DOI: 10.1002/cmdc.200800290

Similarity-Based Virtual Screening with a Bayesian Inference Network

Ammar Abdo* and Naomie Salim[a]

Many methods have been developed to capture the biological similarity between two compounds for use in drug discovery. A variety of similarity metrics have been introduced, the Tanimoto coefficient being the most prominent. Many of the approaches assume that molecular features or descriptors that do not relate to the biological activity carry the same weight as the important aspects in terms of biological similarity. Herein, a novel similarity searching approach using a Bayesian inference network is discussed. Similarity searching is regarded as an inference or evidential reasoning process in which the probability that a given compound has biological similarity with the query is estimated and used as evidence. Our experiments demonstrate that the similarity approach based on Bayesian inference networks is likely to outperform the Tanimoto similarity search and offer a promising alternative to existing similarity search approaches.

Introduction

Similarity searching is one of the most widely used virtual screening approaches. The basic idea underlying similarity searching is the similar property principle which states that structurally similar molecules tend to have similar properties.[1] Over the years, many types of similarity measures have been introduced,[2,3] with similarity measures based on the number of sub-structural fragments common to a pair of molecules and a simple association coefficient being the most common.[4,5] There is an extensive and continuing debate about what sorts of measures are most appropriate.[5] The performance of different similarity coefficients with regard to their use in molecular similarity searching has earlier been analyzed.[4–8] Several methods have been used to further optimize the measures of similarity between molecules, which include weighting,[9] standardization,[10] and data fusion.[4,5,11] Probability-based similarity searching has also been developed on top of the industry-standard vector-space models.[12]

Many studies reported that Bayesian classifiers can improve the enrichment of virtual screening of chemical databases.[13–21] For instance, Bender et al.[14] introduced a novel technique to describe the similarity of molecules, by combining atom environments, information-gain-based feature selection, and a naïve Bayesian classifier. Molecules with the highest probability values are most likely to belong to the active class. Molecules with the lowest probability values are most likely to belong to the inactive class. This approach performed as well as the best commonly used 2D algorithms and outperformed all 3D methods when using single molecule queries. The method obtained close-to-ideal hit rates when multiple molecule queries are used. Klon et al.[15] employed naïve Bayesian classifier to further enrich results from high-throughput docking (HTD). The application of naïve Bayesian classifier was shown to improve the enrichment obtained by HTD alone. In their experiment, they applied consensus scoring after HTD, prior to the application of the naive Bayesian classifier to rescue poor HTD result.[16] Glick et al.[17,18] applied the Laplacian-modified naïve Bayesian to enrich noisy high-throughput screening data when screening in mixtures. They found that the computationally inexpensive Laplacian-modified naïve Bayesian model works surprisingly well, particularly in high levels of noise. Nidhli et al.[19] applied the multiple-category naïve Bayesian model in WOMBAT (World of Molecular BioActivity) database to predict the most likely activities for all compounds in the MDDR database. The objective of this study is to classify targets rather than compounds. In another study, unsupervised clustering using a Bayesian model has been applied by Yan Li[20] for clustering protein conformations sampled during a molecular dynamics simulation of the HIV-1 integrase catalytic core. She found the Bayesian clustering method eliminates the number of clusters in advance, and takes into account the probability distribution of data. Vogt et al.[22] introduced a distance function defined in Bayesian formulations to navigate high-dimensional descriptor space. This function was earlier developed by Godden and Bajorath[23] to address the difficulties associated with the use of high-dimensional chemical spaces for compound classification and ligand-based virtual screening. The distance function ranks compounds according to their distances from the center of a subspace of known active compounds. In other words, the increasing distance from the center of a subspace correlates with decreasing probability that a retrieved compound is active. The Bayesian approach has also been able to directly combine the property descriptors and molecular fingerprints in one framework to search for active compounds.[21]

[a] *A. Abdo, Dr. N. Salim*
Department of Information Systems
Faculty of Computer Science and Information Systems
Uinveristi Teknologi Malaysia, 81310, Skudai, Johor (Malaysia)
Fax: (+6) 7553 2210
E-mail: ammar_utm@yahoo.com

In this study, similarity searching is treated as an inference or evidential reasoning for decision making under uncertainty where probabilistic networks are used for similarity searching. A probabilistic network is a graphical representation of dependence and independence relations between random variables representing a domain. A domain of random variables form the basis of a decision support system to help decision makers identify the best decision in a given situation. A probabilistic network represents and processes probabilistic knowledge between these variables.

A probabilistic network consists of two components, qualitative and quantitative. The qualitative component encodes a set of conditional dependence and independence statements between a set of random variables, information precedence and preference relations using a graphical representation. The quantitative component, on the other hand, specifies the strengths of dependence using probability theory. Probabilistic network uses the graphical representation to describe the knowledge of a problem domain in a precise manner. The graphical representation is intuitive and easy to comprehend, making it an ideal tool for communicating domain knowledge between experts, users, and systems. For these reasons, the formalism of probabilistic networks is becoming an increasingly popular domain knowledge representation for reasoning and decision making under uncertainty.

In this paper, we consider the subclass of probabilistic networks known as Bayesian inference network (BIN). A Bayesian network is a graphical model of a probability distribution.[24] BINs are also known as "belief networks", "causal probabilistic networks" "casual nets" and "graphical probability networks". Many researchers use these networks as a possible solution to problems of decision support under uncertainty. Recent research in information retrieval has proved that retrieval models based on BINs give significant improvements in retrieval performance compared with conventional models.[25–28] In essence, two reasons attracted us to apply these networks models in molecular similarity searching. First, the processes of retrieval of active compounds in a chemical database suffers from uncertainty and incomplete information as in textual information retrieval. Second, these network models allow involvement of opinions, statistical weights and other supplementary information to similarity problems.

BIN is able to represent the main (in)dependence relationships between compound features as conditional probabilities with the degree of resemblance between pairs of such features computed to represent the probability. Similarity searching is regarded as an inference or evidential reasoning process in which the probability that a given compound is similar to a query compound is estimated and used as evidence.

Results and Discussion

Taking into account the different search methods and weighted fingerprint types, there is a total of 12 different similarity procedures available for evaluation. Each procedure was used for each of 12 activity classes from the MDL Drug Data Report[29] (MDDR) database, with ten searches being carried out for the queries in each particular activity class. A different set of ten query structures was used for each of the ten searches. The results of the searches are shown in Tables 1–4, with the first two listing the average recall obtained from the top 5% of the rankings for each activity class using the BIN method, and the next two listing the average recall from the top 5% of the rankings using the Tanimoto similarity method. The mean recalls, averaged over all of the 12 activity classes, are shown in Figure 1. Investigation of the results reported in Tables 1–4 reveals that on average, the BIN method is superior

Table 1. Comparison of the average percentage of active compounds recalled over the top 5% of the ranked test set using the BIN method with EHFC_4, EEFC_4, and ECFC_4.

	BIN					
	EHFC_4		EEFC_4		ECFC_4	
Activity Class	Mean	SD	Mean	SD	Mean	SD
5HT3 antagonists	35.70	2.64	35.99	2.55	34.43	2.94
5HT1A agonists	36.35	5.74	36.61	5.51	35.98	4.64
D2 antagonists	28.23	2.60	28.83	2.44	28.13	1.68
Renin inhibitors	85.08	5.86	84.90	5.49	84.27	6.01
Angiontensin II AT1 antagonists	62.23	5.08	59.21	6.03	58.50	10.69
Thrombin inhibitors	30.98	5.68	30.41	6.34	30.86	9.14
Substance P antagonists	33.32	15.30	33.71	14.88	28.59	14.58
HIV-1 protease inhibitors	43.59	8.97	41.75	10.78	37.05	10.24
Cyclooxygenase inhibitors	18.45	6.32	15.23	6.37	12.84	5.88
Tyrosine protein kinase inhibitors	17.15	9.25	17.30	10.01	20.66	10.88
PAF antagonists	13.33	5.90	13.00	6.22	14.55	7.08
HMG-CoA reductase inhibitors	45.46	15.09	39.95	15.24	35.99	15.56
Average over all classes	37.65	7.80	36.41	7.655	35.15	8.28

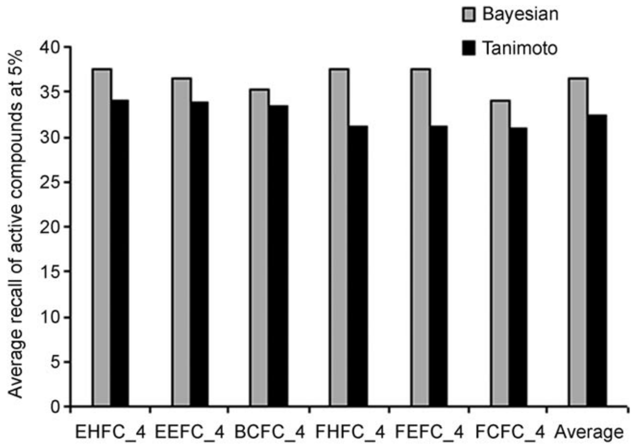

Figure 1. Comparison of the average percentage recalls obtained in the top 5% of the ranked test set using Bayesian inference network and Tanimoto similarity methods.

to the Tanimoto similarity measure regardless of the weighted fingerprint types used. The results reported in Tables 1 and 2 demonstrate that the BIN approach with EHFC_4 produced the highest overall average recall rate relative to other weighted fingerprint types. Similarly, the results in Tables 3 and 4 show that the Tanimoto similarity method with EHFC_4 produced the highest average recall.

Inspection on the results reported in Table 5 show that BIN with EHFC_4 obtained average recall rates of 13–85 %, higher than the Tanimoto similarity method. In only two instances, for activity classes Cyclooxygenase inhibitors and Tyrosine protein kinase inhibitors, BIN produced a considerably lower recall rate than the Tanimoto similarity approach. It is noticeable that these inferior results are associated with the most diverse data sets. Figure 1 graphically shows that BIN outperforms the Tanimoto similarity method over all weighted fingerprint types considered. In addition, the difference between the performance of the Tanimoto similarity method and BIN method remained fairly constant during most of the weighted fingerprint types. Hence, a high recall rate with any activity class in the Tanimoto similarity approach normally reflected the same relative performance using the BIN approach.

Thus far, we have evaluated the various approaches solely in terms of the numbers of active compounds that have been retrieved. It is, however, also important to evaluate the various approaches in terms of their ability to identify a range of different scaffolds, as it is clearly preferable for the outputs to also maximize the number of chemotypes that are identified.

Table 2. Comparison of the average percentage of active compounds recalled over the top 5 % of the ranked test set using the BIN method with FHFC_4, FEFC_4, and FCFC_4.

			BIN			
	FHFC_4		FEFC_4		FCFC_4	
Activity Class	Mean	SD	Mean	SD	Mean	SD
5HT3 antagonists	30.49	2.66	30.45	2.56	29.10	1.73
5HT1A agonists	39.96	6.67	40.02	6.63	40.05	7.43
D2 antagonists	38.17	3.48	37.89	3.47	35.45	2.97
Renin inhibitors	83.86	5.75	84.13	5.38	81.61	7.57
Angiontensin II AT1 antagonists	59.09	8.15	58.76	8.15	55.52	6.65
Thrombin inhibitors	34.03	7.92	34.04	7.46	24.25	8.13
Substance P antagonists	31.87	14.62	31.77	14.63	31.47	19.32
HIV-1 protease inhibitors	30.77	11.74	30.74	11.80	31.61	9.75
Cyclooxygenase inhibitors	19.68	8.23	19.34	8.16	12.91	9.68
Tyrosine protein kinase inhibitors	24.29	10.74	24.29	10.78	18.07	13.10
PAF antagonists	18.24	7.00	18.27	7.09	15.47	9.89
HMG-CoA reductase inhibitors	41.04	18.56	40.98	18.56	32.46	22.67
Average over all classes	37.62	8.79	37.56	8.72	34.00	9.91

Table 3. Comparison of the average percentage of active compounds recalled over the top 5 % of the ranked test set using Tanimoto similarity method with EHFC_4, EEFC_4, and ECFC_4.

			Tanimoto Similarity			
	EHFC_4		EEFC_4		ECFC_4	
Activity Class	Mean	SD	Mean	SD	Mean	SD
5HT3 antagonists	30.93	4.39	31.74	4.78	28.69	4.23
5HT1A agonists	28.77	3.39	28.96	3.52	27.90	3.59
D2 antagonists	24.78	2.62	25.56	2.72	25.11	2.88
Renin inhibitors	75.63	9.95	74.61	9.71	74.32	9.64
Angiontensin II AT1 antagonists	57.32	4.72	58.20	4.84	59.48	4.68
Thrombin inhibitors	19.61	3.66	19.27	3.44	18.97	3.35
Substance P antagonists	35.27	4.99	35.55	4.96	34.98	4.62
HIV-1 protease inhibitors	37.99	3.78	38.48	3.98	40.64	3.03
Cyclooxygenase inhibitors	24.94	1.92	24.45	1.86	23.01	2.02
Tyrosine protein kinase inhibitors	22.18	5.47	21.65	5.68	21.44	5.69
PAF antagonists	14.52	1.45	14.63	1.49	14.63	1.48
HMG-CoA reductase inhibitors	33.26	3.17	33.02	3.38	30.13	3.16
Average over all classes	34.02	4.10	33.84	4.20	33.28	4.03

Table 4. Comparison of the average percentage of active compounds recalled over the top 5 % of the ranked test set using Tanimoto similarity method with FHFC_4, FEFC_4, and FCFC_4.

			Tanimoto Similarity			
	FHFC_4		FEFC_4		FCFC_4	
Activity Class	Mean	SD	Mean	SD	Mean	SD
5HT3 antagonists	24.49	3.65	24.45	3.66	23.19	3.33
5HT1A agonists	28.06	3.02	28.09	3.04	28.75	3.40
D2 antagonists	24.01	3.92	24.21	3.91	23.93	3.88
Renin inhibitors	74.52	7.83	74.51	7.84	74.32	8.27
Angiontensin II AT1 antagonists	37.91	4.58	38.24	4.52	39.68	5.10
Thrombin inhibitors	22.88	3.54	22.88	3.54	20.77	3.18
Substance P antagonists	34.76	4.80	34.69	4.77	33.80	4.58
HIV-1 protease inhibitors	39.00	3.68	38.98	3.67	39.31	3.72
Cyclooxygenase inhibitors	23.85	1.54	23.82	1.56	23.95	1.59
Tyrosine protein kinase inhibitors	18.52	5.50	18.50	5.50	18.86	5.57
PAF antagonists	13.96	2.12	13.98	2.14	13.70	1.90
HMG-CoA reductase inhibitors	30.72	2.42	30.72	2.44	28.84	2.06
Average over all classes	31.06	3.88	31.09	3.88	30.76	3.88

We have considered two measures of the number of different scaffolds retrieved in a similarity search, the number of different atomic frameworks or cyclic systems[30] and molecular frameworks or skeletal cyclic systems.[31,32] These methods were used by Hert et al. to quantify the diversity of hits retrieved in a similarity search.[33]

Figure 2 shows the percentage of the atomic frameworks in the complete set of actives that are retrieved in the top 5% of the rankings for each activity class using the BIN and Tanimoto methods. Similarly, Figure 3 shows the percentage of the molecular frameworks in the complete set of actives that are retrieved in the top 5% of rankings for each activity class using the BIN and Tanimoto methods. Figures 2 and 3 also give the mean number of unique atomic and molecular frameworks found by 12 actives, averaged over all 12 activity classes for each similarity method. It can be observed that the relative performance of the BIN and Tanimoto approaches in terms of retrieving active compounds that belong to different lead series (scaffold hop) mirrors closely the relative performance based on numbers of actives (as shown in Table 5). This finding is in agreement with the study of comparison of topological descriptors by Hert et al.[33] However, in the D2 antagonist activity class, BIN performs better than the Tanimoto similarity method, whereas the D2 antagonist data set is the most diverse measurement using the average number of compounds per scaffold. Thus, there is some evidence that the BIN method is more effective at scaffold hopping for the more diverse data set.

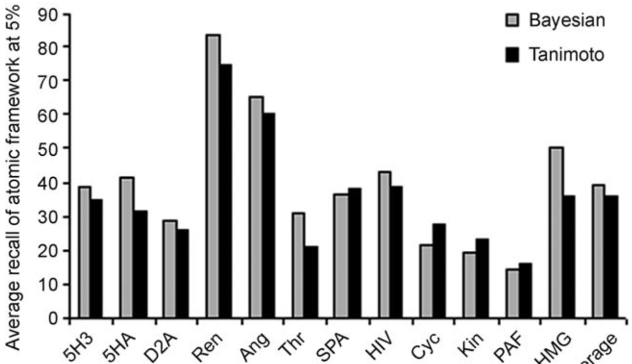

Figure 2. Comparison of the average percentage of atomic frameworks retrieved in the top 5% of the ranked test set using Bayesian and Tanimoto methods with EHFC_4.

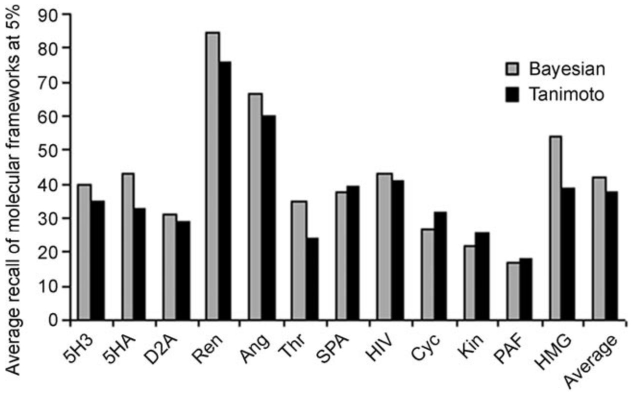

Figure 3. Comparison of the average percentage of molecular frameworks retrieved in the top 5% of the ranked test set using Bayesian and Tanimoto methods with EHFC_4′.

Table 5. Comparison of the average percentage of active compounds retrieved over the top 5% of the ranked test using Bayesian and Tanimoto similarity methods with EHFC_4.

	BIN		Tanimoto Similarity	
Activity Class	Mean	SD	Mean	SD
5HT3 antagonists	35.70	2.64	30.93	4.39
5HT1A agonists	36.35	5.74	28.77	3.39
D2 antagonists	28.23	2.60	24.78	2.62
Renin inhibitors	85.08	5.86	75.63	9.95
Angiontensin II AT1 antagonists	62.23	5.08	57.32	4.72
Thrombin inhibitors	30.98	5.68	19.61	3.66
Substance P antagonists	33.32	15.30	35.27	4.99
HIV-1 protease inhibitors	43.59	8.97	37.99	3.78
Cyclooxygenase inhibitors	18.45	6.32	24.94	1.92
Tyrosine protein kinase inhibitors	17.15	9.25	22.18	5.47
PAF antagonists	13.33	5.90	14.52	1.45
HMG-CoA reductase inhibitors	45.46	15.09	33.26	3.17
Average over all classes	37.65	7.80	34.02	4.10

However, the superiority of the BIN method is ascribed to the ability to understand the content of the compounds and queries and using this understanding to infer the relationship between compounds and queries. This understanding is achieved by applying statistical approaches on descriptors of the entire database and queries, with the information gained from the statistical approach used to assign different weights to individual descriptors.

Finally, the results presented above included only the top 5% of experiments because the conclusions that can be drawn from these results are the same as those that can be drawn from the top 1% of experiments.

Conclusions

One of the disadvantages in simple similarity searching is that molecular features or descriptors that are not related to biological activity carry the same weight as the important ones. To overcome this limitation, we introduced a novel approach based on a Bayesian inference network where the features carry different statistical weights. Features that are statistically less relevant are being de-prioritized. In this study, we look at the similarity searching problem using inference or evidential reasoning and decision making under uncertainty. Overall the results show that the Bayesian inference network method outperforms the Tanimoto similarity method. In addition the Bayesian inference network method is more efficient than the Tanimoto similarity method because the calculations are conducted for the features in common between the compound and query. There was also some evidence to suggest that the Bayesian inference network method is more effective at scaffold hopping, for the more diverse data sets.

Experimental Section

Fingerprint designs

In order to make the evaluation of an approach independent of the characteristics of the specific fingerprint design, we included six different weighted fingerprints in our experiments: atom type extended-connectivity counts (ECFC), functional class extended-connectivity counts (FCFC), atom type atom environment counts (EEFC), functional class atom environment counts (FEFC), atom type hashed atom environment counts (EHFC), and functional class hashed atom environment counts (FHFC) from SciTegic.[34]

Extended-connectivity fingerprint generate higher-order features, with each feature representing the presence of a structural unit. Atom environment counts generate higher-order features using a method developed by Bender et al.[14] Hashed atom environment counts use a hashing algorithm to create an integer representation of the atom environment counts. The only difference in the generation of a functional class or atom type is in the assignment of the initial atom code for each heavy, non-hydrogen atom in the molecule. The initial code assigned to an atom type is based on the number of connections to the atom, the element type, the charge, the atom mass, and the valence. Hence, atoms that differ in any of these features generate a different ECFC, EEFC, and EHFC initial atom code.

For the functional class, the initial atom code is based on the quick estimate of the functional role the atom plays. This role indicates that the atom must be a combination of the hydrogen-bond acceptor, hydrogen-bond donor, positively ionized or positively ionizable, aromatic, and halogen.

These experiments used the ECFC_4, FCFC_4, EEFC_4, FEFC_4, EHFC_4, and FHFC_4, where the numeric code denotes the diameter in bonds up to which features are generated. To make the computational task manageable, we employed a diameter size of four for all weighted fingerprint types in this study, and the fingerprint types are folded to a fixed length of 1024 bits. All the weighted fingerprint types above were generated by Pipeline Pilot software[35] from SciTegic.

Database preparation

For evaluation of the various approaches above, simulated virtual screening searches have been conducted on the MDDR database. After removal of duplicates and molecules that could not be processed using Pipeline Pilot software, a total of 40 751 compounds were available for forming our test database, including 6804 compounds belonging to 12 different activity classes. The activity classes and number of compounds per class are reported in Table 6, which also contains anumeric estimates of the level of structural diversity in each activity class.

Pipeline Pilot software[35] is used to conduct a rigorous search on each of the chosen sets of bioactivity, by matching each compound with every other compounds in its activity class, calculating the class diversity using the ECFP_6 fingerprints, the Tanimoto coefficient, and computing the mean and standard deviation for these intra-set diversities. The resulting diversity scores are listed in Table 6, where it can be observed that the renin inhibitors are the most homogenous and the cyclooxygenase inhibitors are the most heterogeneous.

Table 6. MDDR compound activity classes used in this study.

Code	Activity Class	Actives	Diversity	
			Mean	SD
5H3	5HT3 antagonists	213	0.8537	0.008
5HA	5HT1A agonists	116	0.8496	0.007
D2A	D2 antagonists	143	0.8526	0.005
Ren	Renin inhibitors	993	0.7188	0.002
Ang	Angiontensin II AT1 antagonists	1367	0.7762	0.002
Thr	Thrombin inhibitors	885	0.8283	0.002
SPA	Substance P antagonists	264	0.8284	0.006
HIV	HIV-1 protease inhibitors	715	0.8048	0.004
Cyc	Cyclooxygenase inhibitors	162	0.8717	0.006
Kin	Tyrosine protein kinase inhibitors	453	0.8699	0.006
PAF	PAF antagonists	716	0.8669	0.004
HMG	HMG-CoA reductase inhibitors	777	0.8230	0.002

All compounds in each activity class were reduced to scaffolds (atomic and molecular frameworks) using Pipeline Pilot software.[35] The number of unique atomic and molecular frameworks in each activity class is listed in Table 7, which also contains the average number of molecules per scaffolds, where it can be seen that the HMG-CoA reductase inhibitors are the most homogeneous and the D2 antagonists are the most heterogeneous.

Table 7. The number of unique atomic and molecular frameworks presented in each activity class.

Activity Class	Unique		Average	
	AF[a]	MF[b]	AF	MF
5HT3 antagonists	133	87	1.60	2.45
5HT1A agonists	67	54	1.73	2.15
D2 antagonists	109	75	1.31	1.91
Renin inhibitors	542	328	1.83	3.03
Angiontensin II AT1 antagonists	698	396	1.96	3.45
Thrombin inhibitors	528	335	1.68	2.64
Substance P antagonists	119	78	2.22	3.38
HIV-1 protease inhibitors	455	330	1.57	2.17
Cyclooxygenase inhibitors	83	44	1.95	3.68
Tyrosine protein kinase inhibitors	247	162	1.83	2.80
PAF antagonists	381	252	1.88	2.84
HMG-CoA reductase inhibitors	337	168	2.31	4.63

[a] Unique AF is the number of unique atomic frameworks[30] present in the class. [b] Unique MF is the number of unique molecular frameworks[31,32] present in the class.

For each of the 12 activity classes, 10 different sets of 10 active compounds were randomly selected as query sets. Each search method was repeated 10 times using 10 different query sets for each type of the weighted fingerprint. For each combination of a weighted fingerprint and activity class, the Tanimoto similarity and BIN methods were applied and the percentage of the recall active molecules monitored at the top 1 % and the top 5 % of the ranking list were generated.

The results presented in this study are the mean and standard deviations for these recall values, averaged over each set of the 10 searches. The Tanimoto similarity method was applied in combination with non-binary Tanimoto coefficient to compute the similarity scores. If w_{ik} represents the i^{th} feature of the weighted fingerprint and w_{jk} is the i^{th} feature of the compound to be evaluated, the Tanimoto coefficient is given by:

$$sim(w_i, w_j) = \frac{\sum_{k=1}^{n} w_{ik} w_{jk}}{\sum_{k=1}^{n} (w_{ik})^2 + \sum_{k=1}^{n} (w_{jk})^2 - \sum_{k=1}^{n} w_{ik} w_{jk}} \qquad (1)$$

Similarity inference network model

The basic model for similarity inference network, shown in Figure 4, consists of two component networks: a compound network and a query network. The compound network represents the compound collection. The compound network is built once for a given collection and its structure does not change during query processing. The query network consists of a single node which represents the user's activity requirement and one or more query node representations. A query network is built for each activity required in the query and is modified during query processing as the query is refined or additional representations are added in an attempt to better characterize the activity requirement. The compound and query networks are connected though links between their feature nodes. Each node is binary-valued and takes on one of two values from the set {*true*, *false*}.

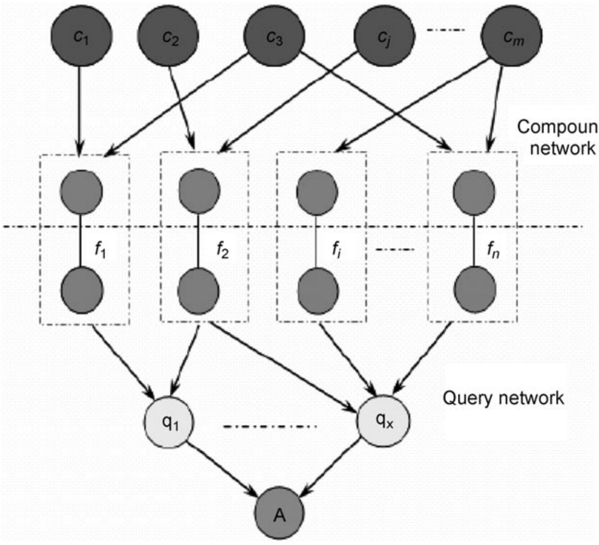

Figure 4. Similarity inference network model.

Compound network

The compound network shown in Figure 4 is a simple direct acyclic graph consisting of compound nodes (c_j) as roots, and feature nodes (f_i) as leaves. Each compound node represents an actual compound in the collection. A compound node corresponds to the event that a specific compound has been observed. Each compound node has one or more feature nodes as children. Each feature node has one or more compound nodes as parents. The feature nodes can be divided into several subsets, each corresponding to a single molecular descriptor type that has been applied to the compound. For example, descriptors that represent properties of whole molecules such as log *P* and molar reactivity, descriptors that can be calculated from 2D graph representations of structures such as topological indices and 2D fingerprints, and descriptors such as pharmacophore keys that require 3D representations of structures. For simplicity, we only consider a weighted 2D fingerprint in which each feature is being weighted by the frequency of its occurrence in the molecule. Therefore, the number of the feature nodes corresponds to the length of the molecular descriptors used to characterize the compound. For instance, 1052 feature nodes are needed for BCI 1052-bits fingerprint.[36]

We represent the assignment of a specific feature to a compound by drawing a directed arc to the feature node from the compound node. In this case, the presence or absence of a link corresponds to the binary assignment of features to compounds. Each compound node has a prior probability associated with it that describes the probability of observing that compound. This prior probability will generally be set to *1/*(*collection size*) and this probability will be small for real collections. Each feature node contains a specification of the conditional probability associated with the node given its set of parent compound nodes. This specification incorporates the effect of any weighting scheme (e.g., feature frequency or inverse compound frequency) associated with the feature node.

Query network

The query network is an "inverted" DAG with a single leaf that corresponds to the event that an activity requirement is met and multiple roots that correspond to the features that express the activity requirement. A set of intermediate query nodes may also be used when multiple queries are used to express the activity requirement. Here, we only use a single molecule as a query. The roots of the query network are query features. A single query feature node has a single compound feature node as parent. A query feature node contains a specification of its dependence on a single parent compound feature node. The query feature nodes define the mapping between the features used to represent the compound collection and the features used to describe the query. In our model, the relation between query and compound feature node is 1:1 and completely dependent because the same descriptor is used to describe compound and query.

Weighting scheme

A weighting Scheme is used to differentiate between different features in a molecule, based on how important they are in determining the similarity of that molecule with another molecule. Certain molecular features can be emphasized by associating higher weights to them when calculating similarity.

Different types of statistical information can be extracted from computerized representations of molecules to form the basis for a feature weighting scheme. These are as follow, (a) Feature Frequency (*ff*), the number of occurrences of a particular feature within a compound, with more frequently occurring features being given greater weights than those that occur less frequently. (b) Inverse Compound Frequency (*icf*), the frequency of the feature in the whole compound collection, with less frequently occurring features being given a greater weight than those that occur more frequently throughout the molecule collection. (c) Compound size (compound length), the number of features assigned to a compound, with features in a smaller compound being assigned a greater weight than the same features in a larger compound. The assignment of weights has been used at the National Cancer Institute.[37] Willett and Winterman[7] found that giving more weight to features that occur more frequently in a molecule did seem to give good results and other weighting schemes had little significance.

Interpretation of inference network

The conditional probability and the Bayes rule play a central role in our inference model. The topology of the inference network model is intended to capture all of the significant probabilistic dependencies among the variables represented by nodes in the entire network. Once the Bayesian network has been created, it can be used to predict the values that certain variables can take. Given the prior probabilities associated with the compounds and the conditional probabilities associated with the interior nodes, we can compute the posterior probability associated with each node in the network.

The main aim of this model is to obtain the probability of biological similarity of each compound in the collection to a given query. When the query network is first built and attached to the compound network, we compute the belief associated with each node in the query network. The initial value at the node representing the activity requirement is the probability that the activity requirement is met given that no specific compound in the collection has been observed to be more similar to the query relative to the other compounds. If a single compound c_j is instantiated and evidence is attached to the network asserting $c_j = true$ with all remaining compound nodes set to *false*, we can compute a new belief for every node in the network given $c_j = true$. In particular, we can compute the probability that the activity requirement is met given that c_j has been observed in the collection. We can now remove this evidence and observe another compound c_k, where $k \neq j$. By repeating this process, we can compute the probability that the activity requirement is met given each compound in the collection and then rank the compounds accordingly. Here, we consider only compounds in isolation for simplicity reasons. The compound network is built once for a given collection. Given one or more queries representing the activity requirement, we then build a query network that attempts to characterize the dependence of the activity requirement on the collection.

Encoding the probabilities using link matrices

Once the structure of the network has been created, the information will be propagated toward the node representing the activity requirement. The process of propagation is known as inference. For this process, we need to estimate the strength of the relationships represented by the network. This process involves estimating and encoding a set of conditional probability distributions. For any of the non-root nodes ***A*** in the network, where ***A*** has a set of parent nodes $\{P_1, P_2,\ldots,P_n\}$, we must estimate and encode $P(\boldsymbol{A}|P_1,P_2,\ldots,P_n)$. The conditional probability can be estimated by many types of weighting schemes. This estimation can be encoded using link matrix form. Unfortunately, the evaluation of link matrix for node ***A*** with *n* parents require $O(2^n)$ floating-point operation and space for all combination of parent values. To overcome the computational complexity, we use canonical link matrix forms[24,25] to encode this estimate, so that the space and time complexity is reduced to $O(n)$.

More specifically, we use the *weighted-sum* canonical link matrix to implement a variety of weighting schemes, including feature frequency, inverse compound frequency, compound size or any combination of them. We assign a weight to the child node ***A***, which is, in essence, the maximum belief that can be associated with that node. Moreover, weights are also assigned to its parents, reflecting their influence on the child node. Consequently, our belief in the node depends on the specific parents that are true. To illustrate how *weighted-sum* link matrix can implement various weighting schemes, let node ***A*** have only three parents P_1, P_2, and P_3 and let w_1, w_2 and w_3, be the parent weights, and let w_A be the child weight ***A*** and $\mathrm{P}(P_1 = true) = p_1$, $\mathrm{P}(P_2 = true) = p_2$, and $\mathrm{P}(P_3 = true) = p_3$, then the full 2×2^n link matrix L_A is as follows:

$$L_A = \begin{bmatrix} 1 & 1-w_3w_A & 1-w_2w_A & 1-(w_2+w_3)w_A & 1-w_1w_A & 1-(w_1+w_3)w_A & 1-(w_1+w_2)w_A & 1-(w_1+w_2+w_3)w_A \\ 0 & w_3w_A & w_2w_A & (w_2+w_3)w_A & w_1w_A & (w_1+w_3)w_A & (w_1+w_2)w_A & (w_1+w_2+w_3)w_A \end{bmatrix}$$

In this representation the values of the first row corresponds to the case that ***A*** = *false* and the second row corresponds to ***A*** = *true*. We use the binary representation of the column number to index the values of the parents, so that the highest order bit reflects the value of the first parent, the second highest order bit the value of the second parent and so on.

Evaluation of this link matrix form results in the following.

$$\begin{aligned} P(A = true) &= w_3 w_A \bar{p}_1 \bar{p}_2 p_3 + w_2 w_A \bar{p}_1 p_2 \bar{p}_3 + (w_2 + w_3) w_A \bar{p}_1 p_2 p_3 \\ &\quad + w_1 w_A p_1 \bar{p}_2 \bar{p}_3 + (w_1 + w_3) w_A p_1 \bar{p}_2 p_3 \\ &\quad + (w_1 + w_2) w_A p_1 p_2 \bar{p}_3 + (w_1 + w_2 + w_3) w_A p_1 p_2 p_3 \\ &= (w_1 p_1 + w_2 p_2 + w_3 p_3) w_A \\ P(A = false) &= 1 - (w_1 p_1 + w_2 p_2 + w_3 p_3) w_A \end{aligned} \tag{2}$$

In the case of a node ***A*** having *n* parents, the link matrix at Equation 2 become **NP** hard, therefore the derived link matrix can be evaluated using the following closed form expression:

$$bel(A) = w_A \sum_{i=1}^{n} (w_i p_i) \tag{3}$$

To illustrate how the *weighted-sum* canonical link matrix can be used to implement a variety of weighting schemes, let ***A*** be a feature node, and P_1, P_2, and P_3 be the compound nodes. Let w_1, w_2 and w_3 be the *ff* values for P_1, P_2, and P_3, and w_A is the *icf* value for ***A***. Given our model, when compound P_1 is instantiated, belief in ***A*** can be given by

$$\begin{aligned} bel(A) &= w_1 w_A \\ &= ff_1 \times icf_A \end{aligned} \tag{4}$$

Similarly, when P_2 is instantiated, $bel(\boldsymbol{A}) = ff_2 \times icf_{\boldsymbol{A}}$. Equation 4 is appropriate for estimating feature probabilities, in which only one compound is instantiated at a time. In general, when a compound is instantiated all feature nodes to which it is attached take on the *ff* × *icf* weight associated with the compound/feature pair. Consequently, each feature node has one compound node as its parent at a given time. Further, in case node ***A*** has a small number *n* of parents (less 5 or 6), the full link matrix could be employed instead of using evaluated canonical link matrix forms.

Probability estimation

Given the link matrix form, we need to provide estimates that characterize the dependence of the random variables in our model. The roots in Figure 4 are compound nodes, with the prior probability associated with these nodes set to 1/(collection size). Estimates are required for three different types of nodes: features, query and activity requirement.

Feature nodes

Compound and query feature nodes are viewed as identical under the assumption that the user knows the set of compound features and can formulate queries using the compound features directly by using similar molecular descriptors. For the features involved in compound and not in query, we assign "false" beliefs, to achieve the identical assumption. To estimate the probability that a feature node is good for discriminating a chemical compound's structure, different weighting function can be incorporated in the estimation equation. We use two different types of equations for this estimate. This estimate is given by the following equation:

$$\begin{aligned} &P(f_i|c_j = true) = \alpha + (1-\alpha) \times nff_{ij} \times nicf_i \\ &P(f_i|allparentfalse) = 0 \end{aligned} \tag{5}$$

where α is a constant and experiments using the inference network show that the best value for α is 0.4,[25,38] nff_{ij} is the normalized frequency of the i^{th} feature within the j^{th} compound and $nicf_i$ is the normalized inverse compound frequency of the i^{th} feature in the collection.

$$P(f_i|c_j = true) = \alpha + (1-\alpha) \times \frac{ff_{ij}}{ff_{ij} + 0.5 \times 1.5 \times \frac{cl_j}{avg\_cl_j}} \times \frac{\log\left(\frac{m+0.5}{icf_i}\right)}{\log(m+1.0)} \tag{6}$$

where ff_{ij} is the frequency of the i^{th} feature within the j^{th} compound, icf_i is the inverse compound frequency of the i^{th} feature in the collection, cl_j is the size of the j^{th} compound, avg_cl_j is the average compound size (over collection), and m is the number of compounds in the collection. In essence, Equation (6) is similar to Equation (5) but has been adapted from the equation developed and used in the Okapi retrieval system.[39,40] Empirical result shows that, Equation (5) and (6) yield results that are almost the same, with slight improvement when Equation (6) is used.

Query node

We need to encode the dependency of each query formulation upon the feature nodes. To encode this probability, we use weighted-sum canonical link matrix forms, as described in Equation (3). By using weighted-sum canonical link matrix forms, we can assign a weight to each of the n parents of the query node, reflecting their influence on the query node. The parents with larger weights have more influence on our belief bel(q). The belief in the query node is then determined by the parents that are true and evaluated as

$$bel(q_k|f_{i..n} = true) = \frac{c_{jk}}{cl_j} \times \sum_{i=1}^{n} (nff_{ik} \times nicf_i \times p_i) \tag{7}$$

where c_{jk} is the set of feature in common between k^{th} query and j^{th} compound (always, $n = c_{jk}$), cl_j is the size of j^{th} compound, nff_{ik} is the normalized frequency of the i^{th} feature within k^{th} query, $nicf_i$ is the normalized inverse compound frequency of the i^{th} feature in the collection and p_i is the estimated probability at the i^{th} feature node.

Activity requirement node

Since we considered only a single query in this work, the activity requirement node coincides with the query node. However, if the user's activity requirement is expressed by multiple queries, the activity requirement node can be viewed as a way of forming a query that is a composite of the individual query formulations connecting to it. These can be combined using a weighted-sum link matrix as before, with weights expressing the importance of each query.

Finally, for feature i that occur ff_i times in the entire collection, the normalized feature frequency (ff_{ij}) and the normalized inverse compound frequency ($nicf$) are given by

$$nff_{ij} = \frac{ff_{ij}}{Max\_ff_j} \qquad nicf_i = \frac{\log(\frac{collectionsize}{ff_i})}{\log(collectionsize)} \tag{8}$$

where Max_ff_j is the maximum feature frequency value of any feature in the j^{th} compound.

Keywords: Bayesian networks · drug discovery · inference networks · similarity searching · virtual screening

[1] M. A. Johnson, G. M. Maggiora, *Concepts and Application of Molecular Similarity*, Wiley, New York, **1990**.

[2] P. Willett, J. M. Barnard, G. M. Downs, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.

[3] A. Bender, R. C. Glen, *Org. Biomol. Chem.* **2004**, *2*, 3204–3218.

[4] J. D. Holliday, C. Y. Hu, P. Willett, *Comb. Chem. High Throughput Screening* **2002**, *5*, 155.

[5] N. Salim, J. Holliday, P. Willett, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 435–442.

[6] G. W. Adamson, J. A. Bush, *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 55–58.

[7] P. Willett, V. Winterman, *Quant. Struct. Activ. Relat.* **1986**, *5*, 18–25.

[8] C. Cheng, G. Maggiora, M. Lajiness, M. Johnson, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 909–915.

[9] G. M. Downs, A. R. Poirrette, P. Walsh, P. Willett, *Chemical Structures 2: The International Language of Chemistry: Proceedings of the Second International Conference* (Berlin, Heidelberg), **1993**.

[10] P. A. Bath, C. A. Morris, P. Willett, *J. Chemom.* **1993**, *7*, 543–550.

[11] N. Daut, R. Mohemad, N. Salim, *3rd International Conference on Artificial Intelligence in Engineering and Technology* (Sabah, Malaysia), **2006**.

[12] N. Salim, W. W. P. Godfrey, *J. Advancing Inf. Manage. Stud.* **2005**, *2*, 56–74.

[13] X. Xia, E. G. Maliski, P. Gallant, D. Rogers, *J. Med. Chem.* **2004**, *47*, 4463–4470.

[14] A. Bender, H. Y. Mussa, R. C. Glen, S. Reiling, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 170–178.

[15] A. E. Klon, M. Glick, M. Thoma, P. Acklin, J. W. Davies, *J. Med. Chem.* **2004**, *47*, 2743–2749.

[16] A. E. Klon, M. Glick, J. W. Davies, *J. Med. Chem.* **2004**, *47*, 4356–4359.

[17] M. Glick, A. E. Klon, P. Acklin, J. W. Davies, *J. Biomol. Screening* **2004**, *9*, 32–36.

[18] M. Glick, J. L. Jenkins, J. H. Nettles, H. Hitchings, J. W. Davies, *J. Chem. Inf. Model.* **2006**, *46*, 193–200.

[19] Nidhi, M. Glick, J. W. Davies, J. L. Jenkins, *J. Chem. Inf. Model.* **2006**, *46*, 1124–1133.

[20] Y. Li, *J. Chem. Inf. Model.* **2006**, *46*, 1742–1750.

[21] M. Vogt, J. Bajorath, *Chem. Biol. Drug Des.* **2008** *71*, 8–14.

[22] M. Vogt, J. W. Godden, J. Bajorath, *J. Chem. Inf. Model.* **2007**, *47*, 39–46.

[23] J. W. Godden, J. Bajorath, *J. Chem. Inf. Model.* **2006**, *46*, 1094–1097.

[24] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann Publishers Inc., **1988**.

[25] R. T. Howard, PhD Thesis, University of Massachusetts (USA), **1991**.

[26] R. T. Howard, W. B. Croft, *Comput. J.* **1992**, *35*, 279–290.

[27] A. N. R. Berthier, M. Richard, *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Zürich, Switzerland), **1996**.

[28] L. M. de Campos, J. M. Fernández-Luna, J. F. Huete, *Int. J. Approx. Reasoning* **2003**, *34*, 265–285.

[29] The MDL Drug Data Report Database is available from MDL Information Systems Inc. at http://www.mdli.com (accessed November 17, 2008).

[30] G. W. Bemis, M. A. Murcko, *J. Med. Chem.* **1996**, *39*, 2887–2893.

[31] Y. Xu, M. Johnson, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 181–185.

[32] Y. J. Xu, M. Johnson, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 912–926.

[33] J. Hert, P. Willett, D. J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby, A. Schuffenhauer, *Org. Biomol. Chem.* **2004**, *2*, 3256–3266.

[34] SciTegic Accelrys Inc.: http://www.SciTegic.com (accessed November 17, 2008).

[35] Pipeline Pilot Basic Chemistry Component collection, SciTegic Inc.: http://www.SciTegic.com (accessed November 17, 2008).

[36] Barnard Chemical Information Ltd.: http://www.bci.gb.com/ (accessed November 17, 2008).

[37] L. Hodes, *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 66–71.

[38] R. T. Howard, W. B. Croft, *ACM Trans. Inf. Syst.* **1991**, *9*, 187–222.

[39] S. E. Robertson, S. Walker, *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Dublin, Ireland), **1994**.

[40] S. E. Robertson, S. Walker, M. M. Hancock-Beaulieu, *Inf. Process. Manage.* **1995**, *31*, 345–360.

Received: September 7, 2008
Revised: October 28, 2008
Published online on December 12, 2008