Prediction of New Bioactive Molecules using a Bayesian Belief Network

Ammar Abdo, Valérie Leclère, Philippe Jacques, Naomie Salim, Maude Pupin

JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Prediction of New Bioactive Molecules using a Bayesian Belief Network

Ammar Abdo,*[,†,‡] Valérie Leclère,[§] Philippe Jacques,[§] Naomie Salim,[∥] and Maude Pupin[†]

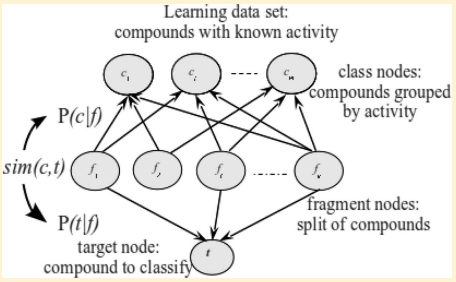
[†]LIFL UMR CNRS 8022 Université Lille1 and INRIA Lille Nord Europe, 59655 Villeneuve d'Ascq cedex, France

[§]ProBioGEM, UPRES EA 1026, Polytech'Lille, Av P. Langevin, Univ Lille 1- Sciences et Technologies, 59655 Villeneuve d'Ascq cedex, France

[‡]Computer Science Department, Hodeidah University, Hodeidah, Yemen

[∥]Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia, 81310, Skudai, Malaysia

ABSTRACT: Natural products and synthetic compounds are a valuable source of new small molecules leading to novel drugs to cure diseases. However identifying new biologically active small molecules is still a challenge. In this paper, we introduce a new activity prediction approach using Bayesian belief network for classification (BBNC). The roots of the network are the fragments composing a compound. The leaves are, on one side, the activities to predict and, on another side, the unknown compound. The activities are represented by sets of known compounds, and sets of inactive compounds are also used. We calculated a similarity between an unknown compound and each activity class. The more similar activity is assigned to the unknown compound. We applied this new approach on eight well-known data sets extracted from the literature and compared its performance to three classical machine learning algorithms. Experiments showed that BBNC provides interesting prediction rates (from 79% accuracy for high diverse data sets to 99% for low diverse ones) with a short time calculation. Experiments also showed that BBNC is particularly effective for homogeneous data sets but has been found to perform less well with structurally heterogeneous sets. However, it is important to stress that we believe that using several approaches whenever possible for activity prediction can often give a broader understanding of the data than using only one approach alone. Thus, BBNC is a useful addition to the computational chemist's toolbox.



INTRODUCTION

Due to the similar property principle,[1] structurally similar compounds are expected to exhibit similar properties and similar biological activities. This principle is exploited for in silico discovery of new drugs with the emergence of an activity prediction technology based on chemical structures. A variety of computational approaches for target or activity prediction were published over the past several years. For example, quantitative structure−activity relationship (QSAR)[2−5] was established on the hypothesis that compounds with similar physicochemical properties and/or structure share similar activities. The effectiveness of a QSAR analysis relies both on selecting the relevant descriptors for modeling the biological activity of interest and on the choice of a good quantitative model that maps the compound descriptors to chemical property or biological activity by means of statistical techniques. In similarity searching strategies, an unknown compound (the target) is compared to a set of compounds with known activities. The activity for which the compounds are the most similar to the target is assigned to it. Binary kernel discrimination (BKD),[6,7] naïve Bayesian classifier (NBC),[8−11] artificial neural networks (ANN),[12−16] and support vector machine (SVM)[17−19] are machine learning methods employed for activity prediction by compound classification. They can only be applied when annotated compounds are available to design a training set divided in an active class (compounds having a specific activity) and an inactive one (compounds belonging to other activities). The training set is then analyzed to develop decision rules that classify new compounds (the test set) into one of the two classes (active or inactive). The learning step, needed to determine the criteria for the classification of an unknown compound is time-consuming because many criteria are explored.

Overall, while target prediction approaches exhibit several successes, some issues need to be addressed. In many studies, different approaches predict a different subset of targets for the same compound.[20−22] Moreover, given that the approach might work better on specific targets or databases in a way that is difficult to predict beforehand, should we ignore approaches with low or similar performance? We think the answer is no because it is hard to predict which approach will do better on a given combination of database and query. Also, each approach may be able to predict specific targets that all others could not. It is important to stress that we believe that using several approaches whenever possible for target prediction can often give a broader understanding of the data than using only one approach alone.[6]

Received: June 24, 2013
Published: January 6, 2014

   dx.doi.org/10.1021/ci4004909 | J. Chem. Inf. Model. 2014, 54, 30−36

Previously, Abdo and Salim[23] developed a ligand-based virtual screening method that used a Bayesian inference network (BIN) for similarity searching. Experiments with a subset of the MDDR and WOMBAT[24] databases demonstrated that the BIN provided an interesting alternative to existing tools for ligand-based virtual screening. It substantially outperformed a conventional Tanimoto-based similarity searching system when the active molecules have a high degree of structural homogeneity. To overcome this limitation, an alternative Bayesian network model called Bayesian belief network (BBN) has been introduced by Abdo et al.[25] In this paper, we introduce a new activity prediction approach using Bayesian belief network, but this time for classification. We suggest that the BBN model introduced by Abdo et al.[25] for similarity searching can be applied to classification of small molecules in activity classes. The purpose of this paper is to introduce BBN model as a useful tool for activity prediction. We show that it provides a useful method for using the prior knowledge of target class information (active and inactive compounds) to predict the activity of orphan compounds. We apply this new approach on eight well-known data sets extracted from literature and compare its performance with three classical machine learning algorithms.

MATERIALS AND METHODS

Data Sets. We evaluate the quality of our prediction model on eight data sets already used to validate the classification of molecules based on structure−activity relationship. The five data sets described in Table 1 are taken from the studies of Sutherland et al.[26] and Helma et al.[27] with compounds classified as active or inactive: (1) 467 cyclooxygenase-2 (COX2) inhibitors, (2) 405 ligands of the benzodiazepine receptor (BZR), (3) 756 dihydrofolate reductase (DHFR) inhibitors, (4) 393 ligands of the estrogen receptor (ER), and (5) 683 mutagens (MUT) of molecular structures. These data sets have been used by many studies for validating prediction models.[2,28−30]

Table 1. Summary of the First Fifth Data Sets[a]

data set	no. compounds		pairwise similarity (mean)	
	active	inactive	active	inactive
cyclooxygenase-2 inhibitors	303	164	0.687	0.690
benzodiazepine receptor	306	99	0.536	0.538
dihydrofolate reductase	393	363	0.502	0.537
estrogen receptor	141	252	0.468	0.456
mutagens	340	343	0.143	0.125

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

The sixth data set (Table 2) of this study is taken from the Norine database (version of August 2013), which contains 1122 peptides with 11 distinct activities.[31] We have restricted our data set to 605 peptides, belonging to five different activity classes.[32] The last two data sets (Tables 3 and 4) are chosen from the MDL Drug Data Report (MDDR)[33] and have been extensively used by many studies for validating ligand-based virtual screening approaches.[34−37] The data sets MDDR1 and MDDR2 contain 10 homogeneous activity classes and 10 heterogeneous ones, respectively. Each row of data set tables (Tables 2−4) contains an activity class, the number of molecules/peptides belonging to the class, and the diversity of classes, which is computed as the mean pairwise Tanimoto similarity calculated across all pairs of molecules/peptides in the class using ECFC4 (extended connectivity) for data sets in Tables 1, 3, and 4 and monomer composition fingerprints (MCFP)[32] in the Norine data set. For Norine and MDDR data sets, each activity class is considered, one by one, as active set and the others as the inactive set.

Table 2. Summary of Norine Data Set

activity class	NRPs number	pairwise similarity (mean)
antibiotics	319	0.09
toxin	157	0.09
siderophore	82	0.18
antitumor	25	0.27
protease inhibitors	22	0.26

Table 3. MDDR Activity Classes for MDDR1 Data Set

activity index	activity class	active molecules	pairwise similarity (mean)
07707	adenosine (A1) agonists	207	0.424
07708	adenosine (A2) agonists	156	0.484
31420	renin inhibitors	1130	0.584
42710	monocyclic beta-lactams	111	0.596
64100	cephalosporins	1346	0.512
64200	carbacephems	113	0.503
64220	carbapenems	1051	0.414
64300	penicillin	126	0.444
65000	antibiotic, macrolide	388	0.673
75755	vitamin D analogous	455	0.569

Table 4. MDDR Activity Classes for MDDR2 Data Set

activity index	activity class	active molecules	pairwise similarity (mean)
09249	muscarinic (M1) agonists	900	0.257
12455	NMDA receptor antagonists	1400	0.311
12464	nitric oxide synthase inhibitors	505	0.237
31281	dopamine beta-hydroxylase inhibitors	106	0.324
43210	aldose reductase inhibitors	957	0.370
71522	reverse transcriptase inhibitors	700	0.311
75721	aromatase inhibitors	636	0.318
78331	cyclooxygenase inhibitors	636	0.382
78348	phospholipase A2 inhibitors	617	0.291
78351	lipoxygenase inhibitors	2111	0.365

All compounds in the first five data sets and MDDR data sets are converted to Pipeline Pilot's ECFC4 fingerprints and folded to a size of 1024.[38] ECFC4 fingerprints have been successfully used by many previous studies. For the Norine data set, all the peptides are converted to MCFP. MCFP is a novel representation we have developed for obtaining an appropriate description of nonribosomal peptides in relation to their activity. So, the compounds of all data sets are represented by fingerprints containing the frequency of each fragment within it. However, it should be noticed that it is not our intention to tackle here the question of which is the most appropriate fingerprint/descriptor.

dx.doi.org/10.1021/ci4004909 | J. Chem. Inf. Model. 2014, 54, 30−36

As classes contain a set of compounds, we calculated a representative fingerprint. For each bit of the fingerprint representing a given fragment f, we have taken the average number of occurrences of f within a class c according to the following formula:

$$\mathrm{av}(\mathrm{nb}_f(c)) = \frac{1}{\mathrm{nb}_{\mathrm{co}}(f, c)} \sum^{\mathrm{nb}_{\mathrm{co}}(c)} \mathrm{nb}_f(\mathrm{co}) \quad (1)$$

where $\mathrm{nb}_{\mathrm{co}}(f, c)$ is the number of compounds containing the fragment f within the class c and $\mathrm{nb}_{\mathrm{co}}(c)$ is the number of compounds in the class c and $\mathrm{nb}_f(\mathrm{co})$ is the number of occurrences of the fragment f in the compound co.

Bayesian Belief Network Classifier Model (BBNC). The BBNC model is based on Bayesian belief network model (for more details about BBN see ref 25) but adapted to activity prediction (classification rather than similarity searching).

The BBNC model, shown in Figure 1, consists of three types of nodes. The roots are the fragment nodes ***f***. Each node

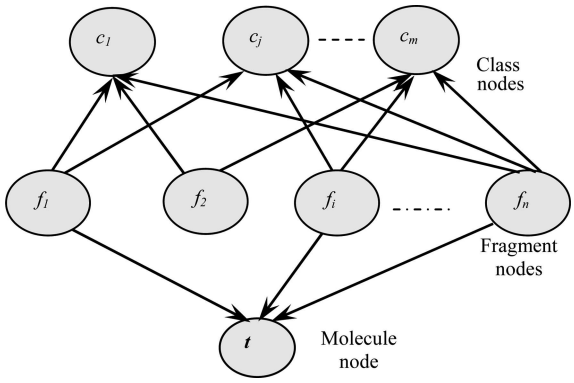


Figure 1. Bayesian belief network classifier model.

represents a bit of the fingerprints generated for the compounds, i.e., the presence/absence of particular chemical substructures. Both class and molecule nodes, the leaves, are described using a set of fragment nodes. The class nodes ***c*** represents the compound collections for the different biological activities that are part of the training set, summarized by the average frequency of formula 1. The first five data sets contain 2 class nodes, 5 class nodes for Norine, and 10 for MDDR1 and MDDR2. The other type of leaf is the molecule node ***t*** representing the molecule for which we want to predict an activity (unknown natural compound or target of an unknown protein). 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 and BBN works. An edge is traced from one fragment node to one class or molecule node if the fragment appears in that class/molecule.

To build a robust model that distinguish efficiently the classes and then give the right prediction, we excluded the very rare and predominant fragments that introduce noise in the data. We determined empirically the rare fragments as the ones covering less than 3% of the active molecules in the training set and the predominant ones covering more than 50% of the active set and 70% of the inactive one. The corresponding fragment nodes are set to 0.

To construct our belief network classifier, we needed to estimate the strength of the relationships represented by the network by encoding a set of conditional probability distributions. Specifically, for any nonroot node A in the network, where A has a set of parent nodes $\{P_1, P_2, ..., P_n\}$, we must estimate the probability $P(A|P_1, P_2, ..., P_n)$, i.e., we need to specify the conditional probability for the nonroot nodes and $P(t|f)$. We adapted the probabilities to the specificity of chemical data. Our weighting scheme is derived of the InQuery system and has been used successfully in our previous studies. Here, it is proposed to use the following probabilities $P(c|f)$ for each fragment in a class c or molecule t vector.

$$P(c|f) = \alpha + (1 - \alpha) \times \left(\frac{\mathrm{av}(\mathrm{nb}_f(c))}{\max(\mathrm{av}(\mathrm{nb}_f(c)))} \times \frac{\mathrm{nb}_{\mathrm{co}}(f, c)}{\mathrm{nb}_{\mathrm{co}}(c)} \right) \times \frac{\log\left(\frac{\mathrm{nb}_{\mathrm{co}}(\overline{c}) + 0.5}{\mathrm{nb}_{\mathrm{co}}(f, \overline{c})} \right)}{\log(\mathrm{nb}_{\mathrm{co}}(\overline{c}) + 1.0)} \quad (2)$$

$$P(t|f) = \alpha + (1 - \alpha) \times \left(\frac{\mathrm{nb}_f(t)}{\max(\mathrm{nb}_f(t))} \times \frac{\mathrm{nb}_{\mathrm{co}}(f, c)}{\mathrm{nb}_{\mathrm{co}}(c)} \right) \times \frac{\log\left(\frac{\mathrm{nb}_{\mathrm{co}}(\overline{c}) + 0.5}{\mathrm{nb}_{\mathrm{co}}(f, \overline{c})} \right)}{\log(\mathrm{nb}_{\mathrm{co}}(\overline{c}) + 1.0)} \quad (3)$$

where

- ***α*** is a constant (which was set to 0.4 in our preliminary experiments) that avoids the probabilities being equal to 0 for fragments not observed in the data set which is only a sample of the existing compounds.
- $\mathrm{av}(\mathrm{nb}_f(c))$ and $\mathrm{nb}_f(t)$ are, respectively, the average number of occurrences of the fragment f within the class c (see (1)) and the number of occurrences of the fragment f within the molecule t (only one molecule).
- $\mathrm{nb}_{\mathrm{co}}(f, c)$ and $\mathrm{nb}_{\mathrm{co}}(f, \overline{c})$ are the numbers of compounds containing the fragment f within the class c or within the other classes $\overline{c}$ representing the inactive training set.
- $\mathrm{nb}_{\mathrm{co}}(c)$ and $\mathrm{nb}_{\mathrm{co}}(\overline{c})$ are the numbers of compounds in, respectively, the class c and the other classes $\overline{c}$.

The ratio between $\mathrm{nb}_{\mathrm{co}}(f, c)$ and $\mathrm{nb}_{\mathrm{co}}(c)$ is the fragment f frequency among the class c. It modulates the weight of the fragments by decreasing it when f is rare and increasing it when f is ubiquitous within the class. At the same time, the ratio of logs is near 1 when f is rare and near 0 when f is ubiquitous outside the class c. This increases the probability of the fragments that are specific of a class c.

Once the probabilities of each nodes of the network are obtained, we calculate a similarity between the target t and each of the classes by

$$\mathrm{sim}(c, t) = \frac{\mathrm{nb}(f, c) \cap \mathrm{nb}(f, t)}{\min(\mathrm{nb}(f, c), \mathrm{nb}(f, t))} \times \sum^{n} \left((1 - |P(c|f) - P(t|f)|) / \frac{\mathrm{nb}_{\mathrm{co}}(f, \overline{c})}{\mathrm{nb}_{\mathrm{co}}(\overline{c})} \right) \quad (4)$$

The sum concerns the n fragments between the class node and the target node and $\mathrm{nb}(f, c)$ and $\mathrm{nb}(f, t)$ are the numbers of single fragments in the tested class c and the molecule t, respectively.

Here, $\mathrm{nb}(f, c) \cap \mathrm{nb}(f, t)$ is the set of fragments in common between the tested class and the molecule. If $\mathrm{nb}_{\mathrm{co}}(f, \overline{c})$ is equal to zero (meaning that this fragment never occurs in an inactive training set), then the $\mathrm{nb}_{\mathrm{co}}(f, \overline{c})$ in eq 4 is substituted with a

dx.doi.org/10.1021/ci4004909 | J. Chem. Inf. Model. 2014, 54, 30−36

Table 5. Sensitivity, Specificity, AUC, and Accuracy Rates for the Prediction Models for Different Data Sets

	BBNC				NaïveB				RBFN				LSVM			
activity class	Sens	Spec	AUC	Acc	Sens	Spec	AUC	Acc	Sens	Spec	AUC	Acc	Sens	Spec	AUC	Acc
COX2	0.99	0.99	0.99	98.19	1.00	0.99	1.00	99.79	0.93	0.76	0.84	86.94	1.00	1.00	1.00	100.00
BZR	0.99	0.91	0.95	92.11	0.94	0.61	0.78	86.17	0.99	0.65	0.82	90.62	1.00	1.00	1.00	100.00
DHFR	1.00	0.98	0.99	94.00	0.99	1.00	0.99	99.47	0.86	0.80	0.84	83.86	1.00	1.00	1.00	100.00
ER	1.00	0.90	0.95	89.13	1.00	1.00	1.00	100.00	0.62	0.70	0.64	64.89	1.00	1.00	1.00	100.00
MUT	0.74	0.85	0.80	78.59	0.59	0.84	0.72	71.74	0.69	0.50	0.59	59.00	0.72	0.71	0.71	71.60
mean	0.94	0.93	0.93	90.40	0.90	0.89	0.90	91.43	0.82	0.68	0.75	77.06	0.94	0.94	0.94	94.32

Table 6. Sensitivity, Specificity, AUC, and Accuracy Rates for the Prediction Models with the Norine Data Set

	BBNC			NaïveB			RBFN			LSVM		
activity class	Sens	Spec	AUC	Sens	Spec	AUC	Sens	Spec	AUC	Sens	Spec	AUC
antibiotics	0.96	0.91	0.94	0.88	0.77	0.82	0.90	0.96	0.93	0.96	0.95	0.96
toxin	0.90	0.97	0.94	0.66	0.92	0.79	0.87	0.93	0.90	0.92	0.97	0.95
siderophore	0.99	0.99	0.99	0.94	0.99	0.97	0.96	0.99	0.98	0.99	1.00	0.99
antitumor	0.23	0.99	0.61	0.48	0.97	0.48	0.60	0.97	0.79	0.64	0.98	0.81
protease inhibitors	0.92	1.00	0.96	0.73	0.97	0.49	0.82	0.99	0.90	0.91	1.00	0.95
mean	0.80	0.97	0.89	0.74	0.92	0.71	0.83	0.97	0.90	0.88	0.98	0.93
accuracy	92.10			80.99			88.60			93.88		

very small value given by $1/(\mathrm{nb}_{co}(c) + \mathrm{nb}_{co}(\overline{c}))$. However, it should be noted that the calculation was carried out only for nodes representing common fragments between the molecule and the tested class. Once the similarity calculations against each class are completed, the target molecule is assigned to the most similar class.

BBNC is designed for classification while BIN and BBN was used for similarity searching. The main difference between those methods is the nature of the network. In BBNC, the leaves are, on one hand, the classes composed of numerous compounds extracted from the training data sets and, on the other end, the target compound; whereas in BBN each leaf represents one compound of the data set and the target. So, in BBNC the number of leaves corresponds to the number of biological activity classes (generally less than hundred); whereas, in BIN and BBN the number of leaves equals the total number of compounds in the data set (generally several thousands or more). Another difference is the calculation of the belief of the fragments that is used for similarity searching in BIN and BBN and for classification in BBNC.

Machine Learning Algorithms. We have compared our method to three machine learning algorithms available in WEKA-Workbench:[39] The naive Bayesian classifier (NaiveB),[40] the support vector machine classifier, called LibSVM (LSVM),[41] and the neural network classifier (RBFN).[42] Details of these algorithms can be found in their references. Finding the optimal parameters for a classifier is somewhat a tedious process. However, WEKA-Workbench offers the possibility of automatically finding the best possible setup for LSVM classifier. Here, LSVM was used with linear kernel and the following values: 0.1, 1.0, and 0.001 were assigned to the Gamma, Cost, and Epsilon parameters, respectively. We used supervised discretization to convert numeric attributes to nominal ones in NaiveB classifier and the minimum standard deviation parameter was set to 0.01 in RBFN classifier. The remaining parameters were left as they are in default setup for each classifier in WEKA-Workbench.

Validation. Ten-fold cross-validation was used to validate the results of BBNC and the other machine learning algorithms. In this cross-validation, the data set was split into 10 parts; 9 were used for training and the remaining 1 was used for testing. This process was repeated 10 times, so all the compounds were used in the test set once. Each activity class was tested against all the others, grouped. As in the case of many prediction methods, we used the area under the reciever operating characteristic (ROC) curve (AUC) as quality criterion to quantify the performance of classification algorithms. AUC is defined as (sensitivity + specificity)/2. The sensitivity is defined by sens = tp/(tp + fn) and specificity is defined by spec = tn/(tn + fp), where tp, tn, fp, and fn are the number of true positives, true negatives, false positives, and false negatives, respectively. A ROC curve describes the trade-off between sensitivity and specificity, where the sensitivity and specificity are defined as the effectiveness of a model to identify positive and negative labels, respectively. The area under the ROC curve (AUC) is a measure of the model performance: the closer to 1, the better the performance of the prediction. We also used an accuracy (ac) measurement to quantify the performance of the classification models. Accuracy is the overall effectiveness of a model and is calculated as the sum of correct classifications divided by the total number of classifications ac = (tp + tn)/(tp + tn + fp + fn).

RESULTS AND DISCUSSION

The main aim of this study was to introduce the Bayesian belief network classifier approach into ligand-based target fishing or activity prediction and then identify the prediction accuracy of such an approach. To achieve this aim, the BBNC approach was compared with three machine learning algorithms available in WEKA-Workbench: NaiveB, LSVM, and RBFN, used with optimal parameters.

Visual inspection of the AUC and accuracy in Tables 5−8 enabled comparing the performances of the four prediction algorithms. However, a more quantitative approach is possible using the Kendall W test of concordance.[43] This test was developed to quantify the level of agreement between multiple sets ranking the same set of objects. Here, we used this approach to rank the performance of the tested algorithms. In the present context, the data sets in Table 5 and the activity classes in Tables 6−8 were considered as judges and the AUC

dx.doi.org/10.1021/ci4004909 | J. Chem. Inf. Model. 2014, 54, 30−36

Table 7. Sensitivity, Specificity, AUC, and Accuracy Rates for the Prediction Models with the MDDR1 Data Set

	BBNC			NaïveB			RBFN			LSVM		
activity index	Sens	Spec	AUC	Sens	Spec	AUC	Sens	Spec	AUC	Sens	Spec	AUC
07707	1.00	1.00	1.00	0.99	1.00	0.99	0.63	1.00	0.82	0.98	1.00	0.99
07708	1.00	1.00	1.00	0.97	1.00	0.99	0.51	1.00	0.75	0.98	1.00	0.99
31420	0.99	1.00	1.00	1.00	1.00	1.00	0.96	0.96	0.96	1.00	1.00	1.00
42710	0.95	1.00	0.98	0.94	1.00	0.97	0.43	1.00	0.72	0.99	1.00	1.00
64100	0.99	0.99	0.99	0.95	1.00	0.97	0.97	0.90	0.94	1.00	1.00	1.00
64200	0.97	1.00	0.99	0.87	0.99	0.93	0.43	1.00	0.71	0.93	1.00	0.96
64220	0.98	1.00	0.99	1.00	1.00	1.00	0.95	0.97	0.96	1.00	1.00	1.00
64300	1.00	1.00	1.00	1.00	0.99	1.00	0.44	1.00	0.72	0.98	1.00	0.99
65000	1.00	1.00	1.00	1.00	1.00	1.00	0.80	1.00	0.90	1.00	1.00	1.00
75755	0.99	1.00	1.00	0.99	1.00	1.00	0.76	1.00	0.88	1.00	1.00	1.00
mean	0.99	1.00	0.99	0.97	1.00	0.98	0.69	0.98	0.84	0.99	1.00	0.99
accuracy	99.10			97.88			86.84			99.43		

Table 8. Sensitivity, Specificity, AUC, and Accuracy Rates for the Prediction Models with the MDDR2 Data Set

	BBNC			NaïveB			RBFN			LSVM		
activity index	Sens	Spec	AUC	Sens	Spec	AUC	Sens	Spec	AUC	Sens	Spec	AUC
09249	0.89	0.98	0.93	0.91	0.99	0.95	0.82	0.98	0.90	0.98	1.00	0.99
12455	0.69	0.98	0.83	0.88	0.97	0.92	0.66	0.98	0.82	0.96	0.99	0.98
12464	0.75	0.98	0.86	0.85	0.99	0.92	0.75	0.95	0.85	0.92	1.00	0.96
31281	0.91	1.00	0.96	0.94	1.00	0.97	0.53	1.00	0.76	0.98	1.00	0.99
43210	0.86	0.95	0.91	0.84	0.99	0.91	0.78	0.97	0.87	0.96	0.99	0.98
71522	0.94	0.87	0.90	0.82	0.99	0.91	0.75	0.97	0.86	0.94	1.00	0.97
75721	0.95	0.98	0.97	0.91	0.99	0.95	0.86	0.98	0.92	0.99	1.00	0.99
78331	0.71	0.97	0.84	0.81	0.96	0.89	0.79	0.93	0.86	0.84	0.99	0.91
78348	0.85	0.93	0.89	0.65	0.99	0.82	0.74	0.96	0.85	0.91	1.00	0.95
78351	0.68	0.98	0.83	0.82	0.94	0.88	0.59	0.96	0.78	0.94	0.98	0.96
mean	0.82	0.96	0.89	0.84	0.98	0.91	0.73	0.97	0.85	0.94	0.99	0.97
accuracy	80.40			83.99			71.23			94.37		

and accuracy measurement of the various prediction algorithms as objects. The outputs of the test are the value of the Kendall coefficient, ranging from 0 (no agreement between set of ranks) to 1 (complete agreement), and the associated significance level, which indicates whether this value of the coefficient could have occurred by chance. If the value is significant (we use cutoff values of 0.01 or 0.05), then it is possible to give an overall ranking of the objects that have been ranked. The results of the Kendall analysis are reported in Table 9 and give the ranking for the various prediction algorithms.

We first considered the results of the first five data sets (COX2, BZR, DHFR, ER, and MUT). The results reported in Table 5 showed that the LSVM produces the highest sensitivity, specificity, AUC, and accuracy values compared to the other approaches, with BBNC and NaiveB also performing well. Table 9 showed that the value of the Kendall coefficient, 0.549, was significant at the 0.05 cutoff value. So, we can conclude that the overall ranking of the four methods is: LSVM > BBNC = NaiveB > RBFN. Here, the results in Table 9 suggested that the BBNC has the second best performance after LSVM.

The good performance for LSVM and BBNC approaches was not restricted to the first five data sets since they also perform best for the Norine and MDDR1 data sets (Tables 6 and 7). The results in Table 6 showed that LSVM and BBNC produced the best performance across the five activity classes in the Norine data set. In only one instance (antitumor class), the performance of BBNC was not satisfying (Sens = 0.23 and AUC = 0.61). This finding was in accordance to our previous study.[32] This is because the peptides with antitumor activity are biologically closed to antibiotic and toxin peptides and can even harbor several of those activities. Table 9 showed that the value of the Kendall coefficient, 0.845, was significant at the 0.01 level; we can conclude that the overall ranking of the different methods is: LSVM > BBNC > RBFN > NaiveB.

Table 9. Rankings of Prediction Approaches Based on Kendall W Test Results[a]

data set	W	P	ranking
AUC (DS1−5)	0.549	<0.05	LSVM > BBNC = NaiveB > RBFN
AUC (Norine)	0.845	<0.01	LSVM > BBNC > RBFN > NaiveB
AUC (MDDR1)	0.751	<0.01	BBNC > LSVM > NaiveB > RBFN
AUC (MDDR2)	0.825	<0.01	LSVM > NaiveB > BBNC > RBFN
Accuracy (all Data sets)	0.622	<0.05	LSVM > BBNC > NaiveB > RBFN

[a]The contents of the columns above show the different performance measure for data set type, the value of the Kendall coefficient, the associated significance probability, and the ranking of the methods, respectively. The methods are ranked in decreasing order of prediction performance.

The best performance of BBNC was seen with MDDR1 (Table 7). Sensitivity, specificity, and AUC values in Table 7 showed that BBNC and LSVM produced the best performance compared to other approaches. Table 9 showed that the value of the Kendall coefficient, 0.751, is significant at the 0.01 level

dx.doi.org/10.1021/ci4004909 | J. Chem. Inf. Model. 2014, 54, 30−36

of statistical significance. Given that the result is significant, we can hence conclude that the overall ranking of the different methods is BBNC > LSVM > NaiveB > RBFN. This outstanding performance could be due to the low diversity of MDDR1 data set.

Returning to Table 5, BBNC and others showed lower prediction results (~72%) for MUT compared to the first four data sets (COX2, BZR, DHFR, and ER). However, that was because the MUT data set involves the most heterogeneous activity class of those data sets. The same results can be observed for MDDR2 in Table 8. The results in Tables 8 and 9 showed that the BBNC ranked third after LSVM and NaiveB. The MUT and MDDR2 results are of particular interest since they involved the most heterogeneous activity classes in the eight data sets used in this study (except for Norine data set) and thus provided a stiff test of the effectiveness of a prediction method.

Visual inspection of the results in Tables 5−9 showed that BBNC provides an interesting and promising method for activity prediction. Beside the good performance of BBNC method, BBNC was also very fast in comparison to other machine learning algorithms, since building of the BBNC model only depends on statistical calculation from active and inactive training data.

Results in Tables 5−9 clearly showed that the performance of each method was different from one data set to another and even from one activity to another. Thus, it is important to stress that no prediction method is better than the others and able to deal with all the data sets. However, we believe that using different methods for target prediction can help a better understanding of data. Thus, we introduce BBNC as a useful addition to the computational chemist's toolbox.

■ CONCLUSION

This paper has further investigated the use of Bayesian belief network classifier for ligand-based target or activity prediction. We applied the BBNC method on eight well-known data sets from the literature and compared its performance to three machine learning algorithms. Experiments showed that BBNC provides interesting prediction rates (from 79% accuracy for a high diverse data set to 99% for low diverse data set) with short time calculation for activity prediction. Experiments also showed that BBNC is particularly effective for homogeneous data sets but performs less well with structurally heterogeneous ones. However, it is important to stress that we believe that using several approaches, whenever possible, for target prediction can often give a broader understanding of the data than using only one approach alone. Thus, we see BBNC being a useful addition to the computational chemist's toolbox.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: ammar_utm@yahoo.com.

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) Johnson, M. A.; Maggiora, G. M. *Concepts and Application of Molecular Similarity*; John Wiley: New York, 1990.

(2) Bruce, C. L.; Melville, J. L.; Pickett, S. D.; Hirst, J. D. Contemporary QSAR Classifiers Compared. *J. Chem. Inf. Model.* **2007**, *47*, 219−227.

(3) Burden, F. R.; Winkler, D. A. New QSAR Methods Applied to Structure−Activity Mapping and Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 236−242.

(4) Rogers, D.; Hopfinger, A. J. Application of Genetic Function Approximation to Quantitative Structure-Activity Relationships and Quantitative Structure-Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854−866.

(5) Walters, W.; Goldman, B. Feature selection in quantitative structure-activity relationships. *Curr. Opin. Drug. Discov. Dev.* **2005**, *8*, 329−333.

(6) Harper, G.; Bradshaw, J.; Gittins, J. C.; Green, D. V. S.; Leach, A. R. Prediction of Biological Activity for High-Throughput Screening Using Binary Kernel Discrimination. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1295−1300.

(7) Willett, P.; Wilton, D.; Hartzoulakis, B.; Tang, R.; Ford, J.; Madge, D. Prediction of Ion Channel Activity Using Binary Kernel Discrimination. *J. Chem. Inf. Model.* **2007**, *47*, 1961−1966.

(8) Xia, X.; Maliski, E. G.; Gallant, P.; Rogers, D. Classification of Kinase Inhibitors Using a Bayesian Model. *J. Med. Chem.* **2004**, *47*, 4463−4470.

(9) Glick, M.; Klon, A. E.; Acklin, P.; Davies, J. W. Enrichment of Extremely Noisy High-Throughput Screening Data Using a Naïve Bayes Classifier. *J. Biomol. Screen.* **2004**, 9.

(10) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Molecular Similarity Searching Using Atom Environments, Information-Based Feature Selection, and a Naïve Bayesian Classifier. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 170−178.

(11) Nidhi; Glick, M.; Davies, J. W.; Jenkins, J. L. Prediction of Biological Targets for Compounds Using Multiple-Category Bayesian Models Trained on Chemogenomics Databases. *J. Chem. Inf. Model.* **2006**, *46*, 1124−1133.

(12) Ajay; Bemis, G. W.; Murcko, M. A. Designing Libraries with CNS Activity. *J. Med. Chem.* **1999**, *42*, 4942−4951.

(13) Balakin, K. V.; Tkachenko, S. E.; Lang, S. A.; Okun, I.; Ivashchenko, A. A.; Savchuk, N. P. Property-Based Design of GPCR-Targeted Library. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1332−1342.

(14) Schneider, G.; Wrede, P. Artificial neural networks for computer-based molecular design. *Prog. Biophys. Mol. Biol.* **1998**, *70*, 175−222.

(15) Winkler, D. A.; Burden, F. R. Application of Neural Networks to Large Dataset QSAR, Virtual Screening, and Library Design. In *Combinatorial Library: Methods and Protocols*; Springer-Verlag, New York, 2002; Vol. *201*, pp 325−367.

(16) Baskin, I.; Zhokhova, N.; Palyulin, V.; Zefirov, A.; Zefirov, N. Multilevel approach to the prediction of properties of organic compounds in the framework of the QSAR/QSPR methodology. *Dokl. Chem.* **2009**, *427*, 172−175.

(17) Yang, Z. R. Biological applications of support vector machines. *Brief. Bioinform.* **2004**, *5*, 328−338.

(18) Kawai, K.; Fujishima, S.; Takahashi, Y. Predictive Activity Profiling of Drugs by Topological-Fragment-Spectra-Based Support Vector Machines. *J. Chem. Inf. Model.* **2008**, *48*, 1152−1160.

(19) Wassermann, A. M.; Geppert, H.; Bajorath, J. r. Searching for Target-Selective Compounds Using Different Combinations of Multiclass Support Vector Machine Ranking Methods, Kernel Functions, and Fingerprint Descriptors. *J. Chem. Inf. Model.* **2009**, *49*, 582−592.

(20) Sheridan, R. P.; Kearsley, S. K. Why do we need so many chemical similarity search methods? *Drug Discovery Today* **2002**, *7*, 903−911.

(21) Ding, H.; Takigawa, I.; Mamitsuka, H.; Zhu, S. Similarity-based machine learning methods for predicting drug−target interactions: a brief review. *Brief. Bioinform.* **2013**, DOI: 10.1093/bib/bbt056.

(22) Jenkins, J. L.; Bender, A.; Davies, J. W. In silico target fishing: Predicting biological targets from chemical structure. *Drug Discovery Today: Technol.* **2006**, *3*, 413−421.

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

(24) Chen, B.; Mueller, C.; Willett, P., Evaluation of a Bayesian inference network for ligand-based virtual screening. *J. Cheminf.* **2009**, *1*, 5; available at http://www.jcheminf.com/content/1/1/5.
(25) Abdo, A.; Chen, B.; Mueller, C.; Salim, N.; Willett, P. Ligand-Based Virtual Screening Using Bayesian Networks. *J. Chem. Inf. Model.* **2010**, *50*, 1012−1020.
(26) Sutherland, J. J.; O'Brien, L. A.; Weaver, D. F. Spline-Fitting with a Genetic Algorithm: A Method for Developing Classification Structure−Activity Relationships. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1906−1915.
(27) Helma, C.; Cramer, T.; Kramer, S.; De Raedt, L. Data Mining and Machine Learning Techniques for the Identification of Mutagenicity Inducing Substructures and Structure Activity Relationships of Noncongeneric Compounds. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1402−1411.
(28) Chavatte, P.; Yous, S.; Marot, C.; Baurin, N.; Lesieur, D. Three-Dimensional Quantitative Structure−Activity Relationships of Cyclooxygenase-2 (COX-2) Inhibitors: A Comparative Molecular Field Analysis. *J. Med. Chem.* **2001**, *44*, 3223−3230.
(29) Sutherland, J. J.; O'Brien, L. A.; Weaver, D. F. A Comparison of Methods for Modeling Quantitative Structure−Activity Relationships. *J. Med. Chem.* **2004**, *47*, 5541−5554.
(30) Sutherland, J. J.; Weaver, D. F. Three-dimensional quantitative structure-activity and structure-selectivity relationships of dihydrofolate reductase inhibitors. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 309−331.
(31) Caboche, S.; Pupin, M.; Leclère, V.; Fontaine, A.; Jacques, P.; Kucherov, G. NORINE: a database of nonribosomal peptides. *Nucleic Acids Res.* **2008**, *36*, D326−D331.
(32) Abdo, A.; Caboche, S.; Leclère, V.; Jacques, P.; Pupin, M. A new fingerprint to predict nonribosomal peptides activity. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 1187−1194.
(33) Symyx Technologies. *MDL Drug Data Report*. http://accelrys.com/products/databases/bioactivity/mddr.html (accessed June 20, 2013).
(34) Abdo, A.; Saeed, F.; Hamza, H.; Ahmed, A.; Salim, N. Ligand expansion in ligand-based virtual screening using relevance feedback. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 279−287.
(35) Abdo, A.; Salim, N. New Fragment Weighting Scheme for the Bayesian Inference Network in Ligand-Based Virtual Screening. *J. Chem. Inf. Model.* **2011**, *51*, 25−32.
(36) Abdo, A.; Salim, N.; Ahmed, A. Implementing Relevance Feedback in Ligand-Based Virtual Screening Using Bayesian Inference Network. *J. Biomol. Screen* **2011**, *16*, 1081−1088.
(37) 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.
(38) *Pipeline Pilot*, Accelrys Software Inc.: San Diego CA, 2008.
(39) Witten, I. H.; Frank, E. *Data Mining: Practical machine learning tools and techniques*, 2nd ed.; Morgan Kaufmann: San Francisco, 2005.
(40) John, G. H.; Langley, P., Estimating continuous distributions in Bayesian classifiers. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, Morgan Kaufmann Publishers Inc.: Montréal, Qué, Canada, 1995; pp 338−345.
(41) Chang, C.-C.; Lin, C.-J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 1−27.
(42) Bugmann, G. Normalized Gaussian Radial Basis Function networks. *Neurocomputing* **1998**, *20*, 97−110.
(43) Siegel, S.; Castellan, N. J. *Nonparametric Statistics for The Behavioral Sciences*; McGraw-Hill: New York, 1988.