JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

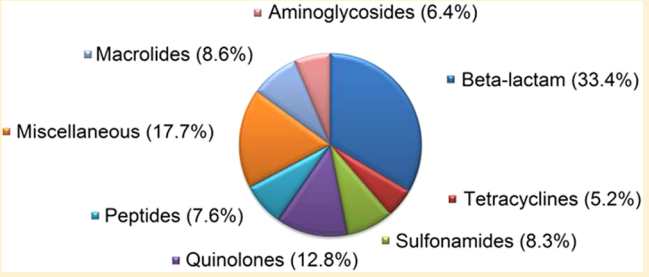
QSAR Classification Model for Antibacterial Compounds and Its Use in Virtual Screening

Narender Singh,* Sidhartha Chaudhury, Ruifeng Liu, Mohamed Diwan M. AbdulHameed, Gregory Tawa, and Anders Wallqvist

DoD Biotechnology High Performance Software Applications Institute, BHSAI/MRMC, ATTN: MCMR-TT, 2405 Whittier Drive, Frederick, Maryland 21702, United States

Supporting Information

ABSTRACT: As novel and drug-resistant bacterial strains continue to present an emerging health threat, the development of new antibacterial agents is critical. This includes making improvements to existing antibacterial scaffolds as well as identifying novel ones. The aim of this study is to apply a Bayesian classification QSAR approach to rapidly screen chemical libraries for compounds predicted to have antibacterial activity. Toward this end we assembled a data set of 317 known antibacterial compounds as well as a second data set of diverse, well-validated, non-antibacterial compounds from 215 PubChem Bioassays against various bacterial species. We constructed a Bayesian classification model using structural fingerprints and physicochemical property descriptors and achieved an accuracy of 84% and precision of 86% on an independent test set in identifying antibacterial compounds. To demonstrate the practical applicability of the model in virtual screening, we screened an independent data set of ∼200k compounds. The results show that the model can screen top hits of PubChem Bioassay actives with accuracy up to ∼76%, representing a 1.5−2-fold enrichment. The top screened hits represented a mixture of both known antibacterial scaffolds as well as novel scaffolds. Our study suggests that a well-validated Bayesian classification QSAR approach could compliment other screening approaches in identifying novel and promising hits. The data sets used in constructing and validating this model have been made publicly available.



INTRODUCTION

It is impossible to determine the number of bacterial infections treated each year worldwide. According to World Health Organization (WHO), the top five infectious diseases with highest death rates are lower respiratory tract infections (3.9 million), diarrhea (1.8 million), tuberculosis (1.6 million), pertussis (290 000), and tetanus (210 000).[1] Considering this staggering number of deaths, there should be a lucrative market for drug therapies for these diseases. Indeed, this was true up until the early 1990s when around 20 pharmaceutical companies were involved in antibacterial research. Today only two are active.[2] In the last 25 years, not a single novel antibacterial drug class has been discovered. Though many scientists consider the last three major classes discovered to be novel. oxazolidinones (2000), lipopeptides (2003), and pleuromutilins (2007), they were, in fact, patented in 1978,[3] 1987,[4] and 1952,[5] respectively.

Multiple reasons have been cited for this drift from the "golden age" (1945−1965) to the "innovation gap" (1987 and onward) of discovering novel antibacterial compounds. Among many, there are three main hurdles to success in this area. First is "scientific difficulties" due to (i) rapid evolution of resistant strains that renders even the newly developed antibacterials ineffective, (ii) lack of novel screening libraries and compounds, for new drug discovery, and (iii) difficult to manage side-effects due to high dose requirements in order to achieve blood levels necessary for efficacy. Second, is "pharmaceutical company disinterest" due to (i) lack of financial gains because of the short-duration treatment regimen typically prescribed for antibacterials, (ii) difficulties in licensing, and (iii) the uncertain future of drugs due to resistance. Last, but not the least, is "regulatory hurdles" due to (i) The Food and Drug Administration's delay in issuing guidance documents regarding acceptable study designs and acceptable efficacy outcomes and (ii) requirements for studies to sufficiently demonstrate the superiority over current treatment regimens of drugs, which leads to costly and difficult clinical trials. Many excellent reviews and articles exist in the literature that highlight these problems in detail.[6−9]

Despite all these difficulties, both scientific and otherwise, there is a perpetual need for new antibacterials, as a result, the antibacterial product pipeline has never been totally empty. Many new antibacterials have been approved since 1970, and almost all of them are improved versions of the previously known scaffold classes. Many of these improvements, some very substantial, have yielded analogues with broader antibacterial spectra to avoid resistance, lesser toxicity, and

Received: July 17, 2012
Published: September 26, 2012

   dx.doi.org/10.1021/ci300336v | J. Chem. Inf. Model. 2012, 52, 2559−2569

low dose regimens. A good example of such antibacterial evolution is cephalosporins. This class is one of the most commonly prescribed class of antibacterials.[10] Over the years, they have constantly evolved and each new generation (from I to V), while retaining the cephem scaffold, is designed to have added spectrum of activity and/or to be active against those bacteria that have become resistant to the previous generation.[11] Additional examples of new antibacterials, approved since 2000, based on known scaffolds include doripenem and ertapenem (carbapenems); tigecycline (tetracyclines); telithromycin and fidaxomicin (macrolides); telavancin (glycopeptides); gemifloxacin (quinolones); linezolid (oxazolidinones); dapromycin (lipopeptides); and retapamulin (pleuromutilin). Such modifications and incremental tailoring are not only necessary to fight the resistant pathogens but they can also be used to maximize the therapeutic potential of each scaffold to the fullest. This shows, that, along with research efforts to discover novel scaffolds we also need to devise ways to further explore the known scaffold properties of current antibacterials. This is especially true, since it is well-known that the currently known small molecule space is sparsely inhabited by antibacterial-like compounds. Hence, strategies that explore the available chemical space would help in finding new antibacterials.

Structurally, antibacterials differ significantly from other drug classes, such as drugs targeting human proteins.[12] Most of these differences are attributed to their need of penetrating and persisting in bacterial cells while avoiding human cells.[13] The differences, such as higher molecular weight and polarity, and other physicochemical properties, have been exploited in a few previous studies that utilize binary quantitative structure−activity relationship (QSAR) classification models to distinguish between antibacterials and non-antibacterial compounds.[14−20] These attempts include the use of techniques such as linear discriminant analysis, binary logistic regression, and artificial neural networks. In all these studies, the training data sets of antibacterials and non-antibacterial compounds range from 24 to 249 and 35 to 731, respectively, where all the non-antibacterials were collected from the Merck Index of compounds.[21] Additionally, many QSAR models for "class specific" antibacterials, such as fluoroquinolones,[22] β-lactams,[23] and aminoglycosides[24] have also been developed and appear to be useful in screening potential hits.

In the current study, we have used a previously unused Bayesian classification approach to build a QSAR model that can distinguish between antibacterial and non-antibacterial compounds. Bayesian modeling is a well-known classification approach, and many examples of its utility as a tool in drug discovery and structure−activity analysis exists. Previously, it has been used successfully in finding inhibitors of kinases,[25] G-protein-coupled receptors (GPCRs),[26] γ amino butyric acid type A ($GABA_A$) ionotropic receptor,[27] *Mycobacterium tuberculosis*,[28] and in identifying important structural features/fragments for microsomal stability[29] and human ether-a-go-go related gene (hERG) protein blockers.[30] Along with being deceptively simple and robust, another major strength of Bayesian approach is its ability to rank the molecules according to their probability of being active. This ranking of molecules is important when prioritizing molecules for screening, i.e., making focused libraries, or for further development.

In our study, we utilized structural fingerprints and selected physiochemical properties of 317 known antibacterials to build a Bayesian model. Unlike previous antibacterial classification studies, our collection of antibacterials is significantly larger. Using a novel strategy, an equal number of non-antibacterials were also collected using inactive compounds from 215 bacterial bioassays deposited in the freely available PubChem repository and provided by the National Center for Biotechnology Information (NCBI).[31] This is different from collecting non-antibacterials from the Merck Index of compounds, which is not open source and where most of the compounds are actually drugs targeting human proteins. Ultimately, since the goal of this model will be to enrich for antibacterial compounds from compound libraries typically used for antibacterial screening, we curated a representative set of well-validated non-antibacterials from data from actual antibacterial screening studies available through PubChem. The developed Bayesian models were validated using independent test set molecules that were not used to train the models. This allowed us to more accurately estimate the prediction power of the models. As mentioned above, since a model would be more useful if the model results could be translated into practical virtual screening strategies, we further validated our approach by successfully filtering out active hits from ∼200 000 screening molecules that were used to find inhibitors for various bacterial pathogens and deposited in PubChem Bioassays. Ultimately, the main purpose of this model is to make predictions, based on known antibacterial and non-antibacterials, for unknown screening compounds in order to identify the analogues that contain the most antibacterial like structural features and properties.

METHODS AND MATERIALS

Workflow. The workflow followed for our classification QSAR model building, its validation, and its use in virtual screening is shown in Figure 1. All the data sets were collected from DrugBank,[32] PubChem, and literature. Our Bayesian model was built using training set molecules, whereas validation was done on separate test set molecules. Finally, virtual screening was done on the collected set of actives and inactives from PubChem Bioassays for pathogens. Each step of this workflow is described in detail in the following subsections.

Data Set Collection. We collected a total of 317 known antibacterials from the literature and the DrugBank database of compounds. Structurally, these can be divided into nine classes as shown in Figure 2. The biggest class is β-lactam antibacterials that constitute roughly 1/3 of all the antibacterials and include subclasses such as cephalosporin, penicillin, carbapenem, monobactem, and oxacephem. After the β-lactams, quinolones, sulfonamides, and macrolides constitute the next three largest compound classes. Unlike antibacterials, collecting non-antibacterial compounds, compounds that are inactive in a broad panel of bacterial species, is difficult and no straightforward approach or database is readily available for this task. In our study, to collect a database of non-antibacterial compounds we used publically available PubChem Bioassay results as provided by the NCBI. A total of 215 different bacterial bioassays were available in PubChem at the time of this study. Out of these, only 30 bioassays screened 10 or more compounds. From this subset of 30 bioassays, we selected all 350 000 unique inactive compounds. We further selected only those, a total of 190 477, compounds that were found to be inactive in at least 7 or more different bacterial bioassays. This is still a huge collection of compounds compared to our antibacterial data set of 317 compounds. Because a QSAR classification model works best if the data set compounds are as

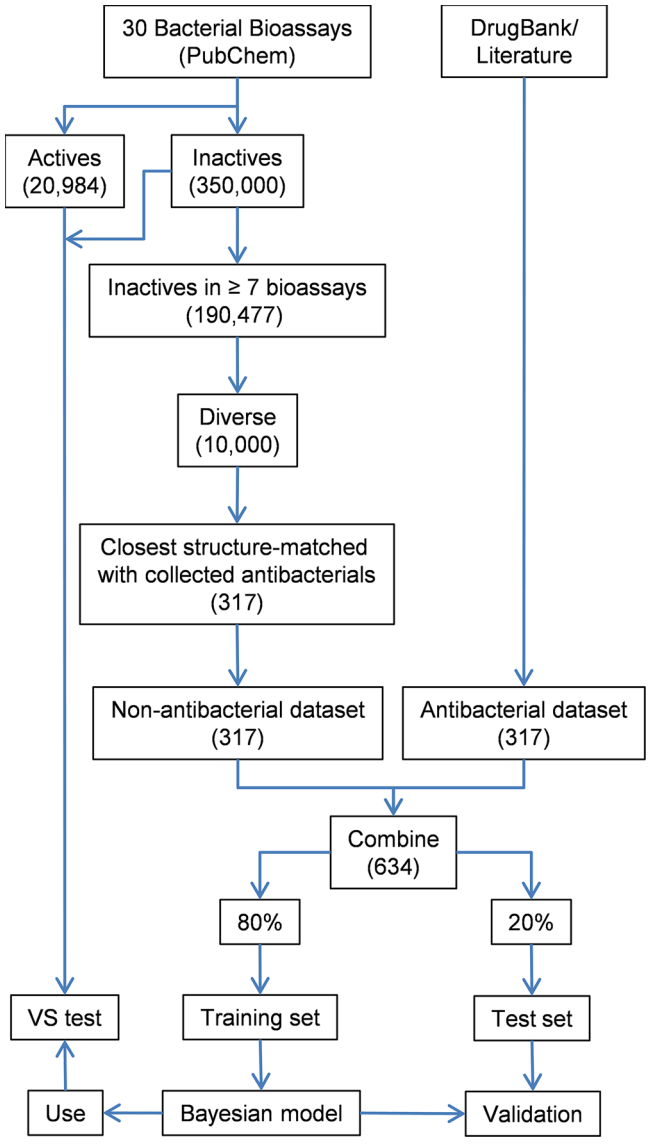


Figure 1. Workflow for QSAR classification model building, validation, and virtual screening (VS) as applied to antibacterial and non-antibacterial data sets. The number of compounds is also shown for some of the steps, in parentheses.

diverse as possible, we did a clustering analysis, and based on Tanimoto coefficient values, selected the top 10 000 most diverse compounds from the pool of 190 477 non-antibacterials. In this clustering task, molecular similarity was done based on the Tanimoto distance between molecules using the ECFP_6 fingerprint property (atom type-based extended connectivity fingerprint).[33] The maximum dissimilarity of center selections were picked from the diverse outer edges of the clusters. Additionally, for better classification models, the two data sets, the antibacterials and the diverse set of inactives, should be as closely matched as possible so that only the best discriminating features between the two sets can be collected. To do this, we structure-matched the 10 000 compounds with the 317 antibacterials and selected the same number of most closely matched inactives to keep a ratio of 1:1 between the two sets. We labeled this inactive set of 317 compounds as the non-antibacterial data set.

Model Building. On the collected data sets of antibacterials and non-antibacterials, we applied the Bayesian classification approach which is based on a learn-by-example protocol, as implemented in Pipeline Pilot, version 8.0.[34] The Bayesian approach is a robust classification approach that can distinguish between active and inactive compound sets. Complete details of the Bayesian method are described elsewhere,[25] but in short the technique is based on the frequency of occurrence of various descriptors that are found in two or more sets of molecules that discriminate best between these sets. The model learning process starts by generating a large set of binary (yes/no) features from the input set of descriptors, structural and/or physicochemical, and then collects the frequency of occurrence of each feature in the "good (active)" subset and among the "all data set" compounds. To apply the model to a particular sample, the features of the sample are generated, and a Laplacian adjusted weight is calculated for each feature based on a probability estimate. Finally, the weights are added to create a weight sum which provides a relative predictor of the likelihood of that sample being from the "good (active)" subset.

In our approach, we selected both the structural descriptors, molecular function class fingerprints of maximum diameter 6 (FCFP_6),[33] and the physiochemical descriptors SlogP, molecular weight, number of rotatable bonds, number of rings, number of aromatic rings, number of hydrogen bond acceptors, number of hydrogen bond donors, and molecular polar surface area. All the physicochemical descriptors were precalculated with Chemical Computing Group's Molecular Operating Environment (MOE), v. 2010.10.[35] The compounds were divided into the training and test sets by randomly selecting 80% of the antibacterials and non-antibacterials for training and the remaining for testing. To test whether the random selection of compounds for training and testing created a bias, we repeated the selection 10 times and applied the algorithm to each set. We did not detect any significant difference between the various data set results. Finally, the model was built using the training data set of compounds.

Model Validation. The model validation was done using leave-one-out cross-validation. In this technique, each compound is left out one at a time, and the model built form the remaining compounds is used to predict the left out compound. Once all the compounds pass through this cycle of prediction, a Receiver Operator Characteristic (ROC) plot is generated and the area under the curve (AUC) is measured. Predictions were made for both the training set and test set compounds. Table 1 gives the definition and relationship of the statistical parameters calculated to determine the quality of the model, i.e., accuracy, sensitivity, specificity, precision, and kappa.

For classification models, the kappa value is considered as true accuracy, because the agreement by chance is corrected for and, hence, it is a better statistical parameter than accuracy to estimate the prediction power of the model. A model is often considered useful if its kappa value is $\geq$0.4.[29]

Virtual Screening. To test how well the model performs in a real virtual screening experiment, we prepared two data sets. In one data set, we mixed 20 984 PubChem active compounds and an equal number of inactive compounds randomly selected from the pool of 190 477 inactive compounds collected from PubChem Bioassays. This gives a ratio of 1:1 for actives versus inactive. In another set, we mixed 20 984 actives with all the 190 477 compounds that were inactives in seven or more PubChem Bioassays. This gives a ratio of ~1:9 for actives versus inactives. It is fairly easy to collect and prepare other ratio data sets also, but we feel these two ratios, 1:1 and 1:9 of actives versus inactives, are sufficient to give an indication of the

dx.doi.org/10.1021/ci300336v | J. Chem. Inf. Model. 2012, 52, 2559−2569

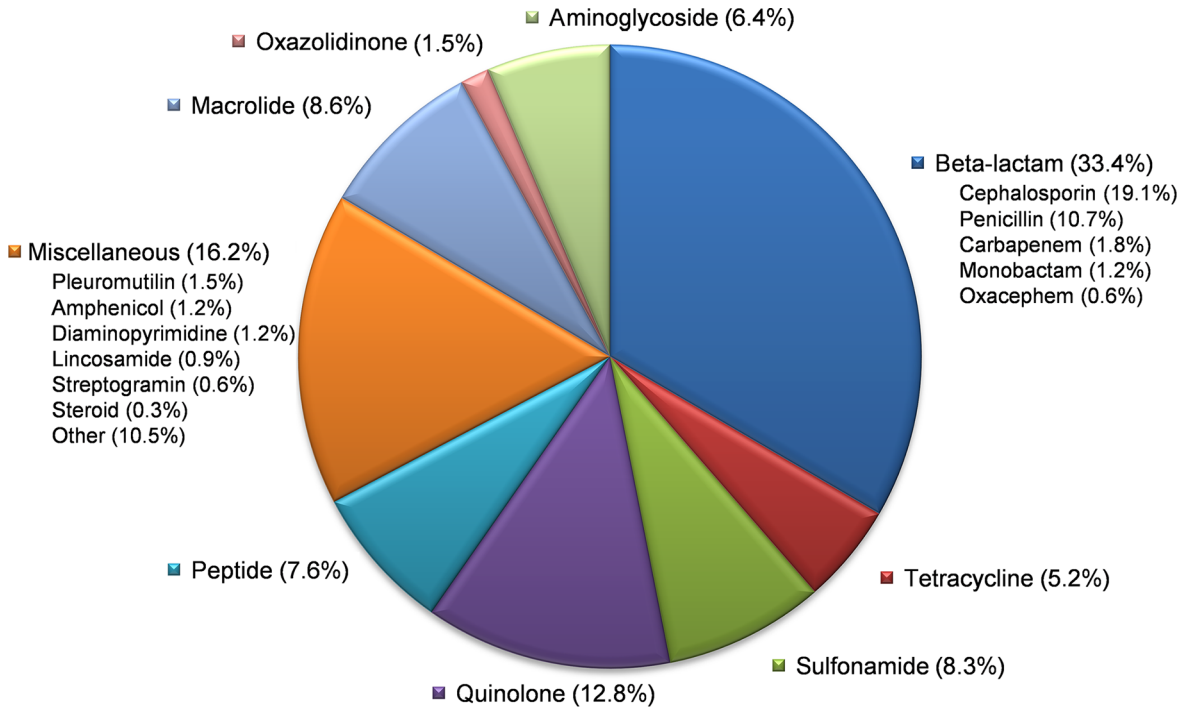


Figure 2. Pie chart of 317 antibacterials from 9 different classes collected from DrugBank and literature. Percentage of total number of compounds for each class is shown.

Table 1. Definition of Classification Model Performance Measures between Predicted and Observed Parameters for Two Data Sets, Antibacterial and Non-Antibacterial[a]

		predicted	
		antibacterial	non-antibacterial
observed	antibacterial	true positive (TP)	false negative (FN)
	non-antibacterial	false positive (FP)	true negative (TN)

[a]Various statistical parameters can be calculated based on this relationship. N (total) = TP + FP + FN + TN. Accuracy (proportion of true prediction in the entire population) = (TP + TN)/N. Sensitivity (ability to correctly predict positive results) = TP/(TP + FN). Specificity (ability to correctly predict negative results) = TN/(FP + TN). Precision (proportion of true prediction against all true results) = TP/(TP + FP). Kappa = ((TP + TN) − (((TP + FN)(TP + FP) + (FP + TN) (FN + TN))/N))/(N − (((TP + FN)(TP + FP) + (FP + TN)(FN + TN))/N)).

quality of the model performance in virtual screening experiments. Moreover the ratio of 1:1 reflects the similarity with a data sets ratio used for training, while the ratio of 1:9 is more reflective of real case virtual screening cases where active and inactive ratio is highly imbalanced, in favor of inactives. The model was tested on both data sets and the number of active compounds found in the top 10, 50, 100, 500, and 1 000 predicted compounds were calculated. A comparison of Bayesian model screening was also performed with a similarity-based screening method on the same two data sets. The 2D similarity screening was carried out using Tanimoto coefficients computed based on the structural fingerprint FCFP_6. In both the data sets, a 1:1 and 1:9 ratio of PubChem actives versus inactives was screened for similarity with the 317 antibacterials. Similar to the Bayesian model assessment, from both the data sets we extracted the top 10, 50, 100, 500, and 1 000 most similar hits (most similar to any of the 317 antibacterials). Finally, the numbers of PubChem actives in those sets, i.e., in top 10, 50, 100, 500, and 1000, were calculated. The numbers of PubChem actives (enrichment) in these sets were compared with the number of PubChem actives obtained from Bayesian screening sets.

RESULTS AND DISCUSSION

The data set collection, model development, and validation procedure described in this study provided a robust and straightforward approach for estimating the antibacterial-like probabilities of a small molecule. This includes an estimate of the accuracy and the predictive power of the developed Bayesian classification model. The data sets required for building such a model requires one active antibacterial data set and one inactive non-antibacterial data set.

Antibacterials and Nonantibacterials. Antibacterials constitutes a very heterogeneous set of compounds. They occupy a unique physicochemical property space, as compared to drugs targeting human proteins and compared with compounds that are commonly found in screening libraries.[12] In Table 2, we provide the mean values of nine physicochemical descriptors, molecular weight (Wt), hydrogen bond acceptors (HBA) and donors (HBD), number of nitrogen (nN) and oxygen atoms (nO), number of rings (Rings), log of the octanol/water partition coefficient (SlogP), topological polar surface area (TPSA), number of rotatable bonds (RB), and two violation counts using Lipinski (LV)[36] and Oprea (OV)[37] rules, for the three sets of compounds classified as antibacterials, non-antibacterials, and drugs targeting human proteins. These mean values amply demonstrate the substantial differences between these three compound categories.

Overall, antibacterials have, roughly, 50% higher weight, 60−130% more acceptors and donors, 30−90% more nitrogen and 120% more oxygen atoms, 40% higher flexibility (RB), 150% lower solubility, 90−120% higher polarity (TPSA), and 30% more ring structures, compared to the non-antibacterial and

dx.doi.org/10.1021/ci300336v | J. Chem. Inf. Model. 2012, 52, 2559−2569

Table 2. Comparison of Average (Mean) Compound Property Values of Three Data Sets Representing Antibacterials, Non-Antibacterials, and Drugs Targeting Human Proteins[a]

		average descriptor value											
DB	class	N	Wt	HBA	HBD	RB	SlogP	TPSA	Rings	nN	nO	LV	OV
AB	AG	24	511	9.9	5.5	7.3	−9.3	271	3.2	4.8	9.9	2.5	4.8
	BL	104	450	4.3	1.7	7.8	−1.5	154	3.5	4.6	5.8	0.8	1.4
	ML	26	726	10.5	3.4	8.8	1.4	184	3.5	1.8	12.0	2.1	3.3
	OX	6	412	4.0	1.5	6.8	1.2	108	3.7	4.5	4.7	0.2	1.2
	PE	24	1225	14.1	12.0	25.7	−4.7	460	5.0	12.6	16.1	2.7	4.5
	QL	41	362	1.9	0.1	3.1	−0.3	88	3.9	3.0	3.5	0.0	0.3
	SL	27	289	3.2	1.9	3.9	1.4	103	2.1	3.6	2.9	0.1	0.1
	TC	17	503	4.3	4.1	4.2	−1.9	195	4.2	2.4	8.6	1.8	3.1
	MS	48	563	6.9	3.3	8.4	2.2	153	3.7	2.6	8.0	1.2	2.4
	all	317	530	6.0	3.1	8.1	−1.1	177	3.6	4.3	7.3	1.1	2.0
NAB		317	354	3.8	1.8	5.9	2.7	92	2.7	3.3	3.3	0.3	0.6
DHP		527	353	2.8	1.3	5.9	1.7	78	2.8	2.2	3.3	0.3	0.7

[a]Antibacterials are further divided into nine classes and their comparison is also shown. [Abbreviations: database (DB), antibacterials (AB), non-antibacterials (NAB), drugs for human proteins (DHP), aminoglycosides (AG), β-lactams (BL), macrolides (ML), oxazolidinones (OX), peptides (PE), quinolones (QL), sulfonamides (SL), tetracyclines (TC), and miscellaneous (MS)].

Table 3. statistical Outcome of the Performance of the Bayesian Classifiers for the Training and Test Sets Molecules

parameters	training	test
N	506	128
good (antibacterials)	253	64
bad (non-antibacterials)	253	64
TP	226	51
TN	236	57
FP	17	8
FN	27	12
accuracy	0.91	0.84
precision	0.93	0.86
sensitivity	0.89	0.81
specificity	0.93	0.88
kappa	0.83	0.69

drugs targeting human proteins. The antibacterial compounds, on average, also violate at least one Lipinski's rule-of-5 and two Oprea's lead-like criteria of small molecules. Among the antibacterials, most nondrug like classes are aminoglycosides, peptides, and macrolides, each showing Lipinski violation counts of 4.8, 4.5, and 3.3, respectively. On other hand, the most druglike are sulfonamides and quinolones, each showing Lipinski violation counts of only 0.1 and 0.3, respectively.

Bayesian Model Development and Validation. To exploit the differences in both structural properties and physiochemical properties and between antibacterial and non-antibacterial compounds, we a used Bayesian classification technique as implemented in Pipeline Pilot. In this classification scheme, 317 antibacterial compounds are classified as "good" samples and 317 non-antibacterial compounds as "bad" samples. Here the "good" and "bad" are arbitrary labels to distinguish the two sets of compounds. The combined data sets were further divided into a 80:20 ratio to make a training set (506 compounds) and a test set (128 compounds). The Bayesian model was built from the training set compounds, using both the structural and the physicochemical property parameters. The model was validated using a leave-one-out cross-validation method where one compound is removed from the data set and its class, good or bad, is predicted using the model derived from the rest of the data set compounds. An accuracy of 91% was obtained with this the model. The same model was also used to predict the test data set of 128 compounds. For the test data set, an accuracy of 84% was obtained. The rest of the statistical parameters are shown in

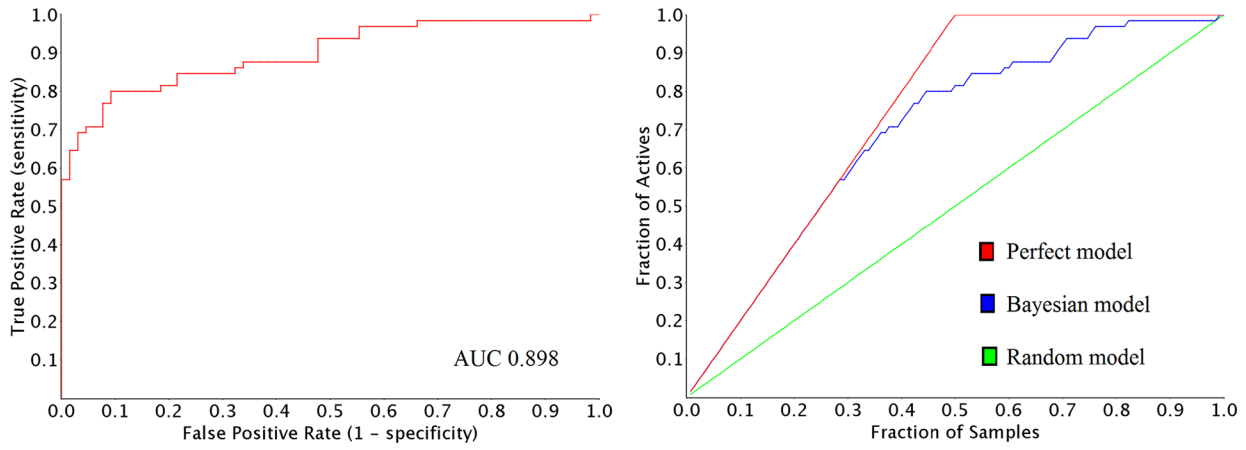


Figure 3. ROC plot (left) showing area under the curve and enrichment plot (right) showing the percentage of top retrieved actives for test set molecules.

dx.doi.org/10.1021/ci300336v | J. Chem. Inf. Model. 2012, 52, 2559−2569

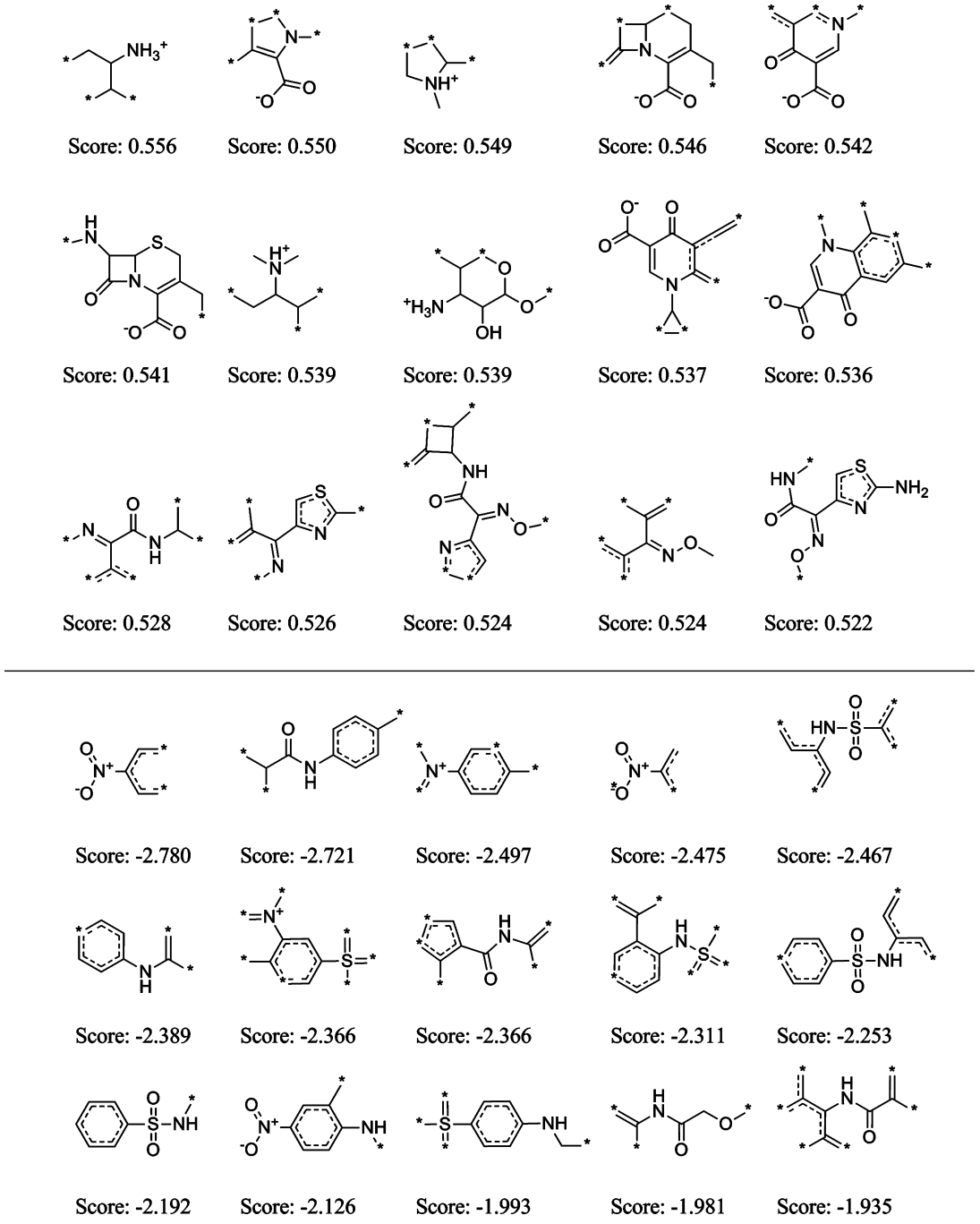


Figure 4. Examples of the top 15 good (top) and bad (bottom) fragments estimated by Bayesian modeling. The Bayesian score (Score) is given for each fragment.

Table 3. The ROC plot and the enrichment plot for the test set are shown in Figure 3. As expected, the model behaved better for the training set, but the outcome was still very good for the independent test set compound classification with a precision of 86% and a kappa value of 0.69 for the classification of the compounds. Sensitivity and specificity were 81% and 88% for the test set compounds.

Antibacterials: Good and Bad Fragments. One of the advantages of using a Bayesian classifier based on structural fingerprints, such as FCFP_6, is that it can identify important fragments or fingerprint features frequently found in two classifying groups. From a total of 7 232 FCFP_6 features that we used in making the model, the top 15 good and top 15 bad diverse fragments, favorable and unfavorable for the antibacterial classification, are shown in Figure 4. [It is important to note that these 30 fragments by no means represent all the antibacterial structural information, since it is a very small percentage (<0.5%) of the total number of features used in building the model].

As expected, many of the top good features contain $-NH_3^+$/$-NH_2^+$/$-NH^+$, thiazole, penem, cephem, or quinolone fragments that are common fragmental features of aminoglycosides, peptides, β-lactams, and quinolones, which constitute a majority of known antibacterials. Interestingly, in the top bad fragments, many O═S(−N)═O containing fragments were found, even though these are part of one of the most populated

dx.doi.org/10.1021/ci300336v | J. Chem. Inf. Model. 2012, 52, 2559−2569

Table 4. Statistical Outcome of the Performance of the Bayesian Classifiers for the Training and Test Sets Molecules That Do Not Include Sulfa Compounds

parameters	training	test
	(no sulfa compounds)	
total	485	124
good (antibacterials)	232	59
bad (non-antibacterials)	253	65
true positives	228	51
true negatives	232	54
false positives	21	11
false negatives	4	8
accuracy	0.95	0.85
precision	0.92	0.82
sensitivity	0.98	0.86
specificity	0.92	0.83
kappa	0.83	0.69

Table 5. Enrichment Results of Two Data Sets, For Two Screening Methods[a]

	screening data set		number of actives in				
screening method	no. of actives	no. of inactives	top 10	top 50	top 100	top 500	top 1 000
Bayesian model screening	20 974	20 974	9	43	85	405	758
	20 974	190 159	9	26	45	155	276
similarity-based screening	20 974	20 974	10	33	55	231	439
	20 974	190 159	9	27	30	70	115

[a]In one data set, the ratio of actives versus inactives is 1:1 (20 974 actives and equal number of inactives), and in another it is 1:9 (20 974 actives and 190 159 inactives). The number of actives retrieved by both the methods in top 10, 50, 100, 500, and 1 000 compounds is shown.

class (sulfa) of antibacterials. The reason that the sulfonamide moiety was selected as a bad fragment is because of its common occurrence in many non-antibacterial compounds. For example, other than antibacterials, sulfa compounds are also used in diuretic, anticonvulsants, and many dermatological drugs. Hence, because of their wide occurrence in both antibacterials, non-antibacterials, and in general screening libraries, these fragments are given a bad Bayesian score as they cannot be used as distinguishing features between the antibacterials and non-antibacterial compounds. [To provide an estimate of sulfonamide scaffold popularity, we calculated that out of ~62 million compounds available in the ChemNavigator database[38] of purchasable compounds, 7.6 million or 12.5% of all compounds were sulfonamides].

In model development, if we exclude the whole sulfa class of antibacterials from the training set compounds and follow exactly the same steps as in the previous model building, the resulting model behaves almost like a perfect model. The ROC-AUC value of such a model is 0.99. Other statistical validation parameters of the sulfa-excluding Bayesian model gives an accuracy of 95%, a precision of 92%, a sensitivity of 98%, and a specificity of 92% in the classification of the training set compounds. For the test set, the accuracy and precision of this model were 85% and 82%, while the sensitivity and specificity were 86% and 83%. The complete set of parameters are provided in Table 4.

Antibacterial Bayesian Model in Virtual Screening. In a study like this, the primary objective of the *in silico* screen is to determine whether the model can distinguish and classify unknown structures as good or bad. This is a common situation in drug discovery where one wants to retrieve active analogues from screening databases based on initial leads. Therefore, to validate the predictive power of our Bayesian model in a real test case scenario, we again used compounds from the PubChem collection of bacterial bioassays. A total of 215 different bacterial assays were selected from PubChem. Any bioassay that screened less than 10 compounds was excluded. From the remaining 30 bioassays, we collected 350 000 screened compounds. From this data set, we further selected a total of 190 477 compounds that were found to be inactive in at least 7 or more different bacterial assays. From this set, the 317 compounds that were used in developing the model as inactives (non-antibacterials) were removed. Finally, we had 190 159 compounds as inactives. From the same 215 assays, we also collected all the compounds that were flagged as active. Any known antibacterial tested in this set was also removed. This gives us 20 974 active compounds. This data set of inactive and active molecules represents a completely independent data set from the one used to build or test the Bayesian model. This data set was further divided into two subsets before model evaluation. This is done because the outcome of high throughput screening assays are highly imbalanced between

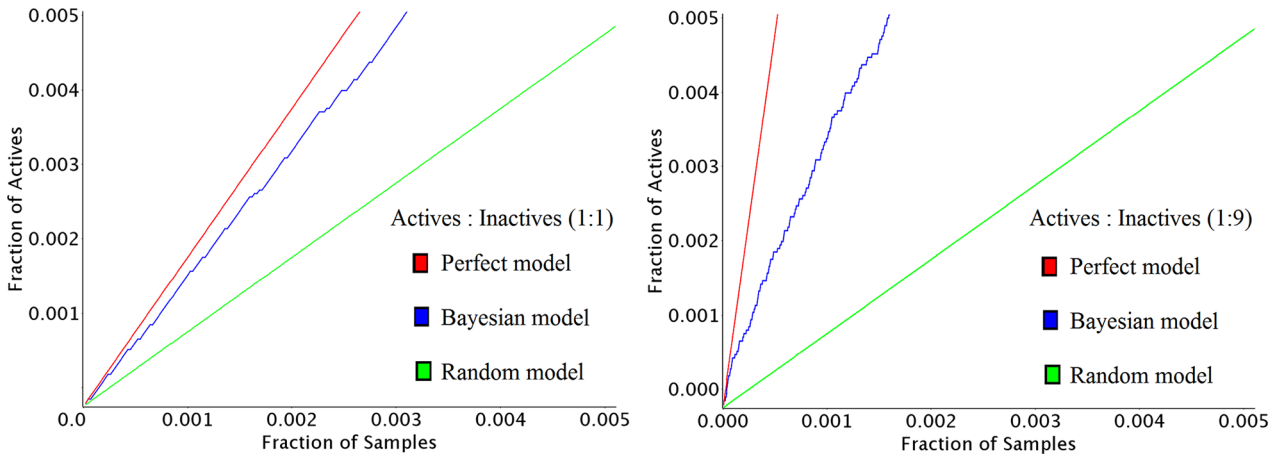


Figure 5. Enrichment plot showing the percentage of top retrieved actives in 0.5% of screened database for two virtual screening data sets selected from PubChem Bioassays for various bacterial species.

dx.doi.org/10.1021/ci300336v | J. Chem. Inf. Model. 2012, 52, 2559−2569

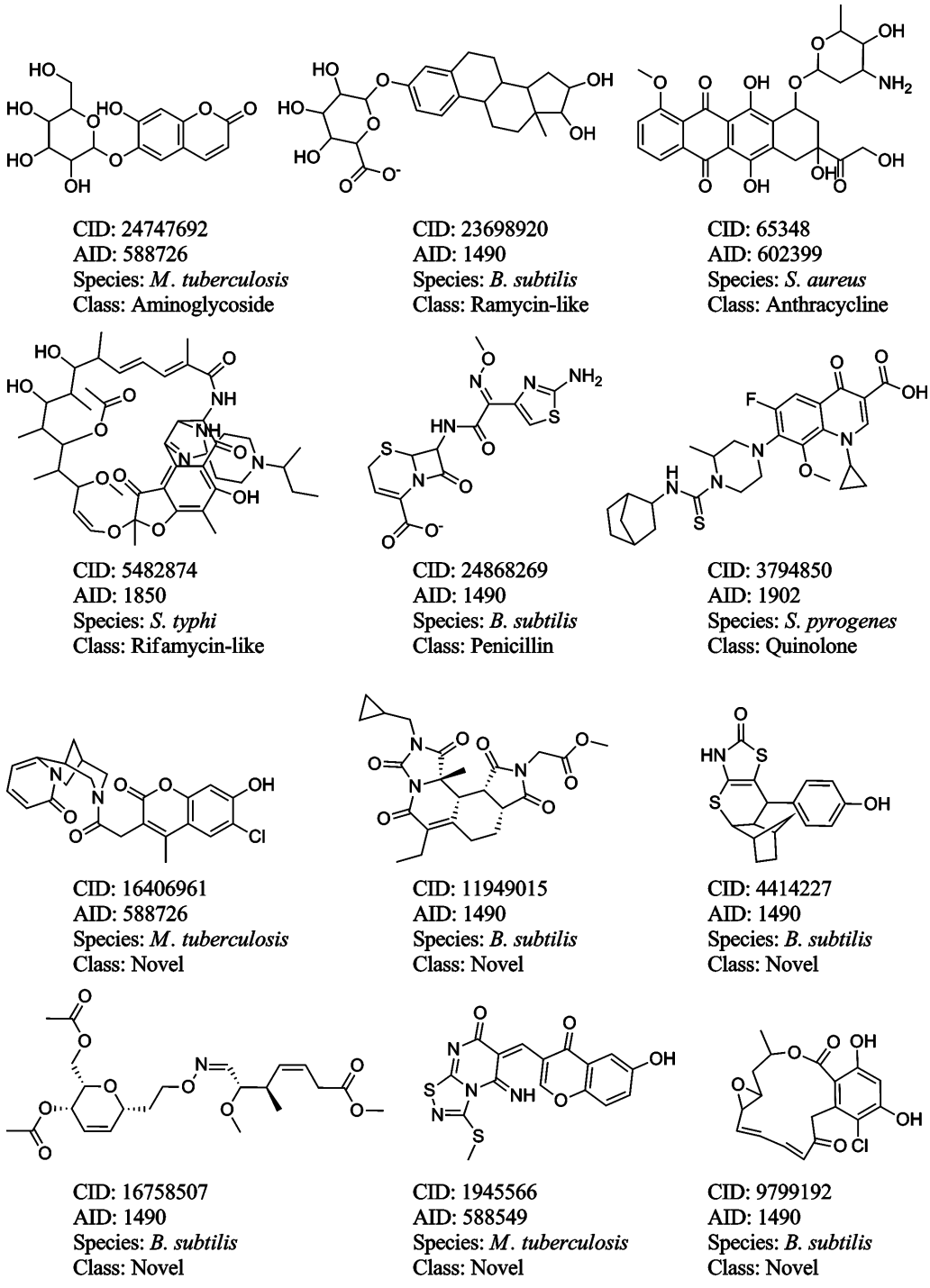


Figure 6. Structures of some of the top retrieved hits by the Bayesian model in virtual screening of the PubChem database. Each of these was experimentally found to be active in different PubChem bioassays. The PubChem identity number (CID), bioassay identity number (AID), species screened, and classification of antibacterial scaffold are provided for each hit.

active and inactives, in favor of inactives. For example, among PubChem Bioassays, in most cases, the experimental hit rate does not exceed 0.5%.[39] This imbalance poses a significant problem for classification models because models that correctly predict the same fraction of objects in each class will have different objective function values. Hence, we developed two independent data sets where in first we kept the ratio of actives versus inactive 1:1 and second where we kept the ratio 1:9 for actives versus inactives.

Next, we performed another test to compare how good our model results were as compared to 2D similarity-based screening. Among the ligand-based screening methods, the 2D similarity-based screening is one of the most popular methods of choice. This is not only because of the 2D method's computational efficiency but also because of its demonstrated effectiveness in multiple studies.[1,40−46] The 2D similarity screening was carried out using Tanimoto coefficients computed from structural fingerprint (FCFP_6). Both the data sets, of 1:1 and 1:9 ratio of PubChem actives versus

inactives, were screened for similarity with 317 antibacterials. The top 1 000 most similar hits for both the data sets were extracted and the number of actives in those sets were calculated.

The results of both the Bayesian model screening and the 2D similarity-based screening, for both the data sets, are shown in Table 5. Since the hit rate in actual high-throughput screening does not exceed 0.5%, we only show the actives extracted from the top 1 000 hits. In the first virtual screen of equal actives and inactives (1:1), the compounds extracted using the Bayesian model contained 90%, 85%, and ∼76% actives from the top 10, 100, and 1 000 hits, respectively. For the 1:9 ratio of actives and inactives of the second set, compounds extracted using the Bayesian model contained 90%, 45%, and 27.6% actives from the top 10, 100, and 1 000 hits, respectively. Compared to the actual high-throughput screen outcome for actives in PubChem Bioassays, there is a significant improvement in extracted hits in Bayesian model screening. For the model, the top 0.5% results are also shown as enrichment plots for both the subsets of 1:1 and 1:9 ratios of active versus inactive (Figure 5). In comparison, the early enrichment of the 2D similarity method was comparable to the Bayesian method. In the top 10 and 50 hits, the similarity method screened 100% to 66% actives in the 1:1 data set of actives and inactives and 90% to 54% actives in the 1:9 data set of actives and inactives. In later enrichment when all the antibacterial-like structures were exhausted in the similarity screen, the method performed no better than random sampling of hits, i.e., the actives screened in top 100 to top 1 000 hits (Table 5).

Another significant advantage of the Bayesian model was evident from the nature of the hits themselves. The model output was not just limited to finding only the existing scaffolds, i.e., similar to the molecules that were used in training of the model, but also included novel scaffolds. Figure 6 shows the structures of some of the top screening hits of Bayesian model that are experimentally active in PubChem Bioassays. The ability to discover novel antibacterial scaffolds is inherent in the Bayesian model formulation and should represent an attractive way to discover novel drug designs. In comparison, the 2D similarity search output is only limited to finding molecules that are similar to known antibacterial scaffolds.

CONCLUSIONS

Despite significant advances both in understanding the biology and the techniques available, antibacterial drug discovery is still an arduous task. In the last couple of years, several *in silico* methods have emerged as important drug-discovery tools. Currently, few studies exist that have described the use of property-based *in silico* classification models for antibacterial activity. Most of these published models show good to acceptable discrimination between antibacterial and non-antibacterial classification.

Our study differs from these previous studies in a number of ways. First, our collection of antibacterials is vast. In previous *in silico* studies, the antibacterial collection range from 24 to 249 compounds. We have collected 317 antibacterials of nine different classes from DrugBank and extensive literature searches and have made them publically available (Supporting Information). This is one of the largest reported and characterized data set of antibacterials. Such a collection is important since the performance of *in silico* classification models, such as Bayesian, heavily depends on the number and diversity of input training molecules. Second, no previous study has ever attempted to effectively describe "how to collect non-antibacterial" compounds. This is mainly because most of the studies tend to describe only the positive results, i.e., the compounds that turned out to be active in bacterial assays. More importantly, even if the data is published concerning inactive compounds, it remains focused only on one or few selected species of bacteria. For a non-antibacterial data set of compounds, the ideal compounds would be those that show inactivity against a panel of different species of bacteria. Our study is unique since we have collected the inactive compounds from 215 PubChem Bioassays results that were screened for a wide panel of bacterial species. Third, the Bayesian classification model described in this study has performed exceptionally well. The model correctly classified 51 of the 64 actives in an independent test set data, showing an overall accuracy of 84% and precision of 86%. Fourth, the model was subjected to an actual virtual screen test case of extracting high-throughput actives from PubChem bacterial bioassays. A comparison of such a virtual screening test case was also made with a 2D similarity search method. The Bayesian model extracted 75.8% of the actives from the top 1 000 extracted hits in a scenario where actives and inactives were mixed in a 1:1 ratio. In a more stringent test case, where actives and inactives were mixed in a 1:9 ratio, the model extracted 27.6% actives from the top 1 000 screened compounds. In comparison, the 2D similarity search only extracted 43.9% (in 1:1 ratio of actives and inactive) and 11.5% (in 1:9 ratio of actives and inactive) of the actives from the top 1 000 screened compounds, which is no better than random sampling. Moreover, while the top actives retrieved by 2D similarity search were all from previously known scaffold classes, Bayesian model screening hits were well populated with both the novel scaffolds as well as previously known scaffolds.

Overall, the Bayesian classification model is a robust method that permits a quick *in silico* discovery of novel antibacterials candidates making use of a minimum of resources, and it may be used as an efficient alternative to high-throughput screening of antibacterial agents.

ASSOCIATED CONTENT

Supporting Information

Four worksheets of (1) antibacterial statistics, showing the statistical details (average and standard deviation) of nine descriptors used in the study for all the data sets, i.e., antibacterials, non-antibacterials, and drugs for human proteins. The ttest comparison between antibacterial and non-antibacterial data sets is also shown. (2) Antibacterial data, lists 317 antibacterials used in the study. (3) Nonantibacterial data, lists 317 non-antibacterials used in the study. (4) Human drugs data, lists 527 human drugs used in the study. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Phone: +1-301-619-1941. Fax: +1-301-619-1983. E-mail: nsingh@bhsai.org.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

Funding of this research was provided by the U.S. Department of Defense Threat Reduction Agency Grant

TMTI0004_09_BH_T. The opinions or assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the U.S. Army or of the U.S. Department of Defense. We would also like to thank Ian Kerman, Dana Honeycutt, and Lee Herman for providing helpful suggestions in developing Pipeline Pilot protocols for Bayesian model development. This paper has been approved for public release with unlimited distribution.

REFERENCES

(1) Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* **2006**, *11*, 1046−1053.
(2) Lifesaving antibiotics face doubtful future. Infectious Diseases Society of America Press Release. April 7, 2011. www.idsociety.org/Content.aspx?id=17577.
(3) Fugitt, R. B.; Luckenbaugh, R. W. *5-Halomethyl-3-phenylu-2-oxazolidinones*. U.S. Patent 4,128,654, December 5, 1978.
(4) Debono, M.; Barnhart, M.; Carrell, C. B.; Hoffmann, J. A.; Occolowitz, J. L.; Abbott, B. J.; Fukuda, D. S.; Hamill, R. L.; Biemann, K.; Herlihy, W. C. A21978C, a complex of new acidic peptide antibiotics: isolation, chemistry, and mass spectral structure elucidation. *J. Antibiot. (Tokyo)* **1987**, *40*, 761−777.
(5) Anchel, M. Chemical studies with pleuromutilin. *J. Biol. Chem.* **1952**, *199*, 133−139.
(6) Gwynn, M. N.; Portnoy, A.; Rittenhouse, S. F.; Payne, D. J. Challenges of antibacterial discovery revisited. *Ann. N.Y. Acad. Sci.* **2010**, *1213*, 5−19.
(7) Livermore, D. M. Discovery research: the scientific challenge of finding new antibiotics. *J. Antimicrob. Chemother.* **2011**, *66*, 1941−1944.
(8) Payne, D. J.; Gwynn, M. N.; Holmes, D. J.; Pompliano, D. L. Drugs for bad bugs: confronting the challenges of antibacterial discovery. *Nat. Rev. Drug Discovery* **2007**, *6*, 29−40.
(9) Williams, K. J.; Bax, R. P. Challenges in developing new antibacterial drugs. *Curr. Opin. Investig. Drugs* **2009**, *10*, 157−163.
(10) Newman, D. J.; Cragg, G. M. Natural products as sources of new drugs over the last 25 years. *J. Nat. Prod.* **2007**, *70*, 461−477.
(11) Gilbert, D. N.; Moellering, R. C. , J. M. D.; Eliopoulos, G. M.; Henry F. Chambers, M. D.; Michael S. Saag, M. D.In *The Sanford Guide to Antimicrobial Therapy*, 39th ed.; Antimicrobial Therapy: Sperryville, VA, 2009.
(12) O'Shea, R.; Moser, H. E. Physicochemical properties of antibacterial compounds: implications for drug discovery. *J. Med. Chem.* **2008**, *51*, 2871−2878.
(13) Leeson, P. D.; Davis, A. M. Time-related differences in the physical property profiles of oral drugs. *J. Med. Chem.* **2004**, *47*, 6338−6348.
(14) Aptula, A. O.; Kühne, R.; Ebert, R.-U.; Cronin, M. T. D.; Netzeva, T. I.; Schüürmann, G. Modeling Discrimination between Antibacterial and Non-Antibacterial Activity based on 3D Molecular Descriptors. *QSAR Comb. Sci.* **2003**, *22*, 113−128.
(15) Cronin, M. T.; Aptula, A. O.; Dearden, J. C.; Duffy, J. C.; Netzeva, T. I.; Patel, H.; Rowe, P. H.; Schultz, T. W.; Worth, A. P.; Voutzoulidis, K.; Schuurmann, G. Structure-based classification of antibacterial activity. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 869−878.
(16) Marrero-Ponce, Y.; Medina-Marrero, R.; Torrens, F.; Martinez, Y.; Romero-Zaldivar, V.; Castro, E. A. Atom, atom-type, and total nonstochastic and stochastic quadratic fingerprints: a promising approach for modeling of antibacterial activity. *Bioorg. Med. Chem.* **2005**, *13*, 2881−2899.
(17) Molina, E.; Diaz, H. G.; Gonzalez, M. P.; Rodriguez, E.; Uriarte, E. Designing antibacterial compounds through a topological substructural approach. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 515−521.
(18) Murcia-Soler, M.; Perez-Gimenez, F.; Garcia-March, F. J.; Salabert-Salvador, M. T.; Diaz-Villanueva, W.; Castro-Bleda, M. J.; Villanueva-Pareja, A. Artificial neural networks and linear discriminant analysis: a valuable combination in the selection of new antibacterial compounds. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1031−1041.
(19) Murcia-Soler, M.; Perez-Gimenez, F.; Garcia-March, F. J.; Salabert-Salvador, M. T.; Diaz-Villanueva, W.; Medina-Casamayor, P. Discrimination and selection of new potential antibacterial compounds using simple topological descriptors. *J. Mol. Graphics Modell.* **2003**, *21*, 375−390.
(20) Tomas-Vert, F.; Perez-Gimenez, F.; Salabert-Salvador, M. T.; Garca-March, F. J.; Jaen-Oltra, J. Artificial neural network applied to the discrimination of antibacterial activity by topological methods. *J. Mol. Struct.: THEOCHEM* **2000**, *504*, 249−259.
(21) Maynard, R. L., *The Merck Index: Occupational and Environmental Medicine*, 12th ed.; 1996; pp 1−288.
(22) Minovski, N.; Vracko, M.; Solmajer, T. Quantitative structure-activity relationship study of antitubercular fluoroquinolones. *Mol. Diversity* **2011**, *15* (2), 417−426.
(23) Kempe, H.; Kempe, M. QSRR analysis of beta-lactam antibiotics on a penicillin G targeted MIP stationary phase. *Anal. Bioanal. Chem.* **2010**, *398*, 3087−3096.
(24) Setny, P.; Trylska, J. Search for novel aminoglycosides by combining fragment-based virtual screening and 3D-QSAR scoring. *J. Chem. Inf. Model.* **2009**, *49*, 390−400.
(25) Xia, X.; Maliski, E. G.; Gallant, P.; Rogers, D. Classification of kinase inhibitors using a Bayesian model. *J. Med. Chem.* **2004**, *47*, 4463−4470.
(26) Lee, J. H.; Lee, S.; Choi, S. In silico classification of adenosine receptor antagonists using Laplacian-modified naive Bayesian, support vector machine, and recursive partitioning. *J. Mol. Graphics Modell.* **2010**, *28*, 883−890.
(27) Vijayan, R. S.; Bera, I.; Prabu, M.; Saha, S.; Ghoshal, N. Combinatorial library enumeration and lead hopping using comparative interaction fingerprint analysis and classical 2D QSAR methods for seeking novel GABA(A) alpha(3) modulators. *J. Chem. Inf. Model.* **2009**, *49*, 2498−2511.
(28) Prathipati, P.; Ma, N. L.; Keller, T. H. Global Bayesian models for the prioritization of antitubercular agents. *J. Chem. Inf. Model.* **2008**, *48*, 2362−2370.
(29) Hu, Y.; Unwalla, R.; Denny, R. A.; Bikker, J.; Di, L.; Humblet, C. Development of QSAR models for microsomal stability: identification of good and bad structural features for rat, human and mouse microsomal stability. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 23−35.
(30) Wang, S.; Li, Y.; Wang, J.; Chen, L.; Zhang, L.; Yu, H.; Hou, T. ADMET evaluation in drug discovery. 12. Development of binary classification models for prediction of hERG potassium channel blockage. *Mol. Pharm.* **2012**, *9*, 996−1010.
(31) PubChem Bioassays. http://pubchem.ncbi.nlm.nih.gov/assay/ (accessed March 15, 2012).
(32) Knox, C.; Law, V.; Jewison, T.; Liu, P.; Ly, S.; Frolkis, A.; Pon, A.; Banco, K.; Mak, C.; Neveu, V.; Djoumbou, Y.; Eisner, R.; Guo, A. C.; Wishart, D. S. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.* **2011**, *39* (Database issue), D1035−D1041.
(33) Rogers, D.; Brown, R. D.; Hahn, M. Using extended-connectivity fingerprints with Laplacian-modified Bayesian analysis in high-throughput screening follow-up. *J. Biomol. Screen.* **2005**, *10*, 682−686.
(34) *Pipeline Pilot*, version 8.0; Accelrys: San Diego, CA, 2010.
(35) *Molecular Operating Environment (MOE)*, version 2010.10; Chemical Computing Group Inc.: Montreal, Canada, http://www.chemcomp.com
(36) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.
(37) Oprea, T. I.; Davis, A. M.; Teague, S. J.; Leeson, P. D. Is there a difference between leads and drugs? A historical perspective. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1308−1315.
(38) ChemNavigator. http://www.chemnavigator.com/cnc/products/iRL.asp (accessed January 25, 2012).

(39) Langdon, S. R.; Mulgrew, J.; Paolini, G. V.; van Hoorn, W. P. Predicting cytotoxicity from heterogeneous data sources with Bayesian learning. *J. Cheminform.* **2010**, *2*, 1−11.
(40) Sheridan, R. P.; Kearsley, S. K. Why do we need so many chemical similarity search methods? *Drug Discovery Today* **2002**, *7*, 903−911.
(41) Eckert, H.; Bajorath, J. Exploring peptide-likeness of active molecules using 2D fingerprint methods. *J. Chem. Inf. Model.* **2007**, *47*, 1366−1378.
(42) Duan, J.; Dixon, S. L.; Lowrie, J. F.; Sherman, W. Analysis and comparison of 2D fingerprints: insights into database screening performance using eight fingerprint methods. *J. Mol. Graphics Modell.* **2010**, *29*, 157−170.
(43) Kruger, D. M.; Evers, A. Comparison of structure- and ligand-based virtual screening protocols considering hit list complementarity and enrichment factors. *ChemMedChem* **2010**, *5*, 148−158.
(44) Sastry, M.; Lowrie, J. F.; Dixon, S. L.; Sherman, W. Large-scale systematic analysis of 2D fingerprint methods and parameters to improve virtual screening enrichments. *J. Chem. Inf. Model.* **2010**, *50*, 771−784.
(45) Heikamp, K.; Bajorath, J. How do 2D fingerprints detect structurally diverse active compounds? Revealing compound subset-specific fingerprint features through systematic selection. *J. Chem. Inf. Model.* **2011**, *51*, 2254−2265.
(46) Hu, G.; Kuang, G.; Xiao, W.; Li, W.; Liu, G.; Tang, Y. Performance Evaluation of 2D Fingerprint and 3D Shape Similarity Methods in Virtual Screening. *J. Chem. Inf. Model.* **2012**, *52*, 1103−1113.