JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

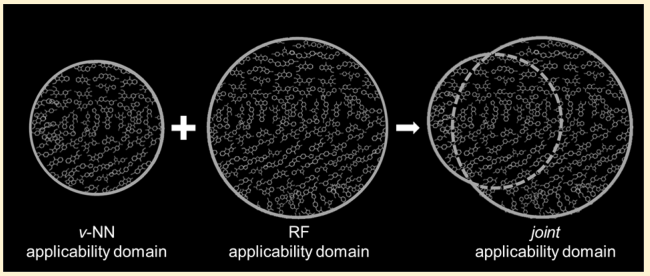
pubs.acs.org/jcim

Merging Applicability Domains for *in Silico* Assessment of Chemical Mutagenicity

Ruifeng Liu* and Anders Wallqvist*

DoD Biotechnology High Performance Computing Software Applications Institute, Telemedicine and Advanced Technology Research Center, U.S. Army Medical Research and Materiel Command, 2405 Whittier Drive, Frederick, Maryland 21702, United States

ABSTRACT: Using a benchmark Ames mutagenicity data set, we evaluated the performance of molecular fingerprints as descriptors for developing quantitative structure−activity relationship (QSAR) models and defining applicability domains with two machine-learning methods: random forest (RF) and variable nearest neighbor (*v*-NN). The two methods focus on complementary aspects of chemical mutagenicity and use different characteristics of the molecular fingerprints to achieve high levels of prediction accuracies. Thus, while RF flags mutagenic compounds using the presence or absence of small molecular fragments akin to structural alerts, the *v*-NN method uses molecular structural similarity as measured by fingerprint-based Tanimoto distances between molecules. We showed that the extended connectivity fingerprints could intuitively be used to define and quantify an applicability domain for either method. The importance of using applicability domains in QSAR modeling cannot be understated; compounds that are outside the applicability domain do not have any close representative in the training set, and therefore, we cannot make reliable predictions. Using either approach, we developed highly robust models that rival the performance of a state-of-the-art proprietary software package. Importantly, based on the complementary approach used by the methods, we showed that by combining the model predictions we raised the applicability domain from roughly 80% to 90%. These results indicated that the proposed QSAR protocol constituted a highly robust chemical mutagenicity prediction model.



INTRODUCTION

Mutagens are chemicals that can cause abnormal genetic mutations that underlie many cancers developed through environmental, drug, or toxicant exposures. A common assay for gauging mutagenicity is the Ames test.[1] This test uses several strains of genetically modified Salmonella bacteria with their histidine synthesis coding genes rendered ineffective through mutations. Thus, when placed in a histidine-deficient medium, these strains of Salmonella cannot survive. However, when placed in a medium with mutagens, reverse mutations that restore the functional capability of the bacteria to synthesize histidine enable bacterial colonies to grow in a histidine-deficient medium. To test a compound for mutagenicity, the bacteria are exposed to the compound in a histidine-deficient medium, and by comparing the numbers of bacterial colonies before and after the exposure, we can classify the compound as a mutagen or nonmutagen. Because many chemicals are metabolized in the liver and their mutagenic effect is due to different metabolic products, compounds are routinely tested in the presence of a mammalian metabolizing system otherwise not natively present in the Salmonella system.[2]

To date, thousands of chemicals have been evaluated using the Ames test, and it has become a standard assay for safety assessment of chemicals and drugs. Although experimental testing will always be required for novel compounds, *in silico* methods for predicting mutagenicity provide an efficient use of existing data for making highly reliable predictions for certain compound sets. On the basis of improved *in silico* predictive toxicology methods, including quantitative structure−activity relationship (QSAR) predictions of molecular toxicity, these methods have gained popularity and acceptance.[3−6] In addition, to meet the regulatory requirements for safety assessments, regulatory agencies have encouraged the use of QSAR predictions when experimental data are not available or as supplementary information.[4,7,8]

The rising popularity of QSAR methods has been tempered by an inadequate understanding of when a prediction should be made and whether it can be trusted. Most QSAR models are based solely on descriptors calculated from structures of the small molecules, but the models are used to predict complex properties that ultimately depend on interactions between these small molecules and intricate biological systems such as proteins, RNA, DNA, membranes, etc. Thus, deviations of a molecule from the training set compounds from which the model parameters were derived rapidly degrade performance, as the model can only capture heuristic relationships derived from a statistical analysis of known molecular components. The practical problem is then transformed into calculating measures

Received: January 9, 2014
Published: February 4, 2014

of prediction reliabilities that can flag an untrustworthy prediction and mark the compound for experimental evaluation. In other words, how do you know what you cannot predict?

Over the past few years, many studies have used the concept of an applicability domain for QSAR models to address the reliability of predictions.[6,9,10] As most QSAR studies use molecular physicochemical parameters and topological indices as descriptors, the applicability domain is typically defined by the value ranges of these descriptors, the value ranges of the principal components of these descriptors, or the distance between a compound and the training set compounds calculated on the basis of these descriptors. In our opinion, the most important premise for QSAR is that molecules with similar structures have similar activities. Hence, molecular descriptors that can capture structural similarities should be appropriate both for developing QSAR models *and* for defining the applicability domain of a model.

Molecular fingerprints are perhaps most efficient for describing molecular structural details, and they are routinely used in molecular similarity searches.[11] Previous studies have shown that when used with suitable statistical methods, high quality regression models can be developed using molecular fingerprints as descriptors.[12,13] It was also shown by Sheridan et al. that fingerprint-based molecular similarity is a good discriminator for QSAR prediction accuracy.[14] Here, we describe our study of using molecular fingerprints as descriptors for developing QSAR models and for defining applicability domains for *in silico* predictions of chemical mutagenicity.

We confirmed that using molecular fingerprints was an efficient way to construct high accuracy QSAR models and to define intuitive applicability domains for two machine-learning methods: random forest (RF) and variable nearest neighbor (v-NN). Importantly, we showed that by combining classifications based on the presence or absence of characteristic structural fragments via the RF model and chemical structural similarity using the v-NN model, we could significantly expand the applicability domain for *in silico* predictions of chemical mutagenicity from roughly 80% to 90% of all tested compounds. This provides a robust tool for *in silico* predictions of Ames mutagenicity and for identifying when a lack of prediction reliability necessitates experimental evaluations.

METHODS AND MATERIALS

Ames Data Set. We used the benchmark Ames mutagenicity data set compiled by Hansen et al.[15] This data set consists of 6512 compounds whose Ames test results were collected from different sources, with 3503 (53.8%) of them classified as Ames-positive. As one of the largest bioactivity data sets in the public domain, it is well suited for the development and validation of *in silico* predictive mutagenicity models. In this study, we divided the data set randomly into 10 equal-sized groups for 10-fold cross-validation; i.e., we used nine groups as a training set for model development and predicted mutagenicity of the 10th group. The process was repeated until each and every group was left out once for the evaluation of model performance. To compare performance with other studies, we also ran 5-fold cross-validation calculations. For the 5-fold cross-validation, we used the data set splits of Hansen et al.[15] That is, the data set was segregated into a static training set of 1585 compounds and five nearly equal-sized validation groups of close to 1000 compounds each. In the cross validation calculations, we combined the static training set with four of the five validation groups for model training, and the model thus derived was used to predict the mutagenicity of the excluded validation group. We repeated this process so that each of the validation groups was left out once and used to assess model performance.

Molecular Descriptors. In this study, we used the extended connectivity fingerprints[16] (ECFP) as molecular descriptors. The fingerprints were generated iteratively to encode features that represent each atom in larger and larger structural neighborhoods. At iteration 0 (ECFP_0), we encoded the information of individual atoms by turning on a corresponding bit in a binary bit string. The information includes the number of connections (bonds) to the atom, element type, charge, and atom mass. At iteration 1, we encoded the information of all atoms directly bonded to the atom (within a diameter of two chemical bonds, and hence termed ECFP_2). At iteration 2 (ECFP_4), we encoded the information on all atoms within a diameter of four chemical bonds. When the desired neighborhood size was reached, the process was complete, and the set of bits representing all features of the atom was returned as part of the molecular fingerprint. This process was repeated for all the atoms in a molecule. The molecular ECFP_n fingerprint is a collection of all the bits representing atoms in their molecular neighborhoods. Each bit represents a specific molecular structure moiety and is called a bit feature.

With increasing n, ECFP_n gives an increasingly more detailed description of molecular structures. However, with increasing n, the number of unique bit features increases exponentially, and so does the computational cost. A practical strategy to balance the cost and performance is to fold an original fingerprint into a fixed-length bit string by the logical OR operation.[17] Such folding leads to a loss of information due to bit-feature clashing. The degree of information loss is proportional to the degree of folding. For Tanimoto coefficient-based similarity searches of drug-sized molecules, a fixed-length bit string of 1024 bits works reasonably well, and it is the default bit length in many software packages.[17]

Variable Nearest Neighbor Method. On the basis of the premise of similar structures having similar activities, the k-nearest neighbor (k-NN) method should be well suited for QSAR, as it always uses the nearest neighbors to make a prediction. Indeed, it was one of the machine learning methods Hansen et al. used in their study of Ames mutagenicity, but they found that it underperformed compared to the other machine-learning methods.[15] A shortcoming of the method is that it always gives a prediction for a compound based on a constant number of nearest neighbors, irrespective of whether the nearest neighbors are structurally similar enough to ensure similar activity. To correct for this shortcoming, we recently proposed a variable number nearest-neighbor (v-NN) method.[18] Instead of using a constant number of nearest neighbors, v-NN uses all nearest neighbors meeting a structural similarity criterion for making a prediction. When no nearest neighbor meets the similarity criterion, we do not make a prediction in order to maintain the overall reliability of predictions. In essence, the predicted property y is made via a weighted average across structurally similar neighbors, as

$$y = \frac{\sum_{i=1}^{v} y_i \, e^{-\left(\frac{d_i}{h}\right)^2}}{\sum_{i=1}^{v} e^{-\left(\frac{d_i}{h}\right)^2}}, \; d_i \le d_0 \qquad (1)$$

dx.doi.org/10.1021/ci500016v | J. Chem. Inf. Model. 2014, 54, 793−800

where d_i denotes the Tanimoto distance between a target molecule for which a prediction is made and molecule i of the training set, y_i denotes the experimentally measured value of molecule i, v denotes the total number of training set molecules satisfying the condition $d_i \leq d_0$, h is a smoothing factor which dampens the distance penalty, and d_0 is a Tanimoto-distance threshold beyond which two molecules are not considered sufficiently similar to include in the average. For predicting chemical mutagenicity, we assigned a y_i value of 1 to all Ames-positive compounds and a value of 0 to all Ames-negative compounds in the benchmark data set. Using eq 1, the predicted molecular mutagenicity value falls between 0 and 1. A value below 0.5 classified a compound as nonmutagenic; otherwise, a compound was classified as mutagenic.

Random Forest Method. We also used the random forest (RF) method[19,20] for predicting mutagenicity[15] with special focus on developing an intuitive applicability domain using molecular fingerprints. To develop an RF model, we trained 500 decision trees. Each of them used a subset of ECFP_n bit features to recursively partition the training set samples so that mutagenic and nonmutagenic compounds were enriched in different branches. To predict the mutagenicity of a compound, we used all 500 decision trees. A compound was categorized as mutagenic if it was predicted positive by more than 50% of the trees. We used the RF module of the R Project for Statistical Computing[21] implemented in Pipeline Pilot[22] in this study.

Model Quality Measures. We used the following metrics to measure quality of the classification models:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2}$$

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{FP} + \mathrm{TN}} \tag{3}$$

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{4}$$

$$\kappa = \frac{\mathrm{ACC} - \Pr(\mathrm{e})}{1 - \Pr(\mathrm{e})} \tag{5}$$

where TPR is the rate of positives being predicted as positive (sensitivity), TNR is the rate of negatives being predicted as negative (specificity), ACC is the accuracy or probability of correct predictions (concordance), κ is a metric for assessing quality of binary classifiers,[23] and Pr(e) is an estimate of the probability of a correct prediction by chance, calculated by the equation $\Pr(\mathrm{e}) = (((\mathrm{TP} + \mathrm{FN})(\mathrm{TP} + \mathrm{FP}) + (\mathrm{FP} + \mathrm{TN})(\mathrm{TN} + \mathrm{FN}))/(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{TN})^2)$ In essence, κ compares the probability of correct predictions to the probability of correct predictions by chance. Its values range from +1 (perfect agreement between model prediction and experiment) to −1 (complete disagreement), with 0 indicating no agreement above that expected by chance. As a good measure of the quality of a binary classifier, κ's merit over ACC is easy to appreciate with an unbalanced data set, e.g., a data set in which 90% of the samples belong to one class and the remaining 10% of samples belong to another class. A meaningless classifier that simply assigns everything to the majority class would have a decent accuracy of 90% for such a data set, as no more than 10% of the samples would be incorrectly assigned. For such a data set, κ of the meaningless classifier would be 0, as Pr(e) of the meaningless classifier would be 90%.

In addition to the above metrics, we also considered coverage—the percentage of samples within the applicability domain for a given data set—as a quality measure. After all, a model offers little practical value if it has a very small applicability domain, even if it can give perfect predictions for a very small number of samples.

RESULTS AND DISCUSSION

Parameter Selection and Performance of the *v*-NN Method. With the v-NN method, we considered several adjustable parameters that may impact performance, including fingerprint size n, molecular structural similarity threshold d_0, and smoothing factor h.

Previous studies by Hert et al.[24] found that for similarity-based virtual screenings, ECFP_2 is inferior to ECFP_4, and Riniker and Landrum[25] observed that the fingerprints of ECFP_4 and ECFP_6 are highly correlated, with a squared correlation coefficient r^2 as high as 0.999 for most drug-sized molecules. We also performed extensive preliminary calculations using ECFP_2, ECFP_4, and ECFP_6 fingerprints using the Ames mutagenicity data. We found that ECFP_4 and ECFP_6 performed similarly, and both were slightly better than ECFP_2, presumably because ECFP_2 bit features are smaller and fewer and therefore give less detailed descriptions of molecular structures. On the basis of our calculations and previous work, we decided to use only ECFP_4 fingerprints in the v-NN calculations.

To determine an optimal Tanimoto distance threshold d_0 and smoothing factor h, we performed a number of 10-fold cross-validation calculations by increasing h stepwise from 0.1 to 1.0 (step size 0.1) and increasing d_0 stepwise from 0.05 to 0.75 (step size 0.05). We found that overall the model quality strongly depended on d_0 but much less so on h. Figure 1a shows model performance measures versus d_0 obtained at a constant smoothing factor of $h = 0.50$. For low d_0 values, model performance as measured by ACC and κ was high, but the coverage was very low, meaning that the majority of the compounds do not have near neighbors meeting the stringent molecular structural similarity requirement. With increasing d_0, model performance deteriorated gradually, whereas model coverage increased significantly. With a d_0 of 0.55, the model had coverage of 85%, ACC of 80%, sensitivity of 86%, and specificity of 73%. The ACC of 80% was close to the reported Ames assay inter- and intralaboratory reproducibility of 87%.[26]

Figure 1b shows the influence of the smoothing factor h on model performance obtained at a fixed d_0 of 0.55. The results showed that starting at the low end of h, model performance improved gradually with increasing h. All performance measures reached a plateau at around $h = 0.50$. The results indicated that the combination of $d_0 = 0.55$ and $h = 0.50$ gave a good compromise between performance and coverage.

Parameter Selection and Performance of the RF Method. To develop a reliable RF classification model, we evaluated the performance of ECFP_2, ECFP_4, and ECFP_6 fingerprints with different bit string lengths. As there are slightly more Ames-positive than Ames-negative compounds in the data set, we used the "equalize class size" option in the model training process to reduce bias. Each tree was trained with a subset of ECFP_n fingerprint bit features as molecular descriptors. The number of bit features of the subset impacts model quality, as a small number may lead to undersampling of the descriptor space, and a large number increases computational costs and may result in identical trees, which reduces the

dx.doi.org/10.1021/ci500016v | J. Chem. Inf. Model. 2014, 54, 793−800

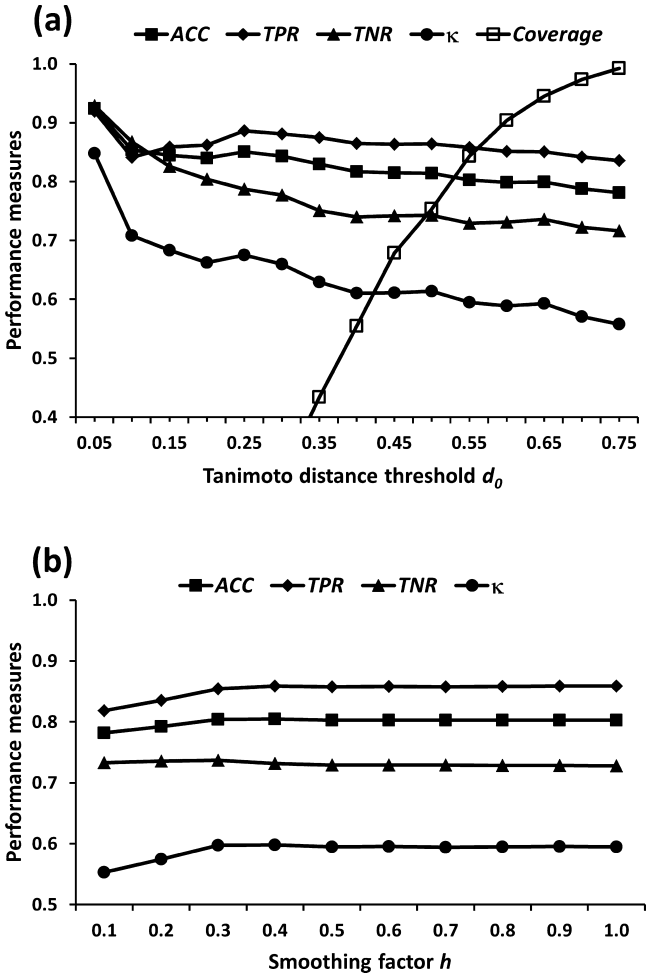


Figure 1. (a) Performance measures of v-NN with respect to Tanimoto distance threshold d_0 at a constant smoothing factor h of 0.50. (b) Performance measures of v-NN with respect to smoothing factor h at a constant Tanimoto distance threshold d_0 of 0.55. The coverage (84.3%) is not shown, as it is constant with a d_0 of 0.55. ACC: concordance. TPR: sensitivity. TNR: specificity. κ: kappa coefficient, an overall model quality measure.

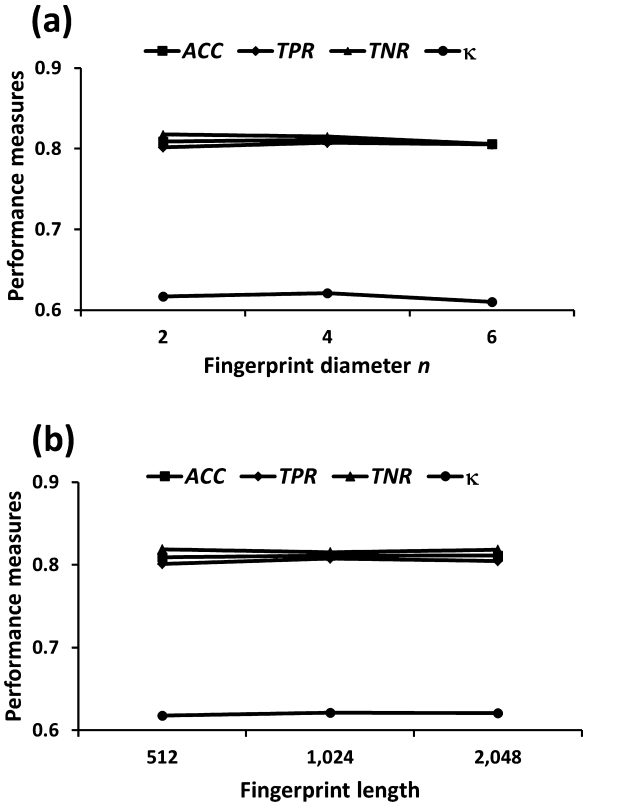


Figure 2. (a) Performance of random forest respective to ECFP_n fingerprint size n at a fixed fingerprint length of 1024 bits. (b) Performance of random forest respective to ECFP_4 fingerprint length. ACC: concordance. TPR: sensitivity. TNR: specificity. κ: kappa coefficient, an overall model quality measure.

power of the RF method. In this study, we used the square root of the total number of bit features in the fixed bit-length fingerprint as the number of descriptors for each tree.

To assess the impact of fingerprint size on RF model quality, we performed 10-fold cross-validation using ECFP_2, ECFP_4, and ECFP_6 fingerprints folded to a fixed length of 1024 bits as molecular descriptors. Figure 2a shows that the three fingerprints performed similarly, with ECFP_4 marginally better as judged by the magnitudes of κ. The similar performance of ECFP_2, ECFP_4, and ECFP_6 indicated that mutagenicity is mainly determined by the presence or absence of some rather small molecular fragments. This observation agrees with the fact that most genotoxicity structural alerts are small reactive structural moieties that react with DNA molecules.[27]

To assess the impact of fingerprint length on model performance, we carried out 10-fold cross-validation calculations using ECFP_4 fingerprints folded into 512 bits, 1024 bits, and 2048 bits, respectively. Figure 2b shows that the performance measures of the resulting models were nearly identical. The κ values were 0.618, 0.621, and 0.621 for models developed from the ECFP_4 fingerprints folded into 512, 1024, and 2048 bits, respectively. Because the performance measures were almost identical, we used a fixed length of 1024 bits for the remainder of the study.

While the performance of the RF model developed from ECFP_4 folded into a fixed length of 1024 bits appeared quite satisfactory, there is no indication of its limitations in terms of an applicability domain outside which the performance may be unreliable. As the decision trees used the presence/absence of fingerprint bit features of the training set compounds to predict the category of test compounds, and none of the training sets covered a significant portion of chemical space, there were certainly structural moieties strongly associated with mutagenicity not present in the training set. For molecules with these structural moieties, we expected the RF predictions to be unreliable. Thus, we hypothesized that the number of fingerprint bit features of a test compound not present in the model training process should give an indication of prediction reliability or the applicability domain of the model.

To investigate whether this hypothesis was correct, we performed 10-fold cross-validation calculations using ECFP_4 fingerprints as descriptors and keeping track of the number of ECFP_4 bit-features of the test set molecules that were missing from the training set molecules. Table 1 and Figure 3a show that for the 2898 compounds without missing bit features in the training set, the ACC is as high as 0.855 and κ is 0.694, the highest achieved in this study. With an increasing number of missing bit features, model quality initially deteriorated, then fluctuated somewhat to finally fall off. The performance measures were less meaningful toward higher numbers of

dx.doi.org/10.1021/ci500016v | J. Chem. Inf. Model. 2014, 54, 793−800

Table 1. Performance Measures of the Random Forest Model Respective to the Number of ECFP_4 Bit-Features Not Present among the Training Set compounds[a]

missing bit-features	number of compounds	ACC	TPR	TNR	κ
0	2898	0.855	0.894	0.794	0.694
1	1261	0.785	0.770	0.803	0.570
2	811	0.760	0.720	0.798	0.518
3	446	0.803	0.742	0.860	0.604
4	285	0.775	0.675	0.852	0.535
5	262	0.805	0.684	0.903	0.599
6	169	0.751	0.623	0.859	0.490
7	119	0.664	0.469	0.800	0.280
8	77	0.831	0.704	0.900	0.620
9	37	0.730	0.667	0.813	0.465
10	32	0.781	0.667	0.826	0.477
11	27	0.778	0.833	0.762	0.481
12	22	0.682	0.400	0.765	0.154
13	11	0.636	0.667	0.600	0.267
14	10	0.600	0.333	0.714	0.048
15	9	0.778	0.000	1.000	0.000
16	9	0.889	0.000	1.000	0.000
17	3	1.000	ND	1.000	ND
18	5	0.600	0.500	0.667	0.167
19	1	0.000	0.000	ND	0.000
20	3	0.667	0.000	1.000	0.000
21	3	1.000	ND	1.000	ND
22	2	1.000	ND	1.000	ND
23	1	1.000	ND	1.000	ND
24	1	1.000	ND	1.000	ND
25	1	1.000	ND	1.000	ND
28	2	1.000	ND	1.000	ND
29	1	1.000	ND	1.000	ND
32	2	1.000	ND	1.000	ND
35	1	0.000	0.000	ND	0.000
38	1	1.000	ND	1.000	ND

[a]The random forest consisted of 500 decision trees; each of the trees was based on 32 ECFP_4 fingerprint bit features as molecular descriptors. The performance measures were derived from 10-fold cross-validation calculations. ACC: concordance. TPR: sensitivity. TNR: specificity. κ: kappa coefficient, an overall model quality measure; ND: not defined.

missing bit features, as the number of compounds with a large number of missing bit features quickly approached 0. The results confirmed that the number of missing bit-features was a reasonable measure of the applicability domain of our RF classification models for mutagenicity.

As the description of the applicability domain does not have to be strictly connected to the fingerprints used for the prediction, we examined the ECFP_2 bit features as an alternative set. Although for a given molecule, ECFP_4 had considerably more unique bit features than ECFP_2, all ECFP_2 bit-features were contained in the ECFP_4 bit features. Figure 2 shows that both fingerprints performed roughly equally well in distinguishing mutagenic and non-mutagenic compounds, and hence, one might also use the number of missing ECFP_2 bit-features as an indicator of the applicability domain of the RF models. To confirm this, we tracked the missing ECFP_2 bit features and the performance measures of the RF models. Table 2 and Figure 3b show clear model deterioration with an increasing number of missing ECFP_2 bit features accompanied by diverging sensitivity and

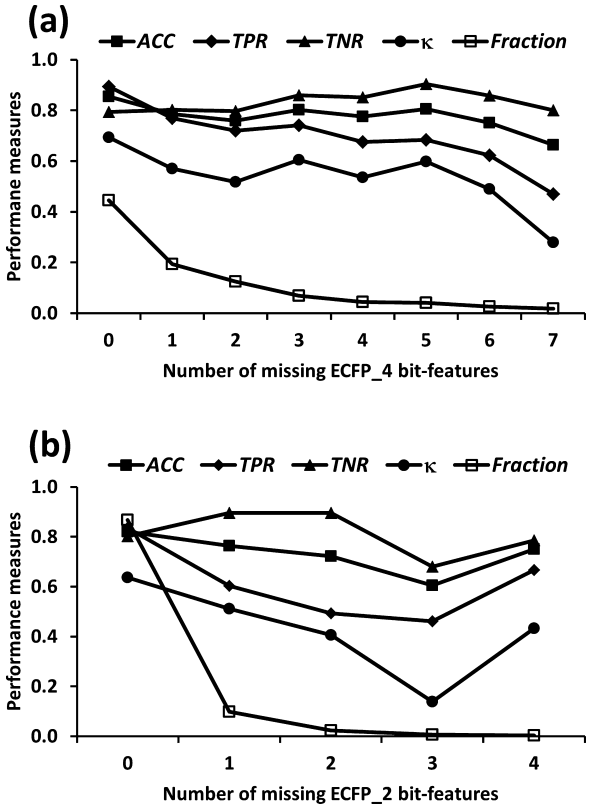


Figure 3. (a) Performance of random forest respective to the number of missing ECFP_4 fingerprint bit features in the training set. (b) Performance of random forest respective to the number of missing ECFP_2 fingerprint bit features in the training set. *Fraction* denotes the fraction of compounds in category. ACC: concordance. TPR: sensitivity. TNR: specificity. κ: kappa coefficient, an overall model quality measure.

Table 2. Performance Measures of Random Forest Model with Respective to the Number of ECFP_2 Bit Features Not Present among the Training Set Compounds[a]

missing bit features	number of compounds	ACC	TPR	TNR	κ
0	5651	0.821	0.835	0.802	0.637
1	644	0.764	0.603	0.895	0.511
2	151	0.722	0.492	0.895	0.406
3	38	0.605	0.462	0.680	0.139
4	20	0.750	0.667	0.786	0.432
5	6	1.000	ND	1.000	ND
6	2	0.000	0.000	ND	0.000

[a]The random forest consisted of 500 decision trees; each of the trees was based on 32 ECFP_4 fingerprint bit features as molecular descriptors. The performance measures were derived from 10-fold cross-validation calculations. ACC: concordance. TPR: sensitivity. TNR: specificity. κ: kappa coefficient, an overall model quality measure. ND: not defined.

specificity. For the 86.8% of the compounds without any missing ECFP_2 bit features, the RF model has an ACC value of 0.821 and a very small gap between sensitivity (0.835) and specificity (0.802), indicating that the model predicted positives and negatives with nearly the same high accuracy. The advantages of using the smaller ECFP_2 bit features as a measure of the applicability domain lie both in the sensitivity of the model with respect to the fingerprints themselves and in

dx.doi.org/10.1021/ci500016v | J. Chem. Inf. Model. 2014, 54, 793−800

their interpretation. Given a missing fingerprint, it is easy to identify molecules or a scaffold that will complement the training data set. Thus, for the benchmark Ames mutagenicity data set, we defined the applicability domain of the RF model as the chemical space occupied by molecules whose ECFP_2 bit features were all present in the training set molecules.

Union of the *v*-NN and RF Applicability Domains. One of the limiting factors for developing a high-quality QSAR model is the availability of a large number of structurally diverse compounds for model training. Because of limited coverage of the chemical space by small training sets, the applicability domains of QSAR models are usually small. For practical applications, expanding the applicability domain is as important as improving the accuracy of a QSAR model.

To predict molecular mutagenicity, we initially evaluated the performance of *v*-NN and RF methods and defined their applicability domains separately. Fundamentally, each method focuses on different structural/chemical aspects of chemically induced mutagenicity. Thus, the RF predictions were based on the presence or absence of certain molecular structural moieties, without consideration of the structure of the whole molecule. This model framework captures mechanisms of chemical mutagenicity due to the formation of chemical bonds between the reactive structural moieties of mutagens and DNA molecules. The *v*-NN method, on the other hand, makes predictions based on whole-molecule structural similarity between a target compound and mutagens in the training set. This framework is intended to identify mutagens that share noncovalent binding to DNA or to proteins involved in DNA damage, repair, and maintenance functions. Efficient noncovalent binding typically involves both a shape and interaction complementarity that allows the mutagen to bind to structurally compatible regions of the DNA or protein. Because the RF and *v*-NN predictions may relate to different mechanisms of chemical mutagenicity, a nonconsensus combination of the two should provide reliable predictions for a broader range of chemicals than either method. To test this hypothesis, we implemented the following prediction procedure for a given compound: (1) If the compound belongs to the applicability domain of the RF model, mutagenicity of the compound is predicted by the RF model; (2) *else*, if the compound is within the applicability domain of the *v*-NN model, mutagenicity of the compound is predicted by the *v*-NN model; (3) *else*, the compound is considered outside the union of the applicability domains of both models and no prediction is made.

Table 3 shows the results of 10-fold cross-validation calculations based on this procedure. The union of the applicability domains provided the study-highest coverage of 93%, an ACC of 82%, and a κ value of 0.63. These results indicated that the proposed protocol constitutes a highly robust and broadly applicable chemical mutagenicity prediction model.

Table 3. Results of the Combined RF/*v*-NN Approach in the Applicability Domain Derived from 10-Fold Cross-Validation Calculations[a]

number of compounds	ACC	TPR	TNR	κ	coverage
6051	0.816	0.830	0.799	0.628	0.929

[a]ACC: concordance. TPR: sensitivity. TNR: specificity. κ: kappa coefficient, an overall model quality measure. Coverage: fraction of test set compounds in the applicability domain.

Comparison with MultiCase for PC (MC4PC). To compare performance with previous studies, we performed 5-fold cross-validation calculations using the splits of the data set of Hansen et al.[15] The five validation groups contain 984 to 987 compounds each. To replicate their evaluation metrics, we combined the static training set of 1585 compounds with four of the five validation groups to develop a model and used the model to make predictions for the excluded validation group. Table 4 shows the results of the 5-fold cross-validation using the RF, *v*-NN, and combined RF/*v*-NN methods with those of MC4PC.[28] The performance measures of the RF model were nearly identical to those of MC4PC, except the RF model had a higher coverage (81% vs 85%). The *v*-NN model had a slightly larger gap between TPR and TNR than the MC4PC model, but with a higher coverage of 84%. The combined RF/*v*-NN approach achieved nearly identical prediction accuracy as MC4PC in terms of ACC, TPR, and TNR but had a significantly higher coverage of 92%. Considering that the inter- and intralaboratory consistency of Ames assays is about 87%, the prediction accuracy of 80% achieved by these methods is close to the upper limit of a practical prediction model. The 13% higher coverage of the combined RF/*v*-NN approach indicated that it could give the same highly reliable predictions for a significantly broader range of chemicals, making it a robust method for chemical mutagenicity predictions.

Table 4. Performance of Different Methods in Their Respective Applicability Domains Derived from 5-Fold Cross-Validation Calculations

method	ACC	TPR	TNR	κ	coverage
MC4PC[a]	0.791	0.842	0.732		0.810
RF[b]	0.797	0.847	0.723	0.574	0.853
v-NN[c]	0.788	0.864	0.668	0.546	0.843
RF/*v*-NN[d]	0.794	0.842	0.723	0.570	0.921

[a]MultiCase for PC, a proprietary software of MultiCase, Inc. [b]Random forest. [c]Variable nearest neighbor method. [d]Combined RF/*v*-NN approach. ACC: concordance. TPR: sensitivity. TNR: specificity. κ: kappa coefficient, an overall model quality measure. Coverage: fraction of test set compounds in applicability domain.

CONCLUSIONS

We have shown that both the *v*-NN and RF models yielded results comparable to those of MC4PC, a state-of-the-art computational toxicology software package, on a large benchmark Ames mutagenicity data set. More importantly, we showed that the combination of the models provided a more robust approach for computational prediction of chemical mutagenicity, with an enhanced applicability domain compared to either individual methods or MC4PC. The enhancement is not due to a consensus prediction; rather, the improvement stems from extending the applicability domain. The results also demonstrated that extended connectivity fingerprints are excellent molecular descriptors for both QSAR model development and for defining their applicability domains. A distinct advantage of molecular fingerprints as QSAR descriptors is that the results are directly associated with readily interpretable and actionable structural information. For example, when defining the applicability domain of a QSAR model by the number of missing ECFP_2 bit features, a straightforward approach to broaden the applicability domain is to bring molecules with ECFP_2 bit features not present in the training set into the

training set and retrain the model. In contrast, when defining the applicability domain by conventional molecular descriptors (such as molecular physicochemical parameters, topological indices, E-state indices, etc.), there is no such intuitive approach to understand and expand the applicability domain.

Generally speaking, it may be preferable to define an applicability domain independently of the statistical methods used for model development, in order to use a single applicability domain to gauge the reliability of predictions by different QSAR models. A universal applicability domain, however, will not work well when multiple and significantly different molecular mechanisms of action are involved. Because each mechanism has its own structure−activity relationship, an applicability domain valid for one QSAR model accounting for a particular mechanism is unlikely to be valid for another QSAR model accounting for a different mechanism.

Because discovery projects constantly break new ground and venture into previously unexplored chemical space, the performance of most QSAR models tends to deteriorate over time. A commonly adopted approach to keep QSAR models up to expectations is to periodically retrain the models with up-to-date experimental data. In this respect, the *v*-NN method is superior, as it does not build a static model but makes predictions on the fly. As long as the *v*-NN model can retrieve data from a repository of up-to-date experimental results, the *v*-NN predictions are always up to date. In addition, *v*-NN is perhaps the least complex machine learning method, yet its predictive performance on the benchmark mutagenicity data set rivals that of other methods. Another advantage of *v*-NN is that the method incorporates the concept of an applicability domain directly in its predictions and flags when a compound is outside the applicability domain, thus alerting the user that an alternative assessment or new experimental measurements are required.

AUTHOR INFORMATION

Corresponding Author

*Phone: 301-619-1989. Fax: 301-619-1983. E-mail: RLiu@bhsai.org (R.L.), AWallqvist@bhsai.org (A.W.).

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors were supported by the U.S. Army Medical Research and Materiel Command (Ft. Detrick, MD), as part of the U.S. Army's Network Science Initiative, and the Defense Threat Reduction Agency grant CBCall14-CBS-05-2-0007. The opinions and 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 the U.S. Department of Defense. This paper has been approved for public release with unlimited distribution.

REFERENCES

(1) Ames, B. N.; Durston, W. E.; Yamasaki, E.; Lee, F. D. Carcinogens are mutagens: a simple test system combining liver homogenates for activation and bacteria for detection. *Proc. Natl. Acad. Sci. U. S. A.* **1973**, *70*, 2281−2285.

(2) Mortelmans, K.; Zeiger, E. The Ames Salmonella/microsome mutagenicity assay. *Mutat. Res.* **2000**, *455*, 29−60.

(3) Bailey, A. B.; Chanderbhan, R.; Collazo-Braier, N.; Cheeseman, M. A.; Twaroski, M. L. The use of structure-activity relationship analysis in the food contact notification program. *Regul. Toxicol. Pharmacol.* **2005**, *42*, 225−235.

(4) Kruhlak, N. L.; Benz, R. D.; Zhou, H.; Colatsky, T. J. (Q)SAR modeling and safety assessment in regulatory review. *Clin. Pharmacol. Ther.* **2012**, *91*, 529−534.

(5) Naven, R. T.; Greene, N.; Williams, R. V. Latest advances in computational genotoxicity prediction. *Expert Opin. Drug Metab. Toxicol.* **2012**, *8*, 1579−1587.

(6) *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationships [(Q)SAR] Models*; OECD Environment Health and Safety Publications: Paris, 2007.

(7) U. S. Food and Drug Administration: Center for Drug Evaluation and Research. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH). M7 (2013): Assessment and Control of DNA Reactive (Mutagenic) Impurities in Pharmaceuticals to Limit Potential Carcinogenic Risk. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM347725.pdf (accessed January 23, 2014).

(8) Sutter, A.; Amberg, A.; Boyer, S.; Brigo, A.; Contrera, J. F.; Custer, L. L.; Dobo, K. L.; Gervais, V.; Glowienke, S.; van Gompel, J.; Greene, N.; Muster, W.; Nicolette, J.; Reddy, M. V.; Thybaud, V.; Vock, E.; White, A. T.; Muller, L. Use of in silico systems and expert knowledge for structure-based assessment of potentially mutagenic impurities. *Regul. Toxicol. Pharmacol.* **2013**, *67*, 39−52.

(9) Netzeva, T. I.; Worth, A.; Aldenberg, T.; Benigni, R.; Cronin, M. T.; Gramatica, P.; Jaworska, J. S.; Kahn, S.; Klopman, G.; Marchant, C. A.; Myatt, G.; Nikolova-Jeliazkova, N.; Patlewicz, G. Y.; Perkins, R.; Roberts, D.; Schultz, T.; Stanton, D. W.; van de Sandt, J. J.; Tong, W.; Veith, G.; Yang, C. Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. The report and recommendations of ECVAM Workshop 52. *Altern. Lab. Anim.* **2005**, *33*, 155−173.

(10) Sahigara, F.; Mansouri, K.; Ballabio, D.; Mauri, A.; Consonni, V.; Todeschini, R. Comparison of different approaches to define the applicability domain of QSAR models. *Molecules* **2012**, *17*, 4791−4810.

(11) Willett, P. Similarity searching using 2D structural fingerprints. *Methods Mol. Biol.* **2011**, *672*, 133−158.

(12) Liu, R.; Zhou, D. Using molecular fingerprint as descriptors in the QSPR study of lipophilicity. *J. Chem. Inf. Model.* **2008**, *48*, 542−549.

(13) Zhou, D.; Alelyunas, Y.; Liu, R. Scores of extended connectivity fingerprint as descriptors in QSPR study of melting point and aqueous solubility. *J. Chem. Inf. Model.* **2008**, *48*, 981−987.

(14) Sheridan, R. P.; Feuston, B. P.; Maiorov, V. N.; Kearsley, S. K. Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1912−1928.

(15) Hansen, K.; Mika, S.; Schroeter, T.; Sutter, A.; ter Laak, A.; Steger-Hartmann, T.; Heinrich, N.; Muller, K. R. Benchmark data set for in silico prediction of Ames mutagenicity. *J. Chem. Inf. Model.* **2009**, *49*, 2077−81.

(16) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

(17) Daylight Chemical Information Systems, Inc. Fingerprints - Screening and Similarity. http://www.daylight.com/dayhtml/doc/theory/theory.finger.html (accessed January 9, 2014).

(18) Liu, R.; Tawa, G.; Wallqvist, A. Locally weighted learning methods for predicting dose-dependent toxicity with application to the human maximum recommended daily dose. *Chem. Res. Toxicol.* **2012**, *25*, 2216−2226.

(19) Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5−32.

(20) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random forest: a classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947−1958.

(21) R: The R Project for Statistical Computing. http://www.r-project.org/ (accessed January 9, 2014).

(22) Accelrys: Pipeline Pilot. http://accelrys.com/products/pipeline-pilot/ (accessed January 9, 2014).
(23) Dunn, G.; Everitt, B. *Clinical Biostatistics: An Introduction to Evidence-based Medicine*; Edward Arnold: London, 1995.
(24) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *Org. Biomol. Chem.* **2004**, *2*, 3256−3266.
(25) Riniker, S.; Landrum, G. A. Open-source platform to benchmark fingerprints for ligand-based virtual screening. *J. Cheminform.* **2013**, *5*, 26.
(26) Kamber, M.; Fluckiger-Isler, S.; Engelhardt, G.; Jaeckh, R.; Zeiger, E. Comparison of the Ames II and traditional Ames test responses with respect to mutagenicity, strain specificities, need for metabolism and correlation with rodent carcinogenicity. *Mutagenesis* **2009**, *24*, 359−366.
(27) Rosenkranz, H. S.; Klopman, G. Structural alerts to genotoxicity: the interaction of human and artificial intelligence. *Mutagenesis* **1990**, *5*, 333−361.
(28) Saiakhov, R. D.; Klopman, G. Benchmark performance of MultiCASE Inc. software in Ames mutagenicity set. *J. Chem. Inf. Model.* **2010**, *50*, 1521−1521.