Computational Prediction of the Chromosome-Damaging Potential of Chemicals

Andreas Rothfuss,*,† Thomas Steger-Hartmann,† Nikolaus Heinrich,‡ and Jörg Wichard‡,§

Experimental Toxicology, Schering AG, D-13342 Berlin, Germany, Computational Chemistry, Schering AG, D-13342 Berlin, Germany, and Molecular Modeling Group, FMP, D-13125 Berlin, Germany

Received June 21, 2006

We report on the generation of computer-based models for the prediction of the chromosome-damaging potential of chemicals as assessed in the *in vitro* chromosome aberration (CA) test. On the basis of publicly available CA-test results of more than 650 chemical substances, half of which are drug-like compounds, we generated two different computational models. The first model was realized using the (Q)SAR tool MCASE. Results obtained with this model indicate a limited performance (53%) for the assessment of a chromosome-damaging potential (sensitivity), whereas CA-test negative compounds were correctly predicted with a specificity of 75%. The low sensitivity of this model might be explained by the fact that the underlying 2D-structural descriptors only describe part of the molecular mechanism leading to the induction of chromosome aberrations, that is, direct drug−DNA interactions. The second model was constructed with a more sophisticated machine learning approach and generated a classification model based on 14 molecular descriptors, which were obtained after feature selection. The performance of this model was superior to the MCASE model, primarily because of an improved sensitivity, suggesting that the more complex molecular descriptors in combination with statistical learning approaches are better suited to model the complex nature of mechanisms leading to a positive effect in the CA-test. An analysis of misclassified pharmaceuticals by this model showed that a large part of the false-negative predicted compounds were uniquely positive in the CA-test but lacked a genotoxic potential in other mutagenicity tests of the regulatory testing battery, suggesting that biologically nonsignificant mechanisms could be responsible for the observed positive CA-test result. Since such mechanisms are not amenable to modeling approaches it is suggested that a positive prediction made by the model reflects a biologically significant genotoxic potential. An integration of the machine-learning model as a screening tool in early discovery phases of drug development is proposed.

Introduction

Screening approaches for determining the genotoxic potential of new compounds play a pivotal role during hit validation and lead characterization phases of drug development in pharmaceutical companies. Traditionally, the assessment of the genotoxic potential of drug substances was typically performed during early developmental stages by conducting a standard set (battery) of genotoxicity tests that support the submission of novel drugs to regulatory agencies. As outlined in the respective ICH[1] guidelines (*1*), this standard set generally consists of a bacterial gene mutation test (Ames test), an *in vitro* cytogenetic assay in mammalian cells for the detection of chromosomal damage (e.g., a chromosome aberration (CA-) test) and an *in vivo* cytogenetic assay in rodent hematopoietic cells.

Today, pre-regulatory genotoxicity tests are frequently performed in pharmaceutical companies because of increased compound throughput and in order to avoid late stage termination of a cost-intensive drug development due to unforeseen genotoxicity. Such screening strategies primarily rely on *in vitro* assays, which often represent a cut down version of the respective regulatory tests (e.g., Ames II) or make use of alternative assays (e.g., the in vitro micronucleus test for the detection of chromosomal damage). In principle, the concordance between screening assays and regulatory tests is relatively high (*2*, *3*). However, in particular with respect to screening assays for chromosomal damage, they are at best medium throughput and as such their use in early discovery stages is restricted because of costs and compound availability. Additionally, genotoxicity screens might be biased by the frequent presence of (genotoxic) impurities in early research drug batches leading to potentially false positive results.

As an alternative, computational (*in silico*) structure−activity models have gained increasing importance in the assessment of a genotoxic potential. They have the clear advantage that no compound is needed for testing and that they can be applied in a true high-throughput manner. Computational programs used for genotoxicity prediction are mainly focusing on the prediction of the outcome of the Ames test and relatively good predictive accuracies (>70%) can be reached for this endpoint (*4*). In practice, however, it is not sufficient to solely predict bacterial mutagenicity because results from in silico genotoxicity predictions are frequently used as part of the decision process during drug discovery. Instead, it is desirable to also be able to model the chromosome-damaging potential of compounds in order to fully cover the basic regulatory mutagenicity tests.

* To whom correspondence should be addressed. Phone: +49-(0)30 46815268. Fax: +49-(0)30 46815091. E-mail: andreas.rothfuss@schering.de.
† Experimental Toxicology, Schering AG.
‡ Computational Chemistry, Schering AG.
§ Molecular Modeling Group.

[1] Abbreviations: CA-Test, chromosome aberration test; ICH, International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use; knn, k-nearest neighbour; QSAR, quantitative structure-activity relationship; SAR, structure-activity-relationship; SVM, support vector machine.

10.1021/tx060136w CCC: $33.50
Published on Web 09/29/2006 PAGE EST: 6.7

However, in contrast to the Ames test prediction, no models with comparable performance are currently available for the CA-test. Several reasons might account for this situation. The good correlation for the Ames test is based on the abundance of (publicly) available data for this test system as well as on the fact that most of the molecular mechanisms underlying this genetic endpoint are fairly well understood and can be directly related to the chemical structure (*5*). The situation is clearly more complex for the CA-test. It is well-established that different mechanisms can lead to the microscopically visible formation of aberrant chromosomes. Structural chromosome aberrations can be formed by direct drug−DNA interactions as a result of incorrect DNA repair processes (*6*) or an interaction of drugs with enzymes involved in DNA replication and transcription (*7*). Numerical chromosome aberrations such as the gain or loss of chromosomes are generally a result of the interaction with cellular proteins involved in chromosome segregation (*8*). In addition, it is well-known that nonphysiological stimuli during cell culture, such as those induced by excessive cytotoxicity, osmolarity, pH and temperature, can also lead to structural chromosome aberrations (*9*).

Furthermore, the CA-test is experimentally less standardized than the Ames test (i.e., different cells from different species are used), and publicly available experimental data is significantly less abundant than Ames test data and almost purely qualitative (i.e., aberration frequencies are hardly available). Most importantly, the quality of available CA-test data is frequently compromised by incomplete assay data sets and differences in the judgment of a positive effect, in particular in the presence of cytotoxicity (*10*). High-qualitative CA-test data might, in principle, be derived from publicly available data on pharmaceuticals because they are likely to be conducted using ICH and GLP-compliant methods. However, such public data are relatively scarce, and in particular, the number of positive results is limited.

Consequently, only few publications are available in which the performance of computational models for the prediction of CA-test data has been assessed. Using the MULTICASE (MCASE, Beachwood, USA) methodology for constructing experimental databases that can be used to predict the bioactivity of compounds, Rosenkranz et al. (*11*) reported the construction of a CA-test prediction model based on 233 compounds. These, mostly organic compounds, were assessed in a CA-test as part of the National Toxicology Program (NTP), with approximately 40% of the compounds being tested positive. Using an internal validation strategy, the observed sensitivity and specificity (i.e., the correct prediction of positives and negatives, respectively) of the model were 53% and 71%, respectively (*12*).

More recently, Serra et al. (*13*) reported on the generation of an automated machine-learning approach to generate classification models for the prediction of CA-test data. Support vector machines (SVM) and k-nearest neighbor (knn) models were developed on a set of molecular descriptors calculated for 346 mostly organic compounds (29% positives). Using a prediction set of 37 compounds that were not included in model formation, sensitivity and specificity values of 73% and 92%, respectively, were obtained for knn classification models. Similar values were obtained for SVM models.

Despite the respectable performance characteristic, of the latter model in particular, their value for a routine in silico CA-test screening during early drug development seems to be questionable. First, the number of CA-test positive compounds used for model building and evaluation in the Serra model (*13*) appears to be critically low. Less than one-third of the compounds tested positive in the CA-test, and thus, it seems questionable as to whether similar performance characteristics and conclusions had been obtained using a more balanced data set containing equal numbers of active and inactive compounds. Second, the structural diversity (chemical space) of compounds represented in the MCASE and machine learning model (*12*, *13*) is clearly limited to mainly organic compounds, such as agrochemicals, known carcinogens, and industrial chemicals. It was already noticed during the course of Ames-test modeling that computational models, which were predominantly constructed using industrial and environmental compounds, performed in a clearly poorer manner when applied to pharmaceutical compounds (*14*−*16*). This is an important implication if a computational prediction model for the CA-test has to be developed as a screening tool during early drug discovery.

In the present study, we therefore aimed to construct and evaluate two different computational models based on a heterogeneous data set including a significant number of pharmaceutical compounds to be used in genotoxicity screening approaches in a pharmaceutical environment. The recent publication of two data collections (*10*, *16*) containing qualitative CA-test information on more than 650 compounds, including a significant number of pharmaceuticals and drug-like compounds, allowed us to readdress the issue of modeling a chromosome-damaging potential on the basis of the largest high-quality data collection currently publicly available.

Materials and Methods

CA-Test Data Information. The CA-test data used in this study were obtained from two recently published data collections (*10*, *16*). Further details on the original data source can be obtained from the references of both data compilations.

The genotoxicity data collection from Snyder et al. (*16*) contains *in vitro* cytogenetics data for 248 marketed pharmaceuticals, with positive (i.e., chromosome-damaging) results being reported for 48/248 compounds (19%). Structural information could be retrieved for 229 of the 248 compounds. Altogether, 189 negative and 40 positive data records from this data source could be used for model-building purposes. As outlined in the article (*16*) and described in more detail in a previous collection effort (*17*), the *in vitro* cytogenetic data represents CA-test results obtained with diverse cell types (Chinese hamster ovary cells, Chinese hamster lung cells, V79 cells, MCL-5 human lymphoblastoid cells, and human blood peripheral lymphocytes). Despite this obvious methodological diversity, the overall quality of the data set and the reliability of the test result are judged to be high because the data has been generated according to standardized ICH- and GLP-compliant methods.

The CGX database collected by Kirkland et al. (*10*) contains CA-test data for 488 structurally diverse compounds, consisting of industrial, environmental, and pharmaceutical compounds. Out of a total number of 488 chemicals, 292 (60%) were considered positive, and 28 were judged to be equivocal. The latter were excluded from our model building. Structural information was retrieved for 450 out of the 460 remaining compounds. Altogether, 168 negative and 282 positive data records from this data source could be used for model-building purposes. Similar to the Snyder CA-test collection, results obtained with all cell types are included in this compilation. With respect to data quality, considerable effort was undertaken to review collected test results (*10*) suggesting an overall consistent evaluation of test data. In order to estimate the number of drug-like compounds contained in this dataset, we analyzed all 450 compounds for drug-likeness using a proprietary in-house software based on the model proposed by Sadowski and Kubinyi (*18*). Less than one-third of the compounds taken from Kirkland et al. (*10*) were considered as drug-like (data not shown), thus confirming that both data sources can roughly be separated

Table 1. Data Sets Used for Model Generation

	CA-test positive	CA-test negative
Total Data		
Kirkland et al. (2005)	282	168
Snyder et al. (2004)	40	189
	322 (47%)	357 (53%)
MCASE Model		
	CA-test positive	CA-test negative
training set	251	286
prediction set	47	53
ML Model		
	CA-test positive	CA-test negative
training set	252	282
prediction set	70	75

into drug-like (*16*) and less drug-like (Kirkland et al., 2005) compounds.

As summarized in Table 1, 679 compounds were used in total for model generation, of which 322 tested positive (47%) and 357 tested negative (53%) in the CA-test.

Collection of Structures. CAS numbers of identified substances were collected from the respective data collections (*10*, *16*) and queried in the MDL Toxicity database (MDL Information Systems Inc., San Leandro, CA). The retrieved chemical structures were stored as an sd file (MDL ISIS sdf file). For MCASE prediction model construction, SMILES notations of all compounds were generated by running the sd files through an existing prediction module in MCASE (Muticase Inc, Beachwood, OH), which generated a text file containing the respective SMILES code of the queried compounds.

Model Construction and Validation in MCASE. A hallmark of the MCASE software is its capability to automatically generate prediction modules on the basis of structural information and associated bioactivity (*19*). Details on model generation and software algorithms are published elsewhere (*20*). In essence, the program identifies structural fragments, ranging from 2 to 10 atoms length, in combination with2D distances between atoms, which are statistically correlated with activity (biophores) and inactivity (biophobes), respectively. In addition, the program detects fragments that act as modulators of activity and takes into account basic physicochemical descriptors for the module development process. A limitation of MCASE is that compounds containing ions, molecular clusters (such as hydrates), and rare atoms (such as Mn, Ca, or K) are not accepted for model generation. Consequently, compounds containing such structural features were automatically eliminated from the training set by the program during model construction.

From the overall data set containing 679 data records, 100 compounds (15%; 53 negative and 47 positive compounds) were randomly removed before model building and used as a prediction set to assess model predictivity. A training set was created out of the remaining 579 compounds (304 negative and 275 positive compounds). Because of MCASE's structural limitations, the automatically generated MCASE model for CA-test prediction contained 537 compounds (286 negatives and 251 positives). The predictivity of the generated model was assessed by internal and external validation. For the internal validation, 10 separate, non-overlapping sets consisting of 53 compounds (10% of the training set) were randomly selected from the training set and compiled as test sets. The remaining 90% of the individual learning sets were then used to predict the 53 compounds of the test set. For external validation, the initially removed 100 compounds (prediction set) were predicted by the MCASE model. As performance parameters, average values for sensitivity (ratio of correctly predicted positive compounds to all positives), specificity (ratio of correctly predicted negative compounds to all negatives), and concordance (ratio of correctly predicted compounds to total number of compounds) were assessed.

Table 2. Performance Characteristics for the MCASE Model

	coverage[b] [%]	sensitivity	specificity	concordance	X^2
training set[a]	93	52.8%	75.0%	64.9%	4.90 ($p < 0.05$)
prediction set[a]	94	56.8%	71.7%	65.1%	6.89 ($p < 0.01$)

[a] Mean values of 10 independent validations. [b] Percentage of 2−8 atom fragments structurally represented in the training set.

Table 3. List of Some Significant Biophores Identified in the MCASE Model

fragment	present in no. of cmpds: CA-positive	CA-negative	structural representation
NH2-c═cH-cH═	23	4	aromatic amine
NO-N	13	1	N-nitroso
Cl-CH2	16	5	halogenated alkane
C═C-CH═C-	13	1	α,β-unsaturated
cH═cH-c═cH-cH═c•—	7	1	aromatic ring with ET-drawing group (e.g., NH2)
O^-CH2	4	0	epoxide

Machine Learning (ML) Model. For the machine-learning model, 10% of the data was randomly removed and used to assess the performance of the final ML model (prediction set, see below). The remaining 90% of the data was designated as a training set and used for model generation.

The process of ML model generation can be separated into three distinct processes. First, a broad set of molecular descriptors encoding a variety of properties of the molecules are calculated for each compound of the training set. Next, redundant information of descriptors is removed via a process called feature selection, resulting in a small subset of the most useful descriptors. Finally, a classification model is built on the basis of the identified descriptors and validated using a set of data that was not previously included in the model-building effort.

Descriptor Generation and Feature Selection. All descriptors used in the ML model were calculated with the dragonX software (*21*) that was originally developed by Milano Chemometrics and the QSAR Research Group. The software generates a total number of 1664 molecular descriptors that are group into 20 different blocks, such as constitutional descriptors, topological descriptors, and walk and path counts (*22*). For each compound in the training set, all 1664 descriptors were calculated. Because many of these descriptors are redundant or carry correlated information, feature selection processes need to be performed in order to select the most useful subset of descriptors to build a ML model.

Our feature selection approach follows the method of variable importance as proposed by Breiman (*23*). The underlying idea is to select descriptors on the basis of the decrease of classification accuracy after the permutation of the descriptors (*24*). Briefly, an ensemble of decision trees is built, which uses all descriptors as input variables and associated activity (CA-test result) as output variables using 90% of the data (training set). The prediction accuracy of the classification model on an out of training portion of the data (test set) is recorded. In a second step, the same is done after the successive permutation of each descriptor. The relative decrease of classification accuracy is the variable importance following the idea that the most discriminative descriptors are the most important ones. We first separately calculated the variable importance of each descriptor of the 20 blocks of molecular descriptors and selected the most important ones. This descriptor set was reduced in a second iteration, resulting in a final set of 14 descriptors (Table 4).

Building the Machine Learning Classification Model. An ensemble approach was used to build the final classification model

Table 4. List of DragonX Descriptors Used in the Machine-Learning Model

descriptor ID	type	description
GGI5	topological charge index	topological charge index of order 5
IAC	information indices	information index of atomic composition
TIC1	information indices	total information content index
DPO3	randic molecular profile	randic molecular profiles nr.3
SPO2	randic molecular profile	shape profile nr.2
W3D	geometrical	3-D Wiener index
nCs	functional	number of total C (sp3)
nCrs	functional	number of ring C (sp3)
nROH	functional	number of hydroxyl groups
H4e	getaway descriptors	H autocorrelation of lag4/weighted by atomic Sanderson electronegativities
R4e+	getaway descriptors	R maximal autocorrelation of lag5/weighted by atomic Sanderson electronegativities
BID	walk and path	Balaban ID number
ATS6m	2D autocorrelations	Broto moreau autocorrelation of a topological structure lag6/weigthed by atomic masses
MW	constitutional	molecular weight

for the prediction of the chromosome-damaging potential of the chemical compounds. An ensemble is the average output of several different individual models, which were trained on different subsets of the entire training data (sometimes called Bootstrap Aggregating or Bagging, (*25*)). Building ensembles is a common way to improve classification and regression models in terms of stability and accuracy. We built heterogeneous ensembles consisting of several different model classes to achieve diverse ensembles (*26*). The model classes were as follows: (1) classification and regression trees (CART), where we used the implementation in the MATLAB Statistics Toolbox (The MathWorks, Natick, USA); (2) support vector machines (SVM) with Gaussian kernels (*28*); (3) linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), and linear ridge models (*29*); (4) feedforward neural networks (NN) with two hidden layers trained with a simple gradient descend (*30*); and (5) k-nearest-neighbor models (knn) with adaptive metrics (*30*).

The selection of the different model classes used for the construction of the final classification model was based on cross-validation (CV) approaches. This means that the training set (i.e., 90% of the total data) was split randomly into a training-learning set (80% of the data) and a training test set (20% of the data). Each of the different model classes was then trained on the training-learning set and assessed for their prediction accuracy on the training test set. This procedure was repeated 21 times using a novel randomly selected training-learning and training test set each time. In each of the runs, only the best model (i.e., the one showing the lowest classification error) was selected to become a member of the final ensemble. In this way, all model classes had to compete with each other because they are trained and tested on the same data set. Our approach, thus, resulted in a final classification model (the ML model) consisting of 21 individual models. The prediction output of this ML model is based on the counting of the vote of each of the individual 21 models and the determination of the majority vote, which then constitutes the final prediction.

Performance Evaluation of the ML Model. In a final step, the performance of the ML model was assessed on the entire training set (90% of the total data) and on the 10% of data (prediction set), which was initially removed and never included in model building. The procedure was independently repeated 20 times. This means that all model-building processes, that is, the random removal of 10% of the data, the construction of a classification model ensemble on the remaining 90% of the data as outlined above (always using the same 14 dragon descriptors determined in the feature selection step), and the prediction of both training and prediction sets were performed each time. For a final output, the mean average prediction values were calculated.

For the analysis of misclassified compounds, 50 independent model-building rounds were performed, and the number of incorrect classifications of each compound was recorded irrespective of its presence in the training or test sets.

Results and Discussion

On the basis of a data collection of high-quality CA-test results of more than 650 pharmaceuticals and industrial chemicals, we investigated the usefulness of two different computational approaches to predict the chromosome-damaging potential of compounds. We used a functionality of the commercially available MCASE system to automatically generate predictive models from a training set of compounds with associated qualitative (negative/positive) CA-test results. The predictivity of an in-house prediction model built in MCASE and an in-house machine-learning model in their ability to qualitatively predict the outcome of the CA-test was assessed.

MCASE Prediction Model. The performance characteristics for the MCASE prediction model are listed in Table 2. Both the training set and prediction set were predicted with comparable performances. Sensitivity values for the training set and prediction set were 53% and 57%, respectively. Clearly, higher values (i.e., 75% and 72%, respectively) were determined for the correct classification ratio of inactive compounds (specificity). Altogether, a concordance value of 65% was reached for both sets.

Interestingly, the performance characteristics obtained in our study were very similar to those reported by Rosenkranz et al. (*12*), although their data set was much smaller in size ($n = 233$ vs $n = 537$ in our study). The Danish-EPA reports on their website (http://www.mst.dk/) on the creation of an MCASE model based on approximately 500 chromosome aberration test data taken from the Ishidate data collection (*31*). Although overall higher performance values for this model were reported (76% concordance), similar unbalanced values for sensitivity (59%) and specificity (82%) were achieved. The persistently low sensitivity of the MCASE models indicates that the underlying 2D-fragment-based descriptors do not sufficiently describe the mechanism(s) leading to a positive result in the chromosome-aberration test. One way to further assess this possibility is the analysis of identified structural fragments that are statistically correlated with activity (biophores).

A list of the most significant biophores identified in our MCASE model is given in Table 3. As can be seen from the respective structural representation, almost all identified biophores represent known structural alerts for DNA reactivity. This implies that the structural determinants that on the basis of our MCASE analysis contribute to a positive effect in the chromosome aberration test reflect a direct drug−DNA (i.e., electrophilicity) interaction and, thus, are identical to the structural fragments identified from Ames test data (*32*).

The low sensitivity of the MCASE prediction model clearly limits its application as a decision tool during lead characterization phases. Companies developing new compounds are primarily dependent on prediction tools that have a relatively low false negative prediction rate (i.e., high sensitivity) in order to focus further development on those compounds that are presum-

Table 5. Performance Characteristics for the Machine-Learning Model

	TP[a]	FN	TN	FP	sensitivity	specificity	concordance
training set[b]	190 ± 10	65 ± 7	215 ± 10	63 ± 7	75.1%	76.8%	76.0%
prediction set[b]	46 ± 5	20 ± 5	50 ± 5	18 ± 5	70.8%	71.4%	71.6%

[a] TP, true positive; FN, false negative; TN, true negative; FP, false positive. [b] Values represent mean ± SD of 20 independent validations.

ably safe. However, false positive predictions could result in the loss of valuable candidates. Therefore, a balanced performance between sensitivity and specificity is desirable, resulting in ideal predictive tools that show equally high values for sensitivity, specificity, and concordance. This, however, is clearly not the case for the MCASE model, where the acceptable concordance value is primarily based on the low false positive rate. In other words, the particular descriptor applied in MCASE seems to be limited to pick up only one mechanism of CA induction, that is, the direct interaction of a drug with DNA. In order to overcome this apparent limitation, we investigated whether the use of more complex molecular descriptors in combination with a machine-learning approach might enable us to generate more predictive classification models.

Machine Learning Model. Statistical learning methods, such as support vector machines (SVM) or k-nearest-neighbor (knn) approaches are currently being used as a new approach in *in silico* toxicity prediction (*33*, *34*). Compared to traditional QSAR modeling approaches, statistical learning methods are often superior in terms of performance (*35*). As outlined in detail in the Materials and Methods section, we used a novel approach by building a classification model based on a heterogeneous ensemble of SVM, knn, neuronal networks, and other model classes.

A list of the 14 molecular descriptors selected for model building purposes are given in Table 4. As outlined in detail in the Materials and Methods section, these 14 descriptors were selected from more than 1600 dragonX descriptors after eliminating those that are redundant and choosing those that had the highest impact for classification. Several of the identified descriptors can be directly related to genotoxicity and, thus, present a mechanistically sound basis of the molecular features. Several functional descriptors as well as (electro)topological indices specify characteristics of structures involved in DNA modifications. Generic descriptors, such as geometrical and general information indices, describe the shape, size, and composition of molecules. A recent study on the prediction of genotoxicity by using statistical methods, such as SVMs and knn, indicates that such generic descriptors can be valuable for describing the DNA-reactive property of compounds (*33*). Molecular weight was selected as a discriminating feature probably because of the heterogeneous data base, consisting of many small organic chemicals that are chromosome-damaging (*10*) and an equally large amount of pharmaceutical compounds that are mostly not chromosome-damaging (*16*).

The performance values of our machine-learning model for the training set and prediction set are given in Table 5. As outlined in Materials and Methods, 20 independent cross validations were performed by removing each time 10% of the data (prediction set), building the ML model using the remaining 90% of the data (training set), and then predicting the removed compounds. The values for true positive, false negative, true negative, and false positive predictions of both training and prediction sets as well as for the other performance characteristics outlined in Table 5, thus, represent the mean ± standard deviation of 20 independent evaluations. Compared to the data obtained with the MCASE model, the ML approach led to a clearly improved prediction of CA-positive compounds (53% vs 75% for the training set), resulting in a balanced prediction model with almost equal performance values for sensitivity, specificity, and concordance.

Table 6. Performance Comparison between the Present ML model and the knn and SVM models published by Serra et al. (*13*)

	sensitivity	specificity
knn[a]	72.7%	92.3%
SVM[a]	72.7%	88.5%
ML model[b]	**90.5%**	**92.7%**

[a] Values taken from Serra et al. (*13*). [b] Only part of the Serra dataset was used for the analysis. For further details, see the Results and Discussion section.

Although this improvement reflects the usefulness of applying various molecular features as discriminators for the prediction of chromosome-damaging potential, a comparison with the values reported by Serra et al. (*13*) might lead to the conclusion that our ML model has a lower performance. However, a direct comparison of performance values between this study and Serra et al. (*13*) is difficult because of the differences in the data set and statistical evaluations. As mentioned before, Serra et al. used a smaller (and structurally less diverse) prediction set in which the proportion of known chromosome-damaging compounds was lower than that in our study (11 out of 37 compounds vs 70 out of 145 in our study). Thus, it remains open as to whether similarly good performance values would be achieved if a more extensive prediction set containing more CA-test positive compounds had been used. Second, the model characteristics described by Serra et al. seem to be based on a single cross-validation effort only, whereas we used a 20-fold CV to perform our validation procedure.

Despite these differences in model construction, we attempted to get a more objective comparison of the predictive value of our ML model by applying it to the compound data used by Serra et al. (*13*). In order to not be biased by our training set, only those compounds that were not included in our training set were extracted and, thus, represent novel compounds. Altogether, 291 compounds fulfilled this criterion, out of which 74 were reported with a positive result. These compounds were then collected as sd files, computed with our set of 14 molecular descriptors. and classified using our ML model. The resulting performance characteristics are given in Table 6 in comparison to the values reported by Serra et al. (*13*). Because the selected compounds were not previously included in our ML model, they can be seen as an independent prediction set, which we compared to our data. Overall, this comparison shows that our ML model reaches comparable prediction accuracies to those of the learning models reposted by Serra et al., although the latter were trained on a structurally less diverse set of compounds. Nevertheless, the sensitivity of our ML model clearly outscores the performance characteristics of the knn and SVM models.

Although tentative in nature, several conclusions can be drawn from this comparison. First, it is reasonable to assume that the lower prediction accuracies observed with our test set data compared to that of Serra et al. (*13*) is a consequence of the extension of the chemical space in our training set by adding a significant amount of pharmaceutical compounds to the less drug-like compounds contained in the Kirkland data set (*10*).

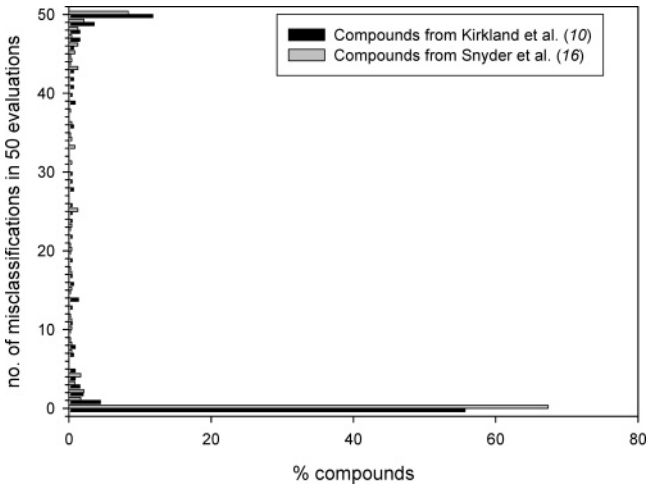

Figure 1. Percentage of compounds from both data sources (Kirkland et al., (*10*); Snyder et al. (*16*)) plotted against the number of incorrect predictions (misclassification) in a series of 50 independent evaluations. A compound that was correctly predicted in all of the runs, thus, falls into the group of zero misclassifications, whereas a consistently incorrect predicted compound is classified into the group of 50 misclassifications.

Because the majority of CA-test positive compounds in our study originates from the Kirkland data compilation, which from a chemical diversity point of view resembles the Serra data, it is not surprising that our ML model performs particularly well in terms of sensitivity on the latter data set. The development of prediction models for diverse data sets, such as those in our study, is generally considered to be problematic (*36*), and in theory, the construction of two local models (i.e., one for each data set) would have been favorable. Such an approach, however, is currently not feasible, because only few CA-test positive data for drug-like compounds are publicly available, and sufficiently large training sets for CA-test modeling, therefore, need to be compiled from structurally diverse compounds, as has been done in our study.

Given the structural diversity of our training set used for model construction, we investigated whether our ML model performed differently on the two underlying data sets. As a measure for predicting accuracy, we determined the number of misclassifications for each compound of both data sources in 50 independent evaluations. This means that an ML model was generated 50 times, and in each run, the classification result (i.e., true or false) was recorded for each compound. A compound that was correctly predicted in all of the 50 runs would, thus, be categorized with zero misclassifications, whereas a compound showing 50 misclassifications would have always been predicted incorrectly. The results of this exercise are shown in Figure 1. As can be seen, almost 70% of all compounds from the pharmaceutical class (*16*) were correctly predicted in 50 out of 50 evaluations (i.e., zero misclassifications). In comparison, the same was true for less than 60% of the less drug-like class (*10*). However, approximately 10% of pharmaceuticals were never predicted correctly (50 misclassifications), which to a slightly higher degree was also true for the Kirkland compounds. Altogether, it can be stated that compounds from the pharmaceutical class were predicted with higher accuracies than those from the less drug-like class.

Of the 20 compounds from the pharmaceutical class that were consistently misclassified in all 50 evaluations, 15 are false negatives, that is, a chromosome-damaging potential was missed. These 15 false negatives are listed in Table 7. Compounds were incorrectly classified as chromosome-damaging (false positives)

Table 7. List of 15 False Negative Classified Pharmaceuticals

compd	CAS	other genotoxicity information[a,b]
Chloroquine	54-05-7	Ames positive
Citalopram	59729-33-8	Ames positive
Clofibrate	637-07-0	Ames negative
Deoxycycline	7261-97-4	Ames negative BM negative
Fosinopril	98048-97-6	Ames negative BM negative
Fosphenytoin	93390-81-9	Ames negative BM negative
Imatinib	152459-95-5	Ames negative BM negative
Imipramine	50-49-7	N/A
Letrozole	112809-51-5	Ames negative BM negative
Oxacarbazepine	28721-07-5	Ames positive
Pentoxyphyline	6493-05-6	Ames negative BM negative
Rivastigmine	123441-03-2	Ames negative BM negative
Temozolomide	85622-93-1	Ames positive
Tiagabine	115103-54-3	Ames negative BM negative
Ziprasidone	146939-27-7	Ames positive

[a] Genotoxicity information taken from Snyder et al. (*10*). [b] Ames, Ames test; BM, mouse bone marrow micronucleus test; N/A: not available.

in only five cases (not listed). Mechanistic information on a possible mode of action of chromosome-damage induction of the 15 known genotoxic compounds is limited. Most of the compounds do not contain structural alerts for mutagenicity, suggesting that they do not primarily act genotoxic through direct drug−DNA interaction. A review of other mutagenicity test results obtained for the false-negative-predicted compounds shows 5 out of 14 compounds (no mutagenicity data were available for imipramine) were also tested positive in an Ames test, suggesting a genotoxic potential that was missed by our ML model. Surprisingly, 9 out of the 14 compounds were tested positive uniquely in the CA-test, whereas they yielded negative results in the Ames-test and the in vivo mouse micronucleus test (Table 7). This suggests that the positive CA-test result of these misclassified compounds might not be due to an inherent genotoxic potential but instead induced by biologically non-significant effects detected by this test system.

As outlined before, nonphysiological stimuli during cell culture can lead to structural chromosome aberrations (*9*). It is likely that other yet unknown mechanisms that are not directly related to the chemical structure can result in a (biologically not significant) positive result in the CA-test. Because these artificial effects are not directly related to the chemical structure of the compound, they are not amenable to modeling and, therefore, automatically decrease the predictivity of computational models applied to such data.

In conclusion, our data show that the chromosome-damaging potential of pharmaceuticals can be predicted using machine-learning approaches, albeit with lower predictivity than that previously reported for industrial chemicals (*13*). Nevertheless, the inclusion of a significant amount of pharmaceutical compounds into our model and the concomitant expansion of the chemical space covered by the model now makes it a potentially useful tool that can be incorporated in compound selection processes during early phases of drug development. A balanced prediction accuracy of 70−75% is sufficiently high during these developmental phases to filter out potential genotoxic compounds. Together with an experimental screening test (e.g., the *in vitro* micronucleus test) for the follow-up testing of compounds with a negative call, such a tool can significantly

contribute to a more targeted development of non-genotoxic drug candidates. In addition, given the high concordance between the *in vitro* micronucleus test and the CA-test, data obtained during the experimental screening of drug compounds could be fed back in order to train improved models solely based on drug-like compounds.

References

(1) *ICH 2SB*: Genotoxicity: a standard battery for genotoxicity testing for pharmaceuticals. CPMP/ICH/174/95.

(2) Miller, B., Potter-Locher, F., Seelbach, A., Stopper, H., Utesch, D., and Madle, S. (1998) Evaluation of the in vitro micronucleus test as an alternative to the in vitro chromosomal aberration assay: position of the GUM working group on the in vitro micronucleus test. *Mutat. Res. 410*, 81−116.

(3) Diehl, M. S., Willaby, S. L., and Snyder, R. D. (2000) Comparison of the results of a modified miniscreen and the standard bacterial reverse mutation assays. *Environ. Mol. Mutagen. 35*, 72−77.

(4) White, A. C., Mueller, R. A., Gallavan, R. H., Aaron, S., and Wilson, A. G. (2003) A multiple in silico program approach for the prediction of mutagenicity from chemical structure. *Mutat. Res. 539*, 77−89.

(5) Simon-Hettich, B., Rothfuss, A., and Steger-Hartmann, T. (2006) Use of computer-assisted prediction of toxic effects of chemical substances. *Toxicology 224*, 156−62.

(6) Obe, G., Pfeiffer, P., Savage, J. R., Johannes, C., Goedecke, W., Jeppesen, P., Natarajan, A. T., Martinez-Lopez, W., Folle, G. A., and Drets, M. E. (2002) Chromosomal aberrations: formation, identification and distribution. *Mutat. Res. 504* 17−36.

(7) Degrassi, F., Fiore, M., and Palitti, F. (2004) Chromosomal aberrations and genomic instability induced by topoisomerase-targeted antitumour drugs. *Curr. Med. Chem.: Anti-Cancer Agents 4*, 317−25.

(8) Parry, E. M., Parry, J. M., Corso, C., Doherty, A., Haddad, F., Hermine, T. F., Johnson, G., Kayani, M., Quick, E., Warr, T., and Williamson, J. (2002) Detection and characterization of mechanisms of action of aneugenic chemicals. *Mutagenesis 17*, 509−21.

(9) Kirkland, D., and Müller, L. (2000) Interpretation of the biological relevance of genotoxicity test results: the importance of thresholds. *Mutat. Res. 464*, 137−147.

(10) Kirkland, D., Aardema, M., Henderson, L., and Müller, L. (2005) Evaluation of the ability of a battery of three in vitro genotoxicity tests to discriminate rodent carcinogens and non-carcinogens. I. Sensitivity, specificity and relative predictivity. *Mutat. Res. 584*, 1−256.

(11) Rosenkranz, H. S., Ennever, F. K., Dimayuga, M., and Klopman, G. (1990) Significant differences in the structural basis of the induction of sister chromatid exchanges and chromosomal aberrations in Chinese hamster ovary cells. *Environ. Mol. Mutagen. 16*, 149−177.

(12) Rosenkranz, H. S. (2004) SAR modelling of genotoxic phenomena: the consequence on predictive performance of deviation from a unity ratio of genotoxicants/non-genotoxicants. *Mutat. Res. 559*, 67−71.

(13) Serra, J. R., Thompson, E. D., and Jurs, P. C. (2003) Development of binary classification of structural chromosome aberrations for a diverse set of organic compounds from molecular structure. *Chem. Res. Toxicol. 16*, 153−163.

(14) Cariello, N. F., Wilson, J. D., Britt, B. H., Wedd, D. J., Burlinson, B., and Gombar, V. (2002) Comparison of computer programs DEREK and MCASE to predict bacterial mutagenicity. *Mutagenesis 17*, 321−329.

(15) Greene, N. (2002) Computer systems for the prediction of toxicity: an update. *Adv. Drug Delivery Rev. 54*, 417−431.

(16) Snyder, R. D., Pearl, G. S., Mandakas, G., Choy, W. N., Goodsaid, F., and Rosenblum, I. Y. (2004) Assessment of the sensitivity of the computational programs DEREK, TOPKAT and MCASE in the prediction of the genotoxicity of pharmaceutical molecules. *Environ. Mol. Mutagen. 43*, 143−158.

(17) Snyder, R. D., and Green, J. W. (2001) A review of the genotoxicity of marketed pharmaceuticals. *Mutat. Res. 488*, 151−169.

(18) Sadowski, J., and Kubinyi, H. (1998) A scoring scheme for discriminating between drugs and nondrugs. *J. Med. Chem. 41*, 3325−3329.

(19) Klopman, G., and Rosenkranz, H. S. (1994) International Commission for Protection Against Environmental Mutagens and Carcinogens. Approaches to SAR in carcinogenesis and mutagenesis. Prediction of carcinogenicity/mutagenicity using MULTI-CASE. *Mutat. Res. 305*, 33−46.

(20) Rosenkranz, H. S., Cunningham, A. R., Zhang, Y. P., Claycamp, H. G., Macina, O.T., Sussmanm, N. B., Grant, G. S., and Klopman, G. (1999) Development, characterization and application of predictive toxicology models. *SAR QSAR Environ. Res. 10*, 277−298.

(21) http://www.talete.mi.it/dragon_exp.htm

(22) Todeschini, R., and Consonni V. (2000) Handbook of Molecular Descriptors. In *Series of Methods and Principles in Medicinal Chemistry*, (Mannhold, R., Kubinyi, H., and Timmerman, H., Eds.) Vol. 11, Wiley-VCH, Weinheim, Germany.

(23) Breiman, L. (2001) Random forests. *Machine Learning 45*, 5−32.

(24) Breiman, L. (1998) Arcing classifiers. *Annals of Statistics 26*, 801−849.

(25) Breiman, L. (1996) Bagging predictors. *Machine Learning 24*, 123−140.

(26) Wichard, J., and Ogorzalek, M. (2006) Time series prediction with ensemble models applied to the cats benchmark. *Neurocomputing*, in press.

(27) Breiman, L. (1993) *Classification and Regression Trees.* Chapman & Hall, Boca Raton, FL.

(28) Chang, C., and Lin, C. (2001) Libsvm - A library for support vector machines. http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

(29) Hastie, T., Tibshirani, R., and Friedman, T. (2001) The Elements of Statistical Learning. In *Springer Series in Statistics* (Bickel, P., Diggle, P., Fienberg, S., Gather, U., Olkin, I., and Zeger, S., Eds.) Springer-Verlag, Heidelberg, Germany.

(30) Merkwirth, C., and Wichard, J. (2002) ENTOOL - A MATLAB toolbox for ensemble modelling, http://chopin.zet.agh.edu.pl/∼wichtel/.

(31) Sofuni, T. Ed. (1998) *Data Book of Chromosomal Aberration Test in Vitro.* Life Science Information Center, Japan.

(32) Ashby, J., and Styles, J. A. (1978) Does carcinogenic potency correlate with mutagenic potency in the Ames assay? *Nature 271*, 452−455.

(33) Li, H., Ung, C. Y., Yap, C. W., Xue, Y., Li, Z. R., Cao, Z. W., and Chen, Y. Z. (2005) Prediction of genotoxicity of chemical compounds by statistical learning methods. *Chem. Res. Toxicol. 18*, 1071−1080.

(34) Zhao, C. Y., Zhank, H. X., Zhang, X. Y., Liu, M. C., Hu, Z. D., and Fan, B. T. (2006) Application of support vector machine (SVM) for prediction toxic activity of different data sets. *Toxicology 217*, 105−119.

(35) He, L., Jurs, P. C., Custer, L., Durham, S. K., and Pearl, G. M. (2003) Predicting the genotoxicity of aromatic compounds from molecular structure with different classifiers. *Chem. Res. Toxicol. 16*, 1576−1580.

(36) Richard, A. M., and Benigni, R. (2001) AI and SAR approaches for predicting chemical carcinogenicity: survey and status report. *SAR QSAR Environ. Res. 13*, 1−19.

TX060136W