J Comput Aided Mol Des (2011) 25:455–467
DOI 10.1007/s10822-011-9431-3

Effect of training data size and noise level on support vector machines virtual screening of genotoxic compounds from large compound libraries

Pankaj Kumar · Xiaohua Ma · Xianghui Liu · Jia Jia · Han Bucong · Ying Xue · Ze Rong Li · Sheng Yong Yang · Yu Quan Wei · Yu Zong Chen

Received: 21 May 2010 / Accepted: 17 April 2011 / Published online: 10 May 2011

Abstract Various in vitro and in-silico methods have been used for drug genotoxicity tests, which show limited genotoxicity (GT+) and non-genotoxicity (GT−) identification rates. New methods and combinatorial approaches have been explored for enhanced collective identification capability. The rates of in-silco methods may be further improved by significantly diversified training data enriched by the large number of recently reported GT+ and GT− compounds, but a major concern is the increased noise levels arising from high false-positive rates of in vitro data. In this work, we evaluated the effect of training data size and noise level on the performance of support vector machines (SVM) method known to tolerate high noise levels in training data. Two SVMs of different diversity/noise levels were developed and tested. H-SVM trained by higher diversity higher noise data (GT+ in any in vivo or in vitro test) outperforms L-SVM trained by lower noise lower diversity data (GT+ in in vivo or Ames test only). H-SVM trained by 4,763 GT+ compounds reported before 2008 and 8,232 GT− compounds excluding clinical trial drugs correctly identified 81.6% of the 38 GT+ compounds reported since 2008, predicted 83.1% of the 2,008 clinical trial drugs as GT−, and 23.96% of 168 K MDDR and 27.23% of 17.86M PubChem compounds as GT+. These are comparable to the 43.1–51.9% GT+ and 75–93% GT− rates of existing in-silico methods, 58.8% GT+ and 79% GT− rates of Ames method, and the estimated percentages of 23% in vivo and 31–33% in vitro GT+ compounds in the "universe of chemicals". There is a substantial level of agreement between H-SVM and L-SVM predicted GT+ and GT− MDDR compounds and the prediction from TOPKAT. SVM showed good potential in identifying GT+ compounds from large compound libraries based on higher diversity and higher noise training data.

Keywords Bioinformatics · Genotoxicity · Computer aided drug design · Drug safety · Genotoxicity · Machine learning · Statistical learning · Support vector machine

Our SVM genotoxicity virtual screening models can be accessed at http://bidd.nus.edu.sg/gtox/gtox.html.

Electronic supplementary material The online version of this article (doi:10.1007/s10822-011-9431-3) contains supplementary material, which is available to authorized users.

P. Kumar · X. Ma · X. Liu · J. Jia · H. Bucong · Y. Z. Chen (✉)
Bioinformatics and Drug Design Group, Centre for Computational Science and Engineering, Department of Pharmacy, National University of Singapore, Blk S16, Level 8, 3 Science Drive 2, Singapore 117546, Singapore
e-mail: phacyz@nus.edu.sg

Y. Xue · Z. R. Li
College of Chemistry, Sichuan University, Chengdu 610064, People's Republic of China

S. Y. Yang · Y. Q. Wei · Y. Z. Chen
State Key Laboratory of Biotherapy, Sichuan University, Chengdu 610064, People's Republic of China

Introduction

Genotoxicity has been rigorously tested in drug discovery processes and evaluated by regulatory agencies in drug approval processes [1]. The molecular mechanisms of genotoxicity include DNA intercalation by aromatic ring of a compound, DNA methylation, DNA adduct formation and strand break, and unscheduled DNA synthesis [2]. Some of these actions, such as DNA strand break, may be triggered by chemicals that facilitate certain enzyme–DNA interactions such as stabilization of DNA topoisomerase

II—DNA cleavage complex [3]. Some genotoxic compounds require metabolic activation and their genotoxic effect are mediated via N-dialkylation [4]. DNA repair inhibition may also be a predominant mechanism of some genotoxic compounds [5].

Both in vivo and in vitro methods have been extensively used for genotoxicity tests [6, 7], and some in silico methods have been used in combination with in vitro analysis to aid in the early detection of genotoxicity [8]. Significant interest to in vitro and in silico methods arises because in vivo methods are costly, time-consuming, labor intensive, dependent on animal resources of public concerns, and in some cases show substantial false-positive and false-negative rates [6, 9]. The high false hit rates of in vitro tests and the inconsistency between in vitro and in vivo tests coupled with the high cost of in vivo tests make it more desirable to develop high performance in silico tools (by utilizing all available genotoxicity assay data and derived knowledge) to enable the screening of large compound libraries for gentoxicity tests at sufficiently good accuracy levels to complement genotoxicity assays.

Progress has also been made in developing in silico methods of good performance [8], which include DEREK, MCASE, TOPKAT, QSPR, DNA docking, kinase inhibitor profiling, and machine learning models such as decision trees, linear discriminant analysis, k-nearest neighbor classification, neural networks, and support vector machines (SVM) [10]. Nonetheless, testing results of some of these in-silico methods show limited sensitivity (true positive rate i.e. number of correctly predicted true genotoxic positive compounds divided by the number of genotoxic positive compounds) of 43.1–51.9% for Ames and 21.3–31.9% for in vivo cytogenetics positive compounds, but good specificity (the true negative rate i.e. number of correctly identified true non-genotoxic compounds divided by the number of non-genotoxic compounds) of 75–93% [4, 8]. Specificity may be further improved by using consensus models of multiple in-silico methods at the expense of decreased chemical coverage [8, 11]. The performance of individual in-silico methods may also be improved by incorporating the profiles of the large number of recently reported genotoxic (GT+) and non-genotoxic (GT−) compounds. Our literature search identified 4,763 compounds tested GT+ in at least one in vivo or in vitro test and 2,113 compounds tested GT− in Ames or rodent non-carcinogenic tests, which substantially diversified the training set of ~200 GT+ and ~600 GT− compounds used in earlier works [10].

However, incorporation of these compounds is expected to introduce significantly higher level of noise into the training data because of the high-false-positive rates in in vitro data [12, 13]. In vitro methods such as Ames tend to show limited sensitivity of 58.8–78.7% and specificity of 30.8–73.9% when used individually, and improved sensitivity of 75.3–92.3% but reduced specificity of 5.0–34.6% when used in combinations, leading to high false-positive rates of 26.1–69.2% for individual and 65.4–95.0% for combination tests [12]. Although factors contributing to and strategies for overcoming false-positives in in vitro data have been discussed [13], and progress has been made in improving the sensitivity and specificity of in vitro methods [14] and in high-throughput [15] tests, high volume in vitro data of sufficiently low false-positive rate is not yet available. Therefore, it is desirable to explore in-silico methods that tolerate high noise training data for improved genotoxicity identification capability.

SVM has been found to tolerate training data false positives and false negatives up to the levels when the ratio of the intentionally mislabeled positives to true positives is 5:1 and that of the intentionally mislabeled negatives to true negatives is 1:1 [16]. A question remains as to what extent the performance gain from data diversity outweighs the performance loss from data noise. In this work, we evaluated the effect of training data size and noise level on SVM identification of GT+ compounds from large compound libraries. Two SVM models of different diversity/noise levels were developed based on parameters determined by five-fold cross validation test. Assays for detecting genotoxicity are not perfect and constitute specific advantages and disadvantages. In vivo tests are mandatory for follow-up studies of in vitro GT+ compounds [6, 17, 18], and are thus expected to be significantly higher in accuracy than in vitro methods. Among various in vitro tests Ames tests has shown the best performance [12]. A dataset of lower percentage of false positives and thus lower noise levels can be constructed from the data derived from these two more accurate assay types. On the other hand, an expanded dataset can be generated that include the data derived both from the two more accurate assay types and from additional in vitro assays of higher false positive rates. This dataset is of higher diversity because of more diverse compounds included, and is of high noise levels because of the higher number of false positive compounds included.

One SVM model is H-SVM, a higher diversity higher noise SVM trained by GT+ compounds tested positive in any in vivo or in vitro test (which covers 1,191 chemical families and has a diversity index of 0.3498). Another is L-SVM, a lower noise but lower diversity SVM model trained by GT+ compounds tested positive in in vivo or Ames test only (which covers 899 chemical families and has a diversity index of 0.3600). The sensitivity of H-SVM and L-SVM were evaluated by GT+ compounds reported in journals published in year 2008 and 2009 to be positive in in vivo or Ames in vitro tests only and not included in the training sets. These were compared to those of in vitro methods [7] and other in-silico methods [4, 8]. The

capability of H-SVM and L-SVM in searching large compound libraries of interest was tested by screening 168 K MDDR (MDL Drug Data Report) and 17.86M PubChem (http://pubchem.ncbi.nlm.nih.gov/) compounds to determine whether the percentages of the SVM identified GT+ compounds are consistent with the estimated percentages of in vivo and in vitro Ames positive compounds in the "universe of chemicals" [9].

Methods

Collection of GT+ and GT− compounds and dataset construction

We collected from literatures and databases 4,801 GT+ compounds tested positive in at least one in vivo or in vitro genotoxicity assay, and 2,113 GT− compounds tested negative in Ames or non-carcinogenic in rodent (summarized in Table 1). The sensitivity and specificity in identifying these compounds are 58.8 and 73.9% for Ames test, 78.7 and 30.8% for micronucleus test (MN), 73.1 and 39.0% for mouse lymphoma assay (MLA), 65.6 and 44.9% for chromosomal aberration test (CA) [7]. In vivo tests have been used for follow-up studies of compounds positive in in vitro genotoxicity tests [6, 17, 18]. We have not found publically reported sensitivity and specificity of in vivo tests. Nonetheless as the go-to methods for follow-up studies of in vitro GT+ compounds, the accuracies of in vivo methods are expected to be significantly higher than those of in vitro methods. The collected GT− compounds are insufficient for covering GT− chemical space. Thus, putative GT− compounds were selected from two chemical groups likely containing high percentages of GT− compounds. The first group contains 1,293 FDA approved and 2,008 clinical trial drugs from TTD (http://bidd/group/ttd/ttd.asp). A test of 300 drugs has shown that 12% of these drugs are positive in in vivo cytogenetics tests [19]. The estimated percentage of GT+ compounds in FDA approved drugs is thus ~12% and that of clinical trial drugs is not significantly higher than 12%. The second group includes 2,328 food additives from FDA EAFUS((Everything Added to Food in the US) database (http://www.fda.gov/Food/FoodIngredientsPackaging/ucm115326.htm) and 177 GRAS (Generally Regarded As Safe) compounds from FDA GRAS Substances Page (http://www.fda.gov/Food/FoodIngredientsPackaging/GenerallyRecognizedasSafeGRAS/GRASSubstancesSCOGSDatabase/ucm084104.htm). No literature report has been found to give a reasonable estimate of the percentage of GT+ compounds in these chemicals. A survey of literature-reported safety evaluation of food additives and GRAS compounds suggests that most (89%) of the evaluated compound groups show no significant genotoxic activities at conditions relative to humans. Overall, the ratio of falsely and correctly collected GT+ and GT− compounds are expected to be lower than the levels of 5:1 and 1:1 tolerated by SVM [16].

These compounds were further separated into training and testing sets by two different ways depending on the testing tasks. For five-fold cross validation test of H-SVM, 4,763 GT+ compounds positive in at least one in vivo or in vitro test and 8,232 GT− compounds were randomly divided into five GT+ and GT− subsets of approximately equal size. For five-fold cross validation test of L-SVM, 3,321 GT+ compounds positive in in vivo or Ames only and 8,232 GT− compounds were randomly divided into five GT+ and GT− subsets of approximately equal size. Four GT+ and GT− subsets were selected as the training set and the fifth GT+ and GT− subset as the testing set. This process was repeated five times such that every subset is selected as a testing set once. For the test of screening large compound libraries, 4,763 GT+ agents (positive in at least one in vivo or in vitro test and reported before year 2008) were included in the training sets of H-SVM and 3,321 GT+ agents (positive in in vivo or Ames only and reported before year 2008) were included in the training set of L-SVM. A total of 8,232 GT− compounds excluding clinical trial drugs were included in the training sets of H-SVM and L-SVM. A total of 38 GT+ compounds (in vivo or Ames positive only) reported since 2008 and 2,008 clinical trial drugs were used as an independent testing set for both H-SVM and L-SVM.

Molecular descriptors and molecular fingerprints

Molecular descriptors are quantitative representations of structural and physicochemical features of molecules, which have been extensively used in deriving SAR, QSAR, virtual screening (VS) [20], acute toxicity and genotoxicity prediction [10]. A total of 522 1D and 2D descriptors derived by using our software MODEL (http://jing.cz3.nus.edu.sg/cgi-bin/model/model.cgi) were used in this work for developing H-SVM and L-SVM models. These descriptors and the relevant references are given in Supplementary Table S1, which include 58 descriptors in the class of simple constitutional properties, 14 descriptors in the class of electrochemical properties, 41 descriptors in the class of molecular connectivity and shape, 42 descriptors in the class of electrotopological state. In computing the descriptors of these compounds, their 3D structures were generated by using the Concord software [21], and optimized by using the semi-empirical AM1 method. In evaluating the similarity between compounds, we used molecular fingerprints to represent each compound. A total of 881 fingerprints including 166 structural fragments in MACCS keys were generated by using our software Molfeat (http://jing.cz3.nus.edu.sg/cgi-bin/molfeat/molfeat.cgi).

Table 1 Statistics of the collected genotoxic and non-genotoxic compounds, distribution of these compounds in the training and testing datasets, and the corresponding sources, experimental methods, and references

Compound class	Test platform or data source	Number of compounds	References source or pubmed ID
Genotoxic compounds in training datasets			
In vivo positive	TOXNET (GENE-TOX)	442	NLM leased genotox data http://www.nlm.nih.gov/databases/leased.html
Ames positive	CPDB (Carcinogenic potency database)	394	Pubmed ID: 19042710
	Rodent carcinogenic and Ames positive	311	Pubmed ID: 18585956
	Ames positive	2,401	Pubmed ID: 15634026
		42	Pubmed ID: 17906314
Positive in at least one other in vitro test		1,547	NLM leased genotox data http://www.nlm.nih.gov/databases/leased.html
		122	Pubmed ID: 18585956
Training data for H-SVM	Positive in at least one in vivo or in vitro	4,763	442 + 394 + 311 + 2,401 + 42 + 1,547 + 122 = 5,259 => (remove duplicates) => 4,763
Training data for L-SVM	Positive in in vivo or Ames only	3,321	442 + 394 + 311 + 2,401 + 42 = 3,590 => (remove duplicates) => 3,321
Genotoxic compounds in test dataset			
In vivo	In vivo enotoxin and positive in in vitro mammalian cell genotoxicity tests	19	Mutagenesis vol. 24 no. 3 pp. 279–284, 2009 Mutation research 650 (2008) 104–114 Mutation research 675 (2009) 51–59
Ames	Ames positive	19	Mutation research 653 (2008) 99–108
Testing data for H-SVM and L-SVM	In vivo or Ames positive	38	19 + 19 = 38
Non-genotoxic compounds in training dataset			
In vivo negative	Rodent non-carcinogenic	177	Pubmed ID: 18585956
In vitro Ames negative	Ames	1,936	Pubmed ID: 15634026
Food additives	Everything added to food (EAFUS)	2,328	www.fda.gov
GRAS	GRAS	177	www.fda.gov
FDA approved drugs	TTD database	1,293	Pubmed ID: 11752352
Drugs in PDR	Physician desk reference (PDR) drugs	545	Pubmed ID: 19334052
Food related	Food standards agency	1,128	http://www.food.gov.uk/
Food-contact substances	Cumulative estimated daily intakes (CEDIs) and acceptable daily intakes (ADIs)	802	http://www.fda.gov/Food/FoodIngredientsPackaging/FoodContactSubstancesFCS/CEDIADIDatabase/default.htm
Training data for H-SVM and L-SVM	In vivo or Ames negative, Food additives, GRAS compounds, FDA approved drugs	8,232	177 + 1,936 + 2,328 + 177 + 1,293 + 545 + 1,128 + 802 = 8,582 (remove duplicates = 8,232)
Non-genotoxic compounds in testing dataset			
Clinical trial drugs	TTD database	2,008	Pubmed ID: 11752352

Support vector machines method

The process of training and using SVM for screening compounds based on their molecular descriptors is schematically illustrated in Fig. 1. SVM is based on the structural risk minimization principle of statistical learning theory, which consistently shows outstanding classification performance, is less penalized by sample redundancy, has lower risk for over-fitting [22], and tolerate high noise levels in training data [16]. In linearly separable cases, SVM constructs a hyper-plane to separate GT+ and GT− classes of compounds with a maximum margin. A compound is represented by a vector $\boldsymbol{x}_i$ composed of its molecular descriptors. The hyper-plane is constructed by finding another vector $\mathbf{w}$ and a parameter b that minimizes $\|\mathbf{w}\|^2$ and satisfies the following conditions:

$$w \cdot x_i + b \geq +1, \quad \text{for } y_i = +1 \quad \text{Class 1(active)} \tag{1}$$

$$w \cdot x_i + b \leq -1, \quad \text{for } y_i = -1 \quad \text{Class 2(inactive)} \tag{2}$$

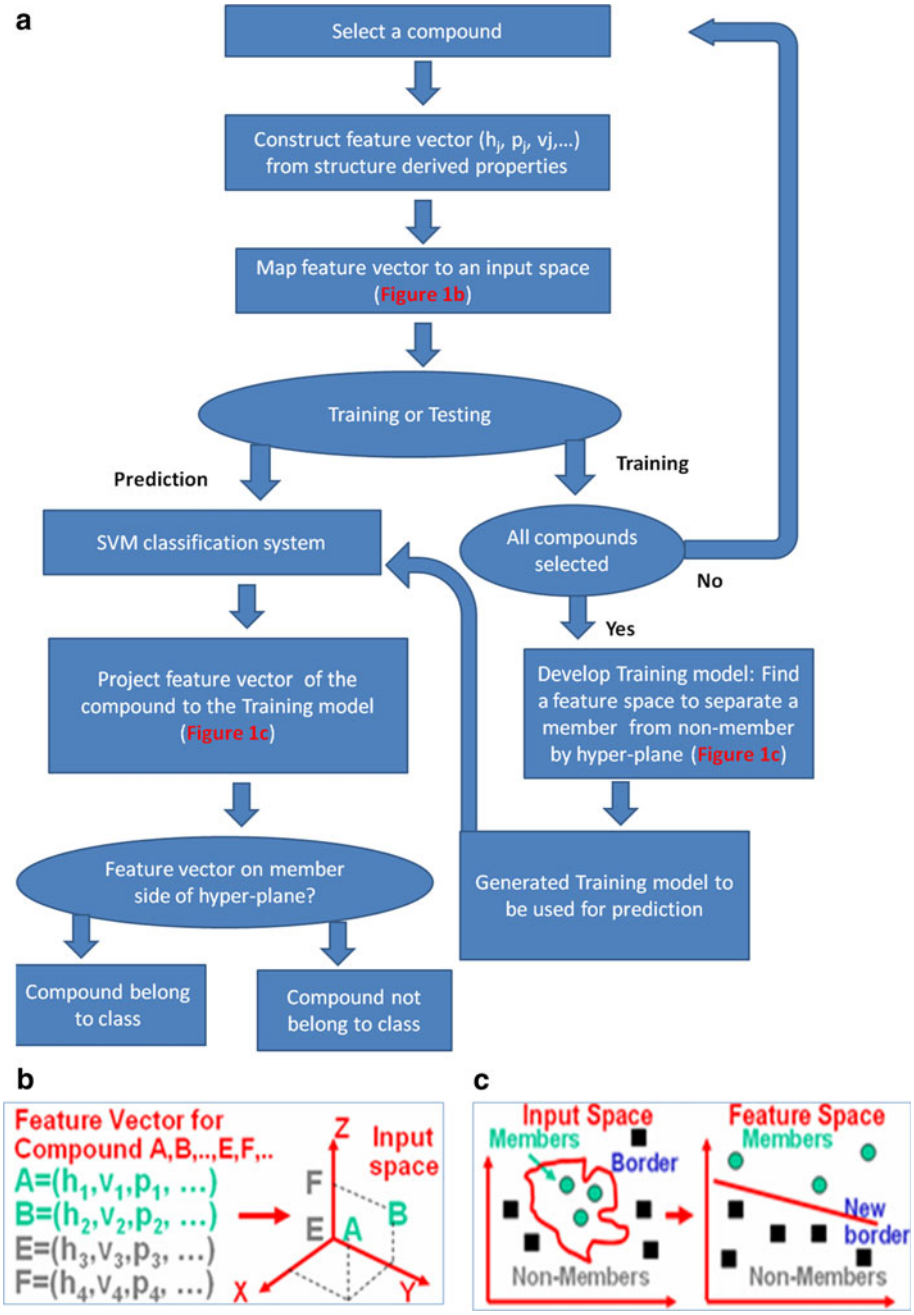

Fig. 1 Schematic diagram illustrating the process of the training a SVM prediction model and using it for predicting compounds of genotoxic class from their structurally-derived properties (molecular descriptors) by using support vector machines. *A*, *B*, *E*, *F* and (h_j, p_j, *vj*,…) represents such structural and physicochemical properties as hydrophobicity, volume, polarizability, etc

where y_i is the class index, $\mathbf{w}$ is a vector normal to the hyperplane, $|b|/\|\mathbf{w}\|$ is the perpendicular distance from the hyperplane to the origin and $\|\mathrm{w}\|^2$ is the Euclidean norm of $\mathbf{w}$. Base on $\mathbf{w}$ and b, a given vector $\boldsymbol{x}$ can be classified by $f(x) = \mathrm{sign}[(w \cdot x) + b]$. A positive or negative $f(\boldsymbol{x})$ value indicates that the vector $\mathbf{x}$ belongs to the GT+ or GT− class respectively.

In nonlinearly separable cases, which frequently occur in classifying compounds of diverse structures, SVM maps the input vectors into a higher dimensional feature space by using a kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$. We used Radial Basis Function (RBF) kernel $K(x_i, x_j) = e^{-\|x_j - x_i\|^2/2\sigma^2}$ which has been extensively used and consistently shown better performance than other kernel functions. In this RBF kernel, σ is the Parzen window which represents the kernel width and defines the scope of influence for a support vector over the data space. Higher σ value allows a support vector to have strong influence over a large area and thus will reduce the number of support vectors needed to make a separating boundary in the data space. Linear SVM can then applied to this feature space based on the following decision function: $f(\mathbf{x}) = sign(\sum_{i=1}^{l} \alpha_i^0 y_i K(\mathbf{x}, \mathbf{x}_i) + b)$, where the coefficients α_i^0 and b are determined by maximizing the following Langrangian expression: $\sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$ under the conditions $\alpha_i \geq 0$ and $\sum_{i=1}^{l} \alpha_i y_i = 0$. A positive or negative $f(\boldsymbol{x})$ value indicates that the vector $\boldsymbol{x}$ belongs to the GT +or GT− class respectively. The Parzen windown σ in RBF kernel determines the area of influence this support vector has over the data space. Higher σ value allows a support vector to have strong influence over a large area and thus will reduce the number of support vectors needed to make a separating boundary in the data space (Fig. 1).

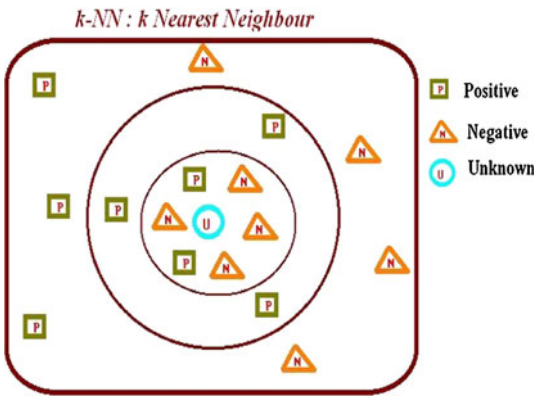

Fig. 2 k-Nearest Neighbour (kNN). The input labeled data (e.g. *positive* and *negative*) is classified by kNN by grouping close neighbors together which is then used to determine the class of unknown

In developing SVM, a hard margin $c = 100{,}000$ was used, and the σ value was found to be 5.6 based on five-fold cross validation test. As in the case of all discriminative methods, the performance of SVM is measured by the quantity of true positives (*TP*), true negatives (*TN*), false positives (*FP*), false negatives (*FN*), sensitivity $SEN = TP/(TP + FN) \times 100$ which is the prediction accuracy for the GT+ compounds in this work, and specificity $SP = TN/(TN + FP) \times 100$ which is the prediction accuracy for the GT− compounds in this work. The overall prediction accuracy (ACC) and Matthews correlation coefficient (*MCC*) are also used to measure the prediction accuracies [23] and can be given by:

$$SEN = TP/(TP + FN) \times 100 \tag{3}$$

$$SP = TN/(TN + FP) \times 100 \tag{4}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$MCC = \frac{TP \times TN - FN \times FP}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}} \tag{6}$$

Matthews' Correlation Coefficient returns a number between −1 and 1, with a value of 1 indicating a perfect prediction, 0 indicating a random prediction and values below 0 indicate a worse than random prediction.

k-nearest neighbour (KNN) method

KNN is a supervised machine learning method which classifies data by grouping close neighbors together. Based on the label of input training data points, the new test data is classified by the count of labeled of k nearest neighbored training data (Fig. 2). Ideally, the value of k should be decided on the number of labeled training data and is optimized during training. The algorithm implementing kNN can vary in number of ways e.g. on the basis distance calculation methods like Euclidian or Manhattan. Different K-nearest neighbor algorithm have been used for the classification of biological and chemical data [24–26]. In this work, k-NN algorithm of WEKA class IBk was used [27].

Feedforward backpropagtion neural network (FBNN) method

Neural networks are a type of supervised machine learning method and Feed forward Neural Network is one of its subtypes. FBNN has been applied in this work by WEKA class Multi Layer Perceptron [28]. A multilayer perceptron maps sets of input data onto a set of appropriate output. Multilayer perceptron is a modified standard linear

perceptron in that it uses three or more layers of neurons (nodes) with nonlinear activation functions, and is more powerful than the perceptron in that it can distinguish data that is not linearly separable. Multilayer perceptron design has three or more layers which are input, output, and one or more hidden layers. Nodes of one layer connects o every node in the following layer with certain weight. Learning occurs in the perceptron by changing connection weights after each piece of data is processed, based on the amount of error in the output compared to the expected result. Generalization of the least mean squares algorithm in the linear perceptron results in back propagation.

Tanimoto similarity searching method

Compounds similar to at least one GT+ agent in a dataset can be identified by using the Tanimoto coefficient *sim*(*i,j*) [29].

$$sim(i,j) = \frac{\sum_{d=1}^{l} x_{di}x_{dj}}{\sum_{d=1}^{l} (x_{di})^2 + \sum_{d=1}^{l} (x_{dj})^2 - \sum_{d=1}^{l} x_{di}x_{dj}} \quad (7)$$

where l is the number of molecular fingerprints. A compound i is considered to be similar to a known GT+ agent j in the GT+ dataset if the corresponding *sim*(*i,j*) value is greater than a cut-off value. In this work, the similarity search was conducted for MDDR compounds. Therefore, in computing *sim*(*i,j*), the molecular fingerprint vectors $\mathbf{x}_i$s were scaled with respect to all of the MDDR compounds. The cut-off values for similarity compounds are typically in the range of 0.8–0.9 [30]. A stricter cut-off value of 0.9 was used in this study.

Determination of structural diversity

Structural diversity of a collection of compounds can be measured by using the diversity index (DI) value, which is the mean value of the similarity between pairs of compounds in a dataset:

$$DI = \frac{\Sigma_{i=1}^{N} \Sigma_{j=1, i\neq j}^{N} sim(i,j)}{N(N-1)} \quad (8)$$

Where *sim(i,j)* is a measure of the similarity between compound i and j, and N is the number of compounds in the dataset. The structural diversity of a dataset increases with decreasing DI value. In this work, *sim(i,j)* is computed by using the Tanimoto coefficient (T):

$$T = \frac{C}{A + B - C} \quad (9)$$

where C is number of bits set to 1 in binary fingerprints of compound i and j, A is number of bits set to 1 in binary fingerprints of compound i, and B is number of bits set to 1 in binary fingerprints of compound j.

Results and discussion

Genotoxicity prediction Performance of SVM, kNN and FBNN evaluated by five-fold cross validation test.

Table 2 gives the five-fold cross validation test results of H-SVM and L-SVM in identifying GT+ and GT− compounds. The accuracies for predicting GT+ and GT− are 81.6–84.1% and 92.1–94.3% for H-SVM, and 79.2–81.9% and 90.9–92.5% for L-SVM respectively. The overall prediction accuracy Q and Matthews correlation coefficient C are 88.3–89.6% and 0.75–0.78 for H-SVM, and 87.6–88.7% and 0.49–0.52 for L-SVM respectively. Both H-SVM and L-SVM appears to show good prediction capability in identifying GT+ compounds at low false-positive rates, with H-SVM outperforming L-SVM. Similar prediction accuracies were also found from two additional five-fold cross validation studies conducted by using training–testing sets separately generated from different random number seed parameters.

The accuracies of our H-SVM and L-SVM models are comparable to or better than the reported 75.5% GT+ and 90.6% GT− rates of SVM [10], and 43.1–51.9% GT+ and 75–93% GT− rates of other existing in-silico methods [4, 8]. Caution needs to be raised about straightforward comparison of these results, which might be misleading because the outcome of genotoxicity prediction strongly depends on the datasets, molecular descriptors and machine learning methods used. To evaluate the effect of different machine learning methods on genotoxicity prediction performance, we developed and tested kNN [27] and FBBNN [28] genotixicity prediction models by using the same sets of molecular descriptors and the same training and testing datasets in the five-fold cross validation studies of the H-SVM and L-SVM. In correspondence with H-SVM and L-SVM, we developed H-kNN and L-kNN, and H-FBNN and L-FBNN for the high diversity high noise training dataset and low diversity low noise training dataset respectively. The five-fold cross validation results for the H-kNN and L-kNN are in Table 3 and those for the H-FBNN and L-FBNN are in Table 4 respectively. The GT+ and GT− accuracies of H-kNN are 72.3–74.6% and 85.1–87.8%, L-kNN are 65.5–67.5% and 83.2–84.1%, H-FBNN are 75.2–77.1% and 85.3–88.2%, and L-FBNN are 62.3–65.5% and 87.8–89.4%, respectively. Consistent with the results of H-SVM and L-SVM, H-kNN outperforms L-kNN and H-FBNN outperforms L-FBNN, suggesting that the positive contribution of data diversity somehow outweighs the negative effect of data noise in developing machine learning models for predicting

Table 2 The performance of support vector machines trained by higher diversity higher noise data (H-SVM) and by lower noise and lower diversity data (L-SVM) in prediction of genotoxic (GT+) and non-genotoxic (GT−) compounds based on five-fold cross validation tests

Method	Cross -validation	Genotoxicity			Non-genotoxicity			ACC (%)	G-Mean	MCC
		TP	FN	SEN (%)	TN	FP	SP (%)			
H-SVM	1	791	161	83.1	1,501	127	92.2	88.5	87.5	0.76
	2	777	175	81.6	1,509	119	92.7	88.3	87.0	0.75
	3	771	181	81.0	1,535	93	94.3	89.0	87.4	0.77
	4	795	157	83.5	1,499	129	92.1	88.6	87.7	0.76
	5	801	151	84.1	1,521	107	93.4	89.6	88.6	0.78
	Mean			82.7			92.9	88.8	87.7	0.77
	SD			1.32			0.92	0.55	1.1	0.01
	SE			0.59			0.41	0.25	0.5	0.01
L-SVM	1	535	129	80.6	1,500	146	91.1	88.1	85.7	0.51
	2	544	120	81.9	1,496	150	90.9	88.3	86.3	0.52
	3	526	138	79.2	1,524	122	92.6	88.7	85.6	0.52
	4	544	139	81.9	1,524	150	92.6	88.7	87.1	0.52
	5	525	120	79.1	1,496	122	90.9	87.6	84.8	0.49
	Mean			80.5			91.6	88.3	85.9	0.51
	SD			1.4			0.9	0.5	1.1	0.01
	SE			0.6			0.4	0.2	0.5	0.01

The performance for each fold is reported by true positive (*TP*), false negative (*FN*), accuracy of identifying GT+ or sensitivity (*SEN*), accuracy of identifying GT− or specificity (*SP*), overall accuracy (*ACC*), geometric mean of sensitivity and specificity (*G-Mean*), Matthews correlation coefficient (*MCC*), and their mean, standard deviation (*SD*), standard error (*SE*)

Table 3 The performance of *k*-Nearest Neighbour trained by higher diversity higher noise data (H-SVM) and by lower noise and lower diversity data (L-SVM) in prediction of genotoxic (GT+) and non-genotoxic (GT−) compounds based on five-fold cross validation tests

Method	Cross-validation	Genotoxicity			Non-genotoxicity			ACC (%)	G-Mean	MCC
		TP	FN	SEN (%)	TN	FP	SP (%)			
H-kNN (kNN with dataset same as H-SVM)	1	690	262	72.5	1,386	242	85.1	80.5	78.5	0.58
	2	688	264	72.3	1,399	229	85.9	80.9	78.8	0.59
	3	710	242	74.6	1,424	204	87.5	82.7	80.8	0.63
	4	695	257	73.0	1,429	199	87.8	82.3	80.1	0.62
	5	698	254	73.3	1,407	221	86.4	81.6	79.6	0.60
	Mean			73.1			86.5	81.6	79.5	0.60
	SD			0.9			1.1	0.9	1.0	0.02
	SE			0.4			0.5	0.4	0.4	0.01
L-kNN (kNN with dataset same as L-SVM)	1	445	219	67.0	1,379	267	83.8	79.0	74.9	0.50
	2	448	216	67.5	1,370	276	83.2	78.7	74.9	0.49
	3	435	229	65.5	1,384	262	84.1	78.7	74.2	0.49
	4	439	225	66.1	1,382	264	84.0	78.8	74.5	0.49
	5	437	227	65.8	1,375	271	83.5	78.4	74.1	0.48
	Mean			66.4			83.7	78.7	74.5	0.49
	SD			0.8			0.3	0.2	0.5	0.01
	SE			0.4			0.2	0.1	0.3	0.002

The performance for each fold is reported by true positive (*TP*), false negative (*FN*), accuracy of identifying GT+ or Sensitivity (*SEN*), accuracy of identifying GT− or specificity (*SP*), overall accuracy (*ACC*), geometric mean of sensitivity and specificity (G-Mean), Matthews correlation coefficient (*MCC*), and their mean, standard deviation (*SD*), standard error (*SE*)

Table 4 The performance of feedforward backpropagation neural network (FBNN) by higher diversity higher noise data (H-SVM) and by lower noise and lower diversity data (L-SVM) in prediction of genotoxic (GT+) and non-genotoxic (GT−) compounds based on five-fold cross validation tests

Method	Cross -validation	Genotoxicity			Non-genotoxicity			ACC (%)	G-Mean	MCC
		TP	FN	SEN (%)	TN	FP	SP (%)			
H-FBNN (FBNN with dataset same as H-SVM)	1	724	228	76.1	1,388	240	85.3	81.9	80.6	0.61
	2	732	220	76.9	1,431	197	87.9	83.8	82.2	0.65
	3	721	231	75.7	1,436	192	88.2	83.6	81.7	0.65
	4	716	236	75.2	1,391	237	85.4	81.7	80.1	0.61
	5	734	218	77.1	1,413	215	86.8	83.2	81.8	0.64
	Mean			76.2			86.7	82.8	81.3	0.63
	SD			0.8			1.4	1.0	1.1	0.02
	SE			0.4			0.6	0.5	0.5	0.009
L-FBNN (FBNN with dataset same as L-SVM)	1	435	229	65.5	1,445	201	87.8	81.4	75.8	0.54
	2	421	243	63.4	1,467	179	89.1	81.7	75.2	0.54
	3	414	250	62.3	1,461	185	88.8	81.2	74.4	0.53
	4	423	241	63.7	1,471	175	89.4	82.0	75.5	0.55
	5	418	246	63.0	1,465	181	89.0	81.5	74.9	0.54
	Mean			63.6			88.8	81.6	75.2	0.54
	SD			1.2			0.6	0.3	0.8	0.01
	SE			0.5			0.3	0.1	0.4	0.003

The performance for each fold is reported by true positive (*TP*), false negative (*FN*), accuracy of identifying GT+ or sensitivity (*SEN*), accuracy of identifying GT− or specificity (*SP*), overall accuracy (*ACC*), geometric mean of sensitivity and specificity (*G-Mean*), Matthews correlation coefficient (*MCC*), and their mean, standard deviation (*SD*), standard error (*SE*)

genotoxicity. Moreover, H-SVM produces slightly better performance than H-FBNN which in turn is slightly better than H-kNN, and L-SVM produces slightly better performance than L-FBNN which in turn is slightly better than L-kNN. This is consistent with a number of studies that suggests that in some classification studies including genotixicity and acute toxicity predictions SVM is superior than other machine learning methods [10, 31, 32].

Performance of SVM in searching genotoxic compounds from large compound libraries

The performance of H-SVM and L-SVM trained by GT+ compounds reported before 2008 in identifying GT+ compounds reported since 2008, and in classifying GT+ compound hits from clinical trial drugs, and MDDR and PubChem databases is summarized in Table 5. 81.6 and 57.9% of the 38 GT+ reported since 2008 was correctly identified, and 83.1 and 79.8% of the 2,008 clinical trial drugs as GT− by H-SVM and L-SVM respectively. H-SVM outperforms L-SVM by a substantial margin. Moreover, the performance of both H-SVM and L-SVM is comparable to the five-fold cross validation rates given in the previous section and the reported 50–94% active agent and 80–99% non-active agent identification rates of various VS tools [20] and the 58.8% GT+ and 79% GT− rates of Ames tests [12]. Strictly speaking, direct comparison with the reported performances is inappropriate because of the differences in the type, composition and diversity of compounds screened, and in the molecular descriptors, VS tools and their parameters used. The comparison cannot go beyond the statistics of accuracies as the reports are not detailed enough to address questions of whether all methods detect the same hit.

H-SVM and L-SVM identified as GT+ compounds 27.23 and 25.32% of the 17.86M PubChem compounds, 23.96 and 19.52% of the 168K MDDR compounds, and 36.76 and 31.59% of the 6,216 and 4,966 MDDR compounds similar to a GT+ agent in the training dataset. These are comparable to the estimated percentages of 23% in vivo and 31–33% in vitro GT+ compounds in the "universe of chemicals" [9]. This suggests that both H-SVM and L-SVM is capable of achieving relatively low false-positive rates in identifying GT+ compounds from large compound libraries, which are consistent with the reported low false-hit rates of 0.0054–8.3% of SVM in virtual screening large compound libraries [20]. Overall, H-SVM outperforms L-SVM, sometimes by a substantial margin, in identifying GT+ compounds in both five-fold cross validation and large compound libraries screening

Table 5 Performance of support vector machines trained by higher diversity higher noise data (H-SVM) and by lower noise and lower diversity data (L-SVM) for identifying genotoxic compounds from large compound libraries

Performance of H-SVM	
Percent of 38 newly reported genotoxic compounds correctly identified by SVM	81.6%
No and percent of identified genotoxic compounds outside training chemical families	11 (35.5%)
No and percent of 2,008 clinical trial drugs identified as genotoxic compounds	337 (16.8%)
No and percent of 17.86M pubchemcompounds identified as genotoxic compounds	4,864,006 (27.2%)
No and percent of the 168K MDDR compounds identified as genotoxic compounds	40,257 (24%)
No and percent of the 6,216 MDDR compounds similar to the known GT+ compounds identified as genotoxic compounds	2,285 (36.8%)
No and percent of the 15,484 MDDR compounds similar to the known GT− compounds identified as genotoxic compounds	12,083 (78%)
Performance of L-SVM	
Percent of 38 newly reported genotoxic compounds correctly identified by SVM	57.9%
No and percent of identified genotoxic compounds outside training chemical families	6 (27.3%)
No and percent of 2,008 clinical trial drugs identified as genotoxic compounds	406 (20.2%)
No and percent of 17.86M pubchemcompounds identified as genotoxic compounds	4,522,812 (25.3%)
No and percent of the 168K MDDR compounds identified as genotoxic compounds	32,803 (19.5%)
No and percent of the 4,966 MDDR compounds similar to the known GT+ compounds identified as genotoxic compounds	1,569 (31.6%)
No and percent of the 15,484 MDDR compounds similar to the known GT− compounds identified as genotoxic compounds	12,665 (81.8%)

Table 6 MDDR classes that contain higher percentage (≥3%) of H-SVM classified GT+ compounds in screening 168K MDDR compounds

MDDR classes that contain higher percentage (>3%) of H-SVM classified GT+ compounds	No and percentage of H−SVM classified GT+ compounds	Percentage of class members selected as H-SVM classified GT+ compounds (%)
Anti-neoplastic	4,848 (12.04%)	22.47
Anti-allergic/anti-asthmatic	2,326 (5.78%)	21.68
Anti-hypertensive	2,095 (5.2%)	19.59
Anti-arthritic	1,948 (4.84%)	25.32
Cognition disorders, agent for	1,752 (4.35%)	23.02
Anxiolytic	1,363 (3.39%)	20.16
Anti-depressant	1,232 (3.06%)	19.87
Anti-inflammatory	1,227 (3.05%)	22.04

The total number of SVM classified GT+ compounds is 40,257 (23.96%)

tests, suggesting that the performance gain from the enhanced training data outweighs the performance loss from increased noise levels.

The H-SVM and L-SVM identified GT+ and GT− MDDR compounds can be compared with the TOPKAT (Discover Studio) identified genotoxic and carcinogenic MDDR compounds. The comparison is presented as a confusion matrix in Table 7. There are 164,358 MDDR compounds uncategorized by TOPKAT because these compounds were determined by TOPKAT to be outside Optimal Prediction Space (OPS) or the distance from OPS was regarded by TOPKAT as exceeding permissible limits. An additional set of 3,455 MDDR compounds were not processed by TOPKAT due to TOPKAT identified errors. Nonetheless, analysis of TOPKAT identified genotoxic/carcinogenic agents (TOPKAT+) and non-genotoxic/non-carcinogenic agents (TOPKAT−) from the remaining 203 MDDR compounds indicates substantial level of agreement between the TOPKAT prediction results and those of H-SVM and L-SVM. TOPKAT identified 16.26% of the remaining 203 MDDR compounds as TOPKAT+. The H-SVM and L-SVM predicted percentages of 23.96 and 19.52% GT+ agents out of the 168K MDDR compounds are comparable to the TOPKAT predicted percentage. Moreover, 24 (72.72%) and 22 (66.67%) of the 33 TOPKAT+ agents were predicted by H-SVM and L-SVM as

Table 7 Comparison of prediction accuracies of our H-SVM and L-SVM with TOPKAT 168K MDDR compounds

	TOPKAT positive	TOPKAT negative	TOPKAT uncategorized	TOPKAT unprocessed	Total
Confusion matrix between L-SVM and TOPKAT					
L-SVM positive	23	16	31,984	780	32,803
L-SVM negative	10	154	132,374	2,675	135,213
Total	33	170	164,358	3,455	168,016
Confusion matrix between H-SVM and TOPKAT					
H-SVM positive	24	23	39,515	694	40,256
H-SVM negative	9	147	124,843	2,761	133,489
Total	33	170	164,358	3,455	168,016

H-SVM identifies 23.96% compound as GT+, *L-SVM* identifies 19.52% compounds as GT+ and *TOPKAT* identifies 19.41% as GT+ of total categorized compounds of MDDR

GT+, and 147 (86.47%) and 154 (90.59%) of the 170 TOPKAT- agents were predicted by H-SVM and L-SVM as GT−, suggesting substantial level of agreement between TOPKAT prediction and the prediction from H-SVM and L-SVM.

Evaluation of SVM classified MDDR genotoxic compound hits and possible mechanisms leading to their identification

H-SVM classified GT+ compound hits in MDDR were evaluated based on the known biological or therapeutic target classes specified in MDDR. Table 6 gives the MDDR classes that contain higher percentage (≥3%) of SVM classified GT+ compound hits and the percentage values. We found that 4,848 or 12.04% of the 40,257 classified GT+ compound hits belong to the anti-neoplastic class, which represent 22.47% of the 21,557 MDDR compounds in the class. The molecular mechanisms of genotoxicity include DNA intercalation by aromatic ring of a drug, DNA methylation, and DNA adduct formation [33]. Inhibitors of kinases involved in mitosis or chromosomal segregation may also cause chromosomal damage leading to genotoxicity. Many of the anti-cancer drugs are based on the same mechanisms and some of them have been found to cause genotoxicity in persons frequently exposed to some anti-cancer drugs [34]. It is thus not surprising that many compounds in the anti-neoplastic class were classified by H-SVM as GT+ compound hits because of similarities in mechanisms of compound actions or similarity in chemical structures.

Substantial percentages of the SVM classified compound hits belong to the anti-allergic/anti-asthmatic (5.78%), anti-hypertensive (5.2%), anti-arthritic (4.84%), and anti-inflammatory (3.05%) therapeutic classes. Some mitosis and cell cycle regulators have been found to regulate allergy-associated cytokine and chemokine production [35]. Inhibitors of these regulators may thus produce both anti-allergic [36] and genotoxic effects. Some anti-hypertensive compounds are known to cause chromosome delay and to a lesser extent chromosome breakage [37], induce DNA damage [38], and inhibit DNA methylation [39]. Reduction of DNA methylation may activate a cascade of genotoxic stress checkpoint proteins, resulting in phosphorylation of Chk1 and 2, gammaH2AX focus formation, and CDC25a degradation [40]. A substantial percentage of the marketed anti-hypertensive drugs (18%) have been tested positive in at least one genotoxicity assay [41]. Some anti-arthritic compounds also target regulators involved in mitosis and cell cycle [42]. It has been reported that nonsteroidal anti-inflammatory drugs may induce DNA damage via radiation absorption and subsequent energy transfer to DNA [43]. Therefore, some of the H-SVM classified GT+ comound hits in the anti-arthritic, anti-hypertensive, anti-inflammatory, and anti-allergic/anti-asthmatic classes may be selected because of similar mechanisms on mitosis, chromosome organization and DNA structure with respect to those of genotoxic compounds, or because of their structural similarity to the relevant compounds.

Moreover, 4.35, 3.39 and 3.06% of the H-SVM classified GT+ compound hits are in the cognition disorder agent, anxiolytic and anti-depressant classes, respectively. Neurodegenerative diseases associated with learning and memory impairment may be recovered by chromatin modifications that involve increased histone acetylation by histone deacetylase inhibitors [44], which in some cases cause elevated levels of DNA damage leading to genotoxicity [45]. Some compounds with anxiolytic and anti-depressant activities have been found to induce DNA strand breaks and G2/M cell cycle arrest [46]. Some anxiolytic compounds have been found to induce spindle aberrations and displacement of chromosomes from the equator [47]. Some compounds produce anti-depressant effects via inhibition of histone demethylation [48]. A significant percentage of anti-psychotics and anti-depressants (42%) are positive in at least one genotoxicity assay [49]. As many of these reported

mechanisms are similar to those of genotoxic compounds, some of them were expectedly selected by H-SVM as GT+ compounds.

Conclusion

This study showed that SVM is useful for facilitating the search of genotoxic compounds from large compound libraries based on higher diversity and higher noise training data currently available. Another advantage of SVM is that it does not require the explicit knowledge of the intrinsic mechanism of genotoxicity and the structural features of genotoxic compounds. Moreover, SVM is capable of searching large compound libraries at sizes comparable to the 17.86M PubChem and 168K MDDR compounds at false-positive rates comparable to those of other in vitro and in-silico methods without the need to define an applicability domain, i.e. it has a broad applicability domain that covers the whole chemical space defined by the PubChem and MDDR databases. Because of its high computing speed and generalization capability for covering highly diverse spectrum compounds, SVM can be potentially explored as early genotoxicity screening tools for facilitating toxicity evaluation in drug discovery.

References

1. Custer LL, Sweder KS (2008) Curr Drug Metab 9:978
2. Bolzan AD, Bianchi MS (2002) Mutat Res 512:121
3. Li Y, Luan Y, Qi X, Li M, Gong L, Xue X, Wu X, Wu Y, Chen M, Xing G, Yao J, Ren J (2010) Toxicol Sci 118(2):435
4. Snyder RD, Pearl GS, Mandakas G, Choy WN, Goodsaid F, Rosenblum IY (2004) Environ Mol Mutagen 43:143
5. Schwerdtle T, Ebert F, Thuy C, Richter C, Mullenders LH, Hartwig A (2010) Chem Res Toxicol 23(2):432–442
6. Tweats DJ, Blakey D, Heflich RH, Jacobs A, Jacobsen SD, Morita T, Nohmi T, O'Donovan MR, Sasaki YF, Sofuni T, Tice R (2007) Mutat Res 627:78
7. Kirkland D, Aardema M, Henderson L, Muller L (2005) Mutat Res 584:1
8. Snyder RD, Smith MD (2005) Drug Discov Today 10:1119
9. Rosenkranz HS (2003) Mutat Res 529:117
10. Li H, Ung CY, Yap CW, Xue Y, Li ZR, Cao ZW, Chen YZ (2005) Chem Res Toxicol 18:1071
11. White AC, Mueller RA, Gallavan RH, Aaron S, Wilson AG (2003) Mutat Res 539:77
12. Kirkland D, Speit G (2008) Mutat Res 654:114
13. Kirkland D, Pfuhler S, Tweats D, Aardema M, Corvi R, Darroudi F, Elhajouji A, Glatt H, Hastwell P, Hayashi M, Kasper P, Kirchner S, Lynch A, Marzin D, Maurici D, Meunier JR, Muller L, Nohynek G, Parry J, Parry E, Thybaud V, Tice R, van Benthem J, Vanparys P, White P (2007) Mutat Res 628:31
14. Hastwell PW, Chai LL, Roberts KJ, Webster TW, Harvey JS, Rees RW, Walmsley RM (2006) Mutat Res 607:160
15. Ritter D, Knebel J (2009) Genotoxicity testing in vitro - development of a higher throughput analysis method based on the comet assay. Toxicol In Vitro 23(8):1570–1575
16. Glick M, Jenkins JL, Nettles JH, Hitchings H, Davies JW (2006) J Chem Inf Model 46:193
17. Vasquez MZ (2010) Combining the in vivo comet and micronucleus assays: a practical approach to genotoxicity testing and data interpretation. Mutagenesis 25(2):187–199
18. Pfuhler S, Kirkland D, Kasper P, Hayashi M, Vanparys P, Carmichael P, Dertinger S, Eastmond D, Elhajouji A, Krul C, Rothfuss A, Schoening G, Smith A, Speit G, Thomas C, van Benthem J, Corvi R (2009) Mutat Res 680:31
19. Brambilla G, Martelli A (2009) Update on genotoxicity and carcinogenicity testing of 472 marketed pharmaceuticals. Mutat Res 681(2–3):209–229
20. Ma XH, Jia J, Zhu F, Xue Y, Li ZR, Chen YZ (2009) Comb Chem High Throughput Screen 12:344
21. Pearlman RS (1988) In: CONCORD User's Manual, Tripos, St. Louis, MO
22. Pochet N, De Smet F, Suykens JA, De Moor BL (2004) Bioinformatics 20:3185
23. Matthews BW (1975) Biochim Biophys Acta 405:442
24. Chin SF, Wang Y, Thorne NP, Teschendorff AE, Pinder SE, Vias M, Naderi A, Roberts I, Barbosa-Morais NL, Garcia MJ, Iyer NG, Kranjac T, Robertson JF, Aparicio S, Tavaré S, Ellis I, Brenton JD, Caldas C (2007) Using array-comparative genomic hybridization to define molecular portraits of primary breast cancers. Oncogene 26(13):1959–1970
25. Chou KC, Shen HB (2007) Large-scale plant protein subcellular location prediction. J Cell Biochem 100(3):665–678
26. Karakoc E, Cherkasov A, Sahinalp SC (2006) Bioinformatics 22:e243
27. Aha DW, Kibler D, Albert MK (1991) Instance-based learning algorithms. Machine Learning 6:37–66
28. Witten IH, Frank E (2005) Data Mining: practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco
29. Willett P (1998) J Chem Inf Comput Sci 38:983
30. Bostrom J, Hogner A, Schmitt S (2006) J Med Chem 49:6716
31. Liu XH, Ma XH, Tan CY, Jiang YY, Go ML, Low BC, Chen YZ (2009) J Chem Inf Model 49:2101
32. Xue Y, Li H, Ung CY, Yap CW, Chen YZ (2006) Chem Res Toxicol 19:1030
33. Bolzan AD BMS (2002) Mutat Res 512:121
34. Cavallo D, Ursini CL, Perniconi B, Francesco AD, Giglio M, Rubino FM, Marinaccio A, Iavicoli S (2005) Mutat Res 587:45
35. Wong WS (2005) Curr Opin Pharmacol 5:264
36. Sugita A, Ogawa H, Azuma M, Muto S, Honjo A, Yanagawa H, Nishioka Y, Tani K, Itai A, Sone S (2009) Int Arch Allergy Immunol 148:186
37. Andrianopoulos C, Stephanou G, Demopoulos NA (2006) Environ Mol Mutagen 47:169
38. Arencibia JM, Del Rio M, Bonnin A, Lopes R, Lemoine NR, Lopez-Barahona M (2005) Int J Oncol 27:1617
39. Csoka AB, Szyf M (2009) Epigenetic side-effects of common pharmaceuticals: a potential new field in medicine and pharmacology. Med Hypotheses 73(5):770–780
40. Unterberger A, Andrews SD, Weaver IC, Szyf M (2006) Mol Cell Biol 26:7575
41. Brambilla G, Martelli A (2006) Mutat Res 612:115
42. Park HJ, Lee SH, Son DJ, Oh KW, Kim KH, Song HS, Kim GJ, Oh GT, Yoon DY, Hong JT (2004) Arthritis Rheum 50:3504
43. Chouini-Lalanne N, Defais M, Paillous N (1998) Biochem Pharmacol 55:441
44. Fischer A, Sananbenesi F, Wang X, Dobbin M, Tsai LH (2007) Nature 447:178

45. Olaharski AJ, Ji Z, Woo JY, Lim S, Hubbard AE, Zhang L, Smith MT (2006) Toxicol Sci 93:341
46. Bezerra DP, Moura DJ, Rosa RM, de Vasconcellos MC, e Silva AC, de Moraes MO, Silveira ER, Lima MA, Henriques JA, Costa-Lotufo LV, Saffi J (2008) Mutat Res 652:164
47. Yin H, Baart E, Betzendahl I, Eichenlaub-Ritter U (1998) Mutagenesis 13:567
48. Lee MG, Wynder C, Schmidt DM, McCafferty DG, Shiekhattar R (2006) Chem Biol 13:563
49. Brambilla G, Mattioli F, Martelli A (2009) Toxicology 261:77