A Maximum Common Subgraph Kernel Method for Predicting the Chromosome Aberration Test

Johannes Mohr,*,† Brijnesh Jain,† Andreas Sutter,‡ Antonius Ter Laak,‡ Thomas Steger-Hartmann,‡ Nikolaus Heinrich,‡ and Klaus Obermayer†

School for Electrical Engineering and Computer Science, Berlin Institute of Technology, Berlin, Germany, and Bayer Schering Pharma AG, Berlin, Germany

Received September 25, 2009

The chromosome aberration test is frequently used for the assessment of the potential of chemicals and drugs to elicit genetic damage in mammalian cells in vitro. Due to the limitations of experimental genotoxicity testing in early drug discovery phases, a model to predict the chromosome aberration test yielding high accuracy and providing guidance for structure optimization is urgently needed. In this paper, we describe a machine learning approach for predicting the outcome of this assay based on the structure of the investigated compound. The novelty of the proposed method consists in combining a maximum common subgraph kernel for measuring the similarity of two chemical graphs with the potential support vector machine for classification. In contrast to standard support vector machine classifiers, the proposed approach does not provide a black box model but rather allows to visualize structural elements with high positive or negative contribution to the class decision. In order to compare the performance of different methods for predicting the outcome of the chromosome aberration test, we compiled a large data set exhibiting high quality, reliability, and consistency from public sources and configured a fixed cross-validation protocol, which we make publicly available. In a comparison to standard methods currently used in pharmaceutical industry as well as to other graph kernel approaches, the proposed method achieved significantly better performance.

1. INTRODUCTION

Besides a bacterial reverse mutation assay (Ames test), a chromosome aberration (CA) test[1] is frequently used for the assessment of the in vitro genotoxic potential of chemicals and drugs. In early drug discovery stages, testing is generally limited due to costs and compound availability. The development of in silico toxicity prediction tools constitutes a major strategy to overcome these hurdles. In silico models for the prediction of Ames mutagenicity are already successfully applied in drug discovery and development.[2] In order to perform a more comprehensive assessment of the mutagenic potential of validated hits and lead or development candidates, a model to predict the outcome of the CA test exhibiting high prediction accuracy and providing structure−activity information is urgently needed.

The purpose of the in vitro CA test is to evaluate the potential of a substance to induce structural chromosome aberrations in cultured mammalian cells. Since many chemicals interact with genetic material only after metabolic activation, the test compounds are typically examined in the absence and presence of a mammalian metabolizing system containing liver microsomes (S9 mix). Following incubation with the test compound for predetermined intervals in the absence or presence of S9-mix, cell cultures are treated with a metaphase arresting substance (e.g., colcemid) and harvested, and metaphase cells are analyzed microscopically for the presence of chromosome aberrations. A compound is classified as clastogen if it significantly induces chromosomal aberrations in the presence and/or absence of S9 mix. A compound is considered nonclastogenic if it does not cause increased chromosomal aberrations, neither in the presence nor absence of S9 mix.

Compared to the Ames test, the in silico prediction of the CA test is more challenging. First, various mechanisms may underlie the induction of chromosomal aberrations. Structural chromosome aberrations might be caused by interactions of compounds with DNA[3] or with enzymes involved in DNA replication or transcription.[4] However, they could also result from unfavorable cell culture conditions:[5] Test compounds leading to chromosome aberrations by an impairment of cell culture conditions (biologically not significant positives) are most probably contained in all CA data sets. However, as these positive test results are not directly related to the chemical structure, they are not amenable to structure-based modeling and thereby lead to decreased predictivity. Second, the quantity of publicly available high-quality experimental data is significantly lower than for the Ames test. Third, the experimental design of the CA test is less standardized than the Ames test (e.g., use of different cell lines) making uniform data comparison difficult.

Several commercial tools suitable for predicting the outcome of the CA test have been developed. For example, Derek for Windows[6] (Lhasa Limited, UK), an expert system providing known structure−activity relationships (SARs), contains 61 structure-based alerts for assessing chromosomal damage. MultiCASE[7] (Multicase Inc., USA), a correlative tool predicting toxicity on the basis of structural fragments statistically correlated with activity offers several modules

* Corresponding author. E-mail: johann@cs.tu-berlin.de.
† Berlin Institute of Technology.
‡ Bayer Schering Pharma AG.

10.1021/ci900367j
Published on Web 09/30/2010

for predicting chromosome aberration [A61 based on data from the national toxicology program (NTP) with $N = 805$ structures and other modules recently developed in co-operation with the FDA].[8] However, these commercial tools have shown limitations regarding predictive performance and adaptability to a company's chemical space.[9]

In the literature, both structure-based approaches applying the commercial software MultiCASE as well as descriptor-based machine learning approaches have been followed to predict the outcome of the CA test.[10−13] In such descriptor-based approaches, each molecule is represented by a large number of numerical descriptors treated as a vector in Euclidean space. On this vectorial representation, standard machine learning predictors, such as K-nearest neighbors (KNN), neural networks, or support vector machines (SVMs), can be trained. Due to the large number of descriptors, feature selection or construction methods need to be applied to reduce the dimensionality of the problem and thus to avoid overfitting. Although they reached a high prediction accuracy (up to 70−75% balanced), a disadvantage of the descriptor-based approaches to predict the outcome of the CA test is that the resulting models are difficult to interpret. They are therefore of limited use for guidance of structure optimization in drug discovery and do not fully meet the Organization for Economic Cooperation and Development (OECD) principles for (Q)SAR validation,[14] as one of these is a mechanistic interpretation of a prediction model, if possible.

In contrast to the above methods, which start out by generating a set of descriptors for each molecule and then compare these descriptors, graph-kernel methods are based on a structural comparison of chemical graphs. Kernel methods[15] are statistically well-founded and efficient pattern recognition methods, which use a kernel function $k(X, X') \in \mathbb{R}$ to measure the similarities of pairs of objects X and X', yielding a kernel matrix $\mathbf{K}$. Kernel methods usually involve solving a convex optimization problem in the often high-dimensional vector space that is implicitly defined by the kernel function. A famous example are SVMs,[15,16] which require the kernel function to be positive semidefinite, i.e., the kernel matrix needs to be symmetric and must not have negative eigenvalues. Although originally applied to vectorial data, kernel methods are now increasingly applied also to structured data, like sequences, trees, or graphs.[17,18] For chemical graphs, several positive semidefinite kernels have been suggested.[19−23] A graph kernel method closely related to our approach has been proposed by ref 24. This graph kernel is however restricted to the class of outerplanar graphs and therefore inapplicable to the more general chemical structures considered in this paper. Another class of kernels which have been proposed for molecules are optimal assignment kernels,[26,27] which, however, lack the property of positive semidefiniteness.[28] Therefore conventional SVMs are no longer applicable, since the underlying theory does only hold for positive semidefinite kernels. In contrast, the potential support vector machine (P-SVM),[29−31] a recently proposed kernel method for dyadic data, does not require the kernel to be square or positive semidefinite. In a previous work,[32] we made use of this fact by applying the P-SVM to building predictive QSAR models with indefinite three-dimensional (3D) kernels based on the optimal alignment of local bipods. While such 3D kernels based on a fixed molecular structure work well for modeling single receptor−ligand or enzyme−substrate bindings, they cannot account for the large number of potential mechanisms involved in chromosome aberration.

The goal of this work was to obtain a method for predicting the outcome of the CA test that yields models having high sensitivity and specificity, while at the same time including information about the molecular connectivities contributing to the prediction. Moreover, we intended to provide a well-defined benchmark environment for method comparison that allows to assess the statistical significance of differences in predictive accuracy.

Our contributions are as follows: (i) We propose a graph kernel method for chemical graphs. The 2D graph kernel is based on the concept of the maximum common subgraph (MCS) of two graphs. Since determining a MCS is NP-complete, exact algorithms that guarantee to return an optimal solution are too time-consuming for large data sets of graphs of even moderate size. We therefore resort to approximate solutions using the graduate assignment algorithm.[33] Since the proposed graph kernel is not positive semidefinite, we apply the P-SVM for model building. The novelty of the graph kernel method consists in successfully combining graph kernels with the P-SVM for the first time. (ii) We propose a novel visualization method that provides a loading of the molecular connectivities contributing to an individual prediction, which is considered of great value to lead optimization, as it gives relevant information for a path forward to toxicologists and chemists. In this sense, the proposed method yields informative instead of black-box models. (iii) We provide an informative model for prediction of the CA test showing a very high prediction accuracy, meeting the needs of lead identification, optimization, and application in a regulatory environment. (iv) For evaluation and model comparison, we compiled a new data set of 940 compounds with CA test results from public sources, thereby containing 261 additional substances compared to a previous collection.[13] Since we were specifically interested in developing a model to predict the outcome of the CA test, existing benchmark data sets could not be used, as they target different end points. In order to facilitate future comparisons with prediction methods of other researchers, we make this data set publicly available, including well-defined random 10-fold cross-validation splits. We also provide the individual values of predictive accuracy (average of sensitivity and specificity) achieved by the different methods on the single cross-validation folds, which allows to assess via a 10-fold cross-validated paired t-test whether the difference in predictive accuracy of two methods is statistically significant. We employed this benchmark environment to compare our method to MultiCASE, the Pipeline Pilot Bayesian classifier, vectorial machine learning methods using DragonX descriptors as inputs, and other graph kernel methods. In this comparison, the proposed MCS kernel method achieved significantly better results than all the other methods. Furthermore, the method correctly identified several electrophilic substructures, overall reflecting well-known structural alerts for DNA reactivity and mutagenicity, and substructures accounting for other mechanisms than a direct interaction with DNA.

2. METHODS AND DATA

2.1. Maximum Common Subgraph Kernel Method. In this section we describe the method for building informative prediction models for chemical compounds, which we call the MCS kernel method. It consists of three parts:

(i) MCS Kernel. A kernel function basically measures the similarity of a pair of given objects. In our case, the objects are chemical graphs representing molecules. Intuitively, we consider two chemical graphs as more similar the larger their common structural overlap is. Guided by this intuition, the MCS kernel asks for the maximum number of common bonds of two graphs. The common bonds of two graphs give rise to isomorphic subgraphs that need not to be connected. Thus multiple functional groups appearing in both molecules at different relative locations can also be matched. The MCS kernel takes only topology information into account, not geometrical distances, therefore it is a 2D graph kernel. Because the computational complexity of calculating this kernel grows exponentially with the number of atoms, an exact calculation of this similarity is intractable even for molecules of moderate size, and therefore, we have to resort to approximate solutions. As an approximate algorithm for determining the proposed MCS kernel, we suggest to apply the graduated assignment algorithm.[33] The MCS kernel and its calculation are described in Section 2.1.1.

(ii) Model Building with the P-SVM. Since it involves a maximization operation over kernels, the MCS kernel is not positive semidefinite. Since it is not a valid Mercer kernel, a conventional SVM cannot be applied without violating its underlying theoretical foundation. Instead, we suggest the use of the P-SVM,[29,31] a recently proposed kernel method for feature selection, classification, and regression on dyadic (relational) data. The P-SVM is not based on a margin concept and does not require positive semidefinite kernels. This method was already applied in our previous work to the indefinite 3D molecule kernels based on matching bipod alignments.[32] Details on the process of model building and prediction are given in Section 2.1.2.

(iii) Visualization of SAR Information. A kernel provides a scalar value for each pair of molecules, which are then used in the model to predict the class label, i.e., the test result, of a compound unknown to the model. However, usually kernel methods do not provide an explanation for the classification decision. In this work, we therefore introduce a method for retaining the information which structural elements contribute to the classification throughout the model building process. This is basically done by storing the responsibilities of each structural element in the kernel calculation for each pair of molecules and weighing these by the respective linear coefficient (Lagrange parameter) in the prediction function. A visualization scheme is introduced that marks the atoms involved in the bonds with the strongest influence on the prediction, allowing to identify functional groups in favor of a positive or negative classification. The details are found in Section 2.1.3.

2.1.1. Maximum Common Subgraph Kernels. Maximum common subgraph (MCS) kernels are a family of similarity functions based on the common structural overlap of two given graphs. We consider two graphs being more similar the larger their common structural overlap. In order to precisely quantify the common structural overlap of two graphs, we introduce some graph theoretical concepts.

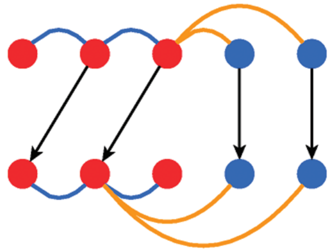


Figure 1. Example of a morphism of two graphs shown in the upper and lower row. Different colors refer to different vertex and edge attributes. Black arrows indicate a partial mapping between both vertex sets. The mapping can be completed to a total mapping by assigning the first vertex of the upper graph to the third vertex of the lower graph, since both vertices have the same color.

An undirected graph is a tuple $X = (V, E)$ consisting of a finite nonempty set V of vertices and a set $E \subseteq V^{[2]}$ of edges, where $V^{[2]} = \{\{i, j\} \subseteq V: i \neq j\}$ is the set of all two element subsets of V. In a chemical structure, the vertices of a graph represent the atoms, and the edges of a graph denote the bonds connecting each pair of covalently bonded atoms. We annotate properties of atoms and bonds by assigning attributes to the corresponding vertices and edges. Suppose that A and B are disjoint sets of attributes representing atom types and bond types, respectively. A chemical graph is a quadruple $X = (V, E, \alpha, \beta)$ consisting of an undirected graph (V, E) together with an attribute function $\alpha: V \rightarrow \mathrm{A}$ that assigns an atom type to each vertex and an attribute function $\beta: E \rightarrow \mathrm{B}$ that assigns a bond type to each edge. In principle any bond type can be used; in the experiments we restricted ourselves to single, double, and triple bonds and did not include aromatic bonds as a separate bond type. In what follows, we refer to chemical graphs briefly as graphs.

A graph $Y = (V', E', \alpha', \beta')$ is a subgraph of a graph $X = (V, E, \alpha, \beta)$ if

(i) $V' \subseteq V$
(ii) $E' \subseteq E \cap V'^{[2]}$
(iii) $\alpha' = \alpha|_{V'}$
(iv) $\beta' = \beta|_{E'}$

A morphism between two graphs $X = (V, E, \alpha, \beta)$ and $X' = (V', E', \alpha', \beta')$ is a bijective mapping:

$$\phi: U \rightarrow U', \qquad i \mapsto i^{\phi}$$

between subsets $U \subseteq V$ and $U' \subseteq V'$ such that $\alpha(i) = \alpha(i^{\phi})$. Thus, a morphism is an one-to-one mapping between subsets of V and V' that maps vertices with the same attribute onto each other. By $M(X, X')$ we denote the set of all morphisms between X and X'. Figure 1 shows an example of a morphism.

We say $\{i, j\} \in E$ and $\{r, s\} \in E'$ have a common edge, written as $\{i, j\} \equiv \{r, s\}$, if there is a morphism ϕ between X and X' satisfying the following properties:

(i) $i^{\phi} = r$, with $\alpha(i) = \alpha(r)$
(ii) $j^{\phi} = s$, with $\alpha(j) = \alpha(s)$
(iii) $\beta(\{i, j\}) = \beta(\{r, s\})$

The set of common edges defined by a morphism ϕ determines isomorphic (i.e., common) subgraphs Y of X and Y' of X'. The score $s(\phi)$ of a morphism ϕ is its number of common edges, which is exactly the number of edges of any graph isomorphic to the subgraph Y (and Y'). An example is shown in Figure 2.

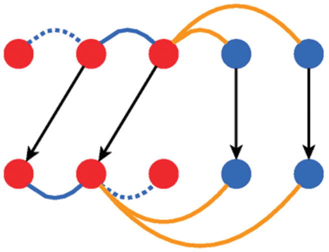


Figure 2. Common edges are highlighted. In this example, the number of common edge of the shown morphism is three, which is also the maximum number of common edges. Note that common edges must have the same color.

The canonical MCS kernel $k_{\text{MCS}}(X, X')$ is then defined by the maximum number:

$$k_{\text{MCS}}(X, X') = \max_{\phi \in M(X,X')} s(\phi) \quad (1)$$

of common edges of X and X'. The canonical MCS kernel is the simplest of the family of MCS kernels. Since the MCS kernel is biased toward larger graphs, we consider the Tanimoto MCS kernel defined by

$$k_{\text{T}}(X, X') = \frac{k_{\text{MCS}}(X, X')}{|E| + |E'| - k_{\text{MCS}}(X, X')} \quad (2)$$

The difficult task in determining the Tanimoto MCS kernel $k_{\text{T}}(X, X')$ is the computation of the MCS of X and X'. It is well-known that this problem is NP complete.[33] Algorithms that guarantee to return an exact value of $k_{\text{MCS}}(X, X')$ are useless for all but the smallest graphs due to their exponential runtime complexity. We therefore have to resort to approximate algorithms that return satisfactory approximations within an acceptable period of time. For approximating the MCS kernel, we have chosen the graduated assignment algorithm.[33] Graduated assignment is a state-of-the-art algorithm that is widely applied for solving graph matching problems because of its high speed and good solution quality. In addition, since graduated assignment exploits sparsity of graphs, it is well suited for molecules.

Let $X = (V, E, \alpha, \beta)$ and $X' = (V', E', \alpha', \beta')$ be graphs. Suppose that X has $|V| = n$ vertices and X' has $|V'| = n'$ vertices. Without loss of generality, we assume that $n \leq n'$. We cast the problem of determining the maximum number of common edges to a continuous optimization problem and then apply the graduated assignment algorithm.[33] To this end, we set up a compatibility matrix $\mathbf{C} = (c_{irjs})$ with elements:

$$c_{irjs} = \begin{cases} 1: & \{i, j\} \equiv \{r, s\}, \text{ where } \{i, j\} \in E, \{r, s\} \in E' \\ 0: & \text{otherwise} \end{cases}$$

Maximizing the number of common edges is then equivalent to minimizing the cost function:

$$E(\mathbf{M}) = -\sum_{i=1}^{n} \sum_{r=1}^{n'} \sum_{j=1}^{n} \sum_{s=1}^{n'} m_{ir} \cdot m_{js} \cdot c_{irjs}$$

subject to the constraints:

$$\mathbf{M} = (m_{ir}) \in \{0, 1\}^{n \times n'} \quad (3)$$

$$\sum_{r=1}^{n'} m_{ir} \leq 1 \qquad \forall i \in \{1, ..., n\} \quad (4)$$

$$\sum_{i=1}^{n} m_{ir} \leq 1 \qquad \forall r \in \{1, ..., n'\} \quad (5)$$

We call a $(n \times n')$ matrix $\mathbf{M} = (m_{ir})$ that satisfies the constraints (3)−(5) a match matrix. A match matrix consists of at most one element with value one in each row and column, and all other elements are zero. Nonzero elements $m_{ij} = 1$ encode a mapping of vertex i of graph X to vertex j of graph X'. If $n = n'$, i.e., if both graphs X and X' have the same number of vertices, then a match matrix is a permutation matrix.

This discrete problem can be formulated as a continuous problem by relaxing constraint (3) to $\mathbf{M} = (m_{ir}) \in [0, 1]^{n \times n'}$, assuming that the elements m_{ir} take continuous values from the interval [0, 1]. The core of the graduated assignment algorithm implements a deterministic annealing process with annealing parameter T by the following iteration scheme:

$$m_{ij}^{(t+1)} = a_i b_j \exp\left(-\frac{1}{T} \sum_{r=1}^{n} \sum_{s=1}^{m} m_{rs}^{(t)} c_{ijrs}\right) \quad (6)$$

where t denotes the time step. The scaling factors a_i, b_i computed by Sinkhorn's algorithm[34] enforce the constraints of the match matrix. After termination, graduated assignment returns a feasible solution $\mathbf{M}' = (m'_{ir})$ of the continuous problem, which is in general not a valid match matrix of the discrete problem. In order to obtain a feasible solution of the original discrete problem, we convert the continuous solution $\mathbf{M}'$ to a discrete one by applying Munkres algorithm.[35] The resulting discrete match matrix $\mathbf{M}$ is a matrix representation of a morphism ϕ_M between X and X', where $\phi_M(i) = r$, if and only if $m_{ir} = 1$. Substituting the score $s(\phi_M)$ of the morphism ϕ_M as an approximation of the canonical MCS kernel $k_{\text{MCS}}(X, X')$ into the Tanimoto MCS kernel gives the approximation:

$$\tilde{k}_{\text{T}}(X, X') = \frac{s(\phi_M)}{|E| + |E'| - s(\phi_M)} \quad (7)$$

2.1.2. P-SVM for Model Building and Prediction. Let G denote the set of all possible chemical graphs of interest for the underlying CA test. Suppose that we are given a training set $X_1, ..., X_m \in G$ of m chemical graphs together with their respective class labels $y_1, ..., y_m \in \{-1,+1\}$ that encode a negative (−1) or positive (+1) CA test outcome. Given a training set, the aim of model building is to obtain a prediction function $f(X)$: $G \rightarrow \{-1, +1\}$ that assigns to each chemical graph $X \in G$ a class label $y \in \{-1, +1\}$ such that the expected classification error (generalization error) is low.

First, the kernel matrix $\mathbf{K}$ on the training set is calculated by evaluating the kernel $k(X_p, X_q)$ between the chemical graphs X_p and X_q of all pairs of compounds from the training set. Here, the kernel k corresponds to $\tilde{k}_{\text{T}}(X, X')$, eq (7), which is the Tanimoto MCS kernel approximated using the graduated assignment algorithm. For a training set containing m compounds, this corresponds to a $m \times m$ symmetric matrix with ones on the diagonal, which requires the calculation of $m(m - 1)/2$ scalar kernel values.

Since the kernel matrix is indefinite, the P-SVM[29,31] is used for building a model. The P-SVM is a recently proposed kernel method for classification, regression, or feature selection, which does not require the kernel matrix to be square or positive semidefinite. Details on the P-SVM are provided in Appendix 6.1.

In order to build a model, the P-SVM is trained using both the labels and the calculated kernel matrix **K**. The P-SVM requires the selection of two hyperparameters (C and ε). This is done by minimizing the 10-fold cross-validated prediction error on the training set over a discrete grid of hyperparameter values. Then the P-SVM is trained at the optimal hyperparameters using the full training set, which yields the final model. The prediction function for a molecule X is specified via the set of α values and an offset b. It takes the form:

$$f(\mathrm{X}) = \mathrm{sgn}\left(\sum_{q=1}^{m} \alpha_q k(X_q, X) + b\right) \tag{8}$$

Note that only the molecules X_q corresponding to nonzero α_q, the support molecules, are needed for prediction.

This model can then be used for prediction in the following way: First, the kernel matrix $\mathbf{K}_{pr}$ is calculated between the set of m_{sv} support molecules and the set of m_{pr} molecules for which we wish to predict the test outcome. This is done by evaluating the MCS kernel for all pairs involving a member from each of these sets. The calculation of the $m_{sv} \times m_{pr}$ kernel matrix $\mathbf{K}_{pr}$ thus requires $m_{sv} \cdot m_{pr}$ kernel evaluations. Predictions of class labels are obtained via eq 8. There are two variants of the MCS kernel method, corresponding to the balanced and unbalanced version of the P-SVM. For training sets with different class size, the balanced version enforces a similar training error rate for each class (see Appendix for details).

2.1.3. Visualizing SAR Information. It is desirable to have interpretable models, that include SAR information. In the following, we propose a method of how such models can be obtained with the MCS kernel method. The method consists of the calculation of so-called influence vectors that describe the impact each bond has on the classification of an unlabeled molecule.

Each kernel entry $k(X_p, X)$ between a molecule X and a support molecule X_p can be decomposed into the sum of the common edges, multiplied by a normalization factor g_p. The prediction function, eq 8, is based on a linear combination of these kernel values with coefficients α_p.

For a given kernel entry $k(X_p, X)$ let us represent the $|E|$ bonds in molecule X by a vector $\mathbf{m}_p$ of length $|E|$, which takes the value g_p, if the respective bond had a match in the optimal morphism and zero else. The relative contribution which the individual bonds in molecule X have on the prediction can then be calculated as an influence vector **w**:

$$\mathbf{w} = \sum_{p=1}^{m} \alpha_p \mathbf{m}_p \tag{9}$$

The predicted class for molecule X can then be simply calculated as the sign of the sum over the components of this vector plus the offset b:

$$f(X) = \mathrm{sign}(\sum_{i=1}^{|E|} w_i + b) \tag{10}$$

Note that the P-SVM transforms the columns of the kernel matrix to zero mean and variance one. Therefore, in order for eqs 9 and 10 to be applicable, the values $\tilde{\alpha}$ and $\tilde{b}$ returned by the P-SVM algorithm have to be adjusted as follows

$$\alpha_p = \frac{\tilde{\alpha}_p}{\sigma_p} \tag{11}$$

$$b = \tilde{b} - \sum_{p=1}^{m} \frac{\tilde{\alpha}_p}{\sigma_p} \mu_p \tag{12}$$

where

$$\mu_p = \frac{1}{m} \sum_{q=1}^{m} k(X_p, X_q) \tag{13}$$

and

$$\sigma_p = \sqrt{\frac{1}{m} \sum_{q=1}^{m} (k(X_p, X_q) - \mu_p)^2} \tag{14}$$

are the mean and standard deviation for each column p of the kernel matrix.

The influence vector directly shows the impact the presence of a certain bond has on the individual classification of an unlabeled substance as positive or negative. In the following, we propose a method to visualize groups of atoms which have a strong positive (i.e., contributing to genotoxicity) or negative (i.e., lowering the probability of a genotoxic effect) impact on an individual classification decision.

In this visualization scheme, a threshold value is chosen for each compound of interest, which can be lowered until functional groups of interest are visible. Bonds having an absolute value of the influence component lower than this value are considered neutral. Bonds above this threshold are considered positive, if their influence is positive, and negative otherwise. Atoms that are only having positive bonds are depicted in red, atoms having only negative bonds are colored in blue, atoms having both positive and negative bonds are shown in purple, and atoms which have only neutral bonds are depicted in gray. Groups of red and blue atoms therefore correspond to connectivities with positive and negative contributions, respectively, to the classification decision.

2.2. CA Test Data Set. In order to compare different methods for generating predictive models from given training data a large and clearly defined data set is needed. For this purpose, we compiled a CA test data set from information contained in the literature[11,36,37] and in VITIC[6] using the software Pipeline Pilot. The compilation by Snyder et al.[36] comprises in vitro CA test data for marketed pharmaceuticals. CA tests were performed using diverse cell types (Chinese hamster ovary and lung cells, V79 cells, MCL-5 human lymphoblastoid cells, and human blood peripheral lymphocytes). The database collected by Kirkland et al.[37] is structurally more diverse, containing in vitro CA test data for industrial and environmental chemicals as well as pharmaceutical compounds. Similar to the Snyder data, results obtained with all different cell types are included in

Table 1. Compiled In Vitro CA Test Data Set

source	positive	negative
Kirkland et al.[37]	282	168
Snyder et al.[36]	40	189
Serra et al.[11]	107	263
VITIC[6]	163	228
total (after merging) $N = 940$	351	589

this compilation. The authors carefully reviewed collected test results in order to ensure a high quality and consistency of the data. The in vitro CA test data collected by Serra et al.[11] is structurally diverse, consisting of industrial, household, cosmetic, and pharmaceutical compounds. Data from this publication were generated using cultured Chinese hamster lung cells only. The VITIC database[6] contained carefully reviewed in vitro CA test data from the Ishidate clastogenicity data collection[38] (database contains results obtained with over 60 different cell types, with the vast majority of the compounds yielding consistent results), National Toxicology Program (NTP)[39] studies (CA test performed with Chinese hamster ovarian cells), the International Uniform ChemicaL Information Database (IUCLID)[40] of the European Chemicals Bureau (ECB) comprising tests generated using human leukocytes and lymphocytes, HeLa cells, mouse mammary carcinoma cells, mouse lymphoma cells, and Chinese hamster ovary and fibroblast cells, and collected from the scientific literature (Chinese hamster ovary cells).

In general, the intra- and interlaboratory reproducibility of the CA test is affected by the purity of the test compound, the cells and reagents used, experimental procedures, level of cytotoxicity, and scoring uncertainty, and the interpretation of results.[41] However, despite these experimental differences, the data used in the present work are well-documented and meet either regulatory standards[11,36] or have been reviewed extensively for quality and consistency.[6,37] Thus, a high quality, reliability, and consistency can be assumed for the used data set. The data set will be provided online.

About 35% of the compounds in the final data set are reported in more than one source. We found that 24 of 1005 compounds (2.4%) had contradictory results in different sources and removed them from the data set. A number of 15 molecules with less than 3 atoms were excluded since they cannot be processed by the MCS kernel method for comparison. A small number of 26 extraordinary or inorganic molecules have been omitted due to MultiCASE's limitations. The final chromosome aberration data set contains 940 unique compounds as canonical SMILES together with the corresponding in vitro CA test results and data sources (see Supporting Information). Additionally, the names of 625 compounds, 403 of which being listed in the World Drug Index (43% of the data set) and Chemical Abstracts Service (CAS) numbers of 794 compounds are provided. The composition of the chromosome aberration data set is shown in Table 1. The table shows the number of compounds and results taken from each source when stepwise extending the data set. Due to overlaps between different sources, the total number of relevant data contained in the individual databases can be higher. The mean molecular weight of the data set is 251 Da (median MW 215 Da, see Figure 3).

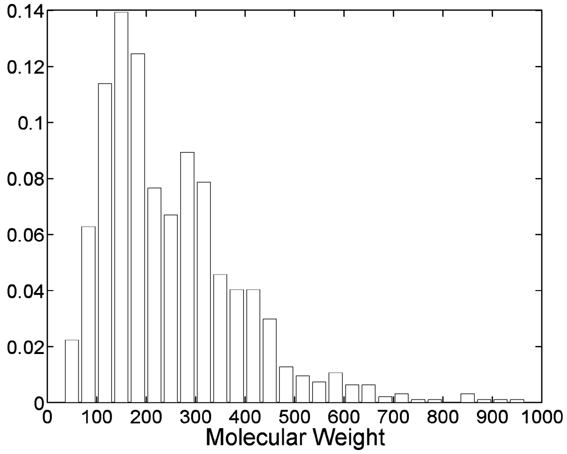


Figure 3. Histogram of the molecular weights of the in vitro CA test data set.

2.3. Methods Used for Comparison. The predictive performance of the proposed MCS kernel method was compared to MultiCASE, four Pipeline Pilot Bayesian classifiers with different fingerprints, nine machine learning methods based on DragonX descriptors, and fourteen graph kernel methods which apply a support vector machine to positive definite graph kernels. In order to compare methods for model building, it is required that they are trained and tested on exactly the same data sets. Therefore only adaptable methods which could be trained from scratch were used in the comparison. This ruled out a comparison to Derek for Windows, which is an expert system based on rules extracted from a given database. Still, Derek for Windows was used for the identification of substructures related to chromosome damage or Ames mutagenicity. In an analysis of 662 compounds, most of the Ames positives also caused chromosome aberrations.[12] Direct interaction of DNA with electrophilic compounds is the most common mechanism for both end points. The SARs in Derek for Windows for chromosome damage and Ames mutagenicity are partly distinct, however, reflecting in most cases electrophilic properties. Thus, by including all SARs for both end points in the analysis, information was obtained for a higher number of compounds. In the following, we briefly review the methods used for comparison.

2.3.1. Descriptor-Based Machine Learning Methods. Since machine learning methods based on a set of extracted descriptors have previously been applied to the problem of in vitro CA test prediction, we also included such methods in our comparison. For this purpose, the software DragonX was used to calculate a total of 1495 descriptors for all compounds. These descriptors were then considered as components of a vector in a Euclidean space, in which standard machine learning methods can be applied. In this work we combine three approaches for dimensionality reduction with three classifiers, resulting in a total of nine algorithms included in the comparison.

In earlier work, it was shown that several blocks of computationally expensive Dragon descriptors can be omitted without significantly impacting the performance.[42] The models described in this paper are based on DragonX, version 1.2, descriptors from blocks 1, 2, 6, 9, 12, 15−18, and 20. These include, among others, constitutional descriptors, topological descriptors, walk and path counts, eigenvalue-based indices, functional group counts, and atom-centered fragments. Most of the descriptors depend only on the 2D structure of the molecule, while some actually take 3D information into account. Therefore the 3D molecule structures were calculated using Corina, version 3.4.[43] A full list of Dragon descriptors including references can be found online.[44]

2.3.1.1. Feature Selection/Construction. Due to the high dimensionality of the descriptor space, techniques for dimensionality reduction were employed. Conceptually, one distinguishes between feature selection methods,[45−47] which chose a subset of the descriptors as an input variable, and feature construction methods, which construct a reduced number of complex features obtained, e.g., by linear combination if the original descriptor variables. We used three different methods: (i) Principal component analysis (PCA) is a feature construction method, where the directions of largest variance in data space are found. Dimensionality reduction is achieved by projecting the data on the H components, corresponding to the largest eigenvalues. H is a hyperparameter, which has to optimized on the training set. (ii) Feature selection via concave minimization (FSV) is a feature selection technique[48] which aims at finding a separating hyperplane that correctly classifies the training set with the minimum possible number of features. In the concave minimization approach, a separating plane is generated by minimizing a weighted sum of the distances of misclassified points to two parallel planes that bound the sets and determine the separating plane midway between them. Additionally, the number of dimensions used to determine the plane is minimized. We used the Spider[49] implementation of FSV and adjusted the number of features F as hyperparameter. (iii) The P-SVM is a kernel method which can be used for classification, regression, and feature selection. While in the MCS kernel method the P-SVM is used for classification, we use it here for feature selection on the set of DragonX descriptors. In the feature selection mode, the P-SVM has only one hyperparameter ε, which implicitly controls the number of selected features. Details on the P-SVM can be found in Appendix 6.1.

2.3.1.2. Classification. After the dimensionality reduction, a classifier was trained on the selected (or constructed) features. We used the following three classifiers having different theoretical backgrounds: (i) KNN is a nonparametric method which bases the classification of a new data point on a voting of the class labels of its K nearest neighbors. Its only hyperparameter is the number K. We used the Spider[49] implementation of KNN. (ii) The C-support vector machine (C-SVM) is based on the principles of statistical learning theory.[16] It tries to find the separating hyperplane with the largest margin, which is the distance from the closest data point to the hyperplane. There is theoretical consideration that links the size of the margin to lose upper bounds on the generalization error, and in fact, SVMs have shown excellent generalization performance in practice. While often kernels are used to nonlinearly transform the data into a higher dimensional feature space, we used a linear kernel to account for the already high dimensionality of the data. The linear C-SVM has only one hyperparameter, C, which accounts for outliers by allowing small violations of the margin. For the experiments, we used the Libsvm[50] implementation, which is able to account for unbalanced class size. (iii) A Random forest[51] is an ensemble classifier that consists of many decision trees. It uses bootstrap-aggregation[52] ("bagging") to generate many new data sets from a given training set. Then the classification and regression trees (CART) algorithm is employed to generate one decision tree for each data set. Classification is done by majority voting over these classifiers. The only hyperparameter was K, the number of trees. We used the Spider[49] implementation for the experiments.

2.3.2. MultiCASE. MultiCASE (version 2.1, Multicase Inc., Beachwood, OH) is a correlative tool predicting toxicity on the basis of structural fragments statistically correlated with activity (QSAR). It builds predictive toxicity models on the basis of structural fragments of 2−10 atoms correlated with activity and inactivity. It provides SAR information in the form of biophores and biophobes. 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 not considered when building the data set. MultiCASE assigns a probability to predictions that is calculated as the ratio of positive examples and the total number of examples containing the fragment of interest. A very high probability is interpreted as a high confidence level, resulting in an unambiguous positive prediction. A low confidence results in an ambiguous prediction ("possibly active"). An ambiguous negative prediction may be due to the absence of biophores in a structure, the fragments of which being partly unknown to the model.

2.3.3. Derek for Windows. Derek for Windows (version 10.0.2, service pack 3, Knowledge Base Release DfW 10.0.0_25_07_2007, Lhasa Limited, UK) is an expert system providing known SARs, each relating to a toxicological end point. All compounds contained in the data set were evaluated if they triggered one or more structural alerts for chromosome aberration or Ames mutagenicity. This analysis was used as a benchmark to evaluate the plausibility of the visualization generated by the MCS kernel method.

2.3.4. Pipeline Pilot Bayesian Classifiers. In another effort we combined Bayesian classifiers with Pipeline Pilot's extended connectivity fingerprint technology (Pipeline Pilot, version 7.0, SciTegic/Accelrys, San Diego, CA). Extended connectivity fingerprint generation begins with the assignment of an initial atom code for each heavy atom in the molecule. The atom codes differ for standard extended connectivity fingerprints ECFP (atom type), FCFP (functional class), and SCFP (Sybyl atom types). A hashing scheme is used to generate numbers from each atom code together with its nonhydrogen attachments, which in turn are defined by pairs of bond- and atom-type codes. An iterative process is used to derive these numbers in larger and larger structural neighborhoods. When the desired neighborhood size is reached, the process is complete, and the set of all hash function numbers is returned as the extended connectivity fingerprint. For example, "FCFP_6" generates features around each atom up to a diameter of six bonds, which requires three iterations, because each iteration increases the diameter of the neighborhood by two bonds. An additional Bayesian classifier was built using MDL public keys as descriptors. MDL public keys are 166 of a total of 960 MDL keys, mostly substructural features, developed for rapid substructural searching of ISIS databases (Symyx Technologies Inc., formerly Molecular Design Limited, Inc.) and are available in Pipeline Pilot, version 7.0.

2.3.5. Other Graph Kernel Methods. Although graph kernel have not previously been applied for the CA test, they are state-of-the-art in machine learning. We therefore included fourteen methods based on positive definite graph kernels and support vector classification in the comparison. Specifically, we evaluated the 3D spectrum kernel,[22] which is a discrete approximation to the pharmacophore kernel,[20] the marginalized graph kernel,[19] the walk-based spectrum kernel,[22] the subtree kernel,[20] the Tanimoto graph kernel,[23] the min/max Tanimoto graph kernel,[23] and the λ-k kernel,[22] a fast approximation of the geometric kernel.[18] These graph kernels were calculated using ChemCpp.[25] In order to separate effects of the kernel and the classifier, we combined the different graph kernels first with a standard C-SVM (using the Libsvm[50] implementation) as classifier and then with the P-SVM, yielding 2 × 7 combinations. We chose the default values of all kernel options, where possible. Where applicable, we allowed the length of molecular fragments (number of covalent bonds) to range from 0 to 20 (the maximum value possible for our data set).

2.4. Procedure for Method Comparison. All methods were evaluated in a 10-fold cross-validation setting. The data set was randomly divided into 10 disjoint sets of equal size, T_1, ..., T_{10}. Ten trials were conducted, where in the i^{th} trial T_i was used as the test set, while the training set corresponded to the union of the other sets T_j, $j \neq i$. For each cross-validation fold, the models were built only on the training sets, and their predictive performance was evaluated on the test set.

The parameters of the MCS kernel method, the other graph kernel methods, and the descriptor-based machine learning algorithms were selected for each training set via a grid search using an inner loop of cross validation. For each training set of the outer cross validation loop (used to assess the generalization performance), an inner cross validation was conducted for all combinations of a set of hyperparameter candidate values lying on a grid. The accuracy (average of sensitivity and specificity) was calculated for each grid point, and the parameters at which the highest accuracy was achieved in the inner cross-validation loop were used to train the model on the respective training set of the outer cross-validation loop. This way the selection of hyperparameters is considered part of the training process, and the results are valid estimates of the generalization performance of the final model on new data from the same distribution as the training data.

For the Pipeline Pilot Bayesian classifiers, several fingerprint connectivity lengths (nCFP4 to nCFP14) have been evaluated, and the optimal length for each fingerprint type was used for the method comparison. Note that this could bias the Pipeline Pilot results to be overly optimistic, since the selection of hyperparameters was not part of the training process (i.e., not carried out separately for each training set of the cross validation).

For the evaluation, predictive values for the following performance measures were calculated on each test set of the cross-validation procedure: sensitivity (TP/(TP + FN)), specificity (TN/(TN + FP)), coverage (percentage of test data for which a prediction was possible), and accuracy (defined as the average of sensitivity and specificity). For the end point of predicting the in vitro CA test, high sensitivity and specificity are equally desirable, therefore, we mainly base our evaluation on their average value, the accuracy.

We then tested for each pair of methods whether differences in the average of the accuracy values over the 10 cross-validation folds were significant. The null hypothesis is that method A does not have a higher accuracy than method B. For testing whether the null hypothesis can be rejected, we used a one-sided k-fold cross validation paired t-test. Let $d_i = \text{acc}(A) - \text{acc}(B)$, $i = 1,..., k$ denotes the difference in accuracy values of methods A and B for each of the $k = 10$ folds, and let $\bar{d} = 1/k\sum_{i=1}^{k} d_i$ be the average difference. Under the assumption that the k differences were drawn independently from a normal distribution, one can apply Student's t-test by computing the statistic:

$$t = \frac{\bar{d}\sqrt{k}}{\sqrt{1/(k-1)\sum_{i=1}^{k}(d_i - \bar{d})^2}} \tag{15}$$

Under the null hypothesis, this statistic has a t-distribution with $k - 1$ degrees of freedom. This is a one-sided test, where H_0 can be rejected at a significance level $\alpha = 0.05$, if $t > t_{9,\,0.95} = 1.833$.

3. RESULTS

The results of the method comparison are summarized in Table 2. Shown are the mean and standard deviation of sensitivity, specificity, coverage, and accuracy (defined as an average of sensitivity and specificity). The values shown are always the mean values over the 10 cross-validation folds ± the standard deviation.

First we will look at the coverage of the methods. Since the unambiguous variant of the MultiCASE prediction refuses to classify about one-fourth of the test data and does not yield a significantly higher accuracy (see below and Table 2), the ambiguous version with 98% coverage should be preferred in practice. The descriptor-based machine learning methods did almost achieve total coverage, except for a few molecules for which DragonX could not calculate some of the descriptors. The Pipeline Pilot Bayesian classifiers, the MCS kernel, and the other graph kernel methods achieved 100% coverage.

The (ambiguous) MultiCASE method achieved on average 43.2% sensitivity and 76.7% specificity, corresponding to an accuracy of 59.9%. The results for the unambiguous method were only slightly higher, however, at the cost of severely reduced coverage. The performance of most of the descriptor-based machine learning methods differs only slightly, with the best method, PCA+C-SVM, achieving on average a sensitivity of 73.1% and a specificity of 67.0%, which gives an accuracy of 70.0%. The Pipeline Pilot Bayesian classifiers methods yielded quite similar average values, the best method being FCFP10 with 77.1% sensitivity, 66.0% specificity, and 71.5% accuracy. For the positive definite graph kernels, the P-SVM classifier consistently giving slightly better average accuracies than the C-SVM classifier. The best value was in each case achieved by the marginalized graph kernel (C-SVM: 69.1 and P-SVM: 70.5%), the worst by the min/max Tanimoto kernel (C-SVM: 56.8 and P-SVM: 67.1%). In contrast, the proposed MCS

Table 2. Results of Method Comparison[a]

method	sensitivity	specificity	coverage	accuracy
1. MCASE (unambiguous)	41.5 ± 12.5%	82.3 ± 5.4%	73.2 ± 5.7%	61.9 ± 7.6%
2. MCASE (ambiguous)	43.2 ± 10.0%	76.7 ± 5.6%	97.7 ± 1.8%	59.9 ± 5.7%
3. PCA + KNN	60.5 ± 9.4%	75.2 ± 6.8%	99.7 ± 0.7%	67.9 ± 6.0%
4. PCA + C-SVM	73.1 ± 11.5%	67.0 ± 7.4%	99.7 ± 0.7%	70.0 ± 6.9%
5. PCA + random forest	50.4 ± 10.9%	84.6 ± 4.1%	99.7 ± 0.7%	67.5 ± 5.4%
6. FSV + KNN	54.1 ± 7.1%	74.6 ± 7.4%	99.7 ± 0.7%	64.3 ± 5.9%
7. FSV + C-SVM	69.2 ± 9.5%	68.7 ± 7.3%	99.7 ± 0.7%	69.0 ± 6.2%
8. FSV + random forest	56.7 ± 8.8%	82.1 ± 3.5%	99.7 ± 0.7%	69.4 ± 4.8%
9. P-SVM + KNN	55.3 ± 7.9%	77.4 ± 8.0%	99.7 ± 0.7%	66.4 ± 5.6%
10. P-SVM + C-SVM	68.7 ± 6.9%	71.1 ± 7.6%	99.7 ± 0.7%	69.9 ± 5.1%
11. P-SVM + random forest	54.6 ± 8.1%	83.8 ± 4.4%	99.7 ± 0.7%	69.2 ± 4.3%
12. PP Bayesian (ECFP10)	76.7 ± 6.6%	65.4 ± 6.2%	100.0 ± 0.0%	71.1 ± 4.4%
13. PP Bayesian (FCFP10)	77.1 ± 10.9%	66.0 ± 7.7%	100.0 ± 0.0%	71.5 ± 4.2%
14. PP Bayesian (SCFP6)	75.7 ± 7.4%	66.3 ± 4.9%	100.0 ± 0.0%	71.0 ± 5.2%
15. PP Bayesian (MDL public keys)	59.2 ± 4.3%	74.9 ± 7.3%	100.0 ± 0.0%	67.0 ± 3.9%
16. 3D spectrum GK + C-SVM	66.4 ± 7.1%	71.5 ± 6.4%	100.0 ± 0.0%	68.9 ± 4.0%
17. marginalized GK + C-SVM	68.8 ± 6.8%	69.5 ± 6.9%	100.0 ± 0.0%	69.1 ± 5.2%
18. spectrum GK + C-SVM	63.7 ± 7.6%	74.3 ± 5.1%	100.0 ± 0.0%	69.0 ± 3.9%
19. subtree GK + C-SVM	64.6 ± 7.0%	68.4 ± 4.7%	100.0 ± 0.0%	66.5 ± 4.9%
20. Tanimoto GK + C-SVM	55.4 ± 10.0%	78.9 ± 7.3%	100.0 ± 0.0%	67.2 ± 7.0%
21. min/max Tanimoto GK + C-SVM	49.2 ± 12.0%	64.3 ± 10.0%	100.0 ± 0.0%	56.8 ± 5.0%
22. λ-k GK + C-SVM	63.7 ± 7.6%	74.3 ± 5.1%	100.0 ± 0.0%	69.0 ± 3.9%
23. 3D spectrum GK + P-SVM	70.4 ± 5.6%	67.6 ± 6.5%	100.0 ± 0.0%	69.0 ± 3.1%
24. marginalized GK + P-SVM	71.3 ± 6.8%	69.8 ± 7.6%	100.0 ± 0.0%	70.5 ± 6.3%
25. spectrum GK + P-SVM	75.0 ± 7.8%	65.2 ± 5.4%	100.0 ± 0.0%	70.1 ± 4.1%
26. subtree GK + P-SVM	75.7 ± 6.5%	63.2 ± 4.9%	100.0 ± 0.0%	69.5 ± 4.2%
27. Tanimoto GK + P-SVM	69.6 ± 8.9%	70.1 ± 10.2%	100.0 ± 0.0%	69.8 ± 5.9%
28. min/max Tanimoto GK + P-SVM	69.1 ± 9.9%	65.0 ± 4.5%	100.0 ± 0.0%	67.1 ± 5.4%
29. λ-k GK + P-SVM	70.1 ± 4.1%	65.2 ± 5.4%	100.0 ± 0.0%	70.1 ± 4.1%
30. MCS kernel method (b)	94.9 ± 2.5%	83.8 ± 4.3%	100.0 ± 0.0%	89.3 ± 2.7%
31. MCS kernel method (u)	85.7 ± 4.6%	93.3 ± 2.8%	100.0 ± 0.0%	89.5 ± 2.3%

[a] Shown are the specificity, sensitivity, coverage and accuracy (average of sensitivity and specificity) of the different methods as mean values ± standard deviation. For the the MCS kernel method, (b) denotes the balanced version of the P-SVM, and (u) denotes the unbalanced one. GK denotes graph kernel.

kernel method achieved an average sensitivity of 94.9 (85.7%) and specificity of 83.8 (93.3%) using the balanced (unbalanced) P-SVM, corresponding to an average accuracy of 89.3 (89.5%).

The most relevant performance statistic for the prediction of the in vitro CA test outcome is the predictive accuracy, since it accounts for sensitivity and specificity by averaging them. In order to apply the 10-fold cross-validation paired *t*-test to test whether differences in accuracies are significant, the accuracies achieved on the individual test sets are needed, which are listed in Appendix 6.3. We used a one-sided test with a significance level of $\alpha = 0.05$. We tested for each of the 31 algorithms whether it had a significantly higher accuracy than each of the other methods. The results of the test are shown in Table 3, where an X marks the case where the row algorithm is significantly better than the column one. The results show that both the balanced as well as the unbalanced MCS kernel method perform significantly better than all other methods. Most other methods achieve significantly higher accuracy than MultiCASE and the min/max Tanimoto graph kernel with the C-SVM. Also FSV + KNN, P-SVM + KNN, the Bayesian classifier with MDL public keys, and the min/max Tanimoto GK with the P-SVM are significantly worse than many of the other methods. Note, however, that the best Bayesian classifier, the best positive semidefinite graph kernel methods with C-SVM and P-SVM, and the best descriptor vector-based method show no significant difference in performance. For the subtree and the min/max Tanimoto graph kernels, the P-SVM classifier performed significantly better than the C-SVM classifier.

A plausibility check was performed to assess if the visualization method highlighted molecular connectivities corresponding to substructures well-known for their potential to confer genotoxic effects. To this end, example compounds containing such substructures were identified in the data set using Derek for Windows. A compilation of examples for the visualization of the structure−activity information of the MCS kernel method is given in Figures 4−8. The models were always trained on the whole CA data set, leaving out the depicted molecule only. Shown are the compounds in their 3D conformation calculated by CORINA as ball-and-stick models. Red and blue atoms correspond to connectivities with positive and negative contribution, respectively, to an individual classification decision, while purple atoms are part of both. It is important to note that the visualization of relevant substructures reflects only part of the information that led to an individual classification. Rather, the prediction of CA test outcome is based on a topological comparison of entire structures, as described in Section 2.1.1. The visualization method provides a loading of the molecular connectivities contributing to an individual prediction and makes a prediction plausible to the user. In this sense, the visualization may help the toxicologist to interpret the prediction if a compound is classified as positive.

In Figures 4−6, almost all connectivities contributing to positive predictions represent well-known structural alerts

Table 3. Results of the 10-fold Cross-Validation Pairwise *t*-Test Used for Method Comparison[a]

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
1	−	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	X	0	0	0	0	0	0	0	0	0	0
2	0	−	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	X	X	−	0	0	X	0	0	0	0	0	0	0	0	0	0	0	0	0	0	X	0	0	0	0	0	0	0	0	0	0
4	X	X	0	−	0	X	0	0	X	0	0	0	0	0	0	0	0	0	0	0	X	0	0	0	0	0	0	0	0	0	0
5	X	X	0	0	−	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	X	0	0	0	0	0	0	0	0	0	0
6	0	X	0	0	0	−	0	0	0	0	0	0	0	0	0	0	0	0	0	0	X	0	0	0	0	0	0	0	0	0	0
7	X	X	0	0	0	X	−	0	0	0	0	0	0	0	0	0	0	0	0	0	X	0	0	0	0	0	0	0	0	0	0
8	X	X	0	0	0	X	0	−	X	0	0	0	0	0	X	0	0	0	X	0	X	0	0	0	0	0	0	0	0	0	0
9	0	X	0	0	0	0	0	0	−	0	0	0	0	0	0	0	0	0	0	0	X	0	0	0	0	0	0	0	0	0	0
10	X	X	0	0	0	X	0	0	X	−	0	0	0	0	0	0	0	0	0	0	X	0	0	0	0	0	0	0	0	0	0
11	X	X	0	0	0	X	0	0	0	0	−	0	0	0	0	0	0	0	0	0	X	0	0	0	0	0	0	0	0	0	0
12	X	X	X	0	0	X	0	0	X	0	0	−	0	0	X	0	0	X	X	X	X	X	0	0	0	0	0	X	0	0	0
13	X	X	X	0	X	X	0	0	X	0	0	0	−	0	X	0	0	0	X	X	X	0	0	0	0	0	0	X	0	0	0
14	X	X	0	0	0	X	0	0	X	0	0	0	0	−	X	0	0	0	X	0	X	0	0	0	0	0	0	X	0	0	0
15	X	X	0	0	0	0	0	0	0	0	0	0	0	0	−	0	0	0	0	0	X	0	0	0	0	0	0	0	0	0	0
16	X	X	0	0	0	X	0	0	0	0	0	0	0	0	0	−	0	0	0	0	X	0	0	0	0	0	0	0	0	0	0
17	X	X	0	0	0	X	0	0	0	0	0	0	0	0	0	0	−	0	0	0	X	0	0	0	0	0	0	0	0	0	0
18	X	X	0	0	0	X	0	0	0	0	0	0	0	0	0	0	0	−	X	0	X	0	0	0	0	0	0	0	0	0	0
19	X	X	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	−	0	X	0	0	0	0	0	0	0	0	0	0
20	0	X	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	−	X	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	−	0	0	0	0	0	0	0	0	0	0
22	X	X	0	0	0	X	0	0	0	0	0	0	0	0	0	0	0	0	X	0	X	−	0	0	0	0	0	0	0	0	0
23	X	X	0	0	0	X	0	0	0	0	0	0	0	0	X	0	0	0	0	0	X	0	−	0	0	0	0	0	0	0	0
24	X	X	0	0	0	X	0	0	0	0	0	0	0	0	0	0	0	0	X	0	X	0	0	−	0	0	0	X	0	0	0
25	X	X	0	0	0	X	0	0	X	0	0	0	0	0	X	0	0	0	X	0	X	0	0	0	−	0	0	X	0	0	0
26	X	X	0	0	0	X	0	0	X	0	0	0	0	0	0	0	0	0	X	0	X	0	0	0	0	−	0	0	0	0	0
27	X	X	0	0	0	X	0	0	0	0	0	0	0	0	0	0	0	0	X	0	X	0	0	0	0	0	−	X	0	0	0
28	X	X	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	X	0	0	0	0	0	0	−	0	0	0
29	X	X	0	0	0	X	0	0	X	0	0	0	0	0	X	0	0	0	X	0	X	0	0	0	0	0	0	X	−	0	0
30	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	−	0
31	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	0	−

[a] Both rows and columns represent the 31 methods (numbers following Table 2). The X's mark the cases where the row method has a significantly higher accuracy than the column method. The 0′s mark the case where the null hypothesis of equal accuracy could not be rejected (at level $\alpha = 0.05$).

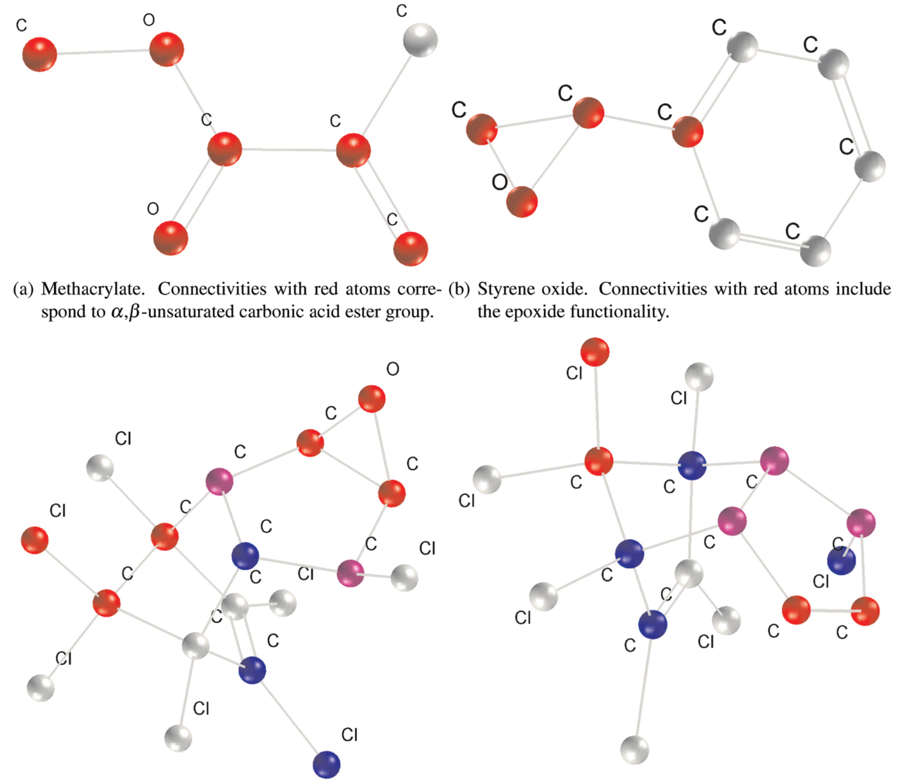
for DNA reactivity, reflecting a direct drug−DNA interaction (i.e., electrophilicity).[53] In agreement, most of these substructures are reflected as structural alerts in Derek for Windows, either linked to chromosome damage or to Ames mutagenicity. This is in line with an analysis showing that about 80% of the Ames positives also caused chromosome aberrations.[12]

The α,*β*-unsaturated carbonic esters, such as in Figure 4a, are capable of Michael addition to DNA.[54] The electron-withdrawing carbonic ester group may polarize the double bond, inducing a positive charge of the *β*-carbon, enabling it to add to an electron-rich species. Epoxides, as in Figure 4b and c[55] or alkyl halides, such as in Figure 4d and c[56] are electrophilic compounds capable of direct DNA alkylation. Primary aromatic amines, such as in Figure 5a and b, can be metabolized involving CYP 1A2-mediated *N*-hydroxylation and *O*-esterification,[57] followed by cleavage to form a reactive nitrenium ion which is capable of binding to DNA. Similar to aromatic amines, aromatic amides, as in Figure 5c, may exhibit mutagenic activity in the presence of S9 mix.[58] Deacylation of the amide may take place prior to *N*-hydroxylation. Phenylhydrazines, which result from hydrolytic cleavage of the respective phenylhydrazides (Figure 6a), are capable of inducing oxidative DNA damage and DNA−adduct formation.[59] *N*-nitro and *N*-nitroso compounds, as in Figure 6b, may be decomposed hydrolytically to form a diazoalkane. Diazomethane results from the hydrolysis of nitrosoguanidine,[60] capable of alkylating DNA[61] and the active agent in nitrosoguanidine mutagenesis.[62] For alkyl nitrites as in Figure 6c mutagenic activity may result from the nitrite ion formed by hydrolysis, from direct alkylation or nitrosation of DNA, or from the production of alkylating agents following nitrosation of other amines or amides.[63−65] Aryl *N*-alkylcarbamate esters as in Figure 6d may be hydrolyzed to release isocyanate intermediates,[66,67] which may in turn contribute to the observed clastogenic activity.[68]

Many of the true positive compounds contained electrophilic substructures that were implemented as structural alerts in Derek for Windows (e.g., aromatic amine/amide, aromatic nitro, *N*-nitroso, and halogenated alkanes). Nevertheless, Derek did not yield any structural alerts for mutagenicity or chromosome damage for some of these compounds (e.g., Figure 4a and b), despite the presence of such a substructure. These compounds were not covered by the respective Derek alerts due to alert refinements that have been set up by the knowledge base department at Lhasa Limited to increase specificity, however apparently at the expense of sensitivity.

Apart from these molecules directly interacting with DNA, the MCS kernel method yielded correct predictions for molecules that may cause chromosome aberrations, at least additionally, via other mechanisms (Figures 7 and 8).

The connectivity with positive contribution to classification in Figure 7a contains a purine substructure. Azathioprine is a purine synthesis inhibitor leading to interference with DNA synthesis and chromosome breaks.[69] The aromatic nitro substructure may additionally contribute to induction of chromosomal abnormalities. In Figure 7b, the connectivity with positive contribution to classification includes part of



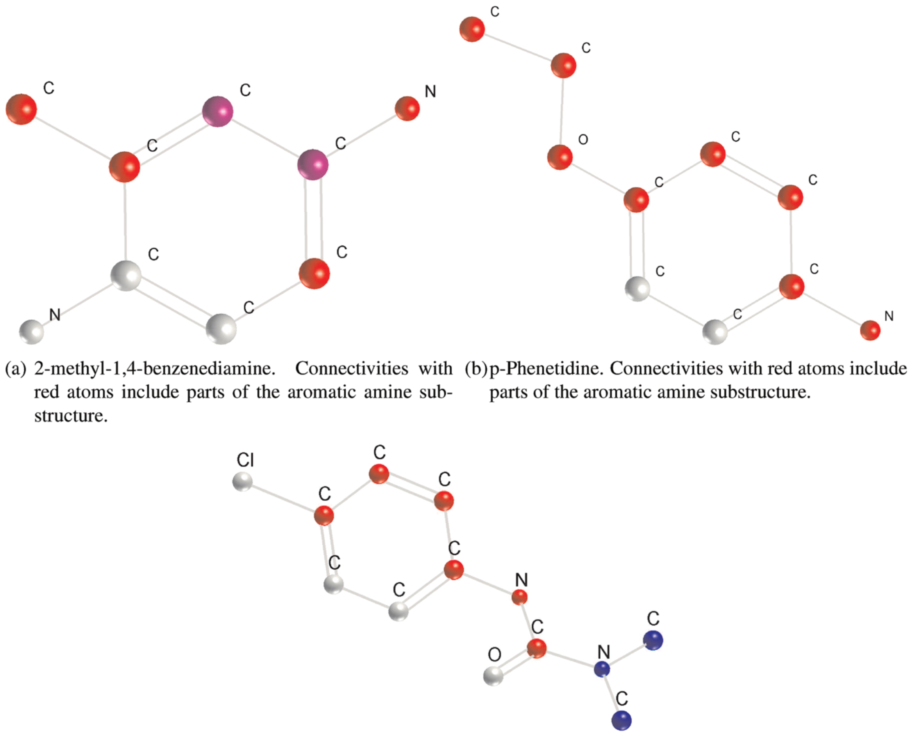
(a) Methacrylate. Connectivities with red atoms correspond to α,β-unsaturated carbonic acid ester group.

(b) Styrene oxide. Connectivities with red atoms include the epoxide functionality.

(c) Epoxyheptachlor. Connectivities with red atoms include parts of the halogenated alkanes and the epoxide substructure.

(d) Heptachlor. Connectivities with red atoms include parts of the halogenated alkane substructures.

Figure 4. Examples for SAR information obtained by the MCS kernel method. All four structures were correctly classified as positive. Red atoms are connectivities with contribution to an individual positive prediction. Structural alerts for chromosome damage (a, c, and d) and Ames mutagenicity (b−d) were raised in Derek for Windows based on the same respective substructures.

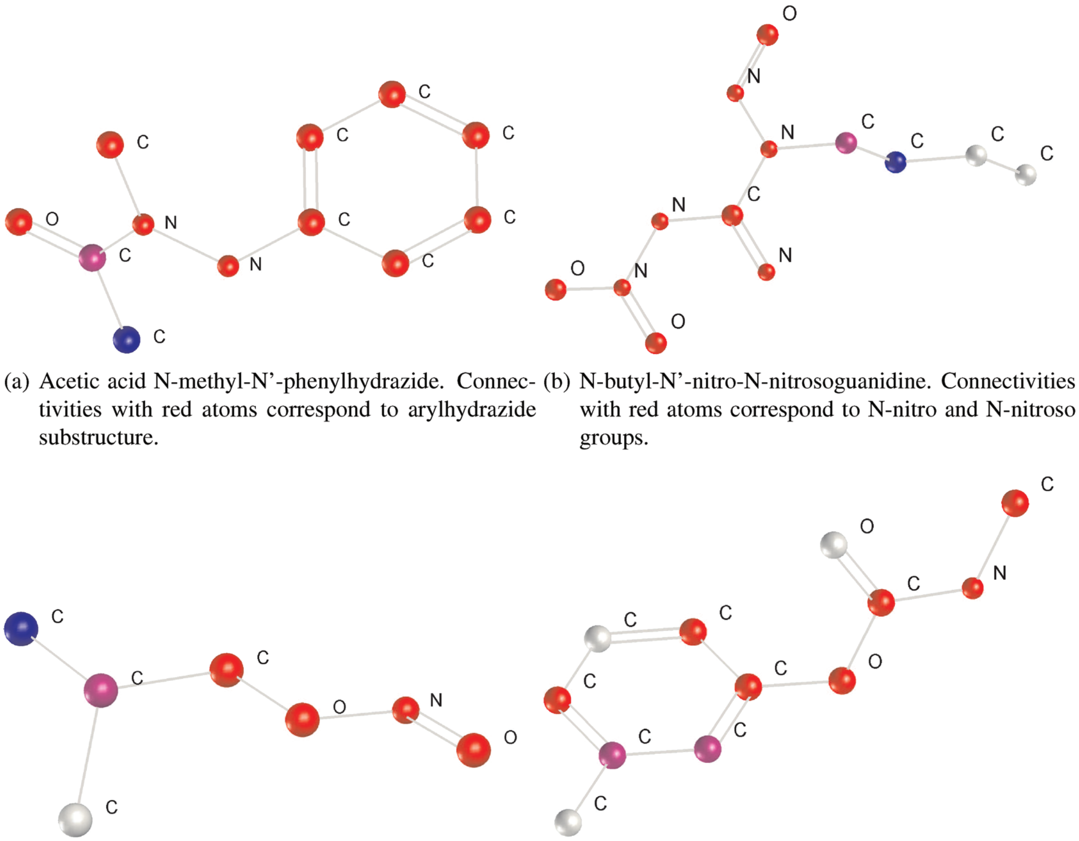


(a) 2-methyl-1,4-benzenediamine. Connectivities with red atoms include parts of the aromatic amine substructure.

(b) p-Phenetidine. Connectivities with red atoms include parts of the aromatic amine substructure.

(c) Monuron. Connectivities with red atoms include parts of the aromatic amide substructure.

Figure 5. Examples for SAR information obtained by the MCS kernel method. All three structures were correctly classified as positive. Red atoms are connectivities with contribution to an individual positive prediction. A structural alert for Ames mutagenicity (a) was raised in Derek for Windows based on the same substructure. (b and c) No structural alerts were raised in Derek for Windows.



(a) Acetic acid N-methyl-N'-phenylhydrazide. Connectivities with red atoms correspond to arylhydrazide substructure.

(b) N-butyl-N'-nitro-N-nitrosoguanidine. Connectivities with red atoms correspond to N-nitro and N-nitroso groups.

(c) Isobutyl nitrite. Connectivities with red atoms correspond to alkyl nitrite substructure.

(d) Metolcarb. Connectivities with red atoms include parts of the aryl N-alkylcarbamate substructure.

Figure 6. Examples for SAR information obtained by the MCS kernel method. All four structures were correctly classified as positive. Red atoms are connectivities with contribution to an individual positive prediction. Structural alerts for chromosome damage (b−d) and Ames mutagenicity (a−c) were raised in Derek for Windows based on the same respective substructures.

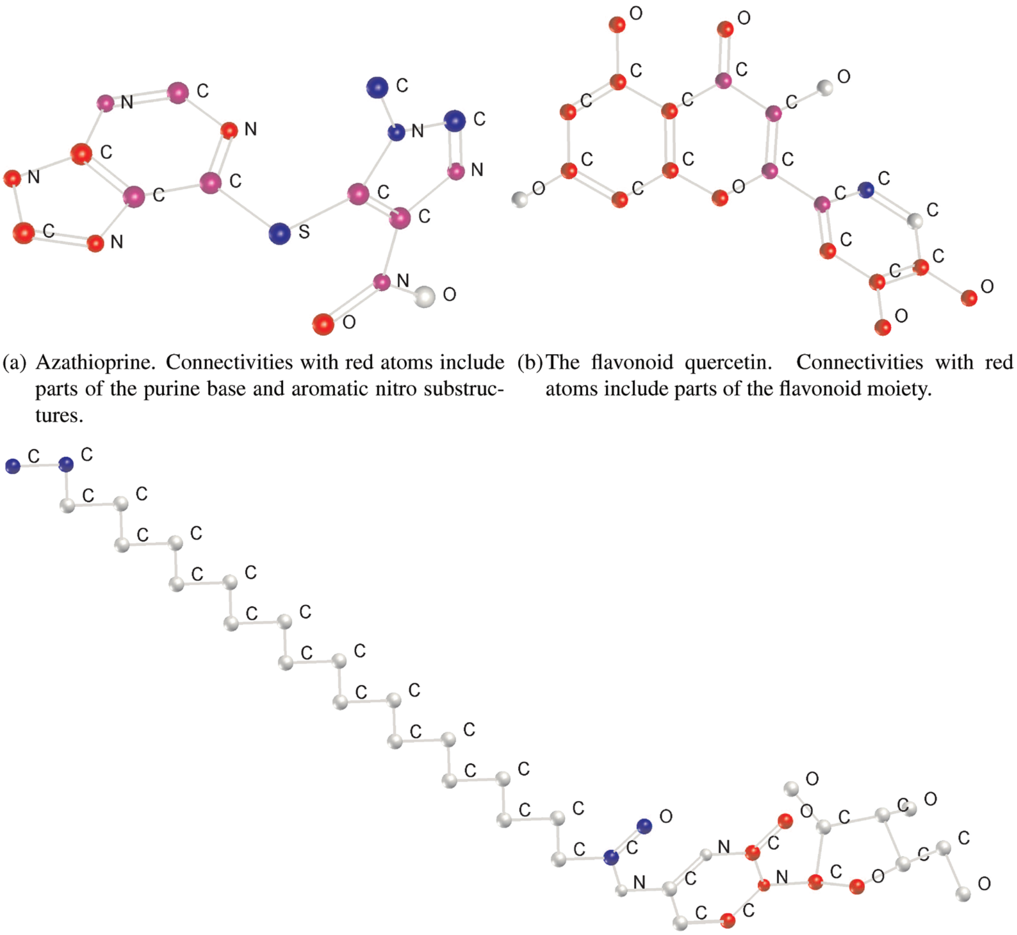
the flavonoid structure, which is related to clastogenicity via several proposed mechanisms (both direct and indirect DNA interaction).[70] The connectivities with positive contribution to classification in Figures 7c and 8 include part of the cytosine moieties. Both enocitabine[71] and cytarabine[72] (Figures 7c and 8, respectively) act as antimetabolites inducing the premature termination of growing DNA chains. In addition, they inhibit several enzymes involved in the synthesis of DNA precursors, such as DNA polymerases and nucleotide reductase enzymes, leading to impairment of DNA synthesis.

4. DISCUSSION

We compared the prediction accuracy of the proposed MCS kernel method to MultiCASE, the Pipeline Pilot Bayesian classifier, several positive semidefinite graph kernel methods, and machine learning methods applied to DragonX descriptor vectors on a large in vitro CA data set we compiled from various public sources. In this comparison, the MCS kernel method achieved about 90% accuracy, where the version using the balanced P-SVM had higher sensitivity but lower specificity than the unbalanced one. This could be expected, since there are more negative examples in the data set than positive ones. The accuracies, however, are almost identical, and both are significantly higher than those achieved by all the other methods included in the comparison, the best of which achieved an average accuracy of about 72%.

As shown in Table 3, the MCS kernel method also significantly outperformed the other graph kernel methods. This boost in performance cannot be attributed to the used classifier alone, because using the P-SVM instead of a C-SVM on the positive semidefinite graph kernels gave only a very small increase in performance (although the differences were significant for two of the seven graph kernels). Instead, we assume the following reason: Positive definite graph kernels that include the full structural information, like the MCS kernel does, are computationally inefficient. In contrast to the MCS kernel method, one cannot simply approximate positive definite graph kernels for the SVM, since their approximate solutions may yield indefinite kernel matrices. Enforcing both, positive definiteness and computationally efficiency, may result in graph kernels that systematically lose structural information relevant for predicting the CA test. More generally, any constraint like positive definiteness that is unrelated to the given problem at hand restricts the class of possible similarity measures. Hence, for some problems, one is forced to choose an actually inappropriate similarity measure in order to meet the required constraints. We therefore assume that in contrast to the positive definite graph kernel methods considered here, the approximated MCS kernel still retains the relevant structural information.

In a previous work based on data collections from Snyder et al.[36] and Kirkland et al.[37]—both of which are contained in the present in vitro CA test data set—a machine learning



(a) Azathioprine. Connectivities with red atoms include parts of the purine base and aromatic nitro substructures.

(b) The flavonoid quercetin. Connectivities with red atoms include parts of the flavonoid moiety.

(c) Enocitabine. Connectivities with red atoms include parts of the cytosine substructure.

Figure 7. Examples for SAR information obtained by the MCS kernel method. All three structures were correctly classified as positive. Red atoms are connectivities with contribution to an individual positive prediction. Structural alerts for chromosome damage (a−c) and Ames mutagenicity (a and b) were raised in Derek for Windows that are based on the same substructures.

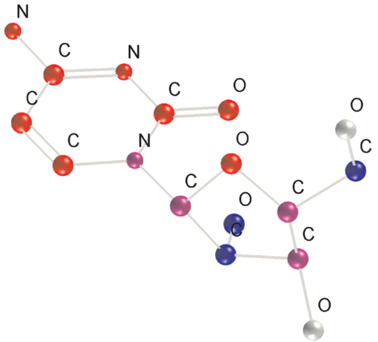


Figure 8. Cytarabine, correctly classified as positive. Red atoms are connectivities with red atoms identified by the MCS kernel method, which includes parts of the thymine substructure. Derek for Windows raised a structural alert for chromosome damage based on the same substructure.

model was developed reaching good sensitivity and specificity values, however, lacking SAR information.[13] The MCS kernel method applied in this work, provides, in addition to a dramatically increased predictive performance, an informative model. We showed how information on the relative loadings of the molecular connectivities contributing to an individual prediction can be obtained by visualizing groups of atoms favoring a positive or negative classification. We illustrated this by showing examples of substructures identified by the method known to be involved in different mechanisms causing chromosome aberration.

For its application in a regulatory environment, e.g., under the European Union's REACH legislation, a (Q)SAR model should be scientifically valid.[73] In their guidance for the implementation of REACH, the European Chemicals Agency (EChA) proposes to assess the model validity by reference to the OECD principles for the validation of (Q)SARs.[74] These comprise the need for a defined end point, an unambiguous algorithm, a defined domain of applicability, appropriate measures of goodness-of-fit, robustness, and predictivity as well as a mechanistic interpretation, if possible.[14] In this paper we provide a well-defined data set, the underlying modeling algorithm, and a 10-fold cross validation providing prediction accuracy values with standard deviations. For a specific application, like the hazard assessment of chemicals in REACH, an external evaluation covering the target chemical space is recommended. Moreover, the MCS kernel approach visually relates the contribution of a certain bond to the classification decision of the present structure, thereby providing structure−activity information that can add plausibility to an individual prediction from a chemistry perspective. The inclusion of a criterion indicating the domain of applicability and the confidence of the model in its prediction, which are recommended for the application of the model in a regulatory environment, will be the topic of future research. These features will help the user to assess the reliability of an individual prediction, which is a more meaningful diagnostic parameter for the user's present analysis than a high accuracy reached in a cross-validation setting. From a drug discovery and development perspective, the MCS kernel model for the prediction of the in vitro CA test needs to be evaluated externally on an independent data set covering the user company's chemical space. Although pharmaceuticals are generally represented in the training data set, a retraining including in-house data may be necessary prior to application in order to reach appropriate prediction accuracies. The visualization of individual SAR information is considered of great value to lead optimization, providing the toxicologists and chemists with relevant information for a path forward.

The visualization step led to a correct yet incomplete coloration of relevant groups linked to aromatic systems or occurring more than once in a molecule (e.g., Figure 5 a−c). One reason could be the fact that a substructure involving only parts of the ring might be already sufficiently indicative for an individual class assignment. In many compounds, functional groups associated with chromosome aberration appear multiple times, e.g., Figure 4c and d. However, often only one or a few of these are marked as contributors to classification. If instead a certain number of appearances contributed to classification, the method was able to detect this and to mark multiple groups.

Interestingly, it has been presumed in the literature[12] that the low performance of hitherto existing structure-based approaches to CA test prediction was due to the limitations of the used 2D structural fragments, which were not sufficient to describe the multitude of mechanisms leading to a positive result in the CA test. In contrast, our results obtained with the MCS kernel method indicate that approaches based on a topological comparison of 2D information only can reach high accuracy in the prediction of the CA test.

In summary, our comprehensive methodological analysis strikingly showed that the MCS kernel method combines the expressiveness of substructure matching with the generalization properties of support vector machines. This is made possible through the use of the P-SVM, which allows to extend the successful support vector approach to the case of relational (dyadic) data. By working with kernels on structures rather than descriptor vectors, it not only achieves much better performance but also retains the interpretability of structure-based approaches.

It should be marked that on top of this, a confidence estimate as well as a matching with the applicability domain of the model would help the user assess the reliability of an individual prediction which cannot be deduced from the cross-validation results. Yet, the 20% higher prediction accuracy compared to other methods was shown to be statistically significant and demands further exploration of the advantages of the MCS kernel method.

5. CONCLUSIONS AND OUTLOOK

In this work, we proposed the maximum common subgraph (MCS) kernel method for building informative prediction models of the outcome of the in vitro chromosome aberration (CA) test. Our contributions were: (i) the use of the indefinite MCS kernel for chemical graphs in conjunction with the potential support vector machine (P-SVM) for model building, (ii) the extraction and visualization of the individual SAR information implicitly contained in the prediction model, (iii) the provision of a model for predicting the in vitro CA test yielding both, a very high prediction accuracy and SAR information, thereby meeting the needs of both structure optimization and application in a regulatory environment, and (iv) the compilation of a large, publicly available in vitro CA test data set, together with all the information required for comparing methods and testing the significance of differences in their estimated generalization performance.

The MCS kernel method unites many advantages of different approaches for predicting the in vitro CA test outcome. Both Derek for Windows and MultiCASE provide structure−activity information, Derek for Windows mostly even with mechanistic interpretation but do not achieve high accuracy. The structural fragments created by MultiCASE are limited to a maximum size of 10 atoms, so that larger potentially discriminating substructures will not be found. Derek for Windows, as a rule-based expert system, is not capable of inductive learning from data. Specifically for assessing chromosomal damage, it demonstrated limited reliability and predictivity for chromosome aberration.[9] It can be assumed that there are unknown SARs which are not contained in the Derek for Windows knowledge database. Moreover, the rules contained in it, although some taking, e.g., inter-relationships between different structural features or physicochemical properties into account, may in part be too generic to reflect the influence of the chemical neighborhood of a functional group on their clastogenic activity. Nevertheless, Derek for Windows and MultiCASE are still essential for drug discovery and development as they provide structure−activity and/or mechanistic information essential for structure optimization and are widely accepted by regulatory authorities. In addition, MultiCASE provides the user with individual estimates of the prediction confidence which includes, e.g., an evaluation of the reliability of individual biophores and a search for structural fragments in the test structure not contained in the training set of the model (match with applicability domain). This helps the user evaluate the reliability of an individual prediction. Pipeline Pilot Bayesian classifiers, positive definite graph kernel, and descriptor-based machine learning methods achieve better accuracies, however their results were still significantly worse then those obtained with the proposed MCS kernel method. Moreover, they only provide black-box models, which are hardly interpretable and therefore may not fully meet the OECD principles for QSAR validation.

In contrast to the rule-based expert system Derek for Windows and the commercial MultiCASE modules for in vitro CA test prediction, the MCS kernel method makes use of inductive learning for model building and is therefore able to adapt itself to new data without supervision. Unlike the Pipeline Pilot Bayesian classifiers and the descriptor-based machine learning methods, it provides information on the molecular connectivities contributing to an individual prediction. Last but not least, it achieved a 20% higher accuracy than even the black-box models.

It should be emphasized that the P-SVM, a kernel method designed for relational data, can be validly applied to kernel matrices that are not positive semidefinite. Such kernels arise from many natural similarity measures[26,27] and pose conceptual problems for conventional SVMs, as the nonfulfillment of positive definiteness violates their theoretical underpinnings. The fact that the P-SVM is not constrained to positive semidefinite kernels allows a valid use of a much wider set of kernels, without resorting to heuristics. Also the proposed visualization scheme can be adapted to other kernels and provide informative models.

From an industry perspective, the MCS model for prediction of the in vitro CA test outcome apparently outperforms both commercial and present in-house solutions as it demonstrated—at least in a cross-validation setting—a very high prediction accuracy and yields structural information meeting the needs of both structure optimization and application in a regulatory environment. Routine application requires prior evaluation of the model using a set of in-house in vitro CA test data to assess the prediction accuracy in the target chemical space. This also applies, besides the determination of the applicability domain and the confidence of predictions, to validation of the model following the OECD principles.

6. APPENDIX

6.1. Potential Support Vector Machine (P-SVM). The P-SVM[29−31] selects models using the principle of structural risk minimization. In contrast to standard SVM approaches, however, the P-SVM is based on a new objective function and a new set of constraints which lead to an expansion of the classification or regression function in terms of "support features". The optimization problem is quadratic, always well-defined, suited for relational (dyadic) data and neither requires square nor positive semidefinite kernel matrices. Therefore, the method can also be used without preprocessing with matrices which are measured and with matrices which are constructed from a vectorial representation using an indefinite kernel function.

In this work, we employ the P-SVM for two tasks: (i) in "dyadic" mode, as a classifier working on the potentially indefinite MCS kernel matrix, and (ii) in "vectorial" mode, as a filter method for feature selection on the descriptor vectors obtained by DragonX. In the experiments, we used the implementation of the P-SVM by Knebel et al.[31,75] In the following, we will briefly outline the mathematical formulation of the P-SVM in these different modes.

6.1.1. P-SVM for Relational Data. In the MCS kernel method, we make use of the fact that the P-SVM can handle potentially indefinite kernel matrices, like the proposed MCS kernel, which describes the set of molecules as relational data, via their pairwise similarity. For training, the MCS kernel matrix **K** is needed, which contains the MCS kernel evaluated for all pairs of molecules in the training set. The prediction function is based on evaluating the kernel between the test molecules and a set of "support molecules", a subset of the training data set.

We consider a two class classification task, where we have a kernel matrix $\mathbf{K} = k(X_p, X_q) \forall p, q \in [1, ..., m]$ evaluating the kernel for all pairs of objects in the training data set. The class labels are summarized in the vector **y**. It is assumed that the kernel matrix is standardized to zero mean and variance one. This can always be enforced by subtracting the mean and dividing by the standard deviation. Note that this standardization of the kernel matrix has to be accounted for when extracting the SAR information from the model (see Section 2.1.1).

The (dual) optimization problem of the P-SVM can be derived as

$$\begin{aligned} \min_{\boldsymbol{\alpha}^+,\boldsymbol{\alpha}^-} \quad & \frac{1}{2}(\boldsymbol{\alpha}^+ - \boldsymbol{\alpha}^-)^{\mathrm{T}}\mathbf{K}^{\mathrm{T}}\mathbf{K}(\boldsymbol{\alpha}^+ - \boldsymbol{\alpha}^-) \\ & - \mathbf{y}^{\mathrm{T}}\mathbf{K}(\boldsymbol{\alpha}^+ - \boldsymbol{\alpha}^-) + \varepsilon\mathbf{1}^{\mathrm{T}}(\boldsymbol{\alpha}^+ + \boldsymbol{\alpha}^-) \\ \text{s.t.} \quad & \mathbf{0} \le \boldsymbol{\alpha}^+ \le C\mathbf{1}, \quad \mathbf{0} \le \boldsymbol{\alpha}^- \le C\mathbf{1} \end{aligned} \tag{16}$$

where $\boldsymbol{\alpha} = (\boldsymbol{\alpha}^+ - \boldsymbol{\alpha}^-)$ denote the Lagrange multipliers for the constraints, and C and ε are regularization parameters.

The dual optimization problem only depends on **K** via $\mathbf{K}^{\mathrm{T}}$ **K**. Therefore, **K** is neither required to be positive semidefinite nor to be square. This allows to extend SVM-based approaches to the new class of indefinite kernel matrices. The set of nonzero α_i, $i \in S$ mark the support molecules X_i, $i \in S$. For a high enough value of ε, a sparse solution will be enforced, where the number of support molecules is much smaller than the size of the training set. If $\mathbf{K}^{\mathrm{T}}\mathbf{K}$ is singular, ε enforces a unique solution. The hyperparameters C and ε need to be adjusted via cross-validation on the training set (see Appendix 6.2).

Equation 16 can be solved efficiently using a new sequential minimal optimization (SMO) technique.[31] The resulting classifier is then given by

$$f(X) = \mathrm{sgn}\left(\sum_{q=1}^{m} \alpha_q k(X, X_q) + b\right) \tag{17}$$

where

$$b = -\frac{1}{m}\sum_{i=1}^{m} y_i \tag{18}$$

Only the kernels $k(X, X_q)$ for the support molecules need to be evaluated for prediction.

Often, training data sets are unbalanced, i.e., one class contains more examples than the other class. In such cases, the P-SVM in the standard *unbalanced mode* will chose a classification boundary favoring the larger class, since this reduces the overall empirical error. If the positive class is much larger than the negative one, then this leads to an increased sensitivity at the cost of a reduction in specificity. If the negative class is larger, then it will be the other way around. However, if used in *balanced mode*, the P-SVM accounts for the unbalancedness of the training data set by multiplying the class label of the smaller class by a factor proportional to the ratio of the class sizes. This procedure will shift the class boundary in the direction of the smaller class, leading to more balanced values of sensitivity and specificity even for highly unbalanced classes.

6.1.2. P-SVM Feature Selection on Descriptor Vectors. In the descriptor-based machine learning approaches, the P-SVM is one of the methods employed for feature selection. For this purpose the P-SVM is used in feature selection mode on vectorial data to select the "support features", a subset of the descriptors that will be used for classification.

If the m (d-dimensional) descriptor vectors are written as columns of a data matrix $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_m)$, the dual

Table 4. Accuracies (Average of Sensitivity and Specificity) on the Individual Cross-Validation Folds

	set									
method	1	2	3	4	5	6	7	8	9	10
MCASE (unamb.)	65.1%	71.4%	58.5%	59.9%	67.5%	59.6%	65.9%	48.9%	50.4%	71.9%
MCASE (ambig.)	59.8%	66.6%	63.0%	61.5%	63.9%	57.6%	61.7%	46.8%	53.3%	65.5%
PCA + KNN	61.4%	72.9%	71.0%	70.9%	75.1%	73.6%	65.2%	62.4%	57.2%	68.8%
PCA + C-SVM	75.8%	79.8%	75.2%	71.9%	77.6%	66.8%	64.1%	66.4%	59.4%	63.4%
PCA + rand. for.	60.7%	67.2%	67.8%	73.1%	71.7%	69.8%	72.0%	66.0%	56.0%	70.5%
FSV + KNN	66.1%	72.0%	67.9%	70.1%	70.8%	63.8%	56.6%	58.5%	60.8%	57.0%
FSV + C-SVM	77.3%	72.2%	72.8%	72.1%	76.8%	64.0%	62.0%	63.7%	68.7%	60.1%
FSV + rand. for.	72.6%	75.3%	74.3%	69.8%	65.7%	73.4%	69.5%	59.9%	65.7%	68.0%
P-SVM + KNN	65.8%	75.1%	69.4%	65.5%	66.8%	71.3%	71.0%	61.7%	59.8%	57.0%
P-SVM + C-SVM	72.4%	76.2%	75.2%	75.0%	73.3%	67.2%	66.1%	65.0%	67.3%	61.5%
P-SVM + rand. for.	67.3%	74.9%	69.5%	71.5%	77.5%	66.5%	63.6%	65.7%	66.1%	69.5%
BC (ECFP10)	71.5%	75.2%	72.8%	74.9%	73.7%	76.0%	67.6%	62.4%	69.9%	66.7%
BC (FCFP10)	71.2%	78.1%	68.8%	69.7%	73.7%	75.3%	70.3%	73.3%	62.5%	72.3%
BC (SCFP6)	70.5%	76.6%	68.8%	63.7%	76.7%	78.3%	68.0%	63.7%	73.0%	70.5%
BC (MDL publ. keys)	70.1%	70.3%	69.5%	68.5%	61.3%	69.0%	61.1%	62.0%	69.6%	68.8%
3D spectrum GK + C-SVM	67.1%	69.9%	68.0%	74.5%	68.9%	59.8%	67.3%	67.9%	72.2%	73.9%
marginalized GK + C-SVM	68.4%	72.4%	65.6%	73.9%	78.7%	69.4%	58.7%	65.9%	71.7%	66.8%
spectrum GK + C-SVM	64.6%	73.5%	69.6%	72.0%	75.8%	72.2%	65.6%	63.8%	67.0%	66.3%
subtree GK + C-SVM	61.5%	74.5%	63.2%	71.1%	68.2%	71.8%	65.9%	57.2%	65.0%	67.0%
Tanimoto GK + C-SVM	62.4%	74.9%	67.9%	81.1%	66.6%	74.7%	59.7%	59.4%	61.7%	63.5%
min/max Tanimoto GK + C-SVM	54.1%	58.8%	59.3%	53.4%	66.9%	51.0%	53.6%	59.7%	50.2%	61.0%
λ-k GK + C-SVM	64.6%	73.5%	69.6%	72.0%	75.8%	72.2%	65.6%	63.8%	67.0%	66.3%
3D spectrum GK + P-SVM	66.2%	69.3%	68.0%	72.2%	66.3%	70.8%	66.5%	64.2%	73.1%	73.7%
marginalized GK + P-SVM	70.9%	73.4%	68.0%	71.2%	78.7%	76.6%	58.7%	62.8%	78.4%	67.0%
spectrum GK + P-SVM	67.8%	72.8%	68.0%	72.0%	75.7%	74.3%	67.1%	61.5%	73.9%	68.5%
subtree GK + P-SVM	64.1%	72.8%	67.2%	72.2%	73.4%	76.0%	68.8%	61.9%	71.2%	67.3%
Tanimoto GK + P-SVM	60.6%	76.2%	67.9%	74.9%	70.4%	74.7%	64.3%	60.2%	74.0%	75.4%
min/max Tanimoto GK + P-SVM	60.9%	68.4%	63.9%	70.3%	75.3%	69.7%	59.3%	61.0%	67.7%	74.5%
λ-k GK + P-SVM	67.8%	72.8%	68.0%	72.0%	75.7%	74.3%	67.1%	61.5%	73.9%	68.5%
MCS kernel method (b)	87.9%	91.4%	89.7%	88.9%	94.3%	85.4%	92.0%	91.1%	85.8%	87.0%
MCS kernel method (u)	87.2%	90.3%	91.2%	92.7%	89.6%	86.4%	86.6%	93.4%	89.4%	88.2%

optimization problem of the P-SVM for feature selection is obtained as

$$\begin{aligned} \min_{\boldsymbol{\alpha}^+,\boldsymbol{\alpha}^-} \quad & \frac{1}{2}(\boldsymbol{\alpha}^+ - \boldsymbol{\alpha}^-)^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}(\boldsymbol{\alpha}^+ - \boldsymbol{\alpha}^-) \\ & - \mathbf{y}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}(\boldsymbol{\alpha}^+ - \boldsymbol{\alpha}^-) + \varepsilon\mathbf{1}^{\mathrm{T}}(\boldsymbol{\alpha}^+ + \boldsymbol{\alpha}^-) \\ \text{s.t.} \quad & \mathbf{0} \le \boldsymbol{\alpha}^+, \quad \mathbf{0} \le \boldsymbol{\alpha}^- \end{aligned} \tag{19}$$

Again, the dual problem is solved by SMO. Here, the nonzero components α mark the support features, i.e., those descriptor variables which are chosen by the feature selection and used in a subsequent classification algorithm. Note that there is only one hyperparameter (ε). If $\mathbf{XX}^{\mathrm{T}}$ is singular and $\mathbf{w}$ is not uniquely determined, then ε enforces a unique solution. The value of ε implicitly controls the size of the set of support features. Increasing ε increases the sparsity of the solution, which means that fewer features will be selected.

6.2. Hyperparameter Optimization. *6.2.1. Descriptor-Based Machine Learning Methods.* An inner 10-fold cross-validation loop was conducted for each training set of the cross-validation. For each training set, the parameter combination yielding the best predictive accuracy (average of sensitivity and specificity) was then used to train the model. The machine learning methods always consisted of one feature selection/construction method and one classifier, each of which had always one hyperparameter. Therefore, a 2D grid search in parameter space was carried out for all nine methods. In the following, the parameter values along the respective grid axis are given.

Feature Selection/Construction Methods:

- PCA: number of components $H \in [5, 10, 15,\ 30, 40, 50, ..., 200]$
- FSV: number of features $V \in [5, 10, 15, 30, 40, 50, ..., 200]$
- P-SVM feature selection: ε: $[2^{-11}, 2^{-9}, 2^{-7}, 2^{-5}, 2^{-3}, 2^{-2}, 2^{-1}, 2^{0.5}, 2^{1}, 2^{1.5}, 2^{2}]$

Classifiers:

- KNN: number of neighbors $K \in [1,\ 3,\ 5,\ 7]$
- C-SVM: $C \in [2^{-15}, 2^{-13}, 2^{-11}, 2^{-9}, 2^{-7}, 2^{-5}, 2^{-3}, 2^{-2}, 2^{-1}, 2^{1}, 2^{2}, 2^{3}]$
- Random forest: number of decision trees $K \in [5, 10, 15, ..., 40, 60, 80, ..., 160]$

6.2.2. Maximum Common Subgraph Kernel Method. Again, an inner 10-fold cross-validation loop was carried out over a 2D search grid exploring the hyperparameter space of the P-SVM classifier. The following values were used for the hyperparameters ε and C:

- $C \in [1, 1.1, 1.2, ..., 2.9, 3]$
- $\varepsilon \in [0.05,\ 0.1,\ 0.15, ..., 0.45, 0.5]$

6.2.3. Positive Semidefinite Graph Kernel Methods. Both P-SVM and C-SVM were used as classifiers in combination with the positive semidefinite graph kernels. The hyperparameters of the P-SVM were optimized exactly as described in Section 6.2. For the C-SVM, the only hyperparameter was C, which was optimized using a 10-fold cross-validation loop on each training set over the following values:

- $C \in [2^{-10},\ 2^{-9}, ...,\ 2^{9},\ 2^{10}]$

6.3. Results on the Individual CV Folds. In Table 4 we report the accuracy (defined as the average of sensitivity and specificity) the different methods achieved on the individual folds of the 10-fold cross-validation. These values are required for assessing whether one method performed significantly better than another one, via the 10-fold cross validation paired *t*-test. These results also allow future comparison with other methods if exactly the same cross-validation sets are used, which we make publicly available. For the MCS kernel method, (b) denotes the balanced version of the P-SVM, (u) the unbalanced one.

ACKNOWLEDGMENT

The authors thank Florian Kirchner for helping with the evaluation of the descriptor-based methods and Lhasa Limited for the provision of high-quality, peer-reviewed CA test data through VITIC.

Supporting Information Available: The compiled CA test data set and the used cross-validation splits are provided for benchmarking purposes. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) *OECD Guidelines for the Testing of Chemicals, Test Guideline No. 473: In vitro Mammalian Chromosome Aberration Test*; Organization for Economic Cooperation and Development: Paris, France, 1997; Vol. 1 (4), pp 1−10.

(2) Muster, W.; Breidenbach, A.; Fischer, H.; Kirchner, S.; Müller, L.; Pähler, A. Computational toxicology in drug development. *Drug Discovery Today* **2008**, *13* (7/8), 303–310.

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

(4) Degrassi, F.; Fiore, M.; Palitti, F. Chromosomal aberrations and genomic instability induced by topoisomerase-targeted antitumour drugs. *Curr. Med. Chem.: Anti-Cancer Agents* **2004**, *4*, 317–325.

(5) Scott, D.; Galloway, S. M.; Marshall, R. R.; Ishidate, M., Jr.; Brusick, D.; Ashby, J.; Myhr, B. C. Genotoxicity under Extreme Culture Conditions. A report from ICPEMC Task Group 9. *Mutat. Res.* **1991**, *257*, 147–204.

(6) Marchant, C. A.; Briggs, Katharine A.; Long, A. In Silico Tools for Sharing Data and Knowledge on Toxicity and Metabolism: Derek for Windows, Meteor, and Vitic. *Toxicol. Mech. Methods* **2008**, *18* (2), 177–187.

(7) Klopman, G. MULTICASE 1. A Hierarchical Computer Automated Structure Evaluation Program. *Quant. Struct.-Act. Relat.* **1992**, *11* (2), 176–184.

(8) Genetic Toxicity; MultiCASE: Beachwood, OH; http://www.multicase.com/products/prod0910.htm. Accessed July 1, 2009.

(9) In-house experience from Bayer Schering Pharma AG and the results section in the present paper.

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

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

(12) Mekenyan, O.; Todorov, M.; Serafimova, R.; Stoeva, S.; Aptula, A.; Finking, R.; Jacob, E. Identifying the structural requirements for chromosomal aberration by incorporating molecular flexibility and metabolic activation of chemicals. *Chem. Res. Toxicol.* **2007**, *20*, 1927–41.

(13) Rothfuss, A.; Steger-Hartmann, T.; Heinrich, N.; Wichard, J. Computational prediction of the chromosome-damaging potential of chemicals. *Chem. Res. Toxicol.* **2006**, *19*, 1313–1319.

(14) *OECD Principles for the Validation, For Regulatory Purposes, of (Quantitative) Structure-Activity Relationship Models*; Organization for Economic Cooperation and Development: Paris, France; http://www.oecd.org/dataoecd/33/37/37849783.pdf. Accessed July 1, 2009.

(15) Schölkopf, B.; Smola, A. J. *Learning with Kernels*; MIT Press: Cambridge, MA, 2002.

(16) Vapnik, V. N. *Statistical Learning Theory*; Wiley: New York, 1998.

(17) Bakir, G. H.; Hofmann, T.; Schoelkopf, B.; Smola, A. J.; Taskar, B.; Vishwanathan, S. V. N. *Predicting Structured Data*; MIT Press: Cambridge, MA, 2007.

(18) Gärtner, T.; Flach, P. A.; Wrobel, S. On graph kernels: Hardness results and efficient alternatives. In *Proceedings of the 16th Annual Conference on Computational Learning Theory and the 7th Kernel Workshop*; Schölkopf, B., Warmuth, M. K., Eds.; Springer: Berlin, Heidelberg, NY, 2003; pp 129−143.

(19) Kashima, H.; Tsuda, K.; Inokuchi, A. Marginalized kernels between labeled graphs. In Proceedings of the *International Conference on Machine Learning*; Morgan Kaufmann: San Francisco, CA, 2003; pp 321−328.

(20) Mahé, P.; Ralaivola, L.; Stoven, V.; Vert, J.-P. The Pharmacophore Kernel for Virtual Screening with Support Vector Machines. *J. Chem. Inf. Model.* **2006**, *46* (5), 2003–2014.

(21) Mahé, P.; Vert, J.-P. Graph kernels based on tree patterns for molecules. *Mach. Learn.* **2009**, *75* (1), 3−35.

(22) Perret, J.-L., Mahé, P. *ChemCpp User Guide*; Bioinformatics Center and Center for Computational Biology: Kyoto, Japan and Paris, France, 2006; http://chemcpp.sourceforge.net/doc/chemcpp_user-guide.pdf. accessed March 9, 2010.

(23) Ravaivola, L.; Swamidass, S. J.; Saigo, H.; Baldi, P. Graph Kernels for chemical informatics. *Neural Networks* **2005**, *18* (8), 1093–1110.

(24) Schietgat, L.; Ramon, J.; Bruynooghe, M.; Blockeel, H. An efficiently computable graph-based metric for the classification of small molecules. In *Lecture Notes in Computer Science*, Vol. 5255, Proceedings of the International Conference on Discovery Science, Budapest, Hungary, October 13−16, 2008; Springer: Berlin, NY, 2008, 197−209.

(25) *ChemCpp*; Perret, J.-L.; Mahé, P.; Vert, J.-P.; Akutsu, T.; Kanehisa, M.; Ueda, N. Bioinformatics Center of Kyoto University, Japan, and Center for Computational Biology of Ecole des Mines de Paris, France. http://chemcpp.sourceforge.net/html/index.html (accessed March 9, 2010).

(26) Jain, B. J.; Geibel, P.; Wysotzki, F. SVM Learning with the SH Inner Product. In *Proceedings of the 12th European Symposium on Artificial Neural Networks*; Verleysen, M., Ed.; D-side Publications: Evere, Belgium, 2004; Vol. 29, pp 9−304.

(27) Fröhlich, H.; Wegner, J. K.; Sieker, F.; Zell, A. Optimal assignment kernels for attributed molecular graphs. In *Proceedings of the 22nd International Conference on Machine Learning*; ACM Press: New York, NY, USA, 2005, pp 225−232.

(28) Vert, J.-P. *The optimal assignment kernel is not positive definite*; Computational Neurobiology Lab: San Diego, CA, 2008; arXiv: 0801.4061v1, arXiv.org ePrint archive; http://aps.arxiv.org/pdf/0801.4061v1. Accessed July 1, 2009.

(29) Hochreiter, S.; Obermayer, K. Support vector machines for dyadic data. *Neural Comput.* **2006**, *18*, 1472–1510.

(30) Hochreiter, S.; Obermayer, K. Nonlinear feature selection with the potential support vector machine. In *Feature Extraction: Foundations and Applications*, Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L., Eds.; Springer: Berlin, Heidelberg, NY, 2006, pp 419−438.

(31) Knebel, T.; Hochreiter, S.; Obermayer, K. An SMO algorithm for the potential support vector machine. *Neural Comput.* **2008**, *20*, 271–287.

(32) Mohr, J.; Jain, B.; Obermayer, K. Molecule Kernels: A Descriptor- and Alignment-Free Quantitative Structure-Activity Relationship Approach. *J. Chem. Inf. Model.* **2008**, *48* (9), 1868–1881.

(33) Gold, S.; Rangarajan, A. Graduated Assignment Algorithm for Graph Matching. *IEEE Trans. Pattern Anal. Mach. Intell.* **1996**, *18*, 377–388.

(34) Sinkhorn, R. A relationship between arbitrary positive matrices and doubly stochastic matrices. *Ann. Math. Stat.* **1964**, *35*, 876–879.

(35) Munkres, J. Algorithms for the Assignment and Transportation Problems. *J. Soc. Indust. Appl. Math.* **1957**, *5* (1), 32–38.

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

(37) Kirkland, D.; Aardema, M.; Henderson, L.; Müller, L. 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.* **2005**, *584*, 1–256.

(38) Ishidate, Jr., M.; Harnois, M. C.; Sofuni, T. A comparative analysis of data on the clastogenicity of 951 chemical substances tested in mammalian cell cultures. *Mutat. Res.* **1988**, *195*, 151–213.

(39) *National Toxicology Program*; Department of Health and Human Services: Research Triangle Park, NC; http://ntp.niehs.nih.gov/. Accessed July 1, 2009.

(40) *IUCLID5*; European Chemicals Agency: Helsinki, Finland, 2007−2009; http://iuclid.echa.europa.eu/. Accessed July 1, 2009.

(41) Abbatt, J. D.; Bora, K. C.; Quastel, M. R.; Lefkovitch, L. P. International reference study on the identification and scoring of human

chromosome aberrations. Results of a WHO comparative study. *Bull. W. H. O.* **1974**, *50* (5), 373–388.

(42) Schwaighofer, A.; Schroeter, T.; Mika, S.; Hansen, K.; Ter Laak, A.; Lienau, P.; Reichel, A.; Heinrich, N.; Müller, K. R. A probabilistic approach to classifying metabolic stability. *J. Chem. Inf. Model.* **2008**, *48* (4), 785–796.

(43) Sadowski, J.; Schwab, C.; Gasteiger, J. *Corina*, v3.4; Molecular Networks GmbH Computerchemie: Erlangen, Germany.

(44) Todeschini, R.; Consonni, V.; Mauri, A.; Pavan, M. *Dragon for Windows and Linux*; Talete SRL: Milano, Italy, 2006; http://www.talete.mi.it/. Accessed July 1, 2009.

(45) Blum, A.; Langley, P. Selection of relevant features and examples in machine learning. *Artif. Intell.* **1997**, *97*, 245–271.

(46) Guyon, I.; Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.

(47) Kohavi, R.; John, G. H. Wrappers for feature subset selection. *Artificial Intelligence* **1997**, *97* (1−2), 273–324.

(48) Bradley, P. S.; Mangasarian, O. L. Feature Selection via Concave Minimization and Support Vector Machines. In *Proceedings of the Fifteenth international Conference on Machine Learning*; Shavlik, J. W., Ed.; Morgan Kaufmann Publishers: San Francisco, CA, 1998; pp 82−90.

(49) *Spider toolbox*, v 1.71; http://www.kyb.mpg.de/bs/people/spider/main.html. Accessed July 1, 2009.

(50) *Libsvm*, v 2.88; http://www.csie.ntu.edu.tw/ cjlin/libsvm/ (accessed July 1, 2009).

(51) Breiman, L. Random forests. *Mach. Learn.* **2001**, *45* (1), 5–32.

(52) Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24* (2), 123–140.

(53) Ashby, J.; Tennant, W. Definitive relationships among chemical structure, carcinogenicity and mutagenicity for 301 chemicals tested by the U.S. NTP. *Mutat. Res.* **1991**, *257*, 229–306.

(54) Eder, E.; Henschler, D.; Neudecker, T. Mutagenic properties of allylic and a, (3-unsaturated compounds: consideration of alkylating mechanisms. *Xenobiotica* **1982**, *12*, 831–848.

(55) Sugiura, K.; Goto, M. Mutagenicities of styrene oxide derivatives on bacterial test systems: relationship between mutagenic potencies and chemical reactivity. *Chem. Biol. Interact.* **1981**, *35* (1), 71–91.

(56) Eriksson, L.; Hellberg, S.; Johansson, E,; Jonsson, J.; Sjöström, M.; Wold, S.; Berglind, R.; Karlsson, B. A strategy for ranking environmentally occurring chemicals. Part VI. QSARs for the mutagenic effects of halogenated aliphatics. *Acta Chem. Scand.* **1991**, *45* (9), 935–44.

(57) Colvin, M. E.; Hatch, F. T.; Felton, J. S. Chemical and biological factors affecting mutagen potency. *Mutat. Res.* **1998**, *400* (1−2), 479–92.

(58) Trieff, N. M.; Biagi, G. L.; Sadagopa Ramanujam, V. M.; Connor, T. H.; Cantelli-Forti, G.; Guerra, M. C.; Bunce, H. 3rd; Legator, M. S. Aromatic amines and acetamides in Salmonella typhimurium TA98 and TA100: a quantitative structure-activity relation study. *Mol. Toxicol.* **1989**, *2* (1), 53–65.

(59) Yamamoto, K.; Kawanishi, S. Site-Specific DNA Damage by Phenylhydrazine and Phenelzine in the Presence of Cu(II) Ion or Fe(III) Complexes: Roles of Active Oxygen Species and Carbon Radicals. *Chem. Res. Toxicol.* **1992**, *5*, 440–446.

(60) Black, T. H. The Preparation and Reactions of Diazomethane. *Aldrichimica Acta* **1983**, *16*, 3.

(61) Süssmuth, R.; Haerlin, R.; Lingens, F. The mode of action of N-methyl-N′-nitro-N-nitrosoguanidine in mutagenesis. VII. The transfer of the methyl group of N-methyl-N′-nitro-N-nitrosoguanidine. *Biochim. Biophys. Acta* **1972**, *269* (2), 276–86.

(62) Cerdá-Olmedo, E.; Hanawalt, P. C. Diazomethane as the Active Agent in Nitrosoguanidine Mutagenesis and Lethality. *Molec. Gen. Genetics* **1968**, *101*, 191–202.

(63) Mirvish, S. S.; Williamson, J.; Babcook, D.; Chen, S. C. Mutagenicity of iso-butyl nitrite vapor in the Ames test and some relevant chemical properties, including the reaction of iso-butyl nitrite with phosphate. *Environ. Mol. Mutagen.* **1993**, *21* (3), 247–52.

(64) Törnqvist, M.; Rannug, U.; Jonsson, A.; Ehrenberg, L. Mutagenicity of methyl nitrite in Salmonella typhimurium. *Mutat. Res.* **1983**, *117* (1−2), 47–54.

(65) Wild, D.; King, M. T.; Gocke, E.; Eckhardt, K. Study of artificial flavouring substances for mutagenicity in the Salmonella/microsome, Basc and micronucleus tests. *Food Chem. Toxicol.* **1983**, *21* (6), 707–19.

(66) Vontor, T.; Socha, J.; Vecera, M. Kinetics and mechanism of hydrolysis of 1-naphthyl N-methyl- and N, N-dimethylcarbamates. *Collect. Czech. Chem. Commun.* **1972**, *37*, 2183–2196.

(67) Bergon, M.; Hamida, N. B.; Calmon, J. P. Isocyanate formation in the decomposition of phenmedipham in aqueous media. *J. Agric. Food Chem.* **1985**, *33* (4), 577–583.

(68) Tamura, N.; Aoki, K.; Lee, M. S. Selective reactivities of isocyanates towards DNA bases and genotoxicity of methylcarbamoylation of DNA. *Mutat. Res.* **1992**, *283* (2), 97–106.

(69) Roy-Burman, P. *Analogues of Nucleic Acid Components: Mechanism of Action*, Recent Results in Cancer Research, Springer-Verlag: Berlin, NY, 1970; Vol. 25.

(70) Snyder, R. D.; Gillies, P. J. Evaluation of the clastogenic, DNA intercalative, and topoisomerase II-interactive properties of bioflavonoids in Chinese hamster V79 cells. *Environ. Mol. Mutagen.* **2002**, *40* (4), 266–276.

(71) Maruo, N.; Horiuchi, H.; Nakabo, T.; Kondo, M.; Nakamura, T. Cytokinetic study on the effects of N4-behenoyl-1- β -d-arabinofuranosylcytosine on murine leukemic cells L1210: A comparison with the effects of 1- β -d-arabinofuranosylcytosine. *Hematol. Oncol.* **1985**, *3*, 39–48.

(72) Major, P. P.; Egan, E. M.; Herrick, D. J.; Kufe, D. W. Effect of ara-C incorporation on deoxyribonucleic acid synthesis in cells. *Biochem. Pharmacol.* **1982**, *31* (18), 2937–2940.

(73) Regulation No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH).

(74) QSARs and grouping of chemicals. *Guidance on information requirements and chemical safety assessment*; European Chemical Agency: Helsinki, Finland, 2008; http://echa.europa.eu/reach_en.asp. Accessed July 1, 2009.

(75) *Potential - Support Vector Machine*, v 1.31; Neural Information Processing Group, Department for Electrical Engineering and Computer Science, Berlin Institute of Technology: Berlin, Germany; http://ni.cs.tu-berlin.de/software/psvm/index.html. Accessed July 1, 2009.

CI900367J