Toxicity-Indicating Structural Patterns

Modest von Korff and Thomas Sander*

Department of Research Informatics, Actelion Ltd., Gewerbestrasse 16, CH-4123 Allschwil, Switzerland

Received August 31, 2005

We describe a toxicity alerting system for uncharacterized compounds, which is based upon comprehensive tables of substructure fragments that are indicative of toxicity risk. These tables were derived computationally by analyzing the RTECS database and the World Drug Index. We provide, free of charge, a Java applet for structure drawing and toxicity risk assessment. In an independent investigation, we compared the toxicity classification performance of naive Bayesian clustering, *k* next neighbor classification, and support vector machines. To visualize the chemical space of both toxic and druglike molecules, we trained a large self-organizing map (SOM) with all compounds from the RTECS database and the IDDB. In summary, we found that a support vector machine performed best at classifying compounds of defined toxicity into appropriate toxicity classes. Also, SOMs performed excellently in separating toxic from nontoxic substances. Although these two methods are limited to compounds that are structurally similar to known toxic substances, our fragment-based approach extends predictions to compounds that are structurally dissimilar to compounds used in the training set.

INTRODUCTION

During the past decade, even early stages of drug discovery were driven by multidimensional optimization.[1] In addition to a compound's binding affinity to the target protein, researchers focused on its adsorption, distribution, metabolism, excretion, and toxicity (ADMET) features increasingly early. Where reliable high-throughput assays to evaluate ADMET properties are expensive or missing, there is a high demand for predictive in silico techniques. Thus, computational methods that deliver reliable indicators of a compound's toxicity potential are of high interest to medicinal chemists. Furthermore, there is a strong regulatory interest in the European Union and in the United States to complement animal tests by toxicity prediction systems.[2,3]

Some examples of available computer programs predicting toxicity are Case,[4] DEREK,[5] Multicase,[6] Hazardexpert,[7] TOPKAT,[8] OnkoLogic,[9] and ESP.[10] Several toxicity prediction algorithms rely on the fact that toxic behavior is often associated with structural motives.[11,12] These prediction systems are rule-based, like DEREK, HazardExpert, and OncoLogic. Another approach to predict toxicity is the use of QSAR models as by TOPKAT and CASE; MCASE and ESP are hybrids. The literature for toxicity prediction was recently reviewed.[13]

Toxicity is the ability of a chemical agent to cause injury.[14] Toxic effects may be caused by many different modes of action. Chemical reactions are only one way to injure an organism. Cancer may be caused by mechanical disturbance from asbestos fibers.[15] Toxic effects can be caused by bioactive molecules, which bind as ligands to a protein and cause an adverse effect. Proteins themselves may also cause toxic effects, the very potent Botulinum toxin being a famous example.[16] To fulfill its toxic potential, a substance needs to reach its place of action. As for drugs, the ADME (absorption, distribution, metabolism, and excretion) profile of a toxin influences it's behavior. Thus, properties such as poor aqueous solubility, the inability to pass through membranes, or susceptibility to rapid metabolism may render an otherwise aggressive compound harmless. These complex relationships complicate the construction of reliable models for predicting toxicity.

Our approach consists of three parts, the first of which is a fragment-based statistical analysis of toxicity databases (DBs), the second applies standard classification algorithms on top of our self-developed chemical fingerprint descriptors, and the third visualizes the chemical space of toxic and nontoxic molecules as an aid to interpretation of the analysis.

Risk Indicator Fragments. Our approach relies on the fact that toxic behavior is often associated with structural motives. We assume that substructure fragments, which are frequently found in toxic compounds but only rarely in nontoxic ones, are likely to cause the toxic effect. To identify those fragments in a manner free from human bias, we implemented a computer program for the task; we used the Registry of Toxic Effects of Chemical Substances database (RTECS), which covers compounds of various toxicity classes (Table 1),[17]as a reference database. For our studies, we considered substances which are known to be mutagenic, tumorigenic, or irritant or to cause reproductive effects. Sufficient data are available on these four kinds of toxicity effects to generate and to validate expressive models. As supposedly nontoxic reference compounds, we took a subset from the World Drug Index (WDI; Table 1).[18]

Analysis of the RTECS structures of the four toxicity categories yielded several hundred thousand raw substructure fragments. Query features were introduced to further generalize fragments or to specify substitution patterns. For every fragment, we counted its occurrence frequency within all molecules of a specific toxicity class and normalized this

* Corresponding author phone: +41 61 565 65 23; fax: +41 61 565 65 00; e-mail: thomas.sander@actelion.com.

10.1021/ci050358k CCC: $33.50
Published on Web 01/07/2006

Table 1. Data Sets Used

name	description	# compounds
RTECS	Registry of Toxic Effects of Chemical Substances subset, test and training data	16 085
IDDB	Investigational Drugs database subset, no anticancer, test, and training data	17 000
WDI cleaned	World Drug Index subset, no anticancer, test, and training data	5876
cytostatics	from WDI, test set	261
supplier data set	supplier claim: new and druglike molecules, test set	10 731

frequency by the fragment's natural occurrence found in the reference database. Fragments with both a statistically relevant overall frequency and significantly higher occurrence in the toxic group of compounds were listed as potentially risky fragments, with the latter property determining whether the fragment is considered an indicator for medium or high risk.

For predicting the toxicity of an uncharacterized molecule, we ran a substructure search for all risky fragments, which, if found, provided an indication of potential toxicity risk. The absence of risky fragments, however, suggests a low risk concerning the toxicity class under investigation. We provide a free Java applet that permits the assessment of toxicity risks of any drawn chemical structure.[19]

Standard Classification Algorithms. To assess the performance of our fragment-based toxicity estimation, we compared the prediction accuracy for mutagenicity with four standard classification algorithms: naive Bayesian, tree-augmented naive Bayesian (TAN Bayes), *k* nearest neighbors (*k*NN), and sequential minimal optimization (SMO).[20,21] SMO belongs to the kernel-based classification methods and is a high-performance implementation of the support vector machine.[22] For the validation of our models, we used leave-multiple-out cross-validation.[23] Both the mutagenic compound data set and the drug data set were split into equally sized test and training data sets after random shuffling. The training data were used together with the algorithms to generate the models. With the models, the toxicity of the test data was predicted. For each algorithm, the model was calculated from the training set and then applied to predict the test data. The average results of seven repetitions for each algorithm were evaluated with Bayesian analysis.[24] The models of the four classification algorithms were based on the Tanimoto similarity of the Actelion fingerprint descriptor (ActelionFp).[25]

Visualization. To visualize the chemical space occupied by all molecules from the RTECS and the Investigational Drugs database (IDDB),[26] we trained a self-organizing map (SOM) with the compounds of both databases.[27] SOMs were already used in several approaches for toxicity prediction. Espinosa and co-workers used the SOM as a clustering tool to find meaningful descriptors.[28] The clustering capability of a SOM was also used to select clusters to generate local models for toxicity prediction.[29] Mazzatorta and co-workers generated a QSAR model with a supervised SOM.[30] We decided to use the SOM in its original context, as a tool to organize and visualize data by structural similarity. The Tanimoto coefficient of our in-house molecular fragment-based fingerprint (ActelionFp) again served as the similarity criterion. The SOM training yielded a two-dimensional grid of SOM neurons (i.e., reference vectors), each coordinate of which constituted an instance of the ActelionFp. It is characteristic of a well-trained SOM that closely located neurons have similar information content and that, for every training molecule, there exists at least one similar neuron on the SOM. Finally, we mapped all compounds to their closest matching neurons on the SOM and displayed appropriately located markers colored according to the compound's toxicity classes. The result shows the distribution of toxic compounds in the chemical descriptor space, reduced to two dimensions. Areas with a high density of toxic compounds indicate regions in the chemical space to be avoided. To highlight these regions, we also tinted the SOM's background on the basis of the frequency ratio of toxic versus nontoxic compounds. By mapping external compound libraries—be it virtual libraries from combinatorial chemistry or commercially available screening compounds—onto the "toxicity-tinged" SOM, one gets a quick visual impression of the risky subsets of those libraries as well as of the libraries' overall structural diversity.

To benchmark the visualization power of the SOM, we used the ISOMAP algorithm.[31] It combines multidimensional scaling with principal component analysis. This algorithm claims to conserve the distances in the original space to the mapped space.

METHODS

Data Sets (Table 1). Toxicity data were extracted from the RTECS database, version June 2004, available with the ChemBank database through Croner, compiled by NIOSH (National Institute of Occupational Safety and Health).[32] Since the database itself does not contain chemical structures, we used the ChemOffice[33] "Name to structure" function to generate the structures' connection tables from the IUPAC names contained in the database. This way, we gained 125 195 records containing chemical structures from 158 340 records in total. The remaining records were plant extracts, mixtures, had invalid IUPAC names, or had a name that the ChemOffice function could not interpret. A total of 16 085 structures belong to at least one of the four considered toxicity categories. These categories comprised 10 040 mutagenic, 3674 tumorigenic, 3283 irritant, and 4732 reproductive effective entries. The nontoxic reference compounds were extracted from the World Drug Index from which all traded drugs were selected by taking those compounds with an available INN (international nonproprietary name). From this subset of 6532 compounds, we removed all compounds indicated with at least one of the following terms: tumor, leukemia, malignant, cytostatics, and antiseptics. After removing all redundant records, 5876 structures remained.

IDDB Subset. The IDDB contains traded drugs and bioactive molecules.[26] Every anticancer compound was removed. From the remaining 35 000 compounds, a diverse subset of 17 000 molecules was selected.

Supplier Data Set. The supplier data set consists of 10 731 molecules offered by one supplier. The supplier claims that

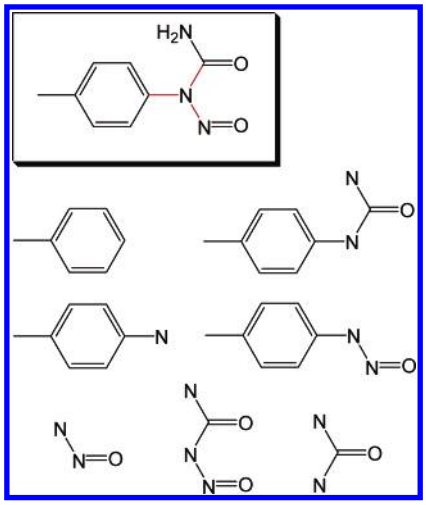

Figure 1. Mutagenic molecule with cuttable bonds, in red, and all raw fragments.

the structures are new, druglike, and well-suited as receptor ligands in bioassay screening experiments.

Cytostatic Data Set. The data set with cytostatic compounds was extracted from the WDI and comprised 261 molecules.

Algorithms. For the three algorithms naive Bayesian, tree-augmented naive Bayesian, and sequential minimal optimization we used the Weka implementation.[34] For the ISOMAP algorithm, we adapted a Java source.[35] All other algorithms were implemented in Java by us. The visualization of the SOMs was done in DataWarrior, a Java-based in-house tool for chemical data analysis, property prediction, and visualization.[36]

Risk Indicator Fragments. The fragment creation process was as follows: For a given chemical structure, the algorithm determined, in a first step, all cuttable bonds, which are the bonds that connect rigid fragments. These were defined as all nonring, nonterminal single bonds of the hydrogen-depleted graph. In a second step, all substructures of the original structure were generated which can be formed by cutting one or more of those cuttable bonds. We, thus, obtained a list of raw fragments (Figure 1).

From the set of raw fragments, we derived a more specific set of fragments by adding the original substitution information. Atoms that had a higher substituent count in the original molecule were marked with a query feature to require at least one more substituent. The remaining atoms were saturated with hydrogen atoms (Figure 2).

A further set of fragments was derived from the raw fragments by replacing every atom by an atom wild card, effectively retaining the skeletal backbone of the original fragment. Finally, we derived from every one of these skeletal fragments a small set of more specific skeletal fragments. For this, we created a fresh copy of the skeletal fragment for every heteroatom in the original fragment. The atom that originally had been a heteroatom was defined to be a non-carbon atom. Altogether, we derived from any given raw fragment $2 + n$ additional fragments, with n being the number of heteroatoms.

All raw fragments, substitution conscious fragments, skeletal fragments, and derived skeletal fragments were added to an overall toxicity-class-specific fragment list.

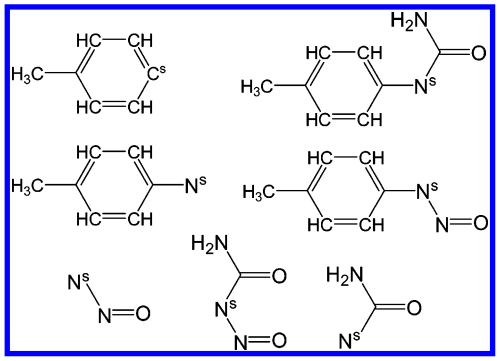

Figure 2. Query feature decorated fragments (s: further substituent required).

For any of our fragments, we determined its absolute frequency within the WDI database and within the appropriate RTECS subset. If a fragment was not present in the reference database, then an absolute frequency in an RTECS subset above 6 or above 10 was considered an indicator for medium or high risk, respectively. In cases where a fragment was present in the reference database, the ratio between its relative frequency in an RTECS subset to its relative frequency in the reference database needed to be above 6 or 10 to consider the fragment indicative of medium or high risk, respectively. To eliminate redundant information, we removed fragments that were superstructures of other fragments in the same list and we removed fragments from the medium-risk fragment list if they were superstructures to any fragment of the high-risk list. In this way, we obtained for the four toxicity classes *mutagenic*, *tumorigenic*, *irritant*, and *reproductive effective* 4972, 1757, 2396, and 2305 fragments, respectively.

A similar approach was described by Lai and co-workers, who dissected the Wiswesser line notations (WLN) of the RTECS records to yield frameworks, ring systems, linkers, and side chains as fragments.[37] Earlier, Bemis and Murcko also applied a WLN-based fragmentation approach.[38] Chittimoori used another fragmentation-based approach for toxicity prediction.[39] Enhanced atom types, which belong to the fragment-based methods,[40] were recently used for toxicity prediction.[41]

Visualization. To visualize potentially toxic areas of entire compound libraries, we generated a binary SOM, that is, a SOM consisting of binary reference vectors. Binary SOMs are scarcely described in the literature.[42] Lobo et al. applied them to classify sonar echoes with moderate success. We were inspired by the fact that our input vectors, which were our training molecules' fingerprints, were binary anyway and that processing 512 bits rather than 512 floating point values promised a tremendous gain in computational performance. On the flipside, however, we were sacrificing arbitrarily small changes during reference vector adaptation for quantized changes, because the smallest possible change on a binary vector is the inversion of one bit. To compensate the loss of precision due to binary reference vectors, we use SOMs large enough to outweigh this effect. To accommodate the binary nature of our SOMs, we also adjusted the adaptation process in the following manner. Instead of slightly adapting every component of a reference vector v_r toward the input vector, we changed a few of the reference vector's bits to more closely resemble the input vector. The count of inverted bits

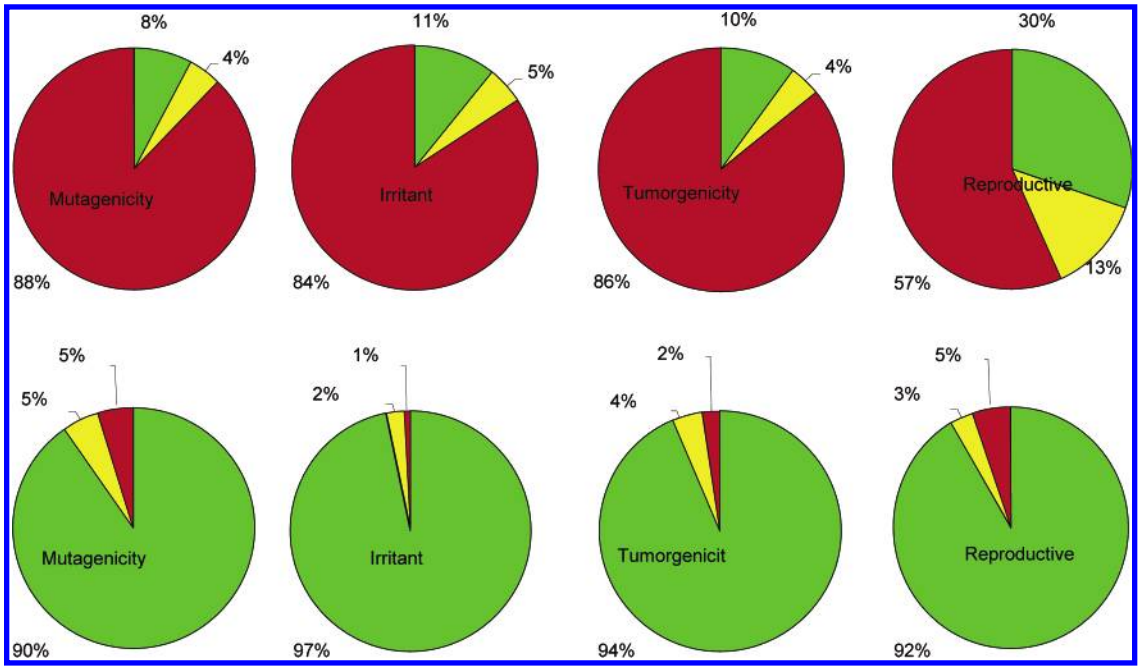

Figure 3. Risk assignments of compound sets. The upper row shows the modeling results of the toxic compounds of the four classes. The row below shows the modeling results of the IDDB compounds. The colors indicate the risk classes. Red: high risk. Yellow: medium risk. Green: low risk.

depended on the adaptation rate. The affected bits were randomly selected. The adaptation rate decreases linearly with time over the training process. It also depends on the distance on the map between the adapted reference vector and the one most closely resembling the input vector. We used a Gaussian curve for this distance dependence. The SOM consists of 200 by 200 reference vectors, also termed neurons. After the training phase, we mapped the training compounds onto the SOM in the usual manner. For any compound, the neuron most similar to the compound's descriptor was selected. The compound is placed on the map at the coordinates of the most similar neuron. Coordinates fine-tuning is done on the basis of the similarity to the neighbor neurons. After the completion of compound mapping, the entire map was colored according to the local densities of mapped compounds belonging to certain classes. Preliminary tests showed a more than 10-fold performance gain of the binary version compared to the classical SOM. A detailed description of this new kind of SOM will be the topic of a future publication.

The ISOMAP algorithm requires optimization of a distance parameter to achieve optimal results. We calculated the ISOMAP with these distance values: 3, 7, 15, 30, 60, 120, and 240. The best separation of toxic from nontoxic compounds was achieved with a distance of 15. The data set used for the SOM and the ISOMAP algorithm was the RTECS database subset and the IDDB subset.

RESULTS

Risk Indicator Fragments. The pie charts in Figure 3 show a satisfying modeling capability of the fragment-based approach. For the three toxicity groups *mutagenic*, *tumorigenic*, and *irritant*, about 90% of the compounds were correctly found to pose a medium or high risk. For reproductive effects, the modeling performance reached 70%. On the other hand, more than 90% of the IDDB compounds were correctly found to be risk-free concerning all four toxicity classes. Thus, we have fewer false positive results than false negative ones.

Table 2. Results of Test Set Classification Mutagenicity, Leave-Multiple-Out Cross-Validation

algorithm	sensitivity	specificity	positive predictive value	negative predictive value
fragments	0.616	0.837	0.863	0.567
naive Bayes	0.772	0.715	0.817	0.657
TAN Bayes	0.811	0.739	0.836	0.704
*k*NN	0.871	0.768	0.861	0.784
SMO	0.874	0.753	0.874	0.753

Standard Classification Algorithms. In Table 2, the results of the leave-multiple-out cross-validation are shown. For all runs, the standard deviation of the repetitions was below 1%. For the evaluation of the results, the following definitions are given: a, true positive count (number of correctly predicted toxic compounds); b, false negative count (toxic but not predicted to be toxic); c, false positive count (not toxic but predicted to be toxic); and d, true negative count (correctly predicted to be not toxic). The relationship sensitivity $[= a/(a + b)]$ describes the probability of correctly detecting a toxic effect (true positive rate).[43,44] Specificity $[= d/(c + d)]$ is the probability of correctly detecting a nontoxic compound (true negative rate). The positive predictive value $[= a/(a + c)]$ is the probability that a molecule is toxic if it is predicted to be. The negative predictive value $[= d/(b + d)]$ is the probability of a molecule to be nontoxic that is predicted to be harmless.

The SVM outperforms the other algorithms in terms of specificity, closely followed by the simple *k*NN classifier. In the ranking by the positive predictive value, the SVM is again the best algorithm, followed by the fragment-based toxicity prediction. The *k*NN algorithm performs best concerning the average of both predictive values. Swamidass and co-workers reported good results for a similar kernel-

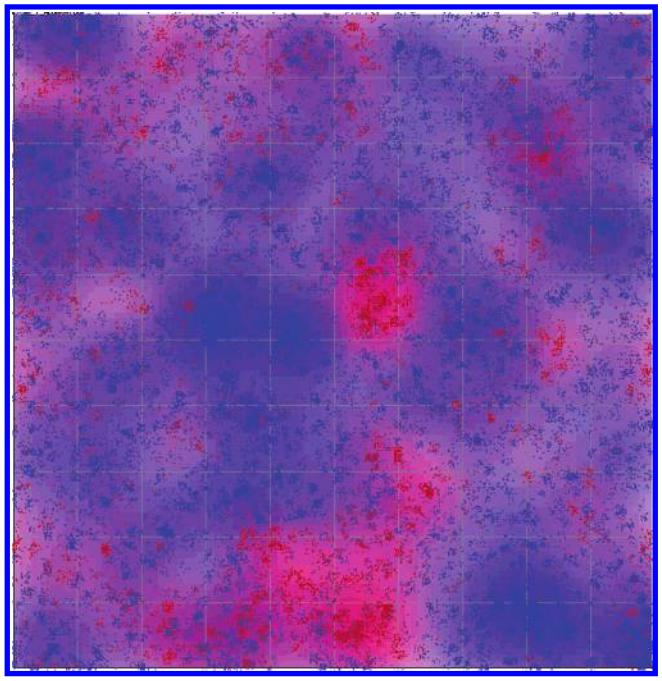

Figure 4. Binary SOM with 40 000 nodes. Descriptor: ActelionFp. Data: RTECS and IDDB compounds. Red: mutagenic compounds. Blue: IDDB.

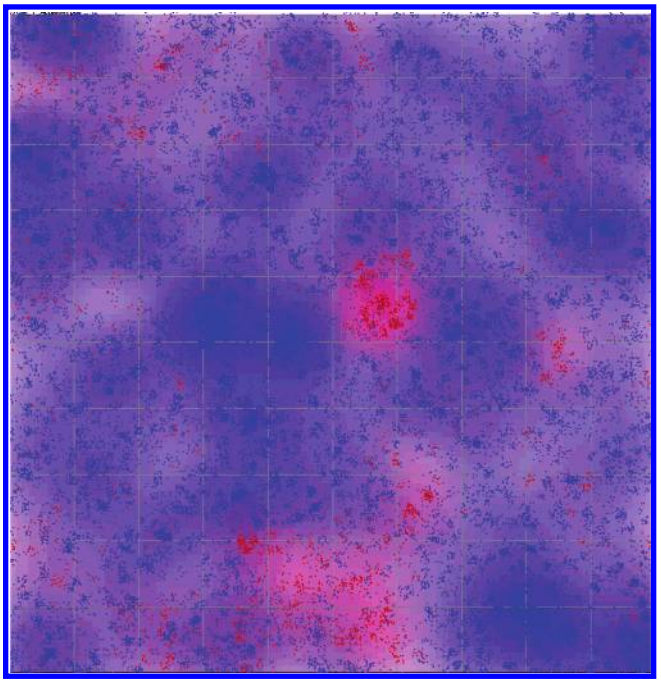

Figure 5. SOM with RTECS and IDDB data. Red: tumorigenic compounds. Blue: IDDB.

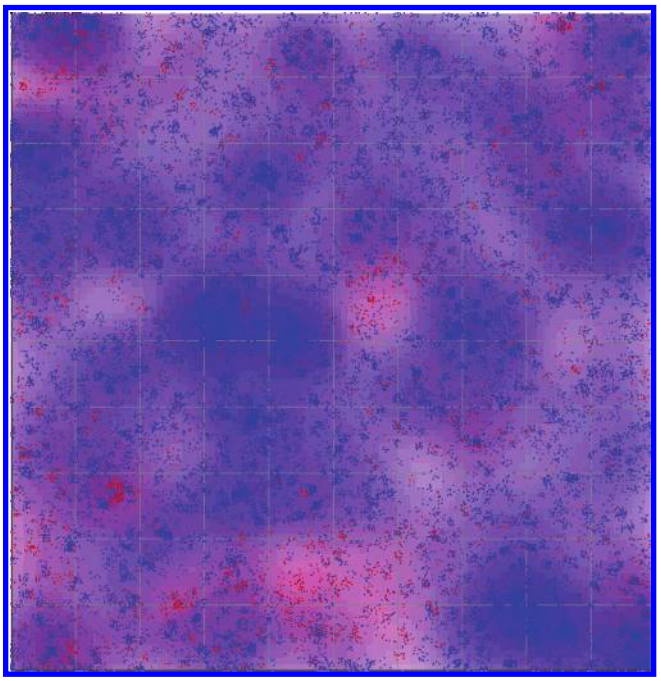

Figure 6. SOM with RTECS and IDDB data. Red: reproductive effectors. Blue: IDDB.

based method. They applied a support vector machine on chemical fingerprint descriptors to predict toxicity.[45]

Visualization. Lobo et al. reported unsatisfying results for applying a binary SOM to classify acoustic signatures of ships.[42] In contrast to this study, we achieved excellent clustering results with our binary SOM implementation. After the SOM had been trained with both the IDDB subset and the RTECS subset, we mapped all compounds of both subsets to the SOM neurons. To visualize mutagenic regions, we colored the background of every SOM neuron from blue to red depending on the ratio of mutagenic to nonmutagenic compounds mapped to the neurons (Figure 4). Analogously, we visualized teratogenic (Figure 5), reproductive effector (Figure 6), and irritant regions (Figure 7). The RTECS compounds not belonging to the toxicity class under consideration do not contribute to the coloring in Figures 4−7. In Figure 8, all toxic compounds contribute to the red coloring. While the members of the toxicity classes tend to huddle in clusters, we see the IDDB substances spread out more homogeneously on the map. Several areas are occupied either by mutagenic compounds or by IDDB molecules, indicating that mutagenic molecules can be distinguished from other bioactive molecules by considering their specific structural features.

To estimate the ability of the SOM to represent out data set, we created a histogram showing the distribution of Tanimoto dissimilarities between SOM reference vectors and the compounds mapped to them (Figure 9). The median of the distribution lies in the bin with 0.2−0.25 units of dissimilarity, which means that most of our compounds are within a similarity of 0.75 or better, well-represented by the SOM. The shape of the distribution is almost normal with a skew to higher dissimilarities.

To examine how the SOM represents external data sets which were not used as training data, we mapped a supplier data set onto the SOM (Figure 10). The toxic RTECS compounds and the nontoxic IDDB compounds from the IDDB are still shown as red and blue background shading. The majority of the supplier compounds avoided the toxic (red) regions of the map, supporting the vendor's claim of the druglike nature of these compounds. Some smaller clusters lie in the border region of toxic clusters.To examine a test data set of compounds with different toxicity types, we mapped 261 cytostatic compounds from the WDI onto the SOM (Figure 11). These compounds are very homogeneously distributed on the map. Only a few sparse clusters can be observed. Toxic areas are as evenly occupied as the nontoxic areas. As expected, alkylating agents such as Mitobronitol, Mitoclomine, and Tretamine are located in

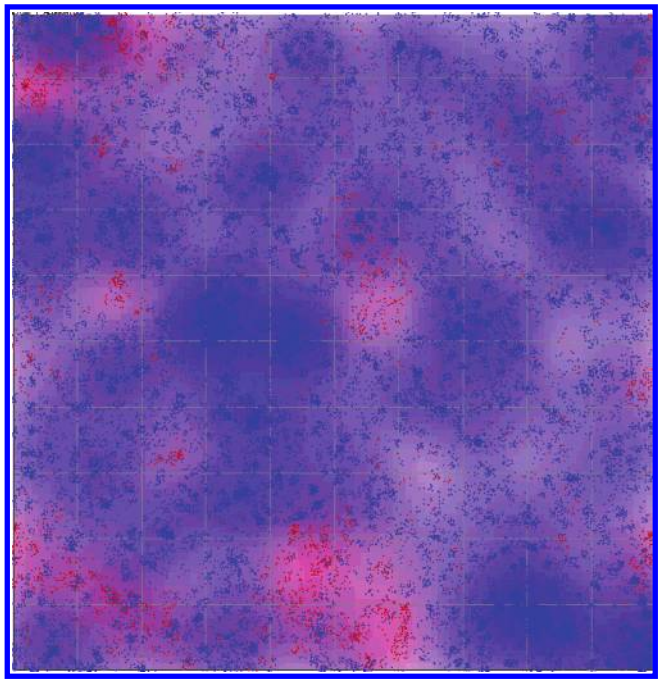

Figure 7. SOM with RTECS and IDDB data. Red: primary irritants. Blue: IDDB.

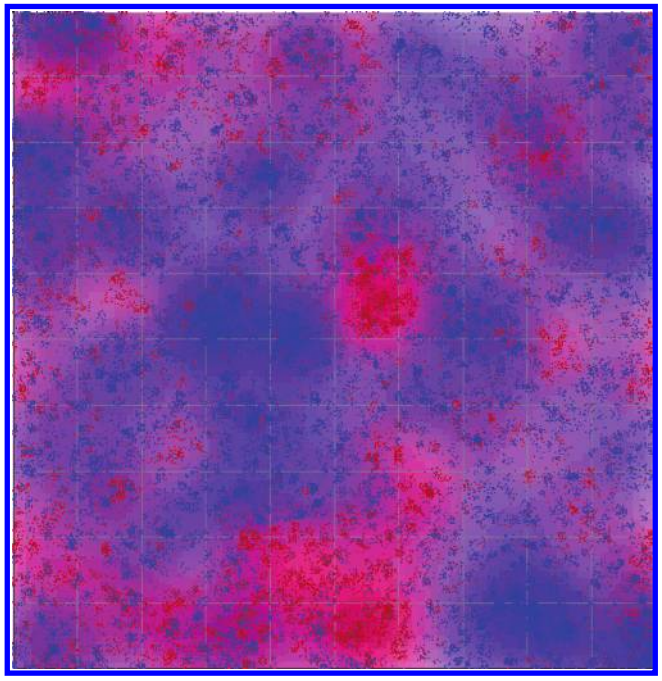

Figure 8. SOM with RTECS and IDDB data. Red: mutagenic, reproductive effector, irritant, and tumorigenic compounds. Blue: IDDB.

toxicity-indicating regions of the map. Molecules with a more sophisticated mode of action like Zindoxifene, which acts as an antiestrogen agent, and Maitansine, which is an antitubulin agent, are not recognized as being toxic. This experiment shows that the SOM is well-suited to reveal toxic compounds by applying its knowledge of toxic molecule structures. Reliable identification of the mechanism of toxicity is more challenging when it is mediated through activation or inactivation of receptors and is only possible if quite similar molecules were used to generate the SOM. A good example is the antimetabolite Fluorouracil that is located in a toxic region and surrounded by similar structures

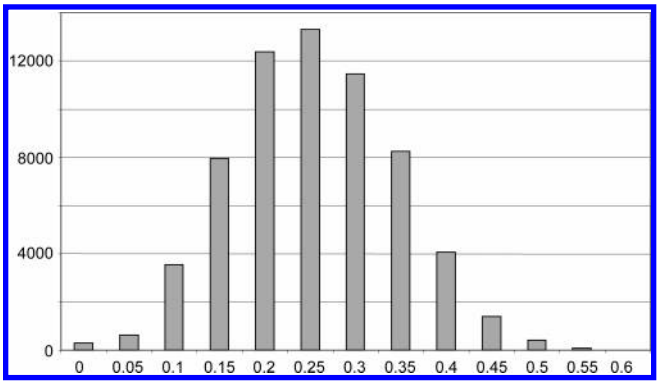

Figure 9. Histogram of dissimilarities between SOM reference vectors and compounds mapped to them. *x* axis: Tanimoto dissimilarity on ActelionFp. *y* axis: frequency.

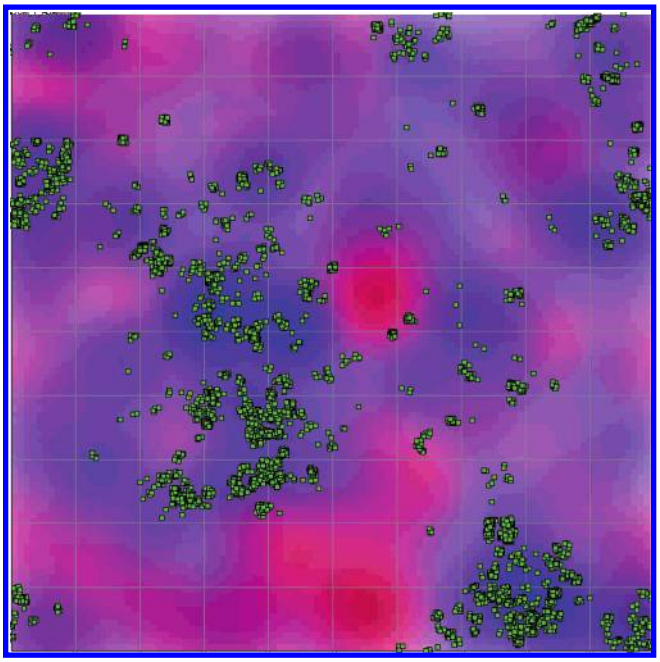

Figure 10. Supplier library mapped onto the SOM. Toxic areas: red. Nontoxic areas: blue. Mapped compounds: green.

of the RTECS DB. However, according to the RTECS DB, these surrounding structures have a different mechanism of toxicity, that is, mutagenesis.

A dissimilarity histogram is a measure of the ability of the SOM to represent the matched data set. A low dissimilarity median is a necessary condition to assume similar properties of closely located compounds. Comparing the dissimilarity histograms (Figures 12and 13) of the two mapped libraries reveals a significant difference. The supplier library is not well-represented by the SOM's neurons, leading to the conclusion that those compounds are substantially different than the toxic training compounds and also the druglike molecules used to train the SOM. This supports the supplier's claim that these compounds are new and, therefore, may be well-suited as potential starting points in drug finding. The majority of the cytostatics, however, map even better to the SOM vectors than the training compounds. Therefore, the toxicity profile of these compounds can be reliably assumed to be similar to those of the closely located training compounds.

The ISOMAP algorithm in Figure 14 shows a less pronounced clustering between the toxic and the nontoxic

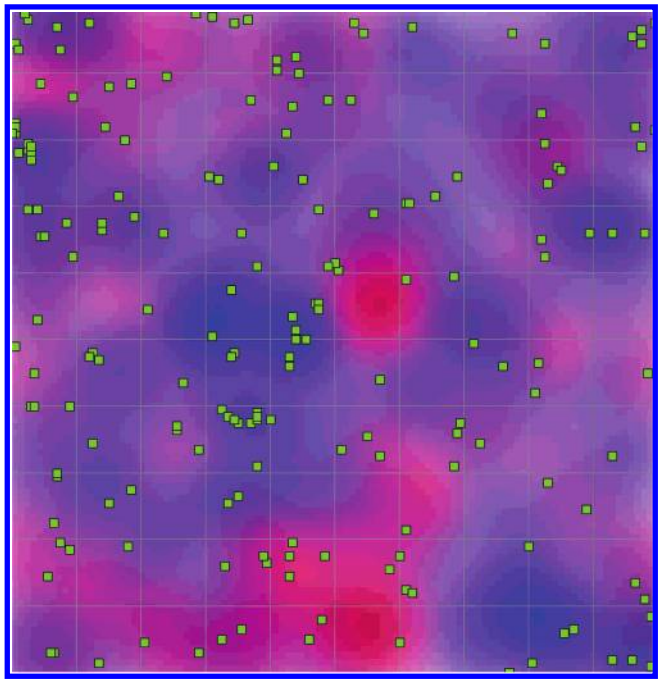

Figure 11. Cytostatic compounds from the WDI mapped onto the SOM. Toxic areas: red. Nontoxic areas: blue. Mapped compounds: green.

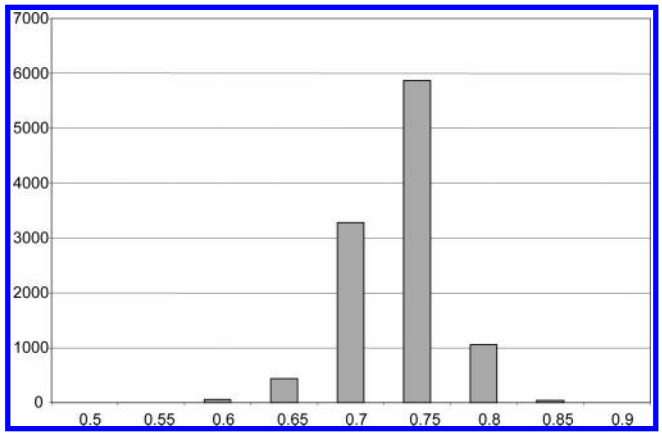

Figure 12. Histogram of dissimilarities between SOM reference vectors and compounds from the supplier library mapped to them. *x* axis: bins for inverse Tanimoto distance. *y* axis: frequency.

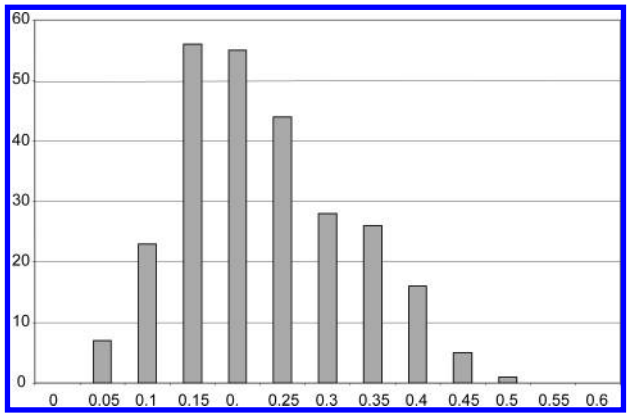

Figure 13. Histogram of dissimilarities between SOM reference vectors and cytostatic compounds from the WDI mapped to them. *x* axis: bins for inverse Tanimoto distance. *y* axis: frequency.

compounds. One can observe one large cluster consisting almost soley of nontoxic compounds. In the other clusters,

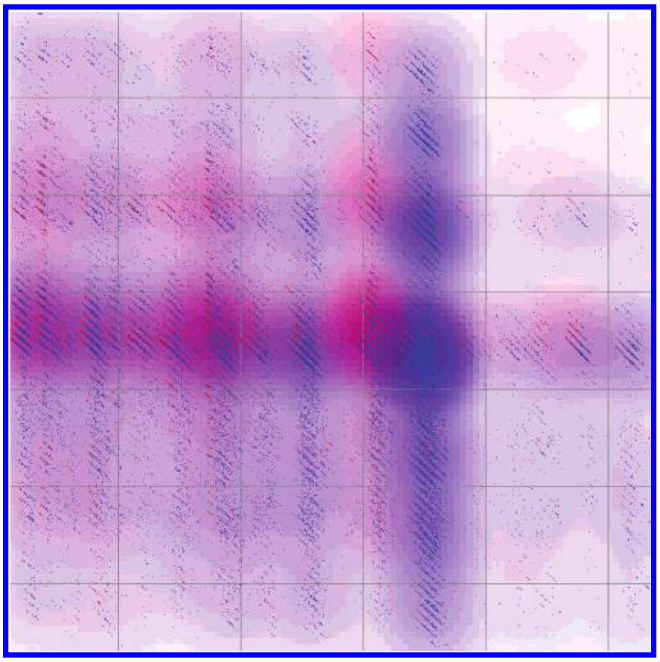

Figure 14. ISOMAP with distance 15. Data: RTECS DB of mutagenic, reproductive effect, irritant, and tumorigenic data and IDDB subset. Descriptor: ActelionFp.

the toxic and the nontoxic compounds are spread almost equally.

SUMMARY AND CONCLUSIONS

A limitation that reduces the reliability of all toxicity prediction methods is that, in the universe of druglike compounds, good quality, experimentally determined toxicity data is available for only a very small proportion of compounds, compared to the actual (unknown) number of toxicity determinants in chemical space. This can be inferred from the small size of our reference data sets, relative to possible structural diversity. Notwithstanding this, and as we expected, the Actelion toxicity prediction method is, therefore, well-suited to flag potentially hazardous molecules but is less reliable for concluding that a compound is probably safe. The fragment-based toxicity predictor was shown to be a conservative estimator of mutagenicity. The positive predictive value of 86% means that a molecule predicted as mutagenic has a high probability to be mutagenic in reality. Contrarily, we cannot conclude that a molecule is harmless if mutagenicity is not predicted. The situation is similar for the three other toxicity categories. Since the prediction model is based on chemical query structures, it is intuitive for chemists and pharmacologists. The algorithm performs well in comparison to the examined standard classification algorithms, but its applicability goes beyond theirs since toxicity risks may be detected for compounds being substantially dissimilar to the RTECS training compounds.

The SMO algorithm wins the classification contest by a narrow margin. The good performances of all applied classification algorithms demonstrate that the ActelionFp descriptor is well-suited to reveal toxic compounds. The high predictive values for the *k*NN classifier with the ActelionFp prove that similar toxic behavior correlates with structural similarity. Thus, organizing compounds in a SOM by structural similarity should separate toxic from nontoxic molecules in distinct regions of the map.

Indeed, our binary SOM did a good job of separating toxic compounds from the IDDB molecules. Since the SOM is an unsupervised learner and doesn't consider toxicity classes of molecules during self-organization, overfitting is not possible. Compounds that were not considered during the SOM training may be mapped onto the SOM. Properties of these compounds can be expected to be similar to those of the nearby compounds used to train the SOM. The property similarity, however, relies on a high structural similarity between SOM neurons and mapped compounds. A high similarity median is a necessary condition to assume similar properties of closely located compounds.

The clustering ability of the ISOMAP algorithm proved to be inferior to the robust clustering ability of the SOM. These results are in accordance with the experiments of Balasubramanian and Schwartz, which concluded that the ISOMAP algorithm tends to fail on complex multidimensional spaces.[46] Another disadvantage of the ISOMAP algorithm is the necessity to optimize its distance parameter, which reduces its general applicability.[47]

We showed that the investigated algorithms together with the ActelionFp descriptor perform well in detecting toxic patterns in highly diverse data sets. The fragment-based approach extends toxicity risk assessment beyond the borders of known toxic compounds. Both methods can be applied to filter new screening compounds before acquisition or synthesis to significantly reduce the risk of pursuing potentially hazardous lead families. SOMs complement these approaches by visualizing the multidimensional chemical space and separating compounds of various toxicity classes from nontoxic ones.

REFERENCES AND NOTES

(1) Oprea, T. Lead Structure Searching: Are We Looking at the Appropriate Property? *J. Comput.-Aided Mol. Des.* **2002**, *16*, 325−334.

(2) Bruner, L. H.; Carr, G.; Curren, R. G.; Chamberlain, M. Validation of Alternative Methods For Toxicity Testing. *Comments Toxicol.* **1997**, *6*, 37−51.

(3) Directive 2003/15/EC of the European Parliament and of the Council of February 27, 2003 amending Council Directive 76/768/EEC on the approximation of the laws of the Member States relating to cosmetic products (Text with EEA relevance). European Union; Official Journal L 066; November 3, 2003; pp 0026−0035.

(4) Klopman, G. Computer Automated Structure Evaluation of Biological Activity of Organic Molecules. *J. Am. Chem. Soc.* **1984**, *106*, 7315−7321.

(5) Sanderson, D. M.; Earnshaw, C. G. Computer Prediction of Possible Toxic Action from Chemical Structure. *Hum. Exp. Toxicol.* **1991**, *10*, 261−273.

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

(7) Smithing, M.; Darvas, F. Hazardexpert: an expert system for predicting chemical toxicity. In *Food Safety Assessment*; American Chemical Society: Washington, DC, 1992; pp 191−200.

(8) Enslein, K.; Gombar, V. K.; Blake, B. W. International Commission for Protection Against Environmental Mutagens and Carcinogens. Use of SAR in computer-assisted prediction of carcinogenicity and mutagenicity of chemicals by the TOPKAT program. *Mutat. Res.* **1994**, *305*, 47−61.

(9) Woo, Y.; Lai, D.; Argus, M.; Arcos, J. Development of structure−activity relationship rules for predicting carcinogenic potential of chemicals. *Toxicol. Lett.* **1995**, *79*, 219−228.

(10) Klopman, G.; Chakravarti, S. K.; Zhu, H.; Ivanov, J. M.; Saiakhov, R. D. ESP: a method to predict toxicity and pharmacological properties of chemicals using multiple MCASE databases. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 704−715.

(11) Ashby, J.; Tennant, R. Chemical structure, Salmonella mutagenicity and extent of carcinogenicity as indicators of genotoxic carcinogenesis among 222 chemicals tested in rodents by the U.S. NCI/NTP. *Mutat. Res.* **1988**, *204*, 17−115.

(12) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activity? *J. Med. Chem.* **2002**, *45*, 4350−4358.

(13) Cronin, M. T. D.; Livingstone, D. J. Predicting Chemical Toxicity and Fate. CRC Press: Boca Raton, FL, 2004.

(14) Katzung, B. Introduction to Toxicology: Occupational & Environmental. In *Basic & Clinical Pharmacology*; McGraw-Hill: New York, 2001; p 988.

(15) Eaton, D. L.; Klaasen, C. D. Principles of Toxicology. In *Casarett & Doul's Toxicology-The Basic Science of Poisons*; McGraw-Hill: New York, 2001; p 13.

(16) Pitot, H. C., III.; Dragan, Y. P. Chemical Carcinogenesis. In *Casarett & Doul's Toxicology-The Basic Science of Poisons*; McGraw-Hill: New York, 2001; pp 245−255.

(17) RTECS, Registry of Toxic Effects of Chemical Substances, U.S. Department of Health and Human Services. http://www.cdc.gov/niosh/rtecs/default.html (accessed Mar 2004).

(18) World Drug Index, Thomson Derwent. www.derwent.com (accessed Mar 2002).

(19) Link to property explorer: http://www.actelion.com/uninet/www/www_main_p.nsf/Content/Technologies+Property+Explorer.

(20) Duda, R. O.; Hart, P. E.; Stork, D. G. *Pattern Classification*; Wiley: New York, 2000.

(21) Platt, J. C. Using Analytic QP and Sparseness to Speed Training of Support Vector Machines. In *Advances in Neural Information Processing Systems 11*; MIT Press: Cambridge, MA, 1999; pp 557−563.

(22) Schölkopf, B.; Burges, C. J. C.; Smola, A. J. *Advances in Kernel Methods: Support Vector Learning*; MIT Press: Cambridge, MA, 1999.

(23) Baumann, K.; Albert, H.; von Korff, M. A systematic evaluation of the benefits and hazards of variable selection in latent variable regression. Part I. Search algorithm, theory and simulations. *J. Chemom.* **2002**, *16*, 339−350.

(24) Sivia, D. *Data Analysis: A Bayesian Tutorial*; Clarendon: Oxford, U. K., 1996.

(25) Sander, T. ActelionFp. Unpublished work, 2002. The Actelion fingerprint is a nonhashed binary fingerprint encoding the existence or absence of 512 predefined substructure fragments that partially contain wild-card atoms and atom query features. These fragments were computationally created by fragmenting large collections of diverse organic molecule structures. Initially, hundreds of thousands of fragments were obtained. From these, we created additional fragment sets by introducing wild cards or query features to retain original substitution patterns. Of all the fragments, we selected computationally 512 fragments, balancing two required criteria. First, they needed to be orthogonal, concerning the occurrence in diverse organic compounds, and second, their frequencies in those compounds had to be reasonably high. Both criteria ensure that these fragments are optimally suited for their original purpose, i.e., to achieve an optimal discrimination in the prescreening of a substructure search. Into the bargain, these fragments also proved to be a very valuable foundation for Tanimoto-based similarity calculations. As a result of the frequent usage of wild cards, calculated similarity values are more tolerant to single atom differences, and two molecules with similar skeletal backbones are still considered similar even if some of the atoms differ.

(26) Thomson Investigational Drugs Database (IDDB). http://scientific.thomson.com/products/iddb/ (accessed Sept 2004).

(27) Kohonen, T. *Self-Organizing Maps*. Springer: New York, 2001; p 30.

(28) Espinosa, G.; Arenas, A.; Giralt, F. An Integrated SOM-Fuzzy ARTMAP Neural System for the Evaluation of Toxicity. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 343−359.

(29) Gini, G.; Craciun, M.; Konig, C.; Benfenati, E. Combining unsupervised and supervised artificial neural networks to predictaquatic toxicity. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1897−902.

(30) Mazzatorta, P.; Vračko, M.; Jezierska, A.; Benfenati, E. Modeling Toxicity by Using Supervised Kohonen Neural Networks. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 485−492.

(31) Tenenbaum, J. B.; de Silva, V.; Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *Science* **2000**, *290*, 2319−2323.

(32) ChemBank DB: http://www.healthandsafety-centre.net/croner/jsp/groupDetails.do?contentId=137592&BV_UseBVCookie=Yes&channelId=-236221.

(33) *ChemOffice 2000 (ChemDraw, ChemBats3D, ChemFinder)*; CambridgeSoft Corp.: Cambridge, MA, 2000. http://www.cambridgesoft.com/.

(34) Witten, I. H.; Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann: San Francisco, CA, 2000.
(35) Payette, F. *ISOMAP Java*; Univeristy of Montreal: Montreal, Canada, 2002. http://www.iro.umontreal.ca/~payettf/index_en.html.
(36) Sander, T. *DataWarrior, a tool for visualizing data*; Actelion: Allschwil, Switzerland, 2000.
(37) Wang, J.; Lai, L.; Tang, Y. Structural Features of Toxic Chemicals for Specific Toxicity. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1173−1189.
(38) Bemis, G.; Murcko, M. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.
(39) Chittimoori, R. N. Discovering Substructures in Chemical Toxicity Domain. Masters Thesis, University of Texas, Arlington, TX, 1999.
(40) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, Germany, 2000.
(41) Sun, H. Prediction of Chemical Carcinogenicity from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **2004**, 44.
(42) Lobo, V. S.; Bandeira, N. F. C.; Moura-Pires, F. Distributed Kohonen Networks for Passive Sonar Based Classification. *Proceedings of the International Conference on Multisource-Multisensor Information Fusion (FUSION98)*, Las Vegas, NV, July 6−9, 1998.
(43) Schoonjans, F. Receiver Operating Characteristic (ROC) curve analysis. In *MedCalc for Windows, Statistics for Biomedical Research, Software Manual*; MedCalc Software: Mariakerke, Belgium, 2005; Vol. 8.1, pp 110−112.
(44) ROC curve analysis, introduction: http://www.medcalc.be/manual/mpage06−13a.php.
(45) Swamidass, S. J.; Chen, J.; Bruand, J.; Phung, P.; Ralaivola, L.; Baldi, P. Kernels for small molecules and the prediction of mutagenicity, toxicity and anti-cancer activity. *Bioinformatics* **2005**, *21*, i359−i368.
(46) Balasubramanian, M.; Schwartz, E. L. The Isomap Algorithm and Topological Stability. *Science* **2002**, *295*, 7.
(47) Ye, J. On Measuring and Correcting the Effects of Data Mining and Model Selection. *J. Am. Stat. Assoc.* **1998**, *93*, 120−131.

CI050358K