Toxicity Prediction using 2D Pharmacophores and Support Vector Machines

Max Pereira[1], **Ademar Schmitz**[1]

[1]Núcleo de Pesquisa e Desenvolvimento em Informática Médica
Universidade do Sul de Santa Catarina (UNISUL)
Tubarão – SC – Brasil

{max.pereira,ademar.schmitz}@unisul.br

Abstract. *In silico methods have been largely used in drug development process to predict the toxicity of molecules. Predicting the toxicity is one of the most important stage in developing new pharmaceuticals and computational methods are being used in order to make this process less time-consuming and decrease its high cost. Here we report a new approach, using two-dimensional pharmacophore fingerprint to encode pharmacophoric features of molecules in string sets, which are then processed by support vector machines (SVM) to predict the toxicity endpoint of a carcinogenic data set with 1547 compounds. Previous studies have shown the use of machine learning approaches in predicting the toxicity of molecules, however, in those cases it was required to calculate a large number of molecular descriptors to be able to make such prediction. Using SVM and only one molecular descriptor it was possible to achieve a satisfactory accuracy rate compared to other machine learning approaches.*

1. Introduction

Drug development is a high cost and time-consuming process which takes about 12 years and US$800 million to bring a drug from discovery to market [Schachter and Ramoni 2007, Suresh and Basu 2008]. This multi-step process investigates the promising compounds, traditionally, using *in vivo* methods for their pharmacokinetic properties, metabolism and potential toxicity [Ekins 2003]. With currently more than 80 000 chemicals in use it is truly necessary the use of computational methods to facilitate the access of this huge amount of data. The use of computers form the core structure-based drug design which increases the chance of success in the development process at a lower cost [Cronin 2001]. The potential toxicity of a compound is investigated in the so-called preclinical stage of the drug development process. Computational methods for predicting toxicity have been employed in the last years and *in silico* techniques as knowledge-based expert systems and (quantitative) structure-activity relationship (Q)SAR [Hansch et al. 1962] modelling approaches have therefore helped to significantly identifying adverse drug reactions in preclinical studies [Muster et al. 2008, Kavlock et al. 2008, Egan et al. 2004]. The increase of computational methods is mainly due to the emergence of new chemical descriptors and new algorithms and statistical perspectives in addition to the increase in the amount of available toxicity data [Benfenati 2007].

Here we report a SAR-based (structure-activity relationship) method for toxicity prediction, which is an approach close to QSAR, but actually in these models the relation

between a chemical property and the biological activity or effect is expressed in a qualitative way without assigning a quantitative value to the toxicity. This approach can identify molecular functionalities (features) that are known to cause toxicity [Kruhlak et al. 2007], therefore the presence of these features means a potential toxicity in the compound. Thus, success using *in silico* techniques would increase efficiency end effectiveness in determining the hazards of the many compounds that must be dealt with, so the need for more sophisticated tools and models for toxicological prediction. In order to accomplish this, it is indispensable an efficient mathematical algorithm and an appropriate way of describing the chemical structure [Rabinowitz et al. 2008].

Some commercially toxicity prediction programs are available including TOPKAT (toxicity-prediction by computer-assisted technology), DEREK (deductive estimation of risk from existing knowledge), CSGenoTox, MetaDrug and HazardExpert [Ekins 2007]. These programs have a common characteristic, they are classified as "global" models [White et al. 2003] since they were developed using a non-congeneric set of chemicals. Actually it is not necessary for the chemicals in these data sets to be congeneric, but they should share structural features. Besides the commercially available programs, another studies concern predictive toxicology have been published using machine learning approaches [Tiwari et al. 2006, Kazius et al. 2005, Amini et al. 2007, Neagu et al. 2005, Dearden 2003].

In this paper, a rapid method for predict carcinogenic compounds that utilizes support vector machines (SVMs) is presented. Although a number of applications of SVM to computational chemistry have been published [Zhao et al. 2006, Jiang et al. 2006, Yap et al. 2006] and similar toxicological prediction studies have been reported as well, there is plenty of room for improvement. The ability of the predicting models is determined primarily by the choice of descriptors that represent the compounds. Thus, we present a simplified method using only 2D pharmacophore fingerprints descriptors to represent compounds and to predict their toxicity endpoint. We have applied the SVM to predict the toxicity of compounds using the carcinogenic potency database (DSSTox CPDBAS Database). The method described here has utility in preclinical stage of drug development.

The rest of the paper is structured as follows. In section 2, we briefly describe molecule pharmacophoric properties and introduce the SVM classifier. We present the method to predict the carcinogenic toxicity in section 3. Numerical test results and discussions can be found in section 4 before de conclusion in section 5.

2. Background

2.1. Two-dimension Pharmacophores

A pharmacophore is a set of structural features in a molecule which represents the interactions between small molecule ligands and a protein receptor, and therefore, is responsible for that molecule's biological activity [McGregor and Muskal 1999]. The hydrogen bonding, charge-charge and hydrophobic interactions are typical in molecules. Thus, the identification of these pharmacophore features requires a chemical structure analysis to determine the presence of some structures, such as: hydrophobicity, aromaticity, a hydrogen bond acceptor/donor and whether cationic or anionic. In order to use this structures in computational models pharmacophore fingerprints were proposed

[Bonachera et al. 2006]. These are pharmacophore models that represent each structure as a point and the relation between these structures as a point pair. In the simplest case, is the three-dimensional Euclidean distance between each point pair or in the two-dimensional case topological relations are used to represent the relative position of pharmacophore points [McGregor and Muskal 1999]. In a 2D atom-based pharmacophore fingerprint representation [Varin et al. 2008], the pharmacophoric point pairs are calculated and each pattern of the fingerprint corresponds to the shortest path between the nodes (atoms-features) of the chemical graph, and the pharmacophore properties of molecules are encoded as frequency counts of these point pairs[1].

2.2. Support Vector Machines

The support vector machines (SVM) is a binary classification technique developed by Vapnik [Vapnik and Cortes 1995]. SVM has become very popular because of its excellent generalization capacity. The main advantage of SVM is the structure risk minimization (SRM) principle, which has been shown to be superior to the traditional empirical risk minimization (ERM) principle, employed by conventional neural networks [Zhao et al. 2006, Jiang et al. 2006]. SVM algorithm attempts to establish a boundary using support vectors (examples in the training) and ignores those examples that are outside the boundary which differs from the neural nets which seek to minimize the errors over the entire training set [Jiang et al. 2006]. SVM has been developed and applied to classification problems, regression problems and time-series estimation [Yap et al. 2006]. Given the training data $\{(x_i, y_i)|y_i = 1 \text{ or } -1, i = 1, ..., N\}$ for a two-class classification, where N is the number of examples; x_i is the input vector and y_i is the class. The decision surface is created by SVM with:

$$f(x) = sign\Big(\sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b\Big), \tag{1}$$

where K is the kernel function that define the feature space, b is the bias value, α_i gives the maximum margin hyperplane by the interval $[0 \leq \alpha_i \leq C]$ where C controls the values between maximizing the margin and minimizing the training error. The kernel function can be linear, polynomial or Gaussian.

3. Method

In this paper we used a data set from the Distributed Structure-Searchable Toxicity (DSSTox) Public Database Network [Richard and Williams 2002] from the U.S.Environmental Protection Agency [2]. The DSSTox database project is targeted to toxicology study areas, with standardized chemical structure annotation.

CPDB: The Carcinogenic Potency Database (CPDB) contains detailed results and analyses of 6540 chronic, long term carcinogenesis bioassays [Gold et al. 2001], which currently contains 1547 compounds. For the purpose of this study the carcinogenicity endpoint was evaluated concerning hamster, mouse and rat species. The experimental results about the remaining species (cynomolgus, dog, rhesus) are insufficient, therefore, they were discarded. In the CPD database when the same compound was tested in more

[1] www.chemaxon.com

[2] http://www.epa.gov/ncct/dsstox/index.html, accessed Dez 2008

than one specie, the experiment results for all these species were stored in a single entry. Thus, for this study, the database was preprocessed in order to split a single entry as many as necessary, accordingly to the number of species. We also discarded the entries with inconclusive results, which actually represented only six examples. We have then 2272 total entries, divided in two classes - active with 1059 examples and inactive with 1213 examples. The structures of chemicals in DSSTox are stored as SDF files as well as SMILES strings.

Molecular descriptors fall into some general categories [Ekins 2007, Todeschini et al. 2000] and, for the purpose of this study, we have chosen a topological-based descriptor (pharmacophore fingerprint). The pharmacophore fingerprint molecular descriptor, for all molecules, were calculated using the GenerateMD software [3].

For each entry (compound) in CPDB data set two additional attributes were generated, for instance, for the molecule in Figure 1 we have the following informations (Table 1).

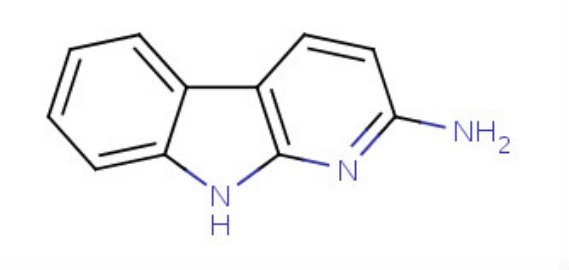

Figure 1. Molecule Structure

Table 1. Molecular Descriptors

Molecular Descriptor	Value
PMAP	d;r;r;r;r;r;a/r;d/r;r;r;r;r;r;r;
PF	d a = 0 2 0 0 0 0 0 0 0 0 , d d = 0 0 0 1 0 0 0 0 0 0 , r a = 2 3 3 2 2 0 0 0 0 0 , r d = 3 6 6 4 2 2 2 0 0 0 , r r = 15 21 19 13 7 3 0 0 0 0

PMAP stands for Pharmacophore Mapping, all molecule's pharmacophoric features present in the molecule were calculated, identified and stored in this attribute, Table 2 describes all possible pharmacophoric informations. PF stands for Pharmacophore Fingerprint, i.e., the frequency and the path length of the pharmacophoric features point pairs, which were all encoded in this strings set. As the pharmacophore fingerprint is an atom-based descriptor, thus each entry, in the strings, was assigned to an attribute. We have so 232 attributes, i.e, all point pairs combinations multiplied by the number of the maximum frequency (10), plus the toxicity information attribute (ActivityOutcome). Based on the toxicity information (active/inactive) available in the CPDB data set and on these pharmacophoric data we can then construct a new data set format with the required examples to be processed by the SVM algorithm.

[3] http://www.chemaxon.com, accessed Nov 2008

Table 2. Pharmacophoric Features

Symbol	Description
-	anionic
+	cationic
a	hydrogen bond acceptor
d	hydrogen bond donor
h	hydrophobic
r	aromatic

For the classification process the data set of 2272 molecules was randomly divided into ten folds with 227-228 molecules in each fold. One fold was used as a testing set, and the nine others folds were for training. Calculations were repeated ten times for a 10-fold cross validation.

4. Results and Discussion

The data set was trained using Gaussian Kernel with a C (trade-off between training error and margin) ranging from 1 to 100.

$$K(x,z) = esp\Big(-\frac{\|x-z\|^2}{2\sigma^2}\Big), \tag{2}$$

we had then two parameters to be specified, σ and C. The selection was made by training for different values of these parameters to identify those which minimized an upper bound on the expected generalization error. If C has a low value then insufficient emphasis will be placed on fitting the training data. On the other hand, if C has a great value it might overfit the training data. In spite of this to make the learning process stable a great value should be used initially. The optimal value of C was 90.

The speed of training, number of support vectors, training and test accuracy, precision and recall were analysed. In spite of the fact that all informations were observed, due to space limit we reported only the best results. The details of predicted values are given in Table 3. According to Table3, the recall presented a large value, which means that a great number of positives (active examples) have been predicted correctly. It is important to mention that these results are better than those obtained by Chemical Descriptors methods, which use various chemical descriptors such as LOGP, LUMO and dipole moments to model toxicity of compounds [Enslein et al. 1994]. Other studies about toxicity prediction have been reported in the recent last years [Cronin 2001, Amini et al. 2007, Dearden 2003, White et al. 2003, Jiang et al. 2006] and their results are very closed to ours. However, in most of the cases it is necessary to calculate a large number of molecular descriptors for the properly toxicity prediction. In this case, only one molecular descriptor is sufficient, which is present in all molecules whatever the toxicity endpoint is. That means no matter what toxicity data set is being used it is possible to calculate the pharmacophore fingerprint to predict the toxicity activity.

Table 3. Prediction of Active and Inactive Toxicity Classes

TP^1	FN^2	TN^3	FP^4	$Recall^5$	$Precision^6$	$Accuracy^7$
803	256	986	227	75.82%	77.96%	78.70%

[1]True Positives. [2]False Negatives. [3]True Negative. [4]False Positive. [5]TP/(TP + FN). [6]TP/(TP + FP). [7](TP + TN)/(TP + TN + FP + FN).

After knowing the results an attribute selection procedure was applied in order to identify those input attributes that were most relevant to toxicity prediction.A ranker method listed all 231 attributes from the most relevant to the less relevant. Table 4 lists the top four features (attributes) and the Figure 2 illustrates the molecules which shows pharmacophoric features represented in these attributes.

Table 4. Ranked Attributes

Attribute	Path (between point pairs)
a +	1
r h	7
r r	2
r r	7

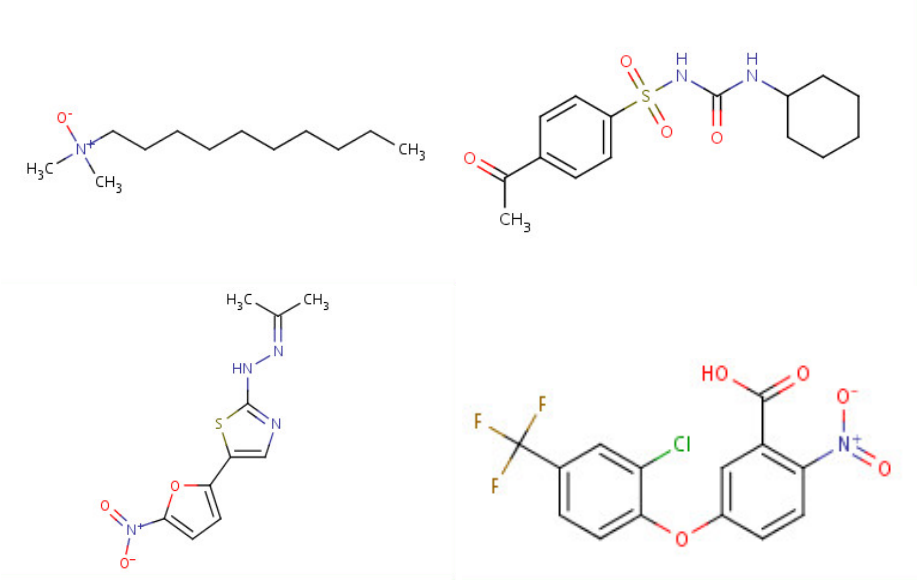

Figure 2. Graphical Representation of Pharmacophoric Features

5. Conclusion

Developing predictive classifiers for biological data sets is a great challenge. Although many works have been done there is room for improvements. We have proposed a new method for toxicity prediction based only on the two-dimensional pharmacophore features. There is no need to calculate many molecular descriptors and deal with other structure descriptions like atom bonds either. It seems obviously that this method must be

improved in order to achieve higher accuracy rates. However, we can do a more detailed feature selection to identify and select those features which are more relevant for the toxicity prediction in a molecule concern these pharmacophoric informations. During the last years a huge number of new molecular descriptors have been calculated to help predicting toxicity of candidate drugs. These new molecular descriptors in fact have helped in this complex process, on the other hand, they have increased the time processing and also increased the biological database's size. Maybe it would be better to investigate further the already known descriptors in order to find out new ways to represent and manipulate them.

References

Amini, A., Muggleton, S., Lodhi, H., and Sternberg, M. (2007). A novel logic-based approach for quantitative toxicology prediction. *J. Chem. Inf. Model.*, 47(3):998–1006.

Benfenati, E. (2007). Predicting toxicity through computers: A changing world. *Chemical Central Journal*, 32(1).

Bonachera, F., Parent, B., Barbosa, F., Froloff, N., and Horvath, D. (2006). Fuzzy tricentric pharmacophore fingerprints. *Journal of Chemical Information and Modeling*, 6(46):2457–2477.

Cronin, M. (2001). Prediction of drug toxicity. *Il Farmaco*, pages 149–151.

Dearden, J. (2003). In silico prediction of drug toxicity. *Journal of computer-aided molecular design*, 17(2-4):119–127.

Egan, W., Zlokarnik, G., and Grootenhuis, P. (2004). In *silico* prediction of drug safety: despite progress there is abundant room for improvement. *Drug Discovery Today: Technologies*, 1(4):381–387.

Ekins, S. (2003). In *silico* approaches to predicting drug metabolism, toxicology and beyond. *Biochemical Society Transactions*, 31(3):611–614.

Ekins, S. (2007). *Computational Toxicology: Risk Assessment for Pharmaceutical and Environmental Chemicals (Wiley Series on Technologies for the Pharmaceutical Industry)*. Wiley-Interscience.

Enslein, K., Gombar, V., and Blake, B. (1994). Use of sar in computer-assisted prediction of carcinogenicity and mutagenicity of chemicals by the topkat program. *Mutat. Res.*, (305):47–61.

Gold, L., Manley, N., Slone, T., and Ward, J. (2001). Compendium of chemical carcinogens by target organ: Results of chronic bioassays in rats, mice, hamsters, dogs, and monkeys. *Toxicologic Pathology*, 29(6):639–652(14).

Hansch, C., Maloney, P., Fujita, T., and Muir, R. (1962). Correlation of biological activity of phenoxyacetic acids with hammett substituent constants and partition coefficients. *Nature*, (194):178–180.

Jiang, Z., Yamauchi, K., Yoshioka, K., Aoki, K., Kuroyanagi, S., Iwata, A., Yang, J., and Wang, K. (2006). Support vector machine-based feature selection for classification of liver fibrosis grande in chronic hepatitis c. *J. Med. Syst.*, (30):389–394.

Kavlock, R., Ankley, G., Blancato, J., Breen, M., Conolly, R., Dix, D., Houck, K., Hubal, E., Judson, R., Rabinowitz, J., Richard, A., Setzer, R., Shah, I., Villeneuve, D., and Weber, E. (2008). Computational toxicology - a state of the science mini review. *Toxicological Sciences*, 103(1):14–27.

Kazius, J., Mcguire, R., and Bursi, R. (2005). Derivation and validation of toxicophores for mutagenicity prediction. *J. Med. Chem.*, 48(1):312–320.

Kruhlak, N., Contrera, J., Benz, R., and Matthews, E. (2007). Progress in qsar toxicity screening of pharmaceutical impurities and other fda regulated products. *Advanced Drug Delivery Reviews*, 59:43–55.

McGregor, M. and Muskal, S. (1999). Pharmacophore fingerprinting: Application to qsar and focused library design. *J. Chem. Inf. Comput. Sci.*, 39(3):569–574.

Muster, W., Breidenbach, A., Fischer, H., Kirchner, S., Müller, L., and Pähler, A. (2008). Computational toxicology in drug development. *Drug Discovery Today*, 8(7).

Neagu, D., Craciun, M., Stroia, S., and Bumbaru, S. (2005). Hybrid intelligent systems for predictive toxicology - a distributed approach. *Intelligent Systems Design and Applications, International Conference on*, pages 26–31.

Rabinowitz, J., Goldsmith, M., Little, S., and Pasquinelli, M. (2008). Computational molecular modeling for evaluating the toxicity of environmental chemicals: Prioritizing bioassay requirements. *Environmental Health Perspectives*, 116(5):573–577.

Richard, A. and Williams, C. (2002). Distributed structure-searchable toxicity (dsstox) public database network: a proposal. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 499:27–52(26).

Schachter, A. and Ramoni, M. (2007). Clinical forecasting in drug development. *Nature Reviews*, 6:107–108.

Suresh, P. and Basu, P. (2008). Improving pharmaceutical product development and manufacturing: Impact on cost of drug development and cost of goods sold of pharmaceuticals. *Journal of Pharmaceutical Innovation*, 3(3):175–187.

Tiwari, A., Knowles, J., Avineri, E., Dahal, K., and Roy, R., editors (2006). *Advances in the Application of Machine Learning Techniques in Drug Discovery, Design and Development*, Advances in Soft Computing. Springer.

Todeschini, R., Consonni, V., Mannhold, R., Kubinyi, H., and Timmerman, H. (2000). *Handbook of Molecular Descriptors*. Wiley-VCH.

Vapnik, V. and Cortes, C. (1995). Support-vectors networks. *Machine Learning*, 20:273–297.

Varin, T., Saettel, N., Villain, J., Lesnard, A., Dauphin, F., Bureau, R., and Rault, S. (2008). 3d pharmacophore, hierarchical methods, and 5-ht4 receptor binding data. *Journal of Enzyme Inhibtion and Medicinal Chemistry*, 23(5):593–603.

White, A., Mueller, R., Gallavan, R., Aaron, S., and Wilson, A. (2003). A multiple in silico program approach for the prediction of mutagenicity from chemical structure. *Mutation Research/Genetic Toxicology and Environmental Mutagenesis*, 539:77–89(13).

Yap, C., Xue, Y., Li, Z., and Chen, Y. (2006). Application of support vector machines to *in Silico* prediction of cytochrome p450 enzyme substrates and inhibitors. *Current Topics in Medicinal Chemistry*, (6):1593–1607.

Zhao, C., Zhang, H., Zhang, X., Liu, M., Hu, Z., and Fan, B. (2006). Application of support vector machine (svm) for prediction toxic activity of different data sets. *Toxicology*, (217):105–119.