Comparison of Linear and Nonlinear Classification Algorithms for the Prediction of Drug and Chemical Metabolism by Human UDP-Glucuronosyltransferase Isoforms

Michael J. Sorich,[†] John O. Miners,*[,‡] Ross A. McKinnon,[†] David A. Winkler,[§] Frank R. Burden,[||] and Paul A. Smith[‡]

University of South Australia, Adelaide, South Australia, Australia, Department of Clinical Pharmacology, Flinders University and Flinders Medical Centre, Bedford Park, South Australia, Australia, Division of Molecular Science, CSIRO, Clayton, Victoria, Australia, and School of Chemistry, Monash University, Clayton, Victoria, Australia

Received May 29, 2003

Partial least squares discriminant analysis (PLSDA), Bayesian regularized artificial neural network (BRANN), and support vector machine (SVM) methodologies were compared by their ability to classify substrates and nonsubstrates of 12 isoforms of human UDP-glucuronosyltransferase (UGT), an enzyme "superfamily" involved in the metabolism of drugs, nondrug xenobiotics, and endogenous compounds. Simple two-dimensional descriptors were used to capture chemical information. For each data set, 70% of the data were used for training, and the remainder were used to assess the generalization performance. In general, the SVM methodology was able to produce models with the best predictive performance, followed by BRANN and then PLSDA. However, a small number of data sets showed either equivalent or better predictability using PLSDA, which may indicate relatively linear relationships in these data sets. All SVM models showed predictive ability (>60% of test set predicted correctly) and five out of the 12 test sets showed excellent prediction (>80% prediction accuracy). These models represent the first use of pattern recognition methods to discriminate between substrates and nonsubstrates of human drug metabolizing enzymes and the first thorough assessment of three classification algorithms using multiple metabolic data sets.

1. INTRODUCTION

The advent of combinatorial chemistry, large chemical libraries, and high-throughput screening methodologies has resulted in the generation of large quantities of qualitative biological data (e.g. active or inactive) for diverse chemicals. To analyze these types of data, classification pattern recognition methods capable of developing models of maximal generalization ability (i.e., for predicting new structurally diverse chemicals correctly) from large and generally noisy data sets are required. There are many pattern recognition methods suitable for classification; three of the most commonly used are partial least squares discriminant analysis (PLSDA),[1] artificial neural networks (ANN),[2] and support vector machines (SVM).[3]

These methods are widely recognized for their ability to generalize well from underdetermined data sets, in which there are more descriptors than chemicals. PLSDA is a variant of partial least squares regression, a commonly used method in quantitative structure−activity relationship (QSAR) modeling. It is a linear technique, and thus determination of the relative importance of descriptors is possible. ANN are capable of modeling nonlinear relationships, including where the form of nonlinearity is unknown a priori. While ANN are commonly used to generate QSAR models, overtraining and overfitting limit their usefulness. A modified form of ANN, the Bayesian Regularized Artificial Neural Network (BRANN), is more resistant to overfitting and overtraining than conventional ANN, resulting in improved generalization performance.[2] As with ANN, the SVM are universal function approximators. Unlike ANN, however, there is only one minimum in the optimization problem, resulting in the rapid generation of a unique solution. There are very few reported applications of the SVM methodology in drug discovery applications[4,5] and none that compare the SVM to BRANN or PLSDA. Some methods will clearly perform better than others, depending on the degree nonlinearity in the relationship, and consequently only a comparison involving a series of data sets will be able to give reliable indications of the general utility of the classification method.

The enzyme UDP-glucuronosyltransferase (UGT) catalyzes the covalent linkage (i.e., "conjugation") of glucuronic acid, derived from the cofactor UDP-glucuronic acid, to a typically lipophilic substrate bearing a suitable "acceptor" functional group according to a second-order nucleophilic substitution mechanism. Glucuronidation serves as a clearance mechanism for drugs from all therapeutic classes.[6] Additionally, glucuronidation provides an elimination pathway for numerous endogenous compounds, dietary chemicals, and environmental pollutants (including some chemical carcinogens) and aids excretion of the products of phase I metabolism. Endogenous compounds metabolized by glucuronidation include bilirubin, bile acids, fatty acids, hydroxysteroids, and thyroid hormones. Consistent with this

* Corresponding author phone: 61-8-82044131; fax: 61-8-82045114; e-mail: john.miners@flinders.edu.au. Corresponding author address: Department of Clinical Pharmacology, Flinders Medical Centre, Bedford Park, SA 5042, Australia.

[†] University of South Australia.

[‡] Flinders University and Flinders Medical Centre.

[§] CSIRO.

[||] Monash University.

10.1021/ci034108k CCC: $25.00
Published on Web 09/06/2003

substrate diversity, UGT comprises a superfamily of enzymes ("isoforms"). UGTs characterized to date have been classified into two families, *UGT1* and *UGT2*, based on amino acid identity and evolutionary divergence.[7] Substrate selectivity has been documented reasonably well for 12 of the 16 functional human isoforms: UGT 1A1, 1A3, 1A4, 1A6, 1A7, 1A8, 1A9, 1A10, 2B4, 2B7, 2B15, and 2B17. The UGT isoforms exhibit distinct, but overlapping substrate selectivity and differ in terms of regulation, incidence and frequency of genetic polymorphism, and patterns of drug−drug interactions.[8,9]

Reaction phenotyping involves identification of the isoform(s) responsible for the metabolism of a given drug or chemical. Together with an understanding of isoform regulation, pharmacogenetics, and drug interactions, reaction phenotyping allows prediction of factors likely to alter drug metabolic clearance (and hence response) in vivo. Currently, in vitro assays are used for reaction phenotyping. However, the low throughput and relatively high costs of in vitro procedures restrict the availability of appropriate data to later stages of the lead development process. In silico methods provide the most promising approach to overcome these problems and thereby allow earlier assessment of drug metabolism.

Recently, the ability of various methodologies (pharmacophore modeling, molecular field based QSAR and 2D-QSAR) to predict the binding ability (competitive inhibition constant or Michaelis constant) of UGT1A1 and UGT1A4 has been reported by this laboratory.[10,11] These analyses demonstrated that 2D-QSAR is the best method for such predictions. However, as a result of the complexity of the relationships, it was apparent that a larger data set would be required in order to predict the substrate selectivity of novel chemicals with confidence. As noted previously, existing in vitro methods for the experimental determination of substrate binding affinity are expensive and time-consuming. In comparison, screening simply whether a chemical can be metabolized by a UGT isoform is faster and cheaper. Consequently, there is considerable data available in the public domain defining the ability of chemicals to be metabolized by individual human UGT isoforms.

Thus the work reported here aimed to (1) determine the feasibility of predicting human UGT isoform specific chemical metabolism (reaction phenotyping) using standard two-dimensional (2D) chemical properties/descriptors and (2) compare the generalization performance of the PLSDA, BRANN, and SVM methodologies for the classification of metabolic data.

2. DATA AND METHODS

2.1. Data Sets. All data in the public domain relating to chemicals tested for metabolism by individual UGT isoforms were collated. Data from 100 publications containing experimental results of chemicals tested in recombinant cell systems expressing a single UGT isoform were compiled in an Access Database (Microsoft Corporation, WA). Isoform-specific data sets were collated in which the chemicals tested for activity were classified as either substrates or nonsubstrates. Twelve data sets, each containing more than 50 chemicals, were generated (Table 1). Chemicals tested on more than one occasion against the same isoform that showed conflicting results were not included in the data sets unless the reason for the discrepancy could be determined.

Table 1. Composition of the Human UGT Isoform Substrate Selectivity Data Sets Used for Classification Analyses

	number of chemicals	percent substrates (%)
1A1	205	38
1A3	178	72
1A4	181	55
1A6	195	38
1A7	69	38
1A8	115	76
1A9	216	63
1A10	156	49
2B4	140	29
2B7	213	64
2B15	141	40
2B17	55	44

Table 2. Molecular Descriptors Used in the Human UGT Isoform Substrate Classification Analyses

atomistic[13]	counts of atom types (numbers after atom symbol represent the number of connections to other atoms): H, C2, C3, C4, C(Aromatic), N1, N2, N3, N4, N(Aromatic), O1, O2, O(Aromatic), F, Si2, Si3, Si4, P2, P3, P4, P5, S1, S2, S3, S4, S(Aromatic), Cl, Br, I, H donor, H acceptor
rings	counts of rings of size 3−8
fragments	count of fragments: H−O−C, H−O−N, C−O−C, N−O−C,N−O−N, C═O, O═C−N, O═C−O, N═O, O═N═O
eigenvalue descriptors[14]	the 10 highest absolute eigenvalues of a modified adjacency matrix of the molecule
connectivity indices[15]	vertex degree and valence vertex degree connectivity indices of path lengths 0−4

2.2. Chemical Descriptors. Chemical structures for the 523 chemicals in the 12 data sets were constructed using ChemDraw (CambridgeSoft, MA). Sixty-seven 2D chemical descriptors (Table 2) were calculated using in-house software written in Matlab (MathsWorks Inc., MA). The Unsupervised Forward Selection (UFS) algorithm[12] was used to select a subset of descriptors for each data set such that redundancy was eliminated and multicollinearity was reduced. This method initially selects the two descriptors which are least well correlated and then additional variables on the basis of their multiple correlation with those already chosen, resulting in a subset of variables that are as close to orthogonal as possible. Subsets of descriptors were selected using this procedure with the R^2_{max} parameter set to 0.99. These descriptors were normalized to have a mean of zero and unit variance prior to generation of classification models.

2.3. Pattern Recognition Methods. 2.3.1. Partial Least Squares Discriminant Analysis (PLSDA). PLS is designed to deal with collinearity among the independent variables.[1] It is broadly similar to principal component regression but with both the independent and dependent variables involved in the generation of the orthogonal latent variables rather than only independent variables. PLS is an iterative algorithm with consecutive estimates obtained using the residuals from previous iterations as the new dependent variable.[1] Each iteration of the algorithm introduces another latent variable, and leave-30%-out cross-validation (the average prediction accuracy of 30% of the training set data, left out in 30 different ways) was used to determine the optimal number of latent variables.

The PLS methodology was implemented in the Python scripting language (www.python.org) using the SAMPLS

algorithm.[16] The number of latent variables was chosen to maximize the training set cross-validation percent predicted correct (i.e., the test set was not used to select the number of latent variables).

2.3.2. ν-Support Vector Machine (ν-SVM). Overfitting of data can be avoided by limiting the complexity of the models that the method can possibly generate. A specific approach for controlling the complexity of the models is given by the Vapnik-Chervonenkis (VC) theory and the structural risk minimization principle.[17] This is applied to the training of a classification SVM by fitting of a hyperplane such that the largest margin is formed between two classes of chemicals while minimizing the classification errors. Nonlinearity in a data set is accounted for with kernel functions, which map the input vectors to some higher dimensional space such that a hyperplane with reduced classification errors can be found.[3] A major advantage is that optimization problems resulting from SVMs have a global minimum and can be solved with standard quadratic programming tools. A recent improvement in the SVM algorithm allows for a more sensible choice of the regularization parameter.[18] The new parameter, ν, represents an upper bound on the fraction of errors (fraction of chemicals misclassified) for a classification problem and can be chosen depending on the inherent error in the data.

The SVM models were generated using the LIBSVM implementation[19] of the ν-SVM algorithm.[18] For all the data sets the radial basis function kernel was used with the default value of the gamma parameter (=1/number of descriptors) and the ν parameter set to 0.1.

2.3.3. Bayesian Regularized Artificial Neural Network (BRANN). The Bayesian framework for neural networks is based on a probabilistic interpretation of network training. In contrast to conventional network training, where an optimal set of weights is chosen by minimizing an error function, the Bayesian approach involves a probability distribution of network weights.[20] As a result, the predictions of the network are also probability distributions. Most importantly, complex models are penalized in the Bayesian approach, reducing the problems of overfitting and overtraining.

The Netlab toolbox[21] of Matlab was used to generate BRANN models. All networks were fully connected, employed sigmoidal transfer functions (in both hidden and output layers), and contained one hidden layer with six nodes. A separate inverse variance hyperparameter was employed for each group of weights (inputs, input bias, outputs, output bias). The scaled conjugate gradient algorithm was used to train the network. The hyperparameters were re-estimated after each 100 iterations. There were eight cycles of the whole algorithm.

2.4. Assessment of Generalization Performance. To assess the ability of the three pattern recognition methods to predict new chemicals (i.e., generalization performance), 30% of each data set was randomly chosen to be the test set using a random number generator implemented in Python. The remaining 70% of chemicals were used to generate the models with the three pattern recognition methods. The test set was not used in any way to influence the training and selection of models. The test set was predicted and compared against the known experimental results only after the models were defined completely. The generalization ability of the models was expressed as the percent of test set chemicals that were correctly predicted (both substrates and nonsubstrates).

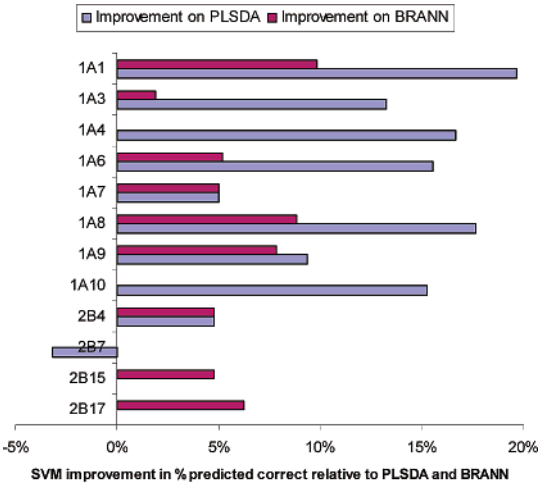

Figure 1. Improvement of SVM over other methods for each data set. Note that for UGT1A4, UGT1A10, and UGT2B7, the SVM model was equivalent to the BRANN model, and for UGT2B15 and UGT2B17, the SVM model was equivalent to the PLSDA model.

Table 3. Percent of Test Set Predicted Correctly for All UGT Substrate Data Sets by Each Classification Method

	SVM	PLSDA	BRANN
1A1	72	52	62
1A3	81	68	79
1A4	83	67	83
1A6	81	66	76
1A7	75	70	70
1A8	65	47	56
1A9	77	67	69
1A10	80	65	80
2B4	88	83	83
2B7	63	67	63
2B15	69	69	64
2B17	75	75	69

3. RESULTS AND DISCUSSION

Table 3 shows the percent test set predicted correctly for each data set using the three different classification methods. In all data sets other than UGT2B7, the SVM generated a model with either equivalent or better predictability as judged by the test set. This observation is highlighted graphically in Figure 1. It is to be expected that in data sets containing significant nonlinearity the SVM and BRANN algorithms would improve on PLSDA. In these cases the SVM is able to predict up to 20% more of the test set correctly. Similarly, the BRANN produces models able to predict 10% more of the test set correctly, compared to PLSDA. The UGT2B7, UGT2B15, and UGT2B17 data sets are modeled well by PLSDA, possibly indicating a relatively linear relationship between the chemical descriptors and chemical lability for metabolism by UGT.

As shown in Table 3 and displayed graphically in Figure 2, the SVM is best able to predict the test sets overall. Unless there is prior indication that the relationships involved will be predominantly linear in nature, it appears that the SVM provides the most appropriate approach for similar classification problems. The SVM algorithm is computationally

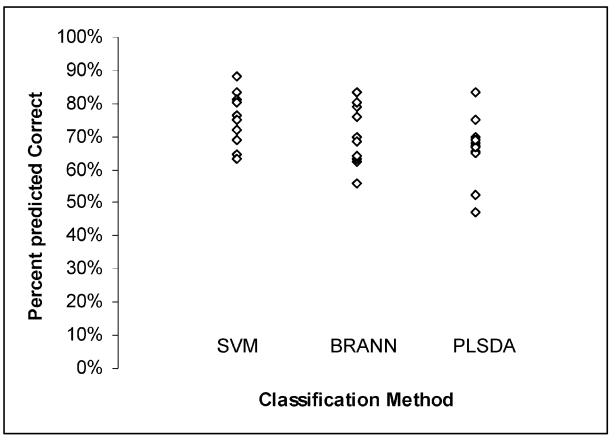

Figure 2. Distribution of generalization performance for the three classification methods. For each classification method there are 12 points, each representing the percent of test set predicted correctly for a UGT isoform.

efficient to train using data sets of this size and, unlike an ANN, determines a unique solution. It is possible, however, that as the size of the data sets increases, the BRANN performance may improve relative to that of SVM.

A BRANN has previously been compared to the PLS methodology for regression problems. Three data sets ranging in size between 55 and 245 chemicals were analyzed.[22] For all three data sets, the BRANN produced significantly better test set predictivity. Results presented here are in general agreement with these previous observations. A SVM has also been compared previously to a conventional ANN and to linear discriminative analysis (LDA) for several biological and simulated classification data sets.[4] The study showed the SVM to slightly outperform the ANN, which in turn slightly outperformed the LDA. The classification methods are similar but not identical to those reported here. The ANN was not Bayesian regularized, and thus its performance may have been suboptimal due to overfitting and/or overtraining. It would be expected that LDA would produce very similar results to PLSDA when there are many more chemicals than descriptors in the data set and there is little redundancy in the descriptors. However, in many QSAR studies this will not be the case and PLSDA is likely to perform better than LDA. A second difference between the studies is the number of data sets tested. In the previous study three simulated and two biological data sets were examined.[4] In this study, 12 biological data sets were used to compare the three classification methods. The desirability of a larger number of data sets is apparent from Figure 1. Large differences in the generalization performance of the classification methods were apparent for some data sets, while there was little difference in others. Clearly, conclusions from the analysis of a small number of data sets may be misleading. Consequently, the results presented here provide a much stronger indication of the performance of the various classification methods in real systems of relevance to drug discovery. It should be noted that covalent bonds are formed and broken in the glucuronidation reaction, distinguishing our data from that published previously.

The SVM and conventional ANN methodologies have also been compared for the classification of 179 molecules capable of penetrating the CNS and 145 molecules that were unable to do so.[5] The SVM (82% predicted correct) was able to classify the molecules better than the ANN (76% predicted correct). While the comparison of classification methods is limited to one data set, the result is consistent with the outcomes reported here.

It is apparent from Figure 2 and Table 3 that the general predictivity displayed by the models generated with the SVM algorithm is very good. It should be noted, however, that the experimental data have inherent noise or error due to measurement error and variation in experimental conditions between laboratories, which will limit the maximum possible predictivity of the models. Since some chemicals were tested for glucuronidation on more than one occasion, it was possible to estimate that data conflicted for about 10% of chemicals in each data set. Thus, experimental error would appear to limit the maximum possible predictability to approximately 90%. All SVM models showed predictive ability (>60% of test set predicted correctly), and five out of the 12 test sets showed excellent prediction with >80% prediction accuracy, demonstrating that standard descriptors trained with a SVM can account for a large amount of the variation.

The 67 descriptors (comprising counts of different atomic features, eigenvalue descriptors, and connectivity indices) used in this study to classify substrates and nonsubstrates were chosen on the basis of simplicity, ease of calculation, and diverse representation of chemical properties. More complex descriptors, such as those based on quantum chemical properties and/or 3D chemical structure were not used, primarily because they were not required. As described previously, the models generated with the simple descriptors approached the maximum possible predictivity given the inherent error in the data. It was thought that incorporation of additional descriptors was unlikely to significantly improve predictivity but would significantly increase the time required to predict isoform selectivity, due to the increased computational demand imposed by the calculation of quantum chemical properties and/or optimization of molecular geometry. Considering the complexity of the event being modeled (binding of a chemical to an enzyme active site in a mode where the reactive site of the molecule is aligned suitably with the bound cofactor), it is perhaps surprising that such generic and simple chemical properties are so predictive. However, these simple descriptors have been shown previously to encode subtle complexity[23] and have been used successfully in the past to predict diverse data sets.[22,24,25] Nevertheless, the models generated here are difficult to interpret because the connectivity indices and eigenvalue descriptors lack a simple physicochemical interpretation. In addition, the models generated using the SVM are nonlinear and hence separating the contribution of each descriptor in a straightforward and meaningful way is not feasible. Since it is generally accepted that UGT-catalyzed conjugation proceeds according to a second-order nucleophilic substitution (S_N2) reaction,[8] electronic properties describing nucleophilicity may prove predictive of metabolism by a UGT; however, this awaits investigation.

Table 4 illustrates that, for some isoforms, there is a significant difference in the percentage of substrates versus nonsubstrates predicted correctly. For example, 92% of the UGT1A3 substrates are predicted correctly, whereas only 50% of the nonsubstrates are predicted correctly. The importance of the various types of predictions (overall vs

Table 4. Percentage of Substrates and Nonsubstrates Predicted Correctly Using the SVM Algorithm[a]

	% correct		
	whole test set (%)	substrates (%)	nonsubstrates (%)
1A1	72	63	78
1A3	81	92	50
1A4	83	83	84
1A6	81	78	83
1A7	75	43	92
1A8	65	82	33
1A9	77	73	83
1A10	80	81	80
2B4	88	83	90
2B7	63	69	50
2B15	69	73	67
2B17	75	80	67

[a] Data are presented as the percentage of each test set predicted correctly and the percentage of substrates and nonsubstrates in the test sets predicted correctly.

substrate vs nonsubstrate) is application dependent. In certain situations it may be more important to ensure all substrates are found, and the consequential increase in false positives (nonsubstrates predicted to be substrates) is of lesser importance. This bias can be achieved by changing the weighting of substrates versus nonsubstrates in the training process. By default, however, the pattern recognition methods will generally attempt to maximize the overall number of training chemicals classified correctly. Thus, data sets trained with more substrates than nonsubstrates will predict substrates better than the nonsubstrates. This is evident from Tables 1 and 4. Throughout the 12 data sets there exists a strong correlation ($r = 0.77$) between the percentage of chemicals that are substrates and the relative predictive ability toward substrates and nonsubstrates (i.e., percent substrates predicted correctly/percent nonsubstrates predicted correctly). While no significant correlation existed between the size of the data set (i.e., number of chemicals) and the percent of the test set predicted correctly, it is likely that the percent of test set predicted correctly is related in some way to the diversity of chemicals tested.

Reaction phenotyping provides useful information for the selection of new chemical entities in the drug discovery process. Much of the interindividual variation in drug response across a population is due to variation in metabolic clearance. Each metabolic enzyme has a different degree of population variability due to differences in levels of expression and activity. For example, promoter polymorphisms affect the expression of UGT1A1 while coding region polymorphisms and drug interactions influence UGT1A1 intrinsic clearance. Our understanding of the factors responsible for population variability in drug-metabolizing enzyme activity is increasing rapidly, and judgment may be made regarding which enzyme(s) is preferable to metabolize a new chemical entity. As indicated previously, in vitro assays used for reaction phenotyping are relatively slow and expensive. Hence, estimates of enzyme activity and reaction phenotyping are generally not available to guide chemical selection until a relatively late stage in the drug discovery process. It is only recently that classification techniques have been applied to isoforms of another important drug metabolizing enzyme, cytochrome P450 (CYP). Two studies have investigated the predictability of chemical inhibition of CYP3A4.[26,27] PLSDA and an ANN were separately used to classify the inhibitors, and both studies reported similar predictive ability (approximately 90%). PLSDA has also been used to generate a model capable of classifying inhibitors of CYP2C9 with approximately 75% accuracy.[28] Classification techniques have not previously been applied to any other enzymes involved in drug metabolism.

This paper is the first to report the use of classification methods to discriminate between substrates and nonsubstrates of drug metabolizing enzymes. Specifically, the use of simple 2D chemical descriptors and pattern recognition methods (especially SVM) has provided predictive models of UGT isoform substrate selectivity. Like UGT, most other drug metabolizing enzymes, particularly CYP, exist as superfamilies of isoforms which exhibit distinct but overlapping substrate selectivities. It is likely that similar approaches to those adopted in this work for UGT will similarly prove useful for other drug metabolizing enzyme families. However, this requires confirmation.

4. CONCLUSIONS

A comparison of three widely used classification methods has shown that, on average, the SVM is able to generate the most predictive models, followed by BRANN-derived models and then PLSDA. It is likely that the classification performance of other noisy data sets produced from data mining or from high-throughput assays will follow similar patterns. The speed, unique solution, and generalization performance of the SVM make it an excellent choice for general classification problems.

Using only standard 2D descriptors, models capable of predicting metabolism by individual human UGT isoforms were generated. This is a significant step toward integrated in silico metabolism models for use in early stages of drug discovery. Studies are in progress to increase predictive ability through use of chemical descriptors that contain information more relevant to the reaction mechanism employed by UGT. Furthermore, increased interpretability of the models is also priority for further work, as this will likely allow greater use of the models in the drug design process.

ACKNOWLEDGMENT

This work was funded by a grant from the National Health and Medical Research Council of Australia. M.J.S. is the recipient of an Australian Postgraduate Award.

REFERENCES AND NOTES

(1) Wold, S.; Ruhe, A.; Wold, H.; Dunn, W. J. The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses. *J. Scientific Stat. Comput.* **1984**, *5*, 735−743.
(2) Bishop, C. M. *Neural networks for pattern recognition*; Oxford University Press: Oxford, 1995.
(3) Burges, C. J. C. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining Knowledge Discovery* **1998**, *2*, 121−167.
(4) Czerminski, R.; Yasri, A.; Hartsough, D. Use of Support Vector Machine in pattern classification: Application to QSAR studies. *Quant. Struct.−Act. Relat.* **2001**, *20*, 227−240.
(5) Doniger, S.; Hofmann, T.; Yeh, J. Predicting CNS permeability of drug molecules: Comparison of neural networks and support vector machine algorithms. *J. Comput. Biol.* **2002**, *9*, 849−864.
(6) Miners, J. O.; Mackenzie, P. I. Drug glucuronidation in humans. *Pharmacol. Ther.* **1991**, *51*, 347−369.
(7) Mackenzie, P. I.; Owens, I. S.; Burchell, B.; Bock, K. W.; Bairoch, A. et al. The UDP Glycosyltransferase gene superfamily − Recom-

mended nomenclature update based on evolutionary divergence. *Pharmacogenetics* **1997**, *7*, 255−269.
(8) Radominska-Pandya, A.; Czernik, P. J.; Little, J. M.; Battaglia, E.; Mackenzie, P. I. Structural and functional studies of UDP-glucuronosyltransferases. *Drug Metab. Rev.* **1999**, *31*, 817−899.
(9) Tukey, R. H.; Strassburg, C. P. Human UDP-glucuronosyltransferases: Metabolism, expression, and disease. *Annu. Rev. Pharmacol. Toxicol.* **2000**, *40*, 581−616.
(10) Sorich, M. J.; Smith, P. A.; McKinnon, R. A.; Miners, J. O. Pharmacophore and quantitative structure activity relationship modelling of UDP-glucuronosyltransferase 1A1 (UGT1A1) substrates. *Pharmacogenetics* **2002**, *12*, 635−645.
(11) Smith, P. A.; Sorich, M. J.; McKinnon, R. A.; Miners, J. O. Pharmacophore and quantitative-structure activity relationship modeling: Complementary approaches for the rationalization and prediction of UDP-Glucuronosyltransferase 1A4 substrate selectivity. *J. Med. Chem.* **2003**, *46*, 1617−1626.
(12) Whitley, D. C.; Ford, M. G.; Livingstone, D. J. Unsupervised forward selection: a method for eliminating redundant variables. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1160−1168.
(13) Winkler, D.; Burden, F. R.; Watkins, A. J. R. Atomistic topological indices applied to benzodiazepines using various regression methods. *Quant. Struct.−Act. Relat.* **1998**, *17*, 14−19.
(14) Burden, F. R. A Chemically Intuitive Molecular Index Based on the Eigenvalues of a Modified Adjacency Matrix. *Quant. Struct.−Act. Relat.* **1997**, *16*, 309−314.
(15) Kier, L. B.; Murray, W. J.; Hall, L. H. Molecular connectivity. 4. Relationships to biological activities. *J. Med. Chem.* **1975**, *18*, 1272−1274.
(16) Bush, B. L.; Nachbar, R. B. Sample-distance partial least squares − PLS optimized for many variables, with application to COMFA. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 587−619.
(17) Muller, K.; Mika, S.; Ratsch, G.; Tsuda, K.; Scholkopf, B. An introduction to kernel-based learning algorithms. *IEEE Trans. Neural Networks* **2001**, *12*, 181−202.
(18) Scholkopf, B.; Smola, A. J.; Williamson, R. C.; Bartlett, P. L. New support vector algorithms. *Neural Comput.* **2000**, *12*, 1207−1245.
(19) Chang, C.-C.; Lin, C.-J. LIBSVM: A library for support vector machines. 2003, Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.
(20) Burden, F. R.; Winkler, D. A. A quantitative structure−activity relationships model for the acute toxicity of substituted benzenes to Tetrahymena pyriformis using Bayesian-regularized neural networks. *Chem. Res. Toxicol.* **2000**, *13*, 436−440.
(21) Nabney, I. T. *Netlab: algorithms for pattern recognition*; Springer: London, 2002.
(22) Burden, F. R.; Winkler, D. A. Robust QSAR models using Bayesian regularized neural networks. *J. Med. Chem.* **1999**, *42*, 3183−3187.
(23) Brown, R. D.; Martin, Y. C. The Information Content of 2d and 3d Structural Descriptors Relevant to Ligand−Receptor Binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1−9.
(24) Burden, F. R.; Ford, M. G.; Whitley, D. C.; Winkler, D. A. Use of automatic relevance determination in QSAR studies using Bayesian neural networks. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1423−1430.
(25) Burden, F. R. Quantitative structure − Activity relationship studies using Gaussian processes. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 830−835.
(26) Zuegge, J.; Fechner, U.; Roche, O.; Parrott, N. J.; Engkvist, O. et al. A fast virtual screening filter for cytochrome P450 3A4 inhibition liability of compound libraries. *Quant. Struct.−Act. Relat.* **2002**, *21*, 249−256.
(27) Molnar, L.; Keseru, G. M. A neural network based virtual screening of cytochrome P450 3A4 inhibitors. *Bioorg. Med. Chem. Lett.* **2002**, *12*, 419−421.
(28) Afzelius, L.; Masimirembwa, C. M.; Karlen, A.; Andersson, T. B.; Zamora, I. Discriminant and quantitative PLS analysis of competitive CYP2C9 inhibitors versus noninhibitors using alignment independent GRIND descriptors. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 443−458.

CI034108K