QSAR & Combinatorial Science QCS

Comparison of Contemporary Feature Selection Algorithms: Application to the Classification of ABC-Transporter Substrates

Michael A. Demel[a], Andreas G. K. Janecek[b], Wilfried N. Gansterer[b], Gerhard F. Ecker[a]*

[a] University of Vienna, Dept. of Medicinal Chemistry, Pharmacoinformatics Research Group, Althanstrasse 14, A-1090 Vienna, Austria
[b] University of Vienna, Research Lab Computational Technologies and Applications, Lenaugasse 2/8, A-1080 Vienna, Austria
*e-mail: Gerhard.F.Ecker@univie.ac.at

Keywords: ABC-transporter, Substrates, ADME/TOX, Feature selection, Classification

This work has been presented in part at the EuroQSAR Conference in Uppsala, 2008

Received: December 15, 2008; Accepted: July 2, 2009

DOI: 10.1002/qsar.200860191

Abstract
Multidrug ABC-transporters are highly polyspecific in their substrate recognition pattern and influence the pharmacokinetics of a broad variety of structurally diverse compounds. Thus, prediction of ABC-transporter substrate properties of compound libraries is of major interest. In this study, we use *k*-nearest neighbor (*k*NN) classification in combination with five different feature subset selection (FS) algorithms to create predictive models for classification of ABCB1, ABCC1, and ABCG2 substrates. Our results show that FS methods that incorporate the classification algorithm give the best results and contain only a small subset of descriptors. For ABCB1 and ABCG2 cross validated accuracies of higher than 80% were achieved. The interpretation of the best performing feature subsets showed that descriptors consisting of simple counts as well as of projections of physicochemical properties on subdivided surface areas have highest contribution to the models.

1 Introduction

ATP-Binding-Cassette (ABC) transporters represent a ubiquitous family of membrane-bound proteins being mainly responsible for conducting chemo-defence mechanisms by extruding xenobiotics out of living cells [1]. Thus, the ABC-transporters ABCB1 (P-glycoprotein), ABCG2 (MXR, BCRP), and ABCC1 (MRP1) confer a multidrug-resistant phenotype to cancer cells [2]. Furthermore, they are expressed in various tissues and thus influence absorption and distribution of a broad variety of structurally and functionally unrelated compounds [3]. In light of this increasing knowledge on the importance of ABC-transporters for bioavailability of candidate compounds prediction of potential substrates is of major interest in the early drug discovery phase [4].

Selecting the most relevant descriptors (features) reflecting the relationship between chemical structure and biological activity is one of the striking challenges in ligand-based design [5]. Roughly, feature subset selection (FS) algorithms can be categorized into two distinct classes: filters and wrappers. Filter methods are fast and classifier-agnostic, i.e. they do not rely on the performance of a specific classifier. Some of the filter methods consider interaction effects among variables and therefore return a selected feature subset, whereas others perform only feature ranking according to the individual predictive power of the respective descriptor. For ranking methods, an additional heuristic (e.g. selection of the *n*-top ranked features) has to be performed to yield a final subset. Wrappers are feedback methods, which rely on a specific classifier to evaluate the quality of a set of features. Thus, wrappers can also be seen as a feature subset selection method.

Previously we were able to show that a different kind of data pre-processing technique, namely principal component analysis (PCA) can be successfully used as dimensionality reduction technique to classification of ABCB1 substrates by computing linear combinations of the original attributes and using them for classification. However, our results showed that classification performance with PCs is highly unstable and depends heavily on the methodology applied to calculate eigenvalues and eigenvectors [6]. Moreover, PC loadings do not always reliably indicate

Supporting information for this article is available on the WWW under www.qcs.wiley-vch.de

which variables are the most relevant ones [7, 8]. In this study we examine the performance of different FS methods that perform either feature subset selection or feature ranking, rather than reducing the input space by linear combinations of these.

2 Methods

2.1 Data Sets

Data sets for ABCB1, ABCC1, and ABCG2 substrates have been obtained from the work of Szarkacs et al. [9]. Potential advantages of using these data for constructing in-silico models over certain other data sets (mostly compiled from different literature sources) are that there are no inter-laboratory differences in the measurements and it provides chemical information and activity of more than 1400 compounds for all 48 human ABC-transporters. These screening data contain Pearson's correlation coefficients of the mRNA levels of the respective transporter and the cytotoxicity of a compound over a panel of 60 tumor cell lines In other words, compounds which give a negative correlation over 60 different cancer cell lines between transporter expression (determined as mRNA level) and their "intrinsic" cytotoxicity (the higher the expression of the transporter the lower is the toxicity of the compound)can be regarded as being transported by the transporter, whereas compounds showing no correlation between these two parameters are not regarded to interact with the protein. In this study we assign compounds with correlation coefficients lower than -0.3 to be substrates and those which show no correlation between toxicity and transporter expression ($-0.02 < r < 0.02$) to be nonsubstrates. This procedure yields a set of 240 (110 (45.8%) substrates and 130 (54.2%) nonsubstrates) compounds for ABCB1, 227 (124 (54.6%) substrates and 103 (45.4%) nonsubstrates) compounds for ABCC1 and 198 (94 (47.5%) substrates and 104 (52.5%) nonsubstrates) compounds for ABCG2.

2.2 Structure Preparation and Molecular Descriptors.

Chemical structures were cleaned from counter ions, and hydrogens and lonepairs are added using the MOE2007.09 [10] wash routine. PEOE partial charges are assigned to each structure. Minimization was carried out using the MMFF99x forcefield. The descriptor classes used in this study cover the collection of all available 2D descriptors contained in the MOE software environment. We have discarded those descriptors that reflect only filter types (e.g.: number of leadlike-violations). In total we utilize 179 descriptors.

2.3 FS Algorithms

The five different FS techniques considered here embody one unsupervised subset selection method (RM), two filter ranking procedures (IG, ReliefF), one filter subset selection technique (CFS), and one wrapper for the *k*NN classifier (*k*NN-wrapper).

RM: The RM criterion, as proposed by McCabe [11, 12], is a proximity (or similarity) measure that uses the concept of matrix correlation of the PC matrix and the *p*-optimal feature subset of the original matrix. Matrix correlation is defined as the cosine between two $n \times p$ matrices [12]. This cosine origins from dividing the inner product of two matrices (which is calculated similarly to the "usual" inner product of two vectors) by the norm induced by this inner product. According to this, the RM coefficient represents the cosine of the PC matrix of the original matrix and a selected feature subset of the original data matrix. This feature subset is selected by a genetic algorithm as search heuristic. The genetic algorithm utilizes the RM criterion as fitness function to identify the global optimal subset [13].

Information Gain (IG): IG, also known as Kullback-Leibler divergence, originally used to compute splitting criteria in decision tree algorithms, is often applied to find out how well each single feature separates a data set [14]. It can be seen as a supervised analogue of Claude Shannon's information theoretical entropy calculation. The relevance of each attribute is measured in terms of entropy reduction. The underlying theory of this algorithm is to eliminate those descriptors whose value distributions are relatively random across the class labels, i.e., have only a small entropy. A drawback of this filter-ranking method is that each descriptor is evaluated independently of the context of other descriptors.

ReliefF: The ReliefF algorithm, as introduced by Kira and Rendell [15, 16], is a ranking method which, utilizes instance based learning to determine a relevance weight for each descriptor. Each of these weights reflect the descriptor's ability to discriminate between the classes. The output of this method is a ranked quality weight for each feature in the range $[-1,1]$.

Correlation-based Feature Selection (CFS): Contrary to IG and ReliefF, CFS as introduced by Hall [17] is a filter method that performs feature subset selection. The resulting feature subset contains features that show a high degree of correlation with the class label (i.e., they are supposed to be predictive of the class label), while having a low degree of intercorrelation (i.e., are not supposed to be predictive of each other).

kNN-Wrapper: Contrary to the methods above, wrappers incorporate the machine learning algorithm in the FS process, rather than being independent of it [18]. Wrappers, generally, tend to achieve better classification results than filter methods based on the fact that they are tuned towards the classification algorithm and its training data.

However, they tend to be much slower than filters because they must repeatedly call the induction algorithm. In our case, the resulting feature subset from a wrapper relies on the performance of the *k*NN algorithm evaluated by 10-fold cross validation.

2.4 Classification Algorithm

For comparing the effectiveness of our FS methods, we apply *k*NN classification modeling. A 10-fold cross validation is carried out to determine the performance of the different models. Each data set containing the features selected by the respective FS method is randomly split into 90% training compounds and 10% test compounds. A kNN model is constructed on the selected 90% training compounds and the 10% test compounds are predicted. This is repeated 10 times. It is noteworthy to mention, that sampling is done without replacement in order to assure that each compound is one time in the test set and the remaining nine times in the training sets. Additionally, we used *Y*-randomization to estimate the relationship between the derived feature subsets and the binary biological activity. For this, we randomly split our data sets into 90% training and 10% test set and permuted the class label of the training set. Consecutively, a *k*NN model was built on this data set of the selected features for each method and this permuted activity. This procedure was necessary since *k*NN modelling has no intrinsic training step (*k*NN is a 'lazy' learning method or 'instance-based' learning method, which does classification in the same step as learning). Afterwards, this model was used to predict the class label of the remaining 10% test compounds. This was repeated for 100 times and classification results were averaged for each method and each data set. The overall classification accuracy (given in percent) was calculated as: $((true\ positives + true\ negatives)/all\ compounds) \times 100$.

2.5 Software

The R software package [19] was used to generate the RM subsets (function genetic in the subselect package) as well as for generating the 2D radial visualization plots (function radviz2d in the dprep package). The *Y*-randomized training sets for each data set were generated using an in-house R script. All other FS methods as well as the classification were performed using the WEKA3.5.8 software [20].

3 Results and Discussion

3.1 Comparison of FS Methods

We compared five different FS algorithms with respect to their classification accuracy in 10-fold cross validation runs as well as in 100 replicates of *Y*-randomization. Furthermore, we compared the derived models also in terms of the numbers of descriptors. Since classification was done on basis of *k*NN we first elucidated the optimal value of *k*. For all data sets it is shown that $k = 1$ retrieved the best results (see Supplementary Information SI 1). Therefore, the results in Table 1 report classification with $k = 1$.

From Table 1 it can be seen that the best models in terms of %-classification accuracy are derived by the wrapper FS methodology with correct classifications of 85.6% for ABCB1, 72.0% for ABCC1, and 88.1% for ABCG2 in 10-fold crossvalidaton. Comparing these *k*NN classification results for ABCB1 with previously published results with *k*NN learning on higher dimensional data sets [21] demonstrates the applicability of our models. Our models also show a relatively low classification accuracy in *Y*-randomization, which further highlights the information content of the selected features. Additionally, the models contain only a small number of descriptors (10 – 12) for all three targets. The unsupervised method (RM) showed the poorest classification performance for the three data sets. The ranking methods IG and ReliefF gave similar classification results, but the best models retrieved with the ReliefF method contained a smaller number of descriptors (see Supplementary Information SI2). However, ReliefF also showed a similar performance for *Y*-randomization as for cross-validation (especially for ABCC1), which renders this method questionable for this application. The CFS algorithm showed a medium performance among the FS methods used.

Table 1. Classification performance expressed as classification accuracy for the three data sets and the five FS methods. Best classification results are shown in bold letters, worst results are underlined; ACC = classification accuracy, 10 × CV = 10-fold cross-validation, Yrand = *Y*-randomization, # descr = number of descriptors.

	ABCB1			ABCC1			ABCG2		
	ACC (10 × CV)	Yrand (test)	# descr.	ACC (10 × CV)	Yrand (test)	# descr.	ACC (10 × CV)	Yrand (test)	# descr.
RM	<u>63.54</u>	54.55	7	<u>57.66</u>	34.78	9	<u>44.76</u>	55.22	2
IG	75.42	59.09	90	62.11	47.83	54	76.91	50.69	72
RELIEF	80.83	72.73	54	65.11	65.22	18	76.56	55.36	18
CFS	74.6	36.36	15	59.44	52.17	3	79.07	70.69	11
WRAPPER	**85.57**	45.46	12	**71.96**	39.13	10	**88.14**	64.55	11

Since the best models were obtained when applying the *k*NN-wrapper algorithm, we also examined the classification performance of the three feature sets with respect to other classification algorithms such as support vector machine and decision tree. For the ABCB1 and the ABCG2 data set classification accuracies of 75% to 79% are obtained. For the ABCC1 set classification results are worse, but still comparable to the *k*NN results. This suggests that the obtained feature sets might also be used successfully in the context of other machine learning systems. For details see Supplementary Information SI3.

3.2 Interpretation of Wrapper Selected Feature Subsets

For the three ABC-transporters the following descriptors have been selected by the wrapper method:

ABCB1: apol, chi0_C, chi0v_C, chi1_C, rings, PEOE_VSA-5, PEOE_VSA_POL, PEOE_VSA PPOS, SlogP_VSA0, SMR_VSA2, TPSA, opr_brigid.

ABCC1: a_count, a_hyd, chi1v, opr_nring, PEOE_VSA+3, PEOE_VSA+5, PEOE_VSA-4, PEOE_VSA-6, Q_VSA_PNEG, vsa_acc.

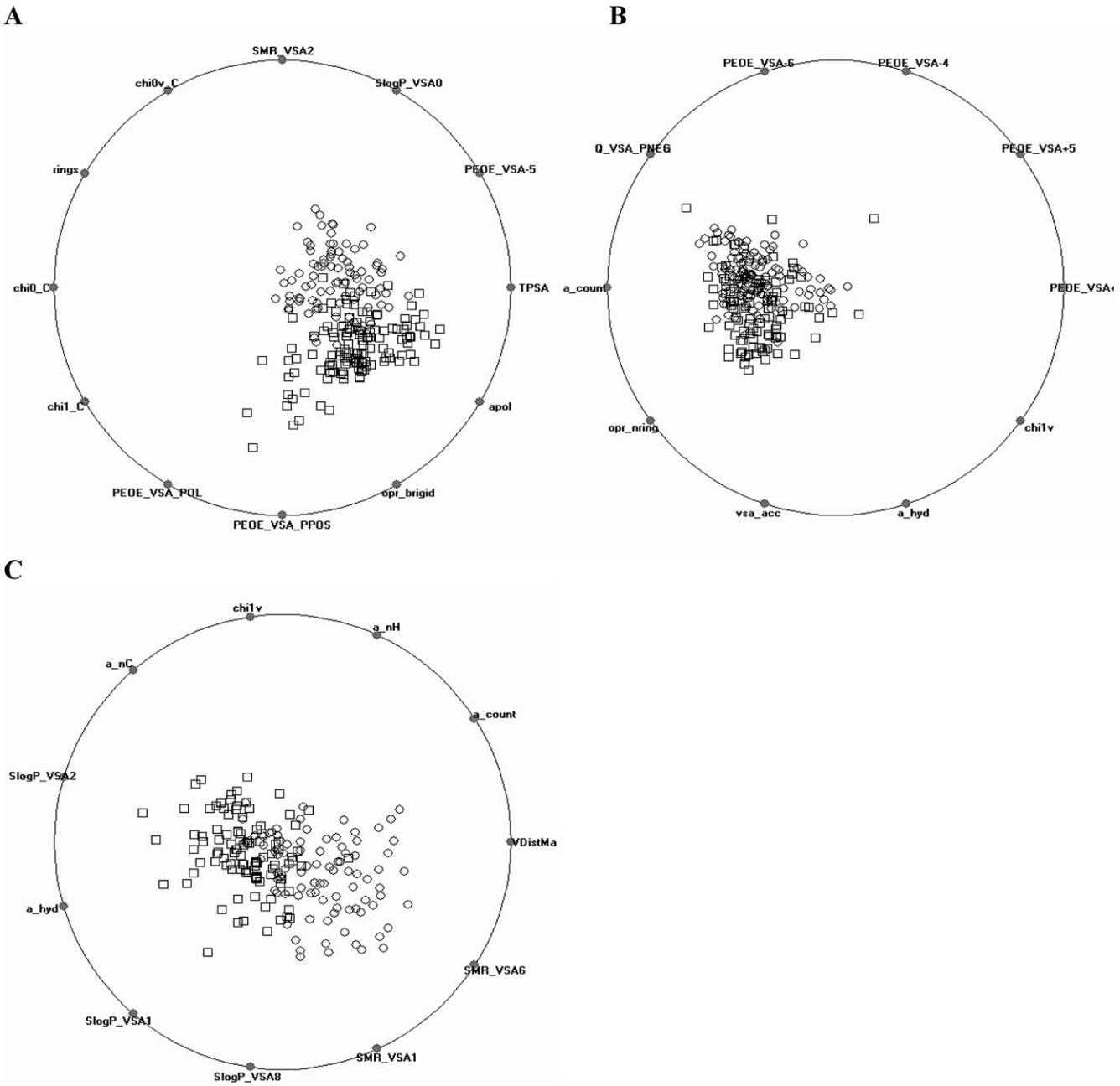

Figure 1. A – C: Radial visualization of the kNN-WRAPPER selected descriptors. Classes are encoded as follows: circles = substrates / actives, squares = nonsubstrates.

ABCG2: a_count, a_hyd, a_nC, a_nH, chi1v, SlogP_VSA1, SlogP_VSA2, SlogP_VSA8, SMR_VSA1, SMR_VSA6, VDistMa.

From this list of selected descriptors it can be deduced that simple atom counts as well as a description in terms of subdivided surface areas (VSA) are in general good means to describe substrates and nonsubstrates of ABC-transporters. Interestingly, for ABCB1 and ABCG2 VSA-descriptors reflecting lipophilicity and size (SlogP_VSA and SMR_VSA) are selected, whereas for ABCC1 the dominant class of VSA-descriptors is the partial charge class (PEOE_VSA). The three descriptor sets are graphically visualized in a 2D-radial visualization in Figure 1A – C, which represents a nonlinear projection of the attribute space with the attributes shown along the perimeter of the circle onto two dimensions. Classes are specified by circles (active/substrates) and squares (inactive/nonsubstrates). For details on this visualization technique see [22].

The graphs show that for ABCB1 and ABCG2 a good separation of the two classes is achieved using the wrapper selected attributes. Maximum class separation is considered to be one of the striking attributes of a feature subset. The results in Figure 1A – C illustrate the class discriminating power of the selected subsets. The interpretation of Figure 1A highlights that substrates of ABCB1 have higher values for SMR_VSA2 and SlogP_VSA8 and the nonsubstrates show a higher number of rigid bonds (Oprea's rigid bond count – opr_brigid), polarizable atoms (a_pol) and higher values for partial charge descriptors (PEOE_VSAPPOS, PEOE_VSA_POL). The importance of lipophilicity for ABCB1 substrates has already been mentioned in other studies [23, 24]. From these results, a reduction of lipophilicity as well as an increase in the number of rigid bonds might be a promising strategy to avoid ABCB1 conferred drug-transport. For ABCG2 (Fig. 1C) SMR_VSA1 and SMR_VSA6 are the dominant descriptors for the active class, whereas the inactive class shows higher values for SlogP_VSA1, SlogP_VSA2, and a_hyd. For the ABCC1 set (Fig. 1B) the separation is only moderate, which is in convergence with its weak classification performance. However, the nonsubstrates seem to be more rigid, which is reflected in Oprea's ring count descriptor.

4 Conclusions

In this paper we concentrated on the classification performance of five different feature subsets for the three ABC-transporters ABCB1, ABCC1, and ABCG2. Our results show that the wrapper method outperforms the other FS methods. Additionally, a comparison of the three feature sets retrieved for ABCB1, ABCC1 and ABCG2 highlights certain general properties (e.g. size, partial charge, rigidity) of ABC-transporter substrates and nonsubstrates that might be useful in shaping chemical libraries to avoid ABC-transporter related ADMET problems.

5 Acknowledgement

This work was supported by the FFG (grant #B1-812074) as well as by the CPAMMS project (FS397001) in the research focus area "Computational Science" of the University of Vienna.

6 References

[1] K. Linton, *Physiology* **2007**, *22*, 122 – 130.
[2] M. M. Gottesman, T. Fojo, S. E. Bates, *Nat. Rev. Cancer* **2002**, *2*, 48 – 58.
[3] G. Szakács, A. Váradi, C. Özvegy-Laczka, B. Sarkadi, *Drug Discov. Today* **2008**, *13*, 379 – 393.
[4] G. Ecker, *Chemistry Today* **2005**, *23*, 39 – 42.
[5] M. A. Demel, A. G. K. Janecek, K.-M. Thai, G. Ecker, W. N. Gansterer, *Curr. Comput.-Aided Drug Design* **2008**, *4*, 91 – 110.
[6] A. G. K. Janecek, W. N. Gansterer, M. A. Demel, G. F. Ecker, *JMLR* **2008**, *4*, 90 – 105.
[7] J. F. C. L. Cadima, I. T. Jolliffe, *J Appl. Statist.* **1995**, *22*, 203 – 214.
[8] J. F. C. L. Cadima, I. T. Jolliffe, *J Agric. Biol. Environ. Statist.* **2001**, *6*, 62 – 79.
[9] G. Szakacs, J. P. Annereau, S. Lababidi, U. Shankavaram, A. Arciello, K. J. Bussey, W. Reinhold, Y. Guo, G. D. Kruh, M. Reimers, J. N. Weinstein, M. M. Gottesman, *Cancer Cell* **2004**, *6*, 129 – 137.
[10] M*olecular Operating Environment (MOE)*, Chemical Computing Group, Version 2007.09.
[11] G. P. McCabe, *Technical Report #86 – 19*, Dept. of Statistics, Purdue University **1986**.
[12] G. P. McCabe, *Technometrics* **1984**, *26*, 137 – 144.
[13] J. F. C. L. Cadima, I. T. Jolliffe, *Comput. Stat. Data Anal.* **2004**, *47*, 225 – 236.
[14] I. Witten, E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, San Francisco, CA **2005**.
[15] K. Kira, L. Rendell, in *Ninth International Workshop on Machine Learning* **1992**, 249 – 256.
[16] I. Kononenko, in *European Conference on Machine Learning*, **1994**, 171 – 182.
[17] M. Hall, *PhD thesis*, Waikato, New Zealand, **1998**.
[18] R. Kohavi, G. H. John, *Artificial Intelligence* **1997**, *97*, 273 – 324.
[19] *http://cran.r-project.org/*
[20] *http://www.cs.waikato.ac.nz/~ml/weka/*
[21] P. De Cerqueira-Lima, A. Golbraikh, S. Oloff, Y. Xiao, A. Tropsha, *J. Chem. Inf. Model.* **2006**, *46*, 1245 – 1254.
[22] M. Ankerst, D. Keim, in *Proc. IEEE Visualization Conf.* **1997**.
[23] R. Didziapetris, P. Japertas, A. Avdeef, A. Petrauskas, *J. Drug Targeting* **2003**, 391 – 406.
[24] J. Huang, G. Ma, I. Muhammad, Y. Cheng, *J. Chem. Inf. Model.* **2007**, *47*, 1638 – 1647.

