Analysis of Activity Space by Fragment Fingerprints, 2D Descriptors, and Multitarget Dependent Transformation of 2D Descriptors

Alireza Givehchi,*[,†,‡] Andreas Bender,[§,||] and Robert C. Glen[§]

Institut für Organische Chemie und Chemische Biologie, Johann Wolfgang Goethe-Universität, Marie-Curie-Strasse 11, D-60439 Frankfurt, Germany, Klinik und Poliklinik für Neurochirurgie, Klinik und Poliklinik für Neurologie, Universitätsklinikum Münster, Albert-Schweitzer-Strasse 33, D-48129 Münster, Germany, and Unilever Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom

Received January 22, 2005

The effect of multitarget dependent descriptor transformation on classification performance is explored in this work. To this end decision trees as well as neural net QSAR in combination with PLS were applied to predict the activity class of 5HT3 ligands, angiotensin converting enzyme inhibitors, 3-hydroxyl-3-methyl glutaryl coenzyme A reductase inhibitors, platelet activating factor antagonists, and thromboxane A2 antagonists. Physicochemical descriptors calculated by MOE and fragment-based descriptors (MOLPRINT 2D) were employed to generate descriptor vectors. In a subsequent step the physicochemical descriptor vectors were transformed to a lower dimensional space using multitarget dependent descriptor transformation. Cross-validation of the original physicochemical descriptors in combination with decision trees and neural net QSAR as well as cross-validation of PLS multitarget transformed descriptors with neural net QSAR were performed. For comparison this was repeated using fragment-based descriptors in combination with decision trees.

Using neural net QSAR with PLS multitarget dependent transformed physicochemical descriptors improves the class-averaged number of correctly classified instances from 87.08% to a value of 96.57%, compared to untransformed descriptors. At the same time only five-dimensional input vectors are employed. Using decision trees in combination with fragment-based descriptors as a benchmark gave a class-averaged correct classification rate of the active class of 74.9% for all classes. Improved performance can also be observed in the PLS models alone if transformed descriptors are employed, suggesting broad applicability of the data treatment procedure presented here. Descriptor transformation, as proposed in this work, represents an improved way to preprocess data for classification purposes. This is particularly relevant since the number of input vector dimensions can be greatly reduced, circumventing problems associated with high-dimensional data spaces such as underdetermined equation systems and overlearning.

INTRODUCTION

Computer-aided drug design methods such as "virtual screening"[1,2] attempt to find new active materials, often small molecules, that are able to act against certain diseases. Many methods which are used successfully in research and development have been developed to this end, and they are often based on the "molecular similarity principle"[3−5] which states that structurally similar molecules tend to exhibit similar properties.

The comparisons of molecules that aim to identify those promising new hits generally take place in three steps, and is often carried out in a way that an active structure is used as a query for searching a large compound database. In the first step, molecules are represented in chemical space using one or several "molecular descriptors".[4,5] Second, features may be selected. In this step the aim is to identify those features which are more relevant to the particular classification, thereby eliminating noise and improving the signal-to-noise ratio, thus improving classification performance. Finally, activity or other molecular properties are estimated using a similarity (or distance) measure[5] between the structures with known properties and the new structures whose properties one would like to predict. The more similar the molecules are structurally the more likely they are to have similar biological properties (as the "molecular similarity principle" states). As was shown recently[6] the confidence with which predictions about new molecules can be made is increased for molecules which are more similar to the training set.

Among the commonly used mathematical transformation and model generation methods are principal component analysis (PCA), partial least squares (PLS),[7−9] nonlinear iterative partial least-squares (NIPALS),[7−11] decision trees,[12] rule-based approaches,[13] neural networks,[14−17] and evolutionary algorithms.[18−20] Improvements have also been achieved using combinations of those methods.[21,22] But as described above, the prediction of a new molecular entity based on its similarity to substances with known properties also comprises

* Corresponding author phone: +49 (69) 798 29824; fax: +49 (69) 798 29826; e-mail: alireza.givehchi@chemie.uni-frankfurt.de.
† Johann Wolfgang Goethe-Universität.
‡ Universitätsklinikum Münster.
§ University of Cambridge.
|| Current address: Lead Discovery Center, Novartis Institutes for BioMedical Research Inc., 250 Massachusetts Ave., Cambridge, MA 02139.

10.1021/ci0500233 CCC: $33.50
Published on Web 03/04/2006

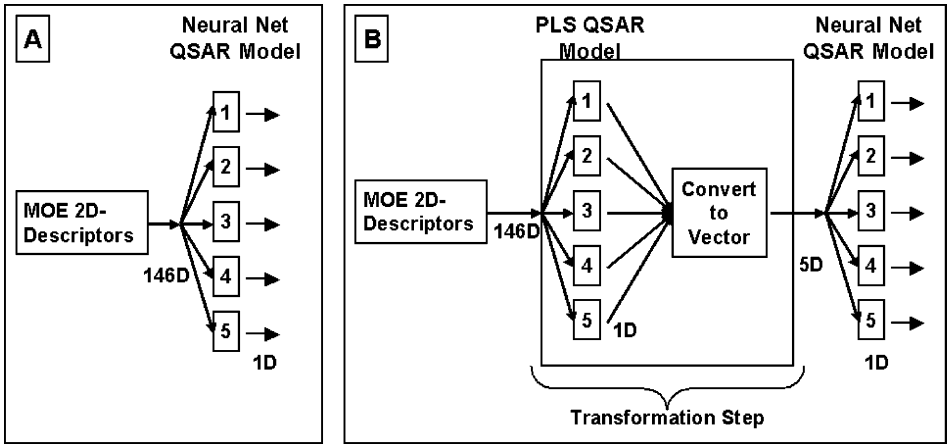

Figure 1. (A) For the untransformed descriptors, 5 neural net QSAR models for the prediction of activity classes are generated directly from MOE 2D descriptors. (B) For the transformed descriptors, the first PLS models are generated and subsequently neural net QSAR models are built. Thus the original MOE descriptors are subject to a multitarget dependent transformation to give five-dimensional vectors which are used for further classification.

the feature selection and the descriptor generation step. Thus, to optimize classification and prediction, new approaches for representing chemical information are also of crucial importance: the descriptor chosen has to capture information about the structures under consideration which is relevant to the problem,[23−25] together with the smallest possible amount of noise (which might clearly be valuable information in another context). This is the purpose of the current work. It explores the possibility to transform multidimensional physicochemical descriptors to low-dimensional multitarget-dependent descriptors, a procedure which (as shown later) is able to improve classification performance.

In addition the problem (or opportunity?) of multitarget activity is addressed in this paper. While seemingly "simple" binary active/inactive classification against one target is, depending on the activity class, already a nontrivial task, another problem exists in drug design: the fact that one molecule could have activity on different targets which cause side effects. In some cases these activities may be desired,[26] in others just suspected,[27] but in reality often not even enough knowledge about a single target involved in a disease is known, let alone multiple receptors or enzymes (depending on the particular target or pathway involved). Therefore ligand-based algorithms need to be optimized so that they are able to find ligands which exhibit the desired activity profile (selectivity) for the particular situation.

In the following we compare the classification performance of two different representations of a compound set comprising five activity classes (plus a set of "decoys"), in combination with two different model generation methods. First, a set of physicochemical descriptors is calculated for the data set, and a cross-validated model generation via a neural net QSAR model is performed on this full set of physicochemical descriptors. Second, the physicochemical descriptors are transformed to lower-dimensional chemical space with the help of PLS models, and again the neural net QSAR model with the transformed descriptors is cross-validated. To establish a performance baseline, fragment-based descriptors and the full set of physicochemical descriptors are used to build a decision tree using a C4.5 implementation.

MATERIAL AND METHODS

A data set of 957 compounds was selected from the MDL Drug Data Report (MDDR) which was introduced by Briem and Lessel[28] who subjected it to a variety of similarity searching methods. The database chosen enables comparison of performance to established methods since it is a well-defined data set that has already been used extensively;[23,24,28] in addition it comprises multiple activity classes. The set contains 49 serotonin receptor ligands (5HT3), 40 angiotensin converting enzyme inhibitors (ACE), 111 HMG CoA reductase inhibitors (HMG), 134 platelet activating factor antagonists (PAF), and 49 thromboxane A2 antagonists (TXA2). In addition, 574 compounds were selected randomly from other activity classes which were added as decoys. While only one activity class is assigned to each compound, no knowledge about definitive inactivity is contained in the MDDR database which generally contributes to the problem of judging performance of virtual screening approaches in retrospective settings. ("False positives" may thus be tested positive in an experimental setting in the future.) Structures were exported from the MDDR database in 2D SD format. Subsequent steps were performed in the program MOE.[29] Salts and solvents were removed using the "strip salts and solvent" option. Next, all available 2D descriptors of the compounds were calculated. The dimension of the resulting property vectors was 146.

To build a QSAR model for the prediction of compound classes we used the programs Rfor PLS model building[30] and WEKA[32] for neural net model building. All PLS models mentioned are purely linear models and do not employ cross terms. In the first series of calculations we calculated neural net QSAR models for each of the 5 classes and performed a 10-fold cross-validation of the models (Figure 1A). In the second series of calculations (Figure 1B) descriptor transformation was performed in the following way, again employing 10-fold cross-validation throughout. First, PLS QSAR models for all classes were calculated. Second, the 5 predicted activity values for each compound were used to create a five-dimensional input vector, which was used to build the predictive neural net QSAR model. Both steps were performed on the five sets of active compounds as well as

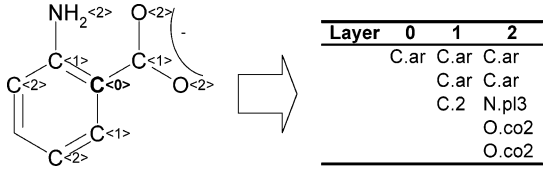

Figure 2. Illustration of MOLPRINT 2D descriptor generation step, applied to an aromatic carbon atom. The distances ("layers") from the central atom are given in brackets. In the first step, Sybyl mol2 atom types are assigned to all atoms in the molecule. In the second step, count vectors from the central atom (here C⟨0⟩) up to a given distance (two bonds from the central atom apart) are constructed. Fingerprints are then binary presence/absence indicators of count vectors of atom types without regarding connectivity information between the atom layers.

on the five sets of active compounds plus the set of "inactive" decoy structures. It should be emphasized that no direct comparison of PLS and neural net QSAR is performed here; instead the effect of multitarget dependent descriptor transformation on classification performance is the purpose of this work.

The options used to generate neural net QSAR models via weka.classifiers.functions. Multilayer perceptron in WEKA were as follows. When no descriptor transformation was employed, parameters were set to L (learning rate) = 0.3, M (momentum) = 0.2, N (number of iterations) = 2000, V (validation set size) = 95, S (seed) = 0, E (validation threshold) = 25, and H (number of neurons in hidden layer) = 10. Descriptor transformation including as well as excluding the set of inactive compounds employed the parameters $L = 0.9$, $M = 0.2$, $N = 2000$, $V = 95$, $S = 0$, $E = 25$, $H = 3$. The option -I (normalize attributes) was set to TRUE.

Thus the difference of the two methods above was the generation of the multitarget dependent five-dimensional vectors from the original 146 dimensional vectors, each representing the activity prediction for one of the five classes, as derived from a PLS model. This explains the term of a "multitarget dependent vector". One might interpret it in the way that each dimension of the new vector presents information about how similar the compound is to a "prototypical" active compound of each class, as described by the training set molecules. In a related method termed "multispace classification"[22] the descriptors were divided into subgroups (spaces), and the prediction results from each space were combined to vectors to use them for subsequent activity prediction. Here we combine the prediction result from different classes to vectors and use them for auxiliary prediction.

In addition to physicochemical descriptors, fragment-based descriptors calculated according to a previously presented method[23,24] (MOLPRINT 2D) were employed in combination with decision trees as a benchmark method. They are calculated in a two-step procedure (see Figure 2):

1. Sybyl atom types[31] are assigned to every heavy atom in the hydrogen-depleted structure of the molecule.

2. An individual atom fingerprint is calculated for every heavy atom in the molecule. A count vector is constructed with the vector elements being counts of atom types at a given distance from the central atom. This calculation is performed using distances from 0 up to 2 bonds and keeping count of the occurrences of the atom types but no information about bonding between the atoms.

Features are stored as binary presence/absence features for each molecule. Since they are calculated for every heavy atom of the structure, as many descriptors are calculated for a molecule as there are heavy atoms present in the structure.

Ten-fold cross-validation of the five classes of active compounds and the whole data set (including "inactives") was performed using a J48 decision tree[32] (the WEKA[32] implementation of the C4.5 classification tree[12] in WEKA version 3.4). A confidence factor of 0.25 and a minimum number of objects of 2 per node were used (default settings). Ten-fold cross-validation with identical parameters was also performed on the full set of physicochemical parameters, again of both the sets of active compounds only and the complete data set.

To determine the influence of multitarget dependent descriptor transformation on the performance of PLS models alone those were calculated both for the original descriptors and the five-dimensional transformed descriptors. Again the purpose is to gauge whether descriptor transformation is beneficial for model performance.

RESULTS AND DISCUSSION

The results of the classification protocol via neural net QSAR on the initial physicochemical descriptors as well as those employing neural net QSAR on the PLS-transformed descriptors are given in Table 1 for a 10-fold cross-validation. On all sets of active compounds we observe improved classification for the transformed descriptor vectors, from a class-average cross-validated root-mean-squared error 0.245 to a value of 0.153 (0.167 if the 'inactive' set of compounds is omitted). The class-averaged cross-validated fraction of correct classified compounds was improved from 87.08% to a value of 96.57% (96.32% if again the class of inactive compounds is omitted). For detailed results for the different classes please see Table 1. We assume that the reason for the prediction improvement is due to the fact that the new descriptor vectors include information differentiating between classes and not only information about the activity class to which they belong.

For comparison with this new method, well-established methods have also been employed for compound classification, namely a C4.5 decision tree in combination with fragment-based descriptors as well as physicochemical descriptors.

Classification results using the five sets of active compounds in combination with the fragment-based descriptor are given in Tables 2 and 3. The term "recall" in those tables denotes the fraction of compounds of a class which were correctly identified as belonging to this class. The confusion matrix shows which classes structures were assigned to, both correctly and incorrectly. A tree of size 33 was built in 0.18 s by the algorithm, containing 17 leaves and correctly classifying 90.1% of the instances (class-average correct prediction 88.0%). Falsely classified compounds were not distributed evenly over the other activity classes, such as in the case of 5HT3 ligands, where false-negatives were consistently classified as PAF antagonists, indicating that some of the activity classes are on average more similar to each other than to other classes.

Including the class of "inactive" compounds and again employing fragment-based descriptors, a decision tree con-

Table 1. Cross-Validation Root Mean Squared Error of the Neural Net QSAR Model without Descriptor Transformation, with PLS Descriptor Transformation, and with PLS Descriptor Transformation but without the Set of Inactive Compounds[a]

classification performance (RMSE/correctly classified data points)						
without descriptor transformation		with descriptor transformation		with descriptor transformation, without inactive compounds		
cross-validated root mean square error (RMSE)	correctly classified instances, %	cross-validated root mean square error (RMSE)	correctly classified instances, %	cross-validated root mean square error (RMSE)	correctly classified instances, %	class
0.217	94.46	0.088	98.54	0.124	98.02	5HT3
0.206	94.67	0.110	97.17	0.152	96.96	ACE
0.284	89.86	0.155	96.76	0.162	96.55	HMG
0.309	87.04	0.236	92.27	0.229	93.42	PAF
0.226	94.67	0.171	96.13	0.165	96.66	TXA2
0.226	61.75	0.159	98.56	n/a	n/a	INACT
0.245	87.08	0.153	96.57	0.167	96.32	average

[a] Results are consistently superior for the transformed descriptors.

Table 2. Confusion Matrix of 10-Fold Cross-Validated Classification of the Five Classes of Active Compounds Using a J48 Decision Tree on the Fragment-Based Descriptors

5HT3	ACE	HMG	PAF	TXA2	prediction
45	0	0	4	0	5HT3
0	30	0	5	5	ACE
0	0	105	3	3	HMG
4	0	7	122	1	PAF
1	4	0	1	43	TXA2

Table 3. Detailed Classification Results over the Five Classes of Active Compounds, Using a J48 Decision Tree on the Fragment-Based Descriptors

TP rate	FP rate	precision	recall	F-measure	class
0.918	0.015	0.900	0.918	0.909	5HT3
0.750	0.012	0.882	0.750	0.811	ACE
0.946	0.026	0.938	0.946	0.942	HMG
0.910	0.052	0.904	0.910	0.907	PAF
0.878	0.027	0.827	0.878	0.851	TXA2
0.880	0.026	0.890	0.880	0.884	average

Table 4. Confusion Matrix of 10-Fold Cross-Validated Classification of All against All Activity Classes Using a J48 Decision Tree on the Fragment-Based Descriptors[a]

5HT3	ACE	HMG	PAF	TXA2	inactive	prediction
27	0	0	6	0	16	5HT3
0	30	0	1	1	8	ACE
0	1	100	3	2	5	HMG
2	0	4	101	0	27	PAF
0	0	1	2	31	15	TXA2
9	9	8	17	11	520	INACT

[a] False-negative predictions of compounds of the five "active" classes are mostly predicted as belonging to the "inactive" data set, a finding which can be explained by the fact the "inactive" data set contains compounds from a larger number of activity classes some of which might be similar to the sets of "active" compounds.

sisting of 43 leaves and 85 nodes was built by the J48 algorithm using fragment-based descriptors. This took around 1 s of computing time; prediction results are shown in Tables 4 and 5. Overall correct classification was 84.54% with a class-average correct prediction rate of 74.90%. False classifications most often occur in the case of 5HT3 ligands (true positive rate of 55.1%) and TXA2 antagonists (63.3%), and in both cases the majority of instances (72.7% and 83.3%, respectively) are wrongly classified as belonging to the "inactive" data set. This is understandable since the "inactive" data set contains a large number of drugs from many different activity classes some of which may resemble more atypical

Table 5. Detailed Classification Results of All against All Classes of Active Compounds Using a J48 Decision Tree on the Fragment-Based Descriptors[a]

TP rate	FP rate	precision	recall	F-measure	class
0.551	0.012	0.711	0.551	0.621	5HT3
0.750	0.011	0.750	0.750	0.750	ACE
0.901	0.015	0.885	0.901	0.893	HMG
0.754	0.035	0.777	0.754	0.765	PAF
0.633	0.015	0.689	0.633	0.660	TXA2
0.906	0.185	0.880	0.906	0.893	INACT
0.749	0.046	0.782	0.749	0.764	average

[a] Predictions are worst for the 5HT3 data set and best for the set of inactive compounds.

Table 6. Confusion Matrix of 10-Fold Cross-Validated Classification of the Five Classes of Active Compounds Using a J48 Decision Tree and the Full Physicochemical Descriptor Set[a]

5HT3	ACE	HMG	PAF	TXA2	prediction
41	0	3	4	1	5HT3
1	30	2	1	6	ACE
3	1	99	6	2	HMG
6	1	4	122	1	PAF
2	3	4	2	38	TXA2

[a] Overall, results are slightly inferior to fragment-based classification.

Table 7. Detailed Classification Results of 10-Fold Cross-Validated Classification of the Five Classes of Active Compounds Using a J48 Decision Tree and the Full Physicochemical Descriptor Set[a]

TP rate	FP rate	precision	recall	F-measure	class
0.837	0.036	0.774	0.837	0.804	5HT3
0.750	0.015	0.857	0.750	0.800	ACE
0.892	0.048	0.884	0.892	0.888	HMG
0.910	0.052	0.904	0.910	0.907	PAF
0.776	0.030	0.792	0.776	0.784	TXA2
0.833	0.036	0.842	0.833	0.837	average

[a] Except for the case of PAF antagonists inferior results are obtained, compared to classification results employing fragment-based descriptors.

compounds of the 5HT3 and TXA2 data sets. Of all false-negative predictions of the five activity classes, prediction that the compounds belong to the "inactive" data set account for more than half (56.8%) of all false-negative predictions.

Classification results using the full set of physicochemical descriptors on the five classes of active compounds are given in Tables 6 and 7. A tree of size 51 was built in 1.34 s by the algorithm, containing 26 leaves and correctly classifying 86.2% of the instances (class-average correct prediction

Table 8. Confusion Matrix of 10-Fold Cross-Validated Classification of All against All Activity Classes Using a J48 Decision Tree and the Full Physicochemical Descriptor Set

5HT3	ACE	HMG	PAF	TXA2	INACT	← prediction
27	0	0	2	0	20	5HT3
0	20	1	0	3	16	ACE
0	2	75	1	1	32	HMG
0	0	3	88	0	43	PAF
0	5	2	0	29	13	TXA2
11	10	36	48	10	459	INACT

Table 9. Detailed Classification Results of 10-Fold Cross-Validated Classification of All against All Activity Classes Using a J48 Decision Tree and the Full Physicochemical Descriptor Set

TP rate	FP rate	precision	recall	F-measure	class
0.551	0.012	0.711	0.551	0.621	5HT3
0.500	0.019	0.541	0.500	0.519	ACE
0.676	0.050	0.641	0.676	0.658	HMG
0.657	0.062	0.633	0.657	0.645	PAF
0.592	0.015	0.674	0.592	0.630	TXA2
0.800	0.324	0.787	0.800	0.793	INACT
0.629	0.080	0.665	0.629	0.644	average

Table 10. Cross-Validated Root Mean Square Error of the PLS Model Generated without Descriptor Transformation, with Descriptor Transformation, and with Descriptor Transformation but without the Class of Inactive Compounds Using *R*

cross-validated root mean square error (RMSE)			
without descriptor transformation	with descriptor transformation	with descriptor transformation, without class of inactive compounds	class
0.646	0.147	0.146	5HT3
0.289	0.143	0.142	ACE
0.237	0.180	0.179	HMG
1.125	0.218	0.219	PAF
0.337	0.166	0.166	TXA2
0.270	0.213		INACT
0.484	0.178	0.170	average

83.3%). Surprisingly, performance on some classes is just as good as in the case of the fragment-descriptor based compound classification (for ACE and PAF; 57.0% and 91.0%, respectively), while for other groups performance deteriorates (5HT3, HMG, TXA2). Overall, performance of the physicochemical classification tree is slightly inferior to that of the fragment-descriptor based one (86.2% vs 90.1% correct classifications), while at the same time needing a much larger number of leaves (26 vs 17), likely leading to less robust models.

Using the full set of physicochemical parameters on all data sets (including the set of "inactive" compounds) the results in Tables 8 and 9 were obtained, giving overall correct classification rate of 72.9% and a class-average correct prediction rate of 62.9%. The model was built within 5.19 s and is comparably large, being of size 195 and containing 98 leaves. Classification performance is by 11.6% lower than that of the fragment-descriptor based model (72.9% vs 84.5% instances which are correctly classified).

To determine the influence of multitarget dependent descriptor transformation on the performance of PLS models alone those were calculated both for the original descriptors and the five-dimensional transformed descriptors. Resulting root-mean-square errors (RMSE) are given in Table 10. Cross-validated root-mean-squared errors improve considerably if descriptor transformation is performed, namely from 0.484 to 0.178 if the set of inactive descriptor is used and to 0.170 if it is not used for building the models. Overall it can be seen that performance of the PLS model can be improved considerably by employing descriptor transformation on the original descriptors.

Comparing neural net QSAR results on transformed and untransformed descriptors to decision tree results on fragment-based and physicochemical descriptors, it can be seen that while classification performance is broadly similar for untransformed descriptors and neural net QSAR, compared to both descriptor types in combination with decision trees, performance on the transformed descriptors improves greatly. This underlines the value of descriptor transformation for classification purposes.

CONCLUSIONS

In this work we have shown that the PLS-based multitarget dependent transformation of 2D-descriptors combined with neural net QSAR improves classification results in a cross-validation study over that of neural net QSAR alone. The class-average cross-validated number of correctly classified instances improves from 87.08% to 96.57%. Improved performance can also be observed if PLS models are built on the transformed descriptors. In this case the class averaged cross-validated root-mean-squared error is improved from 0.484 to 0.178 if the set of inactive structures is used and to 0.170 if not.

Classification employing the transformed descriptors in combination with neural net QSAR outperforms decision tree classification based on both fragment-based descriptors and physicochemical ones. This is not the case for the original (untransformed) descriptor set. Binary QSAR results show similar trends to the ones reported here, thereby showing that results are not due to different local optima identified by the (stochastic) neural network learning algorithms.

It can thus be concluded that descriptor transformation, as proposed here, represents an improved way to preprocess data for classification purposes. This is particularly important since the number of input vector dimensions can be greatly reduced, circumventing problems associated with high-dimensional data spaces such as underdetermined equation systems and overlearning.

ACKNOWLEDGMENT

A.G. thanks the Beilstein-Institut zur Förderung der Chemischen Wissenschaften, Frankfurt. A.B. and R.C.G. thank the Gates Cambridge Trust and Unilever for funding. The anonymous referees are thanked for very helpful comments on earlier versions of the manuscript.

APPENDIX A: PLS AND NIPALS ALGORITHM

Partial Least Squares (PLS) transforms input data into an uncorrelated space of latent variables, similar to PCA, but also taking the variability in the response variable into account. It attempts to maximize the covariance between predictor and response variables. The flavor of PLS employed here is based on the nonlinear iterative partial least squares (NIPALS) algorithm. NIPALS determines the principal components (PCs) of a data matrix **X**. Instead of calculating all PCs at once it proceeds in an iterative manner. Initially the first score and loading vector, t_1 and p_1, are calculated and their product, t_1p_1, is subtracted from **X** to determine

the residual $\mathbf{E}_1$. The subsequent scores and loadings vectors, t_i and p_i, are then calculated from the residual matrix $\mathbf{E}_i$.

The complete NIPALS algorithm can be outlined as follows:

(a) Normalize matrix **X** to a mean of zero

(b) Select the first column from **X**, t_h

(c) Normalize loading vector

$$\mathbf{p}_h^T \text{ to } p_h^T = \frac{t_h^T \mathbf{X}}{t_h^T t_h}$$

(d) Normalize loading vector

$$\mathbf{p}_h^T \text{ to give } \mathbf{p}_{h\text{norm}}^T \text{ with } \mathbf{p}_{h\text{norm}}^T = \frac{\mathbf{p}_h^T}{||\mathbf{p}_h^T||}$$

(e) Calculate score vector

$$t_h \text{ as } t_h = \frac{\mathbf{X}\mathbf{p}_h}{\mathbf{p}_{h\text{norm}}^T \mathbf{p}_{h\text{norm}}}$$

(f) Compare t_h from (e) to t_h from (c). Continue if they are within a defined threshold, otherwise go back to (c)

(g) Calculate the residual matrix **E** as $\boldsymbol{E} = \mathbf{X} - t_h p_h^T$. If the residual matrix does not reach a defined threshold level, set $\mathbf{X} = \mathbf{E}$ and go to (a). Otherwise exit.

After this procedure has been performed we are able to build the score matrix **T** and loading matrix **P**, containing score and loadings vectors, respectively. Now we can perform the algorithm above also on output matrix **Y** and calculate its score matrix, **U**.

The "outer relation" is now given by the two equations

$$\mathbf{X} = \mathbf{T}\mathbf{A}^\mathbf{T} + \mathbf{E}$$

where **A** is the loading matrix of **X**, **T** is the score matrix of **X**, and **E** is the residual matrix of **X** as well as

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^\mathbf{T} + \mathbf{F}$$

where **Q** is the loading matrix of **Y**, **U** is the score matrix of **Y**, and **F** is the residual matrix of **Y**.

The "inner relation" of the scores of **X** and **Y** can then be calculated as

$$u_i = b_i t_i;\ i = 1,2,..,k$$

Finally, the regression vector, **b**, is given via ordinary least squares (OLS) as

$$\mathbf{b}_i = (t_i^T t_i) - {}^1 t_i^T u_i$$

REFERENCES AND NOTES

(1) *Virtual screening for bioactive molecules*; Böhm, H. J., Schneider, G., Eds.; Wiley-VCH: Weinheim, 2000.

(2) Kubinyi, H. In *Computer-assisted lead finding and optimization − Current tools for medicinal chemistry*; van de Waterbeemt, H., Testa, B., Folkers, G., Eds.; Wiley-VCH: Weinheim, 1997; pp 9−28.

(3) Johnson, M. A.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*; John Wiley & Sons: New York, 1990.

(4) Bender, A.; Glen, R. C. Molecular Similarity: A Key Technique in Molecular Informatics. *Org. Biomol. Chem.* **2004**, *2*, 3204−3218.

(5) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput.* Sci. **1998**, *38*, 983−996.

(6) Sheridan, R. P.; Feuston, B. P.; Mairov, V. N.; Kearsley, S. K. Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1912−1928.

(7) Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Wold, S. *Introduction to multi- and megavariate data analysis using projection methods (PCA & PLS)*; Umeå: Umetrics, 1999.

(8) Noonan, R. D.; Wold, H. Partial least squares. In *Educational Research, Methodology, and Measurement: An International Handbook*; Keeves, J. P., Ed.; Pergamon Press: Oxford, 1988; pp 710−716.

(9) Geladi, P.; Kowalski, B. R. Partial Least-Squares Regression: A Tutorial. *Anal. Chim. Act.* **1986**, *185*, 1−17.

(10) Helland, I. S., On the structure of partial least squares. *Commun. Statistic − Simul.* **1988**, *17*, 1581−607.

(11) Givehchi, A.; Dietrich, A.; Wrede, P.; Schneider, G., ChemSpace-Shuttle: A tool for data mining in drug discovery by classification, projection, and 3D visualization. *QSAR Comb. Sci.* **2003**, *5*, 549−559.

(12) Quinlan, J. R. Induction of Decision Trees. *Mach. Learning* **1986**, *1*, 81−106.

(13) Glen, R. C.; A-Razzak, M. Applications of Rule-Induction in the Derivation of quantitative structure−activity relationships. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 349−383.

(14) Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design. An Introduction*; Wiley-VCH: Weinheim, 1999.

(15) Sadowski, J. Database Profiling by Neural Networks. In *Virtual screening for bioactive molecules*; Böhm, H.-J., Schneider, G., Eds.; Wiley-VCH: Weinheim, 2000; pp 117−130.

(16) Schneider, G.; Wrede, P. Artificial neural networks for computer-based molecular design. *Prog. Biophys. Mol. Biol*. **1998**, *70*, 175−222.

(17) Schneider, G.; Nettekoven, M. Ligand-based combinatorial design of selective purinergic receptor (A2A) antagonists using self-organizing maps. *J. Comb. Chem.* **2003**, *5*, 233−237.

(18) Goldberg, D. E. *Genetic Algorithms in search, Optimization and Machine Learning*; Addison-Wesley: MA, 1989.

(19) Schneider, G. Evolutionary Molecular Design in Virtual Fitnet Landscape. In *Virtual screening for bioactive molecules*; Böhm, H. J., Schneider, G., Eds.; Wiley-VCH: Weinheim, 2000; pp 161−186.

(20) Weber, L. Practical Approaches to Evolutionary Design. In *Virtual screening for bioactive molecules*; Böhm, H. J., Schneider, G., Eds.; Wiley-VCH: Weinheim, 2000; pp 187−205.

(21) Givehchi, A.; Schneider, G. Impact of descriptor vector scaling on the classification of drugs and nondrugs with artificial neural networks. *J. Mol. Model.* **2004**, *10*, 204−211.

(22) Givehchi, A.; Schneider, G. Multi-space classification for predicting GPCR-ligands. *Mol. Div.* **2005**, *9*(4), 371−83.

(23) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Molecular similarity searching using atom environments, information-based feature selection, and a naive Bayesian classifier. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 170−178.

(24) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Similarity searching of chemical databases using atom environment descriptors: evaluation of performance. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1708−1718.

(25) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; WILEY-VCH: Weinheim, 2000.

(26) Morphy, R.; Kay, C.; Rankovic, Z. From magic bullets to designed multiple ligands. *Drug Discovery Today* **2004**, *9*, 641−651.

(27) Sheridan, R. P. J. Finding multiactivity substructures by mining databases of drug-like compounds. *Chem. Inf. Comput. Sci.* **2003**, *43*, 1037−1050.

(28) Briem, H.; Lessel, U. F. In vitro and in silico affinity fingerprints: Finding similarities beyond structural classes. *Perspect. Drug Discovery Des.* **2000**, *20*, 231−244.

(29) Chemical Computing Group Inc., 1010 Sherbrooke Street West, Suite 910, Montreal, Quebec, Canada H3A 2R7, http://www.chemcomp.com.

(30) (30) The R Project for Statistical Computing, http://www.r-project.org/.

(31) Witten, I. H.; Frank, E. *Data Mining: Practical machine learning tools with Java implementations*; Morgan Kaufman: San Franciso, CA, 2000.

(32) Clark, M.; Cramer, R. D.; Vanopdenbosch, N. Validation of the General-Purpose Tripos 5.2 Force-Field. *J. Comput. Chem.* **1989**, *10*, 982−1012.

CI0500233