ANALYTICA CHIMICA ACTA 602 (2007) 37–46

available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/aca

Modelling the quality of enantiomeric separations using Mutual Information as an alternative variable selection technique

Sónia Caetano[a], *Catherine Krier*[b], *Michel Verleysen*[b], *Yvan Vander Heyden*[a,*]

[a] *FABI, Department of Analytical Chemistry and Pharmaceutical Technology, Vrije Universiteit Brussel – VUB, Laarbeeklaan 103, Brussels 1090, Belgium*

[b] *Université Catholique de Louvain, Machine Learning Group, Place du Levant 3, Louvain-la-Neuve 1348, Belgium*

ARTICLE INFO

Article history:
Received 26 March 2007
Received in revised form 6 July 2007
Accepted 30 August 2007
Published on line 4 September 2007

Keywords:
Enantioseparation
Mutual Information
Variable selection

ABSTRACT

This paper uses Mutual Information as an alternative variable selection method for quantitative structure–property relationships data. To evaluate the performance of this criterion, the enantioselectivity of 67 molecules, in three different chiral stationary phases, is modelled. Partial Least Squares together with three commonly used variable selection techniques was evaluated and then compared with the results obtained when using Mutual Information together with Support Vector Machines. The results show not only that variable selection is a necessary step in quantitative structure–property relationship modelling, but also that Mutual Information associated with Support Vector Machines is a valuable alternative to Partial Least Squares together with correlation between the explanatory and the response variables or Genetic Algorithms. This study also demonstrates that by producing models that use a rather small set of variables the interpretation can be also be improved.

1. Introduction

In the pharmaceutical field, molecular chirality plays a main role in the activity of drugs, which makes the identification and separation of enantiomers extremely important. The enantiomers of a given compound only differ in their optical activity and their interaction with other chiral molecules. Thus, when interacting with other chiral molecules, enantiomers should be regarded as different chemical compounds, as they may show significant differences in their interactions. Hence, developing techniques that allow the identification, separation and quantification of enantiomers is very important for the pharmaceutical analysis.

High-performance liquid chromatography (HPLC) with chiral stationary phases (CSPs) is one of the most widely used techniques to perform the direct separation of enantiomers [1]. However, at present, the choice of a CSP to perform the chiral separation of a substance is still a trial and error task, making the selection time consuming and uneconomic. Thus, the development of models that are able to predict whether or not a certain CSP is able to perform a given chiral separation would be of great benefit.

Molecular descriptors have been used to study the relationships that exist between the structure of organic compounds and many of their physical, chemical and biological properties. They are numbers that characterize the constitution and configuration of the molecule and can be used to build a model that allows predicting some interesting molecular properties [2].

The simplicity in the determination of molecular descriptors, that nowadays computer software provides, allows the calculation of hundreds of descriptors for a single molecule.

* *Corresponding author*. Tel.: +32 2 477 47 34; fax: +32 2 477 47 35.
E-mail address: yvanvdh@vub.ac.be (Y.V. Heyden).
0003-2670/$ – see front matter © 2007 Published by Elsevier B.V.
doi:10.1016/j.aca.2007.08.048

Therefore, the quantitative analysis of these large amounts of data leads to an increasing use of chemometrical techniques, such as multivariate calibration methods, in the quantitative structure–activity relationships (QSAR) field. Nevertheless, and even if chemometrical techniques are able to deal with large datasets, sometimes it is essential to reduce the number of variables that are used to model a certain property, as there are some methods, like multiple linear regression (MLR), that are unable to deal with data containing more variables than objects. Indeed using techniques that allow the selection of a reduced set of variables containing the most important information enables, not only a better interpretation and comprehension of the model, but also a better model itself, as the exclusion of uninformative variables can be seen as elimination of noise.

The Mutual Information (MI) is a statistical measure of the relation that exists between two variables. Unlike other parametric estimators, such as the correlation, the MI does not make any assumption about what type of relation could exist between the variables. It can thus be used in a wide range of contexts, including for the selection of variables.

The aim of this study is first to demonstrate that MI can be applied to quantitative structure–property relationship (QSPR) data, as an alternative variable selection technique, and secondly that even though the number of variables is significantly reduced, it is still possible to produce models with a good performance and that can be interpreted. Finally, the last goal is to demonstrate that different groups of molecular descriptors can lead to models with similar predictive power. This second property opens the possibility to an intelligent selection of molecular descriptors, since some are easier to obtain than others, for a similar prediction power.

This paper will show that the MI can be used as an interesting alternative for variable selection, when a nonlinear prediction model is used. For this purpose, after a brief description of the molecular descriptors in Section 2.1, Section 2.2.1 will remind the conventional linear Partial Least Squares (PLS) method, Section 2.2.2 will explain the use of the MI for variable selection, and Section 2.2.3 will describe Support Vector Machines (SVMs).

2. Theory

2.1. Molecular descriptors

Molecular descriptors have seen a great increase on their use, especially in the field of QSAR. The success of QSAR encouraged scientists, particularly in the pharmaceutical area, to investigate the relation between molecular parameters and properties other than activity.

A molecular descriptor is a number extracted from a defined molecular representation or a well-specified experimental procedure, i.e., it is a value that results from the transformation of the information contained in a symbolic representation of a molecule, or from a standardized experiment. This leads to a classification of the molecular descriptors in two main groups, experimental descriptors, such as dipole moment, and theoretical descriptors, like GETAWAY descriptors, which are calculated based on the symbolic representation of the molecule [2].

Since a molecule can be represented in different ways, theoretical molecular descriptors are divided in 0D-, 1D-, 2D-, 3D- and 4D-descriptors [2]. In this study, all types of descriptors, except for 4D-descriptors, are used.

The main advantage of using 0-, 1- and 2D descriptors is that the user does not have to optimize the geometry of the molecule, which is a time consuming task. However, because they do not consider the molecule as a three-dimensional object, information like the spherosity of a molecule cannot be accessed.

2.2. Methods

2.2.1. Partial Least Squares (PLS)

The theory of PLS has been thoroughly discussed in numerous publications [3–5]. Here, only a short description will be made.

The aim of Partial Least Squares is to model the relationship between $\mathbf{X}$ and $\mathbf{y}$ by using a set of latent variables that maximize the covariance between them. PLS is a latent variables regression method, i.e. the model regression vector is estimated by means of factors, which are linear combinations of the original variables. The PLS factors are computed in order to maximize the covariance between $\mathbf{X}$ and $\mathbf{y}$. Hence, $\mathbf{X}$ information orthogonal to $\mathbf{y}$ is not extracted by the PLS approach, and the corresponding models usually use less factors than their PCR equivalent. The PLS model can be written as:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \tag{1}$$

$$\mathbf{y} = \mathbf{T}\mathbf{q}^{\mathrm{T}} + \mathbf{f} \tag{2}$$

where $\mathbf{y}$ ($m \times 1$) is the vector of responses $\mathbf{T}$ ($m \times r$) is the score matrix, $\mathbf{P}$ ($n \times r$), and $\mathbf{q}$ ($1 \times r$) are the loading matrix and vector, $\mathbf{E}$ and $\mathbf{f}$ are the $\mathbf{X}$ residual matrix and $\mathbf{y}$ residual vector after the projection of $\mathbf{X}$ and $\mathbf{y}$, respectively, m represents the number of objects, r is the number of selected factors, and n is the number of variables. These components can then be used for visual inspection of the data. The dimensionality of the model, m, i.e., the optimal number of PLS components, can be determined using cross-validation (CV).

For new data, it is possible to predict the response based on the PLS model that was previously built. The predicted $\mathbf{y}$ values ($\hat{\mathbf{y}}$) can be determined by:

$$\hat{y} = \mathbf{m}_{\mathbf{y}} + (\mathbf{x}_{new} - \mathbf{m}_{\mathbf{X}})^{\mathrm{T}}\mathbf{b} \tag{3}$$

where $\mathbf{m}_{\mathbf{y}}$ and $\mathbf{m}_{\mathbf{X}}$ are the mean of $\mathbf{y}$ and $\mathbf{X}$, respectively, $\mathbf{x}_{\mathbf{new}}$ the new incoming data, and $\mathbf{b}$ the regression coefficient vector, that can be calculated according to Eq. (4):

$$\mathbf{b} = \mathbf{W}(\mathbf{P}^{\mathrm{T}}\mathbf{W})^{-1}\mathbf{q} \tag{4}$$

where $\mathbf{W}$ is the matrix of the loading weights, and $\mathbf{P}$ and $\mathbf{q}$ have the same meaning as in Eqs. (1) and (2).

2.2.2. Mutual Information

When a large number of descriptors are available, it is usually not a good idea to use all of them as inputs to a prediction (classification or regression) model. Indeed too many inputs lead to the so-called curse of dimensionality, an expression that gathers all difficulties linked to the fact that designing a model in a high-dimensional space is much more difficult than in a lower-dimensional one. For example, the number of data necessary (in theory) to build a model with a predefined quality grows exponentially with the number of inputs, although in practice more than a few tens or hundreds of data are seldom available in chemometric applications.

There is thus a need to reduce the number of inputs to a model. This can be done by selecting the most relevant ones from the initial set (in this case, the set of molecular descriptors). To assess the relevance of variables, two ingredients are necessary. First, a measure is needed to assess in an objective way to which extend a variable, or a set of variables, is relevant to the problem. Second, once this measure is available, one has to build successive candidate variables sets and assess them individually. The most relevant set is then chosen. However, as all combinations of variables cannot be used as candidate sets because of computational reasons, this means that a procedure has to be chosen to scan possible sets among all possible ones. The relevance criterion and the selection procedure are detailed in the following paragraphs. As a measure of relevance, the correlation (between each of the inputs or molecular descriptors, and the quantity to predict) might be used. Nevertheless the correlation suffers from two drawbacks. First, it is restricted to the measure of linear relations; as an example, the correlation between a uniform centered random variable and the same variable to the power two is zero, while there is obviously a relation between both of them. Second, the correlation is usually limited to two variables, while there is a need (see forward procedure below) for a measure between groups of variables.

The MI is a better measure of the relations between variables (therefore, of the relevance of a variable as input to a prediction model). The MI is based on Shannon's information theory [6]. More information about the MI concepts and definition can be found in [7–9]. The MI estimator works on normalized variables $\mathbf{x}$ and $\mathbf{y}$, and therefore, even though the MI values are not limited by an upper value (such as the correlation which is bounded by 1, for example), comparisons between estimated MI are relevant. In the following variable selection procedure, only comparisons between MI values are performed.

An important comment must be made here. If a linear prediction model is built after the variable selection, there is no need to use a nonlinear relevance measure such as the MI. Indeed, methods such as PLS provide an optimal choice of the factors, given the fact that the subsequent prediction model is chosen (in this case, a linear regression on the factors). The interest of a nonlinear criterion like MI rises when a nonlinear prediction model, a priori more powerful than a linear one, is used. In this case, the use of a linear selection criterion would be suboptimal.

The concept of MI is particularly suited to measure the relevance of a variable $\mathbf{x}$ for predicting $\mathbf{y}$, i.e. of the adequacy of $\mathbf{x}$ as input to a model aiming at predicting $\mathbf{y}$. However, the exact measure of the MI is not possible in practice. Indeed, the exact measure of MI is possible only when the probability density functions (PDF) of $\mathbf{x}$ and $\mathbf{y}$ are known. In practice, the PDF are not known (we only know a few samples, not the distribution of data) and must be estimated. MI estimators have thus been developed to compute an approximation of the MI between $\mathbf{x}$ and $\mathbf{y}$, i.e., $I(\mathbf{x},\mathbf{y})$, in the finite sample case. Histograms and kernels may be used for that purpose. However, a recently published k-nearest neighbours (k-NN) based estimator gives better performances in particular when $\mathbf{x}$ is a vector instead of a scalar variable [7]. This estimator is used in this work. It necessitates to fix k, i.e. the number of neighbours in the k-NN procedure embedded in the MI estimator. The choice of k will be discussed in Section 3 of this paper.

Once a relevance criterion for variables is defined, a procedure to select the variables has to be designed. Let us denote by $\{\mathbf{x}_i\}$, $1 \le i \le n$ the set of all variables (among which a few of them should be selected). The first variable $\mathbf{x}_1$ to select is obviously the one that maximizes the MI with the response $\mathbf{y}$ to predict. To select the second variable, two options are possible. First, the second variable may be selected as the one that maximizes the MI with $\mathbf{y}$, with the exception of $\mathbf{x}_1$ (already selected). This choice may however lead to a variable $\mathbf{x}_2$ that is highly correlated (or highly similar in terms of Mutual Information) with the first selected variable $\mathbf{x}_1$. Although the MI with $\mathbf{y}$ is important, the information added by $\mathbf{x}_2$ may be small, as $\mathbf{x}_1$ is already known. Another option is therefore to select $\mathbf{x}_2$ such that the MI between the group $\{\mathbf{x}_1, \mathbf{x}_2\}$ and i is large. These two options will be referred to as *ranking* and *forward* procedures, respectively, in the following. The two procedures have advantages and drawbacks. The forward procedure selects variables that are as independent as possible between them, maximizing the information added in the set. However, it may fail when the important information is contained in the differences between $\mathbf{x}_i$ variables, even if the variables themselves are largely dependent. As, it will be shown in Section 3 of this paper, both procedures may lead to interesting results, and should thus be considered.

2.2.3. Support Vector Machines for regression

Support Vector Machines (SVMs) can be applied both to classification and regression tasks, as several works have shown [10–12]. The number of SVM applications has been growing in the last years, mainly due to their ability to model complex non-linear relationships by using a suitable kernel function, which transforms the input space to a higher-dimensional feature space where the non-linear relationships can be represented in a linear form.

The theory of Support Vector Regression (SVR), i.e., SVM applied to regression, has been developed by Vapnik in 1995 and is extensively described in the literature [10,13,14]. Here, only a brief description is included.

In SVR, given a training set $\mathbf{X}$ and the property of interest (output variable) $\mathbf{y}$, which in this case are the matrix of molecular descriptors and the selectivities (α) of the molecules in a certain CSP, respectively, the predicted value of the selectivity

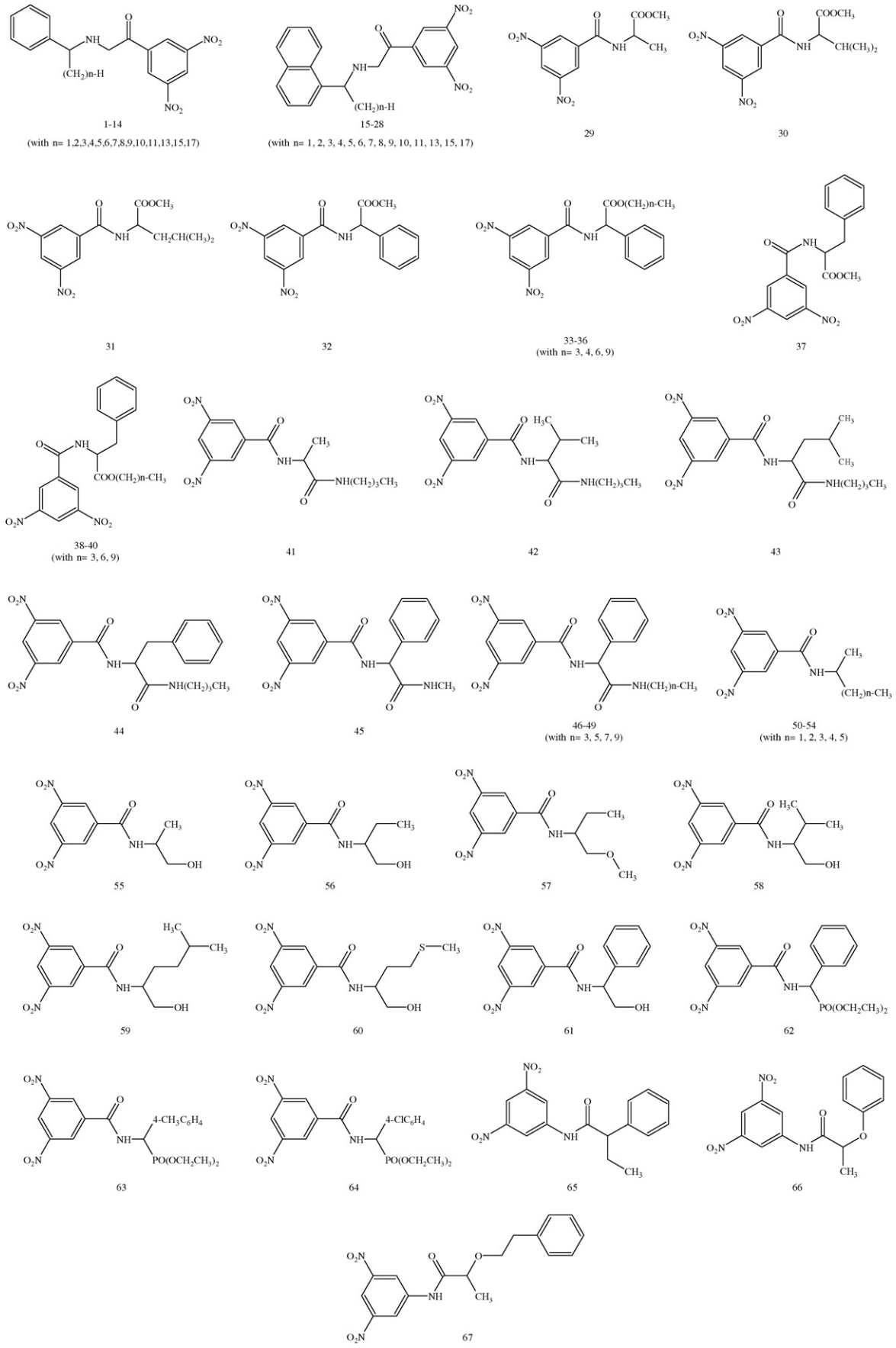

Fig. 1 – Structure of the molecules used.

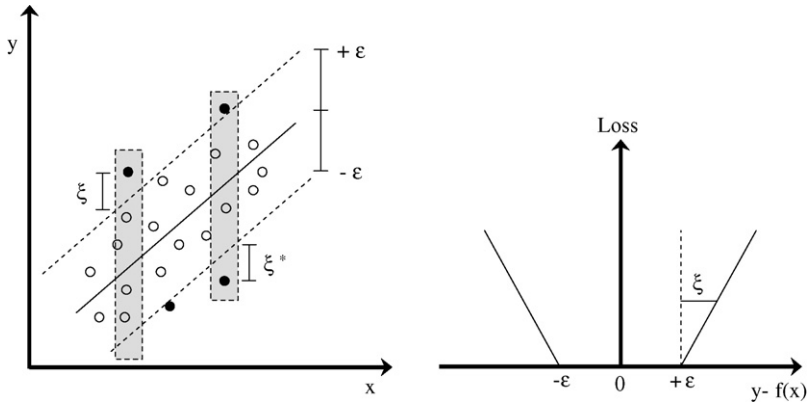

Fig. 2 – Left: In SVR a tube with radius ε is fitted to the data. Predictions that are larger than $\pm\varepsilon$ are taken into account by ξ and ξ^*, respectively. The black dots represent the support vectors, which are located outside of the ε tube. Right: The ε-insensitive loss function showing that deviations from the tolerance band are penalised. The slopes are determined by C (see text).

($\hat{\mathbf{y}}$) for an unknown molecule will be determined as:

$$\hat{\mathbf{y}} = \mathbf{f}(\mathbf{x}, \beta, \beta^*) = \sum_{i=1}^{n} (\beta_i - \beta_{i*}) K(\mathbf{x}, \mathbf{x}_i) + e, \quad (5)$$

where β and β^* are the Lagrange multipliers, **e** is bias and **K** ($\mathbf{x}$,$\mathbf{x}_i$) a generic kernel function. The kernel map function transforms the input space into a high-dimensional feature space in which the solution for the problem is linear. Several kernel functions can be used, among which the radial basis function (RBF) ($K(\mathbf{x}, \mathbf{x}_i) = \exp(-\gamma\|\mathbf{x} - \mathbf{x}_i\|^2)$) and the polynomial are the most employed. In this study RBF kernels are used [15].

Vectors β and β^* are the solutions of the following constrained quadratic programming problem:

$$\text{maximize} \sum_{i=1}^{n} y_i(\beta_i^* + \beta_i) - \varepsilon \sum_{i=1}^{n} (\beta_i^* + \beta_i) - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (\beta_i^* + \beta_i)(\beta_j^* + \beta_j) \mathbf{K}(\mathrm{xi}_{,\mathrm{x}_j}) \quad (6)$$

subject to:

$$0 \le \beta_i, \beta_i^* \le C \quad i = 1, 2, \ldots, n$$
$$\sum_{i=1}^{n} \beta_i^* = \sum_{i=1}^{n} \beta_i \quad , \quad (7)$$

where ε is the size of the ε-insensitive zone of the loss function which penalises errors that are larger than $\pm\varepsilon$. Predictions deviating more than $\pm\varepsilon$ are taken into account by the slack variables (ξ and ξ^*) (see Fig. 1). It should be noted that ε does not indicate the desired predicted error of the model, but is a characteristic of the prediction error penalty. C is a regularization constant which is used to determine the trade-off between the model complexity and the amount up to which deviations larger than ε are tolerated.

Most of the elements of vectors β and β^* are equal to zero [10]. The non-zero values are associated to a few x_i that are called support vectors.

The choice of the necessary parameters for the application of SVM to regression (γ, ε and C) must be done by the user.

3. Experimental results

3.1. Dataset

The molecules used in this study were selected from the literature and consists of 67 hydantoins (Fig. 2), separated on HPLC columns with three urea-linked α-arylalkylaminine derived CSPs [16]. For all CSPs, the mobile phase used was ispropanol-*n*-hexane, with composition 20:80 (v/v) for the first and second (datasets 1 and 2), and 10:90 (v/v) for the third (dataset 3).

3.2. Descriptors

The calculation of the descriptors for each molecule was based on their geometrical structure optimized using Hyperchem® 6.03 professional software (Hypercube, Gainesville, FL, USA). Geometry optimization was done using the Molecular Mechanics Force Field method (MM+) and the Polak–Ribière conjugate gradient algorithm with an RMS gradient of 0.05 kcal (Å mol)$^{-1}$ as stopping criterion. The matrices with the positions of the atoms, as Cartesian coordinates, resulting from the geometrical representation of the molecule, were used to calculate the molecular descriptors. For each substance, 1630 descriptors were calculated with Dragon 5.0 Professional version [17]. From these 313 were deleted since they presented constant or nearly constant values.

Since the molecules used are the same, the descriptors, i.e., the **X** matrix used in the study, are always the same. The difference between datasets is the response variable. By using different CSPs, the time that a given enantiomer is retained by CSP is different and thus, different selectivity values are obtained.

The data used in this study is available to download from http://www.ucl.ac.be/mlg/index.php?page=DataBases.

4. Results and discussion

4.1. *Modelling selectivity*

To model the α of the enantiomeric separations of the 67 molecules, three different methodologies were used: PLS, associated with different variable selection techniques, stepwise-MLR and SVM together with MI. The results of stepwise-MLR are not shown here as the models obtained clearly over-fitted.

The samples were divided into calibration (40 samples) and test (27 samples) sets by using the DUPLEX algorithm [18]. For all methodologies and CSPs the same training and testing sets were used.

4.1.1. *Partial Least Squares*

The PLS models for α were built by using the calibration set and considering all variables, i.e., 1317 molecular descriptors, and then were tested by using the remaining samples. The optimal number of factors, when building the model with the calibration set, was determined using both Leave One Out CV (LOOCV) [19] and Monte Carlo CV (MCCV) [20], the latter to prevent over-fitting. Table 1 shows the results obtained. It can be seen that the Root Mean Squared Error of MCCV (RMSEMCCV) is always higher than the Root Mean Squared Error of Prediction, for the test set, (RMSEP). The main reason for this behaviour is that when some of the samples, in particular samples 9, 16 and 35, of the calibration set, are left out during the CV procedure, they are very badly predicted, increasing enormously the RMSEMCCV. The reason for the special behaviour of these samples is not yet known.

Since a model built with 1317 variables is impossible to interpret, three variable selection techniques that are commonly associated with PLS were applied. For each variable selection technique, 12 variables were selected. The reason for choosing 12 variables, besides comparison purposes, is given in Section 4.1.2.

The first technique applied was the selection of the variables corresponding to the highest absolute **b** coefficients. The variables selected for each dataset are given in Table 2.

From the results presented in Table 1 it can be seen that by decreasing the number of variables considered, from the initial 1317 (no variable selection) to the 12 variables with the highest **b** coefficients, the RMSEMCCV is improved, and the number of latent variables needed to build the model is reduced, but the RMSEP increases for both dataset 1 and 2.

The second variable selection technique that was taken into account was the correlation between the molecular descriptors and the response. Thus, the 12 variables that presented the highest absolute correlation values were kept (see Table 2). Table 1 shows again a decrease of both the RMSEMCCV and the number of factors, and also of the RMSEP. When comparing the results obtained using the **b** coefficients and the correlations, it can be seen that correlation performs better. When looking at the descriptors selected by both techniques, with the exception of dataset 3, quite different descriptors can be found.

Finally, genetic algorithms were used to select the best set of 12 variables to model α for each dataset. The best set of descriptors, i.e., the one that presented the smallest RMSEMCCV, was chosen after 10 runs, having each run 500 generations. The results shown in Table 1 demonstrate that by using PLS coupled with GA it is possible to obtain models with a better RMSEMCCV but, with the exception of dataset 3, no improvement of the RMSEP. Concerning the variables that are selected (see Table 2) it can be seen that there are hardly any variables in common between those that were kept using GA and the other techniques.

Even though the results obtained for PLS-GA, when considering the RMSEMCCV, are much better that those obtained when using the complete dataset and interpretation of the 12 variables selected in each model is now possible, there is still the problem of reproducibility of the GA results. Also, the use of the variables with the highest correlation together with PLS provides better results, then when no variable selection is considered. Nevertheless, the fact that the selected variables might have also a high correlation between themselves is not being taken into account, meaning that it can happen that some variables are providing the same information.

Table 1 – PLS results for the three datasets, using different variable selection techniques

Dataset	Variable selection	No. of factors	RMSEMCCV	RMSEP
1	None	6	1.27	0.73
	b-values	2	0.97	0.93
	Correlation	4	0.87	0.60
	GA	3	0.80	0.82
2	None	6	1.18	0.87
	b-values	3	0.93	1.23
	Correlation	3	0.98	0.78
	GA	6	0.69	1.49
3	None	2	0.56	0.48
	b-values	1	0.48	0.48
	Correlation	1	0.50	0.46
	GA	2	0.46	0.45

The number of variables chosen with the variables selection techniques was 12. For dataset 1 the **y**-values vary between 1 and 6.45, for dataset 2 between 1 and 8.44, and for dataset 3 between 1 and 3.39.

Table 2 – Variables used to build PLS models for each dataset after the application of the three variable selection techniques

Dataset	Technique	Variables selected				
		0D	1D	2D	3D	Others
1	b-values	nN	N-072, C-005, O-060, nHDon, H-050, nCONHR, nCOOR	T(N..N)	RDF040e, DISPe, G(N..N)	–
	Correlation	nO	H-051, O-060, nCOOR	ESpm13d, ESpm14d, ESpm15d, SEigp, T(O..O), ESpm01d, GATS7e	–	–
	GA	R6e+	nCOOR	GATS7e, EEig15r, ESpm14d, ESpm11r, piID, Dz	RDF040u, DISPe, RDF105u, Mor14u	–
2	b-values	nN	O-060, H-050, nHDon, nCONHR, nCOOR	T(N..N)	E2e, DISPm, DISPe, E3s, G(N..N)	–
	Correlation	nO	O-060, nCOOR	T(O..O), ZM1V, DELS	RDF040u, RDF045m, RDF050u, RDF040m, RDF050e, RDF040e	–
	GA	–	–	piPC03, piPC06, ESpm10r, CID	Mor09p, G1, SP03, E3s, DISPm, R1u	PSA, ALOGP
3	b-values	RBF	–	PW3, X2A, PCR, GATS5m, PW5, X3A, PW3, X1A,	Mor15v, DISPv, Mor15p	–
	Correlation	Mv	–	PW5, X4A, PW4, X3A, GATS2p, PW3, X1A	Mor15v, Mor13p, Mor15p, DISPv	–
	GA	Se	–	BEHp5, ATS1m, EEig06d	L2m, H3m, RDF065u, R8m+, Mor15p, R6u, Mor17m, H6u	–

0D, 1D, 2D, 3D and others (charge descriptors and molecular properties) correspond to the type of molecular descriptor selected. The definition and explanation of the descriptors can be found in [2].

Thus, to try to solve this problem MI was used to determine the most informative variables to model α.

4.1.2. *Variable selection with Mutual Information*

To model the α of the enantiomeric separations of the 67 molecules, for the three different CSPs, both ranking and forward procedures for the MI were applied in order to select the most suitable variables to describe the selectivity (see Fig. 3). For both procedures the MI was estimated using six neighbours, i.e., $k = 6$ in the procedure described in [4]. The choice of $k = 6$ was based on references [9,21].

The selection of the variables was done using the calibration sets, i.e., 40 molecules, and that the maximum number of variables that could be extracted was 12. This maximum number of variables has been set in order not only to limit the computation time of the MI estimator, but also the complexity

Fig. 3 – (a) Ranking and (b) forward procedures to determine the MI for dataset 1 using $k = 6$. Solid black and grey lines: MI calculated for the molecular descriptors and random variables respectively, (—) cut-off (maximum value of MI for the random variables); (●) optimal number of variables selected, i.e., group of descriptors that gives the maximum information.

Table 3 – Variables selected by the MI for $k = 6$ and for k between 5 and 7 for the ranking (average) procedure

Dataset	Procedure	0D	1D	2D	3D	Others
1	Ranking	–	–	ESpm01d, GGI3, MWC10, EEig05x, GGI2, ESpm07u, ESpm05u, MWC07, Qindex, MWC08, ESpm09u, MWC04	–	–
	Forward	–	nCOOR, O-060, C-005, C-008, C-001	ESpm01d	–	LAI
	Ranking (average)	–	–	GGI3, ESpm01d, MWC10, GGI2, ESpm07u, EEig05x, Qindex, ESpm05u, ESpm09u, MWC09, MWC07, ESpm11u	–	–
2	Ranking	nO, Mp	H-047	JGI9, MWC02, MWC03, X5A, SEigp, SEigv, Qindex, MWC08	R8m	–
	Forward	–	O-060, C-008	JGI9	–	Hypnotic-50
	Ranking (average)	nO	H-047	JGI9, MWC03, MWC02, SEigv, T(O..O), SEigp, X5A, Qindex, MWC08	R8m	–
3	Ranking	Mv, ARR	–	PW5, X5A, MWC10, MWC04, GGI3, X4A, RBF	R7v+, R7p+, MEcc	–
	Forward	–	nCOOR, C-026	RBF	G(N..P)	Psychotic-50
	Ranking (average)	Mv, ARR	nCs, H-052	RBF, X5A, PW5, MWC10, X1A, X4A	R7p+, MEcc	–

0D, 1D, 2D, 3D and others (charge descriptors and molecular properties) correspond to the type of molecular descriptor selected. The definition and explanation of the descriptors can be found in [2].

of the final models. In addition, when the forward procedure is used, the quality of the MI estimator decreases with the number of variables. It is thus reasonable to fix a limit that takes both the computation time and accuracy issues into account.

Fig. 3 shows how the descriptors for dataset 1 are selected by each procedure. In the ranking procedure (Fig. 3a), the 12 variables with the highest MI are kept, while in forward procedure (Fig. 3b), the optimal number of variables is determined by the maximum value of MI, which in this example is seven.

Table 3 shows the number of variables selected for each dataset.

4.1.3. *Support Vector Machines for regression*

For comparison purposes, three SVM models, one for each dataset, using all variables, i.e., the initial 1317 molecular descriptors, were built. The models were built using the LIBSVM package [15] and a grid search to optimize γ (width of the Gaussian) and C (the error weight). The parameter ε (insensitivity) was kept constant and equal to 0.1 in all models. This option was taken due to the fact that a full optimization of γ, C and ε would have been extremely time consuming. The results obtained from this partial optimization were satisfactory, and can be seen in Table 4.

The variables selected by the MI using both the ranking and forward procedures (see Table 3) are then used to build SVM models, using 10-fold CV. The choice of SVM as modelling technique is related with the fact that when using MI as a variable selection technique, the choice of the variables to be kept is done in a non-linear way. Therefore, SVM was chosen, since it is a modelling technique that is able to deal with non-linear relationships.

As can be seen in Table 4 the Root Mean Squared Error of Cross-Validation (RMSECV) is always higher than the RMSEP. The reason for this behaviour was already explained in the previous section. Also the reduction of the number of variables, in particular in dataset 1, improves the models, as it is possible to see a decrease of the errors values.

Table 4 – Parameters and errors of the SVM models using all variables, i.e., 1317 molecular descriptors, and those selected by MI

Dataset	Procedure	No. of selected variables	γ	C	RMSECV	RMSEP
1	None	None	8.3×10^{-7}	16,384	1.05	0.77
	Ranking	12	0.37	256	0.80	0.65
	Forward	7	3.4×10^{-4}	512	0.92	0.31
	Average	12	0.37	222.86	0.79	0.73
2	None	None	6.9×10^{-6}	4,096	1.05	0.74
	Ranking	12	0.14	5.28	1.01	0.80
	Forward	4	0.02	128	0.99	0.95
	Average	12	0.14	4.59	0.91	0.78
3	None	None	1.6×10^{-3}	4	0.45	0.38
	Ranking	12	0.27	4	0.48	0.44
	Forward	5	0.33	3.78	0.34	0.35
	Average	12	0.29	4	0.47	0.42

It can be seen in Table 4 that again, by doing a variable selection, the models are improved and that, with the exception of dataset 2, the RMSEP is higher for the ranking procedure than for the forward. It can also be verified that the RMSEP obtained for the dataset 1 (forward procedure) and dataset 3 (both ranking and forward procedures) are comparable or better to those of PLS together with GA or correlation (see Table 1), and that the variables selected by using each method are quite different. In fact, for dataset 1, if someone wants to determine the α of a new enantiomeric separation using the model obtained with PLS-GA he has to determine 0D, 1D, 2D and 3D molecular descriptors, while for the SVM-MI model only 2D descriptors are necessary. This means that the last model will be easier to use, as there is no need of optimizing the geometry of the molecule to determine the necessary variables to apply the model.

When comparing the variables selected by the MI procedure (Table 3) it is seen that, except for dataset 2, the variables selected by the ranking procedure originate models with lower predictive power, i.e., RMSEP, which shows that the selection of variables based on their combined information is a more efficient technique. The definition and explanation of the descriptors can be found in [2].

As it was mentioned before, MI was applied using $k = 6$, based on references [9,21]. Nevertheless, there is so far no way of finding the optimal k value. Therefore, the MI, using the ranking procedure, between the explanatory variables and the response was also calculated on the calibration set using $k = 5$ and $k = 7$. To try to make the variable selection using MI more robust, the average of the MI for k between 5 and 7 was determined, and the 12 variables that presented the highest values were kept (Table 3).

The models obtained by using the average MI and SVM are, in general, better than those determined for the ranking procedure with $k = 6$, and they are as good as or better than those found with PLS-correlation or PLS-GA. When comparing Tables 2 and 3 it can be seen that MI always selects a majority of 2D-descriptors, while for GA and correlation, with the exception of dataset 1, it is 3D. The variables selected by the ranking procedure using the average between $k = 5$ and $k = 7$ and $k = 6$ have at least nine variables in common, which shows that the initial choice of $k = 6$ was suitable.

4.1.4. *Interpretation of the best models*

For each dataset the model that presented the best predictive ability was chosen to be interpreted. Therefore, in the following paragraphs, a short description of the variables used in each of these models will be made. It should be noticed that this description is organized in groups, i.e., GETAWAY, 2D autocorrelations, etc., and not individually. This option was taken due to the fact that an individual interpretation is rather difficult, as for theoretical molecular descriptors not always is possible to directly link them to physicochemical properties.

The model with the best performance for dataset 1 was obtained when using SVM together with the forward procedure for the MI. This model contains seven variables (Table 3): nCOOR, O-060, C-005, C-008, C-001, ESpm01d and LAI. The nCOOR is a functional group counts descriptor related with the number of aliphatic esters in the molecule [17]. The descriptors O-060, C-005, C-008 and C-001 belong to the atom-centred fragments group. They give information about the number of predefined structural features in the molecule, which in this case are Al–O–Ar/Ar–O–Ar/R..O..R/R–O–C = X, for O-060, CH3X, for C-005, CHR2X, for C-008 and CH3R/CH4, for C-001 (R represents any group linked through the carbon atom, X any electronegative atom (O, N, S, P, Se, halogens) and Ar an aromatic group) [2,17]. The ESpm01d is an edge adjacency index and it corresponds to the first spectral moment from the edge adjacency matrix weighted by dipole moments [2,17]. This descriptor accounts for size of the molecule and heteroatoms. The last descriptor selected in this model, LAI, is a molecular property named Lipinski alert index, which is a drug-like index.

For dataset 2, the model that has the best predictive ability was that obtained using SVM without variable selection. Since for this model, 1317 variables are considered, it is impossible to make any kind of interpretation based on the most important descriptors.

The best model for dataset 3 was found when using SVM together with the forward procedure for the MI. The five variables selected for the model are nCOOR, which was also selected for best model of dataset 1, C-026, RBF, G(N..P) and Psychotic-50 (Table 3). The descriptor C-026 is an atom-centred fragments group, and, as said before, it provides information about the number of predefined structural features in the molecule, which in this case is R–CX–R [2,17]. The constitutional descriptor RBF determines the rotatable bond fraction of the molecule [17]. G(N..P) is a geometrical descriptor that estimates the sum of geometrical distances between N..P. Finally, the molecular property Psychotic-50 is a drug-like index and it gives information about the Ghose–Viswanadhan–Wendoloski antipsychotic-like index at 50% [17].

5. Conclusion

It was seen that the reduction of the initial set of variables provides models with better predictive power. Of the three variable selection techniques applied together with PLS, correlation is the one that gives the best results. The selection of variables based on the **b** coefficients results in models with similar or worse predictive abilities. Concerning GA, only for dataset 3, slightly better results than those obtained using correlation were found.

This study shows that MI can be used as an alternative variable selection technique for QSPR data. MI is able to provide reduced subsets of variables, with similar predictive power to those obtained with the correlation between the variables and the response, with the advantages of being reproducible, fast to calculate and consider groups of variables instead of each one individually. SVM together with the forward procedure produced the best models for datasets 1 and 3. These results show that by considering the information provided by groups of variables not only more accurate models can be found, but also smaller, making their interpretation possible.

The MI estimator used in this study relies on an internal parameter k for which there is no clear choice, and therefore is a limitation. However, the choice of $k = 6$ seems to provide good results, and the average of the MI values, for the ranking procedure, for k values between 5 and 7 appears to be a

good alternative to cope with the problem. Nevertheless, more effort should be put in trying to find a way to determine the optimal k value.

It can also be seen that 0-, 1-, and 2D-descriptors can provide similar information to 3D, and therefore one can avoid the time consuming determination of 3D-descriptors.

REFERENCES

[1] Y. Vander Heyden, D. Mangelings, N. Matthijs, C. Perrin, Chiral separations, in: S. Ahuja, M. Dong (Eds.), Handbook of HPLC in Pharmaceutical Analysis, Academic Press/Elsevier, 2005, pp. 447–498.
[2] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley-VCH, Weinheim, 2000.
[3] M. Martens, T. Naes, Multivariate Calibration, Wiley, New York, 1989.
[4] R. Manne, Chemom. Intell. Lab. Syst. 2 (1987) 187–197.
[5] S. Wold, M. Sjöström, L. Eriksson, Chemom. Intell. Lab. Syst. 58 (2001) 109–130.
[6] C.E. Shannon, W. Weaver, The Mathematical Theory of Communication, University of Illinois Press, Urbana, IL, 1949.
[7] A. Kraskov, H. Stögbauer, P. Grassberger, Phys. Rev. E69 (2004) 066138.
[8] N. Benoudjit, D. Francois, M. Meurens, M. Verleysen, Chemom. Intell. Lab. Syst. 74 (2004) 243–251.
[9] F. Rossi, A. Lendasse, D. François, V. Wertz, M. Verleysen, Chemom. Intell. Lab. Syst. 80 (2006) 215–226.
[10] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods, Cambridge University Press, 2003.
[11] C.J.C. Burges, Data Min. Knowledge Discov. 2 (1998) 121–167.
[12] S.R. Gunn, Support Vector Machines for Classification and Regression (Technical Report), University of Southampton, UK, 1998.
[13] V. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 1995.
[14] A.L. Smola, B. Schölkopf, A Tutorial on Support Vector Regression, NeuroCOLT2 Technical Report Series, NC2-TR-1998-030, Royal Holloway University of London, UK, 1998.
[15] C.-C. Chang, C.-J. Lin, LIBSVM: A Library for Support Vector Machines, 2001 (software available to download on: http://www.csie.ntu.edu.tm/∼cjilin/libsvm).
[16] W.H. Pirkle, M. Ho Hyun, J. Chromatogr. 322 (1985) 295–307.
[17] R. Todeschini, V. Consonni, A. Mauri, M. Pavan, Dragon Professional Version 5.0, Milano Chemometrics and QSAR Research Group, 2004 (software available on: http://www.talete.mi.it/dragon.htm).
[18] R.D. Snee, Technometrics 19 (1977) 415–428.
[19] R.R. Picard, R.D. Cook, J. Am. Stat. Assoc. 79 (1984) 575–583.
[20] Q.S. Xu, Y.Z. Liang, Chemom. Intell. Lab. Syst. 56 (2001) 1–11.
[21] S. Harald, K. Alexander, A.A. Sergey, G. Peter, Phys. Rev. E 70 (2004), 066123, 1–17.