J. Braz. Chem. Soc., Vol. 13, No. 6, 742-753, 2002.
Printed in Brazil - ©2002 Sociedade Brasileira de Química
0103 - 5053 $6.00+0.00

Review

Multivariate QSAR

Márcia M. C. Ferreira*

Instituto de Química, Universidade Estadual de Campinas, CP 6154, 13083-862 Campinas - SP, Brazil

Neste trabalho, são apresentadas as técnicas usuais de quimiometria em estudos de relações quantitativas estrutura-atividade biológica (QSAR). Elas são introduzidas em ordem cronológica, iniciando pela análise de Hansch, e os métodos de análise exploratória de componentes principais e agrupamento hierárquico (PCA e HCA). Os métodos de regressão que usam a análise de componentes principais como fundamento (PCR e PLS) são apresentados a seguir. São introduzidos também os dois métodos de reconhecimento de padrões usados para classificação (KNN e SIMCA). Diferentes aplicações são usadas para ilustrar os diferentes métodos quimiométricos. É feito então um resumo da metodologia de regressão usada em QSAR-3D. Finalmente é apresentada uma nova metodologia, denominada PLS multilinear, já utilizada em química analítica, mas não explorada ainda pela comunidade de QSAR. Este método de regressão mantém a estrutura de ordem superior dos dados em lugar do desdobramento utilizado em QSAR-3D e apresenta vantagens computacionais, maior estabilidade e simplicidade do modelo, uma vez que o número de parâmetros calculados é bastante reduzido.

In this work, the chemometric techniques most frequently used in QSAR (quantitative structure-activity relationships) studies are reviewed. They are introduced in chronological order, beginning with Hansch analysis and the exploratory data analysis methods of principal components and hierarchical clustering (PCA and HCA). Principal component regression and partial least squares regression methods (PCR and PLS) are discussed, followed by the pattern recognition methods (KNN and SIMCA). Different applications are presented to illustrate these chemometric techniques. The methodology used for regression in 3D-QSAR is presented (unfolding PLS). Finally, the higher order method called Multilinear PLS, already used in analytical chemistry but not yet explored by the QSAR community, is introduced. This method maintains the multiway structure of the data and has several advantages over bilinear PLS including speed in calculation, simplicity and stability, since the number of parameters to be estimated can be greatly reduced.

Keywords: chemometrics, principal component analysis, partial least squares, SIMCA, KNN

1. Introduction

It is well accepted nowadays that physicochemical properties and structural attributes of chemical compounds govern their biological activities (biological response). The former include the ability of a molecule to cross cell membranes or to be taken up by fatty material (which is strongly influenced by solubility and lipophylicity), as well as its capacity of forming stable complexes and/or react with molecules in biological systems (this is governed by the electronic distribution via hydrogen bond, dipole and other intermolecular interactions), among others. The latter are related to the nature of the atoms assembled together by electronic forces and the way they are arranged in 3D space (which can be described in terms of chemical bonds, bond angles, torsion angles, molecular size and topology and other structural parameters), as well as the conformational flexibility of the functional groups.

The search for quantitative relations between the chemical "structure", *i.e.* physicochemical, structural and conformational properties and the biological response is the subject of quantitative structure-activity relationship (QSAR) studies. These relationships will hopefully help to understand and explain the driving forces behind the drugs action, ultimately supporting the development of new compounds exhibiting desirable biological properties.

QSAR and QSPR (quantitative structure-property relationships) have been known for a long time in organic chemistry, from observations that within a structural class, the smaller the variation in the chemical structure, the less the change in its physicochemical properties. If it is

* e-mail: marcia@iqm.unicamp.br

assumed that a uniform mechanism of action can be elicited for all the compounds within the class, it would be natural to expect that compounds with similar chemical and physical properties, would also have similar biological properties.

The foundation of QSAR as a practical tool of drug design had its concrete beginning in the early 1960s with the works of Hansch[1] and Free and Wilson.[2] One important contribution from the Free-Wilson method is the quantification of observation that changing a substituent at one position of a molecule is often independent of the effect of changing the substituent at a different position.

In the past 25 years, several other chemometric methods have broadened the arsenal of tools that can be applied to QSAR studies. Among them, the most widely used are: *i*) the exploratory data analysis methods of Principal Component Analysis (PCA)[3-6] and Hierarchical Cluster Analysis (HCA);[4,6] *ii*) the regression methods of Multiple Linear Regression (MLR),[3] Principal Component Regression (PCR)[3,4,7] and Partial Least Squares (PLS);[7,8] and finally, *iii*) the pattern recognition methods of Kth Nearest Neighbor (KNN)[9] and the Soft Independent Modeling of Class Analogy (SIMCA).[10] The most prominent application of PLS regression in QSAR is the Comparative Molecular Field Analysis (CoMFA) developed by Cramer *et al.*[11,12]

2. Chemometric methods in QSAR

2.1 Hansch analysis

Hammett,[13] in 1940*,* was the first who recognized that one chemical reaction could be taken as a standard process by proposing an equation (equation 1) to describe the ionization of benzoic acids in H_2O. Since steric effects do not play an important role in this reaction, the substituent constant σ was considered to represent some kind of electronic or polar effect when replacing –H by some substituent at *ortho*, *meta* or *para* position of the benzene ring:

$$\log K = \log K_0 + b_1\sigma \qquad (1)$$

In this equation log K_o and log K are the dissociation constants of the unsubstituted and substituted benzoic acid, respectively, σ is the Hammett electronic parameter and b_1 is the proportionality constant. Hansch extended Hammett's equation in terms of the electronic properties of the substituents as

$$\log {}^{1}\!/\!_{C} = b_o + b_1\sigma + b_2 \log P \qquad (2)$$

where log 1/C in equation 2 stands for the biological activity defined as the logarithm of the reciprocal of the effective dose ($C = ED$) or inhibitory concentration ($C = IC$) that is needed to produce a certain biological effect, and P is the octanol/water partition coefficient. The coefficients b_o, b_1 and b_2 can be determined by means of MLR. In the matrix form, Hansch's equation and the least square solution for the estimation of the regression vector **b** can be written as

$$\left[\mathbf{log}\frac{\mathbf{1}}{\mathbf{C}}\right] = \begin{array}{c} \begin{array}{ccc} & \sigma & \log P \end{array} \\ \begin{bmatrix} 1 & - & - \\ \vdots & \vdots & \vdots \\ 1 & - & - \end{bmatrix} \end{array} \begin{bmatrix} b_o \\ b_1 \\ b_2 \end{bmatrix} + \begin{bmatrix} \mathbf{e} \end{bmatrix} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} b_o + \begin{array}{c} \mathbf{x}_1 \\ \begin{bmatrix} - \\ \vdots \\ - \end{bmatrix} \end{array} b_1 + \begin{array}{c} \mathbf{x}_2 \\ \begin{bmatrix} - \\ \vdots \\ - \end{bmatrix} \end{array} b_2 + \begin{bmatrix} \mathbf{e} \end{bmatrix} \qquad (3)$$

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e} = b_o + b_1\mathbf{x}_1 + b_2\mathbf{x}_2 + \mathbf{e} \qquad \Rightarrow \qquad \hat{\mathbf{b}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}.$$

where, **y**, and **e** are respectively the vector of biological activities (physicochemical property, reactivity etc.) and residuals between measured and estimated activities. The vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ contain respectively the values of σ and log P for each compound. **X** is the matrix of independent variables with a juxtaposed column of ones and the superscripts T and –1 indicate the transpose and inverse of the matrix, respectively. Hansch's equation established a breakthrough in the area of QSAR.

The advantage of this equation is that it accounts for both electronic and hydrophobic interactions. The fundamental assumption in this approach is that contributions to σ and log P from several substituent groups in the same parent compound are additive, which means that the substituents are independent (the effect of a substituent group at one position in the parent molecule does not affect the substitution at other position). It is known that this assumption is violated for example, when hydrogen bonding occurs between two adjacent groups.

However, very frequently the data set does not fit well to this linear model and so the method was extended by the addition of non-linear terms and experimental/calculated parameters, among them the following: *i*) several quantum chemical properties related to the electron distribution in the molecule (dipole and quadrupole moments, polarizabilities, atomic charges, frontier orbital energies, heat of formation, etc.); *ii*) steric parameters (molecular volume, surface area, etc.); *iii*) molecular weight, molar refractivity; *iv*) connectivity (topological) indices given by a single number which expresses how atoms are arranged in a molecule; *v*) indicator variables, accounting for the presence or absence of specific chemical functions.

After including all these descriptors, the next question to be answered is: Which among all these parameters and functions are actually related to the biological activity, and should be included in the regression equation? By employing MLR for the regression analysis, two points must be observed: firstly, the ratio compounds/descriptors should be at least 5, and secondly, the chances are high that high correlations (multicolinearity) can appear among descriptors and this may cause spurious solutions. These problems may be solved in two ways: by making a variable selection and then employing the multiple linear regression as in Hansch's equation, or one could make use of multivariate techniques based on principal components such as PLS or PCR.

Another crucial point to be considered in any data analysis is preprocessing. The original data matrix usually does not have optimal value distribution for the analysis (*e. g.* has different units and variances in variables), which requires some pretreatment prior to data analysis. In general, the autoscale preprocessing, which results in scaled variables with zero mean and unit variance, is used.

2.2 Principal Component Analysis, PCA[3-6]

Most chemical applications of data analysis are by nature multivariate and one of the most suitable methods for such cases is PCA. This is a data compression method based on the correlation among variables. Its aim is to group those correlated variables, replacing the original descriptors by a new set called principal components, PCs, onto which the data is projected. These PCs are completely uncorrelated and are built as a simple linear combination of original variables. It is important to point out here that the PCs contain most of the variability in the data set, albeit in a much lower dimensional space. The first principal component, PC1, is defined in the direction of maximum variance of the whole data set. PC2 is the direction that describes the maximum variance in the orthogonal subspace to PC1. The subsequent components are taken orthogonal to those previously chosen and describe the maximum of the remaining variance. Once the redundancy is removed, only the first few principal components are required to describe most of the information contained in the original data set. The data matrix $\mathbf{X}(I \times J)$ corresponding to I molecules and J descriptors, is decomposed into two matrices, **T** and **L**, such that

$$\mathbf{X} = \mathbf{T}\mathbf{L}^{\mathrm{T}} \qquad (4)$$

The matrix **T**, known as the "score" matrix, represents the positions of the compounds in the new coordinate system whereas the PCs are the axes. **L** is the "loading" matrix whose columns describe how the new axis, *i. e.* the PCs, are built from the old axes. This is best done by using the singular value decomposition technique, SVD, which decomposes **X** into three matrices **U**, **S** and **V**, where **U** and **V** are orthogonal eigenvector square matrices and **S** is a diagonal matrix containing the singular values (equivalent to the square root of the eigenvalues). The product **U*S** is the score matrix **T**, while **V** corresponds to the loading matrix **L**. The square of each diagonal element of **S**, is equal to the amount of variance in the original data described by the corresponding principal component.

2.3 Hierarchical Cluster Analysis, HCA[3,6]

Hierarchical Cluster Analysis is another important multivariate method of data analysis. Its primary purpose is to display the data in such a way as to emphasize its natural clusters and patterns. The results, which are of qualitative nature, are presented in the form of a dendogram allowing one to visualize the samples or variables in a 2D space. The distances between samples or variables are calculated and transformed into a similarity matrix **S** (a usual notation that shouldn't be mistaken by the matrix of singular values from PCA) whose elements are the similarity indexes. For any two samples *k* and *l*, the similarity index is defined as

$$S_{kl} = 1.0 - \frac{d_{kl}}{d_{max}} \qquad (5)$$

where S_{kl} is an element of **S**, d_{max} is the largest distance for any pair of samples in the data set. d_{kl} is the Euclidean distance between samples *k* and *l* calculated as

$$d_{kl} = \sqrt{(x_{k1} - x_{l1})^2 + (x_{k2} - x_{l2})^2 + \cdots + (x_{kJ} - x_{lJ})^2} \qquad (6)$$

where x_{ij} is an element of the data matrix **X**.

The similarity scale ranges from zero to one. It is clear that the larger the index S_{kl}, the smaller the distance between *k* and *l*. Therefore, S_{kl} directly reflects their similarity.

2.3.1 Example: Antiviral activity of HIV-1 peptidic protease inhibitors[14]

The use of PCA and HCA is demonstrated by using a data set on forty eight peptidic HIV-1 protease inhibitors, having four substituents (P1, P1', P2, P2').[14] The biological activities are given in log units, pIC_{50}, (where IC_{50} is the required concentration of drug to reduce the viral production by 50%) ranging from 5.158 to 10.267. The molecules are classified as low/moderately (5.158 – 8.268 log units) and highly active (8.886 – 10.267 log units). The descriptors are fourteen *a priori* non-computer generated molecular properties recently described in reference 14. The analysis is carried out on autoscaled data. HCA shows that the samples are grouped into two main groups (Figure 1). In the smaller cluster **G1**, are the low/moderately active compounds. These are the smallest compounds where one or two of the substituents (P1, P1', P2, P2') are small or missing, and there is no -OH on P2'. P2' can be a small ring or acyclic system, what obviously reduces the activity.[14] The other group has three subgroups (**G1 - G4**). **G2** is a mixture of compounds covering the full range of activities. However, **G3** and **G4** contain only highly active compounds (exception are the compounds **40** and **42**). These are the largest compounds and have large P1', P2 and P2' substituents.

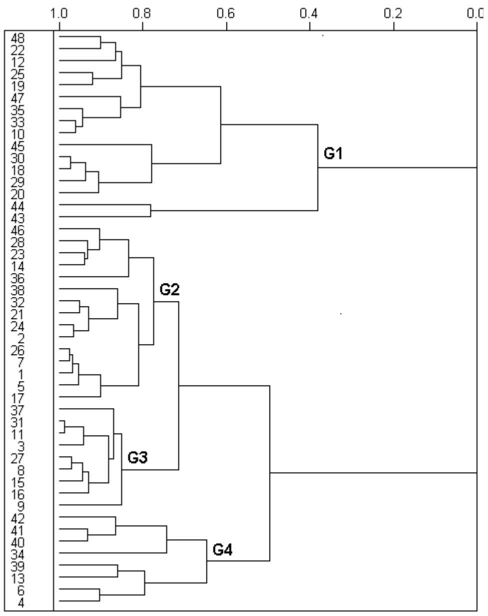


Figure 1. Hierarchical Cluster Analysis on 48 HIV-1 inhibitors.

PCA results show that PC1 discriminates low/moderately from highly active compounds (scores in Figure 2 separated by the curved line). The exceptions are compounds **14**, **23**, **28**, **40** and **42**. The contribution of molecular descriptors to particular PCs are in the loadings plot (Figure 2). They vary from 0.21 to 0.33, with higher contribution from X_2, X_3, X_9, X_{11}, X_{13} and X_{14}. PC1 is well expressed in terms of molecular size (cavity, bulk properties) and contents of various types of valence electrons (electronic and hydrophobic properties).[14] On the left side of the scores plot are the least active samples **43**, **44**, **47**, and on the right side of PC1 are the highest active **34**, **39**, **41**. This situation corresponds to the fact that the best inhibitors have maximal effective number of rings and substituents, and are rich in π-electrons and lone pair electrons from heteroatoms. PC2 mainly includes shape and electronic variables (X_1, X_4, X_6 and X_8) pointing out the complexity of steric and electrotopological properties. PC2 separates more branched molecules (top: **40-45**) from more compact ones (bottom: **9**, **10**, **12**, **35**, **22**, **58**, etc.) as is in Figure 2. Polar groups are important for PC3 where the most significant variables are X_{10} and X_{12} (loadings in Figure 3). From the scores in Figure 3, it can be seen that PC3 separates molecules (relatively to the size of their hydrocarbon parts) rich in electronegative atoms and polar groups (top; **43**, **22**, **48**, **6**, **13**, **39** etc.) from those having more hydrocarbon (aromatic or aliphatic) fragments (bottom: **20**, **29**, **30**, **18**, **34**, etc.). It can be said that PC3 has meaning of fine (valence electron) distribution of electron density, namely, polarity and hydrogen bond properties.

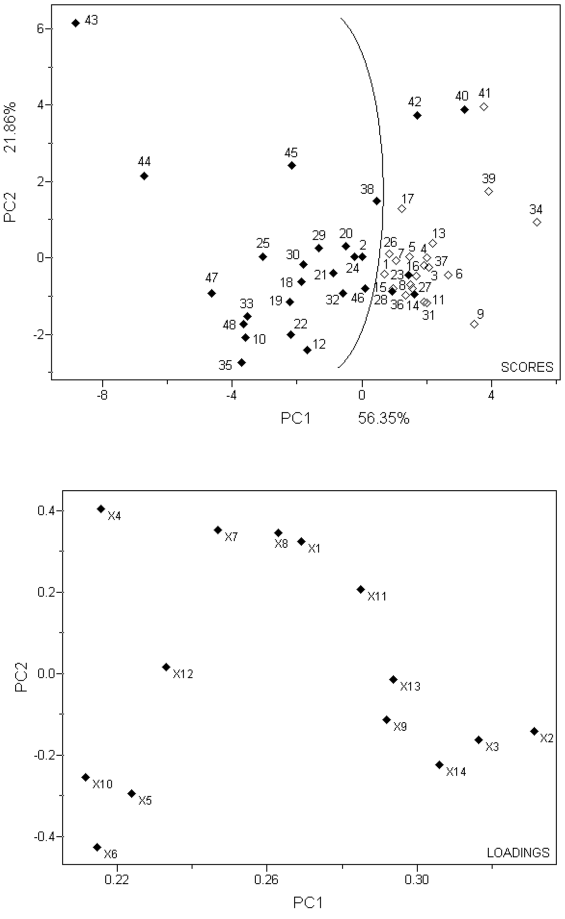


Figure 2. PC1 against PC2 scores and loadings plots from Principal Component Analysis.

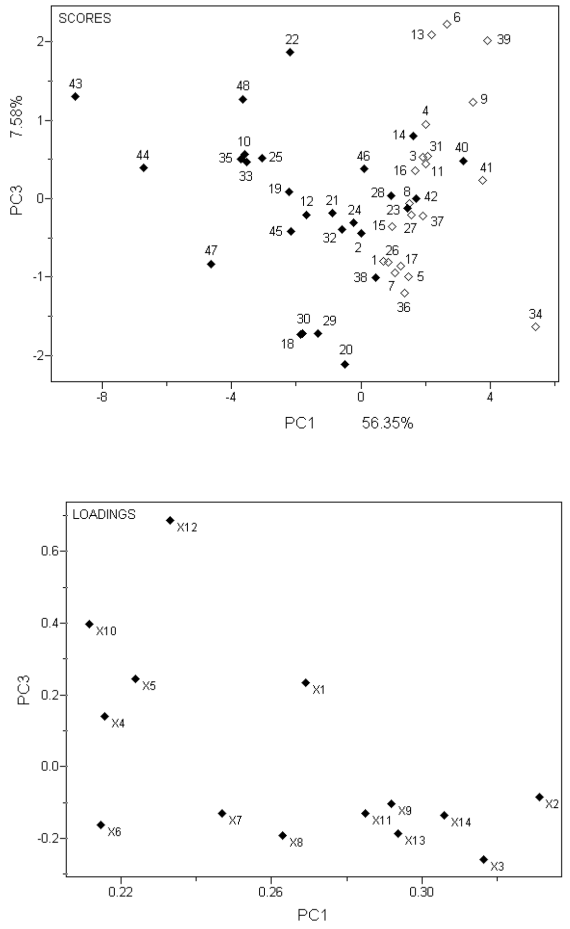


Figure 3. PC1 against PC3 scores and loadings plots from Principal Component Analysis.

As already mentioned, the goal of any QSAR study is to build a correlation between physicochemical and structural properties and biological activities. The biological activities can be represented in two ways: either they are continuously valued or discrete category assignments. In case they are continuous, it is possible to develop regression models using PCR or PLS. However, in other cases, the biological activity might be given only as a categorical property (highly-active/moderately-active/ non-active), in which case, pattern recognition methods such as KNN and SIMCA are the appropriate approach. The final goal in this case, is to predict the most likely category of new compounds.

2.4 Principal Component Regression PCR[3,4,7]

Principal component regression uses steps similar to those in PCA to decompose the data matrix into principal components, which describe the latent structure of **X**, followed by a regression step. Based on the singular values decomposition technique introduced in PCA analysis, one can write the PCR models as:

$$\begin{bmatrix}\mathbf{y}\end{bmatrix}=\begin{bmatrix}\quad\mathbf{X}\quad\end{bmatrix}\begin{bmatrix}\mathbf{b}\end{bmatrix}+\mathbf{e}_y \Rightarrow \begin{bmatrix}\mathbf{y}\end{bmatrix}=\begin{bmatrix}\quad\mathbf{U}\quad\end{bmatrix}\begin{bmatrix}\quad\mathbf{S}\quad\end{bmatrix}\begin{bmatrix}\quad\mathbf{V}^{\mathrm{T}}\quad\end{bmatrix}\begin{bmatrix}\mathbf{b}\end{bmatrix}+\mathbf{e}_y$$

$$\begin{bmatrix}\hat{\mathbf{b}}\end{bmatrix}=\begin{bmatrix}\mathbf{V}\quad\mathbf{S}^{-1}\quad\mathbf{U}^{\mathrm{T}}\end{bmatrix}^{A}\begin{bmatrix}\mathbf{y}\end{bmatrix} \tag{7}$$

where the superscript A is the number of principal components included in the model ($\mathbf{V}(A\times A)$, $\mathbf{S}^{-1}(A\times A)$ and $\mathbf{U}^{\mathrm{T}}(A\times I)$). Since the principal components are orthogonal to each other, they are suitable as regressors for **y** using MLR. In this case, the least squares solution for the estimation of the regression coefficients **a** when the original data matrix, **X**, is replaced with a subset of A columns of the score matrix **T** is:

$$\mathbf{y} = \mathbf{Ta} + \mathbf{e} = a_o + a_1\mathbf{t}_1 + \ldots + a_A\mathbf{t}_A + \mathbf{e} \Rightarrow \hat{\mathbf{a}} = (\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1}\,\mathbf{T}^{\mathrm{T}}\mathbf{y} \tag{8}$$

Interpretation of the PCR model would be greatly simplified if **a** might be given in terms of regression coefficients **b**. From Equations 8 and 3, follows:

$$\mathbf{X}\hat{\mathbf{b}} = \mathbf{T}\hat{\mathbf{a}} \tag{9}$$

Also, **T** = **XL** (**X** = $\mathbf{TL}^{\mathrm{T}}$ from Equation 4 and **L** is an orthogonal matrix). Replacing **T** with **XL** in Equation 9, it is clear that one possible solution for $\hat{\mathbf{b}}$ is $\mathbf{L}\hat{\mathbf{a}}$. Since $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ can be close to singular, the solution for $\hat{\mathbf{b}}$ is hardly unique. When all the principal components are kept in the model (no data compression; $A = J$, for $J < I$), the PCR model would match MLR's.

2.5 Partial Least Squares, PLS[7,8]

In PLS regression, the **X** block of independent variables (given by the descriptors) is correlated with the **y** vector (given by the activities) in such a way that the projected coordinates, **T**, are good predictors of **y**. In other words, the biological activities are included in the decomposition procedure. In PCR regression, the loading matrix, **L**, is determined in such way as to maximize the variance with **X**, not to predict **y**. In PLS, on the contrary, the loading matrix **W**, made by the "weight loading vectors" are defined in such a way that the product (variance in **X**) times (the correlation **XW** to **y**) is maximized.

Below is a PLS algorithm from Martens *et al.*[7]

```
PREPROCESS X and y
ESTIMATE the first “weight loading vector” w.
LOOP
    BLOCK X
        $\mathbf{X} = \mathbf{y}\,\mathbf{w}^{\mathrm{T}} + \mathbf{E}_{\mathrm{X}}$    $(\mathbf{X}^{\mathrm{T}} = \mathbf{w}\mathbf{y}^{\mathrm{T}})$
        $\hat{\mathbf{w}} = \mathbf{X}^{\mathrm{T}}(\mathbf{y}^{\mathrm{T}}\mathbf{y})^{-1}\mathbf{y}^{\mathrm{T}}$
        $\hat{\mathbf{w}} = \hat{\mathbf{w}}^{\mathrm{T}}/\sqrt{\hat{\mathbf{w}}^{\mathrm{T}}\hat{\mathbf{w}}}$    (normalization)
        $\mathbf{t} = \mathbf{X}\hat{\mathbf{w}}$    (scores)
        $\mathbf{l} = \mathbf{X}^{\mathrm{T}}\mathbf{t}(\mathbf{t}^{\mathrm{T}}\mathbf{t})^{-1}$    (loadings)
    Relate the vector t with the biological activity y
    $\mathrm{q} = (\mathbf{t}^{\mathrm{T}}\,\mathbf{t})^{-1}\mathbf{t}^{\mathrm{T}}\mathbf{y}$
    Update X and y
        $\mathbf{E} = \mathbf{X} - \mathbf{t}\,\mathbf{l}^{\mathrm{T}}$
        $\mathbf{e} = \mathbf{y} - \mathrm{q}\,\mathbf{t}$
        $\mathbf{X} = \mathbf{E}$
        $\mathbf{y} = \mathbf{e}$
    Repeat for the next “weight loading vector”.
CALCULATE the regression vector
    $\hat{\mathbf{b}} = \mathbf{W}\left(\mathbf{L}^{\mathrm{T}}\,\mathbf{W}\right)^{-1}\mathbf{q}^{\mathrm{T}}$
```

The regression model is validated usually through an internal leave-one-out cross-validation procedure in order to verify its ability for future predictions. The optimal number of PLS-components, *A*, to be included in the model, is chosen as the number of latent variables that minimize the *PRESS* (Predicted Residual Error Sum of Squares; see equation 10). The Standard Error of Prediction (*SEP*; equation 10) can also be used to assess the predictability of the model

$$SEP = \left[\frac{\sum\left(y_{iexp} - y_{ipred}\right)^2}{n}\right]^{1/2} = \left[\frac{PRESS}{n}\right]^{1/2} \qquad (10)$$

where n is the number of compounds used in the cross validation procedure, y_{iexp} is the known value of activity for the i^{th} sample and y_{ipred} is the activity predicted by the model built without sample i. The parameters usually used to assess the statistical quality of the model are the correlation coefficient R^2 and Q^2, the crossvalidated correlation coefficient (both of them close to unity)

$$R^2 = 1 - \frac{\sum\left(y_{ical} - y_{iexp}\right)^2}{\sum\left(y_{iexp} - y_{mean}\right)^2},\ Q^2 = 1 - \frac{\sum\left(y_{iexp} - y_{ipred}\right)^2}{\sum\left(y_{iexp} - y_{mean}\right)^2} \qquad (11)$$

where y_{cal} is the value estimated by this model, y_{mean} is the mean experimental activity and y_{pred} is the crossvalidated predicted activity. It shouldn’t be forgotten that by definition, a model possesses predictability only when the prediction errors are less than the variation of the response variables.

However, it should be pointed out that besides good prediction, interpretation and understanding of the structural features that affect the biological activity is an important issue in QSAR, especially from the chemical point of view. It is not an easy task to interpret models which include a big set of descriptors. In order to reduce the number of features considered during the model building, the descriptors must be analyzed and those containing little useful information (correlation coefficient smaller than a given cutoff) should be removed. When doing a QSAR study, one should always keep in mind to consider descriptors that could give information about the influence of electronic, steric, hydrophobic and hydrogen bonding features on the biological activity of drug molecules. There are several algorithms for variable selection available in the current literature.

2.5.1 Example: Bioconcentration Factor (BCF) of Polycyclic Aromatic Hidrocarbons (PAHs)[15]

A quantitative relationship between structural features of PAHs and their bioconcentration in aquatic organisms is built as an illustration of the regression method PLS. The bioconcentration factor is derived from measured concentration of a chemical present in the aquatic organism and the ambient environment as

$$\mathrm{BCF} = \frac{\mathrm{C_{org}}}{\mathrm{C_w}}$$

where $\mathrm{C_{org}}$ = concentration in target organism (mg/kg) and $\mathrm{C_w}$ = concentration in (pure) water (mg/l) or calculated from rate constants.

The data set was taken from reference 15. Four descriptors were used to model the BCF: Experimental log Kow and log S (*n*-octanol/water partition coefficient and water solubility, respectively); calculated molecular volume Vol and the calculated vertex connectivity index Xv. The BCFs for some PAHs in the crustacean *Daphnia-Magna* are listed in Table 1. The correlation between BCF and the descriptors can be visualized in Figure 4. Note that BCF in this aquatic invertebrate shows a high correlation with log Kow, and log S. With an increase in the hidrophobicity (inversely related to water solubility), it is expected a higher mobility of PAHs to the target organism. BCF also increases with the PAH size.

A good PLS model was obtained using autoscaled bioconcentration data and the four descriptors. One latent variable describing 95.09% of the original information (Table 2) is the optimum number of factors to build the model because it gives the lowest standard error of validation (0.295). The correlation coefficient R^2 and

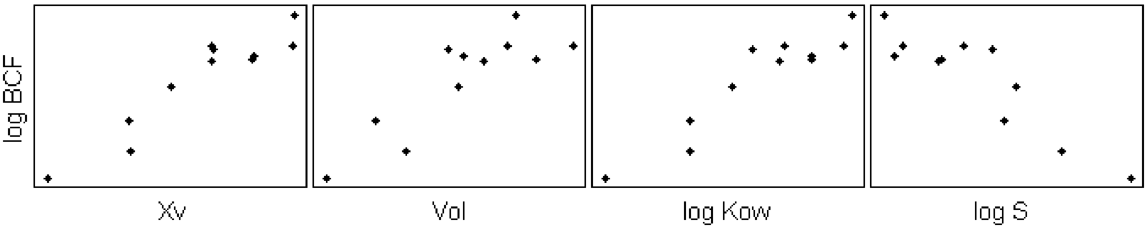


Figure 4. Scatter plots of Bioconcentration Factor (log BCF) versus descriptors.

Table 1. Experimental and estimated[a] values of bioconcentration factor of PAHs in aquatic invertebrate *Daphnia-Magna*

	Exp.[b] log BCF	Est. log BCF	Res.
Naphthalene	2.12	2.167	-0.0468
Anthracene	2.95	2.948	0.00201
Phenanthrene	2.51	2.887	-0.377
Benz[a]anthracene	4.00	3.744	0.256
Chrysene	3.785	3.738	0.0469
Triphenylene	3.96	3.478	0.482
Pyrene	3.43	3.295	0.135
Perylene	3.86	3.965	-0.105
Benzo[a]pyrene	3.82	4.019	-0.199
Dibenz[a,h]anthracene	4.00	4.355	-0.355
Benzo[ghi]perylene	4.45	4.288	0.162
SEP		0.295	
R^2 0.87	Q^2	0.81	
BCF= 0.694 + 0.0959 Xv + 0.0022 V + 0.163 log Kow - 0.109 log S			

[a]PLS model using autoscaled data, 1 latent variable and leaving-one-out crossvalidation; [b]From *Daphnia-Pulex*. From reference 15.

Table 2. Percent and cumulative percent variance from PCA analyis and SEP from leave-one-out cross validation

LV	Percent	Cumulative	SEP[a]
1	95.09	95.09	0.295
2	3.50	98.59	0.323
3	1.30	99.89	0.406
4	0.11	100.00	0.478

[a]Standard error of prediction for crossvalidation. See equation 10.

crossvalidation correlation coefficient Q^2 were respectively 0.87 and 0.81. The regression equation, estimated values and residuals of BCF are in Table 1 and the plot of experimental versus estimated BCF it is in Figure 5. The residuals obtained for two compounds, phenanthrene and triphenylene, are greater than 10%.

2.6 Pattern Recognition Methods KNN and SIMCA [9,10]

Both these methods are similarity-based classification techniques. Classification with KNN is based on distance comparison among samples. In a first step, the multivariate Euclidean distance (equation 6) between every pair of compounds in the training set (set of samples with known classes) is calculated. Other distance measures can also be used. The predicted class of a test compound is determined based on the multivariate distance of this compound with respect to the closest *K* compounds in the training set. Leave-one-out crossvalidation is used to select the optimal number of nearest neighbors, *K*, to pool for future classification. In this procedure, each sample in the training set is excluded and then classified using the remaining training set. This is repeated for different values of *K*. Each of the *K* nearest samples "votes" once for its class. The class receiving the highest number of votes is assigned to that sample.

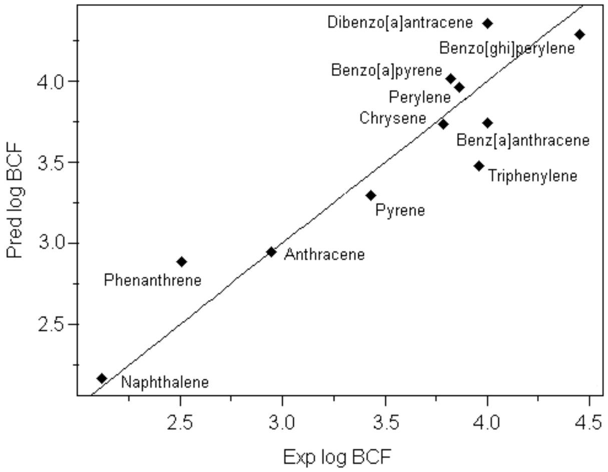


Figure 5. Predicted versus measured log bioconcentration factor.

In contrast with KNN, which is based only on physical distances among samples, SIMCA method builds principal component models, one for each class in the training set. The number of optimal PCs is determined for each class and the model is completed by defining boundary regions for each PCA model. In other words, a hyperbox is constructed for each class. The principal components define the orientation of the box and the range of the scores defines the boundaries inside which the training set samples reside.

Once the class boundaries are defined, it is important to determine whether any of the class overlaps. The “distance between classes” and the “residuals between classes” accomplish the discriminating power of the SIMCA models to distinguish between classes. The classification of a test sample to a given class is based in its projection into the scores space and its distance from the boundaries of that class. This is repeated for all the SIMCA classes and at the end, the test sample may be a member of one or more than one (when SIMCA models do not have enough discrimination power) and even none of the classes (when the sample is unusual with respect to the training set).

2.6.1 Example: Lapachol and derivatives of 1,4-naphthoquinones against carsinosarcoma W 256[16]

In this example, lapachol (extracted from the heartwood of certain Asian and South American bignoniaceous) and several derivatives of 1,4-naphthoquinone are used to investigate possible relationships between the structural parameters and activity against carcinosarcoma Walker 256 (W 256) using KNN and SIMCA. The data set (25 compounds) taken from Reference 16, classifies the compounds as active and inactive. The descriptors are HOMO (the highest occupied molecular orbital) coefficients for carbon atoms *b*, *c*, *m*, *n*, *o*, *p*, *q*, *s*, *t* and *u* of the side chain double bond (see Ref 16 for the atomic center labels). The descriptors were obtained from semiempirical AM1 molecular orbital calculations. PC1 discriminates between active and inactive compounds. From scores plot (Figure 6) and loadings (Table 3), it can be concluded that active compounds have a high contribution from descriptors *p-u* (negative loadings) which means high electron density in the side-chain double bond and terminal groups. Inactive compounds have high contribution from *b-n* atoms. HCA results in Figure 6 confirm PC analysis, providing a diagnostic of modeling strength.

Table 3. PC1 and PC2 loadings

Variables[a]	PC1	PC2
b	0.365	0.212
c	0.382	0.0335
m	0.370	0.220
n	0.372	0.225
o	0.0888	-0.602
p	-0.294	0.362
q	-0.357	-0.270
s	-0.358	0.102
t	-0.184	0.424
u	-0.249	0.312

[a]See Reference 16 for variable’s definition.

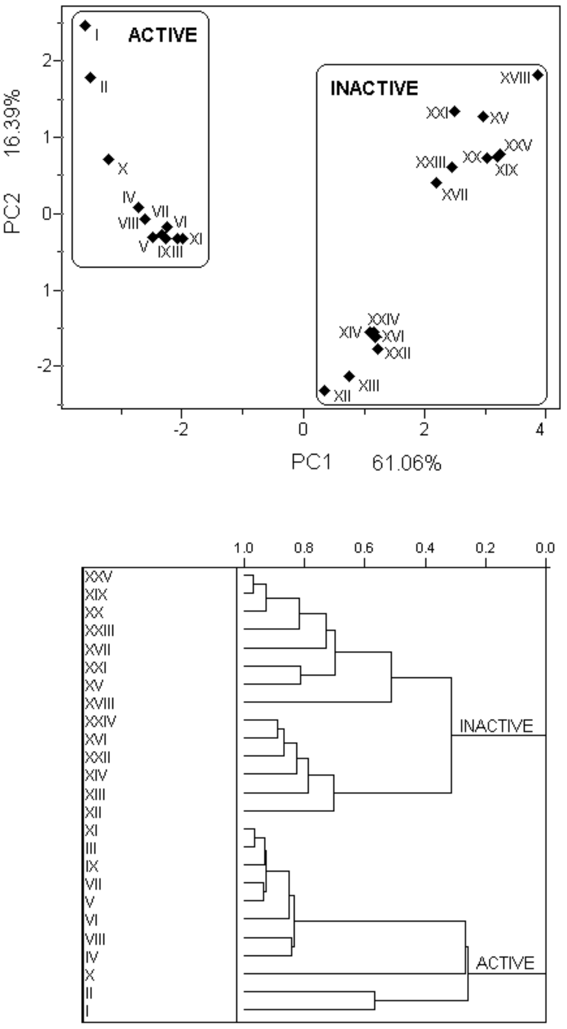


Figure 6. PC1 against PC2 scores and dendogram for 25 naphthoquinone derivatives.

Once the data structure is known, a classification model can be built using the training set (set of 25 compounds). KNN model was built on autoscaled data using three nearest neighbors, $K = 3$, with zero misses. Figure 7 shows the PC1xPC2xPC3 scores plot for the training set with the classes’ boundaries for SIMCA modeling. Both classes are well modeled by 2PCs on locally autoscaled data. SIMCA models have a good discriminating power, *i. e.*, high interclass distances and high between classes’ residuals. The interclass residuals are shown in Table 5. As expected, they are significantly smaller when the class is fit to itself than when fit to the model of the other class.

Once the number of PCs and the boundaries are defined for each class, the model can be used for future prediction. A set of 16 samples is used as test. The results for both models are in Table 5. For KNN prediction, the compounds

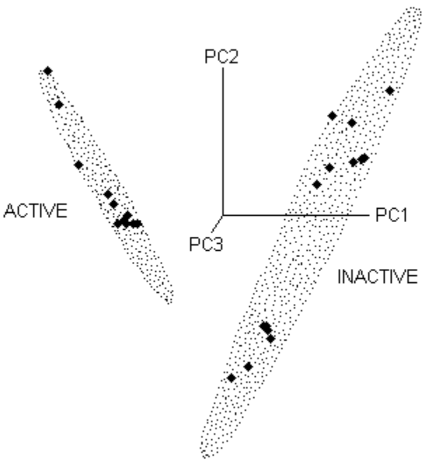


Figure 7. 3D Scores plot showing boundaries classes for 25 naphthoquinone derivatives.

Table 4. Interclass Residuals

	Class1	Class2
Class1	0.5702	7.1075
Class2	13.911	0.5618

were definitly assigned to one of the two classes. Contrary to KNN, in SIMCA prediction 10 samples are assigned to either one of the classes, and six of them did not fit any of the two classes (predicted class number is set equal to zero).

In Figure 8 is shown the plot of PC1 against PC2 for Class1, where one can see the boundaries for this class and those eight compounds predicted as active projected inside. Only compound XXVIII which was expected to be active, is outside the limits of the hyperbox. The other compounds are far distant and don't belong to this class. Figure 9 is a biplot of the class distances. Samples in the NW quadrant belong to Class1 (active). Those in the SE quadrant belong to Class2 (inactive). Samples on the SW quadrant can be a member of either Class1 or Class2, and those in the NE don't belong to any of the classes. Compound XXXI not assigned to any class is in the border of the non-active compounds.

Table 5. KNN and SIMCA predicted class of activity[a]

Compound	Class	Number
XXVII	1	1
XXVIII	1	0
XXIX	1	1
XXX	1	1
XXXI	2	0
XXXII	2	2
XXXIII	2	2
XXXIV	2	0
XXXV	2	0
XXXVI	2	0
XXXVII	2	0
XXXVIII	1	1
XXXIX	1	1
XL	1	1
XLI	1	1
XLII	1	1

[a]Samples which did not fit in either the two classes are assigned to class 0.

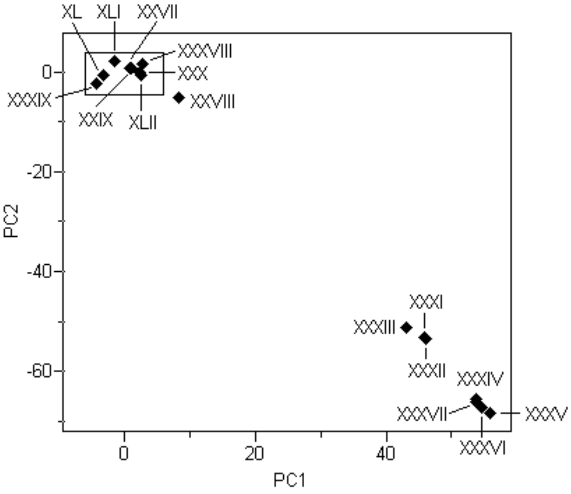


Figure 8. PC1 against PC2 scores plot for test set showing the active compound of Class1 inside its boundaries.

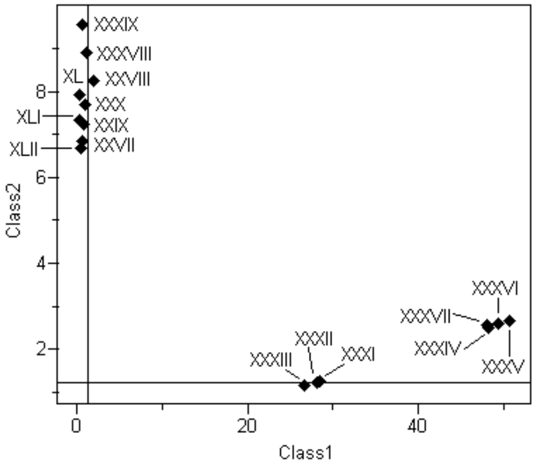


Figure 9. Class distance plot for active (Class1) and inactive (Class2) compounds in the test set.

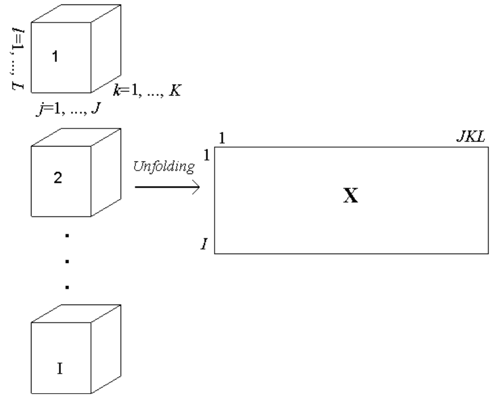
2.7 3D QSAR [11,12]

Since Cramer *et al.* introduced the 3D Comparative Molecular Field Analysis (CoMFA) in 1988, it has become a key stone in 3D QSAR. Here, PLS is the regression method used to model the relationship between the biological activity of a set of compounds with a specified alignment and their 3D interaction energy fields (electronic, steric, hydrophobic and H bonding). These fields are determined on a 3D grid that is laid over the chemical structures. Another similar 3D QSAR methodology available is the GRID/GOLPE[17] combination, where the GRID force field[18] is combined with the variable selection method GOLPE.[19]

To generate a CoMFA model[20,21] on a diverse set of compounds, the first two steps: *i*) conformational analysis to find the low energy or the global minimum conformations; and *ii*) alignment of all molecules in the same coordinate system; are time consuming and frequently the determining step. Once all molecules are properly aligned, one can *iii*) generate the molecular fields that would describe the intermolecular interactions with the target biomolecule. Then, the last step is *iv*) the regression analysis with biological activities using bilinear PLS.

Molecular fields are 3D representations of the steric, electrostatic, hydrogen bonding and hydrophobic surroundings of the molecule. They are generated by enclosing the molecule into a 3D grid and calculating the interaction energies between a probe (an atom, an ion or/and a molecule) and the target molecule at each grid point. These energies can be used as descriptors of the 3D molecular structure.

In order to model such data sets using bilinear PLS algorithm, it is necessary to unfold and arrange all generated data in a bidimensional matrix as shown in the scheme 1 below.



Scheme 1. The unfolding of molecular fields for *I* compounds originating a bidimensional matrix **X**(I×JKL).

In this scheme, a four-way array $\underline{\mathbf{X}}(I{\times}J{\times}K{\times}L)$ is unfolded into a two-way matrix $\mathbf{X}(I{\times}JKL)$, where each cube corresponds to the molecular field generated by a given probe for compound i (i = 1, ..., I). The other three modes referring to j (j=1, ..., J), k (k=1, ..., K) and l (l=1, ..., L) correspond to the x, y and z coordinates for the grid points.

To make the results of a PLS model in 3D QSAR easy to interpret, the regression coefficients are presented as iso-contour plots.

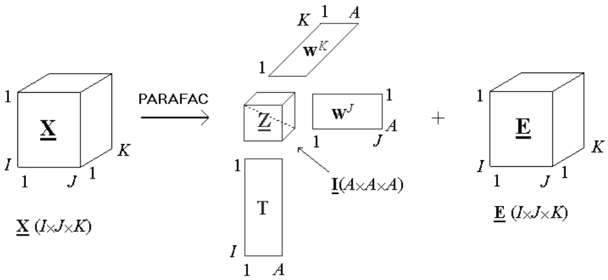
3. Perspectives and future development

It has been proposed since last decade a multilinear PLS algorithm,[22,23] which is an extension of the traditional bilinear PLS. It has the advantage that the multiway (or higher order) structure of the data is maintained. This method is suitable to 3D-QSAR problems but only a few applications have been made so far.[24-26]

3.1 Multilinear PLS[22,23]

In this methodology, the unfolding step is omitted and the PARAFAC method is used to build the model. Multilinear PLS can be superior to the unfolding PLS method since due to its simplicity (the number of variables can be effectively reduced), the predictive ability is in general slightly better for fewer parameters need to be estimated.

For a three-way data $\underline{\mathbf{X}}(I, J, K)$, the Tucker and PARAFAC models decompose the data into three matrices $\mathbf{T}$, $\mathbf{W}^J$ and $\mathbf{W}^K$ and a core matrix $\underline{\mathbf{Z}}$. PARAFAC is a special case of the Tucker model where the core matrix $\underline{\mathbf{Z}}=\underline{\mathbf{I}}$, *i. e*., only the superdiagonal elements are different from zero and equal to unity (see representative Scheme 2 below). The loading matrices are chosen in such a way that the sum of squared residuals is minimized.



Scheme 2. Schematic representation of a multiway array $\underline{\mathbf{X}}(I{\times}J{\times}K)$ into three loadings matrices and a core array $\underline{\mathbf{Z}} = \underline{\mathbf{I}}(A{\times}A{\times}A)$.

In Scheme 2, A is the number of factors in the model and $\underline{\mathbf{E}}(I{\times}J{\times}K)$ is the residual array. The miltilinear PLS regression model is a combination of bilinear PLS with the PARAFAC decomposition.

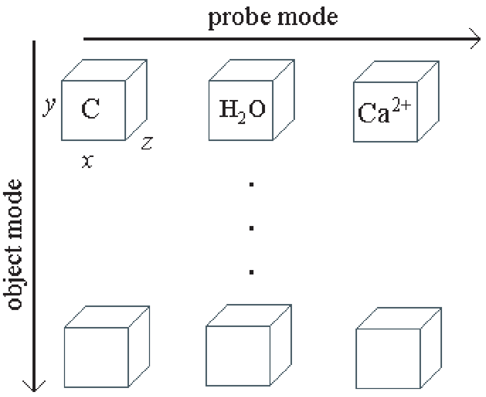
As an example,[21] one can consider a set of 30 compounds. They are aligned, surrounded by a 3D grid large enough to enclose all the aligned molecules with a border in all directions. The directions (*x, y, z*) in the grid are divided into 30, 15 and 20 steps, yielding 9000 grid points. The interactions between the probe and each molecule are calculated at each grid point, resulting in the molecular field.

Suppose that in this example, three different probes are used yielding different types of fields:

The carbon atom is one probe reflecting a steric field. The cation Ca^{2+}, another probe, reflects an electrostatic field and the water molecule, the third probe, describes a hydrogen bonding field.

As shown below, this complete data set defines five modes represented by $\underline{\mathbf{X}}$(*objects*×*x*×*y*×*z*×*probe*) according to Scheme 3:

1 Object mode of 30 dimensions
2 *x*-mode of 30 dimensions
3 *y*-mode of 15 dimensions
4 *z*-mode of 20 dimensions
5 the probe mode of 3 dimensions



Scheme 3. Complete data set with its five modes.

Unfolding $\underline{\mathbf{X}}$ for bilinear PLS analysis, the data set form a matrix (30×27000) which is decomposed into scores $\mathbf{T}(30\times A)$ and loadings $\mathbf{L}(27000\times A)$. In multilinear PLS, this unfolding step is omitted and the decomposition gives the scores $\mathbf{T}(30\times A)$ and four loading matrices, $\mathbf{W}^J(30\times A)$, $\mathbf{W}^K(15\times A)$, $\mathbf{W}^L(20\times A)$, $\mathbf{W}^M(3\times A)$ corresponding to the objects mode, *x*, *y* and *z*-direction grid and the probe mode respectively.

It is visible from the example above, that the number of parameters to be estimated when using PARAFAC, is far less than when doing the unfolding.

Concluding, the objective of this work is to give a brief overview on Multivariate QSAR and above all, to call the attention to the application of multilinear PLS to medicinal chemists which can be superior to traditional PLS and bring some improvement with respect to speed of calculation, simplicity and stability.

Acknowledgments

The author would like to thank financial support from the Brazilian agency FAPESP.

References

1. Hansch, C.; Fujita, T.; *J. Am. Chem. Soc*. **1964**, *86*, 1616.
2. Free, S. M.; Wilson, J. W.; *J. Med. Chem.* **1964**, *7*, 395.
3. Beebe, K. R.; Pell, R. J.; Seasholtz; M. B.; *Chemometrics: A Pratical Guide*; Wiley: New York, 1998.
4. Ferreira, M. M. C.; Antunes, A. M.; Melgo, M. S.; Volpe, P. L. O.; *Quim. Nova* **1999**, *22*, 724.
5. Malinowski, E. R.; *Factor Analysis in Chemistry*, 2nd ed., John Wiley & Sons: New York, 1991.
6. Sharaf, M. A.; Illman, D. L.; Kowalski, B. R.; *Chemometrics*, Wiley-Interscience: New York, 1986.
7. Martens, H.; Naes, T.; *Multivariate Calibration*, Wiley: New York, 1989.
8. Geladi, P.; Kowalski, B. R.; *Anal. Chim. Acta* **1986**, *185*, 1.
9. Kowalski, B. R.; Bender, C. F.; *J. Am. Chem. Soc.* **1972**, *94*, 5632.
10. Wold, S.; Sjostrom, M.; In *Chemometrics: Theory and Application*, Kowalski B. R. ed., ACS Symposium Series No 52, American Chemical Socisty: Washington, DC, 1977, ch. 12.
11. Cramer III, R. D.; Patterson, D. E.; Bunce, J. D.; *J. Am. Chem. Soc*. **1988**, *110*, 5959.
12. Cramer III, R. D.; de Priest, S. A.; Patterson, D. E.; Hecht, P. In *3D QSAR in Drug Design*; Kubinyi H. ed.; Leiden, 1993, p 443-485.
13. Hammett, L. P.; *Physical organic Chemistry*, McGraw-Hill: New York, 1940.
14. Kiralj, R.; Ferreira, M. M. C.; *J. Mol. Graph. Model*. in press.
15. Ferreira, M. M. C.; *Chemosphere* **2001**, *44*, 125.
16. Subramanian, S.; Ferreira, M. M. C.; Trsic, M.; *Struct. Chem.***1998**, *9*, 47.
17. Cruciani, G.; Watson, K. A.; *J. Med. Chem*. **1994**, *37*, 2589.
18. Goodforf, P.; *J. Med. Chem*. **1985**, *28*, 849.
19. Baroni, M; Costantino, G.; Cruciani, G.; Riganelli, D.; Valigi, R.; Clementi, S.; *Quant. Struct.-Act. Relat*. **1993**, *12*, 9.

20. Kubinyi, H.; *3D QSAR in Drug Design. Theory, Methods and Applications*, ESCOM: Leiden, 1993.
21. Martin, C. Y.; Kim, K-H; Lin, C. In *Advances in Quantitative Structure-Property Relationships*; Charton M. ed., JAI Press Inc: London, England, 1996 Volume 1, 1.
22. Bro, R; *J. Chemom*. **1996**, *10*, 47
23. Andersson, C. A.; Bro R.; *Chemom. Intell. Lab. Syst*. **2000**, *52*, 1
24. Nilsson, J.; Jong, S. D.; Smilde, A. K.; *J. Chemom*. **1997**, *11*, 511.
25. Nilsson, J.; Homan, E. J.; Smilde, A. K.; Grol, G. J.; Wikstrom, H.; *J. Comut.-Aided Mol. Des*. **1998**, *12*, 81.
26. Hasegawa, K.; Arakawa, M.; Funatsu, K.; *Chemom. Intell. Lab. Syst*. **1999**, *47*, 33.

Received: June 6, 2002
Published on the web: November 19, 2002

FAPESP helped in meeting the publication costs of this article.