Chemical pattern recognition and multivariate analysis for QSAR studies

Yoshikatsu Miyashita*, Zhiliang Li and Shin-ichi Sasaki
Toyohashi, Japan

Chemical pattern recognition (CPR) and quantitative structure–activity relationships (QSAR) studies based on multivariate analysis and chemometric techniques are reviewed. In particular, applications of the SIMCA classification method to structure–taste problems are discussed. Cluster significance analysis (CSA) is compared with modelling powers for feature selection of asymmetric data sets. A concentric hypersphere model is used to predict candidate new sweeteners. Partial least squares (PLS) modelling methods are employed to antiarrhythmic data of phenylpyridines and fungicidal and herbicidal data of thiocarbamates, respectively. The CoMFA approach to 3-dimensional QSAR using PLS modelling is described as well. In practice, QSAR is an important branch of chemometrics and enhances rational drug design and new agent development. The chemometric techniques described in the article not only work well for QSAR but also are very helpful for solving the problems related to analytical characteristics–chemical structure relationships.

Introduction

Discovering and establishing relationships between the chemical structures of molecules and their activities or properties are always interesting research topics for chemists. The quantitative description of the relations is the so-called quantitative structure–activity/property relationships (QSAR and QSPR) [1–7].

QSAR studies express the biological activity of compounds as a function of their various structural parameters and describe how variation of the biological activity depends on change of the chemical structure. It can be traced back to the last century when chemists realized that some properties of compounds, such as physiological action, are related to their chemical structures and the relationships between them can be described by mathematical tools. Hammett proposed the linear free energy relationships (LFER), which can be regarded as a starting point of QSAR. But indeed, the first and most important work dealing with QSAR was published, based on LFER, by Corwin Hansch and his co-workers in the 1960s [1,2]. In this pioneering framework, the Hansch equation was developed for quantitative approaches to describe relationships between chemical structure and biological activity as dependent variables y based on the establishment of an empirical model of drug action using LFER related parameters as independent variables x:

$$y = a\pi + b\pi^2 + c\sigma + dE_s + e \quad (1)$$

where y is a biological activity, $y = \log(1/C)$, in which C refers to ED_{50}, LD_{50}, I_{50}, MIC, etc. The parameters π, σ and E_s which are widely used substituent constants, are assumed to represent the hydrophobic, electronic and steric factors, respectively. The QSAR models are developed by using multiple linear regression analysis (MLR) which is a popular classical modelling method [1,2].

The Hansch approach is a powerful technique for optimizing the activity of a lead compound. With this method a basic assumption is that all the factors involved in variation in biological activity arising from the modification of the molecular structure with a congeneric series can be correlated with concomitant change of physicochemical parameters. The great advantage of the MLR method is that a causal model is obtained and the physical meaning is obvious. However, the following conditions must be satisfied to apply MLR:

- the descriptor variables/parameters are orthogonal and
- the number of compounds (objects, samples) is greater than that of descriptors (variables, parameters).

So in classical QSAR studies based on MLR analysis, initially the number of samples is required to be at least five times the number of descriptors and preferably

*To whom correspondence should be addressed.

0165-9936/93/$06.00

more than 10 times. Otherwise, either condition is frequently not satisfied. Sometimes overfitting results may be obtained and the predictive power of the model is very poor.

Chemical pattern recognition (CPR) [8,9] is regarded as another method of mathematically modelling QSAR and QSPR. QSAR, together with CPR, is an important branch of chemometrics and has been proven to play an important role in drug design [5–7]. Although the QSAR studies began before the field of chemometrics developed, chemometrics extended QSAR studies and increased the research level of information which can be obtained from QSAR studies.

In QSAR and CPR studies, a fundamental problem is how to quantitatively and accurately predict the chemical and related properties of a compound on the basis of its chemical composition and structure. In scientific research, one certainly hopes to establish a global hard model. In quantum chemistry and molecular mechanics, finding a global hard model of relationships between chemical structure and property is desired. But for some complex molecules, it is difficult to perform quantum mechanical calculations [6]. For these reasons, much quantum mechanical research is semi-empirical. Where a global hard model cannot be obtained, chemists frequently utilize other local soft models, such as analogy and similarity methods. For instance, an empirical rule "like dissolves like" for solubility and the well-known periodic table for chemical elements are classical analogy methods and are commonly used in chemistry. Chemical phenomena are more complex than physical ones and are affected by many unknown factors. So, the real chemical world is typically multivariate. Facing this multivariate chemical world, we must make many assumptions and/or hypotheses in order to obtain a global hard model and then the model loses its strictness or the advantage of rigour. The Hansch approach using MLR is regarded as a hard model.

In QSAR and CPR, local soft modelling becomes a powerful approach because soft models can be used to predict the related property and activity. The SIMCA and softer partial least squares (PLS) modelling methods, recently developed in chemometrics, can lead to local soft model solutions to chemical problems [10]. In this article the applications of SIMCA and PLS to QSAR will be discussed.

Data structure: symmetric and asymmetric

In QSAR and CPR studies, substances of known biological activity constitute the training or learning sets while substances of unknown activity are the test or prediction sets. The QSAR data mainly consist of two basic parts: the biological activity(ies) and chemical descriptors.

The biological activity data are measured and describe the pharmacological profile of the training set. They may be either continuous, *e.g.* $y = \log(1/C)$, or discrete values, *e.g.* active–inactive, agonist–antagonist, sweet–bitter, weak–moderate–strong activity, etc. The chemical descriptor data may be both experimentally measured and/or theoretically calculated, and may be both physicochemical parameters and computational parameters or graph invariants. If graph invariants are used the model is only one-way predictive.

Understanding the data structure of descriptor space is quite important for QSAR analysis, because the relative position of an object in multidimensional space becomes clearer. Usually the Karhanen-Loeve plots (principal component analysis) or non–linear mapping plots are employed to display the multidimensional descriptor data space.

In general, the data structure of QSAR may be complex. In exceptional cases the data structure is symmetric as shown in Fig.1a, *i.e.*, the samples in active and inactive classes are linearly separable in descriptor space. In this case linear discriminant analysis (LDA) can be used. The data structure is as a rule asymmetric (embedded) as shown in Fig.1b; in other words, the two class samples in descriptor space are not linearly separable. The symmetric and asymmetric cases were defined by Wold and Dunn [11].

The asymmetric data structure shows more restricted structure requirements for activities. As an example, a plot of the Hammett $\Sigma\sigma$ against the Hansch $\Sigma\pi$ for carcinogenic dimethylaminoazobenzenes using the original data in ref. 1 is shown in Fig. 2. Interestingly, the carcinogenic (active) class is also clustered in the data space but the non-carcinogenic (inactive) group is scattered "randomly" in the parameter ($\Sigma\pi$–$\Sigma\sigma$) space.

This data structure may also frequently be encountered in QSAR studies. For this embedded case, LDA may fail to obtain a model while SIMCA gives a successful model for such a classification problem.

The SIMCA classification method

Many varied methods have been developed for classification. For example, the LDA method has several limits or disadvantages and does not suit some cases as follows:

- asymmetric data structure,

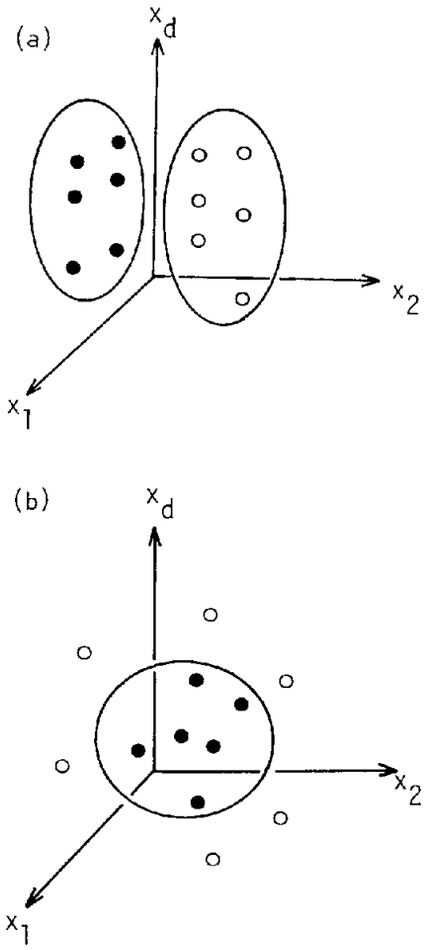


Fig. 1. (a) A symmetric data structure: the classes can be distinguished linearly in the pattern space. (b) An asymmetric data structure: the classes cannot be separated linearly in pattern space or the active class to form a cluster is imbedded in the inactive class to spread in various directions. • = active, o = inactive.

- the number of compounds is less than that of descriptors,
- overfitting problem,
- sensitive to groupings of objects.

Those cases are often encountered in practical drug design.

The SIMCA class modelling technique [12–14] based on disjoint principal component analysis (PCA) is one of the most useful methods in chemical pattern recognition, which can be given by several names. Among these, soft independent modelling of class analogy most accurately reflect the characteristics of the SIMCA pattern recognition method. All fundamental features and basic aspects of the SIMCA method are thoroughly documented together with numerous chemical examples [12–17]. SIMCA, as a biased method, uses the orthogonal scores (t) and loadings (p) to model the descriptor data for each class and can remove the defects of unbiased methods such as the LDA approach. The LDA approach, like MLR, can work only when the number of samples substantially exceeds the number of variables while SIMCA is not subject to the restriction of few variables and works well in these cases. For instance, SIMCA works

- with as few as five objects per class and any number of variables for classification and
- when variables ($d = 105$) exceed samples ($n = 16$) to classify gas chromatographic profiles of human brain tumor tissues [14].

The basic idea of the SIMCA method is that through disjoint PCA of descriptor data for a class, a local model will be obtained to describe the behavior of this class. For the samples or objects in test sets, the developed soft model can be used to compare the similarity between the object and each class and determine whether or not the object belongs to any class of the training sets. The SIMCA technique is also suitable where the unknown belongs simultaneously to two or more classes. For instance, a compound with sweet and salty tastes can fall into two classes.

For the training sets, each sample i ($i = 1,...,n$) can be described by descriptor variables k ($k = 1,...,d$):

$$x_{ik}^{(q)} = \alpha_k^{(q)} + \sum_{a=1}^{A_q} t_{ia}^{(q)} p_{ak}^{(q)} + e_{ik}^{(q)} \qquad (2)$$

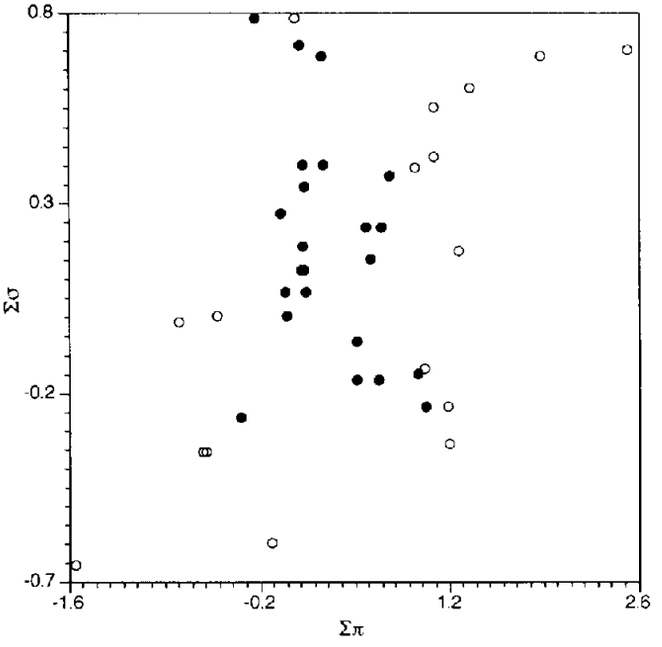


Fig. 2. Plot of the Hammett Σσ against the Hansch Σπ for carcinogenic dimethylaminoazobenzenes. • = carcinogenic, o = non-carcinogenic.

where $x_{ik}^{(q)}$ is descriptor k for sample i in class q; $\alpha_k^{(q)}$ is a mean value of the k-th descriptor in class q; A_q is the dimensionality or the number of components in class q and is usually determined by the cross-validation technique [18]; $t_{ia}^{(q)}$ is the score of sample i; $p_{ak}^{(q)}$ is the loading describing the correlation of parameter k with the a-th principal component (PC); and $e_{ik}^{(q)}$ is the residual of data which can not be explained by PC.

In SIMCA an approximate F-test is used for the classification of objects. In the F-test, the class fit distance $D_j^{(q)}$ between object j and class q model is compared with the residual standard deviation (R.S.D.) of the class q in the training set. $D_j^{(q)}$ is given by the following equation:

$$D_j^{(q)} = \left[\sum_{k=1}^{d} e_{jk}^{(q)2} / (d - A_q) \right]^{1/2} \quad (3)$$

In addition, sometimes there are "outliers" — the examined compounds are found to belong to no defined class because $D_j^{(q)}$ is much larger than the R.S.D. for a given class. SIMCA can be used to detect outliers, if present.

The SIMCA method was applied by Miyashita *et al.* [15,16] to analyze the structure–taste relationships on sweet and bitter dipeptides [15] and sulfamates (R-$NHSO_3Na$) [16]. Fifty sulfamates consist of 14 sweet and 36 non-sweet samples. Seven descriptors, molar refractivity (*MR*),Taft's σ*, and modified Verloop's STERIMOL parameters (L, B_1, B_o, B_r and B_l), as 7-dimensional space were selected to characterize those 50 sulfamates. Here L is the length of a substituent R. The B parameters are the widths in the direction perpendicular to the L axis: B_1 is the minimum width of the substituent, B_o the parameter in the opposite direction to B_1, B_r and B_l are the right and left-hand widths, respectively. The modelling powers of these descriptors were evaluated for feature selection, as shown in Table 1. The modelling power of the k-th descriptor Ψ_k is defined by the following equation:

$$\psi_k = 1 - S_k / S_{k,x} \quad (4)$$

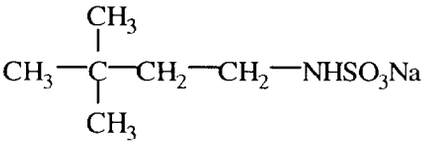
where S_k and $S_{k,x}$ are the residual and corresponding standard deviations of the k-th descriptor for the sweet class, respectively. If Ψ_k is nearly 1 all information on the descriptor k is useful for PLS modelling; if Ψ_k is nearly zero the descriptor k is useless. Then, only four descriptors, L, B_1, *MR* and B_o, were found to be significant on the basis of the modelling power ($\Psi_k > 0.3$) of each descriptor. The data structure of 4-dimensional selected descriptor space was asymmetric. The SIMCA model based on those 4 descriptors was obtained and correctly classed 13 out of the 14 (93%) sweet samples and 24 out of the 36 (67%) non-sweet samples in the training set. By using the SIMCA model six candidate sweet sulfamates were predicted. One of the candidate sulfamates,$(CH_3)_3C(CH_2)_2NHSO_3Na$ (as shown in Scheme 1) was synthesized and its sweetness was found to be about 3 times greater than that of sucrose.

TABLE 1. Modelling power of SIMCA technique

Descriptor variable	Modelling power	Rank
MR	0.496	3
L	0.585	1
B_1	0.513	2
B_o	0.349	4
B_r	0.219	5
B_l	–0.014	6
σ*	–0.354	7

In another study [17], the SIMCA method was used to investigate structure–taste relationships of substituted β-(3-hydroxy-4-methoxyphenyl)ethylbenzenes. The data set consisted of 9 sweet,1 bitter and 16 tasteless compounds. The SIMCA model for the sweet compounds with 2 principal components was obtained based on 15 physicochemical descriptors (MR_1,MR_2,MR_3, L_1,L_2,L_3, B_{1-1},B_{1-2},B_{1-3}, π_1,π_2,π_3, σ_1,σ_2,σ_3). The Karhunen-Loeve plot showed that the sweet compounds form a cluster and the non-sweet compounds are located outside the cluster in 15-variable space. In this case the data structure was asymmetric. The SIMCA box given by a 2-dimensional PC model (t_1 and t_2) is schematically shown in Fig. 3 with the R.S.D. Comparison of F statistics showed that 9 out of the 9 (100%) sweet componds and 17 of the 17 (100%) non-sweet compounds were correctly classified by the PC model.

It is concluded that the SIMCA technique is a powerful predictive tool for the sweetness qualities of sul-



Scheme 1.

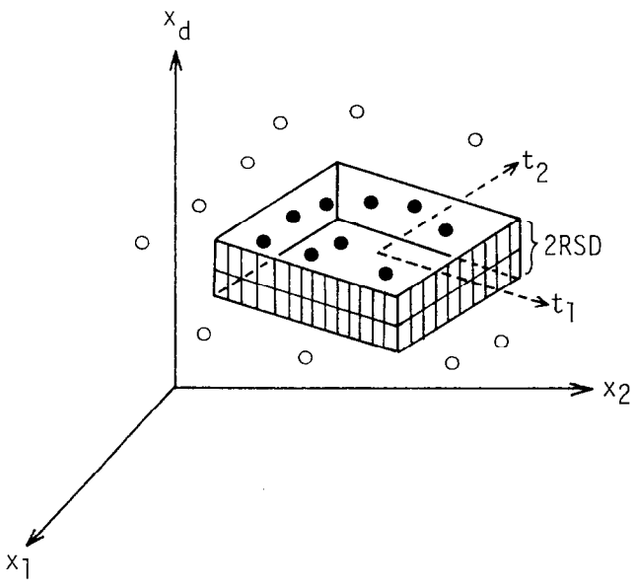


Fig. 3. Principal component model of SIMCA in the descriptor space. • = sweet, o = non-sweet.

famates and ethylbenzenes and can be used to develop the class models that provide the required important information on sweet compounds (*e.g.*, sulfamates and ethylbenzenes). SIMCA can also play an important role in the development of new agents or drug design.

The CSA method, modelling power and feature selection

Cluster significance analysis (CSA) is also an appropriate but conceptually simple statistical method to analyze structure–activity data, including asymmetric or imbedded data. A disadvantage of CSA is that it does not consider the interrelationship between the variables as SIMCA does and thus does not lead to a predictive model [19,20]. The basic principle and application examples of CSA were first described by McFarland and Gans [21–23]. The basic concept and main idea is given as follows. If a descriptor does not affect the activity of agents, then the active and inactive members of the examined compounds, for example the sweet and non-sweet sulfamates, will be scattered randomly along the axis corresponding to the descriptors. But, when the active compounds do form a non-random cluster or group along the axis, the corresponding descriptors affect the activity of the compounds to be examined. The location of the definite cluster in the whole descriptor space provides the approximate region associated with the actives and helps predict whether a new compound will be active or inactive. The CSA as an algorithm calculates the significance probability (P-value) that a perceived cluster might have arisen as a purely random event, and based on the P-values and their change by addition or subtraction of an examined descriptor, judges whether the descriptor is a relevant or non-relevant descriptor. For instance, when a relevant descriptor is subtracted from the system the P-values will be raised; and when a relevant descriptor is added to the system, the P-value will be reduced. In contrast, the addition and subtraction of a non-relevant descriptor will result in increase and decrease of the P-value, respectively.

How to calculate the P-values is a key problem. First, a proper definition of the cluster tightness is required. To this end the mean squared distance (MSD) among all compounds in the active class is obtained by calculating the squared distance between each pair of points in the descriptor space and then dividing the sum by the number of the pairs. Thus MSD is used as the measure of tightness. Then, the number of total combinations in which the same size as the active class can be taken at a time from all the compounds including the active and inactive ones is calculated and designated as NC. The other MSD values for the remaining combinations other than for the active class are computed in the same way and compared to the MSD value of the active class. If the combinations give a value less than or equal to MSD of the active class, then they are called clusters. The number of clusters (including the active one itself) is counted and designated as NA. Finally, the probability (P-value) that a cluster at least as tight as the observed active group would have arisen by chance alone is then given by $P = NA/NC$. This P-value thus reflects the significance of the relationship between the descriptors and activity. As always, the smaller the P-value, the less the chance aggregation. If the P-values are less than 0.05 (the significance level, $P < 0.05$), then this is confirmation that the descriptor parameters are jointly related to activity; otherwise, the idea is not accepted.

CSA was used by McFarland and Gans [21–23] to analyze the same sulfamate data [16]. The significance probability P for each descriptor and combinations of some descriptors within 95% confidence limits is given in Table 2. From Table 2, the CSA P-values for the individual descriptors can also be used as a measure of the modelling ability: the lower the P-value, the larger the modelling power. Based on the contribution of each individual descriptor only three descriptor parameters — L, B_1 and MR — are significant ($P < 0.05$, P = 0.0009, 0.0012, 0.0454, respectively), and the others —

TABLE 2. *P*-values of CSA method and rank of modelling power

Descriptor variable	*P*-value ± 95% confidence limits	Rank of modelling power
MR	0.001260 ± 0.000311	2
L	0.000900 ± 0.000131	1
B_1	0.045400 ± 0.005770	3
B_0	0.061400 ± 0.006654	4
B_r	0.125600 ± 0.009186	5
B_l	0.578600 ± 0.013687	6
σ^*	0.720200 ± 0.012443	7
L+*MR*	0.000027 ± 0.000010	
L+*MR*+B_1	0.000002 ± 0.000003	
L+*MR*+B_1+B_0	0.000010 ± 0.000006	
L+*MR*+B_1+B_0+B_r	0.000050 ± 0.000031	
L+*MR*+B_1+B_0+B_r+B_l	0.000600 ± 0.000215	

B_0, B_r, B_l and σ^* — are non-significant ($P > 0.05$). Although it is encouraging to expect that the combination of all seven descriptors is significant ($P < 0.05$), indeed the lowest of the *P*-values was obtained by the combination of only three variables and what is more interesting is that the three variables are still the same as *L*, *MR* and B_1. So these three descriptors can be selected as the features to describe QSAR of the examined sulfamates.

The results of CSA by McFarland and Gans [21–23] are slightly different from the SIMCA results of Miyashita *et al.* [16]: (1) B_0 may be not a significant parameter for sweet sulfamate compounds and (2) the rank order of modelling power evaluated by CSA differs only slightly from that estimated by the SIMCA method. For these two orders, the Spearman's rank-correlation coefficient is 0.96 with $P = 0.001$. The overall conclusions are very similar: *L*, *MR* and B_1 are the most important variables for estimating the sweetness of these sulfamates. The major difference concerns B_0: B_0 was chosen by the SIMCA method only on the basis of its modelling power ($\Psi_k = 0.349 > 0.3$) while it was discounted as being important by the CSA approach owing to its CSA *P*-value ($P = 0.0614$).

From the above, we draw the following conclusions. CSA can directly determine the significance of the descriptors both in individual and in combinations; however, it does not give a well defined boundary/location for the region of activity. By contrast, SIMCA may not directly deal with the issue of the modelling ability of descriptor combination but with the modelling rank of the individual descriptor; however, it can provide the individual modelling power of every input variable and the defined regional boundary/location in property space for predicting the activity of new drug members. In practice, CSA and SIMCA may be complementary to each other.

Concentric hypersphere model

When adequate quantitative biological data is not available for some drug series, it is often possible to deduce a QSAR model by treating it as a classification problem. The descriptor variables describing the active samples are compared to those of the inactive ones. When all the samples are plotted in the descriptor space, the active class members will frequently form a well defined cluster and the inactive class members will be scattered or distributed randomly in various directions (see Figs. 1 and 2). In this asymmetric case we are interested in modelling active potency data. It is difficult to obtain a statistically significant model equation for the asymmetric case by the classical Hansch-Fujita approach using the above descriptor variables; but the "concentric hypersphere model", derived from the SIMCA method, can cope with problems of this kind. The concentric hypersphere model can be used for both active (sweet) and inactive (non-sweet) compounds.

This fairly new concentric hypersphere concept is established in the following way: The class fit distance $D_j^{(1)}$ between object *j* and the sweet class (1) model is computed. The active potency is plotted aganist the distance $D^{(1)}$. The active potency is obviously a monotonically decreasing function of $D^{(1)}$, and the inactive samples have higher $D^{(1)}$ values than those of the active samples. The structure requirements for activity proposed by the concentric hypersphere model is that the responding $D^{(1)}$ value must be as small as possible.

In the study by Miyashita and his co-workers [16], the concentric hypersphere model was used to investigate the structure–taste correlation of substituted ethylbenzenes and to predict the sweet candidate ethylbenzenes. As is shown in the sweet potency–$D^{(1)}$ distance plot for the sweet compounds, the higher the sweet potency, the smaller the $D^{(1)}$ value; and for both taste-

less and bitter compounds, the $D^{(1)}$ distances are always great. The results are shown in Fig. 4 and Table 3.

We conclude that the concentric hypersphere model and the SIMCA method are useful for the rational design of new active substances, including not only new sweet compounds but also new potent drugs. Another useful approach to addressing such a concentric hypersphere modelling problem may be the CARSO (computer aided response surface optimization) technique developed by Clementi *et al.* [24] and the QPLS (quadratic PLS) procedure by Wold *et al.* [25] for the non-linear PLS modelling, which will not be discussed in detail here.

PLS modelling of univariate and multivariate activity data

In the classical MLR method of QSAR, the chemical descriptors are often assumed to be independently distributed, precise and 100% revelant, but this situation is not common in QSAR problems. Therefore by using the classical QSAR methods, the useful information may not be extracted from the chemical descriptor data with fairly high correlation. In other words, some multivariate analysis methods including MLR often cannot provide predictive models.

Partial least squares regression in latent variables (PLS) is a relatively newly developed multivariate statistical analysis method which is appropriate for investigating systems under indirect observation where the latent variable model extracts the patterns inherent to descriptor variables which have predictive power for

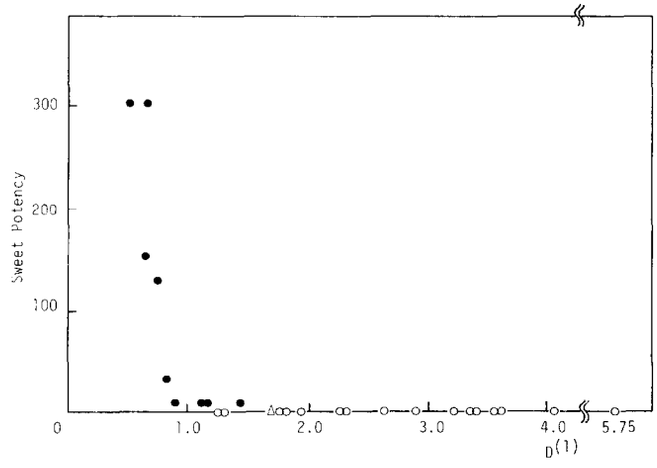
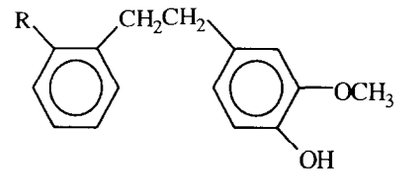


Fig. 4. Plot of the sweetness potency of ethylbenzenes against $D^{(1)}$ distance between each object and the sweet-class center in 15-variables space. • = sweet, o = tasteless, Δ = bitter.

TABLE 3. Basic structure and substituent of ethylbenzenes and $D^{(1)}$ class fit distance of the sweet ortho-substituted candidates



Number	R	$D^{(1)}$ distance
1	F	0.866
2	CH_2CN	0.937
3	CH_2Cl	0.986
4	SH	1.016
5	NHCHO	1.055
6	$OCHCH_2$	1.095
7	Cl	1.109
8	CH_2CH_3	1.127
9	N_3	1.131
10	*trans*-CHNOH	1.149
11	$NHNH_2$	1.161
12	$CHCH_2$	1.170
13	CHO	1.185
14	CCH	1.197

dependent variables [26]. PLS can overcome the shortcomings of MLR and therefore play an important role in QSAR studies. Also, PLS is widely employed to solve multivariate calibration and resolution in analytical chemistry [27].

On the basis of the type of activity data to be processed, the PLS can be divided into two: PLS1 and PLS2 to handle univariate and multivariate activity QSAR problems, respectively. Only a brief description of the PLS method is given here; for details the reader is referred to Dunn *et al.* [26]. The PLS model is derived in a PCA-like expression and gives the relation of descriptor data $\mathbf{X}$ and activity data $\mathbf{Y}$ using latent components t:

$$\mathbf{X} = \mathbf{X}_0 + \sum_{a=1}^{A} \mathbf{t}_a \mathbf{p}_a' + \mathbf{E} \quad (5)$$

$$\mathbf{Y} = \mathbf{Y}_0 + \sum_{a=1}^{A} \mathbf{b}_a \mathbf{t}_a \mathbf{q}_a' + \mathbf{F} \quad (6)$$

here $\mathbf{X}_0$ and $\mathbf{Y}_0$ are the corresponding mean value matrices; $\mathbf{p}_a'$ and $\mathbf{q}_a'$ are the transpose of loading vectors

for the **X** and **Y** blocks in the a-th component, respectively; $\mathbf{b}_a$ is a sensitivity; A is the dimensionality of the PLS model; **E** and **F** are the residual matrices of **X** and **Y**, respectively. The latent variable $\mathbf{t}_h$ is expressed using a linear combination of original chemical descriptors:

$$\mathbf{t}_h = (\mathbf{X} - \mathbf{X}_0 - \sum_{a=1}^{h-1} \mathbf{t}_a \mathbf{p}'_a)\, \mathbf{w}_h \quad (7)$$

here $\mathbf{w}_h$ is a PLS weight vector for **X** block (variables data matrix) of the training set in component h. If eqn. 7 is substituted in eqn. 6, then an MLR-like model equation, similar to MLR,of **Y** in terms of **X** is obtained:

$$\mathbf{Y} = \mathbf{Y}_0 + \mathbf{S}(\mathbf{X} - \mathbf{X}_0) \quad (8)$$

here **S** is a coefficient matrix for the PLS predictive model. Eqns. 7 and 8 provide insight in terms of latent variables and original chemical descriptors, respectively.

In PLS the "best model equation" is obtained by selecting the proper number of latent variables using a cross-validation procedure [18,28]. In the case of cross-validation, one or more examples of the calibration set are omitted and the rederived QSAR is used to predict the omitted dependent activities. This process is repeated until each of the dependent activity values has exactly been predicted once and gives the cross-validated or predictive r^2_{xval}. The cross-validated r^2_{xval} is given by

$$r^2_{\text{xval}} = 1 - \sum (y_{\text{obs}} - y_{\text{pred}})^2 / \sum (y_{\text{obs}} - y_{\text{ave}})^2 \quad (9)$$

where y_{obs}, y_{pred}, y_{ave} are observed, predicted and average activities, respectively. The value r^2_{xval} becomes negative when the difference between y_{pred} and y_{obs} is large. If $r^2_{\text{xval}} = 1.0$ the prediction is perfect; in contrast, if $r^2_{\text{xval}} = 0.0$ the prediction is no better than "no model at all" (physically meaningless); and in the worst case, $r^2_{\text{xval}} < 0.0$, such a QSAR model is worse than the above case (no model at all). This technique is one of the powerful diagnostic tools for eliminating chance correlations and avoiding overfitting in the construction of PLS models. In this way, the number of components or the dimensionality of the model is determined according to its predictive ability rather than its fitting power. Both the fitting and predictive errors can be estimated as a function of model complexity (dimensionality). As the model dimensionality increases, the fitting error becomes small but the predictive error will have a minimum value that is an optimum point. This situation is shown in Fig. 5. Overfitting can be avoided and then the optimal predictive models obtained.

PLS is a robust method that is not sensitive to collinearity; and this method can treat a singular descriptor array with more variables than objects. PLS has the following advantages over the classical unbiased methods (including MLR):

- collinearities exist among the descriptors;
- the number of descriptors is larger than that of samples;
- overfitting is avoided;
- the model is predictive;
- the common latent variables are available for simultaneously processing multivariate activity data by PLS2 analysis.

The PLS method has been applied to solve various QSAR problems [29–31] including binary classification [30] and multivariate activity data processing [31]. For instance, it was successfully used to investigate the discrete binary antiarrhythmic activity data of 14 phenylpyridines using a two-component PLS model based on three measured and computational chemical descriptors, *MR*, *TOR* (torsion angle) and *PA* (proton affinity), and to distinguish the actives from the inactives [30]. The modelling powers of *MR*, *TOR* and *PA* are 0.49, 0.39 and 0.75, respectively. The recognition rate of the proposed PLS binary classification was 86%

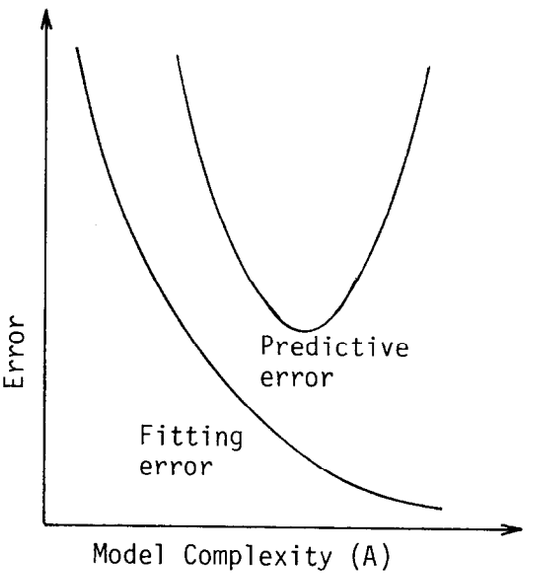


Fig. 5. Plot of PLS modelling errors against the model complexity *A*.

and the predictive ability based on the leave-one-out cross-validation procedure was 79%; the predictive ability using the computerized learning machine was 71%. These results show that the PLS technique is more predictive and promising in multivariate QSAR analysis.

In another example [31], the PLS2 method of a three-component PLS model in terms of five original parameters was used in a multivariate QSAR analysis of 83 thiolcarbamates with both fungicidal and herbicidal activities and provided a total picture of structure–activity relationships according to the leave-one-out procedure. The PLS model equations for both fungicidal activity y_F and herbicidal activity y_H are expressed in terms of the first three latent variables t_1, t_2 and t_3:

$$y_F = -0.484\ t_1 + 0.270\ t_2 - 0.197\ t_3 + 0.732 \quad (10)$$

$$y_H = 0.328\ t_1 + 0.513\ t_2 + 0.445\ t_3 + 2.133 \quad (11)$$

The variables of the PLS model equation indicate how to change the structure to have higher activities of both. Comparing y_F and y_H, the coefficients of t_2 for the two activities possess identical signs and those of t_1 and t_3 opposite signs. If t_2 is increased and both t_1 and t_3 are restricted to a small interval near to the centers of t_1 and t_3, new thiolcarbamates with both more potent fungicidal and herbicidal activities can be predicted. The predictive abilities of the PLS2 model of three significant components based on 5 descriptor parameters are 81% and 76% for fungicidal and herbicidal activities, respectively. In addition, PLS2 makes it quite easy to interpret the structural requirements for the compounds with both activities because the activities are modelled by the common latent variables. On the basis of the PLS model equations, 21 candidate thiolcarbamates with both high fungicidal and herbicidal activities are proposed. The PLS2 results were compared with those of the adaptive least squares method (ALS) [32] and the PLS2 was found to be more powerful than ALS. It is noted that the PLS2 method can be applied to derive structural information for optimizing the activity profile and discover the regularities of the activity profile of biological data. Also, the PLS2 method can improve interpretability of X–Y relationships, give clearer understanding of the QSAR data and provide better prediction of biological activities, which will be important and effective in the development of new potent drugs.

Recently, three-dimensional (3-D) QSAR studies [28,33] have been developed as an interesting and promising field; here the term "3-D" refers only to the three-dimensional structure of the compounds. It should not be confused with multi-way data analysis and with multivariate QSAR analysis. A new and important 3-D QSAR approach, the so-called comparative molecular field analysis (CoMFA), was initially presented by Cramer *et al.* [28,33,34] and this method has become a popular and valuable tool for computer-aided 3-D drug design. Its rationale is two-fold: (1) the interactions which produce biological activity are usually non-covalent and (2) the non-covalent interactions can account for a change in biological activities. MLR is not suitable for CoMFA analysis because the number of descriptors is much greater than that of compounds. However, PLS gives robust model equations for 3-D QSAR problems.

A complete description of CoMFA is given elsewhere [28,34]; only a basic introduction to this method is described below. The characteristics of CoMFA, as implemented currently, are

- a fitting technique for optimal mutual alignment within a series,
- description of ligand molecules by its steric (Lennard-Jones or van der Waals) and electrostatic (Coulombic) fields sampled at the intersections of a grid in 3-D space,
- data processing by PLS analysis of descriptors and biological data using the cross-validation method to maximize the likehood that the results possess predictive ability,
- converting the PLS equation to the MLR-like equation and presenting results as graphic contoured 3-D plots.

The CoMFA approach was used by Cramer *et al.* [28,34] to examine the effect of shape on the binding of steroids to carrier proteins and estimate a possible and useful strategy for 3-D structure recognition. Recently, CoMFA and PLS have been used to address 3-D QSAR of HMG-CoA reductase inhibitors [35]. The inhibitors block biosynthesis of cholesterol. Based on a training set composed of 13 compounds of known pIC_{50} values and CoMFA descriptors, an excellent PLS model was obtained.

In QSAR, hydrophobicity is another important factor in addition to the above-metioned steric and electronic factors. Recently, Kellogg *et al.* [36] integrated a 3-D hydrophobic field into CoMFA by adding a hydrophobic factor. The steroid binding problems reported by Cramer *et al.* [33] were reanalyzed using the CoMFA method with the hydrophobic field incorporated. The CoMFA coefficient counter from the hydrophobic field defined the most active steroid molecules. Miyashita *et al.* [35] also intergrated molecular lipophilicity potential(MLP) into CoMFA.

Conclusions

QSAR has had a remarkable influence on drug design. The MLR method initially proposed by Hansch *et al.* [1,2] is widely used for QSAR studies and regarded as a classical and relatively hard method. Indeed, in QSAR, there are several types of data such as continuous and discrete, symmetric and asymmetric, univariate and multivariate. For various data or problems, different methods should be employed. Certainly, global methods or hard models should be first considered. When they are not suitable, the soft methods or local models are applied. The global hard model and the local soft models are not mutually exclusive but complementary. For some QSAR problems, their combination makes the solution more effective and powerful. Up to now, a lot of methods, including the classical and relatively hard MLR method and new soft PLS method, have been developed for QSAR studies. But no method is perfect in every way, each having individual advantages and shortcomings.

In general, PLS with cross-validation is excellent and predictive and can be recommended; particularly in the cases where the descriptor variables are neither normally nor independently distributed and the number of compounds is much fewer than that of the descriptors. However, PLS has some weaknesses [37]; for instance, there is a risk of overlooking a correlation actually present in the data where the first few PLS components can explain the data variance quite well and the later PLS components may not be considered through the cross-validation procedures [37].

There are many related problems in analytical chemistry, such as multivariate calibration [27], monitoring environmental toxicology and composition [3], relating analytical characteristics with chemical structure, particularly the relationship between reagent structure and spectroscopic response, and quantitative structure–chromatographic retention relationships (QSRR) [38,39], which are similar to QSAR. The data analysis techniques that work well for QSAR are directly applicable to the above analytical problems.

Acknowledgements

One of the authors (Zhiliang Li) is grateful for a postdoctoral fellowship from the Japanese Ministry of Education, Science and Culture (Monbusho). We thank the Computer Center of the Institute for Molecular Science for affording us facilities for computation. We also thank the editors and referees for their helpful comments.

References

1 C. Hansch and T. Fujita, *J. Am. Chem. Soc.*, 86 (1964) 1616.
2 C. Hansch, *Acc. Chem. Res.*, 2 (1969) 232.
3 W.J. Dunn III and S. Wold, *Trends Anal. Chem.*, 5 (1986) 53.
4 W.J. Dunn III, *Chemom. Intell. Lab. Syst.*, 6 (1989) 181.
5 Y.C. Martin, *Quantitative Drug Design*, Marcel Dekker, New York, 1978.
6 D. Hadzi and B. Jerman-Blazic (Editors), *QSAR in Drug Design and Toxicology*, Elsevier, Amsterdam, 1987.
7 J.L. Fauchere (Editor), *QSAR: Quantitative Structure–Activity Relationships in Drug Design*, Alan R. Liss, New Nork, 1989.
8 S. Sasaki, H. Abe, Y. Takahashi, C. Takayama and Y. Miyashita, *Introduction to Pattern Recognition for Chemists*, Kagaku-Dogin, Tokyo, 1984.
9 O. Strouf, *Chemical Pattern Recognition*, Research Studies Press, Letchworth, 1986.
10 W.J. Dunn III, S.L. Emery, W.G. Glen and D.R. Scott, *Environ. Sci. Technol.*, 23 (1989) 1499.
11 S. Wold and W.J. Dunn III, *J. Med. Chem.*, 21 (1978) 1001.
12 S. Wold, *Patt. Recogn.*, 8 (1976) 127.
13 S. Wold and M. Sjostrom, in B.R. Kowalski (Editor), *Chemometrics: Theory and Applicatins (ACS Symposium Series*, Vol. 52), Am. Chem. Soc., Washington DC, 1977, pp. 243–282.
14 S. Wold and E. Johnsson, *Anal. Chim. Acta*, 133 (1981) 251.
15 Y. Miyashita, Y. Tahahashi, C. Takayama, K. Sumi, K. Nakatsuku and S. Sasaki, *J. Med. Chem.*, 29 (1986) 906.
16 Y. Miyashita, Y. Takahashi, C. Takayama, T. Ohkubo, K. Funatsu and S. Sasaki, *Anal. Chim. Acta*, 184 (1986) 143.
17 Y. Miyashita, S. Kanaya, H. Katsumi, C. Takayama, A. Nagakura and S. Sasaki, *Chemical Senses*, 14 (1989) 781.
18 S. Wold, *Technometrics*, 20 (1978) 397.
19 L. Hodes, G.F. Harzard, R.I. Geran and S. Richman, *J. Med. Chem.*, 20 (1977) 469.
20 L. Hodes, *J. Chem. Inf. Comp. Sci*, 21 (1981) 132.
21 J.W. McFarland and D.J. Gans, *J. Med. Chem.*, 29 (1986) 505.
22 J.W. McFarland and D.J. Gans, *J. Med. Chem.*, 30 (1987) 46.
23 J.W. McFarland and D.J. Gans, *Drug Inf. J.*, 24 (1990) 705.
24 S. Clementi, G. Cruciani, G. Curti and B. Skagerberg, *J. Chemom.*, 3 (1989) 499.
25 S. Wold, N. Kettaneh-Wold and B. Skagerberg, *Chemom. Intell. Lab. Syst.*, 7 (1989) 53.
26 W.J. Dunn III, S. Wold, U. Edlund, S. Hellberg and J. Gasteiger, *Quant. Struct.–Activ. Relat.*, 4 (1984) 131.
27 H. Martens and T. Naes, *Multivariate Calibration*,

Wiley, New York, 1989.
28 R.D. Cramer III, D.E. Patterson and J.D. Bunce, *J. Am. Chem. Soc.*, 110 (1989) 5959.
29 P. Berntsson and S. Wold, *Quant. Struct.–Activ. Relat.* 5 (1986) 45.
30 K. Hasegawa, Y. Miyashita, S. Sasaki, H. Sonoki and H. Shigyou, *Chemom. Intell. Lab. Syst.*, 16 (1992) 69.
31 Y. Miyashita, H. Ohsako, C. Takayama and S. Sasaki, *Quant. Struct.–Activ. Relat.*, 11 (1992) 17.
32 O. Kirino, C. Takayama, M. Yoshida, S. Inoue and R. Yoshida, *Agric. Biol. Chem.*, 52 (1988) 561.
33 G.R. Marshall and R.D. Cramer III, *Trends Pharmacol. Sci.*, 9 (1988) 285.
34 M. Clark, R.D. Cramer III, D.M. Jones, D.E. Patterson and P.E. Simeroth, *Tetrahedron Computer Methodology*, 3 (1990) 47.
35 Y. Miyashita, Y. Shiraishi, K. Hasegawa and S. Sasaki, in M. Doyama, S. Sasaki, T. Suzuki, M. Tanaka and R. Yamamoto (Editors), *Proceedings of the 2nd International Conference and Exhibition on Computer Applications to Materials and Molecular Science and Engineering (CAMSE'92), Pacifico Yokohama, Japan, September 22–25, 1992*, Elsevier, Amsterdam, in press.
36 G.E. Kellogg, S.F. Semus and D.J. Abraham, *J. Comput-Aided Mol. Design*, 5 (1991) 553.
37 R.D. Cramer III, J.D. Bunce, D.E. Patterson and I.E. Frank, *Quant. Struct.–Act. Relat.*, 7 (1988) 18.
38 R. Kaliszan, *Quantitative Structure–Chromatographic Retention Relationships*, Wiley, New York (1987).
39 R. Kaliszan, *Anal. Chem.*, 64 (1992) 619A.

Dr. Yoshikatsu Miyashita is an associate professor at the Laboratory for Chemical Information Science, Department of Knowledge-based Information Engineering, Toyohashi University of Technology, Tempaku-cho, Toyohashi 441, Japan. His interests are in chemometrics, chemical pattern recognition (CPR), quantitative structure–activity relationships (QSAR), artificial neural networks, chemical graphy theory.
Dr. Zhiliang Li is a postdoctoral fellow on leave at the Laboratory for Chemical Information Science, Department of Knowledge-based Information Engineering, Toyohashi University of Technology, Tempaku-cho, Toyohashi 441, Japan and is an assistant professor at the Department of Chemistry and Chemical Engineering, Hunan University, Changsha 410082, People's Republic of China. His present interests include analytical chemistry, QSAR, chemometrics, optimal parameters estimation (Kalman filtering), neural networks, multivariate calibration and resolution.
Dr. Shin-ichi Sasaki is a professor and president of Toyohashi University of Technology, Tempaku-cho, Toyohashi 441, Japan. His interests are in chemical information sciences, organic chemistry, chemical expert systems, chemometrics, chemical pattern recognition, QSAR.

Pyrolysis–mass spectrometry under soft ionization conditions

Albert C. Tas* **and Jan van der Greef**
Zeist, Netherlands

Pyrolysis conducted under direct chemical ionization conditions enables the detection of compounds of higher molecular weight, and polar compounds of low volatility. Losses of such compounds during pyrolysate transport are avoided, to a large extent. Because this pyrolysis technique can be used directly in combination with advanced mass spectrometric equipment, such as tandem mass spectrometry, sophisticated analysis of pyrolysis products can improve our insight into the structure of such compounds.

*To whom correspondence should be addressed.

Introduction

The analysis of complex (bio)macromolecular samples and (bio)polymers imposes important challenges to the analytical chemist. Micro-organisms, cells, body fluids, plant tissues, raw materials and cell wall materials are examples of materials which are difficult to define chemically because of their high complexity. Many of the compounds in such matrices cannot be analysed directly by current instrumental analytical techniques such as gas chromatography, liquid chromatography and mass spectrometry, because of their high-molecular-weights and their strongly diverging physico-chemical properties. This complexity often makes a total analysis based on the selective analysis of groups of compounds a tedious job. Moreover, chemical and stereochemical barriers often block the selective action of enzymes or chemical reagents thus

0165-9936/93/$06.00