Journal of Computer-Aided Molecular Design, **13:** 453–467, 1999.
KLUWER/ESCOM

QSAR using 2D descriptors and TRIPOS' SIMCA

Peter A. Hunt
NRC Terlings Park, MSD, Eastwick Road, Harlow, Essex CM20 2QR, U.K.
E-mail: Peter_Hunt@Merck.com

Received 6 November 1998; Accepted 29 January 1999

Key words: atom-pair, classification analysis, principal components, soft modelling, topological-torsion

Summary

The combination of 2-dimensional descriptors and classification analysis has seen limited use within drug design either due to the general nature of the descriptors used or by the drive to use only 3D information. We present the use of SIMCA as implemented by TRIPOS in conjunction with our in-house 2D topological descriptors as a means of giving chemically significant analyses without the need for an alignment step. The TRIPOS method was applied to two published data sets, an in-house data set and two artificial data sets. The results showed that the structural features deemed to be necessary for the desired activity were identified. These experiments also highlighted the significant differences between the TRIPOS and literature versions of SIMCA. The potential uses of the SIMCA/2D technique seem limitless as any activity can be categorised.

Abbreviations: SIMCA, Soft Independent Modelling by Class Analogy (or SIMple Category Analysis); QSAR, Quantitative Structure Activity Relationships; PCA, Principal Component Analysis; NMDA, N-Methyl-D-Aspartic acid; CMR, Calculated Molar Refractivity.

Introduction

The arbitrary assignment of an alignment prior to an analysis is a constant problem with 3D, and certain 2D, QSAR methods. Another is the assumption that the biological data and the structural information are linearly correlated. When these two conditions are satisfied then methods such as CoMFA [1] are extremely useful techniques giving chemists much needed direction to improve activity or to design de novo compounds which overcome some undesirable property without losing the desired activity. However with activities such as efficacy or bioavailability the relationships between structure and activity are often complex and very unlikely to be linear. In this regard the use of discriminant or categorical analysis would appear to be the method of choice as the relationships are derived not with a definite numerical figure, but with a likelihood of belonging to a desired category of activity which is displayed by a variety of compounds.

One such categorical analysis technique is SIMCA – Soft Independent Modelling by Class Analogy, pioneered by Wold et al. [2] (herein called W-SIMCA). It has been used extensively in process applications to help monitor variations in process conditions relative to end product quality [3]. Other uses have been in analytical applications, especially within the food industry, to identify regional variations in products using chromatographic data [4, 5]. Very little has been published in the use of W-SIMCA, or in fact any categorical analysis, for chemical structure activity relationships [6–8]. W-SIMCA takes a precategorised training set and, for each category in turn, models the members of that category by the principal components of the explanatory data for that category; hence a three category data set is expressed as three independent models. The principal components for each category are limited in length and therefore define a hypervolume in which the category members lie. Category membership for a test point then depends upon whether the point falls within one of these defined vol-

umes; if it does then the point belongs in the category. This matching against each category separately means that one point may quite legitimately belong to more than one category if the hypervolumes overlap. It also allows the possibility that the point does not fall in any category (similar to a 'none of the above' answer in multiple choice examinations), which indicates that the test point belongs to an unknown category which has not been expressed by the training set.

It must be stated here that the TRIPOS implementation of SIMCA (herein called T-SIMCA) differs quite markedly from the above, W-SIMCA and whilst still useful, the differences alter the behaviour of the algorithm and the results that are obtained. The major difference lies in the length of the principal components derived for each category. In T-SIMCA these components are infinite and this causes some problems as will be described later. Another difference is that a test point is always assigned to one of the predefined categories, the 'none of the above' option is no longer available, which is unfortunate as such knowledge can be of great importance. Finally the nomenclature used in the TRIPOS implementation is also different as the first component in T-SIMCA refers to the displacement vector of the mean of a category's data from the mean of the whole data set. It is in fact the second component which is actually the first principal component of a particular category's data.

The standard SIMCA output parameters are 'Loadings', 'Scores', 'Discriminating Power' and 'Modelling Power'. The Loadings are the most useful in a multi category data set and are almost equivalent to the Loadings matrix found within normal Principal Component Analysis (PCA). The 'almost' refers to the above-mentioned difference in nomenclature between W-SIMCA and T-SIMCA. The Loadings for the first component in T-SIMCA show how the mean of the data for any category relates to the mean of the whole data set, and the 'Scores' matrix for the first component is unity. If a descriptor influences the bioactivity then the 'active' compounds should consistently contain/lack that descriptor and be suitably displaced from the mean of the data set. A large, usually positive 'Loadings' value in a category means that the presence of that descriptor is beneficial to membership of that category, the opposite, usually negative, sign for a Loading means that the absence of that descriptor benefits membership of that category.

The Modelling Power output describes the fraction of the column variance which is used in deriving the components, therefore values near to zero indicate irrelevant descriptors whilst values near unity indicate important descriptors. The Discriminating Power output describes the ratio of the sum of squared residuals for a descriptor when compounds are fitted onto the models for every category apart from their own, and the sum of squared residuals when compounds are fitted onto the model for their own category. Thus a large value for Discriminating Power indicates that a descriptor can distinguish one category from another. Both these powers can mislead by showing high values. For instance if the variance of a descriptor column is low then all of the column variance is used in a model and this gives a falsely high Modelling power reading. We stipulate a minimum variance for each descriptor column as we feel it is important in T-SIMCA if one is to use the Modelling power output effectively. The usual application of T-SIMCA within SYBYL is with CoMFA columns and here the automatic column filtering process removes the low variance columns as described above; however, this filtering *only* applies to CoMFA columns and therefore the user must be aware of the problems that could arise. Also large Discriminating powers in a multi-category model may arise, for example, with a descriptor which is very good at distinguishing the inactive from the not-so-inactive but has no power to distinguish the highly active from the inactive.

The literature relevant to SAR and W-SIMCA is quite general [9–11] but deals mainly with either the structure-taste relationships for artificial sweeteners [9] or the toxicities of aryl nitro/nitroso compounds [10]. With the artificial sweeteners the two category classification was performed using seven gross structural variables, two of which were STERIMOL [12] derived length parameters which require some sort of prior alignment step. In this paper we wish to introduce the application of our in-house topological molecular descriptors [13] as variables in the categorisation of compounds from designed data sets, an in-house N-Methyl-D-Aspartic acid (NMDA) receptor data set and two literature data sets. Our descriptors represent not only the atom type information within a molecule but also the partial charge, hydrophobic and donor-acceptor properties contained therein. Also the fact that there is virtually no path-length cutoff in our descriptors allows some long range features to be included which some descriptions ignore. We hope to prove that in combination, categorical analysis using our descriptors can give structure activity information which is of chemical use in directing the modification of compounds to elicit the desired biological activity.

We also highlight the differences between the published W-SIMCA and T-SIMCA which is at present implemented within TRIPOS' SYBYL [14].

Method

The T-SIMCA implementation within SYBYL6.3 was used in all these experiments and whilst the author acknowledges the existence of possibly better classification algorithms [15, 16], this T-SIMCA was readily available and hence a convenient starting point. The 2D topological descriptors of atom-pair and torsional relationships for each compound within each data set were generated using the in-house TOPOGEN routine running on Silicon Graphics workstations using either R4000, R4400, or R10000 processors under Irix 6.3. A brief description of the nature of these descriptors will be given in the following subsection covering the halobenzene data set. The compounds were added to data-set specific molecular spreadsheets with each relevant descriptor as a separate column. Categorisation of the data sets was performed by a specifically written Sybyl Programming Language (SPL) macro. For all but the designed datasets, this macro assigns Category 1 around the most negative or smallest values of the desired activity and the other categories towards the larger, more positive values. As the activity scales change with the different examples the most and least active categories will be highlighted in each case in the discussion. The 'relevance' of a descriptor depended upon whether the frequency of occurrence of that descriptor in a molecule (hereafter referred to simply as the descriptor frequency), or just its presence/absence was being used. In the latter case a descriptor was not included if it was present in $\leq$10% or $\geq$90% of the compounds under consideration within a data set; when descriptor frequencies were used the same cutoff at $\leq$10% was applied but no upper limit was set. When descriptor frequencies were used the columns of data were also subject to a minimum Standard Deviation of 0.32 (i.e. a minimum variance of 0.1) and any falling below this limit were rejected. The T-SIMCA analyses were run using all the descriptor columns which survived these cutoffs. Some of the output displays were constructed from specially written pieces of SPL and NAWK which mainly dealt with reformatting either the Loadings or Discriminating Power or Modelling Power output from T-SIMCA.

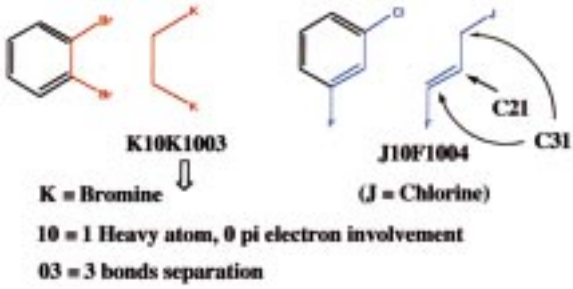


Figure 1a. Illustration of the atom-pair regular descriptor type.

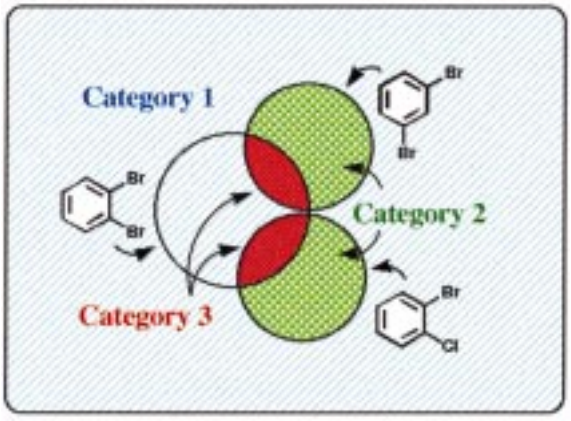


Figure 1b. Venn diagram to illustrate the features required for category membership.

Cubic data set

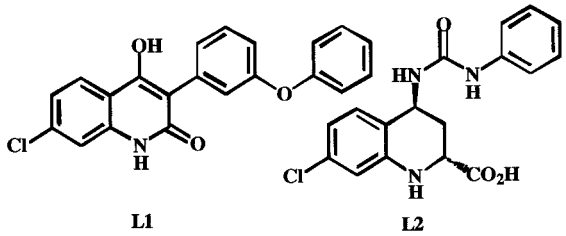
This very simple data set was a collection of 30 points in a Cartesian coordinate system where the explanatory data were the x,y,z coordinates of each point. The points were categorised by their location in space and the purpose of the set was to demonstrate how T-SIMCA classified a series of 216 test points which filled the remaining space. This data set was created after the following data sets had been examined as a simple means of testing T-SIMCA after some unexpected behaviour was observed.

Halobenzene data set

A constructed data set of 64 halobenzenes was built by taking a benzene ring and randomly substituting between two to four positions with a randomly selected halogen from the list of F,Cl,Br. This data set was devised to test whether or not it was possible for T-SIMCA to determine 'rules' for category membership in a very controlled situation. If T-SIMCA could not highlight the desirable features in a simple case where

we knew the answer then we could have no confidence in the method finding the important features in a real situation. Two structural rules were chosen to create an 'active' category, Category 3, (i.e. that there must be a bromine ortho to another bromine) and a 'partially active' category, Category 2, (i.e. that there must be a F meta to a Cl). These are shown graphically in Figure 1a along with their atom-pair regular (APREG) descriptors. Finally the remaining compounds, after the above categories were created (no compounds were allowed to belong in both the 'active' and 'partially active' categories), were designated as being 'inactive'. The 'REGular' descriptors use atom types (C,N,O,S etc.) with certain two-letter elements being represented by single letter codes (e.g. K for bromine) for programming simplicity. The next two numbers describe how many heavy atom connections that atom has and the pi-electron involvement of the atom. For the halogens the heavy atom connection value is 1 but it would be either 2 or 3 in the case of the aromatic carbons (as indicated by the C21 and C31 arrows). The pi-electron involvement is 0 for the halogens and 1 for the aromatic carbons and it would be 2 if an atom were to be involved in an acetylene or allene system. The final two numbers of an atom-pair descriptor indicate how long the through-bond path is between the two atoms. This is not needed for the topological-torsion descriptors because they only deal with the four connected atoms which form a torsion. Hence the ortho bromine relationship could have equally well been chosen using the REGular torsion descriptor (TTREG) K10C31C31K10. For the non-REGular descriptors which deal with the HYDrophobic, Donor/Acceptor (DA), and Partial Charge (PC) properties of the atoms involved in an atom-pair or torsion then numbers are used to represent the atoms. These numbers (from 1 to 9 for HYD and PC descriptors, 1 to 7 for DA descriptors) represent the 'bucket' into which that atom has been placed by the TOPOGEN routine. For example the atom-pair hydrophobic (APHYD) descriptor 9908 means two hydrophobic atoms (bucket 9) separated by 08 bonds.

In a second categorisation of this data set the 'active' category, Category 3, was defined by the presence of the previous ortho bromo relationship and the presence of *either* a bromo ortho to a chlorine *or* a bromo meta to another bromine. The 'partially active' category, Category 2, was defined by the presence of *either* the bromo ortho to a chlorine *or* the bromo meta to another bromine *without* the ortho bromines relationship. The 'inactive' set Category 1 were again the remainder



L1 L2

Figure 2. Active structures whose bicyclic cores were used in the database searches.

which also included compounds which had the ortho bromines relationship but *neither* of the other two features. This more complex relationship can be viewed as a Venn diagram as shown in Figure 1b with the categories being shown by the three colours and the circles indicating the compounds within the whole set which had the structural features indicated.

RPR data set

For the Rhone-Poulenc-Rorer (RPR) Phosphodiesterase (PDE)-IV literature [17] compounds a variety of categorisations were chosen based upon log of the activity in μM at the PDE-IV enzyme. In the three category model the most active compounds were in Category 1 and the least in Category 3. This data set and the following set were examined to see whether literature conclusions could be reproduced by the method and the different categorisations were used to see if any improvement in categorisation quality could be obtained.

Fisons data set

The published [18] Fisons (now Astra Charnwood) Calcium channel set was divided into either 2, 3 or 4 categories based upon their efficacy at increasing the cardiac contractility of guinea pig atria paced at 1 Hz. In the two category case the most active compounds were in Category 2 and the least in Category 1. In addition to the TOPOGEN descriptors there were columns of Calculated Molar Refractivity (CMR) and measured LogP data added to the data table as they had been shown to be important features in the original literature analysis.

NMDA data set

This is a large data set of 405 compounds from our in-house program to find an N-Methyl-D-Aspartic acid receptor glycine site antagonist [19]. The set was derived from core substructure searches of the Merck

Table 1. Halobenzene data set output parameters from the simplest 'indicator variable' type categorisation with the Loadings from component 1

Descriptor	Discrim_Pwr	Model_Pwr	Desc_Name	Cat. 1	Cat. 2	Cat. 3
APREG_920	2.9 e+32	1.00	J10F1004	−0.55	1.78	−0.55
APREG_953	2.7 e+31	1.00	K10K1003	−0.59	−0.59	1.64
TTREG_409	2.7 e+31	1.00	K10C31C31K10	−0.59	−0.59	1.64
TTHYD_9669	2.7 e+31	1.00	TTHYD_9669	−0.59	−0.59	1.64
APREG_41	62.43	0.97	J10C3101	−0.22	0.90	−0.38

Table 2. Halobenzene data set and the effects of varying the maximum number of components allowed when forming the model in the simplest 'indicator variable' type categorisation

Max. no. of components per category	No. correctly classified Cat. 1	No. correctly classified Cat. 2	No. correctly classified Cat. 3
1	23/32	13/15	14/17
2	24/32	14/15	16/17
3	29/32	14/15	17/17
4	30/32	14/15	17/17
5	31/32	15/15	17/17
6	31/32	15/15	17/17 (5 components)
8	32/32 (7 components)	15/15	17/17

compound collection based upon the two active compounds labeled L1 and L2 (Figure 2). The data was split into either five or ten categories with Category 5 being the most active category in the five category case. The reason for using such a set was to be able to have many compounds per category and hence be able to have more categories with narrower affinity ranges and attempt to make affinity predictions which might come close to those from CoMFA/PLS (Partial Least Squares) analyses.

Results and discussion

Cubic data set

To test T-SIMCA this data set of 30 training points and 216 test points was created using their x,y,z coordinates as the explanatory data. The scatter graphs shown in Figure 3 show these data and test points and they are coloured by the categorisation that various T-SIMCA models place upon those points. In Figure 3a the 30 training data points are shown with 'Category 1' in black along one edge, 'Category 2' in blue covering two corners and 'Category 3' in red occupying only one corner. Figures 3b-d show how the categorisation changes as one increases the maximum number of components allowed from 1 to 3, respectively.

The 1 component model Figure 3b has all the training points correctly classified and the classification of the test points seems reasonable, hence one should believe that allowing more components into the model should not change the model greatly. When a more complex 2 component model is tried however, one can see in Figure 3c that some of the black Category 1 training set points are now misclassified as being in the red Category 3 despite being in the opposite far corner. This is due to the fact that in this new model categories 3 and 2 are now defined by 2 infinitely long components whilst Category 1 is only defined by 1 component. As one of the components defining Category 3 points in the direction of the opposite corner it makes those Category 1 points closer to it than to the component which defines Category 1. When one extends the number of allowed components even further to 3 and generates a new model (Figure 3d) then all of the training points are again correctly categorised. With this more complex model the categorisation of the test points is now distinctly different compared to that shown in Figure 3b or 3c with even fewer of

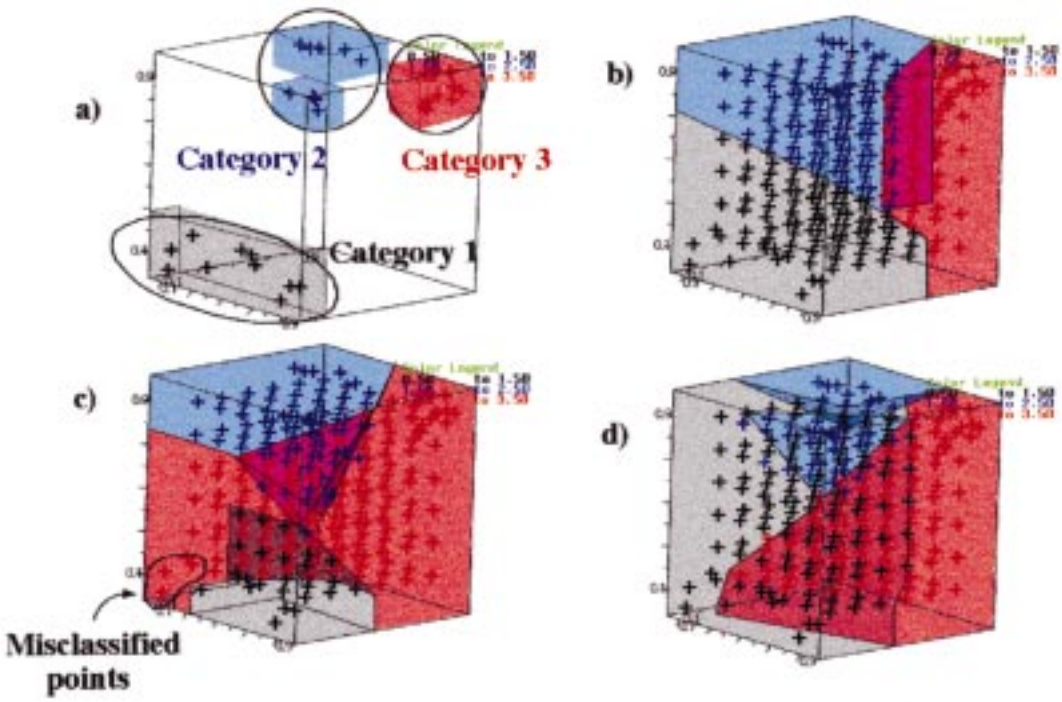


Figure 3. Cubic data set of (a) 30 training points coloured by predefined category; (b) training and test points coloured by predicted category in a 1 component model, the shading is an approximate indication of volume occupied by the predicted points in each category; (c) as (b) but now using 2 components, note the three Category 1 points misclassified as Category 3; (d) as (b) but now using 3 components.

the test points being predicted to be in Category 2. This again is due to the fact that Category 1 now has 3 components defining it whilst the other categories have only 2. A test point will be always as close as or closer to a higher dimensional category than a lower one (i.e. a point is closer to a line than to a point or closer to a plane than to a line etc.).

The model produced by T-SIMCA, if just allowed to use as many components as it wishes, is the 3 component model which seems to go against Wold's approach and indeed Occam's razor of 'simplest is best'. As both Wold and TRIPOS claim to use predicted sum of squared residuals it is difficult to rationalise this discrepancy.

Halobenzene data set

The first categorisation of the designed halobenzene data set had 17 compounds in the 'active' category – Category 3, 15 compounds were 'partially active' – Category 2, with the remaining 32 compounds being 'inactive' – Category 1. T-SIMCA was able to classify correctly all the compounds into their appropriate categories using 8 components. If one considered only the successful categorisation of all the compounds of the 'active' categories then only 5 components were needed (which is consistent with the view that compounds may be inactive for many reasons but active for only a few). This simple case, where category membership was essentially determined by indicator variables, was to test that the method could identify the important features from the noise of the other descriptors.

Table 1 details the output parameters for the 5 component analysis with the very, very high discriminating power values for the atom-pair descriptors J10F1004 and K10K1003, and the topological torsion descriptors K10C31C31K10 and TTHYD_9669 (see earlier); as one would expect from the use of indicator variables. It was clear though that the method had found the features which had been chosen to determine class membership. The values of the Loadings parameters for component 1 are in the final three columns and indicate whether or not these descriptors are beneficial for category 2 (e.g. the positive 1.78 figure for J10F1004) or category 3 membership (e.g. the positive 1.64 figure for K10K1003). Despite the success with this analysis it was disturbing that the analysis required 5 components to correctly classify the 'active'

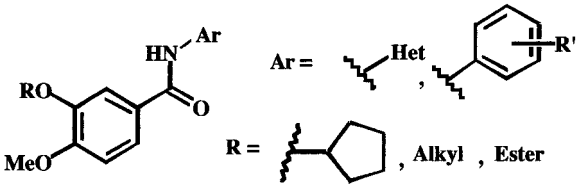


Figure 4a. General structures represented in the RPR PDE-IV data set.

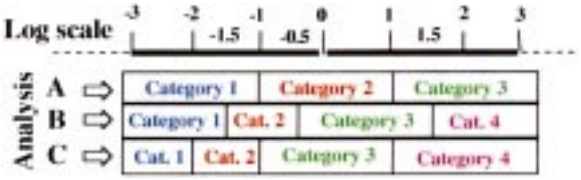


Figure 4b. RPR PDE-IV data set category boundaries.

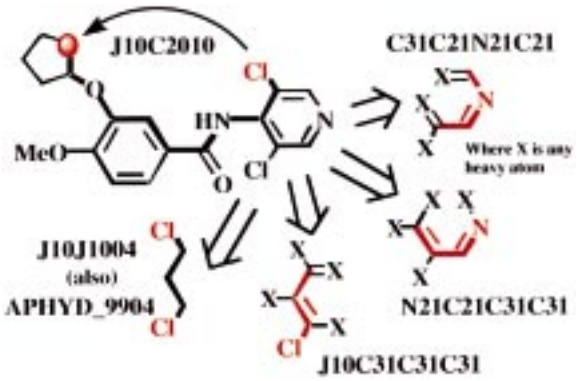


Figure 4c. The relationship between the top descriptors from Table 6 for the RPR data set and the structure of an active example from the data set.

categories where one would expect only two should be needed. Indeed if one looks at the classification results for analyses where the maximum number of components are varied (as shown in Table 2) one can see that after 3 components almost all of the compounds are classified correctly and the improvements gained with each new component are small.

Normal categorisations do not rely on a few indicator variables and so an attempt to make this designed set more realistic was to have an 'either-or' requirement for category membership. The definitions for the 'active' Category 3 and 'partially active' Category 2 are indicated in Figure 1b. The results of this second classification upon the successful categorising ability of T-SIMCA when the maximum number of components is varied are shown in Table 3a. Again one can see that only when at least 4 components are used are the 'active' categories correctly categorised and when the algorithm is allowed free rein then 25 components are required for category 1. If one takes the 4 component model and examines the 5 compounds which were misclassified then they were all placed in Category 2 and all contain either a meta chloro-bromo relationship or a meta fluoro-bromo relationship. The Loadings and Discriminating Power output are less clear cut in this model but they still highlight the atomic relationships which were chosen for category membership (Table 3b). Both of the experiments on this data set seem to produce models that require many components in order to correctly categorise all the molecules. This seems to be in accord with the results of the simple cubic data set and hence we believe, that these overly complex models are as a direct result of the fact that in T-SIMCA the principal components used to define the categories are infinite in length.

RPR data set

The literature data sets had either been part of a formal QSAR analysis [18] or had had some sort of SAR derived [17]. The RPR data set [17] contained 69 compounds with the general structure as shown in Figure 4a. The affinities of these compounds ranged from 1000 to 0.001 μM and were converted to $\log_{10}$ values (hence the most active compounds are in Category 1) before being classified into either 3 or 4 groups using the criteria shown in Figure 4b.

The results of T-SIMCA using the Analysis A categorisation are detailed in Tables 4a–c and show the effects that reducing the maximum number of components have on the categorising ability and the inter- versus intra-category distances. These distances are the RMS distance of one category's members to its own category and then to each of the other categories. What one desires is a small figure on the diagonal (the intra-category distance) and large figures off the diagonal (the inter-category distance) [20]. These distance tables give an indication of how well one category is distinguished from the others.

In Table 4a the data set is classified almost perfectly, but this is probably due to using too many components to define each category relative to the number of compounds within that category. Ideally the number of components defining a category should be less than one third the number of compounds within that category to avoid over-fitting [11]. Such overfitting has occurred with Category 3 which has the same number of components as compounds and it

Table 3a. Halobenzene data set and the effects of varying the maximum number of components allowed when forming the model in the more complex 'either-or' categorisation

Max. no. of components per category	No. correctly classified Cat. 1	No. correctly classified Cat. 2	No. correctly classified Cat. 3
1	34/40	6/11	9/13
2	34/40	7/11	12/13
3	35/40	9/11	13/13
4	35/40	11/11	13/13
5	36/40	11/11	13/13
8	39/40	11/11 (7 components)	13/13 (6 components)
all	40/40 (25 components)	11/11 (7 components)	13/13 (6 components)

Table 3b. Halobenzene data set output parameters from the more complex 'either-or' categorisation with the Loadings from component 1

Descriptor	Discrim_pwr	Desc_name	Cat. 1	Cat. 2	Cat. 3
APHYD_8903	64.01	APHYD_8903	−0.52	1.21	0.58
APREG_2239	64.01	K10J1003	−0.52	1.21	0.58
TTREG_2063	64.01	K10C31C31J10	−0.52	1.21	0.58
TTHYD_8669	64.01	TTHYD_8669	−0.52	1.21	0.58
APREG_954	22.26	K10K1004	−0.50	0.39	1.20
APREG_953	12.05	K10K1003	−0.37	−0.59	1.64
TTREG_409	12.05	K10C31C31K10	−0.37	−0.59	1.64
TTHYD_9669	12.05	TTHYD_9669	−0.37	−0.59	1.64
APREG_325	8.41	K10C3101	−0.34	0.57	0.57

can be seen in the very small intra-category distance of $1.6e^{-6}$ for Category 3. A second run (Table 4b) was carried out with a 10 component restriction for each category. The reduction in the number of available components did not affect Category 1, and all the compounds within Category 3 are still correctly categorised although the correct categorisation of Category 2 compounds suffered. When the maximum number of components was reduced to 4 (Table 4c), a high level of correct classification (circa 87% correct) was retained despite the simplicity of the new models, however the intra-category distances rose markedly indicating that the discrimination between the categories was now much worse.

The results for the first of the four category models (Analysis B) are shown in Tables 5a and b and they show a similar behaviour to the three category models.

A maximum of either 10 or 7 components (Tables 5a and 5b, respectively) was imposed and gave high to reasonable (94% to 87%) classification percentages, with good distinction between the categories. One criticism would be that the active category (Category 1) has too broad a definition of affinity to be useful, ranging as it does from 0 to 32 nM. Unfortunately as the number of compounds per category decreases it restricts the number of components which should be used to form the model for that category. The above models themselves have too many components if judged upon the desired limit of one third of the number of compounds. T-SIMCA does allow one to restrict the number of components used but this maximum component figure is a rather coarse control as it is applied to all categories equally, irrespective of how many members each category has. An independent control on the maximum number of components allowed for each category would be much more useful. The Loadings output from the 7 component model of Analysis B is shown in Table 6 and has been sorted in order of greatest contribution to the active Category 1. The beneficial contribution of the dichloro-substituted pyridyl moiety, as noted within the RPR paper [17],

Table 4a. SIMCA results for the RPR PDE-IV data set, split into 3 categories, with an unlimited number of components

Analysis A	Cat. 1	Cat. 2	Cat. 3
No. components/category	9	20	13
No. compounds	18	38	13
No. correctly classified	18	37	13
Distances between categories			
Projected Category 1	0.27	0.82	0.92
Projected Category 2	0.58	0.24	0.70
Projected Category 3	0.98	0.86	1.6e-6

Table 4b. SIMCA results for the RPR PDE-IV data set, split into 3 categories, and using a maximum number of 10 components

Analysis A	Cat. 1	Cat. 2	Cat. 3
No. components/category	9	10	10
No. compounds	18	38	13
No. correctly classified	18	34	13
Distances between categories			
Projected Category 1	0.27	0.82	0.92
Projected Category 2	0.76	0.44	0.77
Projected Category 3	0.99	0.87	0.19

Table 4c. SIMCA results for the RPR PDE-IV data set, split into 3 categories, and using a maximum number of 4 components

Analysis A	Cat. 1	Cat. 2	Cat. 3
No. components/category	4	4	4
No. compounds	18	38	13
No. correctly classified	17	31	12
Distances between categories			
Projected Category 1	0.55	0.94	1.13
Projected Category 2	0.95	0.69	1.01
Projected Category 3	1.03	0.93	0.51

Table 5a. SIMCA results for the RPR PDE-IV data set, split into 4 categories, and using a maximum number of 10 components

Analysis B	Cat. 1	Cat. 2	Cat. 3	Cat. 4
No. of components/category	7	10	10	10
No. compounds	9	19	30	11
No. correctly classified	9	18	27	11
Distances between categories				
Projected Category 1	0.15	0.90	1.03	1.14
Projected Category 2	0.57	0.30	0.67	0.94
Projected Category 3	0.80	0.67	0.41	0.82
Projected Category 4	0.97	0.96	0.89	0.08

Table 5b. SIMCA results for the RPR PDE-IV data set, split into 4 categories, and using a maximum number of 7 components

Analysis B	Cat. 1	Cat. 2	Cat. 3	Cat. 4
No. of components/category	7	7	7	7
No. compounds	9	19	30	11
No. correctly classified	9	16	24	11
Distances between categories				
Projected Category 1	0.15	0.90	1.03	1.14
Projected Category 2	0.65	0.45	0.73	1.03
Projected Category 3	0.87	0.74	0.53	0.84
Projected Category 4	0.98	0.98	0.90	0.27

can be seen in the combination of these top seven descriptors as illustrated in Figure 4c.

The second 4 category definition, Analysis C, has narrowed Category 1 and redistributed the data set more towards the inactive categories increasing the number of compounds in Categories 3 and 4. The results for this categorisation using a maximum number of 7 or 5 components are shown in Tables 7a and b respectively and show that different category boundaries or reducing the number of components can help the categorisation of some categories.

These analyses demonstrate some interesting features of T-SIMCA. As a consequence of the boundary defining Category 1 moving, 3 compounds no longer lie within Category 1 and yet the significant drop from 7 components in Analysis B to 3 components in Analysis C does not affect the perfect categorisation of Category 1. This may mean that the compounds which changed category in the redefinition were outliers in the original categorisation and now a much simpler model distinguishes the active from the less active compounds. Such outliers could be detected before analysis, by examination of the standard deviation of the category member's residuals, in W-SIMCA but such a feature is not directly available in T-SIMCA. Also the reduction in the number of allowable components from 7 to 5 in the Analysis C models, actually *improves* the categorisation of Category 4. Again we believe this effect to be due to the infinite length of the principal components, but these analyses do show that boundary changes can have a marked influence on

Table 6. SIMCA Loadings output for component 1 with the RPR data set, split into 4 categories, using the 7 component model and sorted by the active Category 1 values

Descriptor	Cat. 1	Cat. 2	Cat. 3	Cat. 4	Desc_names
APHYD_9904	1.95	0.23	−0.33	−0.17	APHYD_9904
APREG_188	1.80	0.38	−0.34	−0.21	J10J1004
TTREG_60	1.46	0.67	−0.35	−0.09	J10C31C31C31
TTREG_30	1.36	0.48	−0.38	−0.36	N21C21C31C31
APREG_1051	1.33	0.42	−0.14	−0.20	J10C2010
APHYD_1908	1.27	0.24	−0.11	−0.37	APHYD_1908
TTREG_582	1.26	0.17	−0.32	−0.45	C31C21N21C21

Table 7a. SIMCA results for the RPR PDE-IV data set, split into 4 new categories, and using a maximum number of 7 components

Analysis C	Cat. 1	Cat. 2	Cat. 3	Cat. 4
No. of components/category	3	7	7	6
No. compounds	6	12	38	13
No. correctly classified	6	11	33	12
Distances between categories				
Projected Category 1	0.35	0.98	1.18	1.29
Projected Category 2	0.55	0.31	0.85	0.97
Projected Category 3	0.89	0.77	0.54	0.87
Projected Category 4	0.99	1.00	0.89	0.38

Table 7b. SIMCA results for the RPR PDE-IV data set, split into 4 new categories, and using a maximum number of 5 components

Analysis C	Cat. 1	Cat. 2	Cat. 3	Cat. 4
No. of components/category	3	5	5	5
No. compounds	6	12	38	13
No. correctly classified	6	10	28	13
Distances between categories				
Projected Category 1	0.35	0.98	1.18	1.29
Projected Category 2	0.62	0.45	0.89	1.03
Projected Category 3	0.92	0.92	0.63	0.97
Projected Category 4	1.00	1.02	0.91	0.44

the categorising ability of the algorithm and should be experimented with during any analysis.

It should be noted that the in-house TOPOGEN routine also gives the frequency of occurrence of the descriptors within a molecule and these frequencies can be used in analyses instead of just their presence or absence. As a consequence more descriptors are usually retained after the variance filters which, in turn, leads to larger molecular spreadsheets and increased manipulation and computational times for the analysis of these data sets. The results of these analyses are not significantly different compared to the presence/absence description, although the use of descriptor frequencies did, at times, reduce the number of components needed for categorisation.

Fisons data set

The Fisons data set [18] was smaller with only 36 compounds involved in the published PLS analysis. The compounds were substituted benzoyl derivatives of 2,5-dimethyl,4-(carbomethoxy) pyrrole as shown in Figure 5a. The efficacies of these compounds (their ability to increase cardiac contractility using guinea pig atria paced at 1 Hz) were measured relative to BAY K8644 (Figure 5b) and these relative measures were converted to a log scale so that a large positive log value (i.e. the higher categories) correspond to the most active compounds. The compounds were categorised into 2, 3 or 4 categories using the criteria indicated in Figure 6.

The Analysis A results are summarised in Table 8 showing almost perfect categorisation of the data set using 6 components.

Such an analysis would be ideally suited to database searching to find new leads and in this manner T-SIMCA is very similar to our existing TrendVector algorithm [21]. The three category analysis (B) is shown in Table 9 and it produces perfect categorisation for the most active category using only 4 components whilst the least active was less well categorised despite having a slightly higher compound to component ratio.

Higher success rates were achieved with Category 2 when the number of components was increased,

Table 8. SIMCA results for the Fisons data set, split into 2 categories, and using a maximum number of 6 components

Analysis A	Cat. 1	Cat. 2
No. of components/category	6	6
No. of compounds	19	17
No. correctly classified	18	17
Distances between categories		
Projected Category 1	0.51	0.76
Projected Category 2	1.04	0.51

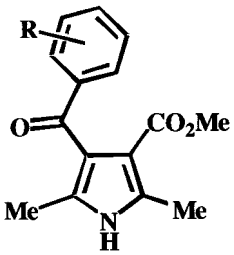


Figure 5a. General structure for the Fisons Calcium channel data set.

but with there being only a limited number of compounds, the predictive use of such an analysis would be much reduced. A similar comment could be leveled at Analysis C which has more categories, the results of which are detailed in Table 10.

One of the most active, Category 4, compounds is incorrectly classified despite the high number of components. This is due to the use of our calculated LogP values as a descriptor, when these are replaced by the measured LogP values quoted within the paper [18] then the categorisation is perfect. The discrepancies in the LogP values occur with the more lipophilic compounds, which are generally the most active, with our calculated values being underestimates of the measured values. Another feature shown to be of importance in their paper was that of Molar

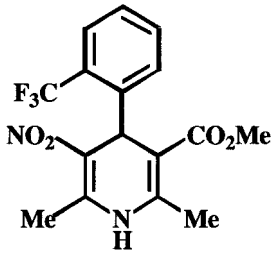


Figure 5b. BAY K8644.

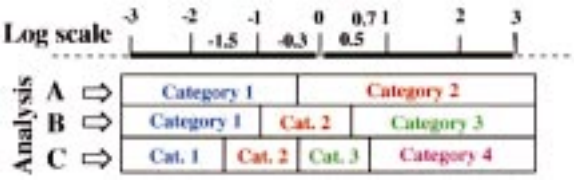


Figure 6. Fisons data set category boundaries.

Table 9. SIMCA results for the Fisons data set, split into 3 categories, and using a maximum number of 4 components

Analysis B	Cat. 1	Cat. 2	Cat. 3
No. of components/category	4	4	4
No. of compounds	8	14	14
No. correctly classified	6	11	14
Distances between categories			
Projected Category 1	0.51	0.86	0.83
Projected Category 2	0.82	0.58	0.86
Projected Category 3	1.04	1.04	0.47

Table 10. SIMCA results for the Fisons data set, split into 4 categories, and using a maximum number of 10 components

Analysis C	Cat. 1	Cat. 2	Cat. 3	Cat. 4
No. of components/category	4	10	5	5
No. of compounds	4	15	8	9
No. correctly classified	4	15	8	8
Distances between categories				
Projected Category 1	9.7 e−8	0.94	0.95	0.9
Projected Category 2	0.69	0.23	0.68	0.70
Projected Category 3	1.03	1.07	0.34	0.81
Projected Category 4	1.07	1.12	0.95	0.34

Refractivity (MR), a measure of molecular volume. It was interesting to note that the use of the total number of either Atom pair or Topological torsion descriptors was an equally good measure of molecular volume and correlated well with the calculated MR figures given in the paper.

NMDA data set

The final example concerns the large data set of 405 compounds from our in-house search for an NMDA glycine site antagonist [19]. The set was based upon the structures of L1 and L2 (shown in Figure 2) and T-SIMCA was used to identify, if any, the features required for good activity. The log of the affinity data was used and categorised into 5 or 10 categories (as

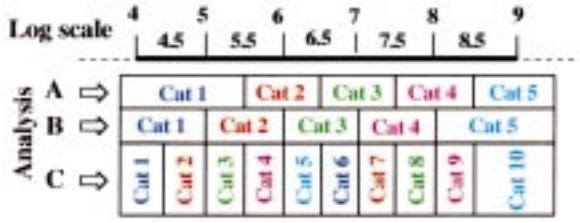


Figure 7. NMDA data set category boundaries.

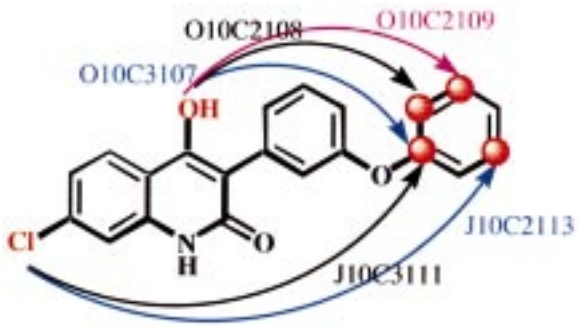


Figure 8. Visualisation of some of the important atom-pair descriptors in the NMDA model.

shown in Figure 7) and so Category 5 (or 10) contained the most active compounds. The division of the data set into ten categories meant that compound affinities to a precision of 0.5 log units could be estimated, comparable perhaps to accuracies obtained in CoMFA/PLS analyses.

The analyses were run either with the whole data set of 405 compounds or with a 5% test set of 21 randomly chosen compounds and a 384 compound training set. The categorisations of the 405 and 384 compound training sets are detailed in Tables 11a–d. As has been seen earlier the positions of the category boundaries can affect the categorisation ability of a model, hence the staggered categorisations were run to see if significantly different success rates were obtained. The problem caused by the presence of borderline compounds in the test set might be overcome if one examined the predictions of a combination of the two staggered analyses rather than relying on one alone.

Despite having more components per compound in each category the 10 category model, Analysis C, classified 4 fewer compounds correctly overall than the 5 category model Analysis B (Tables 11b and 11a, respectively). The categorisation success rate of almost 78% for the Analysis B 10 component model, only rose to 87% using 20 components with the extra 10 components mostly helping to define the compounds in the more populated, middle categories (Table 11a). It was interesting that the staggering of the categories this time did not alter the categorising ability of T-SIMCA (∼90% success) with only 2 compounds more being correctly categorised in the Categorisation **A** 384 compound training set (Table 11c) as compared to the Categorisation **B** 384 compound set (Table 11d).

The results for the 21 compound test set on the two models in Tables 11c and 11d were that, individually, the Categorisation **A** model categorised 9 correctly, 11 one category off and only 1 two categories off; the Categorisation **B** model categorised 10 correctly, 9 one category off and 2 two categories off. Chi-squared values for these results show that there is less than a 2% chance that this was a random outcome. Taking the Categorisation **A** and **B** models together 13 were correctly predicted (6 of those were categorised correctly by both) and 8 compounds one category off. These results are very similar to those obtained in predictions based upon a 10 category model with 6 compounds being correctly predicted, 12 being 1 or 2 categories off and the remaining 3 being 3, 4, or 5 categories off respectively. The combination of the two 5 category models seems to perform slightly better than the single 10 category model but obviously the category ranges are much larger with the 5 category model. With the combination approach one can identify those predictions which are likely to be dubious by seeing where the discrepancies in prediction occur. With a single model one has to rely on interpreting the distances from the test point to each category and estimate from them the possible category error in the original prediction. However both analyses show that the models can provide useful predictions of affinity and in almost a third of the compounds the predicted affinity was correct to within 0.5 log units.

Analysing the Loadings table (Table 12) for Component 1 in the 5 category model shown in Table 11a one can see the importance of the aromatic moiety in being 7 bonds away from the OH of the quinolin-2-one. Also it highlights that the pendant phenyl should be unsubstituted 8 or 9 bonds away. The chlorine substituent is highlighted to be important relative to the pendant aromatic and our CLogP is seen to be important for activity and so on. These relationships are shown more easily on the original search structure L1 in Figure 8. However there are many loadings figures to be considered and it would be impossible to display them all in such a representation.

Hence we have written code to apply the Loadings figures for a category to the heavy atoms of a probe

Table 11a. SIMCA correct classification results for the whole NMDA data set, split into 5 categories, and using a maximum number of either 10 or 20 components

Analysis B	Cat. 1	Cat. 2	Cat. 3	Cat. 4	Cat. 5
No. of compounds (405 in total)	58	66	103	123	55
No. correctly classified 10 comp. max	51	51	73	88	52
No. correctly classified 20 comp. max	52 (19 components)	58	86	104	54

Table 11b. SIMCA correct classification results for the whole NMDA data set, split into 10 categories, and using a maximum number of 10 components

Analysis C	Cat. 1	Cat. 2	Cat. 3	Cat. 4	Cat. 5	Cat. 6	Cat. 7	Cat. 8	Cat. 9	Cat. 10
No. of compounds (405 in total)	37	21	24	42	44	59	65	58	41	14
No. correctly classified 10 comp. max	31	18 (9 components)	18 (8 components)	32	37	42	42	42	37	12 (5 components)

Table 11c. SIMCA correct classification results for the training NMDA data set, split into 5 categories, and using a maximum number of 20 components

Analysis B	Cat. 1	Cat. 2	Cat. 3	Cat. 4	Cat. 5
No. of compounds (384 in total)	55	63	97	117	52
No. correctly classified 20 comp. max	52	58	83	101	51

Table 11d. SIMCA correct classification results for the training NMDA data set, split into 5 displaced categories, and using a maximum number of 20 components

Analysis A	Cat. 1	Cat. 2	Cat. 3	Cat. 4	Cat. 5
No. of compounds (384 in total)	78	81	117	95	13
No. correctly classified 20 comp. max	73	76	102	90	6 (4 components)

molecule to give some indication as to which atoms could be sites for change as these atoms have more 'bad' descriptors than 'good' ones. An example of this is with the pair of compounds L3 (pKi = 8.59) and L4 (pKi = 6.39) shown in Figure 9. Here the Loadings for each category from a five category model (Table 11a) have been applied and the atoms have been coloured by the colour scheme shown. The 'warmer' colours of orange, red or black denote that that atom is present in more descriptors which are beneficial for that compound's membership of that category and the converse for the 'cooler' colours of cyan, blue and purple. One can see from the Category 5 images that more of the L3 structure is black compared to L4 where the cooler colours for the nitrogens of the pyrazine ring show that they do not contribute well to the high affinity

Table 12. SIMCA Loadings output for component 1 with the NMDA data set, split into 5 categories, sorted by the active Category 5 values

Descriptor	Cat. 1	Cat. 2	Cat. 3	Cat. 4	Cat. 5	Desc_name
APREG_610	−0.37	−0.40	−0.27	0.11	1.12	O10C3107
APREG_622	−0.26	−0.47	−0.34	0.17	1.10	O10C2109
APREG_377	−0.33	−0.40	−0.25	0.10	1.09	O10C2108
APREG_4815	−0.32	−0.34	−0.40	0.19	1.06	J10C2113
APREG_1143	−0.31	−0.32	−0.33	0.13	1.05	J10C3111
APPC_1509	−0.45	−0.40	−0.33	0.26	0.98	APPC_1509
US_CLOGP	−0.31	−0.53	−0.29	0.24	0.98	US_CLOGP

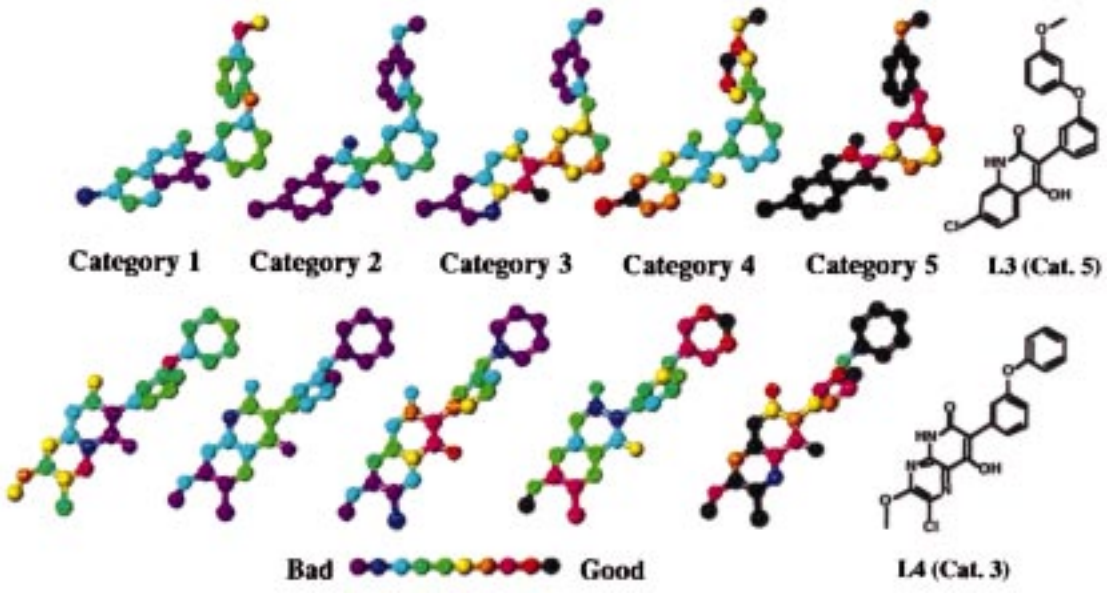


Figure 9. Visualisation of the effect of all the atom-pair and topological torsion descriptors used in the NMDA model on two similar compounds with differing activity.

category. Indeed if one examines the colouration of the Category 1 structures one can see that those nitrogens in L4 are favourable for membership of the non-affine category. The model predicts L3 correctly as Category 5 but misclassifies L4 as Category 4 rather than Category 3. The model does, nevertheless, correctly predict lower activity for L4 compared to L3 and examination of Figure 9 directly implicates one of the nitrogens as a reason for the loss of activity.

The results from all these analyses show that T-SIMCA coupled with our 2D descriptor set is a versatile QSAR tool, but as has been mentioned, a number of discrepancies in T-SIMCA versus W-SIMCA [2] have been discovered.

The lack of restrictions upon the length of the principal components in T-SIMCA seems to produce models which are more complex than need be and it also removes the useful ‘none of the above’ outcome for a test compound that the constraining of a category’s hypervolume allows. Such ‘none of the above’ results are of importance though as one would like to know that a prediction made on such a compound was likely to be inaccurate as the training set has not described the test compound well enough. It would also be of use if one wanted to deliberately explore new regions of descriptor space. The exact restriction on the length of the principal components is a purely arbitrary limit and if one had control over that limit then one could make the category definitions as broad or as narrow as one liked. This would be of interest to see just how vague the definitions of any or all categories needed to be to obtain a useful model. A compound

belonging to two categories simultaneously is also not possible in T-SIMCA as a compound is assigned to the closest category only. For the categorisation of affinity that may seem logical but allowing such a feature may help deal with borderline compounds better than moving the boundaries and rerunning the analysis.

In an attempt to reproduce the T-SIMCA results outside the SYBYL implementation we used TRIPOS' own principal components analysis on the cubic test data set and measured distances from the data points to these components. If one examines the Loadings values for the SIMCA model using only 1 component and compare them to the vectors for the data means then one sees that some normalisation of the data to unit variance has occurred and the 'distances' quoted may have therefore also been normalised. As the exact method for producing these distances is unknown to us, this normalisation may be the reason why the output distances were not the same as those calculated by us. The other test analyses had the categories defined by different numbers of components and therefore again without knowing how the distances are computed in T-SIMCA it is very difficult to compare the 'distances' quoted in the varying number of dimensions.

Conclusions

In this paper we have shown that the TRIPOS implementation of SIMCA (T-SIMCA) can be used to analyse 2D data derived from our in-house routines in a useful manner suitable for the production of structure activity relationships. We believe that the coupling of our 2D descriptors with the classification analysis methodology exemplified by T-SIMCA, is and will be, of great use in the more difficult areas of drug development or wherever non-linear relationships or multiple possible alignments make standard analysis methods unproductive. We have been limited to the use of SIMCA in the form which TRIPOS have implemented the code and have found that this implementation is not as suitable as it could be. However we hope that this work and other 2D based QSAR work using such methods as HQSAR [22] (Holographic QSAR, a 2D/PLS methodology from TRIPOS) would encourage the development of better classification algorithms or implementations.

Acknowledgements

The author would like to thank Dr H. Broughton, Dr M. Miller and Dr B. Bush for their advice and comments during this work and Prof. S. Wold for his views on categorical analysis.

References

1. Cramer III, R.D., Patterson, D. and Bunce, J., J. Am. Chem. Soc., 110 (1988) 5959.
2. Wold, S. and Sjöström, M., ACS Symp. Ser., 52 (Chemometrics: Theory Appl., Symp.) (1977) 243.
3. SIMCA-4000 available from Umetri, Box 7960, S-907 19 UMEÅ, Sweden. http://www.umetri.com
4. Ishiyama, J.-I., Miyashita, Y. and Sasaki, S., In Doyama, M., Kihara, J., Tanaka, M. and Yamamoto, R. (Eds), Comput. Aided Innovation New Mater. 2, Proc. Int. Conf. Exhib. Comput. Appl. Mater. Mol. Sci. Eng., 2nd (1993) Meeting Date 1992, Issue Pt. 2, Elsevier, Amsterdam, pp. 891–894.
5. Van Der Voet, H. and Doornbos, D. A., Anal. Chim. Acta, 161 (1984) 125.
6. Leonards, P.E.G., van Hattum, B., Cofino, W.P. and Brinkman, U.A.T., Environ. Toxicol. Chem., 13 (1994) 129.
7. Dunn, W.J., Wold, S.J. and Martin, Y.C., J. Med. Chem., 21 (1978) 922.
8. Henry, D.R. and Craig, A.M., ACS Symp. Ser., 413 (Probing Bioact. Mech..) (1989) 70.
9. Miyashita, Y., Takahashi, Y., Takayama, C., Sumi, K., Nakatsuka, K.,Ohkubo, T., Abe, H. and Sasaki, S., J. Med. Chem., 29 (1986) 906.
10. Dunn, W.J. and Wold, S., J. Chem. Inf. Comput. Sci., 21 (1981) 8.
11. Dunn, W.J. and Wold, S., Bioorg. Chem., 9 (1980) 505.
12. Verloop, A., Hoogenstraaten, W. and Tipker, J., Med. Chem., 11 (1976) 165.
13. Kearsley, S.K., Sallamack, S., Fluder, E.M., Andose, J.D., Mosley, R.T. and Sheridan, R.P., J. Chem. Inf. Comput. Sci., 36 (1996) 118.
14. SYBYL Molecular Modelling package available through TRIPOS Inc., 1699 S. Hanley Road, St. Louis, MO, USA. http://www.tripos.com
15. Van Der Voet, H., Coenegracht, P.M.J. and Hemel, J.B., Anal. Chim. Acta, 209 (1988) 1.
16. Frank, I.E. and Lanteri, S., Chemom. Intell. Lab. Systems, 5 (1989) 247.
17. Ashton, M.J., Cook, D.C., Fenton, G., Karlsson, J.-A., Palfreyman, M.N., Raeburn, D., Ratcliffe, A.J., Souness, J.E., Thurairatnam, S. and Vicker, N., J. Med. Chem., 37 (1994) 1696.
18. Davis, A.M., Gensmantel, N.P., Johansson, E. and Marriott, D., J. Med. Chem., 37 (1994) 963.
19. Carling, R.W., Leeson, P.D., Moore, K.W., Moyes, C.R., Duncton, M., Hudson, M.L., Baker, R., Foster, A.C., Grimwood, S., Kemp, J.A., Marshall, G.R., Tricklebank, M.D. and Saywell, K.L., J. Med. Chem., 40 (1997) 754.
20. SYBYL Ligand Based Design Manual, version 6.2, 242.
21. Sheridan, R.P., Nachbar, R.B. and Bush, B.L., J. Comput.-Aided Mol. Des., 8 (1994) 323.
22. HQSAR – available through TRIPOS Inc., 1699 S. Hanley Road, St. Louis, MO, USA. http://www.tripos.com.