ELSEVIER

Analytica Chimica Acta 400 (1999) 413–424

ANALYTICA CHIMICA ACTA

www.elsevier.com/locate/aca

Molecular challenges in modern chemometrics

R. Wehrens [a,*], R. de Gelder [b], G.J. Kemperman [c], B. Zwanenburg [c], L.M.C. Buydens [a]

[a] *Department of Analytical Chemistry, NSR/RIM Research Centre, University of Nijmegen, Toernooiveld 1, 6525 ED Nijmegen, The Netherlands*

[b] *Department of Inorganic Chemistry, NSR/RIM Research Centre, University of Nijmegen, Toernooiveld 1, 6525 ED Nijmegen, The Netherlands*

[c] *Department of Organic Chemistry, NSR/RIM Research Centre, University of Nijmegen, Toernooiveld 1, 6525 ED Nijmegen, The Netherlands*

Received 3 June 1999; accepted 21 July 1999

Abstract

Since the very beginning of the discipline, chemometrics has mainly focussed on analytical chemical problems such as calibration. With the growing importance of databases and applications in medicinal and computational chemistry, the domains of analytical chemistry and chemometrics have been enlarged significantly in recent years. Especially the relation between molecular structure and function has become of considerable interest. Despite the huge quantities of data that are available nowadays, it is often difficult to recognise and extract relevant chemical information for the problem at hand. One of the main obstacles is the definition of an appropriate representation of a molecule. Although a variety of different representations are used, none are generally applicable.

This paper focuses on the challenges that arise in the chemometrical analysis of molecular structures, the relation between structure and function and the relation between molecular representation and chemometrical modelling. Exciting opportunities for further research are illustrated using an example concerning the prediction of co-crystallisation behaviour for small organic molecules with cephalosporin antibiotics.

Keywords: Molecular structure; QSAR; Cephalosporins; Co-crystallisation

1. Introduction

The main goals in analytical chemistry are accurate identification and quantification, and in the nearly three decades of the existence of the field of chemometrics, research mainly has concentrated on these areas. For quantitative analysis multivariate calibration methods have become indispensable, especially as a result of the large amounts of data generated by modern instruments. Qualitative analysis, mainly in the context of identification, is performed using mainly clustering and classification methods. In the last few years the set of analytical questions, quantitative (‘how much?’) and qualitative (‘what?’) is extended with questions about localisation (‘where?’) and molecular structure (‘in what form?’). Examples of the localised problems can be found in surface analysis methods, where a location-specific quantification or classification is required. This necessitates the use of image processing techniques as well as chemomet-

* Corresponding author. Tel.: +31-24-365-2053; fax: +31-24-365-2653

E-mail address: rwehrens@sci.kun.nl (R. Wehrens)

0003-2670/99/$ – see front matter
PII: S0003-2670(99)00621-2

rical methods [1]. Questions on molecular structure are the subject of the present paper and will be referred to as 'molecular problems'. For these problems, it is not only important to know which and how many molecules are present in the sample, but knowledge about their conformation(s) will be required, too, to answer the questions posed. A lot of research has been devoted to the relation between molecular structure and function, especially in the fields of pharmaceutical chemistry and quantitative structure–activity relationships (QSAR). Techniques that are often used in these fields include energy minimisation and docking procedures, and linear regression techniques to predict behaviour from structure and physico-chemical properties. In the last decade the interest in molecular problems has increased enormously, and the huge potential of chemometrical methods in this respect has already been acknowledged [2].

Three main driving forces can be identified for the recent increase in attention. First of all, molecular structure can be determined nowadays both in crystalline form and in solution, even for quite large structures such as proteins. Especially high-resolution multidimensional NMR techniques are important in this respect. Second, computational techniques are routinely being employed to assess the quality of experimental structures and to refine experimental results. This generally leads to huge amounts of data. In order to be able to interpret these data, the information must be condensed, e.g., by means of a cluster analysis [3,4]. Finally, more and more experimental and computational results are available to the scientific community in the form of huge data bases, accessible through the internet. The information potential of these data bases is almost unlimited, but again the question arises how to make proper use of the data. In data-mining, simple uni- and bivariate statistics are often used to formulate conclusions on ensembles of structures (see, e.g., ref. [5,6]), but clearly multivariate techniques can yield much more information [7,8].

Standard chemometrical methods cannot always be applied per se, however. In conventional chemometrics the variance–covariance matrix indicating differences and similarities between objects and variables in the data set is the starting point of many analyses. In molecular problems, there is no clear way to define such a matrix and the question is whether it is possible to define chemical similarity in a meaningful way at all. In the next section this question and the related issue of representation of chemical structures will be addressed. Problems with the current state of affairs will be illustrated using an example in which complexation behaviour of small organic compounds is predicted. The paper concludes with a discussion of opportunities for the wider application of chemometrical techniques, and identifies problems and directions for future research.

2. Chemical similarity

In order to apply chemometrical techniques to molecular problems, the concept of chemical similarity takes the place of conventional distance measures such as Euclidean or Mahalanobis distances. However, chemical similarity is not a unique and well-defined concept. In many cases, one would like to relate similarity in structure to similarity in chemical properties or behaviour. However, in the example given later, several compounds with only minor structural differences show quite different properties. Clearly, in many cases only a small part of a molecule is responsible for properties such as biological activity, and whether or not similarity measures can reflect is of crucial importance for database searches to find new, biologically active compounds [9]. Not surprisingly, most successful QSAR applications have focussed on molecules of roughly equal size, preferably with a common skeleton. The structural difference or similarity between a protein and a small organic molecule is hardly relevant.

Calculation of adequate similarity measures for molecules depends on the representation of these structures [10]. For trained chemists a very concise representation like a 2D structure or even a line notation is sufficient to infer chemical properties such as functional groups and partial charges. For a computer program, there is not yet a way to automatically derive the desired parameters from a simple 2D graph, so they must be provided explicitly. Unfortunately, for many applications it is not clear which parameters are needed, and in what form they should be cast. How should one represent molecular shape, for instance? Many representations are in use today, for a large number of different applications [11]. Examples of comparisons of various descriptors and prediction

Table 1
Several molecular representations and their attributes[a]

Representation	Local versus global	General versus specific	Relative versus absolute	Variable versus fixed size
Molecular properties	G	G	A	F
Similarities	G	G	R	F
Atomic coordinates	L	G	A	V
Torsion angles	L	S	A	V
CoMFA	L	G	R	F

[a] Molecular properties: dipole, polarisability, etc. Measured or synthetic spectrum-like descriptors fall in this category as well. Atomic coordinates: Cartesian coordinates for each atom (sometimes excluding H). Torsion angles: using standard values for bond lengths and bond angles, centre of mass at origin with a random orientation. Substituent lists fall in the same category. CoMFA: methods using a grid at which properties of interest are sampled. The size of the grid is determined by the largest molecule in the set.

methods can be found in the work of Brown and Martin [12,13]. Since each representation captures a different part of the description of the molecule, it is often not possible to go from one representation to another. Several representations can be distinguished according to a number of criteria (Table 1).

2.1. Local versus global approaches

Global descriptors describe properties of molecules as a whole. Sometimes they are expressed by one number, e.g., the dipole moment, charge, or the Wiener topological index, sometimes by a vector or a matrix. Examples of the latter type are distance matrices, giving the distance from each atom of the structure to another, experimental spectra such as IR spectra, or synthetic spectra. These synthetic spectra may emulate measured spectra or may capture completely different information, such as the mass distribution relative to the centre of mass. In these global descriptors, all local information, such as the vicinity of groups or partial charges, is lost. Therefore, these descriptors usually have a weak performance when trying to predict molecular properties such as biological activity. As outlined above, often only a part of the molecule is responsible for the specific processes taking place.

Local descriptors, on the other hand, contain information on properties at different locations within the molecule. This makes it possible to concentrate on, e.g., the active site of a molecule only, and therefore, local descriptors are much more suited for QSAR modelling. Often, approaches such as comparative molecular field analysis (CoMFA) are used (see [14] for an overview of CoMFA and related techniques) in which properties like electrostatic potential are evaluated in a three-dimensional grid around the molecules of interest. The values at the grid points are then used to predict the property of interest, usually by partial least squares (PLS) [15].

2.2. Relative versus absolute descriptors

Relative descriptors are based on some kind of alignment of two molecules being compared. An example is the amount of overlap for several molecules attached to the same receptor site (the similarity of the molecules with respect to shape or charge distribution), or the RMS deviation relative to a crystal structure. The alignment may be done with respect to each other but also with respect to external factors, e.g., the structure of a binding site, that are not part of the analysis themselves. Usually, though, the alignment is performed by identifying key atoms that should be in the same position, (e.g., for a data set with a common skeleton), or by overlaying centres of mass and subsequent rotation. A wrong alignment may lead to severe errors. The reason why relative measures are very popular is that they can often be related directly to a property such as biological activity. Absolute descriptors, such as atomic coordinates or internal distance atomic matrices may be able to show similarities and dissimilarities in a set of molecules, but often fail to pinpoint the relevant ones.

2.3. General versus specific descriptors

Application-specific descriptors are able to define the structure of a molecule in a very concise way.

Examples are torsion angles defining the structure of the backbone in nucleic acids and proteins, or substituent lists in a homologous series of compounds. To what extent the three-dimensional structure can be coded depends on the descriptors and the rigidity of the molecules. The price that has to be paid for using specific descriptors is, of course, generality. General descriptors such as Cartesian coordinates can describe any chemical structure whereas specific operators cannot. However, in most QSAR applications the usefulness of predictive models is limited to a homologous class of molecules, and the lack of generality is not necessarily a disadvantage. In most cases where a more diverse set of molecules is investigated, specific descriptors are useless.

2.4. Static versus dynamic descriptors

It is evident that molecules are not static entities, and indeed in some cases the dynamics of the molecular conformations are crucial for the function of the molecule. No concise representations exist that give a dynamic account of molecular structure. If dynamic behaviour is important, one usually performs the analysis a number of times with different snapshots of the conformational distribution. Although this is far from ideal, it has the advantage that the method can be applied for most of the descriptors currently available. One of the disadvantages is that the relevant conformation can easily be overshadowed by a large number of other, irrelevant conformations.

2.5. Fixed-size and variable-size descriptors

For many types of analysis it is important that all structures under investigation are represented by the same number of variables. This means that variable-size descriptors such as atomic coordinates, where the number of parameters increases with larger molecules, are not useful. Examples of fixed-size descriptors are spectrum-like descriptors, and CoMFA-like descriptors. An overview of the usefulness of several often used representations can be found in [16], where it was concluded that no single representation was best overall in a variety of applications.

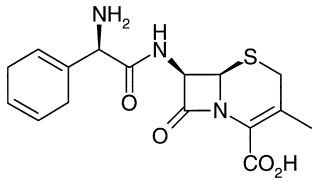

Fig. 1. The structure of cephradine.

3. Predicting complexation behaviour

To illustrate the problems that can be encountered when applying chemometrical methods to molecular problems, we here present an example in which the aim is to predict whether or not co-crystallisation will occur between two compounds. In cases where large-scale purification and isolation of a compound is difficult, e.g., because of instability under certain conditions, crystallisation or co-crystallisation often is a convenient alternative. In this case, the compound of interest is the antibiotic cephradine (see Fig. 1). The other compound, the so-called complexing agent, is used to isolate the cephradine from the reaction mixture by forming a (micro)crystalline material incorporating the antibiotic. In the solid state, the complexing agent is present in cavities formed by the cephradine host molecules [17]. It is known that a range of compounds can act as suitable agents [18]. Naturally, it is important to find compounds with an optimal complexation efficiency, also satisfying boundary conditions such as a low price and low toxicity. Ideally, a large database of candidate molecules is evaluated using a chemometrical model to identify new complexing agents. Due to the large number of candidates and the sometimes capricious behaviour of seemingly similar compounds, chemometrical models offer potential advantages over human chemical intuition.

Earlier investigations were performed using molecular modelling to identify compounds that would fit in the cavity created by a related host molecule, the antibiotic cefadroxil. The shape of the cavity was obtained by removing a guest molecule from the crystal structure of a complex. Next, 1000 compounds were generated that showed a good fit to the cavity. Although the original guest molecule was found too, none of a random selection of 50 of these compounds showed any form of complexation. Cephradine forms an even more difficult case than cefadroxil because

several different crystal structure types were found for various complexing agents. This feature distinguishes the problem from the more common situation in which active molecules should be identified in a data base that interact with some receptor.

Since the cephradine molecule is a constant factor, the representation of the candidate compounds is of crucial importance here, and one of the main results should be which representation, if any, contains the most relevant information for the prediction of complexation behaviour. This question will be tackled using both unsupervised and supervised methods. The former consists of principal component analysis (PCA) [19] and hierarchical clustering methods [20]; classification methods such as linear discriminant analysis (LDA) and K-nearest-neighbour (KNN) classification [15] are used for the latter.

3.1. Experimental

A data set of 99 small organic compounds was constructed and the occurrence of co-crystallisation with the antibiotic cephradine was investigated experimentally. Due to the specific shape of the cavities present in the cephradine host skeleton, all compounds consisted of relatively flat aromatic structures with one, two or three substituted rings. Fifty-five of the molecules were found to form complexes with cephradine. A few examples of compounds in the data set are depicted in Fig. 2.

Two types of variables, calculated with the semi-empirical program Tsar [21], are used to describe the organic compounds. First of all, a set of 19 physico-chemical variables such as total dipole, dipole moments along the three principal axes, molecular surface area and molecular volume is used to capture information about shape and other global molecular characteristics. This data matrix will be indicated as Xparam. The second type of data set consists of similarities. Such a relative representation (cf. Table 1) may be able to capture implicitly information relevant to whether or not complexation will occur. The first step in the calculation of these similarities is the alignment of the molecules with respect to shape. Centres of mass are overlayed and a full rigid search is performed using rotation angle–angle increments of 18°. A simplex optimisation is used to fine-tune the optimal alignment. Next, the similarity of the two molecules with respect to shape, charge, refractivity and lipophilicity can be calculated. After some initial experimentation it was decided to generate similarity matrices for shape and charge distribution. Two matrices of dimensions 99×99 called Xshape and Xcharge, respectively, are obtained in this way. Alignment on other parameters than the shape of the molecules in most cases led to substantially worse models, and in no case to better ones. This lends support to the hypothesis that for a successful fit of the guest molecule, shape and charge distribution should simultaneously match the cavity in the host molecule. Settings of Tsar remained fixed during all experiments.

Often, models with too many variables lead to bad predictions because of overfitting. Especially with the similarity data this is a real danger since the number of variables equals the number of objects. Two strategies for decreasing the number of variables (columns) in the similarity matrices are applied here. The first is based on chemical considerations and is meant to build models for more homogeneous subsets of compounds. The second uses a genetic algorithm to find the set of variables that leads to a minimal prediction error. The subsets that are obtained can be viewed as 'reference' compounds. Such a set, if it can be found, offers several other advantages. First of all, for new untested compounds, only a few similarities have to be calculated to be able to predict whether or not complexes will be formed. Furthermore, it is easier to interpret the similarities in such a small set. This should make it possible to formulate design constraints, that new untested compounds should satisfy.

Models are validated with a set of 20 additional, experimentally tested compounds, not used in the model building. For each of these test compounds the 19 physico-chemical variables as well as the similarities with the 99 training set compounds (both Xshape and Xcharge) were calculated in exactly the same way as described before.

All analyses were performed on each of the three individual data matrices as well as on the combined data matrix (Xall, dimension 99×217). Statistical calculations were performed using R [1] version 0.63 [22] on a Linux Pentium II machine (266 MHz).

[1] The main R site is http://www.ci.tuwien.ac.at/R/contents.html.

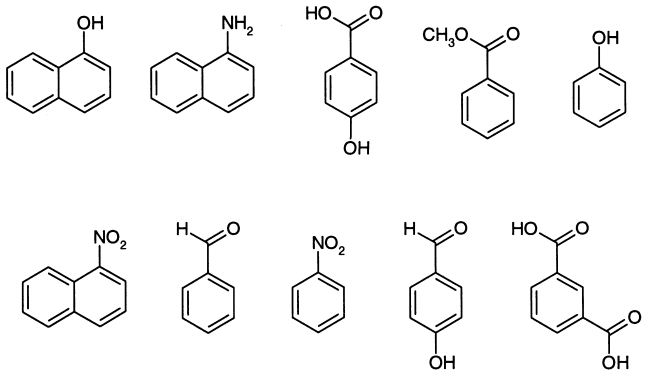

Fig. 2. Examples of compounds in the data set. The compounds in the top row form complexes with cephradine, compounds in the bottom row do not.

Genetic algorithms for reference compound selection were performed on SUN workstations using the PGA Pack library. [2]

3.2. Results

3.2.1. Data validation

The first step in the analysis of the molecular data consists of a validation of the data. Data, whether they are obtained experimental, from a database or by calculations, are never completely error-free, and, unfortunately, seldomly validated afterwards (an exception can be found in, e.g., ref. [23]). It is important to identify objects that do not conform to the general trend in the data, especially since they may have a large (and usually disturbing) influence in the modelling phase. Most multivariate outlier detection methods obtain a robust estimate of mean and covariance matrix from a subset of the data, and use these to calculate Mahalanobis distances for all objects in the data set [24,25]. Since the Mahalanobis distance can not be computed from the molecular similarities, multivariate outlier detection methods were only applied to the Xparam data set. Three methods were used: Rousseeuw's MCD [24] and two methods proposed by Egan and Morgan, SHV and RHM [25]. Several outliers were identified by these methods, most of them because of energy-minimised conformations that were not flat. This is a direct result of the composition of the data set, where attention was focussed in the first place on compounds fitting in the cavity. In all cases, the outlying observations did not form complexes with cephradine and as such provide valuable negative examples for statistical models. Since the similarities with the other compounds in the set, as measured by row or column means in Xcharge and Xshape, were not different from the ones of the outlying observations, it was decided to retain all observations in the data set.

Although a direct validation of matrices Xshape and Xcharge is not possible, one would expect them to be symmetrical around the diagonal. Remarkably, notable differences exist, indicating that aligning molecule A with B does not always yield the same results as aligning B with A. In one instance in Xcharge, the difference was found to be as large as 0.67 on a scale from -1 to 1. Similarities from the lower triangle of Xcharge are plotted against corresponding similarities in the upper triangle in Fig. 3. The largest deviations are found in regions of low similarity. This could indicate that two non-resembling molecules can be aligned in several ways, with similar overlap with respect to shape, but with different charge distribution overlap.

In favour of this hypothesis is the fact that the Xshape matrix is much more symmetric, with a mean difference between the upper and lower triangles of 0.006 and a maximum difference of 0.09. Moreover, calculating the shape similarity after aligning on charge distributions also shows significant differences between upper and lower triangles of the similarity matrix, clearly indicating that the shape and charge spaces are not congruent.

[2] This library is available from ftp://ftp.mcs.ano.gov/pub/pgapack/pgapack.tar.Z.

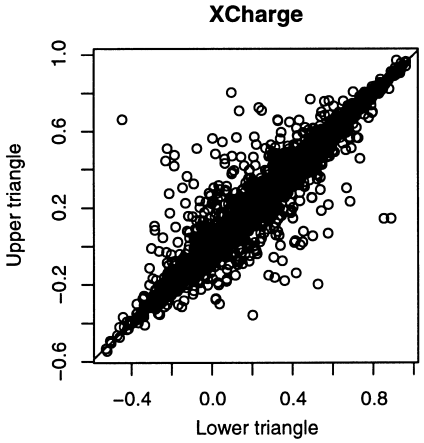

Fig. 3. Asymmetry in the similarity matrix Xcharge. Especially with very high or very low similarities, the order of the alignment does not seem very important. With non-similar compounds, aligning A on B may give different similarities than aligning B on A.

Although the data calculated by Tsar appear to have some degree of error, a manual validation showed that molecules perceived as being very similar from a chemical point of view indeed show high similarities.

3.2.2. Unsupervised methods

The next step is to apply unsupervised methods like PCA and hierarchical clustering. The results of the PCA on the three data matrices and the combined data matrix are depicted in Fig. 4, where the scores on the first two PC's are plotted. Prior to PCA, all matrices are scaled to zero mean and unit variance. Clearly, the complexing and non-complexing compounds cannot be distinguished in this way. Applying other preprocessing techniques yields similar results.

A similar result is obtained with clustering. Again, the data matrix Xparam is scaled to zero mean and unit variance and Euclidean distances are calculated; Xshape and Xcharge are transformed to dissimilarities by the equation

$$DS_{ij} = S_{ii} + S_{jj} - S_{ij} - S_{ji} = 2 - S_{ij} - S_{ji}$$

These dissimilarities are then used as distances in the clustering algorithm. An added advantage is that the dissimilarity matrices are symmetrical.

The cluster tree obtained using Ward's method on the Xparam data set is shown in Fig. 5. The two main branches of the cluster tree both contain complexing as well as non-complexing compounds. Ward's clustering was found to perform quite well in other structure–activity applications [12], but several other clustering methods such as average and complete linkage show similar plots. Clusterings on Xcharge and Xshape confirmed that no easy discrimination between complexing and non-complexing compounds was possible.

3.2.3. Supervised methods

Supervised methods like k-nearest neighbours (KNN) and linear discriminant analysis (LDA) were applied to see if complexation behaviour could be predicted. In Fig. 6 the results of a LDA leave-one-out cross-validation (LOOM) are gathered. The broad bars indicate the number of correct predictions and the thin bars the erroneous predictions. Clearly, all four data sets yield unsatisfactory results, with no method achieving more than 60% correct predictions. Interestingly, the complexing compounds are predicted slightly better than the non-complexing compounds: in all cases the thick bars on the left are much higher than the thin bars on the left whereas for the right bars the differences are much smaller and even negative in two cases.

The results of a leave-one-out validation for KNN are depicted in Fig. 7 as a function of the number of neighbours considered. For this, the distance matrices mentioned earlier with the unsupervised clustering analysis are used. The results are slightly better than the LDA results, which is not surprising given the apparent class overlap in the PCA plots of Fig. 4.

3.2.4. Variable selection

Using chemical considerations, it could be argued that the complexing compounds form a single, well-defined group unlike the non-complexing compounds. The latter may have any kind of shape or charge distribution and is therefore, much more diverse in nature. From this it might be concluded that the similarities with non-complexing compounds may not contain useful information and may in fact disturb the analysis. To verify this, the Xshape and Xcharge similarity matrices are modified in order to remove all columns associated with non-complexing compounds (55 columns are retained). Another way to decrease the inter-class variability is to concentrate on only

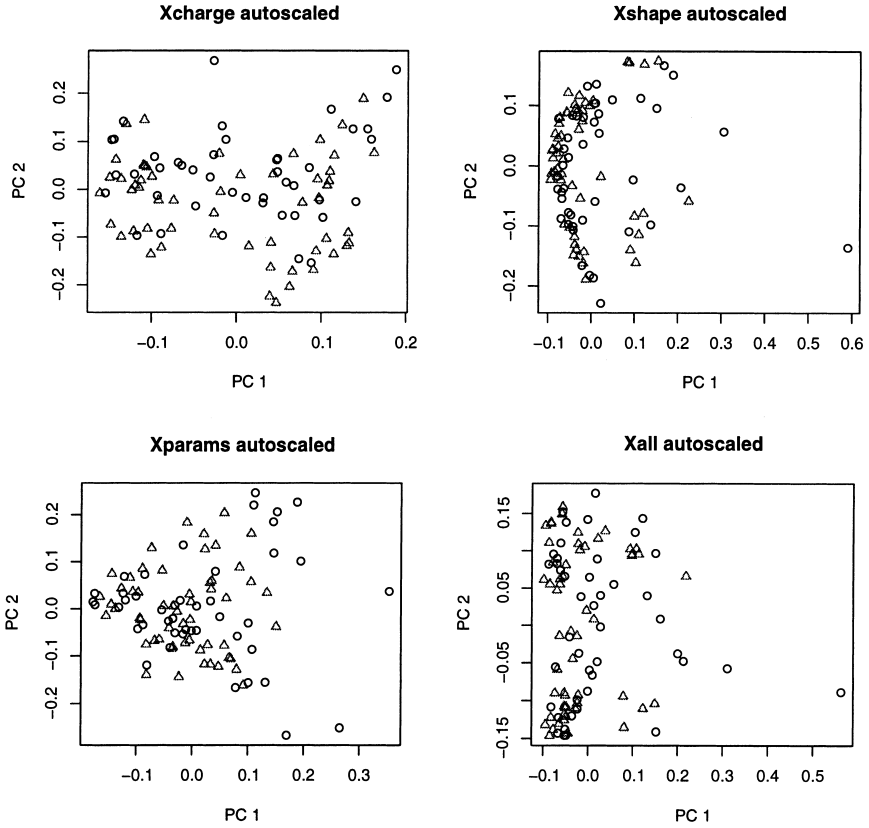

Fig. 4. Scores on the first two PCs for the three data matrices and the combined data matrix. Complexing compounds are indicated by triangles, non-complexing compounds by circles.

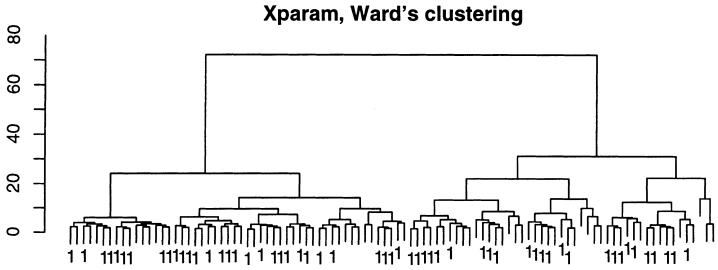

Fig. 5. Clustering according to Ward's method of Xparam. Euclidean distances are used calculated from the autoscaled data matrix. Labelled leaves indicate complexing compounds; non-labelled leaves indicate compounds that do not show complexing behaviour.

one class of compounds, e.g., the benzene derivates, and to remove polycyclic compounds or heterocycles from the data sets Xshape and Xcharge, leading to data matrices of size 99×69. Finally, these two criteria can be combined to retain only columns from the similarity matrices corresponding to complexing benzene derivates. In that case, only 36 variables remain.

The smaller data sets are indicated with Xshape* and Xcharge*, respectively. The combination of these two data sets is indicated by Xchsh*. The results of the LDA analysis on these data sets are gathered in Table 2. Again, data matrix Xcharge* appears to contain the most relevant information. Focusing on complexing compounds seems to have a beneficial effect

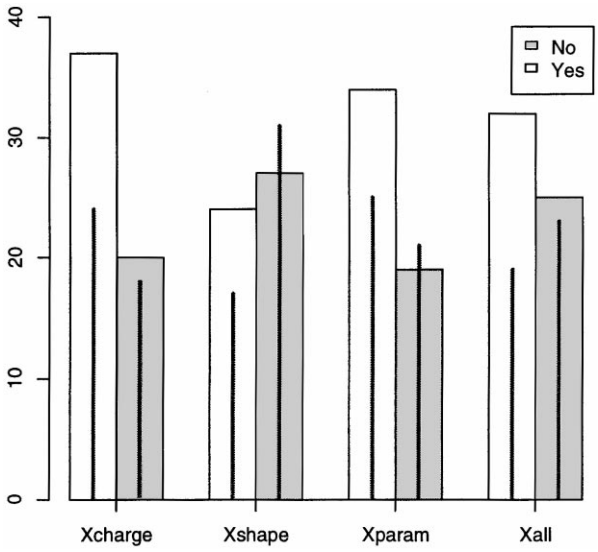

Fig. 6. Leave-one-out results for LDA using the four data sets. The left bars indicate the number of cases that complexation is predicted; the right bars indicate the number of cases in which no complexation is predicted. Thick bars indicate correct predictions, thin bars indicate errors. The sum of each set of four bars equals 99, the number of objects in the data set.

on the predictions, even when also the non-benzene derivates are predicted with these models. Interestingly enough, the predictions of the benzene derivates alone are not better than the predictions of the other compounds. The latter conclusion is also reached when using Xparam* and Xchsh*.

To assess whether the similarities with an even smaller set of 'reference' compounds could be used to predict complexation behaviour, a genetic algorithm (GA) [26–28] was used to select a maximum of 10 columns from data set Xcharge, which seemed to be the most important data set. For the LDA models, the LOOM error was used as evaluation function; for the KNN classification the mean prediction error of five random subsets of 50 compounds (rows in Xcharge) was used. Both $k=1$ and $k=3$ were used, but since the results were very similar, only the results for $k=1$ are given below. For both the LDA and KNN subset selection, 10 GA runs were performed, each starting from a different random initial population. In all LDA models, success rates of 85% or higher were obtained. Results for KNN appeared to be even better by approximately 5%. Whereas the KNN subsets D′-M′ always contained 10 reference compounds, the LDA

Table 2
Prediction results (LOOM success rates in LDA) for complexation behaviour using several subsets[a]

	Xcharge*	Xshape*	Xchsh*
A	65.6	61.6	58.6
B	67.7	50.5	50.5
C	70.7	63.6	60.6

[a] A: only complexing compounds; B: only benzene derivates; C: only complexing benzene derivates.

subsets were smaller, sometimes containing as few as two reference compounds. However, seven compounds were selected in five GA runs or more in the KNN case, whereas in the LDA case no compound was selected in more than three out of 10 GA runs.

The variable selection results were compared with the models A–C from Table 2 using the Xcharge* data set on an independent test set of 20 compounds. For the LDA prediction of the test set, the columns indicated by the subsets A–M are scaled using the mean and variance of the training sets. The prediction results are displayed in Fig. 8.

From the figure, it is clear that variable selection based on chemical considerations (A–C) does not improve prediction ability, whereas variable selection with a GA, on the whole, does. The improvement is not large, however, and certainly not as large as indicated by the LOOM values: success rates are lower than the GA LOOM estimates by 20–30%. One possible cause for this is that the objects in the training set, in this case the Xcharge* variants, are not quite representative for the test set. In Fig. 9 it is shown that for both the Xcharge and Xshape data sets there are small but systematic differences in column means in the training and test set for approximately the first 80 compounds, whereas differences can be quite large for similarities with compounds 80–99. This may indicate that the training set is not quite representative for the test set. Additional evidence for this hypothesis is given by the fact that no variable selection leads to a 'random' [3] prediction.

Another reason for the low success rate for the test set could be that there is a kind of overfitting involved in the variable selection. The selected columns do a good job of predicting the training set but little

[3] Since the numbers of complexing and non-complexing agents are approximately equal, a random classification would also lead to 50% success.

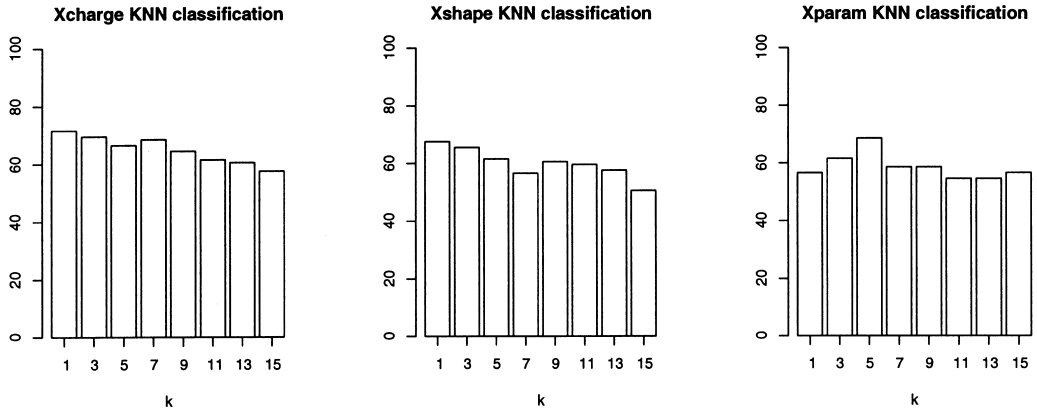

Fig. 7. Leave-one-out results for KNN prediction with a varying number of neighbours (k). The best predictions are obtained with Xcharge using one neighbour (71.7% correct).

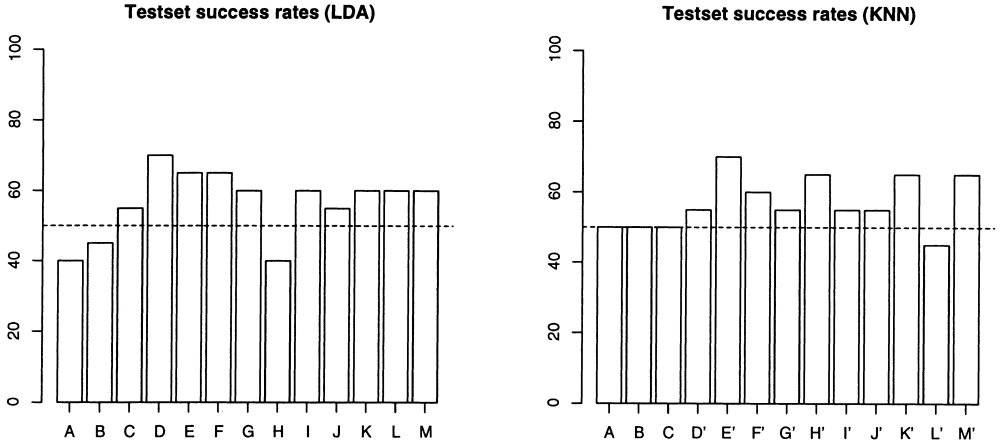

Fig. 8. Success rates of prediction of an independent test set of 20 compounds with LDA and KNN. Letters A–C refer to the Xcharge* models from Table 2; letters D–M and D′–M′ refer to ten subsets obtained with genetic algorithms for LDA and KNN, respectively. Each subset contains at most ten columns from data set Xcharge. Dashed lines indicate LDA and KNN performance, respectively, for the test set without variable selection.

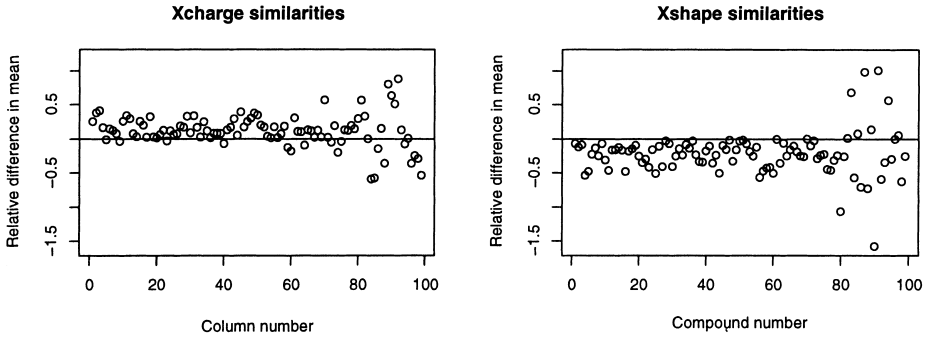

Fig. 9. Differences in column means between training and test sets for the Xcharge and Xshape data sets. Differences in means are scaled by the standard deviation of the corresponding columns in the training set.

predictive ability is achieved. However, the subset selection with the GA does seem to increase the predictive ability of both KNN and LDA models. There appears to be no relation between the number of reference compounds selected (2–10 with LDA) and prediction error.

4. Discussion

As the example shows, analysis of a problem in which molecular structure plays an important role can be quite different from classical chemometrical problems. First of all, the representation of the molecules in the data set is not straightforward. In some cases such as this one, indirect data are available in the form of similarities or dissimilarities. The interpretation of, e.g., a principal component loading or score vector of such data is unclear.

Second, the data may not contain sufficient information to solve the problem. In the example given, it is expected that molecular shape and charge distribution play an important role in whether or not complexes are formed. However, the similarity data and the physico-chemical parameters only describe global characteristics of the molecules. This means that small, perhaps very important differences in structure may be overshadowed by apparent similarities (the structures in Fig. 2 seem to underline this point). Switching to local descriptors that take into account molecular structure at a more detailed level has the disadvantage that we do not know the relevant structure precisely, at least not the exact conformation of both the guest molecule and the cavity. To make things even more complicated, cavities formed by cephradine may vary to a certain extent in order to adjust to the size and shape of the guest molecule. Although this is more or less taken into account implicitly by using a training set in which all known cavity shapes are represented, it might lead to difficulties if new complex types are present in the test set.

In an ideal situation we should therefore, be able to predict the structure of the host and guest molecules as well as the role of the solute, and use this knowledge to predict whether or not complexation will occur. Closely related to this is the issue of flexibility: molecules are clearly not the rigid structures they might seem to be from computer representations! Although in this case most molecules have a quite rigid basic structure, the flexibility of functional groups may play an important role.

Finally, for this kind of problem it is very difficult to define the concept of a ‘representative’ training set. One usually has to use whatever is available (and in the example the number of positive examples was roughly half of the complete set). In many cases a predictive model then is used to screen a data base of candidate structures. Such a ‘test set’ is, almost by definition, of a much wider range than the compounds in the training set, and the number of expected hits is much lower.

The most positive outcome one can hope for is a model that gives insight in why some compounds form complexes and why others do not. Despite all the objections against local descriptors, these are the only ones enabling such an interpretation. Therefore, further work is concentrating on such matters. The fact that multivariate outlier detection was able to identify compounds with less flat shapes indicates that the combination of several descriptor types may be beneficial.

5. Conclusions

Modern analytical laboratories, whether in industry or otherwise, are asked to provide answers to a much wider range of questions than a few years ago. This of course also has an impact on the chemometrical methods employed, and one important class of questions involving chemical structure has been highlighted in this paper. The main impediments for a successful application of chemometrics in this type of research are the inadequacy of many computer representations of chemical structures and the inability of chemometrical methods to fully utilise the information that is contained in these representations.

The important message is that chemistry returns to chemometrics: whereas in the past in many cases chemometrics might have resembled just another branch of applied statistics, the challenges highlighted in this paper can only successfully be tackled with a good knowledge of the underlying chemistry [29]. Combination of chemometrical methods with state-of-the-art computational chemistry packages opens up exciting prospects in QSAR. Modern techniques such as three-way PLS analysis have been

applied to chemical structures in order to find more easily interpretable representations [30]. With all these new challenges, the future for chemometrics in the next millennium looks broad and bright.

Acknowledgements

DSM Life Science Products and the Dutch Ministry of Economic Affairs are kindly acknowledged for their financial support.

References

[1] P. Geladi, H. Grahn. Multivariate Image Analysis, Wiley, Chichester, 1996.
[2] H. van de Waterbeemd (Ed.), Chemometric Methods in Molecular Design, vol. 2 of Methods and Principles in Medicinal Chemistry, VCH, Weinheim, 1995.
[3] P.S. Shenkin, D.Q. McDonald, Cluster analysis of molecular conformations, J. Comput. Chem. 15(8) (1994) 899–916.
[4] C.H. Reynolds, R. Druker, L.B. Pfahler, Lead discovery using stochastic cluster analysis (SCA): a new method for clustering structurally similar compounds, J. Chem. Inf. Comput. Sci. 38(2) (1998) 305–312.
[5] B. Schneider, S. Neidle, H.M. Berman, Conformations of the sugar-phosphate backbone in helical DNA crystal structures, Biopolymers 42 (1997) 113–124.
[6] C.M. Duarte, A.M. Pyle, Stepping through an RNA structure: a novel approach to conformational analysis, J. Mol. Biol. 284 (1998) 1465–1478.
[7] M.L.M. Beckers, L.M.C. Buydens, Multivariate analysis of a data matrix containing A-DNA and B-DNA dinucleotides. Multidimensional Ramachandran plots for nucleic acids, J. Comp. Chem. 19 (1998) 695–715.
[8] J. Giraldo, S.J. Wodak, D. van Belle, Conformational analysis of GpA and GpAp in aequeous solution by molecular dynamics and statistical methods, J. Mol. Biol. 283 (1998) 863–882.
[9] G.M. Downs, P. Willett, Similarity searching in databases of chemical structures, in: K.B. Lipkowitz, D.B. Boyd (Eds.), Reviews in Computational Chemistry, vol. 7, Wiley, New York, 1995, pp. 1–66.
[10] P. Willett, J.M. Barnard, G.M. Downs, Chemical similarity searching, J. Chem. Inf. Comput. Sci. 38(6) (1998) 983–996.
[11] G.A. Arteca, Molecular shape descriptors, in: K.B. Lipkowitz, D.B. Boyd (Eds.), Reviews in Computational Chemistry, vol. 9, VCH, New York, 1996, chap. 5, pp. 191–253.
[12] R.D. Brown, Y.C. Martin, Use of structure–activity data to compare structure-based clustering methods and descriptors for use in compound selection, in: J. Chem. Inf. Comput. Sci. [13], pp. 572–584.
[13] R.D. Brown, Y.C. Martin, The information content of 2D and 3D structural descriptors relevant to ligand–receptor binding, J. Chem. Inf. Comput. Sci. 37 (1997) 1–9.
[14] T.I. Oprea, C.L. Waller, Theoretical and practical aspects of 3D QSARs, in: K.B. Lipkowitz, D.B. Boyd (Eds.), Reviews in Computational Chemistry, vol. 11, VCH, New York, 1997, chap. 3, pp. 127–182.
[15] B.G.M. Vandeginste, D.L. Massart, L.M.C. Buydens, S. de Jong, P.J. Lewi, J. Smeyers-Verbeke (Eds.), Handbook of Chemometrics and Qualimetrics: Part B, volume 20B of Data Handling in Science and Technology, Elsevier, Amsterdam, 1998.
[16] K. Baumann, Uniform-length molecular descriptors for quantitative structure–property relationships (QSPR) and quantitative structure–activity relationships (QSAR): classification studies and similarity searching, Trends Anal. Chem. 18(1) (1999) 36–46.
[17] G.J. Kemperman, R. de Gelder, F.J. Dommerholt, P.C. Raemakers-Franken, A.J.H. Klunder, B. Zwanenburg, Clathrate type complexation of cephalosporins with β-naphtol, Chem.: a European J. 5 (1999) 1205–1210.
[18] G.J. Kemperman, R. de Gelder, F.J. Dommerholt, P.C. Raemakers-Franken, A.J.H. Klunder, B. Zwanenburg, Design of inclusion compounds of submitted cephalosporin derivatives, submitted for publication.
[19] J.E. Jackson, A User's Guide to Principal Components, Wiley, New York, 1991.
[20] L. Kaufman, P.J. Rousseeuw, Finding Groups in Data, An Introduction to Cluster Analysis, Wiley, New York, 1989.
[21] Tsar version 3.2, Oxford Molecular Group PLC, Oxford, UK.
[22] R. Ihaka, R. Gentleman, R: a language for data analysis and graphics, J. Comput. Graphic. Statist. 5(3) (1996) 299–314.
[23] J.F. Doreleijers, J.A.C. Rullmann, R. Kaptein, Quality assessment of NMR structures: a statistical survey, J. Mol. Biol. 281 (1998) 149–164.
[24] P.J. Rousseeuw, A.M. Leroy, Robust Regression and Outlier Detection, Wiley, New York, 1987.
[25] W.J. Egan, S.L. Morgan, Outlier detection in multivariate analytical chemical data, Anal. Chem. 70 (1998) 2372–2379.
[26] C.B. Lucasius, G. Kateman, Understanding and using genetic algorithms Part 1. Concepts, properties and context, Chemometr. Intell. Lab. Syst. 19 (1993) 1–33.
[27] C.B. Lucasius, G. Kateman, Understanding and using genetic algorithms Part 2. Representation configuration and hybridization, Chemometr. Intell. Lab. Syst. 25 (1994) 99–145.
[28] R. Wehrens, L.M.C. Buydens, Evolutionary optimisation: a tutorial, Trends Anal. Chem. 17(4) (1998) 193–203.
[29] S. Wold, M. Sjöström, Chemometrics, present and future success, Chemometr. Intell. Lab. Syst. 44(1) (1998) 3–14.
[30] J. Nilsson, S. de Jong, A.K. Smilde, Multiway calibration in 3D QSAR, J. Chemometr. 11 (1997) 511–524.