

Tutorial

Molecular data-mining: a challenge for chemometrics

Lutgarde M.C. Buydens^{*}, Theo H. Reijmers, Mischa L.M. Beckers¹, Ron Wehrens

Laboratory of Analytical Chemistry, University of Nijmegen, Toernooiveld 1, 6525 ED, Nijmegen, Netherlands

Abstract

The discipline of chemometrics is rooted in analytical chemistry. Currently, however, the application range of chemometrical techniques is being widened to molecular questions as well, addressing molecular conformations and behaviour. With the increasing availability of databases through the world-wide web, the need for techniques that help extracting information from data is greater than ever. This problem is generally termed data-mining. Several aspects of the application of chemometrics in this domain are highlighted and a worked example is given. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: Chemometrics; Analytical chemistry; Data-mining

Contents

1. Introduction	121
2. Data-mining: specific issues	122
2.1. Database size	122
2.2. Contaminated data	123
2.3. Methods	123
3. Representation	124
4. Example: conformation of DNA	125
4.1. Outliers	126
4.2. Representation	128
4.3. Incorporation of prior knowledge	131
5. Conclusions and future research	132
Acknowledgements	132
References	133

^{*} Corresponding author. Tel.: +31-24-3653192; fax: +31-24-3652653; E-mail: ibuydens@sci.kun.nl

¹ Present address: Kluwer Academic Publishers, P.O. Box 990 3300 AZ Dordrecht, Netherlands.

1. Introduction

One new challenge for chemometrics in chemistry is the spectacular growth of databases that contain a

large amount of data on molecular structures and their properties, for instance the Cambridge Structural Database (CSD) [1] for organic and organometallic crystal structures, the Brookhaven Protein Data Bank (PDB) [2], mainly for protein structures, and the Nucleic Acid Database (NDB) [3] for nucleic acids. These and many other databases are now readily available through the internet. A worrying trend is that more and more databases are exploited commercially, thus making it more difficult for universities and non-profit research organisations to assess their contents.

In the beginning, the primary use of these databases was storage and retrieval of data. These are, of course, valid objectives. However, the more data was gathered, the more it was realised that these databases contained important information that was not explicitly brought into them. It is the information hidden in the relations between all these data, such as similarities and dissimilarities, that may reveal important new chemical knowledge. Finding these hidden relations in databases is sometimes called data-mining or knowledge discovery. In this new field an explosion of activities has occurred, especially in computer science and applied statistics. Numerous web-sites and journals have emerged, such as *Data-mining and knowledge discovery: an international journal*. Old and new techniques, mainly in the area of classification, are proposed as knowledge discovery tools for all kind of database searches. Also in chemistry, more and more applications of data-mining techniques are appearing ([4] reviews some examples in the field of chromatographic databases).

Since the mission statement of chemometrics can be paraphrased as ‘‘to maximise the yield of chemical information and knowledge from chemical data’’, chemometricians are in an ideal position to extend the application area of data-mining techniques to chemical problems (see, e.g., Ref. [5]). It is clear that chemometrical techniques such as cluster analysis can contribute in finding relations in databases containing numerical results of chemical measurements. However, a class of questions that is growing more and more important and that will be termed loosely ‘‘molecular questions’’ in this tutorial, is concerned with molecular structure. What similarities can be found in a database, and how do they relate to differences in chemical behaviour? Can we discern group-

ings in the data? What substructures are consistently present, or perhaps more importantly, absent? It is these kinds of questions that are often tackled using the databases mentioned above. They typically contain molecular conformations, sometimes together with results from physico-chemical measurements such as log P values.

Chemometrical methods, as they are now, are not suited for handling molecular structures, nor are the molecular structure representations employed in the different databases very suitable for direct analysis. The information is there, but not in a usable form. This tutorial is meant to demonstrate how chemometrical techniques can be used for discovering new chemical knowledge from (large) databases. By means of an example we will show the important questions, steps and pitfalls that one may encounter in this new research field for chemometrics. It is not our intention, nor is it within our current knowledge, to provide ready-made solutions for all problems. Rather, we will try to highlight areas where more research is needed before one can truly speak of discovery of chemical knowledge.

2. Data-mining: specific issues

2.1. Database size

Data-mining is a term that is used to describe the process of extracting information and identifying interesting patterns or features out of large masses of data. This is certainly quite similar to an important branch of chemometric research, namely exploratory data analysis. The main difference lies in the size of the databases and all problems related to this. Chemometricians are used to deal with relatively clean, often more or less designed datasets. These datasets are analysed with all kinds of pattern recognition and other exploratory data-analysis tools. The results can be evaluated statistically in order to draw chemically relevant conclusions.

The number of objects in databases, interesting for data-mining, is orders of magnitude larger. Consider, e.g., the database resulting from the human genome project already containing gigabytes of data. This leads to problems where neither chemometricians nor statisticians are used to. One obvious consequence is

that with current computer technology it is not possible to keep all data in memory. This means that if one wants to process all data, new algorithms and strategies will have to be developed that can process the data sequentially, or that do not use all data but a limited subset. Examples of the latter approach can be found in techniques for clustering large datasets; if the number of objects is too large for the clustering algorithm, clusters are created using a subsample, randomly selected from all objects, and the other objects can be classified if necessary [6,7]. A further complicating factor is that information on objects may not be stored in a single datafile but rather in different, perhaps interrelated, databases. This makes the access to the data even more complicated.

A second problem is the high dimensionality of the data. Clearly, not all variables will be relevant in identifying new patterns. A distance measure in the full dimensionality may easily overlook patterns differing in only a small number of variables. Specialised clustering techniques using only a small subset of variables to discriminate between classes have been proposed [8]. In the case of molecular databases, the problem of finding an appropriate distance metric is directly related to the representation of the data, which will be treated in a separate paragraph.

Another less obvious problem associated with the large amounts of data lies in the statistical evaluation of the results. Most statistical tests are based on a fixed level of significance, e.g., $\alpha = 0.01$ or 0.05 (the probability for a type I error or for wrongly rejecting the null hypothesis). However, given the large amounts of data, any tiny difference will become statistically significant, even when it has no real meaning. The power of tests (the ability of correctly rejecting the null hypothesis) will be much more important in data-mining research.

2.2. Contaminated data

A further difference with data sets usually encountered in chemometrics is the way databases are constructed. Data are not gathered according to a design, but rather are stored when measured and interpreted. There are several consequences of this fact. Firstly, old data are probably of a different quality than more recent data. Especially with recent improvements in structure determination by NMR and

X-ray crystallography, the conformation of much larger structures can be obtained than was possible a few years ago, and with better resolution. Moreover, not all scientists submitting data use the same techniques and the same equipment, so that even with the newest data some inherent quality differences will be present. Furthermore, the objects in a database are almost certainly a non-random sample of all possible objects, a fact which may lead to unwarranted conclusions. The data almost always will deviate from normal distributions, and are easily contaminated by errors.

For these reasons, strategies to cope with outliers are extremely important. The further development of robust multivariate methods, in which the outlying observations do not significantly alter the observed structure in the data, is crucial. It is also important to reconsider the treatment of outlying observations. Traditionally, when an outlier is identified, the most straightforward action is to go back to the source and try to uncover the reasons for the outlying observation, if possible correct it, and otherwise remove it before further data analysis. In knowledge discovery these outliers may represent novelties, exactly what one is looking for. In this view, outliers are observations that deserve special attention, rather than to be disposed. In the example below we will try to illustrate this aspect.

Missing data represent a further problem. In cases where not all variables are measured or in cases where some variables are impossible to measure, conventional statistical methods to model the dependencies in the data may break down. Taking only those variables and objects without missing values into account may lead to a substantial decrease of the data set size. Apart from the smaller generality of the conclusions, the chance that something interesting is missed increases. Maybe there is a reason for the missing values!

2.3. Methods

Methods for data-mining are closely related to those used traditionally in exploratory data analysis. For smaller data sets, visualisation methods in two or three dimensions are very important (e.g., Ref. [9]). As John Tukey once said, “numerical summaries focus on expected values, graphical summaries on un-

expected values''. The human capability of visually recognising regularities in data is still unsurpassed by computer methods. Common techniques include Principal Component Analysis (PCA), projection pursuit and pattern recognition techniques, unsupervised (cluster analysis) as well as supervised. Empirical modelling techniques are potentially useful in data-mining, too. All these methods will have to cope with the problems described in the previous paragraph.

With larger databases, visualisation in two or even three dimensions is less important because of the dimensionality of the data and the sheer number of data points. In such a case, it is imperative that the hidden structure in the data is detected automatically. While automatic search is certainly not new in chemometrics (an example is the automatic variable selection by means of simulated annealing or genetic algorithms such as applied in Ref. [10]) new research is clearly necessary. Inductive algorithms such as the ID3 family also receive renewed attention in knowledge discovery research and are commercially available in several computer packages [4]. Applications of both inductive logic and genetic algorithms to derive rules for the selection of a detection method in ion chromatography have been described in chemical literature [11,12].

3. Representation

The way in which chemical structures are represented in a computer is often crucial for success or failure of the subsequent analysis. For trained chemists a very concise representation like a 2D structure is sufficient to infer chemical properties such as chemical similarities, reactive groups and partial charges. Most databases store chemical structures in the form of coordinate files containing the Cartesian coordinates of the individual atoms, sometimes accompanied by the connectivities. Smaller organic molecules are often stored in the well-known Wiswesser line notation [13]. For data-mining, these representations offer only limited possibilities, and therefore many different secondary representations have been developed.

Some describe the molecule as a whole using one number or a vector of numbers. Examples of these

so-called *global* descriptors are the dipole moment, total charge, several topological indices, or spectral-like representations. A disadvantage of these descriptors is that information on the local structure of the molecule is lost. One approach which does not suffer from this is Comparative Molecular Field Analysis (COMFA), in which quantities like mass distribution or electron density are evaluated on a three-dimensional grid over the molecule. The resulting 3D data matrix then is used in conjunction with PLS to predict the desired properties. This preserves locality, in that locally important features such as the presence of charges at particular locations in the molecule are not averaged into one isotropic descriptor, but has the disadvantage that it is relative to some fixed point in space, usually the centre of mass. This necessitates an alignment of all molecules in the data set. Although aligning a set of molecules with a common skeleton is not that difficult, a more diverse data set may pose some problems. Several approaches have been proposed, the DISCO set of programs [14] perhaps being the most widely used strategy. An overview of these 3D QSAR techniques can be found in Refs. [15,16].

The ideal molecular descriptor for a specific application contains all relevant information in such a form that the data-mining techniques employed are able to utilise this information. The most important point is that similarities and dissimilarities computed from the descriptor should reflect chemical reality, either because of similar observed behaviour or other characteristics. The work of Downs and Willett is important in this respect [17,18]. Furthermore, the descriptor should be concise, and if necessary should allow for flexibility in molecular structures. The number of degrees of freedom may be decreased significantly by considering the structures as flexible aggregates of rigid building blocks such as chemical bonds or larger entities such as functional groups, amino acids and nucleosides. Since most molecular questions are relevant in the context of a limited subset of structures, such a specialised representation is not necessarily a disadvantage. In this paper, we will illustrate this with the example of DNA sequences which are described by a limited number of torsion angles. A final prerequisite is that most modelling techniques require that the size of the descriptor is independent of the size of the molecule. This is quite

an important restriction, and so far none of the known fixed-length representations is consistently better than all others [19].

4. Example: conformation of DNA

The example that we will use to demonstrate some of the above mentioned aspects concerns the conformational analysis of DNA (deoxyribose nucleic acid). DNA is the carrier of genetic code, and during the last decade the human genome project has made huge efforts to elucidate the complete sequence of human DNA. Since the three-dimensional structure of biomacromolecules for a large part determines their biological functioning, it is very important to discover the relation between sequence and structure, be it of DNA, proteins, or other biomacromolecules.

This type of analysis serves several purposes which are listed below.

- To develop rules that will limit the dimensionality of the data. The relation between certain torsion angles may be indicative of a specific DNA class. This has been investigated in the past by bivariate relations between torsion angles or other structural parameters (e.g., Refs. [20,21]). A multivariate approach can reveal new information.

- The development of multivariate Ramachandran-like conformational maps that indicate “forbidden” areas [22,23]; these are useful in classifying and validating new structures. If the torsion angles of a tentative structure do not fall in the allowed regions, there may be reason to investigate the structure further to see whether this is a physically impossible structure, or just different from most other structures.

- The multivariate relationships that are uncovered in this approach may serve as a guide for future experiments.

In double-stranded DNA, two backbones consisting of alternating sugar and phosphate groups are kept together by hydrogen bonds between complementary bases in the opposing strands. The four regular bases in DNA and their pairings are depicted in Fig. 1, and a view of the backbone and the relevant torsion angles is given in Fig. 2. Sometimes derivatives of these bases occur in DNA sequences, or suboptimal base pairings in which the full potential for hydrogen bonding is not used. The binding between the two strands at these locations is less strong which may have an effect on the molecular structure.

Whereas the bases in the DNA carry the genetic information, the deoxyribose and phosphate groups that form the backbone of a DNA strand perform a structural role. The classical Watson–Crick double helix is right-handed and contains about 10 residues per turn. This form, called B-DNA, is sometimes divided in two classes: BI- and BII-DNA [24]. In this paper we use the criterion from Ref. [22], where a conformation is classified as BI-DNA if torsion angle ζ is smaller than ε . Otherwise, the conformation is BII-DNA. This criterion is used to be able to classify the borderline cases into one of the two classes. Several other helix conformations exist: a right-handed helix with 11 residues per turn, called A-DNA, and a left-handed helix containing about 12 residues per turn (Z-DNA). The preferred conformation is determined by the DNA sequence and external factors (e.g., A-DNA can be formed from B-DNA upon dehydration).

Given the Cartesian coordinates of a piece of DNA or a graphical representation of it, a trained chemist

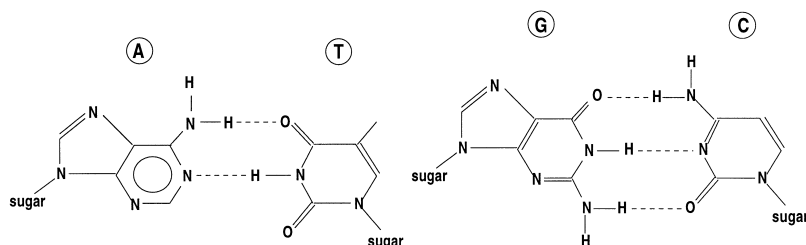


Fig. 1. Base-pairing in DNA. The pair adenine (A) and thymine (T) is bound by two hydrogen bonds (left) and the pair guanine (G) and cytosine (C) by three (right).

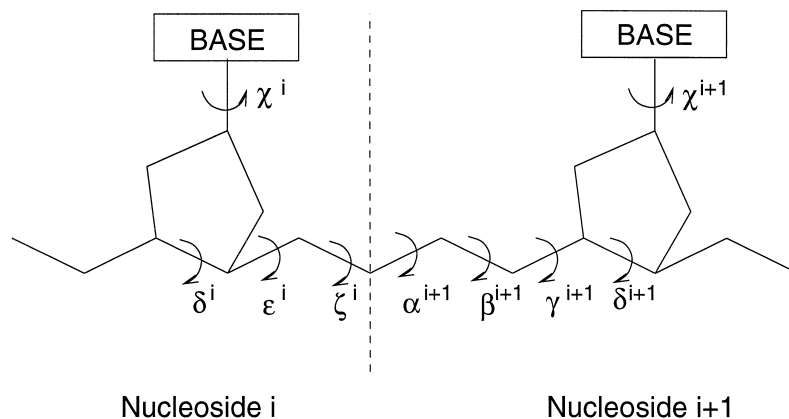


Fig. 2. The nine torsion angles used as parameters in this paper. Six angles (α – ζ) determine the conformation of the DNA backbone; the seventh (χ) the orientation of the base. Superscripts indicate in what nucleoside the torsion angle is found.

is able to determine which helix conformation is present. In the present example, this classification is performed automatically. Several parameters can be used to describe DNA structure [25]; in this case, we more or less arbitrarily use nine torsion angles in the range of 0–360° to describe dinucleoside monophosphates. That is, we group the DNA sequence in steps of two nucleosides at a time. Only one strand is analysed; the other is assumed to have the complementary three-dimensional structure. This process is illustrated in Fig. 3.

The data are taken from the Cambridge Crystallographic Database and the NDB. When obtaining a dataset from such databases, care must be taken to ensure that the set is sufficiently homogeneous, i.e., that the individual objects in the set can be compared to each other in a meaningful way. For this end, only structures of comparable quality (resolution 1.4–2.6

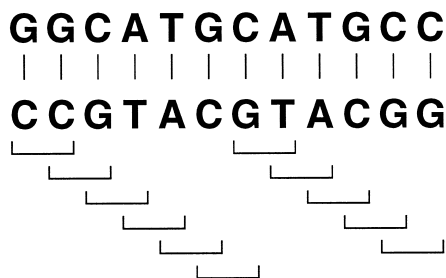


Fig. 3. Definition of objects in the data set. Each object is a dinucleoside monophosphate, represented by nine torsion angles. A DNA structure of 12 base pairs thus leads to 11 objects (CC, CG, GT, etcetera). Only the lower strand is analysed.

Å and R -factors in the range of 11.5–23.2% [22,20]) have been selected. This screening yielded a data set consisting of 33 crystal structures of DNA sequences, in total containing 287 dinucleoside monophosphate steps. The data are described in more detail in Ref. [22]. Four classes can be discerned, A-DNA, both BI- and BII-DNA, and a special class in which both the α and ζ torsion angles are *trans*, instead of the usual *gauche* – and *gauche* + conformations. This is called the crankshaft effect, and is most often observed in A-DNA. In total, 105 dinucleosides belong to the A-DNA class, 134 to BI-DNA, 43 to BII-DNA and 5 are crankshaft objects. Of the latter, four would be classified as A-DNA and one as B-DNA, but for simplicity we will treat them as one separate class.

The aim is to distinguish the different structural forms present in the dataset using multivariate analysis and to investigate which torsion angles are responsible for this subdivision. This approach is much more versatile and informative than approaches focusing on uni- and bi-variate analysis (e.g., Refs. [20,21]). In the present example we will elaborate on that analysis, show alternative ways to obtain information from these data and identify some pitfalls. In all cases, the data are median-scaled prior to analysis [22].

4.1. Outliers

Outliers are observations which, for some reason, do not conform to the general pattern present in a data

set. In general, the presence of outliers may have several causes. First, there is always the chance that the outliers are genuine errors. Of course, these must be removed prior to analysis. A potentially more difficult situation is the one in which the outliers are formed by unsuspected structure in the data. Since a large group of outliers is more difficult to detect than a few outlying observations, one should always be aware of this possibility. One further cause for outliers is the possibility that there is a reason that these data are different from the bulk. These outliers should be removed prior to analysis and possibly analysed separately. The problem is to recognise which kind of outliers we are dealing with.

Several aspects of outlier detection for the application of finding structure in database contents merit further attention. First of all, univariate outlier detection is not appropriate. As depicted in Fig. 4, outliers in a multivariate sense are not likely to be detected at all, whether it concerns true aberrant values or subgroupings in the data.

Second, for larger datasets the outlier detection cannot be performed by hand. Robust, automatic methods should be used. Which method is to be preferred depends on the goal of the analysis. In this case, where the goal is to find structure in the data set of DNA sequences, robust PCA as described in Ref. [26] will be applied. A robust distance is calculated for all objects, and objects further away from the centre are marked as outliers. These will receive a

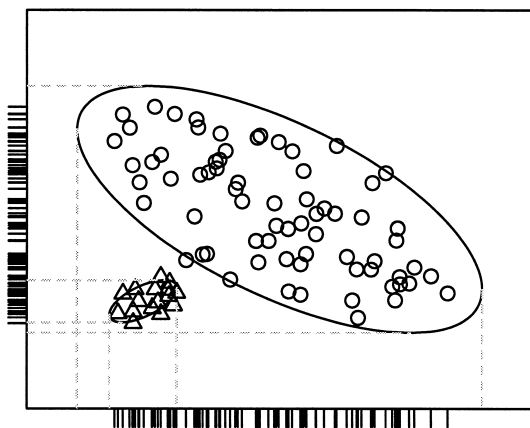


Fig. 4. Failure of univariate outlier detection for multivariate data. The distributions of both individual variables show no signs of outliers, while clearly two separate groups are present.

weight of 0.01, whereas the objects within the threshold distance will receive the full weight of 1. Finally, PCA is performed taking these weights into account. By downweighting the outliers, the PCA loadings are mostly determined by the non-outlying observations. In this way, aberrant data will not influence the model. Further analysis of the outliers may reveal causes for the non-standard behaviour. The threshold deciding whether objects are outliers or not can be set by the user. It should be noted that when no outliers are present the robust PCA yields slightly different results from classical PCA [26]. Many other approaches are possible [27].

After the first preselection step, in which only structures of comparable quality are extracted from the database, 287 objects, each represented by the nine torsion angles depicted in Fig. 2 are present in the data matrix. In Ref. [22], a manual validation procedure reduced the number of objects to 244 by removing those containing non-standard bases, mismatched base pairs, and torsion angles that would lead to Van der Waals clashes. It is not certain that especially the first two criteria are necessary, nor can we hope to eliminate all outlying observations in this way. A more safe strategy is to rely on robust outlier identification.

A comparison of classical and robust PCA on our data is shown in Fig. 5. The upper left plot in this figure shows the classical PCA object scores on the first two PC's, obtained from the manually validated set of 244 objects. Although there is some overlap between the different classes, the structure is clearly visible. On the PC 1 axis the A-DNA is separated from the B-forms. The A-crankshaft objects are on the edge of the A-DNA cluster, while the B-crankshaft object is right in the middle of the BI cluster. The comparison with the upper right plot shows how the classical PCA score plot changes upon inclusion of the 43 manually removed objects. Although the general structure of the three main clusters is still recognisable, the scatter, especially in the PC 2 direction is significantly larger. More importantly, the location of the non-outlying objects changes because of the inclusion of the 43 suspect objects. This is especially clear in the location of the crankshaft objects.

Similarly, the bottom left plot shows the robust scores from the 244 manually validated objects, and

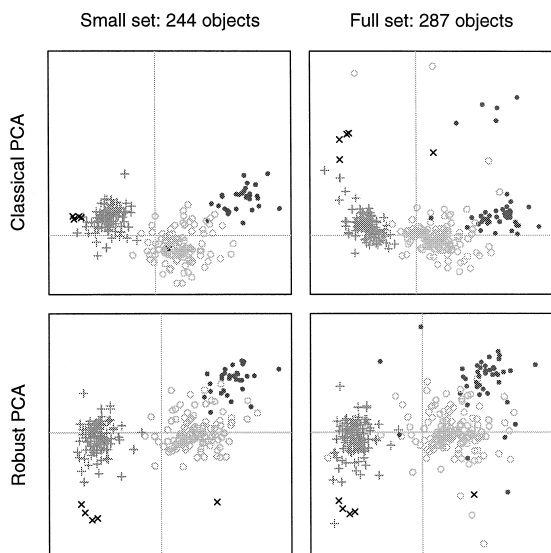


Fig. 5. Score plots (PC 1 vs. PC 2) of classical and robust PCA. In the two plots on the left, only the 244 manually validated objects are used to calculate scores and loadings. In the plots on the right, the complete set (287 objects) is used. Symbols used: A-DNA (+), BI-DNA (○), BII-DNA (●) and cranks shafts (×).

the bottom right plot right shows the robust score plot, produced using all 287 objects. The agreement between the two robust plots is clear. The objects, manually identified as outliers, indeed have no effect on the scores of the other objects, which is exactly what is required. Of course, objects identified as outliers by the robust procedure may still be scattered in the plot. A plot without the outlying observations will be shown in Fig. 8 and is very similar to the bottom left plot in this figure.

The weights that are assigned to the objects in robust PCA identify the observations marked as outlying. Depending on the threshold setting, a smaller or larger fraction of the data is marked as outliers. Here, the threshold is chosen in such a way that representatives of all four classes are retained. In total, 34 outliers were identified, all of them B-DNA. The majority (18 objects) were also excluded in the manual validation. Of the remaining 16 outliers, 11 were identified in the first two dinucleosides of a DNA sequence. This indicates that the larger freedom of movement in these parts may lead to somewhat unexpected conformations. These outliers are very hard

to find by manual validation. Interestingly, not all manually identified outliers are recognised as such by the robust procedure; therefore it can be concluded that the manual validation was too cautious and several mismatched base pairs and unusual bases do conform to the general trends.

In Fig. 6 it is shown more clearly that the robust scores of the 244 validated observations are not affected upon inclusion of the manually identified outliers. Especially for PC 2 and PC 3 the scores of the classical PCA are heavily disturbed, whereas the robust PCA still shows the same behaviour. The only significant deviations from the straight line in the robust score plots are objects marked as outliers in the manual validation step. Loading plots (not shown) exhibit the same behaviour.

This shows that robust analysis is less sensitive to outliers and more apt to uncover the underlying structure in data. Moreover, the robust analysis is able to assess whether perceived differences, such as the occurrence of uncommon bases, should be repaired prior to analysis. However, many different robust methods exist and the degree of permitted outlying observations is determined by the user by adjusting the threshold, so that a certain subjectivity is unavoidable.

The main point is that application of robust procedures can lead to the identification of (groups of) data points that do not conform to the pattern set by the majority: in some cases these “outliers” are genuine errors (databases do contain a lot of errors), in other cases they represent a subgroup with properties different from the bulk. Often, such a small subgroup is the most interesting object of study! In the DNA example presented here, the robust PCA is tuned in such a way that all DNA classes are modelled and the “outliers” can be regarded as individual aberrant objects.

4.2. Representation

In general, several different representations can be used to describe chemical data. This may greatly influence the subsequent analysis, so some thought or preliminary experimentation is required. In this case we want to analyse chemical structures, and obvious candidates are Cartesian coordinates, since these are most often found in databases. They form the most

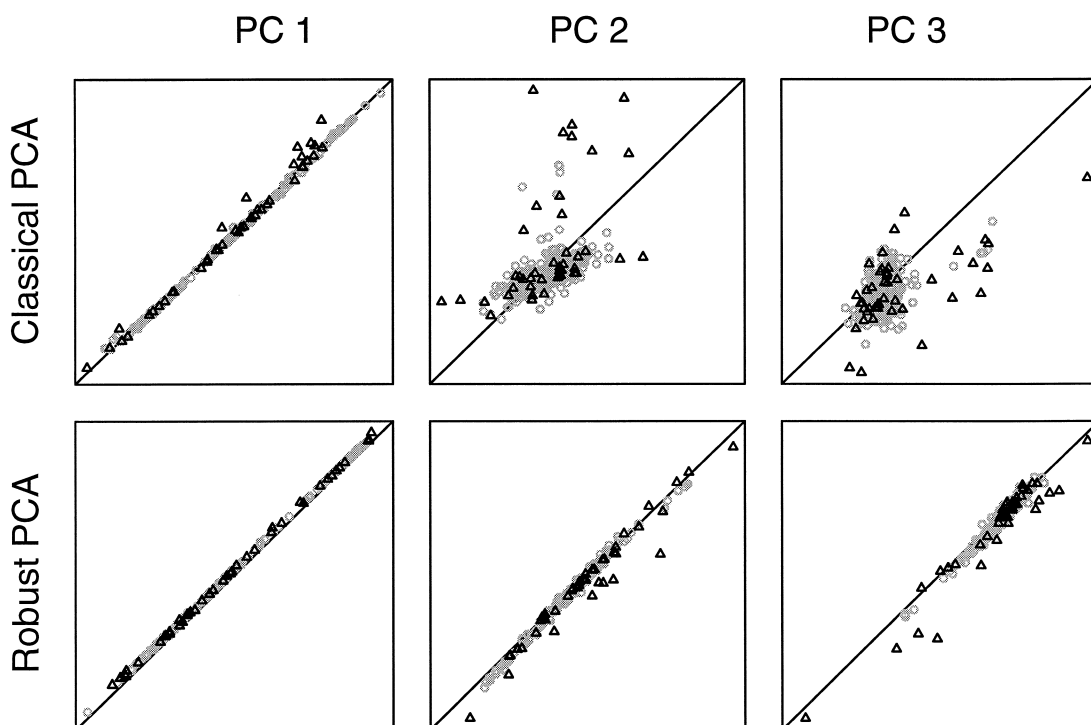


Fig. 6. The influence of outliers on PCA scores. X-axis: scores calculated from the 244 manually validated objects (○); the 43 other observations (△) are projected in this space. Y-axis: scores calculated from the complete set of 287 objects.

flexible representation, in that they can be used to describe any chemical structure, no matter how unusual. The main disadvantage is that for each atom three coordinates are needed, which leads to large data sets when many molecules are considered at the same time. Often, therefore, internal coordinates (torsion angles) are used. Because bond lengths and bond angles are kept fixed on standard values, the number of parameters needed to describe a chemical structure is much reduced. The disadvantage is that molecules with slightly deviating geometries cannot be described in this representation.

A second point is the choice of objects. In the DNA example, the A- or B-type DNA behaviour is observed on a scale larger than individual nucleosides. How many of them should we take into account? Is it necessary to analyse one complete turn in the helix (i.e., 10–11 base pairs) or would much smaller objects such as dinucleosides yield essentially the same information? Some experimentation may be needed to resolve this problem.

4.2.1. Choice of variables

Torsion angles may be represented in a number of ways: the first possibility is to use values between 0 and 360°. The obvious disadvantage of this is that the difference between angles of 20 and 340° is perceived as being much larger than it is in reality (see Fig. 7). Techniques using variance measures, such as PCA, can be severely affected by this effect. If a certain parameter of objects in the same class has values at the very high and very low end of the scale, chances are that this class will be split in two. Of course, inspection of the individual torsion angle distributions may reveal whether problems are to be expected, and it may be possible to choose another angle range for which no edge effects are present. However, for large numbers of parameters this is no longer feasible; moreover, it may be impossible to find a range for which no parameter distributions are on both sides of the edge. A more fundamental remedy is to represent each torsion angle by two values, a sine and cosine value. Difficulties with the subse-

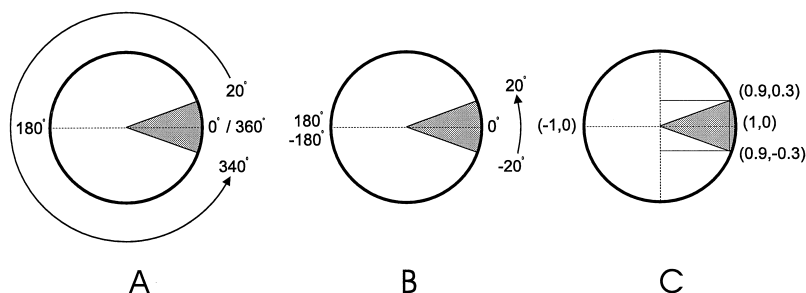


Fig. 7. The effect of the representation on the distance between two torsion angles. With conventional representations such as degrees or radians, angles that are in effect very close may appear to be quite different. The sine/cosine representation does not suffer from this.

quent analysis may be introduced by the non-linear sine/cosine transformation, however. Another disadvantage is that the number of parameters doubles.

To illustrate the effect of the choice for angles between 0 and 360°, we also analysed the same data set when torsion angles are values between -180 and 180° and sine/cosine tuples. The robust score plots for the three representations are shown in Fig. 8. Only those objects are plotted that are not identified as outliers. Again, the threshold for the robust PCA is chosen in such a way that all four DNA classes are represented. The representation with angles between -180 and 180° is especially sensitive to the choice of the threshold that defines the number of outliers; however, for all values of the threshold the apparent structure in the data is markedly different from the other representations and splittings of clusters are observed. This is caused by torsion angle values in

one class at both sides to 180° , which lie far apart in this representation. The sine/cosine representation shows a structure very much like the structure of angles represented between 0 and 360° , indicating that the latter representation in this specific case is not bothered by the cyclicity of the data. One peculiarity of the sine/cosine representation is that much fewer objects are seen as outliers. This is a result of the condition that all classes should be represented. If we were to relax this condition, the crankshaft objects would directly be identified as outlying observations. The overall division between the classes seems to be of slightly less quality than the representation using torsion angles between 0 and 360° . Probably this is caused by the non-linearity of the sine/cosine transformation. In the remainder of the paper, we will therefore stick to the original choice of torsion angles in the range of 0– 360° .

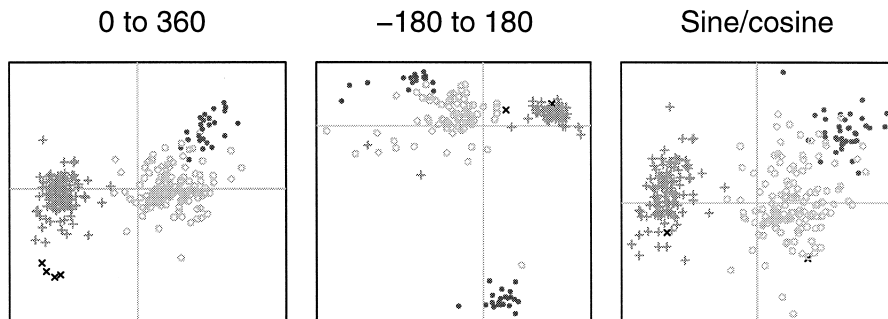


Fig. 8. Score plots (robust PCA) of the complete dataset in three different representations. Outliers, as identified by the robust procedure, are not plotted. The left plot shows the scores when torsion angles are in a range of 0– 360° (253 objects plotted). Torsion angles from -180 to 180° are used to produce the middle plot (170 plotted), and sine/cosine tuples for the right plot (275 objects). Symbols as in Fig. 5.

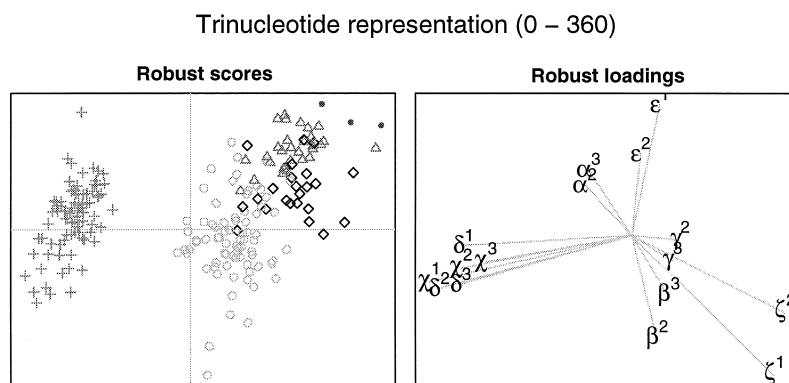


Fig. 9. Score plot (left) and loadings plot (right) for the robust PCA on the trinucleoside data set. In total, 17 objects are identified as outliers and are not plotted. Symbols are identical to the ones used in Fig. 5; furthermore mixed BI–BII (\diamond) and BII–BI (\triangle). Crankshafts are not shown. Superscripts indicate the number of the nucleoside in the object.

4.2.2. Object representations

To be able to include more information on the direct environment of a nucleoside, one can use trinucleosides instead of dinucleosides. This increases the number of torsion angles from 9 to 16, and reduces the number of objects from 287 to 234. Moreover, the number of classes increases, because also mixed BI–BII and BII–BI classes can be identified. Mixed classes containing crankshafts are not taken into account because they consist of single objects in this representation. The scores and the loadings from the corresponding robust PCA are shown in Fig. 9. Again, the same underlying structure as seen in Fig. 5 can be discerned, indicating that both representations capture the same relevant information. The mixed B-classes are nicely positioned in between the pure BI and BII classes. Moreover, the loadings plot shows significant correlations between identical torsions in subsequent nucleosides. The same correlations are found as in Ref. [22]. It can be concluded that larger objects of study do not convey new information; since the number of objects decreases rapidly and the interpretation of the results gets more difficult (mixed classes), the choice of analysing dinucleosides appears to be a good one.

4.3. Incorporation of prior knowledge

One final question is how to incorporate prior chemical knowledge in the analysis. In some cases, constraints may be defined beyond which chemical reality is violated. In this case, this may be done by

already distinguishing several groupings beforehand and analysing each of these separately. The nucleosides can be divided into purine derivatives (the joint five- and six-membered rings A and G) and pyrimidine derivatives (C and T). This leads to four dinucleoside combinations: purine–purine, pyrimidine–pyrimidine, purine–pyrimidine and pyrimidine–purine, and robust PCA can be performed on these four groups. The results are depicted in Fig. 10. In-

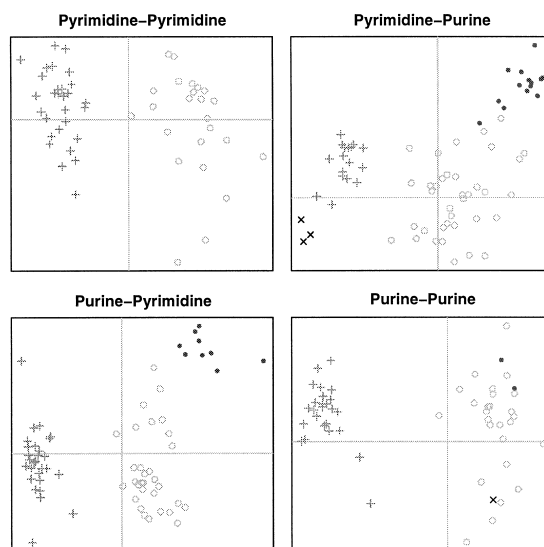


Fig. 10. Score plots (robust PCA) of the complete dataset, divided in four groups: purine–purine, pyrimidine–pyrimidine, purine–pyrimidine and pyrimidine–purine combinations. Classes seem to be further apart in this representation. Symbols as in Fig. 5.

terestingly, all BII-DNA is found in mixed purine–pyrimidine dinucleosides except for two purine–purine dinucleosides. The B-crankshafts are all located in the pyrimidine–purine plot. This kind of analysis may show unexpected distributions in data sets that are difficult to see otherwise. In this case, the class separations appear to have improved somewhat, but no real deviations from the previous experiments have been found.

5. Conclusions and future research

In molecular data mining, several constraints must be satisfied for methods to be useful. First of all, since the methods often operate on large databases, they should be very fast. This requirement is more important than the level of sophistication, so that simple methods will often be preferred to more complicated ones. In the example, the goal was to uncover hidden relations in the data. This was achieved by analysing a carefully selected subset of all available data. The models obtained in this way can be easily applied to the remainder of the data base with much less computational effort. This is an example of a typical data mining solution in cases where the amount of data is too large to take into account at once. However, there are cases where it is imperative to consider all potential objects in a data base, for instance in identifying potentially active molecules in pharmaceutical applications. In that case, one has the possibility to limit the number of variables that is taken into account instead of the number of objects in order to speed up the process.

Data mining methods should not be sensitive to outliers or unexpected distributions of the data. This means that truly aberrant values should have no effect on the outcome, and that clusters of data, identified as outliers should be stored automatically for further inspection. Especially in databases containing heterogeneous data (such as collections of physical constants), appropriate scaling procedures are mandatory. One should be aware that maybe not all parameters are useful when looking for structure in the data. A better clustering, for instance, may be obtained with only a small fraction of the parameters.

Chemical knowledge is essential for the correct application of data-mining techniques. Not only must

the results of the data mining exercise be evaluated in the context of the chemical question, in many respects the chemistry behind the data determines the steps that have to be taken. In the example presented here, the preselection of the data and the treatment of outliers as identified by the robust procedure are based on chemical considerations. The correlations found between identical torsion angles in neighbouring bases appeal to chemical common sense and are therefore all the more acceptable. Because of the importance of the inclusion of chemical knowledge, it is imperative not just to copy and apply data mining techniques from other sciences such as computer science, but to assess the applicability of each in the context of the chemical problem. The discipline of chemometrics has an important task in this respect. Firmly rooted in chemistry, chemometricians are also familiar with most data mining techniques. The ease with which these methods can be adapted for chemical applications will be a major criterion on which to judge the different approaches.

Finally, a thorough understanding of the interplay between the representation of the chemical data and the working of the data mining algorithms is required. As we have seen, inappropriate representations may lead to spurious relationships (such as the splitting of the BII-DNA class in Fig. 8). A representation capturing all *relevant* information in such a way that it can be processed by the model-building algorithm is essential for the maximum use of all databases that are at our disposal today. Of course, different representations may be needed for different applications, and this points to the necessity of being able to quickly transform data from one representation to another.

In what way flexible structures should be handled is even more unclear. The analysis of a set of possible conformers is time-consuming and only to a degree satisfying. This last item possibly forms the most important challenge of the next decade that has to be overcome to truly utilise the vast potential of the combination of large databases and chemometrics.

Acknowledgements

T.H. Reijmers is supported by the Dutch Organisation for Scientific Research (NWO-CW).

References

- [1] F.H.S. Allen, M.D. Brice, B.A. Cartwright, A. Doubleday, H. Higgs, T. Hummelink, B.G. Hummelink-Peters, O. Kennard, W.D.S. Motherwell, J.R. Rogers, D.G. Watson, The Cambridge Crystallographic Data Centre: computer-based search, retrieval, analysis and display of information, *Acta Cryst. Sect. B* 35 (1979) 2331–2339.
- [2] F.C. Bernstein, T.F. Koetzle, G.J.B. Williams, E.F. Meyer, M.D. Brice, J.R. Rogers, O. Kennard, T. Shimanouchi, M. Tasumi, The Protein Data Bank: a computer-based archival file for macromolecular structures, *J. Mol. Biol.* 112 (1977) 535–542.
- [3] H.M. Berman, W.K. Olson, D.L. Beveridge, J. Westbrook, A. Gelbin, T. Demeny, S.H. Hsieh, A.R. Srinivasan, B. Schneider, The Nucleic Acid Database: a comprehensive relational database of three-dimensional structures of nucleic acids, *Biophys. J.* 63 (1991) 751–759.
- [4] C.H. Bryant, R.C. Rowe, Knowledge discovery in databases: application to chromatography, *Trends Anal. Chem.* 17 (1998) 18–24.
- [5] S. Wold, M. Sjöström, Chemometrics, present and future success, *Chemom. Intell. Lab. Syst.* 44 (1) (1998) 3–14.
- [6] C.B. Lucasius, A.D. Dane, G. Kateman, On k-medoid clustering of large data sets with the aid of a genetic algorithm, background, feasibility and comparison, *Anal. Chim. Acta* 282 (1993) 647–669.
- [7] L. Kaufman, P.J. Rousseeuw, *Finding Groups in Data, An Introduction to Cluster Analysis*, Wiley, New York, 1989.
- [8] R. Agrawal, J. Gehrke, D. Gunopoulos, P. Raghavan, Automatic subspace clustering of high-dimensional data for data mining applications, *SIGMOD* 27 (2) (1998) 94–105.
- [9] D.F. Swayne, D. Cook, A. Buja, Xgobi: Interactive dynamic data visualization in the x window system, *Journal of Computational and Graphical Statistics* 7 (1) (1998) 113–130.
- [10] B.E. Mitchell, P.C. Jurs, Prediction of infinite dilution activity coefficients of organic compounds in aqueous solution from molecular structure, *J. Chem. Inf. Comput. Sci.* 38 (1998) 489–496.
- [11] M. Mulholland, D.B. Hibbert, P.R. Haddad, C. Sammut, Application of the C4.5 classifier to building an expert system for ion chromatography, *Chemom. Intell. Lab. Syst.* 27 (1995) 95–104.
- [12] A.H.C. van Kampen, Z. Ramadan, M. Mulholland, D.B. Hibbert, L.M.C. Buydens, Learning classification rules from an ion chromatography database using a genetic-based classifier system, *Anal. Chim. Acta* 344 (1997) 1–15.
- [13] W.J. Wiswesser, How the WLN began in 1949 and how it might be in 1999, *J. Chem. Inf. Comput. Sci.* 22 (1982) 88–93.
- [14] Y.C. Martin, M.G. Bures, E.A. Danaher, J. DeLazzer, I. Lico, P.A. Pavlik, A fast new approach to pharmacophore mapping and its application to dopaminergic and benzodiazepine agonists, *J. Comput.-Aided Mol. Des.* 7 (1993) 83.
- [15] 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, Chap. 3, VCH, New York, 1997, pp. 127–182.
- [16] G. Greco, E. Novellino, Y.C. Martin, Approaches to 3d qsar, in: K.B. Lipkowitz, D.B. Boyd (Eds.), *Reviews in Computational Chemistry*, Vol. 11, Chap. 4, VCH, New York, 1997, pp. 183–240.
- [17] 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, Chap. 1, Wiley/VCH, 1995, pp. 1–66.
- [18] P. Willett, J.M. Barnard, G.M. Downs, Chemical similarity searching, *J. Chem. Inf. Comput. Sci.* 38 (6) (1998) 983–996.
- [19] 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.
- [20] M.A. El Hassan, C.R. Calladine, Propeller-twisting of base-pairs and the conformational mobility of dinucleotide steps in DNA, *J. Mol. Biol.* 259 (1996) 95–103.
- [21] B. Schneider, S. Neidle, H.M. Berman, Conformations of the sugar-phosphate backbone in helical DNA crystal structures, *Biopolymers* 42 (1997) 113–124.
- [22] 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.
- [23] M.M.W. Mooren, On nucleic acid structure analysis by NMR. PhD thesis, University of Nijmegen, The Netherlands, 1993.
- [24] G.G. Privé, U. Heinemann, S. Chandrasegaran, L.S. Kan, M.L. Kopja, R.E. Dickerson, Helix geometry, hydration and G·A mismatch in a B-DNA decamer, *Science* 238 (1987) 498–504.
- [25] Anonymous, Definitions and nomenclature of nucleic acid structure parameters, *J. Mol. Biol.* 205 (1989) 787–791.
- [26] H. Hove, Y. Liang, O.M. Kvalheim, Trimmed object projections: a nonparametric robust latent-structure decomposition method, *Chemom. Intell. Lab. Syst.* 27 (1995) 33–40.
- [27] W.J. Egan, S.L. Morgan, Outlier detection in multivariate analytical chemical data, *Anal. Chem.* 70 (1998) 2372–2379.