

# Circular effects in representations of an RNA nucleotides data set in relation with principal components analysis

T.H. Reijmers, R. Wehrens, L.M.C. Buydens\*

*Laboratory for Analytical Chemistry, University of Nijmegen, 6525 ED Nijmegen, Netherlands*

Received 4 September 2000; accepted 12 February 2001

---

## Abstract

During the last few years, the main reason for using molecular structure databases has changed. Instead of using databases as a storage medium, databases now are also used as a source for data-mining applications. The large number of objects and variables in these databases induced that besides univariate techniques, multivariate techniques are also applied to search for knowledge hidden in the data. A popular multivariate technique that is used to explore the underlying structure in data is called principal component analysis (PCA). Because structure data are often represented as torsion angles and PCA is not originally designed to deal with this kind of circular data, the outcome of PCA experiments can be misleading. This article describes several alternative representations of circular data and its effect on the outcome of PCA experiments. A worked example is given using a database of RNA nucleotides. © 2001 Elsevier Science B.V. All rights reserved.

*Keywords:* PCA; Data mining; RNA nucleotides; Multivariate analysis

---

## 1. Introduction

The increasing popularity of the internet has resulted in an enormous growth of the size and usage of publicly available molecular structure databases. Apart from storage and reference purposes, it is also realized nowadays that these databases contain hidden information. Especially, the discovery and investigation of relations between the objects and variables in the data set are important. Are there any trends visible in the data? Can the data be clustered into groups of objects belonging to the same class? Do similar objects/variables in the data show similar behaviour? The new field of data-mining explores

ways to extract this kind of information from the data. Often, instead of Cartesian coordinates, torsion angles are used to represent the molecular structures. The resulting reduction in dimensionality is justified because bond lengths and angles often do not change significantly in the structures. An additional advantage of the torsion angles representation over the Cartesian coordinates representation is that no alignment is necessary to be able to compare different structures. Mainly concentrating on protein structures, this has resulted in many different database studies, focussing on the analysis of the distributions of individual torsion angles (e.g. Refs. [1–4]). Because the backbone of a nucleic acid (DNA, RNA) contains more torsion angles than the backbone of a protein (six torsion angles in a nucleotide against three in an amino acid), the univariate approach is no

---

\* Corresponding author.

*E-mail address:* L.Buydens@sci.kun.nl (L.M.C. Buydens).

longer sufficient to give a complete picture of the structure, although interesting results have been obtained (e.g. Refs. [5–11]). Recently, besides univariate techniques, multivariate techniques have also been applied (e.g. Refs. [12–15]).

Principal component analysis (PCA) is a popular multivariate technique that is often used to visualize the underlying structure of a data set [16]. Beckers and Buydens [12] used PCA to derive additional information on a data set consisting of the torsion angles of DNA dinucleotides. A low-dimensional map of the data was constructed that could be used for DNA classification. However, problems may arise when PCA is offered data for which the usual definition of variance does not apply. In the case of circular data, e.g. torsion angle data, PCA can give wrong results because distances between two circular data objects are defined differently as is the case for regular noncircular data. This is illustrated in Fig. 1 where for both noncircular and circular examples, the distances between two objects are visualized.

This article describes several different representations of circular data and examines the effect on the outcome of PCA experiments. The aim is to identify which representation allows standard PCA techniques to be used, and what types of artefacts can be expected in cases where an inappropriate representation has been selected. To observe the effect of these different representations, all results are compared with a study performed on the same data set with noncircular data, i.e. Cartesian coordinates. All experiments are performed using a data set that consists of RNA sequence fragments.

## 2. Experimental

### 2.1. Data

To examine how the PCA results are influenced by the circularity of the data, a subset is selected from the database used in Ref. [7]. The original set consists of 1480 RNA mononucleotides extracted from 52 RNA sequences, originated from the nucleic acid (NDB) [17] and protein (PDB) [18] databases. In Fig. 2A, the mononucleotides are visualized by means of the two pseudo torsion angles  $\eta$  and  $\theta$ , defined by the atoms  $C4'_{N-1}-P_N-C4'_N-P_{N+1}$  and  $P_N-C4'_N-P_{N+1}-C4'_{N+1}$ , respectively (see Fig. 3). On the basis of Fig. 2A and some manual verifications, Duarte and Pyle [7] defined eight different RNA classes: helical, C2 bend, base twist, chi switch, cross-strand stack, flip turn, stacked turn, and stack switching (see Fig. 2B). It should be noted that the differences between the clusters are small and that, to some degree, cluster border definition is arbitrary. Nevertheless, the clustering by Duarte and Pyle is retained, mainly for visualizing the effect of the different object representations. To bring the size of the data set to manageable proportions and to remove the bias towards helices (over 70% of the data set consist of helical objects), a subset is selected, so that each class would be represented by at least five objects. Because for the determination of pseudo torsion angles  $\eta$  and  $\theta$ , information on the neighbouring nucleotides is also taken into account, trinucleotides are taken as objects in the subset.

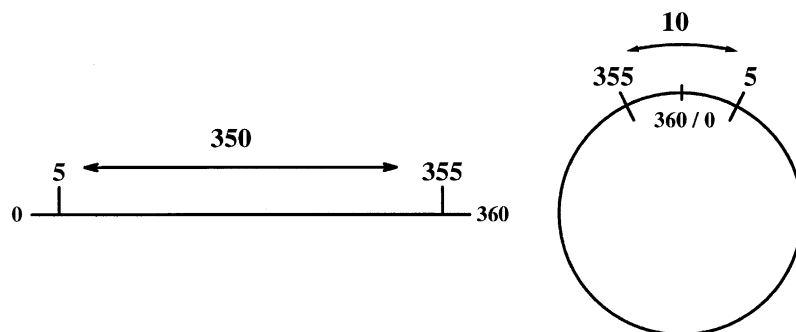


Fig. 1. For both noncircular (left) and circular examples (right) the distance between two objects is visualized.

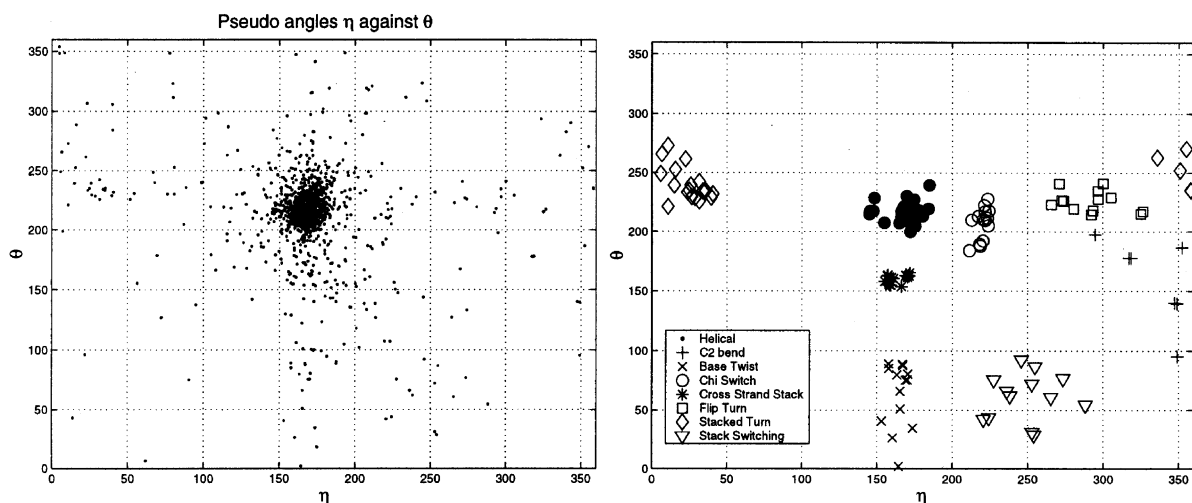


Fig. 2. On the left (A), an overview is given of the complete molecular data set of Duarte and Pyle [7] of the RNA structures in the  $\eta$  and  $\theta$  pseudo torsion space. On the right (B), eight subsets representing eight different RNA classes of the same data set are depicted.

The variables of the data set consist of the six backbone torsion angles ( $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $\delta$ ,  $\varepsilon$ , and  $\zeta$ ) of each nucleotide (see Fig. 3). This leads to a data set of 121 objects (trinucleotides), each described by 18 variables ( $3 \times 6$  torsion angles).

## 2.2. Methods

### 2.2.1. Analysis of circular data

The potential effects of the circularity of the data on the PCA procedure depend on the representation of the torsion angles. Besides the most often used representations, where the torsion angles have values

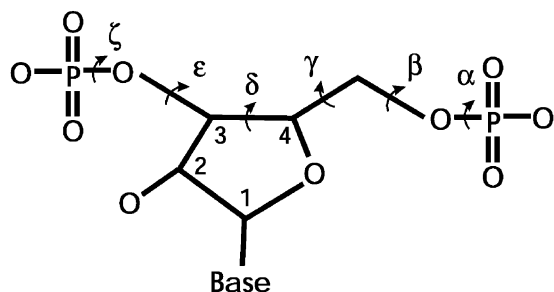


Fig. 3. Molecular structure of an RNA nucleotide with the corresponding torsion angles. The pseudo torsion angles are defined by the C4 and phosphor atoms of this nucleotide and its neighbours.

between  $0^\circ$  and  $360^\circ$  (representation I) and  $-180^\circ$  and  $180^\circ$  (representation II), respectively, three other representations are considered. The first alternative representation is already described in Ref. [19]. If each torsion angle is represented by two values, the sine and cosine values of that torsion angle (representation III), the data are no longer circular. A disadvantage of this representation is that the nonlinear sine/cosine transformation can cause problems in the subsequent analysis. Also, the doubled number of variables is not favourable for this representation. Finally, if this representation is analysed by PCA, the technique will treat the sine and cosine variables independently of each other, despite the fact that these variables are strongly related. A similar alternative representation would be the complex notation ( $\sin(\text{angle}) + i \times \cos(\text{angle})$ ) of each torsion angle. By using this representation, PCA is forced to deal with a combination of the sine and cosine variables. A problem of this complex representation is that PCA also returns complex results and interpretation is rather difficult. Therefore, these results will not be shown here.

The other representations that are investigated here is to apply PCA to either a correlation matrix (representation IV) or a variance/covariance matrix (representation V), where circularity is taken into account while calculating these matrices. Instead of the

regular correlation coefficient, the circular correlation coefficient as described in Ref. [20] is used:

$$r_{ab} = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^n \sin(a_i - a_j) \sin(b_i - b_j)}{\sqrt{\sum_{i=1}^{n-1} \sum_{j=i+1}^n \sin^2(a_i - a_j) \sum_{i=1}^{n-1} \sum_{j=i+1}^n \sin^2(b_i - b_j)}}, \quad (1)$$

wherein  $r_{ab}$  is the circular correlation coefficient between vectors  $\mathbf{a}$  and  $\mathbf{b}$  with length  $n$ . To calculate the variance/covariance matrix for circular data, the circular mean [20] is used:

$$\tan \bar{a} = \frac{\sum_{i=1}^n \sin a_i}{\sum_{i=1}^n \cos a_i}. \quad (2)$$

The importance of using the right definition for calculating the mean for circular data is shown in Fig. 4, where the construction of the circular mean for three objects is visualized. If the objects cover less than half of the full scale (ranging from  $0^\circ$  to  $360^\circ$ ), the circular mean differs not from the regular mean. However, if the objects cover more than half of the scale, like the objects in Fig. 4, the means differ and the circular mean should be used. To calculate circular (co)variances, differences larger than half of the scale used are given an additional subtraction of  $180^\circ$ ,

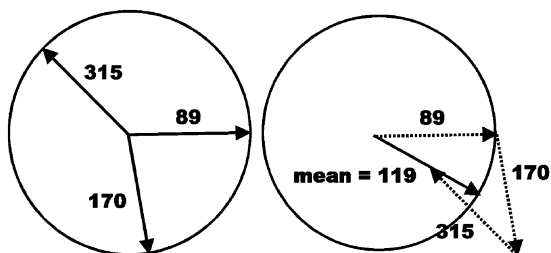


Fig. 4. Visualization of the construction of the circular mean for three objects with values: 89, 170 and 315. The noncircular mean value of the three objects is 191.

because the difference between two circular objects cannot be more than half of the scale used (i.e.  $360^\circ$ ).

### 2.2.2. Analysis of reference (noncircular) data

To see whether changes in the representation of the data resulted in better handling of circularity of the data, all results of the PCA experiments are compared with a so-called reference plot. The reference plot is created using the same objects as in the circular data set; however, now, Cartesian coordinates (noncircular variables) are applied. Starting point of this plot is a data set consisting of the 3D coordinates of the same trinucleotides as in the torsion angles data set. For all possible pairs of trinucleotides, the backbone atoms are translated and rotated by means of Procrustes analysis [21,22] in such a way that maximum overlap between the involved structures is obtained. A similarity measure is calculated by summation of all squared distances between the backbone atoms of the two trinucleotides. This results in a squared distance matrix with size  $121 \times 121$ . By using principal coordinates analysis (PCoA), the distance matrix can be visualized in a 2D plot [23]. The objects are placed in a low-dimensional space in such a way that the majority of the distances between the objects are kept intact (see Fig. 5). Notice that, although some clustering can be detected for several classes (especially the helical, cross-strand stack and stacked turn classes), Fig. 5 differs significantly from Fig. 2. This is not surprising because the Cartesian coordinates representation considers structure information on a different structure level than the pseudo torsion angle representation, i.e. a complete picture of the backbone that describes each backbone atom, against a more global picture of the backbone considering only several backbone atoms. This will be more thoroughly examined in the Results and Discussion section.

Ultimately, all PCA score plots resulting from the different representations, used in the torsion angles data set (circular data), are compared with the reference plot, obtained from the 3D coordinates data set (noncircular data). The representation corresponding to the most similar score plot deals best with the circular data. Either a 2D or 3D reference plot is used. These explain 73% and 84% of the variance in the data set, respectively.

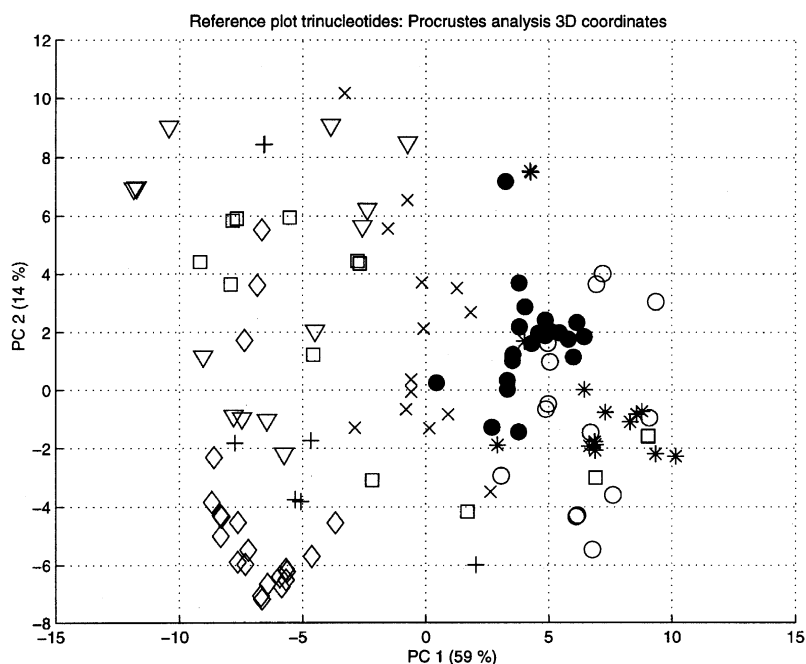


Fig. 5. Principal coordinates analysis plot of the dissimilarity matrix of the 121 aligned RNA nucleotides. Cartesian coordinates are used as representation. The same symbols are used as in Fig. 2B.

### 2.2.3. The comparison of circular and noncircular data

Because visual inspection of the plots does not give unambiguous results, usage is made of two statistical techniques. By means of another Procrustes analysis, the objects in the two plots that have to be compared are translated and rotated in such a way that optimal overlap between the corresponding objects is obtained. After summation of the squared distances between the aligned objects, a measure of similarity can be assigned to the two plots. One difficulty of this approach is that the scale of the compared plots should be the same to get a valid distance measure. To accomplish this, all plots are range-scaled before Procrustes analysis. In this way, the possibility exists that the internal distances are disturbed, so this approach of comparing two score plots is only used to get a first impression.

A more thorough approach to compare the scores is based on canonical correlation analysis [23,24]. Canonical correlation analysis is a multivariate method that searches for a linear combination of variables (canonical variables) in one data set that

yields the highest multiple correlation with a linear combination of variables in the other data set. The higher the multiple correlation coefficient, the more the compared data sets are related. An important advantage of this method is that no scaling has to be used and that the compared data sets may consist of different numbers of variables. However, the number of objects should be equal in both data sets.

All calculations are performed using Matlab for Unix Workstations, version 5.3, by Mathworks.

## 3. Results and discussion

The first step in any statistical analysis is to look at the data. In Fig. 6, an overview of the backbone torsion angles values for each trinucleotide in the data set is given. No distinction is made whether the torsion angle is situated in the first, second or third nucleotide of the trinucleotide. Fig. 6A shows the torsion angles ranging from  $0^\circ$  to  $360^\circ$ , while in Fig. 6B the torsion angles have values between  $-180^\circ$  and  $180^\circ$ . Both plots give a first indication to what extent

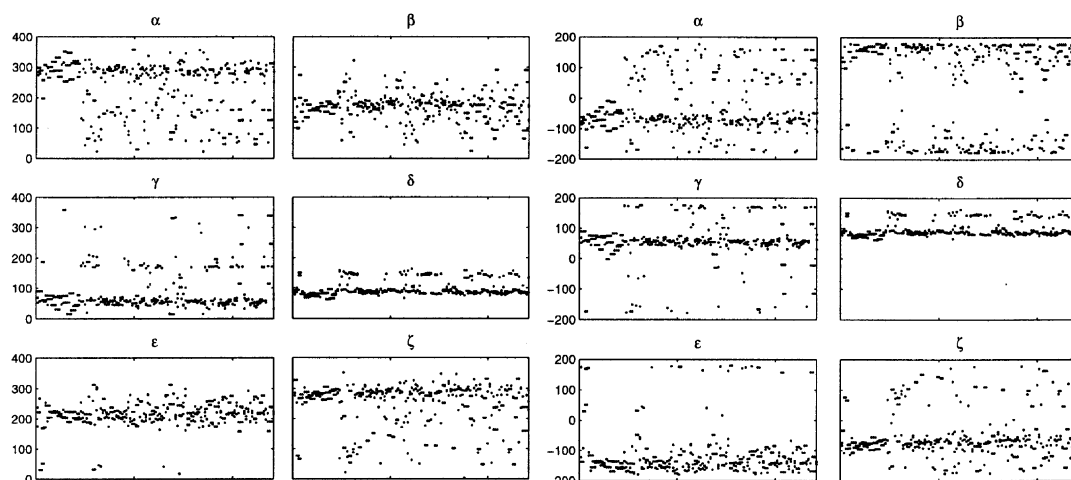


Fig. 6. Overview of the six types of backbone torsion angles for each trinucleotide. On the left, the torsion angles have values between  $0^\circ$  and  $360^\circ$ , while on the right the values range from  $-180^\circ$  to  $180^\circ$ .

circularity manifests itself in the data. For regular data plots, the maximum observed distance between two objects is the difference between the highest and the lowest possible values. For circular data, the highest possible distance is half the maximum distance as observed for noncircular/regular data (i.e.  $180^\circ$ ). Now, if a plot like Fig. 6 contains many objects separated further than the maximal distance (objects close to the minimum and maximum ranges of the

data), a technique that does not take into account the circularity of the data will give bad results, so another representation should be preferred. Torsion angles for which this may lead to problems are  $\alpha$ ,  $\gamma$  and  $\zeta$  in Fig. 6A. Fig. 6B shows an extreme example of this kind of behaviour for backbone torsion angle  $\beta$ , whereas  $\epsilon$  and  $\zeta$  also may cause some problems.

Fig. 7 shows the results of the PCA on the median-centered nucleotides using the  $0^\circ$  to  $360^\circ$  torsion

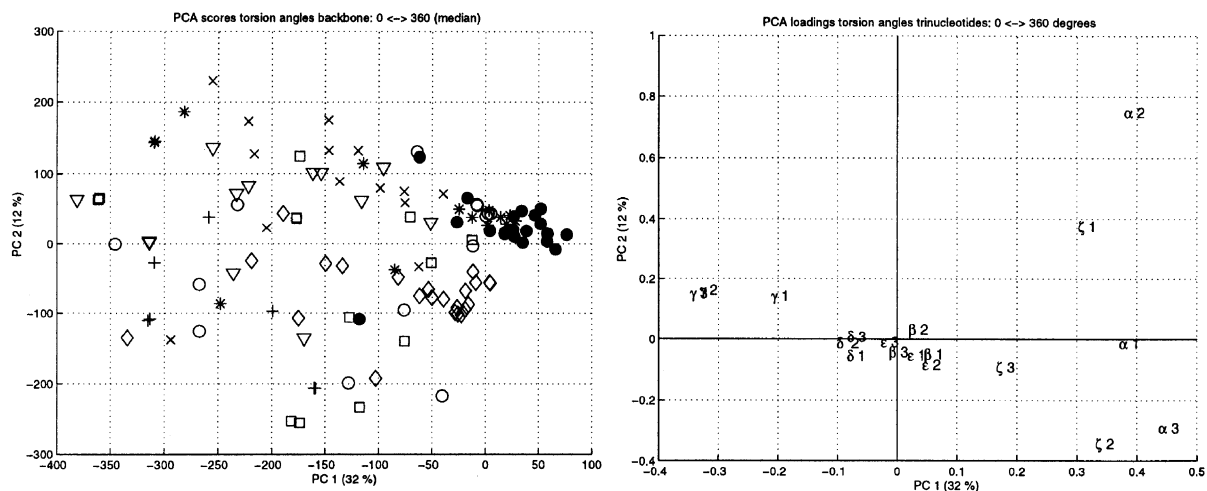


Fig. 7. Results of application of PCA on median-scaled torsion angles ranging from  $0^\circ$  to  $360^\circ$  (representation I). On the left, the score plot of the first two principal components are depicted. On the right, the corresponding loadings are visualized. Symbols as in Fig. 2.

angle representation (representation I). On the left, the scores of the first two principal components are visualized, while on the right, the corresponding loadings are depicted. Because PCA maximizes the amount of explained variance of the total data set for the determination of the first principal components, variables that exhibit a wide range of values have a large influence on the first principal components. Fig. 6A shows that this is the case for backbone torsion angles  $\alpha$ ,  $\gamma$  and  $\zeta$ . As a consequence, these torsion angles have high loadings on the first few principal components (see loadings plot Fig. 7). In the score plot, no separate clusters for the eight classes can be detected. As in the reference plot (Fig. 5), the helical objects (the filled circles) are situated close together. The same is true for the stacked turn objects (diamonds). Other groups seem to be more scattered than in the reference plot.

In Fig. 8, the results of the PCA using the  $-180$ – $180^\circ$  representation torsion angle data set are given (representation II). Again, on the left, the score plot and on the right, the loadings plot are depicted. These plots show that the variable representation has a big influence on the outcome of the PCA results. The difference between the results depicted in Figs. 7 and 8 is mainly caused by torsion angle  $\beta$ . While for the  $0 \leftrightarrow 360$  representation, the variance of the  $\beta$  torsion angle values is rather small (especially when compared with  $\alpha$ ,  $\gamma$  and  $\zeta$  distributions), the variance of

$\beta$  in the  $-180 \leftrightarrow 180$  representation is much larger (see Fig. 6). In both loadings plots, this is also clearly visible: whereas for the  $0 \leftrightarrow 360$  representation, the loadings for  $\beta$  in the first two principal components is rather small, for the  $-180 \leftrightarrow 180$  representation, the loadings for  $\beta$  are extremely large. This once more confirms the conclusion drawn on the basis of Fig. 6. The bad choice of representation for  $\beta$  also has a disastrous effect on the scores of the objects in the score plot. Clearly, two clusters, or better, ellipses, are formed, one in the upper and one cluster in the lower part of the figure. The helical and stacked turn trinucleotides that are more or less clustered in the score plot of Fig. 7 are now partly situated in both the upper and lower ellipses.

The results of the experiments described above clearly show that because PCA is applied to variances, it experiences problems if instead of regular data, circular data are processed. One simple solution for the problem could be to transform each variable separately until no perceptible circular effect is present in the data. For example, if the  $-180 \leftrightarrow 180$  representation data set is used, the representation of the  $\beta$  torsion angle could be replaced by the  $0 \leftrightarrow 360$  representation. The problem with this approach is that for large data sets, it is not feasible to test each variable separately, and because most databases are contaminated with outliers and errors, there is no guarantee that there is an acceptable transformation.

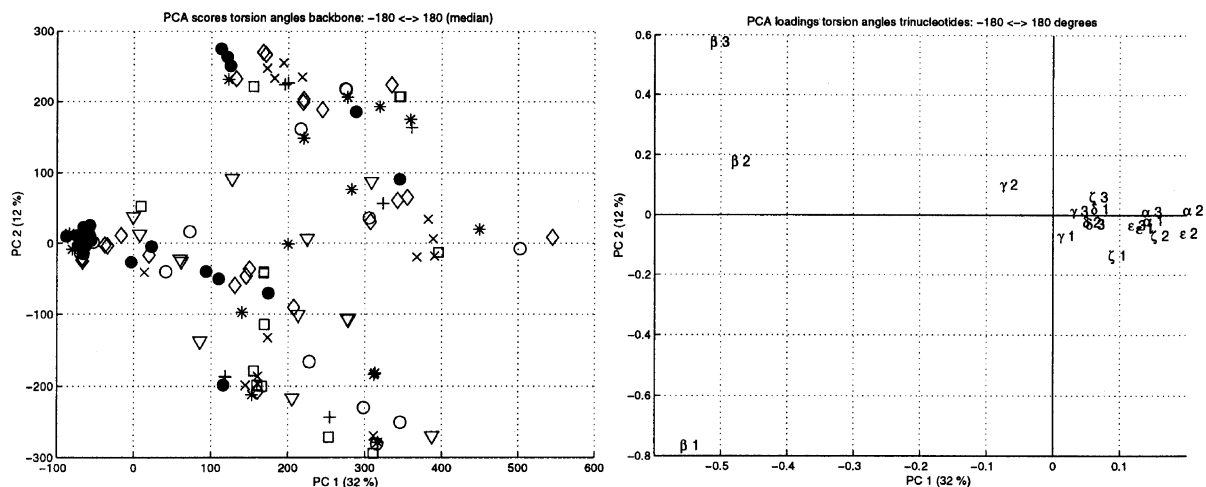


Fig. 8. Results of application of PCA on median-scaled torsion angles ranging from  $-180^\circ$  to  $180^\circ$  (representation II). On the left, the score plot of the first two principal components are depicted. On the right, the corresponding loadings are visualized. Symbols as in Fig. 2.

Therefore, it is better to look for other solutions. In the next part, the influence of three other representations of the original data set on the outcome of the PCA results will be discussed. The representations are: the sine and cosine representation (representation III), the circular correlation matrix representation (representation IV) and the circular variance/covariance matrix representation (representation V), respectively. More information about the exact definition and calculation of these representations can be found in the Experimental section.

In Table 1, the result shows the comparison of the different PCA score plots with the reference plot using Procrustes analysis. For each different representation of the torsion angle data set, the dissimilarity with the coordinates data (see Fig. 5) is calculated for two and three principal components, respectively. Low numbers indicate a good agreement.

The conclusions drawn earlier from Figs. 6–8 are confirmed when the distances of representation II are compared with the distances of representation I. When only two dimensions of the score plots are considered, representation V gives the best and representation I the second best results. If three principal components are considered for the reference plot, the circular variance/covariance matrix representation is still the best (the summed distances are only marginally larger than those for two principal components); however, now, second best is representation III. This is probably caused by the fact that this representation is composed of twice as much variables and, as a consequence, more principal components are needed to describe the underlying structure. The circular correlation matrix representation can be

Table 1

Comparison of 2D (first row) and 3D (second row) reference plots with different torsion angle representations by means of Procrustes analysis

Number of PCs	Data representation				
	I	II	III	IV	V
2	4.4511	5.1628	4.6851	4.5622	4.3363
3	4.8882	6.0399	4.8658	4.9848	4.3447

Numbers indicate dissimilarities. Representations: I =  $0^\circ$ – $360^\circ$ ; II =  $-180^\circ$ – $180^\circ$ ; III = sine/cosine representation; IV = circular correlation matrix representation; V = circular variance/covariance matrix representation.

Table 2

Comparison of the 2D reference plot and different torsion angle representations with canonical correlation analysis

Number of PCs	Data representation									
	I	II	III	IV	V					
1–2	0.18	0.05	0.09	0.00	0.17	0.04	0.07	0.02	0.13	0.08
1–3	0.18	0.06	0.09	0.02	0.18	0.14	0.11	0.02	0.37	0.12
1–4	0.19	0.13	0.17	0.09	0.20	0.16	0.12	0.05	0.37	0.17
1–5	0.19	0.15	0.18	0.14	0.32	0.19	0.12	0.05	0.38	0.18
1–6	0.37	0.18	0.18	0.15	0.35	0.19	0.33	0.12	0.42	0.18

Numbers indicate the squared multiple correlations for the first and second canonical variables, respectively. Representations as in Table 1.

considered as a cross-product matrix of auto-scaled data, and the circular variance/covariance matrix representation as the cross-product of mean-scaled data. Because some of the variables in the original data set are nonnormally distributed, mean scaling can be expected to retain the data structure better than auto-scaling [12]. This is also seen in Table 1: representation IV gives inferior results in comparison with representation V. For reasons already mentioned in the Experimental section, the results in Table 1 should be used only to get a first impression of the suitability of the experimented representations.

Tables 2 and 3 reflect the results of the canonical correlation analysis. The score plots of each representation of the torsion angle data set is subsequently compared with the 2D (Table 2) and the 3D (Table 3) reference plots. In these calculations, the dimensionalities of the torsion angle score plots are varied. For example, the last row of both tables shows the results where canonical correlation analysis searches for the combination of the first six principal components of the torsion angle data sets that displays the highest correlation with a combination of the first two (Table 2) or three (Table 3) components of the reference data set. Apart from the squared multiple correlation of the first canonical variable, the squared multiple correlation is also listed for the second (Tables 2 and 3) and third canonical variables (Table 3).

The first row in Table 2 shows that when the 2D reference plot is compared with the 2D torsion score plots, the  $0 \leftrightarrow 360$  and sine and cosine representations give the best results. With so few variables, the second canonical variables show only very small

Table 3

Comparison of the 3D reference plot and different torsion angle representations with canonical correlation analysis

Number of PCs	Data representation														
	I			II			III			IV			V		
1–3	0.38	0.14	0.04	0.09	0.03	0.00	0.49	0.18	0.07	0.42	0.08	0.00	0.54	0.34	0.09
1–4	0.46	0.19	0.06	0.20	0.09	0.01	0.52	0.19	0.08	0.44	0.08	0.03	0.55	0.34	0.14
1–5	0.47	0.19	0.10	0.29	0.18	0.07	0.52	0.27	0.18	0.46	0.08	0.03	0.56	0.36	0.16
1–6	0.52	0.37	0.16	0.34	0.18	0.11	0.55	0.33	0.19	0.47	0.33	0.08	0.56	0.41	0.16

Numbers indicate the squared multiple correlations for the first, second and third canonical variables, respectively. Representations as in Table 1.

multiple correlations for all representations. The multiple correlation of representation V becomes much larger if instead of two, three principal components are considered (second row of Table 2). The squared multiple correlations for the remaining representations do not change much, except for representation IV, which is no longer the worst performing representation. From further increase of the number of principal components that are taken into account during the comparisons with the reference plot, it can be concluded that the circular variance matrix representation deals best with the circularity of the data. On the whole, the sine and cosine representation performs second best, followed by the  $0 \leftrightarrow 360$  representation. The bad results for the  $-180 \leftrightarrow 180$  rep-

resentation in Table 2 again confirm the conclusions drawn previously.

In Table 3, the same trends can be detected: representation V performs best, followed by representations III, I, IV and II, respectively. For nearly all representations, no large increase in the squared multiple correlation of the first canonical variable is found when the dimensionality of the torsion plot is increased. Probably, all relevant information present in the 3D reference plot can be explained by the first three principal components of the PCA. This certainly is not the case for the reference plot in Table 2.

Fig. 9 shows two score plots where the circular variance/covariance matrix representation is used to

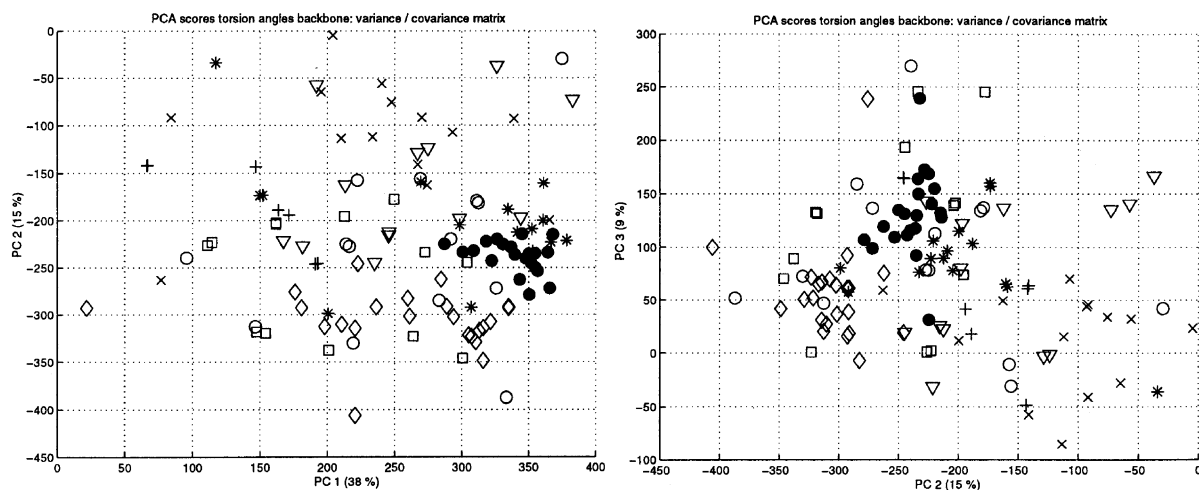


Fig. 9. Results of the application of PCA on torsion angles using the circular variance/covariance representation. On the left, the score plot of the first two principal components is depicted, while on the right the scores of principal components two and three are visualized. Symbols as in Fig. 2.

get a low-dimensional picture of the trinucleotide torsion angle data set. On the left, the trinucleotides are depicted using the first two principal components. This plot does not differ much from the score plot of the  $0 \leftrightarrow 360$  representation in Fig. 7. The results in Tables 1 and 2 already suggested this. Both tables also indicate that the difference between the representations becomes clearer after taking an extra dimension into account. Therefore, on the right of Fig. 9, principal components two and three are used to visualize the objects for the circular variance/covariance matrix representation. In both plots, the helical and stacked turn objects are nicely clustered, whereas the scatter of the other groups is quite substantial. In spite of the fact that the torsion angle data set is chosen in such a way that it contains as much as possible the same information as the pseudo torsion angle data set, Figs. 2 and 9 are totally different. Apparently, application of PCA to the trinucleotide torsion angles data set leads to different information, resulting in a divergent positioning of the objects compared to the objects in Fig. 2.

Table 4 shows the results of the canonical correlation analysis using the Cartesian coordinates reference and the pseudo torsion angles plots, where the pseudo torsion angles are represented as values between  $0^\circ \leftrightarrow 360^\circ$  and  $-180^\circ \leftrightarrow 180^\circ$ . While the  $0 \leftrightarrow 360$  representation of the pseudo angles gives a correlation for the 2D reference plot similar to the torsion angles in Table 2, a much larger correlation is found for the 3D reference plot in comparison with the regular torsion angles data (Table 3) for the first canonical variable. Again, the correlation for the second canonical variable is much lower and the  $-180 \leftrightarrow 180$  representation performs badly. On the basis of

Table 4, the conclusion can be made that, because of the higher squared multiple correlation values, the pseudo angle torsion plot is closer to the reference plot than the torsion angle plots. This seems to support the statement made by Duarte and Pyle [7] that torsion angles, in some cases, provide too detailed information and a more simple representation should be used.

A disadvantage of both the reference and the pseudo torsion angles plots is that some subjective choices have to be made before the plots can be generated. For the pseudo torsion angles plot, several definitions for pseudo torsion angles describing the rough structure of the RNA backbone are possible. The definition for the pseudo torsion angles  $\eta$  and  $\theta$  was obtained after trying some and keeping the best one. In the case of the reference plot, freedom exists in the choice of the atoms, which will be used in the Procrustes analysis and the atoms that are used to calculate a similarity measure between the aligned structures. An additional disadvantage of the 3D coordinates plot is that when new trinucleotides are being added to the data set, some calculations have to be repeated. Besides the alignment experiments of the new trinucleotide with the other trinucleotides in the data set, the principal coordinates analysis has to be also repeated using the expanded distance matrix. For both the torsion and pseudo torsion angle plots, the new trinucleotide can be projected in the already existing torsion angle space.

All low-dimensional plots have their own properties and, therefore, their own positioning of the RNA sequences. Further research on this topic is in course.

#### 4. Conclusion

Because most molecular structure databases contain many related variables, multivariate techniques, like PCA, can give good results when information hidden in the database has to be discovered. However, the way in which the molecular structure data are represented and used in PCA must be chosen carefully. Circular data may behave unexpectedly, when PCA is applied. The aim of this article, to find a representation for the circular data that does not affect PCA negatively, has been achieved. From the

Table 4

Comparison of different representations of the pseudo torsion angles with the reference plot (2D: first row, 3D: second row) by means of canonical correlation analysis

Reference plot	Data representation pseudo torsion angles data			
	$0 \leftrightarrow 360$		$-180 \leftrightarrow 180$	
2D	0.16	0.01	0.07	0.01
3D	0.65	0.06	0.14	0.07

Numbers indicate the squared multiple correlations for the first and second canonical variables, respectively.

experiments described above, it is concluded that torsion angles can best be represented by means of a variance matrix that takes into account the circular character of the molecular structure variables. Together with this appropriate representation, some techniques are presented to examine the influence of the representation on the outcome of PCA experiments. If a reference plot is available, canonical correlation analysis appears to be the most valid technique to examine the behaviour of the representation used.

## References

- [1] S.R. Griffiths-Jones, G.J. Sharman, A.J. Maynard, M.S. Searle, *J. Mol. Biol.* 284 (1998) 1597–1609.
- [2] K. Gunasekaran, C. Ramakrishnan, P. Balam, *J. Mol. Biol.* 264 (1996) 191–198.
- [3] L. Serrano, *J. Mol. Biol.* 254 (1995) 322–333.
- [4] N.J. West, L.J. Smith, *J. Mol. Biol.* 280 (1998) 867–877.
- [5] H.M. Berman, *Biopolymers* 44 (1997) 23–44.
- [6] M.A. El Hassan, C.R. Calladine, *J. Mol. Biol.* 259 (1996) 95–103.
- [7] C.M. Duarte, A.M. Pyle, *J. Mol. Biol.* 284 (1998) 1465–1478.
- [8] V.L. Murthy, R. Srinivasan, D.E. Draper, G.D. Rose, *J. Mol. Biol.* 291 (1999) 313–327.
- [9] M.J. Packer, M.P. Dauncey, C.A. Hunter, *J. Mol. Biol.* 295 (2000) 71–83.
- [10] M.J. Packer, M.P. Dauncey, C.A. Hunter, *J. Mol. Biol.* 295 (2000) 85–103.
- [11] B. Schneider, S. Neidle, H.M. Berman, *Biopolymers* 42 (1997) 113–124.
- [12] M.L.M. Beckers, L.M.C. Buydens, *J. Comput. Chem.* 19 (1998) 695–715.
- [13] M.L.M. Beckers, W.J. Melssen, L.M.C. Buydens, *J. Comput.-Aided Mol. Des.* 12 (1998) 53–61.
- [14] M.J. Packer, C.A. Hunter, *J. Mol. Biol.* 280 (1998) 407–420.
- [15] S. Wold, J. Jonsson, M. Sjöström, M. Sandberg, S. Rännar, *Anal. Chim. Acta* 277 (1993) 239–253.
- [16] J.E. Jackson, *A User's Guide to Principal Components*. Wiley, New York, 1991.
- [17] H.M. Berman, W.K. Olson, D.L. Beveridge, J. Westbrook, A. Gelbin, T. Demeny, S.H. Hsieh, A.R. Srinivasan, B. Schneider, *Biophys. J.* 63 (1991) 751–759.
- [18] 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, *J. Mol. Biol.* 112 (1977) 535–542.
- [19] L.M.C. Buydens, T.H. Reijmers, M.L.M. Beckers, R. Wehrens, *Chemom. Intell. Lab. Syst.* 49 (1999) 121–133.
- [20] J.H. Zar, *Biostatistical Analysis*. Prentice-Hall, London, 1996.
- [21] J.C. Gower, *Psychometrika* 40 (1975) 33–51.
- [22] J.M.F. ten Berge, *Psychometrika* 42 (1977) 267–276.
- [23] B.G.M. Vandeginste, D.L. Massart, L.M.C. Buydens, S. de Jong, P.J. Lewi, J. Smeyers-Verbeke, *Handbook of Chemometrics and Qualimetrics*. Elsevier, Amsterdam, 1998.
- [24] H. Hotelling, *Biometrika* 28 (1936) 321–327.