Pergamon

PII: S0097-8485(97)00041-7

Computers Chem. Vol. 21, No. 6, pp. 377–390, 1997

Printed in Great Britain
0097-8485/97 $17.00 + 0.00

A self-organizing feature map for clustering nucleic acids Application to a data matrix containing A-DNA and B-DNA dinucleotides

M. L. M. Beckers, W. J. Melssen and L. M. C. Buydens*

Laboratory for Analytical Chemistry, Faculty of Science, University of Nijmegen, Toernooiveld 1, 6525 ED, Nijmegen, The Netherlands

(Received 24 March 1997; Accepted 9 June 1997)

Abstract—A self-organizing feature map to cluster DNA dinucleotides is presented. During a training session 244 training patterns, each consisting of nine torsion angles, are clustered in a 10 by 10 map. The method is successful for separating the four known DNA classes in the training set. Contour plots of the weights after a training session indicate gradients in torsion angles corresponding to class separation. Moreover, certain units in the map probably correspond to unfavourable torsion angle combinations resulting in, e.g. van der Waals clashes. Hence, although no direct relation to a conformation's energy (as in a Ramachandran plot) is present in the map, it may provide a multidimensional interpretation of accessible and forbidden areas for dinucleotides. The applicability of the method on this DNA data matrix shows its potential to be used in more extensive structural analysis studies, e.g. in a case of comparing DNA with RNA. Several test patterns resulting from molecules with unusual structural characteristics are identified with the map. Copyright © 1998 Elsevier Science Ltd

Key words: self-organizing feature map, clustering, nucleic acids

1. INTRODUCTION

Most biomacromolecules, such as proteins and nucleic acids, occur in preferred conformations. Examples include α-helices and β-sheets in proteins and A-DNA, or various types of B-DNA, in nucleic acids. These preferred conformations are the basic keys for structural stability and biological activity of the molecules. It is therefore of biological importance to understand the relation between a preferred conformation and the responsible structural parameters. Which structural parameters can be used to study this relation? Possible candidates in the case of nucleic acids are helical parameters, such as, roll, twist and rise. For example, El Hassan and Calladine (1996) studied the relation between propeller-twisting of base pairs and the conformational mobility of dinucleotide steps in DNA. Torsion angles provide another possibility in the research of underlying principles for preferred conformations.

Beckers and Buydens (1998) used a data matrix with 244 objects derived from structures given in El Hassan and Calladine (1996). Each object is a dinucleotide which is represented by a vector of nine torsion angle values. The authors searched for multiple correlations between the torsion angles in the dinucleotides with multivariate analysis on the basis of singular value de-composition (SVD) (Beckers and Buydens, 1998). They demonstrated that score plots revealed a clear separation of DNA classes and provide a means of constructing multidimensional Ramachandran-like plots (Ramachandran *et al.*, 1963) for nucleic acids.

Traditional multivariate techniques, such as SVD methods, are often based on data-reduction by means of linear combinations of the original variables. These approaches assume some kind of parametric model. Artificial neural networks (ANNs) construct models without the inclusion of explicit parameters and are, therefore, often referred to as black boxes. However, they are an interesting alternative for traditional multivariate approaches in both calibration and classification problems (Rubner and Tavan, 1989; Melssen *et al.*, 1993; Smits *et al.*, 1993; Walczak and Wegscheider, 1993). Moreover, ANNs are capable of dealing with non-linearities.

* Author to whom correspondence should be addressed. Tel: + 31 24 3653192; Fax: + 31 24 3652652; E-mail: lbuydens@sci.kun.nl.

In this paper a self-organizing feature map (Kohonen, 1989; Kohonen, 1995) is introduced to reveal the structure of the 9-dimensional data matrix of Beckers and Buydens (1998). For this purpose the data matrix, referred to as the training set in this paper, is mapped onto a rectangular grid, holding 9-dimensional weight vectors, during the training phase of the network. The basic principle is quite simple. First of all, the map of neurons, or units, is constructed. Each unit contains a 9-dimensional vector of weights. These weights are initialised in a random manner. Then a 9-dimensional input vector, or pattern, is compared successively with each of the weight vectors. According to some measure of similarity, the unit with a weight vector that compares best with the input vector is assigned the winner. This process is repeated for all input vectors. In an iterative manner the weights are adapted in such a way that the topology of the data matrix is preserved in the best possible manner. After training, input patterns resembling each other will be located on the same unit, or units, in close proximity. Hence, a clustering of input patterns is accomplished by mapping a matrix of high-dimensional input patterns onto a grid of weight vectors.

Our aim was to show the applicability of this new method to the clustering of a large DNA data set. It is demonstrated that by means of a self-organizing feature map the DNA classes present in the training set are clearly separated. Moreover, by comparing the mean torsion angles of patterns that are clustered on each unit, some kind of gradient in individual torsion angles is observed that defines the DNA-class separation. Hence, the method has the potential to be used in detailed nucleic acid structural analysis studies, e.g. comparison of DNA with RNA, or the analysis of loop structures.

The gradient in individual torsion angles is seen even more clearly when the contour plots of the final weights per unit are studied. Weights reflect characteristics of patterns clustered on a certain unit. This is an advantage over SVD-like methods where the gradient between classes is not seen that clear. Moreover, on some units, no patterns are clustered at all.

Probably these units represent torsion angle combinations that would produce, e.g. too much sterical hindrance or other unfavourable structural aspects. Hence, these units may represent forbidden areas for dinucleotides. Although no direct relation between conformation and energy is present in the map this aspect strongly brings to mind the information that is represented in a Ramachandran plot.

The final weights of the Kohonen map are used to identify unknown patterns of several test sets. Most of the test sets contain one or more patterns of uncommon dinucleotide steps, e.g. bases that bulge out of a helix or bases that belong to a loop structure. The 244-pattern training set was constructed in such a manner that it covered most known (naturally occurring) DNA-class patterns. Therefore all phenomena present in the test sets can be explained with the 10 by 10 Kohonen map.

2. MATERIALS AND METHODS

2.1. Data Sets

2.1.1. Training set

Table 1 summarizes the data that was used in this study. It is based on the study of conformational mobility of dinucleotide steps by El Hassan and Calladine (all references, in which the crystal structures of sequences given in Table 1 were published, are given in El Hassan and Calladine (1996)). Each single strand sequence was subdivided into dinucleotide steps. We describe the dinucleotide steps by the nine torsion angles depicted in Fig. 1a. Hence, the training set contains nine columns or variables. The individual dinucleotides, represented by nine variables, are the rows, or patterns, of the data matrix. Dinucleotides containing the bases inosine and uracil as well as dinucleotides with bases in mismatched base pairs were not added to the data matrix. They are indicated in boldface in Table 1. Dinucleotides that had one or more torsion angle combinations in forbidden areas (see Mooren, 1993) were not added to the data matrix.

According to El Hassan and Calladine (1996) the dinucleotides derived from the structures depicted in Table 1 could be labeled as A-DNA or B-DNA.

Table 1. Sequence of the structures from which the dinucleotide steps were taken and corresponding helix types

Sequence*	Helix type
Dodecamers	
d(CCGTACGTACGG)	A
d(GCGTACGTACGC)	A
d(CGCIAATTAGCG)	B
d(CGCGAATTCGCG)	B
d(CGCAAATTTGCG)	B
d(CGCGAATTCGCG)	B
d(CGCAAATTCGCG)	B
d(CGCGAATTTGCG)	B
d(CGCAAATTIGCG)	B
d(CGCGAATTGGCG)	B
Decamers	
d(CGATCGATCG)	B
d(CGGTATACGC)	A
d(GCGTATACGC)	A
d(GCGTATACGC)	A
d(CGATCGATCG)	B
d(CCAGGCCTGG)	B
d(CCAGGCCTGG)	B
d(CCAACITTGG)	B
d(CCAACITTGG)	B
d(CCAAGATTGG)	B
d(CCAACGTTGG)	B
d(CGATTAATCG)	B
d(CGATATATCG)	B
Octamers	
d(GGIGCTCC)	A
d(GGGTACCC)	A
d(CCCCGGGG)	A
d(GCCCGGGC)	A
d(GCCCGGGC)	A
d(GGGGCTCC)	A
d(CTCTAGAG)	A
d(GGUAUACC)	A
d(GGGGTCCC)	A
d(GGGATCCC)	A
d(GGGGCCCC)	A
d(GTACGTAC)	A
Tetramer	
d(CCGG)	A

*Dinucleotides from mismatched base pairs as well as base pairs containing inosine and uracil were not taken into the data matrix and are indicated in boldface.

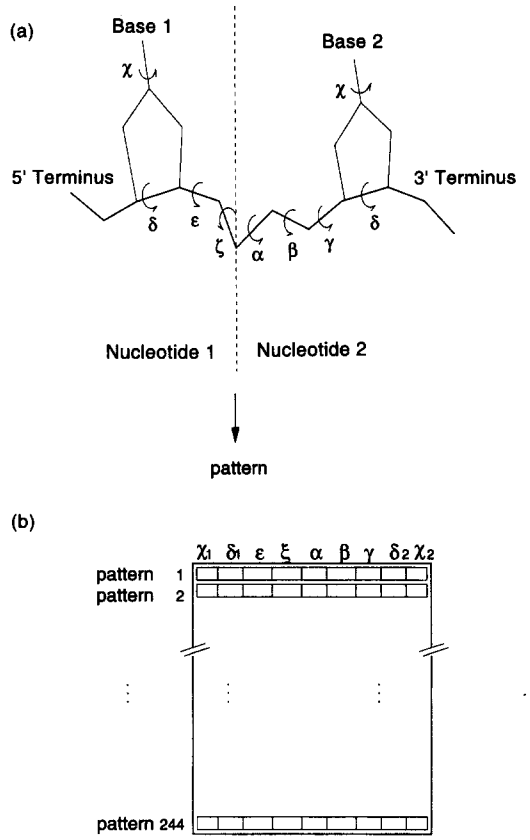

Fig. 1. (a) Torsion angle representation of a dinucleotide. The backbone conformation is represented by torsion angles α, β, γ, ϵ and ζ. The complete conformation of the furanose ring is not taken into account and, instead, it is represented by δ. The orientation of the base with respect to the sugar ring is given by torsion angle χ. (b) Training set with 244 patterns of dinucleotides represented by 9 torsion angles.

Subsequently we provided the B-DNA dinucleotides with additional labels B_I, with an $\epsilon(tr)/\zeta(g\text{-})$ combination, and B_{II}, having an $\epsilon(g\text{-})/\zeta(tr)$* combination, following the definition given in Privé *et al.* (1987). This resulted in 112 B_I dinucleotides and 32 B_{II} dinucleotides, respectively. The remaining 100 A-DNA dinucleotides contained four objects that were labeled as "crankshaft". In the small group of objects exhibiting the so-called crankshaft effect the combination $\alpha(g\text{-})/\gamma(g^+)$ is switched to $\alpha(tr)/\gamma(tr)$. Although crankshaft effects are most often reported for A-DNA, it appeared that one of the B_I dinucleotides (object 73) also had a $\alpha(tr)/\gamma(tr)$ combination. No data pre-treatment was performed on the resulting training set of 244 dinucleotides.

2.1.2. Test sets

To validate the classification of the Kohonen network we constructed several test sets (Table 2). Torsion angles and other information were extracted from the Nucleic Acid Database (NDB, (Berman *et al.*, 1992)).

For each of the sequences the first and last nucleotide were not taken into account. From the sequences marked with "†" the torsion angles of only one strand were used. Spermine was added to bdf062. This resulted in a base pair opening with the accompanying CCACCG strand. A novel non-Watson Crick hydrogen bonding scheme for a T·A base-pair was the result. bdfp24 has a chiral phosphorothioate linkage in the backbone instead of the usual phosphoro-di-ester. A loop structure is formed by self-association of complementary bases in udg028. Flipped-out bases and bulges concern bases that are directed towards the outside of a double-helix.

2.2. Self-organizing Feature Maps

In this study a Kohonen network, which belongs to the class of self-organizing feature maps, is used. Unlike most ANNs, Kohonen networks are trained unsupervised, i.e. input patterns are presented but no associated output patterns are offered. First a framework of units in a grid is set up. This is called the actual Kohonen map. To choose the grid size, corresponding to the number of units in this map, one can use the following rule of thumb:

$$2 \times \text{number of classes} < \text{number of units} \ll \text{number of patterns}$$

Then each unit, j, is assigned a n-dimensional vector of weights, $\vec{w}_j$. This procedure is depicted in Fig. 2a. The weight values are initialised in a random manner, provided that they lie between the minimum and maximum of the original variable values.

Now the network is ready for the training phase. A pattern, $\vec{x}_i$, is chosen randomly from the training set. It is compared with each of the weight vectors according to some similarity criterion, e.g. the Euclidean distance, see Fig. 2b. The unit with a weight vector, $\vec{w}_j$, most similar to the pattern, $\vec{x}_i$, is assigned the winner. According to a pre-defined neighbourhood criterion, some of the units surrounding the winner are also selected. In the next step the weight values of the winning unit and corresponding neighbourhood units are adapted according to equation (1):

$$\vec{w}_j(t+1) = \vec{w}_j(t) + \eta(t)N(t,r)[\vec{x}_i - \vec{w}_j(t)] \quad (1)$$

$\vec{w}_j(t)$ = the weight vector of unit j at iteration t
$\vec{x}_i$ = input pattern i
$\eta(t)$ = the learning rate at iteration t
$N(t,r)$ = neighbourhood function at iteration t.

The learning rate is decreased in time. The neighbourhood function shrinks in time and according to the distance r between the winning unit and other units in the map that need to be updated. In other words, the number of units of which the weight values are updated decreases in time and eventually only the weight vector of the winning unit is adapted. After all the input patterns have been matched, one cycle of training is completed. Training usually stops when weight vectors do not change significantly any more. A trained network contains

* The expressions in parentheses are used to indicate the corresponding torsion angle ranges, i.e. g^+: $60 \pm 60°$; tr: $180 \pm 60°$; $g\text{-}$: $300 \pm 60°$.

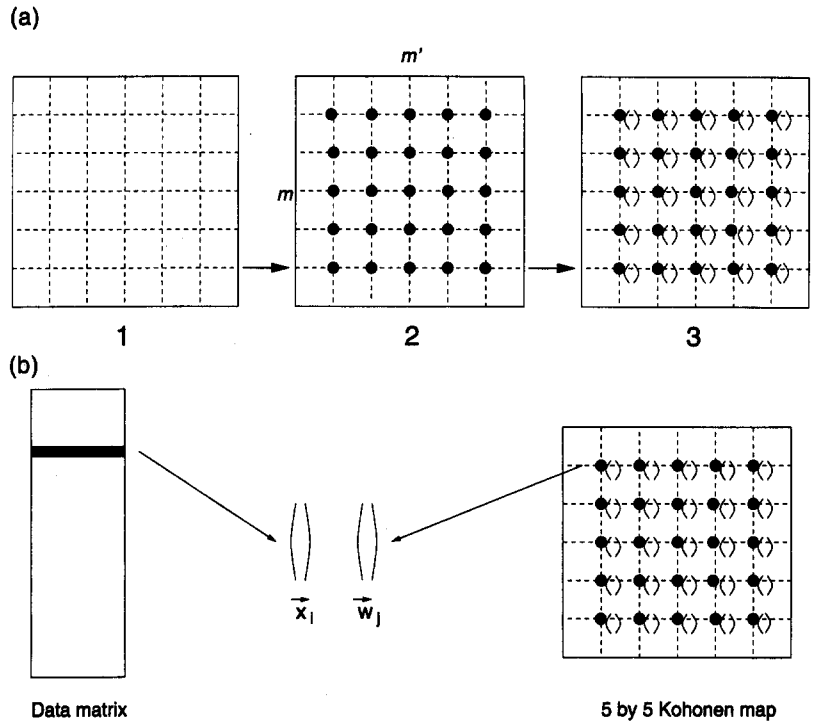

Fig. 2. (a) Initialization of m by m' Kohonen map; 1 — a grid of size, in this case, 5 by 5 is defined; 2 — the nodes represent units; 3 — each unit is assigned a n-dimensional weight vector. (b) Pattern $\vec{x}_i$ is compared with each of the weight vectors according to some similarity criterion.

weight vectors that approximate the distribution of the patterns in the original data matrix. In such a trained network each unit might be associated with an object class. In this way, the resulting map can be used for classification purposes.

3. EXPERIMENTAL

3.1. Configuration

3.1.1. Training session

As indicated in Section 2.1 we expected four classes of dinucleotides to be present in the training set. The number of patterns in this training set is 244. Therefore, we decided to build Kohonen maps ranging from 3 by 3 to 10 by 10. Results will be given for 5 by 5, 7 by 7 and 10 by 10 Kohonen maps. These maps were represented by a rectangular grid with the index of the units starting from 0.

The neighbourhood function was a so-called "bubble" function, i.e. a set of array points around the winning unit is chosen which decreases linearly in time. During the first 3000 cycles, the initial radius of the bubble function was 5. This radius was decreased to a radius of 3 for the following cycles. Training was allowed for a maximum of 100 000 cycles. During the initial training phase the starting value of the learning rate was 0.3 while this was 0.05 for the remaining part of the training phase.

3.1.2. Testing session

After the network is trained the resulting weights are stored. For a test pattern a similarity measure, in this case the Euclidean distance, is calculated with each of the weight vectors. The test pattern is clustered on the unit having a weight vector that results in the smallest Euclidean distance.

Table 2. Sequence, NDB code definition of helix type with corresponding remarks and reference of the test sets

Sequence/code	Type/remarks	Reference
d(GCCGGC)/adf073†	A	Mooers *et al.*, 1995
d(AGGCATGCCT)udj032†	A/flipped-out bases	Nunn and Neidle, 1996
d(GCGTGG)bdf062	B/spermine	Tari and Secco, 1995
d(CGCAATTGCG)udj031†	B/helix–helix junction	Spink *et al.*, 1995
d(CGCAGAATTCGCG)udm010†	B/bulges	Joshua-Tor *et al.*, 1992
d(GGCCAATTGG)udj049†	B/overhanging bases	Vlieghe *et al.*, 1996
d(GCGCGC)bdfp24	B/modified backbone	Cruse *et al.*, 1986
d(GCATTGCT)udg028	B/loop	Leonard *et al.*, 1995

†Torsion angles of only one strand used.

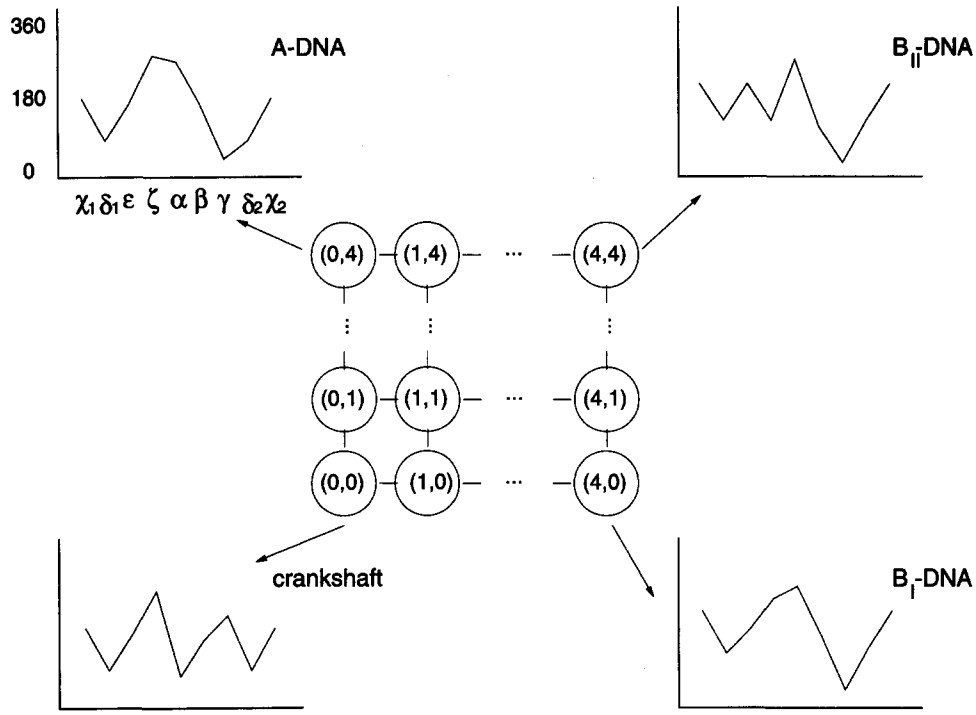

Fig. 3. Part of a 5 by 5 Kohonen map with the indices of the units. For some units a mean torsion angles "spectrum" (see text) of patterns typical for A-DNA, B_I-DNA, B_{II}-DNA and crankshaft dinucleotides are drawn.

3.1.3. Display

To analyse the results of training session and testing session, several display modes are used. In the first, for each map the number of patterns clustered on each unit is displayed. The indices of the patterns, and hence the torsion angles of these patterns, clustered on a unit, can be retrieved. In the second display mode, labels representing a DNA class and the mean torsion angles of patterns clustered on a unit are depicted in the same map. This leads to a kind of torsion angle "spectrum" typical for the patterns clustered on a unit. An alternative is to display the weight vectors of each unit. In this manner clustering of specific pattern characteristics is displayed, see Fig. 3. To study the distribution of torsion angles on the units, contour plots of the weights are displayed. The input patterns are not scaled. Therefore, the final weights values after a training session are on the same scale as the torsion angles. For each weight, corresponding to a certain torsion angle, the value on each of the units can be retrieved. These values can be displayed in contour plots. E.g., the third value on a weight vector corresponds to ϵ. After training the 10 by 10 map the values of the third value on the weight vectors of all units were stored. The minimum value was 170 and the maximum value 260. This range was subdivided into 10 parts and the weight values corresponding to ϵ were displayed as grey tones according to the range.

3.2. Hardware and Software

In this study we used SOM_PAK (Self-Organizing Map Program Package) Version 3.1 (SOM_PAK, 1995). The program was executed on a Sun Sparc™ machine. A training session of 100 000 cycles costs about 10 s real time. Additional software to analyse the outcome of SOM_PAK and to construct figures was programmed with Matlab for Unix Workstations, version 4.2c, by The MathWorks, Inc.

4. RESULTS AND DISCUSSION

4.1.1. Training session

Figure 4a depicts the number of input patterns clustered on each unit after training the 5 by 5 network. As can be seen in Fig. 4b, the B_{II}-DNAs are concentrated in the upper right corner of the map. There is some overlap with B_I-DNAs. On unit (3,3) one B_I-DNA is clustered with four B_{II}-DNAs and on unit (3,4) three B_I-DNAs occur with five B_{II}-DNAs. Besides the labels of DNA classes the mean torsion angles of the patterns clustered on each unit are depicted in Fig. 4b. These show a gradient of increasing ϵ values and decreasing ζ values in the direction (4,0) to (4,4). The same gradient, albeit less pronounced, as expected, can be observed in the direction (3,0) to (3,4). The A-DNAs are clustered entirely on the left side of the map. There is some overlap with B_I-DNAs. On unit (1,0) one B_I-DNA is clustered with two A-DNAs while unit (2,0) holds only one A-DNA with nine B_I-DNAs. The crankshaft entries are clustered with three A-DNAs in unit (0,0). Not surprisingly these three A-DNAs have a α–γ combination that is intermediate between real crankshaft dinucleotides and complete non-crankshaft dinucleotides. The gradient observed in going from (0,0) to (0,4) is an increasing α. The

column of units (2,0) to (2,4) clearly represents the B_I-DNA structure. The main difference is the column of units (1,0) to (1,4), which also holds some A-DNA nucleotides. The latter has higher mean ζ values and lower χ and δ values which is typical of A-DNA.

Although there is an acceptable class separation and much additional information from the mean torsion angle gradients in the 5 by 5 map, possibly more structure can be retrieved in a map with larger grid size. The orientation of the patterns that are clustered on a larger map may be different. This is caused by a renewed initialisation for larger maps. In the 7 by 7 map in Fig. 5, overlap in the B_{II}-DNA class is still observed. However, this overlap is small and the bulk of the B_{II}-DNAs is clustered on unit (6,0). There is only one unit with overlapping A-DNAs and B_I-DNA (unit (2,6) holding two A-DNAs and three B_{II}-DNAs) and now the crankshafts are completely separated from all other

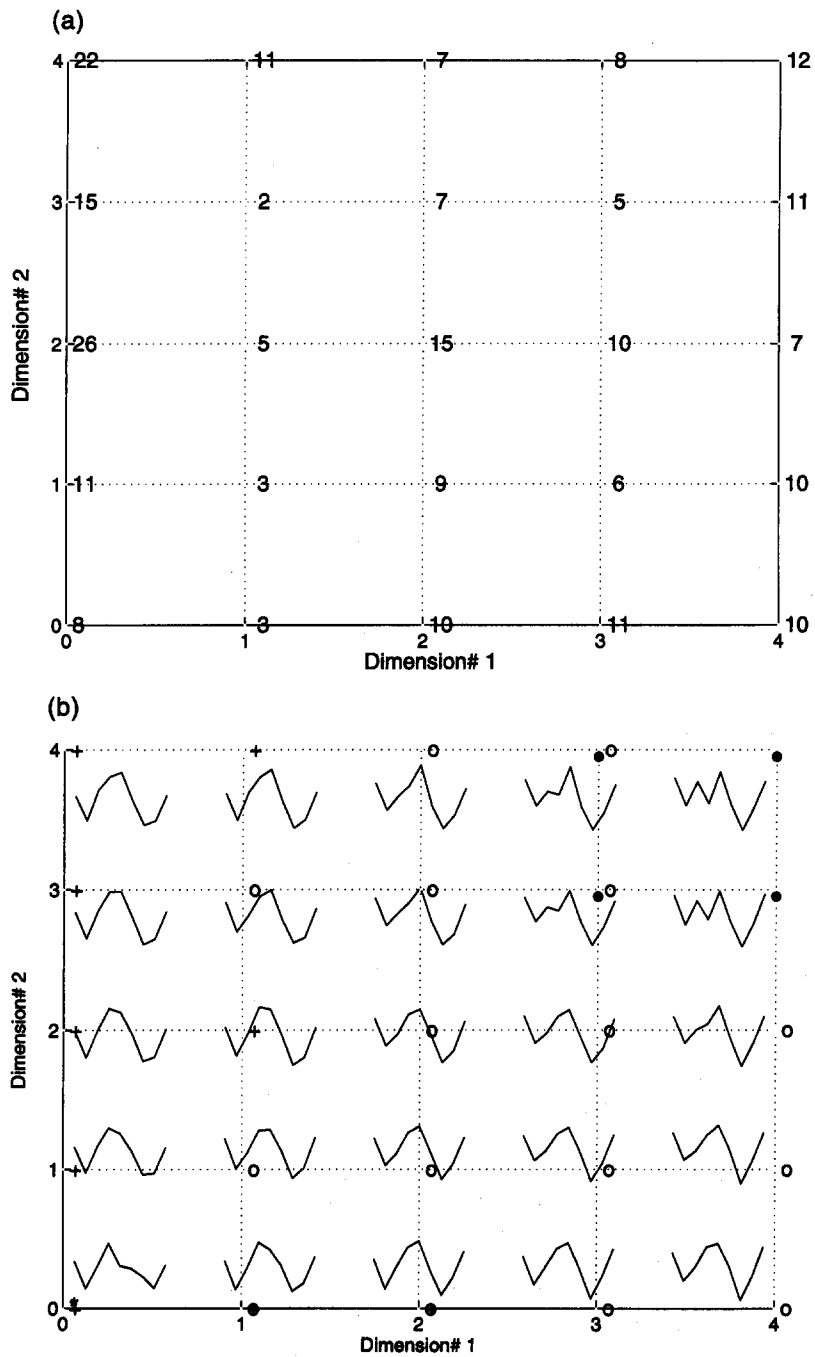

Fig. 4. 5 by 5 Kohonen map, (+) A-DNA, (○) B_I-DNA, (●) B_{II}-DNA and (*) crankshaft; (a) number of clustered patterns on each unit; (b) labels and mean torsion angles of clustered patterns on each unit.

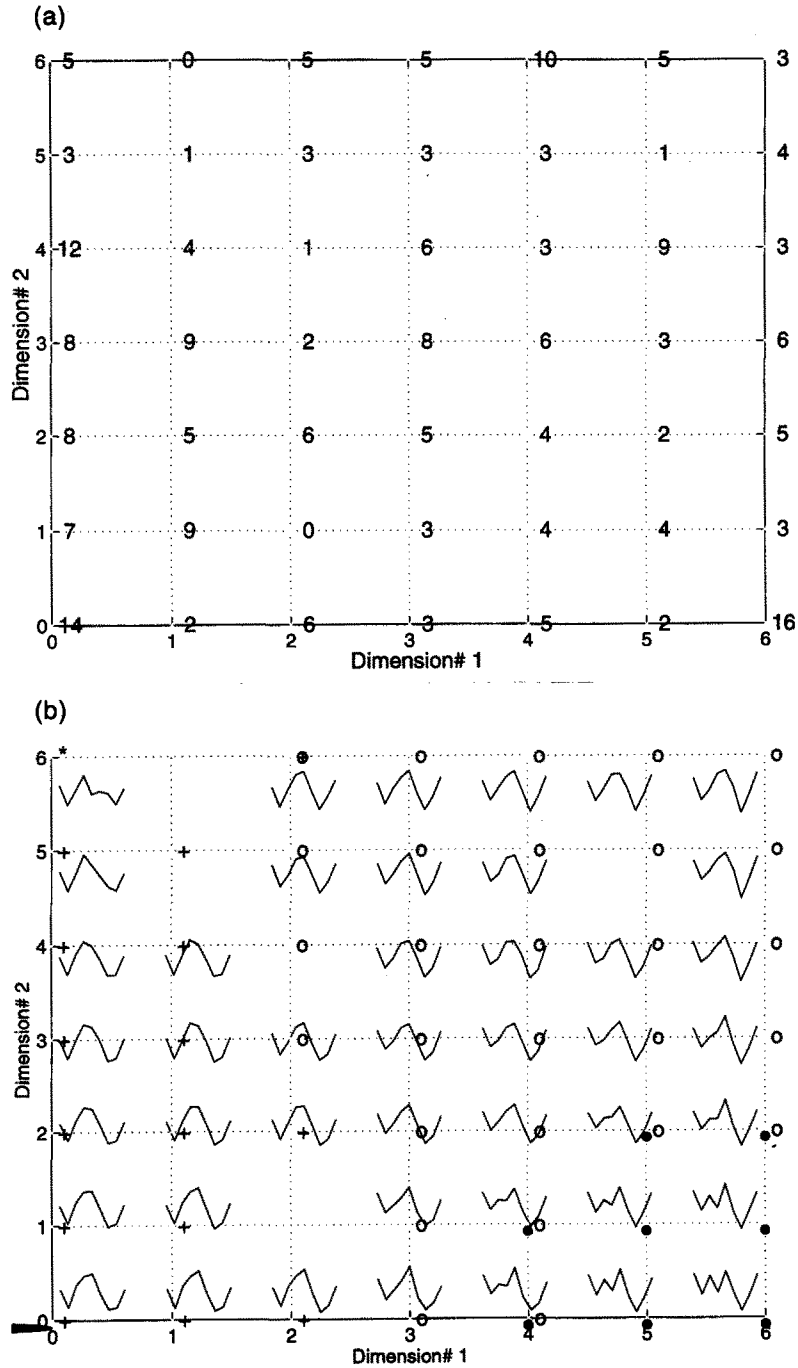

Fig. 5. 7 by 7 Kohonen map, (+) A-DNA, (○) B_I-DNA, (●) B_{II}-DNA and (*) crankshaft; (a) number of clustered patterns on each unit; (b) labels and mean torsion angles of clustered patterns on each unit.

dinucleotides. The three patterns in unit (0,5) are indeed the patterns that have a α–γ combination intermediate between real crankshaft dinucleotides and complete non-crankshaft dinucleotides. A more strict separation of A-DNA from B-DNA can be seen in this map. Of course, the gradients of mean torsion angles are still present here.

In the 10 by 10 map the class separation is very clear (Fig. 6). Not only is A-DNA separated from B-DNA but also B_I-DNA from B_{II}-DNA and, again, there is the crankshaft separation. Very small amounts of patterns are clustered on the three units that still have overlapping patterns. Hence, the 5 by 5 map does not yet give a proper separation of the crankshaft entries. For maps with more points (starting from 7 by 7) there is good crankshaft separation although there is some overlap of B_I-DNA with B_{II}-DNAs and of B-DNA with A-DNA. This latter overlap decreases when the number of points increases but as shown earlier (see Section 2) there is

an upper limit to this number of points (we took 10 by 10 as the upper limit).

Figure 7 depicts contour plots of the weights corresponding to ϵ, ζ, α and δ_1, respectively, after training the 10 by 10 map. Although no patterns are clustered on some units, the weights of these units were nevertheless adapted during the training session as a result of the neighbourhood function. After a training session is ended the weight vectors will show nice gradients. In this 10 by 10 map we see ϵ decrease and ζ increase in a vertical direction starting from the lower left corner. This is also the case when we go from the lower left corner to the right. In the same manner δ (and the positively correlated χ) decrease in going from left to right. Starting from the lower right corner α decreases (while the negatively correlated γ increases) when we move to the upper right corner. These are the main gradients. Of course, when having

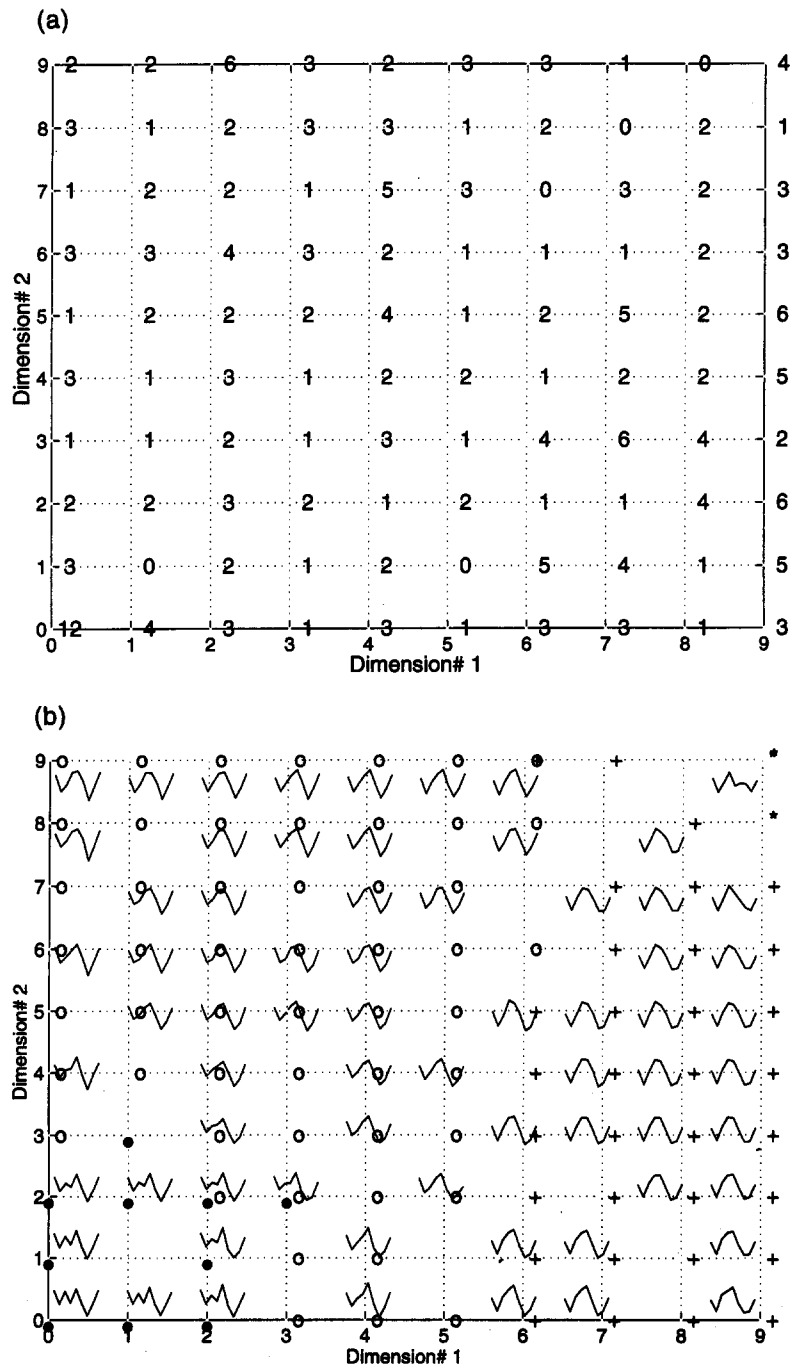

Fig. 6. 10 by 10 Kohonen map, (+) A-DNA, (○) B_I-DNA, (●) B_{II}-DNA and (*) crankshaft; (a) number of clustered patterns on each unit; (b) labels and mean torsion angles of clustered patterns on each unit.

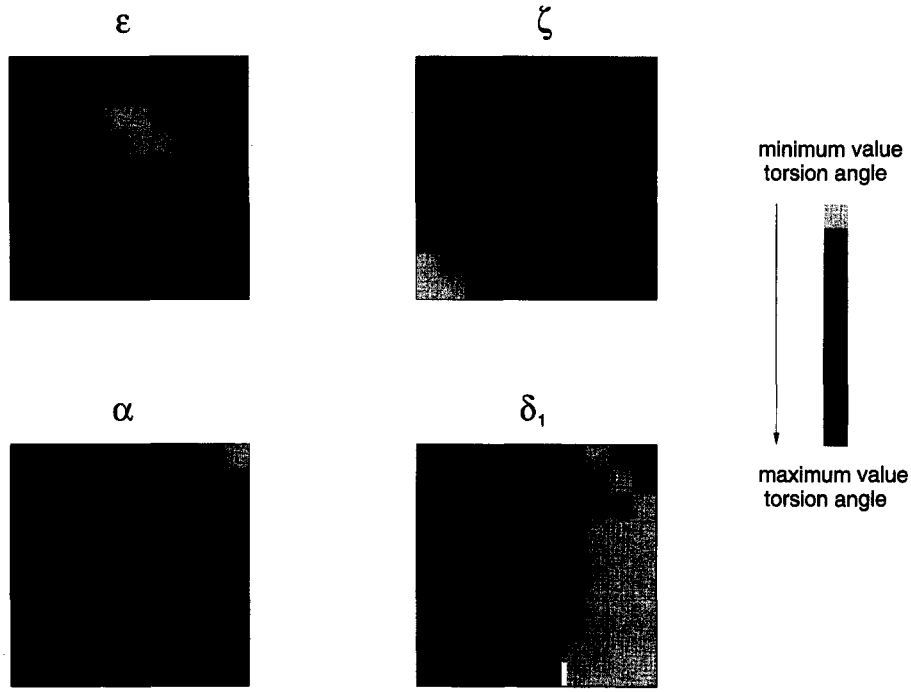

Fig. 7. Contour plots for weight values corresponding to ϵ, ζ, α and δ_1 after training the 10 by 10 Kohonen map. Light tones correspond to low values and dark tones correspond to high values (white corresponds to the minimum and black corresponds to the maximum of the weight value).

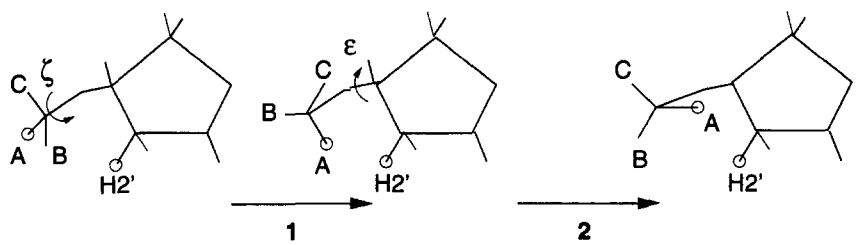

Fig. 8. When ζ moves from its normal *g-* value towards a *tr* value, atom A approaches atom H2′ which inevitably results in a van der Waals clash (step 1). However, the molecule avoids this by, at the same time, moving ϵ in the opposite direction from its normal *tr* value towards a *g-* value (step 2).

a more detailed look some other gradients can be detected.

On some of the units zero or only one pattern is classified. These units have no mean torsion angle spectrum and probably represent patterns that have torsion angle combinations which would produce severe sterical hindrance or other unfavourable structural aspects. Hence, these units may be considered forbidden areas for dinucleotides.

If a direct relation between conformations in the data matrix and their energies existed this would reflect the equivalent of a Ramachandran plot for nucleic acids. Because all the information from the original 9-dimensional torsion angle space is mapped onto a grid space a multidimensional Ramachandran plot would be the result. However, we can only speak in terms of accessible and forbidden areas in the map without making a comparison with the well-known Ramachandran plot*.

The data is not scaled. Hence, the weights are on the same scale as the torsion angles. Therefore, when a unit corresponds to torsion angle combinations that are considered forbidden, one looks at the weights and immediately has insight in the structural reason for this, e.g. it is known that ϵ and ζ are negatively correlated. Typical B_f-DNA ϵ and ζ values are 180° and 270°, respectively. It can be demonstrated that when ϵ starts to move towards a higher value, a corresponding decrease in ζ is necessary to avoid a van der Waals clash between two specific atoms in the backbone. This is depicted in Fig. 8. Maybe some of the units in the lower left corner of the 10 by 10 map on which none or only one pattern is classified represent ϵ–ζ combination in which one of the two angles has not changed enough to avoid the clash.

* A realistic plot might be constructed by producing all conformations for a dinucleotide with a certain resolution, calculating the energies, and subjecting the resulting data matrix to a self-organizing feature map analysis. However, depending on the resolution of the torsion angle variations, this leads to a huge data matrix.

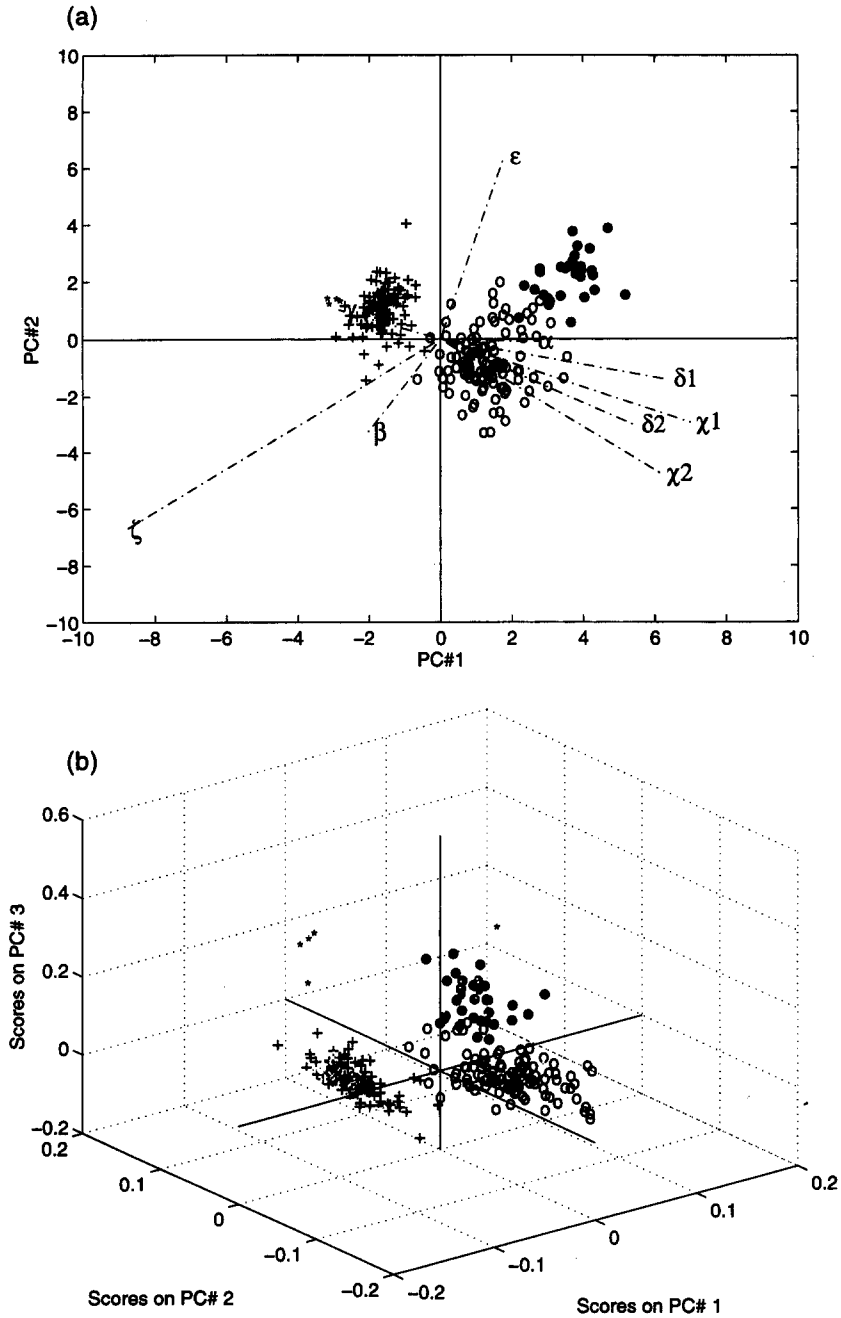

Fig. 9. (a) 2-Dimensional biplot; class separation of (+) A-DNA, (○) B_I-DNA, (●) B_{II}-DNA and (*) crankshaft dinucleotides as indicated by score points; (b) 3-dimensional score plot after SVD analysis.

However, we have to be careful with the interpretation of the maps in this manner. Training patterns that were used contained torsion angle combinations that are accessible. At least this is assumed because they are X-ray resolved structures. Although there are areas in the maps that seem to represent forbidden torsion angle combinations, because no patterns are clustered on them, the network was not explicitly trained for these kind of combinations by actually presenting corresponding patterns.

In an earlier study we used SVD to arrive at so-called biplots of the data matrix (Fig. 9). For a detailed discussion of this method see Beckers and Buydens (1998) In a biplot both scores and loadings are depicted in the same figure. Scores separate the data matrix into the known classes. Loading vectors correlate torsion angles. The projections of scores on loading vectors indicate which torsion angles play a dominant role in DNA class separation. The results of the biplots are supported by (simple) physical interpretations. The

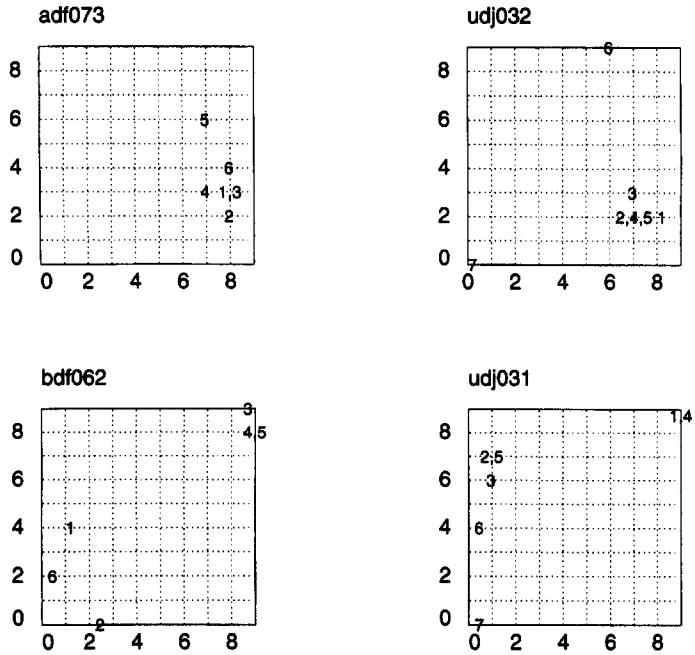

Fig. 10. Classification of unknown patterns from adf073, udJ032, bdf062 and udj031 on the 10 by 10 map.

advantage of the self-organizing feature map is that an easy to interpret class separation is obtained, i.e. the individual dinucleotides are assigned to a single unit. This may aid in a better interpretation of structural aspects of the clustering.

From a 3-dimensional score space the nine original torsion angles can be reconstructed using the SVD method. Hence, the interpretation of accessible and forbidden areas is again possible.

4.1.2. Testing session

The adf073^{+} strand is stored in the NDB as A-DNA. All the dinucleotides are correctly clustered in the A-DNA region of the Kohonen map (Fig. 10). Not much diversity is seen between the nucleotides which was expected by looking at the corresponding torsion angles (Table 3).

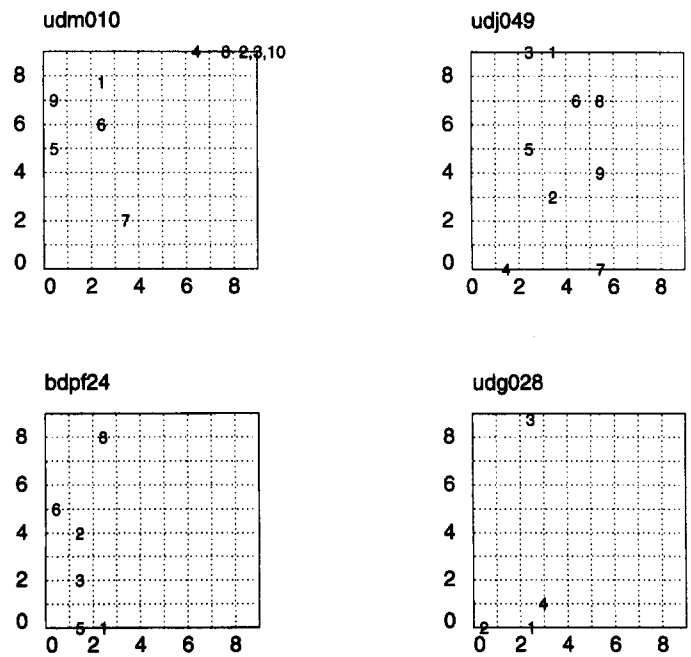

Fig. 11. Classification of unknown patterns from udm010, udj049, bdfp24 and udg028 on the 10 by 10 map.

Table 3. Torsion angles for test patterns

Pattern#	χ_1	δ_1	ϵ	ζ	α	β	γ	δ_2	χ_2
adf073									
1	204	84	206	293	292	164	59	80	202
2	202	80	216	285	283	183	53	89	201
3	201	89	202	288	277	178	64	84	194
4	212	82	211	298	277	181	59	79	211
5	211	79	189	291	275	186	55	76	197
6	197	76	198	283	284	191	54	87	203
udj032									
1	186	89	212	297	288	181	51	80	200
2	200	80	214	284	294	166	51	84	203
3	203	84	203	288	289	181	50	87	205
4	205	87	212	286	292	166	53	85	207
5	207	85	204	290	295	174	51	83	204
6	204	83	178	268	298	176	59	140	239
7	239	140	250	127	292	204	43	134	236
bdf062									
1	229	161	232	249	277	148	75	163	275
2	275	162	250	166	327	112	30	131	269
3	269	131	241	165	106	247	160	170	251
4	207	146	211	243	340	171	342	161	258
5	258	161	188	210	60	57	43	102	237
6	237	102	233	198	319	115	57	133	280
udj031									
1	250	129	191	240	34	182	302	152	270
2	270	152	159	265	298	186	64	156	268
3	268	156	175	247	305	195	36	130	259
4	259	130	122	300	94	164	249	156	246
5	246	156	180	279	285	187	54	145	263
6	263	145	221	231	308	163	18	161	292
7	292	161	285	150	254	159	53	140	260
udm010									
1	236	173	183	238	191	231	107	137	297
2	297	137	230	166	243	90	308	143	267
3	267	143	322	301	98	150	180	96	233
4	233	96	165	291	295	195	40	151	260
5	260	151	179	242	324	160	43	149	267
6	267	149	157	264	261	179	89	147	255
7	255	147	226	213	297	99	93	110	236
8	236	110	215	290	352	232	280	153	245
9	245	153	194	275	311	188	45	144	282
10	282	144	223	169	63	253	199	177	204
udj049									
1	245	112	207	252	317	158	40	137	286
2	286	137	258	287	308	145	181	95	223
3	223	95	213	286	296	171	51	148	274
4	274	148	244	162	314	149	32	146	270
5	270	146	164	263	283	182	70	116	235
6	235	116	176	271	309	168	62	114	233
7	233	114	189	254	318	160	48	113	242
8	245	112	177	274	283	180	59	121	244
9	244	121	188	261	301	169	51	90	232
bdfp24									
1	260	144	244	167	318	122	45	130	255
2	255	130	209	247	310	157	43	132	257
3	257	133	229	198	299	145	51	133	251
4	248	134	187	279	290	181	44	145	267
5	267	145	262	150	295	148	40	138	250
6	250	138	202	256	318	155	35	146	278
udg028									
1	255	117	238	153	312	143	43	138	220
2	220	138	217	70	79	194	67	146	231
3	231	146	279	283	302	192	60	147	256
4	256	147	213	229	323	139	39	113	250

Udj032$^+$ has a rather common A-DNA structure except that one of the bases is flipped out of the helix. Dinucleotide 6 of this strand has higher δ and χ values than previously examined dinucleotides. It is like B_I-DNA and is indeed clustered in the border region between A-DNA and B_I-DNA. Dinucleotide 7 has B-DNA δ and χ values and besides this, $\epsilon - \zeta > 0$. It is correctly identified as B_{II}-DNA.

In the bdf062 sequence there is base-pair opening and shearing of a T·A base pair. This non-Watson Crick hydrogen bonding scheme is induced by the addition of spermine. The δ and χ values of bdf062 indicate that all the dinucleotides belong to the B-DNA class. There are two distinct B_{II}-DNA dinucleotides. Dinucleotide 1 is an intermediate between B_I-DNA and B_{II}-DNA. The other patterns

are identified as crankshafts although in dinucleotide 4 and 5 the anti-correlated effect between α and γ is not observed. Instead it seems that only a low α value or high γ value, without any other extreme effects, leads to clustering in the crankshaft region.

Udj031$^+$ is part of a pseudo four-way helix–helix junction. All the dinucleotides in udj031$^+$ have high δ and χ values typical of B-DNA. Two of the dinucleotides show a clear crankshaft effect although they have mainly B-DNA characteristics. There are three B_I-DNA objects and one obvious B_{II}-DNA dinucleotide. Dinucleotide 6 is intermediate in that it has a high ϵ value but no corresponding low B_{II}-DNA ζ.

In udm010$^+$ one of the bases, which is in dinucleotide 2 and 3, is bulged or points towards the outside of the helix. Dinucleotides 2 and 3 have B-DNA-like δ and χ values but show a pronounced crankshaft effect. The other dinucleotides that are clustered in the crankshaft region have lower δ and χ values and hence show more A-DNA characteristics. Dinucleotide 7 obviously is intermediate between B_I-DNA and B_{II}-DNA. The other dinucleotides are in the B_I-DNA region.

Udj049$^+$ has overhanging bases which expresses itself in a diverse identification of patterns all of which have indeed more or less B-DNA characteristics although, in particular, dinucleotides 8 and 9 are in regions between A-DNA and B-DNA.

In bdpf24 the normal phosphoro-di-ester linkage in the backbone was replaced with a phosphorothioate linkage. Nevertheless, the B-DNA characteristics, with three dinucleotides in a B_I-DNA region and three dinucleotides in a B_{II}-DNA region, remained intact.

The udg028 sequence is a loop structure. This is accomplished by some specific torsion angle combinations. Dinucleotide 2 has a very low α but no corresponding high γ value. A high ϵ value is accompanied by a very low ζ which explains the identification as B_{II}. Also dinucleotide 3 has a high ϵ value but because there is also a high ζ it is clustered in the B_I region.

5. CONCLUSIONS AND OUTLOOK

We used a self-organizing feature map to cluster DNA dinucleotides. It is demonstrated that 244 patterns in a training session are distributed over the maps corresponding to their DNA-class characteristics. An advantage over traditional multivariate techniques, such as principal component analysis, is that no parametric model is assumed. Moreover, by studying the weights that result from a training session, one finds how the characteristics of specific patterns change while going from one unit to another. This results in a kind of DNA-class borders.

By displaying the numbers of patterns clustered on each unit, a density map is created. On some units zero or one pattern will be clustered and hence these units have a low density. These units possibly correspond to dinucleotides that have torsion angle combinations that may be forbidden on sterical grounds. Nevertheless, the largest map used in this study, the 10 by 10 map, has only a small number of units on which no patterns were clustered. On the one hand enlarging the grid size would probably result in a better DNA class separation and hence a better indication of forbidden and accessible areas. However, we have to keep in mind that "number of units < number of patterns". Therefore, enlarging the grid size beyond 10 by 10 leads to creating some kind of memory rather than a low-dimensional map for clustering purposes. On the other hand, the indication of accessible and forbidden areas in the present map is exclusively based on training patterns that are assumed to have acceptable torsion angle combinations.

It is demonstrated in this paper that unknown patterns are identified well using the weights that result from the training session. Some of the unknown patterns are clustered on units corresponding to only one training pattern but none of the unknown patterns are clustered on units holding no training patterns.

The applicability of the method is shown by means of the elegant manner it characterizes the variety in the DNA dinucleotide data set. This shows that the method has the potential to be used in more complex clustering or in characterization tasks concerning other types of nucleotides.

Acknowledgements—The authors would like to thank Dr. R. Wehrens for fruitful discussions.

REFERENCES

Beckers, M. L. M. and Buydens, L. M. C. (1988) Multivariate analysis of a data matrix containing A-DNA and B-DNA dinucleotides. Multidimensional Ramachandran plots for nucleic acids (accepted for publication in *Journal of Computational Chemistry*).

Berman, H. M., Olson, W. K., Beveridge, D. L., Westbrook, J., Gelbin, A., Demeny, T., Hsiek, S., Srinivasan, A. R. and Schneider, B. (1992) The nucleic acid database; a comprehensive relational database of three-dimensional structures of nucleic acids *Biophysical Journal* **63,** 751.

Cruse, W. B. T., Salisbury, S. A., Brown, T., Cosstick, R., Eckstein, F. and Kennard, O. (1986) Chiral phosphorothioate analogues of B-DNA: the crystal structure of Rp-d(Gp(S)CpGp(S)CpGp(S)C) *Journal of Molecular Biology* **192,** 891.

El Hassan, M. A. and Calladine, C. R. (1996) Propeller-twisting of base-pairs and the conformational mobility of dinucleotide steps in DNA *Journal of Molecular Biology* **259,** 95.

Joshua-Tor, L., Frolow, F., Apella, E., Hope, H., Rabinovich, D. and Sussman, J. L. (1992) The three-dimensional structures of bulge-containing DNA fragments *Journal of Molecular Biology* **225,** 397.

Kohonen, T. (1989) *Self-organization and associative memory*, 3rd edn. Springer–Verlag, Berlin–Heidelberg.

Kohonen, T. (1995) *Self-organizing maps*. Springer–Verlag, Heidelberg.

Leonard, G. A., Zhang, S., Peterson, M. R., Harrop, S. J., Helliwell, J. R., Cruse, W. B. T., Langlois D'Estaintot, B., Kennard, O., Brown, T. and Hunter, W. N. (1995) Self-association of a DNA loop creates a quadruplex: crystal structure of d(GCATGCT) at 1.8 Angstrom resolution *Structure* **3,** 335.

Melssen, W. J., Smits, J. R. M., Rolf, G. H. and Kateman, G. (1993) Two-dimensional mapping of IR spectra using a parallel implemented self-organizing feature map *Chemometrics and Intelligent Laboratory Systems* **18,** 195.

Mooers, B. H., Schroth, G. P., Baxter, W. W. and Ho, P. S. (1995) Alternating and non-alternating dG-dC hexanucleotides crystallize as canonical A-DNA *Journal of Molecular Biology* **249,** 772.

Mooren, M. M. W. (1993) On nucleic acid structure analysis by NMR. Thesis, University of Nijmegen, the Netherlands.

Nunn, C. M. and Neidle, S. (1996) The high resolution crystal structure of the DNA decamer d(AGGCATGCCT) *Journal Molecular Biology* **256,** 340.

Privé, G. G., Heinemann, U., Chandrasegaran, S., Kan, L. S., Kopja, M. L. and Dickerson, R. E. (1987) Helix geometry, hydratation, and G·A mismatch in a B-DNA decamer *Science* **238,** 498.

Ramachandran, G. N., Ramakrishan, C. and Sasisekharan, V. (1963) Stereochemistry of polypeptide chain configurations *Journal of Molecular Biology* **7,** 95.

Rubner, J. and Tavan, P. (1989) A self-organizing network for principal component analysis *Europhysics Letters* **10,** 693.

Smits, J. R. M., Schoenmakers, P., Stehmann, A., Sijstermans, F. and Kateman, G. (1993) Interpretation of infrared spectra with modular neural network systems *Chemometrics and Intelligent Laboratory Systems* **18,** 27.

SOM_PAK, the self-organizing map program package (1995) Version 3.1. Helsinki University of Technology, Finland.

Spink, N., Nunn, C. M., Vojtechovsky, I., Bermann, H. M. and Neidle, S. (1995) Crystal structure of a DNA decamer showing a novel pseudo four-way helix–helix junction *Proceedings of the National Academic of Science USA* **92,** 10767.

Tari, L. W. and Secco, A. S. (1995) Base-pair opening and spermine binding B-DNA features displayed in the crystal structure of a gal operon fragment: implications for protein-DNA-recognition *Nucleic Acids Research* **23,** 2065.

Vlieghe, D., van Meervelt, L., Dautant, A., Gallois, B., Precigoux, G. and Kennard, O. (1996) Parallel and anti-parallel $(G.GC)_2$ triple helix fragments in a crystal structure *Science* **273,** 1702.

Walczak, B. and Wegscheider, W. (1993) Non-linear modelling of chemical data by combination of linear and neural network methods *Analytica Chimica Acta* **283,** 508.