Classification of DNA structures by means of chemometrical methods

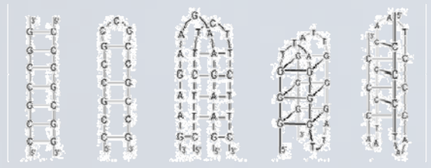
Miquel del Toro[1], Raimundo Gargallo[1], Ramon Eritja[2], Joaquim Jaumot[1]

1. Chemometrics Group, Analytical Chemistry Dept., Universitat de Barcelona.
2. Chemistry of Nucleic Acids Group, Structural Biology Dept., IBMB-CSIC.

INTRODUCTION

Nucleic acids could adopt several structures different from the famous Watson-Crick double helix[1] In order to study this variability of structures it is better to work with short strands (generally referred as oligonucleotides with a chain length between 10 and 25 bases) than with larger strands. Despite these short strands different structures can be built up.

On one hand, there are the unordered structures that are known as random coil strands. On the other hand there are the ordered structures. In this group we have several types of structures. An easy classification is obtained when considering the number of interacting bases.



So, duplex structures are obtained when only two different bases are bonded. Higher order structures such as triplex or quadruplex structures are obtained when, respectively, three or four bases are interacting simultaneously

An intesting fact of these ordered structures is the variety of topologies that can adopt in solution. This can be explained because these structures can be built up by the association of one, two or four different strands that allow to obtain the intramolecular (a single strand) or intermolecular (two, three or four strands) structures. In addition, depending on the length and the composition of the DNA sequence, it can be formed parallel or antiparallel structures[2].

Several spectroscopic techniques can be used to study the formation and stability of these structures such as UV molecular absorption, circular dichroism or NMR. However, circular dichroism in the UV region could be considered the most appropriate technique because it allows detecting differences in the structures adopted by the DNA sequences[3].

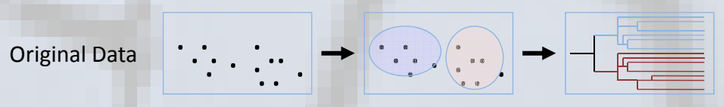
In this work it will be tested the ability of the combination of circular dichroism spectroscopy and chemometrics tools to classify a dataset made up by different DNA structures.

CHEMOMETRICAL METHODS

The goal data clustering is to group together samples that have similar spectral profiles. Different methods have been tested in this work:

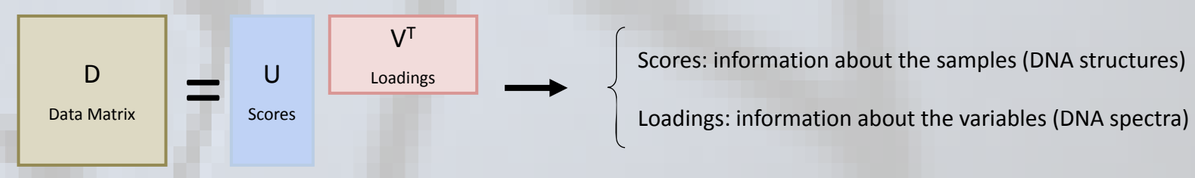
CLUSTERING

Hierarchical clustering algorithms typically build a tree (dendogram) that represents the similarity between the different samples evaluated using a "distance" measure [4]. Several linkage methods (centroids, median, Ward's method, nearest neighbour, …) with different forms of measure distances (Euclidean, SemiEuclidean, Chevychev, Pearson Correlation, Manhattan, …) have been tested.



PRINCIPAL COMPONENT ANALYSIS.

Principal Component Analysis (PCA) [5] is another standard method for clustering different samples. This method finds new axes explaining maximum variance. It allows the compression of the original information. Projection on the new PCA axes provides information about how samples are clustered.



EXPERIMENTAL METHODS

A database of CD spectra of DNA oligonucleotides has been obtained using a JASCO 810 spectropolarimeter.

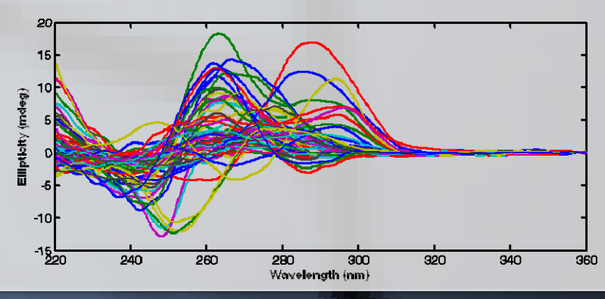
This data set is composed by 57 samples in which different DNA structures could be expected. Different experimental conditions were used depending on the DNA sequence and the expected structure:

- 2 different pH values: acetate buffer (pH=5) and phosphate buffer (pH=7)
- 2 different saline media: potassium or sodium
- 2 different temperatures: low (20ºC) and high (85ºC) temperatures

Using these experimental conditions , different DNA structures could be expected:

DNA structure	Expected Spectra
Single strand or High temperature structures	No significant features
Duplex	Positive maximum at approximately 260 nm.
Triplex	Strong variation depending on the conformation.
Quadruplex	Parallel structures: positive maximum at 260 nm Antiparallel structures: positive maximum at 290 nm

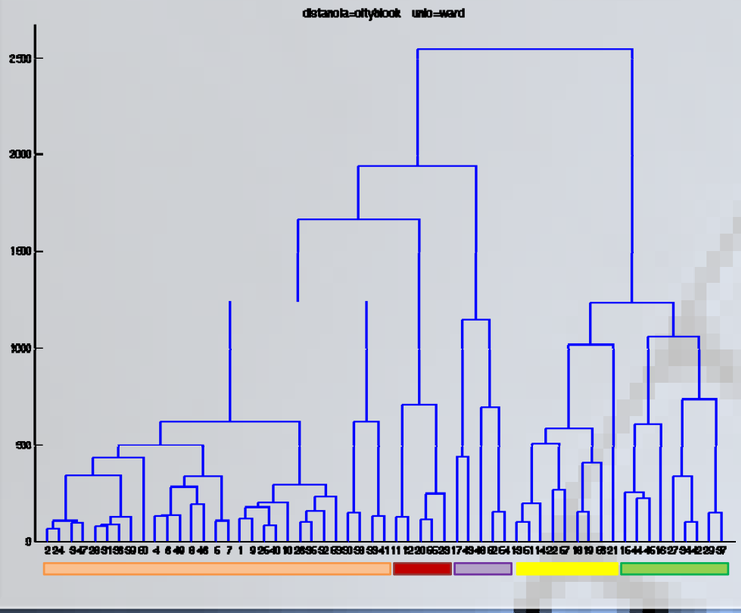
CD Spectra of all the DNA samples are:



CLUSTERING

The best results in order to explain the data variation were obtained by using the linkage Ward method and the Cityblock measurement of the distances.

Two main groups can be clearly observed allowing classification of the different structures.



LEFT BRANCH

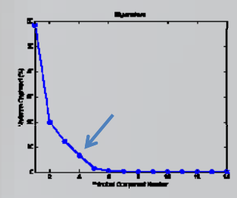
There is a main group corresponding to the samples measured at high temperature (1-10) and the samples where single strand structure can be expected. In all these cases only small signals are observed (less than 3 mdeg)

Two subgroups are also observed:
- Samples 11-23: Samples that show small positive bands between 260-300 and a strong negative band at approximately 260 nm.
- Samples 17-54: Samples that show strong positive bands between 280 and 295 nm.

RIGHT BRANCH

This cluster groups the samples that show a strong positive signal at approximately 260 nm. However, two subgroups can be observed:
- Samples 13-21 that show a positive signal at 260 nm.
- Samples 15-37 that show a positive signal between 260-270 nm and shoulders (or small contributions) between 275-290 nm.

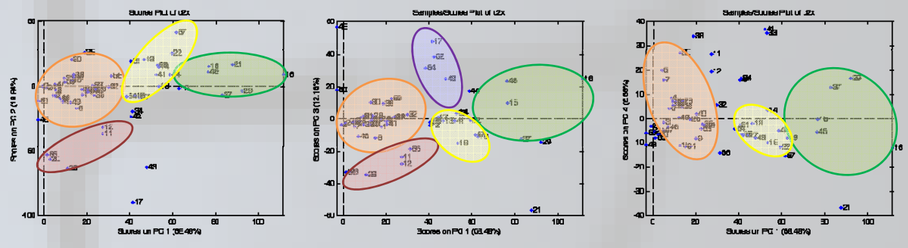
PRINCIPAL COMPONENT ANALYSIS



A small number of components is needed to explain most of the variability of the data. Four components explains approximately 97% of variance of the data.

ANALYSIS OF SCORES

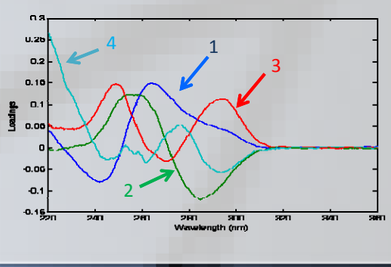
Similar groups to the ones observed using the hierarchical clustering methods can be identified.



ANALYSIS OF LOADINGS

Loadings profiles clearly identifies the most influential wavelengths in each component.

Most important wavelengths for the classification are 294, 285 and 264 nm.



Simultaneous analyisis of scores and loadings allows to identify the main trends of the different factors (i.e. the high temperature samples only show big contributions in the factors that explain less variance) .

CONCLUSIONS

Application of different chemometrical methods to a circular dichroism spectra dataset of several nucleic acids structures has allowed to obtain a classification of these samples.

Both chemometrical methods used have allowed to separate in different groups the ordered and unordered DNA structures. In order to obtain the classification of the samples both chemometrical methods (hierarchical clustering and Principal Component Analysis have detected key wavelengths.

These classification key wavelengths can be used to achieve a biological interpretation of the results. So, unordered structures do not show any significant signal upon the wavelength range studied. However, structured samples have been classified depending on the wavelength of the maxima (or minima) of the CD signals.

Further work should be addressed in order to improve the classification of the samples and to predict the structure of an unknown nucleic acid only from its circular dichroism spectra.

REFERENCES

[1] M. Blackburn; M.J. Gait: *Nucleic Acids in Chemistry and Biology*. Oxford University Press, 1990
[2] V.A. Bloomfield; D.M. Crothers: *Nucleics Acids. Structure, Properties and Functions*. University Science Books, 1999
[3] G.D. Fasman: *Circular Dichroism and the conformational analysis of biomolecules*. Plenum Press, 1996
[4] D.L. Massart, L. Kaufman. The interpretation of analytical chemical data by the use of cluster analysis. Wiley, 1983
[5] D.L. Massart; L.M.C. Buydens; B.G.M. Vandegiste: *Handbook of Chemometrics and Qualimetrics*. Elsevier, 1997

ACKNOWLEDGEMENTS

This research was supported by MEC (CTQ2006-15052-C02-01/BQU, CTQ2007-28940-E/BQU and BFU2007-63287/BMC).

Dr. Susana Navea is acknowledged for helpful discussion.