Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 132 (2014) 345–354

Contents lists available at ScienceDirect

Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy

journal homepage: www.elsevier.com/locate/saa

Application of supervised Kohonen map and counter propagation neural network for classification of nucleic acid structures based on their circular dichroism spectra

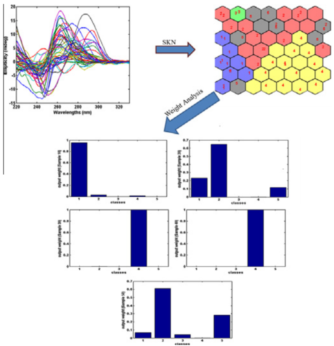
Mohadeseh Zarei Ghobadi [a], Mohsen Kompany-Zareh [a,b,*]

[a] Department of Chemistry, Institute for Advanced Studies in Basic Sciences (IASBS), Zanjan 45137-66731, Iran
[b] Department of Food Science, Faculty of Life Sciences, University of Copenhagen, Rolighedsvej 30, 1958 Frederiksberg C, Denmark

HIGHLIGHTS

- We applied the supervised Kohonen network for classification of nucleic acid structures.
- The classification was fulfilled based on CD spectra of different DNA structures with high similarity.
- It is shown that the spectral similarities between the weights of each group and their CD spectra lead to accurate prediction of unknown samples.
- The weight analysis was used for accurate prediction of unknown samples.
- Prediction of mixture samples was done well by applying the introduced weight analysis method.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 28 January 2014
Received in revised form 26 March 2014
Accepted 25 April 2014
Available online 9 May 2014

Keywords:
Classification
DNA
Chemometrics
SKN
CPNN

ABSTRACT

One of the most popular instrumental methods to detect the DNA structure is circular dichroism. Specific experimental conditions are required to form different structures of DNA. However, there is the possibility of different structures establishing in the similar circumstance. So, methods development to improve the classification and prediction of structures using their spectra information are needed. To this end, we applied unsupervised (PCA) and supervised (PLS-DA, SKN, and CPNN) approaches to classify CD spectra dataset of different DNA sequences (random coil (ss-DNA), duplex, hairpin, reversed and normal triplex, parallel and antiparallel G-quadruplex, and i-motif). The main part of this work concentrates on the application of artificial neural networks and weight analysis to obtain more classification and prediction accuracy. For this purpose, the trained network was run 10 times, and the average weights were taken. Also, weight analysis was done for the prediction of mixture samples include different structures. The results prove that new method of weights analysis based on SKN and CPNN is useful for classification of complicated data such as different types of DNA structures.

Introduction

Since the discovery of DNA in 1953 by Watson–Crick [1], various DNA structures have been known. The formation of these

[*] Corresponding author at: Department of Chemistry, Institute for Advanced Studies in Basic Sciences (IASBS), Zanjan 45137-66731, Iran. Tel.: +98 02414153123.
E-mail address: kompanym@iasbs.ac.ir (M. Kompany-Zareh).

http://dx.doi.org/10.1016/j.saa.2014.04.159
1386-1425/© 2014 Elsevier B.V. All rights reserved.

structures is depending on two parameters: DNA sequence and the experimental conditions. Disordered single stranded DNAs are the predominant form in high pH and temperature. In other hands, DNA secondary structures include duplex, G-quadruplex, i-motif, and triplex are stable in low temperature. The complementary base pairs (guanine/cytosine and adenine/thymine) allow the DNA helix to has a regular helical structure [1]. Besides, both G-rich and C-rich DNA strands can make unusual DNA structures. The G-rich strand constitutes a planar four-stranded G-quadruplex structure [2,3], whilst the C-rich strand can form the so-called i-motif with C/C^+ base pairs in acidic pH [4]. Triplex DNA is another high order structure due to association of a target duplex with a strand named triplex-forming oligonucleotide (TFO) [5]. The pyrimidine motif can bind parallel to the purine strand of the duplex by Hoogsteen hydrogen bonding (T.AT and C^+.GC) [6]. Also, the reverse Hoogsteen hydrogen bonding is the result of binding the purine motif to the purine strand of the duplex (T.AT, A.AT, and G.GC) [7].

One of the best spectroscopic techniques used for structural studies of DNA is circular dichroism (CD) [8,9] that has many advantages for conformational analysis such as: sensitive to low concentration of DNA, applicable for low soluble and film samples, suitable for both short and long oligonucleotides, and usable for detection of induced conformational changes affected by various factors [10].

Several supervised and unsupervised methods have been applied for classification and computational prediction of a biological dataset [11–14]. In supervised methods like partial least square-discriminant analysis (PLS-DA) [15], supervised Kohonen network (SKN) [16], and counter propagation network (CPNN) [17], the class label information is utilized to discrete between different groups. On the other hand, unsupervised methods as principal component analysis (PCA) [18] can differentiate between classes without need to groups information. Until now, most people have studied on the application of unsupervised self-organized

Table 1
Description of the CD spectra dataset.

Code	DNA sequence	Expected structure	Class code	Conditions
1	5′-CCGGCCGG-3′	Single strand	1	HT, K, pH 7
2	5′-TCTCCTCCTTC-3′	Single strand	1	HT, K, pH 7
3	5′-GAAGGAGGAGA-3′-(EG)6-3′-TCTCCTCCTTC-5′	Single strand	1	HT, K, pH 7
4	5′-GAAGGAGGAGA-T4-TGTGGTGGTTG-3′	Single strand	1	HT, K, pH 7
5	5′-Phos-AGGAGA-T4-AGAGGAGGAAG-T4-GAAGG	Single strand	1	HT, K, pH 7
6	Mixture of DNA sequence 2 and 4	Single strand	1	HT, K, pH 7
7	Mixture of DNA sequence 2 and 45	Single strand	1	HT, K, pH 7
8	5′-CGCGCGCG-3′	Single strand	1	HT, Na, pH 7
9	5′-CCGGCCGG-3′	Single strand	1	HT, Na, pH 7
10	5′-CCCCGGGG-3′	Single strand	1	HT, Na, pH 7
11	5′-TCTCCTCCTTC-3′	Single strand	1	LT, K, pH 7
12	5′-ACCCTAACCCTA-3′	Single strand	1	LT, K, pH 7
13	5′-CGCGCGCG-3′	Duplex inter-antiparallel	2	LT, K, pH 7
14	5′-CGCGCGCG-3′	Duplex inter-antiparallel	2	LT, Na, pH 7
15	5′-CCGGCCGG-3′	Duplex inter-antiparallel	2	LT, K, pH 7
16	5′-CCGGCCGG-3′	Duplex inter-antiparallel	2	LT, Na, pH 7
17	5′-CCCCGGGG-3′	Duplex inter-antiparallel	2	LT, K, pH 7
18	5′-CCCCGGGG-3′	Duplex inter-antiparallel	2	LT, Na, pH 7
19	5′-CGCGAATTCGCG-3′	Duplex inter-antiparallel	2	LT, K, pH 7
20	5′-GAAGGAGGAGA-3′-(EG)6-3′-TCTCCTCCTTC-5′	Duplex intra-parallel	2	LT, K, pH 7
21	3′-AGANGGANGGAAG-5′-5′-T4-CTTCCTCCTCT-3′	Duplex intra-parallel	2	LT, K, pH 7
22	3′-AGANGGANGGAAG-CTTTG-5′-5′-CTTCCTCCTCT-3′	Duplex intra-parallel	2	LT, K, pH 7
23	5′-GAAGGANGGANGA-T4-AGAGGAGGAAG-3′	Duplex intra-antiparallel	2	LT, K, pH 7
24	5′-GAAGGAGGAGA-T4-TGTGGTGGTTG-3′	Duplex intra-antiparallel	2	LT, K, pH 7
25	5′-GAAGGANGGANGA-T4-TGTGGTGGTTG-3′	Duplex intra-antiparallel	2	LT, K, pH 7
26	5′-Phos-AGGAGA-T4-TGTGGTGGTTG-T4-GAAGG-3′	Duplex intra-antiparallel	2	LT, K, pH 7
27	5′-Phos-AGGAGA-T4-AGAGGAGGAAG-T4-GAAGG-3′	Duplex intra-antiparallel	2	LT, K, pH 7
28	5′-T12-(EG)6-A12-3′-(EG)6-3′-T12-5′	Triplex reversed	3	LT, K, pH 7
29	5′-T12-(EG)6-A12-(EG)6-T12-3′	Triplex normal	3	LT, K, pH 7
30	5′-CGGGCACGGGAGGAAGGGGGCGGG-3′	G-quadruplex parallel	4	LT, K, pH 5
31	5′-CGGGCACGGGAGGAPAGGGGGCGGG-3′	G-quadruplex parallel	4	LT, K, pH 7
32	5′-GGCGCGGGAGGAATTGGGCGGG-3′	G-quadruplex parallel	4	LT, Na, pH 7
33	5’-GCGCGGGAGGAATTGGGCGGG-3′	G-quadruplex parallel	4	LT, K, pH 7
34	5′-TGGGGGT-3′	G-quadruplex parallel	4	LT, Na, pH 7
35	5′-TGGGGGT-3′	G-quadruplex parallel	4	LT, K, pH 7
36	5′-GGNGTTGGGTGTGGGTTGGG-3′	G-quadruplex antiparallel	4	LT, K, pH 7
37	5′-GGGNTTGGGTGTGGGTTGGG-3′	G-quadruplex antiparallel	4	LT, K, pH 7
38	5′-GGTTGGTGTGGTTGG-3′-biot	G-quadruplex antiparallel	4	LT, K, pH 7
39	5′-TAGGGTTAGGGT-3′	G-quadruplex antiparallel	4	LT, K, pH 7
40	5′-GGNGTTGGGTGTGGGTTGGG-3′	G-quadruplex antiparallel	4	LT, Na, pH 7
41	5′-GGGNTTGGGTGTGGGTTGGG-3′	G-quadruplex antiparallel	4	LT, Na, pH 7
42	5′-CCCGCCCAATTCCTCCCGCGCCCG-3′	i-Motif	4	LT, K, pH 5
43	5′-CCCGACCCCTTCAPTCCCGAGCCCG-3′	i-Motif	4	LT, K, pH 5
44	Mixture of DNA sequence 33 and 42	Duplex + quadruplex	5	LT, K, pH 7
45	Mixture of DNA sequence 33 and 42	Duplex + quadruplex	5	LT, K, pH 5
46	Mixture of DNA sequence 2 and 20	Duplex	5	LT, K, pH 7
47	Mixture of DNA sequence 2 and 25	Duplex	5	LT, K, pH 7
48	Mixture of DNA sequence 2 and 22	Duplex + triplex	5	LT, K, pH 7
49	Mixture of DNA sequence 2 and 26	Duplex + triplex	5	LT, K, pH 7
50	Mixture of DNA sequence 2 and 27	Duplex + triplex	5	LT, K, pH 7

(EG)6 denotes hexaethyleneglycol linker; biot denotes biotine tetraethylenglycol (biotine-TEG); phos denotes phosphate; AN denotes 8-aminoadenine; AP denotes 2-aminopurine and GN denotes 8-aminoguanine. In conditions column HT refers to high temperature (85 °C), LT to low temperature (20 °C), K to a 150 mM potassium medium, Na to a 150 mM sodium medium.