Physica A 409 (2014) 87–96

Contents lists available at ScienceDirect

Physica A

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

A novel representation of DNA sequence based on CMI coding

Wenbing Hou [a], Qiuhui Pan [a,b], Mingfeng He [a,*]

[a] School of Mathematical Sciences, Dalian University of Technology, Dalian, Liaoning, 116024, China
[b] School of Innovation Experiment, Dalian University of Technology, Dalian, Liaoning, 116024, China

HIGHLIGHTS

- We propose a novel representation based on CMI coding.
- We use a new method to extract the features of the DNA sequences.
- Different data sets are used to prove our model's effectiveness.

ARTICLE INFO

Article history:
Received 9 December 2013
Received in revised form 11 March 2014
Available online 6 May 2014

Keywords:
DNA sequence
Graphical representation
CMI coding
Similarity analysis
Phylogenetic tree

ABSTRACT

Graphical representation of DNA sequences provides a simple and intuitive way of analyzing and sorting various gene sequences. It is attractive to researchers to propose much more appropriate methods. In this study, a new graphical representation is presented. The method adopts the CMI coding to represent four nucleotides-A, G, C and T. Our approach considers not only the sequences' structure but also the chemical structure for DNA sequence. We take several sets of data to test our method. The results of our experiment demonstrate that our representation is effective.

1. Introduction

It is a challenge for biologists to understand the structure and function of DNA sequence. In recent years, numbers of approaches [1–12] have been suggested to obtain the characteristic information from the sequences by analyzing the DNA. The comparison between the DNA sequences of different species helps determine the phylogenetic relationship among species. As graphical representation is a simple way of viewing, sorting and comparing various gene sequences by providing intuitive pictures and patterns, it has become a research focus for bioinformatics. H-curve, the first graphical representation, proposed in 1983 by Hamori and Ruskin [13], demonstrates a visible 3D curve to tell the difference of species directly. Following the H-curve, various methods have been proposed [14–27]. A 2D representation of DNA sequence is proposed by Nandy [28]. However, as there are overlapping and intersections in the curve, some information is inevitably lost. Another graphical representation by Randic [29,30] is based on four horizontal lines. He constructed the L/L matrix, M/M matrix and other high order matrices. Assisted with this method, Randic first characterized the DNA sequence by extracting the leading eigenvalues of the matrices. Jafarzadeh et al. [31] proposed a curve by the name of C-curve. They characterize the sequences by the codons in the exons. According to the chemical structure, Liao [32] reduced a DNA sequence into six molecular topological indices of six graphical structures by a novel graphical coding of DNA sequence. Until now, some of

* Corresponding author.
E-mail address: mfhe@dlut.edu.cn (M. He).

http://dx.doi.org/10.1016/j.physa.2014.04.030
0378-4371/

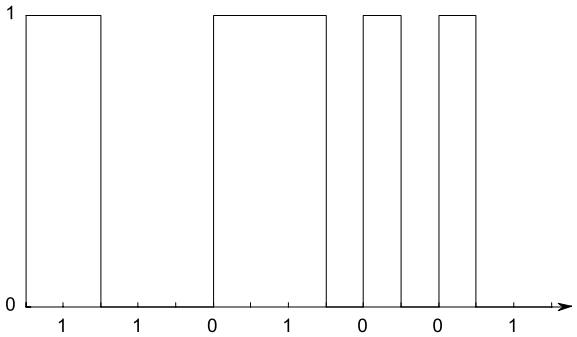

Fig. 1. The CMI coding for sequence "1101001".

the representations have been used in further research with great achievements [33,34]. For example, the Z-curve, presented by Zhang [35,36], is widely used in gene identification and prediction of different functional regions.

In this paper, we present a novel representation of DNA sequences. First, according to the different classifications, we translate the primary sequence into the "source sequence". Then the source sequence is translated into 0–1 coding sequence based on the CMI coding. With the application of cross correlation function, we calculate the similarities among thirteen species. Our result shows that the scheme is effective and it is in agreement with evolutionary relation satisfactorily.

2. CMI coding

In telecommunication, coded mark inversion (CMI) is a non-return-to-zero (NRZ) line code [37]. It encodes zero bits as a half bit time of zero followed by a half bit time of one, and while one bits are encoded as a full bit time of a constant level. The level used for one bits alternates each time one is coded. To clarify the definition, we take the sequence "1101001" as an example, and the coding result is shown in Fig. 1.

From the definition, we know that three digital pairs are used in the coding. The pair "1 0" is forbidden in this coding. It is an interesting feature and we will also make use of it in our work later.

3. The representation of DNA sequence

We construct a new mapping to translate primary DNA sequence into the source signal sequence. Considering there are two kinds of digits in the CMI coding, we need to build a mapping between the digits and the four kinds of nucleotides (A, G, T, C) by some classification. Based on the chemical characters, the nucleotides can be classed into groups: purine R = {A, G}/ pyrimidine Y = {C, T}; on the functional groups of the bases, we have amino M = {A, C}/keto K = {G, T}; on the strength of the hydrogen bonds between paired bases, the classification should be weak H-bond W = {A, T}/ strong H-bond S = {C, G} [31].

There is a DNA primary sequence P which will be translated into source signal sequence with the length of n:

$$P = p_1 p_2 p_3 \ldots p_n, \quad p_n \in \{\mathrm{A}, \mathrm{G}, \mathrm{T}, \mathrm{C}\}\,.$$

The source sequence we get from the primary sequence is S:

$$S = s_1 s_2 s_3 \ldots s_n, \quad s_n \in \{0, 1\}\,.$$

According to the classification we mentioned before, we have six kinds of mapping between the primary and source sequence:

$$s_i = \begin{cases} 0 & p_i \in W \\ 1 & p_i \in S \end{cases}; \qquad s_i = \begin{cases} 0 & p_i \in S \\ 1 & p_i \in W \end{cases};$$

$$s_i = \begin{cases} 0 & p_i \in M \\ 1 & p_i \in K \end{cases}; \qquad r_i = \begin{cases} 0 & p_i \in K \\ 1 & p_i \in M \end{cases}; \quad 1 \le i \le n.$$

$$s_i = \begin{cases} 0 & p_i \in R \\ 1 & p_i \in Y \end{cases}; \qquad s_i = \begin{cases} 0 & p_i \in Y \\ 1 & p_i \in R \end{cases};$$

As we introduced above, there are totally six kinds of mapping for one primary sequence. By the source sequences we get from the mapping, we can get the unique primary sequence. Here we make use of all these mappings to ensure the information could be kept completely. We translate the source sequence into the coding sequence under the rules of CMI coding. For example, we translate the first exons of β-globin gene from humans into a coding sequence. Some of the coding sequences are shown in Fig. 2.