Physica A 392 (2013) 4663–4667

Contents lists available at SciVerse ScienceDirect

Physica A

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

A new graphical coding of DNA sequence and its similarity calculation

Bo Liao *, Qilin Xiang, Lijun Cai *, Zhi Cao

College of Information Science and Engineering, Hunan University, Changsha, Hunan, 410082, China

HIGHLIGHTS

- We present a novel coding method that the DNA primary sequence.
- We use a novel method to display the features of those DNA sequences.
- We calculate the similarities/dissimilarities among the eleven species using the Euclidean distance.

ARTICLE INFO

Article history:
Received 28 January 2013
Received in revised form 9 April 2013
Available online 22 May 2013

Keywords:
DNA primary sequence
Structure graphic
Topological indices
Similarities/dissimilarities

ABSTRACT

The DNA primary sequence is translated into "structure graph" based on classification of its four bases. Then the invariants, such as topological indices, are extracted from those graphic representations of DNA primary sequences, and used to calculate the similarities among the eleven species. From these similarities/dissimilarities the homologies can be revealed that they are in agreement with evolutionary relation satisfactorily.

1. Introduction

The DNA sequence is a sequence of nucleotides over the four-letters alphabet {A, C, G, T}. The DNA sequence of different organisms has its own unique base sequence of genetic information, so the comparison of DNA sequences, such as comparing the order of 4 bases in the main chain, is very important. By comparing the DNA sequences of different species, we obtain their similarity, which can be inferred phylogenetic relationship between species distances.

At present, based on this, there have been some methods of comparing DNA sequences [1,2], however, those method are not perfect, such as different lengths of DNA sequence comparisons, which remains a failure to properly solve the problem.

With the development of graph theory, people image that using the graph represent the DNA sequence, and then predict the similarity between sequences by graphical comparison [3–29]. Based on this idea, Nandy [28] at first proposes a graphical representation of DNA sequences, so that DNA sequences comparison of different lengths is possible. Follow Nandy, some scientists according to certain rules use A, T, C and G to be different coordinates, and compare with the first exon of some species, which is a successful application [30–39]. For example, Randic [17,18] have considered kinds of condensed matrices. Bo Liao [6,14] have presented three characteristic representations corresponding to the three classifications of the bases of DNA and constructed several sets of matrices to represent DNA primary sequences. However, those methods only considered "graph", and completely ignored the "chemical".

* Corresponding authors. Tel.: +86 731 8821390; fax: +86 731 8821715.
E-mail addresses: in_ljcai@yeah.net (L.J. Cai), boliao@yeah.net, dragonbw@163.com (B. Liao).

0378-4371/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.physa.2013.05.015

In this paper, we present a novel coding method that the DNA primary sequence is translated into "structure graph". As an example of the first exon of β-globin gene of eleven different species, we use a novel method to display the features of those DNA sequences, and calculate the similarities/dissimilarities among the eleven species using the Euclidean distance. This method can also compare different lengths of DNA sequences, and improve the unique correspondence between graphs and sequences.

2. Graphical codes of DNA sequences

In chemistry, given a molecular formula, we can draw the corresponding molecular structure, which is consisted of a main chain and some side chains. Follow this idea; we present a group of graphical codes which translate a DNA sequence into "molecular structure".

As we all know, a DNA sequence is consisted of four bases A, C, G and T. Based of DNA sequences can be classed this four bases into groups: purine (A, G)/pyrimidine (C, T), amino(A, C)/keto(G, T) and week-*H* bond(A, T)/strong-*H* band(G, C) [25]. We can obtain only three representations corresponding to the three classifications.

For two groups of every classification, we set a group as the main chain, the other as the side chain. For example, in purine (A, G)/pyrimidine (C, T), A and G are used for main chain atoms, and C and T are used for side chain atoms. Using these codes, we can obtain an encoding way which is called the AG_CT structure. If C and T are used for main chain atoms, and A and G are used for side chain atoms, we can obtain another encoding structure. So every classification has two encoding structures. There are in total six different encoding structures, which are called AG_CT structure, CT_AG structure, AC_GT structure, GT_AC structure, AT_GC structure, and GC_AT structure, respectively. For example, the sequence GGTGCACCTGAC is translated into six molecular structures, which are listed in Fig. 1.

On the basis of the above six structure codes, we set 1 on the main chain atoms, and 0 on the side chain atoms. The distances between the marked 1 atoms and the marked 1 atoms are calculated, which is used for D_{1i1j}, as D_{1216} is the distance of the second main chain atom and the sixth main chain atom. Because there are 4 stripes between the second main chain atom and the sixth main chain atom, so $D_{1216} = 4$. The distances between the marked 1 atoms and the marked 0 atoms are calculated, which is used for D_{1i0j}, as D_{1206} is the distance of the second main chain atom and the sixth side chain atom. Because from the second main chain atom to the sixth side chain atom, there are 4 stripes in the main chain, and 1 stripe in the side chain, so $D_{1206} = 1+4 = 5$. So from every molecular structure a two distance matrix can be obtained. For example, we encode Fig. 1(a) into 0 and 1 and obtain D_{1i1j} and D_{1i0j}, which is listed in Fig. 2.

3. Graphical invariant extraction

There are a lot of parameters which are used to characterize the molecular structure [29]. The molecular topological index method [4], which is derived from the structure graph (such as distance matrix). The algorithm is relatively simple, while often getting better results in the correlation analysis. However, there are many parameters in the molecular topological index method. To solve this problem, we improve the molecular topological index method, and use only a total molecular topological index of distance matrix as the graphical invariant. This total molecular topological index of distance matrix is listed as follows:

$$S = \sum_{1i}\sum_{0j} D_{1i0j} \Big/ \frac{1}{2}\sum_{1i}\sum_{1j} D_{1i1j}.$$

In order to evaluate the new graphical codes and the improved molecular topological index, we list the first exon-1 of the *b*-globin gene for 11 different species which were reported by Randic et al. [11] in Table 1. In Table 2, we list the improved molecular topological index of six structures belonging to 11 species (see Table 1).

4. Similarity calculation

In order to facilitate the quantitative comparison of different species in terms of their collective parameters, we compute the similarity of different species. There are a lot of measure methods [29], such as Hamming distance, Euclidean distance, TaniMott factor, the correlation angle and so on. We use Euclidean distance to compute the similarity. Suppose that there are two species A and B, six parameters are $S_1^A, S_2^A, S_3^A, S_4^A, S_5^A, S_6^A, S_1^B, S_2^B, S_3^B, S_4^B, S_5^B, S_6^B$, respectively, where $S_1^A, S_2^A, S_3^A, S_4^A, S_5^A, S_6^A$ are the six improved molecular topological indices of six structures which correspond to the species A. The distance d_{ij} between two species is computed by the following formula:

$$d_{ij} = \left[\sum_{k=1}^{6}(S_k^i - S_k^j)^2\right]^{1/2}$$

where S_k^i is the mean of the improved molecular topological indices of the kth structures corresponding to the species i. d_{ij} is the measure of the similarity between the species i and species j. Obviously, the smaller d_{ij}, the more similar are the DNA sequences.