Journal of Molecular Structure: THEOCHEM 909 (2009) 33–42

Contents lists available at ScienceDirect

Journal of Molecular Structure: THEOCHEM

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

Applications of representation method for DNA sequences based on symbolic dynamics

Shiyuan Wang, Fengchun Tian *, Wenjiang Feng, Xiao Liu

College of Communication Engineering, Chongqing University, No 174 Shazheng Street, Chongqing 400044, China

ARTICLE INFO

Article history:
Received 10 April 2009
Received in revised form 16 May 2009
Accepted 16 May 2009
Available online 28 May 2009

Keywords:
DNA sequences
Symbolic dynamics
Biological information
Alignment
Similarity/dissimilarity

ABSTRACT

This paper addresses the applications of the improved representation method of DNA sequences based on symbolic dynamics. This method can visualize DNA sequences in three-dimensional coordinates with no loss of information in the transfer of data from a DNA sequence to its mathematical representation, and be applied to solve fundamental problems in bioinformatics fields. Graphical alignment of two shorter DNA sequences is exemplified for the application of the symbolic-dynamics-based representation method to sequence alignment. With the truncated length of DNA representation being chosen as the number of nucleotides in a codon, the application can be then extended to codon alignment. Based on the applications of alignment, the examination of similarities/dissimilarities among the coding sequences of the first exon of β-globin gene of different species illustrates the utility of the improved representation approach.

1. Introduction

The graphical representations of DNA sequences [1–4] were initiated to arrive at an efficient ways to intuitively inspect their data and structure, which can help in analysis of DNA sequence, e.g., sequence alignment and recognizing major differences among similar DNA sequences. However, in order to take full advantage of the graphical representations, an efficient representation method with numerical characterization is required, particularly when the analyzed DNA sequences have thousands of nucleic acid bases.

The graphical representations often start by assuming certain basic geometrical object or format, and then adopt a specific algorithm, used for describing individual nucleotides. Jeffrey [1] Zhang et al. [2], Hao [3], Randić [4–7] and others [8–11] developed the graphical representation with embodying different characteristics of DNA sequences. With the modification of the algorithm of the mathematical "chaos game", Jeffrey [1] designed chaos game representation (CGR) of DNA sequences for construction of various patterns with fractal characterizations, by assigning to the four bases, which are adenine (A), thymine (T), guanine (G) and cytosine (C), the four corners of a selected square. Then a two-dimensional compact graphical DNA representation in a limited space is obtained with no loss of information [1]. The *Z* curve proposed by Zhang et al. [2], is a three-dimensional curve which is a unique representation for a given DNA sequence in the sense that each can be uniquely reconstructed given the other [2]. Therefore, the *Z* curve contains all the information that the corresponding DNA sequence carries and also embodies the numerical characterizations of DNA sequences. Hao [3] proposed a global visualization method of DNA sequences based on counting and coarse-graining the frequency of appearance of nucleic acid bases of a given length. The Hao method can show distinctive patterns for different genomes, and also reveal fractal-like patterns in DNA sequences [3]. The four-line graphical representation, in which to the four nucleic acid bases are assigned four horizontal lines on which, at equal intervals are placed sequentially points corresponding to nucleotides as they appear in the DNA sequence, is proposed and expanded by Randić [4,5,7]. The advantage of the four-line method is to allow the construction of mathematical invariants characterizing a DNA sequence, resulting in a numerical characterization of a DNA sequence. The L/L matrix, M/M and their high order matrices are usually selected for DNA analysis with calculating the eigenvalues of these matrices [7].

As we have known that, the protein coding regions of DNA sequences tend to exhibit a period-3 pattern which has been regarded as a good indicator of gene position. Period-3 also implies chaos [12]. Hence, there probably exist chaotic dynamic characterizations within exons. The novel representation with embodying chaotic features of DNA sequences, which maps DNA sequences into chaotic sequences, has been proposed based on the principle of symbolic dynamics by Wang et al. [13].

In this paper, the symbolic-dynamics-based approach is expanded further by allowing visual inspection and numerical analysis of DNA sequences. In the improved graphical representation,

* Corresponding author.
E-mail addresses: wangsy@cqu.edu.cn (S. Wang), fengchuntian@cqu.edu.cn (F. Tian), fwj@cqu.edu.cn (W. Feng), liuxiao@cqu.edu.cn (X. Liu).

0166-1280/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.theochem.2009.05.025

there is no loss of information in the transfer of data from a DNA sequence to its mathematical representation. The features of improved representation method can be applied to facilitate solving complex problems of genome analysis. Therefore, the researches on applications of the improved representation are utilized to illustrate its efficiency and reliability in genome analysis.

2. Theoretical methods and materials studied

Genomic information being digital in a very real sense, is represented in the form of sequences of which each element can be one out of a finite number of entities. Therefore, such sequences, like DNA and proteins, have been mathematically denoted by character strings, in which each character is a letter of an alphabet. For instance, the alphabet for DNA representation is size 4 and consists of the letters A, T, C and G; the size of the corresponding alphabet for proteins representation is 20.

Consider a symbolic sequence space of DNA being defined as $\sum$, which is composed of the four letters. Define a shift operation σ, which is a map from the space $\sum$ to itself, shown as follows.

$$\sigma : \sum \to \sum. \tag{1}$$

Assume that a symbolic sequence starting at the position n, which belongs to the space $\sum$, is represented by $\boldsymbol{S}_n$. And this symbolic sequence can be governed by a left shift operation σ_l, where the leftmost symbol is discarded at each iteration [14], shown as follows.

$$\boldsymbol{S}_n = \sigma_l(s_{n-1}s_n s_{n+1}\ldots) = s_n s_{n+1} s_{n+2}\ldots \tag{2}$$

A one-dimensional discrete-time dynamic system is expressed as a difference equation, which is given by the following dynamical equation.

$$x_n = f(x_{n-1}), x_n \in I \tag{3}$$

where $f(\cdot)$ is a nonlinear function mapping points from a region I onto the same interval and possessing certain properties, such as sensitivity to initial values, which are the intrinsic feature of a chaotic sequence.

Hence, with the phase space of a dynamic system divided into a finite number of partitions, $I = \{I_1, I_2, \ldots, I_M\}$, and each partition labeled by a symbol, i.e., $s_n \in I_k$, a symbolic sequence can be generated by using a chaotic sequence, as shown by

$$s_n = \beta(x_n) = k-1, \quad x_n \in I_k, \quad (k = 1, 2, \ldots, M). \tag{4}$$

Accordingly, a symbolic sequence bearing a one-to-one relationship with the states of phase space for a dynamic system, can be represented by the dynamic system.

Conversely, rather than generating a chaotic sequence with (3) directly, an equivalent chaotic sequence can be obtained by exploiting symbolic dynamics, which provide a method for finite precision representation of chaotic system [14,15].

Define a truncated length of symbolic sequences being N, which is pertinent to the representation precision. As the truncated length increases, the representation is more accurate. Therefore, the mapping of symbol sequences $\boldsymbol{S}_n$ into chaotic sequences x_n can be expressed as

$$x_n = \beta^{-1}(\boldsymbol{S}_n) = \sum_{k=n}^{n+N-1} s_k M^{-(k-n+1)} \tag{5}$$

where M is the number of partitions in the space of a dynamical system. Actually, the symbolic sequence represents the M-ary expansion of a chaotic sequence.

According to symbolic dynamics, the relationship between symbolic sequences and dynamic systems is summarized as follows [13].

$$\begin{array}{ccc} \mathbf{S}_{n-1} & \xrightarrow{\sigma} & \mathbf{S}_n \\ \beta^{-1}\downarrow\uparrow\beta & & \beta^{-1}\downarrow\uparrow\beta. \\ x_{n-1} & \xrightarrow{f} & x_n \end{array} \tag{6}$$

Therefore, without consideration of biological information, DNA sequences can be directly mapped into one chaotic sequence with M in (5) being set to 4, that is

$$x_n = \sum_{k=n}^{n+N-1} s_k 4^{-(k-n+1)}. \tag{7}$$

Consider the DNA sequence F56F11.4a [13] in *Caenorhabditis elegans* chromosome III (base number 7949–14625, Accession No. AF099922) being denoted by (7). Fig. 1 shows the phase space diagram of the coordinate with the truncated length N being set to 9. It can be seen from this figure that the discrete time sequence generating from (7) has the same phase space diagram with the sawtooth map.

However, in the molecular biology, nucleic acids are linear macromolecules. Analysis and research of DNA sequences should consider not only the strings' structures but also their chemical structures. In DNA sequences, based on the biological information existing in the nucleotides, the four bases can be divided into two groups in three ways: purine (R=A, G) and pyrimidine (Y=C, T), amino (M=A, C) and keto (K=G, T), and weak hydrogen-bonds (W=A, T) and strong hydrogen-bonds (S=G, C). Assuming the DNA sequence being represented by $\{u_n\}$, the DNA sequence is first mapped into three symbolic sequences shown as follows.

$$\begin{aligned} s_n^1 &= \begin{cases} 1, & u_n = \mathrm{R} \\ 0, & u_n = \mathrm{Y} \end{cases} \\ s_n^2 &= \begin{cases} 1, & u_n = \mathrm{M} \\ 0, & u_n = \mathrm{K} \end{cases} \\ s_n^3 &= \begin{cases} 1, & u_n = \mathrm{W} \\ 0, & u_n = \mathrm{S} \end{cases} \end{aligned} \tag{8}$$

According to this classification, we map a DNA sequence into a three-dimensional representation by defining the three coordinates of the position n. Since each classification has two different groups, $M = 2$ in (5) is chosen herein. Therefore, a DNA sequence can be decomposed into three series of real-number, where the mapped point $X_n(x_n^1, x_n^2, x_n^3)$ with the three coordinates x_n^i being defined by

$$x_n^i = \sum_{k=n}^{n+N-1} s_k^i 2^{-(k-n+1)}, \quad i = 1, 2, 3. \tag{9}$$

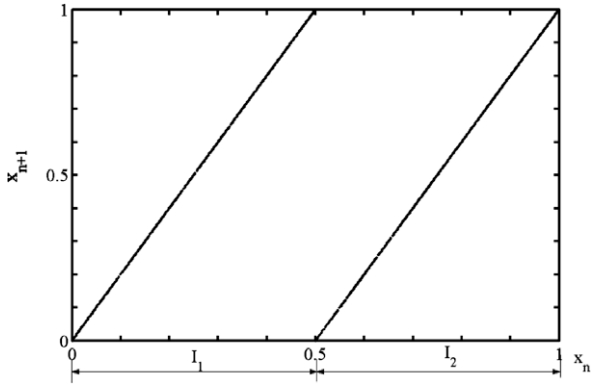

Fig. 1. The phase space diagram of the coordinate from (7), where the N being set to 9.