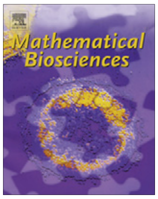
Contents lists available at ScienceDirect

Mathematical Biosciences

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



Vector representations and related matrices of DNA primary sequence based on *L*-tuple

Ying-zhao Liu [a,*], Tian-ming Wang [b]

[a] *Department of Mathematics, Luoyang Normal University, Luoyang 471022, PR China*
[b] *School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, PR China*

ARTICLE INFO

Article history:
Received 2 November 2007
Received in revised form 24 July 2010
Accepted 27 July 2010
Available online 3 August 2010

Keywords:
Vector representation
Related matrix
L-tuple
Numerical characterizations
H5N1 avian influenza virus

ABSTRACT

We consider to construct 4^L-components vectors for a DNA primary sequence based on the *L*-tuple. For two DNA sequences, using the corresponding vectors, we construct a set of $L \times L$ matrices called related matrix. The mathematical characterization from the constructed matrices have been selected to characterize the degree of similarity between the two DNA sequences. The search for similar sequences of a query sequence from a database of 39 library sequences and the construction of phylogenetic tree of H5N1 avian influenza virus illustrate the utility of the matrices for DNA sequences.

1. Introduction

The overwhelming burst in DNA primary sequences data has made the sequence analysis be a fundamental and challenging discipline that grew enormously in recent years. The “numerical characterizations approach” for the sequence analysis has recently been proposed to characterize the DNA sequences to help in comparison, identification, cataloguing and retrieval from the databases. From a mathematical point of view, these approaches can be divided into two groups – one is geometrical and the other is algebraic. The schemes in the first group is to convert a DNA primary sequences to a (or several) geometrical curve (is also called graphical representation of sequences in many references) and next according to the geometrical representation, to facilitate comparison of the corresponding DNA sequences and observing differences in their structure. In the second type of approach, with some algebraic methods, a DNA sequences is transformed into a (or several) numerical sequence, vectors or matrices, etc., and then some numerical characterizations are selected as invariants to analyze the DNA primary sequences.

Graphical representations of DNA sequences that have been developed within the past two decade years provide a way of gaining an understanding of the underlying genomic language. Following the pioneering work of Hamori and Ruskin [1], in recent years some novel graphical representations of DNA sequences based on 2D [2–12] and 3D [13–15] have been outlined. The advantage of such representations is that they allow visual inspection of data, helping in recognizing major differences among similar DNA sequences, but also to derive numerical characterization for DNA primary sequences. So, once the graphical representation of a DNA sequence is obtained, we can use it to extract useful knowledge from the sequence, such as long-range correlation information, sequence periodicities, local nucleotide composition and other sequence characteristics [5,10,15,17–20]; furthermore, according to the representation, some mathematical characterizations have been selected to quantitatively analyze the corresponding sequence, such as gene recognition and constructions of phylogenetic trees [21–25].

As for the algebraic methods, they, instead of the direct use of DNA sequences, consider a set of invariants of DNA sequence and uses these invariants for establishing the degree of similarity/dissimilarity among DNA sequences. After a DNA sequence is transformed into number sequences [16,26], vetors [28,29], or matrices [30–33], some mathematical theories, such as linear algebra, matrix analysis and random processes, are employed to extract some numerical characterizations from the “algebraic representations” as invariants for sequence analysis. In comparison with the sequence alignment methods, the time and space complexity of these approaches is relatively lower and they are more suitable for the whole genomes sequences comparisons, because the usage of the numerical characterizations condenses the information in the DNA primary sequences.

[*] Corresponding author.
E-mail addresses: liuyingzhaoyz@yahoo.com.cn (Y.-z. Liu), wangtm@dlut.edu.cn (T.-m. Wang).

0025-5564/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.mbs.2010.07.004

Similar to the numerical characterizations approaches, another alternative category is mainly based on the word (oligomer) frequency within a DNA sequence [34–40]. The basic idea of methods in this category is firstly to map the DNA sequence to vectors defined by the counts of each L-tuple, where L is the word length considered; secondly, linear algebra and statistical theory are further employed to define some distance functions; and finally based on the distance measure, the distance between the vectors quantify the similarities among the DNA sequences with the assumption that similar sequences will share word composition to some extent. Likewise, the computational complexity of the methods is relatively low and they have been applied to the rapid search for similar sequences in large databases, the detection of coding regions, the clustering of EST sequences, the evolutionary tree reconstruction and so forth (see Ref. [34] and references therein for a full description on these methods and their applications).

In Ref. [28], we provided an 8D representation of DNA primary sequences and a kind of related matrices for two different DNA sequences. In the method, the cumulative numbers of triplets of nucleic acid bases are used to extra biological information from the sequences and the normalized leading eigenvalues of the related matrices are selected as mathematical characterizations to perform the analysis of the sequences. It is obvious that the triplets of DNA sequences corresponds to the three-tuples and therefore in this study, we present a novel approach for sequence analysis with the intention to generalize the 8D representation of DNA sequences and combine the use of L-tuples and the numerical characterizations approach.

According to the cumulative numbers of different L-tuples, a 4^L-components vector representation of DNA primary sequences is proposed, and then using the non-overlapping sliding window method and the different start positions of the window, a DNA primary sequence is converted L different 4^L-components vectors. For two DNA sequences, using the different relationship among their vectors, we construct a set of $L \times L$ matrices called related matrix. Search for similar sequences of a query sequence from a database of 39 library sequences and construction of phylogenetic tree of H5N1 avian influenza virus showed in Table 2 illustrate the utility of the matrices for DNA sequences.

2. Vector representations and related matrices of DNA primary sequence

2.1. 4^L-dimensional representation of DNA primary sequences

A DNA sequences, of length n, can be viewed as a linear sequence of n symbols from a finite alphabet $\mathcal{N} = \{\mathcal{A}, \mathcal{C}, \mathcal{G}, \mathcal{T}\}$. A segment of L symbols, with $L \leqslant n$, is designated an L-tuple (in some references is also defined as L-word or L-plet). The set W_L consists of all possible L-tuples that can be extracted from the DNA sequence and has K elements,

$$W_L = \{w_1, w_2, \ldots, w_K\}, \qquad (1)$$
$$K = 4^L.$$

Let $S = s_1 s_2 \cdots s_n$ be an arbitrary DNA primary sequence, where n is the length of S. A window is a substring of the entire sequence and the step length is the number of base pairs that the window is moved each time. For a fixed L, we take the size of the window and the length of the moving step equally as L and it is obvious that the substring corresponding to the window at each moving is some L-tuple out of $\{w_k, k = 1,2,\ldots,K\}$. Because the window size and the step length are equal, the windows at each of moving steps are non-overlapping and hence we will obtain the different L-tuples for the different starting positions of the window. When the window is shifted along the DNA sequence S step by step with the start at position $u(u = 1,2,\ldots,L)$ in S, we denote the L-tuple occurring at the mth position in S as r_m^u, that is $r_m^u = s_m s_{m+1} \cdots s_{m+L-1}$. Based on all of the occurring L-tuples, we define the vector representing the S as $P_u(\alpha_1^u, \alpha_2^u, \ldots, \alpha_K^u)$, conveniently abbreviated to $P_u(\alpha_k^u)(u = 1,2,\ldots,L, k = 1,2,\ldots,K)$, where α_k^u is a number related to the cumulative numbers of the kth L-tuples in W_L. In detail, we define the number α_k^u by the following formulas:

$$\alpha_k^u = \frac{1}{n}\sum_{i=0}^{n'-1} \delta_{Li+u}^u, \qquad (2)$$
$$\delta_{Li+u}^u = \begin{cases} \sum_{j=0}^{i} \beta_j^u, & r_{Li+u}^u = w_k, \\ 0, & \text{else}, \end{cases}$$
$$\beta_j^u = \begin{cases} 1, & r_{Lj+u}^u = w_k, \\ 0, & \text{else}, \end{cases}$$

where n' is the quotient of $(n - u + 1)$ divided by L and the use of n in Eq. (2) is to normalize the vector P_u, reducing the effect of the length of S upon P_u. It is notable that if we were to consider a sliding window starting at position $L + 1$, the L-tuples obtained would be a subset of the ones for starting position 1, so this is actually the same sliding window staring at the first position. Consequently, when the sliding window starts at position $1,2,\ldots,L$, respectively, we can construct the corresponding L different 4^L-components vectors $P_1(\alpha_k^1), P_2(\alpha_k^2), \ldots, P_L(\alpha_k^L)$ to represent the S.

For example, consider the sequence $S = ATATACATAT$, where $n = 10$. We take $L = 3$, and it is evident that the set W_3 is comprised of all 64 triplets of nucleotide bases and the three different starting positions of the sliding window are just three reading frames, i.e., three possible ways of grouping bases to form codons in DNA or RNA sequences. Therefore the set W_3 and the three 64-components vectors $P_1(\alpha_i)$, $P_2(\alpha_i)$ and $P_3(\alpha_i)$ calculated by Eq. (2) would be

$$W_3 = \{ATA, TAT, TAC, CAT, ACA, AAA, \ldots\},$$
$$P_1 = (0.3, 0, 0.1, 0, 0, 0, \ldots),$$
$$P_2 = (0, 0.3, 0, 0, 0.1, 0, \ldots),$$
$$P_3 = (0.1, 0, 0, 0.1, 0, 0, \ldots).$$

Note that α_k^u depends on the number of appearances of w_k in the sequence S. Similar to the proof in Appendix of Ref. [28], we have that for two different DNA sequences S and E, when S and E are of unequal lengths, they must yield the different vector P; when S and E are of equal lengths, they will yield the different vector P except for some special circumstances (the detailed discussion not shown due to space limitations). In case of real DNA sequences, and specifically when comparing gene sequences with their wide variations in base distribution and compositions, the special circumstances could be expected to be very rare. So, once the DNA sequence is given, the $P_u(\alpha_k^u)$ is generally uniquely calculated by Eq. (2).

2.2. Related matrices

For any two DNA sequences S and E, by the vector definitions in sub Section 2.1, we can obtain the L vectors corresponding to each of the two sequences. Thus the similarity between the two DNA sequences can be obtained by the similarity among these vectors. Let $P_1(\alpha_k)$, $P_2(\alpha_k)$, …, $P_L(\alpha_k)$ and $Q_1(\alpha_k)$, $Q_2(\alpha_k)$, …, $Q_L(\alpha_k)$ be the vectors representing S and E, respectively. For obtaining the exact similarity between S and E and avoiding the loss of information in DNA sequence as possible, we will consider the similarity between any two vectors respectively from the vectors P and Q. We let d_{uv} be the value measuring the similarity between the vectors P_u and Q_v $(u,\ v = 1,2,\ldots,L)$.

Clearly, there are L^2 possible combinations between the vectors P_u and Q_v. Based on the similarity among these vectors, we