Journal of Theoretical Biology 281 (2011) 107–112

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

# Alignment-free comparison of genome sequences by a new numerical characterization

Guohua Huang [a,*], Houqing Zhou [a], Yongfan Li [b], Lixin Xu [a]

[a] Department of Mathematics, Shaoyang University, Shaoyang, Hunan 422000, China
[b] Hunan First Normal College, Changsha, Hunan 410002, China

ARTICLE INFO

*Article history:*
Received 4 December 2010
Received in revised form
1 April 2011
Accepted 2 April 2011
Available online 28 April 2011

*Keywords:*
Alignment-free comparison
Graphical representation
DNA sequence
Numerical characterization
Phylogenetic tree

ABSTRACT

In order to compare different genome sequences, an alignment-free method has proposed. First, we presented a new graphical representation of DNA sequences without degeneracy, which is conducive to intuitive comparison of sequences. Then, a new numerical characterization based on the representation was introduced to quantitatively depict the intrinsic nature of genome sequences, and considered as a 10-dimensional vector in the mathematical space. Alignment-free comparison of sequences was performed by computing the distances between vectors of the corresponding numerical characterizations, which define the evolutionary relationship. Two data sets of DNA sequences were constructed to assess the performance on sequence comparison. The results illustrate well validity of the method. The new numerical characterization provides a powerful tool for genome comparison.



## 1. Introduction

Nucleotide molecules are basic data material to explore the origin of life and metabolism of tissue, while comparative method is a more important strategy to investigate sequences. Conventional comparison of sequences is evaluated within the well-established framework of alignment (Kantorovitz et al., 2007), which would frequently be liable to lead to complicated computations, especially for multiple sequences. Hence, alignment-free approach is desirable. Recently, an alignment-free comparison based on numerical characterizations of sequences has been developed to compensate for ineffectiveness of traditional alignment of sequence. The numerical characterization quantifying the intrinsic nature of sequences determines the validity and quality of comparisons. During the past ten years, many numerical characterizations for DNA/protein sequences have been introduced, most of which are extracted from the string and graphical representations. The simpler and more important feature from string representations first used for comparison of genome sequence by Blaisdell (1986) and later for alignment-free comparison of regulatory sequences by Kantorovitz et al. (2007) is counts of $k$-letter words that is defined as the times of all possible $k$-letters words appearing in a sequence. Obviously, the characterization is a $4^k$-dimensional vector. Wu et al. (1997) further converted counts of $k$-letters words into a frequency vector as the numerical characterization of DNA sequences. Let $S=s_1s_2\ldots s_n$ be a DNA sequence, the $k$-letter words frequency distribution $F=(F_1,F_2,\ldots,F_{4^k})$are counted as follows: $F_i=C_i/\Sigma C_j$, where $C_i$ denotes the sum of the $i$th word with length $k$ in a genome sequence. Various frequency-based algorithms later have introduced for sequence comparisons as indicated in Wu et al. (1997, 2001, 2005), Korf and Rose (2009), Sims et al. (2009a, 2009b) and Jun et al. (2010). However, the parameter $k$ has a great influence on the results of sequence comparisons, so it is very critical how to pick a suitable length $k$. Some researchers have investigated selection of $k$. For example, Wu et al. (2005) proposed an optimal word size for dissimilarity measurement which is determined by length of sequence considered, i.e., increasing when the sequence length increases. They listed explicitly the optimal length in case the number of residues would be less than 5000 bp. For sequences with more than 5000 bp bases, the optimal $k$ has not been yet worked out. Sims et al. (2009a) reported another solution, thinking that the optimal length of word would appear in the range of the lower limit and upper limit. The lower limit is equal to $\log_4(n)$ approximately, while the upper one is determined by the criterion that polygenetic tree topology for length $k$ parallel that of $k+1$. Hence, there are a few numbers to select freely as the optimal sizes of word between the lower limit and upper one.

The features from graphical representations of DNA or protein sequences have been developed to capture the essence of the base composition and distribution of the sequences in a quantitative manner recently. There are mainly two methods of defining the

* Corresponding author.
*E-mail address:* guohuahhn@163.com (G. Huang).

0022-5193/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2011.04.003

numerical characterizations: geometry-based method and graph-theoretical one (Nandy et al., 2006). The geometry-based method originated by Raychaudhury and Nandy (1999) reckoned the central coordinate of the graphical representation as the characterizations of DNA sequences. The central coordinate picturing well positions of base in a geometrical curve already has attracted extensive attention. For example, Liao and Ding (2006), Wen and Zhang (2009) and Abo El Maaty et al. (2010) employed them for measuring similarity/dissimilarity of DNA/protein sequences. Charactering the main information of a DNA sequence, the central coordinate yet does not mirror all the nature hidden in a sequence. The graph-theoretical approach based on distance matrices, was first proposed by Randić et al. (2000), Randić and Vračko (2000) and further developed by Song and Tang (2005), Randić et al. (2003) and Liao and Wang (2004). These distance matrices are exactly or approximately equivalent descriptors for graphs, so some invariants such as leading eigenvalue, average row element, etc. are derived from them and viewed as the quantitative features for similarity comparison of sequences. When genome sequences become much longer, the corresponding distance matrices will occupy large memory in a computer and even seriously give rise to computation unavailable in common computers. Furthermore, the leading eigenvalues are usually involved in more complicated calculation. Yu et al. (2010) reported a numerical characterization which is different from the above method. The quantitative feature is defined by $M_j = \Sigma_{j=1}^{n}(x_i - y_i)^j$, $j=1, 2,\ldots, n$, where $n$ is the number of nucleotides in a DNA sequence, and $(x_i,y_j)$ is the coordinate of the $i$th vertex in a graphical curve. They proved that the numerical characterization is an equivalent form of a graph absolutely. The dimension of the characterization equals length of sequences, so the dimension of data of sequence does not reduce at all.

In the article, we presented a new numerical characterization of DNA sequences abstracted from the graphical representation which is a 10-dimensional vector, and then used them for alignment-free comparison of genome sequences.

## 2. Computational method

### 2.1. A 2D graphical representation of DNA sequences

As shown in Fig. 1, four vectors correspond to four groups of nucleotide A, T, G, and C, namely T($1, \sqrt[3]{2}$), C($1, \sqrt{3}$), G($1, -\sqrt{5}$) and A($1, -\sqrt[3]{3}$). On the basis of the above design, we can change a DNA sequence into a graphical curve in the first and fourth quadrants of Cartesian coordinate system using the same method as in Yu et al. (2010) and Huang et al. (2008, 2009). Here, we highlighted the difference of four vectors in $y$-values. By comparison with other designs for vectors, its advantage is that we can recover numbers of nucleotides A, T, G and C in the first $i$ residues in terms of the coordinate of the $i$th point in the graphical curve. Suppose that the number of nucleotides A, T, G and C is a variable labeled as $a$, $b$, $c$, $d$, respectively, and the coordinate is $(x, y)$. Obviously, $a+b+c+d=x$, and $a\sqrt[3]{2} + b\sqrt{3} - c\sqrt{5} - d\sqrt[3]{3} = y$. Because $y$ is a combination of these particular numerals, in theory, we can solve the above equations and then achieve the unique number of bases. We chose the particular vectors for the two purposes. For one thing, we assure the graphical representation produced by the vectors would not overlap. For another, the number of nucleotides A, T, G, and C can be computed out by the $i$th coordinate. Any array of vectors which satisfies the above two purposes can theoretically utilized to represent the nucleotides and differentiate different DNA sequences. In Fig. 2, we plotted the graphical graphs of the mitochondrial complete genomes from eight mammal species. The graph shows intuitively the similarity and dissimilarity of eight DNA sequences on the whole. In addition, the representation displays the two excellent qualities in the following: (i) there is no circuit in a curve that is to say, the graph does not overlap or intersect and (ii) correspondence between DNA sequences and curves proves one-to-one.

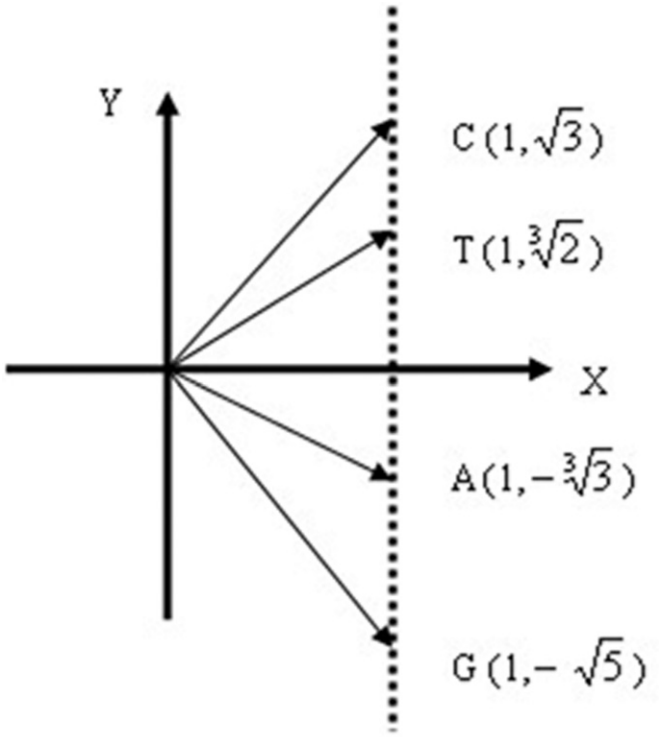


**Fig. 1.** Four vectors corresponding to nucleotides.

### 2.2. A new numerical characterization for genome sequences

Because a graphical curve uniquely denotes a DNA sequence, characterizations of representations can represent that of DNA sequences. The tallest peak and the lowest point are one of most important traits of a graphical graph. The central position of each kind of nucleotide in the representation also describes well geometrical feature of a curve, which is defined as follows:

$$\overline{M}_\alpha = \sum_{M_j = \alpha} P_j, \tag{1}$$

where $P_j$ is the coordinate of the $j$th residue $M_j$ in the curve, and $\alpha$ is one of four groups of nucleotide. Hence, we combined the tallest peak, the lowest point and the central points together as the numerical characterization for genome sequences which is a 10-dimensional vector or point. The characterization depicts the information about distribution and positions of bases in a sequence.

### 2.3. The quantitative measure of alignment-free comparison for sequences

After the numerical characterizations are obtained, the similarity among various sequences can be quantified by computing distance between either vectors or points. In the section, we reviewed the four important distance measurements that usually are utilized for alignment-free comparison of sequences. A more frequently used method is certainly Euclidean distance. Given two vectors or points $A=(a_1,a_2,\ldots,a_n)$ and $B=(b_1,b_2,\ldots,b_n)$ where $n$