Computational Biology and Chemistry 53 (2014) 331–338

Contents lists available at ScienceDirect

Computational Biology and Chemistry

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

A novel k-word relative measure for sequence comparison

Jie Tang, Keru Hua, Mengye Chen, Ruiming Zhang*, Xiaoli Xie*

College of Science Northwest A&F University, Yangling, Shaanxi 712100, PR China

ARTICLE INFO

Article history:
Received 26 April 2014
Received in revised form 10 August 2014
Accepted 25 October 2014
Available online 7 November 2014

Keywords:
DNA sequences
Discriminate analysis
Phylogenetic analysis
Phylogenetic trees

ABSTRACT

In order to extract phylogenetic information from DNA sequences, the new normalized k-word average relative distance is proposed in this paper. The proposed measure was tested by discriminate analysis and phylogenetic analysis. The phylogenetic trees based on the Manhattan distance measure are reconstructed with k ranging from 1 to 12. At the same time, a new method is suggested to reduce the matrix dimension, can greatly lessen the amount of calculation and operation time. The experimental assessment demonstrated that our measure was efficient. What's more, comparing with other methods' results shows that our method is feasible and powerful for phylogenetic analysis.

1. Introduction

The comparison of sequences is to find similarity, and to get biological features previously unknown. The pioneering approaches for sequence comparison were based on sequence alignment either global or local, pairwise or multiple sequence alignment. Waterman (1995) and Durbin et al. (1998) provided comprehensive reviews about this method. These approaches generally give excellent results when the sequences under study are closely related and can be reliably aligned, but when the sequences are divergent, a reliable alignment cannot be obtained. Hence, applications of sequence alignment are limited. Another limitation of these approaches is computational complexity and time-consuming, thus, they are limited when dealing with large-scale sequence data (Pham and Zuegg, 2004). To overcome these limitations, it is necessary to propose reliable and effective alignment-free sequence comparison methods.

Although many efficient alignment-free methods have been developed, they are still in early development compared with alignment-based measure. These alignment-free methods can be categorized into several classes: (i) Methods based on substrings employ the similarity and difference of substrings in a pair of sequences (Domazet-Louso and Haubold, 2011; He, 2006; Ukkonen, 1985; Ulitsky et al., 2006). (ii) Information Theory has provided successful methods for alignment-free sequence analysis and comparison. Existing applications of Information Theory include global and local characterization of DNA, RNA and proteins, estimating genome entropy to motif and region classification. It holds promise in gene mapping, next-generation sequencing analysis and metagenomics (Vinga, 2013). Methods based on Information Theory include: Base–base correlation (Cheng et al., 2013; Liu and Sun, 2008; Liu et al., 2008) and Lempel–Ziv compress (Li et al., 2001; Otu and Sayood, 2003). (iii) Graphical approaches can provide intuitive insights and the overall structure property, and therefore they are extremely useful in dealing with various biological problems, especially for very complicated biological systems, as indicated by various studies on a series of important biological topics, such as enzyme-catalyzed reactions (Andraos, 2008; Chou, 1980, 1989), protein folding kinetics and folding rates (Chou, 1990), drug metabolism systems (Chou, 2010), protein–protein interactions (Chou et al., 2011, 2011; Zhou, 2011), analysis of DNA sequence (Huang and Wang, 2012; Liao et al., 2006; Pandit and Sinha, 2010; Qi et al., 2004; Zhang et al., 2003), and graphic representation of protein sequence (Wu et al., 2010). (iv) The popular methods based on k-word frequencies include feature frequency profile (Sims et al., 2009; Sims and Kim, 2011), the D_2 score (Kantorovitz et al., 2007), frequency chaos game representation (Hatje and Kollmar, 2012), return time distribution (Kolekar et al., 2012) and composition vector (Gao and Qi, 2007; Hao et al., 2003; Qi et al., 2004; Lu et al., 2008; Wang et al., 2009; Wu et al., 2006). (v) Composition vectors based on k-word position is a new method. Many researchers have begun to extract the position information of a k-word (Afreixo et al., 2009; Ding

* Corresponding authors at: School of College of Science, Northwest A&F University, Yangling, Shaanxi 712100, PR China. Tel.: +86 13032906582.
E-mail addresses: tangjie6733033@163.com (J. Tang), huakeru@nwsuaf.edu.cn (K. Hua), chenmengye1988@126.com (M. Chen), ruimingzhang@yahoo.com (R. Zhang), xxlyuan@Hotmail.com (X. Xie).

http://dx.doi.org/10.1016/j.compbiolchem.2014.10.007
1476-9271/

et al., 2013; Gao and Luo, 2012; Huang and Wang, 2011; Yang and Wang, 2013). According to this classification, our method belongs to the composition vector based on k-word position. Based on Bonham-Carter et al. (2013), the basic steps of creating composition vectors based on k-word position are the following: (i) find the positions of the motifs in a sequence, (ii) create a vector by organizing the positions in some order, (iii) compute the distance between every two composition vectors to form a distance matrix, and optionally (iv) construct the phylogenic tree based on the differences.

In this paper, a new alignment-free method is presented based on the normalized k-word average relative distance to capture evolutionary information for sequence comparison. In our method, the effects of k-word counts, every k-word position distribution and the length of the sequence are combined together to capture more k-word distribution information. In Section 3, alignment-free method $E(w_1w_2\ldots w_k)$ (Ding et al., 2013) was compared with our method using discriminate analysis. The result demonstrated that each of our measure, with word length k range from 1 to 5, performs better than $E(w_1w_2\ldots w_k)$. Our method was further used to construct phylogenetic trees on two separate sets. Both the results were in good agreement with the authoritative phylogenies, which indicate that our measure is efficient for phylogenetic analysis. In sum, all these results demonstrated that our measure provides more information and greatly improves the efficiency of sequence comparison.

2. Materials and methods

2.1. New normalized k-word average relative distance

Supposing that $w_1w_2\ldots w_k$ is a k-word, where $w_i \in \{A, T, C, G\}$. Let $P_{w_1w_2\ldots w_k}$ represents the position vector of $w_1w_2\ldots w_k$ in a DNA sequence. If the $w_1w_2\ldots w_k$ occurs in a given DNA sequence, then $P_{w_1w_2\ldots w_k}$ is composed by positions of $w_1w_2\ldots w_k$ in the given sequence and $P_{w_1w_2\ldots w_k}(i)$ denotes its ith element. If $w_1w_2\ldots w_k$ does not exist in the given sequence, $P_{w_1w_2\ldots w_k} = 0$. For example, the 2-word position sequence for a short DNA sequence of length 20 *attcggcgtaatcgaccaaa*, so we can get $P_{aa} = (10, 18, 19)$, $P_{ct} = (0), \ldots$

These k-word position sequences can effectively capture distribution information of each k-word in the given sequence. For a fixed k, we can reverse this sequence by some of k-word position sequences. Furthermore, if a k-word exists in the given sequence, the counts of this k-word in the DNA sequence are equal to the length of its corresponding position sequence. We can use the following 2-word position sequences to reconstruct the DNA sequence used in the previous example:

$P_{aa} = (10, 18, 19)$, $P_{ac} = (15)$, $P_{ag} = (0)$, $P_{at} = (1, 11)$, $P_{ca} = (17)$, $P_{cc} = (16)$, $P_{cg} = (4, 7, 13)$, $P_{ct} = (0)$, $P_{ga} = (14)$, $P_{gc} = (6)$, $P_{gg} = (5)$, $P_{gt} = (8)$, $P_{ta} = (9)$, $P_{tc} = (3, 12)$, $P_{tg} = (0)$, $P_{tt} = (2)$.

Clearly, we do not find *ag*, *ct* and *tg* in the previous example. Now, we start to reverse the given DNA sequence as follows:

$P_{at} = (1, 11)$	at$*******$at$*******$
$P_{tc} = (3, 12)$	attc$******$atc$*******$
$P_{gg} = (5)$	attcgg$****$atc$*******$
$P_{cg} = (4, 7, 13)$	attcggcg$**$atcg$******$
$P_{ta} = (9)$	attcggcgtaatcg$******$
$P_{ac} = (15)$	attcggcgtaatcgac$****$
$P_{ca} = (17)$	attcggcgtaatcgacca$**$
$P_{aa} = (10, 18, 19)$	attcggcgtaatcgaccaaa

In order to calculate the similarity distances between different sequences, we should assign a signature to each k-word based on the k-word position sequence. In this paper, we try to extract more evolutionary information contained in position sequences. We use the following formula to extract evolutionary information from the DNA sequence. Suppose $P_{w_1w_2\ldots w_k} = (p_1, p_2, \ldots, p_n)$, where n is the counts of $w_1w_2\ldots w_k$ in the given DNA sequence. The new normalized k-word average relative distance of $w_1w_2\ldots w_k$ is denoted by $D(w_1w_2\ldots w_k)$ and it is defined as follows:

$$D(w_1w_2\ldots w_k) = \begin{cases} \dfrac{\sum_{i=1}^{n}(P_{w_1w_2\ldots w_k}(i) - P_{w_1w_2\ldots w_k}(1))}{n(l-k+1)}, & n \neq 0 \\ 0, & n = 0 \end{cases},$$

where l is the length of the given sequence.

Comparison of our method with another method $E(w_1w_2\ldots w_k)$ (Ding et al., 2013) shows that our method not only depends on the counts of $w_1w_2\ldots w_k$, but also the length of a DNA sequence and all the occurring positions of $w_1w_2\ldots w_k$. The method presented here combines the distribution information and the counts of the k-word together. It will capture more phylogenetic information from DNA sequences. For example, for two sequences $s = ggtcaacggg$, $t = gggtcaacgg$, the counts of *gg* in s and t are both 3. If we only consider the frequency of *gg*, $F_s(gg) = F_t(gg)$, or if we use the method $E(w_1w_2\ldots w_k)$, $E_s(gg) = E_t(gg)$, so phylogenetic information of *gg* captured by $F(gg)$ and $E(gg)$ is not enough. But when we use our method $D(w_1w_2\ldots w_k)$, $D_s(gg) = 0.5$, $D_t(gg) = 0.3$. Hence, more phylogenetic information of *gg* can be captured by $D(gg)$.

2.2. Distance calculations

For a fixed k, there are total 4^k distinct k-words to be considered. Putting these k-words in a fixed order, we can get a 4^k-dimension feature representation vector denoted by $(x_1, x_2, \ldots, x_{4^k})$. Then, according to the feature representation vector, we can get the vector $(D_1, D_2, \ldots, D_{4^k})$. For given L DNA sequences, we can get a $L \times 4^k$ matrix:

$$\begin{bmatrix} D_{11} & D_{12} & \ldots & D_{14^k} \\ D_{21} & D_{22} & \ldots & D_{24^k} \\ \vdots & \vdots & \ddots & \vdots \\ D_{L1} & D_{L2} & \ldots & D_{L4^k} \end{bmatrix}$$

With the increase of k, the dimensions of the vectors become very large, so as the matrix. We need much time to calculate their distance in the general computer. In fact, when k become larger than a certain number, a lot of k-words will not appear in the given sequences, and the final matrix is a sparse matrix.

For the sake of reducing the amount of calculation and operation time, we should reduce the dimension of the matrix, so a new method is proposed. On the basis of the feature representation vector, we can get L vectors of the k-words count $(c_1, c_2, \ldots, c_{4^k})$, where c_i denotes the count of its ith k-word element $w_1w_2\ldots w_k$. Then, we remove all the k-words that do not appear in each vector $(c = 0)$, which means that if a k-word is removed, the count of this k-word must be 0 in all vectors. So we can get L new vectors with the same number of components. The dimension of the new vectors N is less than or equal to 4^k. When k is small, N is equal to 4^k, when k becomes large, N is far less than 4^k. According to these new vectors, we can get L N-dimension vectors denoted by $(D_1, D_2, \ldots, D_N)$,