Physica A 398 (2014) 162–171

Contents lists available at ScienceDirect

Physica A

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

Phylogenetic analysis of DNA sequences based on *k*-word and rough set theory

Chun Li [a,b,*], Yan Yang [a], Meiduo Jia [a], Yingying Zhang [a], Xiaoqing Yu [c], Changzhong Wang [a]

[a] *Department of Mathematics, Bohai University, Jinzhou 121013, PR China*

[b] *Engineering and Technology Research Center of Food Preservation, Processing and Safety Control of Liaoning Province, Jinzhou 121013, PR China*

[c] *Department of Applied Mathematics, Shanghai Institute of Technology, Shanghai 201418, PR China*

HIGHLIGHTS

- We associate a DNA sequence with a 3×2^k dimensional complete word-based vector.
- It reflects information on both word frequencies and the order relation among them.
- We present a feature selection scheme on the basis of rough set theory.
- Based on the selected *k*-words, a much lower dimensional feature vector is obtained.

ARTICLE INFO

Article history:
Received 12 July 2013
Received in revised form 14 October 2013
Available online 22 December 2013

Keywords:
DNA
Feature selection
k-word
Phylogenetic analysis
Rough set theory

ABSTRACT

Among alignment-free methods for sequence comparison, the model of *k*-word frequencies is a well-developed one. However, most existing word-based methods neglect relationships among *k*-word frequencies, while a few others focus on the correlation of *k*-words but ignore the word frequency itself. In this paper, we propose a new *k*-word method which succeeds in conquering the two problems.

By means of characteristic sequences of a DNA sequence, we construct a 3×2^k dimensional complete word-based vector. Then we present a feature selection scheme based on rough set theory (RST) to extract the most informative *k*-words and use only these selected features to represent the DNA sequence. To evaluate the effectiveness of our method, we test it by phylogenetic analysis on three datasets. The first one is used as a training set, by which 869 top ranked *k*-words are selected. The other two are used as the testing set. The results demonstrate that the proposed method can capture more important information and is more efficient for molecular phylogenetic analysis.

1. Introduction

With the third-generation sequencing technology rapidly approaching, it becomes more feasible to obtain large sequence datasets of organisms at the whole genome scale. These enormous quantities of data issue the increasing challenge regarding the previous methods for molecular phylogenetic analysis. An important reason is that the most existing approaches for phylogenetic inference require a multiple alignment of sequences. As pointed out by Otu and Sayood [1], in addition to the bottleneck problems (the computational complexity and the inherent ambiguity of the alignment cost criteria), the multiple

[*] Corresponding author at: Department of Mathematics, Bohai University, Jinzhou 121013, PR China. Tel.: +86 416 3719118.
E-mail address: lchlmb@163.com (C. Li).

0378-4371/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.physa.2013.12.025

alignment strategy does not work for whole genome phylogeny due to gene rearrangements, inversion, transposition and translocation at the substring level, unequal length of sequences, etc. Therefore, it is necessary to develop the effective alignment-free method for sequence comparison.

During the past 30 years, there exist many attempts to characterize bio-sequences and compare them bypassing the multiple sequence alignment. All of these methods are intended to extract the hidden evolutionary information from different angles. For instance, graphical techniques of DNA sequences have emerged as a powerful tool for the visualization and analysis of long DNA sequences [2–21], which provide intuitive insights into local and global characteristics and the occurrences, variations and repetition of the nucleotides along a sequence. Another useful tool for characterization and comparison of bio-sequences is the matrix representation, such as the adjacency matrix, the condensed matrix, M/M matrix, L/L matrix and their "higher order" matrices [8,9,11,12,20–24]. The third class of methods, such as the Lempel–Ziv (LZ) complexity, and the Burrows–Wheeler (BW) transform [1,25–35], are based on compression algorithm, but do not actually apply the compression. There are also some other important methods which attempt to extend single nucleotide or single amino acid composition to study string composition. Within these methods each sequence is associated with a vector whose components are related to the k-word [36–46]. Some of them use strings of a fixed length whereas the others use strings with multiple lengths, in a range $[1, k]$ for some k [43,46]. However, most of these methods neglected the relationships among word counts in a sequence. Yang and Wang [45] considered rank orders of values of word counts and used 4^5 or 4^6 dimensional vectors to characterize a DNA sequence, but they ignored the word count itself.

Taking into account the two aspects above, we propose a novel k-word based measure to extract more information from biological sequences. Our method is based on three binary sequences of a DNA sequence, and corresponding to any given k, a 3×2^k dimensional complete vector is constructed. Since not every k-word contributes equally to the evolutionary distance calculation, we present a string scoring scheme to select the important k-words that heuristically identify the richest evolutionary information. Using these selected k-words, we obtain a much lower dimensional feature vector to characterize a DNA sequence. The proposed method is tested by phylogenetic analysis. In order to determine the value of k and the number of selected k-words (denoted by n_k), we regard a dataset composed of 19 Hantaviruses as the training set. Then, using the fixed k and n_k, we construct phylogenetic trees on two independent datasets commonly used in the literature [43,45,47–49]. The results demonstrate that our method provides more information and improves the efficiency of sequence comparison.

2. Methods

2.1. Complete word-based vector

For a (0, 1)-sequence S with length m, the count of a k-word w, denoted by $c(w)$, is the number of occurrences of w in the sequence S. Since there are $m - k + 1$ (overlapping) k-words in S in total, the frequency of appearance of word w in sequence S is $f(w) = c(w)/(m - k + 1)$. Once frequencies of all the n possible k-words are given, one often uses them to construct a frequency vector:

$$F_k = \left(f(w_{k,1}), f(w_{k,2}), \ldots, f(w_{k,n})\right).$$

Clearly, such a vector views word frequencies as discrete units separately, in spite of their correlations. To reflect the information on the order relation among k-word frequencies, we sort the k-word frequencies in F_k by ascending order, and denote the new vector by F_s:

$$F_s = \left(f(w_{k,i_1}), f(w_{k,i_2}), \ldots, f(w_{k,i_n})\right).$$

When two k-word frequencies equal, they are sorted by lexicographic order of the words. It implies that the following order relationships are satisfied in F_s.

$$f(w_{k,i_1}) \leq f(w_{k,i_2}) \leq \cdots \leq f(w_{k,i_n}).$$

Consequently, for each k-word frequency $f(w)$, there is a unique "position" in F_s, denoted by $g(w)$. Combining the position information with the frequency itself, we characterize a (0, 1)-sequence by an n-dimensional vector V_{FP}:

$$V_{FP} = \left(g(w_{k,1}) * e^{f(w_{k,1})}, g(w_{k,2}) * e^{f(w_{k,2})}, \ldots, g(w_{k,n}) * e^{f(w_{k,n})}\right).$$

Obviously, in the case of $f(w) = 0$, the component corresponds only to the position.

For instance, suppose

$S = 101101101110000001111110010010110110001011111110001000000110101010111110110001011011010100$ and $k = 3$.

Then there are 8 possible 3-words in total:

000, 001, 010, 011, 100, 101, 110, 111.

It is easy to find that

$$\begin{aligned} F_k &= (f(000), f(001), f(010), f(011), f(100), f(101), f(110), f(111)) \\ &= (0.1250, 0.0795, 0.1136, 0.1364, 0.0909, 0.1705, 0.1364, 0.1477), \end{aligned}$$