Journal of Theoretical Biology 395 (2016) 23–30

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

An estimator for local analysis of genome based on the minimal absent word

Lianping Yang [a], Xiangde Zhang [a,*], Haoyue Fu [a], Chenhui Yang [b]

[a] College of Sciences, Northeastern University, Wenhua Road, 110004 Shenyang, China
[b] School of Information Science and Engineering, Xiamen University, China

HIGHLIGHTS

- A comparison model is proposed based on the minimal absent word.
- Smooth-local-analysis-curve and similarity-distribution are constructed.
- A distance measure is deduced based on probability model.
- The method has potential advantages over the local alignment method.

ARTICLE INFO

Article history:
Received 12 October 2015
Received in revised form
17 January 2016
Accepted 19 January 2016
Available online 29 January 2016

Keywords:
Visual comparison
Alignment-free
Relative feature

ABSTRACT

This study presents an alternative alignment-free relative feature analysis method based on the minimal absent word, which has potential advantages over the local alignment method in local analysis. Smooth-local-analysis-curve and similarity-distribution are constructed for a fast, efficient, and visual comparison. Moreover, when the multi-sequence-comparison is needed, the local-analysis-curves can illustrate some interesting zones.

1. Introduction

With the avalanche of biological sequences generated in the post-genomic age, one of the most challenging problems in computational biology is how to formulate a biological sequence with a discrete model or vector, yet still keep considerable sequence order information. This is because almost all the existing machine-learning algorithms were developed to handle vector but not sequence samples (Chou, 2015). However, a vector defined in a discrete model may completely lose all the sequence-order information. To avoid completely losing the sequence-order information for proteins, the pseudo-amino acid composition or PseAAC (Chou, 2001, 2005) or Chou's PseAAC (Cao et al., 2013; Lin and Lapointe, 2013; Zhong and Zhou, 2014; Du et al., 2014) was proposed. Ever since the concept of PseAAC was proposed in 2001, it has been widely used in nearly all the areas of computational proteomics (see e.g. Cao et al., 2013) and a long list of references cited in a recent paper (Du et al., 2014). Encouraged by the successes of using PseAAC to deal with protein/peptide sequences, a natural question has occurred: how to use the similar approach to deal with DNA/RNA sequences? To address this problem, recently the pseudo-*k*-tuple nucleotide composition or PseKNC (Chen et al., 2014b) and PseKNC-General (Chen et al., 2014c) were developed.

We find there are two major sequence comparison frames among the alignment or alignment-free methods. One is characterization methods and the others are relative feature methods. Characterization methods attempt to introduce many technologies such as pattern recognition, artificial intelligence and so on into the sequence analysis field. The features of biological sequences can be extracted using different methods; hence, sequence comparison becomes feature vector comparison. In this case, we can formalize the sequence characterization process by $S \rightarrow \mathfrak{F}(S)$. After obtaining $\mathfrak{F}(S)$, various tools can be used to calculate the distance and identify similarities between sequences. These methods include Euclidean distance, angle measurement, correlation coefficient, relative entropy, etc. (Leimeister et al., 2014; Mantaci et al., 2008; Pham and Zuegg, 2004; Vinga and Almeida, 2003; Xia et al., 2013; Dai et al., 2013).

* Corresponding author.
E-mail addresses: yanglp@mail.neu.edu.cn (L. Yang), zhangxdneu@163.com (X. Zhang).

http://dx.doi.org/10.1016/j.jtbi.2016.01.023
0022-5193/

At the same time another kind of methods known as relative feature methods are used for sequence analysis. Relative feature means it depends on the compared object. That is to say the feature changes as the compared object does. In this situation, we put the pairwise sequences as a whole to obtain the similarity or dissimilarity information. Therefore, we can formalize this kind of methods by $(S,T)\rightarrow\mathfrak{F}(S,T)$ for the given biological sequences S and T. $\mathfrak{F}(S,T)$ is a similarity-object, from which we can deduce a similarity-score or a visualization. For example, the alignment method is a relative feature method because if the given (S,T) is aligned then the alignment is a similarity-object $\mathfrak{F}(S,T)$.

Characterization methods play an important role in the sequence analysis. A well-characterization method can reduce irrelevant factors, highlight important features and provide a better representation of the sequences. For example, graphical representation constructs a bijection between biological sequences and graphs in a plane or space (Liao et al., 2011; Randic et al., 2010; Yu and Huang, 2012; Zhang and Wang, 2000; Yao et al., 2010, 2014a,b). Therefore, $\mathfrak{F}(S)$ is a graph which gives us an illustration to detect the essence of the sequence. Some recent researches convert the DNA sequences into discrete signals and the Fourier analysis is followed (Yin and Yau, 2015; Hoang et al., 2015). A biological sequence can also be considered as a Markov chain in modeling (Wang et al., 2014). The composition vector model, which is based on Markov chain, has been applied in genome phylogenetic analysis (Qi et al., 2004). The pseudo-amino acid model has been successfully used in protein structure class prediction, subcellular localization, etc. (Chou, 2011, 2013; Chen et al., 2015b).

Since the concept of pseudo-amino acid composition or Chou's PseAAC (Du et al., 2012; Cao et al., 2013; Lin and Lapointe, 2013) was proposed, it has penetrated into many biomedicine and drug development areas (Zhong and Zhou, 2014) and nearly all the areas of computational proteomics (see e.g. Khan et al., 2015; Dehzangi et al., 2015; Kumar et al., 2015; Mondal and Pai, 2014; Wang et al., 2015) as well as a long list of references cited in Du et al. (2014). Because it has been widely and increasingly used, recently three powerful open access softwares, called "PseAAC-Builder" (Du et al., 2012), "propy" (Cao et al., 2013), and "PseAAC-General" (Du et al., 2014), were established: the former two are for generating various modes of Chou's special PseAAC; while the third one for those of Chou's general PseAAC (Chou, 2011), including not only all the special modes of feature vectors for proteins but also the higher level feature vectors such as "Functional Domain" mode (see Eqs. (9) and (10) of Chou, 2011), "GeneOntology" mode (see Eqs. (11) and (12) of Chou, 2011), and "Sequential Evolution" or "PSSM" mode (see Eqs. (13) and (14) of Chou, 2011). Encouraged by the successes of using PseAAC to deal with protein/peptide sequences, three web-servers (Chen et al., 2014b,c; Liu et al., 2015c) were developed for generating various feature vectors for DNA/RNA sequences. Particularly, recently a powerful web-server called Pse-in-One (Liu et al., 2015d) has been established that can be used to generate any desired feature vectors for protein/peptide and DNA/RNA sequences according to the need of users' studies.

In either vector or graph representation, the key idea is to convert S to another object $[\mathfrak{F}(S)]$ for comparison. However, the conversion process might lose some information that is seemingly irrelevant but potentially important. For example, if the widely used statistics are based on k-tuple as the $\mathfrak{F}(S)$ represents the biological sequence, then the order information of the original sequence will be considerably affected.

Relative feature analysis, which considers the original data (S,T) as a whole, can obtain sufficient information and effectively solve the sequence comparison problem. There is always a similarity-hypothesis when we convert (S,T) to $\mathfrak{F}(S,T)$. For example, the similarity-hypothesis of the alignment method indicates the similarity between the sequences is in the cost of insertion, deletion, and substitution. Lower cost indicates higher similarity. Another method for relative feature analysis is based on text compression. The corresponding similarity-hypothesis is the scale of one sequence that belongs to the other, and this sequence can be obtained by compressing the joint sequences. The joint sequences are significantly compressed, and the compression rate is much higher than that of the separate compression of two sequences when these sequences are close. Hence, many techniques in information compression can be used to analyze the sequences. These techniques include Kolmogorov complexity (Li et al., 2001), Lempel–Ziv complexity (Otu and Sayood, 2003), and Burrows–Wheeler transform (Yang et al., 2010).

Furthermore, we can discover the inherent weakness of the method by analyzing the similarity-hypothesis of one method. For example, the similarity-hypothesis of the alignment method focuses on the local mutations; thus, integral information such as segment rearrangement is neglected. When we use the compression rate to describe the similarity, the local information will lose much.

Many methods for relative feature analysis (Cohen and Chor, 2012; Comin et al., 2012; Haubold et al., 2009; Mantaci et al., 2005; Ulitsky et al., 2006; Yang et al., 2013a) depend on the similarity-hypothesis and the corresponding similarity-object. Yang et al. (2013b) employ the rearrangement of big k-words as the similarity-object. Their similarity-hypothesis is that the k-word-switchs waiting time is short if the two sequences are close to each other. Ulitsky et al. (2006) define the similarity by the longest common prefix. The corresponding similarity-hypothesis is that the bigger the total sum of the longest common prefix is, the more similar the pair-wise sequences is. Haubold et al. (2009) used the shortest unique substring (shustring) as the similarity-object $\mathfrak{F}(S,T)$. The similarity-hypothesis indicates that longer shustring implies lesser similarities.

The present study amplifies relative feature analysis on the basis of the minimal absent words. We propose a novel similarity-hypothesis and introduce a scoring system to describe how much one segment belongs to the other. We also design smooth-local-analysis-curve (SLAC) and similarity-distribution for sequence local or integral analysis and visualization. Tests on HIV-1 genome sequences show that our tools are powerful and efficient. We also compare our work to the local alignment method. Results show that our method has more potential advantages than local alignment on the local analysis problem.

2. Problem and method

2.1. Local similarity problem

The local similarity between two sequences has drawn our attention. The general method for local similarity involves local alignments that seek to optimize both locations and lengths of aligned substrings. Behnam et al. (2013) altered the local similarity problem using fixed-sized windows. Specifically, fixed-width windows are identified, with one in each of two sequences. Thus, the similarity between each pair of windows is maximal over all possible pairs of substrings. However, we find that there is an implicit assumption when we do the traditional local analysis. If we want to detect one segment in sequence S which is similar to some part of sequence T, the assumption is that there exists a continuous segment in T which is similar to one segment in S. The results obtained under this assumption are not always the best one because of the restriction of the continuity.

For example, sequence S includes a segment A which consists of two major parts α and β. At the same time there is a segment B