Journal of Theoretical Biology 335 (2013) 276–282

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Alignment free comparison: *k* word voting model and its applications

Lianping Yang, Xiangde Zhang*, Hegui Zhu

College of Sciences, Northeastern University, Shenyang 110004, China

AUTHOR-HIGHLIGHTS

- *k* word voting model is to compare the biological sequences without alignment.
- The model does not use the *k* word frequency or its statistics.
- The information entropy is employed to characterize the difference between the sequences.

ARTICLE INFO

Article history:
Received 22 December 2012
Received in revised form
25 April 2013
Accepted 26 June 2013
Available online 10 July 2013

Keywords:
Large scale comparison
Information entropy
Gamma distribution

ABSTRACT

Alignment free sequence comparison is widely used in sequence analysis, especially in computational biology for large scale similarity comparison. In this paper, we propose a word voting model to compare the biological sequences without alignment. Unlike many comparison methods based on the *k* word, this model does not use the *k* word frequency or statistics. Thus there is no limitation on the choice of *k*. Instead, we used information entropy of gamma distribution to characterize the differences among biological sequences in this model. Finally, we employed the model to do the similarity search and phylogenetic tree construction to further validate the model.

1. Introduction

One of the important problems in the field of biological sequence analysis is the appropriate definition and the accurate estimation of the similarity between the biological sequences. The similarity estimation could be applied to many potential fields of computational biology and bioinformatics. Many studies have indicated that sequence based prediction approaches (Chen et al., 2012; Chen and Li, 2013; Chen et al., in press; Du et al., 2012), such as prediction of protein domain (Chou et al., 1998; Feng et al., 2005; Li et al., 2012), prediction of protein secondary structure (Ding et al., 2009; Jones, 1999), protein subcellular location prediction (Chou et al., 2011; Du et al., 2009, 2011; Fan and Li, 2012; He et al., 2010; Liao et al., 2010; Xiao et al., 2011a), protein interaction based on sequences similarity (Chou and Cai, 2006; Huang et al., 2010; Wang et al., 2009; Zhao et al., 2012), protein quaternary attribute prediction (Sun et al., 2012; Xiao et al., 2009), protein 3D structure prediction (Chou, 2004), protein folding pattern prediction (Shen and Chou, 2009a; Shen et al., 2008), identification of membrane proteins and their types (Chou and Shen, 2007; Wang et al., 2008), identification of GPCR and their types (Xiao et al., 2011b), identification of proteases and their types (Chou and Shen, 2008, 2009b), protein cleavage site prediction (Shen and Chou, 2008), signal peptide prediction (Shen and Chou, 2007), predicting RNA editing sites (Du et al., 2007) can provide very timely and useful information and insights for both basic researches and drug design and hence are widely welcomed by the development of a novel method to study sequence similarity in the hope that it may become a useful tool in the relevant areas.

Although the alignment methods are the basically and popularly used methods during the process of the sequence comparison, many alignment free methods are investigated by many researchers to overcome some disadvantages of the alignments (Cheng et al., 2005; Dai et al., 2012; Edgar, 2004; Ferragina et al., 2007; Gao and Luo, 2012; Goeke et al., 2012; Mitrophanov and Borodovsky, 2006; Pham and Zuegg, 2004; Sims et al., 2009; Yang et al., 2012; Yang et al., 2013; Yang and Wang, 2013). In the society of the alignment free sequence analysis, the *k* words are frequently discussed (Dai et al., 2008, 2012; Ding et al., 2013; Jun et al., 2010; Liu and Wang, 2010; Wu and Ye, 2011). These methods transform a sequence into an object on which the tool commonly used in linear algebra and statistical theory can be applied (Mantaci et al., 2008). The basic hypothesis is that similar sequences share many common words. After mapping each sequence into a 4^k-

* Corresponding author. Tel.: +86 24 83671318.
E-mail addresses: yangmath@aliyun.com (L. Yang), zhangxdmath60@aliyun.com (X. Zhang).

0022-5193/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jtbi.2013.06.037

dimensional vector according to the k word frequency, the similar score is obtained by some distance measures belonging to Euclidean distance, Pearson correlation coefficient, Kullback–Leibler discrepancy and Cosine distance (Vinga and Almeida, 2003). For example, the D2 statistic is such a kind of method that compares the sequences using the k-word frequency. To improve the accuracy of comparison, Reinert et al. (2009) and Wan et al. (2010) suggested two variants of the D_2 word count statistic, which are called D_2^S and D_2^*, and they showed that the statistic is asymptotically normally distributed and not dominated by the noise in the individual sequences. Moreover, (Liu et al., 2011) developed an alignment free statistics based on D_2^S and D_2^* by comparing local sequence pairs and then summing over all the local sequence pairs of certain length.

But the use of the k words is under the restriction that k cannot be too large. For the DNA sequence, the 4^k words should be considered often, but 4^k is very large while k is not very big. As for the amino acids sequence, it is much worse. Apart from the big number, the vector consisting of the frequency of the 4^k words are sparse, which is an unstable factor during the comparison. In this paper, we propose a novel model called k word voting model. The idea is as follows: in a long queue election, the votes of two candidates are an upward spiral during a close contest. A landslide victory rarely happens in a close contest because if it happens, then it is overmatched. We let the k words be the voters. Then the evaluation of the similarity is the analysis on the contest. We make use of the information entropy of Gamma distribution to characterize the difference between the sequences. In the k word voting model, the choice of k is under a broad limitation and meanwhile some experiments show that it is an effective alignment free comparison model.

2. Method

2.1. k word voting model

Let $S=s_1s_2\cdots s_n$ be a sequence on an alphabet. If integer $i+k-1\leq n$, let $(S,i,k)=s_is_{i+1}\cdots s_{i+k-1}$ and $W_k(S)=\{(S,i,k)|1\leq i+k-1\leq n,i,k$ are postive integers$\}$. Note that $W_k(S)$ is the set of all the k words of sequence S. What we do next is the evaluation of the similarity between two sequences. The core assumption is that the votes of two candidates show an upward spiral during the close contest. The supporters of S could be regarded as the elements of the set $W_k(S)$ while the supporters of T are in $W_k(T)$. From that point of view, if all the voters of two sequences are lined by some kind of order, the evaluation of the similarity is the analysis on the contest. We arrange the set $W(S)\cup W(T)$ in the lexicographic order and call it "voting sequence". A pattern is called a switch if it forms as $(S,\cdot,\cdot)(T,\cdot,\cdot)$ or $(T,\cdot,\cdot)(S,\cdot,\cdot)$ in the voting sequence. We will utilize the switch's waiting time, that is, the number of the k words between the two adjacent switches to describe the relationship between S and T.

There are two basic hypotheses in our k word voting model:

Hypothesis 1. The switch waiting time is modeled with a gamma distribution due to the fact that the gamma distribution is frequently used to model waiting times;

Hypothesis 2. The entropy of the distribution of the waiting time is small if the two sequences considered are close to each other.

Hypothesis 3. In a fierce contest, the switch will appear more frequently in the voting sequence. As a consequence, this makes the uncertainty of the switch waiting time drop down and we have the hypothesis 2.

2.1.1. Gamma distribution parameter estimation

The gamma distribution is a continuous probability distribution with two parameters. One is shape parameter $\alpha>0$ and the other is scale parameter $\beta>0$. According to the hypothesis 1, $X\sim\Gamma(\alpha,\beta)$ if the switch waiting time is denoted by X. The equation defining the probability density function of X is

$$f(x;\alpha,\beta)=\begin{cases} x^{\alpha-1}\exp(-x/\beta)/(\beta^\alpha\Gamma(\alpha)) & x\geq 0 \\ 0 & otherwise \end{cases} \tag{1}$$

where $\Gamma(\alpha)=\int_0^{+\infty}x^{\alpha-1}\exp(-x)\,dx$.

Given two sequences, we can obtain the samples (the switch waiting times) which could be used to estimate the parameters of the gamma distribution. Let $x_1,x_2,\ldots,x_m$ be those samples. According to the maximum likelihood estimation method, the likelihood function is

$$\begin{aligned} L(x_1,\ldots,x_m;\alpha,\beta) &= \prod_{i=1}^{m}(x_i^{\alpha-1}\exp(-x_i/\beta))/\beta^\alpha\Gamma(\alpha) \\ &= (\Gamma(\alpha))^{-m}\beta^{-\alpha m}\left(\prod_{i=1}^{m}x_i\right)^{\alpha-1}\exp\left(-\sum_{i=1}^{m}x_i/\beta\right) \end{aligned} \tag{2}$$

The log likelihood function is

$$\begin{aligned} \ln(L(x_1,\ldots,x_m;\alpha,\beta)) &= -m\ln\Gamma(\alpha)-m\alpha\ln\beta \\ &\quad +(\alpha-1)\sum_{i=1}^{m}\ln x_i-\sum_{i=1}^{m}x_i/\beta \end{aligned} \tag{3}$$

Further, we obtain the likelihood equations as follows:

$$\begin{cases} f_1(\alpha,\beta)=\dfrac{\partial\ln L(x_1,\ldots,x_m;\alpha,\beta)}{\partial\alpha}=-m\psi(\alpha)-m\ln\beta+\sum_{i=1}^{m}\ln x_i=0 \\ f_2(\alpha,\beta)=\dfrac{\partial\ln L(x_1,\ldots,x_m;\alpha,\beta)}{\partial\beta}=-m\alpha/\beta+1/\beta^2\sum_{i=1}^{m}x_i=0 \end{cases} \tag{4}$$

where $\psi(x)=d(\ln\Gamma(x))/dx$ is the digamma function.

There is no closed-form solution for the parameters. But the numerical solutions could be obtained by, for example, Newton's method.

Following the estimation of the parameters, Kolmogorov Smirnov test is employed to compare the switch waiting times with the gamma distribution.

2.1.2. The entropy of Gamma distribution

Entropy is a measurement of the uncertainty associated with a random variable. The entropy $H(X)$ of the $X\sim\Gamma(\alpha,\beta)$ can be derived as

$$\begin{aligned} H(X) &= E(-\ln(f(X;\alpha,\beta))) \\ &= -E(-\alpha\ln\beta-\ln\Gamma(\alpha)+(\alpha-1)\ln X-X/\beta) \\ &= \alpha\ln\beta+\ln\Gamma(\alpha)+(1-\alpha)E(\ln X)+E(X)/\beta \end{aligned} \tag{5}$$

The $E(X)$ can be obtained by (see more details in Appendix):

$E(X)=\alpha\beta$.

We compute the $E(\ln X)$ as follows (see more details in Appendix)

$E(\ln X)=\ln\beta+\psi(\alpha)$

Hence,

$$H(X)=\alpha+\ln\beta+\ln\Gamma(\alpha)+(1-\alpha)\psi(\alpha) \tag{6}$$

2.1.3. The sample entropy of the switch waiting time

Let $x_1,x_2,\ldots,x_m$ be the switch waiting time samples collected from the voting sequence. $n_j=\{i|x_i=j,i=1,2,\ldots,m\}$, where the # means the number of the set. Note that $n_1+n_2+\cdots=m$. Then the sample entropy of the switch waiting time X is $H_s(X)=\sum_{n_k>0}(n_k/m)\ln(n_k/m)$.