BioSystems 106 (2011) 67–75

Contents lists available at ScienceDirect

BioSystems

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

A mathematical consideration of the word-composition vector method in comparison of biological sequences

Takuyo Aita[a,*], Yuzuru Husimi[b], Koichi Nishigaki[c]

[a] Graduate School of Science and Engineering, Saitama University, 255 Shimo-okubo, Saitama 338-8570, Japan
[b] Innovative Research Organization, Saitama University, Saitama 338-8570, Japan
[c] Department of Functional Materials Science, Faculty of Engineering, Saitama University, 255 Shimo-okubo, Saitama 338-8570, Japan

ARTICLE INFO

Article history:
Received 28 March 2011
Received in revised form 23 June 2011
Accepted 26 June 2011

Keywords:
Angle metric
Genome distance
Genome space
Optimal resolution
Sequence comparison
Word frequency

ABSTRACT

To measure the similarity or dissimilarity between two given biological sequences, several papers proposed metrics based on the "word-composition vector". The essence of these metrics is as follows. First, we count the appearance frequencies of all the *K-tuple* words throughout each of two given sequences. Then, the two given sequences are transformed into their respective word-composition vectors. Next, the distance metrics, for example the angle between the two vectors, are calculated. A significant issue is to determine the optimal word size *K*. With a mathematical model of mutational events (including substitutions, insertions, deletions and duplications) that occur in sequences, we analyzed how the angle between the composition vectors depends on the mutational events. We also considered the optimal word size (=resolution) from our original approach. Our results were verified by computational experiments using artificially generated sequences, amino acid sequences of hemoglobin and nucleotide sequences of 16S ribosomal RNA.

1. Introduction

When comparing biological sequences, many authors utilize the alignment-free sequence comparison (Vinga and Almeida, 2003; Mantaci et al., 2008; Guyon et al., 2009). Particularly, many researchers have adopted a method based on the "word-composition vector", the elements of which represent the appearance frequencies of all the *K-tuple* words throughout a given sequence. As distance metrics between the word-composition vectors for two sequences, we can find the Euclidean distance (Blaisdell, 1986), Mahalanobis distance (Wu et al., 1997), Kullback–Leibler divergence distance (Wu et al., 2005; Sims et al., 2009), correlation coefficient (van Heel, 1991), and angle metric (Stuart et al., 2002; Qi et al., 2004; Yu et al., 2010). For all of these studies, the effectiveness of the distance metrics is dependent on the word size, which is termed the "resolution" (Vinga and Almeida, 2003). A significant issue is to determine the optimal resolution *K* (Wu et al., 2005; Sims et al., 2009).

In this paper, we focused on the angle metric, which is defined as the angle between the word-composition vectors. By introducing a mathematical model of mutational events that occur in sequences, we analyzed how the angle depends on the mutational events and we considered the optimal resolution from our original approach.

Our approach is similar to several previous studies. Zharkikh and Rzhetsky (1993) gave a relationship between *K*-tuple metric and the number of point mutations. We extended their approach to more complicated mutational events that include substitutions, insertions, deletions and duplications. Wu et al. (2005) and Sims et al. (2009) examined the optimal resolution numerically and analytically. As for the former study, according to their criterion, the optimal resolution is determined to give the maximum values of Spearman's rank correlation between *K*-tuple metric and mutation rates. We adopted a similar criterion that the optimal resolution is determined to give the maximum correlation between *K*-tuple metric (angle metric) and the number of mutational events. As for the latter study, they presented the lower and upper limits of the optimal resolution analytically. According to their criterion, the lower limit gives the maximum number of different *K*-tuples that can be found in a sequence, while the upper limit makes the phylogenetic tree topologies converged. They tested the performance of their method under widely varying mutation rates. We also gave the lower and upper limits based on our original criterion.

This paper has four parts: the first part describes how to transform a sequence into a word-composition vector; the second part describes our model of mutational events and gives a mathematical relation between *K*-tuple metric (angle metric) and the mutational events; the third part shows the results of computational experiments using a set of artificially generated sequences and a set of naturally occurring sequences, to confirm the validity of our theory; the fourth part discusses the optimal resolution range.

* Corresponding author.
E-mail address: taita@mail.saitama-u.ac.jp (T. Aita).

0303-2647/$ – see front matter © 2011 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.biosystems.2011.06.009

2. Representation of a Sequence as a Word-composition Vector

Consider a biological sequence s with length of $\nu(s)$. The λ letters are available at each site. For example, $\lambda = 4$ for DNA or RNA sequences and $\lambda = 20$ for protein sequences. We define the "window" of length K along the sequence (see Fig. 1(a)). By sliding the window from the leftmost position to the rightmost position by shifting one position at a time, we count the appearance frequencies of all the "*K-tuple* words" (or simply "*K-tuples*"[1]) with fixed length $K \ll \nu(s)$. The parameter K is called the "resolution" (Vinga and Almeida, 2003; Mantaci et al., 2008; Sims et al., 2009). There are $W(s) = \nu(s) - K + 1$ windows throughout the whole sequence, while there are a total of $N = \lambda^K$ possible types of *K-tuple*. Let $f_i(s)$ be the appearance frequency of a *K-tuple* i throughout the whole sequence. The N-dimensional vector $(f_1(s), f_2(s), \cdots, f_N(s))$ represents the composition of all the *K-tuples* that appear in the sequence s: $\sum_{i=1}^{N} f_i(s) = W(s)$. As a metric to measure the similarity between sequences s and t, we use the cosine-angle metric, $\cos\theta(s, t|K)$, between their respective composition vectors with a given value of K (Stuart et al., 2002):

$$\cos\theta(s, t|K) \equiv \frac{1}{R(s)R(t)} \sum_{i=1}^{N} f_i(s) f_i(t), \tag{1}$$

where

$$R(s) \equiv \sqrt{\sum_{i=1}^{N} f_i(s)^2}.$$

3. Relationship Between $\cos\theta(s, t|K)$ and Differences Between Sequences

In this section, we describe how $\cos\theta(s, t|K)$ reflects the differences between two compared sequences. We consider general cases where a mutant sequence "t" is derived from an original sequence "s" by introducing the "mutations" that includes: substitutions, insertions, deletions and duplication. To describe this system mathematically, we define the "mutation-introduced regions (MIRs)" and four types of the "windows" as follows (see Fig. 1(b)).

Mutation-introduced regions (MIRs): The architectural changes of a sequence (s or t) are classified into substitutions, insertions and deletions (note that duplications are special cases of insertions). For example, we consider the following changes in a sequence: a consecutive substitution is introduced into the j_1 th–j_2 th sites, a deletion is introduced into the k_1 th–k_2 th sites, and an insertion is introduced at the joint between l th site and $l+1$ th site ($j_1 < j_2 < k_1 < k_2 < l$). We collectively designate these sites and joints as the "mutation-introduced regions" (MIRs) in this paper. We denote the m th MIR from the leftmost one through the sequence by $\mathrm{MIR}^{(m)}$ and denote the number of all the MIRs by M. That is, M is the number of the mutational events. In this study, we consider the M-value as the main measure to indicate the difference between the two sequences s and t. In the above case, the j_1 th–j_2 th sites are $\mathrm{MIR}^{(1)}$, the k_1 th–k_2 th sites are $\mathrm{MIR}^{(2)}$ and the joint between l th site and $l+1$ th site is $\mathrm{MIR}^{(3)}$, and $M = 3$. As special cases, if two or more different types of MIRs border each other, these MIRs are defined as a single MIR. For the above example, if $k_1 = j_2 + 1$, then the $\mathrm{MIR}^{(1)}$ and $\mathrm{MIR}^{(2)}$ are merged into the $\mathrm{MIR}^{(1)}$, and then $M = 2$.

Inter-MIRs distance: This distance is defined as the length of a region which is surrounded by two neighboring MIRs along the sequence (see Fig. 1(c)). We denote the inter-MIRs distance between $\mathrm{MIR}^{(m)}$ and $\mathrm{MIR}^{(m+1)}$ by H_m ($m = 0, 1, 2, \cdots, M$), where $\mathrm{MIR}^{(0)}$ and $\mathrm{MIR}^{(M+1)}$ represent the left and right termini of the sequence, respectively. It holds that $\sum_{m=0}^{M} H_m = \nu(s) - (L^{\mathrm{sub}} + L^{\mathrm{del}})$, where L^{sub} and L^{del} represents the total length of all the substituted regions and that of all the deleted regions in the sequence s, respectively.

Conserved windows: These are windows which do not include MIRs and are conserved after the mutagenesis. These windows appear in both sequences s and t.

Disappearing windows: These are defined in the original sequence s. Each of these windows includes at least one of the MIRs in s. Therefore, these windows disappear after the mutagenesis.

Appearing windows: These are defined in the mutant sequence t. Each of these windows includes at least one of the MIRs in t. Therefore, these windows appear after the mutagenesis. However, the appearing windows excludes the "duplicated windows" defined below.

Duplicated windows: These are defined in the mutant sequence t only when duplications are introduced in t. A duplication is conducted by copying a local string in s and pasting it at a joint somewhere in t. Each of these windows is located within the duplicated (inserted) regions in t. Therefore, these windows appear at least twice after the mutagenesis.

We denote the number of: conserved windows by W^{O}, disappearing windows by W^{-}, appearing windows by W^{+} and duplicated windows by $W^{\times}$. Obviously,

$$W(s) = W^{\mathrm{O}} + W^{-} \quad \text{and} \quad W(t) = W^{\mathrm{O}} + W^{+} + W^{\times}.$$

Examples of W^{-}, W^{+} and $W^{\times}$ for four simple cases are shown in Fig. 1(b).

Let Δf_i^{-} be the decrement of the appearance frequency of the *K-tuple* i coded in the disappearing windows. Let Δf_i^{+} and $\Delta f_i^{\times}$ be the increment of the appearance frequency of the *K-tuple* i coded in the appearing windows and that coded in the duplicated windows, respectively. Then, we obtain the following relationship:

$$f_i(t) = f_i(s) - \Delta f_i^{-} + \Delta f_i^{+} + \Delta f_i^{\times}, \tag{2}$$

$$W^{-} = \sum_{i=1}^{N} \Delta f_i^{-}, \qquad W^{+} = \sum_{i=1}^{N} \Delta f_i^{+}, \qquad W^{\times} = \sum_{i=1}^{N} \Delta f_i^{\times}, \tag{3}$$

$$W^{+} + W^{\times} - W^{-} = W(t) - W(s) = \nu(t) - \nu(s). \tag{4}$$

The values of Δf_i^{-}, Δf_i^{+} and $\Delta f_i^{\times}$ are dependent on the distribution of $f_i(s)$ ($i = 1, 2, \cdots, N$). Several papers, for example, Reinert et al. (2000) and Schbath (2000), discussed the statistical and probabilistic properties and frequency distributions of words in biological sequences. We take an original approach to description of the frequency distributions.

Hereafter, our theoretical scheme is based on

Assumption 1: The original sequence s is a randomly generated one. The *K-tuple* in each window is randomly chosen from a set of all possible $N = \lambda^K$ *K-tuples* (that is, we neglect the effects of overlaps of consecutive windows).

Then, the expected number of an arbitrary *K-tuple* that appears in s is given by $W(s)/N$. Therefore, we deal with two typical cases: the case of $W(s)/N \ll 1$ and that of $W(s)/N \gg 1$. These conditions are

[1] In some references, *K*-mer, *K*-plet, *K*-string or *K*-word.