Journal of Theoretical Biology 407 (2016) 318–327

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Comparison of genomic data via statistical distribution

Saeid Amiri [a,*], Ivo D. Dinov [b]

[a] University of Wisconsin-Green Bay, Department of Natural and Applied Sciences, Green Bay, WI, USA
[b] Statistics Online Computational Resource (SOCR), Michigan Institute for Data Science (MIDAS), School of Nursing, University of Michigan, Ann Arbor, MI 49109, USA

ARTICLE INFO

Article history:
Received 7 April 2016
Received in revised form
22 June 2016
Accepted 20 July 2016
Available online 25 July 2016

MSC:
62Gxx
62H10

Keywords:
Alignment-free
Clustering
Distance
K-tuple

ABSTRACT

Sequence comparison has become an essential tool in bioinformatics, because highly homologous sequences usually imply significant functional or structural similarity. Traditional sequence analysis techniques are based on preprocessing and alignment, which facilitate measuring and quantitative characterization of genetic differences, variability and complexity. However, recent developments of next generation and whole genome sequencing technologies give rise to new challenges that are related to measuring similarity and capturing rearrangements of large segments contained in the genome.

This work is devoted to illustrating different methods recently introduced for quantifying sequence distances and variability. Most of the alignment-free methods rely on counting words, which are small contiguous fragments of the genome. Our approach considers the locations of nucleotides in the sequences and relies more on appropriate statistical distributions. The results of this technique for comparing sequences, by extracting information and comparing matching fidelity and location regularization information, are very encouraging, specifically to classify mutation sequences.

1. Introduction

Sequence comparisons are paramount in genomic and clinical research, as such methods are widely used for functional annotation, phylogenetic studies, and assessments of disease risk. However, it is hard to agree on a unique and canonical distance metric because sequence distances may be dependent on different targets. This problem has attracted significant attention in the computational sciences, mainly driven by problems regarding the structure of biological sequences such as DNA, RNA, and proteins. Such biological structures can be represented by unidimensional sequences defined over a specific alphabet. Structural homologies between genomic sequences correspond to similar features as well as the functionality of the enzymes or proteins they represent. It is believed that common genetic features may be shared between different species, which reflects common evolutionary and functional mechanisms. Hence, researchers are looking for definitions of robust and efficient distance metrics defined on genetic sequences that are able to identify and quantify these common analogs. Locally, genomic data are typically stored as linked lists of categorical values, $\{A, T, C, G\}$, demanding scientific inference based categorical statistical analyses. Globally, it is hard to find a short signature vector representing the overall genomic sequence features. Hence, examining the empirical statistical distribution of sequence data is important and may provide complementary information to local base-pair measurements.

Recent studies (Chen et al., 2014c, 2015; Guo et al., 2014; Liu et al., 2015b,c) were focused on developing innovative methods for comparing DNA/RNA sequences. A review (Chen et al., 2015) of several successful applications of these techniques was used in genome analysis problems (Guo et al., 2014; Chen et al., 2012, 2014a,c, 2013; Yang et al., 2012; Ding et al., 2013; Qi et al., 2010; Chou, 2006; Lin et al., 2014). Computational methods for comparing protein/peptide sequences (Chou, 2001, 2005; Du et al., 2012; Cao et al., 2013; Du et al., 2014; Shen and Chou, 2008) have been employed to address various computational proteomics challenges (Chou, 2011; Liu et al., 2015a). A recently developed method, *Pse-in-One* (Liu et al., 2015a), was introduced to compare both DNA/RNA and protein/peptide sequences.

The traditional methods for comparing biological sequences are mostly based on an initial sequence alignment process, which fragments the data sequence to make it homologous to a reference (target) sequence using various cost functions and string matching algorithms (Ding et al., 2013; Qi et al., 2010). Some alignment-free sequence comparison methods have recently been introduced based on promoter frequency distance measures (Yang et al., 2012). Most sequence aligners consider only local mutations of the genome, which may not be suitable for measuring events and mutations that involve longer segments of genomic arrangements, our numerical study admits it. Furthermore, the aligning

* Corresponding author.
E-mail addresses: saeid.amiri1@gmail.com (S. Amiri), dinov@umich.edu (I.D. Dinov).

http://dx.doi.org/10.1016/j.jtbi.2016.07.032
0022-5193/

algorithms are very time-consuming for large scale data, see review given in Alimehr (2013) and Deonier et al. (2005) and references therein. For these reasons, alignment-free distance measures are of interest. Much of the literature on alignment-free sequence comparison addresses the features of word frequencies in the data; see Sims et al. (2009), Chor et al. (2009), Bonham-Carter et al. (2014) and references therein. Sequences are often represented by word-count vectors, and subsequent statistical inference relies on similarity scores defined for such feature vectors. Since these bases are not randomly distributed, it is natural to count the number of K-letter words (K-tuple) or any possible patterns that a pair of sequences have in common.

Most alignment-free methods are based on word-counting, which only considers the frequency of joint neighboring words without any correction of their locations in the sequence. Often, it is of interest to track the position of nucleotides in sequence and use the underlying distribution of data. A few studies have attempted to either develop new sequence distance measures based on word locations or to propose new entropy-based statistical tests. The focus of our research is to address this gap and enable modeling sequence distances using the distributions of various distance metrics (Chou, 2006). We compare our method to several alternative techniques using mammalian, bacterial, and viral genome sequences. The accuracy of the distance-based sequence homologies is evaluated using a clustering method. This clustering method represents an unsupervised technique for identifying natural classes within a set of data. The main idea is to group unlabeled data into subsets where the within-group sub-sequences are fairly homogeneous (in terms of their paired distance measures). By using a dissimilarity matrix and a distance matrix, we designed a hierarchical algorithm that may be used to combine clusters and thereby obtain a phylogenetic tree. The output of this algorithm is a dendrogram, which provides a way to explore the resulting hierarchical classification. In order to achieve a hierarchical clustering, linkage-based algorithms (average linkage) are used; see Amiri and Clarke (2015).

Consider two sequences, X and Y, with different lengths, L_X and L_Y, that can be represented

$$X = x_1 \ldots x_{L_X},$$
$$Y = y_1 \ldots y_{L_Y},$$

where $x_i, y_j \in \Sigma = \{A, T, C, G\}$, $i = 1, \ldots, L_X$, $j = 1, \ldots, L_Y$. We are looking for a distance metric $D(X, Y)$ representing the difference between these two sequences. For n sequences, we can compute a paired distance matrix where each entry corresponds to the pairwise distance $D(.,.)$. This dissimilarity matrix represents the distances between the n sequences. The organization of this paper is as follows: Section 2 proposes the distance frequency and the distance in terms of the neighborhood, K-tuple. Section 3 presents the distances that account for the location/position of nucleotides in whole sequence. The proposed methods are studied numerically in Section 4, where the results of a clustering classification are reported. Our concluding remarks are in Section 5.

2. *K*-tuple distances

Let p_i^X and p_i^Y, $i = 1, \ldots, 4$ be the relative frequency of A, T, C and G in the sequences X and Y, respectively. The distance between two sequences can be calculated via the Euclidean distance:

$$D(X, Y) = \sum_{i=1}^{4} (p_i^X - p_i^Y)^2.$$

To generalize it, one can split the sequence to $\mathcal{T}$ partitions such that:

$$X = \{\mathcal{X}_1, \ldots, \mathcal{X}_{\mathcal{T}}\},$$
$$Y = \{\mathcal{Y}_1, \ldots, \mathcal{Y}_{\mathcal{T}}\},$$

where $\mathcal{X}_1 = \{X_1, \ldots, X_{m_1}\}$, $\mathcal{X}_2 = \{X_{m_1+1}, \ldots, X_{m_2}\}$, …, $\mathcal{X}_{\mathcal{T}} = \{X_{m_{\mathcal{T}-1}+1}, \ldots, X_{m_{\mathcal{T}}}\}$, and $\mathcal{Y}_1 = \{Y_1, \ldots, Y_{n_1}\}, \ldots, \mathcal{Y}_{\mathcal{T}} = \{Y_{n_{\mathcal{T}-1}+1}, \ldots, Y_{n_{\mathcal{T}}}\}$. For brevity, only subsets with equal lengths are considered, i.e., $m_1 = \ldots = m_{\mathcal{T}} = \left[\frac{L_X}{\mathcal{T}}\right] = u$ and $n_1 = \ldots = n_{\mathcal{T}} = \left[\frac{L_X}{\mathcal{T}}\right] = v$. For each subset, there are $(n_{1t}, \ldots, n_{4t})'$, $t = 1, \ldots, \mathcal{T}$, that are frequencies corresponding to the nucleotides A, T, C and G, respectively. The distance of the X and Y can be calculated by

$$D(X, Y) = \sum_{i=1}^{4} \sum_{t=1}^{\mathcal{T}} \left(p_{t,i}^X - p_{t,i}^Y\right)^2,$$

where $p_{X,t}(i)$ and $p_{Y,t}(i)$ are the relative frequencies in the tth partition. Instead of using single nucleotides, one may consider a short word length ℓ and map each sub-sequence of length L_X, length of X into vectors of length ℓ to assess the similarity of sequences, which is referred to as K-tuple. For the K-tuple with 2 sliding windows, there are 4^2 situations, i.e., $\{AA, AT, \ldots, GG\}$; and for ℓ sliding windows, there are 4^ℓ situations. Let us define a string of size ℓ at the location i as $X[i, i + \ell - 1]$. Then, all possible (or interesting) K-tuples are defined by:

$$\mathcal{A}^\ell = \{\mathcal{A}_1, \ldots, \mathcal{A}_{\mathfrak{L}}\}, \tag{1}$$

where $\mathfrak{L} = 4^\ell$, define the count of them by

$$\nu_i = \{j: X[j \ldots j + \ell - 1] = \mathcal{A}_i\}, \quad i \in \{1, \ldots, \mathfrak{L}\},$$
$$v_i = |\nu_i|,$$

where $|.|$ is the cardinality of a set, in this case the number of elements of ν_i. Thus, the relative frequency can be found:

$$p_i^X = \frac{v_i}{L_X - \ell}, \quad i \in \{1, \ldots, \mathfrak{L}\},$$

where L_X is the length of sequence. Using these values, the distance can be obtained as follows:

$$D(X, Y) = \sum_{i=1}^{\mathfrak{L}} \left(p_i^X - p_i^Y\right)^2. \tag{2}$$

This is referred to as frequency distance DFR, and its performance is examined in Section 4. For large $\mathfrak{L}$, the relative frequencies become very small, and consequently, $D(X, Y)$ becomes too small as well. Hence, the absolute difference can be used to get less small distance. For example:

$$D(X, Y) = \sum_{i=1}^{\mathfrak{L}} |p_i^X - p_i^Y|. \tag{3}$$

Unlike the proposed approach, one can calculate the correlation,

$$\rho(X, Y) = \frac{\sum_{i=1}^{\mathfrak{L}} p_i^X p_i^Y}{\sqrt{\sum_{i=1}^{\mathfrak{L}} (p_i^X)^2 \sum_{i=1}^{\mathfrak{L}} (p_i^Y)^2}}.$$

The distance $D(X, Y)$ between the two sequences is defined as

$$D(X, X) = \frac{1 - \rho(X, Y)}{2}.$$

Such idea is used in CVTree (Xu and Hao, 2009), however they used different approaches to calculate the probabilities. This technique assumes the Markov property indicating that the conditioning of the probability distribution on past and present states depends only upon the present state, not on the sequence of events that preceded it, i.e., memoryless process assumption.