Published in final edited form as:

J Theor Biol. 2016 October 21; 407: 318–327. doi:10.1016/j.jtbi.2016.07.032.

Comparison of genomic data via statistical distribution

Saeid Amiri and

University of Wisconsin-Green Bay, Department of Natural and Applied Sciences, Green Bay, WI, USA

Ivo Dinov

Statistics Online Computational Resource (SOCR), School of Nursing, University of Michigan, MI 49109, Ann Arbor, USA

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 the mutation sequences.

Keywords

Alignment-free; clustering; distance; *K*-tuple

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

Correspondence to: Saeid Amiri.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

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 analogues. 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 [1]–[5] were focused on developing innovative methods for comparing DNA/RNA sequences. A review [3] of several successful applications of these techniques were used in genome analysis problems [[2]; [6]–[14]]. Computational methods for comparing protein/peptide sequences [[15]–[20]] have been employed to address various computational proteomics challenges [[21]–[22]]. A recently developed method, *Pse-in-One* [22], 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 [9] and [10]. Some alignment-free sequence comparison methods have recently been introduced based on promoter frequency distance measures, [8]. 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 admit it. Furthermore, the aligning algorithms are very time-consuming for large scale data, see review given [23] and [24] 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 [25]–[27] 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 and 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 [11]. 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 [28].

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

$$\begin{aligned} X &= x_1 \dots x_{L_X}, \\ Y &= y_1 \dots y_{L_Y}, \end{aligned}$$

where $x_i, y_j \in \Sigma = \{ATCG\}$, $i = 1, \dots, L_X, j = 1, \dots, L_Y$. 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

Let p_i^X and p_i^Y , $i = 1, \dots, 4$ 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:

$$\begin{aligned} X &= \{\mathcal{X}_1, \dots, \mathcal{X}_{\mathcal{T}}\}, \\ Y &= \{\mathcal{Y}_1, \dots, \mathcal{Y}_{\mathcal{T}}\}, \end{aligned}$$

where $\mathcal{X}_1 = \{X_1, \dots, X_{m_1}\}$, $\mathcal{X}_2 = \{X_{m_1+1}, \dots, X_{m_2}\}, \dots, \mathcal{X}_{\mathcal{T}} = \{X_{m_{\mathcal{T}-1}+1}, \dots, X_{m_{\mathcal{T}}}\}$, and $\mathcal{Y}_1 = \{Y_1, \dots, Y_{n_1}\}, \dots, \mathcal{Y}_{\mathcal{T}} = \{Y_{n_{\mathcal{T}-1}+1}, \dots, Y_{n_{\mathcal{T}}}\}$. For brevity, only subsets with equal lengths are considered, i.e., $m_1 = \dots = m_{\mathcal{T}} = \lceil \frac{L_X}{\mathcal{T}} \rceil = u$ and $n_1 = \dots = n_{\mathcal{T}} = \lceil \frac{L_Y}{\mathcal{T}} \rceil = v$. For each subset, there are $(n_1, \dots, n_4)'$, $t = 1, \dots, \mathcal{T}$, that are frequencies corresponding the nucleotides A, T, C and G , respectively. The distance of the X and Y can be calculated by

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

where $p_{X,t}(i)$ and $p_{Y,t}(i)$ are the relative frequencies in the t th 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., $\{AA, AT, \dots, GG\}$; and for ℓ sliding windows, there are 4^ℓ situations. Let's 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:

$$\mathcal{A}^\ell = \{\mathcal{A}_1, \dots, \mathcal{A}_{\mathfrak{L}}\}, \quad (1)$$

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

$$\begin{aligned} \nu_i &= \{j: X[j \dots j+\ell-1] = \mathcal{A}_i\}, & i \in \{1, \dots, \mathfrak{L}\}, \\ v_i &= |\nu_i|, \end{aligned}$$

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

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

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

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

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:

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

Unlike the proposed approach, one can calculate the correlation,

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

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

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

such idea is used in CVTree, [29], they used different approach to calculate the probabilities. This technique assumes the Markov property indicating that the conditioning 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.

3. Location based distances

Unlike the regular K -tuple that ignores information within a sequences, the distance in terms of the location of nucleotides or elements of K -tuple in a sequence is presented here. Let us consider a sequence of length L ; each nucleotide A , T , C and G can appear in different spots within sequence. Let's present the locations of these nucleotides as $L_A = \{a_1, a_2, \dots, a_{n_A}\}$, $L_T = \{t_1, t_2, \dots, t_{n_T}\}$, $L_C = \{c_1, c_2, \dots, c_{n_C}\}$ and $L_G = \{g_1, g_2, \dots, g_{n_G}\}$, respectively. The information of location can be used in the K -tuple's elements that are discussed in the below.

One study [30] tried to find the codon positions in DNA primary sequence, [31] presents the inter-nucleotide distance, these ideas were generalized in [32], who calculated the distance between sequences including some location information. To calculate this distance, let's consider the nucleotide A and define:

$$L_r = \frac{1}{a_r - a_{r-1}}, \quad r=1, \dots, n_A, \quad (4)$$

where $L_r \in (0, 1)$ and $a_0 = 0$ and clearly $a_r - a_{r-1}$, $r=1, \dots, n_A$, is the difference of sequential locations. Calculate the partial sum corresponding to

$$S_A = \{L_1, L_1 + L_2, \dots, \sum_r L_r\} = \{s_1, s_2, \dots, s_{n_A}\},$$

and $Z = \sum_{u=1}^{n_A} \sum_{r=1}^u L_r$, the discrete probability at the position r is defined as

$$p_r = \frac{s_r}{Z}, \quad r \in \{1, \dots, n_A\}.$$

The proposed entropy of nucleotides A of sequence X can be calculated as

$$h^X(A) = - \sum_{r=1}^{n_A} p_r \log p_r. \quad (5)$$

Using the same procedure, one can calculate the proposed entropy of other nucleotides. The distance of two sequences X and Y based on all nucleotides can be obtained by

$$D(X, Y) = \sum_{i \in \{A, T, C, G\}} (h_i^X - h_i^Y)^2.$$

The same procedure can be used for the K -tuple, i.e., find locations of elements of $\mathcal{A}^\ell$ in the sequence and calculate the proposed entropy, the distance is defined as

$$D(X, Y) = \sum_{i \in \mathcal{A}^\ell} (h_i^X - h_i^Y)^2, \quad (6)$$

where h_i^X and h_i^Y are the entropy values of the i th K -tuple in sequences X and Y , respectively. This method is referred to as DMk, (Distance Measure based on K -tuples).

[33] converted a DNA sequence into three sequences, representing the chemical bases by replacing $\{A, G\}$ and $\{C, T\}$ in sequence with R and Y , respectively. For the sake of simplicity refer to it as α . Other possibilities are; changing $\{A, C\}$ to M and $\{G, T\}$ to K and refer to β and finally change $\{A, T\}$ to W and $\{C, G\}$ to S and refer as γ . They carry out the DMk on each the based-converted sequences. For instance, consider $\ell = 2$, then the entropy of each combination of the new converted base via the approach DMk can be found according to the following:

$$\begin{aligned} H' &= (h_{RR}^X, h_{RY}^X, h_{YR}^X, h_{YY}^X) \quad \text{on } \alpha, \\ H'' &= (h_{MM}^X, h_{MK}^X, h_{KM}^X, h_{KK}^X) \quad \text{on } \beta, \\ H''' &= (h_{WW}^X, h_{WS}^X, h_{SW}^X, h_{SS}^X) \quad \text{on } \gamma, \end{aligned}$$

Using H' , H'' and H''' one can calculate the distance between the sequences, let denote as DPF

$$D(X, Y) = \sum_{i \in \{RR, RY, \dots, SS\}} (h_i^X - h_i^Y)^2.$$

Comparing distances in terms of difference of entropies may not be appropriate since two different sequences might have the same entropy, which would render the data

indistinguishable, see Section 4. Furthermore, it is not trivial to show how much sequence information is carried in the proposed quantity (5). An alternative approach is to use the mean (4), however, that might be too simple. Instead of relying on a single univariate statistic or parameter, a multivariate statistic may be used to capture these differences (e.g., quantiles, moments, etc.). The key issue of the alignment-free methods is to consider various sequence characteristics, extract appropriate features and generate an appropriate statistic that contains sufficient information representing the intrinsic characteristics of the genome. For instance, calculate L_r or it's reciprocal

$$\begin{aligned} L_r(X) &= \frac{1}{x_r - x_{r-1}} \in (0, 1), \\ L'_r(X) &= x_r - x_{r-1} \in (1, L_1). \end{aligned}$$

To have complete information of sequence should find $L_r(X)$ and $L'_r(X)$ for all combinations given in $\mathcal{A}^\ell$, hence it is more appropriate to present as $L_{i,r}(X)$ and $L'_{i,r}(X)$ and denote the set of them as

$$L_{i,+}(X) = \{L_{i,1}(X), L_{i,2}(X), \dots\}, \quad (7)$$

$$L'_{i,+}(X) = \{L'_{i,1}(X), L'_{i,2}(X), \dots\}. \quad (8)$$

This operation converts the sequence into a signature-vector of continuous or discrete values. Another approach is to consider the empirical distribution function of $L_{i,+}(X)$ or $L'_{i,+}(X)$ instead the single entropy proposed in DMk and DPF. We use a kind of Kolmogorov-Smirnov (KS) and Kullback-Leibler (KL) to measure sequence distribution distances. Other functions of the proposed quantities (7) and (8) may also be considered, for instance,

$\frac{1}{|x_r - x_{r-1}| + |x_{r+1} - x_r|}$ and $\frac{1}{x_{r+1} - x_{r-1}}$ may form a basis for computing distances in terms of these quantities. To use the information can use distance via KS is defined as:

$$D(X, Y) = \sup_z |\hat{F}_n^X(z) - \hat{F}_n^Y(z)|,$$

where $\hat{F}_n^X(z)$ and $\hat{F}_n^Y(z)$ are the empirical distribution function, let DKS denote the KS distance between $L_{i,+}(X)$ s and $L_{i,+}(Y)$ s. Since the KS on $L_{i,+}(\cdot)$ and $L'_{i,+}(\cdot)$ are the same, just the KS distance based on L_r is considered, the distance can be presented

$$\widehat{KS}(X, Y) = \sum_{i \in \mathcal{A}^\ell} D_i(X, Y) = \sum_{i \in \mathcal{A}^\ell} \sum_z \sup_z |\hat{F}_{n,i}^X(z) - \hat{F}_{n,i}^Y(z)|. \quad (9)$$

That is the summation of KS distance on $L_{i,+}(X)$ for all possible combinations of K -tuple that are represented A^l .

To see the formal presentation of Kullback-Leibler divergence, let $(\chi, \mathcal{F}, p_\theta)_{\theta \in \Theta}$ be a probability space, where Θ is an open convex subset of R^L and f_θ be the distribution p_θ with respect to σ -finite measure λ , such that $\int f_\theta(z) d\lambda(z) = 1$, where the integral is taken over the entire space. Relative entropy that refers to as the Kullback-Leibler divergence provides a non-symmetric measure of the difference between two pdfs that is defined as

$$D(f \| g) = \int f_\theta(z) \log \left(\frac{f_\theta(z)}{g_\theta(z)} \right) d\lambda(z), \quad (10)$$

where $\log$ is the natural logarithm. It is well-known that $D(f \| g) \geq 0$ and the equality holds if and only if $f_\theta(z) \stackrel{d}{=} g_\theta(z)$ almost everywhere. It is effectively measuring dissimilarity between $f_\theta(z)$ and $g_\theta(z)$, it is obvious that if $g_\theta(z)$ is constant, it becomes essentially entropy and if $f_\theta(z)$ and $g_\theta(z)$ are same then $D(f \| g) = 0$ hence it can be considered as distance between pdfs. $D(f \| g)$ tells us how surprising observed values of x will appear if we consider them as being generated according to $f_\theta(z)$ when in fact they come from $g_\theta(z)$. Under the null hypothesis, $f_\theta(z) \stackrel{d}{=} g_\theta(z)$, $D(f \| g) = 0$ and larger values favor the alternative research hypothesis, H_1 , $f_\theta(z) \stackrel{d}{\neq} g_\theta(z)$, the statistical test of entropy can be found in [34]. Let consider the exponential distribution for these distributions

$$z_1, \dots, z_n \sim \exp(\theta' T(x)) c(\theta) h(x) d\mu$$

and assume they difference between them are the parameter, then

$$D(f \| g) = \int f_{\theta_1}(z) \log \left(\frac{f_{\theta_1}(z)}{g_{\theta_0}(z)} \right) d\lambda(z) = (\theta_1 - \theta_0)' E_{\theta_1}(T(x)) + \log(c(\theta_1)) - \log(c(\theta_0)),$$

hence the Kullback-Leibler divergence is just the difference of parameters. Let p^X and p^Y be the empirical density function obtained from L_r , the distance of the X and Y can be found by:

$$\widehat{KL}(X, Y) = \sum_{i \in \mathcal{A}^l} D(p_i^X \| p_i^Y) = \sum_{i \in \mathcal{A}^l} \sum_z p_i^X(z) \log \left(\frac{p_i^X(z)}{p_i^Y(z)} \right). \quad (11)$$

Let DKL and DKLD be dissimilarity matrix of n sequences obtained via $\widehat{KL}(X, Y)$ of L_r and L'_r , respectively. The algorithm of DKLD is given in Algorithm 1 and next section examines the performance of the proposed methods.

4. Evaluation

We compare and evaluate the proposed methods on different datasets. The first dataset is the short Mitochondrial genome sequence. We generated several mutations of the genome and compared the algorithms in terms of their accuracy. The second dataset is the Herpesviridae family of viruses and finally the family of bacteria that has long sequences are considered. The discussed methods, DFR, DMk, DPF, DKS, DKL and DKLD provide the distance that can be used in the agglomerative method, it is used to combine clusters to unify them. The output of this algorithm is dendrogram that provides a way to explore the combining cluster. In order to achieve the hierarchical, the linkage-based algorithms is used that start with singleton clustering and combine them to get a dendrogram hierarchy (a tree diagram).

Algorithm 1

Calculate distance of two sequences X and Y via DKLD

-
- 1 For a given ℓ sliding windows, find $\mathcal{A}^\ell$ in (1)
 - 2 For $i \in \mathcal{A}^\ell$, calculate $L'_{i,+}(X)$ and $L'_{i,+}(Y)$ in (8)
 - 3 Calculate the empirical density function of $L'_{i,+}(X)$ and $L'_{i,+}(Y)$ and present as p_i^X and p_i^Y .
 - 4 Calculate $\widehat{KL}(X, Y)$ in (11)
-

4.1. Evaluation on Mitochondrial genome sequences

We compared the results of all different methods and assessed their ability to distinguish different genomes via a clustering method using the proposed distances. The Mitochondrial genome sequences of seven different mammalian species are considered that were downloaded from NCBI, see Table 1 for details. Such data might be appropriate to study the proposed methods, because the Mitochondrial DNA is not highly conserved and has a rapid mutation rate, thus it is very useful for study the evolutionary relationships of organisms, see [35]. To obtain a criterion to compare the alternative methods, here we evaluate the prediction of the proposed method and it is expected the accurate one provides better prediction, to carry out it, we need sequences with known labels. We considered the sequences given in Table 1 as clusters and four new sequences are generated via mutation process from them. To generate a mutated sequence, we consider a sequence of length L and 20% of its nucleotides are selected randomly, then they are replaced with A, T, C, G randomly, since we know the origin of mutated sequences, we expect an accurate clustering method assigns the mutated sequences to same group. So there are 28 mutated sequences with known origin (clusters). They are appropriate for the evaluation of proposed methods, because one can compare the prediction via proposed methods with the true label. We obtained distance matrix using the proposed methods and carried out the clustering method, UPGMA. To find the underlying cluster, we used `cutree` to select k clusters in R that for the given k , in this case $k = 7$, because the mutations are generated from 7 species. We generated many mutations so it expected better performance for accurate one in average. It should be noted and the dendrogram provide a visual tree and researcher might prefer selecting clustering by eyes and not by `cutree`, the `cutree` does not perform accurately

under chaining problem in dendrogram, although we have not had such problem in our data, because we generated mutation of original sequences. These data are appropriate for the study of the proposed methods, since we know the origin of sequences and we expect an accurate clustering put them in the same group. To explore the proposed method, which one recognize the true label, here 28 sequences with label of $\{I, II, \dots, V III\}$ are generate, the proposed methods are run on these sequences and then four clusters with label $\{A, \dots, G\}$ are generated, to match $\{A, \dots, G\}$ with the true labels, using the Hungarian algorithm [36], and the results are shown in Table 2. The matching quality is assessed using the classification rate (CR), which is defined as the proportion of observations from a dataset that were correctly assigned to their cluster or class

$$CR = \frac{\sum_{i=1}^4 a_{ii}}{\sum_{i=1}^4 \sum_{j=1}^4 a_{ij}}.$$

To decrease the random effect, the whole procedure (generating mutation and clustering them) is repeated 500 times and the averages of CRs are given in the Table 3, the standard deviation of CRs is presented in parenthesis. These results, presenting different K -tuples, illustrate that the DKLD method outperforms the alternatives on the mutated sequences. We also run the "ClustalOmega", see [37]) which is a multiple sequence alignment and is implemented in *EMBL-EBI*¹, the accuracy is 0.590(0.04) that is a bit low, so obviously the distances obtained via the K -tuple recognize the underlying cluster better than the alignment method. It is of interest to explore the original Mitochondrial genome sequences, the dendrogram using alignment and DKLD with $\ell = 1, \dots, 4$ are presented in Figure 1, the result is interesting, it is known that Human and chimpanzee are close, under the alignment, they separated, but under DKLD with $\ell = 3$ and 4, they are closer than other species.

In Table 3, we considered different sliding window sizes. It is hard to discuss about the appropriate choice of sliding window sizes, in earlier work, [32] and [33] suggested DMk and DPF with $\ell = 2$ and $\ell = 3$, however they only considered short sequences and the study of proposed models on a large scale DNA data and DNA sequence similarity are postponed to future. Obviously wider sliding windows sizes might not be appropriate specially for short sequences because a lot of elements of $\mathcal{A}^\ell$ would have zero frequencies and the proposed distance metric may degenerate – it may not distinguish between closer and farther sequences. Furthermore, the calculation complexity increases rapidly in these situations. In order to see the effect of increasing the sliding window, the proportion of the low window sizes is calculated,

$$r_\ell = \frac{\sum_{i=1}^{\mathcal{L}} I(|\nu_i| < 15)}{\mathcal{L}}. \quad (12)$$

¹<http://www.ebi.ac.uk/Tools/msa/clustalo/>

This value shows the proportion of sliding windows for ℓ that have small size, less 15, for the calculation of goodness of fit. Our empirical results, now shown, indicate that the proposed method should not be used for ℓ where $\eta > 0.50$. In the Table 3, for $\ell = 5$, we have $\mathcal{L}$ and 57% of them have few elements, less 15, so obviously for 57% of our sliding windows can not provide appropriate information for the proposed distance, hence our distance become degenerate the distance and it can not provide good distance, for $\ell > 5$, we get more sliding windows with small size and more degenerate distance. For the calculation of η we consider ($|\mathbf{v}_i| < 15$), although it make sense to consider a bit larger, for instance ($|\mathbf{v}_i| < 25$) because we are using the empirical distribution.

4.2. Evaluation on Herpesviridae family of viruses

We consider the Herpesviridae family of viruses that cause diseases many species. Majority of adult humans (over 90%) have been infected with some form of herpesvirus which may remain dormant in a latent form in most previously infected subjects, see [38]. In total, there are 8 herpesvirus types that infect humans: Herpesviridae is a large and varied family in length among virus family. According to the pathogen virus resource ², the family of Herpesviridae is divided into three main families (Alphaherpesvirinae, Betaherpesvirinae and Gammaherpesvirinae). We restricted our analysis to the distinct and complete genome of Alphaherpesvirinae. Alphaviridae has more complete genomes than either Betaherpesviridae or Gammaherpesviridae. Within Alphaherpesviridae, there are five genera: Iltovirus, Mardivirus, Scutavirus, Simplexvirus, and Varicellovirus.

We disregard the 'Scutavirus' as we did not find its complete genome. The other forms were denoted by 'Ilt', 'Mar', 'Sim' and 'Var' with frequencies of 19, 17, 20, 40, respectively. The genomes are from different hosts, human, Monkey, chicken, Duck, cow, Bat (Fruit), equidae (horse), Boar, Cat family (denoted hum, mon, chi, duc, cow, bat, equ, boa, cat, respectively in the dendrograms). These viruses have different lengths, the minimum and maximum length are 124,784 bp and 178,311 bp, respectively.

Figure 2 represents the dendrogram obtained from the multiple alignment done using the MAFF³. The proposed distances with $K = 5$ were calculated and their corresponding dendrograms are presented in Figures 3–8. The leave nodes are labeled with the genome Genus or subpopulation and host. Assuming accurate distance calculations we expected to observe clear separation between genera and host species. Inspecting the results we see that DKLD provides most separation, for instance 'Sim' and 'Var' are separated better in dendrogram given in Figure 8.

To compare to method precisely, we generate mutations using the procedure explained in Section 4.1, we consider a sequence of length L and 20% of its nucleotides are selected randomly and replaced with A, T, C, G randomly, since we know the origin of mutated sequences, we expect an accurate clustering method assigns the mutated sequences to same group. So there are 36 mutated sequences with known origin (clusters). The result are given in Table 4,

²<http://vprbrc.org>

³<http://mafft.cbrc.jp/alignment/software/>

4.3. Evaluation on Bacteria family

Here the performance of proposed methods are examined on the genomes of different groups of bacteria provided⁴. We selected 'Bacteria-Aquificae', 'Bacteria-Bacteroidetes' and 'Bacteria-Cyanobacteria', the group and genome are given in Table 5, for the sake of simplicity, they are labeled as A1, ..., C8 and the proposed family has the long lengths, these sequences are diverse and appropriate to evaluate the proposed method. The distance are calculated via the proposed methods and corresponding dendrograms are given in Figures 9–11, obviously, DKLd provides better separation of the bacteria, for instance the bacteria with label A, 'Bacteria-Aquificae', are well-separated using DKLd and obviously they are put under same branch in the dendrogram presented in Figure 11 that means the calculated distance via DKLd provides more specificity, the same discussion can be done for other bacteria. It should be denoted that we tried couple multiple alignment MAFFT, CLUSTAL and MUSCLE on these data but after one week none of them succeeded to get alignment of them although the proposed method provide result in less than hour in Mac 2.3 GHz Intel Core i7 with 16 GB, it can be done in few moments see conclusion section for the discussion.

5. Conclusion

In this paper, we discuss the importance of considering the locations of the nucleotides or K -tuples in the sequence when considering homologous sites. Unlike the regular K -tuple methods that calculate distance of sequences in terms of euclidean, entropy of them, or correlation, we explore a way of computing sequence distances using the empirical density function. We show that this method can provide more accurate distance metrics. We employed sequence clustering based on new distance metrics to compare various alternative algorithms for establishing homologous segments. Our experiments used raw and mutated sequences where the distances are exactly known.

We showed that locational K -tuple can extract additional genomic information. K -tuple and locational K -tuple can be used simultaneously, since both approaches are not very computationally time consuming compared to alignment-based methods. Under suspicious situation, both methods can be used to get appropriate.

As dedicated supporters of open-science, we have deployed a website⁵. to release, maintain and support the complete end-to-end software implementing all steps of the proposed analytic protocol. We are committed to open-science and support the recent trend to develop, validate and share new analysis methods using user-friendly and publicly accessible web-servers. We will make an effort in our future work to provide a web-server for the prediction method presented in this paper.

In terms of computational complexity, we were unsuccessful in completing the alignment of the Bacteria genomes (Subsection 4.3) within one week using a dedicated HPC server (20-core, 2GHz, 1.5TB RAM). On the other hand, our locational K -tuple method is quite fast

⁴<http://archaea.ucsc.edu/cgi-bin/hgGateway>

⁵<https://sites.google.com/site/saeidamiri1/codes/lktuple>

and can generate results on a desktop machine in about 2 hours for the same bacterial genome. In many cases, the algorithm generates results in less than 10 minutes. This computational efficiency of the proposed method is driven by the low complexity of the statistics for the elements of $\mathcal{A}^\ell$, which can be calculated in parallel. Our implementation is based on R and accommodate parallel calculations that speed up the calculations by adding more cores, see example provided in the website of paper. Further optimization may be achieved by utilizing the high-throughput computational resources that are freely accessible and widely available, see OSG⁶ and Pipeline⁷, see [39].

There is no exact protocol on how to choose the parameter ℓ . However, larger ℓ values yield increasing the computational complexity and improve the accuracy of the algorithm. Whereas smaller ℓ values expedite the computational process but may fail to find homologues in the reference sequence. In order to explore the proposed method we examined different ℓ values. As it is shown numerically the accuracy of K -tuple may depend on the ratio of window-size of ℓ to the best of our knowledge, it is not explored clearly and as discussed earlier, the existing method suggest $\ell = 2$ and 3 and unfortunately the insufficient number ℓ may impact the distance measures. We explored ℓ and showed that r_ℓ can be used to select the informative ℓ for locational K -tuple, when $r_\ell > 0.5$, it means that 50% of observation is not informative to calculate the distance via locational K -tuple.

Under suspicious of ℓ a combination of ℓ parameters may be considered via the ensemble approach. Here the distance matrix (a measure of dissimilarity) may be calculated and the sum of different ℓ may be considered as a total distance metric. In order to avoid noise or other artifacts, the standardized or normalized distance matrices should be used.

Acknowledgments

We gratefully acknowledge the constructive comments and suggestions of the anonymous referees and the associate editor. We would also like to thank to Jennifer Clarke for her comments on the beginning project.

References

1. Chen W, Lei TY, Jin DC, Lin H, Chou KC. PseKNC: a flexible web server for generating pseudo K -tuple nucleotide composition. *Analytical biochemistry*. 2014; 456:53–60. [PubMed: 24732113]
2. Guo SH, Deng EZ, Xu LQ, Ding H, Lin H, Chen W, Chou KC. iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k -tuple nucleotide composition. *Bioinformatics*. 2014:btu083.
3. Chen W, Lin H, Chou KC. Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences. *Molecular BioSystems*. 2015; 11(10):2620–2634. [PubMed: 26099739]
4. Liu B, Liu F, Fang L, Wang X, Chou KC. repDNA: a Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects. *Bioinformatics*. 2015; 31(8):1307–1309. [PubMed: 25504848]
5. Liu B, Liu F, Fang L, Wang X, Chou KC. repRNA: a web server for generating various feature vectors of RNA sequences. *Molecular Genetics and Genomics*. 2015:1–9.

⁶<http://www.opensciencegrid.org/>

⁷<http://pipeline.loni.usc.edu/>

6. Chen W, Lin H, Feng PM, Ding C, Zuo YC, Chou KC. iNuc-PhysChem: a sequence-based predictor for identifying nucleosomes via physicochemical properties. *PLoS one*. 2012; 7(10):e47843. [PubMed: 23144709]
7. Chen W, Feng PM, Lin H, Chou KC. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. *Nucleic acids research*. 2013; gks1450.
8. Yang L, Zhang X, Zhu H. Alignment free comparison: Similarity distribution between the DNA primary sequences based on the shortest absent word. *Journal of theoretical biology*. 2012; 295:125–131. [PubMed: 22138094]
9. Ding S, Li Y, Yang X, Wang T. A simple k-word interval method for phylogenetic analysis of DNA sequences. *Journal of theoretical biology*. 2013; 317:192–199. [PubMed: 23085256]
10. Qi ZH, Qi XQ, Liu CC. New method for global alignment of 2 DNA sequences by the tree data structure. *Journal of theoretical biology*. 2010; 263(2):227–236. [PubMed: 20025888]
11. Chou K. Structural bioinformatics and its impact to biomedical science and drug discovery. *Frontiers in medicinal chemistry*. 2006; 3:455.
12. Chen W, Feng PM, Deng EZ, Lin H, Chou KC. iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. *Analytical Biochemistry*. 2014; 462:76–83. [PubMed: 25016190]
13. Chen W, Feng PM, Lin H, Chou KC. iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition. *BioMed research international*. 2014; 2014
14. Lin H, Deng EZ, Ding H, Chen W, Chou KC. iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. *Nucleic acids research*. 2014; 42(21):12961–12972. [PubMed: 25361964]
15. Chou KC. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins: Structure, Function, and Bioinformatics*. 2001; 43(3):246–255.
16. Chou KC. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics*. 2005; 21(1):10–19. [PubMed: 15308540]
17. Du P, Wang X, Xu C, Gao Y. PseAAC-Builder: A cross-platform stand-alone program for generating various special Chou's pseudoamino acid compositions. *Analytical biochemistry*. 2012; 425(2):117–119. [PubMed: 22459120]
18. Cao DS, Xu QS, Liang YZ. propy: a tool to generate various modes of Chou's PseAAC. *Bioinformatics*. 2013; 29(7):960–962. [PubMed: 23426256]
19. Du P, Gu S, Jiao Y. PseAAC-General: fast building various modes of general form of Chou's pseudo-amino acid composition for large-scale protein datasets. *International Journal of Molecular Sciences*. 2014; 15(3):3495–3506. [PubMed: 24577312]
20. Shen HB, Chou KC. PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition. *Analytical biochemistry*. 2008; 373(2):386–388. [PubMed: 17976365]
21. Chou KC. Some remarks on protein attribute prediction and pseudo amino acid composition. *Journal of theoretical biology*. 2011; 273(1):236–247. [PubMed: 21168420]
22. Liu B, Liu F, Wang X, Chen J, Fang L, Chou KC. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic acids research*. 2015; 43(W1):W65–W71. [PubMed: 25958395]
23. Alimehr, Leila. The Performance of Sequence Alignment Algorithms. Uppsala University, Department of Information Technology; 2013.
24. Deonier, RC., Tavaré, S., Waterman, M. Computational genome analysis: an introduction. Springer Science & Business Media; 2005.
25. Sims GE, Jun SR, Wu GA, Kim SH. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proceedings of the National Academy of Sciences*. 106(8): 2677–2682.
26. Chor B, Horn D, Goldman N, Levy Y, Massingham T. Genomic DNA k-mer spectra: models and modalities. *Genome Biol*. 2009; 10(10):R108. [PubMed: 19814784]
27. Bonham-Carter O, Steele J, Bastola D. Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis. *Briefings in bioinformatics*. :btt052.

28. Amiri S, Clarke B, Clarke J, Koepke HA. A General Hybrid Clustering Technique. 2015 arXiv. 1503.01183.
29. Xu Z, Hao B. CVTree update: a newly designed phylogenetic study platform using composition vectors and whole genomes. *Nucleic acids research*. 2009; 37(suppl 2):W174–W178. [PubMed: 19398429]
30. Li C, Wang J. Relative entropy of dna and its application. *Physica A: Statistical Mechanics and its Applications*. 2005; 347:465–471.
31. Nair ASS, Mahalakshmi T. Visualization of genomic data using inter-nucleotide distance signals. *Proceedings of IEEE Genomic Signal Processing*. 2005:408.
32. Wei D, Jiang Q, Wei Y, Wang S. A novel hierarchical clustering algorithm for gene sequences. *BMC bioinformatics*. 13(1):174.
33. Bao J, Yuan R, Bao Z. An improved alignment-free model for dna sequence similarity metric. *BMC bioinformatics*. 15(1):1.
34. Amiri S. The resampling of entropies with the application of biodiversity. *Journal of Applied Statistics*. 2014; 41(2):407–422.
35. Brown WM, Prager EM, Wang A, Wilson AC. Mitochondrial DNA sequences of primates: tempo and mode of evolution. *Journal of molecular evolution*. 1982; 18(4):225–239. [PubMed: 6284948]
36. Kuhn HW. The Hungarian method for the assignment problem. *Naval Research Logistics (NRL)*. 2005; 52(1):7–21.
37. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, ... Thompson JD. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Molecular systems biology*. 2011; 7(1):539. [PubMed: 21988835]
38. Pellett PE, Ablashi DV, Ambros PF, Agut H, Caserta MT, Descamps V, ... Kuehl U. Chromosomally integrated human herpesvirus 6: questions and answers. *Reviews in medical virology*. 2012; 22(3):144–155. [PubMed: 22052666]
39. Dinov ID, Petrosyan P, Liu Z, Eggert P, Hobel S, Vespa P, Moon SW, Van Horn JD, Franco J, Toga AW. High-throughput neuroimaging-genetics computational infrastructure. *Frontiers in neuroinformatics*. 2014; 8:41–41. [PubMed: 24795619]

Figure 1.
Dendrogram for the alignment and DKLd with different ℓ on the Mitochondrial.

J Theor Biol. Author manuscript; available in PMC 2017 October 21.

J Theor Biol. Author manuscript; available in PMC 2017 October 21.

J Theor Biol. Author manuscript; available in PMC 2017 October 21.

J Theor Biol. Author manuscript; available in PMC 2017 October 21.

Figure 7.
Dendrogram for the Alphaherpesviridae data set using DKL.

Figure 8.
Dendrogram for the Alphaherpesviridae data set using DKLd

Figure 9.
Dendrogram for the Bacteria family using DFR.

Figure 10.
Dendrogram for the Bacteria family using *CPF*

Figure 11.
Dendrogram for the Bacteria family using DKLd

Table 1

Mitochondrial genome of species used in the clustering method

Species	GenBank	Length
Ape	NC 002764.1	16586
Baboon	Y18001.1	16521
Chimpanzee	D38113.1	16554
Gibbon	X99256.1	16472
Gorilla	D38114.1	16364
Human	V00662.1	16569
Orangutan	NC 002083.1	16499

Table 2

Cross-table of estimated cluster vs true cluster

estimated cluster	true cluster			
	I	II	...	VII
I	a_{11}	a_{12}	...	a_{17}
II	a_{21}	a_{22}	...	a_{27}
.	.	.	.	.
.	.	.	.	.
IV	a_{71}	a_{72}	...	a_{77}

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 3

The accuracy of proposed methods on the Mitochondrial genome mutated sequences

Method	ℓ			
	2	3	4	5
DFR	0.284 (0.03)	0.358(0.03)	0.438(0.04)	0.581(0.05)
DMk	0.274(0.02)	0.337(0.03)	0.429(0.04)	0.492(0.03)
DPF	0.288(0.02)	0.307(0.03)	0.332(0.03)	0.377(0.03)
DKS	0.347(0.03)	0.426(0.04)	0.515(0.03)	0.421(0.03)
DKL	0.544(0.08)	0.532(0.05)	0.813(0.06)	0.557(0.07)
DKLd	0.784(0.05)	0.939(0.05)	0.906(0.09)	0.575(0.08)
$\mathcal{L}$	16	64	256	1024
η	0	0	0.12	0.57

The value inside the parenthesis is the standard deviation of accuracy of simulated mutations.

Table 4

The accuracy of proposed methods on the Herpesviridae family of mutated viruses

Method	ℓ					
	2	3	4	5	6	
DFR	0.525(0.029)	0.519(0.025)	0.507(0.028)	0.518(0.028)	0.527(0.043)	
DMk	0.638(0.036)	0.614(0.038)	0.572(0.034)	0.559(0.026)	0.490(0.0143)	
DPF	0.629(0.013)	0.611(0.001)	0.613(0.012)	0.628(0.027)	0.601(0.057)	
DKS	0.550(0.032)	0.561(0.023)	0.556(0.014)	0.491(0.024)	0.398(0.014)	
DKL	0.490(0.021)	0.575(0.024)	0.572(0.029)	0.704(0.092)	0.407(0.057)	
DKLd	0.575(0.048)	0.856(0.035)	0.750(0.053)	0.804(0.019)	0.550(0.040)	
χ^2	16	64	25	1024	4096	
r_t	0	0	0	0.145	0.493	

The value inside the parenthesis is the standard deviation of accuracy of simulated mutations.

Table 5

Bacteria genome used in the clustering method

group	Genome	GenBank	Length	Label
Bacteria-Aquificae	Aquifex aeolicus	NC 000918.1	1551335	A1
	Hydrogenobacter thermophilus TK-6	NC 013799.1	1743135	A2
	Hydrogenobaculum sp. 3684	NC 015557.1	1552775	A3
	Hydrogenobaculum sp. Y04AAS1	NC 011126.1	1559514	A4
	Persephonella marina EX-H1	NC 012440.1	1930284	A5
	Sulfurihydrogenibium sp. YO3AOP1	NC 010730.1	1838442	A6
	Sulfurihydrogenibium azorense Az-Fu1	NC 012438.1	1640877	A7
Bacteria-Bacteroidetes	Bacteroides thetaiotaomicron VPI-5482	NC 004663.1	6260361	B1
	Cytophaga hutchinsonii ATCC 33406	NC 008255.1	4433218	B2
	Flavobacterium johnsoniae UW101	NC 009441.1	6096872	B3
	Gramella forsetii KT0803	NC 008571.1	3798465	B4
	Porphyromonas gingivalis W83	NC 002950.2	2343476	B5
	Salinibacter ruber DSM 13855	NC 007677.1	3551823	B6
Bacteria-Cyanobacteria	Anabaena variabilis ATCC 29413	NC 007413.1	6365727	C1
	Nostoc sp. PCC 7120	NC 003272.1	6413771	C2
	Gloeobacter violaceus PCC 7421	NC 005125.1	4659019	C3
	Prochlorococcus marinus CCMP1375	NC 005042.1	1751080	C4
	Synechococcus elongatus PCC 7942	NC 007604.1	2695903	C5
	Synechococcus sp. WH 8102	NC 005070.1	2434428	C6
	Thermosynechococcus elongatus	NC 004113.1	2593857	C7
	Trichodesmium erythraeum IMS101	NC 008312.1	7750108	C8

Sulfurihydrogenibium yellowstonense SS-5, Hydrogenivirga sp. 128-5-R1-1 and Crocosphaera watsonii WH 8501 are not added because the complete genome is not available or has not identified