Information Sciences 181 (2011) 1484–1492

Contents lists available at ScienceDirect

Information Sciences

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

DNA sequence comparison by a novel probabilistic method

Chenglong Yu [a], Mo Deng [b], Stephen S.-T. Yau [b,*]

[a] The Institute of Mathematical Sciences, The Chinese University of Hong Kong, Shatin, Hong Kong
[b] Institutes of Mathematics, East China Normal University, Shanghai, China

ARTICLE INFO

Article history:
Received 14 April 2010
Received in revised form 21 October 2010
Accepted 17 December 2010
Available online 24 December 2010

Keywords:
DNA
Sequence comparison
Graphical representation
Probability distribution
Kullback–Leibler divergence

ABSTRACT

This paper proposes a novel method for comparing DNA sequences. By using a graphical representation, we are able to construct the probability distributions of DNA sequences. These probability distributions can then be used to make similarity studies by using the symmetrised Kullback–Leibler divergence. After presenting our method, we test it using six DNA sequences taken from the threonine operons of *Escherichia coli* K-12 and *Shigella flexneri*. Our approach is then used to study the evolution of primates using mitochondrial DNA data. Our method allows us to reconstruct a phylogenetic tree for primate evolution. In addition, we use our technique to analyze the classification and phylogeny of the Tomato Yellow Leaf Curl Virus (TYLCV) based on its whole genome sequences. These examples show that large volumes of DNA sequences can be handled more easily and more quickly by our approach than by the existing multiple alignment methods. Moreover, our method, unlike other approaches, does not require human intervention, because it can be applied automatically.

1. Introduction

With the development of biotechnology, more and more biological sequence information has been acquired. The number of sequences in GenBank has been growing exponentially in the past 20 years (http://www.ncbi.nlm.nih.gov). Many computational and statistical methods have been proposed for comparing biological sequences. Nevertheless, biological sequence comparison remains one of the most active and important research areas in bioinformatics and computational biology. Existing methods for sequence comparison (i.e., studying the similarity/dissimilarity of sequences) can be classified into alignment-based methods and alignment-free methods.

Alignment-based methods use dynamic programming, a regression technique that finds an optimal alignment by assigning scores to different possible alignments and picking the alignment with the highest score. Several algorithms have been developed that target specific goals such as global alignment, local alignment, with or without overlap [20,27,12]. Subsequently, some heuristic approaches were proposed, based on the recognition of alignment "seeds", with BLAST [1,2] and FASTA [21,22] being the most widely applications. However, the search for optimal solutions using sequence alignment turns out to be computationally difficult with large biological databases, especially when comparing three or more biological sequences at a time, i.e., multiple sequence alignment. Therefore, alignment-free approaches have been developed to overcome the critical limitations of alignment-based methods.

Among all existing alignment-free methods for comparing biological sequences [10,28,23,5,25], sequence graphical representation provides a simple way to view, sort, and compare gene structures [11,13,29,24,18,15,14]. The aim of graphical representation is to display DNA or protein sequences graphically so that we can easily find out visually how similar or

* Corresponding author. Current address: Department of Mathematics, Statistics and Computer Science, University of Illinois at Chicago, IL 60607-7045, USA. Tel./fax: +1 312 996 3065.
E-mail address: yau@uic.edu (S.S.-T. Yau).

0020-0255/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2010.12.010

how different they are. Of course, only performing visual comparison of sequences is not enough for the follow-up research work. We need more a precise way of making the comparison. In our previous work, we transformed a DNA sequence into a vector (feature vector [19,6] or moment vector [30,31]), and then used the Euclidean or Mahalanobis distance between these vectors as an index for comparing DNA sequences. Recently Pham and Zuegg [23] introduced a probabilistic measure of similarity between two DNA sequences without alignment. Their method is based on the concept of comparing the similarity/dissimilarity between two constructed Markov models. Their work gave us a fresh idea that we can apply to DNA sequence comparison. For each DNA sequence, we can construct its "probability distribution" from its graphical representation. Then we use the Kullback–Leibler divergence (relative entropy) to get a new measure of similarity/dissimilarity among various DNA sequences based on their probability distributions.

In our previous work [29], we constructed a pyrimidines–purine graph using two quadrants of the Cartesian coordinate system, with pyrimidines (T and C) in the 1st quadrant and purines (A and G) in the 4th quadrant. In this paper, we make a minor modification of the previous method. We assign the four nucleotides only in the 1st quadrant of Cartesian coordinate system. This small change in the graphical representation gives us a breakthrough because it allows us to construct a probability distribution for the DNA sequence. After obtaining the probability distributions of DNA sequences we use the symmetrised kullback–Leibler divergence [16] to perform similarity studies. Our approach is tested with six DNA sequences taken from the threonine operons of *E. coli* K-12 and *S. flexneri*. We also use our method to study primate evolution using mitochondrial DNA data. A phylogenetic tree of primate evolution is reconstructed using our new method. Finally, we apply our new technique to analyze the classification and phylogeny of Tomato Yellow Leaf Curl Virus (TYLCV) based on whole genome sequences of the virus. The results show that our approach can be used to study the clustering and phylogenetic relationship when you have a large volume of DNA sequence data.

2. Materials and methods

2.1. New graphical representation of DNA sequence

We have constructed the new DNA sequence graphical representation in the first quadrant of the Cartesian coordinate system. Fig. 1 shows the four vectors corresponding to the four nucleotides A, G, C, and T are as follows: A (1,0.8), G (1,0.6), C (1,0.4), T (1,0.2).

The points in the graphical representation are obtained by summing the vectors representing nucleotides in the sequence. The endpoint of every vector represents one nucleotide. Fig. 2 shows the graphical representation of the DNA sequence (ATGGTGCACC) which consists of the first 10 nucleotides of human beta-globin coding sequence. The graphical curve has no circuits or degeneracy and the correspondence between the sequence and the graphical curve can be mathematically proved to be one-to-one [29].

2.2. Probability distribution of DNA sequence

For a DNA sequence of length n, we define its probability distribution as $(p_1, p_2, \ldots, p_n)$,

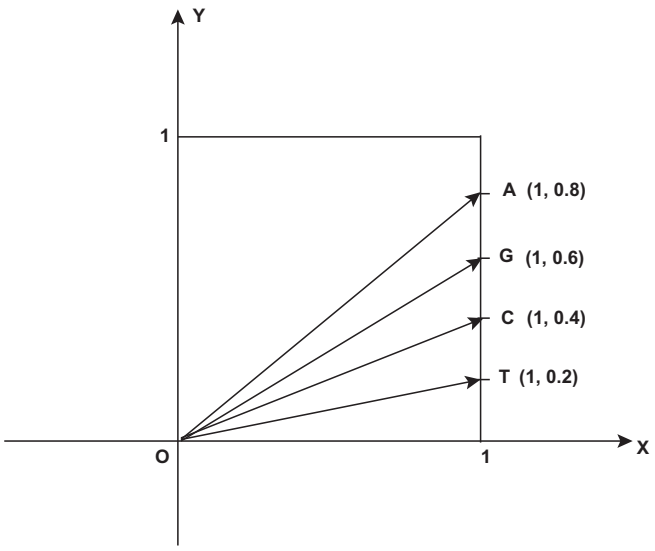

Fig. 1. Nucleotide vector system based on A (1,0.8), G (1,0.6), C (1,0.4), and T (1,0.2).

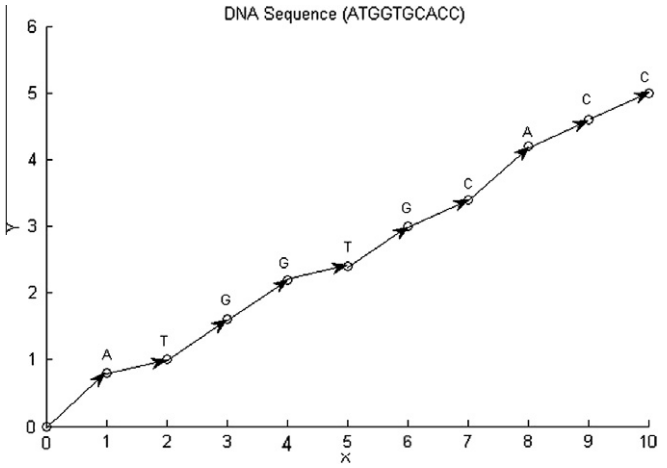

Fig. 2. Graphical representation of DNA sequence (ATGGTGCACC) based on the vector system shown in Fig. 1.

$$p_i = \frac{x_i - \vec{y}_i}{\frac{1}{2}n(n+1) - y_n},$$

where (x_i, y_i) represents the position of the ith nucleotide in the DNA graphical curve, $\vec{y}_i$ represents the choice of y-coordinate value at the ith nucleotide in the DNA graphical curve according to Fig. 1. For example, for DNA sequence (ATGGT),

$$\vec{y}_1 = 0.8, \quad \vec{y}_2 = 0.2, \quad \vec{y}_3 = 0.6, \quad \vec{y}_4 = 0.6, \quad \vec{y}_5 = 0.2; \quad y_5 = 2.4;$$

$$(p_1, p_2, p_3, p_4, p_5) = \left(\frac{1 - 0.8}{\frac{1}{2} \cdot 5 \cdot 6 - 2.4}, \frac{2 - 0.2}{\frac{1}{2} \cdot 5 \cdot 6 - 2.4}, \frac{3 - 0.6}{\frac{1}{2} \cdot 5 \cdot 6 - 2.4}, \frac{4 - 0.6}{\frac{1}{2} \cdot 5 \cdot 6 - 2.4}, \frac{5 - 0.2}{\frac{1}{2} \cdot 5 \cdot 6 - 2.4} \right)$$
$$= (0.0159, 0.1429, 0.1905, 0.2698, 0.3810).$$

Next, we prove that this distribution is a discrete probability distribution:

(1) $\sum_{i=1}^{n} p_i = \sum_{i=1}^{n} \frac{x_i - \vec{y}_i}{\frac{1}{2}n(n+1) - y_n} = \frac{\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \vec{y}_i}{\frac{1}{2}n(n+1) - y_n} = \frac{\frac{1}{2}n(n+1) - y_n}{\frac{1}{2}n(n+1) - y_n} = 1.$

(2) Since $0 < \vec{y}_i < 1$ and $1 \leqslant x_i \leqslant n$, $x_i - \vec{y}_i \leqslant x_i \leqslant n$.

$$y_n = \sum_{i=1}^{n} \vec{y}_i < n, \quad \text{so } \frac{1}{2}n(n+1) - y_n > \frac{1}{2}n(n+1) - n.$$

Thus $p_i = \frac{x_i - \vec{y}_i}{\frac{1}{2}n(n+1) - y_n} < \frac{n}{\frac{1}{2}n(n+1) - n} = \frac{1}{\frac{n+1}{2} - 1} = \frac{2}{n-1}$. So, when $n \geqslant 3$, $p_i < 1$.

$x_i - \vec{y}_i > 0$, and $\frac{1}{2}n(n+1) - y_n > \frac{1}{2}n(n+1) - n = \frac{n(n-1)}{2} > 0$ when $n \geqslant 3$. So, $p_i > 0$.

Therefore, when $n \geqslant 3$, $0 < p_i < 1$.

By (1) and (2) we have proved that $(p_1, p_2, \ldots, p_n)$ is a discrete probability distribution.

2.3. Similarity measure by symmetrised Kullback–Leibler divergence

Now that we have a discrete probability distribution for DNA sequences, we want to find a similarity/dissimilarity measure between two discrete probability distributions $\lambda_1 = (p_1, p_2, \ldots, p_n)$ and $\lambda_2 = (q_1, q_2, \ldots, q_n)$. A well-known dissimilarity measure between two probability distributions is the Kullback–Leibler divergence [7].

Let P_1 and P_2 be two discrete probability distributions on a universe X, the Kullback–Leibler divergence (KLD) or the relative entropy, denoted as $H(P_1, P_2)$ of P_1 with respect to P_2 is defined by $H(P_1, P_2) = \sum_{x \in X} p_1(x) \log \frac{p_1(x)}{p_2(x)}$. $H(P_1, P_2) = 0$ if and only if $P_1 = P_2$. $H(P_1, P_2)$ is often called a distance, but it is not a true metric because $H(P_1, P_2) \neq H(P_2, P_1)$. Moreover, it does not satisfy the triangle inequality.

Thus, given two discrete probability distributions $\lambda_1 = (p_1, p_2, \ldots, p_n)$ and $\lambda_2 = (q_1, q_2, \ldots, q_n)$ for two DNA sequences, we now can define the symmetried similarity measure, denoted by $d(\lambda_1, \lambda_2)$, as $d(\lambda_1, \lambda_2) = \frac{H(\lambda_1, \lambda_2) + H(\lambda_2, \lambda_1)}{2}$. Clearly, $d(\lambda_1, \lambda_2) = d(\lambda_2, \lambda_1)$. Therefore, we have obtained a symmetrised similarity measure between two DNA sequences with the same length.

To test that the measure obtained in this way truly incorporates clustering and phylogenetic analysis, we apply it to the complete coding sequence of beta-globin genes from 10 different species, which are human (U01317), woolly monkey (AY279114), tufted monkey (AY279115), rat (X06701), rabbit (V00882), hare (Y00347), gallus (NM_001081704), duck (X15739), opossum (J03642), salmon (NM_001123672). All these DNA sequences have 444 nucleotides. The similarity ma-

Table 1
the similarity matrix of the complete coding sequence of beta-globin genes from 10 different species.

1.0e−005*	Human	Woolly monkey	Capuchin monkey	Rat	Rabbit	Hare	Gallus	Duck	Opossum	Salmon
Human										
Woolly monkey	0.0192									
Capuchin monkey	0.0132	0.0090								
Rat	0.0679	0.0666	0.0774							
Rabbit	0.0560	0.0539	0.0482	0.0956						
Hare	0.0482	0.0501	0.0449	0.0924	0.0090					
Gallus	0.0896	0.0961	0.0911	0.1324	0.0720	0.0684				
Duck	0.0882	0.0936	0.0910	0.1200	0.1064	0.1028	0.0511			
Opossum	0.0684	0.0646	0.0716	0.1005	0.0889	0.0858	0.1101	0.1012		
Salmon	0.1582	0.1525	0.1658	0.1459	0.1726	0.1677	0.1438	0.1145	0.1519	

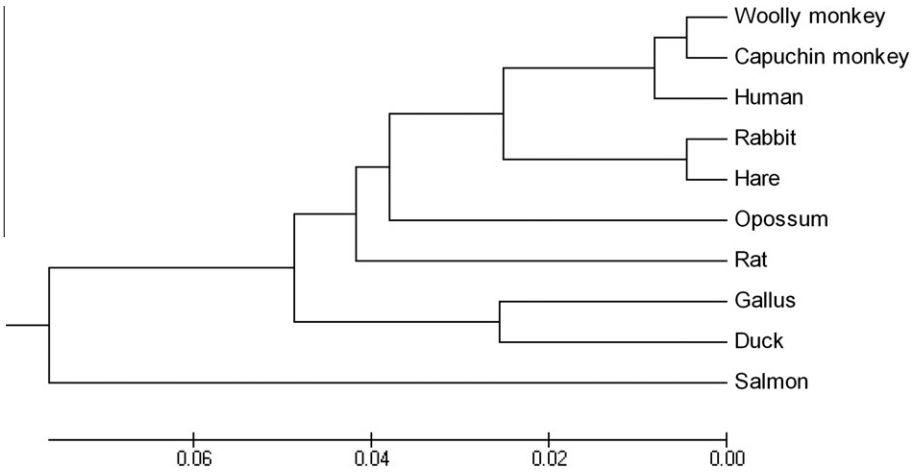

Fig. 3. Phylogenetic tree of 10 different species based on their complete coding sequence of beta-globin genes by using our new approach.

trix of these 10 DNA sequences is shown in Table 1, and the phylogenetic tree for them by using UPGMA algorithm of MEGA 4 package [17] is represented in Fig. 3. Here we should point out that the phylogenetic relationship of these 10 species may not be accurate because we did not use the whole genome information for constructing the tree, but the figure still clearly shows similarity of these 10 DNA sequences.

2.4. Normalized probability distribution of DNA sequence

In Section 2.2, we transformed a DNA sequence into a discrete probability distribution using our graphical representation. However, the probabilistic distribution of a DNA sequence $(p_1, p_2, \ldots, p_n)$ is related to its length n. This limits the comparison of DNA sequences with different lengths. To overcome this limitation, we need to develop a normalized probability distribution for all DNA sequences. For a DNA sequence of length n and a specific $N < n$, consider the $n - N + 1$ subsequences of length N. By using the approach in Section 2.2, we can get the probability distributions $(p_1, p_2, \ldots, p_N)$ for each of subsequences of length N. Then we can average over these probabilistic distributions to obtain a normalized probability distribution for this DNA sequence. For example, DNA sequence (ATGGTGCACC) has length 10, and we take $N = 6$. Then it is separated into 5 subsequences of length 6: ATGGTG, TGGTGC, GGTGCA, GTGCAC, and TGCACC. Thus, by using the approach in Section 2.2, we obtain their probability distributions: (0.0111, 0.1000, 0.1333, 0.1889, 0.2667, 0.3000), (0.0435, 0.0761, 0.1304, 0.2065, 0.2391, 0.3043), (0.0225, 0.0787, 0.1573, 0.1910, 0.2584, 0.2921), (0.0222, 0.1000, 0.1333, 0.2000, 0.2333, 0.3111) and (0.0440, 0.0769, 0.1429, 0.1758, 0.2527, 0.3077). So, the normalized probability distribution of this DNA sequence is the mean value of them (0.0286, 0.0863, 0.1395, 0.1924, 0.2501, 0.3031). It should be pointed out that the choice of N depends on the length of the shortest sequence in the tested group of DNA sequences. For example, suppose we are comparing a group of DNA sequences with different lengths. If the shortest sequence in this group has length N, then we use N to get the normalized probability distributions for all DNA sequences in this group.

3. Results

The method is tested with six DNA sequences [23], taken from the threonine operons of *E. coli* K-12 (gi: 1786181) and *S.* flexneri (gi: 30039813). The three sequences taken from each threonine operon are thrA (aspartokinase I-homoserine

dehydrogenase I), thrB (homoserine kinase) and thrC (threonine synthase), using the open reading frames (ORFs) 337–2799 (ec-thrA), 2801–3733 (ec-thrB) and 3734–5020 (ec-thrC) in the case of *E. coli* K-12, and 336–2798 (sf-thrA), 2800–3732 (sf-thrB) and 3733–5019 (sf-thrC) in the case of *S. flexneri*. All the sequences were obtained from GenBank. In addition, we compared all six sequences with a randomly generated sequence (randomA), using the same length and base composition as ec-thrA.

The length of ec-thrA and sf-thrA is 2463 nt, the length of ec-thrB and sf-thrB is 933 nt, and the length of ec-thrC and sf-thrC is 1287 nt. As mentioned above, here we take $N = 933$, then these 7 DNA sequences are transformed into 7 normalized probability distributions. By using the symmetrised Kullback–Leibler divergence, the similarity matrix of these 7 DNA sequences is displayed in Table 2. This table shows that (ec-thrA, sf-thrA), (ec-thrB, sf-thrB), and (ec-thrC, sf-thrC) have very high similarity. Although the simulated randomA sequence has high similarity with ec-thrA (0.000065723537205), the true gene sf-thrA has higher similarity with ec-thrA (0.000000482115678). This is true even though randomA has the same length and base composition as ec-thrA. This implies that our method can discriminate true matches from random DNA sequences.

In order to further illustrate the efficiency of our approach we then focus on an interesting question about human origins. The 19th century discovery of fossilized Neanderthal skeletons in Europe raised many problems about the origin of human beings, among them the issue of our relation to this species. Now we can answer many questions about human and primate origins by studying their mitochondrial genomes. Mitochondrial DNA is not highly conserved and has a rapid mutation rate, thus it is very useful for studying the evolutionary relationships of organisms [4].

Table 2
the similarity matrix of 6 DNA sequences taken from the threonine operons of *Escherichia coli* K-12 and *Shigella flexneri* and one random DNA sequence with the same length and base composition as ec-thrA.

1.0e−005*	ec_thrA	ec_thrB	ec_thrC	sf_thrA	sf_thrB	sf_thrC	randomA
ec_thrA							
ec_thrB	0.103515801528209						
ec_thrC	0.000036502094026	0.102437899212019					
sf_thrA	0.000000482115678	0.103581825228513	0.000037714891986				
sf_thrB	0.103192589271809	0.000799729968234	0.102114309213345	0.103258812312576			
sf_thrC	0.000042030295348	0.102455599565548	0.000001246963514	0.000043042184261	0.102133014685575		
randomA	0.000065723537205	0.102297937628119	0.000074700320295	0.000076709835796	0.101973258727072	0.000083620997234	

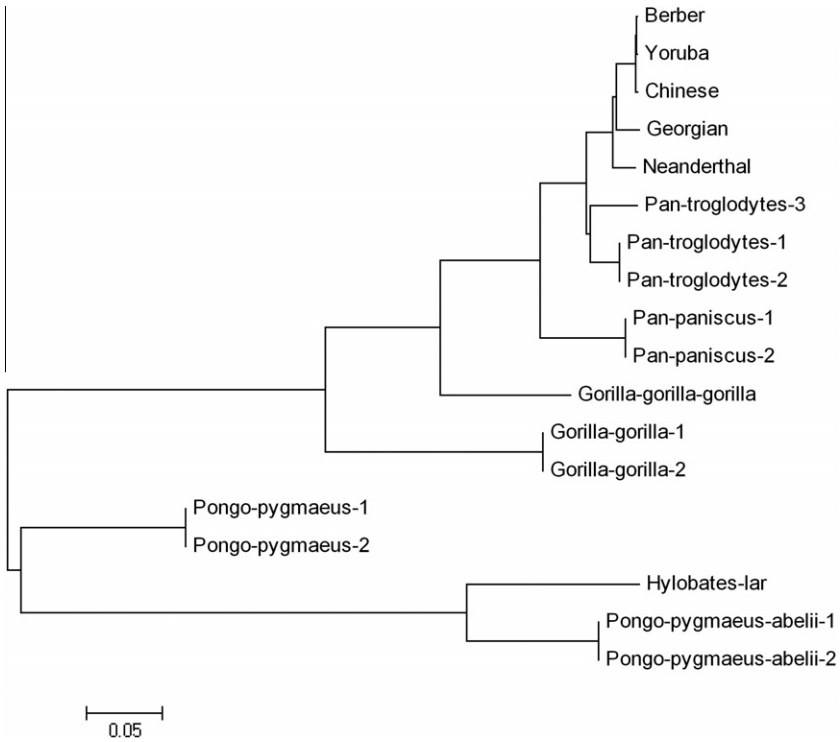

Fig. 4. Phylogenetic tree of 18 primate species based on DNA sequences of hyper variable region II.

In this study we have particular interest for a specific region of mtDNA. This region is the only real stretch of non-coding sequence in the mitochondrial genome and is known as the D-loop. The D-loop does not have any genes; however, it does contain necessary features including the origin of replication and the mitochondrial promoter. The origin of replication is where the replication of the circular genome begins; the promoter is where transcription of all of the mtDNA genes begins The D-loop contains two hyper variable regions I and II (HVR-I and HVR-II). Because HVR is very quickly evolving we can study variation in the mitochondrial genome by just studying the hyper variable regions [8].

Our method is tested with 18 DNA sequences. They are the Hyper Variable Region II of human, Neanderthal, chimpanzee, bonobo, gorilla, orangutan, and gibbon. This data can be obtained from the online material of Cristianini and Hahn's book [8]. Here we take $N = 337$, then these 18 DNA sequences are transformed into 18 normalized probability distributions. By using the symmetrised Kullback–Leibler divergence, the similarity matrix of these DNA sequences can be obtained. The phylogenetic tree among them by using neighbor-joining algorithm [26] of MEGA 4 package [17] is represented in Fig. 4. Our result coincides with those found by Cristianini and Hahn. In fact, our tree also shows that Neanderthal are more closely related to

Table 3
TYLCD-causing virus sequences used in this study.

Isolate	Accession No.	Length
TYLCV_IL	X15656	2787
TYLCV_DO	AF024715	2781
TYLCV_CU	AJ223505	2781
TYLCV_Flo	AY530931	2781
TYLCV_Omu	AB116630	2774
TYLCV_Alm	AJ489258	2781
TYLCV_Mis	AB116631	2774
TYLCV_EG_Ism	AY594174	2781
TYLCV_Miy	AB116629	2774
TYLCV_PR	AY134494	2781
TYLCV_MA	EF060196	2781
TYLCV_TR_Mer1_04	AJ812277	2781
TYLCV_Tosa_H	AB192966	2781
TYLCV_Tosa	AB192965	2781
TYLCV_RE4	AM409201	2781
TYLCV_Sic	DQ144621	2781
TYLCV_TN	EF101929	2781
TYLCV_JO	EF054893	2781
TYLCV_MX_Cul	DQ631892	2781
TYLCV_Mld_PT	AF105975	2793
TYLCV_Mld_Aic	AB014347	2787
TYLCV_Mld_Shi	AB014346	2791
TYLCV_Mld_ES7297	AF071228	2791
TYLCV_Mld_ES	AJ519441	2790
TYLCV_Mld_Sz_Yai	AB116632	2791
TYLCV_Mld_Atu	AB116633	2787
TYLCV_Mld_Kis	AB116634	2787
TYLCV_Mld_Sz_Dai	AB116635	2787
TYLCV_Mld_Sz_Osu	AB116636	2787
TYLCV_Mld_RE	AJ865337	2791
TYLCV_Mld_JO	EF054894	2791
TYLCAxV_Alg	AY227892	2772
TYLCMalV	AF271234	2782
TYLCMLV	AY502934	2794
TYLCMLV_ET	DQ358913	2785
TYLCSV	X61153	2773
TYLCSV_Sic	Z28390	2773
TYLCSV_ES1	Z25751	2777
TYLCSV_ES2	L27708	2777
TYLCSV_MA	AY702650	2777
TYLCSV_TN	AY736854	2772
TYLCCNV	AF311734	2734
TYLCCNV_Tb_Y25	AJ457985	2738
TYLCCNV_YM	DQ256460	2731
TYLCKaV_TH_Kan1	AF511529	2752
TYLCKaV_TH_Kan2	AF511530	2752
TYLCKaV_VN	DQ169054	2751
TYLCTHV	X63015	2743
TYLCTHV_MM	AF206674	2746
TYLCTHV_Y72	AJ495812	2748
TYLCTHV_ChMai	AY514630	2747
TYLCTHV_NoK	AY514631	2744
TYLCTHV_SaNa	AY514632	2747

modern humans than any of the other extant Great Apes, including our closest living relatives, the chimpanzees and bonobos.

In addition, we apply our new method to study the classification and phylogeny of Tomato Yellow Leaf Curl Virus (TYLCV) [9]. This virus possesses a linear single-stranded DNA genome. 53 complete genome sequences of viruses casing TYLCD were

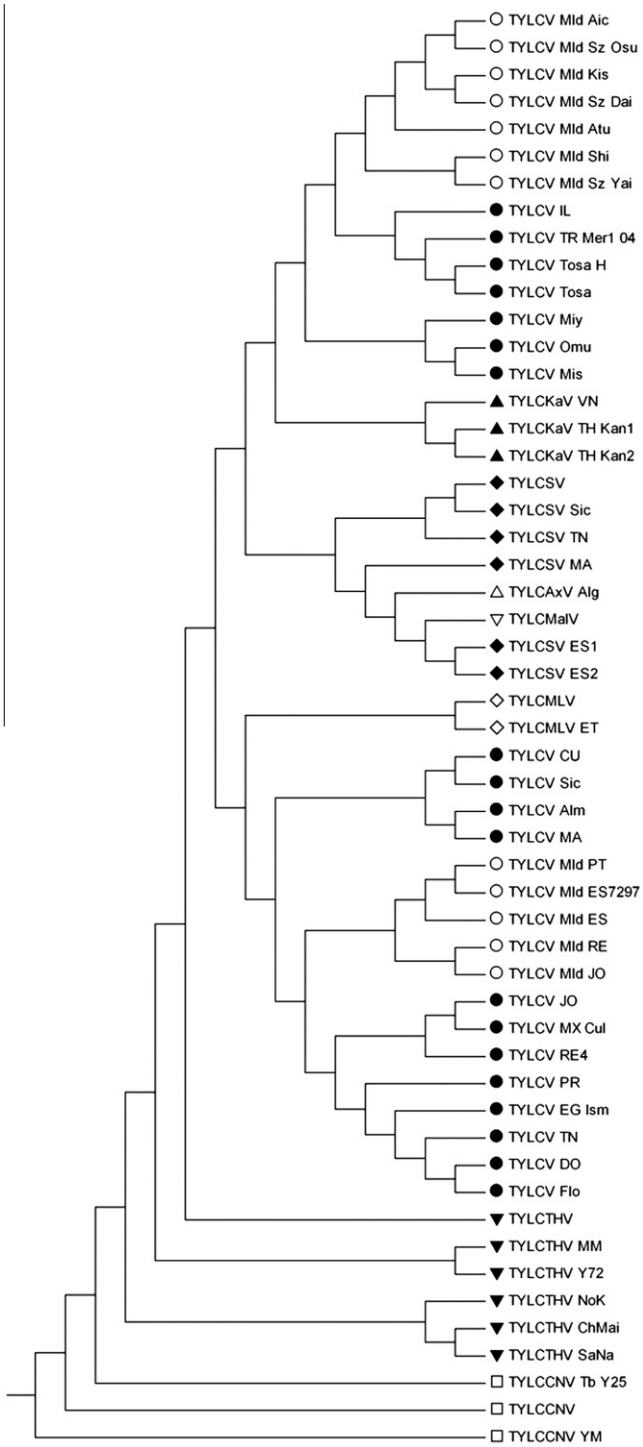

Fig. 5. Phylogenetic tree of 53 TYLCD-causing viral genomes. TYLCV Severe phenotype (•), Mild phenotype (○), and the viruses from Axarquia (△), Malaga (▽), Mali (◇), Sardinia (♦), China (□), Kanchanaburi (▲), Thailand (▼) are shown in this tree. The details about these 53 TYLCV genomes can be found in Table 3.

downloaded from GenBank, each having lengths of more than 2700 nucleotides (Table 3). The shortest sequence is from *Tomato Yellow Leaf Curl China Virus* (TYLCCNV-YM, DQ256460). It has 2731 nucleotides. So, here we take $N = 2731$, then these 53 DNA sequences are transformed into 53 normalized probability distributions. We used the symmetrised Kullback–Leibler divergence to get the similarity matrix of these DNA sequences. Fig. 5 shows the phylogenetic tree relating them. This tree was obtained by using the UPGMA algorithm of MEGA 4 [17], and it agrees with Duffy and Holmes's result [9]. The tree can clearly identify the viruses from *Axarquia, Malaga, Mali, Sardinia, China, Kanchanaburi, Thailand*. For TYLCV Mild phenotype, we find not only Mld ES and Mld JO are far away from other Mild phenotype viruses as shown by Duffy and Holmes, but also Mld ES7297, Mld RE and Mld PT. We also suggest TYLCV IL, TYLCV TR Mer1 04, TYLCV Tosa H, TYLCV Tosa, TYLCV Miy, TYLCV Omu and TYLCV Mis should form a new subcluster of TYLCV Severe phenotype. Furthermore, in order to show the computational efficiency of our approach we use the existing multiple alignment tool ClustalW to do the same work with those 53 DNA sequences. It took us about 20 min to get the result on our Intel (R) Core (TM)2 Duo CPU E8400@3.00 GHz, 2.99 GHz Windows PC with 1.93 GB RAM. However, our new approach needs only about 2 min by a Matlab program on the same computer. The codes used to prepare this paper are available from the author upon request.

4. Discussion

It should be pointed out that the construction of our new approach depends on four parameters (the y-coordinates of the A, C, T, and G in Fig. 1). If we change these four parameters, we shall get a different probability distribution for the DNA sequence. Because the nucleotide content (especially GC content) of DNA molecule is found to vary with different organisms, nucleotide content should be considered when we assign the y-coordinate values of nucleotide vectors. Because most DNA sequences analyzed in this paper have low AG-content (40%–50%), we have assigned larger y-coordinate values to A and G. However, the y-coordinate values of the four nucleotides must be between 0 and 1 to assure that we can get the probability distribution of the DNA sequence. Therefore, in order to obtain a universal probability distribution for DNA sequences without considering nucleotide content, further studies will be needed to determine universal y-coordinate values.

Our aim in this paper is not to conform or refute the previous studies for DNA sequence comparison but rather to bring a novel direction to comparative genomic analysis at the sequence level. Most existing methods for phylogenetic inference require multiple alignment of sequences and assume some sort of an evolutionary model [3]. The choice of evolutionary model totally depends on the researchers. Consequently, the results obtained from different models must be different. In other word, these results require human intervention and are usually controversial. Our approach does not use any evolutionary model. It does not need this type of human intervention. The results are naturally and automatically generated. Our new approach also can handle large volumes of DNA sequences more quickly and more easily than multiple alignment methods . For the normalized probability distribution of DNA sequence the choice of N is very important. In this paper we let N equal to the length of the shortest sequence in the tested group of DNA sequences. Thus, N may be very large when dealing with very long DNA sequences such as whole genome sequences. Further studies will be needed to reduce the size of N under the condition with not losing biological information. In addition, our method may also extend to the protein sequence study in future.

5. Conclusion

In this paper, we have proposed a novel probabilistic method for DNA sequence comparison that uses a graphical representation. After constructing the graphical representation, we were able to construct a probability distribution for a DNA sequence. After obtaining the probabilistic distributions of DNA sequences we use the symmetrised Kullback–Leibler divergence to perform the similarity studies. The results show that our approach provides a new, powerful tool to analyze the similarity and dissimilarity among various DNA sequences for both molecular biologists and computational scientists.

Acknowledgements

We gratefully acknowledge the anonymous reviewers who read our paper and gave some constructive comments. We also thank Dr. Max Benson for critically reading and editing the manuscript.

References

[1] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, Journal of Molecular Biology 215 (1990) 403–410.
[2] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSIBLAST: a new generation of protein database search programs, Nucleic Acids Research 25 (1997) 3389–3402.
[3] S. Ando, E. Sakamoto, H. Iba, Evolutionary modeling and inference of gene network, Information Sciences 145 (2002) 237–259.
[4] W. Brown, E. Prager, A. Wang, A. Wilson, Mitochondrial DNA sequences of primates: tempo and mode of evolution, Journal of Molecular Evolution 18 (1982) 225–239.
[5] R.J.G.B. Campello, E.R. Hruschka, On comparing two sequences of numbers and its applications to clustering analysis, Information Sciences 179 (2009) 1025–1039.
[6] k. Carr, E. Murray, E. Armah, R.L. He, S.S.-T. Yau, A rapid method for characterization of protein relatedness using feature vectors, PLoS One 5 (2010) e9550. doi:10.1371/journal.pone.000955.
[7] T.M. Cover, J.A. Thomas, Elements of Information Theory, John Wiley and Sons, NY, 1991.

[8] N. Cristianini, M.W. Hahn, Introduction to Computational Genomics: A Case Studies Approach, Cambridge University Press, 2007. pp. 78–94.
[9] S. Duffy, E.C. Holmes, Phylogenetic Evidence for Rapid Rates of Molecular Evolution in the Single-Stranded DNA Begomovirus Tomato Yellow Leaf Curl Virus, Journal of Virology 82 (2008) 957–965.
[10] M. Elloumi, Comparison of strings belonging to the same family, Information Sciences 111 (1998) 49–63.
[11] M.A. Gates, Simpler DNA sequence representations, Nature 316 (1985) 219.
[12] O. Gotoh, An improved algorithm for matching biological sequences, Journal of Molecular Biology 162 (1982) 705–708.
[13] E. Hamori, Novel DNA sequence representations, Nature 314 (1985) 585–586.
[14] G. Huang, B. Liao, Y. Li, Y. Yu, Similarity studies of DNA sequences based on a new 2D graphical representation, Biophysical Chemistry 143 (2009) 55–59.
[15] X. Jiang, D. Lavenier, S.S.-T. Yau, Coding region prediction based on a universal DNA sequence representation method, Journal of Computational Biology 15 (2008) 1237–1256.
[16] B.H. Juang, L.R. Rabiner, A probabilistic distance measure for hidden Markov models, AT& T Technical Journal 64 (1985) 391–408.
[17] S. Kumar, M. Nei, J. Dudley, K. Tamura, MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences, Briefings in Bioinformatics 9 (2008) 299–306.
[18] B. Liao, T. Wang, New 2D graphical representation of DNA sequences, Journal of Computational Chemistry 25 (2004) 1364–1368.
[19] L. Liu, Y. Ho, S.S.-T. Yau, Clustering DNA sequences by feature vectors, Molecular Phylogenetics and Evolution 41 (2006) 64–69.
[20] S.B. Needleman, C.D. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins, Journal of Molecular Biology 48 (1970) 443–453.
[21] W.R. Pearson, D.J. Lipman, Improved tools for biological sequence comparison, Proceedings of the National Academy of Sciences of the United States of America 85 (1988) 2444–2448.
[22] W.R. Pearson, Rapid and sensitive sequence comparison with FASTP and FASTA, Methods Enzymology 183 (1990) 63–98.
[23] T.D. Pham, J. Zuegg, A probabilistic measure for alignment-free sequence comparison, Bioinformatics 20 (2004) 3455–3461.
[24] M. Randic, M. Vracko, N. Lers, D. Plavsic, Novel 2-D graphical representation of DNA sequences and their numerical characterization, Chemical Physics Letters 368 (2003) 1–6.
[25] G. Reinert, D. Chew, F. Sun, M.S. Waterman, Alignment-free sequence comparison (I): statistics and power, Journal of Computational Biology 16 (2009) 1615–1634.
[26] N. Saitou, M. Nei, The neighbor-joining method: a new method for reconstructing phylogenetic trees, Molecular Biology and Evolution 4 (1987) 406–425.
[27] T.F. Smith, M.S. Waterman, Identification of common molecular subsequences, Journal of Molecular Biology 147 (1981) 195–197.
[28] S. Vinga, J. Almeida, Alignment-free sequence comparison – a review, Bioinformatics 19 (2003) 513–523.
[29] S.S.-T. Yau, J. Wang, A. Niknejad, C. Lu, N. Jin, Y. Ho, DNA sequence representation without degeneracy, Nucleic Acids Research 31 (2003) 3078–3080.
[30] S.S.-T. Yau, C. Yu, R. He, A protein map and its application, DNA and Cell Biology 27 (2008) 241–250.
[31] C. Yu, Q. Liang, C. Yin, R.L. He, S.S.-T. Yau, A novel construction of genome space with biological geometry, DNA Research 17 (2010) 155–168.