Evolutionary Bioinformatics

ORIGINAL RESEARCH

A Novel Model for DNA Sequence Similarity Analysis Based on Graph Theory

Xingqin Qi[2], Qin Wu[1], Yusen Zhang[2], Eddie Fuller[1] and Cun-Quan Zhang[1]

[1]Department of Mathematics, West Virginia University, Morgantown, WV, USA, 26506. [2]School of Mathematics and Statistics, Shandong University at Weihai, Weihai, China, 264209.
Corresponding authors email: ef@math.wvu.edu; cqzhang@math.wvu.edu

Abstract: Determination of sequence similarity is one of the major steps in computational phylogenetic studies. As we know, during evolutionary history, not only DNA mutations for individual nucleotide but also subsequent rearrangements occurred. It has been one of major tasks of computational biologists to develop novel mathematical descriptors for similarity analysis such that various mutation phenomena information would be involved simultaneously. In this paper, different from traditional methods (eg, nucleotide frequency, geometric representations) as bases for construction of mathematical descriptors, we construct novel mathematical descriptors based on graph theory. In particular, for each DNA sequence, we will set up a *weighted directed graph*. The adjacency matrix of the directed graph will be used to induce a representative vector for DNA sequence. This new approach measures similarity based on both ordering and frequency of nucleotides so that much more information is involved. As an application, the method is tested on a set of 0.9-kb mtDNA sequences of twelve different primate species. All output phylogenetic trees with various distance estimations have the same topology, and are generally consistent with the reported results from early studies, which proves the new method's efficiency; we also test the new method on a simulated data set, which shows our new method performs better than traditional global alignment method when subsequent rearrangements happen frequently during evolutionary history.

Keywords: DNA sequence, mathematical descriptor, similarity analysis, weighted graph

Evolutionary Bioinformatics 2011:7 149–158

doi: 10.4137/EBO.S7364

This article is available from http://www.la-press.com.

Introduction

The number of DNA sequences is rapidly increasing in the DNA database. It is one of the challenges for bio-scientists to analyze the large volume of genomic DNA sequence data. Many schemes have been proposed to numerically characterize DNA sequences and analyze their similarities.

Sequence alignment has been frequently used as a powerful tool to accomplish the comparison of two closely related genomes at the base-by-base nucleotide sequence level. This method is mainly based on the orderings of nucleotides appearing in the sequence. But with the divergence of species over time, subsequence rearrangements occurring during evolution make sequence alignment similarity scores less reliable. Improvements[1,2] have been proposed to overcome the difficulty, but most of these improvements mainly rely on the correct definition and selection of common genes to be compared, and significant homology among aligned gene sequences.

Blaisdell B.E.[3] introduced a measure of similarity of sets of sequences not requiring sequence alignment. It is the first usage of features (l-mers) counts for biological sequence comparison. Geometric representations of DNA sequences has been regarded as another powerful alignment-free tool for the analysis of DNA sequences recently since Hamori and Ruskin[4] first proposed a 3D geometric representation for DNA sequences. This methodology always starts with a graphical representation of DNA sequence, which could be based on 2D,[5–13] 3D,[14–20] 4D,[21] 5D,[22] and 6D[23] spaces, and represents DNA as matrices by associating with the selected geometrical objects, then vectors composed of the invariants of matrices are used to compare DNA sequences.

Sequence alignment method is mainly based on the orderings of nucleotides appearing in the sequence. But with the diverge of species over time, subsequence rearrangements (eg, reversal, transposition or block-exchange) occurring during evolution would make sequence alignment similarity scores less reliable. Features count methods only focus on the appearing frequencies (l-mers), which would lose significant amounts of information. Geometric representation schemes have an advantage in that they order an instant, though visual and qualitative summary of the lengthy DNA sequences. But this approach also involves many unresolved questions. For example, how to obtain suitable matrices to characterize DNA sequences and how to select invariants suitable for sequence comparisons. Another difficulty we must face is that the calculation of the matrices or the invariants will become more and more difficult with the length of the sequences.

It has been one of major challenges for computational biologists that low time-complexity alignment-free methods are needed for proper measurements of sequence similarity, which should not only take into account the happenings of single nucleotide mutations but also the happenings of subsequence rearrangements.

In this paper, we introduce a novel method, which is based on graph theory, to represent DNA sequences mathematically for similarity analysis. In particular, for each DNA sequence, we will set up a *weighted directed graph,* whose adjacency matrix will give us a representative vector. Three distance measurements for representative vectors are then defined to assess the similarity/dissimilarity analysis for DNA sequences. As an application, the method is tested on a set of 0.9-kb mtDNA sequences of twelve different primate species, and the output phylogenetic trees based on these three distance measurements have the same topology, and are all generally consistent with the results reported in previous studies.[24–26] Furthermore, to show its robustness and tolerance to the happening of rearrangements of DNA subsequence, we test it on one synthetic data set by showing the fact that based on our method offspring after various generations could still find its original ancestor with high probability. This method is significantly different from all traditional methodologies and is a promising approach in future studies.

The paper is organized as follows. In Section 2, we describe the method of constructing the weighted directed graph and the representative vector for a given DNA sequence; in Section 3, three distance measurements are introduced to assess the similarity/dissimilarity of DNA sequences; the experimental results for 0.9-kb mtDNA sequences of twelve different primate species are presented in Section 4; and the simulated test is discussed in Section 5; conclusions are made in Section 6.

Construction of Representative Vector for DNA Sequence

The alphabet representation of a DNA sequence is a string of letters *A*, *C*, *G* and *T*. Assume $S = s_1\, s_2 \ldots s_n$ is a DNA sequence of length *n*, where $s_i \in \{A, C, G, T\}$.

Directed multi-graph

We will show how to construct the weighted directed multi-graph for $S = s_1\, s_2\, \ldots\, s_n$, which is denoted by $G_m = (V(G_m), A(G_m))$. The vertex set $V(G_m) = \{A, C, G, T\}$. For each pair of nucleotides s_i and s_j in *S* with $i < j$, put a arc from s_i to s_j, and define the weight of the arc as $\left(1/(j-i)^{\alpha}\right)$[a], where $\alpha > 0$ so that $1/(j-i)^{\alpha}$ is an decreasing function of $(j-i)$ which would reflect the fact that the two nucleotides with smaller distance would have stronger interactive relationship than those with bigger distance.

An example with parameter $\alpha = 1/2$ is illustrated in Figure 1.[b]

Theorem 1. It is an one-to-one mapping between a DNA sequence *S* and its corresponding weighted directed multi-graph G_m.

Proof. It is sufficient to show that we can get only one DNA sequence from the graph G_m. Let n_W be the number of nucleotide base $W\ (\in \{A, C, G, T\})$ appearing in the DNA sequence and x_W be the number of loops incident with the vertex W in G_m, respectively. Clearly $x\_w = (n_W - 1)*n_W/2$ for every $W \in \{A, C, G, T\}$, thus we can get each n_W by x_w. The length of DNA sequence can be obtained by $n = n_A + n_G + n_C + n_T$. Note that there is only one arc (W', W'') with weight $1/(n-1)^{\alpha}$ in G_m. Thus, the first nucleotide base in the sequence *S* is W'. The *j*th nucleotide base W^* in *S* is determined by the arc (W', W^*) with weight $1/(j-1)^{\alpha}$.

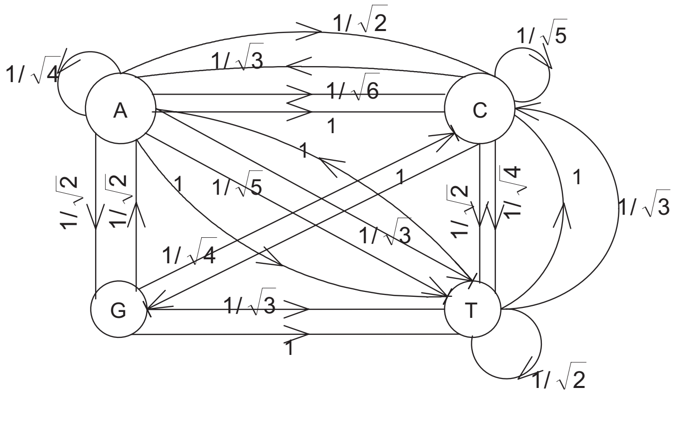

Figure 1. Directed multi-graph G_m for $S = ACGTATC$ with $\alpha = 1/2$.

The simplified weighted directed graph

G_m is a directed multi-graph. That is, there may be parallel arcs from one vertex to anther. In the following, we will simplify G_m to G_s by merging parallel arcs into one arc.

Let the vertex set $V(G_s) = V(G_m)$. Denote $A^m_{u,v}$ as the set of all arcs from the vertex *u* to *v* in G_m; for any pair of vertices *u* and *v*, if $A^m_{u,v} \neq \emptyset$, put an arc (u, v) from *u* to *v* in G_s, and assign the weight of the arc (u, v) in G_s as

$$w_s(u,v) = \sum_{(u,v)\in A^m_{u,v}} w_m(u,v),\ A^m_{u,v} \neq \emptyset$$

Based on this simplification rule, the directed multi-graph in Figure 1 is simplified and illustrated in Figure 2.

Note that the one-to-one mapping between a DNA sequence *S* and the simplified graph does not exist then, which is also the source of error of our strategy,

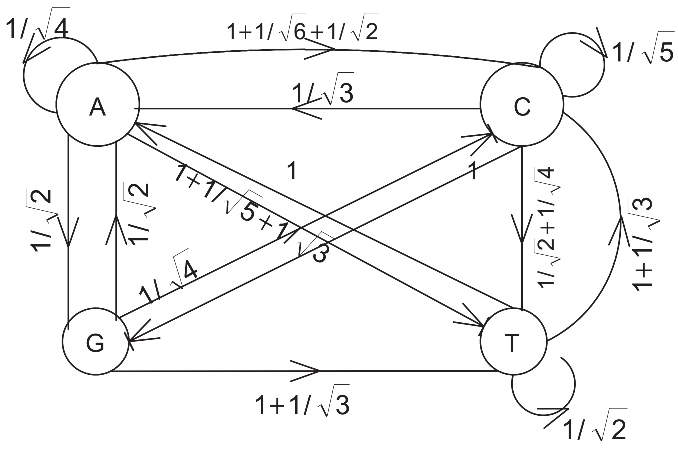

Figure 2. Simplified graph G_s for $S = ACGTATC$.

[a] α is a user specified parameter.

[b] As an approximation and for practical purpose, we only keep 4 decimals for $w_m(s_i,s_j)$.

but we will see later that the simplified graph still contains enough accurate information to characterize DNA sequences.

The representative vector

From above subsections, we get one weighted directed graph associated with a DNA sequence. The weighted directed graph G_s corresponds to a (4 × 4) **adjacency matrix** M, which is defined as follows:

$$M = \begin{pmatrix} w_s(A,A) & w_s(A,C) & w_s(A,G) & w_s(A,T) \\ w_s(C,A) & w_s(C,C) & w_s(C,G) & w_s(C,T) \\ w_s(G,A) & w_s(G,C) & w_s(G,G) & w_s(G,T) \\ w_s(T,A) & w_s(T,C) & w_s(T,G) & w_s(T,T) \end{pmatrix}$$

Then we rewrite matrix M as one 16-dimensional vector $\vec{R}$ by the row order,

$$\vec{R}^T = [w_s(A,A),\ldots,w_s(A,T),w_s(C,A),\ldots,\\ w_s(C,T),\ldots,w_s(T,A),\ldots,w_s(T,T)]$$

We call the 16-dimensional vector the **representative vector** of a DNA sequence. We admit that there is a loss of information when one condenses sequence S to a 16 dimensional vector, but we will see later that it is still enough to make comparisons for DNA sequences.

For the given example of $S = ACGTATC$, when the weight function is $f(l) = 1/\sqrt{l}$, the (4 × 4)-matrix and the 16-dimensional vector are as follows.

$$M = \begin{pmatrix} 0.5000 & 2.1154 & 0.7071 & 2.0246 \\ 0.5774 & 0.4472 & 1.0000 & 1.2071 \\ 0.7071 & 0.5000 & 0 & 1.5774 \\ 1.0000 & 1.5774 & 0 & 0.7071 \end{pmatrix}$$

$$\vec{R}_S = [0.5000, 2.1154, 0.7071, 2.0246, 0.5774\\ 0.4472, 1.0000, 1.2071, 0.7071, 0.5000,\\ 0, 1.5774, 1.0000, 1.5774, 0, 0.7071].$$

We call the method of constructing the representative vector for a DNA sequence *directed euler tour* (DET) method.

Three Distance Measurements for Similarity Calculation

In the above section, we obtain a mapping from a set of DNA sequences to a set of vectors in the 16-dimensional linear space by DET method. Comparison between DNA sequences becomes comparison between these 16-dimensional vectors. We will introduce three popular measurements of defining the distance between two 16-dimensional vectors to reflect the dissimilarity of the two corresponding DNA sequences. The smaller the distance is, the more similar the two sequences are. For two DNA sequences s and h, we denote the representative vectors by $\vec{R}_s$ and $\vec{R}_h$ respectively.

The first distance measurement $d_1(s,h)$ is defined to be the Euclidean distance between the end points of $\vec{R}_s$ and $\vec{R}_h$, which is based on the assumption that two DNA sequences are similar if the corresponding 16-vectors have similar magnitudes, ie,

$$d_1(s,h) = \sqrt{\sum_{i=1}^{16} (\vec{R}_s(i) - \vec{R}_h(i))^2}.$$

The second distance measurement $d_2(s, h)$ between s and h is defined to be one minus the cosine of the included angle between $\vec{R}_s$ and $\vec{R}_h$, which is based on the assumption that two DNA sequences are similar if the corresponding 16-dimensional vectors in the 16-dimensional space have similar directions, ie,

$$d_2(s,h) = 1 - cos(\vec{R}_s, \vec{R}_h)\\ = 1 - \frac{\sum_{i=1}^{16} \vec{R}_s(i) \cdot \vec{R}_h(i)}{\sqrt{\sum_{i=1}^{16} (\vec{R}_s(i))^2 \cdot \sum_{i=1}^{16} (\vec{R}_h(i))^2}}$$

The third distance measurement is based on the correlation coefficients. The calculation of the linear correlation coefficient $r(s, h)$ between $\vec{R}_s$ and $\vec{R}_h$ uses the conventional Pearson formalism as detailed in the following:

$$r(s,h)=\frac{K\sum_{i=1}^{K}\left[\vec{R}_s(i)\cdot\vec{R}_h(i)\right]-\sum_{i=1}^{K}\vec{R}_s(i)\cdot\sum_{i=1}^{K}\vec{R}_h(i)}{\sqrt{K\sum_{i=1}^{K}(\vec{R}_s(i))^2-\left[\sum_{i=1}^{K}\vec{R}_s(i)\right]^2}\times\sqrt{K\sum_{i=1}^{K}(\vec{R}_h(i))^2-\left[\sum_{i=1}^{K}\vec{R}_h(i)\right]^2}}$$

where K is the dimension of $\vec{R}_s$ or $\vec{R}_h$ (here $K = 16$). Thus we define the third distance measurement as:

$$d_3(s, h) = 1 - r(s, h).$$

Then a comparison between a pair of DNA sequences to judge their similarity or dissimilarity could be carried out by calculating the distances between the corresponding mathematical descriptors. We will give a test of the utility of DET method and the proposed distance measurements in the following Section 4.

Applications and Experimental Results

Data description

To test the utility of DET method and the proposed distance measurements, we will use the 0.9-kb mtDNA fragments of twelve species of four different groups of primates for a test, which were reported by Hayasaka[24] firstly and subsequently used by Zhang.[25,26] The data source consists of four species of old-world monkeys (*Macaca fascicular, Macaca fuscata, Macaca sylvanus, Macaca mulatta*), one specie of new-world monkeys (*Saimiri scirueus*), two species of prosimians (*Lemur catta, Tarsisus syrichta*), and five hominoid species (*Human, Chimpanzee, Gorilla, Orangutan and Hylobates*), for detailed information please see Table 1.

Previous experiments results for these species based on different methods

In Hayasaka et al[24] calculated the number of nucleotide substitutions for a given pair of species by the six-parameter method. Using the calculated methods, they gave phylogenetic trees for these twelve species with the same topology depending on three different grouping methods. Thus the phylogenetic relationships derived from these mtDNA comparisons appear reliable. In References Zhang et al[25,26] also obtained consistent results with[24] based on their new proposed methods for DNA sequence comparison, where only eleven species except human were involved. For the sake of later comparison, we re-construct these previous

Table 1. 0.9-kb mtDNA fragments of 12 species.

Species	ID/accession	Abbreviation	Length (bp)	Database
Macaca fascicular	M22653	M.fas	896	NCBI
Macaca fuscata	M22651	M.fus	896	NCBI
Macaca mulatta	M22650	M.mul	896	NCBI
Macaca sylvanus	M22654	M.syl	896	NCBI
Saimiri scirueus	M22655	S.sci	893	NCBI
Chimpanzee	V00672	Chi	896	NCBI
Lemur catta	M22657	Lemur	895	NCBI
Gorilla	V00658	Gorilla	896	NCBI
Hylobates	V00659	Hyl.	896	NCBI
Orangutan	V00675	Ora	895	NCBI
Tarsisus syrichta	M22656	T.syr	895	NCBI
Human	L00016	Human	896	NCBI

phylogenetic trees based on the upper triangular part of dissimilarity matrix reported in the references, see Figure 3.

Selection of the parameter α

In this subsection, we show how to choose the value of α for this data set. Denote the weight function $f(l)=1/l^{\alpha}$, where l is an integer. Because the maximum value of $f(l)$ for any α is just 1, the arcs with weights not less than 0.1 should be thought as relatively important. We thus define l_0 as the *preference distance* of function $f(l)$ when $f(l_0) \geq 0.1$ while $f(l_0+1) < 0.1$. Pairs of nucleotides within l_0 would be assigned bigger weights (at least 0.1) when we construct the representative vector.

Here we list the preference distance for $f(l)$ with different α. Considering the lengths of these twelve species (890 ~ 900), when α = 2 or α = 1, l_0 is too small; while when $\alpha=1/3$ or $\alpha=1/4$, l_0 is too big. Thus, for this data set, we prefer to use $\alpha=1/2$ with l_0 = 100. The nucleotides with distance 100 would be considered to have stronger interactive relationships. But we admit, for data sets with very long DNA sequences, to make l_0 bigger correspondingly, one could choose higher order roots.

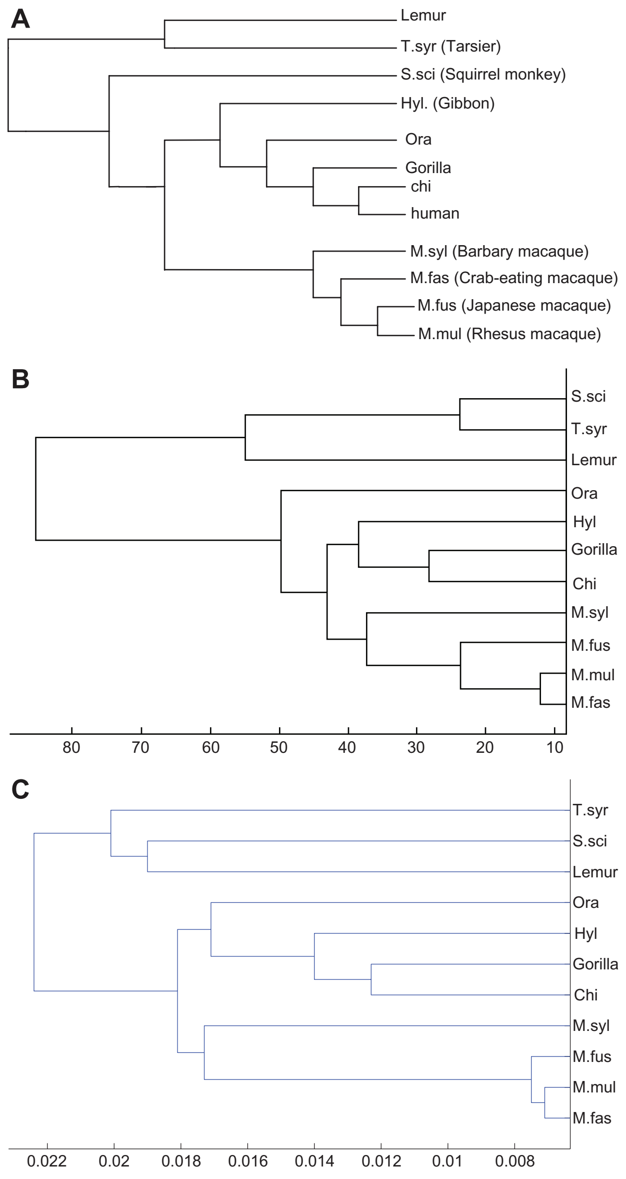

Figure 3. Previous phylogenetic trees for these 12 species based on different methods. (**A**) Figure 3 in[24] (**B**) Figure 1 in[26] (**C**) Figure 1 in.[25]

	α = 2	α = 1	α = 1/2	α = 1/3	α = 1/4
l_0	3	10	100	1000	10000

Similarity matrix based on DET

By the DET method of Section 2, each sequence could be represented by a 16-dimensional vector, and then the similarities between each pair of these twelve mtDNA fragments could be computed under the proposed distance measurements. In Table 2, we present the upper triangular part of the similarity matrix among these twelve species by the DET method with weighted function $f(l)=1/\sqrt{l}$ based on the first distance measurement d_1.

When we examine Table 2, we notice that the smallest entries are associated with the pairs (Gorilla, Human), (M.fas, M. Mul), (M. Fus, M. Mul), (Chi., Gorilla), (M. fas, M. fus), (Human, Chi.) and (M.Syl) and (M.mul). Those observed facts are similar to that reported in previous studies.[25,26] And also consistent with biological classification in[27] and[28] that *Gorilla, Chimpanzee, Human* are in the same family *hominidae* and the same subfamily *homininae*; and *Macaca fascicular*, *Macaca fuscata*, *Macaca mulatta, Macaca sylvanus* are in the same family *Cercopithecidae* and the same genus *Macaca*.

We also present the upper triangular part of the similarity matrices based on the second and the third distance measurement in Tables 3 and 4 respectively. We will see that there is a whole qualitative agreement among similarities based on these three distinct distance measurements. It provides a strong evidence that DET method works well for DNA representation and comparison.

Table 2. The upper triangular part of similarity/dissimilairty matrix based on d_1.

Species	Lemur	Chi	S.sci	M.fas	Gorilla	M.fus	M.mul	M.syl	Hyl	Ora	T.syr	Human
Lemur	0	0.0511	0.0169	0.0358	0.0539	0.0373	0.0327	0.0221	0.0510	0.0702	0.0171	0.0591
Chi		0	0.0528	0.0183	0.0072	0.0171	0.0211	0.0325	0.0179	0.0264	0.0654	0.0098
S.Sci			0	0.0362	0.0545	0.0395	0.0347	0.0286	0.0496	0.0716	0.0201	0.0592
M.fas				0	0.0210	0.0085	0.0059	0.0172	0.0196	0.0391	0.0488	0.0255
Gorilla					0	0.0186	0.0233	0.0354	0.0133	0.0211	0.0679	0.0058
M.fus						0	0.0063	0.0181	0.0174	0.0342	0.0514	0.0238
M.mul							0	0.0131	0.0212	0.0397	0.0463	0.0284
M.syl								0	0.0326	0.0509	0.0355	0.0406
Hyl									0	0.0243	0.0631	0.0169
Ora										0	0.0841	0.0198
T.syr											0	0.0730
Human												0

Construction of dendrogram tree

To see the phylogenetic relationships more easily, we use the average linkage clustering method for the phylogenetic tree construction. In Figure 4, the dendrogram trees based on Tables 2–4 are presented respectively. One can find that the three dendrogram trees of these twelve species have the same topology, which are generally consistent with the previous works in Figure 3.

Simulated Test for DET Method

In a DNA sequence of four letters, there are sixteen possible ordered *XY* pairs: *AA, AC, AT, AG, GC* …, etc. In Ref. Randić[29] introduced a condensed characterization of DNA sequences by (4 × 4)-matrix M^d that give the count of occurrences of all pairs *XY* of bases at distance precisely *d*. For example, when the distance $d = 1$, the matrix will give the frequencies of all such pairs *XY* that *X* and *Y* are adjacent in the DNA sequence; when the distance $d = 2$, the matrix will give the frequencies of all such pairs *XY* that *Y* and *X* are separated by one nucleic base. The advantage of such representations of DNA sequences are that it offers upon inspection useful information that is hidden in the lengthy sequence of the DNA. We should notice that, in Randić’*s* work, information from all matrices M^d were not combined together for analysis, thus not enough information is obtained from such characterization. In our method, for a DNA sequence of length *n*, we have considered all pairs *XY* under possible distance $d = 1, 2 \dots, (n - 1)$ apart, and assign different weights to pair *XY* according to their locations and distributions. Furthermore, all information of ordered pairs of nucleotides are assembled in one matrix (ie, the adjacency matrix *M*). So in some senses, the DET method is an extension of Randić’*s*, but it involves more information hidden in the DNA sequence.

Table 3. The upper triangular part of similarity matrix based on d_2.

Species	S.sci	Chi	Lemur	M.fas	Gorilla	M.fus	M.mul	M.syl	Hyl	Ora	T.syr	Human
S.sci	0	0.0163	0.0016	0.0080	0.0182	0.0087	0.0067	0.0030	0.0163	0.0305	0.0018	0.0219
Chi		0	0.0177	0.0021	0.0003	0.0018	0.0028	0.0066	0.0020	0.0043	0.0269	0.0006
Lemur			0	0.0084	0.0190	0.0099	0.0076	0.0050	0.0160	0.0321	0.0024	0.0225
M.fas				0	0.0028	0.0004	0.0002	0.0018	0.0025	0.0094	0.0150	0.0042
Gorilla					0	0.0022	0.0034	0.0078	0.0011	0.0027	0.0290	0.0002
M.fus						0	0.0002	0.0020	0.0019	0.0072	0.0166	0.0036
M.mul							0	0.0011	0.0028	0.0098	0.0135	0.0051
M.syl								0	0.0066	0.0160	0.0079	0.0103
Hyl									0	0.0034	0.0252	0.0018
Ora										0	0.0438	0.0023
T.syr											0	0.0336
Human												0

Table 4. The upper triangular part of similarity matrix based on d_3.

Species	S.sci	Chi	Lemur	M.fas	Gorilla	M.fus	M.mul	M.syl	Hyl	Ora	T.syr	Human
S.sci	0	0.0749	0.0024	0.0360	0.0840	0.0396	0.0302	0.0136	0.0752	0.1348	0.0082	0.1015
Chi		0	0.0869	0.0098	0.0014	0.0087	0.0134	0.0302	0.0081	0.0183	0.1252	0.0027
Lemur			0	0.0429	0.0959	0.0473	0.0366	0.0196	0.0851	0.1485	0.0078	0.1142
M.fas				0	0.0137	0.0016	0.0007	0.0068	0.0123	0.0409	0.0709	0.0205
Gorilla					0	0.0103	0.0166	0.0358	0.0046	0.0103	0.1367	0.0011
M.fus						0	0.0012	0.0090	0.0076	0.0316	0.0773	0.0171
M.mul							0	0.0044	0.0127	0.0431	0.0629	0.0247
M.syl								0	0.0284	0.0708	0.0357	0.0476
Hyl									0	0.0109	0.1210	0.0083
Ora										0	0.1956	0.0086
T.syr											0	0.1586
Human												0

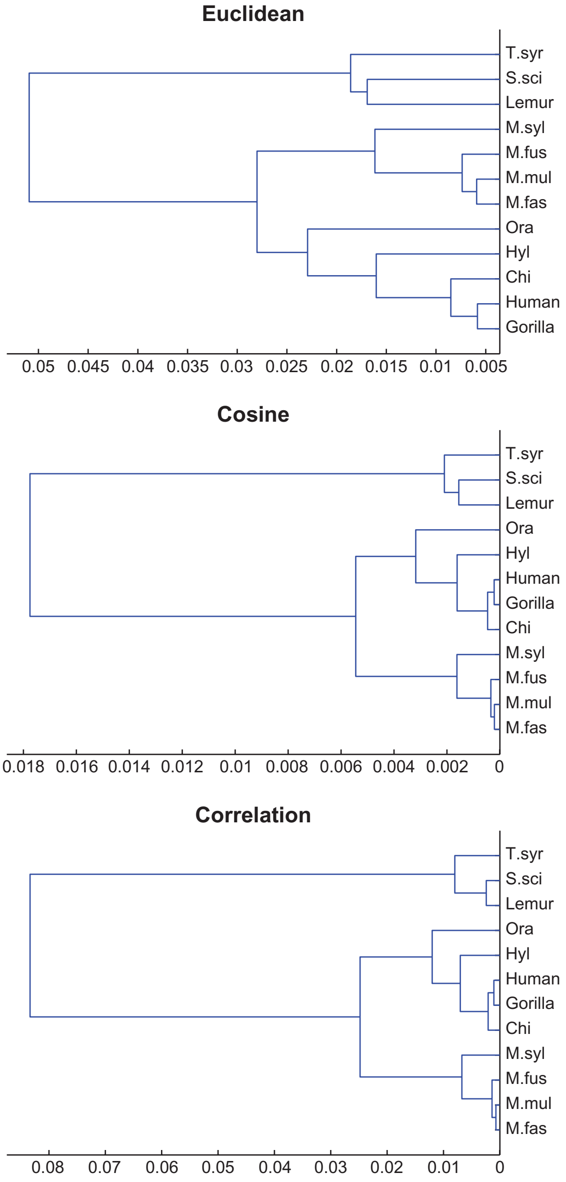

Figure 4. Phylogetic tree for these 12 species based on Tables 2–4.

It is known that the similarity of two sequences determined by alignment is completely based on the ordering of nucleotides, while the similarity by DET is based on both the frequency of pairs of nucleotides and the distance between them. Assume that x_1 is a DNA sequence, x_2 is its descendant after several rearrangements. The alignment method would estimate a relatively low similarity between x_1 and x_2 because of the signifcant change in ordering such that x_1 and x_2 might be regarded as two distinct related species. While the following simulated test is designed to show, by DET method, with high probability, x_2 still could regard x_1 as its ancestor even if x_2 is the offspring of x_1 after many generations.

Constructing the Simulated Data set. This simulated test is designed to test the advantage of the DET method over the alignment method when sequence rearrangements happen frequently. One child genome is copied from the initial parent with *shuffle* model, which involves two operations: (i) *transposition,* exchange the positions of two adjacent random sequence fragments and (ii) *reversal,* reverse the order of a random sequence fragment and reinsertion of it in the same position. To accelerate the speed of achieving the generations that alignment method is no longer useful for offspring to find out its ancestor, we require that the size of involved random sequence fragments for both operations should be at least 0.1*n* (where *n* is the length of genome sequence).

For computational expediency, 0.9 kb mtDNA sequence of Human (L00016) is chosen as a root ancestor test sequence, and 0.9 kb mtDNA sequence of Chimpanzee (V00672) is chosen as its "brother" sequence. We would generate up to 20th generation of Human (L00016) using *shuffle* model. For simplicity, we select its 1st; 5th; 10th; 15th and 20th generation to use.

We then compute the similarities based on both Alignment[c] and DET method[d] between these five generations with Chimpanzee (V00672) and Human (L00016), respectively. Because these five generations are generated randomly, to be fair, we repeat the above process for several times. At each time, if all of these five generations have bigger similarities with Human (L00016) than with Chimpanzee (V00672) based on some method, this method would get one score. Let *suc_rate = score/l*, where *l* is the total test times. Clearly, the method with higher *suc_rate* would be regarded as much more robust when shuffing happens frequently.

In the following table we list the result of *suc_rate* when we run the above simulated test once for different l values. To our surprise, the average *suc_rate* for DET method is 1, while the average *suc_rate* for alignment is just 0.2709, which proves the utility and tolerance of DET method when shuffing happens frequently.

Suc_rate	l = 10	l = 20	l = 30	l = 40	l = 50	l = 60	l = 70	l = 80	l = 90	l = 100
DET method	1	1	1	1	1	1	1	1	1	1
Alignment	0.2000	0.3000	0.4000	0.2000	0.3000	0.2667	0.3857	0.2250	0.2222	0.2500

Concluding Remarks

The graph model presented in this paper is a new method for mathematically characterization of DNA sequences. When it is used to compare DNA sequences, we get consistent results with previous studies and biological classifications. We have no intention to compete with other existing methods, since as we know, for any particular research project we will have to identify which measure is most meaningful or useful. The first advantage of our new method is greater efficiency when compared with alignment method when shuffing happens frequently during the evolutionary process; the second advantage of our method is low time complexity compared with other alignment-free methods. The core of the method is the construction of the representative vector for each DNA sequence and then to get the similarity matrix for a set of DNA sequences. The representative vector for one DNA sequence of length n can be obtained in $O(n^2)$ time units. Then, for a set of m sequences with the maximum length of n, the similarity matrix can be computed in $O(mn^2)$ time units. The similarity matrix is relatively simple for calculation. Our novel method is very different to all traditional methods and is proven to be effective and accurate for similarity comparison of DNA sequences.

Acknowledgement

We thank the Editor and the two anonymous referees for their valuable comments, which help improve this work a lot. This work is supported partly by the Shandong Province Natural Science Foundation of China with No. ZR2010AQ018 and partly by the Independent Innovation Foundation of Shandong University with No. 2010ZRJQ005, and has been partially supported by a WV EPSCoR grant.

Disclosure

This manuscript has been read and approved by all authors. This paper is unique and is not under consideration by any other publication and has not been published elsewhere. The authors and peer reviewers of this paper report no conflicts of interest. The authors confirm that they have permission to reproduce any copyrighted material.

References

1. Ma H, Bork P. Measuring genome evolution. *Proc Natl Acad Sci U S A*. 1988;95:5849–56.
2. Wildman DE, et al. Genomics, biogeography, and the diversification of placental mammals. *Proc Natl Acad Sci U S A*. 2007;104:14395–400.
3. Blaisdell BE. A measure of the similarity of sets of sequences not requiring sequence alignment. *Proc Natl Acad Sci U S A*. 1986;83:5155–9.

[c] Here, we use Needleman-Wunsch globally align algorithm to get the alignment of any two DNA sequences, and define the similarity between them as r/N, where r is the number of matching in the alignment and N is the length of alignment.

[d] The DET similarity between two sequences is defind as $1 - d_2$, ie, the *cosine* value of the included angle between their representative vectors.

4. Hamori E, Ruskin J. H curves: a novel method of representation of nucleotide series especially suited for long DNA sequences. *Journal of Biological Chemistry*. 1983;258:1318–27.
5. Nandy A. A new graphical representation and analysis of DNA sequence structure: I. Methodology and Application to Globin Genes. *Curr Sci*. 1994;66:309–14.
6. Guo X, Randić M, Basak SC. A novel 2-D graphical representation of DNA sequences of lowde generacy. *Chemical Physics Letters*. 2001;350:106–12.
7. Randić M, Vraćko M, Lerś N. Novel 2-D graphical representation of DNA sequences and their numberical characterization. *Journal of Chemical Information and Computer Science*. 2003a;368:1–6.
8. Randić M, Vraćko, M, Lerś N, Plavśić D. Analysis of similarity/dissimilarity of DNA sequence based on novel 2-D graphical representation. *Journal of Chemical Information and Computer Science*. 2003b;371:202–7.
9. Randić M, Vraćko M, Zupan J, Novic M. Compact 2-D graphical representation of DNA. *Chemical Physics Letters*. 2003c;373:558–62.
10. Randić M. Graphical representations of DNA as 2-D map. *Chemical Physics Letters*. 2004;386:468–71.
11. Zhang Y, Liao B, Ding K. On 2D graphical representation of DNA sequence of nondegeneracy. *Journal of Chemical Information and Computer Science*. 2005;411:28–32.
12. Liu X, Dai Q, Xiu Z, Wang T. PNN-curve: a new 2D graphical representation of DNA sequences and its application. *Journal of Theoretical Biology*. 2006;243:555–61.
13. Huang G, Liao B, Li Y, Liu Z. H curves: a novel 2D graphical representation for DNA sequences. *Chemical Physics Letters*. 2008;462:129–32.
14. Randić M, Vracko M, Nandy A. Basak SC. On 3-D graphical representation of DNA primary sequences and their numerical characterization. *Journal of Chemical Information and Computer Science*. 2000;40:1235–44.
15. Liao B, Wang T. 3-D graphical representation of DNA sequences and their numerical characterization. *Journal of Molecular Structure (THEOCHEM)*. 2004;681:209–12.
16. Zhang Y, Liao B, Ding K. On 3D D-curves of DNA sequences. *Mol Simul*. 2006;32:29–34.
17. Qi X, Wen J, Qi Z. New 3D graphical representation of DNA sequence based on dual nucleotides. *Journal of Theoretical Biology*. 2007;249:681–90.
18. Qi Z, Fan T. PN-curve: a 3D graphical representation of DNA sequences and their numerical characterization. *Chemical Physics Letters*. 2007;442:434–40.
19. Cao Z, Liao B, Li R. A group of 3D graphical representation of DNA sequences based on dual nucleotides. *International Journal of Quantum Chemistry*. 2008;108:1485–90.
20. Yu J, Sun X, Wang J. TN curve: A novel 3D graphical representation of DNA sequence based on trinucleotides and its applications. *Journal of Theoretical Biology*. 2009;261:459–68. doi:10.1016/j.jtbi.2009.08.005.
21. Chi R, Ding K. Novel 4D numerical representation of DNA sequences. *Chemical Physics Letters*. 2005;407:63–7.
22. Liao B, Li R, Zhu W. On the similarity of DNA primary sequences based on 5-D representation. *Journal of Mathematical Chemistry*. 2007;42:47–57.
23. Liao B, Wang T. Analysis of similarity/dissimilarity of DNA sequences based on nonoverlapping trinucleotides of nucleotide bases. *Journal of Chemical Information and Computer Science*. 2004;44:1666–70.
24. Hayasaka K, Gojobori T, Horai S. Molecular phylogeny and evolution of primate mitochondrial DNA. *Mol Biol Evol*. 1998;5:626–44.
25. Zhang Y, Chen W. New invariant of DNA sequences. *Match Commun Comput Chem*. 2007;58:197–208.
26. Zhang Y. A simple method to construct the similarity matrices of DNA sequence. *Match Commun Comput Chem*. 2008;60:313–24.
27. Goodman M, et al. Primate evolution at the DNA level and a classification of hominoids. *Journal of Molecular Evolution*. 1990;30:260–6.
28. Groves C, Wilson DE, Reeder DM, editors. Mammal Species of the World (3rd ed.), Johns Hopkins University Press; 2005:161–5.
29. Randić M. Condensed representation of DNA primary sequences. *Journal of Chemical Information and Computer Science*. 2000;40:50–6.