MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **60** (2008) 313-324

ISSN 0340 - 6253

A Simple Method to Construct the Similarity Matrices of DNA Sequences

Yusen Zhang*
School of Mathematics and Statistics, Shandong University at Weihai

Weihai 264209, China

(Received April 11, 2008)

Abstract. In this paper we propose a method to construct three similarity matrices. This approach is illustrated on the primate mitochondrial DNA sequences for 11 different species. Analysis shows an overall qualitative agreement among the three similarity matrices of the primate mitochondrial DNA sequences for 11 different species. We also construct the dendrogram tree for primate mitochondrial DNA sequences. The phylogeny obtained is generally consistent with evolutionary trees constructed in previous studies.

1 Introduction

The primary structure of DNA consists basically of a nitrogenous base of four nucleotides, the two purines, adenine (A) and guanine (G), and the two pyrimidines, cytosine (C) and thymine (T). Thus the DNA sequence can be simply considered as a symbolic sequence on the four symbols A,C,G,T. Comparison of different DNA primary sequences remains one of the important aspects of the analysis of DNA data banks. For a long time the computer science approach was the only methodology. Therefore, direct comparisons of sequences are made using some simplifications in the search for an approximate optimal alignment, which is based on assuming particular scoring functions, that introduce various penalties for the existence of insertions or deletions in the alignment. As has been described by Godzik [1], the outcome of such searches need

*Corresponding author:zhangys@sdu.edu.cn

not be unique. More recently, alternative routes for quantitative measure of the degree of similarity of DNA sequences were considered [2,3]. The novel methodology starts with a graphical representation of DNA, such as proposed by Nandy [4], which are subsequently numerically characterized by associating with the selected geometrical object that represents DNA a matrix [5-7]. For example, one can consider distance matrix in which matrix elements are given as the distances between the vertices which form the geometrical representation of the sequence. Alternatively, one can consider the quotients of distances measured through space and measure along the shortest path between pairs of vertices [2,3]. Finally, we should add that one can arrive at a matrix representation of DNA sequence also without graphical representation. One such representation is based on using the overall sequential labels and sequential labels of each of the four nucleotides A, T, G, and C separately for construction of matrix elements [9]. Construction of matrices to represent DNA has an important advantage for characterization of DNA in that instead of direct comparison of sequences one can construct vectors, the components of which are various matrix invariants. The similarity between sequences is then transformed in calculation of similarities between n-dimensional vectors, which of course is computationally relatively straightforward [10].

In this contribution we also consider non-graphical representation of DNA by matrices the elements of which indicate some relations between the nucleotides measured by the number of nucleotides between successive pairs of nucleotides of the same kind. In this way with each DNA sequence one can associate a vector. The components of the vector indicate the leading eigenvalue of six distance matrices. The similarity among two DNA sequences can be measured by calculating the Euclidean distance between the end points of the 6-component vectors. Clearly, the smaller is the Euclidean distance the more similar are the two DNA sequences. This approach is illustrated on the primate mitochondrial DNA sequences for 11 different species. Analysis shows an overall qualitative agreement among the three similarity matrices of the primate mitochondrial DNA sequences for 11 different species.

2 Matrices of DNA sequences

Given a DNA sequence with N bases $S = s_1 s_2 \cdots s_N$, $s_i \in \{A, C, G, T\}$, inspect it by stepping one base at a time. Let the number of steps be denoted by i ($i = 1, 2, ..., N$). In the i-th step, count the cumulative numbers of the bases A, C, G and T, denoted by the four positive integers A_i, C_i, G_i and T_i, respectively, occurring in the subsequence from the first s_1 to the i-th base s_i in the DNA sequence inspected. We define $A_0 = C_0 = G_0 = T_0 = 0$.

2.1 Euclidean-Distance Matrix ED

Chemical properties of the DNA bases can be used to classify the four DNA bases A, C, G, and T. As we know, the four DNA bases A, C, G, and T can be divided into three classes, purine {A, G}/pyrimidine {C, T}, amino {A, C}/keto {G, T}, and weak-H bond {A, T}/strong-H bond {C, G}. Then we can define six Euclidean-Distance matrices AG, AC, AT, CT, GT and GC corresponding to the six groups, respectively.

According to the classes: purine {A, G}/pyrimidine {C, T}, the Euclidean-Distance matrices AG and CT can be constructed as follow:

the (i, j) element of matrix AG is defined as:

$$[AG]_{ij} = \sqrt{(A_{ij} - mG_{ij})^2 + (T_{ij} - mC_{ij})^2},$$

the (i, j) element of matrix CT is defined as:

$$[CT]_{ij} = \sqrt{(C_{ij} - mT_{ij})^2 + (A_{ij} - mG_{ij})^2},$$

According to the classes: amino {A, C}/keto {G, T}, the Euclidean-Distance matrices AC and GT can be constructed as follow:

the (i, j) element of matrix AC is defined as:

$$[AC]_{ij} = \sqrt{(A_{ij} - mC_{ij})^2 + (T_{ij} - mG_{ij})^2},$$

the (i, j) element of matrix GT is defined as:

$$[GT]_{ij} = \sqrt{(G_{ij} - mT_{ij})^2 + (A_{ij} - mC_{ij})^2},$$

According to the classes: weak-H bond {A, T}/strong-H bond {C, G}, the Euclidean-Distance matrices AT and GC can be constructed as follow:

the (i, j) element of matrix AT is defined as:

$$[AT]_{ij} = \sqrt{(A_{ij} - mT_{ij})^2 + (C_{ij} - mG_{ij})^2},$$

the (i, j) element of matrix GC is defined as:

$$[GC]_{ij} = \sqrt{(G_{ij} - mC_{ij})^2 + (A_{ij} - mT_{ij})^2},$$

where m is a positive real number. $A_{ij} = A_j - A_i$, $C_{ij} = C_j - C_i$, $G_{ij} = G_j - G_i$ and $T_{ij} = T_j - T_i$ and $i, j = 1, 2, ..., N$.

It should be notice that we should try to find out suitable parameters m so that the mathematical model most appropriate to the problem under consideration. We don't think one kind of parameters m can suit all the biological problems.

2.2 Path-Distance matrix PD

The (i,j)-matrix element of PD is defined as:

$$[PD]_{ji} = [PD]_{ij} = [ED]_{i,i+1} + [ED]_{i+1,i+2} + \ldots + [ED]_{j-1,j}, i < j; [PD]_{ii} = 0,$$

where matrix ED is one of the Euclidean-Distance matrices AG, AC, AT, CT, GT and GC.

2.3 Quotient matrix E/P

The (i,j)-matrix element of E/P is defined to be the quotient of the corresponding elements of the ED matrix and the PD matrix:

$$[E/P]_{ij} = [ED]_{ij}/[PD]_{ij}, i \neq j; [E/P]_{ii} = 0.$$

2.4 Quotient matrix E/G

The (i,j)-matrix element of E/G is defined to be the quotient of the corresponding elements of the ED matrix and the graph theoretical distance between i and j :

$$[E/G]_{ij} = [ED]_{ij}/|i-j|, i \neq j; [E/G]_{ii} = 0.$$

3 6-component vectors of DNA sequences

The leading eigenvalue of the matrix associated with a DNA sequence is an important invariant and is proved to be highly effective for characterization of DNA sequences. We choose the leading eigenvalue of Euclidean-Distance matrices or Quotient matrices as mathematical descriptors of DNA sequence. A disadvantage with graphical representations in general is that comparisons of sequences by visual inspection are an inexact method when sequences have different lengths. As DNA primary sequences usually vary enormously in their lengths, we need use the normalized forms of the leading eigenvalue $\eta = \lambda/N$ instead of the leading eigenvalue λ, where N is the number of bases making up the corresponding DNA sequence, so that we can eliminate the influence of the different lengths of the DNA sequences.

For a given DNA sequence, let $\mu_1, \mu_2, \mu_3, \mu_4, \mu_5, \mu_6$ be the normalized leading eigenvalues of Euclidean-Distance matrices AG, AC, AT, CT, GT and GC (or corresponding Quotient matrices E/P and E/G) of DNA sequence, respectively,

we construct a 6-component vector

$$\eta = (\mu_1, \mu_2, \mu_3, \mu_4, \mu_5, \mu_6)$$

Then we get a correspondence between the DNA sequences and 6-component vectors $(\mu_1, \mu_2, \mu_3, \mu_4, \mu_5, \mu_6)$ of AG, AC, AT, CT, GT and GC (or corresponding Quotient matrices E/P and E/G). So $(\mu_1, \mu_2, \mu_3, \mu_4, \mu_5, \mu_6)$ can characterize the corresponding DNA sequences. Comparison between DNA sequences becomes comparison between these 6-component vectors.

The analysis of similarity/dissimilarity among these DNA sequences represented by the 6-component vectors is based on the assumption that two DNA sequences are similar if the corresponding 6-component vectors in the 6D-space have similar magnitudes and directions.

Let $\eta_i = (\mu_{i1}, \mu_{i2}, \mu_{i3}, \mu_{i4}, \mu_{i5}, \mu_{i6}), i = 1, ..., N$, denote all 6-component vectors of Euclidean-Distance matrices AG, AC, AT, CT, GT and GC (or corresponding Quotient matrices E/P and E/G) of s DNA sequences, then the similarity matrix SED (or SE/P, SE/G) can be formulated as the symmetric matrix whose (i,j) element is defined as the Euclidean distance between the vector η_i and η_j. That is:

$$\sqrt{(\mu_{i1} - \mu_{j1})^2 + (\mu_{i2} - \mu_{j2})^2 + (\mu_{i3} - \mu_{j3})^2 + (\mu_{i4} - \mu_{j4})^2 + (\mu_{i5} - \mu_{j5})^2 + (\mu_{i6} - \mu_{j6})^2},$$

where $i, j = 1, 2, ..., N$.

Table 1: Database Source

species	ID/ ACCESSION	Abbreviation	$length(bp)$	database
Saimiri sciureus	M22655	S. sci	893	$NCBI$
Hylobates	V00659	Hyl	896	$NCBI$
Lemur catta	M22657	Lemur	895	$NCBI$
Macaca fascicular	M22653	M. fas	896	$NCBI$
Gorilla	V00658	Gorilla	896	$NCBI$
Macaca fuscata	M22651	M. fus	896	$NCBI$
Macaca mulatta	M22650	M. mul	896	$NCBI$
Macaca sylvanus	M22654	M. syl	896	$NCBI$
Chimpanzee	V00672	Chi	896	$NCBI$
Orangutan	V00675	Ora	895	$NCBI$
Tarsius syrichta	M22656	T. syr	895	$NCBI$

4 Three similarity matrices of DNA sequences

In this section, we will make a comparison for the sequences of homologous 0.9-kb mtDNA fragments from seven species of primates (four old-world monkeys, a new-world

monkey, and two prosimians) and the 0.9-kb mtDNA fragments from four hominoid species (chimpanzee, gorilla, orangutan and Hylobates). In table 1, the sequences for 11 different species are listed, which are used by Hayasaka, Gojobori and Horai [11]. To examine the present method, we substitute m with different values, and at last find that $m = 2$ is most appropriate for this application.

Table 2: The upper triangular part of the similarities matrix SED for the primate mitochondrial DNA sequences

Species	Chi	Gorilla	Hyl	Lemur	M. Fas	M. Fus	M. Syl	Ora	S. Sci	T. Syr	M. Mul
Chi	0	13.5376	25.3391	65.3855	27.2520	22.9948	38.8056	28.2514	82.6335	86.0183	27.6144
Gorilla		0	18.8291	72.3114	34.4077	25.4363	46.2208	16.4231	88.0730	92.7466	32.3864
Hyl			0	61.2295	27.8937	16.1049	38.1046	29.6900	74.6174	80.7194	23.1335
Lemur				0	38.5190	47.6070	27.0206	87.5401	20.8680	20.8154	40.0837
M. Fas					0	12.7543	13.0356	50.2815	55.4754	59.1904	6.2750
M. Fus						0	22.6386	40.0574	63.2814	67.9122	7.9470
M. Syl							0	61.5513	45.7585	47.6606	15.3071
Ora								0	102.9111	107.8056	47.6113
S. Sci									0	14.3562	56.1565
T. Syr										0	60.5339
M. Mul											0

In Table 2, we give the upper triangular part of the similarities matrix SED . Observing Table 2, we find that, the smallest entries are associated with the pairs (gorilla, chimpanzee), (gorilla, Orangutan),(gorilla, Hylobates), (Macaca fascicular, Macaca fuscata), (Macaca fascicular, Macaca mulatta), (Macaca fascicular, Macaca sylvanus) and (Macaca fuscata, Macaca mulatta).

We give the upper triangular part of the similarity matrix SE/G in Table 3 and the upper triangular part of the similarity matrix SE/P in Table 4. Observing Table 3 and Table 4, we can find 11 species qualitative agreement among similarities based on SED, SE/G and SE/P.

And the main results are similar to that reported in previous studies. So the similarity matrices SED, SE/G and SE/P are suited to numerically characterize DNA sequence.

Table 3: The upper triangular part of the similarities matrix SE/G for the primate mitochondrial DNA sequences

Species	Chi	Gorilla	Hyl	Lemur	M. Fas	M. Fus	M. Syl	Ora	S. Sci	T. Syr	M. Mul
Chi	0	0.0394	0.0724	0.1949	0.0835	0.0717	0.1148	0.0844	0.2491	0.2521	0.0821
Gorilla		0	0.0540	0.2167	0.1062	0.0798	0.1386	0.0510	0.2666	0.2736	0.0986
Hyl			0	0.1862	0.0885	0.0517	0.1179	0.0902	0.2294	0.2408	0.0745
Lemur				0	0.1127	0.1403	0.0832	0.2645	0.0647	0.0592	0.1186
M. Fas					0	0.0401	0.0392	0.1554	0.1662	0.1709	0.0184
M. Fus						0	0.0692	0.1258	0.1884	0.1972	0.0245
M. Syl							0	0.1857	0.1430	0.1394	0.0458
Ora								0	0.3137	0.3208	0.1464
S. Sci									0	0.0506	0.1699
T. Syr										0	0.1761
M. Mul											0

Table 4: The upper triangular part of the similarities matrix SE/P for the primate mitochondrial DNA sequences

Species	Chi	Gorilla	Hyl	Lemur	M. Fas	M. Fus	M. Syl	Ora	S. Sci	T. Syr	M. Mul
Chi	0	0.0252	0.0460	0.1222	0.0516	0.0431	0.0727	0.0531	0.1561	0.1594	0.0504
Gorilla		0	0.0349	0.1370	0.0667	0.0489	0.0890	0.0322	0.1678	0.1741	0.0619
Hyl			0	0.1178	0.0554	0.0317	0.0765	0.0584	0.1438	0.1534	0.0471
Lemur				0	0.0716	0.0902	0.0525	0.1669	0.0435	0.0381	0.0755
M. Fas					0	0.0257	0.0275	0.0974	0.1050	0.1093	0.0113
M. Fus						0	0.0462	0.0777	0.1205	0.1271	0.0163
M. Syl							0	0.1179	0.0920	0.0887	0.0309
Ora								0	0.1978	0.2037	0.0917
S. Sci									0	0.0370	0.1079
T. Syr										0	0.1128
M. Mul											0

5 Construction of the dendrogram tree

Hayasaka, Gojobori and Horai [11] calculated the number of nucleotide substitutions for a given pair of species by the six-parameter method. Using the calculated numbers, they constructed a phylogenetic tree by the NJ method, the distance Wagner method and unweighed pair grouping method, respectively. the algorithms for constructing phylogenetic trees are different from each other. These three different methods give phylogenetic trees with the same topology, the phylogenetic relationships derived from these mtDNA sequence comparisons appear reliable.

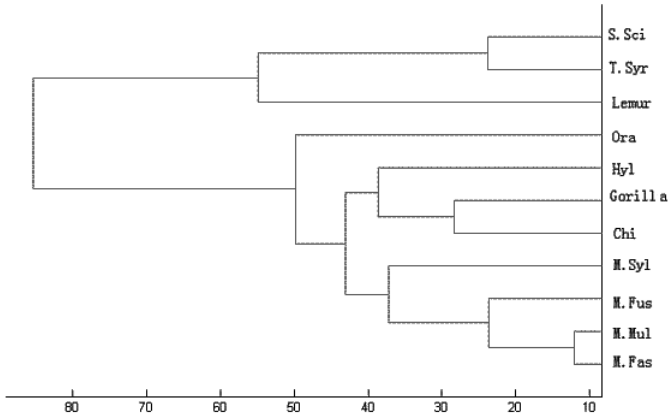

Figure 1: Dendrogram based on SED.

In Figure 1, 2 and 3, we have presented the dendrogram tree based on linkage cluster analysis using Euclidean distances of these 6-dimensional vectors which consist of Table 2, Table 3 and Table 4, respectively, for the 11 different species. The phylogenetic relationships among primate groups shown by our analysis are generally consistent with results in [11,12].

One can also find that the three Dendrogram trees that are constructed based on similarity matrices SED, SE/P and SE/G give us phylogenetic trees with the same topology. The topology of the tree is generally in agreement with previous works.

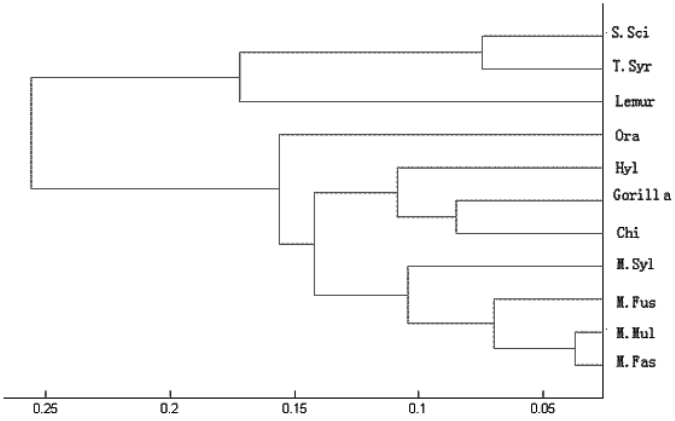

Figure 2: Dendrogram based on SE/G.

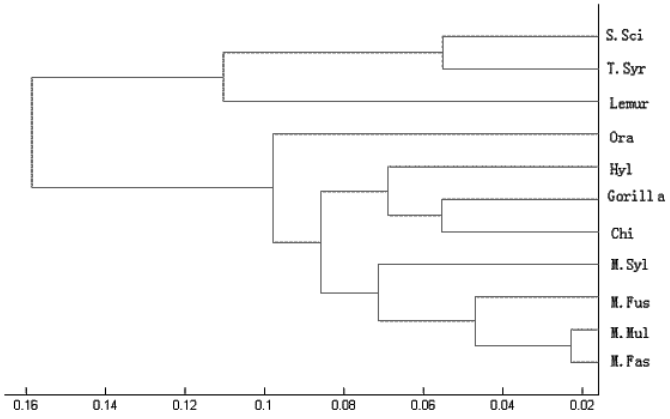

Figure 3: Dendrogram based on SE/P.

6 Conclusions

Three similarity matrices have been constructed to mathematically characterize the DNA sequences. Such matrix allow one to make quantitative comparisons between

different DNA sequences. Our analysis of the sequences of mtDNAs based on the new similarities matrix has provided new insights into evolutionary relationships among primates. Most existing phylogeny construction methods require a multiple alignment of the sequences and assume some sort of an evolutionary model. The proposed method does not require multiple alignment.

Acknowledgements

This work was supported in part by the Shandong Natural Science Foundation (Y2006A14). The author thanks the anonymous referees and editor for their corrections and valuable comments.

References

[1] A. Godzik, The structural alignment between two proteins: is there a unique answer? *Protein Sci.* **5** (1996) 1325–1338.

[2] M. Randić, M. Vračko, N. Lerš, D. Plavšić, Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation, *Chem. Phys. Lett.* **371** (2003) 202–207.

[3] B. Liao, Y. Zhang, K. Ding, T. Wang, Analysis of similarity/dissimilarity of DNA sequences based on a condensed curve representation, *J. Mol. Struct. (Theochem)* **717** (2005) 199–203.

[4] A. Nandy, A new graphical representation and analysis of DNA sequence structure: I. Methodology and Application to Globin Genes, *Curr. Sci.* **66** (1994) 309–314.

[5] M. Randić, M. Vračko, N. Lerš, D. Plavšić, Novel 2-D graphical representation of DNA sequences and their numberical characterization, *Chem. Phys. Lett.* **368** (2003) 1–6.

[6] Y. Zhang, B. Liao, K. Ding, On 2D graphical representation of DNA sequence of nondegeneracy, *Chem. Phys. Lett.* **411** (2005) 28–32.

[7] Y. Zhang, B. Liao, K. Ding, On 3DD-curves of DNA sequences, *Mol. Simul.* **32** (2006) 29–34.

[8] B. Liao, K. Ding, A graphical approach to analyzing DNA sequences, *J. Comput. Chem.* **26** (2005) 1519–1523.

[9] M. Randić, X. F. Guo, S. C. Basak, On the characterization of DNA primary sequence by triplet of nucleic acid bases, *J. Chem. Inf. Comput. Sci.* **41** (2001) 619–626.

[10] G. Jaklič, T. Pisanski, M. Randić, Characterization of complex biological systems by matrix invariants, *J. Comput. Biol.* **13** (2006) 1558–1564.

[11] K. Hayasaka, T. Gojobori, and S. Horai, Molecular phylogeny and evolution of primate mitochondrial DNA, *Mol. Biol. Evol.* **5** (1988) 626–644.

[12] Y. Zhang, W. Chen, New invariant of DNA sequences, *MATCH Commun. Math. Comput. Chem.* **58** (2007) 207–218.