Phylogenetic Analysis by Graphic Representation of DNA Sequences

FENGLAN BAI [a,b]   TIANMING WANG [b,c]
[a] Department of Applied Mathematics
[b] College of advanced science and technology
Dalian university of Technology
Dalian 116024
CHINA
[c] Department of mathematics
Hainan Normal University
Haikou 571158
CHINA

Abstract: - In this paper, we proposed a new method for phylogenetic analysis, based on graphic representations of DNA sequences. Utilizing the invariants of graphs, we give the distance measure of DNA sequences and define the distance between species. We have chosen mitochondrial DNA sequences of 30 species and constructed their phylogenies successfully. The method does not require sequence alignment and is totally automatic.

Key-Words: - DNA sequence; Graphic representation; Invariant of graph; Phylogenetic tree

1 Introduction

Phylogenetic analysis using biological sequences can be divided into two groups. The algorithms in the first group calculate a matrix representing the distance between each pair of sequences and then transform this matrix into a tree. In the second type of approaches, instead of building a tree, the tree that can best explain the observed sequences under the evolutionary assumption is found by evaluating the fitness of different topologies.

Some of the approaches in the first category utilize various distance measures which use different models of nucleotide substation or amino acid replacement. The second category can further be divided into two groups based on the optimal criterion used in tree evaluation: parsimony and maximum likelihood methods. All of these methods require a multiple alignment of the sequences and assume some sort of an evolutionary model. In addition to problems in multiple alignment (computational complexity and inherent ambiguity of the alignment cost criteria), these methods become insufficient for phylogenies using complete genomes. Multiple alignment become misleading due to gene rearrangement, inversion, transposition and translocation at the substring level, unequal length of sequences, etc. and statistical evolutionary models are yet to be suggested for complete genomes. On the other hand, whole genome-based phylogenic analyses are appearing because single gene sequences generally do not possess enough information to construct an evolutionary history of organisms. Factors such as different rates of evolution and horizontal gene transfer make phylogenetic analysis of species using single gene sequences difficult. To overcome these problems, since Sankoff et al.(1992) defined an evolutionary edit distance, several authors have defined various similar distances to apply to genome-based phylogeny. However, these approaches are computationally expensive and do not produce correct results on events such as non-contiguous copies of a gene on the genome or non-decisive gene order. Gene content was proposed by Snel et al. (1999) as a distance measure in genome phylogeny. Similar approaches are taken by others, but such methods fail to work when the gene content of the organisms are very similar. In the early 1990s, various data compression algorithms were applied to the analysis of genetic sequences. Data compression algorithms function by identifying the regularities in the given sequence, and in case of DNA sequences, these regularities would have biological implications. Varre et al.(1999) defined a transformation distance. After that there were a few information distance measures were proposed which used different compressions and complexities.

In this paper we will propose a graphic distance measure based on a graphic representation of DNA sequences for phylogenetic tree construction. In paper [9,10], Liao and Wang gave a tree dimensional representation of DNA sequences, we will use the representation to study phylogenies. Instead of using structural information of DNA sequences in previous measures, we will consider the chemical structure and biological information of DNA sequences.

As for a DNA sequence, it is defined over the set {*A, C, G, T*}. In general, the four bases *A, C, G, T* can be classified into two groups in tree ways, namely by their chemical structures and the strength of the hydrogen bonds. The result is (1) purine $R=\{A,G\}$ and pyrimidine; $Y=\{C,T\}$; (2) amino group $M=\{A,C\}$ and keto group $K=\{G,T\}$; (3) weak H-bond $W=\{A,T\}$ and strong H-bond $S=\{C,G\}$. To give a DNA sequence a graphic representation, first we need to establish a coordinate system *O-xyz,* then put four bases *A, G, T* and *C* on *-x* axis, *+x* axis, *-y* axis and *+y* axis, respectively. The cumulate number of a base is the sequence is put on *z* axis. According to three ways of classification of four bases, every DNA sequence has three graphic representatives, namely, three curves called characteristic curves, see [9]. It can be described mathematically as follows. Let $G=g_1g_2...g_n$ be any DNA sequence, there exist three mappings φ_j, j=1,2,3:

$$\varphi_1(g_i)=\begin{cases}(-1,0,A_i), & \text{if } g_i=A,\\ (1,0,G_i), & \text{if } g_i=G,\\ (0,-1,T_i), & \text{if } g_i=T,\\ (0,1,C_i), & \text{if } g_i=C,\end{cases}$$

$$\varphi_2(g_i)=\begin{cases}(-1,0,A_i), & \text{if } g_i=A,\\ (1,0,G_i), & \text{if } g_i=C,\\ (0,-1,T_i), & \text{if } g_i=T,\\ (0,1,C_i), & \text{if } g_i=G,\end{cases}$$

$$\varphi_3(g_i)=\begin{cases}(-1,0,A_i), & \text{if } g_i=A,\\ (1,0,G_i), & \text{if } g_i=T,\\ (0,-1,T_i), & \text{if } g_i=G,\\ (0,1,C_i), & \text{if } g_i=C,\end{cases}$$

Where $A_i...C_i$ represent the cumulate appearance number of *A, G, T, C* in the sequence, respectively. The mappings φ_j, j=1,2,3 represent patterns *AGTC, ACTG* and *ATGC* respectively. Thus, we have transformed a DNA sequence into a set of points, called a characteristic set of points. Joining the points in the set in turn, we get a curve called characteristic curve. We have established a bijection between DNA sequences and graphic representations. This method got a wide range of applications in similarity analysis of DNA sequences and gene identifying [1-13].

2 Material and method

2.1 Material

We choose mitochondrial DNA sequences of 30 species as research objects, because they are conservative sequences. The variant rate of mitochondrion is 2.2 percent per million years and the difference of mitochondrial DNA sequences is related to variance. The name and serial number of species, see Table 1. All of data are downloaded from website: http://www.ncbi.clm.nih.gov freely.

Table 1 Name and serial number of the species.

species	serial number	species	serial number
human	V00662	rat	X14848
c chimp	D38116	mouse	V00711
p chimp	D38113	opossum	Z29573
gorilla	D38114	wallaroo	Y10524
orangutan	D38115	platypus	X83427
gibbon	X99526	squirrel	AJ238588
baboon	Y18001	fat dormouse	AJ001562
horse	X79547	guinea pig	AJ222767
white rhin	Y07726	donkey	X97337
harbor seal	X63726	Indian rhin	X97336
gray seal	X72004	dog	U96639
cat	U20753	sheep	AF010406
fin whale	X61145	pig	AJ002189
blue whale	X72204	hippopotamus	AJ010957
cow	V00654	rabbit	AJ001588

2.2 Method

Here, we use a numerical descriptor of a graphic representation of a DNA sequence, called average of divergence of a graph, that was proposed by A. Nandy in [3,4,7]. First, defined three components,

$$\mu_x=\tfrac{1}{n}\sum_{i=1}^{N}x(n_i),$$

$$\mu_y=\tfrac{1}{n}\sum_{i=1}^{N}y(n_i),$$

$$\mu_z=\tfrac{1}{n}\sum_{i=1}^{N}z(n_i),$$

where $x(n_i)$, $y(n_i)$ and $z(n_i)$ represent coordinates of the *i*-th point in the curve, respectively. Let $G_R=(\mu_x{}^2+\mu_y{}^2+\mu_z{}^2)^{1/2}$ denote the radii of the graph of a DNA sequence, and $d(G_1,G_2)=[(\mu_x-\mu_x')^2+\ (\mu_y-\mu_y')^2+(\mu_z-\mu_z')^2]^{1/2}$ the distance of two DNA sequences, which was used to construct distance matrix as follows. We take the first ten characters of human's and gray seal's mitochondrial DNA sequences , *GATCACAGGT* and *ACTAATGACT* as examples and gave corresponding characteristic sets und mapping φ_1, φ_2 und φ_3,see Table 2.

Table 2 The first ten characters and coordinates of human's and gray seal's mitochondrions.

human	nucletic	x	y	z	gray seal	nucletic	x	y	z
1	G	1	0	1	1	A	-1	0	1
2	A	-1	0	1	2	C	0	1	1
3	T	0	-1	1	3	T	0	-1	1
4	C	0	1	1	4	A	-1	0	2
5	A	-1	0	2	5	A	-1	0	3
6	C	0	1	2	6	T	0	-1	2
7	A	-1	0	3	7	G	1	0	1
8	G	1	0	2	8	A	-1	0	4
9	G	1	0	3	9	C	0	1	2
10	T	0	-1	2	10	T	0	-1	3

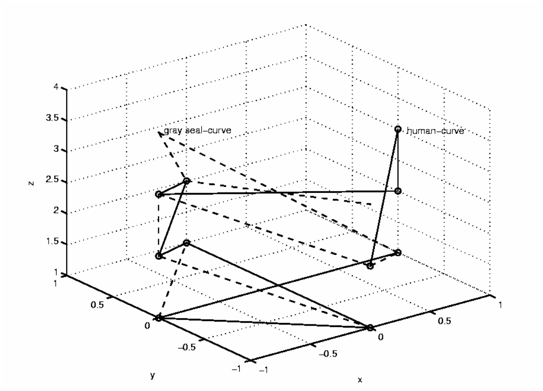

Fig.1 Characteristic curve of the sequence based on the pattern *AGTC*.

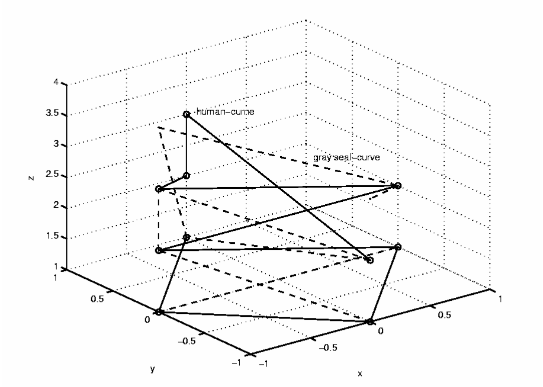

Fig.2 Characteristic curve of the sequence based on the pattern *ACTG*.

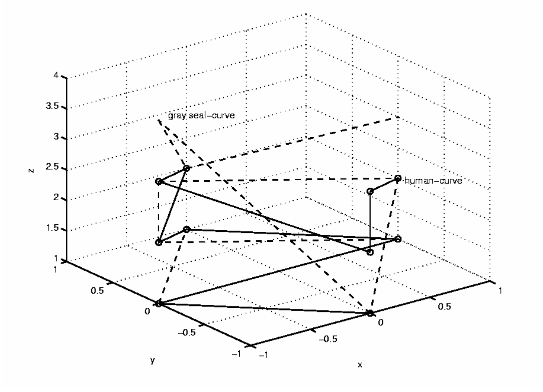

Fig.3 Characteristic curve of the sequence based on the pattern *ATGC*.

To construct the distance matrix of 30 species, we calculated the averages of divergences of the curves corresponding mitochondria under three mappings, see Table 3.

Constructing nine dimension vectors, which components are the corresponding averages of divergences, we calculated the distances between each pair of vectors and got a distance matrix. Because the table too large to put in the paper, we just took a part of the distance matrix show below, see Table 4.

3 Construction of phylogenetic tree

According to the distance matrix in the last paragraph, using Neighbor-Joining method, we constructed the phylogenetic tree of 30 species, see Fig. 4.

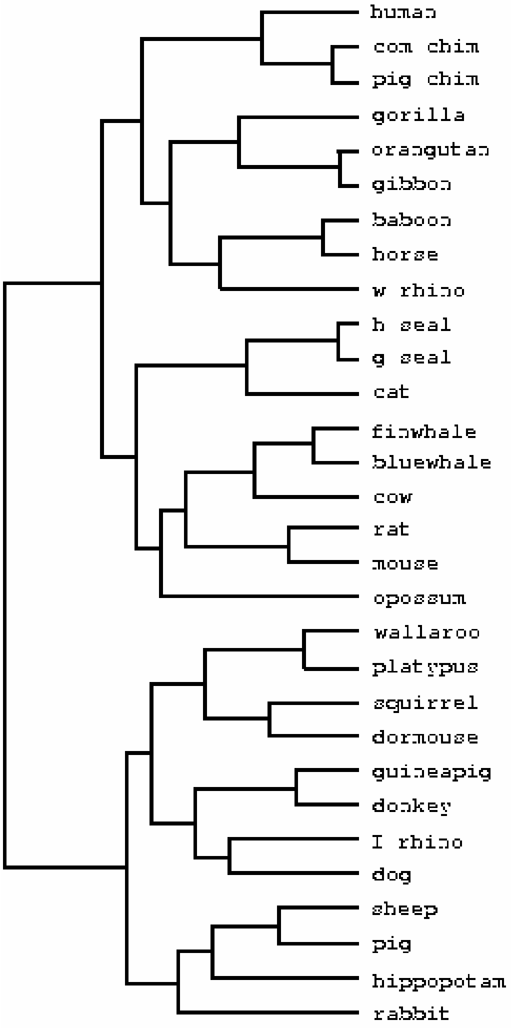

Fig.4 Phylogenetic tree of the 30 species

Table 3 The averages of divergences of the curves corresponding mitochondria of 30 species under the three mappings

species	based on the pattern AGTC			based on the pattern ACTG			based on the pattern ATGC		
human	-0.0002	0.0001	2.2496	0.0000	-0.0001	2.2496	-0.0001	0.0002	2.2496
c chimp	-0.0002	0.0001	2.2562	0.0000	-0.0001	2.2562	-0.0001	0.0002	2.2562
p chimp	-0.0002	0.0001	2.2503	0.0000	-0.0001	2.2503	-0.0001	0.0002	2.2503
gorilla	-0.0002	0.0001	2.2151	0.0000	-0.0001	2.2151	-0.0001	0.0002	2.2151
orangutan	-0.0002	0.0001	2.2345	0.0000	-0.0001	2.2345	-0.0001	0.0002	2.2345
gibbon	-0.0002	0.0001	2.2290	0.0000	-0.0001	2.2290	-0.0001	0.0002	2.2290
baboon	-0.0002	0.0001	2.2418	0.0000	-0.0001	2.2418	-0.0001	0.0002	2.2418
horse	-0.0002	0.0000	2.2485	0.0000	-0.0001	2.2485	-0.0001	0.0002	2.2485
white rhin	-0.0002	0.0000	2.2953	-0.0001	-0.0001	2.2953	-0.0001	0.0002	2.2953
g seal	-0.0002	0.0000	2.2591	-0.0001	-0.0001	2.2591	-0.0001	0.0001	2.2591
cat	-0.0002	0.0000	2.2551	-0.0001	-0.0001	2.2551	-0.0001	0.0001	2.2551
f whale	-0.0002	0.0000	2.2807	-0.0001	-0.0001	2.2807	-0.0001	0.0001	2.2807
b whale	-0.0002	0.0000	2.2175	-0.0001	-0.0001	2.2175	-0.0001	0.0001	2.2175
cow	-0.0002	0.0000	2.2250	-0.0001	-0.0001	2.2250	-0.0001	0.0001	2.2250
rat	-0.0002	0.0000	2.2142	-0.0001	-0.0001	2.2142	-0.0001	0.0001	2.2142
mouse	-0.0002	0.0000	2.2393	-0.0001	-0.0001	2.2393	-0.0001	0.0001	2.2393
opossu	-0.0002	0.0000	2.2536	-0.0001	-0.0002	2.2536	-0.0001	0.0001	2.2536
wallaroo	-0.0002	-0.0001	2.4193	-0.0001	-0.0002	2.4193	0.0000	0.0001	2.4193
platypus	-0.0002	0.0000	2.3001	-0.0001	-0.0001	2.3001	-0.0001	0.0001	2.3001
squirrel	-0.0002	-0.0001	2.3104	-0.0001	-0.0002	2.3104	0.0000	0.0001	2.3104
F dormo	-0.0002	-0.0001	2.2622	-0.0001	-0.0002	2.2622	0.0000	0.0001	2.2622
G pig	-0.0002	-0.0001	2.2842	-0.0001	-0.0002	2.2842	0.0000	0.0001	2.2842
donkey	-0.0002	0.0000	2.2449	-0.0001	-0.0001	2.2449	0.0000	0.0001	2.2449
I rhin	-0.0002	0.0000	2.2587	0.0000	-0.0001	2.2587	-0.0001	0.0002	2.2587
dog	-0.0002	0.0000	2.3016	-0.0001	-0.0001	2.3016	-0.0001	0.0001	2.3016
sheep	-0.0002	0.0000	2.2384	-0.0001	-0.0001	2.2384	0.0000	0.0001	2.2384
pig	-0.0002	0.0000	2.2624	-0.0001	-0.0001	2.2624	-0.0001	0.0001	2.2624
hippopo	-0.0002	0.0000	2.2812	-0.0001	-0.0001	2.2812	-0.0001	0.0001	2.2812
rabbit	-0.0002	0.0000	2.2115	0.0000	-0.0001	2.2115	-0.0001	0.0001	2.2115

Table 4 A part of the distance matrix of mitochondrion DNA sequences of 30 species

species	human	c chim	p chim	gorilla	orangut	gibbon	baboon	horse	w rhin	h seal
human	0	0.0114	0.1069	0.3324	0.5771	0.7605	0.8722	0.9339	0.9696	0.9848
c chimp		0	0.0102	0.1236	0.3536	0.5965	0.7727	0.8792	0.9401	0.9696
p chimp			0	0.0610	0.2484	0.4998	0.7071	0.8409	0.9203	0.9595
gorilla				0	0.0336	0.1849	0.4325	0.6602	0.8243	0.9111
orangutan					0	0.0095	0.0984	0.3147	0.5707	0.7567
gibbon						0	0.0222	0.1527	0.4073	0.6403
baboon							0	0.0116	0.1421	0.3782
horse								0	0.0811	0.2853
white rhin									0	0.0627
harbor sea										0

4 Conclusion

From the Figure 4, we can see that common chimpanzee and pigmy chimpanzee are the most close in phylogenetic relation, then human to them is more close, while orangutan and gibbon are most close then gorilla to then is more close and the relationship among baboon, horse and w rhino is like above, All the nine species mentioned above are in the same clade. The phylogenies inferred using the graphic distance measure can confirm that our method can successfully construct evolutionary histories.

In this paper, the distance measure does not only consider the structure of DNA sequences, but also consider the chemical and biological properties of DNA sequences. Unlike most existing phylogeny construction methods, the proposed method does not require multiple alignments and is full automatic. Therefore, we are able to perform comparison at the whole genome level where multiple alignments based strategies fail. Unequal sequence length or the relatively different positioning similar regions between sequences are not problematic.

The results show that the proposed method can successfully construct phylogenies using either whole genome or single gene. This is quite promising as the genome level phylogeny construction become important with the arrival of such data. Finally, it is worth noting that our distance measure do not use any evolutionary model and seem to be more fitting for whole genome phylogenies where current evolutionary models do not apply directly.

References:

[1] C.T. Zhang, R. Zhang, Analysis of distribution of bases in the coding sequences by a diagrammatic techniqut, *Nucleic Acids Review*, Vol.19, No.22, 1991, pp. 6313-6317.

[2] R. Zhang and C.T. Zhang, Z-curve, an intuitive tool for visualizing and analyzing the DNA sequences, *Journal of Biomolecule Struct and Dynamics*, Vol.11, 1994, pp. 767-782.

[3] A. Nandy, P. Nandy, On the uniqueness of quantitative DNA difference descriptors in 2D graphical representation models, *Chemical Physics Letters*, Vol.368, 2003, pp. 102-107.

[4] A. Nandy, Graphical representation of DNA sequence, *Journal of Chemical Information Computer Science,* Vol.40, 2000, pp. 915-919.

[5] M. Randic, M. Vracko, On the similarty of DNA primary sequence, *Journal of Chemical Information Computer Science*, Vol.40, 2000, pp. 599-606.

[6] C.X. Yuan, B. Liao, T.M. Wang, New 3D graphical representation of DNA sequence and their numerical characterization, *Chemical Physics Letters*, Vol.397, 2003, pp. 412-417.

[7] M. Randic, M. Vracko, A. Nandy, S. C.Basak, On 3-D graphical representation of DNA primary sequence and their numerical characterization, *Journal of Chemical Information Computer Science*, Vol.40, 2000, pp. 1235-1244.

[8] M. Randic, Condensed representation of DNA primary sequence, *Journal of Chemical Information Computer Science,* Vol.40, 2000, pp. 50-56.

[9] B. Liao, T.M. Wang, 3-D graphical representation of DNA sequences and their numerical characterization, *Journal of Molecular Structure: THEOCHEM*, Vol.681, 2004, pp. 209-212.

[10] B. Liao, T.M. Wang, Analysis of similarity of DNA sequences based on 3D graphical representation, *Chemical Physics Letters*, Vol.388, 2004, pp. 195-200.

[11] B. Liao, T.M. Wang, New 2D graphical representation of DNA sequences, *Journal Computational Chemistry*, Vol.25, No.11, pp. 1364-1368.

[12] B. Liao, T.M. Wang, Analysis of similarity of DNA sequences based on triplets, *Journal of Chemical Information Computer Science*, Vol.44, 2004, pp. 1666-1670.

[13] B. Liao, T.M. Shu, K.Q. Ding, A 4D representation of DNA sequences and its application, *Chemical Physic Letters*, Vol.402, 2005，pp. 380-383.

[14] H.O. Hasan, K. Sayood, A new sequence distance measure for phylogenetic tree construction, *Bioinformatics*, Vol.19, No.16, 2003, pp. 2122-2130.

[15] N. Saitou, M. Nei, The neighbor-joining method: a new method for reconstructing phylogenetic tree, *Molecular Biology and Evolution*, Vol.4, 1987, pp. 406-425.

[16] L. Ming, H.B. John, K. Paul, An information based sequence distance form unaligned whole genome protein squence, *Bioinformatics*, Vol.18, No.1, 2002, pp. 100-108.

[17] D.L. Rowe, R.L. Honeycutt, Phylogenetic relationships, ecological correlates, and molecular evolution within the cavioidea (Mammalia,Rodentia), *Molecular Biology and Evolution*, Vol.19, 2002, pp. 263-277.