MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **63** (2010) 691-700

ISSN 0340 - 6253

A Vertical and Horizontal Method for Constructing Phylogenetic Tree

Bo Liao[1*], Lijiao Liao[1*], Guangxue Yue[2], Ronghui Wu[1],Wen Zhu[1]
[1]School of computer and communication, Hunan University,
Changsha Hunan, 410082, China
[2]College of Mathematics and Information Engineering, Jiaxing University,
Jiaxing Zhejiang, 314001, China

(Received September 24, 2009)

Abstract. Phylogenetic reconstruction help us better understand evolutionary relationship through the analyses of DNA sequences. In this paper, as an example of 8 mitochondrial DNA sequences, by means of a 3D graphical representation and the graph radius of the four characteristic curves to construct the similarity matrix, we proposed a vertical and horizontal method for constructing phylogenetic tree. Using our proposed method, a phylogenetic tree can be constructed easily.

1 Introduction

It is an important topic in bioinformatics to study evolution relationship between different species. Constructing their phylogenetic tree can help to reveal the essence of evolutionary force by inter- species implied germ-line relationship. Starting from biological DNA sequence data, there are two types[1]: one of which is the based on the most superior principle approach, the other is the based on algorithm approach for reconstruction phylogenetic trees. At present, the most commonly used two kinds of methods are based on optimal principle of maximum parsimony[2] and maximum

* Corresponding author Fax: +86 731 8821715
E-mail address: dragonbw@126.com L. Liao)
* Corresponding author Fax: +86 731 8821715
E-mail address: dragonbw@163.com (B. Liao)

likelihood[3,4], but it is proved that construction of the greatest parsimony and maximum likelihood tree for n objects is NP[5] .

The traditional distance measure includes p distance, Kimura distance, τ distance and so on. Their common feature is based on between sequences alignment for distance calculation. But the alignment analysis is very demanding for the experimental data, at the same time, it is considerable empirical to use the scoring matrix of alignment. Based on this, many scholars try to use the non-matching approach to align the DNA sequences.

The most common algorithm-based method for building a phylogenetic tree is the distance method. The traditional method such as UPGMA [6], NJ [7], as well as the improved algorithm proposed in recent years [8, 9]. In this paper, we proposed a new distance method—vertical and horizontal method. As an example of 8 mitochondrial DNA sequences , we combined the 3D graphical representation of the DNA sequences proposed by Li and Wang [10], with the graph radius for constructing familiarity matrix to construct the phylogenetic tree. It is proved the feasibility and availability of our method.

2 Similarity analysis based on sequence descriptors comparison

2.1 The 3-D graphical representation and numerical characterization of DNA sequences

As we all know that DNA sequence have four bases A, C, G, T. The four bases are assigned to the following original codes respectively: A(1,0,0);G(0,1,0);C(0,0,1); T(1,1,1). A point in a 3-D space is represented by three coordinates x,y,z, and a base corresponds a unique point in a 3-D space. Using these codes, we can obtain an encoding curve which is called AGC-T curve [10] as follows.

Let sequence $S = S_1 S_2 \cdots S_N$, we can reduce sequence S into a series of nodes $P_0, P_1, \cdots, P_N$, whose coordinates x_i, y_i, z_i (i=0,1,2,....,N, where N is the length of the studied sequence) satisfy the following equation[12]:

$$\begin{cases} x_i = \sum_{j=1}^{i} S_j^1 \\ y_i = \sum_{j=1}^{i} S_j^2 \\ z_i = \sum_{j=1}^{i} S_j^3 \end{cases}$$

where S^k_j(k = 1, 2, 3) represents the k-th component of the vector corresponding to Sj.

If we assign T,A,C,G to (1,0,0),(0,1,0) (0,0,1),(1,1,1), respectively, we can obtain another encoding curve. Although the four bases can be assigned in 24 ways, there are only four essentially different encoding curves, which are called AGC-T curve, TAC-G curve, TGA-C curve, TGC-A curve, respectively. For example, the coordinates of sequence GTTTATGTAG are listed in Table 1.

Table 1 the 3-D coordinates for the first 10 bases of mitochondrial DNA sequence of p chimp

number	base	AGC-T			TAC-G			TGA-C			TGC-A		
		x	y	z	x	y	z	x	y	z	x	y	z
1	G	0	1	0	1	1	1	0	1	0	0	1	0
2	T	1	2	1	2	1	1	1	1	0	1	1	0
3	T	2	3	2	3	1	1	2	1	0	2	1	0
4	T	3	4	3	4	1	1	3	1	0	3	1	0
5	A	4	4	3	4	2	1	3	1	1	4	2	1
6	T	5	5	4	5	2	1	4	1	1	5	2	1
7	G	5	6	4	6	3	2	4	2	1	5	3	1
8	T	6	7	5	7	3	2	5	2	1	6	3	1
9	A	7	7	5	7	4	2	5	2	2	7	4	2
10	G	7	8	5	8	5	3	5	3	2	7	5	2

According to the geometry of the DNA sequence, we give a comparable indicator, which is called as the geometrical center of the points corresponding DNA curve, formula as follow:

$$\begin{cases} \mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i \\ \mu_y = \frac{1}{N}\sum_{i=1}^{N} y_i \\ \mu_z = \frac{1}{N}\sum_{i=1}^{N} z_i \end{cases}$$

2.2 Experimental data

Because of species differences of mitochondrial DNA sequences only related to their variation, the mitochondrial DNA sequences are mutating at the rate of the percentage of 2.2 per million years, and they are the conserved sequences. So we selected 8 species of mitochondrial DNA sequence for this study. This experimental data are downloaded from GenBank (http://www.ncbi. nlm.nih.gov/). Species name and serial number as follow Table 2**:**

Table 2 eight species information of mitochondrial DNA sequence

No	Species Scientific Name	abbreviation	Accession
1	Pan paniscus	p chimp	D38113
2	Gorilla gorilla	gorilla	D38114
3	Pongo pygmaeus	oranguta	D38115
4	Pan troglodytes	c chimp	D38116
5	Equus caballus	horse	X79547
6	Hylobates lar	gibbon	X99256
7	Ceratotherium simum	w rhino	Y07726
8	Papio hamadryas	baboon	Y18001

2.3 Calculation of descriptors

Similar with A. Nandy's graph radius [11], we proposed a descriptor to construct similarity matrix, the formula is listed as follow:

$$\rho = \sqrt{\mu_x{}^2 + \mu_y{}^2 + \mu_z{}^2}$$

The geometrical centers are listed in Table 3 associated with four different curves representing each of the encoding mitochondrial DNA sequences.

Table 3 The geometrical center associated with four different curves representing each of the encoding mitochondrial DNA sequences.

species	AGC-T	TAC-G
p chimp	(4685.027832,3214.665527,4521.959961)	(3214.665527,3755.540283,3592.472412)
gorilla	(4617.371582,3204.585205,4466.686523)	(3204.585205,3715.813721,3565.128174)
oranguta	(4471.153320,3101.115479,4486.245605)	(3101.115479,3708.754639,3723.846680)
c chimp	(4708.655273,3212.718994,4520.610840)	(3212.718994,3761.389160,3573.344971)
horse	(4906.453613,3311.989746,4403.009766)	(3311.989746,3927.489990,3424.046631)
gibbon	(4501.492188,3159.785400,4475.025391)	(3159.785400,3761.474365,3735.007813)
w rhino	(4999.175293,3308.447510,4402.015625)	(3308.447510,4014.484375,3417.324707)
baboon	(4672.000488,3201.637695,4472.509277)	(3201.637695,3787.490723,3587.999756)
species	TGA-C	TGC-A
p chimp	(4521.959961,3592.472412,5062.834473)	(4685.027832,3755.540283,5062.834473)
gorilla	(4466.686523,3565.128174,4977.914551)	(4617.371582,3715.813721,4977.914551)
oranguta	(4486.245605,3723.846680,5093.884766)	(4471.153320,3708.754639,5093.884766)
c chimp	(4520.610840,3573.344971,5069.280762)	(4708.655273,3761.389160,5069.280762)
horse	(4403.009766,3424.046631,5018.510254)	(4906.453613,3927.489990,5018.510254)
gibbon	(4475.025391,3735.007813,5076.714844)	(4501.492188,3761.474365,5076.714844)
w rhino	(4402.015625,3417.324707,5108.052734)	(4999.175293,4014.484375,5108.052734)
baboon	(4472.509277,3587.999756,5058.362305)	(4672.000488,3787.490723,5058.362305)

We gave the graph radius associated with four different curves representing each of the encoding mitochondrial DNA sequences in Table 4.

Table 4 The graph radius associated with four different curves representing each of the encoding mitochondrial DNA sequences.

species	AGC-T	TAC-G	TGA-C	TGC-A
p chimp	7261.658203	6110.975098	7680.251953	7854.034728
gorilla	7179.190430	6065.210449	7578.988281	7739.962869
oranguta	7052.270996	6102.361328	7742.163574	7726.172079
c chimp	7275.226563	6102.328125	7674.785156	7875.093037
horse	7377.604980	6174.021973	7503.068359	8042.630779
gibbon	7090.382813	6171.160156	7729.758789	7757.909185
w rhino	7437.426758	6224.147949	7559.633301	8197.563122
baboon	7216.745117	6121.218262	7646.182617	7858.734233

2.4 Similarity Analysis

In order to compute the similarity of the studied DNA sequences, we will construct a four-component vector consisting of the graph radius associated with four different patterns of the characteristic curves. The distance Dij between two vectors is computed by the following formula [12].

$$Dij = \left[\left(\frac{\rho_1^i}{N^i+1} - \frac{\rho_1^j}{N^j+1} \right)^2 + \left(\frac{\rho_2^i}{N^i+1} - \frac{\rho_2^j}{N^j+1} \right)^2 + \left(\frac{\rho_3^i}{N^i+1} - \frac{\rho_3^j}{N^j+1} \right)^2 + \left(\frac{\rho_4^i}{N^i+1} - \frac{\rho_4^j}{N^j+1} \right)^2 \right]^{\frac{1}{2}}$$

The similarity matrix based on the Euclidean distances between the points of the four-curve of the graph radius of table 3 is shown in Table 5.

Table 5 The similarity matrix based on the Euclidean distances between the points of the four-curve of the graph radius of Table 4.

species	P chimp	gorilla	oranguta	c chimp	horse	gibbon	w rhino	baboon
p chimp	1.000000	0.002236	0.012673	0.001492	0.016522	0.011745	0.019713	0.002836
gorilla	0.002236	1.000000	0.012715	0.003369	0.016606	0.011217	0.020109	0.003314
oranguta	0.012673	0.012715	1.000000	0.013897	0.027819	0.003916	0.030418	0.012483
c chimp	0.001492	0.003369	0.013897	1.000000	0.015485	0.013072	0.018581	0.003248
horse	0.016522	0.016606	0.027819	0.015485	1.000000	0.025802	0.004624	0.015529
gibbon	0.011745	0.011217	0.003916	0.013072	0.025802	1.000000	0.028586	0.011020
w rhino	0.019713	0.020109	0.030418	0.018581	0.004624	0.028586	1.000000	0.018495
baboon	0.002836	0.003314	0.012483	0.003248	0.015529	0.011020	0.018495	1.000000

3 Constructing phylogenetic tree by vertical and horizontal method

Next, we will propose a new method to construct phylogenetic tree, which is called Vertical and horizontal method. This approach is that starting from a point to find another point of the smallest distance in its horizontal, and then from another point to find the edge of the smallest distance between this point and the marked point in its vertical. Concrete steps are as follow:

1) For given two sets: point set V and edge set E.

2) Starting from the any line(for example: the first line),where $V=\{v_1\}$, to find the minimum value of this line e_i, marking the coding number v_i and the first coding

number v_1, where V={v_1.v_i} and E={e_i}and connecting $v_1$with v_i ;

3) Starting from the v_i line, to find the minimum value of this line Min_i, marking the coding number v_j, where V={ v_1,v_i , v_j };

4) Starting from the v_j column, to find the value of the marked code, comparing the size between them, and choosing the minimum value e_j, where E={e_i ,e_j } remembering the coding number v_k, and connecting v_j with k_k ;

5) Repeating the 2) step , until all coding number are marked.

Based on the above steps, we draw maximum tree from the data in table 5, as shown in figure 1. In order to facilitate making the drawing, we use 1,2,…,8 the eight digits to represent the eight species: 1 means p chimp, 2 means gorilla, 3 means orangutan, 4 means c chimp,5 means horse, 6 means gibbon, 7 means w rhino, 8 means baboon.

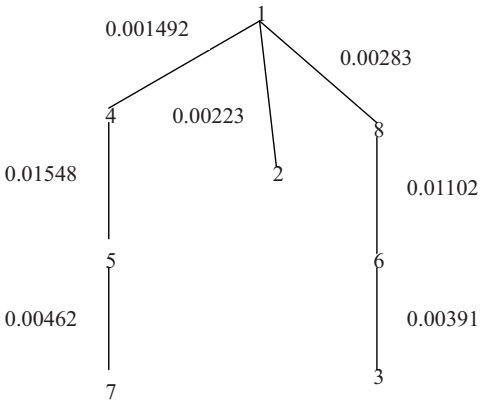

Figure 1: the maximum tree

According to the order from small to large for the tree weight in figure 1, we construct a phylogenetic tree, as shown in figure 2. Using the DRAWGRAM program in the PHYLIP package (http://evolution.genetics.\\washington.edu/ phylip.html), we can obtain the similar phylogenic tree. Compared with figure 2 and figure 3, we can

find that the phylogenetic tree constructed by this method is very familiar. The similarities are P chimp and c chimp, gibbon and orangutan, and horse and w rhino, these evolutionary relationships are very close. It is proved the feasibility and availability of our method.

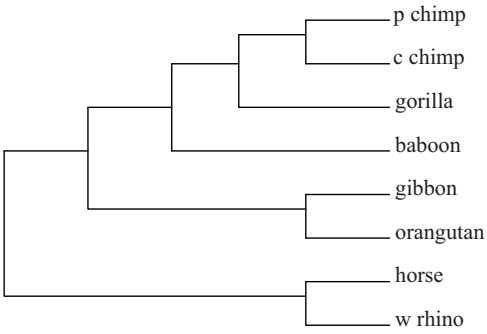

Figure 2: constructing phylogenetic tree using our method

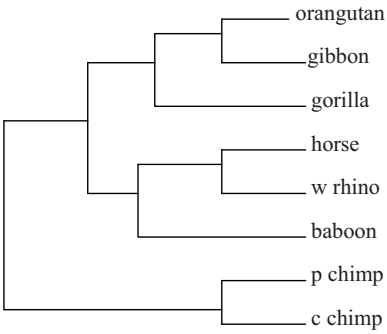

Figure 3: phylogenetic tree using the NEIGHBOR program in the PHYLIP

Conclusion

In order to visually reflect the evolutionary relationships between species, the construction of evolutionary tree is proposed. Graphical representation of

DNA sequences give distance metric can be used as infer evolutionary distance metric of evolutionary relationships between species. Based on this, we according to the characteristics of matrix, proposed a new vertical and horizontal method, which is simple, computing less, fast and so on. Using this method and combination of graphical representation of the DNA sequences, we can construct a phylogenetic tree easily.

Acknowledgment

This work is supported in part by the National Nature Science Foundation of China (Grant 60973082), the National Nature Science Foundation of Hunan province (Grant 07JJ5080), the Planned Science and Technology Project of Hunan Province (Grant 2009FJ3195) and the National Nature Science Foundation of Zhejiang province (Grant Y1090264).

References

[1] M. Nei, S. Kumar, *Molecular Evolution and Phylogenetics*, Oxford University Press, Oxford, 2002.

[2] J. Sourdis, M. Nei, Relative efficiencies of the maximum parsimony and distance matrix methods in obtaining the correct phylogenetic tree, *Mol. Biol. Evol.* **5** (1988) 298 – 311.

[3] M. Holder, P.O. Lewis, Phylogeny estimation: traditional and Bayesian approaches, *Nat. Rev. Genet.* **4** (2003) 275-284.

[4] W. H. Li, Evolutionary change of restriction cleavage sites and phylogenetic inference, *Genetics* **113** (1986) 187-213.

[5] S. Roch, A short proof that phylogenetic tree reconstruction by maximum likelihood is hard, *ACM T. Comput. Biol. Bioinformatics* **3** (2006) 92-94.

[6] P. H. A.Sneath, R. R. Sokal, *Numerical Taxonomy: The Principles and Practice of Numerical Classification*, Freeman, San Francisco, 1973.

[7] N. Saitou, M. Nei, The neighbor-joining method: A new method for reconstructing phylogenetic trees, *Mol. Biol. Evol.* **4** (1987) 406-425.

[8] I. Elias, J. Lagergren, Fast neighbor joining, *Theor. Comput. Sci.* **410** (2009)

1993-2000.

[9] J. H Lee, R. Liu, A fuzzy clustering algorithm based on fuzzy distance norms for asynchronously sampled data, *11th IEEE International Conference on Computational Science and Engineering*, **10** (2008) 361-368.

[10] C. Li, J. Wang. On a 3-D representation of DNA primary sequences, *Comb. Chem. High. T. Scr.* **7** (2004) 23-27.

[11] C. Raychaudhury, A. Nandy, Indexing scheme and similarity measures for maciomolecular sequences, *J. Chem. Inf. Comput. Sci.* **39** (1999) 243-247.

[12] Y. H. Yao, X. Y. Nan, T. M. Wang, Analysis of similarity/dissimilarity of DNA sequences based on 3-D graphical representation, *Chem. Phys. Lett.* **411** (2005) 248-255.