Coronavirus Phylogeny Based on 2D Graphical Representation of DNA Sequence

BO LIAO,[1,*] XUYU XIANG,[2] WEN ZHU[1]

[1]*Laboratory of Embedded Computing and System, School of Computer and Communication, Hunan University, Changsha Hunan 410082, People's Republic of China*

[2]*Department of Computer Science, Hunan City University, Yiyang 413000, People's Republic of China*

Received 25 November 2005; Accepted 22 January 2006

DOI 10.1002/jcc.20439

Published online in Wiley InterScience (www.interscience.wiley.com).

Abstract: A novel coronavirus has been identified as the cause of the outbreak of severe acute respiratory syndrome (SARS). Previous phylogenetic analyses based on sequence alignments show that SARS-CoVs form a new group distantly related to the other three groups of previously characterized coronaviruses. In this aritcle, a new approach based on the 2D graphical representation of the whole genome sequence is proposed to analyze the phylogenetic relationships of coronaviruses. The evolutionary distances are obtained through measuring the differences among the two-dimensional curves.

 J Comput Chem 27: 1196–1202, 2006

Key words: phylogenetic tree; Coronavirus; SARS-CoV; graphical representation

Introduction

The outbreak of atypical pneumonia, referred to as severe acute respiratory syndrome (SARS) was first identified in Guangdong Province, China, and spread to several countries later. A novel coronavirus was isolated and found to be the cause of SARS. The SARS-coronavirus is a new member of the order Nidovirales, family Coronaviridae, and genus Coronavirus. Some researchers have considered the mutation analysis and phylogenetic analysis.[1–6]

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. For example, Jukes and Cantor,[7] Kimura,[8] Barry and Hartigan,[9] Kishino and Hasegawa,[10] and Lake[11] proposed various distance measures. Camin and Sokal,[12] Eck and Dayhoff,[13] Cavalli-Sforza and Edwards,[14] and Fitch[15] gave parsimony methods. Felsenstein et al.[16–18] proposed maximum likelihood methods.

But, 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.

Mathematical analysis of the large volume genomic DNA sequence data is one of the challenges for bioscientists. Graphical representation of DNA sequence provides a simple way of viewing, sorting, and comparing various gene structures. In recent years several authors outlined different graphical representation of DNA sequences based on 2D, 3D, or 4D.[19–32] Graphical techniques have emerged as a very powerful tool for the visualization and analysis of long DNA sequences. These techniques provide useful insights into local and global characteristics and the occurrences, variations, and repetition of the nucleotides along a sequence that are not as easily obtainable by other methods.[29,33] Based on these graphical representation several authors outlined some approaches to make comparison of DNA sequences[34–38]. Recently, we present a new

Correspondence to: B. Liao; e-mail: dragonbw@163.com

Contract/grant sponsor: National Natural Science Foundation of China; contract/grant number: 10571019

Contract/grant sponsor: China Postdoctor of Science Foundation

Contract/grant sponsor: National Natural Science Foundation of Human University

two-dimensional graphical representation of DNA sequences, which has no circuit or degeneracy.[19]

Here, a new approach based on the 2D graphical representation of the whole genome sequence is proposed to analyze the phylogenetic relationships of genomes. The evolutionary distances are obtained through measuring the differences among the 2D curves. The examination of the phylogenetic relationships of coronaviruses illustrates the utility of our approach.

2D Graphical Representation of DNA Sequences

As shown in Figure 1, which is similar with Yan's[34] method, we construct a pyrimidine–purine graph on two quadrants of the cartesian coordinate system, with pyrimidines(T and C) in the first quadrant and purines(A and G) in the fourth quadrant. The unit vectors representing four nucleotides A,G,C, and T are as follows:

$$(m, -\sqrt{n}) \longrightarrow A, (\sqrt{n}, -m) \longrightarrow G, (\sqrt{n}, m) \longrightarrow C, (m, \sqrt{n}) \longrightarrow T$$

where m is a real number, n is a positive real number but not a perfect square number. Using this representation, we will reduce a DNA sequence into a series of nodes $P_0, P_1, P_2, \ldots, P_N$, whose coordinates $x_i, y_i (i = 0, 1, 2, \ldots, N$, where N is the length of the DNA sequence being studied) satisfy

$$\begin{cases} x_i = a_i m + g_i\sqrt{n} + c_i\sqrt{n} + t_i m \\ y_i = -a_i\sqrt{n} - g_i m + c_i m + t_i\sqrt{n} \end{cases}$$

where a_i, c_i, g_i and t_i are the cumulative occurrence numbers of A, C, G, and T, respectively, in the subsequence from the first base to the i-th base in the sequence. We define $a_0 = c_0 = g_0 = t_0 = 0$.

We called the corresponding plot set a characteristic plot set. The curve connecting all plots of the characteristic plot set, in turn, is called the characteristic curve, which is determined by m,n, that satisfy the above mentioned condition. In Figure 2, we show the chimpanzee corresponding curves with different parameters n and m. Observing Figure 2, we find that chimpanzees have similar curves despite corresponding different parameters of n and m. They have the same tendency despite different lengths. In Figure 3, we present the 2D curves for 24 complete coronavirus genomes (see Table 1) with parameters $n = 1/2$ and $m = 3/4$ chosen initially by Yan et al.[34]

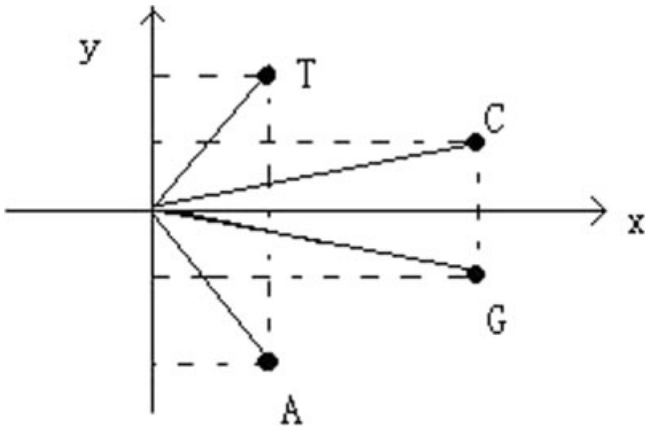

Figure 1. Pyrimidine–purine graph.

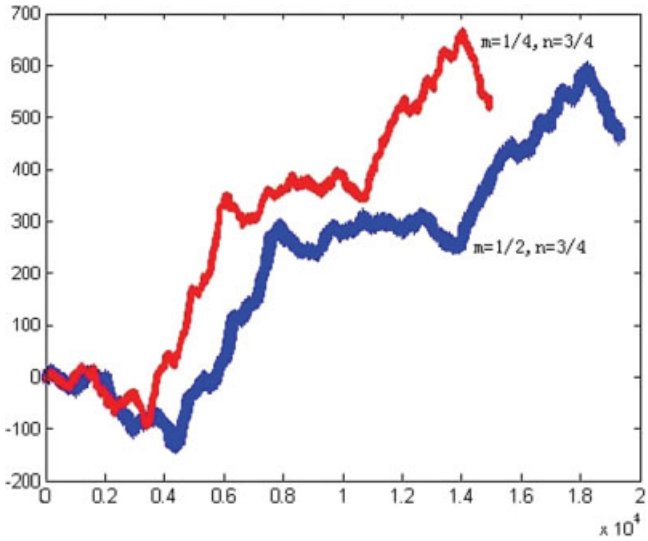

Figure 2. The chimpanzee corresponding curves with different parameters n and m.

Observing Figure 3, we find that the curves of BCoV, BCoVL, BCoVM, and BCoVQ have some similar tendencies. The curves of MHV2, MHV, MHVM, and MHVP have some similar tendencies. The curves of BJ01, CUHK-Su10, CUHK-W1, SIN2679, SIN2748, SIN2774, HKU-39849, SIN2500, SIN2677, TW1, Urbani, and TOR2 have some similar tendencies.

Phylogenetic Tree of Coronaviruses

For any sequence, we have a set of points $(x_i, y_i), i = 1, 2, 3, \ldots, N$, where N is the length of the sequence. The coordinates of the geometrical center of the points, denoted by x^0 and y^0, may be calculated as follows[29]

$$x^0 = \frac{1}{N}\sum_{i=1}^{N} x_i,\ y^0 = \frac{1}{N}\sum_{i=1}^{N} y_i. \tag{1}$$

The element of the covariance matrix CM of the points are defined:

$$\begin{cases} CM_{xx} = \frac{1}{N}\sum_1^N (x_i - x^0)(x_i - x^0) \\ CM_{xy} = \frac{1}{N}\sum_1^N (x_i - x^0)(y_i - y^0) = CM_{yx}\,. \\ CM_{yy} = \frac{1}{N}\sum_1^N (y_i - y^0)(y_i - y^0) \end{cases} \tag{2}$$

The above four numbers give a quantitative description of a set of point $(x_i, y_i), i = 1, 2, \ldots, N$, scattering in a 2D space. Obviously,

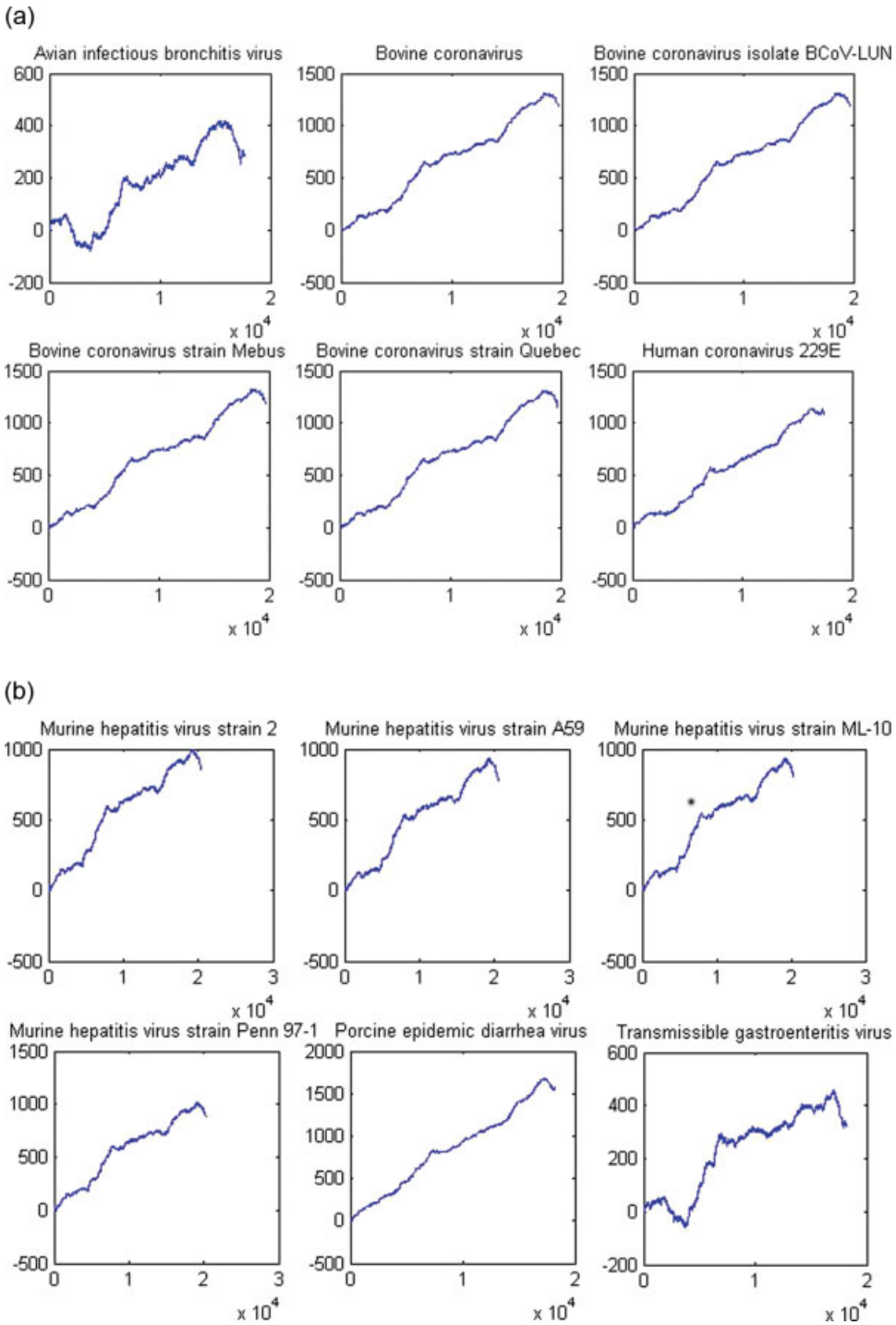

Figure 3. (A) IBV, BCoV, BCoVL, BCoVM, BCoVQ, HCoV-229E complete genome. (B) MHV2, MHV, MHVM, MHVP, PEDV, TGEV complete genome. (C) BJ01, CUHK-Su10, CUHK-W1, SIN2679, SIN2748, SIN2774 complete genome. (D) HKU-39849, SIN2500, SIN2677, TW1, Urbani, TOR2 complete genome. The two-dimensional curves for 24 complete coronavirus genomes. (A–D) The curves of IBV, BCoV, BCoVL, BCoVM, BCoVQ, HCoV-229E, MHV2, MHV, MHVM, MHVP, PEDV, TGEV, BJ01, CUHK-Su10, CUHK-W1, SIN2679, SIN2748, SIN2774, HKU-39849, SIN2500, SIN2677, TW1, Urbani, and TOR2, respectively. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

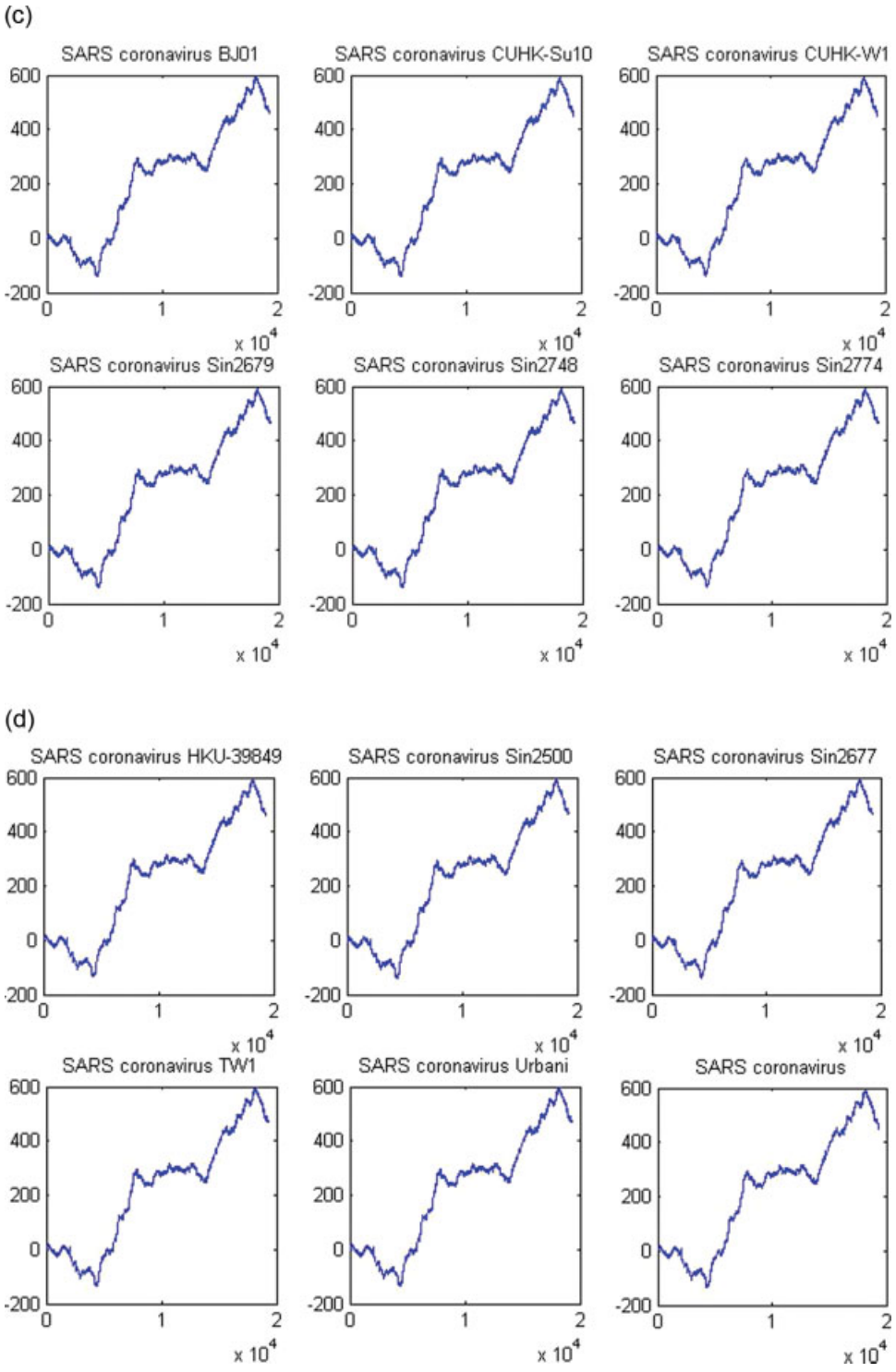

Figure 3. (*continued*)

Table 1. The Accession Number, Abbreviation, Name, and Length for the 24 Coronavirus Genomes.

No.	Accession	Abbreviation	Genome	Length(nt)
1	*NC*_002645	*HCoV*_229*E*	Human coronavirus 229E	27,317
2	*NC*_002306	TGEV	Transmissible gastroenteritis virus	28,586
3	*NC*_003436	PEDV	Porcine epidemic diarrhea virus	28,033
4	U00735	BCoVM	Bovine coronavirus strain Mebus	31,032
5	AF391542	BCoVL	Bovine coronavirus isolate BCoV-LUN	31,028
6	AF220295	BCoVQ	Bovine coronavirus Quebec	31,100
7	*NC*_003045	BCoV	Bovine coronavirus	31,028
8	AF208067	MHVM	Murine hepatitis virus strain ML-10	31,233
9	AF101929	MHV2	Murine hepatitis virus strain 2	31,276
10	AF208066	MHVP	Murine hepatitis virus strain Penn 97-1	31,112
11	*NC*_001846	MHV	Murine hepatitis virus	31,357
12	*NC*_001451	IBV	Avian infectious bronchitis virus	27,608
13	AY278488	BJ01	SARS coronavirus BJ01	29,725
14	AY278741	Urbani	SARS coronavirus Urbani	29,727
15	AY278491	HKU-39849	SARS coronavirus HKU-39849	29,742
16	AY278554	CUHK-W1	SARS coronavirus CUHK-W1	29,736
17	AY282752	CUHK-Su10	SARS coronavirus CUHK-Su10	29,736
18	AY283794	SIN2500	SARS coronavirus Sin2500	29,711
19	AY283795	SIN2677	SARS coronavirus Sin2677	29,705
20	AY283796	SIN2679	SARS coronavirus Sin2679	29,711
21	AY283797	SIN2748	SARS coronavirus Sin2748	29,706
22	AY283798	SIN2774	SARS coronavirus Sin2774	29,711
23	AY291451	TW1	SARS coronavirus TW1	29,729
24	*NC*_004718	TOR2	SARS coronavirus	29,751

Table 2. The Geometric Center and Two Eigenvectors for each of the 24 Coronavirus Genomes.

i	x^0	y^0	$EV^i_{\lambda_1}$	$EV^i_{\lambda_2}$
1	8.7251e+003	567.4895	(0.0671,−0.9977)	(−0.9977,−0.0671)
2	9.1181e+003	231.8617	(0.0265,−0.9996)	(−0.9996,−0.0265)
3	9.1658e+003	854.0672	(0.0891,−0.9960)	(−0.9960,−0.0891)
4	9.8471e+003	678.7491	(0.0682,−0.9977)	(−0.9977,−0.0682)
5	9.8494e+003	669.8507	(0.0683,−0.9977)	(−0.9977,−0.0683)
6	9.8708e+003	671.8188	(0.0678,−0.9977)	(−0.9977,−0.0678)
7	9.8504e+003	667.9839	(0.0684,−0.9977)	(−0.9977,−0.0684)
8	1.0225e+004	508.6553	(0.0456,−0.9990)	(−0.9990,−0.0456)
9	1.0217e+004	560.8241	(0.0484,−0.9988)	(−0.9988,−0.0484)
10	1.0166e+004	571.4215	(0.0492,−0.9988)	(−0.9988,−0.0492)
11	1.0266e+004	503.3193	(0.0457,−0.9990)	(−0.9990,−0.0457)
12	8.8359e+003	177.6139	(0.0271,−0.9996)	(−0.9996,−0.0271)
13	9.6653e+003	217.7081	(0.0348,−0.9994)	(−0.9994,−0.0348)
14	9.6644e+003	220.2759	(0.0347,−0.9994)	(−0.9994,−0.0347)
15	9.6693e+003	219.4720	(0.0345,−0.9994)	(−0.9994,−0.0345)
16	9.6690e+003	217.1652	(0.0346,−0.9994)	(−0.9994,−0.0346)
17	9.6687e+003	217.0494	(0.0346,−0.9994)	(−0.9994,−0.0346)
18	9.6602e+003	216.5541	(0.0347,−0.9994)	(−0.9994,−0.0347)
19	9.6587e+003	216.9280	(0.0347,−0.9994)	(−0.9994,−0.0347)
20	9.6601e+003	216.0181	(0.0346,−0.9994)	(−0.9994,−0.0346)
21	9.6583e+003	216.5654	(0.0347,−0.9994)	(−0.9994,−0.0347)
22	9.6601e+003	216.0584	(0.0346,−0.9994)	(−0.9994,−0.0346)
23	9.6656e+003	220.1538	(0.0347,−0.9994)	(−0.9994,−0.0347)
24	9.6724e+003	219.6501	(0.0346,−0.9994)	(−0.9994,−0.0346)

the matrix is a real symmetric 2×2 one. The eigenvectors and their associated eigenvalues are defined as follows:

$$CM \cdot EV_k = \lambda_k \cdot EV_k,\ EV_k = (EV_{k,1},\ EV_{k,2})^T,\ k = 1, 2.$$

Corresponding to each eigenvalue λ_k, there's an eigenvector EV_k. Corresponding to $\lambda_1 < \lambda_2$, the two eigenvectors are denoted by $EV_{\lambda_1}, EV_{\lambda_2}$, respectively. In Table 2, we list the (x^0, y^0) and eigenvectors belonging to 24 species with parameters $m = \frac{1}{2}, n = \frac{3}{4}$.

To facilitate the quantitative comparison of different species in terms of their collective parameters, we introduce a distance scale and an angle scale as defined below. Suppose that there are two species i and j, the parameters are $x_i^0, y_i^0, \lambda_1^i, \lambda_2^i, x_j^0, y_j^0, \lambda_1^j, \lambda_2^j$, respectively, where (x_i^0, y_i^0) is the geometrical center of the curve belonging to species i. λ_1^i, λ_2^i are the two eigenvalues of matrix CM_i corresponding to species i. The distance d_{ij} between the two points is.[39]

$$d_{ij} = \sqrt{(x_i^0 - x_j^0)^2 + (y_i^0 - y_j^0)^2},\ i,j = 1, 2, \ldots, M \qquad (3)$$

where d_{ij} denotes the distance between the geometric centers of the ith and the jth genomes, and M is the total number of all genomes ($M = 24$, here). Then we obtain a real $M \times M$ symmetric matrix whose elements are d_{ij}.

To reflect the differences between the trends of every two 2D curves, the angles between the corresponding eigenvectors of every two genomes are used. The 2D vectors are denoted as follows:

$$EV_k^i = \left(EV_{k,1}^i,\ EV_{k,2}^i\right)^T,\ i,j = 1, 2, \ldots, M,\ k = \lambda_1, \lambda_2. \qquad (4)$$

The angle between the two vectors is denoted as follows:

$$\theta_{ij}^k = \arccos\left(\frac{EV_k^i \cdot EV_k^j}{|EV_k^i| \cdot |EV_k^j|}\right),\ i,j = 1, 2, \ldots, M,\ k = \lambda_1, \lambda_2. \qquad (5)$$

The sum of θ_{ij}^k over k for given i,j can be used to reflect the trend information of the eigenvectors involved

$$\theta_{ij} = \theta_{ij}^{\lambda_1} + \theta_{ij}^{\lambda_2},\ i,j = 1, 2, \ldots, M. \qquad (6)$$

Consequently, two sets of parameters are obtained. The first reflects the difference of center positions represented by the Euclidean distance between the geometric centers. The second indicates the difference of the trends of the 2D curves represented by the related eigenvectors. The overall distance D_{ij} between the species i and j is defined by

$$D_{ij} = d_{ij} \times \theta_{ij},\ i,j = 1, 2, \ldots, M. \qquad (7)$$

Accordingly, a real symmetric $M \times M$ matrix D_{ij} is obtained and used to reflect the evolutionary distance between the species i and j. The clustering tree is constructed using the UPGMA method in PHYLIP package (http://evolution.genetics.washington.edu/phylip.html). The final phylogenetic tree is drawn using the DRAWGRAM program in the PHYLIP package. In Figure 4, we present the phylogenetic tree belonging to 24 species.

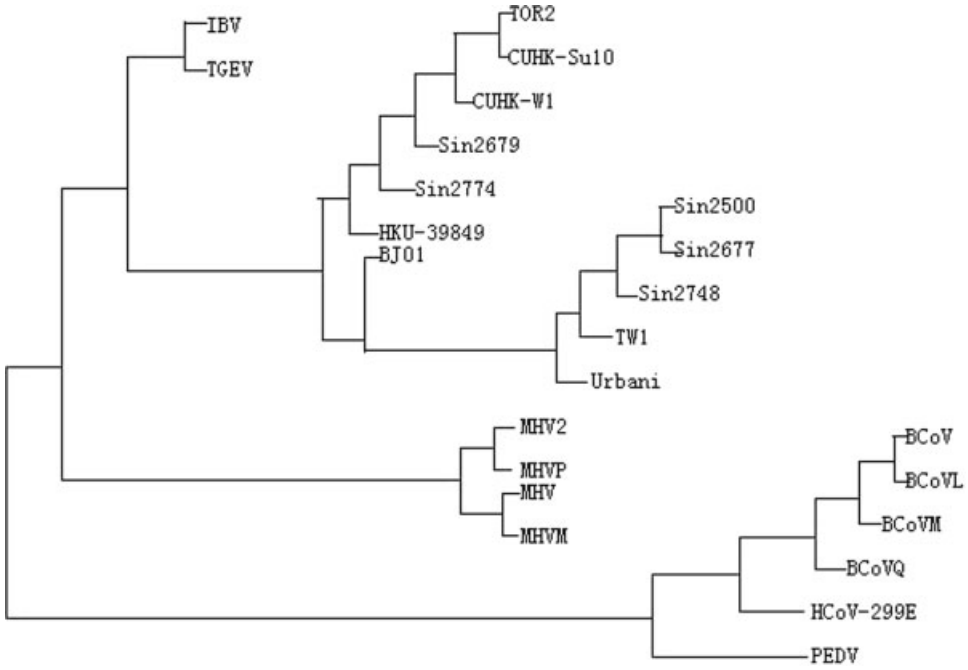

Figure 4. Phylogenetic tree.

Conclusion

Most existing approaches for phylogenetic inference use multiple alignment of sequences and assume some sort of an evolutionary model. The multiple alignment strategy does not work for all types of data, for example, whole genome phylogeny, and the evolutionary models may not always be correct. Our representation provides a direct plotting method to denote DNA sequences without degeneracy. From the DNA graph, the A, T, G, and C usage as well as the original DNA sequence can be recaptured mathematically without loss of textual information. The current 2D graphical representation of DNA sequences provides different approaches for constructing the phylogenetic tree. Unlike most existing phylogeny construction methods, the proposed method does not require multiple alignment. Also, both computational scientists and molecular biologists can use it to analysis DNA sequences efficiently with different parameters of n and m.

Acknowledgments

The authors thank the anonymous referees for many valuable suggestions that have improved this manuscript.

References

1. Lan, Y.-C.; Liu, H.-F.; Shi, Y.-P.; Yang, J.-Y.; Chen, H.-Y.; Arthur Chen, Y.-M. Infect Genet Evol 2005, 5, 261.
2. Palovic–Lazetic, G. M.; Mitic, N. S.; Beljanski, M. V. BMC Bioinformat 2004, 5, 65.
3. Hu, L.-D.; Zheng, G.-Y.; Jiang, H.-S.; Xia, Y.; Zhang, Y.; Kong, X.-Y. Acta Pharmacol Sin 2003, 24, 741.
4. Gao, L.; Qi, J.; Wei, H.; Sun, Y.; Hao, B. Chin Sci Bull 2003, 48, 1170.
5. Rota, P. A.; Oberste, M. S.; et al. Science, 2003, 300, 1394.
6. Gibbs, A. J.; Gibbs, M. J.; Armstrong, J. S. Arch Virol 2004, 149, 621.
7. Jukes, T. H.; Cantor, C. R. Mammalian Protein Metabolism; Academic Press: New York, 1969, p. 21.
8. Kimura, M. J Mol Evol 1980, 16 111.
9. Barry, D.; Hartigan, J. A. Stat Sci 1987, 2, 191.
10. Kishino, H.; Hasegawa, M. J Mol Evol 1989, 29, 170.
11. Lake, J. A. Proc Natl Acad Sci USA 1994, 91, 1455.
12. Camin, J.; Sokal, R. Evolution 1965, 19, 311.
13. Eck, R. V.; Dayhoff, M. O. Atlas of Protein Sequence and Structure; National Biomedical Research Foundation: Silver Spring, MD, 1966, p. 161.
14. Cavalli–Sforza, L. L.; Edwards, A. W. F. Evolution 1967, 21, 550.
15. Fitch, W. M. Syst Zool 1971, 35, 406.
16. Felsenstein, J. Syst Zool 1973, 22, 240.
17. Felsenstein, J. J Mol Evol 1981, 17, 368.
18. Felsenstein, J.; Churchill, G. A. Mol Bio Evol 1996, 13, 93.
19. Liao, B. Chem Phys Lett 2005, 401, 196.
20. Yuan, C.; Liao, B.; Wang, T. Chem Phys Lett 2003, 379, 412.
21. Liao, B.; Wang, T. J Comput Chem 2004, 25, 1364.
22. Liao, B.; Wang, T. J Mol Struct (Theochem), 2004, 681, 209.
23. Liao, B.; Wang, T. Chem Phys Lett 2004, 388, 195.
24. Randic, M.; Vracko, M.; Nandy, A.; Basak, S. C. J Chem Inf Comput Sci 2000, 40, 1235.
25. Randic, M.; Vracko, M.; Lers, N.; Plavsic, D. Chem Phys Lett 2003, 368, 1.
26. Hamori, E.; Ruskin, J. J Biol Chem 1983, 258, 1318.
27. Hamori, E. Nature 1985, 314, 585.
28. Gates, M. A. Nature 1985, 316, 219.
29. Nandy, A. Curr Sci 1994, 66, 309.
30. Nandy, A. Comput Appl Biosci 1996, 12, 55.
31. Liao, B.; Tan, M.; Ding, K. Chem Phys Lett 2005, 402, 380.
32. Liao, B.; Zhang, Y.; Ding, K.; Wang, T. J Mol Struct (Theochem) 2005, 717, 199.
33. Peng, C. K.; Buldyrev, S. V.; Goldberger, A. L. et al. Nature 1992, 356, 168.
34. Yan, S. S.-T.; Wang, J.; Niknejad, A.; Lu, C.; Jin, N.; Ho, Y.-k. Nucleic Acids Res 2003, 31, 3078.
35. Randic, M.; Vracko, M.; Lers, N.; Plavsic, D. Chem Phys Lett 2003, 371, 202.
36. Randic, M.; Vracko, M. J Chem Inf Comput Sci 2000, 40, 599.
37. Nei, M.; Kumar, S. Molecular Evolution and Phylogenetics; Oxford University Press: New york, 2000.
38. Raychaudhury, C.; Nandy, A. J Chem Inf Comput Sci 1999, 39, 243.
39. Hasan, H. O.; Khalid, S. Bioinformatics 2003, 19, 2122.