Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Available online at www.sciencedirect.com
SCIENCE DIRECT
Chemical Physics Letters 421 (2006) 313–318
CHEMICAL PHYSICS LETTERS
www.elsevier.com/locate/cplett

Coronavirus phylogeny based on triplets of nucleic acids bases

Bo Liao *, Yanshu Liu, Renfa Li, Wen Zhu

Laboratory of Embedded Computing and System, School of Computer and Communication, Hunan University, Changsha, Hunan 410082, China

Received 27 December 2005
Available online 20 February 2006

Abstract

We considered the fully overlapping triplets of nucleotide bases and proposed a 2D graphical representation of protein sequences consisting of 20 amino acids and a stop code. Based on this 2D graphical representation, we outlined a new approach to analyze the phylogenetic relationships of coronaviruses by constructing a covariance matrix. The evolutionary distances are obtained through measuring the differences among the two-dimensional curves.

1. Introduction

Compilation of DNA primary sequence data continues unabated and tends to overwhelm us with voluminous outputs that increase daily. Comparison of primary sequences of different DNA strands remains one of the important aspect of the analysis of DNA data banks. Mathematical analysis of the large volume genomic DNA sequence data is one of the challenges for bio-scientists. There are three class methods for the analysis of DNA sequences: (i) Alignment [1,2]. (ii) Matrices: (1) matrices in which an individual entry corresponds to an individual pair of bases [3,6,7] and (2) matrices in which entries summarize information of different X–Y pairs of bases [4,5,7]. (iii) Graphical representation: Graphical representation of DNA sequence provides a simple way of viewing, sorting and comparing various gene structures. 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 which are not as easily obtainable by other methods. In recent years several authors outlined different graphical representation of DNA sequences based on 2D, 3D or 4D [8–20]. Based on these graphical representation, several authors outlined some approaches to make comparison of DNA sequences [21–25].

All this methods are based on the (four letter alphabet, A, C, G, and T standing for nucleotide bases adenine, cytosine, guanine, and thymine, respectively). We will change to consider the fully overlapping triplets of nucleotide bases. Consideration of triplets of nucleotide bases instead of individual nucleotide bases has several reasons and advantages. There are three of them: (i) The genetic code consists of triplets (codons) of DNA (or RNA in some virus) nucleotides. (ii) The second advantage is that one can easily find the open reading frame as the longest sequence of triplets that contains no stop codons when read in a single reading frame. (iii) The computation will become more simple.

In this Letter, we proposed a 2D graphical representation of the protein sequences consisting of 20 amino acids and a stop code. Based on this 2D graphical representation, we outlined a new approach to analyze the phylogenetic relationships of coronaviruses. The evolutionary distances are obtained through measuring the differences among the two-dimensional curves. Unlike most existing phylogeny construction methods [26–31], the proposed method does not require multiple alignment.

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

0009-2614/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.cplett.2006.01.030

2. 2D graphical representation of protein sequences and properties

As is known, all of the 64 triplets of nucleotide bases correspond 20 amino acids and a stop code. There are three reading frame start at position 1, 2 and 3, respectively. Using the translate tool, we can obtain three protein sequences consisting of 20 amino acids and a stop code. The 20 amino acids found in proteins can be grouped according to the chemistry of their R groups as in [32]: amino acids A,V,F,P,M,I,L belong to the hydrophobic chemical group; amino acids D,E,K,R belong to charged chemical group; amino acids S,T,Y,H,C,N,Q,W belong to polar chemical group; amino acid belong to glycine chemical group. Then for any DNA sequence, we will transform it into three new sequences defined over alphabet $\{\overline{H},\overline{C},\overline{P},\overline{G}\}$. The rule is as follows:

$$\phi(g(3i-2,3i-1,3i)) = \begin{cases} \overline{H} & \text{if } g(3i-2,3i-1,3i) = A,V,F,P,M,I,L \\ \overline{C} & \text{if } g(3i-2,3i-1,3i) = D,E,K,R \\ \overline{P} & \text{if } g(3i-2,3i-1,3i) = S,T,Y,H,C,N,G,W \\ \overline{G} & \text{if } g(3i-2,3i-1,3i) = G,- \end{cases}$$

As shown in Fig. 1, we construct a pyrimidine–purine graph on two quadrants of the cartesian coordinate system, with pyrimidines ($\overline{P}$ and $\overline{C}$) in the first quadrant and purines ($\overline{H}$ and $\overline{G}$) in the fourth quadrant. The unit vectors representing four alphabets $\overline{H},\overline{G},\overline{C}$ and $\overline{P}$ are as follows:

$$(m,-\sqrt{n}) \to \overline{H}, (\sqrt{n},m) \to \overline{G}, (\sqrt{n},m) \to \overline{C}, (m,\sqrt{n}) \to \overline{P}$$

where m is a real number and $m \neq \sqrt{n}$, n is a positive real number but not a perfect square number. So that we will reduce a DNA sequence into a series of nodes $P_0,P_1,P_2,\ldots,P_{\lfloor N/3 \rfloor}$, whose coordinates x_i, y_i $(i = 0,1,2,\ldots,\lfloor N/3 \rfloor$, where N is the length of the DNA sequence being studied) satisfy

$$\begin{cases} x_i = \bar{h}_i m + \bar{g}_i\sqrt{n} + \bar{c}_i\sqrt{n} + \bar{p}_i m \\ y_i = -\bar{h}_i\sqrt{n} - \bar{g}_i m + \bar{c}_i m + \bar{p}_i\sqrt{n} \end{cases} \quad (1)$$

$\bar{h}_i, \bar{c}_i, \bar{g}_i$ and $\bar{p}_i$ satisfy

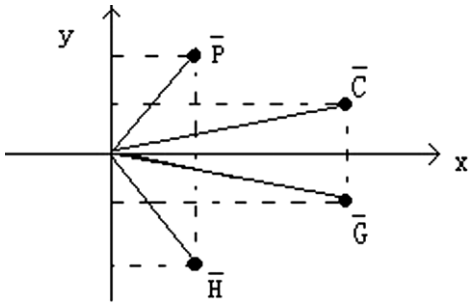

Fig. 1. Pyrimidine–purine graph.

$$\begin{cases} \bar{h}_i = A_i + \sqrt{s_1}V_i + \sqrt{s_2}F_i + \sqrt{s_3}P_i \\ \qquad +\sqrt{s_4}M_i + \sqrt{s_5}I_i + \sqrt{s_6}L_i \\ \bar{c}_i = D_i + \sqrt{s_7}E_i + \sqrt{s_8}K_i + \sqrt{s_9}R_i \\ \bar{g}_i = S_i + \sqrt{s_{10}}T_i + \sqrt{s_{11}}Y_i + \sqrt{s_{12}}H_i \\ \qquad +\sqrt{13}C_i + \sqrt{s_{14}}N_i + \sqrt{s_{15}}Q_i + \sqrt{s_{16}}W_i \\ \bar{p}_i = G_i + \sqrt{s_{17}}\Omega_i \end{cases} \quad (2)$$

where $A_i,V_i,F_i,P_i,M_i,I_i,L_i,D_i,E_i,K_i,R_i,S_i,T_i,Y_i,H_i,C_i$, $N_iQ_iW_i,G_i,\Omega_i$; are the cumulative occurrence numbers of A, $V,F,P,M,I,L,D,E,K,R,S,T,Y,H,C,N,Q,W,G$ and $-$(or stop code), respectively, in the subsequence from the 1st base to the ith base in the sequence. And $s_k, k = 1,\ldots,17$ are positive real number but not perfect square number, $s_i \neq s_j, i,j = 1,\ldots,17$, and $m \neq \sqrt{s_k}, m \neq \sqrt{ns_k}, m\sqrt{s_k} \neq \sqrt{n}$, $1,\ldots,17$. We define $A_0 = V_0 = F_0 = P_0 = M_0 = I_0 = L_0 = D_0 = E_0 = K_0 = R_0 = S_0 = T_0 = Y_0 = H_0 = C_0 = N_0 = Q_0 = W_0 = G_0 = \Omega_0 = 0$.

We called the corresponding plot set be characteristic plot set. The curve connected all plots of the characteristic plot set in turn is called characteristic curve, which is determined by m, n, that satisfy above mentioned condition. In Figs. 2–4, we show the SARS corresponding curves with different parameters n and m, where $s_1 = 2/3; s_2 = 3/4; s_3 = 4/5; s_4 = 5/6; s_5 = 6/7; s_6 = 7/8; s_7 = 8/9; s_8 = 9/10; s_9 = 10/11; \quad s_{10} = 11/12; s_{11} = 12/13; s_{12} = 13/14; s_{13} = 14/15; s_{14} = 15/16; s_{15} = 16/17; s_{16} = 17/18; s_{17} = 18/19$. Observing Figs. 2–4, we find SARS have similar curves despite with different parameters n and m.

Property 1. *For a given DNA sequence there are three 2D representations corresponding to it.*

Proof. Using the translate tool, one can obtain three protein sequences consisting of 20 amino acids and a stop code corresponding three reading frame start at position 1, 2 and 3. In a single reading frame, let (x_i, y_i) be the coordinates of the ith amino acid of protein sequence, then we have

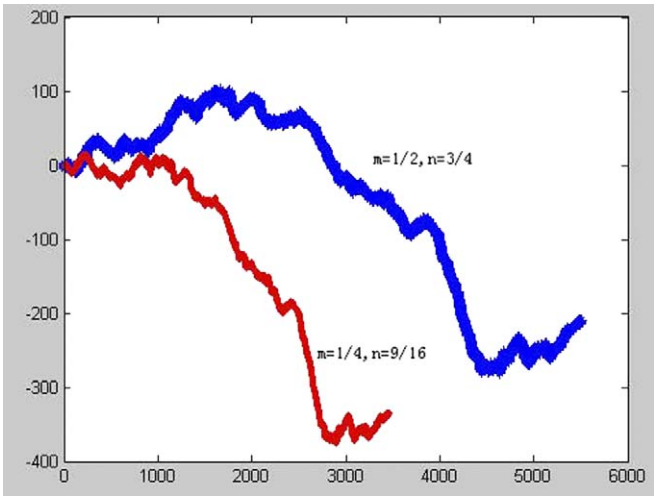

Fig. 2. SARS corresponding curve with different parameters n and m based on the first reading frame.

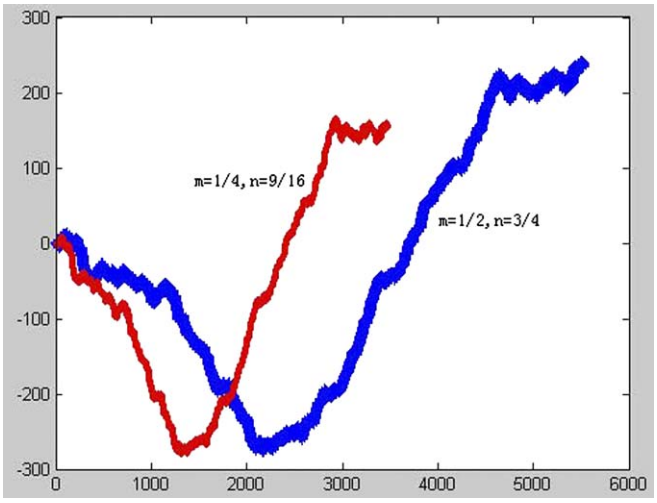

Fig. 3. SARS corresponding curve with different parameters n and m based on the second reading frame.

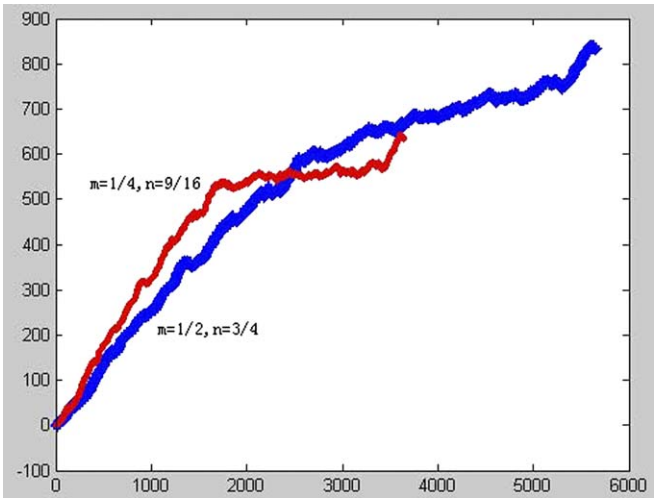

Fig. 4. SARS corresponding curve with different parameters n and m based on the third reading frame.

$$\bar{h}_i(m,-\sqrt{n})+\bar{g}_i(\sqrt{n},-m)+\bar{c}_i(\sqrt{n},m)+\bar{p}_i(m,\sqrt{n})=(x_i,y_i)$$

i.e.,

$$\begin{cases}\bar{h}_i m+\bar{g}_i\sqrt{n}+\bar{c}_i\sqrt{n}+\bar{p}_i m=x_i\\ -\bar{h}_i\sqrt{n}-\bar{g}_i m+\bar{c}_i m+\bar{p}_i\sqrt{n}=y_i\end{cases}\tag{3}$$

□

Obviously, x_i and y_i are irrational numbers of form $sm+k\sqrt{n}$, where s and k are integers. We suppose

$$x_i=s_x m+k_x\sqrt{n}$$
$$y_i=s_y m+k_y\sqrt{n}$$

then we have

$$\begin{cases}\bar{h}_i+\bar{p}_i=s_x\\ \bar{g}_i+\bar{c}_i=k_x\\ -\bar{g}_i+\bar{c}_i=s_y\\ -\bar{h}_i+\bar{p}_i=k_y\end{cases}\tag{4}$$

So, for given x-projection and y-projection of any point $P=(x,\ y)$ on the sequence, after uniquely determining s_x,k_x,s_y,k_y from x and y, the number $A_p,V_p,F_p,P_p,M_p,I_p,L_p,D_p,E_p,K_p,R_p,S_p,T_p,Y_p,H_p,C_p,N_p,Q_p,W_p,G_p,\Omega_p$ of $A,V,F,P,M,I,L,D,E,K,R,S,T,Y,H,C,N,Q,W,G$ and $-$(or stop code) from the beginning of the sequence to the point P can be found by solving linear system (2) and (4).

The vector pointing to the point P_i from the origin O is denoted by r_i. The component of r_i, i.e. x_i and y_i are calculated by Eqs. (1) and (2). Let $\Delta r_i=r_i-r_{i-1}$, then we have Property 2.

Property 2. *For any $i=1,2,\ldots,N'$, where $N'$is the length of protein sequence corresponding the studied DNA sequence, the vector Δr_i has only twenty one possible direction. Furthermore, the length of Δr_i,* i.e.,*$|\Delta r_i|$, is always equal to $s_k(m^2+n)$, for any $i=1,2,\ldots,N$, $k=0,1,\ldots,17,s_0=1$.*

Proof. Actually, the components of Δr_i, i.e., Δx_i and Δy_i can be calculated for each possible residue (A, $V,F,P,M,I,L,D,E,K,R,S,T,Y,H,C,N,Q,W,G$ and $-$) at the ith position of the protein sequence by using Eqs. (1) and (2). For example, when the ith residue is A, we find $\Delta x_i=m$ and $\Delta y_i=-\sqrt{n}$. This result is independent of the conformation state of the $(i-1)$th residue. The two numbers $(m,-\sqrt{n})$ are called the direction of Δr_i. The direction number and the length of Δr_i for each possible residue type at the ith position are summarized. □

Property 3. *There is no circuit or degeneracy in our two-dimensional graphical representation.*

Proof. We assume that: (1) the number of amino acid forming a circuit is l; (2) the number of $A,V,F,P,M,I,L,D,E,K,R,S,\ T,Y,H,C,N,Q,W,G$ and $-$(or stop code) in a circuit is $a',v',f',p',m',i',l',d',\mathrm{e}',k',r',s',t',y',h',c',n',q',w',g'$ and δ', respectively. So $a'+v'+f'+p'+m'+i'+l'+d'+\mathrm{e}'+k'+r'+s'+t'+y'+h'+c'+n'+q'+w'+g'+\delta'=l$. Because $a'A,v'V,f'F,p'P,m'M,i'I,l'L,d'D,e'E,k'K,r'R,s'S,t'T,y'Y,h'H,c'C,\ n'N,q'Q,w'W,g'G$ and δ' $-$(or stop code) form a circuit, the following equation holds:

$$\begin{cases}\bar{h}'=a'+\sqrt{s_1}v'+\sqrt{s_2}f'+\sqrt{s_3}p'+\sqrt{s_4}m'+\sqrt{s_5}i'+\sqrt{s_6}l'\\ \bar{c}'=d'+\sqrt{s_7}e'+\sqrt{s_8}k'+\sqrt{s_9}r'\\ \bar{g}'=s'+\sqrt{s_{10}}t'+\sqrt{s_{11}}y'+\sqrt{s_{12}}h'+\sqrt{13}c'+\sqrt{s_{14}}n'\\ \qquad +\sqrt{s_{15}}q'+\sqrt{s_{16}}w'\\ \bar{p}'=g'+\sqrt{s_{17}}\delta'\end{cases}\tag{5}$$

$$\bar{h}'(m,-\sqrt{n})+\bar{g}'(\sqrt{n},-m)+\bar{c}'(\sqrt{n},m)+\bar{p}'(m,\sqrt{n})=(0,0)$$

i.e.,

$$\begin{cases}\bar{h}'m+\bar{g}'\sqrt{n}+\bar{c}'\sqrt{n}+\bar{p}'m=0\\ -\bar{h}'\sqrt{n}-\bar{g}'m+\bar{c}'m+\bar{t}'\sqrt{n}=0\end{cases}\tag{6}$$

Clearly Eqs. (5) and (6) hold if, and only if $a' = v' = f' = p' = m' = i' = l' = d' = e' = k' = r' = s' = t' = y' = h' = c' = n' = q' = w' = g' = \delta' = 0$. Therefore, $l = 0$, which means no circuit exists in this graphical representation. □

Property 4. *The 2D representation possesses the reflection symmetry.*

Proof. usually the sequence is expressed in the order from 5′ to 3′. Suppose that the 2D representation for protein sequence is described by $(x_i, y_i), i = 0,1,2,\ldots,N$. Suppose again that the 2D representation for the reverse sequence, i.e, the same sequence but from 3′ to 5′ is described by $(\hat{x}_i, \hat{y}_i)$, we find

$$\begin{cases} \hat{x}_i = x_N - x_{N-i} \\ \hat{y}_i = y_N - y_{N-i} \end{cases} \tag{7}$$

□

3. Phylogenetic tree of coronaviruses

For any DNA sequence, we have three translating protein sequences. For any protein 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:

$$x^0 = \frac{1}{N}\sum_{i=1}^{N} x_i, y^0 = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{8}$$

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

$$\begin{cases} CM_{xx} = \frac{1}{N}\sum_{1}^{N}(x_i - x^0)(x_i - x^0) \\ CM_{xy} = \frac{1}{N}\sum_{1}^{N}(x_i - x^0)(y_i - y^0) = CM_{yx} \\ CM_{yy} = \frac{1}{N}\sum_{1}^{N}(y_i - y^0)(y_i - y^0) \end{cases} \tag{9}$$

(See Table 1)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 two-dimensional space. Obviously, the matrix is a real symmetric 2×2 one. There is a leading eigenvalue for a matrix CM. So that there are three geometrical centers and three leading eigenvalue corresponding a DNA sequence. In Table 2, we list the geometrical centers $(x_k^0, y_k^0), k = 1,2,3$ and leading eigenvalues belonging to 24 species with parameter $m = \frac{1}{2}, n = \frac{3}{4}, s_1 = 2/3; s_2 = 3/4; s_3 = 4/5; s_4 = 5/6; s_5 = 6/7; s_6 = 7/8; s_7 = 8/9; s_8 = 9/10; s_9 = 10/11; s_{10} = 11/12; s_{11} = 12/13; s_{12} = 13/14; s_{13} = 14/15; s_{14} = 15/16; s_{15} = 16/17; s_{16} = 17/18; s_{17} = 18/19$ (See Table 3).

In order to facilitate the quantitative comparison of different species in terms of their collective parameters, we introduce a distance scale as defined below. Suppose that there are two species i and j, the parameters are $\lambda_1^i, \lambda_2^i, \lambda_3^i, \lambda_1^j, \lambda_2^i, \lambda_3^j$, respectively, where $\lambda_1^i, \lambda_2^i, \lambda_3^i$ are the three leading eigenvalues of matrix CM_i corresponding to species i. The distance d_{ij} between the two points is

$$d_{ij}\sqrt{(\lambda_1^i - \lambda_1^j)^2 + (\lambda_2^i - \lambda_2^j)^2 + (\lambda_3^i - \lambda_3^j)^2}, i, j = 1,2,\ldots,M \tag{10}$$

Table 1
The accession number, abbreviation, name and length for the 24 coronavirus geneomes

No.	Accession	Abbreviation	Genome	Length (nt)
1	*NC_002645*	*HCoV_229E*	Human coronavirus 229E	27317
2	*NC_002306*	TGEV	Transmissible gastroenteritis virus	28586
3	*NC_003436*	PEDV	Porcine epidemic diarrhea virus	28033
4	U00735	BCoVM	Bovine coronavirus strain Mebus	31032
5	AF391542	BCoVL	Bovine coronavirus isolate BCoV-LUN	31028
6	AF220295	BCoVQ	Bovine coronavirus Quebec	31100
7	*NC_003045*	BCoV	Bovine coronavirus	31028
8	AF208067	MHVM	Murine hepatitis virus strain ML-10	31233
9	AF101929	MHV2	Murine hepatitis virus strain 2	31276
10	AF208066	MHVP	Murine hepatitis virus strain Penn 97-1	31112
11	*NC_001846*	MHV	Murine hepatitis virus	31357
12	*NC_001451*	IBV	Avian infectious bronchitis virus	27608
13	AY278488	BJ01	SARS coronavirus BJ01	29725
14	AY278741	Urbani	SARS coronavirus Urbani	29727
15	AY278491	HKU-39849	SARS coronavirus HKU-39849	29742
16	AY278554	CUHK-W1	SARS coronavirus CUHK-W1	29736
17	AY282752	CUHK-Su10	SARS coronavirus CUHK-SulO	29,736
18	AY283794	SIN2500	SARS coronavirus Sin2500	29711
19	AY283795	SIN2677	SARS coronavirus Sin2677	29705
20	AY283796	SIN2679	SARS coronavirus Sin2679	29711
21	AY283797	SIN2748	SARS coronavirus Sin2748	29706
22	AY283798	SIN2774	SARS coronavirus Sin2774	29711
23	AY291451	TW1	SARS coronavirus TW1	29729
24	*NC_004718*	TOR2	SARS coronavirus	29751

Table 2
Twenty one possible direction

	Δx_n	Δy_n	$\lvert\Delta r_n\rvert$
A	m	$-\sqrt{n}$	m^2+n
D	$\sqrt{n}$	m	m^2+n
S	$\sqrt{n}$	$-m$	m^2+n
G	m	$\sqrt{n}$	m^2+n
V	$m\sqrt{s_1}$	$\sqrt{s_1 n}$	$s_1(m^2+n)$
F	$m\sqrt{s_2}$	$\sqrt{s_2 n}$	$s_2(m^2+n)$
P	$m\sqrt{s_3}$	$\sqrt{s_3 n}$	$s_3(m^2+n)$
M	$m\sqrt{s_4}$	$\sqrt{s_4 n}$	$s_4(m^2+n)$
I	$m\sqrt{s_5}$	$\sqrt{s_5 n}$	$s_5(m^2+n)$
L	$m\sqrt{s_6}$	$\sqrt{s_6 n}$	$s_6(m^2+n)$
E	$\sqrt{ns_7}$	$m\sqrt{s_7}$	$s_7(m^2+n)$
K	$\sqrt{ns_8}$	$m\sqrt{s_8}$	$s_8(m^2+n)$
R	$\sqrt{ns_9}$	$m\sqrt{s_9}$	$s_9(m^2+n)$
T	$\sqrt{ns_{10}}$	$-m\sqrt{s_{10}}$	$s_{10}(m^2+n)$
Y	$\sqrt{ns_{11}}$	$-m\sqrt{s_{11}}$	$s_{11}(m^2+n)$
H	$\sqrt{ns_{12}}$	$-m\sqrt{s_{12}}$	$s_{12}(m^2+n)$
C	$\sqrt{ns_{13}}$	$-m\sqrt{s_{13}}$	$s_{13}(m^2+n)$
N	$\sqrt{ns_{14}}$	$-m\sqrt{s_{14}}$	$s_{14}(m^2+n)$
Q	$\sqrt{ns_{15}}$	$-m\sqrt{s_{15}}$	$s_{15}(m^2+n)$
w	$\sqrt{ns_{16}}$	$-m\sqrt{s_{16}}$	$s_{16}(m^2+n)$
–	$m\sqrt{s_{17}}$	$\sqrt{ns_{17}}$	$s_{17}(m^2+n)$

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}.

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 Fig. 5, we present the phylogenetic tree belonging to 24 species.

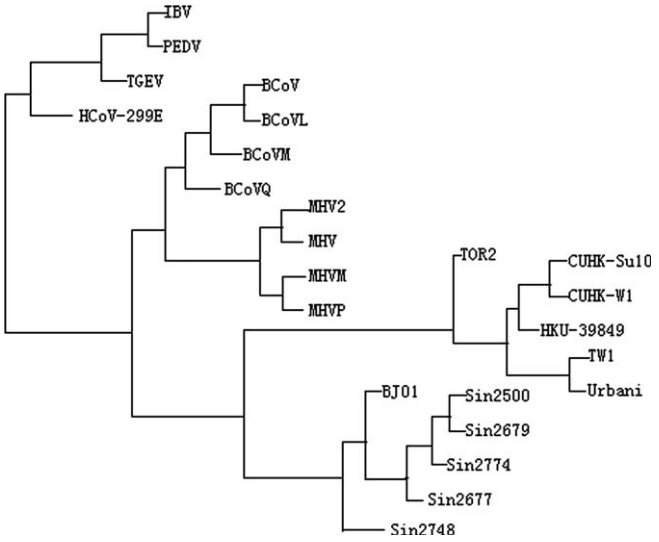

Fig. 5. Phylogenetic tree.

4. Conclusion

We made a analysis of DNA sequences by considering the fully overlapping triplets of nucleotide bases. The presented graphical representation can be recaptured mathe-

Table 3
The geometric centers and three leading eigenvalues for each of the 24 coronavirus genomes

i	x_1^0	y_1^0	x_2^0	y_2^0	x_3^0	y_3^0	λ_1	λ_2	λ_3
1	2.5692e + 003	−159.0439	2.5566e + 003	−342.5873	2.6794e + 003	389.8249	2.1520	2.2321	2.3707
2	2.8619e + 003	−230.4309	2.8245e + 003	−723.2605	2.9971e + 003	128.9913	2.6999	2.8393	2.9157
3	2.8626e + 003	−233.0932	2.8231e + 003	−724.5553	2.9976e + 003	130.5104	2.7034	2.8386	2.9178
4	2.8602e + 003	−245.6989	2.8245e + 003	−743.4898	2.9985e + 003	133.2708	2.7056	2.8453	2.9209
5	2.8688e + 003	−294.6379	2.8364e + 003	−709.6245	3.0012e + 003	146.3851	2.7519	2.8561	2.9158
6	2.6263e + 003	415.1362	2.5204e + 003	−204.5027	2.4666e + 003	−516.9428	2.2817	2.0813	2.1269
7	2.8773e + 003	−476.9658	2.8773e + 003	−476.9658	2.9006e + 003	−252.7994	2.8910	2.8910	2.7932
8	2.8902e + 003	−446.8927	2.8902e + 003	−446.8927	2.9139e + 003	−227.7537	2.9004	2.9004	2.8179
9	2.8853e + 003	−459.6862	3.0344e + 003	82.5446	2.8912e + 003	−273.7115	2.9146	2.9739	2.7829
10	2.8582e + 003	−528.7428	3.0320e + 003	34.9426	2.8807e + 003	−253.2886	2.8697	2.9882	2.7408
11	2.5137e + 003	−415.8854	2.6893e + 003	244.2464	2.5817e + 003	−222.8666	2.2271	2.3287	2.1831
12	2.7670e + 003	−48.3996	2.7276e + 003	−34.7759	2.8570e + 003	524.7574	2.4705	2.5740	2.6849
13	2.7255e + 003	−35.7080	2.8550e + 003	526.4976	2.7646e + 003	−43.8066	2.5698	2.6804	2.4654
14	2.7656e + 003	−45.9837	2.7262e + 003	−35.1151	2.8557e + 003	528.0186	2.4675	2.5711	2.6821
15	2.7659e + 003	−45.2775	2.7260e + 003	−36.4889	2.8558e + 003	530.0127	2.4680	2.5710	2.6828
16	2.7656e + 003	−47.8004	2.7267e + 003	−33.6628	2.8560e + 003	527.4290	2.4680	2.5725	2.6838
17	2.7239e + 003	−35.1426	2.8535e + 003	527.3351	2.7632e + 003	−45.2702	2.5669	2.6777	2.4630
18	2.7233e + 003	−36.1921	2.8529e + 003	527.2583	2.7627e + 003	−45.4289	2.5657	2.6766	2.4620
19	2.7239e + 003	−34.4434	2.8535e + 003	527.8162	2.7633e + 003	−45.2775	2.5667	2.6780	2.4630
20	2.7239e + 003	−35.6707	2.8525e + 003	525.5247	2.7621e + 003	−43.2715	2.5678	2.6737	2.4587
21	2.7241e + 003	−35.5425	2.8535e + 003	527.2287	2.7634e + 003	−45.5734	2.5675	2.6777	2.4636
22	2.7647e + 003	−48.0684	2.7258e + 003	−35.7184	2.8553e + 003	523.7099	2.4661	2.5700	2.6815
23	2.7647e + 003	−47.8421	2.7252e + 003	−35.8263	2.8547e + 003	524.8910	2.4661	2.5692	2.6808
24	2.6110e + 003	−251.1068	2.7585e + 003	459.3175	2.6727e + 003	−97.0235	2.3573	2.4587	2.3322

matically without loss of textual information. And our representation provides a direct plotting method to denote DNA sequences without degeneracy.

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, e.g., whole genome phylogeny, and the evolutionary models may not always be correct. The current two-dimensional graphical representation of DNA sequences provides different approach for constructing 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 protein sequences efficiently. We can obtain some graphical representation of protein sequence based on 2D, 3D and 4D using the following transform: $a_i \to \bar{h}_i, g_i \to \bar{g}_i, c_i \to \bar{c}_i, t_i \to \overline{P}_i$. $\bar{h}_i, \bar{c}_i, \bar{g}_i$ and $\bar{p}_i$ satisfy Eq. (2). 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 1st base to the ith base in the sequence.

Acknowledgments

This work is supported in part by the China Postdoctoral Science Foundation and the National Natural Science Foundation of Hunan University.

References

[1] W.R. Pearson, D.J. Lipman, Proc. Natl. Acad. Sci. USA 85 (1988) 2444.
[2] D. Sankofff, J.B. Kruskal (Eds.), String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison, Addison-Wesley Publ. Co., Reading, MA, 1983, p. 1.
[3] M. Randic, M. Vracko, A. Nandy, S.C. Basak, J. Chem. Inf. Comput. Sci. 40 (2000) 1235.
[4] M. Randic, J. Chem. Inf. Comput. Sci. 40 (2000) 50.
[5] M. Randic, Chem. Phys. Lett. 317 (2000) 29.
[6] M. Randic, M. Vracko, J. Chem. Inf. Comput. Sci. 40 (2000) 599.
[7] M. Randic, S.C. Basak, J. Chem. Inf. Comput. Sci. 41 (2001) 561.
[8] Liao Bo, Chem. Phys. Lett. 401 (2005) 196.
[9] Yuan Chunxin, Liao Bo, Wang Tianming, Chem. Phy. Lett. 379 (2003) 412.
[10] Liao Bo, Wang Tianming, J. Comput. Chem. 25 (11) (2004) 1364.
[11] Liao Bo, Wang Tianming, J. Mol. Struct. THEOCHEM 681 (2004) 209.
[12] S.-T. Yan Stephn, Wang JiaSong, Niknejad Air, Lu Chaoxiao, Jin Ning, Ho Yee-kin, Nucl. Acid Res. 31 (12) (2003) 3078.
[13] M. Randic, M. Vracko, A. Nandy, S.C. Basak, J. Chem. Inf. Comput. Sci 40 (2000) 1235.
[14] Randic Milan, Vracko Majan, Lers Nella, Plavsic Dejan, Chem. Phys. Lett. 368 (2003) 1.
[15] E. Hamori, J. Ruskin, J. Biol. Chem. 258 (1983) 1318.
[16] E. Hamori, Nature 314 (1985) 585.
[17] M.A. Gates, Nature 316 (1985) 219.
[18] A. Nandy, Curr. Sci. 66 (1994) 309.
[19] A. Nandy, Comput. Appl. Biosci. 12 (1996) 55.
[20] Liao Bo, Tan Mingshu, Ding Kequan, Chem. Phy. Lett. 402 (2005) 380.
[21] Liao Bo, Wang Tianming, Chem. Phys. Lett. 388 (2004) 195.
[22] Liao Bo, Zhang Yusen, Ding Kequan, Wang Tianming, J. Mol. Struct.: THEOCHEM 717 (2005) 199.
[23] M. Randic, M. Vracko, N. Lers, D. Plavsic, Chem. Phys. Lett. 371 (2003) 202.
[24] Liaoa Bo, Tan Mingshu, Ding Kequan, Chem. Phys. Lett. 414 (2005) 296.
[25] Liaoa Bo, Ding Kequan, J. Comput. Chem. 14 (26) (2005) 1519.
[26] T.H. Jukes, C.R. Cantor, Mammalian Protein Metabolism, Academic Press, New York, 1969, 21-132.
[27] M. Kimura, J. Mol. Evol. 16 (1980) 111.
[28] D. Barry, J.A. Hartigan, Stat. Sci. 2 (1987) 191.
[29] H. Kishino, M. Hasegawa, J. Mol. Evol. 29 (1989) 170.
[30] J.A. Lake, Proc. Natl Acad. Sci. USA 91 (1994) 1455.
[31] Nei Masatoshi, Kumar Sudhir, Molecular Evolution Phylogeny, Oxford University Press, 2000.
[32] D.W. Mount, Bioinformatics: Sequence and Genome Analysis, Cold Spring Harbor Laboratory Press, 2001.