DNA Research **17**, 155–168, (2010)
Advance Access Publication: 1 April 2010
doi:10.1093/dnares/dsq008

A Novel Construction of Genome Space with Biological Geometry

Chenglong Yu[1], Qian Liang[2], Changchuan Yin[2], Rong L. He[3], and Stephen S.-T. Yau[2,*]

Department of Mathematics, The Chinese University of Hong Kong, Shatin, Hong Kong[1]; Department of Mathematics, Statistics and Computer Science, University of Illinois at Chicago, Chicago, IL 60607-7045, USA[2] and Department of Biological Sciences, Chicago State University, Chicago, IL 60628, USA[3]

*To whom correspondence should be addressed. Tel./Fax. +1 312-996-3065. E-mail: yau@uic.edu

Edited by Hiroyuki Toh
(Received 11 December 2009; accepted 2 March 2010)

Abstract

A genome space is a moduli space of genomes. In this space, each point corresponds to a genome. The natural distance between two points in the genome space reflects the biological distance between these two genomes. Currently, there is no method to represent genomes by a point in a space without losing biological information. Here, we propose a new graphical representation for DNA sequences. The breakthrough of the subject is that we can construct the moment vectors from DNA sequences using this new graphical method and prove that the correspondence between moment vectors and DNA sequences is one-to-one. Using these moment vectors, we have constructed a novel genome space as a subspace in R^N. It allows us to show that the SARS-CoV is most closely related to a coronavirus from the palm civet not from a bird as initially suspected, and the newly discovered human coronavirus HCoV-HKU1 is more closely related to SARS than to any other known member of group 2 coronavirus. Furthermore, we reconstructed the phylogenetic tree for 34 lentiviruses (including human immunodeficiency virus) based on their whole genome sequences. Our genome space will provide a new powerful tool for analyzing the classification of genomes and their phylogenetic relationships.

Key words: genome space; graphical representation; moment vector; classification; phylogeny

1. Introduction

Comparative genomics at the sequence level has existed for ~20 years, since the time genome sequencing started in earnest. Already there have been many proposals to compare genomes. Boore and Brown[1] used gene order to study the evolutionary relationships of metazoan mitochondrial genomes. Snel *et al.*[2] later constructed the phylogenetic tree for completely sequenced prokaryotic genomes based on gene content. In Snel *et al.*'s method, the similarity between two genomes is defined as the number of genes that they have in common divided by their total number of genes. However, such techniques are time-consuming as they must first identify all the genes in one genome. These approaches, together with G + C content, edit distance, and reversal and rearrangement distances,[3–5] compare genomes using only partial genomic information. Thus, these results are usually controversial because single-gene sequences generally do not contain enough information to construct an evolutionary history of organisms.

In Koonin's editorial[6] about the emerging paradigm and open problems in comparative genomics, he pointed out '... but within these superfamilies, there is nothing like a straight one-to-one correspondence between genomes, and in distant genomes, most of the members may not be orthologous'. This sentence enlightened us to get a novel idea of comparing genomes. We can construct a genome space. In this space, each point corresponds to a genome uniquely. The natural distance between two points in the genome space reflects the biological distance between these two genomes. Currently, there is no method to represent genomes by a point in a space without losing biological information.

We introduce graphical representation of DNA sequence to construct the genome space. The graphical representation of DNA sequence provides a simple way of viewing, sorting, and comparing various gene structures. Thus, it is an attractive and promising research direction. The first important method in this direction is due to Hamori.[7] He used a three-dimensional curve to represent a DNA sequence. Gates[8] later constructed a two-dimensional graphical representation that is simpler than the Hamori curve. However, Gates' graphical representation has high degeneracy. Recently, we reported a new two-dimensional graphical representation of gene sequences[9] which has no circuit or degeneracy, so that the correspondence between gene sequences and gene graphs is one-to-one. In this way, the original DNA sequence can be recovered from its graph mathematically without loss of biological information. In this paper, we make a minor modification of our previous method and obtain a new graphical representation approach for DNA sequences. The breakthrough of the subject is that we can construct the moment vectors from DNA sequences using this new graphical method, and we can prove that the correspondence between moment vectors and DNA sequences is one-to-one. The novelty and uniqueness of our approach is that by using these moment vectors of DNA sequences, we have constructed a genome space as a subspace in Euclidean space. Each genome sequence can be represented as a point in this space. Therefore, this genome space can be used to make comparative analysis to study the clustering and phylogenetic relationship among genomes. The biological (evolutionary) distance between two genomes can be obtained through the Euclidean distance among the corresponding points in the genome space.

2. Materials and methods

2.1. Graphical representation of DNA sequence

We constructed a new DNA sequence graph in two quadrants of the Cartesian coordinate system, with pyrimidines (C and T) in the first quadrant and purines (A and G) in the fourth quadrant. As shown in Fig. 1, the vectors corresponding to the four nucleotides G, A, T, and C are as follows:

$$\left(1,-\frac{2}{3}\right)\rightarrow \mathrm{G},\quad \left(1,-\frac{1}{3}\right)\rightarrow \mathrm{A},\quad \left(1,\frac{1}{3}\right)\rightarrow \mathrm{T},\quad \left(1,\frac{2}{3}\right)\rightarrow \mathrm{C}.$$

Here, we should emphasize that the specific ordering of the four nucleotides in the Cartesian coordinate system is related to the GC content of genomes (as it is explained in the 'Discussion' section). Points in the graphical representation are obtained by the sum of vectors representing nucleotides in the sequence. In Fig. 2, we give the graphical representation of whole mitochondrial genome sequences of human, common chimpanzee, Norway rat, and hedgehog, which are based on the vector system shown in Fig. 1. Since human and chimpanzee belong to the same order (primates), their mitochondrial genome graphical representations are very similar visually. These four DNA sequence graphical curves have no circuits or degeneracy and the correspondence between the sequence and the graphical curve can be mathematically proved to be one-to-one.[9]

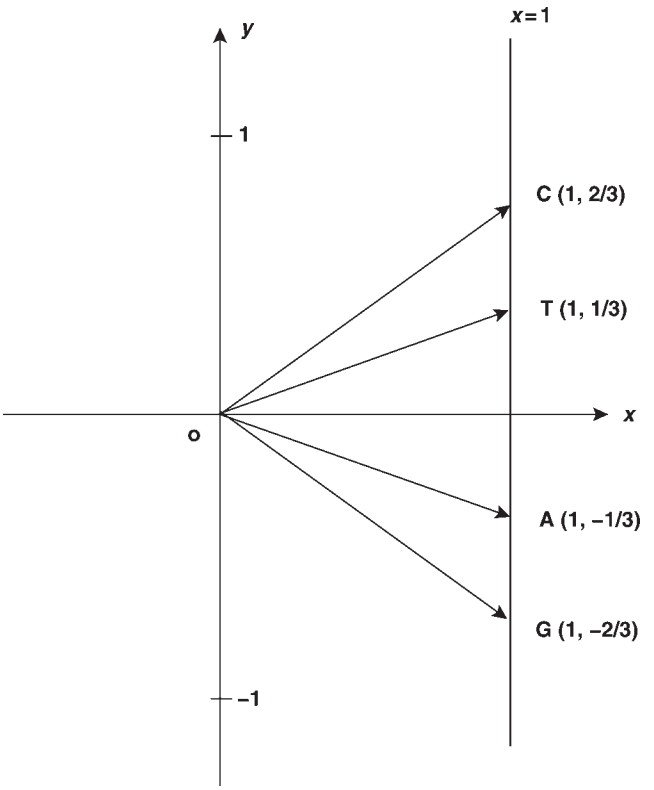

Figure 1. Nucleotide vector system based on G(1, −2/3), A(1, −1/3), T(1, 1/3), and C(1, 2/3).

2.2. Moment vector of DNA sequence

Motivated by our previous work for protein sequence analysis,[10] we use moment vectors to characterize a DNA graphical curve. Given the graphical curve of a DNA sequence, which can be represented by a sequence of points $(1, y_1)$, $(2, y_2), \ldots,$ (n, y_n), we can compute a sequence of numbers $1 - y_1,\ 2 - y_2, \ldots,\ n - y_n$. Conversely, if we know the sequence of numbers $1 - y_1,\ 2 - y_2, \ldots,\ n - y_n$, we can recover the graph $(1,\ y_1),\ (2,\ y_2), \ldots,\ (n,\ y_n)$. Therefore, we want to find a sequence of numbers, each of which uses the global information of the sequence of numbers $1 - y_1,\ 2 - y_2, \ldots,\ n - y_n$ in

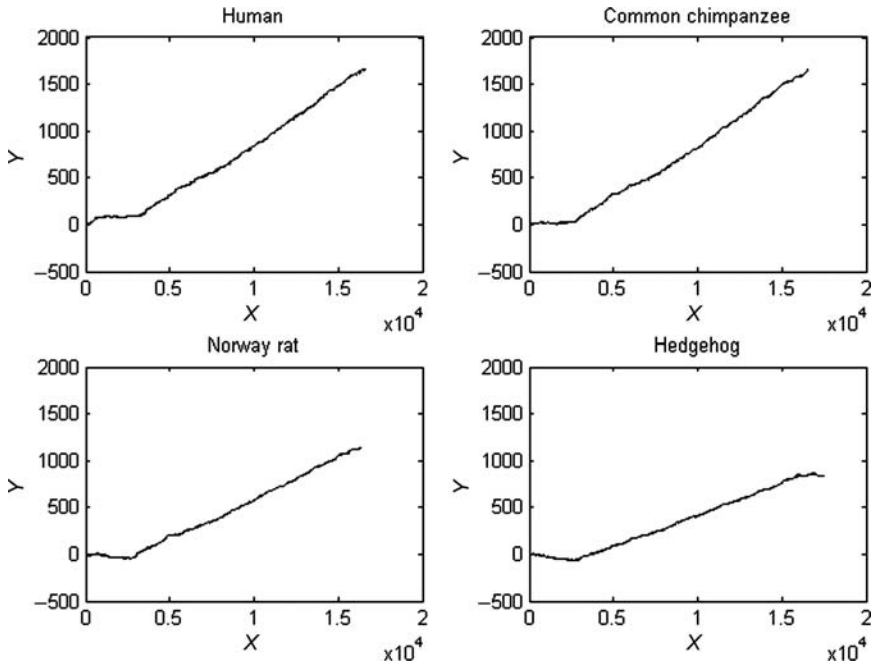

Figure 2. Graphical representation of DNA sequences. Graphical representations of whole mitochondrial genome sequences of four species (human, common chimpanzee, Norway rat, and hedgehog) based on the vector system shown in Fig. 1. *X*-value stands for the number of nucleotides in the DNA sequence. *Y*-value is the cumulative *y*-values of nucleotides G, A, T, and C in Fig. 1. Because the genome lengths of these four species are similar (around 17 000 nt), the *X*-values of the end points of their graphical curves are very similar. For the Y-values of the end points, human and common chimpanzee are very close (more than 1600), but mouse is below 1500 and hedgehog is below 1000.

such a way that this new sequence of numbers determines and is determined by the sequence of numbers $1 - y_1, 2 - y_2, \ldots, n - y_n$. For this purpose, we consider the moments which are defined as follows:

$$M_j = \sum_{i=1}^{n} \frac{(x_i - y_i)^j}{n^j}, \quad j = 1, 2, \ldots, n,$$

where n is the number of nucleotides contained in a DNA sequence, and (x_i, y_i) represents the position of the ith nucleotide in the DNA graphical curve. According to this definition, each DNA sequence has an n-dimensional moment vector $(M_1, M_2, \ldots, M_n)$ associated with it.

The crucial point in this paper is that the correspondence between a DNA sequence and its moment vector obtained from its sequence graph is one-to-one. To obtain this conclusion, we need to prove the following theorem.

Theorem

Consider the set of DNA sequences having the same number (n) of nucleotides. Then, the correspondence between a DNA sequence and its n-dimensional moment vector $(M_1, M_2, \ldots, M_n)$ is one-to-one.

Proof

We have demonstrated that the correspondence between a DNA sequence and its graphical curve is one-to-one.[9] In order to prove the theorem, we will need to prove that the correspondence between a DNA graphical curve and its moment vector is one-to-one.

By the definition, one DNA sequence graph has an n-dimensional moment vector $(M_1, M_2, \ldots, M_n)$. Hence, we need to demonstrate that from any given DNA moment vector, we can recover the DNA curve, which means all (x_i, y_i) $(i = 1, 2, \ldots, n)$ can be recovered from any given DNA moment vector.

x_i is the x-coordinate value of ith nucleotide on a DNA graph. Based on our assignment, x_i should be equal to i. y_i is the y-coordinate value of ith nucleotide on a DNA graph. The next step is to obtain y_i from moment vector. Let $z_i = x_i - y_i$, then the moments can be simplified as:

$$M_j = \sum_{i=1}^{n} \frac{z_i^j}{n^j}, \quad j = 1, 2, \ldots, n.$$

To solve for z_i, let $\delta_j = M_j n^j$, then the δ_j can be obtained by M_j and n. Clearly, δ_j and z_i have the below relation:

$$\begin{aligned} \delta_1 &= z_1 + z_2 + \cdots + z_n \\ \delta_2 &= z_1^2 + z_2^2 + \cdots + z_n^2 \\ &\vdots \\ \delta_n &= z_1^n + z_2^n + \cdots + z_n^n. \end{aligned}$$

These $z_1, z_2, \ldots, z_n$ are roots of the polynomial $(z-z_1)(z-z_2)\cdots(z-z_{n-1})(z-z_n) = z^n + a_{n-1} z^{n-1} + \cdots + a_1 z + a_0$.

Let s_d $(d = 1, 2, \ldots, n)$ be the elementary symmetric polynomials in $z_1, z_2, \ldots, z_n$, i.e. $s_1 = \sum_1^n z_i, s_2 = \sum_{i<j} z_i z_j$, $s_3 = \sum_{i<j<k} z_i z_j z_k, \ldots, s_n = z_1 z_2 \cdots z_n$, then $s_1 = -a_{n-1}$, $s_2 = a_{n-2}, \ldots, s_n = (-1)^n a_0$. By using Newton's famous identities[11]: $\delta_d - s_1 \delta_{d-1} + \cdots + (-1)^{d-1} s_{d-1}\delta_1 + (-1)^d d s_d = 0$, where $d = 1, 2, \ldots, n$, a_i can be obtained by δ_j as shown below:

$$a_{n-1} = (-1)\delta_1$$
$$a_{n-2} = \frac{1}{2}(\delta_1^2 - \delta_2)$$
$$a_{n-3} = (-1)^3 \frac{1}{6}(\delta_1^3 - 3\delta_1\delta_2 + 2\delta_3)$$
$$a_{n-4} = \frac{1}{24}(\delta_1^4 - 6\delta_1^2\delta_2 + 3\delta_2^2 + 8\delta_1\delta_3 - 6\delta_4)$$
$$\vdots$$

As a result, the coefficients of the polynomial $(z-z_1)(z-z_2)\cdots(z-z_{n-1})(z-z_n) = z^n + a_{n-1} z^{n-1} + \cdots + a_1 z + a_0$. can be confirmed, and the set of all roots can be obtained. Next, we need to identify each root $z_1, z_2, \ldots, z_n$.

Because the y-coordinate values of four nucleotides are between -1 and 1, $x_i - y_i \geq 0$. As we have defined that the position of kth nucleotide on a graph is (x_k, y_k) or (k, y_k), the position of $(k+1)$th nucleotide on a graph (x_{k+1}, y_{k+1}) can be represented as $(k+1, y_k + u_{k+1})$, where u_{k+1} may be any of y-coordinate value of the four nucleotides. Thus, $z_{k+1} = x_{k+1} - y_{k+1} = (k+1) - (y_k + u_{k+1}) = (k - y_k) + (1 - u_{k+1}) \geq (k - y_k)$. Because $z_k = x_k - y_k = k - y_k$, $z_{k+1} \geq z_k$. As a consequence, z_i is increasing and each root can be identified by this property, which means each value of y_i can be obtained. With all (x_i, y_i), a DNA graph can be recovered.

Therefore, we have successfully proved that the correspondence between a DNA sequence and its moment vector obtained from its sequence graph is one-to-one.

2.3. *Construction of genome space*

We have already obtained a good numerical characterization, moment vector, to represent a DNA sequence. Now, we will use this tool to construct a genome space. Here, we emphasize that the structure of genomes is complicated. It may be single-stranded or double-stranded and in a linear or circular structure. Thus, we should consider the different structures when constructing the genome space.

For the simplest genome structures, linear single-strand forms, we can treat them as linear DNA sequences. That is, every genome corresponds to a general DNA sequence. Thus, we can utilize our moment vector to construct the genome space. In order to use whole genome information to make comparative analysis among genomes, we can use the first N components $(M_1, M_2, \ldots, M_N)$ of the moment vector of a genome sequence graph to represent a genome as a point in N-dimensional space. Thus, we obtain an N-dimensional genome space as a subspace in R^N. Using the Euclidean distance between two points as an index for comparison, we can perform phylogenetic and clustering analysis for genome sequences in this genome space.

For the circular single-strand genomes, the construction of genome space is more complicated because we do not know which point is the start point in this circular DNA sequence. In this case, we treat every point as the start point in this circular sequence of length n, and then we get n linear single-strand genomes. For every linear single-strand genome sequence, we can compute its n-dimensional moment vector. Then, we take average by n for these n n-dimensional moment vectors to get a normalized moment vector $(M_1, M_2, \ldots, M_n)$. For circular single-strand genomes, we use the first N components $(M_1, M_2, \ldots, M_N)$ of this normalized moment vector to represent a genome as a point in an N-dimensional space. Thus, we obtain an N-dimensional genome space as a subspace in R^N.

For the double-stranded genomes, we need to point out that the moment vector of reverse complementary sequence is not the same as the original sequence. Generally, when meeting the double-stranded genomes, we treat them as two single-stranded genomes. We use the above method (linear or circular) to get two n-dimensional moment vectors for these two single-stranded sequences, and then take average to get a general moment vector $(M_1, M_2, \ldots, M_n)$. By using the first N components $(M_1, M_2, \ldots, M_N)$ of this general moment vector to represent a genome as a point in N-dimensional space. Thus, we obtain an N-dimensional genome space as a subspace in R^N. Here, we need to point out that the two strands of some genomes (e.g. mitochondrial genomes, some bacterial genomes) are differentiated by their nucleotide content, which are called the heavy strand and the light strand, respectively. The two strands have different masses because one has a higher proportion of heavier nucleic acids and its complement a lower proportion. In this case, we just treat them as the single-stranded (by using the heavy strand) genomes to make the genome space.

3. Results

To verify that the biological distance obtained in this way truly incorporates biological utility, we

apply our new genome space to the phylogenetic analysis of organisms. Most existing methods for phylogenetic inference using biological sequences can be divided into two groups. The algorithms in the first group utilize various distance measures[12–15] which are based on different models of nucleotide substitution or amino acid replacement, and then transform the distance matrix into a tree. In the second group of approaches, instead of building a tree, the tree that can best explain the observed sequences under the evolutionary assumption is found by evaluating of different topologies. This category includes parsimony[16–18] and maximum likelihood methods.[19–21] All these methods require a multiple alignment of the sequences and assume some sort of evolutionary model, which require human intervention. Thus, the results are usually controversial. However, our genome space does not need sequence alignment and any evolutionary model. It is totally automatically generated and avoids computation repetition.

First, we consider the phylogeny of mammals. Mitochondrial DNA is not highly conserved and has a rapid mutation rate, thus it is very useful for studying the evolutionary relationships of organisms.[22] We extracted 35 complete mammalian mitochondrial genome sequences from the GenBank, each of which has length of more than 16 000 nucleotides. Moreover, they have double-stranded and circular structures. As mentioned in the previous section, because we have already known the gene content of both strands of these genomes, we just treat them as the single-stranded (by using the heavy strand) circular genomes. For this case, we treat every point as the start point in this circular sequence of length n, and then we get n linear single-strand genomes. For every linear single-strand genome sequence, by using the nucleotide vector system shown in Fig. 1, we can compute its n-dimensional moment vector. Then, we take average by n for these n n-dimensional moment vectors to get a normalized moment vector $(M_1, M_2, \ldots, M_n)$. Here, we use the first 60 components of the moment vector $(M_1, M_2, \ldots, M_{60})$ to characterize these 35 genome graphical curves and obtained 35 points in 60-dimensional genome space. By computing the Euclidean distances between these points, we got the distance matrix for these 35 organisms. The phylogenetic tree for them shown in Fig. 3 is generated using UPGMA program in the MEGA 4 package.[23] The last 10 mammals are grouped into a cluster because they are primates, and the phylogenetic relationship among them coincides with those found by Raina *et al.*[24] We also found that Norway rat, vole, and squirrel are grouped into a cluster for the reason that they are rodent species.

In order to further illustrate the efficiency of our genome space, we then focus on the origins and evolution of human immunodeficiency virus (HIV). HIV is a lentivirus that can lead to acquired immune deficiency syndrome (AIDS), a condition in humans in which the immune system begins to fail, leading to life-threatening opportunistic infections. To develop the anti-HIV drugs and vaccines, the research into the origins and evolution of this virus has become very important. Rambaut *et al.*[25] reconstructed the phylogenetic tree of the primate lentiviruses including HIV-1, HIV-2, and the simian immunodeficiency viruses (SIVs). It was discovered that the two human viruses are related to different SIVs and therefore have different evolutionary origins. However, in Rambaut *et al.*'s paper, the tree was constructed using the maximum likelihood method on an alignment of the nucleotide sequences of polymerase gene in these lentiviruses. Generally, a single-gene sequence does not possess enough information to construct an evolutionary history of organisms, thus we use our new method to reconstruct the phylogenetic tree based on the whole genome sequences.

The genomes of lentiviruses are single-stranded linear RNA. RNA has the base uracil (U) rather than thymine (T) that is present in DNA. In fact, these RNA genome sequences downloaded from GenBank have already been transformed into DNA sequences (change U by T). Thus, we treat them as linear DNA sequences. Using the nucleotide vector system shown in Fig. 1, the sequence graphs of the genomes of the 33 lentiviruses were obtained. Here, we use the first 12 components of the moment vector to characterize these 33 genome graphical curves, and thus we obtained 33 twelve-dimensional vectors. These 33 vectors can be viewed as 33 points in a 12-dimensional genome space. By computing the Euclidean distance between these points, we reconstructed the phylogenetic tree of these primate lentiviruses (Fig. 4) using UPGMA program in the MEGA 4 package.[23] The figure illustrates that both the HIV-1 and HIV-2 lineages fall within that of the SIVs which are isolated from other primates, thus they represent the independent cross-species transmission events. In agreement with Rambaut *et al.*'s result, our phylogenetic tree shows that HIV-1 group N is a new HIV-1 type between the HIV-1 groups M and O (somewhat closer to the M group). It also indicates that HIV-1 groups M, N, and O are closely related to SIVs from chimpanzees (*Pan troglodytes troglodytes* and *Pan troglodytes schweinfurthii*) because there is a mixing of the HIV-1 and SIV lineages. Moreover, our result suggests that *P. t. troglodytes* (SIV_{cpz2} and SIV_{cpz3} in Fig. 4) is the primary reservoir for HIV-1 group O, which is not clear in Rambaut *et al.*'s evolutionary tree. In Rambaut *et al.*'s evolutionary tree, the position of HIV-1 group O is just between

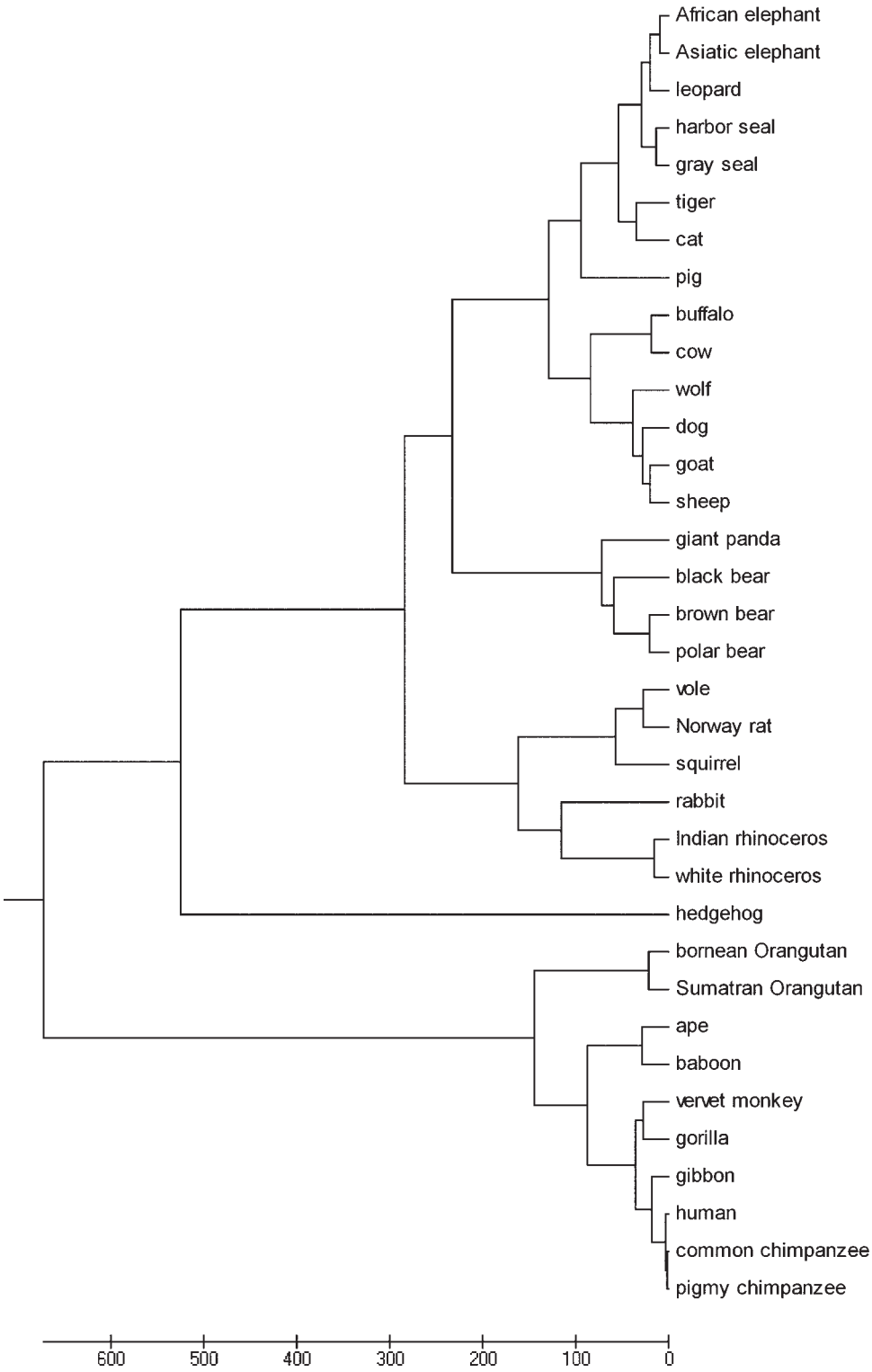

Figure 3. Phylogenetic tree of the mitochondrial genome sequences of 35 mammal species. This tree was reconstructed by the first 60 components of the moment vector ($M_1, M_2, \ldots, M_{60}$). The accession numbers of these 35 species in the GenBank are as follows: human, V00662; pigmy chimpanzee, D38116; common chimpanzee, D38113; gorilla, D38114; gibbon, X99256; baboon, Y18001; vervet monkey, AY863426; ape, NC_002764; Bornean orangutan, D38115; Sumatran orangutan, NC_002083; cat, U20753; dog, U96639; pig, AJ002189; sheep, AF010406; goat, AF533441; cow, V00654; buffalo, AY488491; wolf, EU442884; tiger, EF551003; leopard, EF551002; Indian rhinoceros, X97336; white rhinoceros, Y07726; harbor seal, X63726; gray seal, X72004; African elephant, AJ224821; Asiatic elephant, DQ316068; black bear, DQ402478; brown bear, AF303110; polar bear, AF303111; giant panda, EF212882; rabbit, AJ001588; hedgehog, X88898; Norway rat, X14848; vole, AF348082; squirrel, AJ238588.

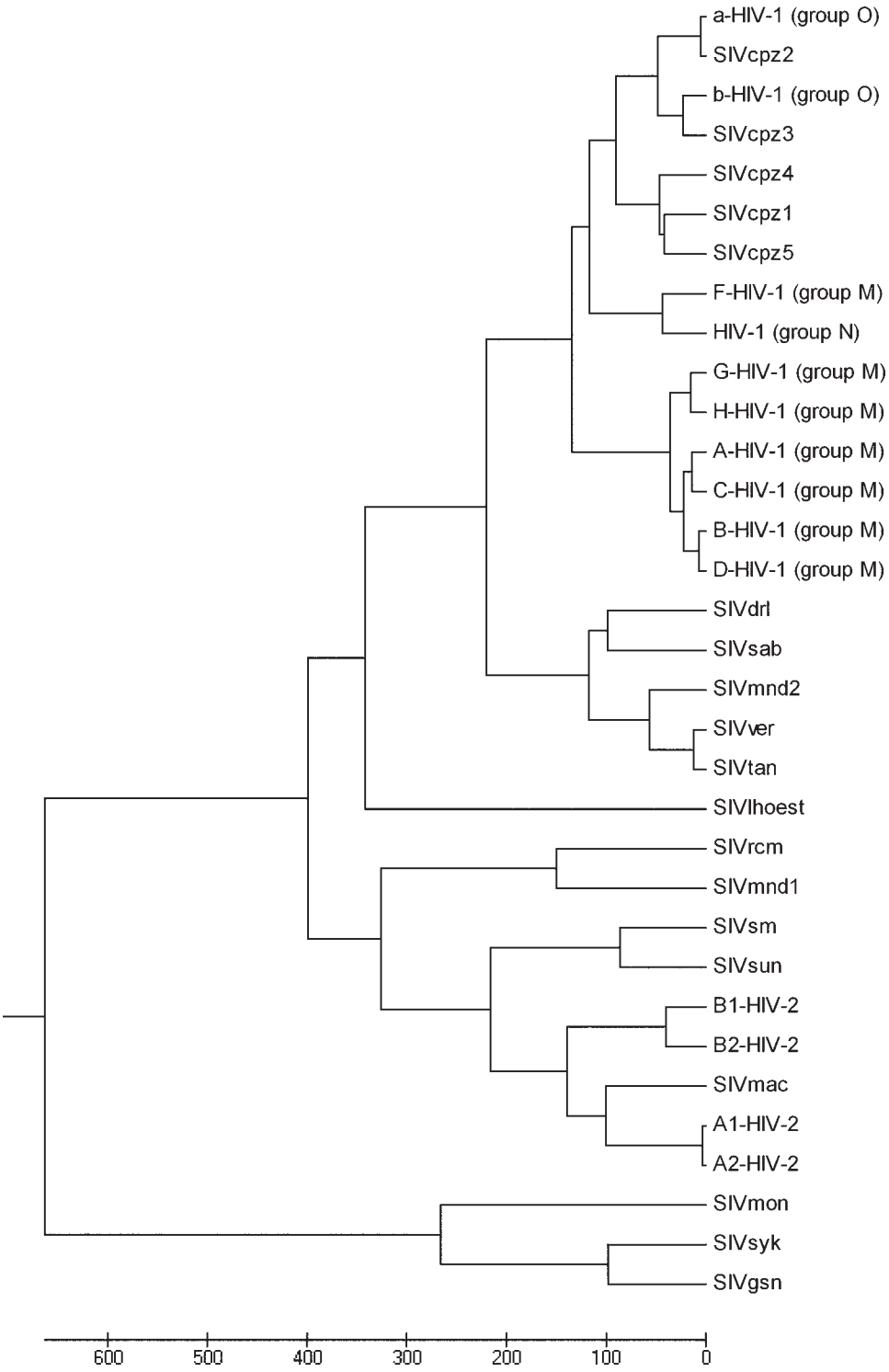

Figure 4. Evolutionary tree of the primate lentiviruses. This tree was reconstructed by using the first 12 components of the moment vector ($M_1, M_2, \ldots, M_{12}$). The subtypes of HIV-1 and HIV-2 are shown. The virus abbreviations, their primate hosts and genome accession numbers in the GenBank are as follows: HIV-1, group M: A, AF004885; B, A04321; C, AF443079; D, K03454; F, AY173957; G, AY772535; H, AF190127; group N: DQ017382; group O: a, AY169802; b, AY169803; HIV-2: A1, AF082339; A2, M30502; B1, L07625; B2, X61240; SIVcpz1-3, chimpanzee (*Pan troglodytes troglodytes*): AY169968, AJ271369, DQ373063; SIVcpz4-5, chimpanzee (*Pan troglodytes schweinfurthii*): DQ374657, DQ374658; SIVdrl, drill: AY159321; SIVgsn, greater spot-nosed monkey: AF468659; SIVlhoest, L'Hoest monkey: AF188114; SIVmac, macaque: D01065; SIVmnd1-2, mandrill: M27470, AY159322; SIVmon, Campbell's mona monkey: AY340701; SIVrcm, red-capped monkey: AF382829; SIVsab, Sabaeus monkey: U04005; SIVsm, sooty mangabey monkey: U72748; SIVsun, sun-tailed monkey: AF131870; SIVsyk, Sykes' monkey: L06042; SIVtan, tantalus monkey: U58991; SIVver, vervet monkey: M29975.

P. t. troglodytes and *P. t. schweinfurthii*, but not closer to either of them. Our suggestion coincides with those found by Gao *et al*.[26] In Gao *et al*.'s work, the authors concluded that the HIV-1 group O is closely related to the just one of these SIVcpz lineages, found in *P. t. troglodytes*. For HIV-2, we find that the SIVs that are closely related to HIV-2 are not only SIVsm (sooty mangabey monkey) and SIVmac (macaque) as shown by Rambaut *et al*.'s tree, but also SIVsun (sun-tailed monkey).

In addition, we apply our genome space to another field of virology: the taxonomy of coronavirus. To study the classification and phylogeny of coronaviruses clearly, we apply our genome space to a large set of 30 complete coronavirus genomes from GenBank, including the two newly sequenced human coronaviruses, HCoV-NL63 and HCoV-HKU1, along with four genomes from *Flaviviridae* and *Togaviridae* which are not coronaviruses (outgroups). The coronavirus genomes are also single-stranded linear RNA. So, similar to the above lentivirus case, we treat these coronavirus genomes as linear DNA sequences. Their abbreviation, accession number, description, and classification are shown in Table 1.

Table 1. Thirty coronavirus genomes and four outgroup genomes from the GenBank

Accession No.	Abbreviation	Length (nt)	Description/Classification
AF304460	HCoV-229E	27 317	Human coronavirus 229E, complete genome/Group 1
AY391777	HCoV-OC43	30 738	Human coronavirus OC43, complete genome/Group 2
AF353511	PEDV	28 033	Porcine epidemic diarrhea virus strain, complete genome/Group 1
U00735	BCoVM	31 032	Bovine coronavirus strain Mebus, complete genome/Group 2
AF391542	BCoVL	31 028	Bovine coronavirus isolate BCoV-LUN, complete genome/Group 2
AF220295	BCoVQ	31 100	Bovine coronavirus strain Quebec, complete genome/Group 2
NC_003045	BCoV	31 028	Bovine coronavirus, complete genome/Group 2
AF208067	MHVM	31 233	Murine hepatitis virus strain ML-10, complete genome/Group 2
AF201929	MHV2	31 276	Murine hepatitis virus strain 2, complete genome/Group 2
AF208066	MHVP	31 112	Murine hepatitis virus strain Penn 97-1, complete genome/Group 2
NC_001846	MHV	31 357	Murine hepatitis virus, complete genome/Group 2
NC_001451	IBV	27 608	Avian infectious bronchitis virus, complete genome/Group 3
EU095850	TCoV	27 657	Turkey coronavirus isolate MG10, complete genome/Group 3
AY278488	BJ01	29 725	SARS coronavirus BJ01, complete genome/Group 4
AY278741	Urbani	29 727	SARS coronavirus Urbani, complete genome/Group 4
AY278491	HKU-39849	29 742	SARS coronavirus HKU-39849, complete genome/Group 4
AY278554	CUHK-W1	29 736	SARS coronavirus CUHK-W1, complete genome/Group 4
AY282752	CUHK-Su10	29 736	SARS coronavirus CUHK-Su10, complete genome/Group 4
AY283794	SIN2500	29 711	SARS coronavirus isolate SIN2500, complete genome/Group 4
AY283795	SIN2677	29 705	SARS coronavirus isolate SIN2677, complete genome/Group 4
AY283796	SIN2679	29 711	SARS coronavirus isolate SIN2679, complete genome/Group 4
AY283797	SIN2748	29 706	SARS coronavirus isolate SIN2748, complete genome/Group 4
AY283798	SIN2774	29 711	SARS coronavirus isolate SIN2774, complete genome/Group 4
AY291451	TW1	29 729	SARS coronavirus TW1, complete genome/Group 4
NC_004718	TOR2	29 751	SARS coronavirus TOR2, complete genome/Group 4
AY297028	ZJ01	29 715	SARS coronavirus ZJ01, complete genome/Group 4
AY572034	Civet007	29 540	SARS coronavirus civet007, complete genome/Group 4
AY572035	Civet010	29 518	SARS coronavirus civet010, complete genome/Group 4
NC_005831	HCoV-NL63	27 553	Human coronavirus NL63, complete genome/Group 1
NC_006577	HCoV-HKU1	29 926	Human coronavirus HKU1, complete genome/Group 5
NC_001564	CellF	10 695	Cell fusing agent virus, complete genome/*Flaviviridae* (outgroup)
NC_004102	HepaCF	9646	Hepatitis C virus, complete genome/*Flaviviridae* (outgroup)
NC_001512	NyongT	11 835	O'nyong-nyong virus, complete genome/*Togaviridae* (outgroup)
NC_001544	RossT	11 657	Ross River virus, complete genome/*Togaviridae* (outgroup)

First, we used our two-dimensional genome space (actually, it is a two-dimensional plane with the first two moments M_1 and M_2 being x-axis and y-axis) to characterize these '34' virus genomes and calculated '34' points in Fig. 5A. Four groups, group 1, group 2, group 3, and outgroups, can be seen in this figure as four distinct clusters. To study the classification for coronavirus clearly, we expanded it and obtained Fig. 5B.

Based on genotypic and serological characterization, coronaviruses are divided into three distinct groups, with human coronavirus 229E (HCoV-229E) being a group 1 coronavirus and human coronavirus OC43 (HCoV-OC43) being a group 2 coronavirus. There is one additional branch, the group 3 coronaviruses, which are found exclusively in birds. These three distinct groups of coronaviruses clearly form three distinct clusters in our two-dimensional

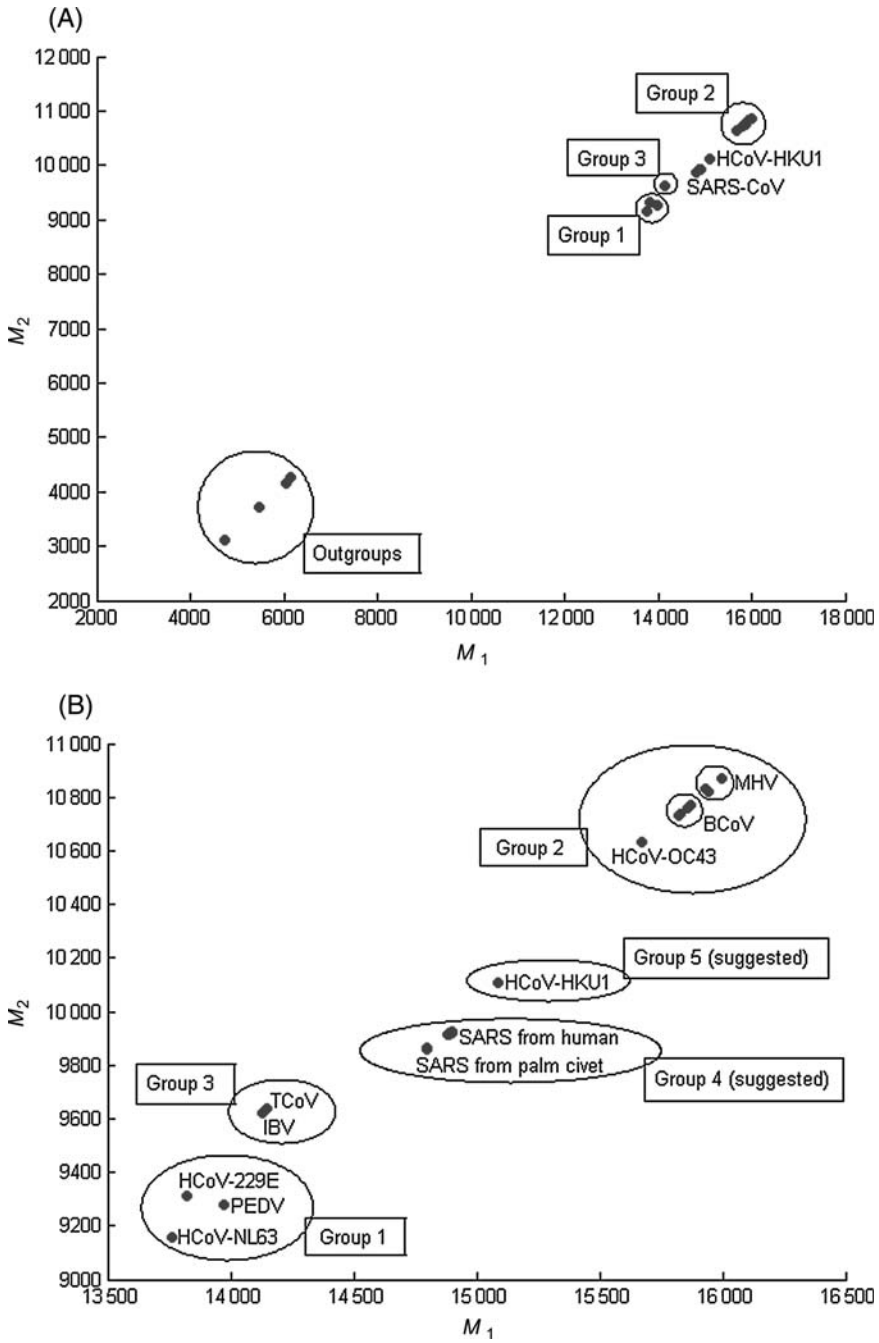

Figure 5. Two-dimensional genome space. (A) We used the first two components of the moment vector (M_1, M_2) to characterize 34 virus genomes from Table 1 and obtained 34 two-dimensional vectors. These 34 vectors can be directly viewed as 34 points in a two-dimensional genome space with M_1 and M_2 being x-axis and y-axis, respectively. (B) Thirty coronavirus genomes from Table 1 are viewed as 30 points in a two-dimensional genome space.

genome space. We also find that the newly discovered human coronavirus NL63 is very close to the human coronavirus 229E, and thus it is classified into group 1. This result is the same as the identification by van der Hoek *et al*.[27] Our two-dimensional genome space reveals that for human SARS-CoV, the most closely related coronavirus is from a small infected mammal, the palm civet (not a bird as initially suspected). This result coincides with those found by Guan *et al*.[28] and Wang *et al*.[29] In Guan *et al*.'s work, the authors also suggested that SARS viruses were isolated from Himalayan palm civets found in a live-animal market in Guangdong, China. Similarly, in Wang *et al*.'s work, the researchers found six palm civets at the restaurant were positive for SARS-associated coronavirus. They think that SARS cases at the restaurant were the result of recent interspecies transfer from the putative palm civet reservoir and not the result of continued circulation of SARS-CoV in the human population.

For another newly discovered human coronavirus HCoV-HKU1, Woo *et al*.[30] classified it into group 2 because HCoV-HKU1 contains certain features that are characteristic of group 2 coronaviruses. However, these authors also stated that the proteins of HCoV-HKU1 formed distinct branches in the phylogenetic trees, indicating that HCoV-HKU1 is a distinct member within the group and is not very closely related to any other known member of group 2 coronaviruses. In our two-dimensional genome space, we found that HCoV-HKU1 is an individual coronavirus between the SARS group (which we suggest as group 4) and the traditional group 2, thus we propose that HCoV-HKU1 belongs to a new group 5. By computing the distances between the genomes in Fig. 5A, we also reconstructed the phylogenetic tree of these coronaviruses (Fig. 6) using UPGMA program in the MEGA 4 package.[23] Thus, the evolutionary relationship of these genomes is clearer.

4. Discussion

It should be pointed out that the construction of our genome space depends on four parameters (the y-coordinates of the A, C, T, and G in Fig. 1). If we change these four parameters, we shall have different embedding of the genome space. Because the GC content of DNA molecule is found to be variable with different organisms, GC content should be considered when we assign the y-coordinate values of nucleotide vectors. The genomes of three groups in this paper have low GC content. Most of them have 30–50% GC content, thus we assign larger y-coordinate absolute values for G and C. However, the y-coordinate values of the four nucleotides must be between -1 and 1 to assure that the correspondence between a DNA sequence and its corresponding moment vector is one-to-one. Therefore, in order to obtain a universal genome space for all species, further studies will be needed to determine the universal y-coordinate values.

In the study of mammalian mitochondrion, lentiviruses (including HIV), and coronavirus genomes, we used 60, 12, and 2 moments to construct the genome space, respectively. Here, we should emphasize that we do not need to calculate all the moments to determine the biological information of genomes. Remember that in the central limit theorem in probability and statistics, the limiting process is Gaussian. For Gaussian, the first two moments determine the density function. Thus, we just use the first N moments to get the results, where N is much less than n (the length of genome). Thus, for coronavirus genomes, we only used the first two components of the moment vector (M_1, M_2) because these two moments have allowed us to obtain the stable classified result—when higher moments are included the relationship of being close or farther away remains unchanged. To make this point clearer, we also use the first 20 components of the moment vector (M_1, $M_2, \ldots,$ M_{20}) to construct the 20-dimmensional genome space. By computing the Euclidean distances between these points in this genome space, we reconstructed the phylogenetic tree of these coronaviruses (Fig. 7). Comparing Figs 6 and 7, we found that the classification relationship of these genomes are the same—group 1, group 2, group 3, group 4, group 5, and outgroups can still be seen in the two trees as six distinct clusters. This means that when using the higher moments, the relationship of being close or farther away remains unchanged. In other word, two moments are already enough to give the right classifying relationship for these genomes. For the same reason, we used the first 60 components of the moment vector (M_1, $M_2, \ldots, M_{60}$) and the first 12 components of the moment vector (M_1, $M_2, \ldots,$ M_{12}) to generate the genome space for mammalian mitochondrion and lentiviruses genomes and obtained the stable phylogenetic analysis result.

Gene rearrangements play important roles in evolution. The order of genes and transcription regions are changed during evolution by gene rearrangements such as DNA inversions and transpositions,[31] which do not affect the gene content of the chromosomes. For example, inversions of large genomic fragments are often observed even between closely related species. The phylogenetic analysis based on the gene order is challenging as it requires detailed complex gene order data in genomes and intensive computation. In order to test whether our genome

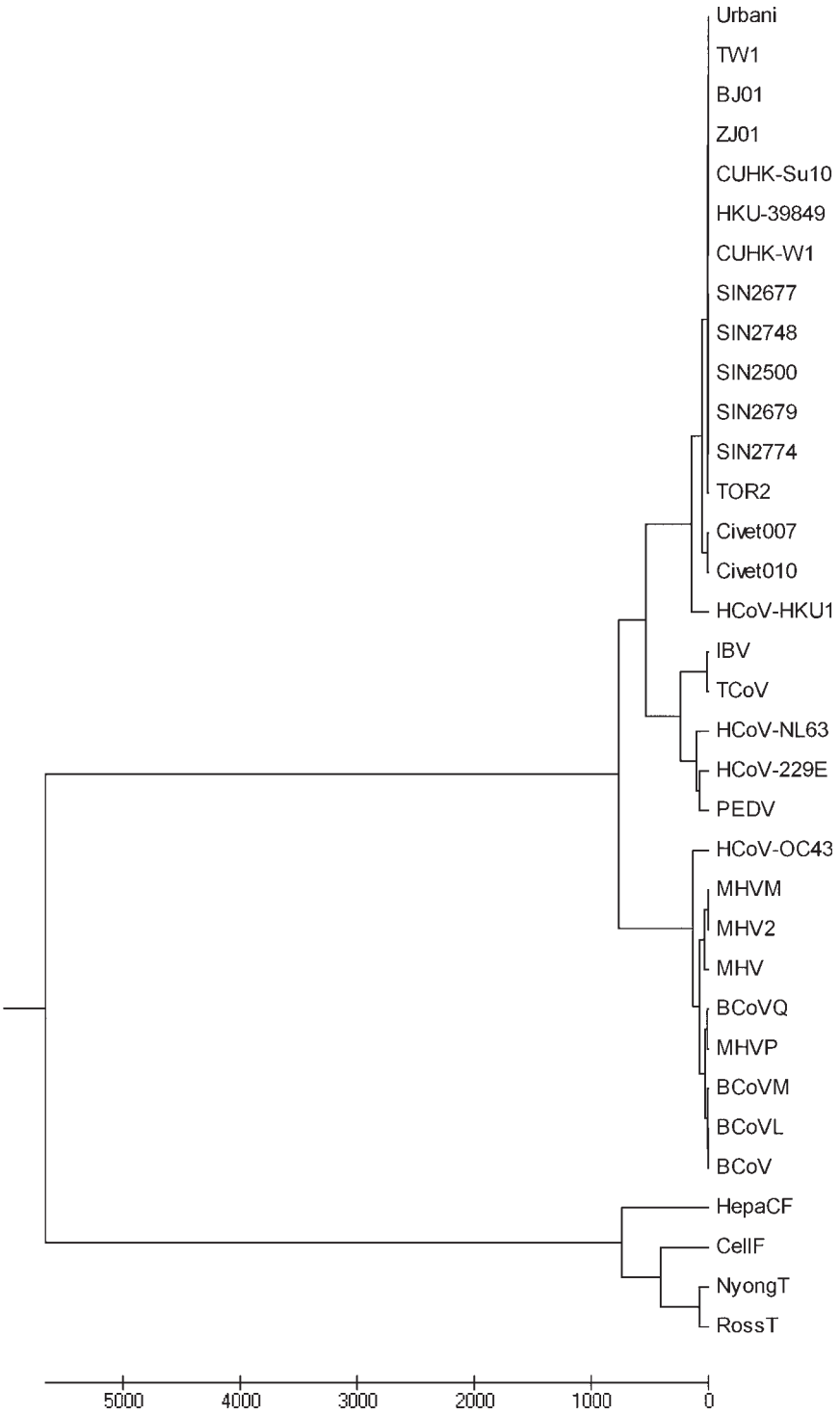

Figure 6. Phylogenetic tree of the coronavirus genomes. This tree was reconstructed by the first two components of the moment vector (M_1, M_2).

space is stable for genomic rearrangement, we consider the following simulated experiment. We choose the human mitochondrion genome and then invert its two genes ATPase 6 and Cytochrome oxidase III to get a simulated genome, which we called human AC genome. We can treat this new genome as the result of the inversion of genes from the human mitochondrion genome. Next, we randomly generate a genome sequence which has the same length and nucleotide content as the human mitochondrion genome, which we called random genome. Thus, we have three genomes of the same length and

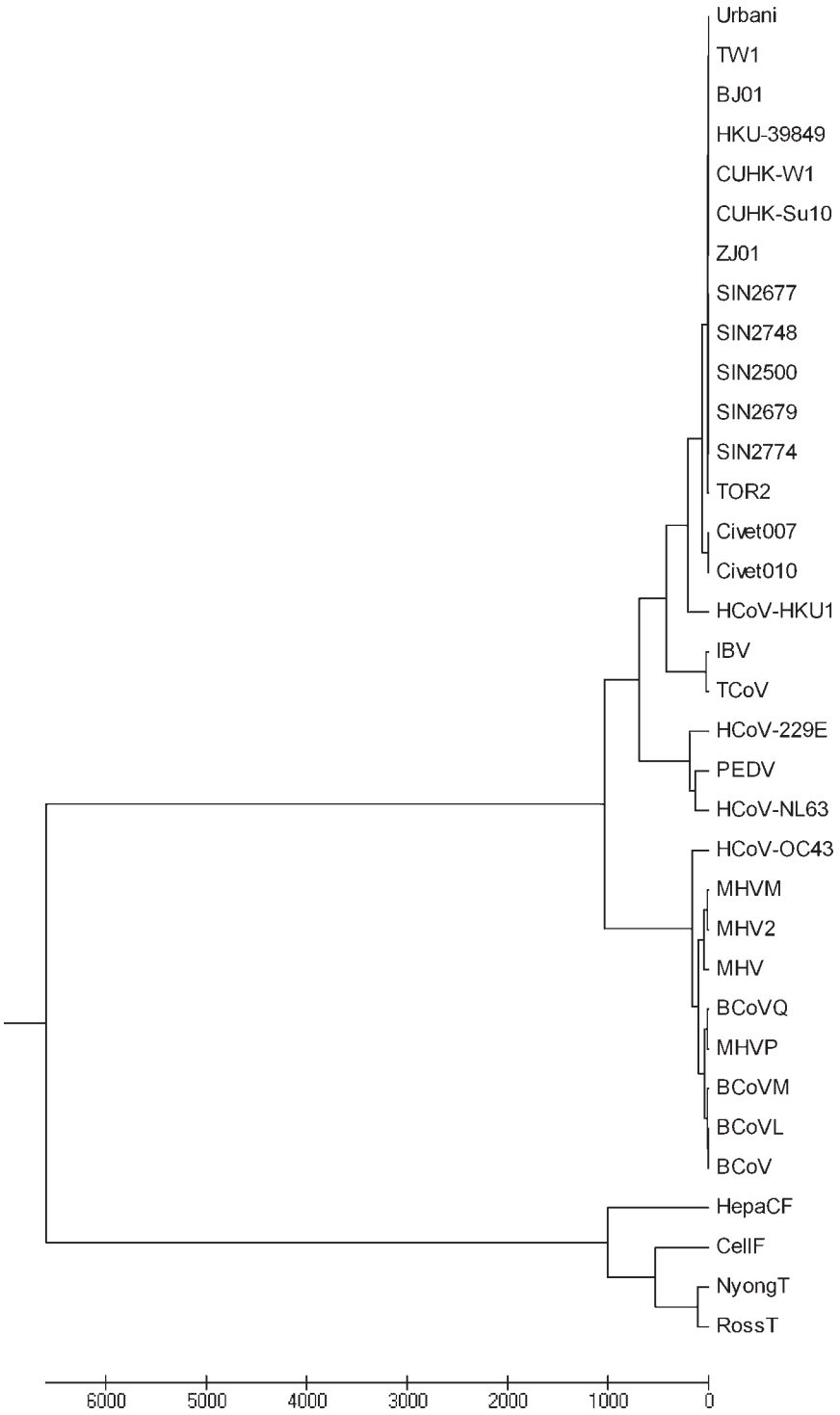

Figure 7. Phylogenetic tree of the coronavirus genomes. This tree was reconstructed by the first 20 components of the moment vector (M_1, $M_2, \ldots, M_{20}$).

nucleotide content: human, human AC, and random. In addition, we also choose the common chimpanzee mitochondrion genome as comparison. By using the 60-dimmensional genome space, we calculate the distance among these four genomes and get a distance matrix in Table 2. In Table 2, we found that the distance between human and human AC is very small. This means that even some genomic rearrangement happened in a genome, the new produced genome still has very close distance from the original genome. Thus, in our genome space, the original genome and its genomic rearrangement genome

Table 2. The distance matrix of four genomes in the 60-dimmensional genome space

	Human	Human AC	Random	Chimpanzee
Human				
Human AC	1.778773			
Random	223.018945	224.777237		
Chimpanzee	20.171072	19.112686	237.476133	

represent different points but have very close distance. Large-scale genomic rearrangements in closely related organisms make the genome sequences very different from the original genome sequences. Although the moment vectors of the original and rearranged DNA sequences are different, if we treat double DNA genomes as two single-stranded genomes to get two n-dimensional moment vectors for these two single-stranded sequences and take the average to get a general moment vector (M_1, $M_2, \ldots, M_n$), the computed distances among related species by gene rearrangements will be small. The genome spaces constructed by the moment vector have advantages over the gene order phylogenetic analysis since the genome spaces do not depend on gene orders and are capable of using all gene families. We will apply the moment vector method on the detailed phylogenetic analysis using the large-scale genomic rearrangement data.

In this paper, we report a two-dimensional graphical representation for DNA sequences. A moment vector system to represent a DNA sequence is introduced, and the correspondence between a DNA sequence and its moment vector is mathematically proven to be one-to-one. With this moment vector system, each genome sequence can be represented as a point in a Euclidean space, and the genome space is constructed as a subspace of this Euclidean space. Genomes with close evolutionary relationship and similar properties plot close together in this genome space. Thus, it will provide a new powerful tool for analyzing the classification of genomes and their phylogenetic relationships. Our method is easier and quicker in handling whole or partial genomes than multiple alignment methods. There are two major advantages to our method. (i) Once a genome space has been constructed, it can be stored in a database. There is no need to reconstruct the genome space for any subsequent application, whereas in multiple alignment methods, realignment is needed for add-on new sequences. (ii) One can have global comparison of all genomes simultaneously, which no other existing method can achieve. Furthermore, in our method, the results in two-dimensional genome space can be displayed and viewed graphically; this is user-friendly and allows even non-expert to understand the relationship among different genomes via viewing the graph of genome space.

Acknowledgements: We gratefully acknowledge Professor Yuen-Ling Chan from University of Chicago, Department of Biochemistry and Molecular Biology who read our paper and gave some constructive comments. We also thank Dr Max Benson for critically reading and editing the manuscript.

References

1. Boore, J.L. and Brown, W.M. 1998, Big tree from little genomes: mitochondrial gene order as a phylogenetic tool, *Curr. Opin. Genet. Dev.*, **8**, 668–74.
2. Snel, B., Bork, P. and Huynen, M.A. 1999, Genome phylogeny based on gene content, *Nat. Genet.*, **21**, 108–10.
3. Kececioglu, J. and Sankoff, D. 1995, Exact and approximation algorithms for the inversion distance, *Algorithmica*, **13**, 180–210.
4. Hannenhalli, S. and Pevzner, P. 1995, Transforming cabbage into turnip. In: *Proceedings of the 27th ACM Symposium on Theory of Computing*, pp. 178–89.
5. Nadeau, J.H. and Sankoff, D. 1998, Counting on comparative maps, *Trends Genet.*, **14**, 495–501.
6. Koonin, E.V. 1999, The emerging paradigm and open problems in comparative genomics, *Bioinformatics*, **15**, 265–6.
7. Hamori, E. 1985, Novel DNA sequence representation, *Nature*, **314**, 585–6.
8. Gates, M.A. 1985, Simpler DNA sequence representations, *Nature*, **316**, 219.
9. Yau, S.S., Wang, J., Niknejad, A., Lu, C., Jin, N. and Ho, Y. 2003, DNA sequence representation without degeneracy, *Nucleic Acids Res.*, **31**, 3078–80.
10. Yau, S.S., Yu, C. and He, R. 2008, A protein map and its application, *DNA Cell Biol.*, **27**, 241–50.
11. Jacobson, N. 1974, *Basic Algebra*, vol. 1. Hindustan Publishing Corporation: India, 135 pp.
12. Jukes, T.H. and Cantor, C.R. 1969, Evolution of protein molecules in mammalian protein metabolism, In: Munro, H.N. and Allison, J.B. (eds.), *Mammalian Protein Metabolism*, Academic Press: New York, pp. 21–132.
13. Kimura, M. 1980, A simple model for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences, *J. Mol. Evol.*, **16**, 111–20.

14. Barry, D. and Hartigan, J.A. 1987, Statistical analysis of ominoid molecular evolution, *Stat. Sci.*, **2**, 191–210.
15. Lake, J.A. 1994, Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances, *Proc. Natl Acad. Sci. USA*, **91**, 1455–9.
16. Camin, J. and Sokal, R. 1965, A method for deducing branching sequences in phylogeny, *Evolution*, **19**, 311–26.
17. Cavalli-Sforza, L.L. and Edwards, A.W.F. 1967, Phylogenetic analysis: models and estimation procedures, *Evolution*, **21**, 550–70.
18. Fitch, W.M. 1971, Toward defining the course of evolution: minimum change for a specific tree topology, *Syst. Zool.*, **35**, 406–16.
19. Felsenstein, J. 1973, Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters, *Syst. Zool.*, **22**, 240–9.
20. Felsenstein, J. 1981, Evolutionary trees from DNA sequences: a maximum likelihood approach, *J. Mol. Evol.*, **17**, 368–76.
21. Felsenstein, J. and Churchill, G.A. 1996, A hidden Markov model approach to variation among sites in rate of evolution, *Mol. Bio. Evol.*, **13**, 93–104.
22. Brown, W.M., Prager, E.M., Wang, A. and Wilson, A.C. 1982, Mitochondrial DNA sequences of primates: tempo and mode of evolution, *J. Mol. Evol.*, **18**, 225–39.
23. Kumar, S., Nei, M., Dudley, J. and Tamura, K. 2008, MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences, *Brief. Bioinform.*, **9**, 299–306.
24. Raina, S.Z., Faith, J.J., Disotell, T.R., Seligmann, H., Stewart, C.B. and Pollock, D.D. 2005, Evolution of base-substitution gradients in primate mitochondrial genomes, *Genome Res.*, **15**, 665–73.
25. Rambaut, A., Posada, D., Crandall, K.A. and Holmes, E.C. 2004, The causes and consequences of HIV evolution, *Nat. Rev. Genet.*, **5**, 52–61.
26. Gao, F., Bailes, E., Robertson, D.L., et al. 1999, Origin of HIV-1 in the chimpanzee *Pan troglodytes troglodytes*, *Nature*, **397**, 436–41.
27. van der Hoek, L., Pyrc, K., Jebbink, M.F., et al. 2004, Identification of a new human coronavirus, *Nat. Med.*, **10**, 368–73.
28. Guan, Y., Zheng, B.J., He, Y.Q., et al. 2003, Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China, *Science*, **302**, 276–8.
29. Wang, M., Yan, M., Xu, H., et al. 2005, SARS-CoV infection in a restaurant from palm civet, *Emerg. Infect. Dis.*, **11**, 1860–5.
30. Woo, P.C., Lau, S.K., Chu, C.M., et al. 2005, Characterization and complete genome sequence of a novel coronavirus, coronavirus HKU1, from patients with pneumonia, *J. Virol.*, **79**, 884–95.
31. Sankoff, D. and Nadeau, H.J. 2003, Chromosome rearrangements in evolution: From gene order to genome sequence and back, *Proc. Natl Acad. Sci. USA*, **100**, 11188–9.