
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.

Journal of Theoretical Biology 271 (2011) 44–50

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

# Genome analysis with the conditional multinomial distribution profile

Guisong Chang [a,b,*], Tianming Wang [a]

[a] School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, PR China
[b] Department of Mathematics, Northeastern University, Shenyang 110004, PR China

ARTICLE INFO

*Article history:*
Received 31 May 2010
Received in revised form
24 November 2010
Accepted 24 November 2010
Available online 1 December 2010

*Keywords:*
Complete multinomial composition vector
Phylogenetic tree
Reference multinomial distribution

ABSTRACT

The focus of the research is on the analysis of genome sequences. Based on the inter-nucleotide distance sequence, we propose the conditional multinomial distribution profile for the complete genomic sequence. These profiles can be used to define a very simple, computationally efficient, alignment-free, distance measure that reflects the evolutionary relationships between genomic sequences. We use this distance measure to classify chromosomes according to species of origin, to build the phylogenetic tree of 24 complete genome sequences of coronaviruses. Our results demonstrate the new method is powerful and efficient.



## 1. Introduction

A great volume of available genomic data has made possible analysis of large sets of organisms at the whole genome scale. However, given that most genomes contain millions to billion nucleotides, traditional molecular analysis methods based on multiple sequence alignment become impractical due to their high computation complexity (Vinga and Almeida, 2003). Consequently, considerable efforts have been made to seek for the alignment-free method for sequence analysis. The first, mainly based on graphical representation of the sequence, is very convenient for studying several selected cases (Liao et al., 2005; Jeffrey, 1990; Nandy, 1994; Nandy and Nandy, 1995, 2003; Randic and Vracko, 2000; Randic et al. 2001, 2003a,b, 2006; Randic and Balaban, 2003; Randic, 2008; Zhang and Zhang, 1994). One of the aim of graphical representation is to identity regions of interest or the distribution of base along the sequence visually (Zhang and Zhang, 1994). The second approach, has been proposed to characterize the DNA sequence (Akhtar et al., 2007). For this purpose one has to find representative descriptors that characterize an abstract mathematical representation of the biological sequence (Dai et al., 2006; He and Wang, 2002). A commonly used numerical characterization of the sequence is to consider binary sequences that describe the position of each nucleotide (Voss, 1992). Different approaches are described in a recent review (Nandy et al., 2006).

Nair and Mahalashmi (2005) proposed the inter-nucleotide distance as a new DNA numerical profile. Any DNA sequence can be converted into a unique numerical sequence with the same length. In the representation, each number represents the distance of a nucleotide to the next occurrence of the same nucleotide. Meanwhile Nair and Mahalashmi (2005) employed discrete Fourier transformation to the inter-nucleotide distance sequence and indicated that this method has a discriminatory capability for highlighting the promoter region of gene sequence. However, Akhtar and Epps (2008) proved that it has poor accuracy in the exon prediction. Afreixo et al. (2009) developed a new method to analyze the inter-nucleotide distance sequence and extracted some interesting features of the DNA sequence. Four nucleotide inter-nucleotide distance distributions and a global distance distribution were given to each genome sequence. In each nucleotide inter-nucleotide distance distribution, only the total number of three other nucleotides was considered (Afreixo et al., 2009). In fact, we can extract more information about the genome sequence from their inter-nucleotide distance sequences.

Motivated by the aforementioned work, we construct four conditional multinomial distributions from four inter-nucleotide distance sequences. In case of the inter-nucleotide distance sequence about nucleotide $A$, the number of the nucleotide $C$, the number of the nucleotide $G$ and the number of the nucleotide $T$ would follow multinomial distribution given that inter-nucleotide distance is $k$. This multinomial distribution will be called the conditional multinomial distribution. The relative error vector derived from the conditional multinomial distribution then can be used as a genomic signature that identifies each species. This approach allows us to perform comparative analysis between complete genome sequences. In fact, we propose a new evolutionary information representation, *complete multinomial composition vector* (*CMCV*), by using a collection of multinomial composition vectors. These multinomial composition vectors are built on the relative error vectors of conditional multinomial distributions with $k$, where $k$ is within a range. The range of $k$ is determined to ensure that the *CMCV* contains the largest amount of evolutionary information hidden in the whole genomic data. We then define the evolutionary

* Corresponding author at: School of Mathematical Sciences, Dalian University of Technology, Dalian, Liaoning 116024, PR China. Tel.: +86 024 83686202; fax: +86 411 84708354.
*E-mail address:* gschang@mail.neu.edu.cn (G. Chang).

0022-5193/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2010.11.034

distance between two genomes based on their complete multinomial composition vectors. The proposed method is tested by phylogenetic analysis on 24 coronavirus genomes. Our results demonstrate that the new method is powerful and efficient.

## 2. Materials and methods

### 2.1. Inter-nucleotide distance sequence

A DNA sequence, of length $n$, can be viewed as a linear sequence of $n$ symbols from a finite alphabet $N=\{A,C,G,T\}$. The inter-nucleotide distance was originally introduced by Nair and Mahalashmi (2005). The global inter-nucleotide distance sequence referred to as *GIN*, was defined as follows:

Given a DNA sequence $S=s_1,s_2,\ldots,s_n$, $GIN(m)=k$, *where* $k=$min value of $i$ such that $s_m=s_{m+i}, m+i\le n$ *else* $k=n-m$. We show below, as an example, the *GIN* for a short DNA fragment *AGTTCTACCAGC* is given as

$GIN=6,9,1,2,3,6,3,1,3,2,1,0.$

From the global inter-nucleotide distance sequence *GIN*, we can get the inter-nucleotide distance sequence to the nucleotide $x\in N$. Four inter-nucleotide distance sequences for the same short DNA segment used previously were given as

$IN^A=6,3,2,\quad IN^C=3,1,3,0,\quad IN^G=9,1,\quad IN^T=1,2,6.$

A similar inter-nucleotide distance sequence to the nucleotide $x\in N$ was defined by Afreixo et al. (2009). The four inter-nucleotide distance sequences for the short DNA fragment *AGTTCTACCAGC*:

$CIN^A=6,3,3,\quad CIN^C=3,1,3,5,\quad CIN^G=9,3,\quad CIN^T=1,2,3,$

considering that the symbolic sequence is circular. The corresponding global distance sequence referred to as *CIN* is exemplified below for the same short DNA segment used previously,

$CIN=6,9,1,2,3,9,2,1,3,3,3,5,$

which is slightly different from the non-circular approach used by Nair and Mahalashmi (2005).

### 2.2. Conditional multinomial distribution

From the definition of the inter-nucleotide distance sequence, we clearly see that the total number of three other nucleotides was only considered in each inter-nucleotide distance sequence (Afreixo et al., 2009). In fact, we can count the number of each nucleotide about the genome sequence from its inter-nucleotide distance sequence. Consequently, we can derive four conditional multinomial distributions from the corresponding inter-nucleotide distance sequences.

We take the inter-nucleotide distance sequence about nucleotide $A$ ($CIN^A$) as an example. Considering the case of $CIN^A=k$ ($k=1,2,\ldots$), let $p_{C|A}$,$p_{G|A}$ and $p_{T|A}$ be the occurrence probabilities of nucleotides $C$, $G$, and $T$, respectively, between the nearest two nucleotide $A$. If the nucleotide sequence was generated by an independent and identically distributed (i.i.d) random process, the number of nucleotide $C$, $G$ and nucleotide $T$ between the nearest two nucleotide $A$ would follow a multinomial distribution. In fact, the joint probability function of $(N_{C|A},N_{G|A},N_{T|A})$ (the number of $C$, $G$, $T$, between the nearest two nucleotide $A$, respectively, given that $CIN^A=k$) is

$$P(N_{C|A}=k_1,N_{G|A}=k_2,N_{T|A}=k_3|CIN^A=k)=\frac{(k-1)!}{k_1!k_2!k_3!}(p_{C|A})^{k_1}(p_{G|A})^{k_2}(p_{T|A})^{k_3},\quad(1)$$

where $k_1+k_2+k_3=k-1$, $k_i=0,1,\ldots,k-1$, $i=1,2,3$. The nucleotide occurrence probability $p_{C|A}$ is estimated by the relative frequency $TN_{C|A}/((k-1)\cdot N^A)$, where $N^A$ is the times of $CIN^A=k$, $TN_{C|A}$ is the total number of $C$ between the nearest two nucleotide $A$ when the inter-nucleotide distance $CIN^A=k$. The nucleotide occurrence probability $p_{G|A}$ and $p_{T|A}$ can be obtained in the similar method. The term *reference conditional multinomial distribution*, applied to a DNA sequence, describes the number of nucleotide $C$, $G$ and nucleotide $T$ would follow that the inter-nucleotide distance sequence about nucleotide $A$ is given, if its nucleotides are randomly determined, with probabilities equal to the relative conditional frequencies, independently of each other.

### 2.3. Complete multinomial composition vector

From the perspective of molecular evolution, conditional multinomial distribution may reflect both the results of random mutation and selective evolution. Mutations have been taking place randomly at molecular level and natural selections shape the direction of evolution. Many neutral mutations may remain and play a role of random background. One should subtract the random background from the simple counting result in order to highlight the contribution of selective evolution (Chang and Wang, 2011; Ding et al., 2010; Gao et al., 2006). In this work, we propose a new conditional multinomial distribution representation which reveals the relative difference of biological sequence from sequence generated by an independent random process to remove the random background.

For a fixed $k$, we can obtain a measured conditional multinomial distribution and a reference conditional multinomial distribution for a certain nucleotide $x\in\{A,C,G,T\}$. For a certain pattern $\alpha$ from the conditional multinomial distribution, we can define the *multinomial composition value* $\pi(\alpha)$ as follows:

$$\pi^x(\alpha|k)=\frac{f_0^x(\alpha|k)-f^x(\alpha|k)}{f^x(\alpha|k)},$$

where the $f_0^x(\alpha|k)$ is the measured relative frequency of the pattern $\alpha$, the relative frequency of the pattern $\alpha$ from the reference conditional multinomial distribution $f^x(\alpha|k)$ can be computed by (1). All these multinomial composition values can be sorted in some order to form a vector $V^x(S|k)=(\pi^x(\alpha_1|k),\pi^x(\alpha_2|k),\ldots,\pi^x(\alpha_m|k))$ for the genome $S$, where $m$ denotes the total number of patterns under consideration. Moreover, four vectors $V^A(S|k)$, $V^G(S|k)$, $V^C(S|k)$ and $V^T(S|k)$ are sorted in some order to form a vector $V(S|k)$ that represents the whole genome $S$. The vector defined by all these multinomial composition values is referred to as the $k$-order multinomial composition vector ($k-MCV$).

For only a fixed $k$, the $k$-order multinomial composition vector of the whole genome $S$ may lost some evolutionary information. The complete multinomial composition vector (*CMCV*) of the whole genome is the concatenation of $V(S|3),V(S|4),\ldots,V(S|k)$, denoted by $CMCV(S,k)$, with the intention to use as much genomic information as possible.

## 3. Results and discussions

### 3.1. The conditional multinomial distribution profile of chromosomes

We begin with the largest fragments of available DNA sequences, the chromosomes of eukaryotes listed in Table 1. The conditional multinomial distribution profile shown in Fig. 1 corresponds to three different chromosomes of *Saccharomyces cerevisiae*. In the case of inter-nucleotide distance sequence $CIN^A$ ($k=5$), we firstly convert the possible value of the $(N_{C|A},N_{G|A},N_{T|A})$ into one-dimensional value by the order of alphabet. We secondly plot the measured conditional multinomial distribution by bar and the reference conditional multinomial distribution by line. We clearly see a pattern of peaks and valleys which occur at the identical

locations for all chromosomes of *S. cerevisiae*. Three other cases are obtained in the similar way. We see again the similarity between the conditional multinomial distribution profiles about a certain nucleotide for the various chromosomes.

If we repeat this experiment for the chromosomes of *Caenorhabditis elegans* we get the same result. Again, when we plot the conditional multinomial distribution profile about a certain nucleotide we see a pattern of peaks and valleys which occur at the identical locations for all chromosomes of *C. elegans*. We demonstrate this with three chromosomes of *C. elegans* in Fig. 2. Again, while the pattern of peaks and valleys in the conditional multinomial distribution profile about a certain nucleotide is the same for all chromosomes of *C. elegans*, this pattern is distinctly different from the pattern of peaks and valleys in the *S. cerevisiae* profile about the same nucleotide.

Finally we repeat the experiment for Mouse. The result is shown in Fig. 3. Once more we obtain a sequence of peaks and valleys in the conditional multinomial distribution profile about a certain nucleotide which are the same for all chromosomes of Mouse, and this pattern of peaks and valleys is different from the pattern in the *S. cerevisiae* and *C. elegans* profiles.

**Table 1**
Labels for chromosomes.

| No. | Strain name | Accession | Chromosome |
|---|---|---|---|
| 1 | m14 | NT_002582 | M. musculus chromosome 14 |
| 2 | m17 | NT_002588 | M. musculus chromosome 17 |
| 3 | MX | NT_003030 | M. musculus chromosome X |
| 4 | sc3 | NC_001135 | S. cerevisiae chromosome 3 |
| 5 | sc5 | NC_001137 | S. cerevisiae chromosome 5 |
| 6 | sc9 | NC_001141 | S. cerevisiae chromosome 9 |
| 7 | ce1 | NC_000965 | C. elegans chromosome 1 |
| 8 | ce2 | NC_000966 | C. elegans chromosome 2 |
| 9 | ce3 | NC_000967 | C. elegans chromosome 3 |

### 3.2. Phylogenetic analysis

The complete multinomial composition vector of each complete genome provides a simple, easily computable signature that identifies each species. The signature can be used in application where evolutionary relationships need to be deduced using large genomic sequence. Distances between sets of genomic sequences can be obtained without the need for multiple sequence alignment. Phylogenetic trees are generated by putting the pairwise distance matrix into UPGMA method in the PHYLIP package (Felsensein, 1989).

The outbreak of atypical pneumonia referred as severe acute respiratory syndrome coronavirus (SARS-CoVs) in 2003 had caught more attention to the relationship between the SARS-CoVs and the others coronaviruses. The 24 complete coronavirus genomes used in this paper were downloaded from GenBank, of which 12 are SARS-CoVs and 12 are from other groups of coronaviruses. The name, accession number, abbreviation, and genome length for the 24 genomes are listed in Table 2. Generally, coronavirus can be classified into three groups according to serotypes. Group I and group II contain mammalian viruses, whereas group III contains only avian. Many investigations have attempted to identify the phylogenetic position of SARS-CoVs. However, this is still a controversial topic-alignment-based methods showed that SARS-CoVs are not closely related to any groups and form a new group (Marra et al., 2003; Rota et al., 2003); maximum likelihood tree built from a fragment of the spike protein preferred SARS-CoVs clustering with group II (Liò and Goldman, 2004); while an information-based method, which makes use of the whole genome sequences, indicated that SARS-CoVs are close to the group I rather than from a new group (Yang et al., 2005). Based on the complete multinomial

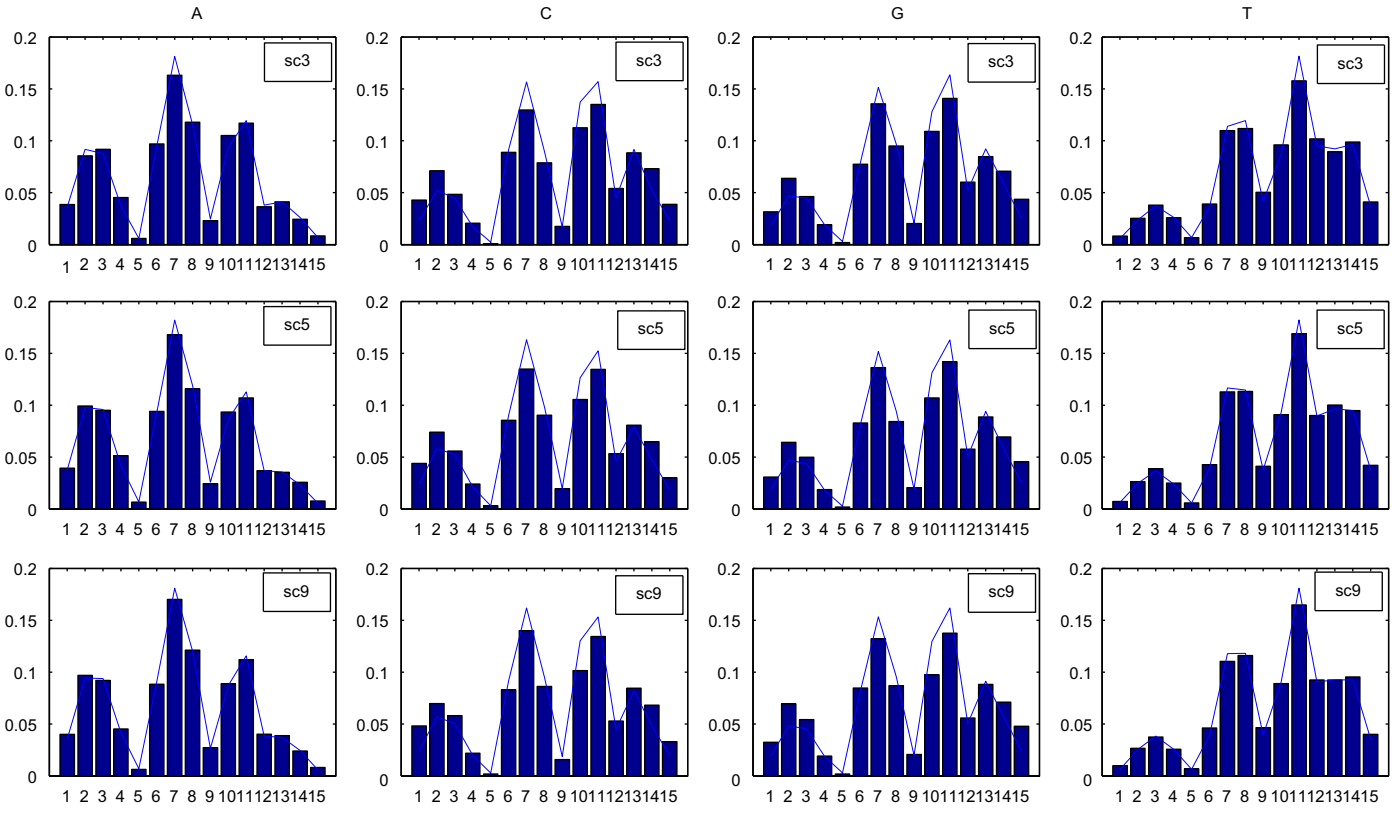


**Fig. 1.** Conditional multinomial distribution profile for three different chromosomes of *Saccharomyces cerevisiae*. The histogram is from the measured conditional multinomial distribution and the line indicates the reference conditional multinomial distribution with parameters estimated from the data ($5-MCV$).

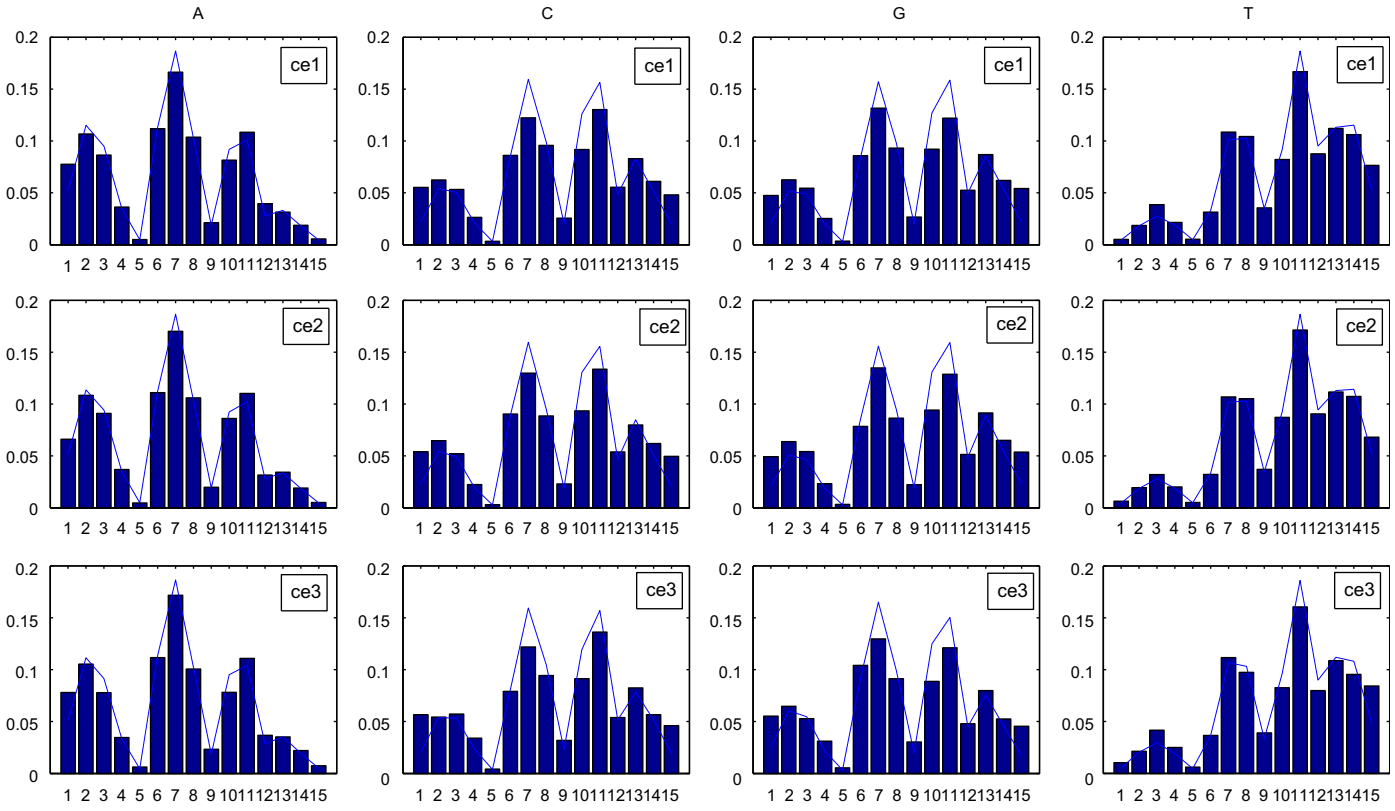


**Fig. 2.** Conditional multinomial distribution profile for three different chromosomes of *C. elegans*. The histogram is from the measured conditional multinomial distribution and the line indicates the reference conditional multinomial distribution with parameters estimated from the data ($5-MCV$).

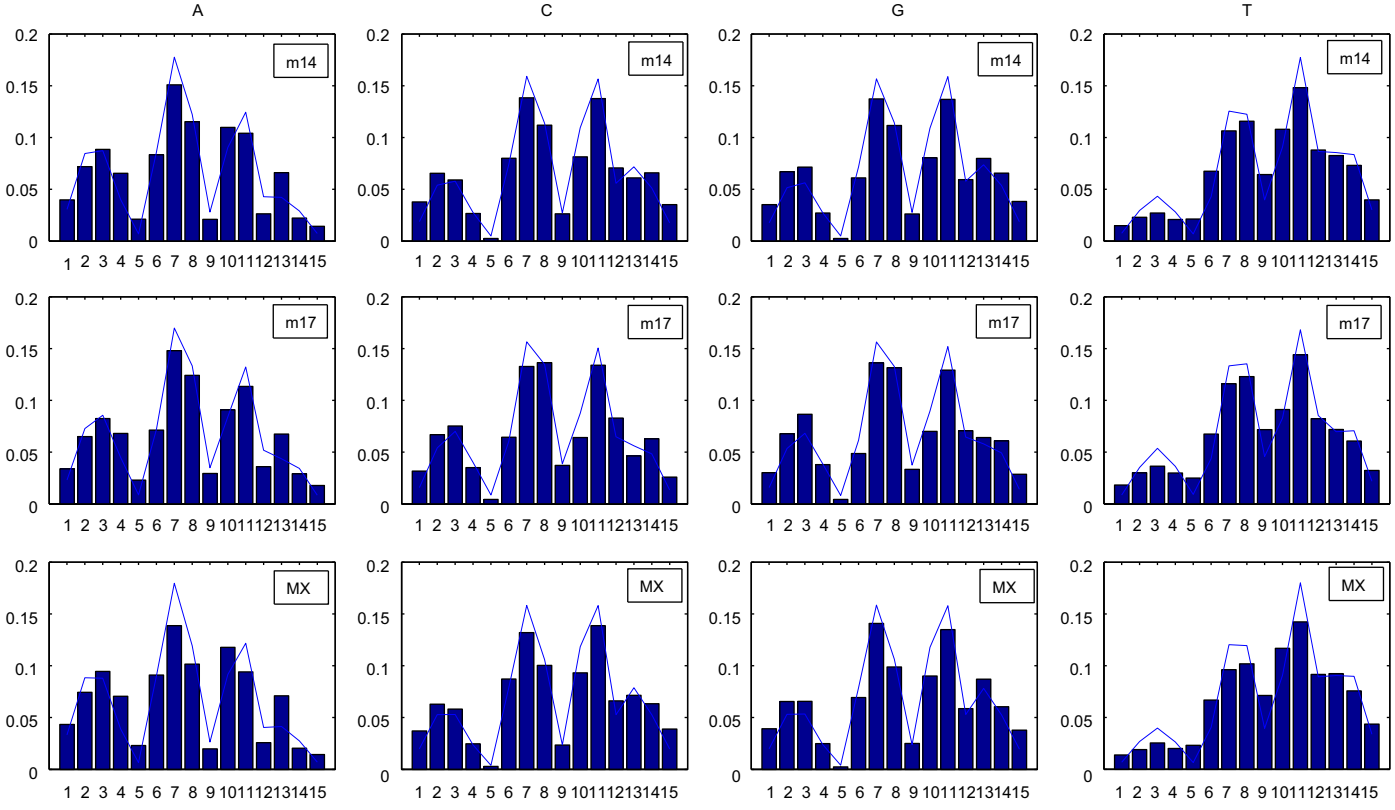


**Fig. 3.** Conditional multinomial distribution profile for three different chromosomes of Mouse. The histogram is from the measured conditional multinomial distribution and the line indicates the reference conditional multinomial distribution with parameters estimated from the data ($5-MCV$).

**Table 2**
The accession number, abbreviation, name, and length for each of the 24 coronavirus genomes.

| No. | Accession | Abbreviation | Genome | Length (nt) |
|---|---|---|---|---|
| 1 | NC_002645 | HCov-229E | 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_4030 | BCoV | Bovine coronavirus | 31,028 |
| 8 | AF208067 | MHVM | Murine hepatitis virus strain ML-10 | 31,233 |
| 9 | AF201929 | 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-SulO | 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 |

composition vector, we build the phylogenetic tree of the 24 coronaviruses listed in Table 2. The phylogenetic tree is built using the UPGMA programs in the PHYLIP package and the distance matrix is computed using the Euclidean distance (Felsensein, 1989).

Our results based on analysis of the complete multinomial composition vector of 24 coronavirus genomes have some notable distinction from the previous phylogenetic study using an information-based similarity index (Yang et al., 2005). As can be seen from Fig. 4, our method indicates that SARS-CoVs are not closely related to any of the previously characterized coronaviruses and form a distinct group (group IV). Our results also show that group II, BCoV, BCoVL, BCoVM, etc., are grouped in a monophyletic clade. This result is also mainly in accordance with the conclusions from the alignment-based method (Marra et al., 2003; Rota et al., 2003) and the alignment-free method (Liu et al., 2007). Moreover the Robinson–Foulds distance between our tree and the result of Liu's is only 26.

### 3.3. The selection of k in CMCV (S, k)

The selection of $k$ in $CMCV(S,k)$ is very important to capture rich evolutionary information of DNA sequence. In the case of $k=1$, there is no nucleotide between two adjacent nucleotides. In the case of $k=2$, there is only one nucleotide between two adjacent nucleotides. Therefore the multinomial composition value of a certain pattern $\alpha$ is zero. The *CMCV* does not contain these multinomial composition values. Certainly, a large value of $k$ will give a vector containing finer evolutionary information. However, many patterns will not occur in the conditional multinomial distribution with a large value of $k$. From the view of information theory, some information may be lost and noise will dominate if a large value of $k$ is considered. To determine the upper bound of the value of $k$, we will introduce a scoring scheme to estimate how important a conditional multinomial distribution is.

$\chi^2-$Test scoring scheme: For a fixed $k$, let $\alpha$ be a pattern in the conditional multinomial distribution, with its multinomial composition value $\pi(\alpha,i|k)$ in genome $i$ (could be found in $k-MCV$). Define the expected multinomial composition value for pattern $\alpha$ to be the average of all composition values across all whole genomes, and denoted as, $E[\pi(\alpha|k)]$ i.e. $E[\pi(\alpha|k)]=(1/n)\sum_{i=1}^{n}\pi(\alpha,i|k)$—assuming $n$ genomes in the dataset. The standard $\chi^2-$test measures the deviation of a set of values from its expected value by summing up the deviations of each element. Clearly, the higher value it has, the more valuable pattern $\alpha$ is. Thus, we may define a score for the conditional multinomial distribution with a fixed $k$ as

$$Score(k)=\frac{1}{\binom{k+1}{2}}\sum_{\alpha}\sum_{i=1}^{n}\frac{(\pi(\alpha,i|k)-E[\pi(\alpha|k)])^2}{|E[\pi(\alpha|k)]|},$$

where the first sum is for all patterns of the conditional multinomial distribution with a fixed $k$.

We believe by considerably extending the basic pattern counting idea and thus studying their underlying distribution, we are able to discover unusual patterns to automatically distinguish their roles in shaping the evolution. In this case, the largest score of conditional multinomial distribution, the $k-MCV$ might be considered as the most representative for the species, while not as abnormal outliers from the pure statistical analysis.

We listed the score of the conditional multinomial distribution with a fixed $k$ (within the range [3,9]) from the dataset of 24 complete coronavirus genomes in Table 3. The score of *CMCV* can be defined as sum of scores of $k-MCV$ involved in the *CMCV*. From Table 3, it is clearly that there is no large difference after the $7-MCV$ is added in the *CMCV*. Moreover, we can define the relative ratio of information involved in a certain conditional multinomial distribution with a fixed $k$ as the $k-MCV$ to the *CMCV* which will involve the $k-MCV$. Form Table 3, we can clearly see that the relative ratio of $7-MCV$ is the maximum $\frac{839}{1408}$. Therefore, we select $CMCV(S,7)$ to represent the genome $S$ in the phylogenetic analysis of the 24 coronaviruses.

## 4. Conclusion

Description and comparison of DNA sequences are still important subjects in bioinformatics. DNA sequence databases have accumulated much data on biological evolution during billions of years, consequently novel concepts and methods are urgent need to reveal the biological functions of DNA sequences information, to investigate relationships of DNA sequences with biological evolution, cellular function, genetic mechanism and occurrence of

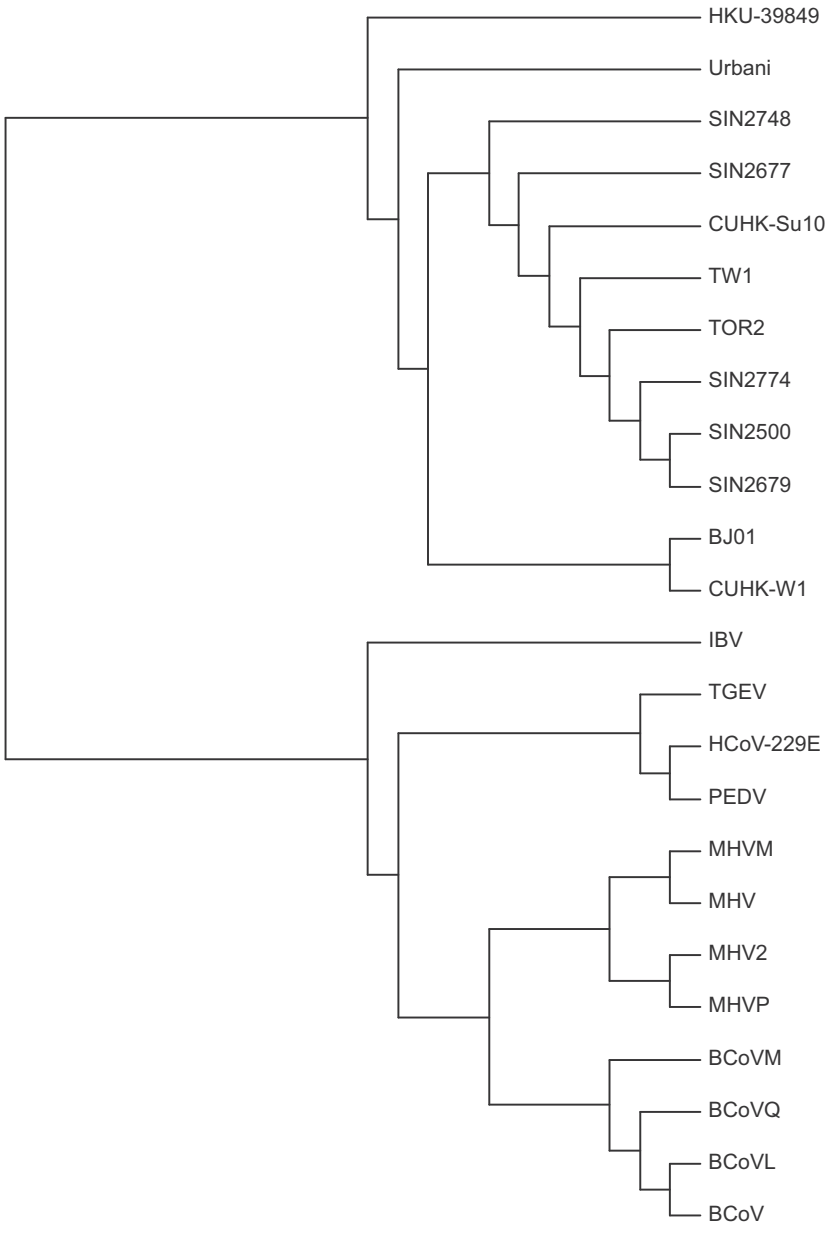


**Fig. 4.** The NJ tree of 24 complete coronavirus genomes is constructed by *CMCV* (*S*, 7).

**Table 3**
Scores for some conditional multinomial distributions.

| *k* | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|
| *Score* | 1194 | 26 | 104 | 84 | 839 | 211 | 313 |
| *SumScore* | 1194 | 1220 | 1324 | 1408 | 2247 | 2458 | 2771 |
| Ratio | – | $\frac{26}{1194}$ | $\frac{104}{1220}$ | $\frac{84}{1324}$ | $\frac{839}{1408}$ | $\frac{211}{2247}$ | $\frac{313}{2458}$ |

illness. In this paper, we propose four conditional multinomial distributions about each nucleotide for complete genome sequence based on the inter-nucleotide distance sequences. From the conditional multinomial distribution profiles about nine chromosomes, we note that the relative error vector between the measured conditional multinomial distribution and the reference conditional multinomial distribution can be used as a genomic signature, thus allowing the comparison of species. Therefore, it is straightforward to generate a phylogenetic tree based on the Euclidean distances of complete multinomial composition vectors.

In order to test the validity of our method, we select the complete genome sequences of 24 coronaviruses which were used by Liu et al. (2007). The phylogenetic tree can be gotten through the distance matrices using the UPGMA method. Fig. 4 is the phylogenetic tree of the 24 genome sequences based on the distance matrix of the complete multinomial composition vector, using UPGMA method. We find that the tree is mainly consistent with the tree constructed by Liu et al. (2007). Fig. 4 also indicates that SARS-CoVs are not closely related to any groups and form a new group.

Overall our results highlight that the conditional multinomial distribution profiles have the ability to extract more information from the genome sequence. Thus this opinion can then be used to guide the development more powerful measures for sequence comparison with future possible improvement on the correlation structure of DNA.

## Acknowledgments

We would like to thank the reviewers for their useful and critical comments, all of which have greatly improved the quality of the paper. This work is supported by the National Natural Science Foundation of China (Grant no. 10871219).

## References

Afreixo, V., Bastos, C., Pinho, A., Garcia, S., Ferreira, P., 2009. Genome analysis with inter-nucleotide distances. Bioinformatics 25, 3064–3070.
Akhtar, M., Epps, J., 2008. Signal processing in sequence analysis: advances in eukaryotic gene prediction. IEEE J. Sel. Top. Signal Process. 2, 310–321.
Akhtar, M., Ambikairajah, E., Epps, J., 2007. On DNA numerical representation for period-3 based exon prediction. In: 5th International Workshop on Genomic Signal Processing and Statistics, Tuusula, Finland.
Chang, G.S., Wang, T.M., 2011. Weighted relative entropy for alignment-free sequence comparison based on Markov model. J. Biomol. Struct. Dyn. 28, 545–555.
Dai, Q., Liu, X.Q., Wang, T.M., 2006. Numerical characterization of DNA sequences based on the k-step Markov chain transition probability. J. Comput. Chem. 27, 1830–1842.
Ding, S.Y., Dai, Q., Liu, H.M., Wang, T.M., 2010. A simple feature representation vector for phylogenetic analysis of DNA sequences. J. Theor. Biol. 265, 618–623.
Felsensein, J., 1989. PHYLIP-phylogeny inference package (version 3.2), vol. 5, pp. 164–166.
Gao, L., Qi, J., Hao, B.L., 2006. Simple Markov subtraction essentially improves prokaryote phylogeny. AAPPS Bull. June, 3–7.
He, P., Wang, J., 2002. Numerical characterization of DNA primary sequence. Int. Electron. J. Mol. Des. 12, 668–674.
Jeffrey, H.J., 1990. Chaos game representation of gene structure. Nucleic Acids Res. 18, 2163–2170.
Liò, P., Goldman, N., 2004. Phylogenomics and bioinformatics of SARS-CoV. Trends Microbiol. 12, 106–111.
Liao, B., Tan, M.S., Ding, K.Q., 2005. Application of 2-D graphical representation of DNA sequence. Chem. Phys. Lett. 401, 196–199.
Liu, Y.Z., Yang, Y.C., Wang, T.M., 2007. Characteristic distribution of L-tuple for DNA primary sequence. J. Biomol. Struct. Dyn. 25, 85–91.
Marra, M.A., et al., 2003. The genome sequence of SARS-associated coronavirus. Science 300, 1399.
Nair, A.S., Mahalashmi, T., 2005. Visualization of genomic data using inter-nucleotide distance signal. In: Processing of IEEE Genomic Signal Processing, Bucharest, Romania.
Nandy, A., 1994. A new graphical representation and analysis of DNA sequences structure: I. Curr. Sci. 66, 309–314.
Nandy, A., Nandy, P., 1995. Graphical analysis of DNA sequences structure: II Relative abundances of nucleotide in DNAs, gene evolution and duplication. Curr. Sci. 68, 75–85.
Nandy, A., Nandy, P., 2003. On the uniqueness of quantitative DNA difference descriptors in 2D graphical representation model. Chem. Phys. Lett. 368, 102–107.
Nandy, A., Harle, M., Basak, S.C., 2006. Mathematical descriptors of DNA sequences: development and applications. Arkivoc ix, 211–238.
Randic, M., Vracko, M., 2000. On the similarity of DNA primary sequences. J. Chem. Inf. Comput. Sci. 40, 599–606.
Randic, M., Guo, X., Basak, S.C., 2001. On the characterization of DNA primary sequences by triplet of nucleic acid bases. J. Chem. Inf. Comput. Sci. 41, 619–626.
Randic, M., Balaban, A.T., 2003. A four-dimensional representation of DNA primary sequences. J. Chem. Inf. Comput. Sci. 43, 532–539.
Randic, M., Vracko, M., Lers, N., Plavsic, D., 2003a. Novel 2-D graphical representation of DNA sequences and their numerical characterization. Chem. Phys. Lett. 368, 1–6.
Randic, M., Vracko, M., Lers, N., Plavsic, D., 2003b. Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation. Chem. Phys. Lett. 371, 202–207.
Randic, M., Butina, D., Zupan, J., 2006. Novel 2-D graphical representation of proteins. Chem. Phys. Lett. 419, 528–532.
Randic, M., 2008. Another look at the chaos-game representation of DNA. Chem. Phys. Lett. 456, 84–88.
Rota, P.A., et al., 2003. Characterization of a novel coronavirus associated with severe acute respiratory syndrome. Science 300, 1394–1399.
Vinga, S., Almeida, J., 2003. Alignment free sequence comparison—a review. Bioinformatics 19, 513–523.
Voss, R.F., 1992. Evolution of long-range fractal correlation and $1/f$ noise in DNA base sequence. Phys. Rev. Lett. 68, 3805–3806.
Yang, A.C., Goldberger, A.L., Peng, C.K., 2005. Genomic classification using an information-based similarity index: application to the SARS coronavirus. J. Comput. Biol. 12, 1103–1116.
Zhang, R., Zhang, C.T., 1994. Z curve, an intuitive tool for visualising and analysing the DNA sequences. J. Biomol. Struct. Dyn. 11, 767–782.