Coronavirus phylogeny based on Base-Base Correlation

Zhi-Hua Liu

State Key Laboratory of Bioelectronics,
Southeast University,
Nanjing 210096, PR China

Harvard Medical School,
Dana-Farber Cancer Institute,
Department of Biostatistics and Computational Biology,
44 Binney St., Boston, Massachusetts 02115, USA

Harvard School of Public Health,
677 Huntington Avenue,
Boston, Massachusetts 02115, USA
E-mail: zhliu@jimmy.harvard.edu E-mail: zhliu@seu.edu.cn

Xiao Sun*

State Key Laboratory of Bioelectronics,
Southeast University,
Nanjing 210096, PR China
E-mail: xsun@seu.edu.cn
*Corresponding author

Abstract: With more and more complete genome sequences having been released, phylogenetic analysis is entering a new era – that of phylogenomics. In this paper, a novel phylogenomic method, named as Base-Base Correlation (BBC), has been proposed to infer phylogenetic relationships from complete genomes, with particular emphasis on coronavirus phylogeny. Following the high-profile publicity of SARS outbreaks, a renewed interest in coronavirus has been promoted and two novel human coronaviruses (NL63 and HKU1) have been identified. Coronavirus phylogenomics based on BBC is well consistent with that of previous studies. BBC, to study genome information structure based on information theory, provides a novel alignment-free phylogenomic methodology in postgenome informatics.

Keywords: Base-Base Correlation; BBC; phylogenomics; genome information structure; coronavirus phylogeny; SARS; NL63; HKU1; bioinformatics.

Reference to this paper should be made as follows: Liu, Z-H. and Sun, X. (2008) 'Coronavirus phylogeny based on Base-Base Correlation', *Int. J. Bioinformatics Research and Applications*, Vol. 4, No. 2, pp.211–220.

Biographical notes: Zhi-Hua Liu is a PhD candidate in the State Key Laboratory of Bioelectronics at Southeast University. Now he is a visiting scholar in Harvard School of Public Health and Dana Farber Cancer Institute, Harvard Medical School. His research interests are in the areas of feature-based

sequence analysis and classification, molecular evolution and genome evolution, exon-intron structure and noncoding sequence analysis.

Xiao Sun, PhD, is a Professor of State Key Laboratory of Bioelectronics at Southeast University. He holds a PhD Degree from Southeast University. His primary areas of scientific expertise include computational biology and bioinformatics. His recent academic interests include the application and development of feature-based methods to analyse nucleotide sequence.

1 Introduction

Until recently, the traditional phylogeny was mainly based on 16S small ribosomal RNA (16S rRNA) sequence comparisons. Although such molecules have proved to be universal distribution and evolutionary conservation, mutational saturation is a problem, due to their restricted lengths (Henz et al., 2005; Moreira and Philippe, 2000). Moreover, it has been shown that rRNA-based phylogeny can be sometimes grossly misleading in inferring phylogenetic relationships in the presence of unequal rates of evolution or differences in base composition (Philippe and Laurent, 1998). To overcome this limitation, it is tempting to apply a genome-scale approach to phylogenetic inference (phylogenomics). The rapidly increasing availability of complete genome sequence has also prompted an interest in using whole genome information to infer phylogenetic relationships.

In addition, traditional phylogenetic methods include a model of multiple sequence alignment. However, when large genome sequences are analysed, the traditional alignment methods appear to be time consuming. Moreover, in sequence alignment, insertions and deletions are poorly evaluated due to the assumption of regular evolutionary models (Grasso and Lee, 2004; Lee et al., 2002; Raphael et al., 2004). Thus, there is a need for an efficient alignment-free way to transcribe whole genome sequence into pertinent phylogenetic information.

Here we developed a novel phylogenomic approach without alignment, named BBC, which is inspired from using Mutual Information Function (MIF) to analyse DNA sequence. Compared with MIF, BBC emphasised the information of different base pairs within the range of k. It improved the resolving power and provided a more appropriate description of sequence dissimilarity (Liu et al., 2007). In this paper, we present our study of applying BBC to phylogenetic inference, with particular emphasis on coronavirus phylogeny.

Coronavirus is a genus of animal virus belonging to the family *Coronaviridae*. Coronaviruses are enveloped viruses with a positive-sense, single-stranded RNA genome and a helical symmetry. The genome size of coronaviruses ranges from approximately 16–31 kilobases, extraordinarily large for an RNA virus (Rota et al., 2003). Coronaviruses can be divided into three groups according to serotypes. Groups I and II contain mammalian viruses, while group II coronaviruses contain a hemagglutinin esterase gene homologous to that of Influenza C virus. Group III contains only avian viruses. In 2003, a novel coronavirus was isolated and found to be the cause of severe acute respiratory syndrome, which had begun the prior year in Asia, and secondary cases elsewhere in the world. The virus was officially named the SARS Coronavirus

(SARS-CoV). For many years, scientists knew only about the existence of two human coronaviruses (HCoV-229E and HCoV-OC43). The discovery of SARS-CoV has promoted a renewed interest in coronavius in the field of virology. By the end of 2004, three independent research labs reported the discovery of a fourth human coronavirus (Hofmann et al., 2005). It has been named NL63, NL or the New Haven coronavirus by the different research groups. Early in 2005, a research team at the University of Hong Kong reported finding a fifth human coronavirus in two pneumonia patients, and subsequently named it HKU1 (Woo et al., 2005).

2 Methods

2.1 Materials

A total of 26 complete coronavirus genomes used in this study were retrieved from NCBI (http://www.ncbi.nlm.nih.gov/). The name, abbreviation, accession number, genome length, and the existing taxonomic groups for the 26 coronavirus genomes are shown in Table 1.

Table 1 The name, abbreviation, accession number, and genome length for each of the 26 genomes

No.	*Genomes*	*Abbreviation*	*Accession*	*Length* (nt)	*Group*
1	Human coronavirus 229E	HCoV-229E	NC_002645	27,317	I
2	Transmissible gastroenteritis virus	TGEV	NC_002306	28,586	I
3	Porcine epidemic diarrhea virus	PEDV	NC_003436	28,033	I
4	Bovine coronavirus strain Mebus	BCoVM	U00735	31,032	II
5	Bovine coronavirus isolate BCoV-LUN	BCoVL	AF391542	31,028	II
6	Bovine coronavirus strain Quebec	BCoVQ	AF220295	31,100	II
7	Bovine coronavirus	BCoV	NC_003045	31,028	II
8	Murine hepatitis virus strain ML-10	MHVM	AF208067	31,233	II
9	Murine hepatitis virus strain 2	MHV2	AF201929	31,276	II
10	Murine hepatitis virus strain Penn 97-1	MHVP	AF208066	31,112	II
11	Murine hepatitis virus	MHV	NC_001846	31,357	II
12	Avian infectious bronchitis virus	IBV	NC_001451	27,608	III
13	SARS coronavirus BJ01	BJ01	AY278488	29,725	IV
14	SARS coronavirus Urbani	Urbani	AY278741	29,727	IV
15	SARS coronavirus HKU-39849	HKU-39849	AY278491	29,742	IV
16	SARS coronavirus CUHK-W1	CUHK-W1	AY278554	29,736	IV
17	SARS coronavirus CUHK-Su10	CUHK-Su10	AY282752	29,736	IV
18	SARS coronavirus Sin2500	SIN2500	AY283794	29,711	IV
19	SARS coronavirus Sin2677	SIN2677	AY283795	29,705	IV
20	SARS coronavirus Sin2679	SIN2679	AY283796	29,711	IV
21	SARS coronavirus Sin2748	SIN2748	AY283797	29,706	IV

Table 1 The name, abbreviation, accession number, and genome length for each of the 26 genomes (continued)

No.	*Genomes*	*Abbreviation*	*Accession*	*Length* (nt)	*Group*
22	SARS coronavirus Sin2774	SIN2774	AY283798	29,711	IV
23	SARS coronavirus TW1	TW1	AY291451	29,729	IV
24	SARS coronavirus	TOR2	NC_004718	29,751	IV
25	Human coronavirus NL63	NL63	NC_005831	27,553	I
26	Human coronavirus HKU1	HKU1	NC_006577	29,926	II

2.2 Base-Base Correlation (BBC)

DNA sequences can be viewed as symbolic strings composed of the four letters $(B_1, B_2, B_3, B_4) \equiv (A, C, G, T)$. The probability of finding the base B_i is denoted by $p_i (i = 1, 2, 3, 4)$. Then BBC is defined as the following:

$$T_{ij}(k) = \sum_{l=1}^{k} p_{ij}(l) \cdot \log_2 \left(\frac{p_{ij}(l)}{p_i p_j}\right) \quad i, j \in \{1, 2, 3, 4\}. \tag{1}$$

Here, $p_{ij}(l)$ means the joint probabilities of bases i and j at a distance of l. $T_{ij}(k)$ represents the average relevance of the two-base combination with different gaps from 1 to k. It reflects a local feature of two bases within the range of k. For each genome sequences m, BBC has 16 parameters and constitutes a 16-dimensional vector $V_m^z (z = 1, 2, \ldots, 16)$.

Let L be a whole genome sequence length ($1 \le k \le L$). Thus, $T_{ij}(L)$ contains all base pairs information for this genome sequence. Theoretically, BBC feature extract more fully genome information when k is larger. However, we find that BBC has no considerable changes when $k > 147$ (Liu et al., 2007). Biological significance of k value may be related to the fact that nucleosomal DNA contains a core DNA region with a stable length of 147 bp, which is relatively resistant to digestion by nucleases. So, we take $k = 147$ in BBC calculation for genome sequence in the present study.

Statistical independence of two bases in a distance l is defined by $p_{ij}(l) = p_i p_j$. Thus, deviations from statistical independence is defined by

$$D_{ij}(l) = p_{ij}(l) - p_i p_j. \tag{2}$$

We expand $T_{ij}(k)$ using a Taylor series in terms of equation (2)

$$\begin{aligned} T_{ij}(k) &= \sum_{l=1}^{k} p_{ij}(l) \cdot \log_2 \left(\frac{p_{ij}(l)}{p_i p_j} \right) \\ &= \sum_{l=1}^{k} [D_{ij}(l) + p_i p_j] \cdot \ln \left[1 + \frac{D_{ij}(l)}{p_i p_j} \right] \\ &= \sum_{l=1}^{k} [D_{ij}(l) + p_i p_j] \cdot \left[\frac{D_{ij}(l)}{p_i p_j} - \frac{D_{ij}^2(l)}{2 p_i p_j} + \cdots \right] \\ &= \sum_{l=1}^{k} D_{ij}(l) + \frac{D_{ij}^2(l)}{2 p_i^2 p_j^2} + o[D_{ij}^3(l)]. \end{aligned} \tag{3}$$

This mathematical transformation further increases the calculation speed and solves effectively the problem of $0 \cdot \log_2 0$ (i.e., $p_{ij}(l) = 0$ in equation (1)).

2.3 The distance matrix

Given two sequences m and n, the distance H_{mn} between two sequences m and n is defined as the following:

$$H_{mn} = \sqrt{\sum_{z=1}^{16} (V_m^z - V_n^z)^2} \quad m, n = 1, 2, \cdots, N. \tag{4}$$

Here, V_m and V_n represent the 16-dimensional vectors of sequences m and n. N is the total number of all sequences analysed. According to equation (4), H_{mn} satisfies the definition of distance: (π) $H_{mn} > 0$ for $m \neq n$; (θ) $H_{mm} = 0$; (ρ) $H_{mn} = H_{nm}$ (symmetric); (σ) $H_{mn} \leq H_{mq} + H_{nq}$ (triangle inequality). For N sequences, a real symmetric $N \times N$ distance matrix is then obtained.

2.4 Clustering

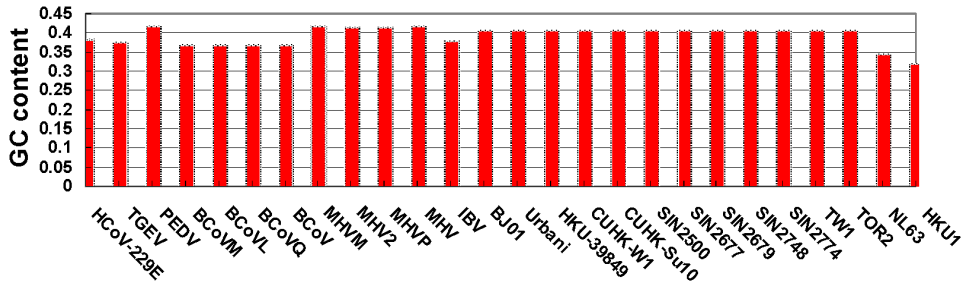
Accordingly, a real symmetric $N \times N$ matrix is used to reflect the evolutionary distance between N sequences. Then, the clustering tree is constructed using neighbour-joining method. The reliability of the branches is assessed by performing 100 resamplings. Bootstrap values are shown on nodes.

3 Results

3.1 GC content of 26 coronavirus genomes

GC content for each of 26 coronavirus genomes is shown in Figure 1. The GC content of coronavirus is below the value of 0.5. The GC content of 12 SARS-CoVs remain relatively stable at 0.4. NL63 and HKU1, which were identified after the outbreak of SARS, are two novel human coronaviruses with GC content below the value of 0.35. The GC content of HKU1 is 0.32, the lowest among all known coronaviruses. Previous studies have revealed a statistical relationship between gene density and GC content, whereas genome sequences with low GC content were also found to correlate with long intron length and a high LINE repeat density (Versteeg et al., 2003).

Figure 1 GC content for each of 26 coronavirus genomes

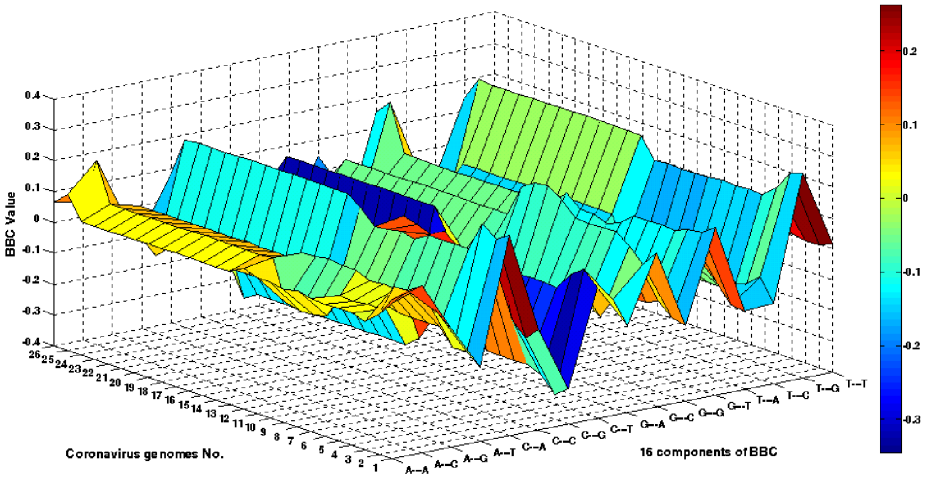


3.2 BBC curves of 26 coronavirus genomes

For each genome sequence, 16 parameters of BBC are calculated and linked to a continuous curve, which is designated BBC curve. BBC curve is then represented as a unique feature for a given genome sequence, providing an intuitionistic and general description for genome sequence.

BBC curves of 26 coronavirus genome sequences are displayed in Figure 2. Each curve represents a full-length coronavirus genome. It is found that BBC curves of SARS-CoVs (genome Nos. 13–24) are distinct from other coronaviruses.

Figure 2 BBC curves of 26 coronavirus genomes



3.3 The distance matrix of 26 coronavirus genomes

Figure 3 shows the distance matrix for 26 coronavirus genomes. This figure has two interesting features. First, a clear block structure indicates that coronavirus is divided into four groups. The blocks of SARS-CoVs (genome Nos. 13–24) are significantly different from the other blocks. Second, the blocks of NL63 and HKU1 (genome Nos. 25 and 26), identified as two novel human coronaviruses, are also distinct from the blocks of SARS-CoVs.

3.4 Coronavirus phylogeny based on Base-Base Correlation (BBC)

As shown in Figure 4, four groups of coronavirus can be seen from the phylogram. The SARS-CoVs appear to cluster together and form a separate branch, which can be distinguished easily from other three groups of coronavirus. NL63 and HCoV-229E tend to cluster together. PEDV and TGEV join them and result in group I. In another branch, the group II coronaviruses, including three subgroups (Bovine coronavirus, Murine hepatitis virus strain and Human coronavirus HKU1), tend to cluster together. Moreover, groups I and II, which are all mammalian viruses, cluster together forming a bigger group. IBV, belonging to group III, is situated at an independent branch. The resulting monophyletic clusters agree perfectly with the established taxonomic groups. Our results also show NL63 and HKU1 belong to groups I and II, respectively.

Figure 3 Density plot of the distance matrix for 26 coronavirus genomes

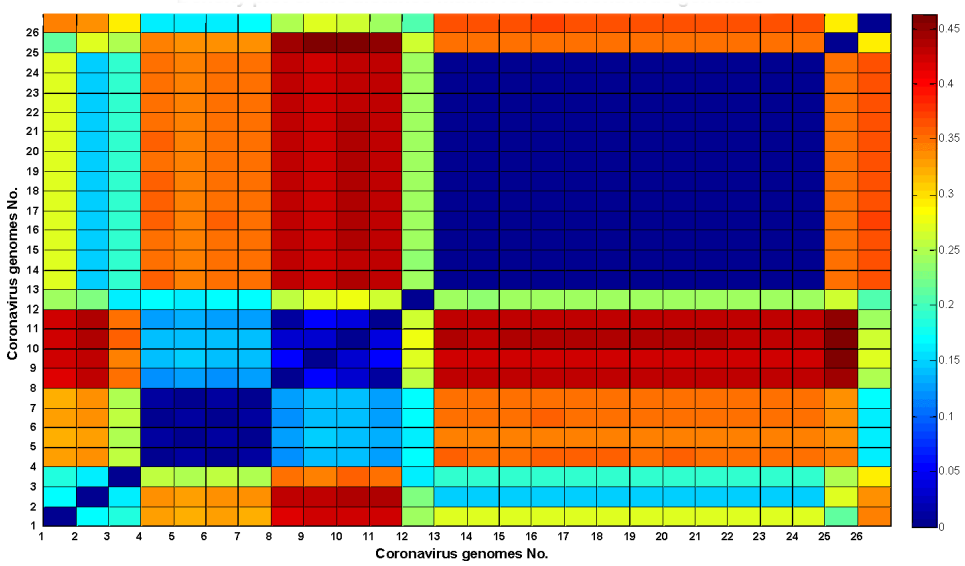
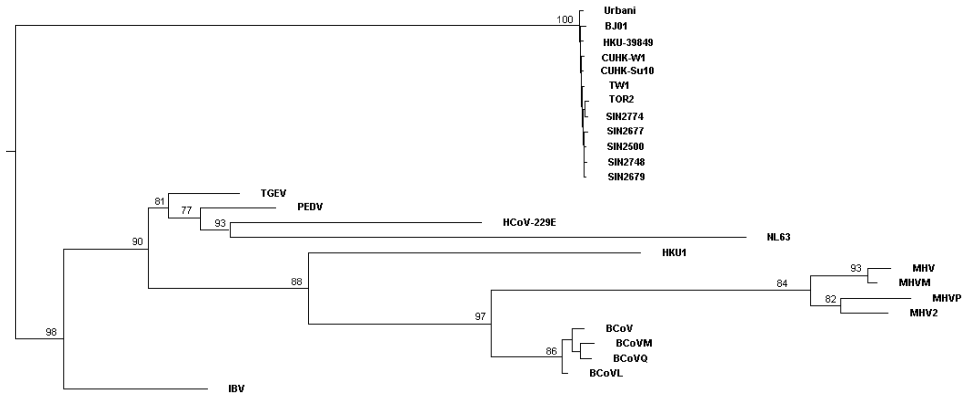


Figure 4 Coronavirus phylogeny based on base-base correlation



4 Discussion

In the present study, a novel algorithm based on BBC is proposed. Then, this algorithm is used for coronavirus phylogeny. The phylogenetic tree constructed by BBC algorithm can well agree with that of previous study.

Previous phylogenetic inference is based on multiple sequence alignment. However, a global multiple alignment of whole genome sequences appears to be time consuming. BBC vectors of 26 coronavirus genomes were calculated within a few seconds on a regular PC. However, multiple sequence alignment of 26 coronavirus genome sequences was performed with a few hours using ClustalX on the same PC.

In addition, most tools for multiple sequence alignment need extra operation such as "exclude positions with gaps", "correct for multiple substitutions" before constructing trees (Chenna et al., 2003; Thompson et al., 1997; Jeanmougin et al., 1998). These operations may throw away the most ambiguous parts of the alignment and underestimate actual evolutionary distances. Especially for sequences with very large divergence, the evolutionary distance cannot be reliably corrected by these alignment tools (Chenna et al., 2003; Thompson et al., 1997; Jeanmougin et al., 1998). BBC, as an alignment-free method, can overcome this limitation. A nucleotide sequence, regardless of its length is kilobases, megabases, or even gigabases, corresponds to a unique 16-dimensional vector. The procedure is actually a normalisation operation to compare genomes of different scales, which are difficult to obtain a good sequence alignment. Changes in the values of 16 parameters reflect different genome length and content. It is usually thought that higher sequence similarity may represent closer genetic relationships between virus strains. It also implies that BBC vectors tend to be more similar if virus strains are in closer genetic relationships. The evolutionary distance matrix is obtained by arithmetic operations between these 16-dimensional vectors, and then is used for the construction of phylogenetic tree.

Moreover, most phylogenetic analysis is always based on some special genes or some conserved fragments because these conservative regions tend to be more evolutionarily conserved. But analysis based on various parts of the genome may lead to different phylogenetic inferences. It is valuable to develop methods of whole genome phylogeny to overcome the biases. As a phylogenomic method, BBC has been applied to whole genome analysis (Liu and Sun, 2007; Liu et al., 2008). Actually, BBC considers full-length genome sequence as a whole, including coding and noncoding regions. The latter is associated with biological functions and may play an important role in the virus evolution. Former study found that BBC differed significantly between coding regions and noncoding regions (Liu et al., 2005). Phylogenetic analysis based on BBC that considers whole genome information including coding and noncoding regions, is likely to be more objective.

In addition, several genome-wide phylogenetic methods such as gene order (Boore and Brown, 1998) and gene content (Snel et al., 1999; Huson and Steel, 2004) need to identify gene. However, identification of gene is a time-consuming procedure. BBC method does not require gene identification or any human intervention.

5 Conclusions

With fast development of worldwide genome sequencing project, more and more completely sequenced genomes become available. However, traditional sequence alignment tools and regular evolutionary models are impossible to deal with large-scale genome sequence. In the present study, a novel phylogenomic method, named BBC, is proposed. We applied BBC to the coronavirus phylogeny. The result is well consistent with that of previous analysis. BBC, not limited to coronavirus phylogeny, provides a fast and intuitionistic tool for whole genome sequence comparison analysis. BBC, based on information theory, provides a new phylogenomic methodology without alignment in postgenome informatics.

Acknowledgements

This work was supported by the National High-Tech Research and Development Program (863 Program) of China (No. 2002AA231071), the Natural Science Foundation of China (No. 60671018; 60121101). We gratefully acknowledge the support of K.C. Wong Education Foundation, Hong Kong.

References

Boore, J.L. and Brown, W.M. (1998) 'Big trees from little genomes: mitochondrial gene order as a phylogenetic tool', *Curr. Opin. Genet. Dev.*, Vol. 8, pp.668–674.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G. and Thompson, J.D. (2003) 'Multiple sequence alignment with the clustal series of programs', *Nucleic Acids Res.*, Vol. 31, pp.3497–3500.

Grasso, C. and Lee, C. (2004) 'Combining partial order alignment and progressive multiple sequence alignment increases alignment speed and scalability to very large alignment problems', *Bioinformatics*, Vol. 20, pp.1546–1556.

Henz, S.R., Huson, D.H., Auch, A.F., Nieselt-Struwe, K. and Schuster, S.C. (2005) 'Whole-genome prokaryotic phylogeny', *Bioinformatics*, Vol. 21, pp.2329–2335.

Hofmann, H., Pyrc, K., van der Hoek, L., Geier, M., Berkhout, B. and Pohlmann, S. (2005) 'Human coronavirus NL63 employs the severe acute respiratory syndrome coronavirus receptor for cellular entry', *Proc. Natl. Acad. Sci. USA*, Vol. 102, pp.7988–7993.

Huson, D.H. and Steel, M. (2004) 'Phylogenetic trees based on gene content', *Bioinformatics*, Vol. 20, pp.2044–2049.

Jeanmougin, F., Thompson, J.D., Gouy, M., Higgins, D.G. and Gibson, T.J. (1998) 'Multiple sequence alignment with Clustal X', *Trends Biochem. Sci.*, Vol. 23, pp.403–405.

Lee, C., Grasso, C. and Sharlow, M.F. (2002) 'Multiple sequence alignment using partial order graphs', *Bioinformatics*, Vol. 18, pp.452–464.

Liu, Z.H., Jiao D. and Sun, X. (2005) 'Classifying genomic sequences by sequence feature analysis', *Genomics Proteomics Bioinformatics*, Vol. 3, pp.201–205.

Liu, Z.H. and Sun, X. (2007) 'Informational structure of agrobacterium tumefaciens C58 genome', *Lecture Notes in Bioinformatics*, Vol. 4689, pp.153–161.

Liu, Z.H., Liu, H.D., Li, J.R., Sun, X. and Jiao, D. (2007) 'Base-base correlation: a novel sequence feature and its applications', *Bioinformatics and Biomedical Engineering*, pp.370–373.

Liu, Z.H., Meng J.H. and Sun, X. (2008) 'A novel feature-based method for whole genome phylogenetic analysis without alignment: application to HEV genotyping and subtyping', *Biochemical and Biophysical Research Communications*, Vol. 368, pp.223–230.

Moreira, D. and Philippe, H. (2000) 'Molecular phylogeny: pitfalls and progress', *Int. Microbiol.*, Vol. 3, pp.9–16.

Philippe, H. and Laurent, J. (1998) 'How good are deep phylogenetic trees?', *Curr. Opin. Genet. Dev.*, Vol. 8, pp.616–623.

Raphael, B., Zhi, D., Tang, H. and Pevzner, P. (2004) 'A novel method for multiple alignment of sequences with repeated and shuffled elements', *Genome Res.*, Vol. 14, pp.2336–2346.

Rota, P.A., Oberste, M.S., Monroe, S.S., Nix, W.A., Campagnoli, R., Icenogle, J.P., Penaranda, S., Bankamp, B., Maher, K., Chen, M.H., Tong, S., Tamin, A., Lowe, L., Frace, M., DeRisi, J.L., Chen, Q., Wang, D., Erdman, D.D., Peret, T.C., Burns, C., Ksiazek, T.G., Rollin, P.E., Sanchez, A., Liffick, S., Holloway, B., Limor, J., McCaustland, K., Olsen-Rasmussen, M., Fouchier, R., Gunther, S., Osterhaus, A.D., Drosten, C., Pallansch, M.A., Anderson, L.J. and Bellini, W.J. (2003) 'Characterization of a novel coronavirus associated with severe acute respiratory syndrome', *Science*, Vol. 300, pp.1394–1399.

Snel, B., Bork, P. and Huynen, M.A. (1999) 'Genome phylogeny based on gene content', *Nat. Genet.*, Vol. 21, pp.108–110.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. and Higgins, D.G. (1997) 'The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools', *Nucleic Acids Res.*, Vol. 25, pp.4876–4882.

Versteeg, R., van Schaik, B.D., van Batenburg, M.F., Roos, M., Monajemi, R., Caron, H., Bussemaker, H.J. and van Kampen, A.H. (2003) 'The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes', *Genome Res.*, Vol. 13, pp.1998–2004.

Woo, P.C., Lau, S.K., Huang, Y., Tsoi, H.W., Chan, K.H. and Yuen, K.Y. (2005) 'Phylogenetic and recombination analysis of coronavirus HKU1, a novel coronavirus from patients with pneumonia', *Arch. Virol.*, Vol. 150, pp.2299–2311.