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.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
WSE, a new sequence distance measure based on word frequencies

Jun Wang\textsuperscript{a,b}, Xiaoqi Zheng\textsuperscript{a,*}

\textsuperscript{a} Department of Applied Mathematics, Dalian University of Technology, Dalian 116024, PR China
\textsuperscript{b} College of Advanced Science and Technology, Dalian University of Technology, Dalian, Liaoning 116024, PR China

\textbf{Abstract}

In this article, we present a new distance metric, the Weighted Sequence Entropy (WSE), based on the short word composition of biological sequences. As a revision of the classical relative entropy (RE), our metric (1) works equivalently with RE in the case of small k; (2) avoids the degeneracy when some word types are absent in one sequence but not in the other. Experiments on 25 viruses including SARS-CoVs show that our method and RE give exactly the same phylogenetic tree when word length $k < 3$. When $k > 3$, our method still works and gets convergent phylogenetic topology but the RE gives degenerate results.

\section{Introduction}

The elucidation of the evolutionary history of different species is a major concern to biological science. Early approaches to deal with it were mainly based on the alignment of a gene or protein sequence. But on one hand, “different genes may tell different stories”—unequal mutational rates of different genes or different pieces of genes and lateral gene transfer are widely detected, so a single gene or protein sequence usually does not have enough information to determine its phylogenetic position. On the other hand, traditional alignment methods are computationally intensive and meaningless to whole genome comparison because each genome has its own genes and gene order. In a word, there is an urgent need to develop new phylogenetic methods utilizing the ever-increasing genome data.

In recent years, several alignment-free phylogenetic methods have been proposed [1–15]. Among the earliest attempts is the gene content method [1,2], which utilizes the ratio between the number of genes two species share and the total number of genes. This method is very simple and can be applied to long genomes. But it fails when species are closely related to each other, e.g., mt genome sequences of some placental mammals used by Cao et al. [3]. Because these genomes share exactly the same genes and gene order. Similar work along this line is the gene order method [4], which also fails in the above case. Besides, some methods based on graphical representations of DNA sequences are outlined [5–10], which usually translate a DNA sequence to a set of plots in 2D/3D space, and use some graphical invariants to characterize this sequence. These methods provide a simple way of viewing, sorting and comparing various gene structures.

Another widely studied method is based on the short word composition of biological sequences. Observing that relative abundances of all dinucleotide are remarkably constant across the genome, Karlin et al. [11–13] proposed the genome signature consisting of the array of dinucleotide relative abundances $p_{xy} = f_{xy}/f_{xx}$ extended over all dinucleotides, where $f_{xy}$ is the frequency of nucleotides $x$ and $f_{xx}$ is the frequency of dinucleotide $xy$. Alternatively, relative abundance of tri-, tetra- or even $k$-nucleotides can also be calculated. Besides phylogenetic inference, these signatures have been used in several applications, e.g., assessment of general relatedness of genomes [12], genetic sequence classification [16], clustering of genes from different genomes [17], and recognition of species-specific sequence blocks [18].

After counting frequencies of all $k$-nucleotides (or $k$-mer, $k$-word, etc.) in a DNA or protein sequence, we can represent this sequence as a vector in the $4^k$ or $20^k$-dimensional Euclidean space. Since the frequency of a $k$-nucleotide measures the probability of its occurrence in the given sequence, it is intuitive to use the relative entropy between two frequency vectors as a measure of distance. Let $p_{ij}$ and $p_{ik}$ be the observed frequencies of the $i$-th $k$-nucleotide in sequences A and B, respectively. The relative entropy [or Kullback–Leibler distance] [19] is defined as

$$\text{RE}(A, B) = \sum_i p_{ij} \log \frac{p_{ij}}{p_{ik}}$$

(1)
The relative entropy is the inefficiency of assuming that the distribution is \( p_0 \) when the true distribution is \( p_\pi \). It does not satisfy the symmetry condition and the triangle inequality, so it is not strictly a distance measure. The relative entropy, however, satisfies many important and useful mathematical properties. For example, it is a convex function of \( p_\pi \), always nonnegative, and vanishes if and only if \( p_\pi = p_0 \) for all possible \( i \).

Apparently, vector representations of biological sequences will lose some information in the case of small \( k \), e.g., only nucleotide composition is known when \( k = 1 \). More information (especially the sequence-order information) will be captured if we consider higher-order words. But unfortunately, as the word order increases, some word types in one sequence may be absent in the other sequence. This leads to the degenerate result when Formula (1) is used. Meanwhile, it is widely detected that some “word spaces” are not saturated for an individual genome, especially for protein sequences. A research of the 101,602 protein sequences in SWISS-PROT database Rel. 40 (2000) showed that all these proteins have taken only less than 26% of the 6-aminoacids strings. In particular, EcoliK has taken less than 2% and Mycge less than 0.3% of the 6-aminoacids strings [20].

Traditionally, in order to avoid the degeneracy, one can unite some \( k \)-words (i.e., words with frequencies up to a particular threshold) into one event. However, this will also be accompanied by a loss of sequence information. In this paper, we give a new distance metric to overcome the degeneracy. As a revision of the classical relative entropy, this metric works equivalently to the relative entropy in phylogenetic inference when \( k \) is small, and avoids the degeneration in the case of large \( k \). In the main text, we give the mathematical description and a simple implementation of this metric. As to its application, we construct phylogenetic trees of 25 viruses including SARS-CoVs.

2. Methods

2.1. Frequencies of \( k \)-words

Given a DNA or protein sequence of length \( L \), we count the occurrences of words of length \( k \). Partial overlap of words is allowed. There are in total \( 4^k \) possible words for DNA sequences and \( 20^k \) words for protein sequences. The frequency of a word \( x_1 x_2 \ldots x_k \) is obtained from its occurrence \( c(x_1 x_2 \ldots x_k) \) by dividing the total number of words \( (L-k+1) \) in the given sequence

\[
 f(x_1 x_2 \ldots x_k) = \frac{c(x_1 x_2 \ldots x_k)}{L-k+1}.
\]

2.2. Subtraction of random background

The above frequency reflects both the results of natural selection and random mutations [20]. In order to highlight the selective diversification of sequence composition, one should subtract the background from the simple frequency results by dividing the corresponding expected value:

\[
 p_{\pi}(x_1 x_2 \ldots x_k) = \frac{f(x_1 x_2 \ldots x_k)}{f_{\exp}(x_1 x_2 \ldots x_k)},
\]

where \( f_{\exp}(x_1 x_2 \ldots x_k) \) is the background frequency of word \( x_1 x_2 \ldots x_k \). There are many approaches to estimate background frequency of a word in a sequence. In the present paper, we use the simplest which is based on the Markov model of order 0 [11–13,21]. For example, the trinucleotide ACT has the expected frequency \( f_{\exp}(ACT) = f(A)/C/f(T) \). More elaborate estimations, e.g., models of higher-order Markov process, can be defined accordingly.

Note that \( p_{\pi} \) is actually not a probability distribution since the sum of its components is not 1. We normalize it here by dividing the sum: \( p' = p_{\pi}/\sum p_{\pi} \), for each word \( i \) of length \( k \).

2.3. Weighted sequence entropy and distance matrix

In order to overcome the degeneracy, we replace \( p_{\pi} \) in Formula (1) by \( (p_{\pi}^2 + p_b^2)/2 \), then

\[
 RE_1(A, B) = \sum_i p_i^2 \log_2 \frac{2p_{\pi}^i}{p_{\pi}^i + p_b^i}. 
\] (2)

Our remedy is meaningful since \((p_{\pi}^2 + p_b^2)/2\) is also a distribution. Actually, Formula (2) can be treated as the relative entropy between \( p_{\pi} \) and the arithmetical average value of distributions \( p_{\pi} \) and \( p_b \). Here \( RE_1(A, B) \neq RE_1(B, A) \). In order to ensure the symmetry condition, we take the sum

\[
 D_1(A, B) = RE_1(A, B) + RE_1(B, A)
\]

\[
 = \sum_i p_i^2 \log_2 \frac{2p_{\pi}^i}{p_{\pi}^i + p_b^i} + \sum_i p_i^2 \log_2 \frac{2p_b^i}{p_{\pi}^i + p_b^i}
\]

\[
 = \sum_i p_i^2 \log_2 p_{\pi}^i + \sum_i p_i^2 \log_2 p_b^i - 2 \sum_i \frac{p_{\pi}^i + p_b^i}{2} \log_2 \frac{p_{\pi}^i + p_b^i}{2}.
\] (3)

Recall that Shannon’s information entropy

\[
 H(A) = -\sum_i p_i^2 \log_2 p_i
\]

and write

\[
 H\left(\frac{A + B}{2}\right) = -\sum_i \frac{p_{\pi}^i + p_b^i}{2} \log_2 \frac{p_{\pi}^i + p_b^i}{2}
\]

as the entropy of arithmetical average value of distributions \( p_{\pi} \) and \( p_b \). Then \( D_1(A, B) \) can be rewritten as

\[
 D_1(A, B) = 2H\left(\frac{A + B}{2}\right) - H(A) - H(B). \] (4)

Alternatively, we can give more elaborate weights to \( p_{\pi}^i \) and \( p_b^i \), e.g., the sequence lengths \( L_A \) and \( L_B \), or their square roots \( \sqrt{L_A} \) and \( \sqrt{L_B} \) (the accuracy in estimation of probability in multinomial distribution is proportional to square root of the corresponding sequence length). This gives another two relative entropy functions

\[
 RE_2(A, B) = \sum_i p_i^2 \log_2 \frac{L_A + L_B}{L_A p_{\pi}^i + L_b p_b^i};
\] (5)

\[
 RE_3(A, B) = \sum_i p_i^2 \log_2 \frac{\sqrt{L_A} + \sqrt{L_B}}{\sqrt{L_A} p_{\pi}^i + \sqrt{L_b} p_b^i};
\] (6)

Their corresponding distance measures can be defined and calculated similarly,

\[
 D_2(A, B) = (L_A + L_b)H\left(\frac{L_A + L_B}{L_A + L_b}\right) - L_A \cdot H(A) - L_B \cdot H(B); \] (7)

\[
 D_3(A, B) = \left(\sqrt{L_A} + \sqrt{L_B}\right)H\left(\frac{\sqrt{L_A} + \sqrt{L_B}}{\sqrt{L_A} + \sqrt{L_B}}\right) - \sqrt{L_A} \cdot H(A) - \sqrt{L_B} \cdot H(B). \] (8)

Since \( D_1, D_2 \) and \( D_3 \) are all linear combinations of Shannon entropies, we call them the Weighted Sequence Entropy (WSE). According to their definitions, these distances can avoid the degeneracy as their upper bounds are \( 2, L_A + L_b \) and \( \sqrt{L_A} + \sqrt{L_B} \), respectively.

To illustrate the virtues of WSE, we give an example as follows. The first exon of \( \beta \)-globin gene of Lemur is \( S = ATGACTTTGCTG \ldots \).
AGTGTGAGGAGAATGTCATGCACCTCTTCGCGCGGAGTCGAT GTAGAGAAATTGGTGGGAGCGCTTGGCCAG. There is only one dinucleotide TA in this sequence. If a point mutation which changes T in the only dinucleotide TA to C occurs (denotes the new sequence by $S'$), then the relative entropy between $S$ and $S'$ is $\infty$, which contradicts the biological meaning. While our methods give the values 0.0245821, 2.26155 and 0.23578 instead.

3. Results

3.1. Phylogenetic trees of 25 viruses

The outbreak of severe acute respiratory syndrome coronavirus (SARS-CoV) in 2003 has had a tremendous impact on worldwide public health and healthcare systems [22,23]. Coronaviruses are members of a family of enveloped viruses that replicate in the cytoplasm of animal host cell. According to the type of the host, coronaviruses isolated previously can be classified into three groups. Groups I and II contain mammalian viruses, whereas group III contains only avian viruses. After genome sequencing of some SARS-CoVs, many efforts have been made to identify the phylogenetic position of SARS-CoVs in coronavirus tree using molecular data. However, this is still a controversial topic—alignment-based methods showed that SARS-CoVs are not closely related to any previously isolated groups and form a new group [24,25]; maximum likelihood tree built from a fragment of the spike protein preferred SARS-CoVs clustering with group II coronaviruses (murine hepatitis virus and rat coronavirus) [26]; while an information-based method, which makes use of the whole genome sequences, indicated that SARS-CoVs are close to the group I coronaviruses rather than form a new group [27].

In the present work, we select 25 complete virus genomes: 12 coronaviruses from the three isolated typical groups, 12 SARS-CoV strains, and a torovirus, which serves as the outgroup for coronaviruses [28] (Table 1). Pairwise distance matrices for these 25 sequences are calculated using our distances ($D_1$, $D_2$ and $D_3$) and the classical relative entropy. Notice that RE does not satisfy the symmetry condition. We use the sum of RE($A, B$) and RE($B, A$) instead. Then, phylogenetic trees are built from these matrices using the UPGMA program in the PHYLPACK package [29]. Finally, rooted phylogenetic trees are drawn by TREEVIEW program [30], and some of them are shown in Figs. 1 and 2.

Our three WSE distances ($D_1$, $D_2$ and $D_3$) perform similarly for all values of $k$ (from 2 to 7). They get nearly the same phylogeny except for the length of some branches. This is under our expectation since these genomes have no significant difference in length (except for torovirus).

Trees constructed by RE and our distances are of the same topology at the word lengths $k = 2$ and 3 (these common topologies are shown in Fig. 1). But when $k \geq 4$, the relative entropy fails as some words are absent in these genomes (Table 2). Explicitly,

1. Phylogenetic result got for $k = 2$ deviates significantly from the commonly accepted one. This phylogeny prefers murine hepatitis viruses as the outgroup (Fig. 1a);
2. The word order $k = 3$ seems to make a remarkable improvement (Fig. 1b): three typical groups of coronaviruses, and 12 SARS-CoVs, all cluster accordingly. Two big branches are formed: one is the group I coronaviruses and all SARS genomes, the other is the torovirus and group II and III coronaviruses. But this topology also does not support the outgroup status of torovirus.

With the definition of WSE, we can study higher-order words and check whether they are suitable to estimate phylogenies. WSE trees at word lengths $k = 4, 5, 6, 7$ are shown in Fig. 2. As can be seen from Figs. 1 and 2, topologies of evolutionary trees converge with the increase of $k$, and become stable at $k = 4, 5$ and 6. These stable trees maintain the main topologies of Fig. 1b, and also successfully support the outgroup status of torovirus relative to all coronaviruses. When $k \geq 6$, the average frequency of words is too small for the present statistical analysis. In conclusion, we prefer $k = 6$ giving the best phylogeny, that is, all 12 SARS-CoV strains are grouped together and form a new fourth group, which is distinctly related to the group I coronaviruses (TGEV, explicitly).

3.2. Comparisons with some classical methods

Besides the relative entropy, there are a great many of alignment-free comparisons on the basis of short word composition. In this part, we will list some of them and compare them with our method through phylogenetic experiments on the same data set.

1. Euclidean distance [31,32]:
$$d_{\text{euclid}}(A, B) = \sqrt{\sum (f_i - f'_i)^2},$$
where $f_i$ and $f'_i$ are the frequencies of the ith word in sequences A and B, respectively.
2. Linear correlation coefficient [33]:
$$d_{\text{ccc}}(A, B) = 1 - \frac{\sum (f_i - \overline{f})(f'_i - \overline{f}')}{\sqrt{\sum (f_i - \overline{f})^2 \sum (f'_i - \overline{f}' )^2}},$$
where
$$\overline{f} = \frac{1}{4} \sum f_i, \quad \overline{f}' = \frac{1}{4} \sum f'_i.$$
3. Cosine function [20,34]:
$$d_{\text{cos}}(A, B) = \frac{1}{2} \left( 1 - \frac{\sum p_i p'_i}{\sqrt{\sum (p_i)^2 \sum (p'_i)^2}} \right),$$

Table 1: Coronaviruses and a torovirus used to construct phylogenetic tree

| No. | Accession | Abbreviation | Genome | Group | Length (nt) |
|-----|-----------|--------------|--------|-------|-------------|
| 1   | NC_002654 | HCoV-229E    | Human coronavirus 229E | I     | 27317       |
| 2   | NC_002306 | TGEV         | Transmissible gastroenteritis virus | I     | 28586       |
| 3   | NC_003416 | PEDV         | Porcine epidemic diarrhea virus | I     | 28033       |
| 4   | U00735    | BCoVm       | Bovine coronavirus strain | II    | 31032       |
| 5   | AF391542  | BCoV        | Bovine coronavirus isolate BCoV-L | II    | 31028       |
| 6   | AF220295  | BCoVQ       | Bovin coronavirus strain Quebec | II    | 31100       |
| 7   | NC_003045 | BCoV         | Bovine coronavirus | II    | 31028       |
| 8   | AF208067  | MHVM         | Murine hepatitis virus strain ML-10 | II    | 31233       |
| 9   | AF201929  | MHV2         | Murine hepatitis virus strain 2 | II    | 31276       |
| 10  | AF208066  | MHVP         | Murine hepatitis virus strain Penn | II    | 31112       |
| 11  | NC_001846 | MHV          | Murine hepatitis virus | II    | 31357       |
| 12  | NC_001451 | IBV          | Avian infectious bronchitis virus | III   | 27608       |
| 13  | AV278488  | BJ01         | SARS coronavirus BJ01 | —     | 29725       |
| 14  | AV278741  | Urbani       | SARS coronavirus Urbani | —     | 29727       |
| 15  | AV278491  | HKU-39849    | SARS coronavirus HKU-39849 | —     | 29742       |
| 16  | AV278554  | CUHK-W1      | SARS coronavirus CUHK-W1 | —     | 29736       |
| 17  | AV282752  | CUHK-Su10    | SARS coronavirus CUHK-Su10 | —     | 29736       |
| 18  | AV283794  | SIN2500      | SARS coronavirus SIN2500 | —     | 29711       |
| 19  | AV283795  | SIN2677      | SARS coronavirus SIN2677 | —     | 29705       |
| 20  | AV283796  | SIN2679      | SARS coronavirus SIN2679 | —     | 29711       |
| 21  | AV283797  | SIN2748      | SARS coronavirus SIN2748 | —     | 29706       |
| 22  | AV283798  | SIN2774      | SARS coronavirus SIN2774 | —     | 29711       |
| 23  | AV291451  | TW1          | SARS coronavirus TW1 | —     | 29728       |
| 24  | NC_004718 | TOR2         | SARS coronavirus | —     | 29751       |
| 25  | X52374    | ETov         | Equine torovirus | —     | 7920        |
Fig. 1. The common phylogenetic trees of 25 viruses constructed by the relative entropy and our distances ($D_k$). For word orders $k = 2$ and 3, the corresponding trees are listed as (a) and (b). When $k \geq 4$, the relative entropy method fails as some words are absent in these genomes.

Fig. 2. Phylogenetic trees of 25 viruses for different values of $k$ with pairwise distances evaluated by our distance (for $k$ increasing from 4 to 7, the respective trees are shown as (a), (b), (c) and (d)). In this case, the classical relative entropy method gets degenerate result.
where $p_i = (f_i^A - F_i)/F_A$, and $F_A$ is the expect frequency of word $i$ in sequence $A$ estimated from the frequencies of appropriate shorter subwords under a Markov model of order $k - 1$.

(4) Information-based similarity index [35]:

$$d_{ISI} = \frac{1}{4n - 1} \sum_i \left| R_A^i - R_B^i \right| \frac{H_A^i + H_B^i}{\sum_i (H_A^i + H_B^i)},$$

where $H_A^i = -f_i^A \log f_i^A$ is the Shannon entropy for word $i$ in sequences $A$. $R_A^i$ and $R_B^i$ represent the ranks of word $i$ in sequences $A$ and $B$.

We repeated our experiment using the above classical distances for different values of $k$, and some phylogenetic trees (at word length $k = 6$) are shown in Figs. 3–5.

Phylogenetic tree obtained using Euclidean distance $d_{Eucli}$ as the dissimilarity metric is shown in Fig. 3. This tree is mainly identical to our tree, i.e., torovirus stays outside of all coronaviruses, and SARS-CoVs are closer to the transmissible gastroenteritis virus (TGEV). But an obvious default in this tree is that it fails to cluster the three group I coronaviruses: HCoV-229E, PEDV and TGEV. Fig. 4 lists the phylogenetic tree constructed by $d_{COS}$. This topology supports SARS-CoVs and group II as a closer pair, which is in accordance with the result of Liò and Goldman using a fragment of the spike protein [26]. However, this tree also fails to cluster the group I coronaviruses. In Fig. 5, we list the tree built by linear correlation coefficient between pairwise frequency vectors. This tree perfectly clusters species within each typical group, and supports the outgroup status of SARS-CoVs relative to other coronaviruses. But it fails to identify the outgroup status of torovirus relative to coronaviruses.

4. Discussion

Phylogenetic analysis from $k$-word composition is a good alternative approach to the traditional alignment methods. It has relatively low computational complexity, and does not suffer from genetic rearrangements and transposon activity, which serve as common ways of genome evolution. In most cases, biological sequences are represented as occurrence or composition vectors in a high dimensional Euclidean space, and phylogenetic results are high reliable on the quality of this vector representation. In order to capture more sequence information and obtain a better vector representation, the suitable word length is of critical importance. According to the information theory, this length reflects the balance between noise and information—some information may be lost if one uses overshort words, while noise will dominate when long words are considered. Moreover, suitable length of word is often species-specific. In a word, a well-defined distance measure which can be applied to any word length is of great importance.

Relative entropy is one of the classical approaches to measure the dissimilarity between two composition vectors. However, as word order increases, this approach will get degenerate results due to the absent of some word types. In the present paper, we propose three revisions of the classical relative entropy (RE). These modifications perform equally with RE in phylogenetic inferences in the case of short words. But when long words are considered, our method still works and gets convergent phylogenetic topologies (in such case, RE gets degenerate results as some word types are absent). Through constructing phylogenetic trees using WSE and some classical distance measures, we find that, as far as the set of 25 virus sequences is concerned, our distances are not inferior to those classical methods. According to our tree, all SARS-CoVs form a group, which is more related to the group I coronavirus: TGEV. This result is mainly in accordance with the phylogeny given by Yang et al. [27], and is also supported by the experimental evidence, which showed that group I coronaviruses (transmissible gastroenteritis virus, TGEV) specific antibodies are able to recog-

Table 2
Numbers of words that are absent in some species for different values of $k$

| Species | BCoV | HCoV-229E | IBV | MHV | PEDV | SARS | TGEV |
|--------|------|----------|-----|-----|------|------|------|
| $k = 5$ | 2    | 6        | 6   | 1   | 1    | 1    | 9    |
| $k = 6$ | 288  | 384      | 338 | 100 | 207  | 259  | 408  |
| $k = 7$ | 5607 | 6205     | 5861| 4456| 5215 | 5150 | 6049 |

Fig. 3. Phylogenetic tree of 25 viruses constructed using Euclidean distance as the dissimilarity metric.

Fig. 4. Phylogenetic tree of 25 viruses constructed using $d_{COS}$ as the dissimilarity metric.
nize antigens in SARS-CoV infected cultured cells [37]. We hope our modification can be served as an alternative approach to avoid the degeneracy, and has potential to study information expressed especially in the case that sequences have significant differences in length, we prefer the revision (15) giving relatively competitive phylogenetic results.

In order to further assure our approach, and evaluate the performances of different WSE, we will do the following experimental and theoretical discussions. Experimentally, we will "generate" phylogenetic trees under different evolutionary models, and reconstruct the phylogenetic topologies from their corresponding external nodes (OTU) using our distances and RE. Then performance of different WSE can be evaluated by checking the consistency of these phylogenetic topologies. Theoretically, we hypothesize that order relations between pairwise sequence distances calculated by RE and WSE are in accordance, i.e., \( \text{WSE}(A,B) < \text{WSE}(C,D) \) if and only if \( \text{RE}(A,B) < \text{RE}(C,D) \), for any sequences A, B, C and D (proof will be presented in our future paper). This phenomenon, on one hand, can partially explain the equivalence of WSE and RE in phylogenetic inferences in the case of short words, and on the other hand, can help to analyze the performances of different weight methods.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China. The authors also thank the editor and the anonymous reviewers for their comments and suggestions.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.mbs.2008.06.001.

References

[1] S.T. Fitz-Gibbon, C.H. House, Whole genome-based phylogenetic analysis of free-living microorganisms, Nucleic Acids Res. 27 (1999) 4218.
[2] B. Snell, P. Bork, M.A. Haynen, Genome phylogeny based on gene content, Nat. Genet. 21 (1999) 108.
[3] Y. Cao, J. Adachi, A. Janke, S. Pääbo, M. Hasegawa, Phylogenetic relationships among eutherian orders estimated from inferred sequences of mitochondrial proteins: instability of a tree based on a single gene, J. Mol. Evol. 39 (1994) 319.
[4] J.I. Boore, W.M. Brown, Big trees from little genomes: mitochondrial gene order as a phylogenetic tool,Curr. Opin. Genet. Dev. 8 (1998) 668.
[5] M. Randic, M. Vracko, A. Nandy, S.C. Basak, J. Chem. Inf. Comput. Sci. 40 (2000) 1235.
[6] M. Randic, M. Vracko, N. Lers, D. Plavsic, Big trees from little genomes: mitochondrial gene order as a phylogenetic tool, Chem. Phys. Lett. 371 (2003) 202.
[7] E. Hamori, J. Ruskin, H. curves, a novel method of representation of nucleotide series especially suited for long DNA sequences, J. Biol. Chem. 258 (1983) 1318.
[8] E. Hamori, Novel DNA sequence representations, Nature 314 (1985) 585.
[9] M.A. Gates, Simpler DNA sequence representations, Nature 316 (1985) 219.
[10] A. Nandy, Graphical representation of long DNA sequences, Curr. Sci. 66 (1994) 309.
[11] S. Karlin, J. Mrázek, Compositional differences within and between eukaryotic genomes, Proc. Natl. Acad. Sci. USA 94 (1997) 10227.
[12] S. Karlin, I. Ladunga, Compositions of eukaryotic genomic sequences, Proc. Natl. Acad. Sci. USA 91 (1994) 12832.
[13] A. Campbell, J. Mrázek, S. Karlin, Genome signature comparisons among prokaryotes, plasmids, and mitochondrial DNA, Proc. Natl. Acad. Sci. USA 96 (1999) 9184.
[14] G.W. Stuart, K. Moffett, S. Baker, Integrated gene and species phylogenies from unaligned whole genome protein sequences, Bioinformatics 18 (2002) 100.
[15] P. Chaudhuri, S. Das, Swords: a statistical tool for analyzing large DNA sequences, J. Biosci. 27 (2002) 1.
[16] N.A. Chuzhanova, A.J. Jones, Feature selection for genetic sequence classification, Bioinformatics 14 (1998) 139.
[17] H. Nakashima, M. Ota, K. Nishikawa, T. Ooi, Genes from nine genomes are separated into their organisms in the dinucleotide composition space, DNA Res. 5 (1998) 251.
[18] J. Qi, B. Wang, B.L. Hao, Whole proteome prokaryote phylogeny without sequence alignment: a k-string composition approach, J. Mol. Evol. 58 (2004) 1.
[19] C. Nikolaou, Y. Almirantis, Measuring the coding potential of genomic sequences through a combination of triplet occurrence patterns and rny preference, J. Mol. Evol. 59 (2004) 399.
[20] N. Lee, D. Hui, A. Wu, P. Chan, L. Cameron, G.M. Joynt, A. Ahuja, M.Y. Chung, C.B. Leung, K.F. To, S.F. Lui, C.C. Szeto, S. Chung, J.J.Y. Sun, A major outbreak of severe acute respiratory syndrome in Hong Kong, N. Engl. J. Med. 348 (2003) 1986.
[21] S.M. Poutanen, D.E. Low, B. Henry, S. Finkelstein, D. Rose, K. Green, R. Tellier, R. Draker, D. Adachi, M. Ayers, A.K. Chan, D.M. Skowronski, I. Salit, A.E. Simor, A.S. Slutsky, P.W. Doyle, M. Krajden, M. Petric, R.C. Brunham, A.J. McGeer, Identification of severe acute respiratory syndrome in Canada, N. Engl. J. Med. 349 (2003) 1995.
[22] M.A. Marra, S.J. Jones, C.R. Astell, et al., The genome sequence of the SARS-associated coronavirus, Science 300 (2003) 1399.
[23] P. Lio, M.S. Osterri, S.S. Monroe, et al., Characterization of a novel coronavirus associated with severe acute respiratory syndrome, Science 303 (2003) 1394.
[24] N. Lee, D. Hui, A. Wu, P. Chan, L. Cameron, G.M. Joynt, A. Ahuja, M.Y. Chung, C.B. Leung, K.F. To, S.F. Lui, C.C. Szeto, S. Chung, J.J.Y. Sun, A major outbreak of severe acute respiratory syndrome in Hong Kong, N. Engl. J. Med. 348 (2003) 1986.