Journal of Theoretical Biology 336 (2013) 52–60

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Study of LZ-word distribution and its application for sequence comparison

Qi Dai [a,*], Zhaofang Yan [a], Zhuoxing Shi [a], Xiaoqing Liu [b], Yuhua Yao [a], Pingan He [a]

[a] College of Life Sciences, Zhejiang Sci-Tech University, Hangzhou 310018, People's Republic of China
[b] College of Sciences, Hangzhou Dianzi University, Hangzhou 310018, People's Republic of China

HIGHLIGHTS

- With the components' length in mind, we revised Lempel–Ziv complexity.
- We first investigated the whole distribution of LZ-words.
- We defined transition and extension operations among the revised LZ-word sets.
- We calculated numerical characteristics of the sorted union LZ-word set.

ARTICLE INFO

Article history:
Received 12 April 2013
Received in revised form
6 July 2013
Accepted 10 July 2013
Available online 19 July 2013

Keywords:
Lempel–Ziv complexity
Word set
Set operation
Phylogenetic analysis

ABSTRACT

Lempel–Ziv complexity has been widely used for sequence comparison and achieved promising results, but until now components' distribution in exhaustive history has not been studied. This paper investigated the whole distribution of LZ-words and presented a novel statistical method for sequence comparison. With the components' length in mind, we revised Lempel–Ziv complexity and obtained various sets of LZ-words. Instead of calculating the LZ-words' contents, we defined a series of set operations on LZ-word set to compare biological sequences. In order to assess the effectiveness of the proposed method, we performed two sets of experiments and compared it with alignment-based methods.

1. Introduction

With high-throughput production of gene and protein sequences, the rate of addition of new sequences to the databases increased exponentially. Such a collection of sequences does not, by itself, increase the scientist's understanding of the biology of organisms. However, comparing new sequences to those with known functions is a key way of understanding the biology of an organism.

Many methods have been proposed for sequence comparison. They can be categorized into two classes. One is alignment-based methods, in which dynamic programming is used to find an optimal alignment by assigning scores to different possible alignments and picking the alignment with the highest score. Several alignment-based algorithms have been proposed such as global alignment, local alignment, with or without overlap (Gotoh, 1982; Needleman and Wunsch, 1970; Smith and Waterman, 1981; Randic, 2013a, 2013b). Waterman (Waterman, 1995) and Durbin et al. (Durbin et al., 1998) provided comprehensive reviews about this method. However, the search for optimal solutions using sequence alignment has problems in: (i) computationally load with large biological databases and (ii) choice of the scoring schemes (Pham and Zuegg, 2004; Vinga and Almeida, 2003). Therefore, the emergence of research into the second class, alignment-free methods, is apparent and necessary to overcome critical limitations of alignment-based methods.

Up to now, many alignment-free methods have been proposed, but they are still in the early development compared with alignment-based methods. One of the most widely used alignment-free approaches is statistical model, in which each sequence is first mapped into an m-dimensional vector according to its k-word frequencies, and sequence similarity can then be measured by distance measures, such as Euclidean distance (Blaisdell, 1986), Pearson's correlation coefficient (Fichant and Gautier, 1987), Kullback–Leibler discrepancy (Wu et al., 2001), Cosine distance (Stuart et al., 2002) among their corresponding vectors. Recently, Ewens and Grant (2005) studied probabilistic properties of words in sequences,

[*] Corresponding author. Tel.: +86 571 86843746; fax: +86 571 86843198.
E-mail address: daiailiu2004@yahoo.com.cn (Q. Dai).

0022-5193/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jtbi.2013.07.008

deducted the exact distributions and evaluated its asymptotic approximations. When the k-words occurring in biological sequence are estimative probabilities rather than frequencies, they are more easily described by more complex models such as Markov model (Pham and Zuegg, 2004; Hao and Qi, 2004; Wu et al., 2006; Apostolico and Denas, 2008), mixed model (Kantorovitz et al., 2007) and Bernoulli model (Lu et al., 2008).

Graphical representation is another widely used alignment-free method. It provides a simple way to view, sort and compare various gene sequences with their intuitive pictures and pattern. Randic et al. gave a comprehensive review on these methods (Randic et al., 2011,2013). In order to facilitate comparison of different biological sequences, they transformed graphical representations into some mathematical objects such as E matrix (Yao and Wang, 2004; Liao and Wang, 2004; Song and Tang, 2005), D/D matrix(Li and Wang, 2003; Yao and Wang, 2004; Liao and Wang, 2004; Song and Tang, 2005), L/L matrix (Randic et al., 2003; Li and Wang, 2003; Yao and Wang, 2004; Liao and Wang, 2004; Song and Tang, 2005) and their "high order" matrices (Yao and Wang, 2004; Song and Tang, 2005). Once a matrix was given, they calculated matrix invariants as descriptors of the sequence, such as the average matrix element, the average row sum, the leading eigen value and the Wiener number. But, for long sequences, these methods become less useful because they require complex repetitive computation to get matrix invariants.

Recently, Out and Sayood introduced Lempel–Ziv (LZ) complexity to compute the distance between two DNA sequences (Out and Sayood, 2003). Because it is based on exact direct repeats, the LZ complexity works well with the small DNA alphabet. Unlike DNA sequences, protein sequences and RNA secondary structures consist of more complex alphabets and structure information, which poses more of a challenge for LZ complexity. So Bacha and Baurain, Liu and Wang, Chen and Zhang presented several strategies in which protein sequences or RNA secondary structures were encoded to a new alphabet prior to computation of the LZ complexity (Bacha and Baurain, 2005; Liu and Wang, 2006; Zhang and Chen, 2010; Zhang and Wang, 2010; Chen and Zhang, 2012). Zhang et al. found that the LZ complexity is strongly correlated with sequence length and proposed a normalized LZ complexity for sequence comparison (Zhang et al., 2009). Taking into account a specific kind of the inexact copy in the text, Li et al. generalized the LZ complexity and proposed a new sequence distance measure for sequence comparison (Li et al., 2010). Liu et al. introduced relative LZ complexity to depict the complexity relationship between two sequences (Liu et al., 2012).

All above LZ-based methods have achieved promising results in biological sequence comparison, but they generally placed a heavy emphasis on the number of components in the exhaustive history, so little attention has been paid to the components themselves. In this paper, we used the proposed revised LZ complexity to obtain a series of LZ-words from the exhaustive history of biological sequences. Based on the LZ-word distributions, we constructed a sorted union LZ-word set from which an indicator sequence was obtained. We then calculated numerical characteristics of the indicator sequence to compare biological sequences. The performance of the proposed method was evaluated by the phylogenetic analysis and comparison with alignment-based method.

2. Method

2.1. LZ-words of DNA sequences

Given a finite alphabet Ω, let U, V and W be sequences over it. $L(U)$ is the length of the sequence U, $U(i)$ is the i-th element of U, and $U(i,\ j)$ is the subsequence of U starting at position i and ending at position j. Here, $U(i,\ j)=\varnothing$, for $i>j$. Concatenating V and W can construct a new sequence $U{=}VW$, in this equation, V is named "a prefix" of U, and U is called "an extension" of V if there exists an integer i such that $V=U(1,\ i)$. An extension $U{=}VW$ of V is reproducible from V denoted by $V\rightarrow U$, if there exists an integer $P{\le}L(V)$ such that $W(k)=U(p+k{-}1)$, for $k{=}1, 2,\ldots, L(W)$. A non-null sequence U is producible from its prefix $U(1,\ j)$, denoted by $U(1,\ j){\Rightarrow}U$, if $U(1,\ j)\rightarrow U(1,\ L(U){-}1)$. For example: 01⇒0100 with $p{=}1$. Note that, the producibility allows for an extra different symbol at the end.

Usually, a DNA primary sequence can be taken as a string of letters A, G, C, and T, which denote the four nucleic acid bases: adenine, guanine, cytosine, and thymine, respectively. Let $S=s_1s_2\ldots s_n$ to be a DNA sequence. To indicate a substring of S that starts at position i and ends at position j, we write $S(i,\ j)$, where is, $S(i,\ j)=s_is_{i+1}\ldots s_j$ for $i{\le}j$. Any sequence S can be built using a production process where at its ith step $S(1,\ h_{i-1}){\Rightarrow}S(1,\ h_i)$, which is described as following:

(1) At the beginning, we had a null-sequence, denoted by $\varnothing$. We then added a prefixs_1 to $\varnothing$ and obtained a new sequence S. If $L(S)>1$, we added a symbol "$\ast$"after $S(1,\ 1)$.
(2) Let a prefix $Q=S(1,\ h_1){\ast}S(h_1+1,\ h_2){\ast}\cdots{\ast}S(h_{m-1}+1,\ h_m)$, checked if $R{=}S(h_m, h_m+1)$ can be reproduced from the sequence $Q=S(1,\ h_m)$. If R could not be reproduced from the set, then joined Q and R to get a new prefix QR, and added a symbol "$\ast$"following QR. If R could be reproduced from the set, then checked again if $R{=}S(h_m, h_m+2)$ can reproduced from the sequence $Q=S(1,\ h_m)$. If so, checked again if $R{=}S(h_m, h_m+3)$ can reproduced from the sequence $Q=S(1,\ h_m)$, $\cdots$ and so on. There two possible cases: in the case $R{=}S(h_m, L(S))$, we ended the procedure and got new prefix $QR=S$; in another case $R{=}S(h_m, h_{m+1})$ cannot be reproduced from the sequence$Q=S(1,\ h_m)$, we got a new prefix QR and added a symbol"$\ast$" behind it.
(3) Repeated the step (2) until produce S.

Instead of focusing on the total number of components in the exhaustive history, we analyzed the components themselves. For convenience, we denoted a component in the exhaustive history as a LZ-word, and all the components in the exhaustive history as a LZ-word set. For example, the LZ-words of S=ATGGTCGGTTTC can be gotten through the following steps, where $\ast$ is used to separate the decomposition component:

- Generate a novel symbol **A**: Ø+A → A.
- Generate a novel symbol **T**: A+T → AT.
- Generate a novel symbol **G**: A∗T+G → A∗T∗G.
- Copy the longest fragment+generate a additional symbol G**T**: A∗T∗G+GT → A∗T∗G∗GT.
- Generate a novel symbol **C**: A∗T∗G∗GT+C → A∗T∗G∗GT∗C.
- Copy the longest fragment+generate a additional symbol GGTT: A∗T∗G∗GT∗C → A∗T∗G∗GT∗C∗GGTT.
- Copy the longest fragment T**C**: A∗T∗G∗GT∗C∗GG∗TT → A∗T∗G∗ GT∗C∗ GG∗TT∗TC.

A, T, G, GT, C, GGTT and TC are the LZ-words of the sequence S. And {A, T, G, GT, C, GGTT, TC} is the LZ-word set of the sequence S.

2.2. Revised LZ-words of DNA sequences

LZ complexity of a sequence is measured by the minimal number of steps required for its synthesis in a certain process. For each step only two operations are allowed in the process: either generating an additional symbol which ensures the uniqueness of each component or copying the longest fragment from the part of a synthesized sequence. When a new decomposition component $S(1,\ h_i)$ is generated, it should be checked whether it is copied from the longest fragment of the $S(1,\ h_{i-1})$. Consequently, the length of LZ-word inevitably becomes large as production process going on. With this problem in mind, we proposed a revised LZ complexity that is described as following:

(4) At the beginning, we had a null-sequence, denoted by$\varnothing$. We then added a prefix s_1 to $\varnothing$ and obtained a new sequence S. If $L(S)>1$, we added a symbol"$*$"after $S(1,\ 1)$.

(5) Let a prefix $Q=S(1,\ h_1)*S(h_1+1,\ h_2)*\cdots*S(h_{m-1}+1,\ h_m)$, checked if $R=S(h_m, h_m+1)$ can be reproduced from the set$\{S(1,\ h_1),\ S(h_1+1,\ h_2),\cdots,S(h_{m-1}+1,\ h_m)\}$. If R could not be reproduced from the set, then joined Q and R to get a new prefix QR, and added a symbol"$*$"following QR. If R could be reproduced from the set, then checked again if $R=S(h_m,$ $h_m+2)$can reproduced from the set$\{S(1,\ h_1),\ S(h_1+1,\ h_2),$ $\cdots,S(h_{m-1}+1,\ h_m)\}$. If so, checked again if $R=S(h_m, h_m+3)$ can reproduced from the set$\{S(1,\ h_1),\ \ S(h_1+1,\ h_2),\cdots,$ $S(h_{m-1}+1,\ h_m)\}$, $\cdots$ and so on. There two possible cases: in the case $R=S(h_m, L(S))$, we ended the procedure and got new prefix $QR=S$; in another case $R=S(h_m, h_{m+1})$ cannot be reproduced from the set $\{S(1,\ h_1),\ S(h_1+1,\ h_2),\cdots,$ $S(h_{m-1}+1,\ h_m)\}$, we got a new prefix QR and added a symbol "$*$" behind it.

(6) Repeated the step (2) until produce S.
Take the above sequence $S=$ATGGTCGGTTTC as an example, we obtained its revised LZ-words through the following steps, where $*$ is used to separate the decomposition component:

- Generate a novel symbol **A**: Ø+A → A.
- Generate a novel symbol **T**: A+T → AT.
- Generate a novel symbol **G**: A$*$T+G → A$*$T$*$G.
- Copy the fragment G+ generate a additional symbol **T**: A$*$T$*$G +GT → A$*$T$*$G$*$GT.
- Generate a novel symbol **C**: A$*$T$*$G$*$GT+C → A$*$T$*$G$*$GT$*$C.
- Copy the fragment G+generate a additional symbol **G**: A$*$T$*$G$*$ GT$*$C → A$*$T$*$G$*$GT$*$C$*$GG.
- Copy the fragment T+generate a additional symbol **T**: A$*$T$*$G$*$ GT$*$C → A$*$T$*$G$*$GT$*$C$*$GG$*$TT.
- Copy the fragment T+generate a additional symbol **C**: A$*$T$*$G$*$ GT$*$C$*$GG$*$TT → A$*$T$*$G$*$GT$*$C$*$ GG$*$ TT$*$TC.

A, T, G, GT, C, GG, TT and TC are revised LZ-words of the sequence S. And {A, T, G, GT, C, GG, TT, TC} is revised LZ-word set of the sequence S. It is interesting to note that the maximum length of the revised LZ-word set is 2, significantly smaller than that of the LZ-word set.

2.3. Operation measure between different revised LZ-word sets

Given a DNA sequence, we can get a revised LZ-word set. Here, we are interested not only in using the revised LZ-word set to numerically characterize the biological sequences, but also in facilitating comparison of biological sequences.

There is a large body of literatures on word statistics, where a sequence is interpreted as a succession of symbols (Reinert et al., 2000). A k-word is a series of k consecutive letters in a sequence. The word statistical analysis consists of counting occurrences of words and calculating their numerical characteristics. The standard approach for counting k-words in a sequence of length m is to use a sliding window of length k, shifting the frame one base at a time from position1 to m-k+1. Instead of counting the LZ-words' content, we analyzed the distribution diversity of revised LZ-words and designed an operation measure to compare biological sequences.

Given two DNA sequences X and Y, we obtained their revised LZ-word sets $RLZSet_X$ and $RLZSet_Y$. We then blended $RLZSet_X$ and $RLZSet_Y$ to compose anther set $RLZSet_{X-Y}$

$$RLZSet_{X-Y}=RLZSet_X \oplus RLZSet_Y.$$

According to the length of revised LZ-words, the $RLZSet_{X-Y}$ set is divided into several mutually exclusive sets $RLZSet^t_{X-Y}$

$$RLZSet_{X-Y}=\bigcup_t RLZSet^t_{X-Y},$$

where

$$RLZSet^t_{X-Y}=\{x \in RLZset_{X-y} | length(x)=t\}.$$

We then lined the elements of the $RLZSet^t_{X-Y}$ set in the lexicographic order and got an ordered $\uparrow RLZSet_{X-Y}$ set

$$\uparrow RLZSet_{X-Y}=\left\{\uparrow RLZSet^1_{X-Y}, \uparrow RLZSet^2_{X-Y}, \uparrow RLZSet^3_{X-Y}, \cdots\right\}.$$

For example, if X=$ATGCGTCGGTCCACCCACGTA$ and Y=$ATCGGT$ $CTGTTACAGACTACG$ are two given DNA sequences, we can get there $\uparrow RLZSet_X$, $\uparrow RLZSet_X$ and $\uparrow RLZSet_{X-Y}$ sets:

$\uparrow RLZSet_X$={A, C, G, T, CA, CC, CT, GT, CAC, GTA, GTC},
$\uparrow RLZSet_Y$={A, C, G, T, AC, AG, CT, GT, ACT, GTC, GTT},
$\uparrow RLZSet_{X-Y}$={A, A, C, C, G, G, T, T, AC, AG, CA, CC, CT, CT, GT, GT, ACT, CAC, GTA, GTC, GTC, GTT}.

Now we focus on the blend degree of two biological sequences. Given any pair of neighboring elements in $\uparrow RLZSet_{X-Y}$ set, there are two possible cases: if one is from $RLZSet_X(RLZSet_X)$ and the other is from $RLZSet_Y(RLZSet_X)$, we suppose there is transition operation (╱, ╲) between them. Otherwise, they may both come from the same set $RLZSet_X(RLZSet_Y)$, we suppose there is extension operation (—) between them. Take above two sequences X and Y for an example, we first listed all the elements of the $\uparrow RLZSet^t_X$ sets in a line with "•" denoting them, and list all the elements of the$\uparrow RLZSet^t_Y$ sets in a line with "∘" denoting them. We then presented all the operations between the $\uparrow RLZSet^t_X$ sets and the $\uparrow RLZSet^t_Y$ sets based on the$\uparrow RLZSet_{X-Y}$ set, which is shown in Fig. 1.

It is interesting to note that the transition operations in the operation figure indicate the similarity between the $\uparrow RLZSet^t_X$ sets and the $\uparrow RLZSet^t_Y$ set, and the extension operations imply their diversity. That is to say, the more the extension operations are, the more similar the $\uparrow RLZSet^t_X$ sets and the $\uparrow RLZSet^t_Y$ set are. According to that, we define length of the operations $L(O)$ as follows:

$$L(O)=\begin{cases} 1\ , if\ \ operation\ \ is\ \ transition \\ t+1,\ if\ \ t\ \ consecutiveoperations\ \ are\ \ extensition \end{cases}$$

Given an operation $L(O)$ with lengthξ, we counted its total appearances ($N_{L(O)=\xi}$) in operation figure. Since $L(O)$ varies with different valueζ, it can be regarded as a discrete random variable. Given a random variable$L(O)$, and a positive integer n, $P(L(O)=n)$ is the probability that $L(O)$takes the value n

$$P(L(O)=n)=N_{L(o)=n}/\sum_{\xi} N_{L(o)=\xi}$$

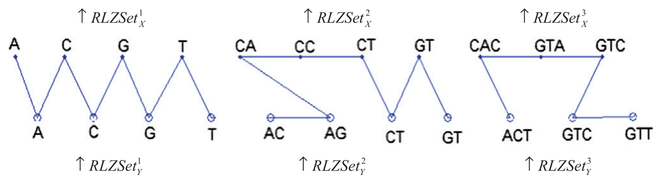

Fig. 1. All the transition operations and extension operations between the $\uparrow RLZSet^t_X$ sets and the $\uparrow RLZSet^t_Y$ sets according to the $\uparrow RLZSet_{X-Y}$ set.

The collection of pairs$(P(L(O)=n),\ n)$, for all positive integer n, is the probability distribution of the listed as follows:

$L(O)$	$L(O)=1$	$L(O)=2$	…	$L(O)=\mathrm{n}$	…
P	$P(L(O)=1)$	$P(L(O)=2)$	…	$P(L(O)=\mathrm{n})$	…

Take all the operations in Fig. 1 for an example, the probability distribution of the operations is

$L(O)$	$L(O)=1$	$L(O)=2$	$L(O)=3$
P	11/15	2/15	2/15

Based on the operation distribution function, we calculated its expectation and propose an operation measure (*OMeasure*) between two sequence X and Y,

$$OMeasure(X,Y)=\sum_{z\geq 1} z\times P(L(O)=z)-1.$$

OMeasure, the average length of the operation, is depended on both the extension operations and transition operations. It is important to note that *OMeasure* only satisfies the identity and symmetry, it does not satisfy inequality conditions. So it is only a dissimilarity measure for sequence comparison.

We are interested in *OMeasure* for two reasons. First of all, it provides an opportunity to study the components' distribution which is, in some ways, more singular than the total number of components in the exhaustive history. The second reason involves the lengths of the operations because differencing lengths of the operations strengthens the effects of the different operations.

3. Results and discussion

3.1. Comparison of component distribution in the exhaustive history between Lempel–Ziv (LZ) complexity and revised Lempel–Ziv complexity

One of the characteristics of the revised Lempel–Ziv complexity is to check whether$R=S(h_m,h_{m+1}-1)$can be reproduced from the set $\{S(1,\ h_1),\ S(h_1+1,\ h_2),\cdots,S(h_{m-1}+1,\ h_m)\}$ instead of from the set $S(1,\ h_m)$. To find their difference, we compared their LZ-word's distributions.

We first compared their component difference in the exhaustive histories. For example, HCoV-229E is a given sequence of Human coronavirus, its length is 27,317 with accession number NC_002645. With Lempel–Ziv complexity and revised Lempel–Ziv complexity, we got two exhaustive histories.

We then deleted all the symbols "*" in the exhaustive histories and obtained two new deduced sequences HCoV-229E_LZ and HCoV-229E_RLZ. It is difficult for us to observe sequence difference directly, but we can calculate *k*-word counts of the deduced sequences to assess their difference. Fig. 2 is the *k*-word counts of the deduced sequences HCoV-229E_LZ and HCoV-229E_RLZ with *k* from 1 to 4. Interestingly, the *k*-word counts of the deduced sequences HCoV-229E_LZ and HCoV-229E_RLZ are similar in Fig. 2. That is to say, the sequence information held through Lempel–Ziv (LZ) complexity and revised Lempel–Ziv complexity operation is similar.

In statistics, one-way analysis of variance (abbreviated one-way ANOVA) is a technique used to compare means of two or more samples (using the F distribution). Here, we used a one-way ANOVA to test whether the *k*-word counts of the deduced sequences HCoV-229E_LZ and HCoV-229E_RLZ differ from each other. The F value obtained by one-way ANOVA test tells us whether the data is significantly different from the Gaussian distribution or not. We rejected the hypothesis if the test is significant at the 0.05 level. Since the F-value is 0.91 > $F_{0.05}$, there is no significant difference between the *k*-word counts of the deduced sequences HCoV-229E_LZ and HCoV-229E_RLZ. That is to say, the components in the exhaustive history between Lempel–Ziv complexity and revised Lempel–Ziv complexity are similar.

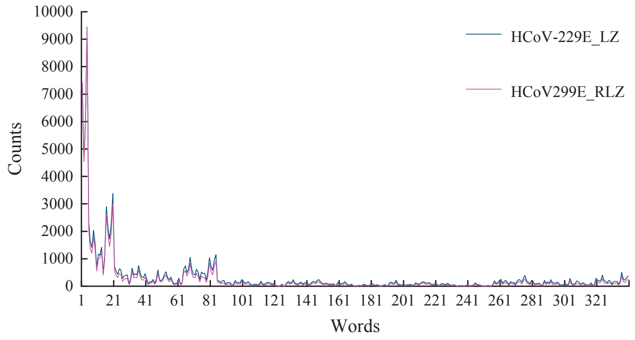

Fig. 2. The comparison of k-word counts of the deduced sequences HCoV-229E_LZ and HCoV-229E_RLZ with k from 1 to 4.

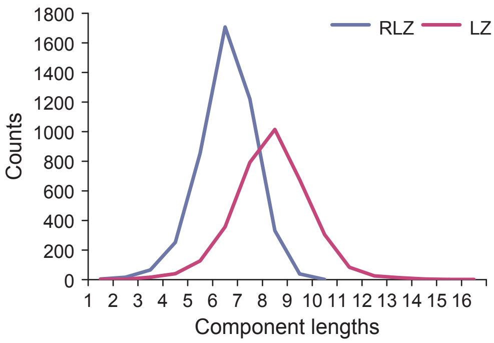

Fig. 3. Comparison of length distribution of the components in the exhaustive history obtained by Lempel–Ziv (LZ) complexity and revised Lempel–Ziv complexity.

We found that the total number of components in the exhaustive history with the revised Lempel–Ziv (LZ) complexity algorithm is 4491, which is 1026 larger than Lempel–Ziv (LZ) complexity algorithm. In addition to components' distribution in the exhaustive history, we also compared the distribution of lengths of the components in the exhaustive history. Take the HCoV-229E as an example, Fig. 3 is the comparison of length distribution of the components in the exhaustive history obtained by Lempel–Ziv complexity and revised Lempel–Ziv complexity. It is interesting to note that there is a great difference between the lengths of the components in the exhaustive history. The maximum length of the components obtained by Lempel–Ziv complexity is 16, while that of the components obtained by revised Lempel–Ziv complexity is only 10. The most appearance length of the components obtained by revised Lempel–Ziv complexity is 6, which is 2 smaller than that of the components obtained by Lempel–Ziv complexity.

Comparison between Lempel–Ziv (LZ) complexity and revised Lempel–Ziv complexity illustrates that they can both extract the similar information of the primary sequences, but the component lengths in the exhaustive history obtained by the revised Lempel–Ziv complexity are obviously smaller than Lempel–Ziv complexity. So the revised Lempel–Ziv complexity is a better way to make the components easier to handle.

3.2. Influence of set splitting methods on operation measure

Given two DNA sequences X and Y, we obtained their revised LZ-word sets $RLZSet_X$ and $RLZSet_Y$ with the revised Lempel–Ziv complexity. We then blended $RLZSet_X$ and $RLZSet_Y$ to compose anther set $RLZSet_{X-Y}$. In order to highlight the influence of different LZ-words' size, we divided the $RLZSet_{X-Y}$set into several mutually exclusive sets $RLZSet^{t}_{X-Y}$ according to the length of revised LZ-word t. It is worthy to note that mutually exclusive sets $RLZSet^{t}_{X-Y}$ rely heavily on set splitting methods. In order to evaluate the influence of the set splitting methods, we adopted the operation measure to classify HEV Genotypes with step-wise refinement of set splitting methods.

HEV (Hepatitis E virus) is a non-enveloped, positive-sense, single-stranded RNA virus and belongs to Hepevirus genus under the separate family of Hepeviridea (Lu et al., 2006). The genome of HEV is approximately 7.2 kb in length and contains a short 5′ untranslated region (5′ UTR), three overlapped open reading frames (ORF1, ORF2, and ORF3′) and a short 3′ UTR. We retrieved a total of 48 full-length HEV genome sequences from NCBI (http://www.ncbi.nlm.nih.gov/). Abbreviation for the strains, accession number, nucleotide length, country, and genotype of all HEV genomes (Lu et al., 2006) are described in Table 1. And the 48 HEV genomes were distinctly clustered into four genotypes by the traditional classification (Liu et al., 2008).

This experiment aims at assessing how well the operation measure with step-wise refinement of set splitting methods performs on classification. Here, set splitting methods with the step-wise refinement (SSM) are:

$$SSM_1 = RLZSet_{X-Y},\ \ SSM_2 = RLZSet^{1}_{X-Y} \cup RLZSet^{t\geq 2}_{X-Y},$$

$$SSM_3 = RLZSet^{1}_{X-Y} \cup RLZSet^{2-4}_{X-Y} \cup RLZSet^{t\geq 5}_{X-Y},$$

$$SSM_4 = RLZSet^{1}_{X-Y} \cup RLZSet^{2-3}_{X-Y} \cup RLZSet^{4-5}_{X-Y} \cup RLZSet^{6-7}_{X-Y} \cup RLZSet^{t\geq 8}_{X-Y},$$

$$SSM_5 = \bigcup_{t=1}^{9} RLZSet^{t}_{X-Y}.$$

In relation to the clustering literature (Handl et al., 2005), Neighbor-joining (Felsenstein, 1989), a classic tree construction algorithm, can be considered as hierarchical methods. These results are represented in Fig. 4.

Table 1
Abbreviation for the strains, accession number, nucleotide length, genotype, acronym and country for each of the 48 complete HEV genomes.

No	Strain name	Accession	Length	Genotype	Abbreviation	Country
1	B1 (Bur-82)	M73218	7207	I	AA	Burma (Rangoon)
2	B2 (Bur-86)	D10330	7194	I	AB	Burma (Rangoon)
3	I2 [Mad-93]	X99441	7194	I	AC	India (Madras)
4	I3	AF076239	7194	I	AD	India (Hyderabad)
5	Np1(TK15/92)	AF051830	7199	I	AE	Nepal (Kathamandu)
6	P2[Abb-2B]	AF185822	7143	I	AF	Pakistan (Abbottabad)
7	Yam-67	AF459438	7206	I	AG	India (Yamuna Nagar)
8	C1(CHT-88)	D11092	7207	I	AH	China (Xinjiang, Hetian)
9	C2(KS2–87)	L25595	7221	I	AI	China (Xinjiang, Kashi)
10	C3(CHT-87)	L08816	7176	I	AJ	China (Xinjiang, Hetian)
11	C4(Uigh179)	D11093	7194	I	AK	China (Xinjiang, Uighur)
12	China Hebei	M94177	7200	I	AR	China (Hebei)
13	P1(Sar-55)	M80581	7138	I	AM	Pakistan (Sargodha)
14	I1(FHF)	X98292	7202	I	AN	India
15	Morocco	AY230202	7212	I	AO	Morocco
16	T3	AY204877	7170	I	AP	Chad
17	M1	M74506	7180	II	BB	Mexico (Telixtac)
18	HE-JA10	AB089824	7262	III	CA	Japan (Tokyo)
19	JKN-Sap	AB074918	7256	III	CB	Japan (Sapporo)
20	JMY-HAW	AB074920	7240	III	CC	Japan (Sapporo)
21	swUS1	AF082843	7207	III	CD	USA
22	US1	AF060668	7202	III	CE	USA (Minnesota)
23	US2	AF060669	7277	III	CF	USA (Tennessee)
24	JBOAR1-Hyo04	AB189070	7247	III	CG	Japan (Hyogo)
25	JDEER-Hyo03L	AB189071	7230	III	CH	Japan (Hyogo)
26	JJT-KAN	AB091394	7218	III	CI	Japan (Kanagawa)
27	JMO-Hyo03L	AB189072	7180	III	CJ	Japan (Hyogo)
28	JRA1	AP003430	7230	III	CK	Japan (Tokyo)
29	JSO-Hyo03L	AB189073	7180	III	CR	Japan (Tokyo)
30	JTH-Hyo03L	AB189074	7180	III	CM	Japan (Tokyo)
31	JYO-Hyo03L	AB189075	7180	III	CN	Japan (Tokyo)
32	swJ570	AB073912	7257	III	CO	Japan (Tochigi)
33	Kyrgyz	AF455784	7239	III	CP	Kyrgyzstan
34	Arkell	AY115488	7255	III	CQ	Canada (Ontario, Guelph)
35	HE-JA1	AB097812	7258	IV	DA	Japan (Hokkaido)
36	HE-JK4	AB099347	7250	IV	DB	Japan (Tochigi)
37	HE-JI4	AB080575	7186	IV	DC	Japan (Tochigi)
38	JAK-Sai	AB074915	7236	IV	DD	Japan (Saitama)
39	JKK-Sap	AB074917	7235	IV	DE	Japan (Sapporo)
40	JSM-Sap95	AB161717	7202	IV	DF	Japan (Hokkaido)
41	JSN-Sap-FH	AB091395	7234	IV	DG	Japan (Hokkaido)
42	JSN-Sap-FH02C	AB200239	7251	IV	DH	Japan (Hokkaido)
43	JTS-Sap02	AB161718	7202	IV	DI	Japan (Hokkaido)
44	JYW-Sap02	AB161719	7202	IV	DJ	Japan (Hokkaido)
45	swJ13–1	AB097811	7258	IV	DK	Japan (Hokkaido)
46	swCH25	AY594199	7270	IV	DR	China (Uighur)
47	T1	AJ272108	7232	IV	DM	China (Beijing)
48	CCC220	AB108537	7193	IV	DN	China (Changchun)

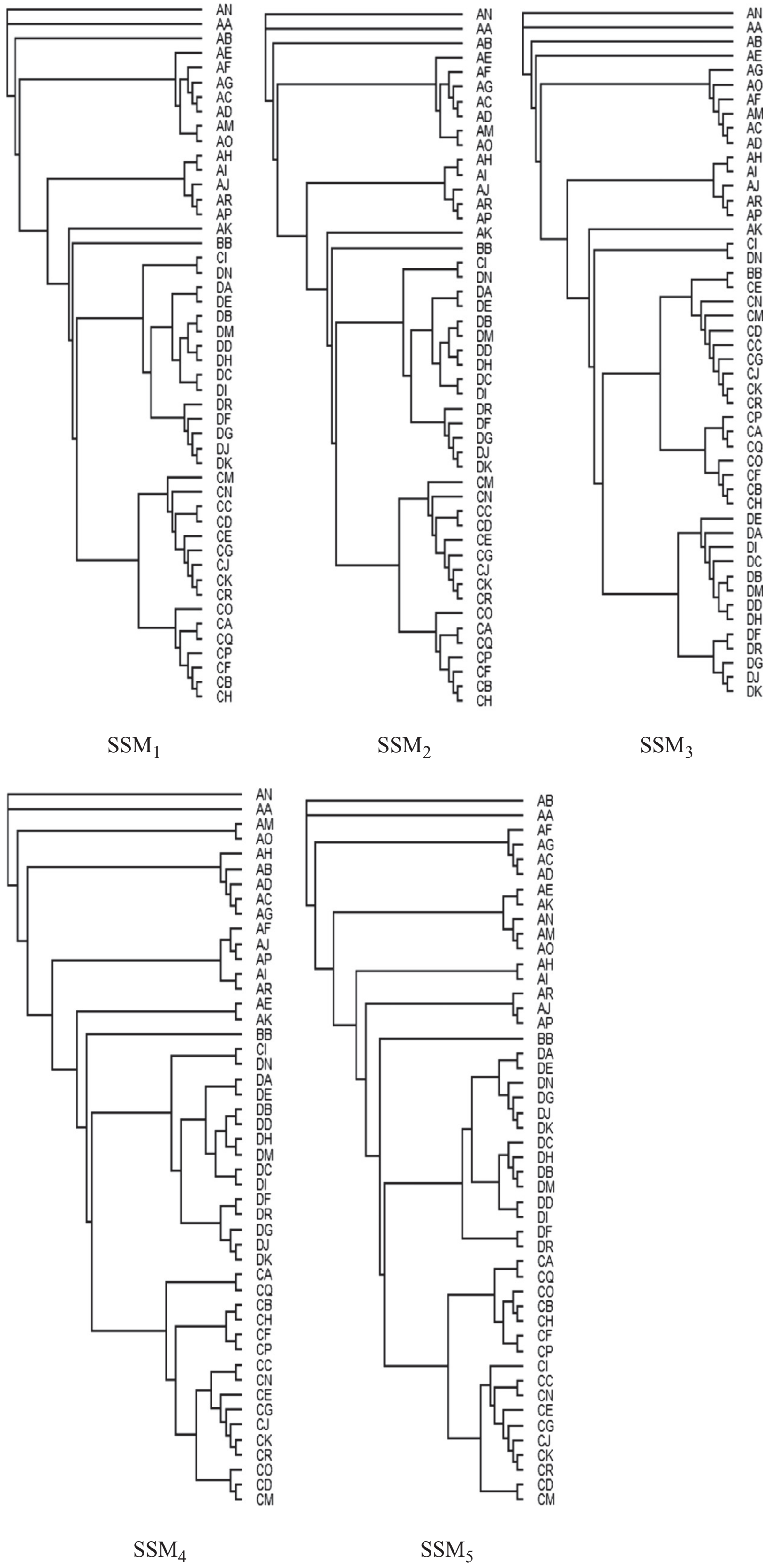

Fig. 4. Cluster trees of 48 HEV genomes using tree construction algorithm Neighbor-joining based on the proposed operation measure with SSM_1, SSM_2, SSM_3, SSM_4, and SSM_5.

To evaluate the performance of the operation measure for HEV genotypes classification, we counted the number of misplaced HEV genotype against a gold standard. For the classification of HEV genotypes, we took the traditional classification as the gold standard (Lu et al., 2006). The numbers of misplaced HEV genotype for the operation measure with SSM_1, SSM_2, SSM_3,

SSM_4, and SSM_5 are 1, 1, 2, 1 and 0, respectively. These results indicate that the higher the refinement scheme is, the higher the operation measure efficiency is.

3.3. Phylogenetic analysis of coronaviruses

Since the outbreak of atypical pneumonia referred to as severe acute respiratory syndrome (SARS), more attentions have been paid to the relationships between the SARS-CoVs and the other coronaviruses, which would be helpful to discover drugs and develop vaccines against the virus. Generally, coronaviruses can be divided into three groups according to serotypes. Group I and group 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.

Based on the operation measure, we next considered to infer the phylogenetic relationships of coronaviruses with the complete coronavirus genomes. The 24 complete coronavirus genomes used

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

No	Accession	Group	Abbreviation	Genome	Length(nt)
1	NC_002645	I	HCoV-229E	Human coronavirus 229E	27,317
2	NC_002306	I	TGEV	Transmissible gastroenteritis virus	28,586
3	NC_002436	I	PEDV	Porcine epidemic diarrhea virus	28,033
4	U00735	II	BCoVM	Bovine coronavirus strain Mebus	31,032
5	AF391542	II	BCoVL	Bovine coronavirus isolate BCoV–LUN	31,028
6	AF220295	II	BCoVQ	Bovine coronavirus strain Quebec	31,100
7	NC_003045	II	BCoV	Bovine coronavirus	31,028
8	AF208067	II	MHVM	Murine hepatitis virus strain ML–10	31,100
9	AF201929	II	MHV2	Murine hepatitis virus stain 2	31,028
10	AF208066	II	MHVP	Murine hepatitis virus strain Penn 97–1	31,233
11	NC_001846	II	MHV	Murine hepatitis virus	31,276
12	NC_001451	III	IBV	Avian infectious bronchitis virus	27,608
13	AY278488	IV	BJ01	SARS coronavirus BJ01	29,725
14	AY278741	IV	Urbani	SARS coronavirus Urbani	29,727
15	AY278491	IV	HKU-39849	SARS coronavirus HKU-39849	29,742
16	AY278554	IV	CUHK-W1	SARS coronavirus CUHK–W1	29,736
17	AY282752	IV	CUHK-Su10	SARS coronavirus CUHK–Su10	29,736
18	AY283794	IV	SIN2500	SARS coronavirus Sin2500	29,711
19	AY283795	IV	SIN2677	SARS coronavirus Sin2677	29,705
20	AY283796	IV	SIN2679	SARS coronavirus Sin2679	29,711
21	AY283797	IV	SIN2748	SARS coronavirus Sin2748	29,706
22	AY283798	IV	SIN2774	SARS coronavirus Sin2774	29,711
23	AY291451	IV	TW1	SARS coronavirus TW1	29,729
24	NC_004718	IV	TOR2	SARS coronavirus	29,751

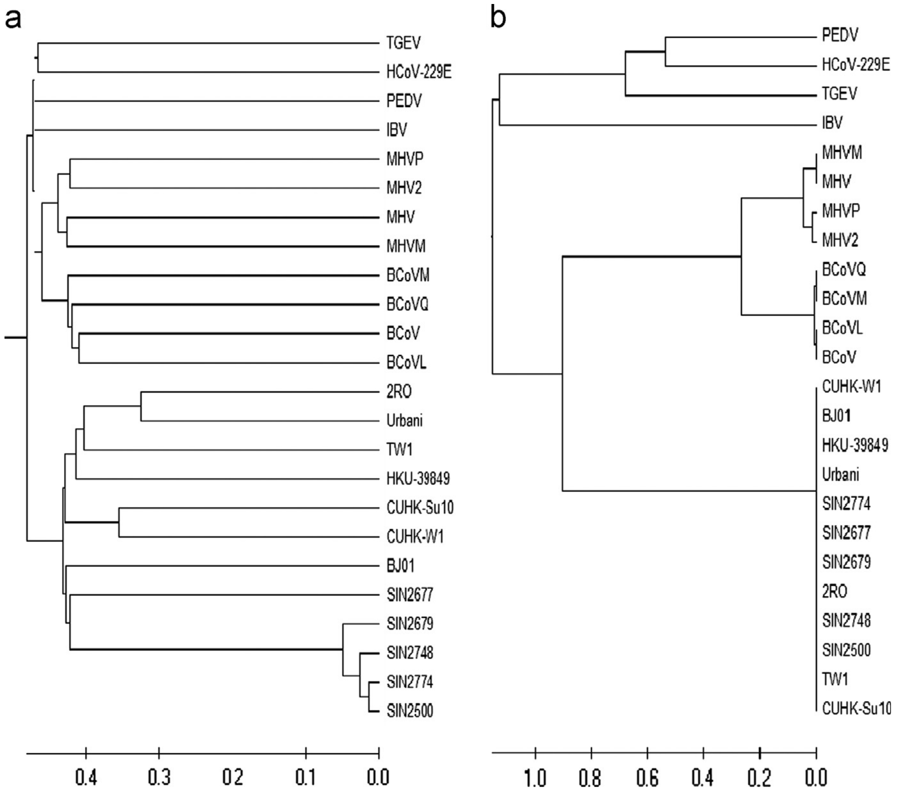

Fig. 5. Phylogenetic tree of 24 coronavirus genomes based on (a) the proposed operation measure and (b) multiple alignment CLUSTAL X.

in this article 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. Given a set of biological sequences, their phylogenetic relationship can be obtained through the following main operations: firstly, we construct the LZ-word set with revised Lempel–Ziv complexity and calculate the similarity/dissimilarity using operation measure; secondly, by arranging all the similarity/dissimilarity into a matrix, we obtain a pair-wise matrix; finally, we put the pair-wise distance matrix into the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) program in the PHYLIP package (Felsenstein, 1989). Fig. 5(a) is phylogenetic tree of the 24 coronavirus genomes obtained using the proposed operation measure with SSM_5.

Generally, an independent method can be developed to evaluate the accuracy of a phylogenetic tree, or the validity of a phylogenetic tree can be tested by comparing it with authoritative ones. Here, we adopted the form one to test the validity of our phylogenetic tree. Both two data sets were aligned with the multiple alignment CLUSTAL X and constructed the phylogenetic tree presented in Fig. 5(b).

Fig. 5(a) shows that our results are quite consistent with the authoritative results (Gu et al., 2004; Zheng et al., 2005) and that of the multiple alignment Fig. 5(b) in the following aspects. First of all, all SARS-CoVs are grouped in a separate branch, which appear different from the other three groups of coronaviruses. Secondly, BCOV, BCOVL, BCOVM, BCOVQ, MHV, MHV2, MHVM, and MHVP are grouped into a branch, which is consonant with the fact that they belong to group II. Thirdly, HCoV-229E, TGEV, and PEDV are closely related to each other, which is consistent with the fact that they belong to group I. Finally, IBV forms a distinct branch within the genus Coronavirus, because it belongs to group III. Rota et al. (Rota et al., 2003) found out that the overall level of similarity between SARS-CoVs and the other coronaviruses is low. Our tree also reconfirms that SARS-CoVs are not closely related to any previously isolated coronaviruses and form a new group, which indicates that the SARS-CoVs have undergone an independent evolution path after the divergence from the other coronaviruses.

4. Conclusion

Sequence comparison is one of the major goals of sequence analysis, which could serve as evidence of structural and functional conservation, as well as of evolutionary relations among the sequences. Despite the prevalence of the alignment-based methods, it is also noteworthy that it is computationally intensive and consequently unpractical for querying large data sets. Therefore, considerable efforts have been made to seek for alternative methods for sequence comparison.

This work presented a novel method to compare biological sequence with the revised Lempel–Ziv complexity. Instead of focusing on the total number of components in the exhaustive history, we analyzed the distribution of components themselves. Then we defined transition and extension operations among the revised LZ-word sets and represented them in the operation figure. With the length of operations in mind, we designed an operation measure to estimate the similarity/dissimilarity of two biological sequences. To assess the effectiveness of the proposed method, two sets of evaluation experiments were taken, and its performance was further compared with alignment-based methods. The results demonstrate that the proposed method is efficient, which highlight the necessity for LZ-based method to consider the whole distribution of the components in the exhaustive history. Thus, this understanding can then be used to guide development of more powerful alignment-free for biological sequence comparison.

Acknowledgments

We thank the referees for many valuable comments that have improved this manuscript. This work is supported by the National Natural Science Foundation of China (61170316, 61001214, 61003191).

References

Apostolico, A., Denas, O., 2008. Fast algorithms for computing sequence distances by exhaustive substring composition. Algorithms for Molecular Biology 3, 13.

Bacha, S., Baurain, D., 2005. Application of Lempel-Ziv complexity to alignment-free sequence comparison of protein families. Benelux Bioinformatics Conference 2005.

Blaisdell, B.E., 1986. A measure of the similarity of sets of sequences not requiring sequence alignment. In: Proceedings of the National Acadamy of Sciences USA. vol. 83, pp. 5155–5159.

Chen, W., Zhang, Y.S., 2012. Comparative analysis of RNA molecules. MATCH communications in mathematical and in computer chemistry 67, 253–268.

Durbin, R., Eddy, S.R., Krogh, A., Mitchison, G., 1998. Biological Sequence Analysis. Cambridge University Press.

Ewens, J., Grant, G., 2005. Statistical Methods in Bioinformatics: An Introduction. Springer Science, New York.

Felsenstein, J., 1989. PHYLIP-Phylogeny inference package (version 3.2). Cladistics 5, 164–166.

Fichant, G., Gautier, C., 1987. Statistical method for predicting protein coding regions in nucleic acid sequences. Computional Applied Biosciences 3, 287–295.

Gotoh, O., 1982. An improved algorithm for matching biological sequences. Journal of Molecular Biology 162, 705–708.

Gu, W., Zhou, T., Ma, J., Sun, X., Lu, Z., 2004. Analysis of synonymous codon usage in SARS Coronavirus and other viruses in the Nidovirales. Virus Research 101, 155–161.

Handl, J., Knowles, J., Kell, D.B., 2005. Computational cluster validation in post-genomic data analysis. Bioinformatics 21 (15), 3201–3212.

Hao, B., Qi, J., 2004. Prokaryote phylogeny without sequence alignment: from avoidance signature to composition distance. Journal of Bioinformatics and Computational Biology 2, 1–19.

Kantorovitz, M.R., Robinson, G.E., Sinha, S., 2007. A statistical method for alignment-free comparison of regulatory sequences. Bioinformatics 23, i249–i255.

Li, C., Wang, J., 2003. Numerical characterization and similarity analysis of DNA sequences based on 2-D graphical representation of the characteristic sequences. Combinatorial Chemistry and High Throughput Screening 6, 795–799.

Li, C., Li, Z.X., Zheng, X.Q., Ma, H., Yu, X.Q., 2010. A generalization of Lempel-Ziv complexity and its application to the comparison of protein sequences. Journal of Mathematical Chemistry 48, 330–338.

Liao, B., Wang, T.M., 2004. 3-D graphical representation of DNA sequences and their numerical characterization. Journal of Molecular Structure Theochem 681, 209–212.

Liu, L.W., Li, D.B., Bai, F.L., 2012. A relative Lempel–Ziv complexity: Application to comparing biological sequences. Chemical Physics Letters 530, 107–112.

Liu, N., Wang, T.M., 2006. A method for rapid similarity analysis of RNA secondary structures. BMC Bioinformatics 7, 493.

Liu, Z., Meng, J., 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 368 (2), 223–230.

Lu, G.Q., Zhang, S.P., Fang, X., 2008. An improved string composition method for sequence comparison. BMC Bioinformatics 9 (6), S15.

Lu, L., Li, C., Hagedorn, C.H., 2006. Phylogenetic analysis of global hepatitis E virus sequences: genetic diversity, subtypes and zoonosis. Reviews in Medical Virology 16, 5–36.

Needleman, S.B., Wunsch, C.D., 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology 48, 443–453.

Out, H.H., Sayood, K., 2003. A new sequence distance measure for phylogenetic tree construction. Bioinformatics 19, 2122–2130.

Pham, T.D., Zuegg, J., 2004. A probabilistic measure for alignment-free sequence comparison. Bioinformatics 20, 3455–3461.

Randic, M., 2013a. Very efficient search for protein alignment–VESPA. Journal of Computational Chemistry 33, 702–707.

Randic, M., 2013b. Very efficient search for nucleotide alignments. Journal of Computational Chemistry 34, 77–82.

Randic, M., Vracko, M., Lers, N., Plavsic, D., 2003. Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation. Chemical Physics Letters 371, 202–207.

Randic, M., Zupan, J., Balaban, A., Vikic-Topic, D., Plavsic, D., 2011. Graphical representation of proteins. Chemical Reviews 111, 790–862.

Randic, M., Novic, M., Plavsic, D., 2013. Milestones in graphical bioinformatics. International Journal of Quantum Chemistry , http://dx.doi.org/10.1002/qua24479.

Reinert, G., Schbath, S., Waterman, M.S., 2000. Probabilistic and statistical properties of words: an overview. Journal of Computational Biology 7, 1–46.

Rota, P.A., Oberste, M.S., Monroe, S.S., 2003. Characterization of a novel coronavirus associated with severe acute respiratory syndrome. Science 300, 1394.

Smith, T.F., Waterman, M.S., 1981. Identification of common molecular subsequences. Journal of Molecular Biology 147, 195–197.

Song, J., Tang, H., 2005. A new 2-D graphical representation of DNA sequences and their numerical characterization. Journal of Biochemical and Biophysical Methods 63, 228–239.

Stuart, G.W., Moffett, K., Baker, S., 2002. Integrated gene and species phylogenies from unaligned whole genome protein sequences. Bioinformatics 18, 100–108.

Vinga, S., Almeida, J., 2003. Alignment-free sequence comparison-a review. Bioinformatics 19, 513–523.

Waterman, M.S., 1995. Introduction to Computational Biology: Maps, Sequences, and Genomes: Interdisciplinary Statistics. Chapman and Hall/CRC, Boca Raton, FL.

Wu, T.J., Hsieh, Y.C., Li, L.A., 2001. Statistical measures of DNA dissimilarity under Markov chain models of base composition. Biometrics 57, 441–448.

Wu, X., Wan, X., Wu, G., Xu, D., Lin, G., 2006. Phylogenetic analysis using complete signature information of whole genomes and clustered Neighbour-Joining method. International Journal of Bioinformatics Research and Applications 2, 219–248.

Yao, Y.H., Wang, T.M., 2004. A class of new 2-D graphical representation of DNA sequences and their application. Chemical Physics Letters 398, 318–323.

Zhang, S., Wang, T., 2010. A complexity-based method to compare RNA secondary structures and its application. Journal of Biomolecular Structure Dynamics 28, 247–258.

Zhang, Y., Hao, J.K., Zhou, C.J., Chang, K., 2009. Normalized Lempel–Ziv complexity and its application in bio-sequence analysis. Journal of Mathematical Chemistry 46, 1203–1212.

Zhang, Y.S., Chen, W., 2010. Comparisons of RNA secondary structures based on LZ complexity. MATCH Communications in Mathematics and Computer Chemistry 63, 513–528.

Zheng, W.X., Chen, L.L., Ou, H.Y., et al., 2005. Coronavirus phylogeny based on a geometric approach. Molecular Phylogenetics Evolution 36, 224–232.