MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **67** (2012) 253-268

ISSN 0340 - 6253

Comparative Analysis of RNA Molecules

Wei Chen, Yusen Zhang*

School of Mathematics and Statistics, Shandong University at Weihai Weihai 264209, China

(Received March 28, 2011)

Abstract. In this paper we propose a novel method to compare RNA molecules. We transform an RNA secondary structure into a linear structural sequence not only differentiating paired bases from free bases but also considering the hydrogen bonds between paired bases. We also propose two suitable distance measures based on both the linear structural sequences and RNA primary sequences using Lempel-Ziv complexity. The obtained pair distance matrix can be used to construct phylogenetic trees. The proposed approach does not require sequence alignment. The algorithm has successfully constructed phylogenies for nine 3'-terminal structures of viruses and three simulated second structures and phylogenies for 15 species of protozoa.

1 Introduction

Comparison of biological sequences is an important research area of bioinformatics. New methods are emerging for sequence comparison of nucleic acids and proteins [1–5]. The investigation of RNA secondary structures is a challenging task in molecular biology. RNA molecules are integral components of the cellular machinery for protein synthesis and transport, transcriptional regulation, chromosome replication, RNA processing and modification, and other fundamental biological functions [6–8]. Due to the special role of RNA in biological system, RNA has recently become the center of much attention because of its functions as well as catalytic properties, leading to a substantially increased interest in identifying new RNAs and obtaining their structural information.

There are many algorithms for computing the similarity of RNA molecules [9–13]. Previously, almost all such comparisons are based on the alignment, in which a distance function or a score function is used to represent insertion, deletion, and substitution of letters in the compared structures. Such approaches, which have been hitherto widely

*Corresponding author: zhangys@sdu.edu.cn

used, are computer intensive. For a few years, several tree comparison algorithms have been developed [14,15]. Nevertheless, these methods do not take into account the pseudo-knots. Recently, some researchers presented different graphical representations for RNA secondary structures, and used the invariants of matrices constructed from the graphs to characterize and compare RNAs [16–19]. The advantage of graphical representations is that they allow visual inspection of data, helping in recognizing major differences among RNA secondary structures. However, in literatures, almost all the schemes merely regarded the free base (denoted as A, U, G, C) and paired base (denoted as A', U', G', C') as different characters, then the RNA secondary structure is converted into a special sequence and the comparison of RNA molecules is reduced to compare the corresponding strings of the letters ($A, U, G, C, A', U', G', C'$) or other letters. It might lead to neglecting the difference of their roles in determining RNA stability and by this way identical RNA primary sequences with different secondary structures will have identical special sequences that cannot ly (Fig. 1).

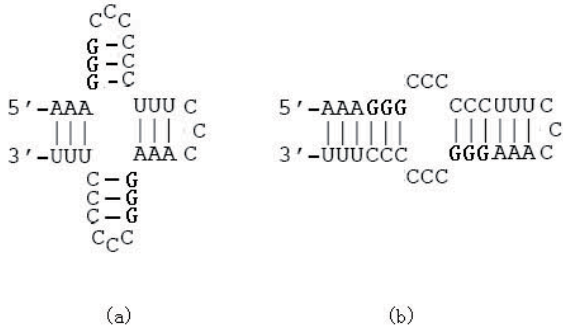


Fig. 1: Two simulated RNA secondary structures with similar RNA sequence.

Complexity is one of the most basic properties of a symbolic sequence. In respect that DNA sequences can be treated as finite-length symbol strings over a four-letter alphabet (A, C, T, G), DNA sequence complexity is much attractive to many researchers. Kolmogorov complexity, the first formal theoretical description of sequence complexity, was proposed by Kolmogorov from the view of algorithm information theory [20]. The authors [21] first introduced Kolmogorov complexity to DNA sequence analysis and proposed

a DNA sequence distance matrix based on it. Because Kolmogorov complexity is not computable, the authors [22] made use of data compression gain to approximate Kolmogorov complexity. However, the generalization of the approximate method is greatly limited because the data compression gain varies evidently with the object to be compressed and the algorithm that a certain compressor uses [23]. In contrary, LZ complexity, another significant complexity measure proposed by Lempel and Ziv [24], is easily computable and is also a universal depiction of sequence complexity. Otu et al [25] have used the Lempel-Ziv algorithm to successfully construct phylogenetic trees from DNA sequences, which verifies the efficiency of Lempel-Ziv algorithm in analyzing the similarity of linear biological sequences. Motivated by this work, Some authors develop a method to compare RNA secondary structures and analyze their similarities [26, 27]. The key idea is that to transform the RNA secondary structures into a linear characteristic sequences (It is called shadow sequences in [27]), these linear sequences are decomposed according to the rule of Lempel-Ziv algorithm to evaluate the LZ complexity. But we find that sometimes these characteristic sequences transformed from RNA secondary structures cannot characterize the corresponding RNA secondary structures effectively. That will result in quit different RNA secondary structures with identical RNA sequence having identical characteristic sequences. For example, the different RNA secondary structures with identical sequences shown in Fig. 1a and Fig. 1b will have identical characteristic sequences according to the method in [27]. We [4] proposed an algorithm to compare RNA secondary structures which can avoid such problems. However, the stacks in RNA molecule structure must be first labeled.

In this paper we propose a novel method for the similarity analysis of RNA secondary structures. In our approach, each secondary structure is transformed into a linear structural sequence by a novel idea that need not label the stacks. We not only differentiate paired bases from free bases but also consider the hydrogen bonds between paired bases. So both the RNA primary sequence and its linear structural sequence contains all the information corresponding RNA molecules. Furthermore, standard and famous Lempel-Ziv algorithm [24] is employed for the similarity analysis. Of course, we have tested the validity of our method by analyzing nine 3'-terminal structures of viruses and three simulated second structures. The results obtained by our method are reasonable and are generally in agreement with the previous studies.

2 Methods and algorithms

2.1 Structural sequence of RNA secondary structures

Because of the complexity of RNA secondary structures, many efficient operations used for the analysis of DNA sequences cannot be applied to the analysis of RNA secondary structures. In order to facilitate the analysis of RNA secondary structures, we proposed a coding algorithm to transform a complex secondary structure into a linear sequence, containing the information on primary sequence and paired bases. The coding algorithm is designed in this section.

The primary sequence of RNA molecule R, reading from $5'$-terminal to the $3'$-terminal, can be represented by $L(R) = R[1]R[2]...R[n]$, where $R[i]$ represents the ith nucleotide of R. The secondary structure of an RNA molecule is the collection of free bases and stacks that consist of series of consecutive base pairs. In RNA molecule, guanine and cytosine pair (G-C) forms a triple hydrogen bond, and adenine and uracil pair (A-U) forms a double hydrogen bond. Additionally, guanine and uracil (G-U) can form a single hydrogen bond base pair [28].

We use $R[i...i+p]$ denotes the consecutive bases $R[i], R[i+1], ...$ and $R[i+p]$, then we can use $R[i...i+p]$ and $R[j...j-p]$ express the two strands of the stack that consists of consecutive bases $R[i], R[i+1], ...$ and $R[i+p]$ with their corresponding pairing partner $R[j], R[j-1], ...$ and $R[j-p]$, respectively.

For example, the primary sequences of the RNA secondary structures R_a (Fig. 1a) and R_b (Fig. 1b) are identical sequence, that is,

$L(R_a) = L(R_b) = AAAGGGCCCCCCUUUCCCAAAGGGCCCCCCUUU.$

$R[1...6]$ and $R[33...28]$, $R[10...15]$ and $R[24...19]$ construct two stacks in Fig.1b.

For a given RNA secondary structure R, we construct its structural sequence $S(R)$ by the following rules:

(1) If $R[i]$ is a free base, that means $R[i]$ doesn't form hydrogen bond with other base, 0 appended to $S(R)$.

(2) If $R[i]$ is a paired base, r_i appended to $S(R)$, where r_i denotes the number of hydrogen bonds between $R[i]$ and its pairing partner.

(3) If consecutive bases $R[i], R[i+1], ...$ and $R[i+p]$ form one strand of a stack, then $\cdot r_i r_{i+1}...r_{i+p}\cdot$ appended to $S(R)$, where $r_i, r_{i+1}, ..., r_{i+p}$ denote the numbers of hydrogen bonds between $R[i], R[i+1], ...$ and $R[i+p]$ with their corresponding pairing partner,

respectively.

For example, in Fig. 1, the structural sequence R_a is different from that of R_b.

$S(R_a) = \cdot 222 \cdot \cdot 333 \cdot 000 \cdot 333 \cdot \cdot 222 \cdot 000 \cdot 222 \cdot \cdot 333 \cdot 000 \cdot 333 \cdot \cdot 222 \cdot$ (Fig. 1a)

$S(R_b) = \cdot 222333 \cdot 000 \cdot 333222 \cdot 000 \cdot 222333 \cdot 000 \cdot 333222 \cdot$ (Fig. 1b)

The RNA primary sequence R and its structural sequence $S(R)$ contain all the information that the RNA secondary structure contains.

2.2 Sequence LZ complexity

Let R, R' and Q be sequences defined over an alphabet $\mathcal{A}$, $\ell(R)$ be the length of R, $R(i)$ denote the ith element of R and $R(i, j)$ define the substring of R composed of the elements of R between positions i and j (inclusive). An extension $R' = RQ$ of R is reproducible from R (denoted $R \rightarrow R'$) if there exists an integer $p \leq \ell(R)$ such that $Q(k) = R'(p+k-1)$ for $k = 1, ..., \ell(Q)$. For example AACGT $\rightarrow$ AACGT CGT CG with $p = 3$ and AACGT $\rightarrow$ AACGT AC with $p = 2$. Another way of looking at this is to say that R' can be obtained from R by copying elements from the pth location in R to the end of R. As each copy extends the length of the new sequence beyond $\ell(R)$, the number of elements copied can be greater than $\ell(R) - p + 1$. Thus, this is a simple copying procedure of R starting from position p, which can carry over to the added part, Q. A sequence R is producible from its prefix $R(1, j)$ (denoted $R(1, j) \Rightarrow R$), if $R(1, j) \rightarrow R(1, \ell(R) - 1)$. For example, AACGT $\Rightarrow$ AACGTAC and AACGT $\Rightarrow$ AACGTACC both with pointers $p = 2$. Note that production allows for an extra different symbol at the end of the copying process which is not permitted in reproduction. Therefore, an extension which is reproducible is always producible but the reverse may not always be true.

Any sequence R can be built using a production process where at its ith step $R(1, h_{i-1})$ $\Rightarrow R(1, h_i)$ [note that $\epsilon = R(1, 0) \Rightarrow R(1, 1)$]. An m-step production process of R results in a parsing of R in which $H(R) = R(1, h_1) \cdot R(h_1 + 1, h_2), ..., R(h_{m-1} + 1, h_m)$ is called the history of R and $H_i(R) = R(h_{i-1} + 1, h_i)$ is called the ith component of $H(R)$. For example for R = AACGTACC, A$\cdot$ A$\cdot$ C $\cdot$ G$\cdot$ T $\cdot$ A$\cdot$ C $\cdot$ C, A $\cdot$ AC $\cdot$ G $\cdot$ T $\cdot$ A $\cdot$ C $\cdot$ C and A $\cdot$ AC $\cdot$ G $\cdot$ T $\cdot$ ACC are three different (production) histories of R. If $R(1, h_i)$ is not reproducible from $R(1, h_{i-1})$, then $H_i(R)$ is called exhaustive. In other words, for $H_i(R)$ to be exhaustive the ith step in the production process must be a production only, meaning that the copying process cannot be continued and the component should be halted with

a single letter innovation. A history is called exhaustive if each of its components (except maybe the last one) is exhaustive. For example the third history given in the preceding paragraph is an exhaustive history of R = AACGTACC. Moreover, every sequence R has a unique exhaustive history [24].

Let $c(R)$ be the number of components in the exhaustive history of R. It is the least possible number of steps needed to generate R according to the whole Lempel-Ziv algorithm, so $c(R)$ becomes an important complexity indicator. For example, $c(L(R_a)) = c(L(R_b)) = 6, c(S(R_a)) = 10$ and $c(S(R_b)) = 9$.

2.3 Proposed distance and pairwise distance matrix

Lempel et al have proposed that, for any given sequences R_2 and R_1, $c(R_2R_1) \leq c(R_2) + c(R_1)$ always holds. This formula shows that the steps required to extend R_2 to R_2R_1 are always less than the steps required to build R_1 from empty string ϕ. Recently, Otu et al [25] concluded that the more similar the sequence R_1 is to sequence R_2, the smaller $c(R_2R_1) - c(R_2)$ is. That is $c(R_2R_1) - c(R_2)$ depends on how much R_1 is similar to R_2. Based on this hypothesis, we use following relative distance measures between sequences R_2 and R_1.

$$d_1(R_1, R_2) = \frac{max\{c(R_1R_2) - c(R_1), c(R_2R_1) - c(R_2)\}}{max\{c(R_1), c(R_2)\}} \tag{1}$$

$$d_2(R_1, R_2) = \frac{max\{c(R_1R_2) - c(R_1), c(R_2R_1) - c(R_2)\}}{max\{c(R_2R_1), c(R_1R_2)\}} \tag{2}$$

The first formula belongs to [25]. The second one is slightly different from the formulas in [25]. We choose to use these formulas mainly because the results are more precise when short RNA primary sequences are compared and analyzed.

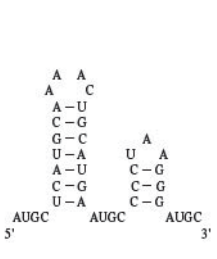
Given RNA molecules R_1 and R_2, we have $d_i(L(R_1), L(R_2))$ and $d_i(S(R_1), S(R_2))$, then the distance measure between RNA molecules R_1 and R_2 is defined as

$$d_i(R_1, R_2) = \sqrt{d_i(L(R_1), L(R_2))^2 + d_i(S(R_1), S(R_2))^2}, \tag{3}$$

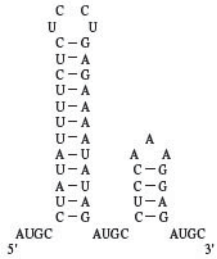
where $i = 1, 2$.

Generally, given n RNA molecules R_1, R_2,..., R_n, the primary sequences are $L(R_1)$, $L(R_2)$,..., $L(R_n)$ and the structural sequences are $S(R_1)$, $S(R_2)$,..., $S(R_n)$. They are linear sequences defined over alphabet $\{A, C, G, U\}$ and $\{0, 1, 2, 3\}$, respectively, and carry the

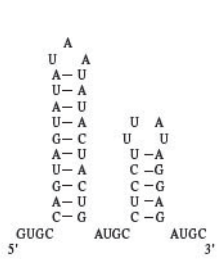
information on RNA molecules. By using Lempel-Ziv algorithm, the distances $d_i(R_k, R_j)$ $(k, j = 1, 2, ..., n)$ between any pair of structures may be rapidly computed. By arranging them into a matrix, a pairwise distance matrix is obtained, denoted by PDM. $PDM_{kj} = (d_i(R_k, R_j))$ contains the information on the similarity/dissimilarity between any pair of RNA secondary structures.



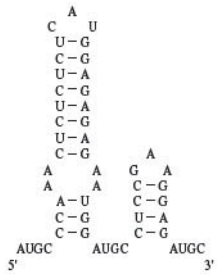
AMV-3



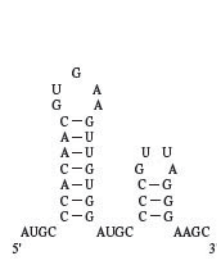
CRV-3



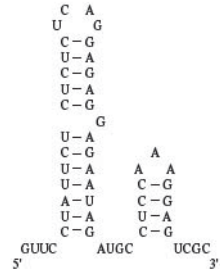
TSV-3



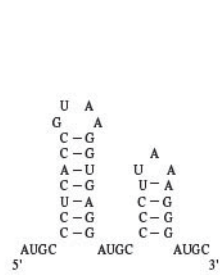
CVV-3



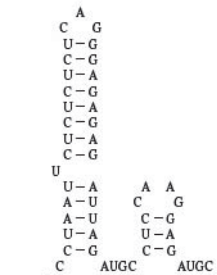
APMV-3



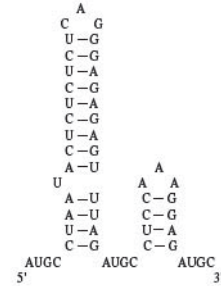
LMV-3



PDV-3



EMV-3



AVII

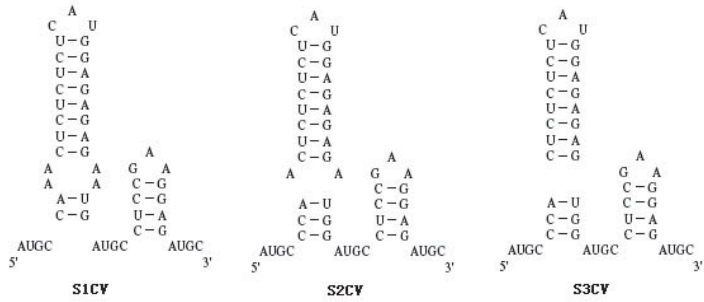


Fig. 2: Secondary structure at the 3'-terminus of RNA 3 of alfalfa mosaic virus (AMV-3), citrus leaf rugose virus (CRV-3), tobacco streak virus(TSV-3), citrus variegation virus (CVV-3), apple mosaic virus (APMV-3), lettuce mosaic virus (LMV-3), prune dwarf ilarvirus (PDV-3), elm mottle virus (EMV-3) and asparagus virus II (AVII). S1CV, S2CV and S3CV are simulated secondary structures which are similar with CVV-3.

3 Comparing RNA secondary structures

Phylogenetic relationships among different organisms are of fundamental importance in biology, and one of the prime objectives of RNA secondary structure is phylogeny reconstruction for understanding evolutionary history of organisms. The goal of our study is to compare RNA secondary structures and analyze their similarity. The utility of our approach in similarity analysis is illustrated by the examination of the similarities/dissimilarities of two sets of the secondary structures.

3.1 Experiment No.1

The utility of our approach in similarity analysis is illustrated by the examination of the similarities/dissimilarities of the RNA secondary structures at the 3'-terminus belonging to nine different viruses, which is used to indicate the validation of their method by many authors [26–28]. In Fig. 2, the nine secondary structures are listed, which were reported by Reusken and Bol [29]. Three simulated secondary structures also are shown in Fig. 2.

Given a set of RNA molecules, our method requires the following main operations for the similarity analysis:

1. The non-linear complex RNA secondary structures are transformed into linear

structural sequences.

2. The primary sequences and structural sequences are decomposed according to the rule of Lempel-Ziv algorithm to evaluate the LZ complexity, respectively.

3. The similarity degree between any two RNA molecules is measured by the distance formulas $d_i(R_1, R_2)$, $i = 1, 2$.

4. After computing the pair distances, we arrange all the values into a matrix for clear and systematical display. This pairwise distance matrix (PDM) is listed in the form of table. It contains the information on the similarity of these RNA molecules.

Table 1: The part of the PDM based on distance measure $d_1(R_1, R_2)$ for 12 RNA secondary structures

Species	CRV-3	TSV-3	CVV-3	APMV-3	LMV-3	PDV-3	EMV-3	AVII	S1CV	S2CV	S3CV
AMV-3	0.9735	1.0152	0.9914	0.9566	1.0207	0.8746	0.9947	0.9638	1.0317	0.9771	1.0328
CRV-3	0	0.7790	0.8201	1.1199	0.7023	1.0135	0.7151	0.6365	0.7023	0.7720	0.8022
TSV-3		0	1.0440	1.1157	0.8940	1.0237	0.9091	0.9253	0.8966	1.0237	0.8781
CVV-3			0	0.9092	0.8018	1.0357	0.7197	0.5943	0.2847	0.2847	0.4359
APMV-3				0	1.0743	0.8109	1.0472	0.9638	0.9327	0.8794	0.8303
LMV-3					0	1.0743	0.7929	0.5908	0.7787	0.8004	0.8771
PDV-3						0	0.9073	0.9638	1.0743	0.9261	0.9667
EMV-3							0	0.5198	0.6938	0.6839	0.7125
AVII								0	0.5943	0.5943	0.5451
S1CV									0	0.4359	0.4956
S2CV										0	0.4093

We have used our method to analyze the similarity of the 3'-terminus belonging to nine different viruses and three simulated secondary structure. The pairwise distance matrix (PDM) based on distance measure $d_1(R_1, R_2)$ for them is listed in Table 1.

We know that S3CV are obtained by deleting four bases in CVV-3 and S1CV and S2CV are obtained by deleting two different bases in CVV-3 secondary structures, respectively. table 1 show us the smaller entries are $d_1(CVV-3, S1CV) = 0.2847$, $d_1(CVV-3, S2CV) = 0.2847$ and $d_1(CVV-3, S3CV) = 0.4359$, which is consistent with the fact above mentioned. Except for S1CV, S2CV and S3CV, we find that the smaller entries ($d_1 < 0.6$) are associated with AVII and CVV-3 [$d_1 = 0.5943$], AVII and LMV-3 [$d_1 = 0.5908$], AVII and EMV-3 [$d_1 = 0.5198$]. All these secondary structures, AVII, CVV-3, LMV-3 and EMV-3, have three base-paired regions and three loop regions. Observing the row corresponding to the EMV-3, we also can find that the smaller entries ($d_1 < 0.8$) are CRV-3, CVV-3, LRMV-3 and AVII which have the similar characterization

Table 2: The part of the PDM based on distance measure $d_2(R_1, R_2)$ for 12 RNA secondary structures

Species	CRV-3	TSV-3	CVV-3	APMV-3	LMV-3	PDV-3	EMV-3	AVII	S1CV	S2CV	S3CV
AMV-3	0.6090	0.6423	0.6455	0.5833	0.6435	0.5644	0.6105	0.6326	0.6609	0.6290	0.6384
CRV-3	0	0.5147	0.5200	0.6466	0.4906	0.6285	0.4742	0.4496	0.4906	0.5315	0.5372
TSV-3		0	0.5938	0.6632	0.5579	0.6550	0.5534	0.5670	0.5487	0.5938	0.5524
CVV-3			0	0.5763	0.4906	0.6325	0.4964	0.4184	0.2368	0.2368	0.3617
APMV-3				0	0.6514	0.5150	0.6294	0.6267	0.5920	0.5858	0.5217
LMV-3					0	0.6690	0.5159	0.4143	0.5021	0.5207	0.5979
PDV-3						0	0.5808	0.6450	0.6690	0.6021	0.5741
EMV-3							0	0.3955	0.4638	0.4730	0.5025
AVII								0	0.4299	0.4184	0.4333
S1CV									0	0.3305	0.3860
S2CV										0	0.3381

that the first hairpin has long stack and second hairpin has similar stack. The row corresponding to the PDV-3 shows us that the smaller entries ($d_1 < 0.9$) associated with PDV-3 are AMV-3 and APMV-3. These three secondary structures have two base-paired regions and two loop regions that are different from above mentioned four secondary structures in topology.

In Table 2, we present the pairwise distance matrix (PDM) based on distance measure $d_2(R_1, R_2)$ for the nine RNA secondary structures and three simulated secondary structures.

Comparing Table 1 and 2, we can find that there exists an overall qualitative agreement among similarities although there is small difference.

To further demonstrate the interrelationships of the 12 secondary structures, we use pairwise distance matrix (PDM) to construct the hierarchical clustering of these secondary structures because the quality of a clustering analysis may verify whether our method of abstracting information from RNA molecules is efficient. in Fig. 3, we present the phylogenetic tree based on linkage cluster analysis using the pairwise distance matrix (PDM) based on distance measure $d_1(R_1, R_2)$.

The relationship of nine 3'-terminal structures of viruses and three simulated second structures is shown reasonably: All the 12 second structures are clustered into two group: AMV-3, PDV-3 and APMV-3 are clustered into one group closely. AVII, CVV-3, LMV-3, EMV-3, CRV-3, TSV-3, S1CV, S2CV and S3CV are clustered into another group. S1CV, S2CV, S3CV and CVV-3 are grouped very closely, which is consistent with the fact that

S1CV, S2CV and S3CV are obtained by deleting two or four bases from CVV-3. But they are less closely with CRV-3 and TSV-3, which is consistent with the fact that the former and latter have some different loop and stem regions. It is not difficult to see that the similarity obtained by our method is coincident with that implicated in the tree.

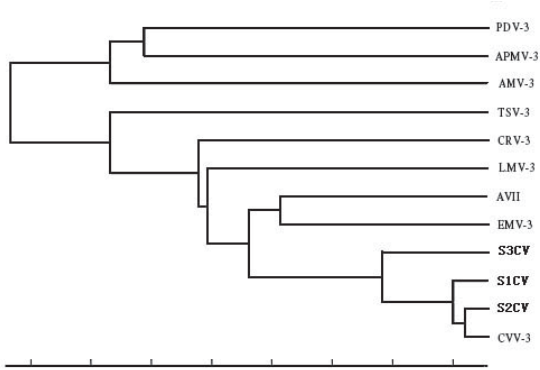


Fig. 3: Phylogenetic tree of the hierarchical clustering of nine 3'-terminal structures of viruses and three simulated second structures which are very similar with CVV-3.

3.2 Experiment No.2

In order to verify the generality of our method, we apply it to the second set. The 15 complex RNA secondary structures of second set are from [30, 31]. These 5S rRNA are: Crithidia fasciculata, Bresslaua vorax, Paramecium tetraurelia, Tetrahymena thermophila, Euplotes woodruffi, Acanthamoeba castellanii, Scenedesmus obliquus, Chlamydomonas sp., Chlorella sp., Spinach (a representative of vascular plants), Aspergillus nidulans (a representative of fungi), Euglena gracilis, Chilomonas paramecium, Physarum polycephalum and Animals (a representative of multicellular animals). The nucleotide sequences of 5S rRNAs from three protozoa, Bresslaua vorax, Euplotes woodruffi and Chlarnydomonas sp. have been determined and aligned together with the sequences of 12 protozoa species including unicellular green algae. Using this alignment, a phylogenic tree of the 15 species of protozoa has been constructed. That can be found in [30].

After computing the pair distances by using $d_1(R_1, R_2)$, we obtain the pair-wise distance matrix (PDM). We present the phylogenetic tree based on linkage cluster analysis

using these pairwise distance matrices and construct two phylogenic trees of these protozoa (Fig.4).

We omit the corresponding results for the distance measure $d_2(R_1, R_2)$ as they are identical to the ones obtained by using $d_1(R_1, R_2)$.

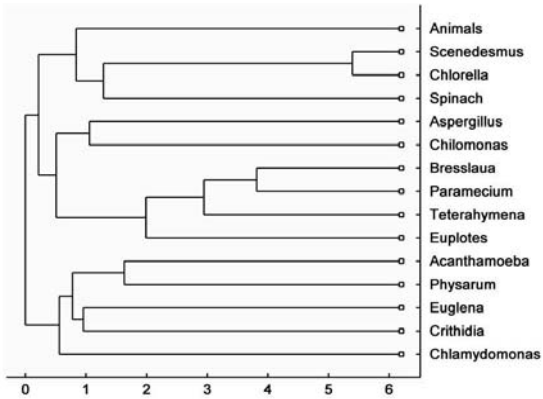


Fig. 4: phylogenetic tree of the hierarchical clustering of 15 5S rRNA . It is obtained by the pair-wise distance matrix based on distance measure $d_1(R_1, R_2)(d_2(R_1, R_2))$ and drawn by Phytreetool.

The phylogenetic tree in Fig.4 shows that Bresslaua, Paramecium, tetraurelia and Euplotes are grouped closely, which is consistent with the fact that they belong to Ciliophora; Physarum, Acanthamoeba, Crithidia and Euglena are intimately related to one another, which is consistent with the fact that they belong to Sarcomastigophora.

In the subphylum Sarcodina, Acanthamoeba and a true-slime mold Physantm are also close to each other from the 5S rRNA sequence data. The slime molds have been often classified in old textbooks as members of fungi, but the 5S rRNA data support the classification system that the slime molds, at least Physarum, and amoeba are placed in the same subphylum Sarcodina. In spite of such a partial consistency, the phylum Sarcomastigophora appears to be, as a whole, composed of a number of mutually unrelated species and thus seems to be quite artificial. Green algae Chlorella, Scenedesmus and plants belong to the same branch in the tree. This view is consistent with the notion that the plants originated from some type of green-flagellated protists.

Corliss (1979) has suggested that in the ciliates the class I species are the "most

primitive" and firstly separated from the ancestor common to "more advanced" class II and the "most advanced" class III species, followed by the separation of these two classes. However, the 5S rRNA tree [30] suggests that the class III (Blepharisma and Euplotes) separated first from the ancestor common to the class I (Bresslaua) and the class II species (Paramecium and Tetrahymena), followed by the separation of the class I and class II more recently. Our result is consistent with the last view. The topology of the tree, except for the positions of the Chilomonas, is generally in agreement with the classification by taxonomic criteria [30].

4 Conclusions

The famous Lempel-Ziv algorithm can efficiently extract the information on repeated patterns encoded in DNA sequence and be used to analyze the similarity of DNA sequence. In order to numerically characterize RNA molecules using these techniques, it is necessary to transform an RNA molecules into linear sequence. As we know that the RNA molecule basically can be described by its primary sequence and secondary structure. Unlike most existing methods, we transform an RNA secondary structure into linear sequences not only differentiating paired bases from free bases but also considering the numbers of hydrogen bonds between paired bases. So the linear sequence contains all the information that contained in the secondary structure. Then comparison of two RNA molecules is now transformed into a comparison of both the RNA primary sequences and linear structural sequences of the corresponding RNA secondary structures. When these linear sequences are obtained, we can use the Lempel-Ziv algorithm to analyze the RNA molecules and build the phylogenetic tree. The proposed method does not require gene identification nor any prior biology knowledge such as an accurate alignment score matrix. To show the utility of the method, we use it to examine the similarities and construct the corresponding phylogenic tree for two data sets.

Acknowledgements

The authors would like to thank the anonymous referees and editors for their corrections and valuable comments. The authors also thank Hasan H. Otu for providing their algorithms. This work is supported by Shandong Natural Science Foundation (ZR2010AM020).

References

[1] Y. Zhang, W. Chen, A new measure for similarity searching in DNA sequences, *MATCH Commun. Math. Comput. Chem.* **65** (2011) 477–488.

[2] Y. Zhang, W. Chen, A dissimilarity measure based on free energy of DNA nearest–neighbor interaction, *J. Biomol. Struct. Dyn.* **28** (2011) 557–565.

[3] Y. Liu, Y. Zhang, New invariant of DNA sequences based on a new matrix representation, *Comb. Chem. High T. Scr.* **14** (2011) 61–71.

[4] Y. Zhang, W. Chen, Comparisons of RNA secondary structures based on LZ complexity, *MATCH Commun. Math. Comput. Chem.* **63** (2010) 513–528.

[5] C. Jia, T. Liu, X. Zhang, H. Fu, Q. Yang, Alignment-free comparison of protein sequences based on reduced amino acid alphabets, *J. Biomol. Struct. Dyn.* **26** (2009) 763–769.

[6] S. R. Eddy, Non-coding RNA genes and the modern RNA world, *Nature Rev. Genet.* **2** (2001) 919–929.

[7] G. Storz, An expanding universe of noncoding RNAs, *Science* **296** (2002) 1260–1263.

[8] T. Schlick, *Molecular Modeling and Simulation: An Interdisciplinary Guide*, Springer–Verlag, New York, 2002.

[9] V. Bafna, S. Muthukrishnan, R. Ravi, Comparing similarity between RNA strings, *6th Annual Symposium on Combinatorial Pattern Matching*, **937** (1995) 1–16.

[10] F. Corpet, B. Michot, RNAlign program: alignment of RNA sequences using both primary and secondary structures, *Comput. Appl. Biosci.* **10** (1994) 389–399.

[11] B. Shapiro, An algorithm for comparing multiple RNA secondary structures, *Comput. Appl. Biosci.* **4** (1988) 387–393.

[12] S. Y. Le, R. Nussinov, J. V. Maizel, RNA secondary structures: comparison and determination of frequently recurring substructures by consensus, *Comput. Appl. Biosci.* **5** (1989) 205–210.

[13] S. Y. Le, R. Nussinov, J. V. Maizel, Tree graphs of RNA secondary structures and their comparisons, *Comput. Biomed. Res.* **22** (1989) 461–473.

[14] B. Shapiro, K. Zhang, Comparing multiple RNA secondary structures using tree comparisons, *Comput. Appl. Biosci.* **6** (1990) 309–318.

[15] S. Dulucq, L. Tichit, RNA secondary structure comparison: exact analysis of the Zhang-Shasha tree edit algorithm, *Theor. Comput. Sci.* **306** (2003) 471–484.

[16] B. Liao, T. Wang, A 3D Graphical representation of RNA secondary structure, *J. Biomol. Struc. Dyn.* **21** (2004) 827–832.

[17] Y. Zhang, On 3D graphical representation of RNA secondary structure, *MATCH Commun. Math. Comput. Chem.* **57** (2007) 157–168.

[18] Y. Zhang, On 2D graphical representation of RNA secondary structure, *MATCH Commun. Math. Comput. Chem.* **57** (2007) 697–710.

[19] J. Feng, T. Wang, A 3D graphical representation of RNA secondary structures based on chaos game representation, *Chem. Phys. Lett.* **454** (2008) 355–361.

[20] M. Li, P. Vitanyi, *An Introduction to Kolmogorov Complexity and Its Applications*, Springer–Verlag, New York, 1997.

[21] M. Li, J. H. Badger, X. Chen, S. Kwong, P. Kearney, H. Zhang, An information–based sequence distance and its application to whole mitochondrial genome phylogeny, *Bioinformatics* **17** (2001) 149–154.

[22] X. Chen, S. Kwong, M. Li, A compression algorithm for DNA sequences and its applications in genome comparison, *Genome Inform. Ser. Workshop Genome Inform.* **10** (1999) 51–61.

[23] H. Sato, T. Yoshioka, A. Konagaya, T. Toyoda, DNA data compression in the post genome era, *Genome Informatics* **12** (2001) 512–514.

[24] J. Ziv, A. Lempel, Compression of individual sequences by variable rate coding, *Inform. Theory IEEE Trans.* **24** (1978) 530–536.

[25] H. H. Otu, K. Sayood, A new sequence distance measure for phylogenetic tree construction, *Bioinformatics* **19** (2003) 2122–2130.

[26] N. Liu, T. Wang, A method for rapid similarity analysis of RNA secondary structures, *BMC Bioinformatics* **7** (2006) 493–503.

[27] C. Li, A. H. Wang, L. Xing Similarity of RNA secondary structures, *J. Comput. Chem.* **28** (2007) 508–512.

[28] D. H. Turner, N. Sugimoto, S. M. Freier, RNA structure prediction, *Ann. Rev. Biophys. Biophys. Chem.* **17** (1988) 167–192.

[29] C. B. Reusken , J. F. Bol, Structural elements of the 3'-terminal coat protein binding site in alfalfa mosaic virus RNAs, *Nucleic Acids. Res.* **24** (1996) 2660–2665.

[30] T. Kumazaki, H. Hori, S. Osa, Phylogeny of protozoa deduced from 5S rRNA sequences, *J. Mol. Evol.* **19** (1983) 411–419.

[31] N. D. Levine, J. O. Corliss, F. E. Cox, G. Deroux, J. Grain, B. M. Honigberg, G. F. Leedale, A. R. Loeblich, J. Lom, D. Lynn, E. G. Merinfeld, F. C. Page, G. Poljansky, V. Sprague, J. Vavra, F. G. Wallace, A newly revised classification of the protozoa, *J. Protozool* **27** (1980) 37–58.