

Comparisons of RNA Secondary Structures Based on LZ Complexity

Yusen Zhang*, Wei Chen

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

(November 11, 2008)

Abstract. We propose a relatively simple and effective method to compare RNA secondary structures. We transform an RNA secondary structure into linear sequences not only differentiating paired bases from free bases but also considering the stacks and loops of the RNA secondary structure. Then we use two suitable distance measures based on the relative information between the linear sequences using Lempel-Ziv complexity, the pair-distance matrix thus obtained can be used to construct phylogenetic trees. The proposed approach does not require sequence alignment. The algorithm has been successfully constructed phylogenies for two data sets.

1 Introduction

RNA is a chain molecule, mathematically a string over a four letter alphabet. It is built from nucleotides containing the bases A(denine), C(ytosine), G(uanine), and U(racil). By folding back onto itself, an RNA molecule forms structure, stabilized by the forces of hydrogen bonds between certain pairs of bases (A-U, C-G, G-U), and dense stacking of neighboring base pairs. 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 [12–14]. Due to the special role of RNA in

*Corresponding author: zhangys@sdu.edu.cn

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 [1–5]. 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 used, are computer intensive. For a few years, several tree comparison algorithms have been developed [6, 7]. 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 [8–11]. 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 secondary structures 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 characterize the corresponding RNA secondary structures effectively (Fig. 1).

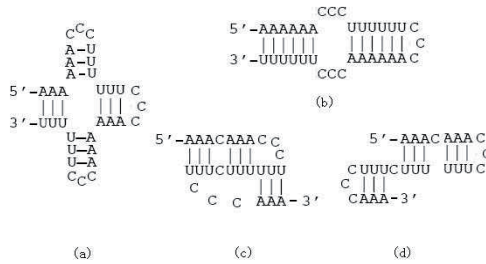


Fig. 1: Four simulated RNA secondary structures.

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 [15]. Li et al. [16] first introduced Kolmogorov complexity to DNA sequence analysis and proposed a DNA sequence distance matrix based on it. Because Kolmogorov complexity is not computable, Chen et al. [17] 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 [18]. In contrary, LZ complexity, another significant complexity measure proposed by Lempel and Ziv [19], is easily computable and is also a universal depiction of sequence complexity. Otu et al. [20] 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 [21, 22]. The key idea is that to transform the RNA secondary structures into a linear characteristic sequences (It is called shadow sequences in [22]), 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 uniquely. 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 [22] and RNA secondary structures shown in Fig. 1c and Fig. 1d will have identical shadow sequences according to the method in [21].

In this paper we propose a novel method for the similarity analysis of RNA secondary structures, where pseudoknots are really taken into account. In our approach, each secondary structure is transformed into a linear sequence by a novel idea. We

not only differentiate paired bases from free bases but also consider the stacks and loops. So the linear sequence contains all the information corresponding RNA secondary structure. Furthermore, standard and famous Lempel-Ziv algorithm [15] is employed for the similarity analysis. Of course, we have tested the validity of our method by analyzing two sets of real data. The results obtained by our method are reasonable and are generally in agreement with the previous studies.

2 Methods and algorithms

2.1 Characteristic 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 secondary structure of an RNA molecule is the collection of free bases and stacks that consist of series of consecutive base pairs. So an RNA sequence, reading from 5'-terminal to the 3'-terminal, can be represented by $R = R[1]R[2]...R[n]$, where $R[i]$ represents the i th nucleotide of R . 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. The stacks are labeled alphabetically from 5'-terminal to 3'- terminal: The first stack is called stack a, and the second stack is called stack b, ... etc.

For example, the RNA sequence shown in Fig. 1 is

$$R = AAAAAACCCUUUUUCCCCAAAAAACCCUUUUU.$$

$R[1...6]$ 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 characteristic sequence $L(R)$ by the following rules:

- (1) If $R[i]$ is a free base, $R[i]$ appended to $L(R)$.

(2) If $R[i]$ is a paired base, $R[i] \cdot$ appended to $L(R)$, where $\cdot$ denote the bond between $R[i]$ and its pairing partner.

(3) If $R[i \dots i + p]$ is one strand of stack with label X , ($X \in \{a, b, c, \dots\}$), then $XR[i] \cdot R[i + 1] \cdot \dots R[i + p] \cdot X$ appended to $L(R)$.

For example, the four characteristic sequences of the RNA secondary structure (Fig. 1) are

$L(R_a) = aA \cdot A \cdot A \cdot abA \cdot A \cdot A \cdot bCCCbU \cdot U \cdot U \cdot bcU \cdot U \cdot U \cdot cCCCcA \cdot A \cdot A \cdot cdA \cdot A \cdot A \cdot dCCcdU \cdot U \cdot U \cdot dcU \cdot U \cdot U \cdot c$ (Fig. 1a)

$L(R_b) = aA \cdot A \cdot A \cdot A \cdot A \cdot A \cdot aCCCbU \cdot U \cdot U \cdot U \cdot U \cdot bCCCbA \cdot A \cdot A \cdot A \cdot A \cdot A \cdot bCCCaU \cdot U \cdot U \cdot U \cdot U \cdot U \cdot a$ (Fig. 1b)

$L(R_c) = aA \cdot A \cdot A \cdot aAAACCCbU \cdot U \cdot U \cdot baU \cdot U \cdot U \cdot aCCCAAAbA \cdot A \cdot A \cdot b$ (Fig. 1c)

$L(R_d) = aA \cdot A \cdot A \cdot aAAACCCaU \cdot U \cdot U \cdot abU \cdot U \cdot U \cdot bCCCAAAbA \cdot A \cdot A \cdot b$ (Fig. 1d)

One can observe that it is easy to restore the RNA sequence and its secondary structure from the characteristic sequence. That means there exists a one-to-one correspondence between the RNA secondary structure and the characteristic sequence. Therefore, the characteristic sequence we proposed is the representative of the RNA secondary structure. The characteristic sequence contains all the information that the RNA secondary structure contains, or vice versa.

2.2 Sequence LZ complexity

Let S, Q and R be sequences defined over an alphabet $\mathcal{A}$, $\ell(S)$ be the length of S , $S(i)$ denote the i th element of S and $S(i, j)$ define the substring of S composed of the elements of S between positions i and j (inclusive). An extension $R = SQ$ of S is reproducible from S (denoted $S \rightarrow R$) if there exists an integer $p \leq \ell(S)$ such that $Q(k) = R(p + k - 1)$ for $k = 1, \dots, \ell(Q)$. For example $AACGT \rightarrow AACGT \text{ CGT}$ CG with $p = 3$ and $AACGT \rightarrow AACGT \text{ AC}$ with $p = 2$. Another way of looking at this is to say that R can be obtained from S by copying elements from the p -th location in S to the end of S . Since each copy extends the length of the new sequence beyond $\ell(S)$, the number of elements copied can be greater than $\ell(S) - p + 1$. Thus,

this is a simple copying procedure of S starting from position p , which can carry over to the added part, Q . A sequence S is producible from its prefix $S(1, j)$ (denoted $S(1, j) \Rightarrow S$), if $S(1, j) \rightarrow S(1, \ell(S) - 1)$. 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 S can be built using a production process where at its i -th step $S(1, h_{i-1}) \Rightarrow S(1, h_i)$ [note that $\epsilon = S(1, 0) \Rightarrow S(1, 1)$]. An m -step production process of S results in a parsing of S in which $H(S) = S(1, h_1) \cdot S(h_1 + 1, h_2), \dots, S(h_{m-1} + 1, h_m)$ is called the history of S and $H_i(S) = S(h_{i-1} + 1, h_i)$ is called the i -th component of $H(S)$. For example for $S = AACGTACC$, $A \cdot A \cdot C \cdot G \cdot T \cdot A \cdot C \cdot C$, $A \cdot AC \cdot G \cdot T \cdot A \cdot C \cdot C$ and $A \cdot AC \cdot G \cdot T \cdot ACC$ are three different (production) histories of S . If $S(1, h_i)$ is not reproducible from $S(1, h_{i-1})$, then $H_i(S)$ is called exhaustive. In other words, for $H_i(S)$ to be exhaustive the i -th step in the production process must be a production only, meaning that the copying process cannot be continued and the component should be stopped 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 $S = AACGTACC$. Moreover, every sequence S has a unique exhaustive history (Lempel and Ziv, 1976).

Let $c(S)$ be the number of components in the exhaustive history of S . It is the least possible number of steps needed to generate S according to the whole Lempel-Ziv algorithm, so $c(S)$ becomes an important complexity indicator. For example, $c(L(R_a)) = 20$, $c(L(R_b)) = 13$, $c(L(R_c)) = 18$, $c(L(R_d)) = 17$.

2.3 Proposed distance and pair-wise distance matrix

Lempel et al have proposed that, for any given sequences Q and S , $c(QS) \leq c(Q) + c(S)$ always holds. This formula shows that the steps required to extend Q to QS are always less than the steps required to build S from ϕ . Recently, Otu et al [20] concluded that the more similar the sequence S is to sequence Q , the smaller $c(QS) -$

$c(Q)$ is. That is $c(QS) - c(Q)$ depends on how much S is similar to Q . Based on this hypothesis, Otu et al have defined five distance measures and used them to successfully construct phylogenetic trees from DNA sequences, which verifies the efficiency of Lempel- Ziv algorithm in analyzing the similarity of linear biological sequences. In order to eliminate the effect of the length on the distance measure $d(S, Q)$, we use following relative distance measures in our applications.

$$d_1(S, Q) = \frac{\max\{c(SQ) - c(S), c(QS) - c(Q)\}}{\max\{c(S), c(Q)\}} \quad (1)$$

$$d_2(S, Q) = \frac{c(SQ) - c(S) + c(QS) - c(Q)}{\max\{c(QS), c(SQ)\}} \quad (2)$$

The first formula corresponds to $d^*(S, Q)$ in [20]. The second one is slightly different from $d_1^*(S, Q)$. We choose to use these formulas mainly because the results are more precise when short RNA primary sequences are compared and analyzed.

Generally, given n RNA secondary structures $R_1, R_2, \dots, R_n$, we can obtain their linear characteristic sequences by the above-mentioned rule, which are $L(R_1), L(R_2), \dots, L(R_n)$. They are linear sequences defined over alphabet $\{A, C, G, U\}$ and $\{a, b, c, d, \dots\}$ and carry the information on RNA secondary structures. Then, by using Lempel-Ziv algorithm, the distance between any pair of structures may be rapidly computed. By arranging them into a matrix, a pairwise distance matrix is obtained, denoted by RD. $RD = (d(R_i, R_j))$ contains the information on the similarity/dissimilarity between any pair of RNA secondary structures.

3 Comparing RNA secondary structures

Phylogenetic relationships among different organisms are of fundamental importance in biology, and one of the prime objectives of DNA sequence analysis 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.

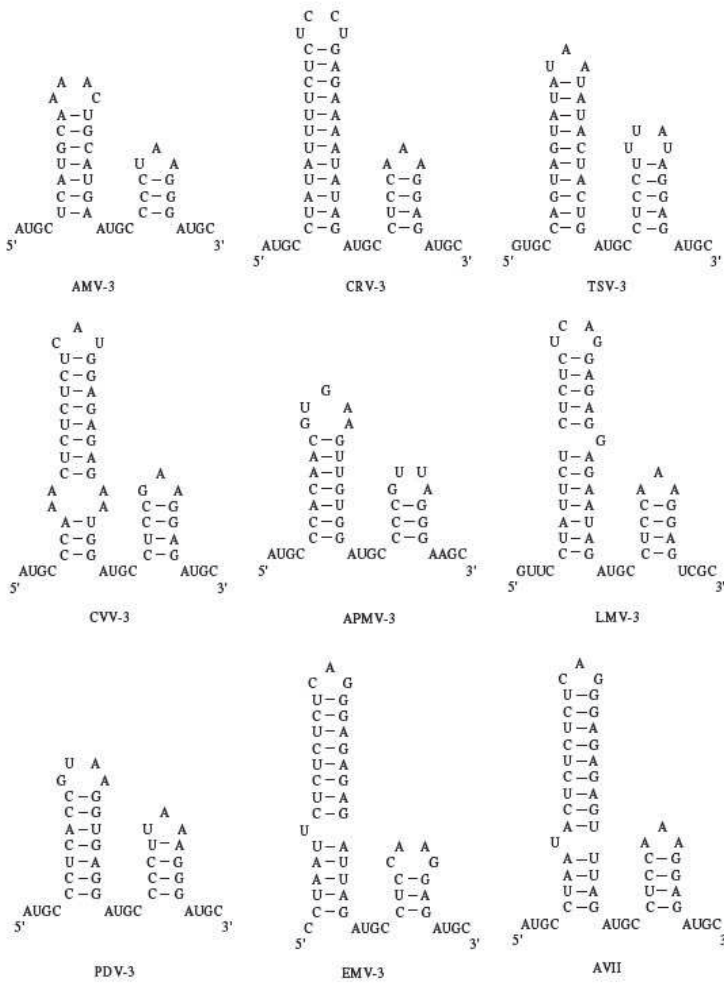


Fig. 2: Secondary structure at the 3'-terminus of RNA 3 of alfalfa mosaic virus (AlMV-3), citrus leaf rugose virus (CiLRV-3), tobacco streak virus(TSV-3), citrus variegation virus (CVV-3), apple mosaic virus (APMV-3), prune dwarf ilarvirus (PDV-3), elm mottle virus (EMV-3) and asparagus virus II (AVII).

The set I consists 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 [8–11]. In Fig. 1, these secondary structures are listed, which were reported by Reusken and Bol [23]. 5S Ribosomal RNA is the smallest RNA component of the ribosomes. Since the nucleotide sequences of 5S rRNA are relatively conserved throughout evolution and are useful for establishing the evolutionary relationship of organisms, the set II consists of the 5S rRNA sequences from 15 representative protozoa species.

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

1. The non-linear complex RNA secondary structures are transformed into linear characteristic sequences.
2. These linear sequences are decomposed according to the rule of Lempel-Ziv algorithm to evaluate the LZ complexity.
3. The similarity degree between any two structures is measured by the distance formulas $d_1(S, Q)$, $d_2(S, Q)$ and $d_3(S, Q)$.
4. After computing the pair distances, we arrange all the values into a matrix for clear and systematical display. This pair-wise distance matrix is listed in the form of Table. It contains the information on the similarity of this set of RNA secondary structures.

We have used our method to analyze the similarity of set I and set II, respectively. The pair-wise distance matrix based on distance measure $d_1(S, Q)$ for nine RNA secondary structures of set I is list in Table 1.

We find that the smaller entries ($d_1 < 0.5$) are associated with AVII and CVV-3 [$d_1 = 0.4074$], AVII and LRMV-3 [$d_1 = 0.4194$], AVII and EMV-3 [$d_1 = 0.4444$], CVV-3 and LRMV-3 [$d_1 = 0.4516$] and CVV-3 and EMV-3 [$d_1 = 0.4815$]. All these secondary structures, AVII, CVV-3, LRMV-3 and EMV-3, have three base-paired regions and three loop regions. We also can observe that AIMV-3, PDV-3 and APMV-3 ($d_1(\text{AIMV} - 3, \text{PDV} - 3) = 0.5652$, $d_1(\text{PDV} - 3, \text{APMV} - 3) = 0.4400$ and $d_1(\text{AIMV} - 3, \text{APMV} - 3) = 0.5600$) are similar to each other. But these three secondary structures have two base-paired regions and two loop regions that are different from the four mentioned secondary structures in topology. Especially notable is that their second hairpins is almost the same.

Table 1: The upper triangular part of the pair-wise distance matrix based on distance measure $d_1(S, Q)$ for nine RNA secondary structures of set I

<i>Species</i>	AIMV-3	CiLRV-3	TSV-3	CVV-3	APMV-3	LRMV-3	PDV-3	EMV-3	AVII
AIMV-3	0	0.6296	0.6818	0.6296	0.5600	0.7419	0.5652	0.7500	0.6667
CiLRV-3		0	0.5185	0.5926	0.6296	0.5484	0.5926	0.5926	0.5556
TSV-3			0	0.5556	0.6400	0.6452	0.5217	0.7083	0.6667
CVV-3				0	0.5926	0.4516	0.5926	0.4815	0.4074
APMV-3					0	0.7742	0.4400	0.7600	0.6296
LRMV-3						0	0.7097	0.5161	0.4194
PDV-3							0	0.7500	0.7407
EMV-3								0	0.4444
AVII									0

In Table 2, we present the pair-wise distance matrix based on distance measure $d_2(S, Q)$ for nine RNA secondary structures of set I.

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

Table 2: The upper triangular part of the pair-wise distance matrix based on distance measure $d_2(S, Q)$ for nine RNA secondary structures of set I

<i>Species</i>	AIMV-3	CiLRV-3	TSV-3	CVV-3	APMV-3	LRMV-3	PDV-3	EMV-3	AVII
AIMV-3	0	0.7179	0.7297	0.7436	0.6944	0.8222	0.6571	0.8250	0.7805
CiLRV-3		0	0.6750	0.7209	0.7619	0.6818	0.7179	0.7442	0.6905
TSV-3			0	0.7073	0.7317	0.7660	0.6857	0.8293	0.7805
CVV-3				0	0.7209	0.6136	0.7317	0.6410	0.5263
APMV-3					0	0.8163	0.5833	0.8605	0.7674
LRMV-3						0	0.8043	0.6667	0.5682
PDV-3							0	0.8293	0.7907
EMV-3								0	0.5833
AVII									0

To further check our result, we use such matrix 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. We construct a relationship tree for the nine RNA secondary structure sequences using the NEIGHBOR method in Phylip 3.66 package. The pairwise evolu-

tionary distance matrix, was formed using distance measure $d_1(S, Q)(d_2(S, Q))$, was fed into the Neighbor-Joining algorithm provided in PHYLIP to construct a phylogeny, respectively. The phylogenies were drawn using TreeView 1.6.6 and shown in Fig. 3 and Fig. 4.

Comparing Fig. 3 and Fig. 4, we can find that the trees obtained by the proposed distance measures $d_1(S, Q)$ and $d_2(S, Q)$ resulted in almost identical topologies. The topology of the tree, except for the positions of the CiLRV-3 and TSV-3, is in agreement for all the part.

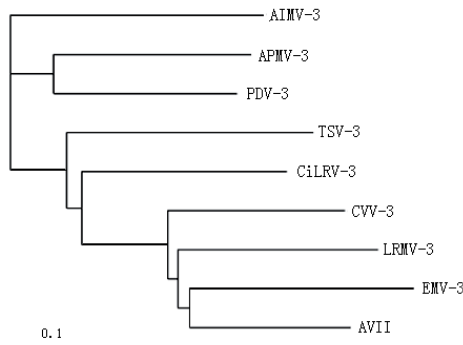


Fig. 3: Neighbor-joining tree for the data in set I. It is obtained by the pair-wise distance matrix based on distance measure $d_1(S, Q)$ and drawn by Treeview program.

It is not difficult to see that the similarity obtained by our method is coincident with that implicated in the tree.

The relationship of nine 3'-terminal structures of viruses is shown reasonably: All the nine molecules are clustered into two group: AIMV-3, PDV-3 and APMV-3 are clustered into one group closely. AVII, CVV-3, LRMV-3, EMV-3, CiLRV-3 and TSV-3 are clustered into another group. AVII, CVV-3, LRMV-3 and EMV-3 are grouped closely. But they are less closely with CiLRV-3 and TSV-3, which is consistent with the fact that the former and latter have some different loop and stem regions.

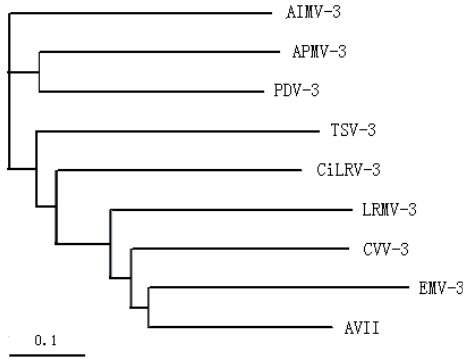


Fig. 4: Neighbor-joining tree for the data in set I. It is obtained by the pair-wise distance matrix based on distance measure $d_2(S, Q)$ and drawn by Treeview program.

In order to verify the generality of our method, we apply it to the set II. The 15 complex RNA secondary structures of set II are from [23] [24]. These 5S rRNA are: *Crithidia fasciculata*, *Bresslaia 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).

After computing the pair distances by using $d_1(S, Q)$, we obtain the pair-wise distance matrix. When it was fed into the Neighbor-Joining algorithm provided in PHYLIP, we constructed a phylogenetic tree of these protozoa (Fig. 5).

We omit the corresponding results for the distance measure $d_2(S, Q)$ as they are identical to the ones presented here.

The tree shows that unicellular green flagellate *Chlamydomonas* and Unicellular green algae *Chlorella*, *Scenedesmus* and plants belong to the same branch in the tree, *Chlorella* and *Scenedesmus* are closely related to each other as expected. *Chlamydomonas* is closer to the sequences of green algae and plants than to other protozoa and animals.

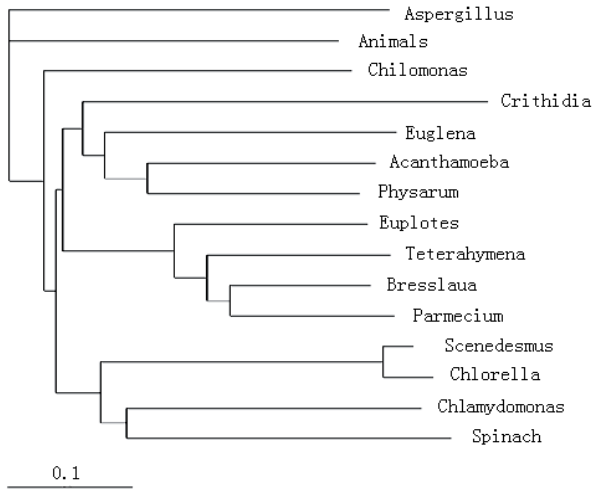


Fig. 5: Neighbor-joining tree for the data in set II. It is obtained by the pair-wise distance matrix based on distance measure $d_1(S, Q)(d_2(S, Q))$ and drawn by Treeview program.

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; Euglena has been classified either as related to plants (by botanists) or to animals (by zoologists) [23]. We also suggest that it is much closer to animals than to plants. This has been also suggested by other biochemical data. The topology of the tree, except for the positions of the Chilomonas, is generally in agreement with the classification by taxonomic criteria [23].

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 secondary structure using these techniques, it is necessary to transform an RNA secondary structure into linear

sequence. 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 stacks and loops of the RNA secondary structure. So the linear sequence contains all the information that contained in the RNA primary sequence and RNA secondary structure. Then comparison of two RNA secondary structures is now transformed into a comparison of linear sequences of the corresponding RNA secondary structures. When these linear sequences are obtained, we can use the Lempel-Ziv algorithm to analyze the RNA secondary structure 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 phylogenetic tree for two data sets, which is consistent with results reported in other literature.

Acknowledgements

The authors would like to thank Hasan H. Otu for providing their algorithms. This work is supported by Shandong Natural Science Foundation (Y2006A14).

References

- [1] V. Bafna, S. Muthukrishnan, R. Ravi, Comparing similarity between RNA strings, *6th Annual Symposium on Combinatorial Pattern Matching* **937** (1995) 1–16.
- [2] F. Corpet, B. Michot, RNAlign program: alignment of RNA sequences using both primary and secondary structures, *Comput. Appl. Biosci.* **10** (1994) 389–399.
- [3] B. Shapiro, An algorithm for comparing multiple RNA secondary structures, *Comput. Appl. Biosci.* **4** (1988) 387–393.
- [4] S. Y. Le, J. Owens, R. Nussinov, J. H. Chen, B. Shapiro, J. V. Maizel, RNA secondary structures: comparison and determination of frequently recurring sub-structures by consensus, *Comput. Appl. Biosci.* **5** (1989) 205–210.

- [5] S. Y. Le, R. Nussinov, J. V. Maizel, Tree graphs of RNA secondary structures and their comparisons, *Comput. Biomed. Res.* **22** (1989) 461–473.
- [6] B. Shapiro, K. Zhang, Comparing multiple RNA secondary structures using tree comparisons, *Comput. Appl. Biosci.* **6** (1990) 309–318.
- [7] S. Dulucq, L. Tichit, RNA secondary structure comparison: exact analysis of the Zhang-Shasha tree edit algorithm, *Theor. Comput. Sci.* **306** (2003) 471–484.
- [8] B. Liao, T. Wang, A 3D graphical representation of RNA secondary structure, *J. Biomol. Struct. Dynamics* **21** (2004) 827–832.
- [9] Y. Yao, B. Liao, T. Wang, A 2D graphical representation of RNA secondary structures and the analysis of similarity/dissimilarity based on it, *J. Mol. Struct. (Theochem)* **755** (2005) 131–136.
- [10] Y. S. Zhang, On 3D graphical representation of RNA secondary structure, *MATCH Commun. Math. Comput. Chem.* **57** (2007) 157–168.
- [11] Y. S. Zhang, On 2D graphical representation of RNA secondary structure, *MATCH Commun. Math. Comput. Chem.* **57** (2007) 697–710.
- [12] S. R. Eddy, Non-coding RNA genes and the modern RNA world, *Nature Rev. Genet.* **2** (2001) 919–929.
- [13] G. Storz, An expanding universe of noncoding RNAs, *Science* **296** (2002) 1260–1263.
- [14] T. Schlick, *Molecular Modeling and Simulation: An Interdisciplinary Guide*, Springer-Verlag, New York, 2002.
- [15] M. Li, P. Vitanyi, *An Introduction to Kolmogorov Complexity and Its Applications*, Springer-Verlag, New York, 1997.
- [16] 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.
- [17] X. Chen, S. Kwong, M. Li, A compression algorithm for DNA sequences and its applications in genome comparison, *Genome Informatics* **10** (1999) 51–61.
- [18] H. Sato, T. Yoshioka, A. Konagaya, T. Toyoda, DNA data compression in the post genome era, *Genome Informatics* **12** (2001) 512–514.

- [19] J. Ziv, A. Lempel, Compression of individual sequences by variable rate coding, *Inform. Theory, IEEE IPrans.* **24** (1978) 530–536.
- [20] H. H. Otu, K. Sayood, A new sequence distance measure for phylogenetic tree construction, *Bioinformatics* **19** (2003) 2122–2130.
- [21] N. Liu, T. Wang, A method for rapid similarity analysis of RNA secondary structures, *BMC Bioinformatics* **7** (2006) 493–503.
- [22] C. Li, A. Wang, L. Xing, Similarity of RNA secondary structures, *J. Comput. Chem.* **28** (2007) 508–512.
- [23] T. Kumazaki, H. Hori, S. Osawa, Phylogeny of protozoa deduced from 5S rRNA sequences, *J. Mol. Evol.* **19** (1983) 411–419.
- [24] M. Szymanski, Z. Mirosława, V. A. Barciszewska, E. J. Barciszewski, 5S ribosomal RNA database, *Nucleic. Acids. Res.* **30** (2002) 176–178.
- [25] J. Felsenstein, PHYLIP-Phylogeny inference package (version 3.2), *Cladistics* **5** (1989) 164–166.
- [26] R. D. M. Page, Tree View: An application to display phylogenetic trees on personal computers, *Comput. Appl. Biosci.* **12** (1996) 357–358.
- [27] 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.
- [28] N. D. Levine, J. O. Corliss, F. E. G. Cox, G. Deroux, J. Grain, B. M. Honigberg, G. F. Leedale, A. R. Loeblich III, J. Lom, D. H. 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.