Comparing RNA Secondary Structures Based on LZ Complexity

Liwei Liu¹, Fenglan Bai¹, Tianming Wang²

¹College of Science
Dalian Jiaotong University
Dalian, 116028, China
²School of Mathematical Sciences
Dalian University of Technology
Dalian, 116024, China
daliguowei@163.com, wangtm@dlut.edu.cn

Abstract

In this paper, we propose a new similarity measure to compare RNA secondary structure. We first transform an RNA secondary structure into three characteristic sequences. Then, based on these characteristic sequences, we calculate their LZ complexity. Finally, we obtain the similarity/dissimilarity matrix based on the LZ complexity, and make a comparison for the secondary structures at the 3'-terminus belonging to nine different species. The proposed method does not require multiple alignments and is easy to operate. This method will also be useful to researchers who are interested in evolutionary analysis.

Keywords: LZ Complexity; RNA secondary structure; Comparison; Similarity.

1. Introduction

It is well known that ribonucleic acid (RNA) is an important molecule which performs a wide range of functions in the biological system. RNA has recently become the center of much attention because of its catalytic properties, leading to an increased interest in obtaining structural information. Recently Liao et al. [1], [2] and Liu et al. [3] used graphs to represent RNA secondary structures and then derived some invariants from graphs to compare RNA secondary structures. Zhang and Wang proposed another method to analyze RNA secondary structures based on the LZ algorithm [4]. Shapiro et al. used tree models to compare RNA secondary structures [5]. Hofacker et al. compared RNA secondary structures by aligning
the corresponding base pairing probability matrices that were computed by McCaskill's partition function algorithm [6].

In this paper, similar with the comparison of TOPS strings [3], we use the LZ Complexity to compare RNA secondary structure. In Fig. 1, the secondary structures at the 3'-terminus belonging to nine different viruses are listed, which were reported by John F. Bol [7].

2. RNA secondary structure sequences

The secondary structure of an RNA is a set of free bases and base pairs formed bonds between A-U and G-C. Following Zuker, we assume a model where there are no knots in the secondary structure. This means that for the secondary structure, the bonds are non-crossing. In this paper we think of base pair G-U as free bases, although the pairing of G and U is frequently allowed. Let A', U', G', C' denote A, U, G, C in the base pair A-U and G-U, respectively. Then we can obtain a special sequence representation of the secondary structure. We call it characteristic sequence of the secondary structure.

Fig. 1 Secondary structure at the 3'-terminus of RNA 3 of alfalfa mosaic virus (AMV-3 [8]), citrus leaf rugose virus (CiLRV-3 [9]), tobacco streak virus (TSV-3 [10], [11]), citrus variegation virus (CVV-3 [9]), apple mosaic virus (APMV-3 [12]), prune dwarf
ilarvirus (PDV-3 [13]), lilac ring mottle virus (LRMV-3 [14]), elm mottle virus (EMV-3 [15]) and asparagus virus II (AVII [16]). Numbering of nucleotides is from the 3' end of RNA 3.

For example, the corresponding characteristic sequence of the substructure of AlMV-3 (Fig. 2) is CGUAG’G’G’AAUC’C’C’CG (from 3’ to 5’).

![Fig. 2 Substructure of AlMV-3](image)

Let $G = g_1g_2 \cdots$ be a characteristic sequence of an RNA secondary structure. Given characteristic sequence of an RNA secondary structure, its bases can be divided into three classes according to their chemical structure, i.e., non-$A'(A') = G, C, U, G', C', U'$, non-$G(G') = A, C, U, A', C', U'$ and non-$C(C') = A, G, U, A', G', U'$ by labelling the elements of non-$A'(A')$, non-$G(G')$ and non-$C(C')$ by $N$, and the elements of $A(A'), G(G'),$ and $C(C')$ by itself, respectively. We thus obtain three sequences, named as non-$A'(A')$, non-$G(G')$ and non-$C(C')$, respectively, for convenience. Similarly, we can define non-$U(U')$, but it is dependent on the others. For example, the corresponding non-$G(G')$ sequence of the substructure of AlMV-3 (Fig. 2) is NGNNG’G’G’NNNNNG (from 3’ to 5’).

3. LZ algorithm and LZ complexity

LZ algorithm was developed to analyze the complexity of linear sequences by Lempel and Ziv in 1976 [17]. In recent years, some authors applied the algorithm construct phylogenetic analysis by using conditional LZ complexity. Otu et al. applied LZ algorithm to phylogenetic analysis and had successfully constructed phylogenetic trees for real and simulated DNA data sets [19]. Liu and Wang used LZ algorithm to compare TOPS strings, and construct phylogenetic trees [20]. Then, we will give some basic definition.

Let $S, Q$ and $R$ be sequences over a finite alphabet $A$, $L(S)$ be the length of $S$, $S(i)$ be the $i$th element of $S$ and $S(i, j)$ be the subsequence of $S$ that starts at position $i$ and ends at position $j$. Note that $S(i, j) = \phi$, for $i > j$. The concatenation of $Q$ and $R$ forms a new sequence $S = QR$, where $Q$ is called a prefix of $S$, and $S$ is called an extension of $Q$ if there exists an integer $i$ such that $Q = S(1, i)$.

An extension $S = QR$ of $Q$ is reproducible from $Q$ denoted by $Q \rightarrow S$, if there exists an integer $p \leq L(Q)$ such that $R(k) = S(p + k - 1)$, for $k = 1, 2, \ldots, L(R)$. For example: $CCA \rightarrow CCACACAC$ with $p = 2$. A non-null sequence $S$ is producible from its prefix $S(1, j)$, denoted by $S(1, j) \rightarrow S$, if $S(1, j) \rightarrow S(1, L(S) - 1)$. For example: $GC \Rightarrow GCGG$ with $p = 1$. 
Any non-null sequence $S$ can be built from a production process by iterative self-deleting-building process where at the $i$th step $S(1,h_{i-1}) \Rightarrow S(1,h_i)$, $\phi = S(1,0) \Rightarrow S(1,1)$. An m-step production process of $S$ leads to a parsing of $S$ into $H(S) = S(1,h_1) \cdot S(h_1+1,h_2) \cdot \cdots \cdot S(h_{m-1}+1,h_m)$, which is called the history of $S$, and $S(h_{i-1}+1,h_i) = H_i(S)$ is called the $i$th component of $H(S)$.

A component $H_i(S)$ and the corresponding production step $S(1,h_{i-1}) \rightarrow S(1,h_i)$ are called exhaustive if $S(1,h_{i-1}) \rightarrow S(1,h_i)$ is not true. A history is called exhaustive if each of its components is exhaustive.

Let $S = S_1S_2 \cdots S_n$ be a non-null sequence, then produce $S$ from null sequence according the following algorithm.

1. At beginning we have a null-sequence $\phi$, then add prefix $S = S_1$, if $n > 1$, need add a dot after $S_1$.

2. Let a prefix $Q = S_1S_2 \cdots S_r$, $0 < r < n$ be available, check if $R = S_r+1$ can be reproduced from $S(1,r)$. If $R$ cannot be reproduced from a subsequence of $S(1,r)$, then join $Q$ and $R$ to get a new prefix $QR$, and add a dot following $QR$. If $R = S_r+1$ can be reproduced from a subsequence of $S(1,r)$, then check again if $R = S_{r+1}S_{r+2}$ can be reproduced from $S(1,r+1)$. If so, check again if $R = S_{r+1}S_{r+2}S_{r+3}$ can reproduce from $S(1,r+2) \cdots$ and so on. There are two possible cases. In the case $R = S_{r+1} \cdots S_n$, then we end the procedure, and get new prefix $QR = S$, in another case $R = S_{r+1} \cdots S_k$ cannot be reproduced from any subsequence of $S(1,k-1)$, then get a new prefix $QR$ and add a dot behind it.

3. Repeat the step (2) until produce $S$.

This algorithm is LZ algorithm. From LZ algorithm we know that, history of $S$ is exhaustive because each of its components is exhaustive. We know from production process that what’s more important, the exhaustive history of any non null sequence is unique [17].

Let $c(S)$ be the number of components in the exhaustive history of sequence $S$, then $c(S)$ is called LZ complexity of $S$. It has been proved that $c(S)$ is the least possible number of steps needed to generate $S$ according to LZ algorithm of production process [17]. For instance, LZ complexity of the sequence $S = GGGUGUUGUGU$ is 4, and this sequence can be generated through the following steps, where $\bullet$ is used to separate the decomposition component:

1. generating a novel symbol $G$: $\phi + G = G \bullet$
2. copying the longest component + generating a additional symbol $U$: $G + GGU = G \bullet GGU \bullet$
3. repeat the step 2: $G \bullet GGU + GGU = G \bullet GGU \bullet GGU \bullet$
4. repeat the step 2: $G \bullet GGU \bullet GUU + UGUG = G \bullet GGU \bullet GUU \bullet UGUG$.

Then, we obtained the exhaustive history of $S$, denoted by $EH(S) = G \bullet GGU \bullet GUU \bullet UGUG$, so $c(S) = 4$.

The flow diagram for the algorithm to calculate the $c(S)$ of a string $S = S_1S_2 \cdots S_n$ is shown by Fig. 3.
4. Computation of similarity/dissimilarity matrix

According to the paper [17], for any given sequences $Q$ and $S$, the following property always remains valid: $c(QS) \leq c(Q) + c(S)$. This formula shows that the steps extend $Q$ to $QS$ are always less than the steps required to build $S$ from $\phi$. Furthermore the more similar the sequence $S$ is to sequence $Q$, the smaller $c(QS) - c(Q)$ is [19]. That is $c(QS) - c(Q)$ depends on the degree to which $S$ is similar to $Q$.

For example, let $S = AAGCUAGGAUUC$, $R = GUACCCAGUUU$ and $Q = AGCCUGAGGAA$. The exhaustive histories of these sequences would be: $EH(S) = A \cdot AG \cdot C \cdot U \cdot AGG \cdot AU \cdot UC$, $EH(R) = G \cdot U \cdot A \cdot C \cdot CCA \cdot GUU \cdot AU$ and $EH(Q) = A \cdot G \cdot C \cdot CU \cdot GA \cdot GG \cdot AA$. Yielding $c(S) = c(R) = c(Q) = 7$. The exhaustive histories of the sequences $SQ$ and $RQ$ would be: $EH(SQ) = A \cdot AG \cdot C \cdot U \cdot AGG \cdot AU \cdot UC \cdot AGCC \cdot UG \cdot AGGAA$ and $EH(RQ) = G \cdot U \cdot A \cdot C \cdot CCA \cdot GUU \cdot AU \cdot AGC \cdot CU \cdot GA \cdot GG \cdot AA$, respectively. So, $c(SQ) = 10$, $c(RQ) = 12$. Note that it took three steps to build $Q$ in the production process of $SQ$. On the other hand, we used five steps to generate $Q$ in the production process of $RQ$. The reason that it took more steps in the second case is $Q$ is 'closer' to $S$ than $R$. In this example we can observe this by looking at the patterns AGC and AGG which $Q$ and $S$ share. We can formulate the number of steps it takes to generate a sequence $Q$ from a sequence $S$ by $c(SQ) - c(S)$. Thus, if $S$ is closer to $Q$ than $R$ then we would expect $c(SQ) - c(S)$ to be smaller than $c(RQ) - c(R)$ as is the case in the above example.
We define the similarity measure as

\[
d(S, Q) = \begin{cases} 
\sqrt{(c(SQ) - c(S))^2 + (c(QS) - c(Q))^2}, & S \neq Q \\
0, & S = Q 
\end{cases}
\]  

(1)

Next, we will consider characteristic sequences of 9 RNA secondary structures and calculate their similarity measure. By arranging all these similarity measure into a matrix, a pair-wise similarity/dissimilarity matrix is derived. This similarity/dissimilarity matrix contains the similarity information on 9 RNA secondary structures.

Table 1 The similarity/dissimilarity matrix for the 9 RNA secondary structures of Fig. 1 based on the non-A(A') sequences

| Species | AIM V-3 | CILR V-3 | TSV-3 | CVV-3 | APM V-3 | LRM V-3 | PDV-3 | EMV-3 | AVII |
|---------|---------|----------|-------|-------|---------|---------|-------|-------|------|
| AIMV-3  | 0       | 2.231    | 0.2839| 0.2861| 0.2703  | 0.2676  | 0.2496| 0.2704|      |
| CILRV-3 | 0       | 0.2784   | 0.2121| 0.2426| 0.2170  | 0.2564  | 0.2231| 0.1911|      |
| TSV-3   | 0       | 0.2562   | 0.2646| 0.2925| 0.2575  | 0.3023  | 0.2803|       |      |
| CVV-3   | 0       | 0.2737   | 0.2305| 0.2575| 0.2364  | 0.2070  |       |       |      |
| APMV-3  | 0       | 0.3129   | 0.2726| 0.2901| 0.2803  |         |       |       |      |
| LRMV-3  | 0       | 0.2492   | 0.2001| 0.2233|         |         |       |       |      |
| PDV-3   | 0       | 0.2797   | 0.2638|         |         |         |       |       |      |
| EMV-3   | 0       | 0.1883   |       |         |         |         |       |       |      |
| AVII    | 0       |          |       |         |         |         |       |       |      |

From Table 1, we find that the most similar are EMV-3 and AVII with the lowest value 0.1883. The more similar are CILRV-3 and AVII with a value of 0.1911, LRMV-3 and EMV-3 with a value of 0.2001, CVV-3 and AVII with a value of 0.2070.

Table 2 The similarity/dissimilarity matrix for the 9 RNA secondary structures of Fig. 1 based on the non-G(G') sequences

| Species | AIM V-3 | CILR V-3 | TSV-3 | CVV-3 | APM V-3 | LRM V-3 | PDV-3 | EMV-3 | AVII |
|---------|---------|----------|-------|-------|---------|---------|-------|-------|------|
| AIMV-3  | 0       | 2.676    | 2.957 | 2.658 | 2.703   | 2.800   | 2.531 | 2.800 | 2.978|
| CILRV-3 | 0       | 0.2407   | 0.2416| 0.3020| 0.2020  | 0.2614  | 0.2144| 0.1953|      |
| TSV-3   | 0       | 0.2571   | 0.2839| 0.2618| 0.2479  | 0.2514  | 0.2472|       |      |
| CVV-3   | 0       | 0.2850   | 0.2531| 0.2472| 0.2206  | 0.2111  |       |       |      |
| APMV-3  | 0       | 0.3991   | 0.1964| 0.2966| 0.2977  |         |       |       |      |
| LRMV-3  | 0       | 0.2618   | 0.2121| 0.2305|         |         |       |       |      |
| PDV-3   | 0       | 0.2708   | 0.2767|         |         |         |       |       |      |
| EMV-3   | 0       | 0.1779   |       |         |         |         |       |       |      |
| AVII    | 0       |          |       |         |         |         |       |       |      |
From Table II, we find that the most similar are EMV-3 and AVII with the lowest value 0.1779. The more similar are CiLRV-3 and AVII with a value of 0.1953, APMV-3 and PDV-3 with a value of 0.1964, CiLRV-3 and LRMV-3 with a value of 0.2020.

Table 3 The similarity/dissimilarity matrix for the 9 RNA secondary structures of Fig. 1 based on the non-C(C') sequences

| Species | AIMV-3 | CiLRV-3 | TSV-3 | CVV-3 | APMV-3 | LRMV-3 | PDV-3 | EMV-3 | AVII |
|---------|--------|---------|-------|-------|--------|--------|-------|-------|------|
| AIMV-3  | 0      | 0.3010  | 0.2800| 0.2797| 0.2297 | 0.2676 | 0.2676| 0.3104| 0.3049|
| CiLRV-3 | 0      | 0       | 0.2631| 0.2850| 0.2726 | 0.2564 | 0.2828| 0.2357| 0.2170|
| TSV-3   | 0      | 0       | 0     | 0.2828| 0.2708 | 0.2652 | 0.2737| 0.2767|
| CVV-3   | 0      | 0       | 0     | 0.2726| 0.2893 | 0.2492 | 0.2632| 0.2538|
| APMV-3  | 0      | 0       | 0     | 0     | 0.2593 | 0.2231 | 0.2937| 0.2867|
| LRMV-3  | 0      | 0       | 0     | 0     | 0     | 0.2593 | 0.2475| 0.2100|
| PDV-3   | 0      | 0       | 0     | 0     | 0     | 0     | 0.2937| 0.2638|
| EMV-3   | 0      | 0       | 0     | 0     | 0     | 0     | 0     | 0.1563|
| AVII    | 0      | 0       | 0     | 0     | 0     | 0     | 0     | 0     | 0    |

From Table III, we find that the most similar are EMV-3 and AVII with the lowest value 0.1563. The more similar are CiLRV-3 and AVII with a value of 0.2170, APMV-3 and PDV-3 with a value of 0.2231, LRMV-3 and AVII with a value of 0.2305.

Observing Tables I, II, and III, we find that the most similar are EMV-3 and AVII.

5. Conclusions and Discussions

Comparing RNA secondary structures is a key topic in bioinformatics when analyzing the functional similarities of different RNA secondary structures. In this paper, we have proposed a novel method to compare RNA secondary structures. We map an RNA secondary structure into non-A(A'), non-G(G') and non-C(C') sequences. Then, we obtain the similarity/dissimilarity matrix based on the LZ complexity, and compare the similarity of RNA secondary structures. The main advantage is that this algorithm can extract repeated when two sequences are compared, the subsequences that they share can be detected.

Acknowledgment

This work is partially supported by the Scientific Research Fund of Liaoning Province Educational Department Grant #L2010074, NSFC Grant #10871219, NSFC Grant #61003191 and the Scientific Research Fund of Liaoning Province Educational Department Grant #2009A125.

References

[1] B. Liao, J. Luo, R. Li and W. Zhu, "RNA secondary structure 2D graphical representation without degeneracy," Int. J. Quantum. Chem., vol. 106, no. 8, pp. 1749-1755, 2006.
[2] W. Zhu, B. Liao and K. Ding, "A condensed 3D graphical representation of RNA secondary structures," J. Mol. Struct. THEOCHEM., vol. 757, pp. 193-198, 2005.

[3] L. Liu and T. Wang, "On 3D graphical representation of RNA secondary structures and their applications," J. Math. Chem., vol. 42, pp. 595-602, 2007.

[4] S. Zhang and T. Wang, "A Complexity-based Method to Compare RNA Secondary Structures and its Application," J. Biomol. Struct. Dyn., vol. 28, no. 2, pp. 247-258, 2010.

[5] B. Shapiro and K. Zhang, "Comparing multiple RNA secondary structures using tree comparisons," Comput. Appl. Biosci., vol. 6, pp. 309-318, 1990.

[6] I.L. Hofacker, S.H.F. Bernhart, and P.F. Stadler, "Alignment of RNA base pairing probability matrices," Bioinformatics, vol. 20, pp. 2222-2227, 2004.

[7] Chantal B.E.M. Reusken and J.F. Bol, "Structural elements of the 3’-terminal coat protein binding site in alfalfa mosaic virus RNAs," Nucl. Acids. Res., vol. 14, pp. 2660-2665, 1996.

[8] E.C. Koper-Zwarthoff, F.Th. Brederode, P. Walstra and J.F. Bol, "Nucleotide sequence of the 3’-noncoding region of alfalfa mosaic virus RNA 4 and its homology with the genomic RNAs," Nucl. Acids. Res., vol. 7, pp. 1887-1900, 1979.

[9] S.W. Scott and X. Ge, "The complete nucleotide sequence of RNA 3 of citrus leaf rugose and citrus variegation ilarviruses," J. Gen. Virol., vol. 76, pp. 957-963, 1995.

[10] E.C. Koper-Zwarthoff, F.Th. Brederode, P. Walstra and J.F. Bol, "The complete nucleotide sequence of RNA 3 of citrus leaf rugose and citrus variegation ilarviruses," Nucl. Acids. Res., vol. 8, pp. 3307-3318, 1980.

[11] B.J. Cornelissen, H. Janssen, D. Zuidema and J.F. Bol, "Complete nucleotide sequence of tobacco streak virus RNA 3," Nucl. Acids. Res., vol. 12, pp. 2427-2437, 1984.

[12] R.H. Alrefai, P. Shicl, L.L. Domier, C.J. D’Arcy, P.H. Berger and S.S. Korban, "The nucleotide sequence of apple mosaic virus coat protein gene has no similarity with other Bromoviridae coat protein genes," J. Gen. Virol., vol. 75, pp. 2847-2850, 1994.

[13] S.W. Scott and X. Ge, "The complete nucleotide sequence of the RNA 3 of lilac ring mottle ilarivirus," J. Gen. Virol., vol. 76, pp. 1801-1806, 1995.

[14] E.J. Bachman, S.W. Scott, G. Xin and V. Bowman Vance, "The complete nucleotide sequence of prune dwarf ilarivirus RNA 3: implications for coat protein activation of genome replication in ilarviruses," Virology, vol. 201, pp. 127-131, 1994.

[15] F. Houser-Scott, M.L. Baer, K.F. Liem, J.M. Cai and L. Gehrke, "Nucleotide sequence and structural determinants of specific binding of coat protein or coat protein peptides to the 3’ untranslated region of alfalfa mosaic virus RNA 4," J. Virol., vol. 68, pp. 2194-2205, 1994.

[16] EMBL/GenBank/DDBJ databases. Accession no.X86352.

[17] A. Lempel and J. Ziv, "On the complexity of finite sequences," IEEE Trans. Inform. Thory., vol. 22, pp. 75-81, 1976.

[18] S. Zhang and T. Wang, "Phylogenetic Analysis of Protein Sequences Based on Conditional LZ Complexity," MATCH Commun. Math. Comput. Chem., vol. 63, pp. 701-716, 2010.

[19] H.H. Otu and K. Sayood, "A new sequence distance measure for phylogenetic tree construction," Bioinformatics, vol. 19, pp. 2122-2130, 2003.

[20] L. Liu and T. Wang, "Comparison of TOPS Strings Based on LZ Complexity," J. Theor. Biol., vol. 251, pp. 159-166, 2008.