BMC Bioinformatics

Methodology article

A method for rapid similarity analysis of RNA secondary structures

Na Liu*[1,2] and Tianming Wang[2,3]

Address: [1]Department of Applied Mathematics, Dalian University of Technology, Dalian 116024, China, [2]College of Advanced Science and Technology, Dalian University of Technology, Dalian 116024, China and [3]Department of Mathematics, Hainan Normal University, Haikou 571158, China

Email: Na Liu* - liunasophia@163.com; Tianming Wang - wangtm@dlut.edu.cn

* Corresponding author

Published: 08 November 2006

BMC Bioinformatics 2006, **7**:493 doi:10.1186/1471-2105-7-493

Received: 24 March 2006
Accepted: 08 November 2006

This article is available from: http://www.biomedcentral.com/1471-2105/7/493

Abstract

Background: Owing to the rapid expansion of RNA structure databases in recent years, efficient methods for structure comparison are in demand for function prediction and evolutionary analysis. Usually, the similarity of RNA secondary structures is evaluated based on tree models and dynamic programming algorithms. We present here a new method for the similarity analysis of RNA secondary structures.

Results: Three sets of real data have been used as input for the example applications. Set I includes the structures from *5S rRNAs*. Set II includes the secondary structures from *RNase P* and *RNase MRP*. Set III includes the structures from *16S rRNAs*. Reasonable phylogenetic trees are derived for these three sets of data by using our method. Moreover, our program runs faster as compared to some existing ones.

Conclusion: The famous Lempel-Ziv algorithm can efficiently extract the information on repeated patterns encoded in RNA secondary structures and makes our method an alternative to analyze the similarity of RNA secondary structures. This method will also be useful to researchers who are interested in evolutionary analysis.

Background

RNA secondary structures play an important role in determining the functions of RNA molecules. Some of them have been accepted as good data for evolutionary analysis. With the completion of the sequencing of the genomes of human and other species, major structural biology resources have been harnessed to predict functions. More and more RNA structures are accumulated and we know little about their functions. This calls for the development of cost-effective computational methods to predict RNA functions, which will provide preliminary information for biologists and guide biological experiments. Earlier studies usually adopt dynamic programming algorithms and tree models. Shapiro et al [1] proposed to compare RNA secondary structures by using tree models. Hofacker et al [2] compared RNA secondary structures by aligning the corresponding base pairing probability matrices that were computed by McCaskill's partition function algorithm [3]. Because these methods rely on dynamic programming algorithms, they are compute-intensive. Constructing tree models is based on the idea that the stems or helices dominantly stabilize the secondary structures. So they ignore their primary sequences and focus on so-called elementary units (stem and loop, etc) for the similarity analysis. There are other works, in which tree models were constructed to analyze the similarity of RNA secondary struc-

tures [4-8]. Recently Liao et al [9] have proposed to use graphs to represent RNA secondary structures and then derive some invariants from graphs to compare RNA secondary structures. This idea is from the study of DNA sequences [10-13]. It has been stated [10] that invariants actually reflect some characterizations of biological structures or sequences and may be regarded as indicators. Some information will be lost, however, and how to obtain and select suitable invariants to characterize biological sequences so as to compare DNA sequences effectively is still unsolved. What's more, the graphical representations don't work well when the size of the RNA secondary structure is large. Obviously, for complex RNA secondary structures, more information is lost, which will affect the similarity analysis. Popular tools for optimal alignment of RNA secondary structures include RNAdistance [1], RNAforester [14] etc. RNAdistance uses the tree models to coarsely represent RNA secondary structures, and compares RNA secondary structures based on tree edit distance measure. RNAforester supports the computation of pairwise and multiple alignment of structures based on tree alignment measure.

In this paper we propose a novel method for the similarity analysis of RNA secondary structures, where pseudoknots are also taken into account. In our approach, each secondary structure is transformed into a linear sequence. The linear sequence not only contains the information on the corresponding RNA primary structure, but also contains the information on the base pairing.

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 three sets of real data. The results obtained by our method are comparable to those given by other authoritative methods. What's more, the whole process is easy to operate. It can yield results rapidly.

Results

Materials

Three sets of real data are used to test our method. RNA secondary structures in set II are from *RNase P* and *RNase MRP*. They are distantly related and there is little sequence homology between them. These secondary structures are used to test distant RNA secondary structures. They are mainly obtained from the *RNase P* Database [16] and the remaining secondary structures are obtained from [17]. The names of the RNA secondary structures from *RNase P* are: *Synechocystis sp.PCC6803*, *Anacystis nidulans PCC6301*, *Pseudoanabaena sp.PCC6903*, *Anabaena sp.PCC7120*, *Porphyra purpurea chloroplast*, *Thermotoga maritima*, *Agrobacterium tumefaciens*, *Rhodospirillum rubrum*, *Bacillus subtilis*, *Reclinomonas americana mitochondria*, *Sulfolobus acidocaldarius*, *Methanococcus jannaschii*, *Halobacterium cutirubrum*, *Human (nuclear) P*. The RNA secondary structures from *RNase MRP* are obtained from [17], whose names are: *Human*, *Bovine*, *Mouse*, *Rat*. RNA secondary structures in set I are from *5S rRNAs*. They are provided by Maciej Szymanski, who has developed the 5S Ribosomal RNA Database [18]. The names of the *5S rRNAs* used in our study are *Halobacterium spl*, *Pyrodictium occultum*, *Sulfolobus spl*, *Actinia equina*, *Dicyema misakiense*, *Basidiobolus magnus*, *Chrysaora quinque*, *Christiansenis pallida* and *Planocera recticulata*. RNA secondary structures in set III are from *16S rRNAs*. The names of the *16S rRNAs* are *Thermoproteus tenax*, *Halobacterium*, *Bacteoides*, *Bacillus*, *Mus musculus*, *Synechococcus*, *Thermotoga*, *Saccharomyces cerevisiae*, *Homo sapiens*, *Escherichia coli*, *Methanococus vannielli*, *Thermococcus celer*, *Vairimorpha* and *Methanobacterium*.

The similarity analysis of set I and set II by using our method

The goal of our study is to compare RNA secondary structures and analyze their similarity. Given a set of RNA secondary structures, our method requires the following main operations for the similarity analysis: Firstly the non-linear complex RNA secondary structures are transformed into linear characteristic sequences. Secondly, these linear sequences are decomposed according to the rule of Lempel-Ziv algorithm to evaluate the LZ complexity. Thirdly, the similarity degree between any two structures is measured by our distance formula, as shown in Method section. Lastly, by arranging all the values into a matrix, we obtain a pair-wise distance matrix. 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. Its validity may be better reflected by its application to reconstruct phylogenetic trees. Hence, for the two sets of data, we input their pair-wise distance matrices, obtained by our methods, into the Neighbor program in the Phylip package [19], respectively. By choosing Neighbor-joining option, we obtain two phylogenetic trees for the two sets, which are drawn by Treeview program [20] and are shown in Figure 1 and Figure 2.

Discussion

Lempel-Ziv algorithm is an algorithm that is related to minimal length encoding. Its successful application to the evolutionary analysis of DNA sequences has indicated that Lempel-Ziv algorithm is an alternative to the similarity analysis of biological sequences. To our knowledge, the concept of applying Lempel-Ziv algorithm to the similarity analysis of RNA secondary structures hasn't been adopted by any other researcher. The introduction of our method in Method section indicates that this is a relatively simple and rapid method for the similarity analysis of RNA secondary structures. We owe the efficiency of this method mainly to the Lemple-Ziv algorithm, which can

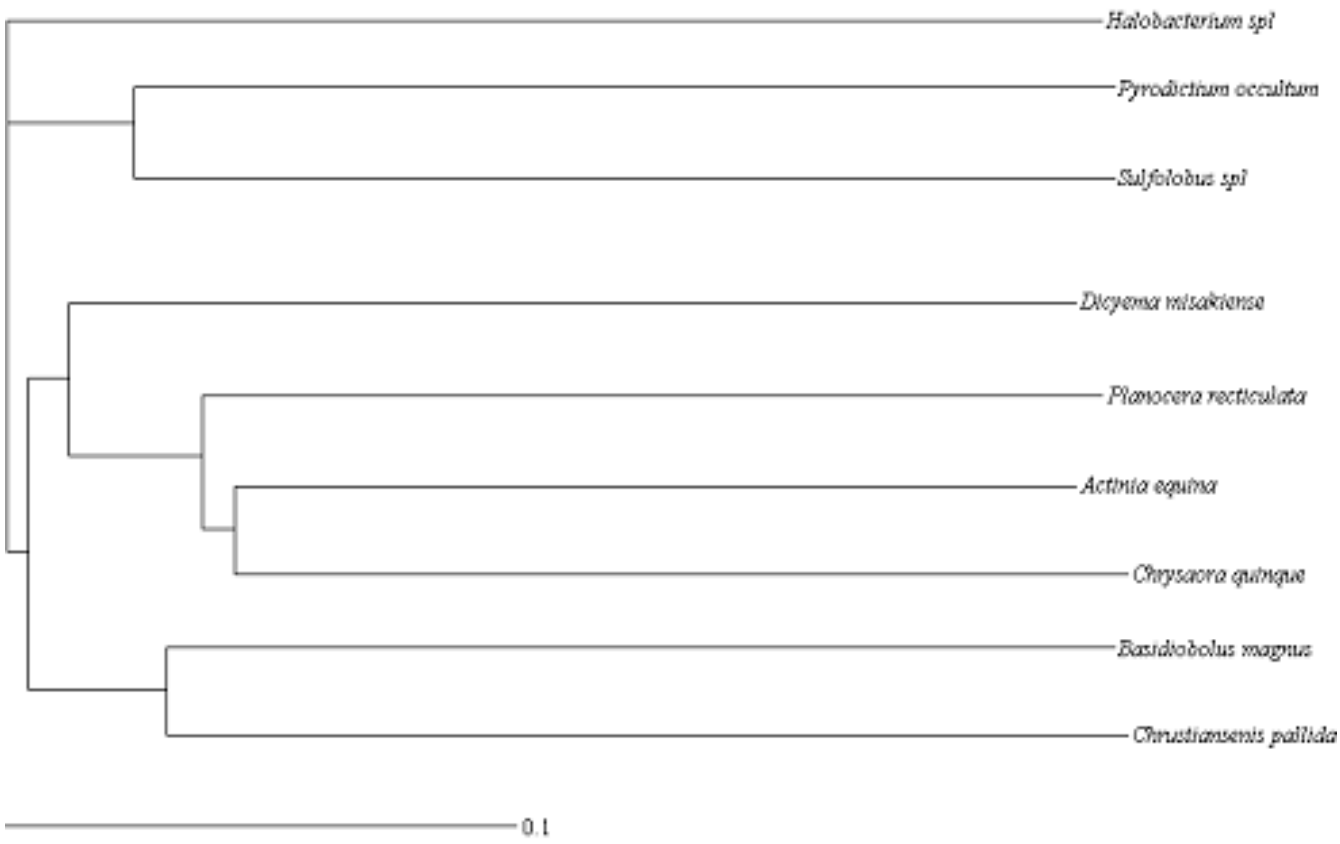

Figure 1
Neighbor-joining tree for the data in set I. It is obtained by our method and drawn by Treeview program.

effectively extract the repeated patterns encoded in linear sequences.

For comparison, we employ RNAforester program to perform the similarity analysis on the same data. This program calculates the similarity score for any pair of RNA secondary structures under the proposed scoring scheme. The similarity relationship is displayed in a cluster tree. By performing the RNAforester program on set I and set II, we obtain two cluster trees, as shown in Figure 3 and Figure 4. The numbers in the interior nodes of the cluster trees usually represent the similarity scores between the two sub-clusters that the interior nodes connect, respectively. Note that we set 0.7 as the clustering threshold when we run RNAforester program. Thus the similarity score that is not less than 0.7 will be replaced by 0 in the cluster tree. The efficiency of RNAforester program in analyzing the data from set I and set II is evaluated by Figure 3 and Figure 4.

At first, we compare Figure 1 with Figure 3. From Figure 1, we observe that: 1. *Actinia equina, Chrysaora quinque and Planocera recticulata* are grouped closely (they belong to *Animalia*); 2. *Basidiobolus magnus* and *Christiansenis pallida* are grouped closely (they belong to *fungi*); 3. *Pyrodictium occultum, Halobacterium spl* and *Sulfolobus spl* (they belong to *Archaebacteria*) are clearly separated from the rest; 4. *Dicyema misakiense* is placed closer to *Animalia* than to *fungi* (it belongs to *mesozoa*). The relationship described by our method is in accordance with the one described in [21,22]. In contrast to Figure 1, we find in Figure 3, obtained by using RNAforester program, that *Halobacterium spl* is separated from the cluster that *Pyrodictium occultum* and *Sulfolobus spl* belong to. Obviously this is not reasonable. Then we compare Figure 2 with Figure 4. From Figure 2, we observe that our result is consistent with the theory that is suggested in [23-26]: MRP evolved from a Eukaryotic Nuclear P in the nucleus of an early Eukaryote. Figure 2 indicates that mrpRNA are more similar to eukaryotic pRNA than to prokaryotic pRNA. Furthermore, *Synechocystis sp.PCC6803*, *Anacystis nidulans PCC6301*, *Pseudoanabaena sp.PCC6903*, *Anabaena sp.PCC7120* and *Porphyra purpurea chloroplast* are grouped closely, named cluster I for convenience; *Thermotoga mar-*

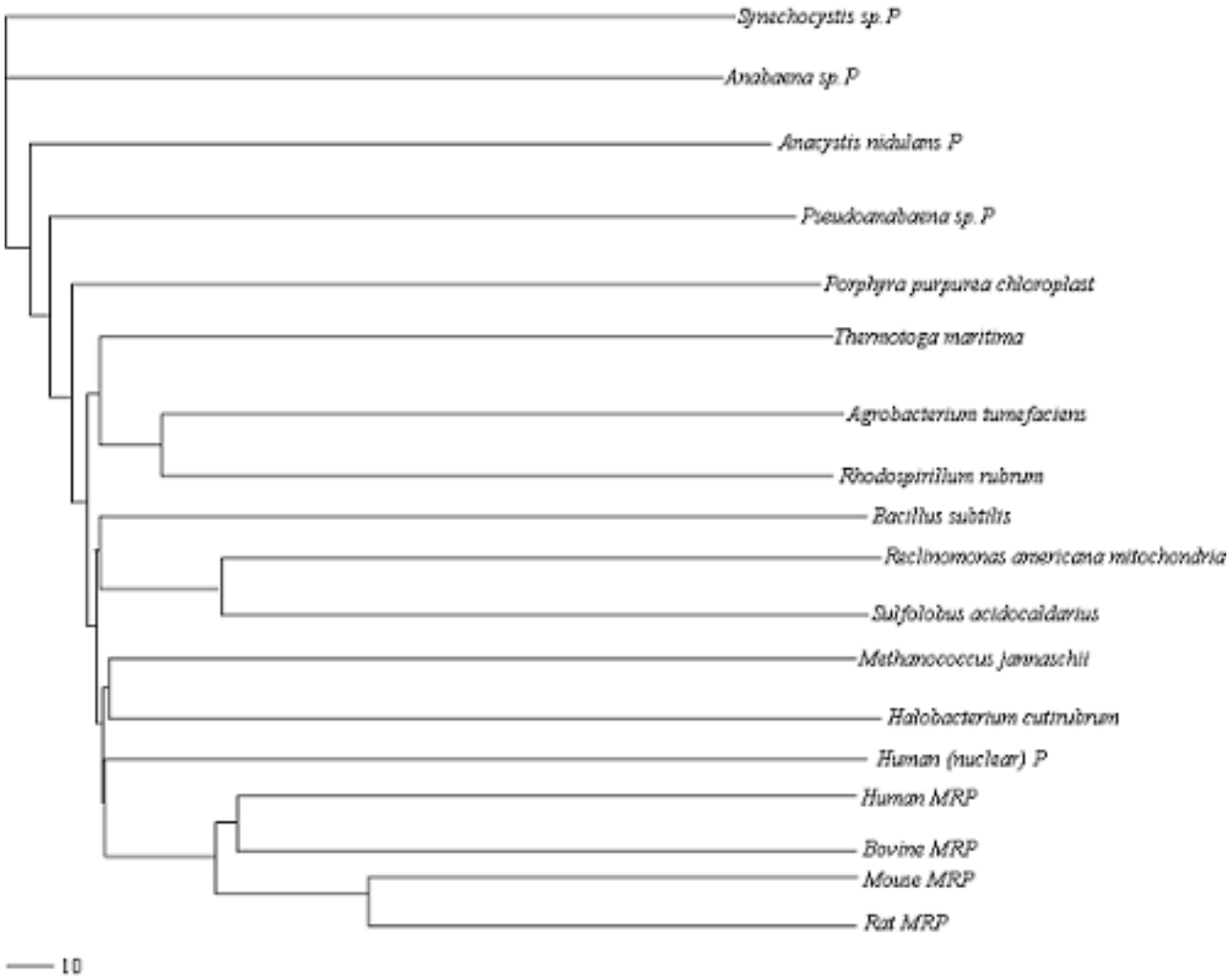

Figure 2
Neighbor-joining tree for the data in set II. It is obtained by our method and drawn by Treeview program.

itima, Agrobacterium tumefaciens and *Rhodospirillum rubrum* are grouped closely, named cluster II. Cluster I and cluster II are adjacent. In Figure 4, *Halobacterium cutirubrum* is put far away from *Methanococcus jannaschii*. Furthermore, *Anacystis nidulans P* is separated far from *Synechocystis sp.P* and *Anabaena sp.P. Bacillus subtilis* and *Reclinomonas americana mitochondria* aren't placed closely. This conformation doesn't accord with the one demonstrated by Collins et al.

In general, our method can compare secondary structures reasonably, with the results consistent with those from [23-26]. For the two data sets, our algorithm performs better than RNAforester program. Additionally, our analysis results favor the proposal that RNA secondary structures are useful materials for evolutionary analysis.

It seems that our method is heavily biased towards comparing sequences, not secondary structures. However, in fact, this is not the truth. We now apply Lempel-Ziv algorithm directly to RNA sequences to see whether the result obtained by this method is better than ours. As a result, the phylogenetic tree for the data in set II has much divergence from ours, shown in Figure 5 (drawn by Treeview).

It's obvious that there exists unreasonable topology that depicts the similarity relationship of these RNA secondary structures in Figure 5. For example, *Thermotoga maritima, Agrobacterium tumefaciens* and *Rhodospirillum rubrum* are placed close to the RNase MRP RNAs and are separated far away from the branch for *Synechocystis sp.P* and *Anabaena sp.P*, etc, which simultaneously leads to the separation of

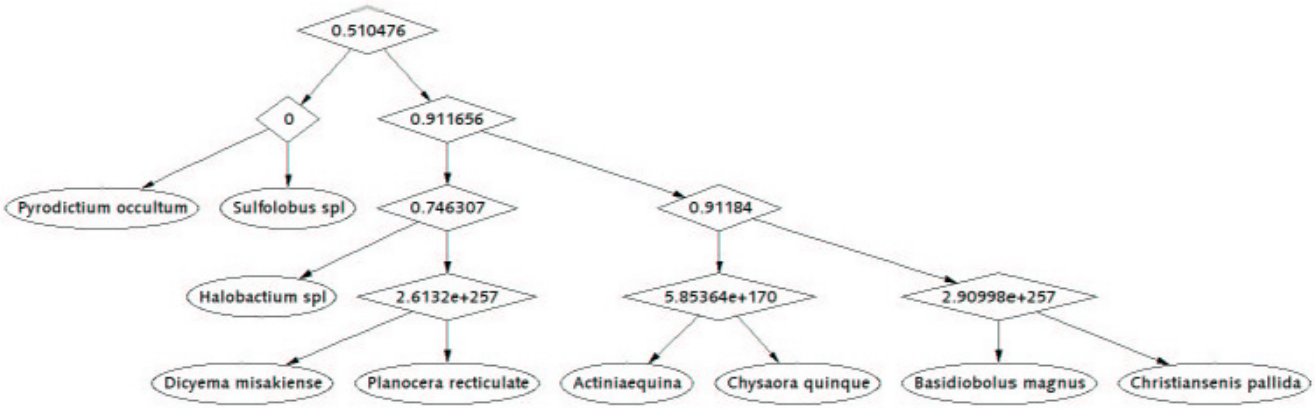

Figure 3
Cluster tree for the data in set I. It is obtained by using RNAforester program. The tree is derived based on the similarity scores between any pair of RNA forests.

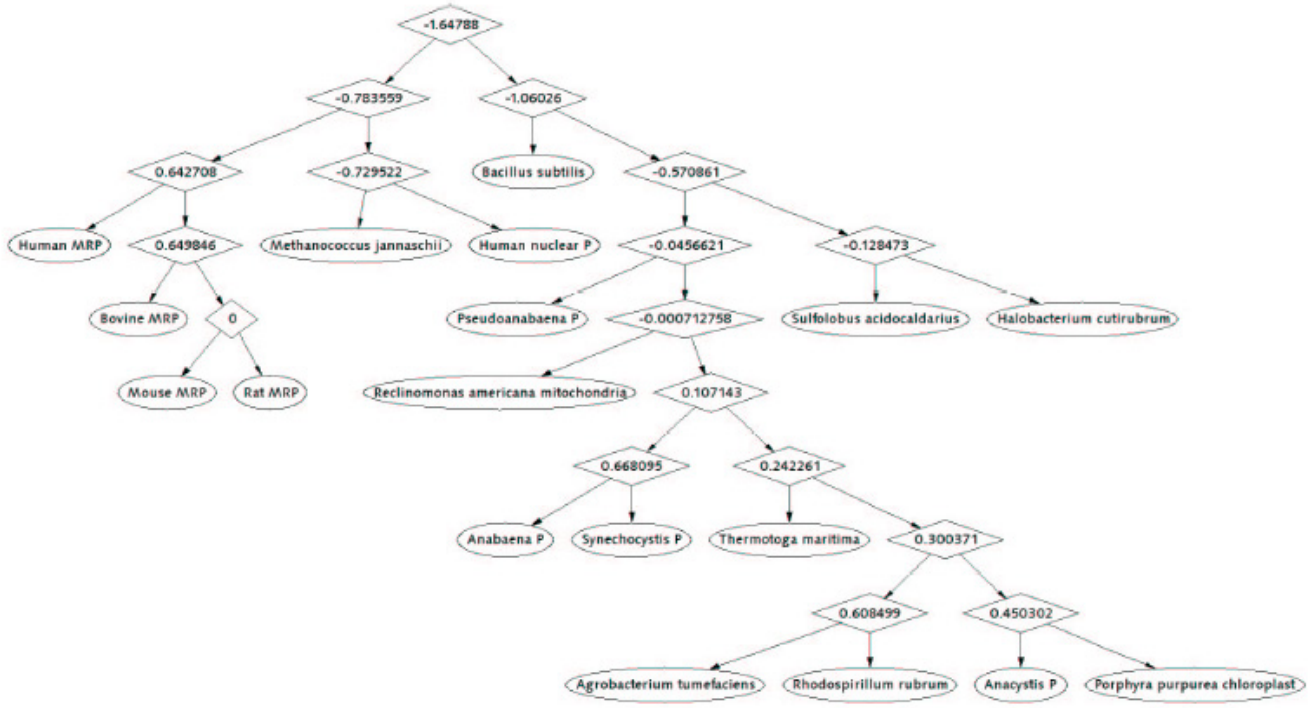

Figure 4
Cluster tree for the data in set II. It is obtained by using RNAforester program. The tree is derived based on the similarity scores between any pair of RNA forests.

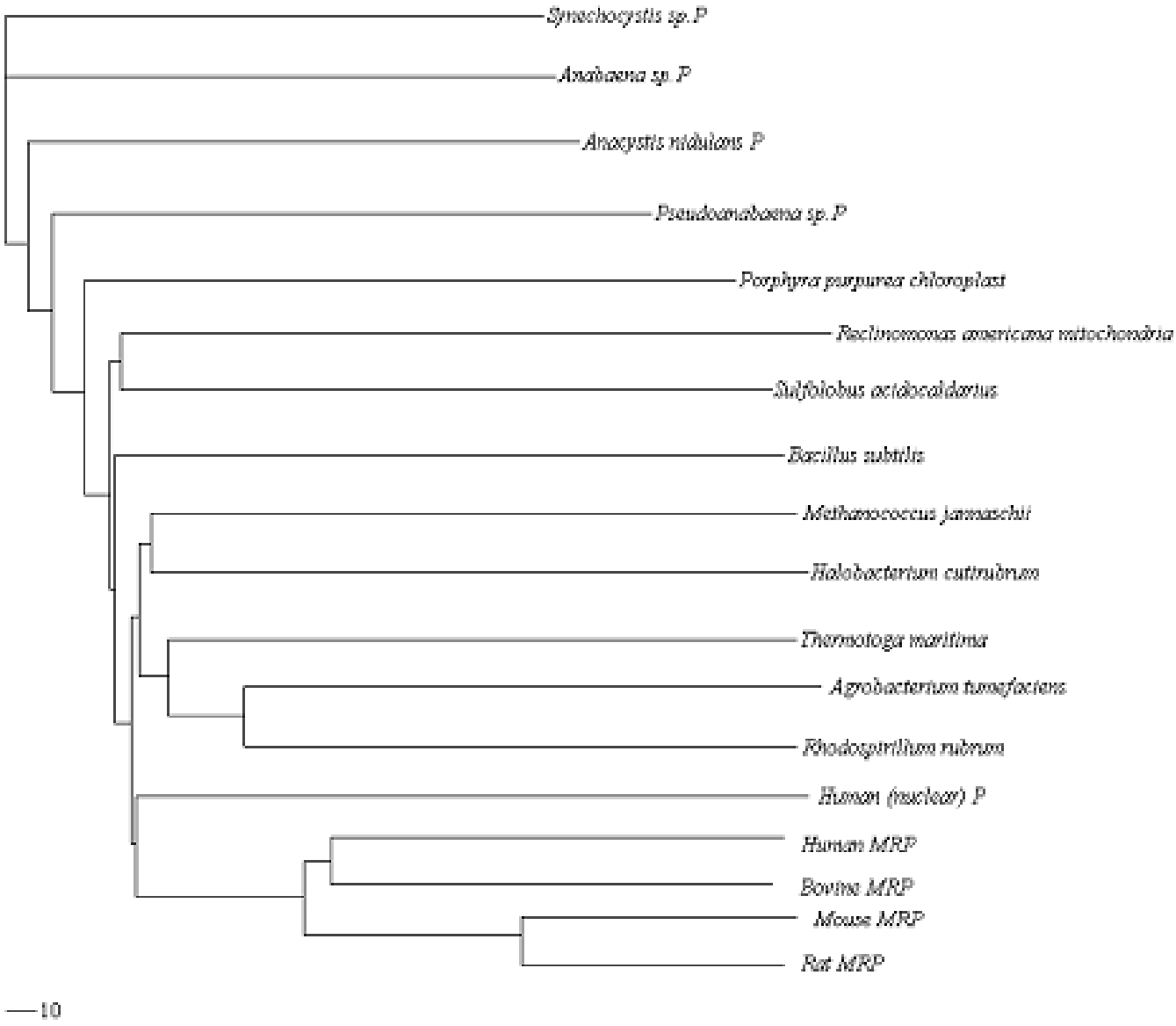

Figure 5
Neighbor-joining tree for the data in set II. It is obtained by performing LZ algorithm on RNA primary structures, i.e. the step to extract linear characteristic sequences from RNA secondary structures has been ignored.

Sulfolobus acidocaldarius from *Methanococcus jannaschii* and *Halobacterium cutirubrum*. In nature, *Thermotoga maritima*, *Agrobacterium tumefaciens* and *Rhodospirillum rubrum* belong to Eubacterial RNase P and should be grouped close to *Synechocystis sp.P* and *Anabaena sp.P*, etc. Figure 5 has favored our claim, i.e. our characteristic sequences do grasp some information on RNA secondary structures (base pairing).

The introduction of the Lempel-Ziv algorithm to the similarity analysis makes our algorithm run fast. Table 1 lists the general time and space complexity of our method and RNAforester program. In Table 1, the relationship between the size (length) of RNA secondary structure and the time complexities hasn't been indicated explicitly for the RNAforester program. We may make approximate estimation. In theory, the total number of the nodes of an RNA forest scales linearly with the size of the RNA secondary structure. For RNA secondary structures that exist in nature, the maximum length of an unpaired region and the branching degree can be considered to be bounded by some constants, which determines that the degree of an

Table 1: Time/Space complexities of our method and the RNAforester program.

Algorithm Name	Running Time	Space requirement	Reference
[a]*RNA forester*	$O(\|F_1\|\|F_2\|deg(F_1)deg(F_2))$	$O(\|F_1\|\|F_2\|max(deg(F_1), deg(F_2)))$	[27]
[b]*Our method*	$O(N^2)$	$O(N^2)$	

[a]$|F_i|$ is the number of nodes in the forest F_i and $deg(F_i)$ is the degree of F_i; [b]N is the average size.

RNA forest is expected to stay a constant. Hence the running time $O(|F_1||F_2|deg(F_1))deg(F_2))$[27] is equivalent to $O(n_1n_2)$, where n_1n_2 is the product of the sizes of the two RNA secondary structures being compared.

On the other hand, we have compared the execution time of our method with that of RNAforester by using some RNA secondary structures of various sizes. The results are listed in Table 2. It's obvious that our algorithm performs faster.

Additionally, we have performed our program on a set of *16S rRNAs*, whose secondary structures are more complex and the sizes of which are relatively larger. The result is shown in Figure 6, drawn by Treeview. Their similarity relationship has been reasonably derived by our method. *Thermoproteus tenax*, *Halobacterium*, *Methanococus vannielli*, *Thermococcus celer* and *Methanobacterium* have been clustered together, which is consistent with the fact that they are of *Archaea*. *Mus musculus*, *Saccharomyces cerevisiae*, *Homo sapiens*, *Vairimorpha* are clustered together, which is consistent with the fact that they are of *Eucaya*. The left are of *Bacteria*.

Conclusion

Here we have proposed a new method to analyze the similarity of RNA secondary structures (pseudoknots are taken into account). It is a simple method that yields results reasonably and rapidly. Our algorithm is not necessarily an improvement as compared to some existing methods, but an alternative for the similarity analysis of RNA secondary structures. The new method doesn't require sequence alignment and the construction of tree models. It is based on linear characteristic sequences that we define for RNA secondary structures and the famous Lempel-Ziv algorithm that relates to minimal length encoding. The characteristic sequences contain the information from RNA primary structures and the base pairs formed in RNA secondary structures. The Lempel-Ziv algorithm effectively extracts the information on the repeated patterns encoded in long sequences. The example applications of our method to three sets of real data and its comparison with other methods verify the validity of our method. From the comparisons, we conclude that our method performs well on distantly related RNA secondary structures. In our approach, complicated computation is avoided. The whole process is easy to operate. What's more, the size of RNA secondary structure is not problematic.

Of course, there is defect in our approach: when non-linear RNA secondary structures are transformed into linear characteristic sequences, some information may be lost. However, our test has indicated that our method can yield results reasonably, i.e. our method can extract some key information from RNA secondary structures.

Methods

Lempel-Ziv algorithm and LZ complexity

Let S, Q and R be sequences over a finite alphabet Λ, $l(S)$ be the length of S, $S(i)$ be the *ith* element of S and $S(i,j)$ be the subsequence of S that starts at position i and ends at position j. Note that $S(i,j) = \emptyset$, for $i > j$. The contatenation 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 $k = 1,2,.., l(R)$. For example: $AACUT \rightarrow AACUTACU$ 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: $CCUA \Rightarrow CCU\ AU\ AUT$ with $p = 3$.

Table 2: Execution times required by our algorithm and the RNAforester program.

Species Name	Execution Time required by RNAforester	Execution Time required by our method
[c]*Two 5S rRNAs*	12.62 s	1.52 s
[d]*Two RNase P RNAs*	35.36 s	6.96 s
[e]*Two 16S rRNAs*	1583.12 s	176.68 s

The two algorithms have been performed on the same representative RNAs for comparison. Letter s represents seconds. In [c], *Halobacterium spl* and *Christiansenis pallida* are chosen to compare. In [d], *Porphyra purpurea chloroplast* and *Bacillus subtilis* are chosen to compare. In [e], *Thermotoga* and *Saccharomyces cerevisiae* are chosen to compare.

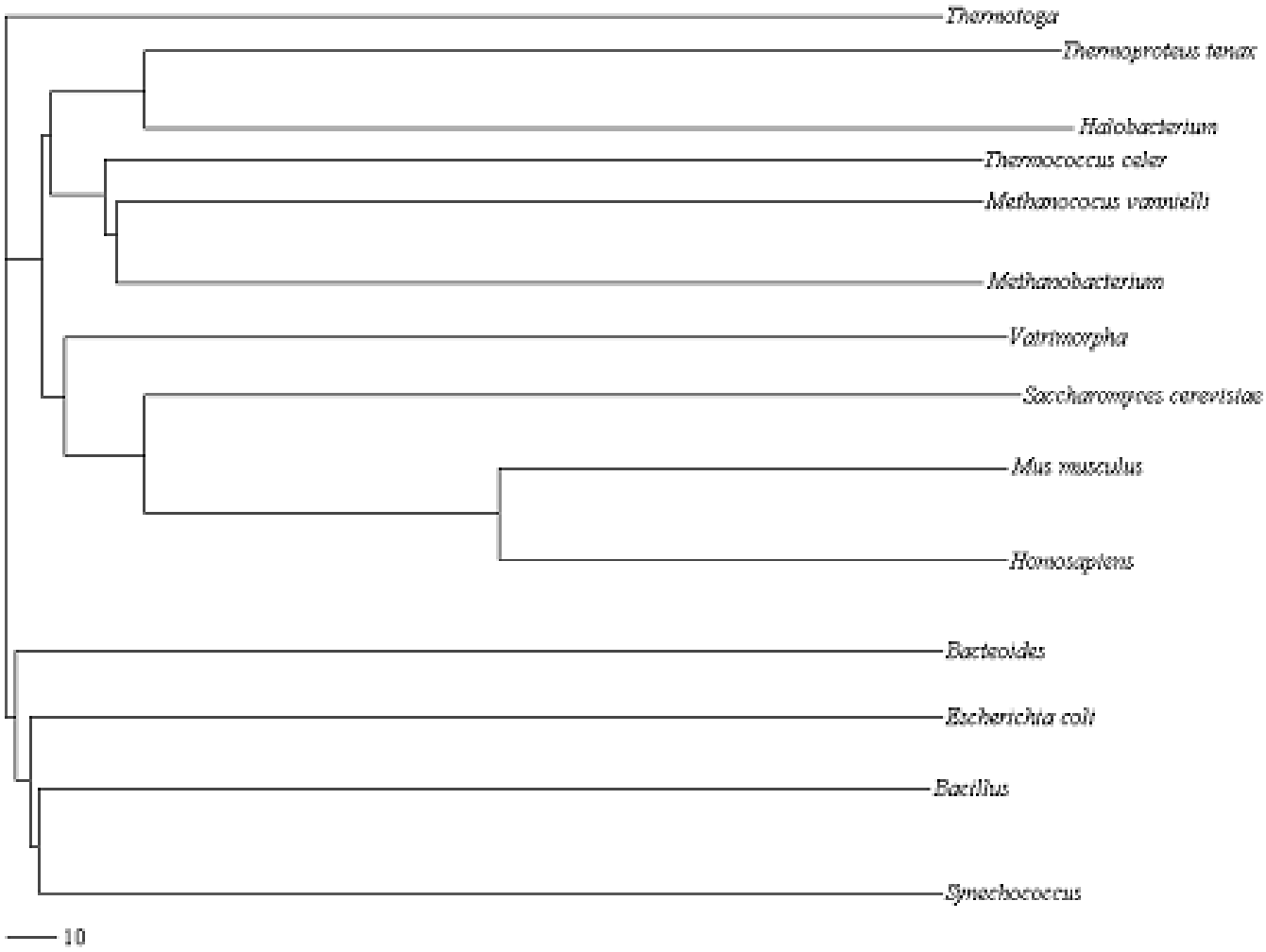

Figure 6
Neighbor-joining tree for the data in set III. It is obtained by our method and drawn by Treeview program.

The difference between producibility and reproducibility is that the former allows for an extra "different" symbol at the end of the extension process which is not permitted in the latter. Therefore an extension which is reproducible is always producible but the reverse may not always be true.

Any non-null sequence S can be built from a production process by iterative self-deleting-building process where at the *ith* step $S(1, h_{i-1}) \Rightarrow S(1, h_i)$, $\varnothing = S(1, 0) \Rightarrow S(1, 1)$. An m-step production process of S leads to a parsing of S into $H(S) = S(1, h_1) \bullet S(h_1 + 1, h_2) \bullet \ldots\ldots \bullet S(h_{m-1} + 1, h_m)$, which is called the history of S, and $H_i(S) = S(h_{i-1} + 1, h_i)$ is called the *ith* 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 (with a possible exception of the last one) is exhaustive. What's more important, the exhaustive history of any non-null sequence is unique. For example, for the sequence S = *UUCGAGGUCGGA*, its exhaustive history is $EH(S)$= *U•UC•G•A•GG•UCGG•A*.

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.

Linear characteristic sequences of RNA secondary structures

Usually, A, C, G, U are used to denote the four bases(nucleotides) in RNA sequences (primary structures). An RNA sequence can thus be represented by R =

$r_1r_2.....r_n$, where r_i is called the i^{th}(ribo)nucleotide. Each r_i belongs to the alphabet {A, C, G, U}. The secondary structure of an RNA molecule is the collection of base pairs that occur in its 3D structure. For each secondary structure, there are two terminals: 5'-terminal and 3'-terminal. Figure 7 shows a simulated RNA secondary structure. Its RNA sequence (from 5'-terminal to 3'-terminal)is:

GGGAAACUGGAAGGCGGGGCGAACGUCGGCCCCAGU GAAGUCAAAUGGAGCGUACACGGACCAUUAUG-GGCUAA.

In this section, we will define linear characteristic sequences for RNA secondary structures. In other words, we will transform non-linear RNA secondary structures into linear sequences. In our research, a group of consecutive base pairs (including one base pair) is called a helix and i is called the position of nucleotide r_i. The open regions surrounded by single stranded bases are called loops. Each helix is numbered from the 5'-terminal to 3'-terminal: The first helix is called Helix1, and the second helix is called Helix2,.....etc. The rule for transformation is as such: For an RNA secondary structure S with N helices, write its RNA sequence with the letters in Helix1 upper case and the rest small case; then write in succession its RNA sequence with the letters in Helix2 upper case and the rest small case;.....then write in succession its RNA sequence with the letters in HelixN upper case and the rest small case. Now a linear sequence has been obtained by following the above-mentioned rule. We call it linear characteristic sequence of *S*, abbreviated to L(S). Take the simulated secondary structure for example, it has five Helices.

According to our rule, the linear characteristic sequence of the simulated secondary structure is as follows:

gggaaacuggaaggcGGGGCgaacgucgGCCCCagugaagucaaaugg agcguacacggaccauuauggguaagggaaacuggaaggcggggcgaACGU CGgccccagugaagucaaaUGGAGCguacacggaccauuauggguaag-ggaaa

cuggaaggcggggcgaacgucggccccagugaAGucaaauggagcguacacg-gaccauuauggguCUaagggaaacuggaa

ggcggggcgaacgucggccccagugaaguCaaauggagcguacacggaccau-uaugGgcuaagggaaacuggaaggcgggg

cgaacgucggccccagugaagucaAAUGgagcguacacggacCAUUaug-ggcuaa

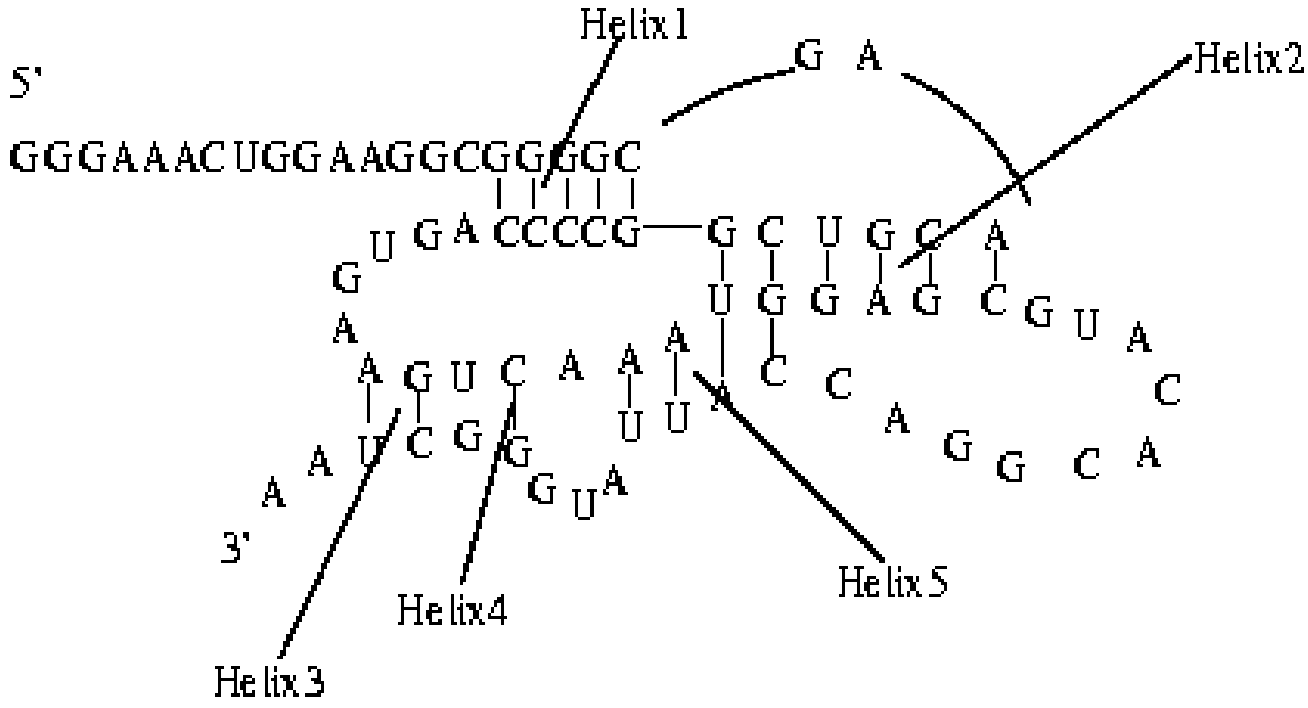

Figure 7
A simulated secondary structure. It contains a pseudoknot. And five Helices are formed according to our definition.

which is obtained by adjoining the following five sequences in succession.

Helix1:
gggaaacuggaaggcGGGGCgaacgucgGCCCCagugaagucaaaug-gagcguacacggaccauuauugggcuaa

Helix2:
gggaaacuggaaggcggggcgaACGUCGgccccagugaagucaaaUG-GAGCguacacggaccauuauugggcuaa

Helix3:
gggaaacuggaaggcggggcgaacgucggccccagugaAGucaaaug-gagcguacacggaccauuauugggCUaa

Helix4:
gggaaacuggaaggcggggcgaacgucggccccagugaaguCaaaug-gagcguacacggaccauuauugGgcuaa

Helix5:
gggaaacuggaaggcggggcgaacgucggccccagugaagucaAAUG-gagcguacacggacCAUUauugggcuaa

Distance computation and pair-wise distance matrix

Lempel et al have proposed that, for any given sequences Q and S, $c(QS) \le c(Q) + c(S)$ always remains valid. This formula shows that the steps required to extend Q to QS are always less than the steps required to build S from $\varnothing$. Recently, Otu et al [28] 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.

For example, let Q, S, R represent three short RNA sequences defined over the alphabet $\{A, C, G, U\}$, where $S = UUACGUAAUGU$, $Q = AGUCCCUAGGA$, $R = UAC-CGAUAAG$. By the rule mentioned above, the corresponding exhaustive histories of S, Q, R, SR, QR, SQ are: $EH(S) = U{\bullet}UA{\bullet}C{\bullet}G{\bullet}UAA{\bullet}UG{\bullet}U$, $EH(Q) = A{\bullet}G{\bullet}U{\bullet}C{\bullet}CCU{\bullet}AGG{\bullet}A$, $EH(R) = U{\bullet}A{\bullet}C{\bullet}CG{\bullet}AU{\bullet}AA{\bullet}G$, $EH(SR) = U{\bullet}UA{\bullet}C{\bullet}G{\bullet}UAA{\bullet}UG{\bullet}UUACC{\bullet}GA{\bullet}UAAG$, $EH(QR) = A{\bullet}G{\bullet}U{\bullet}C{\bullet}CCU{\bullet}AGG{\bullet}AU{\bullet}AC{\bullet}CG{\bullet}AUAA{\bullet}G$, $EH(SQ) = U{\bullet}UA{\bullet}C{\bullet}G{\bullet}UAA{\bullet}UG{\bullet}UAG{\bullet}UC{\bullet}CC{\bullet}UAGG{\bullet}A$. We can find that we need 2 steps to build R from S, 4 steps to build R from Q, 4 steps to build Q from S. So we say R is more similar to S than to Q. The reason is that S and R share the common patterns UAC and UAA.

Based on this hypothesis, Otu et al 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.

Therefore we adopt the following formula to evaluate the distance between secondary structures S and Q, which is slightly different from [28]:

$$Rd(S,Q) = \begin{cases} \dfrac{c(L(S)L(Q)) - c(L(S)) + c(L(Q)L(S)) - c(L(Q))}{c(L(S)L(Q)) + c(L(Q)L(S))} & S \neq Q \\ 0, & S = Q \end{cases}$$

The denominator in [28] is equivalent to $[c(L(S)L(Q))+c(L(S)L(Q))]/2$, which leads to the fact that the distance calculated by the formula proposed in [28] will always be twice as much as the distance calculated by our formula. As you know, a constant will not affect the similarity analysis at all. We choose to use the formula mentioned above mainly because its expression is simpler. The formula in [28] has been proven to be a distance metric by Out et al. Thus $Rd(S, Q)$ also satisfies the conditions required by a distance metric.

It's obvious that the more similar S is to Q, the smaller $c(L(S)L(Q))$-$c(L(S))$ and $c(L(Q)L(S))$-$c(L(Q))$ are, and then the smaller $Rd(S,Q)$ is.

Generally, given n RNA secondary structures $S_1, S_2, \ldots\ldots, S_n$, we can obtain their linear characteristic sequences by the above-mentioned rule, which are $L(S_1), L(S_2), \ldots\ldots, L(S_n)$. They are linear sequences defined over alphabet {A, C, G, U, a, c, g, u,} and carry the information on RNA secondary structures. Then, by using Lempel-Ziv algorithm, the distance between any pair of structures, $Rd(S_i, S_j)$, may be rapidly computed. By arranging them into a matrix, a pair-wise distance matrix is obtained, denoted by RD. $RD = (Rd(S_i, S_j))$ contains the information on the similarity/dissimilarity between any pair of RNA secondary structures.

Based on what Otu et al have proven, we can easily conclude that our distance metric is also valid for inferring phylogeny of species because it satisfies all the conditions for phylogeny analysis. Hence our pair-wise distance matrix may be input the programmes for phylogeny inference to study the phylogeny for RNA molecules based on their secondary structures.

Authors' contributions

NL conceived of the study and drafted the manuscript. NL participated in the design of the algorithm. NL and TMW performed the comparison of RNA secondary structures and tested the algorithm.

Acknowledgements

The authors thank all the anonymous reviewers for their support and suggestions. In particular, the authors thank Maciej Szymanski for providing all the secondary structures of 5S RNAs. The authors also thank Jiang Tian for the technical help. The work is supported by the National Natural Science Foundation of China(10571019).

References

1. Shapiro B, Zhang K: **Comparing multiple RNA secondary structures using tree comparisons.** *Comput Appl Biosci* 1990, **6:**309-318.
2. Hofacker IL, Bernhart SHF, Stadler PF: **Alignment of RNA base pairing probability matrices.** *Bioinformatics* 2004, **20:**2222-2227.
3. McCaskill JS: **The equilibrium partition function and base pair binding probabilities for RNA secondary structure.** *Biopolymers* 1990, **29:**1105-1119.
4. Le SY, Nussinov R, Maizel JV: **Tree graphs of RNA secondary structures and their comparisons.** *Comput Biomed Res* 1989, **22:**461-473.
5. Shapiro B: **An algorithm for comparing multiple RNA secondary structures.** *Comput Appl Biosci* 1988, **4(3):**387-393.
6. Sankoff D: **Simultaneous solution of the RNA folding, alignment, and protosequence problems.** *SIAM J Appl Math* 1985, **45:**810-825.
7. Le SY, Owens J, Nussinov R, Chen JH, Shapiro B, Maizel JV: **RNA secondary structures: comparisons and determination of frequently recurring substructures by consensus.** *Comput Appl Biosci* 1989, **5:**205-210.
8. Jiang T, Lin GH, Ma B, Zhang K: **A general edit distance between RNA structures.** *Journal of Computational Biology* 2002, **9:**371-388.
9. Liao B, Wang TM: **A 3D graphical representation of RNA secondary structure.** *J Biomol Struc Dynamics* 2004, **21:**827-832.
10. Randic M, Basak SC: **Characterization of DNA primary sequences based on the average distances between bases.** *J Chem Inf Comput Sci* 2001, **41:**561-568.
11. Zupan J, Randic M: **Algorithm for coding DNA sequences into "spectrum-like" and "zigzag" representations.** *J Chem Inf Comput Sci* 2005, **45:**309-313.
12. Guo XF, Nandy A: **Numerical characterization of DNA sequences in a 2-D graphical representation scheme of low degeneracy.** *Chemical Physics Letters* 2003, **369:**361-366.
13. Randic M, Vracko M, Lers N, Plavsic D: **Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation.** *Chemical Physics Letters* 2003, **371:**202-207.
14. Hochsmann M, Toller T, Giegerich R, Kurtz S: **Local similarity in RNA secondary structures.** *Proceedings of the IEEE Bioinformatics Conference 2003 (CSB 2003)* 2003:159-168.
15. Lempel A, Ziv J: **On the complexity of finite sequences.** *IEEE T Inform Theory* 1976, **22:**75-81.
16. Brown J: **The ribonuclease P database.** *Nucleic Acids Res* 1998, **26:**351-352.
17. Schmitt M, Bennett J, Dairaghi D: **Secondary structure of RNase MRP RNA as predicted by phylogenetic comparison.** *FASEB J* 1993, **7:**208-213.
18. Szymanski M, Miroslawa Z, Barciszewska VA, Barciszewski EJ: **5S ribosomal RNA database.** *Nucleic Acids Res* 2002, **30:**176-178.
19. Felsenstein J: **PHYLIP-Phylogeny inference package (version 3.2).** *Cladistics* 1989, **5:**164-166.
20. Page RDM: **TREEVIEW: An application to display phylogenetic trees on personal computers.** *Computer Applications in the Biosciences* 1996, **12:**357-358.
21. Hiro H, Osawa S: **Evolutionary change in 5S rRNA secondary structure and a phylogenetic tree of 54 5S rRNA species.** *Proc Natl Acad Sci USA* 1979, **76:**381-385.
22. Hiro H, Osawa S: **Evolutionary change in 5S rRNA secondary structure and a phylogenetic tree of 352 5S rRNA species.** *Proc Natl Acad Sci USA* 1986, **19:**163-172.
23. Morrissey JP, Tollervey D: **Birth of the snoRNPs: the evolution of RNase MRP and the eukaryotic pre-rRNA-processing system.** *TIBS* 1995, **20:**78-82.
24. Reddy R, Shimba S: **Structural and functional similarities between MRP and RNase P.** *Mol Biol Rep* 1996, **22:**81-85.
25. Chamberlain J, Pagan-Ramos E, Kindelberger D, Engelke D: **An RNase P RNA subunit mutation affects ribosomal RNA processing.** *Nucleic Acids Res* 1996, **24:**3158-3166.
26. Collins LJ, Moulton V, Penny D: **Use of RNA secondary structure for studying the evolution of RNase P and RNase MRP.** *J Mol Evol* 2000, **51:**194-204.
27. Höchsmann M, Voss B, Giegerich R: **Pure Multiple RNA Secondary Structure Alignments: A Progressive Profile Approach.** *IEEE ICBB* 2004, **1:**53-62.
28. Otu HH, Sayood K: **A new sequence distance measure for phylogenetic tree construction.** *Bioinformatics* 2003, **19:**2122-2130.