Problem Chosen	2023 MCM/ICM Summary Sheet	Team Control Number
B		

The analysis of the Genetic Process of Sequences

Summary

To address the problem of RNA sequence correlation study, we encoded RNA sequences and used mutual mode entropy and DTW to measure the correlation between two RNA sequences, and on this basis, we explored the gene sequence family analysis.

To address problem 1, in this paper, we encode RNA sequences, and the RNA sequences after completing the encoding can be solved for the shortest distance between two RNA sequences and mutual mode entropy. To solve for the shortest distance, we first calculate the distance matrix between each point of the two sequences, and then find a path from the upper left corner of the matrix to the lower right corner, so that the elements on the path and the smallest first step to obtain the shortest distance path. To solve the mutual mode entropy, in the first step, m data points are taken consecutively for two sets of time series containing N+1 data to form the corresponding m-dimensional vector, in the second step, the baseline value of the m-dimensional vector is calculated and followed by constructing the data vector required for the experiment, in the third step, the probability that these two vectors are similar is calculated and the MME value is solved as the final result. The degree of similarity between different sequences is quantified in two ways.

For problem 2, mRNA sequences from the NCBI database were used to validate the two models constructed in problem 1. 9 sequences were collected, and the shortest distance of 9 sequences and the mutual pattern entropy of 7 sequences were calculated respectively.At the same time, the computational complexity of both models is low, and they are linear operations, which are faster in the process of RNA sequence analysis with tens of thousands of sequences.

In this paper, we constructed a gene tree to visualize the parent-child relationship between sequences using Arabidopsis thaliana gene sequence data (from NCBI database and Pfam database), and used hmmsearch and TBtools to construct a gene tree for protein sequences related to NB-ARC gene family of Arabidopsis thaliana, and obtained good results. We designed a "greedy" algorithm (GIGA) to construct the tree using the sequence distance matrix as a guide. Starting from the two sequences closest to the distance matrix, the subtrees of the sequences are connected together in an iterative manner. The connected subtrees are also "rearranged" at each iteration based on additional (genomic) knowledge. In this way, homologous genes can be quickly determined and gene systems can be inferred.

Finally, the strengths and weaknesses of the model are discussed, and a sensitivity analysis of the mutual mode entropy is performed because there are no parameters to adjust for the shortest distance calculation.

Keywords, RNA sequence comparison; Genetic family analysis; Mutual model entropy

Contents

1 Introduction

1.1 Background and restatement

Homology of DNA, RNA or protein sequences indicates that they have the same ancestor. Similarity of nucleotide or amino acid sequences of DNA, RNA or protein can determine their homology in biological information. Similarity can prove that two sequences evolved from the same ancestor.Nucleotide base mutations can occur by chance during the inheritance of RNA sequences. For convenience, we assume that mutations arise due to changes (conversions or reversals), insertions and deletions of individual bases. Thus, the number of mutation sites can be used to measure the distance between two sequences. Base sequences, multiple closely linked, are able to form a family. This is considered to be homologous.

Based on the above background we need to develop mathematical models to solve the following problems.

- Problem1 Design an algorithm in order to rapidly measure the distance between two long enough (more than 10^3 bases) base sequences.
- Problem2 Evaluate the complexity and precision of the algorithm convincingly, and illustrate with appropriate examples.
- Problem3 Assume that multiple base sequences in a family evolved from a common ancestral sequence, design a competent algorithm to identify the ancestral sequence and draw a genealogical tree.

1.2 Our work

Before starting the modeling, we simulated and observed nucleic acid molecules using chembio3D, a scientific computing and chemical science research software that allows users to generate 3D models of small molecules and biomolecules and perform various calculations and manipulations on the models in order to study the properties and interactions of substances.

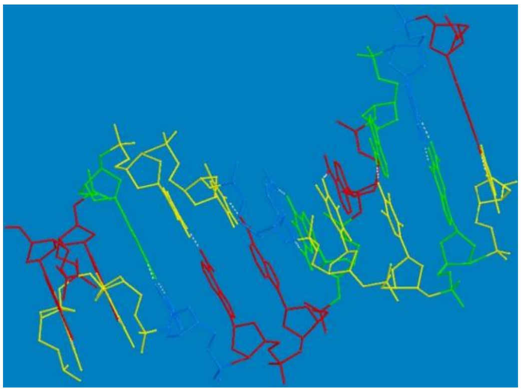


Figure 1:Nucleic acid molecular map

Judging the similarity of base sequences is of great significance for medical treatment, evolution and tracing. In order to obtain the similarity of gene sequences in function and structure, it is often necessary to compare several DNA sequences with different sequences to judge whether the DNA is related. For this purpose, we build models to calculate and measure genetic similarity, infer ancestral base sequences, and construct family trees.

First, based on the data we collected, the entire sequence was modified with different encoding forms to make them more suitable for our model.

Secondly, the Dynamic Time Warping (DTW) algorithm is constructed to calculate the maximum similarity of two time series by combining the sequence planning and distance measure with the nonlinear programming principle.

Thirdly, considering the time complexity, in order to effectively reduce the dependence of the judgment criterion on the tolerance threshold in the similar process of the judgment vector, the number of similar vectors will not be affected by large signal fluctuation or sudden change of signal length. We establish MME algorithm to solve the problem of statistical stability of approximate entropy effectively.

Moreover, in order to construct the genealogy tree, the gene tree was disassembled, GIGA algorithm was established, combining with ‘hmmsearch’ software to analyze the gene sequence of the whole family.

Finally, we analyze the performance of our model and the sensitivity of our model.

2 Assumptions and Justifications

- Boundary conditions, w1 = (1, 1) and wk = (m, n), which means that the first and last parts of the two sequences must match, and the order of the parts must match.

- Continuity, If wk = (a, b) and wk - 1 = (a', b'), then a - a' ⩽ 1 and b - ' ⩽ 1 must be satisfied. this constraint means that the many-to-one and one-to-many cases can only be matched at one time step around the matching process, i.e., it is impossible to match across a certain point, and can only be aligned with its own neighboring points. This ensures that every coordinate in Q and C is present in the wraping path.

- Monotonicity, If wk - 1 = (a', b') and wk = (a, b), then a - a' ⩾ 0 and b - b' ⩾ 0 must be satisfied, indicating that the warping path must be monotonically increasing with time.

There are many warping paths that satisfy the above constraints, so the essence of the problem is the optimization problem one by one to find the optimal warping path. The solution is based on a dynamic programming algorithm, described in mathematical language as,

$$\gamma(i, j) = d(q_i, c_j) + \min\{\gamma in-1, j-1), \gamma(i-1, j), \gamma(i, j-1)\}$$

3 Notations

The primary notations used in this paper are listed in Table 1.

Table 1 Notations

Symbol	Description
D_n	The code of the RNA sequences
X_n	The RNA sequences(C,U,A,G)
j	The RNA sequences(C,U,A,G)
$y(j)$	The code of the RNA sequences
R(t),S(t)	Time series
r_i, s_j	The component section of time series,i or j is its subscript
A_{mxn}	Distance matrix,m is the length of R(t),n is the length of S(t)
a_y	The element of distance matrix,y is its subscript
W	The minimum regularization cost from this matrix
w_k	The element of the minimum regularization cost from this matrix,$w_k=(a_{i,j})_k$
k	The length of the minimum regularization cost from this matrix
$u(i)$, v(j)	Two sets of time series containing data
N	The length of two sets of time series containing data
x(i), y(j)	Corresponding m-dimensional vectors
$B_u(i), B_v(j)$	The calculation of the judgment basis-basis value for each vector
$\Psi_x(i), \Psi_v(j)$	The reconstructed data vector required for the experiment

4 Model preparation

RNA molecules, like DNA, are also polymorphs of nucleotides and are also long chain macromolecules, they consist of four bases, adenine (A), guanine (G), Cytosine (c) and uracil (U). Compared to DNA, RNA has several different natures, (1)RNA is usually single-stranded.(2)It has ribose instead of deoxyribose. It contains four bases among which thymine (T) is replaced by uracil. Although uracil has the same pairing natures as thymine, it also forms hydrogen bonds with adenine (A). The reasons for the above natures (2) and natures (3) are not known with certainty. However, the important characteristic of RNA is that it is single-stranded and structurally similar to DNA, its bases can be in any order. It can also be repeated many times.

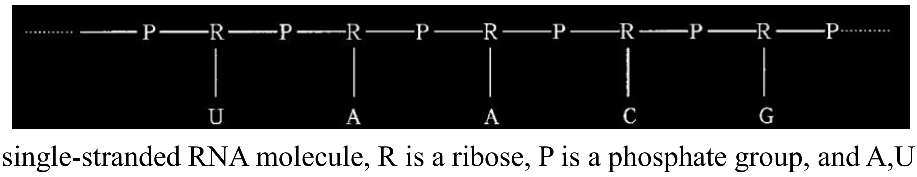


Figure 2:A single-stranded RNA molecule, R is a ribose, P is a phosphate group, and A,U,C,G are the corresponding base groups of the nucleotide molecule

Current RNA sequences still use the four letters U,C,A,G to code the four bases of uracil, cytosine, adenine and guanine, and thus are not yet the most basic coding. If we consider binary coding, the four numbers 0 (00), 1 (01), 2 (10), and 3 (11) are used to code the four bases of RNA molecules, and this coding method has the following outstanding advantages[1],

- Digital encoding is simpler and more abbreviated in representation than character encoding.
- Convenient for computer computation, the digital encoding of RNA sequences can easily perform mathematical operations, such as Fourier transform, Walsh transform, Markov chain transfer probability analysis, and other mathematical operations on RNA sequences.
- The character encoding of each base occupies 8 bits, while the digital encoding of the base occupies only 2 bits, so the encoding can compress the redundancy of information and storage space, and improve the encoding efficiency by 4 times.
- Digital codes can represent various properties of bases, such as base complementarity, strength of hydrogen bonding, etc., and there are rules to follow, while character codes of bases do not show these rules.
- The numerical codes have strict size relationship, i.e., they are fully sequential. So the digits of different RNA sequences can also be sorted according to the order of numerical size.
- The RNA sequence consisting of K bases has a total of 2K bits of digits, which is a point in the N=2K dimensional space. On the one hand, the intersection operation and the concatenation operation of the high-dimensional space can be applied, and the intersection space and concatenation space of multiple RNA digital sequences can be obtained, which is important for analyzing the interrelationship of different RNA sequences.

The RNA sequence consists of an arrangement of four bases, C,U,A,G. Therefore, four bases are digitally encoded. Therefore, the four bases are coded digitally and can be coded in pairs of 0 (00), 1 (01), 2 (10), and 3 (11). From the combinatorial mathematics theory, we know that there are 4!=24 combinations of this coding format. However, since 0 and 1 are complementary in binary numbers, 0 (00) is complementary to 3 (11) and 1 (01) is complementary to 2 (10) in 4 numbers, and G is complementary to C and A is complementary to U in 4 bases, the 4 bases should satisfy the complementarity rule, which can eliminate 16 coding methods.

Geometric interpretation of base digit encoding. Two-dimensional binary numbers encode four fixed points 00, 01, 10, 11 arranged by Hamming distance to form a two-dimensional plane in space.

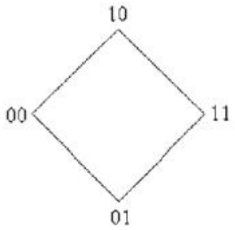


Figure 3:Coding points arranged by Hamming distance

A total of 8 such planes can be formed, and these 8 planes are topologically equivalent, i.e., one coding plane can be transformed into any other coding plane by symmetric transformations such as rotation, central inversion and specular reflection, etc. Using the 0123/CUAG coding method, in the binary digital coding of bases, the first position is called the structural coding bit. When the first position is 1, it encodes a purine base; when the first position is 0, it encodes a pyrimidine base; and the last number is the coding position of the functional group, such as 00 for cytosine and 01 for uracil. The digital coding of two complementary base pairs is also complementary, and there is a distinction between strong hydrogen bonding and weak hydrogen bonding.

The DTW (Dynamic Time Warping) algorithm can be used to calculate the maximum similarity of two time series, i.e., the minimum distance, and is a nonlinear programming technique that combines time series planning and distance measurement.The DTW algorithm was initially used to identify the similarity of speech, and considering its excellent properties in time series correlation analysis, this paper uses it for RNA The DTW algorithm was originally used to identify speech similarity, and considering its excellent properties in time series correlation analysis, it is used in this paper for the analysis of correlation between RNA sequences.

The main steps of the DTW algorithm can be summarized as follows,

- Calculate the distance matrix between each point of two sequences.
- Find a path from the upper left corner of the matrix to the lower right corner such that the elements on the path and the minimum

5 Solution I

5.1 Model I:Dynamic Time Warping (DTW)

There are two ways of transfer here, both of which were tried and found to be little different, and the effect of the transfer on the correlation between the two RNA sequences made no difference from a statistical point of view[2].

$$\mathrm{D}_n = \begin{cases} 0 & X_n = A \\ 1 & X_n = G \\ 2 & X_n = C \\ 3 & X_n = U \end{cases}$$

$$y(j) = \begin{cases} \frac{1}{2} + \frac{\sqrt{3}}{2} i, j = A \\ \frac{\sqrt{3}}{2} + \frac{1}{2} i, j = G \\ \frac{1}{2} - \frac{\sqrt{3}}{2} i, j = U \\ \frac{\sqrt{3}}{2} - \frac{1}{2} i, j = C \end{cases}$$

Classical time series similarity measures are generally classified into two categories, lock-step measures and elastic measures. Lock-step measures are "one-to-one" comparisons of time series; elastic measures allow "one-to-many" comparisons of time series. Probably the simplest way to calculate similarity is to calculate the Euclidean distance of two time series. Suppose there are two time series, Q and C. If the Euclidean distance is used directly to calculate the similarity, there are problems such as unaligned time steps and different lengths of the series.

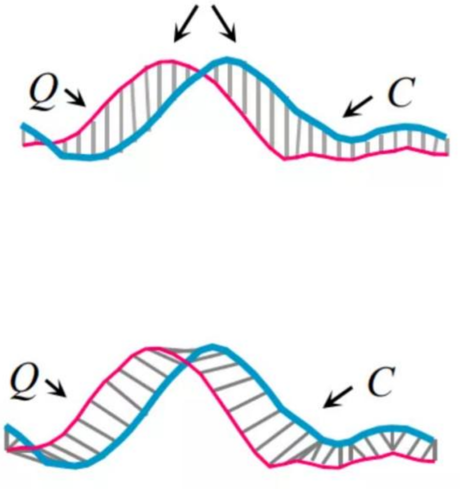
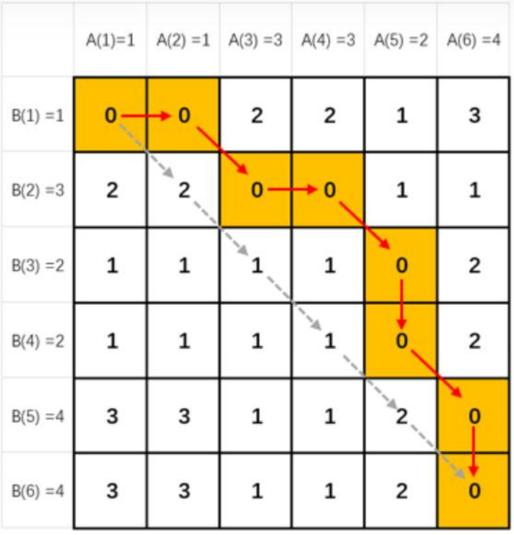


Figure 4:The top is the Euclidean distance calculation, and the bottom is the DTW algorithm calculation

As shown above, the Euclidean distance is not effective in calculating the distance between two time series when the sequences are not of the same length or the time steps are not aligned, especially at the peak. In contrast, DTW can elongate or shorten (compress and expand) the unknown quantity until it is the same length as the reference template, in which the unknown sequence is twisted or bent so that its characteristic quantity corresponds to the standard pattern[3].



	A(1)=1	A(2) =1	A(3) =3	A(4) =3	A(5) =2	A(6) =4
B(1) =1	0	0	2	2	1	3
B(2) =3	2	2	0	0	1	1
B(3) =2	1	1	1	1	0	2
B(4) =2	1	1	1	1	0	2
B(5) =4	3	3	1	1	2	0
B(6) =4	3	3	1	1	2	0

Figure 5:Illustration of DTW algorithm

The DTW (Dynamic Time Warping) algorithm can be used to calculate the maximum similarity of two time series, i.e., the minimum distance, and is a nonlinear programming technique that combines time series planning and distance measurement.The DTW algorithm was initially used to identify the similarity of speech, and considering its excellent properties in time series correlation analysis, this paper uses it for RNA The DTW algorithm was originally used to identify speech similarity, and considering its excellent properties in time series correlation analysis, it is used in this paper for the analysis of correlation between RNA sequences.

The main steps of the DTW algorithm can be summarized as follows.

- Calculate the distance matrix between each point of two sequences.
- Find a path from the upper left corner of the matrix to the lower right corner such that the elements on the path and the minimum

The algorithm principle of DTW,
Construct two time series R(t)={$\mathrm{r}_1, \mathrm{r}_2 \cdots \mathrm{r}_\mathrm{m}$},S(t)={$s_1, s_2, \cdots s_n$}, whose lengths correspond to m,n. Then construct $\mathrm{m} \times \mathrm{n}$ matrix A_{mxn}, whose elements are $a_y = d(r_i, s_j) = \sqrt{(r_i - s_j)^2}$, and finally find the path with the minimum regularization cost from this matrix, denoted as $\mathrm{W} = \mathrm{w}_1, \mathrm{w}_2, \cdots \mathrm{w}_\mathrm{k}$, w_k=$(a_{i,j})_k$. This path must satisfy the following conditions,

- $\max\{\mathrm{m}, \mathrm{n}\} \le \mathrm{k} \le \mathrm{m} + \mathrm{n} - 1$;
- $w_1 = a_{11}, w_k = a_{mn}$;
- For $w_k = a_{ij}, w_{k-1} = a_{i`j`}$, which must satisfy $0 \le \mathrm{i} - \mathrm{i}` \le 1, 0 \le \mathrm{j} - \mathrm{j}` \le 1$, the dynamic time regularization distance is,

$$\mathrm{DTW(R,S)=min(\frac{\sqrt{\sum_{i=1}^{k}w_i}}{k})}$$

5.2 Model II: Mutual Mode Entropy (MME)

Mutual Mode Entropy, or MME, is an extension of the mode entropy algorithm. It is used to measure whether there is a high degree of coupling between different sequences[4]. If the difference between the mutual mode entropy values of two different RNA sequences and the sequence's own mutual mode entropy value is small, it means that the internal complexity of these two sequences is similar, i.e., the degree of similarity is high, i.e., they are homologous. The procedure of calculating the mutual mode entropy is similar to the definition of the mode entropy, and its calculation steps are as follows：

First, for two sets of time series containing data whose number is N+1,

$$\{u(i)\colon 1 \le i \le N+1\}$$
$$\{\mathrm{v(j)\colon 1 \le j \le N+1}\}$$

Take consecutive data points whose number is m from each set of data and compose their corresponding m-dimensional vectors, respectively,

$$\mathrm{x(i)=[u(i),u(i+1),...,u(i+m-1)]}$$
$$\mathrm{y(j)=[v(j),v(j+1),...,v(j+m-1)]}$$

The vectors x(i) and x(j) are both vectors containing consecutive data points whose number is m. And for each point time series whose number is N, there are such vectors whose number is N-m+1, where “m” denotes the so-called encoding length, which is used as the core reference quantity in the whole algorithm calculation process.

The calculation process of the judgment basis-basis value for each vector is,

$$\mathrm{B_u(i)=\frac{\sum_{i=1}^{m}u(i+1)}{m},B_v(j)=\frac{\sum_{i=1}^{m}v(j+1)}{m}}$$

Immediately afterwards, the data vector required for the experiment is reconstructed,

$$\begin{aligned}\Psi_x(i) &= [u(i)-B_u(i), u(i+1)-B_u(i), \ldots u(i+m-1)-B_u(i)] \\ &= [\varphi_u(i), \varphi_u(i+1), \ldots, \varphi_u(i+m-1)]\end{aligned}$$

$$\begin{aligned}\Psi_v(j) &= [v(j)-B_v(j), v(j+1)-B_v(j), \ldots u(j+m-1)-B_v(j)] \\ &= [\varphi_v(j), \varphi_v(j+1), \ldots, \varphi_v(j+m-1)]\end{aligned}$$

Define the distance between the vectors as L_{ij} it is the value that differs the most between the data of the corresponding elements of the two vectors.

$$L_{ij} = L[\Psi_u(i), \Psi_v(j)] = \max_{k=1 \rightarrow m} [|\varphi_u(i+k) - \varphi_v(j+k)|]$$

Since each vector has points whose number is m, there are differences whose number is m. And L_{ij} is determined by the largest one of these differences. After that, the probability that the two vectors are similar is defined as

$$C_i^m(r) = \frac{1}{N-m+1} \sum_{j=1}^{N-m+1} \theta(\mathrm{r} - \mathrm{L_{ij}})$$

Theta(θ) is the unit step function,

$$\theta(z) = \begin{cases} 1, & z > 0 \\ 0, & z \leq 0 \end{cases}$$

In fact, the tolerance “r” is a crucial parameter. Its value is very delicate, because when too large a value of r is chosen as the basis of judgment, it will lead to too few vectors of pattern similarity obtained by statistics, which will produce a great deviation to the actual results. And when the value of r is chosen too small, it has almost no filtering effect on the experimental data, and in serious cases, it even leads to the situation that it is impossible to calculate, which is very unfavorable to the calculation of the experiment.

r takes the value of,

$$0.1 \sim 0.25 * |\mathrm{cov}(\mathrm{u}, \mathrm{v})|$$

Next, take the logarithm of c and divide it by the total number of vectors to get the probability value at encoding length m. Then, let m increase by 1 to m+1 dimensions and recalculate all the above steps to get the probability value at m+1 dimensions. Finally, the two values are subtracted to get the MME value we need.

$$\varphi^m = \frac{1}{N-m+1} \sum_{i=1}^{N-m+1} \ln C_i^m(r)$$

$$MME(m, r, N) = \varphi^m \text{-} \varphi^{m-1}, m \geq 1$$

Or

$$\mathrm{Moden}(\mathrm{m}, \mathrm{r}, \mathrm{N}) = \varphi^m - \varphi^{m+1}, \mathrm{m} \geq 1$$

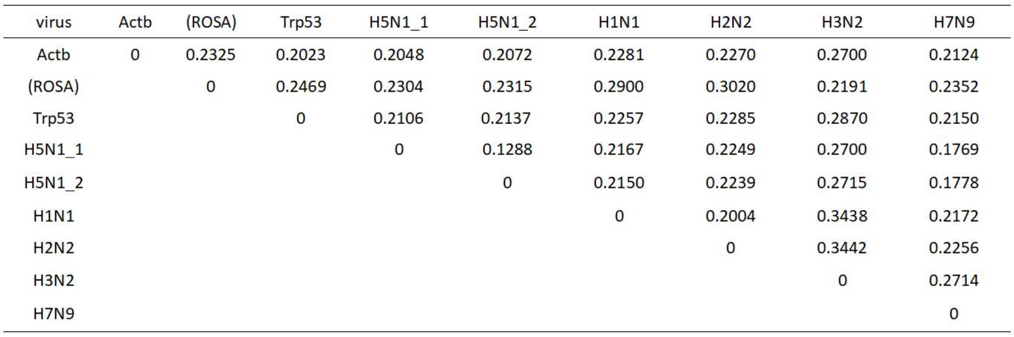
$$=\frac{1}{N-m+1}\sum_{i=1}^{N-m+1}\ln C_r^m(r)-\frac{1}{N-m}\sum_{i=1}^{N-m}\ln C_r^{m+1}(r)$$

$$\overset{N\to\infty}{=}-\frac{1}{N-m+1}\sum_{i=1}^{N-m} ln\frac{C_i^{m+1}(r)}{C_i^{m}(r)}$$

The mutual pattern entropy algorithm is a quantification of the similarity between different sequences, while the pattern entropy is a measure of the complexity of its own sequences. The mutual pattern entropy algorithm can be used not only to calculate the differences between different RNA sequences, but also to analyze the differences between different interval segments of the same sequence, which is of great importance for the analysis and research of RNA sequence similarity.

6 Solution II

6.1 Evaluation of the algorithm by Model I



virus	Actb	(ROSA)	Trp53	H5N1_1	H5N1_2	H1N1	H2N2	H3N2	H7N9
Actb	0	0.2325	0.2023	0.2048	0.2072	0.2281	0.2270	0.2700	0.2124
(ROSA)		0	0.2469	0.2304	0.2315	0.2900	0.3020	0.2191	0.2352
Trp53			0	0.2106	0.2137	0.2257	0.2285	0.2870	0.2150
H5N1_1				0	0.1288	0.2167	0.2249	0.2700	0.1769
H5N1_2					0	0.2150	0.2239	0.2715	0.1778
H1N1						0	0.2004	0.3438	0.2172
H2N2							0	0.3442	0.2256
H3N2								0	0.2714
H7N9									0

Figure 6: results by Model I

No matter which representation method is used, the distance values between H5N1_1 and H5N1_2 are smaller than those between other sequences, which means that the two DNA sequences of H5N1_1 and H5N1_2 have a high degree of similarity, and from the data point of view, these two genes belong to the more correlated genes, so DTW is very good at the task of calculating the correlation between two RNA sequences task. When the shortest distance is less than 0.2, it means that the two sequences already have a relatively strong correlation, from which the rule can be deduced to other sequences.

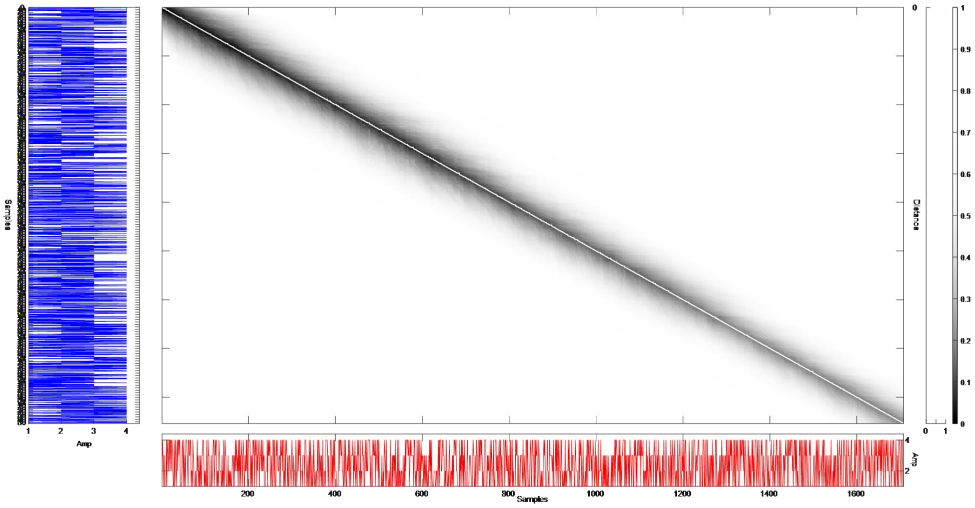


Figure 7: Sequence distances between 4 and 5

The above figure shows the black and white matrix diagram of sequence distances. The shortest distance, 4 and 5, which are the two sequences with the strongest correlation in the previous analysis, is selected by drawing a black and white matrix diagram of the distances between the two sequences, and the darker the color shown on the diagram means the closer the distance is.

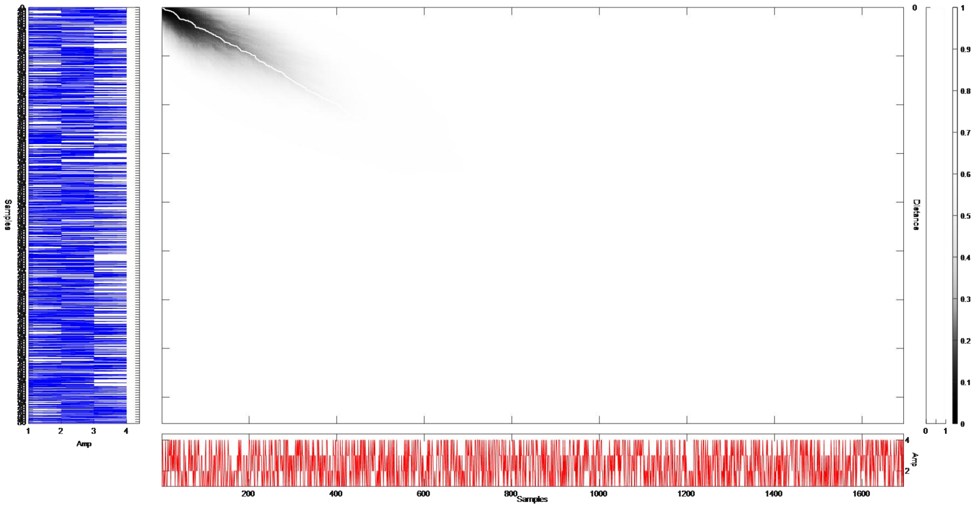


Figure 8: Sequence distances between 5 and 9

Of course, for the black and white distance matrix plots between other sequences, the results don't look as good. The above figure, which is chosen as the shortest distance between 5 and 9, is not as strongly correlated, and a large number of gaps appear near the midline, implying that the two sequences are far apart from each other.

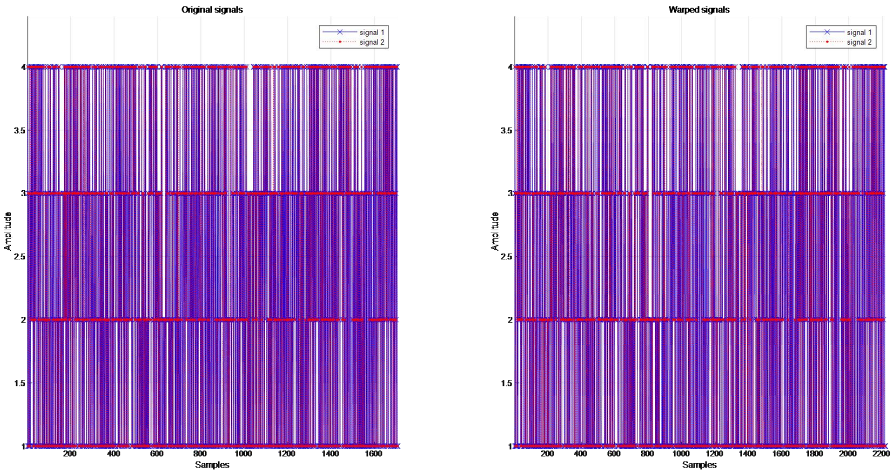
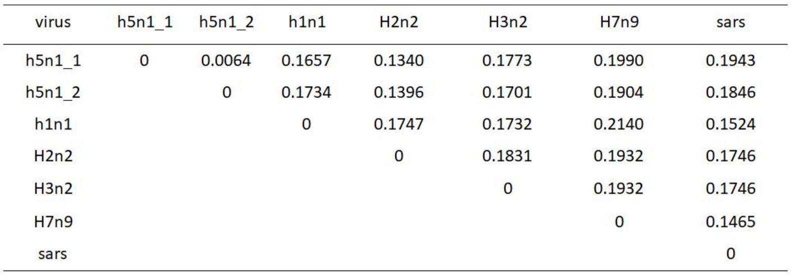


Figure 9: comparison between origin and distort

The above figure shows the original 4 and 5 sequences and the distorted 4 and 5 sequences. Through visualization, we have a deeper knowledge of the inner logic of the DTW algorithm, unlike the Euclidean distance, the advantage of the DTW algorithm is that it can handle two sequences of arbitrary length. When the sequences are of different lengths or the time steps are not aligned, the Euclidean distance is unable to effectively calculate the distance between two time sequences, especially at the peak time. In contrast, DTW can stretch or shorten (compress and expand) the unknown quantity until it is the same length as the reference template, in which the unknown sequence is twisted or bent so that its characteristic quantity corresponds to the standard pattern.

6.2 Evaluation of the algorithm by Model II

The fragment information of these seven viral RNA sequences are in the form of strings, which are not conducive to experimental analysis and research, so they need to be transformed into time series, and in the simulation experiments of this paper, the integer representation (referred to as Method I) will be used, and the mapping relationship is as follows, A=0 c=1 g=2 t=3.



virus	h5n1_1	h5n1_2	h1n1	H2n2	H3n2	H7n9	sars
h5n1_1	0	0.0064	0.1657	0.1340	0.1773	0.1990	0.1943
h5n1_2		0	0.1734	0.1396	0.1701	0.1904	0.1846
h1n1			0	0.1747	0.1732	0.2140	0.1524
H2n2				0	0.1831	0.1932	0.1746
H3n2					0	0.1932	0.1746
H7n9						0	0.1465
sars							0

Figure 10: results by model II

The original RNA sequences were transformed into time series, and the parameters of the mutual mode entropy algorithm were m=2, r=1, and the number of samples N=908. The mutual mode entropy values were calculated for each of the seven viral RNA sequences. According to the definition of mutual mode entropy algorithm, the smaller the mutual mode entropy value between the sequences, the more similar the two sequences are, then the same sequence, its own mutual mode entropy must be the smallest. Therefore, based on the mutual mode entropy of the same sequence, we calculate the difference with this value and get the seven viruses. The matrix of relative values of mutual mode entropy between RNA sequences is shown above.

It can be found that,

- The mutual mode entropy difference between H5N1(1) and H5N1(2) is much smaller than the mutual mode entropy difference between other RNA sequences, indicating that the similarity between H5N1(1) and H5N1(2) is high.

- The mutual mode entropy values of H5N1(1) and h7n9 are the largest in the first row of the table, which means that the degree of similarity between H5N1(1) and the other six viruses is the smallest among the relationships between H5N1(1) and h7n9.

- Under the RNA representation based on integer values, the relative values of mutual mode entropy between H5N1(1) and H5N1(2) are two orders of magnitude different than those between other sequences, which can intuitively show the magnitude of similarity among the seven viral RNA molecules.

Therefore, the mutual mode entropy algorithm is practical for the similarity study between RNA sequences, and it can effectively distinguish the similarity between different RNA sequences.

The macroscopic RNA sequence similarity study is a direct reflection of the sequence similarity by an entropy value of the sequence as a whole, but some detailed information between sequences will be missed. From the perspective of local analysis, dynamic analysis of RNA sequences is used to obtain more detailed information and observe the complexity of RNA sequences from the graphs.

The so-called dynamic analysis is actually to choose a reasonable sliding window to segment the RNA time sequences to calculate their similarity degree. The most critical step of this method is to choose a relatively short and suitable time window, and then use this time window as the scale to slide the RNA time series from the beginning to the end, and then use the data points contained in this time window as new samples for analysis, and use the entropy algorithm to calculate the complexity of this new sample, so as to obtain a dynamic curve to represent the complexity of RNA sequences. This dynamic curve can well demonstrate the locations where different RNA sequence bases exist differently.

The schematic diagram is as follows,

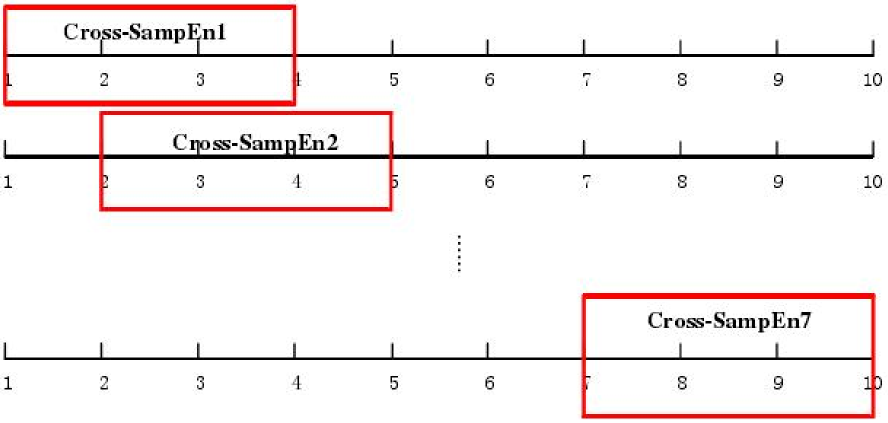
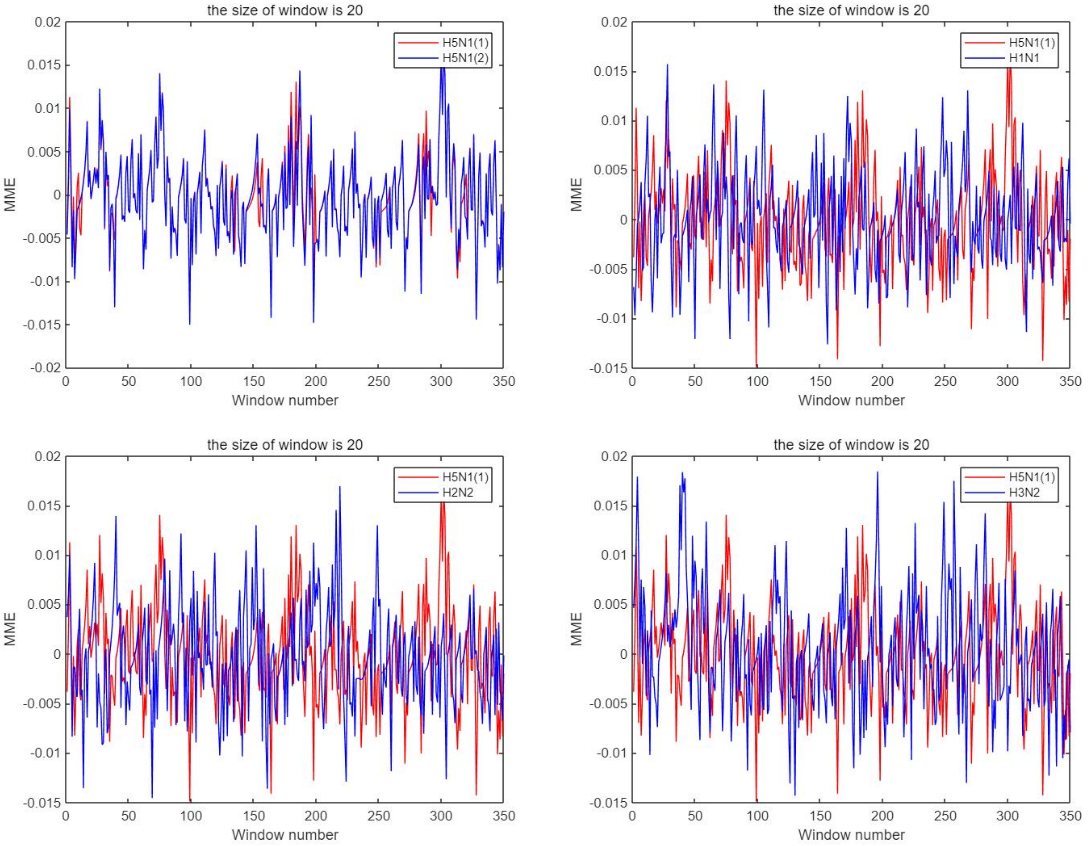


Figure 11: Principle

Since the smaller the scale of the time window, it is easier to determine the position where the differences between sequences occur in the RNA sequence similarity dynamic analysis experiments. Considering the efficiency and accuracy of the experiment, a sliding window of 20 points was taken as the most appropriate when performing RNA sequence similarity dynamic analysis. The number of samples was adopted as 350 points.



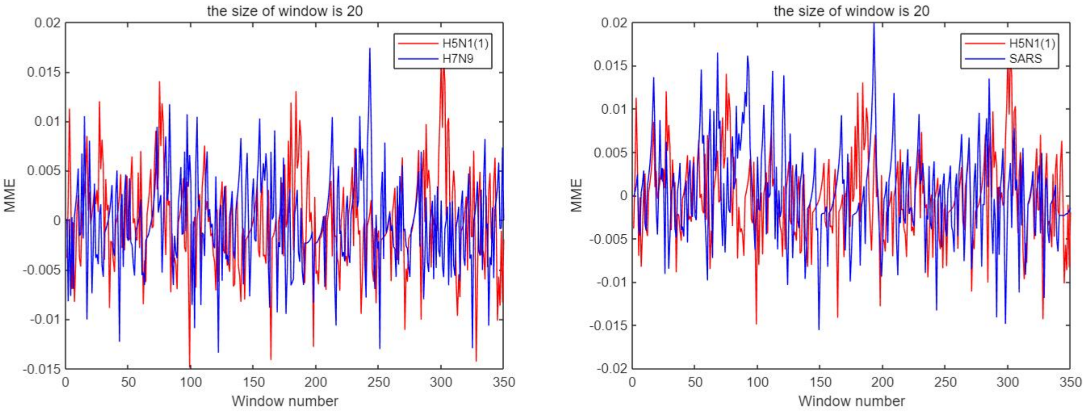


Figure 12: Samples to confirm parameters

It can be seen that,

- The curve overlap ratio in the first graph is much greater than the other five curves for the graph, which means that the patterns of changes in H5N1(1) and H5N1(2) are very similar; while in the other five graphs, there is almost no curve overlap and the pattern entropy values of the sequences differ significantly. This phenomenon once again proves that the degree of similarity between H5N1 (1) and H5N1 (2) is very high, and it can be concluded that these two viruses are homologous, and also shows that the pattern entropy algorithm can effectively distinguish RNA sequences of different complexity.

- The curves in the first figure, we can know that the two curves can be said to be completely overlapping between the 15th point to the 140th point of the sequence, which means that there are obvious differences between the 14 bases of these two sequences, while there is a great possibility that the bases between the 15th point of the sequence and the 160th point of the sequence (140+20) are exactly the same.

The pattern entropy algorithm is suitable for studying the similarity between various large and slowly changing signal sequences. Next, the mutual mode entropy algorithm was introduced to analyze the similarity between seven viral RNA sequences, and it was shown that the mutual mode entropy was feasible to measure the similarity between different RNA sequences. On this basis, the effects of different RNA representation methods and coding length m on the mutual pattern entropy values between sequences were discussed. Finally, the "best match" window for dynamic analysis of RNA sequences with different RNA representation methods was investigated, and the obtained "best match" window length was used to determine the local similarity of RNA sequences.

The best-match window length was used to simulate the local similarity of RNA sequences, and the pattern entropy curve of RNA sequences under this window can well observe the degree of sequence similarity and effectively reflect the base positions where the sequences differ. The similarity intervals between homologous sequences were also analyzed by using the length of the "best fit" window.

7 Solution III

7.1 Algorithm, Sequence Family Tree Construction

Procedure 1 Data collection. The DNA fragment sequence data required for topic III of this paper were downloaded from the NCBI database (http,//www.nubi.nlm.nih.gov) for TAIR10.1 in Arabidopsis (Arabidopsis thaliana). The data types collected were protein coding region files (GCF_000001735.4_TAIR10.1_cds_from_genomic.fna.gz), DNA and protein sequence comparison files in FASTA format (GCF_000001735.4_TAIR10.1_genomic.fna.gz) and genome annotation file (GCF_000001735.4_TAIR10.1_genomic.gff.gz), and then use the pfam database (http,//pfam-legacy.xfam.org/) to collect the NB-ACR gene family of Arabidopsis thaliana, and download the entire NB-ACR gene family through the "curation &model" option to download the entire gene family.

Procedure 2 Download hmmsearch software, add it to your computer path, and use hmmsearch software to compare between gene families by "hmmsearch.exe . \NB-ARC.hmm . \GCF_000001735.4_TAIR10.1_protein.faa >out.txt" can save the data data to "out.txt" and store it to the current path.

Procedure 3 In the third step, only the part of evalue value less than 0.05 was kept, and the results of sequence comparison were extracted by TBtools, input and output paths were set, and the protein sequences related to NB-ARC gene family of Arabidopsis thaliana were obtained after running.

Procedure 4 To perform protein structural domain analysis, the sequence ids of the genes needed in this paper were selected by score, extracted and saved by tbtools, and then motif analysis was performed on the MEME software by loading the file saved by tbtools, setting the structural domain to 10, performing motif analysis, and downloading the "MEME XML output" file obtained from the analysis.

Procedure 5 Using tbtools software, open the gene view in graphics, copy the sequence id of the gene sequence to the id box, and load the file path of the MEME XML output to get the corresponding structural domain position of the protein, as shown in Figure 13.

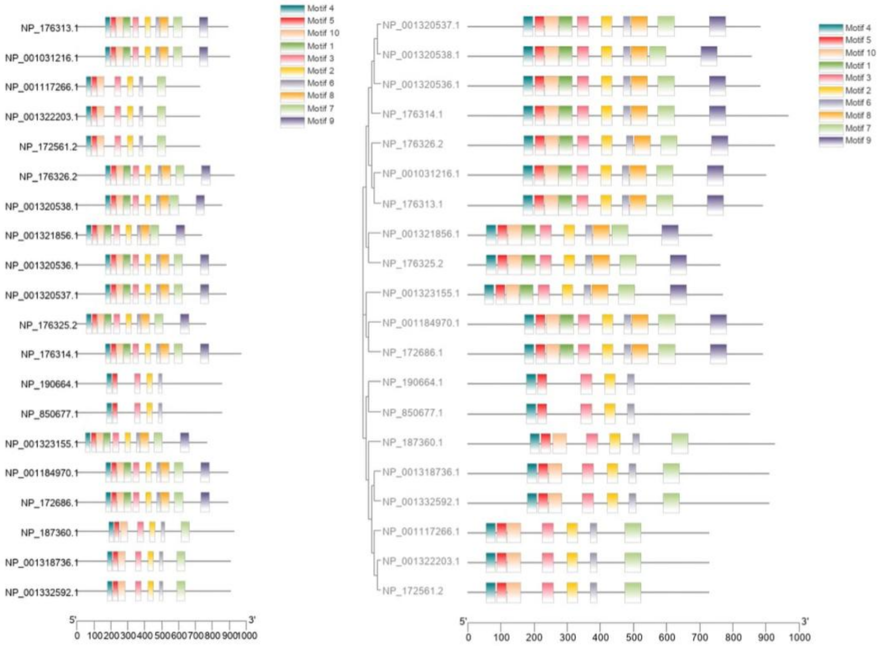


Figure 1 3(left):Protein structural domain analysis, Figure 14(right):Sequence evolutionary tree of Arabidopsis thaliana (with protein structural)

Procedure 6 To construct the evolutionary tree, we first simplified the sequence id of the protein sequences related to the NB-ARC gene family in Arabidopsis, and then used matlab to construct the evolutionary tree to obtain the upper figure 15. Of course, we can also load the information of protein structural domain analysis into the evolutionary tree to get the following figure 14.

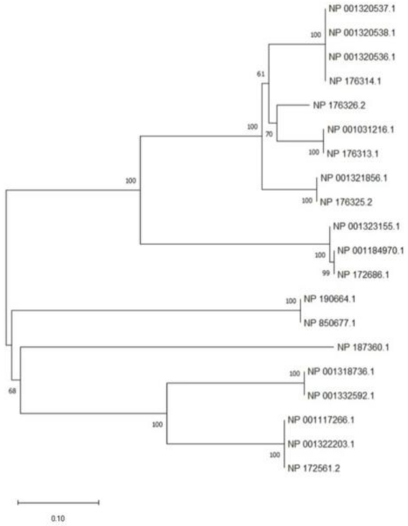


Figure 15: Sequence evolutionary tree of Arabidopsis thaliana

7.2 Algorithm: GIGA

The first step is to describe the new concept of gene trees, which will simplify our interpretation of the rules for determining the tree topology from the order of operations determined by pairwise sequence distances. In this concept, the gene tree consists of "direct homologous subtrees"[5], i.e., sequences that contain sequences related to species formation events.

Each OS contains at most one gene from each organism, and each sequence in the subtree is directly homologous to the other sequences. Different directly homologous subtrees are linked together by events involving the replication (or transfer) of genetic material to produce a gene tree that creates new loci in the ancestral genome. When replication is from the same genome (e.g., tandem gene replication or whole genome replication), the linkage event is a gene replication event; when replication is from another genome, the linkage event is a horizontal transfer event[6].

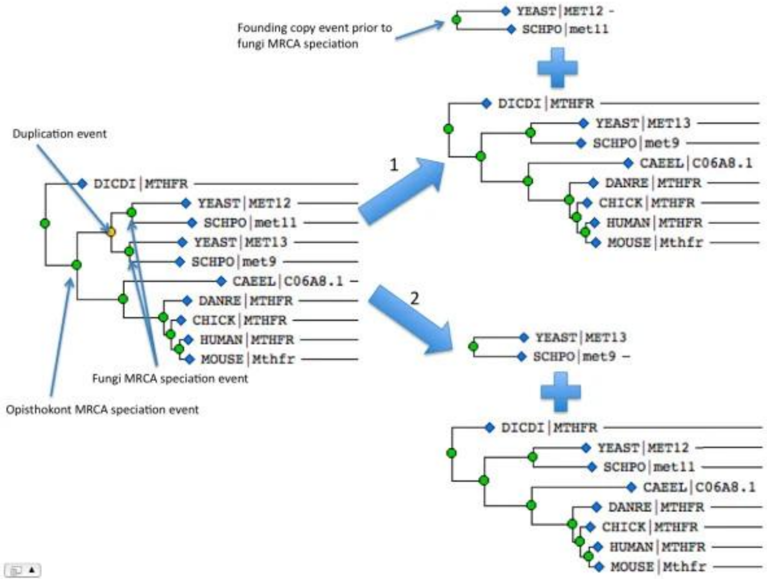


Figure 16:Orthologous subtrees and gene trees

Procedure 1 Pre-processing and setup. Decide on the genomes to be included; construct a "known" species tree for these genomes. For each protein family. Assemble a "complete" set of genes for a given family. Create a multiple sequence alignment of the genes in the family. Select homologous loci for sequence comparison in the alignment. If more than 15% of the weighted sequences are gapped at the site, we "trim" the alignment by removing the site. The sequences were weighted using the procedure of Karplus et al. Representation of sequence divergence at homologous loci. In the spirit of trying the simplest model first, we simply calculate the distance between each pair of sequences as the sequence divergence score of the selected homologous site.

Procedure 2 For each protein family, the gene tree topology is inferred by iteratively defining the immediate homology groups and how these groups are related by gene duplication events. Initialization: each sequence starts in its own operating system. Consider the closest pair of sequences that were not processed in the previous steps and perform one of the following operations on the two operating systems containing them. Connect the two OSes by repeating the event and find the event relative to the morphological event in each OS (rule 3). If two OSs together have two genes from a single organism, then they will be connected by duplicate events (Rule 2) if either of the following conditions also holds. No founding duplicate event for either OS has been found previously. Founding duplicate events were previously found only for OS1 and not OS2, and joining the two operating systems will not conflict with this position. In other words, the phylogenetic span of OS2 must be less than or equal to the phylogenetic span of OS1. This constraint means that joining the two operating systems does not require us to modify our earlier assumptions about the timing of repeating events.

Initial repeating events are initially estimated to minimize the number of implied deletions, i.e., the FDE of the OS with the newer MRCA is set to immediately precede that MRCA. Combining two separate operating systems into one (Rule 1). Merging is only allowed if the sequences are unlikely to be fragments (rule 5). If neither sequence is a fragment, the two operating systems will be merged if one of the following conditions is met. A founding duplicate event for either OS has not been found. A founding duplicate event has been found for OS1 only, not OS2, and merging the two operating systems will not conflict with this position. In other words, the MRCA morphological event for the merged OS is the same as OS1. This constraint means that merging the two OSs does not require us to modify our earlier assumptions about the timing of gene duplication events. The founding duplication event was found only for OS1 but not OS2, and merging the two OSs would conflict with this position, but there is sufficient sequence evidence to support the revised position of the duplication event (Rule 4). We first calculate the standard deviation of the distance between OS2 and OS1 (dist1 and std_dev1) and the distance between OS2 and OS1's sibling (dist2 and std_dev2). If the siblings of OS2 and OS1 have no species overlap and are likely to be directly homologous, we require

$$\text{dist1-dist2}>1.5\ (\text{std\_dev1+std\_dev2})$$

Otherwise, we need a less stringent criterion

$$\text{dist1-dist2}>0.5\ (\text{std\_dev1+std\_dev2})$$

Try to add fragments back into the tree. Allow each fragment to try to merge or join events, depending on the shortest distance between the fragment and any non-fragment.

Procedure 3 Infer the branch length of the tree. We propose to use the tree topology generated by GIGA and estimate branch lengths and ancestor sequences using an ML-based procedure such as PAML. However, in the spirit of simple algorithms, we default to computing an approximate reconstruction of each ancestor sequence (a local, parsimony-like algorithm that reconstructs each node using only its descendants and the nearest outgroup) and then computing the branch length as the sequence difference between neighboring nodes in the tree, including the Jukes-Cantor correction. Infer the ancestral sequence of each node. We do this in a simple way, starting recursively from the leaf nodes (only the existing sequences, the leaves, are known). For each non-leaf node, we consider the descendant nodes and their nearest outgroup nodes. If the sequence of the nearest outgroup nodes has not been determined, define the outgroup using its descendants. If more than half of the descendant nodes align to the same amino acid at the given locus, infer that it is the most likely ancestral amino acid. If the progeny do not agree and the outgroup agrees with one of them, the outgroup amino acid is inferred to be the most likely ancestral amino acid. Otherwise, the ancestral amino acid is considered unknown ("X"). Calculation of branch length based on node sequences. We use a simple metric, i.e. the ratio of sequence differences between parent and child nodes. the Jukes-Cantor correction is applied to this value. However, in one respect we have to be very careful and only calculate distances for a subset of selected sites. After a repetition event, it is usually the case that one of the repetition terms continues to preserve ancestral functions more closely, while the other repetition term diverges more quickly. We can identify the "least divergent" direct homologs by tracking the shorter branches. Because of the rate heterogeneity between sites, relative branch lengths are reliable only when they are calculated at the same site. Therefore, in our algorithm, for branches following repeated events, lengths are calculated using only those sites that are aligned among all descendant nodes.

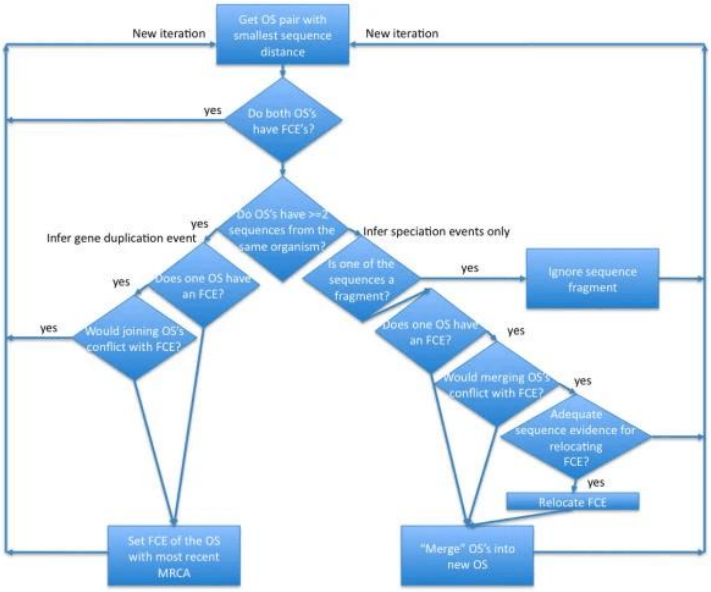

Get OS pair with smallest sequence distance
New iteration
New iteration
yes
Do both OS's have FCE's?
Do OS's have >=2 sequences from the same organism?
yes
Infer gene duplication event
Infer speciation events only
yes
Does one OS have an FCE?
yes
Would joining OS's conflict with FCE?
Is one of the sequences a fragment?
yes
Ignore sequence fragment
Does one OS have an FCE?
yes
Would merging OS's conflict with FCE?
yes
Adequate sequence evidence for relocating FCE?
yes
Relocate FCE
Set FCE of the OS with most recent MRCA
"Merge" OS's into new OS

Figure 17: Core of the algorithm

8 Sensitivity Analysis

As a baseline, we first assessed the robustness of TreeBeST trees, by comparing the TreeFam "clean" and "full" trees. If the TreeBeST algorithm were perfectly robust to the addition of sequences, the topology of the TreeFam full tree would be identical (for the subset of sequences also in the clean tree) to the clean tree. To identify deviations from perfect robustness, we calculated the RF distance between the TreeBeST clean and full trees. Figure below (red bars) shows that TreeBeST is reasonably robust to the additional sequences. In relatively few cases are the TreeBeST trees for the clean and full alignments identical (5.7%) but most are very similar (57% have a distance less than 0.2; 88% have a distance less than 0.4). We note that the TreeBeST algorithm itself is somewhat different for full and clean alignments, as full trees are estimated using only protein sequences, while clean trees can use nucleotide as well as protein sequences.

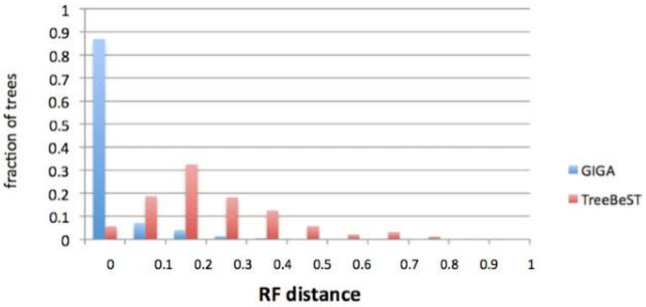


Figure 18: Robustness of tree inference algorithms

Full alignments include additional sequences, but the alignment is the same as for the clean set. An RF distance of 0 indicates that the tree topology is unchanged by adding more sequences. Overall, GIGA is more robust than TreeBeST to the perturbation of adding sequences.

To test the robustness of GIGA, we then constructed two separate GIGA trees for each TreeFam family, one from the "clean" protein alignment, and one from the "full" alignment. We then calculated the RF distance between the two GIGA trees to measure how much the additional "full" sequences changed the topology inferred for the "clean" sequences. We found that GIGA is considerably more robust than TreeBeST to the perturbation of adding sequences (Figure 11, blue bars), with over 85% of the trees being completely unchanged (RF distance of 0) and 98% changing in RF distance by less than 0.2. The robustness of GIGA is remarkable, and is due largely to the strong constraints provided by the rules described above.

9 Strengths and Weaknesses

9.1 Strengths

- Accuracy and stability.We use both mutual mode entropy and shortest distance to evaluate the correlation between sequences, and we know from the calculation results that both models accomplish the analysis task well, and the results are stable and fast.
- Good expansibility and flexibilty. When calculating mutual mode entropy and shortest

distance, any two unequal RNA sequences can be correlated without considering the effect of data length, and at the same time, for any time series, mutual mode entropy and shortest distance models can be applied to calculate the correlation of any time series.

- Visualization Analysis. When calculating the shortest distance, the black and white matrix of the distance between two sequences can be completed at the same time, and the logic within the algorithm is visualized.

9.2 Weaknesses

- Need more data processing. Due to the time problem, nine different sequences and Arabidopsis sequence data were selected for processing in this paper, although the problem was solved, but the addition of more data can more fully demonstrate the excellent nature of the model in calculating the correlation between sequences, and even have some inspirational meaning to the evolutionary process.
- Simplifying assumption. During the assumption process, many problems were ignored, such as the transfer of large base sequences, which are indeterminate factors in the model and have little impact on the correlation calculation, but will affect the sequence tree construction.
- Lack of physiological explanation. In the correlation analysis, although the RNA sequence code was constructed, this way is an assumption, lack of physiological explanation, and the calculated sequence correlation is difficult to explain its specific details meaning from the physiological point of view completely.

10 Conclusion

RNA sequences are changing all the time, and we were asked to construct models to perform correlation analysis between RNA sequences and evaluate their complexity and accuracy, based on which a tree between sequences is constructed so as to label the original and intermediate sequences. In the process of calculating correlation, this paper constructs the coding mode of RNA sequences and performs the calculation of shortest distance and mutual mode entropy, inexamples entropy performs in addition to very good accuracy, and mutual mode entropy completes the correlation analysis task better. In the process of constructing trees between sequences, the construction of Arabidopsis sequence family trees was completed using hmmsearch and Tbtools; finally we used the algorithm of GIGA to rapidly infer gene phylogenies while allowing phylogenetic reconstruction of very large gene families and the determination of large-scale direct homologs.

References

[1] Li, Shuchao, Xu, Jin, Liu, Guangwu. DNA-based computation of RNA sequences with multipoint "mutations" and digital coding of the first base and its algorithm[J]. Computer Engineering and Applications, 2004, 40(8):15-18,128. doi:10.3321/j.issn:1002-8331.2004.08.006.

[2] Fan X.S.,Lei Y.J.,Lu Y.L.,et al. Long-term intuitive fuzzy time series prediction model based on DTW[J]. Journal of Communication,2016,37(8):95-104. doi:10.11959/j.issn.1000-436x.2016160.

[3] Sun To, Xia Fei, Liu Hongbo. Solving time series DTW centers based on dynamic programming[J]. Computer Science,2015,42(12):278-282. doi:10.11896/j.issn.1002-137X.2015.12.060.

[4] An S. J., Zhou S. A., Zhang J., et al. DNA sequence similarity analysis based on mutual pattern entropy[J]. Intelligent Computers and Applications,2019,9(6):52-54. doi:10.3969/j.issn.2095-2163.2019.06.010.

[5] Zou Xinhui,Ge Song. Gene tree conflict and phylogenetic genomics[J]. Journal of Plant Taxonomy,2008,46(6):795-807. doi:10.3724/sp.j.1002.2008.08081.

[6] Tan Yuande,Yan Hengmei. Comparison of topological distances and gene clustering in mtDNA gene trees[J]. Journal of Animal Taxonomy,2002,27(2):205-211. doi:10.3969/j.issn.1000-0739.2002.02.002.