A Survey on RNA Alignment Algorithm

Guosong Jiang

***Abstract*—In bioinformatics, Alignment is an ancient and classic method or technique. The underlying problem of sequence alignment can be attributed to the study of the similarity of two or more symbol sequences. The design of sequence alignment is to reason about its evolutionary relationship by using sequence similarity at the nucleic acid or amino acid level, and then to infer its structural and functional connections. This kind of analysis is of great significance for biochemistry, genetic research, molecular research, and the study of the origin of life. The simplest form of sequence alignment is a pairwise alignment. By this pairwise alignment, the reserved or reserved sites between the two sequences are found, and then the evolutionary relationship between the two is explored. Based on the pairwise alignment, it can be developed as an alignment of multiple sequences, looking for reserved regions or reserved sites between multiple sequences, looking for their evolutionary connections. After determining their evolutionary connections, it is easy to splicing their higher-level structural components based on sequence alignment, and then completing the final structural prediction.**

***Index Terms*— RNA structure prediction, RNA secondary structure, sequence alignment, sequence searching**

I. INTRODUCTION

In early RNA sequence alignments, it is common practice to use global sequence alignments [1]. The shortcoming of this approach is obvious. The amount of data in RNA is usually very large, and the overall comparison is not very practical. Later, the usual practice is to use a local comparison method. In the scoring matrix, the sequence is used as the row vector of the matrix. Each point in the matrix represents the similarity score of the corresponding two bases [2]. The higher the score, the higher the similarity, the lower the score, the lower the similarity, the score The matrix can be used for pairwise alignment of sequences. In this case, the sequence alignment is the problem of finding the most suitable alignment path in the matrix. At present, the most effective method is the dynamic programming algorithm proposed by Needleman Wunsch, and the improved SIM algorithm and Smith Waterman algorithm based on it. Their time complexity is generally $O(m \times n)$, and the space complexity is generally $O(m \times n)$[3]. Such complexity is not suitable for massive RNA sequence comparison tasks, and is only suitable for general laboratory. A small amount of data for the task. How to improve the spatial and temporal execution efficiency of the algorithm to cope with massive RNA sequence alignment is the current research hotspot of RNA sequence alignment.

On the issue of dealing with and solving the complexity of space, the Hirschberg algorithm is proposed. This is an improved dynamic programming algorithm. The space complexity is $O(m + n)$, but the cost is the required time [4]. Longer, he doubled the time required for traditional dynamic programming algorithms. There is also an FA algorithm, which seeks a balance between the traditional dynamic programming algorithm and the Hirschberg algorithm. His time and space complexity are both in the two, so in practice, the algorithm is better than the former two. More valuable. But more worthy of attention is a new fast sequence alignment algorithm proposed by Ukkonen. His time complexity is $O(n + d^2)$ (where n is the sequence length and d is the sequence score), and the space complexity It is $O(n^2)$, which is more powerful when dealing with large-scale sequence databases [5].

With the development of DNA sequence alignment technology, RNA sequence alignment technology has also made great progress in the last 20 years [6]. There are already many open source open-use comparison tools available. For pairwise sequence alignment, in addition to the two algorithms described above: the Needleman-Wunsch method of the classical method of the most global alignment algorithm, and the Smith-Waterman method as the classical method of the local alignment algorithm, FASTA, Two methods of BLAST are two more popular methods. For multi-sequence alignment, the traditional dynamic programming method is quite complex due to the complexity of multi-dimensional data matrix in sequence comparison [7]. The current multi-sequence matching tools that are often used in practice often use heuristic algorithms, using progressive Thought to reduce the complexity of the operation. The CLUSTAL algorithm proposed by Feng and Doolittle in 1987 is widely used [8]. He is a heuristic algorithm using a progressive idea, borrowing the objective fact that similar sequences have evolutionary homology [9].

Compared with the minimum entropy method, the sequence alignment method has higher accuracy and is widely recognized [10]. Among the RNA structures currently accepted by the database, except for a small part which is determined by physicochemical methods, a considerable amount of RNA structure is determined by sequence comparison. In principle, the sequence alignment method can be roughly divided into three types. Classic sequence alignment, Sankoff methods, folding sequence methods[11].

Classical sequence alignment, which predicts the public domain of RNA from both biological evolutionary and energy perspectives. This method is also the most successful sequence alignment method, but he relies more on the quality of the data set. She needs the data set to be sufficiently homologous so that the alignment can be performed. At the same time, there must be sufficient differentiation so that the variation zone can be predicted. Classical sequence alignment methods are usually based on the assumption that there are high quality retention regions between sequences. Such

Manuscript received Oct 09, 2018

Guosong Jiang, School of Computer Science & Software Engineering, Tianjin Polytechnic University, Tianjin, 300387, China

hypotheses are often not true for non-coding RNAs because this type of RNA evolves more frequently and at a faster rate. The Sankoff method combines the alignment and prediction steps. This algorithm requires a very large running space and running time. Generally, its time complexity is $O(n^{3m})$ and its spatial complexity is $O(n^{2m})$[12].

The folding sequence method is a special sequence alignment method. It compares the structure of RNA rather than sequence, which requires the data set to have reliable known structural information of RNA [13]. This method is usually used when high quality sequence retention regions cannot be observed. At this time, the folding sequence method predicts the structure of the entire RNA sequence according to the predicted structure and the structure. Since the structure of RNA has nested features, this type of folding sequence method tends to adopt a tree-like data structure [14]. Tree comparisons or tree comparison models have been widely used.

Among the above three methods, the most successful and most popular is the classic sequence alignment method. The following is a brief introduction to the classic sequence alignment method [15].

II. THE THEORY OF CLASSICAL ALIGNMENT

2.1 The concept of alignment

The principle of the sequence alignment method is as follows: the biological world is divided into many classes according to its own classification [16]. Different classes of organisms are generally considered to be advanced from the same ancestor. The process of evolution is thought to be the result of both the mutation of the gene and the survival of the fittest. The result of evolution is that the closely related organisms have homology to DNA or RNA in many important functional regions, as shown by their great similarity in sequence, in the final formation of gene products (such as RNA and protein). The structures are very similar, and then their functions are the same or similar. Even there are many creatures that cross the gate. In some functional sections, strong homology can also be observed. After all, all living things on Earth are generally thought to have evolved from several primitive creatures with very similar similarities.

2.2 The concept of homologous RNAs

RNA alignment is given by several homologous RNAs, which regions of the RNA are analyzed as belonging regions and which regions belong to the region of variation. It should be emphasized that sequence alignment is the alignment of homologous RNA sequences and does not make any sense for non-homologous RNA alignments [17]. At the same time, it is necessary to distinguish the similarities between homology and similarity. If they are not similar, they must be homologous, but the homology must be similar. Similarity is only a manifestation of homology, and homology is an objectively evolved homolog. Thus, it can be found that given a target RNA, if it can be successfully aligned with the RNA in the database, the structure of the target RNA can be easily confirmed, thereby completing the final RNA structure prediction.

In order to distinguish between similarities and homology between sequences, the following is specifically described. One of the main goals of sequence alignment is to allow people to determine the homology between sequences by having sufficient similarity between the sequences. The similarity of sequences can be abstracted as the degree of similarity between symbol sequences, usually resolved to the appearance of a certain parameter, such as the percentage of bases of the same type, or other more detailed weighting methods. It should be noted that similarity is only a result of comparison between sequences, and does not explain any objective facts in evolution. And homology refers to sequences from the same ancestor, they have the same evolutionary history in the original stage. So in summary, although similarity and homology exist in some cases, they are two different concepts [18]. Similarity refers to an intuitive comparison, such as some measured parameters are similar or similar. Homology is a historical fact of reasoning. It is derived from the data that two sequences have the same evolutionary process in evolutionary history and have common ancestors. Similarity is quantifiable, and depending on the degree of similarity, specific similarity values can be formulated. And homology is not quantifiable, only divided into homologous or different sources. However, practical observations show that significant similarity generally stems from homology.

2.3 The theory of alignment

If you know the whole process of evolution, you can uniquely determine the alignment matrix of these homologous sequences. However, people often only know the results of evolution. Therefore, the goal here is to predict such a comparison matrix and reflect as much as possible the true evolution. In general, obvious similarities in nucleotide sequences are often associated with homology without exception, and thus it is necessary to restore this homologous relationship.

In order to evaluate the relative accuracy of the comparison matrix, the relative accuracy of the matrix is measured by generating a quality score for the comparison matrix, and the matrix with the higher quality score is considered to be the optimal ratio among all possible alignment matrices. For the matrix. The mass score must be able to reasonably reflect the accuracy of the matrix and be easy to calculate. In fact, such a scoring system affects the design of the overall optimal algorithm. Therefore, the first and most important task is to develop a good scoring system. The current common scoring systems include SP scoring system, WSP scoring system, Star scoring system, ME scoring system, Tree scoring system and ML scoring system [19].

lf you think that every step of the biological evolution process is done by one replacement or a single insertion loss. Then in the ML scoring system, the probability of each step needs to be calculated, which will lead to a huge amount of calculation. In other scoring systems, each replacement or insertion deletion is given a certain deduction, and the replacement deduction table is replaced by a replacement matrix. The value of the deduction is derived from the observation and statistics of the occurrence of various variations. It's easy to see that conversions of the same type are more likely to occur than heterogeneous conversions, so fewer pages are deducted. In practice, the insertion or loss of the mutation event is more complicated to deal with than the replacement, because several insertions or losses may occur in a mutation event. The easiest way to deal with this is to ignore this one-time

event and split them into individual insertion or loss events, so the possibility of the final sequence formation is the accumulation of a single event. But in facts this is not reasonable, the probability of this series of insertions or losses occurring as the length of the inserted or lost sequence increases dramatically. Therefore, it is also necessary to consider the gap deduction. In summary, here is a simple comparison example to illustrate the score of the sequence alignment.

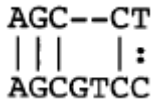

Fig 1. An example of RNA alignment

$$H = S(A,A) + S(G,G) + 2S(C,C) + S(T,C) + g(2) \qquad (1)$$

The final score for this alignment consists of five items, the first four being the replacement score and the last one being the insertion loss score.

After determining how to score the sequence alignment, the remaining task is to find the optimal sequence alignment matrix. Dynamic programming techniques are often used here to construct such optimal sequence alignment matrices. Generally speaking, there are two kinds of dynamic programming techniques applied here, one is the fold line comparison analysis, and the other is the score matrix method. The former applies to sequence alignments of smaller lengths, and the latter applies to sequence alignments of longer lengths, which are not discussed in detail herein.

The above mainly introduces the classical sequence alignment method, which is also called the first-order structure comparison. At the same time, there are secondary structure alignments, which are also used for the prediction of RNA secondary structure. It will not be discussed in detail here.

III. CONCLUSION

The problem of sequence alignment is a difficult but very meaningful research topic. It has made continuous progress in decades of research, but it is still far from perfect, especially some of them have considerable challenge. Some scientists predict that unless mathematics and computers and other disciplines make major substantive breakthroughs in traditional theory, it will be difficult to solve them perfectly. The existing sequence alignment method is based on a dynamic programming algorithm, the data structure and algorithm are complex, the sequence data is too large, and the model design in the comparison makes the running of the program very time consuming, and the storage pressure increases. These have seriously affected the work of sequence alignment, and our understanding of RNA is still at a low level.

REFERENCES

[1] O.Gotoh, Multiple Sequence Alignment: Algorithms and Applications, Adv.Biophys, 1999, 36:159-206,

[2] B.A.Shapiro, Y.G.Yingling, W.Kasprzak, E.Bindewald, Bridging the Gap in RNA Structure Prediction, Current Opinion in Structural Biology, 2007,17:157–165..

[3] M.Rother, K.Rother, T.Puton, J.M.Bujnicki, ModeRNA: A Tool for Comparative Modeling of RNA 3D Structure, 2011, 39(10):4007-4022.

[4] Z.Xu1, D.H.Mathews, Multilign: an Algorithm to Predict Secondary Structures Conserved in Multiple RNA Sequences, Bioinformatics, 2011, 27(5):626–632.

[5] Z.Miao, R.W. Adamiak, M.F.Blanchet, M.Boniecki, RNA-Puzzles Round II: Assessment of RNA Structure Prediction Programs Applied to Three Large RNA Structures, RNA ,2015,21:1066–1084.

[6] S.G.Jones, A.Bateman, M.Marshall, A.Khannal, S.R.Eddy, Rfam An RNA Family Database, Trust Sanger Institute, 2009.

[7] N.B. Leontis, E.Westhof, Geometric Nomenclature and Classification of RNA Base Pairs, RNA, 2001, 7:499–512.

[8] H.Yang, F.Jossinet, N.Leontis, L.Chen, Tools for the Automatic Identification and Classification of RNA Base Pairs, NAR, 2003, 31(13): 3450–3460.

[9] M.Sprinzl, K.S.Vassilenko, Compilation of tRNA Sequences and Sequences of tRNA Genes, NAR, 2005, 33:139-140

[10] F.J.hling, M.Morl, R.K.Hartmann, M.Sprinzl, tRNAdb 2009: Compilation of tRNA Sequences and tRNA Genes, NAR, 2009, 37:159-162.

[11] M.G.Seetin, D.H.Mathews, TurboKnot: Rapid Prediction of Conserved RNA Secondary. Structures including Pseudoknots, bioinformatics, 2012, 28(6): 792–798

[12] K.Sato, Y.Kato, M.Hamada, T.Akutsu, IPknot: Fast and Accurate Prediction of RNA Secondary Structures with Pseudoknots using Integer Programming, BIOINFORMATICS, 2011, 27:i85-i93.

[13] B.Liu, D.H.Mathews, D.H.Turner, RNA Pseudoknots: Folding and Finding, Biology Reports, 2010, 2:8..

[14] R.Nussinov, A.B.Jacobsont, Fast Algorithm for Predicting the Secondary Structure of Single-stranded RNA, Proc.Nati.Acad.Sci., 1980, 77(11): 6309-6313..

[15] B.A.Shapir, K.Zhang, Comparing Multiple RNA Secondary Structures using Tree Comparisons, CABIOS, 1990, 6(4):309-318.

[16] Namhee Kim, Tamar Schlick, Graph Applications to RNA Structure andFunction,RESEARCHGATE,2013.https://www.researchgate.net/publication/278662209

[17] H.H.Gan, D.Fera, J.Zorn, N.Shiffeldrim, RAG: RNA-As-Graphs Database—Concepts,Analysis, and Features, BIOINFORMATICS, 2004, 20(8): 1285–1291.

[18] C.Laing, S.Jung, N.Kim, S.Elmetwaly, Predicting Helical Topologies in RNA Junctions as Tree Graphs, PLOS ONE, 2013, 8(8): e71947.

[19] P.P.Gardner, R.Giegerich, A Comprehensive Comparison of Cmparative RNA Structure Prediction Approaches, BMC Bioinformatics, 2004, 5:140.