A Non-coding RNA Sequence Alignment Algorithm Based on Improved Covariance Model

Xiaodan Liu*, Yu Wang, Na Zhang

Institute of Information and Intelligent Technology
Shaanxi Radio and Television University
Shannxi 710119, China
E-mails: liuxiaodan0003@163.com,
moolan@qq.com, 315634944@qq.com

*Corresponding author

Received: January 11, 2019 **Accepted: July 23, 2019**

Published: September 30, 2019

Abstract: *This paper attempts to overcome the inefficiency of covariance model (CM) in the search for non-coding sequence. For this purpose, the members of non-coding RNA family were compared and the CM of the family was discussed in details. Next, the CM was improved for structural units in the secondary structure, through the addition of the upper and lower limits on subsequence length. Based on the length distribution of each structural unit, the improved model limits the number of insertions and deletions during the evolution of sequences in the same family. After that, the author put forward a novel non-coding RNA sequence alignment algorithm. The experimental results show that the proposed algorithm can greatly reduce the computing time of non-coding RNA sequence comparison.*

Keywords: *Non-coding RNA, Sequence structure, Covariance model, Secondary structure.*

Introduction

Sequence alignment mainly measures the similarity or difference between two or more symbolic sequences. This task cannot be completed without a mathematical model. In fact, different models depict the features of sequences from different angles. The main basis of sequence alignment is the evolutionary theory. During the evolution of biological sequences, some residues remain unchanged, some mutate into other types, and some simply disappear. In addition, some new residues are inserted into or added to the ends of the sequence [1, 2, 7].

Biological sequences can be divided into nucleic acid (DNA and RNA) sequences and protein sequences. Both nucleic acids and proteins are 1D non-branched chain macromolecules. The former is polymerized by four nucleotides, and the latter, by twenty amino acids. Therefore, a biological sequence can be considered as a sequence of symbols selected from an alphabet of four or twenty characters, depending on its category. Despite its simplicity, this symbolic sequence contains the mysteries of life.

Biological sequence analysis is the key and basic problem in bioinformatics, which mainly explores the sequences of various biological macromolecules and derives gene structure, function and evolution from abundant sequence information. Currently, the analysis on nucleic acid sequences mostly focuses on three issues: the gene information in coding and non-coding regions, the gene interactions and the similarity or difference between different types of genomic sequences. As a typical nucleic acid, the RNA is usually described as a linear unstructured sequence [16]. Many non-coding RNAs have advanced 3D structures, some of which may even catalyze biochemical reactions [12].

Non-coding has been proved to have a direct bearing on many biological processes, such as gene regulation [3, 8], chromosome replication [10] and RNA modification [9]. For a non-coding molecule, the biological function mainly depends on its secondary structure. However, the secondary structures of many homologous non-coding sequences are similar or identical, although these sequences seem very different. Hence, the sequence analysis of non-coding RNA is much more complex than that of DNA or proteins, and thus cannot be completed satisfactorily by traditional tools for sequence analysis.

There are three types of detection methods for new non-coding RNA in genomes, namely, the family specific approach, the pattern matching method and the statistical section method. The last approach has been increasingly applied in sequence analysis. Two strategies have been developed based on statistical section for the search of homologous non-coding RNA: easy RNA profile identification (ERPIN) [13] and covariance model (CM) [11]. The CM is the most widely used non-coding RNA sequence analysis model.

The CM can simulate the secondary structure of the non-coding RNA family [15]. During sequence alignment, this method determines whether a sequence fragment in the genome belongs to the non-coding RNA family. Like the hidden Markov model (HMM), the CM describes the statistical distribution of nucleotides in each position or each position pair in a gene family sequence. The alignment between a single sequence and the CM of a single non-coding RNA family can be calculated by a dynamic programming algorithm. Based on the statistics of the CM, the algorithm searches for the alignment with the highest probability, and judges whether the sequence belongs to the non-coding family according to the maximum probability. In this way, the primary sequence and secondary structure of the non-coding family can be illustrated in an accurate manner.

The main bottleneck of the CM lies in the computing speed, because of the low space-time efficiency of the dynamic programming algorithm. In the CM, the computing time of the algorithm soars with the growth in the length of the target non-coding sequence. Since most genome sequences contain numerous nucleotides, it is very inefficient to search for long non-coding sequences using the CM. To solve the problem, this paper compares the members of non-coding RNA family and analyzes the CM of the family. In light of the results, the CM was improved to compute the length constraint of structural units in the secondary structure, and then a novel non-coding RNA sequence alignment algorithm was developed. Finally, the proposed algorithm was verified through experiments.

RNA structure

The RNA structure can be divided into primary structure, secondary structure and tertiary structure. The primary structure is a finite linear sequence of four different nucleotides arranged in a single RNA chain. The four nucleotides, adenine (A), cytosine (C), guanine (G) and uracil (U), can form hydrogen bonded base pairs like G-C, A-U and G-U. These base pairs respectively have three, two and one hydrogen bonds. Since the number of hydrogen bonds is positively correlated with stability, G-C and A-U are usually referred to as typical pairings, and G-U, as atypical pairing. The continuous pairing of bases produces a double helix structure known as stem, which stabilizes the secondary structure [6]. By contrast, the noncircular structure of the RNA molecule weakens structural stability. The secondary structure of RNA strikes the balance between the stabilizing and weakening effect. The seven sub-structural types of the secondary structure of RNA are listed in Table 1 below.

Table 1. The seven sub-structural types of the secondary structure of RNA

Name	Description
Stem	A continuous pair of right-handed double helices formed by hydrogen bonds in complementary chains
Single strand	Unpaired nucleotide chains between two stems
Hairpin loop	A stem ended with unpaired nucleotides
Terminal bulge	An unpaired nucleotide chain at stem ring junction
Lateral bulge	An unpaired nucleotide chain on steam ring
Internal bulge	The absence of a classical pair of nucleotides between two chains
Multibranched loop	A group of nucleotides linked to the base of several stems

The secondary structure is normally illustrated by a 2D graph. For example, the secondary structure of yeast alanine transfer RNA is shown in Fig. 1, where each dot stands for a base paring and each number means the ordinal number of bases.

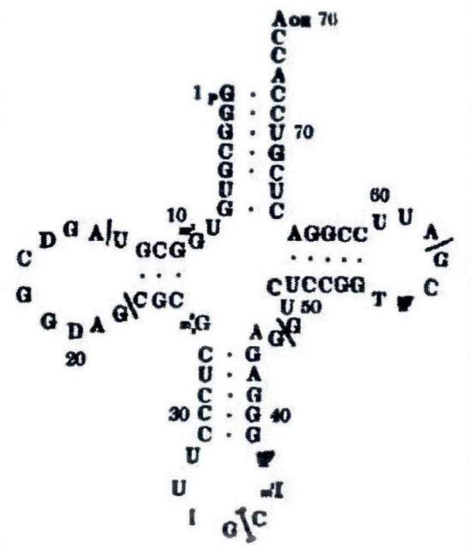

Fig. 1 The secondary structure of yeast alanine transfer RNA

Almost all base pairs in the secondary structure appear to be nested. In some cases, however, the base pairing may take a cross nested form, constituting a pseudoknot. A typical pseudoknot is shown in Fig. 2. In general, the RNA secondary structure has fewer pseudoknots than base pairs. As a result, the pseudoknot information can be neglected to improve the search efficiency.

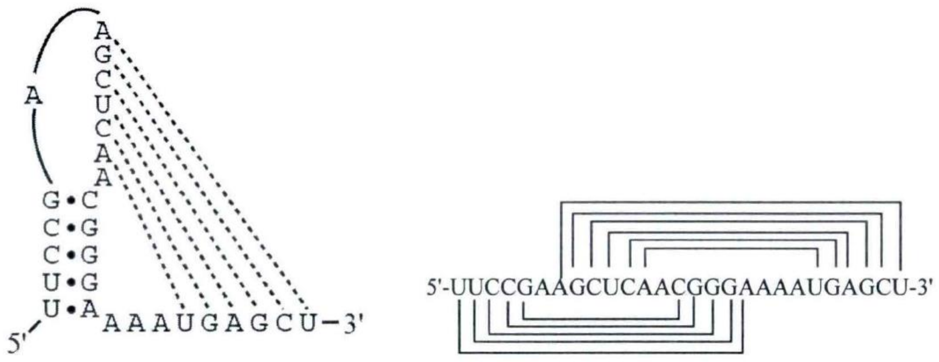

Fig. 2 Pseudoknot structure

Due to the base complementarity of RNA, the secondary structures of many homologous non-coding sequences are similar or identical, although these sequences seem very different. Taking the sequence UUUUGGGAAAA for instance, the base pairs of UUUU and AAAA are paired into the stem, while the unpaired GGG forms signal chain rings. This sequence has the same secondary structure with other sequences like GGGGAAACCCC and GCGCAAACGCG. Despite the difference in primary structure, the secondary structures of these sequences gradually resemble each other through evolution, especially at the occurrence of multiple mutations. Hence, it is more accurate to compare the secondary structures rather than the primary structures in multi-sequence alignment. During the search for homologous RNA, the traditional models that only consider sequence similarity are no longer applicable, and should be replaced with novel tools like the CM.

The covariance model

The CM is a probabilistic model capable of describing the secondary structure [5]. Fig. 3 shows the structure of a non-coding RNA family.

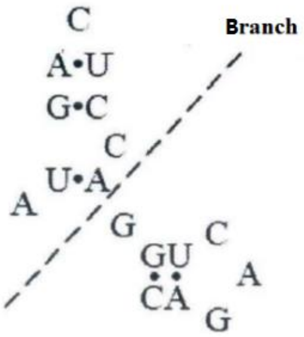

Fig. 3 An example of RNA structure

For a non-coding RNA without pseudoknot, the base pairs appear nested in the structure. As shown in Fig. 4, the paired bases are connected by lines that do not intersect each other.

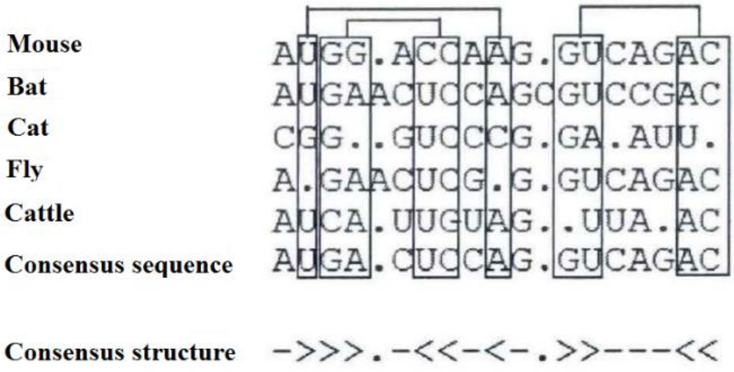
Mouse AUGG.ACCAAG.GUCAGAC
Bat AUGAACUCCAGCGUCCGAC
Cat CGG..GUCCCG.GA.AUU.
Fly A.GAACUCG.G.GUCAGAC
Cattle AUCA.UUGUAG..UUA.AC
Consensus sequence AUGA.CUCCAG.GUCAGAC
Consensus structure ->>>.-<<-<-.>>---<<

Fig. 4 Five example sequences of different animals with the same structure

The CM was extended from the stochastic context-free grammars (SCFG) model [4]. The latter can align multiple sequences of RNA without vacancies and model secondary structure of uniform sequences. Let $p(x|\theta)$ be the probability that SCFG model θ produces a different string x. Then, the probabilities of all production formulas derived from the same nonterminal will total 1.

The alphabet of the SCFG model can be expressed as {A, U, C, G}. The model contains six states: pair launch (*P*), left launch (*L*), right launch (*R*), branch (*B*), start (*S*) and end (*E*).

The production rule of these states is specified in Table 2, where W is a nonterminal character and it is any of the six states.

Table 2. The production rule of SCFG states

State	**Production rule**
P (16 pairs of launch probability)	$P \rightarrow aWb$
L (4 single launch probabilities)	$L \rightarrow aW$
R (4 single launch probabilities)	$R \rightarrow Wa$
B (the probability is 1)	$B \rightarrow SS$
S (the probability is 1)	$S \rightarrow W$
E (the probability is 1)	$E \rightarrow \epsilon$

If each nonterminal character is generated for a uniform sequence, then SCFG model can be obtained for families as Table 3.

Table 3. SCFG model

	Stem 1	**Stem 2**
$\boldsymbol{S_0 \rightarrow L_1}$ ...	$S_3 \rightarrow P_4$	$S_{10} \rightarrow L_{11}$
$\boldsymbol{L_1 \rightarrow \alpha B_2}$ ...	$P_4 \rightarrow \mu R_5 \alpha$ ...	$L_{11} \rightarrow gP_{12}$ ...
$\boldsymbol{B_2 \rightarrow S_3 S_{10}}$	$R_5 \rightarrow P_6 c$ ...	$P_{12} \rightarrow gP_{13}c$ ...
	$P_6 \rightarrow gP_7c$ ...	$P_{13} \rightarrow \mu P_{14}\alpha$ ...
	$P_7 \rightarrow \alpha L_8 \mu$ ...	$L_{14} \rightarrow cL_{15}$ ...
	$L_8 \rightarrow cE_9$ ...	$L_{15} \rightarrow \alpha L_{16}$ ...
	$E_9 \rightarrow \epsilon$ ...	$L_{16} \rightarrow gE_{17}$ ...
		$E_{17} \rightarrow \epsilon$ ...

After linking up the nonterminal characters, SCFG analysis tree can be set up as Fig. 5. The SCFG analysis tree, as a graphic representation of the SCFG model, offers a simple yet intuitive depiction of the structure of the target RNA family. Nevertheless, the SCFG model ignores the vacancies in the sequence, as compared with the alignment family sequence in Fig. 4. The direct application of the non-vacant SCFG model will overlook many homologous sequences. This calls for extension of the SCFG model.

Inspired by the extension of the HMM to profile HMM [17], the SCFG model was expanded into the CM to handle state insertion and deletion. The CM was constructed in the following steps: setting up the non-vacant SCFG analysis tree for uniform sequence, expanding and matching the nodes of the tree, inserting and deleting states, and linking up the states by state transition lines. The directed state graph thus obtained is the CM. The node extension rule of the CM is provided in Table 4, where *MP*, *ML* and *MR* are matching states, *IL* and *IR* are insertion states and *D* is a deletion state.

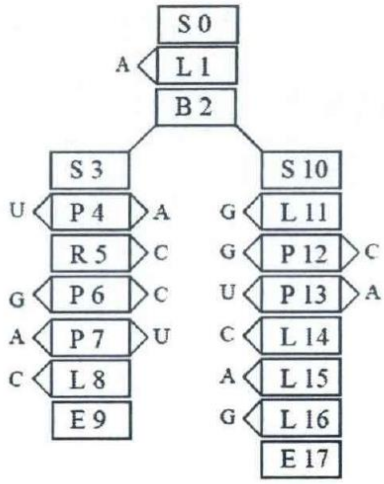

Fig. 5 SCFG analysis tree

Table 4. Node extension rule of the CM

SCFG state	Node type	Extended state
P	MATP	*MP, D, ML, MR, IL, IR*
L	MATL	*ML, IL, D*
R	MATR	*MR, IR, D*
S	ROOT	*S, IL, IR*
	BEGL	*S*
	BEGR	*S, IL*
B	BIF	*B*
E	END	*E*

Taking the nodes *L*11 and *P*12 in Fig. 5 for instance, the extended internal state transition structure of the two nodes is described in Fig. 6 below.

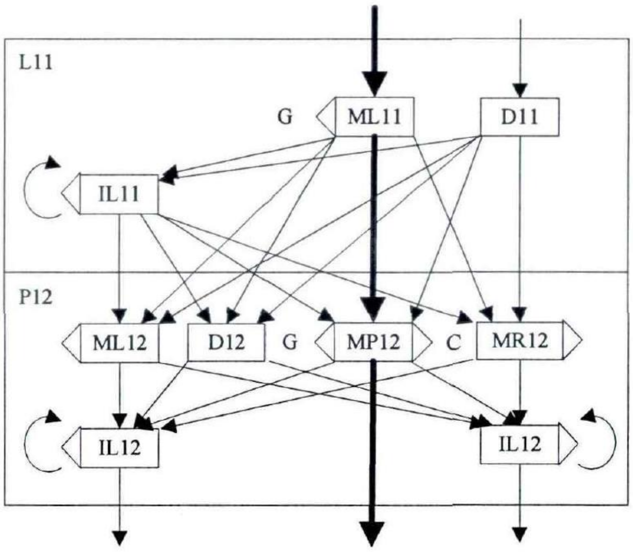

Fig. 6 Internal state transition structure of nodes *L*11 and *P*12

As shown in Fig. 6, the internal state of each node has one or two layers. The upper layer usually includes all states other than the insertion state, while the lower layer contains the insertion state. The transition probability can be expressed by the thickness of state transition lines. For example, the thick line above $ML11$ indicates the high probability of the transition from the previous state to $ML11$. The character G at $ML11$ means this character is more likely to be generated than any other character in the state $ML11$.

The final CM is a directed state graph of K different states. Different states are linked up by state transition lines. Each state has its own state transition probability and character generation probability. Note that character generation probability describes how likely each nucleotide or base pair appears at that location. The different states of the CM, plus their character generation probabilities and state transition probabilities, are listed in Table 5, where Δ_K^L and Δ_K^R are the number of characters generated to left and right in state *v*, respectively, $e_K(a,b)$ is the probability that characters a and b are matched in state *k*, $e_K(a)$ is the probability that character *a* is generated in state *k*, and $t_K(Y)$ is the transition probability from state *k* to state *Y*.

Table 5. The CM states, character generation probabilities and state transition probabilities

State	Rule	$\boldsymbol{\Delta_K^L}$	$\boldsymbol{\Delta_K^R}$	Character generation probability	State transition probability
MP	$P \rightarrow aYb$	1	1	$e_K(a,b)$	$t_K(Y)$
ML	$L \rightarrow aY$	1	0	$e_K(a)$	$t_K(Y)$
IL		n	0	$\prod_{i=1}^{n} e_K(a_i)$	$t_K(Y)$
MR	$R \rightarrow Ya$	0	1	$e_K(a)$	$t_K(Y)$
IR		0	n	$\prod_{i=1}^{n} e_K(a_i)$	$t_K(Y)$
D	$D \rightarrow Y$	0	0	1	$t_K(Y)$
S	$S \rightarrow Y$	0	0	1	$t_K(Y)$
B	$B \rightarrow SS$	0	0	1	1
E	$E \rightarrow \epsilon$	0	0	1	1

A novel non-coding RNA sequence alignment algorithm

Improvement of the CM

According to the recent studies on non-coding RNA families, the optimal alignment of a family member's RNA sequence with the family's CM is like the length of a consistent structure in any state of the model. Here, the ByeB non-coding RNA family is cited to improve the CM [14]. This family has 15 members, whose mean length is about 100.

The author constructed a CM and removed the nodes *S*, *B* and *E*, forming a compressed CM with 88 nodes. Next, all 15 non-coding RNA sequences were adjusted according to the CK. On this basis, the length difference was computed between the subsequences and the uniform structure at each state. The results show that the subsequences and the uniform structure had basically consistent lengths in each state.

If using the traditional CM, all subsequences need to be compared in the meantime of dynamic programming. In the ByeB non-coding RNA family, the upper limit of subsequence length D is 150. If the CM is adopted to compute the 88th node, all subsequences whose length deviations between -1 and +149 should be compared. According to the test results of the ByeB family, 149 of the 150 search sites are useless.

Based on the above analysis, the traditional CM was improved by dividing the secondary structure of RNA family into several basic structural units, each of which represents a stem or a loop. The basic structural units in the secondary structure of RNA family are presented in Fig. 7, where each stem is the cumulation of successive symmetrical bases and each loop is a chain sequence bounded by base pairs.

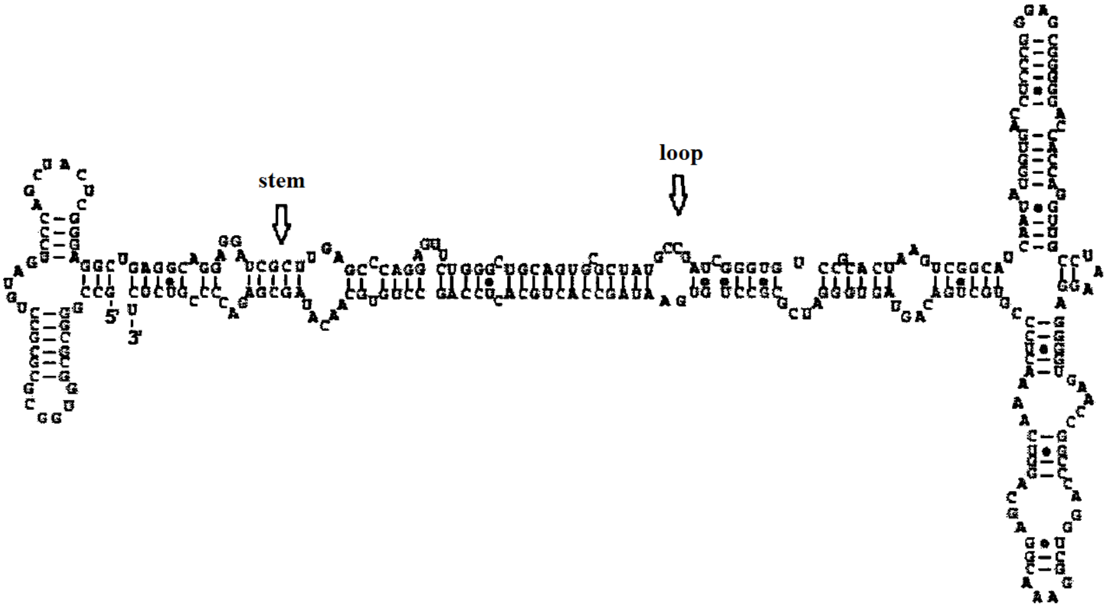

Fig. 7 Basic structural units in the secondary structure of RNA family

In Fig. 7, each structural unit is associated with a pair of integers, which are the lower and upper limits on the length of the structural unit. The two limits form an interval on the length of each structural unit. The addition of the limits is the only difference between the improved CM and the original CM.

Algorithm construction

This subsection mainly develops a non-coding RNA sequence alignment algorithm, which enables the improved CM to determine whether a single RNA sequence L belongs to the target family. In the algorithm, the length interval of the subsequences derived from each state in the CM can be derived from the length limit of each structural unit.

For the improved CM, if state $S(k)$ is not a branch state (B), then the corresponding length interval can derived as $[d_{min}, d_{max}]$. In this case, the dynamic programming only needs to cover all the subsequences whose length is $d_{min} < d < d_{max}$ in L. The required computing time is $O((d_{max} - d_{min})L)$.

Let $[s_{min}, s_{max}]$ and $[l_{min}, l_{max}]$ be the two states corresponding to state $S(k)$, when the latter is a branch state (B). Then, the length interval of the subsequences derived from $S(k)$ is $[s_{min} + l_{min}, s_{max} + l_{max}]$. For each sequence whose length falls within the interval, the branch point's position cannot fall beyond $b = \min\{s_{max} - s_{min}, l_{max} - l_{min}\}$. In this case, the required computing time is $O((s_{max} + l_{max} - s_{min} - l_{min})bL)$.

Considering all the K states in the model and the length interval of the derivable subsequences, the total computing time required for dynamic programming cannot surpass $O(K\Delta^2L)$, where Δ is the maximum length of all these length intervals.

The length interval of a structural unit can be estimated roughly as follows. Let n and δ be the contribution of each nucleotide in the subsequence to the mean and variance of the interval, respectively. If the mean length of the structural unit is l, then every length value $l^{\prime}$ in its length interval must satisfy:

$$|l^{\prime} - l|/l\delta \leq c/\sqrt{n}.$$

Since n, c and δ are constants, the above formula indicates that the length interval of the structural unit is not more than $2cl\delta/(n\sqrt{N}) = O(1/\sqrt{N})$. If $\mathrm{N} = l^2$, then the length interval must be a constant.

Based on the above discussion, the time complexity of dynamic programming can be derived as:

$$\Delta = O(g),$$

where g is the number of branch states in the model. Hence, the computing time of dynamic programming is not more than $O(Kg^2L)$, which is much shorter than that of the traditional CM.

Experimental analysis

The proposed algorithm was contrasted with the CM-based search algorithm through a simulation experiment, using the data from the Rfam database. For each genome family, a maximum of 60 sequences were selected. The similarity between every two sequences was lower than 80% of the minimum similarity between the seed sequences. Several RNA sequences from the same genome family were inserted into randomly generated sequences with the same base composition. Then, the proposed algorithm and the CM-based search algorithm were applied to search for the inserted sequences. To determine the length limit of each structural unit, it is assumed that the lengths of structural units obey the normal distribution. The confidence p and constant c were set to 0.01 and 3, respectively.

The precision of the two algorithms was measured by two indices: sensitivity and specificity. The sensitivity reflects the percentage of non-coding RNA sequences that can be identified by an algorithm out of all sequences in the family. The specificity refers to the percentage of non-coding RNA sequences correctly identified by an algorithm out of all sequences in the family. The sensitivities and specificities of our algorithm and the CM-based search algorithm are compared in Table 6 below.

Table 6 shows that our algorithm achieved comparable or high precision, compared with the CM-based search algorithm, on all tested non-coding RNA families. This means our approach manages to reduce the number of candidate configurations in sequence structure alignment, without sacrificing alignment accuracy.

Table 6. The comparison of sensitivity and specificity between the two algorithms

RNA	Mean length	Sensitivity of our algorithm	Specificity of our algorithm	Sensitivity of the CM-based search algorithm	Specificity of the CM-based search algorithm
Entero_CRE	62	0.76	0.96	0.82	1
Entero_OriR	72	0.92	1	1	1
Let_7	82	1	1	1	1
Lin_4	69	1	1	1	1
Purine	101	0.93	0.98	0.91	1
SECIS	66	0.92	0.86	1	0.96
S_box	108	0.88	1	1	1
Tymo	82	1	1	1	0.96

Table 7 compares computing times of our algorithm and the CM-based search algorithm. It is obvious that our algorithm computed faster than the contrastive algorithm on all tested non-coding RNA families.

Table 7. The comparison of the computing time between the two algorithms

RNA	Computing time of our algorithm	Computing time of the CM-based search algorithm
Entero_CRE	2.32	56.32
Entero_OriR	3.56	101.76
Let_7	8.77	151.68
Lin_4	1.38	128.75
Purine	3.61	176.21
SECIS	6.78	182.37
S_box	22.67	721.89
Tymo	2.68	180.21

Conclusions

This paper improves the traditional CM to speed up the non-coding sequence alignment. The traditional CM describes the sequence and secondary structure of the non-coding RNA gene family through statistical method. In our research, the length of each structural unit in the secondary structure was limited, based on the CM's description of sequence components and secondary structure of non-coding RNA genes. The addition of length constraint aims to shorten the computing time required for sequence alignment. The experimental results show that our approach achieved the same accuracy as the traditional CM in the search for non-coding RNA in the genome, despite consuming much less computing time.

Acknowledgements

Key Topic of Shaanxi Radio & TV University.

Topic No.: 18DA08.

Topic Name: Educational Data Analysis and Teaching Evaluation Research under the Background of Big Data.

References

1. Alejandro O., J. D. Storey, M. Llinás, M. Singh (2015). Beyond the E-value: Stratified Statistics for Protein Domain Prediction, PLOS Computational Biology, 11(11), e1004509.
2. Bissantz C., A. Logean, D. Rognan (2010). High-throughput Modeling of Human G-protein Coupled Receptors: Amino Acid Sequence Alignment, Three-dimensional Model Building, and Receptor Library Screening, Cheminform, 35(31), 1162-1176.
3. Duan J. (2015). Path from Schizophrenia Genomics to Biology: Gene Regulation and Perturbation in Neurons Derived from Induced Pluripotent Stem Cells and Genome Editing, Neuroscience Bulletin, 31(1), 113-127.
4. Folk J. R., R. K. Morris (2003). Effects of Syntactic Category Assignment on Lexical Ambiguity Resolution in Reading: An Eye Movement Analysis, Memory & Cognition, 31(1), 87-99.
5. Han T., J. K. Kim (2014). Driving Glioblastoma Growth by Alternative Polyadenylation, Cell Research, 24(9), 1023-1024.
6. Khan A., H. Ahmed, N. Jahan, S. Raju Ali, A. Amin, M. N. Morshed (2016). An *in silico* Approach for Structural and Functional Annotation of *Salmonella enterica* serovar *typhimurium* Hypothetical Protein R_27, Int J Bioautomation, 20(1), 31-42.
7. Lalwani S., R. Kumar, N. Gupta (2015). A Novel Two-level Particle Swarm Optimization Approach to Train the Transformational Grammar Based Hidden Markov Models for Performing Structural Alignment of Pseudoknotted RNA, Swarm and Evolutionary Computation, 20, 58-73.
8. Laver J. D., X. Li, K. Ancevicius et al. (2013). Genome-wide Analysis of Staufen-associated mRNAs Identifies Secondary Structures that Confer Target Specificity, Nucleic Acids Research, 41(20), 9438-9460.
9. Lin Y., G. E. May, C. J. Mcmanus (2015). Mod-seq: A High-throughput Method for Probing RNA Secondary Structure, Methods Enzymol, 558, 125-152.
10. Lynch V., M. Nnamani, A. Kapusta (2015). Ancient Transposable Elements Transformed the Uterine Regulatory Landscape and Transcriptome during the Evolution of Mammalian Pregnancy, Cell Reports, 10(4), 551-561.
11. Nawrocki E. P. (2014). Annotating Functional RNAs in Genomes Using Infernal, Methods Mol Biol, 1097, 163-197.
12. Nicholas D., F. Magdinier (2018). A Family of Long Intergenic Non-coding RNA Genes in Human Chromosomal Region 22q11.2 Carry a DNA Translocation Breakpoint/AT-rich Sequence, PLoS ONE, 13(4), e0195702.
13. Shao J., J. Zhang, Z. Zhang (2013). Alternative Polyadenylation in Glioblastoma Multiforme and Changes in Predicted RNA Binding Protein Profiles, Omics A Journal of Integrative Biology, 17(3), 136-149.
14. Song J., L. Xu, H. Sun (2014). An Algorithm for Rapid Noncoding RNA Sequence-structure Alignment, Journal of Jiangsu University, 35(1), 69-74.
15. Tabassum R., M. Haseeb, S. Fazal (2016). Structure Prediction of Outer Membrane Protease Protein of *Salmonella typhimurium* Using Computational Techniques, Int J Bioautomation, 20(1), 5-18.
16. Tripathi A. K., M. K. Aparnathi, S. S. Vyavahare (2012). Myostatin Gene Silencing by RNA Interference in Chicken Embryo Fibroblast Cells, Journal of Biotechnology, 160(3), 140-145.
17. Yang J., A. Roy, Y. Zhang (2013). Protein-ligand Binding Site Recognition Using Complementary Binding-specific Substructure Comparison and Sequence Profile Alignment, Bioinformatics, 29(20), 2588-2595.

Xiaodan Liu, M.Sc.
E-mail: gesblee@163.com

Xiaodan Liu has a M.Sc. Degree from the Northwestern Polytechnical University, China. Currently she is working in Shaanxi Radio and Television University. Her research interests are in the field of data mining and data analysis.

Yu Wang, M.Sc.
E-mail: moolan@qq.com

Yu Wang has a M.Sc. Degree in Xidian University, China. Currently he is working in Shaanxi Radio and Television University. His research interests are in the field of educational technology and data analysis.

Na Zhang, M.Sc.
E-mail: 315634944@qq.com

Na Zhang has a M.Sc. Degree of Shaanxi Normal University, China. Currently she is working in Shaanxi Radio and Television University. Her research interests are in the field of data analysis, data mining and machine learning.