AN IMPROVED ALGORITHM FOR RNA SECONDARY STRUCTURE PREDICTION

Oyun-Erdene Namsrai, Kwang Su Jung, Sunshin Kim, Keun Ho Ryu

Database/BioInformatics lab,
School of Electrical & Computer Engineering,
Chungbuk National University
Cheongju, Chungbuk 361-763, Korea
Tel: +82-43-261-2254,Fax: +82-43-275-2254
Email: {oyunerdene, ksjung, sskim04, khryu}@dblab.chungbuk.ac.kr

ABSTRACT:

A ribonucleic acid (RNA) is one of the two types of nucleic acids found in living organisms. An RNA molecule represents a long chain of monomers called nucleotides. The sequence of nucleotides of an RNA molecule constitutes its primary structure, and the pattern of pairing between nucleotides determines the secondary structure of an RNA. Non-coding RNA genes produce transcripts that exert their function without ever producing proteins. Predicting the secondary structure of non-coding RNAs is very important for understanding their functions. We focus on Nussinov's algorithm as useful techniques for predicting RNA secondary structures. We introduce a new traceback matrix and scoring table to improve above algorithm. And the improved algorithm provides better levels of performance than the originals.

KEY WORDS: RNA secondary structure prediction, base pair maximization, dynamic programming algorithm

1. INTRODUCTION

A ribonucleic acid (RNA) is one of the two types of nucleic acids found in living organisms (the other is deoxyribonucleic acid—DNA). An RNA molecule represents a long chain of monomers called nucleotides. RNAs contain four different nucleotides, adenine (A), guanine (G), cytosine (C), and uracil (U). The sequence of nucleotides of an RNA molecule constitutes its primary structure, and the pattern of pairing between nucleotides determines the secondary structure of RNA.

Non-coding RNA genes produce transcripts that exert their function without ever producing proteins. Non-coding RNA gene sequences do not have strong statistical signals, unlike protein coding genes.[10] The RNA structure is essential for its gene function, but interestingly detection of non-coding RNA genes is very hard. However detection of non- coding RNA is can usually not be performed by just considering a predicted structure of a single sequence, predicting the secondary structure of these RNAs is very important for understanding their functions. [3,11,14,19]

In other word, the function of a non-protein-coding RNA is often determined by its structure. Since experimental determination of RNA structure is time-consuming and expensive, its computational prediction is of great interest, and some efficient solutions are known.

2. RELATED WORK

Conventionally, most methods developed so far for predicting the secondary structure of RNAs might be roughly classified into two categories: energy minimization and phylogenetic comparisons.[3,11,14]

The energy minimization technique comprises combinatorial and dynamic programming approaches, and is based on computation of the lowest free energy structure(s) of a sequence. [2,14]

The combinatorial approach first generates all the possible helices of a sequence; and in a second step a branch and bound algorithm combines compatible helices until optimal or suboptimal structures are formed [8,13]. This technique, that exhaustively search the solution space, cannot deal with sequences much longer than 200 nucleotides.

Dynamic programming approach computes the lowest free energy structure by mean of an energy base-pairing optimization based on a recursive relation between the best structures of length k and the best structures of length k-1 [6,7,17,18,19]. Dynamic programming allows to treat sequences containing up to 2000 nucleotides, however such methods method neither consider pseudo-knots nor find sub-optimal solutions.

A more recent method partially solves the last problem [20]. The phylogenetic methods use covariation analysis to identify conserved paired bases among a set of homologous sequences. This is a satisfying procedure that gives excellent results, including pseudo-knots identification. [11,14,19,21]

However the procedure requires a prior alignment of sequences and multiple alignment is, in turn, a difficult problem. One must also quote miscellaneous methods

based either on parallel algorithms, or formal grammar, graph theory and simulation of the RNA folding process.

Recent methods intend both to align RNA sequences and to predict their consensus secondary structure[4,15]. Finally two recent methods have been proposed in order to predict the secondary structure, possibly included pseudoknots, of a single RNA sequence : [16] is based on graph theory approach (Maximum Weighted Matching) while Rivas and Eddy relies on dynamic programming [2,3,9].

3. METHODS

Even though folding of a single sequence in general is not reliable enough for prediction of the structure of a single sequence, the principle of these algorithms are used in almost all other algorithms for predicting RNA secondary structure. Here, we focus on Nussinov's algorithm and the SCFG version of Nussinov's algorithm as useful techniques for predicting RNA secondary structures. We introduce a new improved implementation to these algorithms.

The key to the Nussinov algorithm is that it starts off by examining short subsequences, and in any given subsequence, finds the structure that has the most base-pairings. It then recursively builds upon these subsequences. The key to this recursive algorithm is that there is only four ways to add to subsequences to give the longer sequence [14].

Look at how Nussinov's algorithm based on four possible ways to extend an optimal substructure:

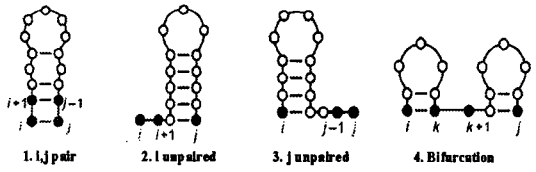

Figure 1. Optimal sub structures.

The Nussinov Algorithm and Nussinov Algorithm using SCFG (Stochastic Context Free Grammar) have some disadvantages. One of the main drawback of the original Nussinov's algorithm is, it considers only the maximum number of base pairs when the algorithm searching for an optimal structure.

Hence, even if the predicted loops are short, the Nussinov's algorithm tends to make base pairs. In reality, since short loops are often thermodynamically unstable structures, we will obtain many incorrect structures.

Another problem with the Nussinov algorithm is that it considers only regular base pairs. RNA base pairs are the canonical Watson-Crick A·U and G·C pairs. Crick proposed, after examining how tRNAs might recognize the genetic code, that G·U is also a valid base pair in RNA secondary structure [1]. The scoring table used by the Nussinov algorithm, however, only counts regular base pairs and equate A-U with G-C.

4. PROPOSED ALGORITHM

To compensate for above two problems and to get more good result when predicting secondary structures, we use the following methods.

Our sub function ***ispair*** begins with checking regular base pairs and it returns true if i-th element paired with j-th element, also it checks given two elements is paired irregularly (G-U). Otherwise our function returns NULL value. Recursive function **traceback** computes traceback starting from position (i, j) and it prints matched pairs to output.txt file when it found. Our **traceback** matrix is logical matrix defined by define preprocessor command and originally we are using tracematrix pointer with integer type. In Nussinov's algorithm usually uses Traceback stack. In our algorithm we defined three values, BASECASE, CASE1, CASE2. We give –3 for BASECASE (in case j next to i), -2 for CASE2 (in case j paired with i) and –1 for CASE1 (in case of unpaired). In case of j paired with k (i<k<j), we call IsPair function with argument k,j and checked fourth condition of Nussinov's algorithm.

Our algorithm input is RNA sequence only, without any comment or additional information and output file contain predicted secondary structure of RNA.

Based on these restrictions we predicted the secondary structure of RNA with the Nussinov algorithm.

5. EXPERIMENTAL REVIEW

5.1 Experimental Setting

We predicted the structure of 15 non-coding RNA sequences taken from various sites of World Wide Web. Average length of these sequences were around 20 bases. For the experiment, we used the original Nussinov, Nussinov algorithm using SCFG and the improved Nussinov algorithm.

5.2 Evaluation Methods and experimental result

Those algorithms were evaluated by measuring accuracy, and shape evaluation. Accuracy is the prediction rate showing whether the position is a base pair or a part of a loop. Accuracy evaluation was having the similar results. The evaluation by the shapes assigns three levels to the result structures. Those are: Perfect match: 5, Having mistakes related to the loop, bulge or hairpin: 7, No match: 3. Evaluation by the shapes indicates that our proposed algorithm able to outperforms the original Nussinov algorithm.

6. CONCLUSION

We presented an improved Nussinov algorithm for the prediction of RNA secondary structure. Our experimental results indicate that this scoring approach is works well.

Nussinov's algorithm is one kind of dynamic programming algorithm. It can't deal with pseudoknots, because pseudoknots violate the recursive definition of the optimal score S(i,j). In the future work we will check

the possibility of new solution for predicting RNA pseudoknotted secondary structure.

ACKNOWLEDGEMENT

This work was supported by the Regional Research Centers Program of Ministry of Education & Human Resources Development in Korea.

REFERENCES

[1] Crick, F.H. Codon–anticodon pairing: the wobble hypothesis. J. Mol.Biol., 19:548–55, 1966.

[2] Eddy, S. R. How do RNA folding algorithms works? Nature BioTechnology Volume 22, Number 11. Nature Publishing Group, November 2004

[3] Eddy, S. R. What is dynamic programming? Nature BioTechnology Volume 22, Number 7. Nature Publishing Group, November,July 2004

[4] Eddy SR, Durbin R: RNA Sequence Analysis Using Covariance Models. Nucl Acids Res 1994, 22:2079-2088.

[5] Gorodkin J, Heyer LJ, Stormo GD: Finding the Most Significant Common Sequence and Structure Motifs in a set of RNA Sequences. Nucl Acids Res 1997, 25:3724-3732.

[6] Nussinov, R. Pieczenk, Eddy, S. R. Griggs, J. R.and Kleitman, D. J. Algorithms for loop matching. SIAM Journal of Applied Mathematics, 35:68–82, 1978.

[7] Nussinov, R. and Jacobson, A. (1980). Fast algorithm for predicting the secondary structure of

single-stranded RNA. Proc Natl Acad Sci USA, 77:6903.

[8] Pipas, J. and McMahon, J. (1975). Method for predicting RNA secondary structure. Proc Natl Acad Sci USA., 72:2017

[9] Rivas, E. and Eddy, S. 1999, A dynamic programming algorithm for RNA structure prediction including pseudoknots, Journal of Molecular Biology, 285, 2053

[10] Rivas,E. and Sean R.Eddy, Noncoding RNA gene detection using comparative sequence analysis, BioMedCentral 2:8, Bioinformatics 2001

[11] Robin D. Dowell. Prepared under the direction of Sean R. Eddy. RNA Structural alignment using stochastic context-free grammars. Ph.D thesis presented to the Sever Institute of Washington University, December, 2004

[12] Samuel, I. Ming-Yang K. Predicting RNA Secondary Structures with ArbitraryPseudoknots by maximizing the Number of Stacking Pairs, Journal of Computational biology, Volume 10, Number 6, 2003;

[13] Sankoff, D. Simultaneous Solution of the RNA Folding, Alignment, and Protosequence Problems. SIAM J Appl Math 1985, 45:810-825.

[14] Stephen McCauley advised by Ian Holmes. An Analysis of the relative efficacy of the Nussinov-Felsenstein, and the Knudsen-Hein, RNA Secondary Structure Prediction, Ph.D thesis presented in October 6, 2003

[15] Tabaska, J. and Stormo, G. Automated alignment of RNA sequences to pseudoknotted structures, Fifth International Conference on Intelligent Systems for Molecular Biology, The AAAI Press, Menlo Park, California (USA), pp. 311(318), 1997.

[16] Tabaska JE, Cary RB, Gabow HN, Stormo GD: An RNA Folding Method Capable of Identifying Pseudoknots and Base Triples. Bioinformatics 1998, 14:691-699.

[17] Waterman, M. S. and Smith, T. F. RNA secondary structure: A complete mathematical analysis," Mathematical Bioscience, Vol. 42, pp. 257:266, 1978

[18] Zuker, M. and Stiegler, P. (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res., 9:133

[19] Zuker, M. and Sankoff, D. RNA secondary structures and their prediction, Mathematical Bioscience, Vol. 46, pp. 591 (621), 1984.

[20] Zuker, M. On Finding All Suboptimal Foldings of an RNA Molecule, Science, 244, 48, 1989.

[21] Zuker, M. Mathews, D. H. and Turner, D. H. Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide in RNA Biochemistry and Biotechnology, ser. NATO ASI Series, J. Barciszewski and B. Clark, Eds. Kluwer Academic Publishers, 1999.