Research Journal of Recent Sciences _____ ISSN 2277-2502
Vol. 3**(IVC-2014),** 20-23 **(2014)** *Res. J. Recent. Sci.*

Comparison of RNA Secondary Structure Prediction Tools in Predicting the Structure

Shaikh S.A. and Trivedi R.A.
Department of Microbiology, Shree Ramkrishna Institute of Computer Education and Applied Sciences, Athwalines, Surat-395001, Gujarat, INDIA

Available online at: www.isca.in, www.isca.me
Received 2rd July 2014, revised 24th August 2014, accepted 10th September 2014

Abstract

Many numbers of software applications (GUIs) are available for the single stranded nucleic acid secondary structure prediction-like Mfold, CONTRA fold, IPknot, Compa RNA, Centroid Alifold, etc. Some uses Minimum Free Energy models (MFE) algorithm and others use stochastic context-free grammars (SCFGs), and rest rely on dynamic programming evolved as an alternative probabilistic methodology for modelling RNA structure. In contrast to physics-based methods, which are dependent on thousands of experimentally-measured thermodynamic parameters, SCFGs require fully-automated statistical learning algorithms to derive model parameters. The performance of 10 single-sequences from a numerous RNA sequences with respective methods were being evaluated. On the whole the most accurate and stable predictions obtained by single-sequence analyses are generated by Mfold, IPknot, RNA Structure and COFOLD.

Keywords: RNA secondary structure, graphical user interface (GUI), minimum free energy (MFE), dynamic programming, IPknot.

Introduction

Apart from the importance of DNA, the equally important molecule is RNA. Because it serves as the bridge between deciphering genetic data and its expression. Noncoding and coding RNAs both play significant role in numerous biological processes[1,2]. From a discovery to up till now the understanding of RNA structure and function have widen the horizon of our knowledge. RNAs are more dynamic in sense of function and mediate many biological processes such as, plays a role in making the reaction faster than usual[3], guiding the post transcriptional modification[4], in regulation of gene[5], in production of target oriented drugs[6,7]. Likewise proteins many types of RNAs are predominantly dependant on the three dimensional structure. Primary structure is an arrangement of nucleotides in linear manner having covalent bonds. When a primary structure folds back a set of canonical base pairing gives rise to secondary structure, and additional folding gives rise to the tertiary structure. 3D Structure of a molecule is mostly determined by its secondary structure which predominantly culminates through base pairing interactions. Secondary structure of RNA can be determined by X-ray crystallography and NMR, but both are time consuming and costlier methods. So to have the better idea of RNA secondary structures they can be predicted by using different thermodynamic models like Turner model[8], stochastic context free grammar and dynamic programming algorithm which projects a structure with minimum free energy (MFE) for a particular RNA.

RNA is having hierarchal arrangement, secondary structure generally can be predicted and analysed before the structure attains its tertiary shape. Mechanism of RNA action can generally be understood by secondary structure prediction as well it aids in design of siRNA and antisense DNA oligonucleotides. In either of the cases it restricts the structure to join with RNA target and also avoids self-folding which prevents hybridization of the structure with that of target.

Secondary structure prediction gives an idea about interacting regions of RNA with proteins[9]. On an average prediction of RNA secondary structure gives an insight into new functional RNA sequences encoded by the genome.

Methodology

Webservers Used for predicting RNA Secondary Structure: As a matter of fact prediction methods are a necessity these days RNAs are being discovered at a faster rate than their structure being resolved. In addition the available techniques as discussed earlier can be applied to a small community of RNAs and the success ratio is uncertain with respect to specific RNA.

Mfold Web server: The method has been developed in late 1980s, where 'm' denotes multiple[10]. The algorithm of a software works by predicting minimum free energy, ΔG, for folding a selected base pair. The server accepts a single RNA or DNA sequence at a time. There are separate forms available for DNA and RNA.

Respected sequence has to be entered in a space provided. For folding of RNA 'T' or't' must be converted to 'U'. Either the sequence of nucleic acid may be circular or linear. By default the software accepts 'linear' but 'circular' sequences can also be considered because it is easier to fold circular than linear ones. The temperature for folding is maintained at 37^0 C.

IPknot: IPknot is for production of RNA secondary structure having pseudoknots which uses integer programming. The method is a mathematical data optimization or feasibility programme in which some or all of the variables are restricted to be integers. For every pseudoknotted structure two steps are mandatory with RNA sequence to be studied. i. Calculation of probabilities of base pairing. ii. For obtaining optimal secondary structure of pseudoknotted RNA solving IP problem is necessary.

MEA-based approach is one of the important aspects to predict RNA secondary structure along with centroid estimation. So the function gain of $\hat{y} \epsilon S(x)$ can be represented as:

$$G_{\gamma(y,\hat{y})} = \gamma TP(y,\hat{y}) + TN(y,\hat{y}) = \sum_{i<j} \gamma I(y_{ij} = 1) I(\hat{y}_{ij} = 1) + I(y_{ij} = 0) I(\hat{y}_{ij} = 0) \quad (1)$$

Where in γ greater than 0 indicates base pair weight parameter, TP and TN are the numbers which show true positives (pairing) and true negatives (non-pairing), and I(condition) shows function that attains a value of either 0 or 1, if a given condition is false or true.

The main focus is to define a secondary structure $\hat{y}$ that shows maximum expected gain function shown in equation 1under the influence of probability distribution of secondary structures having pseudoknots S(x):

$$E_{y/x}[G_\gamma(y,\hat{y})] = \sum_{y \in S(x)} G_\gamma(y,\hat{y}) P(y/x) \quad (2)$$

Here P(y/x) indicates pseudo knotted RNA secondary structure's probability distribution. With the help of γ-centroid estimator secondary structure can be decoded by specific probability distribution.

RNA Structure: Mathews and co-workers had developed RNA Structure programme, which uses dynamic programming algorithm. The software converts chemical modification constraints into the dynamic programming algorithm by which it is able to minimize free energy. Here in this programme both the parameters as chemical modification as well as free energy minimization are taken into consideration so the software runs better than others which rely on only free energy minimization schemes. Following table-1 shows a list of programmes that are used from and are available with the RNA Structure web server:

Co fold: More than 3 decades of research has been invested into devising methods that take a single RNA sequence and predict the corresponding RNA secondary structure. Co fold explicitly takes into account the effects of co-transcriptional folding. COFOLD does not aim to explicitly simulate the folding pathway, but rather to improve RNA secondary structure prediction by considering the implications of kinetic folding. Structure prediction accuracy is measured on a base pair level. True positives (TP) are correctly predicted base pairs. False positives (FP) are incorrectly predicted base pairs. As a performance matrices True positive rate can be defined as (TPR = 100.TP/ (TP+FN). False Positive Rate (FPR=100.FP/ (FP+TN). False positive rate (FPR = 100. FP / (FP + TN)). Positive predictive value (PPV = 100.TP / (TP+FP))

True positive rate is a measurement of sensitivity and indicates the proportion of reference base pairs that were predicted. False positive rate and positive predictive value are both measurements of specificity.

Results and Discussion

For comparison here in this context a glucose oxidase RNA sequence from the available database of organisms like *Apis mellifera* (European honey bee), *Aspergillus niger* (A Fungus); 4 different varieties were taken into consideration, *Helicoverpa zea* (A corn earworm) and *Magneporthe oryzae*(Rice blast fungus) was taken into consideration along with some other enzyme RNA datasets like alkaline protease (*Aspergillus clavatus*), aspartic protease (Aspergillus *oryzae*), esterase (*Bacillus subtillis*). The actual sequences are round about 1000-2000 nt long but for ease of showing the results 50 nt from each are taken.

As clearly visible from the table.2 that the different structures fold back on its own with a characteristic pattern that plays a significant role in producing the active enzyme. The symbols "((( )))" represent pairing of bases whereas "…" shows unpaired bases which will either create bulged loops and pseudoknots, that can be examined with IPKnot and CoFold.

Table-1

Servers	References	Description
AllSub	11,12	Generate all possible low free energy structures for a nucleic acid sequence.
ct2dot	13	Convert a CT-formatted structure into a dot bracket file.
dot2ct	13	Convert a dot bracket file into a CT file.
draw	13	Draw the secondary structure of a nucleic acid strand, with or without color annotation.
MaxExpect	14	Generate a structure or structures composed of highly probable base pairs.
scorer	15,16	Calculate the sensitivity and positive predictive value (PPV) for a predicted as compared to the accepted structure.

Table-2
COFOLD results of RNA sequence from different organisms

Name of Organism	RNA Sequence (50 nt) CoFold Results
Glucose oxidase	
Apis mellifera-	atggcgatcttaaactcaatgtacaacaacgtatccccgctgcagtgcac .(((((((.....((((....((((....(.(((((.(.((..
A. niger-	CUGCAGGUACCUGAAGCCUGCCUAGUUUGAUCACCCUGAAACCAGCACUGC ..((((((.......)))))).((((....(((((.....(((((...((.
Helicoverpa zea -	UAGGAAAAUACCAAGAUGAUUCUGGCGCAGCAAGAUUGCGGCUGCCAAACAG .....(.........)....(((..(((((.........)))))....)))
Magneporthe oryzae -	ATGCGCTCCTCACCGATTCTGCTGCAGCTCCTGCTCTGCGTCTCCGGCCT .........(((........((.((((...))))...))((((.((((..
Alkaline Protease	
Aspergillus clavatus-	ATGCAGTCCATCAAGCGCACCCTTCTGCTCCTTGGAGCCCTCCTGCCGGC ..(((((((((...((((........))))...))))......)))))....
Aspartic Protease	
Aspergillus oryzae-	GGGGTTTCCGCCCATCCGGTGTCATAAGTGGAGGCAACAATTTCGACTTA (..(((...(((..((((.(......).)))))))...)))..).....

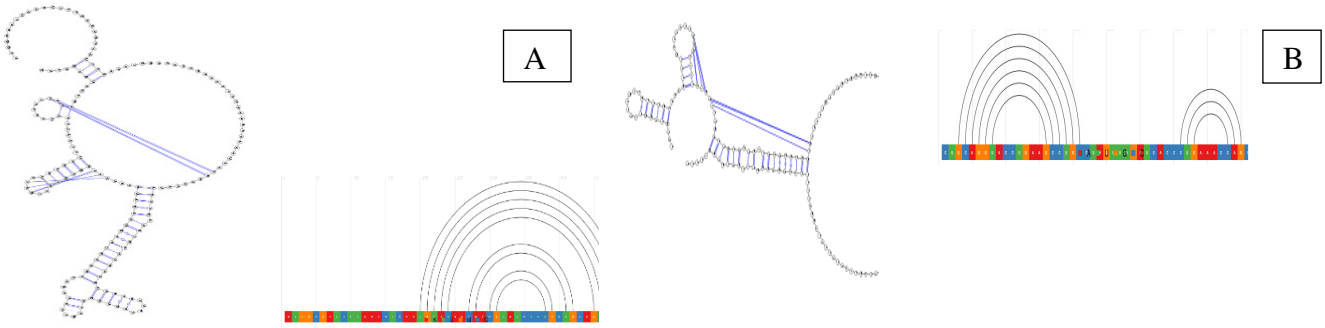

Figure-1
Shows (A) *Apis mellifera* gox folding with IPKnot (blue-2D); structure can also be represented with arc diagrams. (B) *Aspergillus niger* gox folding with IPKnot (blue-2D); along with the arc diagram. Each has got unique minimum free energy (MFE) required for attaining the secondary structure. Below in table.2 are minimum free energy differences are listed

Table-3
MFE obtained through COFOLD, Context Fold and RNA Structure software

Name of Organism	COFOLD Kcal/mol	Context Fold Kcal/mol	RNA Structure Kcal/mol
Glucose oxidase			
Apis mellifera-	-6.00	-1.77	-70.5
A. niger-	-9.60	-5.40	-16.9
Helicoverpa zea -	-11.90	-2.10	-60.6
Magneporthe oryzae -	-13.70	-2.80	-77.9
Alkaline Protease			
Aspergillus clavatus-	-10.10	-5.40	**-**44.0
Aspartic Protease			
Aspergillus oryzae-	-11.70	-8.80	**-**54.6

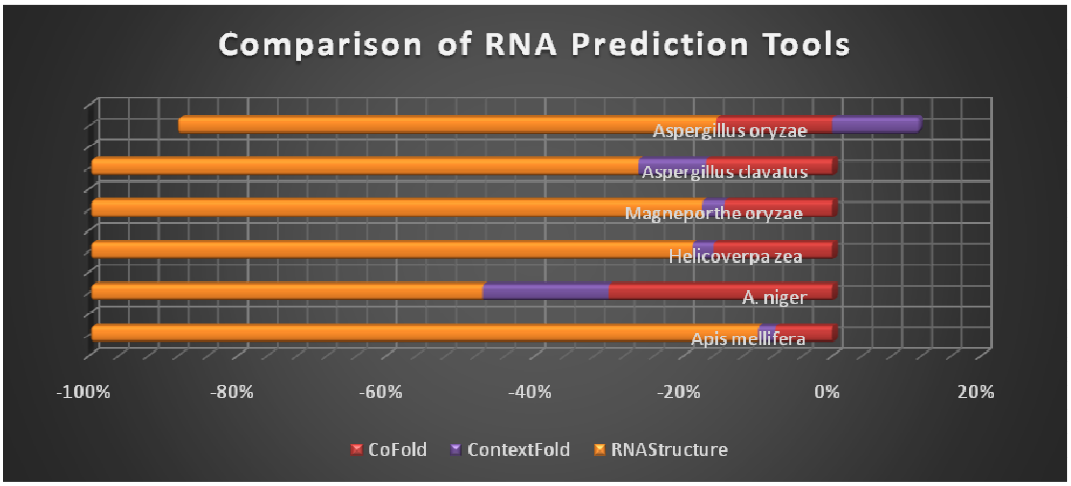

Figure-2
The above chart shows comparison of RNA Secondary structure prediction Tools mainly 3; Co Fold, Context Fold and RNA Structure. Using RNA Structure can give more accurate and reliable results

Conclusion

IPknot predicts a pseudoknotted secondary structure that maximizes the approximate expected gain function, which represents the expectation of the (weighted) number of true predictions of base pairs under a given probability distribution, whereas Compa RNA benchmarks on both data sets offer insight into the relative performance of RNA secondary structure prediction methods on RNAs of different size and with respect to different types of structure. Mfold also gives a vast array of MFE results for Secondary structure prediction. From the three tools compared RNA Structure has evolved to be useful in obtaining various data at a single work station.

References

1. Eddy S.R., Noncoding RNA genes and the modern RNA world, *Nat. Rev. Genet.,* **2,** 919–929 **(2001)**
2. Huttenhofer A. and Schattner P., The principles of guiding by RNA: Chimeric RNA-protein enzymes, *Nat. Rev. Genet*, **7,** 475–482 **(2006)**
3. Doudna J.A. and Cech T.R., The chemical repertoire ofnatural ribozymes, *Nature*, **418,** 222–228 **(2002)**
4. Bachellerie J.P., Cavaille J. and Huttenhofer A., Theexpanding snoRNA world, *Biochimie*, **84,** 775–790 **(2002)**
5. Gong C. and Maquat L.E., lnc RNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements, *Nature*, **470,** 284–288 **(2011)**
6. Sucheck, S.J. and Wong, C.H. RNA as a target for small molecules, *Curr. Opin. Chem. Biol*., **4,** 678–686 **(2000)**
7. Guan L. and Disney M.D., Recent advances in developing small molecules targeting RNA, *ACS Chem. Biol.*, **7,** 73–86 **(2012)**
8. Mathews D., Sabina J., Zuker M. and Turner D., Expanded sequence dependence of thermodynamicparameters improves prediction of RNA secondary structure, *J Mol Biol*, **288(5),** 911–940 **(1999)**
9. Li X., Quon G., Lipshitz H.D. and Morris Q., Predicting in vivo binding sites of RNA-binding proteins using mRNA secondary structure, *RNA*, **16,** 1096–1107 **(2010)**
10. Zuker M., On finding all suboptimal foldings of an RNA molecule, *Science*, **244,** 48–52 **(1989)**
11. Duan S., Mathews D.H. and Turner D.H., Interpreting oligonucleotide microarray data to determine RNA secondary structure: application to the 3' end of Bombyx mori R2 RNA, *Biochemistry*, **45,** 9819–9832 **(2006)**
12. Wuchty S., Fontana W., Hofacker I.L. and Schuster P., Complete suboptimal folding of RNA and the stability of secondary structures, *Biopolymers*, **49,** 145–165 **(1999)**
13. Reuter J.S. and Mathews D.H., RNA structure: software for RNA secondary structure prediction and analysis, *BMC Bioinformatics*, **11**, 129 **(2010)**
14. Lu Z.J., Gloor J.W. and Mathews D.H., Improved RNA secondary structure prediction by maximizing expected pair accuracy, *RNA*, **15,** 1805–1813 **(2009)**
15. Piekna-Przybylska D., DiChiacchio L., Mathews D.H. and Bambara R.A., A sequence similar to tRNA3Lys gene is embedded in HIV-1 U3/R and pro promotes minus strand transfer, *Nat. Struct. Mol. Biol.,* **17,** 83–89 **(2009)**
16. Mathews D.H., Using an RNA secondary structure partition functions to determine confidence in base pairs predicted by free energy minimization, *RNA*, **10,** 1178–1190 **(2004)**