Prediction of RNA Secondary Structure from Random Sequences using ZEM

Cinita Mary Mathew
Dept. Of Computer Science
Amrita Vishwa Vidyapeetham

Meera Krishna G H
Dept. Of Computer Science
Amrita Vishwa Vidyapeetham

G Gopakumar
Dept. Of Computer Science
Amrita Vishwa Vidyapeetham

ABSTRACT

The biological role of many RNA crucially depends on their structure. The in depth understanding of the secondary structure of RNA would provide a better insight in to their functionality. Predicting secondary structure of RNA is the most important factor in determining its 3d structure and functions. This work proposes a model for exploring the features of a number of RNA sequences simultaneously so that comparison of sequences can be made and relevant sequences can be identified. The proposed model accepts RNA sequences in any valid biological file format. For each given sequence, required number of random sequences is generated. The generated sequences should have the same base composition as that of original sequence. ZEM (Zuker's Energy Minimization) Algorithm finds the biologically correct structure of each RNA sequence and its corresponding free energy value. The proposed prototype enables us to experiment with a number of RNA sequences and to study their features so that biologically relevant inferences can be made. An important area where it finds application is in the design of pharmaceutical products.

Keywords

RNA, Secondary structure, ZEM, Traceback Algorithm

1. INTRODUCTION

The central dogma of molecular biology states that the genetic information of an organism is transferred from Deoxyribonucleic Acid (DNA) to Ribonucleic Acid (RNA) and then to Proteins[2]. For a long time DNA was considered as the primary actor in storing the genetic code while RNA had the secondary role of carrier of this information. But a string of discoveries in the last decade have proved that RNA molecules play an important role in controlling cell activities. The knowledge about RNA is expanding rapidly. It is now known that RNA catalyzes reactions, directs the site-specific modification of RNA nucleotides, modulates protein expression and serves in protein localization. Therefore, understanding the function of RNA molecules is a key to unlocking the pathways of disease and biology. These functions can be predicted through the secondary structure of RNA.

RNA is single stranded and it can attain some unique shapes. RNA has three structures which are primary, secondary and tertiary. Primary structure is a linear sequence of bases A G C and U. In secondary structure RNA sequence folds back to itself and bases of RNA form hydrogen bonds. Finally tertiary structure is a complex three dimensional structure. RNA sequence folds back on itself, and complementary bases pair together forming intricate secondary structure. For a given RNA strand, a secondary structure describes which bases are paired. Specifically, the secondary structure of a strand of length n is a set of pairs (i , j) where (i , j) represents a pairing between the i^{th} and j^{th} bases in the strand where the bases in the strand are indexed from 1 to n starting at the 5' end.

In a secondary structure, each base has at most one partner. Base pairs are most often found stacked onto other base pairs in substructures called stems or helices. Stem occurs when two regions of the same molecule, usually palindromes in nucleotide sequence, base-pair to form a double helix that ends in an unpaired loop. The resulting lollipop-shaped structure is a key building block of many RNA secondary structures. Sometimes, unpaired bases are interspersed in stems which are known as internal loops or bulges. When unpaired bases appear only on one side of the structure, the loop is called a bulge loop. When unpaired bases appear on both sides of the structure, the resulting loop is an interior loop. Loops occurring at the ends of stems are called hairpins, and loops from which more than two stems originate are known as multi-branched loops, or simply multiloops[1, 3]. Researches are going in the field of RNA secondary structure prediction to develop efficient algorithms for the same.

2. LITERATURE SURVEY

It is very difficult to compute RNA secondary structure using physical, chemical and biological methods. All computational prediction methods involve some kind of modeling to search for the most stable or possible secondary structure. The criteria for modeling include scores, energy and probabilities. Information obtained from comparative sequence analysis was used which was based on the fact that same family of RNA share a similar secondary structure. One of the most popular dynamic programming algorithm that was used for secondary structure prediction was Nussinov algorithm[6,7,8].

The Nussinov algorithm tries to maximize the possible number of base pairs of a given sequence[10]. The underlying assumption is that the more base pairs there are in a structure the more stable and more likely the structure is [9]. The algorithm takes advantage of the fact that the optimization problem can be solved by breaking it down into smaller sub problems and solving them[9,13]. The recursive solution to the problem is to calculate the best structure for a subsequence S[i …j] with $1 \leq i \leq n$ and $i < j \leq n$ where n is the length of S. The result of the maximum number of pairs is stored in a two-dimensional matrix M at $M[i,j]$. There are four different cases which can occur during the calculation of $M[i,j]$[13] which are

1. An unpaired base i together with the best structure for the smaller subsequence S$[i+1 \ldots j]$ give the best result.

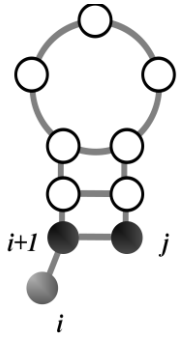

Fig 1. i unpaired structure

2. An unpaired base j together with the best structure for the smaller subsequence S$[i \ldots j-1]$ gives the best result.

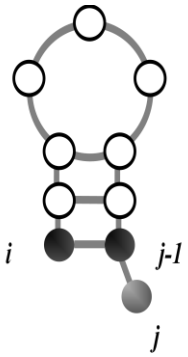

Fig 2. j unpaired structure

3. A base pair (i,j) together with the best structure for the smaller subsequence S$[i+1 \ldots j-1]$ gives the best result. In this case the new base pair (i,j) is added with the score M$[i,j]$ to M$[i+1,j-1]$ receiving the maximum number of base pairs M$[i,j]$.

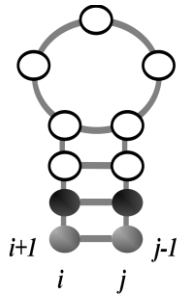

Fig 3. paired structure

4. The combination of the best two structures for the smaller subsequences S$[i \ldots k]$ and S$[k+1 \ldots j]$ gives the best result. In this case k has to be found such that M$[i,k]$ + M$[k+1,j]$ is maximal and then the maximum number of base pairs M$[i,j]$ is M$[i,k]$ + M$[k+1,j]$.

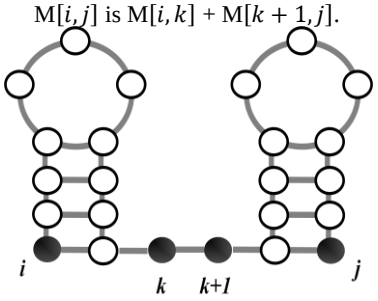

Fig 4. Multiloop structure

The case M$[i,i]$ as well as the case M$[i,i-1]$ have to be made impossible i.e

M$[i,i-1]$= 0 , i = 2 to n.

M$[i,i]$= 0 , i = 1 to n.

M$[i,j]$ = max { M$[i+1,j]$;

M$[i,j-1]$;

M$[i+1,j-1]$ + M$[i,j]$;

$max_{i<k<j}$ M$[i,k]$+ M$[k+1,j]$

}

After the matrix has been filled the solution can be received via backtracking[9]. Beginning at the maximum number of base pairs for the whole sequence, which is stored in M$[1,n]$, one just needs to trace back from which of the four possible cells the maximum has been calculated. Then the subsequences are known and for them again backtracking are done. Like this the base pairs of the optimal folding can be recovered[9]. The overall time complexity of the recursion is O(n^3) and the complexity of space is O(n^2)[13].

Even though Nussinov algorithm is a simplistic approach, it does not give accurate structure prediction i.e. nearest neighbor interactions and stacking interactions are not considered here. It also will not necessarily lead to the most stable structure because of the presence of many interior loops and hairpins which are energetically unfavorable. The basic dynamic programming algorithm was then modified to calculate minimum free energy of the structure. ZEM Algorithm was therefore used which takes into account energy of the structure. Random number of sequences is generated for a particular primary sequence of RNA. Energy corresponding to each and every sequence is obtained using ZEM Algorithm. The secondary structure corresponding to the minimum energy is the most appropriate stable structure of RNA.

3. ARCHITECTURE

The various modules of the system are Process Input, Random Sequence Generation, Energy Matrix Formation, Traceback and Configuration module.

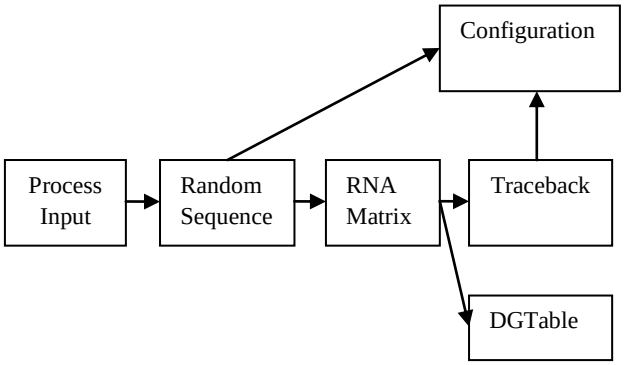

Fig 5. Architecture of RNA Explorer

Process Input module processes input data. It is mainly used to identify the sequence type i.e. whether it is fasta file format or raw format. Random Sequence generation module uses result of Process Input and generates random sequences. This result is used by Energy Matrix Formation module to fill an energy matrix as part of secondary structure prediction. Traceback module performs traceback operation on the energy matrix and base pairs are obtained. Finally the Configuration module draws the secondary structure using the outputs of Random Sequence Generation module and Traceback module. dGTable take the minimum dG computed at RNA matrix. It analysis it and generates dG table.

4. IMPLEMENTATION

The system for prediction of secondary structure of RNA has been developed in JAVA platform using Netbeans IDE 6.9.1. The overall functionality of the proposed system, RNA Explorer is to accept RNA sequences and process them with the intent of exploring RNA features.

4.1 Processing Input

RNA Explorer allows users to enter RNA sequences in two ways. They can either enter a sequence within the text area provided, or can browse for sequences that are already stored. Once the input sequences are received, they are converted to a format that is suitable for further processing. There are two types of input sequences which are in fasta and raw format. Fasta format consists of a line of description followed by the sequence. Raw format consists of the sequence alone.

Fasta Format

>AB000263 |acc=AB000263|descr=Homo sapiens mRNA for prepro cortistatin like peptide, complete cds.|len=368

ACAAGAUGCCAUUGUCCCCCGGCCUCCUGUGCUGC
UGCUCUCCGGGGCCACGGCCACCGCUGCCCUGCCCC
UGGAGGGUGGCCCCACCGGCCGAGACAGCUUUAAU
UACAGACCUGAA

Raw format

CCUGACUUUCCUCGCUUGGUGGUUUGAGUGGACCU
CCAGGCCAGUGCCGGGCCCCUCAUAGAGAGGAAGC
UCGGGAGGUGGCCAGGCGGCAGGAAGGCGCACCCC
CCCAGCAAUCCGCGCGCCGGGACAGAAUGCC

If input sequence is in fasta format, comments are removed. Sequences are checked to see whether there are any invalid characters. If there is any invalid character, user is asked whether to continue or not. If user wants to continue, invalid characters are removed. Else the erroneous sequence is omitted. Then each valid input sequence is written in to a file. New line character is used as delimiter to separates two sequences.

4.2 Random Sequence Generation

After processing the input RNA sequences, next step is to generate random sequences. The produced sequences should have the same base composition as of original RNA sequence. User specifies the number of random sequences to be generated. Validation is performed to ensure that entered random sequence count is valid and it lies within the allowable range i.e. 0 to 1000.

4.3 Secondary Structure Prediction

Secondary structure of each RNA sequence is found using ZEM Algorithm. It calculates the minimum free energy and the base pairs of each sequence.

ZEM Algorithm

ZEM states the secondary structure prediction problem as

E = min E(S) and

E(S) = $\sum e(r_i, r_j)$ $i, j \in$ S

where S - Ranges over all possible secondary structures of an RNA sequence

E(S) - energy of entire RNA structure

$e(r_i, r_j)$ - a function that assigns energy to base pair (r_i, r_j). If r_i and r_j does not form base pairs, energy is set as infinity.

r_i - nucleotide at position i

r_j - nucleotide at position j and

$0 \leq i < j <$ RNA sequence length

Reasonable values of e at 37^0C are -3, -2 and -1 kcal/mole for GC, AU and GU base pairs, respectively.

The secondary structure prediction problem is solved in two steps.

4.3.1. Finding Auxiliary Numbers $E_{i,j}$

In the first step of Zukers algorithm, auxiliary numbers $E_{i,j}$ are computed for all fragments, $r_i \ldots r_j$ of the RNA.

$E_{i,j}$ = 0 for $j - i <$4 otherwise

$E_{i,j}$ = min{ $E_{i+1,j}$;

$E_{i,j-1}$;

e$(i,j) + E_{i+1,j-1}$;

$min_{k=i+1}^{j-1}(E_{i,k} + E_{k+1,j})$ }

That is

- Fragments of length $\leq$ 4 have 0 folding energy, since they cannot fold. Otherwise,
- r_i is unpaired, or
- r_j is unpaired, or
- r_i and r_j pair with each other, or
- r_i and r_j both pair, but not with each other. In this case, r_i pairs with r_{k1} and r_j pairs with r_{k2}, where $i < k1 < k2 < j$. The k in the recursion can be any integer satisfying $k1 \leq k < k2$.

4.3.2. Trace Back Algorithm

A folding with minimum energy is computed using a traceback algorithm.

Initialization: Set $i = 1$ and $j = n$. Put i and j on to the traceback stack.

Steps

1. If the traceback stack is empty, the traceback terminates. Otherwise, take i and j from the traceback stack.
2. If $E_{i+1,j} = E_{i,j}$ then i is not paired.
   (a) If $j - i > 3$, set $i = i + 1$ and continue with 2.
   (b) If $j - i \leq 3$, continue with 1.

   Otherwise, continue with 3.
3. If $E_{i,j-1} = E_{i,j}$ then j is not paired.
   (a) If $j - i > 3$, set $j = j - 1$ and continue with 3.
   (b) If $j - i \leq 3$, continue with 1.

   Otherwise, continue with 4.
4. If $E_{i,j} = e(i,j) + E_{i+1,j-1}$ then r_i pairs with r_j
   (a) Add i,j to the list of base pairs,
   (b) set $i = i + 1$ and $j = j - 1$ and continue with 2

   Otherwise, continue with 5.
5. If $E_{i,j} = E_{i,k} + E_{k+1,j}$ for some k in (i,j), put the fragment $k+1,j$ on to the traceback stack (i is $k + 1$ and j is the current j) and deal with fragment $i \ldots k$ by setting j = k and continue with 2. (Note that some k must exist if this point is reached).

In the first step of Zukers energy minimization algorithm an energy matrix showing the dG values of all fragments in the given RNA sequence is obtained. The dG value of the largest fragment in the energy matrix is the minimum dG value of an RNA sequence. This minimum dG value is stored for performing further analysis on it.

From the trace back algorithm, base pairs contained in the given RNA sequence are found. These base pairs are used for drawing the secondary structure image.

5. PERFORMANCE ANALYSIS

RNAExplorer is the tool that we have designed that predicts the secondary structure of RNA sequences. The correctness of the prediction was checked in order to analyze its efficiency. Vienna RNA Package is a standard tool used for the same and hence it was used to compare the performance. The Package consists of a few standalone programs and a library. Of all the various programs in Vienna RNA Package, we use RNAfold and RNAeval. RNAfold is used to calculate the secondary structures of RNA's. It reads RNA sequences from stdin, calculates the minimum free energy structure (mfe) and prints to stdout the mfe structure in bracket notation and its free energy. RNAeval calculates energy of RNA sequences on given secondary structure. Sequences and structures are read alternatively from stdin. The energy in kcal/mol is written to stdout.

Comparison of basepairs with RNAExplorer and Vienna RNA Package

Base pairs have a big contribution in the stability of RNA. The more the number of base pairs in the RNA structure, more stable the structure is. Various sequence differing in their lengths were given as input to both tools and the number of base pairs obtained were analysed.

Table 1. Base Pair Comparison

Package	Sequence Length				
	16	20	22	24	32
RNA Explorer	3	7	6	8	14
Vienna RNA	2	6	3	6	14

The above table shows that the number of base pairs obtained using our tool RNAExplorer was slightly more than the standard Vienna RNA tool.

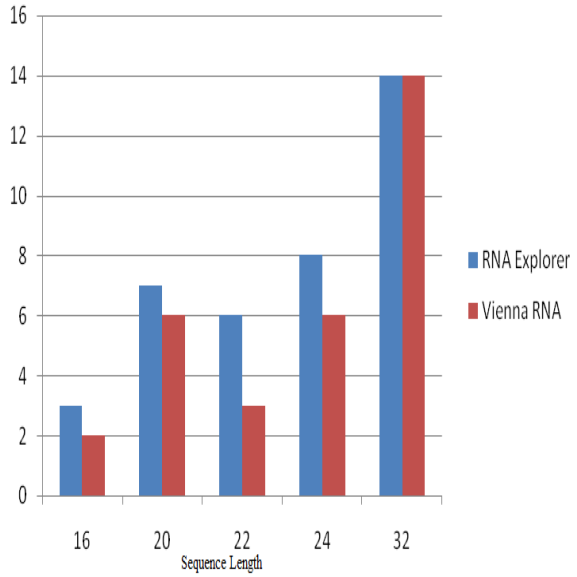

Fig 6. Comparison of Base Pairs

Fig 6 depicts that RNA Explorer gives more number of base pairs for an RNA Secondary Structure compared to RNAfold program in Vienna RNA package. Hence the structure predicted by RNA Explorer is more stable than that predicted by Vienna RNA package. However the base pairs generated by RNA Explorer shows slight variation with that of RNAfold.

SequenceTaken : CCAAGGUUCUAGAAGGCCUU

Base pairs generated by RNA Explorer

Bracket Notation**: . . ((((. (((. . .)))))))**

(3,20) ; (4,19) ; (5,18) ; (6,17) ; (8,16) ; (9,15) ;

(10,14)

Base pairs generated by RNAfold in Vienna RNA

Bracket Notation**: . . ((((((. . . . .))))))**

(3,20) ; (4,19) ; (5,18) ; (6,17) ; (7,16) ; (8,15)

True positive (TP) represents the number of base pairs for both the known and predicted structure. False positive (FP) is the total of predicted base pairs which are not found in the known structure. False negative (FN) indicates the base pairs in the known structures which have not been predicted.

Table 2. Performance Analysis

Sequence length	True Positive(TP)	False Positive(FP)	False Negative(FN)
16	1	2	1
20	4	3	2
22	1	5	2
24	6	2	0
32	14	0	0

6. CONCLUSION AND FUTURE ENHANCEMENT

RNAExplorer is a tool for analysing RNA sequences, extracting relevant information, and presenting it in a way that is useful for the Bioinformaticians. This tool becomes very useful for researchers and other people who are working in this field. Using the random sequence generation, a number of RNA sequences for each input sequence was generated. Secondary structure corresponding to each random sequence was obtained. The most stable structure was the one with the minimum energy. RNAExplorer can be effectively utilized in the design of pharmaceutical products. It effectively reduces the time, effort and cost required for designing drug manually. Once the secondary structure of RNA is known, we can find out the common structural motifs through genetic programming as it has a major role in the functionality of RNA and can be used to find out the common structure of a family of RNA Sequences.

7. ACKNOWLEDGEMENTS

We would like to record our gratitude to Dr. Ramachandra Kaimal of Computer Science Department, Amrita School of Engineering for his motivation and direction towards preperation of this paper. We would also like to express our gratitude to Amrita School of Engineering, Computer Science Department for providing us with facilities to complete our project.

8. REFERENCES

[1] Mirela Andronescu et.al "A New Algorithm for RNA Secondary Structure Design", Journal of Molecular Biology, Volume 336,Isuue3 20 February 2004

[2] Amy C. Anderson **"**The Process of Structure-Based Drug Design", Chemistry & Biology, Vol. 10, 787–797, September, 2003

[3] Rune B. Lyngs Michael Zuker, Christian N.S. Pedersen "Internal loops in RNA secondary structure prediction", Proceedings of the 3th Annual nternational Conference on Computational Molecular Biology (RECOMB)

[4] M. Zuker, D.H. Mathews & D.H. Turner, "Algorithms and Thermodynamics for RNA Secondary Structure Prediction: A Practical Guide". In RNA Biochemistry and Biotechnology, J. Barciszewski & B.F.C. Clark, eds., NATO ASI Series, Kluwer Academic Publishers, (1999)

[5] M. Zuker "Prediction of RNA Secondary Structure by Energy Minimization". in Computer Analysis of Sequence Data A.M. Griffin and H.G. Griffin eds.Methods in Molecular Biology, Humana Press Inc., 267-294 (1994).

[6] Ruth Nussinov and Ann B. Jacobson, "Fast algorithm for predicting the secondary structure of single stranded RNA". Proc. Natl. Acad.Sci. USA, 77(11):6309

[7] N Gautham , "BioInformatics Databases and Algorithms", Narosa Publishing House.

[8] R. Durbin, S. Eddy "Biological Sequence Analysis", Cambridge University Press Inc.

[9] Lisa Yu, "Study of RNA Secondary Structure Prediction Algorithms".

[10] Rivas E., Eddy S.R., "A Dynamic Programming Algorithm for RNA Structure.Prediction Including Pseudoknots", Academic Press, 1999.

[11] RNA structure determination Experimental techniques & Computational prediction [PDF document]. Retrieved from http://www.ibi.vu.nl/teaching/masters/prot_struc/2008/ps-lec12-2008.pdf

[12] D. Huson, "Algorithms in Bioinformatics I", WS'06, ZBIT, December 6, 2006

[13] The web interface to the RNAfold program can be found at http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi