International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622
International Conference on Industrial Automation and Computing (ICIAC) (12th&13th April 2014)

RESEARCH ARTICLE OPEN ACCESS

Secondary Structure Predictions for Long RNA Sequences

Mrs. Ujjwala H. Mandekar*, Ms. Pooja B. Aher**
*(Assistant professor,Computer Science & Engineering, Nagpur University, Nagpur-09
Email:ujjwalaaher@gmail.com)
*(Assistant professor,Computer Science & Engineering, Nagpur University, Nagpur-09
Email:poojabaher786@gmail.com)

ABSTRACT
The RNA (RiboNucleic Acid) is a biological polymer with sugar-phosphate backbone as ribose. It usually found as a single strand and contains the base Uracil(U) , Adenine (A), Cytosine (C) and Guanine (G). RNA bases can bond and form pairs. The canonical pairs are A-U, G-C, G-U, where A-U, G-U are based on two hydrogen bonds and G-C is based on three hydrogen bonds. There are many methods to predict the secondary structure of an RNA molecule like dynamic programming, greedy programming etc. However, the dynamic programming approach usually takes more time. Thus, it is not very practical to solve the problem of long sequences with dynamic programming. Greedy programming does not guarantee the correctness. RPGA (RNA Sequence prediction by the Genetic Algorithm) is a genetic algorithm to align two similar sequences where the structure of one of them, the master sequence, is known and the other (slave sequence) is unknown. It is possible to predict the structure of an RNA molecule by analyzing several homologous sequence alignments. Here a new operation in RPGA is added to mutate the residues of the base pairs in the master sequence and then realign the two sequences again.
Keywords —computational biology, RNA, secondary structure, alignment, genetic algorithm

I. INTRODUCTION

The RNA (RiboNucleic Acid) is a biological polymer with sugar-phosphate backbone as ribose. It usually found as a single strand and contains the base Uracil(U) , Adenine (A), Cytosine (C) and Guanine (G). RNA bases can bond and form pairs. The canonical pairs are A-U, G-C, G-U, where A-U, G-U are based on two hydrogen bonds and G-C is based on three hydrogen bonds (as seen in Figure 1). It plays important roles in many biological processes including gene expression and regulation. Out of these three bonds G-U is highly unstable (referred to as a "wobble" pair) and thus quite rare.

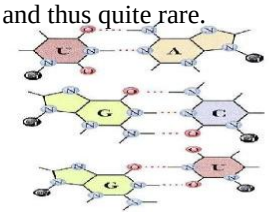
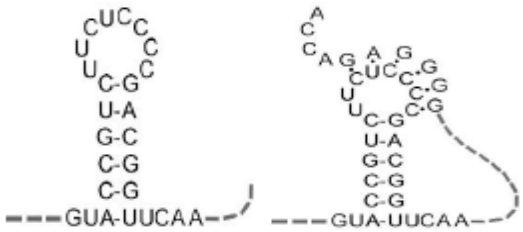


Fig 1: The three possible base pairs form two or three hydrogen bonds [29]

Secondary structural elements in RNA are crucial to their functionality and can be separated into stem loops and pseudoknots (see Figure 2). In both elements, it is well known that an adenine binds with a uracil and a cytosine binds with a guanine. Any stem-loop or pseudoknot contains an inversion, which is a string of nucleotides followed closely by its inverse complementary sequence.



(a) Stem-loop (b) Pseudoknot

Fig2: Two basic elements in RNA secondary structures.[30]

Figure 3 shows an example of an inversion, with the 6-nucleotide string "ACCGCA" followed by its inverse complementary sequence "UGCGGU" after a gap of 3 nucleotides.

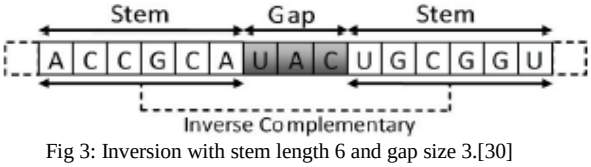


Fig 3: Inversion with stem length 6 and gap size 3.[30]

A. Types of RNA

There are several types of RNA, each with its own structural and functional characteristics:

- Messenger RNA (mRNA): Encodes for the primary sequence of a protein through the genetic code.
- Transfer RNA (tRNA): Binds an amino acid with its anti-sense codon in order to perform translation at the ribosome.
- Ribosomal RNA (rRNA): Binds with proteins to form the ribosome - a complex which translates an mRNA strand to the suitable protein.
- Small Nuclear RNA (snRNA): Short RNA sequences which perform ”maintenance” on RNA (such as splicing and regulation of transcription factors) within the nucleus.
- MicroRNA (miRNA): Short RNAs which can inhibit the translation of mRNA or increase its RNA structure degradation.

B. RNA Structure

- Primary structure - the sequence of RNA bases
- Secondary structure - a two dimensional folding containing an annotation of which base pairs are formed.
- Tertiary structure - a three dimensional folding containing a base sequence with base pair annotation and we describe the spatial location of every atom

II. RNASECONDARY STRUCTURE PREDICTION

Most secondary structure prediction algorithms are based on the minimization of a free energy (MFE) function and the search for a thermodynamically most stable structure starts from the whole RNA sequence. The search for a structure with global minimal free energy may be memory and time demanding, especially for long sequences and for pseudoknot predictions. At the same time, minimal energy configurations may not be most favorable for carrying out the biological functions of RNA, which often require the RNA to react and bind with other molecules (e.g., RNA binding proteins). Our current work suggests that local structures formed by pairings among nucleotides in close proximity and based on local minimal free energies, rather than the global minimal free energy, may correlate better with the real molecular structure of long RNA sequences. The secondary structure consists of the following base-pairing patterns: Single Strand (unpaired bases), stem, hairpin loop, bulge loop, interior loop, junction (multiloop) and pseudoknot, as shown in Figure 4. It is easy to see that if the primary structure and the stem positions are known, it is easy to determine all the other characteristics of the secondary structure, except for the pseudoknots. Normally, the common belief is that pseudoknots contribute very little to the energy balance of the RNA molecule hence a common practice to ignore their locations when predicting RNA folding.

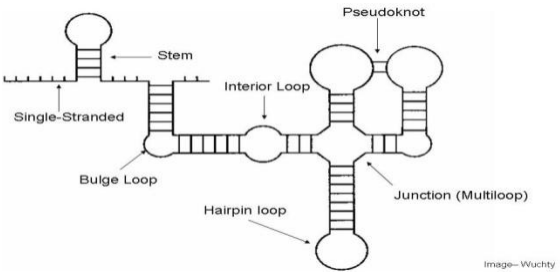


Fig 4: Typical motifs of RNA secondary structure (taken from Wuchty)

III. OVERVIEW OF A GENETIC ALGORITHM

A genetic algorithm has two functions, recombination (or crossover) and mutation. There are also other aspects of a genetic algorithm, which are described below.

(i) Representation of an individual gene: There are many ways of achieving this, such as a discrete representation, real-valued representation, and order based representation.

(ii) Evaluating fitness: In order to determine if some gene represents a good solution, there should be a way of determining the fitness of a gene. Using a fitness function suitable parents can be selected.

(iii) Mutation operators: It must be ensured that a population stays diverse, so some mutation operators are to be applied to the genes. It is obvious that these operators depend on the type of representation of a gene.

(iv) Crossover operator: Creating new genes from a set of genes requires an operator that selects two specific genes (called parents) to create one or more new genes (called children).

(v) Selection of suitable parents: By using a method called fitness proportionate selection, it is possible to determine which genes are most suitable for reproduction and which genes should be excluded from the next generation.

(vi) Stopping criteria: A genetic algorithm can in principle run indefinitely, so there must be some stopping criterion (e.g. due to a limit on computing resources). Detecting convergence of a genetic algorithm can be accomplished by monitoring the gene that represents the best solution of each generation and checking whether it has changed significantly during previous generations.

IV. RNA SECONDARY PREDICTION

The prediction of RNA structures is very important as well as difficult task in bioinformatics. It is very easy to find the primary structure of RNA

molecules, by the technique of sequencing. But the task is more difficult for secondary and tertiary structure. The secondary structure of RNA sequence is a set S of (ri,rj) over the alphabet {A, C, G, U} satisfying the following criteria [12]:

1. $\forall$ (ri,rj)$\in$S,(ri,rj) $\in${(A,U), (U,A), (G,C), (C,G), (G,U), (U,G)}
2. $1 \le i<j \le |S|$
3. $\forall$ (ri,rj), (ri',rj')$\in$ S, i = i'$\Leftrightarrow$j = j'
4. (ri,rj)$\in$ S $\Rightarrow$ |j - i| $\ge$ 4

There are many methods available to determine the structure of RNA molecules. One of them is an exact method which uses experimental techniques such as nucleic magnetic resonance (NMR) and X-ray crystallography. This approach is long, difficult and expensive. Another method predicts the secondary structure starting from the primary structure by using secondary prediction algorithms [13].

To find the RNA structure it is necessary to calculate energy of molecule. The energy of the molecule is the sum of energies of each pair of bases. The free energy of a structure S is given by following formula:

$$E(S) = \sum a(r_i, r_j)_{(r_i,r_j)\in S}$$

a (ri , r j) is the free energy of the pair (ri,rj). This method presents some limits like the relevance of the energy function and biological assumptions are not always true. Another method of score scheme to assess the precision of the conserved secondary structure information contained within the alignment, is the structural conservation index SCI [9]. It is based on the RNAalifold consensus folding algorithm (MFE) [15, 16] which is based upon the sum of a thermodynamic and a covariance term. The Structural conservation index is computed using the following function:

$$SCI = E_A / E'$$

Where EA is the consensus minimum free energy (MFE) of the alignment and E' is the average of the individual MFEs. The SCI is close to zero if RNAalifold identifies no common RNA structure in the alignment, while a set of perfectly conserved structures has an SCI≈1. An SCI >1 shows that there is a conserved RNA secondary structure which is, in addition, supported by compensatory and/or consistent mutations [9].

The GA for RNA secondary structure prediction is divided in two-phases. First, the genetic algorithm is applied i.e., an initial population is created. In the second step, alignment is refined iteratively in order to improve the quality of the conserved secondary structure information of the alignment. In each generation a selection operation is performed to constitute the mature population. Then, the crossover and mutation operators are applied which allow exploring other solutions, only one type of mutations is selected randomly. The mature population is evaluated using the objective function. The global best solution is then updated if better one is found and the whole process is repeated until having satisfaction of stopping criterions.

V. THE GENETIC APPROACH

In this approach, initially it is required to derive representation scheme which includes the definition of an appropriate representation of potential alignments and the definition of evolutionary operators.

A. Genetic representation of alignment

For multiple RNA structural alignment, it is necessary to map potential solutions into a chromosome representation which canbe easily modifiedby genetic operators. The multiple structural sequence alignment Aln= {S1,S2,…,Sn} is viewed as an alphabetic matrix AM where:

Each line i represents a sequence S_i'. The character "–"denotes a gap.

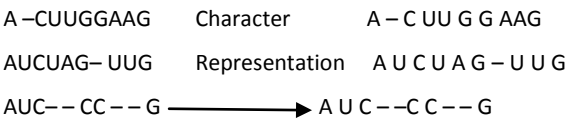


Figure 3: Alphabetic representation of multiple sequence alignment.

B. Population creation

Initial population of chromosomes is generated by encoding the potential sequence alignments. The initial solution is very important and must be significant as a good initial solution can effectively converge faster and consequently cut the computational cost.Hence initial population is created by a progressive alignment method such as ClustalW[19].

C. Selection

Large numbers of selection methods are available having some pron and con. Using Elitism selection [20] it is easy to promote the best individuals of the population, so the best ones will participate in the improvement of population. Elitism method can increase the convergence of genetic algorithm, because it always preserves the best solutions in every generation [20]. But there is the problem of the local optimum.

D. Mutation

Here it is very difficult to place gaps in different RNA sequences. A wrong placement of gaps appears when gap series of the same size occur in different positions, or when an island of characters is surrounded by gaps.

Less significant More significant

GG – – –CAAAU GG– – – CAAAU

AAC– – – CUC– – –UAC AACCUC– – – – – – UAC

To resolve this problem, a simple mutation based on changing randomly the position of gaps can be used. But many times the simple kind of mutation does not improve the solution quality. Hence it is required to use four adaptive mutations. The mutation operators operate on a gap, series of gaps, gap column and gap blocs.

F.1. Gap Mutation

In this type an isolated gap is chosen to a suite of gap in a sequence

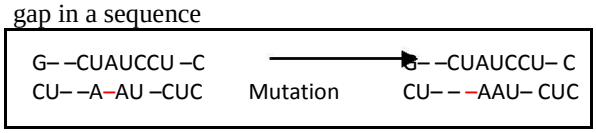


Fig 5: Single gap mutation.

F.2. Gap sequence Mutation

A sequence of gaps is moved to the left or to the right as it's shown in Figure 6.

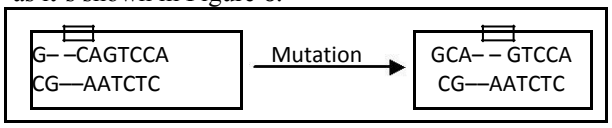


Fig6: Gap series mutation.

F.3. Gap column Mutation

This kind of mutation affects a set of sequences of an alignment. It consists in taking a column of gaps and moves it to the left or to the right (Figure 7).

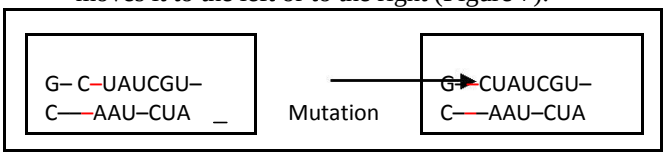


Fig7: Gap column mutation.

F.4. Gap bloc Mutation

A gap bloc is moved left or right. This kind of mutation affects many sequences (Figure 8).

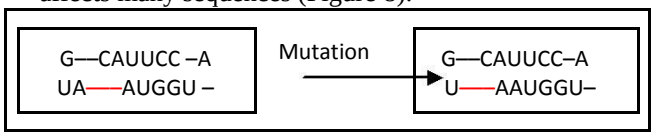


Fig 8: Gap bloc mutation.

E. Crossover operator

For crossover two alignments are taken and then a vertical cut is applied on each alignment. The next step of the crossover operator is to create new individuals by interchanging the parent parts (Figure 9).

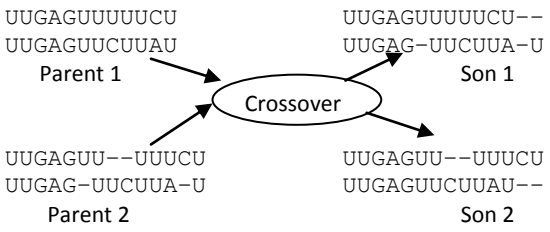


Fig 9: Crossover operator.

F. Fitness evaluation

The fitness function is used to evaluate the alignment quality, and it is the heart of the optimization process. Normally, the objective function is the mathematical tool used to measure the degree to which two or more sequences are similar. For a multiple alignment which is not structurally conserved, the SCI will be near to 0, which predicts that there is no common RNA structures between different sequences. The SCI should be close or greater than 1for an alignment that is structurally conserved. If the alignment is structurally well conserved and compensatory and consistent mutation often occurs, the SCI maybe above 1.

Hence it is possible to use the local search method is the genetic algorithm as follow:

Input:A set of sequences SEQ

(1) Generate population of n chromosomes, POP.

Repeat

Select a subset of the population using the selection operator.

Apply a crossover operation.

Apply a mutation operation.

Evaluate the current population.

If $SCI(Aln_{best}) < SCI(Aln_i)$ then

$Aln_{best} = Aln_i$ and $SCI_{best} = SCI(Aln_i)$.

Apply the replacement operator

Until a termination criterion is reached.

Output: Aln_{best} and $CSI(Aln_{best})$

VI. CONCLUSION

In this paper, an approach to solve the RNA secondary structure prediction problem is explained. The objective of this study is to demonstrate the efficiency of the genetic algorithm and its hybridization with a local search method to deal with the problem at hand. It would be really promising to study this issue as ongoing work.

REFERENCES

[1] Eddy,S. R. 2001. Non-coding RNA genes and the modern RNA world. Nat Rev Genet 2(Dec 2001), 9 19-929.

[2] Mattick,J. S.,Makunin,I. V. 2006. Non-cod ing RNA. Hum Mol Genet 15 Spec N°1, R17(Apr 2006).

[3] Gorodkin,J., Heyer,L., Brunak,S. and Stormo,G. 1997. Displaying the information contents of structural RNA alignments. CABIOS, Vol. 13, 583–586.

[4] Gutell, R., Power, A., Hertz, G., Putz, E. &Stormo, G. 1992. Identifying constraints on the higher-order structure of RNA: continued develop ment and application of comparative sequence analysis methods. Nucleic Acids Res, Vol. 20 (21), 5785–595.

[5] Zuker, M., Jaeger, J. & Turner, D. 1991. A comparison of optimal and suboptimal RN A secondary structures predicted by free energy min imization with structures determined by

phylogenetic comparison. Nucleic Acids Res,Vol. 19 (10),2707–214.

[6] Han,K.-H. andKim,J.-H. 2000. Genetic quantum algorithm and its application to combinat orial optimization problem. Proc. 2000 Congr. Genetic Co mputation, vol. 2, La Jolla, CA, 1354–1360.

[7] Kirkpatrick, C.D. Gelart, a nd P.M. Vecchi. 1983. Optimization by Simulated An nealing. Science, Vol. 220, 671-680.

[8] Balaji, A.N., Jawahar, N. 201 0. A Simulated Annealing Algorithm for a two-stage fixed charge distribution problem of a Supply Chain. International Journal of Operational Research, Vol. 7, No.2, 192 – 215.

[9] Washietl, S., Hofacker, I. and Stadler, P. 2005. Fast and reliable prediction of noncoding RNAs. Proc. Natl Acad. Sci. Vol. 102, 2454–2459.

[10] Gruber, A.R., Bernhart, S.H., H ofacker, I.L., Washietl, S. 2008. Strategies for measuring evolutionary conservation of RNA secondary structures. B MC Bioinformatics 9, 122.

[11] Gardner, P., Wilm, A. and Wash ietl, S. 2005. A benchmark of multiple sequence alignment programs upon structural RNAs. Nucleic Acids Research, Vol. 33(8) 2433–2439.

[12] Sankoff, D. and Kruskal, J. B. 1983.Time warps, string edits, and macromolecules: T he theory and practice of sequence comparison.Addison W esley.

[13] Layeb, A, Meshoul, S., andBatouche, M. 2008. Quantum Genetic Algorithm for Multiple RNA Structural Alignment in the IEEE proceedings of the 2nd Asia International Conference on Modelling& Simulation, pp. 873-877.

[14] Washietl, S. 2010. Sequence and structure analysis of noncoding RNAs. Methods in Molecular Biology, Vol. 609, 285-306.

[15] Washietl, S. and Hofacker,I. 2004. Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparativ e genomics. J. Mol. Biol., Vol. 342, pp. 19–30.

[16] Hofacker, I. L.,Fekete,M., andStadler,P. F. 2002. Secondary structure prediction for aligned RNA sequences. J. Mol. Biol., Vol. 319, 1059–1066.

[17] Hofacker, I. L. 2003. Vienna RNA secondary structure server. Nucleic Acids Res, Vol. 31, 3429-3431.

[18] Okada,Y., Sato, K., and Sakakibara, Y. 2010. Improvement of structure conservation index with centroid estimators. Pacific Symposium on Bio computing, 15:88-97.

[19] Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson,TJ., Higgins, DG., Thompson, JD. 2003. Multiple sequence alignment with the Clustal series of programs.Nucleic Acids Res., Vol. 31, 3497-3500.

[20] Thierens,D.1997.Selection schemes, elitist recombination and selection intensity. in International conference of genetic algorithm, pp. 152-159.

[21] Ziv-Ukelso, M. 2010. A faster algorithm for simultaneous alignment and folding of RNA. Journal of Computational Biology, 17(8), 1051-1065.

[22] Hamada, M., Kiryu, H., Sato, K., Mituyama, T., and Asai, K. 2009. Bioinformatics, 15; 25(4):465-473.

[23] Gesell,T., and Washietl, S. 2008.Dinucleotide controlled null models for comparative RNA gene prediction. BMC Bioinformatics, 9:248.

[24] Tahi, F.,Engelen, S., andRégnier,M. 2003. A Fast Algorithm for RNA Secondary Structure Prediction Including Pseudoknots. Third IEEE Symposium on BioInformatics and BioEngineering (BIBE'03), pp.11.

[25] Engelen, S., andTahi, F. 2010. Tfold: efficient in silico prediction of non-coding RNA secondary structures.Nucleic Acids Res; 38(7):2453-2466.

[26] Engelen, S., andTahi, F. 2007.Predicting RNA secondary structure by the comparative approach: how to select the homologous sequences. BMC Bioinformatics; 8:464.

[27] Washietl, S., Hofacker, I.L., Stadler, P.F.2005. Fast and reliable prediction of noncoding RNAs.ProcNatlAcadSci, 102(7):2454-2459.

[28] Kenza Bouaroudj Abdesslem Layeb Imen Bensetira 2011. A Hybrid Genetic Algorithm for RNA Structural AlignmentInternational Journal of Computer Applications (0975 – 8887)Vol 19– No.7

[29] Computational Genomics. Fall Semester, 2006. Lecture 7: December 5, 2006. Lecturer: Michal Ziv-Ukelson. Scribe: Erez Katzenelson and Ofer Lavi.

[30] Daniel T. Yehdego, Boyu Zhang, Vikram K. R. Kodimala, Kyle L. Johnson, Michela Taufer, Ming-Ying Leung. Secondary Structure Predictions for Long RNA Sequences Based onInversion Excursions and MapReduce.2013 IEEE 27th International Symposium on Parallel & Distributed Processing Workshops andPhD Forum