

Prediction of RNA 3D Structure

Mrs. Ujjwala H. Mandekar^{1*} and Dr. S.P. Khandait²

¹Research Scholar, Priyadarshini Institute of Engineering & Technology, Nagpur, India.

²Professor & Head of Information Technology Department, KDK Engineering College, Nagpur, 440009 India.

email id: spkhandait.kdkce@gmail.com

Corresponding author email id: ujjwalaaher@gmail.com

Date of publication (dd/mm/yyyy): 20/01/2018

Abstract — RNA structures are generally organized hierarchically organized. Prediction of RNA structure is a big challenge for the researchers. Many prediction algorithms have been develop for RNA secondary structure prediction. But prediction of tertiary structure is still a big area of research. Approaches for RNA tertiary structure prediction are broadly categorized into two methods viz. de novo methods and template-based prediction method. The De novo are useful only for the small sequences of the nucleotides. For the sequences with large number of nucleotides template based modelling technique is more useful. These template based method are very common in protein structure prediction but it is not explored in RNA tertiary structure prediction. This paper presents a novel methodology for prediction of RNA tertiary structure using the bank ok known structure (template). First, the RNA secondary structure is analysed on the basis of free nucleotides available region wise which may contribute in the prediction of tertiary structure. Next similar region structure templates are searched in PDB metadata. These templates are then used to for the tertiary structure. As all the templates stored in PDB are predicted structures with minimum negative energy, the resultant tertiary structure is also consequently has minimum negative energy. The proposed method is able to predict even large ribosomal RNA structures. The experimental results have shown for the predicted structures on the basis of percentage similarity as well as time required for prediction.

Keywords — RNA, Base Pair, Kissing Pair, Tertiary Structure.

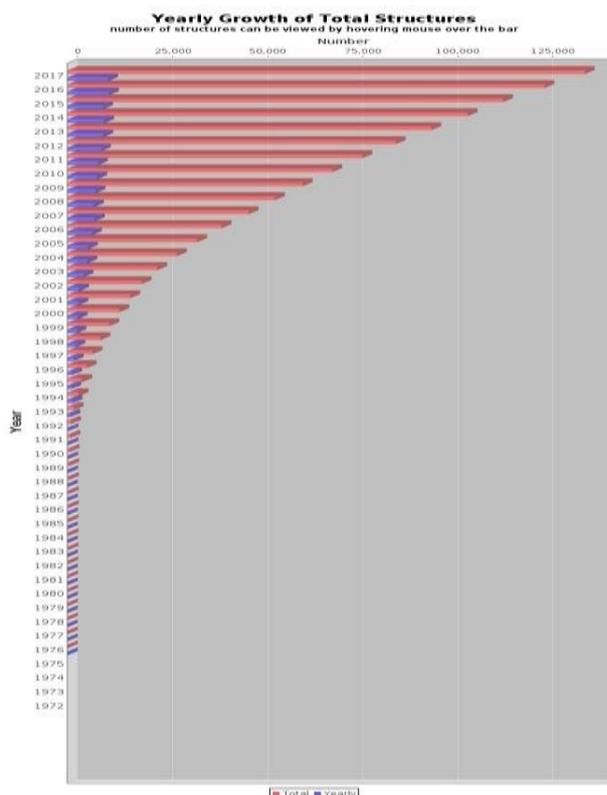
I. INTRODUCTION

Traditionally RNA's biological role is transcription and translation. Beside this, RNA molecules play many important roles in many cellular processes, such as carrying genetic information, participating in the regulation of gene expression, acting as catalysts in many biological pathways [1]. The non-coding RNAs (ncRNAs) play important roles in many cellular activities e.g. Processing of messenger RNAs [5] [6] [7], control of protein stability [8] and RNA interference [9]. They are also concerned in number of human diseases like cancer [10] [11], infectious and neurodegenerative diseases [12] [13] [14], and hence RNA structures are very important in disease detection as well as new remedial agents. Day by day, the number of RNA structures, deposited in the PDB and NDB databases [15] [16], are increasing (graphical representation is as shown in graph below, hence it has been a vital task to develop the computational tools which are helpful in annotation of RNA structures and functions.

Using the primary sequence of RNA, it is possible to predict its secondary structure; consequently from the

secondary structure tertiary structure of the molecule can be predicted [3]. Hence is it essential to analyze the secondary structure in order to know functional characterization of RNA as well as its tertiary structure prediction. For the prediction of the secondary structure Watson-Crick AU and GC base pairs, as well as wobbles pairs are important. These three canonical pairs are the key factors in the RNA folding process. But because of increasing number of known RNA 3D structures, it has become challenge to the researchers to study, in depth, RNA tertiary interactions as well as a variety of other base-base interactions. Generally these base-base interactions are known as non-canonical pairs. From the research study, it has been proved that almost 40 % of all bases in structured RNAs are eligible to take part in non canonical interactions [4].

In this paper, we present an overview of RNA computational models for tertiary structures' predictions and then focus on a recently developed RNA tertiary model. As a lot of investigation has been carried out for secondary structure prediction, the more emphasize is given on the tertiary structure prediction; and hence as an input secondary structure produced be GT fold algorithm is chosen.



II. TERTIARY STRUCTURE PREDICTION

Secondary structure consist of sequence of dots (.) and brackets ('(',')'). Dots indicate free bases whereas brackets indicate paired bases. In kissing (tertiary structure) these

dots (free bases) are more important. Secondary structure consists of hairpin loop bulge loop internal loop, multi loop, stack, single strand and free bases. Except stack all the loops are having free bases which further interact with each other to form kissing pairs. Hence it is important to know number of loops present in secondary structure.

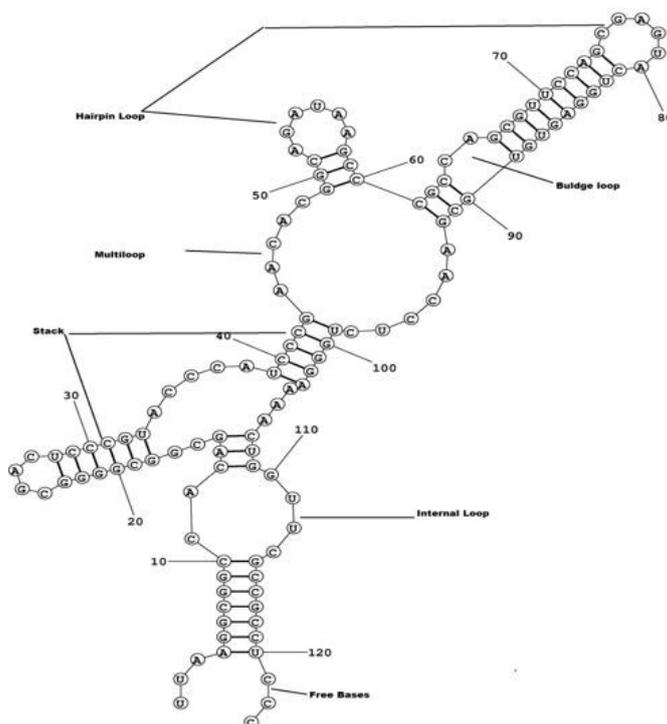


Fig. 1. Different types of loop in secondary structure

After getting the information about number of loops present in secondary structure next step is to find out number of free bases present in each loop. In order to find the free bases present in secondary structure it is divided into different regions, as shown in figure below. The main purpose for splitting structure into regions is to avoid bonding of free bases of same loop. While formation of bonds, region number of free bases are checked. Free bases with different region numbers are allowed to form a pair.

As shown in figure 2 the secondary structure is divided into 7 region i.e. r1 (nucleotides numbered from 1 to 4), r2 (stem has two regions as r2 (a) from 51 to 57 and r2 (b) from 4 to 10), r3 (internal loop region from 10 to 13 and 46 to 51), r4 (stem from 13 to 15 and 44 to 46) etc. From these regions free bases are extracted as volunteers to take part in kissing pair of 3D structure.

1. Algorithm Free Base Count

These free bases can participate in kissing pair formation if they satisfy following conditions.

- Free bases of one loop will form the pairing with free bases of other loop i.e. pair formation within the free bases of the same regions are avoided.
- Free bases can form Random match. In random match free bases of one loop are paired with free bases of two or more loop.

2. Algorithm Kissing-Pair

If we consider the dot sequence of the following secondary structure we get the dot file as shown in figure 4.

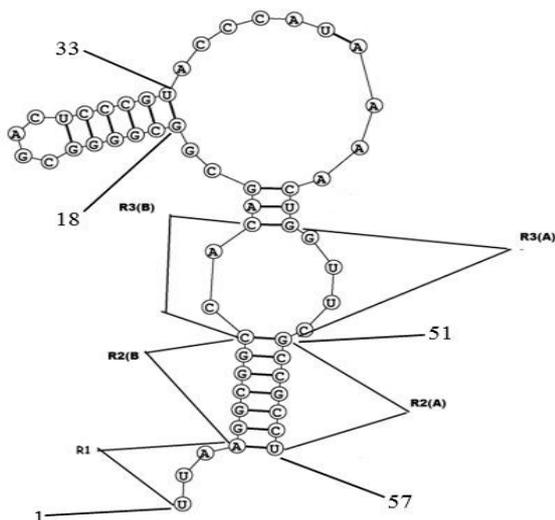


Fig. 2. Secondary structure divided into different regions

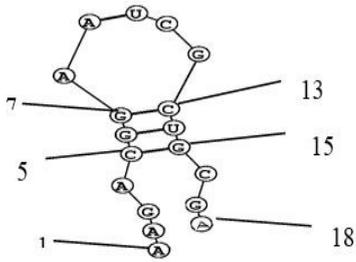


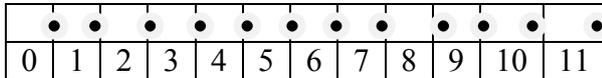
Fig. 3. Sample secondary structure

If regions are formed for the structure as R1(1-5), R2(5-7, 13-15), R3(7-13) and R4(15-18), the dot file becomes

..... (((.....)))

Fig. 4. DOT file of structure

After applying the algorithm free Base Sequence a vector is formed which consists of dots (free bases). Hence dot vector becomes



These free bases forms the data object as shown in table 1; which stores the information as index number, loop to which it belongs and pairable free base index as per Watson-Crick and wobbles pairs. But there is a possibility

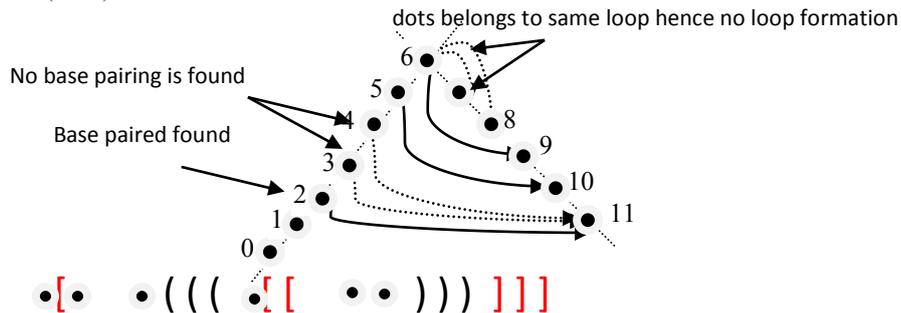
of formation of n^{18} types of structure if 'n' is the total number of nucleotides present in the structure. It is very difficult to select the best structure amongst all possible kissing pair formation. Hence prediction is bases in similarity basis. Concept is to find out the sequence for kissing pair. This dot file sequence is searched in the PDB data bank. PDB data bank consists of more than 13600 sequences. If the dot files matches with any PDB sequence 100%, its corresponding nucleotide sequenced is checked. The sequence, that gives maximum similarity in nucleotides, is selected as most appropriate sequence. Hence algorithm Similarity calculates similarity using the concept of dynamic programming.

Table 1: Contents of data class object of free bases in secondary structure

0	-1	ptr to SingleStrand
1	-1	ptr to SingleStrand
2	-1	ptr to SingleStrand
3	-1	ptr to SingleStrand
4	-1	ptr to hair pin loop
5	-1	ptr to hair pin loop
6	-1	ptr to hair pin loop
7	-1	ptr to hair pin loop
8	-1	ptr to hair pin loop
9	-1	ptr to SingleStrand
10	-1	ptr to SingleStrand
11	-1	ptr to SingleStrand

Algorithm : Similarity

Midpoint of this sequence is 6(= 12/6). Hence there will be formation of two loops; loop1-> mid to 0 (outer) and loop2-> 7 to 11 (inner)

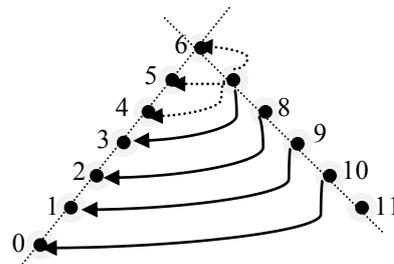


Next step is to find the similarity. Using algorithm Similarity Model and PDB Metadata similarity is calculated.

Now revert the inner and outer loop

loop1-> 7 to 11 (outer) loop1-> mid to 0 (inner)

[[[[[[(((.....))))]]]]]



Now shift mid from 6th position to 5th, 4th, 3rd, 2nd, 1st, 0th position and repeat the same procedure. In next set of iteration mid is shifted to 7th, 8th ...up to last dot position and repeat the same procedure. By the end of all iteration we get the sequence with highest similarity.

The main task of 3D structure is to form PDB file for tertiary structure. A PDB File is a textual file format describing the three-dimensional structures of molecules held in the Protein Data Bank. Protein Data Bank (PDB) format is a standard for files containing atomic coordinates. It is used for structures in the Protein Data Bank and is read and written by many programs. While this short description will suffice for many users, those in need of further details

should consult the definitive description. The complete PDB file specification provides for a wealth of information, including authors, literature references, and the method of structure determination.

PDB format consists of lines of information in a text file. Each line of information in the file is called a record. A PDB file generally contains several different types of records, arranged in a specific order to describe a structure.

Table 2: Contents of Protein data bank
Selected Protein Data Bank Record Types

Record Type	Data Provided by Record
ATOM	atomic coordinate record containing the X,Y,Z orthogonal Å coordinates for atoms in standard residues (amino acids and nucleic acids).
HETATM	atomic coordinate record containing the X,Y,Z orthogonal Å coordinates for atoms in nonstandard residues. Nonstandard residues include inhibitors, cofactors, ions, and solvent. The only functional difference from ATOM records is that HETATM residues are by default not connected to other residues. Note that water residues should be in HETATM records.
TER	indicates the end of a chain of residues. For example, a hemoglobin molecule consists of four subunit chains that are not connected. TER indicates the end of a chain and prevents the display of a connection to the next chain.
HELIX	indicates the location and type (right-handed alpha, etc.) of helices. One record per helix.
SHEET	indicates the location, sense (anti-parallel, etc.) and registration with respect to the previous strand in the sheet (if any) of each strand in the model. One record per strand.
SSBOND	defines disulfide bond linkages between cysteine residues.

Table 3: Protein Data Bank Format
The formats of these record types are given in the tables below.

Protein Data Bank Format: Coordinate Section				
Record Type	Columns	Data	Justification	Data Type
ATOM	1-4	“ATOM”		character
	7-11 [#]	Atom serial number	right	integer
	13-16	Atom name	left*	character
	17	Alternate location indicator		character
	18-20 [§]	Residue name	right	character
	22	Chain identifier		character
	23-26	Residue sequence number	right	integer
	27	Code for insertions of residues		character
	31-38	X orthogonal Å coordinate	right	real (8.3)
	39-46	Y orthogonal Å coordinate	right	real (8.3)
	47-54	Z orthogonal Å coordinate	right	real (8.3)
	55-60	Occupancy	right	real (6.2)
	61-66	Temperature factor	right	real (6.2)
73-76	Segment identifier [¶]	left	character	
77-78	Element symbol	right	character	
79-80	Charge		character	
HETATM	1-6	“HETATM”		character
	7-80	same as ATOM records		
TER	1-3	“TER”		character
	7-11 [#]	Serial number	right	integer
	18-20 [§]	Residue name	right	character
	22	Chain identifier		character
	23-26	Residue sequence number	right	integer
27	Code for insertions of residues		Character	

Older PDB files may not adhere completely to the specifications. Some differences between older and newer files occur in the fields following the temperature factor in ATOM and HETATM records; these fields are omitted from the examples. Some fields are frequently blank, such as the alternate location indicator when an atom does not

have alternate locations. For example, PDB Format of Glucagon which is a small protein of 29 amino acids in a single chain is as given below. The first residue is the amino-terminal amino acid, histidine, which is followed by a serine residue and then a glutamine. The coordinate information (entry 1gen) starts with:

ATOM	1	N	HIS	A	1	49.668	24.248	10.436	1.00	25.00	N
ATOM	2	CA	HIS	A	1	50.197	25.578	10.784	1.00	16.00	C
ATOM	3	C	HIS	A	1	49.169	26.701	10.917	1.00	16.00	C
ATOM	4	O	HIS	A	1	48.241	26.524	11.749	1.00	16.00	O
ATOM	5	CB	HIS	A	1	51.312	26.048	9.843	1.00	16.00	C
ATOM	6	CG	HIS	A	1	50.958	26.068	8.340	1.00	16.00	C
ATOM	7	ND1	HIS	A	1	49.636	26.144	7.860	1.00	16.00	N
ATOM	8	CD2	HIS	A	1	51.797	26.043	7.286	1.00	16.00	C
ATOM	9	CE1	HIS	A	1	49.691	26.152	6.454	1.00	17.00	C
ATOM	10	NE2	HIS	A	1	51.046	26.090	6.098	1.00	17.00	N
ATOM	11	N	SER	A	2	49.788	27.850	10.784	1.00	16.00	N
ATOM	12	CA	SER	A	2	49.138	29.147	10.620	1.00	15.00	C
ATOM	13	C	SER	A	2	47.713	29.006	10.110	1.00	15.00	C
ATOM	14	O	SER	A	2	46.740	29.251	10.864	1.00	15.00	O
ATOM	15	CB	SER	A	2	49.875	29.930	9.569	1.00	16.00	C
ATOM	16	OG	SER	A	2	49.145	31.057	9.176	1.00	19.00	O
ATOM	17	N	GLN	A	3	47.620	28.367	8.973	1.00	15.00	N
ATOM	18	CA	GLN	A	3	46.287	28.193	8.308	1.00	14.00	C
ATOM	19	C	GLN	A	3	45.406	27.172	8.963	1.00	14.00	C

Notice that each line or record begins with the record type ATOM. The atom serial number is the next item in each record. The atom name is the third item in the record. Notice that the first one or two characters of the atom name consists of the chemical symbol for the atom type. All the atom

names beginning with C are carbon atoms; N indicates nitrogen and O indicates oxygen. In amino acid residues, the next character is the remoteness indicator code, which is transliterated according to:

α	A
β	B
γ	G
δ	D
ϵ	E
ζ	Z
η	H

The next character of the atom name is a branch indicator. The next data field is the residue type. Notice that each record contains the residue type. In this example, the first residue in the chain is HIS (histidine) and the second residue is a SER (serine). The next data field contains the chain identifier. The next data field contains the residue sequence number. Notice that as the residue changes from histidine to serine, the residue number changes from 1 to 2. Two like

residues may be adjacent to one another, so the residue number is important for distinguishing between them. The next three data fields contain the X, Y, and Z coordinate values, respectively. The last three fields shown are the occupancy, temperature factor (B-factor), and element symbol. The spacing of the data fields is crucial. If a data field does not apply, it should be left blank.

Table 4: Results of 3D structure prediction algorithm based on similarity
The exact time required for the execution for some sample RNA sequences is as given below.

Sr.	File Name (.dot)	File Size (KB)	Sequence Length	Execution Time (sec)	Similarity (%)
1	yeast trna.dot	0.178	75	21.49	85.69
2	d.5.a.H.morrhuae.2.dot	0.273	120	39.4	80.46
3	d.5.a.H.marismortui.dot	0.281	122	31.24	97.28
4	d.5.a.H.saccharovororum.dot	0.285	123	27.06	93.75
5	d.5.a.H.morrhuae.1.dot	0.281	123	39.51	81.44
6	d.5.a.H.mediterranei.1.dot	0.284	123	45.58	93.89
7	d.5.a.D.mobilis.dot	0.297	133	28.05	89.31
8	X54252.dot	1.4	697	190.79	84.34
9	X54253.dot	1.4	701	133.75	85.63
10	Y00266.dot	2.5	1244	186.47	85.06
11	X98467.dot	2.6	1295	227.96	87.12
12	X65063.dot	2.8	1432	462.85	86.34
13	Z17210.dot	2.8	1435	242.38	90
14	X52949.dot	2.9	1452	203.95	89.89
15	K00421.dot	2.9	1474	220.99	87.59
16	Z17224.dot	3.1	1550	257.10	86.98
17	X59604.dot	3.3	1701	458.64	86.36
18	X00794.dot	3.9	1962	298.29	87.53

III. CONCLUSIONS

As the tertiary structures are increasing consequently their interactions are also increasing it becomes a tedious job to treat free energy. Hence a better solution to this difficulty is to consider already predicted secondary structure to build tertiary structure using molecular dynamics arithmetic. V-fold algorithm [2] is a better choice for 2D as well as 3D structure prediction but the time required to treat large RNA is more. Every day new databases are added with increasing sizes; hence motif template based method gives better result when homologous conformations are to be extracted from the known databases. The main difficulty arises when a new structure is identified. It is necessary to take the backup of such structure if it represents a good structure as homologous conformations cannot be available in PDB database. Hence in this research a 2D structure is predicted using GT-fold algorithm and a novel algorithm is developed using parallel techniques which can its corresponding 3D structure in faster and more accurate way. If the same algorithm is executed using parallel algorithm technique the result may be much better than as shown in table 4.

REFERENCES

- [1] Leontis NB, Westhof E, editors. RNA 3D Structure Analysis and Prediction. Volume 27. Berlin, Heidelberg: Springer Berlin Heidelberg; 2012 [Nucleic Acids and Molecular Biology].
- [2] Gestel and RF. The RNA World, Third Edition. 3rd ed. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 2005.
- [3] Saenger W. Principles of Nucleic Acid Structure. New York, NY: Springer New York; 1984 [Cantor CR (Series editor): Springer Advanced Texts in Chemistry].
- [4] Leontis NB, Westhof E. Geometric nomenclature and classification of RNA base pairs. *RNA*. 2001; 7:499–51
- [5] Eddy, S.R. (2001) Non-coding RNA genes and the modern RNA world. *Nat. Rev. Genet.*, 2, 919 – 929.
- [6] Mattick, J.S. and Makunin, I.V. (2006) Non-coding RNA. *Hum. Mol. Gen.*, 15(Suppl. 1), R17 – R29.
- [7] Mercer, T.R., Dinger, M.E. and Mattick, J.S. (2009) Long non-coding RNAs: insights into functions. *Nat. Rev. Genet.*, 10, 155 – 159.
- [8] Williams, K.P. (2002) The tmRNA Website: invasion by an intron. *Nucleic Acids Res.*, 30, 179 – 182.
- [9] Mello, C.C. and Conte, D. (2004) Revealing the world of RNA interference. *Nature*, 431, 338 – 342.
- [10] Cheetham, S., Gruhl, F., Mattick, J. and Dinger, M. (2013) Long noncoding RNAs and the genetics of cancer. *Br. J. Cancer*, 108, 2419 – 2425.
- [11] Huang, T., Alvarez, A., Hu, B. and Cheng, S.-Y. (2013) Noncoding RNAs in cancer and cancer stem cells. *Chin. J. Cancer*, 32, 582 – 593.
- [12] Zamore, P.D. and Haley, B. (2005) Ribo-gnome: the big world of small RNAs. *Science*, 309, 1519 – 1524.
- [13] Mehler, M.F. and Mattick, J.S. (2006) Non-coding RNAs in the nervous system. *J. Physiol.*, 575, 333 – 341.
- [14] Esteller, M. (2011) Non-coding RNAs in human disease. *Nat. Rev. Genet.*, 12, 861 – 874.
- [15] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) the Protein Data Bank. *Nucleic Acids Res.*, 28, 235 – 242.
- [16] Berman, H.M., Battistuz, T., Bhat, T.N., Bluhm, W.F., Bourne, P.E., Burkhardt, K., Feng, Z., Gilliland, G.L., Iype, L., Jain, S. et al. (2002) The Protein Data Bank. *Acta Crystallogr. D. Biol. Crystallogr.*, 58, 899 – 907.