Sindh Univ. Res. Jour. (Sci. Ser.) Vol. 46 (3) 385-392 (2014)

SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)

CORONATION: Comparison of RibO Nucleic Acid Tertiary's Involving Overlapping Networks

Z. U. A. KHUHRO[++], F. N. MEMON, A. P. HARRISON*

Institute of Mathematics and Computer Science, University of Sindh, Jamshoro

Received 17th July 2014 and Revised 12th September 2014

Abstract*:* A fast approach to compare RNA tertiary structures has been developed. This study considers RNA structures as collection of points (atoms) in three dimensional coordinates. This comparison is completely independent of nucleotide (base) order in sequence of RNA molecules. The CORONATION matches single nucleotides between two different RNA structures by looking at their three dimensional coordinates regardless of their relative sequential position. This method works by using description of geometrical patterns within each RNA structure based on graph theory. It finds similar 3D arrangements of bases among different structures by identifying overlapping between two structures. This overlap can be either a particular motif of RNA or a general matching of RNA chain (surface).

This paper presents extended results of our previous work and further extension in generalization and exactness of the algorithm. This efficient technique compares two RNA molecules in short possible time (mostly in a fraction of a minute). We have compared RNA structures with each other and successfully detected a good number of structures that show their overlapping score. We have particularly compared tRNA structures against the other families of RNA structure present in Protein Data Bank (PDB). Besides detecting similar patterns for tRNA, we spotted the motifs in different RNA families such as Saccharomyces cerevisiae (S.C) and Escherichia coli (E.Coli), etc. These motifs are well known and have been inspected by different procedures before. This study has many advantages and implications for diseases detection, RNA folding, evolution, and can be used to search for pharmacophoric patterns.

Keywords: RNA comparison, nucleotides, graph theoretic method, 3D RNA motifs, tRNA.

1. INTRODUCTION

The RNA structures play a critical role in carrying out the necessary biological functions, so their comparisons give us more insight into evolutionary and functional aspects. We have presented an initial idea (Khuhro, 2009) of the method for comparing RNA tertiary structures. The present paper presents the CORONATION method that is the extension of basic idea of the previous method (Khuhro, 2009). CORONATION is sequence order independent technique for the comparison of three dimensional RNA structures. The method can compare any pair of RNA structures whether determined by X-ray crystallography or NMR.

The CORONATION considers atoms of nucleotides as collection of points in space and finds the cliques between the structures. Detecting structure similarities among RNA 3D structures give more insight into their functional and even evolutionary relationship that would not be detected by sequence similarity alone (Chang, 2008, Petr, 2012, Mattei, 2014). Our 3D approach overcomes a major limitation found in other structural comparison techniques which required sequence similarity as in DIAL (Ferre, 2007). DIAL method is developed to compare RNA structures using dynamic programming algorithm, for computing global, local, and semi-global alignments by sequence similarity, dihedral angles and base pair information. While ARTS (Dror, 2005) is an excellent cubic time method for detecting the structural motifs that describes RNA molecules with a set of 'quadrats' composed of four phosphate atoms of two consecutive base-pairs and uses a bipartite graph to find a maximum number of aligned 'quadrats' between two RNA structures. It should be noted that ARTS does not necessarily preserve linear order within the alignments. Moreover, ARTS takes no account of nucleotide similarity. Another method NASSAM (Harrison, 2003) presents a graph theoretic method to search nucleic acid structures in 3D pattern where each base is represented by two vectors and a whole nucleic acid as a labeled graph. The vectors represented in bases are considered as nodes and the distances between them as edges. Once the structures are searched as graphs, the pattern within the structures is found by using exponential time Ullman algorithm for sub graph isomorphism. Although, it is useful method for searching motifs but it is not suitable to identify new motifs which are not specified earlier (Dror, 2005).

The CORONATION works entirely in a different way and is a unique for RNA structure comparison because it finds spatial similarity between the atoms of phosphate and sugar groups belonging to RNA molecule, regardless of the order and the directionality of the residues in the chain. In particular this approach does not require matches of fragments of residues for particular motif. This shows the generality and applicability of CORONATION. Secondly, it can compare even the recurring motifs without predefinition of the motifs because all the RNA molecules are compared simultaneously.

[++]Corresponding author: Z. U. A. Khuhro, Email: zain@usindh.edu.pk
*Department of Mathematical Sciences, University of Essex, UK

RNA structures give insight into evolution which can be gained by comparison. The similar structures provide an evidence of evolution because such an evolutionary pattern allows biologists to trace the evolutionary path of different species. This method helps to locate the similar structures even though they belong to different RNA families.

CORONATION has been applied to compare the tRNA structure against other RNA structures present in PDB (Berman, 2007). The matches are obtained with other RNA families. Particularly, in the unicellular eukaryote Saccharomyces cerevisiae (S.C) and the prokaryotes Escherichia coli (E.Coli), and other different types of families were found with great similarities. Although such comparison of RNA structures have been described in many methods but CORONATION works efficiently without prediction of any particular motif. This method is able to achieve high level performance because of the approach we have adopted.

2. METHOD

Although there are available methods for solving such problems for comparing protein secondary structures (Grindley, 1993; Harrison, 2003) but there is no such method for comparing RNA structures based on the framework used in our method. Though different approaches prevail which are ARTS (Dror, 2005), and COMPADRES (Wadley, 2004) that also use the three dimensional coordinates framework. But their computational constraints are very high.

The presented framework of CORONATION consists of a series of steps and each step of this pipeline is discussed below.

2.1 RNA Data transformation into Graphs

First of all, the RNA tertiary structures are transformed into graphs that superimposes the largest number of bases containing all the atoms of one structure onto the bases containing all the atoms of another RNA structure within three constraints; length τ, dot product angle θ and dihedral angle φ between the atoms of phosphate and sugar groups.

This method considers only O5′ and C5′ atoms from the phosphate and sugar groups respectively in each base of two RNA structures (Khuhro, 2009). We create a graph as follows:

- Create a vector A between the above said atoms in all nucleotides in every structure.Calculate the dot product angle θ between the two vectors (Suppose a RNA structure contains two nucleotides P and Q, then we have two vectors; vector A between two atoms of nucleotide P and vector B between two atoms of nucleotide Q).
- Calculate the midpoints m_1 and m_2 on vectors A and B respectively.
- Calculate another vector m and its magnitude τ which is nothing but length between the mid points m_1 and m_2.
- Calculate the cross product between vectors A & m and vectors B & m resulting in two vectors A x m and B x m.
- Finally calculate dihedral angle φ between the two vectors A x m and B x m.

2.2 RMSD Calculation

We have calculated the root mean square deviation (rmsd) between the RNA structures as follows:

$$rmsd = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(M_i - N_i\right)^2}$$

rmsd is the measure of the average distance between the atoms of cliques.

2.3 Generating a sequence of matrices to compare RNA structures

This method works by generating a sequence of matrices to compare a pair of structures, each with their molecular sequences $n_1\ n_2\ n_3 \ldots\ n_N$ and $m_1\ m_2\ m_3 \ldots\ m_M$, where N and M are the number of bases in the two structures.

a. The first matrix is termed as the SEQ (sequences) matrix and it stores the information about the matches of bases between two RNA molecules.

b. The second matrix is named as the correspondence matrix and it provides the information of the allowable relationship between pairs of bases. This relationship may correspond to the similar geometric patterns in the given pair of graphs. This matrix depends on the pairs of labels in the SEQ matrix. (Khuhro, 2009) provides further description about the two matrices with detailed examples.

2.4 Two graphs having identical structures are known as "isomorphic" structures. For a RNA structure, the relative geometries between pairs of bases are represented by the edges of its graph. Hence, a geometric arrangement of bases, common in two structures, forms the isomorphic regions. These isomorphic regions among the two RNA structures is identified by finding common clique between the structures and the clique is found by further studying the two matrices (SEQ and correspondence).

2.5 Data Set

A large number of PDB (Berman, 2007) entries of RNA structures of different families have been downloaded and used in this study.

3. RESULTS

In this extended work, we have compared the RNA structures by structured database using hundreds of RNA structures from major RNA families: Saccharomyces cerevisiae (S.C), Escherichia Coli

Table 1: Top ranking matches obtained in the comparison of reference structure (1EHZ.pdb) against other RNA families. Notation is as follows: S is the serial number of top ranked RNAs, PDB id, RNA molecule, Chain, Modified Bases in each RNA molecule, Size, i.e., total number of residues in a specified structure, C is the size of clique, rmsd (section 2.2), Score (section 3.1) and organism. Furthermore RNA scores are given below. In parenthesis, for each RNA structure, we have S, Size, C and Score. The RNAs are: 3bt7 (21,19,8,0.211); 1yz9 (22,11,6,0.208); 3epk (23,69,15,0.207); 398d (24,8,5,0.203); 2gbh (25,16,7,0.201); 2g92 (26,12,6,0.199); 1k6g (27,22,8,0.196); 1kpy (28,22,8,0.196); 3epj (29,69,14,0.193); 1hs8 (30,13,6,0.191).

S. #	PDB	Source and Molecule	Chain	Modified bases	Size	C	rmsd	Score	Organism
1	1EHZ	RNA and tRNA, Phe	A	11	76	76	0.000	1.000	S.C
2	1EVV	RNA and tRNA, Phe	A	11	76	52	0.023	0.69	S.C
3	1G1X	Ribosome and tRNA, Phe	B	12	76	48	0.023	0.63	T.T
4	1ML5	Ribosome and tRNA, Phe	B	11	76	46	0.022	0.61	E.C
5	1I9V	RNA and tRNA, Phe	A	2	76	45	0.023	0.592	E.C
6	4TNA	TRNA and tRNA, Phe	A	11	76	42	0.026	0.553	S.C
7	1TN2	TRNA and tRNA, Phe	A	11	76	41	0.025	0.539	S.C
8	4TRA	TRNA and tRNA, Phe	A	11	76	35	0.024	0.461	S.C
9	1FCW	Ribosome and tRNA, Phe	A	11	76	34	0.023	0.447	E.C
10	1MJ1	Ribosome and tRNA, Phe	D	11	76	31	0.031	0.408	E.C
11	6TNA	TRNA and tRNA, Phe	A	11	76	30	0.023	0.395	S.C
12	1FIR	RNA and HIV-I Reverse Transcription	A	9	78	21	0.03	0.273	B.T
13	2B6G	RNA Binding Protein and RNA	B	0	19	10	0.015	0.263	S.C
14	1OB2	HydrolaseNuclear Protein and tRNA, Phe	B	11	76	19	0.018	0.25	E.C
15	2DLC	LigasetRNA and tRNA	Y	0	69	18	0.023	0.249	S.C
16	3FOZ	TransferaseRNA and tRNA, Phe	C	0	74	18	0.022	0.24	E.C
17	2B7G	RNA and RNA	A	0	19	9	0.014	0.237	S.C
18	157D	RNA and RNA	A	0	12	7	0.011	0.232	E.C
19	1QF6	LigaseRNA and Threonine tRNA	B	5	76	17	0.016	0.224	E.C
20	3EPH	TransferaseRNA and tRNA	E	0	69	16	0.028	0.221	S.C

(E.Coli), Thermus Thermophilus (T.T), Bos Taurus (B.T) and others. In all cases, the results of comparison can be grouped as follows:

1) The highest scores (i.e. the maximum cliques) in all RNA structures correspond to the same family compared to reference tRNA structure; 1EHZ.pdb (Shi, 2000). In this comparison no sequence information or order is used. All the maximum cliques involve the same family members which imply the evolutionary divergence.

2) The other nearer scores correspond to RNA structures which have some similar features. These overlapped RNA structures which contain fewer matching than above usually have a larger rmsd.

3) The smallest clique size of RNA structures show dissimilarity which might contain only a few nucleotides in common. **(Table 1)** includes rmsd computed for specific cliques found in RNA structures and from these comparisons it can be seen clearly the difference between the atomic positions of atoms between the structures. The rmsd between the corresponding atoms of the RNA structures is the simplest way to compare the structures after they have been superimposed. The interesting thing is that rmsd does not even exceeds 1.0 Å for all the RNA structures, which is substantially more than the expected error of the 1.93 Å resolution structure reported as reference structure. Clearly the structures are interpretations of overlapping; all of them place the atoms of each residue in about the same positions. The **(Fig. 1)** shows detailed description.

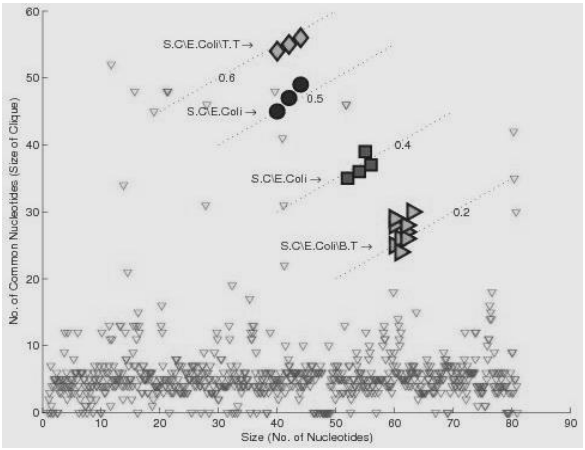

Fig.1: We compared tRNA phenylalanine (phe) (PDB id 1ehz) against a data set of 800 randomly selected RNA structures from PDB (Berman, 2007) having a resolution of 1.93 Å or better. The data set covers the major RNA families. It includes Saccharomyces cerevisae(S. C), Escherichia coli (E. C), Thermus thermophilia (T.T), Archaeoglobus Fulgidus (A, F), HIV-1, Aquifex aeolicus, Homo sapiens, Turnip mosaic virus, Thermoanaerobater tengcongens, Hammerhead ribozymes, Deinococcus radiodurans, Arabidopsis thalina thiamine, Plantia Stali intestine virus and others. Some of the highest score RNAs

are listed below. Each small triangle downwards in the figure represents the comparison between the reference PDB id and one from the data set of 800 RNAs. The x-axis of the figure shows the size (No. of nucleotides) of the RNA compared and the y-axis shows the number of common nucleotides found in the comparison with tRNA of S.C. We also compute the similarity score for each comparison. This score takes into account the overlapping nucleotides and the root of the reference structure and one of the structures from data set. The calculation for score is done by the formula (section 3.1), where clique is the number of overlapping of RNA structures obtained in comparison, reference structure is the number of nucleotides in the structure 1ehz which is 76 here, and target structure is number of nucleotides in the RNA structure compared. Score levels (0.2, 0.4, 0.5, 0.6) are represented by doted lines. The 3 highest ranked scores correspond to S.C, E. Coli and T.T structures (green filled squares), another 3 scores corresponds to S.C and E. Coli (blue filled circles). Four scores correspond to S.C and E. Coli (magenta filled squares) and the rest scores corresponds to S.C, E. Coli and B.T (green filled triangles). There are many RNA structures in data set with zero clique size (number of common nucleotides), i.e. zero scores. The 11 highest ranked scores correspond to phenylalanine (phe) molecule of fungi and bacteria in data set. Ranked 12 is HIV-1 reverse transcription of Bos Taurus organism. The rest ranked molecules belong to either Saccharomyces cerevisae or Escherichia coli. The asterisk shows all comparisons in the figure. Table 1 lists top 20 ranked RNA structures and the legend includes some other RNA structures, having the next largest number of cliques. The comparison of these RNA structures took about 50 seconds or roughly speaking 16 structures per second on Linux operating system.

3.1 tRNA Comparison

We also present here the results of comparing the data set with tRNA phenylalanine (phe) from Saccharomyces Cerevisae (yeast), PDB code 1EHZ (Shi, 2000). In addition to the expected overlapping with all the tRNAs in the database, the cliques, not conserving the order of sequences, are obtained with other RNA families.

Fig. 1 shows the results of the comparison between the reference structure and target structures from the data set of 800 RNA structures. The small triangles downwards in Figure 1 correspond to the comparison of tRNA phe against other RNA structures. We show the relationship between the molecules of data set RNAs and the number of cliques (number of common nucleotides) in Figure 1. We define the similarity score as follows:

$$score = \frac{Cliques}{\sqrt{\mathrm{Re}\,ference\ \ Structure \times\ \ T\arg et\ \ Structure}}$$

Further explanation of score has been given in the legend of Fig. 1. The highest scoring RNAs correspond to phenylalanine (phe) molecule of RNA of S.C and E. Coli in the data set (including 1ehz). As mentioned above, significant overlapping is detected in these two species. Though there are other species, i.e T.T, which is ranked as 3rd molecule and B.T ranked as 12 molecules in RNA structure comparison. Table 1 lists the names, sources, chains, modified bases, size of molecule, and common nucleotides in molecules, scores and rmsd distances between the top 20 RNA structures (some of them are enlarged in Fig. 1).

The comparison of another tRNA phe structure has been carried out with very similar results. The overlapping of top ranked RNAs show higher similarity in structure comparison. The structures of the S.C and E. Coli are very similar to each other as compared to T.T and B.T. The structure tRNA phe of S.C and E. Coli share similar motifs but the overlapping score between them is low. The S.C and E. Coli have long been model organisms and have been the subjects of experiments and it seems that it has been evolved from a common ancestor and has relatively high sequence and structural homology which shows the evidence of existence (for prokaryotes 3.8 billion years ago (bya) and of eukaryotes 2.7 bya (Mojzsis, 1996)).

As the structures of S.C and E. Coli are tRNA phe molecules so their tertiary structures are best described by a compact ***L*** shape where the anti-codon is a single stranded loop at the bottom which base pairs with the triplet codon. The amino acid is attached to the terminal ***A*** on the upper right. The active sites (anticodon and amino acid) are maximally separated. The anticodon stem and acceptor stem form double helix. For secondary structure of yeast phenylalanine (phe) transfer RNA we have obtained some particular overlapping motifs in PDB id 1evv and PDB id 1gix in Table 2. It seems that it is not a random overlapping of nucleotides but consists some particular motifs i.e. Acceptor stem, D stem, T stem, T loop, etc. The overlapping between S.C and E. Coli is very similar and almost the structures compared in Table 2 belong to these organisms.

Table 2: Comparison between PDB id 1evv and PDB id 1gix give us the overlapping motifs and bases positions for tRNA phenylalanine (phe) structure. In this comparison the two structures are almost overlapping with each other. But there are some 11 bases which are excluded because they do not satisfy the constraint conditions.

Name of Motif	Bases positions
Acceptor stem	1-7
Turn	8-9
D stem	10-13
D-loop	14-21 (but base 17 is not overlapped)
D stem	22-25 (but base 24 is not overlapped)
Turn	26
Anti-codon stem	27-31
Anti-codon loop	32-38 (but 35-37 bases are not overlapped)
Anti-codon stem	39-43 (but 39-41 bases are not overlapped)
Var loop	44-48
T stem	49-53
T loop	54-60
T stem	61-65
Acceptor stem	66-72
CCA end	73 (but 74-76 bases are not overlapped)

In **(Table 2)**. tRNA molecule PDB id 1gix (Yusupov, 2001) rank 3rd when compared with PDB id

1EHZ. It also ranked 2nd in the comparison of tRNA molecule 1evv (Jovine, 2000) against the database (not given here). However, it is important to note that although the structures of S.C and T.T are not much similar, but there is high overlapping between them. This is good example of organisms where different ancestors converged to the same structural solution. This was not expected that these two different organisms, S.C and T.T, would rank so high in the structural comparison. The overlapping of structures is three dimensional without any sequential order conservation.

Interestingly, HIV-I (Benas, 2000) ranked very high in comparison with S.C. In Table 1 this HIV-I molecule ranked 12. Similarly, comparing the E. Coli molecule PDB id 1gix with data set of RNA structures, HIV-I found in the top of list and ranked 20. The HIV-I structure is exactly comparable to the well known L-shape structure of tRNA phe (Benas, 2000). The overlapping between the B.T and E Ccoli as well as the match between the S.C and E. Coli are very similar.

The structural similarity between the organisms can be visualized in RasMoL (Sayle, 1993) computer software. Our method succeeded in finding the particular motifs in tRNA molecules automatically, without any prior knowledge of their existence. It is interesting to note that only atoms of phosphorus and sugar groups were compared and the overlapping which we found is not 100% but these overlapping appear very high in (Table 1 and in Fig.1). The 10 S.C, the 8 E. Coli, one T.T and one HIV-I RNA molecules are within top 20 ranking scores. The overlapping of tRNA molecules are completely sequence order independent, comparing single, isolated, though functionally similar nucleotides, lying in particular motifs. They are not adjacent, and thus do not belong to particular fragments, nor do they conserve the linear order of sequences, even though (Fig. 2, 3) shows good overlapping. Also they are not obtained using previously published methods.

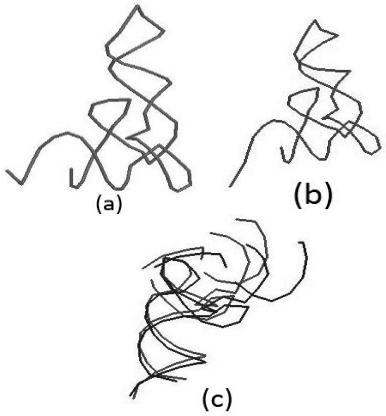

Fig. 2: a) and b) Backbone structure (created in RasMol (Sayle, 1993)) of PDB id 1evv and PDB id 1gix with its nucleotides which are represented by bases i.e. G, C, G, G, A, U, U, U, A, 2MG, C, U, C, A, G, H2U, G, G, G, A, G, A, G, C, M2G, C, C, A, G, A, OMC, U, OMG, A, G, A, G, 7MG, U, C, 5MC, U, G, U, G, 5MU, PSU, C, G, 1MA, U, C, C, A, C, A, G, A, A, U, U, C, G, C, A c) It shows the overlapping of nucleotides between PDB id 1evv with PDB id 1gix. The superposition of the atoms found by the algorithm brings the structure close together. This overlapping contains 65 nucleotides and the other nucleotides (representing their positions in chain i.e. 17,24,35,36,37,39,40,41,74,75,76) are not included because they do not satisfy the constraint conditions. Some segments show slight difference. This difference is because of threshold values which we have chosen as τ = 2Å, θ = 40 degrees and φ = 40 degrees.

Comparison between HIV-I and E. Coli resulted in 16 overlapping at an rmsd of 0.008 and a similarity score of 0.208 but in case of HIV-I with S.C, this comparison resulted with 21 overlapping at rmsd of 0.377 and a similarity score of 0.273. This score is ranked 12 with S.C and 20 with E. Coli in Table 1. This is interesting to note that the S.C and E. Coli are evolved through divergent ancestors but share similar structure. So, this HIV-I molecule also share significant structure homology with these organisms. As our method carries out the comparison without initial equivalences and ignores the sequential order of the nucleotides, this provides stronger evidence in favour of divergent evolution.

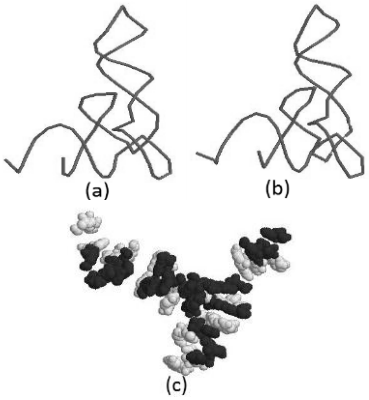

Fig. 3: Backbone structures (created in RasMol) of (a) PDB id 1evv and (b) PDB id 1fir with its nucleotides represented here with bases i.e. U, A, 2MG, G, C, C, A, A, U, G, 7MG, 5MC, G, PSU, C, 1MA, G, C, C (c) Shows the overlapping of three dimensional coordinates of atoms of nucleotides of PDB id 1EVV.pdb with 1FIR.pdb is shown. The atoms O5' and C5' were used for the nucleotide superposition that brings the structures close together. This overlapping contains 19 common nucleotides.

We have also compared other tRNA phe molecules that ranked highly in the comparison with two organisms S.C and E. Coli in order to check the results and score function of our algorithm. Comparing the 4TNA as reference structure with the data set, we found 7 S.C, 5 E. Coli, 2 HIV-I, 2 T.T, 2 Archaeoglobus Fulgidus, and 2 RNA 5'-3'. The top 20 ranked overlapping structures are from the same species along with the Archaeoglobus Fulgidus organism. As expected the maximum overlapping corresponds to S.C in the data set.

3.2 Retroviruses Comparison

RNA viruses (Retroviruses or riboviruses) store the genetic material in the RNA which is usually single-stranded (ssRNA) but may be double stranded RNA (dsRNA). RNA viruses which are responsible for many diseases (human diseases like SARS, influenza and hepatitis C) are classified into different classes and groups. They are ambisense (positive sense and negative sense) RNA viruses. We show here the results of positive sense ssRNA viruses (Group IV) of family leviviridae, PDB code 1aq3 (Van den Worm, 1998) against the other RNA families. In the comparison of RNA viruses, the top ranked RNA structures are found from the same RNA family.

Table 3: Top ranking matches obtained in the comparison of reference structure (1AQ3.pdb) against other RNA families. Notation is as follows: S is the serial number of top ranked RNAs, PDB id, RNA molecule, Chain, Size, i.e., total number of residues in a specified structure, C (size of clique), rmsd (section 2.2), Score (section 3.1) and Organisms. For example, some RNA scores are given in the form of RNA(S, Size, C, Score): 2i82 (21,21,7,0.382); 1s34 (22,23,7,0.365); 1f5u (23,18,6,0.354); 1ik1 (24,14,5,0.334); 2irn (25,9,4,0.333); 1oo7 (26,10,4,0.316); 1jtw (27,16,5,0.312); 1pgl (28,6,3,0.306); 1atv (29,17,5,0.303); 354d (30,12,4,0.289).

S. #	PDB	Source and Molecule	Chain	Size	C	rmsd	Score	Organism
1	1AQ3	Bacteriophage MS2 Mutant (T59s), RNA	R	16	16	0.000	1.000	Bacteriophage MS2
2	1AQ4	Bacteriophage MS2 Mutant (T45A), RNA	R	16	15	0.006	0.938	Bacteriophage MS2
3	2BU1	Bacteriophage MS2, RNA Hairpin (5bru-5) Complex	R	17	14	0.016	0.849	Bacteriophage MS2
4	1ZDI	Bacteriophage MS2, RNA	R	16	13	0.010	0.812	Bacteriophage MS2
5	2BNY	MS2, (N87A-mutant) RNA Hairpin	R	15	12	0.017	0.775	Bacteriophage MS2
6	2BQ5	MS2, (N87A-mutant) RNA Hairpin	R	19	13	0.019	0.746	Bacteriophage MS2
7	2C50	MS2-RNA HAIRPIN (A-5) COMPLEX, RNA	R	14	11	0.017	0.735	Bacteriophage MS2
8	2C4Q	MS2, RNA Hairpin (ONE-5)	R	17	12	0.017	0.728	Bacteriophage MS2
9	2C51	MS2-RNA HAIRPIN (G-5) COMPLEX, RNA	R	15	11	0.017	0.710	Bacteriophage MS2
10	1ZDH	Bacteriophage MS2, RNA Fragment	R	13	10	0.007	0.693	Bacteriophage MS2
11	2IZ8	MS2, Hairpin (C-7) RNA Hairpin	R	16	11	0.015	0.688	Bacteriophage MS2
12	2IZ9	MS2-RNA Hairpin (4ONE-5) Complex	R	15	10	0.014	0.645	Bacteriophage MS2
13	1ZDK	Bacteriophage RNA complex	R	19	11	0.012	0.631	Bacteriophage MS2
14	2B2E	Bacteriophage MS2, (N87S, E89K), RNA	R	16	10	0.012	0.625	Bacteriophage MS2
15	2IZM	MS2, RNA Hairpin	R	13	8	0.018	0.555	Bacteriophage MS2
16	2B2G	MS2, (N87S), RNA	R	16	8	0.017	0.500	Bacteriophage MS2
17	5MSF	MS2, RNA	R	18	8	0.017	0.471	Bacteriophage MS2
18	2IZN	MS2, RNA Hairpin	R	15	7	0.011	0.452	Bacteriophage MS2
19	1ZDJ	MS2, RNA	R	8	5	0.007	0.442	Bacteriophage MS2
20	1U1Y	Bacteriophage MS2, RNA	R	17	7	0.018	0.424	Bacteriophage MS2

Table 3 includes cliques, rmsd and score of top ranked RNAs. The overlapping of top ranked RNAs show higher similarity in the same species as compared to other RNA families. This reference structure which belongs to leviviridae family of RNA virus has major overlapping within the same family. In Table 3, we searched for general 3D similarities allowing any particular motifs but no sequence order is considered.

The results show that a number of RNA virus structures which belong to same family and Group IV have much overlapping. Our result is the set of pair wise similarities. These similarities can be used to position each structure type on other as is given in Figure 4. These similarities have both evolutionary and functional implications.

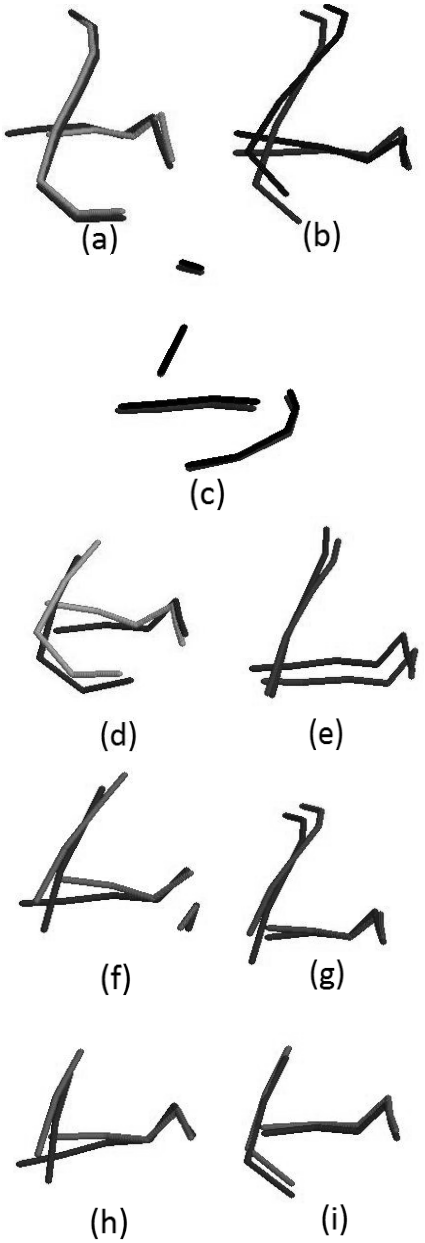

Fig. 4: Backbone structures (created in RasMol) showing the variation in positions of phosphate groups for a variety of

structures with similar superposition for the available 10 RNA structures obtained by using CORONATION. Color codes for the individual RNA structure are as follows with their PDB codes in parenthesis followed by common nucleotide bases. red (1AQ3.pdb) against a) green(1AQ4.pdb, A, U, G, A, G, G, A, U, A, C, C, C, A, U, G) b) blue(2BU1.pdb, A, U, G, A, G, G, A, A, C, C, C, A, U, G), c)black(1ZDI.pdb, A, U, G, A, G, A, U, U, C, C, A, U, G), d)yellow(2BNY.pdb, U, G, A, G, G, A, U, C, C, C, A, U,), e)magenta(2BQ5.pdb, A, U, G, A, G, G, A, U, C, C, C, A, U), f)grey(2C50.pdb, A, U, G, A, G, A, U, C, C, C, A), g)purple(2C4Q.pdb, U, G, A, G, G, A, C, C, C, A, U, G), h) greenblue(2C51.pdb, A, U, G, A, G, G, A, C, C, C, U,), i)pink(1ZDH.pdb, U, G, A, G, G, A, A, C, C, C). These nine examples of overlapping structures were selected from Table 3 which represents a few structures of the PDB data set. We have not shown here the entire structures but only selected nucleotides which form a clique in two structures. Superposition of all the structures matches 3D nucleotides with each other. This overlapping is non sequential. In every structure comparison different number of nucleotides is used as is given in Table 3 with the heading common cliques (C). Different colors are used just for visualization. The whole idea is to see the overlapping of three dimensional coordinates of nucleotides of PDB id 1aq3.pdb with all the given structures. The small difference can be seen because of choice of threshold values.

4. DISCUSSION

We have developed an algorithm CORONATION RNA structure comparison and have applied the algorithm to compare RNA structures given their atoms, i.e. O5′ and C5′ coordinates. Our method is conceptually simple. This method in practice compares the divergently related RNA molecules and detects common 3D motifs in RNA structures, as assessed by visual inspection (Figures 2, 3, 4). We used the threshold values for length, dot product angle, and dihedral angle constraints. These constraints are fixed for getting positions of nucleotides. Our method works less than a minute for the data set we have downloaded and used in our experiments.

The classification of 3D motifs as a result of comparison in the data set may be useful in the analysis, design and prediction of RNA structures. The proposed method of RNA structure comparison is based on geometric invariants which is intrinsic property of 3D structures. Our method is more towards the structural comparison as compared to other methods which uses sequence alignments or RMSD as to find seeds. The RNA structures are calculated by constraint values of length and angles. There is no variation of θ and Ø angles within all the structures. The superposition of structures is considerably better and rmsd differences are also noticeably better. In particular, the structures overlapping with the reference structure show the maximum score with the RNAs which belong to same RNA family **(Fig.5).**

However, the choice of distance $\tau = 2$Å is better defined in structure comparison. When all the structure's common nucleotides were computed for distance constraint against the reference structure the rmsd is below 1 Å as is given in Table 1 and Table 3. Fig. 2, 3 and 4 show the spread of the RNA structures after comparison. It can be seen that the spread of RNA structures with the reference structure and target structure is slight away because of threshold values. The, threshold values, in our method, therefore contribute in RNA structure comparison. This is also surprising that the RMSD difference between the two RNA structures derived from their similar nucleotides is very small. Thus, these constraints calculations have better convergence properties in all RNA structures. This may be possible that, by relaxing the constraints on the RNA structures, many common motifs can be obtained.

Further analysis showed that the majority of overlapping motifs obtained came from structure pairs from same or common RNA families and those with a comparatively low similarity score were from different RNA families.

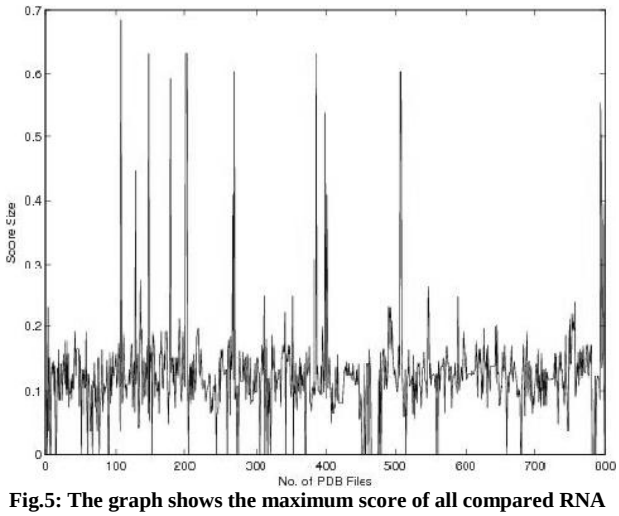

Fig.5: The graph shows the maximum score of all compared RNA structures used in this study.

5. CONCLUSION

Graphs are great because they communicate information visually. The transformation, of RNA structures using graphs, makes it easier to handle the complex structures. These graphs provide the basis for structure comparison.

RNA tertiary structure comparison method CORONATION is developed that does not depend on alignment of primary sequence or secondary structures. This method is purely based on geometric technique which used constraints i.e. length and angles.

CORONATION is applied on around 800 RNA structures and significant similarity ration is found among the RNAs. We have used fix constraints values to find the overlapping cliques in RNA structures. Making use of different measures of similarities can

also extend the use of algorithm to structural comparison. Furthermore, one of the clear applications of fast structure comparison method is the pair wise comparison of all unknown structures and depending upon those similarities order them into different classes.

REFERENCES:

Benas P., G. Bec, G. Keith., R. Marquet. C. Ehresmann., B. Ehresmann., P. Dumas (2000) The crystal structure of HIV reverse-transcription primer tRNA(Lys,3) shows a canonical anticodon loop, *RNA*, **6**(10), 1347-1355.

Berman H., K. Henrick, H. Nakamura and J. L. Markley (2007) The worldwide Protein Data Bank(wwPDB): ensuring a single, uniform archive of PDB data, *Nucleic Acids Research*, 35, Database Issue: D301-D303.

Chang Y.F., Y. L. Huang, C. Lu (2008) SARSA: a web tool for structural alignment of RNA using a structural alphabet, *Nucleic Acids Research*, 36, Web Server Issue:W19-W24.

Dror O., R. Nussinov, and H. Wolfson (2005) ARTS: alignment of RNA tertiary structures, *Bioinformatics*, 21, Suppl.ii47-ii53.

Ferre F., Y. Ponty, W. A. Lorenz, and P. Clote (2007) DIAL: a web server for the pairwise alignment of two RNA three-dimensional structures using nucleotide, dihedral angle and base-pairing similarities, *Nucleic Acids Research*, 35, Web Server Issue:W659-W668.

Grindley H.M.,P. J. Artymiuk, D. W. Ric, and P. Willet (1993) Identification of tertiary structure resemblance in proteins using a maximal common sub-graph isomorphism algorithm, *Journal of Molecular Biology*, 229, 707-721.

Harrison A.M., D. R. South,P. Willett, and P.J. Artymiuk (2003) Representation, searching and discovery of patterns of bases in complex RNA structures, *Journal of Computer-Aided Molecular Design*, 17, 537-549.

Harrison A., F. Pearl, I. Sillitoe, T. Slidel, R. Mott, J. Thornton and C. Orengo (2003) Recognising the fold of a protein structure, *Bioinformatics*, 19(14), 1748-1759.

Jovine L., S. Djordjevic, D. Rhodes (2000) The crystal structure of yeast phenylalanine (phe) tRNA at 2 .0 A resolution: cleavage by Mg(2+) in 15-year old crystals, *Journal of Molecular Biology*, 301(2), 401-414.

Khuhro Z. U. A., F. N. Memon, A. P. Harrison (2009) *Ribonucleic* acid tertiary structure comparison using Graph theory, *Proceedings: 2009 International Workshop on Computational and Integrative Biology (a satellite meeting of the International Conference of Integrative Biology)*, 18-20 September, 2009, Hangzhou, China.

Mattei E. A., Gabriele, F. Fabrizio, and H. Manuela (2014) A novel approach to represent and compare RNA secondary structures, *Nucleic acids research, Oxford Univ Press,* 42(10) , 6146-6157.

Mojzsis S.J., G. Arrhenius, K. D. Mckeegan, T. M. Harrison, A. P. Nutman, and C. R. L. Friend (1996) Evidence for life on earth before 3,800 million years ago, *Nature*, 384, 55-59.

Petr, C., D. Svozil, H. David (2012) SETTER: web server for RNA structure comparison, *Nucleic Acid Research, Pubmed*, *Oxford University Press*, 40(W1), W42-W48.

Sayle R. and A. Bissell, (1993) RasMol: A program for Fast, Realistic Rendering of Molecular Structures with Shadows, *Proceedings: 10th Eurographics, UK.*

Shi H., and P. B. Moore, (2000) The crystal structure of yeast phenylalanine (phe) tRNA at 1.93Å resolution: A classic structure revisited, *RNA*, 6, 1091-1105.

Van den Worm S.H., N. J. Stonehouse, K. Valegârd, J. B. Murray, C. Walton, K. Fridborg, P. G. Stockley., L. Liljas (1998) Crystal structures of MS2 coat protein mutants in complex with wild-type RNA operator fragments, *Nucleic Acids Research*, 26(5),1345-351.

Wadley L.M., A. M. Pyle (2004) The identification of novel RNA structural motifs using COMPADRES: an automated approach to structural discovery, *Nucleic Acids Research*, 32(22), 6650-6659.

Yusupov M.M., G. Z. Yusupova, A. Baucom, K. Liebermna, T. N. Earnest, J. H. D. Cate, and H. F. Noller (2001) Crystal structure of the Ribosome at 5.5 Å Resolution, *Science Express*, 292(5518), 883-896.