A STUDY OF TORSION ANGLES OF RNA MOTIFS

By

Sai Teja Kshir Sagar

Bioinformatics Independent Study

May 2010

Under guidance of Dr. Jason Tsong-Li Wang

WHAT ARE RNA MOTIFS

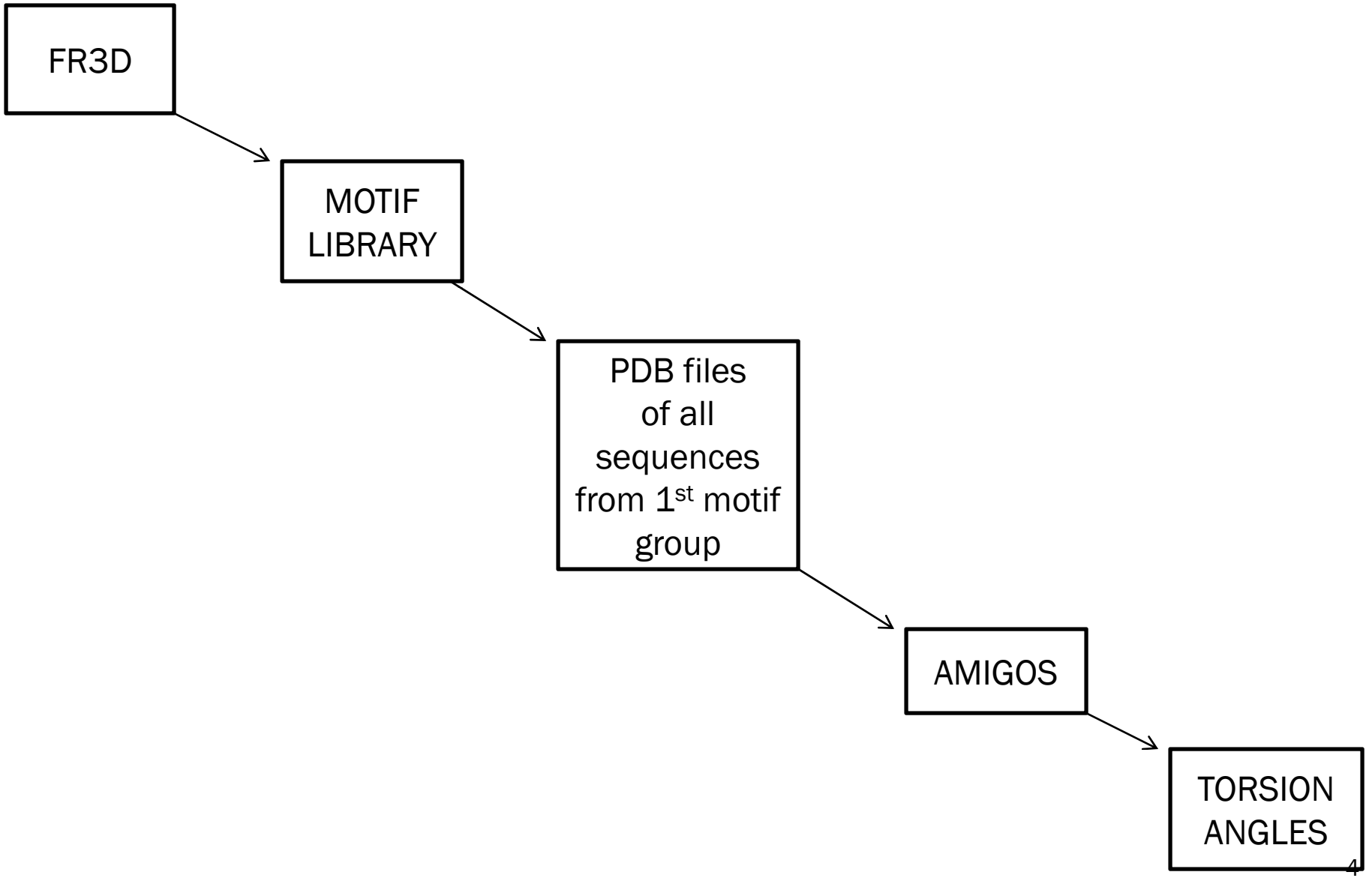
- Small sequence fragments of RNA which are present repeatedly in RNA.
- It is a 3-D structural element or fold within the chain.
- Same motifs can also appear in different other molecules.

MOTIFS

Types of RNA Motifs:

- Hairpin
- Kink Turn
- E-loop
- K-loop

PROCEDURE FOLLOWED FOR AMIGOS



FIND RNA 3D (FR3D)

- Developed by Dept. of Mathematics and Statistics, Bowling Green State University, USA.
- Used for finding recurrent 3-D motifs in RNA.
- Also used as a database of RNA structural motifs.
- Link : http://rna.bgsu.edu/FR3D

CATEGORIES OF RNA MOTIFS ON FR3D

- cWW-tHW-cSW-cWW_C-loop Motif
- tSH-tHH-cSH-tWH-tHS_sarcin/ricin Motif
- tWH-insertion-tHS Motif
- cWW-tWH-cWW_GAAA-receptor Motif
- cWW-(cWW-cSW)-(tWH-cWW)-cWW-cWW Motif
- cWW-tSH-tWH-tHS-cWW Motif
- tHS_C-loop Motif
- tSH-tHS Motif

ALGORITHMIC METHOD FOR IDENTIFYING GROUPINGS OF OVERALL STRUCTURE (AMIGOS)

- Developed by Pyle Lab.
- It is a Perl script which gives tables of torsion angles from nucleic acid PDB files.
- AMIGOS measures standard backbone torsion angles, i.e. alpha, beta, gamma, delta, epsilon, and zeta.
- It also calculates sugar pucker torsion (nu2), chi, and pseudo-torsions eta and theta angles.

INPUT AND OUTPUT FILES

- Amigos accepts only ent or pdb files as input files.
- It generates two output files for each pdb i.e. “filename_area.txt” and “filename_sprd.txt.”
- It also generates two output files (all_sprd.txt & all_area.txt) which contain measurement of all the nucleotides from all the pdb files.
- 2n+2 files are generated, where n is the number of pdb files.

AMIGOS TOOL (POINTS TO BE NOTED)

- HETATM entries in a pdb file are ignored by this tool.
- Bases adjacent to HETATM's torsion are not calculated.
- Only those residues which either contain 'O2' / 'O2*' or are properly named as A,G,C,U or T are considered for geometric calculation.
- Output does not contain the measurements of nucleotide at the start or end of the chain.

POINTS TO BE NOTED

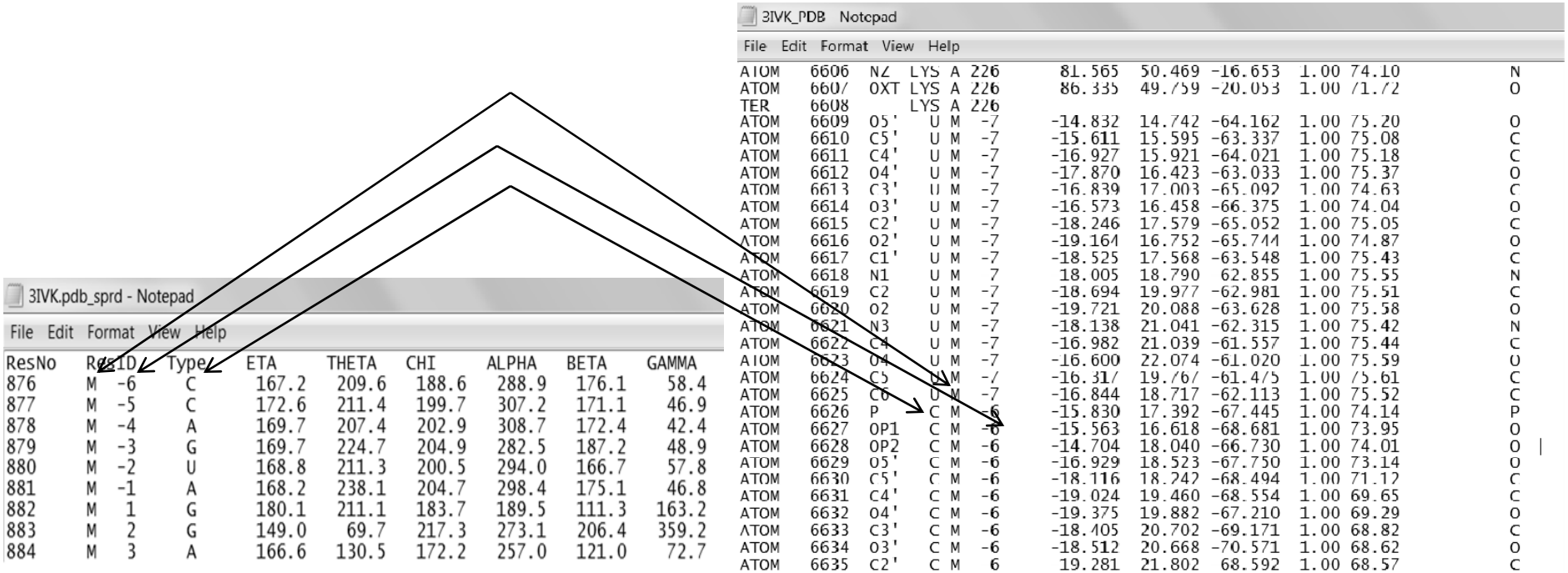
- The tool strictly calculates the measurements of RNA residues ignoring any other protein in the pdb file.
- By default it calculates area of all the nucleotides which fall outside the helical region, but this can be modified in the script according to the need.
- We can also direct the program to calculate measurements of any four user-defined atoms as well, but this has to be modified in the code.

EXAMPLE:

INTERPRETATION OF 3IVK.PDB_SPRD.TXT FILE

- Since the tool reads the file, residue-wise , the res.no does not start with 1 but with 876 because the tool starts reading RNA residues in the pdb file 3IVK from 875th residue.

- Compare the columns Res(2nd), ID(3rd) and type(4th) in the 3IVK.pdb_sprd file with that of the column 5th, 6th and 4th respectively from atom no.6626 in 3IVK.pdb file.



3IVK.pdb_sprd - Notepad

File Edit Format View Help

ResNo	ResID	Type	ETA	THETA	CHI	ALPHA	BETA	GAMMA
876	M -6	C	167.2	209.6	188.6	288.9	176.1	58.4
877	M -5	C	172.6	211.4	199.7	307.2	171.1	46.9
878	M -4	A	169.7	207.4	202.9	308.7	172.4	42.4
879	M -3	G	169.7	224.7	204.9	282.5	187.2	48.9
880	M -2	U	168.8	211.3	200.5	294.0	166.7	57.8
881	M -1	A	168.2	238.1	204.7	298.4	175.1	46.8
882	M 1	G	180.1	211.1	183.7	189.5	111.3	163.2
883	M 2	G	149.0	69.7	217.3	273.1	206.4	359.2
884	M 3	A	166.6	130.5	172.2	257.0	121.0	72.7

Screen shot of 3IVK.pdb_sprd file

3IVK_PDB Notepad

File Edit Format View Help

```
ATOM   6606  NZ  LYS A 226      81.565  50.469 -16.653  1.00 74.10           N
ATOM   6607  OXT LYS A 226      86.335  49.759 -20.053  1.00 71.72           O
TER    6608      LYS A 226
ATOM   6609  O5'   U M  -7     -14.832  14.742 -64.162  1.00 75.20           O
ATOM   6610  C5'   U M  -7     -15.611  15.595 -63.337  1.00 75.08           C
ATOM   6611  C4'   U M  -7     -16.927  15.921 -64.021  1.00 75.18           C
ATOM   6612  O4'   U M  -7     -17.870  16.423 -63.033  1.00 75.37           O
ATOM   6613  C3'   U M  -7     -16.839  17.003 -65.092  1.00 74.63           C
ATOM   6614  O3'   U M  -7     -16.573  16.458 -66.375  1.00 74.04           O
ATOM   6615  C2'   U M  -7     -18.246  17.579 -65.052  1.00 75.05           C
ATOM   6616  O2'   U M  -7     -19.164  16.752 -65.744  1.00 74.87           O
ATOM   6617  C1'   U M  -7     -18.525  17.568 -63.548  1.00 75.43           C
ATOM   6618  N1    U M   7      18.005  18.790  62.855  1.00 75.55           N
ATOM   6619  C2    U M  -7     -18.694  19.977 -62.981  1.00 75.51           C
ATOM   6620  O2    U M  -7     -19.721  20.088 -63.628  1.00 75.58           O
ATOM   6621  N3    U M  -7     -18.138  21.041 -62.315  1.00 75.42           N
ATOM   6622  C4    U M  -7     -16.982  21.039 -61.557  1.00 75.44           C
ATOM   6623  O4    U M  -7     -16.600  22.074 -61.020  1.00 75.59           O
ATOM   6624  C5    U M  -7     -16.317  19.767 -61.475  1.00 75.61           C
ATOM   6625  C6    U M  -7     -16.844  18.717 -62.113  1.00 75.52           C
ATOM   6626  P     C M  -6     -15.830  17.392 -67.445  1.00 74.14           P
ATOM   6627  OP1   C M  -6     -15.563  16.618 -68.681  1.00 73.95           O
ATOM   6628  OP2   C M  -6     -14.704  18.040 -66.730  1.00 74.01           O
ATOM   6629  O5'   C M  -6     -16.929  18.523 -67.750  1.00 73.14           O
ATOM   6630  C5'   C M  -6     -18.116  18.242 -68.494  1.00 71.12           C
ATOM   6631  C4'   C M  -6     -19.024  19.460 -68.554  1.00 69.65           C
ATOM   6632  O4'   C M  -6     -19.375  19.882 -67.210  1.00 69.29           O
ATOM   6633  C3'   C M  -6     -18.405  20.702 -69.171  1.00 68.82           C
ATOM   6634  O3'   C M  -6     -18.512  20.668 -70.571  1.00 68.62           O
ATOM   6635  C2'   C M  -6      19.281  21.802  68.592  1.00 68.57           C
ATOM   6636  O2'   C M  -6     -20.499  21.943 -69.298  1.00 68.17           O
ATOM   6637  C1'   C M  -6     -19.484  21.290 -67.166  1.00 68.52           C
ATOM   6638  N1    C M  -6     -18.409  21.800 -66.287  1.00 68.19           N
ATOM   6639  C2    C M  -6     -18.412  23.155 -65.932  1.00 68.26           C
ATOM   6640  O2    C M  -6     -19.327  23.880 -66.345  1.00 68.56           O
ATOM   6641  N3    C M  -6     -17.418  23.633 -65.139  1.00 67.63           N
ATOM   6642  C4    C M  -6     -16.459  22.806 -64.722  1.00 67.20           C
ATOM   6643  N4    C M  -6     -15.507  23.316 -63.942  1.00 66.94           N
ATOM   6644  C5    C M  -6     -16.435  21.423 -65.079  1.00 67.28           C
ATOM   6645  C6    C M  -6     -17.420  20.964 -65.858  1.00 67.57           C
ATOM   6646  P     C M  -5     -17.350  21.381 -71.399  1.00 68.82           P
ATOM   6647  OP1   C M  -5     -17.390  20.854 -72.780  1.00 69.19           O
ATOM   6648  OP2   C M  -5     -16.105  21.273 -70.605  1.00 68.84           O
ATOM   6649  O5'   C M  -5     -17.794  22.917 -71.410  1.00 68.44           O
ATOM   6650  C5'   C M  -5     -19.090  23.289 -71.845  1.00 68.13           C
ATOM   6651  C4'   C M  -5     -19.310  24.755 -71.546  1.00 67.90           C
```

Screen shot of 3IVK.pdb file

COMBINED INTERPRETATION OF BOTH SPRD FILE AND PDB FILE FOR 3IVK

- RNA residue id in pdb file starts from -7(6th column) or atom no.6609.
- Thus the tool gives the measurement from residue id -6 or atom no.6626.
- All the atoms corresponding to -6 form a single residue of RNA which is 876th residue of that pdb file and is represented as res no.876 in sprd file.
- Thus all the torsion angles of all the residues read by the tool are given in the output file 3IVK.pdb_sprd.
- Residues from 876 to 1129 in output file can be identified in the 3IVK.pdb file from atom 6626 to 12141.

MOTIF GROUP: CWW-THW-CSW-CWW_C-LOOP (1ST GROUP MOTIF)

- AMIGOS Result of 1st Motif group of FR3D (sprd.txt).

- The result below is of PDB: 1KOG, which contains 6 motifs of the same group (1st group).

- All the other pdb file results of the group is given in the excel file.

1	PDB: 1KOG														
2	MOTIF:9														
3	ResNo	ResID	Type	ETA	THETA	CHI	ALPHA	BETA	GAMMA	DELTA	EPSILON	ZETA	NUTWO	TYPEA	
4	8	I 76	G	187.6	221.7	202.5	205.9	254.7	63.5						
5	6	I 74	A	180.8	231.8	212.2	333.6	117	61.2	94.5	217.50%	245	29.3	1	
6	28	I 96	C	169.6	205.9	193.3	287	140.2	80.2	86.8	225.20%	263.6	35.3	1	
7	29	I 97	C	212.8	129.5	207.3	296	151.5	68.5	100	335.20%	99.6	30.4	1	
8	31	I 99	C	174.5	221.1	198.6	303.8	181.5	42.5	80.3	216.80%	290.5	38.7	1	
9	32	I 100	U	165.8	215	200.9	304.1	174.8	43.7	79.3	219.20%	284.2	39	1	
10	MOTIF:5														
11	43	J 74	A	185.5	232.2	217.2	338.1	112.9	61.9	95.9	219.60%	244.1	27.2	1	
12	45	J 76	G	184.6	219.1	201	207.5	251.6	62.7	80.5	201.70%	293.2	40.5	1	
13	65	J 96	C	170.9	206.8	195.4	287.6	144	78.3	90.7	224.60%	263.4	33.7	1	
14	66	J 97	C	211.6	127.1	205	290.3	151	72.5	100.8	326.90%	99.3	30.6	1	
15	68	J 99	C	172.5	218.5	196.1	309.3	179.5	38.4	79.1	217.60%	289.1	37.2	1	
16	69	J 100	U	166.2	217.1	201.9	299.4	175.7	49.3	80.5	219.30%	284	40.8	1	
17	MOTIF:6														
18	101	N 74	A	182.6	227.1	213	336.2	116.9	60.2	93.8	217.30%	243.6	20.2	1	
19	103	N 76	G	188.9	221	202	208.7	254.9	61.9	83.1	202.70%	291	39.1	1	
20	213	N 96	C	171.2	207.1	195.1	290.2	143	78.3	85.4	221.10%	267.4	35.2	1	
21	214	N 97	C	213.5	130	208.1	298.1	151.3	68.2	95.6	336.40%	93.8	29.9	1	
22	216	N 99	C	175.1	219.9	199	308.5	180.9	40.8	77.9	219.00%	289.8	38.5	1	
23	217	N 100	U	165.1	212.5	201.3	306.5	174.3	42.5	78	219.20%	282.7	39.9	1	
24	MOTIF:7														
25	117	L 74	A	182.6	231.4	213.2	334.7	114.9	62.7	95	217.40%	244.5	29.3	1	
26	119	L 76	G	188.3	221.1	202.9	206	256.2	63.8	83.7	198.90%	291.2	39.1	1	
27	139	L 96	C	170.3	205.3	195	281	144	84.1	87.5	223.60%	266.3	33.8	1	
28	140	L 97	C	212.9	129.2	206.7	294.8	150.4	69.3	95.5	336.60%	95.1	29	1	
29	142	L 99	C	174.9	219.1	198.7	302.7	185.2	42.6	81.1	214.40%	289.8	38	1	
30	143	L 100	U	166.5	215.8	197.6	302.4	175.3	44.8	80.5	219.60%	283	39.5	1	
31	MOTIF:8														
32	154	M 74	A	183.1	227.3	217.4	334.8	118	61.3	96.8	216.20%	242.7	26.7	1	
33	156	M 76	G	186.2	220.3	203.7	209	252.4	63.3	82	201.30%	291.8	40.9	1	
34	176	M 96	C	170.7	205.6	194.8	284.1	142	82.9	85.9	220.90%	268.7	34.2	1	
35	177	M 97	C	213.2	127.1	209	292	151.8	71.8	99.4	329.00%	100.6	30.2	1	
36	179	M 99	C	172.9	218.5	198.3	300.5	181.4	45.1	78.4	214.40%	290.3	38.1	1	
37	180	M 100	U	165.5	212.6	200.1	300.6	175.1	46.5	76.6	219.40%	283.3	40.6	1	
38	MOTIF:10-														
39	228	O 74	A	186.1	228.2	214.9	335.5	113	68.4	94.6	217.60%	245.2	28.3	1	
40	230	O 76	G	183	219.4	199.4	206.4	251.5	64.8	82.8	202.20%	291	40	1	
41	250	O 96	C	168.4	204.7	192.8	276.7	143.9	89.3	84.6	221.90%	268.6	34	1	
42	251	O 97	C	216.6	128.7	215	294.6	151.8	71.4	92.6	340.50%	95.4	30.1	1	
43	253	O 99	C	172	217.7	199.3	308.9	175.6	42.4	75.2	214.10%	292.2	39.4	1	
44	254	O 100	U	167.3	214.2	198.1	301.2	173.4	48.1	77.1	218.90%	284.2	39.2	1	

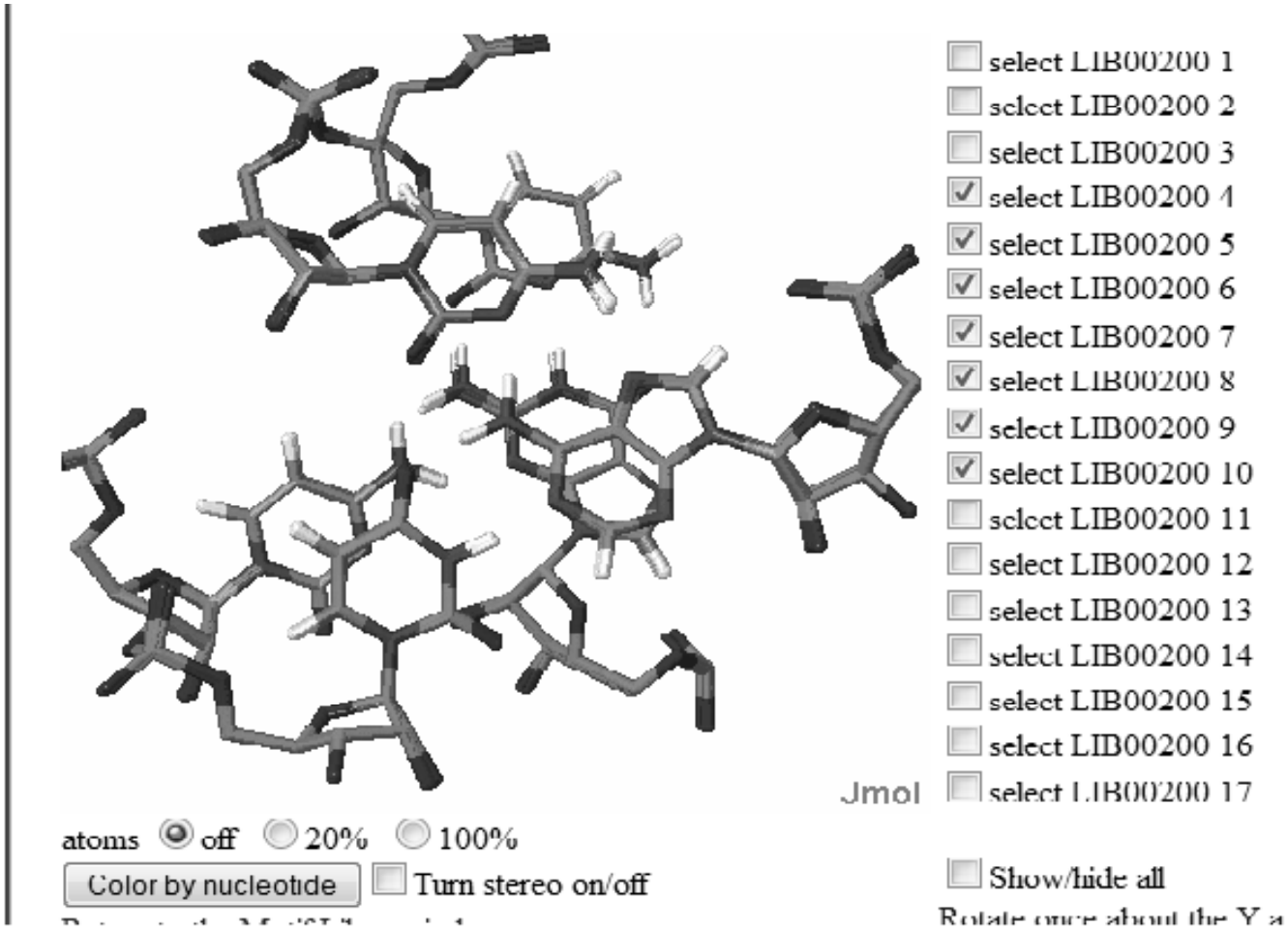
INTERPRETATION OF RESULTS

- There are 6 motifs of 1st group in 1KOG sequence.
- The area marked by the box are the eta and theta angles of the motifs of 1KOG.
- The table shows all the torsion angles of the motif.

INTERPRETATION OF RESULT

- We can see that eta and theta angles of all the motif residues are in a very similar range (+/- 10 degrees).
- In some of the residues the range is very small (+/-2 degrees).
- We can also see that all the other torsion angles of all the residues of the motifs are in same range.
- From the observation we can say that in a given RNA pdb, motifs from the same group have similar torsion angles, irrespective of their chain ID in the sequence.

JMOL VIEW OF ALL THE 6 MOTIFS IN 1KOG FILE



APPLICATION OF AMIGOS

- We can find patterns in the angles of RNA motifs.
- By the help of AMIGOS we can predict the motifs present in any RNA.
- If given an RNA and its motif, we can also classify the motif using AMIGOS, based on its torsion angles.
- By using AMIGOS we can do angle mining of RNA and its motifs.

OTHER TOOLS WHICH I HAVE WORKED ON

- PiRahNA
- PARTS

PIRAHNA

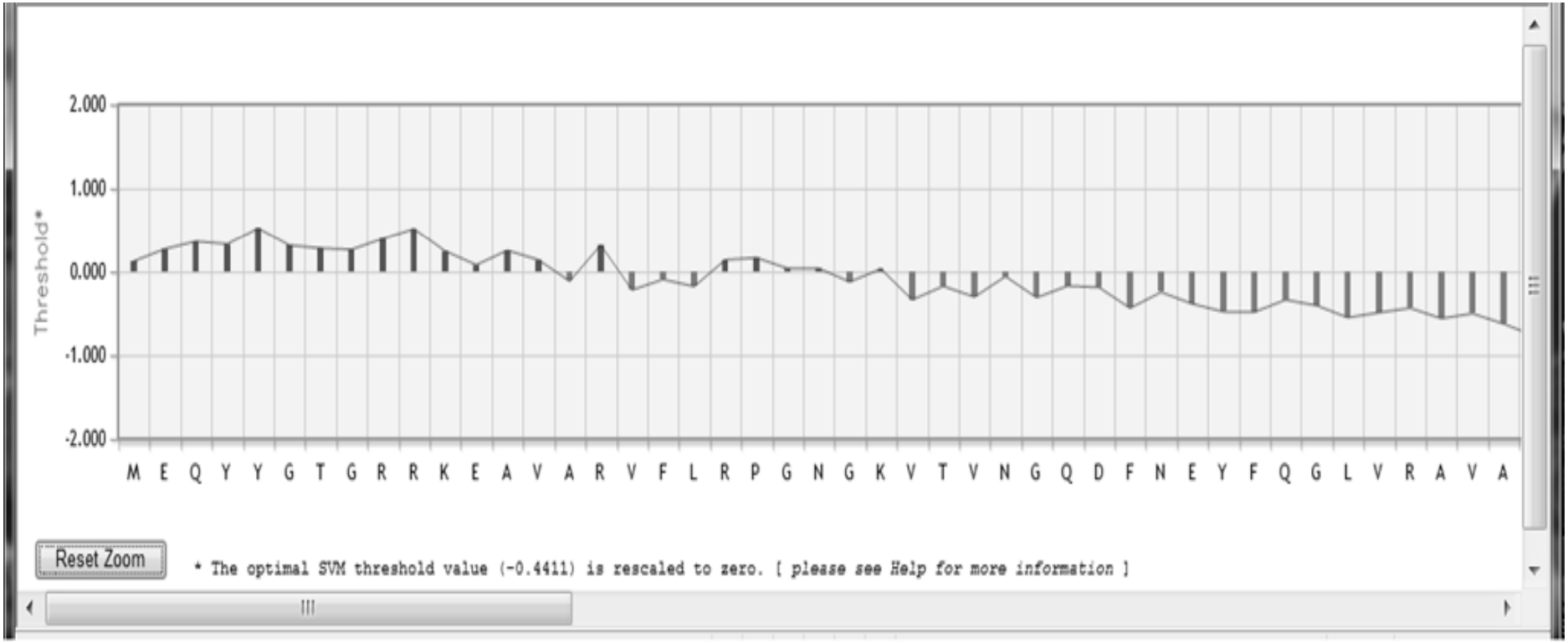
- This tool is based on “Protein Function Annotation from Sequence: Prediction of Residues Interacting with RNA”
- It predicts :
  - RNA-binding residues from protein sequence information
  - RNA-binding function at the protein level

INPUT AND OUTPUT

- Input: Protein sequence

- Output: Graphical representation, where
  - X-axis represents the query sequence
  - Y-axis represents SVM threshold values for individual residues.

OUTPUT RESULTS



OUTPUT INTERPRETATIONS

- Residues which have a SVM threshold above zero are predicted to be RNA binding residues of that sequence.

- In the graph it is represented by **RED** color bars.

- The higher the threshold value of the residue the less is false positive rate and vice versa for false negative rate.

OUTPUT INTERPRETATIONS

- In this tool the optimal threshold value is -0.4411 (which is rescaled to zero in the graph).
- It has a MCC of 0.50 and AUC of 0.86.
- The threshold was obtained by doing 5-fold cross validation of a non-redundant set of 81 RNAs taken from pdb.
- Uniqueness of this tool is that it uses both PSSM and physicochemical properties for RBR prediction.

PARTS

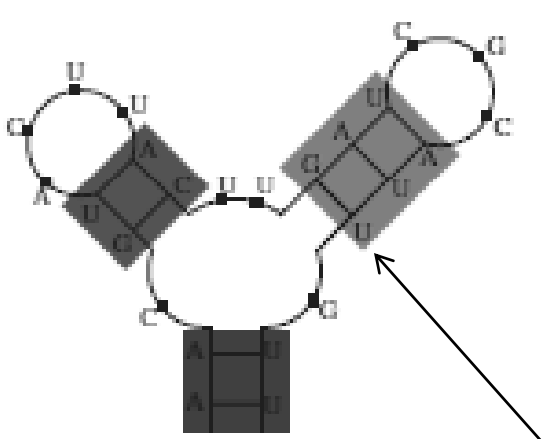
- Probabilistic Alignment for RNA joinT Secondary structure prediction
- Developed by University of Rochester, USA.
- It is a tool to predict alignment and secondary structures of two RNA sequences.

PARTS

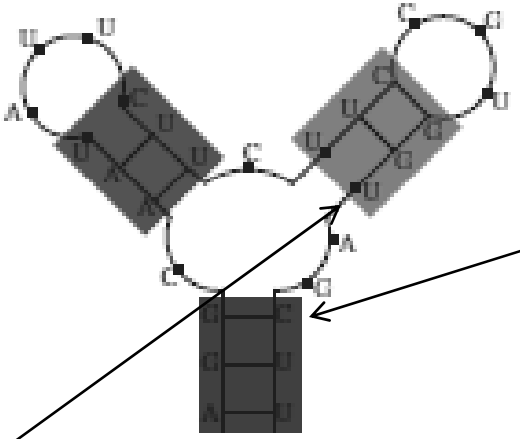
- In this tool the RNA base pairs are aligned first and then they are aligned sequentially.
- This helps in increasing the accuracy of secondary structure prediction.
- It also considers insertion and deletion of base pairs.

PARTS

(a) Sequence alignment



(b) First sequence structure.



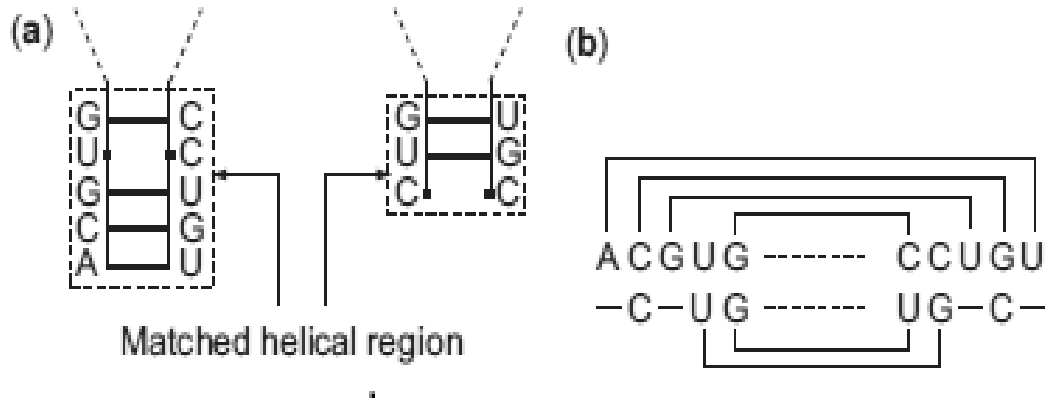
(c) Second sequence Structure.

Base pair insertion

G-U aligned to unpaired nucleotide

PARTS ALIGNMENT

- The alignment of RNA sequences is given below.



SARSA (PARTS)

- Pairwise Alignment of RNA Tertiary Structures
- This tool gives pairwise alignment of RNA tertiary structures.
- This tool converts the 3D structures of RNA to 1D SA (structural alphabet) letters.
- Then it uses classical sequence alignment methods to compare their 1D SA-sequences and find the structural similarities.

INPUT AND OUTPUT

http://bioalgorithm.life.nctu.edu.tw/SARSA/help/examples/parts_show/results.html

PARTS Result(s)

Input RNA 3D Structures

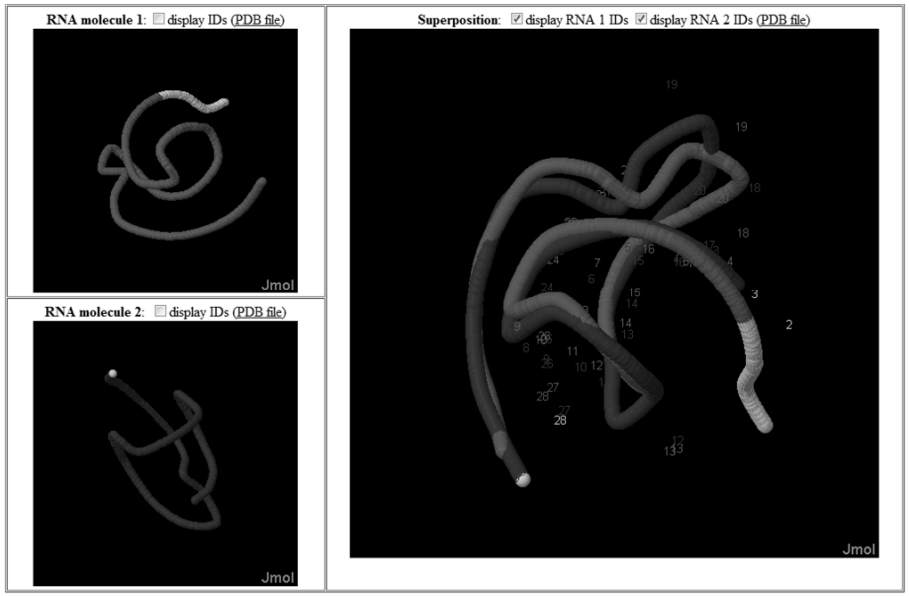
- RNA molecule 1:
  - 1L2X:UR0020 (PDB code:NDB code), Length: 28, Chain ID: A, from 1 to 28, (view Backbone torsions)
- RNA molecule 2:
  - 2A43:UR0066 (PDB code:NDB code), Length: 26, Chain ID: A, from 3 to 28, (view Backbone torsions)

Input Parameters

- Alignment: Global alignment
- Gap open penalty: -5
- Gap extension penalty: -2
- Specified number of suboptimal alignments: 1

```
>Alignment 1
   Alignment score = 75, RMSD =  2.398,  Superposition display
   Alignment of SA-encoded RNA sequences:
      RNA 1     1    FAAAAAAPTAAEVAAAIQDWAAIAMAAP-  28
                        |||||||   ||     |    |
      RNA 2     3    ---TAAAPTAAABLAAARQCRAAIAMAAB  28
   Alignment of original RNA sequences:
      RNA 1     1    GGCGCGGCACCGUCCGCGGAACAAACGG-  28
                        |||||||   ||     |    |
      RNA 2     3    ---GCGGCACCGUCCGCUCAAACAAACGG  28
```

Developed by BioAlgorithm Laboratory, Institute of Bioinformatics & Department of Biological Science and Technology, National Chiao Tung University, Hsinchu, Taiwan



CONCLUSION

- After studying these tools I found:
  - RNA motifs play a very important role in many biological processes.
  - With the help of angle mining we can predict a motif in a given RNA pdb file.
  - We can also classify motifs in a given pdb file by the help of angle mining.
  - Accurate alignment of RNA secondary and tertiary structures would be more significant than sequence alignment.

REFERENCES AND LINKS

- rna.bgsu.edu/FR3D
- pylelab.org/home/index2.html
- R. V. Spriggs, Y. Murakami, H. Nakamura and S. Jones (2009), Protein function annotation from sequence: prediction of residues interacting with RNA
- http://piranha.protein.osaka-u.ac.jp
- Arif Ozgun Harmanci, Gaurav Sharma and David H. Mathews (2008), PARTS: Probabilistic Alignment for RNA joinT Secondary structure prediction
- bioalgorithm.life.nctu.edu.tw/SARSA/index_parts.php