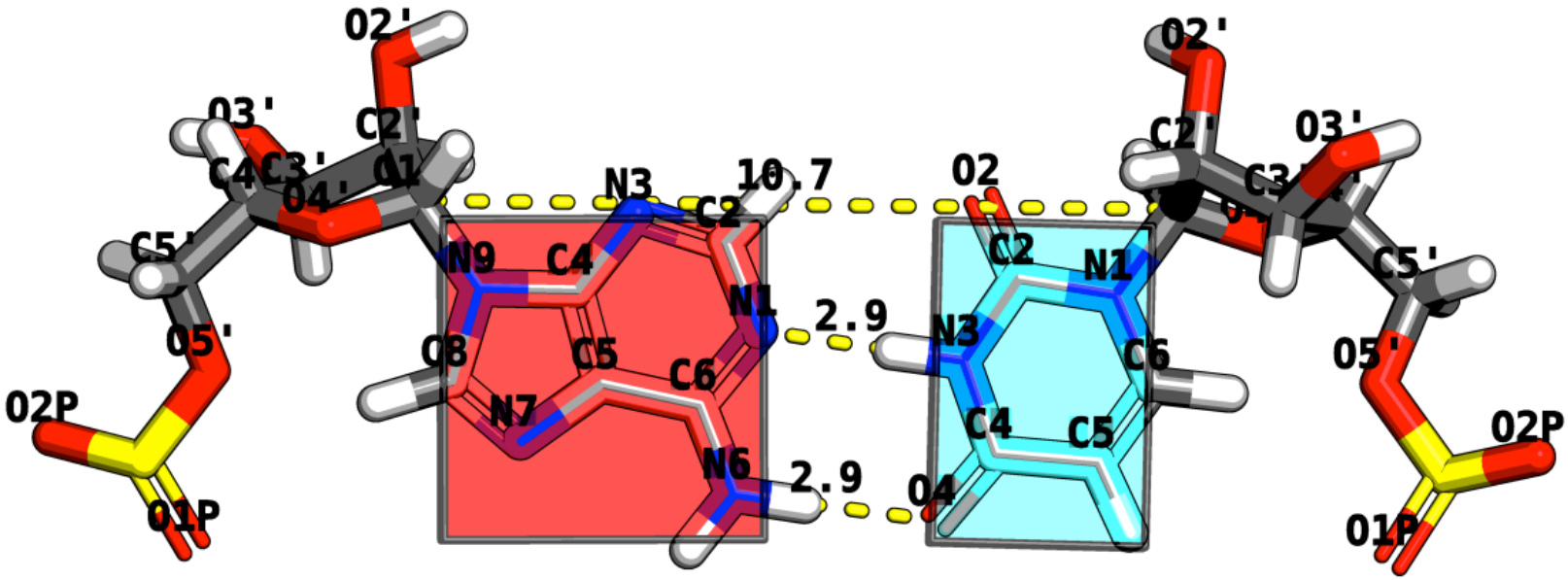
RNA Structure Analysis via the Rigid Block Model



Mauricio Esguerra

October 5th, 2010

Chemistry and Chemical Biology Department

Rutgers, The State University of New Jersey

Program

- **Opening Act**
  - **THE CHALLENGE**: How does RNA fold?
- **Act I**
  - **RNA BASE STEPS**: Dinucleotide step classification using single-stranded base step parameters.
- **Act II.**
  - **RNA BASE PAIRS**: Base pairs in RNA helical regions.
- **Act III**
  - **RNA BASE PAIR STEPS**: From local to global properties in RNA with the aid of rigid-blocks.
- **Act IV**
  - **RNA STRUCTURAL MOTIFS**: RNA structural motifs identification using rigid-blocks.
- **Finale**

Opening ACT

THE CHALLENGE:

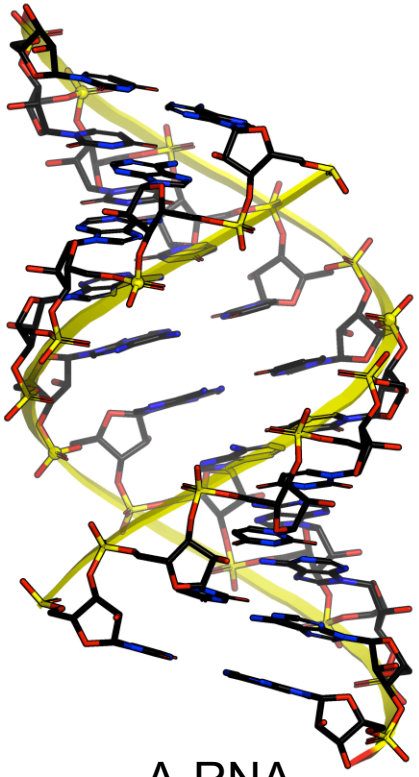
How does RNA fold?

Is this an easy or hard problem?

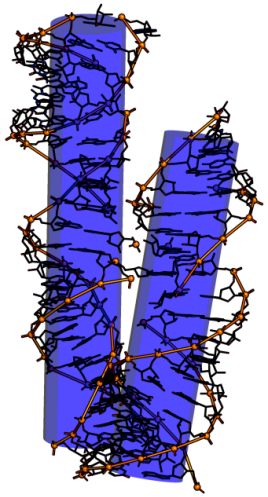
CHALLENGE

What is RNA?

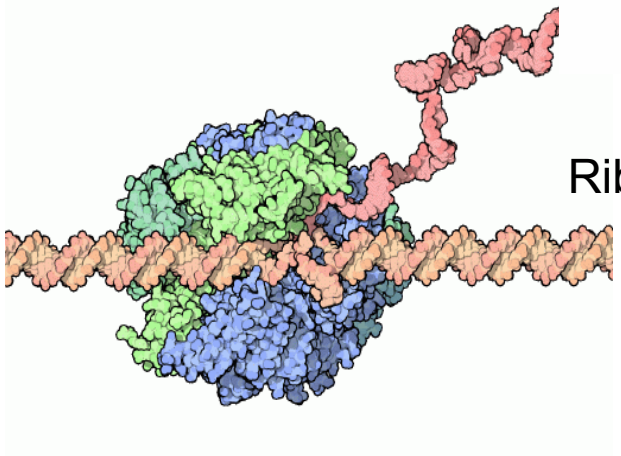
A structure that can give you a Nobel prize or not.



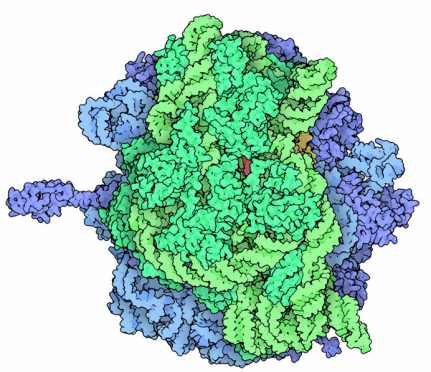
A-RNA
No Nobel



Ribozymes
1989 Nobel



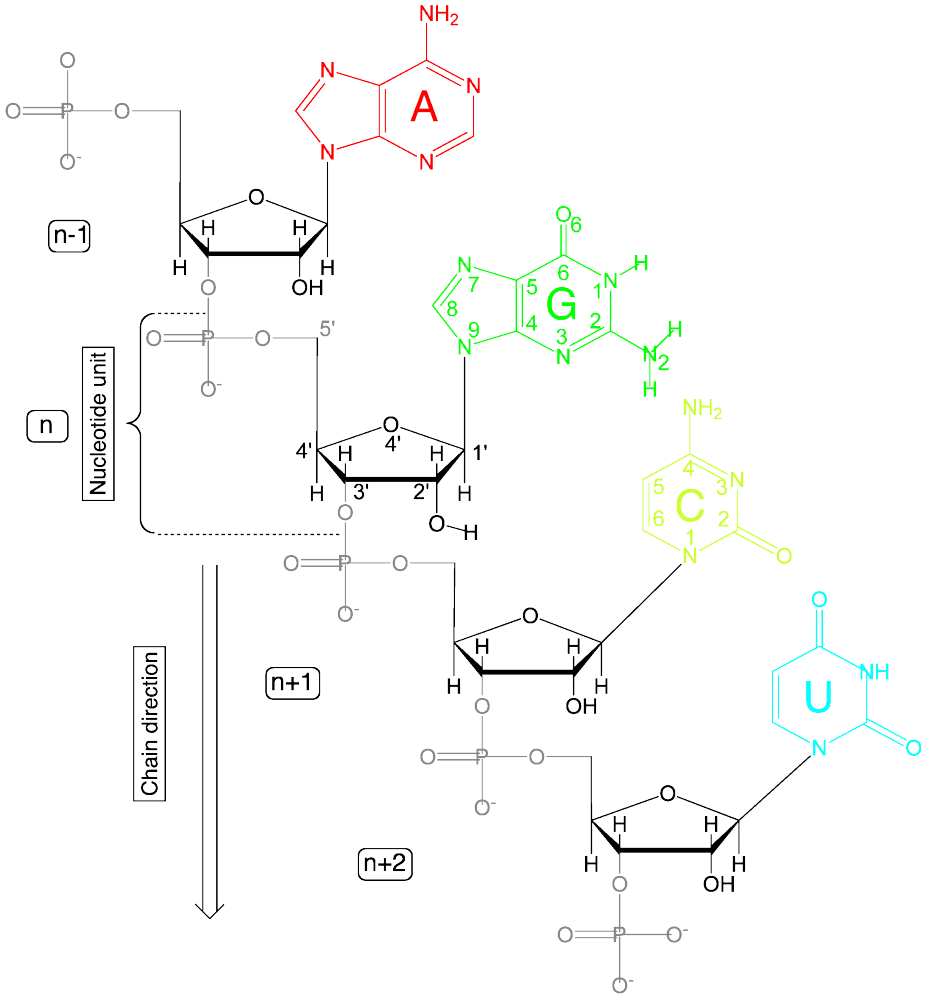
TRANSCRIPTION
RNA Polymerase
mRNA factory
2006 Nobel

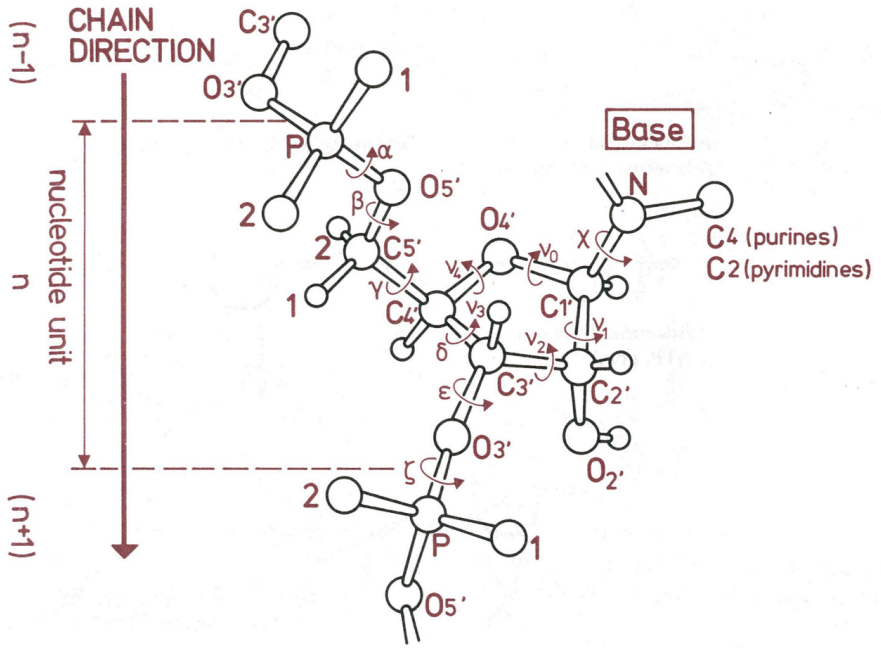


TRANSLATION
Ribosome, aka rRNA
Protein factory
2009 Nobel

CHALLENGE

RNA is composed of units (nucleotides) that have a base, a sugar (ribose), and a phosphate group.





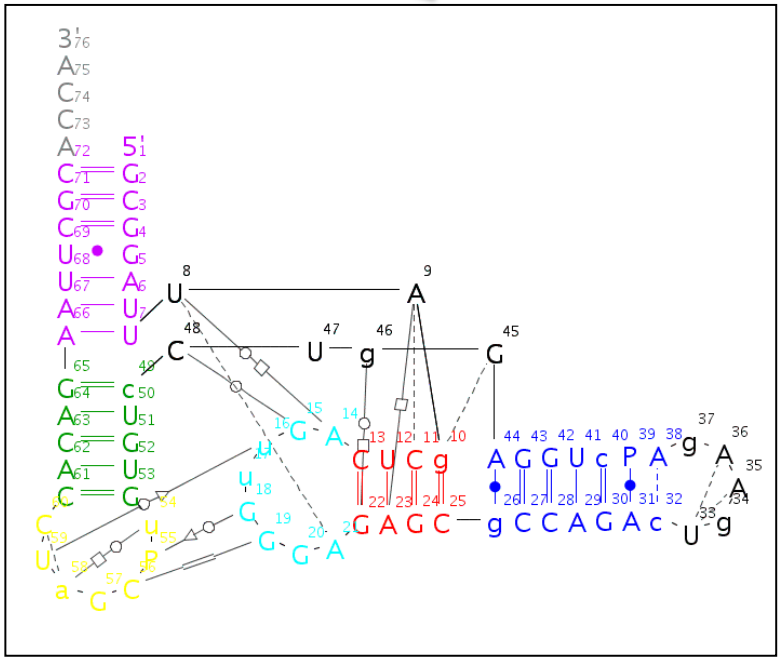
6 Nucleic acid torsion angles.
1 Glycosidic torsion angle.

CHALLENGE

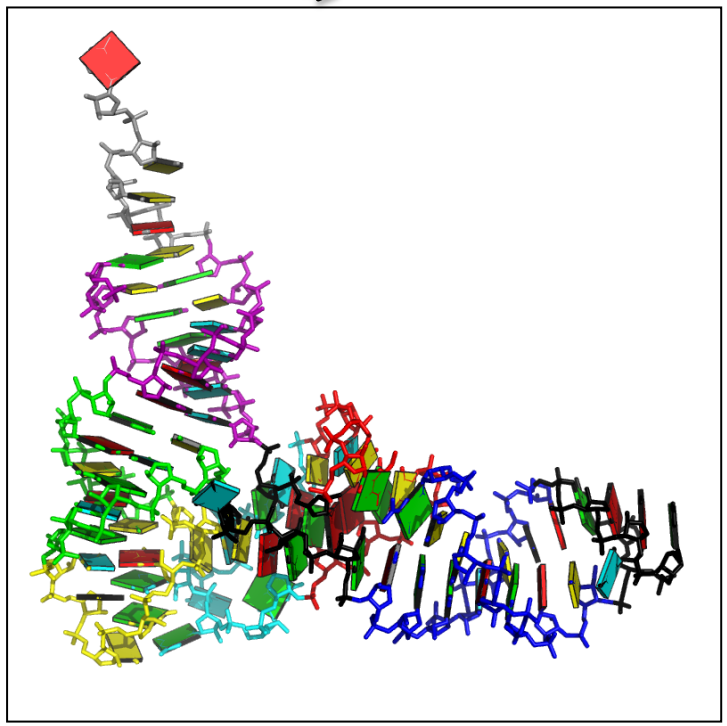
How does the RNA polymer fold?

Primary

GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAGAUCUGGAGGUCCUGUGUUCGAUCCACAGAAUUCGCACCA



Secondary

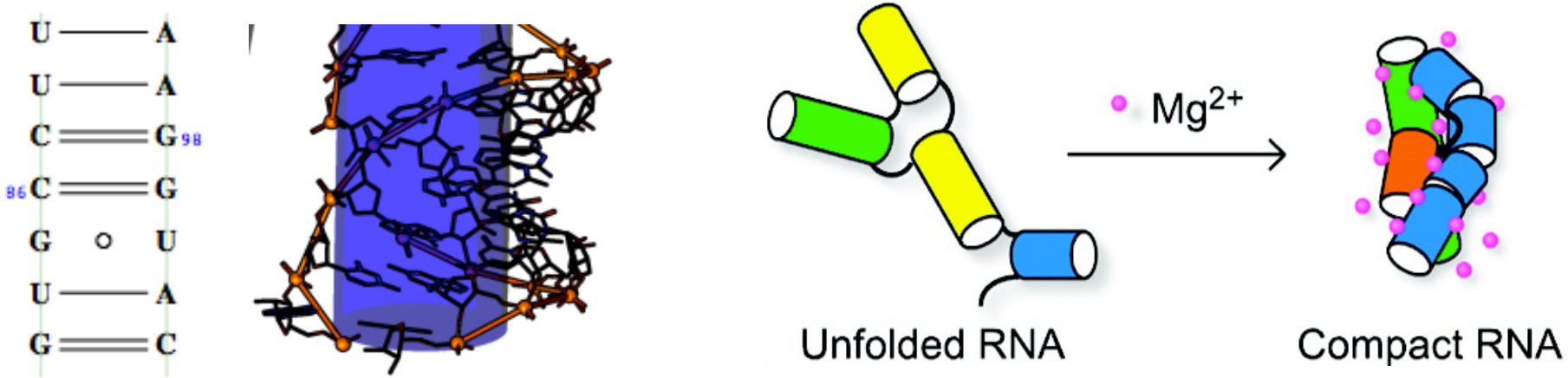


Tertiary

CHALLENGE

Is RNA folding a hard or easy problem to solve? After all, the RNA alphabet has only 4 letters.

- In contrast to proteins, no hydrophobic surface burial.
- Polyanionic sugar-phosphate backbone.
- Secondary and tertiary separated by Mg2+ addition.
- Slow folding compared to protein ~ 1ms.



CHALLENGE

From a structural point of view what can we say or contribute to the RNA folding problem?

- How do RNA structural properties influence RNA folding?
- What can we say about helical regions using common methods used in the Olson lab, e.g. rigid-body models?
- How can we define RNA motifs?
- What are RNA motifs?

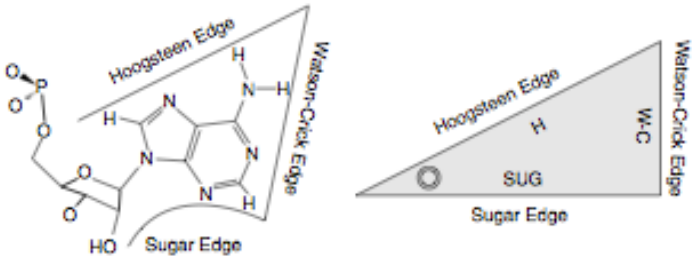
CHALLENGE

Usual ways to classify RNA conformations and an unexplored one -- the rigid-body perspective --.

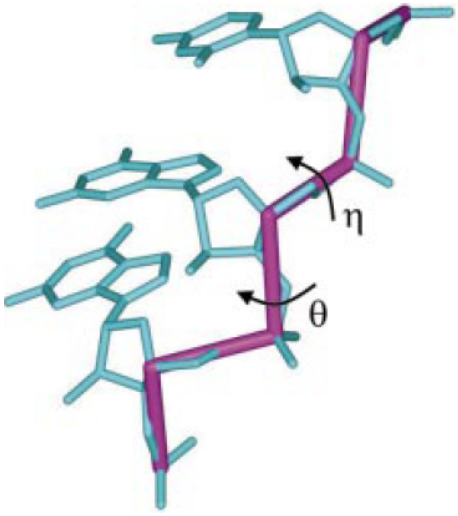
- The atom-based perspective
  - Comparison of backbone atom positions.
  - Comparison from set of backbone, sugar, base, atom positions.
- The bond-based perspective
  - Covalent bonds – backbone + glycosidic torsions.
  - Pseudo-bonds η and θ.
  - Hydrogen bonds forming in edge boundaries (Watson-Crick, Sugar, and Hoogsten).
- The rigid-body-based perspective
  - Base-pair parameters.
  - Base-pair (base) step parameters.

CHALLENGE: Bond perspectives

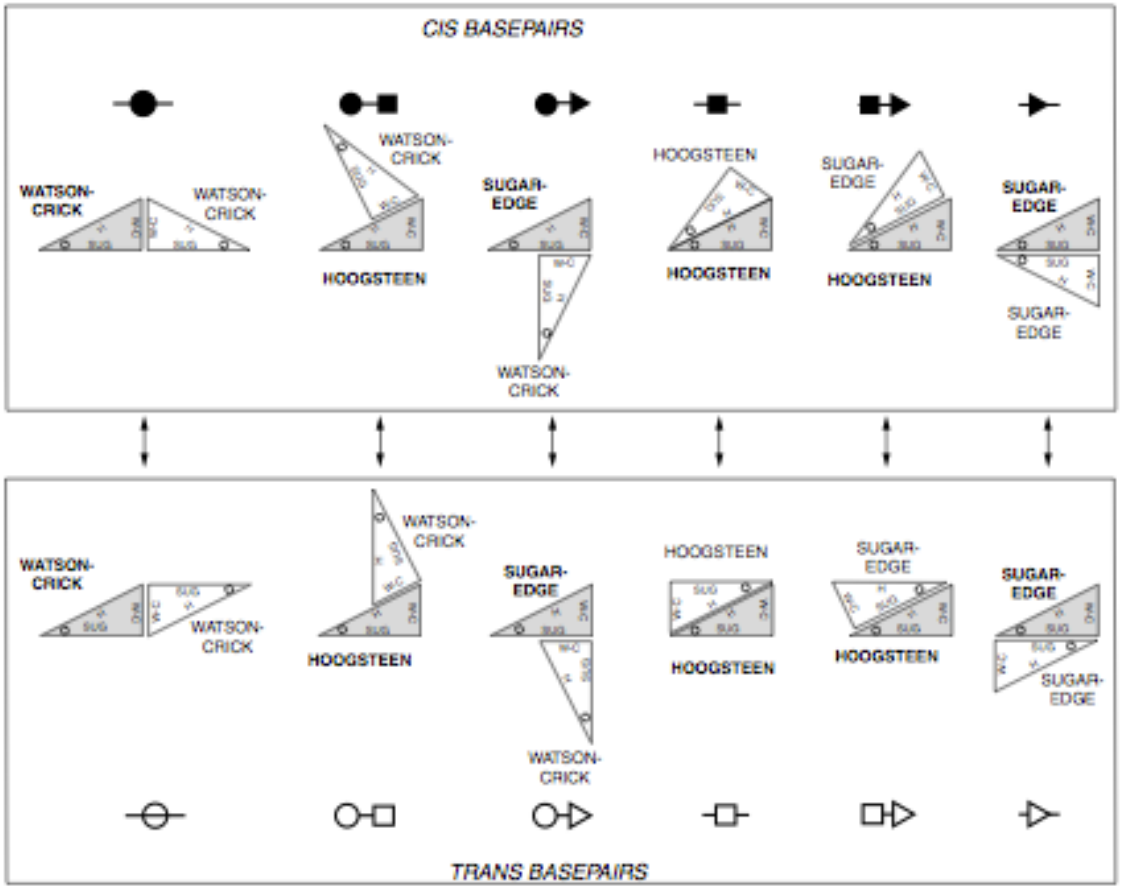
Leontis-Westhof base-pair classification based on base boundaries. Pseudo-bond torsion angles.



Leontis-Westhof boundaries



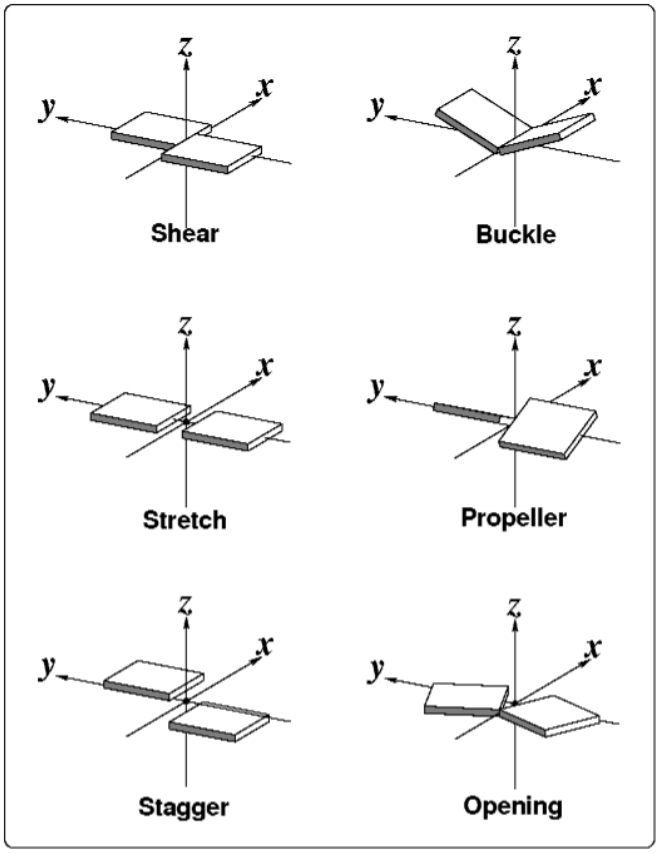
Torsion angles of pseudo-bonds



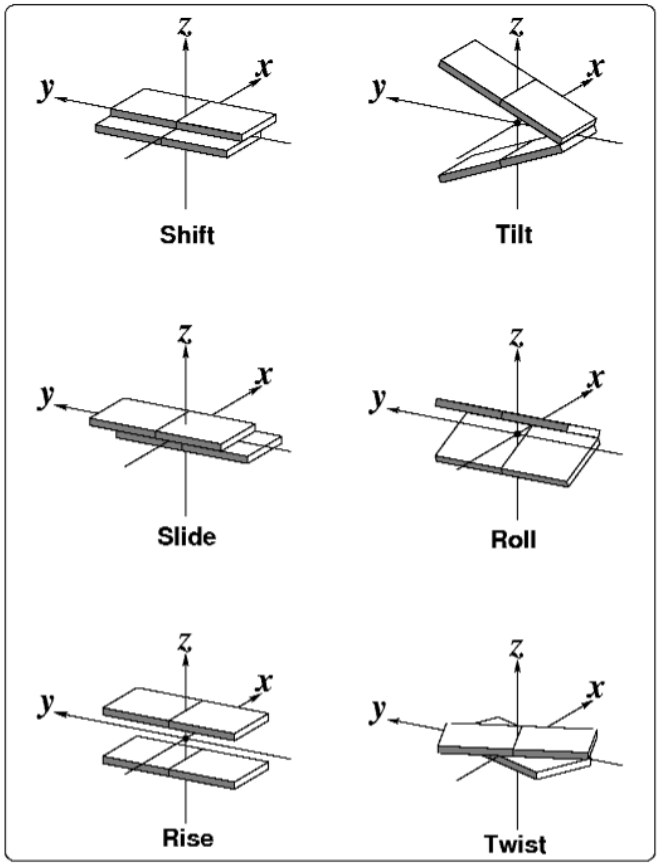
CHALLENGE

The rigid-body perspective: Nucleic acid bases and base-pairs represented as blocks.

Base-pair parameters



Base-pair-step parameters



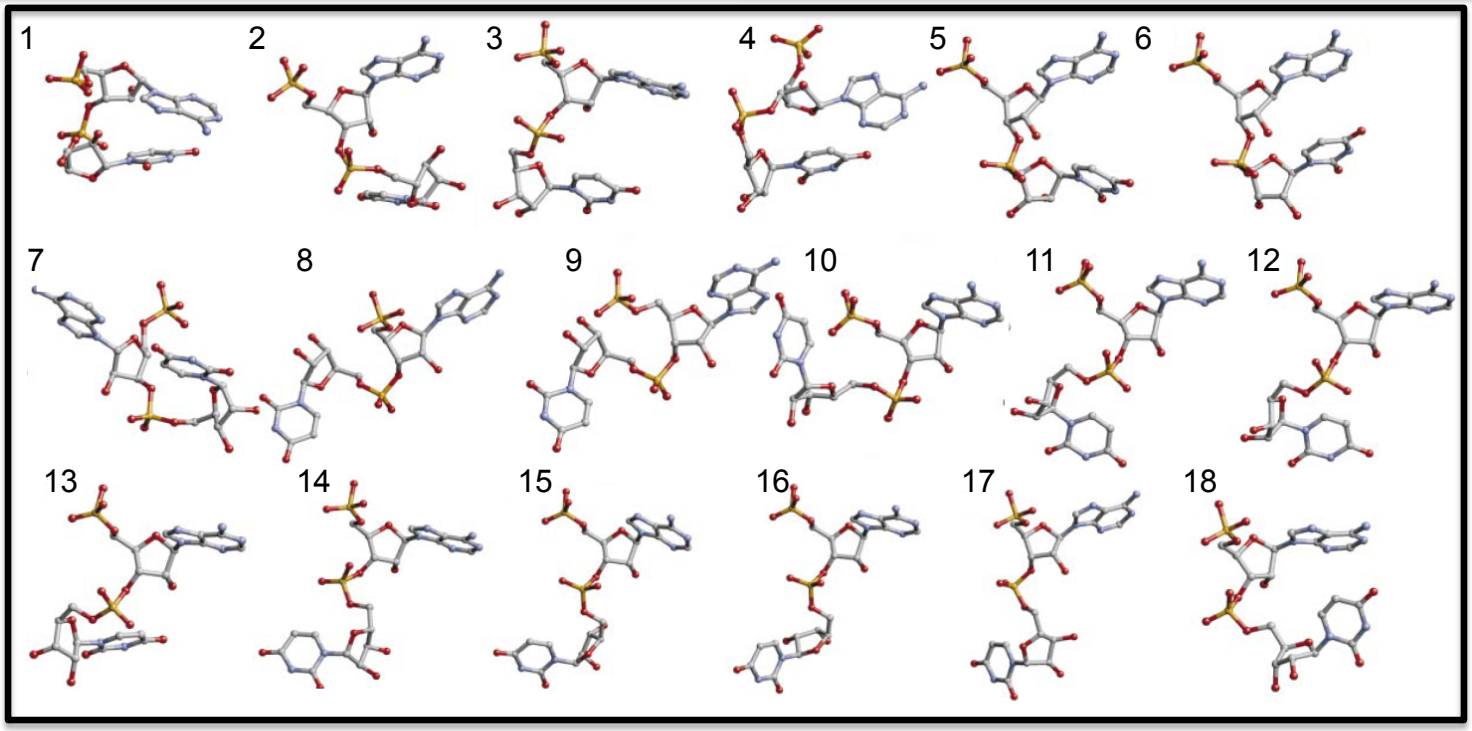
ACT I

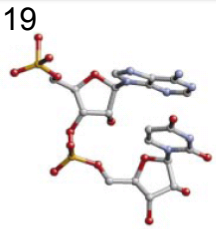
Dinucleotide step classification using single-stranded base step parameters

I. BASE STEPS

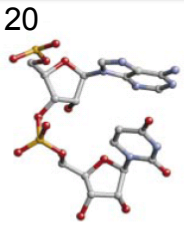
18 non-A-RNA, AII-RNA, and canonical A-RNA conformers kindly provided by Berman and collaborators.

Non A-RNA





AII-RNA



A-RNA

Schneider et al. (2004) *NAR*, **32**, 1667-1677.

Simple example on hierarchical clustering. Consensus clustering uses idea of branch consensus.

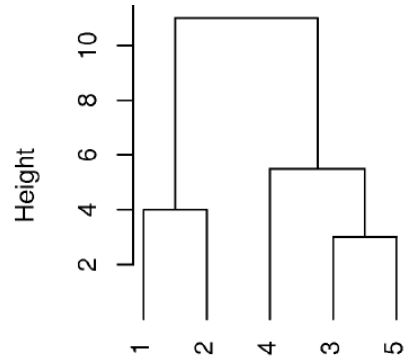
Structure	Property I	Property II
1	1.00	5.00
2	-2.00	6.00
3	2.00	-2.00
4	-2.00	-3.00
5	3.00	-4.00

Dataset

$$d(X,Y) = \begin{vmatrix} & 1 & 2 & 3 & 4 & 5 \\ 1 & 0 & & & & \\ 2 & 4 & 0 & & & \\ 3 & 8 & 12 & 0 & & \\ 4 & 11 & 9 & 5 & 0 & \\ 5 & 11 & 15 & 3 & 6 & 0 \end{vmatrix}$$

Distance matrix (manhattan)

$$D(\{3,5\},1) = \frac{1}{2*1} * (8+11) = 9.5$$
$$D(\{3,5\},2) = \frac{1}{2*1} * (12+15) = 13.5$$
$$D(\{3,5\},4) = \frac{1}{2*1} * (5+6) = 5.5$$

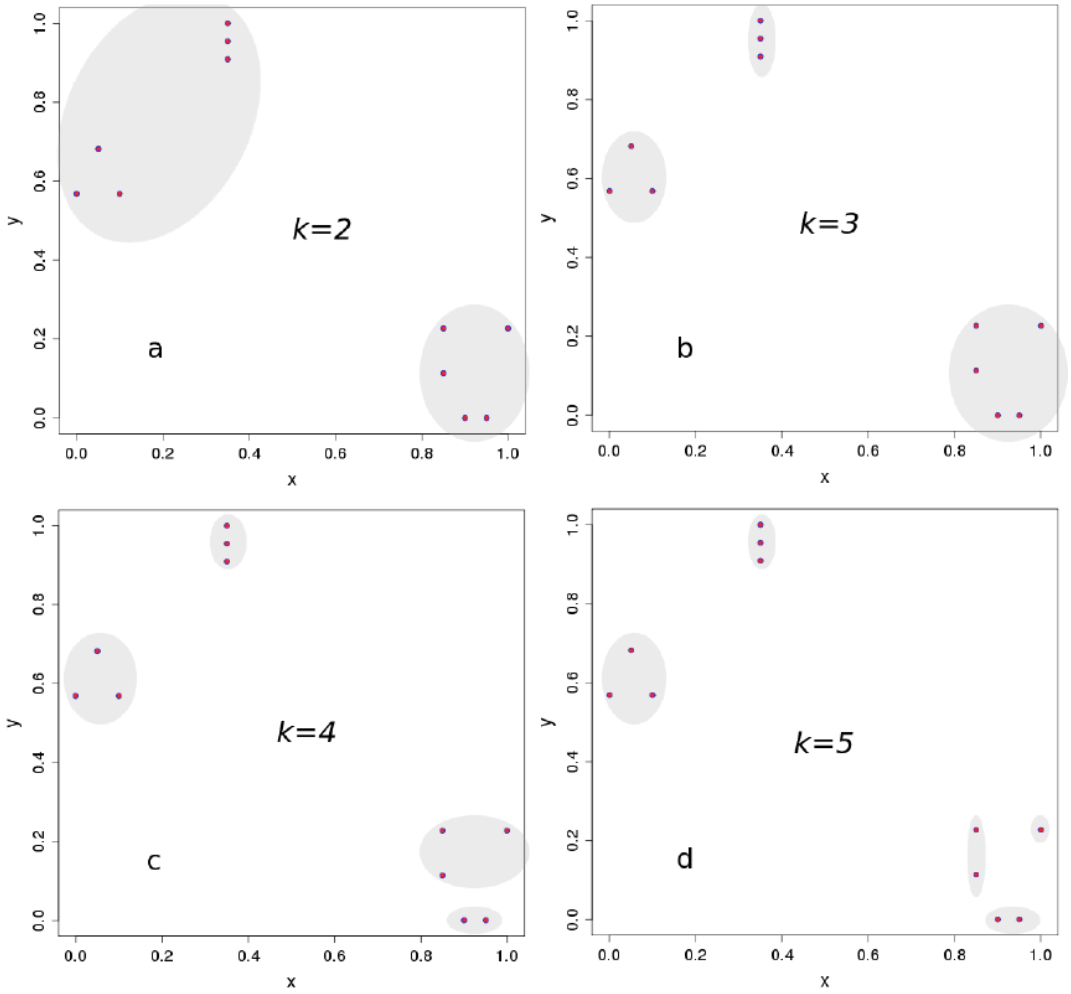


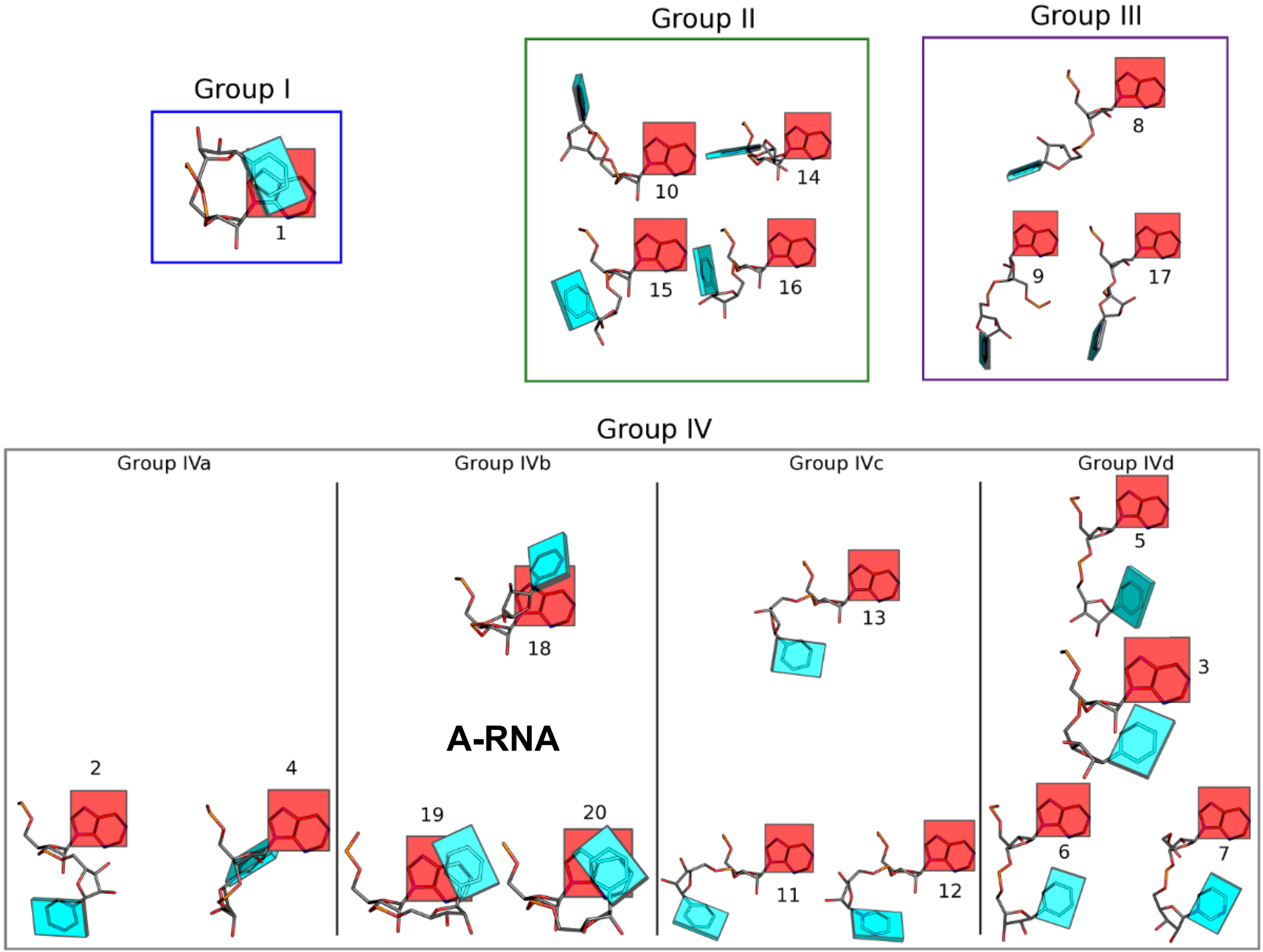
Group using a method. (average linkage)

Validation using scores which quantify compactness, connectedness, and separation.

k=3 is more compact and separated than others. Scores such as Dunn Index and *ASW*, quantify this.

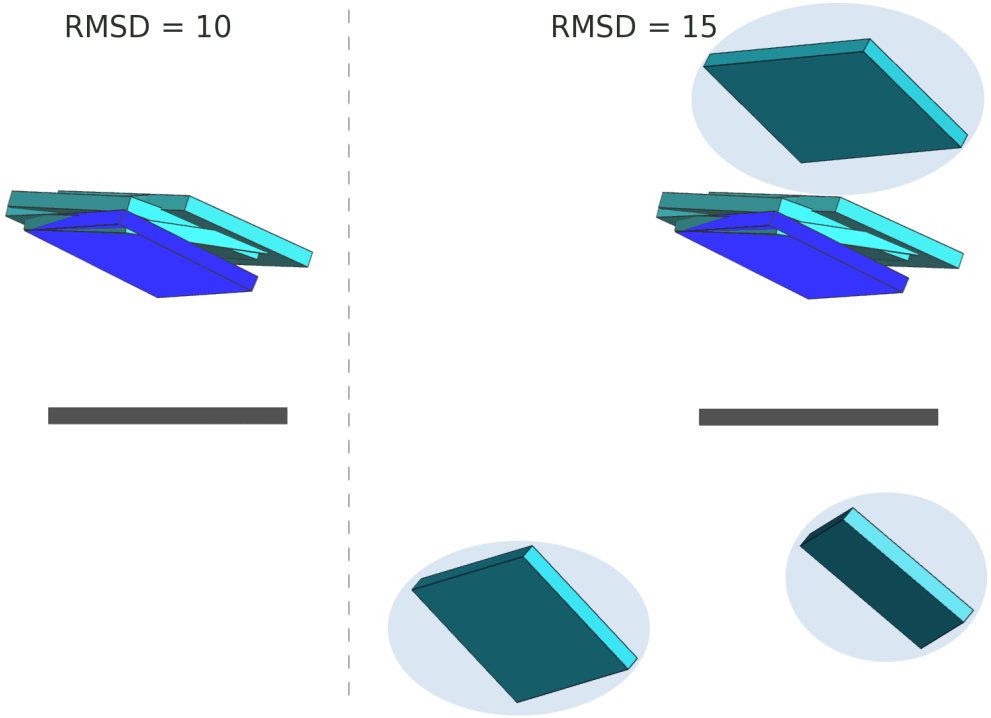
k=2 is more connected. Conectedness decreases progressively as the number of clusters increases.




Group I
1
Group II
10
14
15
16
Group III
8
9
17
Group IV
Group IVa
2
4
Group IVb
18
A-RNA
19
20
Group IVc
13
11
12
Group IVd
5
3
6
7

I. BASE STEPS

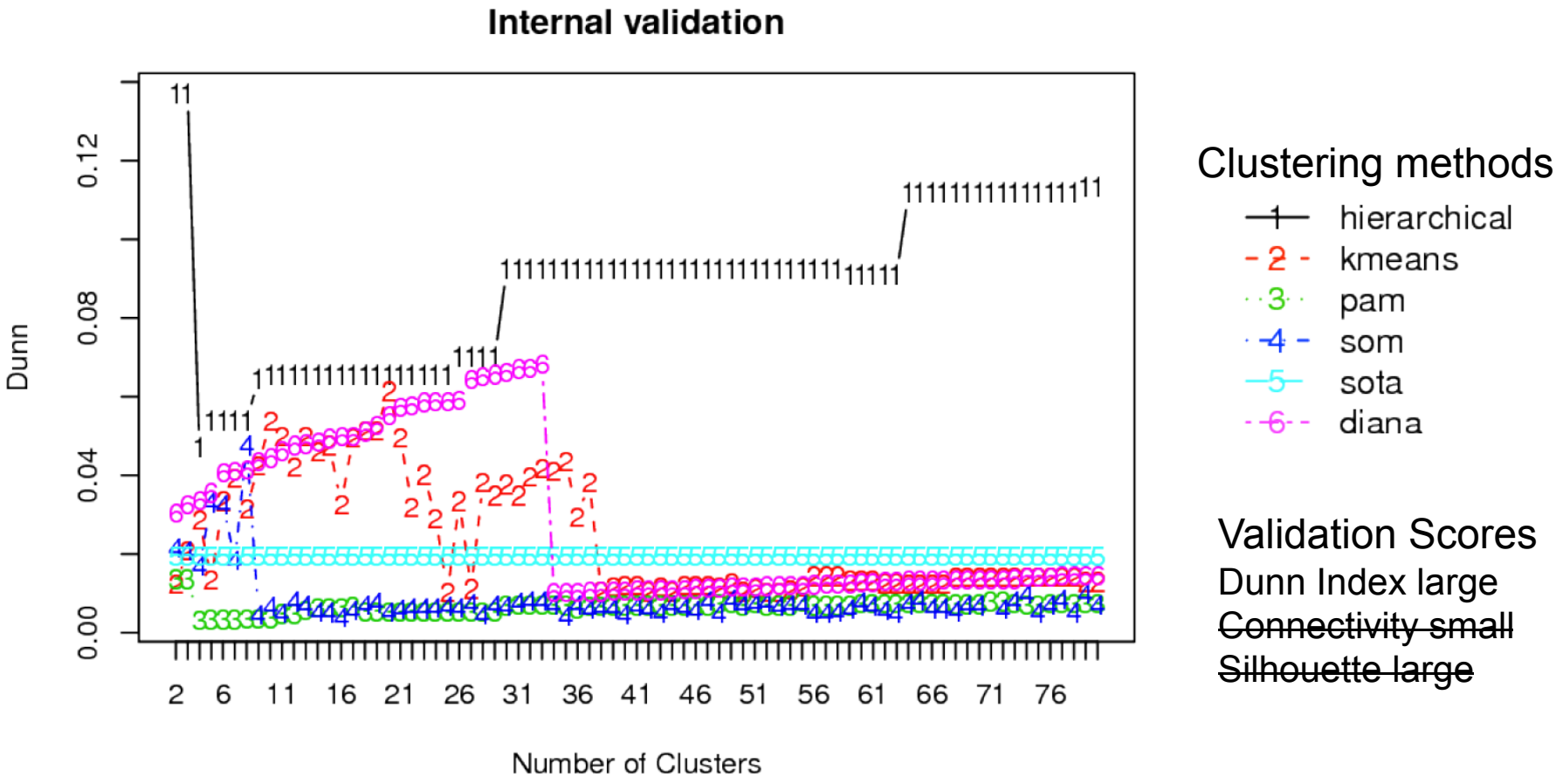
Only a small fraction ~ 30 % of dinucleotide step conformations are represented in the ribosome.



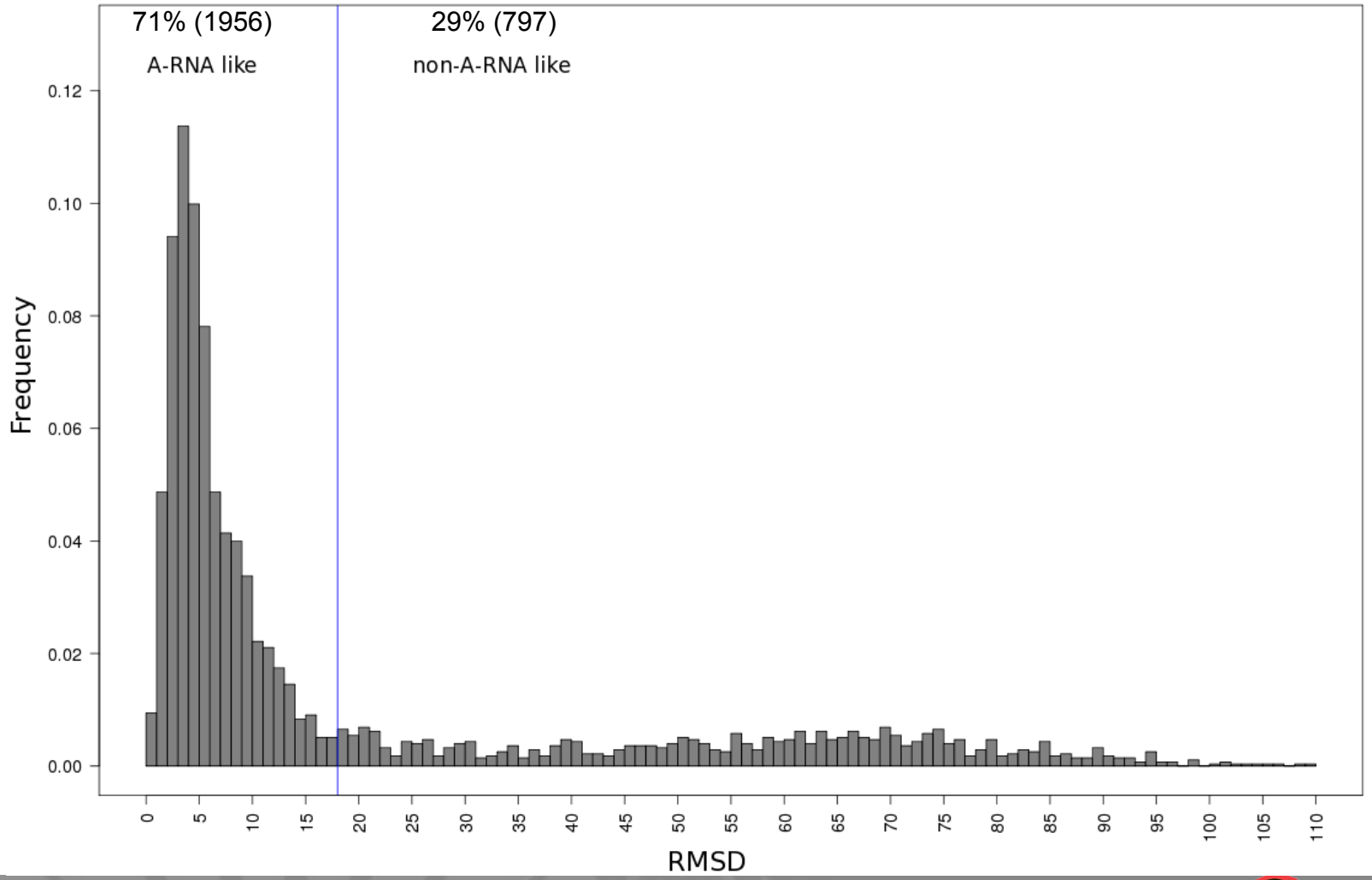
	RMSD Cutoff Value		
Group	10	15	20
I	3 (0.11)	7 (0.25)	7 (0.25)
II	5 (0.18)	13 (0.47)	25 (0.91)
III	1 (0.04)	5 (0.18)	23 (0.84)
IVa	1 (0.04)	1 (0.04)	7 (0.25)
IVb	807 (29.31)	1696 (61.61)	1965 (71.38)
IVc	9 (0.33)	22 (0.80)	41 (1.49)
IVd	35 (1.27)	99 (3.60)	191 (6.94)
Total	861 (31.28)	1843 (66.95)	2259 (82.06)

In parentheses percentages of confomer groups found in the 23S subunit of the ribosome, PDB_ID: 1JJ2

Instead of picking any clustering methodology we use the clValid R package for cluster validation.

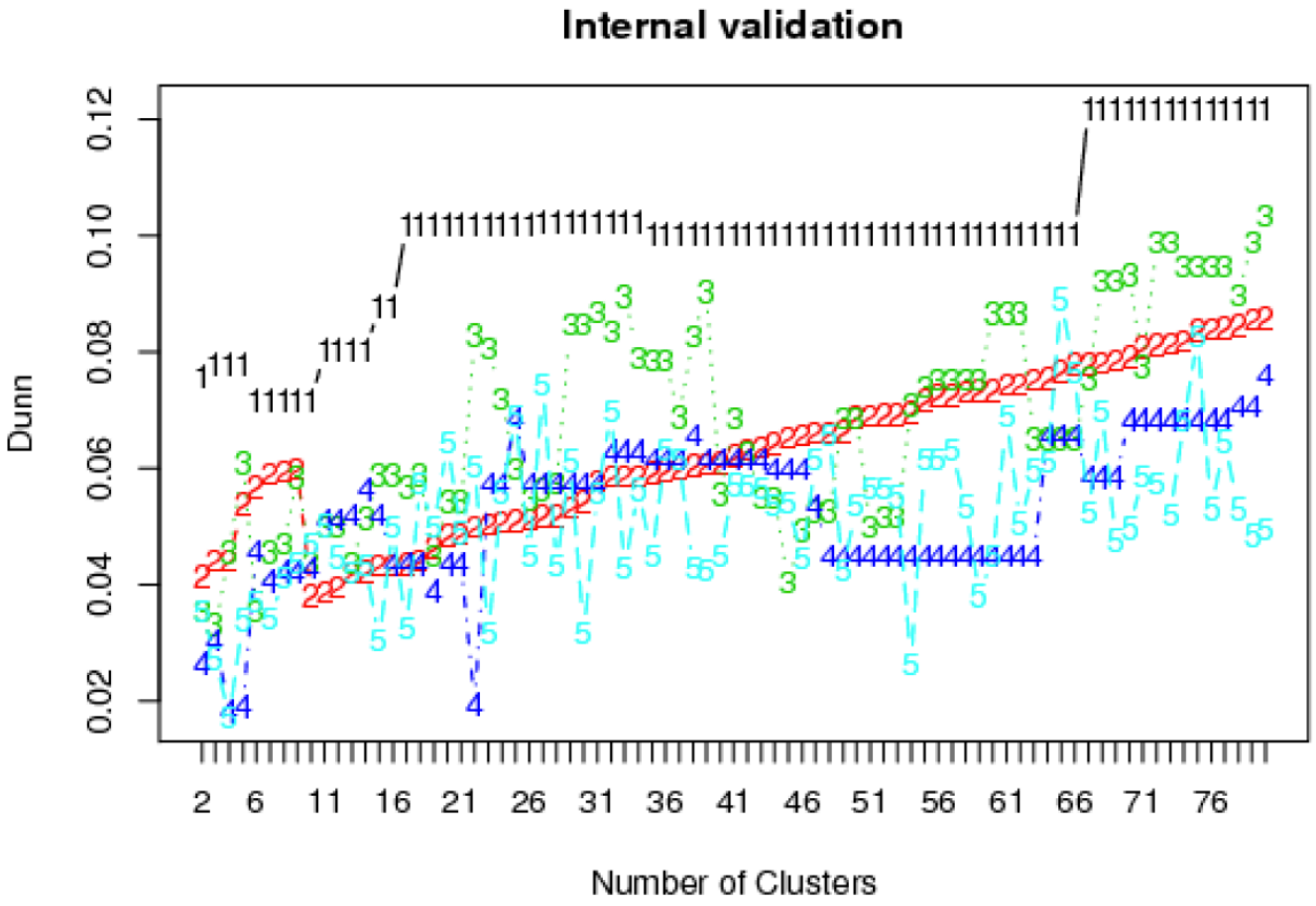


Selecting A-RNA and non-A-RNA-like datasets using distance to canonical A-RNA parameters.



I. BASE STEPS

For the non-A-RNA-like dinucleotide steps the best clustering method is the hierarchical one. k =67.



Clustering methods

- hierarchical
- kmeans
- pam
- som
- sota

Validation Scores

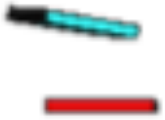
Dunn Index large

~~Connectivity small~~

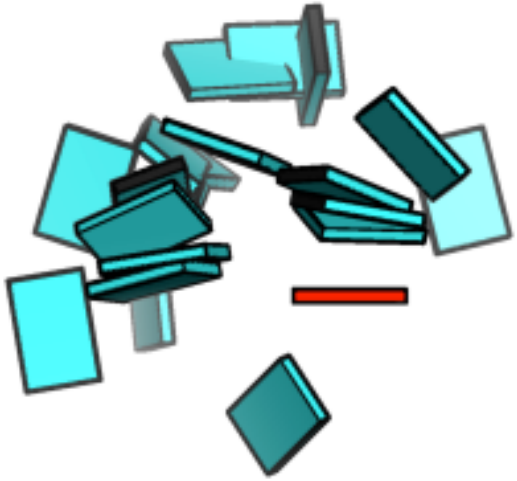
~~Silhouette large~~

Non-A-RNA-like torsion angle conformers do not fill the full space of base geometries.

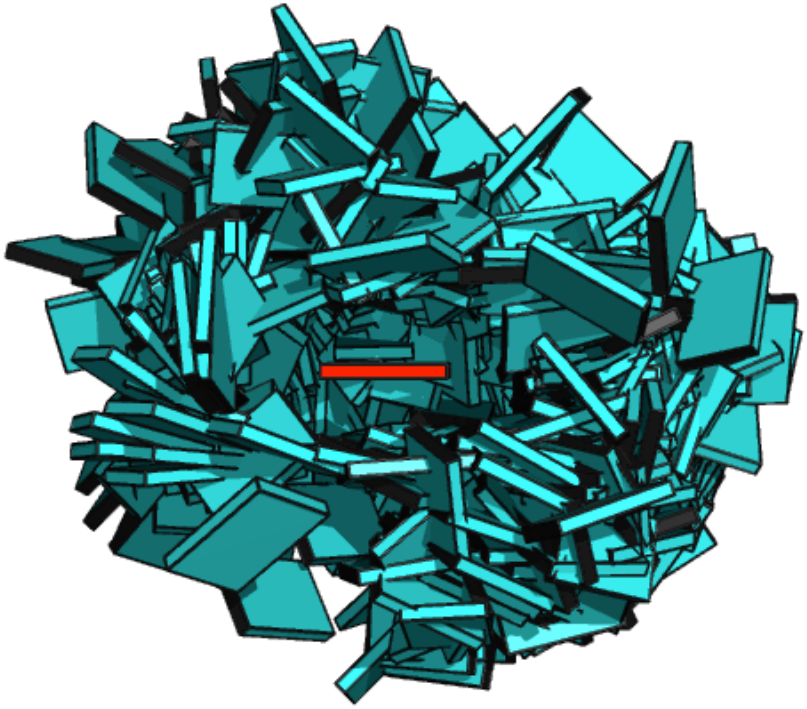
A-RNA



17 non-A-RNA-like

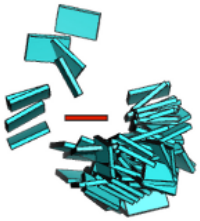


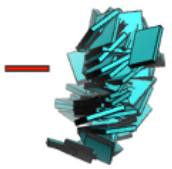
All non-ARNA

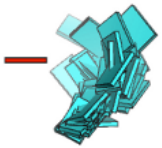


797 (~30% of 23S rRNA)

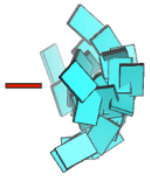
17 Main Non-ARNA-Like Clusters with ten or more members compose 80% of all non-A-RNA steps.

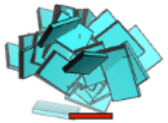

g2(12%)


g5(11%)


g3(8%)

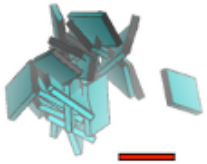

g4(7%)

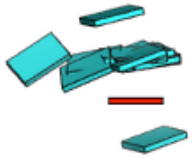

g7(6%)

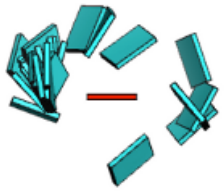

g1(6%)

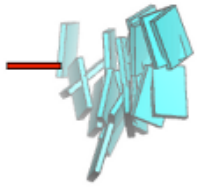

g9(5%)


g16(4%)

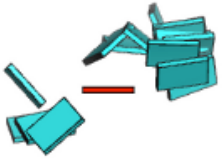

g11(4%)

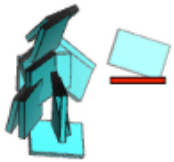

g14(4%)

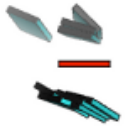

g13(3%)

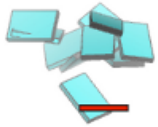

g17(3%)


g8(2%)


g22(2%)


g33(2%)

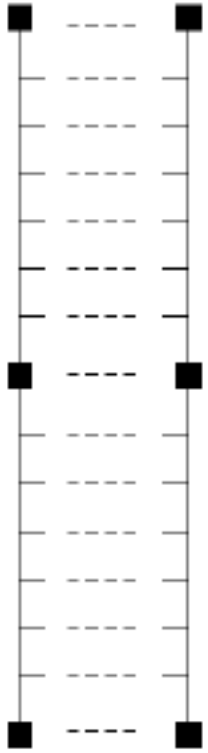

g23(1%)


g32(1%)

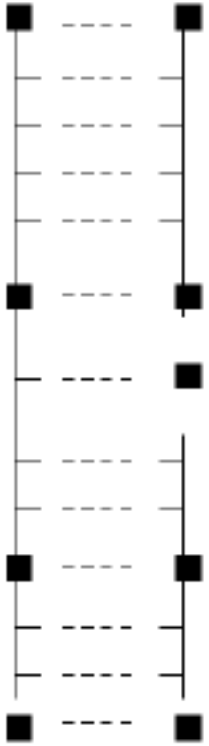
ACT II

Base-pairs in RNA helical regions

Helical regions in 3DNA can be continuous and quasi-continuous.



Continuous helical region



Quasi-continuous helical region

II. BASE PAIRS

Partially non-redundant RNA dataset is GC rich.

		Percent Composition			
RNA Type	Counts	G	C	A	U
small helices	78	36	30	16	18
drug-RNA	36	36	33	14	17
protein-RNA	207	37	32	16	16
protein-tRNA	9	34	30	19	17
rRNA	13	37	28	18	17
tRNA	13	34	27	21	19
ribozyme	113	34	29	20	16
Total	469	36	30	18	16

Helices composed of 3 base-pairs or more

II. BASE PAIRS

Seven predominant base-pairs in RNA helical regions make up to 90% of all base-pairs.

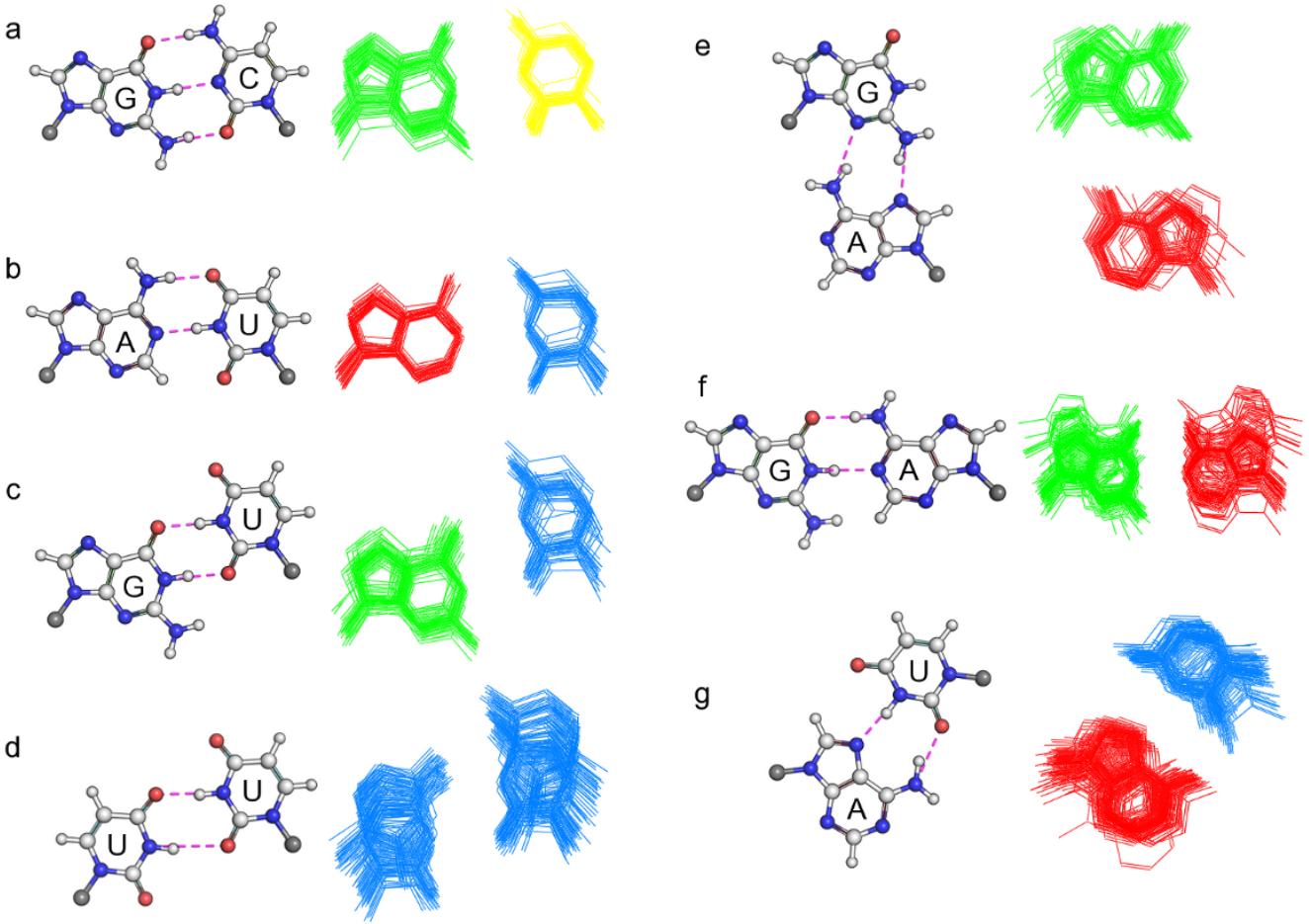
Base-pair		Hydrogen bonds		Number
Canonical				
G·C	Watson-Crick	N2-H···O2	2.79(0.17)	$9500_{x0.90}$
		O6···H-N4	2.92(0.18)	
		N1-H···N3	2.89(0.13)	
A·U	Watson-Crick	N1···H-N3	2.84(0.14)	$3069_{x0.93}$
		N6-H···O4	2.97(0.18)	
Non-canonical				
G·U	Wobble	N1-H···O2	2.79(0.16)	$1049_{x0.69}$
		O6···H-N3	2.85(0.16)	
G·A	Sheared	N2-H···N7	2.89(0.17)	$509_{x0.59}$
		N3···H-N6	3.03(0.18)	
A·U	Hoogsteen	N6-H···O2	2.91(0.21)	$354_{x0.71}$
		N7···H-N3	2.90(0.17)	
G·A	Watson-Crick	N1-H···N1	2.84(0.17)	$185_{x0.85}$
		O6···H-N6	2.91(0.20)	
U·U	Wobble	O2···H-N3	2.95(0.24)	$141_{x0.54}$
		N3···H-O4	2.87(0.15)	

In subscript the percentage of base-pair structures which comply to the hydrogen bonding pattern shown in column II.

In parentheses the standard deviations for the atomic distances of the hydrogen bonding patterns.

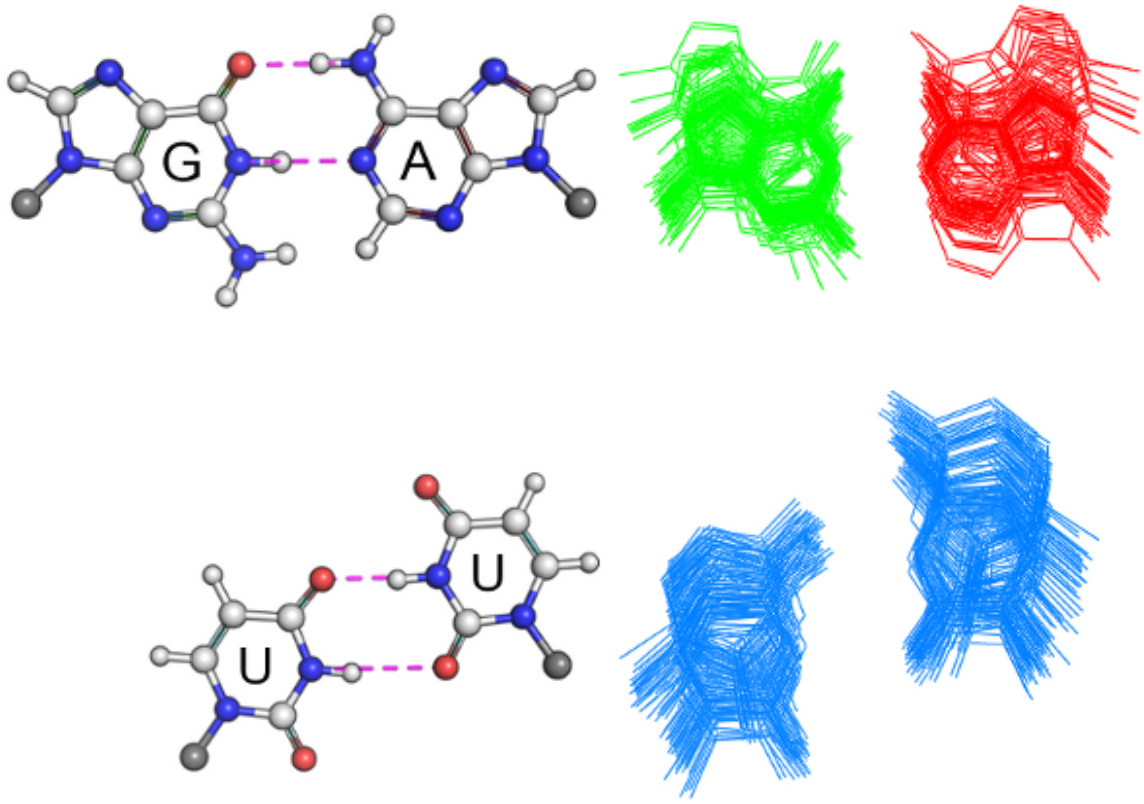
II. BASE PAIRS

Non-canonical base-pairs are more deformable than canonical base-pairs.



II. BASE PAIRS

Most deformed base-pairs seem as oversized or undersized pieces in a puzzle.

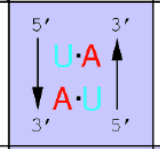
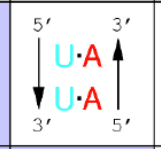
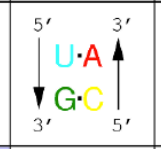
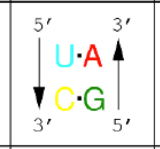
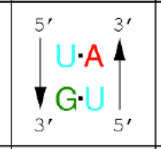
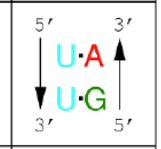
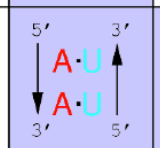
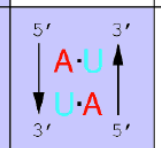
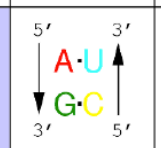
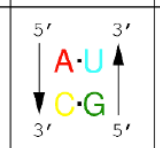
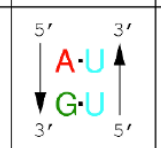
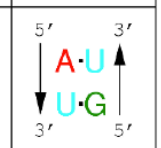
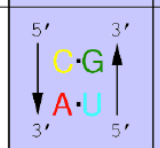
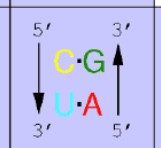
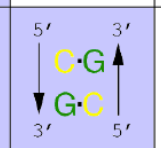
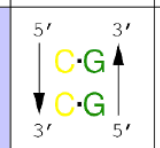
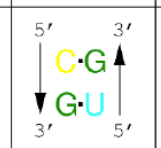
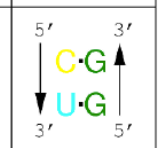
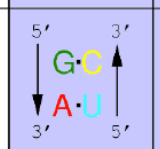
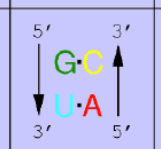
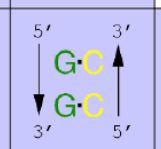
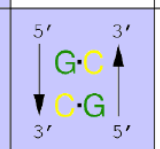
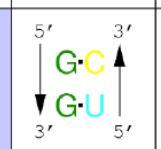
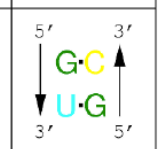
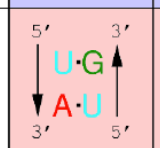
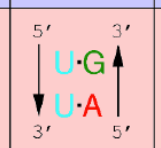
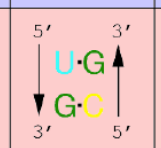
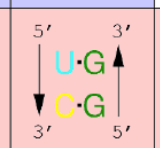
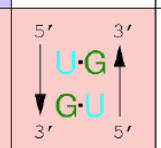
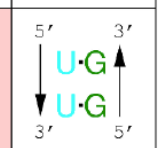
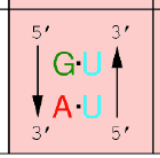
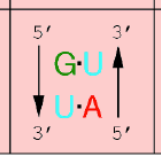
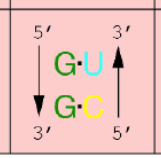
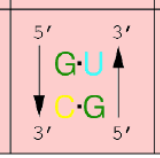
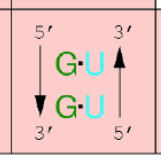
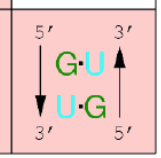


ACT III

Base-pair-steps in RNA helical regions and RNA global properties

III. BASE PAIR STEPS

21 unique base-pair steps formed by canonical Watson-Crick base-pairs and G·U wobble.

i \ i+1	A·U	U·A	G·C	C·G	G·U	U·G
U·A						
A·U						
C·G						
G·C						
U·G						
G·U						

RNA base-pair-steps in helical regions are still few when non-canonicals are considered.

Not all possible combinations of base-pairs into base-pair steps are found in the studied helical regions.
Sheared A·G and U·Uw stand-out.

C·G$_{WC}$	G·C$_{WC}$	U·A$_{WC}$	A·U$_{WC}$	U·G$_{w}$	G·U$_{w}$	A·G$_{s}$	G·A$_{s}$	U·A$_{H}$	A·U$_{H}$	U·U$_{w}$	A·G$_{WC}$	G·A$_{WC}$	bp_i \ bp_{i+1}
604	1335	747	574	77	192	66	–	–	–	18	4	5	G·C$_{WC}$
	608	511	572	161	252	33	–	–	–	69	20	5	C·G$_{WC}$
		97	249	20	45	7	–	–	–	2	–	3	A·U$_{WC}$
			126	48	79	6	–	–	–	20	1	14	U·A$_{WC}$
				31	42	32	1	–	–	5	–	–	G·U$_{w}$
					11	–	–	–	–	–	4	7	U·G$_{w}$
						–	13	–	–	6	–	–	G·A$_{s}$
							20	7	2	–	–	–	A·G$_{s}$
								–	–	–	–	–	A·U$_{H}$
									–	–	–	–	U·A$_{H}$
										3	–	–	U·U$_{w}$
											–	1	G·A$_{WC}$
												–	A·G$_{WC}$

Number of base-pair steps in RNA helical regions

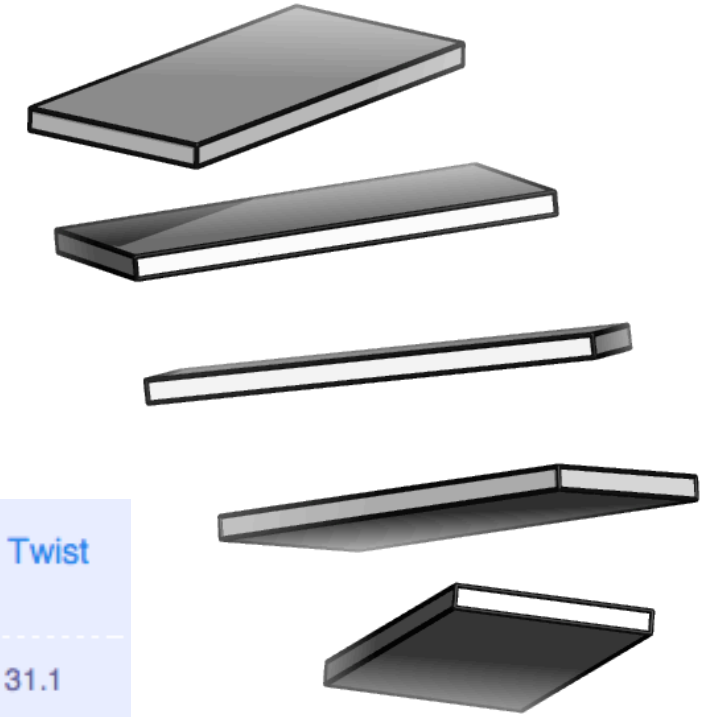
RNA base-pair-steps parameters in helical regions and energy contours. The rnasteps website.

http://rnasteps.rutgers.edu

Intermezzo

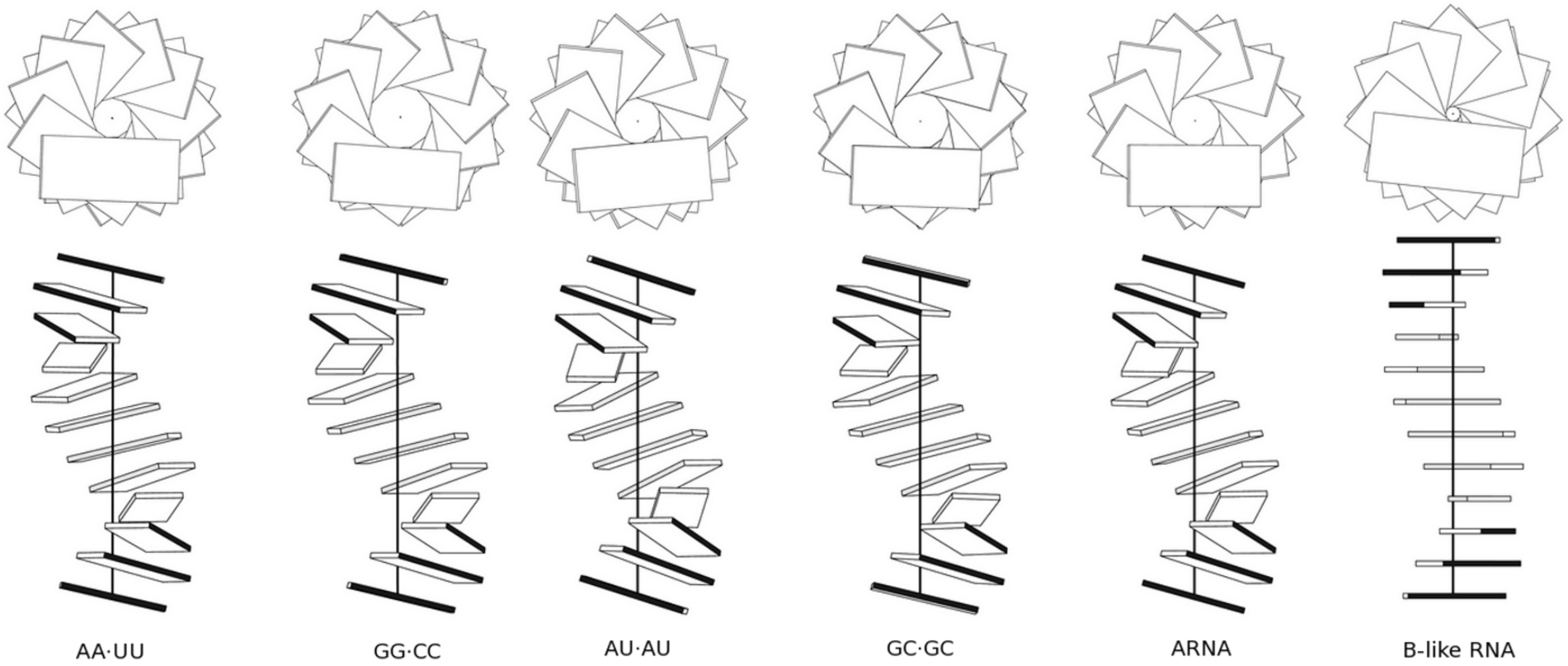
Knowledge of adjacent base-pair steps makes it possible to make up a larger polymer.

With average base-pair-step parameter values from **RNAsteps.rutgers.edu** we can model larger polymers.



Step	Count	Shift	Slide	Rise	Tilt	Roll	Twist
GG.CC	1274	-0.01	-1.85	3.30	0.0	7.4	31.1
UG.CA	700	0.03	-1.59	3.16	0.2	10.6	30.7

RNA sequence has subtle effects on the structure of RNA.



Homopolymers

Block Copolymers

Global properties of polymer chains. Persistence length a.

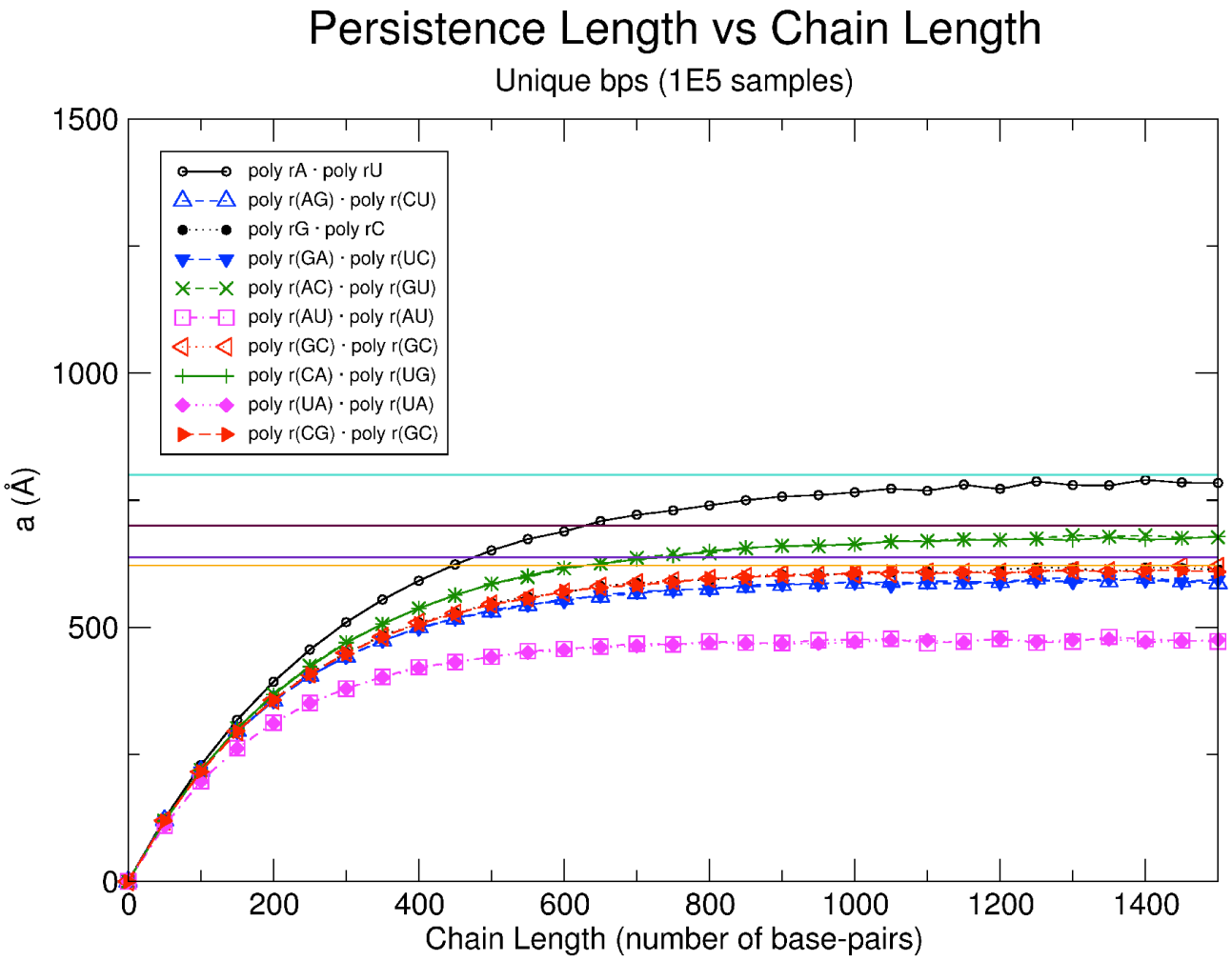
Porod-Kratky

Persistence length is "the average sum of the projections of all bonds $j \geq i$ on bond i in an indefinetely long chain. The bond i is taken to be remote from either end of the chain, i.e., $1 \ll i \ll n$". Paul J. Flory, Statistical Mechanics of Chain Molecules. 1969

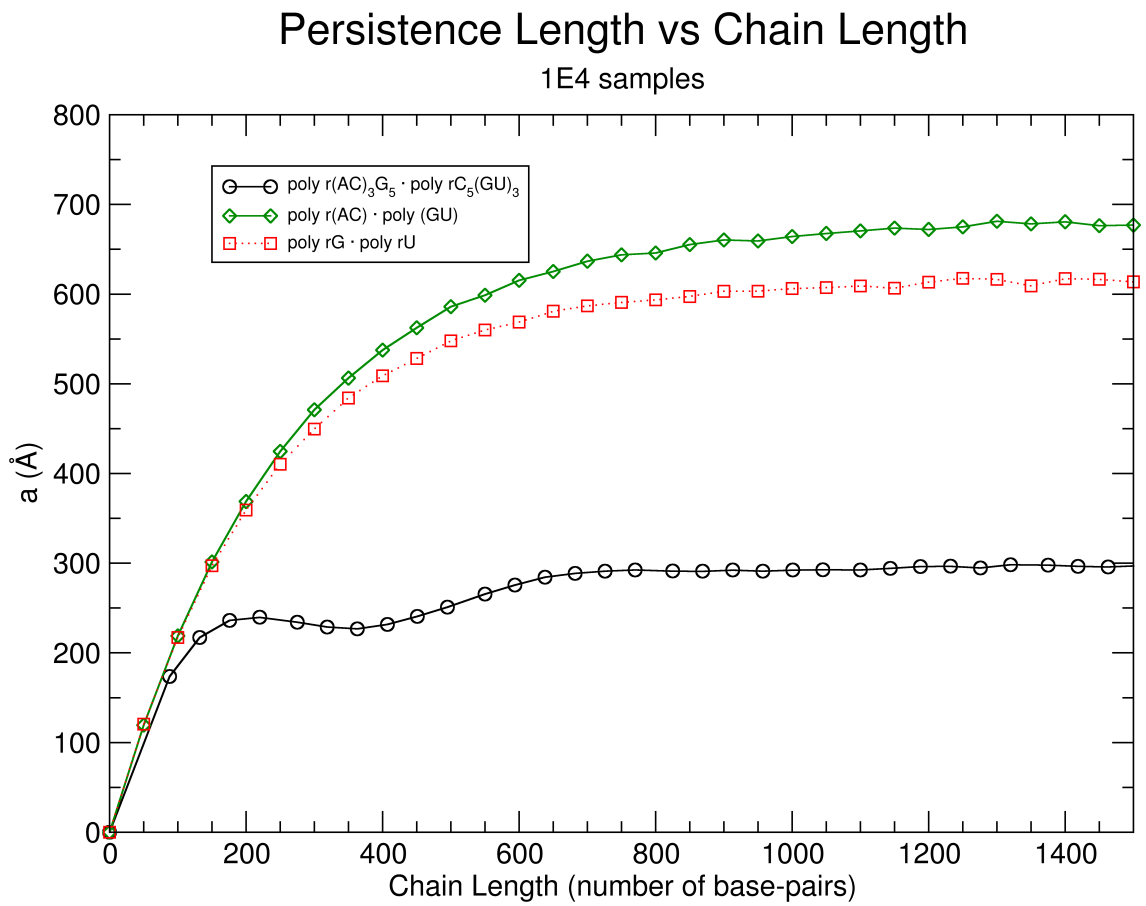
Polymer	a (nm)	Citation
Polymethylene	0.6	Flory[a]
Polystyrene	0.9	Flory[a]
Polyglycine	0.6	Flory [b]
Poly-L-alanine	2	Flory [b]
Poly-L-proline	22	Cantor and Schimel [11]
B-DNA	53	Rivetti [12]
A-RNA	62-64	Abels [13]
α-helix	80-100	Lakkaraju [14]
Coiled-coil	150-300	Lakkaraju [14]
Neurofilament	500	Nelson [2]
Intermediate filament	1000	Lakkaraju [14]
F-actin	17000	Lakkaraju [14]
Microtubule	5200000	Lakkaraju [14]

There might be sequence dependence on the persistence length.

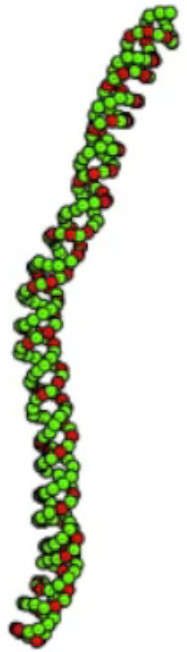
Experimental values
-Cyan and magenta Hagerman
-Purple and yellow Abels et al.



Surprisingly some RNA polymeric chains appear to be curved.



poly $r(AC)_3G_5$ • poly $rC_5(GU)_3$



Modelled vs. Experimental a
In range with Hagerman, higher than Abels et al.

Type	Source	a (Å) RNA	a (Å) DNA
Modelled	Mixed Sequence	818	
Experimental	Hagerman	700-800	
Experimental	Abels et al.	622	
Experimental	Abels et al.	638	
Modelled	Ideal DNA		500
	ζ (scaling factor)	0.75	0.50

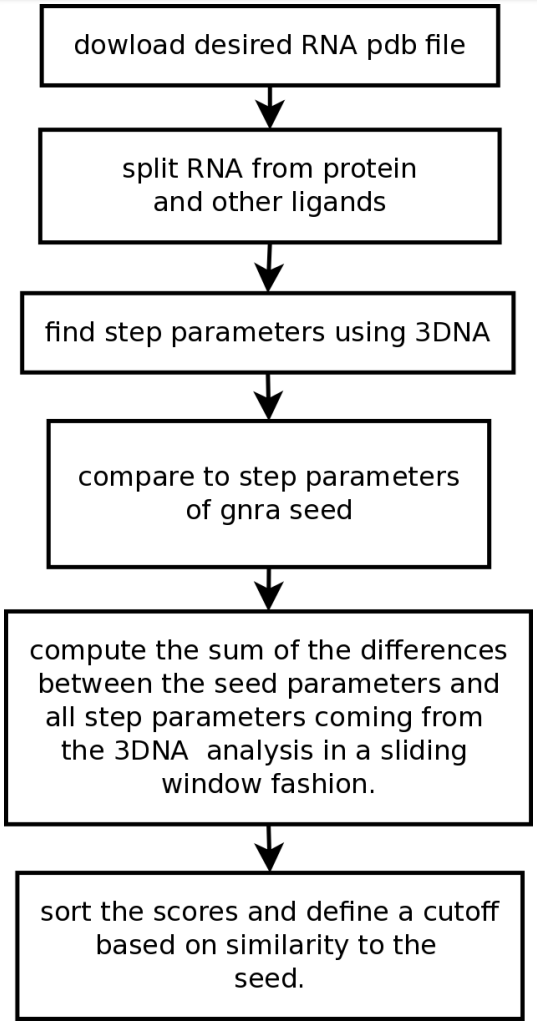
Persistence length derived from X-ray structure analysis suggests than RNA is stiffer than DNA and in the range of Hagerman but slightly higher than Abels and Dekker.

ACT IV

Finding RNA structural motifs using base step parameters.

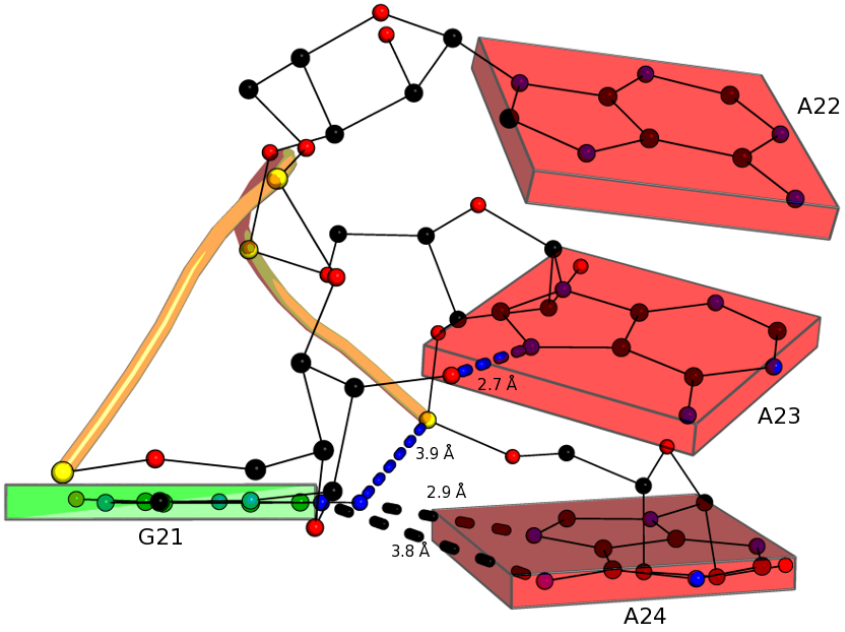
getMotif: A simple program to find RNA motifs based on single stranded base step parameters.

getMotif uses a seed of single-stranded base step parameters to apply a sliding window technique to localize RNA structural motifs in any RNA structure given in pdb format.



Testing the getMotif program using the GNRA Tetraloop Motif

GNRA tetraloop provides good “toy” model to test getMotif



The first step in GNRA is quite far from canonical A-RNA. NR, and RA Steps are closer save for having Greater slide and being over-twisted.

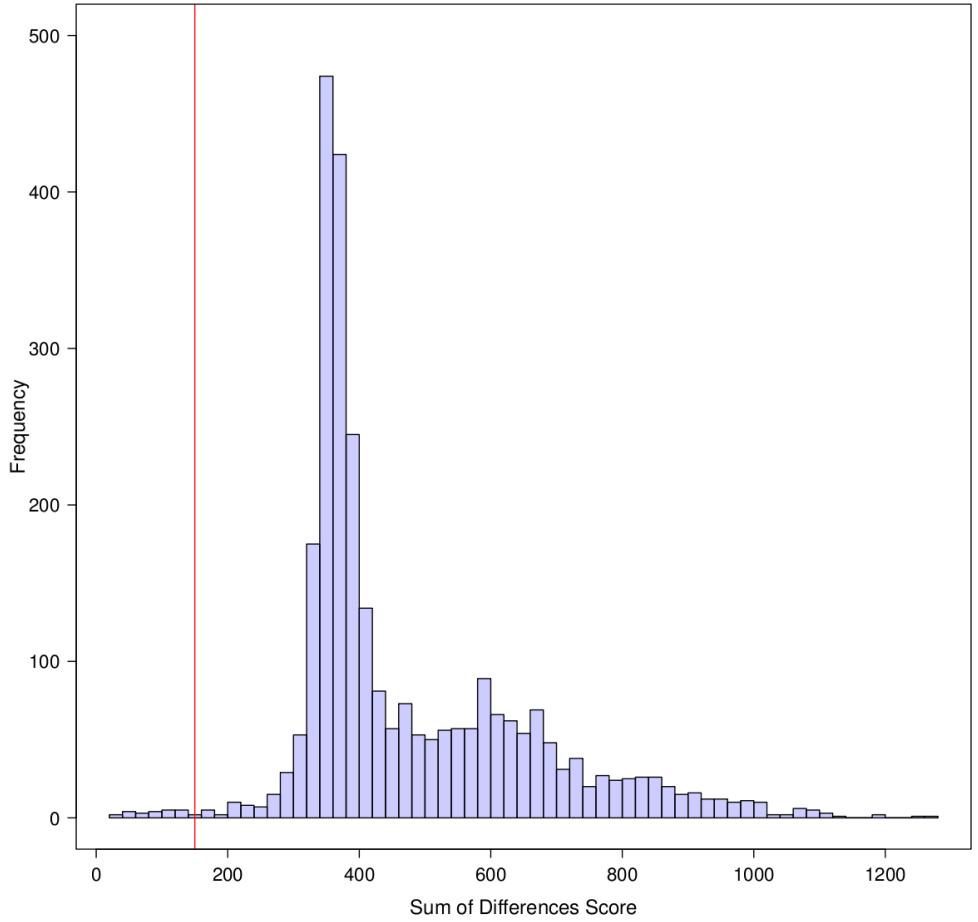
Step	Shift	Slide	Rise	Tilt	Roll	Twist
GN	-9.77	-1.90	-4.93	71.8	124.0	-57.5
NR	2.74	-0.11	3.04	11.5	6.0	50.6
RA	1.11	-0.20	3.01	9.5	6.1	42.4
A-RNA	0.00	-1.48	3.30	0.0	8.6	31.6

$$X_k = |S - W_k|$$

$$\text{score}_k = \sum_{i,j} \{x_{i,j}\}$$

Sum of differences score for a sliding window finds potential GNRA tetraloop motif candidates.

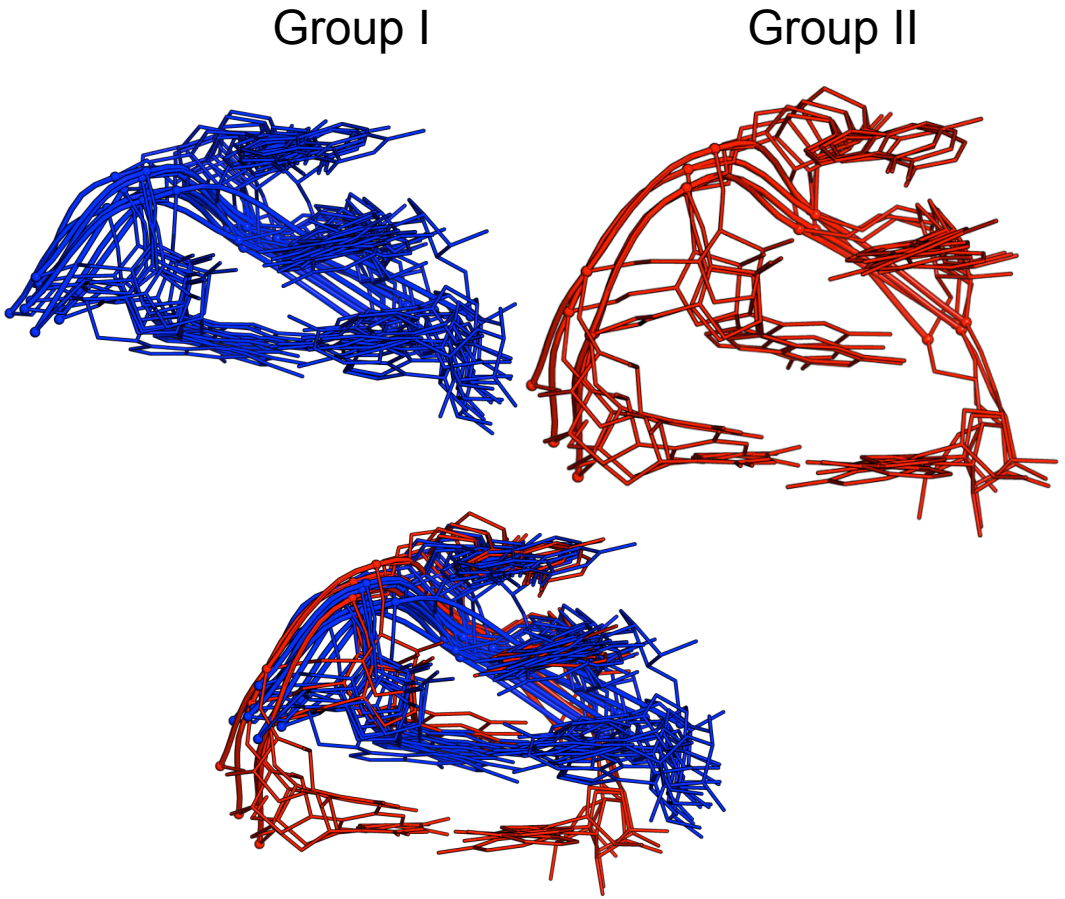
Sum of difference score for the large subunuit of the ribosome, PDB_ID:1FFK, compared to the GNRA tetraloops base steps.



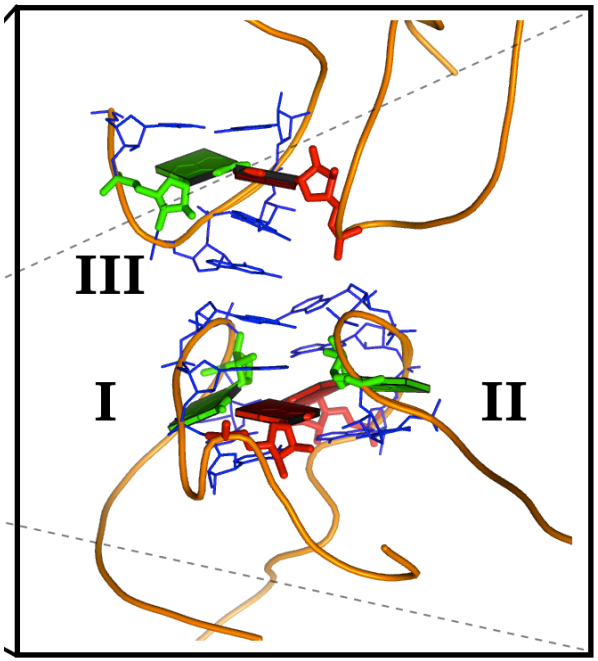
The sum of differences score between single stranded base steps recognizes two GNRA groups.

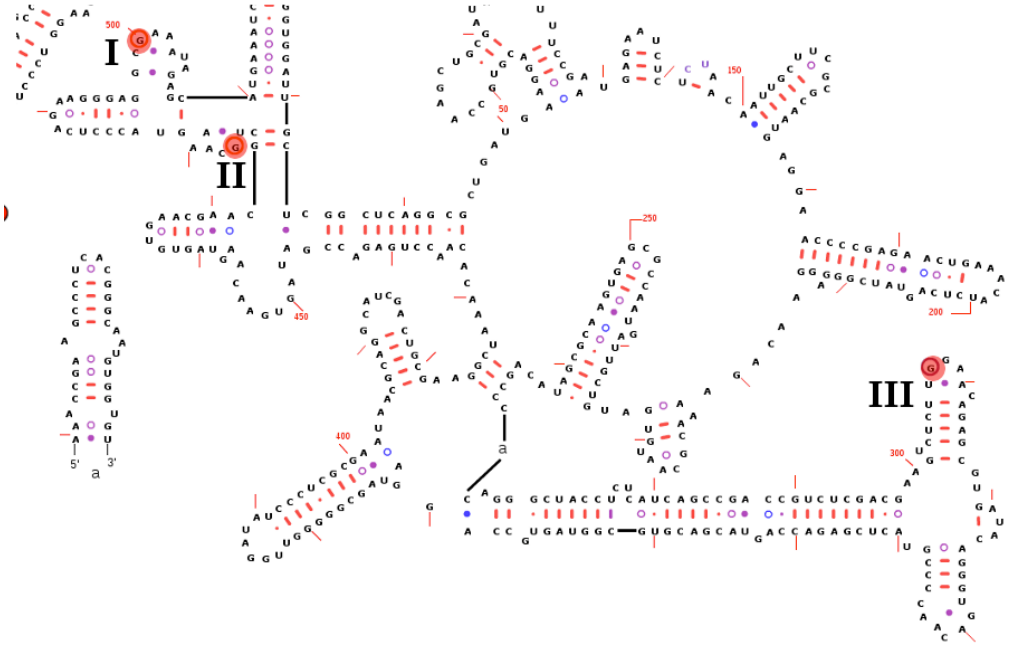
GNRA recognizes two GNRA-like motifs.

In blue the “canonical” GNRA tetraloop motif. In red the lonepair triloop motif.



Group II candidate is the lone-pair triloop motif, or a composite GNRA motif.





The GNRA-like candidates apparently lacking the sheared G·A base-pair, form a lonepair triloop motif, which is just a structural GNRA motif, with a non-sequential closing base-pair.

FINALE

Conclusions

- Rigid-body analysis reveals missing conformations on classification of RNA dinucleotide steps which can be classified using rigid-body parameters.
- Base pairs in RNA helical regions in X-ray structures arrange in seven predominant types.
- Modelling using rigid bodies gives surprising results of RNA curving with specific sequences.
- A new software for finding RNA motifs based on rigid-body parameters successfully finds the GNRA motifs and the closely related lone-pair triloop motif.

FINALE

Future and developing work.

- Expand **RNAsteps.rutgers.edu** to periodically synchronize its data with the PDB in an automated fashion.
- Construction of a dataset of base-step parameters for known RNA motifs to be used in the getMotif software.
- Migrating the getMotif fully to python.
- Develop a web-based front end to getMotif.
- Develop a code for computing rigid-body parameters thinking of helical regions as rigid-bodies.

Thanks to my defense committee members for making it easy to schedule this defense and your valuable time.

Special thanks to all members of the Olson Lab, past and present.

Thanks to all for your attention,

Visit us at: http://dnaserver.rutgers.edu

Global properties of polymer chains. Persistence length.

Persistence length is "the average sum of the projections of all bonds $j \geq i$ on bond i in an indefinetely long chain. The bond i is taken to be remote from either end of the chain, i.e., $1 \ll i \ll n$". Paul J. Flory, Statistical Mechanics of Chain Molecules. 1969

$$a = lim_{n->\infty} \sum_{i=1}^{n} \langle \mathbf{l}_i \cdot (\mathbf{l}_1 / l_1) \rangle$$

$$\mathbf{r} = \mathbf{l}_1 + \mathbf{T}_{12}\mathbf{l}_2 + \mathbf{T}_{12}\mathbf{T}_{23}\mathbf{l}_3 + \ldots + \mathbf{T}_{12}\mathbf{T}_{23}\ldots\mathbf{T}_{N-1,N}\mathbf{l}_N.$$

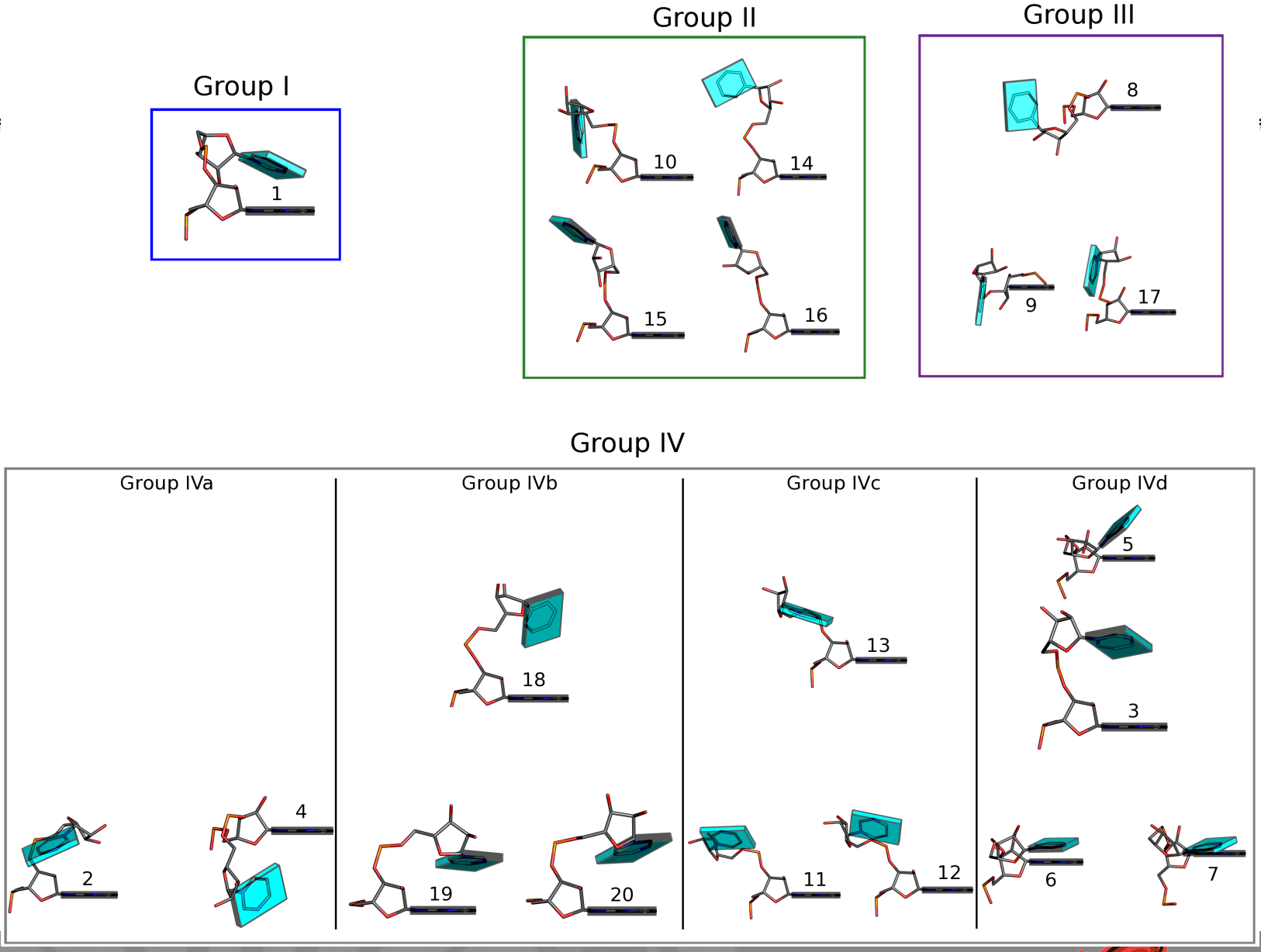
$$\mathbf{r} = \begin{bmatrix} \mathbf{E}_3 & \mathbf{0} \end{bmatrix} \mathbf{A}_{1:N} \begin{bmatrix} \mathbf{0} \\ 1 \end{bmatrix}$$

$$\mathbf{A}_n = \begin{bmatrix} \mathbf{T}_n & \mathbf{l}_n, \\ 0 & 1 \end{bmatrix}$$

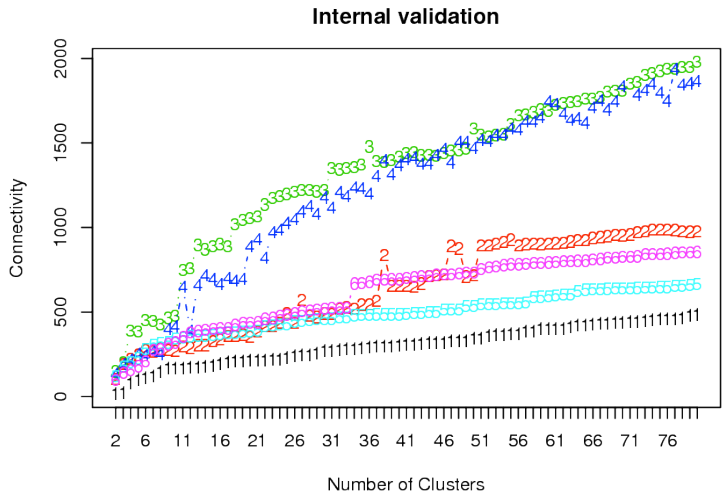
$$\mathbf{A}_{1:N} = \mathbf{A}_1\mathbf{A}_2\ldots\mathbf{A}_N.$$

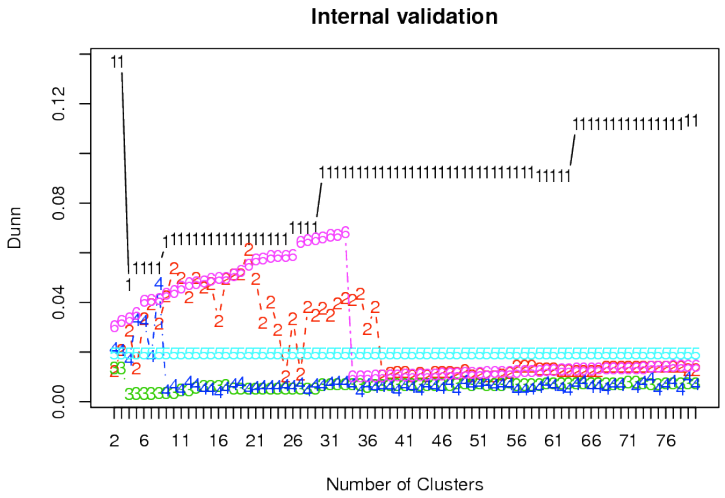
$$\mathbf{P}_N = \langle \mathbf{A}_1 \rangle \langle \mathbf{A}_2 \rangle \ldots \langle \mathbf{A}_{N-1} \rangle \langle \mathbf{A}_N \rangle.$$

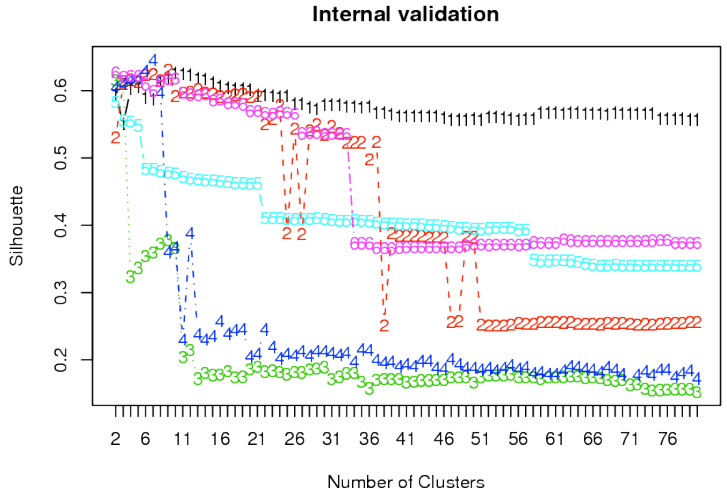
$$a = \lim_{N\to\infty} \begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix} \mathbf{P}_N \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$


Group I
1
Group II
10
14
15
16
Group III
8
9
17
Group IV
Group IVa
2
4
Group IVb
18
19
20
Group IVc
13
11
12
Group IVd
5
3
6
7

Instead of picking any clustering methodology we use the clValid R package for cluster validation.







Clustering methods

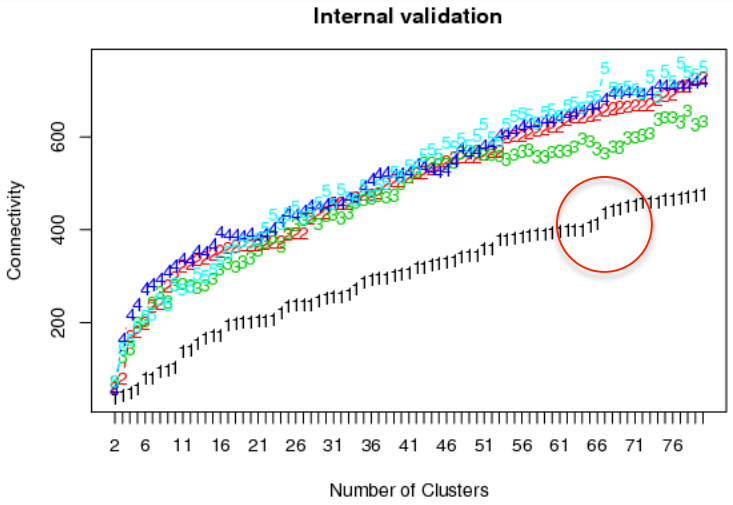
1 hierarchical
2 kmeans
3 pam
4 som
5 sota
6 diana

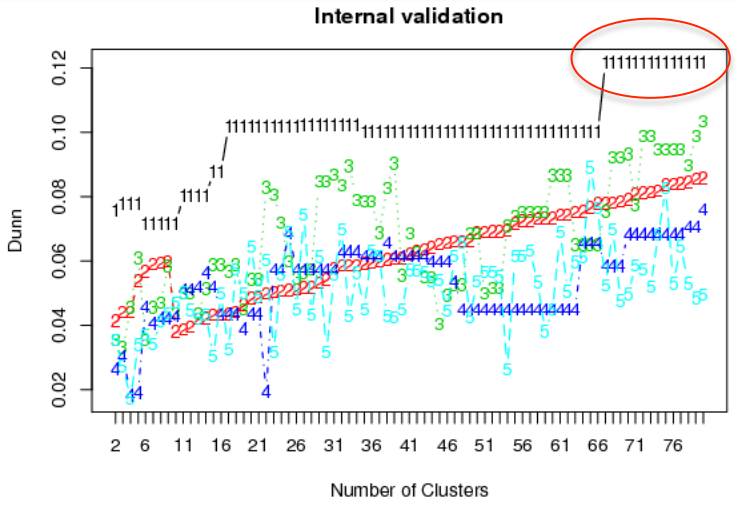
Clustering scores

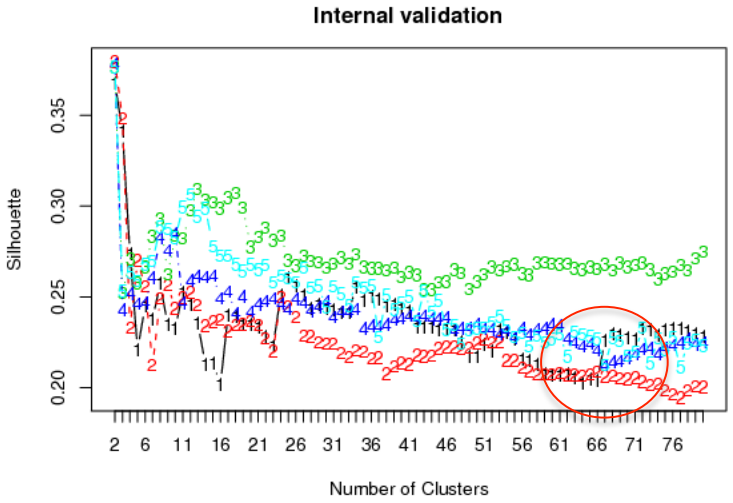
Connectivity small
Silhouette large
Dunn Index large

SAVE IN THE BACK FULL PLOTS

For the non-A-RNA-like dinucleotide steps the best clustering method is the hierarchical one. k =67.



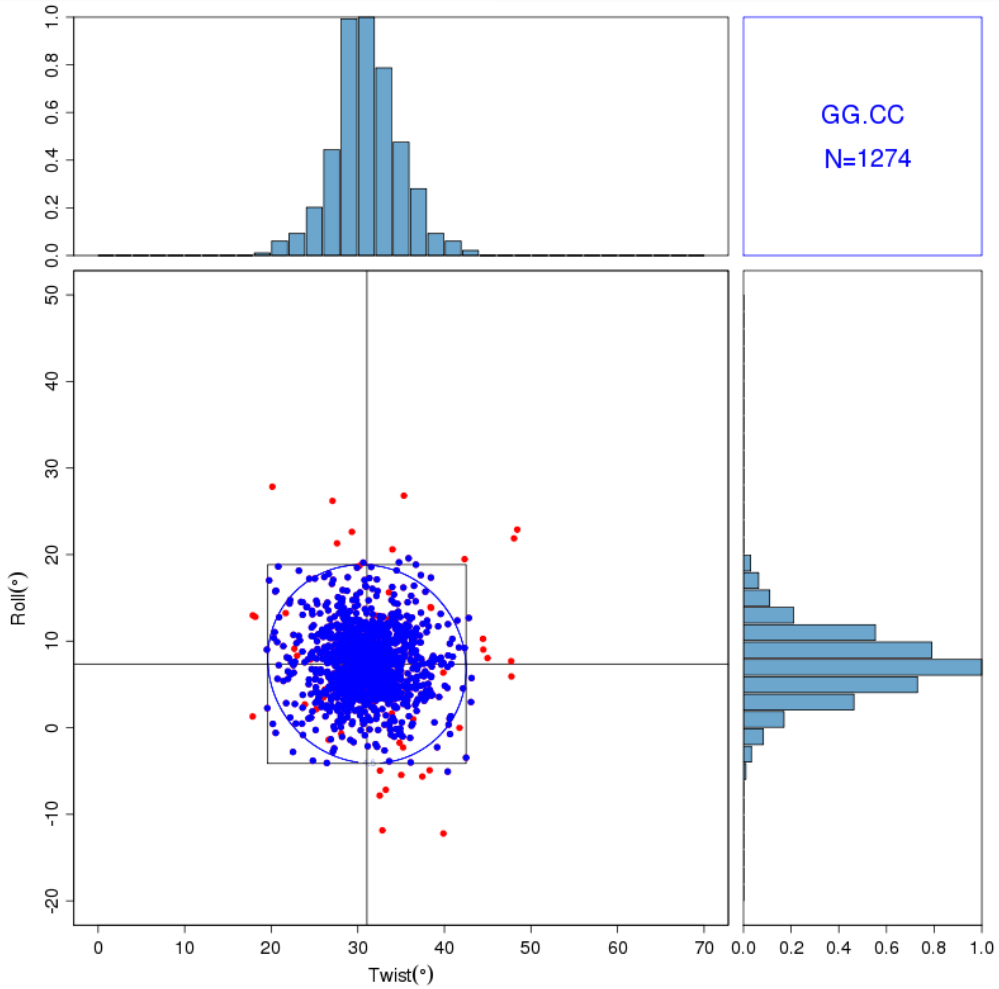




1 hierarchical
2 kmeans
3 pam
4 som
5 sota
6 diana

Connectivity small
Silhouette small
Dunn Index large

Energy contours on data culled by three standard deviations from the mean.



Sum of differences score for a sliding window finds potential GNRA tetraloop motif candidates.

This is for 1FFK

Sliding windows image

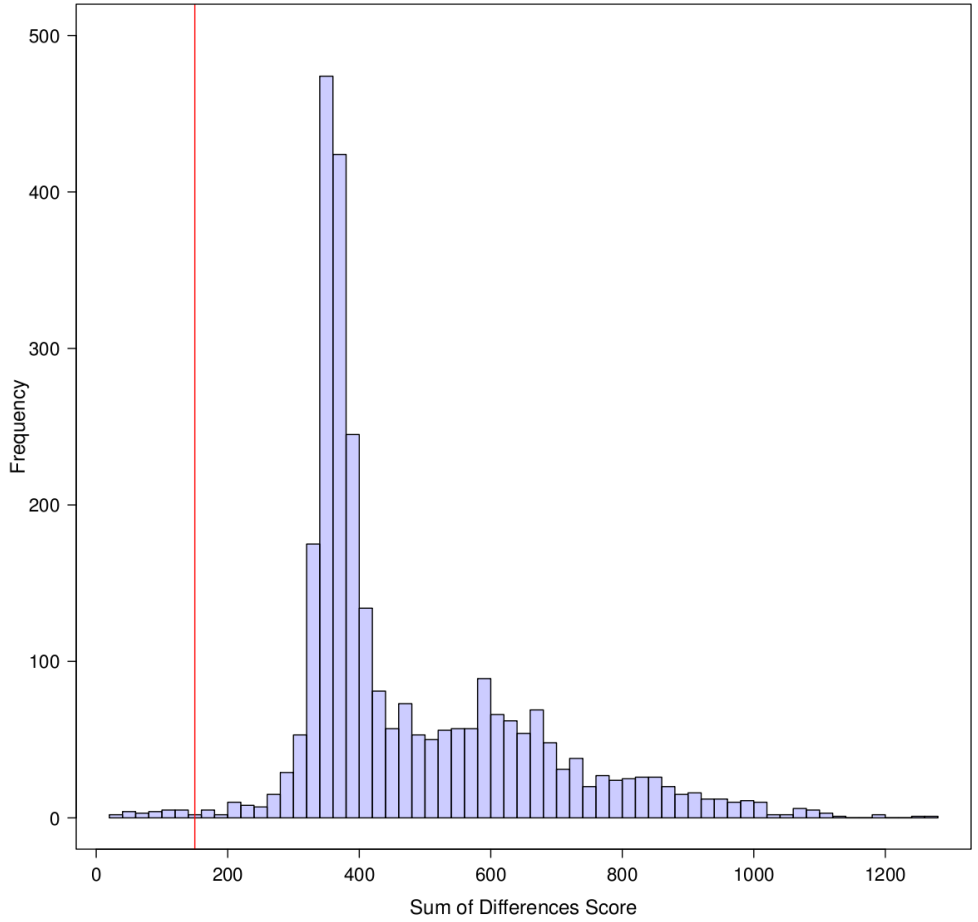
Image with matrices in color Boxes

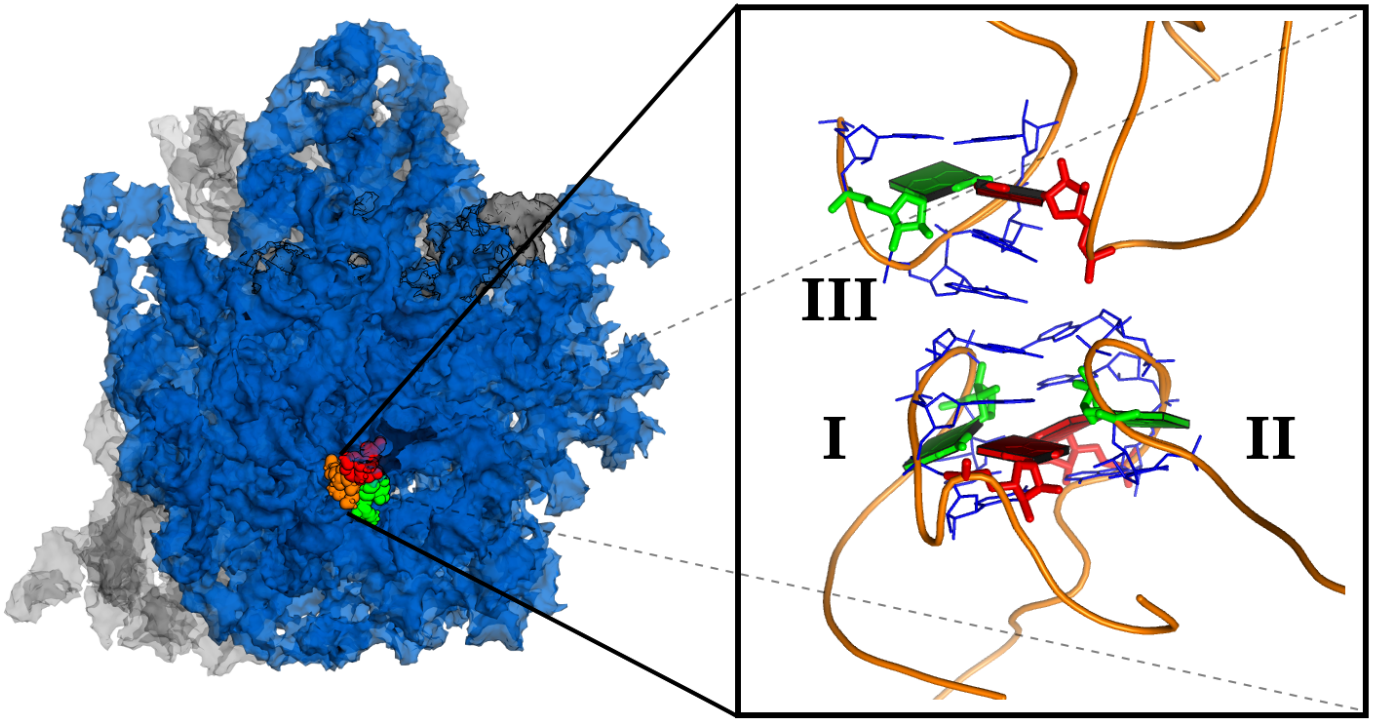
e.g.

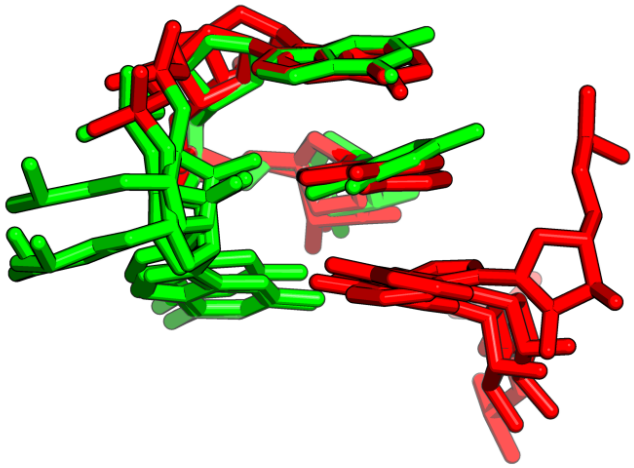
--- ---- ----

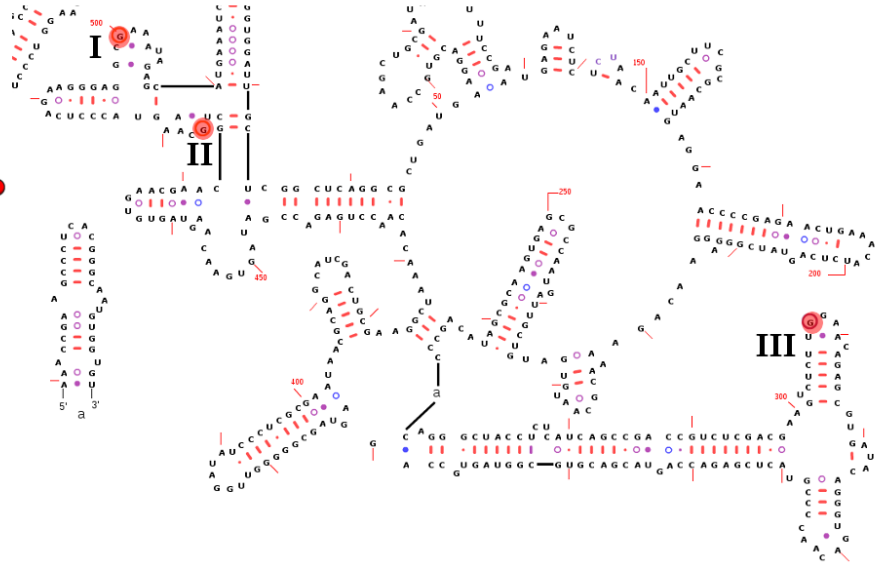
---- ---- -----

--- ---- ----




III
I
II




I
II
III