COMP598: Advanced Computational Biology Methods and Research

Modeling RNA 3D structure

Jérôme Waldispühl

School of Computer Science, McGill

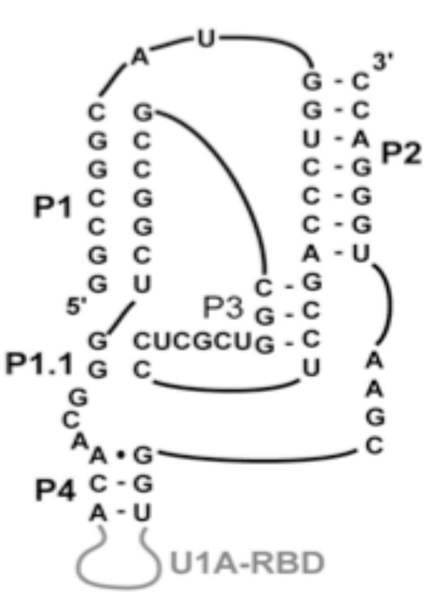
Slides from Neocles Leontis & Jes Frellsen

Motivations and challenges

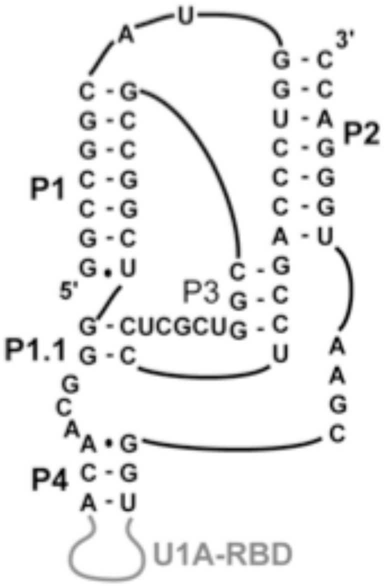
- Secondary structure is a simplification of the three-dimensional structure.
- Function is achieved through the 3D structure.
- Experimental determination the RNA 3D structure is hard.
- Modeling the 3D structure is also hard!
- Before the prediction, a work has to be done on modeling and alignment of 3D structure.

Beyond the secondary structure

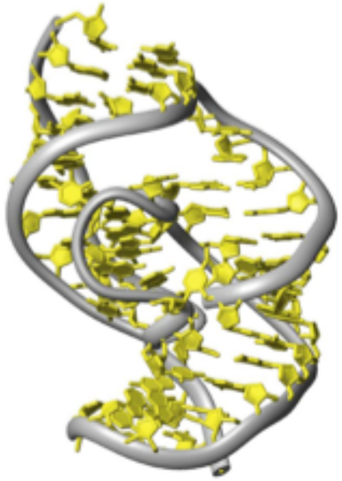
The hierarchy of the model is not as obvious as expected:



Secondary structure without pseudo-knot



Secondary structure with pseudo-knot

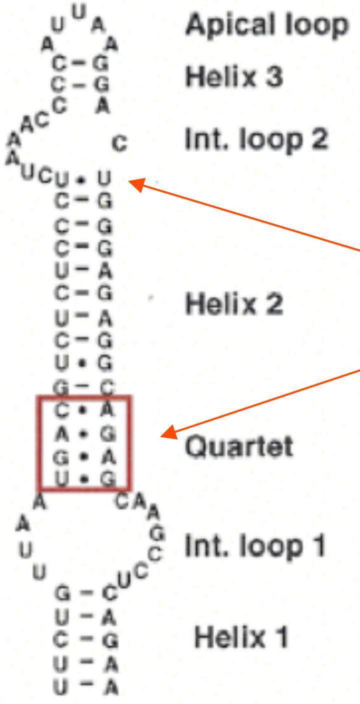


Tertiary structure

- Is the secondary structure with/without pseudo-knot unique?
- Is there other type of interacting motifs? (for instance base triple)

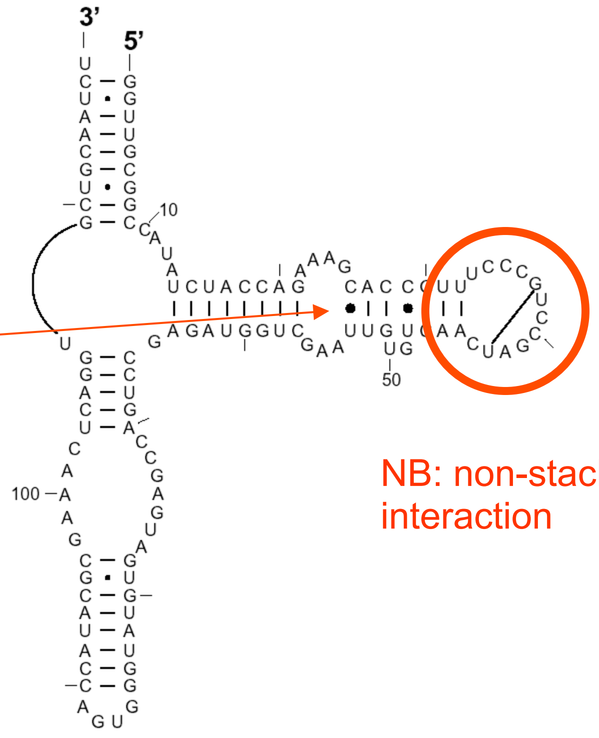
Beyond the secondary structure

The type of interactions is not restricted to Watson-Crick base pairs:



SECIS element

Non-canonical interactions



NB: non-stacking interaction

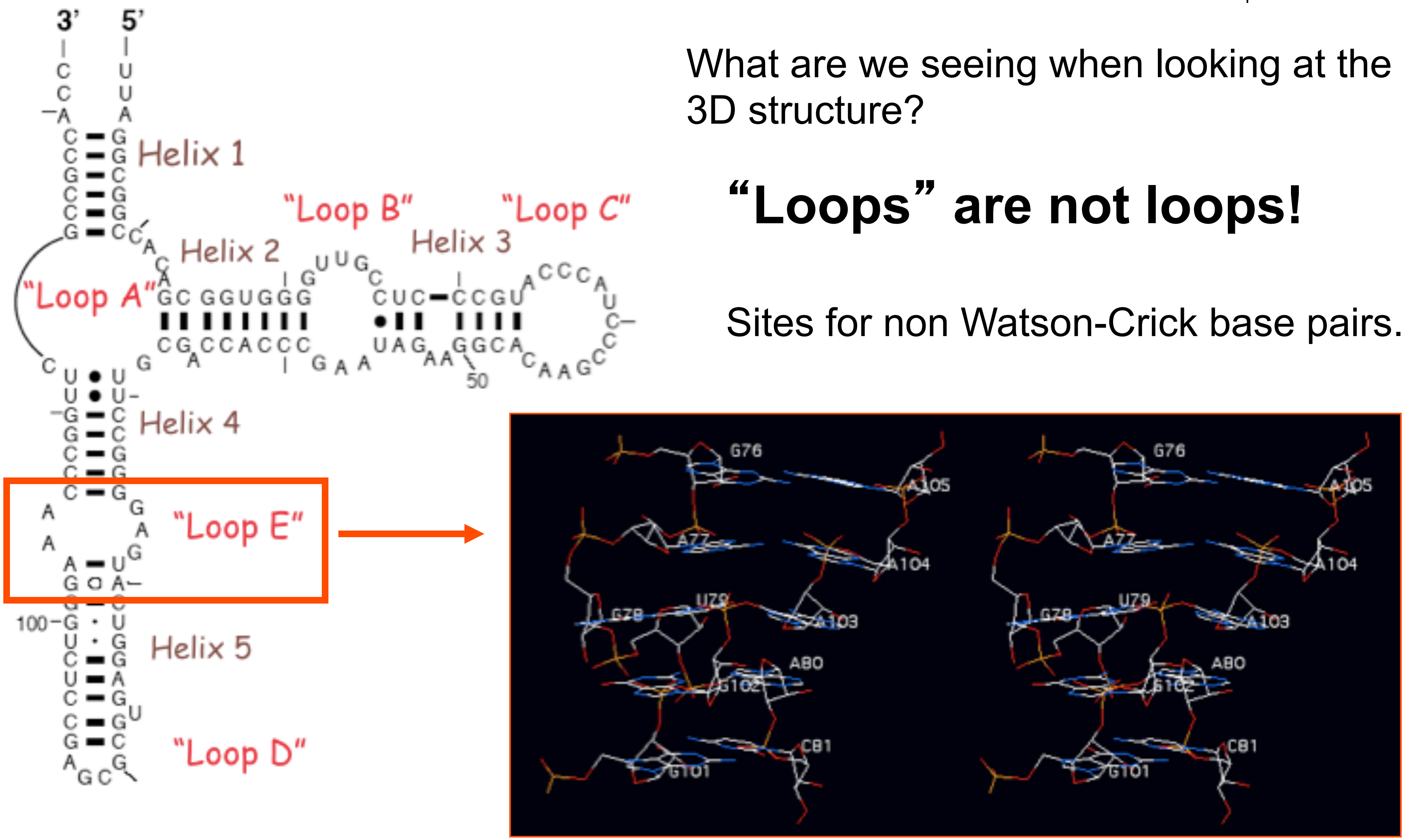
5S ribosomal RNA

Classification of non Watson-Crick base pair interactions

What are we seeing when looking at the 3D structure?

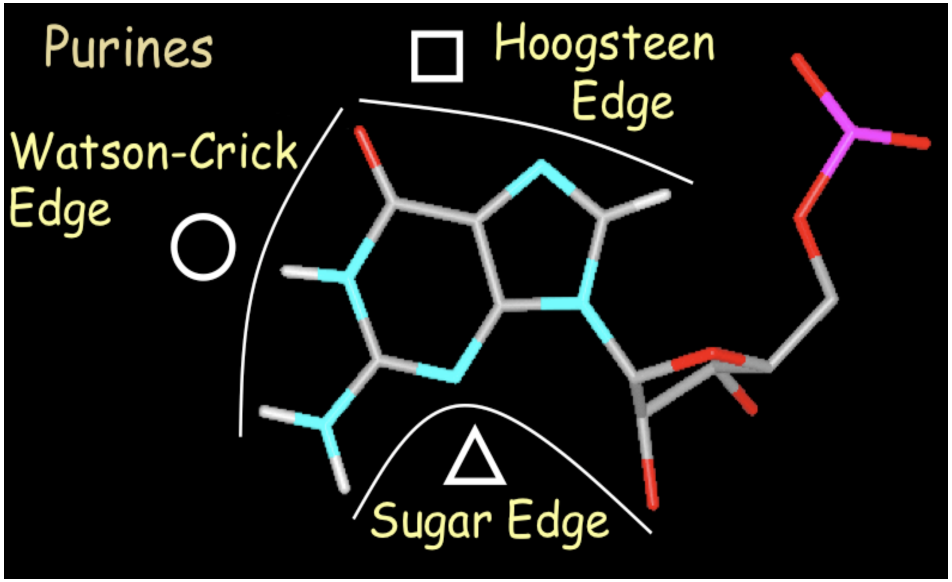
“Loops” are not loops!

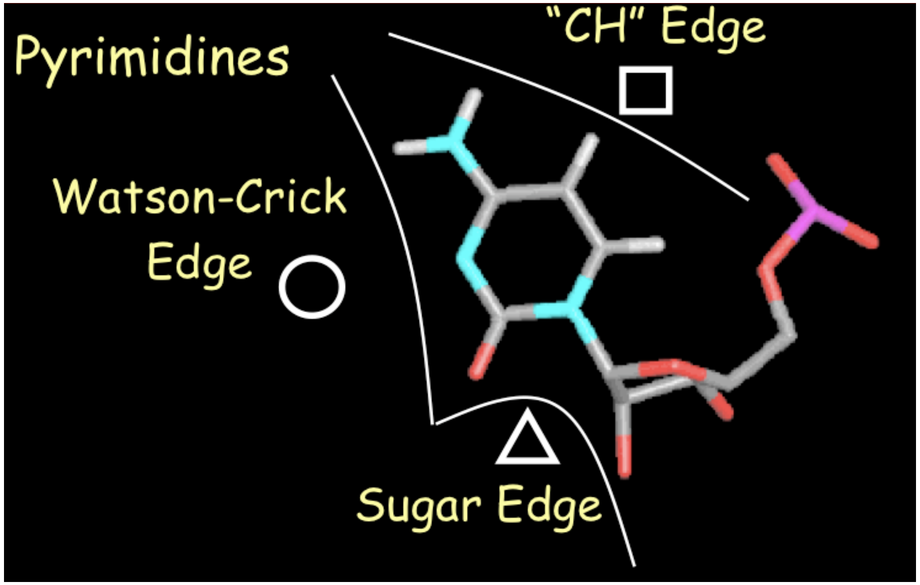
Sites for non Watson-Crick base pairs.



Classification of non Watson-Crick base pair interactions

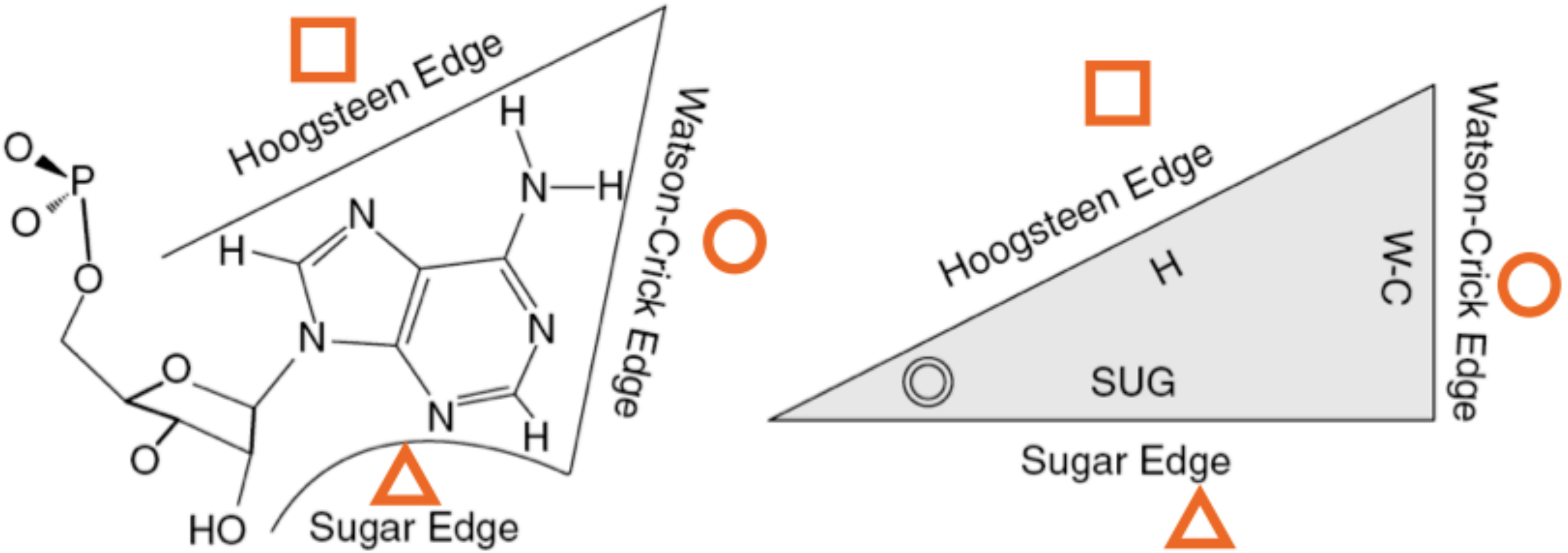
Modeling the nucleotide side-chain with interacting edges





Classification of non Watson-Crick base pair interactions

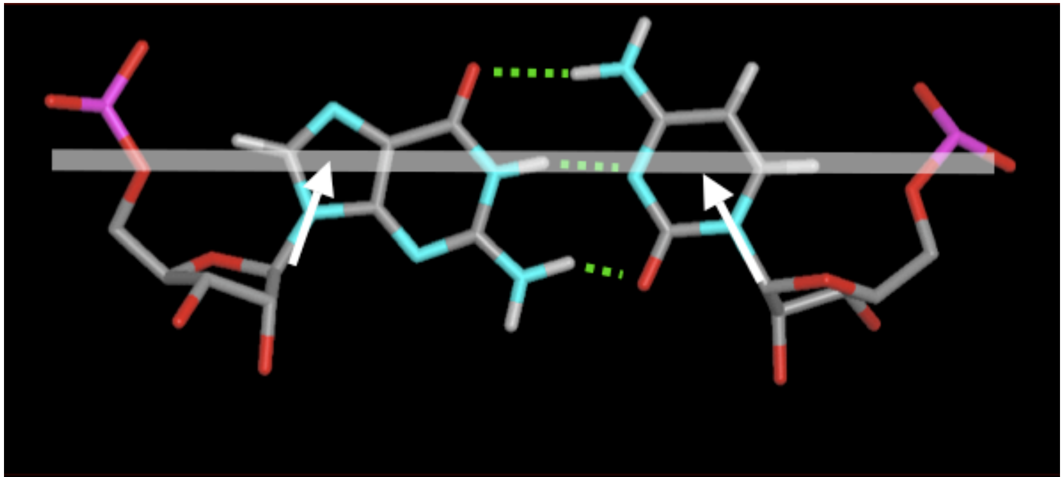
Consequence: 3 edges available for base-pairing.



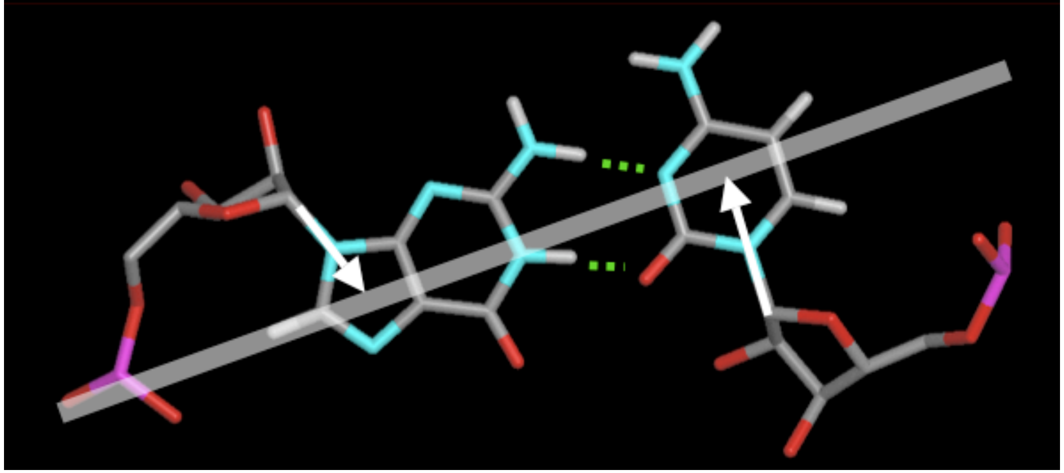
Classification of non Watson-Crick base pair interactions

Orientation of edge interaction is also important: The glycosidic bond orientation.

Cys (default):

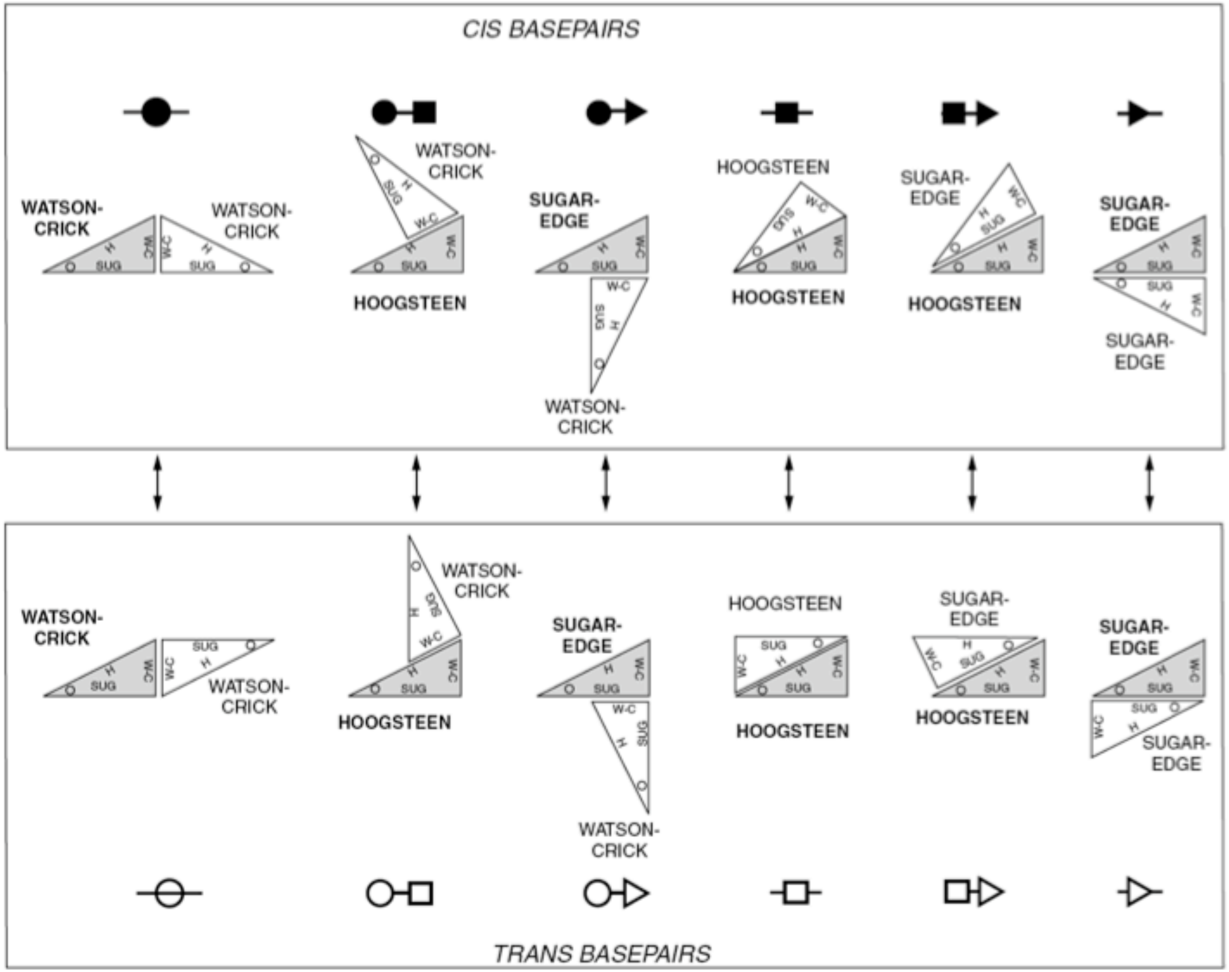


Trans:



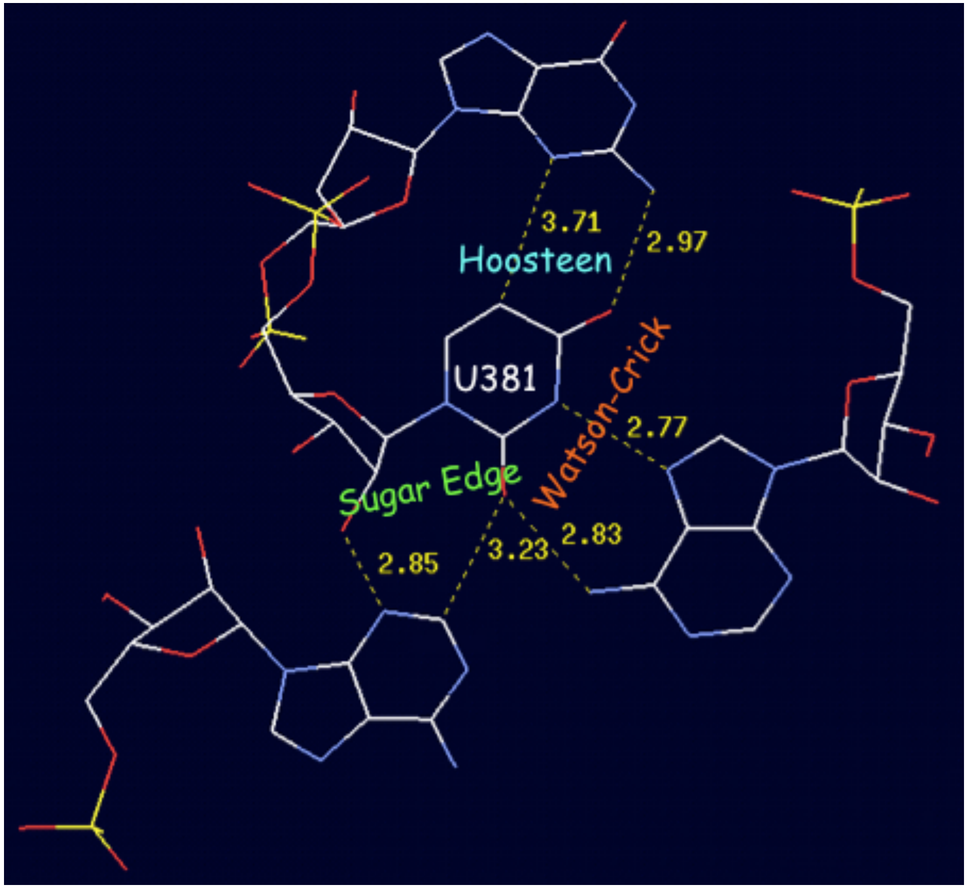
Classification of non Watson-Crick base pair interactions

12 edge-to-edge interacting motifs



Classification of interactions

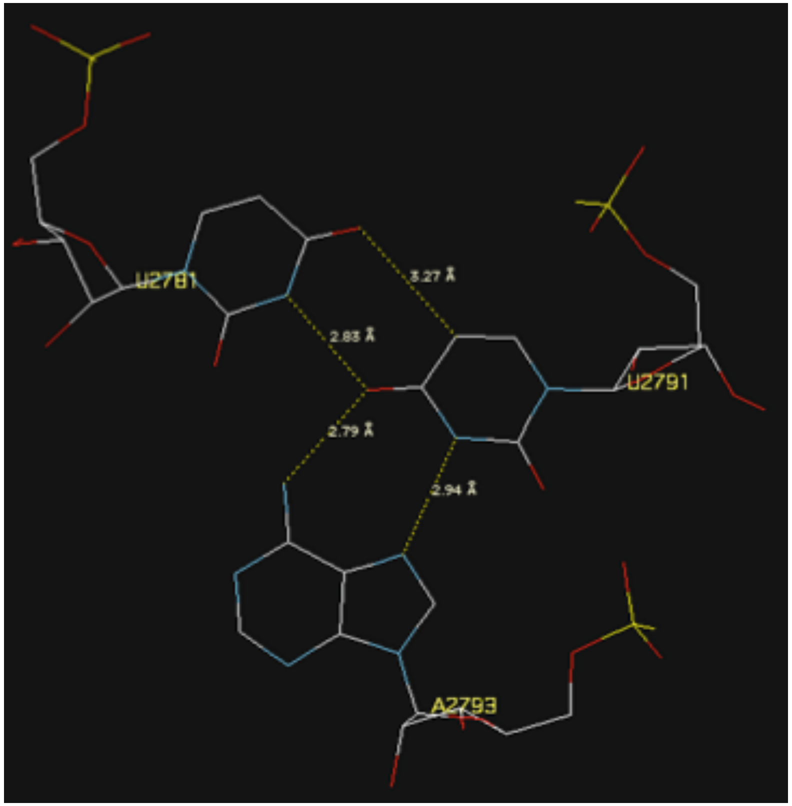
But the puzzle is still far to be completed! ☹



Base interacting with all 3 edges

Classification of interactions

The interacting motif is extended to model base triple.



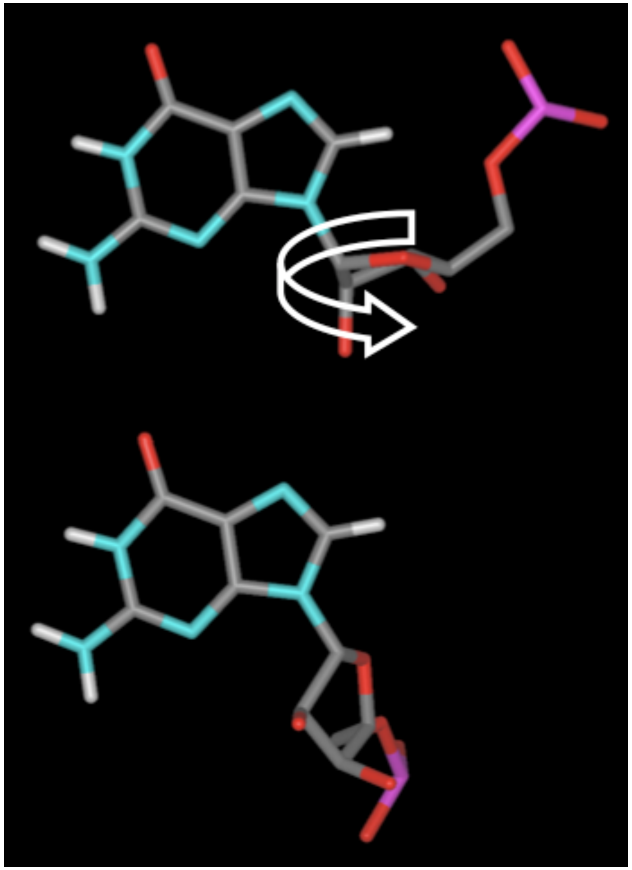
U ○–□ U ●–■ A

More Features…

Base-Sugar conformation.

Anti (default):

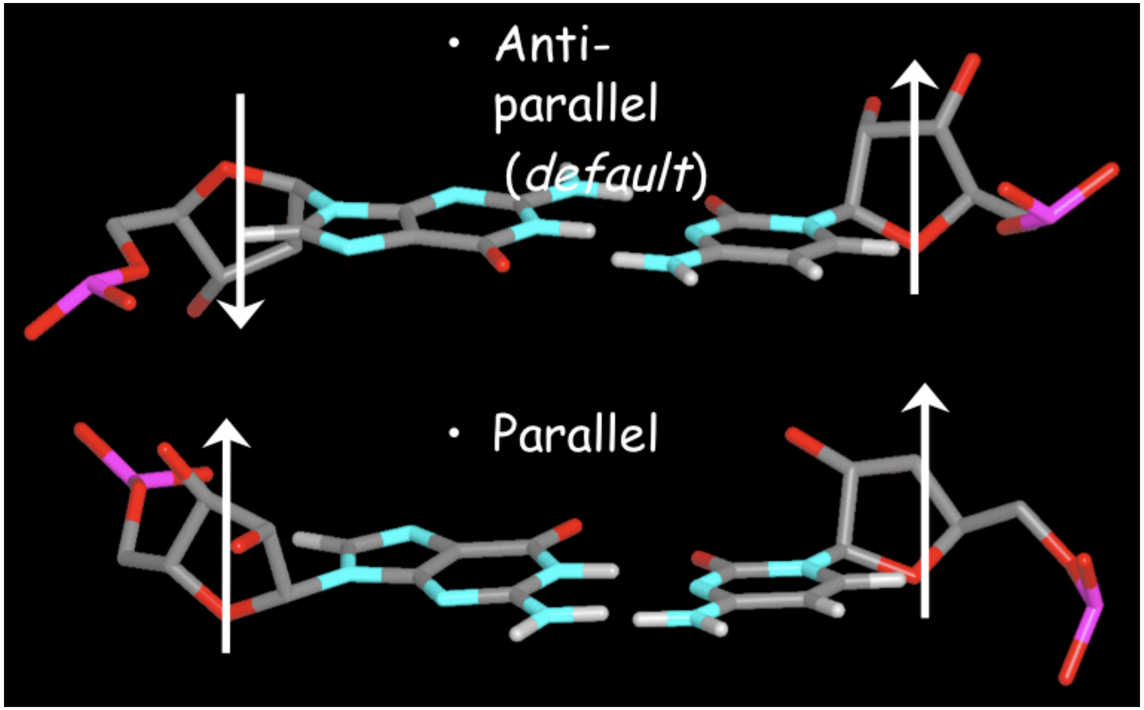
Syn (Purines only):

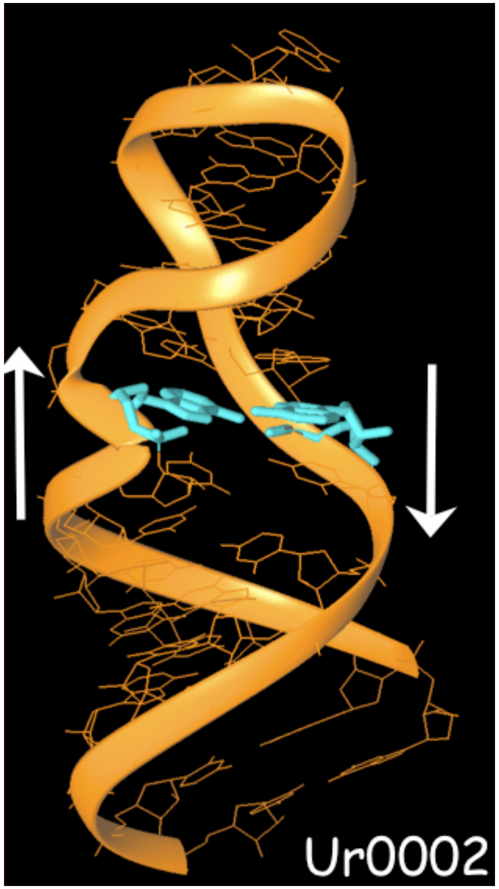


More features…

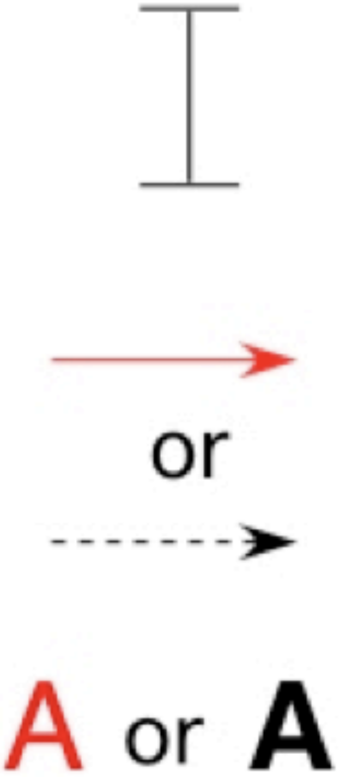
Locally parallel strands:

Local strand orientation:





New symbols

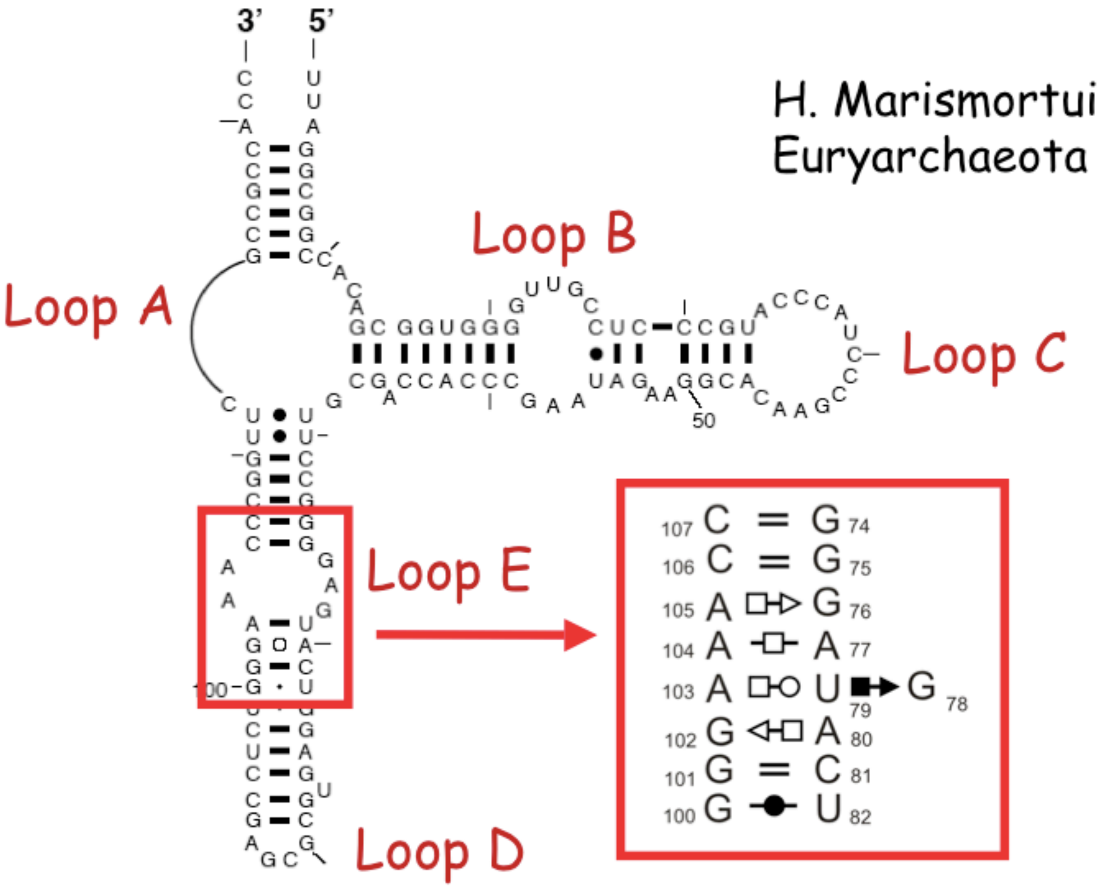


Indicates Base Stacking

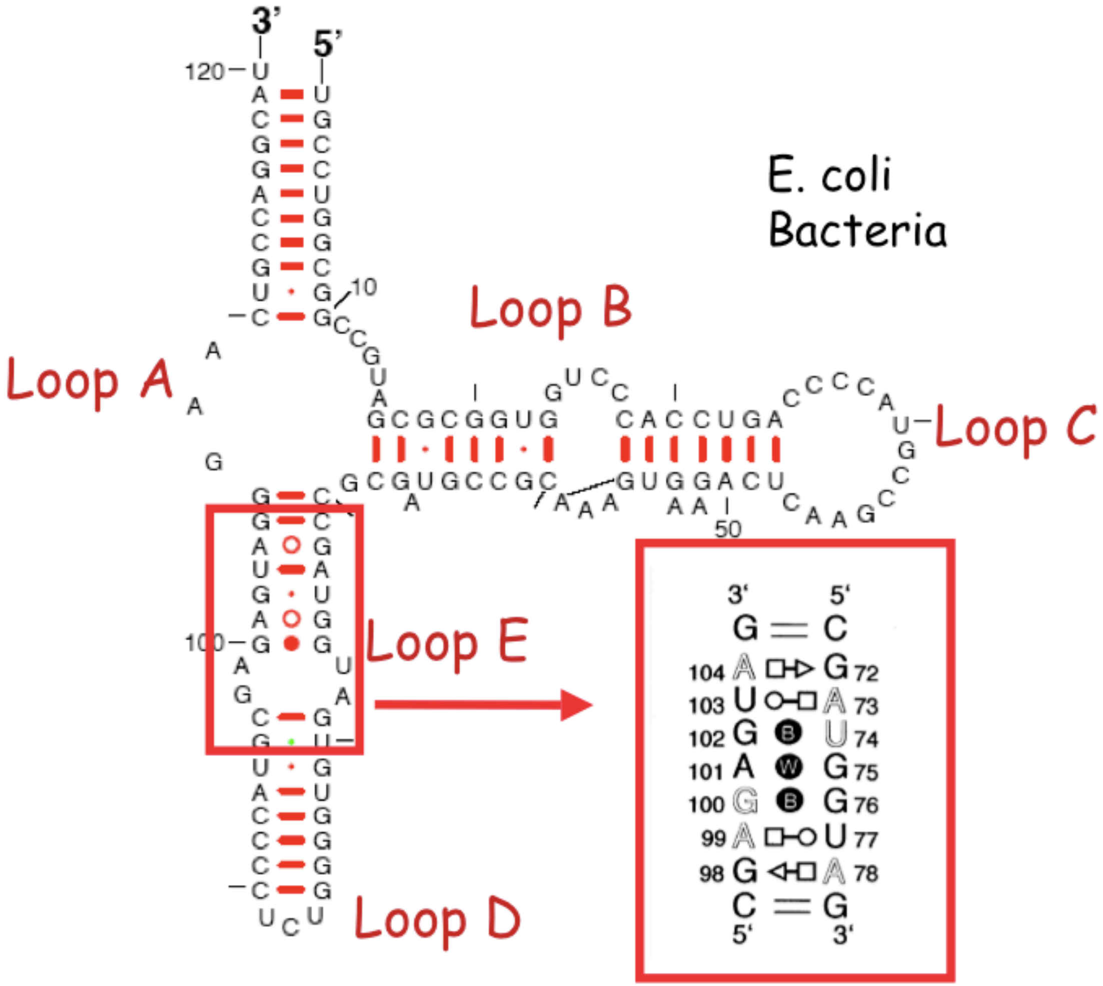
Indicates Change in Strand Orientation

Indicates syn conformation for base

Example: 5S motif



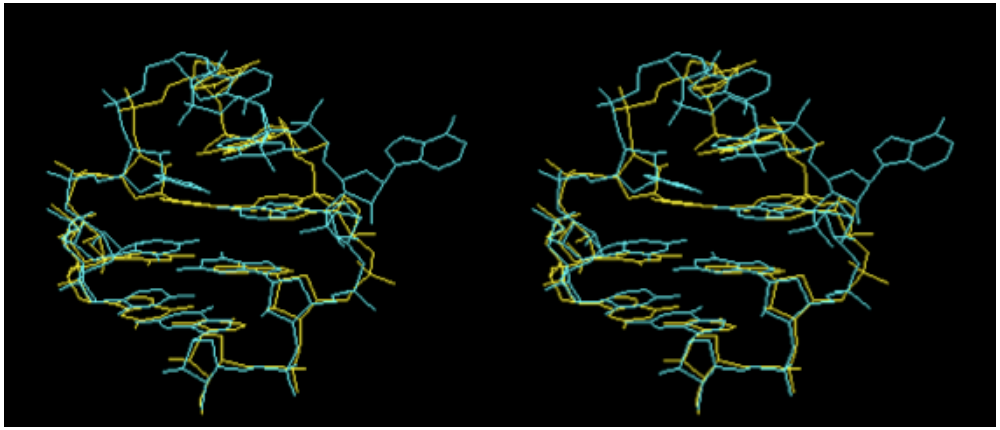
Example: 5S motif

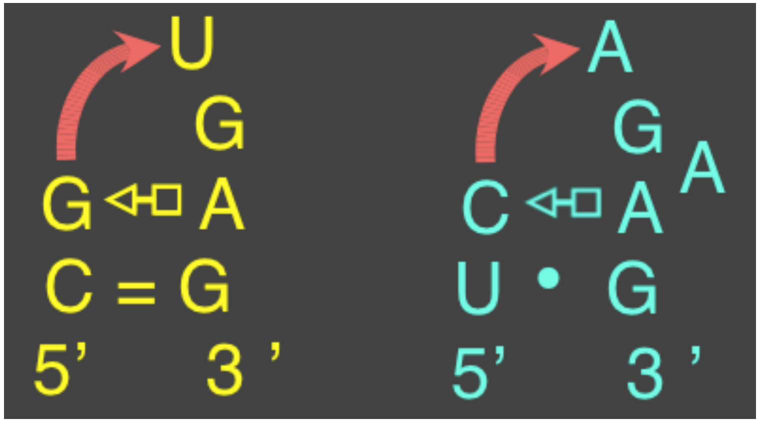


E. coli
Bacteria

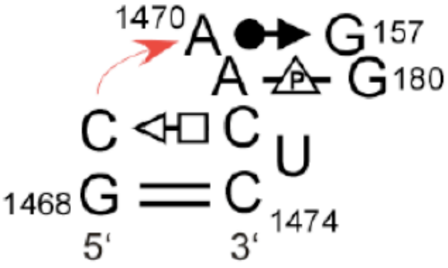
More features (2)…

Superposition of tetra and penta GNRA loops:

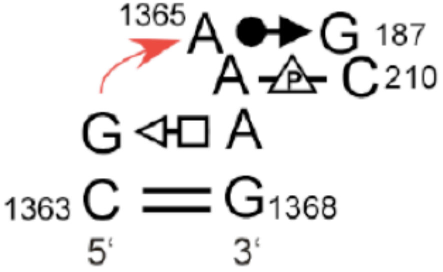




Interaction of GNRA loops are also conserved:



23S H. marismortui



23S T. thermophilus

Finding RNA motifs in 3D structures

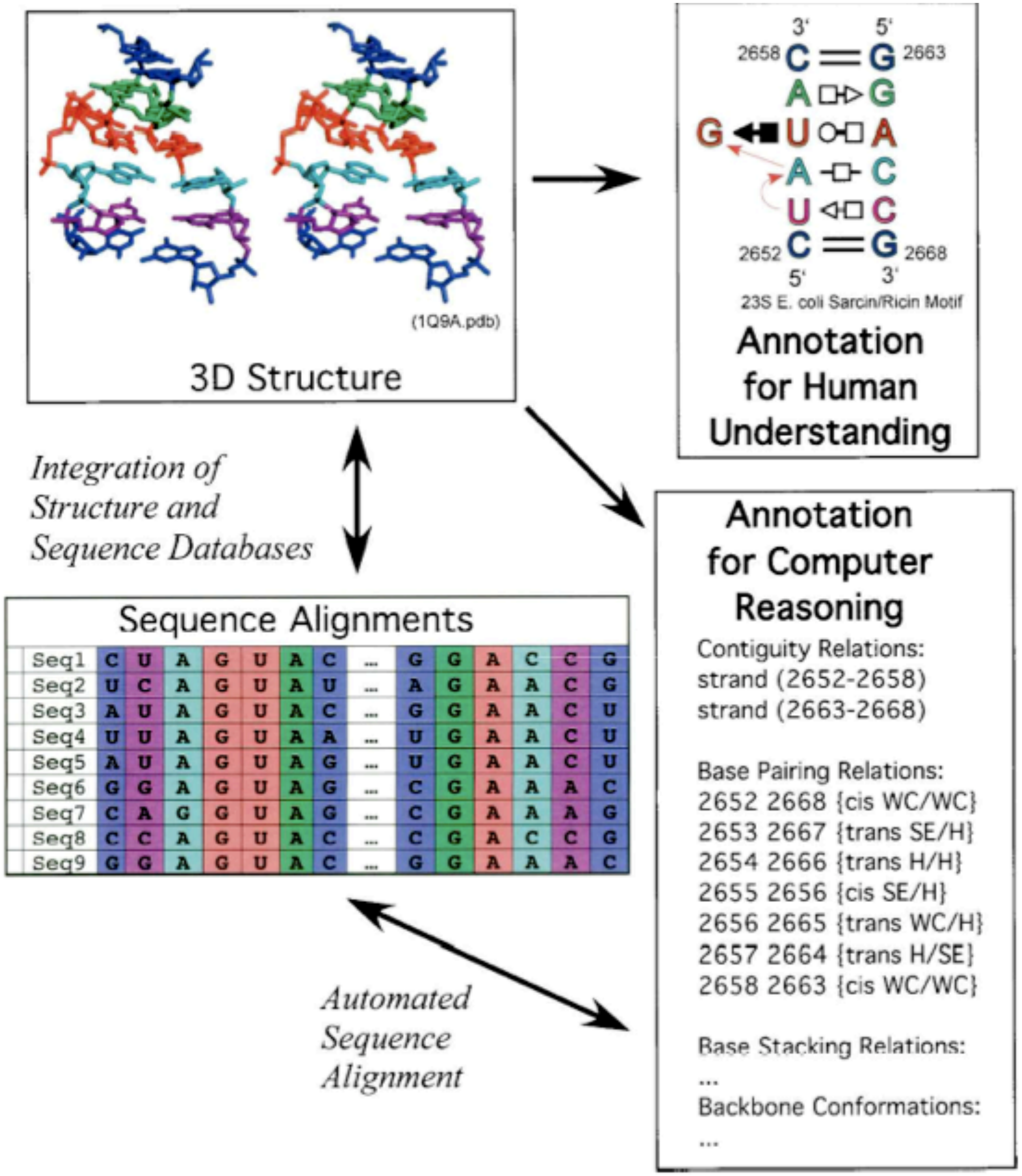
Q: Given a description of a “known” motif, how to identify this motif in target structures?

Use graph theory, the problem of identifying a known pattern in a target graph reduces to the following:

1. Searching for isomorphic occurrences of the pattern (subgraph isomorphism).

2. Finding similar occurrences of the pattern (identifying a maximum common subgraph).

But it’s NP-complete…


(1Q9A.pdb)
3D Structure
3'
5'
2658
C
G
2663
A
G
G
U
A
A
C
U
C
2652
C
G
2668
5'
3'
23S E. coli Sarcin/Ricin Motif
Annotation for Human Understanding
Integration of Structure and Sequence Databases
Automated Sequence Alignment

Sequence Alignments

Seq1	C	U	A	G	U	A	C	...	G	G	A	C	C	G
Seq2	U	C	A	G	U	A	U	...	A	G	A	A	C	G
Seq3	A	U	A	G	U	A	C	...	G	G	A	A	C	U
Seq4	U	U	A	G	U	A	A	...	U	G	A	A	C	U
Seq5	A	U	A	G	U	A	G	...	U	G	A	A	C	U
Seq6	G	G	A	G	U	A	G	...	C	G	A	A	A	C
Seq7	C	A	G	G	U	A	G	...	C	G	A	A	A	G
Seq8	C	C	A	G	U	A	C	...	C	G	A	C	C	G
Seq9	G	G	A	G	U	A	C	...	G	G	A	A	A	C

Annotation for Computer Reasoning

Contiguity Relations:
strand (2652-2658)
strand (2663-2668)

Base Pairing Relations:
2652 2668 {cis WC/WC}
2653 2667 {trans SE/H}
2654 2666 {trans H/H}
2655 2656 {cis SE/H}
2656 2665 {trans WC/H}
2657 2664 {trans H/SE}
2658 2663 {cis WC/WC}

Base Stacking Relations:
...
Backbone Conformations:
...

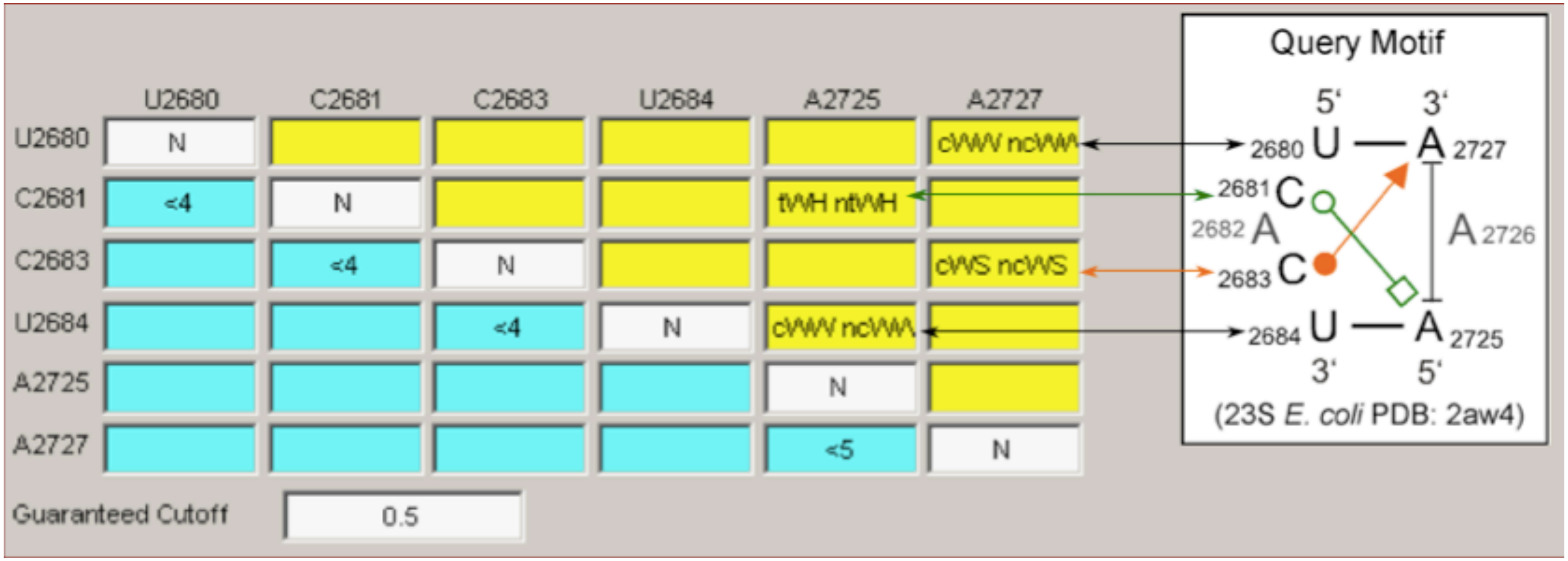
FR3D: Find RNA 3D

(Sarver et al., 2008)

Leontis + Zirbel groups

Find small RNA motifs (two to 20 nucleotides) in PDB files.

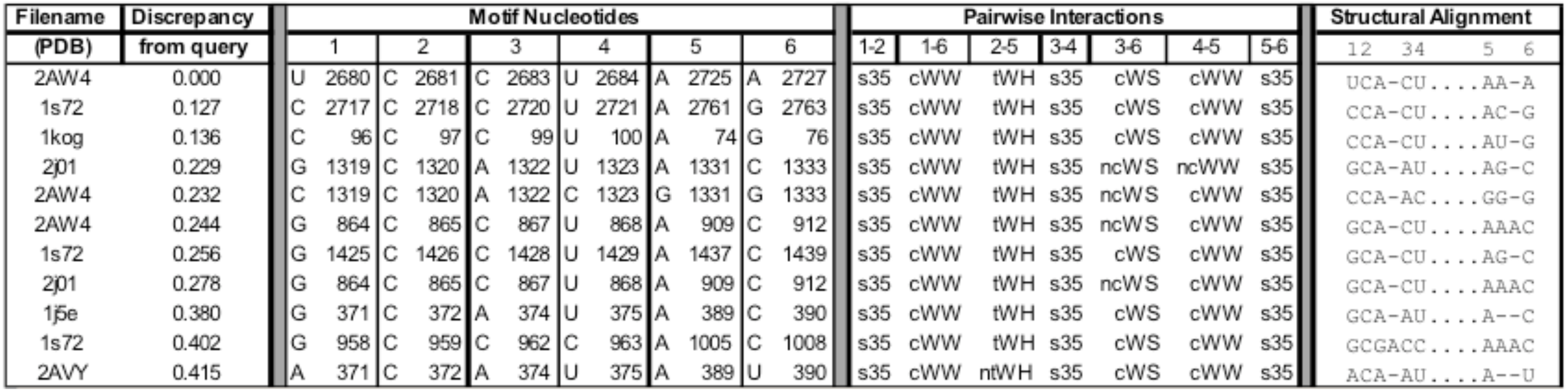
FR3D example: C-loop search



	U2680	C2681	C2683	U2684	A2725	A2727
U2680	N					cWW ncWW
C2681	<4	N			tWH ntWH	
C2683		<4	N			cWS ncWS
U2684			<4	N	cWW ncWW	
A2725					N	
A2727					<5	N

Guaranteed Cutoff 0.5

Output:



Filename (PDB)	Discrepancy from query	Motif Nucleotides 1	2	3	4	5	6	Pairwise Interactions 1-2	1-6	2-5	3-4	3-6	4-5	5-6	Structural Alignment 12 34 5 6
2AW4	0.000	U 2680	C 2681	C 2683	U 2684	A 2725	A 2727	s35	cWW	tWH	s35	cWS	cWW	s35	UCA-CU.....AA-A
1s72	0.127	C 2717	C 2718	C 2720	U 2721	A 2761	G 2763	s35	cWW	tWH	s35	cWS	cWW	s35	CCA-CU.....AC-G
1kog	0.136	C 96	C 97	C 99	U 100	A 74	G 76	s35	cWW	tWH	s35	cWS	cWW	s35	CCA-CU.....AU-G
2j01	0.229	G 1319	C 1320	A 1322	U 1323	A 1331	C 1333	s35	cWW	tWH	s35	ncWS	ncWW	s35	GCA-AU.....AG-C
2AW4	0.232	C 1319	C 1320	A 1322	C 1323	G 1331	G 1333	s35	cWW	tWH	s35	ncWS	cWW	s35	CCA-AC.....GG-G
2AW4	0.244	G 864	C 865	C 867	U 868	A 909	C 912	s35	cWW	tWH	s35	ncWS	cWW	s35	GCA-CU.....AAAC
1s72	0.256	G 1425	C 1426	C 1428	U 1429	A 1437	C 1439	s35	cWW	tWH	s35	cWS	cWW	s35	GCA-CU.....AG-C
2j01	0.278	G 864	C 865	C 867	U 868	A 909	C 912	s35	cWW	tWH	s35	ncWS	cWW	s35	GCA-CU.....AAAC
1j5e	0.380	G 371	C 372	A 374	U 375	A 389	C 390	s35	cWW	tWH	s35	cWS	cWW	s35	GCA-AU.....A--C
1s72	0.402	G 958	C 959	C 962	C 963	A 1005	C 1008	s35	cWW	tWH	s35	cWS	cWW	s35	GCGACC.....AAAC
2AVY	0.415	A 371	C 372	A 374	U 375	A 389	U 390	s35	cWW	ntWH	s35	cWS	cWW	s35	ACA-AU.....A--U

What do we learn?

- Positions of insertions/deletions
- Base-pair co-variations
- Base conservations
- Problem: Limited number of examples

RNA3Dmotif
(Djelloul & Denise, 2008)

Q: Given a structure, how to identify “unknown” motifs within it?

1. Identify all secondary structure elements of the RNA tertiary structure;

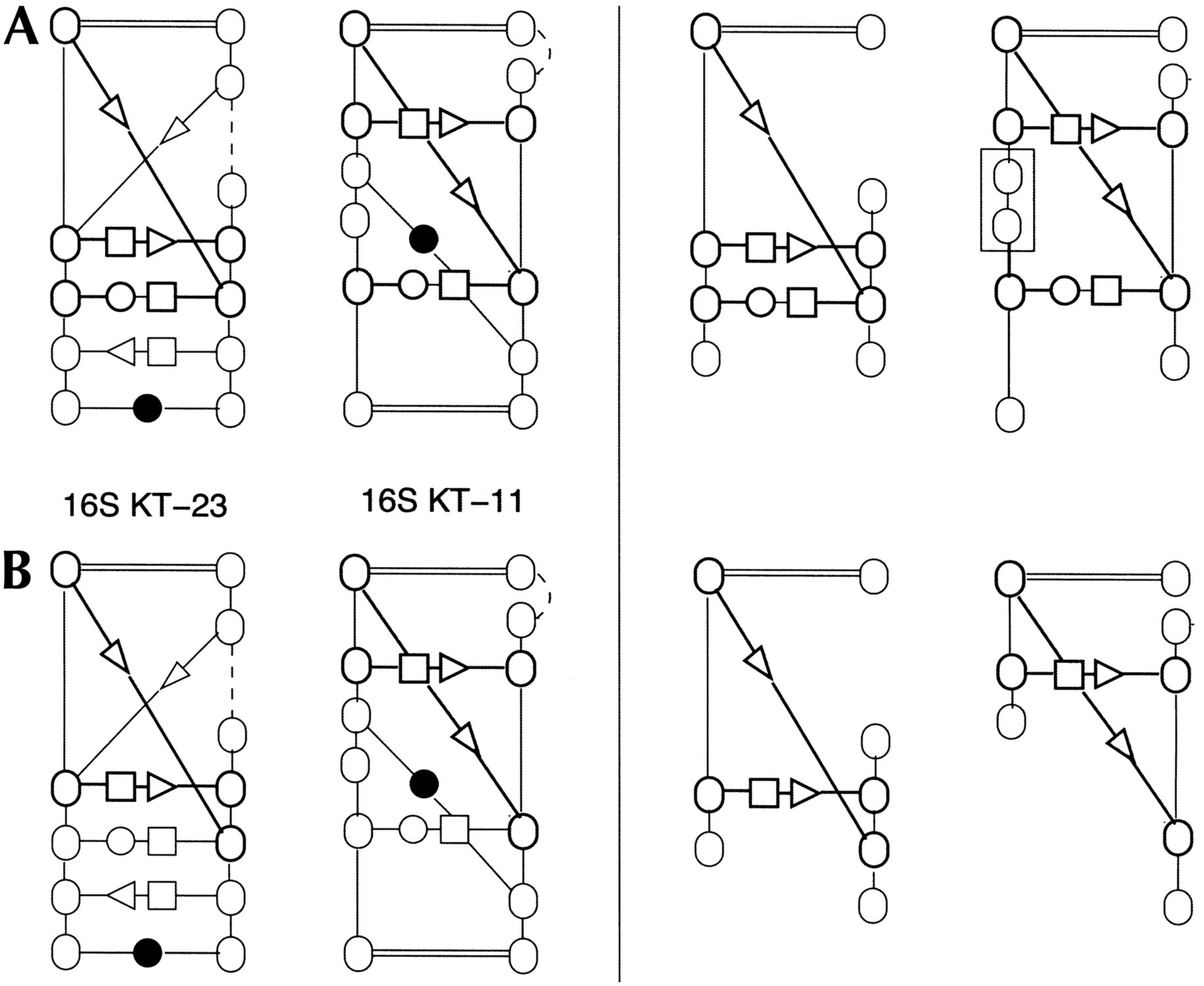
   Rationale: motifs as “often embedded within regular helical regions forming internal loops, but may also comprise hairpin or junction loops.”)

2. Calculate a similarity measure for each pair of structural elements;

   Rationale: Computing the largest extensible common noncanonical subgraph.

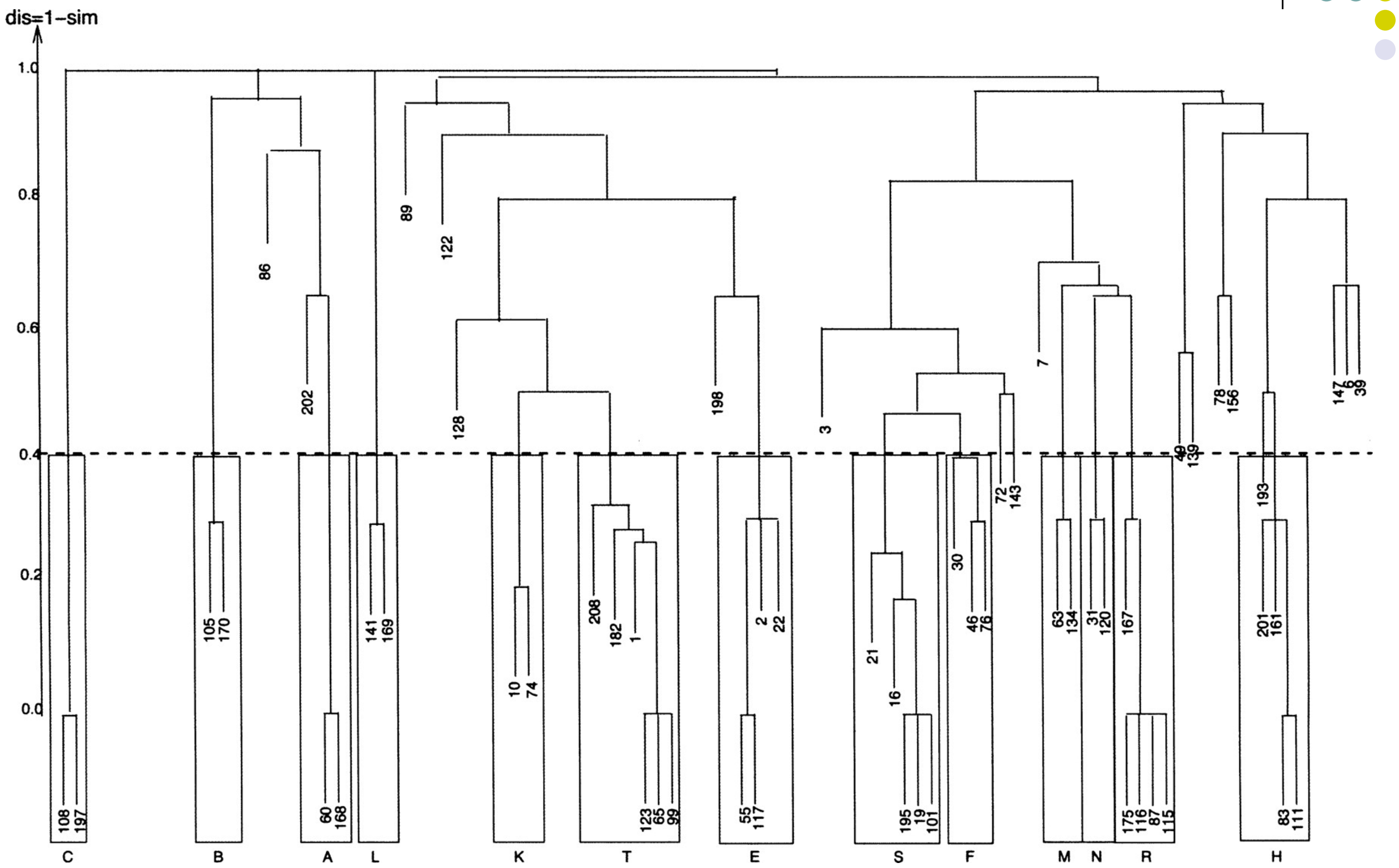
3. Cluster the structural elements according to the similarity measure.

Two structural elements containing 16S K-turn motifs.



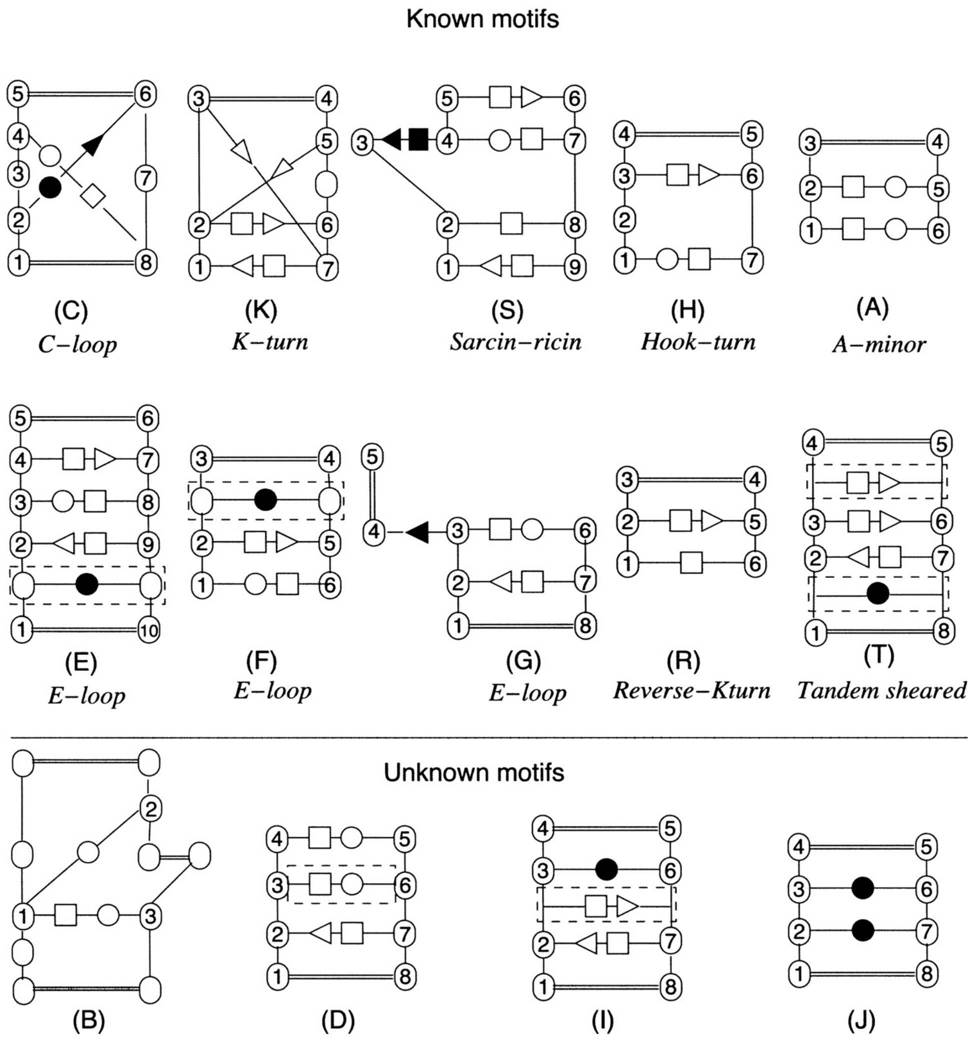
Djelloul M , Denise A RNA 2008;14:2489-2497

Dendrogram of hierarchical clustering of H.m 23S RNA produced with hclust.



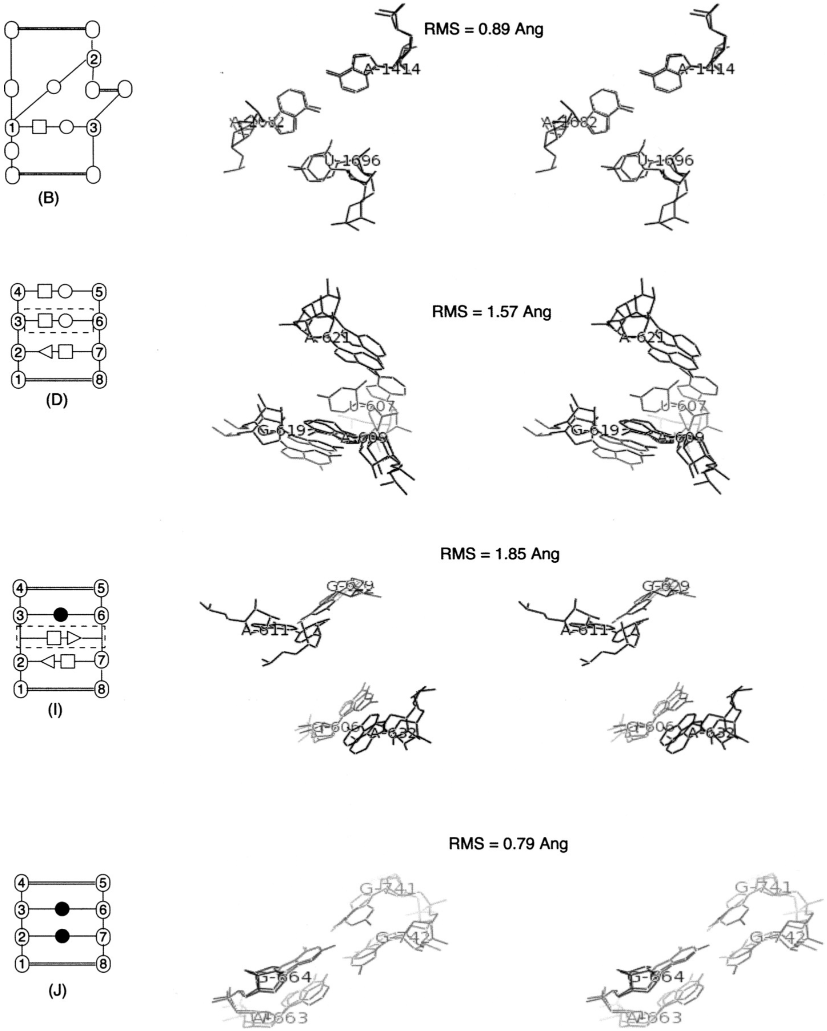
Djelloul M , Denise A RNA 2008;14:2489-2497

Recurrent motifs found in ribosomal structures.



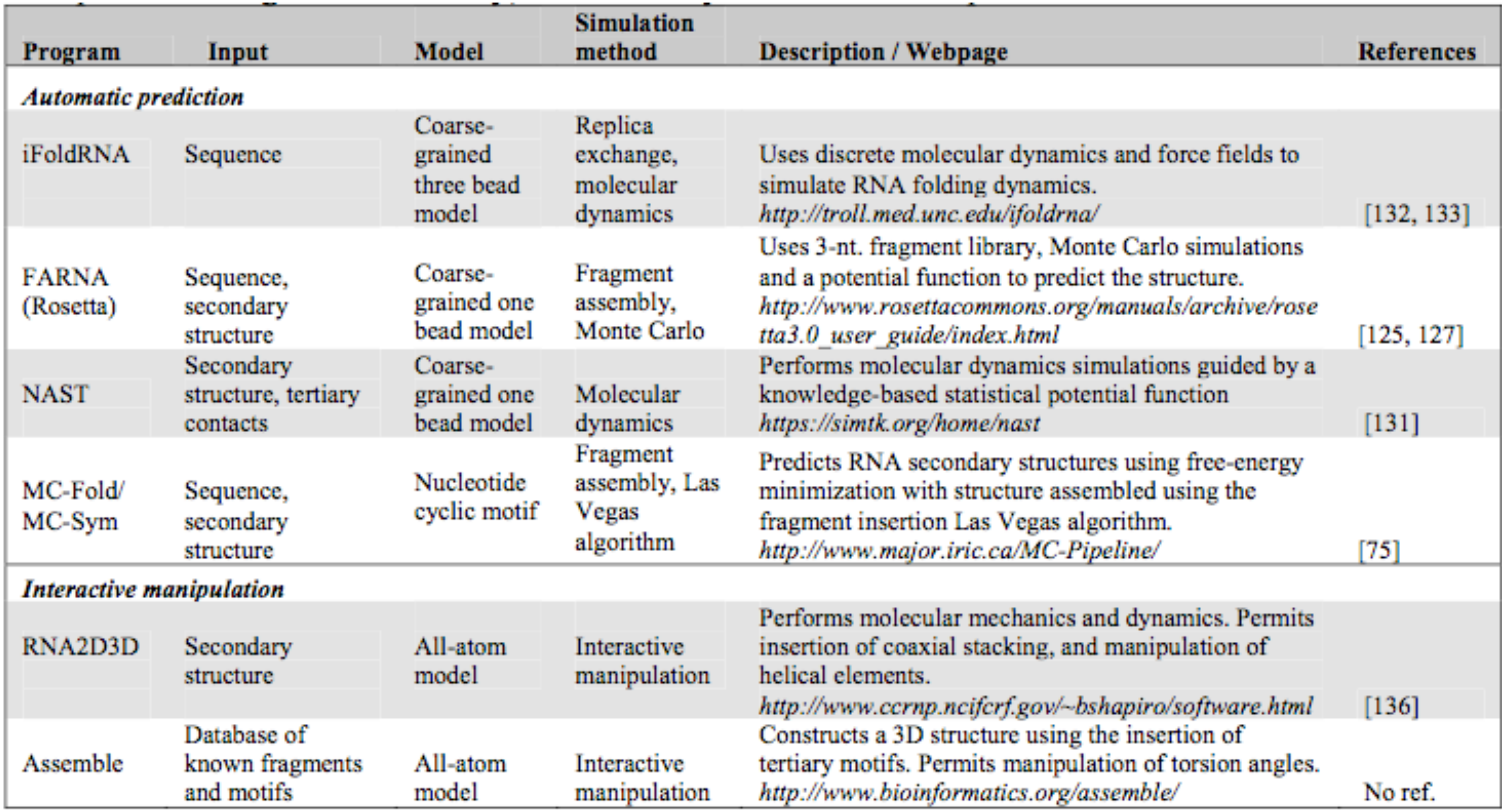
Djelloul M , Denise A RNA 2008;14:2489-2497

Crystal structures of four putative new motifs superimposed.



Djelloul M , Denise A RNA 2008;14:2489-2497

Predicting RNA 3D structures



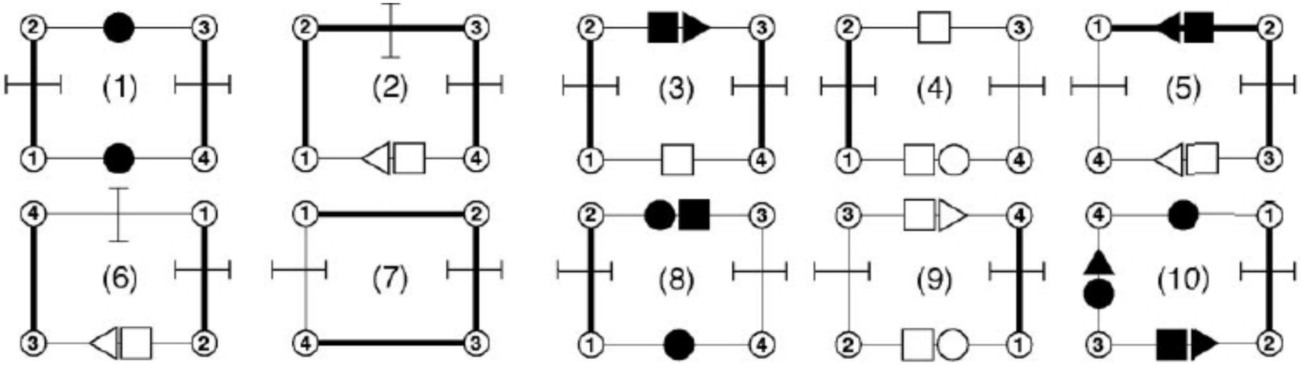
Program	Input	Model	Simulation method	Description / Webpage	References
Automatic prediction					
iFoldRNA	Sequence	Coarse-grained three bead model	Replica exchange, molecular dynamics	Uses discrete molecular dynamics and force fields to simulate RNA folding dynamics. *http://troll.med.unc.edu/ifoldrna/*	[132, 133]
FARNA (Rosetta)	Sequence, secondary structure	Coarse-grained one bead model	Fragment assembly, Monte Carlo	Uses 3-nt. fragment library, Monte Carlo simulations and a potential function to predict the structure. *http://www.rosettacommons.org/manuals/archive/rosetta3.0_user_guide/index.html*	[125, 127]
NAST	Secondary structure, tertiary contacts	Coarse-grained one bead model	Molecular dynamics	Performs molecular dynamics simulations guided by a knowledge-based statistical potential function *https://simtk.org/home/nast*	[131]
MC-Fold/ MC-Sym	Sequence, secondary structure	Nucleotide cyclic motif	Fragment assembly, Las Vegas algorithm	Predicts RNA secondary structures using free-energy minimization with structure assembled using the fragment insertion Las Vegas algorithm. *http://www.major.iric.ca/MC-Pipeline/*	[75]
Interactive manipulation					
RNA2D3D	Secondary structure	All-atom model	Interactive manipulation	Performs molecular mechanics and dynamics. Permits insertion of coaxial stacking, and manipulation of helical elements. *http://www.ccrnp.ncifcrf.gov/~bshapiro/software.html*	[136]
Assemble	Database of known fragments and motifs	All-atom model	Interactive manipulation	Constructs a 3D structure using the insertion of tertiary motifs. Permits manipulation of torsion angles. *http://www.bioinformatics.org/assemble/*	No ref.

(Laing & Schlick, 2010)

Modeling and predicting RNA 3D structure: MC-Fold | MC-Sym pipeline

(F. Major group, UdM)

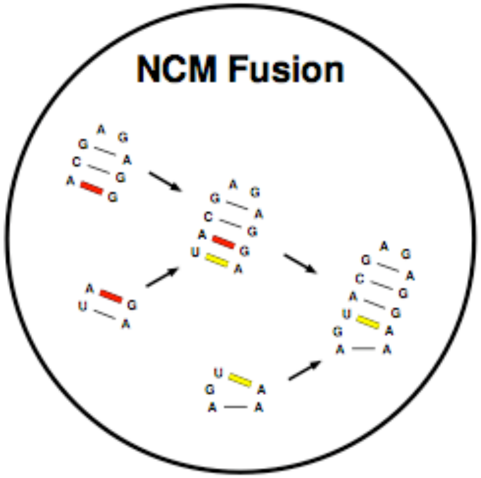
Cycle decomposition of the 3D structure using the Leontis-Westhof nomenclature.



	#	Class	Base pairs	LSU index				Comment
(1)	637	LS-P-LS-P	(W/W,W/W)	02562	02563	02570	02571	Watson-Crick tandem
(2)	21	L-LS-LS-P	(H/S)	02696	02697	02698	02699	GNRA loop
(3)	19	LS-P-LS-P	(H/S,H/H)	01532	01533	01658	01659	Non Watson-Crick tandem
(4)	10	LS-P-S-P	(H/H,W/H)	977	979	9103	9104	Non Watson-Crick tandem
(5)	8	LP-LS-P-S	(S/H,H/S)	01971	01972	01973	02009	Non Watson-Crick tandem
(6)	7	LS-P-L-S	(H/S)	01097	01098	01258	01259	GNRA interior loop
(7)	6	L-LS-L-S		01392	01393	01394	01395	Double-stacked bulge
(8)	6	LS-P-S-P	(W/H,W/W)	02118	02276	02277	02470	Non Watson-Crick tandem
(9)	5	P-S-P-LS	(W/H,H/S)	0481	0485	0486	0482	Non Watson-Crick tandem
(10)	5	LS-P-P-P	(S/H,W/S,W/W)	01231	02498	02522	02523	Base triple

MC-Fold workflow

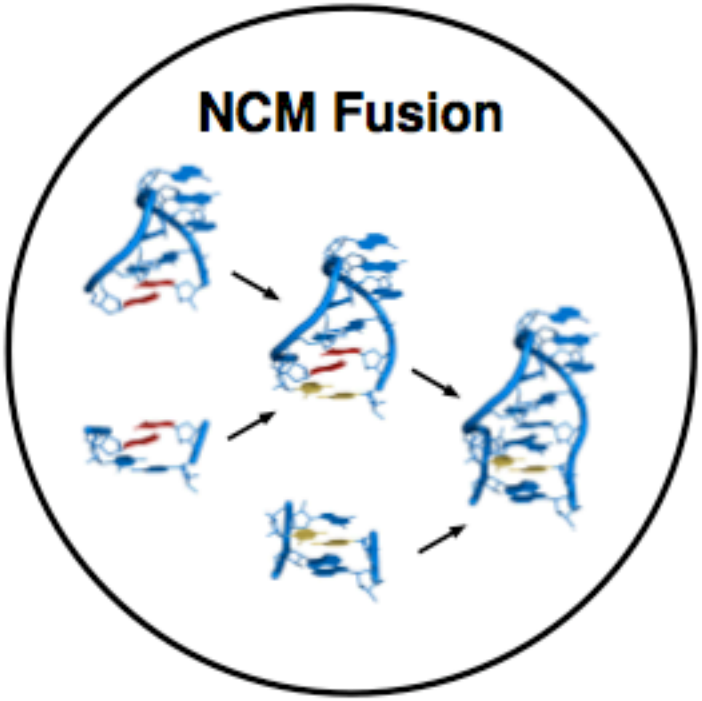
```
>SRL
GGGUGCUCAGUACGAGAGGAACCGCACCC
```



```
>SRL
GGGUGCUCAGUACGAGAGGAACCGCACCC
(((((((((.((((..))))))))))))) -28.38
(((((((.(.((((..))))).))))))) -27.48
((((((..(.((((..)))))..)))))) -27.25
((((((((.(..))))(((..)))))))) -27.17
(((((((((.(((....)))))))))))) -27.15
```

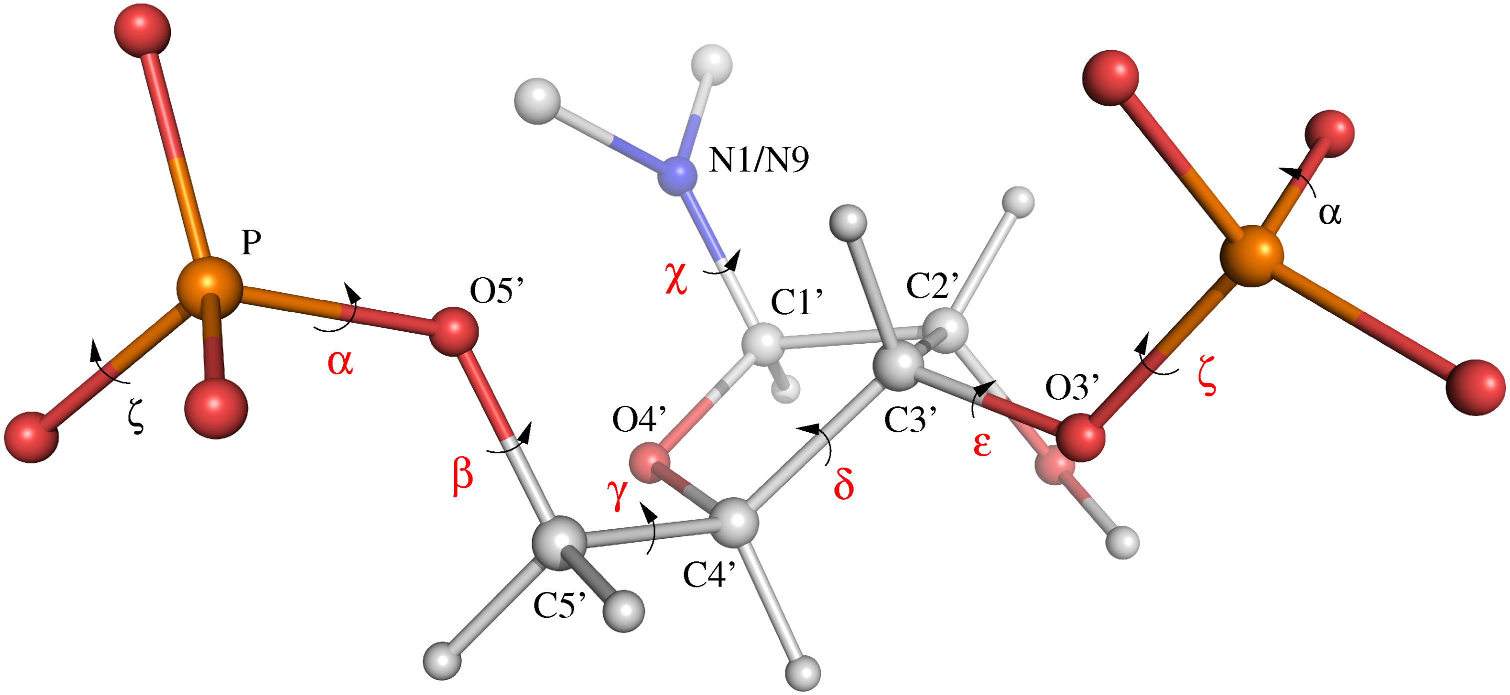
MC-Sym workflow

```
>SRL
GGGUGCUCAGUACGAGAGGAACCGCACCC
(((((((((.((((..)))))))))))))
```



Beyond conserved 3D motifs

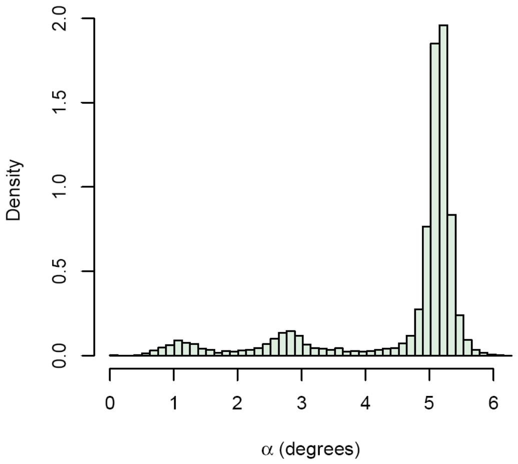
The 3D structure can be modeled by enumeration of the degree of freedom of the polynucleotide.

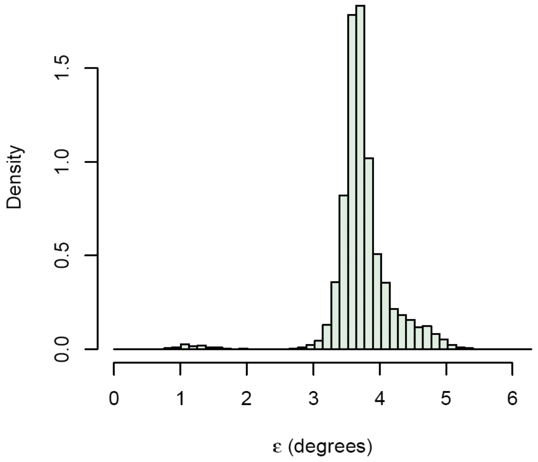


Each nucleotide in an RNA molecule can be represented by the base type and 7 dihedrals angles

A continuous probabilistic model of local RNA 3D structure (Jes Frellsen et al.)

Modeling and estimating the angle distributions.

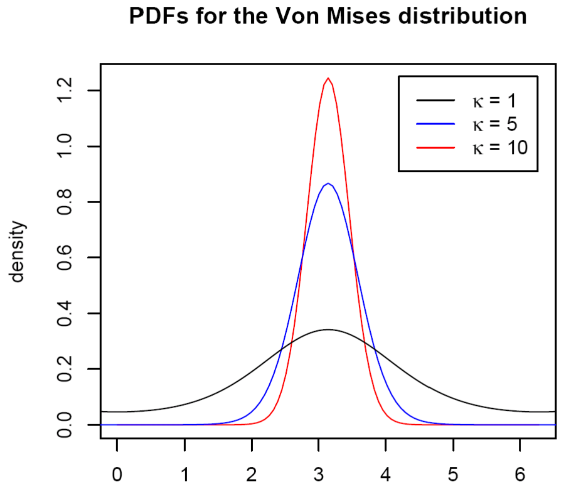




- Each variable lies on a circle
  - Requires directional statistics

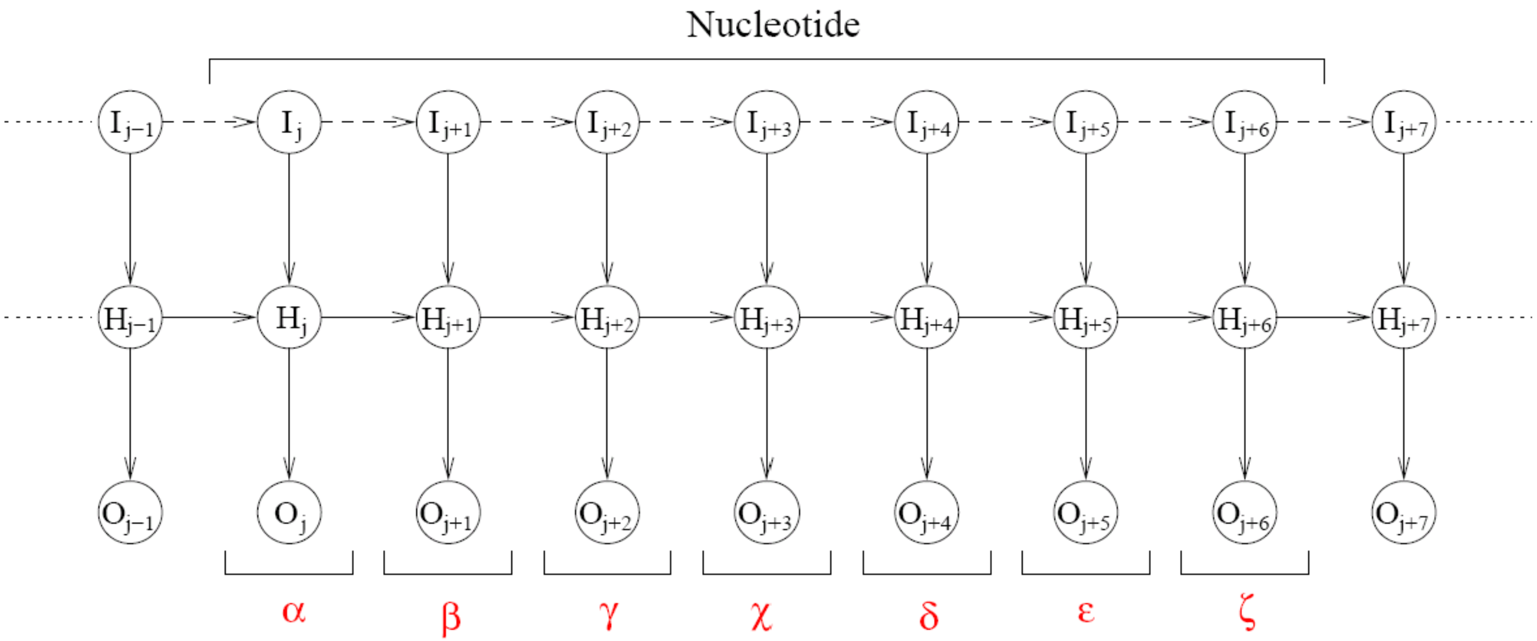
- Each variable is multi-modal
  - Can be described by a mixture of simple distributions
  - Von Mises distribution

- The angles co-vary both within nucleotides and between consecutive nucleotides
  - We model this by a sequential model



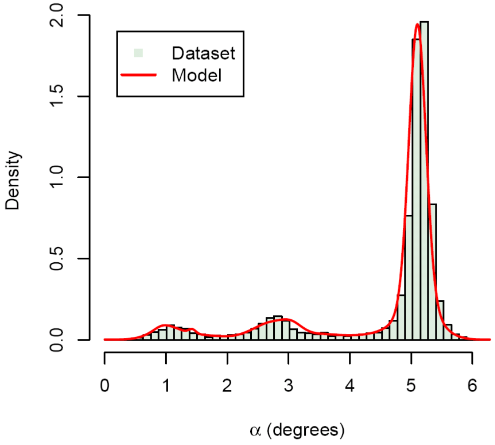
A continuous probabilistic model of local RNA 3D structure

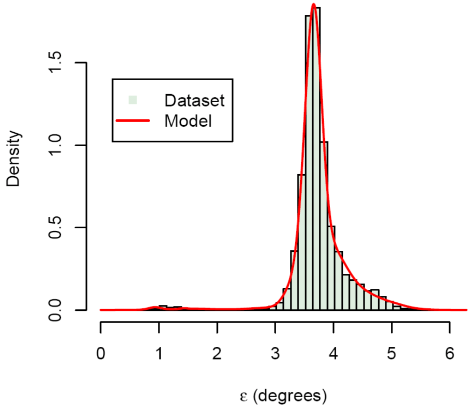
- An DBN with 3 random variables per angle:
  - Discrete input variable indicating angle type (7 states)
  - Hidden variable with 20 states
  - Output variable representation the angle value and the CPDs given the hidden state is modelled by Von Mises distributions
- Structure of an IOHMM with continuous output (except bookkeeping)
- Does not impose a groping of the angles
- Parameters are estimated by stochastic EM from experimental data

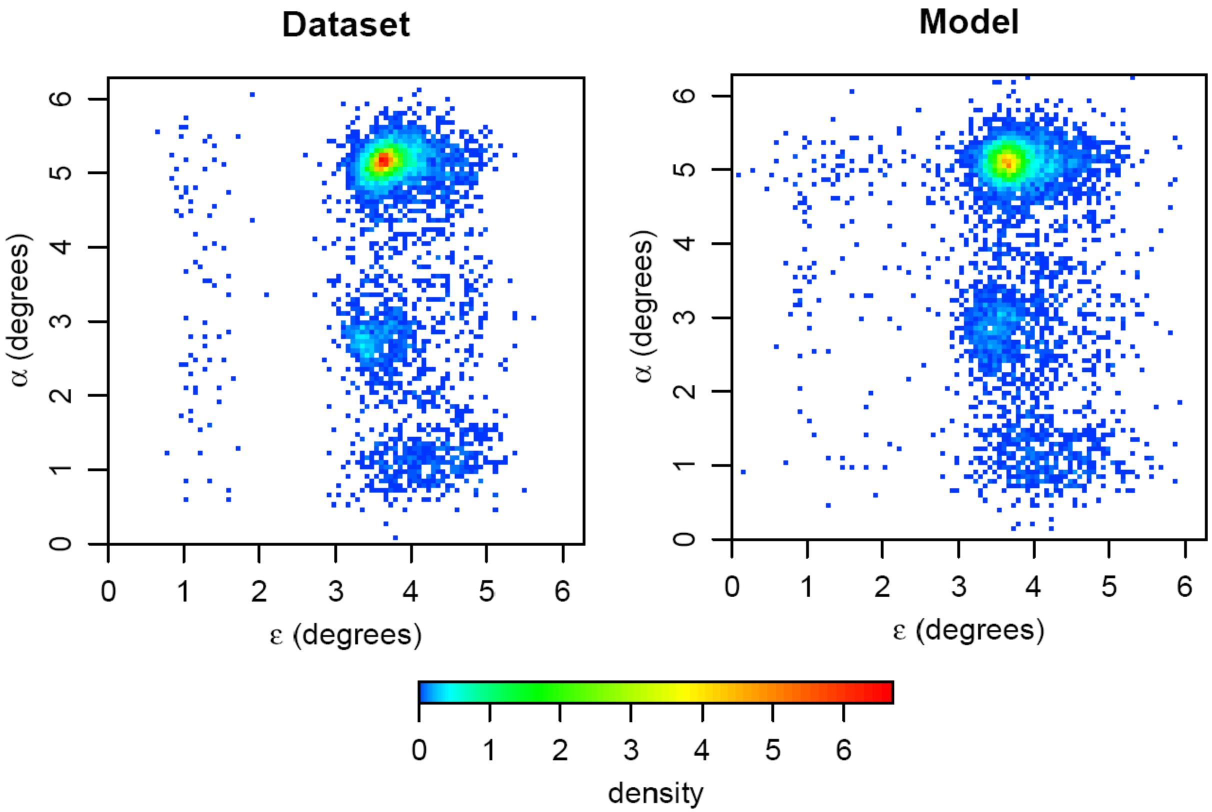


A continuous probabilistic model of local RNA 3D structure

- The model captures the distribution of the individual angles
- The model captures the pairwise dependencies between the angles







A continuous probabilistic model of local RNA 3D structure

Generation of decoy with s simple simulated annealing scheme:

1. Sample a whole structure, *S,* without clashes
2. Make new structure, *S′*, by resampling four consecutive angles in *S* (randomly picked)
3. Evaluate *S′*
   a. If it has clashed it is rejected
   b. If it has a better energy than *S* then *S′* is set to be the new *S*
   c. If it has a worse energy then with probability, *p*, *S′* is set to be the new *S* (otherwise it is rejected)
   d. Go to step 2

In the scheme we used

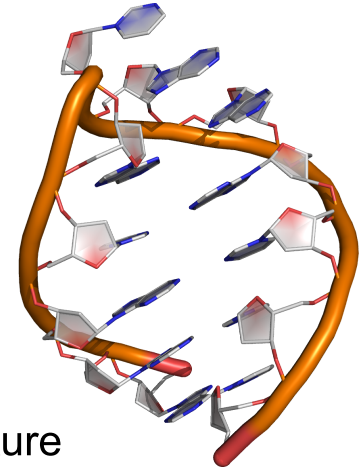
- $p = e^{(E-E')}/T$, where *T* decreases with time
- a simple “energy function” that promotes structure with the same Watson-Crick base pair as are found in the target structure

A continuous probabilistic model of local RNA 3D structure: Results

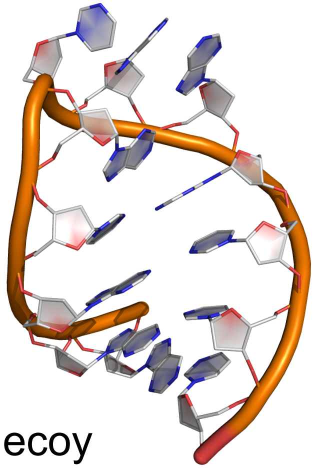
Target Structure	Length (Bases)	Decoys < 4Å	Decoys < 3Å	Lowest RMSD
1ZIH	**12**	58.8%	21.3%	1.55Å
1RNG	**12**	55.1%	3.5%	2.48Å
1XWP	**13**	28.3%	5.8%	2.03Å
1I4B	**13**	34.6%	0.1%	2.91Å
1PJY	**22**	10.0%	1.9%	1.89Å

Results computed from 1500 decoys

1ZIH



Target structure



Best decoy