RNA structure:
motif search; RNA 3D predictions

Comparative RNA structure analysis

A powerful approach in RNA structure prediction, in particular, due to RNA-specific patterns of variation, nucleotide covariations.

An example of two covariations in three related RNA's:

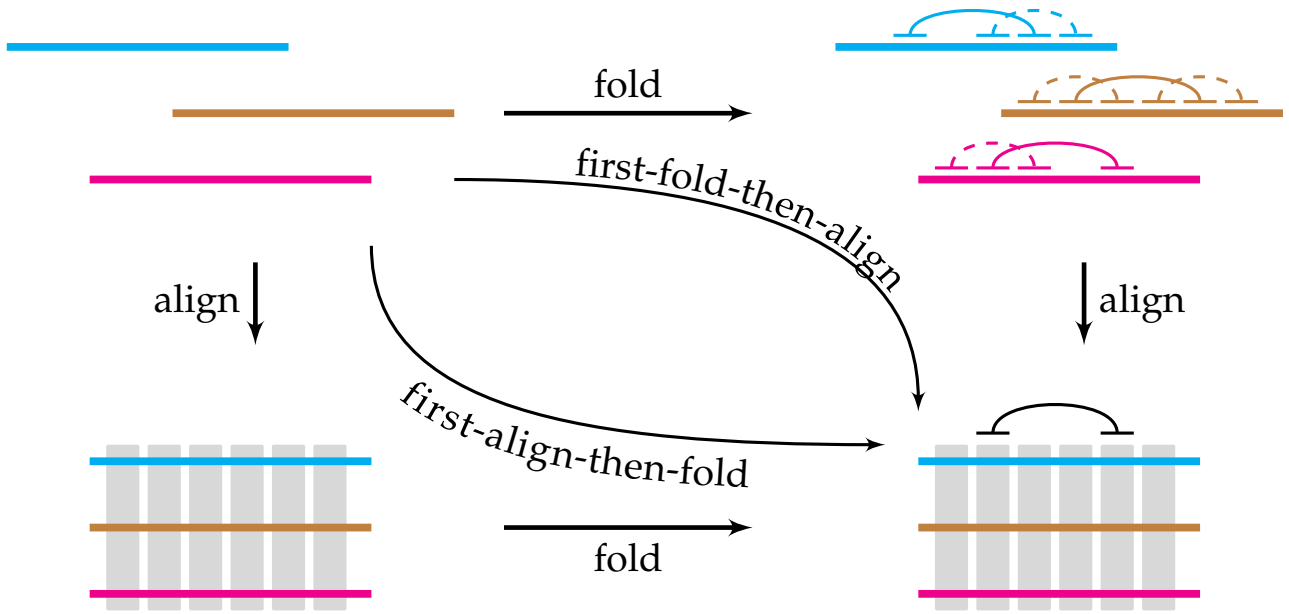
```
 n n              n n              n n
n   n            n   n            n   n
 n-n              n-n              n-n
 G-C              U-A              A-U
 n-n              n-n              n-n
 n-n              n-n              n-n
 A-U              G-C              C-G

RNA 1            RNA 2            RNA 3
```

```
AnnGnnnnnnCnnU    RNA 1
GnnUnnnnnnAnnC    RNA 2
CnnAnnnnnnUnnG    RNA 3
((((....))))     consensus "bracket view"
```

Detecting conserved structures in related RNAs

(prediction of "consensus" structures)

Different strategies:



from Pervouchine (2018)

Detecting conserved structures in related RNAs

(prediction of "consensus" structures)

Consensus structures can be computed from sequence alignments using information from suboptimal structures, base probabilities and covariation patterns

Input: Sequence alignment

Calculation: suboptimal structures/partition functions/base probabilities for individual sequences; detection of common patterns and their scoring

Output: The "consensus" structure, (ideally) conserved in all sequences of the dataset.

Detecting conserved structures in related RNAs

(prediction of "consensus" structures)

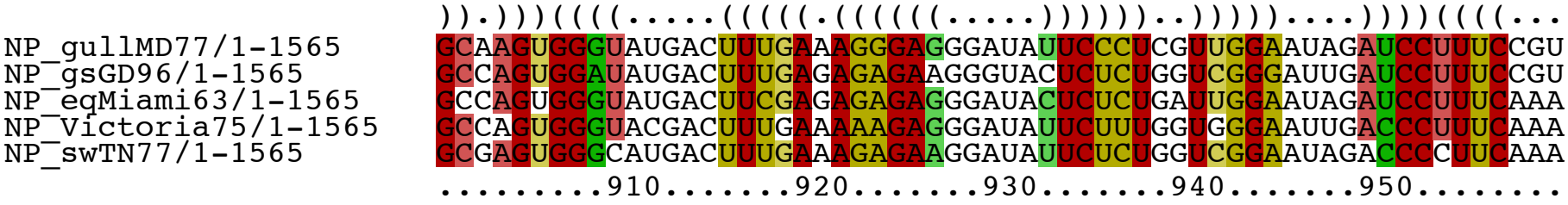
Consensus structures can be computed from sequence alignments using information from suboptimal structures, base probabilities and covariation patterns

Input: Sequence alignment

Calculation: suboptimal structures/partition functions/base probabilities for individual sequences; detection of common patterns and their scoring

Output: The "consensus" structure, (ideally) conserved in all sequences of the dataset.

For instance, a fragment of the output of RNAalifold algorithm:



Detecting conserved structures in related RNAs

(prediction of "consensus" structures)

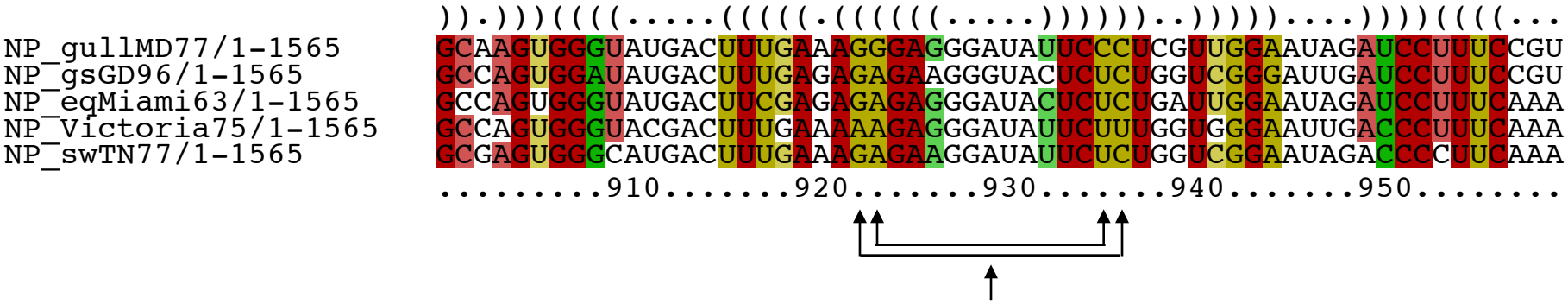
Consensus structures can be computed from sequence alignments using information from suboptimal structures, base probabilities and covariation patterns

Input: Sequence alignment

Calculation: suboptimal structures/partition functions/base probabilities for individual sequences; detection of common patterns and their scoring

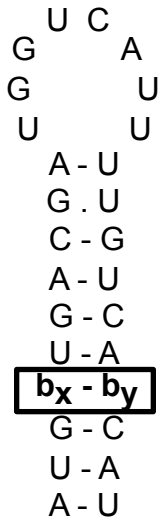
Output: The "consensus" structure, (ideally) conserved in all sequences of the dataset.

For instance, a fragment of the output of RNAalifold algorithm:



Such structure-annotated alignments allow one to identify covariations.

Mutual information and alignment position entropies



Covariation x/y (?)

Mutual information M(x,y):

$$M(x,y) = \sum_{b_x,b_y \in (A,G,C,U)} f(b_x b_y) \cdot log_4 \frac{f(b_x b_y)}{f(b_x) f(b_y)}.$$

Using entropy values at the alignment positions:

$$M(x,y) = H(x) + H(y) - H(x,y),$$

where

$$H = -\sum f(b) \cdot log_4 f(b)$$

(an entropy term, a measure of variability).

Alignment:

```
   x                          y     positions
AUGUUGACGAUGGUCAUUUUGUCACAU  seq1
AUGCUGACGAUGGUCAUUUUGUCGCAU  seq2
AUGCUGACGAUGGUCAUUUUGUCGCAU  seq3
AUGCUGACGAUGGUCAUUUUGUCACAU  seq4
AUGUUGACGAUGGUCAUUUUGUCGCAU  seq5
AUGUUGACGAUGGUCAUUUUGUCACAU  seq6
...
AUGCUGACGAUGGUCAUUUUGUCGCAU  seqN
  f(bx)                    f(by)  base frequencies
```

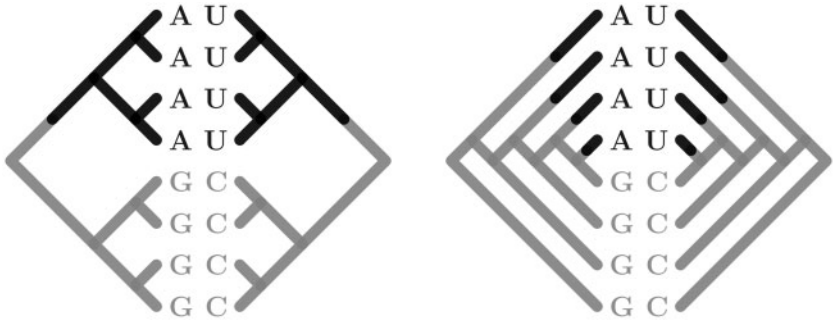
The ratios of *M(x,y)* and entropies can reveal correlations at (biased) positions:

$$R_1(x,y) = \frac{M(x,y)}{H(x)}, \qquad R_2(x,y) = \frac{M(x,y)}{H(y)}.$$

High values (close to 1) indicate to significant correlations.

[Gutell et al., 1992]

Mutual information and numbers of covariation events



$$f_{12}(AU) = f_{12}(GC) = \frac{1}{2} \qquad f_1(A) = f_1(U) = f_2(G) = f_2(C) = \frac{1}{2}$$

$$MI = \sum_X \sum_Y f_{12}(XY) \log_4 \left(\frac{f_{12}(XY)}{f_1(X) \cdot f_2(Y)} \right) = 0.5$$

Similar values of MI may reflect different evolutionary scenarios. The scenario on the right is a stronger case for coevolution hypothesis (multiple covariation events).

from Dutheil (2012)

Alignment:

```
   1                    2     positions
...A..................U...  seq1
...A..................U...  seq2
...A..................U...  seq3
...A..................U...  seq4
...G..................C...  seq5
...G..................C...  seq6
...G..................C...  seq7
...G..................C...  seq8
  f1()                 f2()
```

Covariance models, RNA families and RNA descriptors

One of the core computational problems in RNomics is a so-called "sequence/structure" alignment.

For instance, a problem to align a motif
<<<<.<<<<.....>>>>>>>>
to a sequence:
CCCCACGCGAAAACGCGGGGG

Obviously, a deletion in the sequence yields the best alignment (score):

<<<<.<<<<.....>>>>>>>>
CCCCACGCG-AAAACGCGGGGG

Various algorithms are possible for the search of the optimal sequence/structure alignments (dynamic programming, BLAST-like etc.). They can be used e.g. for the alignment of a structural motif to a sequence (database of sequences), alignment of a sequence to a motif (database of motifs).

Similar ideas can be used in fold/align algorithms (simultaneously folding and aligning RNA sequences).

Multiple sequence/structure alignments lead to definitions of RNA families and descriptors.

Rfam: database of RNA families

http://rfam.sanger.ac.uk/

In Rfam, the related RNAs (families) are stored as sequence/structure alignments (multiple sequence alignments + structure motifs in the Stockholm format)

```
...
Influenza_A_virus_AN.1      UUCCAGGACAUACUAAUGAGGAUGUCAAAAAUGCAAUUGGGAUUCUCA
Influenza_A_virus_Ac.8      UGCCAGGACAUUCUGCGGAGGAUGUCAAAAAUGCAAUUGGGAUCCUCA
Influenza_A_virus_AL.1      UUCCAGGACAUACUGCUGAGGAUGUCAAAAAUGCAGUUGGAGUCCUCA
Influenza_A_virus_Ap.1      UGCCAGGACAUUCUCAUGAGGAUGUCAAAAAUGCAAUUGGAAUCCUCA
...
#=GC SS_cons                .<<<<<..........AAAAAA.............>>>>>aaaaaa.
```

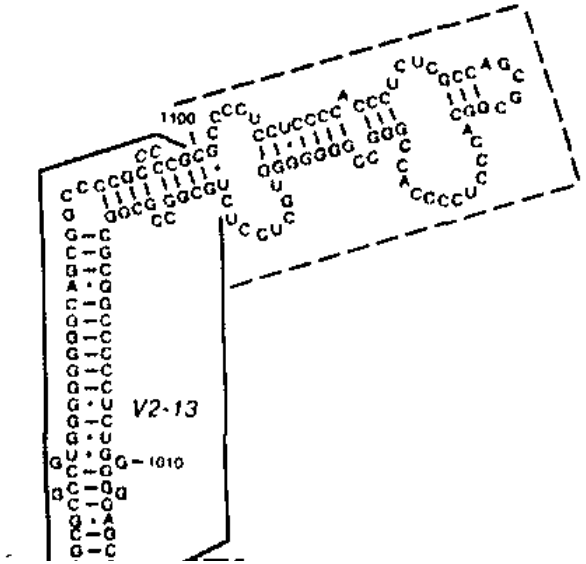
(In Stockholm format, the pseudoknots are shown with AAA...aaa; BBB...bbb etc symbols.)

Every family in Rfam is initially defined by “seed” alignments: representative sequences plus structural motif. These seed alignments define a descriptor (covariance model). The covariance model is further used to search for other family members in a sequence database.

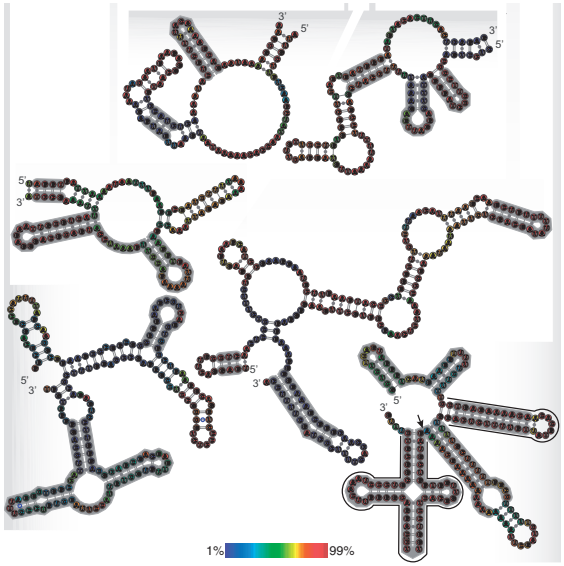
Structured RNA molecules without protein-coding function:

- tRNA
- ribosomal RNA (rRNA)
- small nucle(ol)ar RNA (snRNA, snoRNA)
- microRNA (miRNA)
- long non-coding RNA (lncRNA)
- etc.

Non-coding RNAs (ncRNAs) are usually characterized by a conserved structure.



One of the domains in human 28S rRNA
(Gorski et al., 1987).
Length = 5035 nt



Fragments of conserved structures predicted in human long ncRNA (Smith et al., 2013)
Length ~ 7000 nt

Structured RNA molecules without protein-coding function:

- tRNA
- ribosomal RNA (rRNA)
- small nucle(ol)ar RNA (snRNA, snoRNA)
- microRNA (miRNA)
- long non-coding RNA (lncRNA)
- etc.

Non-coding RNAs (ncRNAs) are usually characterized by conserved structure.

Identification of (non-coding) RNA transcripts and/or structured RNA regions in genomes: RNomics.

Multiple databases of ncRNAs

RNAcentral database (The RNAcentral Consortium, rnacentral.org) integrates data from ncRNA resources

RNAcentral Expert Databases

5SrRNAdb	LncBase	PDBe	SILVA
CRW Site	LNCipedia	piRBase	snOPY
dictyBase	lncRNAdb	PLncDB	snoRNA Database
ENA	LncRNAWiki	PomBase	sRNAmap
Ensembl	MGI	RDP	SRPDB
FlyBase	miRBase	RefSeq	TAIR
GENCODE	miRTarBase	Rfam	TarBase
Greengenes	Modomics	RGD	tmRDB
GtRNAdb	NONCODE	RNApathwaysDB	tmRNA Website
HGNC	NPInter	SGD	tRNAdb
			WormBase

Different search tasks are possible:

Text search

Search by *gene, species, publication, author* or any other keyword

Browse sequences

Sequence search

Search for similar sequences or look up your sequence in RNAcentral

Search by sequence

Genome browser

Explore RNAcentral sequences in your favorite genome locations

Browse genomes

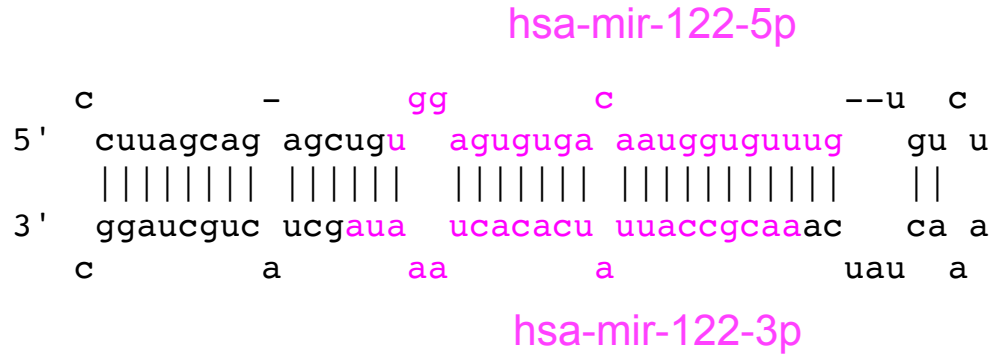
(rnacentral.org)

microRNAs (miRNAs)

MicroRNAs are 21-22 nt RNA's are derived from precursor primary miRNAs (pri-miRNAs).

Pri-miRNAs are extended stem-loops. They are enzymatically processed to yield miRNAs (below shown in colour) that can be produced from both sides of the stem-loop.

The hsa-mir-122 precursor:



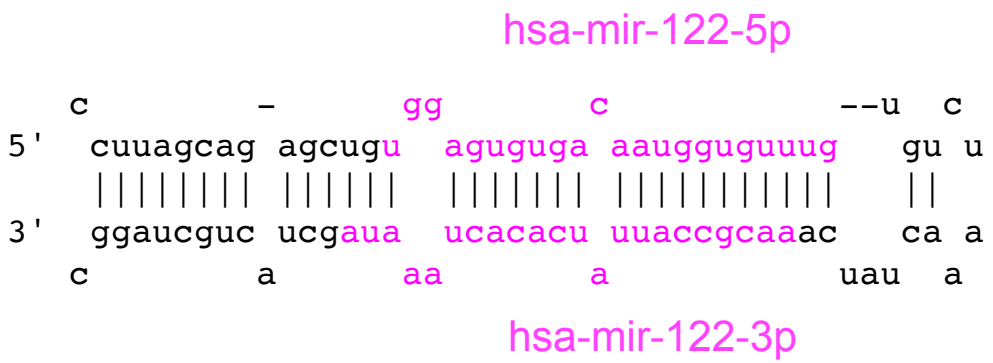
(www.mirbase.org/)

microRNAs (miRNAs)

MicroRNAs are 21-22 nt RNA's are derived from precursor primary miRNAs (pri-miRNAs).

Pri-miRNAs are extended stem-loops. They are enzymatically processed to yield miRNAs (below shown in colour) that can be produced from both sides of the stem-loop.

The hsa-mir-122 precursor:



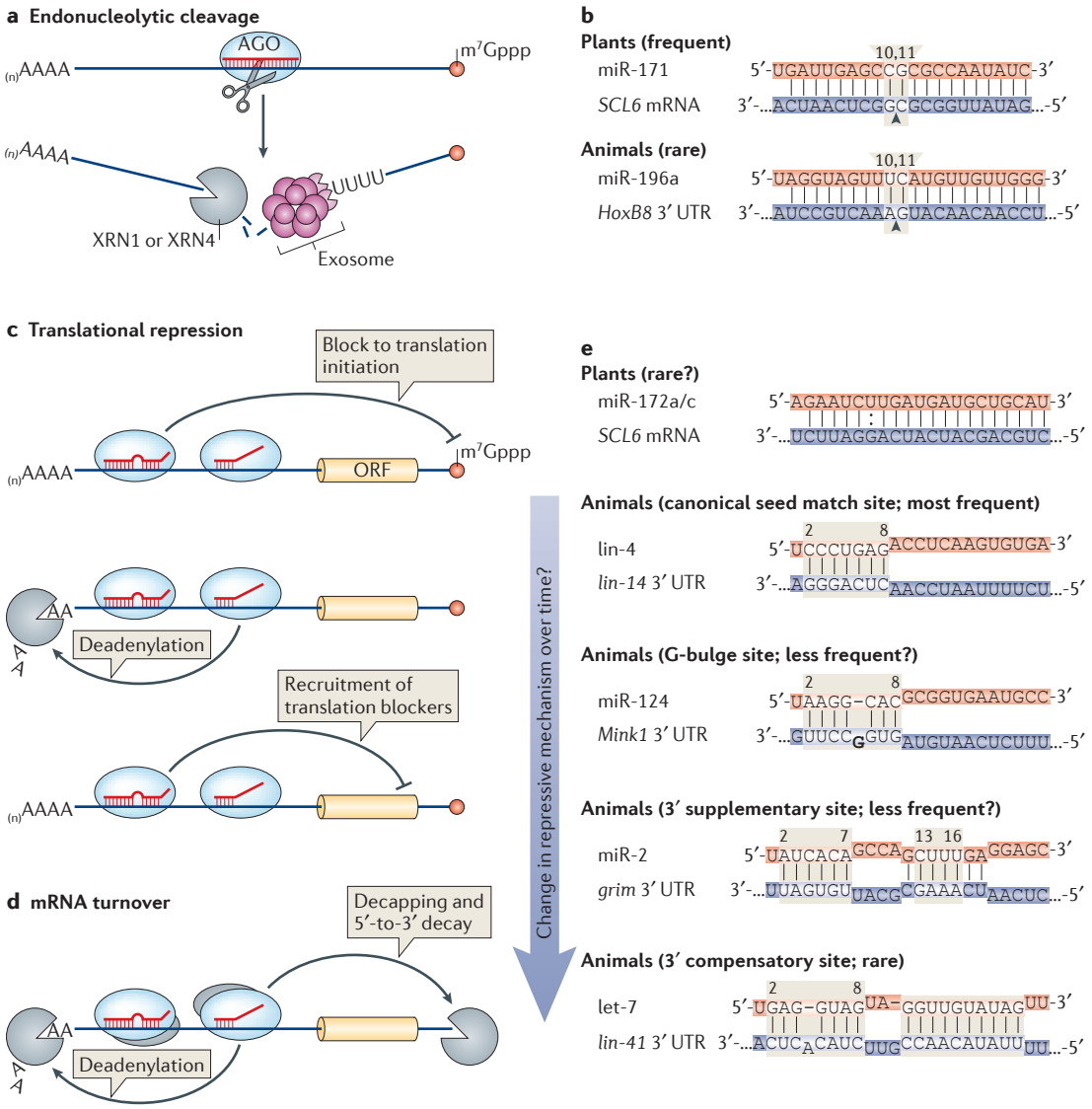
(www.mirbase.org/)

- In animals, the main function of miRNA's is translational repression mediated by miRNA binding to mRNA 3'UTR's.
- This binding is mostly determined by so-called "seed" complementary match of 7-8 base pairs between the miRNA 5'end and target. For instance:

```
5'    ...UGCCCUGGGAGCCCUACACUCCA... target mRNA
                      |||||||
3'        GUUUGUGGUAACAGUGUGAGGU    miRNA
```

microRNAs (miRNAs)

MiRNAs target genes by pairing to mRNAs. Different regulation mechanisms can be used.



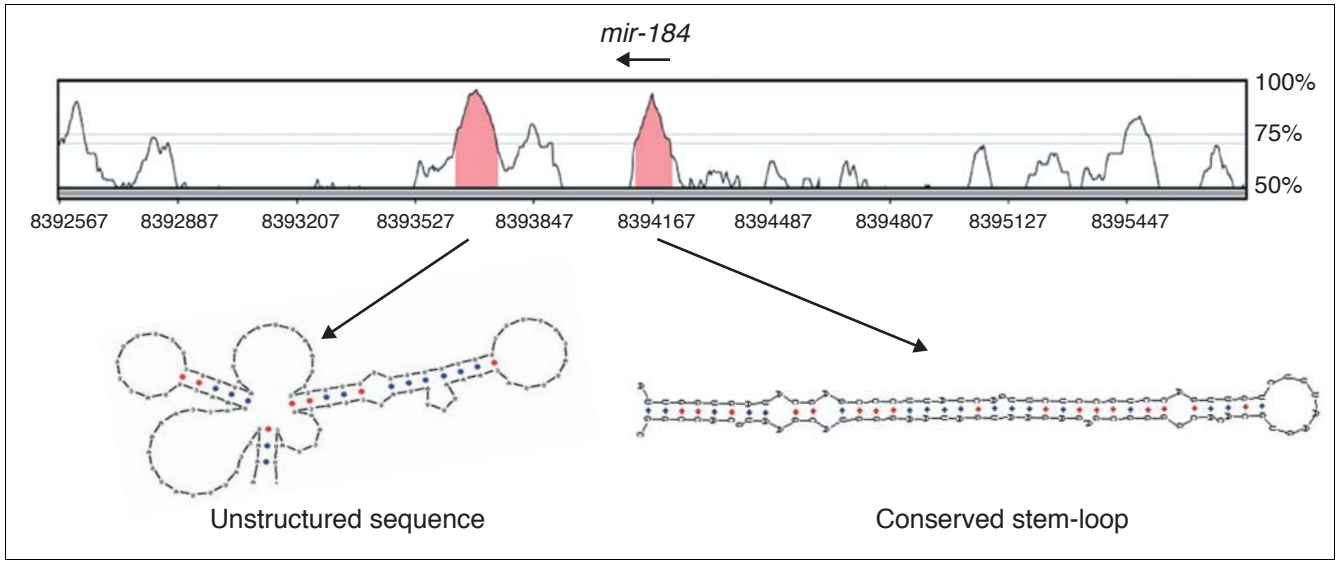
(Ameres & Zamore, 2013)

Prediction of miRNAs and their targets

Predictions of pri-miRNAs are mostly based on finding conserved stem-loop structures encoded in related genomic sequences.

Some sequence preferences can be used in the search.

Due to weak sequence patterns, such an approach may lead to many false-positive results.



(E.C. Lai et al., 2003)

RNAsnp server: predicting SNP effects on RNA folding

(http://rth.dk/resources/rnasnp/ ; Sabarinathan et al., 2013)

RNAsnp Web Server: Predicting SNP effects on local RNA secondary structure

Please fill out the submission form and click the ***Submit*** button given below. Input fields marked with a * are required.
(Load Example Data)

Input sequence*

Enter your input sequence here in either fasta format or linear sequence (without gaps). [?]

(or) Upload sequence file: Choose File no file selected

(or) Select sequence from genome database

Mammal | Human | hg19 | genome | region | chr19:49468565-494695

SNP details*

Enter your SNP details in the required format [?]

- *XposY*, *X* is the wild-type nt., *Y* is the mutant and *pos* is the position
  of nt. (pos=1 for first nucleotide in a sequence)
- In case of multiple SNPs, separate each SNP with hypen "-"
- More than one SNP to test in a single run, provide them in seperate lines

Mode

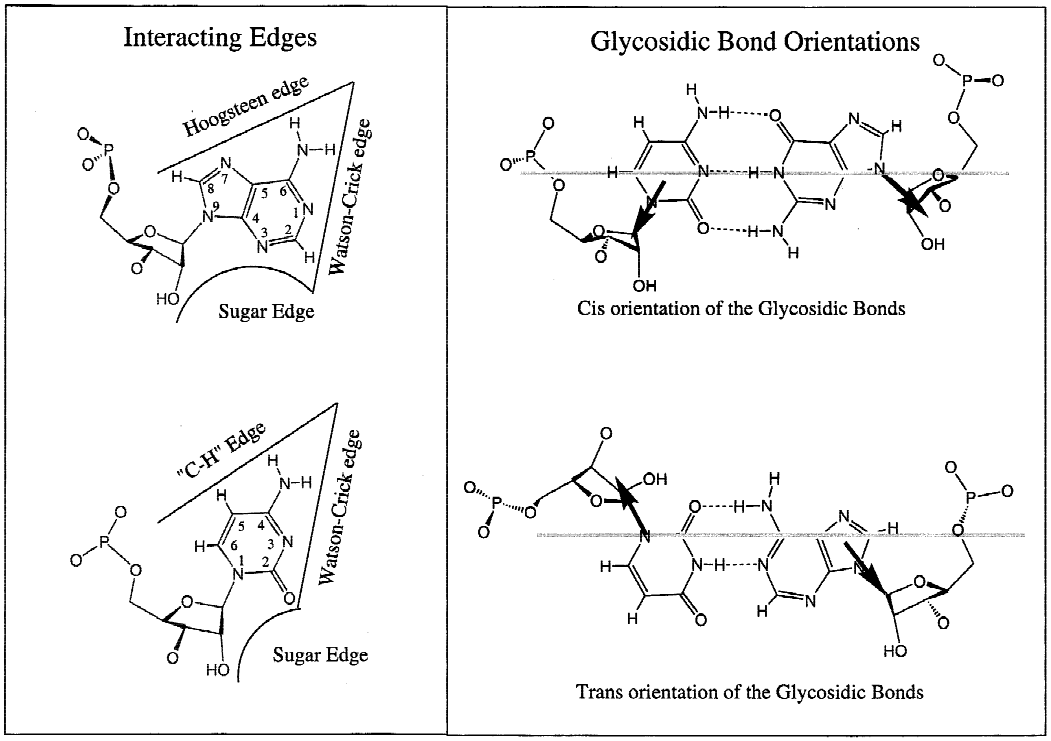
Select the mode of operation [?]

- (•) Mode 1 - based on global folding (RNAfold) [?]
- ( ) Mode 2 - based on local folding (RNAplfold) [?]
- ( ) Mode 3 - to screen putative structure-disruptive SNP [?]

Folding window

Non-canonical base pairs in RNA

In addition to canonical Watson-Crick base pairs (GC and AU), non-canonical edge-to-edge interactions with other base pairs are formed in multiple structured RNAs. These interactions are mediated by hydrogen bonds (H-bonds) and are classified according to geometries of interacting edges.



Twelve main families of *isosteric* base pairs:

No.	Glycosidic Bond Orientation	Interacting Edges	Symbol	Default Local Strand Orientation
1	*Cis*	Watson-Crick / Watson-Crick	●—●	Anti-parallel
2	*Trans*	Watson-Crick / Watson-Crick	○—○	Parallel
3	*Cis*	Watson-Crick / Hoogsteen	●—■	Parallel
4	*Trans*	Watson-Crick / Hoogsteen	○—□	Anti-parallel
5	*Cis*	Watson-Crick / Sugar Edge	●—▶	Anti-parallel
6	*Trans*	Watson-Crick / Sugar Edge	○—▷	Parallel
7	*Cis*	Hoogsteen / Hoogsteen	■—■	Anti-parallel
8	*Trans*	Hoogsteen / Hoogsteen	□—□	Parallel
9	*Cis*	Hoogsteen / Sugar Edge	■—▶	Parallel
10	*Trans*	Hoogsteen / Sugar Edge	□—▷	Anti-parallel
11	*Cis*	Sugar Edge / Sugar Edge	▶—▶	Anti-parallel
12	*Trans*	Sugar Edge / Sugar Edge	▷—▷	Parallel

(Leontis et al,, 2002)

Isosteric base pairs can substitute each other in RNA structure. Frequently a conserved non-canonical pairing can be derived from covariations in alignment.

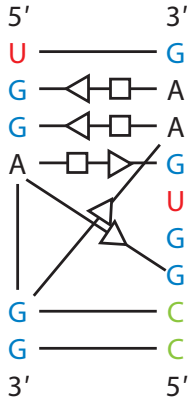
Alignment example:

```
...G.....................U...  seq1
...G.....................U...  seq2
...A.....................G...  seq3
...G.....................U...  seq4
...A.....................G...  seq5
```

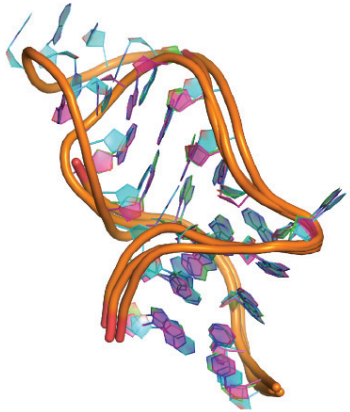
Non-canonical base pairs in RNA

Non-canonical base pairs can determine RNA 3D structure.

The kink-turn or K-turn:



2D diagram



Superimposition of K-turns from three different RNA molecules

From Miao & Westhof (2017)

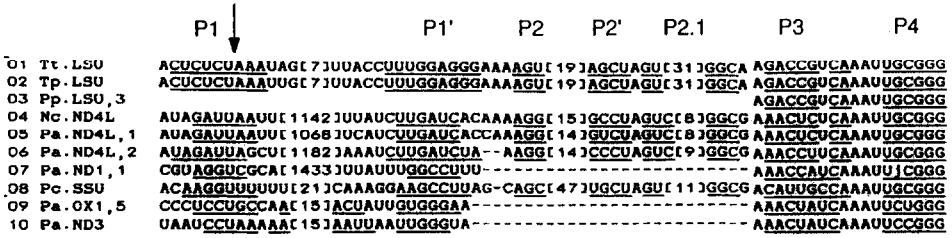
Nomenclature of non-canonical pairs:

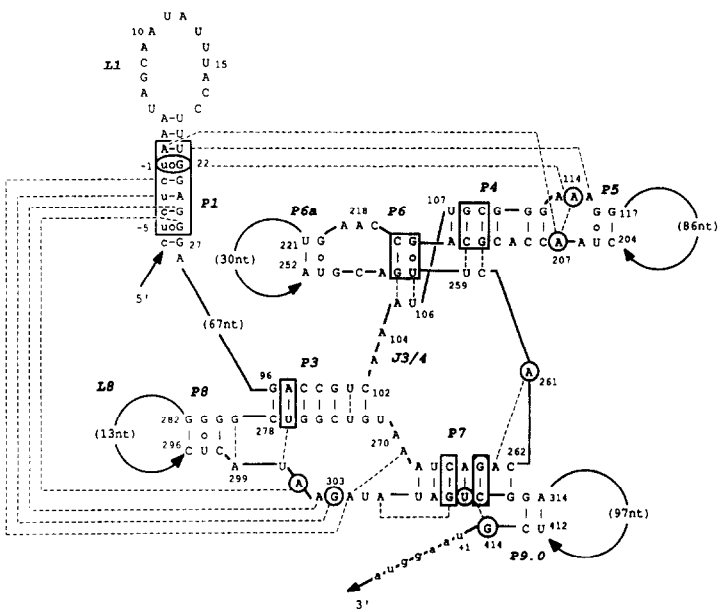
(Leontis et al,, 2002)

No.	Glycosidic Bond Orientation	Interacting Edges	Symbol	Default Local Strand Orientation
1	*Cis*	Watson-Crick / Watson-Crick		Anti-parallel
2	*Trans*	Watson-Crick / Watson-Crick		Parallel
3	*Cis*	Watson-Crick / Hoogsteen		Parallel
4	*Trans*	Watson-Crick / Hoogsteen		Anti-parallel
5	*Cis*	Watson-Crick / Sugar Edge		Anti-parallel
6	*Trans*	Watson-Crick / Sugar Edge		Parallel
7	*Cis*	Hoogsteen / Hoogsteen		Anti-parallel
8	*Trans*	Hoogsteen / Hoogsteen		Parallel
9	*Cis*	Hoogsteen / Sugar Edge		Parallel
10	*Trans*	Hoogsteen / Sugar Edge		Anti-parallel
11	*Cis*	Sugar Edge / Sugar Edge		Anti-parallel
12	*Trans*	Sugar Edge / Sugar Edge		Parallel

RNA 3D modeling using tertiary structure constraints

Comparative analysis (in particular, covariations) can identify important tertiary contacts that constrain the 3D folding. This info can be efficiently used in modeling.
For instance, the first models of ribozymes were produced using constraints derived from covariations identified in alignments guided by secondary structures. Later, these models were confirmed by crystallographic data.





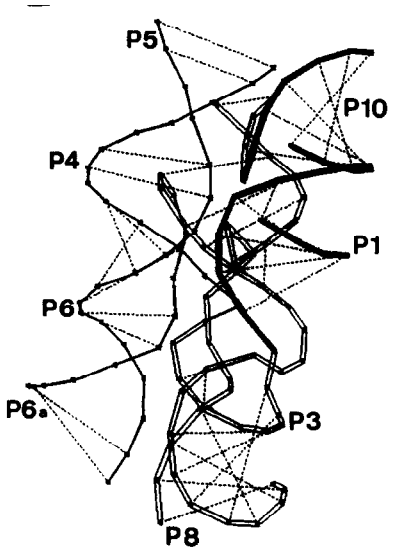
Alignment

↓

Distant contacts

↓

3D →



(Michel & Westhof, 1990)

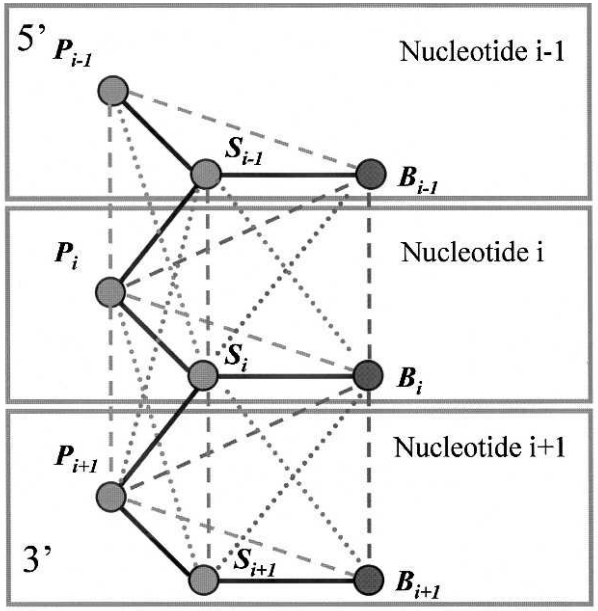
RNA 3D modeling: Molecular Dynamics

RNA 3D folding can be simulated by Molecular Dynamics (MD) approaches.

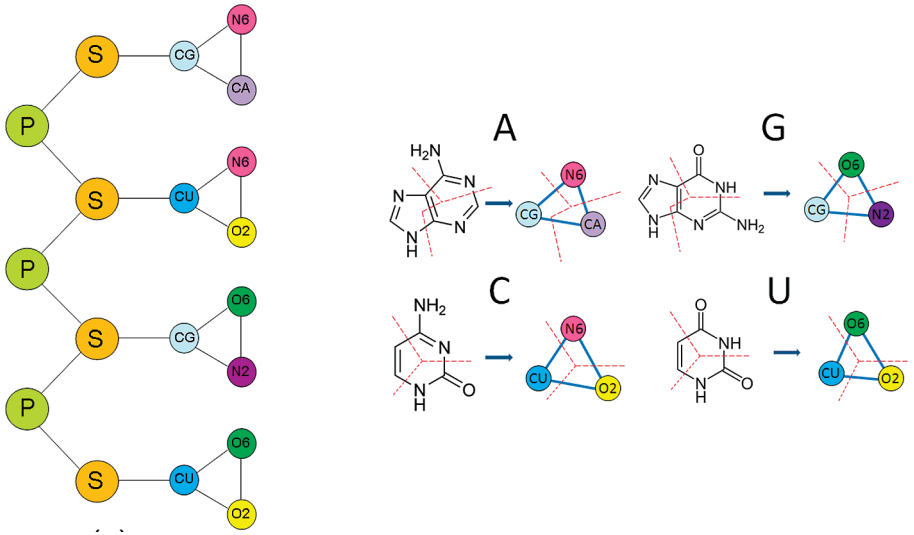
The MD simulations implement the functions (potentials) for interactions ("**force fields**") acting on atoms and molecular groups, that force them to move.

Known or predicted 2D structure is frequently used as a constraint.

A number of algorithms use simplified **coarse-grained** models with "pseudoatoms". For instance, RNA can be considered as a string with beads, with each nucleotide consisting of e.g. three (phosphate, sugar, base) or five (phosphate, sugar, three beads for a base) beads.



(F. Ding et al., 2008)



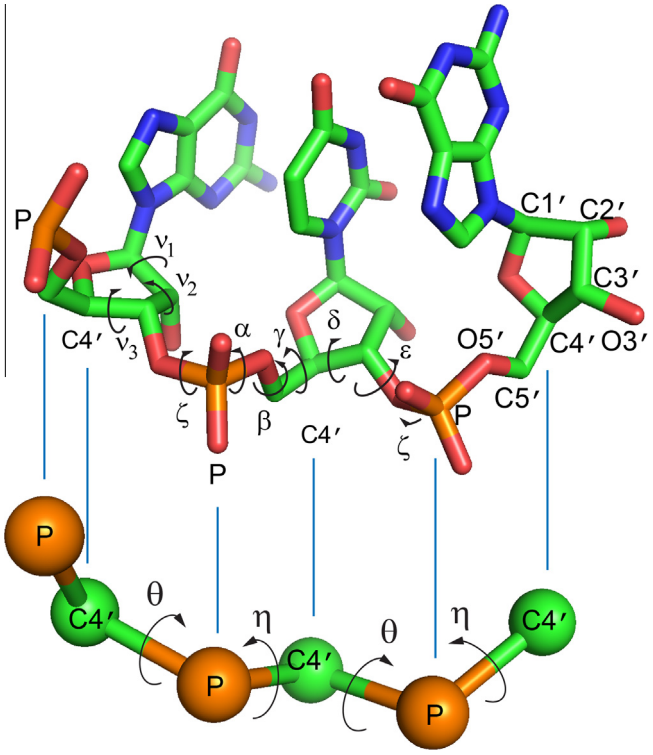
(Z. Xia et al., 2010)

RNA 3D modeling

RNA backbone can be approximated by a coarse-grained representation with virtual bonds, reducing computational complexity.

All-atom (minus H atoms) structure:

Coarse-grained backbone repesentation:

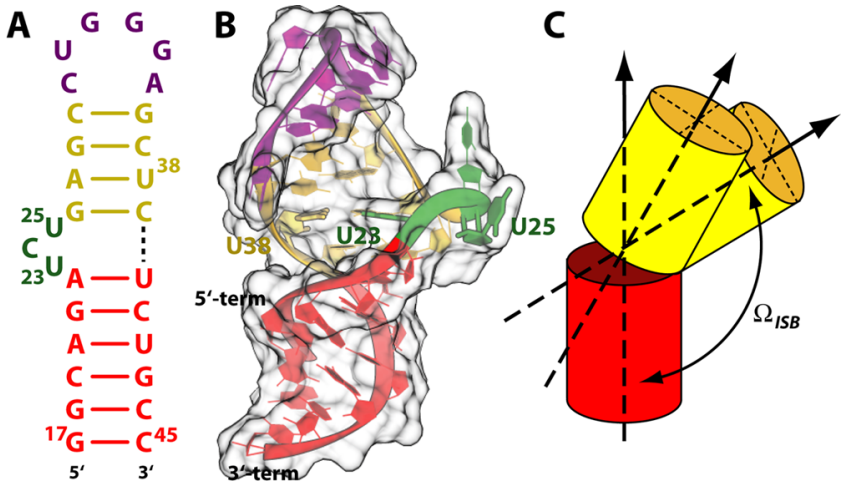


(Dawson et al, 2016)

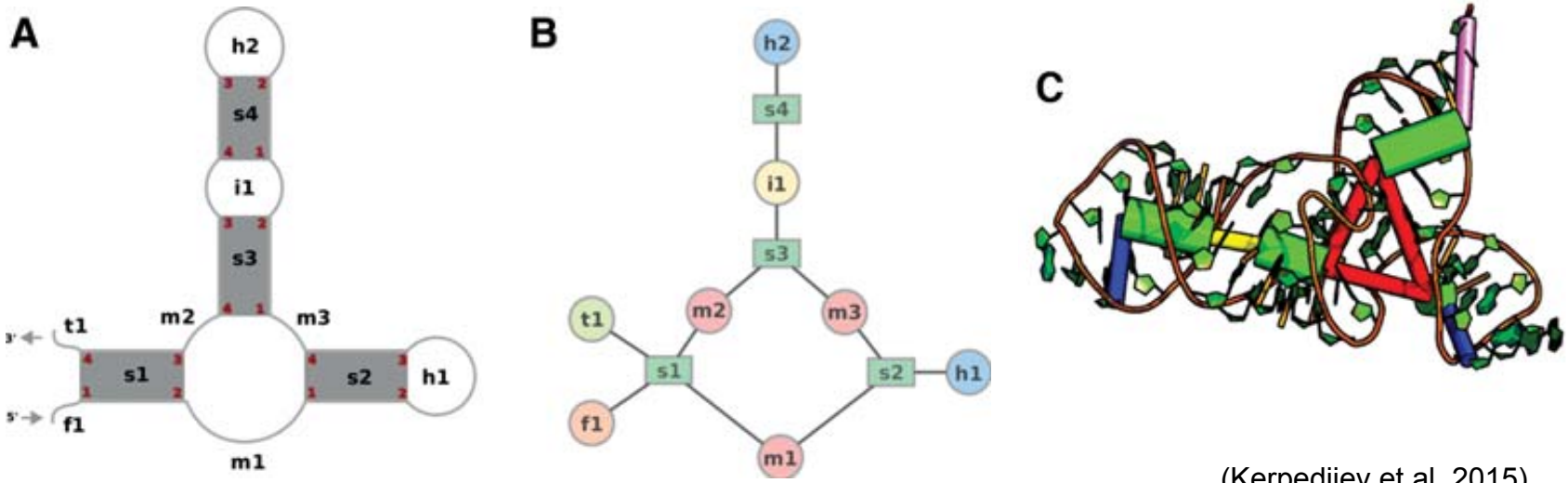
RNA 3D modeling

Molecular Dynamics simulations show that helical stems behave like (quasi-)rigid domains.

(Musiani et al, 2014)



A coarse-grained representation of stems and loops can be used for simulations with energy function defined for their interactions.



(Kerpedjiev et al, 2015)