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
A nn G nnnnnn C n	n U RNA 1	
G nn U nnnnn A n	n C RNA 2	
C nn A nnnnnn U n	n G RNA 3	
((((())))))) consens	sus "bracket view"

Different strategies:



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.

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:

NP_gullMD77/1-1565 NP_gsGD96/1-1565 NP_eqMiami63/1-1565 NP_Victoria75/1-1565 NP_swTN77/1-1565



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

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).

	A - U
Alignment:	
х	y positions
AUGUUGACGAUGGUCAUUUUGU	C <mark>A</mark> CAU seq1
AUGCUGACGAUGGUCAUUUUGU	C <mark>G</mark> CAU seq2
AUGCUGACGAUGGUCAUUUUGU	C <mark>G</mark> CAU seq3
AUGCUGACGAUGGUCAUUUUGU	C <mark>A</mark> CAU seq4
AUG <mark>U</mark> UGACGAUGGUCAUUUUGU	C <mark>G</mark> CAU seq5
AUGUUGACGAUGGUCAUUUUGU	C <mark>A</mark> CAU seq6
•••	
AUGCUGACGAUGGUCAUUUUGU	C <mark>G</mark> CAU seqN
f(b _x) f	(by) base frequencies

Covariation

x/y (?)

υC

A - U G . U C - G A - U

G - C

U - A

U - A

U

G

G บ

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



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)

1	2 positions
- <mark>A</mark>	U seq1
<mark>A</mark>	U seq2
<mark>A</mark>	u seq3
A	U seq4
<mark>G</mark>	c seq5
<mark>G</mark>	C seq6
<mark>G</mark>	c seq7
<mark>G</mark>	C seq8
f1()	f ₂ ()

Covariance models, RNA families and RNA descriptors

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

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.....>>>>>aaaaaaa.
```

(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 (IncRNA)
- 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 (IncRNA)
- 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, <u>rnacentral.org</u>) integrates data from ncRNA resources

5SrRNAdb	LncBase	PDBe	SILVA
CRW Site	LNCipedia	piRBase	snOPY
dictyBase	IncRNAdb	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

RNAcentral Expert Databases

Different search tasks are possible:





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:



hsa-mir-122-5p

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.



- 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

(<u>http://rth.dk/resources/rnasnp/</u>; 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.



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:

GU	. seq1
GU	. seq2
A	seq3
GU	seq4
A	seq5

Non-canonical base pairs in RNA

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



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	-0-	Parallel
3	Cis	Watson-Crick / Hoogsteen	•-	Parallel
4	Trans	Watson-Crick / Hoogsteen	οп	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	-0-	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

F. Michel and E. Westhof

length of aligned segments (see Methods for derivation of subgroups). secondary structure components are Figures 1(a), (b) and/or not discussed xt: (1) P11, a 4 to 7-bp interaction, bgroup IA1, between an internal or

terminal loop distal of stem P6 and the L7.1a terminal loop. (2) Pairings P5a, P5b and P5c (including the A-rich bulge on the 3' side of P5a), distal of stem P5 (Collins, 1988; "c-domain", see Michel & Cummings, 1985). This peripheral extension is typical of many subgroup IC and IB introns; note,

P4

P2.1 P3 OI TU.LSU ACUCUCUARAUAGE 7 JUUACCUUUGGAGGGAAAAGUE 19 JAGCUAGUE 31 JGGCA AGACCGUCAAAUUGCGGG 02 Tp.LSU ACUCUCUAAAUUGE 7 JUUACCUUUGGAGGGAAAAGUE 19 JAGCUAGUE 31 JGGCA AGACCGUCAAAUUGCGGG 03 Pp.LSU.3 AGACCGUCAAAUUGCGGG 04 NC.ND4L AUAGAUUAAUUL 1142 JUUAUCUUGAUCACAAAAGGL 15 JGCCUAGUCL B JGGCG AAACUCUCAAAUUGCGGG AUAGAUUAAUUT 1068 IUCAUCUUGAUCACCAAAGGT 14 JGUCUAGUCTA JGGCG AAACUCUCUCAAUUGCGGG 05 Pa.ND4L,1 AUAGAUUAGCUL 1182 JAAAUCUUGAUCUA -- AAGGL 14 JCCCUAGUCL 9 JGGCG AAACCUUCAAAUUGCGGG 06 Pa.ND4L.2 CGUAGGUCGCAL 1433 JUUAUUUGGCCUUU------- AAACCAUCAAAUUICGGG 07 Pa.ND1,1 ACAAGUUUUUUU21JCAAAGGAAGCCUUAG-CAGC147JUGCUAGU111JGGCGACAAAUUGCCGAGG 08 Pc.SSU 09 Pa.0X1.5 CCCUCCUGCCAAL 15 JACUAUUGUGGGAA-----AAACUAUCAAAUUCUGGG UAAUCCUAAAAAC 15 JAAUUAAUUGGGUA------AAACUAUCAAAUUCCGGG 10 Pa.ND3 13 Pa.ND4 UAUACAUUUAA-----CCUAAAGUGUA-----------AAAGAGCCAAAUACCGGG CANACAUAAAAAE 16 JACUUUGUUGUGUU------AAAGAGCCAAACUCCAAG 4 Pa.ATP6 P5 AAUACCUUAU------GUCCAUAGGGUU------AUAUAGUCAAACUACGGG 15 Pa.ND1,2 CAUGAUUAAUAAC35 JAAAAAGUUGAUCAGAAAAUUAUCAAAC920 JUGAUAUA-U AUAGAGGCAAACUCGAGG 6 Sc. 0X1,5a 17 Pp.LSU 1 UCGUCAUCGUAUL 44)GUAACGGCGUGACGUUL 19)GAAGCCL 52)GCGAGACCU AUGGGAUCUAACCGC AG CAGGGAUUGAUAL 78)UAAACGGGGUCCCAUUAAAAGUCAUCL 104)GAUGGCGG ACGCACCUG-GAUGCUGA 18 Pp LSU 2 P10 AGUUGCUCA-----ACUGGGCCAUUAGUNAK21JUUUGUCUC--CAUUUUGACAAAUUGCUGG 19 Pa.OX1,16 CGGA<u>CUUAA</u>------AAAU<u>UUGAGCDUUAAUUAAGAAAUU-AUUAAG</u>UG--CAA<u>ACUCUCAAAUUCAGGG</u> CGGA<u>CUUAAU------UGUAUUGAGCUUUAGUUGAGAAAUUOACUAAG</u>UG--AUG<u>UUUCAAAUUCAGG</u> CGGA<u>CUUAAU-------UGUAUUGAGCCUUGGUAUGGAAACUUAAG</u>UG--A UAA<u>CUUUCAAAU</u>CAGG CGGA<u>CUUAAU</u>-------UGG<u>AUUGAGCCUUGGUAUGGAAACUUAAG</u>UG--A UCA<u>CUUUCAAAU</u>GAGA 20 Cp.tLeu 21 Mp.tLeu P4 22 Vf.tLeu 23 Nt.tLeu CGGACUUGAU-----UGUAUUGAGCCUUGGUAUGGAAACCUGCUAAGUG--AGUACUUCCAAAUUCAGAG 24 Zm.tLeu AUUCGGUCA-----G CAACAAACUAUGCAGT B JUCUGAAUAAU----G CAACAAACUAUAUGCAGG 25 Pa.0X1,10 CUUCGUUCG-----AGAAUGGCCCUUAAAU---CAAA---AUUUAU--CG CAACAAACUAUAAGCGAG 26 Pa.0X1,9 27 Sc. 0X1.4 CUUD<u>GGUCAA------ACAGUGGCCCUUAUUAUUA</u>L13J<u>UAAUGAUAU----G</u>CA<u>UUUUUUCU</u>AAAU<u>GCUGG</u> 28 Sp. 0X1.2a CUUUGGUCAA-----GGAUGGCCCAAAAUA-----AAAGAUU-----G CAAUGUGCUAUAAGCUGG CUUAGGUCAA------AAUAUGGCCUUAUUAAL123UUAAUAU------UUUAUGUGCUGUAUGCUGG 29 Sc.COB.4 P6 GAAUGGUUA------ACAGGAGCCUUUGUAUUUT 1108 JAAAUACAAU---A AAUUUCAC 3AUGPECUGG AUAUGGCAAAAUAT231 JAUUAAUGCCGUUAAAAAGUGUAAACUUUUUAAUU--AU UACUUCGCUGUAUGUAGA ANT CONTACTS 30 Nc. ATP6 31 An. 0X1, 1 32 Pa. 0X1 3 33 Pa.0X1,2 34 Sc. 0X1,3 UUUAGGUAAAL 1177 JAAUAAAUUGCCGUGAGAGGGGAAUAUCUCUUU- AU UAUAACUAUGUGCGGG P6. 35 Pa.ND5,3 36 Pa.0X2.1 UCACAGUAG-----AGAUGUGCCUAUUCAAAAG(\$3\$ ICUAUGAAUA---AAAUUCCACUAUAUGCUGG AGAGGCUGGUUL 1478 JACGAUUCGGUCUCGUA-----AAAAAUUGCUAUAUGCUGG 37 Pa. ND1.4 UGAGGCUACUAAUA----UGAUUUGGUCUCAUU-----AAAGAUCACAAAUUGCUGG 38 Nc.ND1 UAUGAAUUUCGGE 1200 JCAAAUGAGGUUCAUAUUA-----AA UUUCUAACUAUGCUGG 39 Pa. 0X1.6 UUGAAGUUGAGU-----CAUGAUGGCUUCACCCUAUAGUGAUAUAGGGAL53AUU AAUUUCGCUAUAUGCUGG 40 Pa.OX1,11 CGUGGAUAA-----ACUUGUGUUCACGGGAA----AAAUCUUGCUAUAUGCUGG 41 Pa.OX1,13 GGAGGGUUAAUA---AAUCACUUAGCUCUCCCUUU------AAAGACCACUAUAUGCUGG 42 Pa.0X1,14 **P8** CCACGAUAUUAA--UUUAAUAAGUGUCGUGCUU-----A AAAUUCACUAIAACGGGG 43 Sc. 0X1,5b CCAUGAUAUUAL41 JAUAAUAGGUAUCAUGAAU-----A AAUGCCACCALAACGGUG 44 Sp.0X1,3 CCACGAUACGE 878 JUUAGUAAGUGUCGUGUUG------U CAUGUCACCA I AACGGUG 45 Pa.OX1,15 UACGCUUC------AAAUGGGGCCGUUAGAUAGAAAUAUCUAAUG---AA CAUCUGGCUGUAUGCUGG 46 Cm. psbA 2 47 An.OX1.3 CAGAGGUCAAUU-----UAAUAGGCCUCU------(Michel & Westhof, 1990) -----U AACGUUGCUAUAUGCUGG 48 Sp.0X1,2 AGCCGGUNAACAUA---GANAACUGCCGGCAU-----UA AAUUUGGCUAUAUACUGG 49 An. 0X1.2 50 Pa.0X1,7 GGAAGGUUUUUUU-----UAGUAGGGCUUUCCU--------UA AACUUCACUAUAUGCAGG Nc.ND5 2

cture constraints

tify important tertiary contacts sed in modeling.

using constraints derived from structures. Later, these models

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:



RNA 3D modeling



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

