Improve RNA secondary structure prediction with tertiary motifs

Gregor Entzian

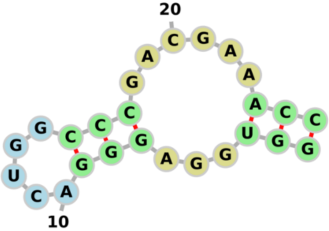
University of Vienna, Faculty of Chemistry, Department of Theoretical Chemistry

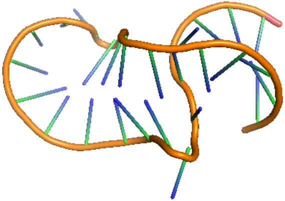
entzian@tbi.univie.ac.at

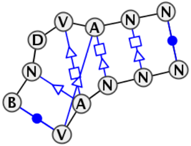
Supervisor: Univ.-Prof. Dipl.-Phys. Dr. Ivo L. Hofacker

February 16, 2018

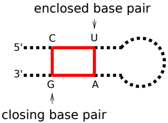
RNA 2D 3D



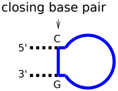




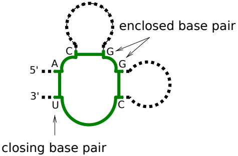
RNA secondary structure motifs



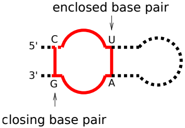
stacking pair



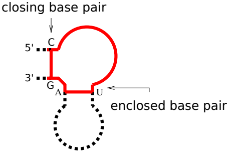
hairpin loop



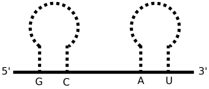
multi loop



interior loop

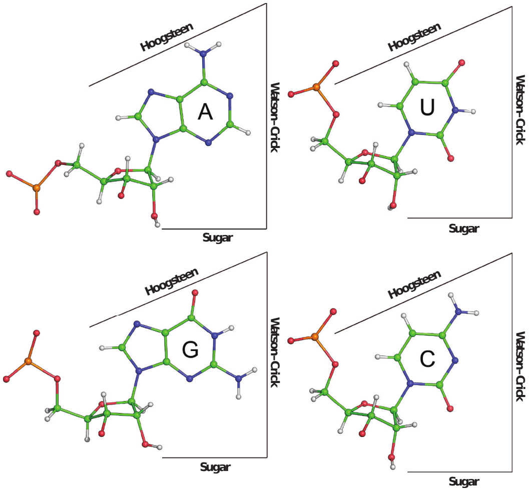


bulge



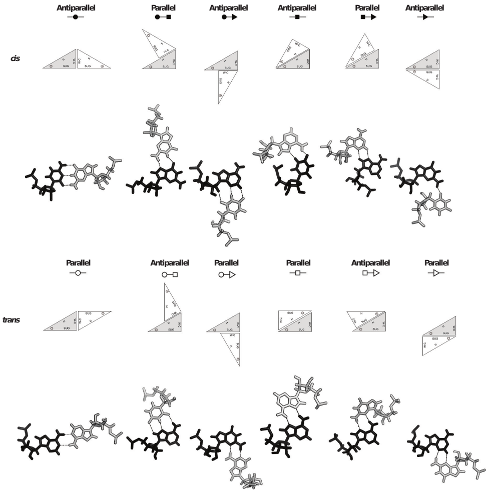
exterior loop

Non-WC base pairs 1/2



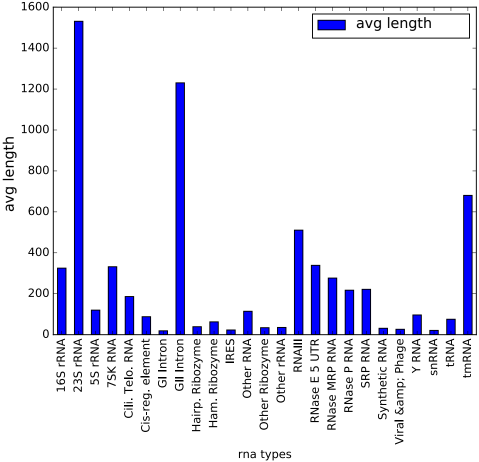
(Zhichao Miao and Eric Westhof 2017)

Non-WC base pairs 1/2

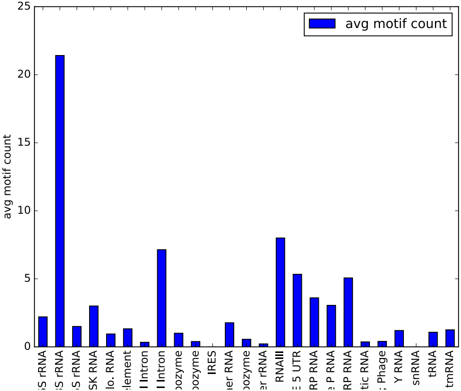


(Zhichao Miao and Eric Westhof 2017)

RNAstrand average lengths

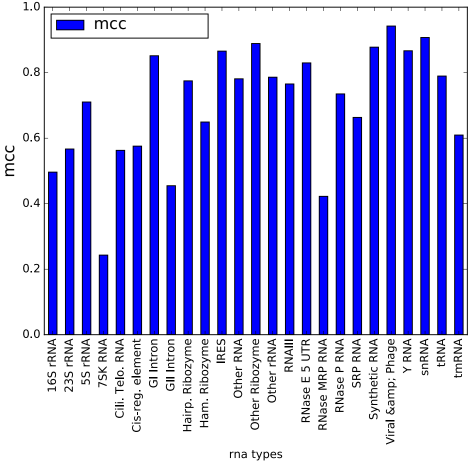


RNAstrand average motifs

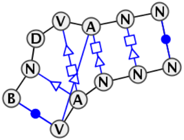


$1.3 motifs/100 nt$

RNAFold mcc


1.0
0.8
0.6
0.4
0.2
0.0
mcc
mcc
16S rRNA
23S rRNA
5S rRNA
7SK RNA
Cili. Telo. RNA
Cis-reg. element
GI Intron
GII Intron
Hairp. Ribozyme
Ham. Ribozyme
IRES
Other RNA
Other Ribozyme
Other rRNA
RNAIII
RNase E 5 UTR
RNase MRP RNA
RNase P RNA
SRP RNA
Synthetic RNA
Viral &amp; Phage
Y RNA
snRNA
tRNA
tmRNA
rna types

Tertiary motifs from motif atlas

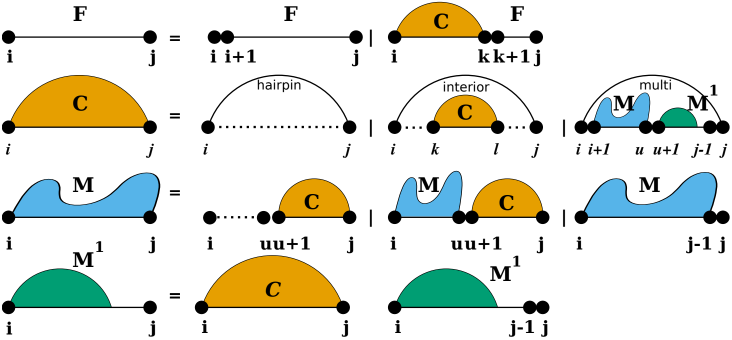


- Automatically detects tertiary motifs in structures from a nonredundant set of PDB files.
- Recurrent motifs discovered with the tool Jar3d. (3d structure alignment, str. that share the same geometry)
- Jar3d is based on SCFG and Markov Random Fields

Useful motifs for secondary structures

- Motifs with not too many wildcars
- Occur often in non-homologous sequences

Folding grammar



How to get energy values?

- Usually UV melting
- No exeperiments available for the detected motifs
- Idea constraint generation (like (M.Andronescu 2007) parameter estimation)

Constraint generation

- Idea:
  Energy(structure with 3d motifs) $<$ Energy(structure without motifs)
- Solver: CPLEX for Quadratic Programming

δ: vector of slack variables (infeasible label constraints)
$c(x, y_x)$: vector with numbers of motifs within structure y_x
Θ: thermodynamic parameter for each motif

Optimization:

$$\textit{minimize } \delta^2$$
$$\textit{subject to}$$
$$c(x, y_x) \cdot \Theta + E(x_{Rest}, y_{xRest}) - \delta < c(x, y') \cdot \Theta + E(x_{Rest}, y'_{xRest})$$
$$\delta \geq 0$$

done:

- Regex-like detection of motifs in iupac notation
- Extend the folding grammar for Motif detection

in progress:

- Conversion from motif sequence alignments into regex-like iupac expressions
- Prepare test and training data: extract motifs, convert pdb files into fasta files with 2d structures, annotate the motifs within the fasta files

todo:

- Find interesting motifs (count occurances in non-homologous rRNA sequences)
- Implement the CG Algorithm
- Compute the prediction accuracy

Thank you!

- Ivo Hofacker
- Craig Zirbel
- Bernhard Thiel
- Andrea Tanzer
- Ronny Lorenz