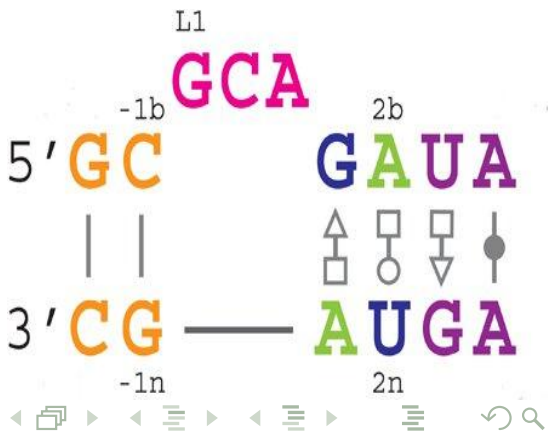
Challenges on the way to answer the question: Can RNA 3D motifs enhance structure predictions?

Corinna Theis

Center for non-coding RNA in Technology and Health
Department of Basic Animal and Veterinary Sciences, IBHV
Faculty of Life Sciences
University of Copenhagen

27^{th} TBI Winterseminar

February 17, 2012



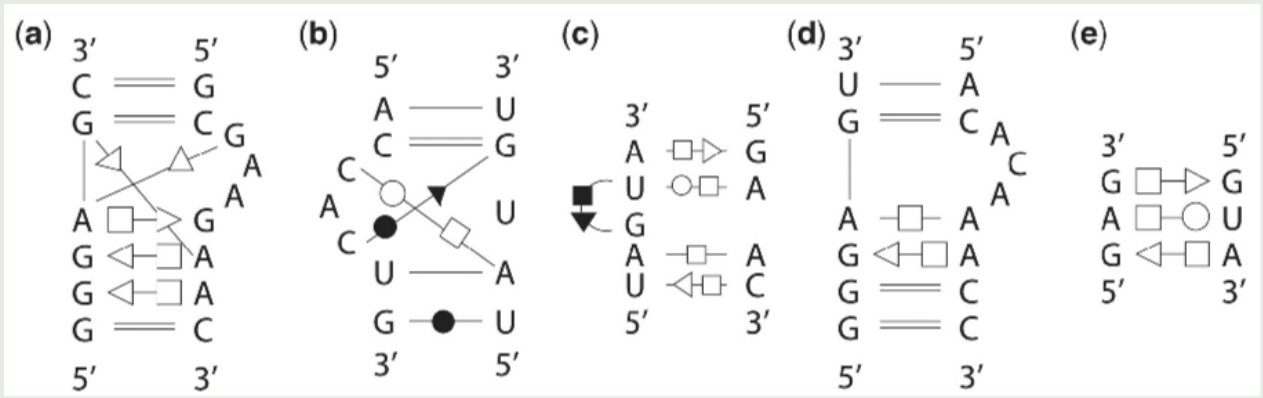
Outline

- RNA 3D motifs
- 4 approaches
- Challenges
- Perspective

RNA 3D motifs

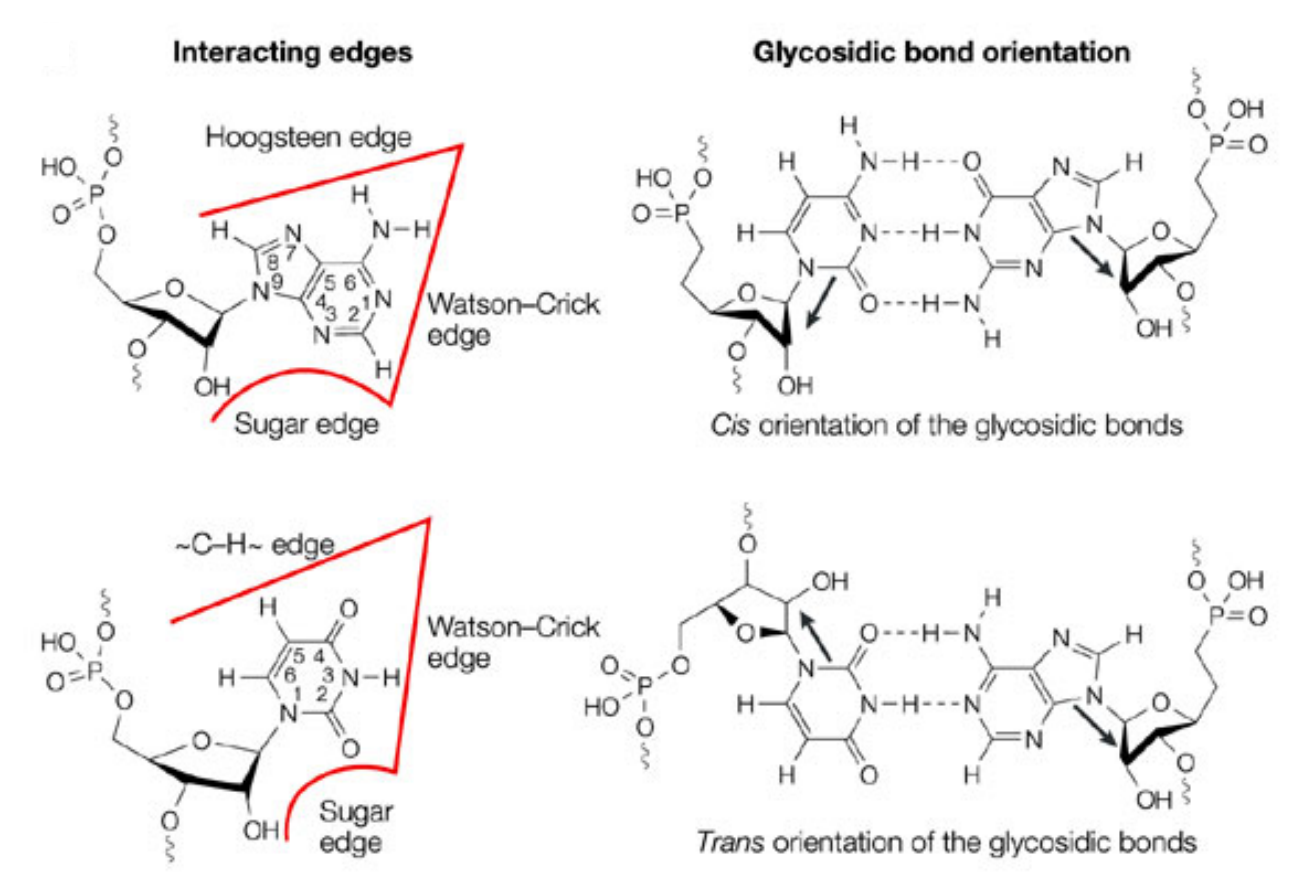
- small ($<$ 20 nt)
- any phylogenetic branch
- architectural organizers, sites of ligand binding

- canonical/non-canonical base pairs
- described by using a (non-planar) graph



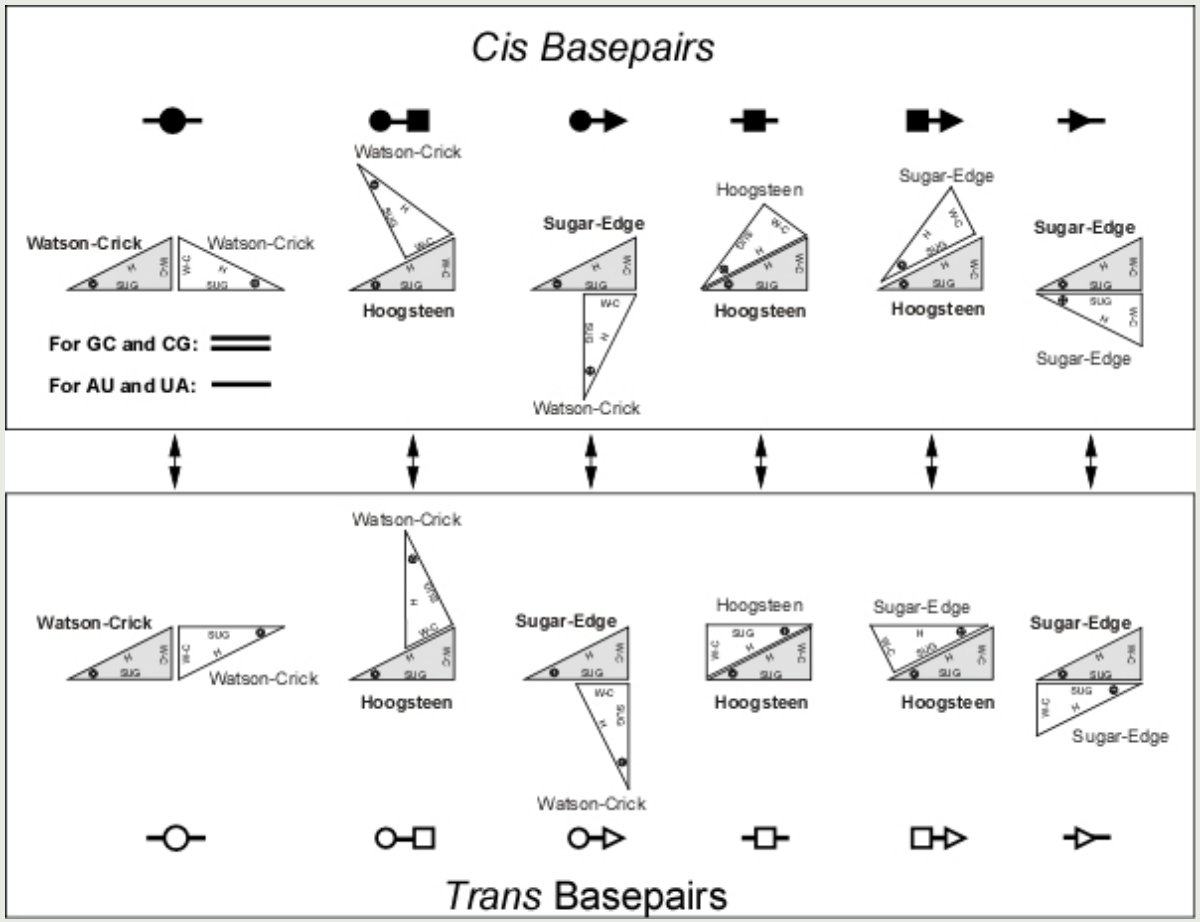
(a) kink-turn (b) C-loop (c) sarcin-ricin (d) reverse kink-turn (e) E-loop (Zhong et al.)

RNA 3D motifs



Base pair geometry (Schroeder et al. 2004)

RNA 3D motifs



Geometric families of base pairs (Leontis et al. 2002)

Motif DB

- $\sim$ 25 different known motifs
- no comprehensive database available

Motif	Submotif	No.	RNA type
kink-turn		17/33	16S,23S rRNA; C/D,H/ACA snoRNA; SAM I,lysine riboswitch; mRNA
G-bulge		5/5	16S,23S rRNA; riboswitch; group I intron; mRNA
C-loop		5/8	16S,23S rRNA; mRNA
E-loop		2/2	5S rRNA; ribozyme
tandem GA		1/1	
sarcin-ricin		3/3	5S,16S,23S rRNA
tetraloop	gnra,ynmg,cuug	3/3	16S rRNA; riboswitch
tetraloop-receptor		1/1	
A-minor	type 0,I,II,III	4/4	
pseudoknots			
reverse kink-turn		1/1	group I intron
hook-turn		1/1	
pentaloop		1/1	
pentaloop-receptor		1/1	
hexaloop		1/1	
hexaloop-receptor		1/1	
T-loop		1/1	
S-turn		1/1	ribozyme
quadruplexes		1/1	
ribose zipper		1/1	
π-turn		1/1	
Ω-turn		1/1	
α-loop		1/1	

4 Approaches

1. 3D Motif DB
   - overview
   - machine-readable format
2. RMdetect (Cruz et al.) on Rfam data
   - proof of concept
   - are true motifs distinguishable from False Positives?
3. Post-filtering of RNAz results
   - can we enhance ncRNA predictions?
4. Extension of 3D motif search tool
   - include more/new motifs

Challenges

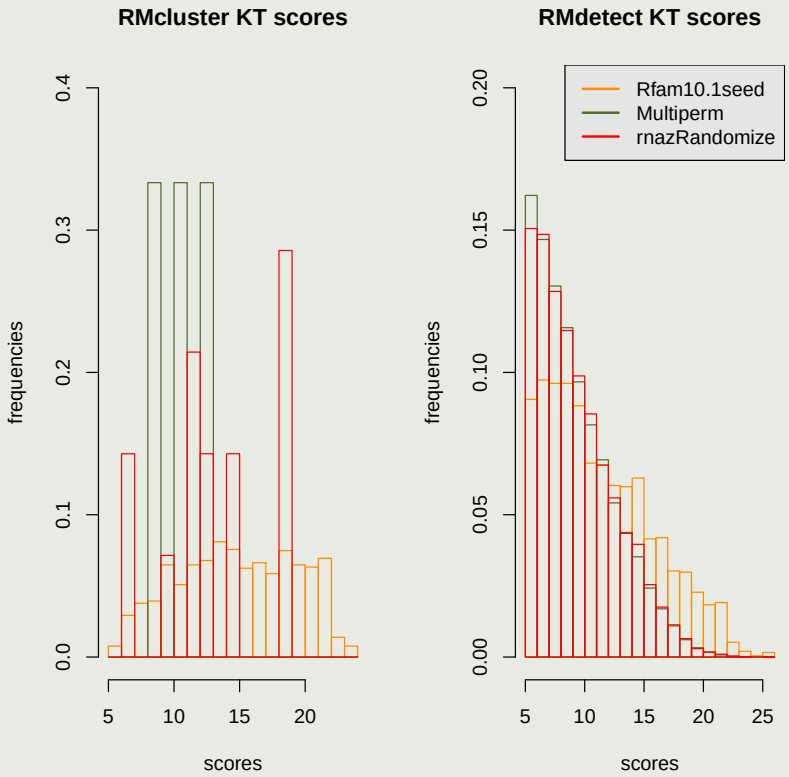
1. 3D Motif DB

- read a lot of papers
- compare positions and sequences against sequence DB
- store in machine-readable format (sql DB)

Challenges

2. RMdetect (Cruz et al.) on Rfam data

- run on 100x shuffled families
  $\longrightarrow$ not enough data
- run on 1000x shuffled families
  $\longrightarrow$ wait 4 weeks for the results
- clean Rfam families
  - no. of sequences per family
  - pairwise sequence identity
  - average pairwise sequence identity per family

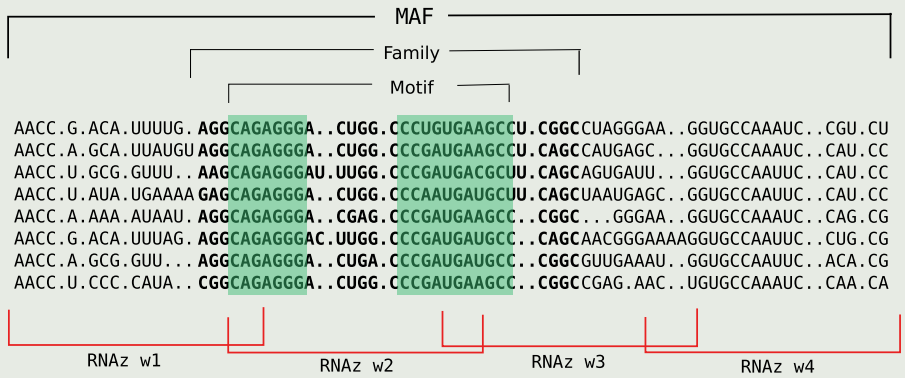


Challenges

3. Post-filtering of RNAz results

- RMdetect on Rfam families
- get MAF blocks for regions with TP motifs
- run RNAz
- combine RMcluster/RNAz scores
- Problem: 238 families
  - # KT clusters: 44
  - # GB clusters: 12
  - # CL clusters: 153
  - # TGA clusters: 117

  but only 3 KTs/ 1 GB in 3 families (TP)



Idea

3. Post-filtering of RNAz results with pseudo TP

- idea: use FR3D results as pseudo TP
- extract motifs $\rightarrow$ 1694 classes, 15290 instances
- not straight-forward: map FR3D motifs to Rfam families
- so far: MAF blocks only for H.sapiens

Challenges

4. Extension of 3D motif search tool

- create new motif models with RMbuild
  - how to rank FR3D classes?
  - need for alignment → 48 Rfam linked to PDB structure, 266 common PDB ids
    - how many contain (high ranked) motifs?
  - map motifs to alignments
  - create Bayesian Networks → manually change base pairs

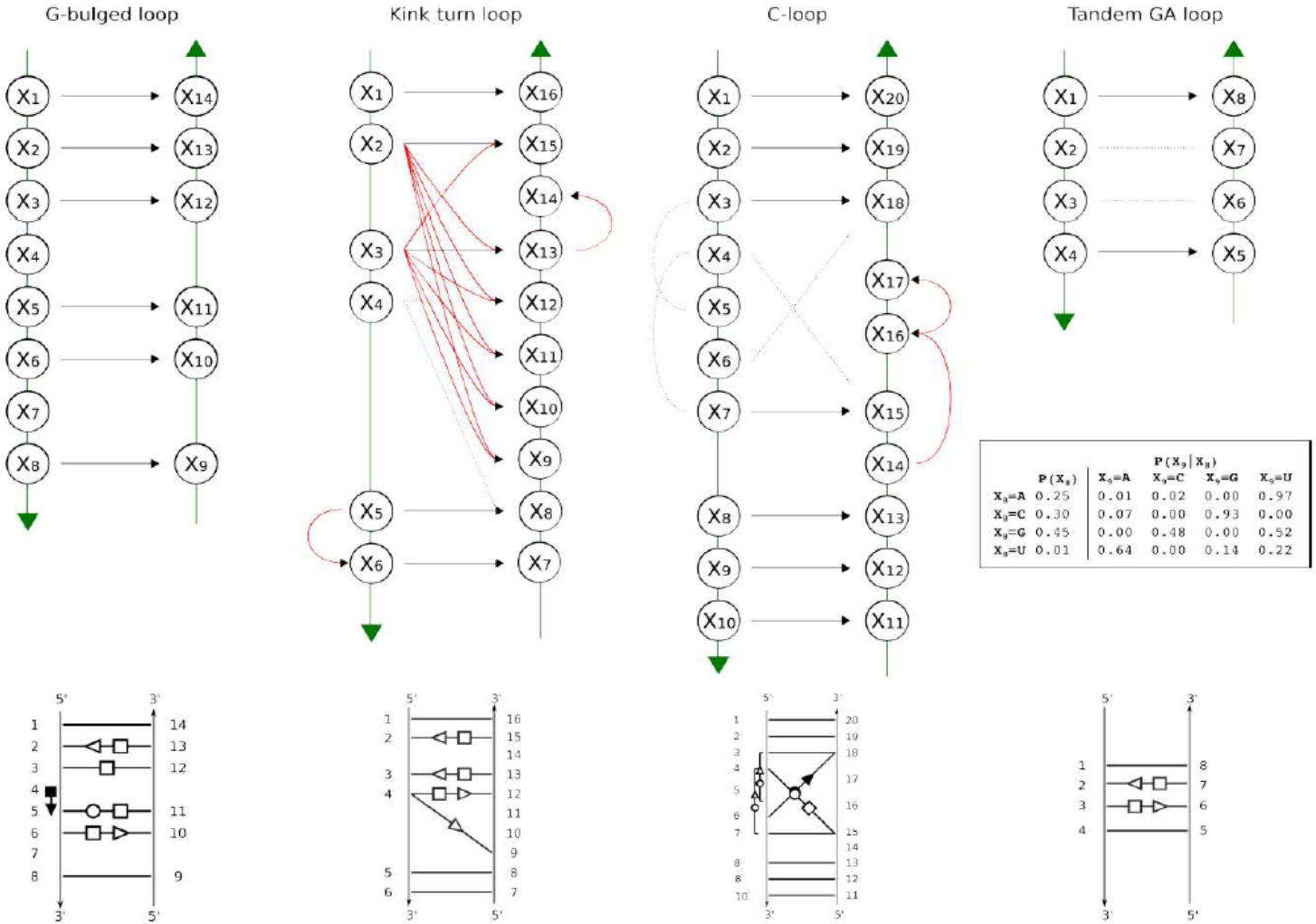
UGGAAG CAGGAAG
(((< .(. > ..)))))
0 13 U-G (c,W,W)
1 12 G-A (t,S,H)
2 11 G-A (t,S,H)
3 7 A-C (t,S,S)
3 10 A-G (t,H,S)
5 7 G-C (c,W,W)

Instances: 59

CGAAGUGGCUGGAAAGCUACACCUCAGAAG CUGUAG
((.....(.)...))
0 36 C-G (c,W,W)
1 35 G-A (t,S,H)
29 31 G-C (c,W,W)

5

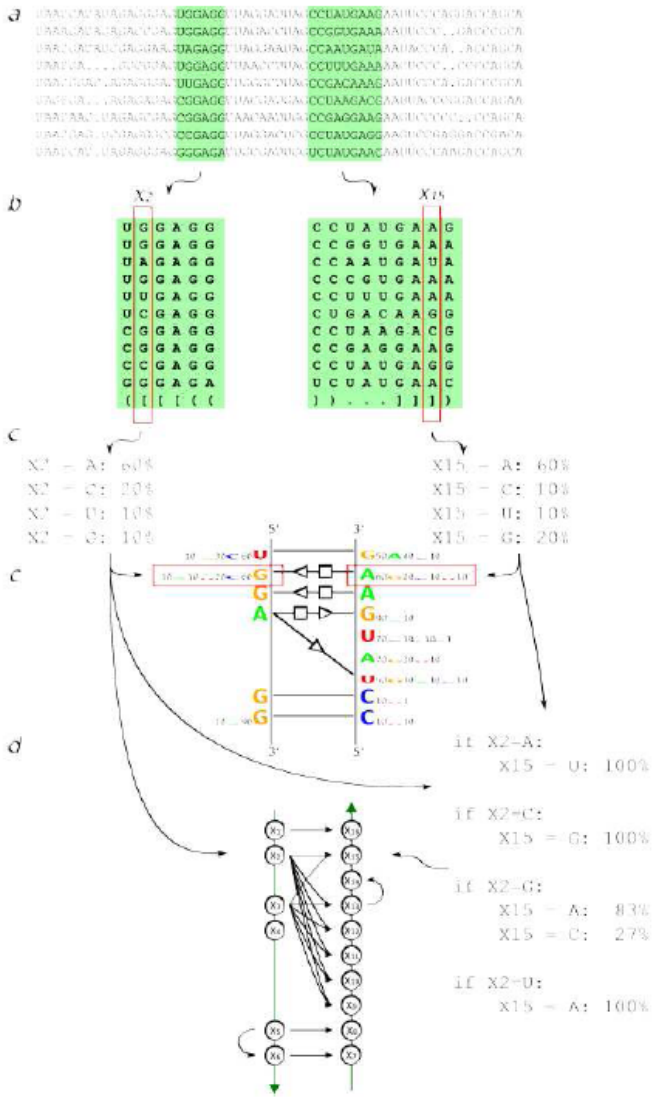
Bayesian Network



	$P(X_8)$	$P(X_9 \mid X_8)$ X_9=A	X_9=C	X_9=G	X_9=U
X_8=A	0.25	0.01	0.02	0.00	0.97
X_8=C	0.30	0.07	0.00	0.93	0.00
X_8=G	0.45	0.00	0.48	0.00	0.52
X_8=U	0.01	0.64	0.00	0.14	0.22

[Cruz et al. 2011]

Bayesian Network



[Cruz et al. 2011]

Perspectives

Question is still...

Can 3D information contribute to improve ncRNA predictions in genomic sequences?

- work on 4 approaches
- after that is done
  - assign 3D information to structures predicted in previous projects (CMfinder results)
    - limit the number of false positives
    - assign functions to the true positives
  - develop methods for predicting and annotating 3D motifs (structures) in ncRNAs
    - on genome wide scale
    - find novel motifs/ncRNAs
  - refine 2D structures
    - bring the structure closer to the biological structure

Acknowledgements

- Christian Höner zu Siederdissen,
  TBI, University of Vienna
- Jan Gorodkin,
  Center for noncoding RNA in Technology and Health (RTH),
  University of Copenhagen
- all the people from RTH

Thanks for your attention!

References

- Leontis, NB. et al., The non-Watson-Crick base pairs and their associated isostericity matrices. Nucleic Acids Res 2002, 30:3497-3531.
- Schroeder, R. et al., Strategies for RNA folding and assembly. Mol. Cell Biol. 2004, 5:908-919.
- Cruz, JA. et al., Sequence-based identification of 3D structural modules in RNA with RMDetect. Nature Methods 2011, 8(6):513-521.
- Sarver, M. et al., FR3D: finding local and composite recurrent structural motifs in RNA 3D structures. J. Math. Biol. 2008, 56:215-252.
- Washietl, S. et al., Fast and reliable prediction of noncoding RNAs. Proc. Natl. Acad. Sci. U.S.A. 2005, 102:2454-2459.
- Zhong, C. et al., RNAMotifScan: automatic identification of RNA structural motifs using secondary structural alignment. Nuc. Acid Res. 2010, 38(18):e176.
- Klein, DJ. et al., The kink-turn: a new RNA secondary structure motif. EMBO 2001, 20(15):4214-4221.