Computational studies of co-transcriptional folding — what we learned so far

Irmtraud M. Meyer

MDC–BIMSB & Freie Universität, Berlin, Germany

Benasque, 24. July 2018

MDC MAX DELBRÜCK CENTER FOR MOLECULAR MEDICINE IN THE HELMHOLTZ ASSOCIATION

Freie Universität Berlin

Co-transcriptional folding: part I

Definitions
Experimental findings
Computational results

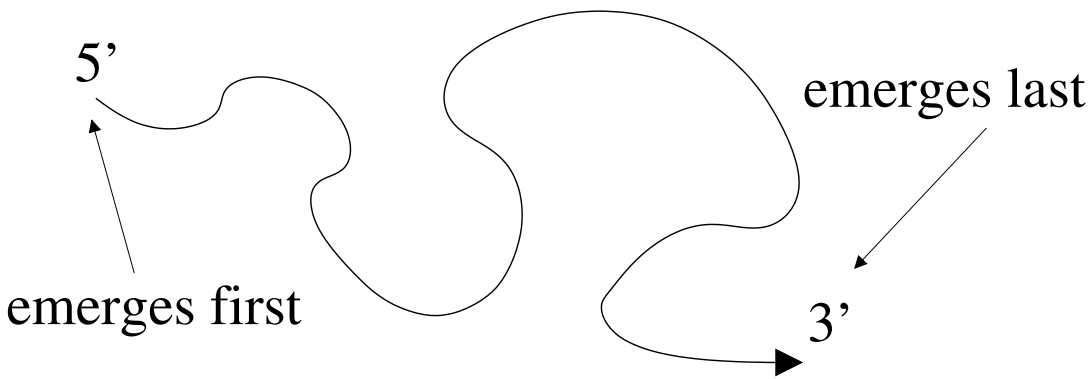
Co-transcriptional folding

Motivation:

RNA genes are *transcribed*, whereas true intergenic sections are not.

Main idea:

RNA genes emerge in a directed process first 5' then 3': do RNA genes fold **co-transcriptionally** i.e. while they are being transcribed ? If yes, what are the effects ?



Experimental evidence for co-transcriptional folding

- RNA molecules fold as they are transcribed

  [Boyle et al., J. of Mol. Biol. 1980, Kramer and Mills, Nucl. Acid Res. 1981]

- transient structures exist and can have a distinct biological function

  [Kramer and Mills, Nucl. Acid Res. 1981, Repsilber et al., RNA 1999, Ro-Choi and Choi, Mol. and Cells 2003]

- wrong speed of transcription can lead to inactive transcripts

  [Lewicki et al., J. of Mol. Biol. 1993, Chao et al., Nucl. Acid Res. 1995]

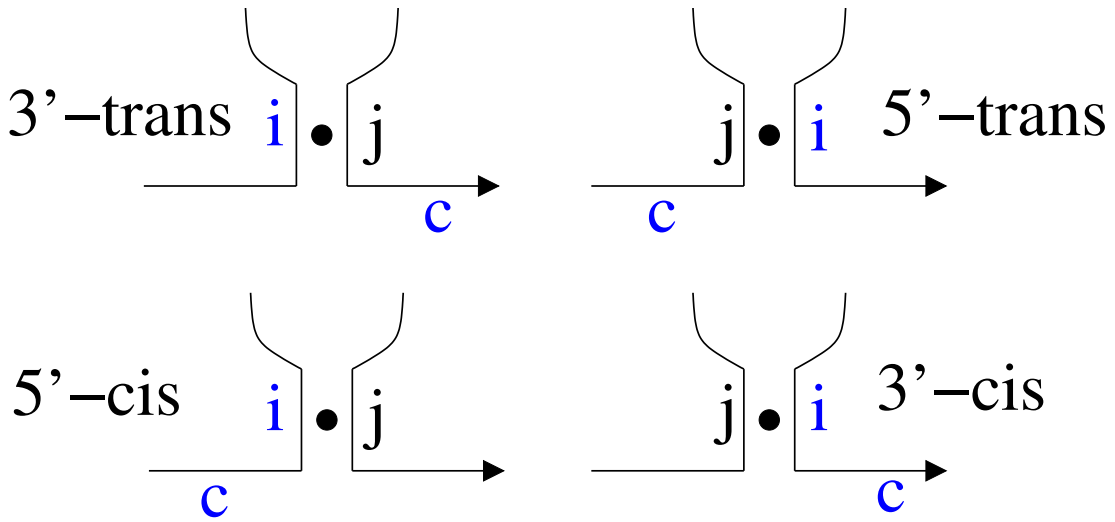
Key questions:

Do RNA genes encode information on their own correct co-transcriptional folding ? If yes, how is this achieved ?

Algorithm for detecting co-transcriptional folding

Idea:

for each helix of the known structure, measure asymmetry between competing helices 5' and 3' of that helix



- `i-j` is a base-pair of the known structure, `i-c` is a base-pair of a competing helix
- a competing helix has to have a minimum length of 9 consecutive base-pairs

Statistics for detecting co-transcriptional folding

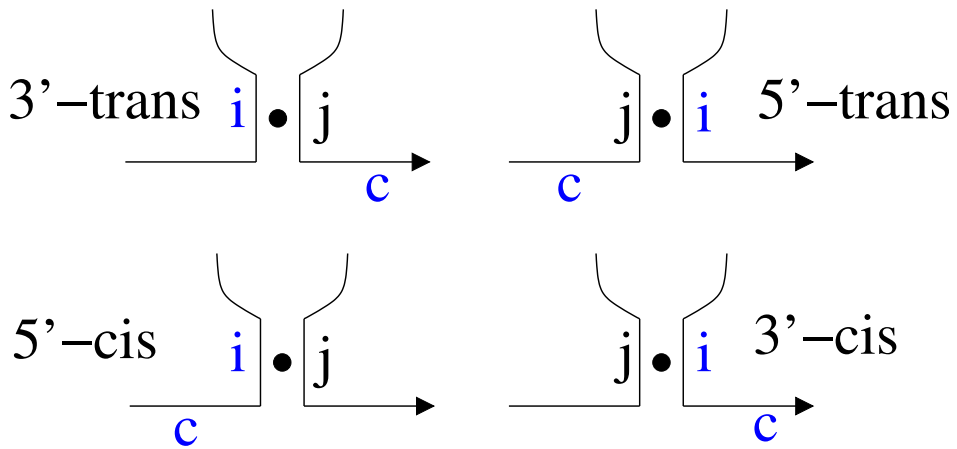
For each RNA sequence, calculate two scalar values:

$$\begin{aligned} \textit{Trans} &:= \sum \textit{3'-trans} - \sum \textit{5'-trans} \\ \textit{Cis} &:= \sum \textit{5'-cis} - \sum \textit{3'-cis} \end{aligned}$$

where *3'-trans*, *5'-trans*, *5'-cis* and *3'-cis* are weights which are proportional to $1/(\text{distance: real} - \text{competing helix})$.

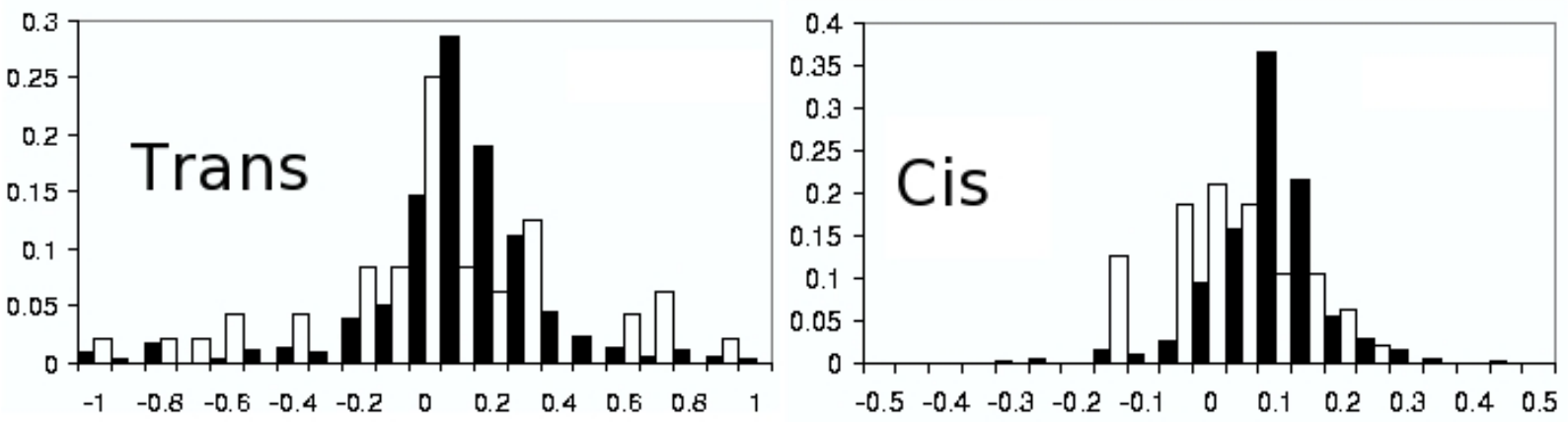
Interpretation:

$$\begin{aligned} \textit{Trans} &> 0 \quad \text{if competing helices are suppressed} \\ \textit{Cis} &> 0 \quad \text{if transient helices are encouraged} \end{aligned}$$



Results:

- data set A: 361 original transcripts (16S rRNAs, 23S rRNAs)
- data set B: 48 sub-sequences of original transcripts (group I and II introns, several 23S rRNAs)



	Trans		*Cis*	
A average (p-val.)	0.079 ± 0.026	(0.0012)	0.070 ± 0.004	(0.0001)
B average (p-val.)	0.041 ± 0.082	(0.3093)	-0.003 ± 0.015	(0.5733)

$\Rightarrow$ co-transcriptional folding is encoded in structured transcripts

[Meyer and Miklós, BMC Bioinformatics (2004) 5:10]

Summary:

Due to co-transcriptional folding, structured RNA genes

(1) suppress competing transient helices that could jeopardize the formation of the final RNA structure

(2) encourage transient structures that would facilitate the formation of the final RNA structure

Key insights:

$\Rightarrow$ RNA genes encode information on their own co-transcriptional folding pathway
$\Rightarrow$ an RNA molecule *in vivo* explores only a reduced folding-space

Goals:

use these insights in order to

- improve RNA structure prediction
- improve RNA gene prediction

Co-transcriptional folding: part II

Can we incorporate co-transcriptional folding into thermodynamic RNA structure prediction?

Key motivation:

1. Can we **conceptually** improve state-of-the-art thermodynamic RNA structure prediction methods such as MFOLD and RNA-FOLD ?

   [Zuker (2003) NAR 31:13, Zuker and Stiegler (1981) NAR 9:133-148]

2. The performance accuracy of thermodynamic methods drops with increased sequence length. Is there a conceptual way to fix this?

   Discrepancies between the conserved RNA secondary structures and predicted MFE structures "cannot simply be put down to errors in the free energy parameters used in the model".

   [Morgan and Higgs (1996) J of Chem Physics 105(16):7152-7157]

3. RNA SEQUENCES *in vivo* FOLD CO-TRANSCRIPTIONALLY. Can we somehow capture this in a thermodynamic method?

   [Boyle1980, Kramer1981, Brehm1983, Lewicki1993, Chao1995, Pan1999, HeilmanMiller2003, HeilmanMiller2003b, Mahen2005, Adilakshmi2009, Mahen2010, Woodson2010]

Existing methods for predicting kinetic folding pathways:

- take a single RNA sequence as input
- make a range of simplifying assumptions
  - transcription speed is constant
  - no interactions with other molecules (ligands, proteins, other transcripts)
  - no modeling of detailed cellular environment (concentrations of different ions, temperature etc)
- further limitations
  - can typically only handle short sequences (typically $\leq$ 1000 bp)

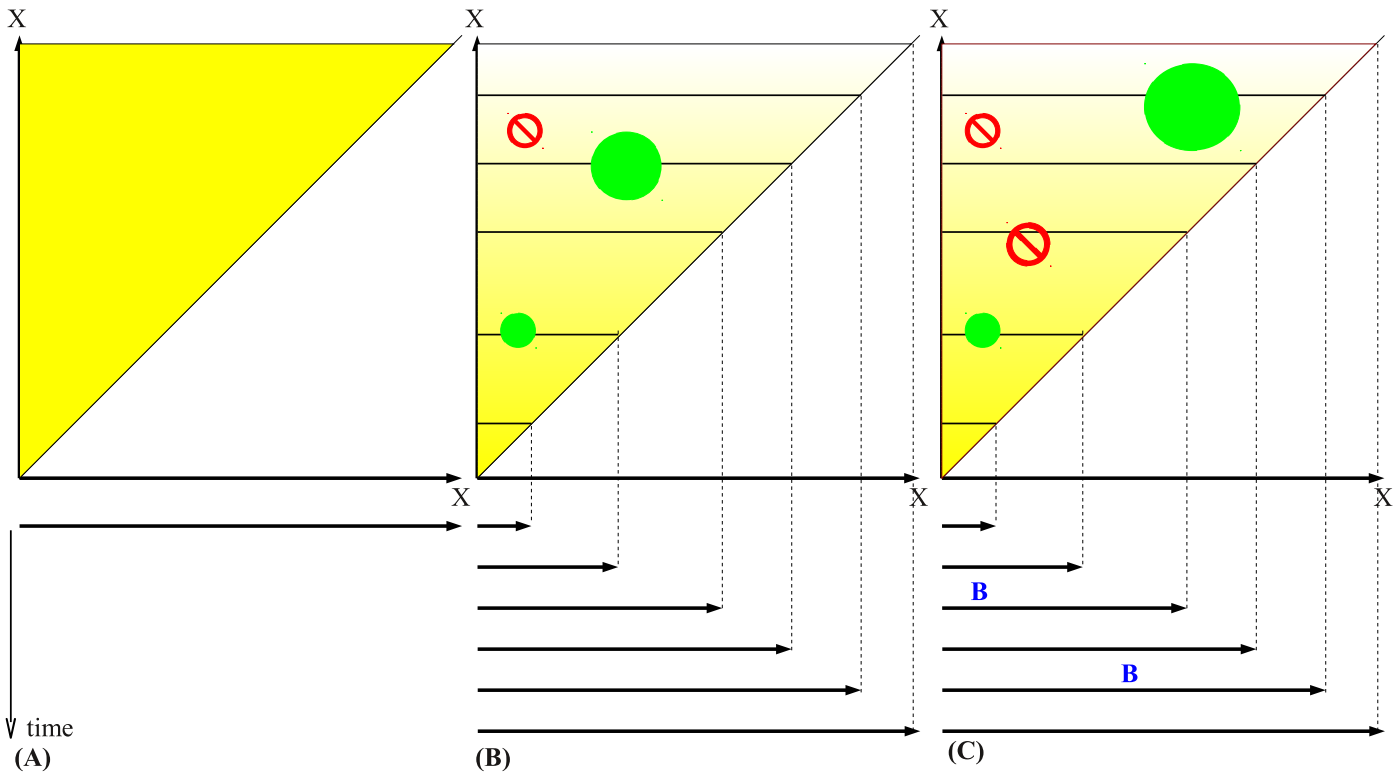
Examples:

- RNAkinetics by Mironov *et al.*
- Kinfold by Flamm *et al.*
- Kinefold by Isambert *et al.*
- Kinwalker by Geis *et al.*

Key challenges:

- Most RNA structure prediction algorithms have no concept of a folding pathway and simply **ignore the process of structure formation**. They assume a fully synthesized transcript.
- *In vivo*, however, a transcript emerging and folding co-transcriptionally **needs to find a way of actually reaching the functional RNA structure**, i.e. the structure formation process is key.

Key challenges:



- co-transcriptional folding reweights the space of all potential RNA structures and makes some potential structures inaccessible and others easier to form

Key features of RNA structure prediction method CoFold:

Design criteria:

- modify RNA-Fold in order to capture some overall effects of co-transcriptional folding
- introduce only modifications with a clear biological interpretation that ...
- depend on as few free parameters as possible.

Key features:

- introduce a scaling-function that **judges the reachability of potential base-pairing partners during kinetic folding**
- introduce scaling-function

$$\gamma(d) := \alpha \cdot (e^{-\frac{d}{\tau}} - 1) + 1$$

which depends on 2 free parameters α and τ and where d is the distance between the two potential pairing partners along the sequence (in nt)

CoFold: data sets

	test set	training set	
	long data set	combined data set	
clade	$>$ 1000 nt	all	$\leq$ 1000 nt
Bacteria	15	69	(54)
Eukaryotes	15	112	(97)
Virus	0	20	(20)
Archea	17	33	(16)
Chloroplast	14	14	(0)
sum	61	248	(187)
av. seq. length	**2397**	**776**	(247)
max. seq. length	**3578**	**3578**	(628)

Selection criteria:

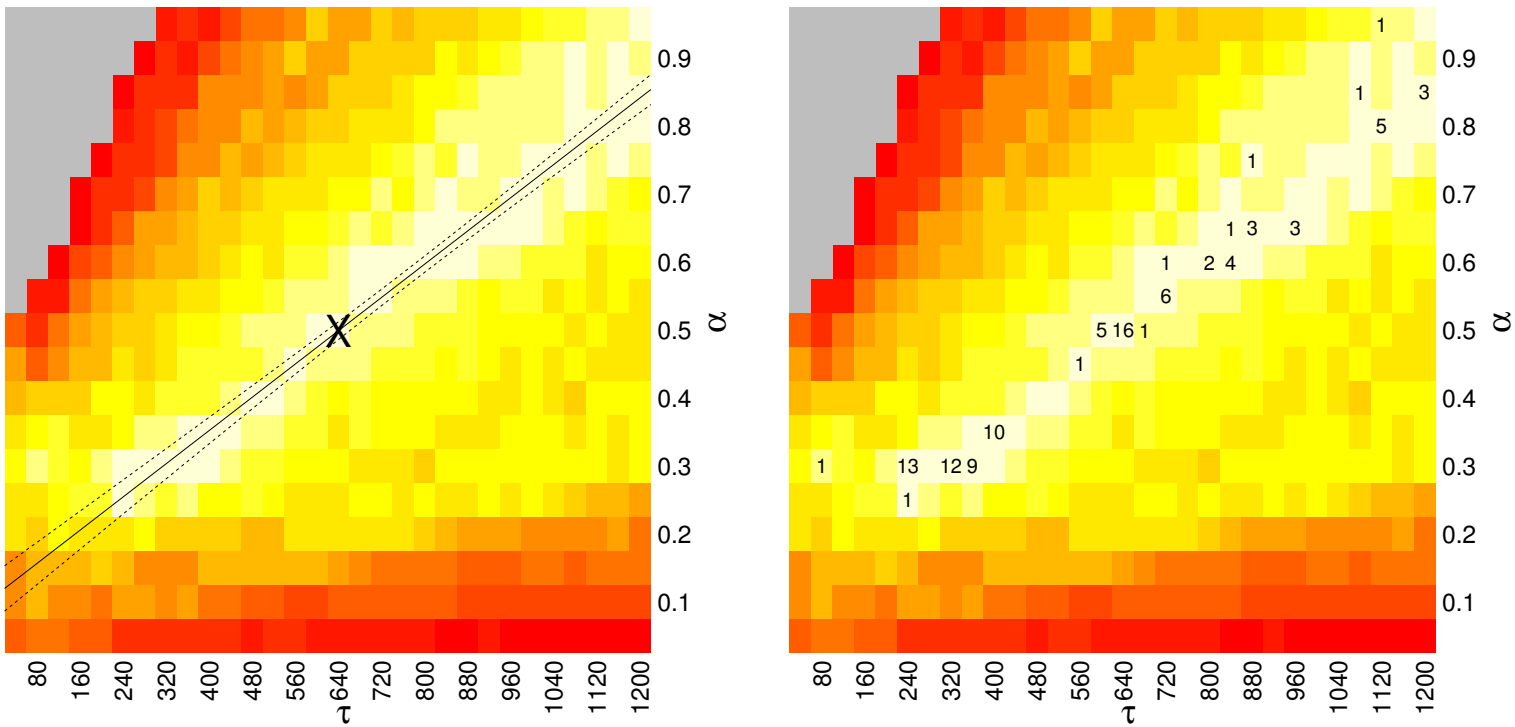
- only biological sequences
- ref. structures supported by strong evol. evidence
- long data set: length $>$ 1000 nt and pairw. % seq. id $\leq$ 85%
- long data set $\Rightarrow$ non-redundant 16S and 23S rRNAs

CRW data b. [Canone (2002) BMC Bioinf 3:2], Rfam data b. [Gardner (2011) NAR]

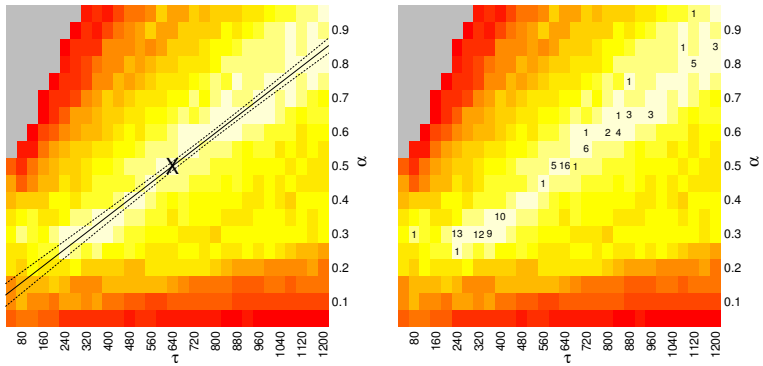
CoFold: parameter training

Strategy:

- task: two parameters to train
- objective: optimize average MCC prediction accuracy
- method: twenty trials of five-fold cross-validation
- use combined data set: non-redundant and diverse data set of 248 sequences (av. length 776 nt, min 110 nt, max 3578 nt)



CoFold: parameter training



Outcome:

- two parameters strongly correlated: $\alpha = a \cdot \tau + b$

  where $a = 6.1 \cdot 10^{-4} \pm 2 \cdot 10^{-5}$ (slope) and $b = 0.105 \pm 0.016$ (intercept) ($R^2 = 98.4\%$)
- $\Rightarrow$ CoFold effectively depends only on **one** parameter
- optimal parameter combinations all fall within or near the 95% confidence interval around the linear fit
- $\Rightarrow$ parameter training robust
- $\Rightarrow$ use $\alpha = 0.50$ and $\tau = 640$ in the following

Introducing CoFold-A and RNAfold-A

Benchmark performance using the following four methods:

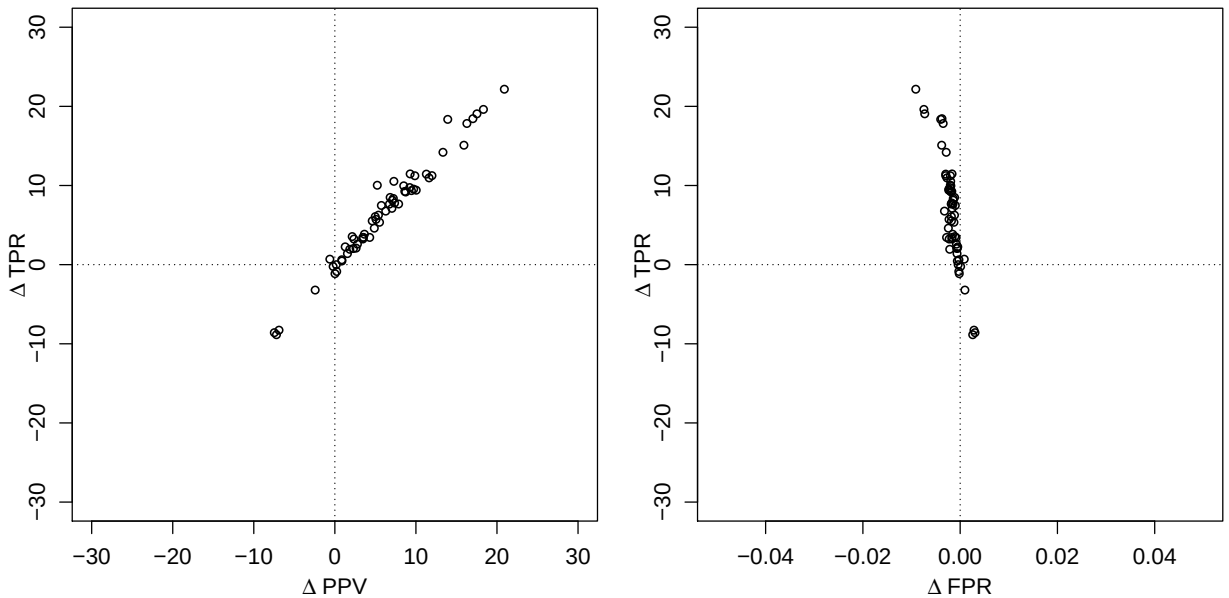
- CoFold and RNAfold: use default energy model (Turner 1999)

  [Mathews et al. (1999) J Mol Biol 288: 5]
- CoFold-A and RNAfold-A: use Andronescu energy model (2007) comprising **363 free parameters** that were trained using sophisticated machine learning techniques.

  [Andronescu et al. (2007) Bioinf 23:13]
- evaluate performance accuracy on long data set: non-redundant, evol. diverse data set of 61 sequences (av. length 2397 nt, min 1245 nt, max 3578 nt)

CoFold: performance accuracy

Absolute (!) changes in prediction accuracy for base-pairs for structures predicted by CoFold for individual sequences w.r.t. RNAfold.



- true positive rate: $TPR = 100 \cdot TP/(TP + FN)$
- positive predictive value: $PPV = 100 \cdot TP/(TP + FP)$
- false positive rate: $FPR = 100 \cdot FP/(FP + TN)$

CoFold: performance accuracy in numbers

Prediction accuracy for base pairs

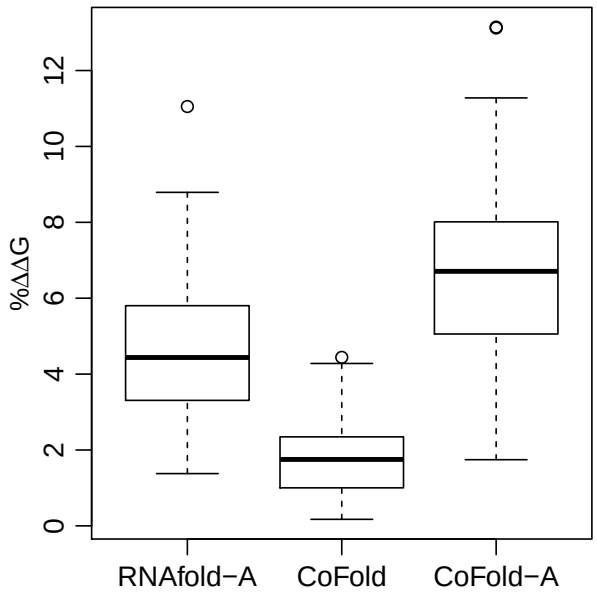
	TPR (%)	FPR (%)	PPV (%)	**MCC** (%)
RNAfold	46.30	0.0176	39.74	**42.81**
RNAfold-A	52.02	0.0160	44.76	**48.17**
CoFold	52.83	0.0159	45.79	**49.10**
CoFold-A	57.80	0.0145	50.06	**53.70**

Bottom line:

- MCC: RNAfold $\rightarrow$ CoFold **+6%** (TPR **+7%**, PPV **+6%**)
- MCC: CoFold $\rightarrow$ CoFold-A +4%
- FPR low for all four methods

CoFold: influence on structures' free energies

Relative free energy differences of the predicted structures w.r.t. the MFE structures predicted by RNAfold.

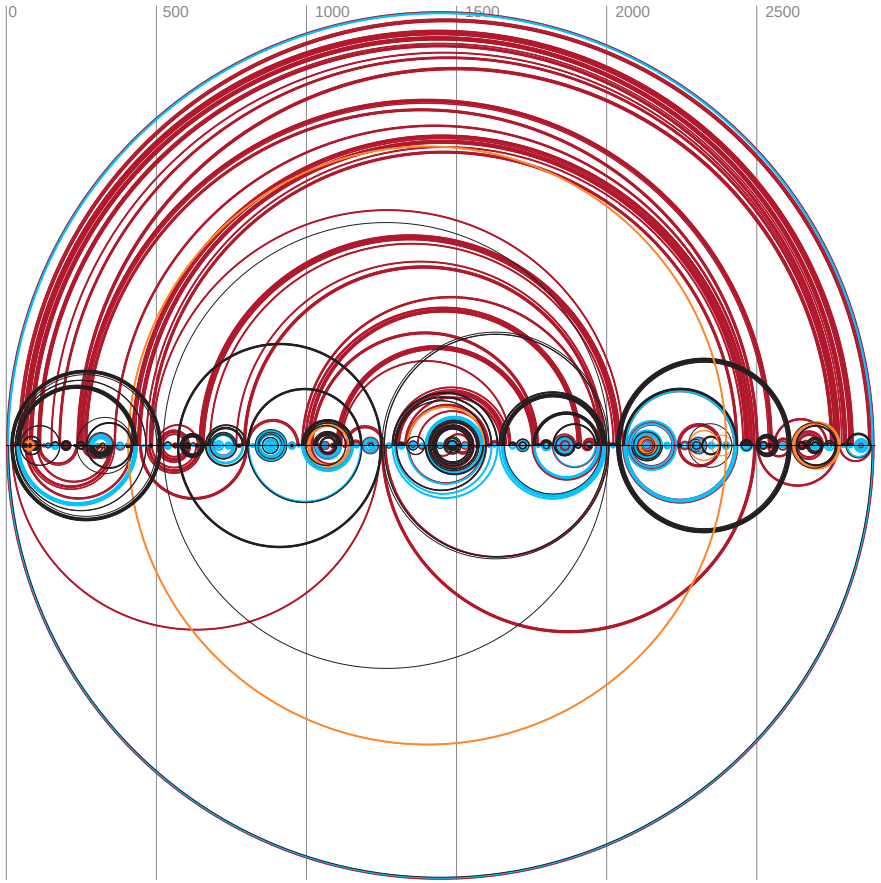


Conclusions:

- Andronescu 2007 parameters result in noticable free energy changes
- scaling-function of CoFold does not significantly (2%) change free energies

⇒ our results support original hypothesis by Morgan & Higgs (1996) that

Example: RNAFOLD versus COFOLD-A predictions for the 23S rRNA of the gamma-proteobacteria *Pseudomonas aeruginosa* (MCC +15%)



[Arc-plot made with R-CHIE, see `www.e-rna.org`]

CoFold: summary

- captures one overall effect of co-transcriptional folding
- depends on only 1 new free parameter (rather than 363)
- parameter training is robust
- improves the prediction accuracy, esp. for long sequences
- free energies of predicted MFE RNA structures hardly changed
- same memory and time complexity as RNAfold
- can confirm hypothesis of Morgan & Higgs (1996)
- to use CoFold, visit

www.e-rna.org

[J.R. Proctor, I.M. Meyer, Nucleic Acids Research (2013) 41(9):e102]

Co-transcriptional folding: part III

Are select, transient RNA structure features of co-folding pathways conserved ?
And, if yes, can we identify them computationally?

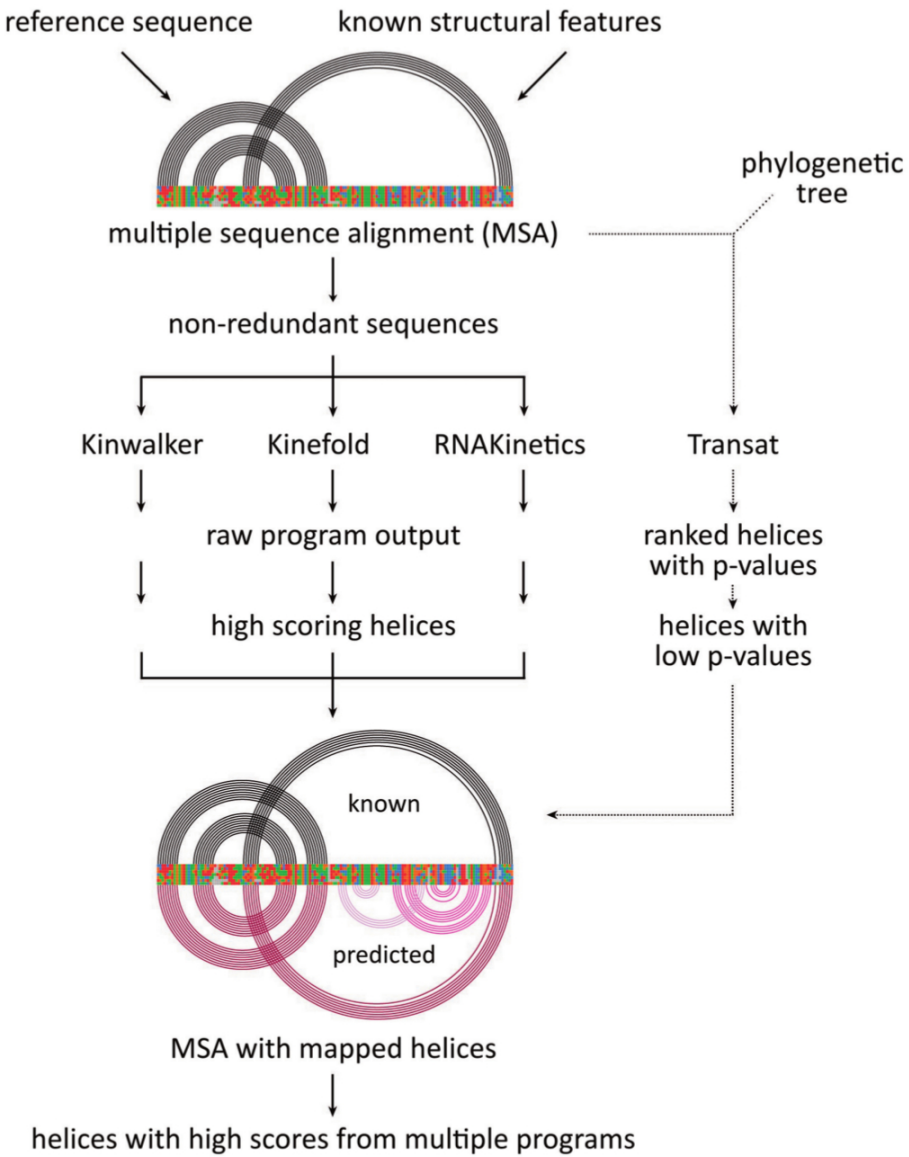
Data:

- non-redundant data set of 32 sequences extracted from 6 multiple-sequence alignments:
  - bacterial ribonuclease P Type A
  - bacterial signal recognition particle 4.5S RNA (SRP)
  - tryptophan operon leader (trp)
  - Hepatitis delta virus ribozyme (HDV)
  - Levivirus maturation gene
  - S-adenosylmethionine riboswitch (SAM)
- key features of known RNA structure features in alignments:

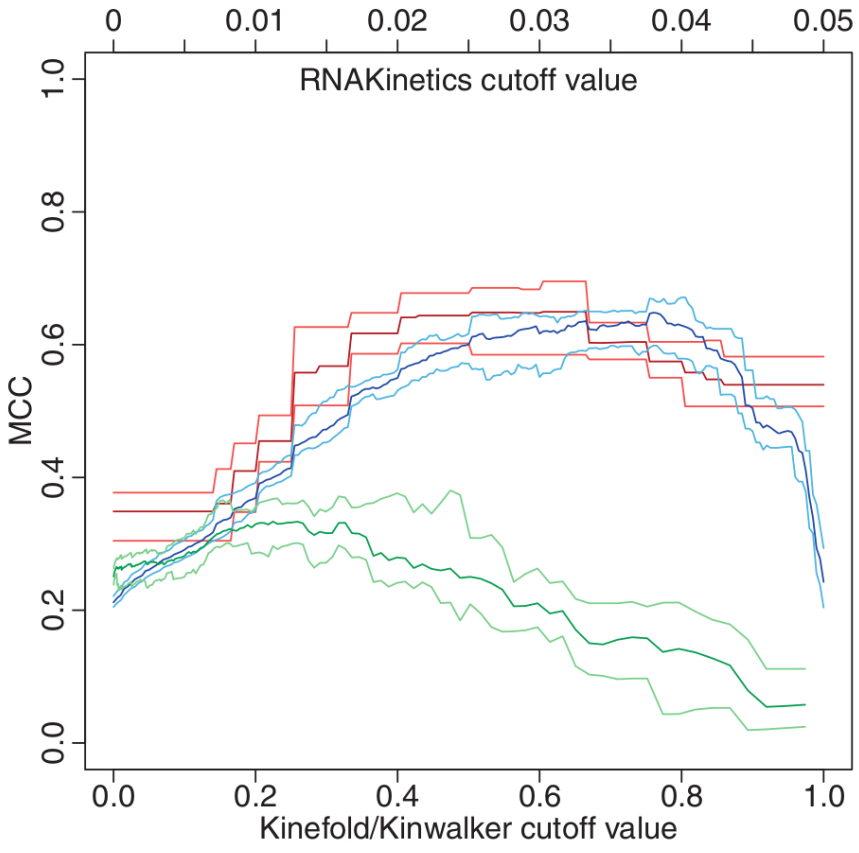
average values	Canonical bps	Covariation	Conservation	Gaps
known transient	0.91	0.10	0.77	0.02
known final	0.96	0.31	0.76	0.02

$\Rightarrow$ transient & final features conserved on approximately same level

Overall strategy:

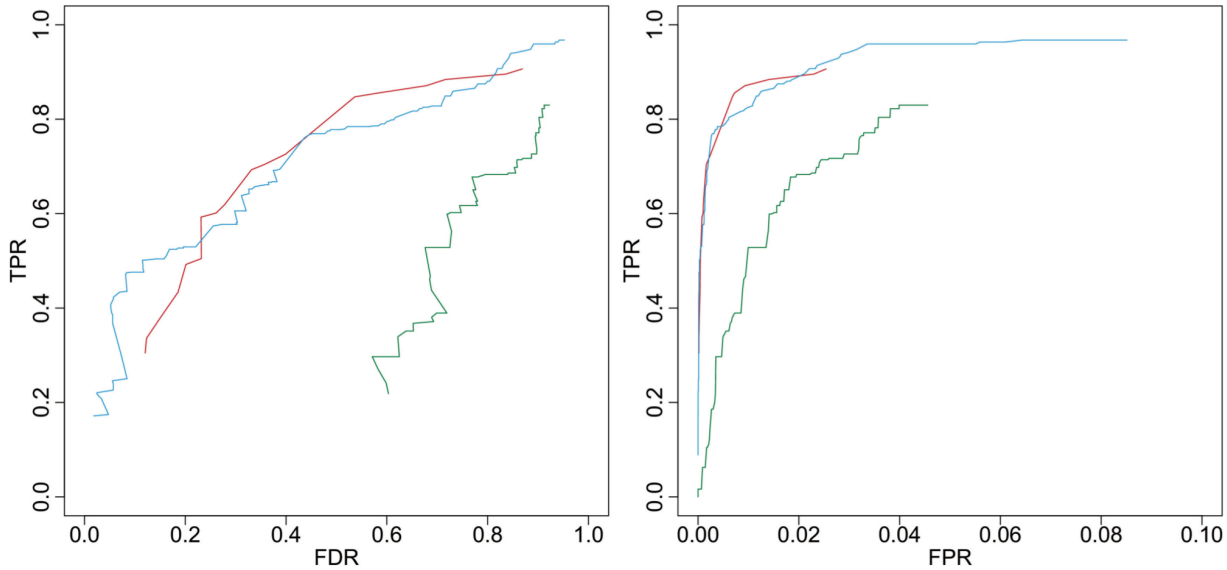


Performance evaluation for known features



Matthews correlation coefficient (MCC) for known transient and final structural features as function of the cutoff value for KINEFOLD (blue), RNAKINETICS (green) and KINWALKER (red).

Performance evaluation for known features (cont'd)



Prediction accuracy for known transient and final structural features as function of the cutoff value for KINEFOLD (blue), RNAKINETICS (green) and KINWALKER (red).

TPR	known transient	known final
KINWALKER	0.428	0.762
KINEFOLD	0.183	0.586
RNAKINETICS	0.722	0.652

TPR for MCC-optimized cut-off values.

Detecting novel transient features:

Strategy:

- use TRANSAT to predict new transient RNA structure features

  [Wiebe and Meyer, PLoS CompBio, 2010]

average values	Canonical bps	Covariation	Conservation	Gaps
new transient	**0.95**	**0.04**	**0.92**	**0.01**
known transient	0.91	0.10	0.77	0.02
known final	0.96	0.31	0.76	0.02

$\Rightarrow$ potential new transient helices are highly conserved

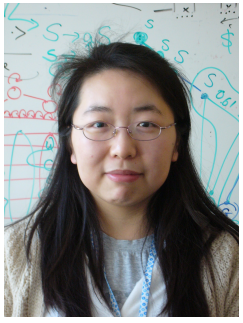
TPR	new transient	known transient	known final
KINWALKER	0.087	0.428	0.762
KINEFOLD	0	0.183	0.586
RNAKINETICS	0.322	0.722	0.652

TPR for MCC-optimized cut-off values.

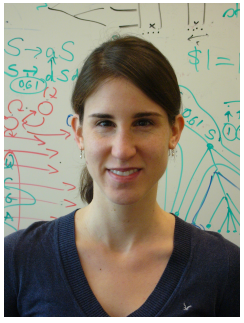
Grand summary:

- structured RNA genes not only encode the final RNA structure, but also information on how to get there co-transcriptionally, i.e. their own folding pathway *in vivo*
- co-transcriptional folding *in vivo* reduces the effective structural search space
- select transient RNA structures are highly conserved and seem to serve as guiding lamp-posts
- the level of conservation of transient RNA structures can be similar to that of final RNA structures
- conserved transient RNA structures can be predicted computationally
- **future: need more experimentally confirmed transient RNA structures and folding pathways in vivo**

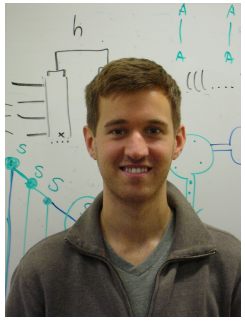
Acknowledgements:



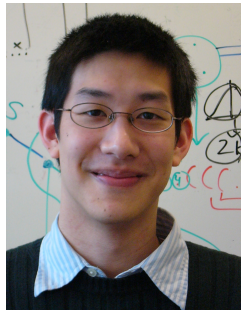
Alice Zhu



Adi Steif



Jeff Proctor



Daniel Lai

- I.M. Meyer, I. Miklós, BMC Molecular Biology (2004) 10:5
- J.Y. Zhu, A. Steif, J.R. Proctor, I.M. Meyer, Nucleic Acids Research (2013) 41(12):6273-85
- D. Lai, J. R. Proctor, I. M. Meyer, RNA (2013) 19: 1461-1473
- J.Y. Zhu, I. M. Meyer, RNA Biology (2015), 12(1):5-20
- I.M. Meyer, Methods (2017) 120:3-16
- S.R. Stefanov and I.M. Meyer, Systems Biology (2018), *in press*

Canada Foundation for Innovation
Fondation canadienne pour l'innovation

NSERC CRSNG

MARIE CURIE ACTIONS

CIHR IRSC
Canadian Institutes of Health Research
Instituts de recherche en santé du Canada

MDC MAX DELBRÜCK CENTER FOR MOLECULAR MEDICINE IN THE HELMHOLTZ ASSOCIATION

Freie Universität Berlin

www.e-rna.org