TRANSAT: Predicting functional RNA structures beyond the one-sequence-one-structure dogma

Irmtraud M. Meyer jointly with Nick Wiebe

Centre for High-Throughput Biology &
Department of Computer Science
University of British Columbia
meyer@chibi.ubc.ca

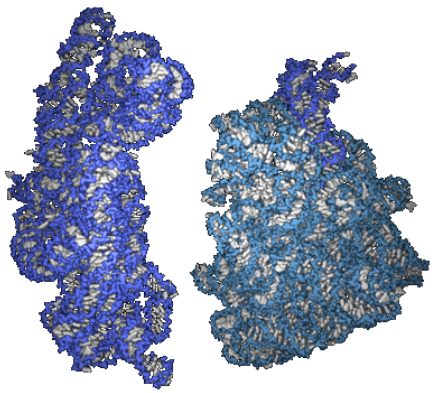
Benasque, summer 2012

Examples of well-known RNA structures:

Examples of global RNA structures, i.e. where most of RNA sequence is structured most of sequence

- tRNAs map codons of mRNA to amino-acids
- rRNAs determine ribosome's structure and function

tRNA



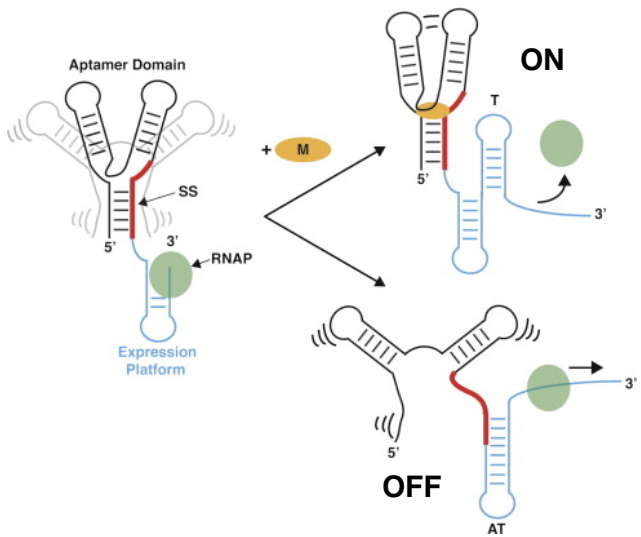
ribosome 30S (left), 50S (right)

However, not all RNA structures are global:

RNA structural elements in transcripts of protein-coding genes

- pre-mRNA: RNA editing sites, riboswitches binding metabolites, structures regulating splicing and alternative splicing
- mRNA: translation initiation and efficiency, degradation, localization, riboswitches binding metabolites

$\Rightarrow$ only part of transcript structured (local RNA structures) and one sub-sequence may encode **multiple** and **mutually exclusive** structures



Garst and Batey (2009) Biochim Biophys Acta

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
  - no detailed modeling of cellular environment (concentrations of different ions, temperature etc)
- further limitations
  - can typically only handle short sequences (typically $<<$ 1000 bp)

Examples:

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

Wishlist and inspiration for TRANSAT:

We would like to have a method that ...

- can detect conserved transient, mutually exclusive and pseudo-knotted structure elements
- does not assume that any input sequence contains a **global RNA structure**
- is fast, i.e. can be employed on a genome-wide scale and long-transcripts such as human pre-mRNAs
- highlights evolutionarily conserved structure elements
- provides more than yes-no predictions, i.e. quantifies reliability of predictions (p-value)

Wishlist and inspiration for TRANSAT:

In order to achieve this, we ...

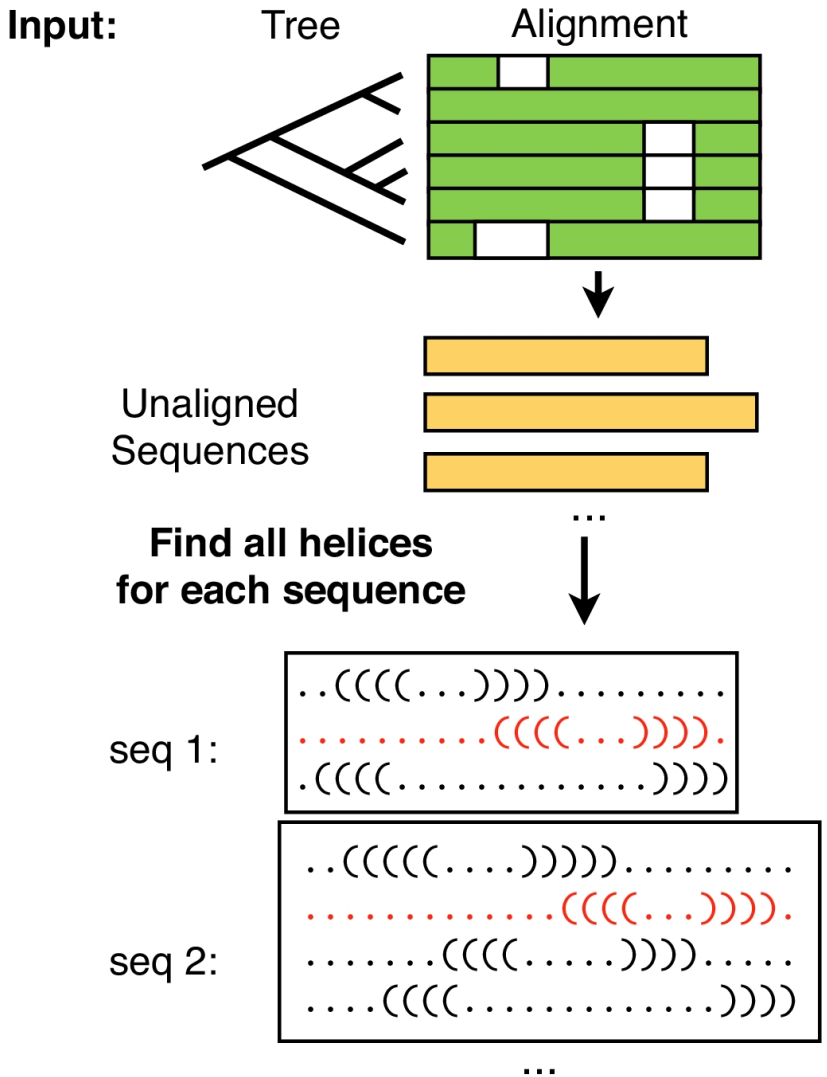
- predict individual helices rather than entire RNA structures that could be realized at the same time
- choose a comparative method that takes a fixed input alignment (this is no real limitation, see e.g. Meyer and Miklós (2007) PLoS Computational Biology)
- employ probabilistic models of RNA structure and of evolution
- use deterministic dynamic programming algorithms to predict features efficiently
- model null-distributions in order to assign reliability values

This means that we

- do not model RNA structual features as function of time (see folding pathway prediction methods), BUT
+ we do not need to model the complex cellular environment *in vivo*

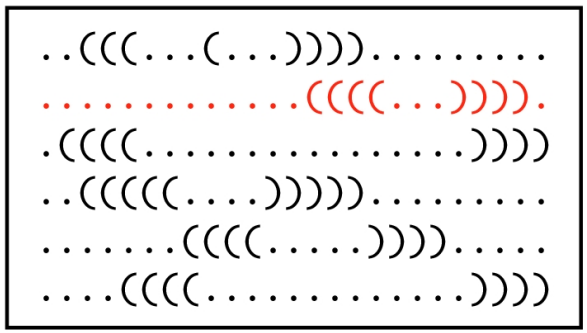
TRANSAT: underlying algorithms

TRANSAT: overall strategy



TRANSAT: overall strategy (cont'd)

Project all helices back onto alignment



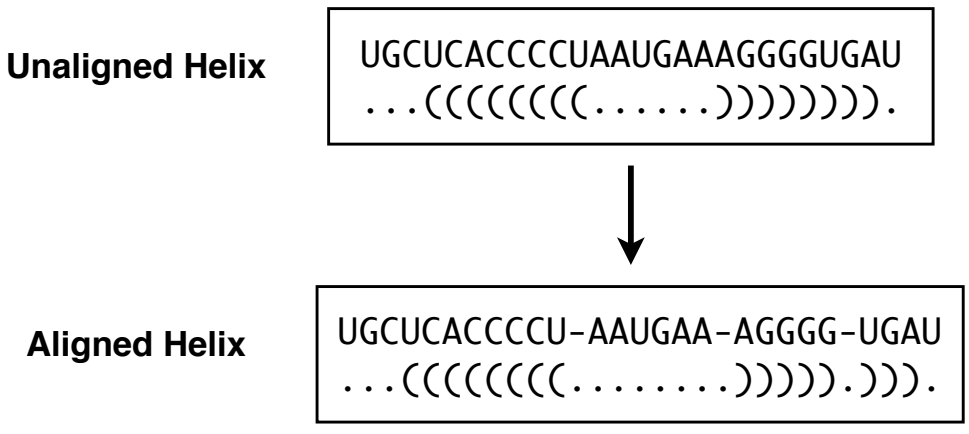
Calculate Log-likelihood score, p-value for each helix

Output Table:

P-value	Log-likelihood Score	Base Pair Positions	...
0.43	-3.45	2:15,3:14,4:13,8:12	...
0.02	1.38	13:23,14:22,15:21,16:20	...
0.62	-4.56	1:24,2:23,3:22,4:21	...

Step 1: finding and mapping helices

- find helices for each sequence individually (min length 4 base-pairs (can be specified by user))
- map helices of sequences back to alignment (conserved helices)



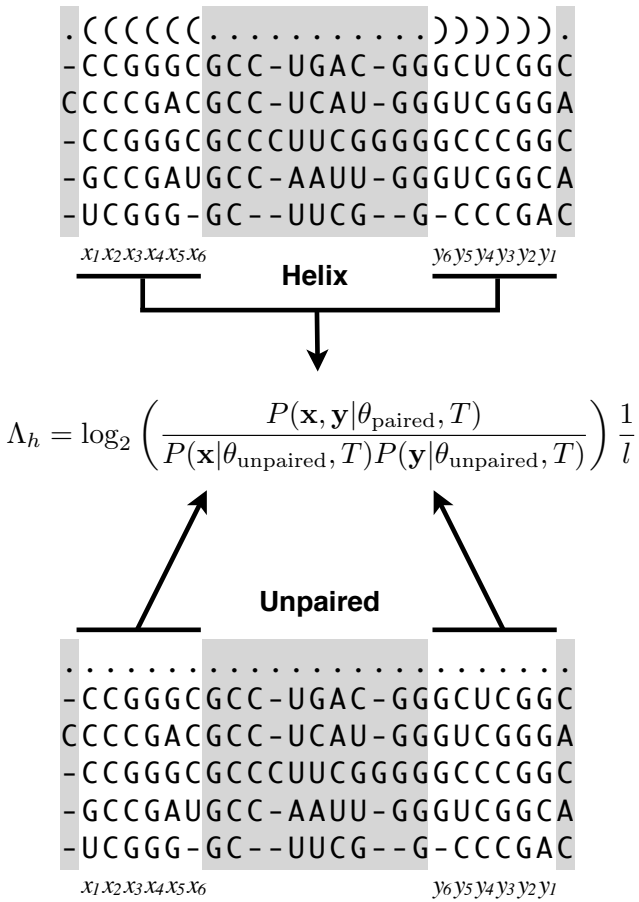
⇒ **Advantage:**

- less dependent on alignment quality than when detecting helices for entire alignment

Step 2: calculating log-likelihood values

For each conserved helix h in alignment, calculate a log-likelihood value Δ_h to test two competing hypotheses:

- $\Delta_h < 0 \Rightarrow$ two regions more likely to be unpaired
- $\Delta_h \geq 0 \Rightarrow$ two regions more likely to form a helix



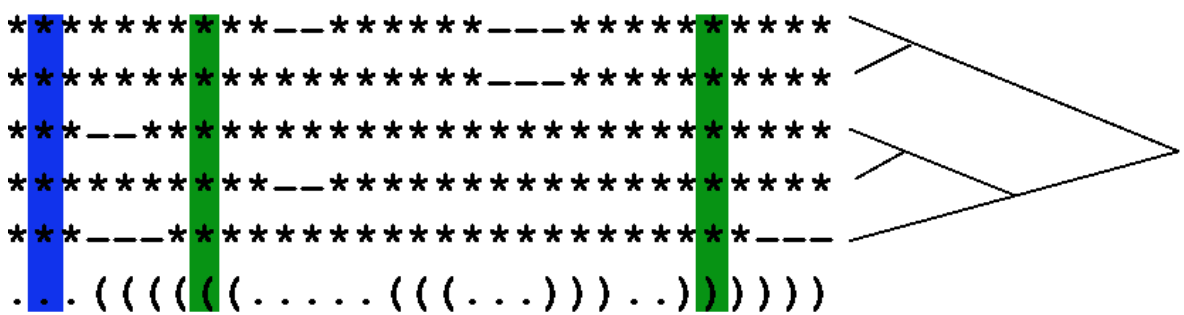
Step 2: calculating log-likelihood values (cont'd)

Have two probabilistic evolutionary models to calculate the log-likelihood values using the Felsenstein algorithm:

- evolutionary model for base-pairs (rate matrix is 16x16 matrix)
- evolutionary model for un-paired nucleotides (rate matrix is 4x4 matrix)

Key ideas of Felsenstein algorithm:

- consider only the *observed* nucleotides/base-pairs at **leaf nodes** of evolutionary tree
- sum over all possibilities for nucleotides/base-pairs at **internal tree nodes** and weight them according to their corresponding evolutionary model



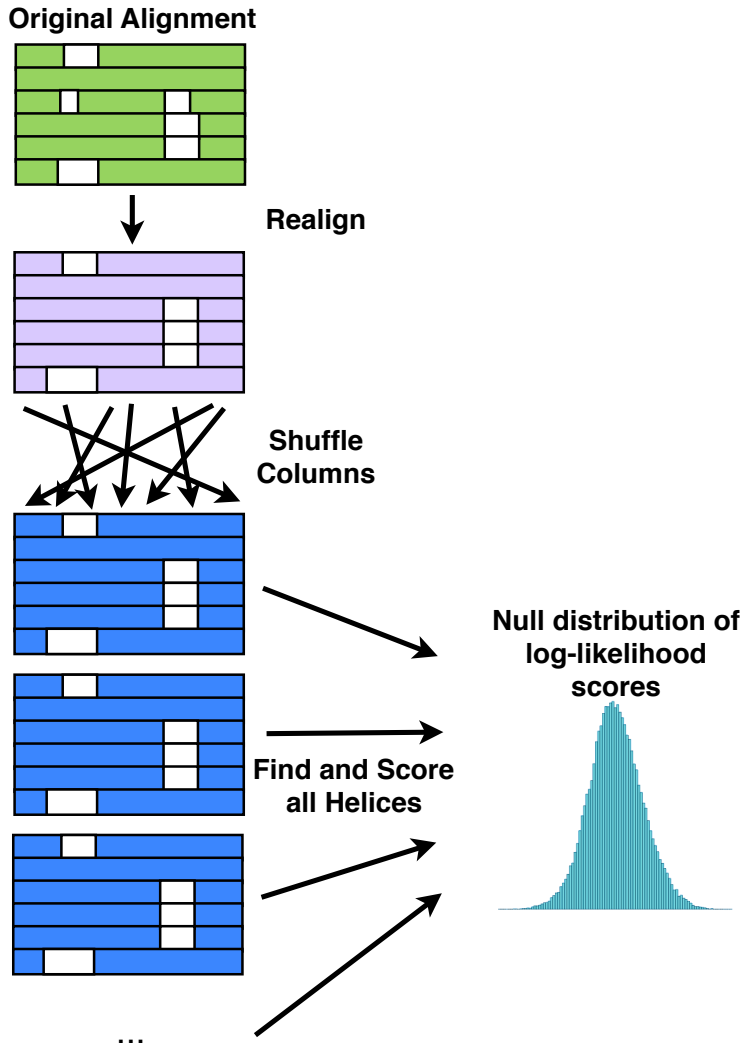
Step 3: estimating p-values for log-likelihood values

Challenge:

- range of log-likelihood values very much depends on properties of each input alignment
- ideally, we would like to know for each log-likelihood value what the probability of seeing it by chance is (i.e. its p-value)

$\Rightarrow$ Solution:

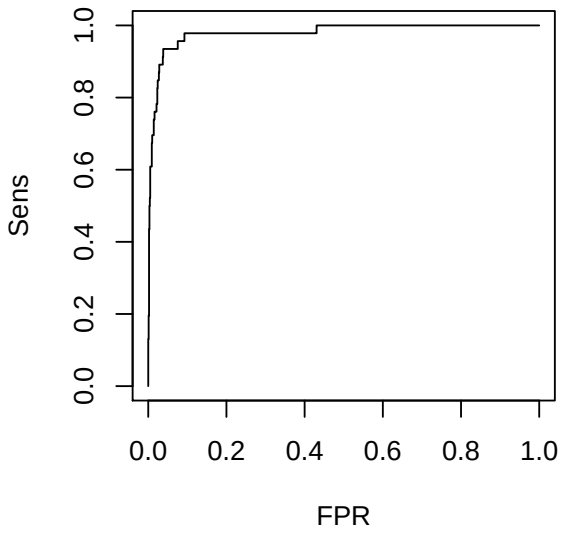
estimate p-values for log-likelihood values in each input alignment

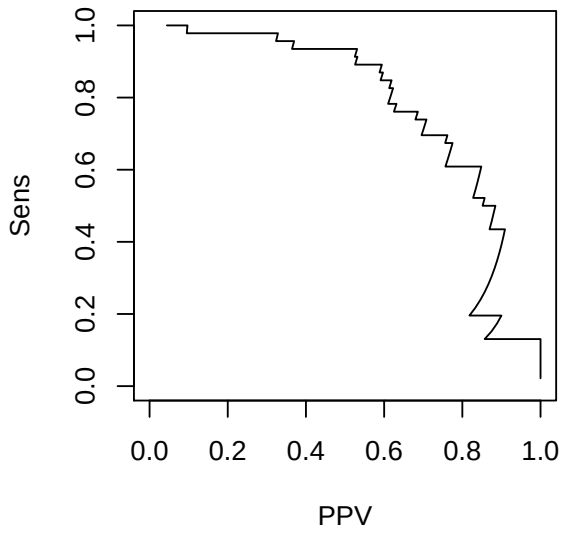


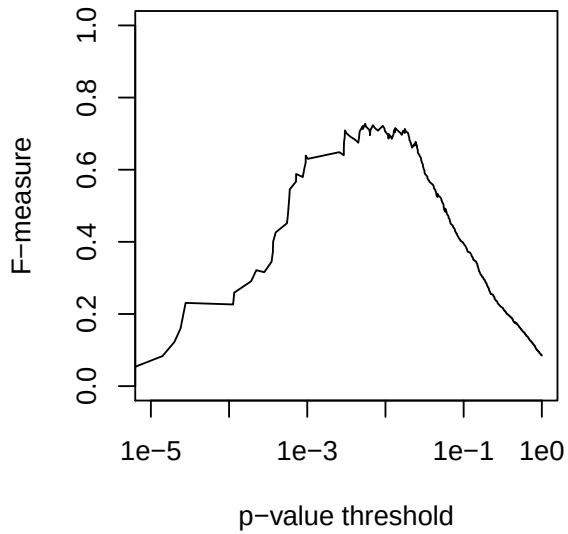
Data sets for performance evaluation:

1. two sequences with known multiple RNA structures
   - hok data set: 9 sequences, 196 bp length, total tree length 2.31
   - trp data set: 8 sequences, 117 bp length, total tree length 2.29
2. set of 134 high-quality alignments from the RFAM database [Gardner *et al.* (2009) NAR 37:D136-140] (Rfam data set)
   - structural annotation is correct, but may not be complete
   - 6 to 712 sequences per alignment, 100 to 1247 bp length, total tree length 0.4 to 116.3 (average 10.0)
3. set of 990 artificially generated alignments (artificial data set) generated by GENERAID (unpublished)
   - structural annotation is correct *and* complete
   - no alignment errors
   - can perform detailed tests
   - derived for known structures from the RNA STRAND database [Andronescu *et al.* (2008) BMC Bioinformatics 9:340]
   - 10 sequences per alignment, 100 to 1000 bp length, total tree length 0.5 to 16

Results: hok data set







Performance definitions:

$$\text{Sens} := \text{TP}/(\text{TP} + \text{FN}) \tag{1}$$

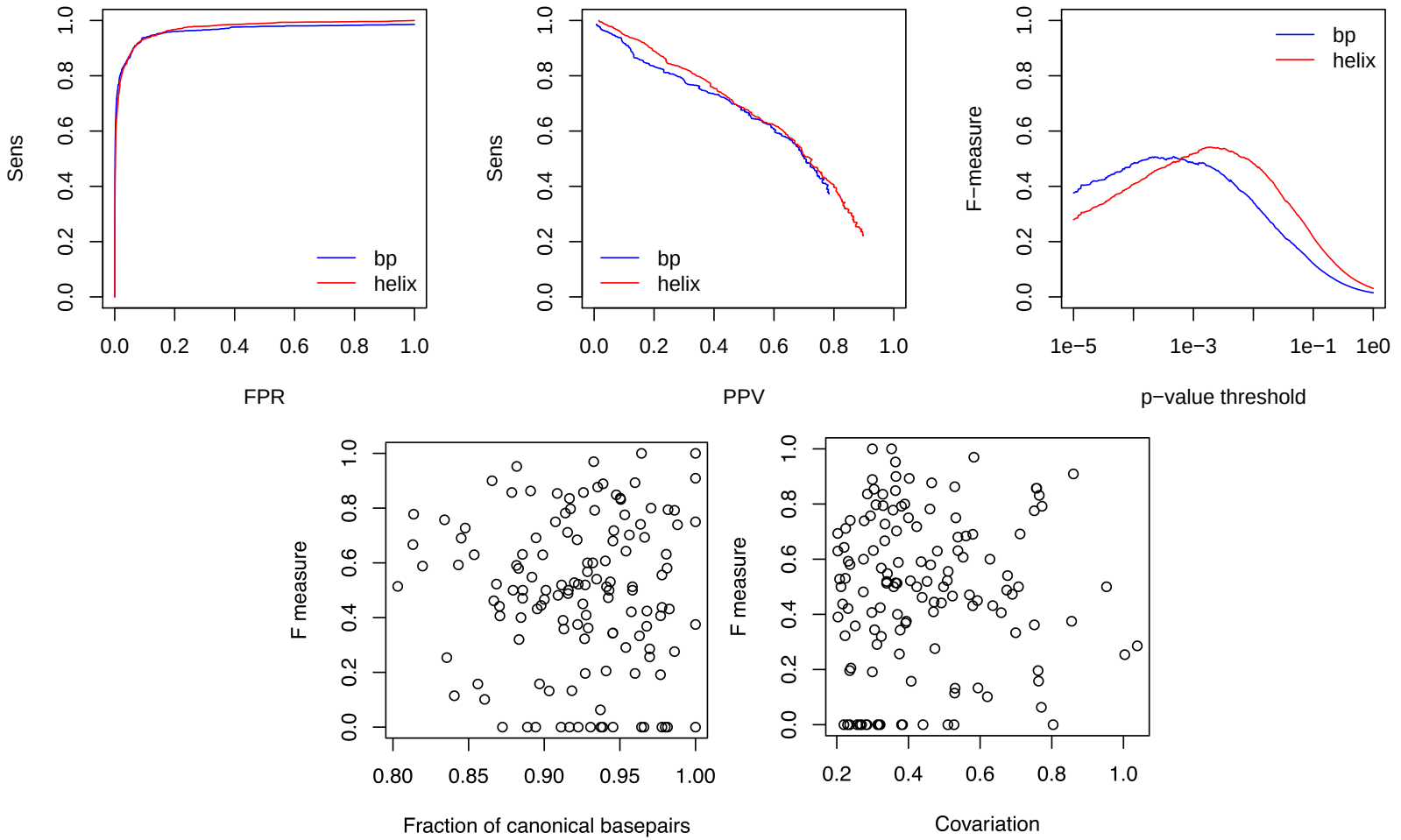
$$\text{FPR} := \text{FP}/(\text{TN} + \text{FP}) \tag{2}$$

$$\text{PPV} := \text{TP}/(\text{TP} + \text{FP}) \tag{3}$$

$$\text{F-measure} := \frac{2 \cdot \text{Sens} \cdot \text{PPV}}{\text{Sens} + \text{PPV}} = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FN} + \text{FP}} \tag{4}$$

where TP (true positives), TN (true negatives), FP (false positives) and FN (false negatives)

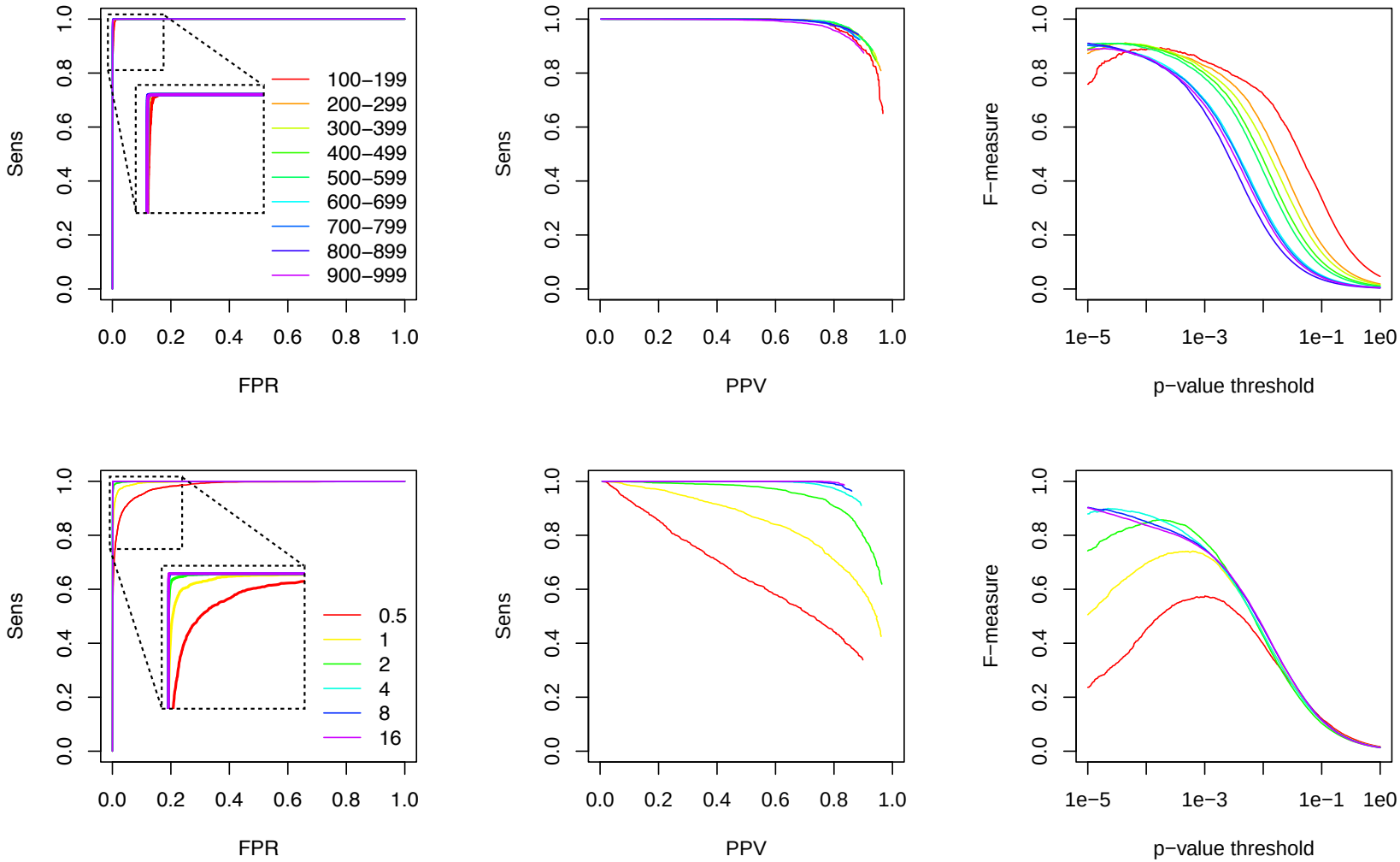
Results: RFAM data set



⇒ select a p-value of 10^{-3} as default threshold

⇒ no correlation with alignment quality (at least in our RFAM data set)

Results: artificial data set

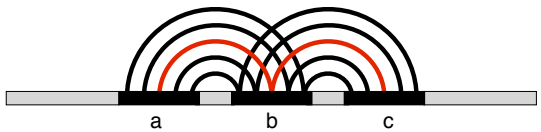


⇒ little dependence on alignment length

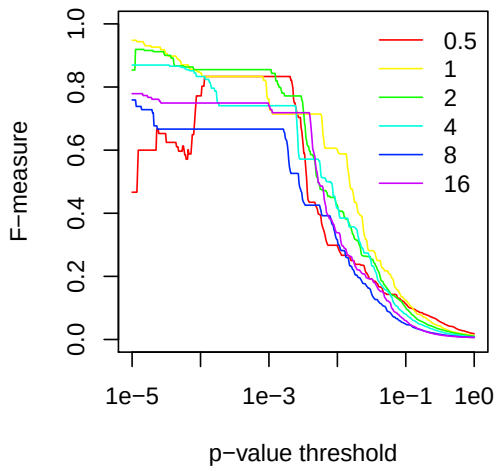
⇒ fairly strong dependence on total tree length

Results: artificial data set for overlapping helices

- develop a novel evolutionary model for overlapping helices
- key feature: **one** nucleotide may be simultaneously base-pair with up to **two** other nucleotides

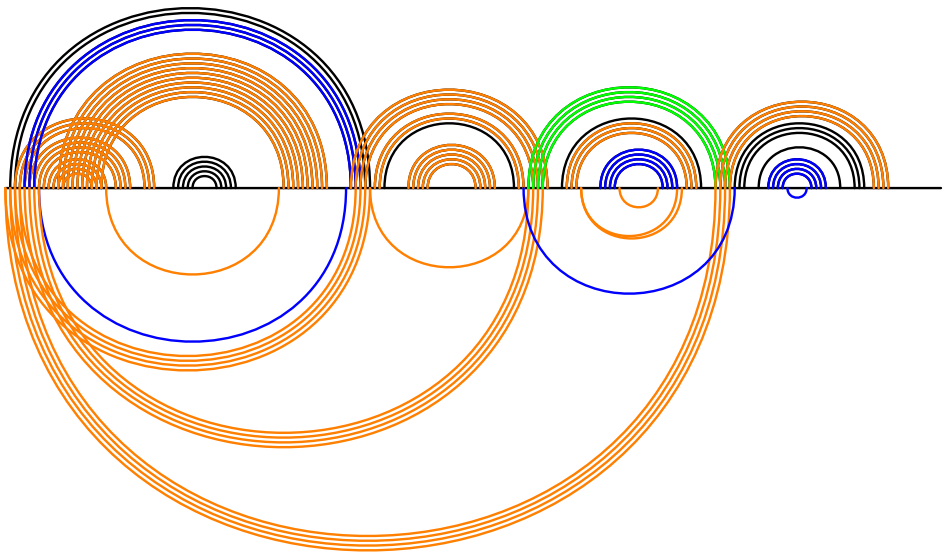


- use this model to generate artificial data set with overlapping helices in order to test if TRANSAT can reliably detect them



$\Rightarrow$ TRANSAT can predict overlapping helices well for a wide range of total tree lengths

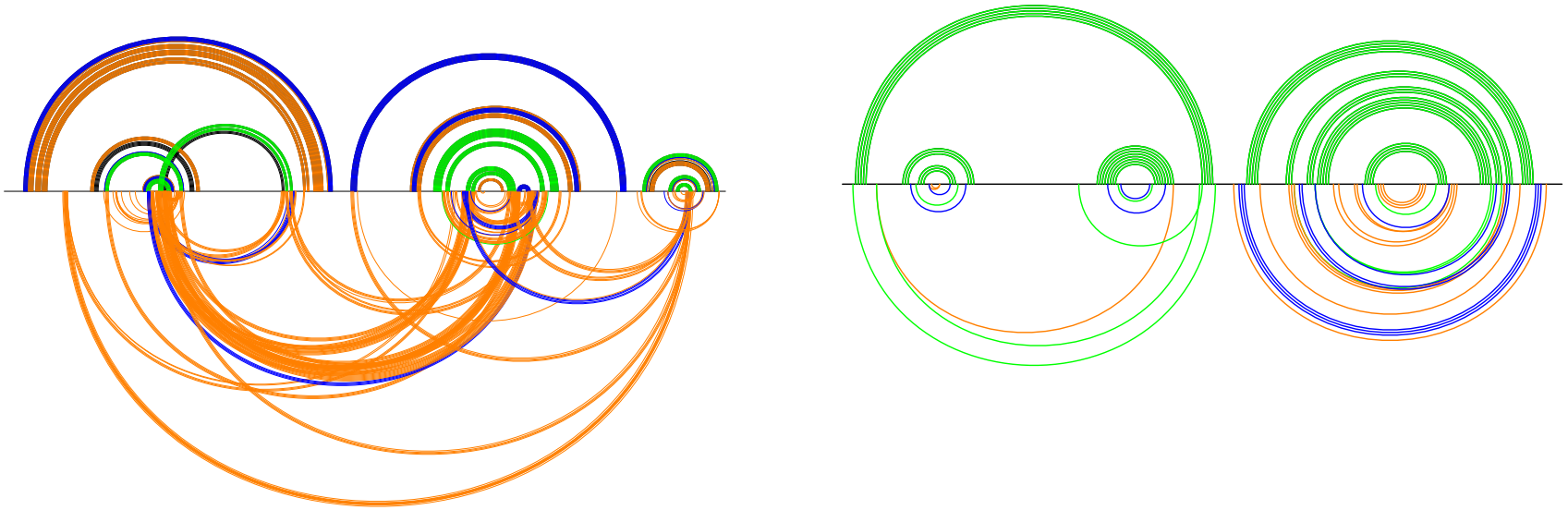
Predictions for *hok* alignment: arc-diagram



- default p-value threshold: 10^{-3}
  - horizontal line: input alignment
  - top arcs: known base-pairs (black if not predicted)
  - bottom arcs: new base-pairs
  - colour coding for predicted base-pairs:
    $< 10^{-5}$ green,
    $< 10^{-4}$ blue,
    $< 10^{-3}$ orange,
    $<$ (p-value threshold) red

- TRANSAT predicts most helices of the two known structures
- in addition, TRANSAT predicts three statistically significant, mutually exclusive conserved helices

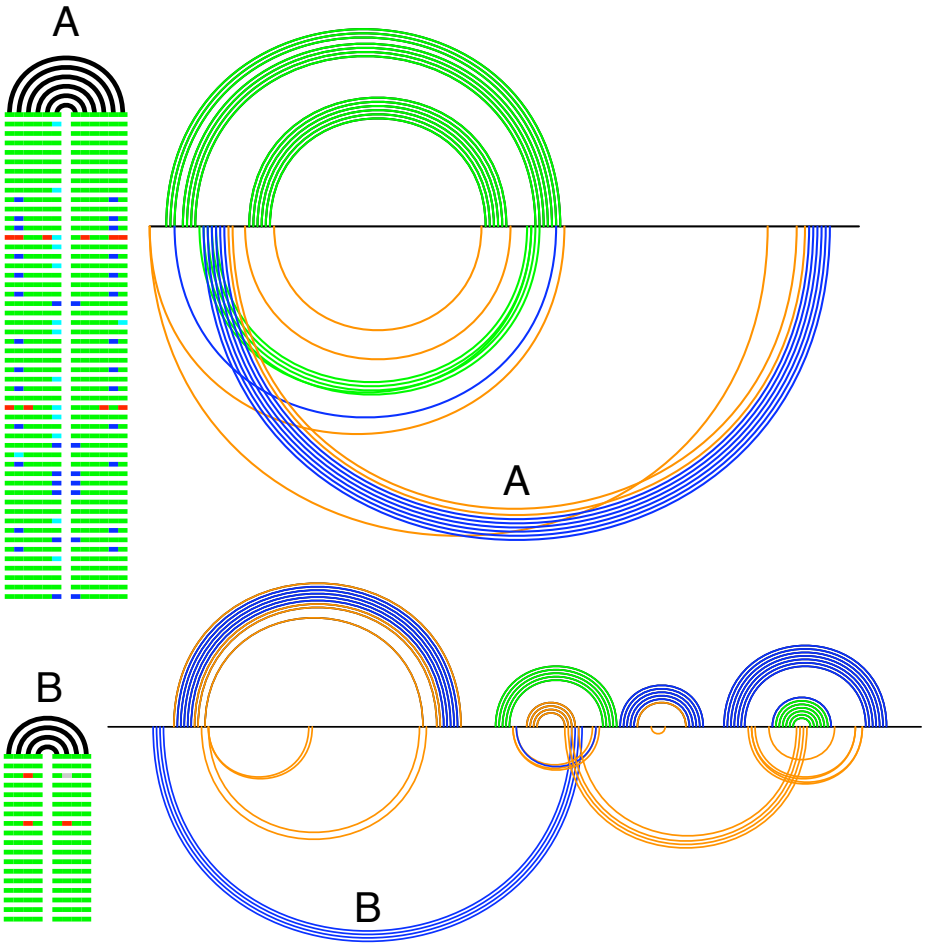
Predictions for vertebrate and ciliate telomerases



Vertebrate telomerase (left, RF00024) and ciliate telomerase (right, RF00025) for a p-value threshold of 10^{-3}.

- known pseudo-knotted structure of vertebrate sequences captured well by TRANSAT prediction
- folding of vertebrate sequences may involve large-range structural arrangements

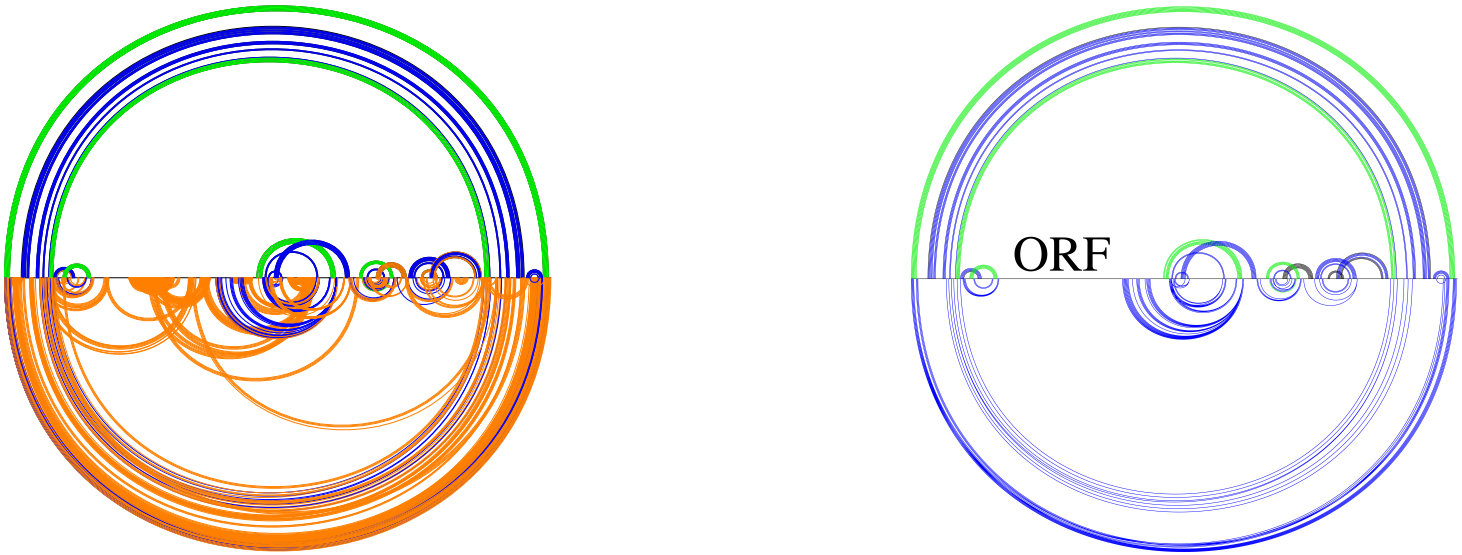
Evidence for new pseudo-knots



- new helices A and B make known structure pseudo-knotted
- little covariation for helix B, but a lot for helix A
- pseudo-knots typically ignored in computational predictions and easily missed in manual annotation

S-adenosyl-L-homocysteine riboswitch family (top, RF01057), a riboswitch found on certain bacterial mRNAs, and the glmS glucosamine-6-phosphate activated ribozyme (bottom, RF00234), a bacterial ribozyme for a p-value threshold of 10^{-3}.

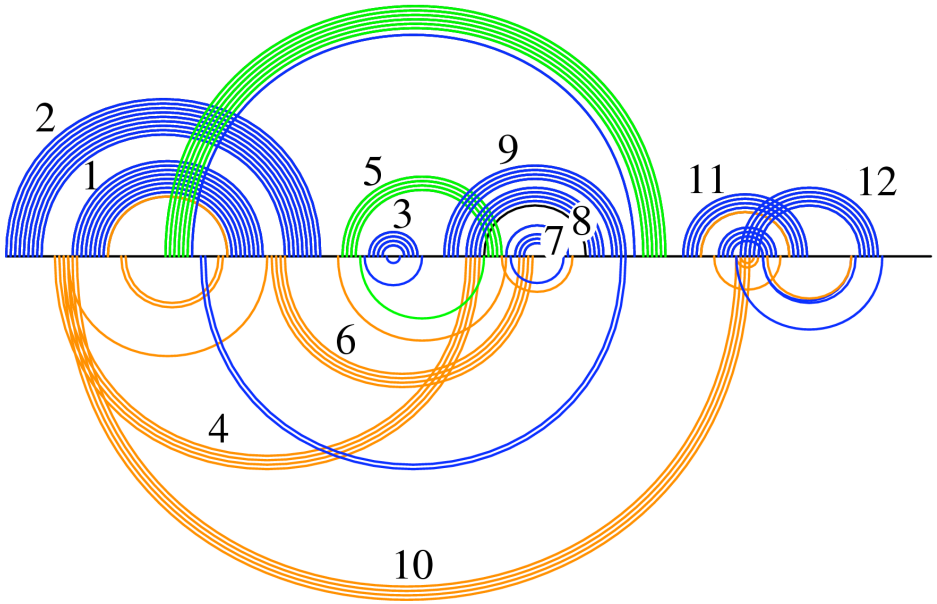
Highlighting un-structured sequence regions

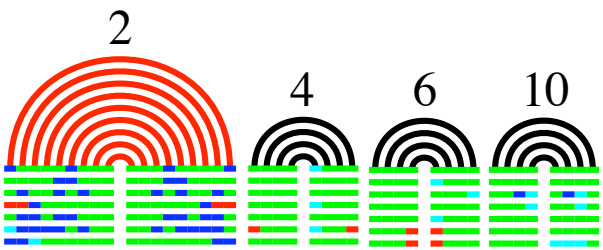


Bacterial transfer-messenger RNA (tmRNA) (RF00023) for a p-value threshold value of 10^{-3} (left) and 10^{-4} (right).

- known pseudo-knotted structure of vertebrate sequences captured well by TRANSAT prediction
- region of tmRNA that contains open reading frame (ORF) is devoid of significant helices

Information on potential folding pathways *in vivo*





- helices 4, 6 and in particular 10 show covariation, but not on the same level as known helix 2
- predicted helices suggest time-wise ordering of potential folding pathway

Cripavirus internal ribosomal entry site (IRES), RF00458, for a p-value threshold of 10^{-3}.

Summary TRANSAT

Main features:

- takes a fixed input alignment and tree and predicts stat. significant, conserved helices

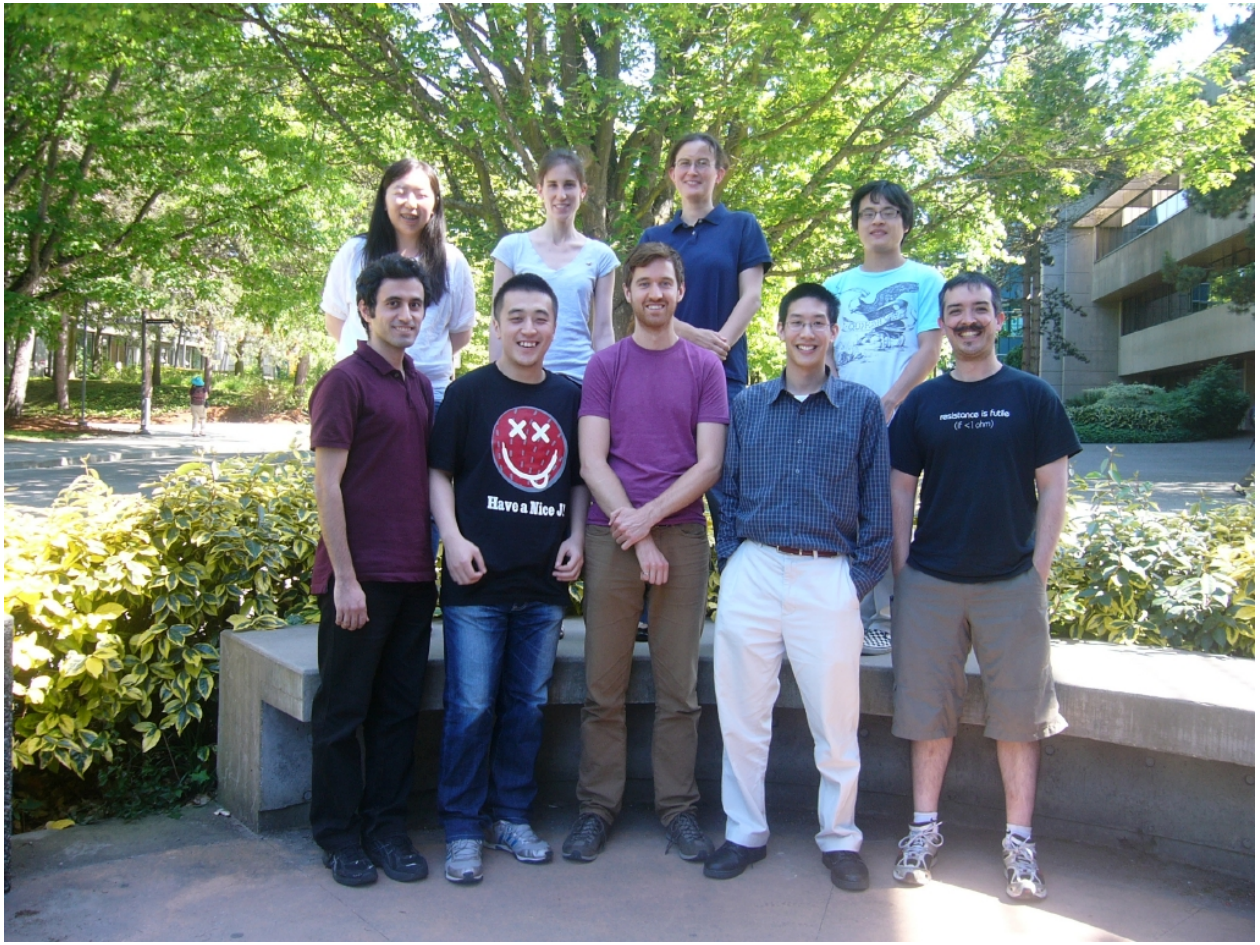
Disadvantage:

- does not predict folding pathway as function of the time

Advantages:

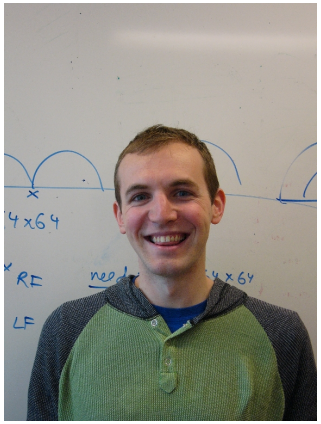
+ capable of detecting transient, competing and pseudo-knotted helices that have been conserved
+ fairly robust w.r.t. aligmnent errors
+ does not require modeling of detailed cellular environment and makes very few assumptions
+ assigns reliability values to its predictions
+ high performance accuracy for a wide range of data
+ fast and memory efficient

Acknowledgements:



My group (including Evan, the photographer ...) enjoying a precious day without rain.

Acknowledgements:

Nick Wiebe 

- TRANSAT: Wiebe & Meyer, PLoS Compbio (2010), 6(6):e1000823.
- TRANSAT: web-page at `www.cs.ubc.ca/~irmtraud/transat/`
- R-CHIE: Lai, Proctor, Zhu and Meyer, NAR (2012) 40(12):e95.
- R-CHIE web-server at `www.e-rna.org/r-chie`

Funding:

- Canadian Foundation for Innovation
- CIHR, Canada
- NSERC, Canada