CoFold: thermodynamic RNA structure prediction with a kinetic twist

Irmtraud M. Meyer
joint work with Jeff Proctor

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

Benasque, July 2012

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 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]

Key motivation:

3. Structured RNA genes not only encode information about their functional structures, but also on their co-transcriptional folding pathway (and, e.g. transient structures).

   [Meyer and Miklós (2004), BMC Mol Biol 10]

4. 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]

Turns out that . . . yes, we can!

Key motivation:

3 Structured RNA genes not only encode information about their functional structures, but also on their co-transcriptional folding pathway (and, e.g. transient structures).

[Meyer and Miklós (2004), BMC Mol Biol 10]

4 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]

Turns out that ... yes, we can!

Key motivation:

3 Structured RNA genes not only encode information about their functional structures, but also on their co-transcriptional folding pathway (and, e.g. transient structures).

[Meyer and Miklós (2004), BMC Mol Biol 10]

4 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]

Turns out that ... yes, we can!

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 $\leq$ 1000 bp)
- no comprehensive performance evaluation yet

Examples:

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

CoFold: key goal

Combine the success of thermodynamic methods with the conceptual beauty of folding pathway prediction methods.

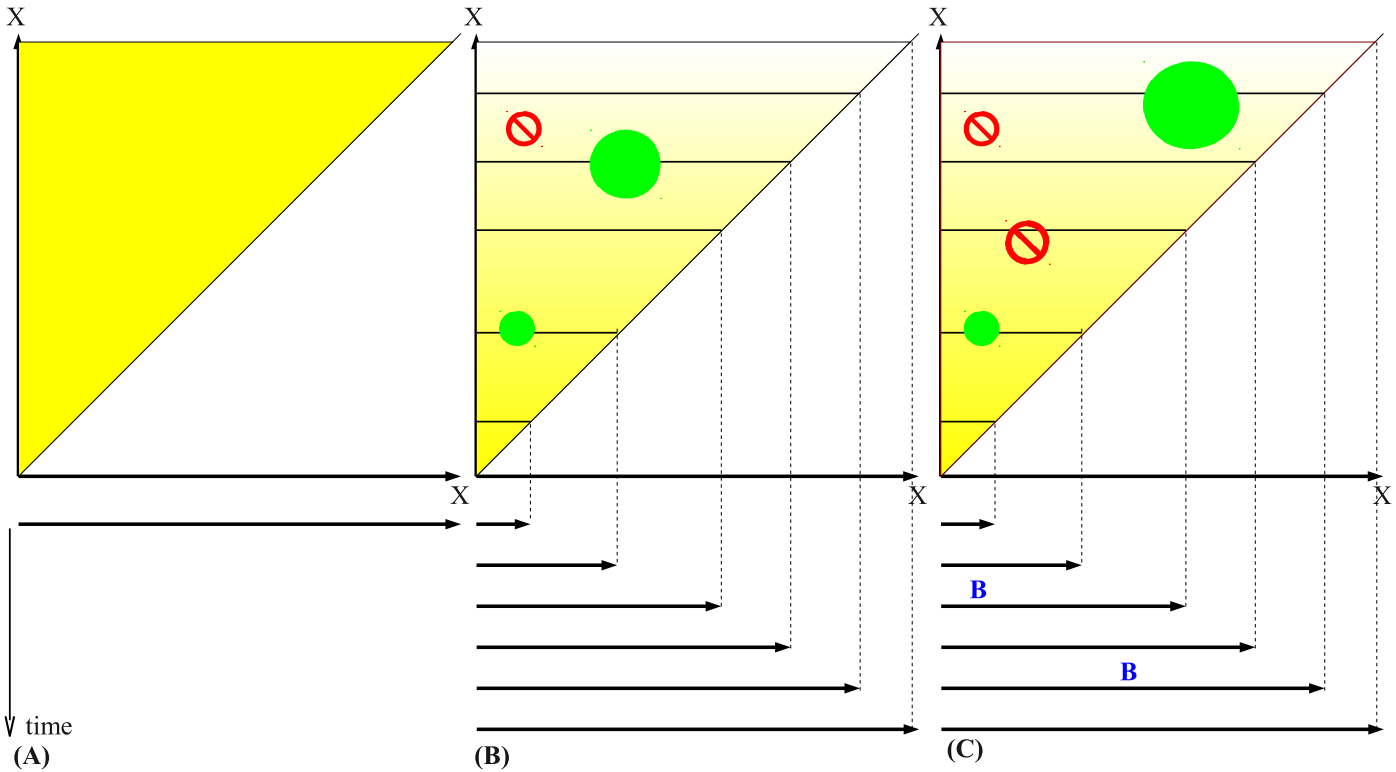
Key challenge:

- RNA structure prediction algorithms such as the one underlying RNA-Fold have no concept of a folding pathway and **ignore** the process of structure formation.
- A transcript emerging and folding co-transcriptionally *in vivo*, however, needs to **find a way of actually reaching the functional RNA structure**, i.e. the folding **process is key**.

Key challenge:

- RNA structure prediction algorithms such as the one underlying RNA-Fold have no concept of a folding pathway and **ignore** the process of structure formation.
- A transcript emerging and folding co-transcriptionally *in vivo*, however, needs to **find a way of actually reaching the functional RNA structure**, i.e. the folding **process is key**.

Key challenge:



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

CoFold: to-do list

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

CoFold: the nitty-gritty details

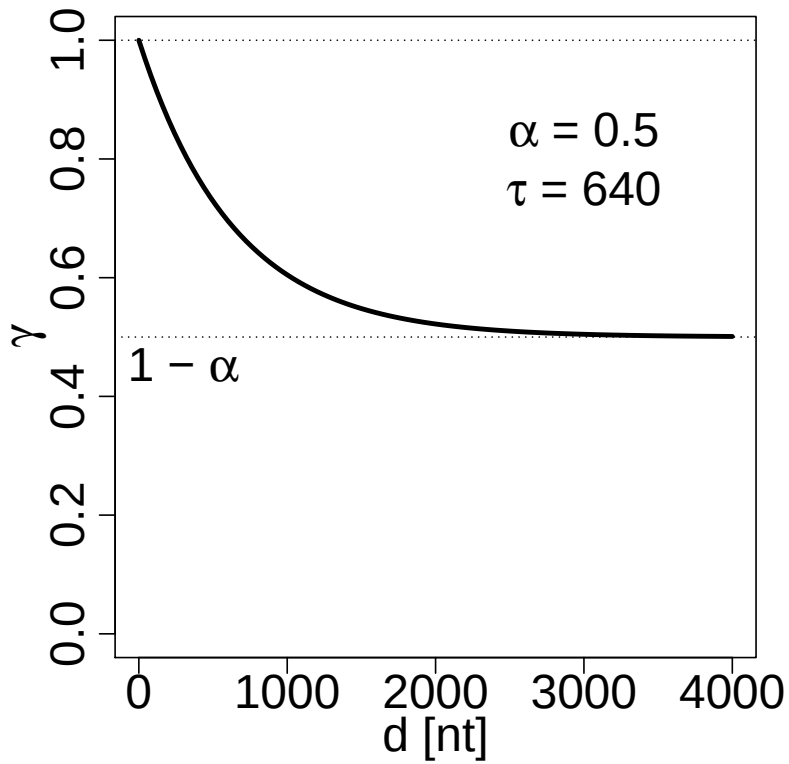
- introduce a scaling-function that **judges the reachability of potential base-pairing partners during kinetic folding**
- justification: potential base-pairing partners nearby are easier to identify than those further apart
- scaling-function

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

which depends on 2 free parameters α and τ

CoFold: the nitty-gritty details

- scaling-function $\gamma(d) := \alpha \cdot (e^{-\frac{d}{\tau}} - 1) + 1$
- apply $\gamma(d)$ to stacking interactions (stab. contrib.) and loops, bulges (dest. contrib.)
- needed to preserve relative magnitude of energy contributions



CoFold

Compiling large and diverse high-quality data sets for training and testing.

Checking the robustness of parameter training.

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:

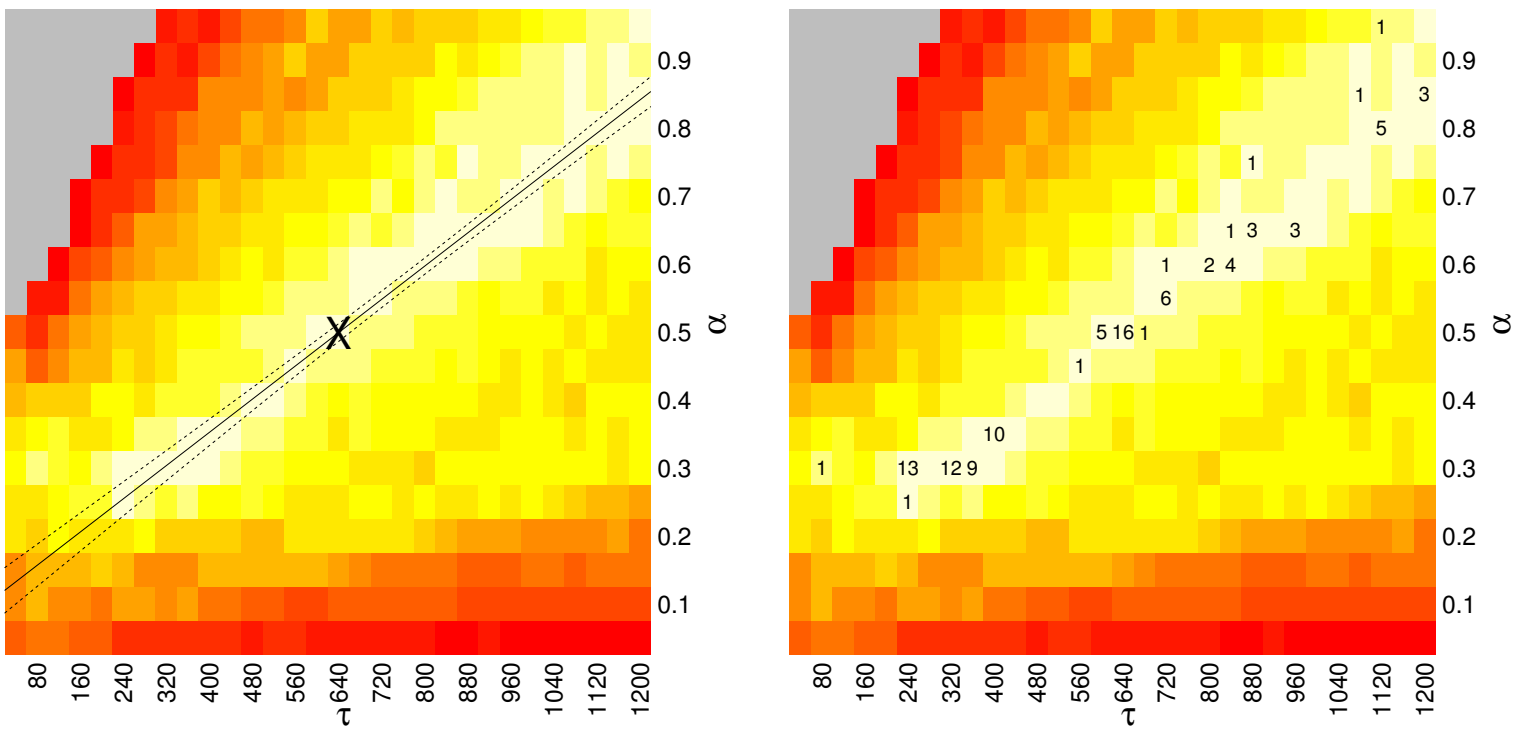
- only biological sequences
- ref. structures supported by strong evol. evidence
- long data set: length $>$ 1000 nt and pairw. % seq. id $\leq$ 85%
- long data set $\Rightarrow$ non-redundant 16S and 23S rRNAs

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

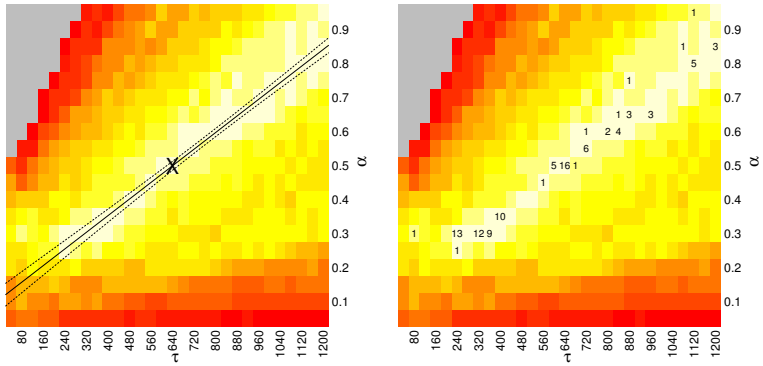
CoFold: parameter training

Method:

- 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 parameter 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

CoFold: *What is the prediction accuracy?*

Introducing CoFold-A and RNAfold-A

Benchmark performance using the following four methods:

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

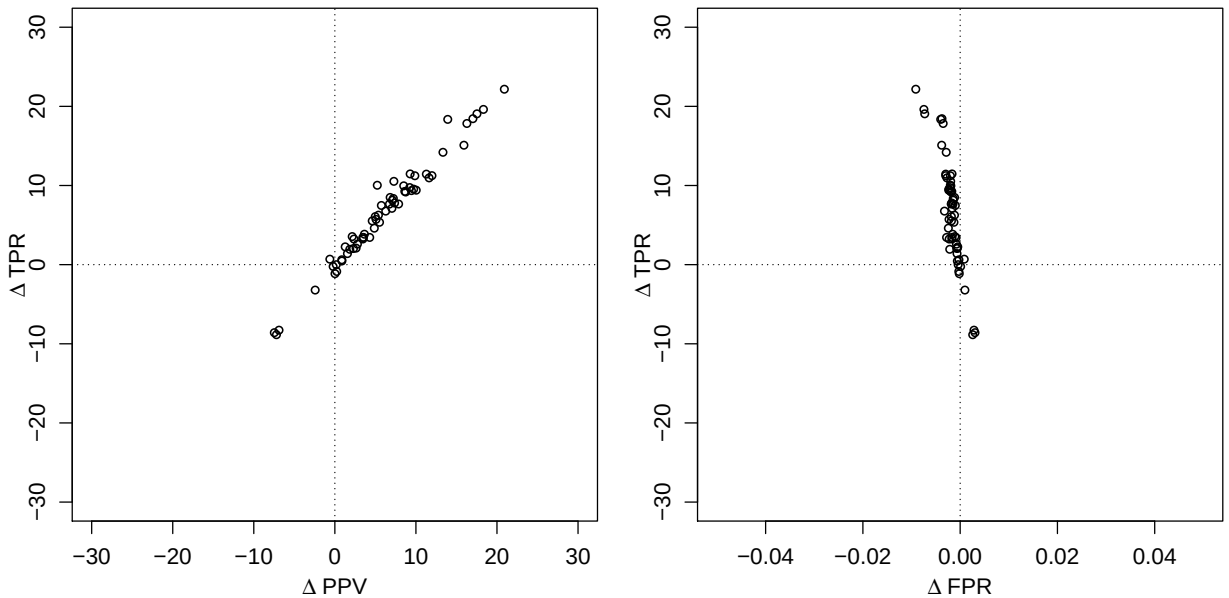
  [Mathews et al. (1999) J Mol Biol 288: 5]
- CoFold-A and RNAfold-A: use Andronescu energy model (2007)

  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

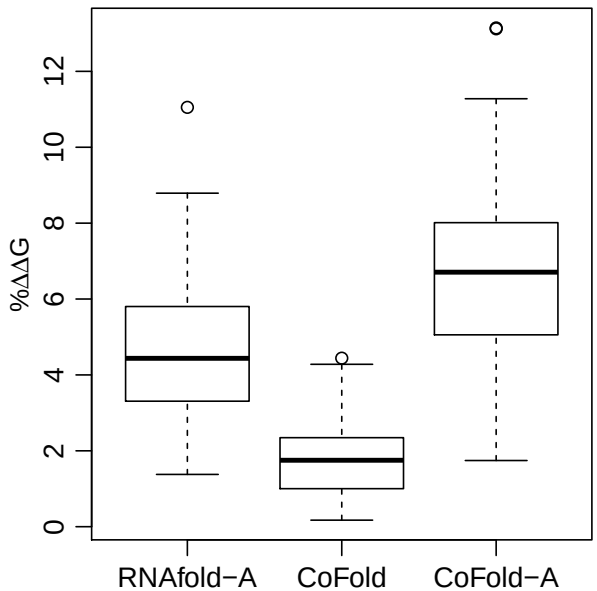
	TPR (%)	FPR (%)	PPV (%)	**MCC** (%)
RNAfold	46.30	0.0176	39.74	**42.81**
RNAfold-A	52.02	0.0160	44.76	**48.17**
CoFold	52.83	0.0159	45.79	**49.10**
CoFold-A	57.80	0.0145	50.06	**53.70**

Bottom line:

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

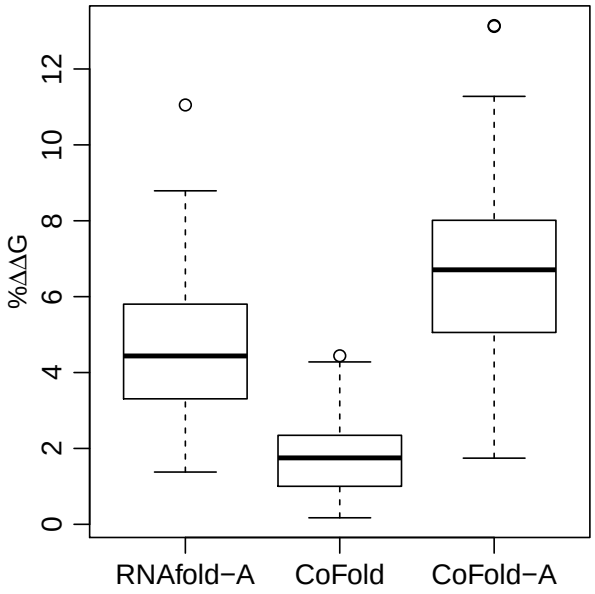
CoFold: influence on structures' free energies

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



	av. (%)	stdev (%)	max (%)
RNAfold-A	5	1.9	11.1
CoFold	2	1.0	4.4
CoFold-A	7	2.4	13.1

CoFold: influence on structures' free energies

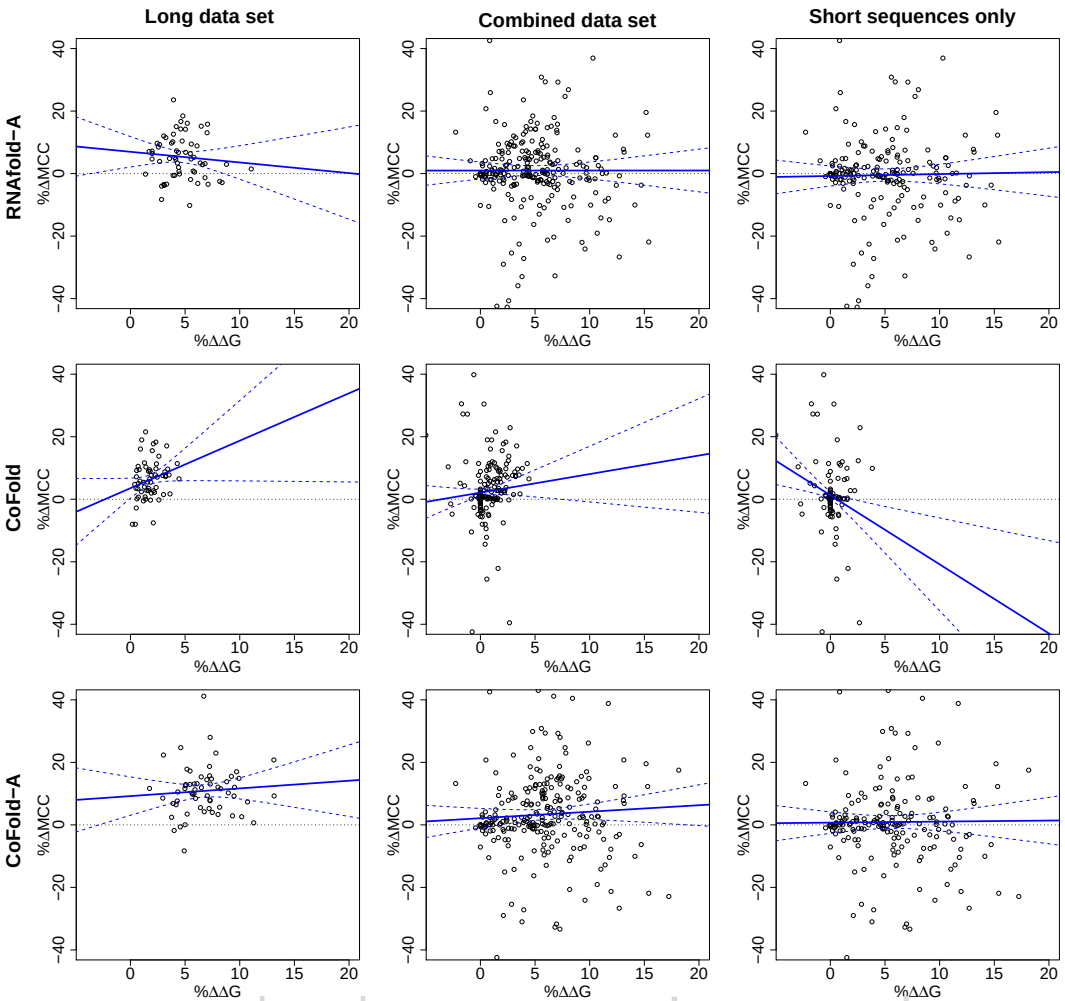


Conclusions:

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

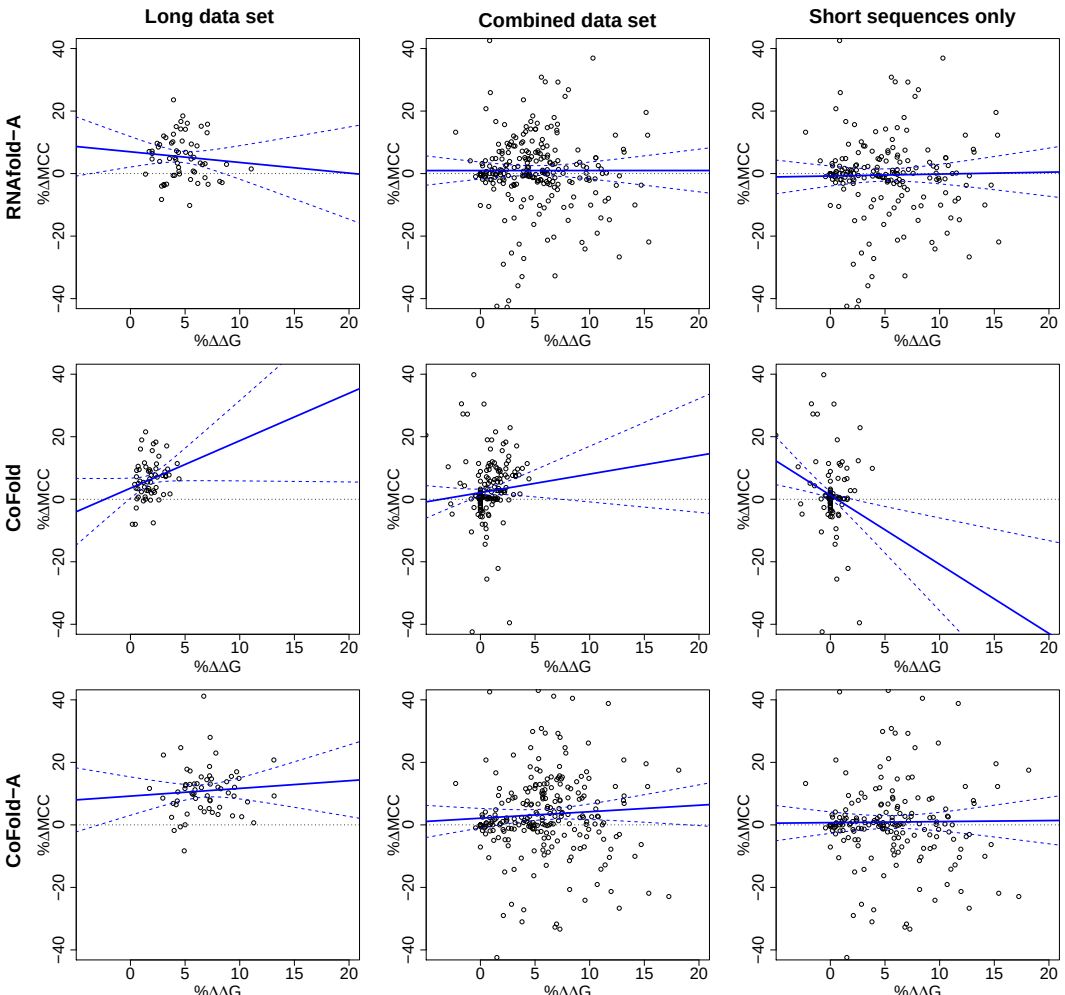
$\Rightarrow$ our results support original hypothesis by Morgan & Higgs (1996) that differences between conserved and predicted MFE structures not primarily due to errors in energy models

Do large energy changes correlate with improved prediction accuracy?



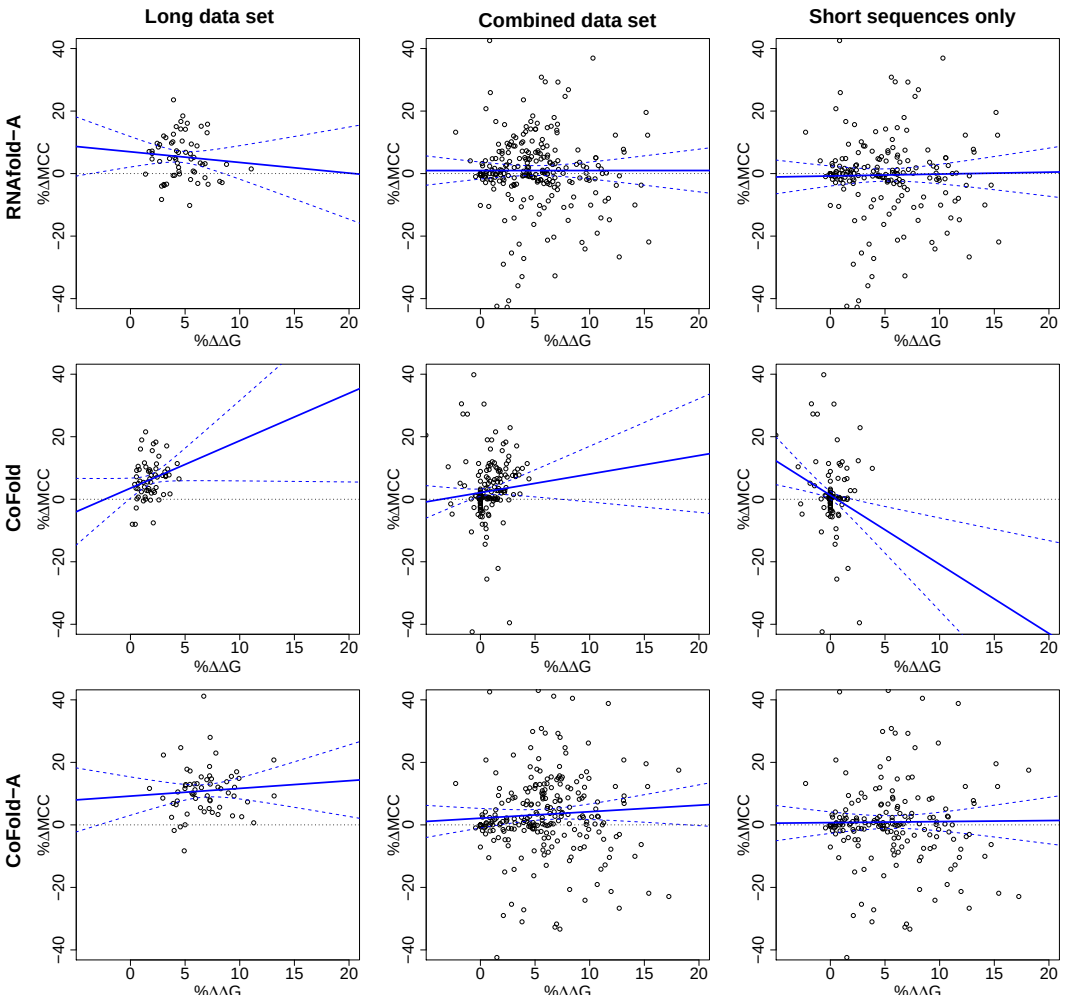
$\Rightarrow$ the short answer is . . . no !

Do large energy changes correlate with improved prediction accuracy?



$\Rightarrow$ the short answer is . . . no !

Do large energy changes correlate with improved prediction accuracy?



$\Rightarrow$ the short answer is . . . no !

Do large energy changes correlate with improved prediction accuracy?

Linear fit to Δ MCC versus % $\Delta\Delta$G distributions

	intercept $\pm$ stdev	slope $\pm$ stdev	R^2 (%)
	long data set ($>$ 1000 nt)		
RNAfold-A	7.0 ± 2.4	-0.34 ± 0.48	0.85
CoFold	3.5 ± 1.6	1.52 ± 0.78	6.06
CoFold-A	9.2 ± 3.1	0.25 ± 0.43	0.56
	combined data set		
RNAfold-A	1.0 ± 1.4	0.0008 ± 0.23	$5.6 \cdot 10^{-06}$
CoFold	2.1 ± 0.6	0.59 ± 0.47	0.64
CoFold-A	2.1 ± 1.6	0.21 ± 0.23	0.34
	short sequences only ($\leq$ 1000 nt)		
RNAfold-A	-0.8 ± 1.6	0.06 ± 0.25	0.03
CoFold	1.3 ± 0.7	-2.21 ± 0.75	4.44
CoFold-A	0.7 ± 1.7	0.03 ± 0.25	0.01

$\Rightarrow$ the long answer is . . . also no (for sequences of all lengths) !

Do large energy changes correlate with improved prediction accuracy?

Linear fit to Δ MCC versus % $\Delta\Delta$G distributions

	intercept $\pm$ stdev	slope $\pm$ stdev	R^2 (%)
	long data set ($>$ 1000 nt)		
RNAfold-A	7.0 ± 2.4	-0.34 ± 0.48	0.85
CoFold	3.5 ± 1.6	1.52 ± 0.78	6.06
CoFold-A	9.2 ± 3.1	0.25 ± 0.43	0.56
	combined data set		
RNAfold-A	1.0 ± 1.4	0.0008 ± 0.23	$5.6 \cdot 10^{-06}$
CoFold	2.1 ± 0.6	0.59 ± 0.47	0.64
CoFold-A	2.1 ± 1.6	0.21 ± 0.23	0.34
	short sequences only ($\leq$ 1000 nt)		
RNAfold-A	-0.8 ± 1.6	0.06 ± 0.25	0.03
CoFold	1.3 ± 0.7	-2.21 ± 0.75	4.44
CoFold-A	0.7 ± 1.7	0.03 ± 0.25	0.01

$\Rightarrow$ the long answer is ... also no (for sequences of all lengths) !

Do large energy changes correlate with improved prediction accuracy?

Linear fit to Δ MCC versus % $\Delta\Delta$G distributions

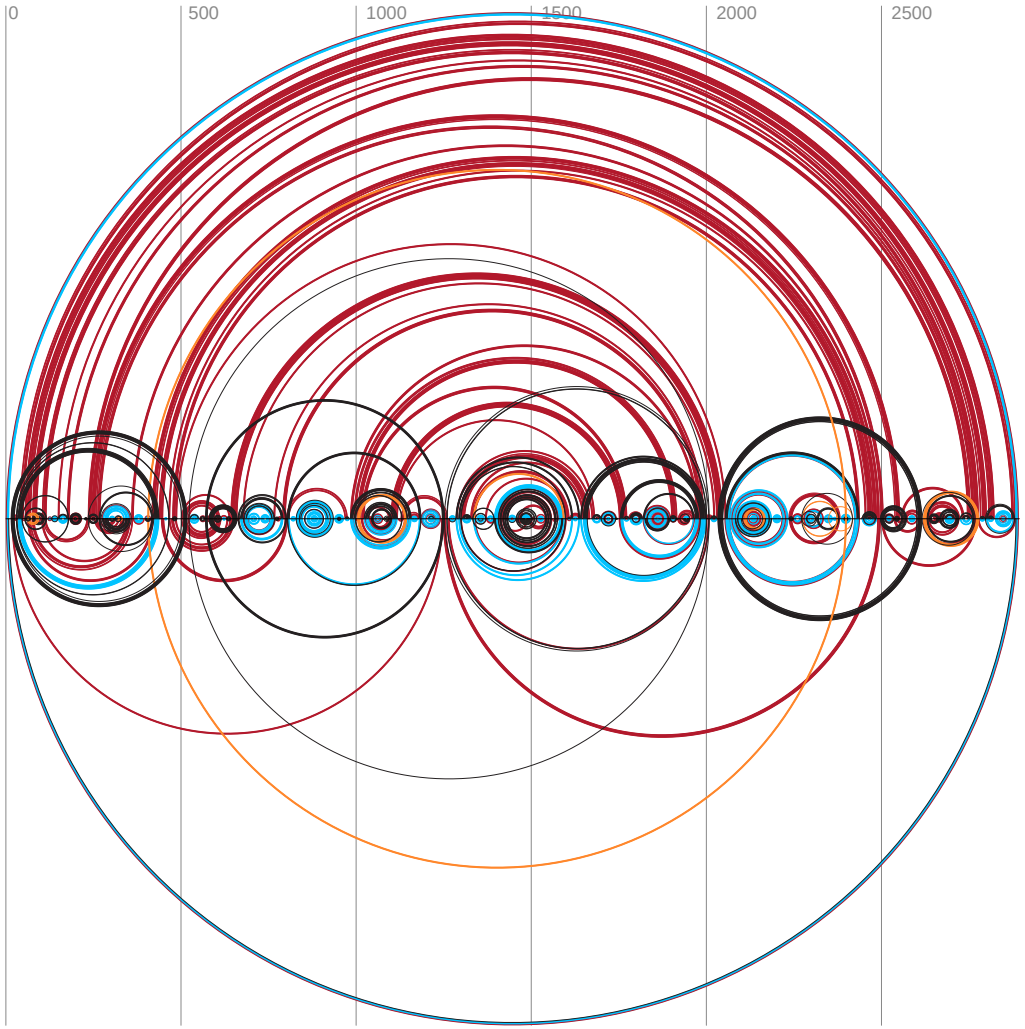
	intercept $\pm$ stdev	slope $\pm$ stdev	R^2 (%)
	long data set ($>$ 1000 nt)		
RNAfold-A	7.0 ± 2.4	-0.34 ± 0.48	0.85
CoFold	3.5 ± 1.6	1.52 ± 0.78	6.06
CoFold-A	9.2 ± 3.1	0.25 ± 0.43	0.56
	combined data set		
RNAfold-A	1.0 ± 1.4	0.0008 ± 0.23	$5.6 \cdot 10^{-06}$
CoFold	2.1 ± 0.6	0.59 ± 0.47	0.64
CoFold-A	2.1 ± 1.6	0.21 ± 0.23	0.34
	short sequences only ($\leq$ 1000 nt)		
RNAfold-A	-0.8 ± 1.6	0.06 ± 0.25	0.03
CoFold	1.3 ± 0.7	-2.21 ± 0.75	4.44
CoFold-A	0.7 ± 1.7	0.03 ± 0.25	0.01

$\Rightarrow$ the long answer is ... also no (for sequences of all lengths) !

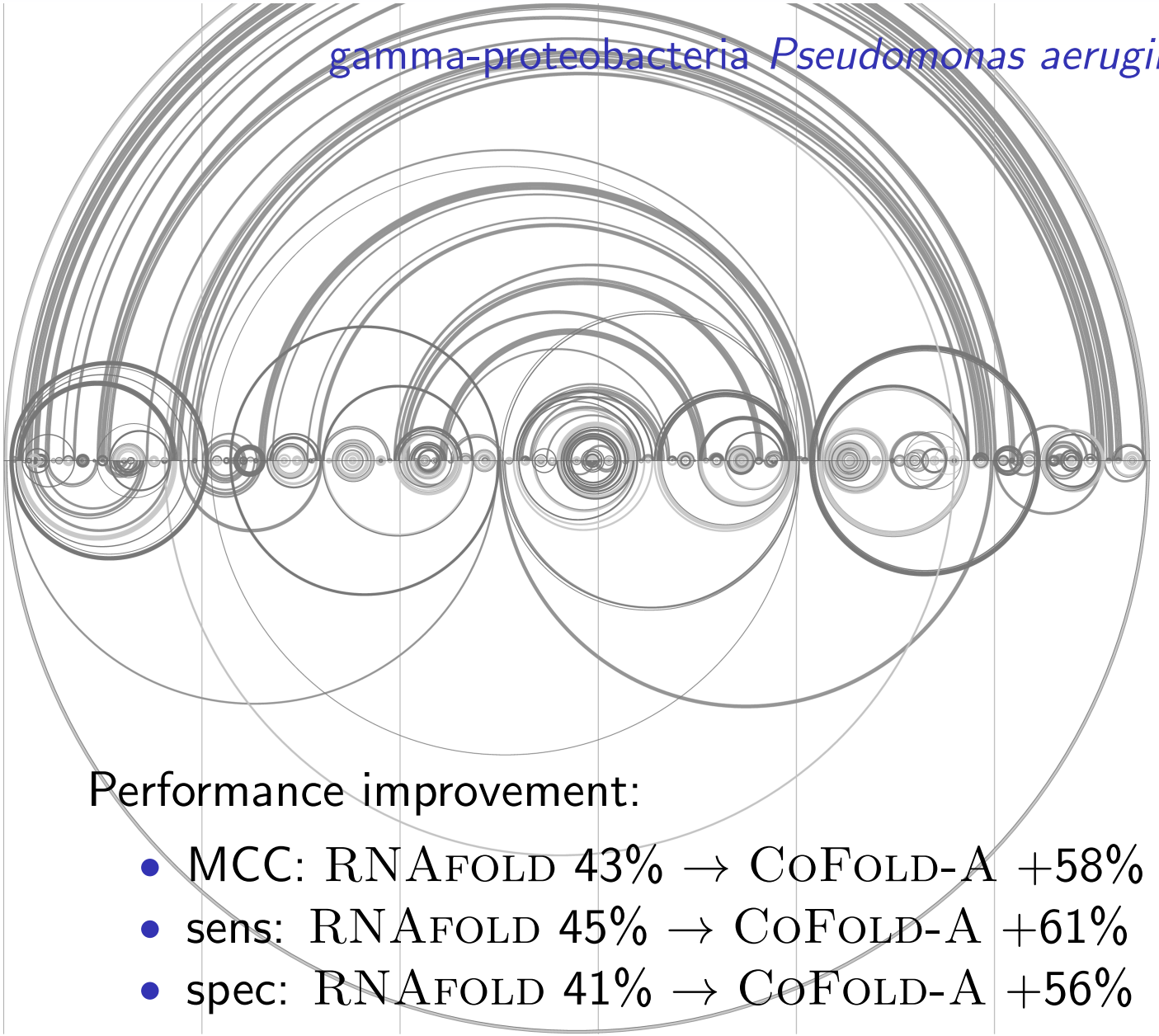
CoFold: 23S rRNAs

- average seq. length 3069 nt (min 2882 nt, max 3578 nt)
- MCC: RNAfold $\rightarrow$ CoFold +8%
- MCC: RNAfold $\rightarrow$ CoFold-A +12%

RNAfold versus CoFold-A predictions for the 23S rRNA of the gamma-proteobacteria *Pseudomonas aeruginosa*



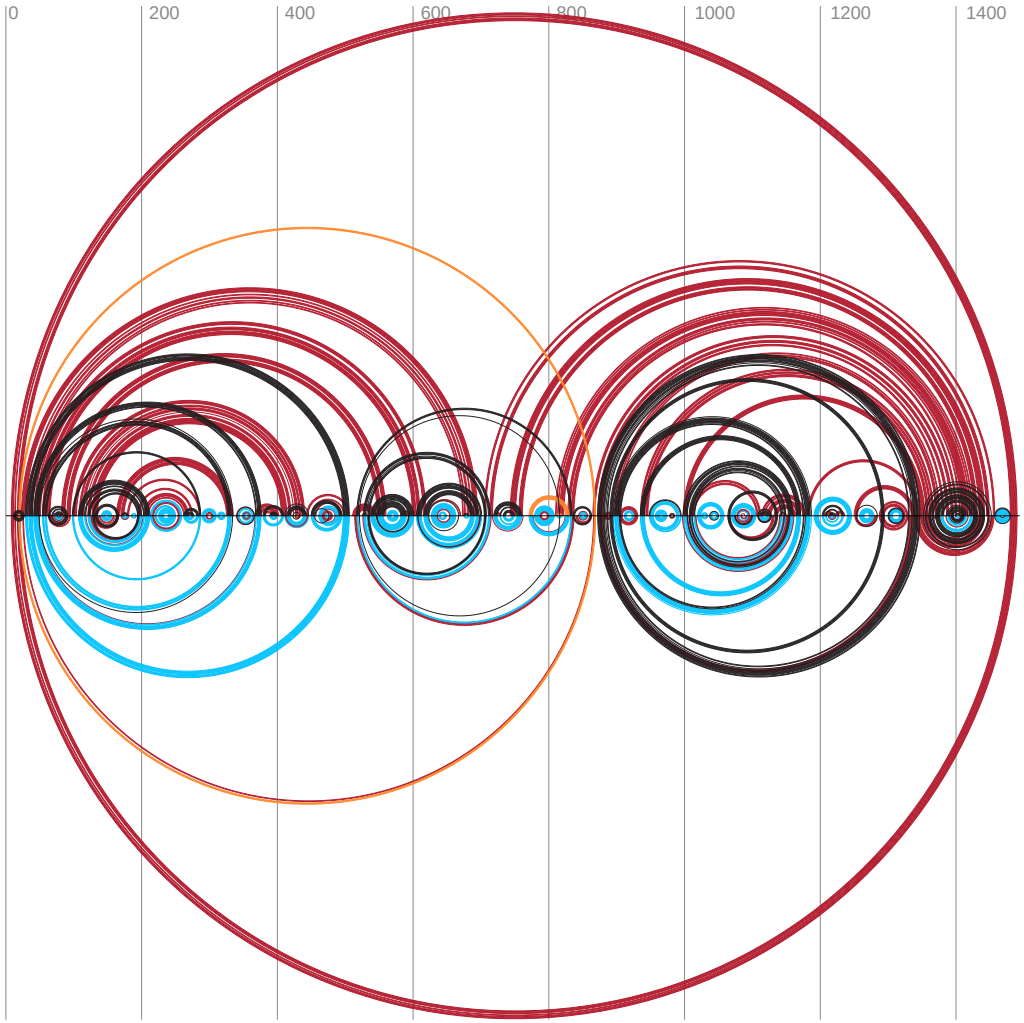
RNAfold versus CoFold-A predictions for the 23S rRNA of the gamma-proteobacteria *Pseudomonas aeruginosa*



Performance improvement:

- MCC: RNAfold 43% → CoFold-A +58% (+15%)
- sens: RNAfold 45% → CoFold-A +61%
- spec: RNAfold 41% → CoFold-A +56%

RNAfold versus CoFold-A predictions for the 16S rRNA of the freshwater algae *Cryptomonas sp.*



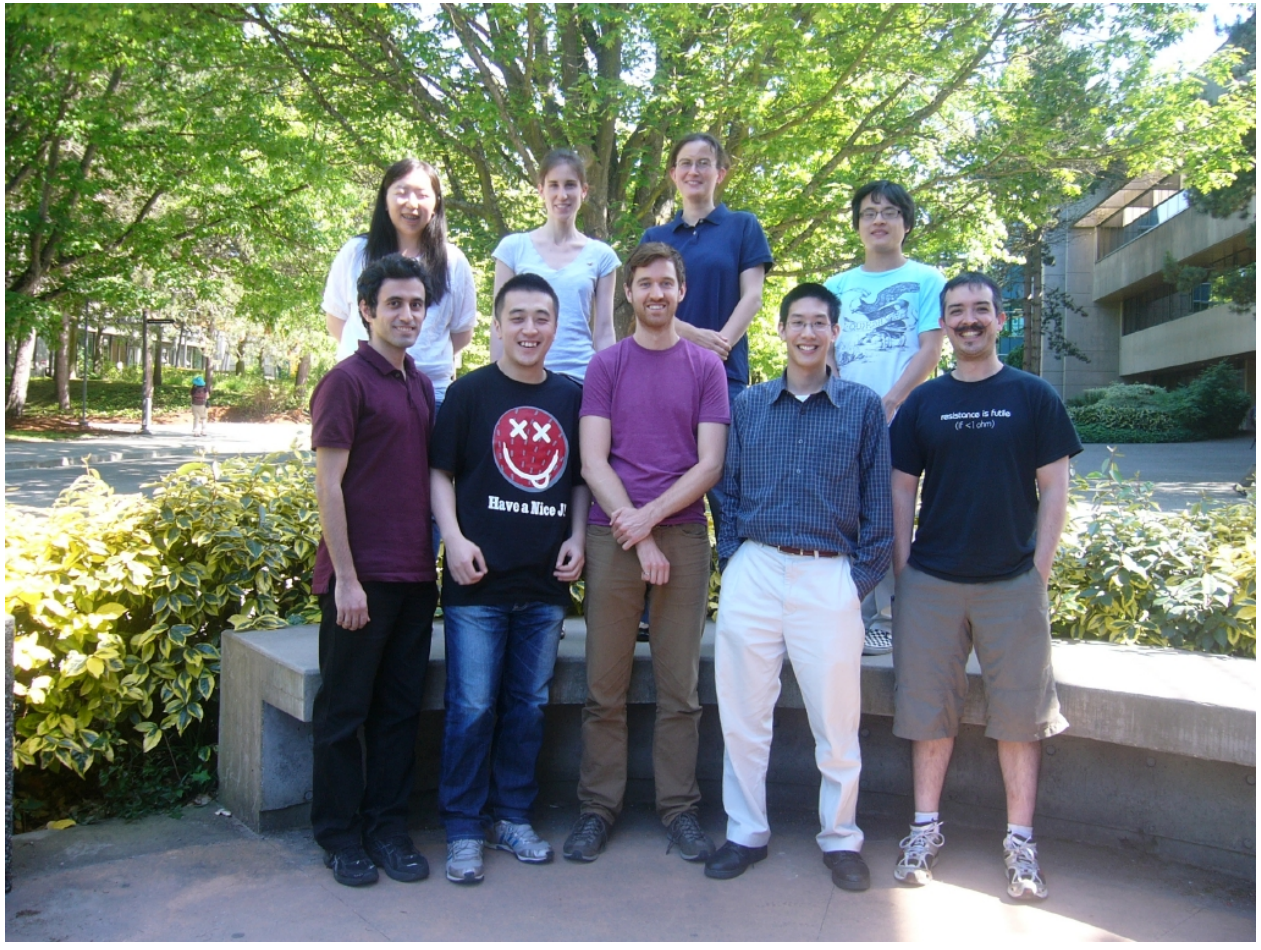
CoFold: summary

+ depends on only 1 free parameter (rather than 363)
+ parameter training is robust
+ compiled non-redundant data set of long sequences
+ improves the prediction accuracy, esp. for long sequences ...
+ ... and also for short sequences, but not as much
  (CoFold and CoFold-A outperform RNAfold and RNAfold-A)
+ free energies of structures hardly changed
+ same memory and time complexity as RNAfold

Key features:

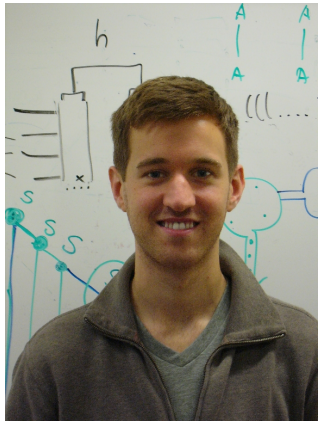
- captures first aspects of kinetic structure formation
- algorithm combines thermodynamic and kinetic considerations
- future: capture more aspects of folding process

Acknowledgements:



My group (including Evan who took the photo).

Acknowledgements:

Jeff Proctor 

- CoFold: submitted, manuscript `arxiv.org/abs/1207.6013`
- CoFold: web-server at `www.e-rna.org/cofold`
- 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