Moritz Kiekeben

Freie Universität Berlin

Seminar presentation

RNAalifold: improved consensus structure prediction for RNA alignments

Stephan H Bernhart et al., 2008, BMC Bioinformatics

Introduction

- About 90% of mammalian genome is transcribed
- < 2% is protein-coding
- Most of the rest is probably functional RNA (tRNA, microRNA, snRNA etc.)

  → Huge amount of RNA molecules to analyse

Introduction

- Most functional RNAs have characteristic secondary structures

  → Highly conserved during evolution

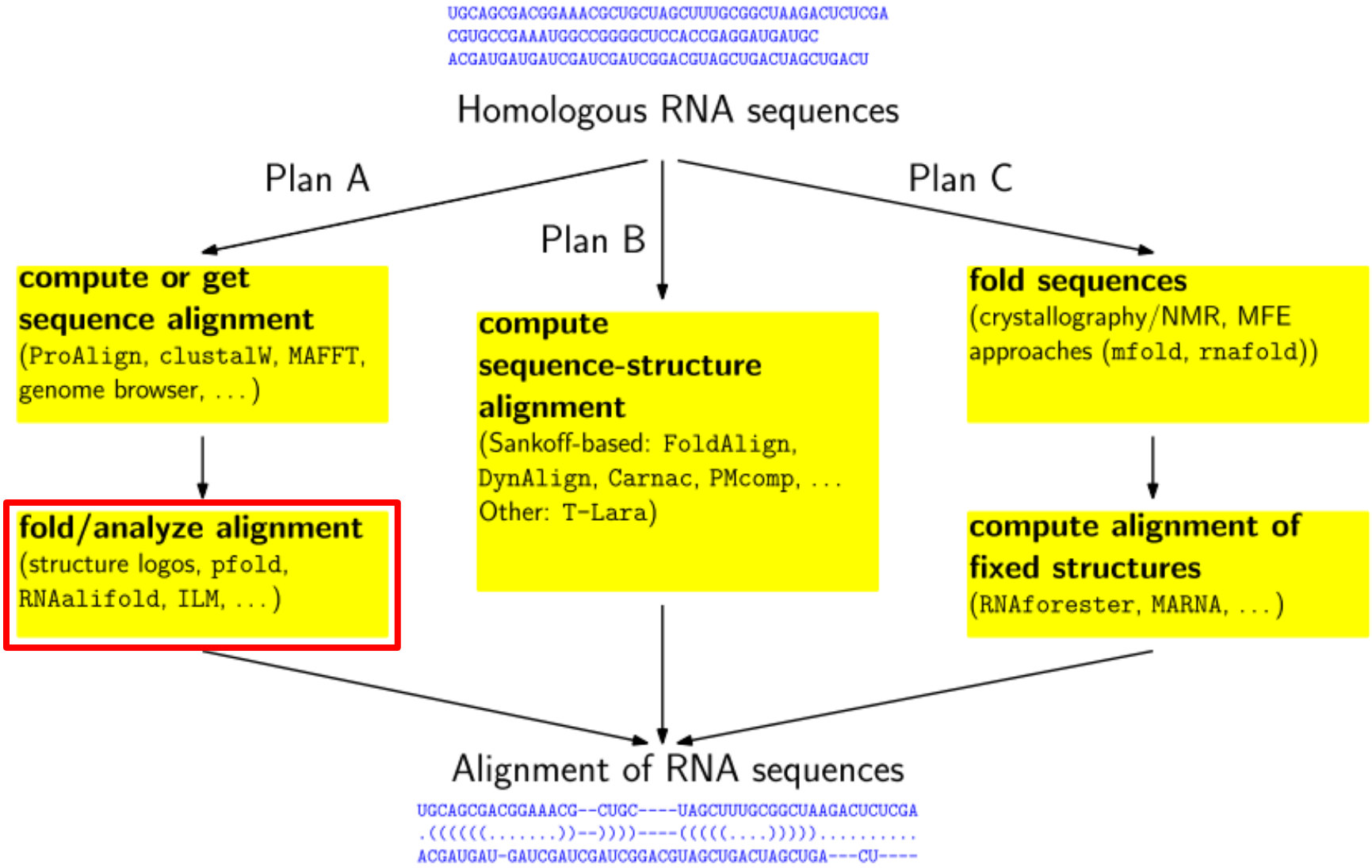
- „Similiar structure“ possibly leads to „similar function“
- Structure is higher conserved than underlying sequences

  → RNA sequences vary between species but structure usually does not

Task: Given a sequence, find the optimal secondary structures or

Given a MSA, find the consensus structure.

Introduction



source: http://www.mi.fu-berlin.de/wiki/pub/ABI/SS14Lecture11Materials/script.pdf, SeqAna RNA script

Introduction

RNAalifold is a DP algorithm, that uses phylogenetic information (sequence covariance) and thermodynamic stability of molecules to predict a consensus structure of a MSA.

Idea behind sequence covariance

example:

seq1	G	**C**	C	U	U	C	C	C	**G**	C
seq2	G	**A**	C	U	U	C	C	C	**U**	C
seq3	G	**G**	C	U	U	C	C	C	**C**	C

Bold printed basepairs are covarying (compensatory mutations).

→ base pair most likely present in consensus structure

Methods

The old RNAalifold

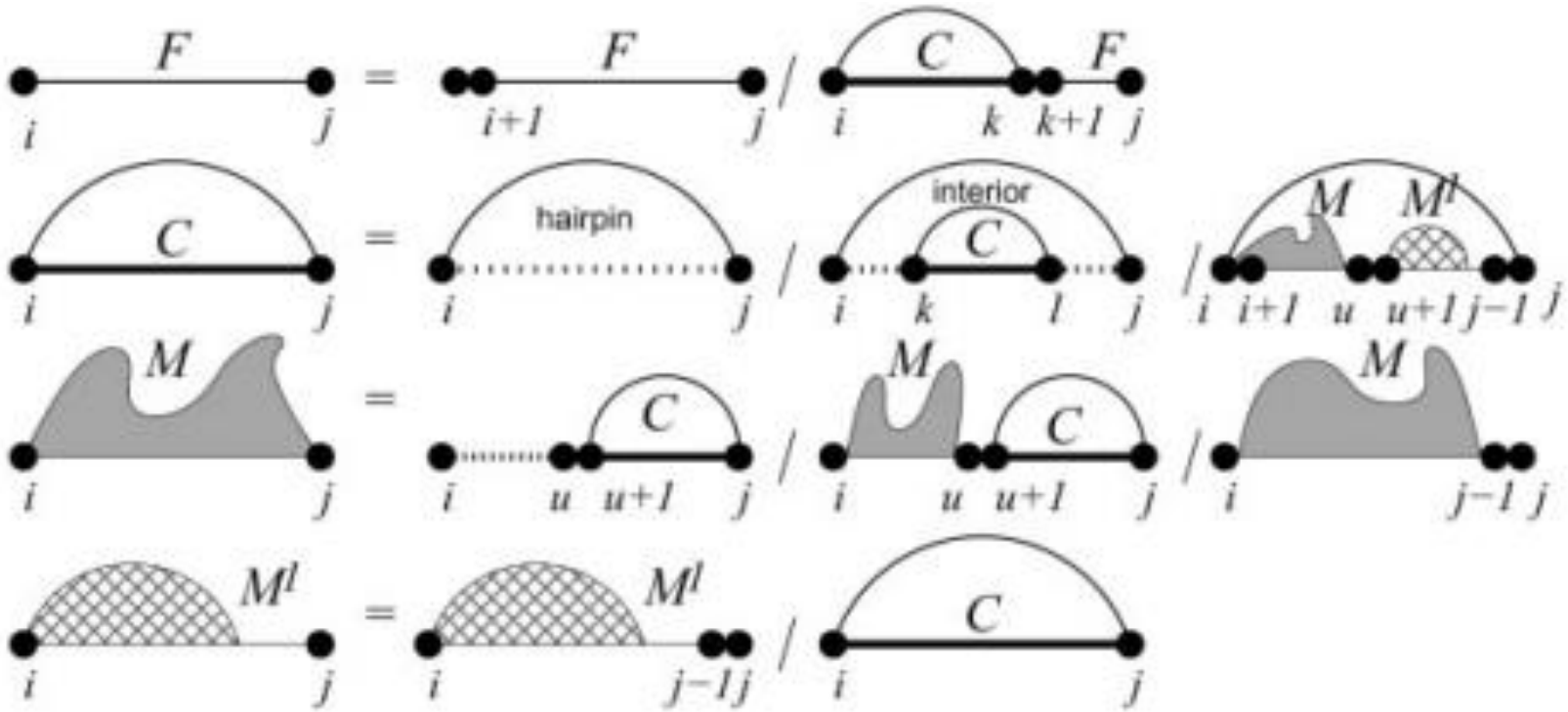
- 4 matrices corresponding to different structural components:
  - F (unconstrained structures)
  - C (constrained structures)
  - M (multi-loop structures)
  - M^1(multi-loop with one branch)
- They hold for every sub-sequence from i to j the optimal folds
- $\gamma(i,j)$ is the covariance score

$$
\begin{aligned}
F_{i,j} &= \min\left(F_{i+1,j}, \min_{i<k\leq j} C_{i,k} + F_{k+1,j} \right) \\
C_{i,j} &= \beta\gamma(i,j) + \\
&+ \min\begin{cases} \sum_{\alpha\in\mathbb{A}} \mathfrak{H}(i,j,\alpha) \\ \min_{i<k<l<j}\left(\sum_{\alpha\in\mathbb{A}} \mathfrak{J}(ij,kl,\alpha) + C_{k,l} \right) \\ \min_{i<k<j}\left(M_{i,k} + M^1_{k+1,j} + \mathfrak{a} \right) \end{cases} \\
M_{i,j} &= \min\begin{cases} M_{i+1,j} + \mathfrak{c} \\ \min_{i<k<j} C_{i,k} + M_{k+1,j} + \mathfrak{b} \\ M^1_{i,j} \end{cases} \\
M^1_{i,j} &= \min\left(M^1_{i,j-1} + \mathfrak{c}, C_{i,k} \right)
\end{aligned}
$$

source: Bernhart et al., 2008, BMC Bioinformatics

Methods

„Visualization“ of the recursions



source: Hofacker/Stadler, 2006

Methods

Improvement of the old RNAalifold

old covariance score:

$$\gamma'(i,j) = \frac{1}{2} \sum_{\substack{\alpha,\beta \in \mathbb{A} \\ \alpha \neq \beta}} \begin{cases} h(\alpha_i,\beta_i) + h(\alpha_j,\beta_j) & \text{if } (\alpha_i,\alpha_j) \in \mathcal{B} \\ & \wedge (\beta_i,\beta_j) \in \mathcal{B} \\ 0 & \text{otherwise} \end{cases}$$

$$\mathcal{B} = \{AU, UA, CG, GC, GU, UG\}$$

- Based on hamming distance h(a, b)
- **Problem**: No quantitative argument, since h(a,b) is either 1 or 0. Some mutations might be more frequent than other...

Methods

Improvement of the old RNAalifold

- **Solution**: Introducing a scoring matrix (RIBOSUM)
- Scores derived from frequencies of aligned and basepaired nucleotides (log odds)
- example: RIBOSUM85-60

	AA	AC	AG	AU	CA	CC	CG	CU	GA	GC	GG	GU	UA	UC	UG	UU
AA	-2.49															
AC	-7.04	-2.11														
AG	-8.24	-8.89	-0.80													
AU	-4.32	-2.04	-5.13	4.49												
CA	-8.84	-9.37	-10.41	-5.56	-5.13											
CC	-14.37	-9.08	-14.53	-6.71	-10.45	-3.59										
CG	-4.68	-5.86	-4.57	1.67	-3.57	-5.71	5.36									
CU	-12.64	-10.45	-10.14	-5.17	-8.49	-5.77	-4.96	-2.28								
GA	-6.86	-9.73	-8.61	-5.33	-7.98	-12.43	-6.00	-7.71	-1.05							
GC	-5.03	-3.81	-5.77	2.70	-5.95	-3.70	2.11	-5.84	-4.88	5.62						
GG	-8.39	-11.05	-5.38	-5.61	-11.36	-12.58	-4.66	-13.69	-8.67	-4.13	-1.98					
GU	-5.84	-4.72	-6.60	0.59	-7.93	-7.88	-0.27	-5.61	-6.10	1.21	-5.77	3.47				
UA	-4.01	-5.33	-5.43	1.61	-2.42	-6.88	2.75	-4.72	-5.85	1.60	-5.75	-0.57	4.97			
UC	-11.32	-8.67	-8.87	-4.81	-7.08	-7.40	-4.91	-3.83	-6.63	-4.49	-12.01	-5.30	-2.98	-3.21		
UG	-6.16	-6.93	-5.94	-0.51	-5.63	-8.41	1.32	-7.36	-7.55	-0.08	-4.27	-2.09	1.14	-4.76	3.36	
UU	-9.05	-7.83	-11.07	-2.98	-8.39	-5.41	-3.67	-5.21	-11.54	-3.90	-10.79	-4.45	-3.39	-5.97	-4.28	-0.02

source: Klein/Eddy, 2003, BMC Bioinformatics

Methods

Improvement of the old RNAalifold

- **new covariance score**: $$\gamma'(i,j) = \frac{1}{2} \sum_{\substack{\alpha,\beta \in A \\ \alpha \neq \beta}} xR(\alpha_i \alpha_j ; \beta_i \beta_j)$$

- Replacement of the hamming distance with the RIBOSUM scores

- *Complete covariance score used in recursions:*

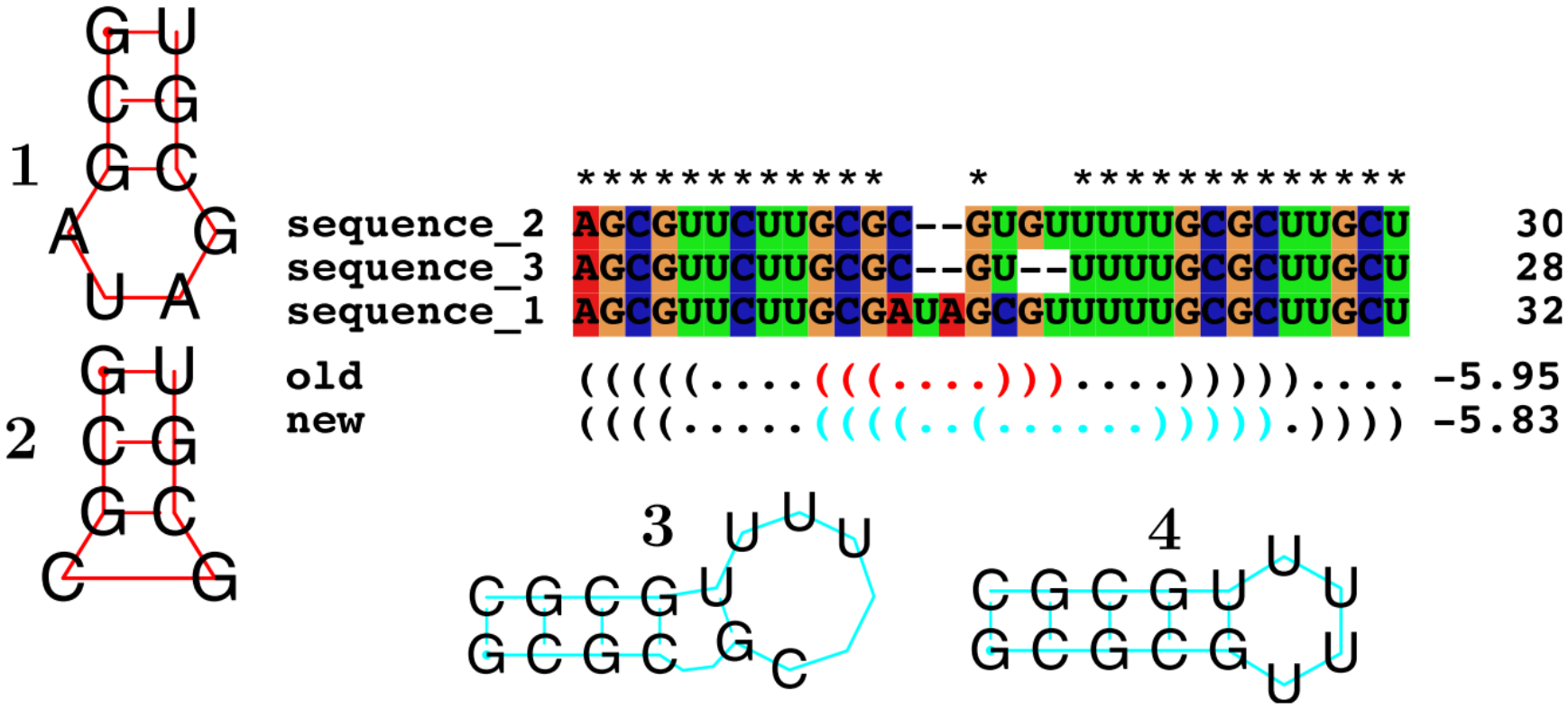
$$\beta \cdot \gamma(i,j) = \beta \cdot \left[\gamma'(i,j) + \delta \sum_{\alpha \in A} \begin{cases} 0, if(\alpha_i \alpha_j) \in B \\ 0.25, if\ \alpha_i and\ \alpha_j are\ gaps \\ 1, otherwise \end{cases} \right]$$

- Parameter β – influence of covariance on folding
- Parameter δ – impact of non-standard base pairs

Methods

Improvement of the old RNAalifold

Removal of gaps in interior loops before calculating energies
→less energetic unfavorable loops



source: Bernhart et al., 2008, BMC Bioinformatics

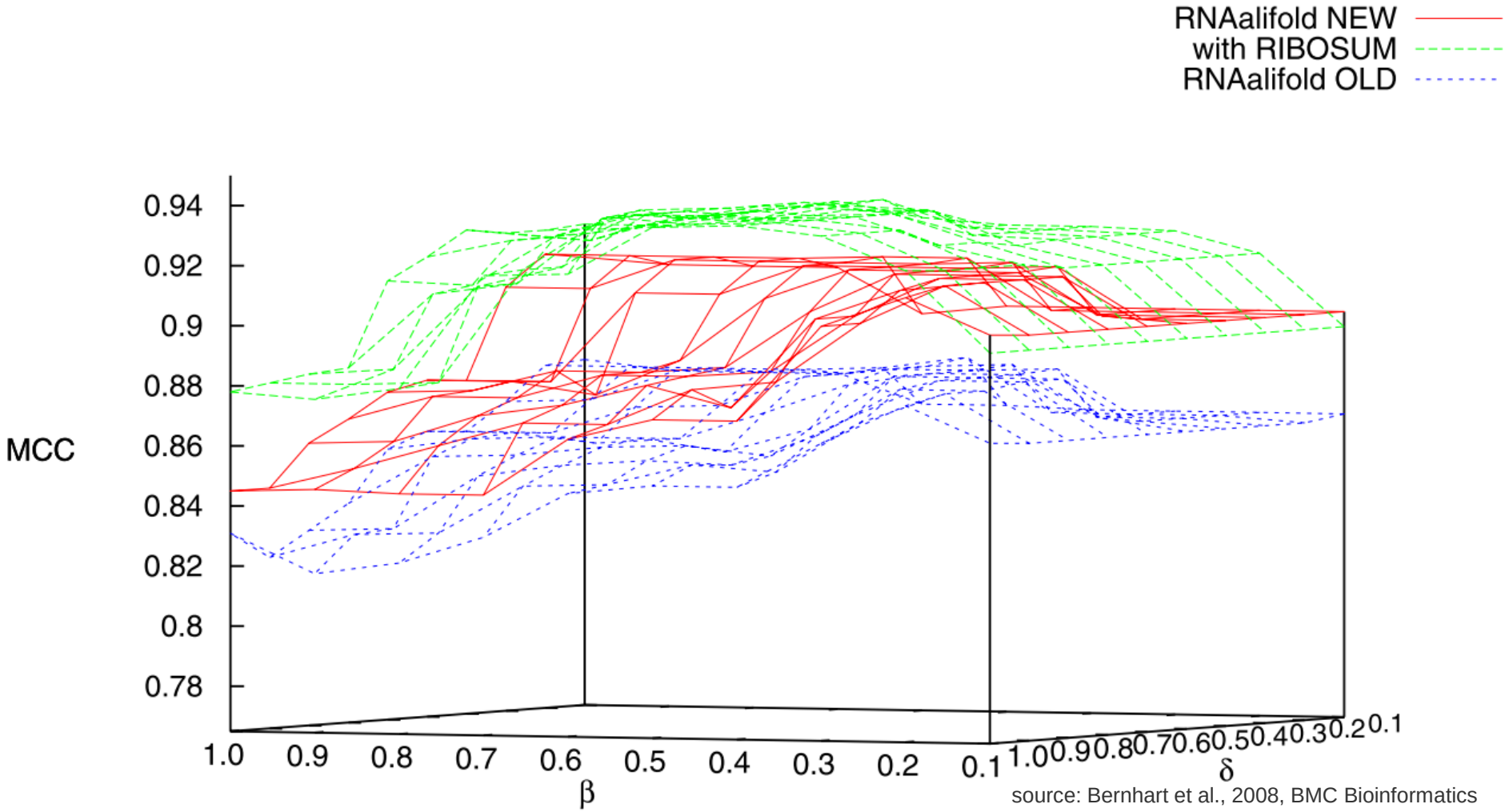
Results

Measuring accuracy

- CMfinder-SARSE dataset for reference structures
- Mathews correlation coefficient (MCC) for accuracy measuring
  - Basically a value to compare reference structure with predicted structure
    - MCC = 1, perfect prediction
    - MCC = 0, not better than random
    - MCC = -1, total disagreement

Results

Influence of β and δ on accuracy



source: Bernhart et al., 2008, BMC Bioinformatics

Results

Measuring accuracy

- New RNAalifold RIBOSUM better than other two (except for $\beta = 1$)
- Determination of optimal values for β and δ (0.6 and 0.5, respectively)
- Old RNAalifold emphazised covariance too much

Results

Comparision to other Folding algorithms

RNA	#seq	MPI	RIBOSUM	RNAalifold	Pfold	KNetFold	McC_mea
Antizyme_FSE	13	87	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
ctRNA_pGA1	15	72	**1.000**	**1.000**	**1.000**	0.976	**1.000**
Entero_5_CRE	160	84	**1.000**	0.848	0.478	**1.000**	0.942
Entero_CRE	56	81	**1.000**	0.736	**1.000**	0.953	0.953
GcvB	17	64	**0.939**	0.799	0.889	**0.939**	0.921
glmS	11	60	**0.986**	0.972	0.972	0.809	0.837
HACA_sno_Snake	22	90	0.871	0.407	0.414	**0.915**	0.884
HCV_SLIV	110	89	**1.000**	0.922	**1.000**	**1.000**	0.961
HDV_ribozyme	15	95	**0.953**	-0.015	0.590	0.460	0.460
HepC_CRE	52	87	**1.000**	0.962	**1.000**	**1.000**	**1.000**
Histone3	64	78	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
Hsp90_CRE	4	98	0.855	0.855	0.413	0.867	**0.874**
IBV_D-RNA	10	96	**1.000**	0.928	0.928	**1.000**	**1.000**
Intron_gpII	114	54	**1.000**	0.779	**1.000**	**1.000**	**1.000**
[.....							.....]
SNORD64	3	94	**1.000**	0.539	0.539	0.661	-0.014
SNORD86	6	82	**0.641**	-0.012	-0.007	0.511	0.000
snoU83B	4	87	**0.927**	**0.927**	0.846	0.895	**0.927**
TCV_H5	3	97	**1.000**	**1.000**	0.685	**1.000**	**1.000**
TCV_Pr	4	95	**1.000**	**1.000**	0.688	**1.000**	**1.000**
Tymo_tRNA-like	28	64	**1.000**	0.916	**1.000**	0.973	**1.000**
ykoK	36	61	0.856	0.756	**0.906**	0.841	0.794
mean			**0.937**	0.831	0.765	0.866	0.837

source: Bernhart et al., 2008, BMC Bioinformatics

Discussion

Comparision with other folding algorithms (RNA STRAND – Rfam set)

RNA	comment	RIBOSUM	RNAalifold	Pfold	KNetFold	McC_mea
7SK		**0.507**	0.456	0.292	0.429	0.306
bicoid_3		**0.949**	0.840	n.a.	0.829	0.927
Corona_pk3	Pk	0.579	0.646	0.674	0.678	**0.705**
CPEB3_ribozyme	Pk	**0.756**	**0.756**	0.663	**0.756**	0.612
Gammaretro_CES		**0.983**	0.948	**0.983**	0.935	**0.983**
Hammerhead_1		**1.000**	0.474	0.621	0.831	0.614
Hammerhead_3		**1.000**	0.960	**1.000**	**1.000**	**1.000**
HDV_ribozyme	Pk	0.709	-0.018	**0.784**	0.388	0.396
IRES_c-myc		-0.004	0.079	0.286	-0.002	**0.350**
[.....]						
Telomerase-vert	pk	**0.918**	0.751	n.a.	n.a.	0.820
Vimentin3		0.741	-0.016	0.184	**0.771**	0.629
Y		**1.000**	**1.000**	0.925	**1.000**	**1.000**
mean		0.759	0.651			0.703
mean	knetfold	0.750	0.645		0.680	0.696
mean	pfold	0.756	0.635	0.693	0.682	0.673

source: Bernhart et al., 2008, BMC Bioinformatics

Results

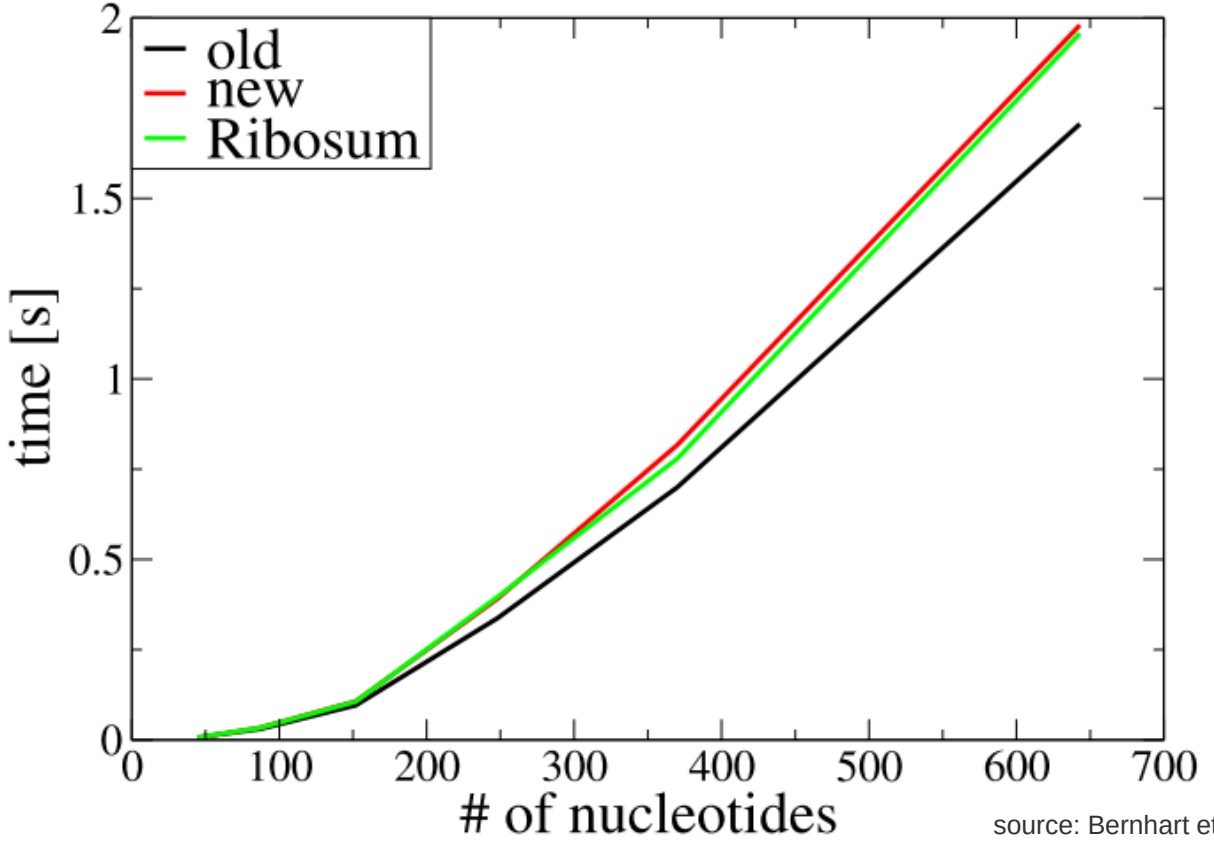
Computational performance

- Runtime: $O(Nn^3)$, N sequences of length n
- Memory: $O(n^2)$
- Tests run on an Intel Xeon 2.8 GHz
- Only runtime was tested

Results

Measuring runtime

- $N = 4$
- n variable

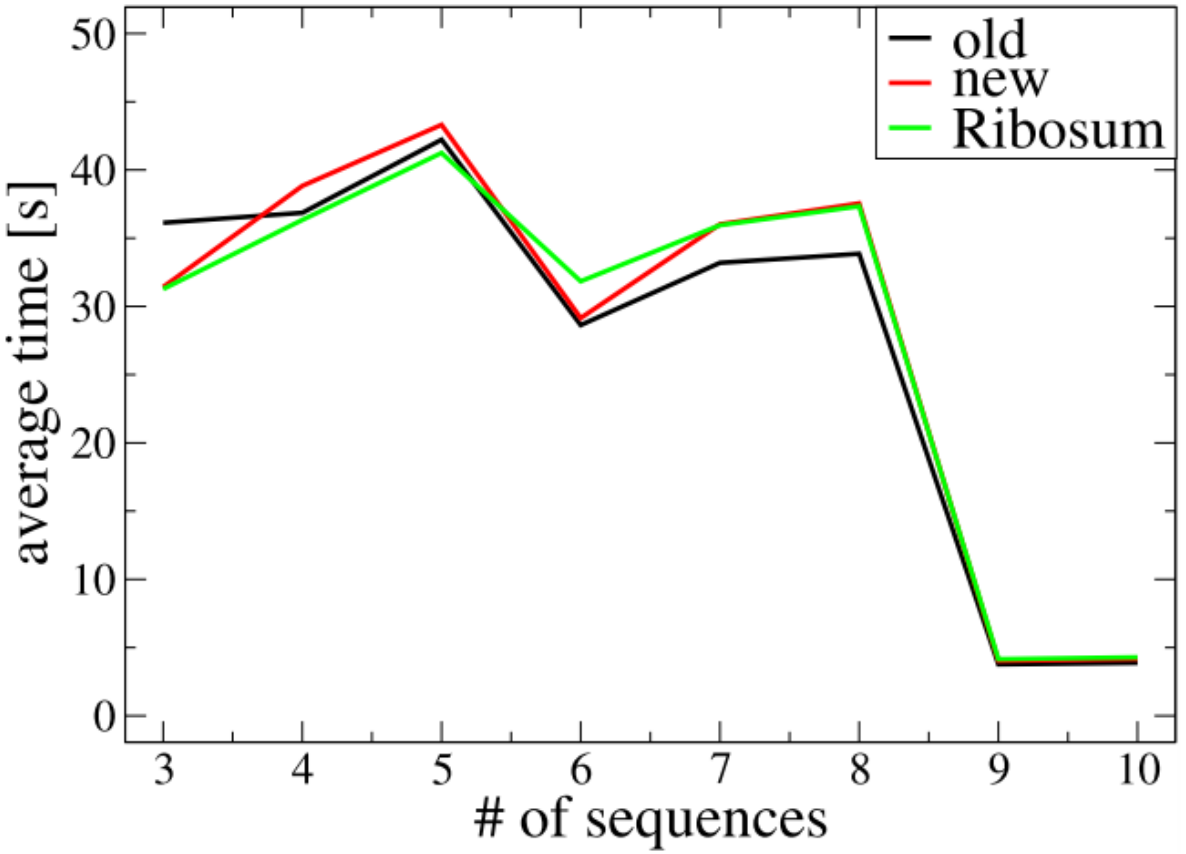


source: Bernhart et al., 2008, BMC Bioinformatics

Results

Measuring runtime

- $n = 1716$
- N variable
- All sequences from same alignment.



source: Bernhart et al., 2008, BMC Bioinformatics

Dropping basepairs that aren‘t supported by sequences (cutoff).

→The more sequences available the less basepairs have to be considered.

Fazit

- Changes to gaps and covariance led to improved accuracy
- Slightly more computational effort
- The new RNAalifold is on par or better than comparitive prediction software (at least on tested data sets)
- No pseudoknots (DP algorithms cannot detect pk)

Thanks for your attention!