

PETfold:

About RNAs which do not fold into the lowest free energy

Stefan E Seemann

Bled (Slovenia), 19th February 2008

Outline

- ❖ Motivation
- ❖ PETfold – Combined Model
- ❖ Pfold Model
- ❖ Parameter tuning
- ❖ Performance
- ❖ Conclusion
- ❖ Further work
- ❖ PETfold v1.0

Motivation
PETfold – Combined Model
Pfold Model
Parameter tuning
Performance
Conclusion
Further work
PETfold v1.0

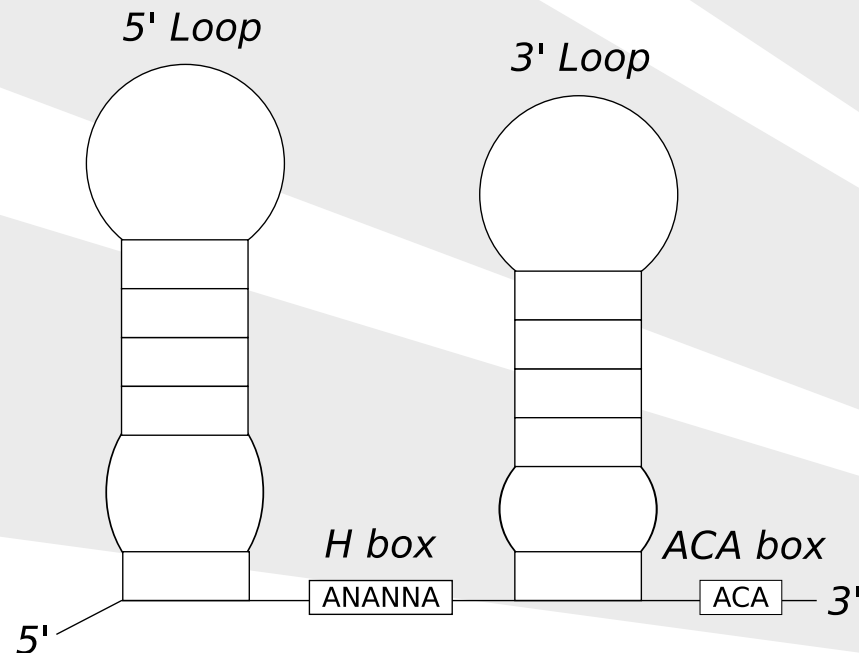
Motivation

❖ Motivation

- ❖ PETfold – Combined Model
- ❖ Pfold Model
- ❖ Parameter tuning
- ❖ Performance
- ❖ Conclusion
- ❖ Further work
- ❖ PETfold v1.0

H/ACA box small nucleolar RNA

- class of 'housekeeping' non-coding RNAs
- guides chemical modifications of RNA genes (methylation or pseudouridylation)
- mostly target ribosomal RNAs and spliceosomal RNAs

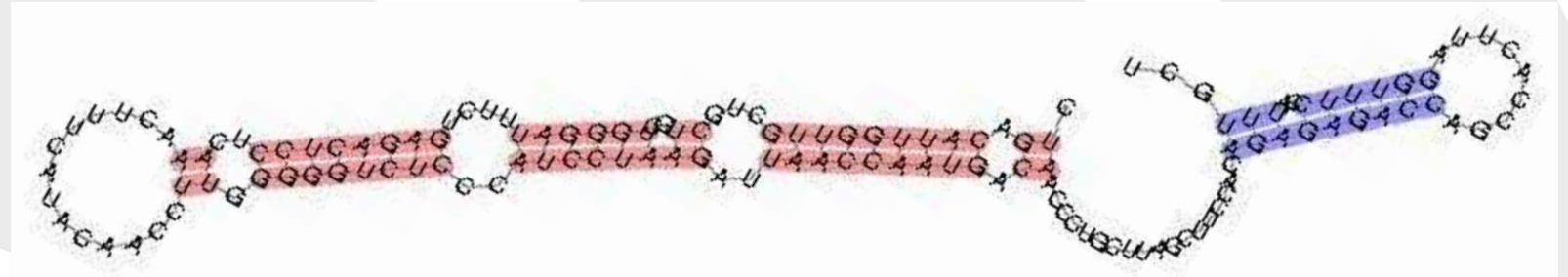


Snake H/ACA box small nucleolar RNA

❖ Motivation

- ❖ PETfold – Combined Model
- ❖ Pfold Model
- ❖ Parameter tuning
- ❖ Performance
- ❖ Conclusion
- ❖ Further work
- ❖ PETfold v1.0

- Rfam family HACA_Sno_Snake (RF00098)
→ 22 sequences, 86.85% MPI



Pfold-.....-.....(.....((.....)).....).....
 RNAalifold(((.....(((.....((((.....(((.....(((.....)))).....)))).....)))).....
 PETfold(.....((((.....)).....((((.....((((.....((-.....-.....))----)).....)))).....)).....
 Rfam((((((((.....(((.....((((.....((((((((.....(((.....)))).....)))).....)))).....)).....

Pfold-----..((((.....))))..... 0.25 MCC
 RNAalifold)))).....))..... 0.41 MCC
 PETfold-----..((((.....))))..... 0.79 MCC
 Rfam((((.....)))).....

PETfold – Combined Model

- ❖ Motivation
- ❖ **PETfold – Combined Model**
- ❖ Pfold Model
- ❖ Parameter tuning
- ❖ Performance
- ❖ Conclusion
- ❖ Further work
- ❖ PETfold v1.0

Former methods combining the duality of energy minimization and evolutionary conservation:

- FOLDALIGN, Dynalign, PMcomp and LocARNA using Sankoff algorithm
- Stemloc and Consan are SCFG based approaches without energy model
- CMfinder combines implicitly energy contribution with SCFG

PETfold– combined model:

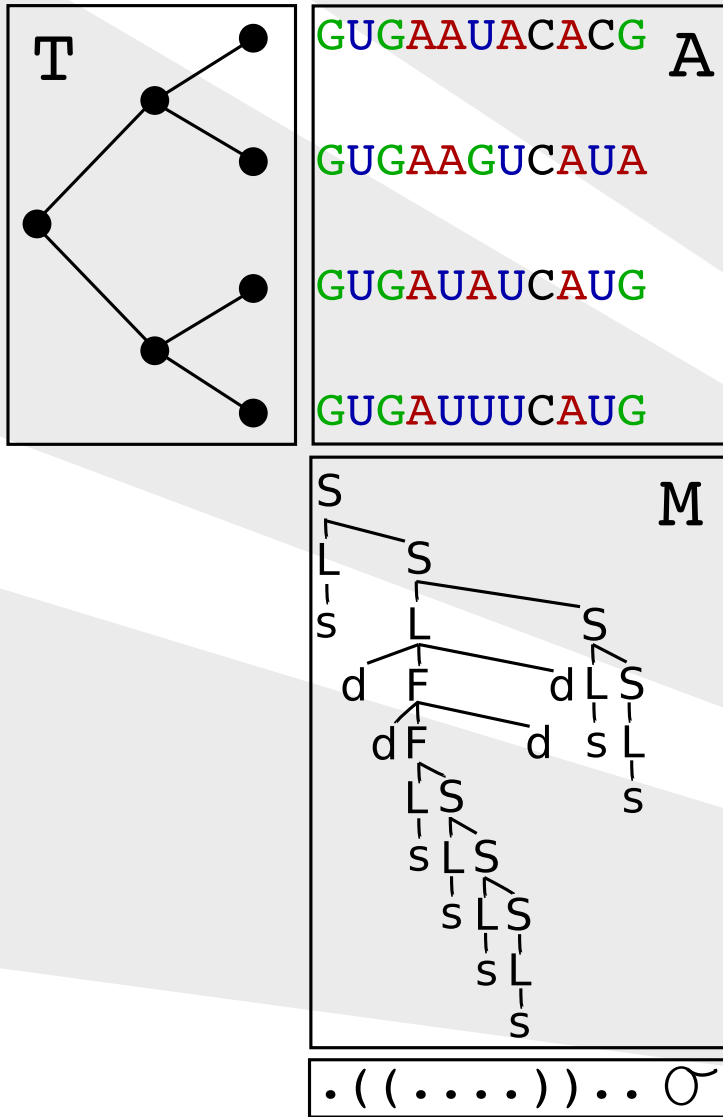
- **Probabilistic Evolutionary and Thermodynamic folding**
- Functional boxes are sequential & structural conserved
- Natural unselected subsequence folds in energetic stable structure

Pfold Model

- ❖ Motivation
- ❖ PETfold – Combined Model

❖ Pfold Model

- ❖ Parameter tuning
- ❖ Performance
- ❖ Conclusion
- ❖ Further work
- ❖ PETfold v1.0



$$\begin{aligned} & \Pr_{\tau_M(\sigma)}(\mathbf{r}(\sigma), A) \\ &= P[A|T, \sigma, M]P[\sigma|T, M] \end{aligned}$$

PETfold – Combined Model

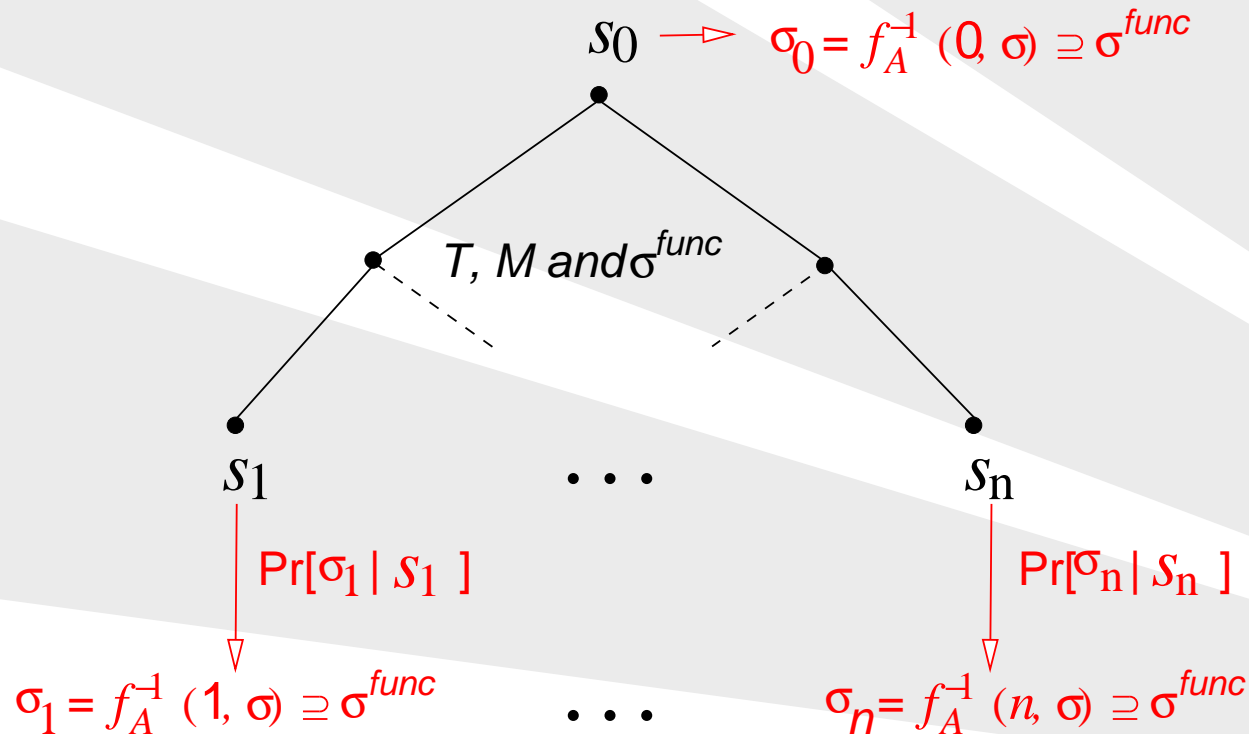
- ❖ Motivation
- ❖ PETfold – Combined Model

❖ Pfold Model

- ❖ Parameter tuning
- ❖ Performance
- ❖ Conclusion
- ❖ Further work
- ❖ PETfold v1.0

Pfold model: Combined SCFG multiplying rule probs with phylogenetic tree probs

PETfold extension of folding energy probs and structural constraints



PETfold – Combined Model

- ❖ Motivation
- ❖ PETfold – Combined Model

❖ Pfold Model

- ❖ Parameter tuning
- ❖ Performance
- ❖ Conclusion
- ❖ Further work
- ❖ PETfold v1.0

Maximum Expected Accuracy

Search consensus structure with maximal expected overlap.

Reliability score in Pfold:

$$\mathcal{R}_{A,T,M}(i, j) = \sum_{(i,j) \in \sigma} \Pr_{\tau_M(\sigma)}(\mathbf{r}(\sigma), A)$$

$\Pr_{\tau_M(\sigma)}(\mathbf{r}(\sigma), A)$... probability distribution on consensus structures

Combined evolutionary and thermodynamic model:

$$\begin{aligned} \text{ex-over}(\sigma) = & \sum_{(i,j) \in \sigma} \left(\mathcal{R}_{A,T,M}(i, j) + \frac{\beta}{n} \sum_u p_{f_A^{-1}}^{s_u}(i, j) \right) \\ & + \sum_{i \in \text{sg}(\sigma)} \alpha \left(\mathcal{R}_{A,T,M}^{\text{sg}}(i) + \frac{\beta}{n} \sum_u q_{f_A^{-1}}^{s_u}(i) \right) \end{aligned}$$

PETfold – Combined Model

- ❖ Motivation
- ❖ PETfold – Combined Model

❖ Pfold Model

- ❖ Parameter tuning
- ❖ Performance
- ❖ Conclusion
- ❖ Further work
- ❖ PETfold v1.0

Free parameters:

- α ... weights single-stranded against base pair positions ($\alpha \leq 0.5$)
- β ... thermodynamic overlap ($\beta = 1$)
- Reliability threshold of conserved functional substructure
→ assumption: highly evolutionary reliable substructures are functional
- ❖ $p_{bp}^{threshold}$... minimal base paired reliability
- ❖ $p_{ss}^{threshold}$... minimal single-stranded reliability

Parameter tuning

- ❖ Motivation
- ❖ PETfold – Combined Model
- ❖ Pfold Model
- ❖ **Parameter tuning**
- ❖ Performance
- ❖ Conclusion
- ❖ Further work
- ❖ PETfold v1.0

CMfinder SARSE dataset

- Rfam seed alignments of 46 RNA families
→ 17 families used in former evaluations ¹ and
29 with high quality alignment according SARSE project²

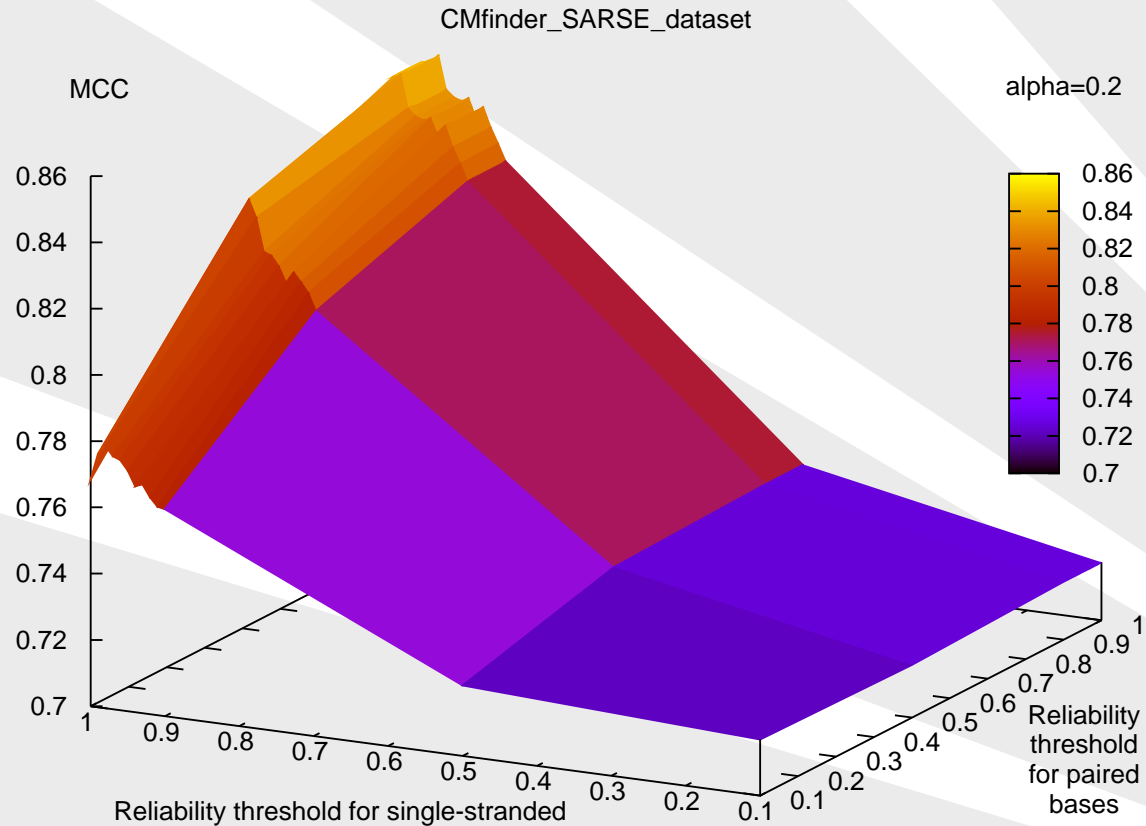
¹Z Yao et al., Bioinformatics 2006; E Torarinsson et al., Bioinf. 2007

²E S Andersen et al., RNA 2007

Parameter tuning

- ❖ Motivation
- ❖ PETfold – Combined Model
- ❖ Pfold Model
- ❖ **Parameter tuning**
- ❖ Performance
- ❖ Conclusion
- ❖ Further work
- ❖ PETfold v1.0

Mean MCCs of test data to Rfam structures

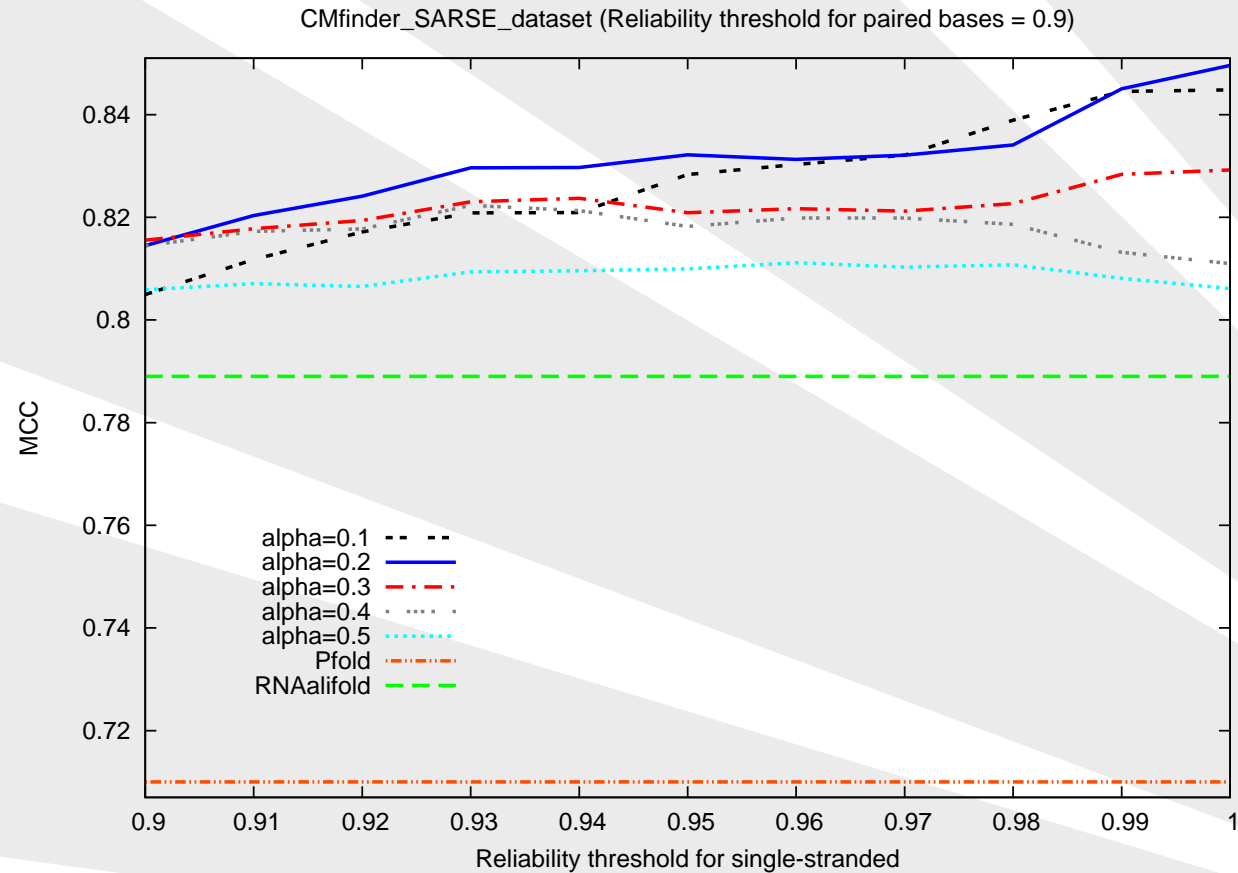


$$(\alpha = 0.2 \text{ and } 0.1 \leq p_{ss}^{threshold} \leq 1 \text{ and } 0.1 \leq p_{bp}^{threshold} \leq 1)$$

Parameter tuning

- ❖ Motivation
- ❖ PETfold – Combined Model
- ❖ Pfold Model
- ❖ **Parameter tuning**
- ❖ Performance
- ❖ Conclusion
- ❖ Further work
- ❖ PETfold v1.0

Mean MCCs of test data to Rfam structures



(several α and $0.9 \leq p_{ss}^{threshold} \leq 1$ and $p_{bp}^{threshold} = 0.9$)

Performance

- ❖ Motivation
- ❖ PETfold – Combined Model
- ❖ Pfold Model
- ❖ Parameter tuning
- ❖ **Performance**
- ❖ Conclusion
- ❖ Further work
- ❖ PETfold v1.0

	PPV ³	SEN ⁴	ACC ⁵	MCC⁶	R_5 ⁷
PETfold	0.852	0.876	0.864	0.850	0.722
Pfold	0.662	0.843	0.747	0.710	0.575
RNAalifold	0.758	0.842	0.799	0.789	0.652

³Positive predictive value

⁴sensitivity

⁵accuracy

⁶Matthews correlation coefficient

⁷ R_5 correlation coefficient

Conclusion

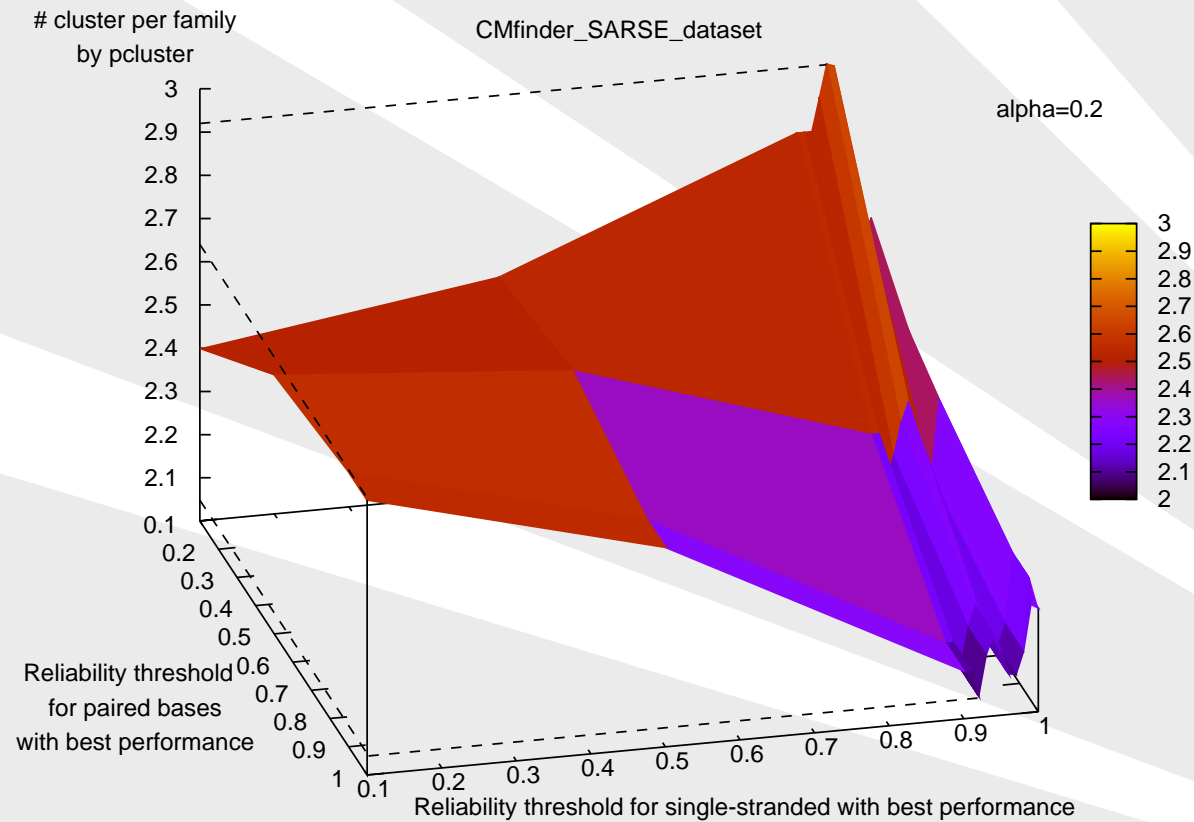
- ❖ Motivation
- ❖ PETfold – Combined Model
- ❖ Pfold Model
- ❖ Parameter tuning
- ❖ Performance
- ❖ **Conclusion**
- ❖ Further work
- ❖ PETfold v1.0

- Base pairs with higher impact than single-stranded positions
- Structurally conserved RNA-motifs are supported by folding energy
- Performance increasement by fusion of evolutionary and thermodynamic information
- Parameter settings as suggestion – cases where evolutionary conservation is more important
- Conservation thresholds are correlated to number of substructures in alignment (as measured by Pcluster)

Further work

- ❖ Motivation
- ❖ PETfold – Combined Model
- ❖ Pfold Model
- ❖ Parameter tuning
- ❖ Performance
- ❖ Conclusion
- ❖ **Further work**
- ❖ PETfold v1.0

● Correlation between the diversity of alignments and evolutionary constraints



PETfold v1.0

- ❖ Motivation
- ❖ PETfold – Combined Model
- ❖ Pfold Model
- ❖ Parameter tuning
- ❖ Performance
- ❖ Conclusion
- ❖ Further work
- ❖ PETfold v1.0

PETfold.pl –fasta example/example.fasta

Pfold RNA sec.struct.:

```
(((((...((...--.))..((((((...))))))....((((.....--.)))))..)).
```

Constraints:

```
<<<<<.....--.....<<<<.....>>>>.....<<<.....--.>>>>>>>>>.
```

PETfold RNA sec.struct.:

```
(((((...((...--.))..((((((...))))))....((((.....--.)))))..)).
```

Score_model,struct(tree,alignm) = 43.667

Reliability_model,struct(tree,alignm) = 0.693

Acknowledgement

- ❖ Motivation
- ❖ PETfold – Combined Model
- ❖ Pfold Model
- ❖ Parameter tuning
- ❖ Performance
- ❖ Conclusion
- ❖ Further work
- ❖ PETfold v1.0

- Rolf Backofen
- Jan Gorodkin
- Bjarne Knudsen
- Ivo Hofacker