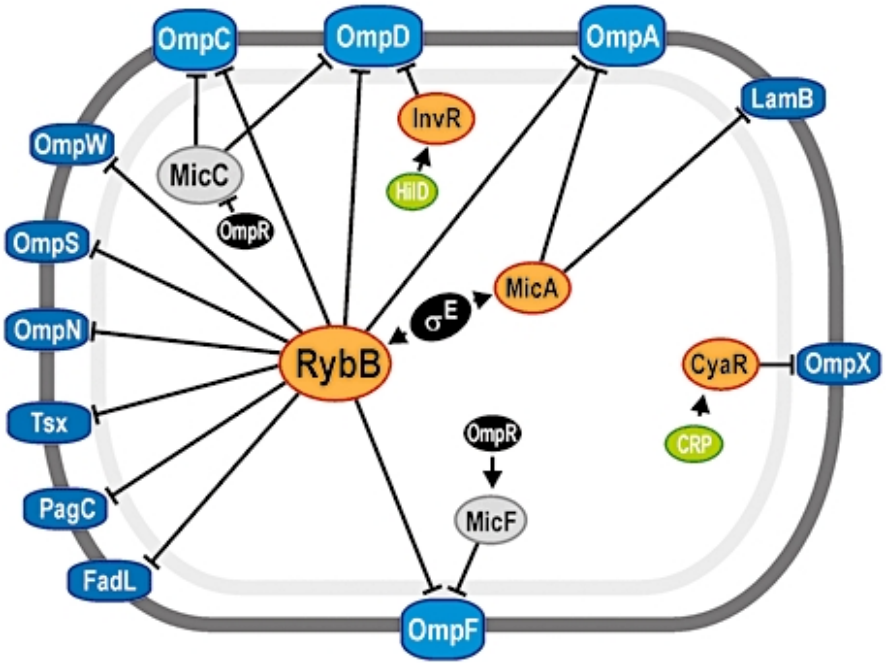
Prediction of small RNA targets incorporating seed regions and target site accessibility

Andreas Richter
Bioinformatics Group
Albert-Ludwigs-University Freiburg

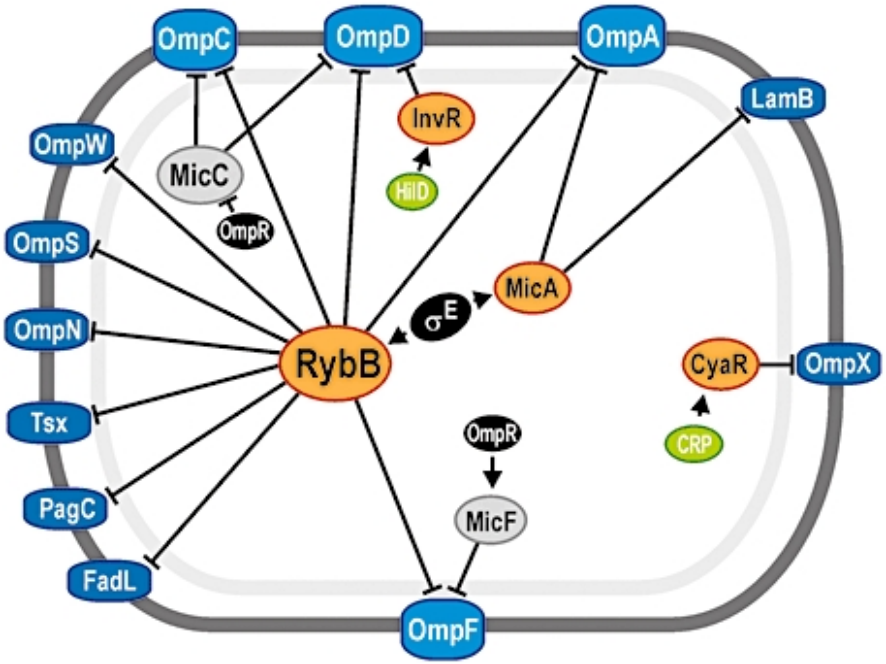
Bled, February 2009

Motivation



(Vogel, Mol. Microbiol. 2008)

Motivation

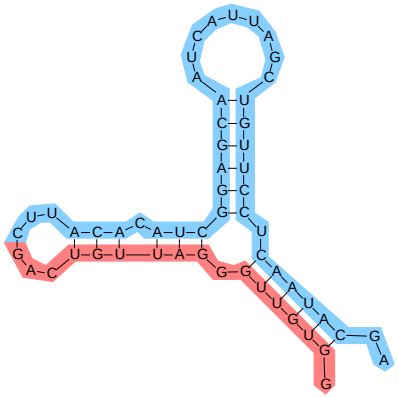


(Vogel, Mol. Microbiol. 2008)

- numerous non-coding RNAs regulate *trans*-encoded mRNAs by antisense base-pairing
- examples: eukaryotic miRNAs and siRNAs, bacterial sRNAs
- finding ncRNAs: biocomputational predictions, microarrays, deep sequencing
- one of many tasks during characterization of these ncRNAs: find putative targets

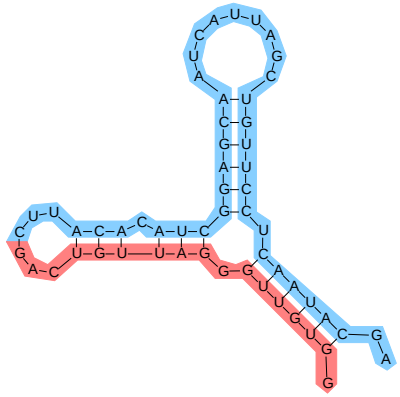
Approaches for predicting RNA-RNA interactions

- PairFold, RNAcofold:
  predict common secondary structure of 2 concatenated sequences

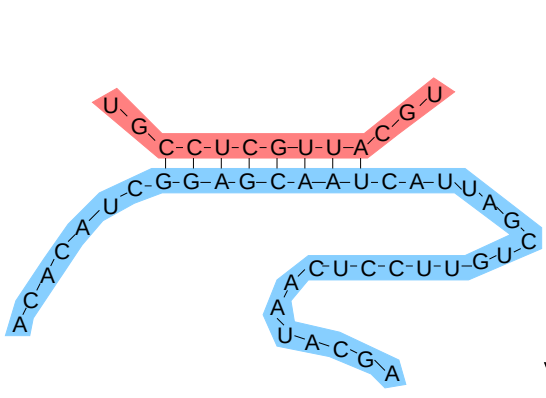


Approaches for predicting RNA-RNA interactions

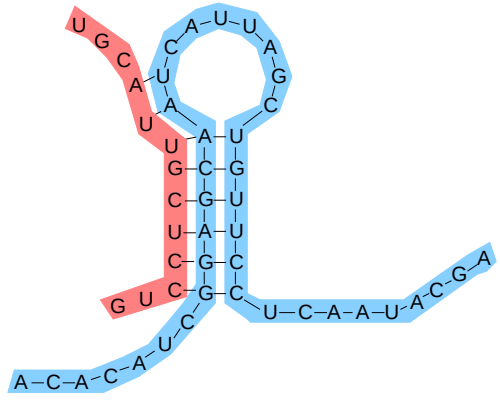
- PairFold, RNAcofold:
  predict common secondary structure of 2 concatenated sequences



- RNAhybrid, RNAduplex, RNAplex, TargetRNA:
  optimize only hybridization energy between 2 sequences



vs.



Approaches for predicting RNA-RNA interactions

- RNAup:
  optimizes hybridization energy + accessibility of hybridized part

Approaches for predicting RNA-RNA interactions

- RNAup:
  optimizes hybridization energy + accessibility of hybridized part
- accessibility:

$$ED_{a,b} = E_{a,b}^{unpaired} - E^{all}$$

$$= \text{energy that is necessary to unfold the subsequence } S_a \ldots S_b$$

E^{all} ... ensemble free energy
$E_{a,b}^{unpaired}$ ... ensemble free energy, given that $S_a \ldots S_b$ is unpaired

Approaches for predicting RNA-RNA interactions

- RNAup:
  optimizes hybridization energy + accessibility of hybridized part
- accessibility:

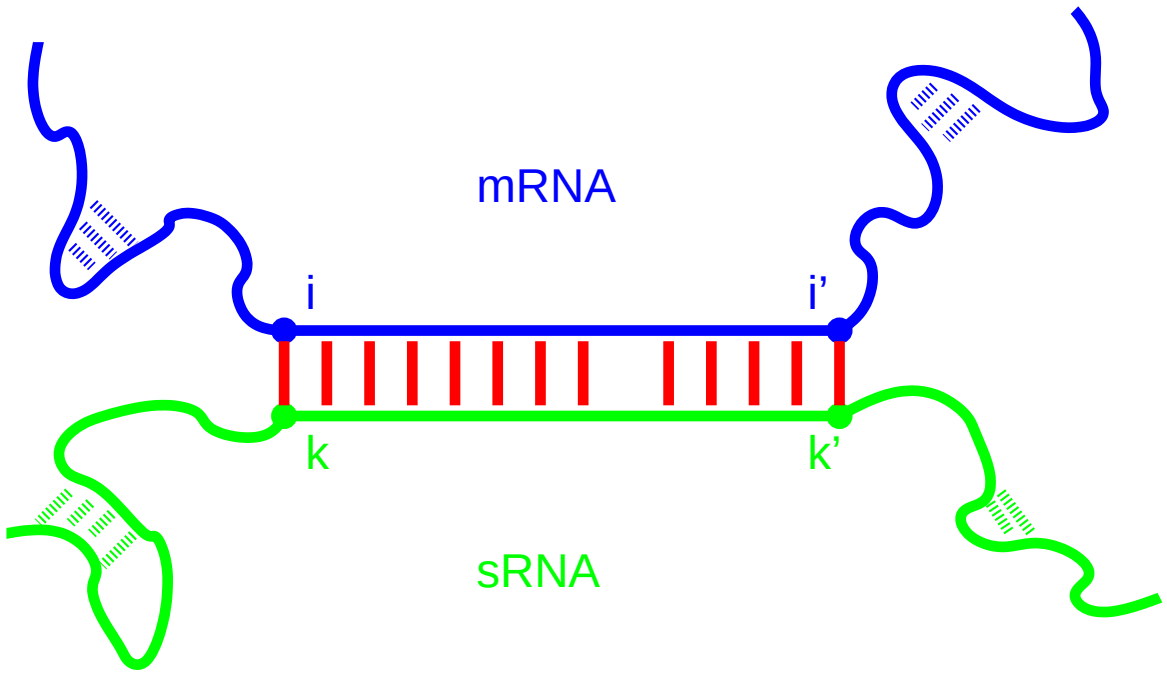
$$ED_{a,b} = E_{a,b}^{unpaired} - E^{all}$$

$$= \text{energy that is necessary to unfold the subsequence } S_a...S_b$$

E^{all} ... ensemble free energy
$E_{a,b}^{unpaired}$ ... ensemble free energy, given that $S_a...S_b$ is unpaired

- but:
  - quite slow for genome-wide predictions
  - does not support seed regions

The Idea of IntaRNA



$$E = E^{hybrid} + ED^{mRNA}_{i,i'} + ED^{sRNA}_{k,k'}$$

$ED^{mRNA}_{i,i'}$ ↓ *RNAplfold*

$ED^{sRNA}_{k,k'}$ ↓ *RNAplfold / RNAup*

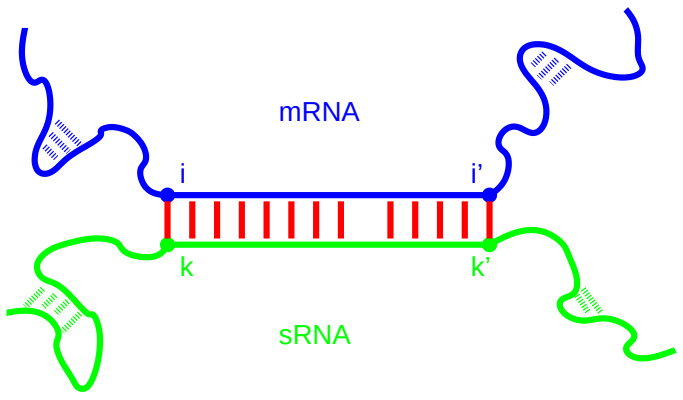
(Busch&Richter&Backofen, Bioinformatics 2008)

The Idea of IntaRNA

- use a matrix for all starts of hybridization (i', k')

$$C^{i',k'}(i,k) = \text{best energy score given that } (i,k) \text{ pair and hybridization } \textbf{starts at base pair } (i',k') \textbf{ and ends at } (i,k)$$

$$= E\left(\right.$$



$$\left. \right)$$

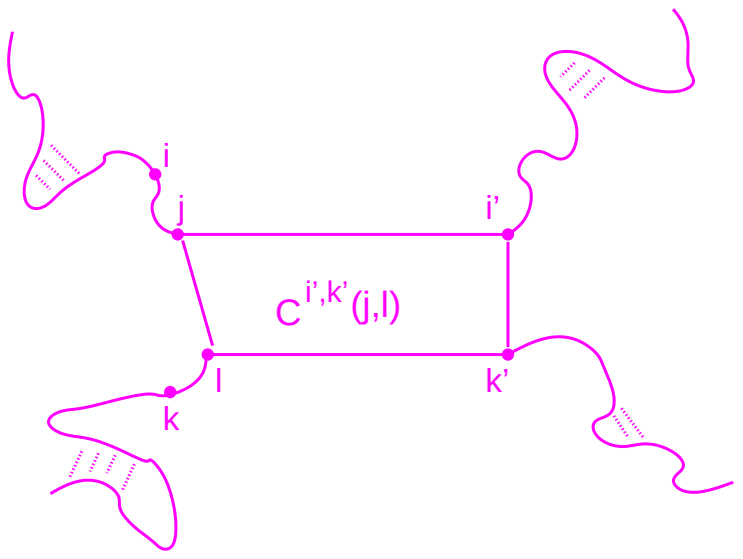
$$= Hybrid(i,k,i',k') + ED^{mRNA}_{i,i'} + ED^{sRNA}_{k,k'}$$

- can be calculated recursively using an RNAhybrid-like approach

The Recursion of IntaRNA

- for all $(i,k) \prec (i',k')$

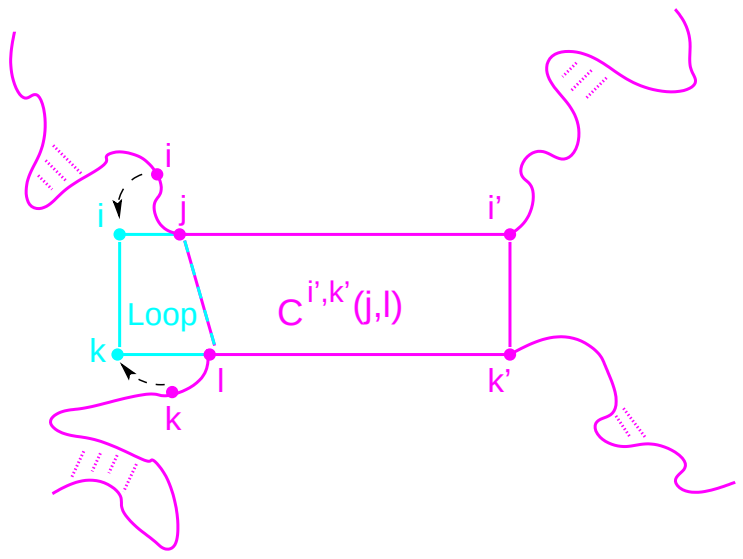
$$C^{i',k'}(i,k) = \min_{\substack{i < j \leq i' \\ k < l \leq k'}} \left(\begin{array}{ll} & + C^{i',k'}(j,l) \\ - & - \\ + & + \end{array} \right)$$



The Recursion of IntaRNA

- for all $(i,k) \prec (i',k')$

$$C^{i',k'}(i,k) = \min_{\substack{i < j \leq i' \\ k < l \leq k'}} \left(\right.$$



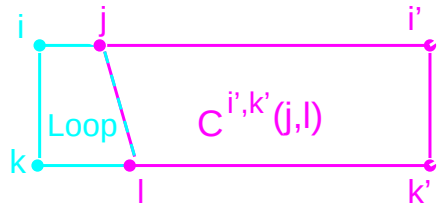
$$\left. \right)$$

$$= \min_{\substack{i < j \leq i' \\ k < l \leq k'}} \begin{pmatrix} Loop(i,k,j,l) + C^{i',k'}(j,l) \\ - \qquad\qquad - \\ + \qquad\qquad + \end{pmatrix}$$

The Recursion of IntaRNA

- for all $(i,k) \prec (i',k')$

$$C^{i',k'}(i,k) = \min_{\substack{i < j \leq i' \\ k < l \leq k'}} \left(\right.$$



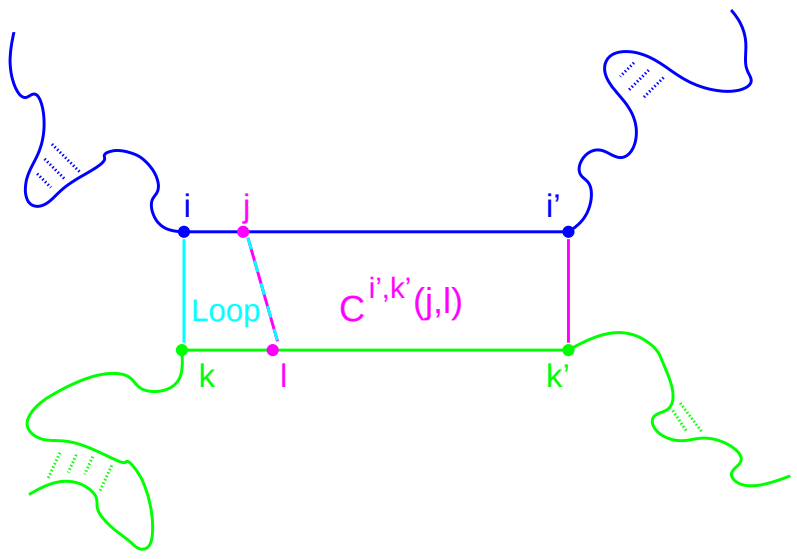
$$\left. \right)$$

$$= \min_{\substack{i < j \leq i' \\ k < l \leq k'}} \left(\begin{array}{ccc} Loop(i,k,j,l) + C^{i',k'}(j,l) \\ -ED^{mRNA}_{j,i'} - ED^{sRNA}_{l,k'} \\ + \qquad\qquad + \end{array} \right)$$

The Recursion of IntaRNA

- for all $(i,k) \prec (i',k')$

$$C^{i',k'}(i,k) = \min_{\substack{i<j\le i' \\ k<l\le k'}} \left(\right.$$



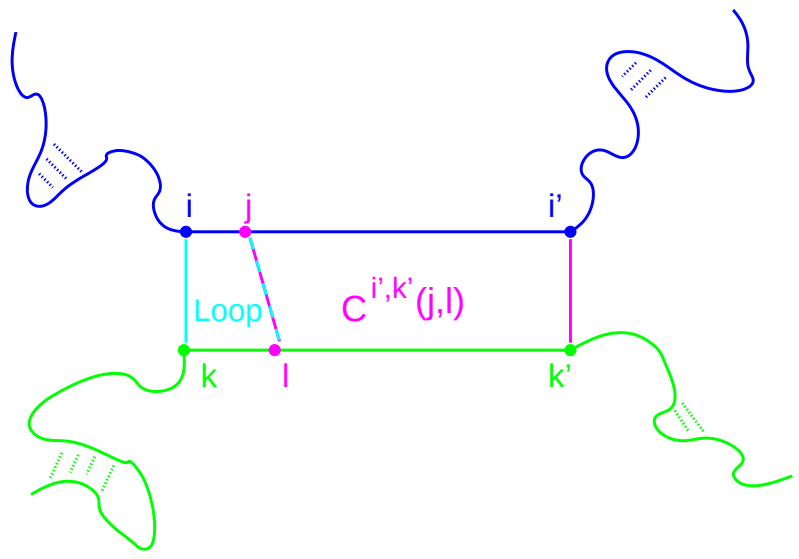
$$\left. \right)$$

$$= \min_{\substack{i<j\le i' \\ k<l\le k'}} \left(\begin{array}{l} Loop(i,k,j,l) + C^{i',k'}(j,l) \\ -ED^{mRNA}_{j,i'} - ED^{sRNA}_{l,k'} \\ +ED^{mRNA}_{i,i'} + ED^{sRNA}_{k,k'} \end{array} \right)$$

The Recursion of IntaRNA

- for all $(i,k) \prec (i',k')$

$$C^{i',k'}(i,k) = \min_{\substack{i<j\le i' \\ k<l\le k'}} \left(\right.$$



$$\left. \right)$$

$$= \min_{\substack{i<j\le i' \\ k<l\le k'}} \left(\begin{array}{l} Loop(i,k,j,l) + C^{i',k'}(j,l) \\ -ED^{mRNA}_{j,i'} - ED^{sRNA}_{l,k'} \\ +ED^{mRNA}_{i,i'} + ED^{sRNA}_{k,k'} \end{array} \right)$$

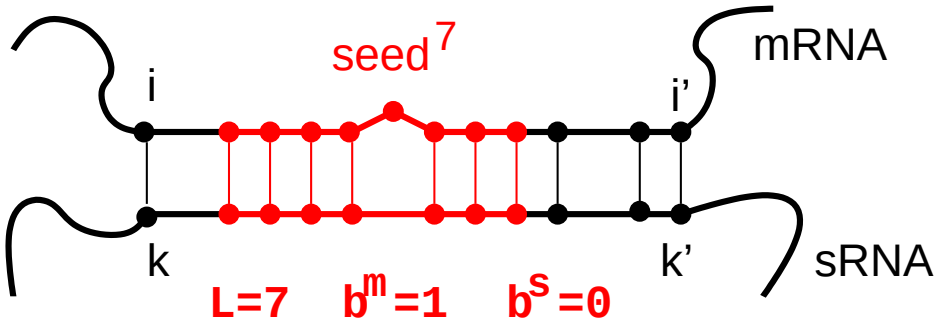
- heuristic simplification:
  - for all pairs (i,k) store only values for one hybridization start (i',k') instead of all possible starts
  - $O(mn)$ time and space for recursion

Incorporating Seeds

- **seed:** short subsequence of nearly perfect complementarity, often conserved
- found in miRNAs and siRNAs near 5'end, in sRNAs at variable positions

Incorporating Seeds

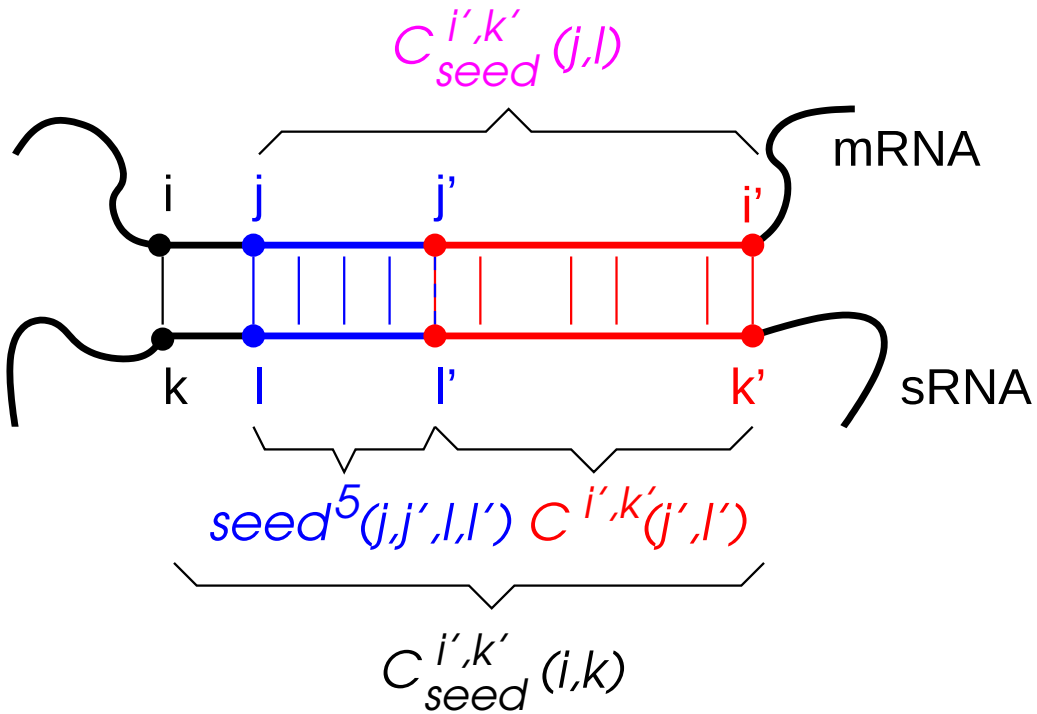
- **seed:** short subsequence of nearly perfect complementarity, often conserved
- found in miRNAs and siRNAs near 5'end, in sRNAs at variable positions
- incorporated into IntaRNA up to now with following features:
  - L ... number of bases perfectly paired in the seed region
  - b^m / b^s ... max. number of mismatches in seed region of mRNA/sRNA
  - position of seed region



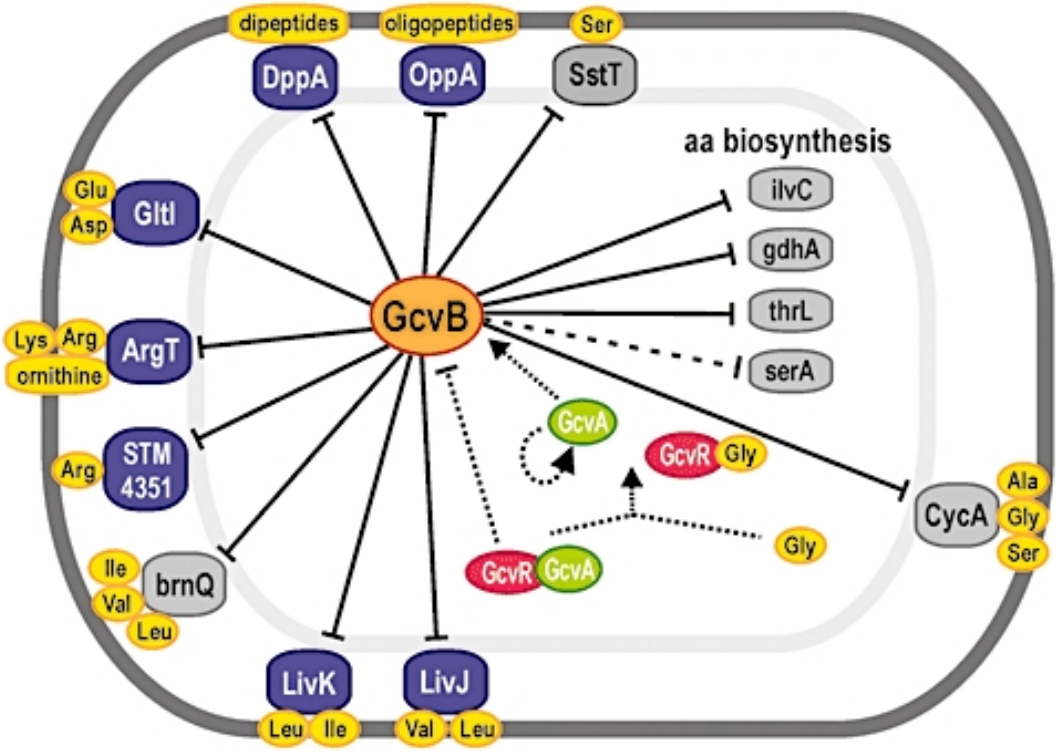
Incorporating Seeds

Incorporating a seed into the recursion to allow it at variable positions:

$C_{seed}^{i',k'}(i,k)$ = best energy score given that (i,k) pair and hybridization starts at base pair (i',k'), ends at (i,k) and **includes a seed region**

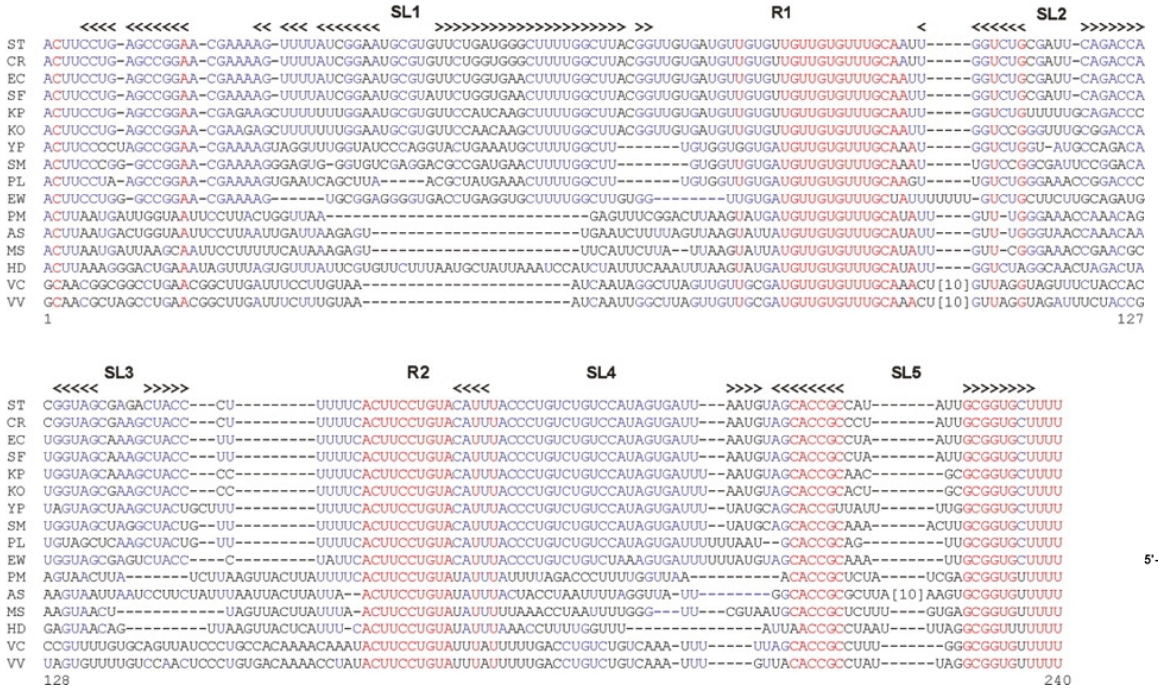


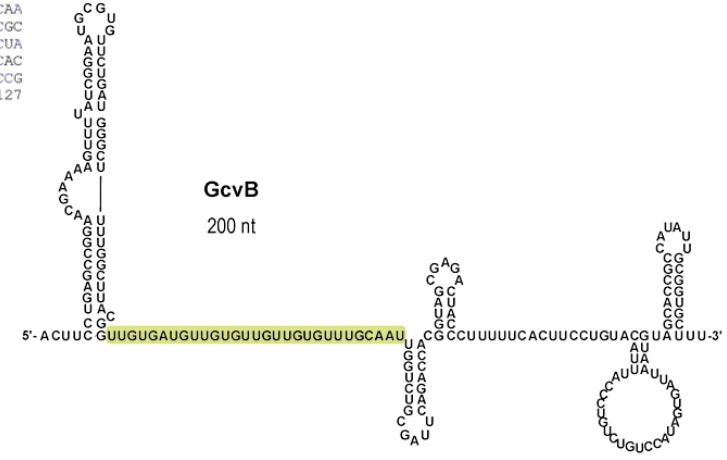
Salmonella sRNA GcvB and its targets



(Vogel, Mol. Microbiol. 2008)

Salmonella sRNA GcvB and its targets

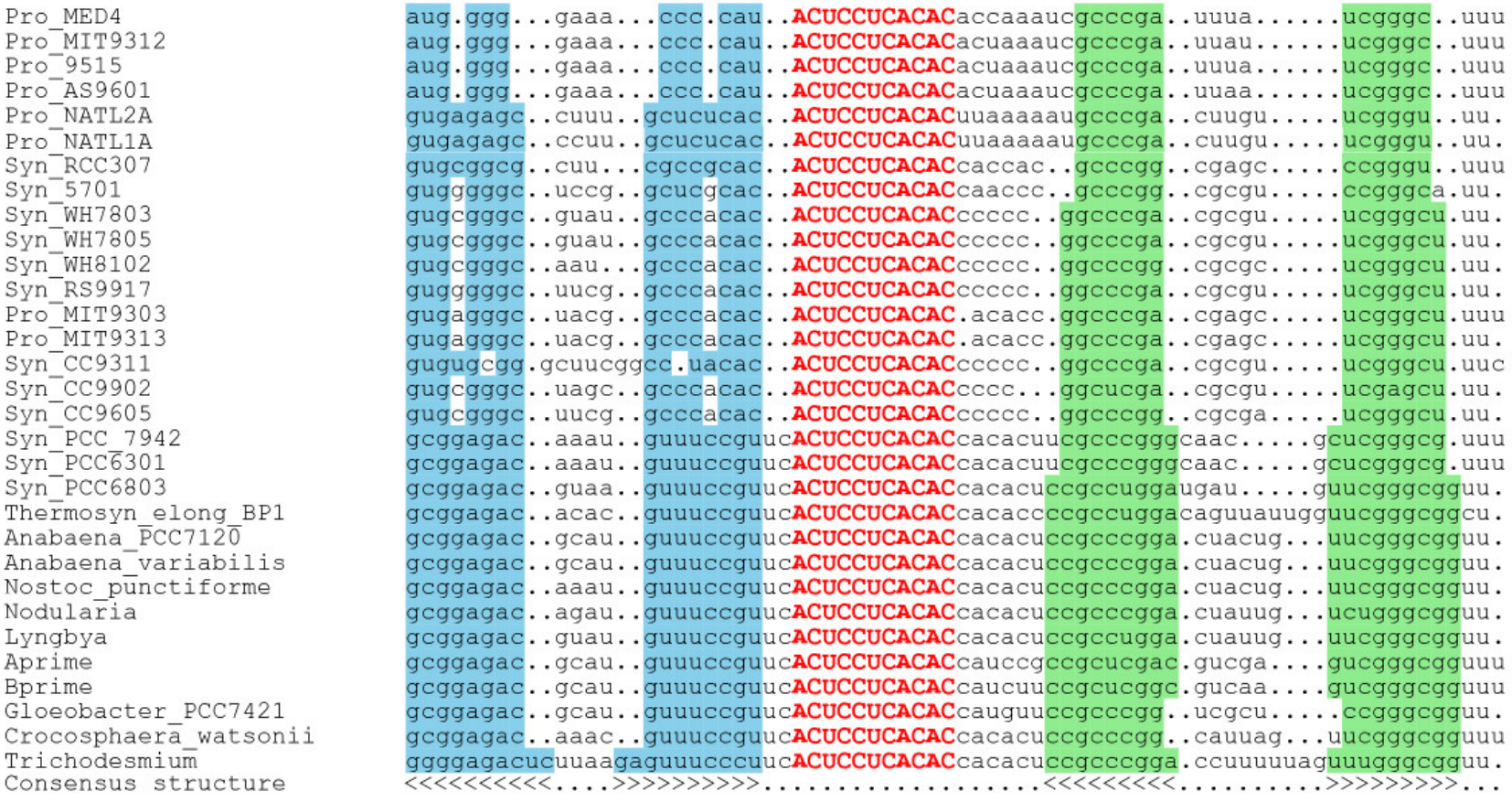




- $\sim$ 30nt G/U-rich region in GcvB interacts with all target mRNAs
- region partially ultra-conserved

(Sharma *et al.*, Genes & Dev. 2007)

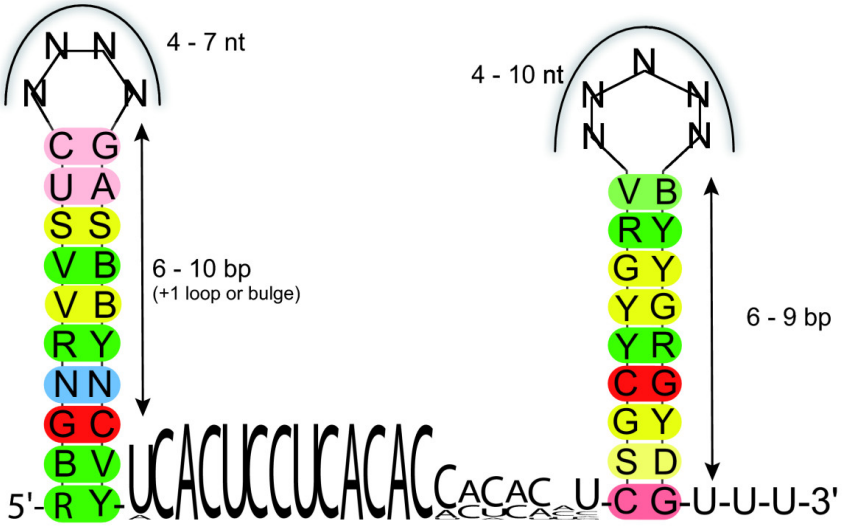
Cyanobacterial sRNA Yfr1



Multiple alignment of sRNA Yfr1 from 31 cyanobacteria

(Voß *et al.*, BMC Genomics 2007)

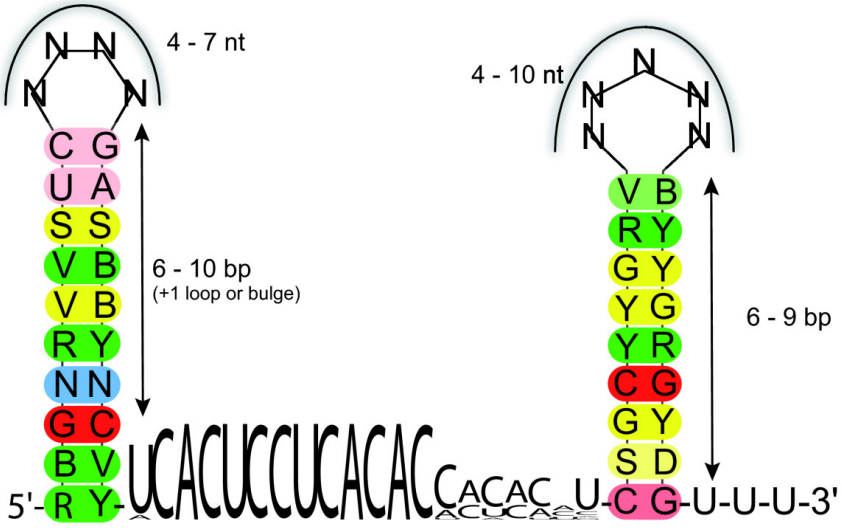
Cyanobacterial sRNA Yfr1



Sequence/structure model of Yfr1 RNA
(Voß *et al.*, BMC Genomics 2007)

sequence motif ACUCCUCACAC perfectly conserved in 31 cyanobacteria

Cyanobacterial sRNA Yfr1



Sequence/structure model of Yfr1 RNA
(Voß *et al.*, BMC Genomics 2007)

sequence motif ACUCCUCACAC perfectly conserved in 31 cyanobacteria

Scan for Yfr1 targets in *Prochlorococcus* MED4 5'-UTR and CDS (200 nt region):

variable seed (no fixed position): 725 putative targets
ultra-conserved sequence motif as seed: 47 putative targets
thereof: 4 high-scoring interactions at RBS

Predicted Yfr1 interactions for experimental validation

som (possible porin) -2 to +8

```
5'-...GAUUAUUUUUAAAUUACUCAAAU          UUUUUAUGAAGCUUU...-3'
                              UGUGUGAGGA
                              ACACACUCCU
3'-UUUCGGGCUAUUUAGCCCGCUAAACC          CAUACCCCAAAGGGGGUA-5'
```

ppa (putative inorganic pyrophosphatase) -35 to -23

```
5'-...AUUUACGUUUUAGAAUUUGAUUGA         U   GGAUAAAUAAAA...-3'
                               GUGUGAGGA GUA
                               CACACUCCU CAU
3'-UUUCGGGCUAUUUAGCCCGCUAAACCA             ACCCCAAAGGGGGUA-5'
```

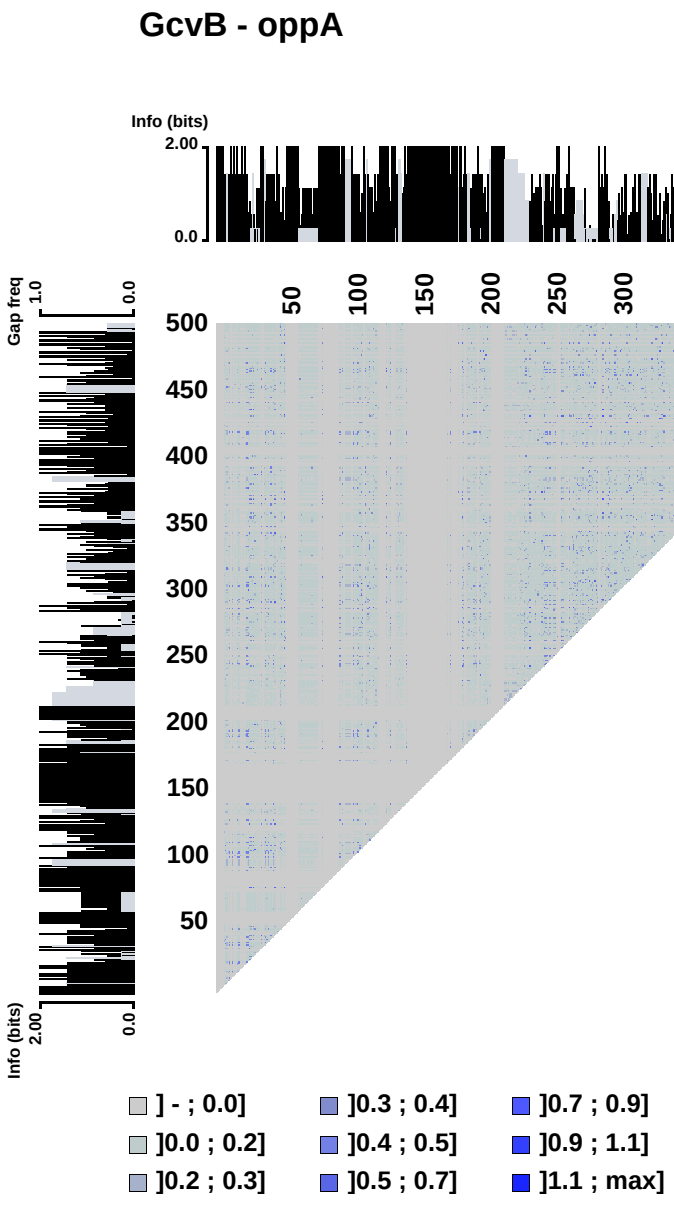
PMM1697 (Type II altern. RNA polymerase sigma factor, sigma-70 family) -4 to +8

```
5'-...AAAAUCCACUUAAAGAGGCCA      A      UCCUUCUGGAAUCUG...-3'
                            GG GUG UGGGGA
                            CC CAC ACUCCU
3'-UUUCGGGCUAUUUAGCCCGCUAAA  A          CAUACCCCAAAGGGGGUA-5'
```

som (possible porin) -2 to +9

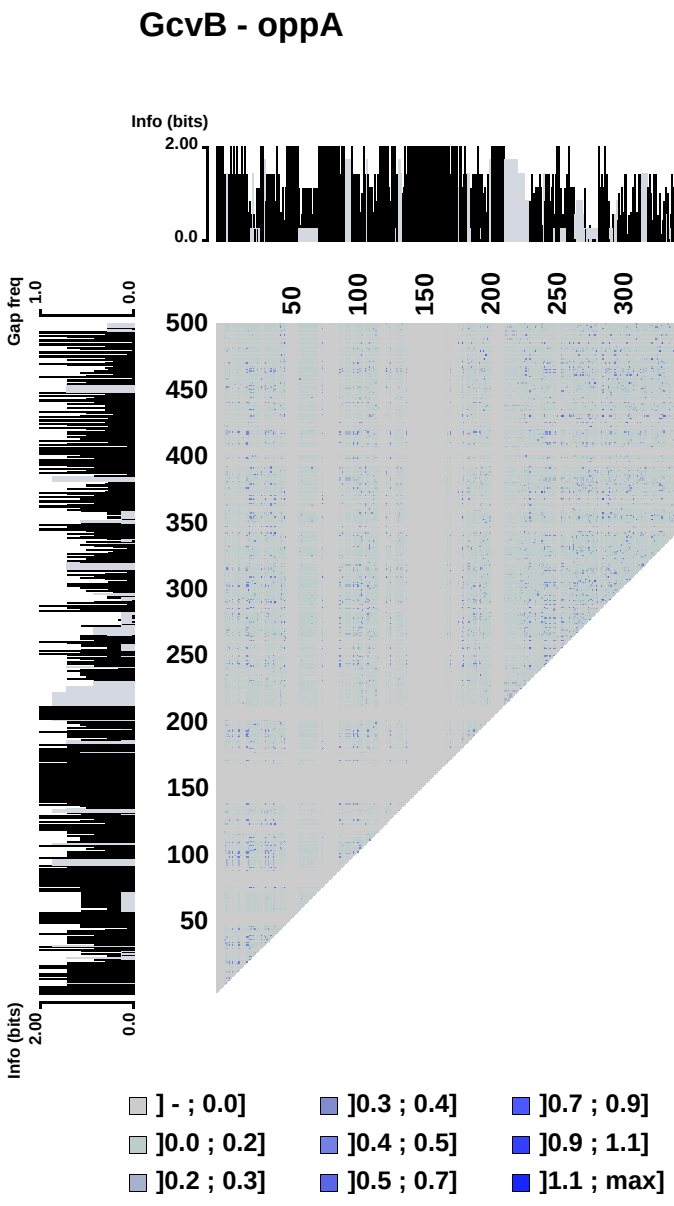
```
5'-...GUAUCUUAAGGUGUCCCUAAUAU         C AUUUAUGAAGCUUUU...-3'
                              UGUGUGAGG A
                              ACACACUCC U
3'-UUUCGGGCUAUUUAGCCCGCUAAACC           CAUACCCCAAAGGGGGUA-5'
```

Covariance at interaction sites



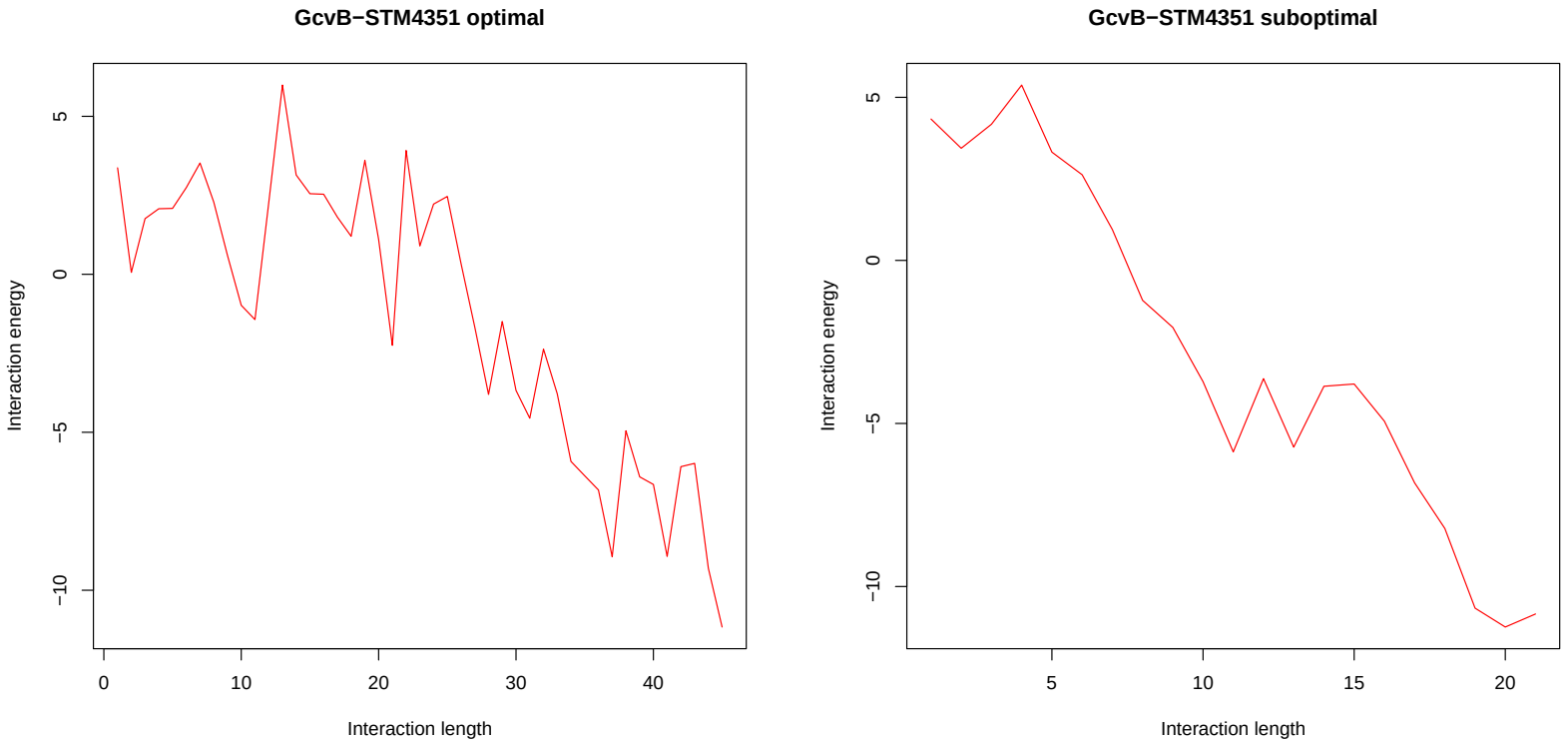
- region of interaction: (68,92), (391,414)

Covariance at interaction sites



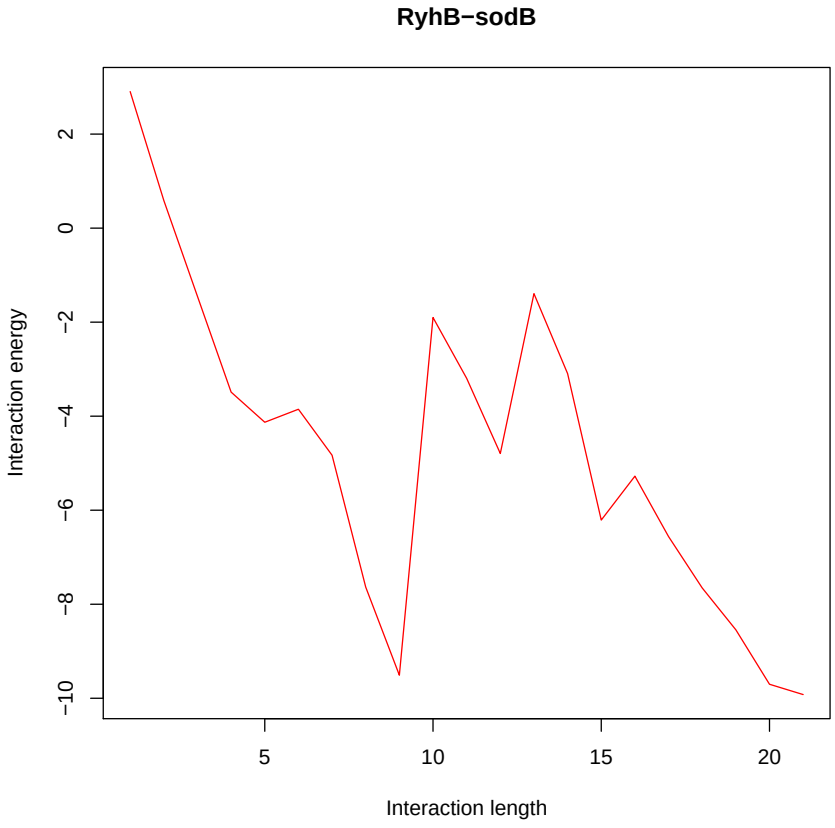
- region of interaction: (68,92), (391,414)
- high sequence conservation at interaction site, but no clear covariance signal
- consequences for target predictions:
  - find seed region with high sequence conservation in mRNA and sRNA
  - start interaction with seed and extend

IntaRNA energy score curves



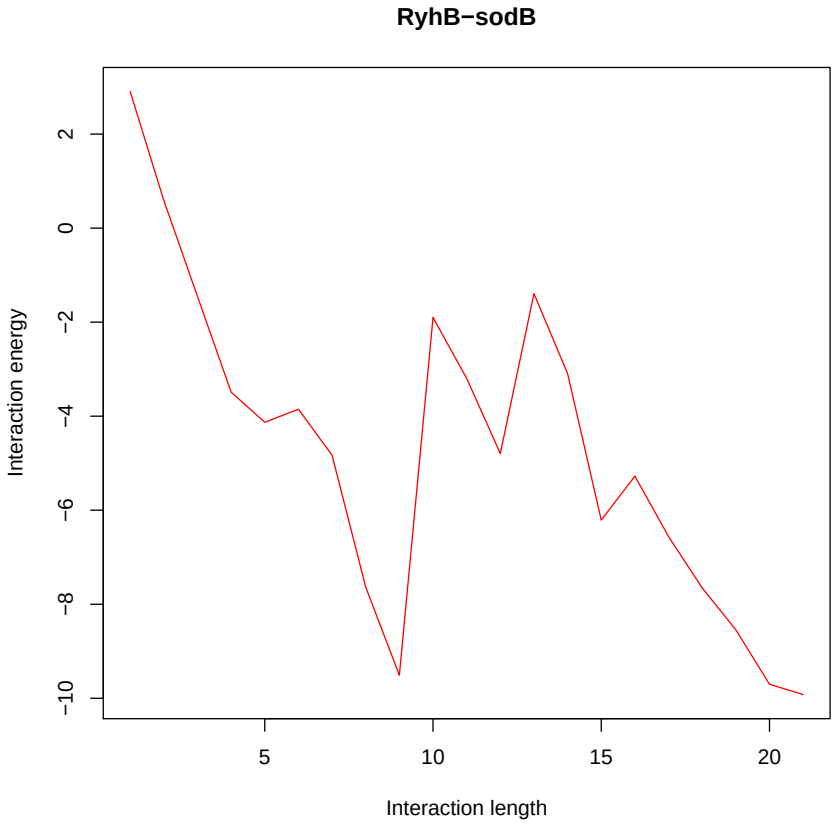
predicted optimal and suboptimal GcvB-STM4351 interactions in *Salmonella*

IntaRNA energy score curves



predicted RyhB-sodB interaction in *E. coli*

IntaRNA energy score curves



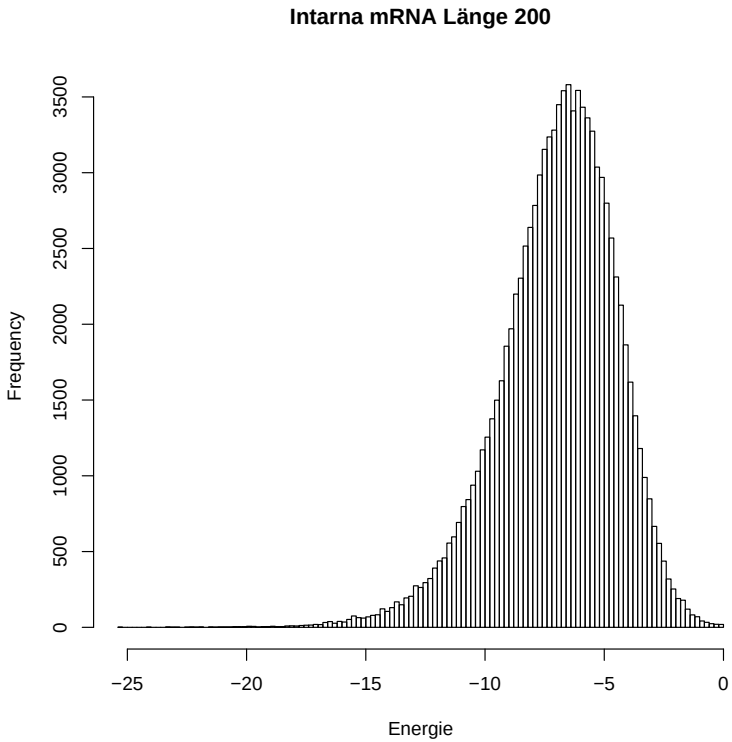
predicted RyhB-sodB interaction in *E. coli*

$\Rightarrow$ forbid distinct increases in energy score of an interaction

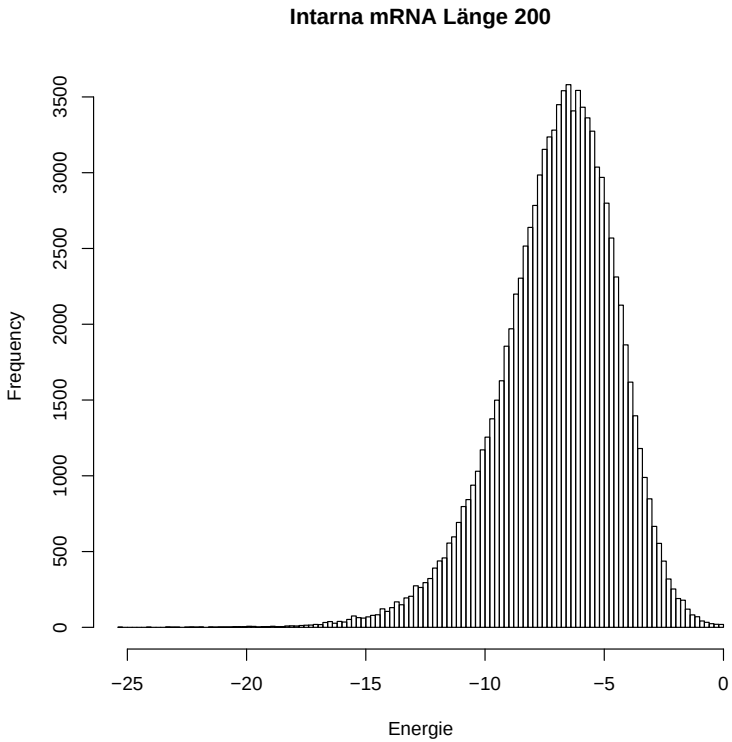
Significance of IntaRNA scores

- statistical significance of IntaRNA scores of interest, especially for genome-wide predictions
- straightforward way: estimate *P*-values for given scores by running simulations with shuffled sequences for the null model
- more desirable: analytic expressions giving dependence between sequence features and *P*-values to avoid simulations
- idea:
  - generate thousands of random sequences of various length and nucleotide compositions
  - try to find dependencies between parameters of IntaRNA score distribution and sequence features

Significance of IntaRNA scores

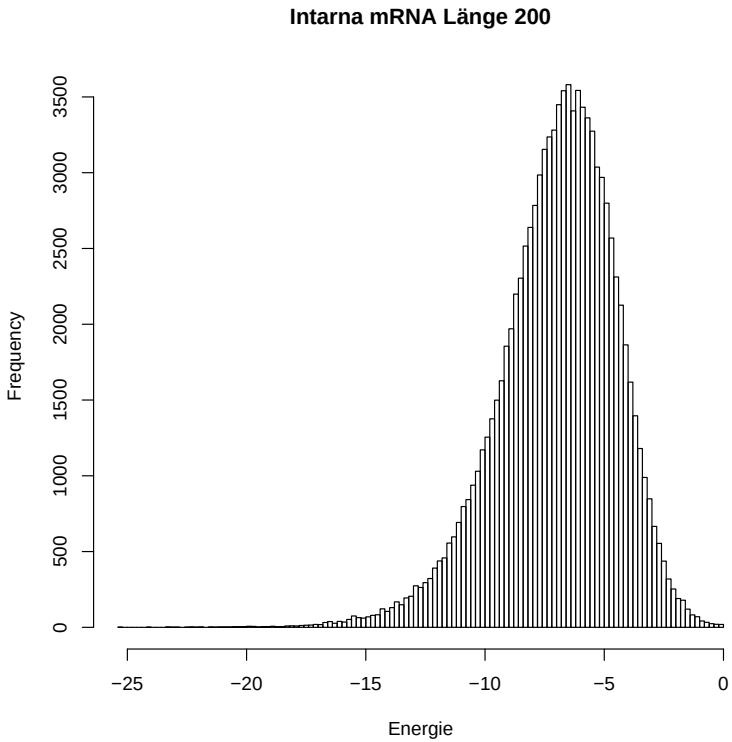


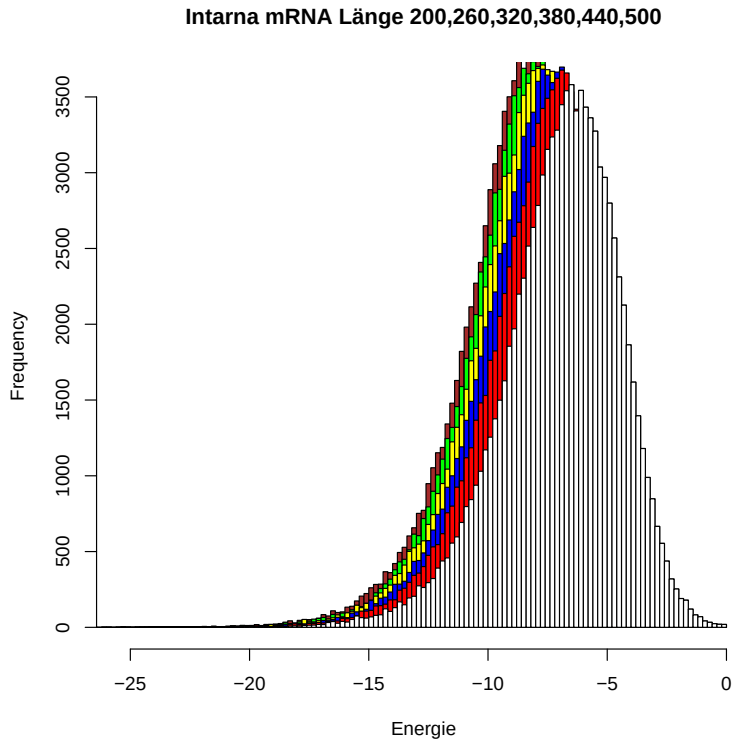
Significance of IntaRNA scores



distribution of IntaRNA scores can be approximated by extreme value distribution ($F(x) = e^{-e^{-\frac{x-b}{a}}}$, b . . . location, a . . . scale)

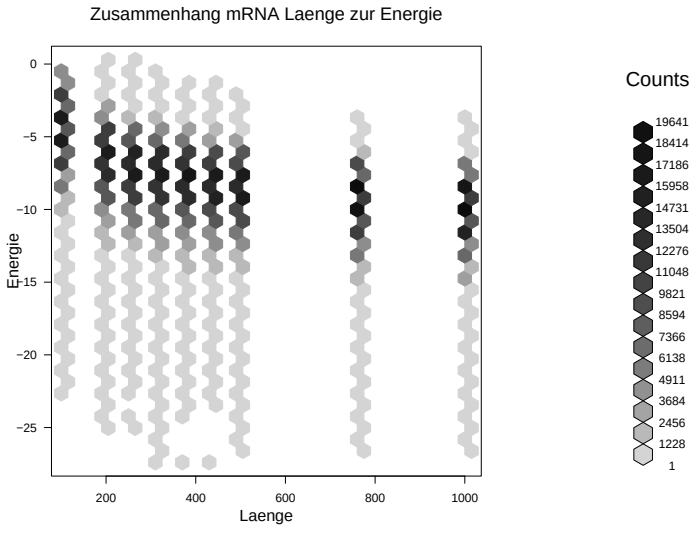
Significance of IntaRNA scores





distribution of IntaRNA scores can be approximated by extreme value distribution ($F(x) = e^{-e^{-\frac{x-b}{a}}}$, b . . . location, a . . . scale)

Significance of IntaRNA scores



it seems reasonable to normalize IntaRNA scores with $\frac{1}{\log(nm)}$
(following Rehmsmeier *et al.*, Bioinformatics 2004)

Thanks

Thanks to:

Rolf Backofen
Anke Busch
Benjamin Schulz
Claudia Steglich