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



• PairFold, RNAcofold:

predict common secondary structure of 2 concatenated sequences





• PairFold, RNAcofold:

predict common secondary structure of 2 concatenated sequences



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





• RNAup:

optimizes hybridization energy + accessibility of hybridized part  $% \left( {{\left[ {{{\left[ {{{c}} \right]}} \right]}_{ij}}}_{ij}} \right)$ 



• RNAup:

optimizes hybridization energy + accessibility of hybridized part  $% \left( {{{\left[ {{{\rm{b}}} \right]}_{{\rm{c}}}}_{{\rm{c}}}} \right)$ 

accessibility:

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

= energy that is necessary to unfold the subsequence  $S_a...S_b$ 

 $E^{all} \qquad \dots$  ensemble free energy  $E^{unpaired}_{a,\,b} \ \dots$  ensemble free energy, given that  $S_a \dots S_b$  is unpaired



• RNAup:

optimizes hybridization energy + accessibility of hybridized part  $% \left( {{{\left[ {{{\rm{b}}} \right]}_{{\rm{c}}}}_{{\rm{c}}}} \right)$ 

accessibility:

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

= energy that is necessary to unfold the subsequence  $S_a...S_b$ 

 $E^{all} \qquad \dots$  ensemble free energy  $E^{unpaired}_{a,\,b} \ \dots$  ensemble free energy, given that  $S_a \dots S_b$  is unpaired

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



### The Idea of IntaRNA



(Busch&Richter&Backofen, Bioinformatics 2008)



### The Idea of IntaRNA

- use a matrix for all starts of hybridization (i', k')
  - $C^{i',k'}(i,k) =$  best energy score given that (i,k) pair and hybridization starts at base pair (i',k') and ends at (i,k)



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

can be calculated recursively using an RNAhybrid-like approach



• for all 
$$(i, k) \prec (i', k')$$
  
 $C^{i',k'}(i, k) = \min_{\substack{i < j \le i' \\ k < l \le k'}} \left( \begin{array}{c} & & \\$ 



• for all 
$$(i, k) \prec (i', k')$$
  
 $C^{i',k'}(i, k) = \min_{\substack{i < j \le i' \\ k < l \le k'}} \left( \underbrace{Loop(i, k, j, l) + C^{i',k'}(j, l)}_{k < l \le k'} \right)$   
 $= \min_{\substack{i < j \le i' \\ k < l \le k'}} \left( \underbrace{Loop(i, k, j, l) + C^{i',k'}(j, l)}_{- - - + + +} \right)$ 



• for all 
$$(i, k) \prec (i', k')$$
  
 $C^{i',k'}(i, k) = \min_{\substack{i < j \le i' \\ k < l \le k'}} \left( \underbrace{Loop(i, k, j, l) + C^{i',k'}(j, l)}_{i = \min_{\substack{i < j \le i' \\ k < l \le k'}} \left( \underbrace{Loop(i, k, j, l) + C^{i',k'}(j, l)}_{-ED_{j,i'}^{mRNA} - ED_{l,k'}^{sRNA}} \right) \right)$ 



• for all 
$$(i, k) \prec (i', k')$$
  
 $C^{i',k'}(i, k) = \min_{\substack{i < j \leq i' \\ k < l \leq k'}} \left( \underbrace{\bigcup_{\substack{i = j \\ k < l \leq k'}} C^{i',k'}(j, l)}_{\substack{i = j \\ k < l \leq k'}} \left( \underbrace{\bigcup_{\substack{i = j \\ k < l \leq k'}} C^{i',k'}(j, l) + C^{i',k'}(j, l)}_{-ED^{mRNA}_{j,i'} - ED^{sRNA}_{j,k'}} \right)$ 



• for all 
$$(i, k) \prec (i', k')$$
  
 $C^{i',k'}(i, k) = \min_{\substack{i < j \leq i' \\ k < l \leq k'}} \left( \underbrace{\bigcup_{\substack{l = 0 \\ k < l \leq k'}} C^{i',k'}(j, l)}_{\substack{l = 0 \\ k < l \leq k'}} \left( \underbrace{\bigcup_{\substack{l = 0 \\ l = 0 \\ j, i' \\ k < l \leq k'}} C^{i',k'}(j, l) + C^{i',k'}(j, l)}_{-ED_{j,k'}^{mRNA} - ED_{j,k'}^{sRNA}}_{j,k'} \right)$ 

• 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

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



- **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<sup>m</sup>/b<sup>s</sup>... 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

Pro MED4	aug.ggggaaaccc.cauACUCCUCACACaccaaaucgcccgauuuaucgggcuuu
Pro MIT9312	aug.ggggaaaccc.cauACUCCUCACACacuaaaucgcccgauuauucgggcuuu
Pro 9515	aug.ggggaaaccc.cauACUCCUCACACacuaaaucgccccgauuuaucgggcuuu
Pro AS9601	aug.ggggaaaccc.cauACUCCUCACACacuaaaucgcccgauuaaucgggcuuu
Pro NATL2A	gugagagecuuugcucucaeACUCCUCACACuuaaaaaugeeegacuuguucggguuu.
Pro NATL1A	gugagageccuugcucucacACUCCUCACACuuaaaaaugeecgacuuguucggguuu.
Svn RCC307	gugeggeg., cuucgeegeac., ACUCCUCACACeaceac., geeeggcgageegggguuuu
Svn 5701	gugggggc.,uccg.,gcucgcac.,ACUCCUCACACcaaccc.,gcccgg.,cgcguccggggca.uu.
Svn WH7803	gugeggge, guau, geeeacac, ACUCCUCACACeeece, ggeeega, egegu,, uegggeu, uu,
Syn WH7805	gugegggeguaugeceacacACUCCUCACACcececeggcccgacgcguucgggcu.uu.
Syn WH8102	gugcgggcaaugcccacacACUCCUCACACccccccggcccggcgcgcucgggcu.uu.
Svn RS9917	gugggggc.,uucg.,gcccacac,ACUCCUCACACcccccc.,ggcccga.,cgcgu,ucgggcu.uu.
Pro MIT9303	gugagggc., uacg., gcccacac, ACUCCUCACAC, acacc, ggcccga, .cgagc, ucgggcu, uuu
Pro MIT9313	gugagggcuacggcccacacACUCCUCACAC.acacc.ggcccgacgagcucgggcu.uu.
Syn CC9311	gugugcgg.geuucggcc.uacac.ACUCCUCACACeccccggcccgacgcguucgggcu.uuc
Syn CC9902	gugegggeuagegeceacaeACUCCUCACACecceggeuegacgeguuegageu.uu.
Syn CC9605	gugegggeuueggeceacaeACUCCUCACACeccccggeceggcgcgauegggeu.uu.
Syn PCC 7942	geggagacaaauguuuceguucACUCCUCACACcacacuucgceegggcaacgeucgggeg.uuu
Syn PCC6301	geggagacaaauguuuceguucACUCCUCACACacacuuegecegggcaacgeucgggeg.uuu
Syn PCC6803	geggagacguaaguuuceguucACUCCUCACACacacacucegecuggaugauguucgggegguu.
Thermosyn elong BP1	geggagacacacguuuceguucACUCCUCACACcacacccccgccuggacaguuauugguucggggggcu.
Anabaena PCC7120	geggagacgeauguuuceguucACUCCUCACACeceacacucegeeegga.cuacuguucgggegguu.
Anabaena variabilis	geggagacgeauguuuceguucACUCCUCACACeceacacucegeeegga.cuacuguucgggeegguu.
Nostoc punctiforme	geggagacaaauguuuceguucACUCCUCACACeceacucegeeegga.cuacuguucgggegguu.
Nodularia	gcggagacagauguuuccguucACUCCUCACACacacuccgcccgga.cuauugucugggcgguu.
Lyngbya	geggagacguauguuuceguucACUCCUCACACacacucegecugga.cuauuguucgggegguu.
Aprime	geggagacgeauguuuceguucACUCCUCACACeaucegeegeuegac.guegaguegggegguuu
Bprime	geggagacgeauguuuceguucACUCCUCACACeaucuucegeucege.gucaagucgggegguuu
Gloeobacter PCC7421	gcggagacgcauguuuccguucACUCCUCACACcauguuccgcccggucgcuccgggcgguu.
Crocosphaera watsonii	geggagacaaacguuuceguucACUCCUCACACcacacuccgcccggcauuaguucgggegguuu
Trichodesmium	ggggagacucuuaagaguuucccuucACUCCUCACACcacacuccgcccgga.ccuuuuuaguuugggcgguu.
Consensus structure	~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

#### 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'-...GAUUAUUUUUAAAAUUACUCAAAU UUUUUUAUGAAGCUUU...-3' UGUGUGAGGA ACACACUCCU 3'-UUUCGGGCUAUUUAGCCCGCUAAAC CAUACCCCAAAGGGGGUA-5'

```
ppa (putative inorganic pyrophosphatase) -35 to -23
```

5'-...AUUUACGUUUUAGAAUUUGAUUGA U GGAUAAAUAAAA....-3' GUGUGAGGA GUA CACACUCCU CAU 3'-UUUCGGGCUAUUUAGCCCGCUAAACCA ACCCAAAGGGGGUA-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'-UUUCGGGCUAUUUAGCCGCCUAAAGGGGUA-5'



### Covariance at interaction sites

GcvB - oppA



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



### Covariance at interaction sites

GcvB - oppA



- 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



- 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









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





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



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



### Thanks to:

Rolf Backofen Anke Busch Benjamin Schulz Claudia Steglich

