COMP598: ADVANCED COMPUTATIONAL BIOLOGY RESEARCH & METHODS

RNA-RNA interaction prediction

Jerome Waldispuhl
School of Computer Science, McGill

From slides from Ivo Hofacker (University of Vienna)

Motivation

- Experimental and bioinformatical methods find novel ncRNAs *en masse*
- Give no hint as to the function of these novel ncRNAs
- Functional characterization of ncRNAs is difficult and slow

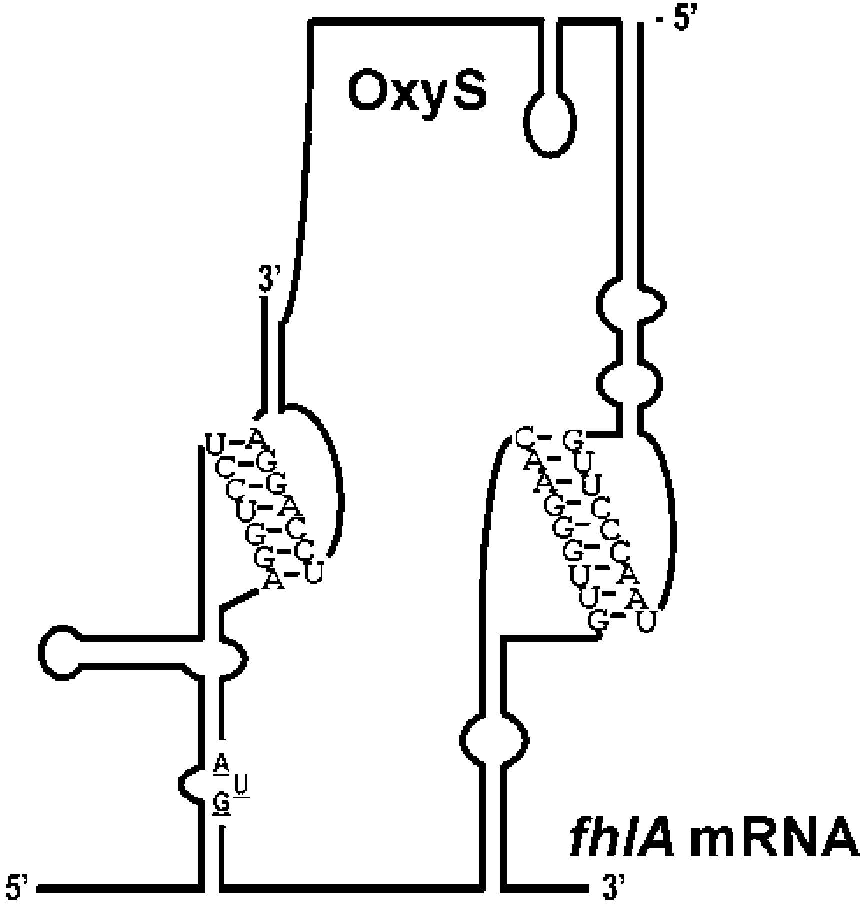
- Most ncRNAs function through interaction with other RNAs
- Identification of interaction partners is the easiest approach to learn about possible functions
- Most obvious in the case of miRNA target prediction

Well known Examples of RNA-RNA Interaction

- micro RNAs regulate mRNA translation
- snoRNAs guide methylation and pseudouridylation of rRNA
- some well studied bacterial examples
  - RyhB is transcribed under low Fe, binds several mRNA of Fe binding proteins (sdh, sodB) and leads to mRNA degradation
  - GadY interacts with the 3' UTR of GadX and inhibits its degradation
  - DsrA is expressed at low temperatures and stimulates the translation of RpoS a translational regulator
  - OxyS is expressed under oxidative stress and inhibits translation of its targets RpoS and flhA
  - T-box motifs bind uncharged tRNAs to control transcription of aminoacyl synthetases

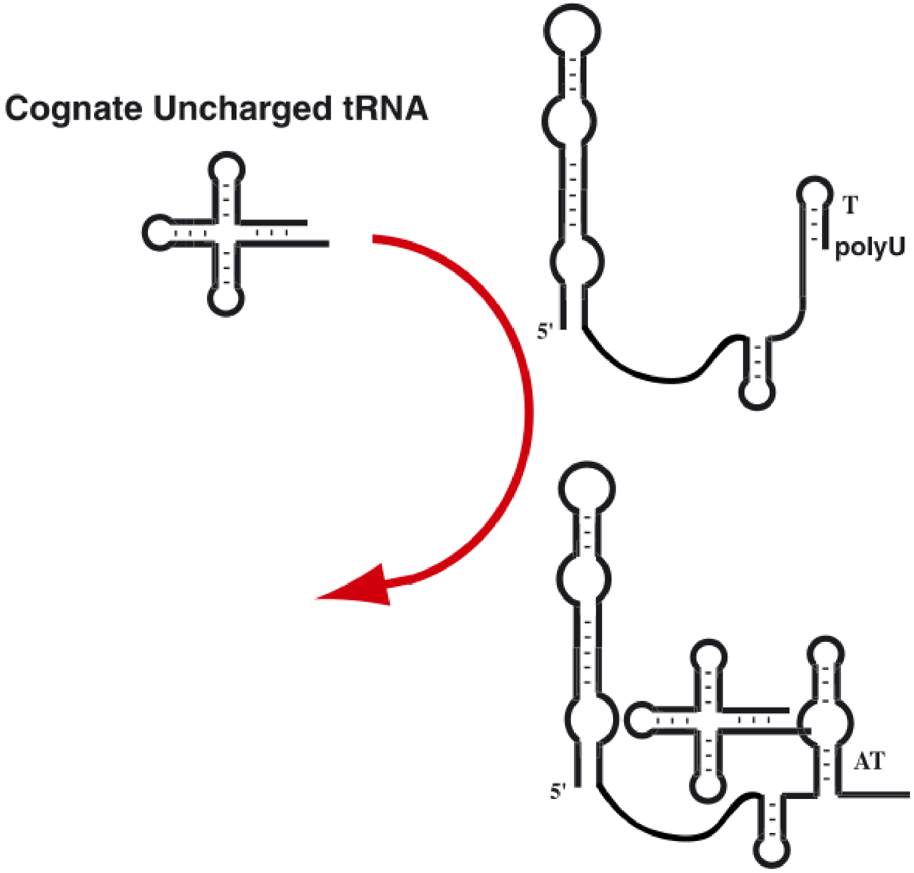
Interaction of OxyS and fhla

Binding of OxyS to fhlA mRNA makes the ribosome binding site (start codon) inaccessible



Transcriptional control by T-box Motifs

Concentration of un-charged tRNAs controls transcription of its aminoacyl synthetase



Challenges

- Few well-studied examples
- Energetics of many interaction motifs are unknown
- Length of the interacting region is often quite small
- Binding is a concentration dependent process
- Folding kinetics rather than thermodynamics may play a role
- A single small RNA may have many targets
- RNA chaperones such as Hfq may be required for binding
- ncRNAs often act within RNPs, what's the influence of the protein?

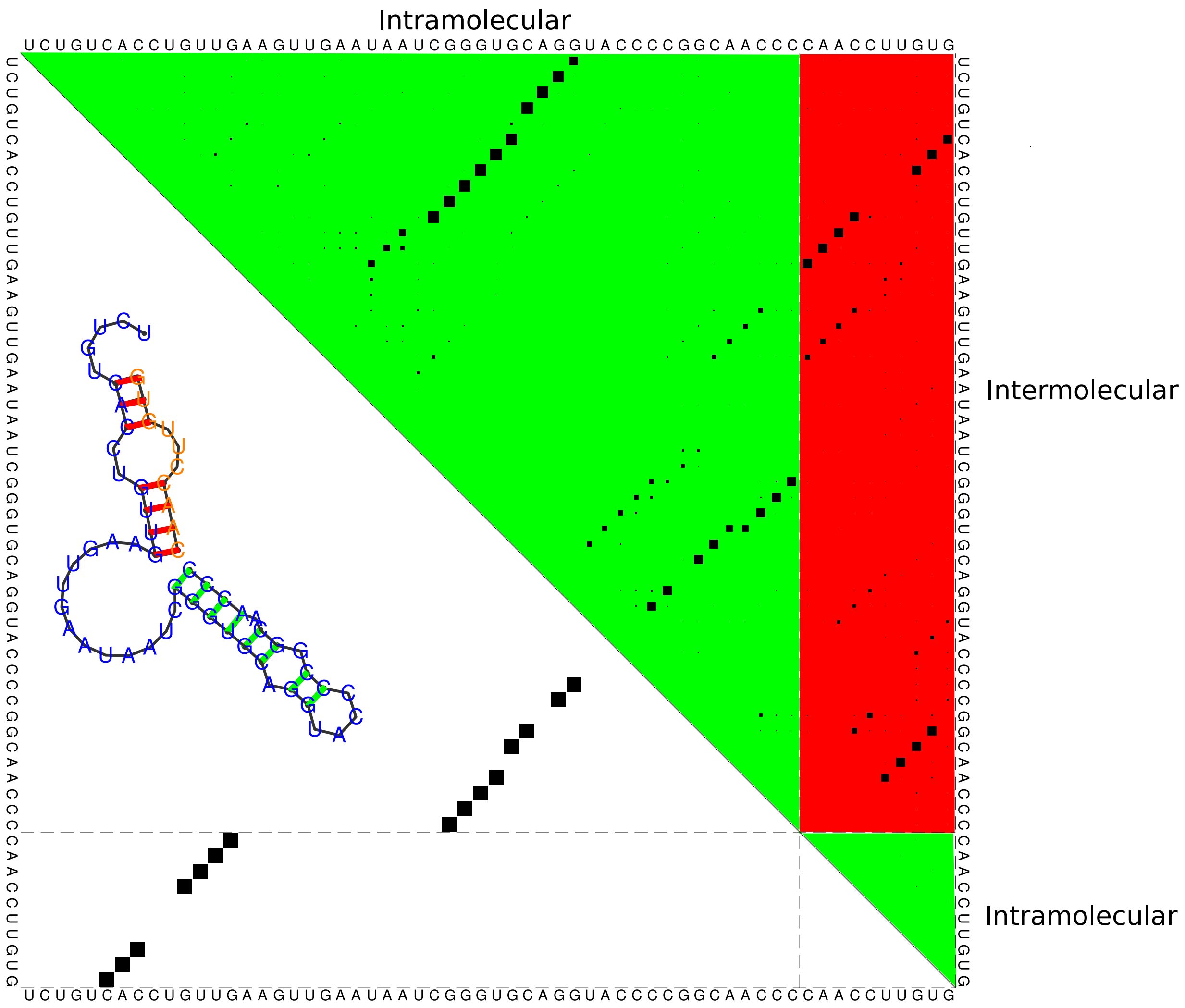
Overview of Prediction Strategies

- Co-folding by concatenation of two sequences, e.g. `RNAcofold`, `pairfold`, `DINAMELT`, `Nupack`
- Co-folding with pseudoknot-like structures, `IRIS`
- Using only inter-molecular interaction, i.e. assume that both molecules are unstructured by themselves. `RNAhybrid`, `RNAduplex`, codeRNAplex
- Combine interaction search with accessibility calculations. `RNAup`, `RNAplfold + RNAplex`, `oligowalk`

Simple Co-folding of two RNAs

- Poor man's approach to cofolding:
  - Concatenate two RNAs using a short linker
  - Use conventional folding programs such as mfold
- Proper way:
  - Use modified folding algorithm that keeps track of the break between the strands
  - Any loop containing the break point is treated specially.
  - Implemented in the RNAcofold program of the Vienna RNA package
- Limited to structures that are pseudo-knot free for concatenated sequences.

Pair Probabilities from RNAcofold



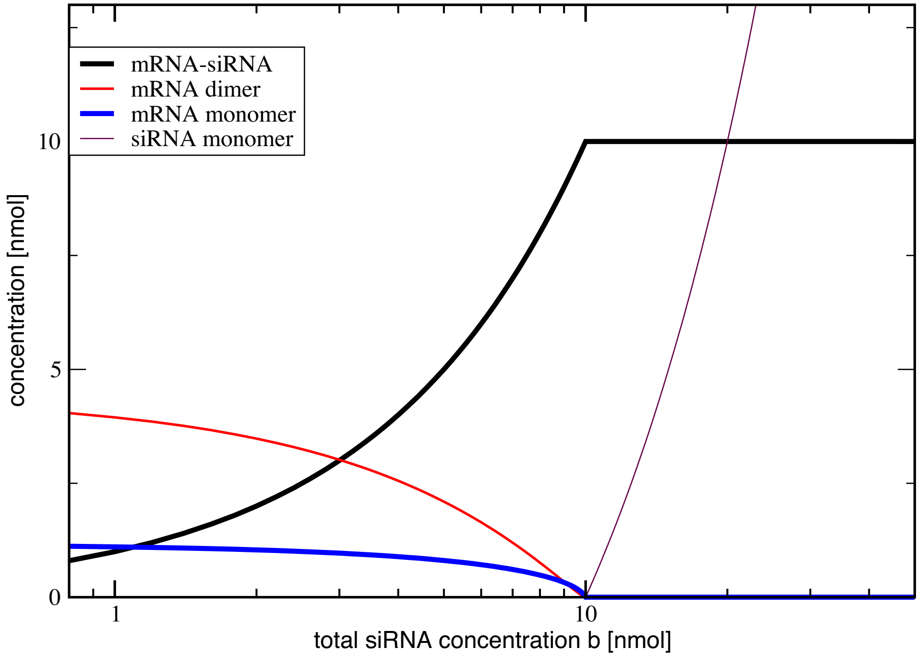
Concentration Dependence of RNA-RNA interactions

Binding processes are always concentration dependent

For two RNAs we have three reactions in equilibrium:

$$A + B \rightleftharpoons AB \qquad A + A \rightleftharpoons AA \qquad B + B \rightleftharpoons BB$$

Compute concentrations of all five monomers and dimers.



UNAFold: prediction of RNA/DNA hybridization

(Dimitrov&Zuker,2004)

Motivation:

Let A and B be two polynucleotide sequences. In solution, UNAFold aims to predict the concentration of single stranded folded and unfolded A and B **AND** hybridization AA, BB and AB.

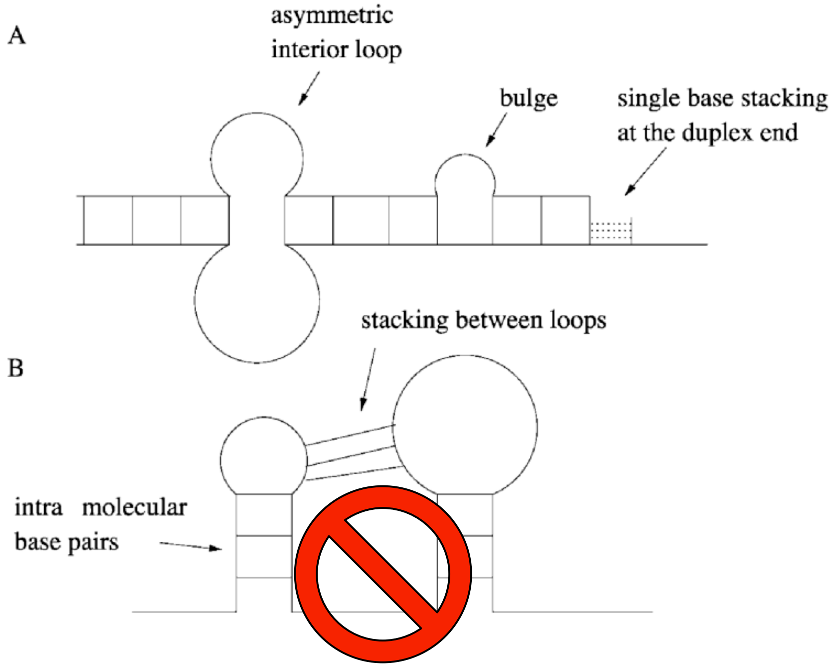
Principles:

- Simple modification of the McCaskill's algorithm.
- Stacking energies computed from experimental measures.

Results:

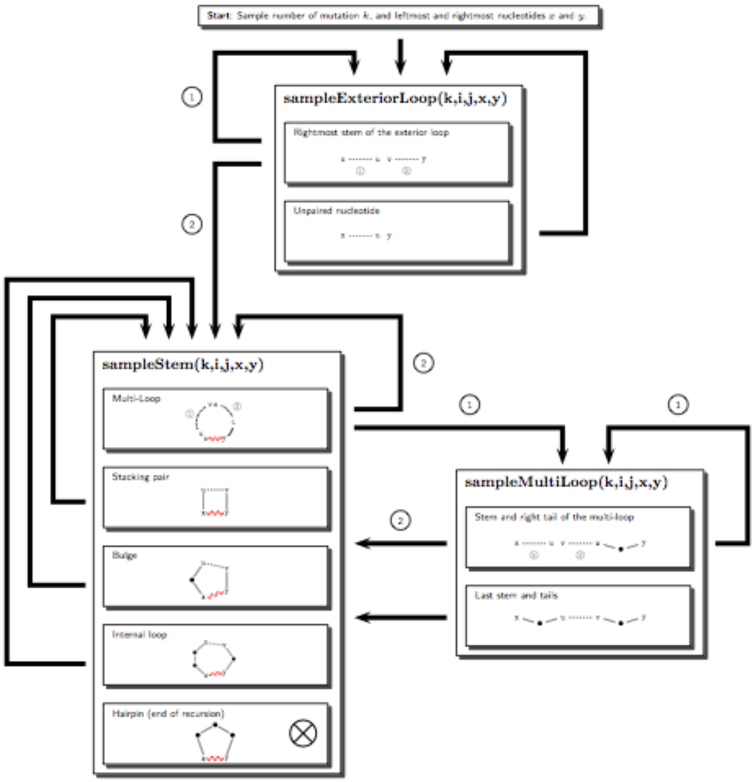
Reproduce experimental observations

Allowed configurations:


A
asymmetric interior loop
bulge
single base stacking at the duplex end
B
stacking between loops
intra molecular base pairs

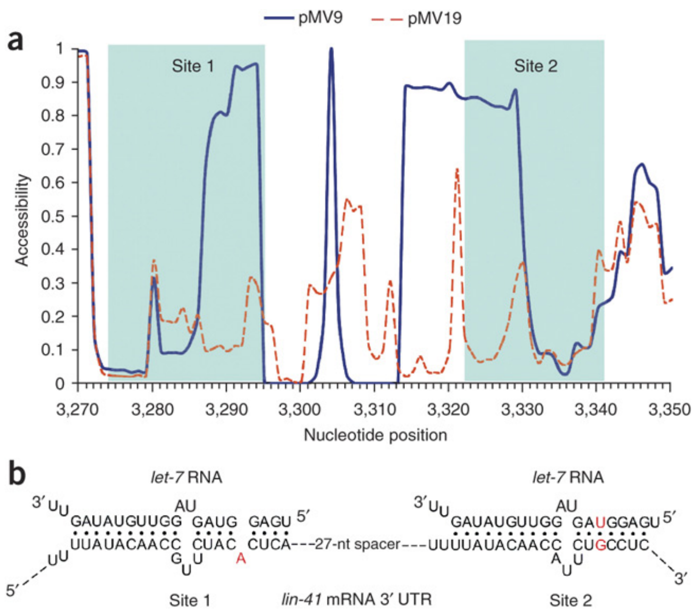
Sfold: Accessibilty prediction through Boltzmann sampling (Ding&Lawrence,2001)

Sample secondary structures using a stochastic backtracking procedure:


Start: Sample number of mutation k, and leftmost and rightmost nucleotides x and y.
sampleExteriorLoop(k,i,j,x,y)
Rightmost stem of the exterior loop
x -------- u v -------- y
①
②
Unpaired nucleotide
x -------- u y
sampleStem(k,i,j,x,y)
Multi-Loop
Stacking pair
Bulge
Internal loop
Hairpin (end of recursion)
sampleMultiLoop(k,i,j,x,y)
Stem and right tail of the multi-loop
Last stem and tails
1
2

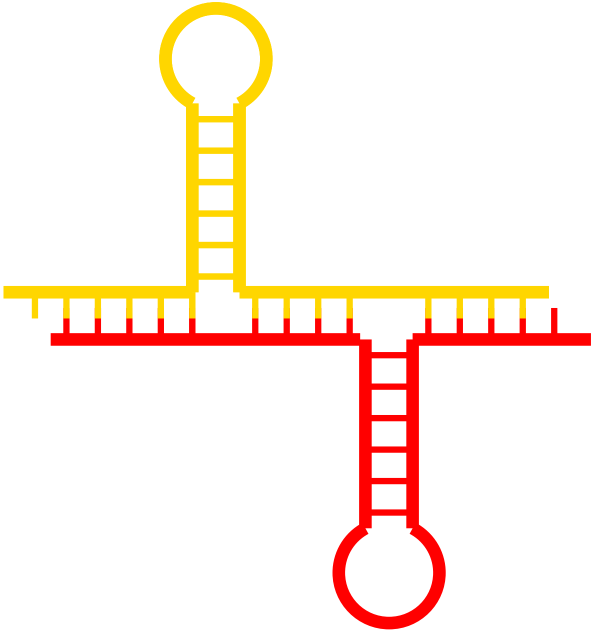
Principle:

- Estimate accessibility (not base paired) of each nucleotide in the sample set.
- Identify the hybridization regions.

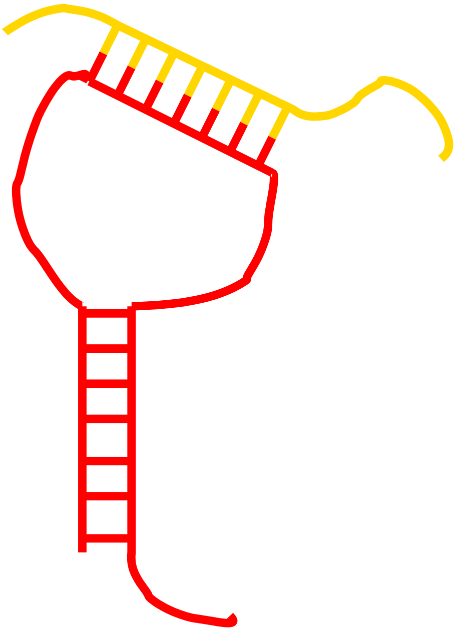


Structures (not) Predicted by RNAcofold

knot-free



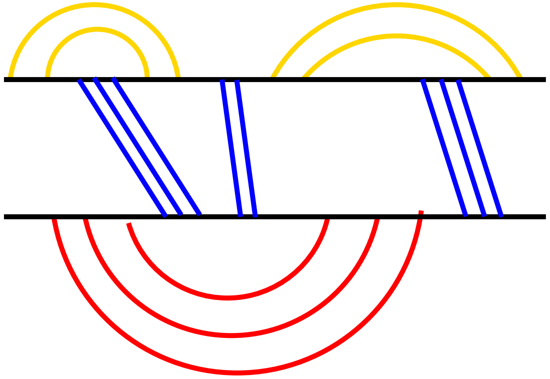
pseudo-knotted



Predicting more complex Structures

Without restricting allowed structure motif RNA-RNA interaction is NP-complete

- The most general algorithms (Alkan 2006, Pervouchine 2004) allow structures where
  - Intra-molecular pairs form pseudo-knot free structures
  - Inter-molecular pairs are not allowed to cross
- Run time is too slow for most purposes ($\mathcal{O}(n^3 \cdot m^3)$)



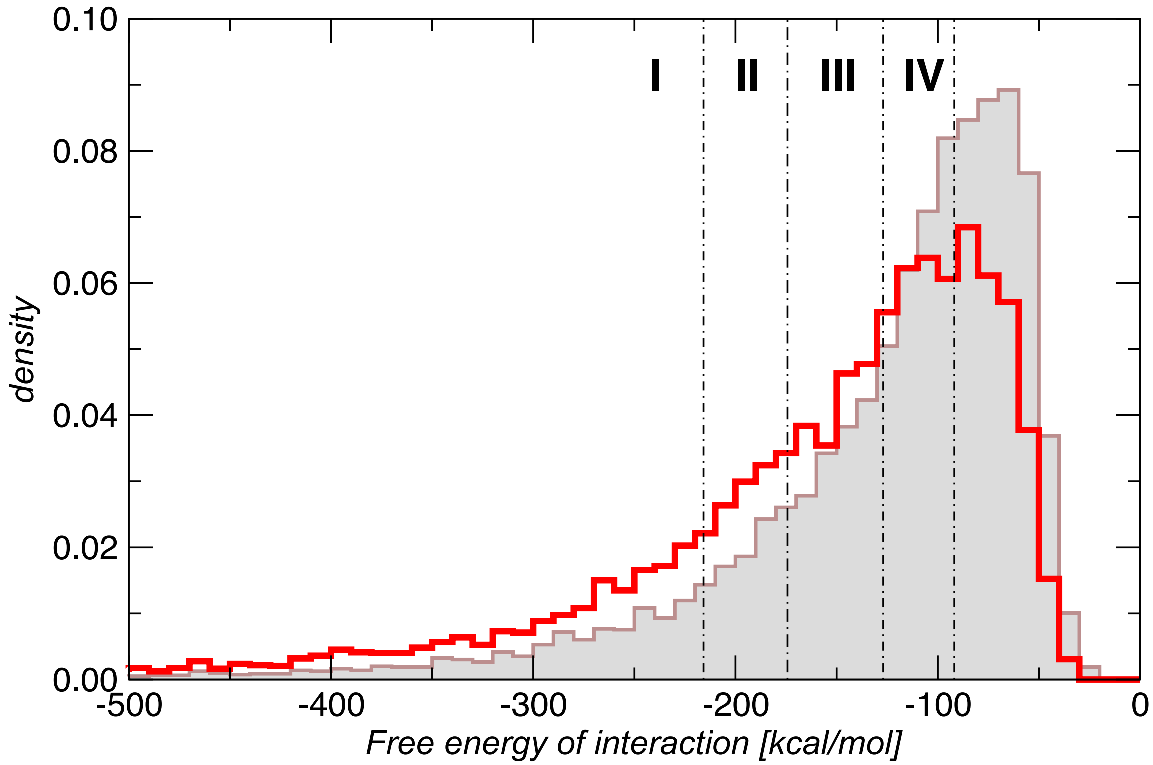
Fast Interaction Search

Methods for fast interaction search

- Search for sequence complementarity by BLAST
- Better: Interaction search using thermodynamics
- Simplified folding algorithm without intra-molecular pairs.
- Runs in $\mathcal{O}(n \cdot m)$ time.
- Used in RNAhybrid (miRNA target prediction), RNAduplex, RNAplex

What's the effect of neglecting intra-molecular structure?

Frequency of ncRNA - mRNA Interactions



RNA-mRNA interaction interaction energies (from `RNAduplex`)
red: ncRNA candidates from `RNAz`, grey: shuffled sequences.
Enrichments relative to randomly chosen conserved regions:
I: 2.3, II: 1.9, III: 1.4, IV: 1.1

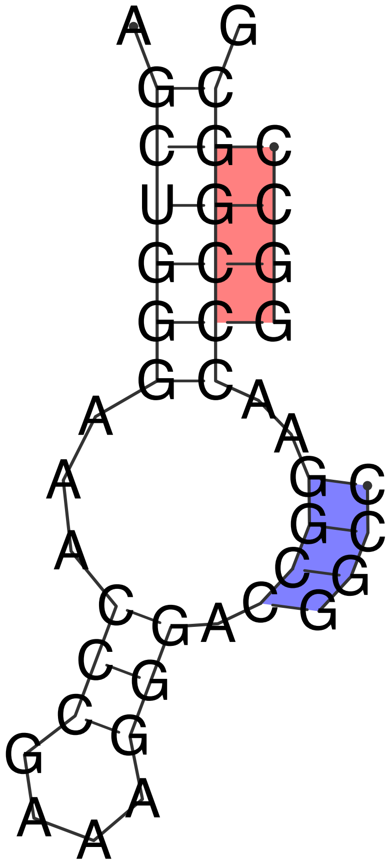
Combining Interaction and Accessibility

Two ingredients for efficient hybridization

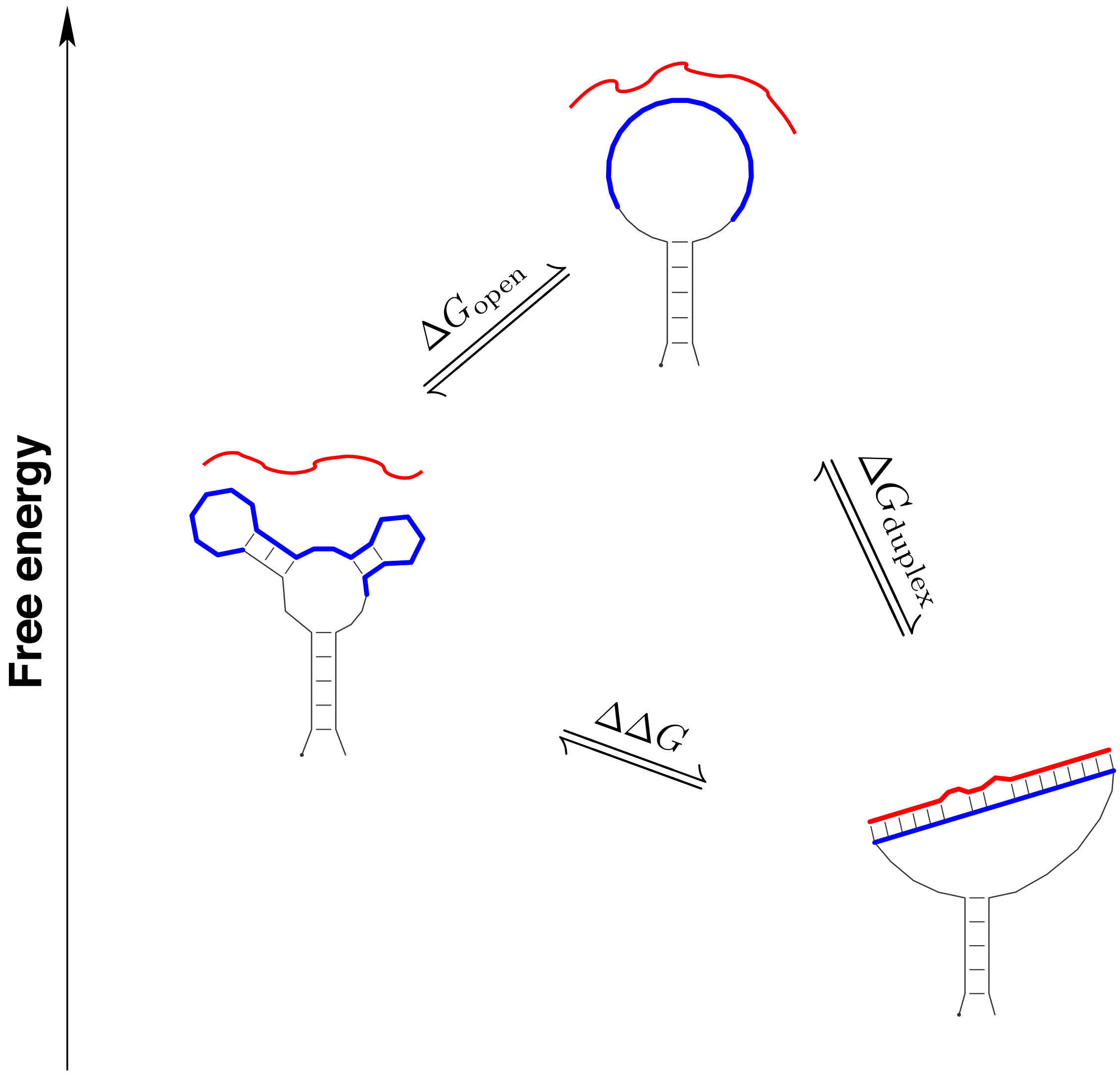
- Complementarity
- Accessibility

How to quantify these?

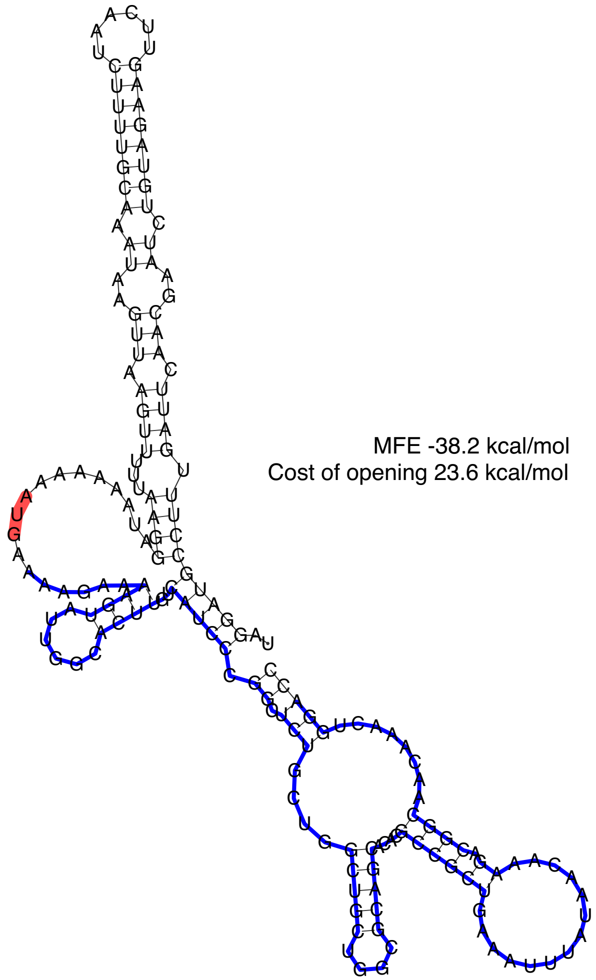
Complementarity → interaction energy
Accessibility → probability to be unpaired



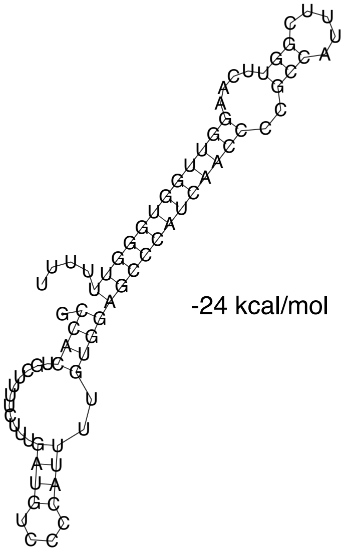
RNA Hybridization as a two Step Process



Example: ompN and RybB



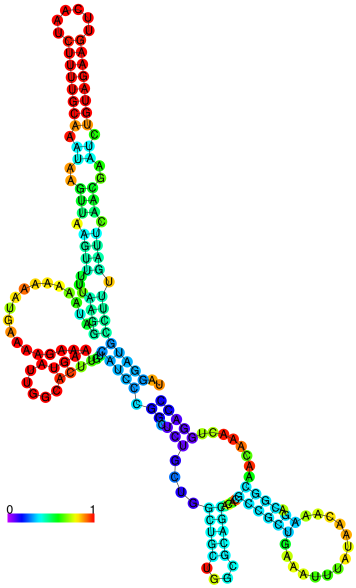
MFE -38.2 kcal/mol
Cost of opening 23.6 kcal/mol

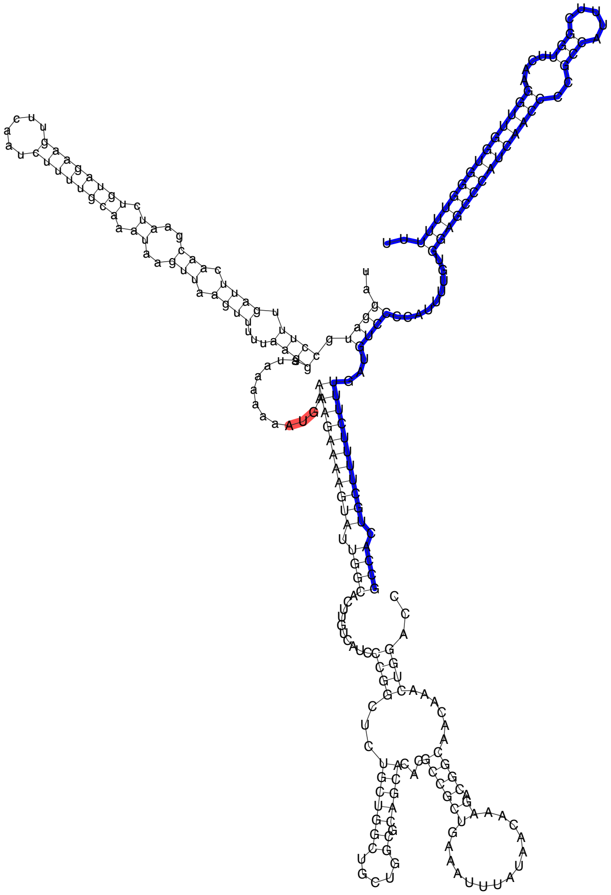


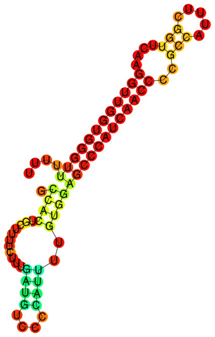
-24 kcal/mol

```
GCCAC-----TGCTTTTCTTTGATGTCCCCATTTT-GTGGA-------GC-CCATCAACCCCGCCATTTCGGTT---CAAG-GTTGGTGGGTTTTTT
 |||      ||||  |||||| |||    ||||| ||||        || ||| || ||  ||    ||||     |||| ||  |||  |||||| -40.30
AGGTCAAACAACGGC-AGAAACAATATT--TAAAGTCGCCGCACACGACGCGGTCGTCGGT-CGTCTCGGCCCTACTGTTCACGGTTATGAAAAGAAACC-3’
```

Example: ompN and RybB

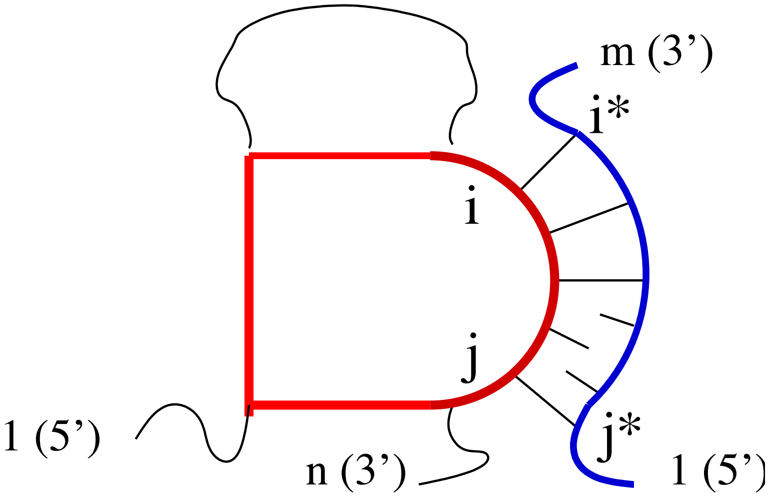






$\Delta G_{\text{open}} = 1.6 + 3.9$ kcal/mol, $\Delta\Delta G = -16$ kcal/mol

The RNAup Approach



- Compute probability that a site at $[i..j]$ is unpaired (equivalent to the energy ΔG_{open} needed to force it open).
- Consider all possible ways of binding to the region $[i..j]$ to compute the interaction energy $\Delta G_{\mathrm{interact}}$
- Total binding energy is the sum of these contributions: $\Delta\Delta G = \Delta G_{\mathrm{open}} + \Delta G_{\mathrm{interact}}$
- Currently, restrict interactions to a single region

Computing Accessibility

ΔG_{open} is equivalent to the probability that the region $[i..j]$ is unpaired in equilibrium $\Delta G_{\mathrm{open}} = -RT \ln P^u[i,j]$

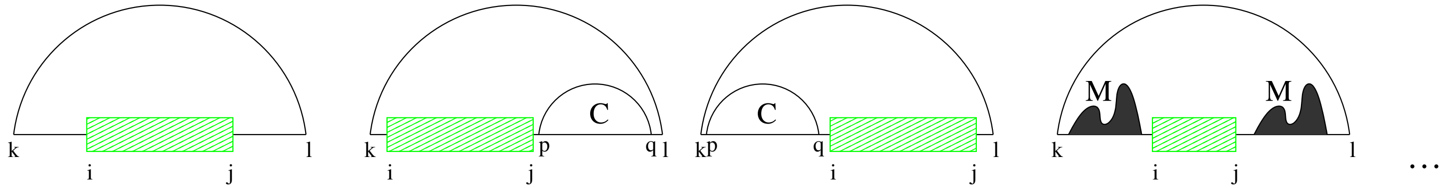
- Constrained folding $\Delta G_{\mathrm{open}} = \Delta G^{\mathrm{constr}} - \Delta G^{\mathrm{free}}$
- Boltzmann sampling, works for short regions only
- Direct computation by modified folding algorithm

Computing Accessibility

ΔG_{open} is equivalent to the probability that the region $[i..j]$ is unpaired in equilibrium $\Delta G_{\mathrm{open}} = -RT \ln P^u[i,j]$

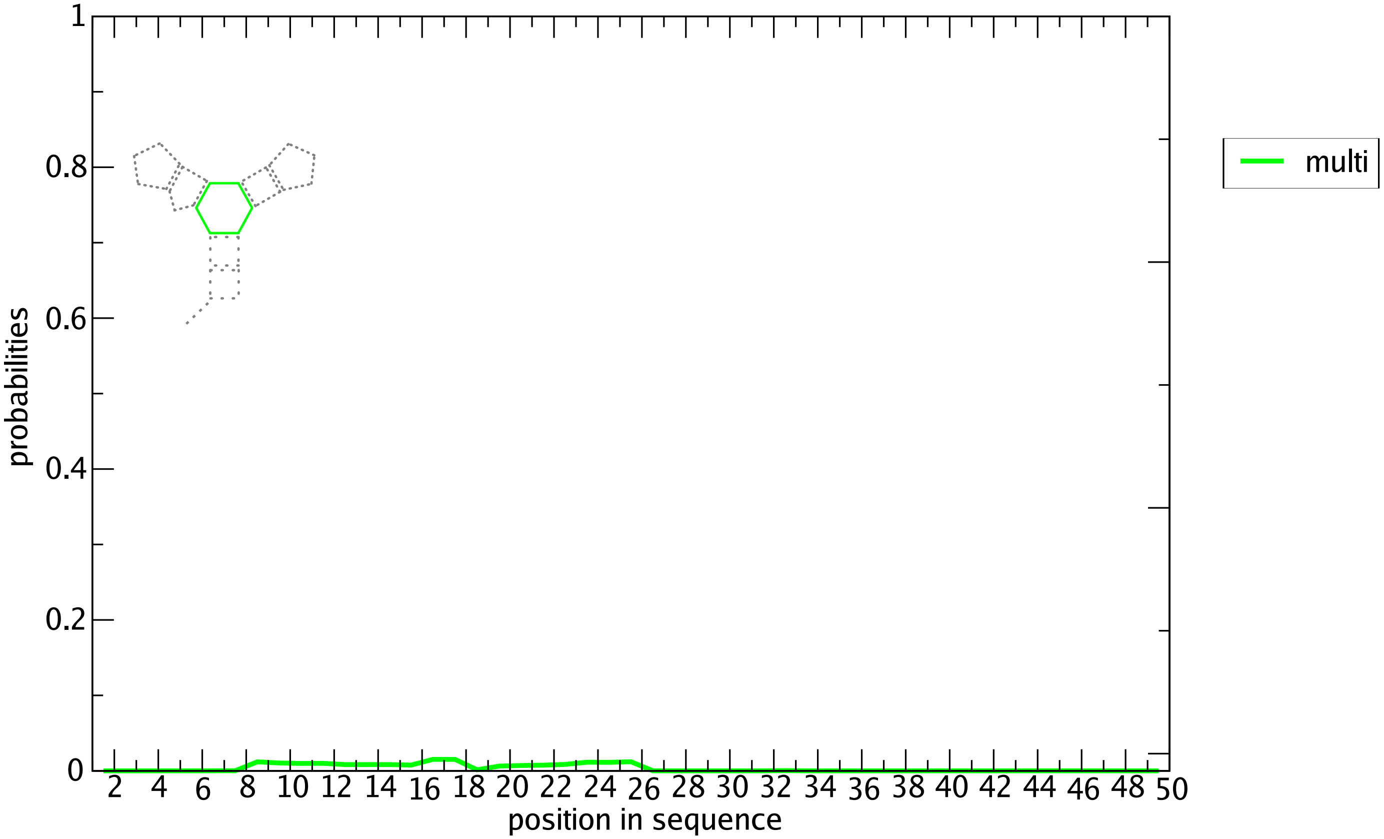
- Constrained folding $\Delta G_{\mathrm{open}} = \Delta G^{\mathrm{constr}} - \Delta G^{\mathrm{free}}$
- Boltzmann sampling, works for short regions only
- Direct computation by modified folding algorithm

$$P^u[i,j] = \frac{Z_{1,i-1} Z_{j+1,n}}{Z_n} + \sum_{h<i,j<l} p_{h,l} \cdot \mathrm{Prob}\left([i,j] | (k,l)\right)$$



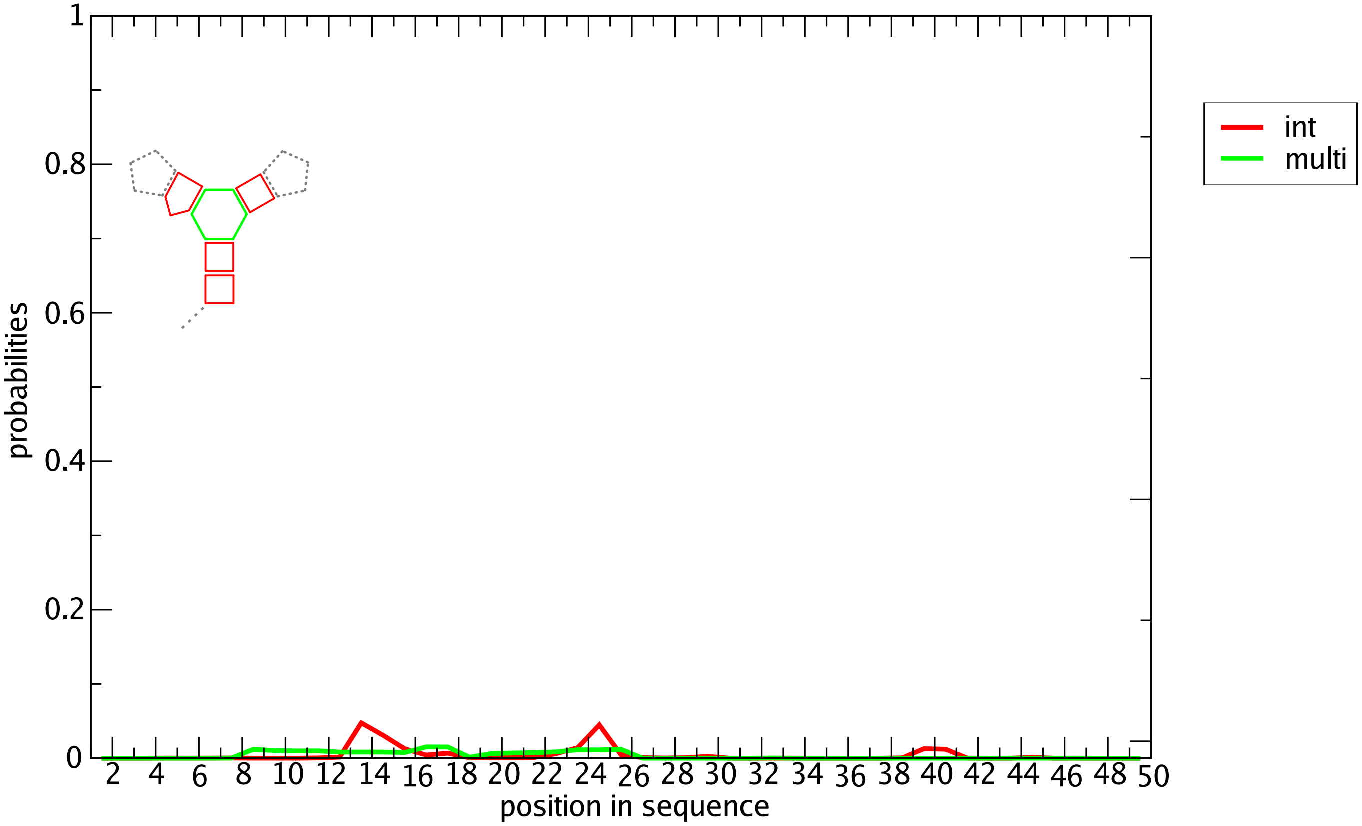
RNAup

Structural Information



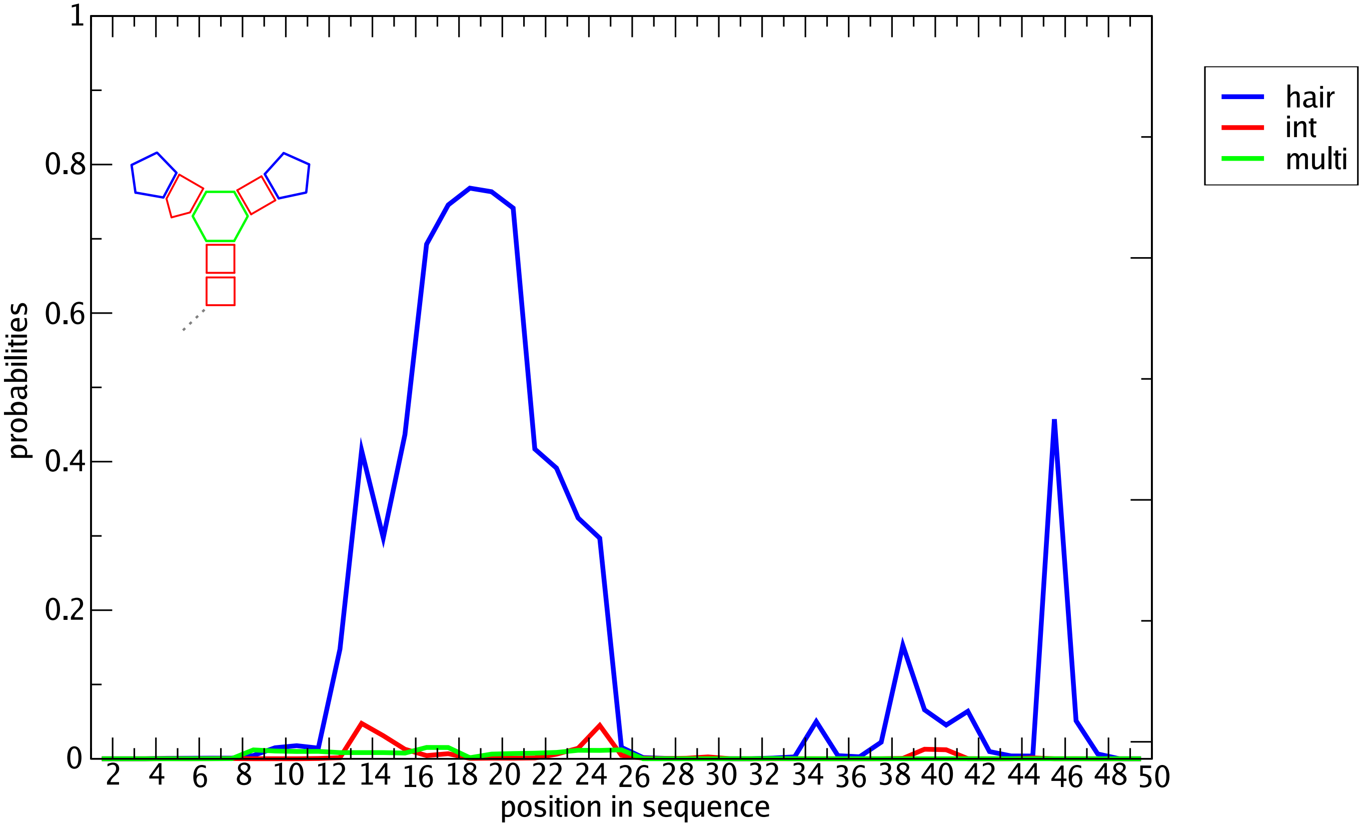
RNAup

Structural Information



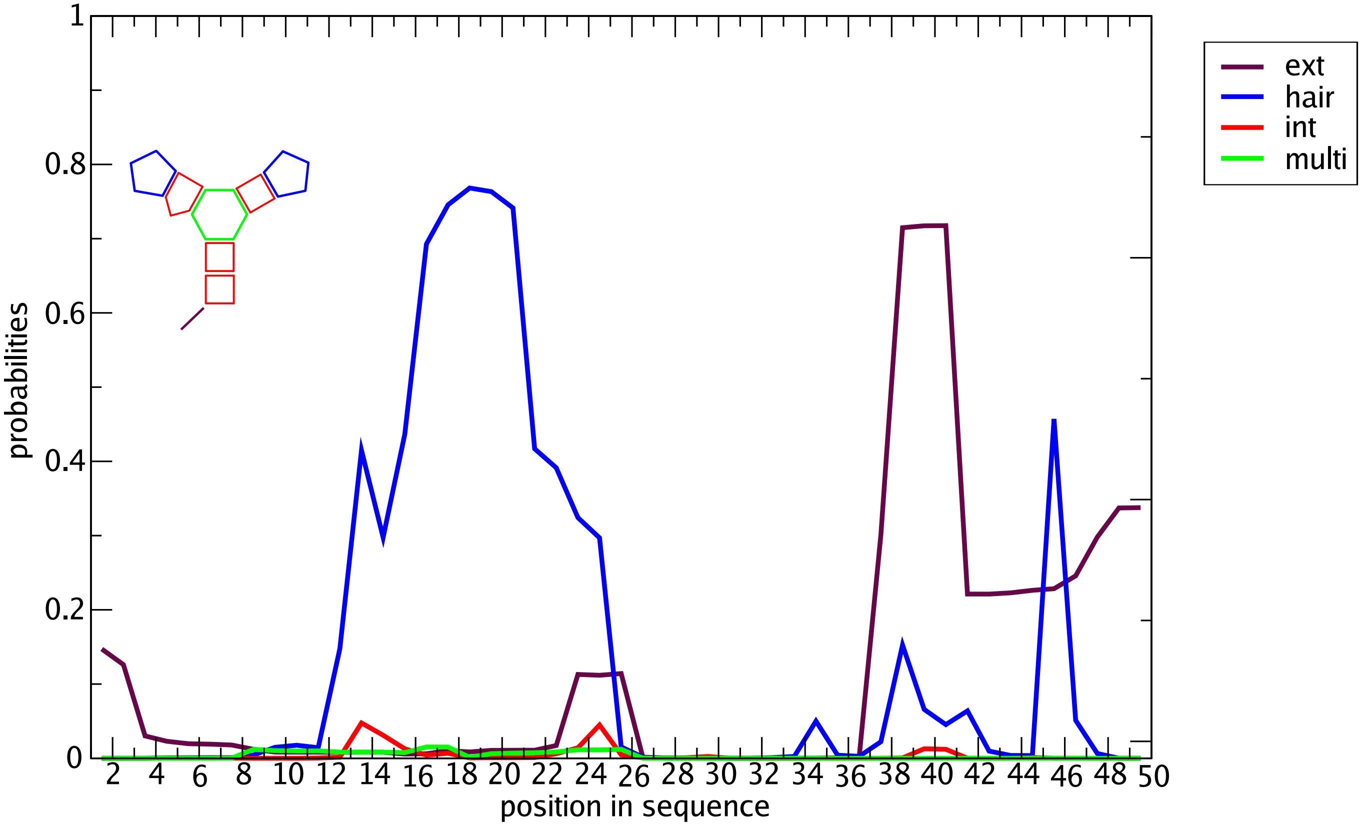
RNAup

Structural Information



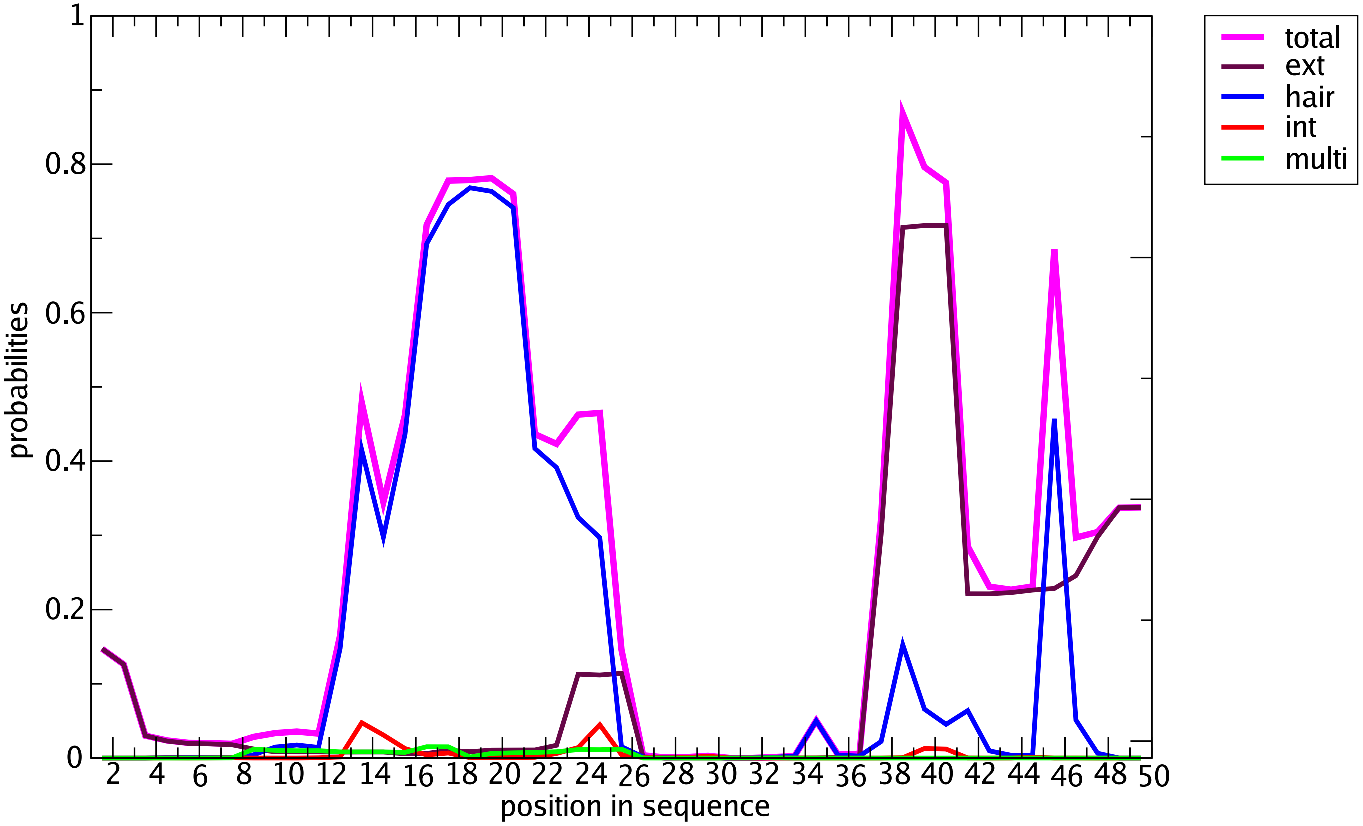
RNAup

Structural Information



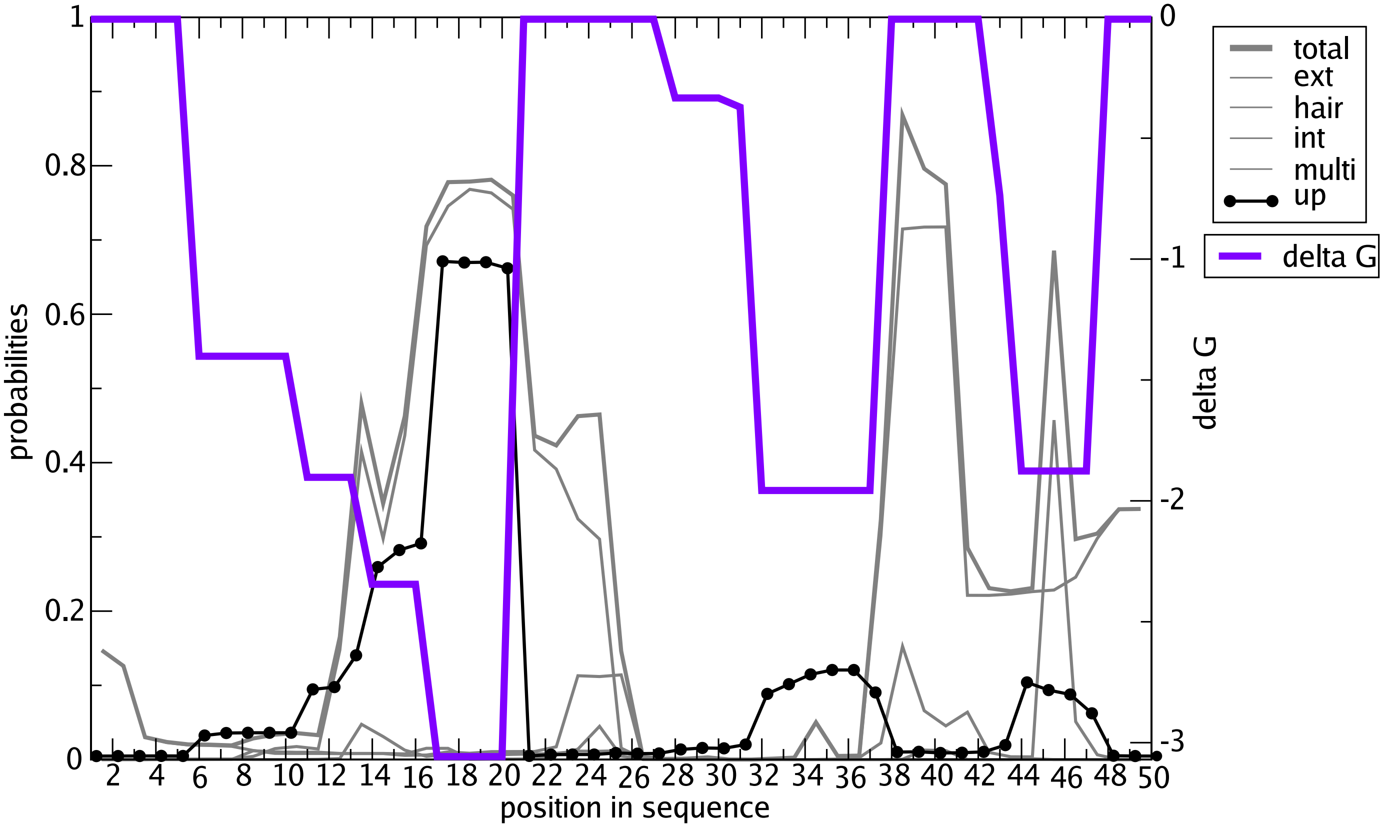
RNAup

Structural Information

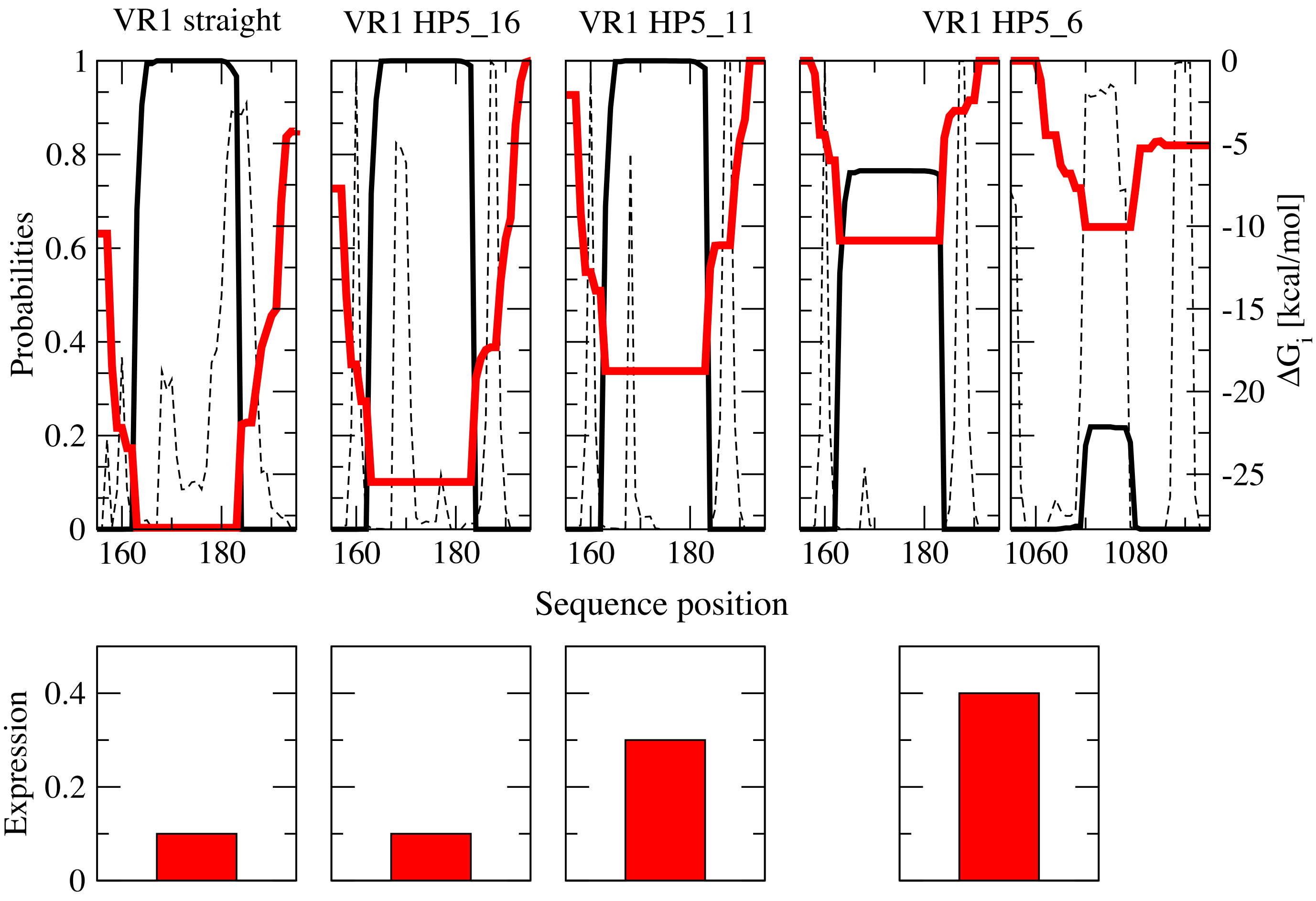


RNAup

Interaction Information



Example: siRNA Binding

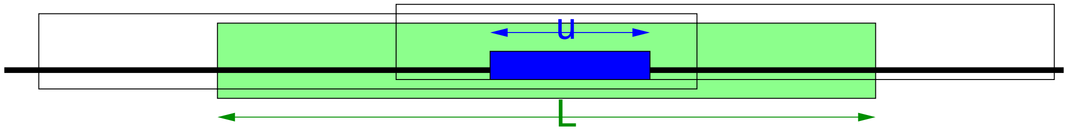


Data taken from Schubert et al 2006

A scanning Version of RNAup

Can we adapt this method for fast searching in large databases?

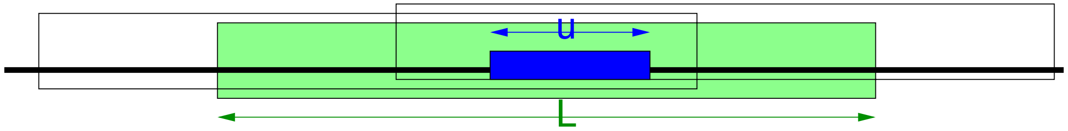
- *Local* folding algorithms can scan very large sequences by restricting the size of local structures to some maximum L.
- `RNAplfold` computes the probability that regions of length u are unpaired by averaging over all windows of length L
- Runtime is linear in the length of the database $\mathcal{O}(n \cdot L^2)$



A scanning Version of RNAup

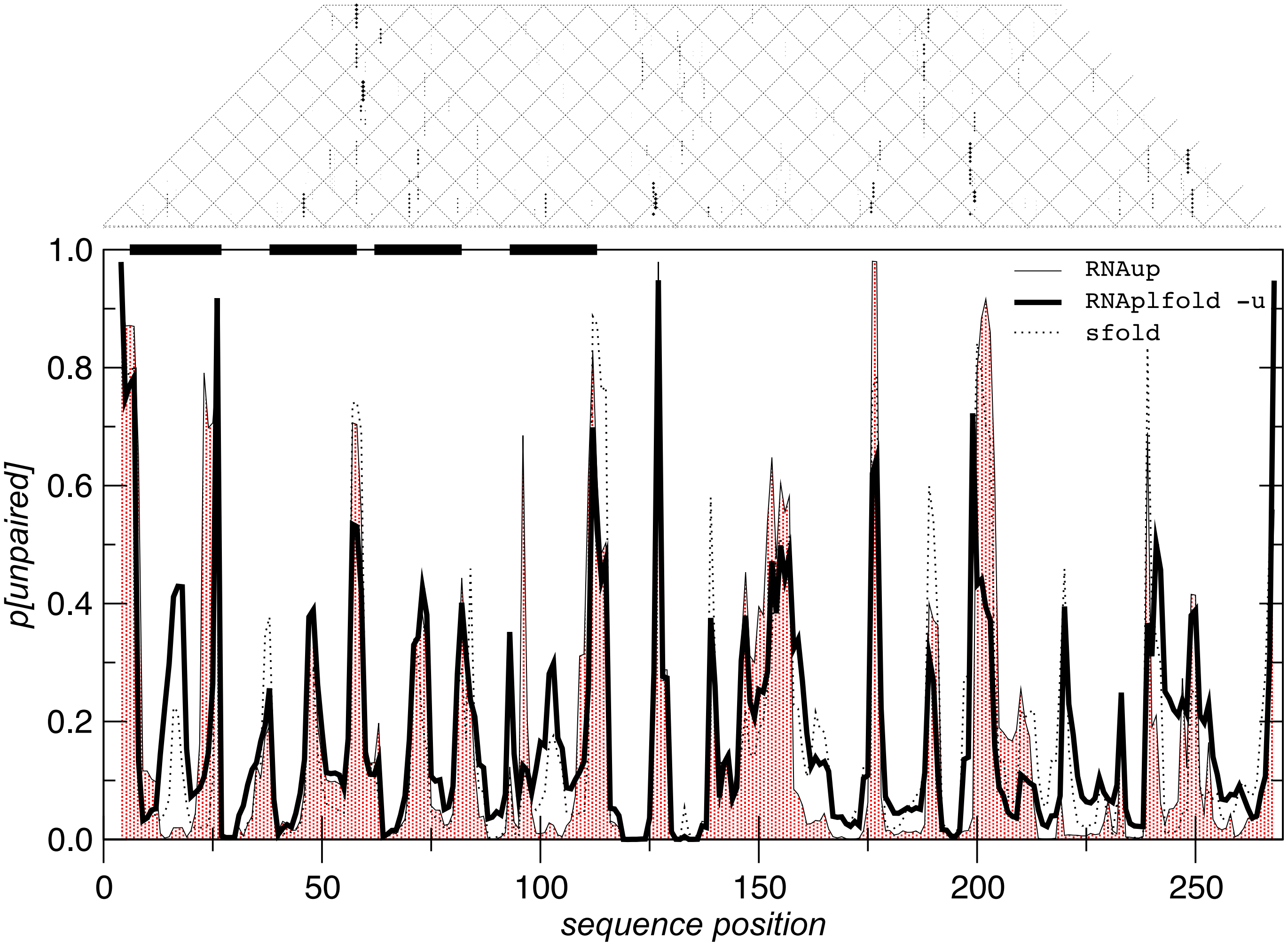
Can we adapt this method for fast searching in large databases?

- *Local* folding algorithms can scan very large sequences by restricting the size of local structures to some maximum L.
- `RNAplfold` computes the probability that regions of length u are unpaired by averaging over all windows of length L
- Runtime is linear in the length of the database $\mathcal{O}(n \cdot L^2)$



Computes average over all windows containing the region

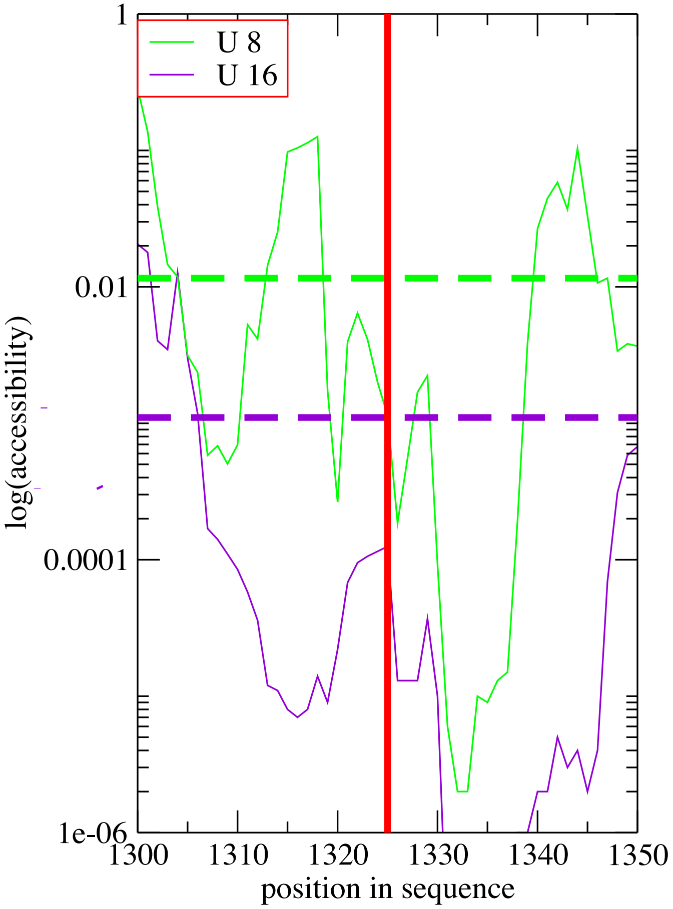
$$\pi^L[i,j] = \frac{1}{L-(j-i)+1} \sum_{u=j-L}^{i} P^{u,L}[i,j]$$



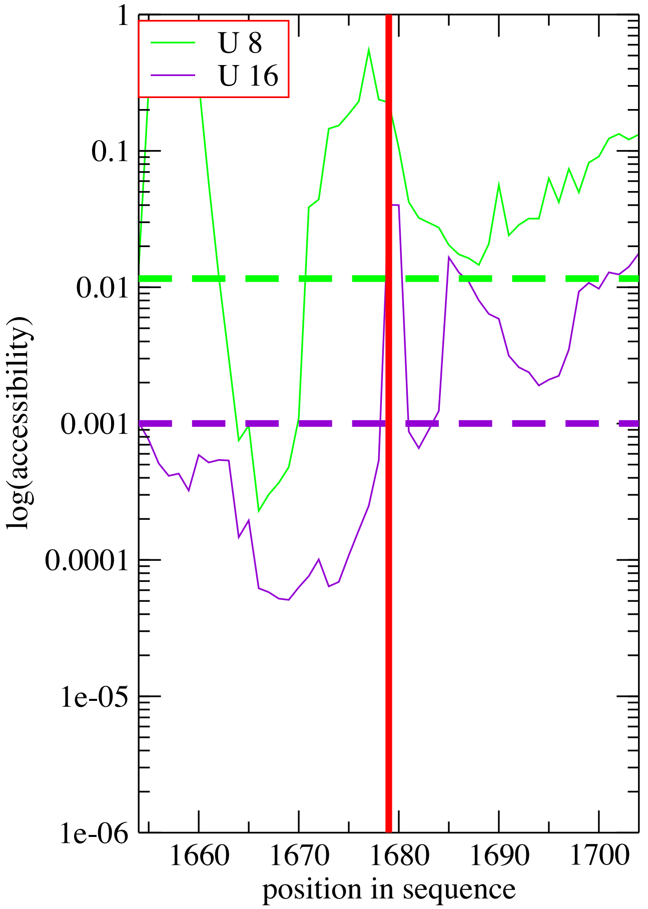
1.0
0.8
0.6
0.4
0.2
0.0
p[unpaired]
0
50
100
150
200
250
sequence position
RNAup
RNAplfold -u
sfold

Accessibility of miRNA targets

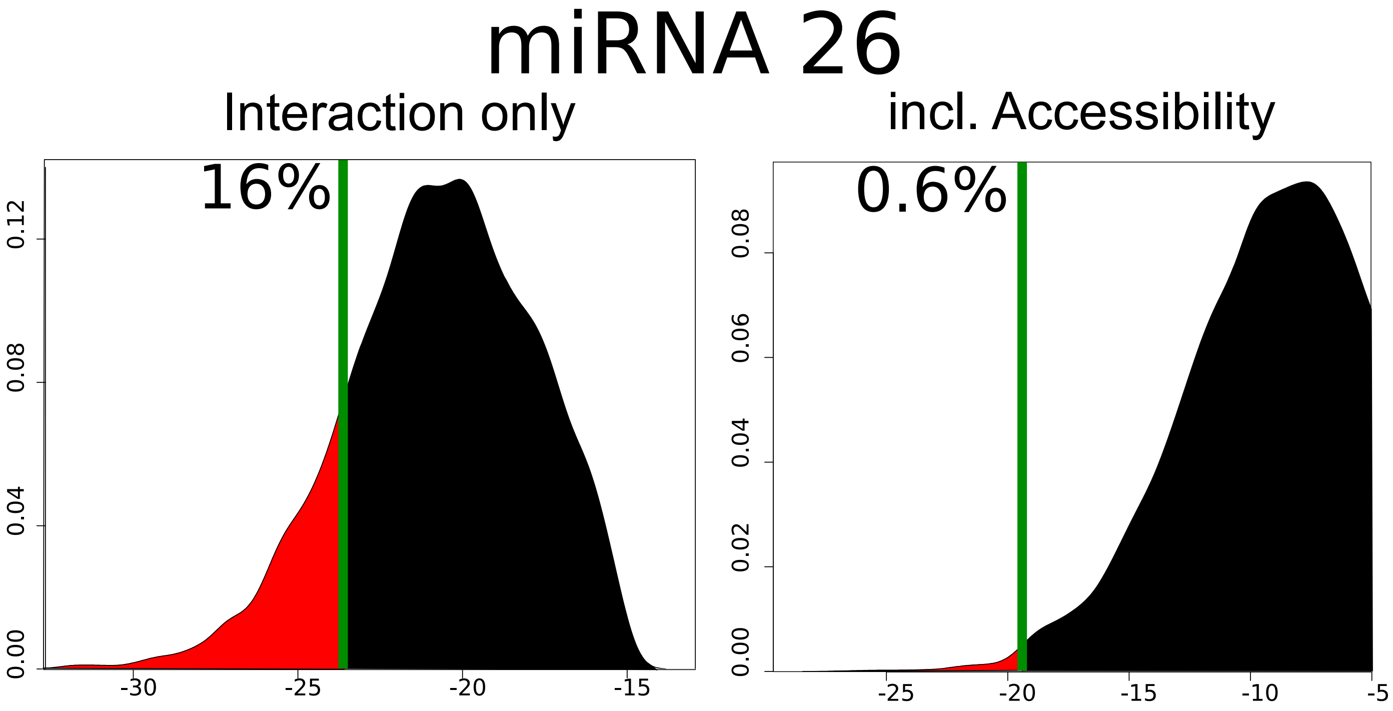
NON WORKING -36.5 kcal/mol



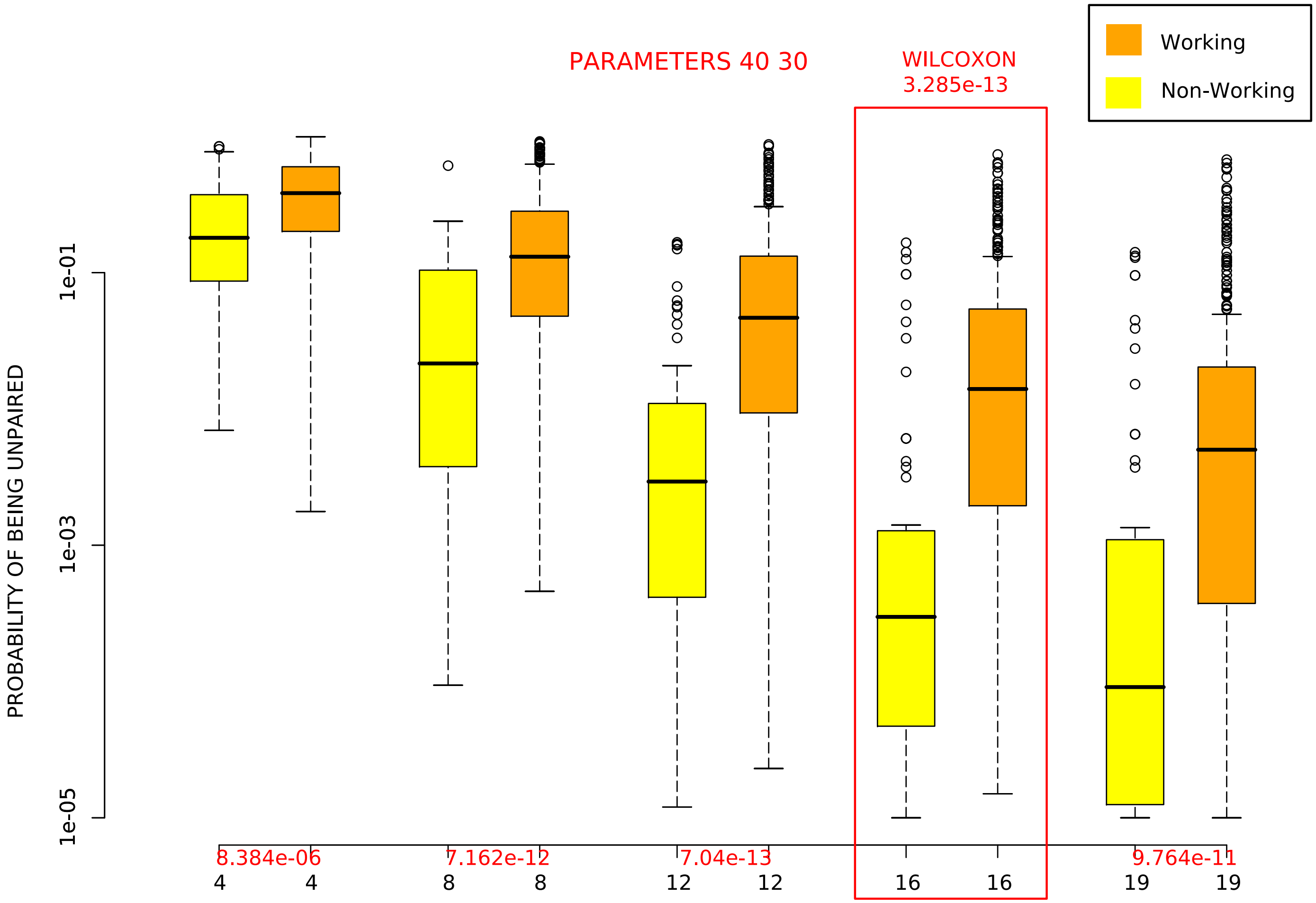
WORKING -28.3 kcal/mol



Accessibility and miRNA targets



Accessibility predicts siRNA efficiency



Data provided by Dharmacon