RNAalifold, bits and pieces

Stephan Bernhart

Institute for Theoretical Chemistry
University of Vienna

Bled, 2007

Outline

1. Locally unpaired structures

2. RNAalifold

Outline

1. Locally unpaired structures

2. RNAalifold

First for something slightly different

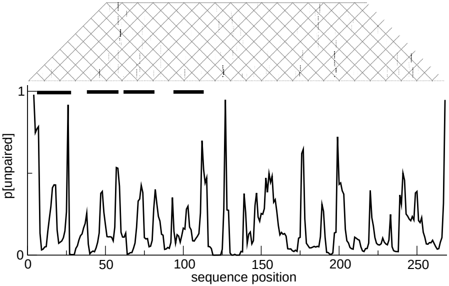
- Last Bled, Jan told Ulli that the unpaired part of RNAup is asked for by lots of scientists.
- They seek information about the accessibility of putative binding sites
- Ulli left the institute, and work on RNAup was not continued.
- However, we decided to create a local version.

RNAplup

- RNAup computes probability of a strech of length l not to form any base pairs, i.e. to be unpaired in the thermodynamic ensemble
- RNAup uses the pair probabilities and matrices computed by RNAfold
- RNAplfold computes these pair probabilities and matrices locally for long sequences.
- As postprocessing step, we compute the average probability of a stretch of length l to be unpaired
- While much slower than RNAplfold, it is still $\mathcal{O}(nW^2)$.

Results

4 artificial binding sites for cxcr4 siRNA, dot plot and probability to be unpaired for a stretch of 4 consecutive bases



Outline

1. Locally unpaired structures

2. RNAalifold

RNAalifold

- Prediction of common secondary structures of Alignments of RNAs
- Bonuses for covariance added
- penalties for non-standard base pairs added
- Every energy evaluation of RNAfold is replaced by a loop over all sequences in alignment

Weighting

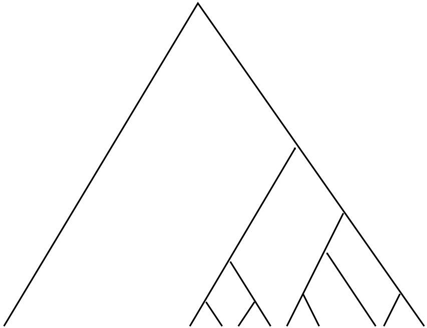
In RNAalifold, every sequence is treated equally important.

- Problem: What if there are two identical sequences?

Weighting

In RNAalifold, every sequence is treated equally important.

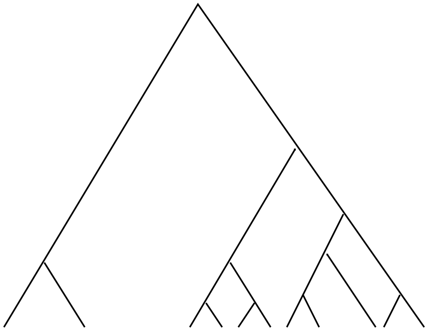
- Problem: Or there is one outlier?



Weighting

In RNAalifold, every sequence is treated equally important.

- Problem: Or there is a “big” and a “small” subtree?

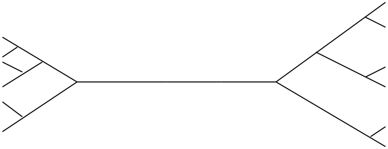


Solution: Weight sequences according to their distance tree.

- Simply weight energy evaluations in loops
- Only useful if tree is highly unbalanced
- Caveats: If there are mistakes in alignment, house numbers will be result.

Weighting Sequences

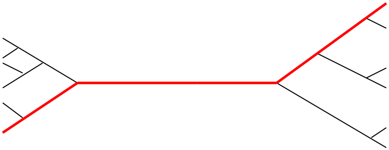
Start with distance weighted trees
First: find root midpoint:



Weighting Sequences

Start with distance weighted trees

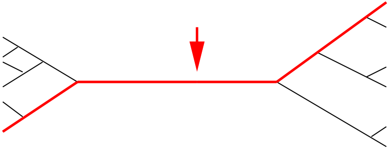
First: find root midpoint:



Weighting Sequences

Start with distance weighted trees

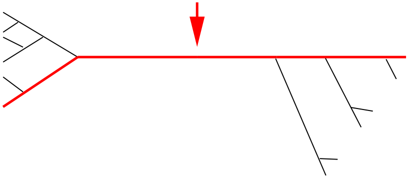
First: find root midpoint:



Weighting Sequences

Start with distance weighted trees

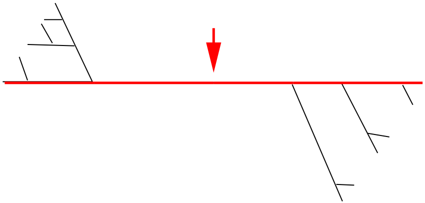
First: find root midpoint:



Weighting Sequences

Start with distance weighted trees

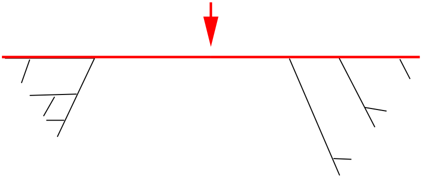
First: find root midpoint:



Weighting Sequences

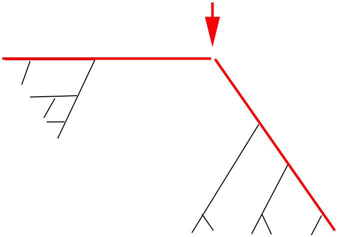
Start with distance weighted trees

First: find root midpoint:



Weighting Sequences

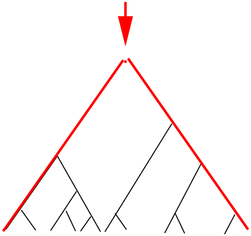
Start with distance weighted trees
First: find root midpoint:



Weighting Sequences

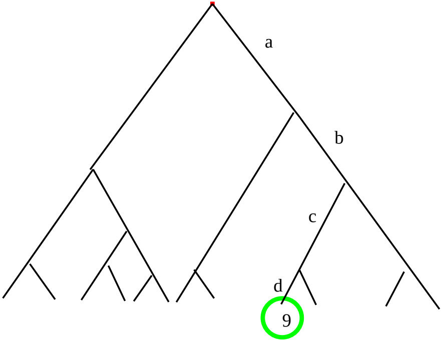
Start with distance weighted trees

First: find root midpoint:



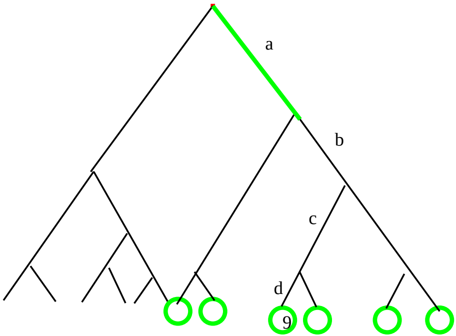
Weighting Sequences

Weight sequences by:
The weights of the edges on the path leaf-root;
Divided by the number of its children leafs *o*:



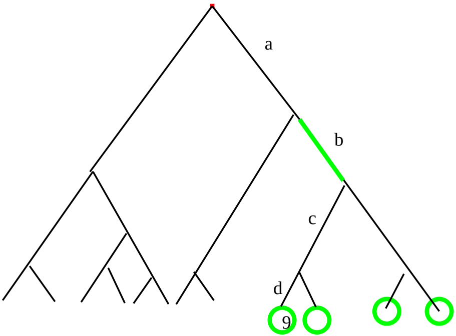
Weighting Sequences

Weight sequences by:
The weights of the edges on the path leaf-root;
Divided by the number of its children leafs *o*:



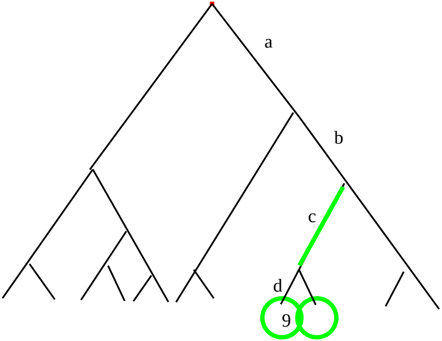
Weighting Sequences

Weight sequences by:
The weights of the edges on the path leaf-root;
Divided by the number of its children leafs *o*:



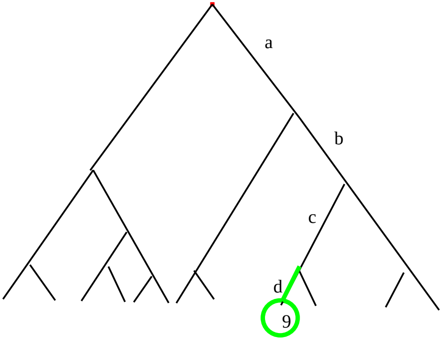
Weighting Sequences

Weight sequences by:
The weights of the edges on the path leaf-root;
Divided by the number of its children leafs *o*:



Weighting Sequences

Weight sequences by:
The weights of the edges on the path leaf-root;
Divided by the number of its children leafs *o*:



Weighting Sequences

Weight sequences by:
The weights of the edges on the path leaf-root;
Divided by the number of its children leafs *o*:

$$W_{\text{seq}} = \sum_{\text{edges}\in\text{path to root}} \frac{w_{\text{edge}}}{o_{\text{edge}}}$$

Weighting Sequences

Weight sequences by:
The weights of the edges on the path leaf-root;
Divided by the number of its children leafs *o*:

Problem:There is a slight difference between the weights of a tree with two identical sequences and the tree where this sequence is counted only once

Weighting Sequences

Weight sequences by:
The weights of the edges on the path leaf-root;
Divided by the number of its children leafs *o*:

Problem:There is a slight difference between the weights of a tree with two identical sequences and the tree where this sequence is counted only once
Solution: Peter??

New energy evaluation

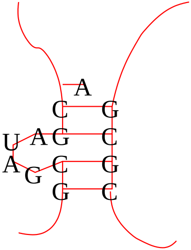
Energy evaluation on alignment sequences includes gaps.

```
===GCGAUAGC-G===GCGC===
===GCG-UCGCAG===GCGC===
===GC----GCAG===GCGC===
```

New energy evaluation

Energy evaluation on alignment sequences includes gaps.

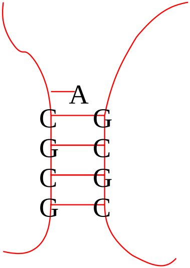
```
===GCGAUAGC-G===GCGC===
===GCG-UCGCAG===GCGC===
===GC----GCAG===GCGC===
```



New energy evaluation

Energy evaluation on alignment sequences includes gaps.

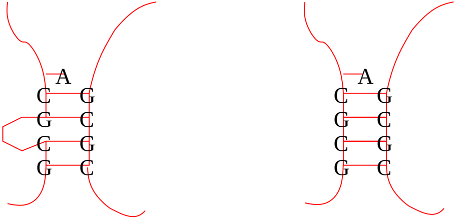
```
===GCGAUAGC-G===GCGC===
===GCG-UCGCAG===GCGC===
===GC----GCAG===GCGC===
```



New energy evaluation

Energy evaluation on alignment sequences includes gaps.

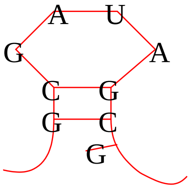
```
===GCGAUAGC-G===GCGC===
===GCG-UCGCAG===GCGC===
===GC----GCAG===GCGC===
```



New energy evaluation

Energy evaluation on alignment sequences includes gaps.

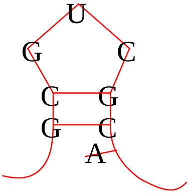
```
===GCGAUAGC-G===GCGC===
===GCG-UCGCAG===GCGC===
===GC----GCAG===GCGC===
```



New energy evaluation

Energy evaluation on alignment sequences includes gaps.

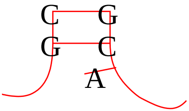
```
===GCGAUAGC-G===GCGC===
===GCG-UCGCAG===GCGC===
===GC----GCAG===GCGC===
```



New energy evaluation

Energy evaluation on alignment sequences includes gaps.

```
===GCGAUAGC-G===GCGC===
===GCG-UCGCAG===GCGC===
===GC----GCAG===GCGC===
```



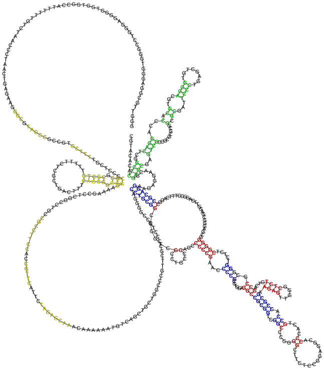
Solution: use ungapped sequences to evaluate the energy.

- Hairpins
- Interior Loops
- Dangles
- Multiloop closing

Easily combined with weighting.

Comparison of Alifolds

Telomerase, unweighted



blue: both wrong, red: both right; green: new right; yellow pseudoknotted mistakes;

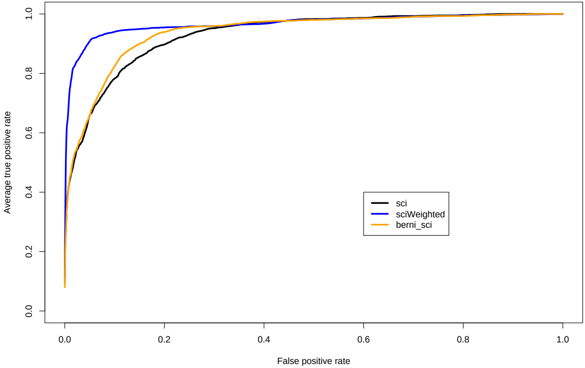
Comparison of Alifolds

BRALIbase

Sequence	TPs (sens.)	FPs (select.)	MCC (approx. corr.)
LSU M Old	478 (57.0)	273 (63.6)	0.602 (60.3)
New	458 (54.6)	270 (62.9)	0.586 (58.8)
Weight	482 (57.4)	295 (62.0)	0.597 (59.7)
LSU.H	429 (51.1)	357 (54.6)	0.528 (52.9)
New	433 (51.6)	356 (54.9)	0.532 (53.2)
Weight	472 (56.3)	299 (61.2)	0.587 (58.7)
SSU.M	383 (81.8)	64 (85.7)	0.837 (83.8)
New	387 (82.7)	62 (86.2)	0.844 (84.4)
Weight	383 (81.8)	64 (85.7)	0.837 (83.8)
SSU.H	314 (67.1)	145 (68.4)	0.677 (67.8)
New	323 (69.0)	139 (69.9)	0.694 (69.5)
Weight	342 (73.1)	117 (74.5)	0.738 (73.8)
RNaseP.H.O.	80 (72.7)	31 (72.1)	0.723 (72.4)
New	78 (70.9)	33 (70.3)	0.705 (70.6)
Weight	80 (72.7)	31 (72.1)	0.723 (72.4)
RNaseP.M.O	88 (80.0)	12 (88.0)	0.838 (84.0)
New	81 (73.6)	27 (75.0)	0.742 (74.3)
Weight	88 (80.0)	13 (87.1)	0.834 (83.6)

Comparison of Alifolds

Artificial Alignment



4 sequences with mpi 95% 1 to reduce it to 65%

Comparison of Alifolds

Speed

	length	factor mfe	factor part. func
2 sequences	73	5.9	2.95
	385	1.73	2.48
	1554	1.48	1.80
	2952	1.41	1.52
	length	factor mfe	factor part. func
5 sequences	73	3.125	2.56
	385	1.63	2.48
	1554	1.55	1.67
	2952	1.55	1.49
	length	factor mfe	factor part. func
9 sequences	73	4.15	2.37
	385	1.78	1.75
	1554	1.70	1.67
	2952	2.07	1.72

Keep it or discard it?

- No positive effect of weighting for AlifoldZ discrimination
- Computationally expansive, but affordable
- Ideas to improve performance appreciated
- How to penalyze too short loops