Combined thermodynamic and evolutionary model for RNA secondary structure prediction

Rolf Backofen[1], Jan Gorodkin[2] and Stefan Seemann[1,2]

[1] Bioinformatics - Inst. of Computer Science, Albert-Ludwigs-University Freiburg
[2] Division of Genetics and Bioinformatics, IBHV, Copenhagen University, Denmark

Bled, Feb 2007

1. Motivation

2. Existing implementations
   - Pfold
   - Vienna RNA Package - RNAfold

3. Combination of two models

4. Application
   - Model performance
   - Alignment dependencies

5. Discussion

Outline

1. Motivation
2. Existing implementations
   - Pfold
   - Vienna RNA Package - RNAfold
3. Combination of two models
4. Application
   - Model performance
   - Alignment dependencies
5. Discussion

Motivation

- non-coding RNA genes provide their functionality through their space conformation
- functional structures are conserved in the evolution
- several independent models to judge consensus secondary structures:
    1. evolutionary model of RNA sequences
    2. probabilistic model for secondary structure
    3. thermodynamic model for folding energy

Outline

1. Motivation
2. Existing implementations
   - Pfold
   - Vienna RNA Package - RNAfold
3. Combination of two models
4. Application
   - Model performance
   - Alignment dependencies
5. Discussion

Pfold

Probabilistic evolutionary model[1], which consists of

1. stochastic evolutionary model (T)
   - $\Pr_{\text{paired}}[\vec{A}^i \vec{A}^{i+j-1} | T]$
   - $\Pr_{\text{single}}[\vec{A}^i | T]$
2. SCFG-based probabilistic model (M) for secondary structure
   - production rules:

$$
\begin{array}{ll}
\mathrm{S} \rightarrow LS \mid L & \textit{(produces loops)} \\
\mathrm{F} \rightarrow dFd \mid LS & \textit{(produces stems)} \\
\mathrm{L} \rightarrow s \mid dFd & \textit{(single base or new stem)}
\end{array}
$$

[1] Knudsen B, Hein J (2003) Pfold: RNA secondary structure prediction using stochastic context-free grammars. Nucleic Acids Res. 31(13):3423-8.

Pfold

Most probable consensus structure σ can be determined by maximize $\Pr[\sigma|A, T, M]$:

$$\sigma^{MAP} = \underset{\sigma}{\operatorname{argmax}} \Pr[A|\sigma, T, M] \Pr[\sigma|T, M]$$

This can be solved using the CYK-algorithm.

Vienna RNA Package - RNAfold[2]

The partition function measures the probability of a secondary structure σ in thermodynamic equilibrium:

$$P_\sigma = \frac{Z_\sigma}{Z} = \frac{e^{-\frac{\Delta G_\sigma}{RT}}}{\sum_{S \in \Omega} e^{-\frac{\Delta G_S}{RT}}}$$

It can be calculated the density probability of

- base pairs $Pr[(A_u^i, A_u^{i+j-1}) \mid s_u]$
- unpaired bases $\mathrm{Pr}_{\text{unpaired}}[A_u^{i+j-1}|s_u]$

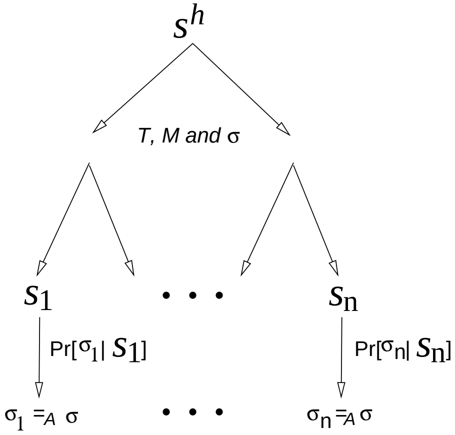
[2] I.L. Hofacker, W. Fontana, P.F. Stadler, S. Bonhoeffer, M. Tacker, P. Schuster (1994) Fast Folding and Comparison of RNA Secondary Structures. Monatshefte f. Chemie 125: 167-188

Outline

1. Motivation
2. Existing implementations
   - Pfold
   - Vienna RNA Package - RNAfold
3. Combination of two models
4. Application
   - Model performance
   - Alignment dependencies
5. Discussion

Extended model

Combination of probabilistic evolutionary information with thermodynamic parameters of the standard energy minimization model:



Probabilistic evolutionary model

Def. of probability of structure σ:

$$\Pr[A|\sigma, T, M] \Pr[\sigma|T, M] = \text{prob}_{M,\tau_M(\sigma)}(r, A)$$

Recursively definition for the probabilistic evolutionary model:

$$\text{prob}_{M,\tau_M(\sigma)}(n, A) =$$

$$\Pr[\text{rule}(n)|M]$$
$$\times \prod_{\ell=1}^{k} \text{prob}_{M,\tau_M(\sigma)}(n_\ell, A)$$
$$\times \begin{cases} \Pr_{\text{paired}}[\vec{A}^i \vec{A}^{i+j-1}|T] & \text{if rule}(n) = F \rightarrow dFd \\ & \text{or rule}(n) = L \rightarrow dFd \\ \Pr_{\text{single}}[\vec{A}^i|T] & \text{if rule}(n) = L \rightarrow s \\ 1 & \text{else} \end{cases}$$

PE thermodynamic model

$$
\begin{aligned}
&\mathrm{prob}_{M,\tau_M(\sigma)}(n, A) = \\
&\qquad \Pr[\mathrm{rule}(n) | M] \\
&\quad \times \prod_{\ell=1}^{k} \mathrm{prob}_{M,\tau_M(\sigma)}(n_\ell, A) \\
&\quad \times \begin{cases}
\Pr_{\mathrm{paired}}[\vec{A}^i \vec{A}^{i+j-1} | T] & \text{if rule}(n) = L \longrightarrow dFd \\
\times \prod_{u=1}^{n} \begin{cases} Pr[(A_u^i, A_u^{i+j-1}) \mid s_u] & \text{if } bp(s_u^i, s_u^{i+j-1}) \\ \Pr_{\mathrm{unpaired}}[A_u^i | s_u] \times \Pr_{\mathrm{unpaired}}[A_u^{i+j-1} | s_u] & \text{if } \neg bp(s_u^i, s_u^{i+j-1}) \end{cases} & \\
& \\
\Pr_{\mathrm{paired}}[\vec{A}^i \vec{A}^{i+j-1} | T] & \text{if rule}(n) = F \longrightarrow dFd \\
\times \prod_{u=1}^{n} \begin{cases} Pr[(A_u^i, A_u^{i+j-1}) \mid (A_u^{i-1}, A_u^{i+j}), s_u] & \text{if } bp(s_u^i, s_u^{i+j-1}) \\ \Pr_{\mathrm{unpaired}}[A_u^i | s_u] \times \Pr_{\mathrm{unpaired}}[A_u^{i+j-1} | s_u] & \text{if } \neg bp(s_u^i, s_u^{i+j-1}) \end{cases} & \\
& \\
\Pr_{\mathrm{single}}[\vec{A}^i | T] \times \prod_{u=1}^{n} \Pr_{\mathrm{unpaired}}[A_u^i | s_u] & \text{if rule}(n) = L \longrightarrow s \\
& \\
1 & \text{else}
\end{cases}
\end{aligned}
$$

Gaps

Treating gaps is a general problem in biological sequence analysis:

- alignment columns with $\geq 25\%$ gaps are removed (like in `Pfold`)
- sequence depended probabilities are calculated without gaps
- gap probabilities are estimated as geometric mean of probabilities in the appropriate column

Outline

1 Motivation

2 Existing implementations
- Pfold
- Vienna RNA Package - RNAfold

3 Combination of two models

4 Application
- Model performance
- Alignment dependencies

5 Discussion

Model performance in U1

Comparison of: `PETfold`, `RNAalifold`, `Pfold`
Test data: Rfam seed alignment of U1 spliceosomal RNA
(av.id= 40.6%; max.id = 50%; #seq = 5)

set rules	set tree	set seq	log2 prob	sensitivity [%]	specificity [%]
0	0	0	0	0	0
0	0	1	-1089	75	64
0	1	0	-12	95	90
0	1	1	-1381	72.5	81
1	0	0	-37	0	0
1	0	1	-1276	72.5	74
1	1	0	-106	90	90
1	1	1	-1540	70	82
`RNAalifold`				62.5	86
`Pfold`				95	90

Optimal consensus structure

mir-9/mir-79 microRNA precursor family

```
Rfam        ...<<<<<<<<<<<<<<<..<.<<<...............>.>>>..>>>>>>>>>>>>>>>...
Pfold       ...(((((((((((((((..(.(((...............).)))..)))))))))))))))...
RNAalifold  ...(((((((((((((((.............................)))))))))))))))...
PETfold     ...(((((((..(.(((((.(-((((.((-.....--))))-))).))))).)..)))))))...
```

sensitivity = 100%|79%|84%; specificity = 100%|100%|80%

U98 small nucleolar RNA

```
Rfam        .......<<<<<.......<<<<<<<........>>>>>>>...........>>>>>......
Pfold       ....................(((((..........)))))).......................
RNAalifold  .....(((..(((...((.((((((((......)))))))).))(((((..............
PETfold     ..--((.((.......(((((((((((......)))))))...))))......)).)).....

Rfam        .......<<<<<<<...........<<<...........>>>................>>>>>>>.......
Pfold       ........................................................................
RNAalifold  (((((.(((......))).)))))......(((......))))))))).....)))...)))...........
PETfold     ......((((((((.((((.....))))..(((......)))..(((.--..)))...)))))))..)....
```

sensitivity = 23%|55%|73%; specificity = 100%|37.5%|52%

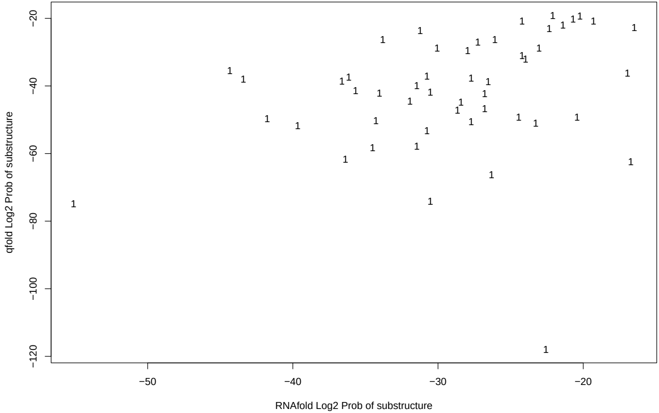
Small nucleolar RNA Z159/U59

```
Rfam        ..<<<<.............~........<<<<<<<<<..~..>>>>>>>>>..............~..>>>>..
Pfold       ((((((.............~........(((((((((..~..)))))))))......((...)).~..))))))
RNAalifold  ..((((.............~........(((((((((..~..))))))))).......(....)..~..))))..
PETfold     (.(((((.....((.((..~..))..))(..(..(((..~..)))..)..)...)-.(....)..~..)))).)
```

sensitivity = 100%|100%|69%; specificity = 76%|93%|53%

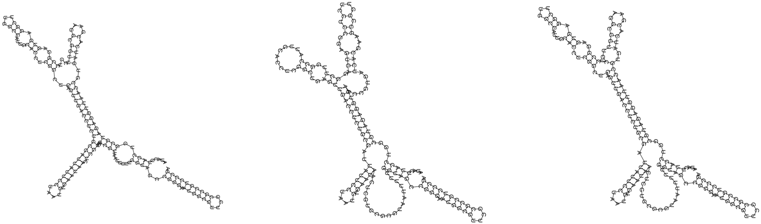
RNAfold vs PETfold

50 suboptimal structures of the U1 spliceosomal RNA *AE003745*



RNAfold vs PETfold

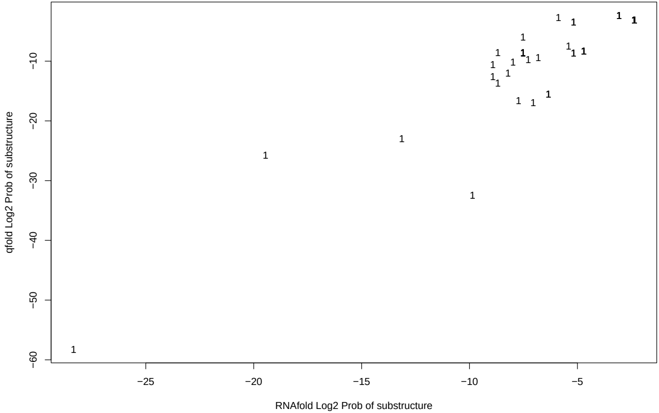
3 most probable structures of the U1 spliceosomal RNA *AE003745* predicted by `RNAfold`



Observation: `PETfold` predicts more probable basepairs as single bases in multiloops

RNAfold vs PETfold

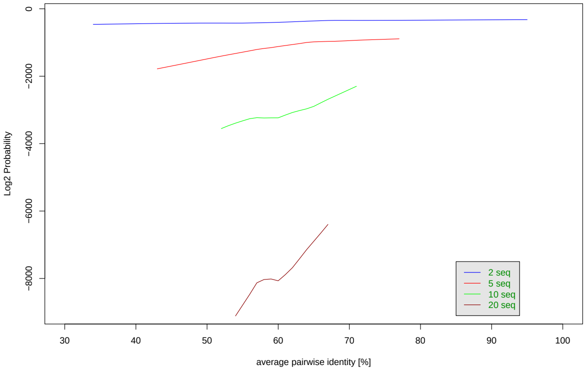
50 suboptimal structures of the mir-9/mir-79 microRNA *Z81467*



Alignment dependencies in U1

Influence of average pairwise identity of an alignment on the optimal structure probability of our model:

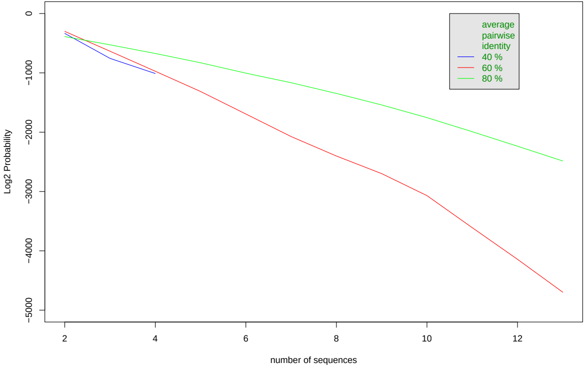
- Test data: Rfam seed alignment of U1 spliceosomal RNA
- the number of sequences in the alignment is fixed



Alignment dependencies in U1

Influence of sequence number in an alignment on the optimal structure probability of our model:

- Test data: Rfam seed alignment of U1 spliceosomal RNA
- the average pairwise identity of the alignment is fixed



Outline

1. Motivation
2. Existing implementations
   - Pfold
   - Vienna RNA Package - RNAfold
3. Combination of two models
4. Application
   - Model performance
   - Alignment dependencies
5. Discussion

Problems and further work

- prior probability of structures without extra information content (uniform distribution)
- low number of basepairs in large alignments (also `Pfold` has problems with large input)
- basepairs with higher probability as single bases in multiloops
- dangling ends are not considered until now (usage of `RNAfold -p2`)

Extended grammar considering stacking probabilities

Modified grammar: single bases are considered in their structural context by changed F rule

$$\begin{aligned} S &\rightarrow LS \mid L \\ F &\rightarrow dFd \mid dFdS \mid sS \\ L &\rightarrow s \mid dFd \end{aligned}$$

Sequence stacking probabilities are estimated by `RNAfold` constraints.

Thank you!!!

:-)