

An MCMC-Method for Sampling RNA Secondary Structures with Pseudoknots

Dirk Metzler

Johann Wolfgang Goethe-Universität Frankfurt am Main
Fachbereich Informatik und Mathematik

Workshop on Bayesian Phylogeny
Budapest, 25.–29. June 2008



D. Metzler, M. Nebel (2007) Predicting RNA Secondary Structures with Pseudoknots by MCMC Sampling
J. Math. Biol., DOI 10.1007/s00285-007-0106-6



D. Metzler, M. Nebel (2006) An MCMC-Method for Sampling RNA Secondary Structures with Pseudoknots
Proceedings of the IASTED International Conference COMPUTATIONAL AND SYSTEMS BIOLOGY November 13–14, 2006, Dallas, TX, USA

Outline

- 1 RNA-Folding with Stochastic Context-Free Grammars
- 2 RNA-Folding with Pseudoknots
- 3 Bayes-Sampling of RNA-Structures with Pseudoknots
- 4 Experiments with Data
- 5 Conclusions

Outline

- 1 RNA-Folding with Stochastic Context-Free Grammars
- 2 RNA-Folding with Pseudoknots
- 3 Bayes-Sampling of RNA-Structures with Pseudoknots
- 4 Experiments with Data
- 5 Conclusions

tmRNA of Escherichia Coli

```
GGGGCUGAUUCUGGAUUCGACGGGAUUUGCGAAACCCAAGGUGCAUGCCGAGGGGCGGUUGGCCUCGUAAAAGC  
CGCAAAAAAUAGUCGCAAACGACGAAAACUACGCUUUAGCAGCUUAAUAACUGCUUAGAGCCUCUCUCCUAG  
CCUCCGCUCUUAGGACGGGGAUCAAGAGAGGUCAAACCCAAAAGAGAUCGCGUGGAAGCCUGCCUGGGGUUGAA  
GCGUUAAAACUAAUCAGGCUAGUUUGUUAGUGGCGUGUCCGCCAGCUGGCAAGCGAAUGUAAAGACUGACU  
AAGCAUGUAGUACCGAGGAUGUAGGAAUUUCGGACGCGGGUUC AACUCCGCCAGCUCACCA
```

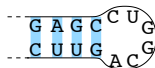
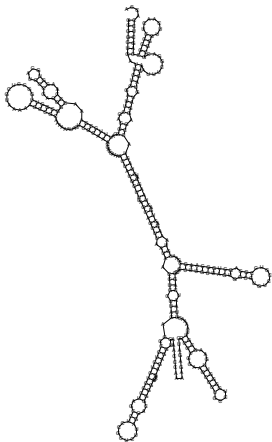
tmRNA of Escherichia Coli

GGGCUGAUUCUGGAUUCGACGGGAUUUGCGAAACCCAAGGUGCAUGCCGAGGGGCGGUUGGCCUCGUAAAAGC
 CGCAAAAAUAGUCGCAAACGACGAAAACUACGCUUUAGCAGCUAAUAACUGCUUAGAGCCUCUCUCCUAG
 CCUCCGCUCUUAGGACGGGGAUCAAGAGAGGUCAAACCCAAAAGAGAUCGCGUGGAAGCCUGCCUGGGGUUGAA
 GCGUUAACCUAAUCAGGCUAGUUUGUAGUGGCGUGUCCGCCAGCUGGCAAGCGAAUGUAAAGACUGACU
 AAGCAUGUAGUACCGAGGAUGUAGGAAUUUCGGACGCGGGUUCAACUCCGCCAGCUCACCA



tmRNA of Escherichia Coli

```
GGGCUGAUUCUGGAUUCGACGGGAUUUGCGAAACCCAAGGUGCAUGCCGAGGGGCGGUUGGCCUCGUAAAAAGC
CGCAAAAAAUAGUCGCAAACGACGAAAAACUACGCUUUAGCAGCUUAAUAACUGCUUAGAGCCUCUCUCCUAG
CCUCCGCUCUUAGGACGGGGAUCAAGAGAGGUCAAACCCAAAAGAGAUCGCGUGGAAGCCUCUGCCUGGGGUUGAA
GCGUAAAAACUAAUCAGGCUAGUUUGUAGUGGCGUGUCCGUCCGACGUGGCAAGCGAAUGUAAAGACUGACU
AAGCAUGUAGUACCGAGGAUGUAGGAAUUUCGGACGCGGGUUC AACUCCGCCAGCUCCACCA
```



RNAfold

implementation of
Zuker's algorithm in
Vienna RNA Package

Zuker (1989)

Zuker et al. (1999)

Stochastic Context-Free Grammar (SCFG)

- terminal symbols

A, C, G, U

- non-terminal symbols

S, L, F

- rules with probabilities

$$S \begin{array}{l} \nearrow \\ \rightarrow \end{array} \begin{array}{l} LS \\ L \end{array}$$

$$F \begin{array}{l} \nearrow \\ \rightarrow \end{array} \begin{array}{l} xLSy \\ xFy \end{array}$$

$$L \begin{array}{l} \nearrow \\ \rightarrow \end{array} \begin{array}{l} x \\ axFyb \end{array}$$

with $x, y, a, b \in \{A, C, G, U\}$

- not permitted: $xFy \rightarrow xLy$

Stochastic Context-Free Grammar (SCFG)

- terminal symbols

A, C, G, U

- non-terminal symbols

S, L, F

- rules with probabilities

$$S \begin{array}{l} \nearrow \\ \rightarrow \end{array} \begin{array}{l} LS \\ L \end{array}$$

$$F \begin{array}{l} \nearrow \\ \rightarrow \end{array} \begin{array}{l} xLSy \\ xFy \end{array}$$

$$L \begin{array}{l} \nearrow \\ \rightarrow \end{array} \begin{array}{l} x \\ axFyb \end{array}$$

with $x, y, a, b \in \{A, C, G, U\}$

- not permitted: $xFy \rightarrow xLy$

generating RNA-structure with SCFG

s

generating RNA-structure with SCFG

S

S→LS

generating RNA-structure with SCFG

LLLLLLLLLLLLLLLLLLLLLLLLL S

S->LS

generating RNA-structure with SCFG

LLLLLLLLLLLLLLLLLLLLLLLLL S

S->x

L->x

generating RNA-structure with SCFG

acg**L**uaagau**L**uau**L**ggcauu**a**

S->**x**

L->**x**

generating RNA-structure with SCFG

acg**L**uaagau**L**uau**L**ggcauu a

L->axFyb

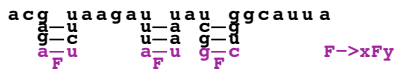
generating RNA-structure with SCFG

acg uaagau uau ggcauua
ga cu ua aa gc ca
F F F F
L->axFyb

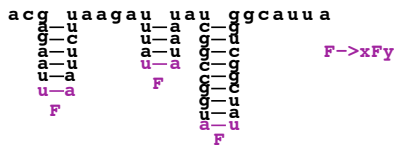
generating RNA-structure with SCFG

acg uaagau uau ggcaua
gac uaa gau uau gagcaua
F F F F → xFy

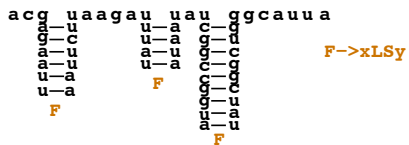
generating RNA-structure with SCFG



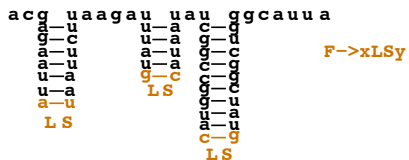
generating RNA-structure with SCFG



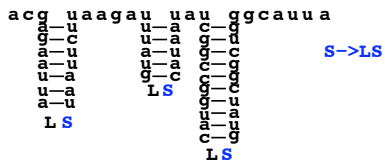
generating RNA-structure with SCFG



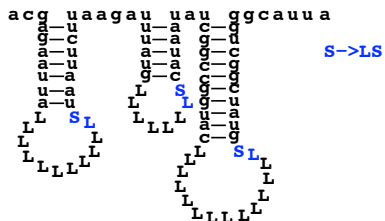
generating RNA-structure with SCFG



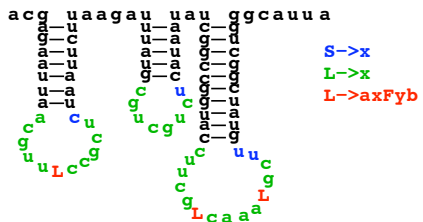
generating RNA-structure with SCFG



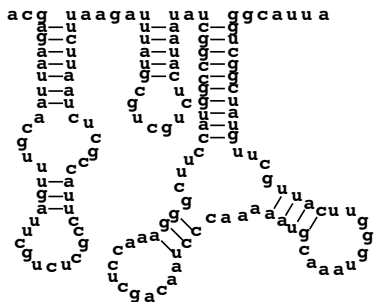
generating RNA-structure with SCFG



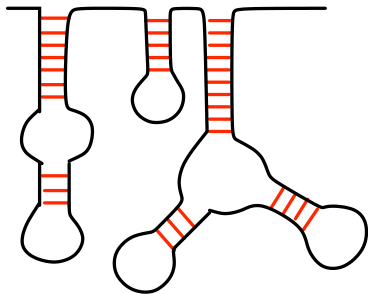
generating RNA-structure with SCFG



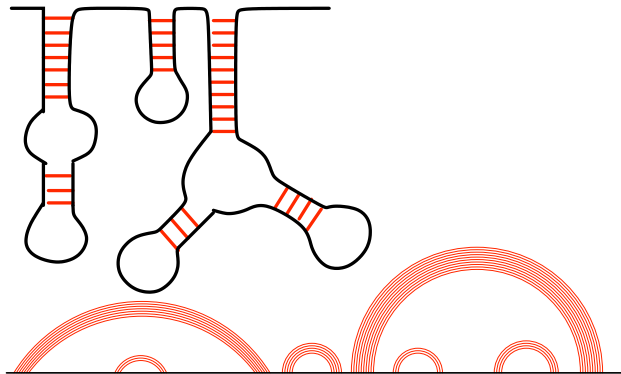
generating RNA-structure with SCFG



generating RNA-structure with SCFG

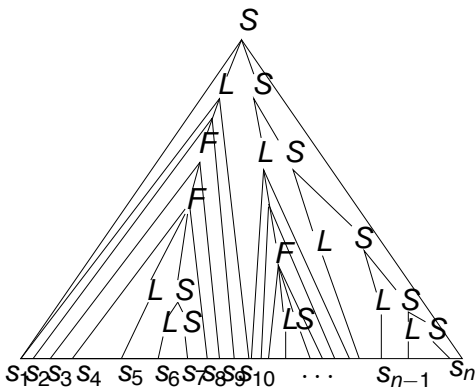


generating RNA-structure with SCFG

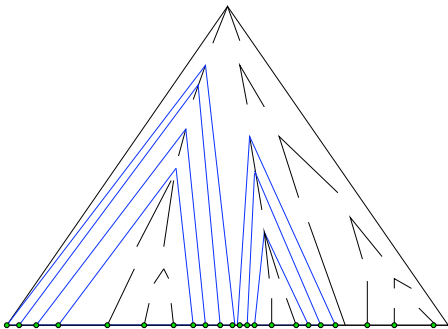


From parse tree to structure

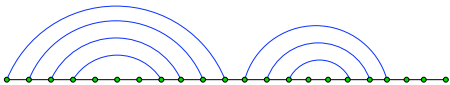
parse tree



From parse tree to structure



From parse tree to structure



efficiently doable in SCFG model

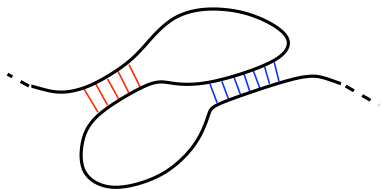
For given RNA sequence efficiently doable by dynamic programming:

- compute most probable secondary structure
- sample parse tree according to posterior distribution
- sample structure according to posterior distribution
- (compute optimal structure)

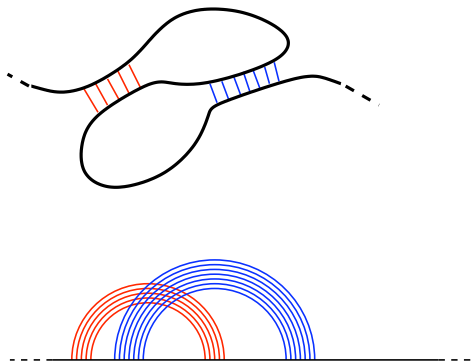
Outline

- 1 RNA-Folding with Stochastic Context-Free Grammars
- 2 RNA-Folding with Pseudoknots**
- 3 Bayes-Sampling of RNA-Structures with Pseudoknots
- 4 Experiments with Data
- 5 Conclusions

Pseudoknots

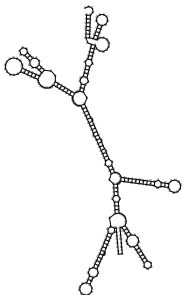


Pseudoknots



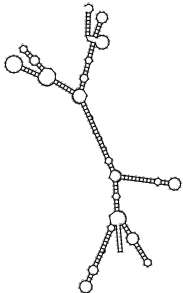
Structure of Escherichia Coli tmRNA

RNAfold estimation

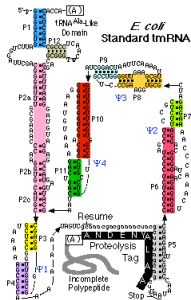


Structure of Escherichia Coli tmRNA

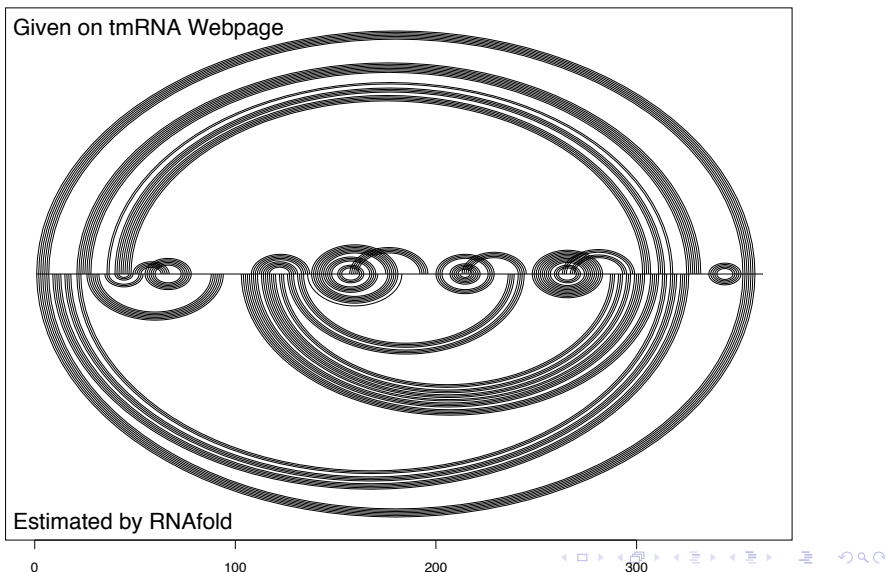
RNAfold estimation






Structure given on tmRNA website



Structure of Escherichia Coli tmRNA



Restricted Pseudoknot Models

-  [L. Cai, R. L. Malmberg, Y. Wu. \(2003\)](#)
Stochastic modeling of RNA pseudoknotted structures: a grammatical approach. *Bioinformatics* 19: i66-i73.
-  [E. Rivas, S. R. Eddy. \(1999\)](#)
A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.* 285:2053-2068.
-  [J. Reeder, R. Giegerich.\(2004\)](#)
Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. *BMC Bioinformatics*, 5:104.

... and many others

- **Restrictions on complexity** of Pseudoknots
- compute **optimal** structure for given RNA sequence

Aims of our model / method

- **a priori no restrictions on pseudoknots**
- Bayes-Sampling of RNA-structures
- efficient, if large number of pseudoknots unlikely
- secondary structure model: prior distribution, rather uninformative
- simple: transparent, parameters interpretable

Aims of our model / method

- a priori no restrictions on pseudoknots
- Bayes-Sampling of RNA-structures
- efficient, if large number of pseudoknots unlikely
- secondary structure model: prior distribution, rather uninformative
- simple: transparent, parameters interpretable

Aims of our model / method

- a priori no restrictions on pseudoknots
- Bayes-Sampling of RNA-structures
- efficient, if large number of pseudoknots unlikely
- secondary structure model: prior distribution, rather uninformative
- simple: transparent, parameters interpretable

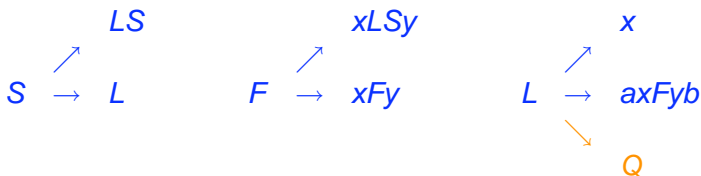
Aims of our model / method

- a priori no restrictions on pseudoknots
- Bayes-Sampling of RNA-structures
- efficient, if large number of pseudoknots unlikely
- secondary structure model: prior distribution, rather uninformative
- simple: transparent, parameters interpretable

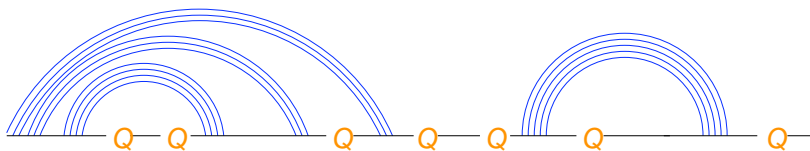
Aims of our model / method

- a priori no restrictions on pseudoknots
- Bayes-Sampling of RNA-structures
- efficient, if large number of pseudoknots unlikely
- secondary structure model: prior distribution, rather uninformative
- simple: transparent, parameters interpretable

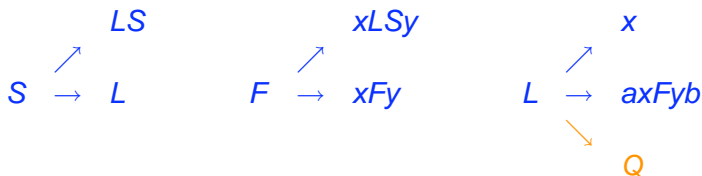
combine SCFG with pseudoknots



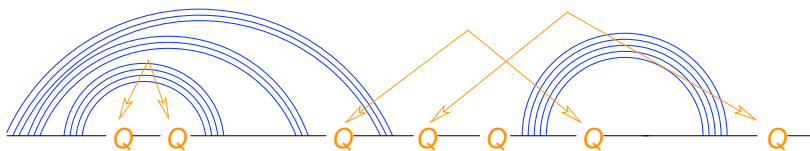
- 1 SCFG \rightsquigarrow RNA with Q-symbols
- 2 random mating of Q-symbols
- 3 Q-Q-pairs generate stems



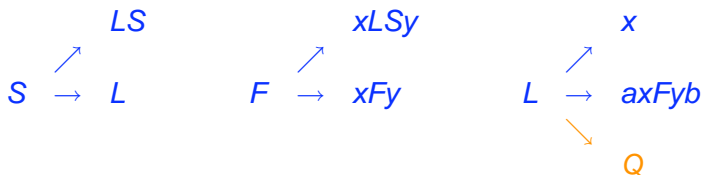
combine SCFG with pseudoknots



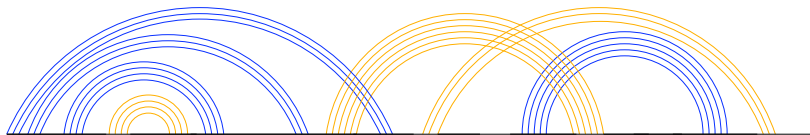
- 1 SCFG \rightsquigarrow RNA with Q-symbols
- 2 random mating of Q-symbols
- 3 Q-Q-pairs generate stems



combine SCFG with pseudoknots



- 1 SCFG \rightsquigarrow RNA with **Q**-symbols
- 2 random mating of **Q**-symbols
- 3 **Q-Q**-pairs generate stems



Notations

\mathcal{S} : given sequence

\mathcal{Q} : configuration of Q-stems

Ψ : SCFG parse tree

$\Omega = [\Psi, \mathcal{Q}]$: structure = $\{(i, j) \mid \text{positions } i \text{ and } j \text{ paired}\}$.

Notations

\mathcal{S} : given sequence

\mathcal{Q} : configuration of Q-stems

Ψ : SCFG parse tree

$\Omega = [\Psi, \mathcal{Q}]$: structure = $\{(i, j) \mid \text{positions } i \text{ and } j \text{ paired}\}$.

Aim: Generate RNA structure according to

$$\Pr(\Omega \mid \mathcal{S}) = \sum_{\Psi, \mathcal{Q} : [\Psi, \mathcal{Q}] = \Omega} \Pr(\Psi, \mathcal{Q} \mid \mathcal{S})$$

Outline

- 1 RNA-Folding with Stochastic Context-Free Grammars
- 2 RNA-Folding with Pseudoknots
- 3 Bayes-Sampling of RNA-Structures with Pseudoknots**
- 4 Experiments with Data
- 5 Conclusions

Bayes-Sampling for Ω

Sampling-Strategy for RNA Structure Ω according to posterior Prob. $\Pr(\Omega | \mathcal{S})$ for given RNA sequence \mathcal{S} :

- 1 Sample \hat{Q} according to $\Pr(Q | \mathcal{S})$ by Markov-chain Monte-Carlo method (MCMC).
- 2 Sample $\hat{\Psi}$ according to $\Pr(\Psi | \hat{Q}, \mathcal{S})$ by dynamic Programming
- 3 This makes $\hat{\Omega} = [\hat{\Psi}, \hat{Q}]$ distributed according to

$$\Pr(\Omega | \mathcal{S}) = \sum_{\Psi, Q : [\Psi, Q] = \Omega} \Pr(\Psi, Q | \mathcal{S})$$

Bayes-Sampling for Ω

Sampling-Strategy for RNA Structure Ω according to posterior Prob. $\Pr(\Omega | \mathcal{S})$ for given RNA sequence \mathcal{S} :

- 1 Sample \hat{Q} according to $\Pr(Q | \mathcal{S})$ by Markov-chain Monte-Carlo method (MCMC).
- 2 Sample $\hat{\Psi}$ according to $\Pr(\Psi | \hat{Q}, \mathcal{S})$ by dynamic Programming
- 3 This makes $\hat{\Omega} = [\hat{\Psi}, \hat{Q}]$ distributed according to

$$\Pr(\Omega | \mathcal{S}) = \sum_{\Psi, Q : [\Psi, Q] = \Omega} \Pr(\Psi, Q | \mathcal{S})$$

Bayes-Sampling for Ω

Sampling-Strategy for RNA Structure Ω according to posterior Prob. $\Pr(\Omega | \mathcal{S})$ for given RNA sequence \mathcal{S} :

- 1 Sample \hat{Q} according to $\Pr(Q | \mathcal{S})$ by Markov-chain Monte-Carlo method (MCMC).
- 2 Sample $\hat{\Psi}$ according to $\Pr(\Psi | \hat{Q}, \mathcal{S})$ by dynamic Programming
- 3 This makes $\hat{\Omega} = [\hat{\Psi}, \hat{Q}]$ distributed according to

$$\Pr(\Omega | \mathcal{S}) = \sum_{\Psi, Q : [\Psi, Q] = \Omega} \Pr(\Psi, Q | \mathcal{S})$$

Markov-chain Monte-Carlo (MCMC)

MCMC: Build Markov chain Q_0, Q_1, Q_2, \dots with stationary distribution $\Pr(Q | S)$ and let it converge.

Metropolis-Hastings:

Given current Q_i propose Q' with probability $p(Q_i \rightarrow Q')$.
Accept $Q_{i+1} := Q'$ with probability

$$\min \left\{ 1, \frac{p(Q' \rightarrow Q_i) \cdot \Pr(Q' | S)}{p(Q_i \rightarrow Q') \cdot \Pr(Q_i | S)} \right\}$$

otherwise $Q_{i+1} := Q_i$

Markov-chain Monte-Carlo (MCMC)

MCMC: Build Markov chain Q_0, Q_1, Q_2, \dots with stationary distribution $\Pr(Q | S)$ and let it converge.

Metropolis-Hastings:

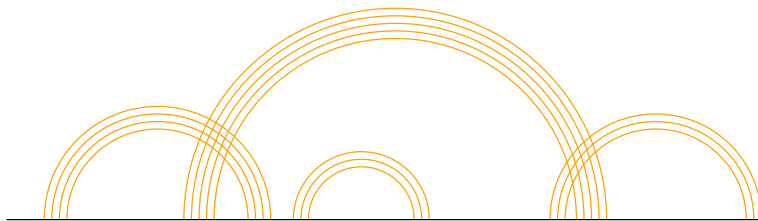
Given current Q_i propose Q' with probability $p(Q_i \rightarrow Q')$.

Accept $Q_{i+1} := Q'$ with probability

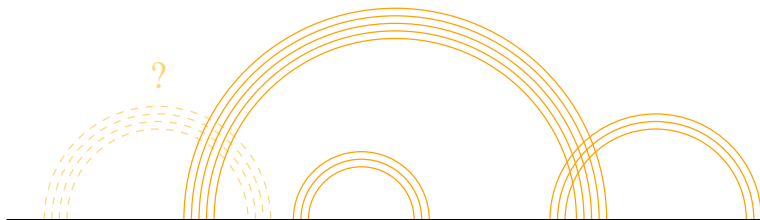
$$\min \left\{ 1, \frac{p(Q' \rightarrow Q_i) \cdot \Pr(Q' | S)}{p(Q_i \rightarrow Q') \cdot \Pr(Q_i | S)} \right\}$$

otherwise $Q_{i+1} := Q_i$

Suggestions for Q_{i+1}



Suggestions for Q_{i+1}

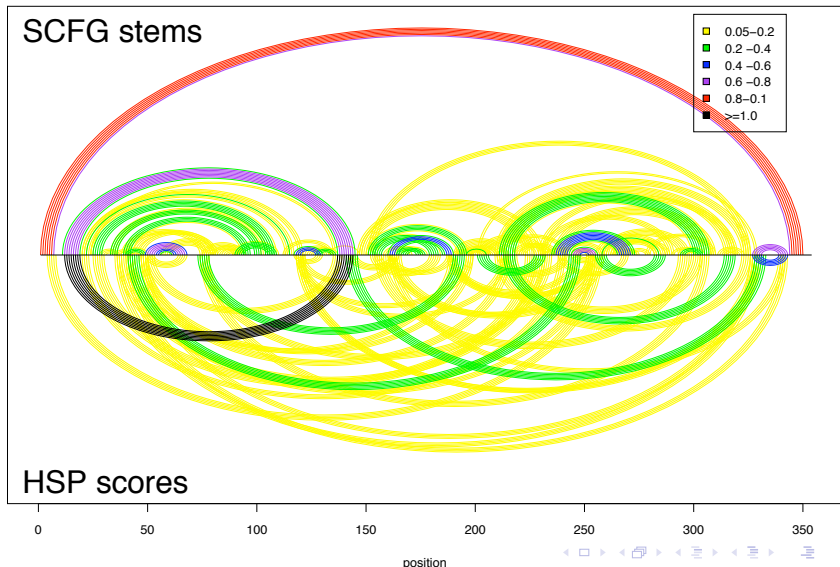


Suggestions for Q_{i+1}



Candidates for Pseudoknots

tmRNA of rice bacterium



Weight for Q -stem proposal

$$\text{proposal prob. for HSP} \propto \frac{1 - e^{(\text{alignment score}) \cdot c_1}}{\max\{(\text{SCFG-stem posterior prob.}), c_2\}}$$

$$c_1 = 10^{-6}, c_2 = 10^{-5}$$

Searching Optima

Search for

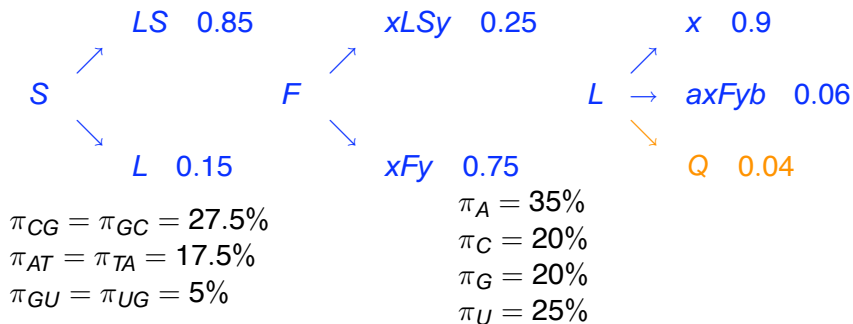
$$\arg \max_{[\Psi, \mathcal{Q}]} \Pr([\Psi, \mathcal{Q}] \mid \mathcal{S})$$

by Simulated Annealing.

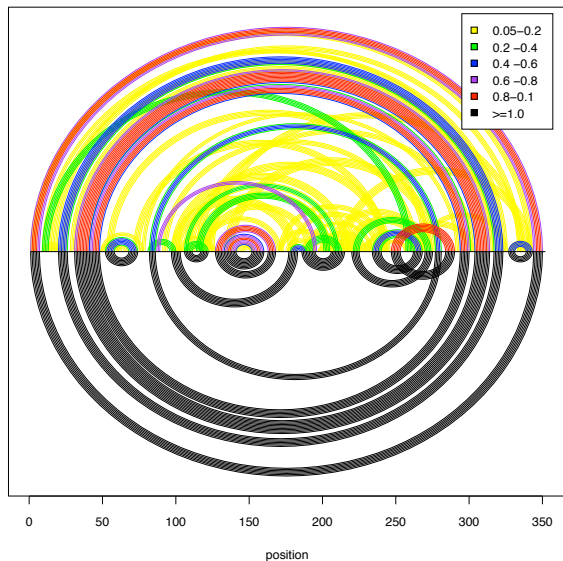
Outline

- 1 RNA-Folding with Stochastic Context-Free Grammars
- 2 RNA-Folding with Pseudoknots
- 3 Bayes-Sampling of RNA-Structures with Pseudoknots
- 4 Experiments with Data**
- 5 Conclusions

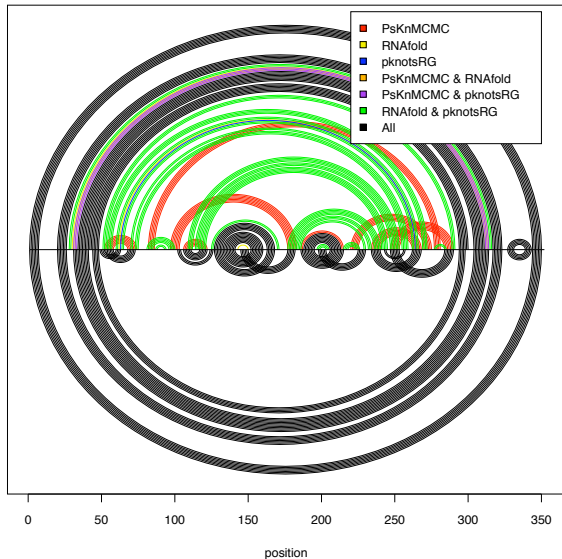
Used model parameter values



Treponema pallidum pre-tmRNA
posterior vs. most probable



Treponema pallidum pre-tmRNA
predictions vs. real



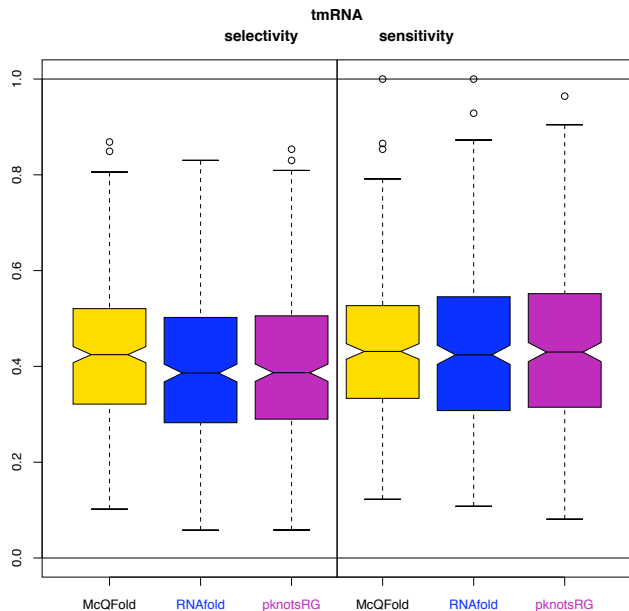
Comparison with RNAfold and pknotsRG

351 tmRNA sequences of length > 200 from
<http://www.indiana.edu/~tmrna/>

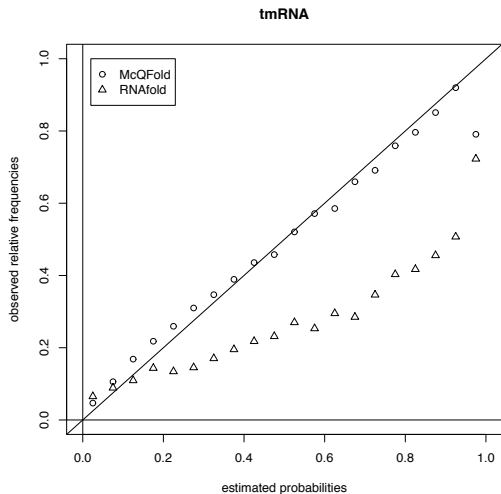
| | estimated paired | estimated not paired |
|---------------------|---------------------|-------------------------|
| actually paired | A | a |
| actually not paired | B | b |

Selectivity: $A/(A + B)$ (also called PPV or Relevance)

Sensitivity: $A/(A + a)$



estimated Pair-Probabilities



More data sets

- 200 simulated sequences
- tRNA
- RNase P
- Pseudoknot-Fragments

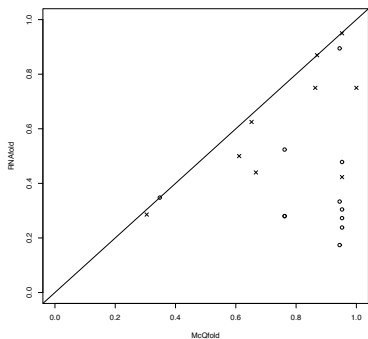
Results:

www.cs.uni-frankfurt.de/~metzler/McQFold/McQFold.pdf

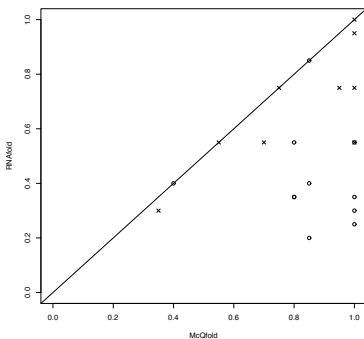
www.cs.uni-frankfurt.de/~metzler/McQFold/McQFoldSupplement.pdf

tRNA: McQFold versus RNAfold

PPV for tRNA



Sensitivity for tRNA



Outline

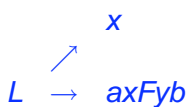
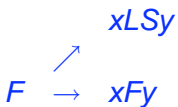
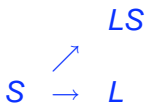
- 1 RNA-Folding with Stochastic Context-Free Grammars
- 2 RNA-Folding with Pseudoknots
- 3 Bayes-Sampling of RNA-Structures with Pseudoknots
- 4 Experiments with Data
- 5 Conclusions**

Conclusions and future directions

- RNA structure estimation uncertain
- Uncertainty should be assessed, e.g. by Bayesian sampling.
- Combine information from homologous sequences.
- Combine SCFGs with Q-Stems to allow pseudoknots in structure alignments and structural sequence profiles.

Berechnung der Wahrscheinlichkeit einer Sequenz

Gegeben: $S = (s_1, \dots, s_n) \in \{a, c, g, u\}^n$ Wahrscheinlichkeiten der Regeln



Gesucht: W'keit, dass S letztlich zu S wird.

Berechnung der W'keit einer Sequenz: Teilprobleme

$$\Phi_{ij}(X) := \Pr(X \text{ führt zu } s_i, \dots, s_j)$$

Berechnung der W 'keit einer Sequenz: Teilprobleme

$$\phi_{ij}(X) := \Pr(X \text{ führt zu } s_i, \dots, s_j)$$

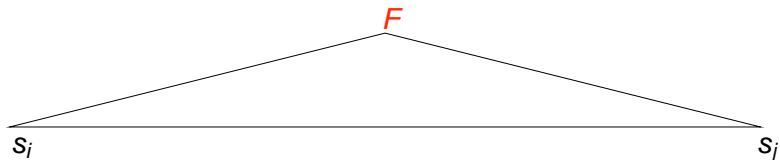
Ziel: berechne $\phi_{1n}(S)$!

Berechnung der W 'keit einer Sequenz: Teilprobleme

$$\Phi_{ij}(X) := \Pr(X \text{ führt zu } s_i, \dots, s_j)$$

Ziel: berechne $\Phi_{1n}(S)$!

$$\begin{aligned} \Phi_{ij}(F) &= \Phi_{i+1,j-1}(F) \cdot \Pr(F \rightarrow xFy) \cdot \pi_{s_i s_j} \\ &+ \sum_k \pi_{s_i s_j} \cdot \Pr(F \rightarrow xLSy) \cdot \Phi_{i+1,k}(L) \cdot \Phi_{k+1,j-1}(S) \end{aligned}$$

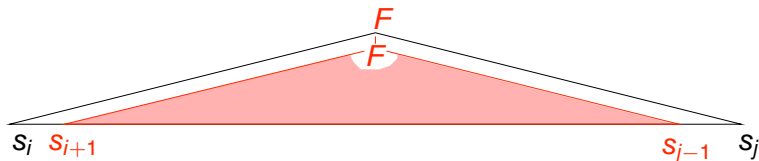


Berechnung der W'keit einer Sequenz: Teilprobleme

$$\Phi_{ij}(X) := \Pr(X \text{ führt zu } s_i, \dots, s_j)$$

Ziel: berechne $\Phi_{1n}(S)$!

$$\begin{aligned} \Phi_{ij}(F) &= \Phi_{i+1,j-1}(F) \cdot \Pr(F \rightarrow xFy) \cdot \pi_{s_i s_j} \\ &+ \sum_k \pi_{s_i s_j} \cdot \Pr(F \rightarrow xLSy) \cdot \Phi_{i+1,k}(L) \cdot \Phi_{k+1,j-1}(S) \end{aligned}$$

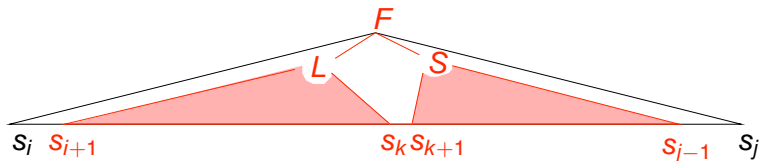


Berechnung der W'keit einer Sequenz: Teilprobleme

$$\Phi_{ij}(X) := \Pr(X \text{ führt zu } s_i, \dots, s_j)$$

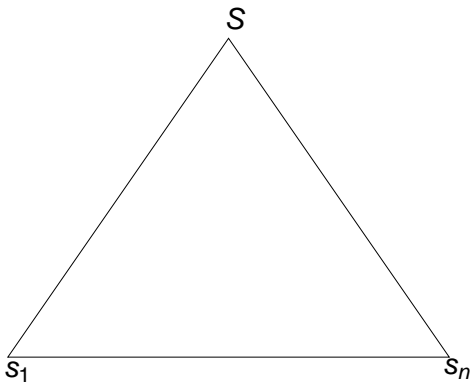
Ziel: berechne $\Phi_{1n}(S)$!

$$\begin{aligned} \Phi_{ij}(F) &= \Phi_{i+1,j-1}(F) \cdot \Pr(F \rightarrow xFy) \cdot \pi_{s_i s_j} \\ &+ \sum_k \pi_{s_i s_j} \cdot \Pr(F \rightarrow xLSy) \cdot \Phi_{i+1,k}(L) \cdot \Phi_{k+1,j-1}(S) \end{aligned}$$



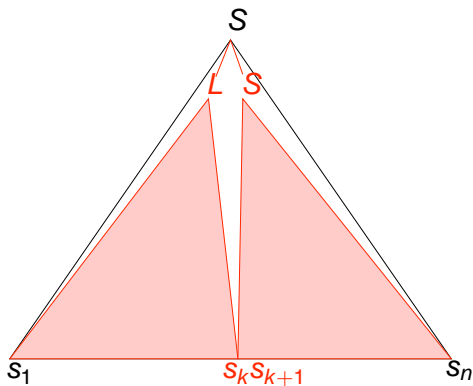
Erzeuge Struktur gemäß a-posteriori-Verteilung in SCFG-Modell

Wähle zufällig gemäß Beitrag zu $\Phi_{ij}(X)$



Erzeuge Struktur gemäß a-posteriori-Verteilung in SCFG-Modell

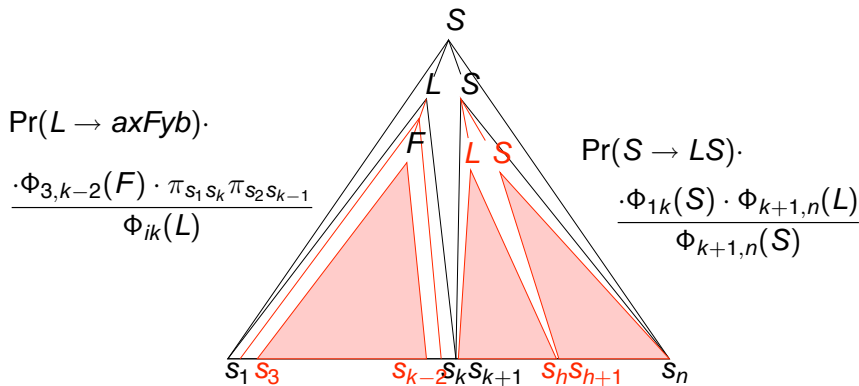
Wähle zufällig gemäß Beitrag zu $\Phi_{ij}(X)$



$$\frac{\Pr(S \rightarrow LS) \cdot \Phi_{1k}(L) \cdot \Phi_{k+1,n}(S)}{\Phi_{1n}(S)}$$

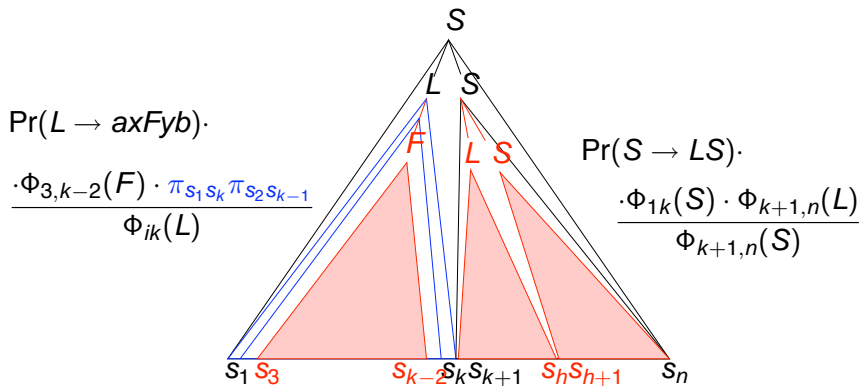
Erzeuge Struktur gemäß a-posteriori-Verteilung in SCFG-Modell

Wähle zufällig gemäß Beitrag zu $\Phi_{ij}(X)$



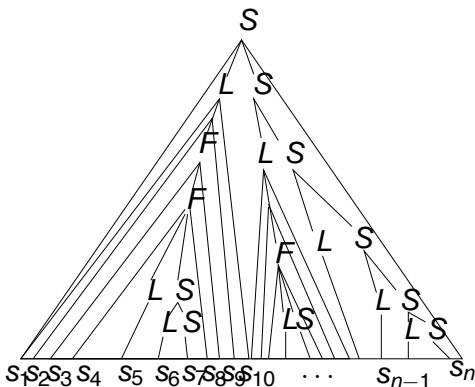
Erzeuge Struktur gemäß a-posteriori-Verteilung in SCFG-Modell

Wähle zufällig gemäß Beitrag zu $\Phi_{ij}(X)$

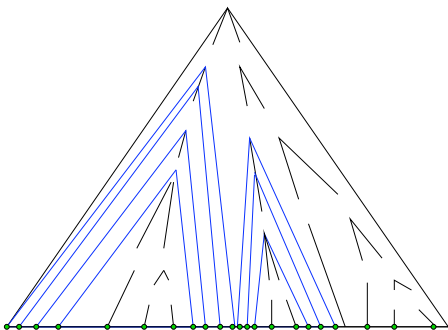


Erzeuge Struktur gemäß a-posteriori-Verteilung in SCFG-Modell

Ableitungsbaum



Erzeuge Struktur gemäß a-posteriori-Verteilung in SCFG-Modell



Erzeuge Struktur gemäß a-posteriori-Verteilung in SCFG-Modell

