MCMC sampling of RNA structures with pseudoknots

Dirk Metzler

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

joint work with Markus Nebel, Universität Kaiserslautern

Outline

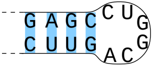
Outline

tmRNA of Escherichia Coli

GGGGCUGAUUCUGGAUUCGACGGGAUUUGCGAAACCCAAGGUGCAUGCCGAGGGGCGGUUGGCCUCGUAAAAAGC
CGCAAAAAAUAGUCGCAAACGACGAAAACUACGCUUUAGCAGCUUAAUAACCUGCUUAGAGCCCUCUCUCCCUAG
CCUCCGCUCUUAGGACGGGGAUCAAGAGAGGUCAAACCCAAAAGAGAUCGCGUGGAAGCCCUGCCUGGGGUUGAA
GCGUUAAAACUUAAUCAGGCUAGUUUGUUAGUGGCGUGUCCGUCCGCAGCUGGCAAGCGAAUGUAAAGACUGACU
AAGCAUGUAGUACCGAGGAUGUAGGAAUUUCGGACGCGGGUUCAACUCCCGCCAGCUCCACCA

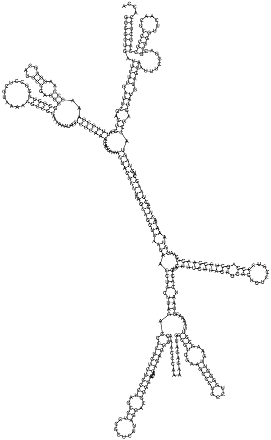
tmRNA of Escherichia Coli

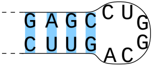
GGGGCUGAUUCUGGAUUCGACGGGAUUUGCGAAACCCAAGGUGCAUGCCGAGGGGCGGUUGGCCUCGUAAAAAGC
CGCAAAAAAUAGUCGCAAACGACGAAAACUACGCUUUAGCAGCUUAAUAACCUGCUUAGAGCCCUCUCUCCCUAG
CCUCCGCUCUUAGGACGGGGAUCAAGAGAGGUCAAACCCAAAAGAGAUCGCGUGGAAGCCCUGCCUGGGGUUGAA
GCGUUAAAACUUAAUCAGGCUAGUUUGUUAGUGGCGUGUCCGUCCGCAGCUGGCAAGCGAAUGUAAAGACUGACU
AAGCAUGUAGUACCGAGGAUGUAGGAAUUUCGGACGCGGGUUCAACUCCCGCCAGCUCCACCA



tmRNA of Escherichia Coli

GGGGCUGAUUCUGGAUUCGACGGGAUUUGCGAAACCCAAGGUGCAUGCCGAGGGGCGGUUGGCCUCGUAAAAAGC
CGCAAAAAAUAGUCGCAAACGACGAAAACUACGCUUUAGCAGCUUAAUAACCUGCUUAGAGCCCUCUCUCCCUAG
CCUCCGCUCUUAGGACGGGGAUCAAGAGAGGUCAAACCCAAAAGAGAUCGCGUGGAAGCCCUGCCUGGGGUUGAA
GCGUUAAAACUUAAUCAGGCUAGUUUGUUAGUGGCGUGUCCGUCCGCAGCUGGCAAGCGAAUGUAAAGACUGACU
AAGCAUGUAGUACCGAGGAUGUAGGAAUUUCGGACGCGGGUUCAACUCCCGCCAGCUCCACCA





RNAfold
Implementation of
Zuker algorithm in
Vienna RNA Package

Zuker (1989)
Zuker et al. (1999)

Stochastic Context-Free Grammar (SCFG)

- Terminal Symbols

$$\texttt{A},\texttt{C},\texttt{G},\texttt{U}$$

- Non-Terminal Symbols

$$S, L, F$$

- Rules with Probabilities

$$\begin{tikzcd} & LS & & xLSy & & x \\ S \arrow[r] \arrow[ur] & L & F \arrow[r] \arrow[ur] & xFy & L \arrow[r] \arrow[ur] & axFyb \end{tikzcd}$$

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

- forbidden: $xFy \rightarrow xLy$

Stochastic Context-Free Grammar (SCFG)

- Terminal Symbols

$$\texttt{A},\texttt{C},\texttt{G},\texttt{U}$$

- Non-Terminal Symbols

$$S, L, F$$

- Rules with Probabilities

$$\begin{array}{ccccccccc} & & LS & & & xLSy & & & x \\ & \nearrow & & & \nearrow & & & \nearrow & \\ S & \rightarrow & L & F & \rightarrow & xFy & L & \rightarrow & axFyb \end{array}$$

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

- forbidden: $xFy \rightarrow xLy$

Generating RNA structure from SCFG

S

Generating RNA structure from SCFG

S

S−>LS

Generating RNA structure from SCFG

L L L L L L L L L L L L L L L L L L L L L S

S−>LS

Generating RNA structure from SCFG

L L L L L L L L L L L L L L L L L L L L L S

S−>x
L−>x

Generating RNA structure from SCFG

acgLuaagauLuauLggcauua

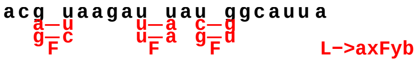
S−>x
L−>x

Generating RNA structure from SCFG

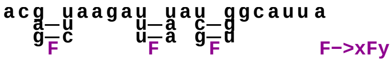
a c g L u a a g a u L u a u L g g c a u u a

L−>axFyb

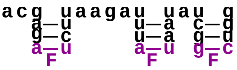
Generating RNA structure from SCFG



Generating RNA structure from SCFG

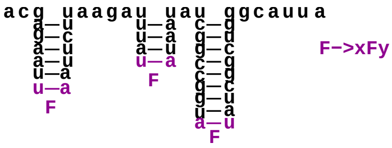


Generating RNA structure from SCFG

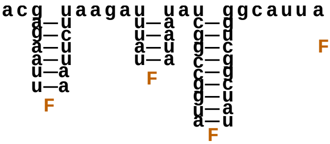


F->xFy

Generating RNA structure from SCFG

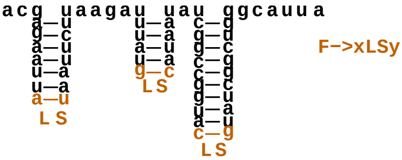


Generating RNA structure from SCFG

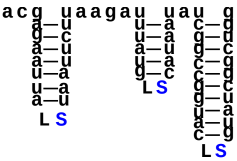


F−>xLSy

Generating RNA structure from SCFG

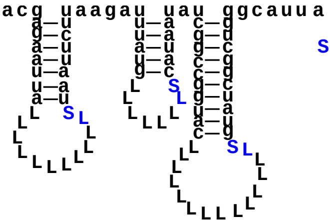


Generating RNA structure from SCFG



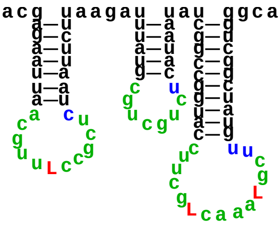
S−>LS

Generating RNA structure from SCFG



S−>LS

Generating RNA structure from SCFG

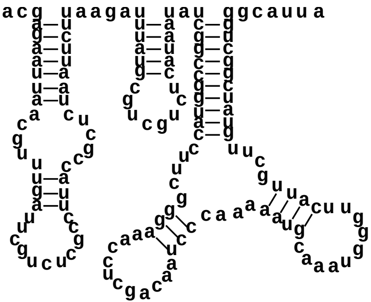


S->x

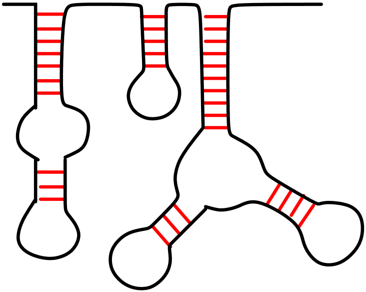
L->x

L->axFyb

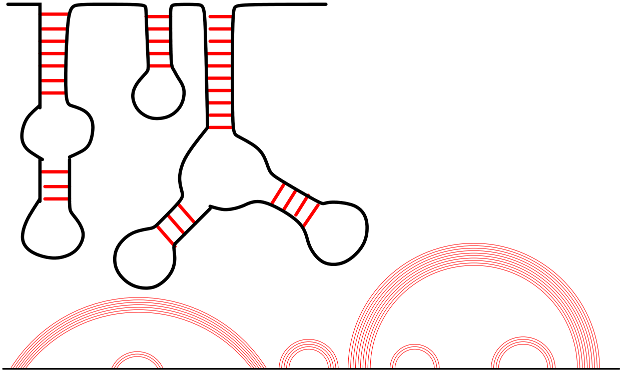
Generating RNA structure from SCFG



Generating RNA structure from SCFG



Generating RNA structure from SCFG



Outline

1. RNA folding with Stochastic Context-Free Grammars
   - Basic Model
   - Dynamic Programming
2. RNA folding with Pseudoknots
   - Pseudoknots
   - Combine SCFG with Pseudoknots
   - Bayesian Sampling of RNA structures with pseudoknots
3. Performance on Data
   - tmRNA
   - Simulated Data

Computing the Probability of a sequence

For given $\mathcal{S} = (s_1, \ldots, s_n) \in \{\mathrm{a}, \mathrm{c}, \mathrm{g}, \mathrm{u}\}^n$ and probabilities of grammar rules

$$\begin{array}{ccccccccccc} & & LS & \quad & & & xLSy & \quad & & & x \\ & \nearrow & & & & \nearrow & & & & \nearrow & \\ S & \rightarrow & L & & F & \rightarrow & xFy & & L & \rightarrow & axFyb \end{array}$$

compute the probability that S is transformed into $\mathcal{S}$.

Computing the Probability of a sequence: partial problems

$$\Phi_{ij}(X) := \Pr(X \text{ is transformed into } s_i, \ldots, s_j)$$

Computing the Probability of a sequence: partial problems

$$\Phi_{ij}(X) := \Pr(X \text{ is transformed into } s_i, \ldots, s_j)$$

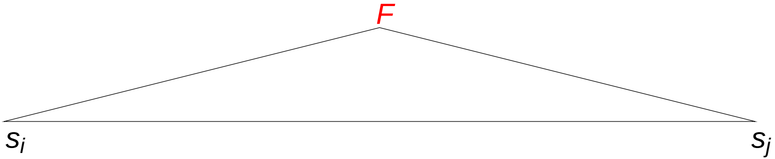
Aim: compute $\Phi_{1n}(S)$!

Computing the Probability of a sequence: partial problems

$$\Phi_{ij}(X) := \Pr(X \text{ is transformed into } s_i, \ldots, s_j)$$

Aim: compute $\Phi_{1n}(S)$!

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

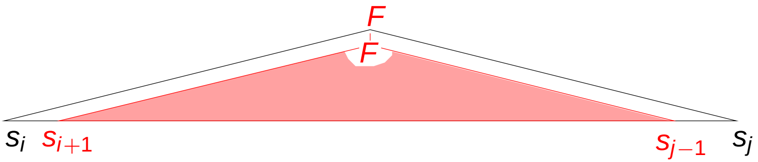


Computing the Probability of a sequence: partial problems

$$\Phi_{ij}(X) := \Pr(X \text{ is transformed into } s_i, \ldots, s_j)$$

Aim: compute $\Phi_{1n}(S)$!

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

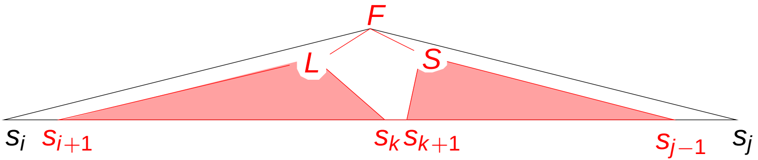


Computing the Probability of a sequence: partial problems

$$\Phi_{ij}(X) := \Pr(X \text{ is transformed into } s_i, \ldots, s_j)$$

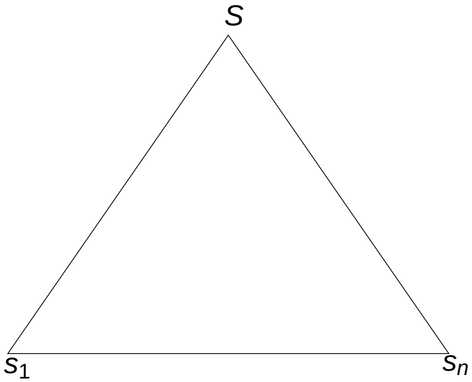
Aim: compute $\Phi_{1n}(S)$!

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



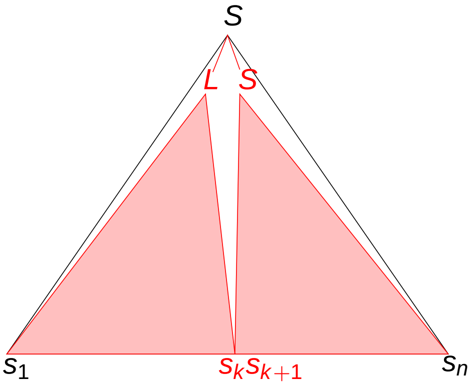
Sampling a Structure from Posterior in SCFG model

Sample according to contribution to $\Phi_{ij}(X)$



Sampling a Structure from Posterior in SCFG model

Sample according to contribution to $\Phi_{ij}(X)$

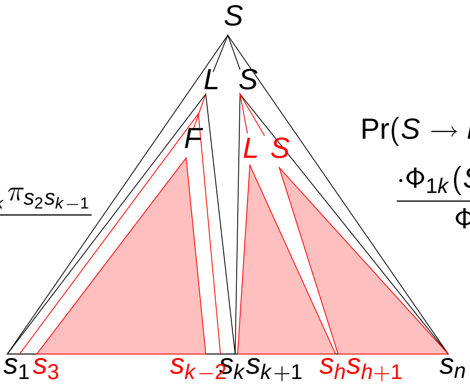


$$\frac{\Pr(S \rightarrow LS) \cdot \Phi_{1k}(L) \cdot \Phi_{k+1,n}(S)}{\Phi_{1n}(S)}$$

Sampling a Structure from Posterior in SCFG model

Sample according to contribution to $\Phi_{ij}(X)$

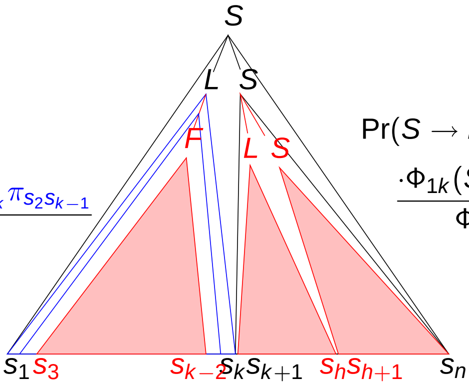
$$\Pr(L \to axFyb)\cdot \frac{\cdot\Phi_{3,k-2}(F)\cdot \pi_{s_1 s_k}\pi_{s_2 s_{k-1}}}{\Phi_{ik}(L)}$$



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

Sampling a Structure from Posterior in SCFG model

Sample according to contribution to $\Phi_{ij}(X)$

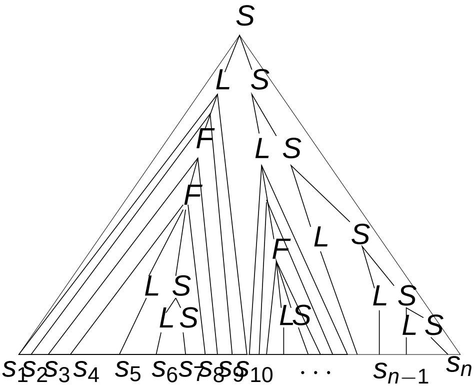


$$\Pr(L \rightarrow axFyb)\cdot \frac{\cdot\Phi_{3,k-2}(F)\cdot\pi_{s_1 s_k}\pi_{s_2 s_{k-1}}}{\Phi_{ik}(L)}$$

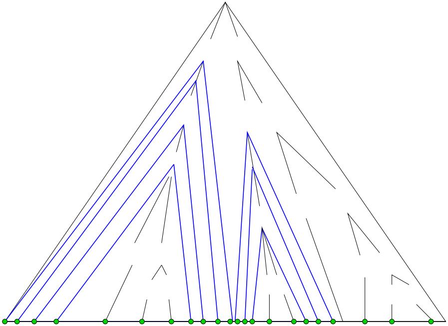
$$\Pr(S \rightarrow LS)\cdot \frac{\cdot\Phi_{1k}(S)\cdot\Phi_{k+1,n}(L)}{\Phi_{k+1,n}(S)}$$

Sampling a Structure from Posterior in SCFG model

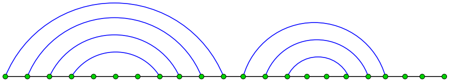
Parse Tree



Sampling a Structure from Posterior in SCFG model



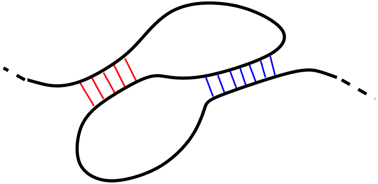
Sampling a Structure from Posterior in SCFG model



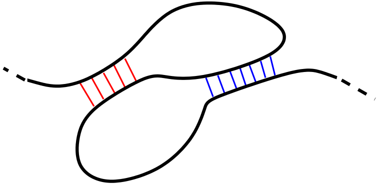
Outline

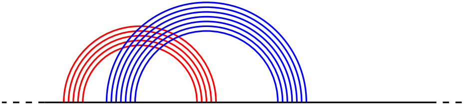
1. RNA folding with Stochastic Context-Free Grammars
   - Basic Model
   - Dynamic Programming
2. RNA folding with Pseudoknots
   - Pseudoknots
   - Combine SCFG with Pseudoknots
   - Bayesian Sampling of RNA structures with pseudoknots
3. Performance on Data
   - tmRNA
   - Simulated Data

Pseudoknots

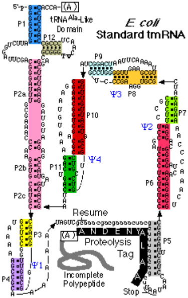


Pseudoknots



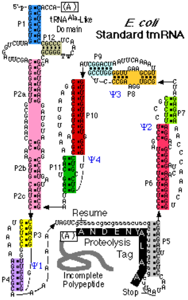


Structure of Escherichia Coli tmRNA

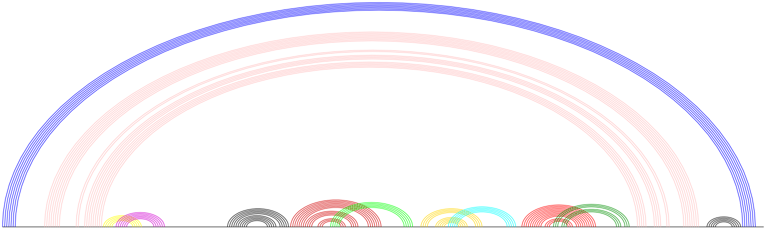


picture stolen from tmRNA website `http://www.indiana.edu/∼tmrna/`

Structure of Escherichia Coli tmRNA

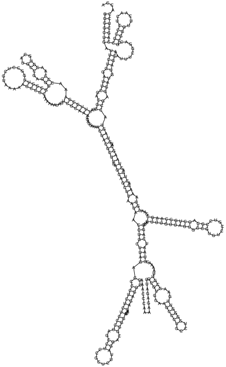


picture stolen from tmRNA website
`http://www.indiana.edu/∼tmrna/`



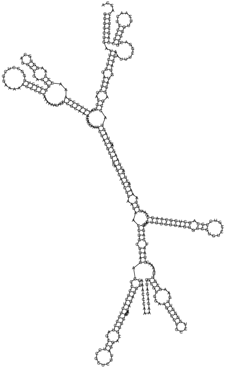
Structure of Escherichia Coli tmRNA

RNAfold estimation

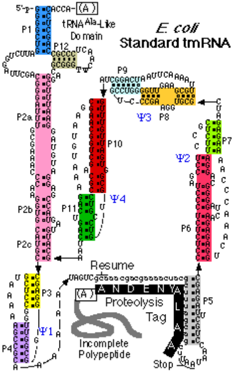


Structure of Escherichia Coli tmRNA

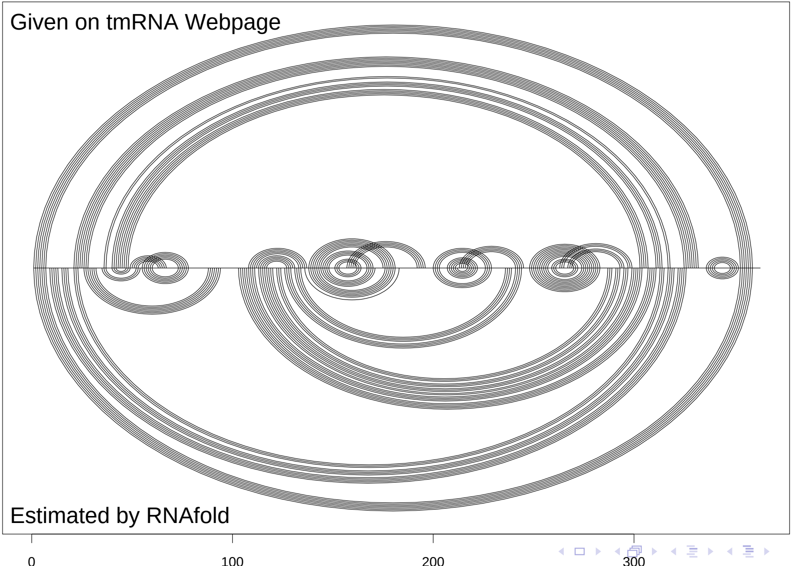
RNAfold estimation



Given on tmRNA Website



Structure of Escherichia Coli tmRNA



Outline

1. RNA folding with Stochastic Context-Free Grammars
   - Basic Model
   - Dynamic Programming
2. RNA folding with Pseudoknots
   - Pseudoknots
   - Combine SCFG with Pseudoknots
   - Bayesian Sampling of RNA structures with pseudoknots
3. Performance on Data
   - tmRNA
   - Simulated Data

Models with restricted types of pseudoknots

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

... and several others

- restrict the way pseuknots may intersect
- compute THE BEST Structure for given RNA sequence

Aims for our Pseudoknot Grammar / Method

- a priori no restrictions on pseudoknot interactions
- sampling structures from posterior distributions
- efficient if high number of pseudoknots is unlikely
- rather prior distribution than biologically meaningful model
- simplicity!

Aims for our Pseudoknot Grammar / Method

- a priori no restrictions on pseudoknot interactions
- sampling structures from posterior distributions
- efficient if high number of pseudoknots is unlikely
- rather prior distribution than biologically meaningful model
- simplicity!

Aims for our Pseudoknot Grammar / Method

- a priori no restrictions on pseudoknot interactions
- sampling structures from posterior distributions
- efficient if high number of pseudoknots is unlikely
- rather prior distribution than biologically meaningful model
- simplicity!

Aims for our Pseudoknot Grammar / Method

- a priori no restrictions on pseudoknot interactions
- sampling structures from posterior distributions
- efficient if high number of pseudoknots is unlikely
- rather prior distribution than biologically meaningful model
- simplicity!

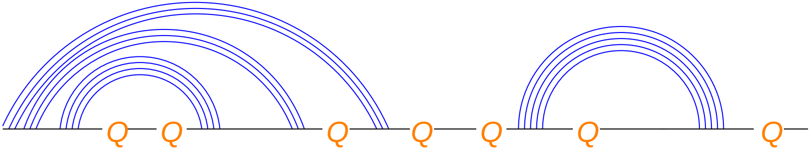
Aims for our Pseudoknot Grammar / Method

- a priori no restrictions on pseudoknot interactions
- sampling structures from posterior distributions
- efficient if high number of pseudoknots is unlikely
- rather prior distribution than biologically meaningful model
- simplicity!

Combine SCFG with pseudoknots

$$\begin{tikzcd} & LS \\ S \arrow[ur] \arrow[r] & L \end{tikzcd} \qquad \begin{tikzcd} & xLSy \\ F \arrow[ur] \arrow[r] & xFy \end{tikzcd} \qquad \begin{tikzcd} & x \\ L \arrow[ur] \arrow[r] \arrow[dr] & axFyb \\ & Q \end{tikzcd}$$

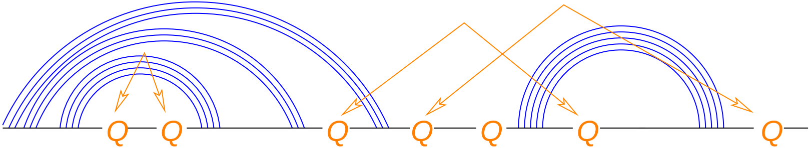
1. SCFG ⇝ RNA with *Q*-Symbols
2. random mating of *Q*-symbols
3. *Q*-*Q*-pairs produce stems



Combine SCFG with pseudoknots

$$\begin{tikzcd} & LS \\ S \arrow[ur] \arrow[r] & L \end{tikzcd} \qquad \begin{tikzcd} & xLSy \\ F \arrow[ur] \arrow[r] & xFy \end{tikzcd} \qquad \begin{tikzcd} & x \\ L \arrow[ur] \arrow[r] \arrow[dr] & axFyb \\ & Q \end{tikzcd}$$

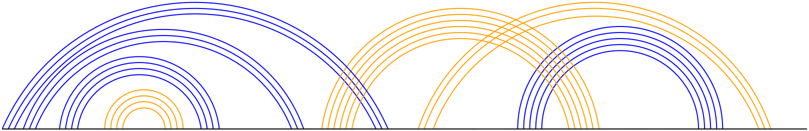
1. SCFG ⇝ RNA with Q-Symbols
2. random mating of Q-symbols
3. Q-Q-pairs produce stems



Combine SCFG with pseudoknots

$$\begin{tikzcd}
 & LS & & & xLSy & & & x \\
S \arrow[ur] \arrow[r] & L & & F \arrow[ur] \arrow[r] & xFy & & L \arrow[ur] \arrow[r] \arrow[dr] & axFyb \\
 & & & & & & & Q
\end{tikzcd}$$

1. SCFG ⇝ RNA with *Q*-Symbols
2. random mating of *Q*-symbols
3. *Q*-*Q*-pairs produce stems



Notations

$\mathcal{S}$: given Sequence

$\mathcal{Q}$: Configuration of Q-stems

Ψ: SCFG Parse Tree

$\Omega = [\Psi, \mathcal{Q}]$: Stucture = $\{(i,j) \mid$ Positions i and j are paired $\}$.

θ: Model parameters

Notations

$\mathcal{S}$: given Sequence

$\mathcal{Q}$: Configuration of *Q*-stems

Ψ: SCFG Parse Tree

$\Omega = [\Psi, \mathcal{Q}]$: Stucture = $\{(i, j) \mid$ Positions *i* and *j* are paired $\}$.

θ: Model parameters

Aims:

- Compute $L_{\mathcal{S}}(\theta) = \text{Pr}_{\theta}(\mathcal{S}) = \sum_{\Psi, \mathcal{Q}} \text{Pr}_{\theta}(\mathcal{S} \mid \mathcal{Q}, \Psi) \cdot \text{Pr}_{\theta}(\Psi) \cdot \text{Pr}_{\theta}(\mathcal{Q} \mid \Psi)$
- Sample RNA Structure according to $\text{Pr}(\Omega \mid \mathcal{S}) = \sum_{\Psi, \mathcal{Q} \,:\, [\Psi, \mathcal{Q}] = \Omega} \text{Pr}(\Psi, \mathcal{Q} \mid \mathcal{S})$

Outline

1. RNA folding with Stochastic Context-Free Grammars
   - Basic Model
   - Dynamic Programming
2. RNA folding with Pseudoknots
   - Pseudoknots
   - Combine SCFG with Pseudoknots
   - Bayesian Sampling of RNA structures with pseudoknots
3. Performance on Data
   - tmRNA
   - Simulated Data

For fixed $\mathcal{Q}$ doable by dynamic programming

- Compute $\Pr(\mathcal{Q} \mid \mathcal{S}) = \sum_{\Psi} \Pr(\Psi, \mathcal{Q} \mid \mathcal{S})$
- Sample SCFG Parse Tree Ψ according to $\Pr(\Psi \mid \mathcal{Q}, \mathcal{S})$
- Sample Structure Ω according to $\Pr(\Omega \mid \mathcal{Q}, \mathcal{S})$
- Compute
$$\arg\max_{\Psi} \Pr(\Psi \mid \mathcal{Q}, \mathcal{S})$$

For fixed $\mathcal{Q}$ doable by dynamic programming

- Compute $\Pr(\mathcal{Q} \mid \mathcal{S}) = \sum_{\Psi} \Pr(\Psi, \mathcal{Q} \mid \mathcal{S})$
- Sample SCFG Parse Tree Ψ according to $\Pr(\Psi \mid \mathcal{Q}, \mathcal{S})$
- Sample Structure Ω according to $\Pr(\Omega \mid \mathcal{Q}, \mathcal{S})$
- Compute

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

For fixed $\mathcal{Q}$ doable by dynamic programming

- Compute $\Pr(\mathcal{Q} \mid \mathcal{S}) = \sum_{\Psi} \Pr(\Psi, \mathcal{Q} \mid \mathcal{S})$
- Sample SCFG Parse Tree Ψ according to $\Pr(\Psi \mid \mathcal{Q}, \mathcal{S})$
- Sample Structure Ω according to $\Pr(\Omega \mid \mathcal{Q}, \mathcal{S})$
- Compute
$$\arg\max_{\Psi} \Pr(\Psi \mid \mathcal{Q}, \mathcal{S})$$

For fixed $\mathcal{Q}$ doable by dynamic programming

- Compute $\Pr(\mathcal{Q} \mid \mathcal{S}) = \sum_{\Psi} \Pr(\Psi, \mathcal{Q} \mid \mathcal{S})$
- Sample SCFG Parse Tree Ψ according to $\Pr(\Psi \mid \mathcal{Q}, \mathcal{S})$
- Sample Structure Ω according to $\Pr(\Omega \mid \mathcal{Q}, \mathcal{S})$
- Compute
  $$\arg\max_{\Psi} \Pr(\Psi \mid \mathcal{Q}, \mathcal{S})$$

Bayesian sampling of Ω

Strategy for sampling RNA structure Ω according to its posterior probability $\Pr(\Omega \mid \mathcal{S})$ for given RNA sequence $\mathcal{S}$:

1. Sample $\mathcal{Q}_i$ according to $\Pr(\mathcal{Q} \mid \mathcal{S})$ by Markov-Chain Monte Carlo (MCMC) Method.
2. Sample Ψ_i according to $\Pr(\Psi \mid \mathcal{Q}_i, \mathcal{S})$ by dynamic programming.
3. Then $\Omega_i = [\Psi_i, \mathcal{Q}_i]$ is sample according to

$$\begin{aligned}\Pr(\Omega \mid \mathcal{S}) &= \sum_{\Psi,\mathcal{Q}\,:\,[\Psi,\mathcal{Q}]=\Omega} \Pr(\Psi, \mathcal{Q} \mid \mathcal{S}) \\ &= \sum_{\Psi,\mathcal{Q}\,:\,[\Psi,\mathcal{Q}]=\Omega} \Pr(\Psi \mid \mathcal{Q}, \mathcal{S}) \cdot \Pr(\mathcal{Q} \mid \mathcal{S})\end{aligned}$$

Bayesian sampling of Ω

Strategy for sampling RNA structure Ω according to its posterior probability $\Pr(\Omega \mid \mathcal{S})$ for given RNA sequence $\mathcal{S}$:

1. Sample $\mathcal{Q}_i$ according to $\Pr(\mathcal{Q} \mid \mathcal{S})$ by Markov-Chain Monte Carlo (MCMC) Method.
2. Sample Ψ_i according to $\Pr(\Psi \mid \mathcal{Q}_i, \mathcal{S})$ by dynamic programming.
3. Then $\Omega_i = [\Psi_i, \mathcal{Q}_i]$ is sample according to

$$\begin{aligned}\Pr(\Omega \mid \mathcal{S}) &= \sum_{\Psi,\mathcal{Q}\,:\,[\Psi,\mathcal{Q}]=\Omega} \Pr(\Psi, \mathcal{Q} \mid \mathcal{S}) \\ &= \sum_{\Psi,\mathcal{Q}\,:\,[\Psi,\mathcal{Q}]=\Omega} \Pr(\Psi \mid \mathcal{Q}, \mathcal{S}) \cdot \Pr(\mathcal{Q} \mid \mathcal{S})\end{aligned}$$

Bayesian sampling of Ω

Strategy for sampling RNA structure Ω according to its posterior probability $\Pr(\Omega \mid \mathcal{S})$ for given RNA sequence $\mathcal{S}$:

1. Sample $\mathcal{Q}_i$ according to $\Pr(\mathcal{Q} \mid \mathcal{S})$ by Markov-Chain Monte Carlo (MCMC) Method.
2. Sample Ψ_i according to $\Pr(\Psi \mid \mathcal{Q}_i, \mathcal{S})$ by dynamic programming.
3. Then $\Omega_i = [\Psi_i, \mathcal{Q}_i]$ is sample according to

$$\begin{aligned}\Pr(\Omega \mid \mathcal{S}) &= \sum_{\Psi,\mathcal{Q}\,:\,[\Psi,\mathcal{Q}]=\Omega} \Pr(\Psi, \mathcal{Q} \mid \mathcal{S}) \\ &= \sum_{\Psi,\mathcal{Q}\,:\,[\Psi,\mathcal{Q}]=\Omega} \Pr(\Psi \mid \mathcal{Q}, \mathcal{S}) \cdot \Pr(\mathcal{Q} \mid \mathcal{S})\end{aligned}$$

Markov-Chain Monte Carlo (MCMC)

MCMC: construct Markov chain $\mathcal{Q}_0, \mathcal{Q}_1, \mathcal{Q}_2, \ldots$ with stationary distribution $\Pr(\mathcal{Q} \mid \mathcal{S})$ and let it converge.

Metropolis-Hastings:

Given current state $\mathcal{Q}_i$ propose $\mathcal{Q}'$ with Prob. $p(\mathcal{Q}_i \to \mathcal{Q}')$

Accept $\mathcal{Q}_{i+1} := \mathcal{Q}'$ with probability

$$\min\left\{1, \frac{p(\mathcal{Q}' \to \mathcal{Q}_i) \cdot \Pr(\mathcal{Q}' \mid \mathcal{S})}{p(\mathcal{Q}_i \to \mathcal{Q}') \cdot \Pr(\mathcal{Q}_i \mid \mathcal{S})}\right\}$$

otherwise $\mathcal{Q}_{i+1} := \mathcal{Q}_i$

Markov-Chain Monte Carlo (MCMC)

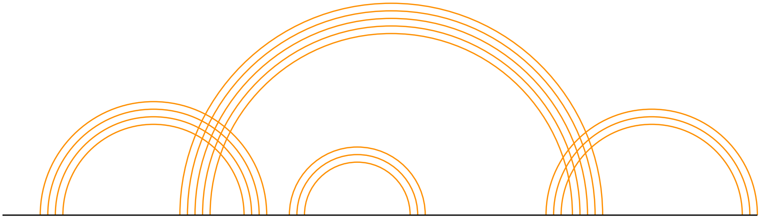
MCMC: construct Markov chain $\mathcal{Q}_0, \mathcal{Q}_1, \mathcal{Q}_2, \ldots$ with stationary distribution $\Pr(\mathcal{Q} \mid \mathcal{S})$ and let it converge.

Metropolis-Hastings:
Given current state $\mathcal{Q}_i$ propose $\mathcal{Q}'$ with Prob. $p(\mathcal{Q}_i \rightarrow \mathcal{Q}')$
Accept $\mathcal{Q}_{i+1} := \mathcal{Q}'$ with probability

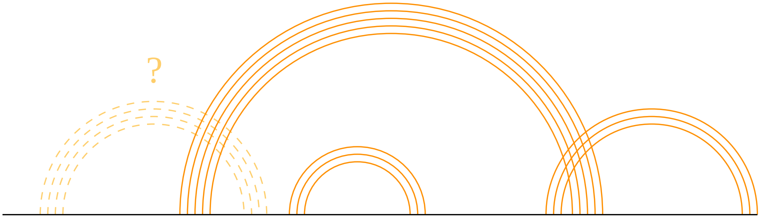
$$\min\left\{1, \frac{p(\mathcal{Q}' \rightarrow \mathcal{Q}_i) \cdot \Pr(\mathcal{Q}' \mid \mathcal{S})}{p(\mathcal{Q}_i \rightarrow \mathcal{Q}') \cdot \Pr(\mathcal{Q}_i \mid \mathcal{S})}\right\}$$

otherwise $\mathcal{Q}_{i+1} := \mathcal{Q}_i$

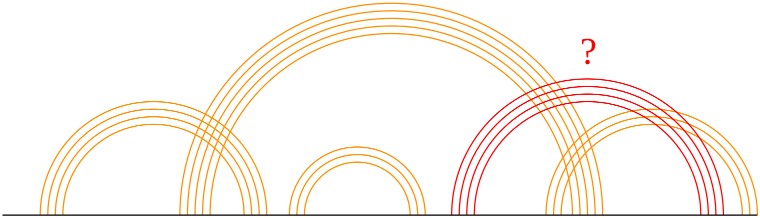
Proposals for $\mathcal{Q}_{i+1}$



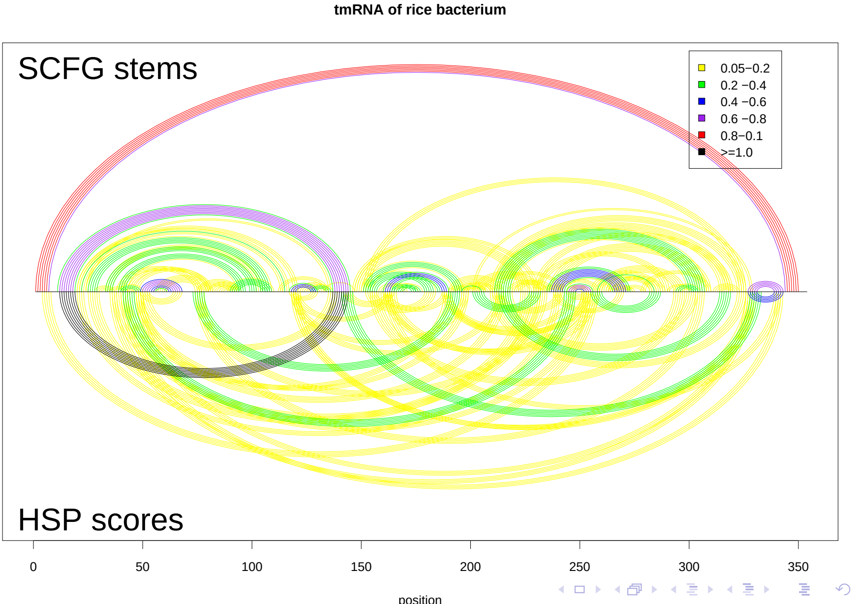
Proposals for $\mathcal{Q}_{i+1}$



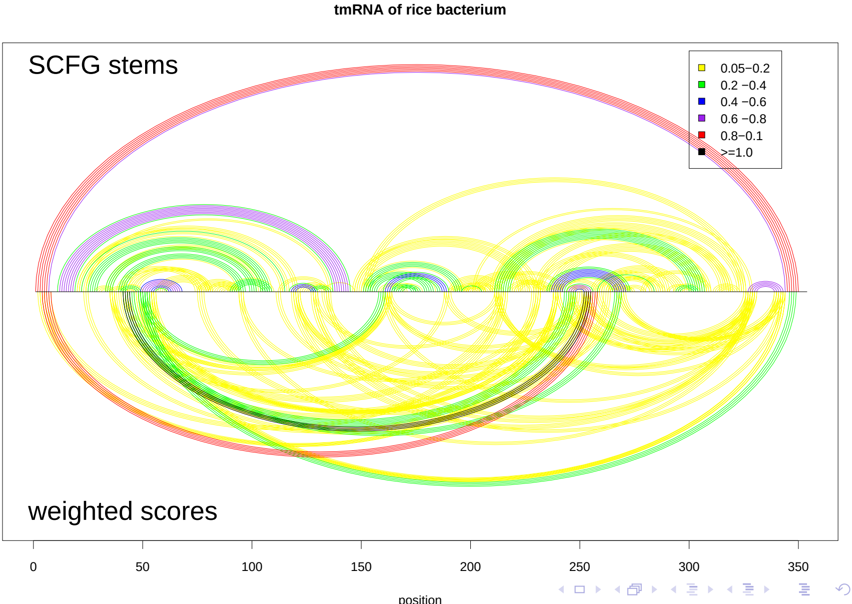
Proposals for $\mathcal{Q}_{i+1}$



Candidates for Pseudoknots



Candidates for Pseudoknots



Weight for $\mathcal{Q}$-stem proposal

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

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

Looking for Optima

We 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
   - Basic Model
   - Dynamic Programming
2. RNA folding with Pseudoknots
   - Pseudoknots
   - Combine SCFG with Pseudoknots
   - Bayesian Sampling of RNA structures with pseudoknots
3. Performance on Data
   - tmRNA
   - Simulated Data

Model parameter values used

$S \nearrow LS \quad 0.85$

$S \searrow L \quad 0.15$

$F \nearrow xLSy \quad 0.25$

$F \searrow xFy \quad 0.75$

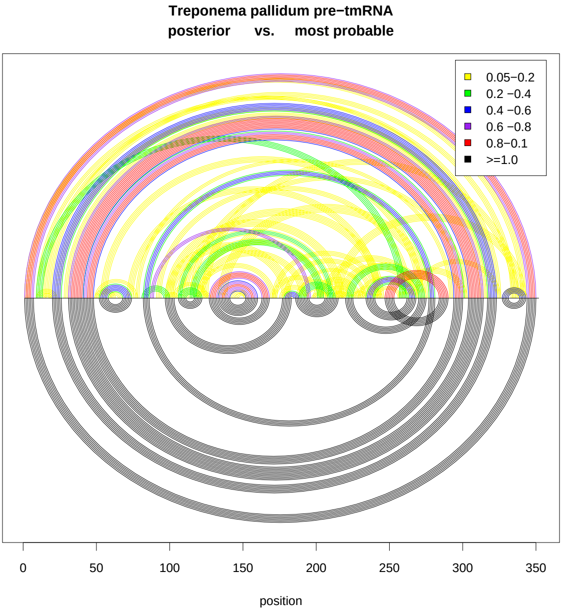
$L \nearrow x \quad 0.9$

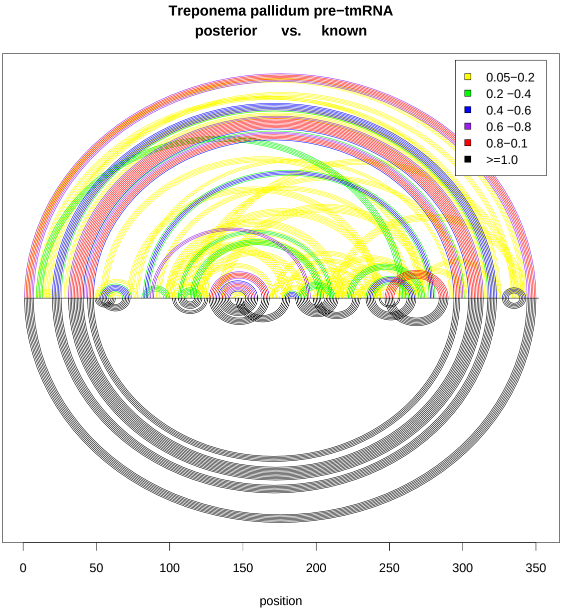
$L \rightarrow axFyb \quad 0.06$

$L \searrow Q \quad 0.04$

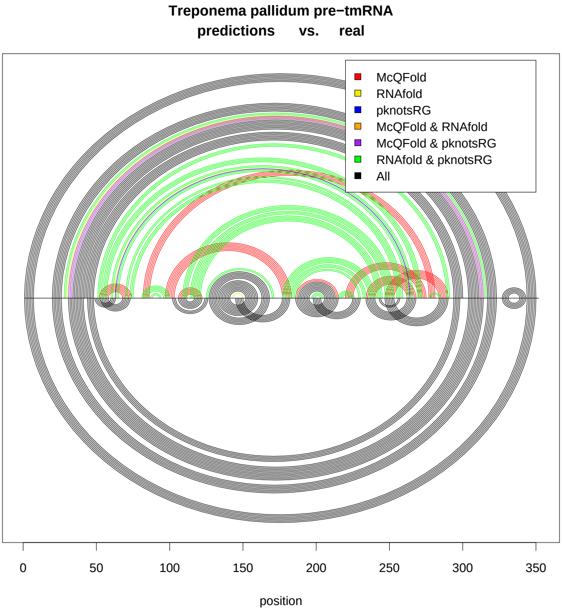
$$\pi_{CG} = \pi_{GC} = 27.5\%$$
$$\pi_{AT} = \pi_{TA} = 17.5\%$$
$$\pi_{GU} = \pi_{UG} = 5\%$$

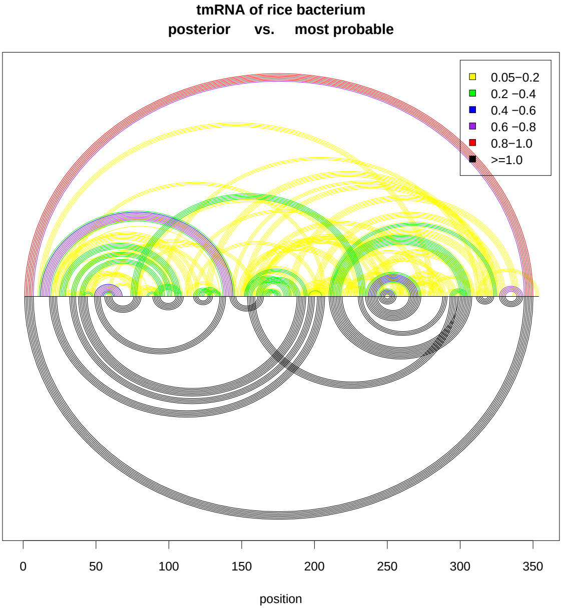
$$\pi_A = 35\%$$
$$\pi_C = 20\%$$
$$\pi_G = 20\%$$
$$\pi_U = 25\%$$


Treponema pallidum pre−tmRNA
posterior vs. most probable
🟨 0.05−0.2
🟩 0.2 −0.4
🟦 0.4 −0.6
🟪 0.6 −0.8
🟥 0.8−0.1
■ >=1.0
0
50
100
150
200
250
300
350
position

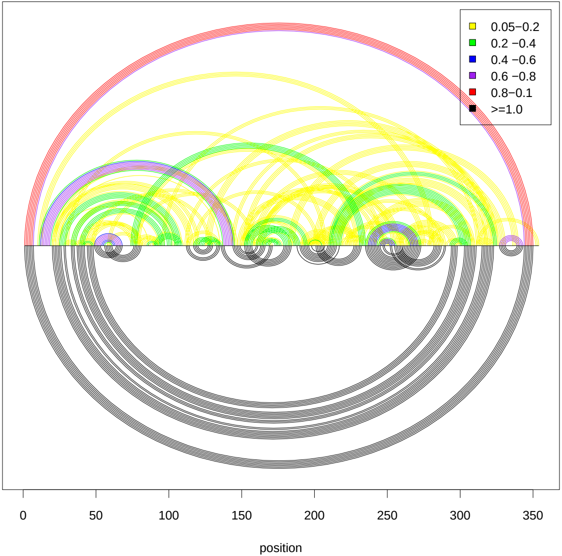


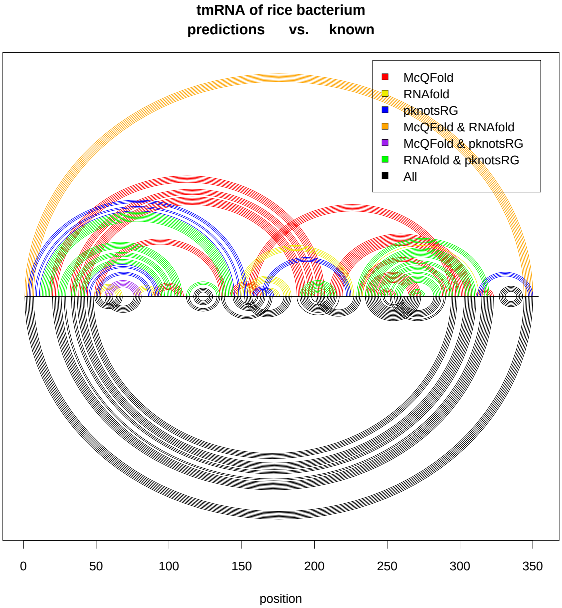
Treponema pallidum pre−tmRNA
posterior vs. known


Treponema pallidum pre−tmRNA
predictions vs. real
🟥 McQFold
🟨 RNAfold
🟦 pknotsRG
🟧 McQFold & RNAfold
🟪 McQFold & pknotsRG
🟩 RNAfold & pknotsRG
⬛ All
0
50
100
150
200
250
300
350
position


tmRNA of rice bacterium
posterior vs. most probable
🟨 0.05−0.2
🟩 0.2 −0.4
🟦 0.4 −0.6
🟪 0.6 −0.8
🟥 0.8−1.0
■ >=1.0
0
50
100
150
200
250
300
350
position

tmRNA of rice bacterium
posterior vs. known



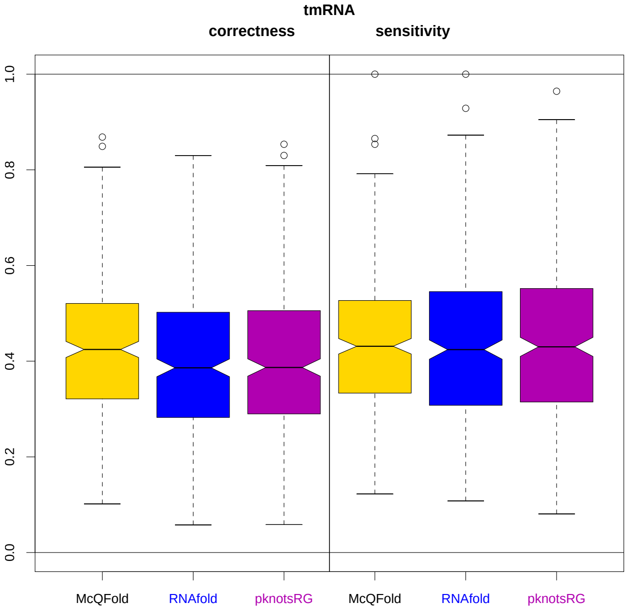

tmRNA of rice bacterium
predictions vs. known
🟥 McQFold
🟨 RNAfold
🟦 pknotsRG
🟧 McQFold & RNAfold
🟪 McQFold & pknotsRG
🟩 RNAfold & pknotsRG
■ All
0
50
100
150
200
250
300
350
position

Comparison with RNAfold and pknotsRG

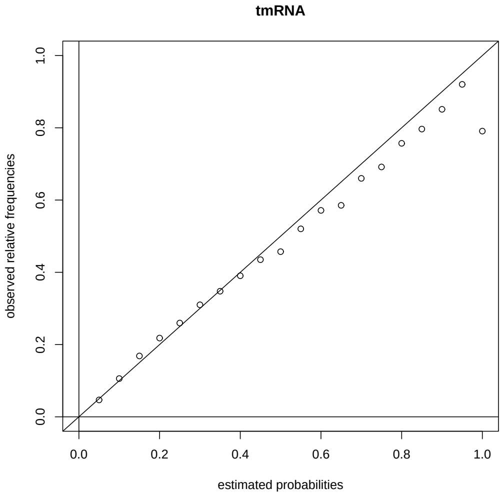
351 tmRNA Sequeces of length$>$ 200 from
`http://www.indiana.edu/~tmrna/`

	estimated to be paired	estimated *not* to be paired
in fact paired	A	a
in fact *not* paired	B	b

correctness: $A/(A+B)$
sensitivity: $A/(A+a)$


tmRNA
correctness
sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
McQFold
RNAfold
pknotsRG
McQFold
RNAfold
pknotsRG

estimated pairing probabilities



Outline

1. RNA folding with Stochastic Context-Free Grammars
   - Basic Model
   - Dynamic Programming
2. RNA folding with Pseudoknots
   - Pseudoknots
   - Combine SCFG with Pseudoknots
   - Bayesian Sampling of RNA structures with pseudoknots
3. Performance on Data
   - tmRNA
   - Simulated Data

Simulated Data

- 200 folded sequences of length 300-460
- generated according to our model
- same parameters as above
- McQFold uses same parameters
- of course unfair against RNAfold and pknotsRG

Simulated Data

- 200 folded sequences of length 300-460
- generated according to our model
- same parameters as above
- McQFold uses same parameters
- of course unfair against RNAfold and pknotsRG

Simulated Data

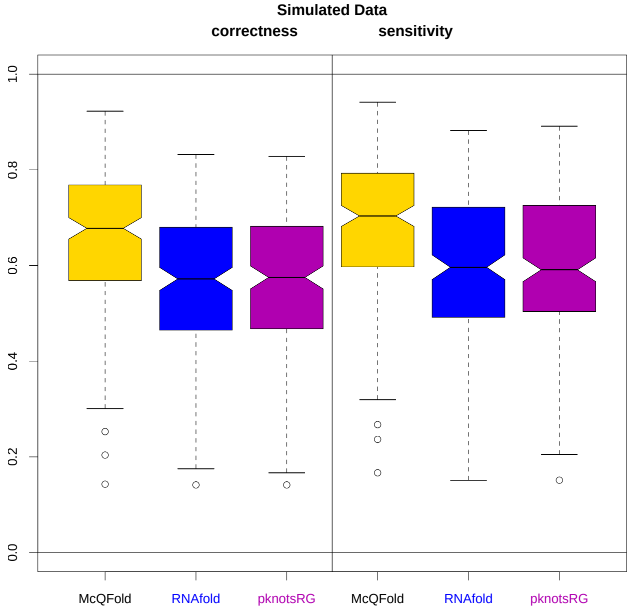
- 200 folded sequences of length 300-460
- generated according to our model
- same parameters as above
- McQFold uses same parameters
- of course unfair against RNAfold and pknotsRG

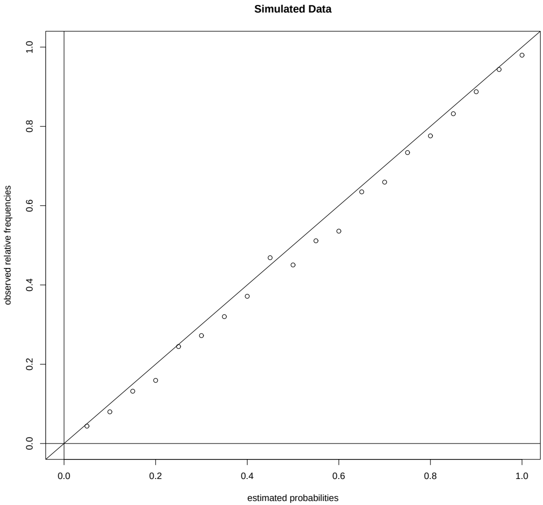
Simulated Data

- 200 folded sequences of length 300-460
- generated according to our model
- same parameters as above
- McQFold uses same parameters
- of course unfair against RNAfold and pknotsRG

Simulated Data

- 200 folded sequences of length 300-460
- generated according to our model
- same parameters as above
- McQFold uses same parameters
- of course unfair against RNAfold and pknotsRG


Simulated Data
correctness
sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
McQFold
RNAfold
pknotsRG
McQFold
RNAfold
pknotsRG


Simulated Data
observed relative frequencies
estimated probabilities
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0

Links

D. Metzler, M. Nebel (2006) Predicting RNA Secondary Structures with Pseudoknots by MCMC Sampling
submitted

Preprint:
`www.cs.uni-frankfurt.de/∼metzler/McQFold/McQFold.pdf`

Homepage of McQFold Software

`www.cs.uni-frankfurt.de/∼metzler/McQFold/`

Conclusions and Future Plans

- Estimation of RNA structure from sequence can be very uncertain.
- Uncertainty should be assessed. This can be done by Bayesian sampling.
- Perhaps better to combine informations from related sequences.
- Combine SCFGs with Q-stems to allow pseudoknots in structural alignments and/or structural sequence profiles.

Conclusions and Future Plans

- Estimation of RNA structure from sequence can be very uncertain.
- Uncertainty should be assessed. This can be done by Bayesian sampling.
- Perhaps better to combine informations from related sequences.
- Combine SCFGs with Q-stems to allow pseudoknots in structural alignments and/or structural sequence profiles.

Thanks: Martina Fröhlich, Christian Färber, Markus Nebel, audience