Blocked Gibbs Sampler for RNA Secondary Structure Prediction from Unaligned Sequences

Chip's Lab Meeting

Donglai Wei and Charles Lawrence[1]

[1]Division of Applied Mathematics,Brown University

Nov.10 2001

Roadmap

1. Background: RNA alignment and structure prediction
2. Algorithm: Blocked Gibbs Sampling
3. Results

Background: the Prediction Problem

1. RNA alignment(A) prediction
2. RNA consensus structure(S) prediction
3. A+S

1) RNA alignment(A) prediction: P(A|S)

a) No S included: profile HMM

b) Align individual S: RNAshape

c) Find and assemble stems: comRNA

d) SCFG grammmar: CM

2) RNA consensus structure(S) prediction: P(S|A)

a) Individual Structure+Mutal Information: RNAlifold

b) Maximum weighted Matching: MWM

3) A+S: P(A,S)

a) Dynamic Programming: Sankoff et.al

b) Iterate between A and S:

i) RNAsampler(stem)
ii) MASTR(MCMC local change)

Probablistic Model

$P(\vec{A}, \vec{S} | \Lambda_A, \Lambda_S, \vec{Q})$

Observation:

$\vec{Q}$: Sequences

Hidden Variables:

$\vec{A}$: Alignment

$\vec{S}$: Consensus Structure

Prior:

$\Lambda_{\vec{A}}$: Prior for Alignment model

$\Lambda_{\vec{S}}$: Prior for Structure model

Visualization:

Q_1: GCUAAUAUCGCUGUGGAAACACCUGGAACCAUCCCGAACCCAGCAGU

A_1:

```
U39694.1/9296-9392   .....GCUA.AUAUCGCUGUGGAAACACCUGGAACCAUCCCGAACCCAGC.AGU
M35167.1/2-111       UCCGGUGACUUUACGCGUGAGGAAACACUCGUUCCCAUUCCGAACACGAC.AGU
AJ131602.1/3-115     CCUGAUGACCAUAGCGACGUGGUACCACUCCUUCCCAUCCCGAACAGGAC.AGU
X52302.1/2-117       GCUGGUGGCUAUGGCGGAAGGGCCACACCCGAUCCCAUCCCGAACUCGGU.CGU
Y00159.1/2-117       CUUGGCGACUAUAGCGAUUUGGAACCACCUGAUACCAUCUCGAACUCAGA.AGU
M58387.1/5-112       ....ACGGCCACAGUCAGCUGAAAACUGGGCAUCCCGUCC.GCUCUGCCAUACA
X67494.1/1-118       AUCCUCGGCCAUAGAAUGACGAAAACGACGCGUCCCGUCC.GAUCUGCGA.AUC
M36316.1/2-120       UCUUACGGCCAUUCACACCAGAAAGCACCAAAUCCCGUCC.GAUCUUUGA.AGU
X02706.1/2-120       UGGAUCGUUCAAACCUUCAAGGCCCCUCCCCAUCCCAUCA.GCACUGGGA.AGA
X05535.1/1-118       AGGAACGGCCAUACCACGUCGAUCGCACCACAUCCCGUCC.GCUCUGUGA.AGU
#=GC SS_cons    S:   <<<<<<<<<<...<<.<<<<<...<<..<<<<<<.......>>..>>>>..>>
```

S_1: <<<<...<<.<<<<<...<<..<<<<<<.......>>..>>>>..>>

Blocked Gibbs Sampling

1. Initial alginment: $\vec{A}^0$

2. Iteration:

a) Sample $\vec{S}^{t+1}$ from $\mathrm{P}(\vec{S}^{t+1}|\vec{A}^t)$

b) Sample $\vec{A}^{t+1}$ from $\mathrm{P}(\vec{A}^{t+1}|\vec{S}^t)$

Cluster Analysis upon samples of S

1. Generalized Centroid Estimator
2. Bias-Variance
3. Credibility Limit
4. Distance between centroids

Test Cases

1. 85 alignments from 17 RNA family (Kiryu et.al.)
   a. PPV-SEN curve for γ-centroid
   b. Effects of number of sequences in the alignment
   c. Detailed look into each family
2. Riboswitch Detection

1.0) Kiryu's Data

From 17 RNA famililies: tRNA,5sRNA,THI,...

5 Subalignments for each Family

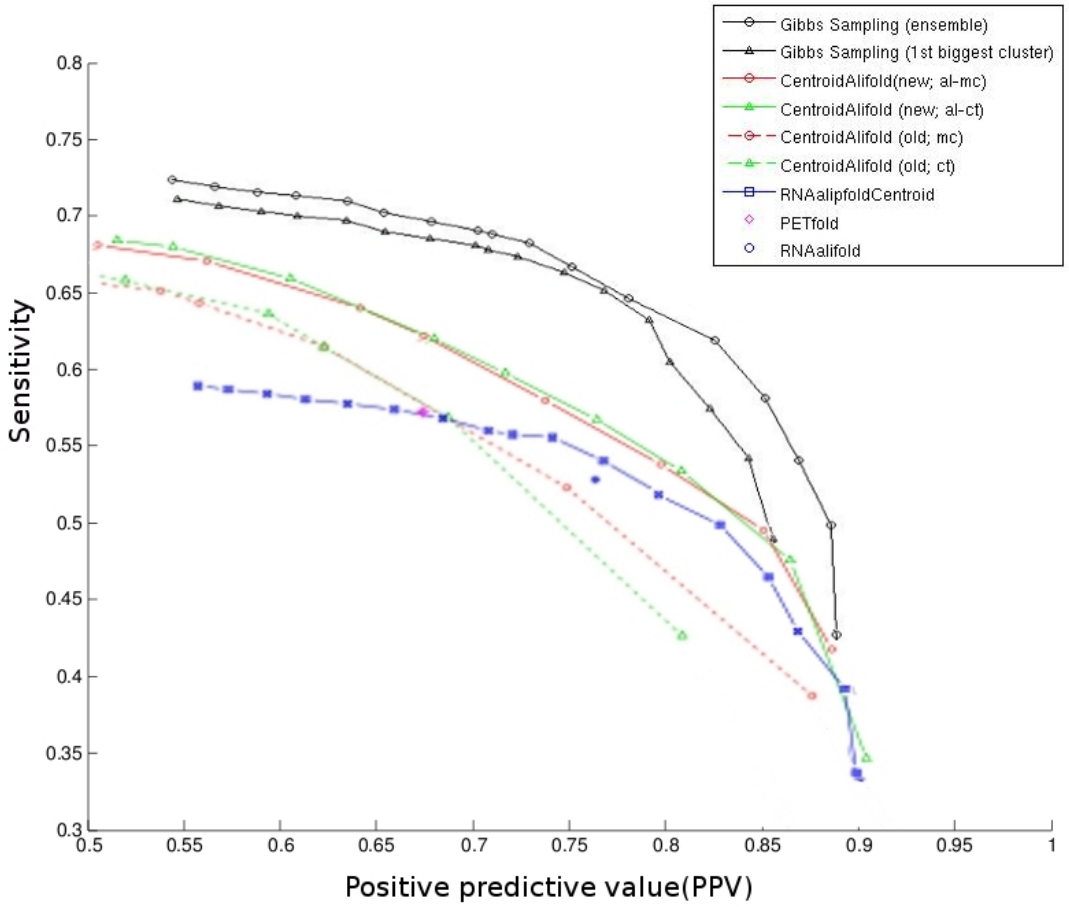
10 homologous Sequences for each Subalignment

```
# STOCKHOLM 1.0


#=GF AC  RF00001:0
#=GF ID  5S_rRNA:0

U39694.1/9296-9392  .....GCUA.AUAUCGCUGUGGAAACACCUGGAACCAUCCCGAACCCAGC.AGU
M35167.1/2-111      UCCGGUGACUUUACGCGUGAGGAAACACUCGUUCCCAUUCCGAACACGAC.AGU
AJ131602.1/3-115    CCUGAUGACCAUAGCGACGUGGUACCACUCCUUCCCAUCCCGAACAGGAC.AGU
X52302.1/2-117      GCUGGUGGCUAUGGCGGAAGGGCCACACCCGAUCCCAUCCCGAACUCGGU.CGU
Y00159.1/2-117      CUUGGCGACUAUAGCGAUUUGGAACCACCUGAUACCAUCUCGAACUCAGA.AGU
M58387.1/5-112      ....ACGGCCACAGUCAGCUGAAAACUGGGCAUCCCGUCC.GCUCUGCCAUACA
X67494.1/1-118      AUCCUCGGCCAUAGAAUGACGAAAACGACGCGUCCCGUCC.GAUCUGCGA.AUC
M36316.1/2-120      UCUUACGGCCAUUCACACCAGAAAGCACCAAAUCCCGUCC.GAUCUUUGA.AGU
X02706.1/2-120      UGGAUCGUUCAAACCUUCAAGGCCCCUCCCCAUCCCAUCA.GCACUGGGA.AGA
X05535.1/1-118      AGGAACGGCCAUACCACGUCGAUCGCACCACAUCCCGUCC.GCUCUGUGA.AGU
#=GC SS_cons        <<<<<<<<<.,,,<<.<<<<<.,,<<.,<<<<<<.,,,,,,>>.,>>>>.,,>>
```

1.1) PPV-SEN curve



1.2) Varying number of Sequences in the Alignment

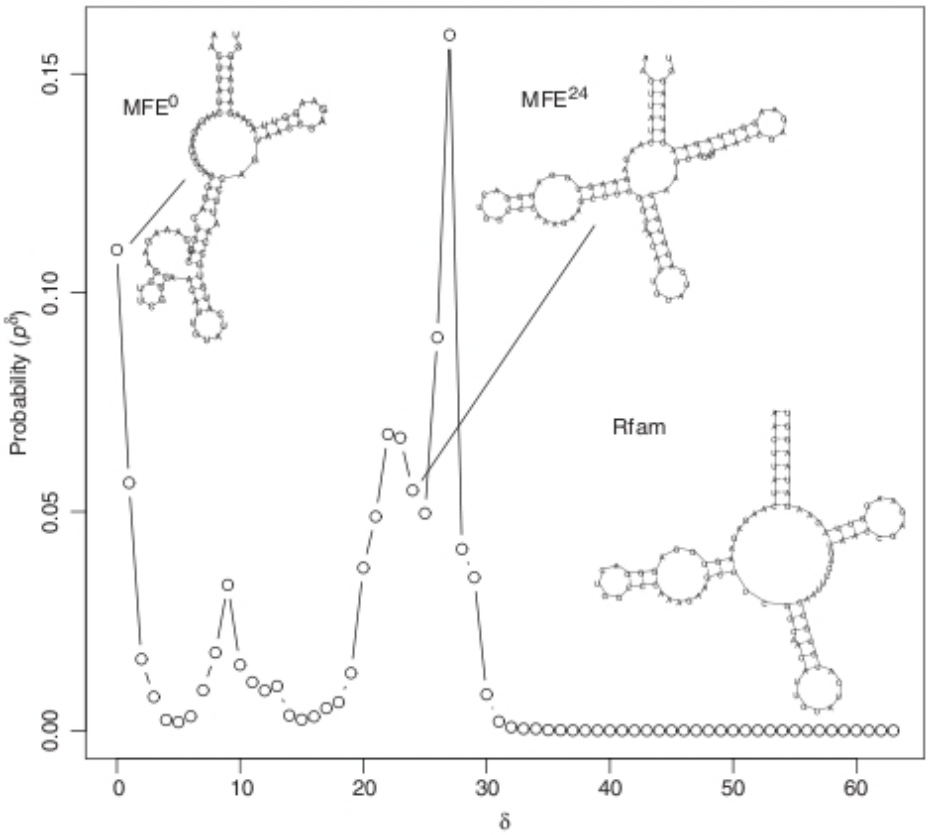
Table 1: effects of the number of sequences

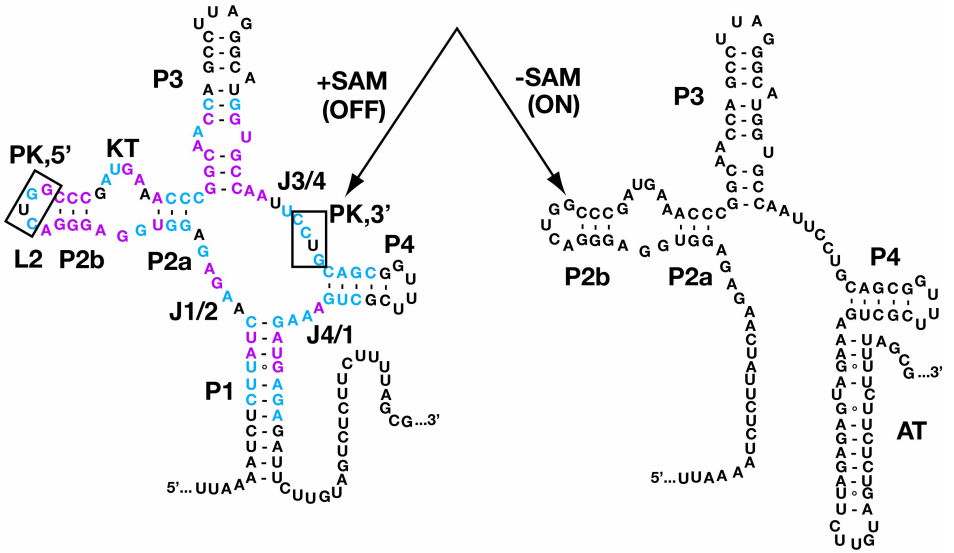
	Area under PPV-SEN					#samples			95% Credibility Limit		
#seqs	ensemble	1st cluster	2nd cluster	Bias	Std	1st cluster	2nd cluster	All	ensemble	1st cluster	2nd cluster
2	0.45	0.45	0.41	0.27	0.04	722.35	149.45	871.8	0.21	0.14	0.11
3	0.61	0.60	0.55	0.20	0.03	788.78	122.64	911.42	0.14	0.10	0.07
4	0.62	0.61	0.57	0.20	0.03	806.71	114.86	921.57	0.14	0.09	0.06
5	0.65	0.65	0.59	0.17	0.03	816.28	114.04	930.32	0.12	0.08	0.05
6	0.68	0.67	0.59	0.16	0.03	806.32	113.45	919.77	0.11	0.07	0.05
7	0.70	0.68	0.64	0.15	0.03	794.54	111.43	905.97	0.10	0.07	0.05
8	0.70	0.69	0.67	0.15	0.03	792.66	114.91	907.57	0.10	0.07	0.04
9	0.70	0.69	0.65	0.14	0.02	793.52	122.76	916.28	0.09	0.06	0.04
10	0.72	0.71	0.66	0.13	0.02	792.85	125.11	917.96	0.09	0.06	0.04

1.3) Look into each Family

RNA Family	RNA type	length	Bias	Std	Credibility Limit			ROC Area			# Samples			Centroid distance
					ensemble	1st cluster	2nd cluster	ensemble	1st cluster	2nd cluster	All	1st cluster	2nd cluster	
T-box	tRNA	244	0.1	0.01	0.06	0.04	0.02	0.54	0.52	0.47	926	826	100	0.0613
t-RNA	tRNA	73	0.02	0.01	0.03	0.01	0.01	1	0.99	0.91	949	888	61	0.0493
5S-rRNA	rRNA	116	0.17	0.02	0.07	0.05	0.03	0.7	0.7	0.67	922	751	171	0.0657
5-8S-rRNA	rRNA	154	0.18	0.03	0.14	0.1	0.08	0.41	0.4	0.26	907	744	163	0.1042
Retroviral-psi	Rviral	117	0.07	0.05	0.15	0.11	0.05	0.99	0.99	0.45	981	952	29	0.1997
U1	sRNA	157	0.16	0.02	0.06	0.06	0.02	0.69	0.69	0.63	988	928	60	0.093
U2	sRNA	182	0.08	0.02	0.05	0.05	0.02	0.91	0.9	0.7	981	941	40	0.081
sno-14q-I-II	sRNA	75	0.07	0.03	0.12	0.08	0.07	1	0.91	0.84	838	636	202	0.0749
Lysine	riboswitch	181	0.07	0.02	0.06	0.05	0.03	0.93	0.93	0.83	983	923	60	0.0693
RFN	riboswitch	140	0.15	0.03	0.11	0.06	0.06	0.67	0.64	0.59	820	574	246	0.07
THI	riboswitch	105	0.08	0.02	0.07	0.06	0.02	0.89	0.88	0.75	968	869	99	0.0936
S-box	riboswitch	107	0.09	0.02	0.07	0.03	0.03	0.88	0.87	0.74	945	806	139	0.0682
IRES-HCV	cis	261	0.25	0.05	0.21	0.16	0.08	0.6	0.57	0.44	936	877	59	0.2435
SECIS	cis	64	0.17	0.02	0.08	0.02	0.02	0.74	0.71	0.72	840	679	161	0.0609
UnaL2	cis	54	0.18	0.03	0.06	0.02	0.02	0.62	0.62	0.61	867	752	115	0.0426
SRP-bact	srpRNA	93	0.16	0.03	0.12	0.04	0.04	0.79	0.78	0.7	834	646	188	0.111
SRP-euk-arch	srpRNA	291	0.23	0.01	0.04	0.03	0.02	0.49	0.48	0.47	921	837	84	0.0407
avg		142	0.13	0.02	0.09	0.06	0.04	0.76	0.74	0.63	926	826	100	0.10

2) Side words about Riboswitch





Take Home Message

1. Sampling, a glimpse of the complicated probability space

2. Reference Structure, a dream never comes true

Acknowledgement

1. Thanks Chip for opening the world of computation to me

2. Thanks Bill for endless technical support

3. Thanks Everyone here for enduring the torture of my presentation :p