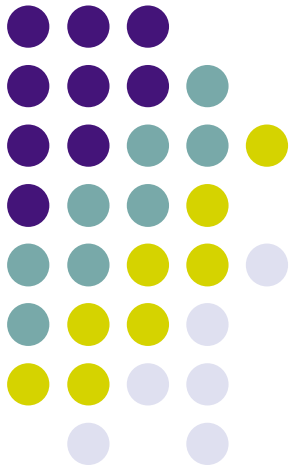
RNA: partition function

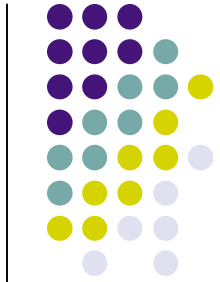
(Institute of Mathematical Sciences, National University of Singapore, 24 July 2007)



P. Clote

Biology and Computer Science

Boston College



Partition function for RNA secondary structure

Boltzmann probability of sequence alignment

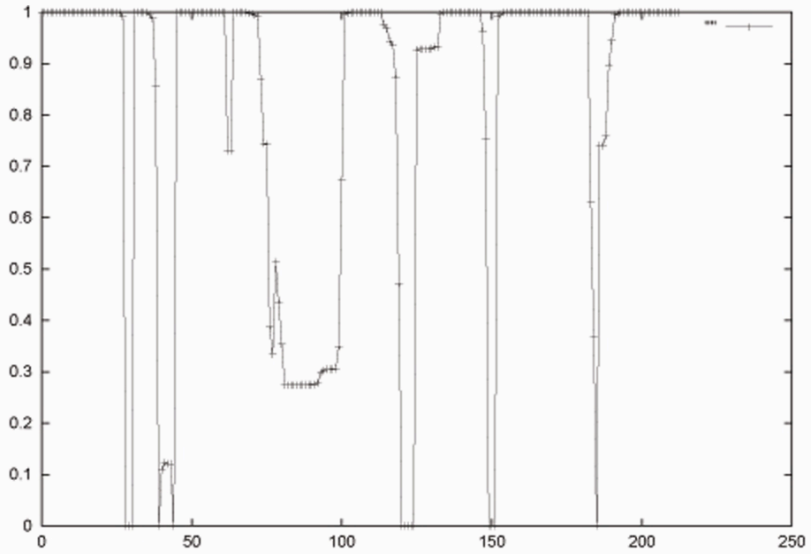
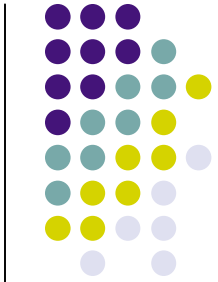


Figure 1: Local alignment Boltzmann probabilities of portions of bovine trypsin and pig elastase (see text)

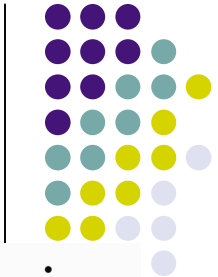
McCaskill's Algorithm



- Partition function

$$Z = \sum_{S \in Q} e^{-\frac{E(S)}{RT}}$$

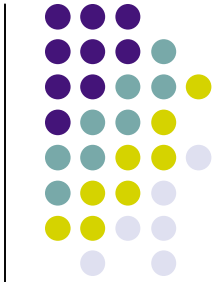
where Q is a set of all possible *states*, $R = 8.3146$ joules per degree Kelvin, or since there are 4.184 joules per calorie, $R \approx 2$ calories per degree Kelvin, and T is absolute temperature in degrees Kelvin. Sometimes, one encounters the Boltzmann constant $k = 13.805 \times 10^{-24}$ J K^{-1} or about 3.3×10^{-24} cal K^{-1} ($k = R/N$, where N is Avogadro's number 6.0229×10^{23}.



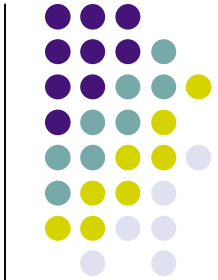
Boltzmann probability of a given secondary structure S_0 is

$$Pr[S_0] = \frac{e^{-\frac{E(S_0)}{RT}}}{Z}$$

Clearly the partition function Z is simply a normalization factor ensuring that the Boltzmann probabilities sum to 1.

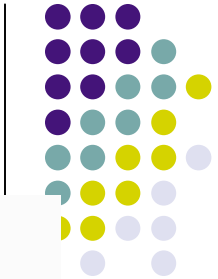


McCaskill's key idea: Additivity of the energy function implies multiplicativity of the partition function. This is best seen by a small example.



Suppose that $s_1, \ldots, s_{k-1}$ has only two secondary structures A,B and $s_{k+1}, \ldots, s_{n-1}$ has only two secondary structures C,D. The partition function contribution, Z^*, for the subensemble of structures where k, n are base paired is

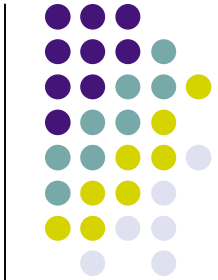
$$
\begin{aligned}
= \quad & e^{-a(k,n)/RT} \cdot \\
& \{ e^{-E(A)/RT} \cdot e^{-E(C)/RT} + e^{-E(A)/RT} \cdot e^{-E(D)/RT} + \\
& e^{-E(B)/RT} \cdot e^{-E(C)/RT} + e^{-E(B)/RT} \cdot e^{-E(D)/RT} \} \\
= \quad & e^{-a(k,n)/RT} \cdot \\
& \{ e^{-E(A)/RT} \cdot \left(e^{-E(C)/RT} + e^{-E(D)/RT} \right) + \\
& e^{-E(B)/RT} \cdot \left(e^{-E(C)/RT} + e^{-E(D)/RT} \right) \} \\
= \quad & e^{-a(k,n)/RT} \cdot \\
& \{ \left(e^{-E(A)/RT} + e^{-E(B)/RT} \right) \cdot \left(e^{-E(C)/RT} + e^{-E(D)/RT} \right) \} \\
= \quad & e^{-a(k,n)/RT} \cdot Z_{1,k-1} \cdot Z_{k+1,n-1}
\end{aligned}
$$



Partition function computation (Nussinov-Jacobson energy)

```
for i=0 to n − 1
  for j=i + 1 to n − 1
    Z(i,j) = 1;  //partition function for empty structure
for d = μ + 1 to n − 1
  for i = 0 to n − 1
     j = i + d;
     if (j < n)
      sum = Z(i, j − 1); //Case 1:  j unpaired
      if (j − i ≥ μ)
        sum+ = e^{−a(i,j,S)/RT} · Z(i + 1, j − 1); //Case 2:  i,j paired
        for k = i+1 to j-μ
           sum+ = e^{−a(k,j,S)/RT} · Z(i, k − 1) · Z(k + 1, j − 1);
           //Case 3:  k,j paired some intermediate k
        Z(i,j) = sum;
```

Probability that i,j base pair (Nussinov-Jacobson energy model)

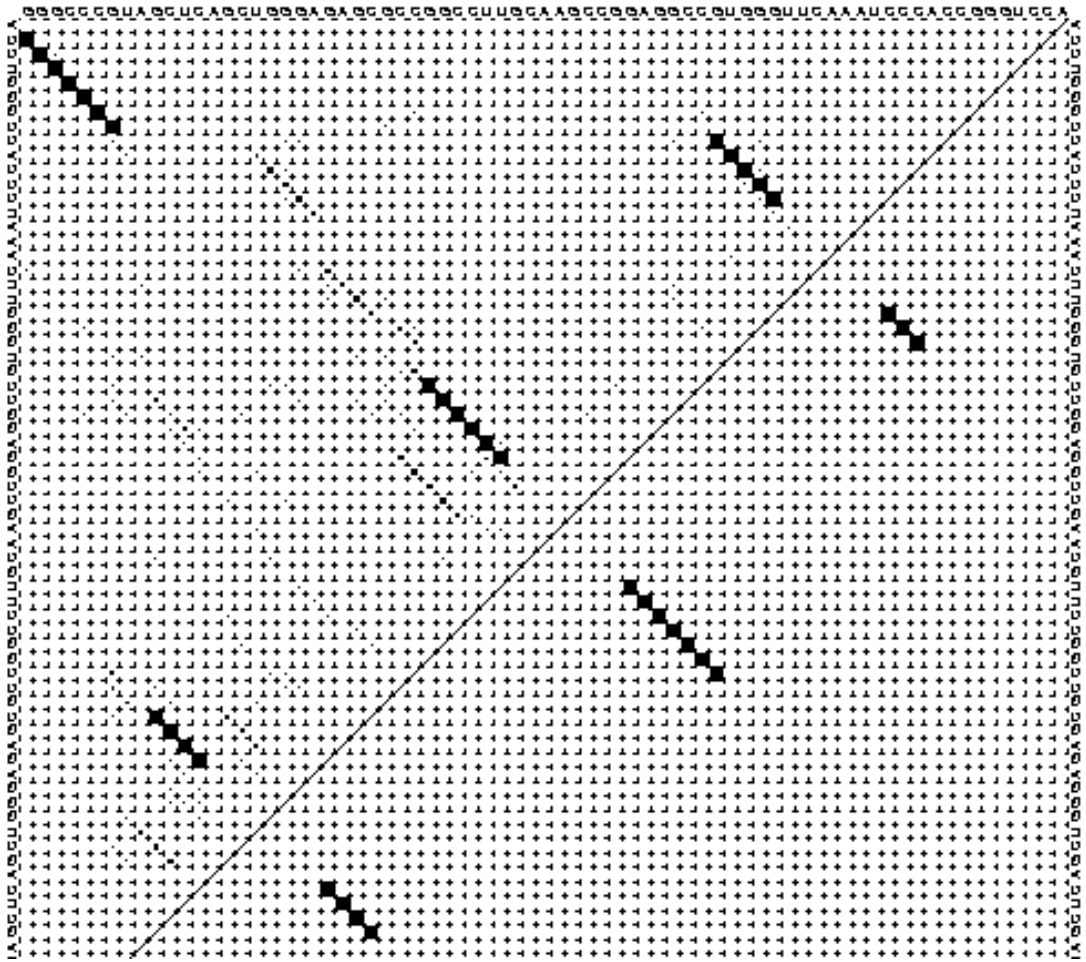


$$p_{i,j} = \sum_{\{i,j\} \in S} Pr[S]$$

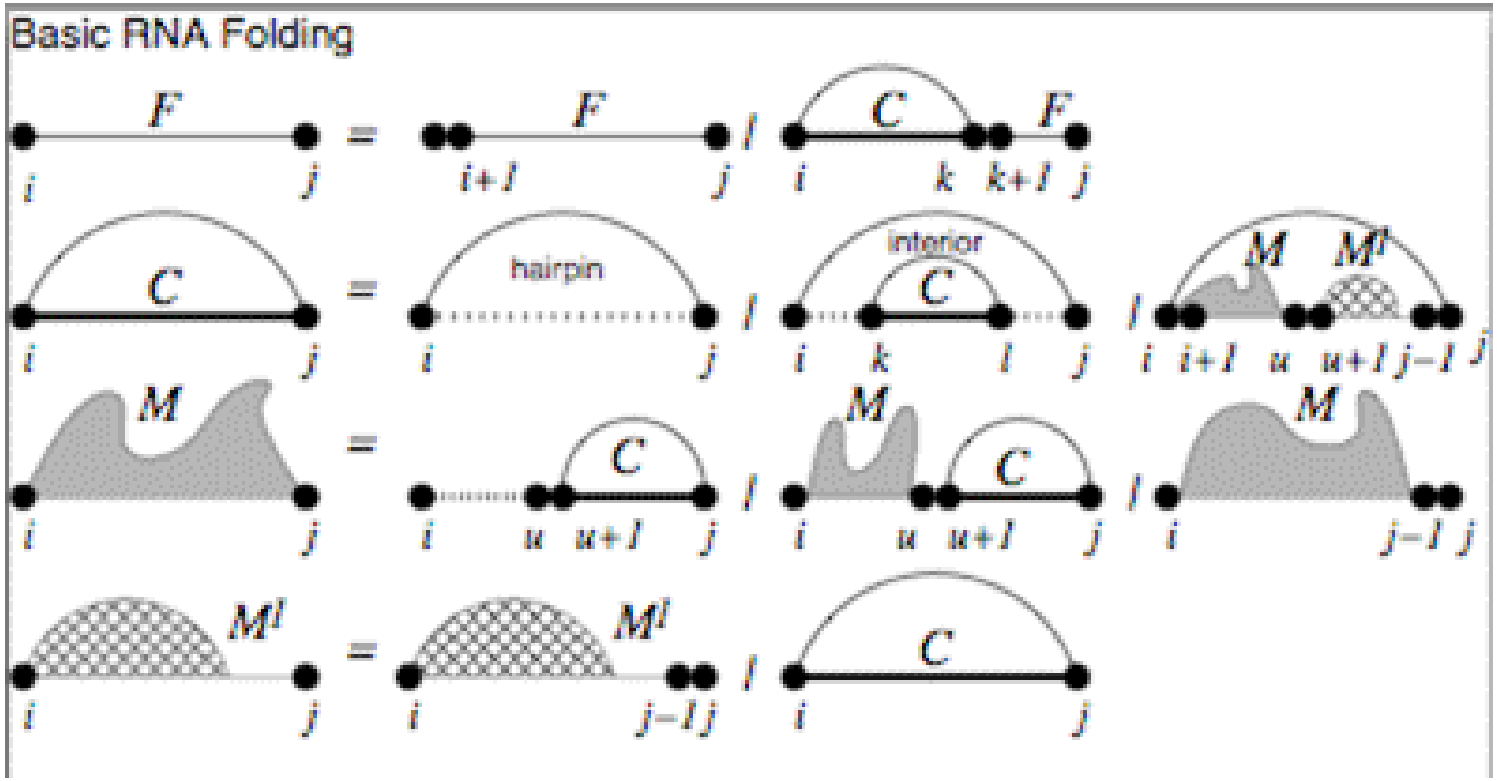
$$p_{i,j} = \frac{Z_{1,i-1} Z^b_{i,j} Z_{j+1,n}}{Z_{1,n}}.$$

Above gives idea, but is missing one case.

McCaskill's algorithm base pairing probabilities for Ala tRNA using Vienna RNA package



“Feynman diagram of recursions for Zuker’s algorithm”



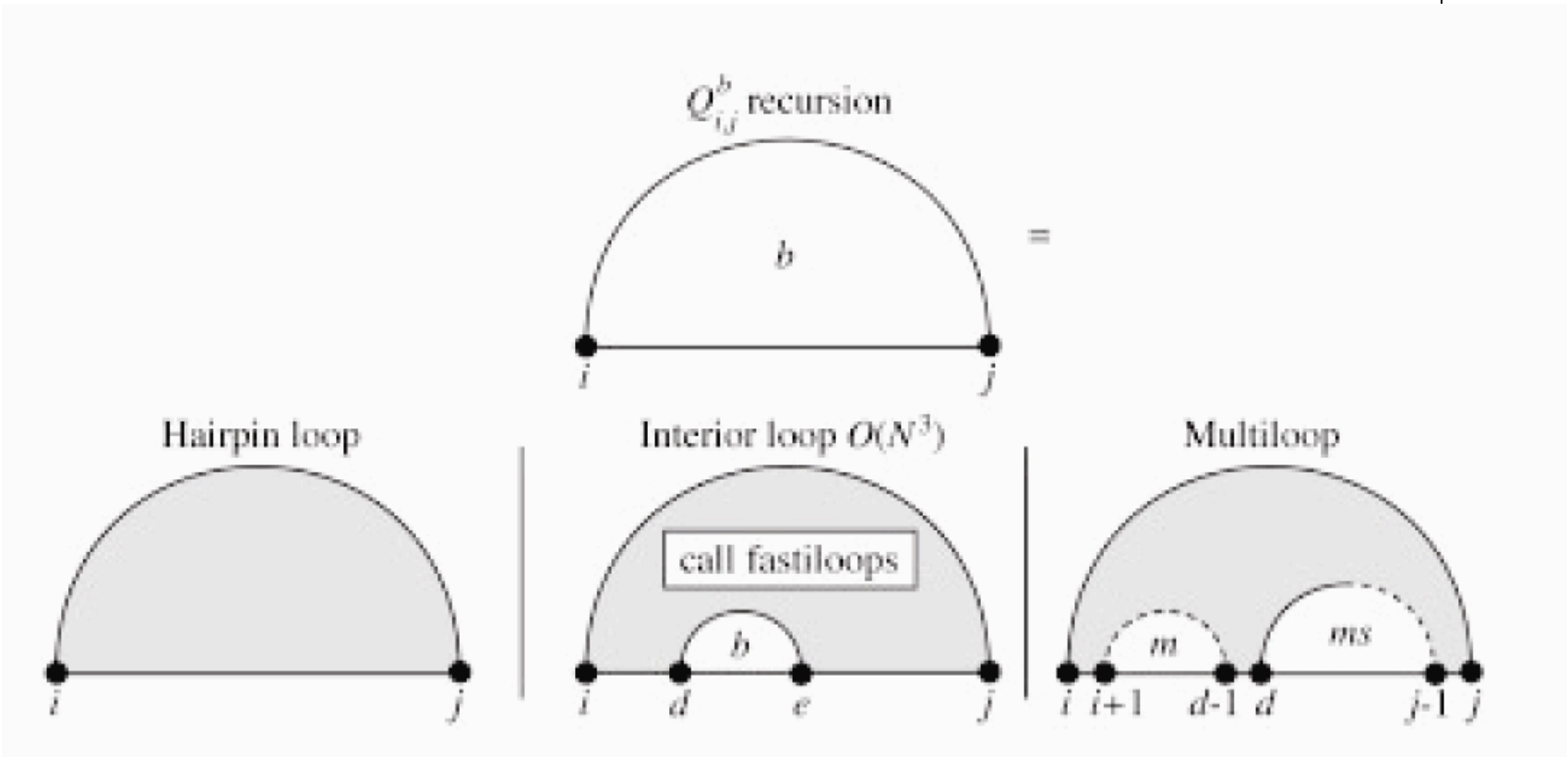
Variations on RNA Folding and Alignment,
Athanasius F. Bompfuenewer, J Math Biol (2007) in press.

Vienna Package implementation of McCaskill's algorithm

$$
\begin{aligned}
Z_{ij} &= Z_{i+1,j} + \textstyle\sum_{i<k\leq j} Z^B_{ik} Z_{k+1,j} \\
Z^B_{ij} &= e^{-\beta \mathcal{H}(i,j)} + \textstyle\sum_{i<k<l<j} Z^B_{kl} e^{-\beta \mathcal{I}(i,j;k,l)} + \textstyle\sum_{i<u<j} Z^M_{i+1,u} Z^{M1}_{u+1,j-1} e^{-\beta a} \\
Z^M_{ij} &= \textstyle\sum_{i<u<j} e^{-\beta(u-i+1)c} Z^M_{u+1,j} + \textstyle\sum_{i<u<j} Z^M_{i,u} Z^B_{u+1,j} e^{-\beta b} + Z^M_{i,j-1} e^{-\beta c} \\
Z^{M1}_{ij} &= Z^{M1}_{i,j-1} e^{-\beta c} + Z^B_{ij} e^{-\beta b} \\
& Z_{ii} = 1,\ Z^B_{ii} = Z^M_{ii} = Z^{M1}_{ii} = 0
\end{aligned}
$$

Variations on RNA Folding and Alignment,
Athanasius F. Bompfuenewer, J Math Biol (2007) in press.

McCaskill's algorithm



Dirks, Pierce, "**A partition function algorithm for nucleic acid secondary structure including pseudoknots***J Comput Chem*, 24(13):1664-1677, 2003

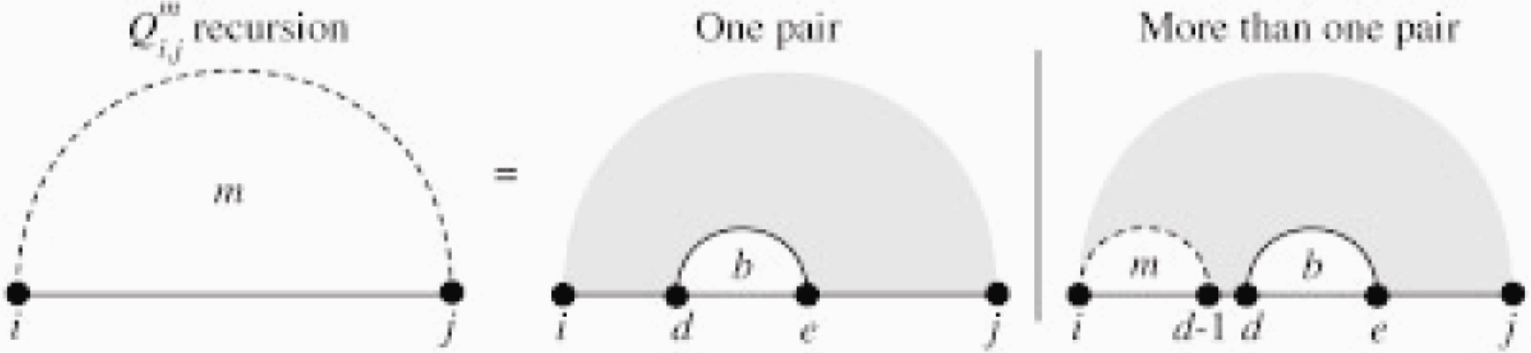
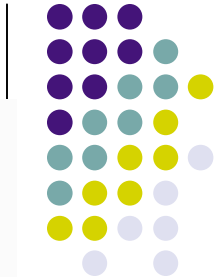


Figure 5. $O(N^4)$ Algorithm: recursion for $Q^m_{i,j}$, the partition function for the subsequence $[i, j]$ inside a multiloop when there is at least one base pair in the subsequence. Either there is only one more base pair $d \cdot e$ defining the multiloop and the recursion energy is $\alpha_2 + \alpha_3(d - i) + \alpha_3(j - e)$, or there is more than one pair with rightmost pair $d \cdot e$ and recursion energy $\alpha_2 + \alpha_3(j - e)$.

Dirks, Pierce, "**A partition function algorithm for nucleic acid secondary structure including pseudoknots***J Comput Chem*, 24(13):1664-1677, 2003



Initialize (Q, Q^b, Q^m) // $O(N^2)$ space
Set all values to 0 except $Q_{i,i-1} = 1$
for $l = 1, N$
  for $i = 1, N-l+1$
    $j = i+l-1$
    // Q^b **recursion**
    $Q^b_{i,j} = \exp\{-G^{\text{hairpin}}_{i,j}/RT\}$
    for $d = i+1, j-5$ // loop over all possible rightmost pairs $d \cdot e$
      for $e = d+4, j-1$
        $Q^b_{i,j}$ += $\exp\{-G^{\text{interior}}_{i,d,e,j}/RT\}\, Q^b_{d,e}$
        $Q^b_{i,j}$ += $Q^m_{i+1,d-1}\, Q^b_{d,e}\, \exp\{-[\alpha_1 + 2\alpha_2 + \alpha_3(j-e-1)]/RT\}$
    // Q, Q^m **recursions**
    $Q_{i,j} = 1$ //empty recursion
    for $d = i, j-4$ // loop over all possible rightmost pairs $d \cdot e$
      for $e = d+4, j$
        $Q_{i,j}$ += $Q_{i,d-1}\, Q^b_{d,e}$
        $Q^m_{i,j}$ += $\exp\{-[\alpha_2 + \alpha_3(d-i) + \alpha_3(j-e)]/RT\}\, Q^b_{d,e}$
        $Q^m_{i,j}$ += $Q^m_{i,d-1}\, Q^b_{d,e}\, \exp\{-[\alpha_2 + \alpha_3(j-e)]/RT\}$
//Partition function is $Q_{1,N}$

Dirks, Pierce, "**A partition function algorithm for nucleic acid secondary structure including**
J Comput Chem, 24(13):1664-1677, 2003

Partition function for pseudoknots (restricted class)

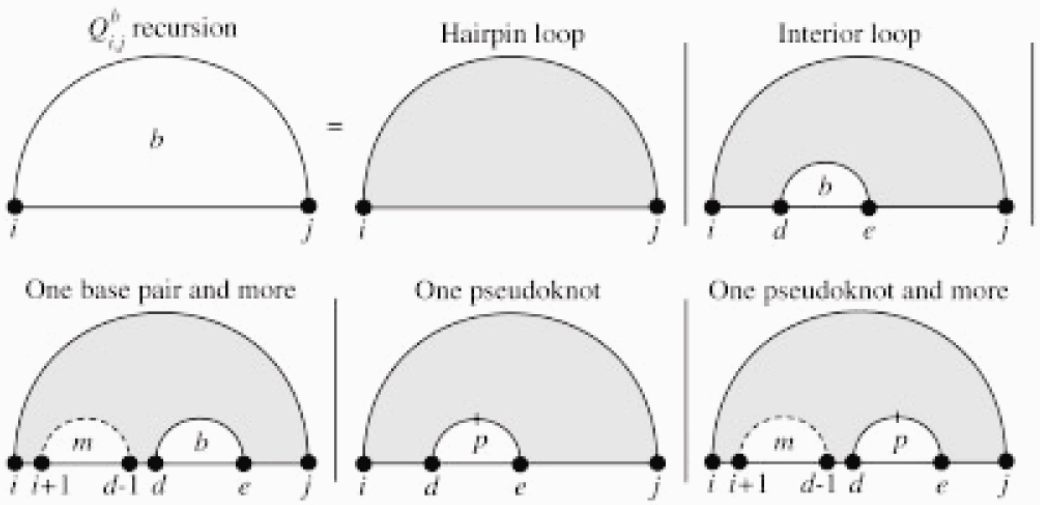


Figure 13. $O(N^8)$ Algorithm: recursion for $Q^b_{i,j}$, the partition function for the subsequence $[i, j]$ assuming i and j are base-paired. Either the subsequence $[i, j]$ is a hairpin loop with recursion energy $G^{\text{hairpin}}_{i,j}$, or there exists one internal base pair $d \cdot e$ forming an interior loop with recursion energy $G^{\text{internal}}_{i,d,e,j}$, or there is more than one base pair or pseudoknot (with rightmost pair $d \cdot e$) forming a multiloop with recursion energy $\alpha_1 + 2\alpha_2 + \alpha_3(j - e - 1)$, or there is one pseudoknot filling the subsequence $[d, e]$ with recursion energy $\alpha_1 + \beta^m_1 + 3\alpha_2 + \alpha_3(d - i - 1) + \alpha_3(j - e - 1)$, or there is more than one pseudoknot or base pair (with rightmost pseudoknot filling the subsequence $[d, e]$) with recursion energy $\alpha_1 + \beta^m_1 + 3\alpha_2 + \alpha_3(j - e - 1)$.

Dirks, Pierce, "**A partition function algorithm for nucleic acid secondary structure including pseudoknots",** *J Comput Chem*, 24(13):1664-1677, 2003

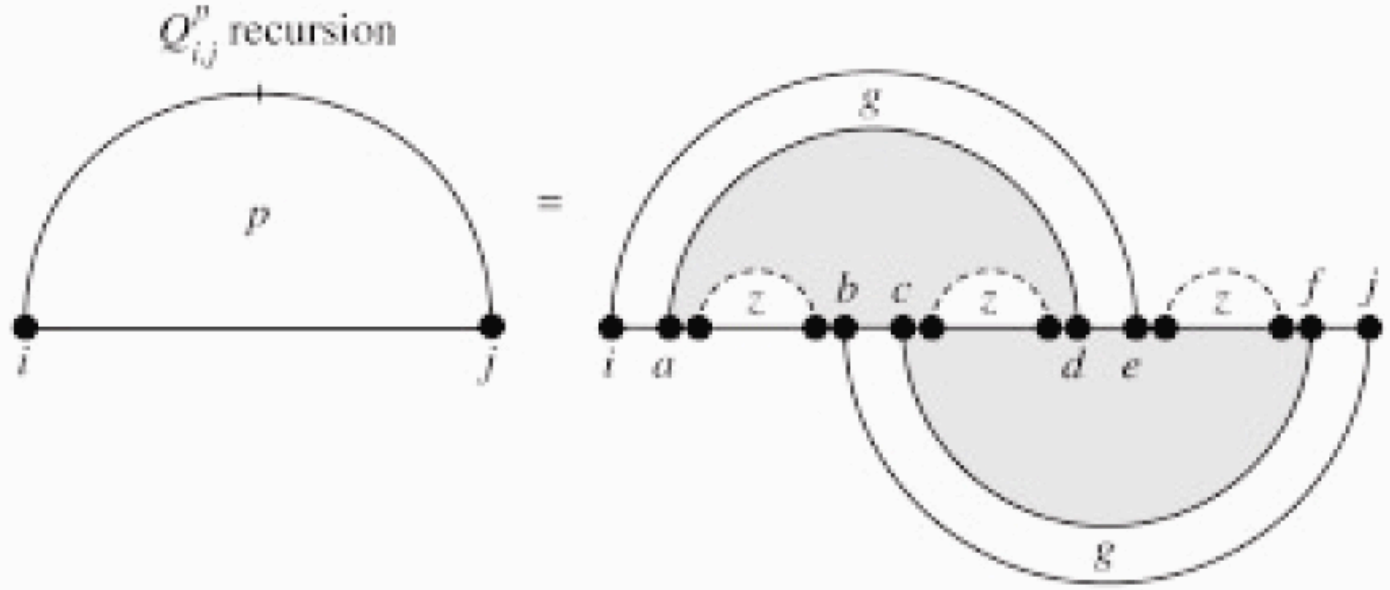


Figure 15. $O(N^8)$ Algorithm: recursion for $Q^p_{i,j}$, the partition function for the pseudoknot filling the subsequence $[i, j]$. The recursion energy is $2\beta_2$, where β_2 is the penalty associated with each base pair bordering the interior of the pseudoknot.

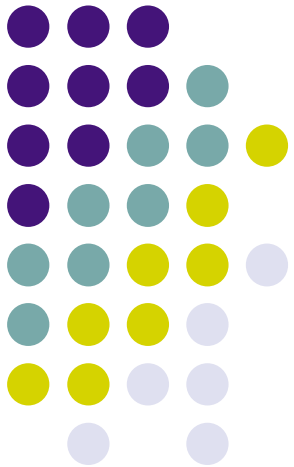
Dirks, Pierce, "**A partition function algorithm for nucleic acid secondary structure including pseudoknots",** *J Comput Chem*, 24(13):1664-1677, 2003

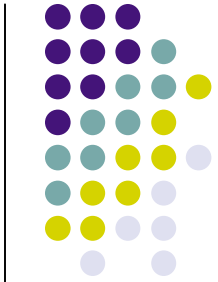
Energy minimization for several species

- “Prediction of hybridization and melting for double-stranded nucleic acids”, Dimitrov, Zuker, Biophysical Journal 87:215-226 (2004) – computation of Boltzmann partition function

- “Secondary structure prediction of interacting RNA molecules”, Andronescu, Zhang, Condon, JMB 345:987-1001 (345)

Applications of Boltzmann partition function





- Base pairs (i,j) with high Boltzmann pair probabilities are more likely to be in the phylogenetic structure (comparative sequence analysis) – observation of D. Mathews in RNA 10:1178-1190 (2004)
- Similar to observation of Vingron and Argos that positions i,j in optimal pairwise sequence alignment with high Boltzmann probability are biologically relevant positions.

Higher probability base pairs are more likely to be in phylogenetic structure

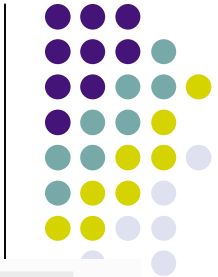


TABLE 2. Sensitivity and positive predictive value for MFE structure prediction

Type of RNA	MFE sensitivity[h]	MFE positive predictive value	Positive predictive value				
			$P_{BP} \geq 0.99$	$P_{BP} \geq 0.95$	$P_{BP} \geq 0.9$	$P_{BP} \geq 0.7$	$P_{BP} \geq 0.5$
SSU rRNA[a,c]	61.4 ± 23.1	54.5 ± 24.5	86.0 ± 23.3	78.1 ± 25.8	74.8 ± 25.8	68.1 ± 26.5	63.2 ± 24.9
	(44.2 ± 14.7)	(37.1 ± 14.4)	(78.3 ± 22.2)	(71.0 ± 19.3)	(67.6 ± 17.5)	(57.6 ± 15.7)	(52.0 ± 15.1)
LSU rRNA[a,c]	74.0 ± 12.3	65.8 ± 12.3	91.8 ± 11.4	87.6 ± 10.5	85.3 ± 9.6	79.5 ± 10.4	75.2 ± 12.3
	(55.2 ± 11.5)	(47.2 ± 11.7)	(78.0 ± 27.4)	(74.1 ± 23.0)	(72.8 ± 19.3)	(67.6 ± 14.5)	(64.0 ± 15.3)
5S rRNA[d]	73.8 ± 26.7	64.6 ± 24.0	94.1 ± 14.4	86.1 ± 20.9	82.2 ± 21.8	72.6 ± 23.1	68.8 ± 23.1
Group I intron[c]	68.9 ± 14.5	61.4 ± 14.2	94.2 ± 12.1	90.4 ± 15.6	85.1 ± 16.7	76.0 ± 17.5	71.4 ± 16.4
Group I intron–2[b,c]	(57.4 ± 13.2)	(54.2 ± 14.5)	(92.4 ± 11.8)	(89.1 ± 11.0)	(81.2 ± 15.9)	(70.8 ± 16.5)	(67.3 ± 16.0)
Group II intron	87.6 ± 2.3	82.7 ± 6.7	89.9 ± 17.5	92.2 ± 13.6	90.8 ± 10.6	90.0 ± 7.5	87.4 ± 6.9
RNase P[e]	63.3 ± 14.4	60.8 ± 13.2	96.0 ± 9.9	95.1 ± 7.9	86.7 ± 15.7	75.4 ± 13.8	72.1 ± 14.2
RNase P–2[b,e]	(58.9 ± 7.6)	(56.6 ± 8.4)	(92.9 ± 12.4)	(92.0 ± 8.6)	(88.6 ± 10.1)	(78.7 ± 10.5)	(72.9 ± 11.8)
SRP[f]	66.4 ± 26.1	50.9 ± 22.3	79.1 ± 20.3	70.2 ± 25.4	67.8 ± 25.0	60.4 ± 24.0	57.0 ± 24.0
tRNA[g]	87.0 ± 17.0	85.5 ± 20.0	96.6 ± 13.3	94.1 ± 14.2	93.1 ± 15.3	90.9 ± 16.0	88.8 ± 17.3
Average	**72.8 ± 9.4**	**65.8 ± 12.4**	**91.0 ± 5.9**	**86.7 ± 8.6**	**83.2 ± 8.3**	**76.6 ± 10.3**	**73.0 ± 10.9**

MFE sensitivity is the percentage of known base pairs that are correctly predicted in the MFE structure:

$$\text{Sensitivity} = \frac{\text{number of known base pairs in predicted structure}}{\text{total number of known base pairs}}$$

where the known base pairs are determined by comparative sequence analysis. MFE positive predictive value is the percentage of base pairs in the predicted MFE structure that are contained in the structure determined by comparative sequence analysis:

$$\text{Positive predictive value} = \frac{\text{number of predicted base pairs in known structure}}{\text{total number of predicted base pairs}}$$

***Mathews, Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization*,**
RNA 10:1178-1190 (2004)

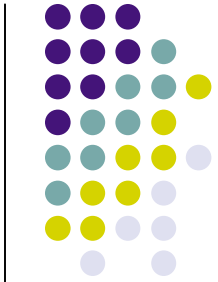
TABLE 4. Sensitivity and positive predictive value for structures constructed of highly probable base pairs

	$P_{BP} \geq 0.99$		$P_{BP} \geq 0.9$		$P_{BP} \geq 0.7$		$P_{BP} \geq 0.5$	
Type of RNA[a]	Sensitivity	PPV	Sensitivity	PPV	Sensitivity	PPV	Sensitivity	PPV
SSU rRNA	22.7 ± 17.1	86.0 ± 23.3	40.7 ± 21.5	74.5 ± 26.4	52.3 ± 22.5	67.0 ± 26.8	60.6 ± 23.5	61.7 ± 25.0
	(13.6 ± 10.3)	(78.2 ± 22.4)	(27.8 ± 13.8)	(67.0 ± 17.6)	(37.3 ± 14.6)	(55.2 ± 16.5)	(44.3 ± 15.8)	(47.5 ± 16.3)
LSU rRNA	25.2 ± 13.0	92.4 ± 11.4	46.4 ± 15.5	85.6 ± 9.5	62.9 ± 13.0	78.8 ± 10.6	71.4 ± 14.1	72.7 ± 13.9
	(17.9 ± 9.9)	(78.0 ± 27.4)	(33.9 ± 11.0)	(72.2 ± 18.4)	(46.1 ± 11.9)	(65.4 ± 13.1)	(53.7 ± 12.3)	(57.5 ± 13.9)
5S rRNA	28.5 ± 17.6	94.1 ± 14.4	47.2 ± 22.9	82.0 ± 22.1	59.7 ± 25.7	71.5 ± 23.3	68.1 ± 26.1	66.2 ± 23.4
Group I intron	22.3 ± 13.1	93.9 ± 12.0	47.9 ± 15.4	85.0 ± 16.7	60.1 ± 15.7	75.5 ± 17.5	67.7 ± 15.4	70.4 ± 16.1
Group I intron–2	(13.7 ± 9.3)	(92.4 ± 11.8)	(33.6 ± 15.4)	(81.2 ± 15.9)	(47.0 ± 12.2)	(69.5 ± 14.5)	(56.4 ± 13.4)	(65.9 ± 13.5)
Group II intron	13.9 ± 3.7	89.9 ± 17.5	54.5 ± 10.9	90.8 ± 10.6	76.3 ± 5.0	89.7 ± 8.2	83.8 ± 3.4	85.5 ± 10.4
RNase P	22.5 ± 7.1	96.0 ± 9.8	37.8 ± 9.2	86.5 ± 16.1	55.3 ± 13.1	72.1 ± 18.6	61.6 ± 16.8	66.5 ± 17.4
RNase P–2	(21.5 ± 10.2)	(92.9 ± 12.4)	(40.4 ± 12.5)	(88.6 ± 10.1)	(50.1 ± 10.4)	(77.0 ± 12.7)	(58.9 ± 8.8)	(68.4 ± 13.8)
SRP	25.3 ± 17.7	78.7 ± 20.6	42.4 ± 24.0	66.5 ± 25.8	53.1 ± 26.6	58.0 ± 24.4	60.6 ± 27.6	52.0 ± 23.1
tRNA	34.4 ± 20.8	96.5 ± 13.2	59.5 ± 23.7	93.0 ± 15.4	76.8 ± 20.0	90.6 ± 16.1	85.8 ± 17.0	87.6 ± 17.8
Average	**24.4 ± 5.8**	**90.9 ± 6.0**	**47.1 ± 7.2**	**83.0 ± 8.7**	**62.1 ± 9.6**	**75.4 ± 11.0**	**70.0 ± 10.0**	**70.3 ± 11.8**

[a]Refer to Table 2 for information about the databases of RNA sequences used in this study.

Mathews, Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization, **RNA** 10:1178-1190 (2004)

Sampling RNA structures



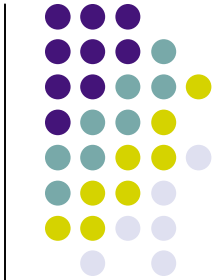
- Y. Ding, C.E. Lawrence, *Computers & Chemistry* 23 (1999) 387-400 use Bayesian method to sample number of loops, stacking energy tables, and secondary structures using the full Turner energy model.

- Simplified IDEA for Nussinov-Jacobson model uses recursive sampling procedure, which

  given sequence

  $$i,i+1,\ldots,j$$

  decides if j is unpaired, paired with i or paired with intermediate $i<k<j$ by using Boltzmann probability.



- Pr[j is unpaired]

$$Z_{i,j-1}/Z_{i,j}$$

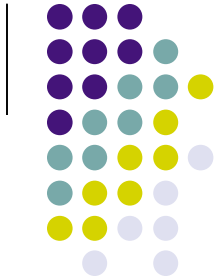
- Pr[j paired with i]

$$e^{-a(i,j)/RT} Z_{i+1,j-1}/Z_{i,j}$$

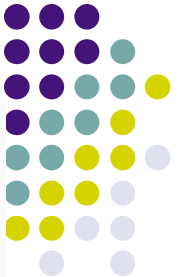
- Pr[j paired with k]

$$e^{-a(k,j)/RT} Z_{i,k-1}Z_{k+1,j-1}/Z_{i,j}$$

Partition function computation (Nussinov-Jacobson energy)



```
for i=0 to n − 1
  for j=i + 1 to n − 1
    Z(i,j) = 1;  //partition function for empty structure
for d = μ + 1 to n − 1 {
  for i = 0 to n − 1 {
     j = i + d;
     if (j < n){
      sum = Z(i, j − 1); //Case 1:  j unpaired
      if (j − i ≥ μ){
        sum+ = e^{−a(i,j,S)/RT} · Z(i + 1, j − 1); //Case 2:  i,j paired
        for k = i+1 to j-μ
           sum+ = e^{−a(k,j,S)/RT} · Z(i, k − 1) · Z(k + 1, j − 1);
           //Case 3:  k,j paired sone intermediate k
        Z(i,j) = sun;
      }
    }
  }
}
```



sample(i,j,paren)

if ($j - i > \mu$){

z = random(0,1);

x = $\frac{Z_{i,j-1}}{Z_{i,j}}$;

// probability that j unpaired

$y_i = e^{-a(i,j,S)/RT} \cdot \frac{Z_{i+1,j-1}}{Z_{i,j}}$;

// probability that i,j paired

for k = $i + 1$ to $j - \mu - 1$

$y_k = e^{-a(i,k,S)/RT} \cdot \frac{Z_{i+1,k-1} \cdot Z_{k+1,j-1}}{Z_{i,j}}$;

// probability that k,j paired for intermediate k

if (z<x)

sample(i,j-1,paren);

else {

sum = x; k = i-1;

while (z<sum){ k += 1; sum += y_k; } //determine k using roulette wheel

paren[k] = '('; paren[j] = ')';

sample(i+1,k-1,paren); //recursive call on left substring

sample(k+1,j-1,paren); //recursive call on right substring

//could additionally ensure at most log space using quicksort trick

}

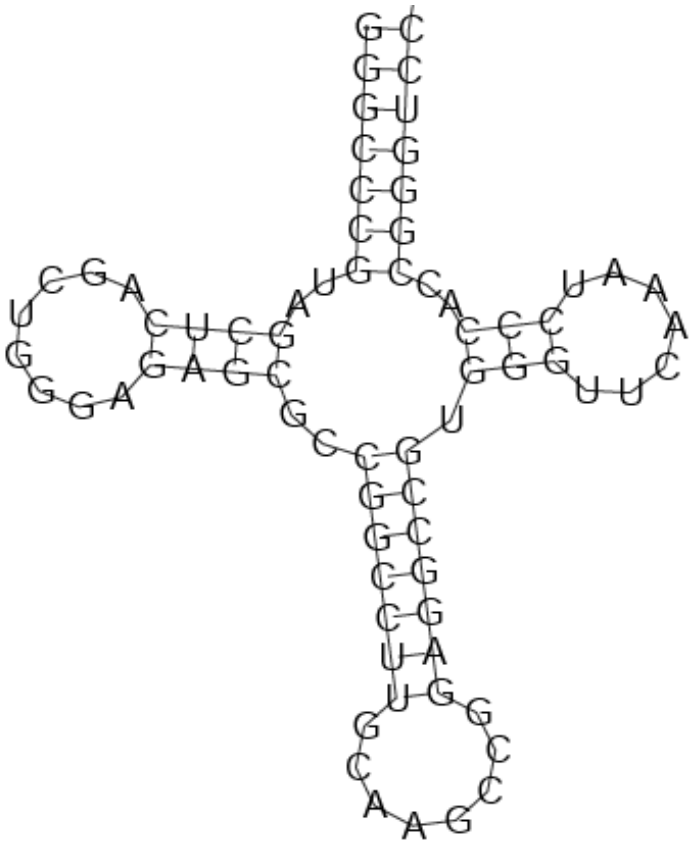
Sfold – Ding's software for sampling RNAs

http://sfold.wadsworth.org/index.pl

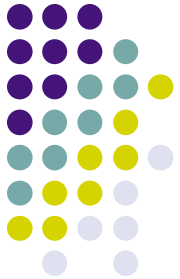
Applications of sampling (Ding-Lawrence)

- Determination of probability that i,i+1,i+2,i+3 is NOT base paired in the statistical ensemble of 1000 samples. Such regions, if they additionally satisfy Tuschl's rules are potential targets for RNAi
- Determination of probability that i is in a hairpin loop, etc.
- Improve RNA secondary structure prediction by finding centroid of cluster of sampled structures. Ding, Chan, Lawrence, *RNA secondary structure by centroids in a Boltzmann weighted ensemble*, RNA **(2005).**

Vienna RNA Package predicted structure for Ala-tRNA from *M. jannaschii*



Sample output from Sfold



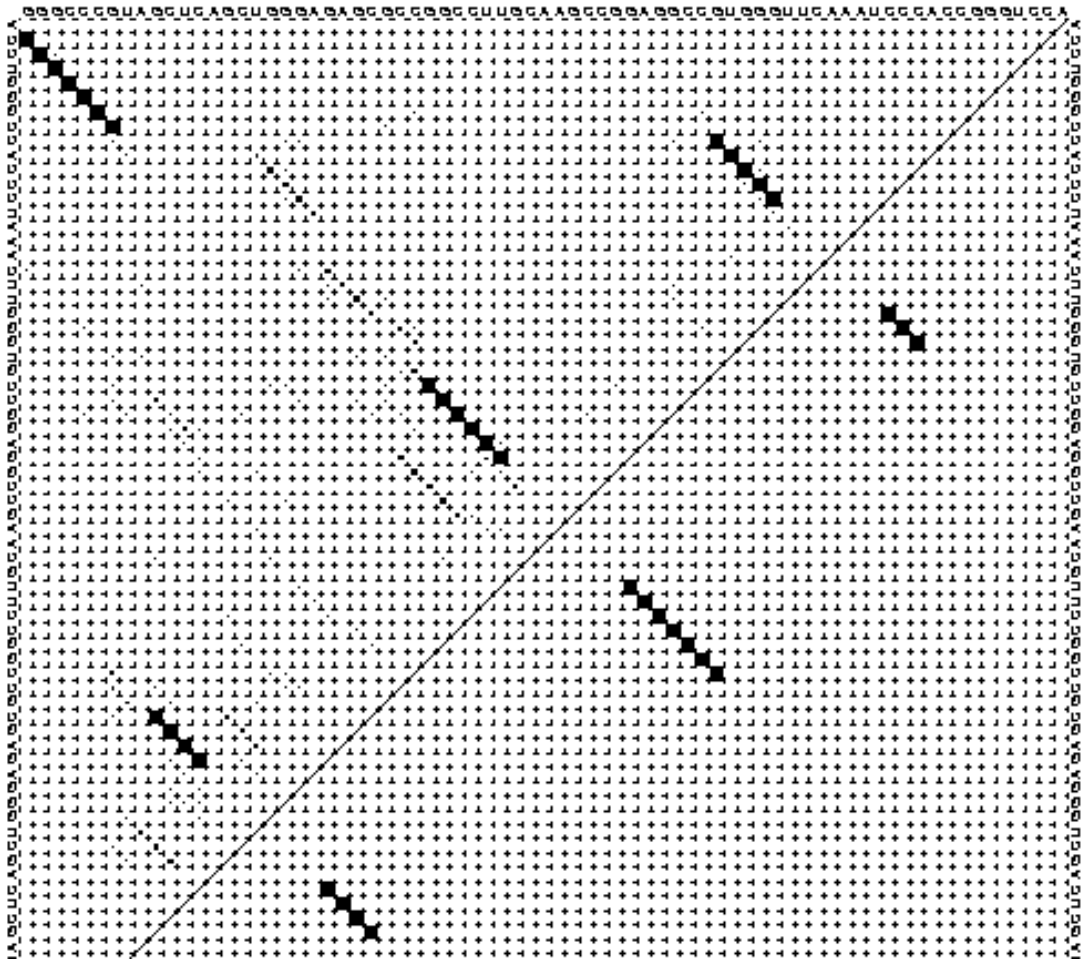
A representation of the structure sample

Two-dimensional histogram of base pairs in the sample

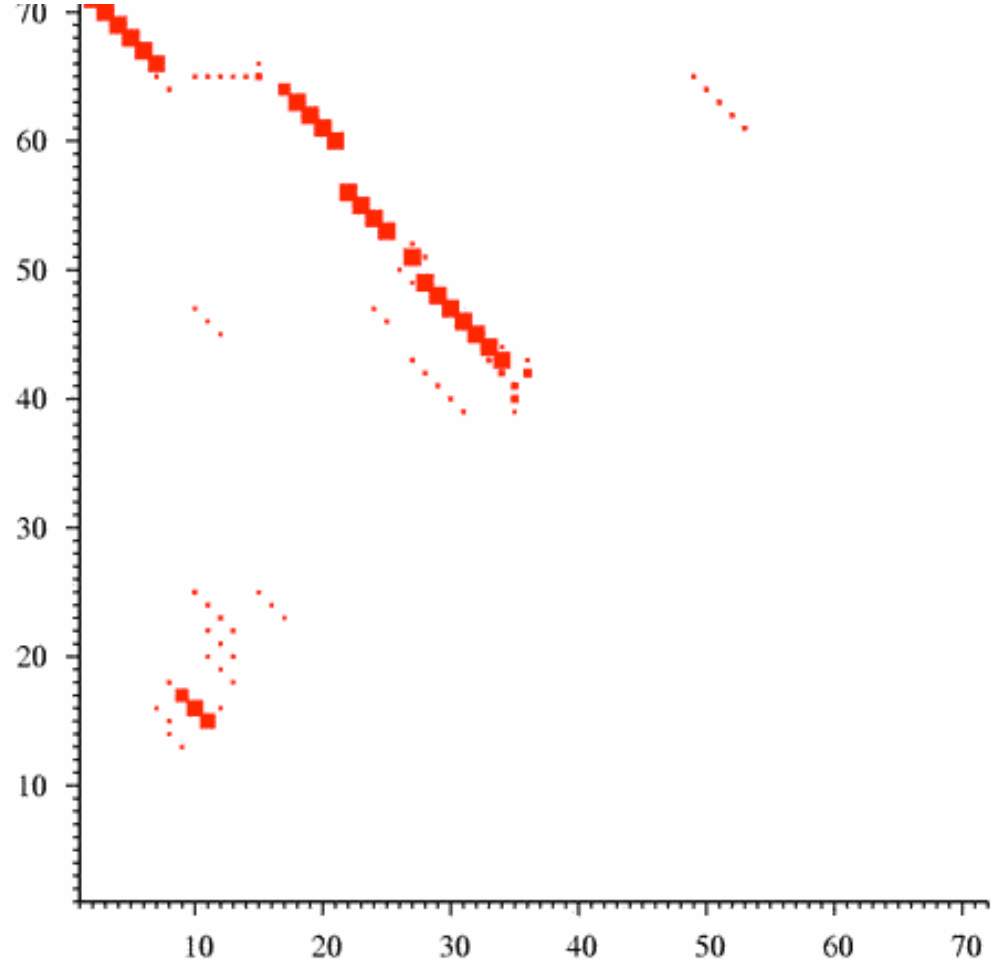
with base pair probabilities View

Structure #	Free energy interval	Sample frequency	Lowest free energy	Structure diagram
1	[-37.80, -36.94)	0.291	-37.80	PNG image View
2	[-36.94, -36.08)	0.342	-36.90	PNG image View
3	[-36.08, -35.22)	0.163	-35.80	PNG image View
4	[-35.22, -34.36)	0.109	-35.20	PNG image View
5	[-34.36, -33.50)	0.034	-34.30	PNG image View
6	[-33.50, -32.64)	0.028	-33.50	PNG image View
7	[-32.64, -31.78)	0.025	-32.60	PNG image View
8	[-31.78, -30.92)	0.004	-31.70	PNG image View
9	[-30.92, -30.06)	0.003	-30.50	PNG image View
10	[-30.06, -29.20]	0.001	-29.20	PNG image View

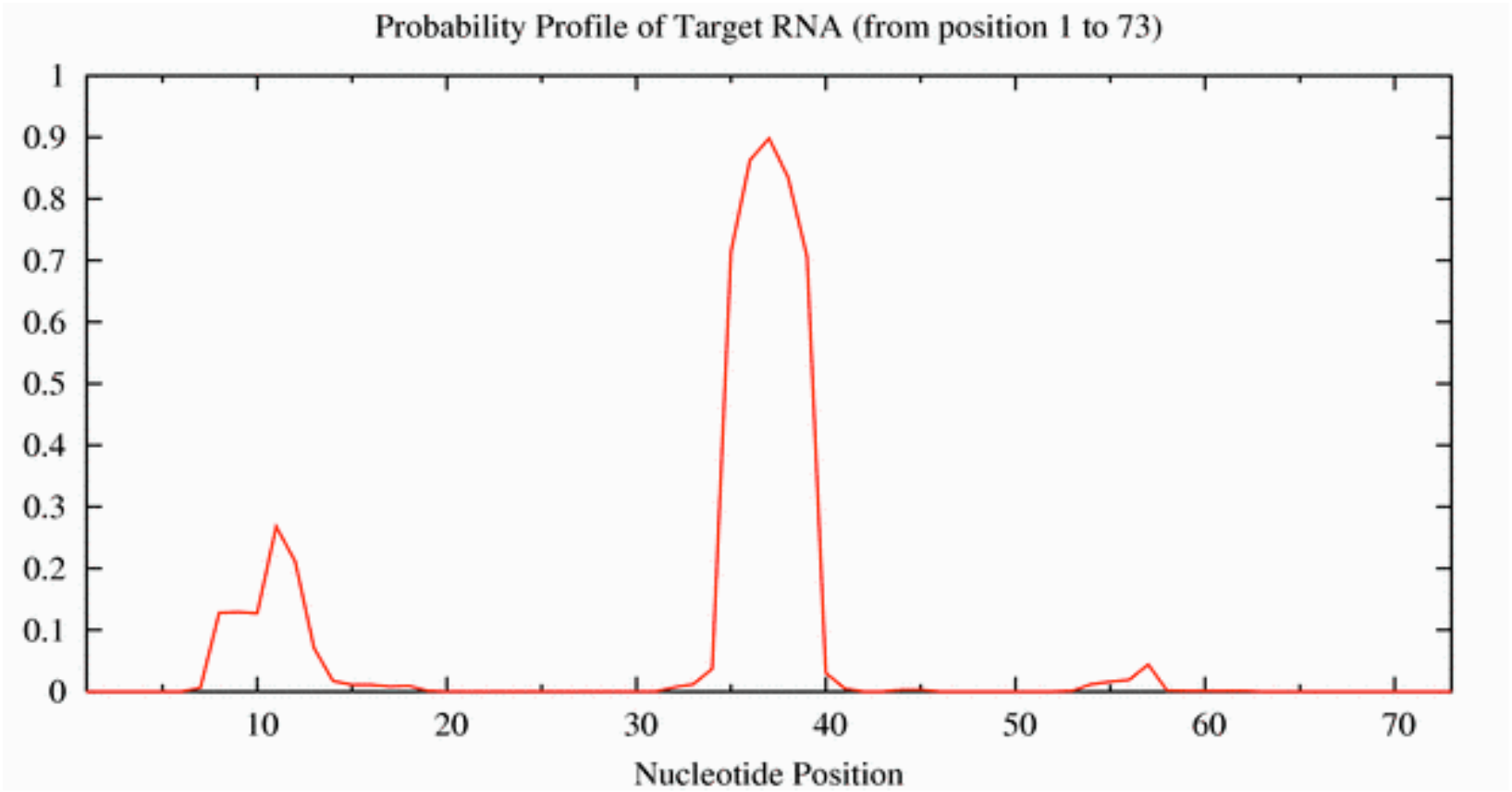
McCaskill's algorithm base pairing probabilities for Ala tRNA using Vienna RNA package



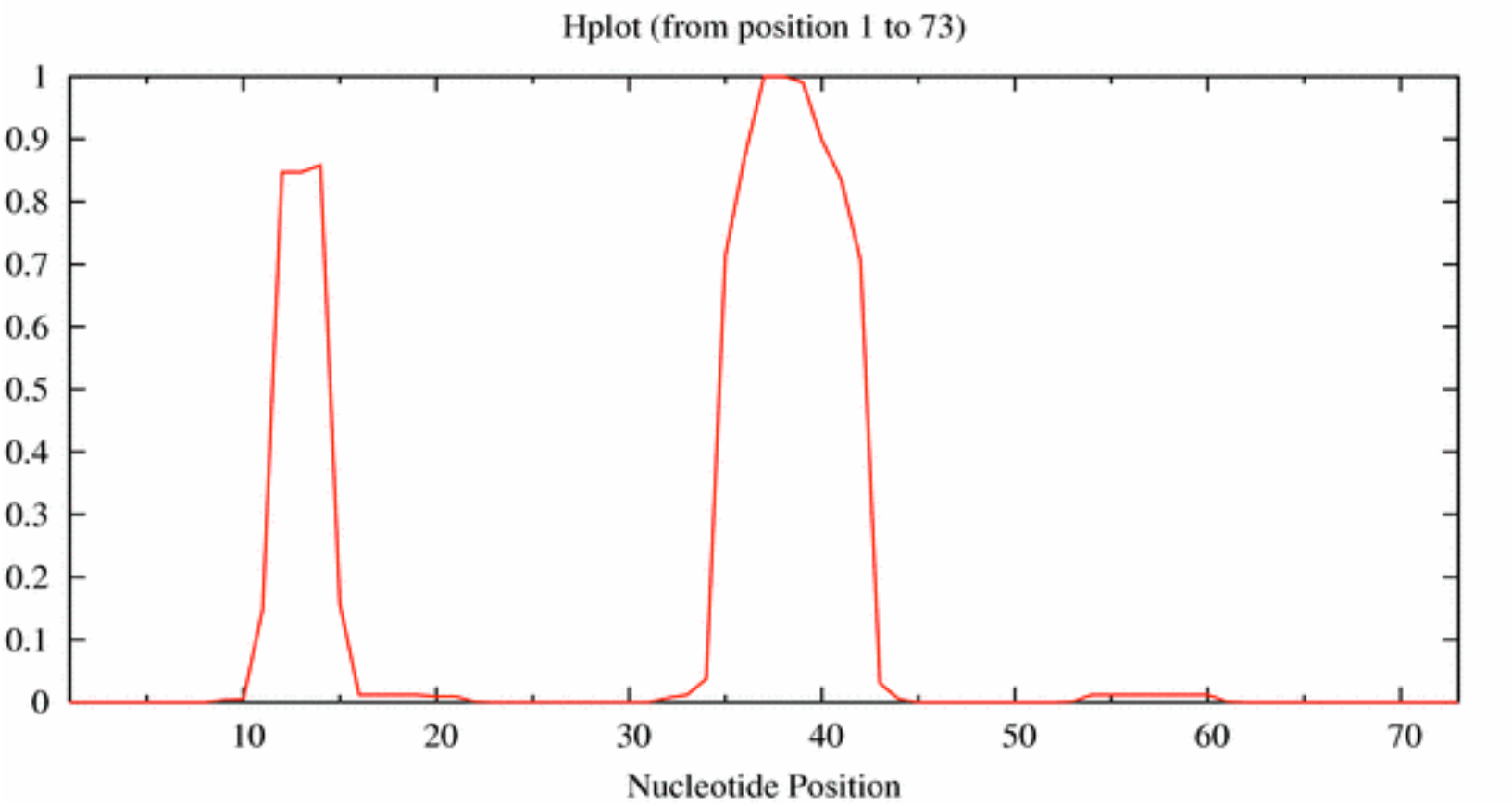
Sampled base pair histogram from 1000 sampled Ala-tRNA structures using Sfold



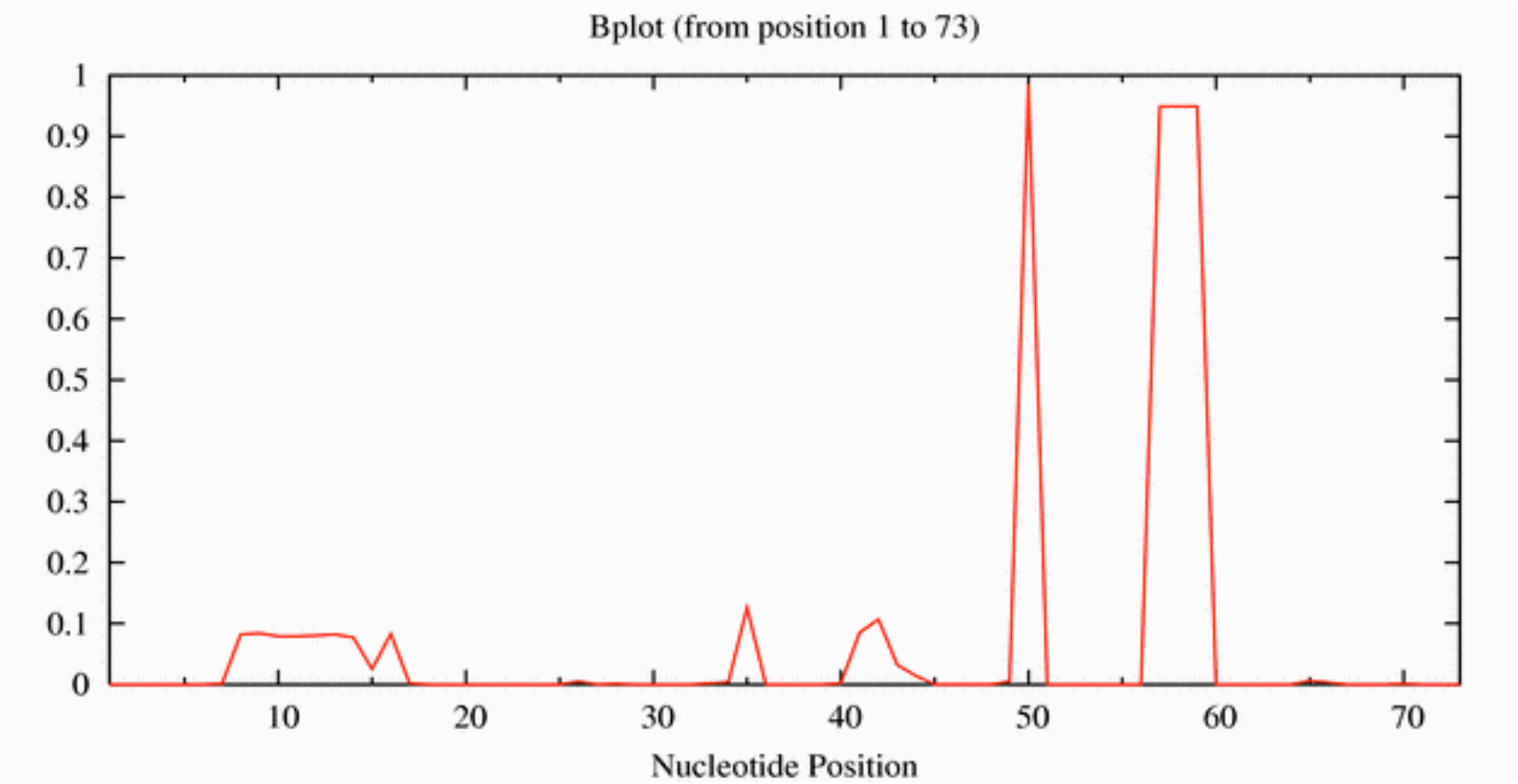
Probability profile of Ala-tRNA using Sirna from Sfold



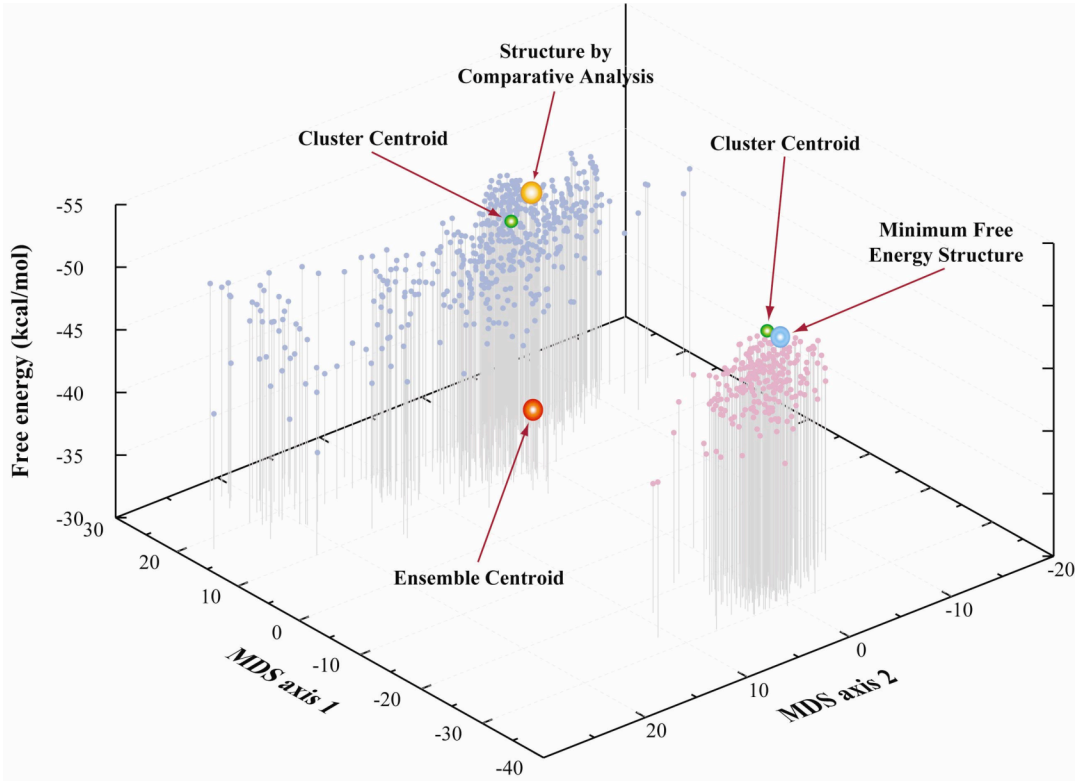
Hairpin frequency in Ala-tRNA samples using Sirna from Sfold



Bulge frequency in Ala-tRNA samples using Sirna from Sfold



Boltzmann centroids



Ding, Chan, Lawrence, RNA secondary structure by centroids in a Boltzmann weighted ensemble, to appear in **RNA** (2005).

UPGMA

Algorithm

INPUT: $n \times n$ distance matrix D

1. initialize set $\mathcal{C}$ to consist of n initial singleton clusters $\{1\}, \ldots, \{n\}$
2. initialize function $dist(c, d)$ on $\mathcal{C}$ by defining for all $\{i\}$ and $\{j\}$ in $\mathcal{C}$

$$dist(\{i\}, \{j\}) = D(i, j)$$

3. repeat $n-1$ times
   (a) determine pair c, d of clusters in D such that $dist(c,d)$ is minimal; define

$$d_{min} = dist(c,d)$$

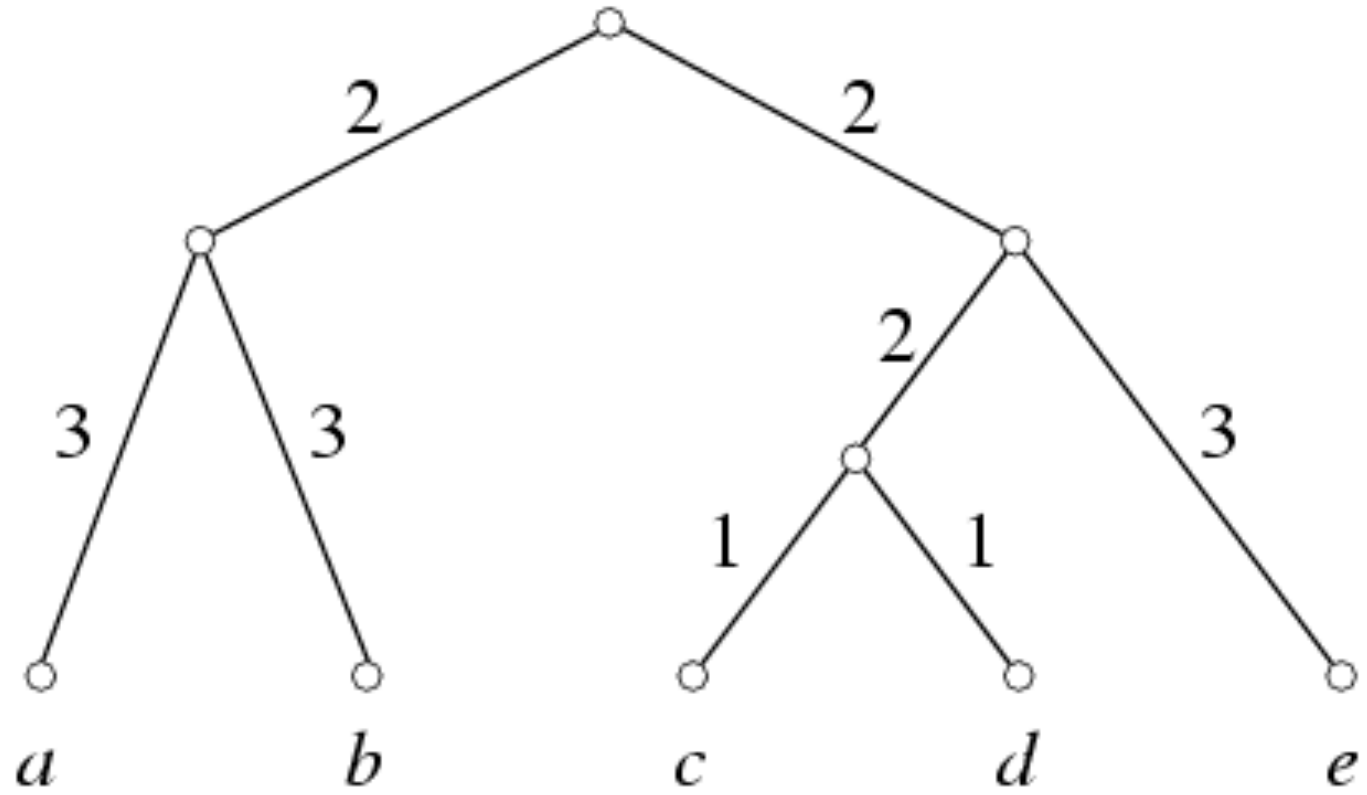
   (b) define new cluster $e = c \cup d$; define

$$\mathcal{C} = \mathcal{C} - \{c,d\} \cup \{e\}$$

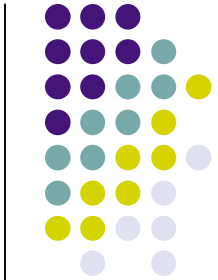
   (c) define a node with label e and daughters c, d, where the e has distance $\frac{d_{min}}{2}$ to its leaves (i.e. the leaves of the subtree rooted at e).
   (d) define for all $f \in \mathcal{C}$ with $f \neq e$

$$dist(e,f) = dist(f,e) = \frac{dist(c,f) + dist(d,f)}{2}$$

Ultrametric tree



FACT. If tree is ultrametric, then UPGMA correctly computes the topology.



Diana hierarchical clustering (R)

Determine optimal number k of clusters by maximizing CH index.

$$CH(k) = \frac{B(k)/(k-1)}{W(k)/(n-k)}$$

Here n is total number of points to cluster, B(k) is the sum of squares between clusters, and W(k) is the sum of squares within clusters.

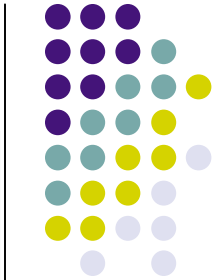


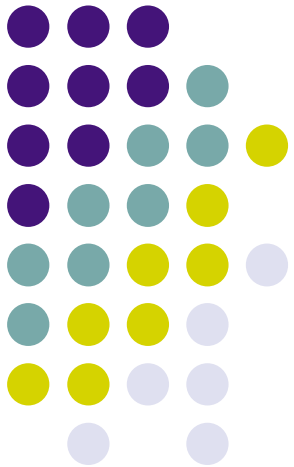
Table 1. Clusters for sampled structures and MFE structure

RNA type	Organism	GenBank Accession no.	Length (nt)	Number of clusters	Cluster probabilities (*cluster of MFE structure)
SSU (16S) rRNA	*Bordetella bronchiseptica*	U04948	1532	2	0.930* 0.070
tRNA	*Crossostoma lacustre*	M91245	70	2	0.906* 0.094
Group I intron	*Acanthamoeba griffini*	U02540	556	2	0.950* 0.050
LSU (23S) rRNA	*Thermus thermophilus*	X12612	2915	3	0.339* 0.335 0.326
Group I intron	*Acanthamoeba griffini*	S81337	526	5	0.464* 0.400 0.052 0.048 0.036
Group II intron	*Saccharomyces cerevisiae*	AJ011856	2520	4	0.557* 0.432 0.007 0.004
5S rRNA	*Agrobacterium tumefaciens*	X02627	120	2	0.591 0.409*
tmRNA	*Dehalococcoides ethenogenes* strain 195	GSP [a]	352	2	0.578 0.422*
RNase P	*Tarsius syrichta*	L08801	286	3	0.552 0.446* 0.002
LSU (23S) rRNA	*Chlamydomonas reinhardtii*	X15727	2902	3	0.907 0.093 0.000*
RNase P	*Dermocarpa* sp.	X97396	359	2	0.803 0.197*
RNase P	*Leptospirillum ferrooxidans*	AF296042	327	2	0.804 0.196*

[a] http://tigrblast.tigr.org/ufmg/index.cgi?database=d_ethenogenes|seq

Ding, Chan, Lawrence, RNA secondary structure by centroids in a Boltzmann weighted ensemble, to appear in **RNA** (2005).

Pseudoknots

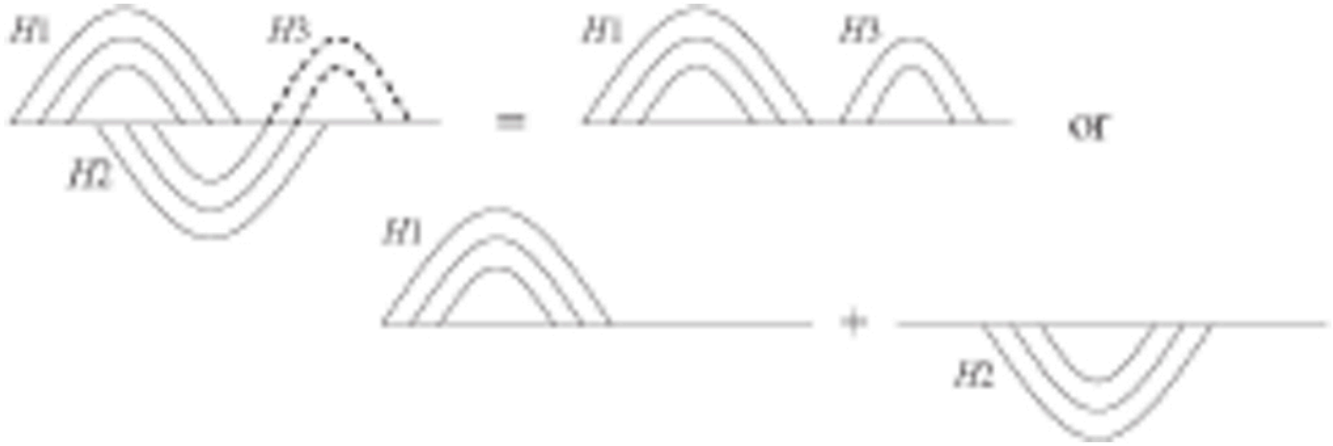


Iterated Loop Model (ILM)

First approach: Given RNA sequence $a_1,...,a_n$
$I = \{1,...,n\}$; $S = \emptyset$
while not finished
S += Nussinov-Jacobson(I)
$I = I - \{i,j : (i,j) \in S\}$

Ruan, Stormo, Zhang, Bioinformatics. 2004 Jan 1;20(1):58-66.

Iterated Loop Model (ILM)



Second approach: Given RNA sequence $a_1,...,a_n$

I = {1,...,n}; S = \emptyset

while not finished

L = Helices (stems) from Nussinov-Jacobson(I)

H = helix with best energy in L

S += H

I = I - {i,j : (i,j) ∈ S}

Ruan, Stormo, Zhang, Bioinformatics. 2004 Jan 1;20(1):58-66.

MCMC sampling (Metzler,Nebel)

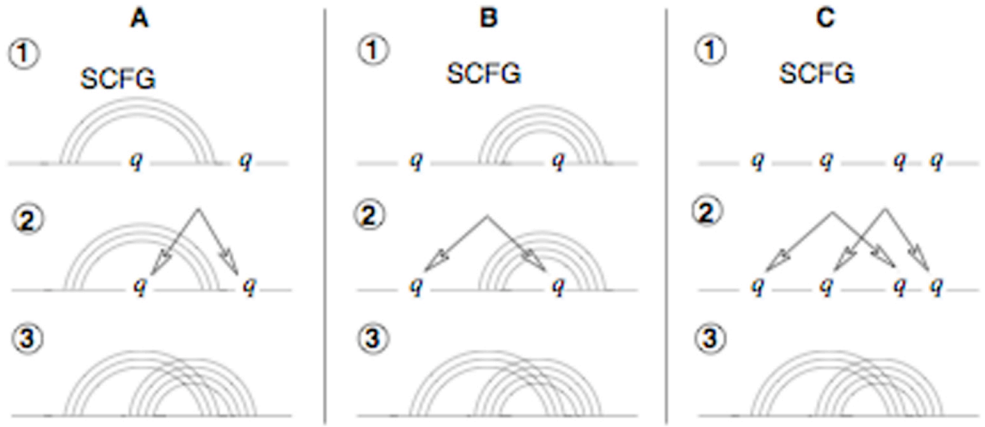
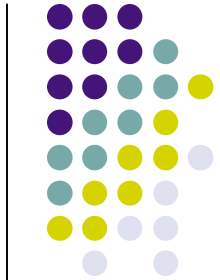


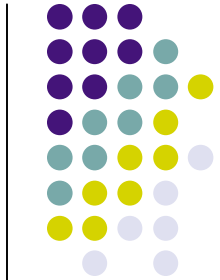
Figure 1: How our grammar works: (1) A sequence that may contain base pairs (represented by arcs connecting the corresponding positions) and q symbols is generated by the SCFG., (2) the q symbols are randomly paired, (3) paired q symbols generate stems and are then replaced by the complementary sequences. Three different ways (A, B, C) lead to the same pseudoknot.

"Predicting RNA secondary structures with pseudoknots by MCMC sampling, Metzler, Nebel, J Math Biol (in press) 2007.



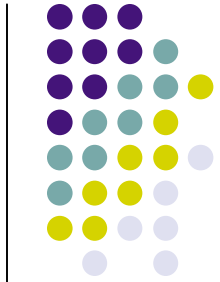
S	$\rightarrow$	LS	0.85
S	$\rightarrow$	L	0.15
F	$\rightarrow$	$xLSy$	0.25
F	$\rightarrow$	xFy	0.75
L	$\rightarrow$	x	0.9
L	$\rightarrow$	$uxFyv$	0.06
L	$\rightarrow$	q	0.04
R	$\rightarrow$	xRy	0.75
R	$\rightarrow$	xy	0.25

Probabilities associated with grammar rules for the MCMC algorithm of Metzler and Nebel to compute RNA secondary structure with p seudoknots.



π_{AU}	0.0175	π_{UA}	0.0175
π_{GC}	0.275	π_{CG}	0.275
π_{GU}	0.05	π_{UG}	0.05
π_{A}	0.35	π_{C}	0.2
π_{G}	0.2	π_{U}	0.25

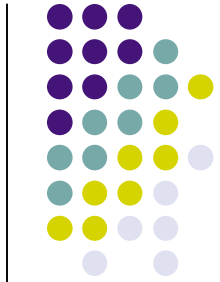
Compositional and base-pairing probabilities for the MCMC sampling pseudoknot algorithm of Metzler and Nebel.



For set μ of base pairs,

$$\begin{aligned} Pr[M = \mu] &= \sum_{\iota,\rho} Pr[M = \mu, I = \iota, P = \rho | S = s] \\ &= \sum_{\iota,\rho} Pr[M = \mu | I = \iota, P = \rho, S = s] \cdot Pr[I = \iota, P = \rho | S = s]. \end{aligned}$$

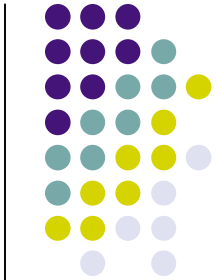
Sample (ι, ρ) from posterior probability $Pr[I = \iota, P = \rho | S = s]$ using MCMC and inside-outside algorithm in $O(n^3)$ time and $O(n^2)$ space. Then approximate $Pr[(i, j) \in M | S = s]$ by averaging $Pr[I = \iota, P = \rho | S = s]$ over all samples pairs (ι, ρ).



In order to sample (ι, ρ) from posterior probability $Pr[I = \iota, P = \rho | S = s]$, use Metropolis-Hastings MCMC to construct a sequence $(I_0, P_0), \ldots, (I_k, P_k), \ldots$, where the randomly generated next candidate (I_k, P_k) is accepted with probability

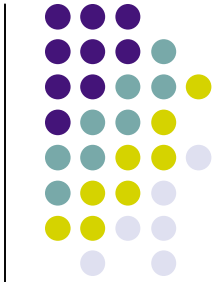
$$\min \left\{ 1, \frac{Q_{\iota', \rho'}}{Q_{\iota, \rho}} \cdot \frac{Pr[I = \iota', P = \rho' | S = s]}{Pr[I = \iota, P = \rho | S = s]} \right\}.$$

Remarks

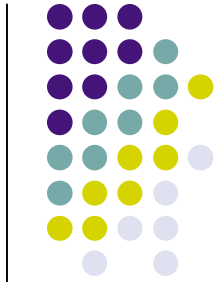


- MCMC sampling approach handles ALL pseudoknots without restriction
- Model requires a minimum of 3 consecutive base pairs to close any loop, including bulges and internal loops
- Algorithm depends on parameters, so improvement possible for particular classes of RNA

Pseudoknots



- Ruan J, Stormo GD, Zhang W. ILM: a web server for predicting RNA secondary structures with pseudoknots.Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W146-9.
- McQFold (Metzler,Nebel) http://www.cs.uni-frankfurt.de/~metzler/McQFold
- NUPACK (Pierce) http://www.nupack.org/
- pknotsRG (Reeder,Giegerich) http://bibiserv.techfak.uni-bielefeld.de/pknotsrg/
- Hot Knots (Condon) http://www.cs.ubc.ca/labs/beta/Software/HotKnots



THANKS!