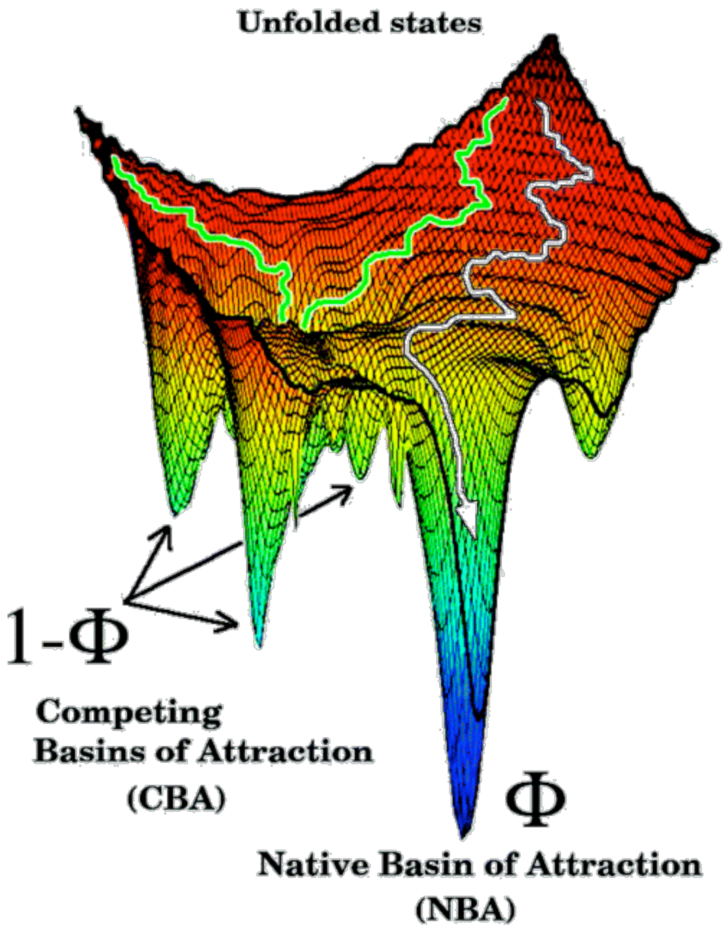
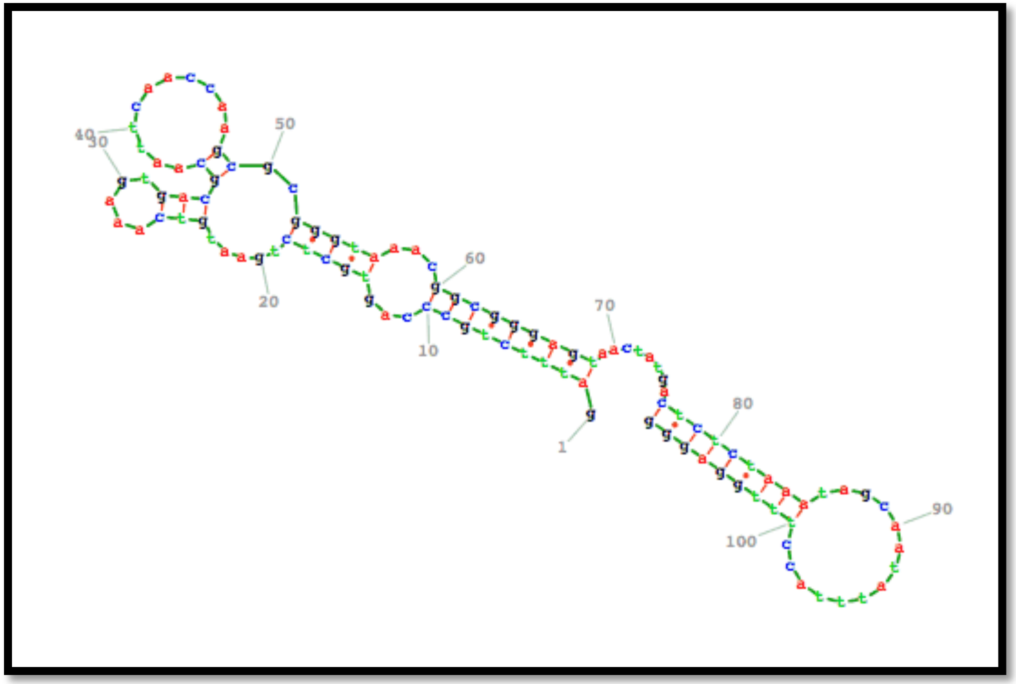
COMP761: Fundation of Computational Structural Biology

RNA STOCHASTIC SECONDARY STRUCTURE PREDICTION

JÉRÔME WALDISPÜHL
SCHOOL OF COMPUTER SCIENCE, MCGILL

Beyond the minimal free energy structure





Beyond the minimal free energy structure

Problems:
- There is potentially several competing structures
- The mfe may not be reachable.

Alternative 1: Look at the suboptimal structure

- P-optimal base pair : base pair $(i \cdot j)$ s.t. $V(i,j) + V(j,i) \geq (1 - \frac{P}{100}) \cdot E_{min}$
- the collection of P-optimal base pairs is the union of all P-optimal foldings
- Compute representative suboptimal foldings by determining the m.f.e. structure of a structure with suboptimal base pair.

Alternative 2: Statistical Mechanics

Definition 1 (Boltzmann partition function) :
Let us label the exact states (microstates) that the system can occupy by *j* (*j*=1,2,3,...), and denote the total energy of the system when it is in microstate *j* as E_j. Additionally, we denote T the temperature of the system and $\mathcal{K}_B$ the Boltzmann constant.

$$Z = \sum_j e^{-\beta \cdot E_j}, \text{ where } \beta = \frac{1}{\mathcal{K}_B T}$$

Definition 2 (Probability of a state) :

$$P_j = \frac{e^{-\beta E_j}}{Z}$$

Application

Theorem 1 *Average energy*

$$\langle E \rangle = \mathcal{K}_B T^2 \cdot \frac{\partial}{\partial T} \ln Z = -\frac{\partial \ln Z}{\partial \beta}$$

Proof :

$$\begin{aligned}
\frac{\partial}{\partial T} \ln Z &= \frac{\partial}{\partial T} \ln \sum_j e^{-E_j/\mathcal{K}_B T} \\
&= \frac{1}{Z(T)} \cdot \sum_j \frac{E_j}{\mathcal{K}_B T^2} \cdot e^{-E_j/\mathcal{K}_B T} \\
&= \frac{1}{\mathcal{K}_B T^2} \cdot \frac{\sum_j E_j \cdot e^{-E_j/\mathcal{K}_B T}}{Z(T)} \\
&= \frac{1}{\mathcal{K}_B T^2} \cdot \sum_j P_j \cdot E_j \\
&= \frac{\langle E_j \rangle}{\mathcal{K}_B T^2}
\end{aligned}$$

Application

Variance (energy fluctuation) :

$$\langle(\delta E)^2\rangle = \frac{\partial^2 \ln Z}{\partial \beta^2}$$

Heat capacity :

$$C_v = \frac{\partial \langle E\rangle}{\partial T} \ln Z = \mathcal{K}_B T^2 \cdot \langle \delta E^2\rangle$$

Entropy :

$$S = -\mathcal{K}_B \sum_j P_j lnP_j = \mathcal{K}_B(\ln Z + \beta\langle E\rangle) = \frac{\partial}{\partial T}(\mathcal{K}_B \ln Z)$$

RNA Secondary Structure Partition Function

Definition 3 *Boltzmann partition function*

- s *is the sequence,*
- $\mathcal{S}(s)$ *is the ensemble of structures over* s*,*
- $E(S)$ *folding energy,*
- R *gas constant,*
- T *temperature.*

$$Q(s) = \sum_{S \in \mathcal{S}(s)} e^{-E(S)/RT}, Q_k(s) = \sum_{S \in \mathcal{S}_k} e^{-E(S)/RT}$$

Definition 4 *density of states*

$$\rho_k(s) = \frac{Q_k(s)}{Q(s)}$$

Principle

Problem 1: How to recursively decompose the partition function?

Observation 1:
Let A,B be 2 secondary structures and E an energy function with the additive property (i.e. E(AB)=E(A)+E(B)), then

$$exp(-(E(A)+E(B))/RT)=exp(-E(A)/RT) \times exp(-E(B)/RT).$$

Conclusion 1:

- The partition function of a sum is the product of the partition function of each component.
- If the seq ω_{ij} has a single secondary structure that can be cut in two parts at index *k* on ω_{ik-1} and ω_{kj} then:

$$Z(i,j)=Z(i,k-1)\times Z(k,j)$$

Problem 2: Can we extend the result if multiple sec. str. can be mapped on subsequences?

Principle

Observation 2: Let $\omega_{i,j}$ be a RNA sequence and let i<k<j. Assume that there is only two sec. str. A,B possible on $\omega_{i,k-1}$ and only two sec. str. C,D on $\omega_{k,j}$. The partition function of this partition is:

$$
\begin{aligned}
Z &= e^{-E(A)/RT} \cdot e^{-E(C)/RT} + e^{-E(A)/RT} \cdot e^{-E(D)/RT} + \\
&\quad e^{-E(B)/RT} \cdot e^{-E(C)/RT} + e^{-E(B)/RT} \cdot e^{-E(D)/RT} \\
&= e^{-E(A)/RT} \cdot \left(e^{-E(C)/RT} + e^{-E(D)/RT}\right) + \\
&\quad e^{-E(B)/RT} \cdot \left(e^{-E(C)/RT} + e^{-E(D)/RT}\right) \\
&= \left(e^{-E(A)/RT} + e^{-E(B)/RT}\right) \cdot \left(e^{-E(C)/RT} + e^{-E(D)/RT}\right)
\end{aligned}
$$

$$\boxed{Z(i,j) = Z(i,k-1) \times Z(k,j)}$$

Comparison with the minimum folding energy algorithm

Principle of Zuker's algorithm:

$$MFE(i,j) = \min_k (MFE(i,k-1) + MFE(k,j))$$

Principle of McCaskill's algorithm:

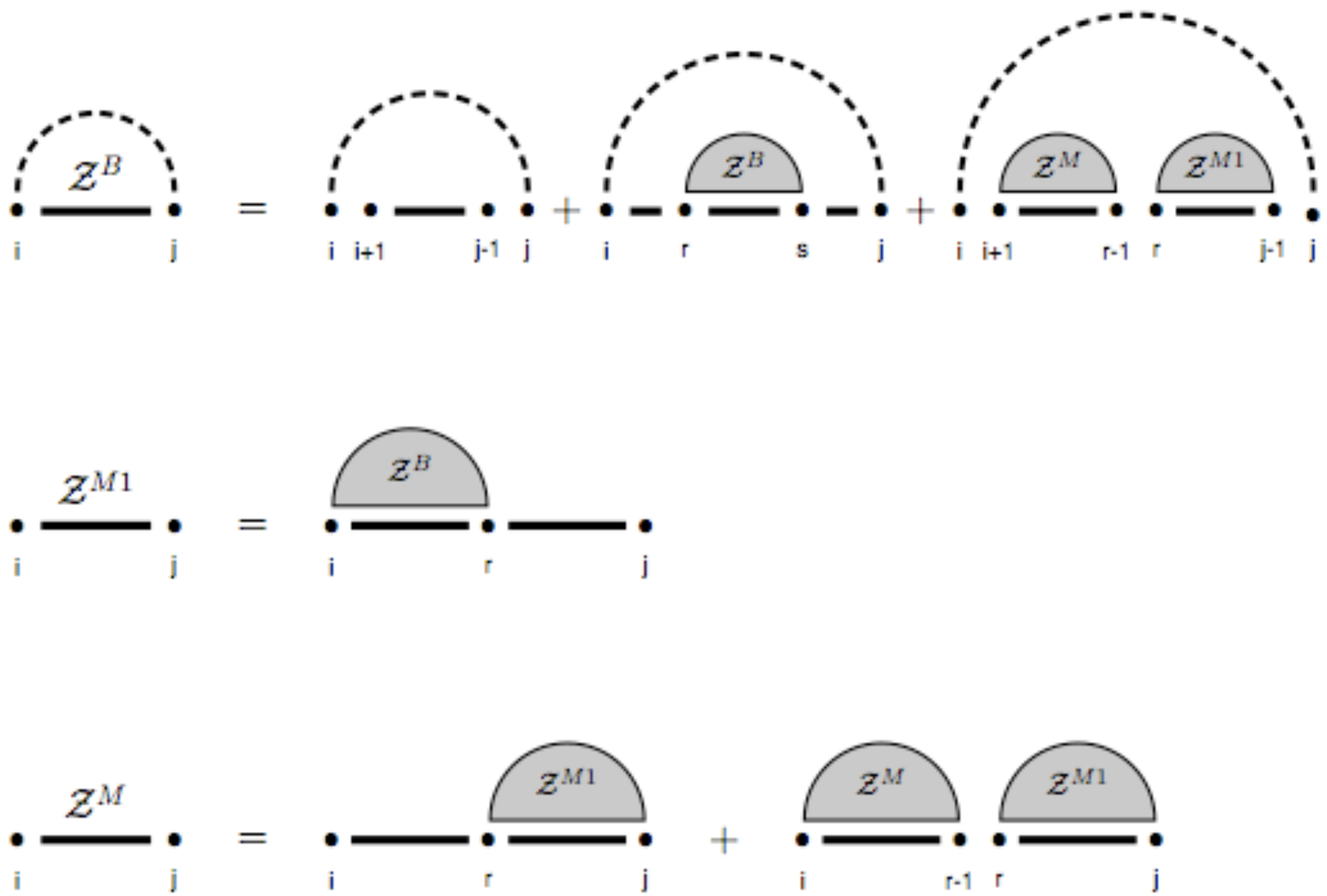
$$Z(i,j) = \Sigma_k \, Z(i,k-1) \times Z(k,j)$$

Conclusion: Conserve the algorithm structure and switch the ring from {min,+} to {+,×}.

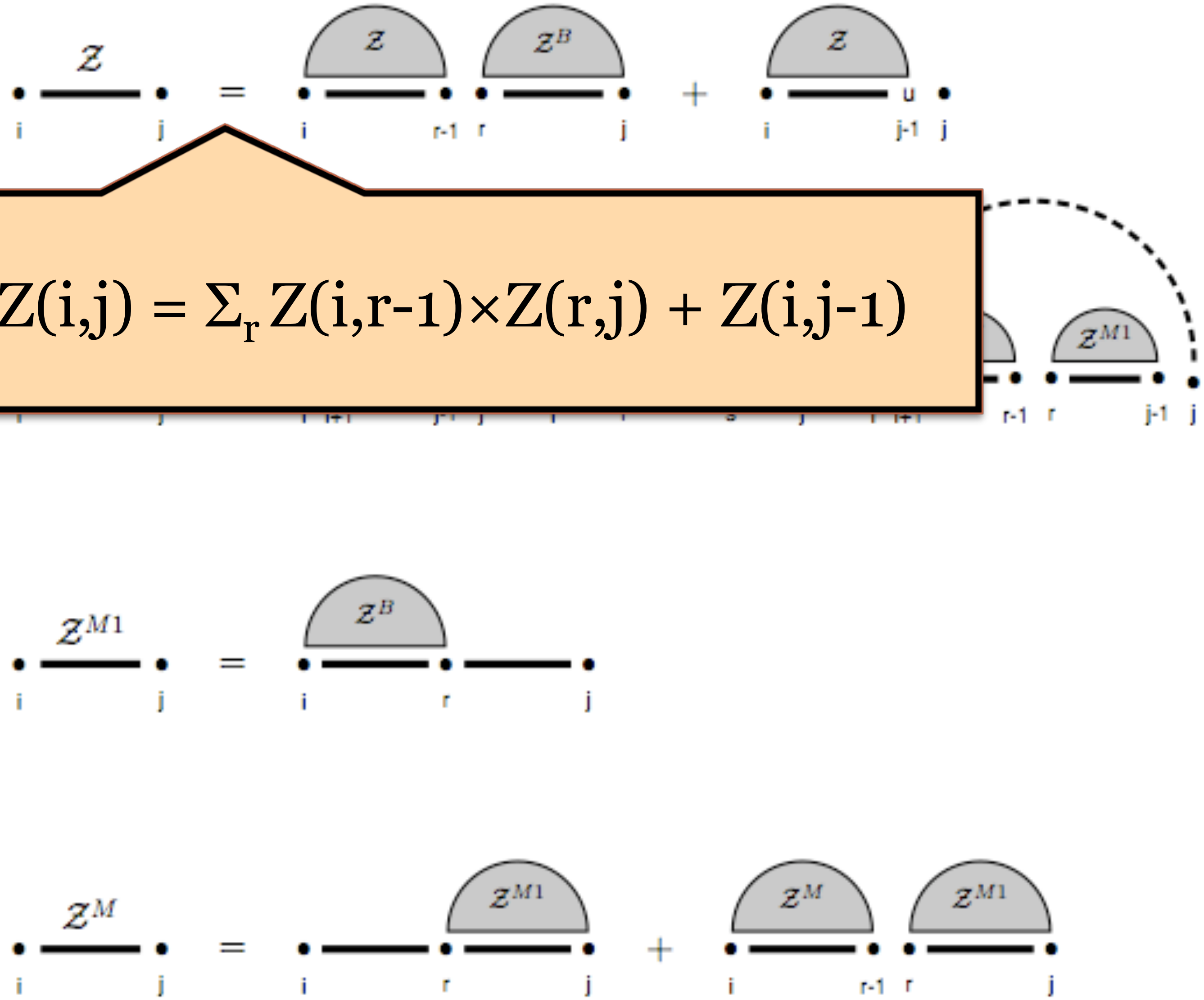
Algorithm (Dynamic tables)

- $Z(i,j)$: partition function over all secondary structures of $a[i,j]$.
- $Z^B(i,j)$: partition function over all secondary structures of $a[i,j]$, which contain the base pair (i,j).
- $Z^M(i,j)$: partition function over all secondary structures of $a[i,j]$, subject to the constraint that $a[i,j]$ is part of a multiloop and has *at least* one component.
- $Z^{M1}(i,j)$: partition function over all secondary structures of $a[i,j]$, subject to the constraint that $a[i,j]$ is part of a multiloop and has at *exactly* one component. Moreover, it is *required* that i base-pair in the interval $[i,j]$; i.e. (i,r) is a base pair, for some $i < r \leq j$.

Algorithm (Feyman Diagrams)

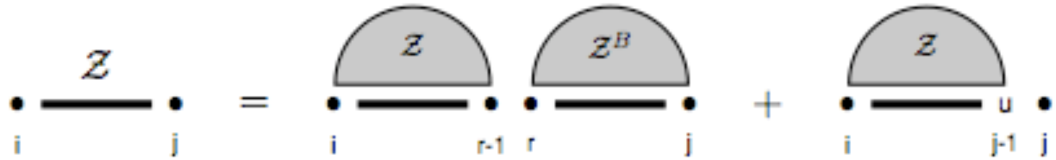


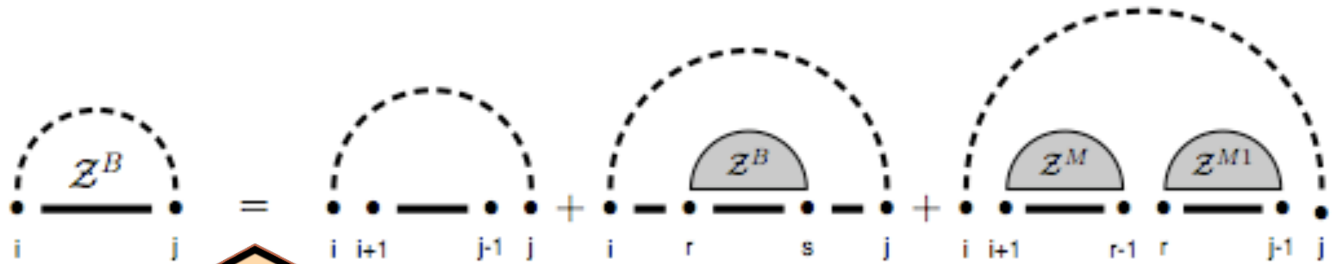
Algorithm (Feyman Diagrams)



$$Z(i,j) = \Sigma_r Z(i,r\text{-}1)\times Z(r,j) + Z(i,j\text{-}1)$$

Algorithm (Feyman Diagrams)

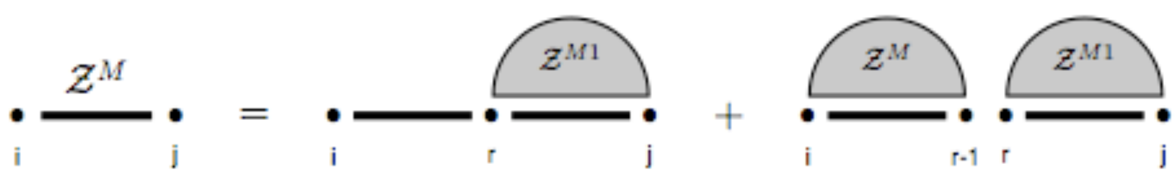




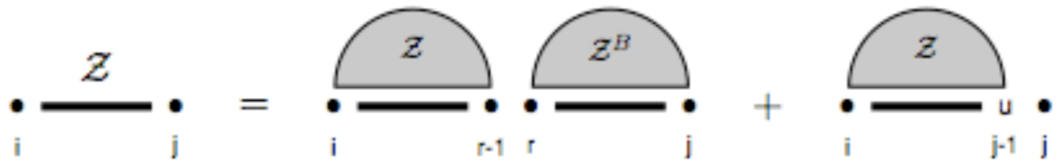
$Z^B(i,j) = \exp(-\text{hairpin}(i,j)) +$

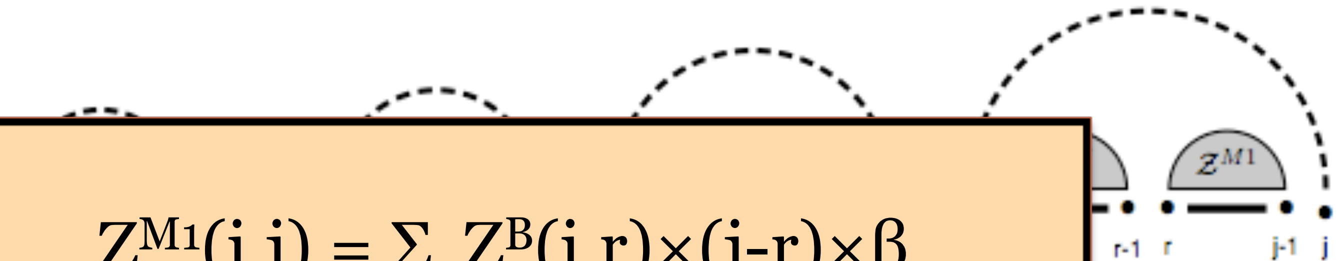
$\Sigma_{r,s}\, Z(i+1,r-1) \times \exp(-E(i,r,s,j)/RT) +$

$\Sigma_r\, Z(i+1,r-1) \times Z(r,j-1) \times \exp(-(\alpha+\Upsilon)/RT)$

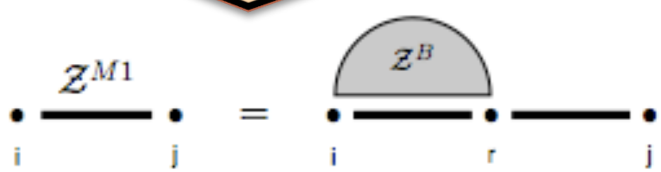


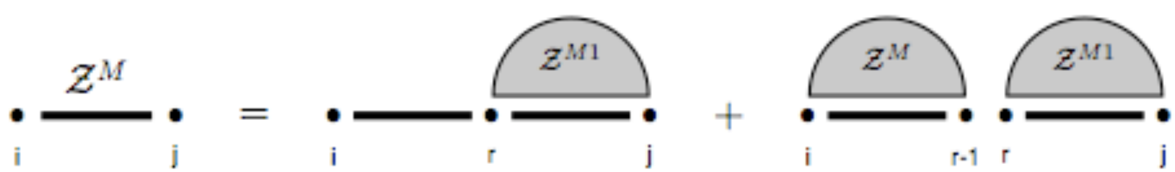
Algorithm (Feyman Diagrams)



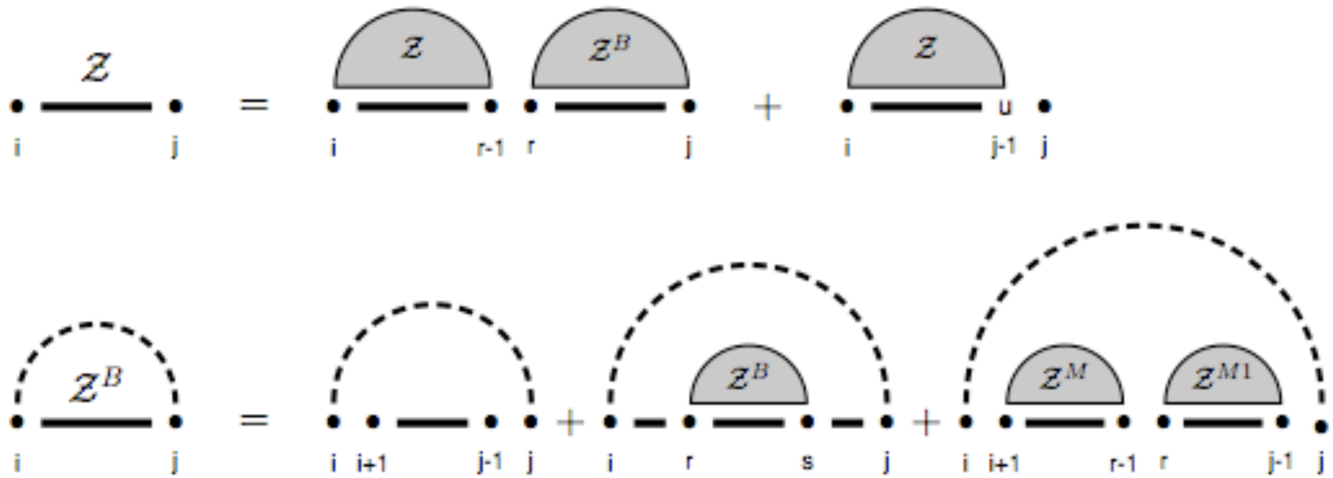


$$Z^{M1}(i,j) = \Sigma_r Z^B(i,r)\times(j-r)\times\beta$$

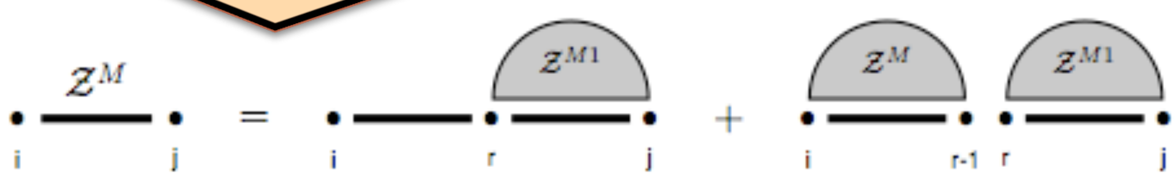




Algorithm (Feyman Diagrams)



$$Z^M(i,j) = \Sigma_r Z^{M1}(r,j)\times(r-i)\times\beta + \Sigma_r Z^{M1}(i,r-1)\times Z^{M1}(r,j)$$



Algorithm (recursive equations)

Unconstrained partition function :

$$Z(i,j) \quad = \quad Z(i,j-1) + \sum_{r=i}^{j-\theta-1} Z(i,r-1) \cdot Z^B(r,j).$$

Partition function such that (i,j) base pair :

$$\begin{aligned} Z^B(i,j) \quad &= \quad e^{-\mathcal{H}(i,j)/RT} + \sum_{i \leq \ell \leq r \leq j} e^{-\mathcal{I}(i,\ell,r,j)/RT} + \\ &\quad e^{-(a+b)/RT} \cdot \left(\sum_{r=i+1}^{j-\theta-2} Z^M(i+1,r-1) \cdot Z^{M1}(r,j-1) \right). \end{aligned}$$

Partition function for single stem multiloop :

$$Z^{M1}(i,j) \quad = \quad \sum_{r=i+\theta+1}^{j} Z^B(i,r) \cdot e^{-c(j-r)/RT}.$$

Partition function for general multiloop :

$$\begin{aligned} Z^M(i,j) \quad &= \quad \sum_{r=i}^{j-\theta-1} Z^{M1}(r,j) \cdot e^{-(b+c(r-i))/RT} + \\ &\quad \sum_{r=i+\theta+1}^{j-\theta-1} Z^M(i,r-1) \cdot Z^{M1}(r,j) \cdot e^{-b/RT} \end{aligned}$$

Applications

Definition (base pair probability) :
Let $S_{i,j}$ be the subset of all sec. str. which contains the base pair (i,j).

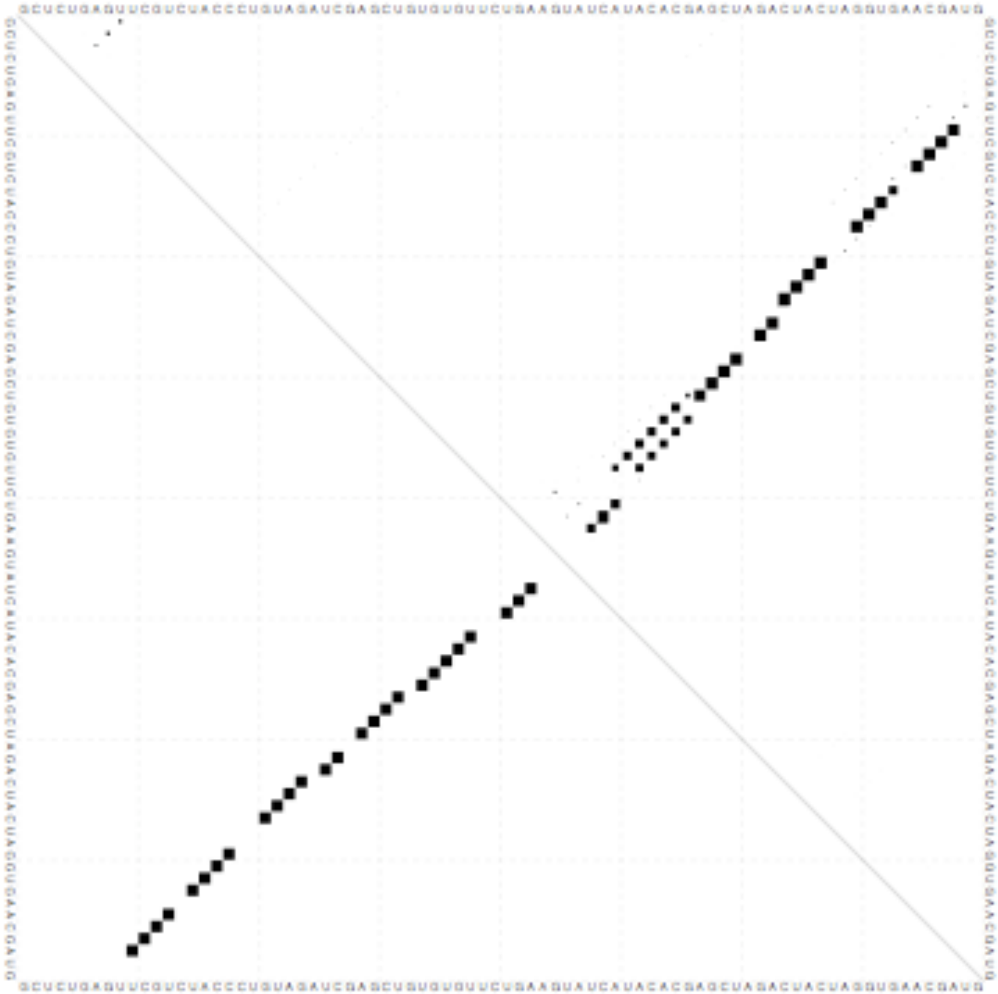
$$P_{i,j} = \frac{Z_{i \cdot j}}{Z} = \frac{\sum_{S \in S_{i,j}} e^{-\beta S}}{Z}$$

$$P_{ij} \;=\; \frac{Z_{1,i-1} \cdot Z'_{i,j} \cdot Z_{j+1,N}}{Z_{1,N}} + \sum_{p<q,q>1} P_{pq} \frac{Z'_{i,j}}{Z'_{p,q}} \cdot$$

$$\Big\{ E_{loop}(p,q,i,j) + Z^{M}_{p+1,j-1} a c^{q-j-1} + Z^{M}_{j+1,q-1} a c^{k-q-1} Z^{M}_{p+1,j-1} Z^{M}_{j+1,q-}$$

Example

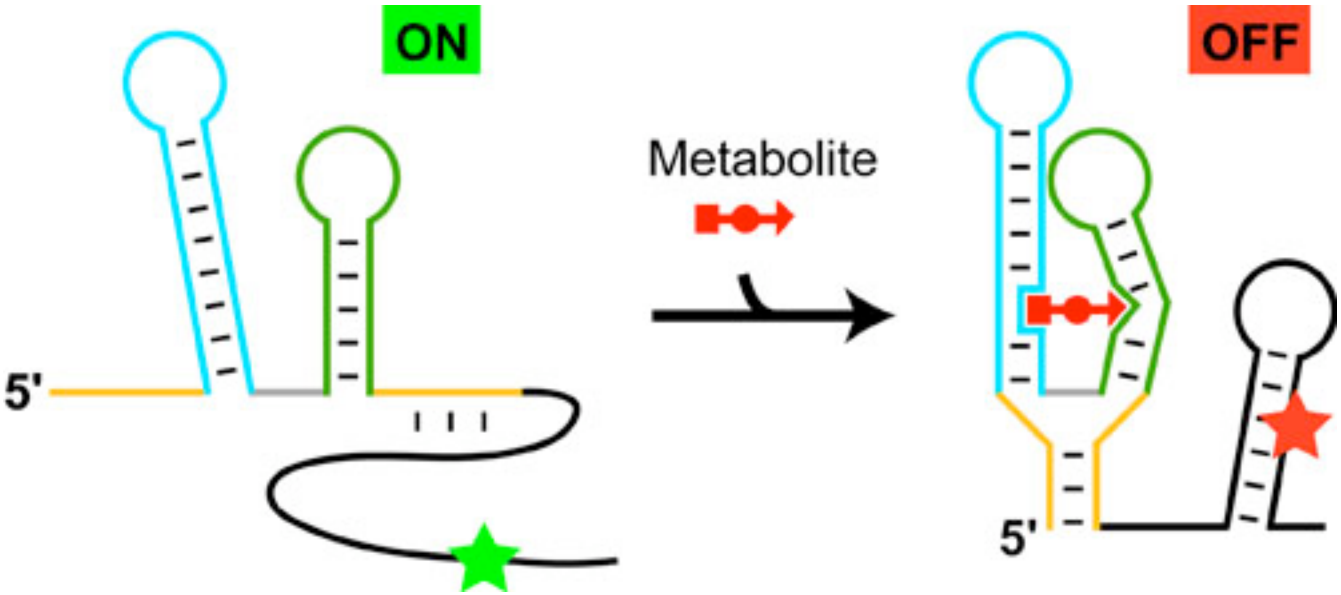
Stochastic Contact Map of pre-miRNA cbr-mir-57



Lower triangle:
MFE Contact Map

(a better) Example

Riboswitch



(Serganov et al.,2006)

RNA with two conformations controlled by a molecule binding.

(a better) Example

Stochastic Contact Map of a riboswitch