The Ensemble of RNA Structures

Example: some good structures of the RNA sequence

```
GGGGGUAUAGCUCAGGGGUAGAGCAUUUGACUGCAGAUCAAGAGGUCCCUGGUUCAAAUCCAGGUGCCCCCU
                                                                 free energy in kcal/mol
((((((..((((.......))))..........((((....))))((((.......))))))))))). -28.10
((((((..((((.......))))....((((.(.......).))))((((.......))))))))))). -27.90
(((((((.((((.......))))(((((((..(((....))))..))))..)))))....))))))). -27.80
(((((((.((((.......))))(((((((..(((....))))..)))..)))))....))))))). -27.80
((((((..((((.......))))....(((..........))))((((.......))))))))))). -27.60
((((((..((((.......))))....((..(.......)..))((((.......))))))))))). -27.50
(((((((.((((.......)))).(((((((..(((....))))..))))..))))....))))))). -27.20
(((((((.((((.......)))).(((((((..(((....))))..)))..)))))....))))))). -27.20
(((((((.((((.......))))..........(((....)))).((((.......))))))))))). -27.20
((((((...((((.......))))..........(((....))))((((.......)))))..))))). -27.20
((((((...(((...(((...(((......)))..)))..)))...((((.......))))))))))). -27.10
(((((((.((((.......))))(((((((...(((....))))...)))..)))))....))))))). -27.00
(((((((.((((.......))))(((((((...(((....))))...)).)))))....))))))). -27.00
((((((..((((.......))))....((((.(.......).)))).((((.......))))))))))). -27.00
((((((..((((.......)))).((((((....).)))))).....((((.......))))))))))). -27.00
((((((..((((.......))))).....((((.....))))((((.......))))))))))). -27.00
((((((...((((.......))))....((((.(.......).))))((((.......)))))..))))). -27.00
(((((((.((((.......))))(((((((..(((......)))..))))..)))))....))))))). -26.70
(((((((.((((.......))))(((((((..(((......)))..)))..)))))....))))))). -26.70
((((((..((((.......))))....((((.(.......).)))).((((.......))))))))))). -26.70
((((((..((((.......)))).((((.(.......).)))))).....((((.......))))))))))). -26.70
((((((...((((.......))))....((((.(.......).))))((((.......)))))..))))). -26.70
```

The set of all nc RNA structures of an RNA sequence S is called *(structure) ensemble $\mathcal{P}$ of S*.

Probability of a Structure

How probable is an RNA structure P for a RNA sequence S?
GOAL: define probability $\Pr[P|S]$.
IDEA: Think of RNA folding as a *dynamic system* of structures (=states of the system). Given much time, a sequence S will form every possible structure P. For each structure there is a probability for observing it at a given time.

This means: we look for a probability distribution!
Requirements: probability depends on energy — the lower the more probable. No additional assumptions!

Distribution of States in a System

Definition (Boltzmann distribution)

Let $\mathcal{X} = \{X_1, \ldots, X_N\}$ denote a system of states, where state X_i has energy E_i. The system is *Boltzmann distributed with temperature* T iff $\Pr[X_i] = exp(-\beta E_i)/Z$ for $Z := \sum_i exp(-\beta E_i)$, where $\beta = (k_B T)^{-1}$.

Remarks

- call $exp(-\beta E_i)$ *Boltzmann weight of* X_i.
- broadly used in physics to describe systems of whatever
- Boltzmann distribution is usually assumed for the *thermodynamic equilibrium* (i.e. after sufficiently much time)
- transfer to RNA easy to see: structures=states, energies
- why temperature?
  - very high temperature: all states equally probable
  - very low temperature: only best states occur
- $k_B \approx 1.38 \times 10^{-23} J/K$ is known as *Boltzmann constant*; β is called *inverse temperature*.

What next?

We assume that the structure ensemble of an RNA sequence is Boltzmann distributed.

- What are the benefits?
  (More than just probabilities of structures . . .)
- Why is it reasonable to assume Boltzmann distribution?
  (Well, a physicist told me . . .)
- How to calculate probabilities efficiently?
  (McCaskill's algorithm)

Benefits of Assuming Boltzmann

Definition

Probability of a structure P for S: $\Pr[P|S] := exp(-\beta E(P))/Z$.

Allows more profound weighting of structures in the ensemble. We need efficient computation of partition function Z!

Even more interesting: probability of structural elements

Definition

Probability of a base pair (i,j) for S:

$$\Pr[(i,j)|S] := \sum_{P \ni (i,j)} \Pr[P|S]$$

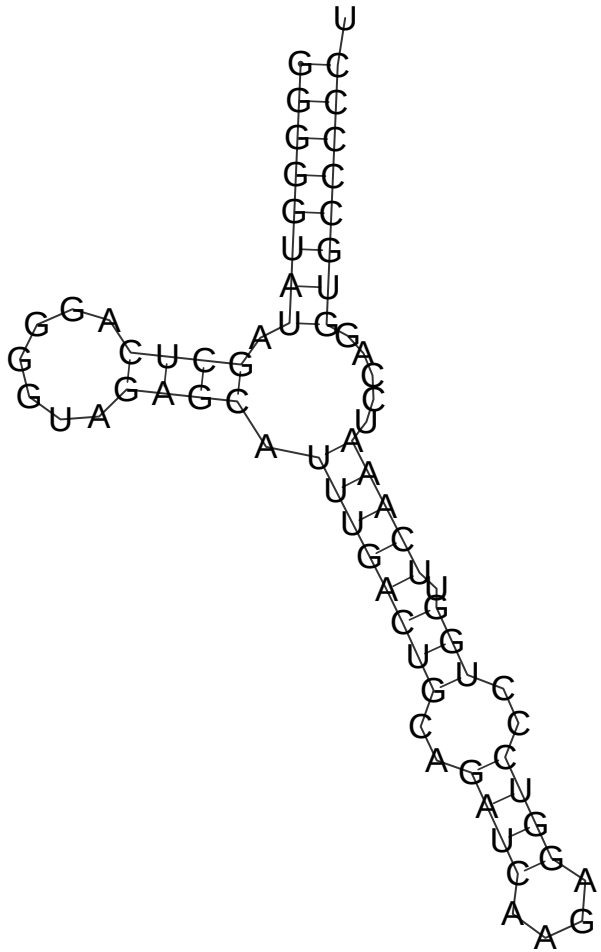
Again, we need Z (and some more). Base pair probabilities enable a new view at the structure ensemble (visually but also algorithmically!).

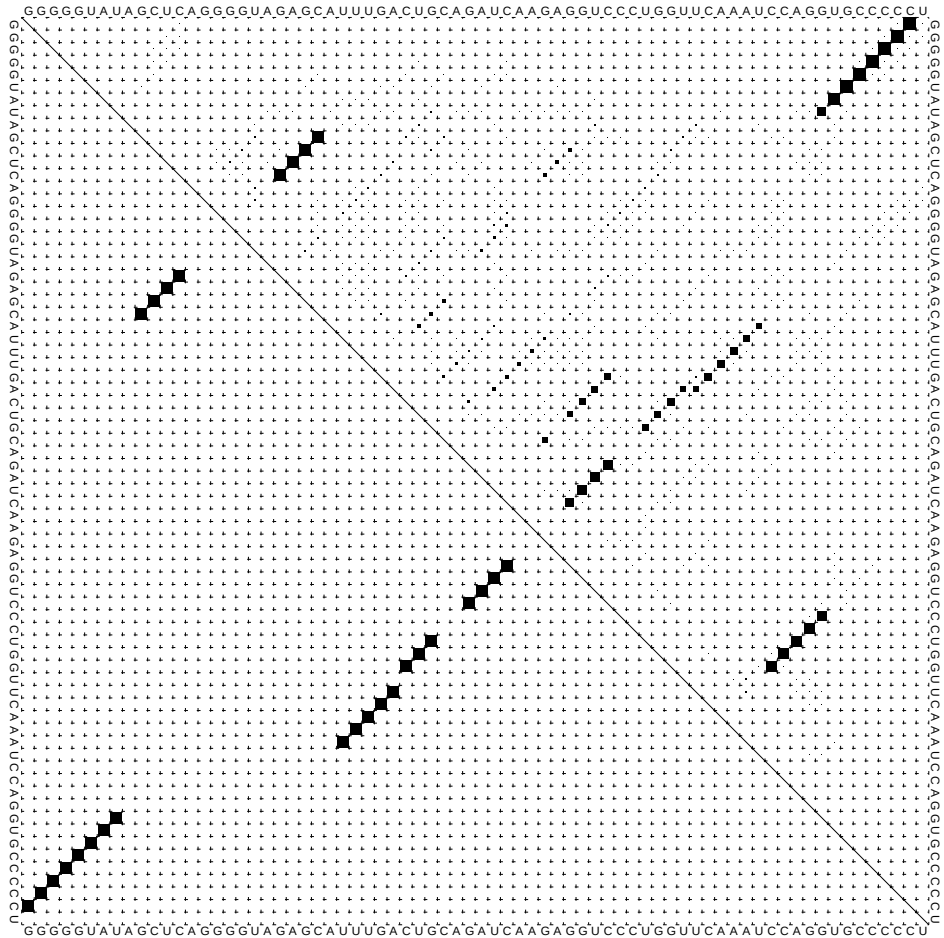
Remark: For RNA, we have "real" temperature, e.g. $T = 37^\circ C$, which determines $\beta = (k_B T)^{-1}$. For calculations pay attention to physical units!

An Immediate Use of Base Pair Probabilities

MFE structure and base pair probability dot plot[1] of a tRNA

GGGGGUAUAGCUCAGGGGUAGAGCAUUUGACUGCAGAUCAAGAGGUCCCUGGUUCAAAUCCAGGUGCCCCCU





[1] computed by "RNAfold -p"

Why Do We Assume Boltzmann

We will give an argument from information theory. We will show: **The Boltzmann distribution makes the least number of assumptions. Formally, the B.d. is the distribution with the lowest information content/maximal (Shannon) entropy.**

As a consequence: without further information about our system, Boltzmann is our best choice.

[What could "further information" mean in a biological context?]
[low information in distribution $\leftrightarrow$ high information in event]

What is Shannon Entropy?

Information of an event

= how much more you know after it happened.

Assume loaded dice that always yield the number 6.

$\Rightarrow$ you know the outcome before you throw the dice.

$\Rightarrow$ no information content (all information in distribution)

Playing lottery, it is very hard to predict the lottery numbers.

$\Rightarrow$ high information content.

This is formalized as:

For some random variable $\mathcal{X} = \{X_1, \ldots, X_N\}$ the amount of information $I[X_i]$ of some event $X_i \in \mathcal{X}$ is defined as

$$I[X_i] = \log_b \frac{1}{\Pr[X_i]} = -\log_b \Pr[X_i]$$

Entropy is the expected amount of information for $\mathcal{X}$:

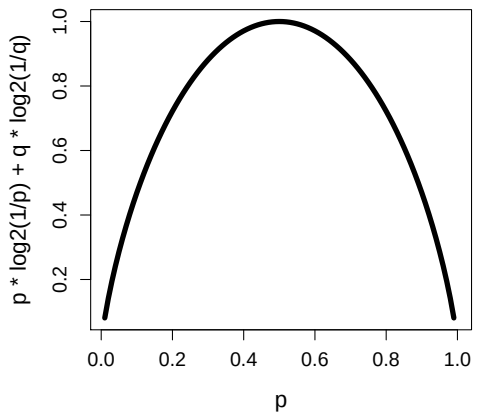
$$H[\mathcal{X}] = \sum_{i=1}^{N} \Pr[X_i] I[X_i]$$

Shannon Entropy (Example)

We toss a coin. For our coin, heads and tails show up with respective probabilities p and q (not necessarily fair).
How uncertain are we about the result?

Answer: expected information

$$H = p \log_b \frac{1}{p} + q \log_b \frac{1}{q}.$$



$p = 0.5, q = 0.5$
$\Rightarrow H = 1$
$\Rightarrow$ max. uncertainty

$p = 1, q = 0$
$\Rightarrow H = 0$
$\Rightarrow$ no uncertainty

In general, define the *Shannon entropy*[2] as

$$H(\vec{p}) := -\sum_{i=1}^{N} p_i \log_b p_i.$$

[2] of a probability distribution $\vec{p}$ over N states $X_1 \ldots X_N$

Formalizing "Least number of assumptions"

Example:
Assume: we have N events. Without further assumptions, we will naturally assume the uniform distribution

$$p_i = \frac{1}{N}.$$

This is the uniquely defined distribution maximizing the entropy $H(\vec{p}) = -\sum_i p_i \log_b p_i$.
It is found by solving the following optimization problem:

maximize the function

$$H(\vec{p}) = -\sum_i p_i \log_b p_i$$

under the side condition $\sum_i p_i = 1$.

Formalizing "Least number of assumptions"

Theorem: Given a system of states $X_1 \ldots X_N$ and energies E_i for X_i. The Boltzmann distribution is the probability distribution $\vec{p}$ that maximizes Shannon entropy

$$H(\vec{p}) = -\sum_{i=1}^{N} p_i \log_b p_i$$

under the assumption that we have a probability distribution

$$\sum_i p_i = 1$$

and under the assumption of known average energy $< E >$ of the system

$$< E >= \sum_{i=1}^{N} p_i E_i.$$

Proof

We show that the Boltzmann distribution is uniquely obtained by solving

$$\textit{maximize function } H(\vec{p}) = -\sum_{i=1}^{N} p_i \ln p_i \qquad [3]$$

under the side conditions

- $C_1(\vec{p}) = \sum_i p_i - 1 \quad = 0$ *and*
- $C_2(\vec{p}) = \sum_i p_i E_i - < E > \quad = 0$

by using the method of Lagrange multipliers.

[3] whether using ln or $\log_b$ is equivalent for maximization

Proof Using Lagrange Multipliers

Following the trick of Lagrange, find the extreme value of

$$L(\vec{p}, \alpha, \beta) = H(\vec{p}) - \alpha C_1(\vec{p}) - \beta C_2(\vec{p}).$$

By construction, $C_1(\vec{p})$ and $C_2(\vec{p})$ are partial derivatives:

$$\frac{\partial L(\vec{p}, \alpha, \beta)}{\partial \alpha} = C_1(\vec{p})$$
$$\frac{\partial L_2(\vec{p}, \alpha, \beta)}{\partial \beta} = C_2(\vec{p})$$

Thus the side conditions hold at the optimum, since there all partial derivatives are 0.

Proof (Ctd.) — Partial Derivatives w.r.t p_j

Futhermore, we need the partial derivatives with respect to p_j

$$\begin{aligned}\frac{\partial L(\vec{p},\alpha,\beta)}{\partial p_j} &= \frac{\partial H(\vec{p})}{\partial p_j} - \alpha\frac{\partial C_1(\vec{p})}{\partial p_j} - \beta\frac{\partial C_2(\vec{p})}{\partial p_j} \\ &= -\frac{\partial \sum_{i=1}^{N} p_i \ln p_i}{\partial p_j} - \alpha\frac{\partial \sum_i p_i - 1}{\partial p_j} - \beta\frac{\partial \sum_i p_i E_i - < E >}{\partial p_j} \\ &= -(\ln p_j + 1) - \alpha - \beta E_j\end{aligned}$$

Proof (Ctd.) — Solve Equations

Finally, we need to solve the system

$$\sum_i p_i E_i - < E > = 0 \qquad (1)$$

$$\sum_i p_i - 1 = 0 \qquad (2)$$

$$\text{for all } j \qquad -(\ln p_j + 1) - \alpha - \beta E_j = 0 \qquad (3)$$

Remarks

- Resolving (3) to p_j and putting into (2) yields a distribution of the same form as the Boltzmann distribution.
- We won't show the dependency of $\beta = k_B T^{-1}$ and $< E >$.

Proof (Ctd)

Equation (3) can be rewritten to:

$$\ln p_j = -\beta E_j - (\alpha + 1).$$

Thus by exponentiation on both sides

$$p_j = \exp(-\beta E_j - \gamma) = \frac{\exp(-\beta E_j)}{\exp(\gamma)}, \tag{4}$$

where $\gamma = (\alpha + 1)$.
By substituting (4) in (2) $\sum_i p_i - 1 = 0$ we get

$$1 = \sum_i \exp(-\beta E_i)/\exp(\gamma) \quad \text{and thus} \quad \exp(\gamma) = \sum_i \exp(-\beta E_i)$$

We insert this in 4 and finally obtain

$$p_j = \frac{\exp(-\beta E_j)}{\sum_i \exp(-\beta E_i)} \tag{5}$$

□

Partition Function

Recall: For probabilities, $\Pr[P|S] = \exp(-\beta E(P))/Z$, we need Z.

Definition

For an RNA sequence S, we call

$$Z := \sum_{P \text{ nc RNA structure for } S} \exp(-\beta E(P))$$

the *partition function (of the RNA ensemble $\mathcal{P}$) of S*.

Remark

Naive computation of Z: exponential, since ensemble size is exponential in $|S|$.

Excursion: Counting of Structures

Computation of Z similar to counting

Problem of computing the partition function is similar to counting the structures in the ensemble $\mathcal{P}$. Partition function is a weighted sum, in counting we "weight" structures by 1.

How to count nc RNA structures for S?

- naïve: enumerate $\Rightarrow$ exponential
- efficient: DP with decomposition a la Nussinov

Example: S=CGAGC for minimal loop length m=0.

Counting of Structures: S=CGAGC

C_1	G_2	A_3	G_4	C_5	
					C_1
					G_2
					A_3
					G_4
					C_5

Counting of Structures: S=CGAGC

C_1	G_2	A_3	G_4	C_5	
{.}	{..,()}	{...,().}	{....,()..,(..)}	{.....,()...,(..)., .(..),...(),().()}	C_1
	{.}	{..}	{...}	{....,(..)}	G_2
		{.}	{..}	{...,.()}	A_3
			{.}	{..,()}	G_4
				{.}	C_5

Subensembles

Definition (Subensemble)

Define the *ij-subensemble* $\mathcal{P}_{ij}$ *of* S (for $1 \leq i \leq j \leq n$) as

$$\mathcal{P}_{ij} := \text{set of all nc RNA } ij\text{-substructures } P \text{ of } S.$$

where:

Definition (RNA Substructure)

An RNA structure P of S is called *ij-substructure of* S iff $P \subseteq \{i, \ldots, j\}^2$.

Remarks

- Example: see last slide, $\mathcal{P}_{14} = \{\{\}, \{(1,2)\}, \{(1,4)\}\}$,
  $\mathcal{P}_{15} = \{\{\}, \{(1,2)\}, \{(1,4)\}, \{(2,5)\}, \{(4,5)\}, \{(1,2),(4,5)\}\}$
- ensemble $\mathcal{P}$ of S: $\mathcal{P} = \mathcal{P}_{1n}$
- $\mathcal{P}_{ij} = \{\{\}\}$ for $j < i + m$ (min. loop size m)

Efficient Counting of Structures

Define: $C_{ij} := |\mathcal{P}_{ij}|$. ($\Rightarrow$ DP-matrix C)

Computation of C_{ij}

for $j - i \leq m$: $C_{ij} = 1$, since $\mathcal{P}_{ij} = \{\{\}\}$

for $j - i > m$: recurse!

$\mathcal{P}_{ij}$ consists of structures

$$\mathcal{P}_{ij-1} \qquad (\ j \text{ unpaired})$$

and structures

$$\mathcal{P}_{ik-1} \otimes \mathcal{P}_{k+1j-1} \otimes \{\{(k,j)\}\} \qquad (\ k,j \text{ paired }),$$

where:

"$\otimes$" combines all structures in one set with all structures in a second set.

Define: $\mathcal{P} \otimes \mathcal{Q} := \{P \cup Q | P \in \mathcal{P}, Q \in \mathcal{Q}\}$.

Computation of C_{ij}

for $j - i > m$:

$$\mathcal{P}_{ij} = \mathcal{P}_{ij-1} \cup \bigcup_{\substack{i \le k < j-m \\ S_k, S_j \text{ compl.}}} \mathcal{P}_{ik-1} \otimes \mathcal{P}_{k+1j-1} \otimes \{\{(k,j)\}\}$$

this means for C_{ij}: recall $C_{ij} = |\mathcal{P}_{ij}|$

$$C_{ij} = C_{ij-1} + \sum_{\substack{i \le k < j-m \\ S_k, S_j \text{ compl.}}} C_{ik-1} \cdot C_{k+1j-1} \cdot 1$$

Remarks

- by DP: compute ensemble size C_{1n} in $O(n^3)$ time and $O(n^2)$ space.
- why "translates" $\cup$ to $+$ and $\otimes$ to $\cdot$? $\Leftarrow$ all unions were disjoint!
  i.e.: 1.) cases in "$\mathcal{P}_{ij}$ consists of ..." are disjoint
  2.) structures combined by $\otimes$ are disjoint

Example

decompose sequence $S_{15} = C_1 G_2 A_3 G_4 C_5$

1. subsequence $C_1 G_2 A_3 G_4$ and C_5 unpaired
   $C_{15} \leftarrow C_{14}$
2. a.) k=2. C_1, $A_3 G_4$, base pair $(2,5)$
   $\mathcal{P}_{15} \leftarrow \mathcal{P}_{11} \otimes \mathcal{P}_{34} \otimes \{\{(2,5)\}\}$
   $C_{15} \leftarrow C_{11} \cdot C_{34} \cdot 1$
   b.) k=4. $C_1 G_2 A_3$, base pair $(4,5)$
   $\mathcal{P}_{15} \leftarrow \mathcal{P}_{13} \otimes \mathcal{P}_{54} \otimes \{\{(4,5)\}\}$
   $C_{15} \leftarrow C_{13} \cdot C_{54} \cdot 1$

ad 2b.)

$$\mathcal{P}_{13} \otimes \mathcal{P}_{54} \otimes \{\{(4,5)\}\} = \{\{\}, \{(1,2)\}\} \otimes \{\{\}\} \otimes \{\{(4,5)\}\}$$
$$= \{\{(4,5)\}, \{(1,2),(4,5)\}\}$$

Example

decompose sequence $S_{15} = C_1 G_2 A_3 G_4 C_5$

1. subsequence $C_1 G_2 A_3 G_4$ and C_5 unpaired
   $C_{15} \leftarrow C_{14}$
2. a.) k=2. C_1, $A_3 G_4$, base pair $(2,5)$
   $\mathcal{P}_{15} \leftarrow \mathcal{P}_{11} \otimes \mathcal{P}_{34} \otimes \{\{(2,5)\}\}$
   $C_{15} \leftarrow C_{11} \cdot C_{34} \cdot 1$
   b.) k=4. $C_1 G_2 A_3$, base pair $(4,5)$
   $\mathcal{P}_{15} \leftarrow \mathcal{P}_{13} \otimes \mathcal{P}_{54} \otimes \{\{(4,5)\}\}$
   $C_{15} \leftarrow C_{13} \cdot C_{54} \cdot 1$

ad 2b.)

$$\mathcal{P}_{13} \otimes \mathcal{P}_{54} \otimes \{\{(4,5)\}\} = \{\{\}, \{(1,2)\}\} \otimes \{\{\}\} \otimes \{\{(4,5)\}\}$$
$$= \{\{(4,5)\}, \{(1,2),(4,5)\}\}$$

Counting vs. Structure Prediction

Counting

init $C_{ij} = 1 \qquad (j - i \leq m)$

recurse $C_{ij} = C_{ij-1} + \sum_{\substack{i \leq k < j-m \\ S_k, S_j \text{ compl.}}} C_{ik-1} \cdot C_{k+1j-1} \cdot 1$

Prediction

init $N_{ij} = 0 \qquad (j - i \leq m)$

recurse $N_{ij} = \max\{N_{ij-1}, \max_{\substack{i \leq k < j-m \\ S_k, S_j \text{ compl.}}} N_{ik-1} + N_{k+1j-1} + 1\}$

Remarks

- "translation" Prediction $\rightarrow$ Counting : $\max \rightarrow +$, $+ \rightarrow \cdot$
- only possible since sets disjoint, i.e.
  - disjoint cases (no "ambiguity")
  - non-overlapping decomposition in each single case

Back to Computing the Partition Function

Recall: For probabilities, $\Pr[P|S] = \exp(-\beta E(P))/Z$, we need Z.
We defined: $Z := \sum_{P \in \mathcal{P}} \exp(-\beta E(P))$
We claimed: Problem of computing the partition function is similar to counting the structures in the ensemble $\mathcal{P}$. Partition function is a weighted sum, in counting we "weight" structures by 1.

Definition (Partition Function of a Set of Structures)

In analogy to $C_{ij} = |\mathcal{P}_{ij}| = \sum_{P \in \mathcal{P}_{ij}} 1$, define the *partition function* $Z_{\mathcal{P}}$ *for the set of RNA structures* $\mathcal{P}$ *of* S by

$$Z_{\mathcal{P}} := \sum_{P \in \mathcal{P}} exp(-\beta E(P)).$$

Idea: compute the $Z_{\mathcal{P}_{ij}}$ recursively $\Rightarrow$ efficient by DP.

Disjoint Decomposition — when to add?

Definition (Disjoint Sets)

Two sets of RNA structures $\mathcal{P}_1$ and $\mathcal{P}_2$ are *(structurally) disjoint* iff $\mathcal{P}_1 \cap \mathcal{P}_2 = \{\}$.

Proposition (Disjoint Decomposition)

Let $\mathcal{P}$, $\mathcal{P}_1$, and $\mathcal{P}_2$ be sets of structures of an RNA sequence S. If $\mathcal{P}_1$ and $\mathcal{P}_2$ are structurally disjoint and $\mathcal{P} = \mathcal{P}_1 \cup \mathcal{P}_2$, then

$$Z_{\mathcal{P}} = Z_{\mathcal{P}_1} + Z_{\mathcal{P}_2}.$$

Proof

Proof.

$$
\begin{aligned}
Z_{\mathcal{P}} &= \sum_{P \in \mathcal{P}} \exp(-\beta E(P)) \\
&=_{\text{disjoint}} \sum_{P \in \mathcal{P}_1 \uplus \mathcal{P}_2} \exp(-\beta E(P)) \\
&= \sum_{P \in \mathcal{P}_1} \exp(-\beta E(P)) + \sum_{P \in \mathcal{P}_2} \exp(-\beta E(P)) \\
&= Z_{\mathcal{P}_1} + Z_{\mathcal{P}_2}
\end{aligned}
$$

□

Independent Decomposition — when to multiply?

Definition (Independent Sets)

Let S be an RNA sequence. Two sets of nc RNA structures $\mathcal{P}_1$ and $\mathcal{P}_2$ for S are *structurally independent* iff for all $P_1 \in \mathcal{P}_1$ and $P_2 \in \mathcal{P}_2$

1. $P_1 \cap P_2 = \{\}$.
2. each loop/secondary structure element of the RNA structure $P = P_1 \cup P_2$ is either a loop of P_1 or one of P_2.

Proposition (Independent Decomposition)

Let $\mathcal{P}_1$ and $\mathcal{P}_2$ be structurally independent sets of nc RNA structures for RNA sequence S and $\mathcal{P} = \mathcal{P}_1 \otimes \mathcal{P}_2$. Then:

$$Z_{\mathcal{P}} = Z_{\mathcal{P}_1} \cdot Z_{\mathcal{P}_2}$$

Remark: Condition (1) suffices for energy functions based on scoring base pairs (like in Nussinov). For loop-based energy models, we need (2), which implies $E(P_1 \cup P_2) = E(P_1) + E(P_2)$.

Proof

Proof.

$$
\begin{aligned}
Z_{\mathcal{P}} &= \sum_{P \in \mathcal{P}} \exp(-\beta E(P)) \\
&=_{\textit{indep.}(1)} \sum_{P_1 \in \mathcal{P}_1, P_2 \in \mathcal{P}_2} \exp(-\beta E(P_1 \cup P_2)) \\
&=_{\textit{indep.}(2)} \sum_{P_1 \in \mathcal{P}_1, P_2 \in \mathcal{P}_2} \exp(-\beta (E(P_1) + E(P_2))) \\
&= \sum_{P_1 \in \mathcal{P}_1} \sum_{P_2 \in \mathcal{P}_2} \exp(-\beta E(P_1)) \exp(-\beta E(P_2)) \\
&= \sum_{P_1 \in \mathcal{P}_1} \exp(-\beta E(P_1)) \left(\sum_{P_2 \in \mathcal{P}_2} \exp(-\beta E(P_2)) \right) \\
&= \sum_{P_1 \in \mathcal{P}_1} \exp(-\beta E(P_1)) Z_{\mathcal{P}_2} \\
&= Z_{\mathcal{P}_1} \cdot Z_{\mathcal{P}_2}
\end{aligned}
$$

Adding and Multiplying of Partition Functions

in the same way as for counts!

Counting

init $C_{ij} = 1 \qquad (j - i \leq m)$

recurse $C_{ij} = C_{ij-1} + \sum_{\substack{i \leq k < j-m \\ S_k, S_j \text{ compl.}}} C_{ik-1} \cdot C_{k+1j-1} \cdot 1$

Partition Function (Nussinov Style)

init $Z^N_{\mathcal{P}_{ij}} = 1 \qquad (j - i \leq m)$

recurse

$$Z^N_{\mathcal{P}_{ij}} = Z^N_{\mathcal{P}_{ij-1}} + \sum_{\substack{i \leq k < j-m \\ S_k, S_j \text{ compl.}}} Z^N_{\mathcal{P}_{ik-1}} \cdot Z^N_{\mathcal{P}_{k+1j-1}} \cdot \exp(-\beta \text{“}E(basepair)\text{”})$$

Remarks

- “E(basepair)”: e.g. -1 or depending on S_i and S_j for base pair (i, j)
- This partitition function variant of the Nussinov algorithm can **not** compute the partition function for the loop-based energy model(!)

Way to RNA Partition Function

- Partition function adding/multiplying like in counting
  Attention: only for disjoint/independent sets
- Loop energy model
  Zuker: how to decompose structure space
  how to compute the energies (as sum of loop energies)

What next?

Develop recursions for partition function using "real" RNA energies

Plan: rewrite Zuker-algo into its partition function variant

What is missing?

Is Zuker's decomposition of structure space

- disjoint?
- independent?

Zuker to Partition Function Variant: Example

$$W_{ij} = \min \begin{cases} W_{ij-1} \\ \min_{i \leq k < j-m} W_{ik-1} + V_{kj} \end{cases}$$

$$Z_{\mathcal{P}_{ij}} = Z_{\mathcal{P}_{ij-1}} + \sum_{i \leq k < j-m} Z_{\mathcal{P}_{ik-1}} \cdot Z_{\mathcal{P}'_{kj}}$$

Zuker to Partition Function Variant: Example

$$W_{ij} = \min \begin{cases} W_{ij-1} \\ \min_{i \leq k < j-m} W_{ik-1} + V_{kj} \end{cases}$$

$$Z_{\mathcal{P}_{ij}} = Z_{\mathcal{P}_{ij-1}} + \sum_{i \leq k < j-m} Z_{\mathcal{P}_{ik-1}} \cdot Z_{\mathcal{P}'_{kj}}?$$

Decomposition by Zuker

$$W_{ij} = \min \begin{cases} W_{ij-1} \\ \min_{i \le k < j-m} W_{ik-1} + V_{kj} \end{cases}$$

$$V_{ij} = \min \begin{cases} \mathrm{eH}(i,j), \min_{i<i'<j'<j} V_{i',j'} + \mathrm{eSBI}(i,j,i',j') \\ \min_{i<k<j} WM_{i+1k-1} + WM_{kj-1} + a \end{cases}$$

$$WM_{ij} = \min \begin{cases} WM_{ij-1} + c, WM_{i+1j} + c, V_{ij} + b \\ \min_{i<k<j} WM_{ik-1} + WM_{kj} \end{cases}$$

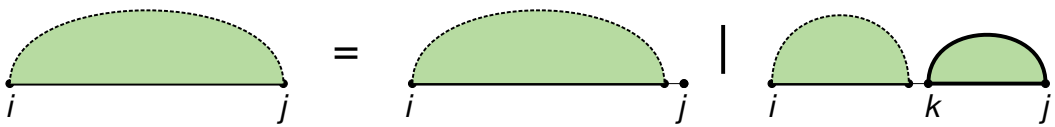
Remark

Combine energy function for stacking and interior loop:

$$\mathrm{eSBI}(i,j,i',j') := \begin{cases} \mathrm{eS}(i,j) & i' = i+1 \wedge j' = j-1 \\ \mathrm{eL}(i,j,i',j') & \text{otherwise} \end{cases}$$

Decomposition by Zuker

$$W_{ij} = \min \begin{cases} W_{ij-1} \\ \min_{i \le k < j-m} W_{ik-1} + V_{kj} \end{cases}$$



$$V_{ij} = \min \begin{cases} \mathrm{eH}(i,j), \min_{i<i'<j'<j} V_{i',j'} + \mathrm{eSBI}(i,j,i',j') \\ \min_{i<k<j} WM_{i+1k-1} + WM_{kj-1} + a \end{cases}$$

$$WM_{ij} = \min \begin{cases} WM_{ij-1} + c, WM_{i+1j} + c, V_{ij} + b \\ \min_{i<k<j} WM_{ik-1} + WM_{kj} \end{cases}$$

Remark

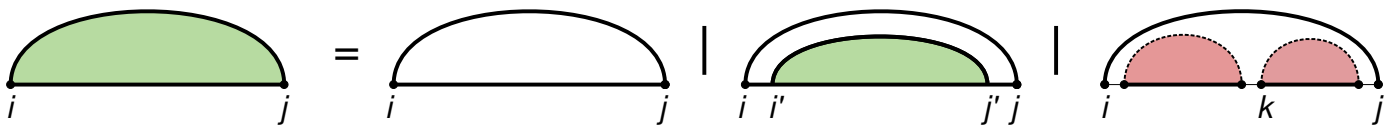
Combine energy function for stacking and interior loop:

$$\mathrm{eSBI}(i,j,i',j') := \begin{cases} \mathrm{eS}(i,j) & i' = i+1 \wedge j' = j-1 \\ \mathrm{eL}(i,j,i',j') & \text{otherwise} \end{cases}$$

Decomposition by Zuker

$$W_{ij} = \min \begin{cases} W_{ij-1} \\ \min_{i \le k < j-m} W_{ik-1} + V_{kj} \end{cases}$$

$$V_{ij} = \min \begin{cases} \mathrm{eH}(i,j), \min_{i<i'<j'<j} V_{i',j'} + \mathrm{eSBI}(i,j,i',j') \\ \min_{i<k<j} WM_{i+1k-1} + WM_{kj-1} + a \end{cases}$$



$$WM_{ij} = \min \begin{cases} WM_{ij-1} + c, WM_{i+1j} + c, V_{ij} + b \\ \min_{i<k<j} WM_{ik-1} + WM_{kj} \end{cases}$$

Remark

Combine energy function for stacking and interior loop:

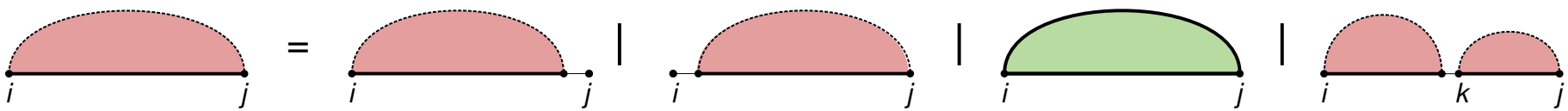
$$\mathrm{eSBI}(i,j,i',j') := \begin{cases} \mathrm{eS}(i,j) & i' = i+1 \wedge j' = j-1 \\ \mathrm{eL}(i,j,i',j') & \text{otherwise} \end{cases}$$

Decomposition by Zuker

$$W_{ij} = \min \begin{cases} W_{ij-1} \\ \min_{i \le k < j-m} W_{ik-1} + V_{kj} \end{cases}$$

$$V_{ij} = \min \begin{cases} \mathrm{eH}(i,j), \min_{i<i'<j'<j} V_{i',j'} + \mathrm{eSBI}(i,j,i',j') \\ \min_{i<k<j} WM_{i+1k-1} + WM_{kj-1} + a \end{cases}$$

$$WM_{ij} = \min \begin{cases} WM_{ij-1} + c, WM_{i+1j} + c, V_{ij} + b \\ \min_{i<k<j} WM_{ik-1} + WM_{kj} \end{cases}$$

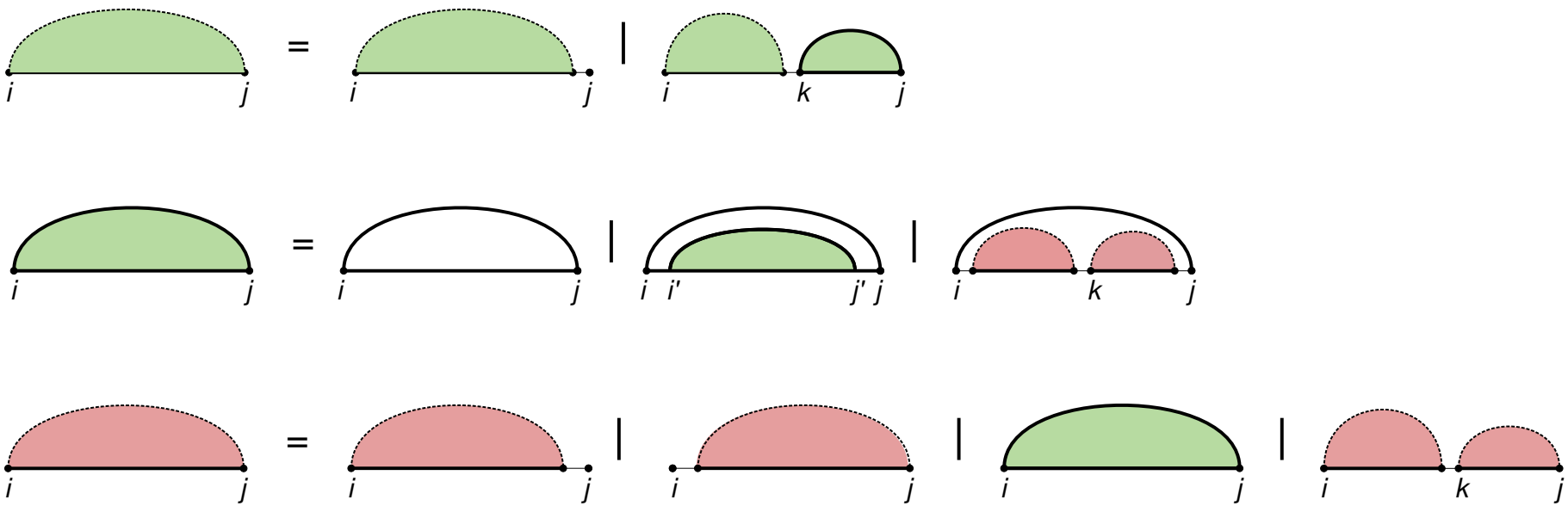


Remark

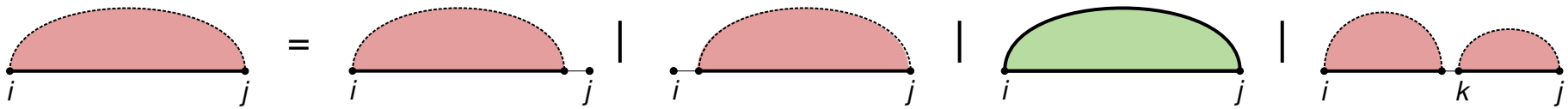
Combine energy function for stacking and interior loop:

$$\mathrm{eSBI}(i,j,i',j') := \begin{cases} \mathrm{eS}(i,j) & i' = i+1 \wedge j' = j-1 \\ \mathrm{eL}(i,j,i',j') & \text{otherwise} \end{cases}$$

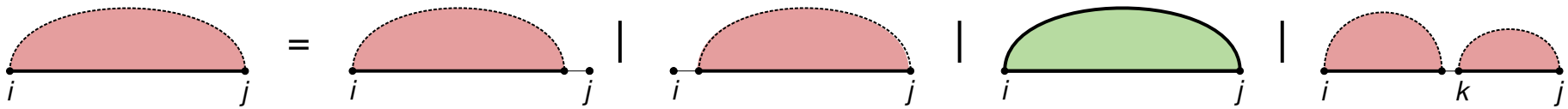
Decomposition of Structure Space by Zuker

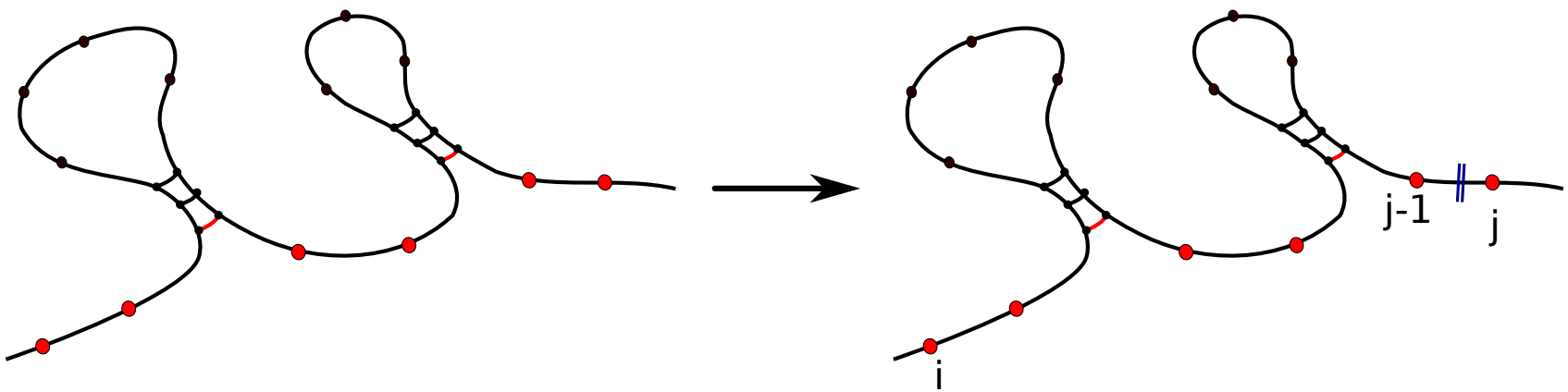


Ambiguity Example

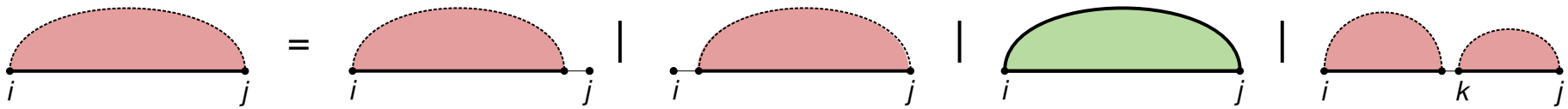


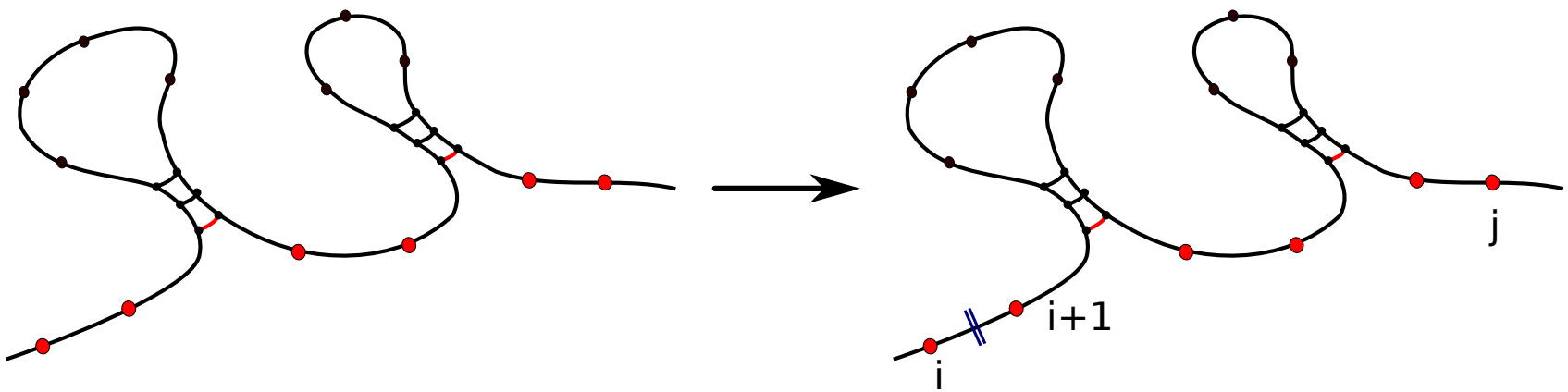
Ambiguity Example



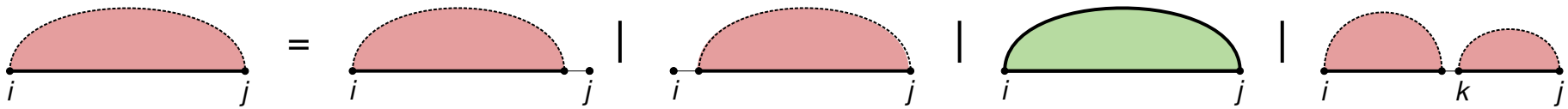


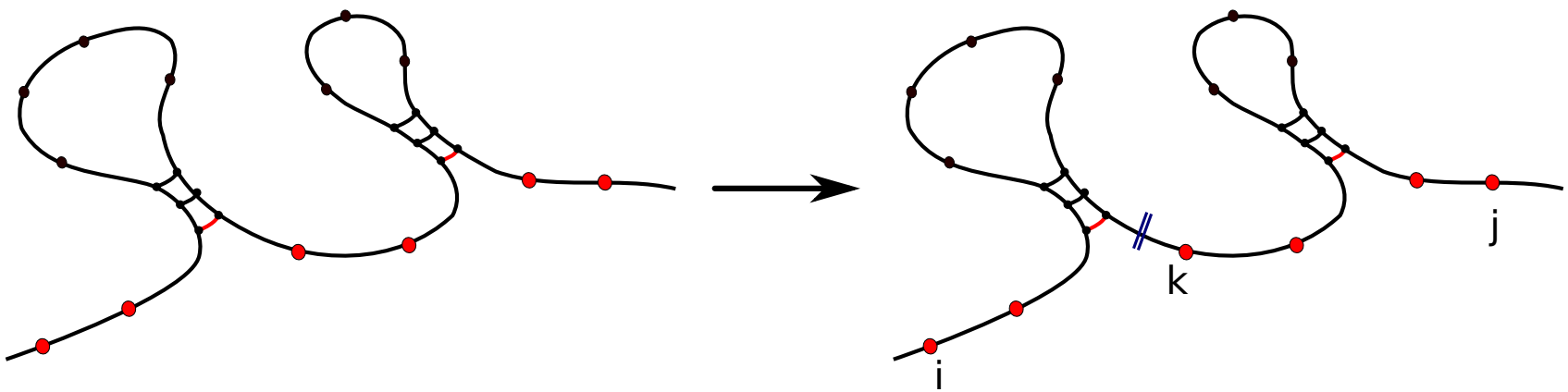
Ambiguity Example



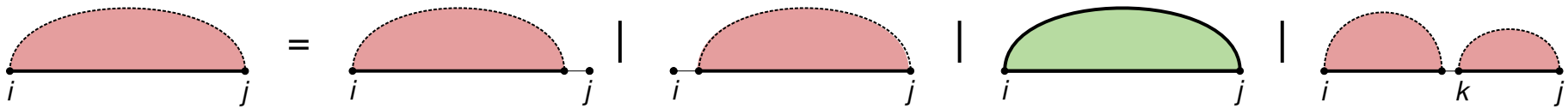


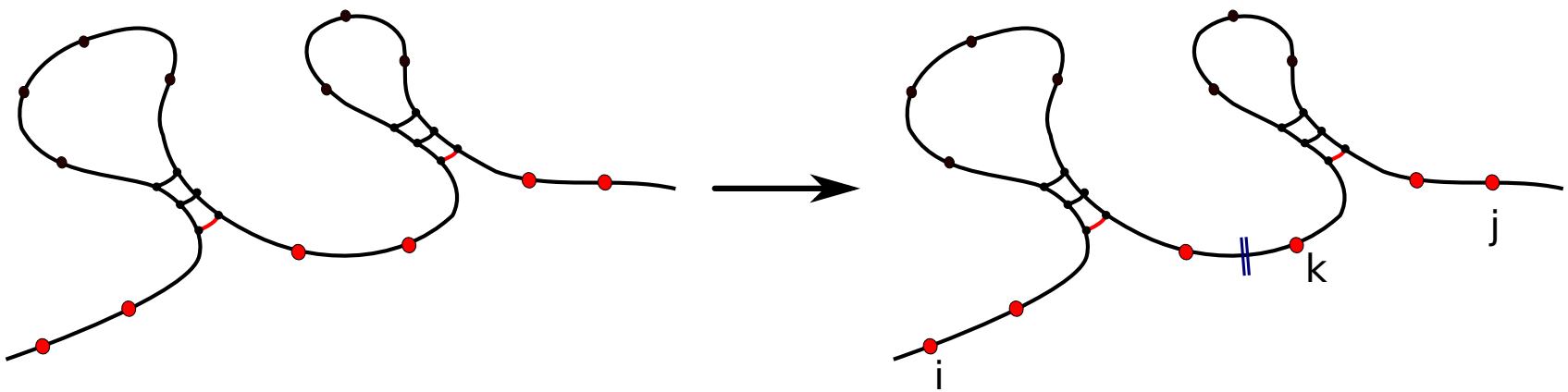
Ambiguity Example



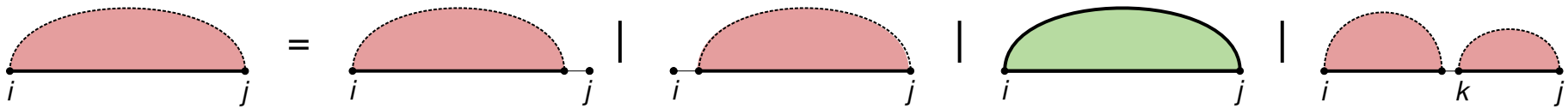


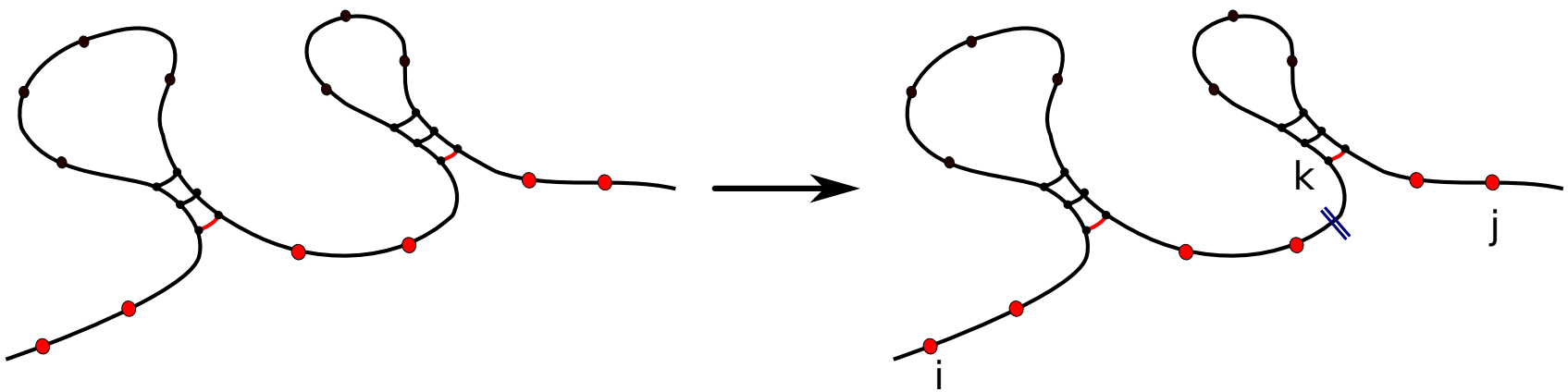
Ambiguity Example



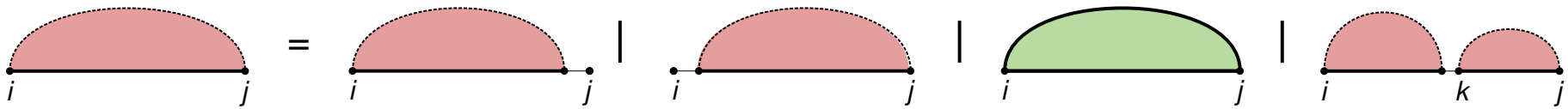


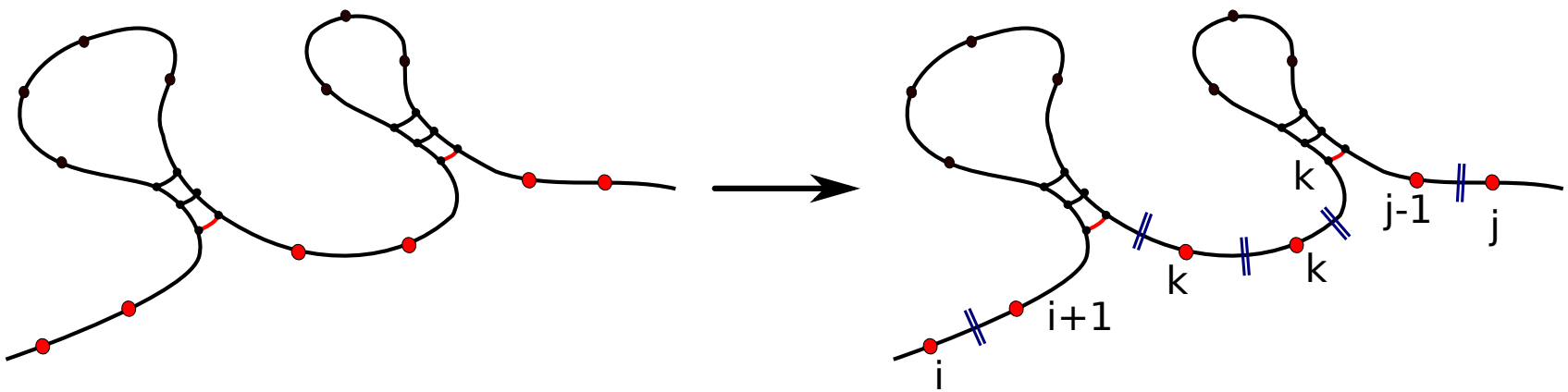
Ambiguity Example



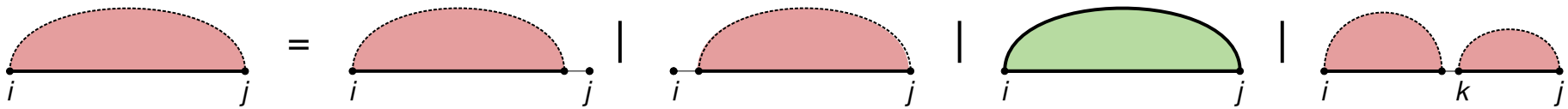


Ambiguity Example

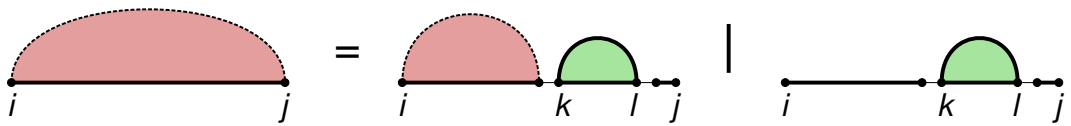




Discarding Ambiguity

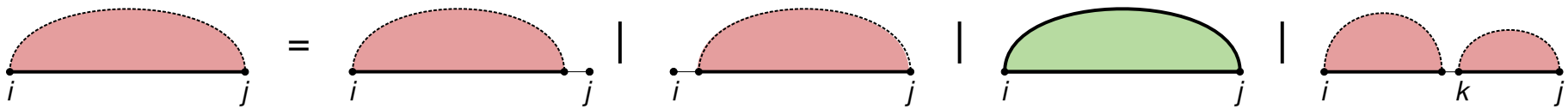


idea for unique choice: always cut at rightmost base pair

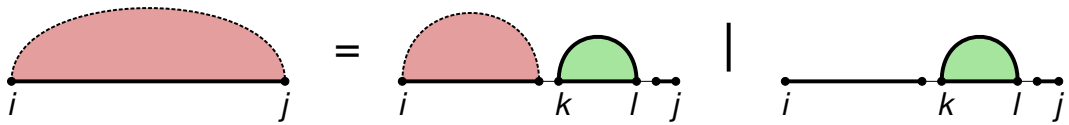


efficient way to do it with additional matrix:

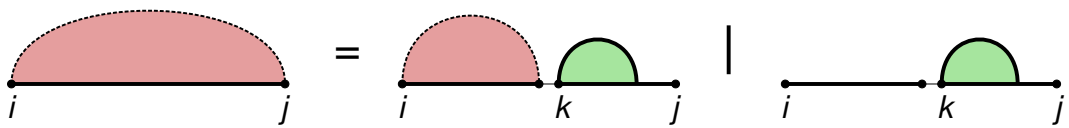
Discarding Ambiguity

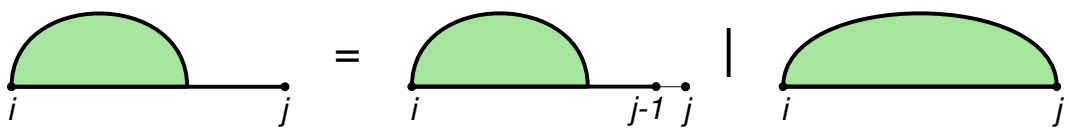


idea for unique choice: always cut at rightmost base pair



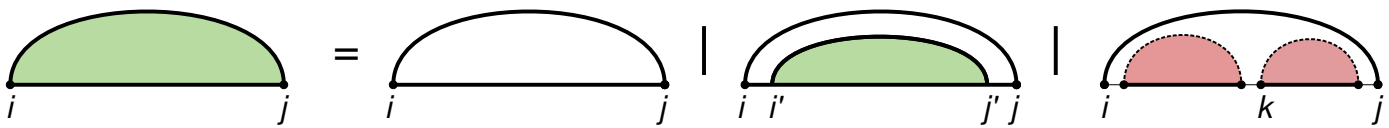
efficient way to do it with additional matrix:





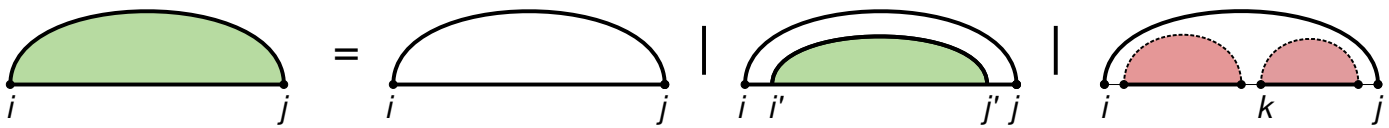
Discarding Ambiguity

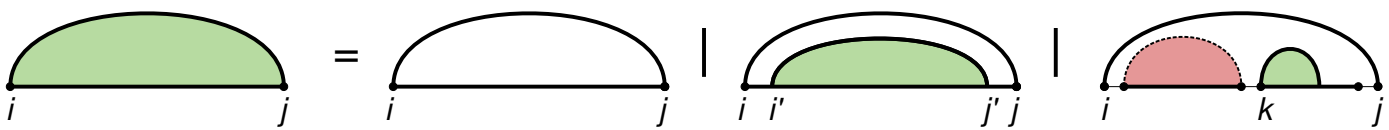
ambiguity in case for closed structures:



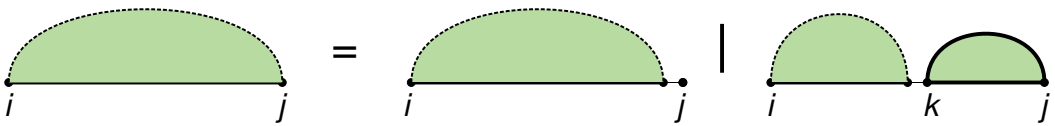
Discarding Ambiguity

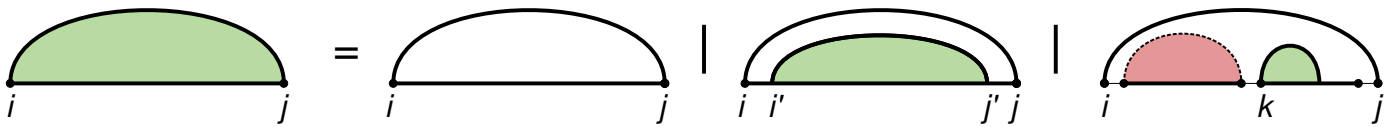
ambiguity in case for closed structures:

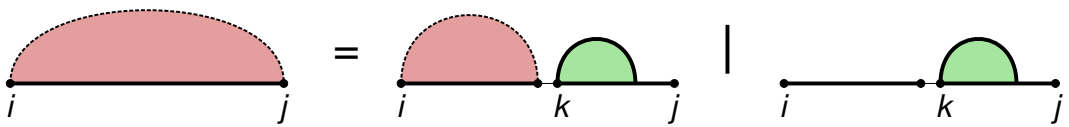


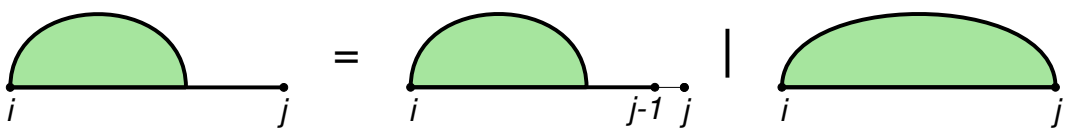


Unambiguous Decomposition: Summary









McCaskill-Algorithm: Matrices

Given RNA sequence $S = S_1 \dots S_n$. As usual, $\mathcal{P}_{ij}$ denotes the ij-subensemble of S.

- Matrix $Q = (Q_{ij})_{1 \le i < j \le n}$
- Matrix $Q^b = (Q^b_{ij})_{1 \le i < j \le n}$
- Matrix $Q^m = (Q^m_{ij})_{1 \le i < j \le n}$
- Matrix $Q^{m1} = (Q^{m1}_{ij})_{1 \le i < j \le n}$

McCaskill-Algorithm: Matrices

Given RNA sequence $S = S_1 \dots S_n$. As usual, $\mathcal{P}_{ij}$ denotes the ij-subensemble of S.

- Matrix $Q = (Q_{ij})_{1 \le i < j \le n}$

$$Q_{ij} := Z_{\mathcal{P}_{ij}}$$

- Matrix $Q^b = (Q^b_{ij})_{1 \le i < j \le n}$

$$Q^b_{ij} := Z_{\mathcal{P}^b_{ij}}, \text{ where } \mathcal{P}^b_{ij} := \{P \in \mathcal{P}_{ij} \mid (i,j) \in P\}$$

- Matrix $Q^m = (Q^m_{ij})_{1 \le i < j \le n}$

$$Q^m_{ij} := Z^m_{\mathcal{P}^m_{ij}}, \text{ where } \mathcal{P}^m_{ij} := \{P \in \mathcal{P}_{ij} \mid P \text{ non-empty}\},$$

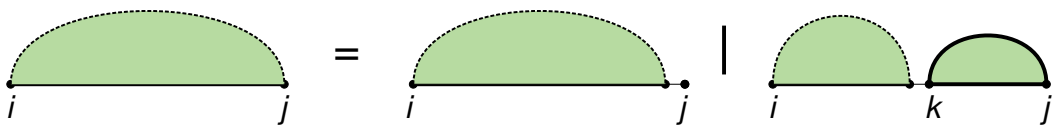
  Z^m includes multiloop-contributions (for unpaired bases, inner base pairs), see earlier modification E^m.

- Matrix $Q^{m1} = (Q^{m1}_{ij})_{1 \le i < j \le n}$

$$Q^{m1}_{ij} := Z^m_{\mathcal{P}^{m1}_{ij}},$$

  where $\mathcal{P}^{m1}_{ij} := \{P \in \mathcal{P}_{ij} \mid \exists k : (i,k) \in P \wedge k+1, \dots, j \text{ unpaired in } P\}$.

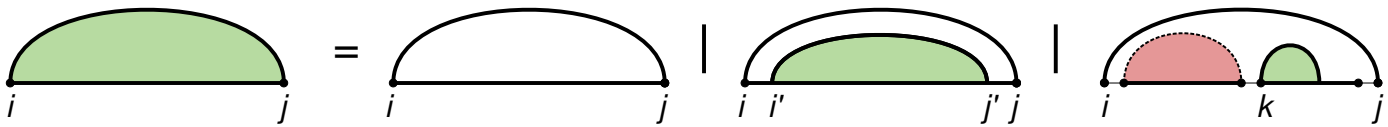
Q-recursion



$$Q_{ij} = 1 \quad (i \geq j - m)$$

$$Q_{ij} = Q_{ij-1} + \sum_{i \leq k < j-m} Q_{ik-1} \cdot Q^b_{kj}$$

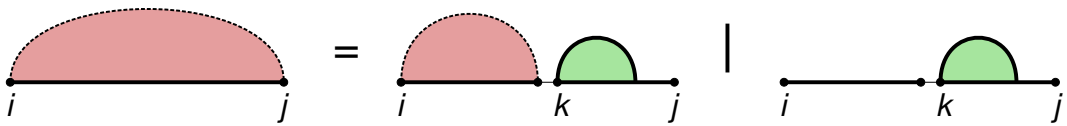
Q^b-recursion



$$Q^b_{ij} = 0 \quad (i \geq j - m)$$

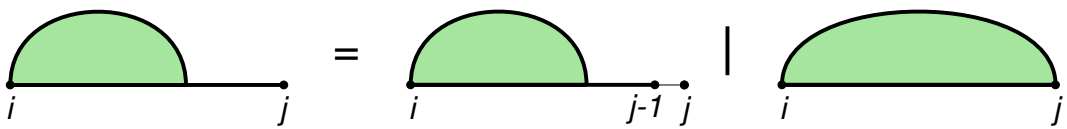
$$Q^b_{ij} = \exp(-\beta\, \mathrm{eH}(i,j)) + \sum_{i<i'<j'<j} \exp(-\beta\, \mathrm{eSBI}(i,j,i',j')) \cdot Q^b_{i'j'} + \sum_{i<k<j-m-1} Q^m_{i+1k-1} \cdot Q^{m1}_{kj-1} \cdot \exp(-\beta a)$$

Q^m-recursion and Q^{m1}-recursion



$$Q_{ij}^{m} = 0 \quad (i \geq j - m)$$

$$Q_{ij}^{m} = \sum_{i \leq k < j-m} (Q_{ik-1}^{m} + \exp(-\beta(k-i)c)) \cdot Q_{kj}^{m1}$$



$$Q_{ij}^{m1} = 0 \quad (i \geq j - m)$$

$$Q_{ij}^{m1} = \sum_{i+m < k \leq j} Q_{ik}^{b} \cdot \exp(-\beta b) \cdot \exp(-\beta(j-k)c)$$

McCaskill — Partition Function — Summary

$$Q_{ij} = 1 \quad (i \geq j - m)$$

$$Q_{ij} = Q_{ij-1} + \sum_{i \leq k < j-m} Q_{ik-1} \cdot Q^b_{kj}$$

$$Q^b_{ij} = 0 \quad (i \geq j - m)$$

$$Q^b_{ij} = \exp(-\beta\, \mathrm{eH}(i,j)) + \sum_{i < i' < j' < j} \exp(-\beta\, \mathrm{eSBI}(i,j,i',j')) \cdot Q^b_{i'j'}$$

$$+ \sum_{i < k < j-m-1} Q^m_{i+1k-1} \cdot Q^{m1}_{kj-1} \cdot \exp(-\beta a)$$

$$Q^m_{ij} = 0 \quad (i \geq j - m)$$

$$Q^m_{ij} = \sum_{i \leq k < j-m} (\exp(-\beta(k-i)c) + Q^m_{ik-1}) \cdot Q^{m1}_{kj}$$

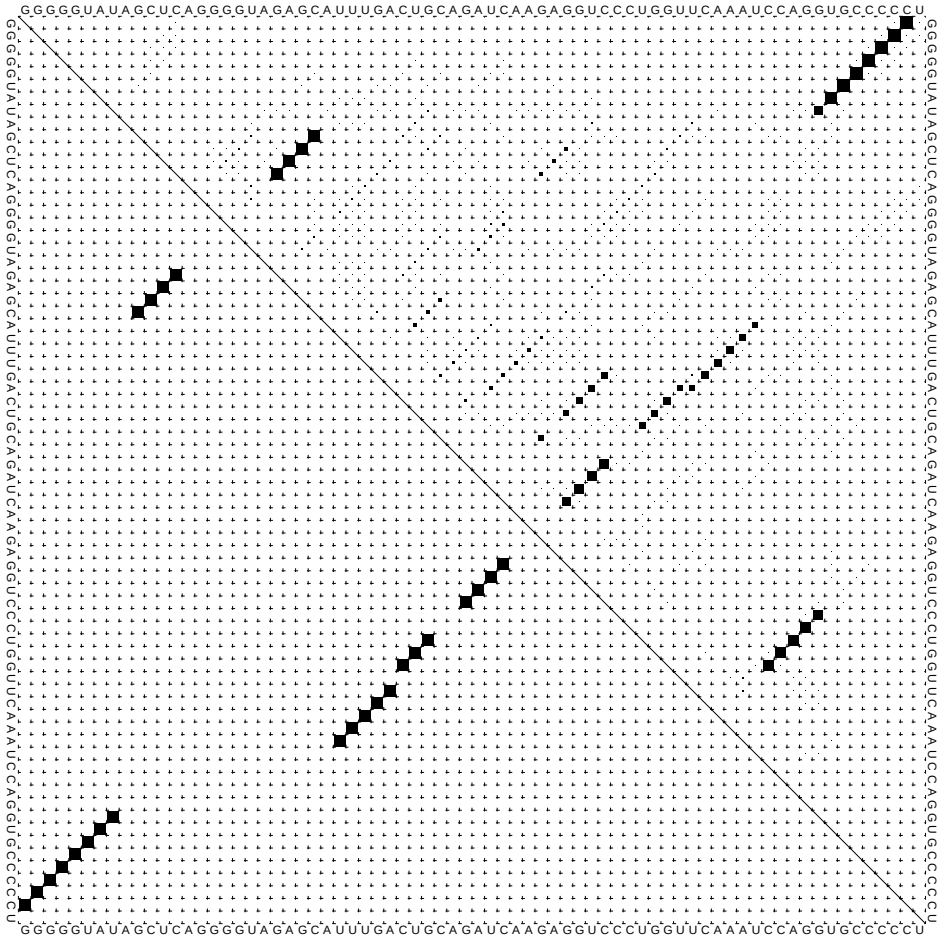
$$Q^{m1}_{ij} = 0 \quad (i \geq j - m)$$

$$Q^{m1}_{ij} = \sum_{i+m < k \leq j} Q^b_{ik} \cdot \exp(-\beta b) \cdot \exp(-\beta(j-k)c)$$

McCaskill Remarks

- Partition function of the ensemble of S in $Z = Q_{1n}$
- Correctness due to disjoint (=unambiguous) and independent decomposition
- Complexity $O(n^2)$ space,
  $O(n^3)$ time (after bounding size of interior loops)
- Probabilities
  - of a structure
    $\Pr[P|S] = Z^{-1} \exp(-\beta E(P))$ (efficient!)
  - of a structure set
    $\Pr[\mathcal{P}|S] = Z^{-1} \sum_{P \in \mathcal{P}} \exp(-\beta E(P))$ (depends on $|\mathcal{P}|$)
  - of a base pair (i, j)
    $\Pr[(i,j)|S] = Z^{-1} \sum_{P \ni (i,j)} \exp(-\beta E(P))$ (?)

Base Pair Probabilities



$$\Pr[(i,j)|S] := \sum_{P \ni (i,j)} \Pr[P|S]$$

McCaskill: Efficient Base Pair Probabilities

Idea: Compute $p_{kl} := \Pr[(k,l)|S]$ recursively (DP!),
recurse from long base pairs (outside) to small ones (inside)

1) simple case (external base pair)

Definition (Probability of external base pair)

$p^E_{kl} := \Pr[\mathcal{P}^E_{kl}]$,
where $\mathcal{P}^E_{kl} := \{P \mid P \in \mathcal{P}, (k,l)$ is external base pair in $P\}$,
where (k,l) is *external base pair in* P iff
$(k,l) \in P$ and $\nexists (i,j) \in P : i < k < l < j$.

$$Z_{\mathcal{P}^E_{kl}} = Q_{1k-1} Q^b_{kl} Q_{l+1n}$$

$$p^E_{kl} = \frac{Z_{\mathcal{P}^E_{kl}}}{Z} = \frac{Q_{1k-1} Q^b_{kl} Q_{l+1n}}{Q_{1n}}$$

McCaskill: Efficient Base Pair Probabilities

2) general case

a) (k, l) is external base pair

b) (k, l) limits stacking, bulge, or interior loop closed by (i, j)

c) (k, l) inner base pair of multiloop closed by (i, j)

McCaskill: Efficient Base Pair Probabilities

2) general case

a) (k, l) is external base pair ✓

b) (k, l) limits stacking, bulge, or interior loop closed by (i, j)

$$
\begin{aligned}
p_{kl}^{SBI}(i,j) &:= p_{ij} \Pr[\text{loop } i,j,k,l | (i,j)] \\
&= p_{ij} \frac{Z_{\{P \in \mathcal{P}_{ij} | P \text{ has loop } i,j,k,l\}}}{Z_{\mathcal{P}_{ij}^b}} \\
&= p_{ij} \frac{\exp(-\beta \, \text{eSBI}(i,j,k,l)) Q_{kl}^b}{Q_{ij}^b}.
\end{aligned}
$$

c) (k, l) inner base pair of multiloop closed by (i, j)

McC: Base Pair Probabilities — Multiloop Case

2) general case

c) (k, l) inner base pair of multiloop closed by (i, j)

$$p_{kl}^{M}(i,j) :=$$

$$p_{ij} \Pr[\text{multiloop with inner base pair } (k,l) \text{ closed by } (i,j) \mid (i,j)]$$

Three cases: position of (k, l) in the multiloop

(i) (k,l) leftmost base pair
(ii) (k,l) middle base pair
(iii) (k,l) rightmost base pair

McC: Base Pair Probabilities — Multiloop Case

2) general case

c) (k, l) inner base pair of multiloop closed by (i, j)

$$p^M_{kl}(i,j) :=$$

$$p_{ij} \Pr[\text{multiloop with inner base pair } (k,l) \text{ closed by } (i,j) \mid (i,j)]$$

Three cases: position of (k, l) in the multiloop

(i) (k,l) leftmost base pair
$Q^b_{kl} Q^m_{l+1j-1} \exp(-\beta(a + b + (k - i - 1)c))$

(ii) (k,l) middle base pair
$Q^m_{i+1k-1} Q^b_{kl} Q^m_{l+1j-1} \exp(-\beta(a + b))$

(iii) (k,l) rightmost base pair
$Q^m_{i+1k-1} Q^b_{kl} \exp(-\beta(a + b + (j - l - 1)c))$

McC — Multiloop Case (Ctd.)

Recall

$$p^M_{kl}(i,j) :=$$
$$p_{ij}\,\Pr[\text{multiloop with inner base pair } (k,l) \text{ closed by } (i,j) \mid (i,j)]$$

putting the three cases of (k,l) position together

$$\begin{aligned} p^M_{kl}(i,j) \;=\; p_{ij}[&Q^b_{kl} Q^m_{l+1j-1} \exp(-\beta(a+b+(k-i-1)c)) \\ &+ Q^m_{i+1k-1} Q^b_{kl} Q^m_{l+1j-1} \exp(-\beta(a+b)) \\ &+ Q^m_{i+1k-1} Q^b_{kl} \exp(-\beta(a+b+(j-l-1)c))]\, Q^{-1}_{ij} \end{aligned}$$

McCaskill — Base Pair Probabilities — Summary

$$p_{kl} = p_{kl}^{E} + \sum_{i<k,l<j} p_{kl}^{SBI}(i,j) + \sum_{i<k,l<j} p_{kl}^{M}(i,j)$$

Remarks

- Recursive formula for p_{kl} furnishes DP
- Efficient calculation of all p_{kl} in $O(n^4)$ time/$O(n^2)$ space
- Time reduction to $O(n^3)$ possible (not shown, but you learned the "trick")
- The algorithm by the p_{kl} recursion is an outside algorithm;
  in contrast the algo for computing Z and the Q_{ij} is inside.
  For getting the probabilities, we combined inside and outside.

Summary Part I

Algorithms

- Nussinov
- Zuker
- McCaskill

Common

- $O(n^3)$ time, $O(n^2)$ space
- non-crossing structure (= "no pseudoknots")

Differences

- realism: base pairs $\leftrightarrow$ free energy (loop-based)
- mfe $\leftrightarrow$ ensemble

Next?

Comparing RNAs