Efficient Computation of Base-Pairing Probabilities in Multi-Strand RNA Folding

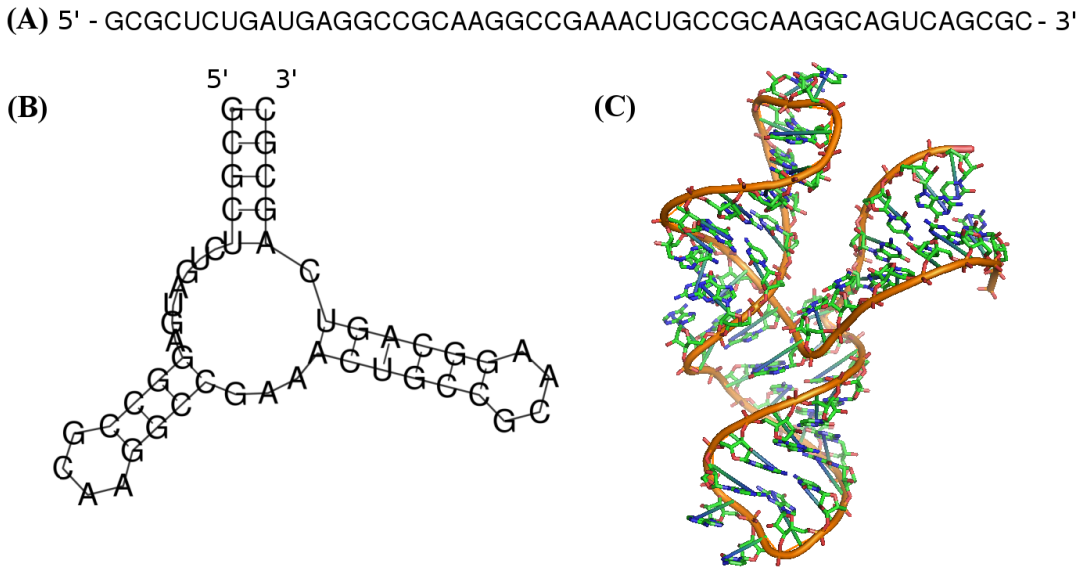
Ronny Lorenz
ronny@tbi.univie.ac.at

University of Vienna, Theoretical Biochemistry Group (TBI)

Valletta, Malta, February 24th 2020

RNA Secondary Structures

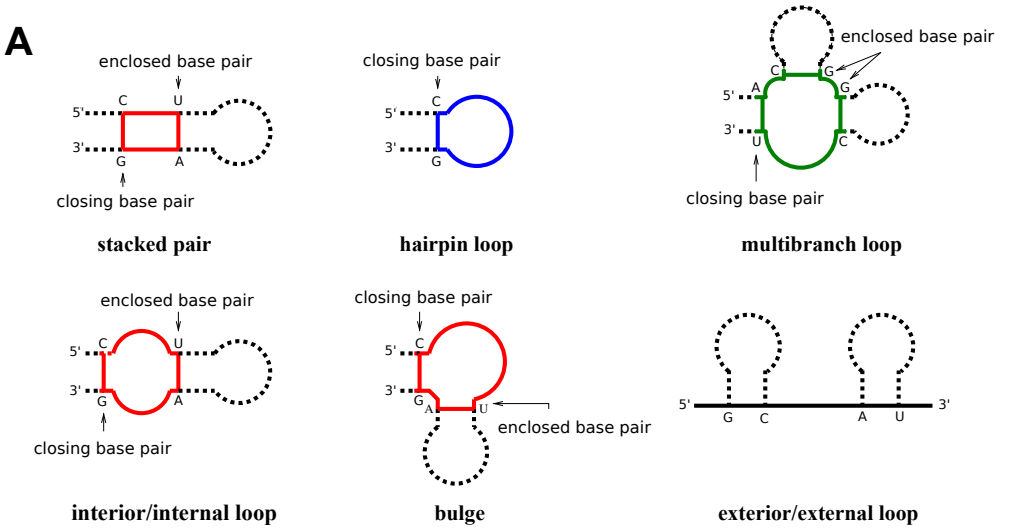
RNAs fold hierarchical (A) $\rightarrow$ (B) $\rightarrow$ (C)



Secondary Structure (B):

- Set of nested base pairs
- Captures majority of stabilizing interactions
- Many thermodynamic properties can be predicted efficiently
- Very good prediction accuracy for small RNAs
- Accuracy drops to 40%-70% for longer sequences

RNA Secondary Structures - Loops vs. Base Pairs



- secondary structures s can be uniquely decomposed into loops L
- stabilizing energy contributions (mostly) from stacked base pairs
- destabilizing contributions from unpaired bases in loops
- each loop L is assigned a free energy contribution E_L[1]

$$E(s) \approx \sum_{L \in s} E_L$$

[1] Turner, DH & Mathews, DH (2009). NNDB: The nearest neighbor parameter database for predicting stability of nucleic acid secondary structure., Nucleic Acids Research 38, D280-D282

RNA Secondary Structures - Statistical Thermodynamics

$$p(F) \propto e^{-\frac{E(F)}{RT}}$$

Most probable structure:

$$MFE = \min_{s} E(s)$$

Partition Function:

$$Q = \sum_{s} e^{-\frac{E(s)}{RT}}$$

Probability of a structure:

$$p(s) = \frac{e^{-E(s)/RT}}{Q}$$

Probability of Base Pair (k,l):

$$p_{k,l} = \frac{1}{Q} \sum_{s|(k,l)\in s} e^{-\frac{E(s)}{RT}}$$

RNA Secondary Structures - Prediction

Dynamic Programming (DP) algorithm[2]

[2]Nussinov, R & Pieczenik, G & Griggs, JR and Kleitman, DJ (1978). Algorithms for Loop Matchings., SIAM J. Appl. Math., 35(1), 68-82

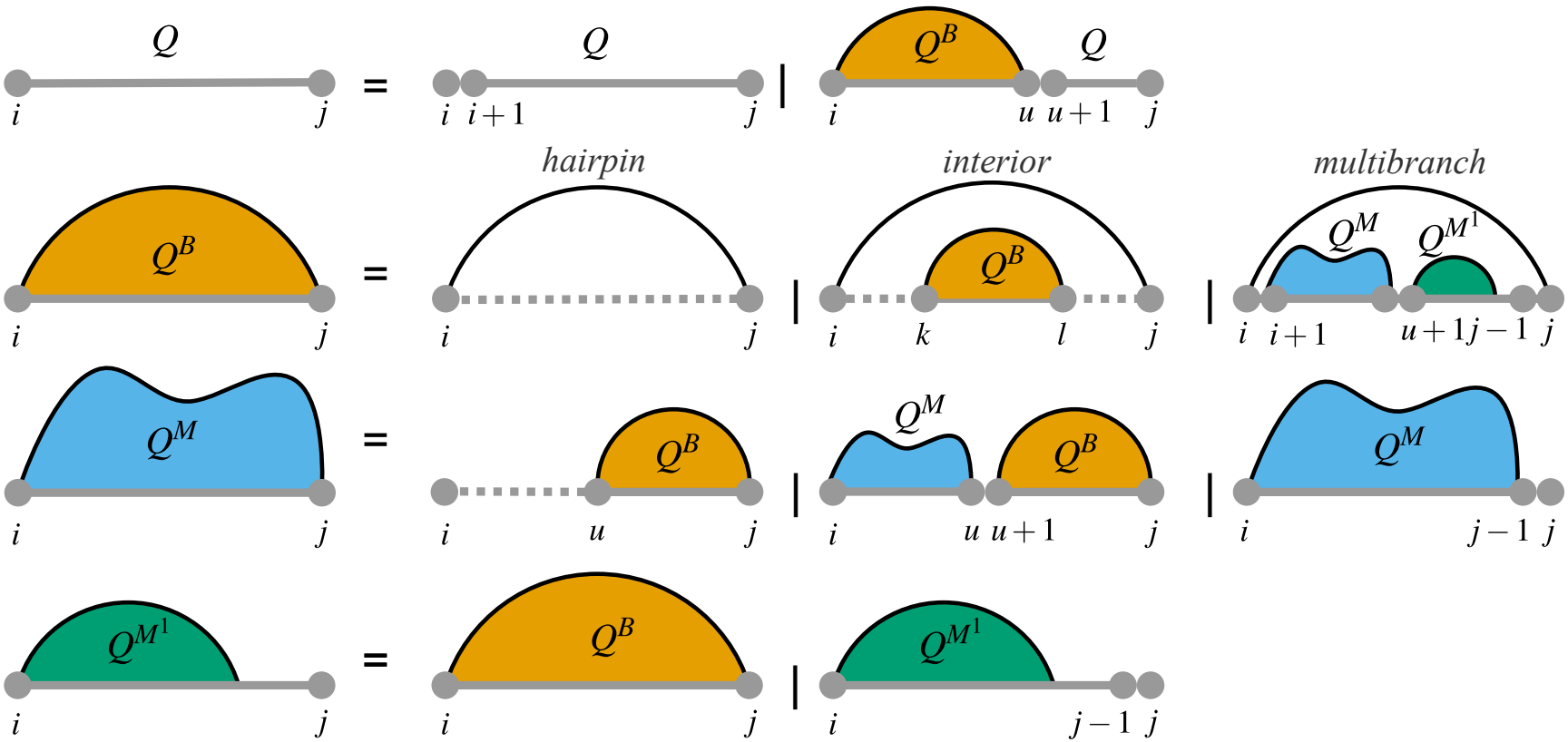
RNA Secondary Structures - Prediction

Dynamic Programming (DP) algorithm[2]

[2]Nussinov, R & Pieczenik, G & Griggs, JR and Kleitman, DJ (1978). Algorithms for Loop Matchings., SIAM J. Appl. Math., 35(1), 68-82

RNA Secondary Structures - Prediction

Dynamic Programming (DP) algorithm[2] [3]



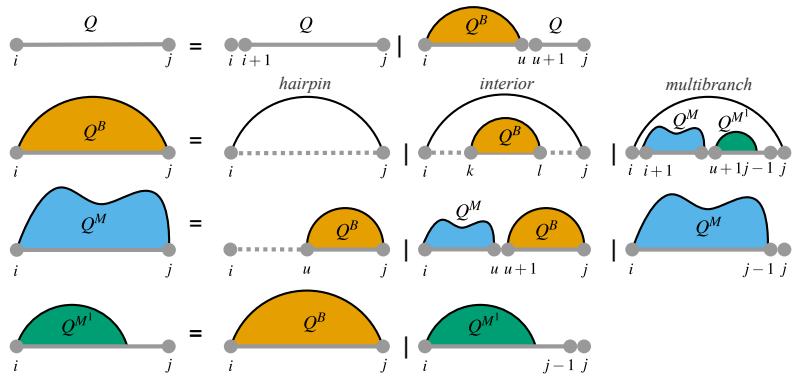
---

[2] Zuker M and Stiegler P (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information., Nucleic Acids Res, 9(1):133-148

[3] McCaskill JS (1990). The equilibrium partition function and base pair binding probabilities for RNA secondary structure., Biopolymers, 29(6-7):1105-1119

RNA Secondary Structures - Prediction

Dynamic Programming (DP) algorithm



$$
\begin{aligned}
Q_{i,j} &= Q_{i+1,j} + \sum_{i<u\leq j} Q^B_{i,u} Q_{u+1,j} \\
Q^B_{i,j} &= e^{-\frac{H(i,j)}{RT}} + \sum_{i<k<l<j} e^{-\frac{I(i,j,k,l)}{RT}} Q^B_{k,l} + e^{-\frac{a+b}{RT}} \sum_{i<u<j} Q^M_{i+1,u} Q^{M^1}_{u+1,j-1} \\
Q^M_{i,j} &= \sum_{i\leq u<j} e^{-\frac{(u-i)c+b}{RT}} Q^B_{u,j} + e^{-\frac{b}{RT}} \sum_{i<u<j} Q^M_{i,u} Q^B_{u+1,j} + e^{-\frac{c}{RT}} Q^M_{i,j-1} \\
Q^{M^1}_{i,j} &= e^{-\frac{b}{RT}} Q^B_{i,j} + e^{-\frac{c}{RT}} Q^{M^1}_{i,j-1}
\end{aligned}
$$

Asymptotic complexity: $O(n^3)$ time and $O(n^2)$ memory

Multiple Interacting Nucleic Acid Strands[2]

Straight-forward extension of single sequence case

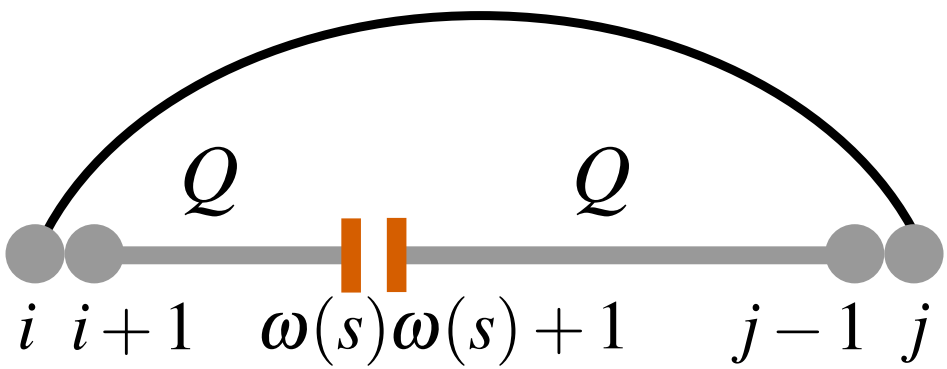
- consider complexes of N strands, i.e. one connected component
- restrict state space to intermolecular base pairs w/o crossings
- concatenate all strands ($n = n_1 + n_2 + \ldots + n_N$)
- prohibit strand nicks in “regular” loops
- treat cases with nicks as “external” loops w/ additional rule
- process all non-cyclic permutations π of strand concatenations
- correct for overcounting of symmetric cases

[2] Dirks, RM, Bois, JS, Schaeffer, JM, Winfree, E, and Pierce, NA (2007). Thermodynamic analysis of interacting nucleic acid strands., SIAM Rev., 49:65–88.

Multiple Interacting Nucleic Acid Strands[2]

Straight-forward extension of single sequence case

- consider complexes of N strands, i.e. one connected component
- restrict state space to intermolecular base pairs w/o crossings
- concatenate all strands ($n = n_1 + n_2 + \ldots + n_N$)
- prohibit strand nicks in “regular” loops
- treat cases with nicks as “external” loops w/ **additional rule**
- process all non-cyclic permutations π of strand concatenations
- correct for overcounting of symmetric cases

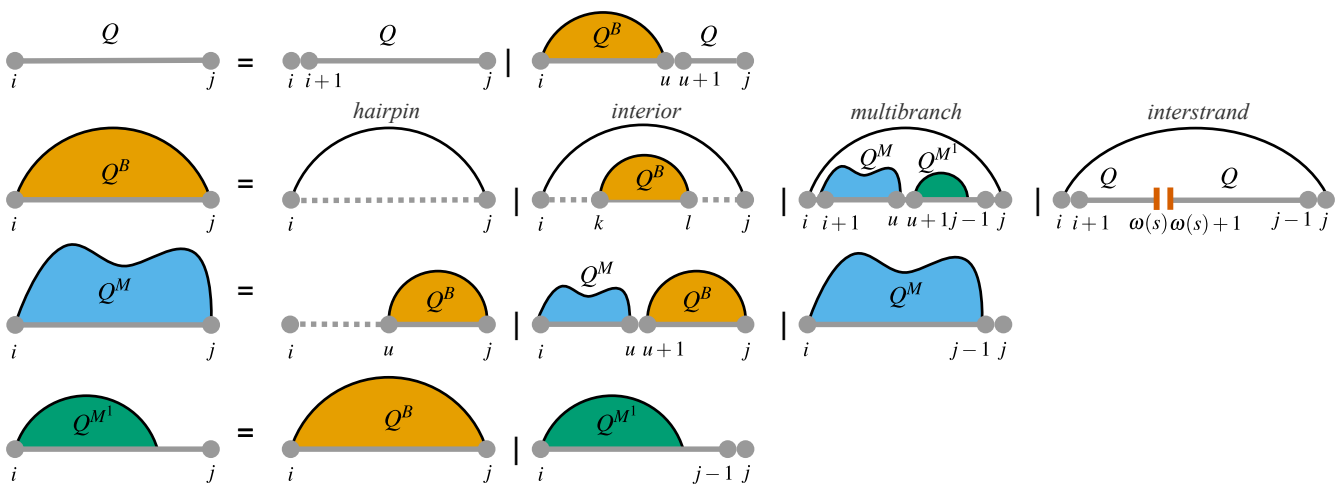


[2] Dirks, RM, Bois, JS, Schaeffer, JM, Winfree, E, and Pierce, NA (2007). Thermodynamic analysis of interacting nucleic acid strands., SIAM Rev., 49:65–88.

Multiple Interacting Nucleic Acid Strands[2]

Straight-forward extension of single sequence case

- consider complexes of N strands, i.e. one connected component
- restrict state space to intermolecular base pairs w/o crossings
- concatenate all strands ($n = n_1 + n_2 + \ldots + n_N$)
- prohibit strand nicks in “regular” loops
- treat cases with nicks as “external” loops w/ **additional rule**
- process all non-cyclic permutations π of strand concatenations
- correct for overcounting of symmetric cases



[2] Dirks, RM, Bois, JS, Schaeffer, JM, Winfree, E, and Pierce, NA (2007). Thermodynamic analysis of interacting nucleic acid strands., SIAM Rev., 49:65–88.

Interacting Nucleic Acid Strands - Base Pairing Probabilities

For *single* strand case (DP algorithm with $O(n^3)$ time, $O(n^2)$ memory):

$$
\begin{aligned}
p_{k,l} &= \frac{1}{Q} \sum_{s|(k,l)\in s} e^{-\frac{E(s)}{RT}} \\
&= \frac{1}{Q} Q^B_{k,l} \widehat{Q}_{k,l}, \quad \text{with} \\
\widehat{Q}_{k,l} &= \underbrace{\bar{Q}_{k,l}}_{\text{not enclosed by any bp}} + \underbrace{\breve{Q}_{k,l}}_{\text{enclosed by bp}}
\end{aligned}
$$

Interacting Nucleic Acid Strands - Base Pairing Probabilities

For *single* strand case (DP algorithm with $O(n^3)$ time, $O(n^2)$ memory):

$$
\begin{aligned}
p_{k,l} &= \frac{1}{Q} \sum_{s|(k,l)\in s} e^{-\frac{E(s)}{RT}} \\
&= \frac{1}{Q} Q^B_{k,l} \widehat{Q}_{k,l}, \quad \text{with} \\
\widehat{Q}_{k,l} &= \underbrace{\bar{Q}_{k,l}}_{\text{not enclosed by any bp}} + \underbrace{\breve{Q}_{k,l}}_{\text{enclosed by bp}}
\end{aligned}
$$

For complexes of N strands:

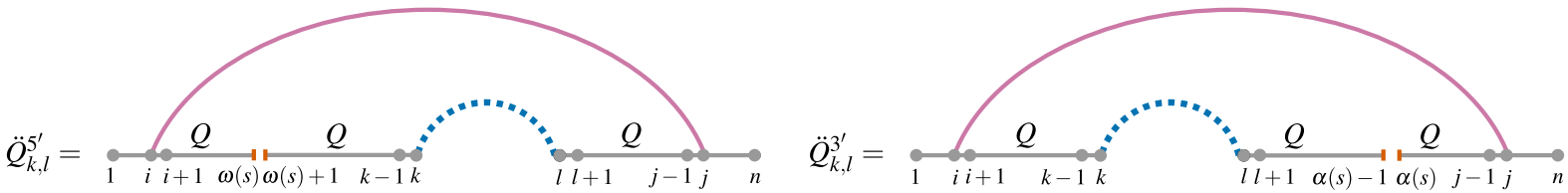
$$
\begin{aligned}
p_{k,l} &= \sum_{\pi} w(\pi) p_{k,l}[\pi] \\
&= \frac{1}{Q} \sum_{\pi} \widehat{Q}_{k,l}[\pi] Q^B_{k,l}[\pi], \quad \text{with} \\
\widehat{Q}_{k,l}[\pi] &= \underbrace{\bar{Q}_{k,l}[\pi]}_{\text{not enclosed by any bp}} + \underbrace{\breve{Q}_{k,l}[\pi]}_{\text{enclosed by bp}} + \underbrace{\ddot{Q}_{k,l}[\pi]}_{\text{enclosed by bp w/ nick in loop}}
\end{aligned}
$$

What is the asymptotic complexity to compute $\ddot{Q}_{k,l}[\pi]$?

Interacting Nucleic Acid Strands - Base Pairing Probabilities

Additional case for nicked loops (for particular π):

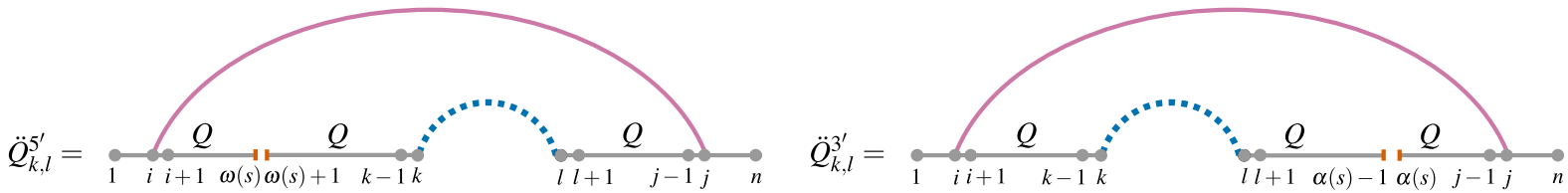
$$\ddot{Q}_{k,l} = \ddot{Q}^{5'}_{k,l} + \ddot{Q}^{3'}_{k,l}$$



Interacting Nucleic Acid Strands - Base Pairing Probabilities

Additional case for nicked loops (for particular π):

$$\ddot{Q}_{k,l} = \ddot{Q}^{5'}_{k,l} + \ddot{Q}^{3'}_{k,l}$$

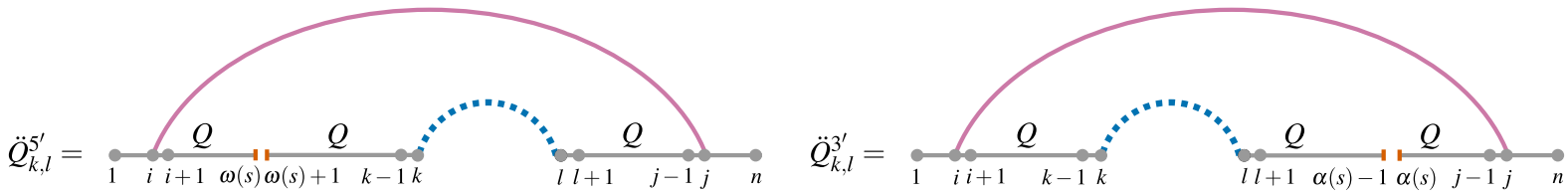


Computing all $\ddot{Q}_{k,l}$ by considering all enclosing pairs (i,j) and $N-1$ strand nicks seems to require $O(n^4 N)$ operations

Interacting Nucleic Acid Strands - Base Pairing Probabilities

Additional case for nicked loops (for particular π):

$$\ddot{Q}_{k,l} = \ddot{Q}^{5'}_{k,l} + \ddot{Q}^{3'}_{k,l}$$



Computing all $\ddot{Q}_{k,l}$ by considering all enclosing pairs (i,j) and $N-1$ strand nicks seems to require $O(n^4 N)$ operations

- Dirks et al., 2007:

  "… equilibrium probability of each intrastrand and interstrand base pair … can be calculated by backtracking through the partition function algorithm … applying a particular algorithmic transformation at each step"

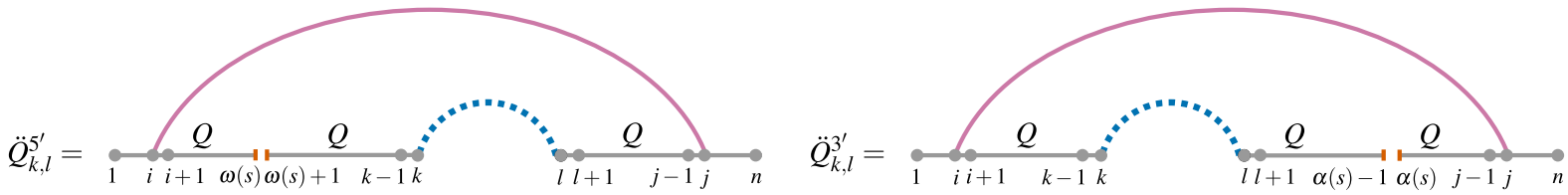
- Wolfe et al., 2017:

  "… the equilibrium base-pairing properties … must be calculated for each complex $j \in \Psi$ *using* $\Theta(|\phi_j|^3)$ *dynamic programs."* (here, $|\phi_j| \equiv n$)

Interacting Nucleic Acid Strands - Base Pairing Probabilities

Additional case for nicked loops (for particular π):

$$\ddot{Q}_{k,l} = \ddot{Q}^{5'}_{k,l} + \ddot{Q}^{3'}_{k,l}$$



Computing all $\ddot{Q}_{k,l}$ by considering all enclosing pairs (i,j) and $N-1$ strand nicks seems to require $O(n^4 N)$ operations

- Dirks et al., 2007:

  "… equilibrium probability of each intrastrand and interstrand base pair … can be calculated by backtracking through the partition function algorithm … applying a particular algorithmic transformation at each step"

- Wolfe et al., 2017:

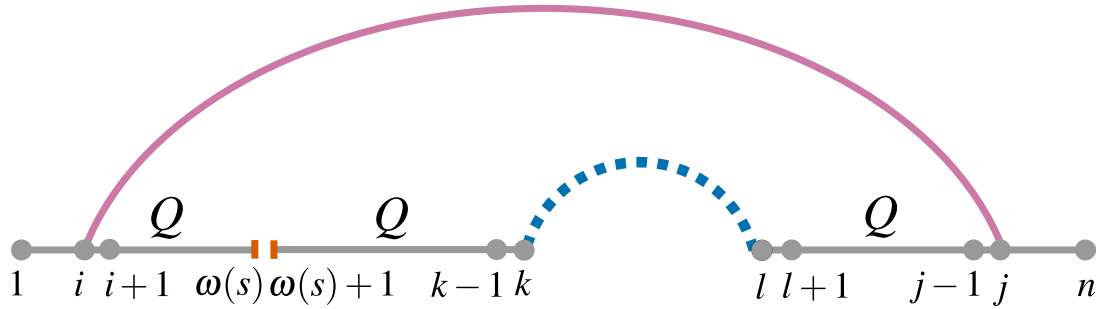
  "… the equilibrium base-pairing properties … must be calculated for each complex $j \in \Psi$ *using* $\Theta(|\phi_j|^3)$ *dynamic programs."* (here, $|\phi_j| \equiv n$)

Still, no reference to the algorithm, so how to achive that?

Interacting Nucleic Acid Strands - Base Pairing Probabilities

case 0: strand nick is on 5' side

$$\ddot{Q}^{5'}_{k,l} =$$



$$\ddot{Q}^{5'}_{k,l} = \sum_{\substack{1\le i<k \\ l<j\le n}} \widehat{Q}_{i,j} Q_{l+1,j-1} \times \sum_{s|i\le\omega(s)<k} Q_{i+1,\omega(s)} Q_{\omega(s)+1,k-1}$$

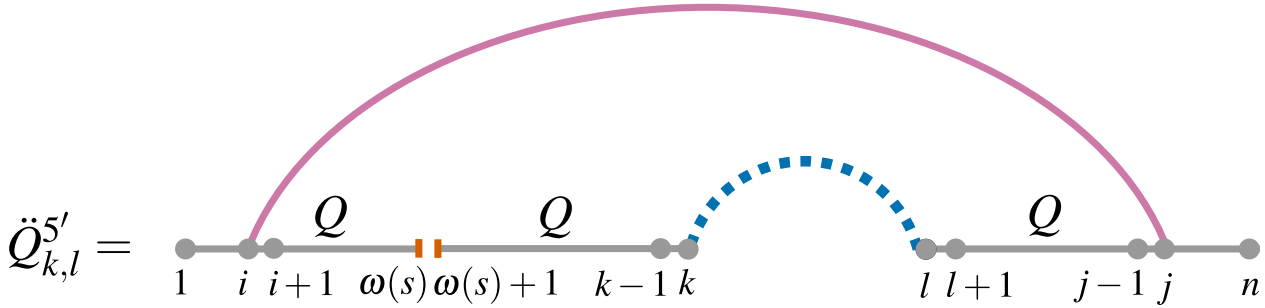
Apply Dynamic Programming paradigm:

- Trade computation time against memory consumption[3]
- Extract parts that are computed redundantly for different $\ddot{Q}^{5'}_{k,l}$

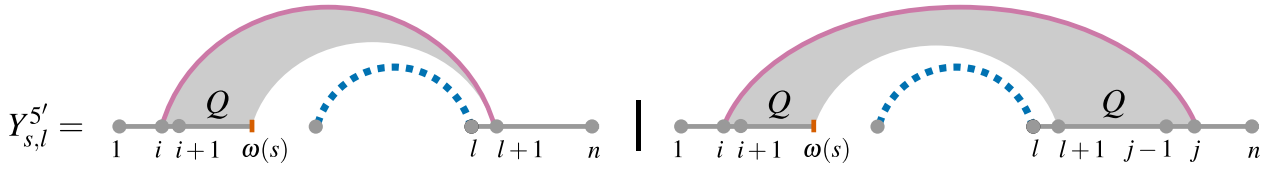
[3] Similar to McCaskill, JS (1990). The equilibrium partition function and base pair binding probabilities for RNA secondary structure., Biopolymers, 29:1105-1119.

Interacting Nucleic Acid Strands - Base Pairing Probabilities

case 0: strand nick is on 5' side



1st step (fixed l):

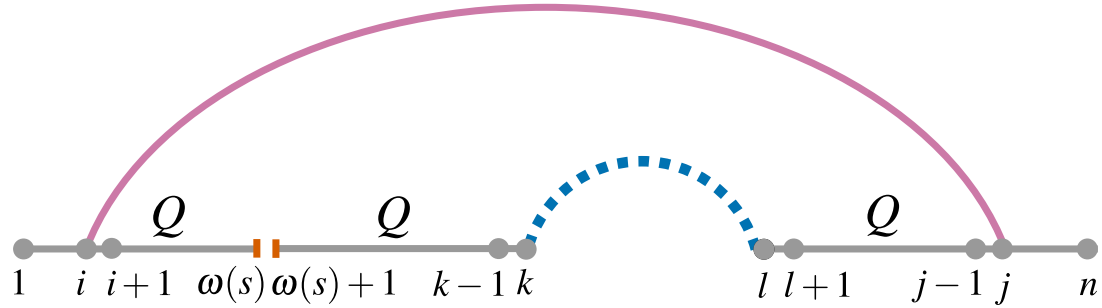


- pre-compute "enclosed" part $Y^{5'}_{s,l}$ up to $\omega(s)$
- re-use $Y^{5'}_{s,l}$ for all k

Interacting Nucleic Acid Strands - Base Pairing Probabilities

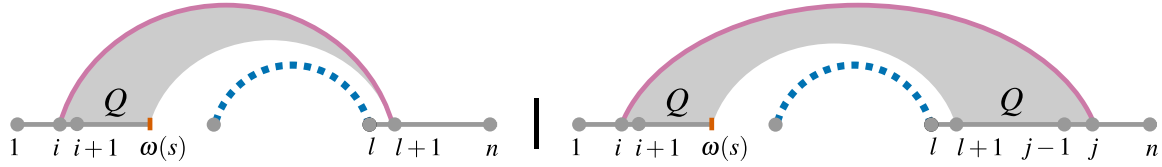
case 0: strand nick is on 5' side

$\ddot{Q}^{5'}_{k,l} =$



1st step (fixed l):

$Y^{5'}_{s,l} =$

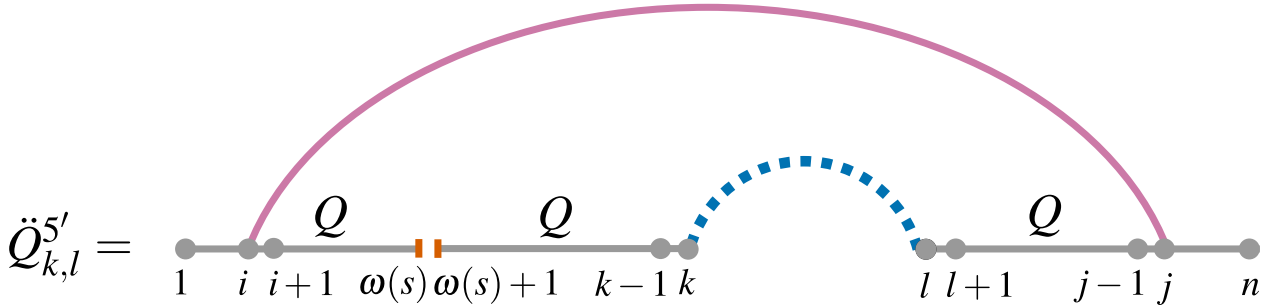


$$\ddot{Q}^{5'}_{k,l} = Y^{5'}_{\sigma(k-1),l} + \sum_{s|\omega(s)<k} Q_{\omega(s)+1,k-1} Y^{5'}_{s,l} \quad \text{(indep. of } i \text{ and } j\text{)}$$

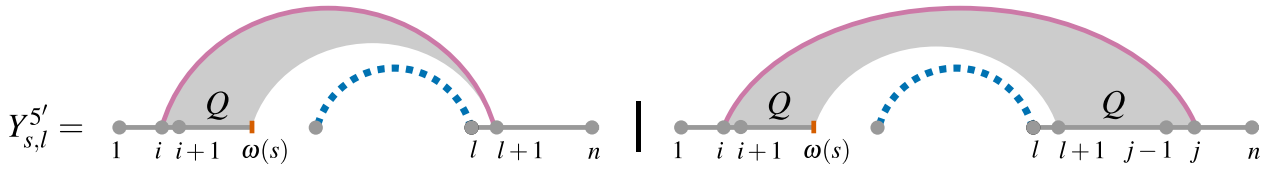
$$Y^{5'}_{s,l} = \sum_{j>l} Q_{l+1,j-1} \times \left(\widehat{Q}_{\omega(s),j} + \sum_{i<\omega(s)} \widehat{Q}_{i,j} Q_{i+1,\omega(s)} \right) \quad \text{(indep. of } k\text{)}$$

Interacting Nucleic Acid Strands - Base Pairing Probabilities

case 0: strand nick is on 5’ side



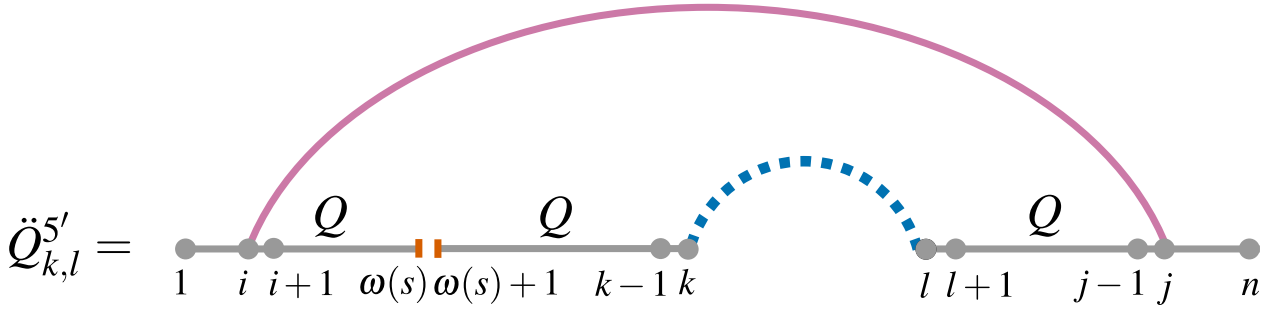
1^{st} step (fixed l):



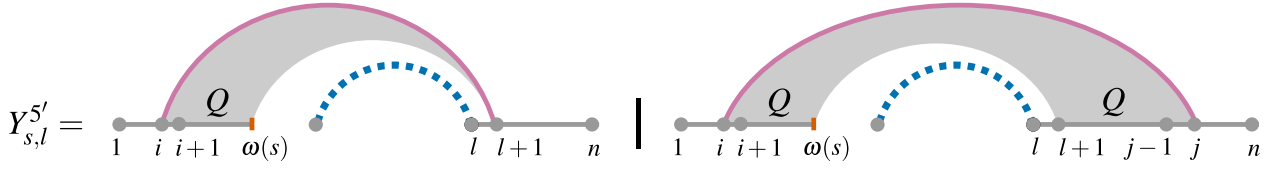
Can we do better than $O(n^3N)$?

Interacting Nucleic Acid Strands - Base Pairing Probabilities

case 0: strand nick is on 5' side



1st step (fixed l):



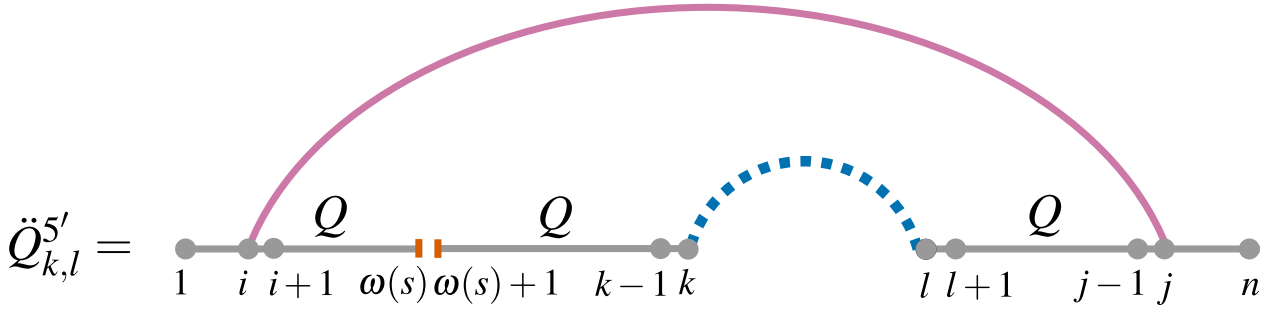
Can we do better than $O(n^3 N)$?

Observations when comparing $Y^{5'}_{s,l-1}$ against $Y^{5'}_{s,l}$:

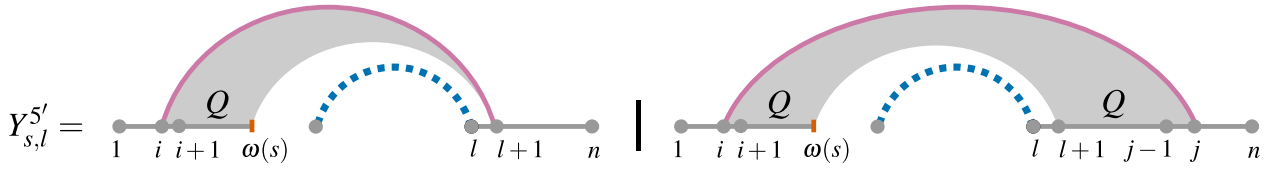
- "left" contribution stays the same
- "right" contribution includes $Q_{l,j}$ instead of $Q_{l+1,j}$
- one more j to account for $(j = l)$

Interacting Nucleic Acid Strands - Base Pairing Probabilities

case 0: strand nick is on 5' side



1st step (fixed l):



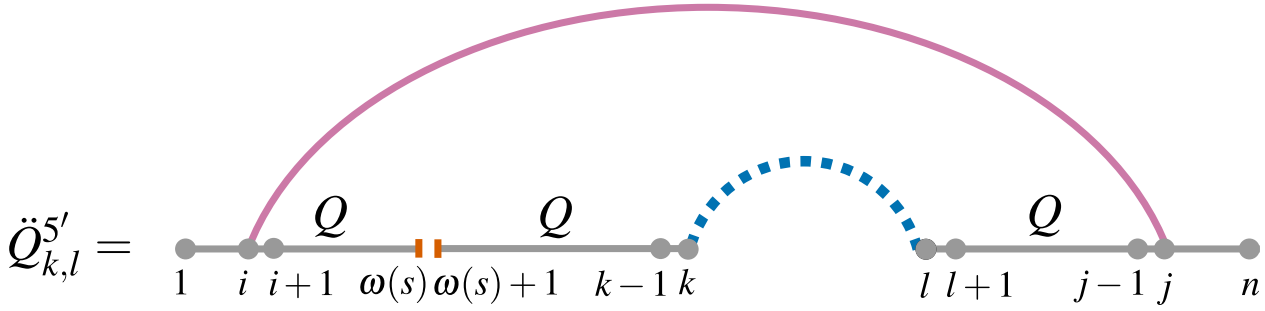
Can we do better than $O(n^3 N)$?

Observations when comparing $Y^{5'}_{s,l-1}$ against $Y^{5'}_{s,l}$:

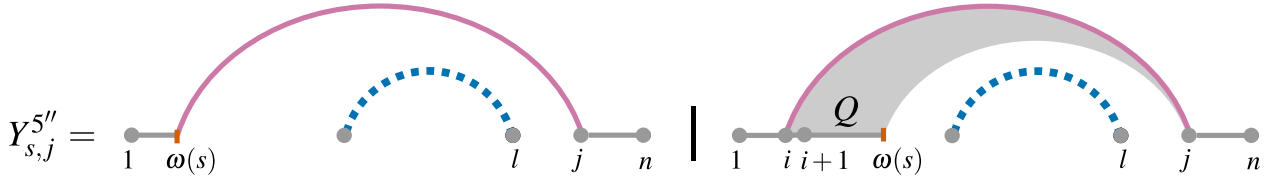
- "left" contribution stays the same **(pre-compute and re-use!)**
- "right" contribution includes $Q_{l,j}$ instead of $Q_{l+1,j}$
- one more j to account for $(j = l)$

Interacting Nucleic Acid Strands - Base Pairing Probabilities

case 0: strand nick is on 5' side



2^{nd} step (pre-compute "left" part of $Y^{5'}_{s,l}$):

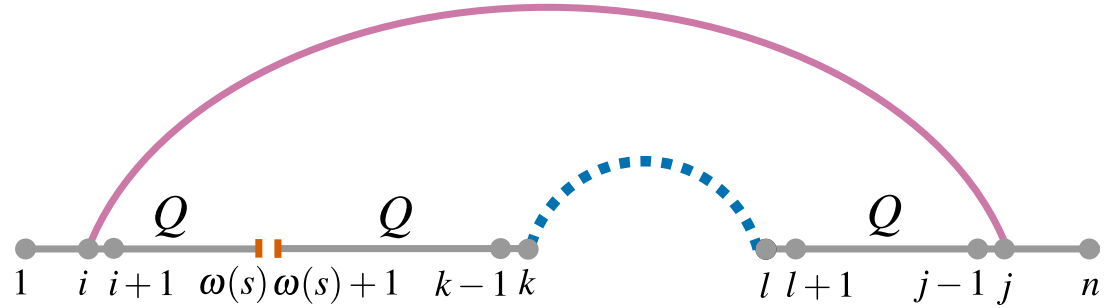


- "left" part ($Y^{5''}_{s,j}$) delimited by $\omega(s)$ and j is independent of k and l
- re-use $Y^{5''}_{s,j}$ to compute $Y^{5'}_{s,l}$ for all l

Interacting Nucleic Acid Strands - Base Pairing Probabilities

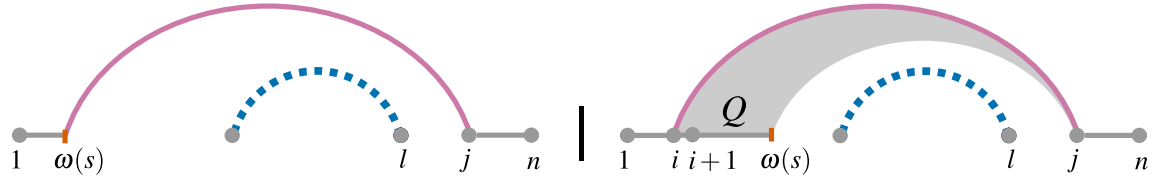
case 0: strand nick is on 5' side

$\ddot{Q}^{5'}_{k,l} =$



2nd step (pre-compute "left" part of $Y^{5'}_{s,l}$):

$Y^{5''}_{s,j} =$



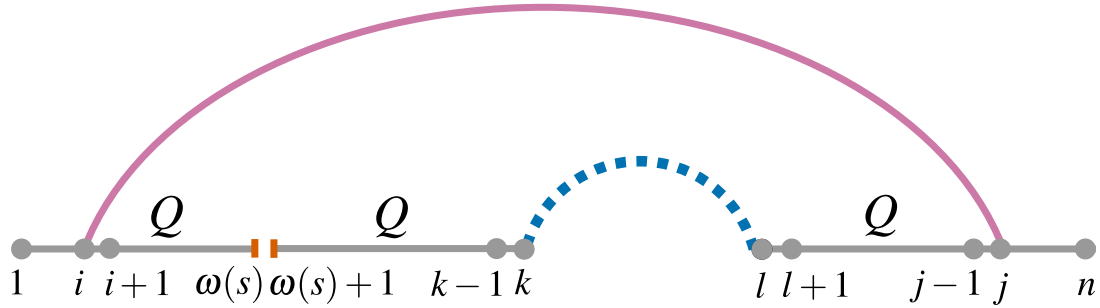
$$Y^{5'}_{s,l} = Y^{5''}_{s,l+1} + \sum_{j>l+1} Q_{l+1,j-1} \cdot Y^{5''}_{s,j} \quad \text{(indep. of } i \text{ and } k\text{)}$$

$$Y^{5''}_{s,j} = \widehat{Q}_{\omega(s),j} + \sum_{i<\omega(s)} \widehat{Q}_{i,j} \cdot Q_{i+1,\omega(s)} \quad \text{(indep. of } k \text{ and } l\text{)}$$

Interacting Nucleic Acid Strands - Base Pairing Probabilities

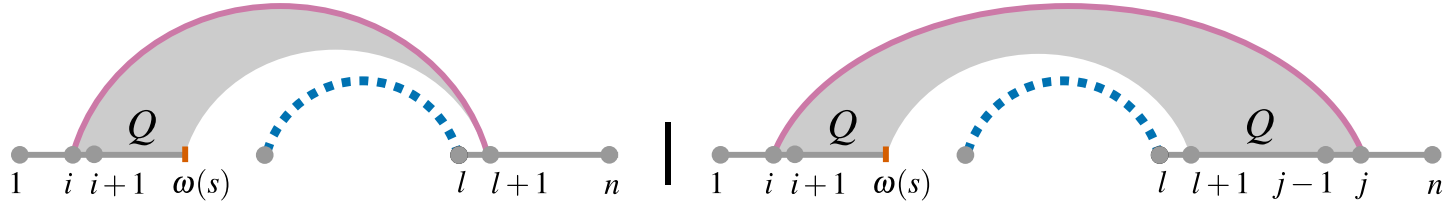
case 0: strand nick is on 5' side

$$\ddot{Q}^{5'}_{k,l} =$$

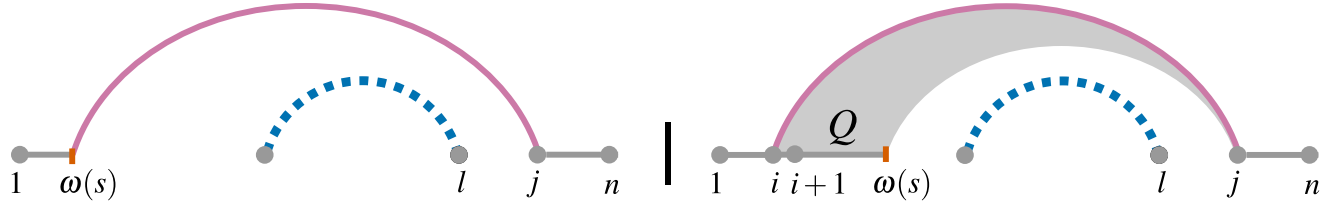


Finally:

$$Y^{5'}_{s,l} =$$



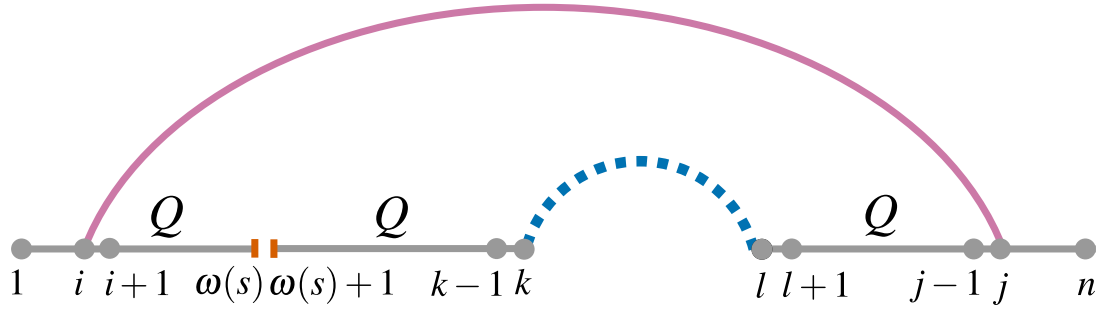
$$Y^{5''}_{s,j} =$$



Complexity: $O(n^2N)$ time and additional $O(nN)$ memory

Interacting Nucleic Acid Strands - Base Pairing Probabilities

case 0: strand nick is on 5' side

$\ddot{Q}^{5'}_{k,l} =$ 

Finally:

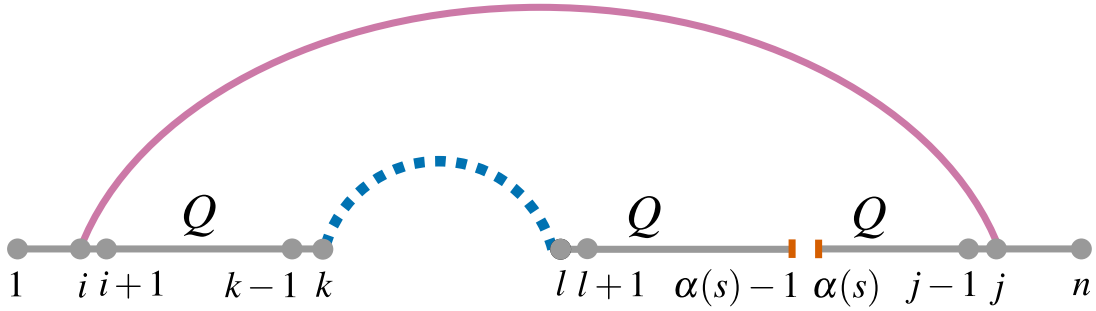
$$
\begin{aligned}
\ddot{Q}^{5'}_{k,l} &= Y^{5'}_{\sigma(k-1),l} + \sum_{s|\omega(s)<k} Q_{\omega(s)+1,k-1} Y^{5'}_{s,l} \\
Y^{5'}_{s,l} &= Y^{5''}_{s,l+1} + \sum_{j>l+1} Q_{l+1,j-1} \cdot Y^{5''}_{s,j} \\
Y^{5''}_{s,j} &= \widehat{Q}_{\omega(s),j} + \sum_{i<\omega(s)} \widehat{Q}_{i,j} \cdot Q_{i+1,\omega(s)}
\end{aligned}
$$

Complexity: $O(n^2 N)$ time and additional $O(nN)$ memory

Interacting Nucleic Acid Strands - Base Pairing Probabilities

case 1: strand nick is on 3' side

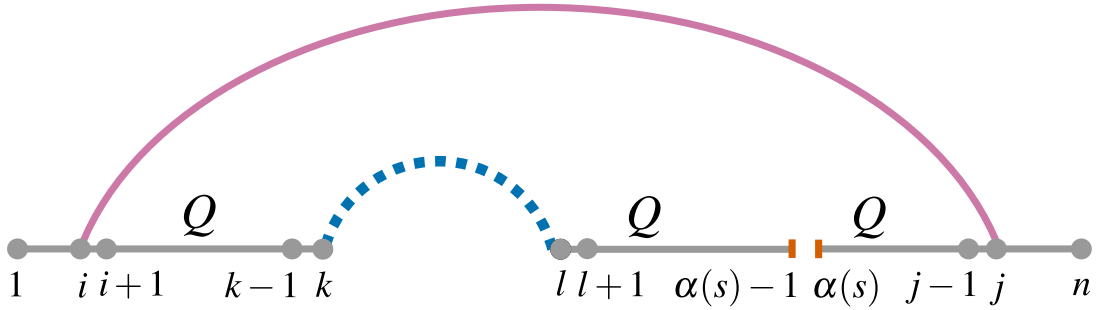
$$\ddot{Q}^{3'}_{k,l} =$$



Interacting Nucleic Acid Strands - Base Pairing Probabilities

case 1: strand nick is on 3' side

$\ddot{Q}^{3'}_{k,l} =$

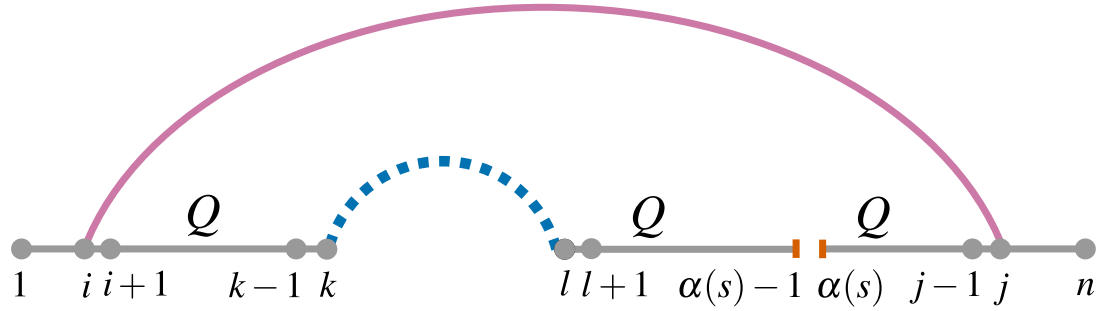


Using similar algorithmic transformations as for $\ddot{Q}^{5'}_{k,l}$

Interacting Nucleic Acid Strands - Base Pairing Probabilities

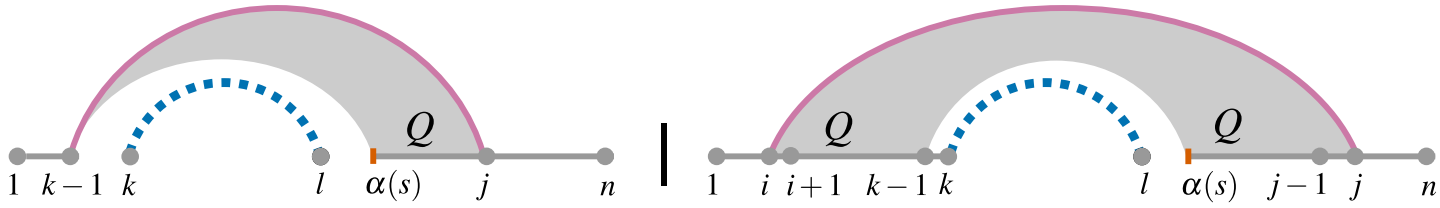
case 1: strand nick is on 3' side

$\ddot{Q}^{3'}_{k,l} =$

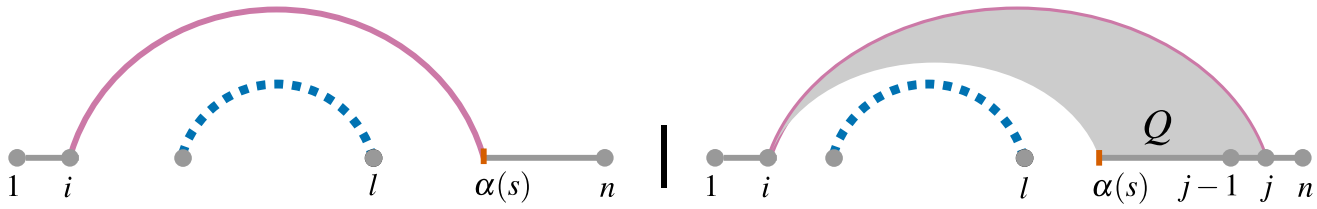


Using similar algorithmic transformations as for $\ddot{Q}^{5'}_{k,l}$

$Y^{3''}_{s,k} =$



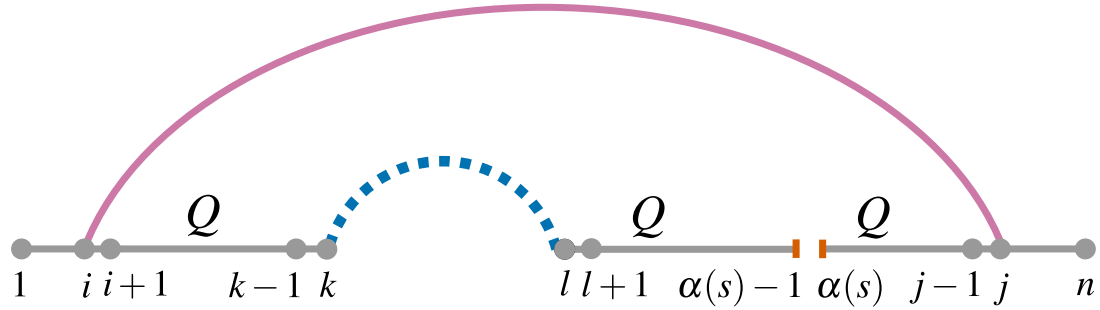
$Y^{3'}_{s,i} =$



Interacting Nucleic Acid Strands - Base Pairing Probabilities

case 1: strand nick is on 3' side

$\ddot{Q}^{3'}_{k,l} =$



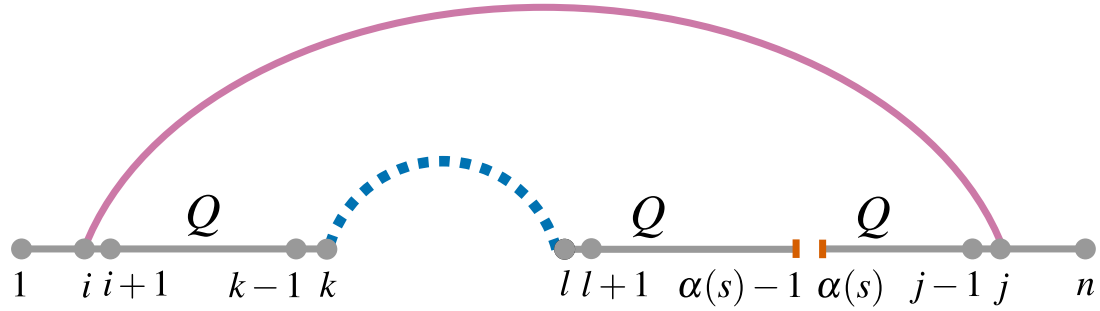
Using similar algorithmic transformations as for $\ddot{Q}^{5'}_{k,l}$

$$
\begin{aligned}
\ddot{Q}^{3'}_{k,l} &= Y^{3''}_{\sigma(l+1),k} + \sum_{s|\alpha(s)>l+1} Q_{l+1,\alpha(s)-1} Y^{3''}_{s,k} \\
Y^{3''}_{s,k} &= \sum_{i<k} Q_{i+1,k-1} Y^{3'}_{s,i} \\
Y^{3'}_{s,i} &= \widehat{Q}_{i,\alpha(s)} + \sum_{j>\alpha(s)} \widehat{Q}_{i,j} Q_{\alpha(s),j-1}
\end{aligned}
$$

Interacting Nucleic Acid Strands - Base Pairing Probabilities

case 1: strand nick is on 3' side

$\ddot{Q}^{3'}_{k,l} =$

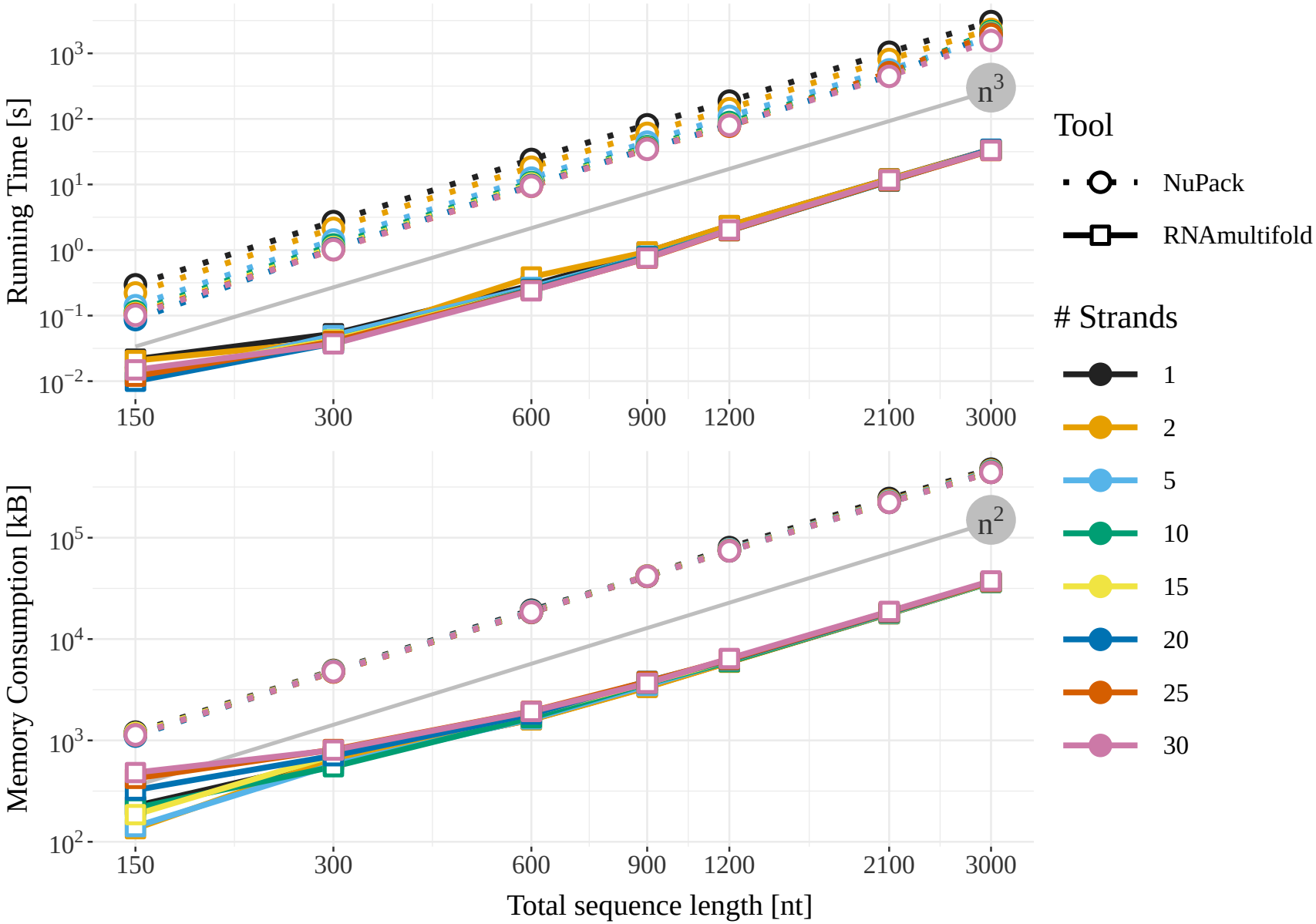


Using similar algorithmic transformations as for $\ddot{Q}^{5'}_{k,l}$

$$
\begin{aligned}
\ddot{Q}^{3'}_{k,l} &= Y^{3''}_{\sigma(l+1),k} + \sum_{s|\alpha(s)>l+1} Q_{l+1,\alpha(s)-1} Y^{3''}_{s,k} \\
Y^{3''}_{s,k} &= \sum_{i<k} Q_{i+1,k-1} Y^{3'}_{s,i} \\
Y^{3'}_{s,i} &= \widehat{Q}_{i,\alpha(s)} + \sum_{j>\alpha(s)} \widehat{Q}_{i,j} Q_{\alpha(s),j-1}
\end{aligned}
$$

Total effort for all $\ddot{Q}_{k,l}$: $O(n^2 N)$ time using additional $O(nN)$ memory

Runtime and Memory Consumption



Conclusion

- overhead of “nicked” loops is negligible
- all $p_{k,l}$ can indeed be computed in $O(|\Pi|n^3)$ time
- implementation available as `RNAmultifold`[4]
- $50-65\times$ faster, $7\times$ less memory than `NUPACK` 3.2.2 (per π)
- full constraints support, e.g. to restrict state space or include experimental probing data (SHAPE, etc.)
- intramolecular G-Quadruplex

Outlook

- automatically compute over all π and complexs of size N
- add MFE, Boltzmann sampling, and suboptimal enumeration
- include ligand binding support, e.g. SSB proteins
- add concentration dependency
- re-use (parts of) DP tables for different π

[4] `ViennaRNA Package 2.5.0alpha`

Thanks to

- **Christoph Flamm**
- **Ivo Hofacker**
- **Peter Stadler**

Thank You for your attention!

This work was funded in parts by the German Federal Ministry of Education and Research (BMBF, project no. 031A538A, de.NBI-RBC, to PFS and project no. 031L0164C, RNAProNet), and the Austrian science fund FWF (project no. I 2874 “Prediction of RNA-RNA interactions”, project no. F 43 “RNA regulation of the transcriptome”).