10 February 2025 RNA Folding

Plan.

* RNA Folding Problem
* Announcements
* DP for RNA Folding.

Another problem from biology

RNA Sequences $m \in \{A, U, C, G\}^*$

$m =$ A U U C G G A C G G A A

Key Question.

What structure does the RNA molecule adopt?

What structure does the RNA molecule adopt?

↳ Base pairing

A — U

C — G

} allowable pairings.

What structure does the RNA molecule adopt?

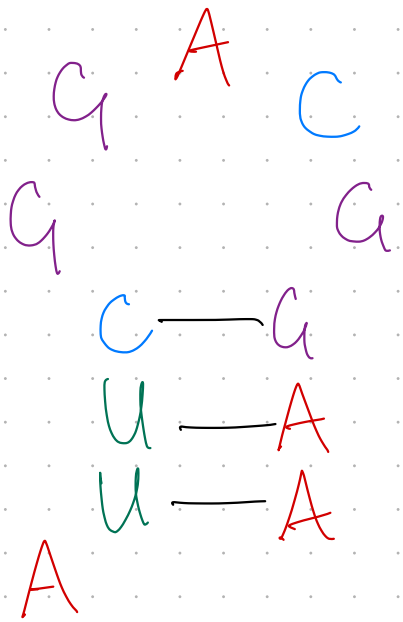
m = A U U C G G A C G G A A

↓ base pairing

What structure does the RNA molecule adopt?

m = A U U C G G A C G G A A

↓ base pairing



Given. RNA sequence $m \in \{A, U, C, G\}^*$

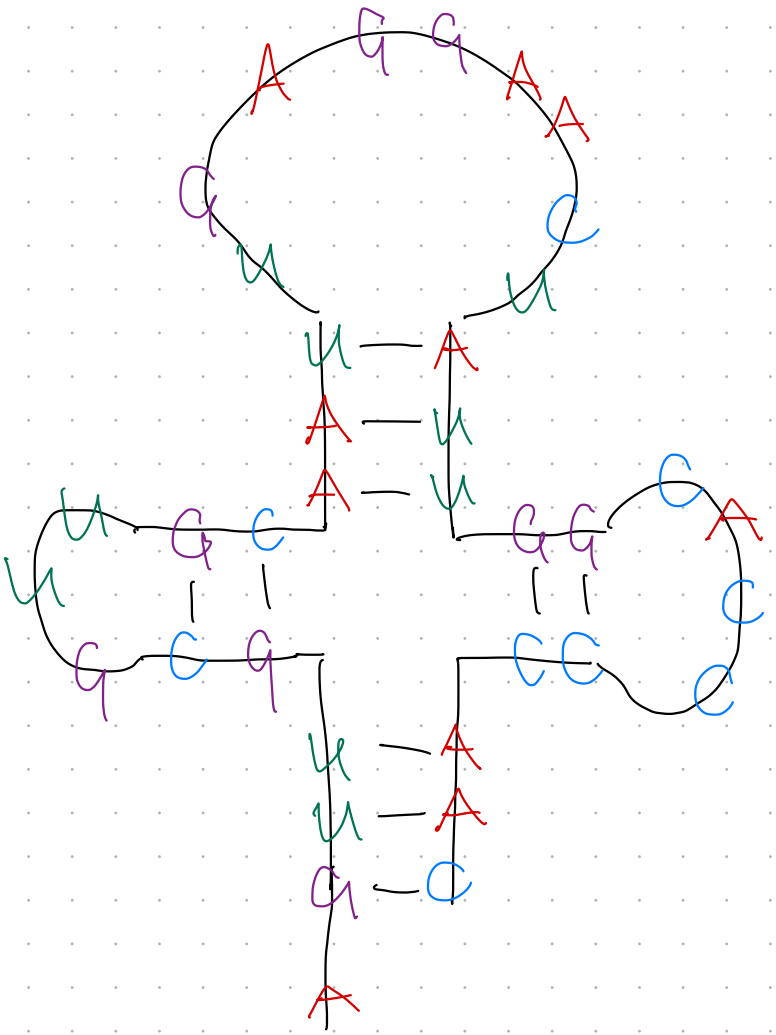
Find. Minimum "energy" structure

AGUUGCGUUGCAAUUGAGGAACUAUUGGCACCCCAAC

Given. RNA sequence $m \in \{A, U, C, G\}^*$

Find. Minimum "energy" structure

AGUUGCGUUGCAAUUGAGGAACUAUUGGCACCCCAAC



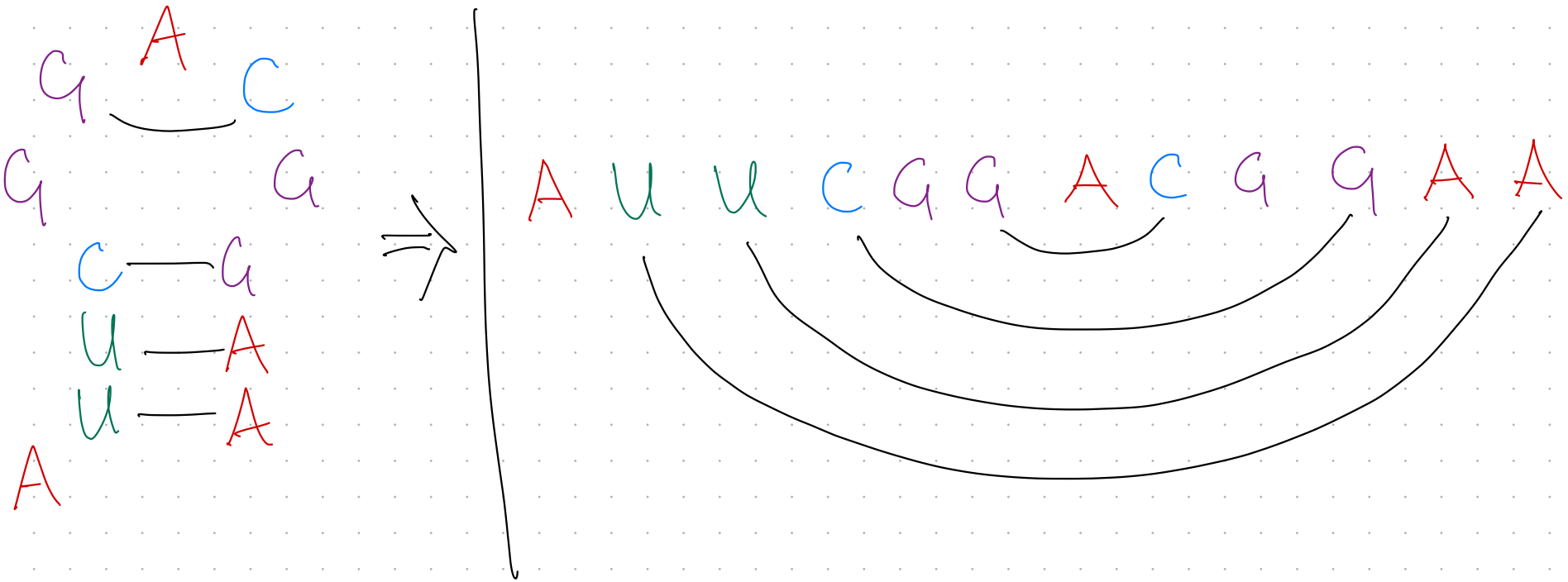
Min "energy"

≡

Max "legal" base pairing

Given. RNA sequence $m \in \{A, U, C, G\}^*$

Compute. Maximum Non-crossing matching E of preferred base pairs within m.

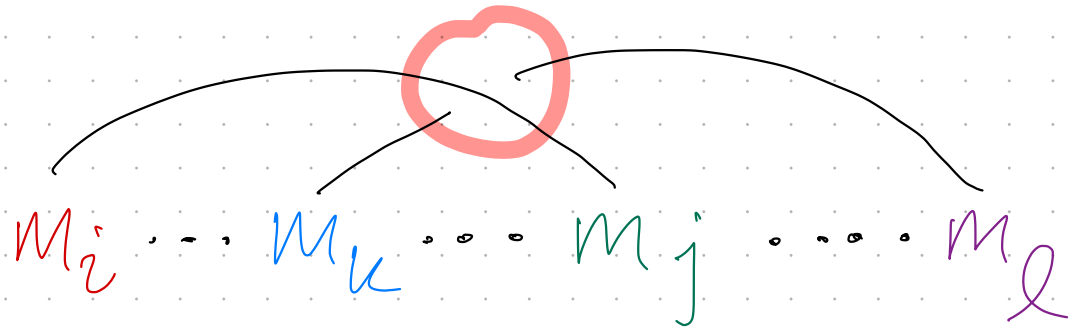


Given. RNA sequence $m \in \{A, U, C, G\}^*$

Compute. Maximum Non-crossing matching E of preferred base pairs within m.

Non-crossing. if $(i, j), (k, \ell) \in E$

then, $\neg(i < k < j < \ell)$.



Announcements

* HW 2 ongoing
* HW 1 grades released after lecture.
* Prelim 1 This Thursday 7:30-9 p.
  * Statler 185
  * Statler 196

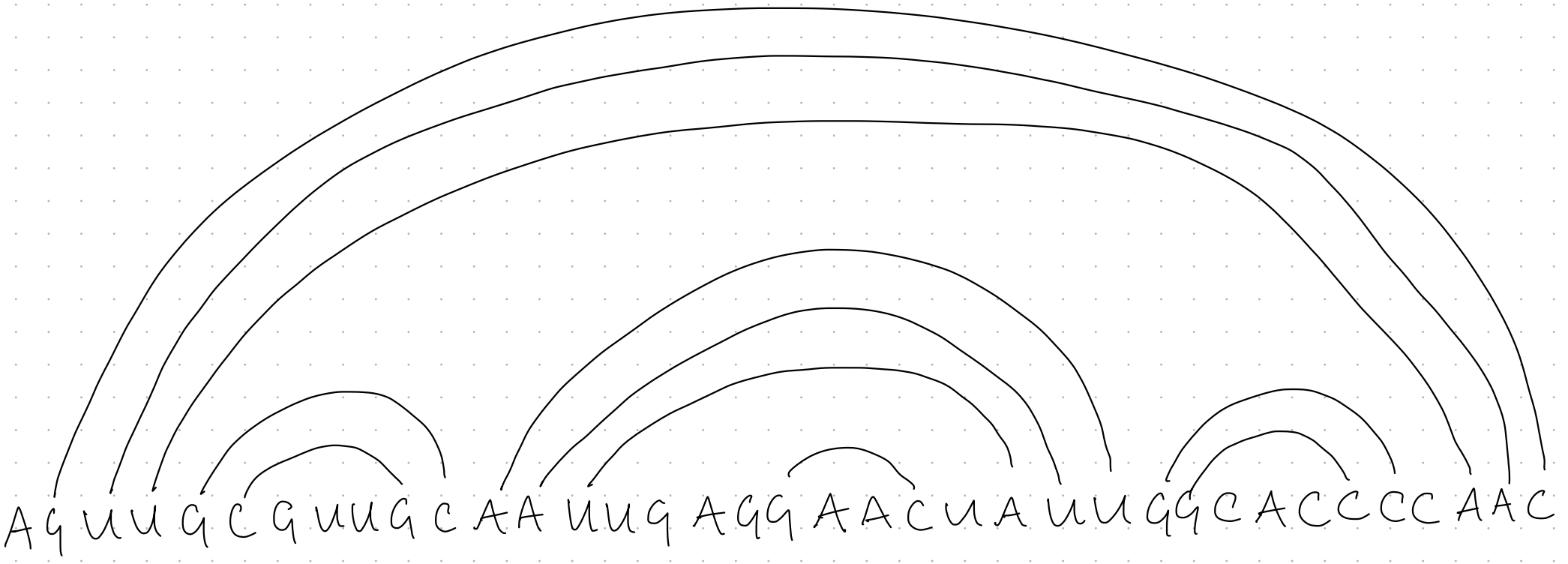
  } See Ed for room assignment.

  ↳ Coverage through last Friday's lecture.
* Review
  * Practice Exam on Canvas.
  * Review Session Tues 7-9 p
    ↳ Gates G01.

Given. RNA sequence $m \in \{A, U, C, G\}^*$

Compute. Maximum Non-crossing matching E of preferred base pairs within m.

$BP = \{(A,U), (C,G)\}$



What are the cases / subproblems?

Cases In optimal folding of M.

*

*

M = UUGCAAUUGAGGAACUAUUGGCACCCCAAC

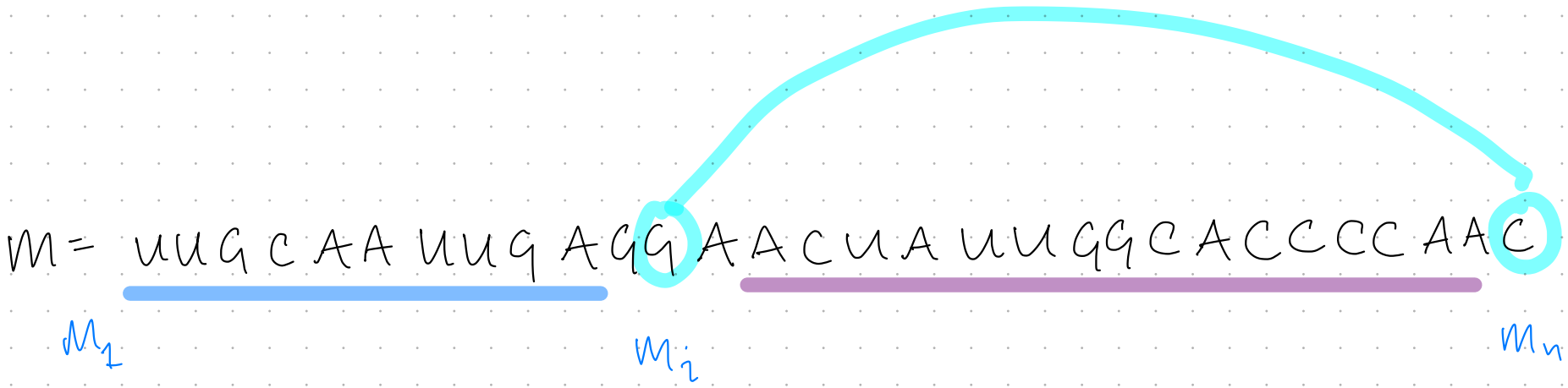
M_1 M_n

What happens to M_n?

Cases In optimal folding of M.

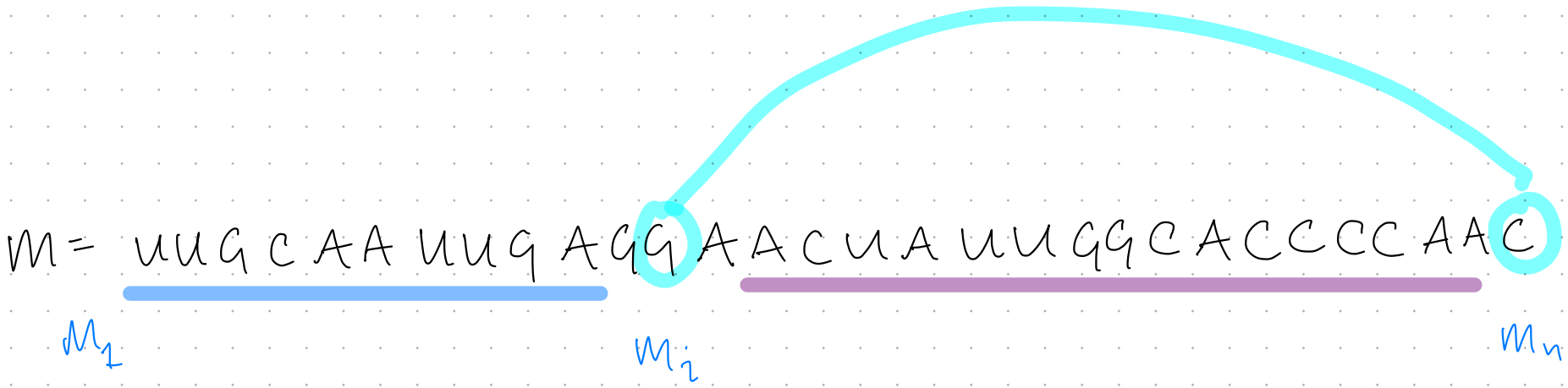
* m_n base pairs w/ some m_i $i < n$, OR

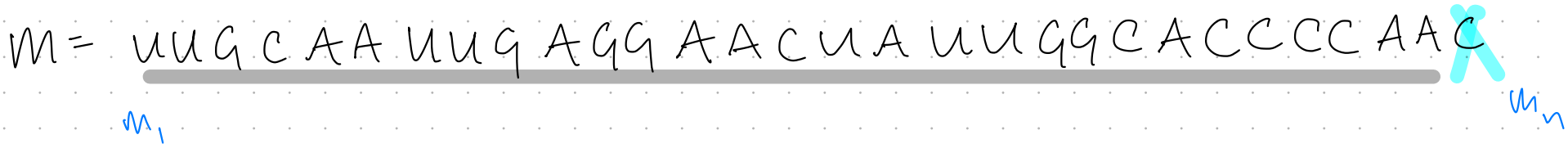
*



Cases In optimal folding of M

* m_n base pairs w/ some m_i $i < n$, OR
* m_n does Not base pair





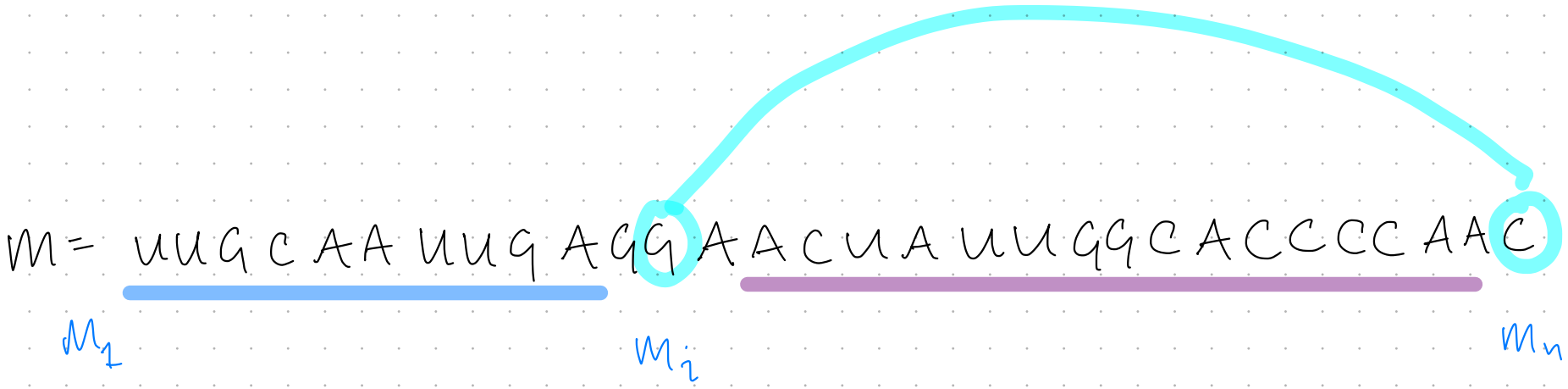
Cases m_n does Not base pair

$M =$ U U G C A A U U G A G G A A C U A U U G G C A C C C C A A G

m_1 m_n

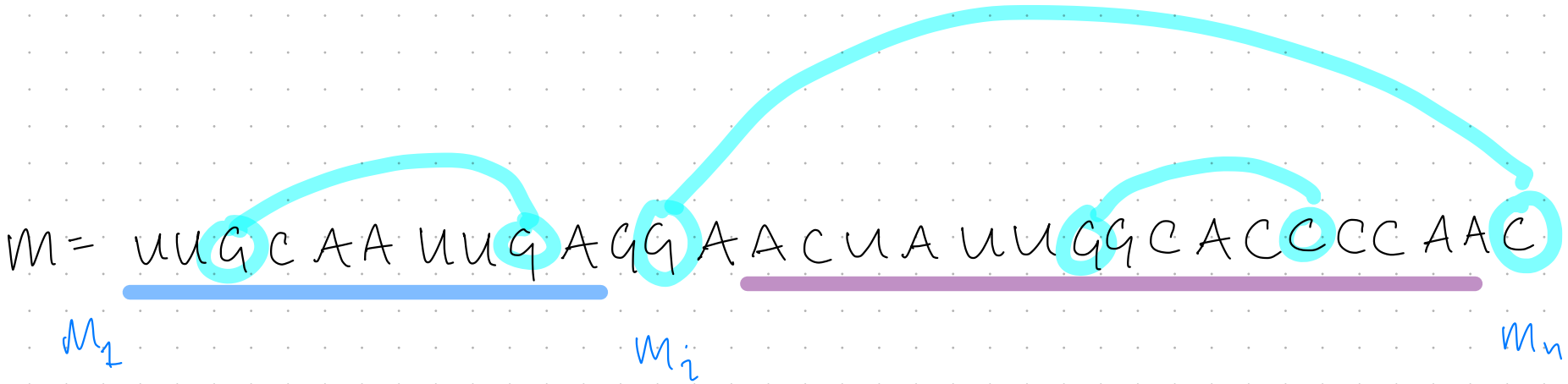
Best up til M_n = Best up til M_{n-1}

Cases m_n base pairs w/ some m_i $i < n$



What base pairs are still legal?

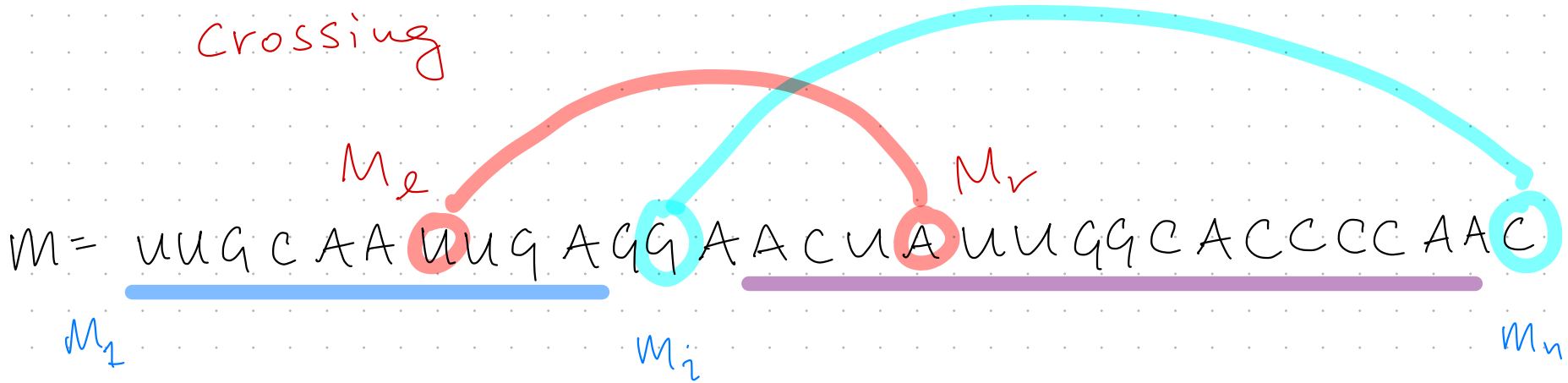
Cases m_n base pairs w/ some m_i $i < n$



What base pairs are still legal?

* ALLOWED: from $M_1 \rightarrow M_{i-1}$ & $M_{i+1} \rightarrow M_{n-1}$

Cases m_n base pairs w/ some m_i $i < n$



What base pairs are still legal?

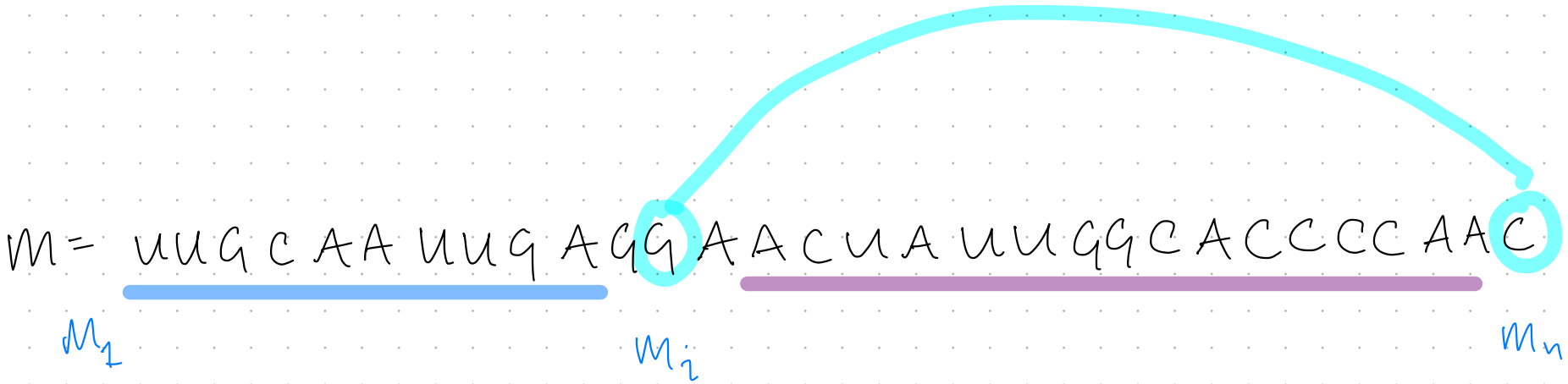
* ALLOWED: from $M_1 \rightarrow M_{i-1}$ & $M_{i+1} \rightarrow M_{n-1}$

* NOT ALLOWED: Crossings!

from $M_\ell \rightarrow M_r$

for $\ell < i < r$.

Cases m_n base pairs w/ some m_i $i < n$



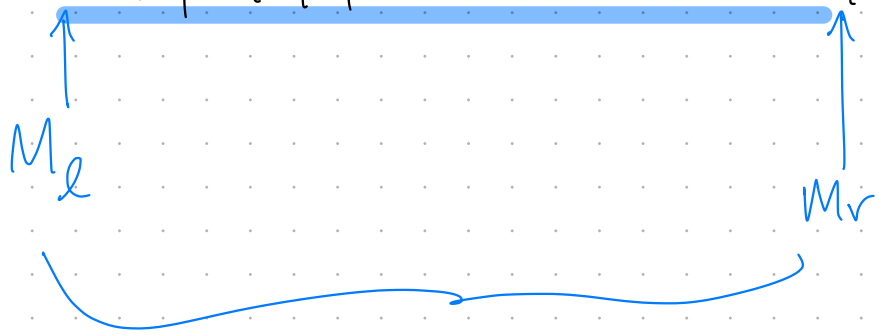
Best for M

=

Best of left + Best of right

Subproblems = Substrings

M = UUGCAAUUGAGGAACUAUUGGCACCCCAAC

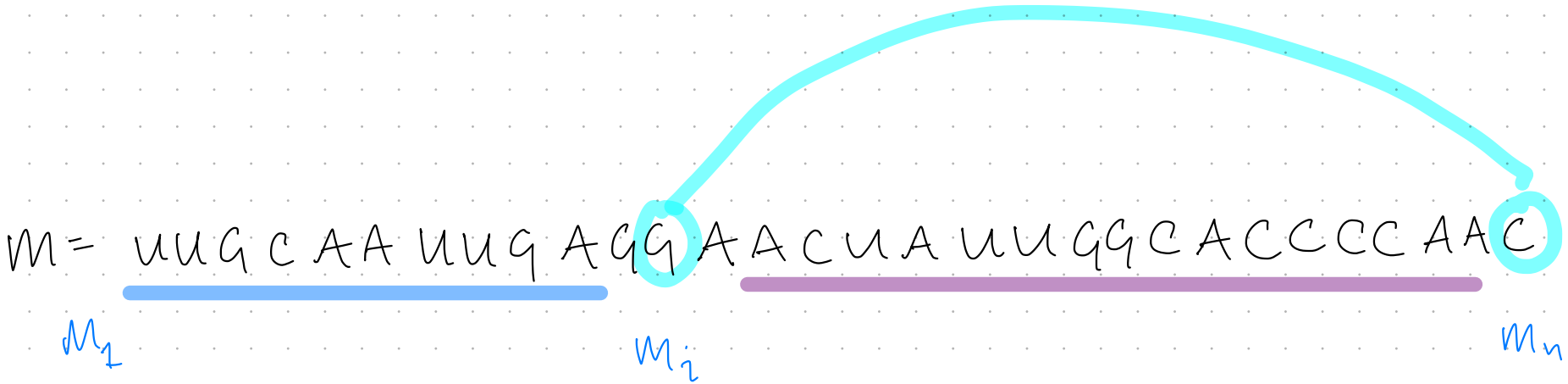


Best folding of substring from left endpoint M_ℓ to right endpoint M_r

= Fold(ℓ, r)

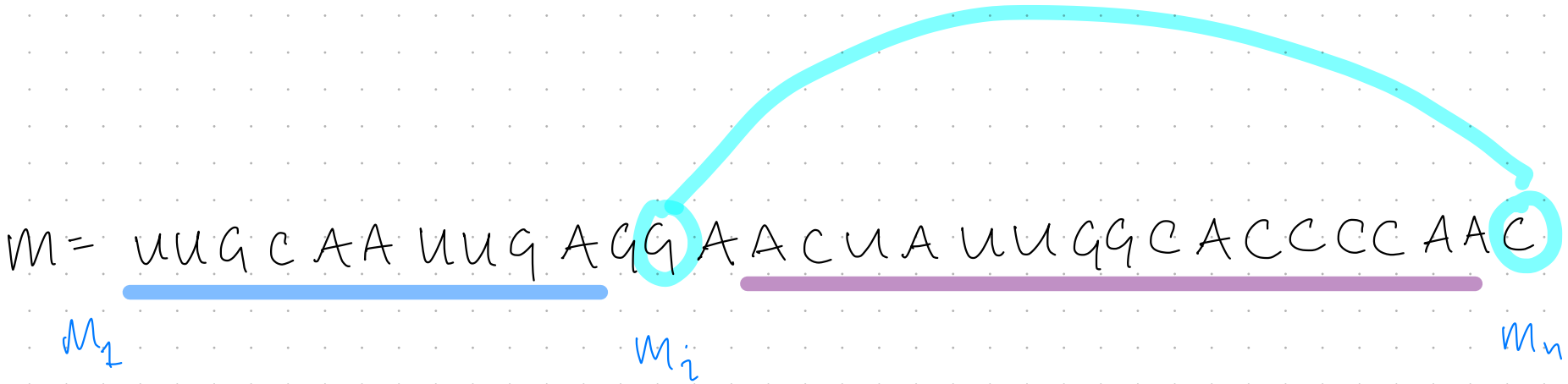
⇒ RNA Folding solved by Fold(1, n)

Cases m_n base pairs w/ some m_i $i < n$



$$Fold(1,n) = 1 + Fold(1,i-1) + Fold(i+1,n-1)$$

Cases m_n base pairs w/ some m_i $i < n$



$$Fold(1,n) = \underline{1} + \underline{Fold(1,i-1)} + \underline{Fold(i+1,n-1)}$$

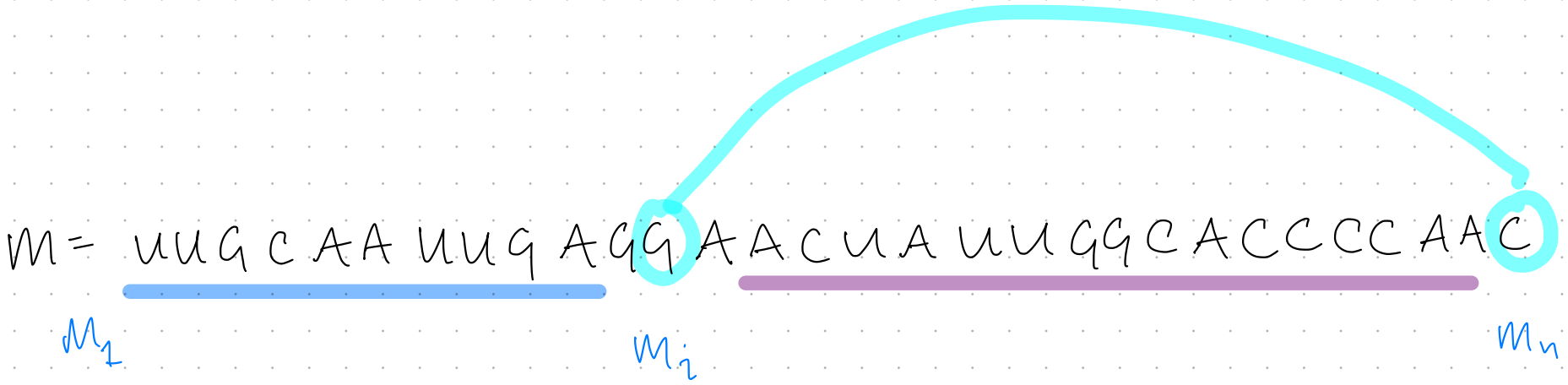
Note: subproblem does NOT start at $l=1$

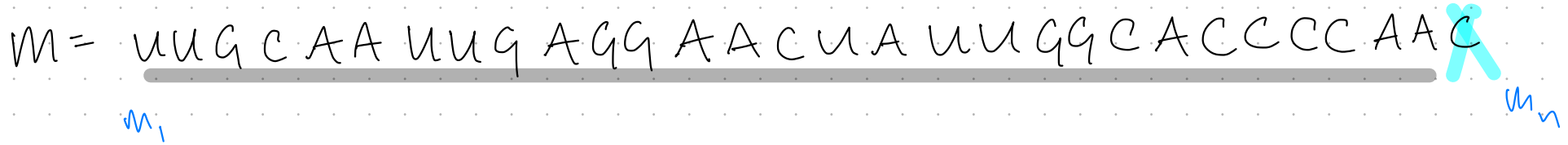
Cases m_n does Not base pair

$M =$ UUGCAAUUGAGGAACUAUUGGCACCCCAAG

m_1 m_n

$$Fold(1,n) = Fold(1,n-1)$$

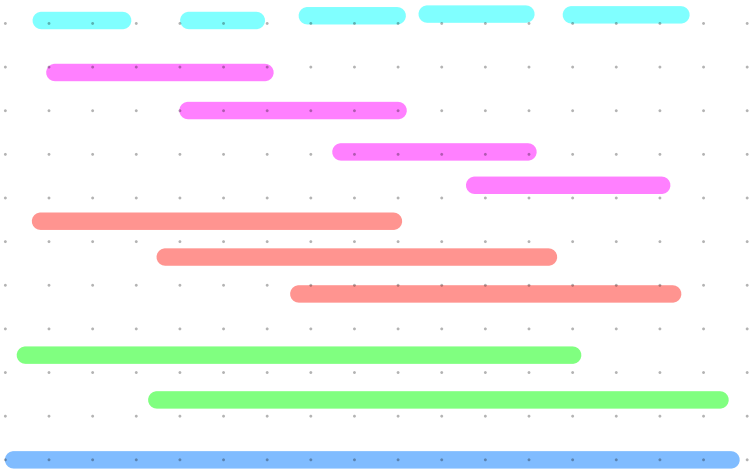




RNA Recurrence

$$Fold(1,n) = Max \begin{cases} 1 + \max\limits_{i<n \text{ s.t. } (M_i, M_n) \in BP} \{ Fold(1, i-1) + Fold(i+1, n) \} \\ Fold(1, n-1) \end{cases}$$

M = A C A G U



Subproblems.

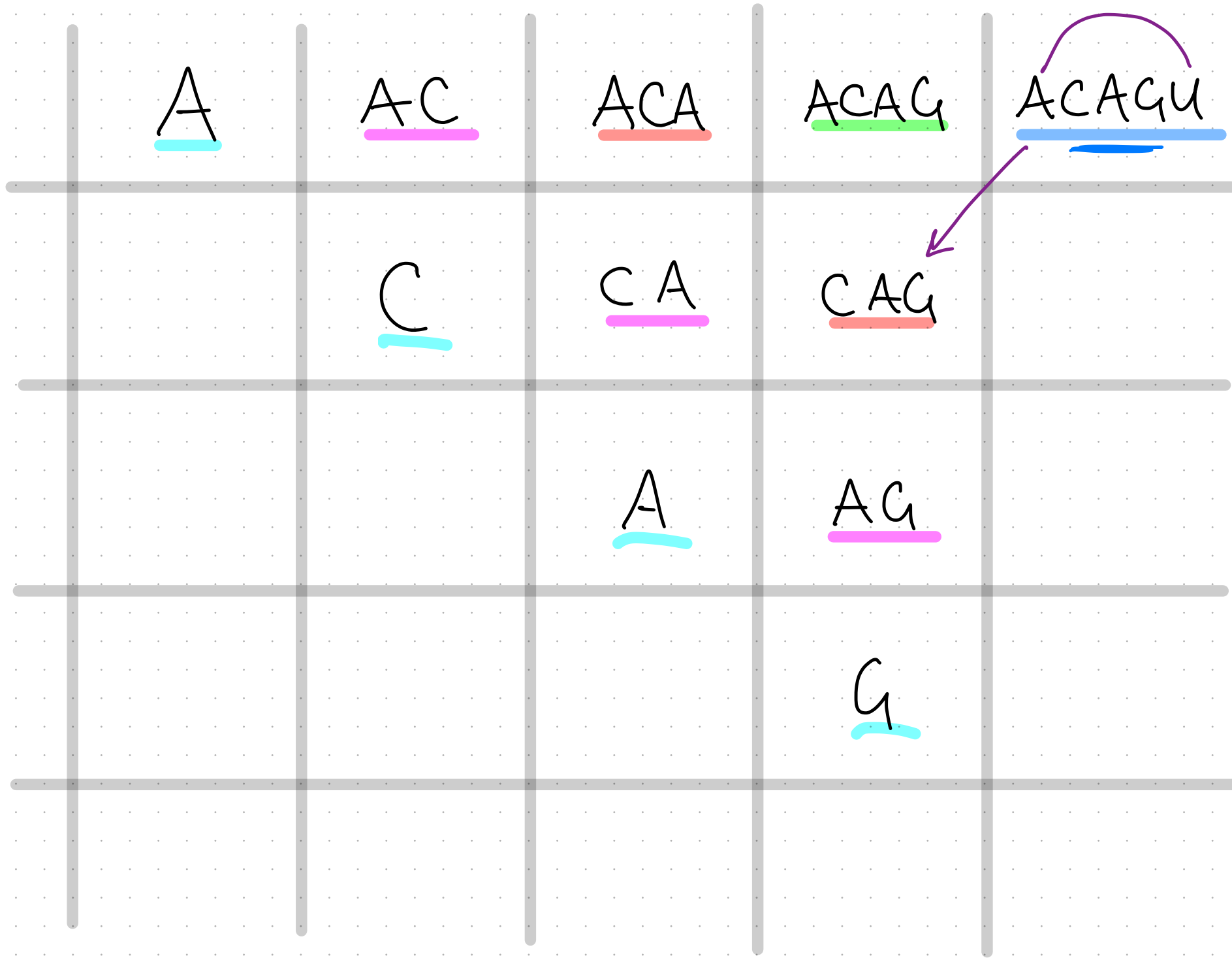
* For each width 1, 2, 3, 4, 5

* For each left endpoint

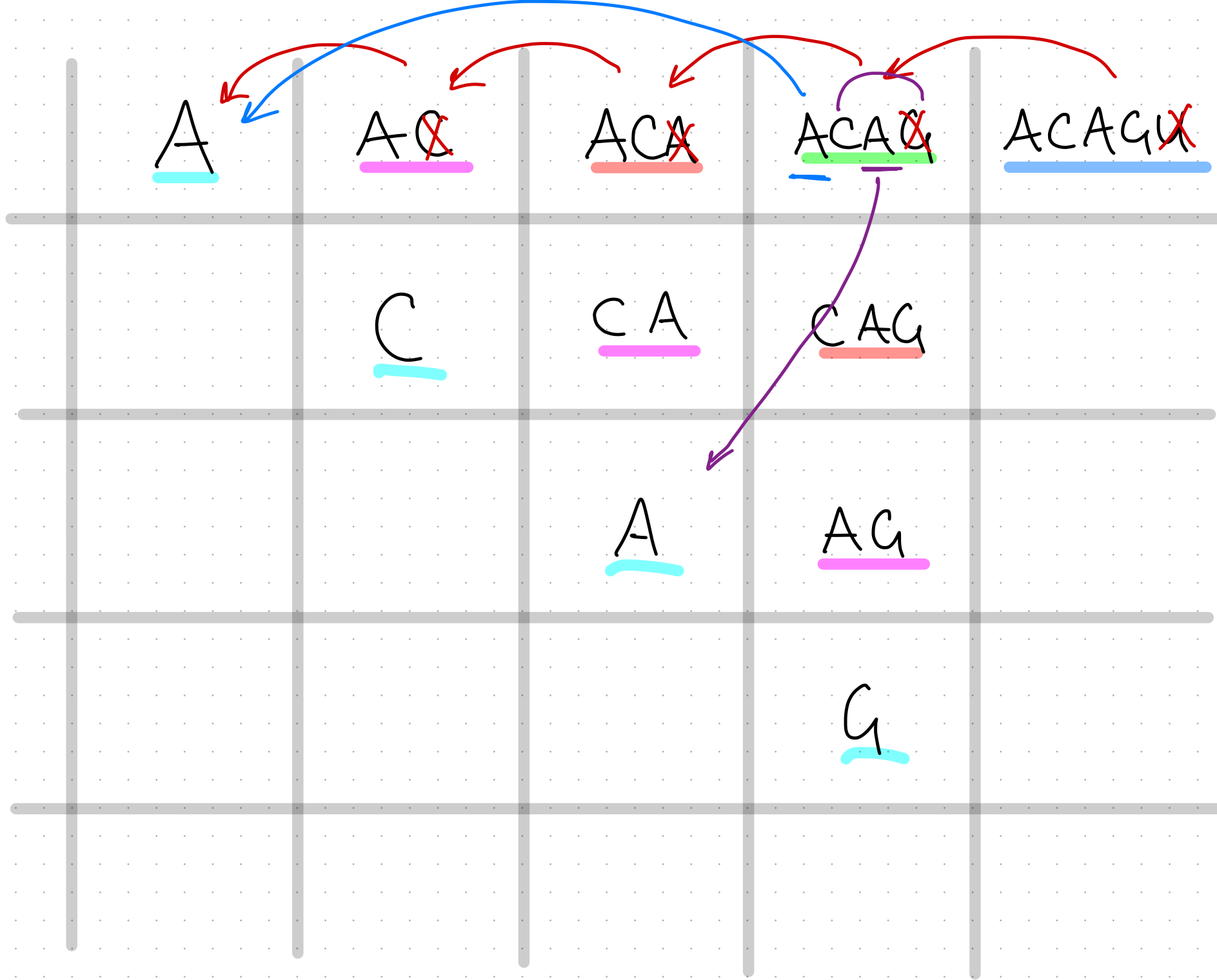
A, C, A, G, U


A
AC
ACA
ACAG
ACAGU
C
CA
CAG
A
AG
G

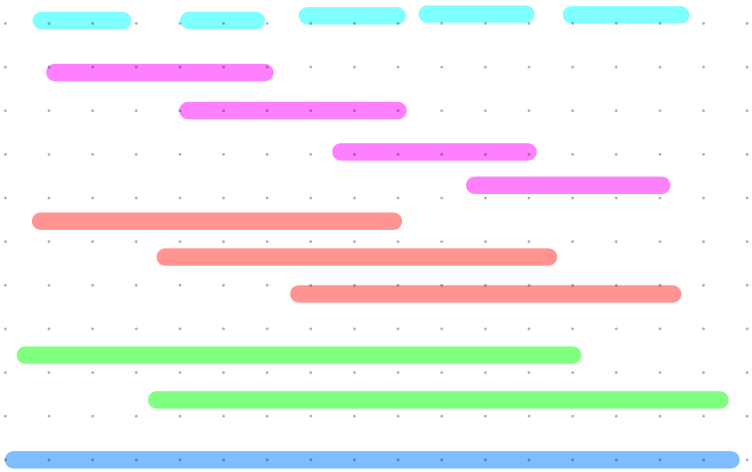

A
AC
ACA
ACAG
ACAGA
C
CA
CAG
A
AG
G


A
AC
ACA
ACAG
ACAGU
C
CA
CAG
A
AG
G


A
AC
ACA
ACAG
ACAGU
C
CA
CAG
A
AG
G


A
AC
ACA
ACAG
ACAGA
C
CA
CAG
A
AG
G

M = A C A G U



Subproblems.

* For each width 1, 2, 3, 4, 5

* For each left endpoint

A, C, A, G, U

RNA-Fold(n)

For $l = 1 \to n$: Fold$(l, l) = 0$.

For width $= 1 \to n-1$

  For $l = 1 \to n - \text{width}$

    $r = l + \text{width}$

    noPair $\leftarrow$ Fold$(l, r-1)$

    bestPair $\leftarrow$

Find best legal base pair between M_q and M_r for some $l < q < r$

Fold$(l, r) = \max\{$ noPair, bestPair $\}$

RNA-Fold(M)

For $l = 1 \to n$: $Fold(l, l) = 0$.

For width $= 1 \to n-1$

  For $l = 1 \to n - width$

    $r = l + width$

    noPair $\leftarrow$ $Fold(l, r-1)$

    bestPair $\leftarrow$ 0

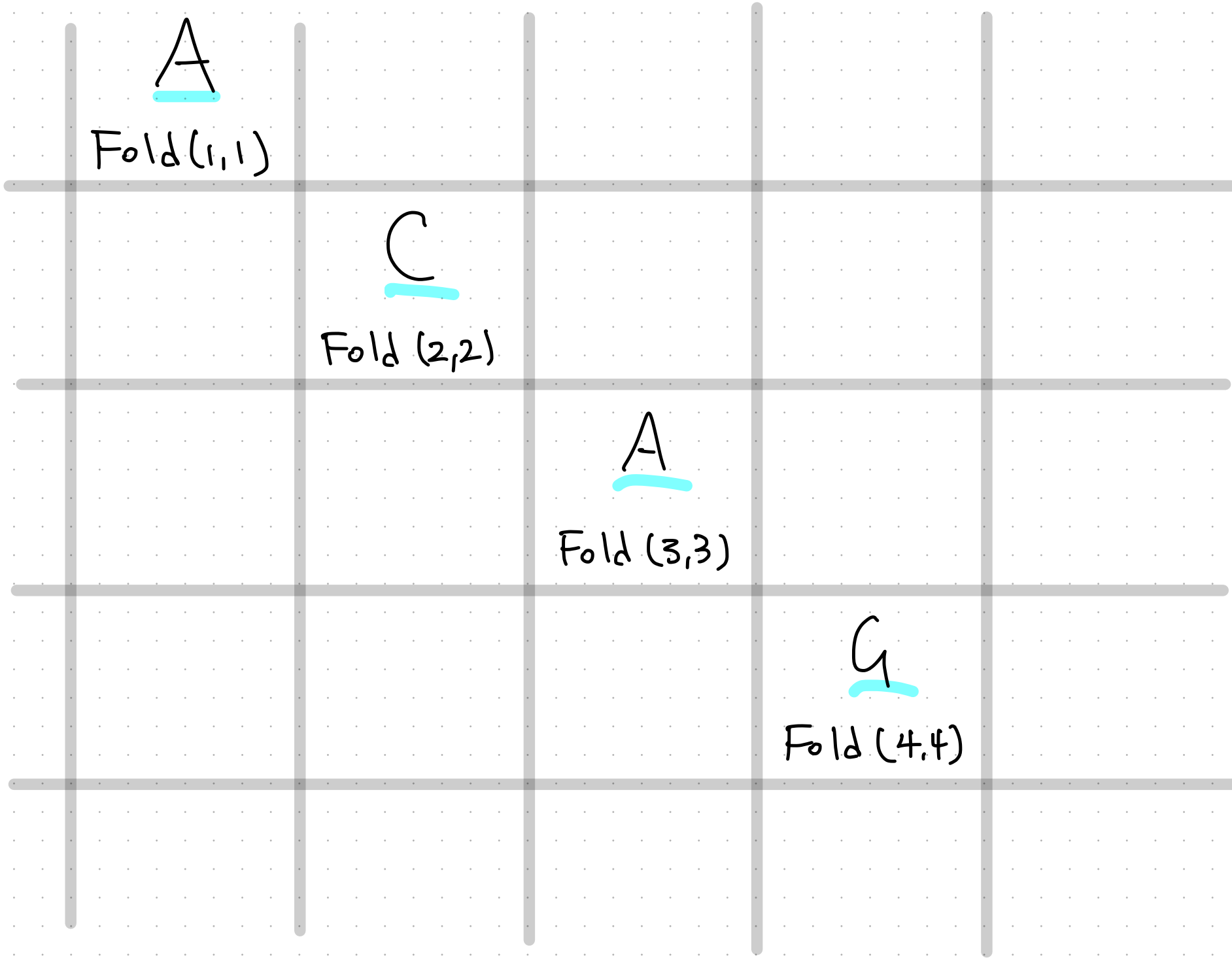
    For $q = l \to r-1$

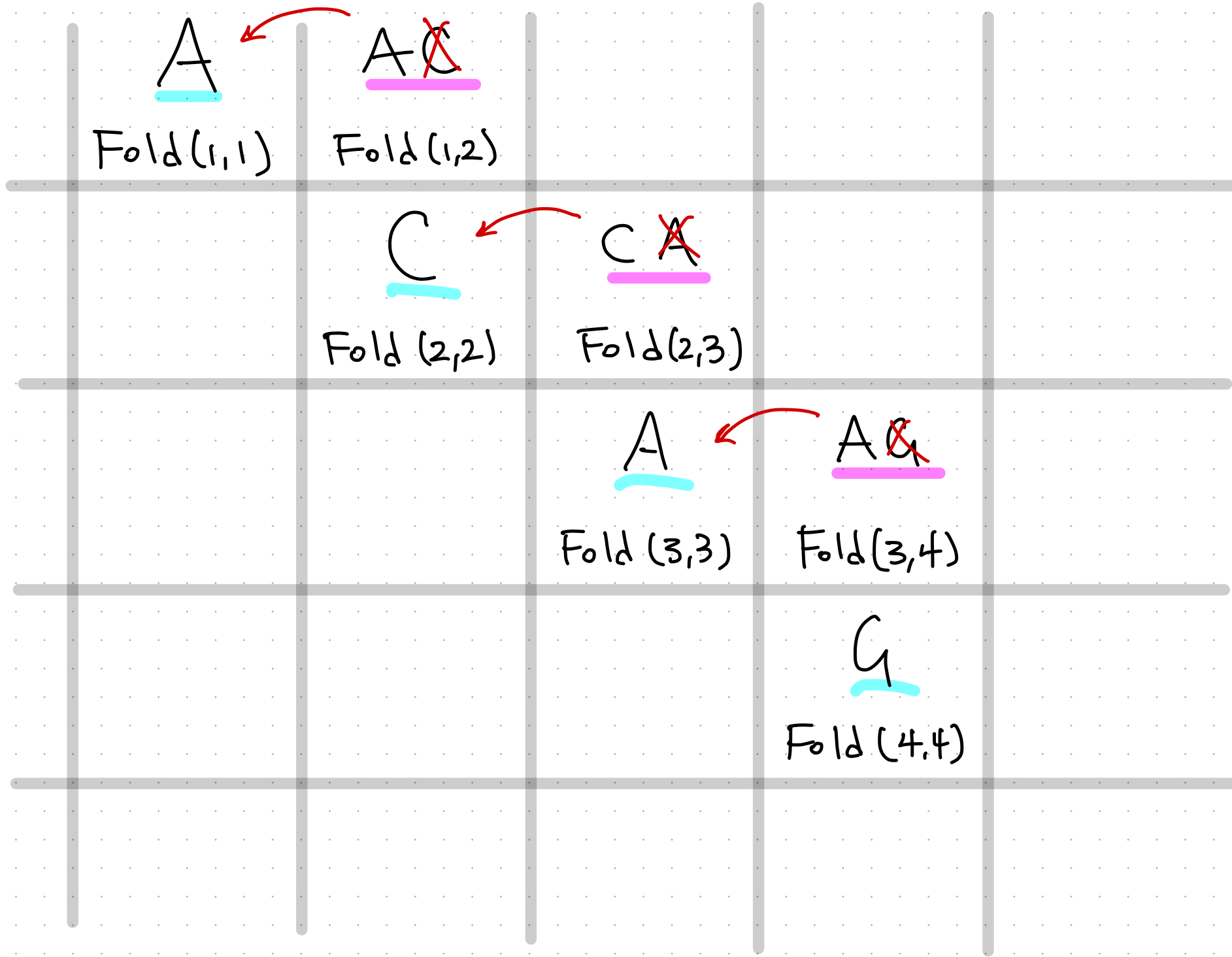
      if $(M_q, M_r) \in BP$:

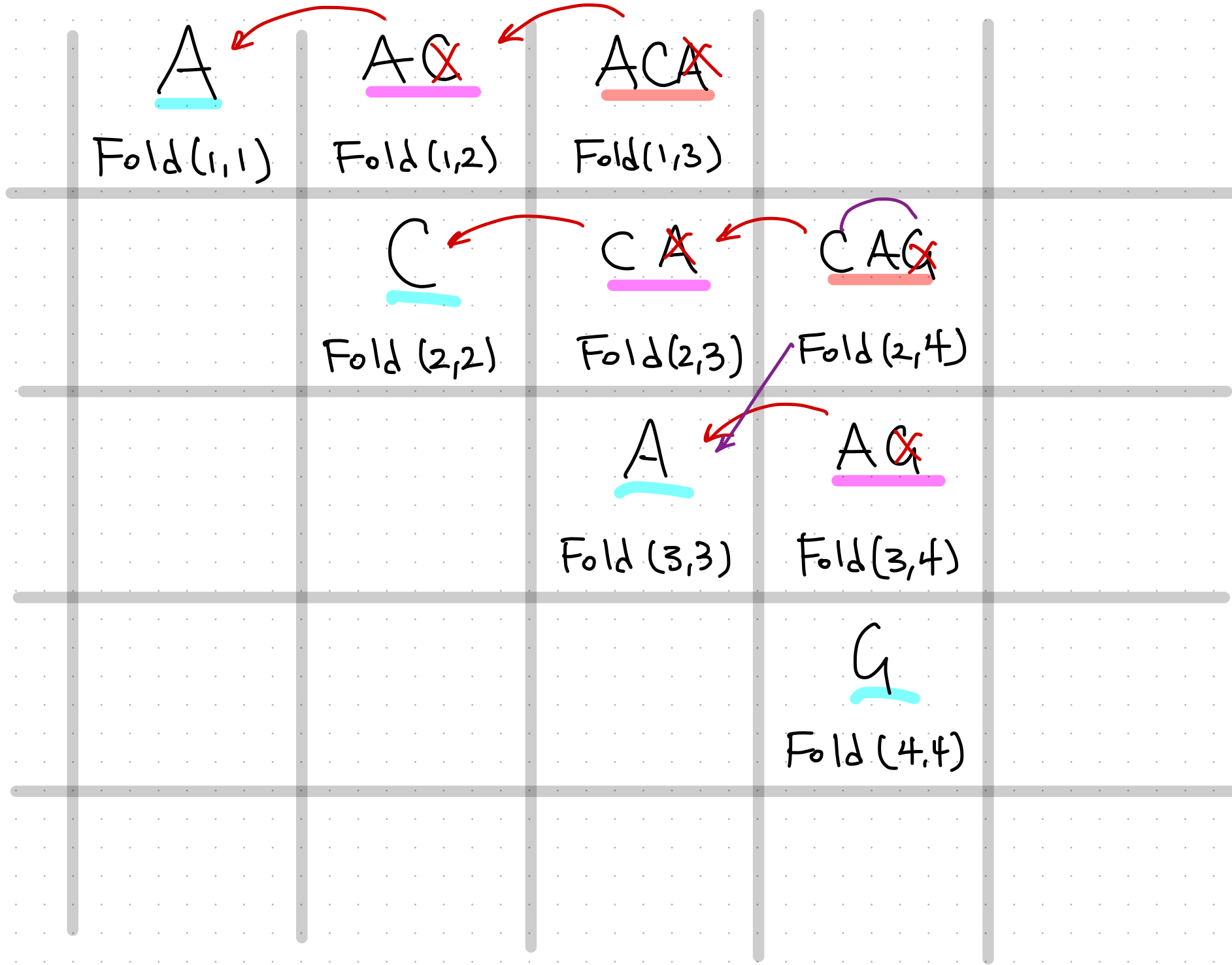
        qPair $\leftarrow$ $1 + Fold(1, q-1) + Fold(q+1, r-1)$

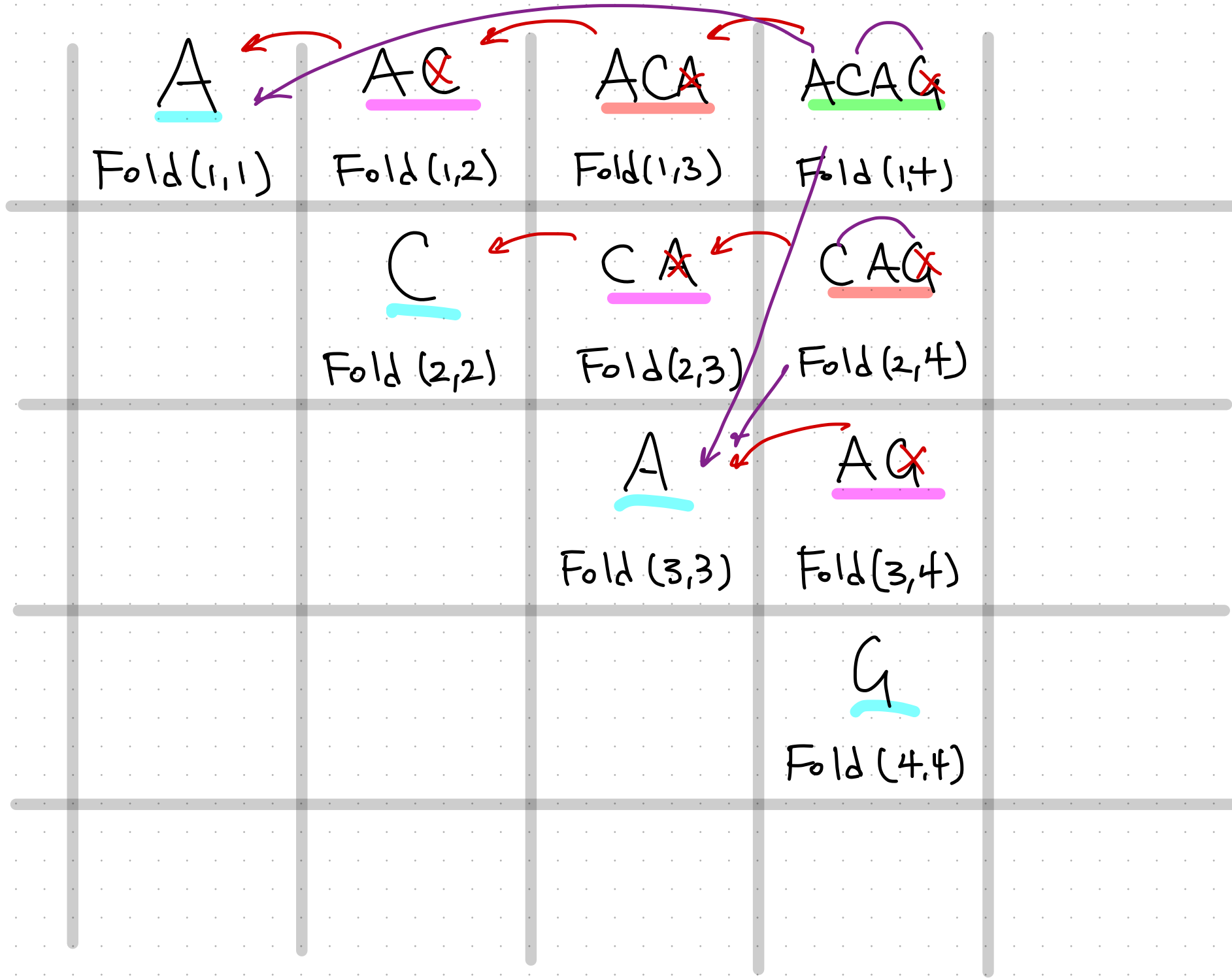
        bestPair $\leftarrow$ $\max\{qPair, bestPair\}$

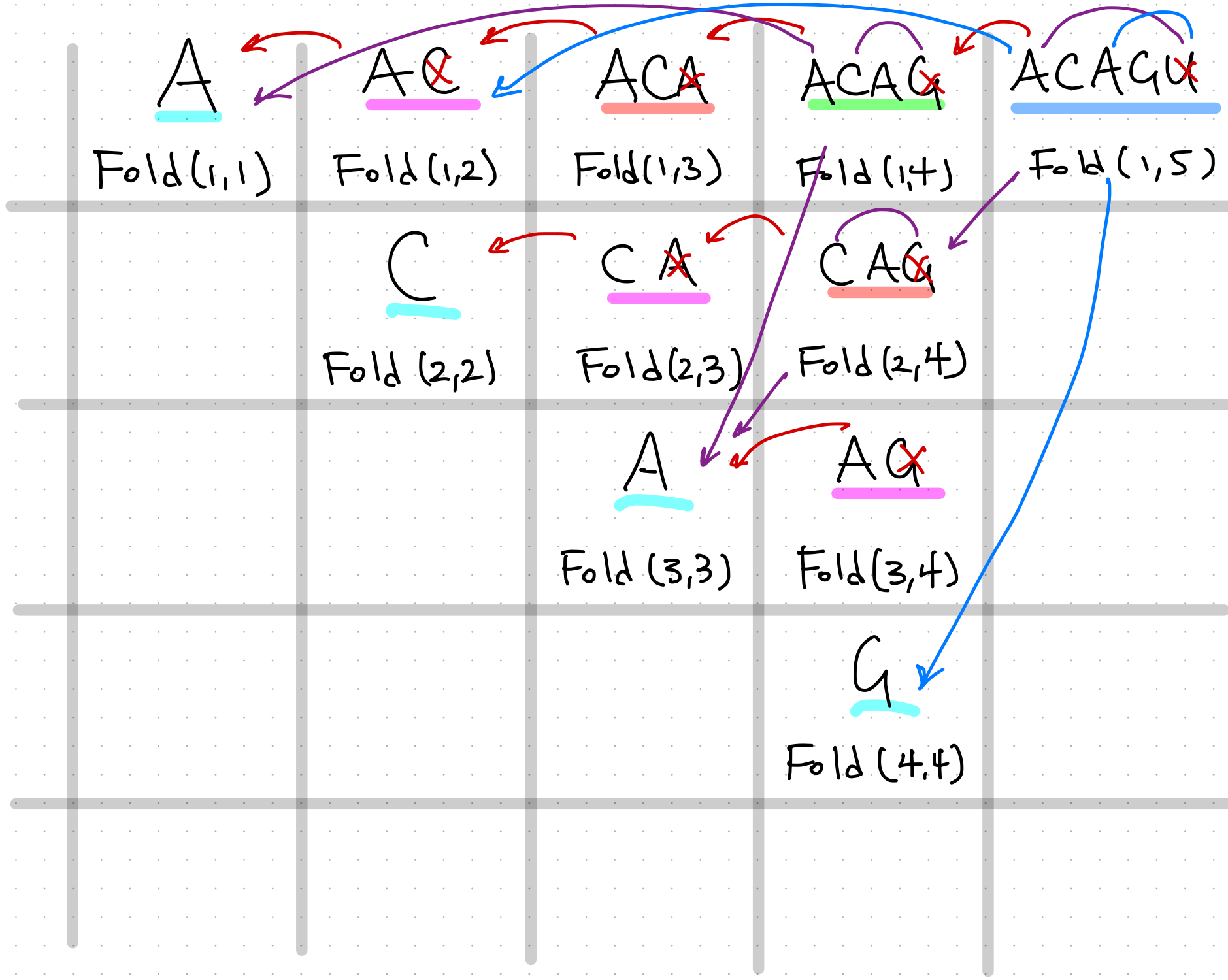
    $Fold(l, r) = \max\{noPair, bestPair\}$


A
Fold (1,1)
C
Fold (2,2)
A
Fold (3,3)
G
Fold (4,4)


A
AC
Fold (1,1)
Fold (1,2)
C
CA
Fold (2,2)
Fold (2,3)
A
AG
Fold (3,3)
Fold (3,4)
G
Fold (4,4)


A
AC
ACA
Fold (1,1)
Fold (1,2)
Fold (1,3)
C
CA
CAG
Fold (2,2)
Fold (2,3)
Fold (2,4)
A
AG
Fold (3,3)
Fold (3,4)
G
Fold (4,4)


A
AC
ACA
ACAG
Fold(1,1)
Fold(1,2)
Fold(1,3)
Fold(1,4)
C
CA
CAG
Fold(2,2)
Fold(2,3)
Fold(2,4)
A
AG
Fold(3,3)
Fold(3,4)
G
Fold(4,4)


A
AC
ACA
ACAG
ACAGU
Fold(1,1)
Fold(1,2)
Fold(1,3)
Fold(1,4)
Fold(1,5)
C
CA
CAG
Fold(2,2)
Fold(2,3)
Fold(2,4)
A
AG
Fold(3,3)
Fold(3,4)
G
Fold(4,4)

RNA-Fold(M)

For $l = 1 \to n$: Fold$(l, l) = 0$.

For width $= 1 \to n-1$

  For $l = 1 \to n - \text{width}$

    $r = l + \text{width}$

    noPair $\leftarrow$ Fold$(l, r-1)$

    bestPair $\leftarrow 0$

    For $q = l \to r-1$

      if $(M_q, M_r) \in BP$:

        qPair $\leftarrow 1 +$ Fold$(l, q-1)$ + Fold$(q+1, r-1)$

        bestPair $\leftarrow \max\{$qPair, bestPair$\}$

Fold$(l, r) = \max\{$noPair, bestPair$\}$

Running Time

Entries for each $(l, r) \in [n] \times [n]$

$\Rightarrow O(n^2)$

To fill in each entry

↳ Probe q s.t. $l < q < r$

$O(n)$

Total

$\Rightarrow O(n^3)$

Proof of Correctness.

For all $n \in \mathbb{N}$, Fold$(1, n)$ returns max non-crossing match.

Proof of Correctness.

For all $n \in \mathbb{N}$, $\text{Fold}(1, n)$ returns max non-crossing match.

For all widths $w \in \mathbb{N}$, for all left endpoints $\ell \leq |M| - w$

$\text{Fold}(\ell, \ell + w)$ = max non-crossing match w/ left endpoint M_ℓ and right endpoint $M_{\ell+w}$.

Proof of Correctness.

For all $n \in \mathbb{N}$, Fold(1, n) returns max non-crossing match.

↓

For all widths $w \in \mathbb{N}$, for all left endpoints $\ell \leq |M| - w$

Fold(ℓ, $\ell + w$) = max non-crossing match w/ left endpoint M_ℓ and right endpoint $M_{\ell+w}$.

Base case. $w = 0$.

* No base pairing possible w/ single character

  ⇒ max match = 0 for all ℓ.

* Fold(ℓ, ℓ) = 0 ✓

Inductive Hypothesis. $\forall w_0 < w$, for all legal l

$Fold(l, l+w_0)$ = best RNA fold from M_l to M_{l+w_0}

Consider $M_l\ M_{l+1}\ M_{l+2} \cdots M_{l+w-1}\ M_{l+w}$.

Two cases.

* M_{l+w} not paired:

Best fold from $M_l \to M_{l+w}$

= Best fold from $M_l \to M_{l+w-1}$ (By IH)

= $Fold(l, l+w-1)$

↓

width $w-1 < w$

Inductive Hypothesis. $\forall w_0 < w$, for all legal l

$Fold(l, l+w_0)$ = best RNA fold from M_l to M_{l+w_0}

Consider $M_l\ M_{l+1}\ M_{l+2} \cdots M_{l+w-1}\ M_{l+w}$.

Two cases. M_{l+w} paired?

Best fold from $M_l \to M_{l+w}$

$= \text{Best } q \quad BP(M_q, M_r) +$

Best fold from $M_l \to M_{q-1}$ + Best fold from $M_{q+1} \to M_{l+w-1}$

Inductive Hypothesis. $\forall w_0 < w$, for all legal l

$Fold(l, l+w_0)$ = best RNA fold from M_l to M_{l+w_0}

Consider $M_l\ M_{l+1}\ M_{l+2} \cdots M_{l+w-1}\ M_{l+w}$.

Two cases. M_{l+w} paired?

Best fold from $M_l \to M_{l+w}$

$= \text{Best}_q\ BP(M_q, M_r) +$

Best fold from $M_l \to M_{q-1}$ + Best fold from $M_{q+1} \to M_{l+w-1}$

Both of width $w' < w$.

Inductive Hypothesis. $\forall w_0 < w$, for all legal l

$Fold(l, l+w_0)$ = best RNA fold from M_l to M_{l+w_0}

Consider $M_l\ M_{l+1}\ M_{l+2} \cdots M_{l+w-1}\ M_{l+w}$.

Two cases. M_{l+w} paired?

Best fold from $M_l \to M_{l+w}$

$= \text{Best } q \quad BP(M_q, M_r) +$

Best fold from $M_l \to M_{q-1}$ + Best fold from $M_{q+1} \to M_{l+w-1}$

Both of width $w' < w$.

$= 1 + Fold(l, q-1) + Fold(q+1, l+w-1)$

By IH

∎