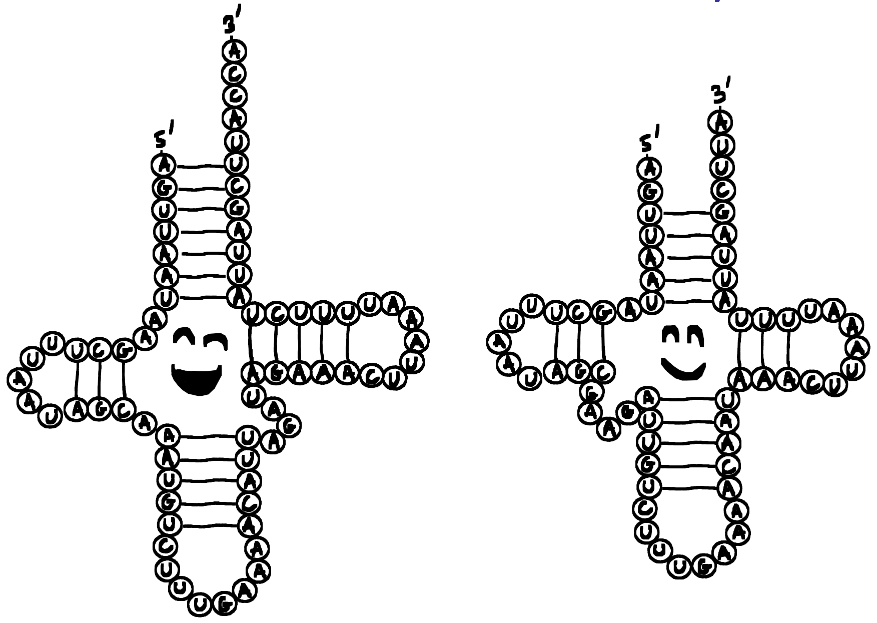
COUNTING, GENERATING AND SAMPLING TREE ALIGNMENTS

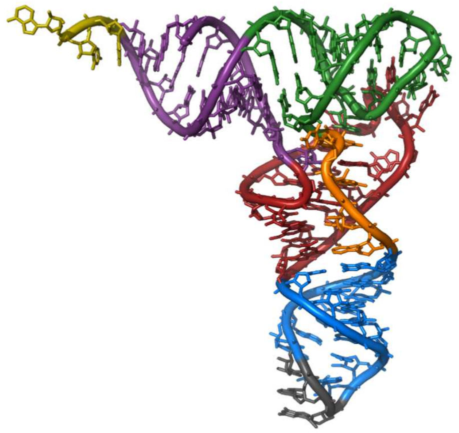
Cedric CHAUVE (Simon Fraser University, Vancouver)
Julien COURTIEL (PIMS/Univ. of British Columbia, Vancouver)
Yann PONTY (CNRS/LIX, Ecole Polytechnique, Inria Saclay)

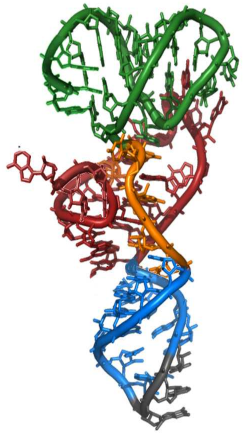


AlCoB 2016

MOTIVATION: RNA COMPARISON

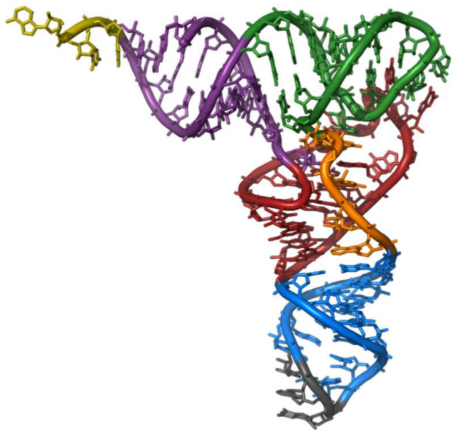
Question: how to measure similarity between two RNAs?

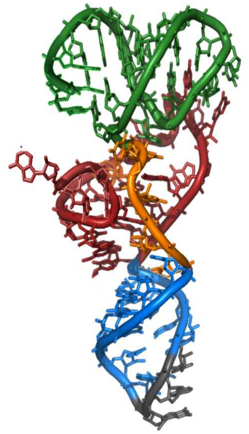




Motivation: RNA comparison

Question: how to measure similarity between two RNAs?



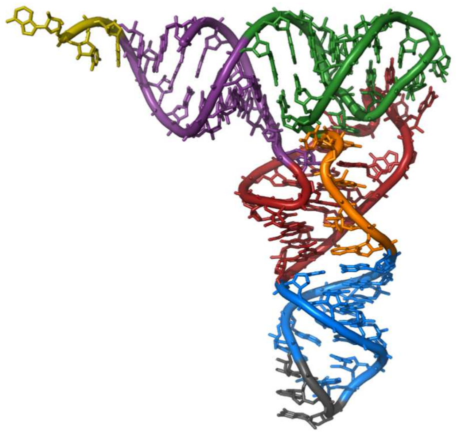


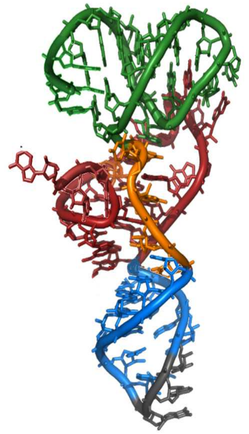
First idea: compare nucleic acid sequences

RNA 1: AUUCGAUUA...
RNA 2: ACCAUGAUUA...

MOTIVATION: RNA COMPARISON

Question: how to measure similarity between two RNAs?





First idea: compare nucleic acid sequences
→ sequence alignment

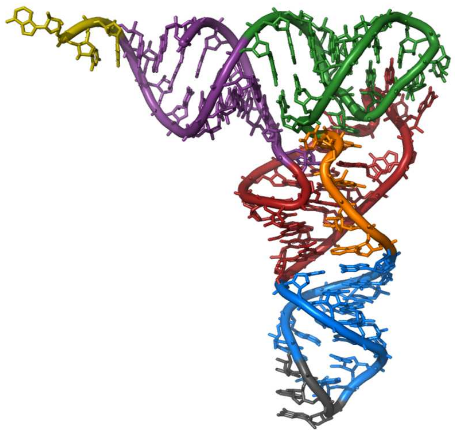
RNA 1: A U U C G A U U A ...

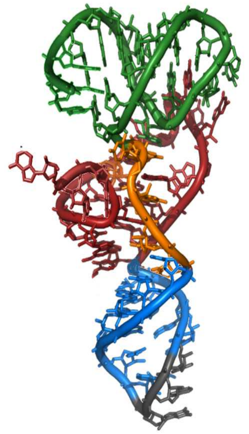
RNA 2: A C C A U G A U U A ...

alignment: $\begin{pmatrix}A\\A\end{pmatrix}\begin{pmatrix}U\\-\end{pmatrix}\begin{pmatrix}-\\C\end{pmatrix}\begin{pmatrix}U\\-\end{pmatrix}\begin{pmatrix}C\\C\end{pmatrix}\begin{pmatrix}-\\A\end{pmatrix}\begin{pmatrix}-\\U\end{pmatrix}\begin{pmatrix}G\\G\end{pmatrix}\begin{pmatrix}A\\A\end{pmatrix}\begin{pmatrix}U\\U\end{pmatrix}\begin{pmatrix}U\\U\end{pmatrix}\begin{pmatrix}A\\A\end{pmatrix}\ldots$

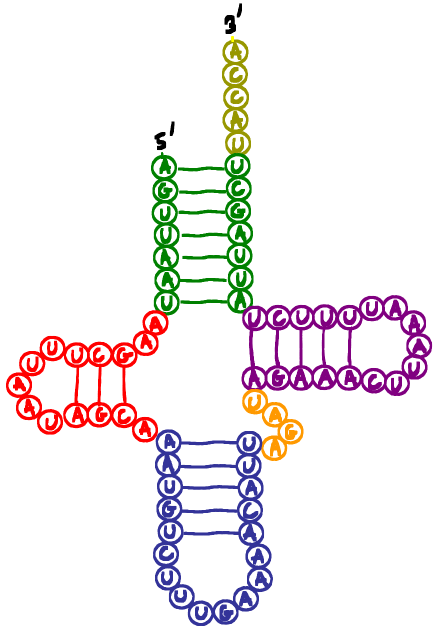
MOTIVATION: RNA COMPARISON

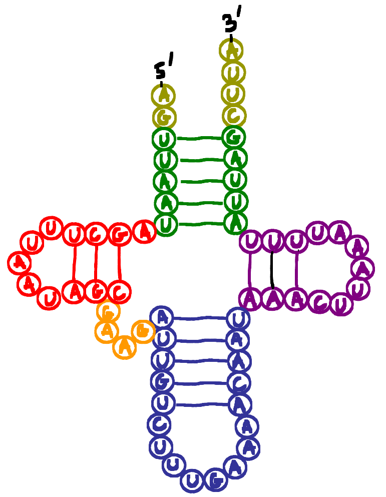
Question: how to measure similarity between two RNAs?





Second idea: compare secondary structures.





→ notion of tree alignment [Jiang, Wang, Zhang]

OUR CONTRIBUTION

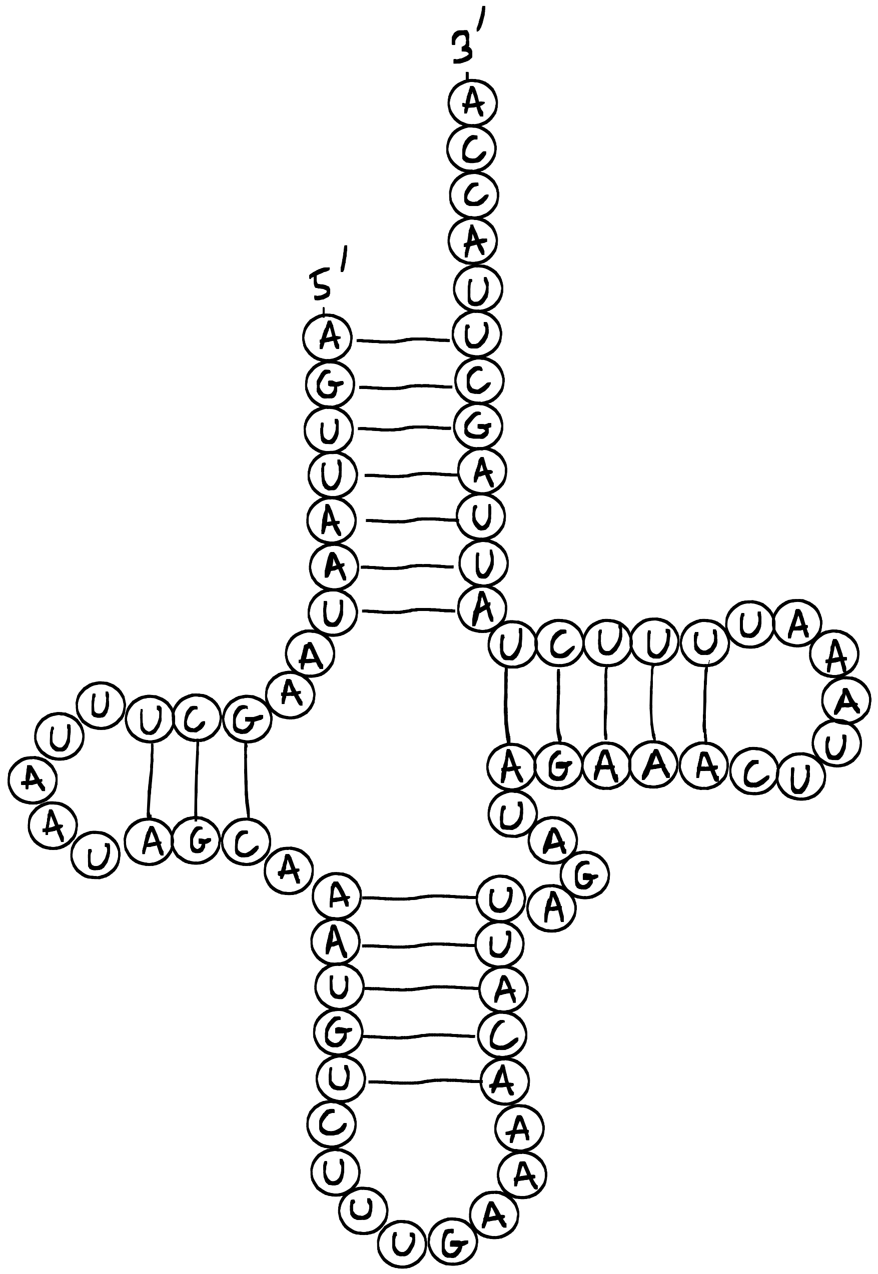
Our main result:

An unambiguous and complete Dynamic Programming scheme for tree alignments

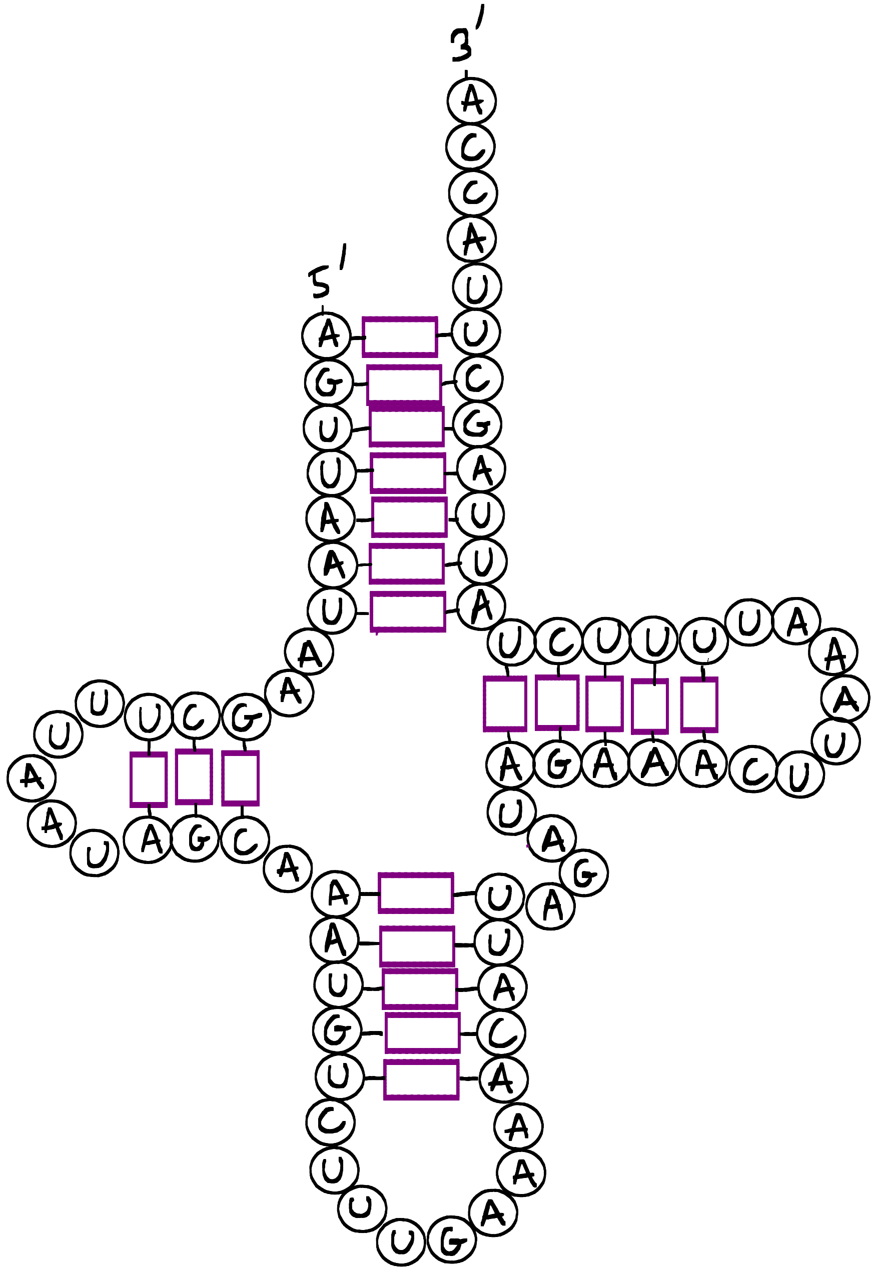
Side-products:

- Boltzmann sampling
- Enumerative results
- Average case complexity (revisited)
- 3D alignment and motif search (upcoming)

FROM SECONDARY STRUCTURES TO TREES

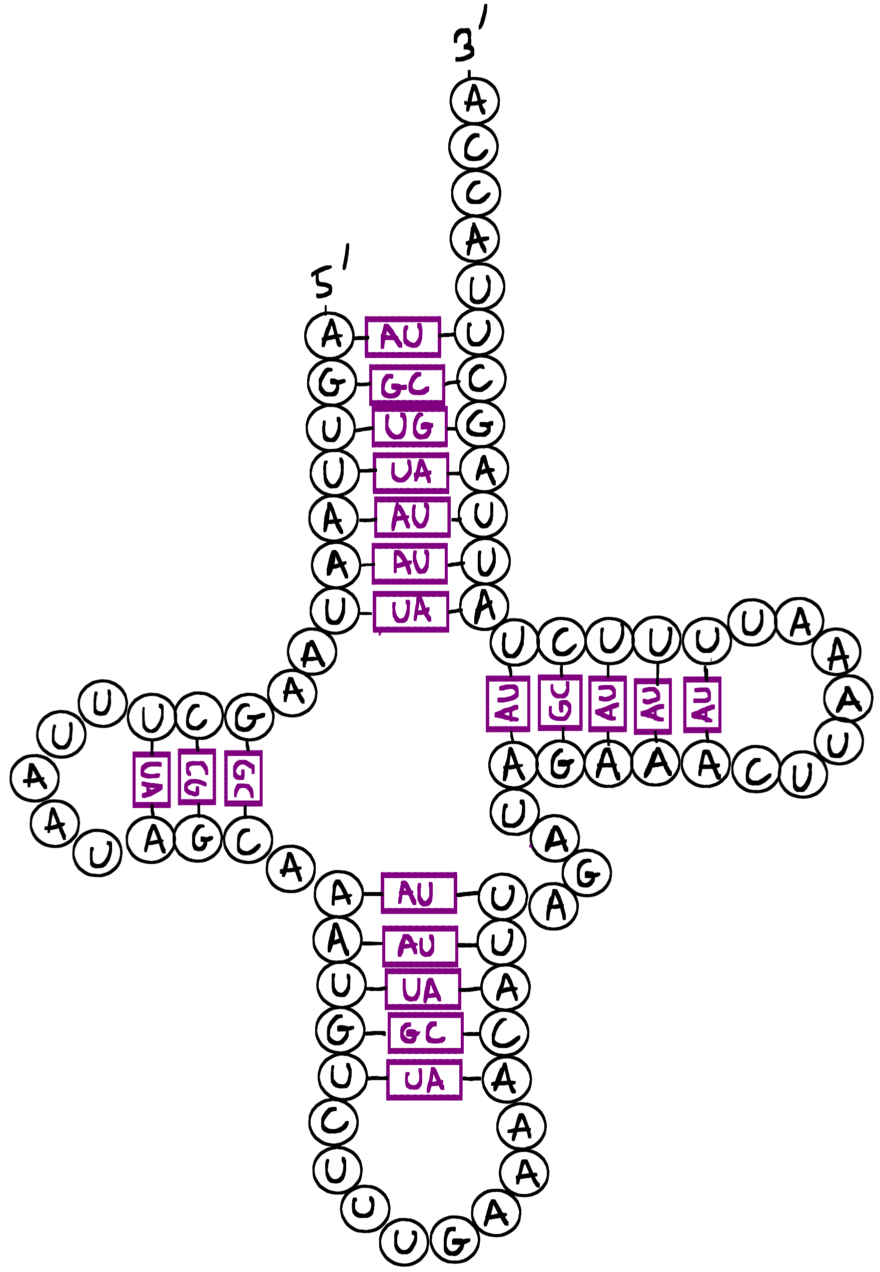


FROM SECONDARY STRUCTURES TO TREES

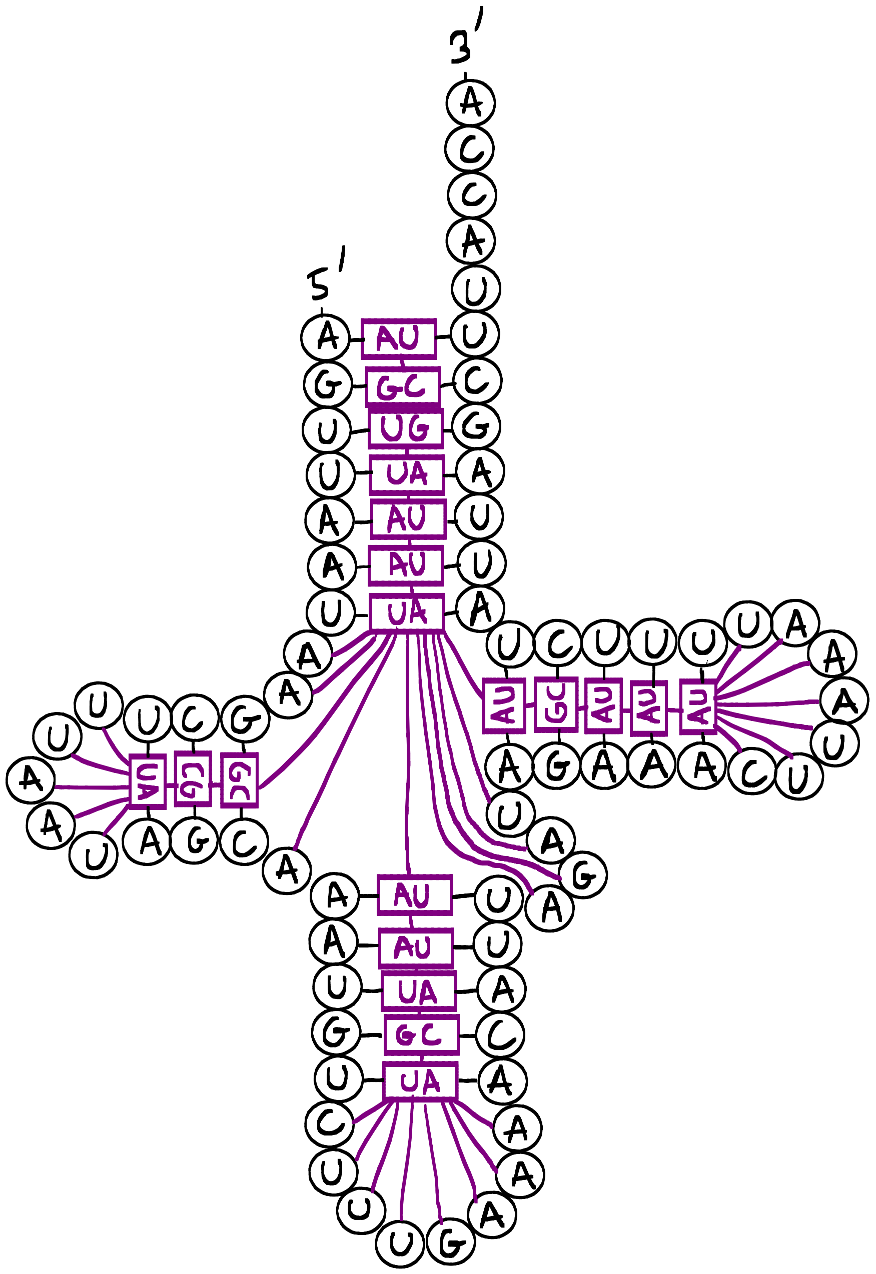


3'
5'
A C C A U U C G A U U A
A G U U A A U
U C U U U U A A A U C A A A G A
A G C U U A A U A G C
A A U G U C U U G A A A C A U U A G A U

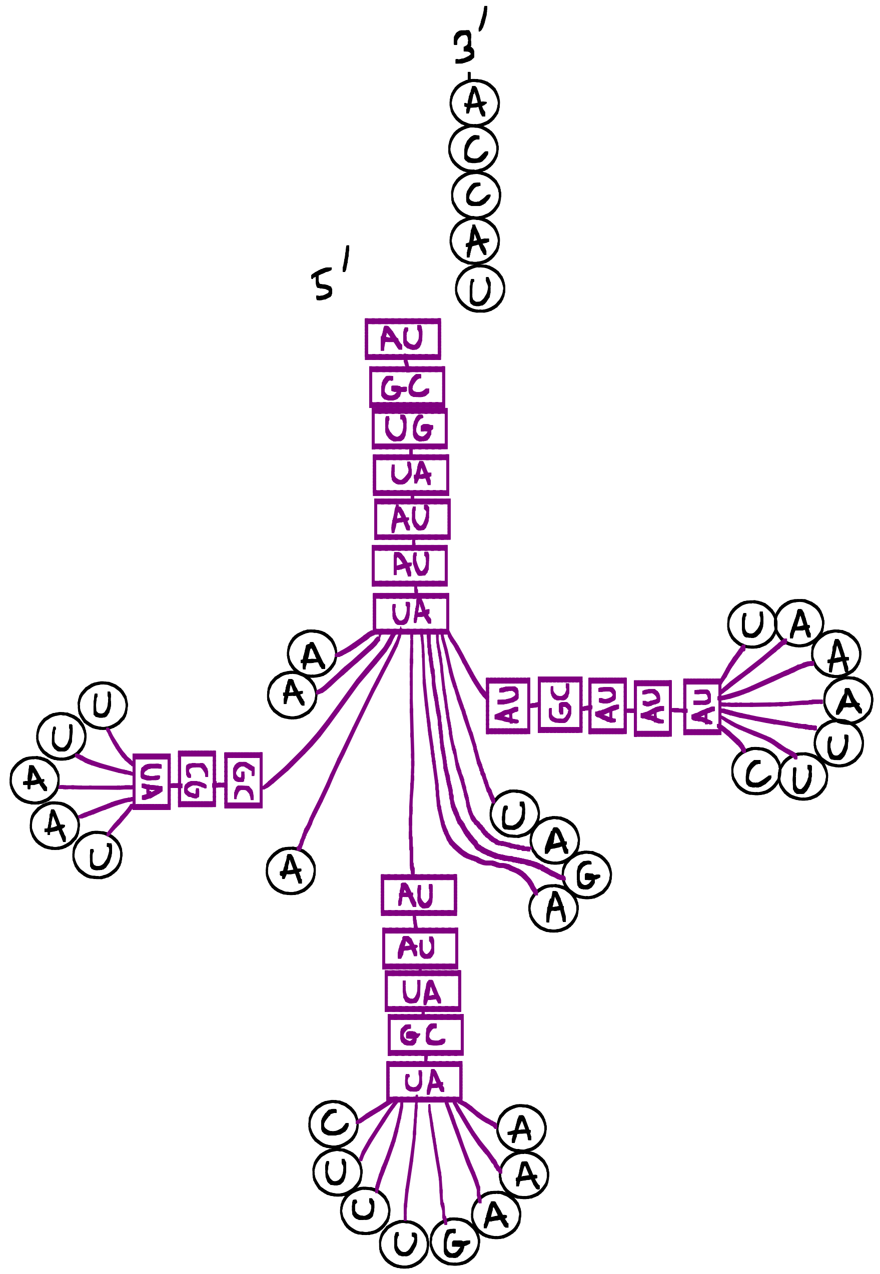
FROM SECONDARY STRUCTURES TO TREES


3'
5'
A C C A U U C G A U U A
A G U U A A U
AU
GC
UG
UA
AU
AU
UA
A A G C U C U U A A U A G C
UA
CG
GC
U C U U U U A A A U U C A A A G A
AU
GC
AU
AU
AU
U A G A
A A U G U C U U G A A A C A U U
AU
AU
UA
GC
UA

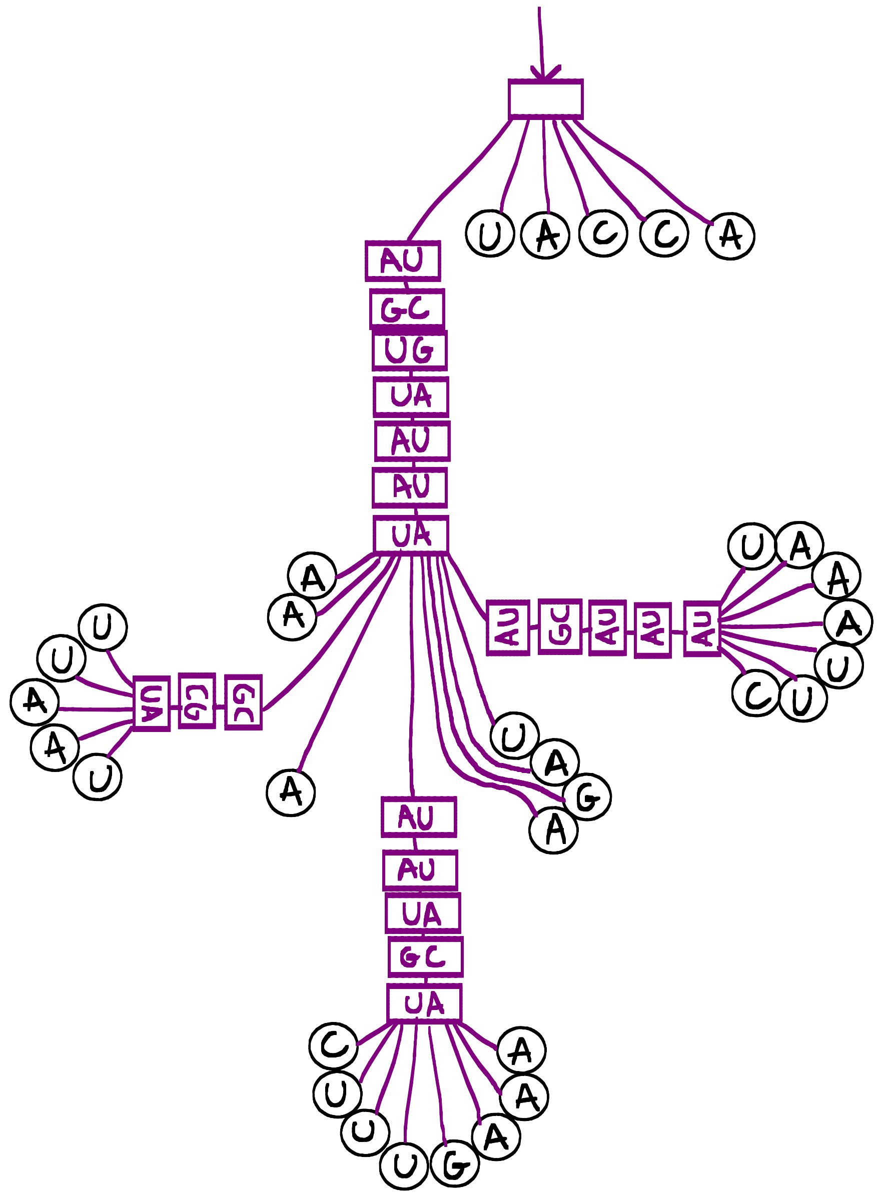
FROM SECONDARY STRUCTURES TO TREES



FROM SECONDARY STRUCTURES TO TREES

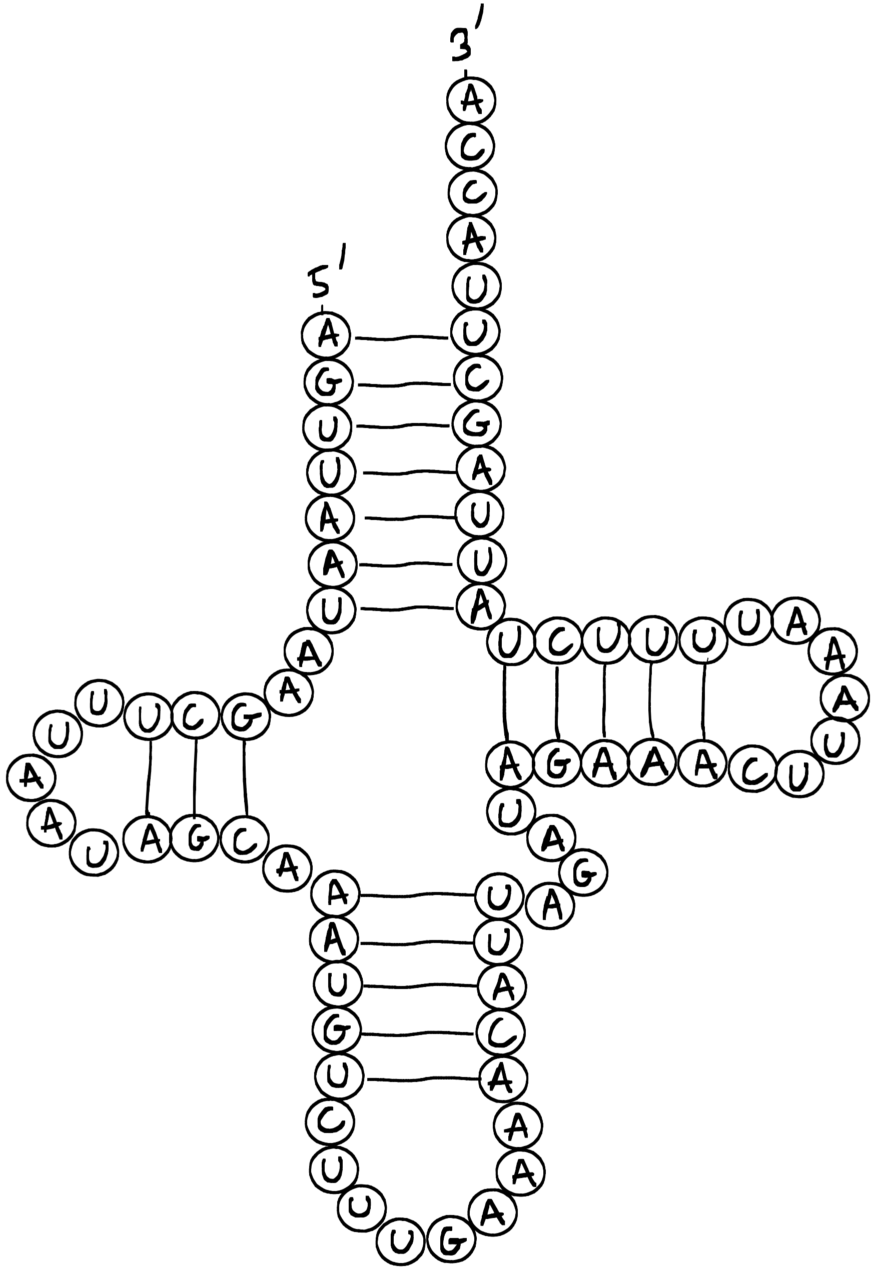


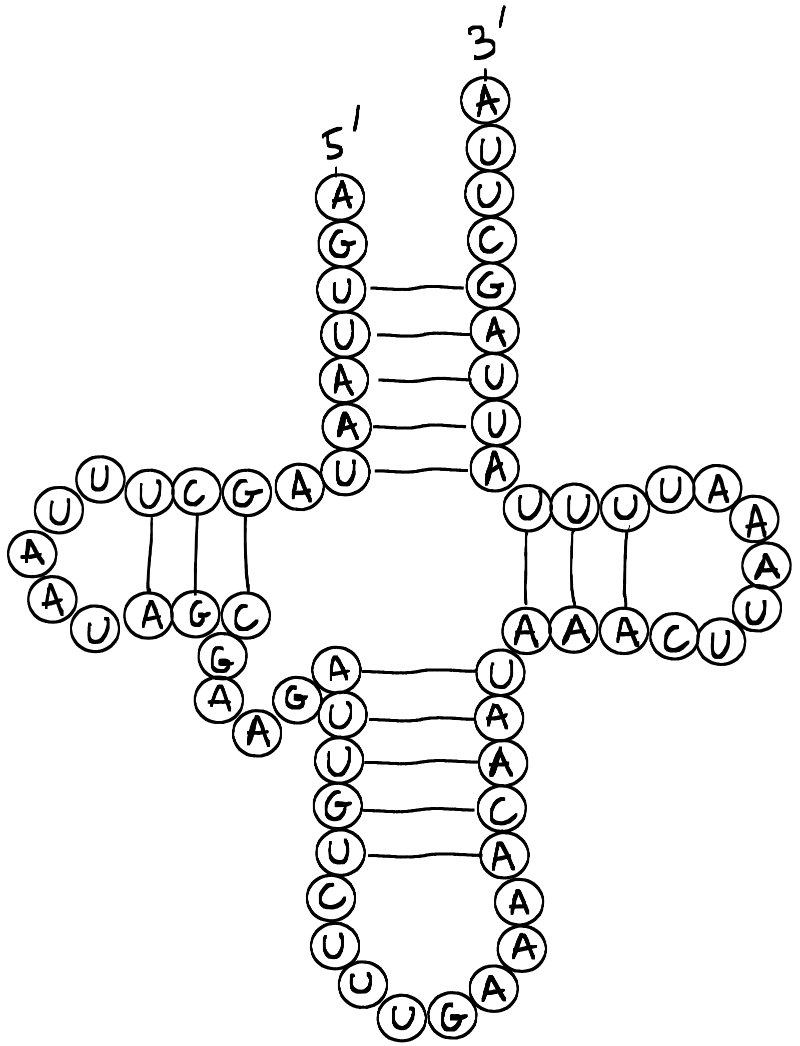
FROM SECONDARY STRUCTURES TO TREES



FROM SECONDARY STRUCTURES TO TREES

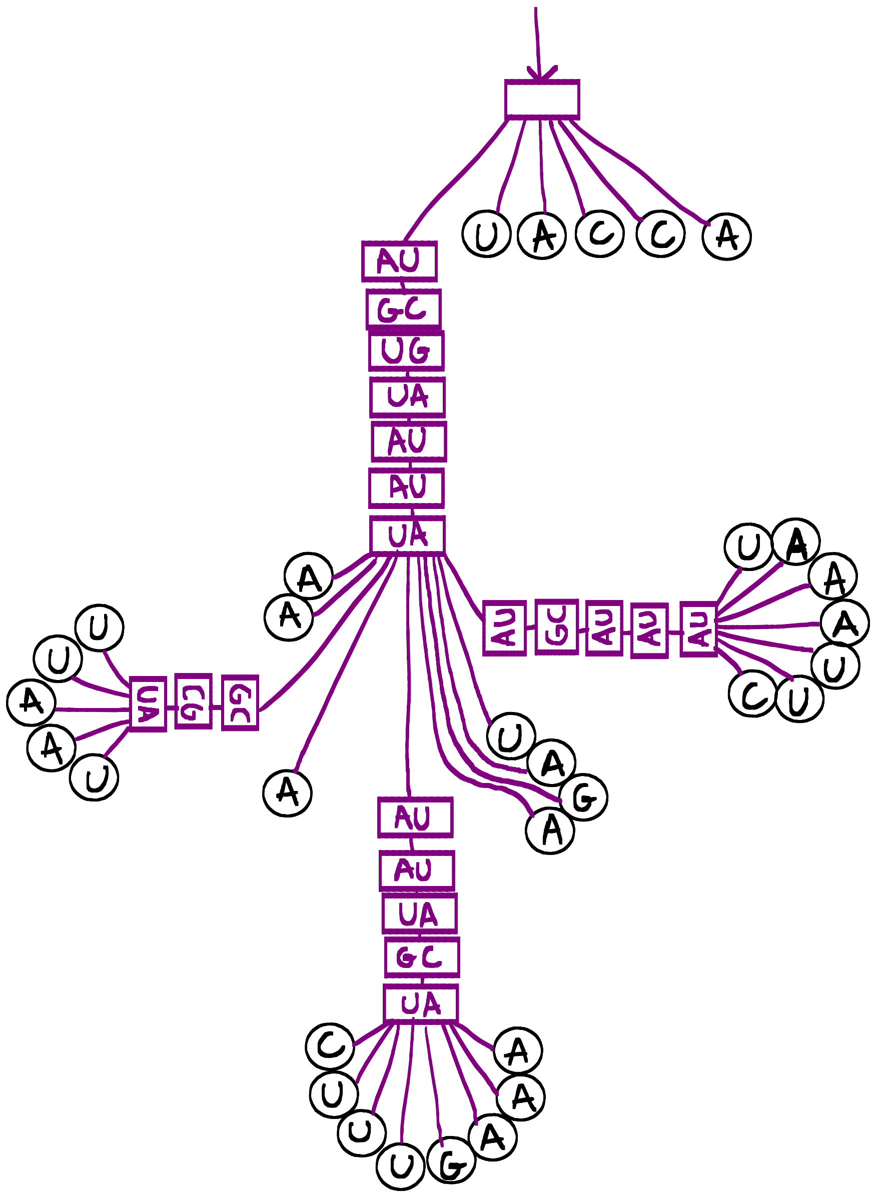
Objective: Align trees coming from RNA 2^{ary} structures

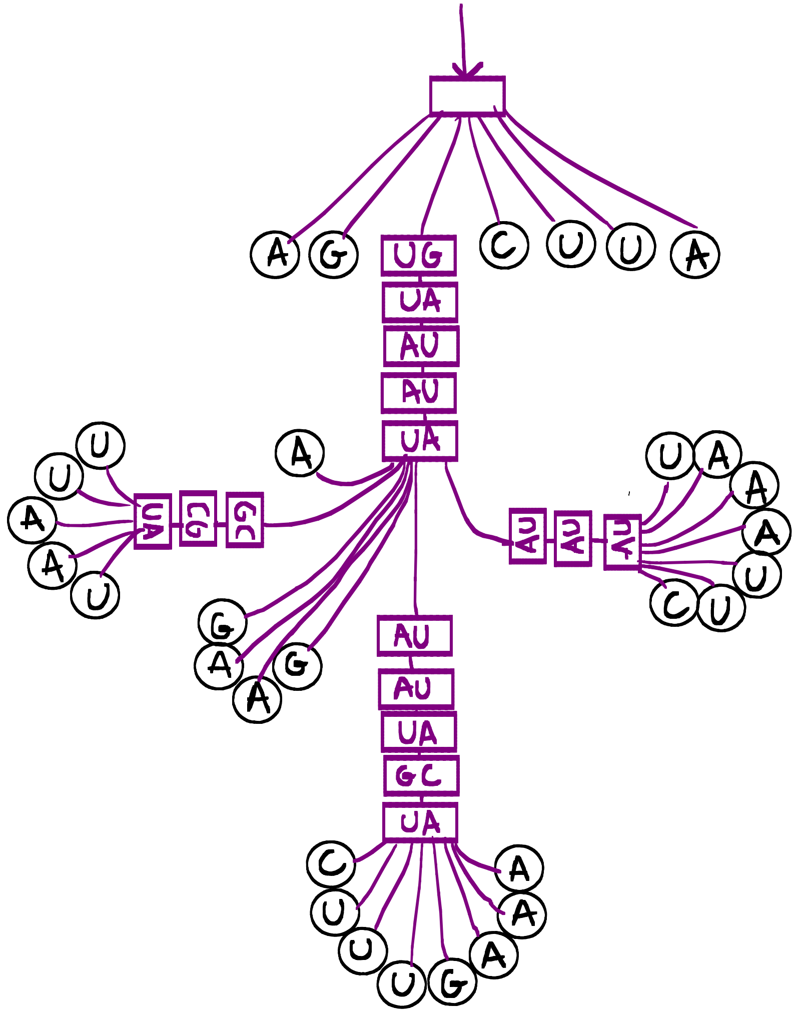




FROM SECONDARY STRUCTURES TO TREES

Objective: Align trees coming from RNA 2ary structures





SEQUENCE ALIGNMENT

super sequence = word on $\Sigma \times \Sigma \oplus \Sigma \times \{-\} \oplus \{-\} \times \Sigma$

$$\underset{\substack{\uparrow\\ \text{match}}}{\begin{pmatrix}A\\A\end{pmatrix}}\begin{pmatrix}U\\-\end{pmatrix}\begin{pmatrix}-\\C\end{pmatrix}\underset{\substack{\uparrow\\ \text{insertion}}}{\begin{pmatrix}U\\-\end{pmatrix}}\begin{pmatrix}C\\C\end{pmatrix}\begin{pmatrix}-\\A\end{pmatrix}\underset{\substack{\uparrow\\ \text{deletion}}}{\begin{pmatrix}-\\U\end{pmatrix}}\begin{pmatrix}G\\G\end{pmatrix}\begin{pmatrix}A\\A\end{pmatrix}\begin{pmatrix}U\\U\end{pmatrix}\underbrace{\begin{pmatrix}A\\U\end{pmatrix}\begin{pmatrix}C\\A\end{pmatrix}}_{\substack{\uparrow\\ \text{mismatch}}}$$

SEQUENCE ALIGNMENT

super sequence = word on $\Sigma\times\Sigma \oplus \Sigma\times\{-\} \oplus \{-\}\times\Sigma$

$$\begin{pmatrix}A\\A\end{pmatrix}\begin{pmatrix}U\\-\end{pmatrix}\begin{pmatrix}-\\C\end{pmatrix}\begin{pmatrix}U\\-\end{pmatrix}\begin{pmatrix}C\\C\end{pmatrix}\begin{pmatrix}-\\A\end{pmatrix}\begin{pmatrix}-\\U\end{pmatrix}\begin{pmatrix}G\\G\end{pmatrix}\begin{pmatrix}A\\A\end{pmatrix}\begin{pmatrix}U\\U\end{pmatrix}\underbrace{\begin{pmatrix}A\\U\end{pmatrix}\begin{pmatrix}C\\A\end{pmatrix}}_{\text{mismatch}} \begin{matrix}\longrightarrow \text{AUUCGAUAC}\\ \longrightarrow \underset{\uparrow\ \text{projections}}{\text{ACCAUGAUUA}}\end{matrix}$$

↑ match ↑ insertion ↑ deletion

SEQUENCE ALIGNMENT

super sequence = word on $\Sigma\times\Sigma \oplus \Sigma\times\{-\} \oplus \{-\}\times\Sigma$

$$\underset{\overset{\uparrow}{\text{match}}}{\begin{pmatrix}A\\A\end{pmatrix}}\begin{pmatrix}U\\-\end{pmatrix}\begin{pmatrix}-\\C\end{pmatrix}\underset{\overset{\uparrow}{\text{insertion}}}{\begin{pmatrix}U\\-\end{pmatrix}}\begin{pmatrix}C\\C\end{pmatrix}\begin{pmatrix}-\\A\end{pmatrix}\underset{\overset{\uparrow}{\text{deletion}}}{\begin{pmatrix}-\\U\end{pmatrix}}\begin{pmatrix}G\\G\end{pmatrix}\begin{pmatrix}A\\A\end{pmatrix}\begin{pmatrix}U\\U\end{pmatrix}\underset{\overset{\uparrow}{\text{mismatch}}}{\underbrace{\begin{pmatrix}A\\U\end{pmatrix}\begin{pmatrix}C\\A\end{pmatrix}}}\begin{matrix}\longrightarrow \text{AUUCGAUAC}\\ \longrightarrow \underset{\overset{\uparrow}{\text{projections}}}{\text{ACCAUGAUUA}}\end{matrix}$$

Given two sequences S_1 and S_2,

alignment between S_1 and S_2 = supersequence with projections S_1 and S_2

cost of an alignment = nb of insertions + deletions + mismatches

SEQUENCE ALIGNMENT

super sequence = word on $\Sigma\times\Sigma \oplus \Sigma\times\{-\} \oplus \{-\}\times\Sigma$

$$\underset{\substack{\uparrow\\ \text{match}}}{\binom{A}{A}}\binom{U}{-}\binom{-}{C}\underset{\substack{\uparrow\\ \text{insertion}}}{\binom{U}{-}}\binom{C}{C}\binom{-}{A}\underset{\substack{\uparrow\\ \text{deletion}}}{\binom{-}{U}}\binom{G}{G}\binom{A}{A}\binom{U}{U}\underbrace{\binom{A}{U}\binom{C}{A}}_{\substack{\uparrow\\ \text{mismatch}}}\begin{array}{l}\longrightarrow \text{AUUCGAUAC}\\ \longrightarrow \underset{\substack{\uparrow\\ \text{projections}}}{\text{ACCAUGAUUA}}\end{array}$$

Given two sequences S_1 and S_2,

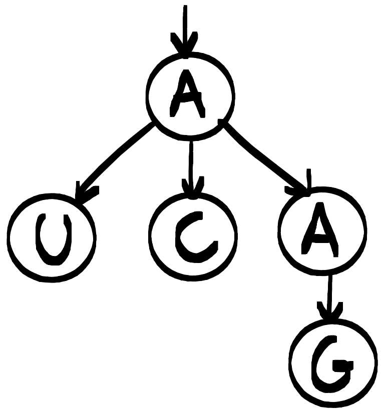
alignment between S_1 and S_2 = supersequence with projections S_1 and S_2

cost of an alignment = nb of insertions + deletions + mismatches

Remark: sequence alignment = sequence edition

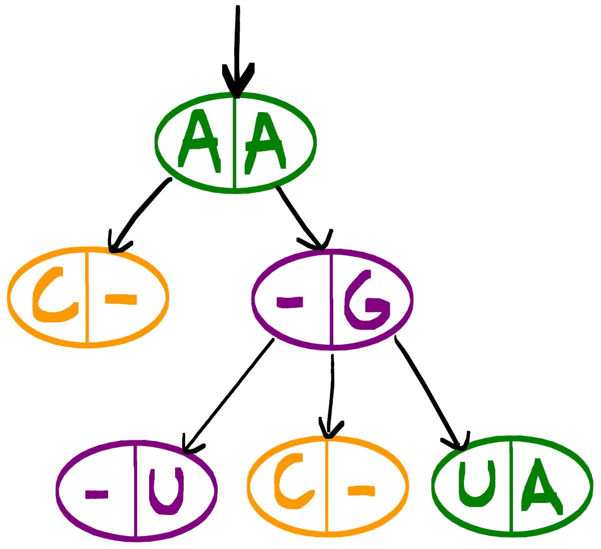
Trees and Supertrees

Trees are plane, rooted, and vertices are labeled by an alphabet Σ.

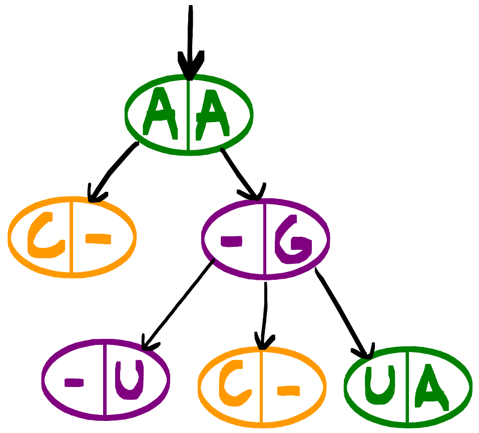


Supertree = tree with 3 types of vertices:

- X | Y (mis)match
- X | - insertion
- - | Y deletion



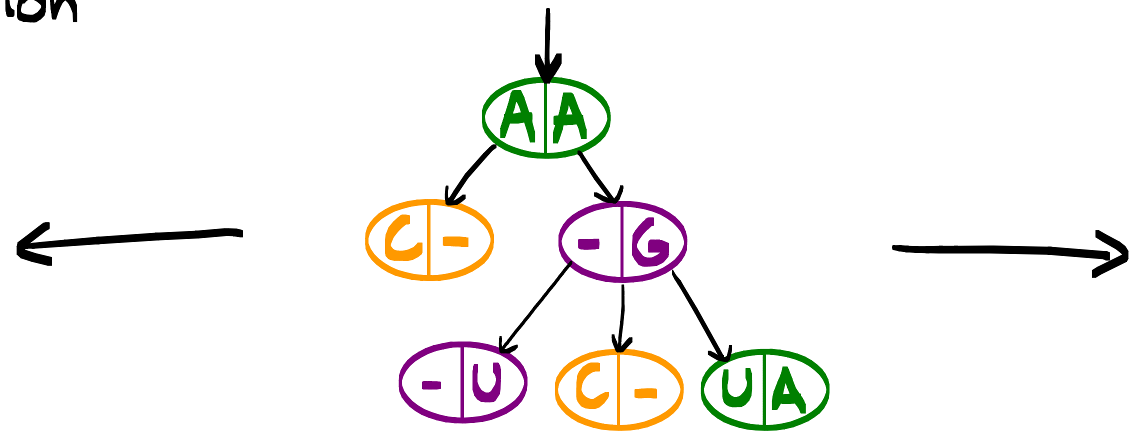
TREE ALIGNMENTS



TREE ALIGNMENTS

first projection

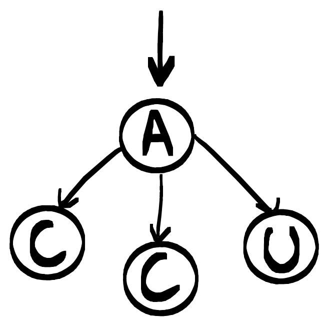
second projection

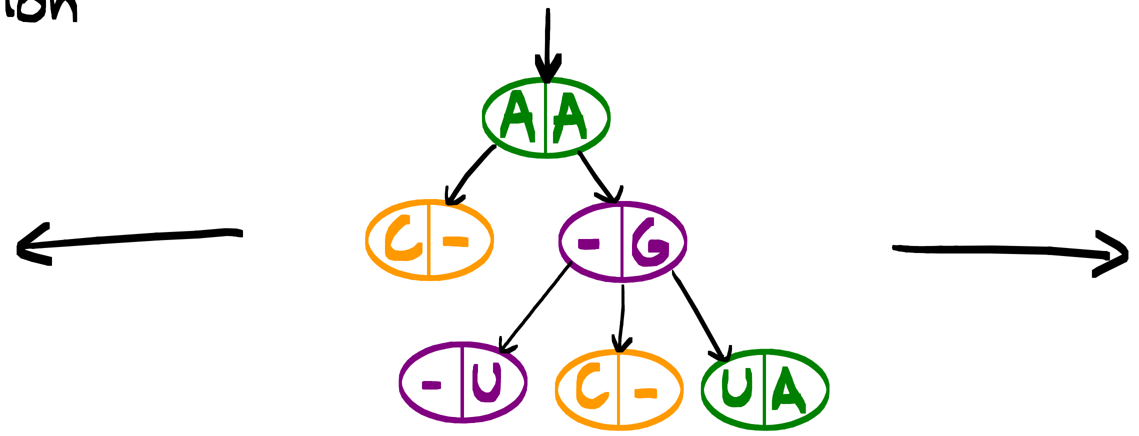


TREE ALIGNMENTS

first projection

second projection



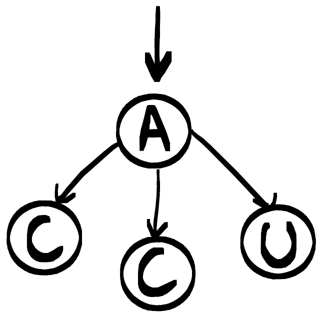


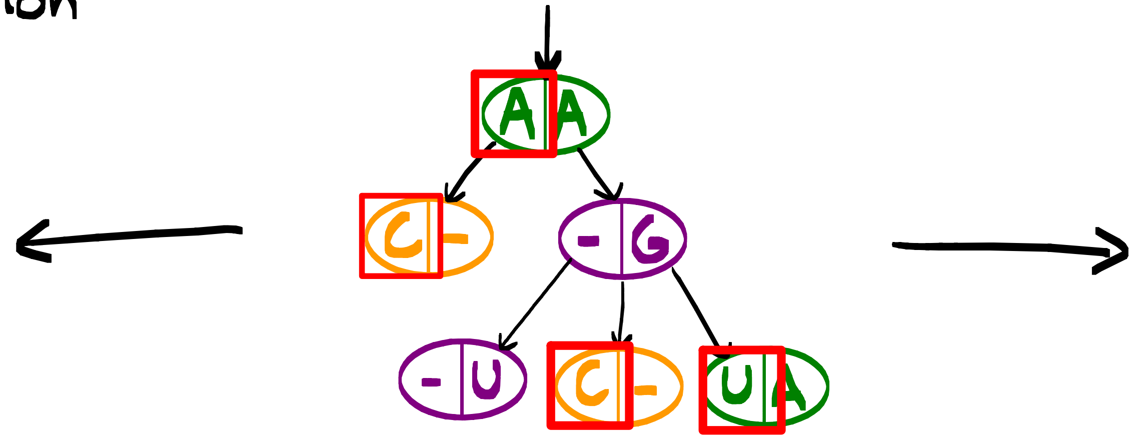
Keep left letters

TREE ALIGNMENTS

first projection

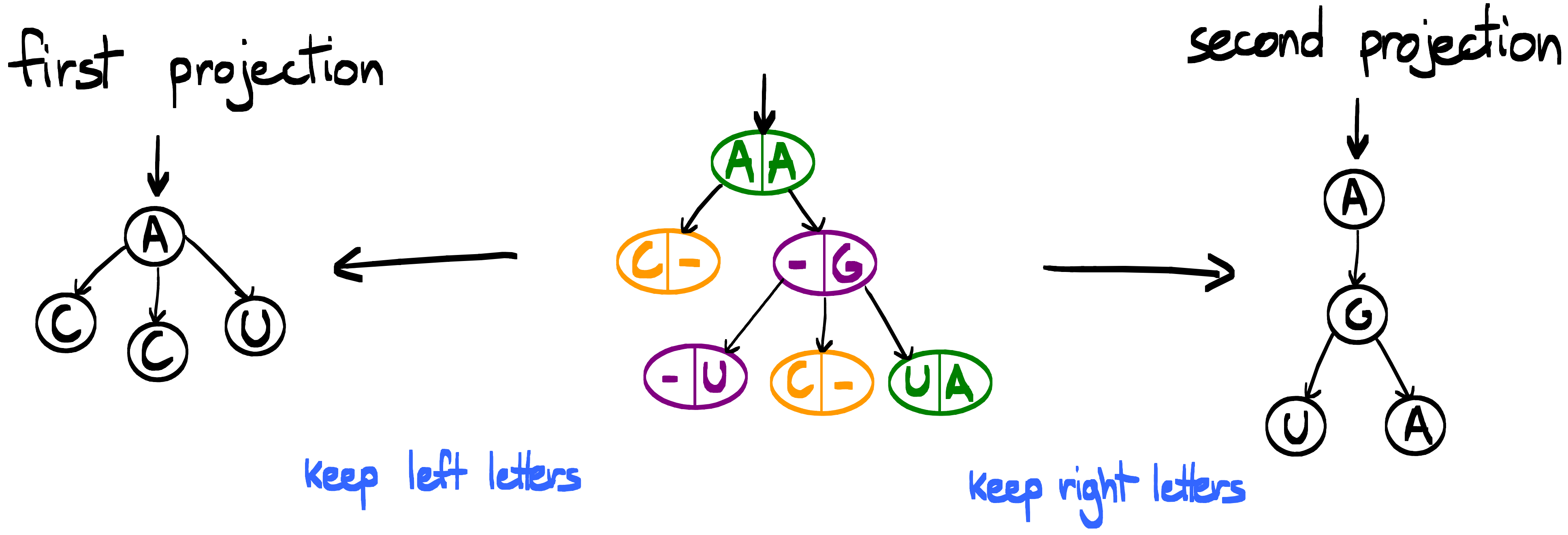
second projection



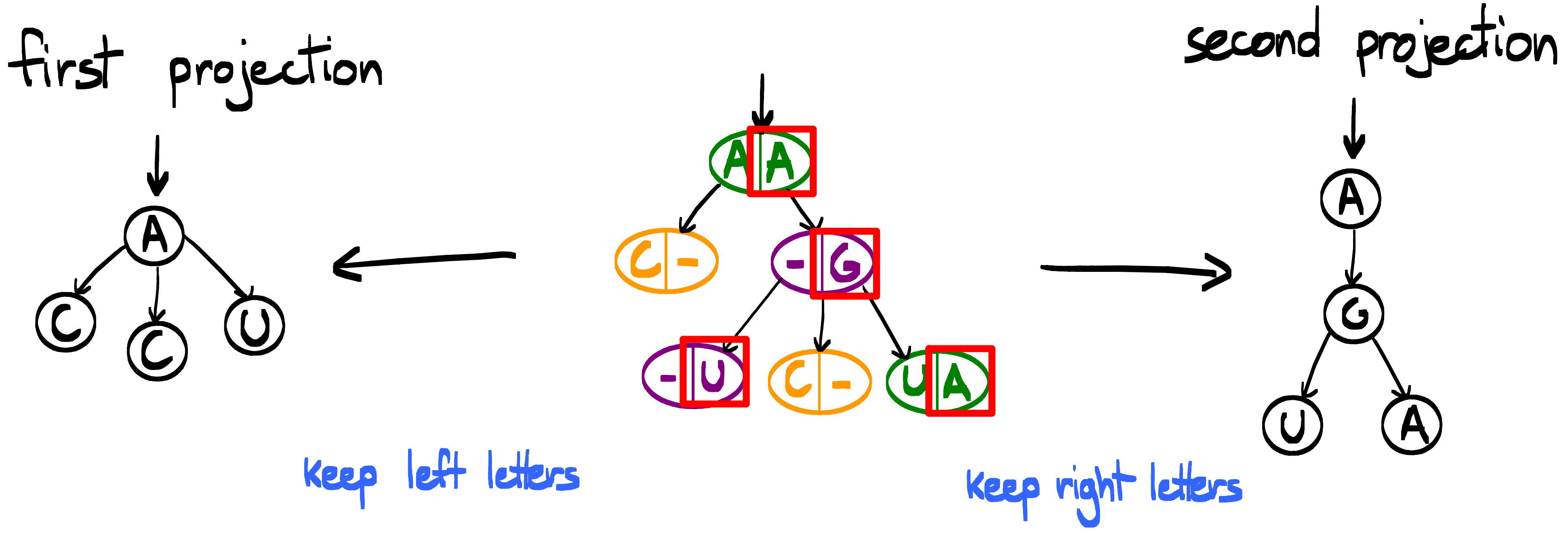


keep left letters

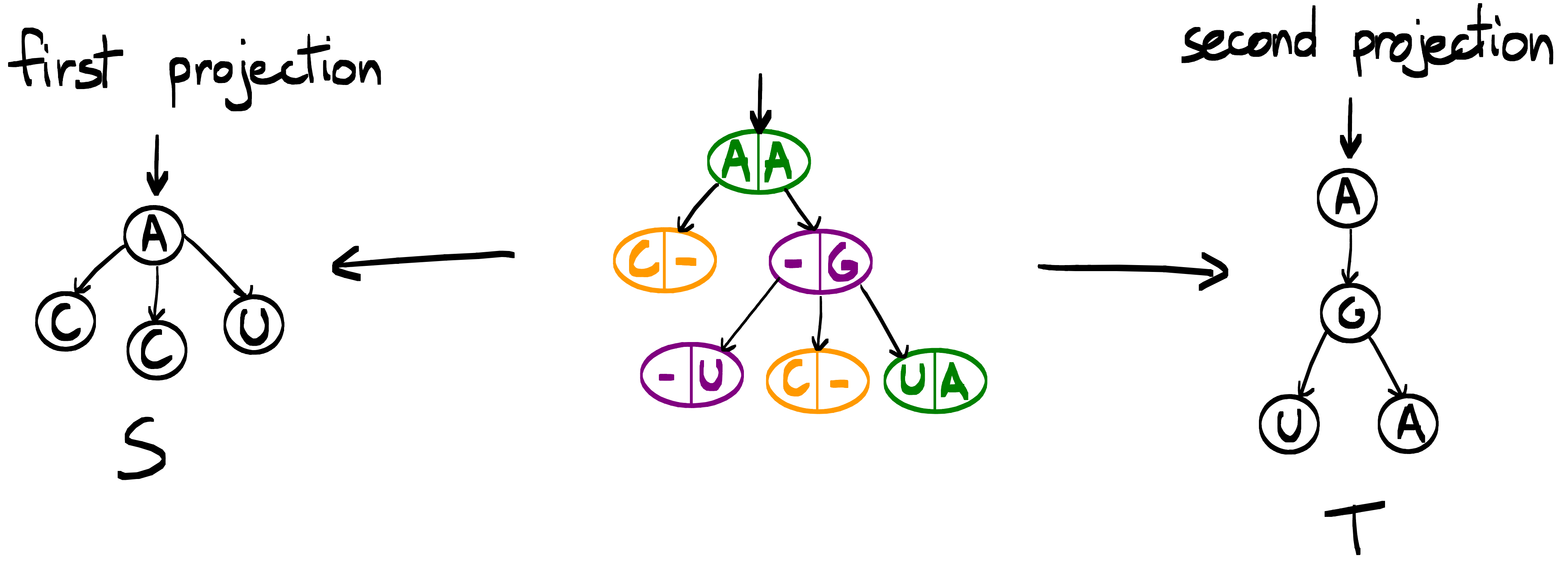
TREE ALIGNMENTS



TREE ALIGNMENTS

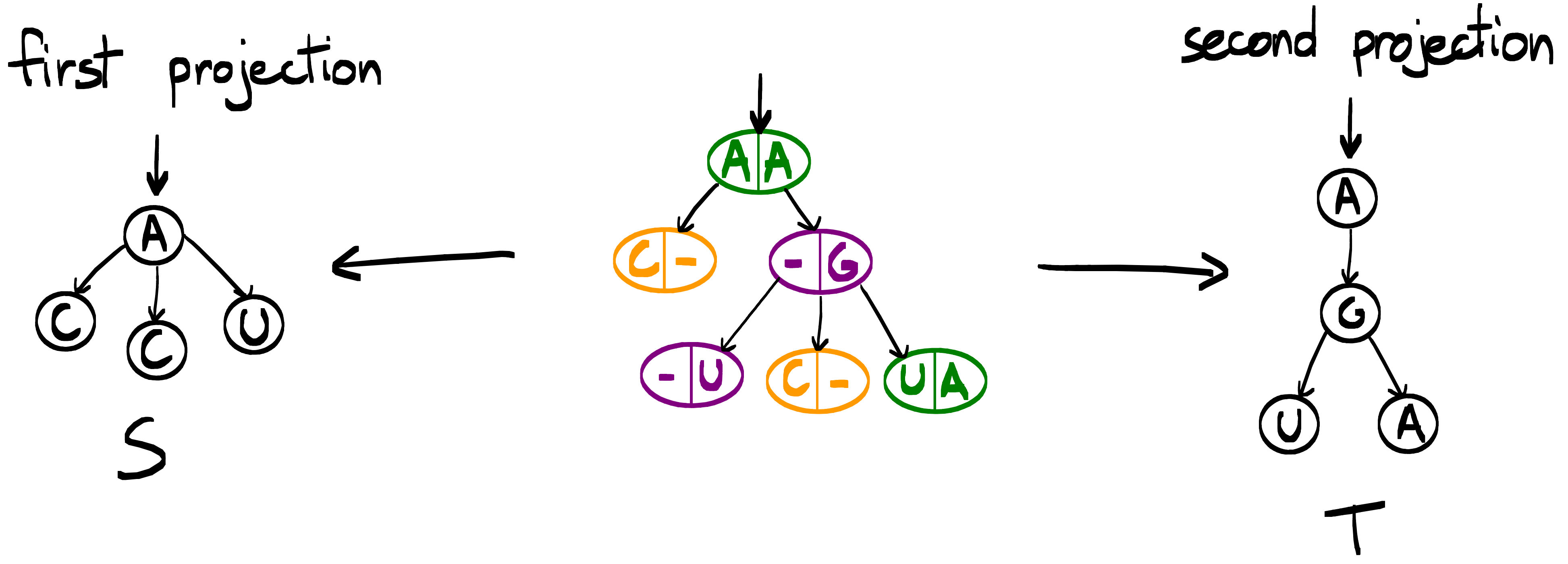


TREE ALIGNMENTS



Given two trees S and T,
<u>alignment</u> between S and T = supertree whose projections are S and T.

TREE ALIGNMENTS



Given two trees S and T,

alignment between S and T = supertree whose projections are S and T.

cost of an alignment = nb of insertions + deletions + mismatches

Remark: tree alignment ≠ tree edition

CONNECTION WITH SEQUENCE ALIGNMENTS

Tree alignments generalize sequence alignments.

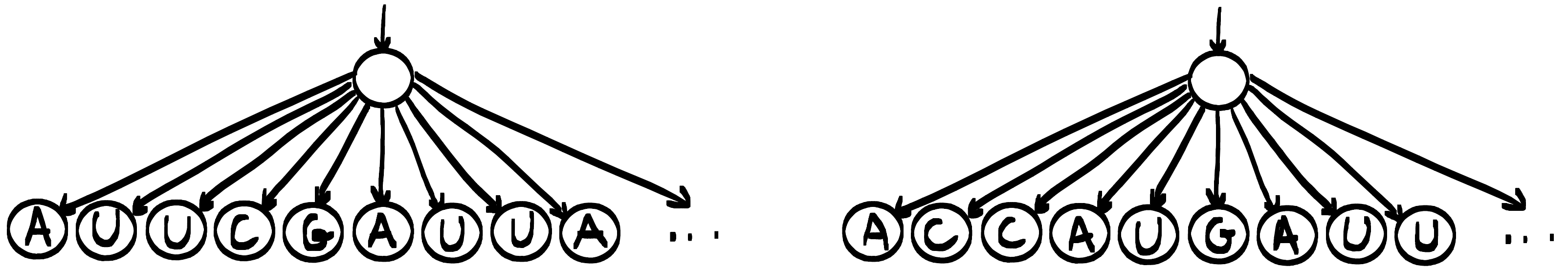
SEQUENCE

AUUCGAUUA...   ACCAUGAUUA...

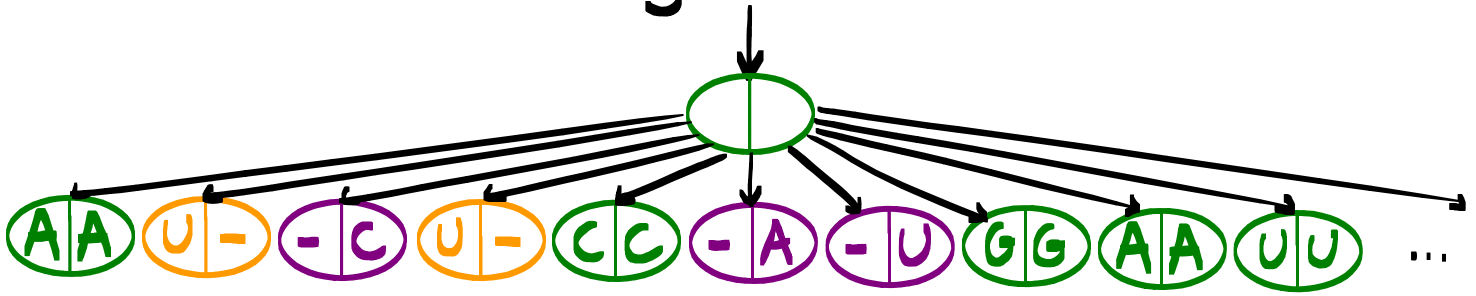
alignment:

$$\binom{A}{A}\binom{U}{-}\binom{-}{C}\binom{U}{-}\binom{C}{C}\binom{-}{A}\binom{-}{U}\binom{G}{G}\binom{A}{A}\binom{U}{U}\binom{U}{U}\binom{A}{A}\cdots$$

TREE



alignment:



OPTIMAL ALIGNMENT

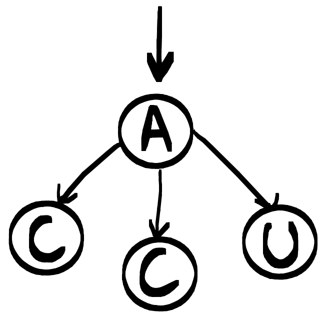
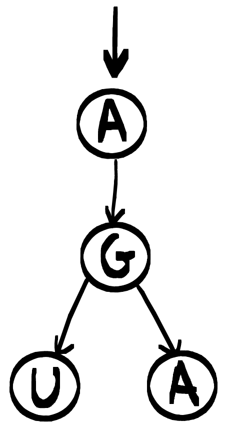
Classical problem: Given S and T, find one optimal alignment between S and T.

Solvable by Dynamic Programming:

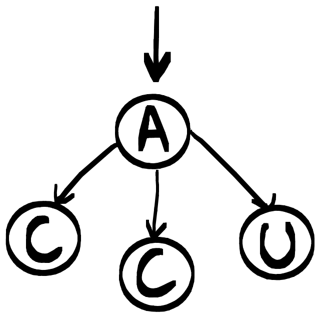
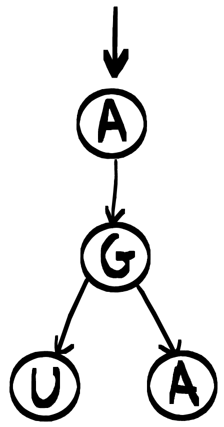
Worst case time $O(n^4)$ [Jiang, Wang, Zhang]

Average time $O(n^2)$ [Herrbach, Denise, Dulucq]

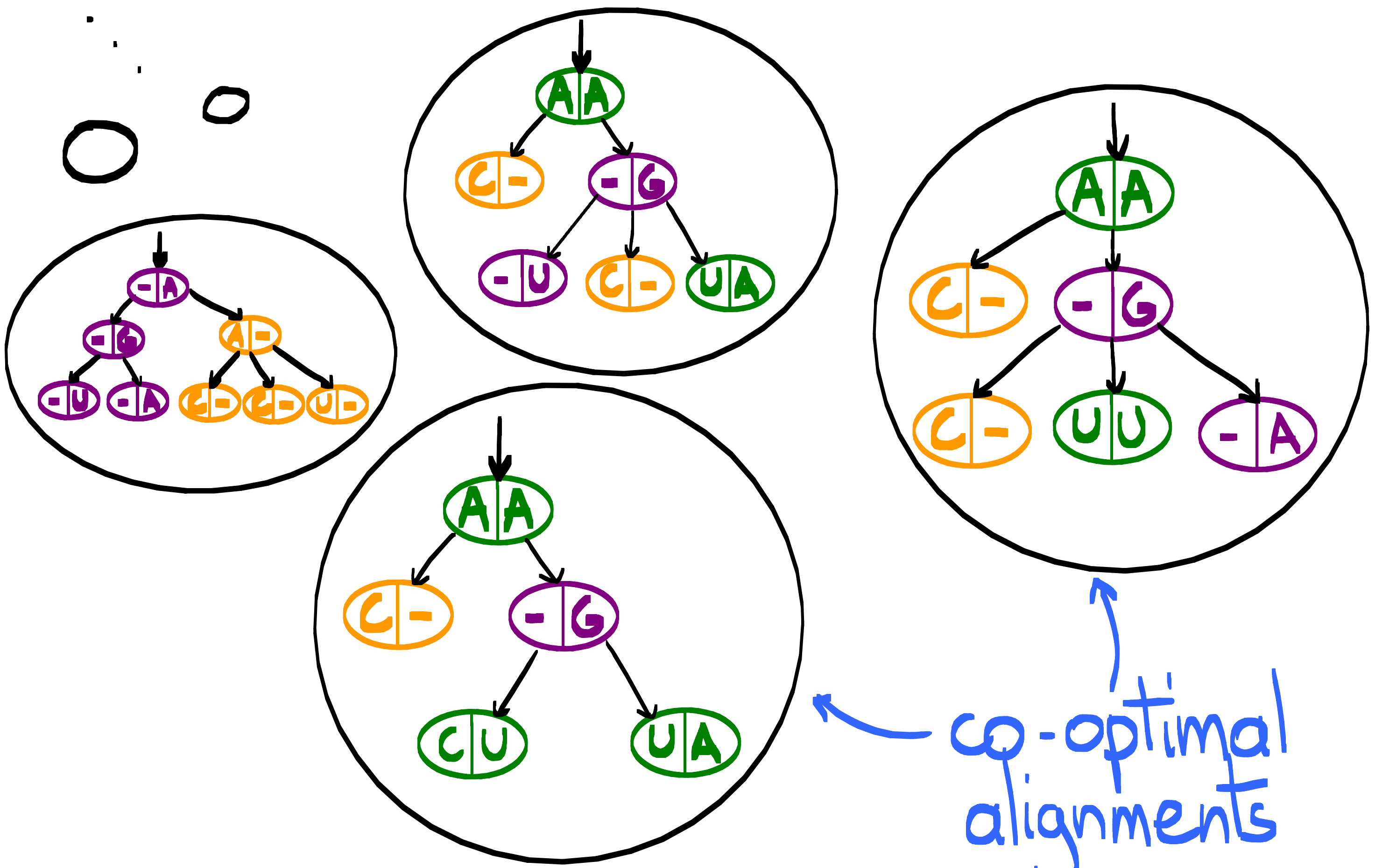
SPACE OF ALIGNMENTS

Which alignment between  and  is the most likely?

SPACE OF ALIGNMENTS

Which alignment between  and 

is the most likely?



SPACE OF ALIGNMENTS

Why finding one optimal alignment may be inadequate :

- Co-optimal alignments can be very different.
  (see for instance [Vingron, Argos, 1990])

- Exploring the space of alignments enables the detection of high probability features.

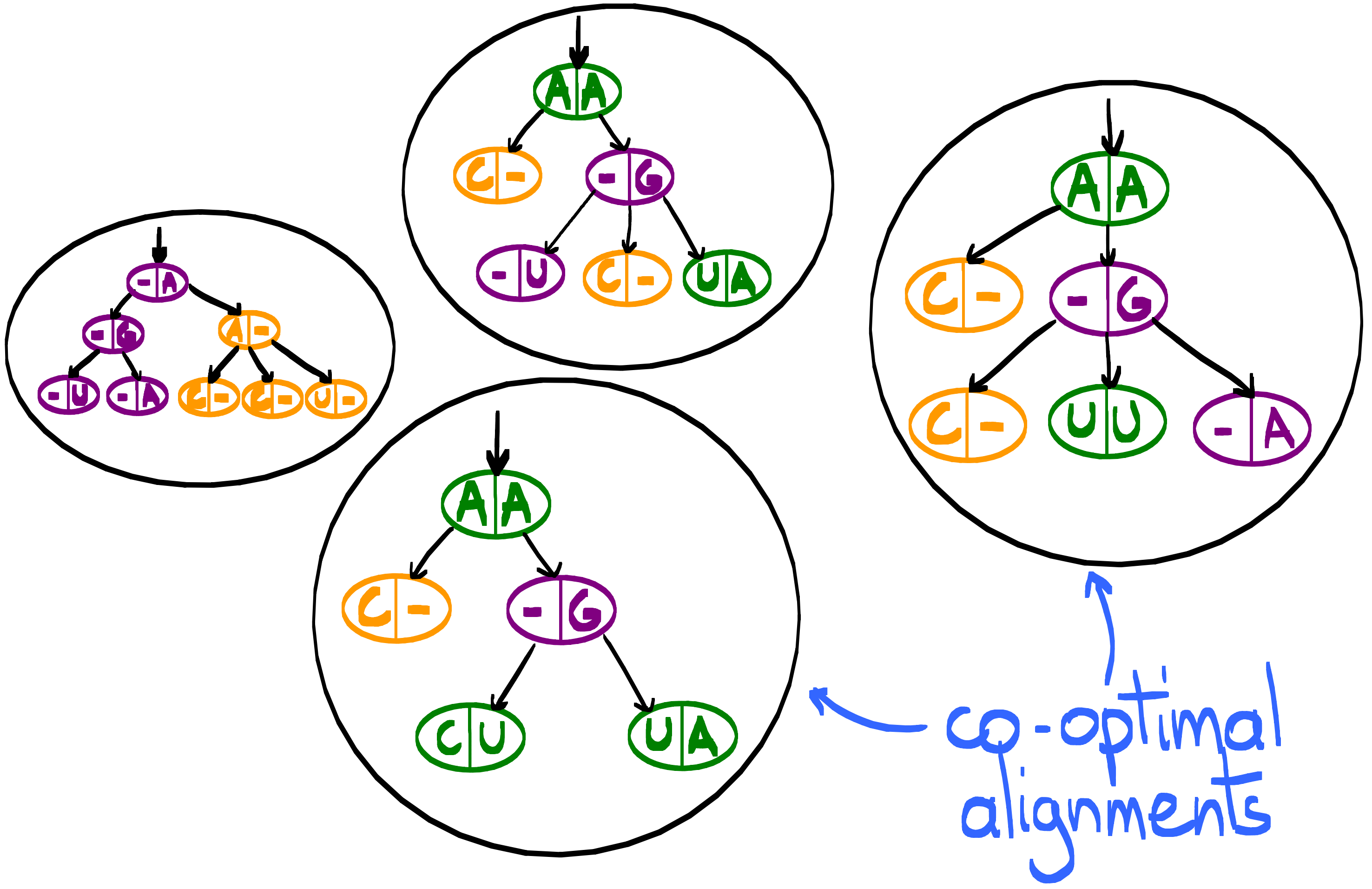
SPACE OF ALIGNMENTS

Objective: Sampling alignments under the Gibbs-Boltzmann probability distribution.

probability of an alignment A

$$\propto e^{-\frac{\text{cost}(A)}{K}}$$

(Gibbs-Boltzmann distribution)



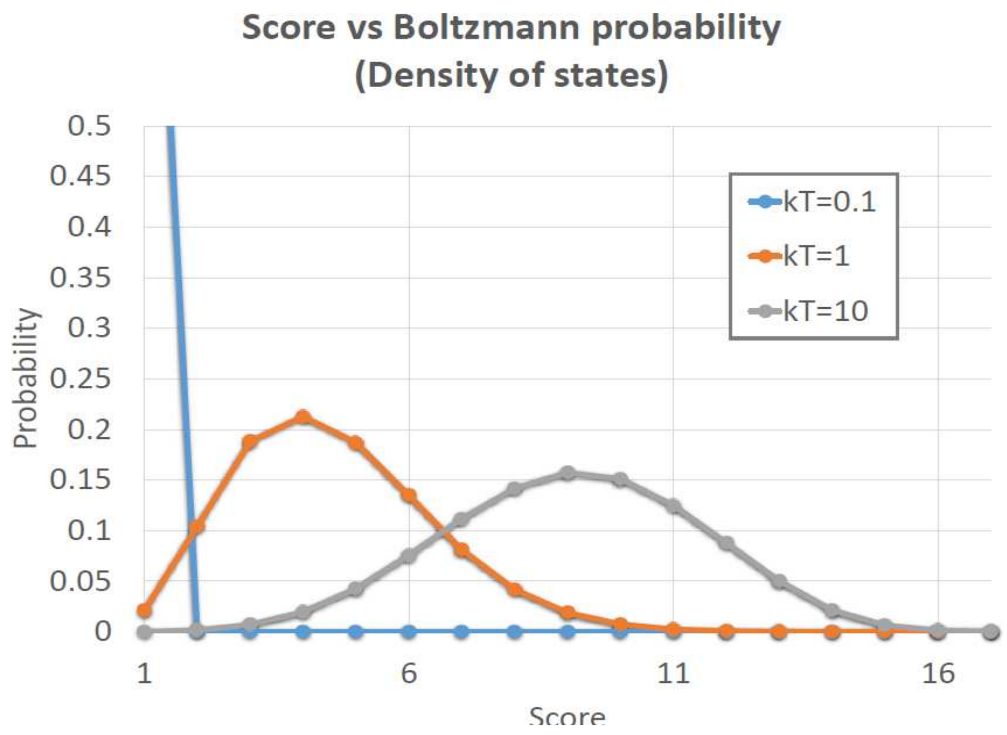
SPACE OF ALIGNMENTS

Objective: Sampling alignments under the Gibbs-Boltzmann probability distribution.

probability of an alignment A

$$\propto e^{-\frac{\mathrm{cost}(A)}{K}}$$

(Gibbs-Boltzmann distribution)



K=0 : Uniform distribution over optimal alignments.

K=+∞ : Uniform distribution over all alignments.

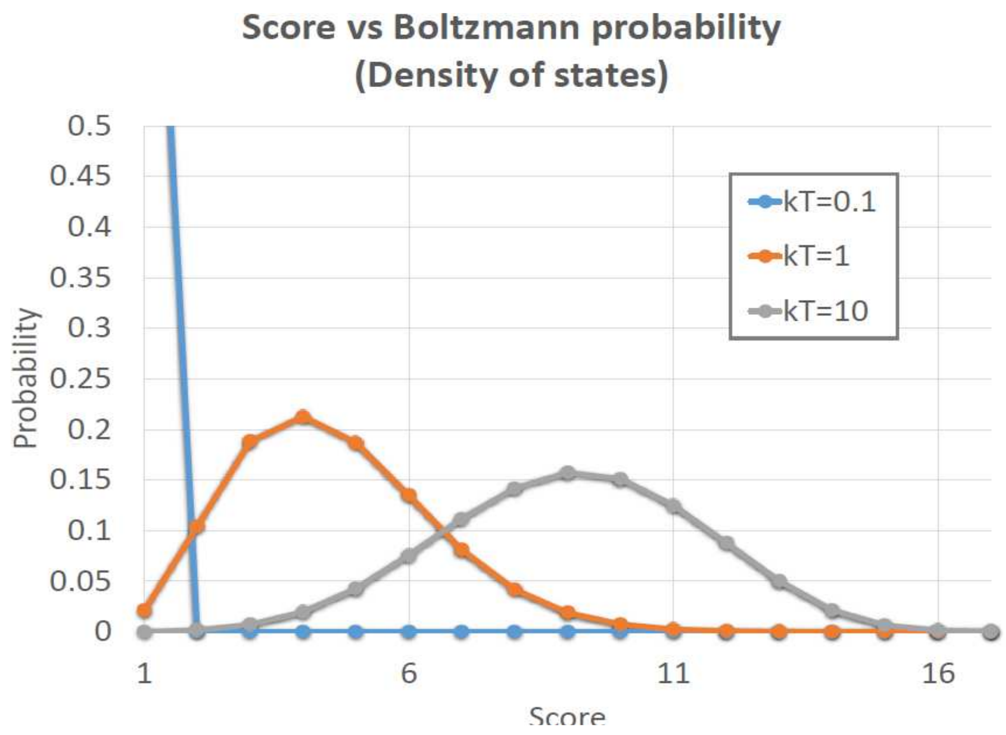
SPACE OF ALIGNMENTS

Objective: Sampling alignments under the Gibbs-Boltzmann probability distribution.

??. probability of an alignment A

$$\propto e^{-\frac{\text{cost}(A)}{K}}$$

(Gibbs-Boltzmann distribution)



$\underline{K=0}$: Uniform distribution over <u>optimal</u> alignments.

$\underline{K=+\infty}$: Uniform distribution over <u>all</u> alignments.

AMBIGUITY OF ALIGNMENTS

For sequences,

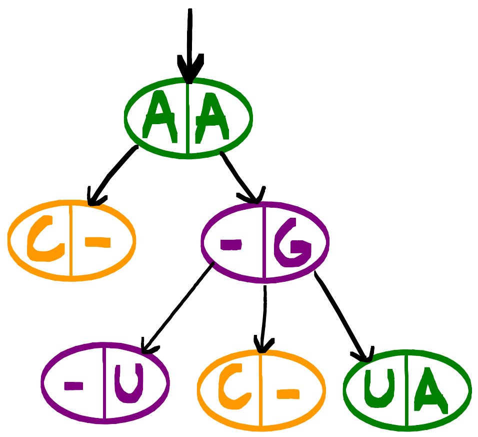
$$\binom{A}{A}\binom{U}{-}\binom{-}{C}\binom{U}{-}\binom{C}{C}\binom{-}{A}\binom{-}{U}\binom{G}{G}\binom{A}{A}\binom{U}{U}\binom{U}{U}\binom{A}{A}$$

is the same alignment as

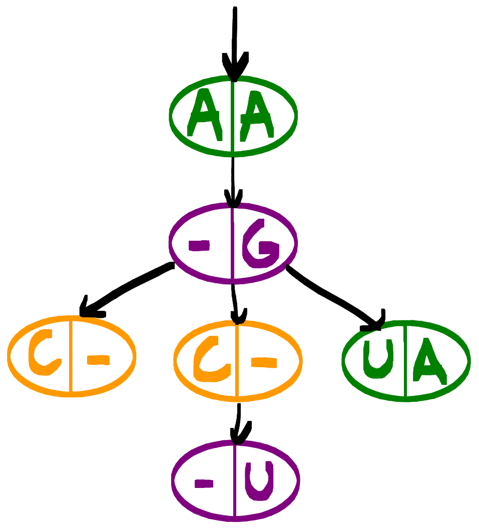
$$\binom{A}{A}\binom{-}{C}\binom{U}{-}\binom{U}{-}\binom{C}{C}\binom{-}{A}\binom{-}{U}\binom{G}{G}\binom{A}{A}\binom{U}{U}\binom{U}{U}\binom{A}{A}$$

AMBIGUITY OF ALIGNMENTS

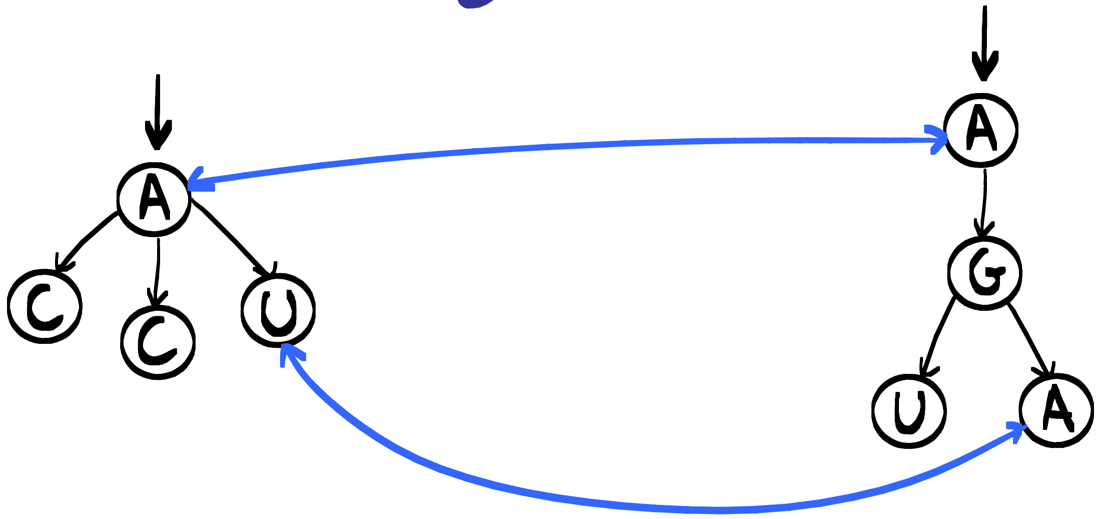
For trees,



and

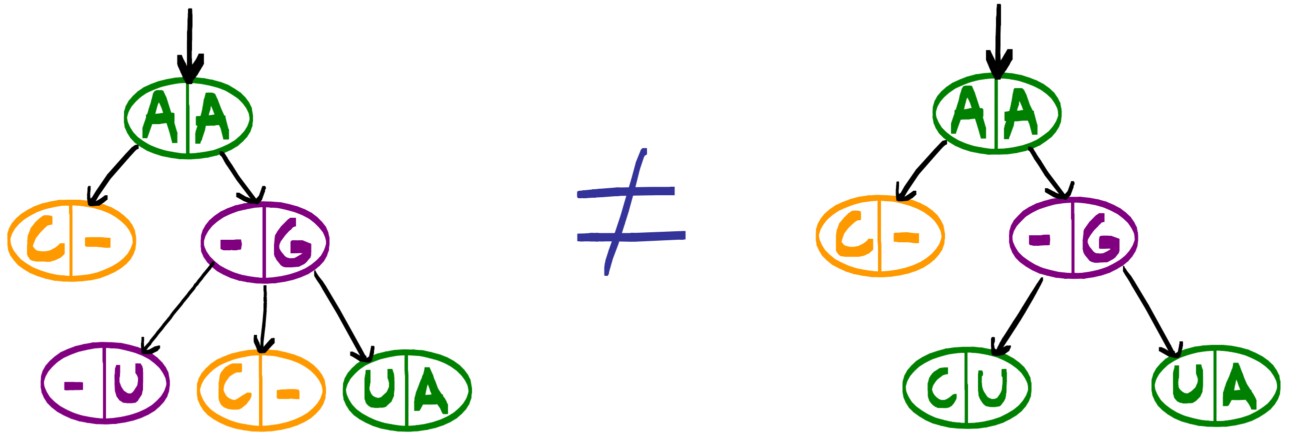


induce the same alignment between

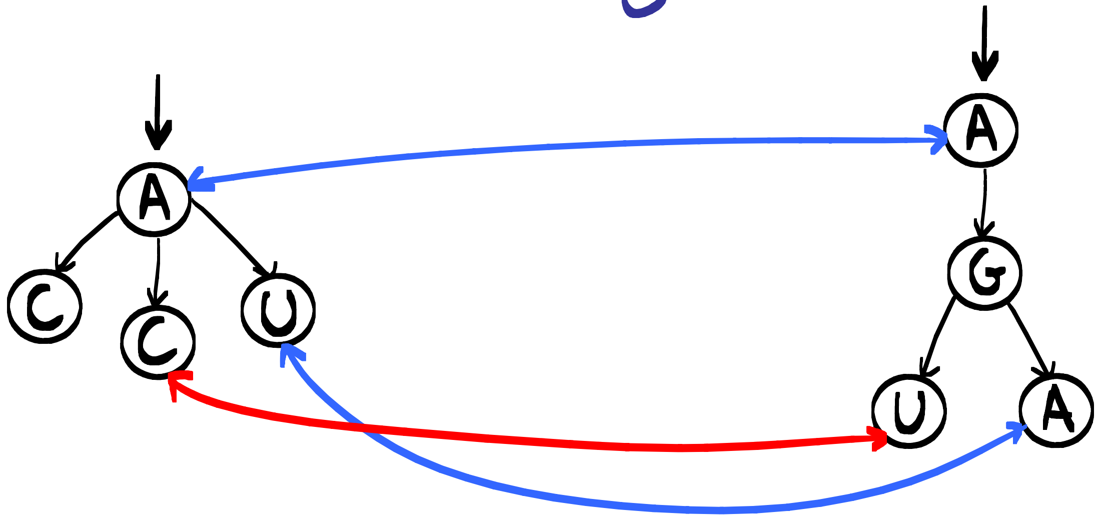


AMBIGUITY OF ALIGNMENTS

The two supertrees



do not induce the same alignment between the trees



PROBLEM RAISED BY THE AMBIGUITY

For sequences, we can deal with the ambiguity by defining canonical alignments.

Ex :

$$\begin{pmatrix}A\\A\end{pmatrix}\begin{pmatrix}U\\-\end{pmatrix}\begin{pmatrix}-\\C\end{pmatrix}\begin{pmatrix}U\\-\end{pmatrix}\begin{pmatrix}C\\C\end{pmatrix}\begin{pmatrix}-\\A\end{pmatrix}\begin{pmatrix}-\\U\end{pmatrix}\begin{pmatrix}G\\G\end{pmatrix}\begin{pmatrix}A\\A\end{pmatrix}\begin{pmatrix}U\\U\end{pmatrix}\begin{pmatrix}U\\U\end{pmatrix}\begin{pmatrix}A\\A\end{pmatrix}$$

PROBLEM RAISED BY THE AMBIGUITY

For sequences, we can deal with the ambiguity by defining canonical alignments.

Ex :

$$\begin{pmatrix}A\\A\end{pmatrix}\begin{pmatrix}U\\-\end{pmatrix}\begin{pmatrix}U\\-\end{pmatrix}\begin{pmatrix}-\\C\end{pmatrix}\begin{pmatrix}C\\C\end{pmatrix}\begin{pmatrix}-\\A\end{pmatrix}\begin{pmatrix}-\\U\end{pmatrix}\begin{pmatrix}G\\G\end{pmatrix}\begin{pmatrix}A\\A\end{pmatrix}\begin{pmatrix}U\\U\end{pmatrix}\begin{pmatrix}U\\U\end{pmatrix}\begin{pmatrix}A\\A\end{pmatrix}$$

Insertions before Deletions.

PROBLEM RAISED BY THE AMBIGUITY

For sequences, we can deal with the ambiguity by defining canonical alignments.

Ex :

$$\binom{A}{A}\binom{U}{-}\binom{U}{-}\binom{-}{C}\binom{C}{C}\binom{-}{A}\binom{-}{U}\binom{G}{G}\binom{A}{A}\binom{U}{U}\binom{U}{U}\binom{A}{A}$$

Insertions before Deletions.

For trees, it is much more complicated!

PROBLEM RAISED BY THE AMBIGUITY

For sequences, we can deal with the ambiguity by defining canonical alignments.

Ex :

$$\begin{pmatrix}A\\A\end{pmatrix}\begin{pmatrix}U\\-\end{pmatrix}\begin{pmatrix}U\\-\end{pmatrix}\begin{pmatrix}-\\C\end{pmatrix}\begin{pmatrix}C\\C\end{pmatrix}\begin{pmatrix}-\\A\end{pmatrix}\begin{pmatrix}-\\U\end{pmatrix}\begin{pmatrix}G\\G\end{pmatrix}\begin{pmatrix}A\\A\end{pmatrix}\begin{pmatrix}U\\U\end{pmatrix}\begin{pmatrix}U\\U\end{pmatrix}\begin{pmatrix}A\\A\end{pmatrix}$$

Insertions before Deletions.

For trees, it is much more complicated!

Strategy: Build a context-free grammar that generates every alignment exactly once

GRAMMARS FOR SEQUENCE ALIGNMENTS

Ambiguous grammar:

$$\mathcal{S} \leftarrow \binom{x}{y}\boxed{\mathcal{S}} \oplus \binom{x}{-}\boxed{\mathcal{S}} \oplus \binom{-}{y}\boxed{\mathcal{S}}$$

Non-ambiguous grammar:

$$\mathcal{S} \leftarrow \binom{x}{y}\boxed{\mathcal{S}} \oplus \binom{x}{-}\boxed{\mathcal{S}} \oplus \binom{-}{y}\boxed{\mathcal{S}^D}$$

$$\mathcal{S}^D \leftarrow \binom{x}{y}\boxed{\mathcal{S}} \oplus \binom{-}{y}\boxed{\mathcal{S}^D}$$

A GRAMMAR FOR ALIGNMENTS

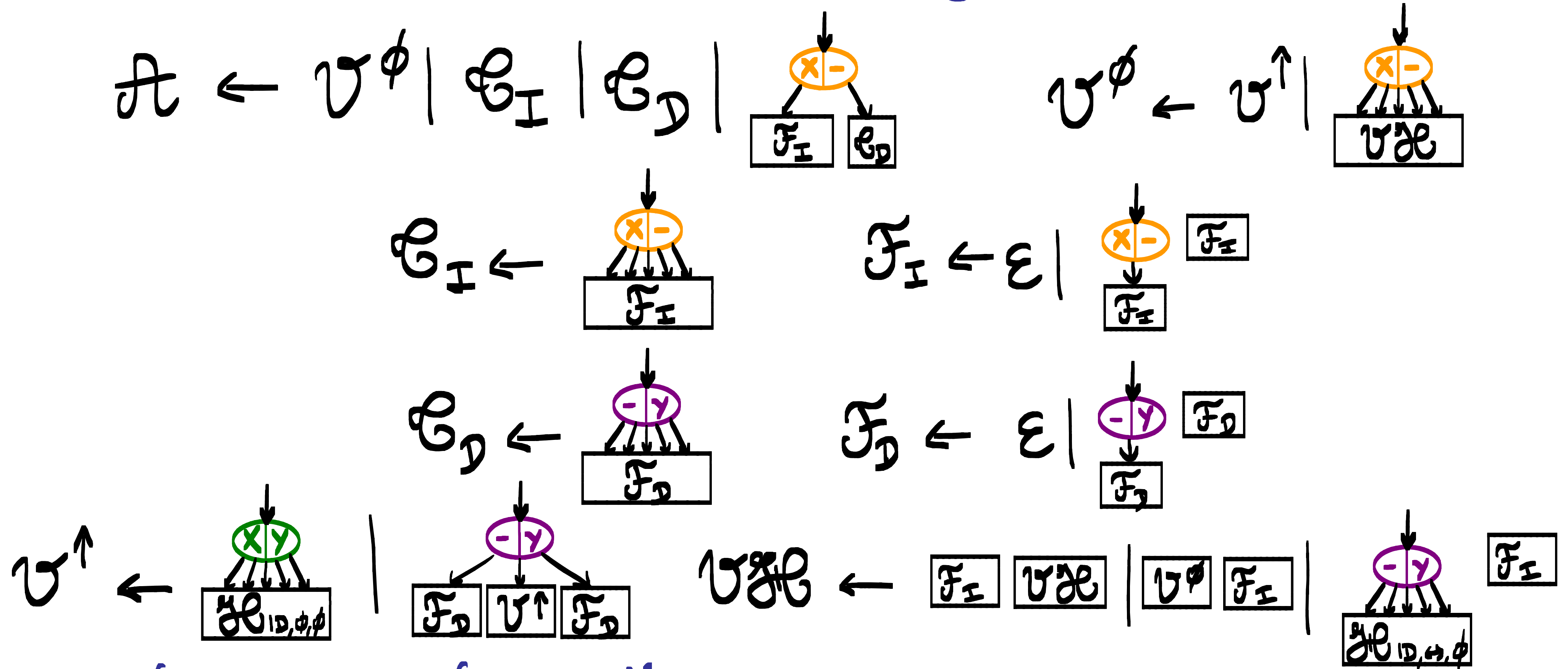
For trees, an ambiguous grammar can be derived from [Jiang, Wang, Zhang].

Our result:

<u>Theorem</u>: The set $\mathcal{A}$ generated by the following grammar contains every tree alignment exactly once.

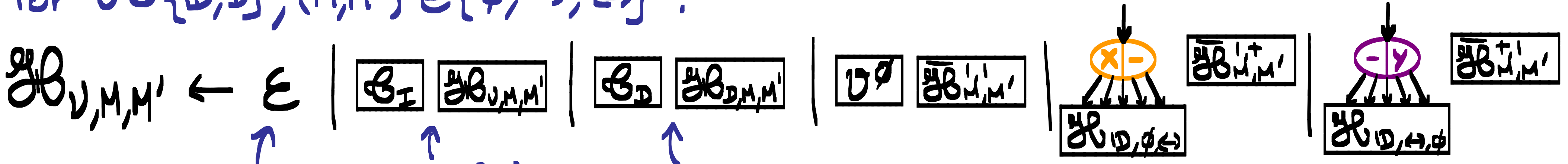
A GRAMMAR FOR ALIGNMENTS

Our (complicated) non-ambiguous grammar:



$\mathcal{A} \leftarrow \mathcal{V}^{\phi} \mid \mathcal{C}_I \mid \mathcal{C}_D \mid$
x|-
$\mathcal{F}_I$
$\mathcal{C}_D$
$\mathcal{V}^{\phi} \leftarrow \mathcal{V}^{\uparrow} \mid$
x|-
$\mathcal{VH}$
$\mathcal{C}_I \leftarrow$
x|-
$\mathcal{F}_I$
$\mathcal{F}_I \leftarrow \varepsilon \mid$
x|-
$\mathcal{F}_I$
$\mathcal{F}_I$
$\mathcal{C}_D \leftarrow$
-|y
$\mathcal{F}_D$
$\mathcal{F}_D \leftarrow \varepsilon \mid$
-|y
$\mathcal{F}_D$
$\mathcal{F}_D$
$\mathcal{V}^{\uparrow} \leftarrow$
x|y
$\mathcal{H}_{\mathbb{D},\phi,\phi}$
-|y
$\mathcal{F}_D$
$\mathcal{V}^{\uparrow}$
$\mathcal{F}_D$
$\mathcal{VH} \leftarrow$
$\mathcal{F}_I$
$\mathcal{VH}$
$\mathcal{V}^{\phi}$
$\mathcal{F}_I$
-|y
$\mathcal{H}_{\mathbb{D},\leftrightarrow,\phi}$
$\mathcal{F}_I$

For $\mathcal{V} \in \{\mathbb{D}, \mathcal{D}\}$, $(M,M') \in \{\phi, \rightarrow, \leftrightarrow\}^2$:



$\mathcal{HB}_{V,M,M'} \leftarrow \varepsilon$
$\mathcal{C}_I$
$\mathcal{HB}_{V,M,M'}$
$\mathcal{C}_D$
$\mathcal{HB}_{D,M,M'}$
$\mathcal{V}^{\phi}$
$\overline{\mathcal{HB}}^{i,j}_{M,M'}$
x|-
$\mathcal{H}_{\mathbb{D},\phi,\leftrightarrow}$
$\overline{\mathcal{HB}}^{i,+}_{M,M'}$
-|y
$\mathcal{H}_{\mathbb{D},\leftrightarrow,\phi}$
$\overline{\mathcal{HB}}^{+,j}_{M,M'}$

only if $(M,M')=(\phi,\phi)$

only if $V \neq \mathbb{D}$ and $M \neq \leftrightarrow$

only if $M' \neq \leftrightarrow$

no room for $\overline{\mathcal{HB}}^{i,j}_{M,M'}$...

APPLICATION 1. COUNTING.

Theorem There are on average $C \times 1.5^n$ alignments between two random trees of cumulative size n where $C = 0.299\ldots$

Corollary: A same alignment was repeated $\sim 0.875 \times 1.412^n$ times on average in Jiang et al.'s ambiguous grammar.

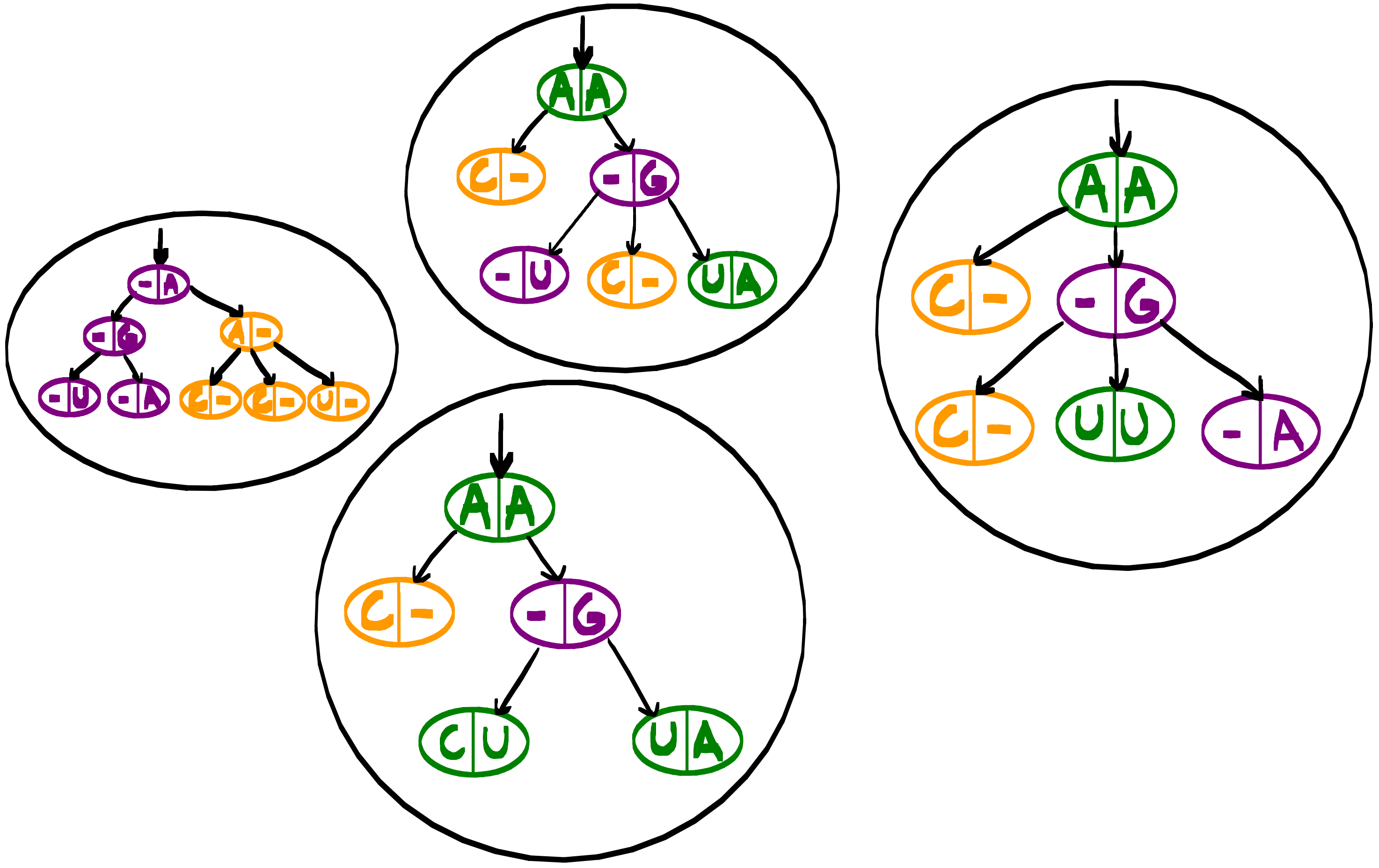
APPLICATION 2. SAMPLING

Objective: Sampling alignments under the Gibbs-Boltzmann probability distribution.

probability of an alignment A

$$\propto e^{-\frac{cost(A)}{K}}$$

(Gibbs-Boltzmann distribution)



APPLICATION 2. SAMPLING

Objective: Sampling alignments under the Gibbs-Boltzmann probability distribution.

Strategy:

→ Filter the grammar to obtain a new grammar that only generates alignments between two fixed trees S and T

→ Use dynamic programming.

SAMPLING

Theorem Let S and T be two trees of size n_1 and n_2. Sampling alignments between S and T under the Gibbs-Boltzmann distribution can be done with worst-case time and space complexities $O(n_1 n_2 (n_1+n_2)^2)$ and with average-case time and space complexities $O(n_1 n_2)$.

SAMPLING

Theorem Let S and T be two trees of size n_1 and n_2.
Sampling alignments between S and T under the Gibbs-Boltzmann distribution can be done with worst-case time and space complexities $O(n_1 n_2 (n_1 + n_2)^2)$ and with average-case time and space complexities $O(n_1 n_2)$.

Proof inspired by

[Herrbach, Denise, Dulucq]

SAMPLING

Theorem Let S and T be two trees of size n_1 and n_2. Sampling alignments between S and T under the Gibbs-Boltzmann distribution can be done with worst-case time and space complexities $O(n_1 n_2 (n_1 + n_2)^2)$ and with average-case time and space complexities $O(n_1 n_2)$.

Upsides:

- No additional complexity cost (except constants, moderate)
- Flexibility of the sampling algorithm.
- Already implemented.

Downside

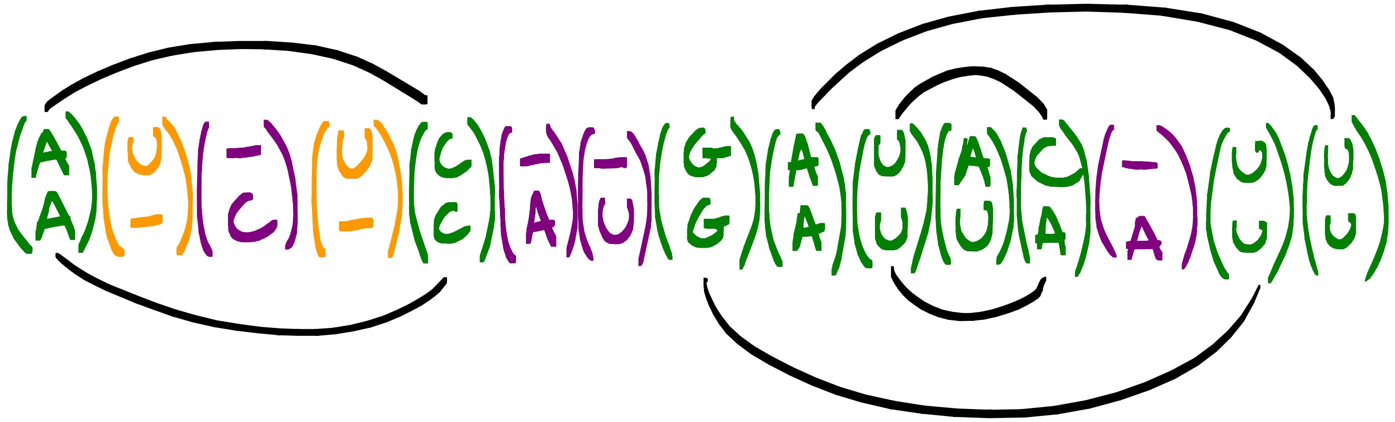
- Complicated DP scheme.

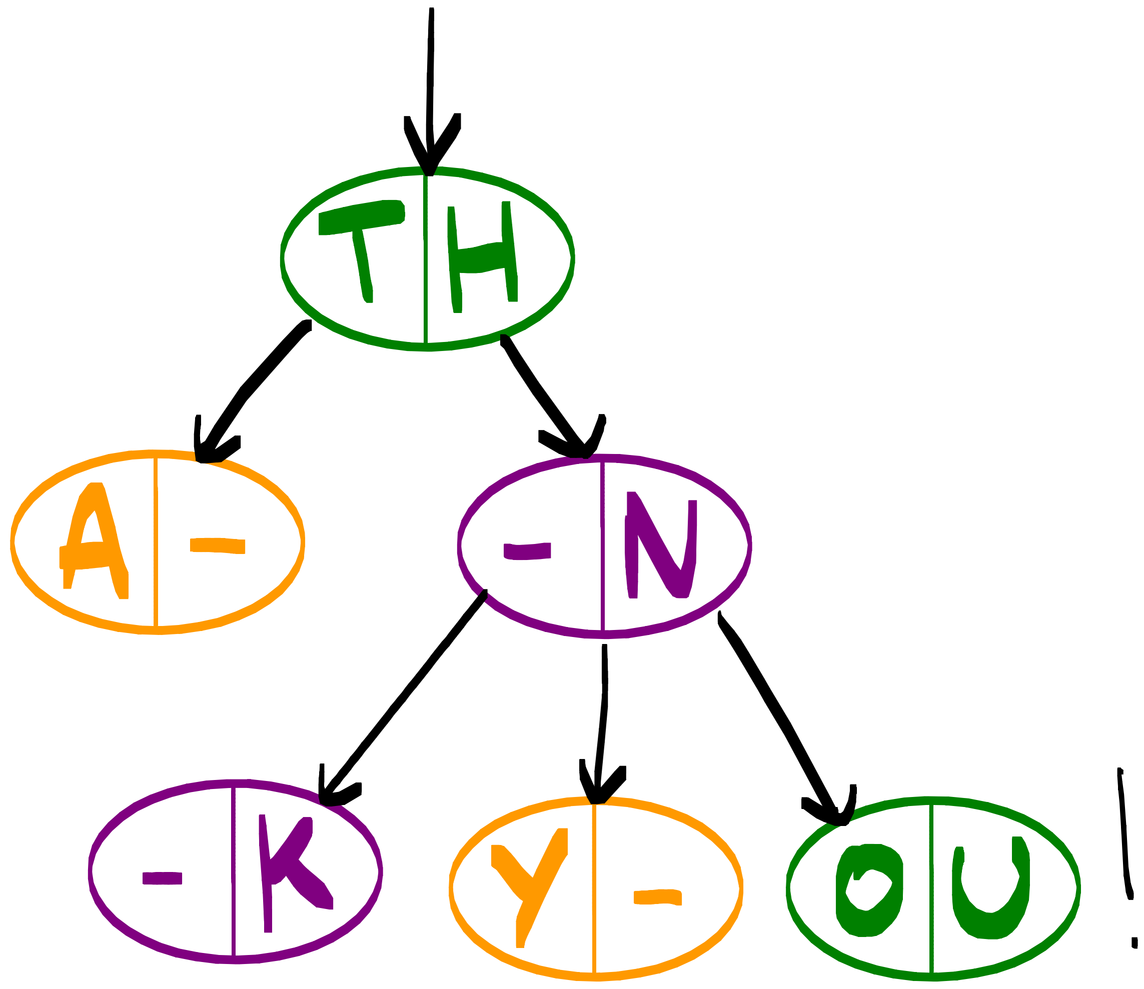
CONCLUSION

→ Algorithm that generates the neighborhood of a tree?

→ Existence of easier decompositions?

→ Alignment problem for arc-annoted sequences?




T | H
A | -
- | N
- | K
Y | -
O | U
!

APPLICATION 1: COUNTING.

Theorem: The generating function $A(z,u)$ of tree alignments satisfies

$$A(z,u) = \left(z^2 + z - uz^2 + \frac{z}{\sqrt{1-4z}}\right) \times B(z,u)$$

where

$$\left(uzC(z)^2 - z^2C(z)^2 + 2z\right)B(z,u)^2 + \left(z^2C^4(z) - 2zC(z)^2 - 1\right)B(z,u) + C^2(z) = 0$$

and

$$C(z) = \frac{1-\sqrt{1-4z}}{2z}$$ Catalan generating function