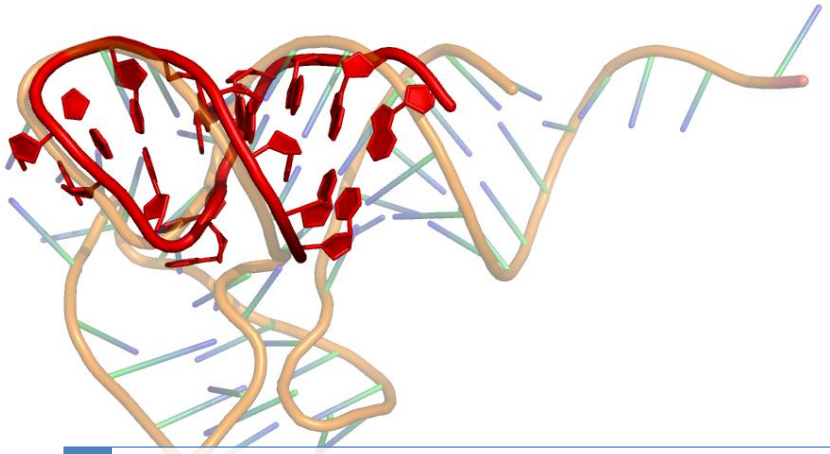
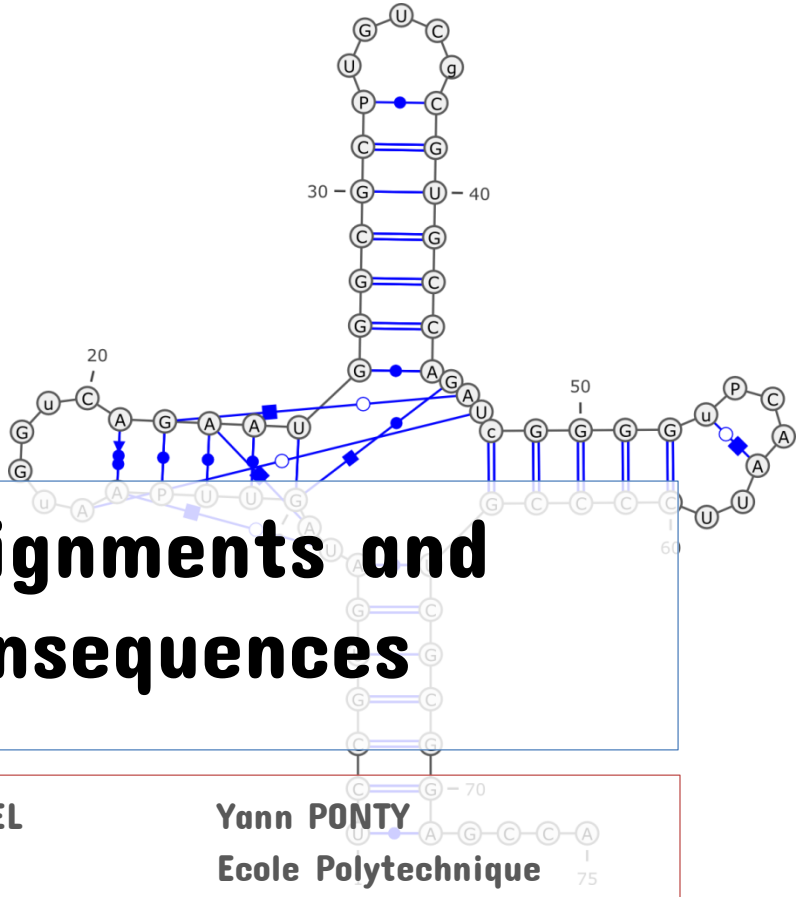




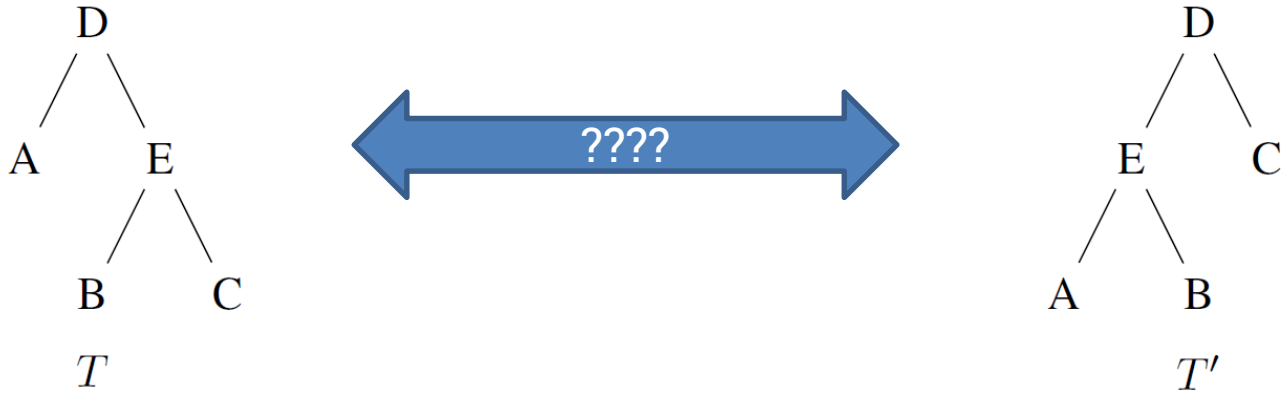
Counting tree alignments and algorithmic consequences

Cédric CHAUVE
Simon Fraser Univ.

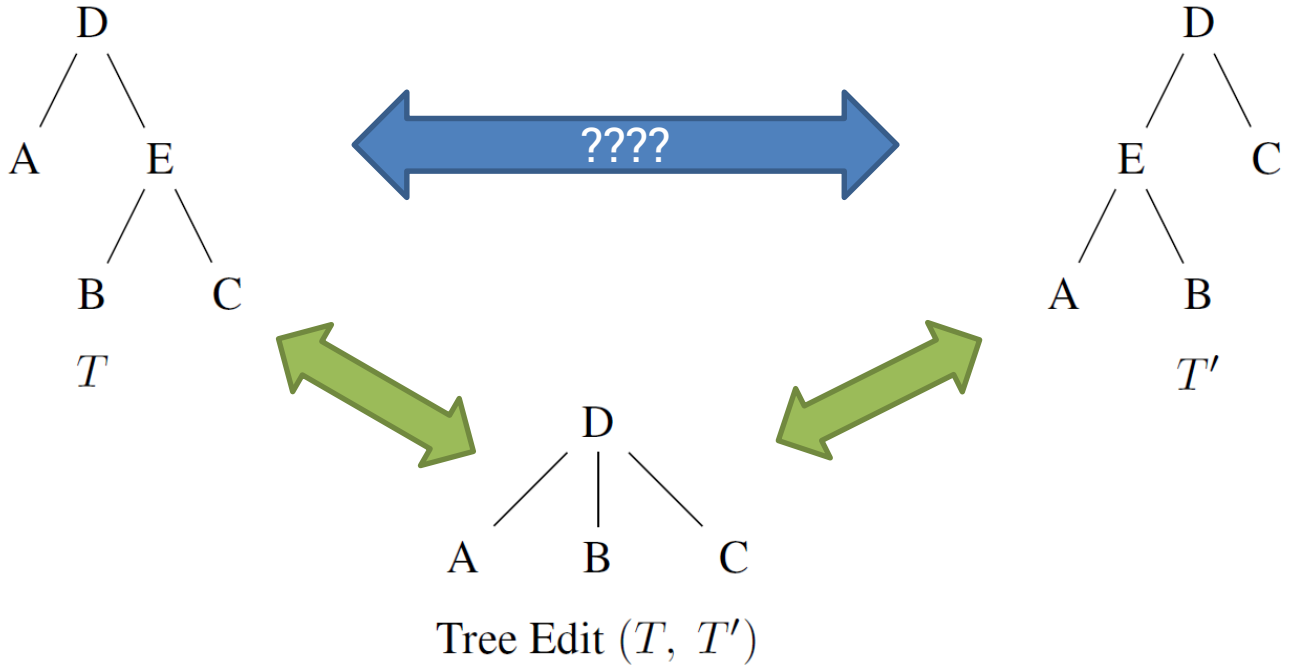
Julien COURTIEL
Univ. de Caen

Yann PONTY
Ecole Polytechnique

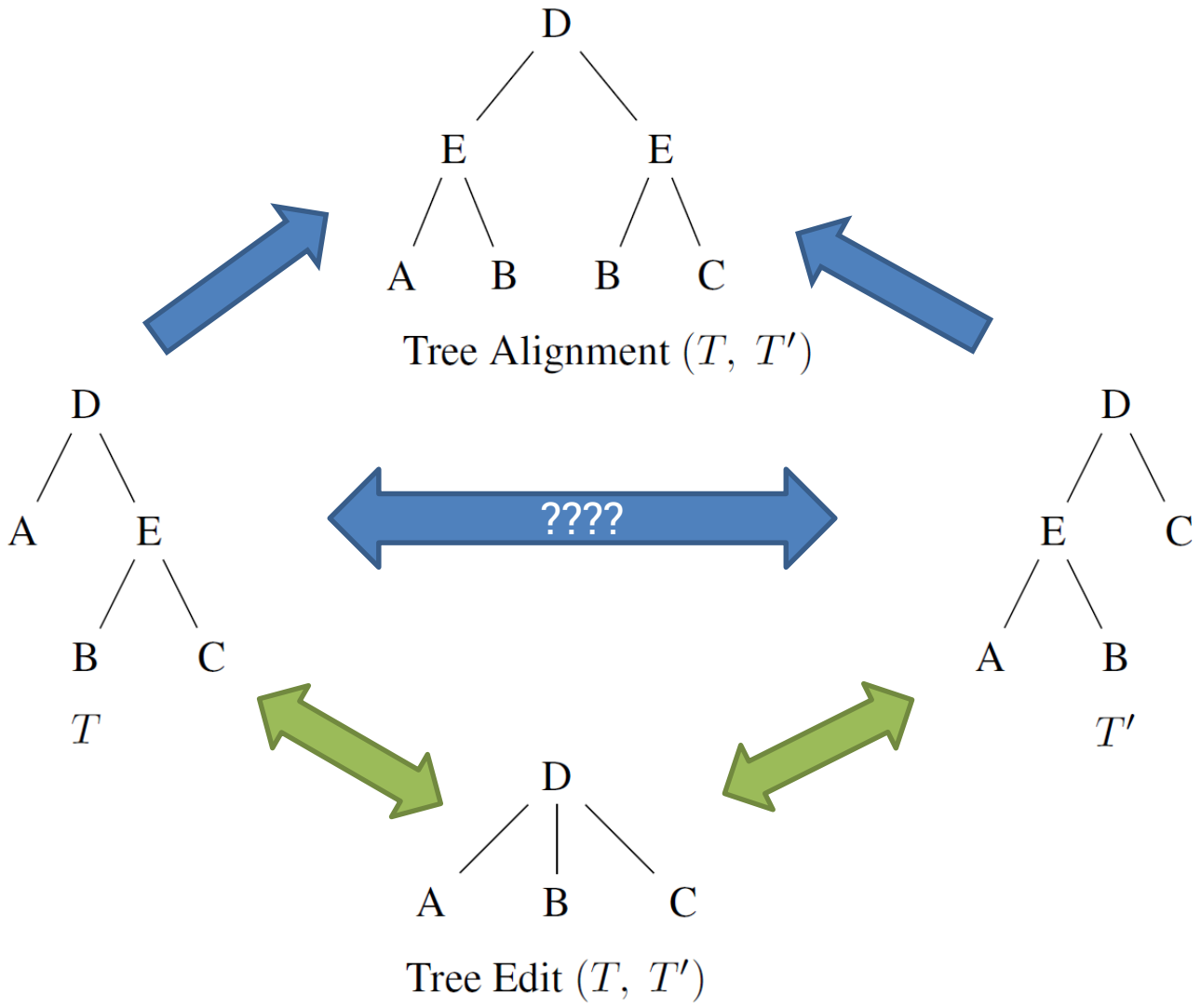
Alignment distance vs edit distance



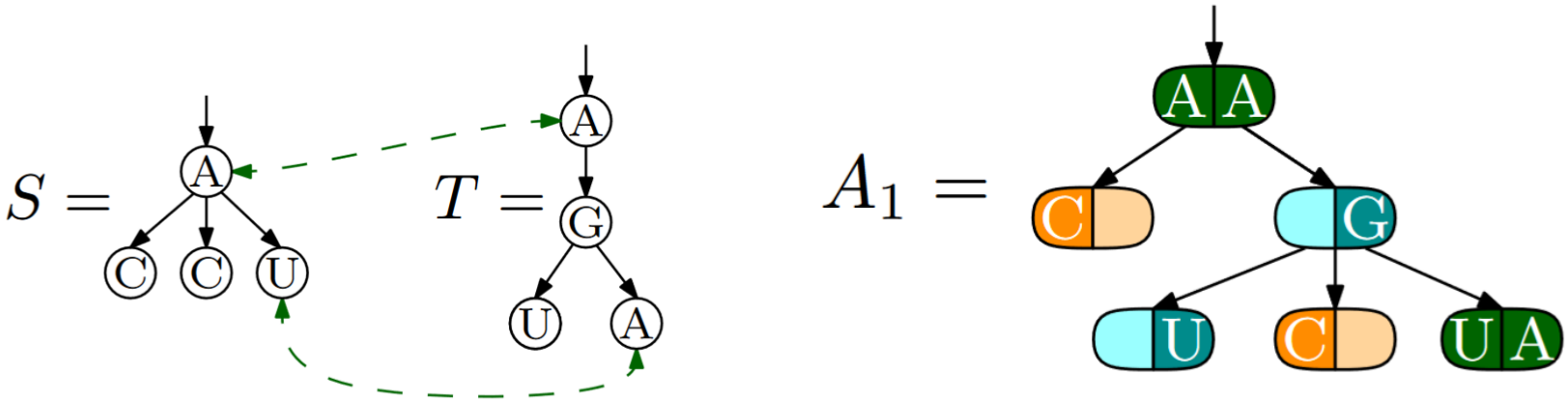
Alignment distance vs edit distance



Alignment distance vs edit distance



Tree alignments and supertrees



- Alignment = Set of correspondences, *aka* (mis)matches (→Set of indels)

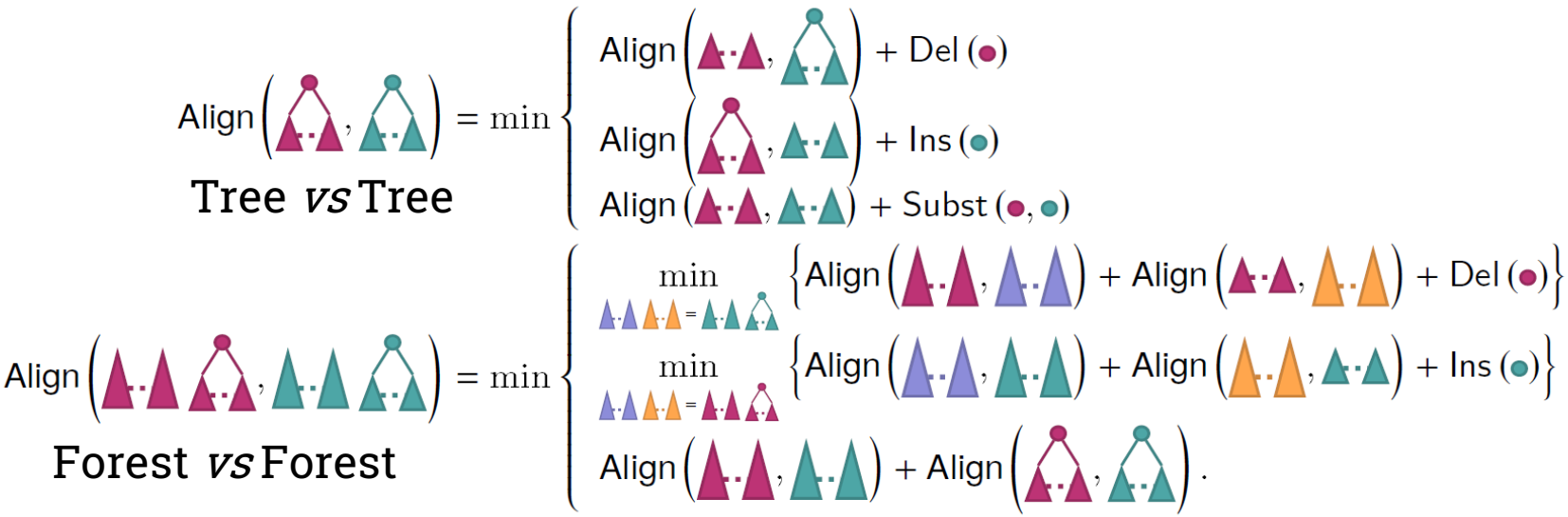
  + Validity: Consistency with ancestries in S and T

- Alt. Alignment = Supertree $\boldsymbol{S}$ such that S and T can be recoved from $\boldsymbol{S}$

Tree alignment problem (Jiang-Wang-Zhang 1995)

Input	Two trees S ($\|S\| = n_1$) and T ($\|T\| = n_2$)
Output	Supertree maximizing weighted sum over (mis)matches

Jiang, Wang and Zhang (JWZ) DP algorithm



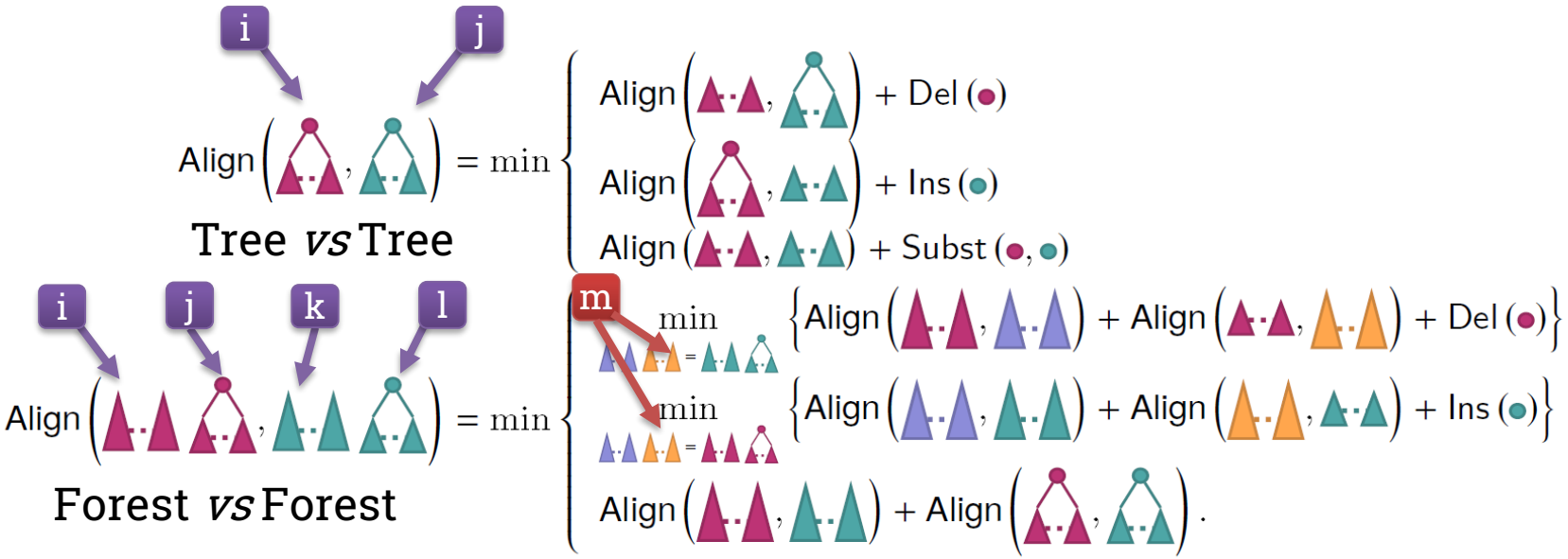
Complexities

- Worst-case → $O(n_1^2.n_2^2)$ space, $O(n_1^2.n_2^2.\max(n_1,n_2))$ time

  But at least one of (i, j, k, l) is first/last of its siblings

  → $\theta(n_1.n_2.\max(n_1,n_2))$ space, $\theta(n_1.n_2.\max(n_1,n_2)^2)$ time

Jiang, Wang and Zhang (JWZ) DP algorithm



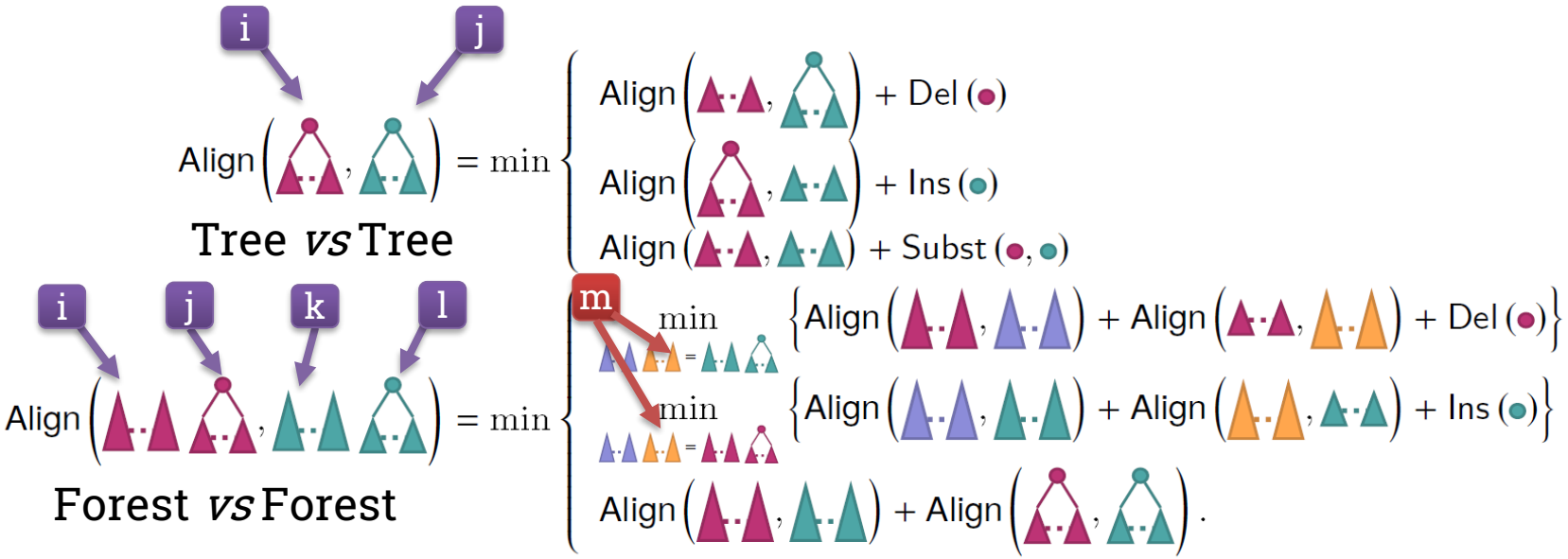
Complexities

- Worst-case → $O(n_1^2.n_2^2)$ space, $O(n_1^2.n_2^2.\max(n_1,n_2))$ time

  But at least one of (i, j, k, l) is first/last of its siblings

  → $\theta(n_1.n_2.\max(n_1,n_2))$ space, $\theta(n_1.n_2.\max(n_1,n_2)^2)$ time

Jiang, Wang and Zhang (JWZ) DP algorithm



Complexities

- Worst-case → $O(n_1^2.n_2^2)$ space, $O(n_1^2.n_2^2.\max(n_1,n_2))$ time

  But at least one of (i, j, k, l) is first/last of its siblings

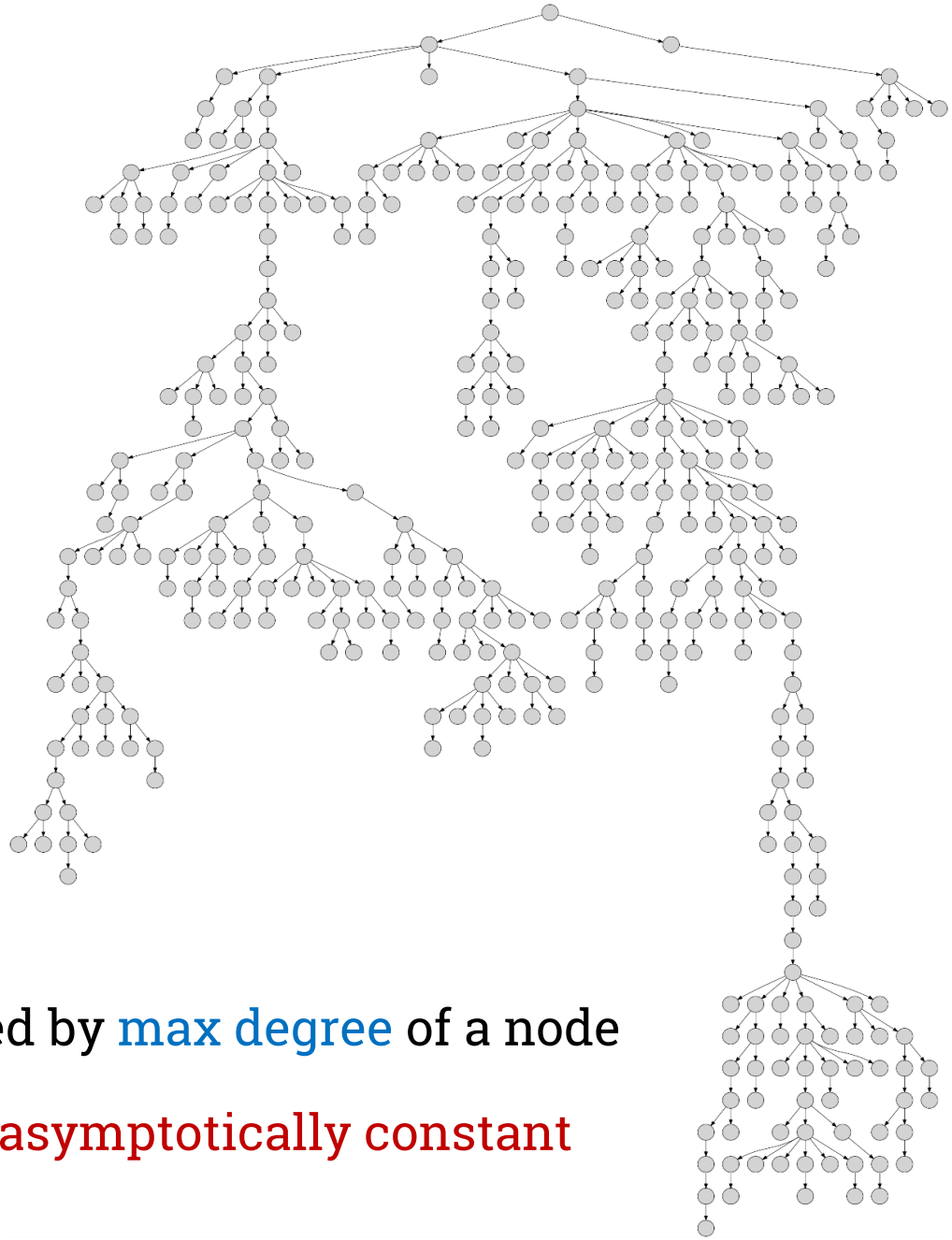
  → $\theta(n_1.n_2.\max(n_1,n_2))$ space, $\theta(n_1.n_2.\max(n_1,n_2)^2)$ time

- Average-case: Random, uniformly distributed, trees of length n_1 and n_2

  → $\theta(n_1.n_2)$ space, $\theta(n_1.n_2)$ time [Herrbach, Dulucq, Denise, TCS 2010]

Remark: Holds for Boltzmann-distributed RNA 2D struct. (homopolymer model)

Intuition



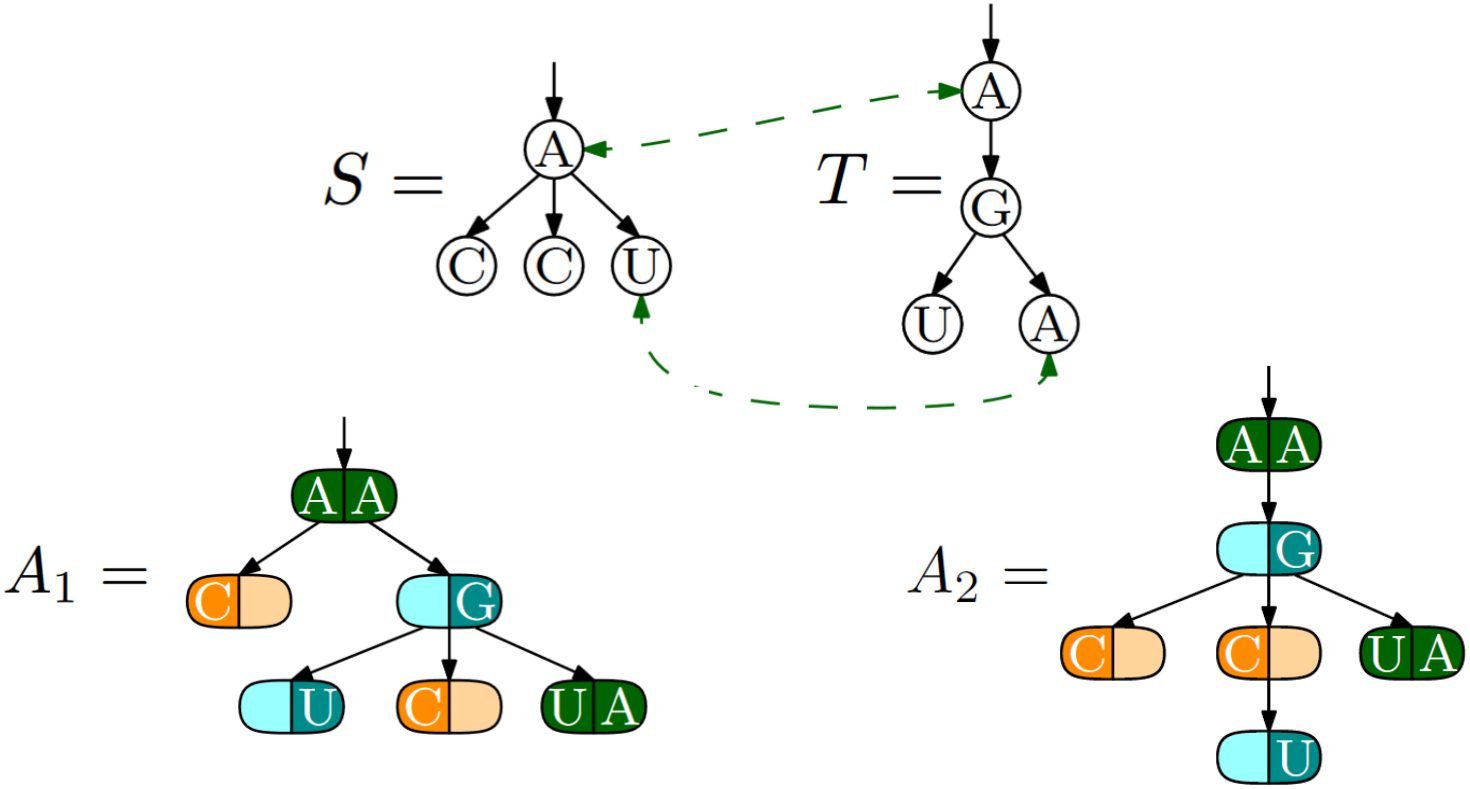
Time complexity dictated by max degree of a node

Max. degree on average asymptotically constant

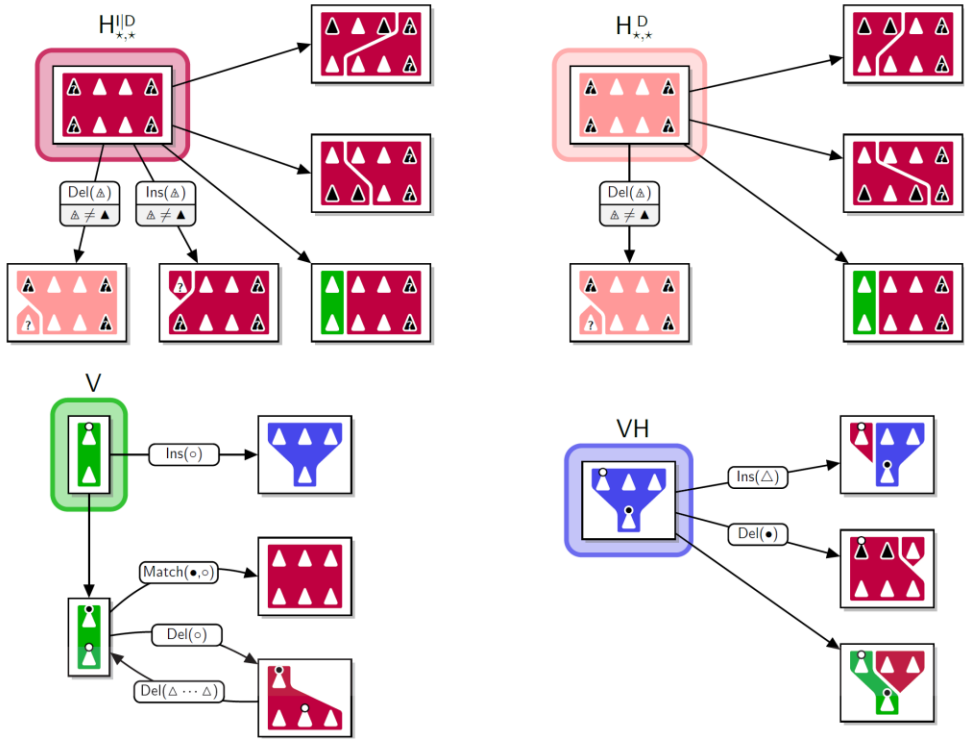
Counting alignments

Motivation: Ensemble analyses (*e.g.* MEA alignment, Bolz. Prob…)

Problem: DP scheme of Jiang-Wang-Zhang is ambiguous



Unambiguous decomposition/DP scheme



Theorem: DP scheme complete, unambiguous + complexities of JWZ

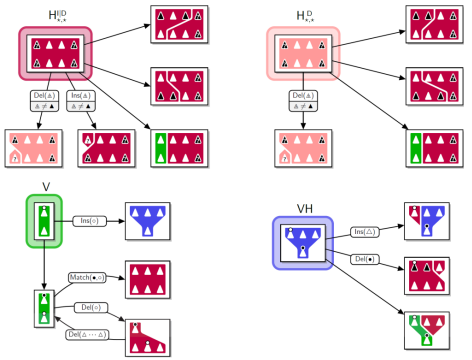
Proof: by intimidation, mainly! (induction, seriously…)

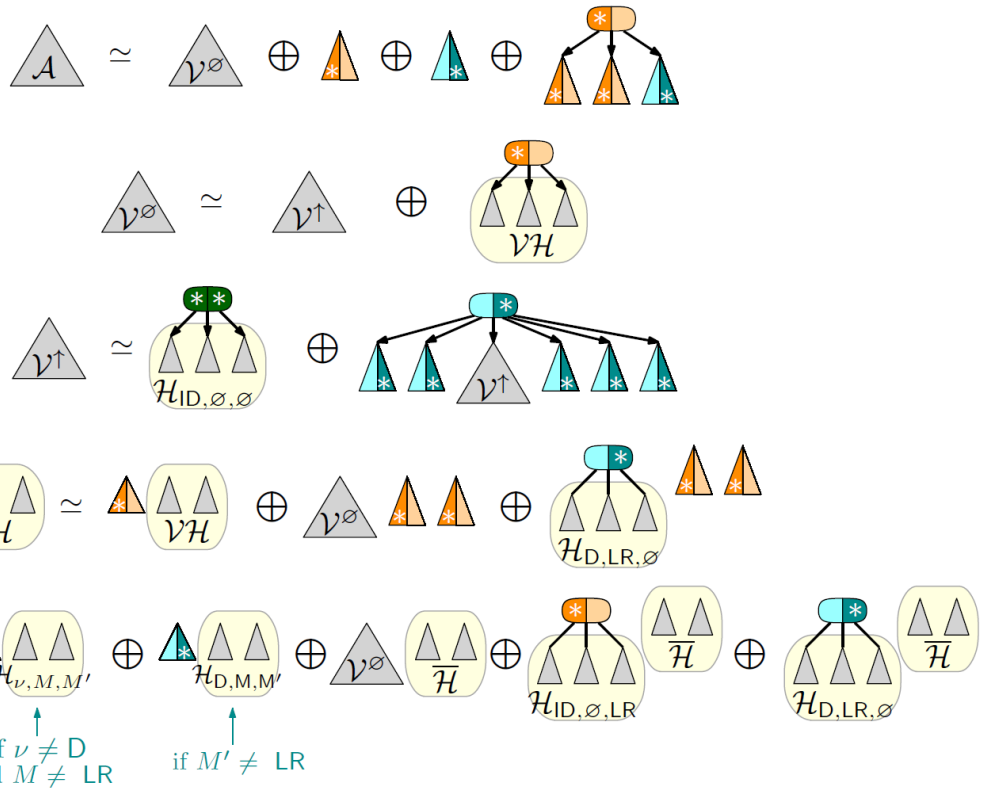
Alternatively: Über-simple unambiguous decomposition (yet mildly incomplete)

[Berkemer, Höhner zu Siederdissen, Stadler, Algorithms 2017]

Counting tree alignments [Chauve Courtiel P, IJFCS 2018]

Scary DP scheme





Scarier combinatorial specification/grammar

System of (algebraic) functional equations → Generating functions

→ Singularity analysis → Asymptotic properties of tree alignments

Asymptotic properties of tree alignments

- #Tree alignments, over a total of n nodes, equivalent to

$$\frac{\sqrt{2}(3-\sqrt{3})}{24\sqrt{\pi}}\frac{6^n}{n\sqrt{n}}$$

  → ≈ 1.5^n tree alignments per pair of tree

- #(Mis)matches in random tree alignment:
  - Expectation ~ n/6
  - Variance ~ n/6

- Avg #supertrees per tree alignment exponential on length
  yet, for all n, unique supertree for certain alignments
  → Exponential bias induced by JWZ decomposition

Trees easier to align than sequences?!

Input: Pair of random uniform sequences of length n_1 and n_2

- Theorem: Needlemann-Wunsch runs in $\theta(n_1.n_2)$ expected time

Input: Pair of random uniform trees of length n_1 and n_2

- Reminder: JWZ algorithm (+ ours) run in $\theta(n_1.n_2)$ expected time

Trees easier to align than sequences?!

Input: Pair of random uniform sequences of length n_1 and n_2

- Theorem: Needlemann-Wunsch runs in $\theta(n_1.n_2)$ expected time

Input: Pair of random uniform trees of length n_1 and n_2

- Reminder: JWZ algorithm (+ ours) run in $\theta(n_1.n_2)$ expected time

---

Input: Random uniform pair of sequences of cumulated length n

- Theorem: Needlemann-Wunsch runs in $\theta(n^2)$ expected time

Input: Random uniform pair of trees of cumulated length n

Trees easier to align than sequences?!

Input: Pair of random uniform sequences of length n_1 and n_2

- Theorem: Needlemann-Wunsch runs in $\theta(n_1.n_2)$ expected time

Input: Pair of random uniform trees of length n_1 and n_2

- Reminder: JWZ algorithm (+ ours) run in $\theta(n_1.n_2)$ expected time

---

Input: Random uniform pair of sequences of cumulated length n

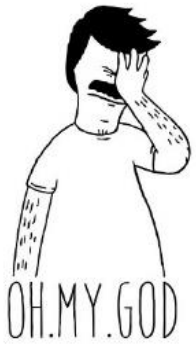
- Theorem: Needlemann-Wunsch runs in $\theta(n^2)$ expected time

Input: Random uniform pair of trees of cumulated length n

- Theorem: JWZ algorithm (+ ours) run in $\theta(n\sqrt{n})$ expected time

Aligning trees easier than aligning sequences?!

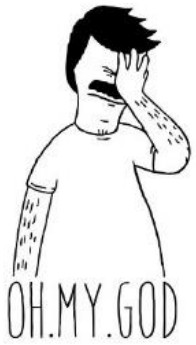
Of course, this is cheating...



n = cumulated length of a pair

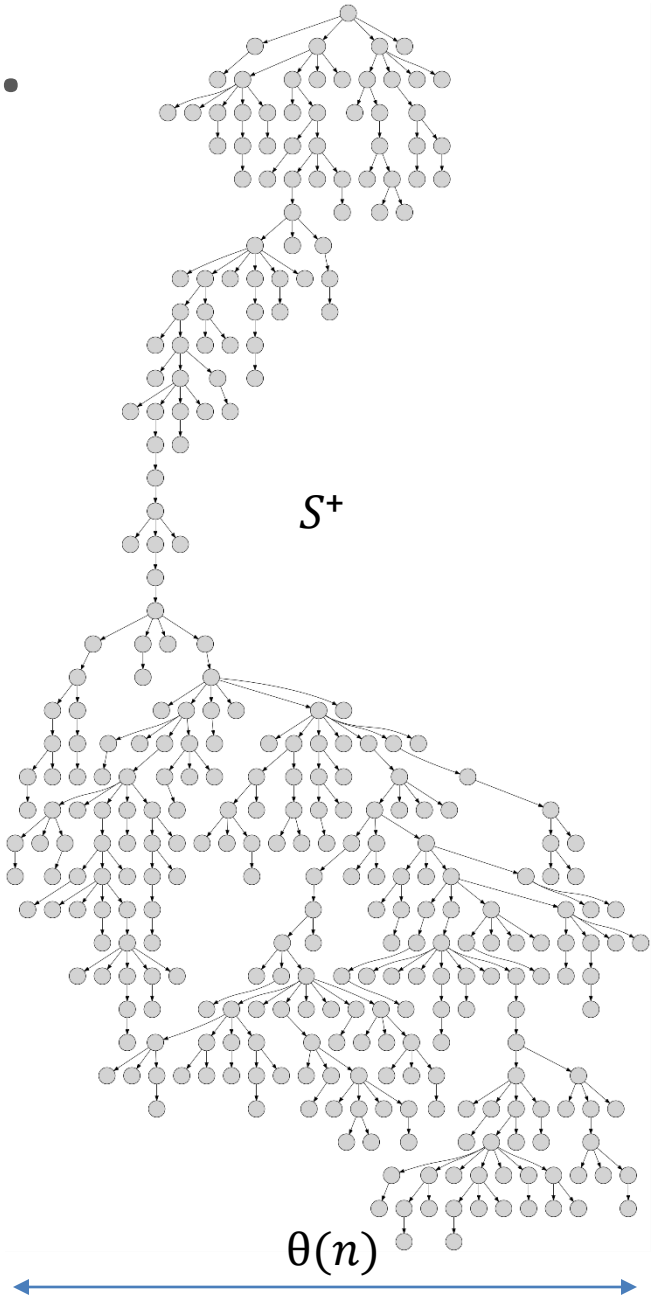
- For sequences:
  - Both have expected length $\frac{n}{2}$
- For trees:

Of course, this is cheating…

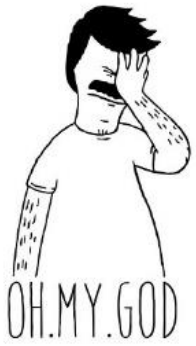


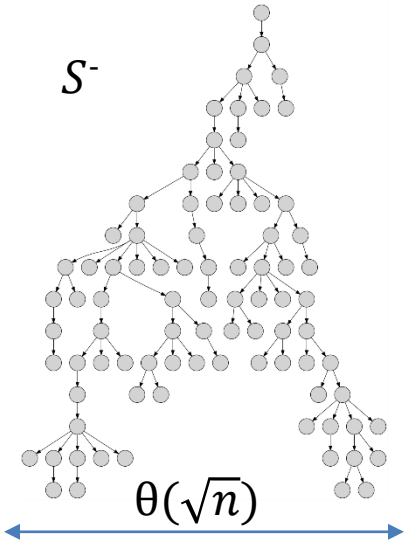
n = cumulated length of a pair

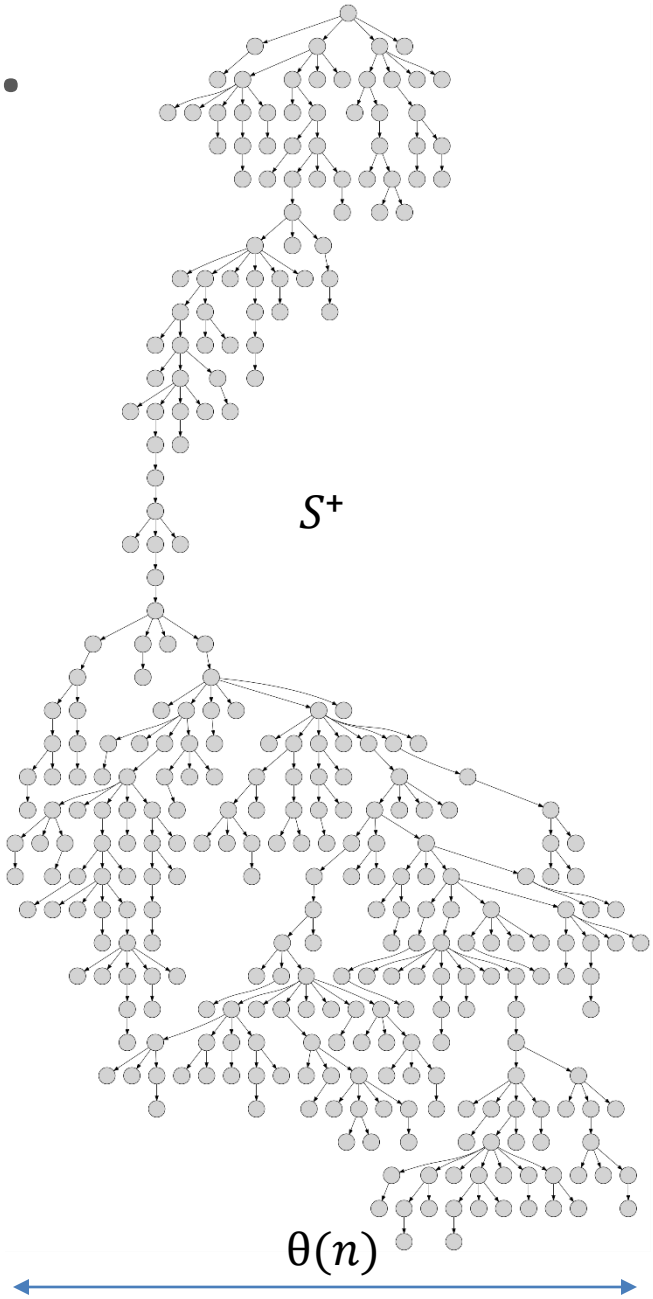
- For sequences:
  - Both have expected length $\frac{n}{2}$
- For trees:
  - Largest tree S^+ has expected length in $\theta(n)$



Of course, this is cheating...







n = cumulated length of a pair

- For sequences:
  - Both have expected length $\frac{n}{2}$
- For trees:
  - Largest tree S^+ has expected length in $\theta(n)$
  - Smallest tree S^- has expected length in $\theta(\sqrt{n})$

The subquadratic complexity of JWZ is only an artifact of the length distribution!

Conclusions/thanks

- Don't be fooled by combinatorics people! (including this one…)
- Tree alignment amenable to ensemble analyses
- Who would like to co-implement our ~~monster~~ DP beauty?

Thanks to Bled Organizers + YOU

Julien Courtiel

Cédric Chauve