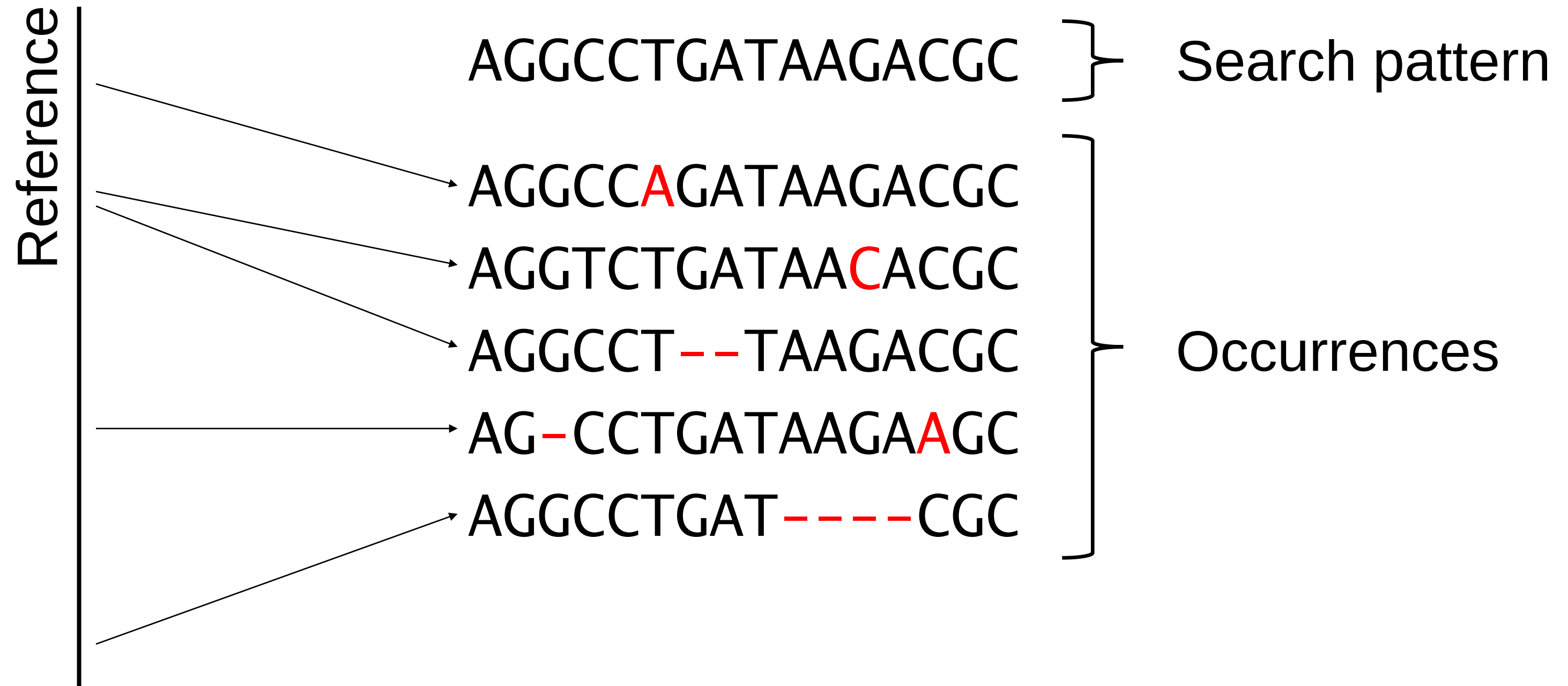


Towards Lossless Approximate Pattern Matching on Pan- genome de Bruijn Graphs

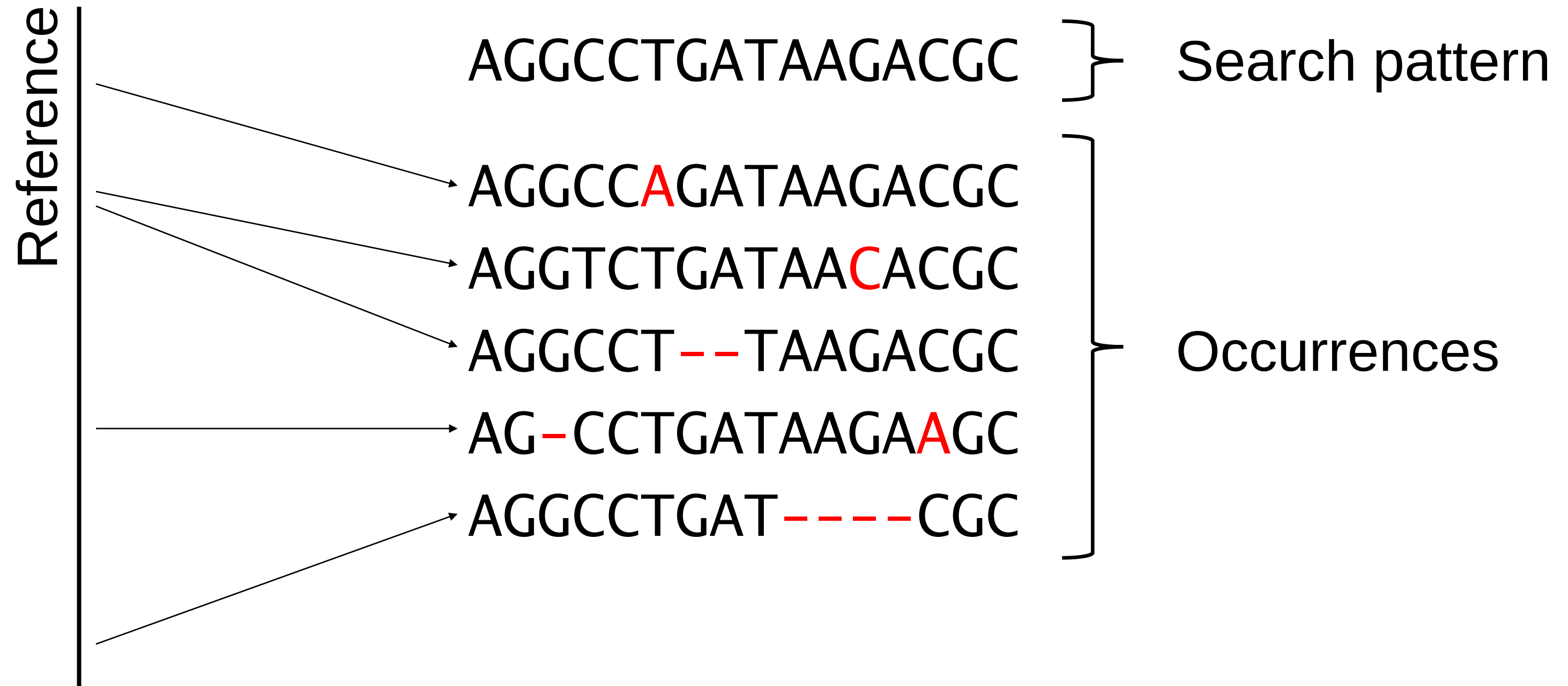
Lore Depuydt, Luca Renders, Thomas Abeel and Jan Fostier – IWBBIO 2022

Lossless Approximate Pattern Matching on Pan-genome de Bruijn Graphs

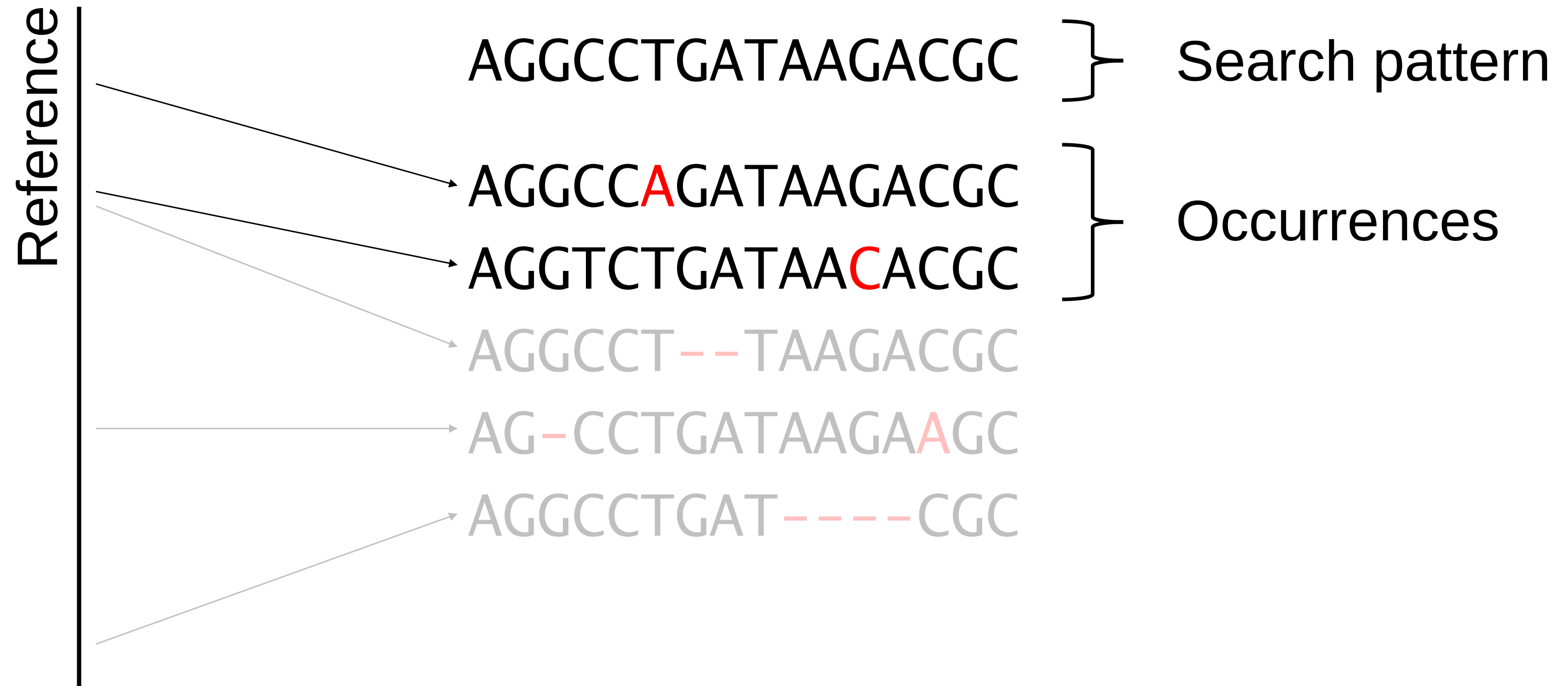
Lossless Approximate Pattern Matching on Pan-genome de Bruijn Graphs



Lossless Approximate Pattern Matching on Pan-genome de Bruijn Graphs



Lossy Approximate Pattern Matching on Pan-genome de Bruijn Graphs

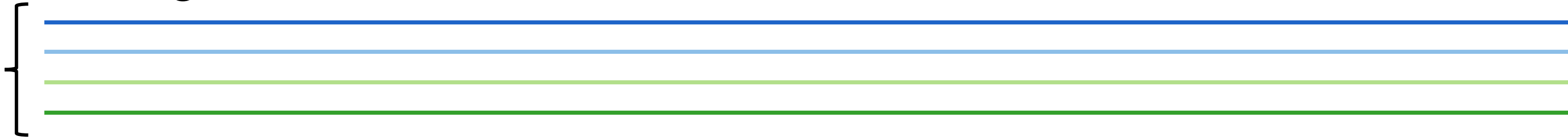


Lossless Approximate Pattern Matching on Pan-genome de Bruijn Graphs

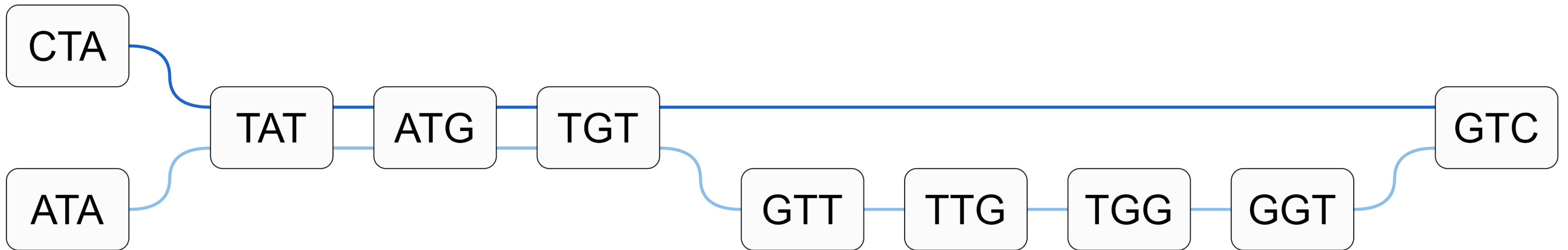
Consensus genome build:



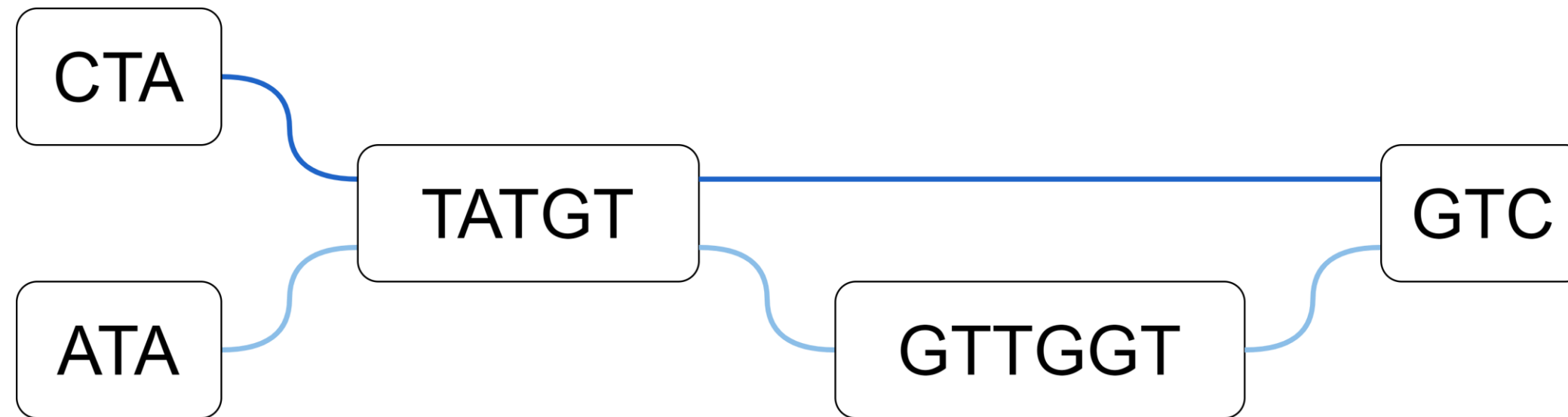
Pan-genome:



Lossless Approximate Pattern Matching on Pan-genome de Bruijn Graphs



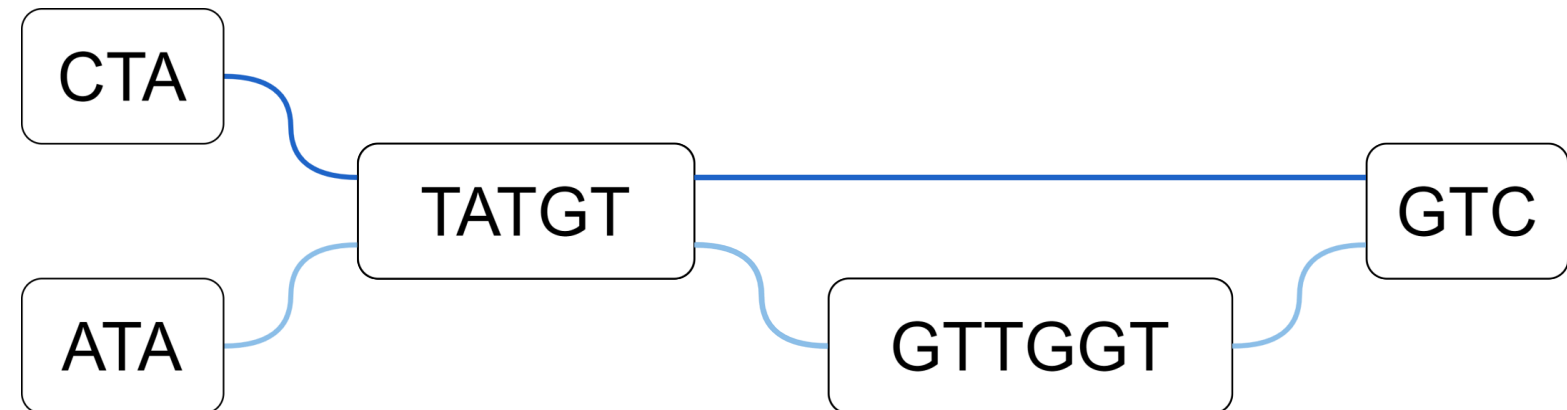
Lossless Approximate Pattern Matching on Pan-genome de Bruijn Graphs



Lossless Approximate Pattern Matching on Pan-genome de Bruijn Graphs

Search pattern: ATGTA

Pan-genome graph:

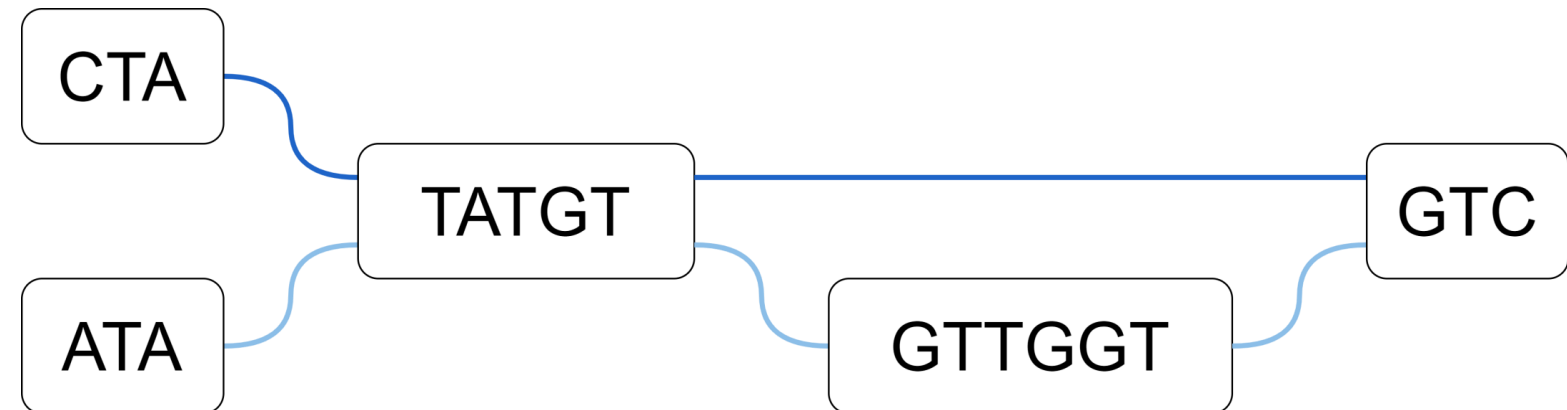


Occurrences:

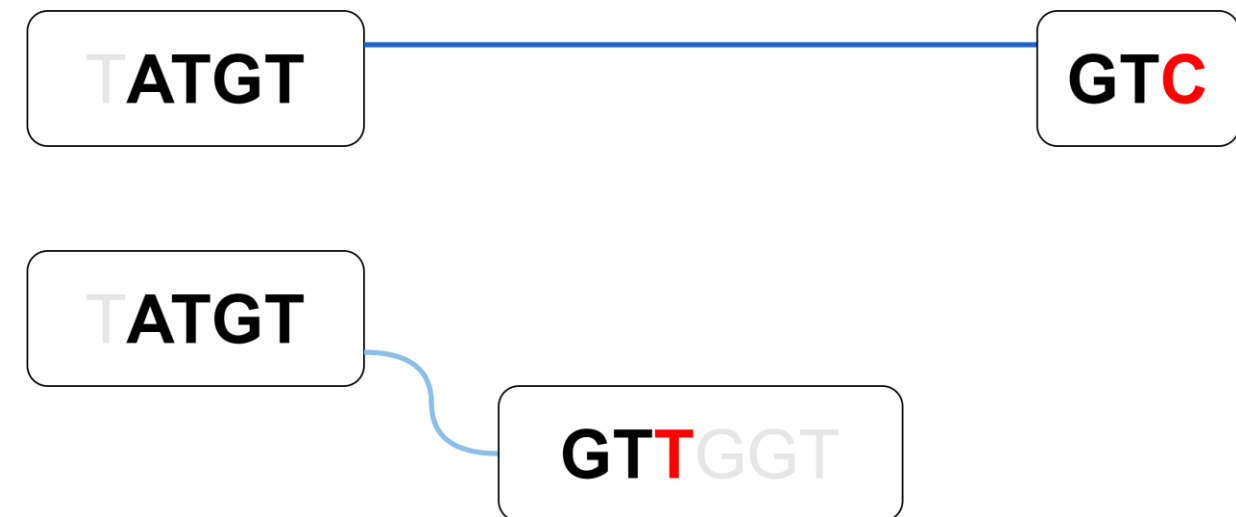
Lossless Approximate Pattern Matching on Pan-genome de Bruijn Graphs

Search pattern: ATGTA

Pan-genome graph:



Occurrences:



Outline

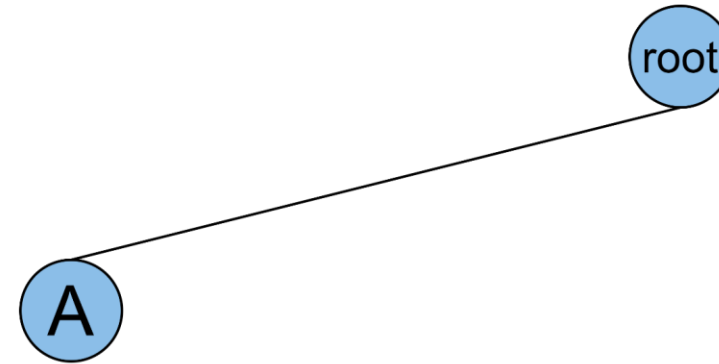
- Search schemes
- Pan-genome graph data structure
- Pattern matching against the pan-genome:
 - Strain-fixed
 - Strain-free
- Results and implementation

Search Schemes

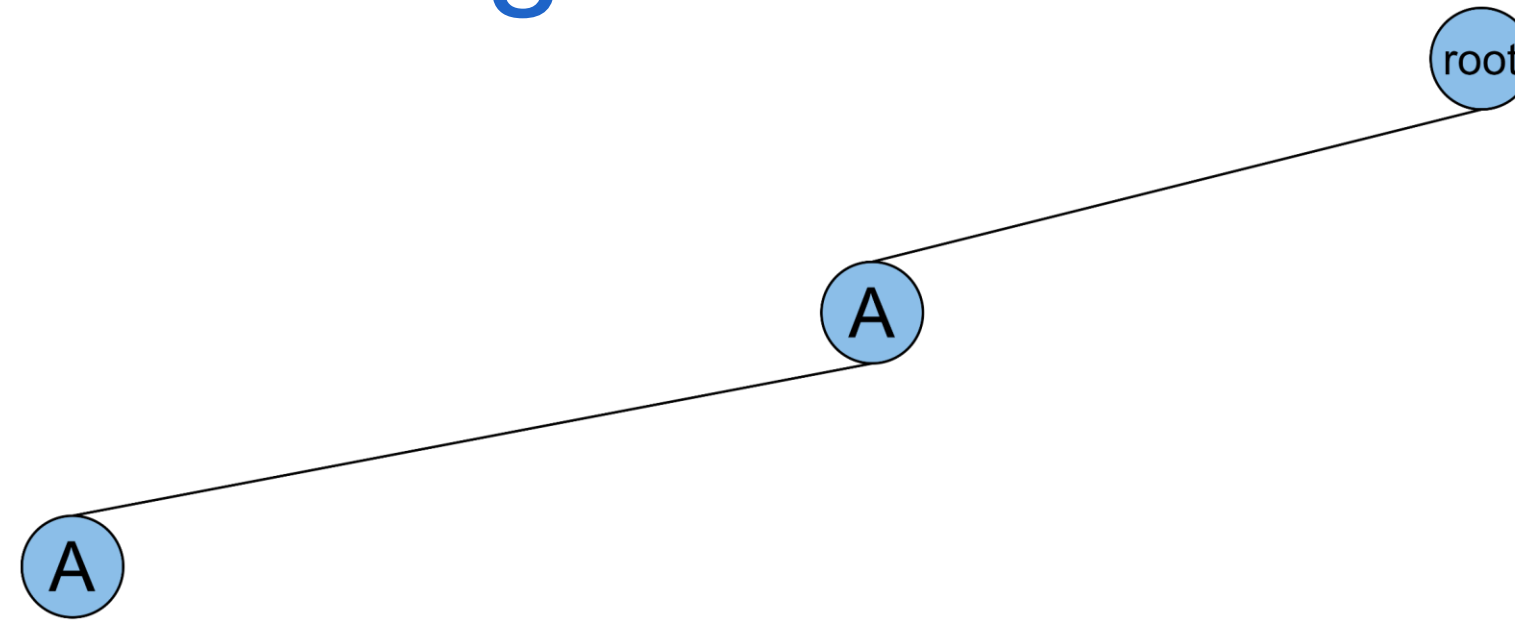
Naive Backtracking



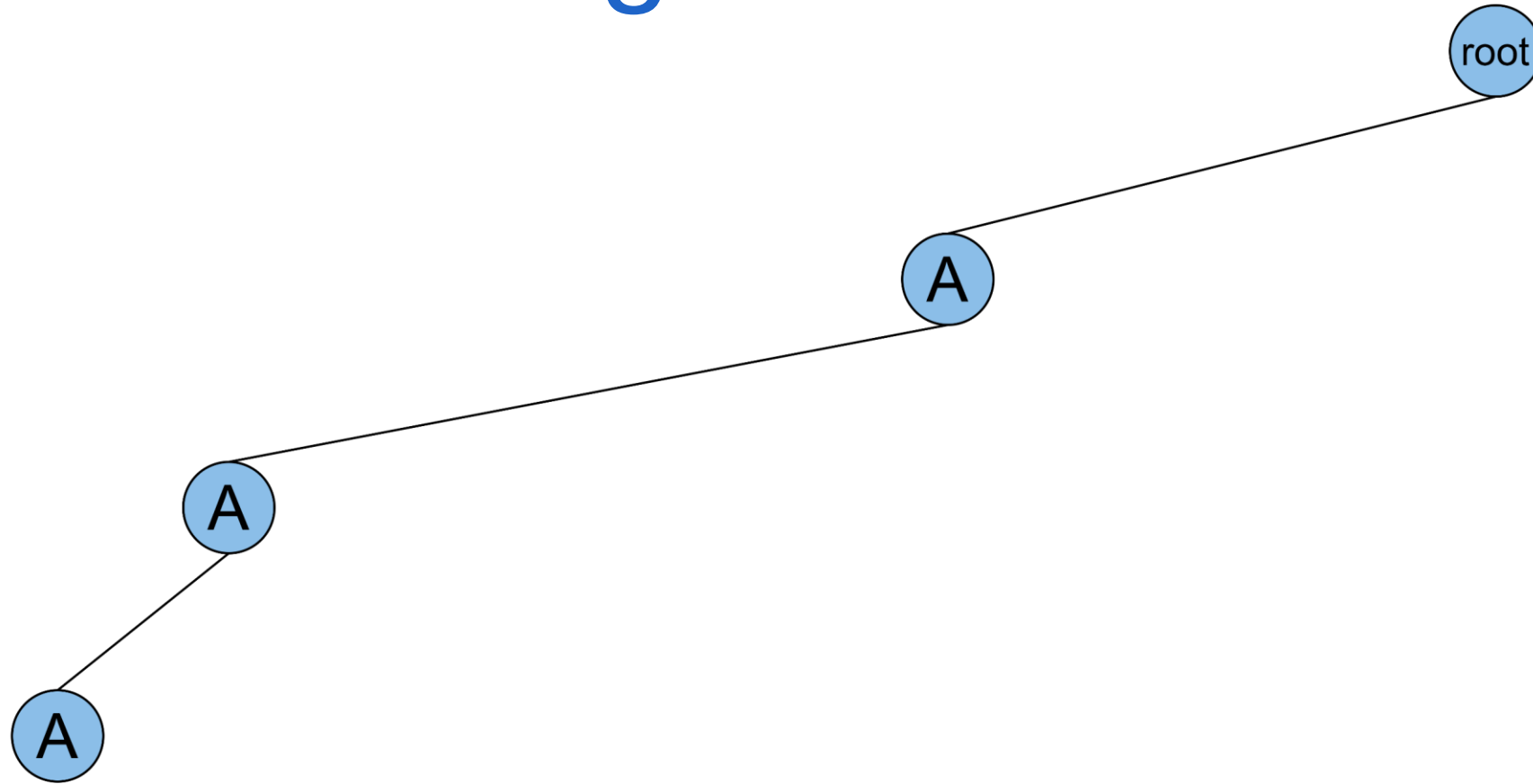
Naive Backtracking



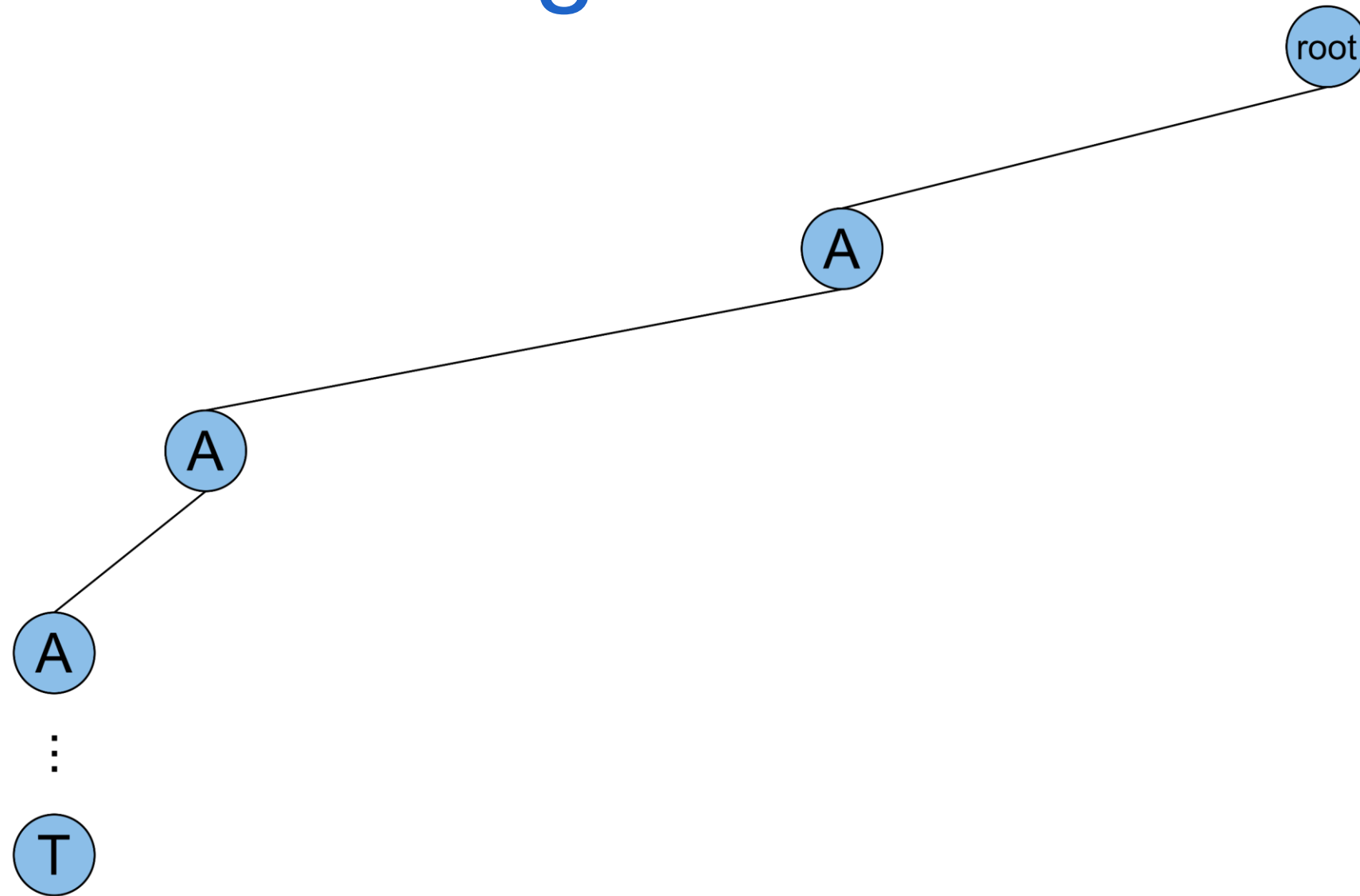
Naive Backtracking



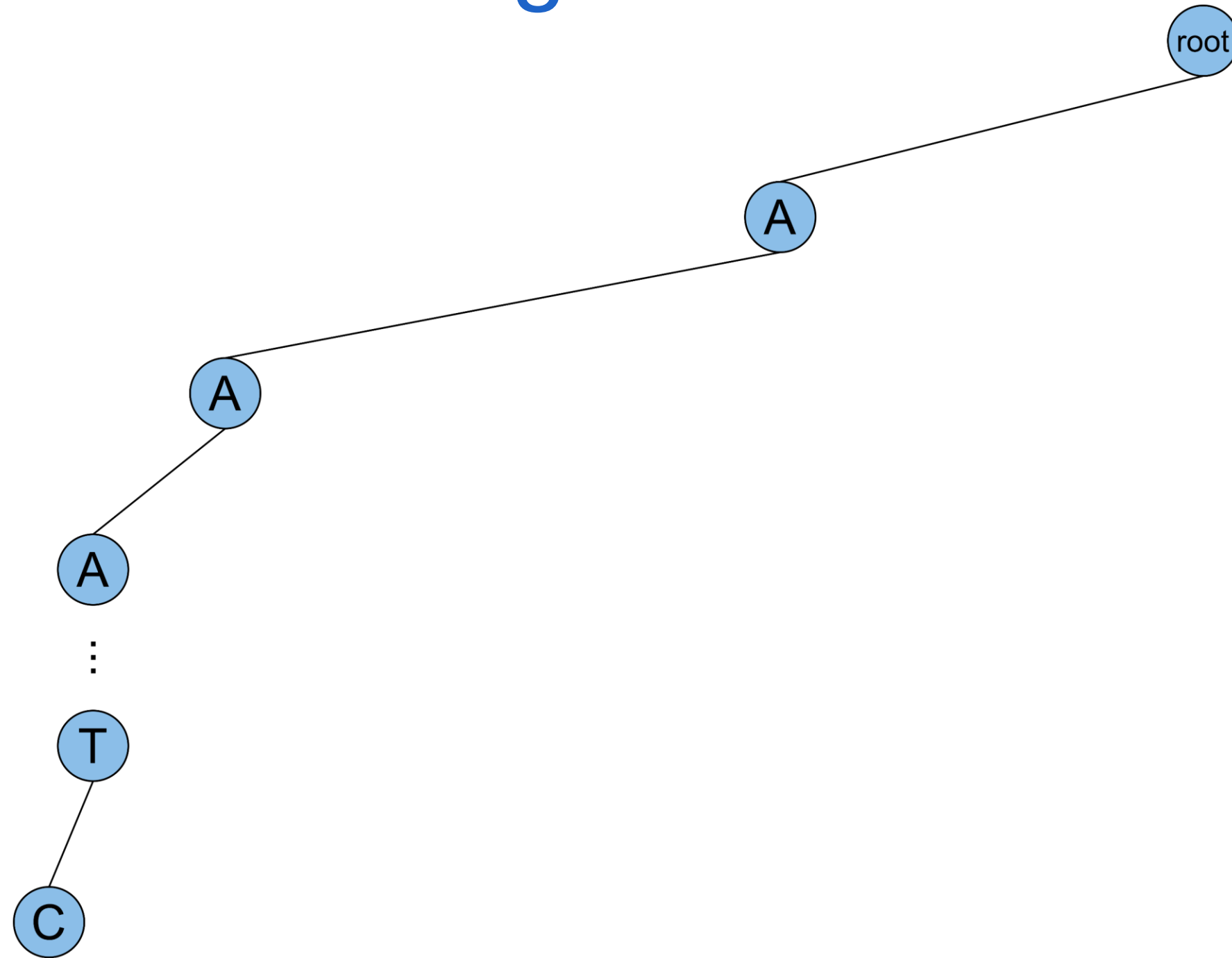
Naive Backtracking



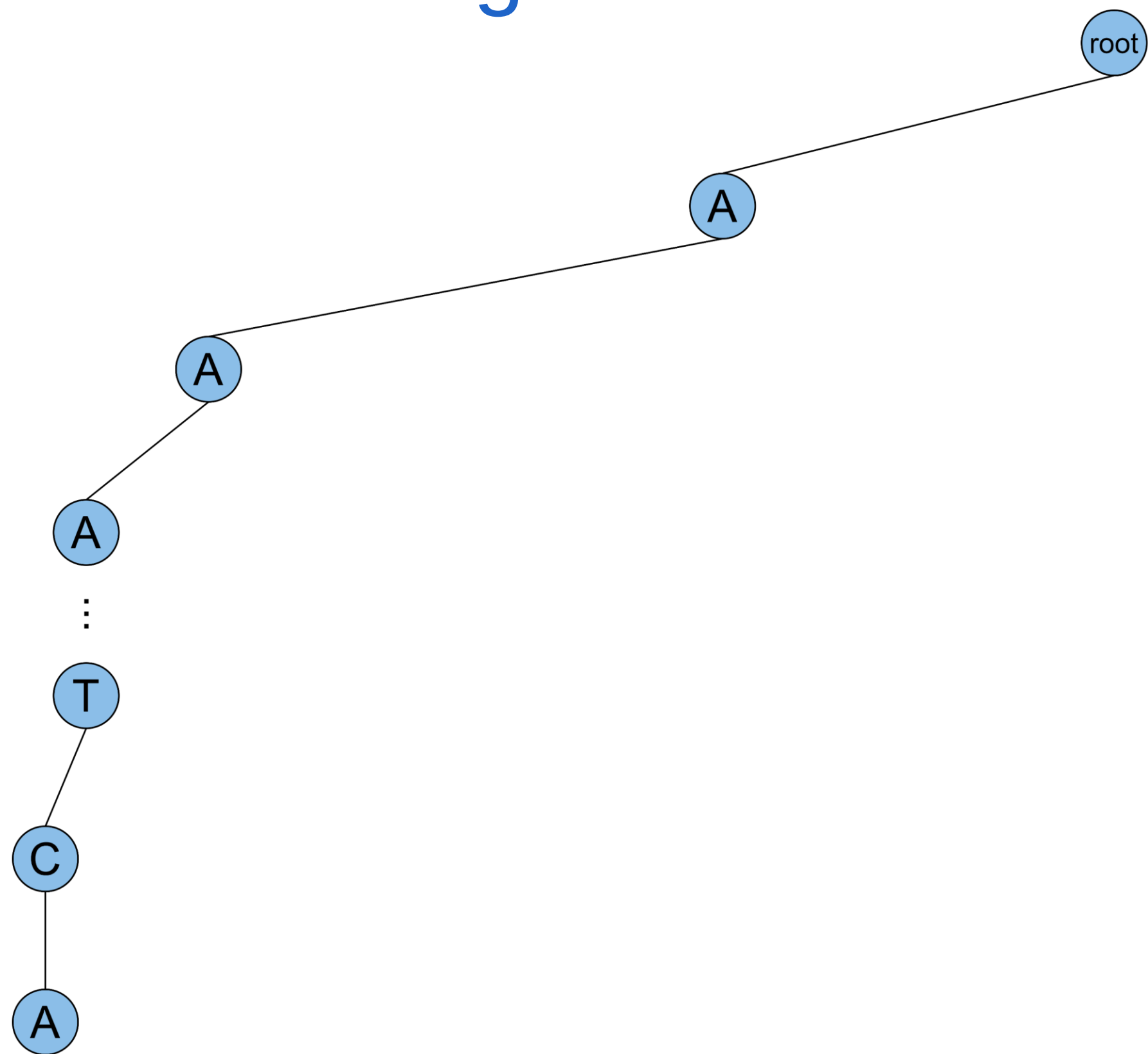
Naive Backtracking



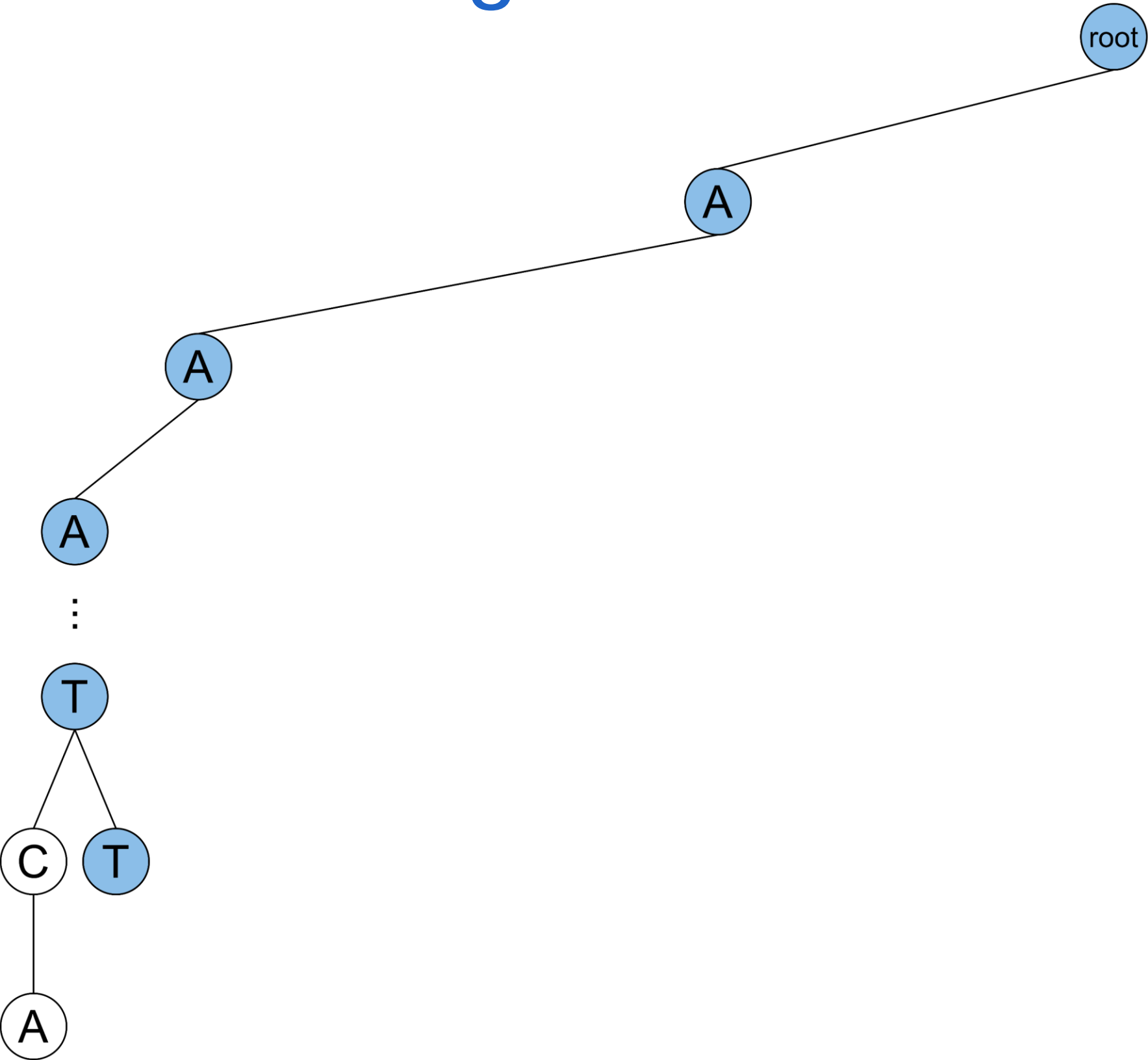
Naive Backtracking



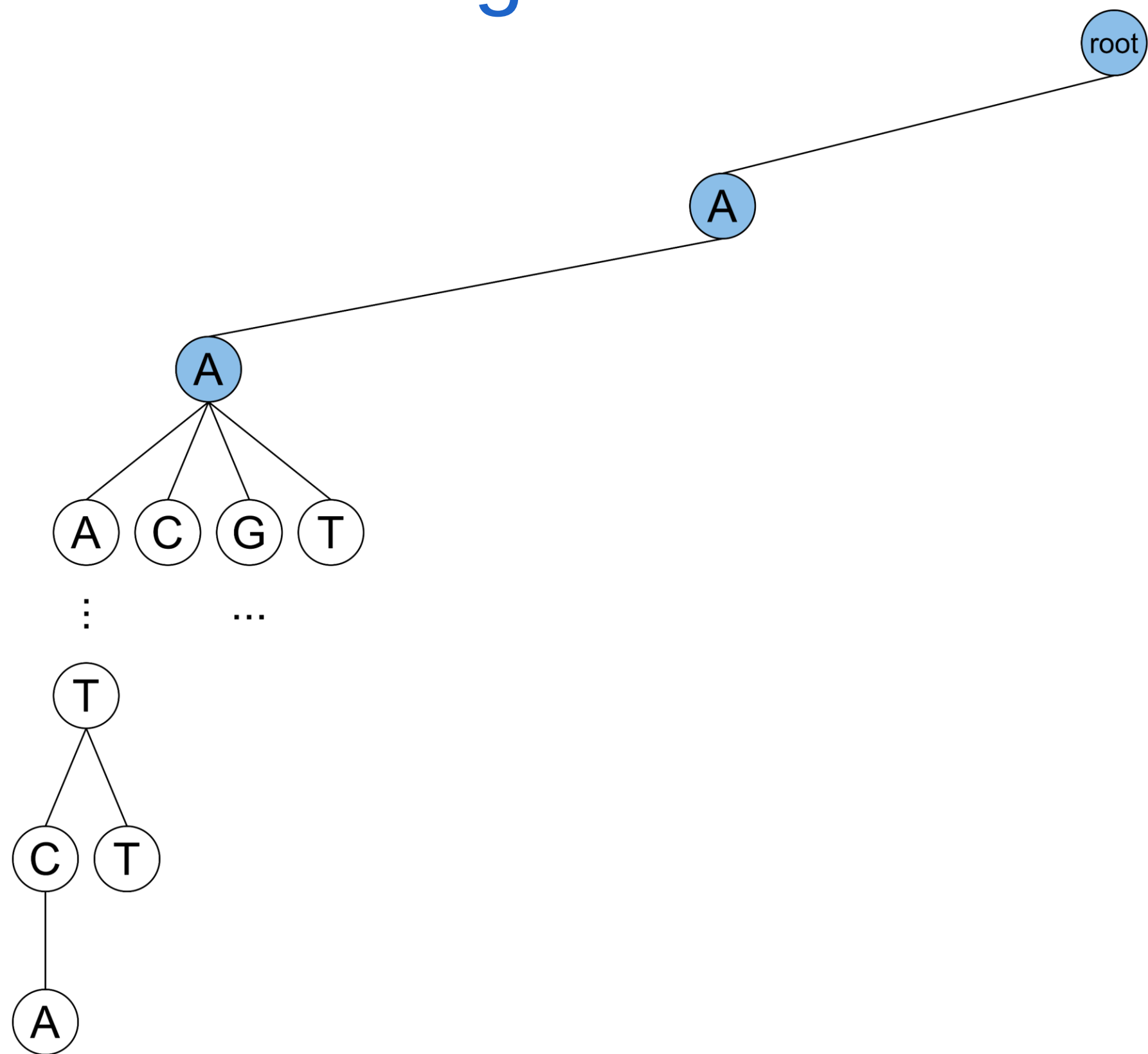
Naive Backtracking



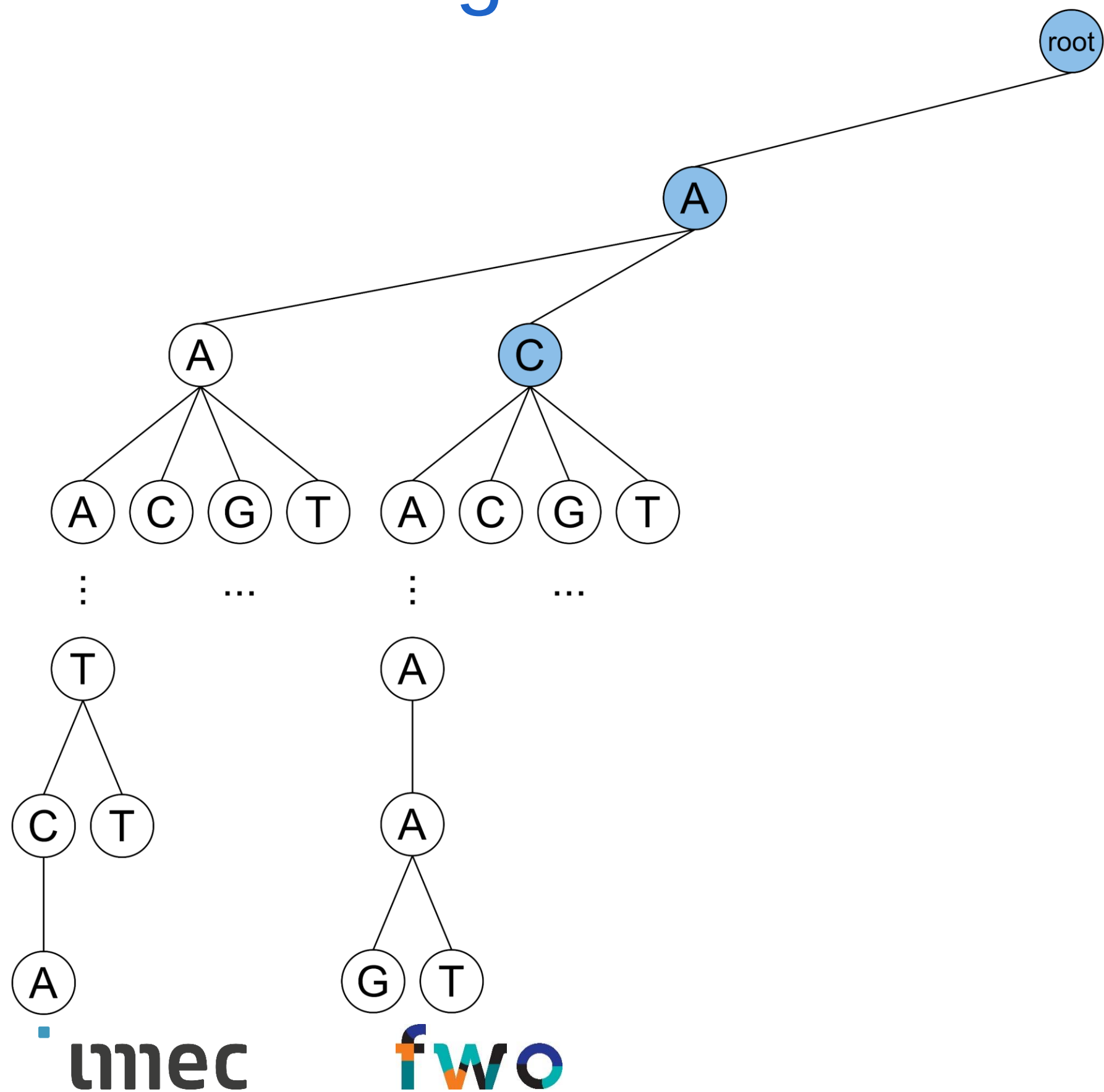
Naive Backtracking



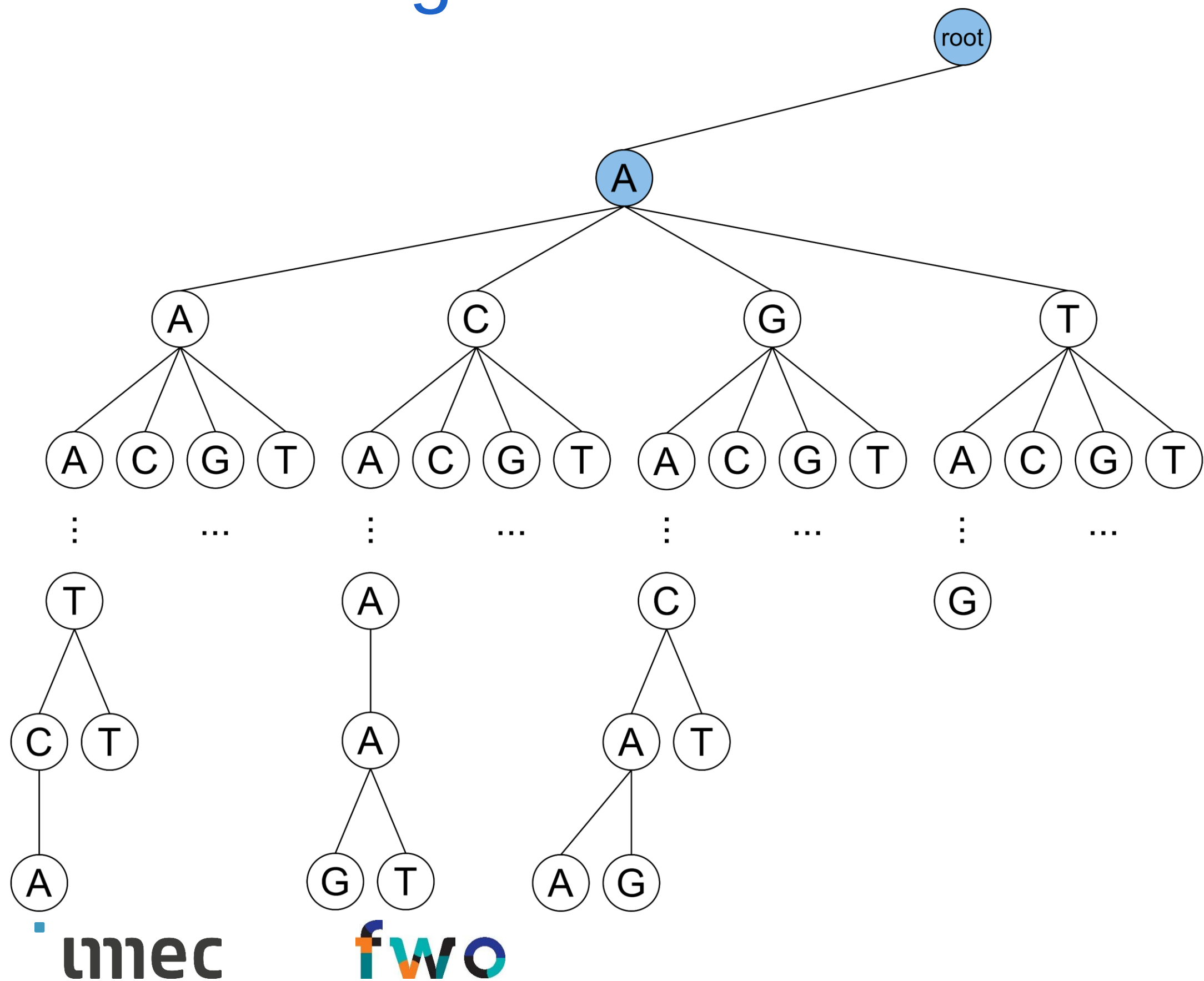
Naive Backtracking



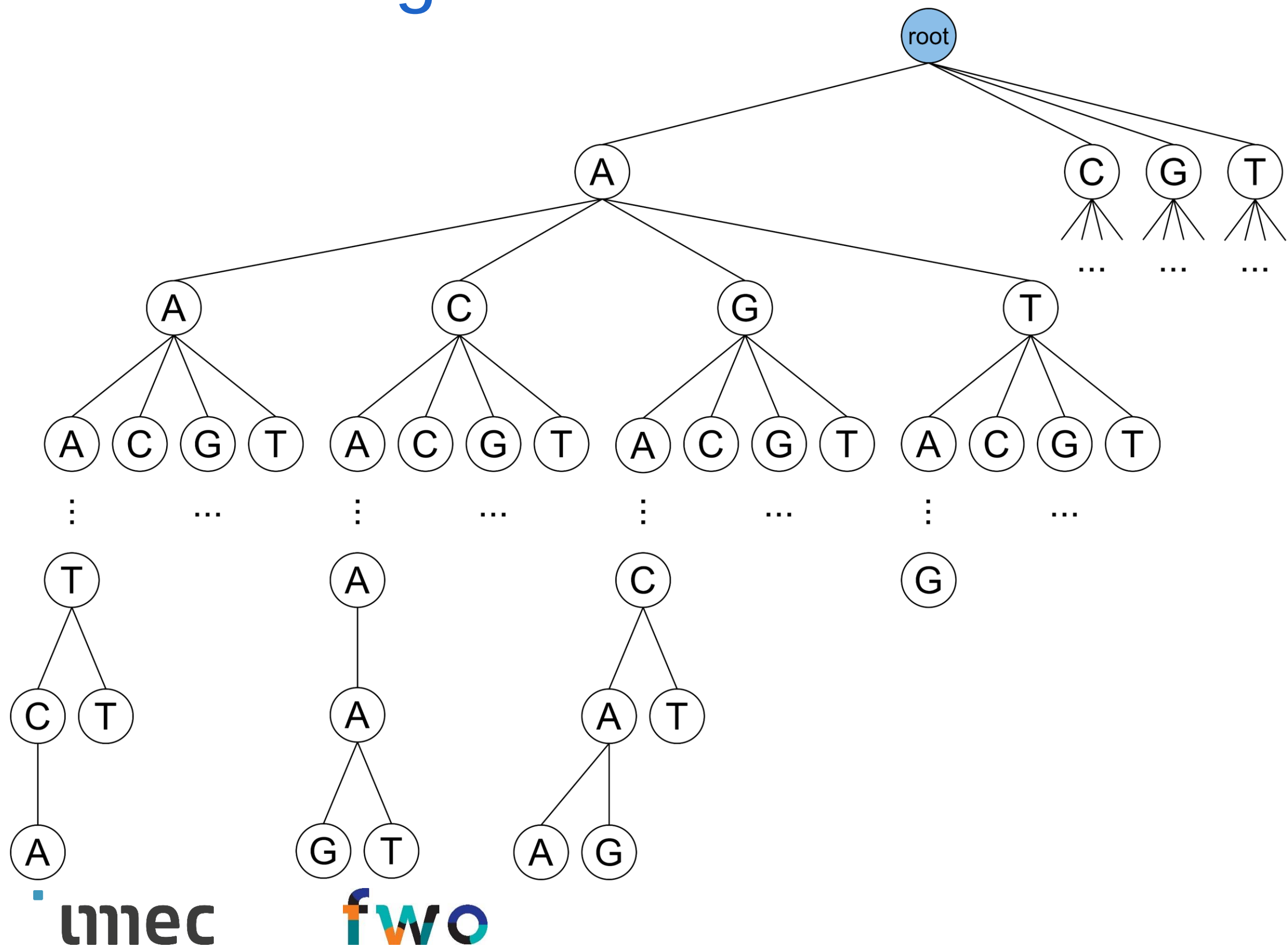
Naive Backtracking



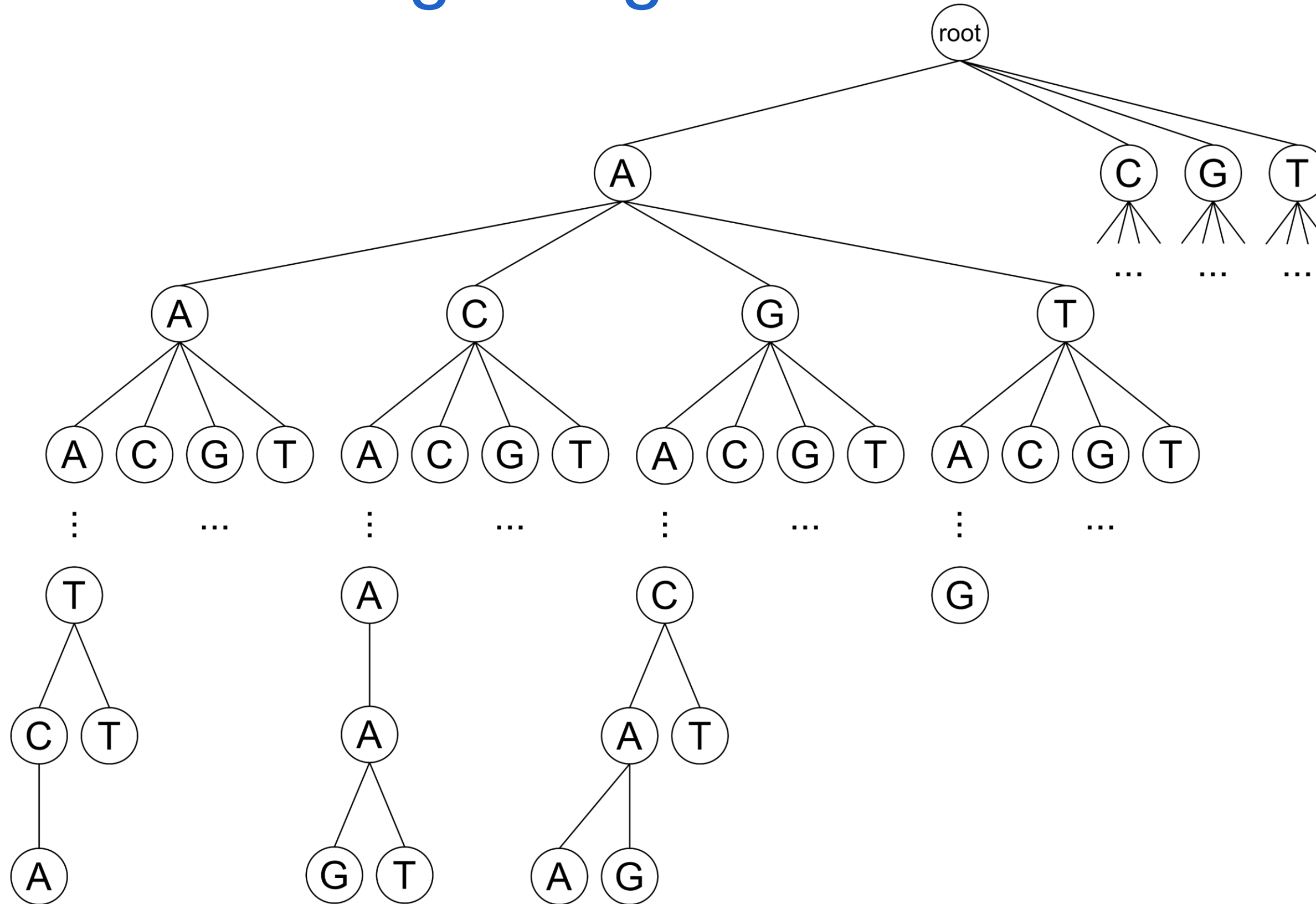
Naive Backtracking



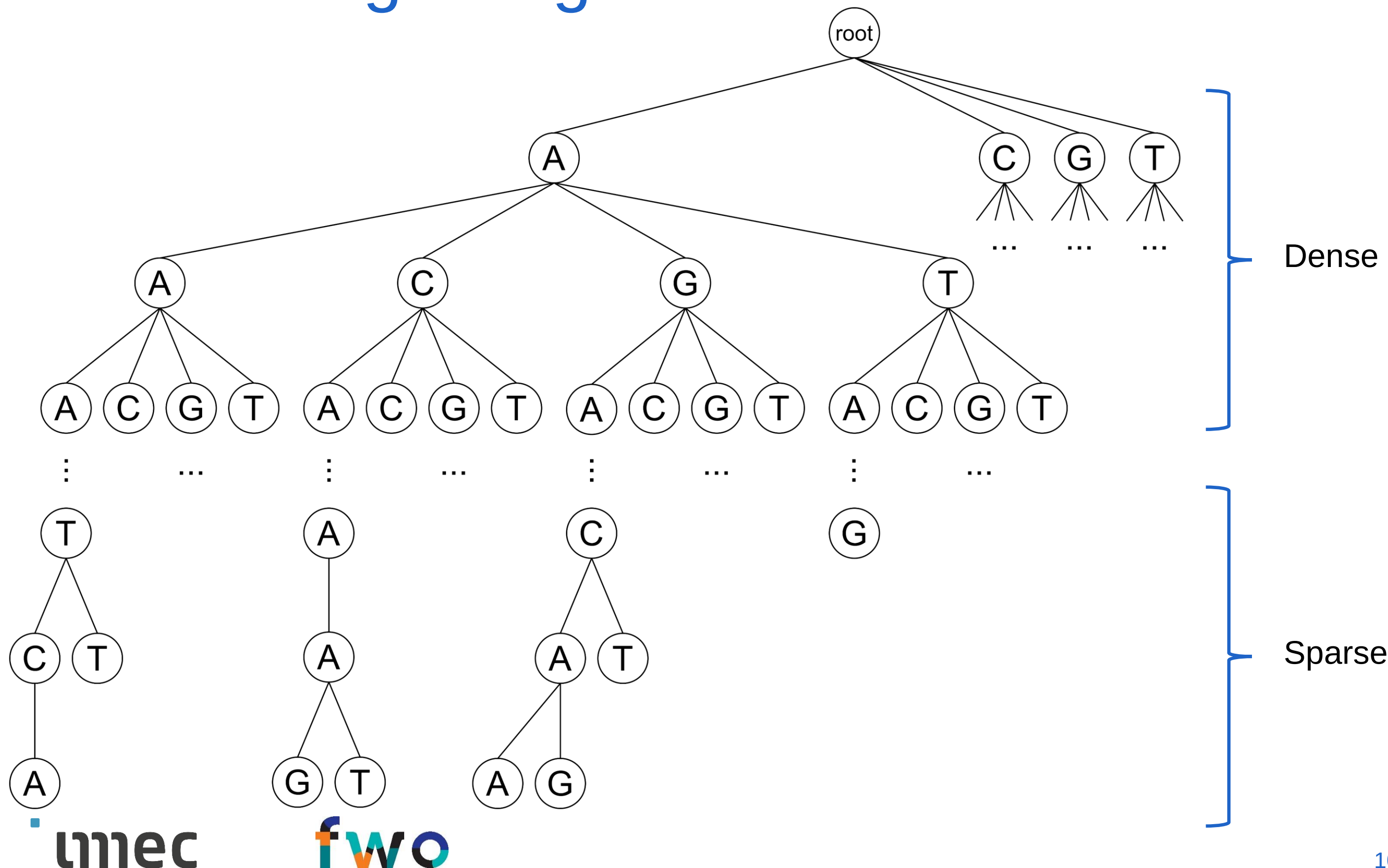
Naive Backtracking



Naive Backtracking using the FM-index



Naive Backtracking using the FM-index



Pigeonhole Principle

Lam et al., 2009

Search pattern: AGACCTGATAAGACG

Occurrences:

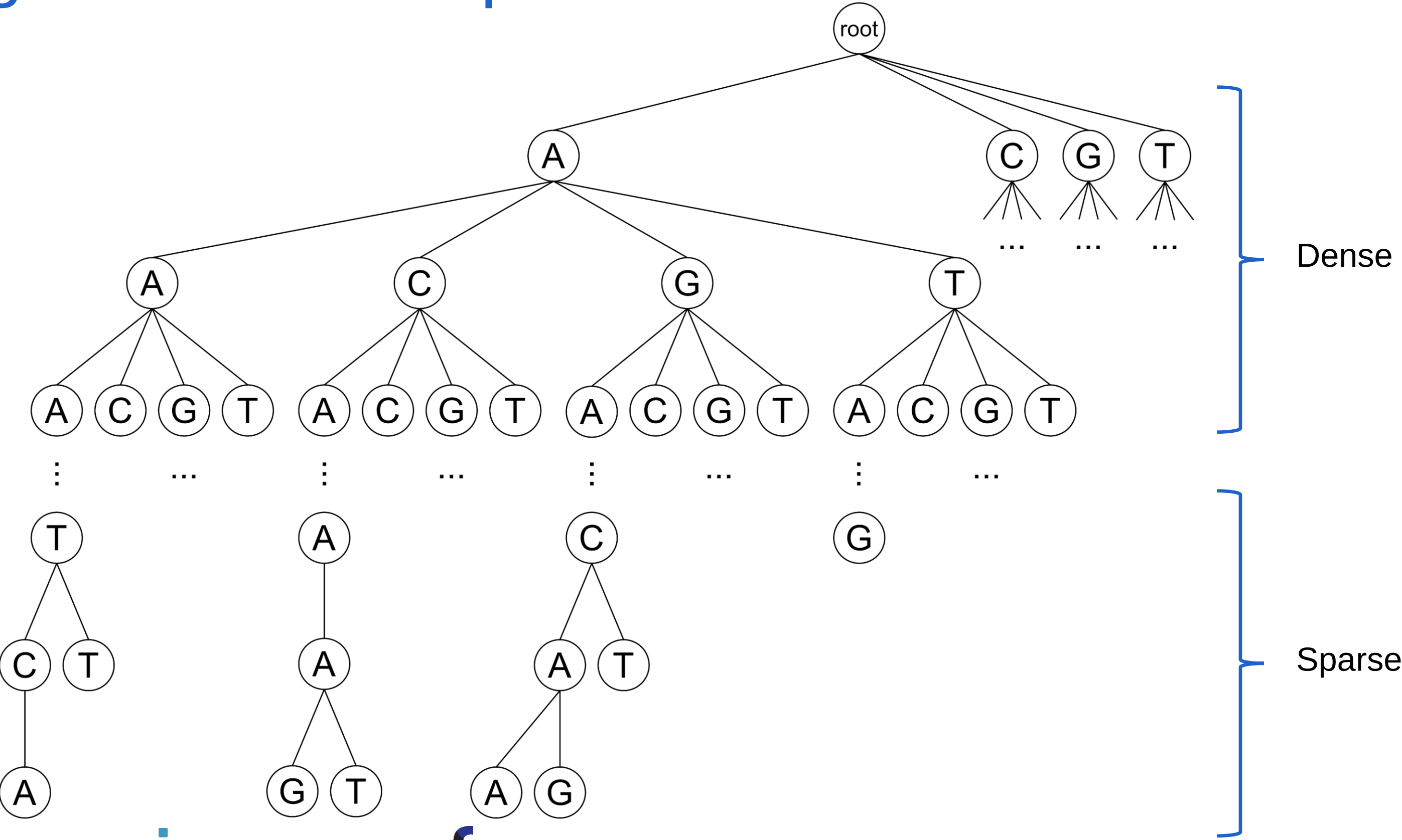
AACCTGATTAGACG
AGACGTGATAAG-CG
AGACCTG-TAGGACG
A--CCTGATAAGACG
AGACCTAATGAGACG
AGACCTGATAAGTTG

Pigeonhole Principle

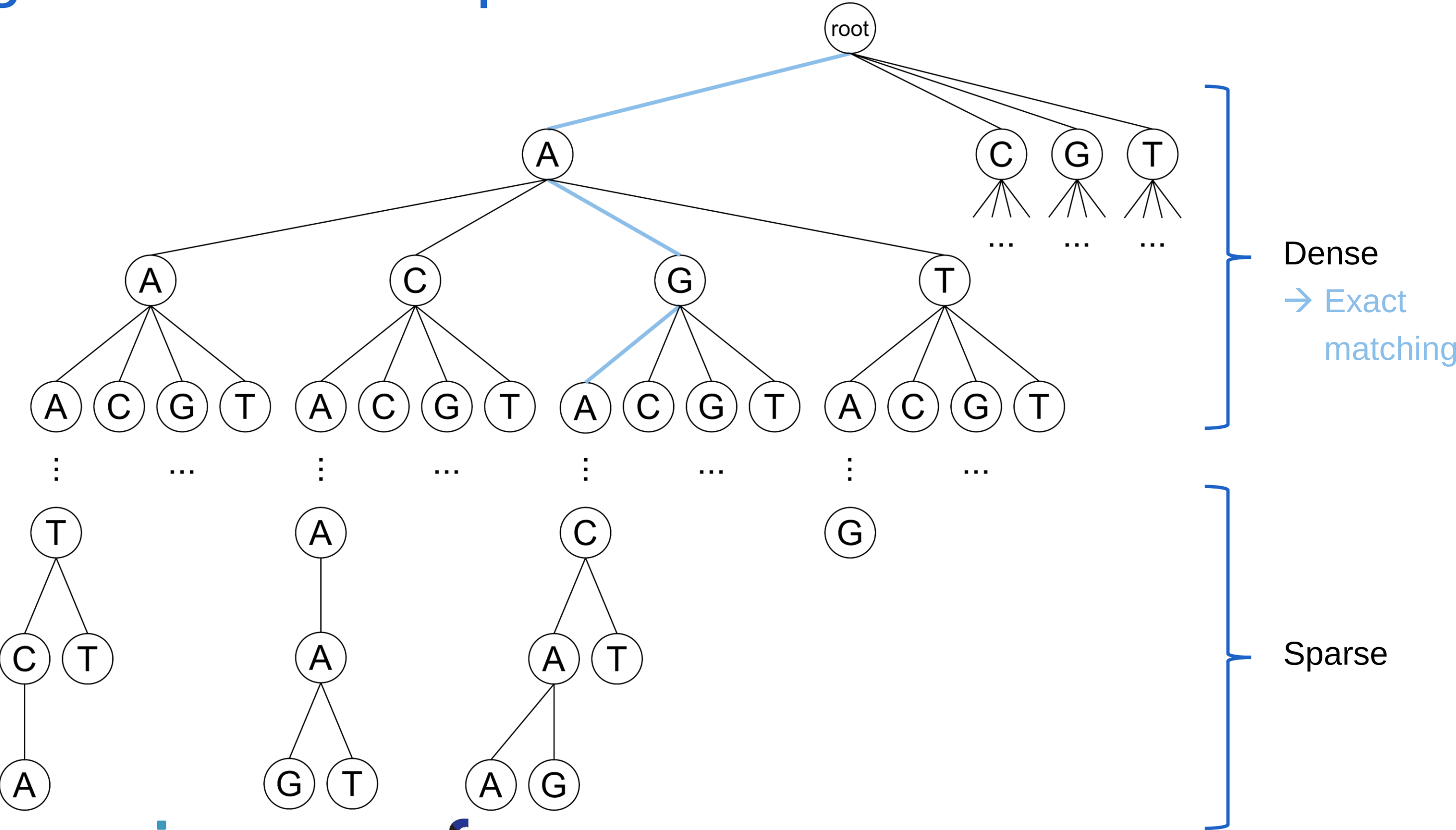
Lam et al., 2009

Search pattern:	AGACC	TGATA	AGACG
Occurrences:	A A ACC	TGATT T	AGACG
	AGAC G	TGATA	AG- C G
	AGACC	TG-TA	G GACG
	A--CC	TGATA	AGACG
	AGACC	T A AT G	AGACG
	AGACC	TGATA	AG TT G

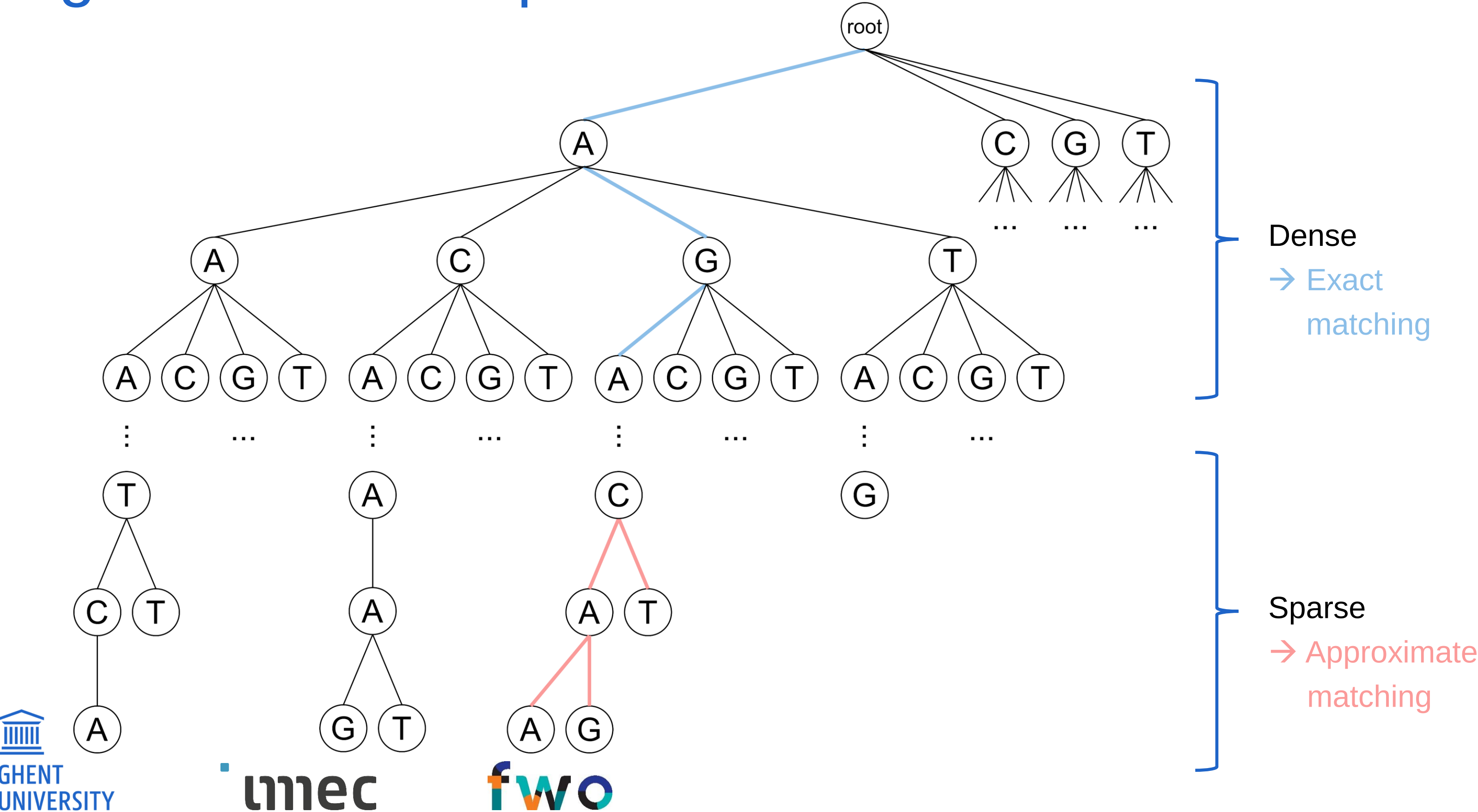
Pigeonhole Principle



Pigeonhole Principle



Pigeonhole Principle



Pigeonhole Principle

Lam et al., 2009

Exact matching

Approximate matching

Search 1:



Search 2:



Search 3:



Bidirectional FM-index

Search Schemes

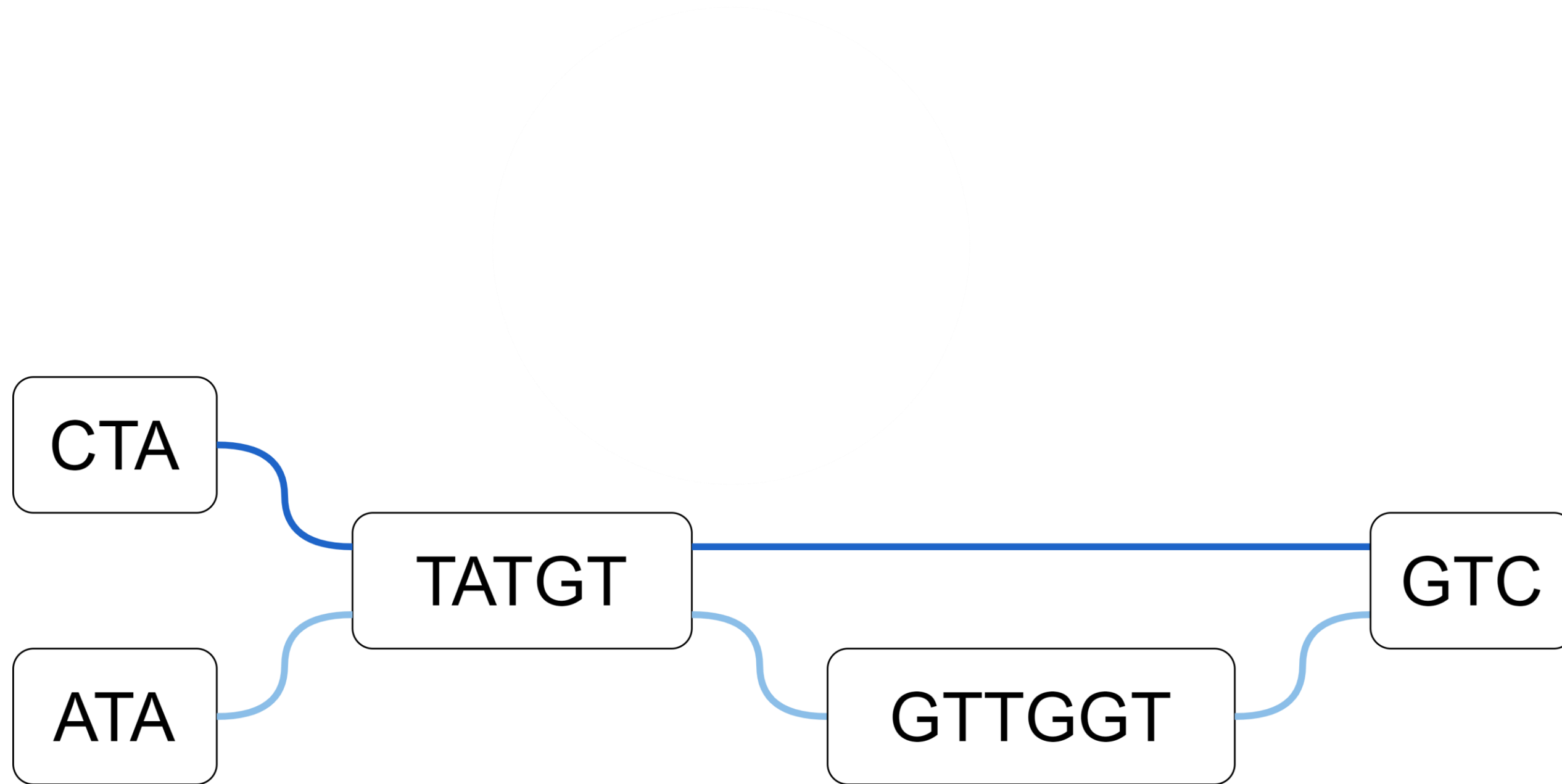
- Search scheme = collection of searches
- We define a search by:
 - The **order** in which the parts are processed
 - The **lower bounds** to the cumulative number of allowed errors after each part has been processed
 - The **upper bounds** to the cumulative number of allowed errors after each part has been processed

Pan-genome Graph Data Structure

Pan-genome Graph Data Structure

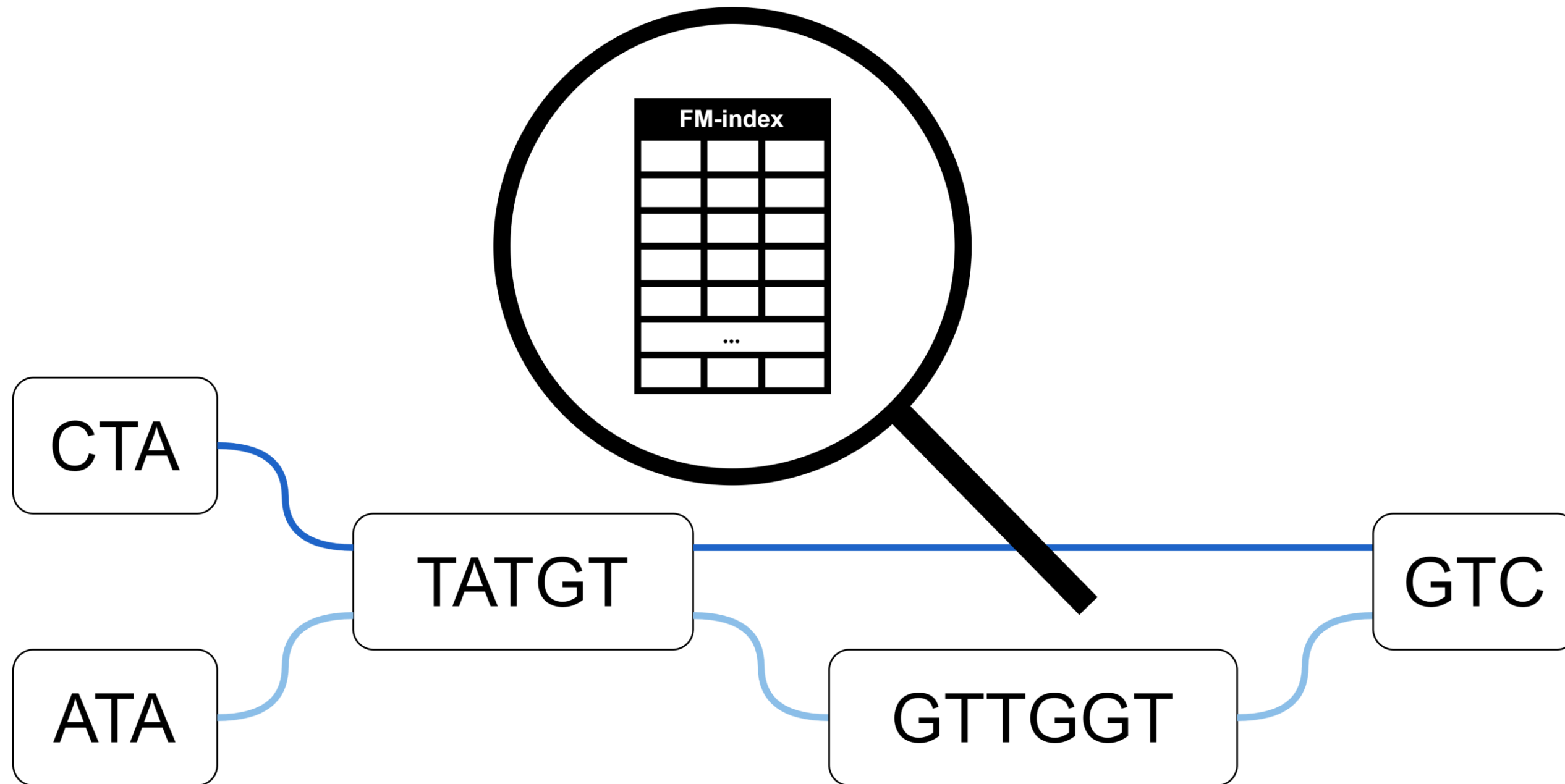
Beller and Ohlebusch, 2016

Dede and Ohlebusch, 2020



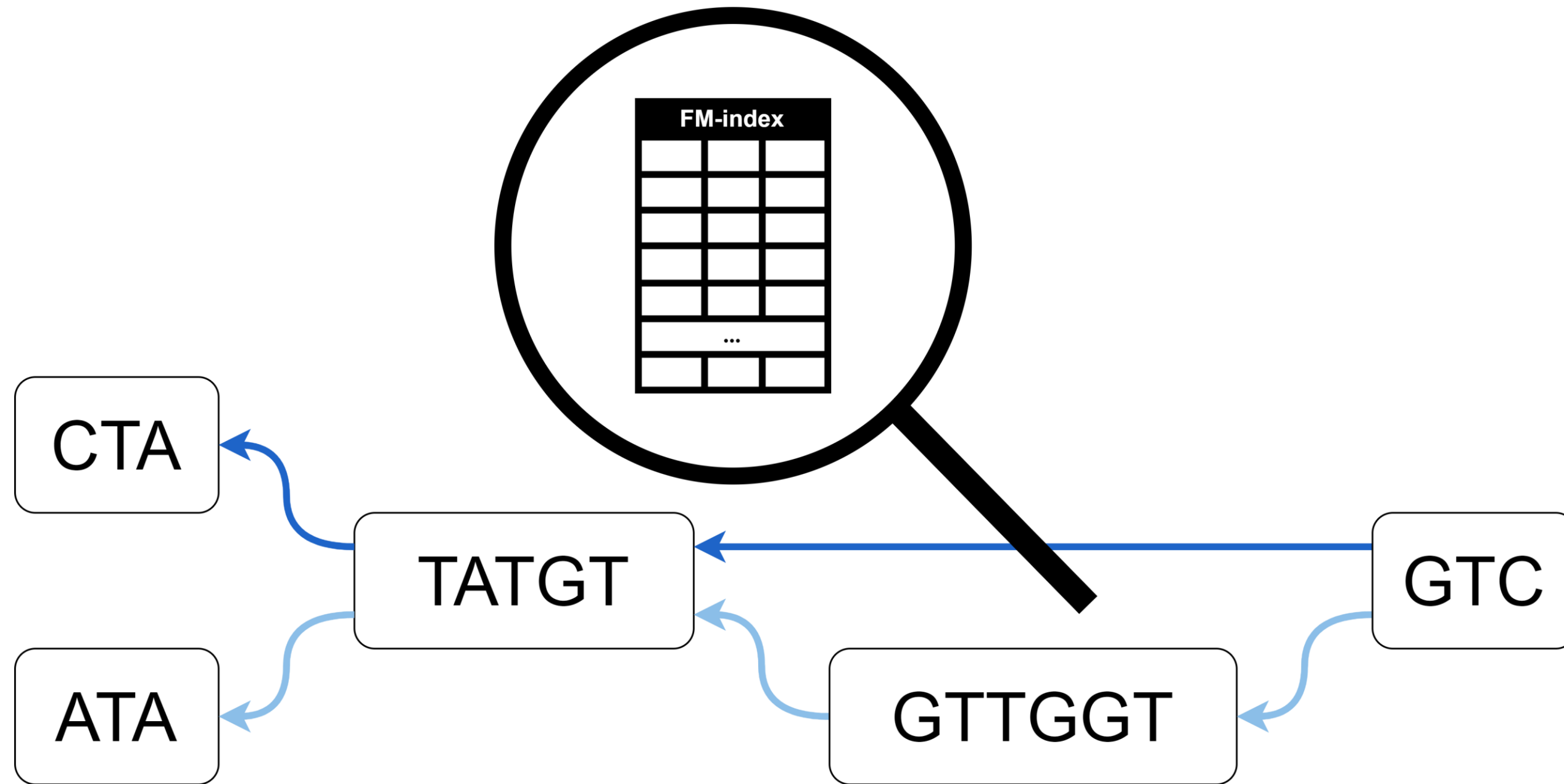
Pan-genome Graph Data Structure

Beller and Ohlebusch, 2016
Dede and Ohlebusch, 2020
Ferragina and Manzini, 2000



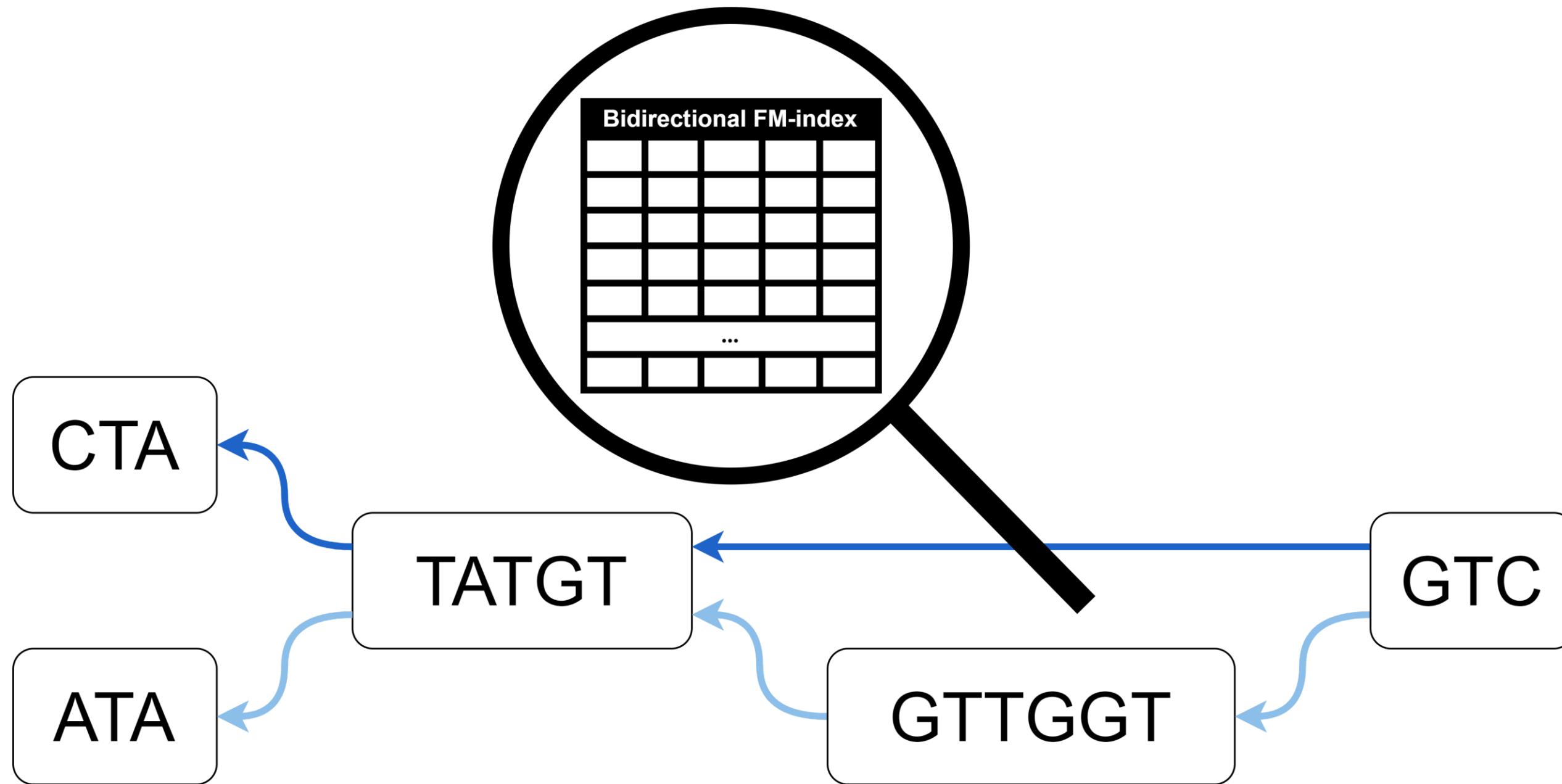
Pan-genome Graph Data Structure

Beller and Ohlebusch, 2016
Dede and Ohlebusch, 2020
Ferragina and Manzini, 2000



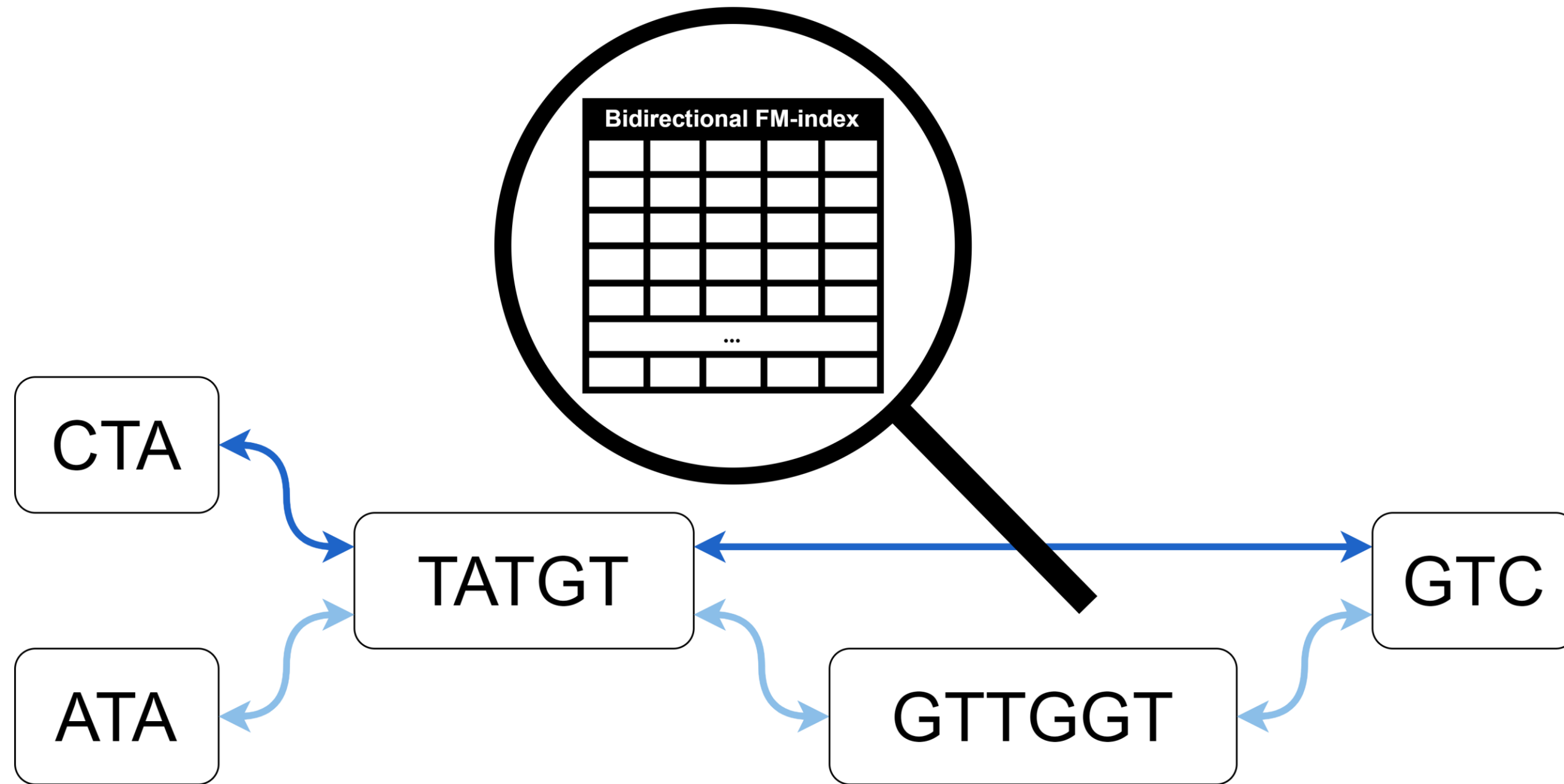
Pan-genome Graph Data Structure

Beller and Ohlebusch, 2016
Dede and Ohlebusch, 2020
Ferragina and Manzini, 2000
Lam et al., 2009



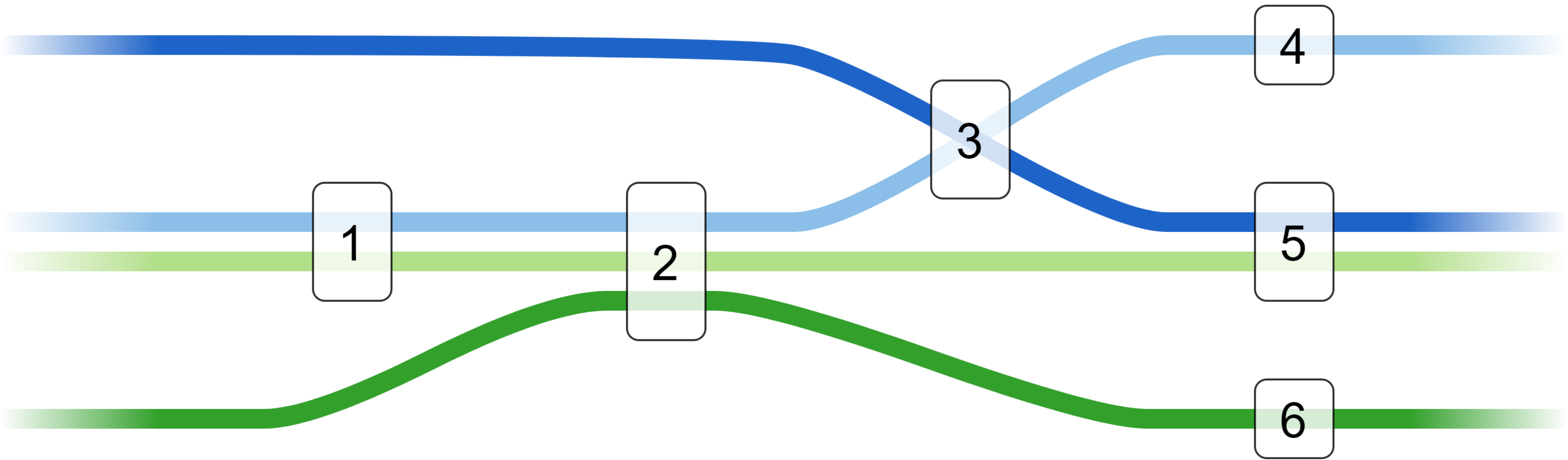
Pan-genome Graph Data Structure

Beller and Ohlebusch, 2016
Dede and Ohlebusch, 2020
Ferragina and Manzini, 2000
Lam et al., 2009

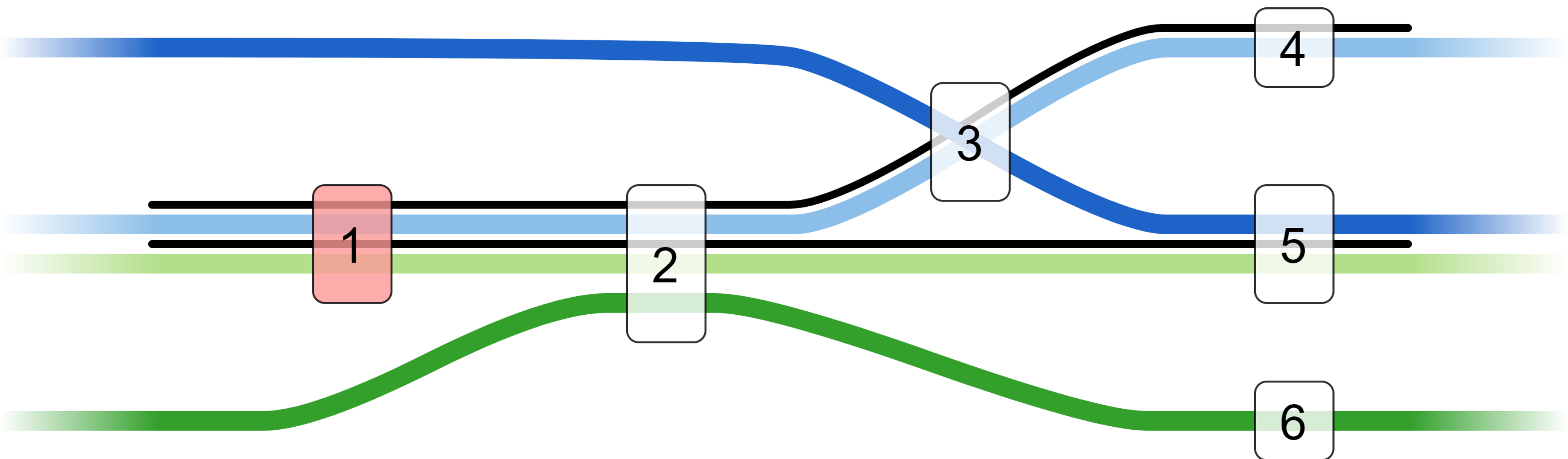


Pattern Matching Against the Pan-genome Graph

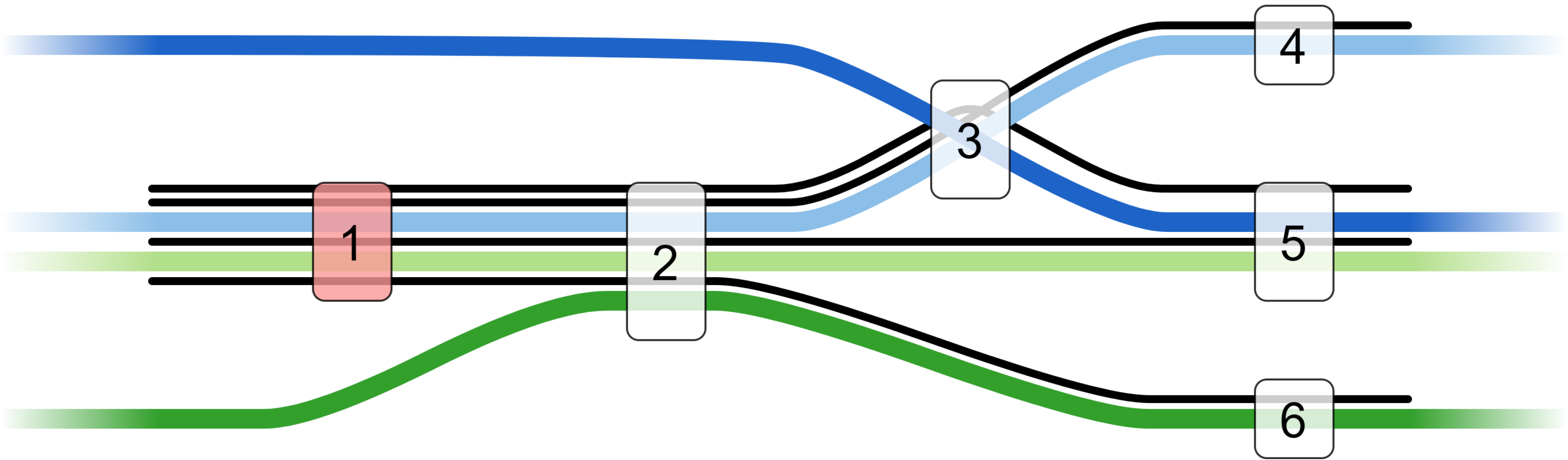
Strain-Fixed Pattern Matching



Strain-Fixed Pattern Matching



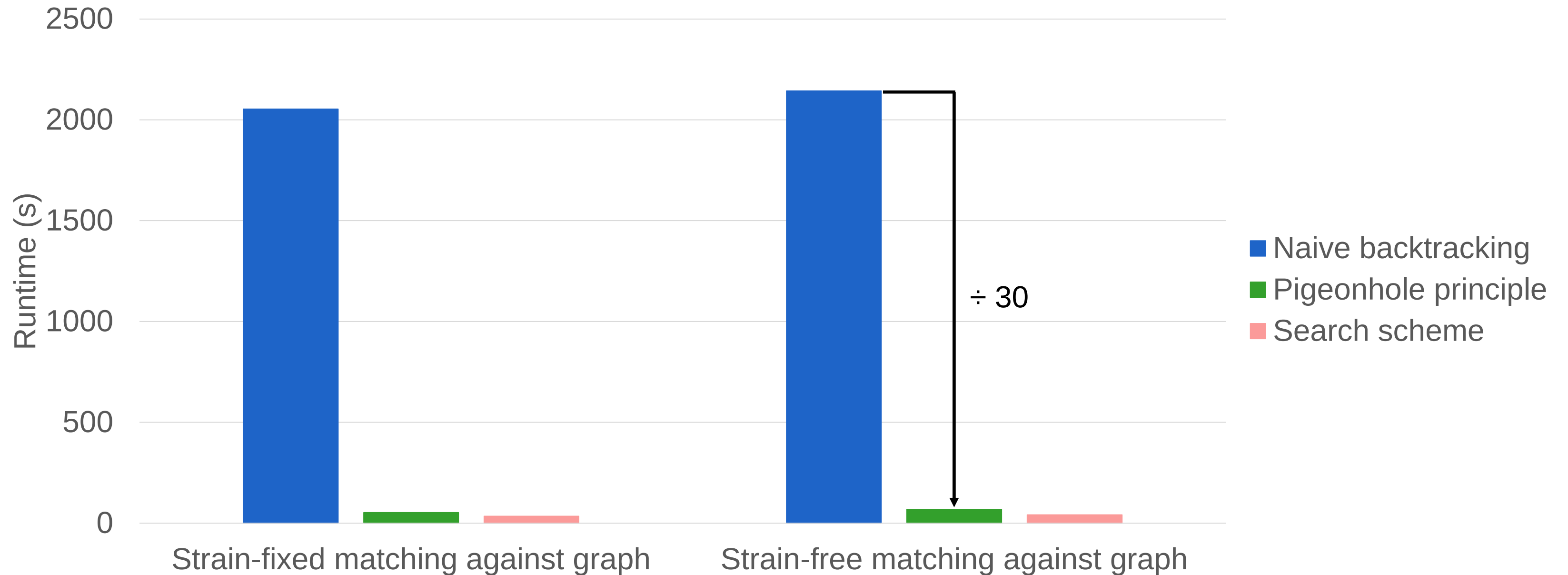
Strain-Free Pattern Matching



Results

Nexus Performance

Matching 100 000 Illumina reads (101 bp) and their reverse complements to a pan-genome of 2 human genome builds (de Bruijn parameter $k = 77$), with a maximum edit distance of 2



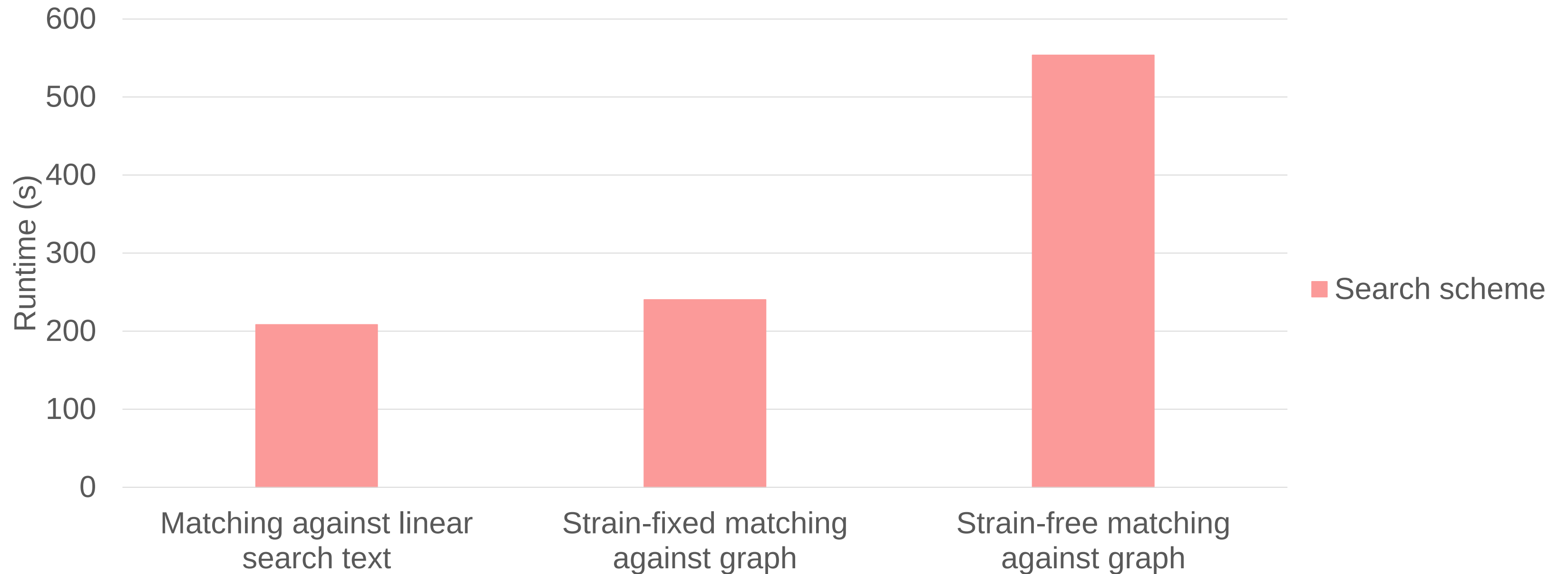
Nexus Performance

Matching 100 000 Illumina reads (101 bp) and their reverse complements to a pan-genome of 2 human genome builds (de Bruijn parameter $k = 77$), with a maximum edit distance of 4



Nexus Performance

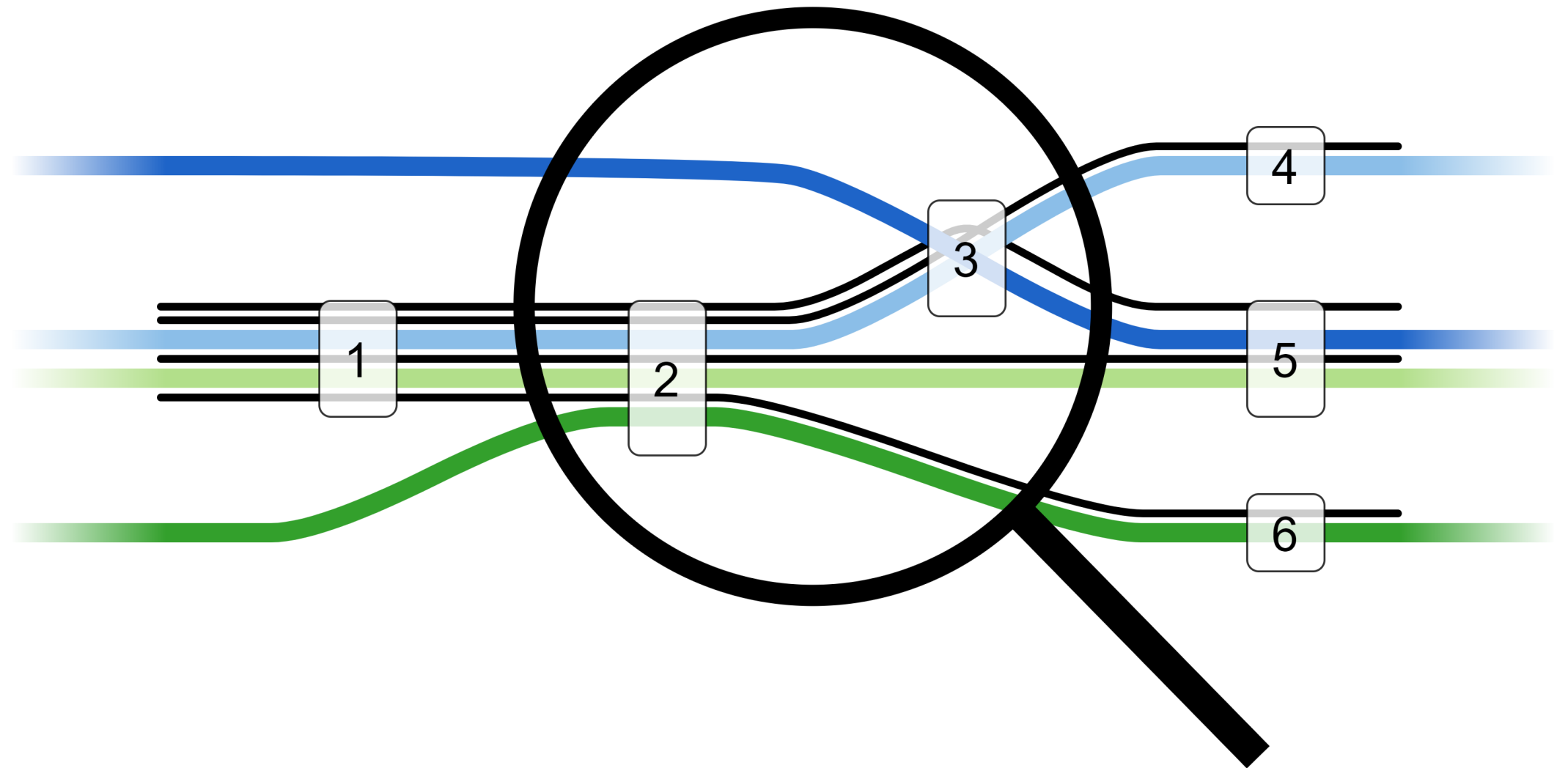
Matching 100 000 Illumina reads (101 bp) and their reverse complements to a pan-genome of 2 human genome builds (de Bruijn parameter $k = 77$), with a maximum edit distance of 4



Nexus



github.com/biointec/nexus



Lore Depuydt

FWO PhD fellow

DEPARTMENT OF INFORMATION TECHNOLOGY (EA05)
IDLAB - INTERNET TECHNOLOGY AND DATA SCIENCE LAB

E Lore.Depuydt@ugent.be

www.ugent.be

 Universiteit Gent
 @ugent
 @ugent
 Ghent University