Biodiversity Computing Research

Alexandros Stamatakis[1,2,3]

Georgios Koutsovoulos[1]

Noah Wahl[1]

1. Institute of Computer Science, Foundation for Research and Technology - Hellas

2. Heidelberg Institute for Theoretical Studies

3. Institute of Theoretical Informatics, Karlsruhe Institute of Technology

www.biocomp.gr (Crete lab)

www.exelixis-lab.org (Heidelberg lab)

Outline

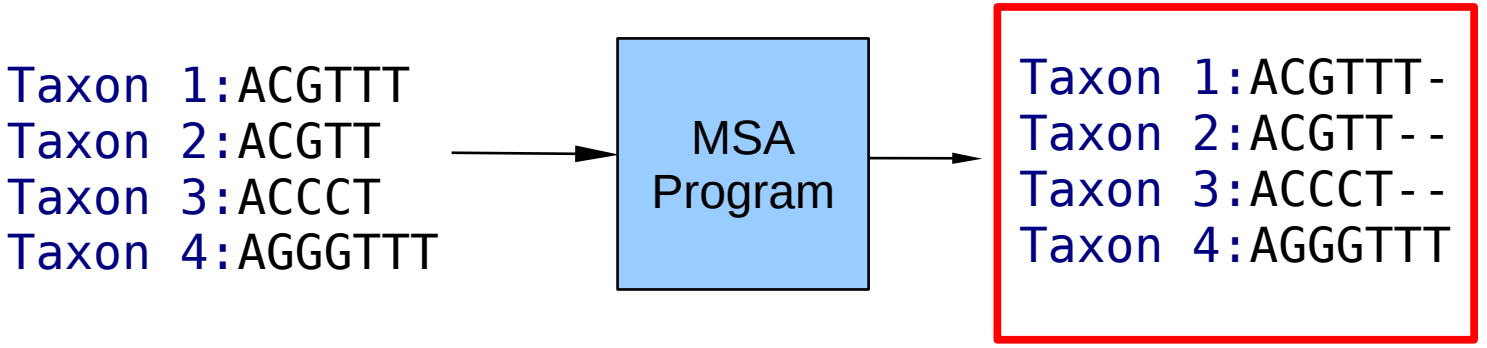
- **Alexis: Predicting, Propagating & Using Uncertainty**
- Giorgos: Making Sense of Metabarcoding Data
- Noah: Gene Tree – Species Tree reconciliation

Tree Inference Pipeline

```
Taxon 1:ACGTTT
Taxon 2:ACGTT
Taxon 3:ACCCT
Taxon 4:AGGGTTT
```

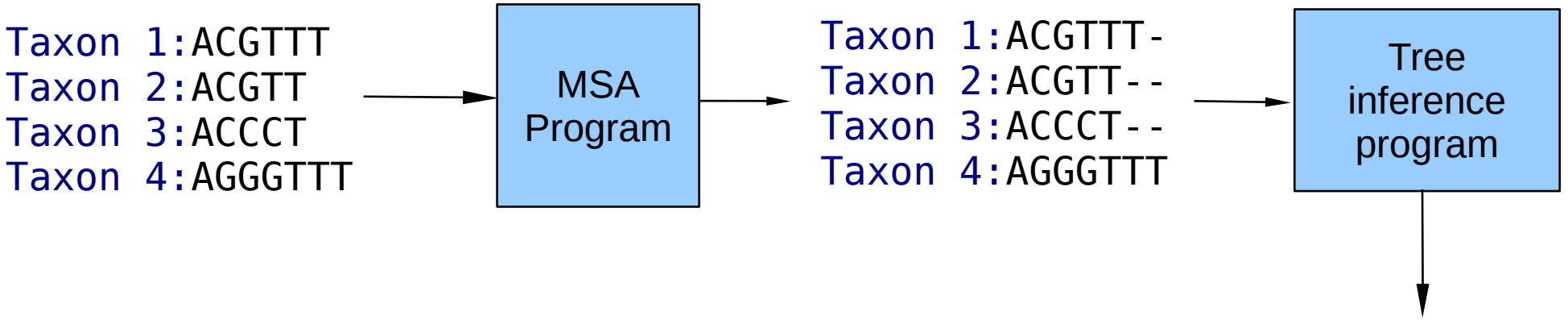
Orthology Clustering:
Mostly *ad hoc* methods →
no widely used uncertainty
quantification approach

Tree Inference Pipeline

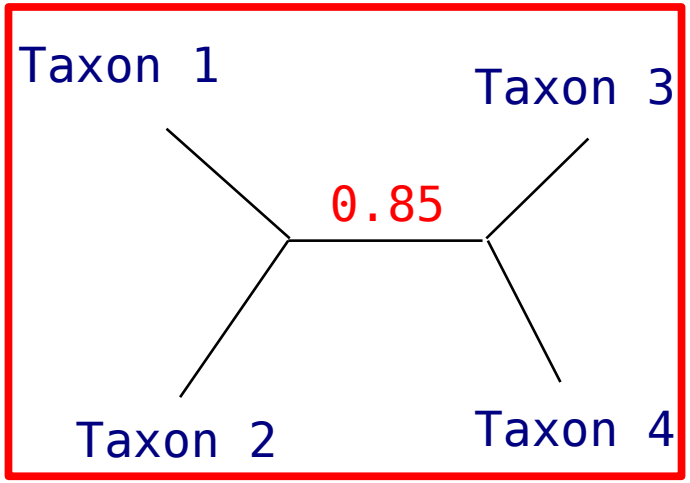


Multiple Sequence Alignment:
Mostly *ad hoc* methods →
no widely used uncertainty quantification approach, but
`Muscle 5` tool → ensembles

Tree Inference Pipeline



Phylogenetic Inference:
A long history of explicit uncertainty models
Bootstrap Methods for Maximum Likelihood
Posterior Probabilities for Bayesian Inference via MCMC



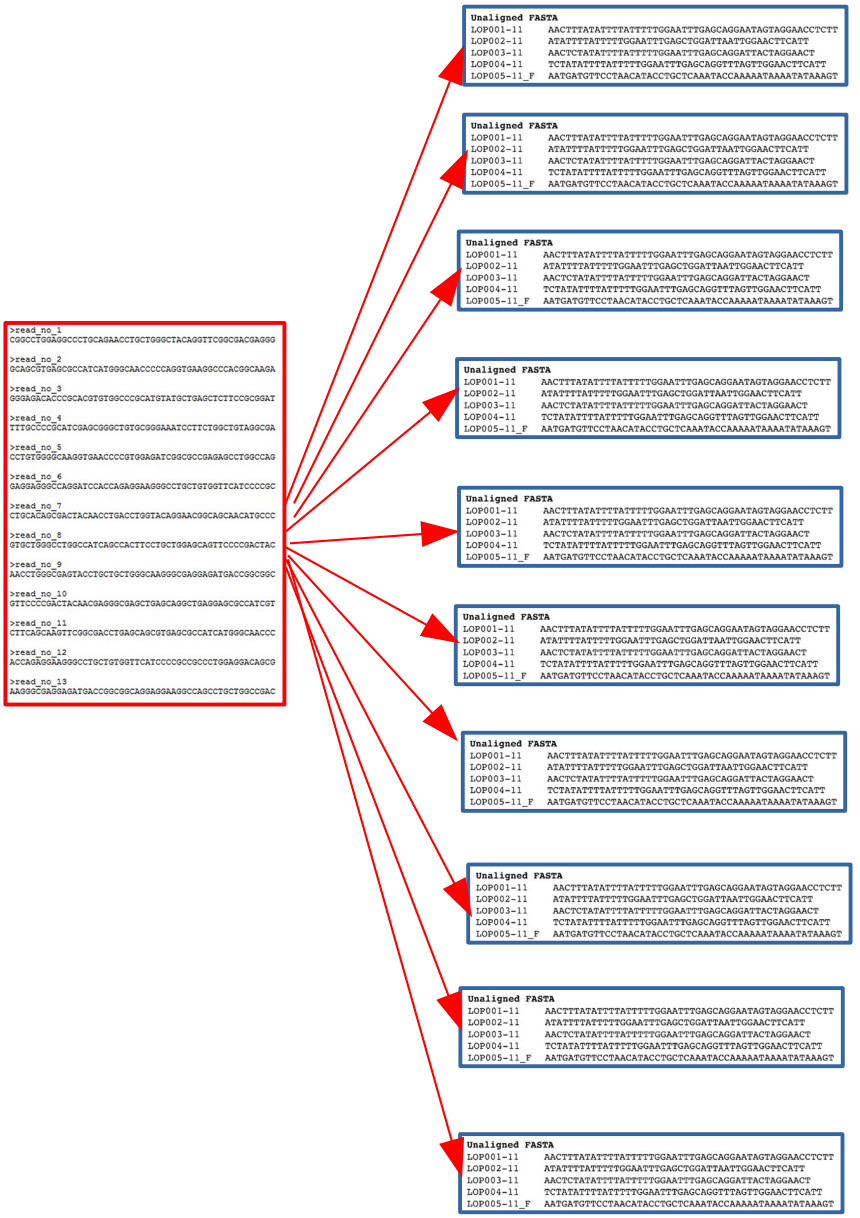
Naively Propagating Uncertainty

One bunch of sequences

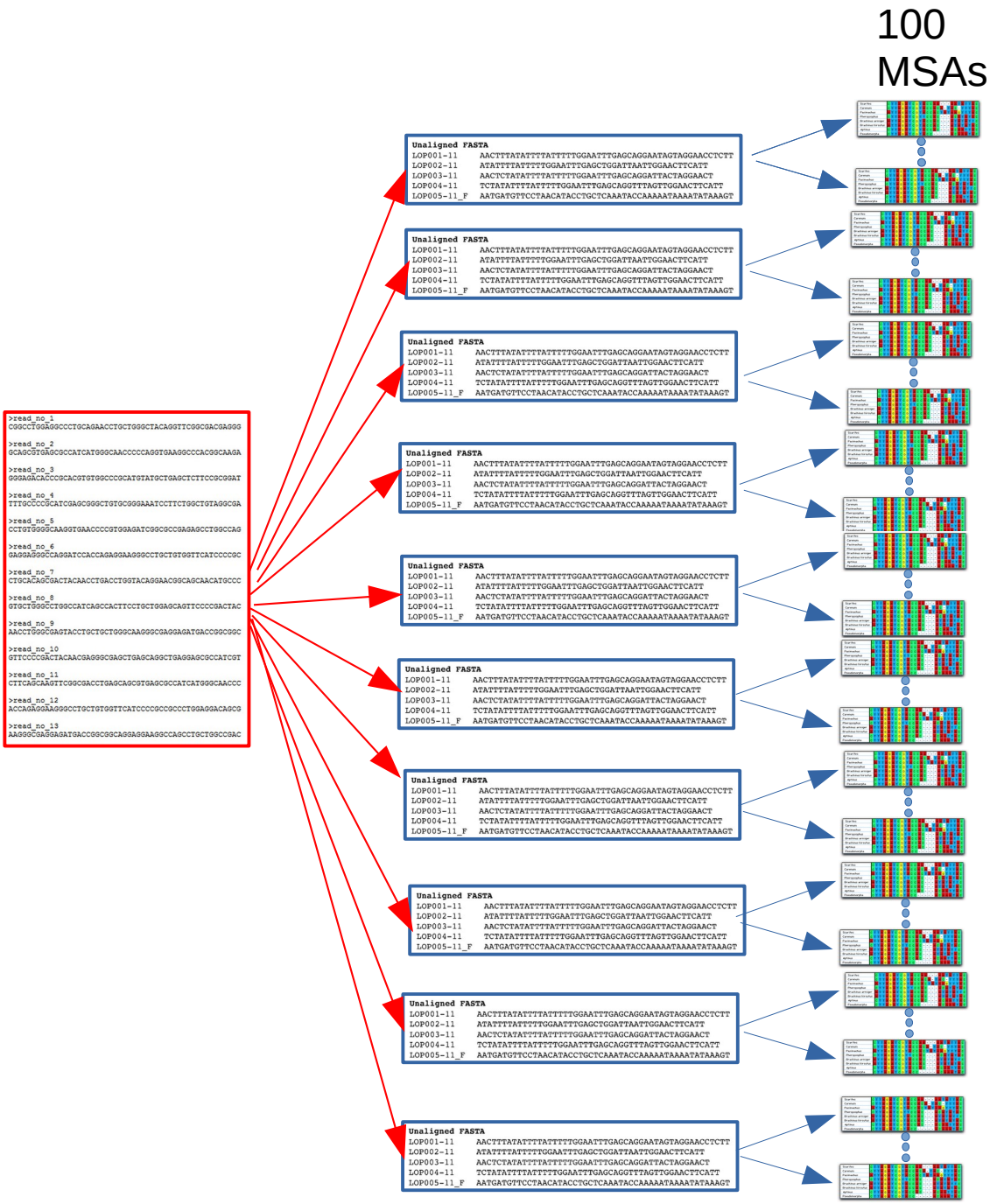


Naively Propagating Uncertainty

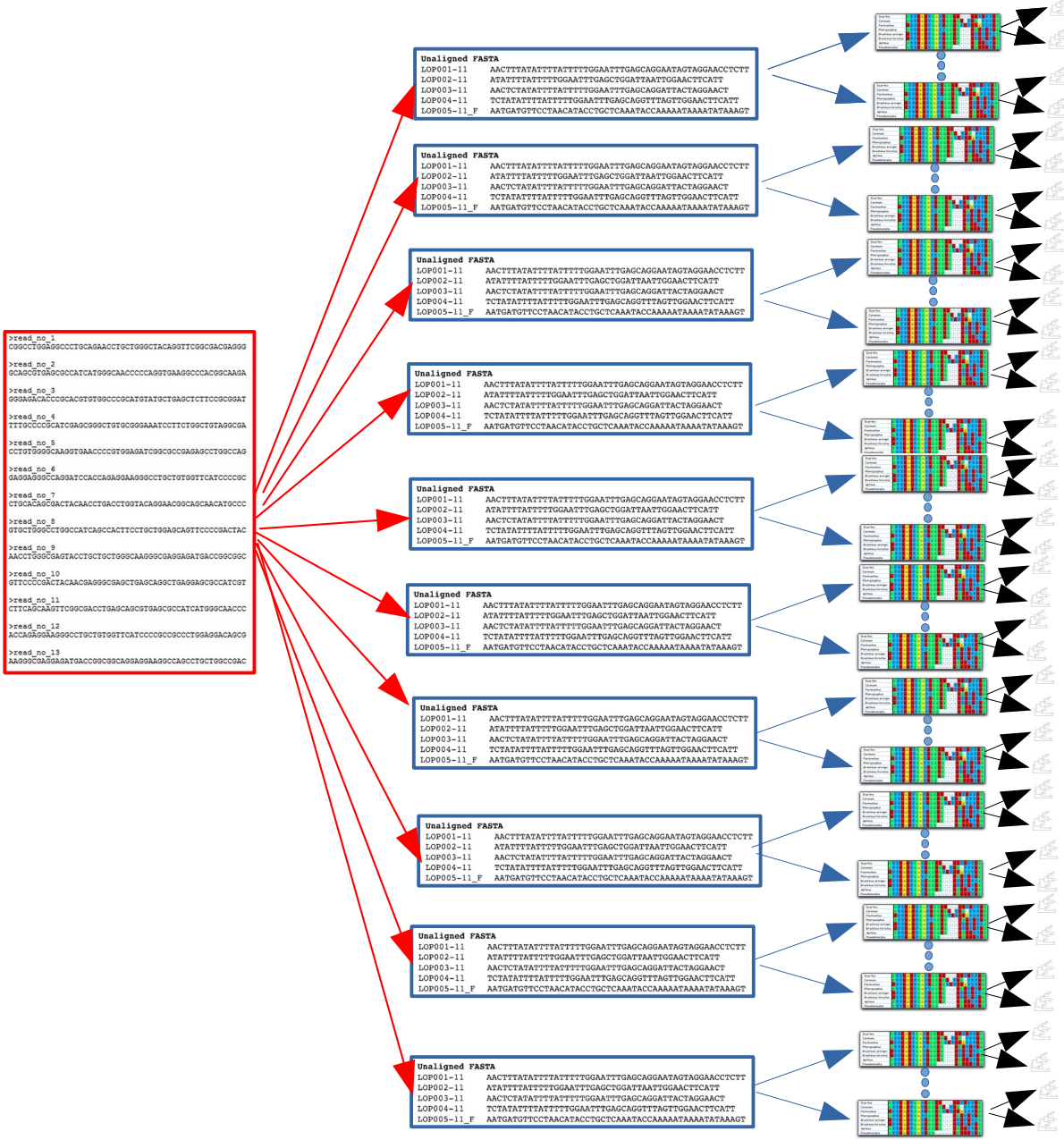
10 orthologous clusters



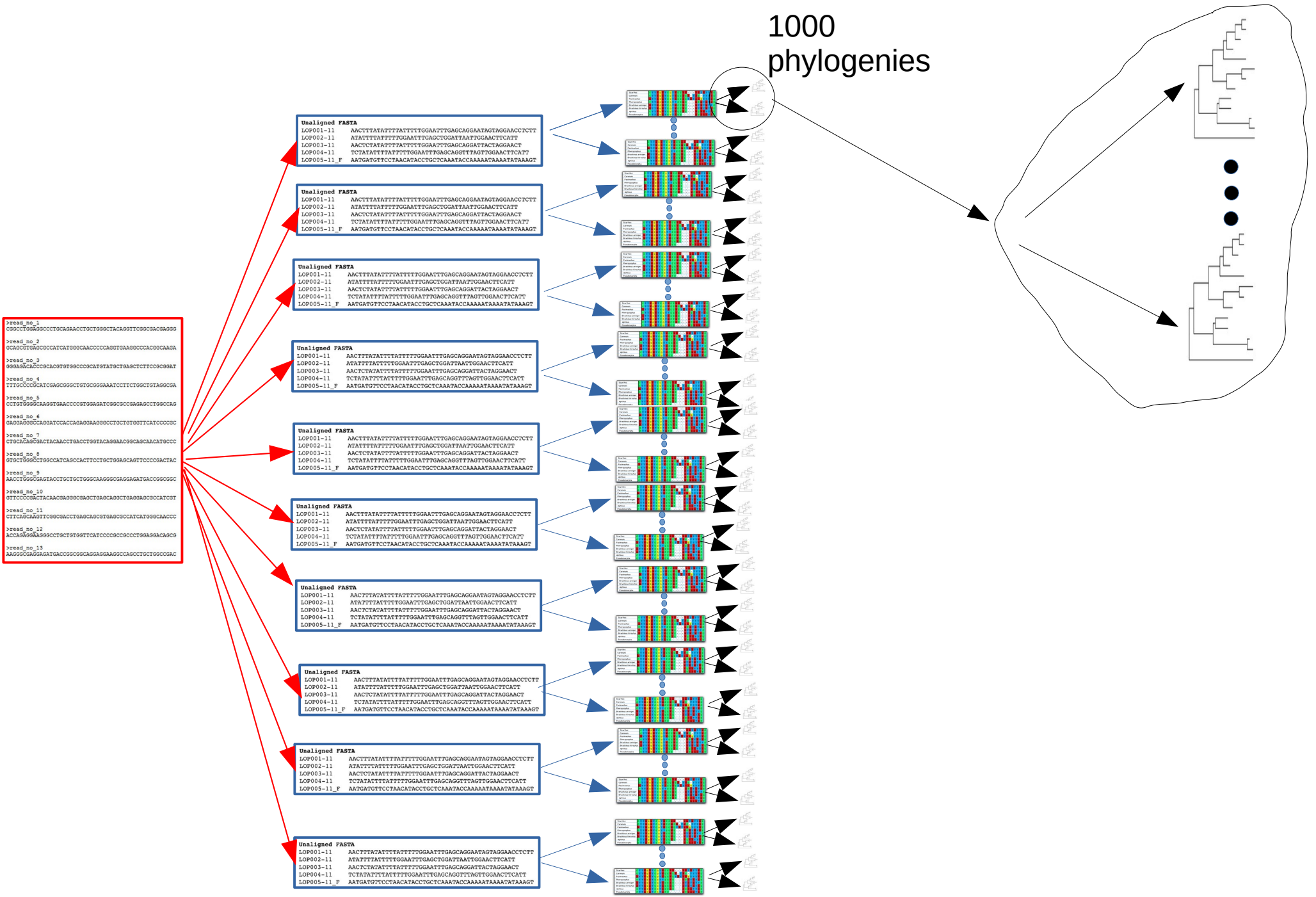
Naively Propagating Uncertainty



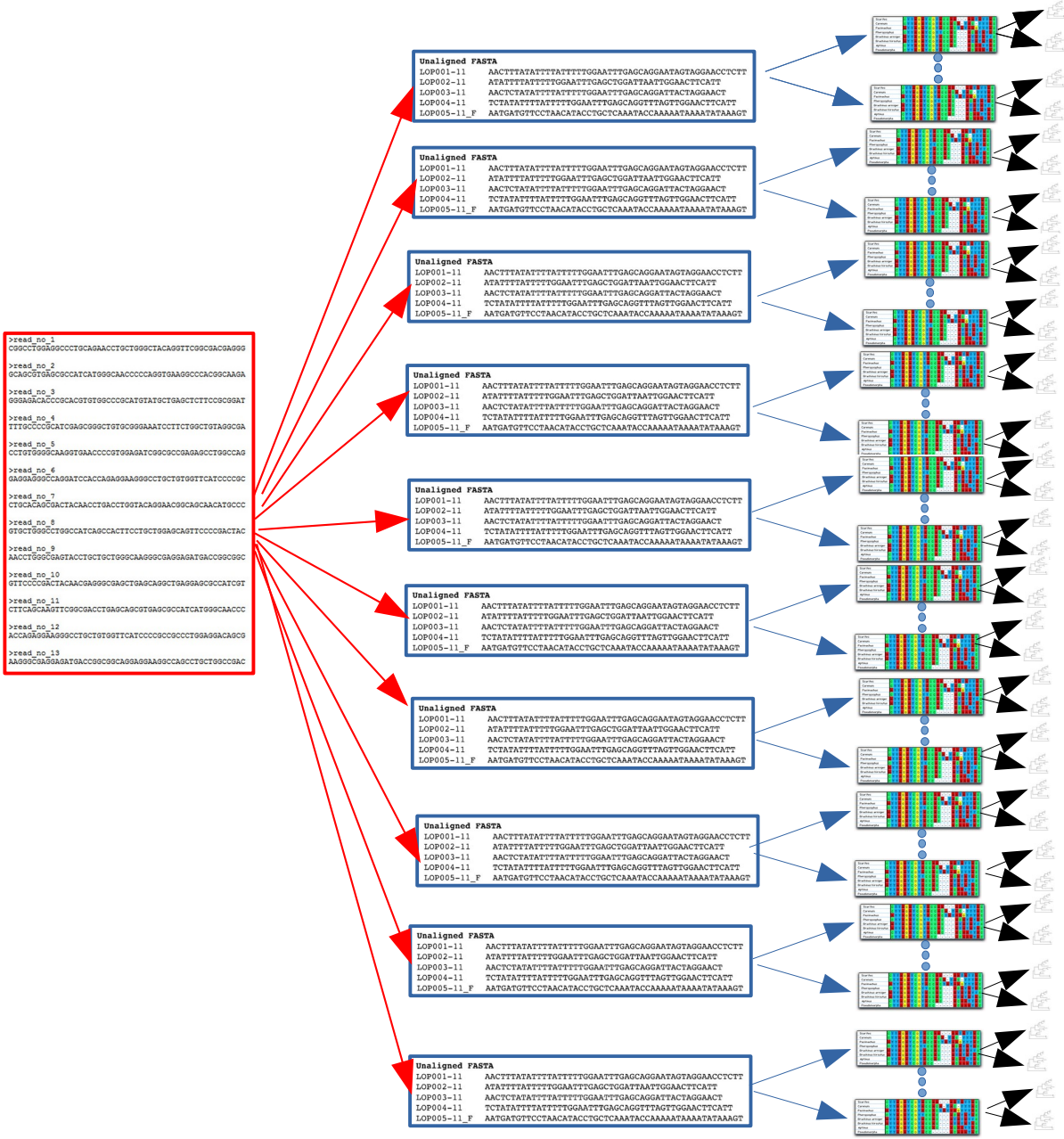
Naively Propagating Uncertainty



Naively Propagating Uncertainty



Naively Propagating Uncertainty



Predicting Uncertainty: Step-by-Step

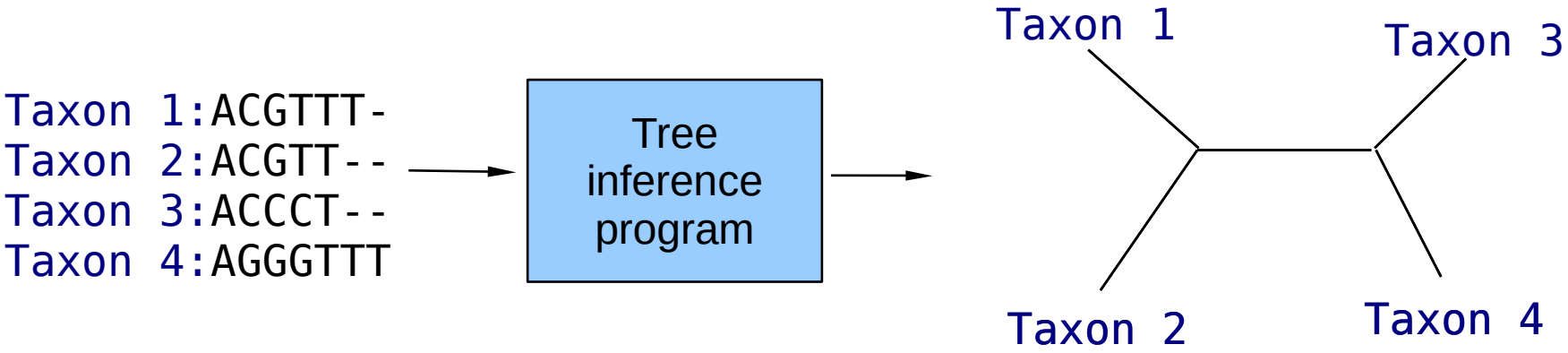
- **Reasonable** definition of difficulty =
    - Uncertainty
    - Variance
    - Dispersion
    - Signal Strength
- **MSA difficulty:** how different are the MSAs in an ensemble for a given unaligned sequence set
- **Phylogenetic difficulty:** How topologically different are equally good trees
- **Easy interpretation:** difficulty between 0 and 1

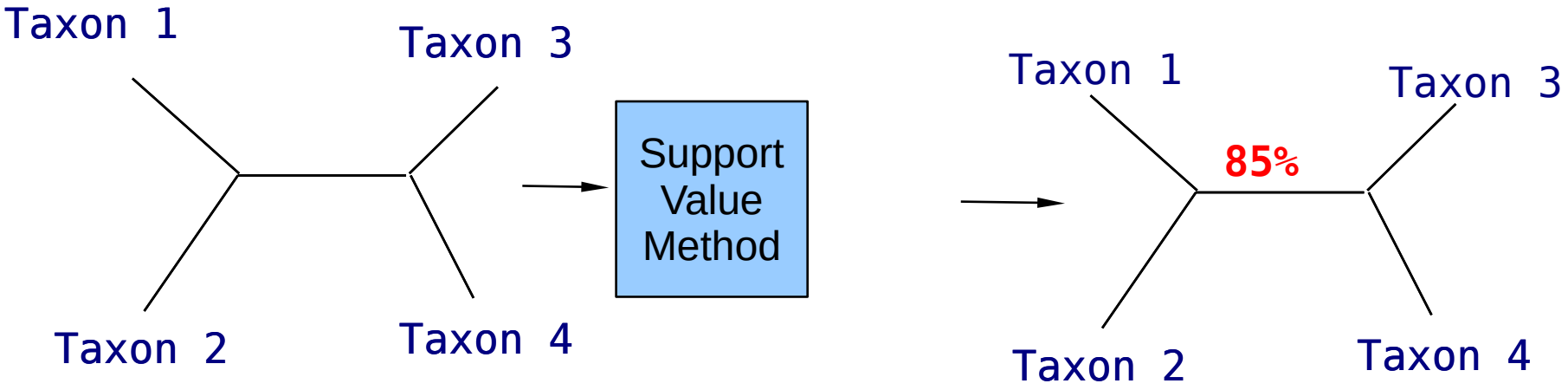
  0 → easy dataset → most programs, models, algorithms will yield the same result

  1 → difficult dataset → a total mess
- Generate labels & train a model

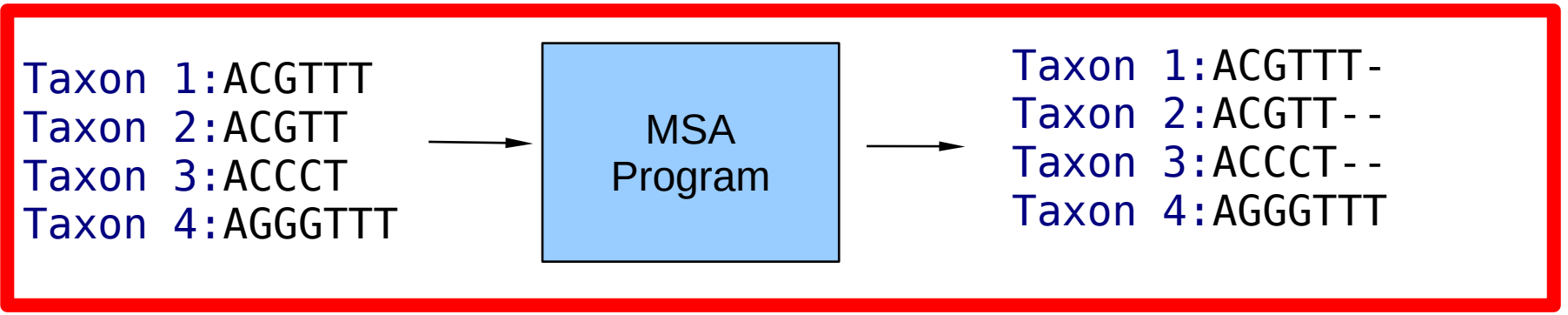
Predicting Uncertainty

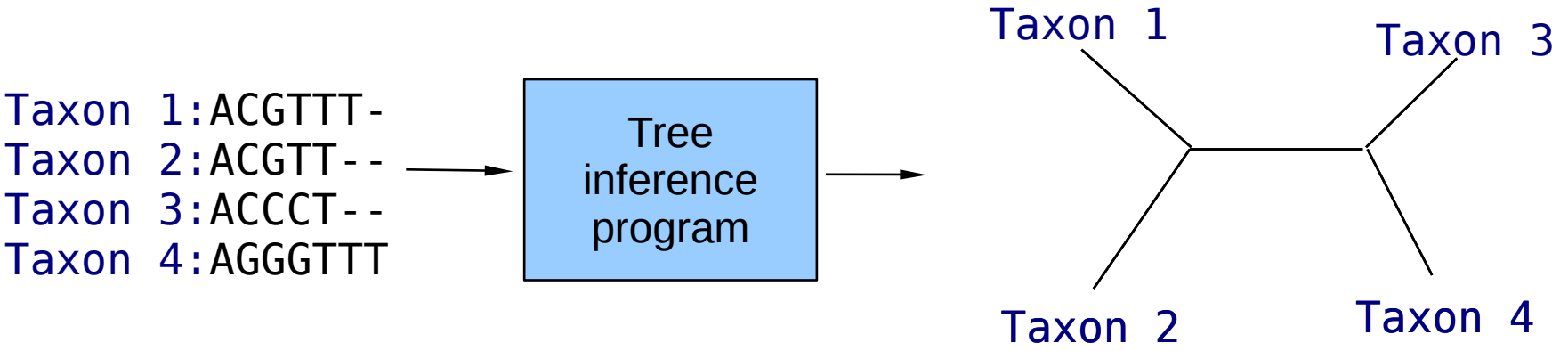


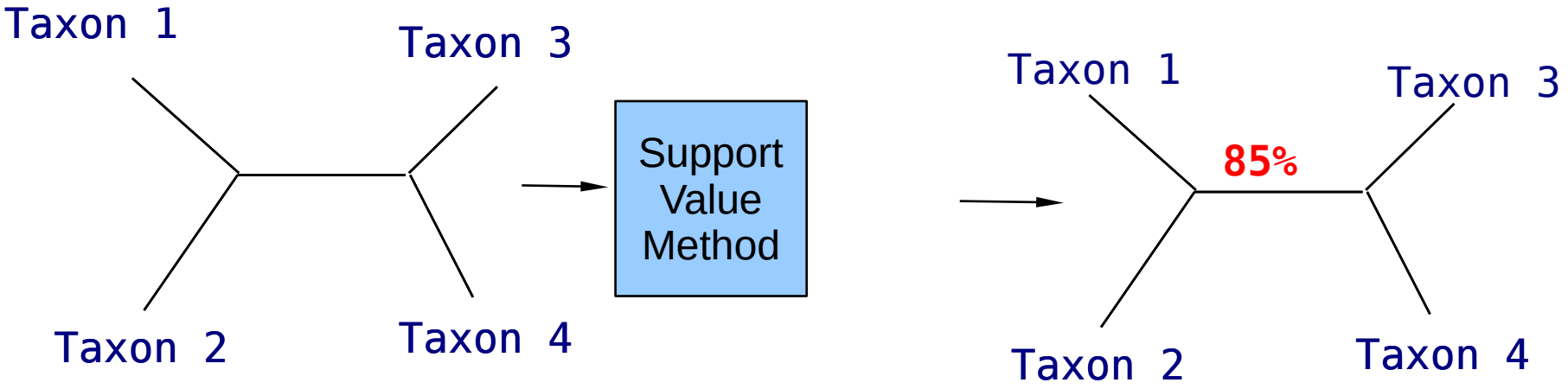




Predicting MSA Difficulty







AlDiScore

New Results

Follow this preprint

Quantifying and Predicting the Difficulty of Multiple Sequence Alignment with AlDiScore

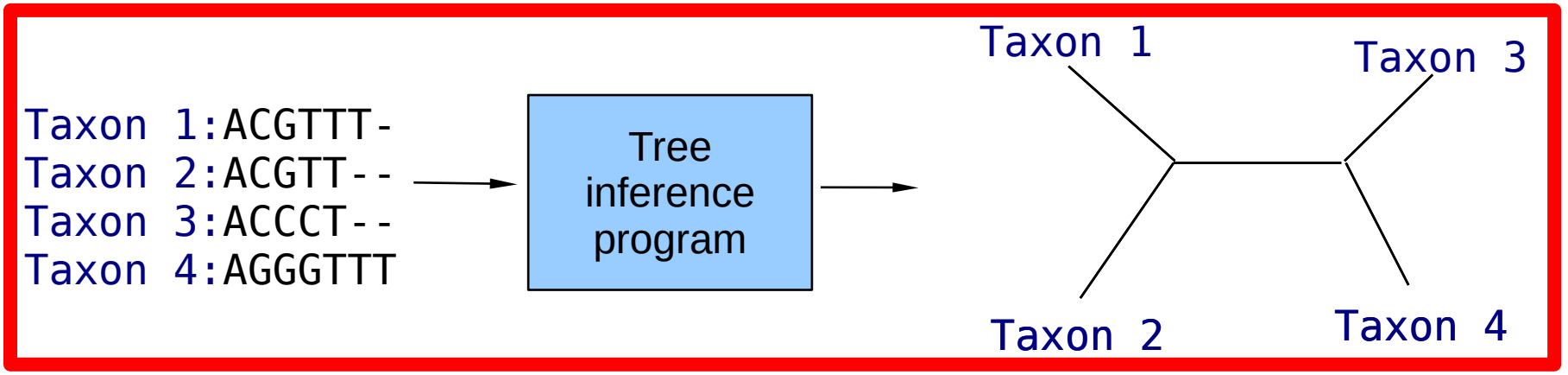
Mattis Bodynek, Lucía Martín-Fernández, Julia Haag, Ben Bettisworth, Alexandros Stamatakis

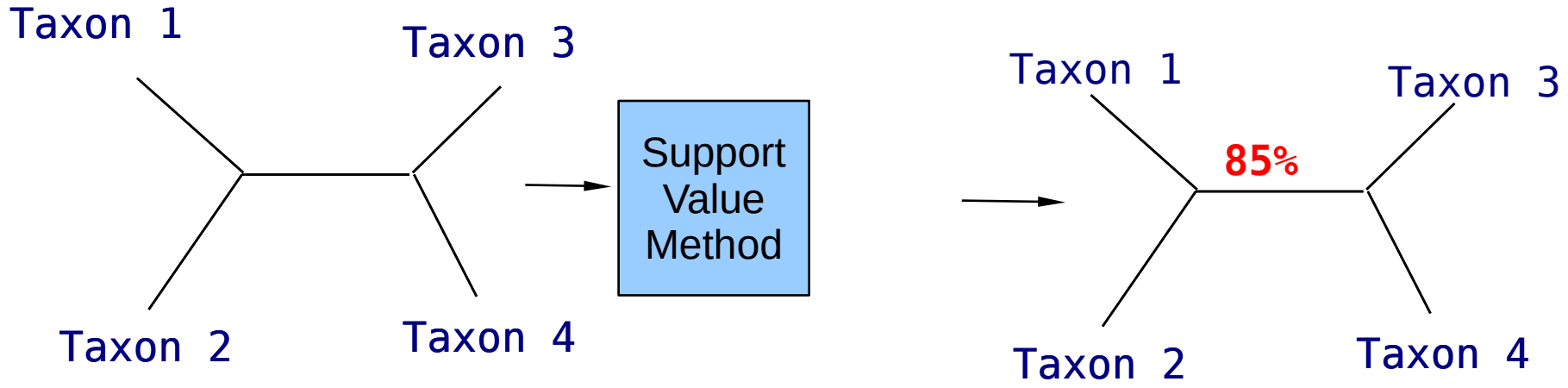
doi: https://doi.org/10.64898/2026.05.29.727837

This article is a preprint and has not been certified by peer review [what does this mean?].

Predicting Phylogenetic Difficulty







Pythia

JOURNAL ARTICLE

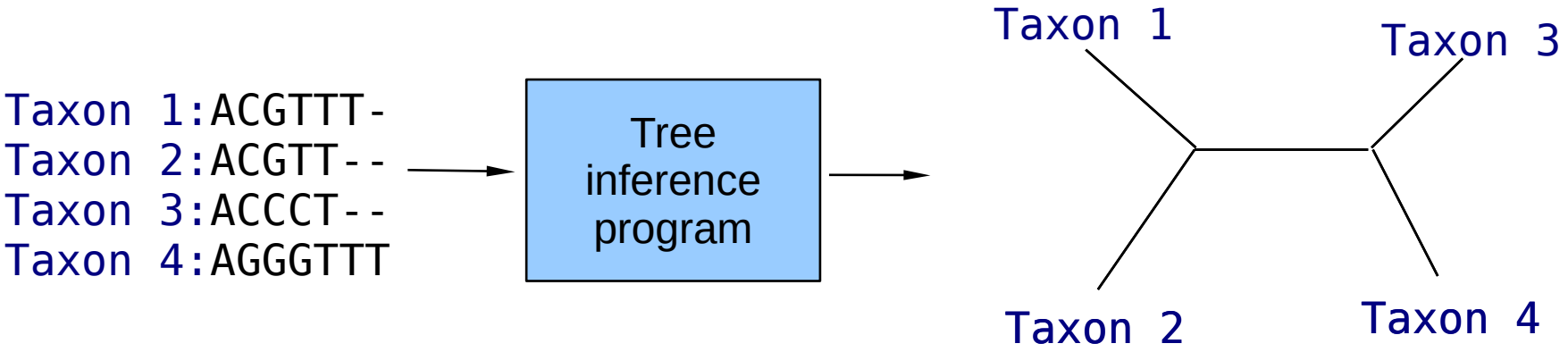
From Easy to Hopeless—Predicting the Difficulty of Phylogenetic Analyses

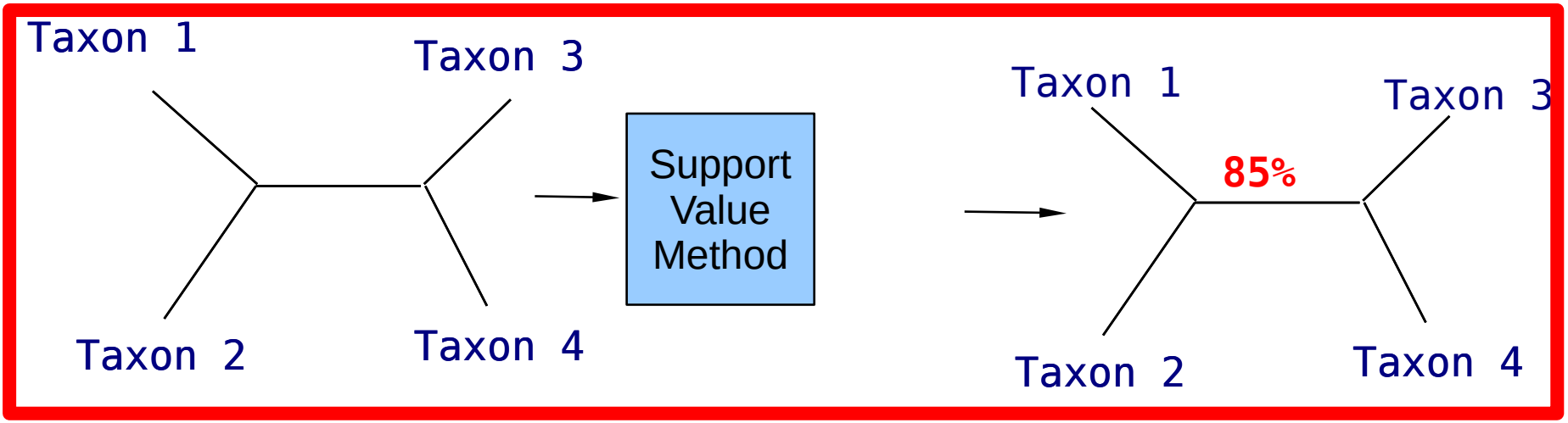
Julia Haag ✉, Dimitri Höhler, Ben Bettisworth, Alexandros Stamatakis

Molecular Biology and Evolution, Volume 39, Issue 12, December 2022, msac254,
https://doi.org/10.1093/molbev/msac254
Published: 17 November 2022

Predicting Phylogenetic Difficulty







EBG: Educated Bootstrap Guesser

JOURNAL ARTICLE

Predicting Phylogenetic Bootstrap Values via Machine Learning

Julius Wiegert ✉, Dimitri Höhler, Julia Haag, Alexandros Stamatakis Author Notes

Molecular Biology and Evolution, Volume 41, Issue 10, October 2024, msae215,
https://doi.org/10.1093/molbev/msae215
Published: 17 October 2024 **Article history** ▾

Impact of Using Phylogenetic Difficulty & Support Prediction

- `RAxML-NG v2.0` → I*n particular, the new fast search heuristic in conjunction with machine learning based branch support prediction induces a* ***65× inference time reduction*** *compared to RAxML-NG 1.2, with minor to no accuracy loss.*


New Results

Follow this preprint

RAxML-NG 2: Automatic model selection, novel tree search heuristics, and fast branch support metrics

Oleksiy M. Kozlov, Anastasis Togkousidis, Christoph Stelz, Dimitri Höhler, Julius Wiegert, Alexandros Stamatakis

doi: https://doi.org/10.64898/2026.09.09.750097

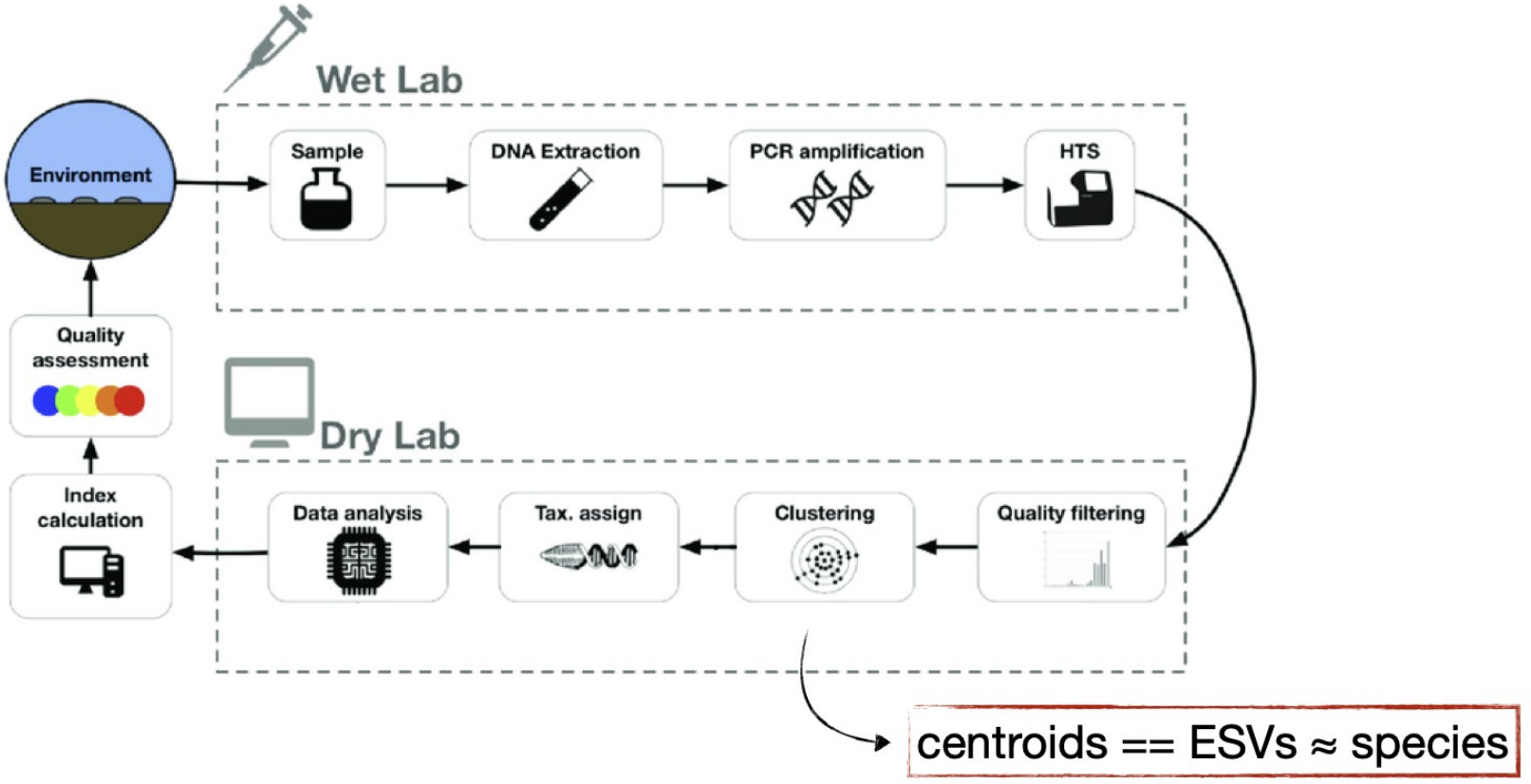
This article is a preprint and has not been certified by peer review [what does this mean?].

Outline

- Alexis: Predicting, Propagating & Using Uncertainty
- **Giorgos: Making Sense of Metabarcoding Data**
- Noah: Gene Tree – Species Tree reconciliation

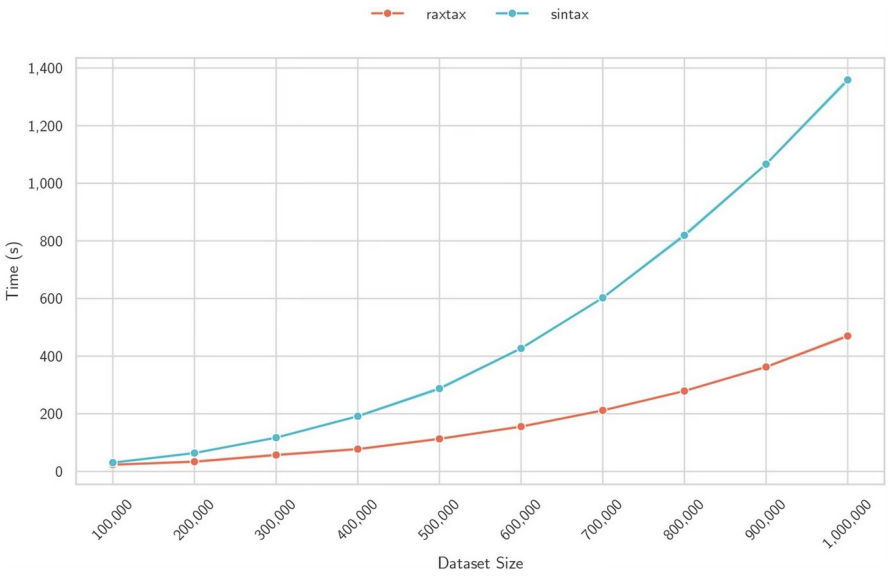
Metabarcoding

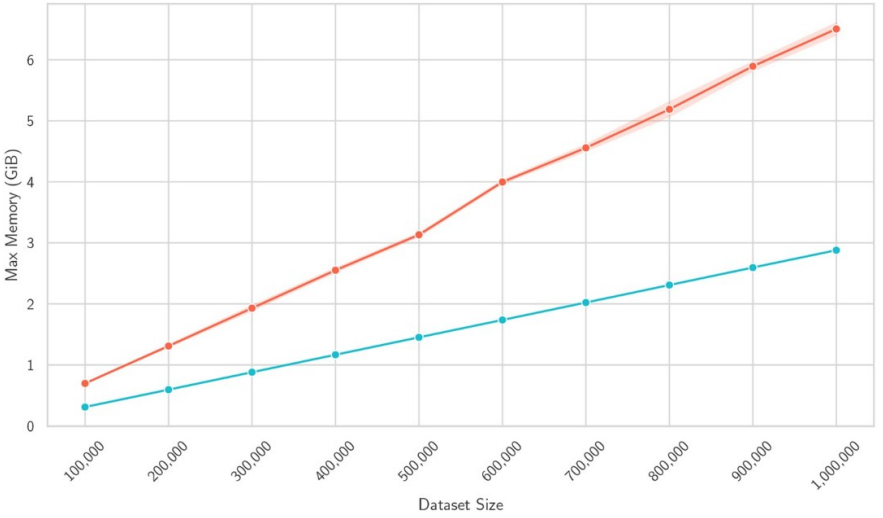
Workflow



Taxonomic Classification

- Rapid growth of reference databases requires computationally efficient tools.
- RAxTAx uses a k-mer-based non-Bayesian approach to provide fast and accurate classifications.
- RAxTAx converts k-mer matching into probabilities and aggregates them into taxonomic clade confidences.
- The output contains also uncertainty metrics which are database specific.



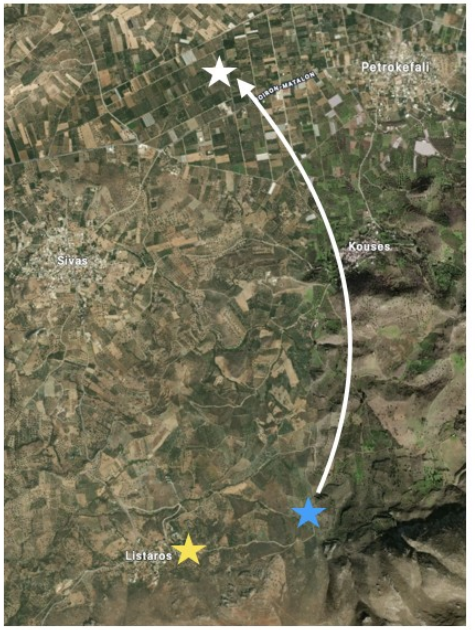


Listaros Insect Biodiversity

Pilot study - How many insect species are in Listaros?

- 3 traps near Listaros village
- 2.5-year period 04/2021-09/2023
- 29 samples
- 2 wet labs
- 1 dry lab
- 300M+ barcodes
- 10K+ ESVs

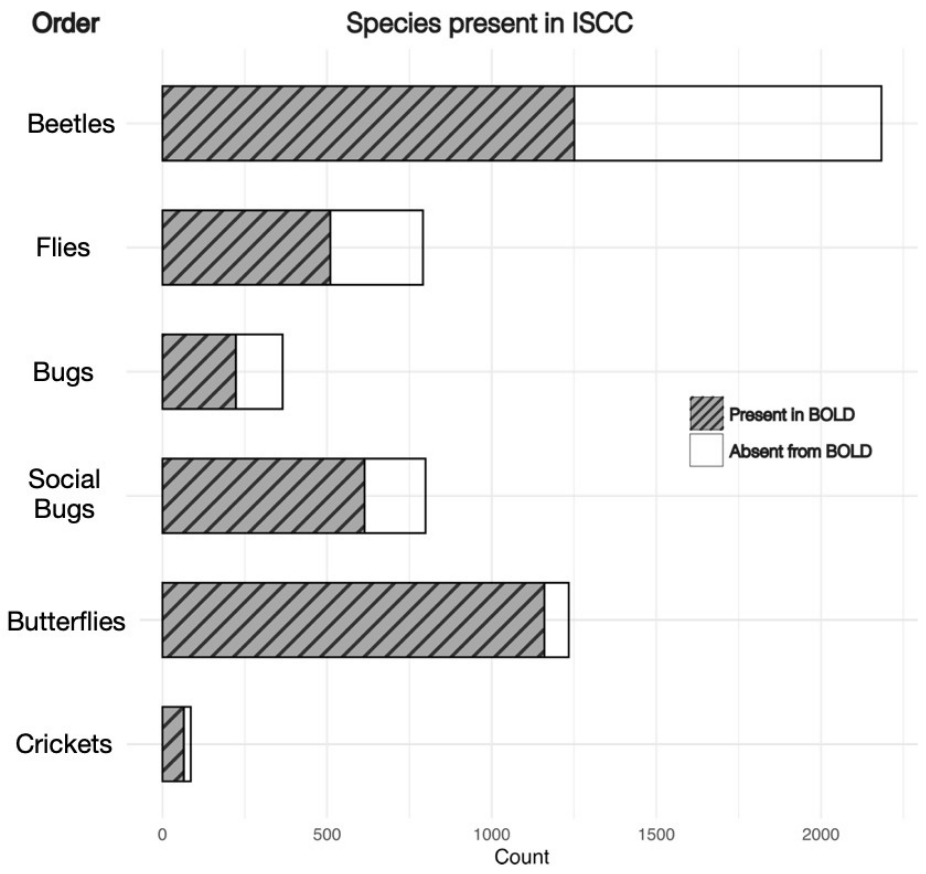






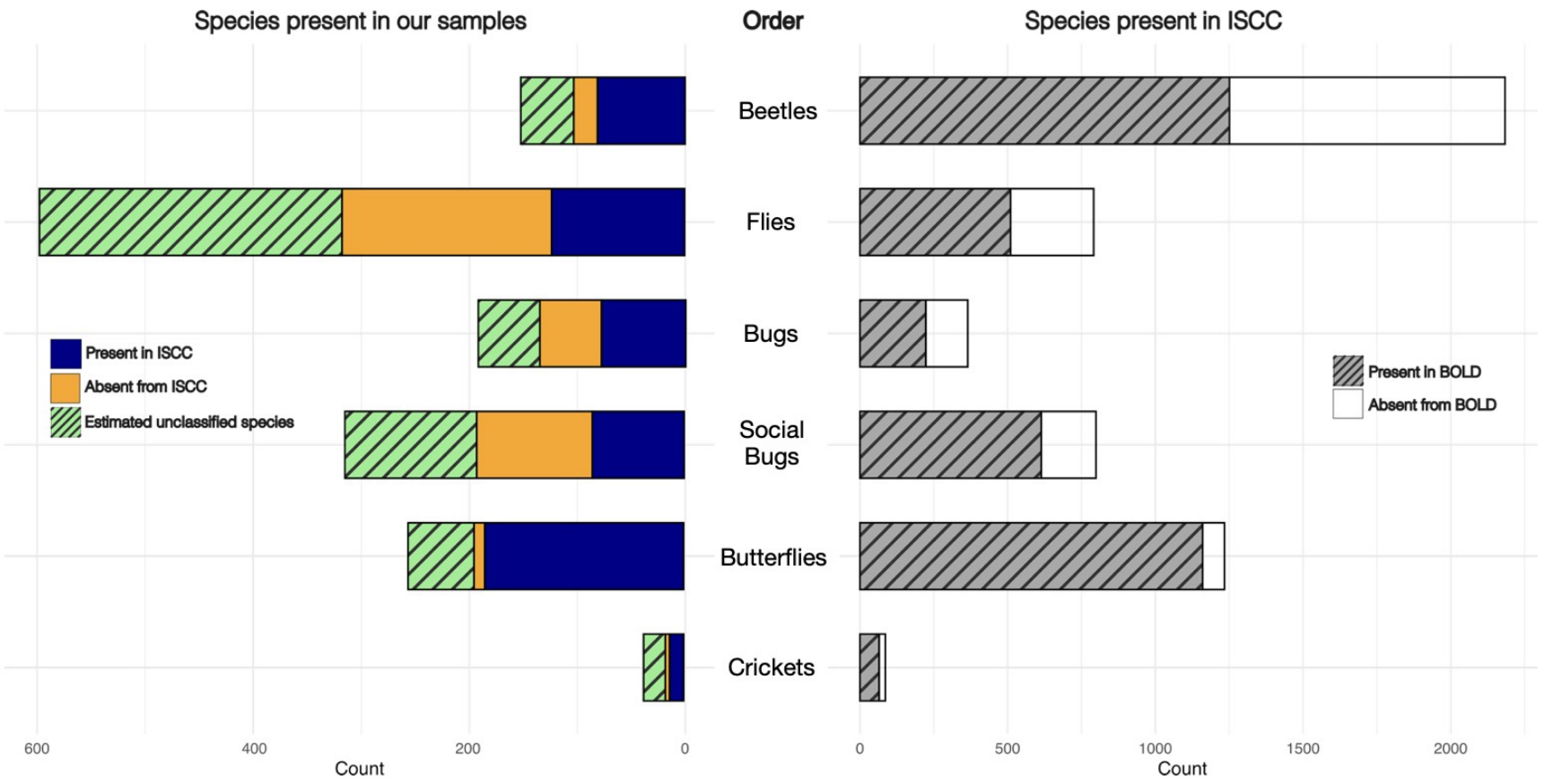
Listaros Insect Biodiversity

Sampled diversity only partly overlaps known records



Listaros Insect Biodiversity

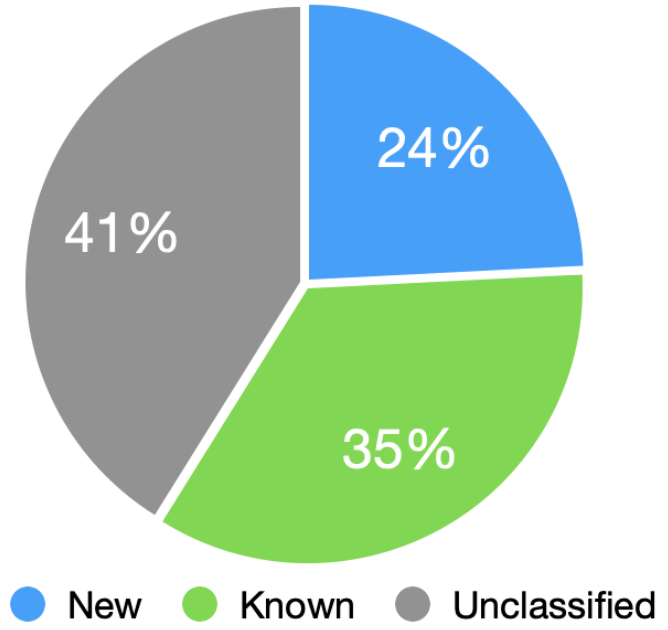
Sampled diversity only partly overlaps known records



Listaros Insect Biodiversity

Pilot study results

- Estimation of 1,678 insect species present in the samples from 14 insect orders
- 582 were already known from Crete
- 406 new observations
- 690 unclassified species



JOURNAL ARTICLE

raxtax: a *k*-mer-based non-Bayesian taxonomic classifier

Noah A Wahl ✉, Georgios Koutsovoulos, Ben Bettisworth, Alexandros Stamatakis

Bioinformatics, Volume 41, Issue 12, December 2025, btaf620,
https://doi.org/10.1093/bioinformatics/btaf620
Published: 19 November 2025 **Article history** ▾

MBMG
Metabarcoding & Metagenomics

Metabarcoding and Metagenomics 10: 1117–1136 (2026)
DOI: 10.3897/mbmg.10.202607

Research Article

Charting the insect biodiversity of southern-central Crete: Insights from a pilot metabarcoding survey

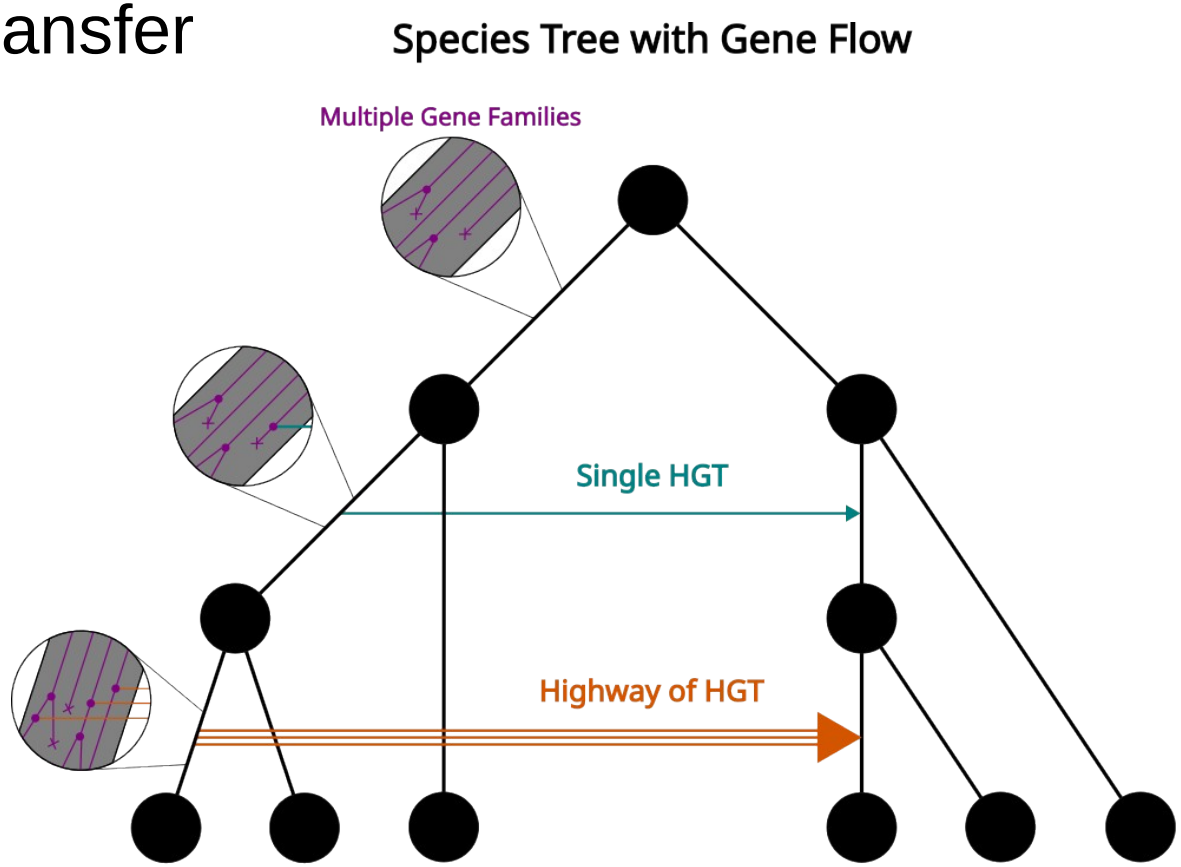
Georgios D. Koutsovoulos[1], Martin Sorg[2], Thomas Hörren[2,3], Dominik Buchner[3], Sarah J. Bourlat[4], Kathrin Langen[4], Apostolos Trichas[5], Florian Leese[3], Alexandros Stamatakis[1,6,7]

Outline

- Alexis: Predicting, Propagating & Using Uncertainty
- Giorgos: Making Sense of Metabarcoding Data
- **Noah: Gene Tree – Species Tree reconciliation**

Highways of Horizontal Gene Transfer

- Gene evolution differs from species evolution via **Duplication**, **Transfer** and **Loss** events
- **Endosymbiotic** events lead to large gene flow events beyond background transfer

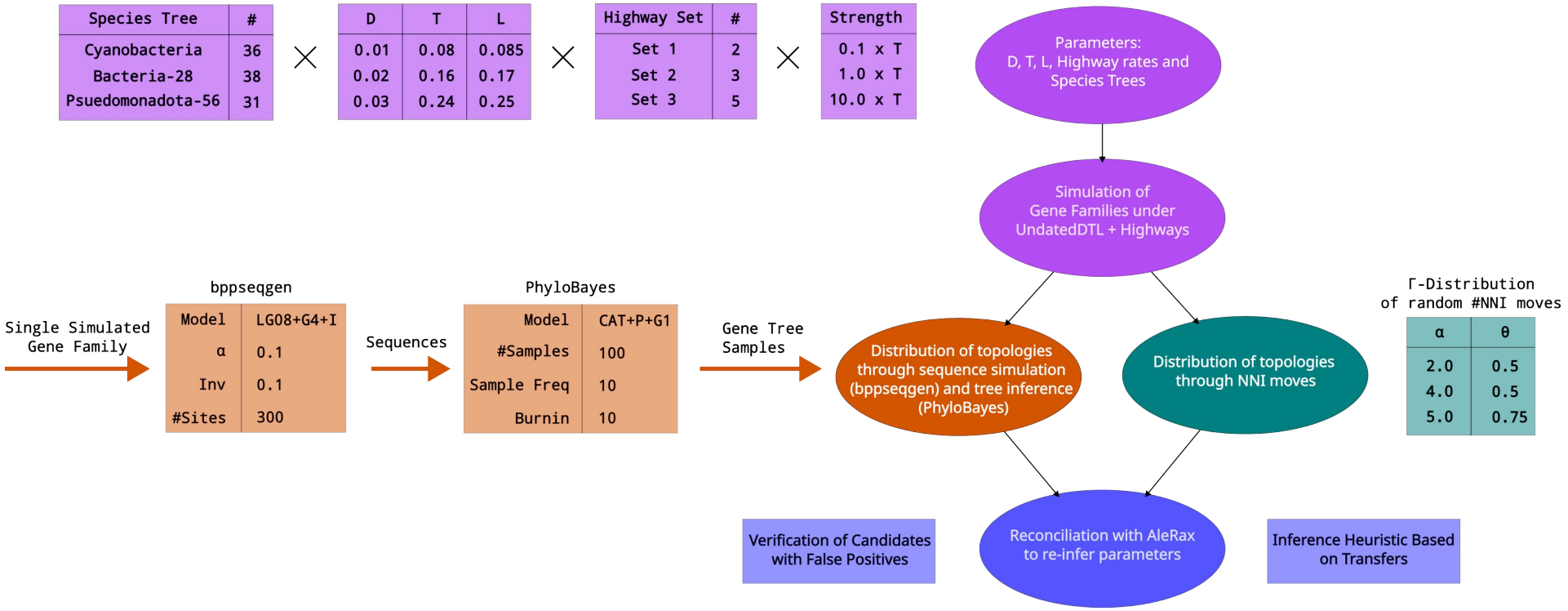


*HGT = Horizontal Gene Tranfers

Highways of Horizontal Gene Transfer

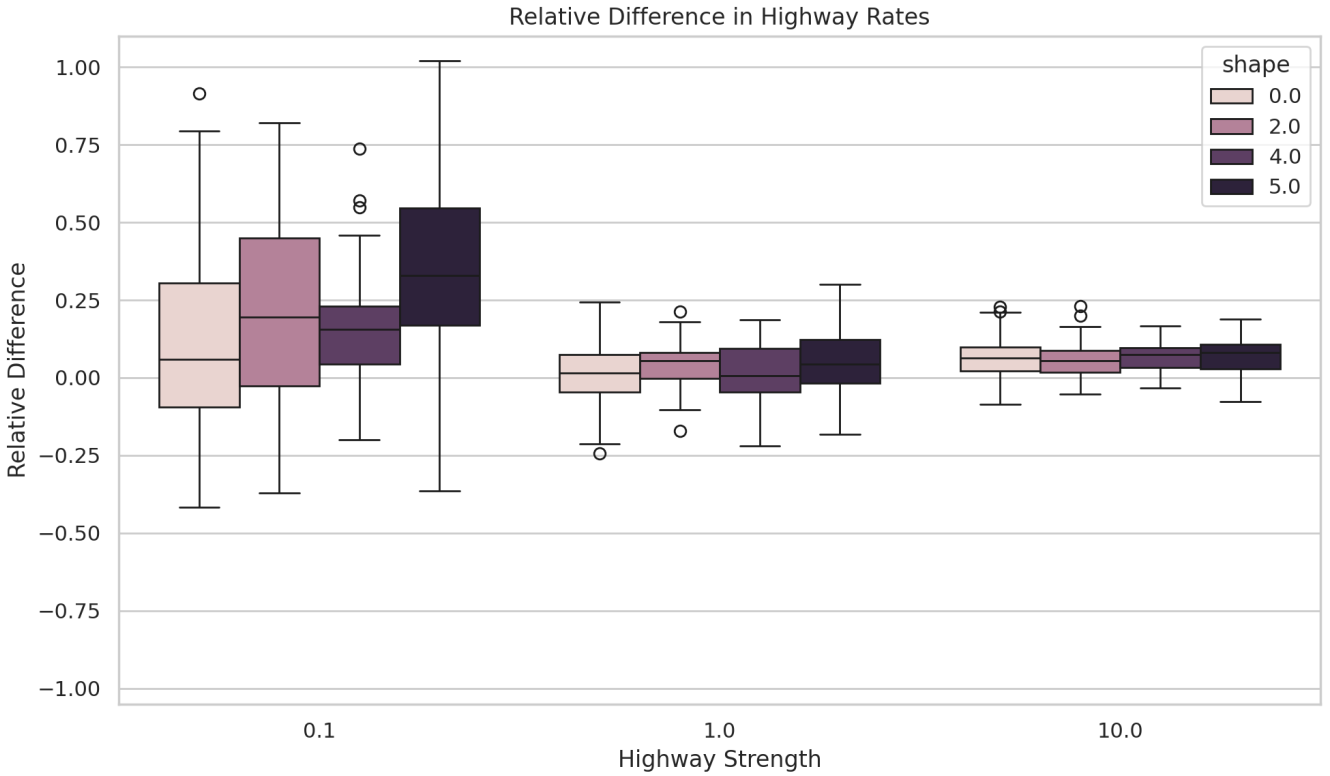
- Current likelihood models do not effienciently account for this variance in transfer rates for specific species pairs
- Highways as an extension of the UndatedDTL model to capture this gene flow
- We use a **heuristic** approach to identify highways with a strong signal and then validate them
- 2 papers in work (technical and applied)

Simulated Highways



- Highways can be inferred from scratch with a heuristic
- Or candidates can be **validated**

Simulated Highways



Empirical Highways

- Project with collaborators across the globe
- Species tree with **1086 species** and **3614 gene families** covering all 3 domains of life
- Validating **existing hypotheses** for highways with the new likelihood model
- Shows promising results for better understanding the **origins of complex life**

Thank You !

