Quantifying Uncertainty in Evolutionary Analyses

Alexandros Stamatakis

ERA Chair, Institute of Computer Science, Foundation for Research and Technology - Hellas
Research Group Leader, Heidelberg Institute for Theoretical Studies
Full Professor, Dept. of Informatics, Karlsruhe Institute of Technology

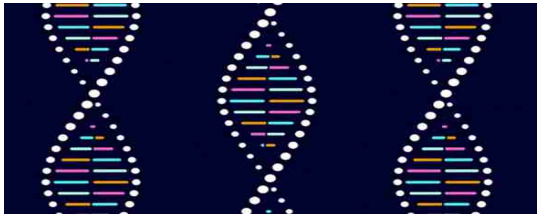
www.biocomp.gr (Crete lab)

www.exelixis-lab.org (Heidelberg lab)

Outline

- **Our Approach to Bioinformatics**
- Introduction to Phylogenetic Inference
- Sources of Uncertainty
- Phylogenetic Difficulty
- Other stuff we are working on

Bioinformatics



How to Write Your First Scientific Paper

Simone Rosa Nunes Reis
Instituto de Informática
UFRGS
Porto Alegre, Brazil
srnunes@inf.ufrgs.br

André Chaves Reis
Instituto de Informática
UFRGS
Porto Alegre, Brazil
andrecr@inf.ufrgs.br

Abstract—this paper presents a method to instruct students on how to write their first scientific paper. The method adopts a Problem Based Learning approach and uses a mixture of lectures and writing. A second issue tackled the choice of the paper subject. Hypothesis formulation and the design of experiments to test it, are the main subjects of the lectures, while the writing is the subject of the first lecture. The method has been successfully used to teach graduate and undergraduate students to write their first articles.

Keywords—writing, paper writing, scientific method

I. Introduction

Engineers normally do not receive formal training on how to write a scientific paper. Nevertheless, it is a fact of life that the second chapter of Carl Sagan's book *Cosmos* (*From Shores to Engineering School* [1]) is about writing, a so time that communication skills, especially written communication skills, may affect the professional success of engineers.

Scientific paper writing is more complex than just writing. This complexity arises from the fact that paper writing involves reporting scientific advances. The process involves several aspects, such as formulating a hypothesis, proposing an experiment to test the hypothesis, performing the experiment and analyzing the data to judge the hypothesis, and finally writing a report about the whole process. The final report has specific structure and is written in a specific language.

Due to the multiple aspects involved in scientific paper writing, statistical discourse is made life difficult normally focus more on a single aspect, or they are too extensive. This paper proposes a short introduction to paper writing, corresponding a lecture, that covers a restricted topics and skills, yet in a concise and practical way.

This paper is organized as follows. Section II presents describes the concept of new knowledge and how it is produced. Advice for conducting the process of investigation and writing a paper is presented in Section III. The choice of a topic is discussed in Section IV, while Section V presents advice on dealt with scientific communities. Finally, conclusions are presented in section VI.

II. Understanding and Producing New Knowledge

Writing a scientific paper is a process that need some steps for ensuring the contents of the paper. Understanding the process helps to produce better papers. This section presents insights on how new knowledge is produced.

A. Starting with New Knowledge

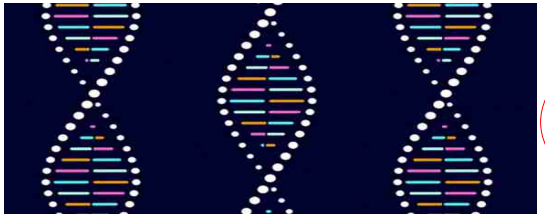
Scientists are in the search for new knowledge, as a paper is supposed to advance the scientific knowledge. Plus minus from imaginative phase on unstructured ideas to scientific discovery. The process of discovery is a process of creation of knowledge, which results in being able to accounts ideas that normally would not be accessed. This is difficult to admit for people starting in science, as normally students are educated to make a mess that they rarely a bread chapter and think they are ready to write a paper. This is a misconception, as it is a fact of life that normally with a set of exercises that most of the students understand what they taught and if so able to use the knowledge to solve the exercises, they are considered to have knowledge. The kind of thinking is called convergent thinking [2] Fig. 1 illustrates the concept that new knowledge has to be produced, and it is not just a matter of learning what is known by some of their elders[1, 2, 3, 4]. Science has to be more dynamic cases, a person imagines something that is new for the person, but it is not new to humankind. Not to the rest of new ideas yet[1, 2, 3, 4].

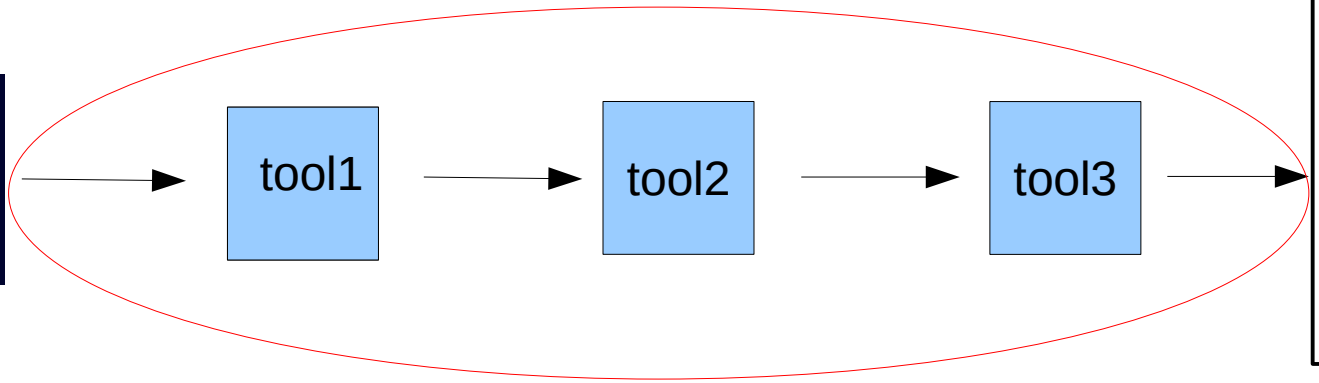
Figure 1. New knowledge is not an imaginary idea.

B. Knowing the Frontiers of Knowledge

Imagination, of divergent thinking, is an important part of the process of creation. It is a fact of life that most of the ideas that decide if a product of one imagination is new or not are in the field of knowledge it is necessary to know the field. For this, it is necessary to read papers. Reading papers helps the author to: (1) understand which is new in each paper, (2) to establish what the contribution of the paper is, (3) to be able to compare

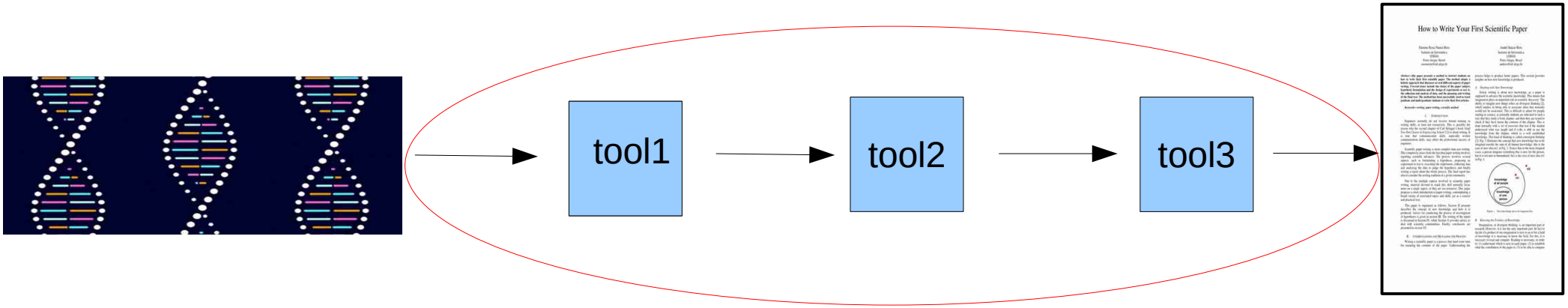
Bioinformatics



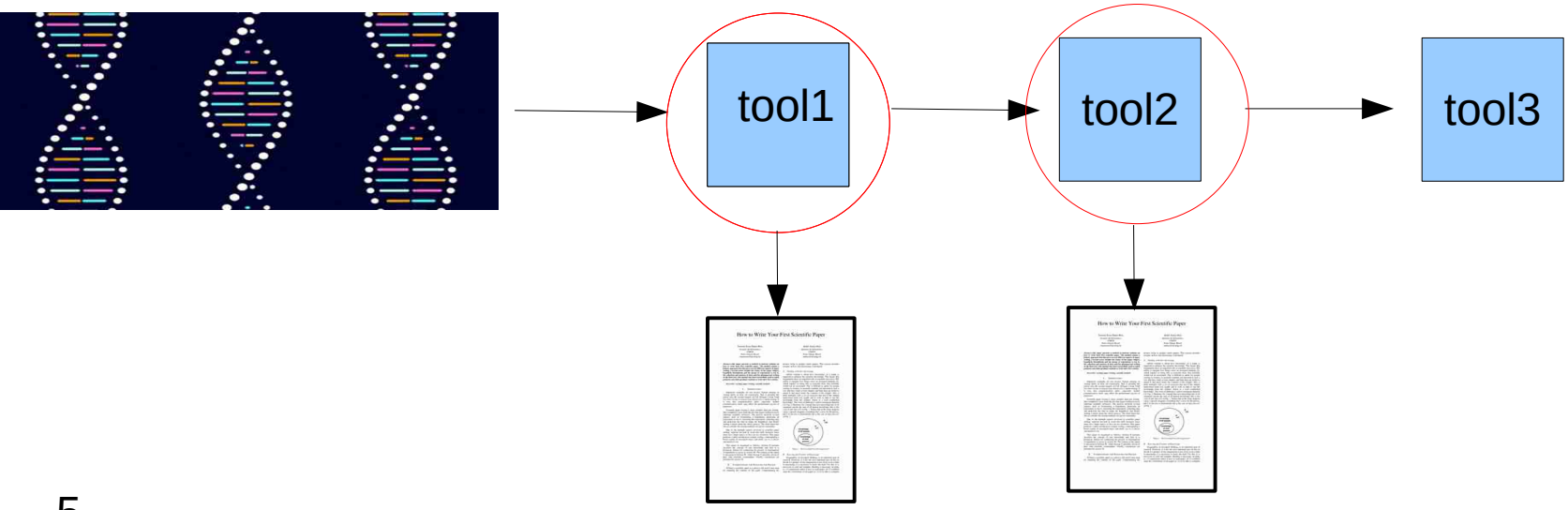


Data-centric: pipeline building

Bioinformatics



Data-centric: pipeline building



Method-centric: tool building

Our Approach

- Focus on *core* tool, model, algorithm, and method development
- Method development better fits the research interests of a computer scientist
- **Goal:** Enable Research in Evolutionary Biology

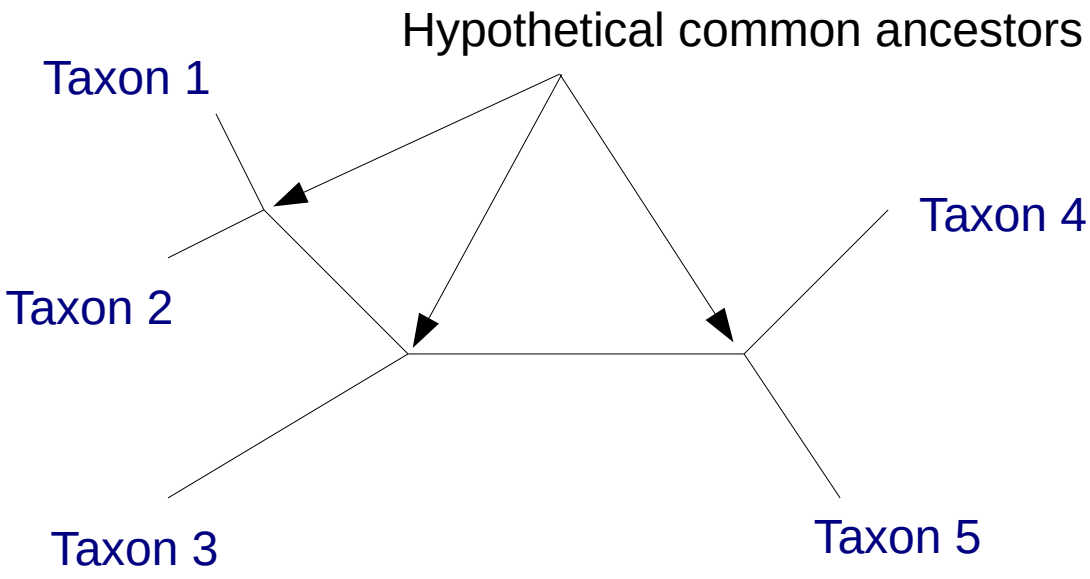
Our Approach

- Focus on *core* tool, model, algorithm, and method development
- Method development better fits the research interests of a computer scientist
- **Goal:** Enable Research in Evolutionary Biology
- Nonetheless, we often conduct data centric research in side projects

Outline

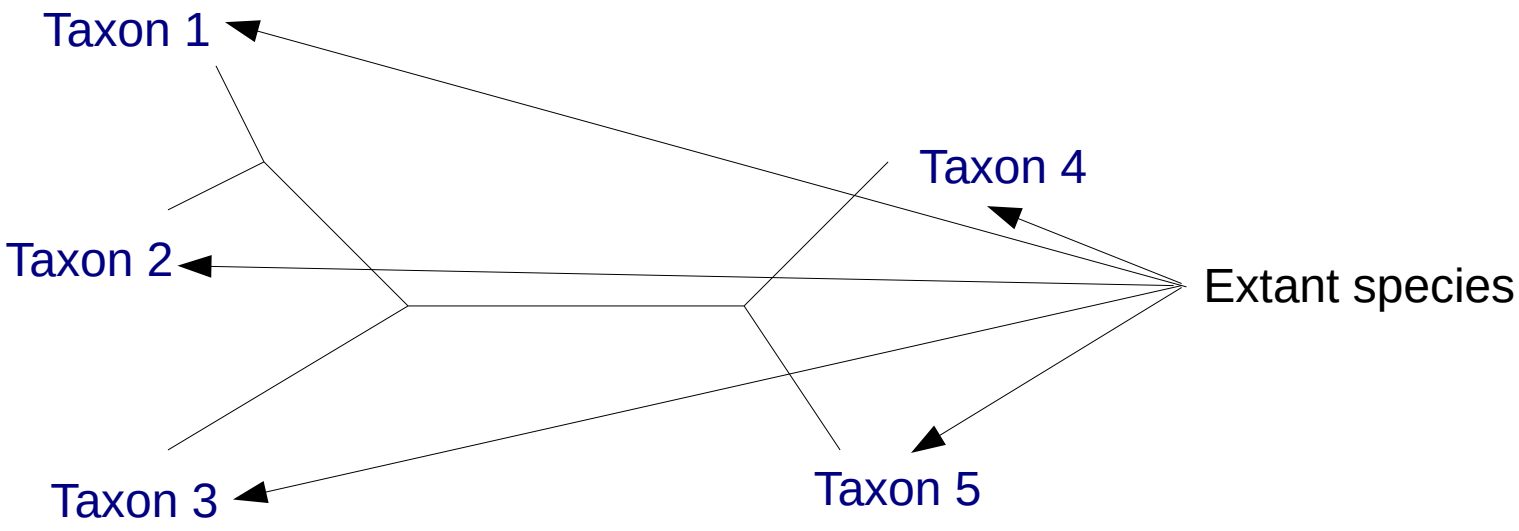
- Our Approach to Bioinformatics
- **Introduction to Phylogenetic Inference**
- Sources of Uncertainty
- Phylogenetic Difficulty
- Other stuff we are working on

A phylogeny

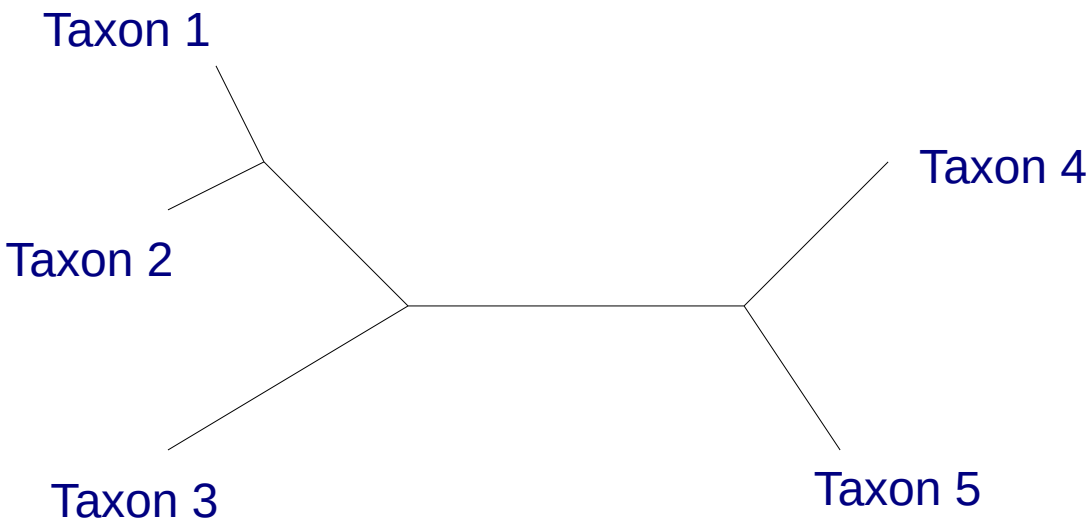


Phylogenies describe evolutionary relationships among **species**

A phylogeny

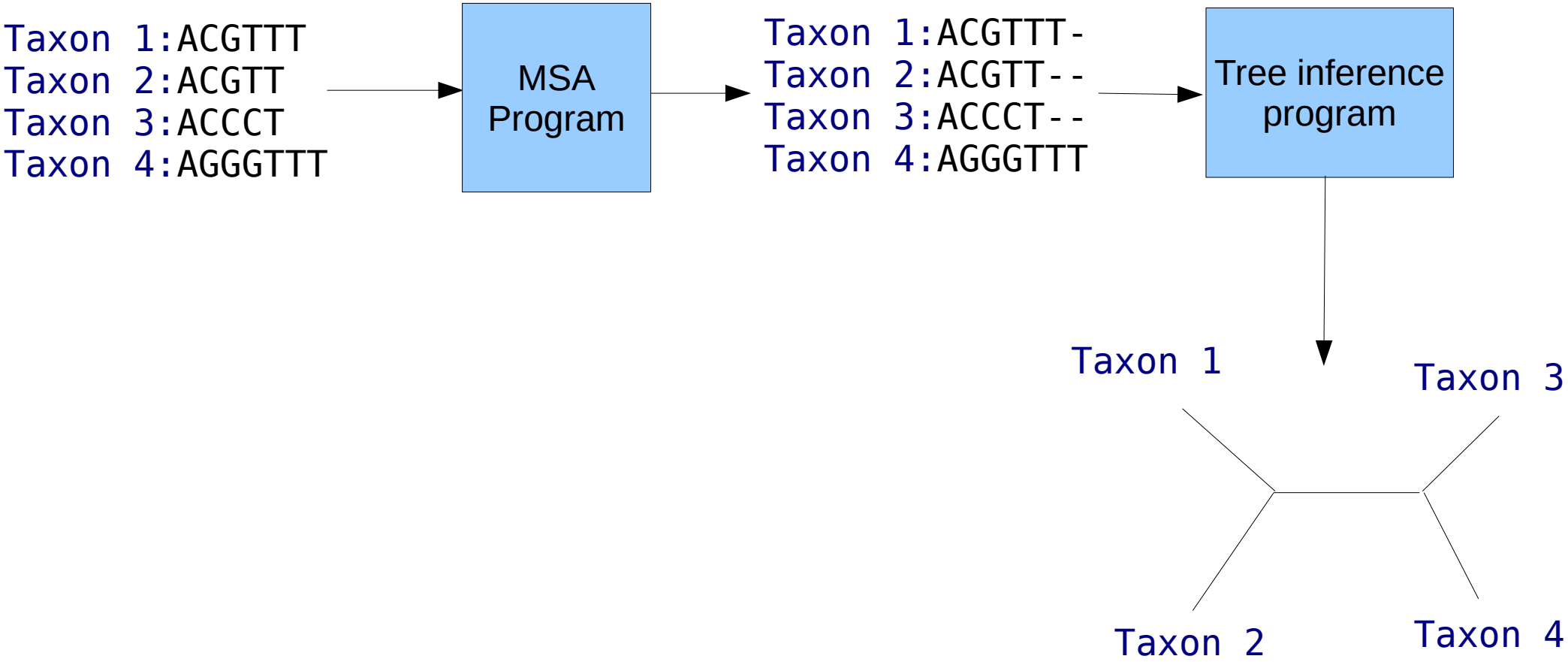


A phylogeny

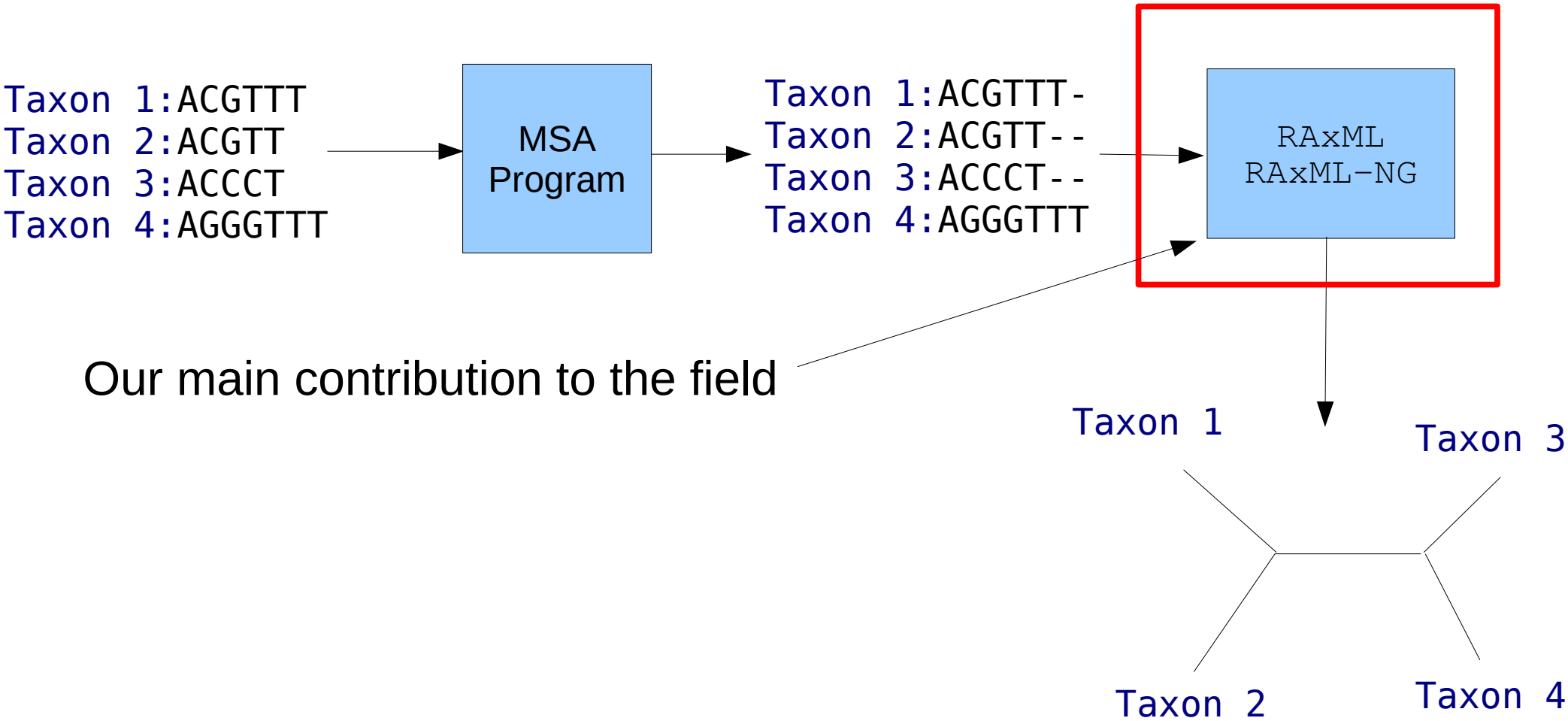


Phylogenetic trees are unrooted binary trees!

Tree Inference Pipeline

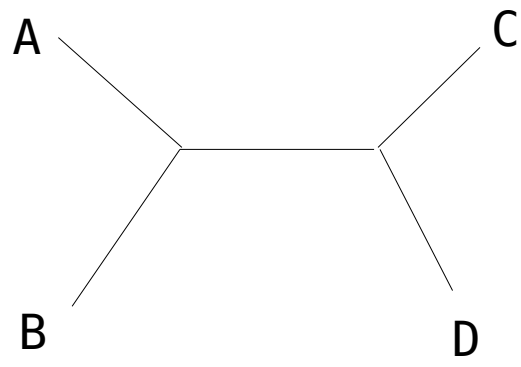


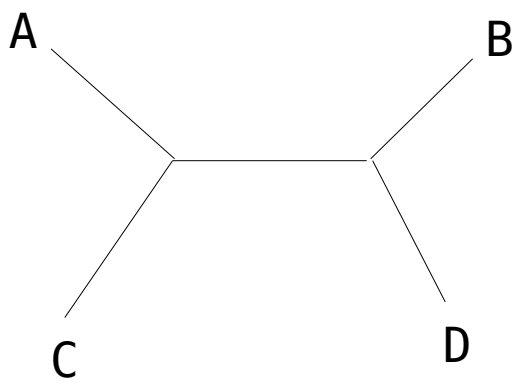
Tree Inference Pipeline

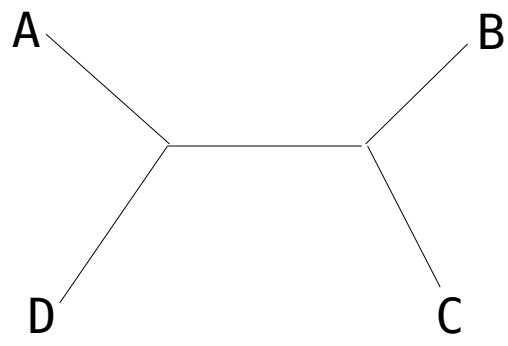


How many unrooted 4-taxon trees exist?

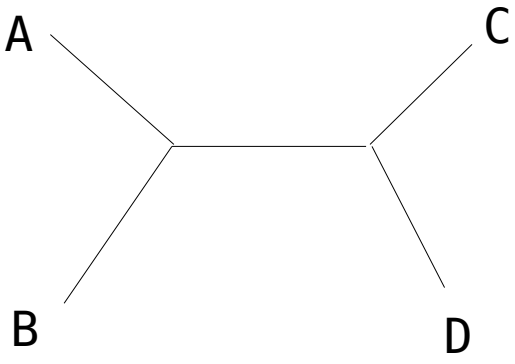
How many unrooted 4-taxon trees exist?

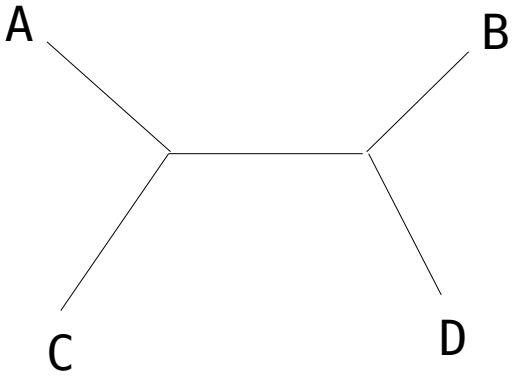


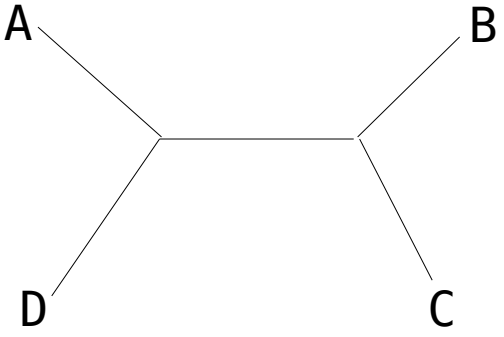




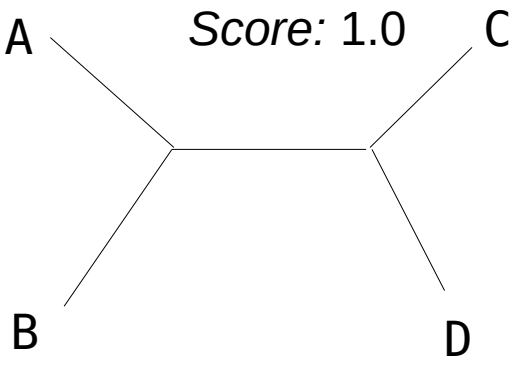
How do we chose among them?

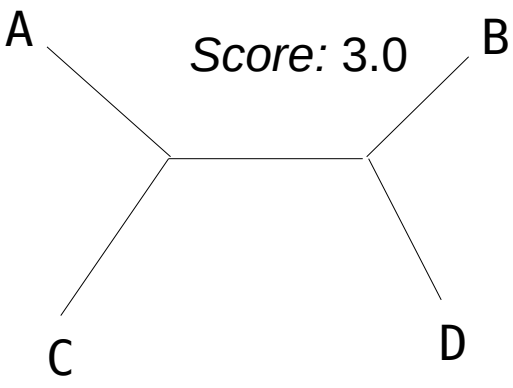


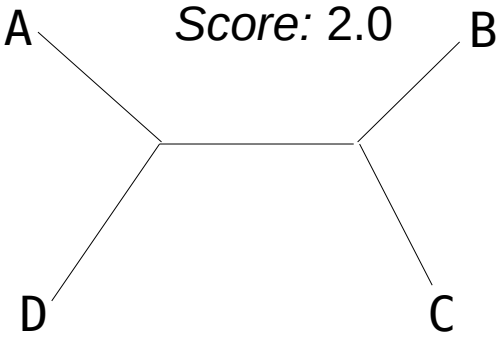




How do we chose among them?

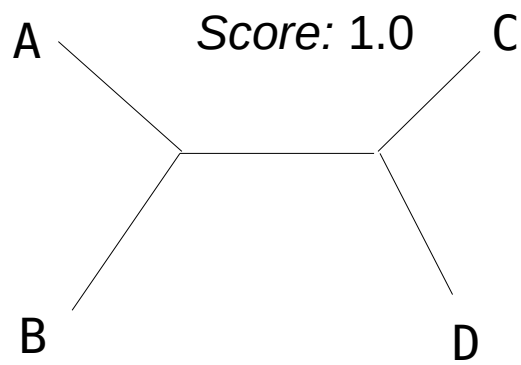


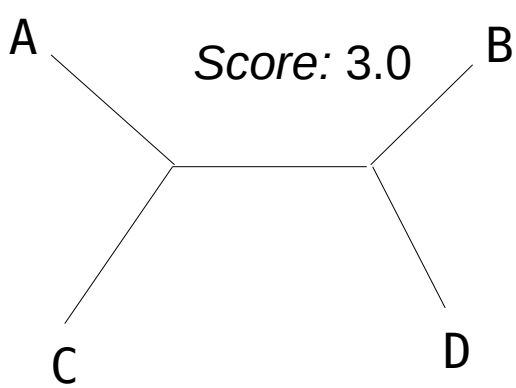


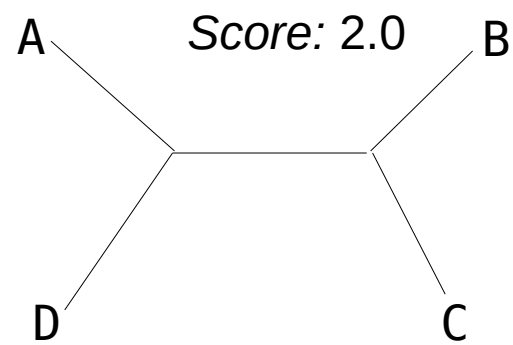


We need scoring criteria!

How do we chose among them?

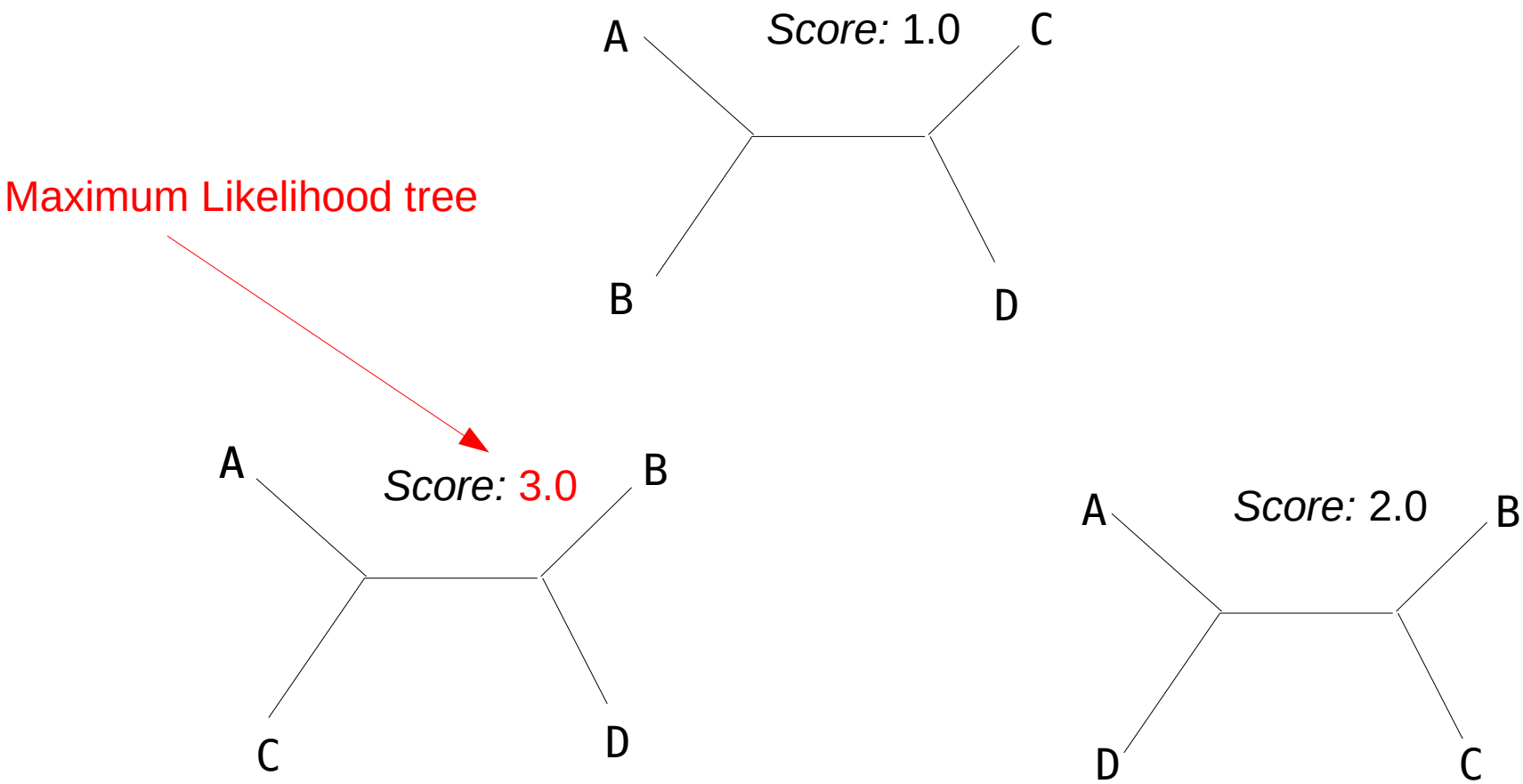






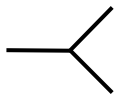
We need scoring criteria!
The currently most widely used criterion is (maximum) likelihood →
How likely is it that the tree, given a model of evolution, generated the observed data?

How do we chose among them?



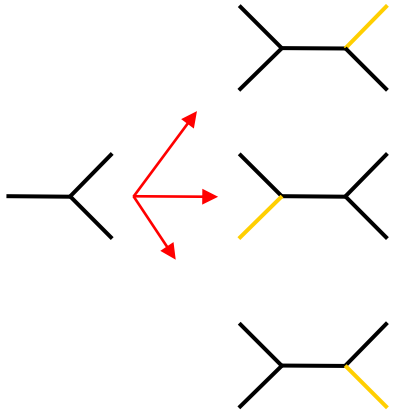
We need scoring criteria!
The currently most widely used criterion is (maximum) likelihood →
How likely is it that the tree, given a model of evolution, generated the observed data?

The number of trees



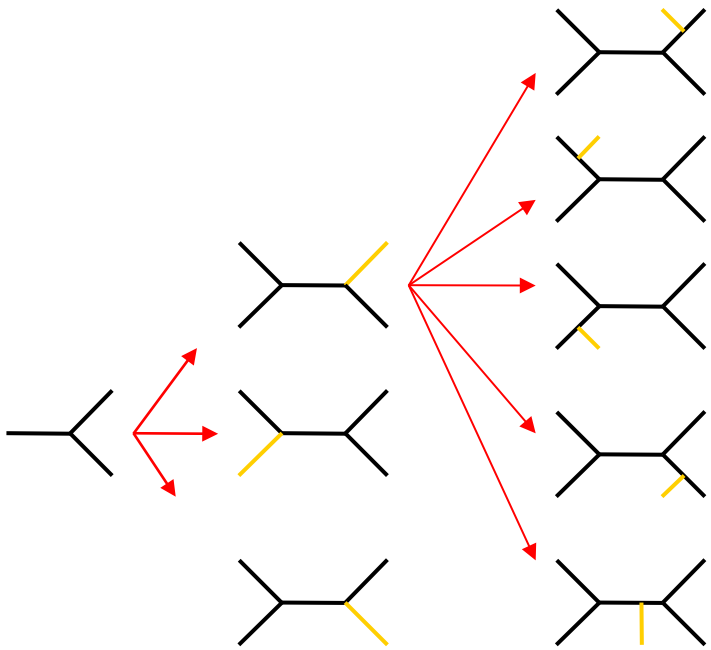
3 taxa → *1 tree*

The number of trees



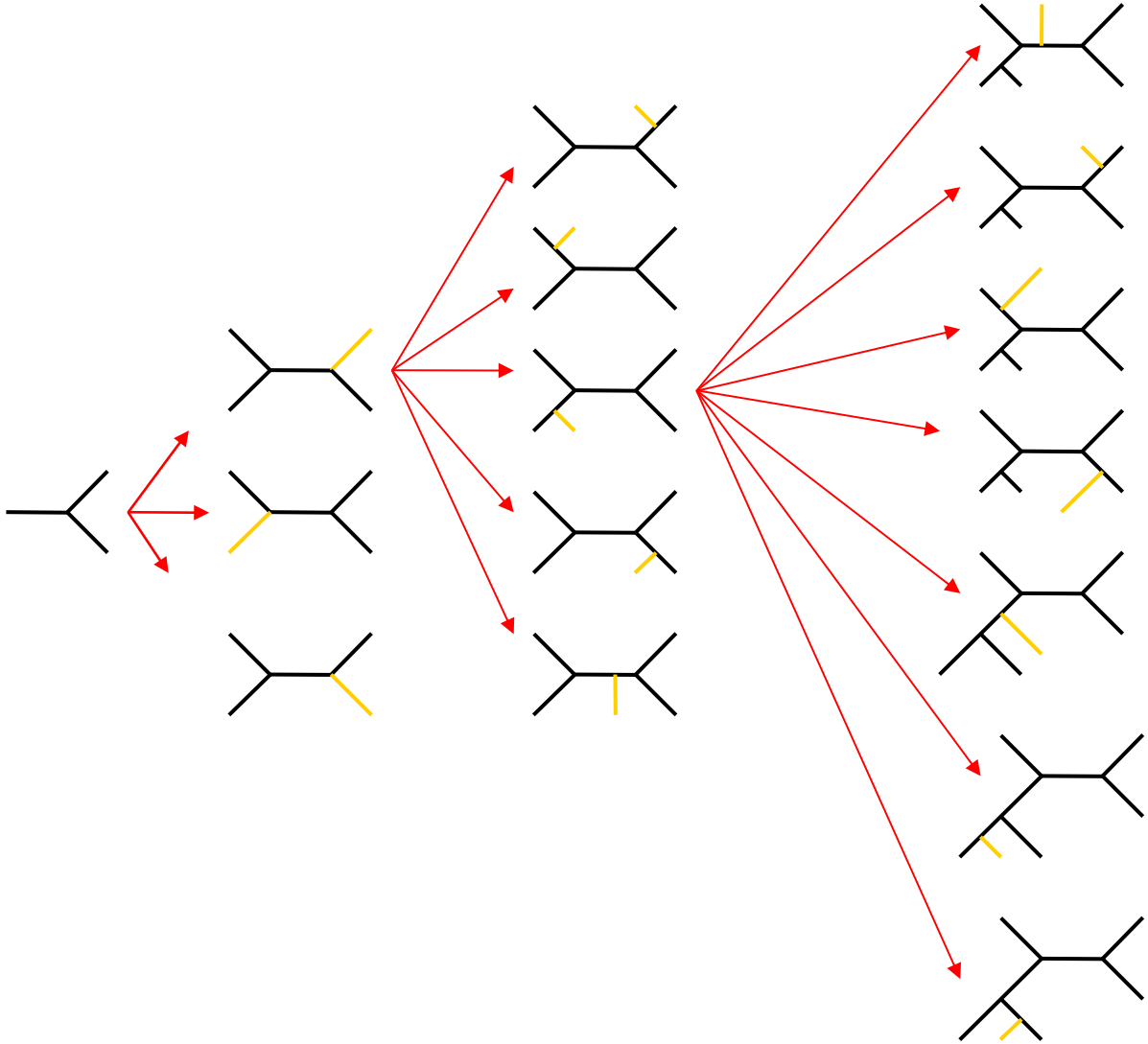
4 taxa → *3 trees*

The number of trees



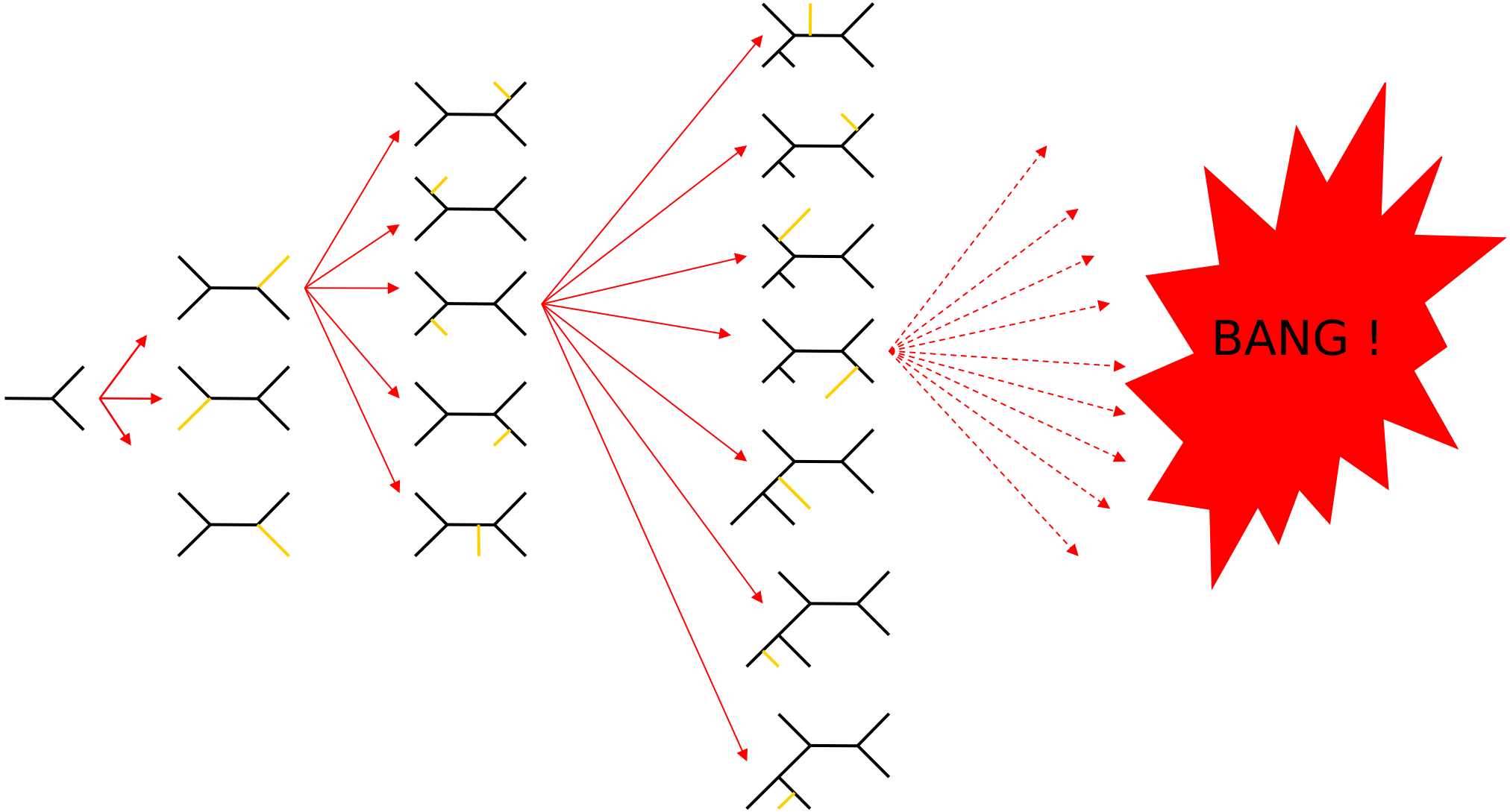
5 taxa → *15 trees*

The number of trees



6 taxa → *105 trees*

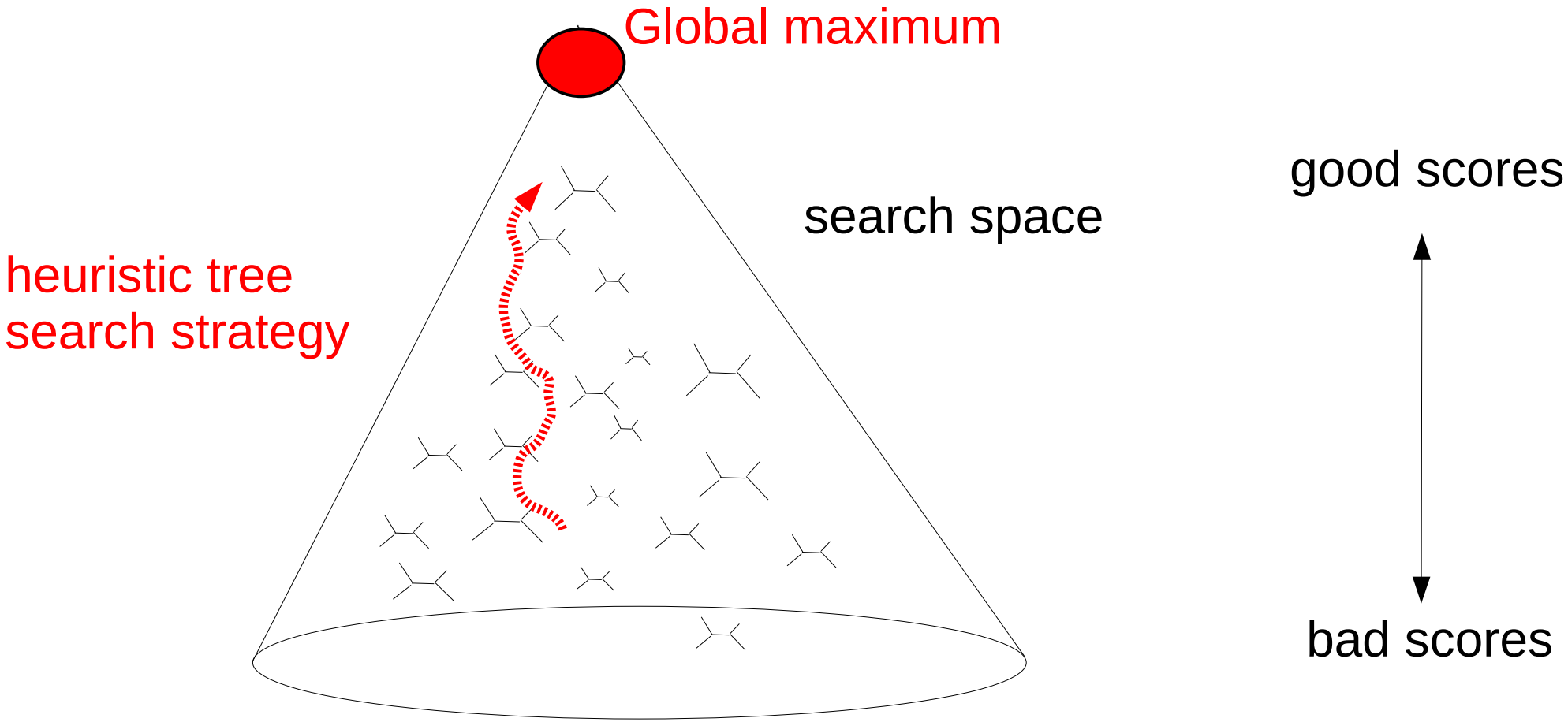
The number of trees explodes!



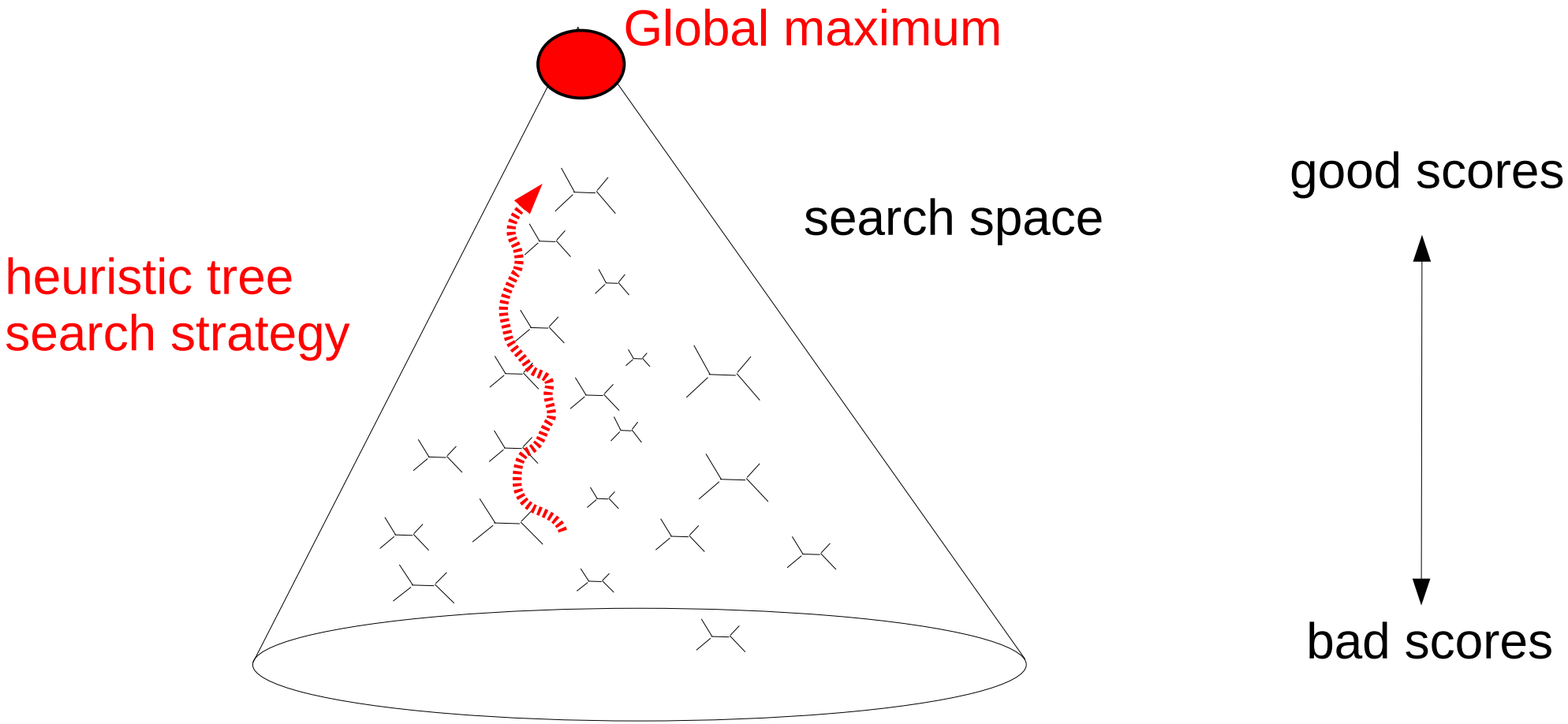
possible trees with 2000 taxa

```
stamatak@exelixis:~/Desktop/GIT/TreeCounter$ ./treeCounter -n 2000


GNU GPL tree number calculator released June 2011 by Alexandros Stamatakis

Number of unrooted binary trees for 2000 taxa: 30049638174211656151632910065681814981377232074237013089504954043012636525258308210827685996688247000464352735214265634288295
89150234460006314939691306329704360561848618774654822779912235368092334555631999108345976931267565250128998674331877528114019609916315223670306091217357097623798477054676677
77953247971826143852733382267277842507372528499166696875844035105795870206865058176870446663181237429010214385064324713609344916670211359697569403006662526464792691245510314
94236619554282411827762511484875825458122791428980113264890267403376129471274576703626757908684316966071860984794181886595721455704474457228866172905358352074425368812312401
06613156948861960941195646736200342575241335277575085829161096422575727699767991408283343210161327401652830993803904592327690690035972919709940739349563486203899010742687
28229759746553771022576726768428580118772249501062181173405232082653973429622273525365905158656313832720311198419874675997386463182903203832523085979979922161012272157808052
48145831206844016760623930600971161672971550472848779963433753134899423037243734787913198908595376407013484944611387757257695240870246172010787429738046227505254570668937
2
31941820644070689188400387059028977219751645449597582166213062050646177610994856637341681835849893290769933820678010524372846149240342296115518260977822861919267207129518
95
89360099591309742330723163825184281103305710174411568843051318658775443763085003114511107238370397074651822320404061547082730786299575493310312752086167006607912980142622
30
05651235227180638195093358726517286235890205200161443617560756542864714221266130044348070840675015892476731663415395405750744749949098314964730310804114018918497359128112
28
37877404988483405621024205664244638600938996508574296194726905430152812375265109658152846997970367921711290355680981807916958795161415928104952817985584729253444786442443
59
98085315372047968149694659917686145337010519859285771574824559433772423695825762426630169463204824951822559392874031776234338810486046309751915569238711675130952134150988
16
71546430786235260623778640683868042469025274911393192768026115159905826038867331729307136739034036186374639806057648364746702744467278808853370742544219227266777470033294
03
32010382880351126890262551830967919483586789293701637681753048206338943871497931152353698229625111163071482945992116208033026847620133356904410896681454361509051558775811
67
97700125639121511162374441704973717046040294811041148222864661319188219975713833683520725260552027698239746132184952492648970507903983602562556062898522888395613578741565
76
64889992608732866126306425432602489792291135600716405739845163752452433769437558573847255455643975996042559146401122211447552355731762399730577471839565312174165322959866
75
90129411612392407220932503696731248844915537592106506560154167207741592362408686676753482865129648887390597075788024733934634708481590116397727977474804173162687009167287
35
61216422684681606831989598012603764856153127816116895872151231233087600634733810972531184233396403909373783950668355787353078863586464005632994994906311874240290927792726
93
30032244537759579722487345689151145855707838505416816676674258113019580636219075007902950310882090972717481364369894739710799327777006763017306175665387397260377771730084
41
34394051236690555449324861650825399577950363267049478442934988531727973481777971465671751511788763964340693324580763461107342143281950499096808740273976889147045174720555
43
89693966687426014772418946932129024533173341882867731946535441133021008665750817132403426475804892186236634616079559437205163951540969498064862423097969472111696659615800
41
78839922322646284989423522769263911243607675089976717896834593378907423463545571935306656153799002779162656363618619740485909382340622354597697331372136593437175855906644
39
64613283001136726019340687064423394891992153043852815416596301185494236348635245857466428346090629162795649265847230036085559898976191629324814009459248989946848688623225
86
01705514689056498372760039274870695504637888197416994290491028531804107765160072633847216389037822700138435995997302657271988043246643123597585552171969605139210102265963
24
78878309774053331315517629781528807186526031763272648280944993704562580999380530584976995700802893798080149029010052938984722799471678048216894241591182842576964647865057
31
53251783023363072982516922103465842658944474649161238546897185079681729091390321828341111848213847677283165486532123173820041319905105189670222018870495856871805095907303
60
69304029372160389689176055876769553823180937058262570838983874090984686566342713975000132918351059433217298798252437075082720879598594371576676601557826996603431977526233
08
89899625878006280095609444169323779449554410336965862615562560106693903032038789709836737860870566414335851061116583145204245132085085899949323648316896711949516716195676
22
70709069738895888555795624666415365617235493018073940047605298017217713916867880002778519661730700612845173075825037356431020651124437308252296250404531605907413438818725
63
47791383066059093188025223100853401768402614015396169891920751471080337577088497401418345997539720598786820648791160649698581776011539720584982226989071813494326918018211
73
31880636539108936898117148913574566805428074851701758582666396335701893544498326697628350926579222017463721902731196417514899440100796368760178267471070199454732188878327
42
60889667243715747134206000937042513098936305374597842799804031329894172664922904257309583685344162156405572902820662240038632375263809102332698978388604237596256015679752
62
69507986398681042948323331602672165551781208992646778049357413263871374084238855465383361586434513054396242813972795597259951107063143059926154956229583202327080576811566
90
48958661052203005737252984721187478271367136660586692710948755639748584894759108197270338782844398644867434562009581619303147273459619004993184243379752436624893633212448
50
59719925236685292493053462527641378534132089431289015237380925560459870909127666623296787033288820591349495800740744731433888007245323217473096597419671144414531271327902
05
10100476710143506388579534784472553898015419233170275198961806351526825431731938329258919315301641305489723111286664654929719304792643282955671909288169209104233412200745
4
24204990087258504620805110487588305949599031118873666850941488217257345763552339640384813182131674083590069164000532622581847837650678044511777173286581898992153583094477
65
350341796875

Approximately 3.00 times 10^6328
```

Problem Complexity

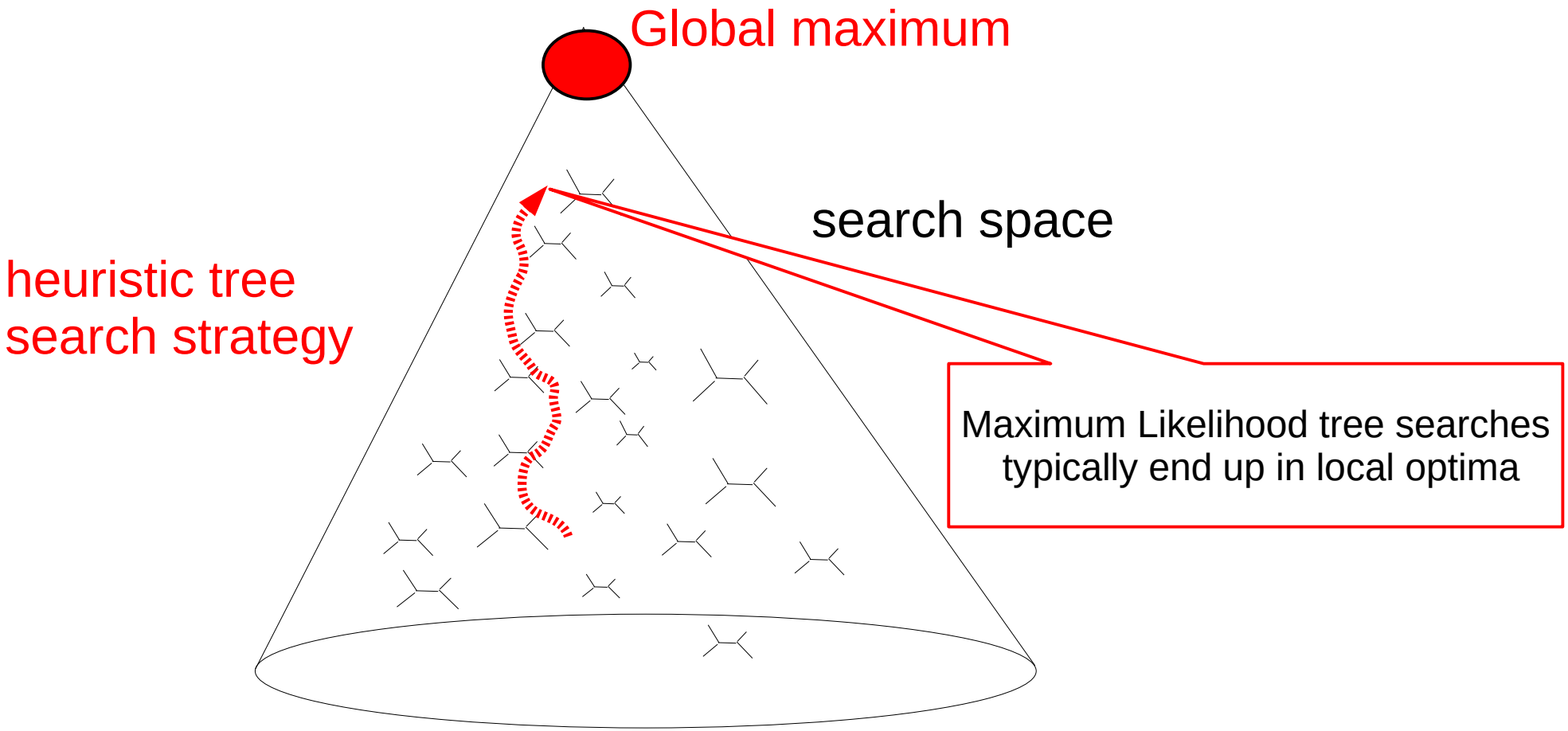


Problem Complexity

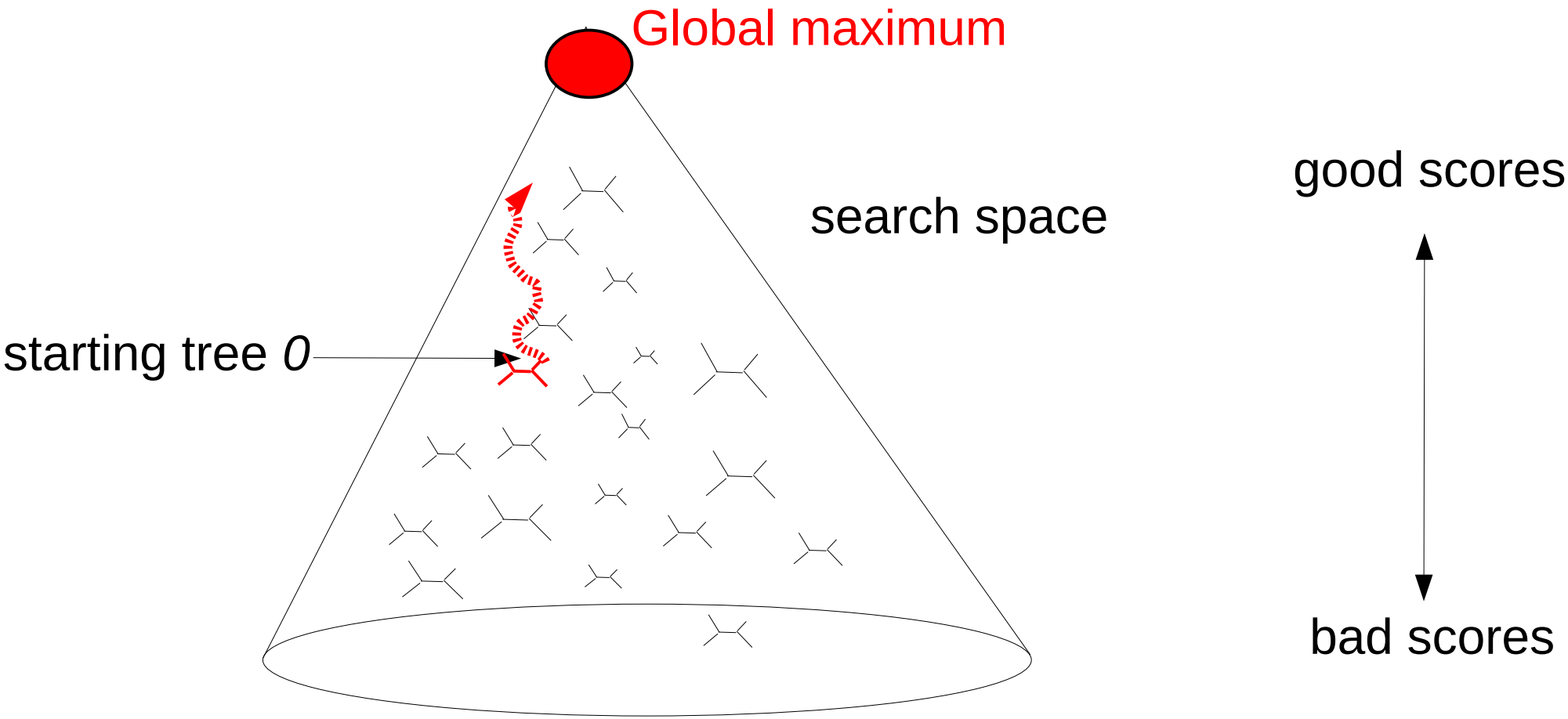


Finding the best tree under Maximum Likelihood is **NP-hard**!

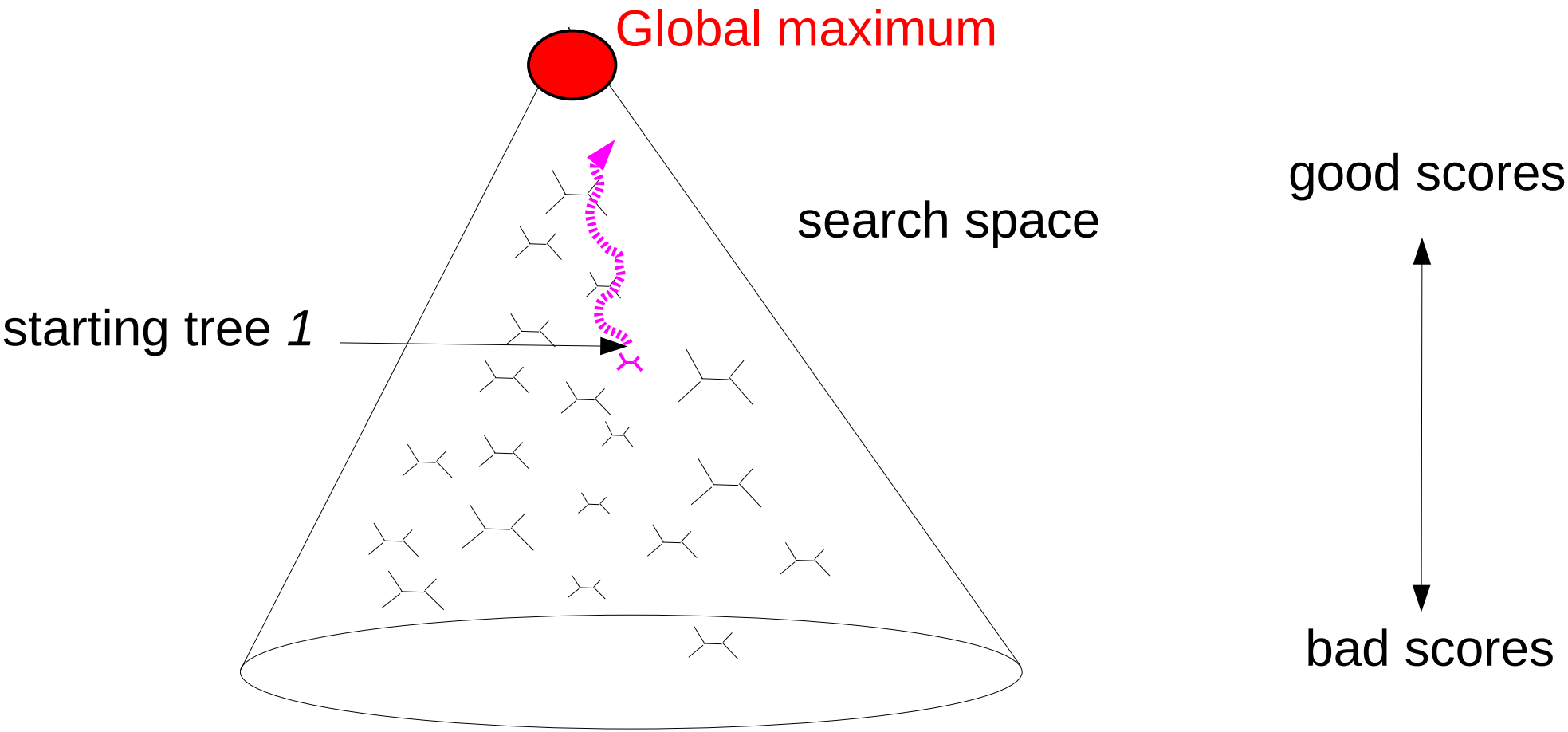
Problem Complexity



Starting Trees

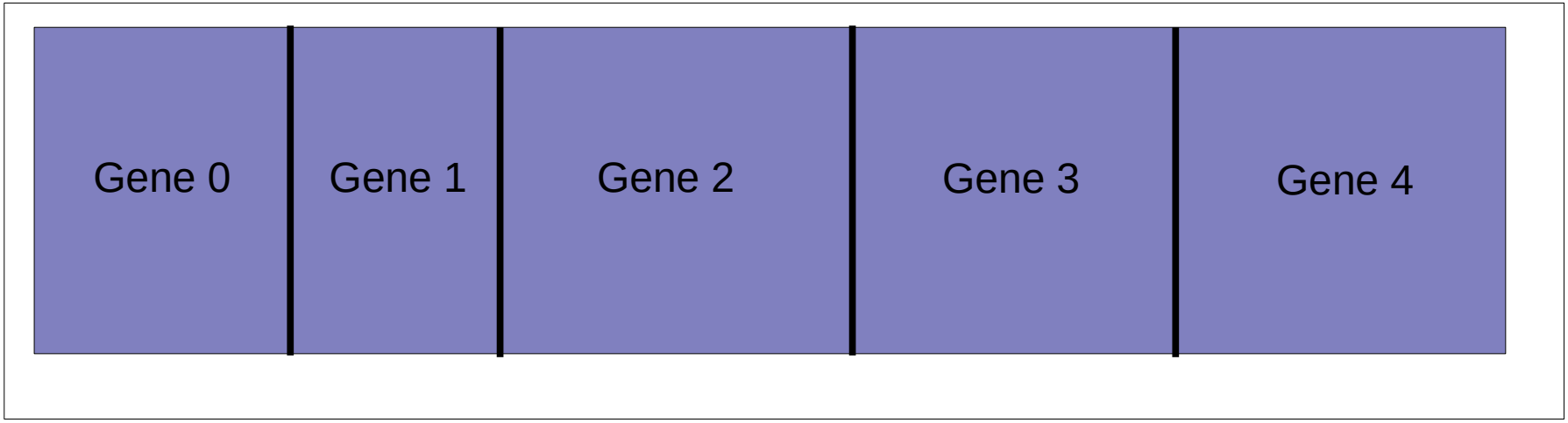


Starting Trees



Main Prior Contributions

- Efficient tree search algorithms
- Low-level **hardware-aware optimization** of likelihood calculations (95% of total execution time)
- **Algorithmic optimization** of likelihood calculations
- Parallelization for analysis of large genomic datasets
  - Optimal data distribution
  - Optimization of parallel I/O



Main Prior Contributions

- Efficient Search Algorithms
- Low-level **hardware-aware optimization** of likelihood calculations (95% of total execution time)
- **Algorithmic optimization** of likelihood calculations
- Parallelization for analysis of large datasets
  - Optimal data distribution
  - Optimization of parallel I/O
- Software for Supercomputers
  - `RAxML-NG` - *scales from the laptop to the supercomputer*
  - `ExaBayes` - *Bayesian inference on extremely large datasets*

Main Prior Contributions

- Efficient Search Algorithms
- Low-level **hardware-aware optimization** of likelihood calculations (95% of total execution time)
- **Algorithmic optimization** of likelihood calculations
- Parallelization for analysis of large datasets
  - Optimal data distribution
  - Optimization of parallel I/O
- Software for Supercomputers
  - `RAxML-NG` - *scales from the laptop to the supercomputer*
  - `ExaBayes` - *Bayesian inference on extremely large datasets*
- **Support, Maintenance, Extension**

Outline

- Our Approach to Bioinformatics
- Introduction to Phylogenetic Inference
- **Sources of Uncertainty**
- Phylogenetic Difficulty
- Other stuff we are working on

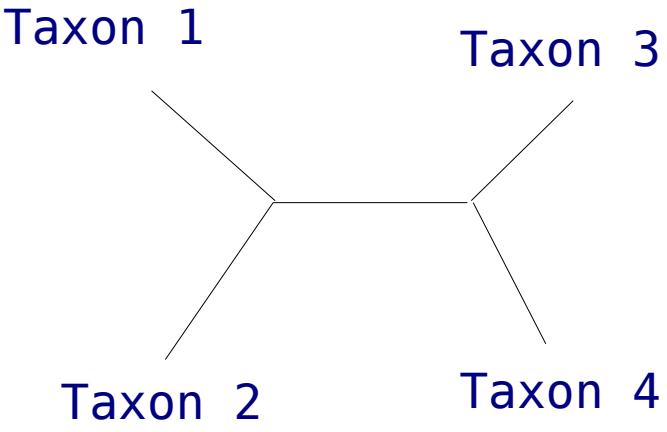
Tree Inference Pipeline

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

→ MSA Program →

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

→ Tree inference program →



Orthology assignment:
Mostly “dirty” *ad hoc* methods
→ no widely used uncertainty quantification approach

Tree Inference Pipeline

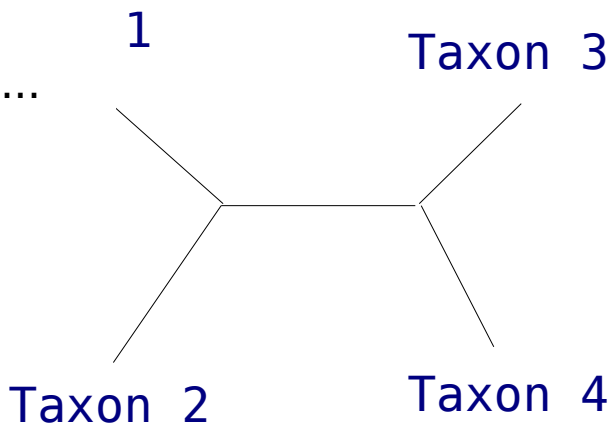
```
Taxon 1:ACGTTT
Taxon 2:ACGTT
Taxon 3:ACCCT
Taxon 4:AGGGTTT
```

→ MSA Program →

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

→ Tree inference program →

Multiple Sequence Alignment:
Mostly *ad hoc* methods →
no widely used uncertainty
quantification approach, but ...



Muscle5

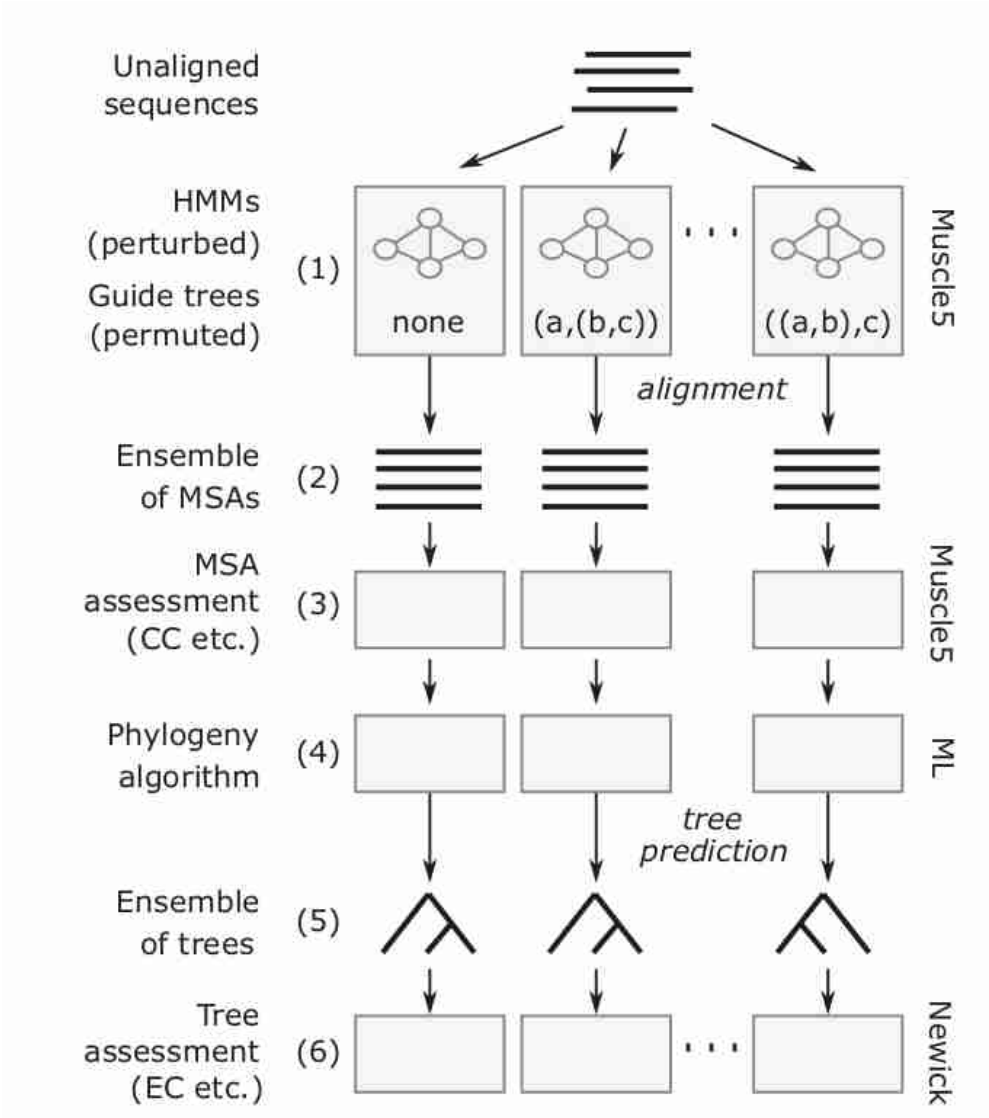
Article | Open Access | Published: 15 November 2022

Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny

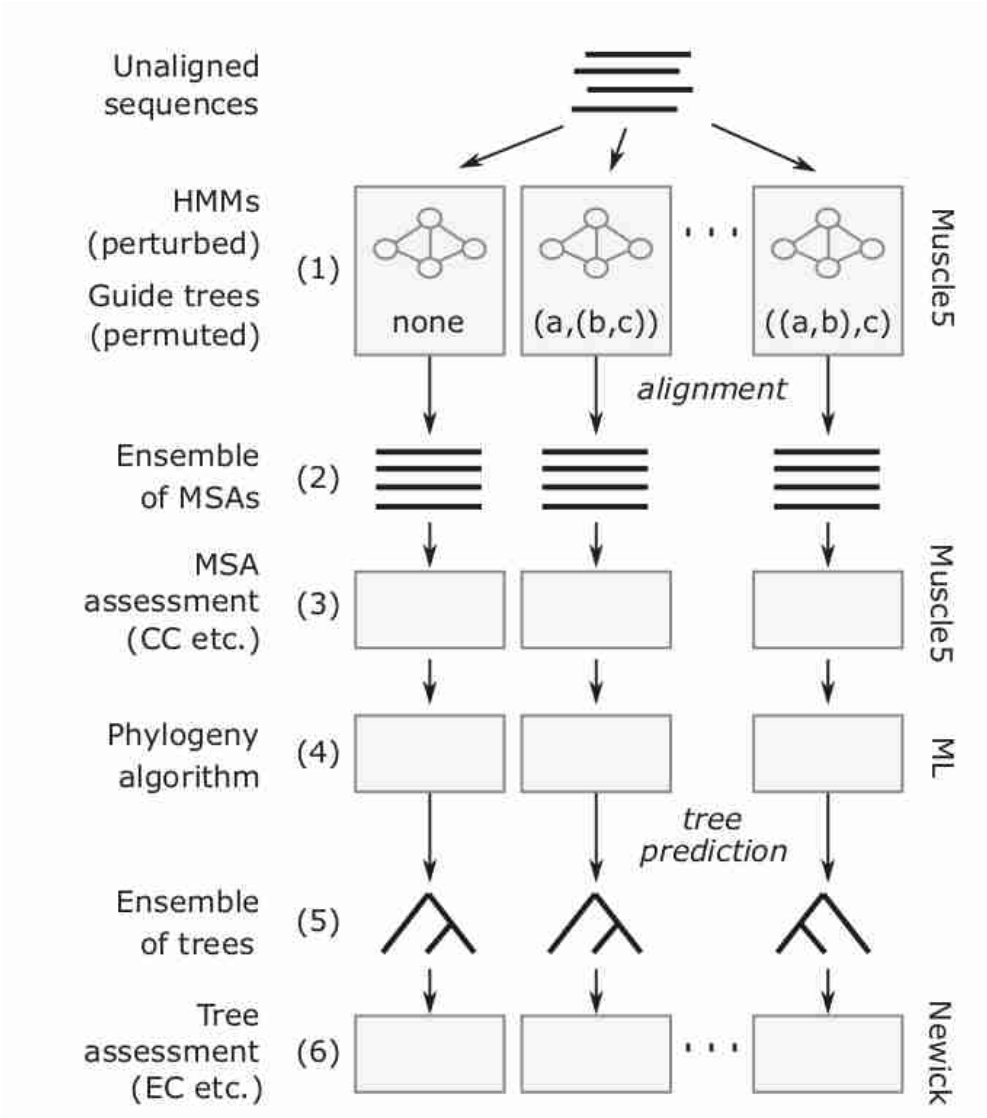
Robert C. Edgar

Nature Communications **13**, Article number: 6968 (2022) | Cite this article

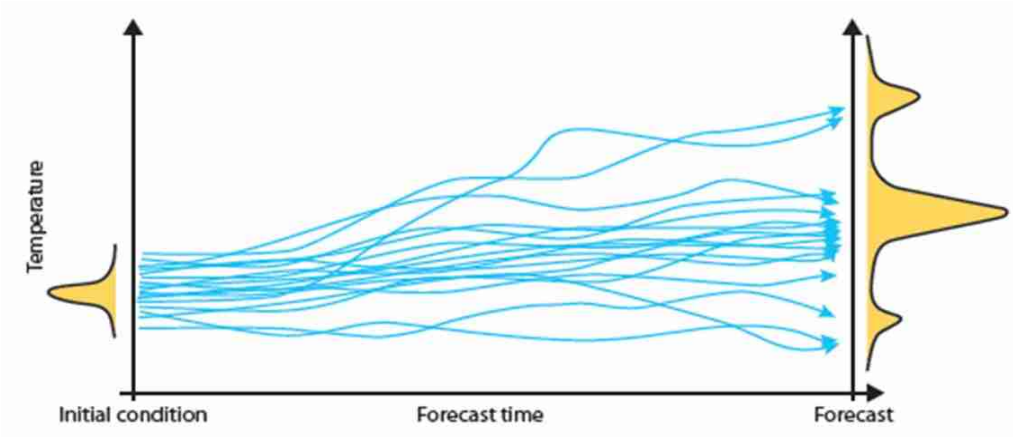
Muscle5


Unaligned sequences
HMMs (perturbed)
Guide trees (permuted)
(1)
none
(a,(b,c))
((a,b),c)
Muscle5
alignment
Ensemble of MSAs
(2)
MSA assessment (CC etc.)
(3)
Muscle5
Phylogeny algorithm
(4)
ML
tree prediction
Ensemble of trees
(5)
Tree assessment (EC etc.)
(6)
Newick

Muscle5



Temperature Ensemble Forecast



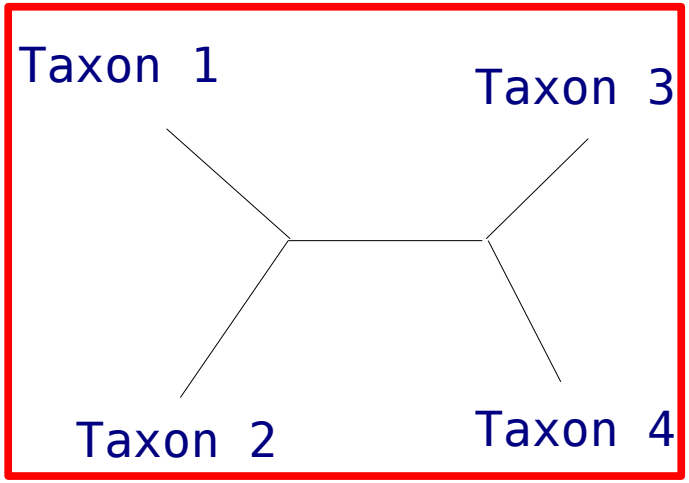
Tree Inference Pipeline

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

MSA Program

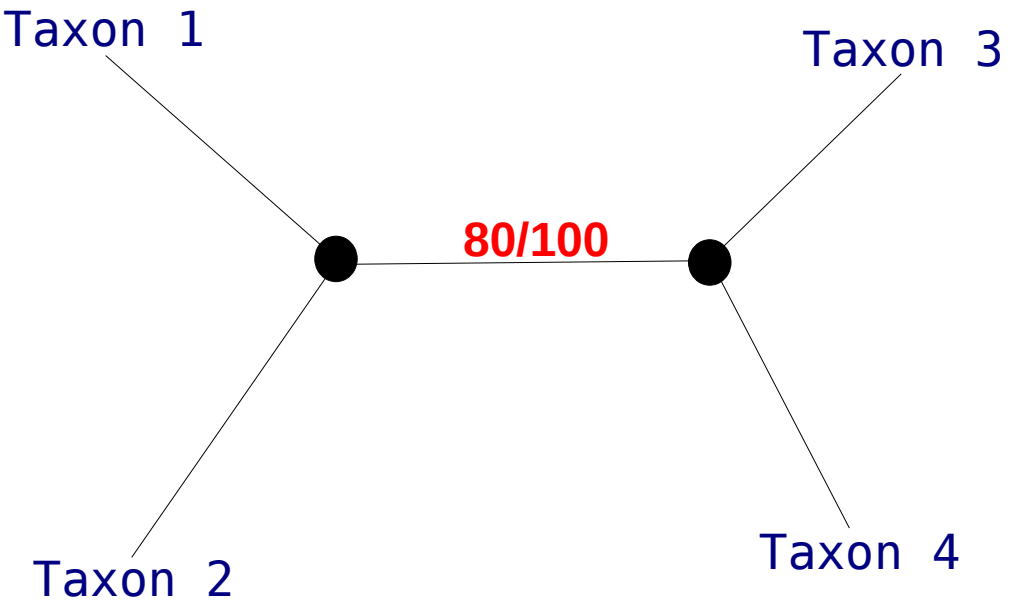
```
Taxon 1:ACGTTT-
Taxon 2:ACGTT--
Taxon 3:ACCCT--
Taxon 4:AGGGTTT
```

Tree inference program



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

A Tree with Support Values



Sources of Uncertainty

1) Orthology Assignment
2) Multiple Sequence Alignment
3) Tree Inference
4) BUT

Software Issues

- Bugs & Software Quality
- Numerical Instability
- Reproducibility
- We re-designed & optimized numerous tools – the *Next Generation* (NG) tools series
  - `RAxML-NG`
  - `ModelTest-NG`
  - `EPA-NG`
  - `Lagrange-NG`

Sources of Uncertainty

1) Orthology Assignment
2) Multiple Sequence Alignment
3) Tree Inference
4) Software issues
5) BUT

Propagating Uncertainty

- Assume
  - *10* alternative orthology assignments
  - *10 x 10* alternative MSAs
  - *10 x 10 x 10* alternative trees

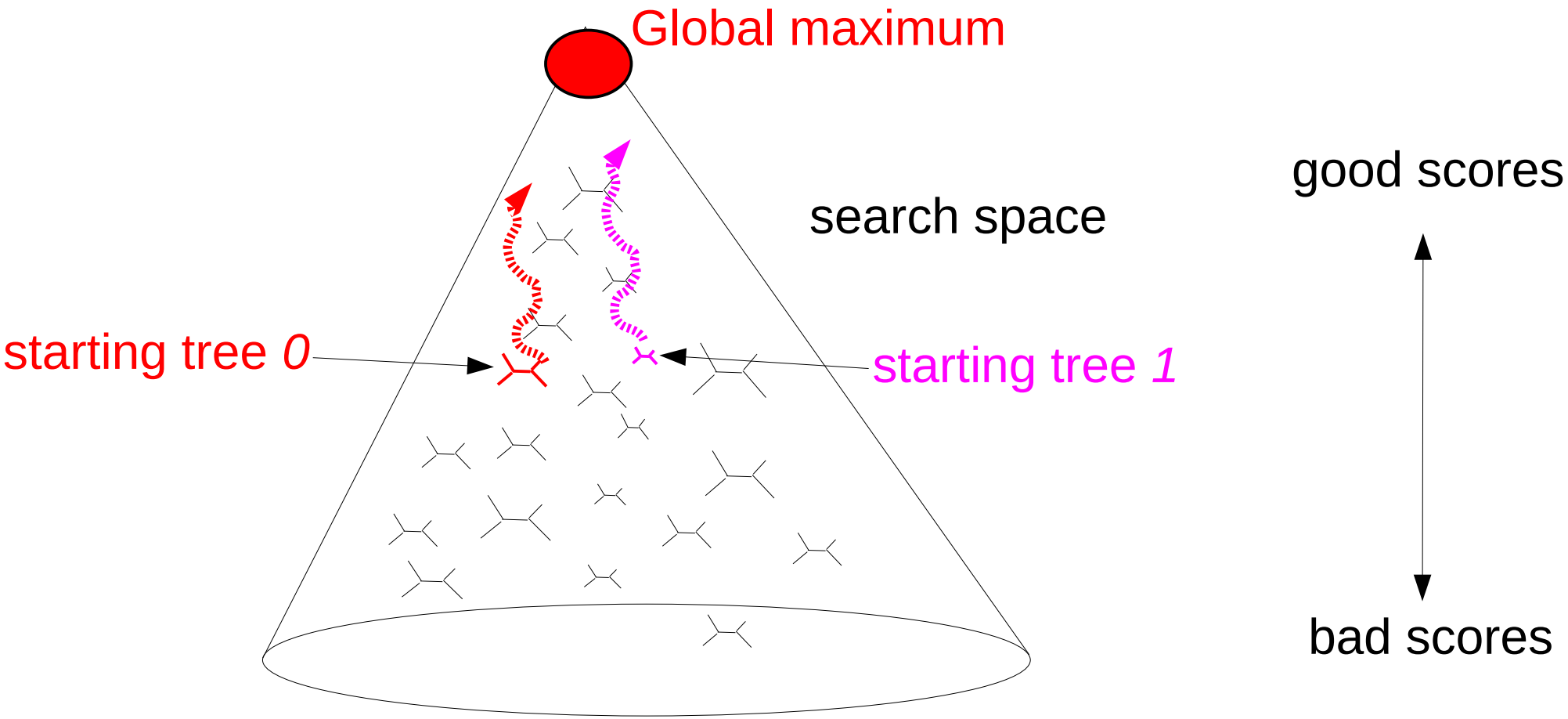
  → exponential explosion with increasing pipeline length

  → intelligent ways to explore parameter space in pipelines needed

Outline

- Our Approach to Bioinformatics
- Introduction to Phylogenetic Inference
- Sources of Uncertainty
- **Phylogenetic Difficulty**
- Other stuff we are working on

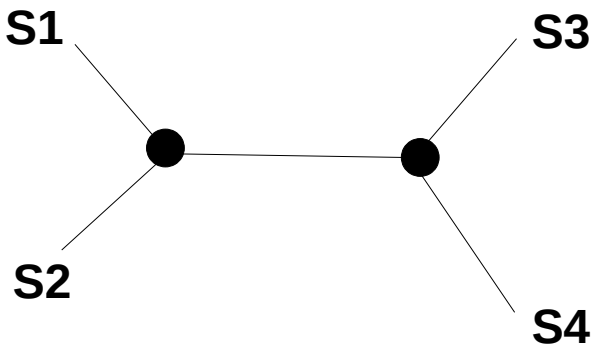
Can we predict how difficult a phylogenetic analysis will be?



Phylogenetic Inference

MSA

S1	ACGTT
S2	ACCGG
S3	TGGAG
S4	GGCTT

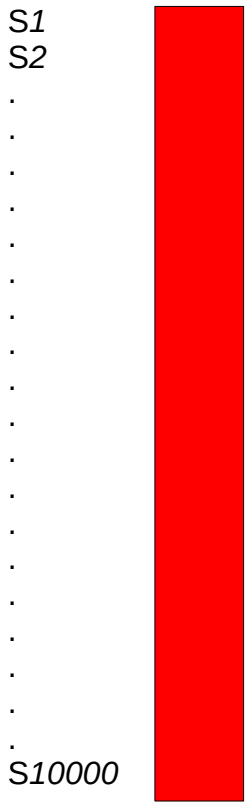


The difficulty of inferring a tree depends on the shape of the multiple sequence alignment

Dataset Shapes

This?

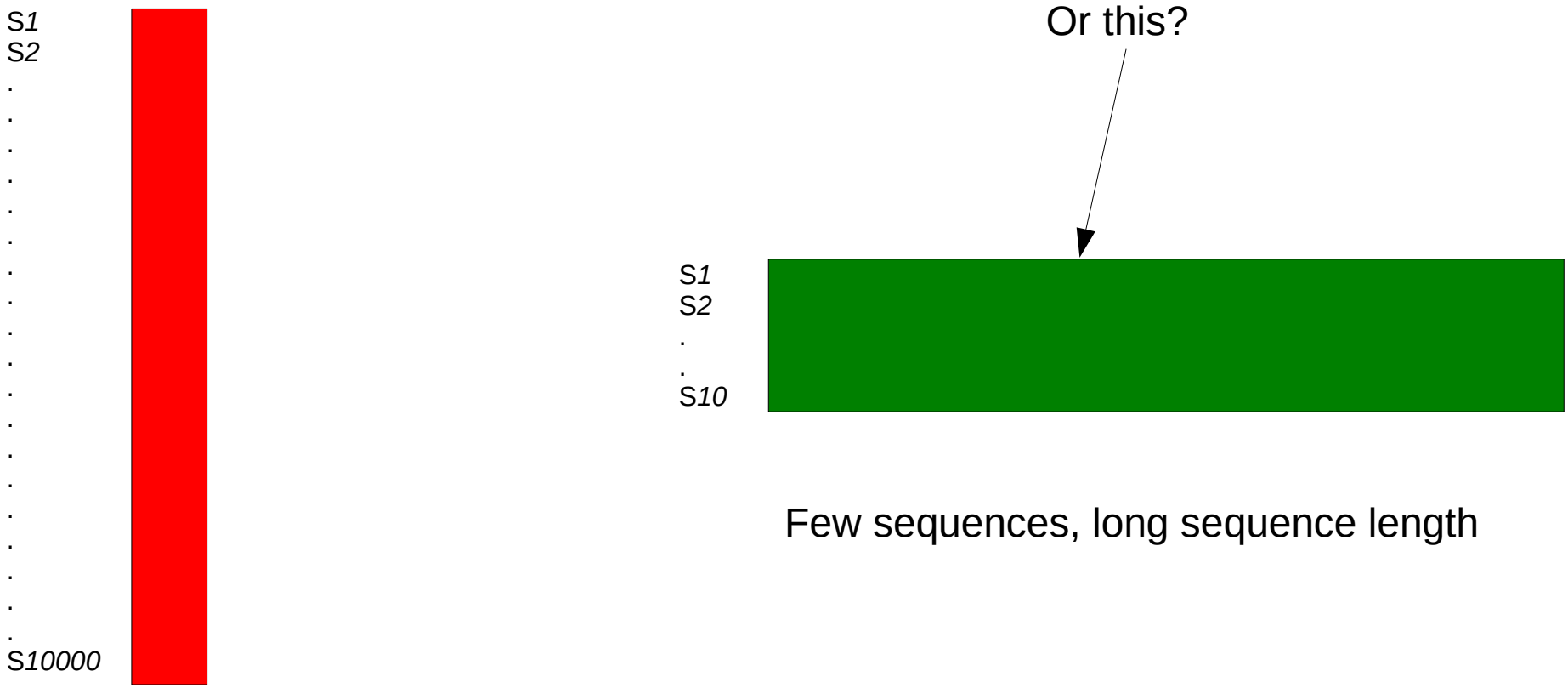
Which data is more difficult to analyze?



Thousands of sequences, short sequence length

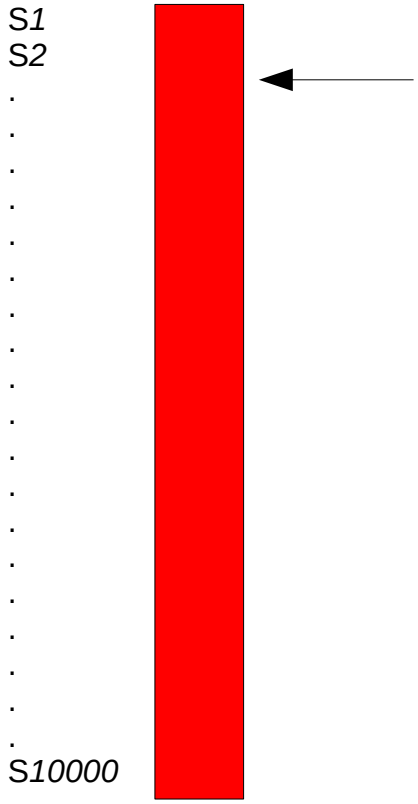
Dataset Shapes

Which data is more difficult to analyze?



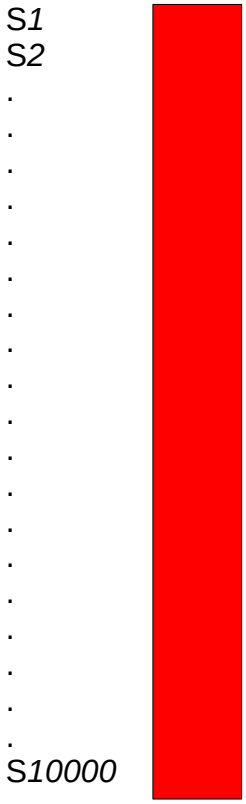
Few sequences, long sequence length

Dataset Shapes



Intuitively it is this dataset here, as it contains much less information for telling apart more sequences

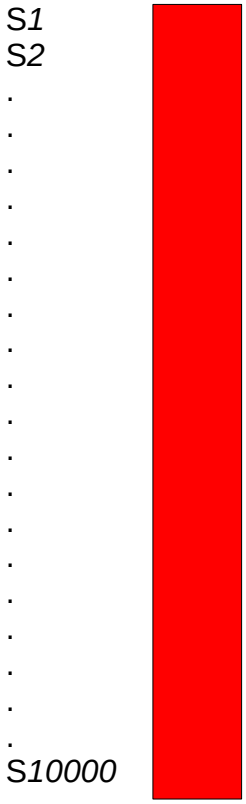
Dataset Shapes



Intuitively it is this dataset here, as it contains much less information for telling apart more sequences

SARS-CoV-2 is such a difficult dataset; it even exhibits some additional difficulties:

Dataset Shapes

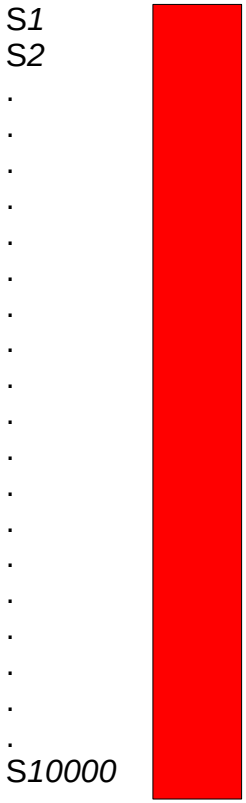


Intuitively it is this dataset here, as it contains much less information for telling apart more sequences

SARS-CoV-2 is such a difficult dataset; it even exhibits some additional difficulties:

1. Due to the low mutation rate (rate at which nucleotides change) sequences are very similar to each other

Dataset Shapes

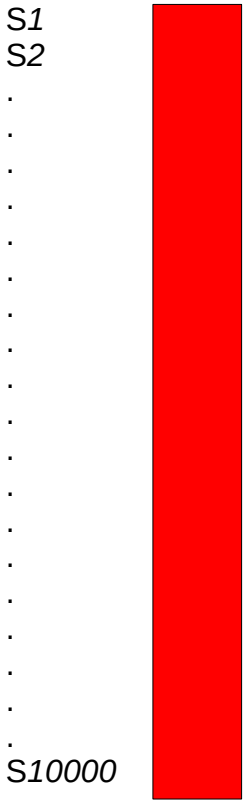


Intuitively it is this dataset here, as it contains much less information for telling apart more sequences

SARS-CoV-2 is such a difficult dataset; it even exhibits some additional difficulties:

1. Due to the low mutation rate (rate at which nucleotides change) sequences are very similar to each other
2. The genome is ≈ *30,000* nucleotides long, but the sequences differ in only *1500-2000* positions → highly similar

Dataset Shapes



Intuitively it is this dataset here, as it contains much less information for telling apart more sequences

SARS-CoV-2 is such a difficult dataset; it even exhibits some additional difficulties:

1. Due to the low mutation rate (rate at which nucleotides change) sequences are very similar to each other
2. The genome is ≈ *30,000* nucleotides long, but the sequences differ in only *1500-2000* positions → highly similar
3. The input sequences are **not from distinct species**!

Consequences

- SARS-CoV-2 data
  - Extremely hard to infer a reliable tree
  - Numerical issues with tree inference tools because the sequences are so closely related
  - Post-analyzing the tree (e.g., determining the root, identifying virus sub-classes) appears to not be feasible using computational tools

Consequences

- SARS-CoV-2 data
  - **Extremely hard to infer a reliable tree**
  - Numerical issues with tree inference tools because the sequences are so closely related
  - Post-analyzing the tree (e.g., determining the root, identifying virus sub-classes) appears to not be feasible using computational tools

For details, see: Benoit Morel, Pierre Barbera, Lucas Czech, Ben Bettisworth, Lukas Hübner, Sarah Lutteropp, Dora Serdari, Evangelia-Georgia Kostaki, Ioannis Mamais, Alexey Kozlov, Pavlos Pavlidis, Dimitrios Paraskevis, Alexandros Stamatakis. **"Phylogenetic analysis of SARS-CoV-2 data is difficult"**, *Molecular Biology and Evolution* 2021

Phylogenetic Inference

- Assembled *4* distinct input datasets
- Per input dataset → executed *100* tree searches
- As we use likelihood models, we determined the trees that are **not statistically significantly different** from each other per set of *100* trees

Results

- For all input datasets about *70* out of *100* trees are not significantly different from each other with respect to their likelihood scores

Results

- For all input datasets about *70* out of *100* trees are not significantly different from each other with respect to their likelihood scores
- But, their pair-wise topological differences (difference in tree shapes) amount on average to **70%** !

Results

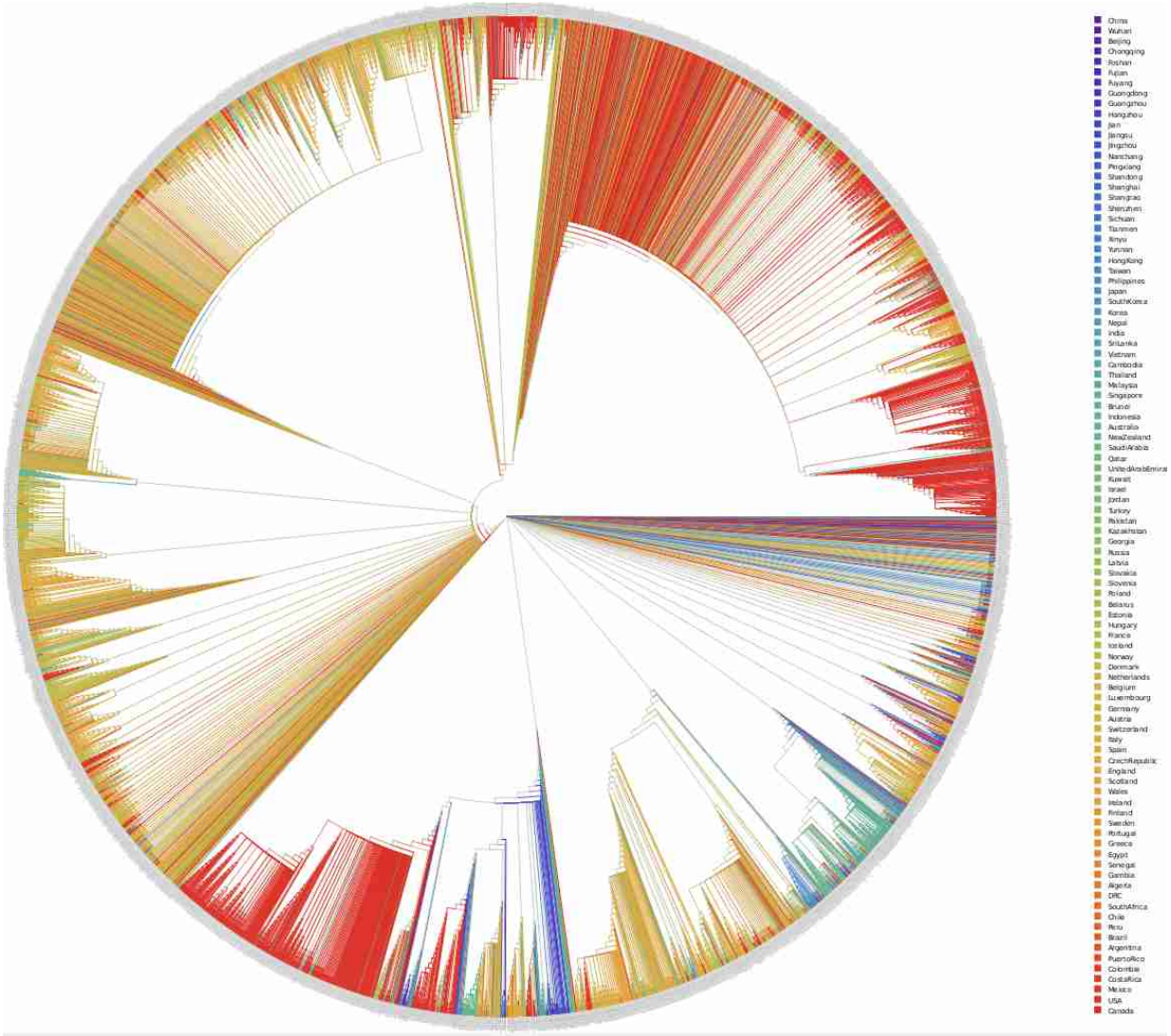
- For all input datasets about *70* out of *100* trees are not significantly different from each other with respect to their likelihood scores
- But, their pair-wise topological differences (difference in tree shapes) amount on average to **70%** !

  → extremely weak signal

  → don't draw conclusions from a single tree!

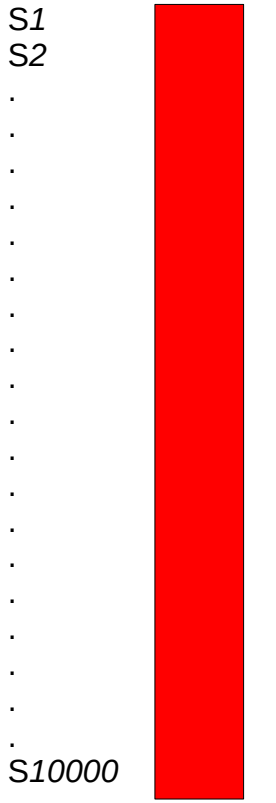
  → try to summarize the trees via summary statistics!

Summarized Trees



SARS-CoV-2 consensus tree colored by country

Difficulty of an MSA

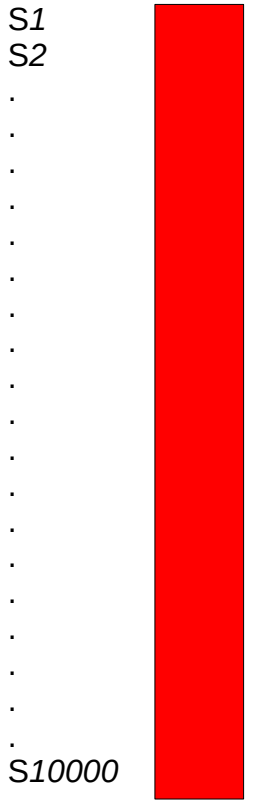


difficult

easy

Difficulty of an MSA

This is very hand-wavy → can we quantify & predict this



difficult

easy

Difficulty Prediction

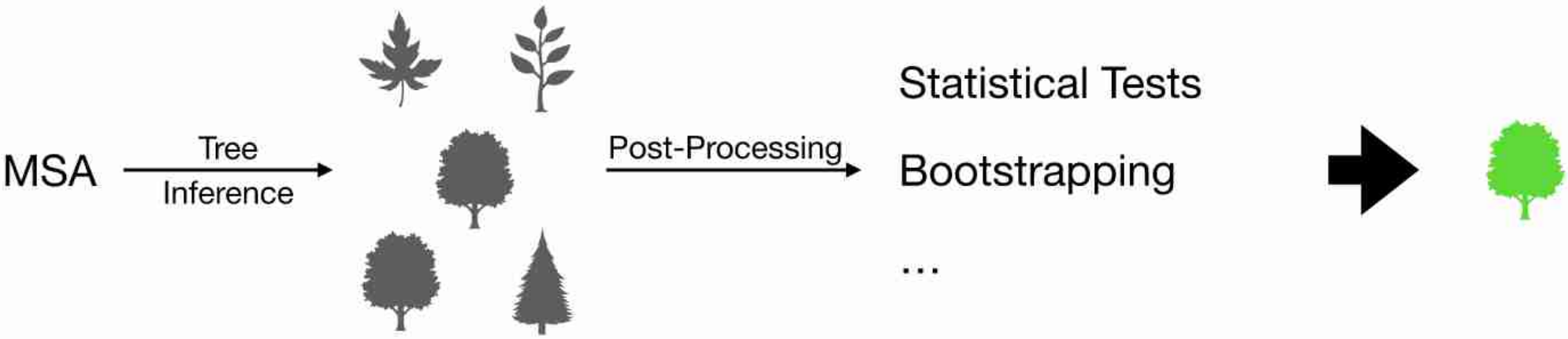
JOURNAL ARTICLE

From Easy to Hopeless—Predicting the Difficulty of Phylogenetic Analyses

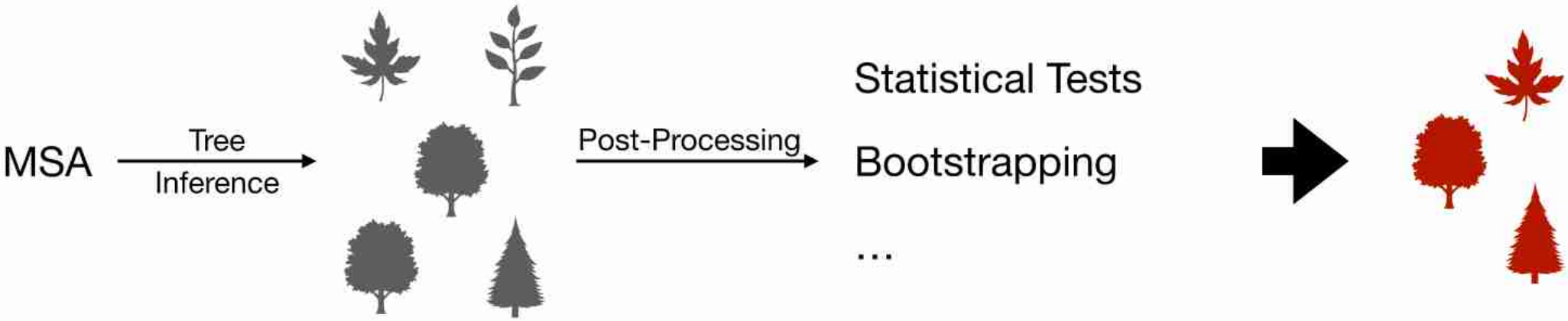
Julia Haag, Dimitri Höhler, Ben Bettisworth, Alexandros Stamatakis

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

Easy


MSA
Tree
Inference
Post-Processing
Statistical Tests
Bootstrapping
...

Difficult


MSA
Tree
Inference
Post-Processing
Statistical Tests
Bootstrapping
...

What does Difficulty mean?

Difficulty = ruggedness of the tree space

Easy → Difficult

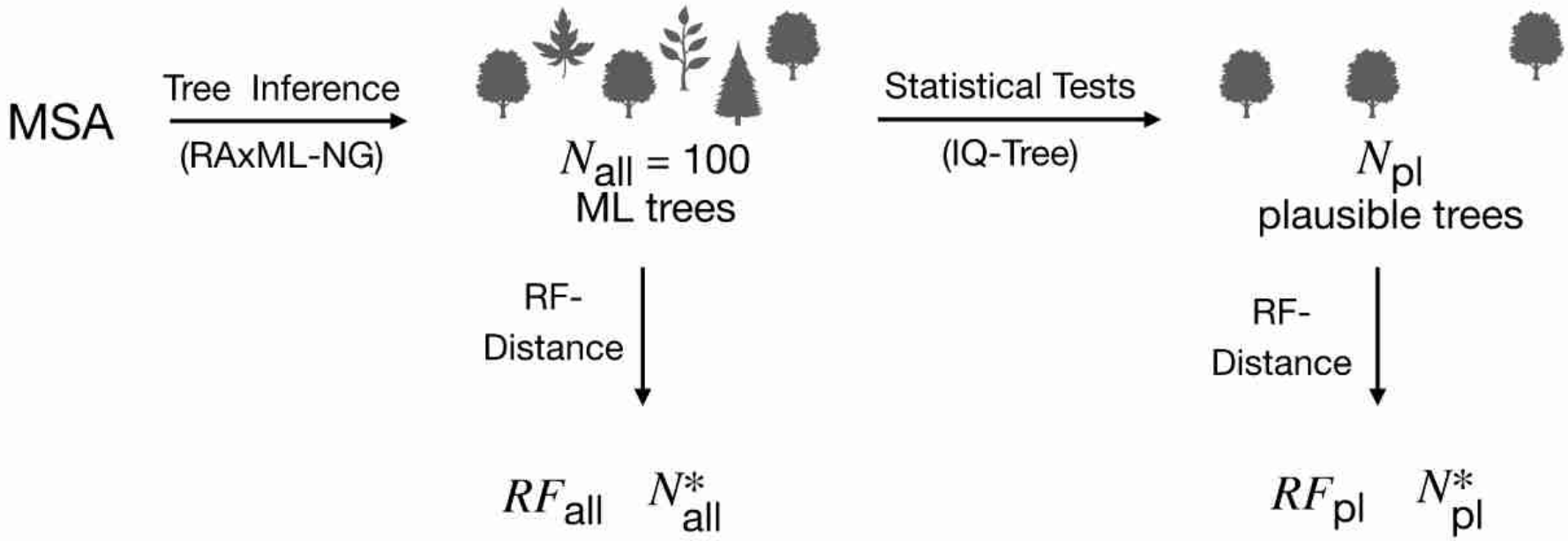
- Few highly similar tree topologies
- Single likelihood peak

- Highly distinct topologies, statistically indistinguishable
- Multiple likelihood peaks

Predicting Difficulty with `Pythia`

- `Pythia` = Boosted Tree Regressor
- Supervised Regression Task
  - Predict difficulty between 0.0 (easy) and 1 (difficult)
  - Ground truth difficulty as training target based on 100 distinct Maximum Likelihood tree inferences
- Trained on 4K empirical MSAs
  - Mean absolute % error: 2.5%

Definition of Difficulty



$$\text{difficulty(MSA)} = \frac{1}{5} \cdot \left[RF_{\text{all}} + \frac{N^*_{\text{all}}}{N_{\text{all}}} + RF_{\text{pl}} + \frac{N^*_{\text{pl}}}{N_{\text{pl}}} + \left(1 - \frac{N_{\text{pl}}}{N_{\text{all}}}\right) \right]$$

Prediction Features

- Eight Features
  - 4 MSA attributes
    - sites-over-taxa, patterns-over-taxa, % gaps, % invariant sites
  - 2 MSA information metrics
    - Shannon entropy, Bollback multinomial test statistic
  - 2 Parsimony-tree-based features
    - Infer 100 parsimony trees → average RF-Distance, % unique topologies

SARS-CoV-2 Example

"*Phylogenetic Analysis of SARS-CoV-2 Data Is Difficult*" (https://doi.org/10.1093/molbev/msaa314)

```
The predicted difficulty for MSA examples/covid.fasta is: 0.84.

FEATURES:

num_taxa: 4869

num_sites: 28361

[ ... ]

num_sites/num_taxa: 5.82

[ ... ]

avg_rfdist_parsimony: 0.79

proportion_unique_topos_parsimony: 1.0

Feature computation runtime:    1830.182 seconds

[ ... ]
```

Using `Pythia`

- **Prior** to tree inference

  → determine analysis & post-analysis setup

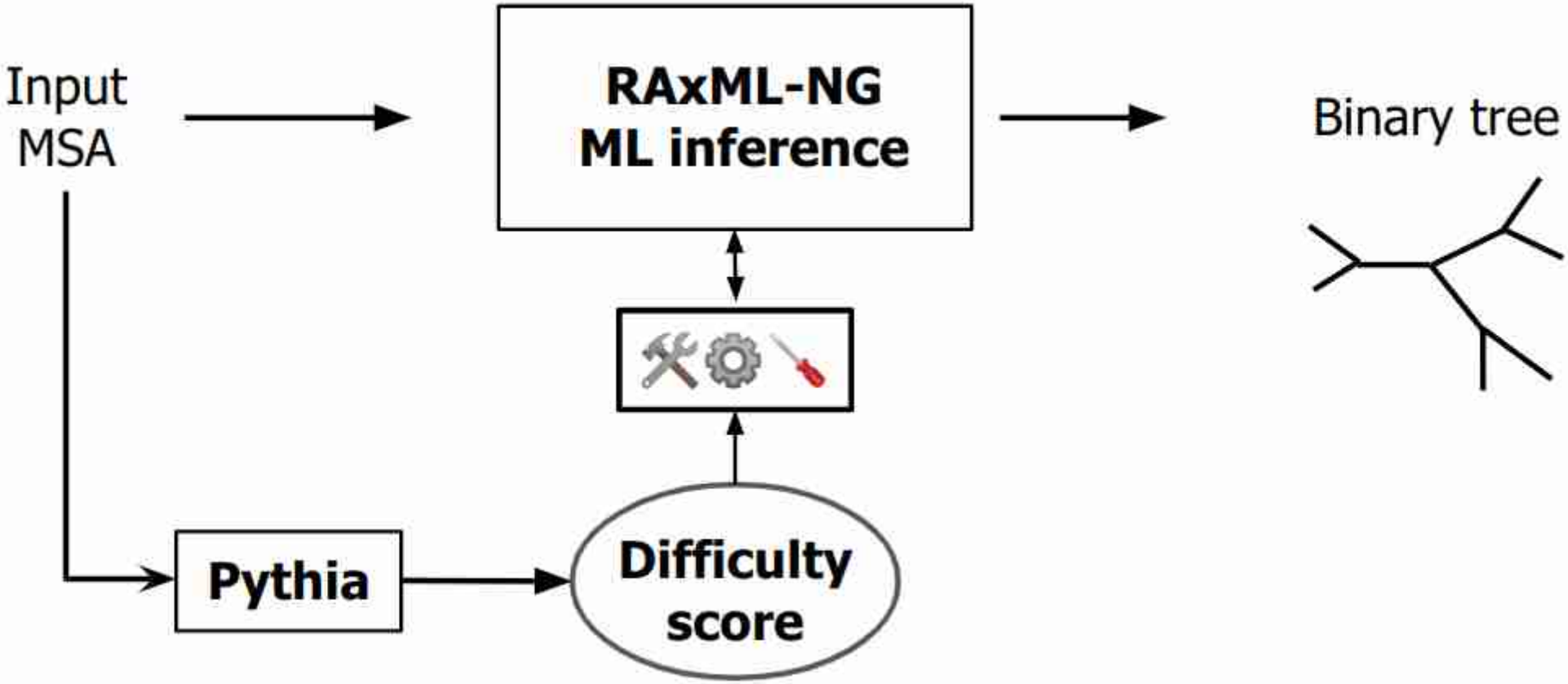
  → adjust/modify MSA

  → adjust user expectations about data

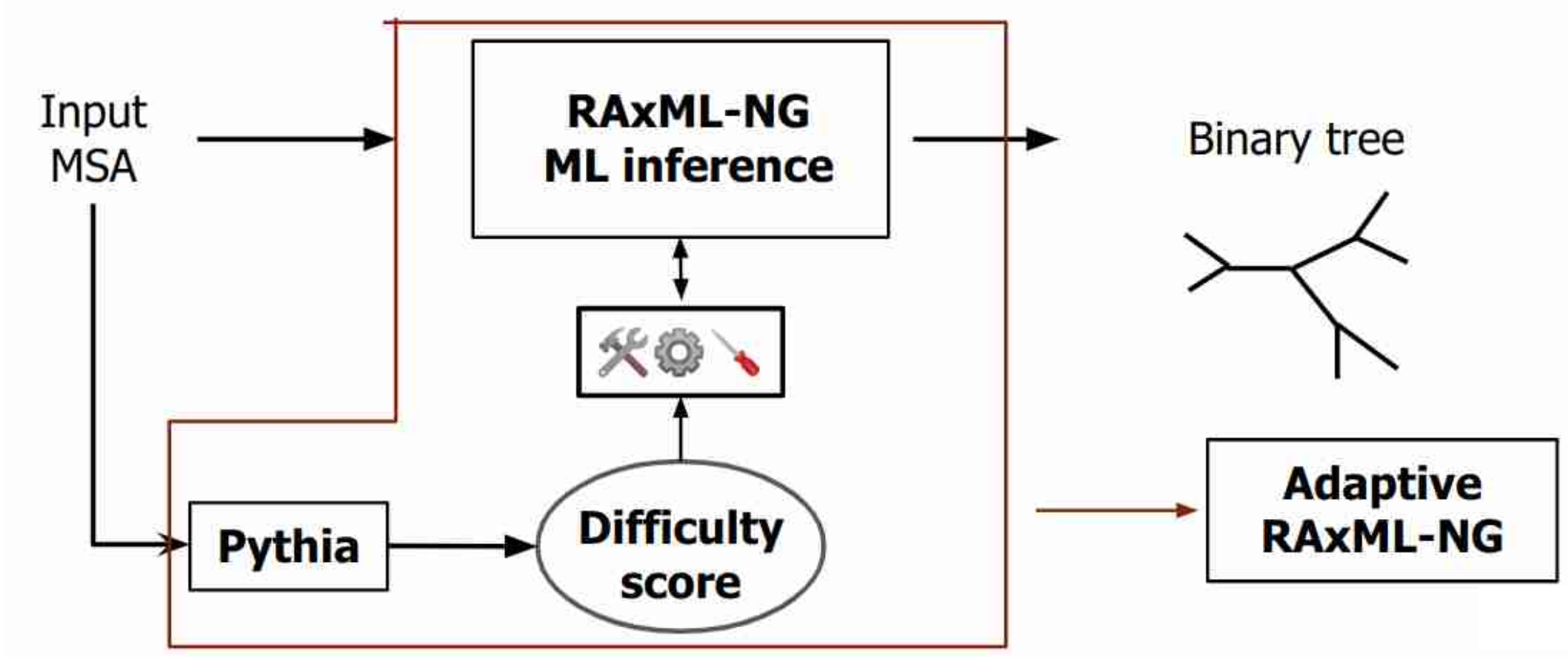
`Pythia` developments

- Next release
  - Trained on 12K datasets (automatic re-training)
  - Additional features
- Deploy to inform tree search heuristics

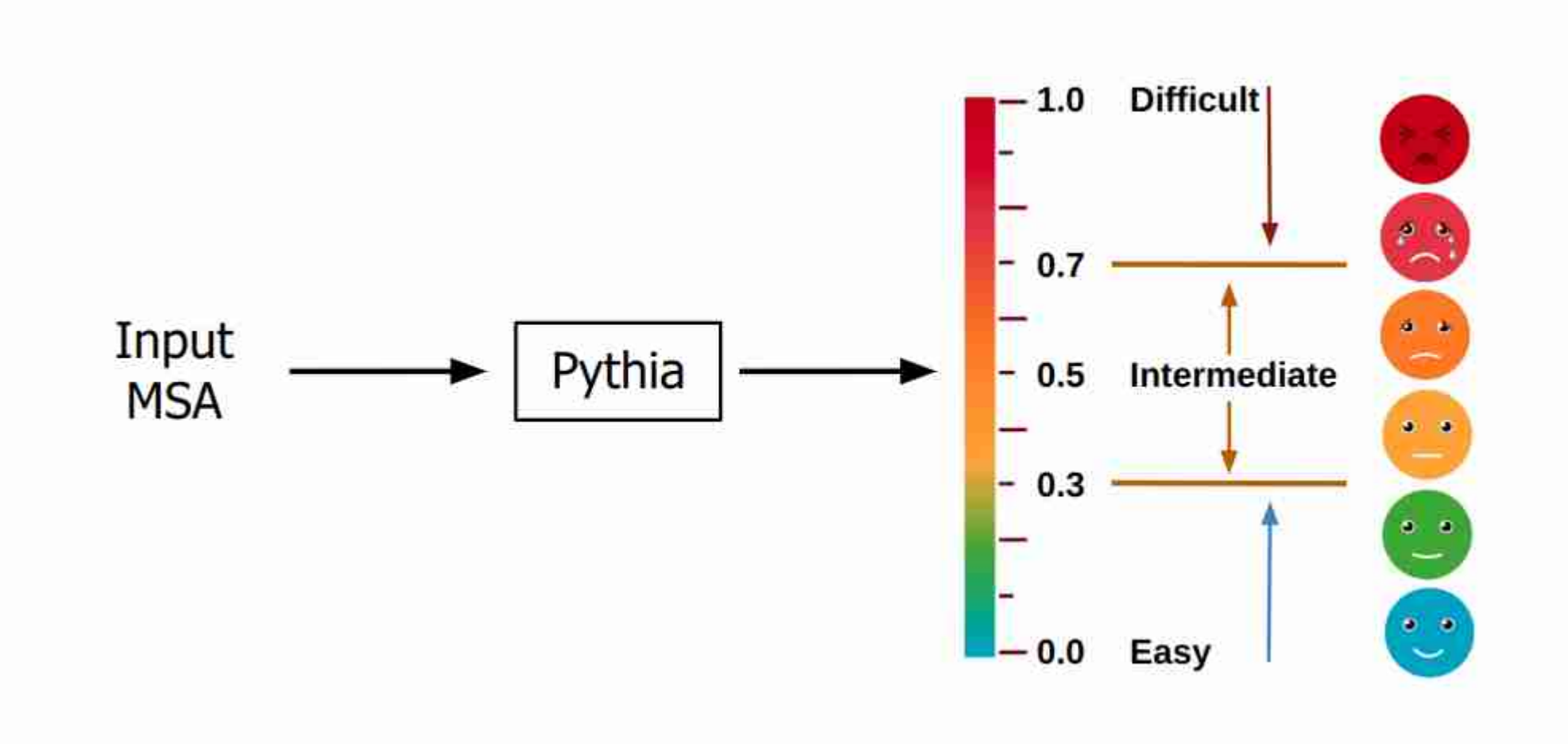
Adaptive RAxML-NG



Adaptive RAxML-NG



Pythia



Adaptive `RAxML-NG` Heuristics

- We modify as a function of difficulty
  - the number of ML tree searches
  - the thoroughness of the search

- And introduce an additional tree search mechanism

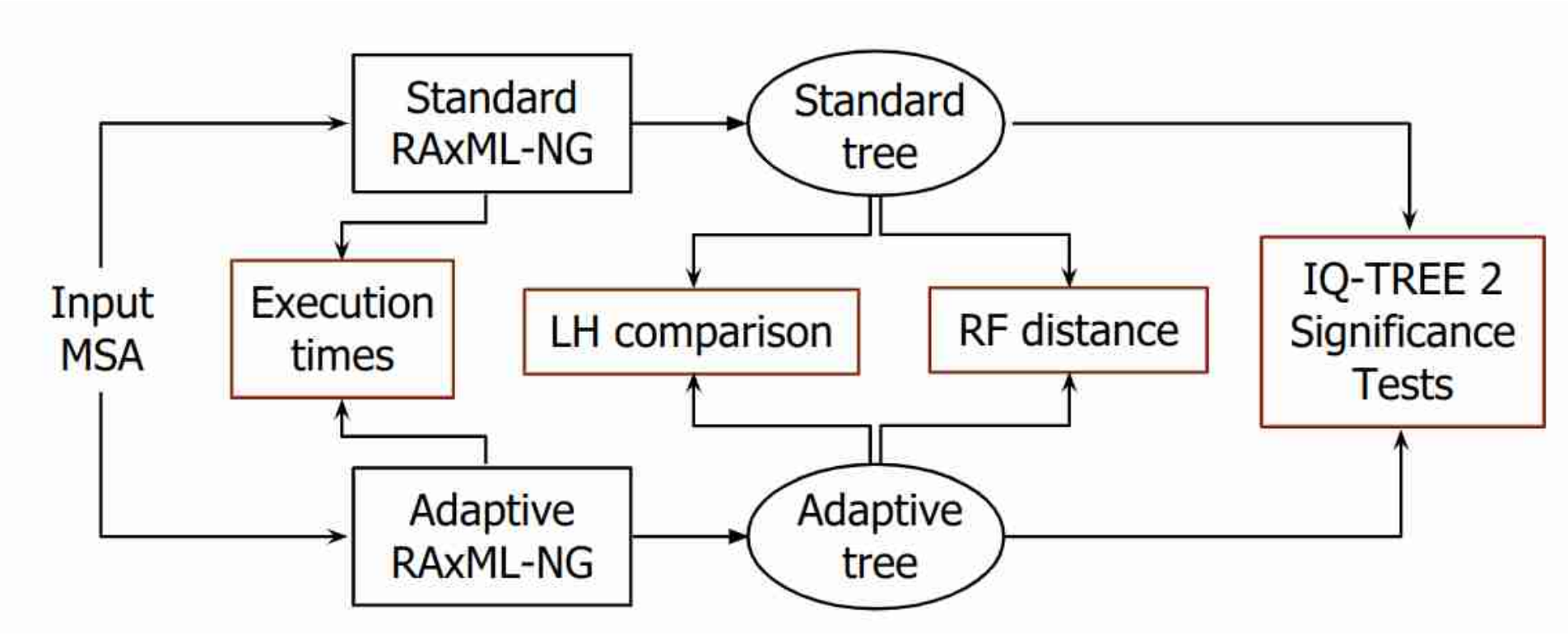
Test Data & Setup

- 10K empirical MSAs from TreeBase

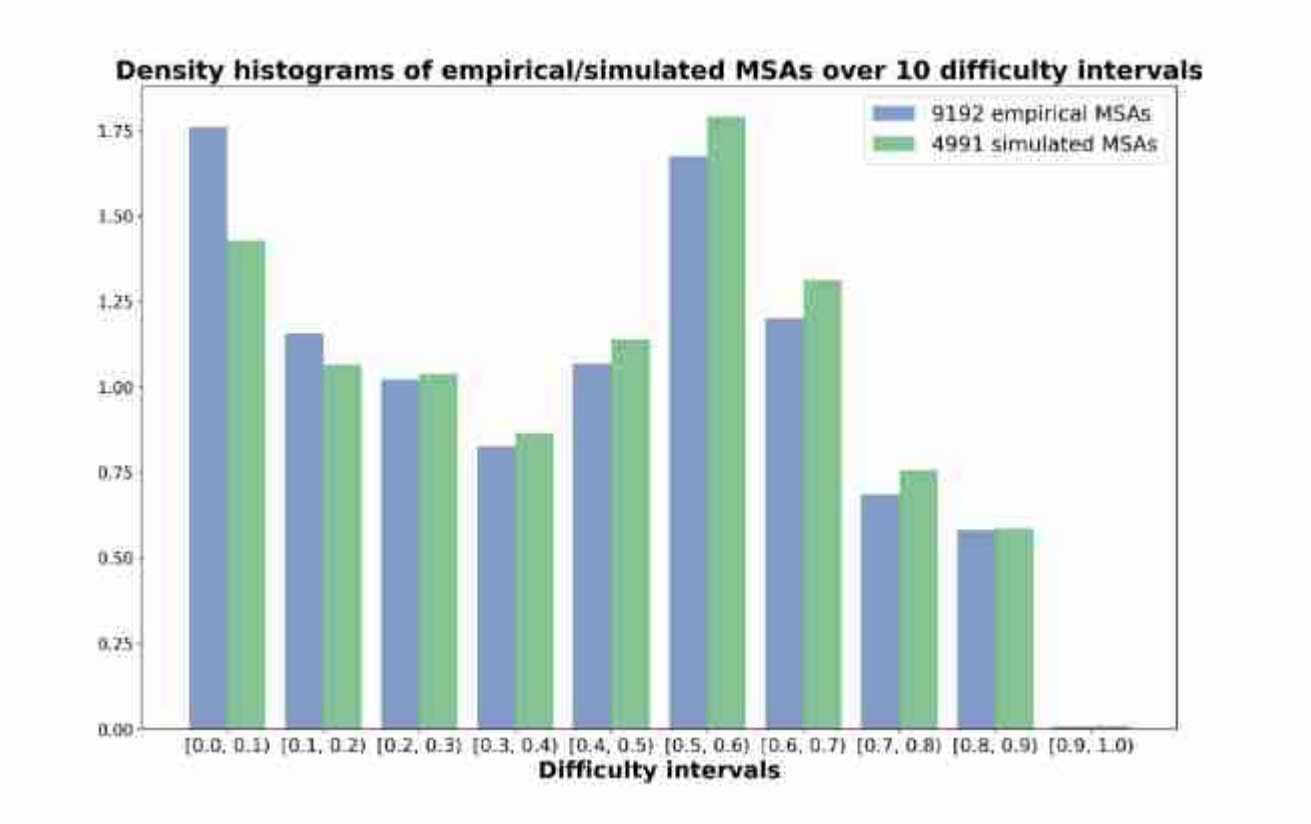
  → 9192 MSAs after filtering

Test Data & Setup

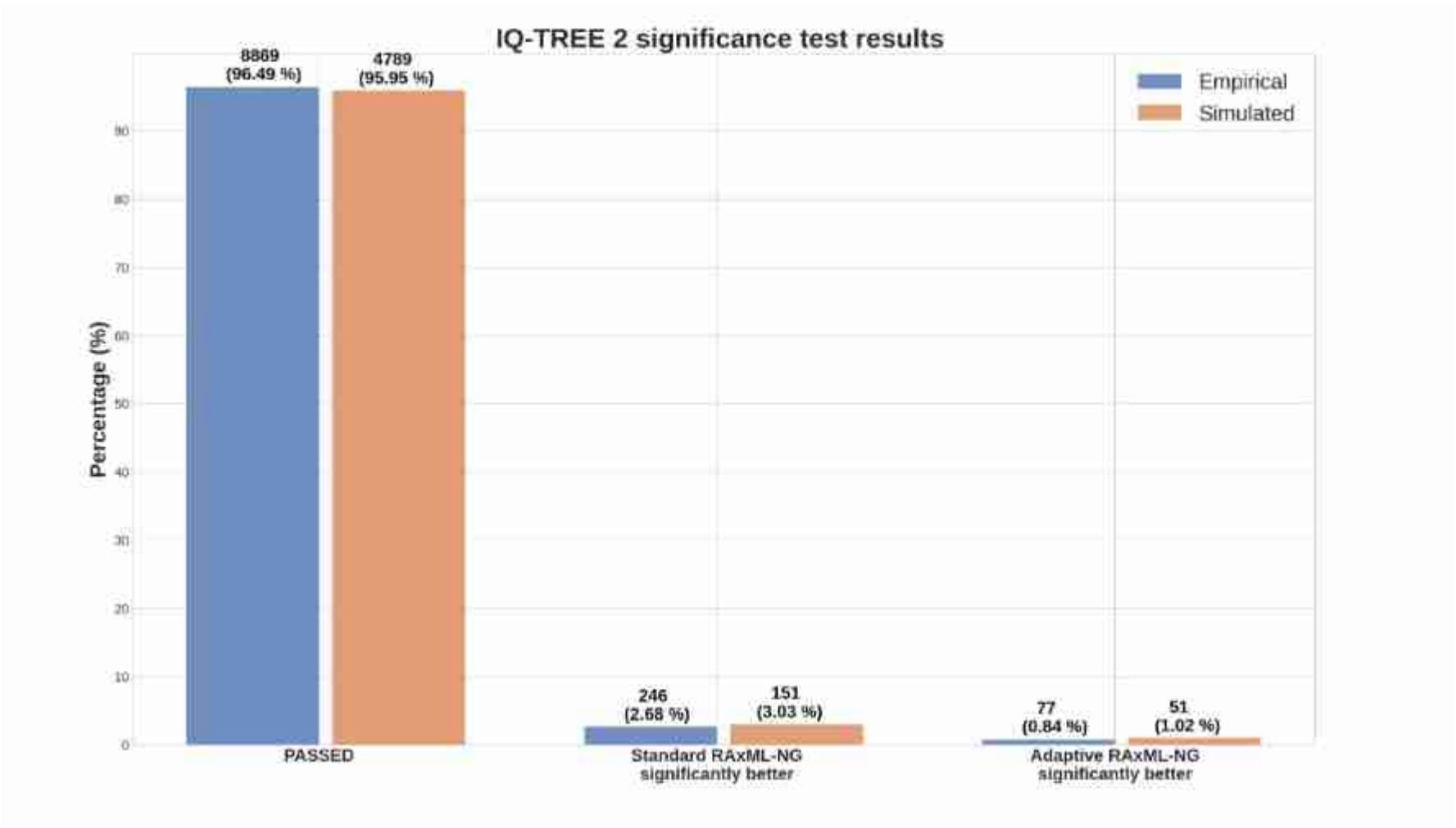
- 10K empirical MSAs from TreeBase
  → 9192 MSAs after filtering



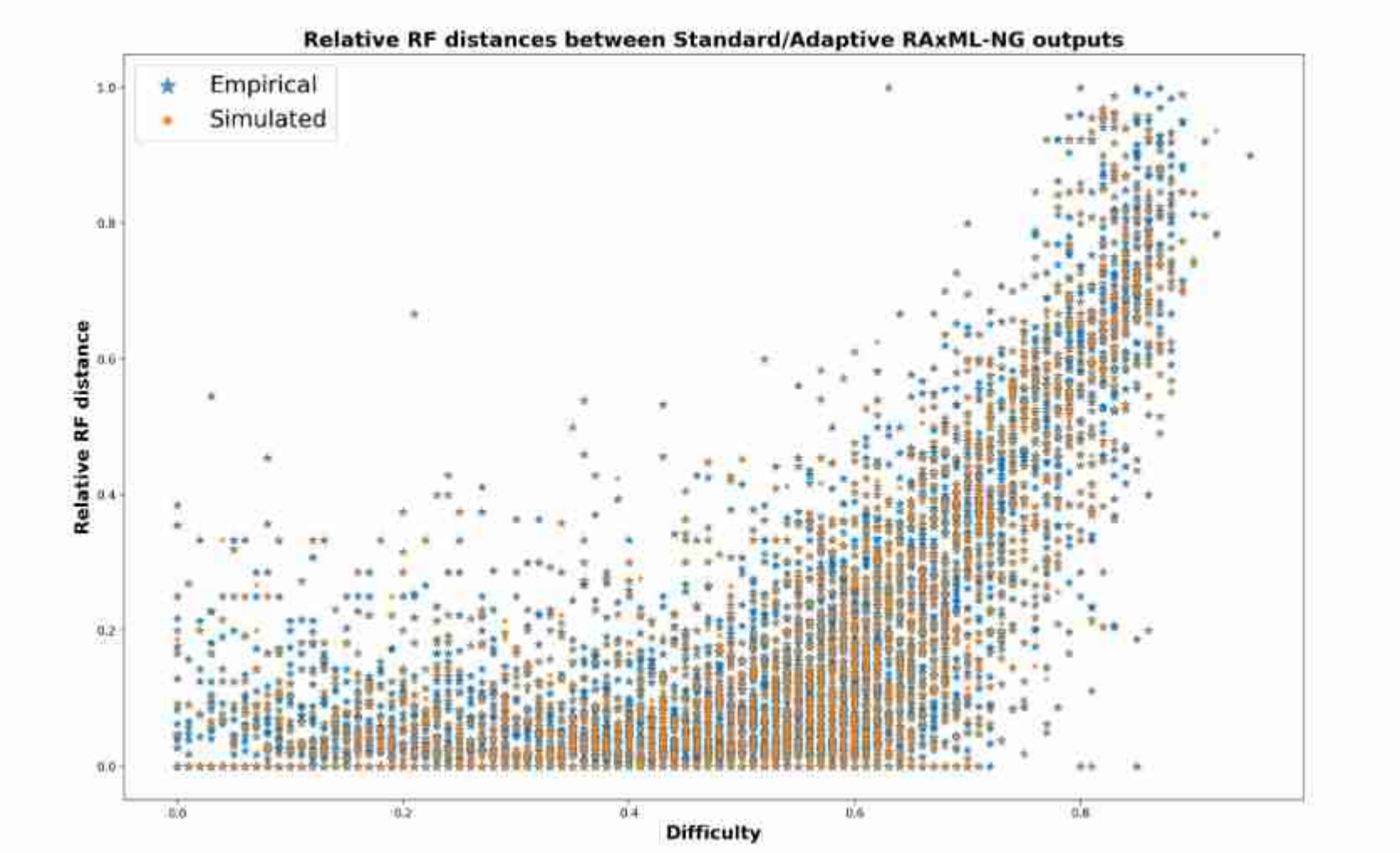
Difficulty Score Distribution



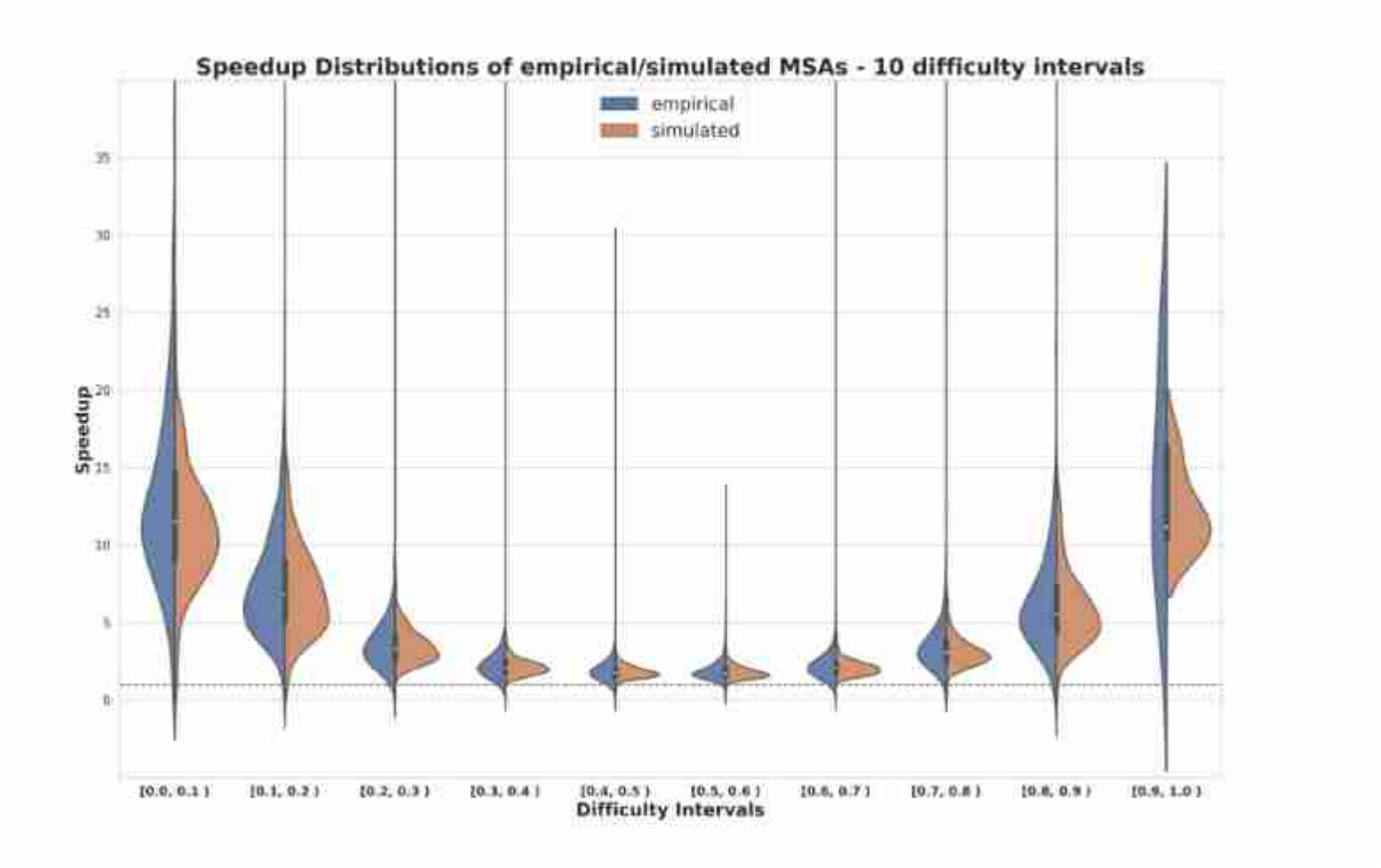
Significance Tests



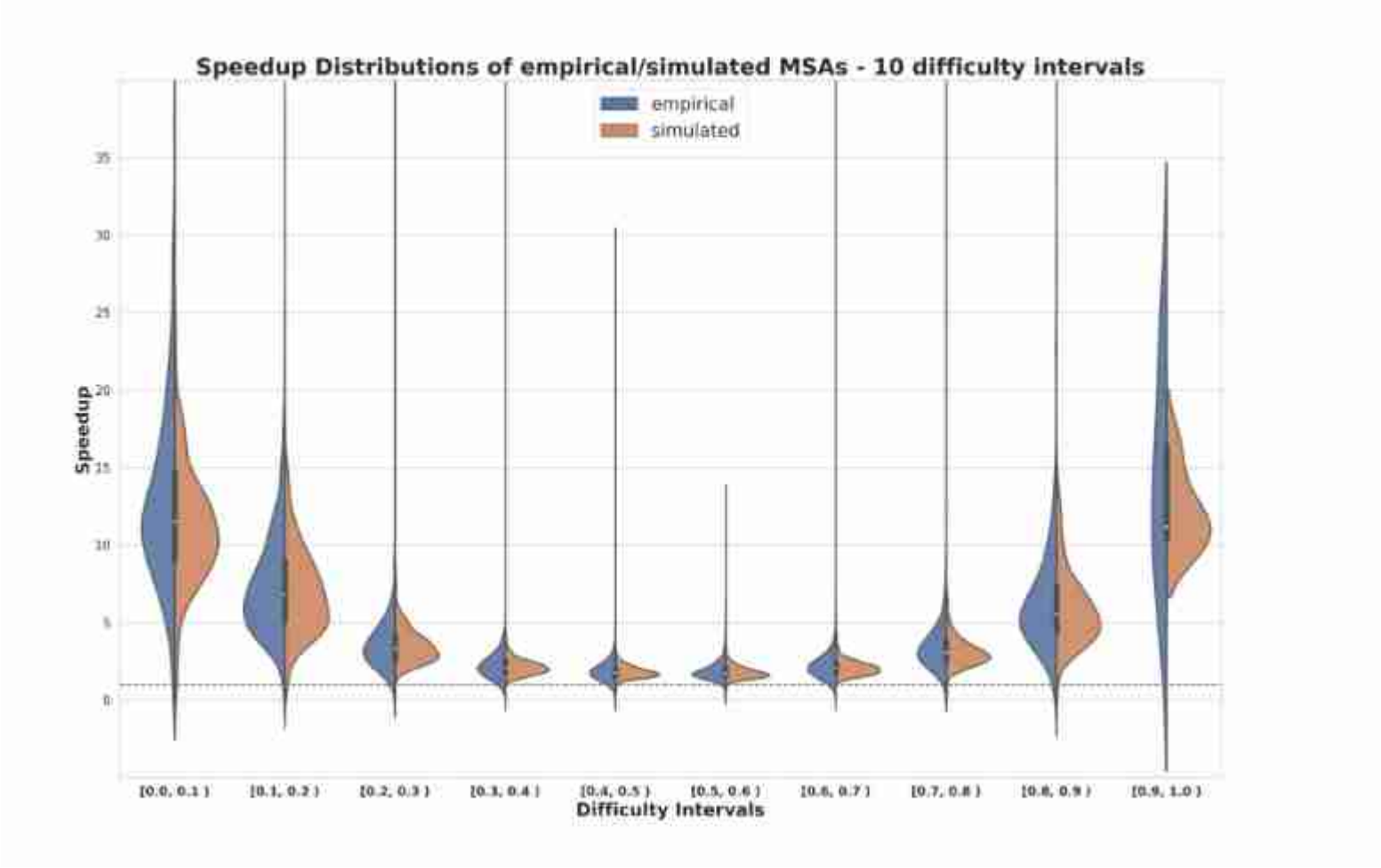
Distances between trees



Speedups


Speedup Distributions of empirical/simulated MSAs - 10 difficulty intervals
empirical
simulated
Speedup
35
30
25
20
15
10
5
0
[0.0, 0.1)
[0.1, 0.2)
[0.2, 0.3)
[0.3, 0.4)
[0.4, 0.5)
[0.5, 0.6)
[0.6, 0.7)
[0.7, 0.8)
[0.8, 0.9)
[0.9, 1.0)
Difficulty Intervals

Speedups

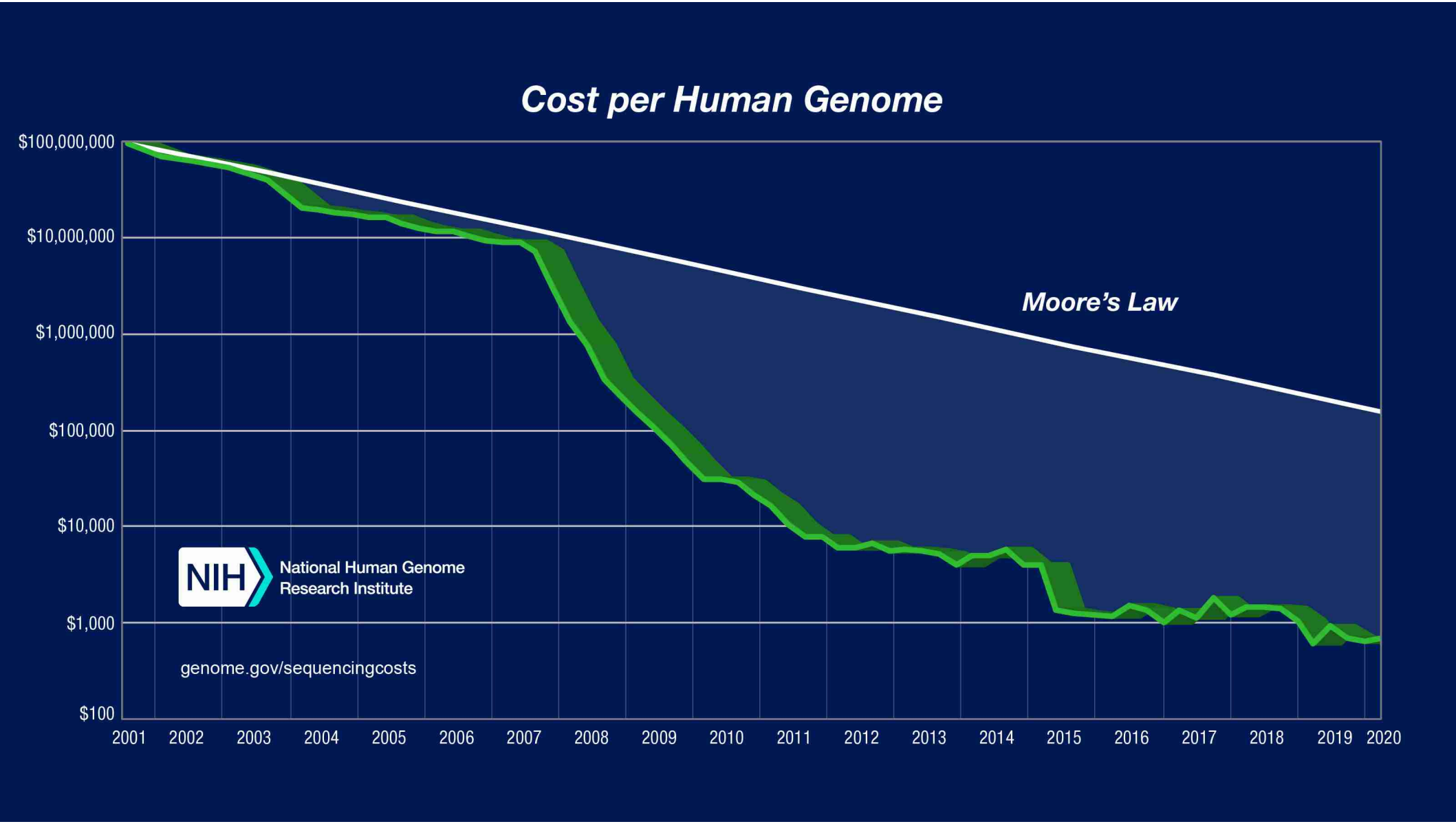


Overall accumulated speedup: approx. 3 on empirical data

Outline

- Our Approach to Bioinformatics
- Introduction to Phylogenetic Inference
- Sources of Uncertainty
- Phylogenetic Difficulty
- **Other stuff we are working on**

Scalability



Single Cell Evolution

- Reconstructing the evolution, e.g., of cancer cells in a single patient is challenging
  - Noisy data
  - Erroneous data
  - Little signal
  - Few & simplistic models

Eleven grand challenges in single-cell data science

David Lähnemann, Johannes Köster, [...] Alexander Schönhuth

Genome Biology **21**, Article number: 31 (2020) | Cite this article

32k Accesses | **16** Citations | **281** Altmetric | Metrics

New Results

Comments (2)

CellPhy: accurate and fast probabilistic inference of single-cell phylogenies from scDNA-seq data

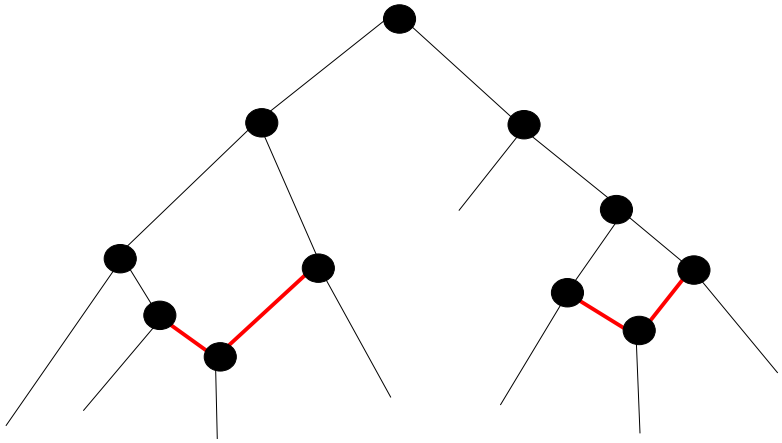
Alexey Kozlov, Joao Alves, Alexandros Stamatakis, David Posada

doi: https://doi.org/10.1101/2020.07.31.230292

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

Phylogenetic Networks

- Evolution does not need to occur in a tree-like manner due to recombination events
- We can model this via so-called phylogenetic networks



Phylogenetic Networks

- Evolution does not need to occur in a tree-like manner due to recombination events
- We can model this via so-called phylogenetic networks
- The likelihood of such a network is substantially more difficult to compute than on a tree

→ computational challenges

JOURNAL ARTICLE

NetRAX: accurate and fast maximum likelihood phylogenetic network inference

Sarah Lutteropp, Céline Scornavacca, Alexey M Kozlov, Benoit Morel, Alexandros Stamatakis

Bioinformatics, Volume 38, Issue 15, August 2022, Pages 3725–3733,
https://doi.org/10.1093/bioinformatics/btac396
Published: 17 June 2022 **Article history**

Gene Tree Species Tree Reconciliation

- There are other phenomena that complicate evolution
  - Gene loss
  - Gene transfer
  - Gene duplication

    → gene tree ≠ species tree
- Infer & correct trees under a joint likelihood model comprising the phylogenetic likelihood and a reconciliation likelihood model

GeneRax

- First full and efficient Maximum Likelihood implementation to infer gene family trees using a given rooted species tree under a joint phylogenetic & reconciliation likelihood model

GeneRax: A Tool for Species-Tree-Aware Maximum Likelihood-Based Gene Family Tree Inference under Gene Duplication, Transfer, and Loss

Benoit Morel, Alexey M Kozlov, Alexandros Stamatakis, Gergely J Szöllősi

Molecular Biology and Evolution, Volume 37, Issue 9, September 2020, Pages 2763–2774, https://doi.org/10.1093/molbev/msaa141
Published: 05 June 2020

SpeciesRax

- **Goal:** Simultaneously infer the gene family trees **and** the species tree under a joint phylogenetic/reconciliation likelihood model

JOURNAL ARTICLE

SpeciesRax: A Tool for Maximum Likelihood Species Tree Inference from Gene Family Trees under Duplication, Transfer, and Loss

Benoit Morel, Paul Schade, Sarah Lutteropp, Tom A Williams, Gergely J Szöllősi, Alexandros Stamatakis

Molecular Biology and Evolution, Volume 39, Issue 2, February 2022, msab365,
https://doi.org/10.1093/molbev/msab365
Published: 11 January 2022

Parallel Fault Tolerance

- Parallel computations on thousands of cores are likely to fail due to failing hardware components
- This applies to tightly coupled massively parallel codes in general and to `RAxML-NG` in particular
- **Goal:** Devise generic and `RAxML-NG` specific strategies for fault tolerance of massively parallel codes

JOURNAL ARTICLE

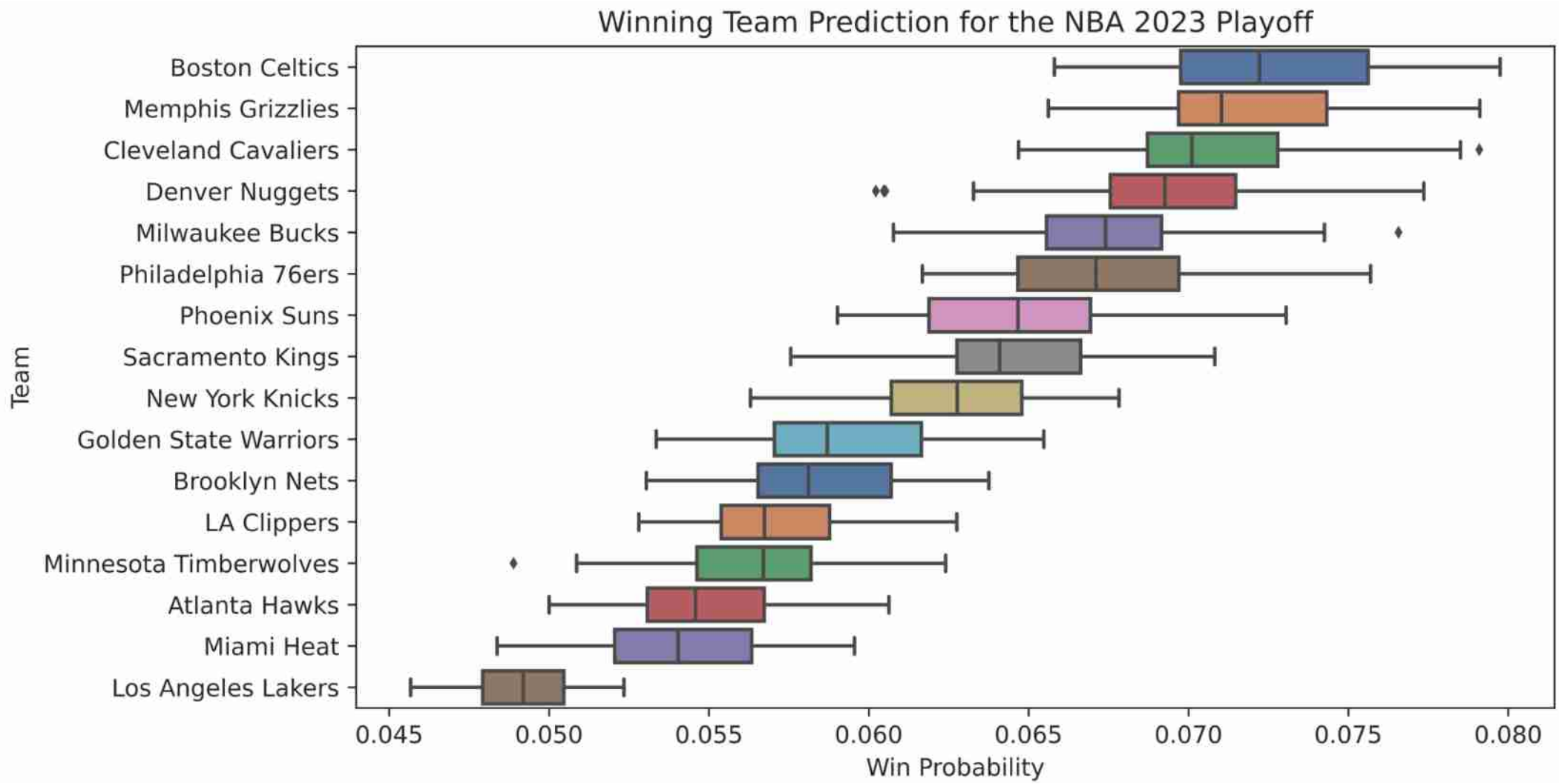
Exploring parallel MPI fault tolerance mechanisms for phylogenetic inference with RAxML-NG

Lukas Hübner, Alexey M Kozlov, Demian Hespe, Peter Sanders, Alexandros Stamatakis

Bioinformatics, Volume 37, Issue 22, 15 November 2021, Pages 4056–4063,
https://doi.org/10.1093/bioinformatics/btab399

Published: 26 May 2021 **Article history**

Tournament Prediction



Software Quality Assessment

- `SoftWipe` tool for automatic scientific software quality assessment (`C` and `C++`)

Article | Open Access | Published: 11 May 2021

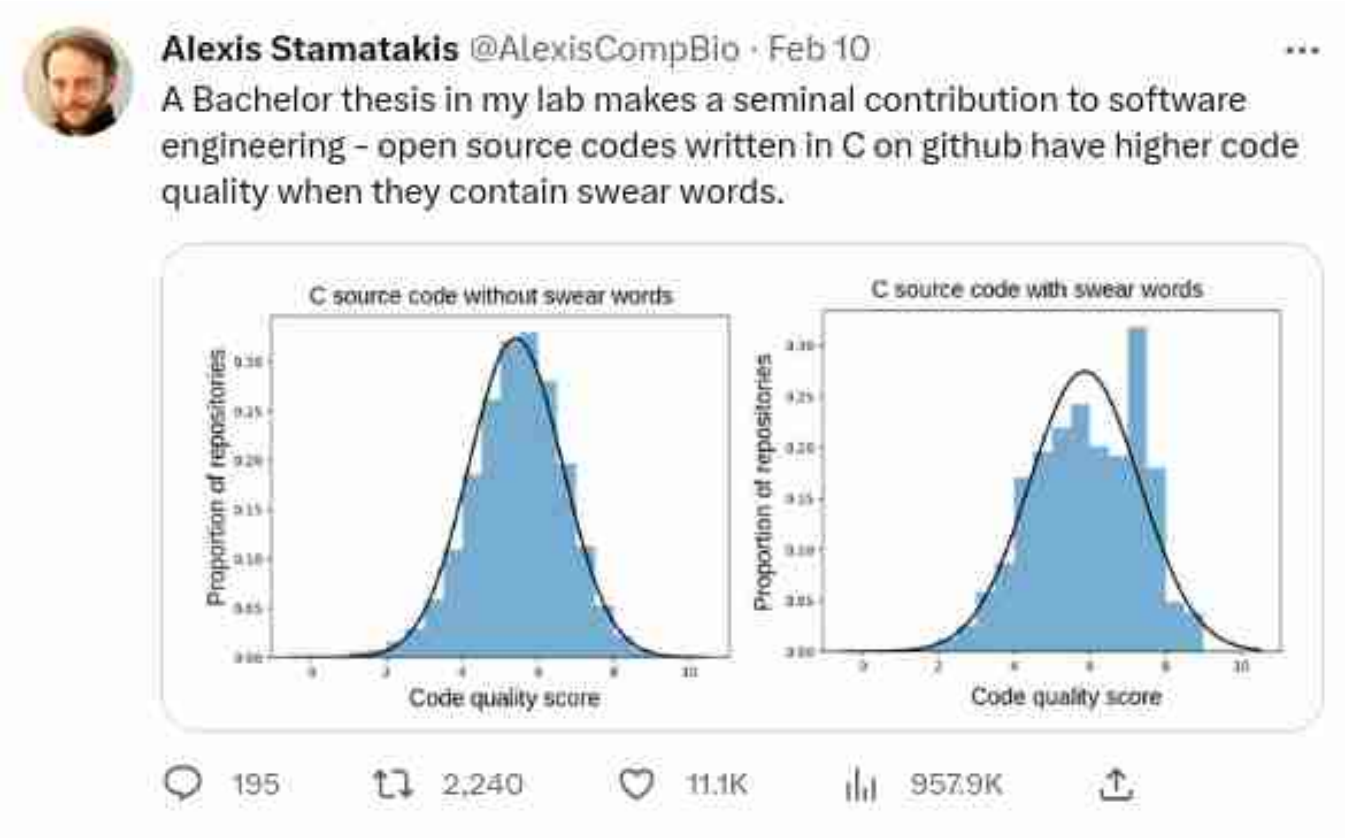
The SoftWipe tool and benchmark for assessing coding standards adherence of scientific software

Adrian Zapletal, Dimitri Höhler, Carsten Sinz & Alexandros Stamatakis

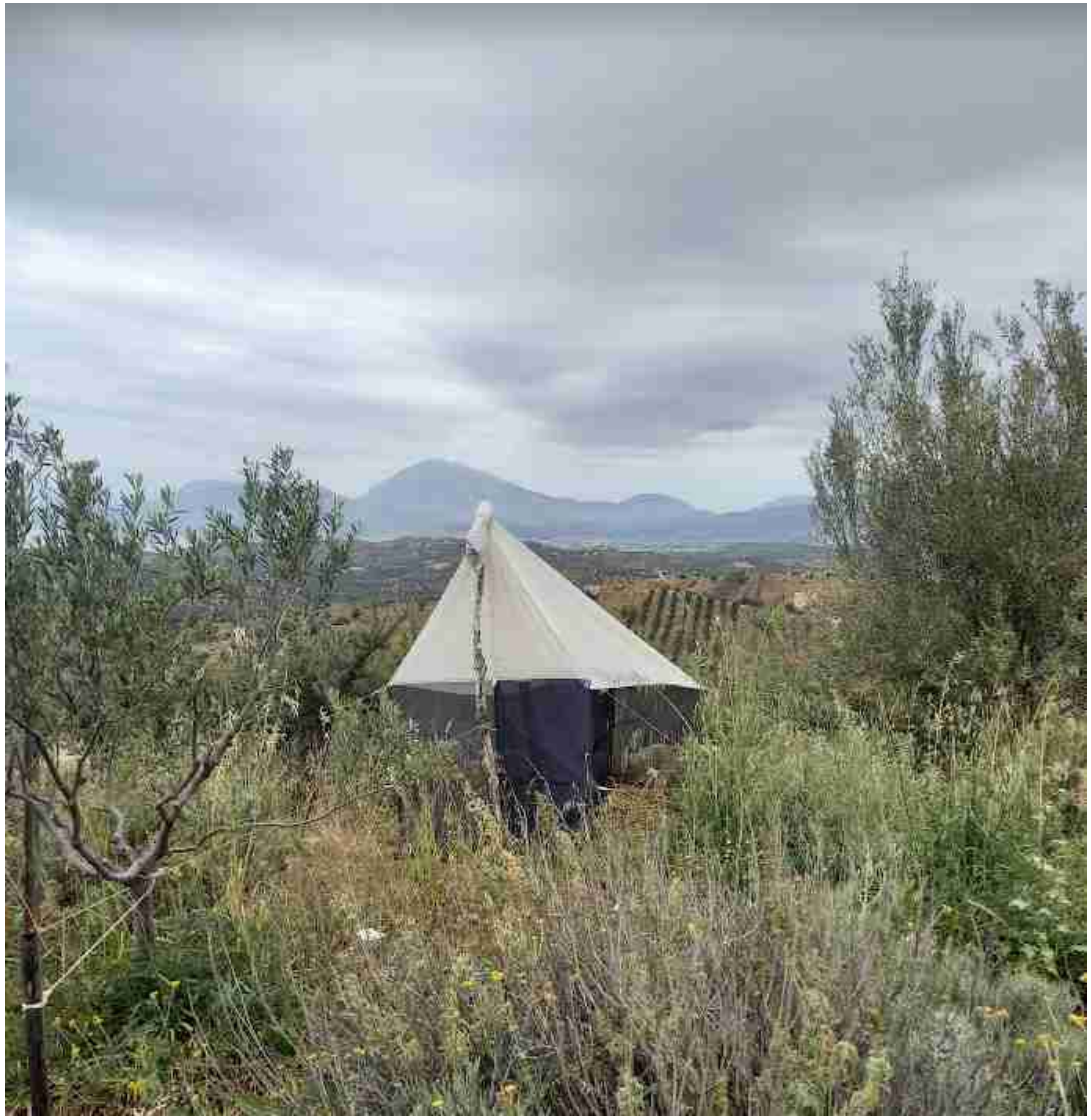
Scientific Reports **11**, Article number: 10015 (2021) | Cite this article

4270 Accesses | **1** Citations | **115** Altmetric | Metrics

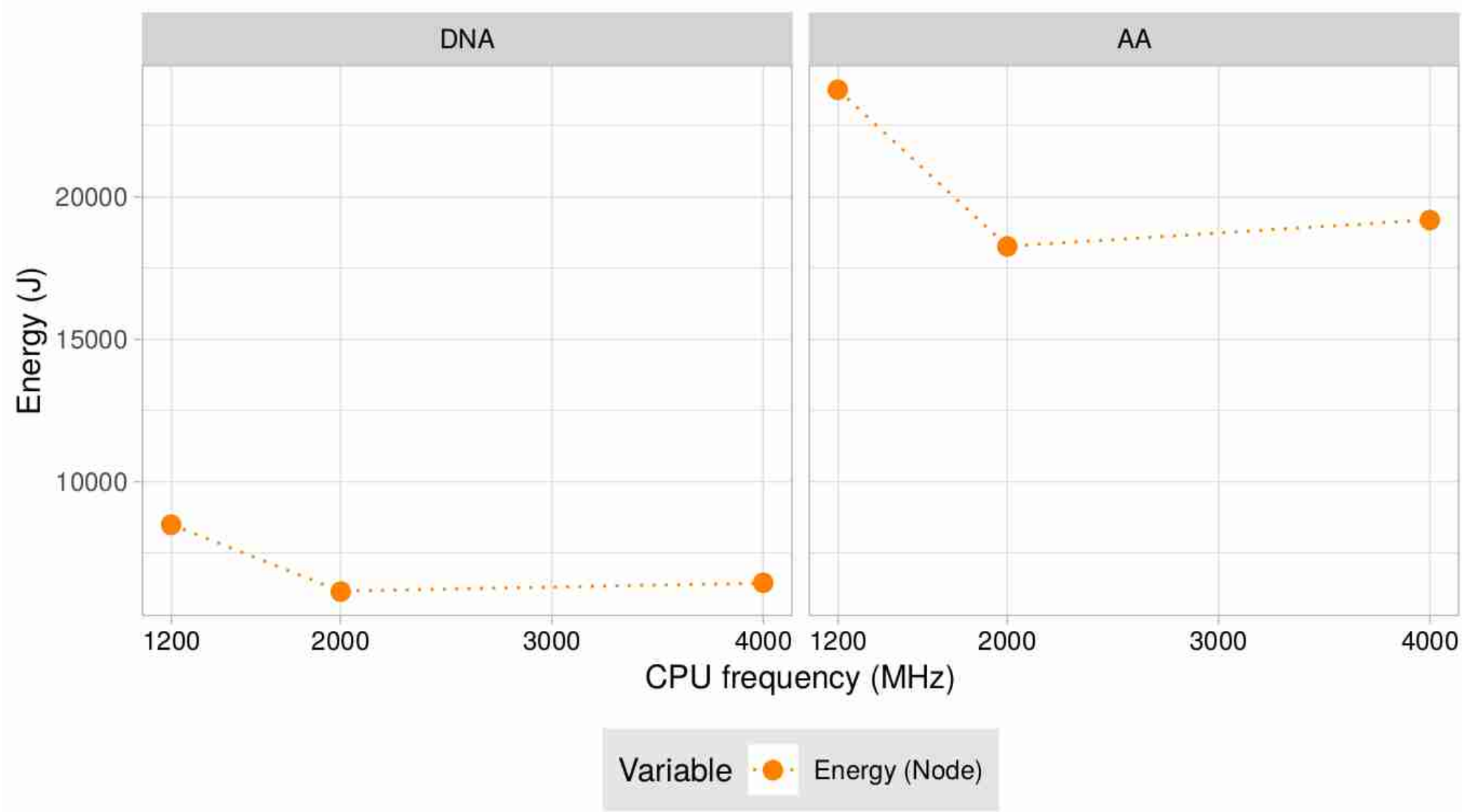
Empirical Software Engineering with `SoftWipe`



Biological Field Work



Energy Efficiency



Ancient DNA

- Better tools for ancient DNA analyses
- Classic aDNA data analyses

Current Biology

CellPress

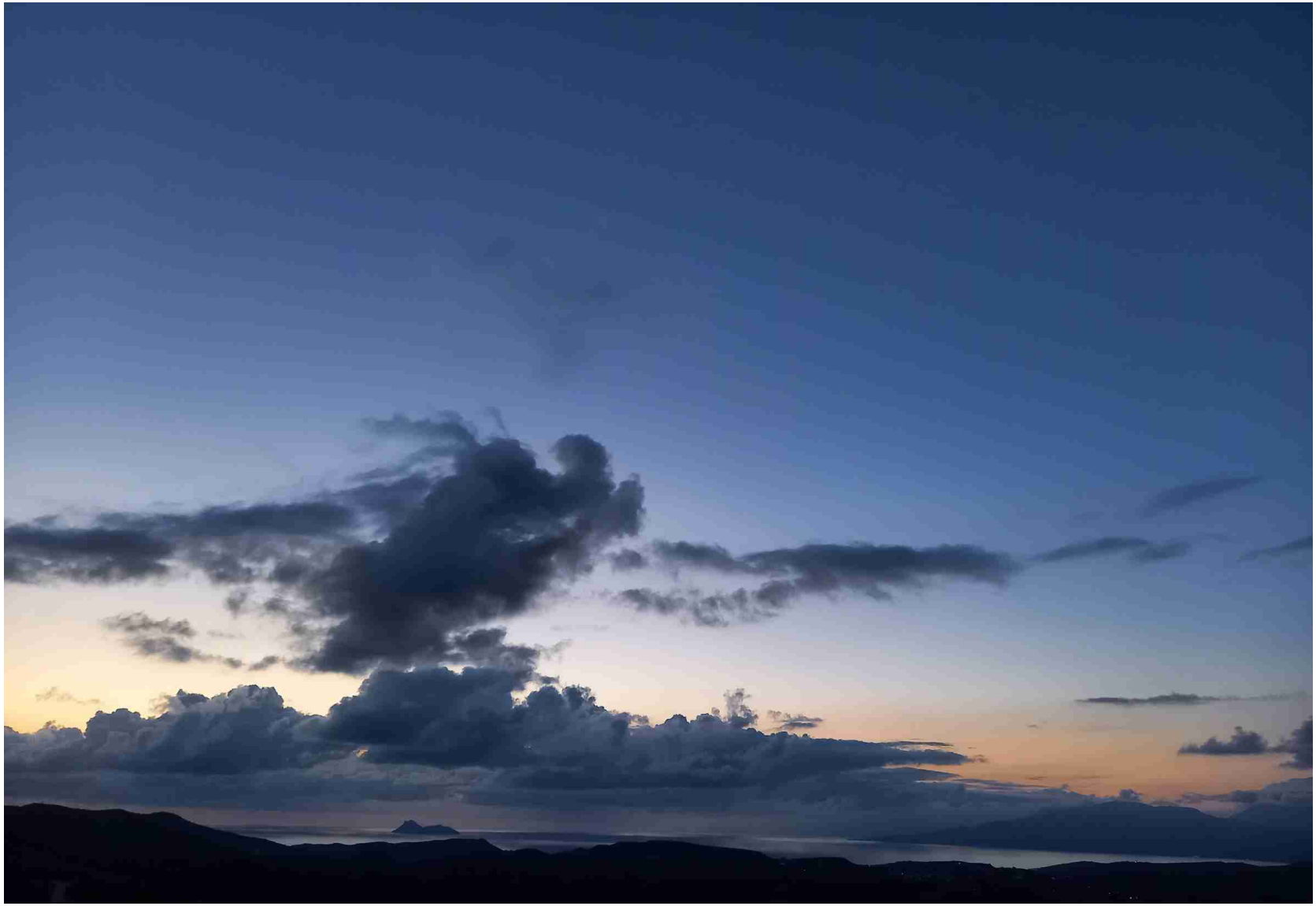
Volume 33, Issue 1, 9 January 2023, Pages 41-57.e15

Article

Spatial and temporal heterogeneity in human mobility patterns in Holocene Southwest Asia and the East Mediterranean

Dilek Koptekin [1 2 45], Eren Yüncü [2], Ricardo Rodríguez-Varela [3 4 42], N. Ezgi Altınışık [5 42], Nikolaos Psonis [6 42], Natalia Kashuba [7], Sevgi Yorulmaz [2], Robert George [3 8], Duygu Deniz Kazancı [2 5], Damla Kaptan [2], Kanat Gürün [2], Kıvılcım Başak Vural [2], Hasan Can Gemici [9], Despoina Vassou [6], Evangelia Daskalaki [4], Cansu Karamurat [9], Vendela K. Lagerholm [3 4], Ömür Dilek Erdal [10], Emrah Kırdök [11], Aurelio Marangoni [3]... Mehmet Somel [1 2 43 44]

Thank you for your attention



Pipeline Complexity

Project Complexity

the good old days

Sequence

```
T1 ACGT
T2 ACC
T3 ACGG
T4 AAGC
```

single gene &
few species

Project Complexity

the good old days

Sequence → Align

```
T1 ACGT
T2 ACC
T3 ACGG
T4 AAGC
```

```
T1 ACGT
T2 ACC-
T3 ACGG
T4 AAGC
```

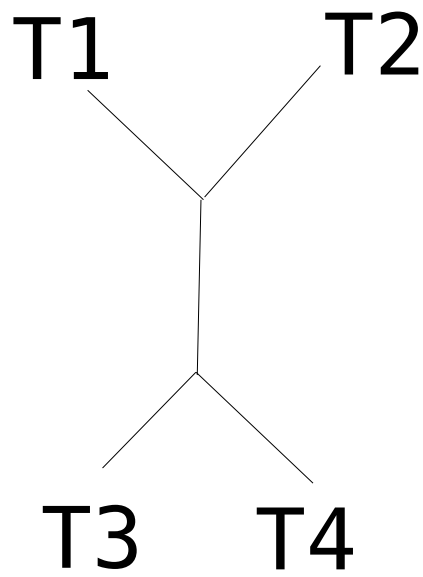
Project Complexity

the good old days

Sequence → Align → Infer Tree

```
T1 ACGT
T2 ACC
T3 ACGG
T4 AAGC
```

```
T1 ACGT
T2 ACC-
T3 ACGG
T4 AAGC
```



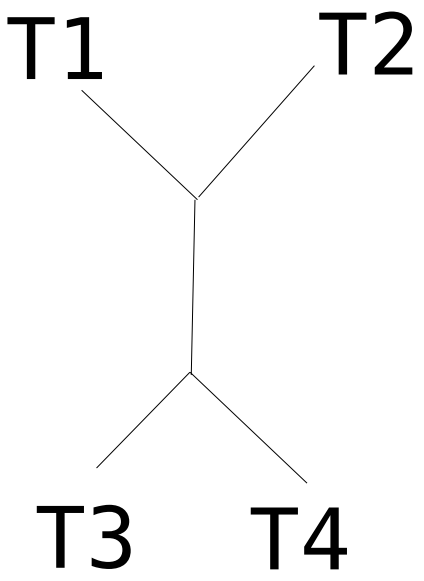
Project Complexity

the good old days

Sequence → Align → Infer Tree → Publish

```
T1 ACGT
T2 ACC
T3 ACGG
T4 AAGC
```

```
T1 ACGT
T2 ACC-
T3 ACGG
T4 AAGC
```

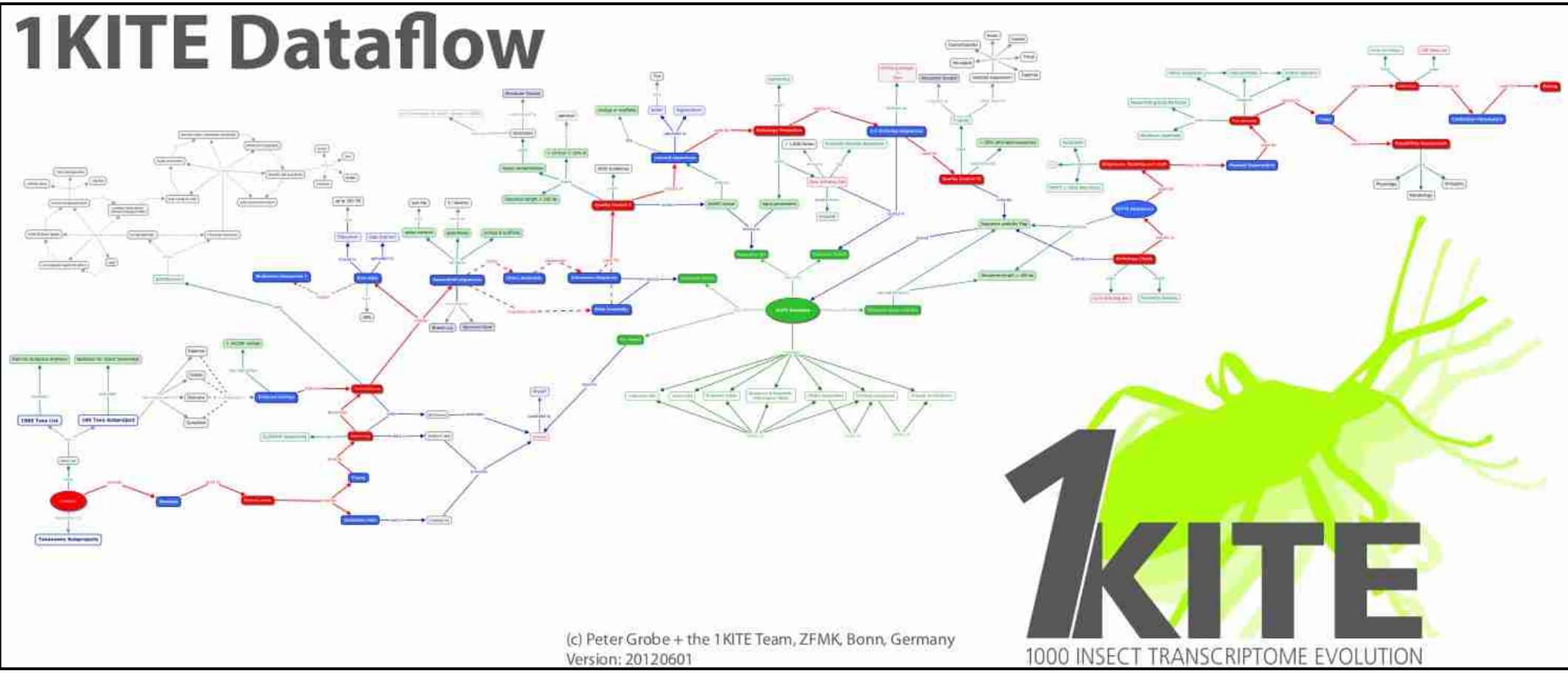


Project Complexity Today

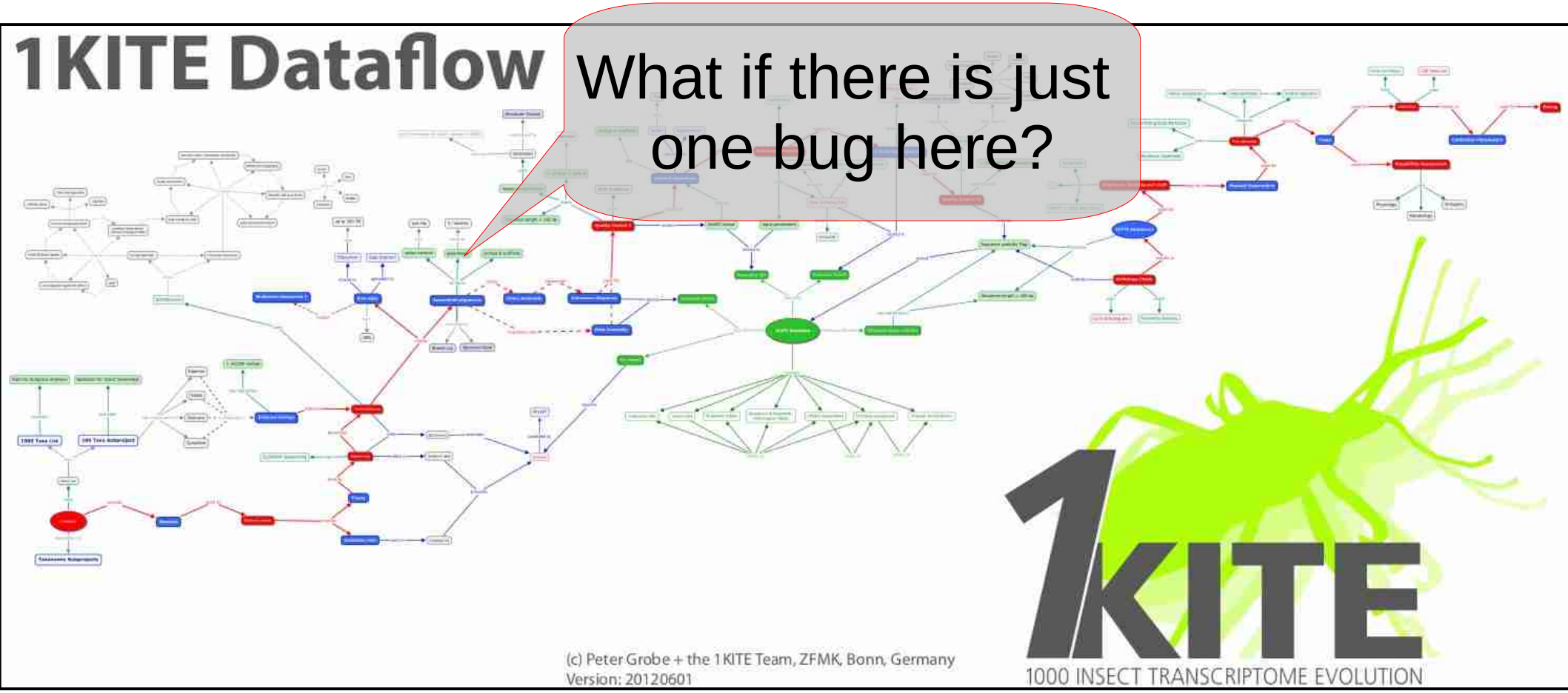
150 insect transcriptomes

50 bird genomes

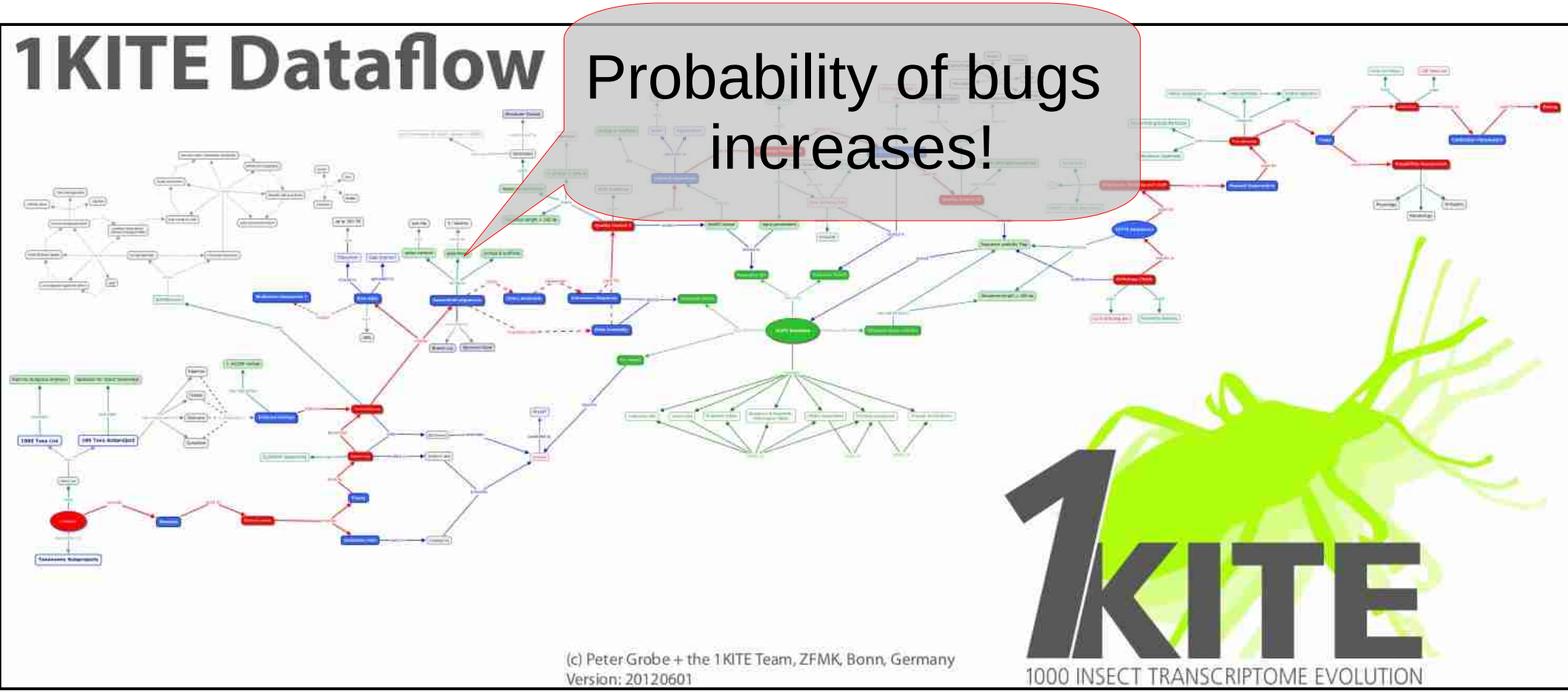
Project Complexity Today



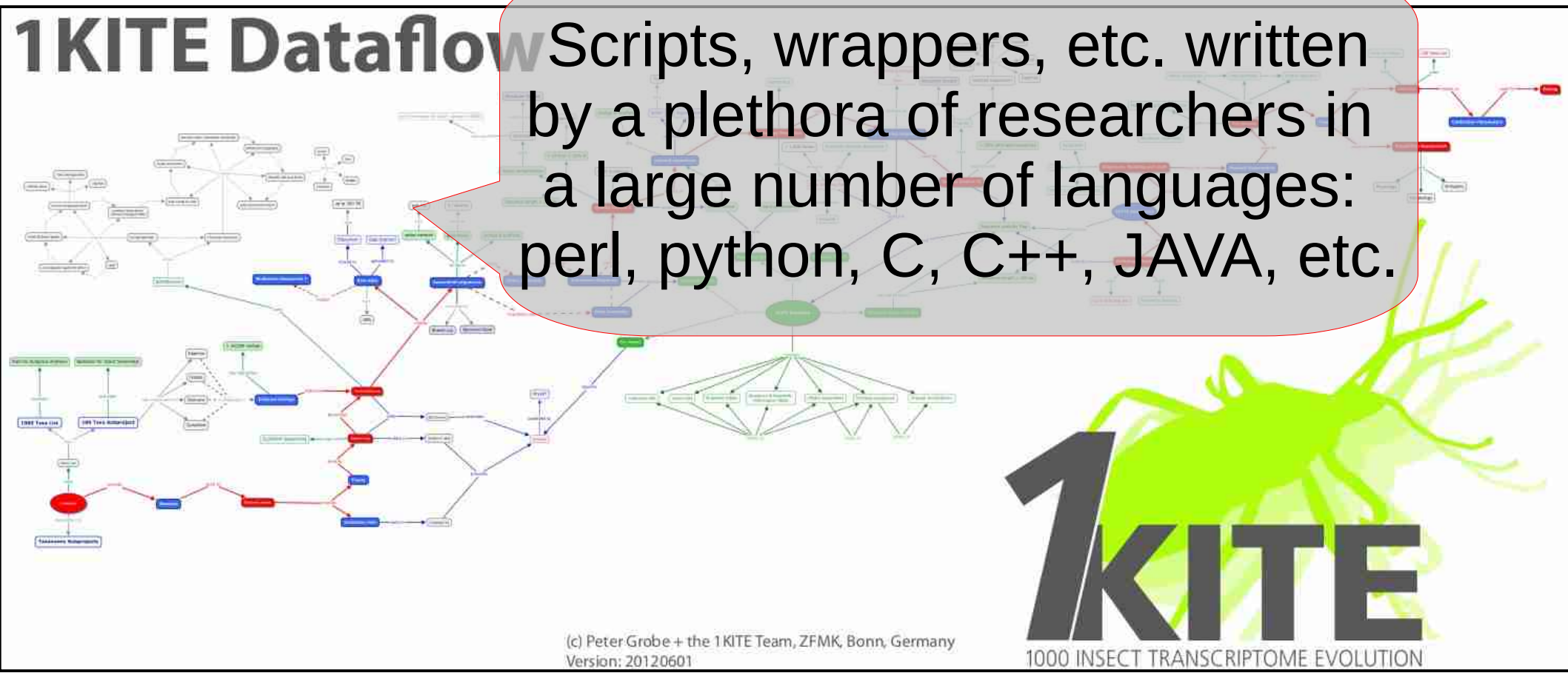
Project Complexity Today



Project Complexity Today



Project Complexity Today



Scripts, wrappers, etc. written by a plethora of researchers in a large number of languages: perl, python, C, C++, JAVA, etc.

The 'crappy' software project

- Analyzed *15* widely-used evolutionary biology tools ≈ *65,000* citations
- Analyses performed
  - Compiled with `gcc` and `clang` with all warnings enabled
  - Memory check with `valgrind`
  - Checked if assertions are used via `assert()`
  - Analyzed degree of code duplication
- **Caution:** “bad” quality does not induce that a tool is faulty, but the probability of it being faulty is higher!

The State of Software for Evolutionary Biology

Diego Darriba, Tomáš Flouri, Alexandros Stamatakis

Molecular Biology and Evolution, Volume 35, Issue 5, May 2018, Pages 1037–1046, https://doi.org/10.1093/molbev/msy014

Published: 29 January 2018

SoftWipe

- Discussion with Science Journalist - *“Can this process be automated?”*
- Development of SoftWipe - An automated tool and benchmark for **relative** quality ranking of scientific software
- Ranking of *51* open source tools written in C or C++ from a wide range of research areas
  - Astrophysics
  - Computer Science
  - Bioinformatics

New Results

Comment on this paper

SoftWipe - a tool and benchmark to assess scientific software quality

Adrian Zapletal, Dimitri Hoehler, Carsten Sinz, Alexandros Stamatakis

doi: https://doi.org/10.1101/2020.10.07.330621

SoftWipe Benchmark

program name	absolute score	relative score
genesis	8.6	8.8
hyperphylo	8.6	8.6
kahypar	8.4	8.5
candy-kingdom	8.2	8.2
bindash-1.0	8.0	7.9
fastspar	7.8	7.9
repeatscounter	7.5	7.7
axe-0.3.3	7.5	7.5
virulign-1.0.1	7.4	7.4
naf-1.1.0/unnaf	7.4	7.5
naf-1.1.0/ennaf	7.4	7.4
ExpansionHunter	7.3	7.5
glucose-3-drup	7.1	7.0
raxml-ng	7.0	7.0
dawg	6.8	6.9
ntEdit-1.2.3	6.4	6.2
defor	6.3	6.4
swarm	6.2	6.2
lemon	6.1	6.0
treerecs	6.1	6.1
IQ-TREE-2.0-rc1	6.1	5.7
BGSA_CPU-1.0	5.9	5.4
emeraLD	5.8	5.5
dr_sasa_n	5.7	6.0
copmem-0.2	5.7	5.7
samtools	5.6	5.6
seq-gen	5.6	5.6
dna-nn-0.1	5.3	5.2
sf	5.2	5.2
cryfa-18.06	5.1	5.1
ngsLD	5.1	5.0
HLA-LA	4.9	4.5
iqtree1.6.10	4.9	4.9
vsearch	4.6	4.6
prank	4.6	4.5
prequal	4.5	4.4
minimap	4.5	4.4
phyml	4.4	4.4
clustal	4.2	4.3
mrbayes	4.1	4.1
tcoffee	4.1	4.2
gadget	4.1	4.0
crisflash	4.0	4.0
PopLDdecay	3.8	3.8
cellcoal	3.8	3.6
bpp	3.8	3.6
ms	3.7	3.7
mafft	3.3	3.1
athena	2.9	2.8
covid-sim-0.13.0	2.5	2.4
indelible	1.4	1.0

SoftWipe Benchmark

program name	absolute score	relative score
genesis	8.6	8.8
hyperphylo	8.6	8.6
kahypar	8.4	8.5
candy-kingdom	8.2	8.2
bindash-1.0	8.0	7.9
fastspar	7.8	7.9
repeatscounter	7.5	7.7
axe-0.3.3	7.5	7.5
virulign-1.0.1	7.4	7.4
naf-1.1.0/unnaf	7.4	7.5
naf-1.1.0/ennaf	7.4	7.4
ExpansionHunter	7.3	7.5
glucose-3-drup	7.1	7.0
raxml-ng	7.0	7.0
dawg	6.8	6.9
ntEdit-1.2.3	6.4	6.2
defor	6.3	6.4
swarm	6.2	6.2
lemon	6.1	6.0
treerecs	6.1	6.1
IQ-TREE-2.0-rc1	6.1	5.7
BGSA_CPU-1.0	5.9	5.4
emeraLD	5.8	5.5
dr_sasa_n	5.7	6.0
copmem-0.2	5.7	5.7
samtools	5.6	5.6
seq-gen	5.6	5.6
dna-nn-0.1	5.3	5.2
sf	5.2	5.2
cryfa-18.06	5.1	5.1
ngsLD	5.1	5.0
HLA-LA	4.9	4.5
iqtree1.6.10	4.9	4.9
vsearch	4.6	4.6
prank	4.6	4.5
prequal	4.5	4.4
minimap	4.5	4.4
phyml	4.4	4.4
clustal	4.2	4.3
mrbayes	4.1	4.1
tcoffee	4.1	4.2
gadget	4.1	4.0
crisflash	4.0	4.0
PopLDdecay	3.8	3.8
cellcoal	3.8	3.6
bpp	3.8	3.6
ms	3.7	3.7
mafft	3.3	3.1
athena	2.9	2.8
covid-sim-0.13.0	2.5	2.4
indelible	1.4	1.0

Does not change over time as more tools are added → can easily be referenced

SoftWipe Benchmark

program name	absolute score	relative score
genesis	8.6	8.8
hyperphylo	8.6	8.6
kahypar	8.4	8.5
candy-kingdom	8.2	8.2
bindash-1.0	8.0	7.9
fastspar	7.8	7.9
repeatscounter	7.5	7.7
axe-0.3.3	7.5	7.5
virulign-1.0.1	7.4	7.4
naf-1.1.0/unnaf	7.4	7.5
naf-1.1.0/ennaf	7.4	7.4
ExpansionHunter	7.3	7.5
glucose-3-drup	7.1	7.0
raxml-ng	7.0	7.0
dawg	6.8	6.9
ntEdit-1.2.3	6.4	6.2
defor	6.3	6.4
swarm	6.2	6.2
lemon	6.1	6.0
treerecs	6.1	6.1
IQ-TREE-2.0-rc1	6.1	5.7
BGSA_CPU-1.0	5.9	5.4
emeraLD	5.8	5.5
dr_sasa_n	5.7	6.0
copmem-0.2	5.7	5.7
samtools	5.6	5.6
seq-gen	5.6	5.6
dna-nn-0.1	5.3	5.2
sf	5.2	5.2
cryfa-18.06	5.1	5.1
ngsLD	5.1	5.0
HLA-LA	4.9	4.5
iqtree1.6.10	4.9	4.9
vsearch	4.6	4.6
prank	4.6	4.5
prequal	4.5	4.4
minimap	4.5	4.4
phyml	4.4	4.4
clustal	4.2	4.3
mrbayes	4.1	4.1
tcoffee	4.1	4.2
gadget	4.1	4.0
crisflash	4.0	4.0
PopLDdecay	3.8	3.8
cellcoal	3.8	3.6
bpp	3.8	3.6
ms	3.7	3.7
mafft	3.3	3.1
athena	2.9	2.8
covid-sim-0.13.0	2.5	2.4
indelible	1.4	1.0

Does change over time as more tools are added →
Difficult to be referenced

SoftWipe Benchmark

program name	absolute score	relative score
genesis	8.6	8.8
hyperphylo	8.6	8.6
kahypar	8.4	8.5
candy-kingdom	8.2	8.2
bindash-1.0	8.0	7.9
fastspar	7.8	7.9
repeatscounter	7.5	7.7
axe-0.3.3	7.5	7.5
virulign-1.0.1	7.4	7.4
naf-1.1.0/unnaf	7.4	7.5
naf-1.1.0/ennaf	7.4	7.4
ExpansionHunter	7.3	7.5
glucose-3-drup	7.1	7.0
raxml-ng	7.0	7.0
dawg	6.8	6.9
ntEdit-1.2.3	6.4	6.2
defor	6.3	6.4
swarm	6.2	6.2
lemon	6.1	6.0
treerecs	6.1	6.1
IQ-TREE-2.0-rc1	6.1	5.7
BGSA_CPU-1.0	5.9	5.4
emeraLD	5.8	5.5
dr_sasa_n	5.7	6.0
copmem-0.2	5.7	5.7
samtools	5.6	5.6
seq-gen	5.6	5.6
dna-nn-0.1	5.3	5.2
sf	5.2	5.2
cryfa-18.06	5.1	5.1
ngsLD	5.1	5.0
HLA-LA	4.9	4.5
iqtree1.6.10	4.9	4.9
vsearch	4.6	4.6
prank	4.6	4.5
prequal	4.5	4.4
minimap	4.5	4.4
phyml	4.4	4.4
clustal	4.2	4.3
mrbayes	4.1	4.1
tcoffee	4.1	4.2
gadget	4.1	4.0
crisflash	4.0	4.0
PopLDdecay	3.8	3.8
cellcoal	3.8	3.6
bpp	3.8	3.6
ms	3.7	3.7
mafft	3.3	3.1
athena	2.9	2.8
covid-sim-0.13.0	2.5	2.4
indelible	1.4	1.0

Written by computer scientists

SoftWipe Benchmark

My lab

program name	absolute score	relative score
genesis	8.6	8.8
hyperphylo	8.6	8.6
kahypar	8.4	8.5
candy-kingdom	8.2	8.2
bindash-1.0	8.0	7.9
fastspar	7.8	7.9
repeatscounter	7.5	7.7
axe-0.3.3	7.5	7.5
virulign-1.0.1	7.4	7.4
naf-1.1.0/unnaf	7.4	7.5
naf-1.1.0/ennaf	7.4	7.4
ExpansionHunter	7.3	7.5
glucose-3-drup	7.1	7.0
raxml-ng	7.0	7.0
dawg	6.8	6.9
ntEdit-1.2.3	6.4	6.2
defor	6.3	6.4
swarm	6.2	6.2
lemon	6.1	6.0
treerecs	6.1	6.1
IQ-TREE-2.0-rc1	6.1	5.7
BGSA_CPU-1.0	5.9	5.4
emeraLD	5.8	5.5
dr_sasa_n	5.7	6.0
copmem-0.2	5.7	5.7
samtools	5.6	5.6
seq-gen	5.6	5.6
dna-nn-0.1	5.3	5.2
sf	5.2	5.2
cryfa-18.06	5.1	5.1
ngsLD	5.1	5.0
HLA-LA	4.9	4.5
iqtree1.6.10	4.9	4.9
vsearch	4.6	4.6
prank	4.6	4.5
prequal	4.5	4.4
minimap	4.5	4.4
phyml	4.4	4.4
clustal	4.2	4.3
mrbayes	4.1	4.1
tcoffee	4.1	4.2
gadget	4.1	4.0
crisflash	4.0	4.0
PopLDdecay	3.8	3.8
cellcoal	3.8	3.6
bpp	3.8	3.6
ms	3.7	3.7
mafft	3.3	3.1
athena	2.9	2.8
covid-sim-0.13.0	2.5	2.4
indelible	1.4	1.0

SoftWipe Benchmark

program name	absolute score	relative score
genesis	8.6	8.8
hyperphylo	8.6	8.6
kahypar	8.4	8.5
candy-kingdom	8.2	8.2
bindash-1.0	8.0	7.9
fastspar	7.8	7.9
repeatscounter	7.5	7.7
axe-0.3.3	7.5	7.5
virulign-1.0.1	7.4	7.4
naf-1.1.0/unnaf	7.4	7.5
naf-1.1.0/ennaf	7.4	7.4
ExpansionHunter	7.3	7.5
glucose-3-drup	7.1	7.0
raxml-ng	7.0	7.0
dawg	6.8	6.9
ntEdit-1.2.3	6.4	6.2
defor	6.3	6.4
swarm	6.2	6.2
lemon	6.1	6.0
treerecs	6.1	6.1
IQ-TREE-2.0-rc1	6.1	5.7
BGSA_CPU-1.0	5.9	5.4
emeraLD	5.8	5.5
dr_sasa_n	5.7	6.0
copmem-0.2	5.7	5.7
samtools	5.6	5.6
seq-gen	5.6	5.6
dna-nn-0.1	5.3	5.2
sf	5.2	5.2
cryfa-18.06	5.1	5.1
ngsLD	5.1	5.0
HLA-LA	4.9	4.5
iqtree1.6.10	4.9	4.9
vsearch	4.6	4.6
prank	4.6	4.5
prequal	4.5	4.4
minimap	4.5	4.4
phyml	4.4	4.4
clustal	4.2	4.3
mrbayes	4.1	4.1
tcoffee	4.1	4.2
gadget	4.1	4.0
crisflash	4.0	4.0
PopLDdecay	3.8	3.8
cellcoal	3.8	3.6
bpp	3.8	3.6
ms	3.7	3.7
mafft	3.3	3.1
athena	2.9	2.8
covid-sim-0.13.0	2.5	2.4
indelible	1.4	1.0

Astrophysics

SoftWipe Benchmark

program name	absolute score	relative score
genesis	8.6	8.8
hyperphylo	8.6	8.6
kahypar	8.4	8.5
candy-kingdom	8.2	8.2
bindash-1.0	8.0	7.9
fastspar	7.8	7.9
repeatscounter	7.5	7.7
axe-0.3.3	7.5	7.5
virulign-1.0.1	7.4	7.4
naf-1.1.0/unnaf	7.4	7.5
naf-1.1.0/ennaf	7.4	7.4
ExpansionHunter	7.3	7.5
glucose-3-drup	7.1	7.0
raxml-ng	7.0	7.0
dawg	6.8	6.9
ntEdit-1.2.3	6.4	6.2
defor	6.3	6.4
swarm	6.2	6.2
lemon	6.1	6.0
treerecs	6.1	6.1
IQ-TREE-2.0-rc1	6.1	5.7
BGSA_CPU-1.0	5.9	5.4
emeraLD	5.8	5.5
dr_sasa_n	5.7	6.0
copmem-0.2	5.7	5.7
samtools	5.6	5.6
seq-gen	5.6	5.6
dna-nn-0.1	5.3	5.2
sf	5.2	5.2
cryfa-18.06	5.1	5.1
ngsLD	5.1	5.0
HLA-LA	4.9	4.5
iqtree1.6.10	4.9	4.9
vsearch	4.6	4.6
prank	4.6	4.5
prequal	4.5	4.4
minimap	4.5	4.4
phyml	4.4	4.4
clustal	4.2	4.3
mrbayes	4.1	4.1
tcoffee	4.1	4.2
gadget	4.1	4.0
crisflash	4.0	4.0
PopLDdecay	3.8	3.8
cellcoal	3.8	3.6
bpp	3.8	3.6
ms	3.7	3.7
mafft	3.3	3.1
athena	2.9	2.8
covid-sim-0.13.0	2.5	2.4
indelible	1.4	1.0

Tools with highly similar functionality

SoftWipe Benchmark

program name	absolute score	relative score
genesis	8.6	8.8
hyperphylo	8.6	8.6
kahypar	8.4	8.5
candy-kingdom	8.2	8.2
bindash-1.0	8.0	7.9
fastspar	7.8	7.9
repeatscounter	7.5	7.7
axe-0.3.3	7.5	7.5
virulign-1.0.1	7.4	7.4
naf-1.1.0/unnaf	7.4	7.5
naf-1.1.0/ennaf	7.4	7.4
ExpansionHunter	7.3	7.5
glucose-3-drup	7.1	7.0
raxml-ng	7.0	7.0
dawg	6.8	6.9
ntEdit-1.2.3	6.4	6.2
defor	6.3	6.4
swarm	6.2	6.2
lemon	6.1	6.0
treerecs	6.1	6.1
IQ-TREE-2.0-rc1	6.1	5.7
BGSA_CPU-1.0	5.9	5.4
emeraLD	5.8	5.5
dr_sasa_n	5.7	6.0
copmem-0.2	5.7	5.7
samtools	5.6	5.6
seq-gen	5.6	5.6
dna-nn-0.1	5.3	5.2
sf	5.2	5.2
cryfa-18.06	5.1	5.1
ngsLD	5.1	5.0
HLA-LA	4.9	4.5
iqtree1.6.10	4.9	4.9
vsearch	4.6	4.6
prank	4.6	4.5
prequal	4.5	4.4
minimap	4.5	4.4
phyml	4.4	4.4
clustal	4.2	4.3
mrbayes	4.1	4.1
tcoffee	4.1	4.2
gadget	4.1	4.0
crisflash	4.0	4.0
PopLDdecay	3.8	3.8
cellcoal	3.8	3.6
bpp	3.8	3.6
ms	3.7	3.7
mafft	3.3	3.1
athena	2.9	2.8
covid-sim-0.13.0	2.5	2.4
indelible	1.4	1.0

Tools with highly similar functionality

SoftWipe Benchmark

program name	absolute score	relative score
genesis	8.6	8.8
hyperphylo	8.6	8.6
kahypar	8.4	8.5
candy-kingdom	8.2	8.2
bindash-1.0	8.0	7.9
fastspar	7.8	7.9
repeatscounter	7.5	7.7
axe-0.3.3	7.5	7.5
virulign-1.0.1	7.4	7.4
naf-1.1.0/unnaf	7.4	7.5
naf-1.1.0/ennaf	7.4	7.4
ExpansionHunter	7.3	7.5
glucose-3-drup	7.1	7.0
raxml-ng	7.0	7.0
dawg	6.8	6.9
ntEdit-1.2.3	6.4	6.2
defor	6.3	6.4
swarm	6.2	6.2
lemon	6.1	6.0
treerecs	6.1	6.1
IQ-TREE-2.0-rc1	6.1	5.7
BGSA_CPU-1.0	5.9	5.4
emeraLD	5.8	5.5
dr_sasa_n	5.7	6.0
copmem-0.2	5.7	5.7
samtools	5.6	5.6
seq-gen	5.6	5.6
dna-nn-0.1	5.3	5.2
sf	5.2	5.2
cryfa-18.06	5.1	5.1
ngsLD	5.1	5.0
HLA-LA	4.9	4.5
iqtree1.6.10	4.9	4.9
vsearch	4.6	4.6
prank	4.6	4.5
prequal	4.5	4.4
minimap	4.5	4.4
phyml	4.4	4.4
clustal	4.2	4.3
mrbayes	4.1	4.1
tcoffee	4.1	4.2
gadget	4.1	4.0
crisflash	4.0	4.0
PopLDdecay	3.8	3.8
cellcoal	3.8	3.6
bpp	3.8	3.6
ms	3.7	3.7
mafft	3.3	3.1
athena	2.9	2.8
covid-sim-0.13.0	2.5	2.4
indelible	1.4	1.0

Covid simulation tool

SoftWipe in Practice

- Leads to healthy competition among lab members → everyone wants to write the cleanest code
- Used by researchers inside and outside of the lab during the development process → potential bugs identified and avoided
- Used as teaching tool in programming practicals
- SoftWipe score already used by us and others in Bioinformatics software paper submissions
- **Vision:** Establish software quality indicators as a necessary prerequisite for software paper submissions

Software Quality and Maintainability

- The Next Generation (**-NG**) projects:
  - Re-design, re-factoring, from scratch re-implementation of flagship tools to ensure maintainability, sustainability, and extensibility & increase scalability/performance
  - `ModelTest-NG` – `model testing of evolutionary models for phylogenetic inference`
  - `RAxML-NG` – phylogenetic inference
  - `EPA-NG` – phylogenetic placement of environmental reads

Energy Efficiency

Innovations

New Results

Comment on this paper

A Fast and Memory-Efficient Implementation of the Transfer Bootstrap

Sarah Lutteropp, Alexey M. Kozlov, Alexandros Stamatakis

doi: https://doi.org/10.1101/734848

480x speedup

EPA-ng: Massively Parallel Evolutionary Placement of Genetic Sequences

Pierre Barbera, Alexey M Kozlov, Lucas Czech, Benoit Morel, Diego Darriba, Tomáš Flouri, Alexandros Stamatakis

Systematic Biology, Volume [illegible], https://doi.org/10.1093/sysbio/syy054

30x speedup

CORRECTED PROOF

RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference

Alexey M Kozlov, Diego Darriba, Tomáš Flouri, Benoit Morel, Alexandros Stamatakis

Bioinformatics, btz305, http[illegible]/bioinformatics/btz305

Publish[illegible]

4x speedup

Multi-rate Poisson tree processes for single-locus species delimitation under maximum likelihood and Markov chain Monte Carlo

P Kapli, S Lutteropp, J Zhang, K Kobert, P Pavlidis, A Stamatakis, T Flouri

Bioinformatics, Volume 33, Issue 11, 1 June 2017, Pages [illegible]-1638, https://doi.org/10.1093/bioinformatics/btx025

Published: 20 [illegible]

~1000x speedup

Oh, wow, this will help save a lot of energy!

The Jevons Paradox

New Results

Comment on this paper

A Fast and Memory-Efficient Implementation of the Transfer Bootstrap

Sarah Lutteropp, Alexey M. Kozlov, Alexandros Stamatakis

doi: https://doi.org/10.1101/734848

480x speedup

EPA-ng: Massively Parallel Evolutionary Placement of Genetic Sequences

Pierre Barbera, Alexey M Kozlov, Lucas Czech, Benoit Morel, Diego Darriba, Tomáš Flouri, Alexandros Stamatakis

Systematic Biology, Volume [illegible], https://doi.org/10.1093/sysbio/syy054

30x speedup

CORRECTED PROOF

RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference

Alexey M Kozlov, Diego Darriba, Tomáš Flouri, Benoit Morel, Alexandros Stamatakis

Bioinformatics, btz305, http[illegible]/093/bioinformatics/btz305

Publish[illegible]

4x speedup

Multi-rate Poisson tree processes for single-locus species delimitation under maximum likelihood and Markov chain Monte Carlo

P Kapli, S Lutteropp, J Zhang, K Kobert, P Pavlidis, A Stamatakis, T Flouri

Bioinformatics, Volume 33, Issue 11, 1 June 2017, Pages 1[illegible]-1638, https://doi.org/10.1093/bioinformatics/btx025

Published: 20 January [illegible]

~1000x speedup

W. S. Jevons „The Coal Question“ (1865)

Improved efficiency ← Increased consumption rate

Lower cost

The Jevons Paradox

New Results

Comment on this paper

A Fast and Memory-Efficient Implementation of the Transfer Bootstrap

Sarah Lutteropp, Alexey M. Kozlov, Alexandros Stamatakis

doi: https://doi.org/10.1101/734848

480x speedup

EPA-ng: Massively Parallel Evolutionary Placement of Genetic Sequences

Pierre Barbera, Alexey M Kozlov, Lucas Czech, Benoit Morel, Diego Darriba, Tomáš Flouri, Alexandros Stamatakis

Systematic Biology, Volume [illegible], [illegible], https://doi.org/10.1093/sysbio/syy054

30x speedup

CORRECTED PROOF

RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference

Alexey M Kozlov, Diego Darriba, Tomáš Flouri, Benoit Morel, Alexandros Stamatakis

Bioinformatics, btz305, http[illegible]/bioinformatics/btz305

Publish[illegible]

4x speedup

Multi-rate Poisson tree processes for single-locus species delimitation under maximum likelihood and Markov chain Monte Carlo

P Kapli, S Lutteropp, J Zhang, K Kobert, P Pavlidis, A Stamatakis, T Flouri

Bioinformatics, Volume 33, Issue 11, 1 June 2017, Pages [illegible]-1638, https://doi.org/10.1093/bioinformatics/btx025

Published: 20 [illegible]

~1000x speedup

We need an alternative solution!

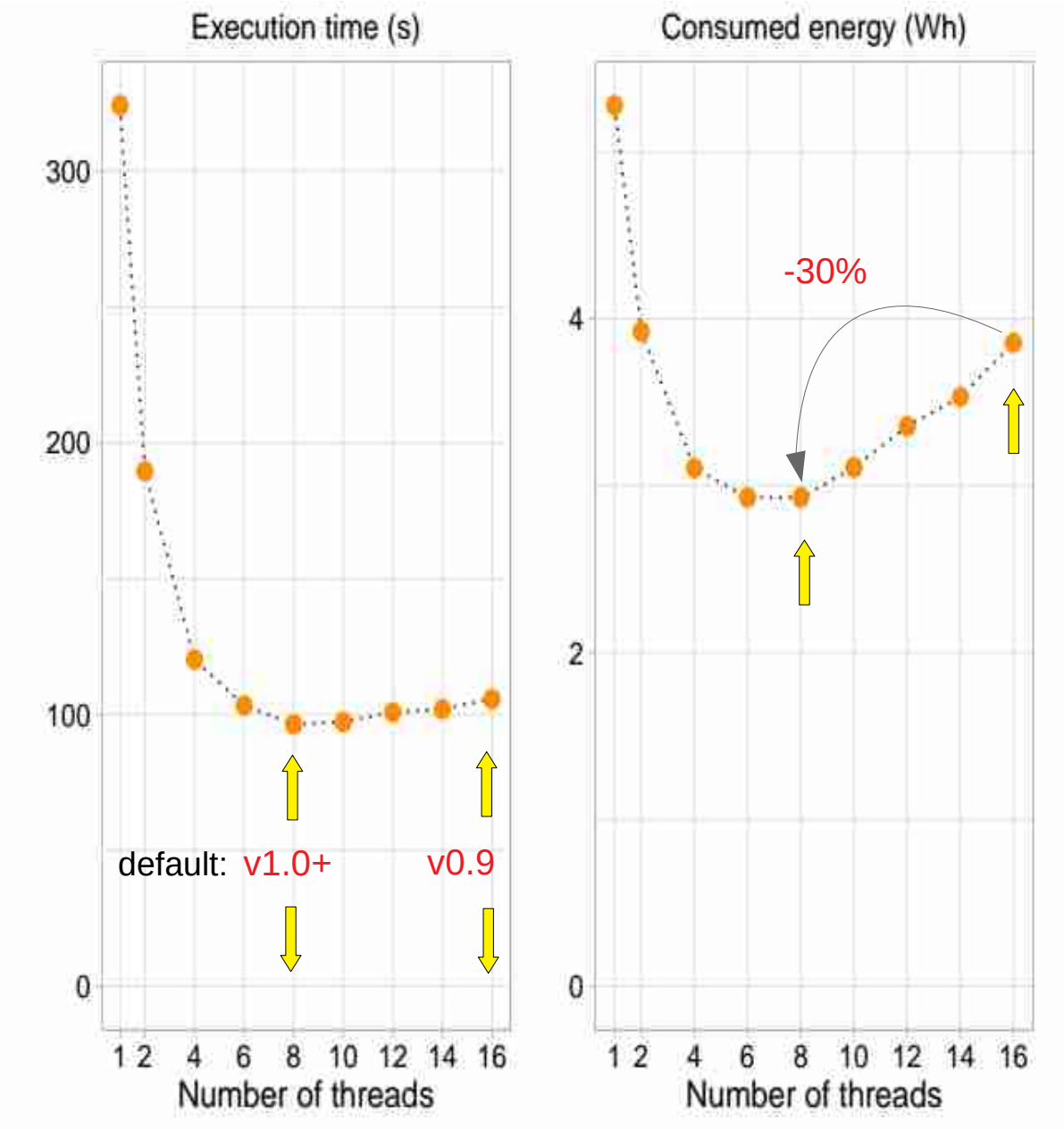
Energy monitoring: RAxML-NG

- New in RAxML-NG v1.0: energy usage report

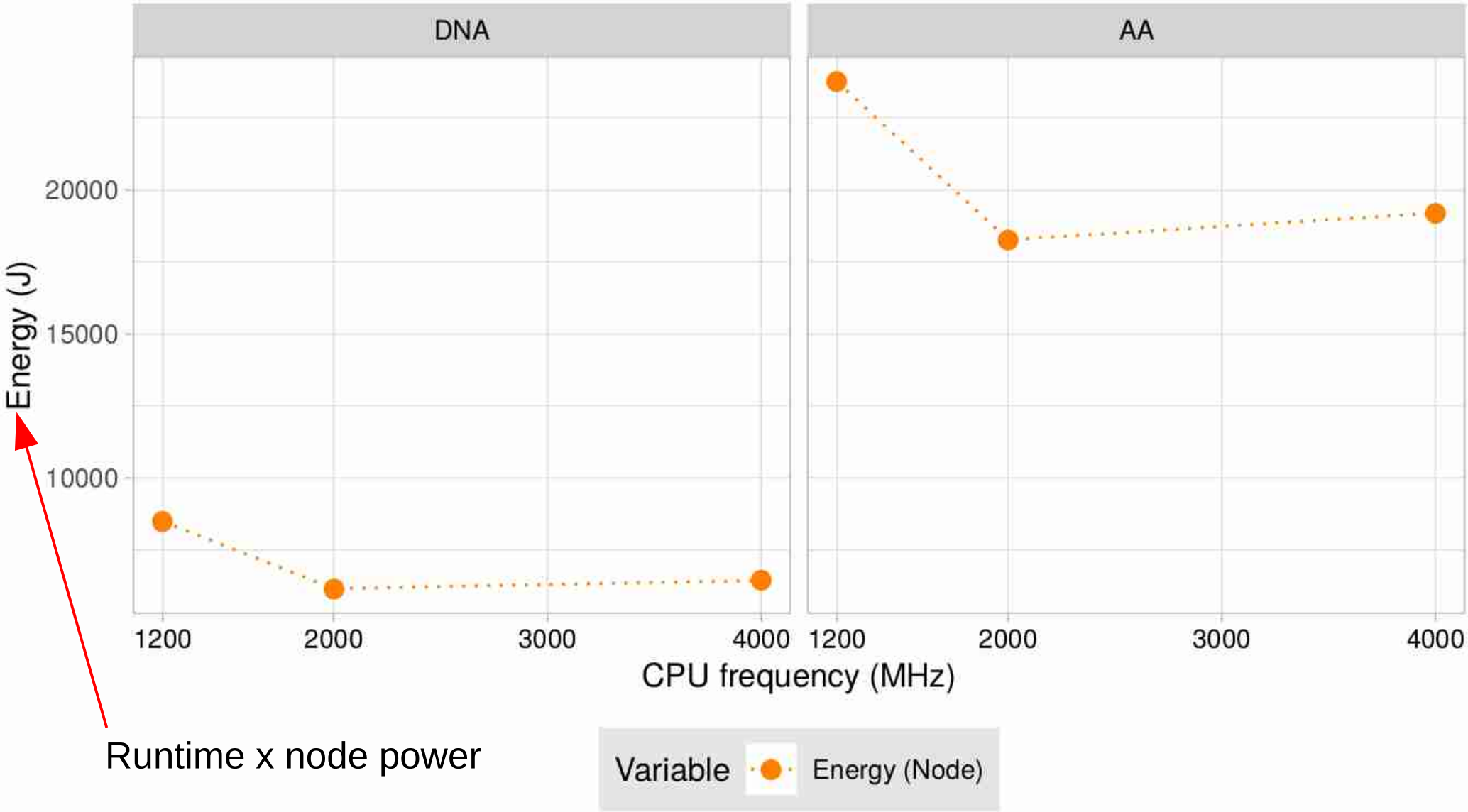
```
Elapsed time: 42846.287 seconds

Consumed energy: 162370.469 Wh (= 812 km in an electric car, or 4059 km with an e-scooter!)
```

Single tree search (96 nodes x 12h):
>160 kWh

Energy Saving Mode in RAxML-NG v1.0



Phylogenetic Inference: Energy as a function of CPU clock frequency



Runtime x node power

Phylogenetic Inference: Energy as a function of CPU clock frequency