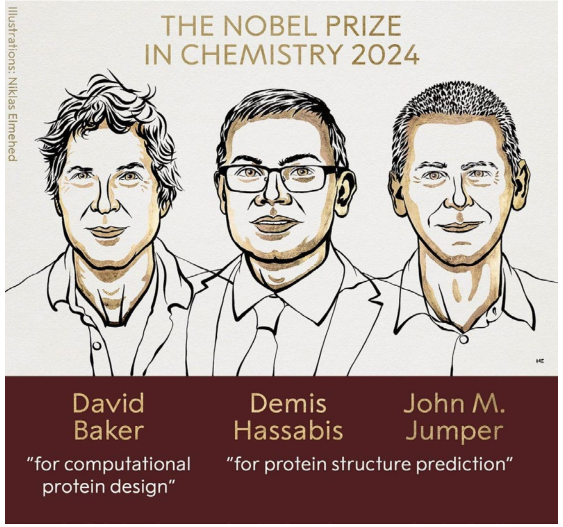
A machine-learning based alternative to phylogenetic bootstrap

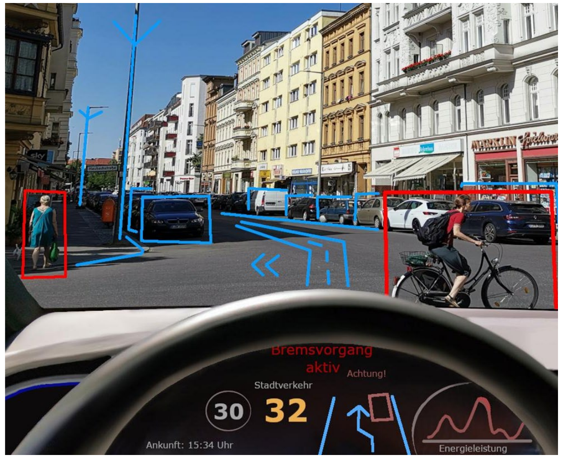
Tal Pupko,
The Shmunis School of Biomedicine and Cancer Research

IMSI, Aug 12 2025
Chicago, USA

AI revolution is transforming the world

- Large Language Models, e.g., ChatGPT
- Autonomous systems (self-driving cars, drones)
- Creativity - first Christie’s AI art sale
- Drug discovery and health care, AlphaFold2


Illustrations: Niklas Elmehed
THE NOBEL PRIZE IN CHEMISTRY 2024
David Baker
“for computational protein design”
Demis Hassabis
John M. Jumper
“for protein structure prediction”



AI phylogeny

Our research aims to harness AI to improve phylogenetic inference

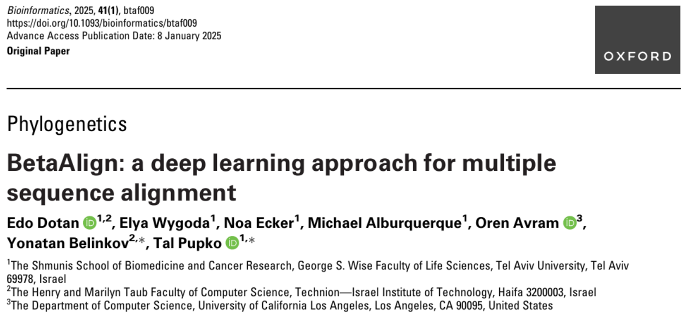
nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-021-22073-8 OPEN

Harnessing machine learning to guide phylogenetic-tree search algorithms

Dana Azouri 1,2, Shiran Abadi 1, Yishay Mansour3, Itay Mayrose 1✉ & Tal Pupko 2✉


Bioinformatics, 2025, **41(1)**, btaf009
https://doi.org/10.1093/bioinformatics/btaf009
Advance Access Publication Date: 8 January 2025
Original Paper

OXFORD

Phylogenetics

BetaAlign: a deep learning approach for multiple sequence alignment

Edo Dotan 1,2, Elya Wygoda1, Noa Ecker1, Michael Alburquerque1, Oren Avram 3, Yonatan Belinkov2,*, Tal Pupko 1,*

[1]The Shmunis School of Biomedicine and Cancer Research, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 69978, Israel
[2]The Henry and Marilyn Taub Faculty of Computer Science, Technion—Israel Institute of Technology, Haifa 3200003, Israel
[3]The Department of Computer Science, University of California Los Angeles, Los Angeles, CA 90095, United States

The Tree Reconstruction Game: Phylogenetic Reconstruction Using Reinforcement Learning

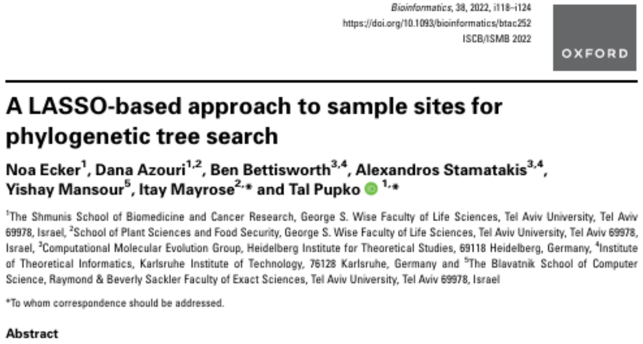
Dana Azouri ,1,2,† Oz Granit ,3,† Michael Alburquerque ,2,† Yishay Mansour ,3,*
Tal Pupko ,2,* and Itay Mayrose 1,*

[1]School of Plant Sciences and Food Security, Tel Aviv University, Ramat Aviv, Tel Aviv 69978, Israel
[2]The Shmunis School of Biomedicine and Cancer Research, Tel Aviv University, Ramat Aviv, Tel Aviv 69978, Israel
[3]Balvatnik School of Computer Science, Tel Aviv University, Ramat Aviv, Tel Aviv 69978, Israel
[†]These author contributed equally.

***Corresponding authors:** E-mails: itaymay@tauex.tau.ac.il; talp@tauex.tau.ac.il; mansour@tauex.tau.ac.il.
Associate editor: Andrey Rzhetsky

Abstract

The computational search for the maximum-likelihood phylogenetic tree is an NP-hard problem. As such, current tree search algorithms might result in a tree that is the local optima, not the global one. Here, we introduce a paradigm shift for predicting the maximum-likelihood tree, by approximating long-term gains of likelihood rather than


Bioinformatics, 38, 2022, i118–i124
https://doi.org/10.1093/bioinformatics/btac252
ISCB/ISMB 2022

OXFORD

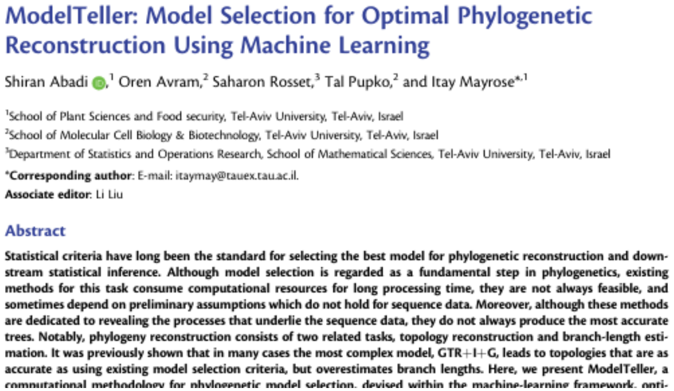
A LASSO-based approach to sample sites for phylogenetic tree search

Noa Ecker1, Dana Azouri1,2, Ben Bettisworth3,4, Alexandros Stamatakis3,4, Yishay Mansour5, Itay Mayrose2,* and Tal Pupko 1,*

[1]The Shmunis School of Biomedicine and Cancer Research, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 69978, Israel, [2]School of Plant Sciences and Food Security, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 69978, Israel, [3]Computational Molecular Evolution Group, Heidelberg Institute for Theoretical Studies, 69118 Heidelberg, Germany, [4]Institute of Theoretical Informatics, Karlsruhe Institute of Technology, 76128 Karlsruhe, Germany and [5]The Blavatnik School of Computer Science, Raymond & Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv 69978, Israel

*To whom correspondence should be addressed.

Abstract


ModelTeller: Model Selection for Optimal Phylogenetic Reconstruction Using Machine Learning

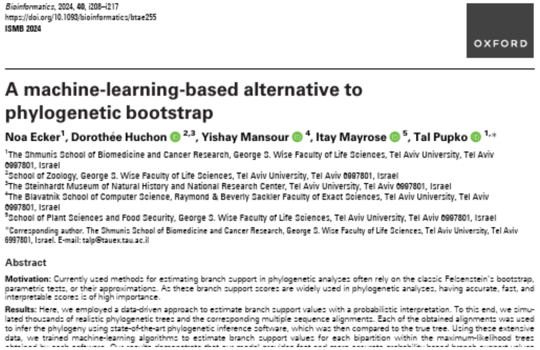
Shiran Abadi ,1 Oren Avram,2 Saharon Rosset,3 Tal Pupko,2 and Itay Mayrose*,1

[1]School of Plant Sciences and Food security, Tel-Aviv University, Tel-Aviv, Israel
[2]School of Molecular Cell Biology & Biotechnology, Tel-Aviv University, Tel-Aviv, Israel
[3]Department of Statistics and Operations Research, School of Mathematical Sciences, Tel-Aviv University, Tel-Aviv, Israel

***Corresponding author**: E-mail: itaymay@tauex.tau.ac.il.
Associate editor: Li Liu

Abstract

Statistical criteria have long been the standard for selecting the best model for phylogenetic reconstruction and downstream statistical inference. Although model selection is regarded as a fundamental step in phylogenetics, existing methods for this task consume computational resources for long processing time, they are not always feasible, and sometimes depend on preliminary assumptions which do not hold for sequence data. Moreover, although these methods are dedicated to revealing the processes that underlie the sequence data, they do not always produce the most accurate trees. Notably, phylogeny reconstruction consists of two related tasks, topology reconstruction and branch-length estimation. It was previously shown that in many cases the most complex model, GTR+I+G, leads to topologies that are as accurate as using existing model selection criteria, but overestimates branch lengths. Here, we present ModelTeller, a computational methodology for phylogenetic model selection, devised within the machine-learning framework, opti-


Bioinformatics, 2024, 40, i208–i217
https://doi.org/10.1093/bioinformatics/btae255
ISMB 2024

OXFORD

A machine-learning-based alternative to phylogenetic bootstrap

Noa Ecker1, Dorothée Huchon 2,3, Yishay Mansour 4, Itay Mayrose 5, Tal Pupko 1,*

[1]The Shmunis School of Biomedicine and Cancer Research, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 6997801, Israel
[2]School of Zoology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 6997801, Israel
[3]The Steinhardt Museum of Natural History and National Research Center, Tel Aviv University, Tel Aviv 6997801, Israel
[4]The Blavatnik School of Computer Science, Raymond & Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv 6997801, Israel
[5]School of Plant Sciences and Food Security, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 6997801, Israel
*Corresponding author. The Shmunis School of Biomedicine and Cancer Research, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 6997801, Israel. E-mail: talp@tauex.tau.ac.il

Abstract

Motivation: Currently used methods for estimating branch support in phylogenetic analyses often rely on the classic Felsenstein's bootstrap, parametric tests, or their approximations. As these branch support scores are widely used in phylogenetic analyses, having accurate, fast, and interpretable scores is of high importance.

Results: Here, we employed a data-driven approach to estimate branch support values with a probabilistic interpretation. To this end, we simulated thousands of realistic phylogenetic trees and the corresponding multiple sequence alignments. Each of the obtained alignments was used to infer the phylogeny using state-of-the-art phylogenetic inference software, which was then compared to the true tree. Using these extensive data, we trained machine-learning algorithms to estimate branch support values for each bipartition within the maximum-likelihood trees

AI phylogeny

Part 1:

Bioinformatics, 2024, 40, i208–i217
https://doi.org/10.1093/bioinformatics/btae255
ISMB 2024

OXFORD

A machine-learning-based alternative to phylogenetic bootstrap

Noa Ecker[1], Dorothée Huchon [2,3], Yishay Mansour [4], Itay Mayrose [5], Tal Pupko [1,*]

[1]The Shmunis School of Biomedicine and Cancer Research, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 6997801, Israel
[2]School of Zoology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 6997801, Israel
[3]The Steinhardt Museum of Natural History and National Research Center, Tel Aviv University, Tel Aviv 6997801, Israel
[4]The Blavatnik School of Computer Science, Raymond & Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv 6997801, Israel
[5]School of Plant Sciences and Food Security, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 6997801, Israel

*Corresponding author. The Shmunis School of Biomedicine and Cancer Research, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 6997801, Israel. E-mail: talp@tauex.tau.ac.il

Abstract

Motivation: Currently used methods for estimating branch support in phylogenetic analyses often rely on the classic Felsenstein's bootstrap, parametric tests, or their approximations. As these branch support scores are widely used in phylogenetic analyses, having accurate, fast, and interpretable scores is of high importance.

Results: Here, we employed a data-driven approach to estimate branch support values with a probabilistic interpretation. To this end, we simulated thousands of realistic phylogenetic trees and the corresponding multiple sequence alignments. Each of the obtained alignments was used to infer the phylogeny using state-of-the-art phylogenetic inference software, which was then compared to the true tree. Using these extensive data, we trained machine-learning algorithms to estimate branch support values for each bipartition within the maximum-likelihood trees obtained by each software. Our results demonstrate that our model provides fast and more accurate probability-based branch support values

Phylogenetic-based tools and algorithms

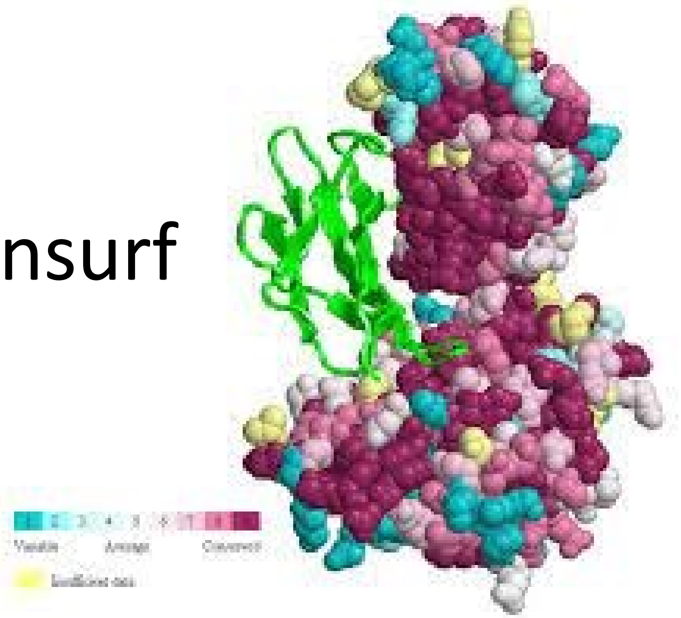
The GUIDANCE2 Server

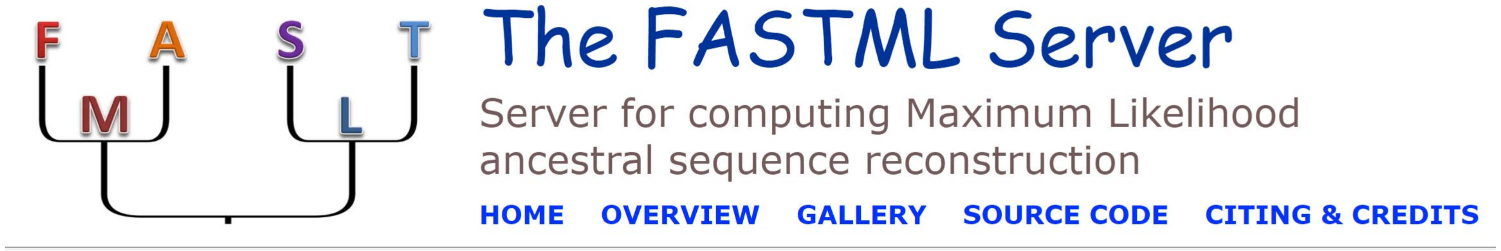
Server for alignment confidence score

M1CR0B1AL1Z3R

A web server for analyzing bacterial genomics data. Easily.

Consurf



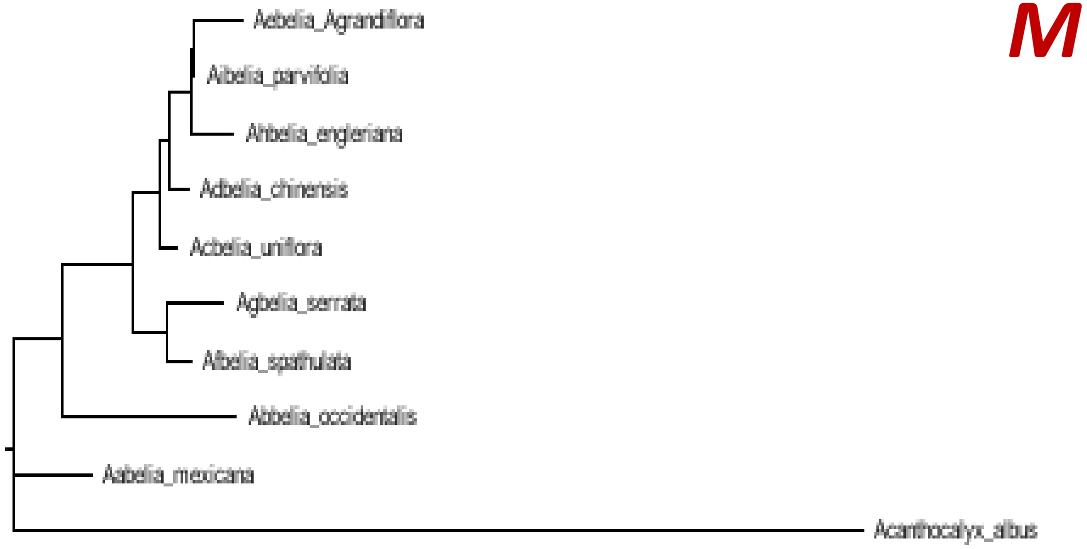


The FASTML Server

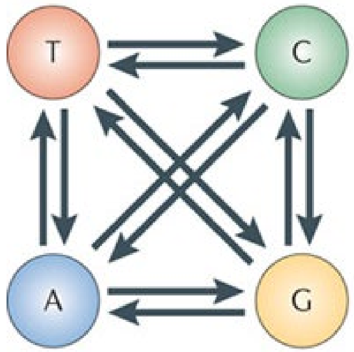
Server for computing Maximum Likelihood ancestral sequence reconstruction

The tree score = log-likelihood

T = Tree



M = Evolutionary model



Tree likelihood

$P(D\,|\,M,T)$: The conditional probability of the data, given the model, *M*, and the tree, *T*

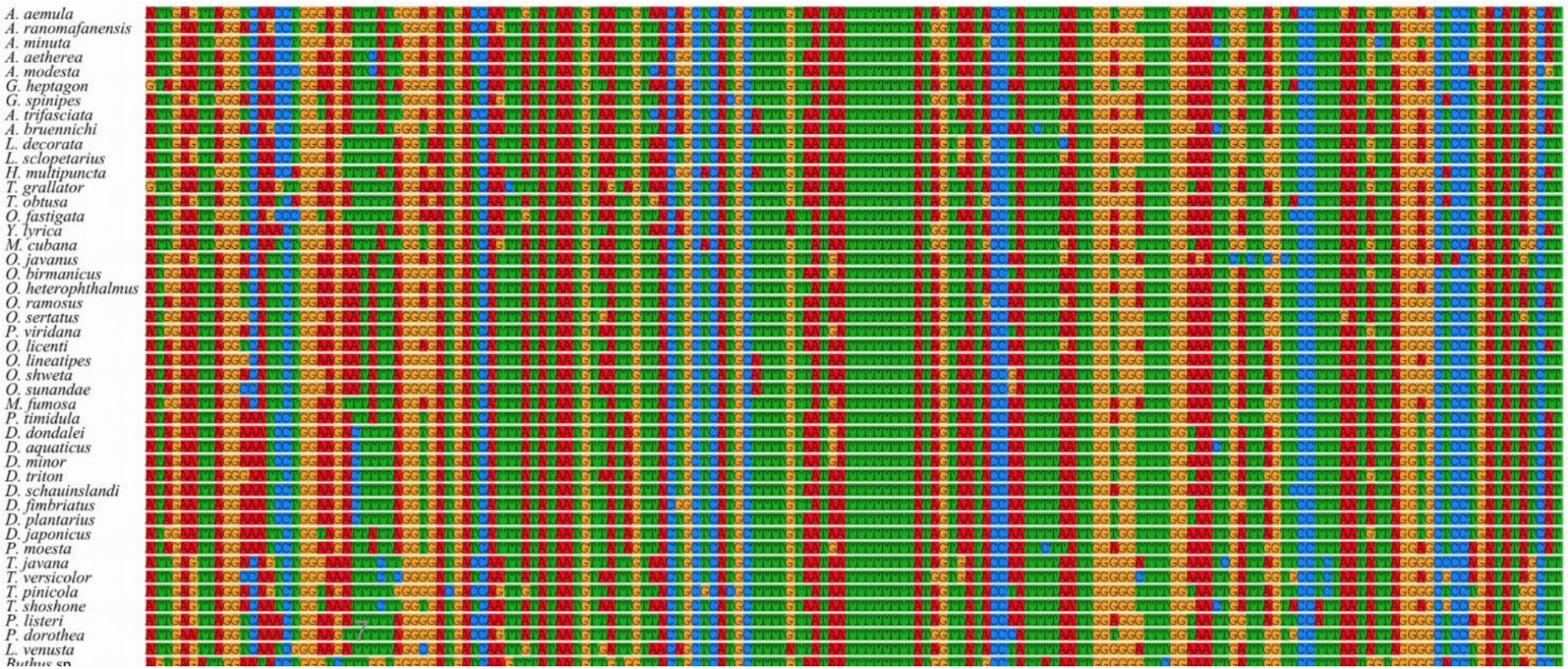
D = Data = Alignment

```
Acanthocalyx_albus     tcgaaacttg cccagcagag cgaccagcga acacctccgt
Abelia_spathulata      tcgaaacctg cacagcagaa cgacccgcga acacgttcgt
Abelia_engleriana      tcgaaacctg cacagcagaa cgacccgcga acacgttcgt
Abelia_chinensis       tcgaaacctg cacagcagaa cgacccgcga acacgttcgt
Abelia_parvifolia      ---------g cacagcagaa cgacccgcga acacgttcgt
Abelia_Agrandiflora    tcgaaacctg cacagcagaa cgacccgcga acacgttcgt
Abelia_mexicana        -tcgaaactg cacagcagaa cgacccgcga acacgttcgt
Abelia_occidentalis    tcgaaacctg cacagcagaa cgacccgcga acacgttcgt
```

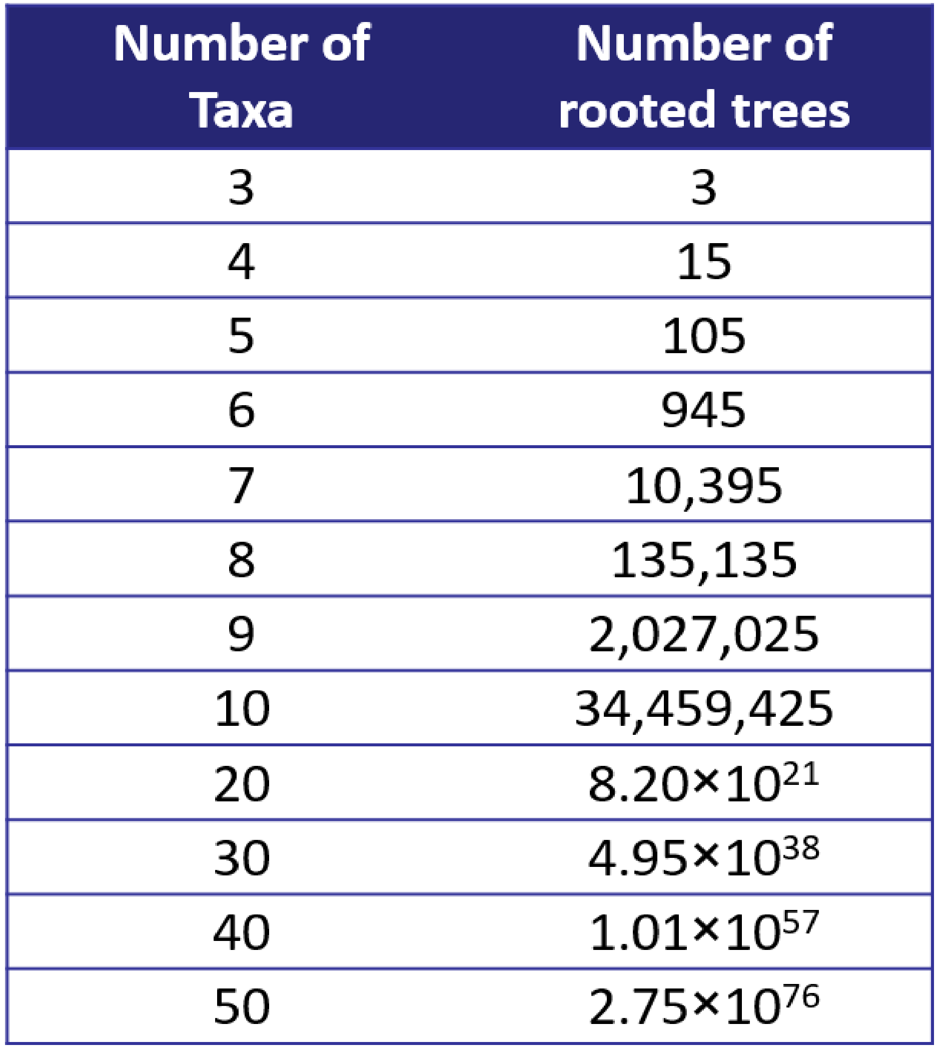
Finding trees keeps becoming more difficult

Today:

- Data: Aligned genomic sequences
- Size: Up to GB of DNA sites, thousands of species



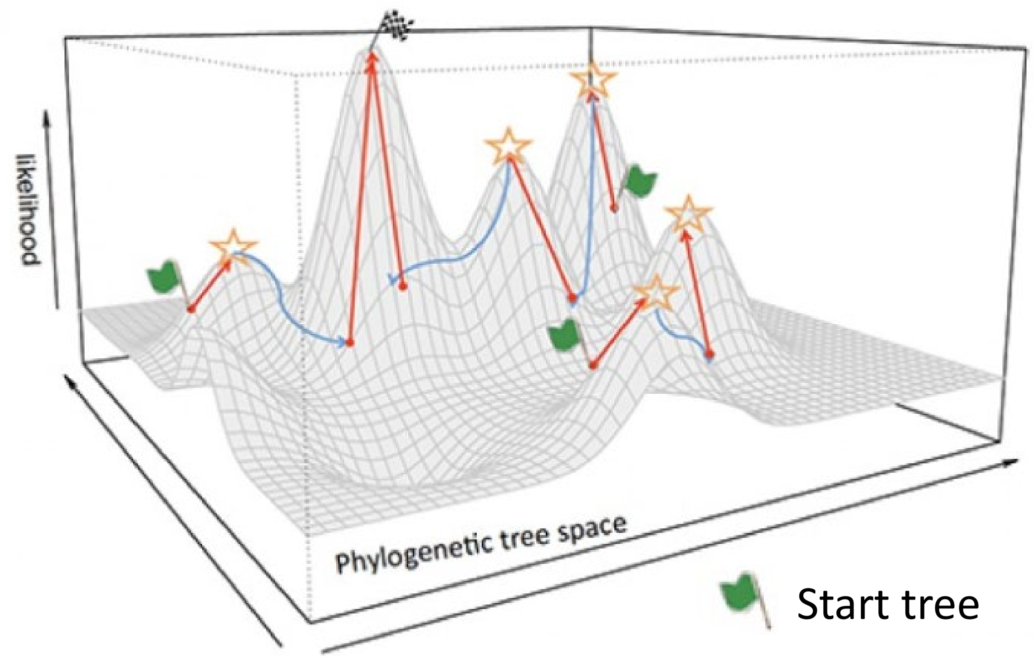
The tree space is huge



Number of Taxa	Number of rooted trees
3	3
4	15
5	105
6	945
7	10,395
8	135,135
9	2,027,025
10	34,459,425
20	8.20×10^{21}
30	4.95×10^{38}
40	1.01×10^{57}
50	2.75×10^{76}

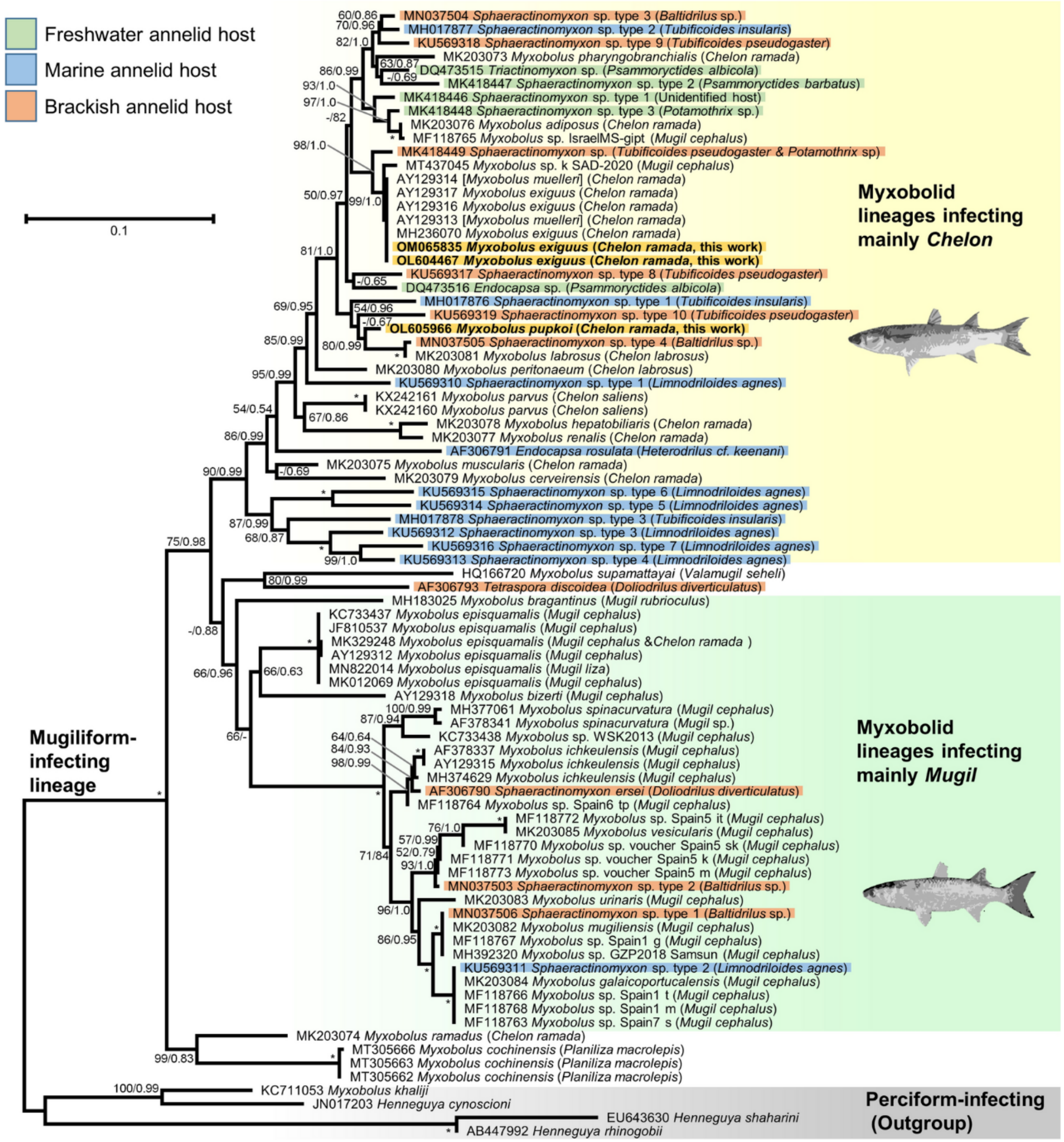
Heuristic search

- Start the search with a good guess for a starting tree
- Examine all "neighboring" trees by making small modifications to the current tree
- Move to the neighbor with the highest (likelihood) score

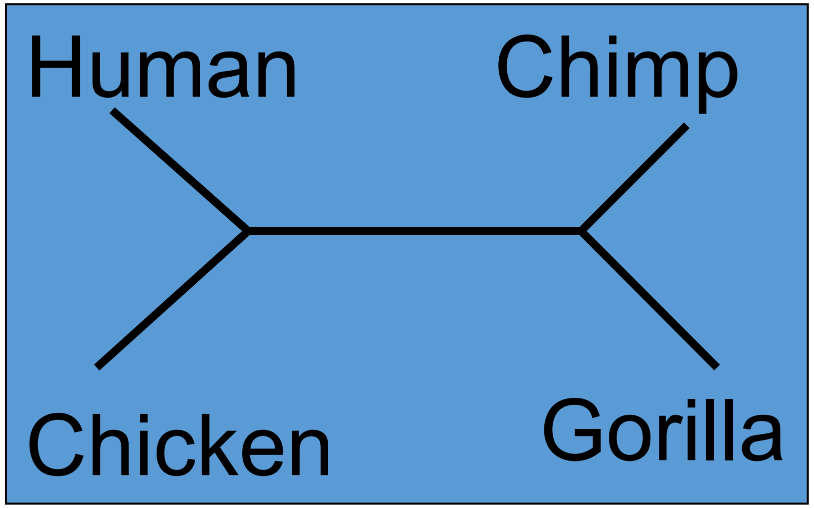


- To avoid local maxima, we start from multiple starting points

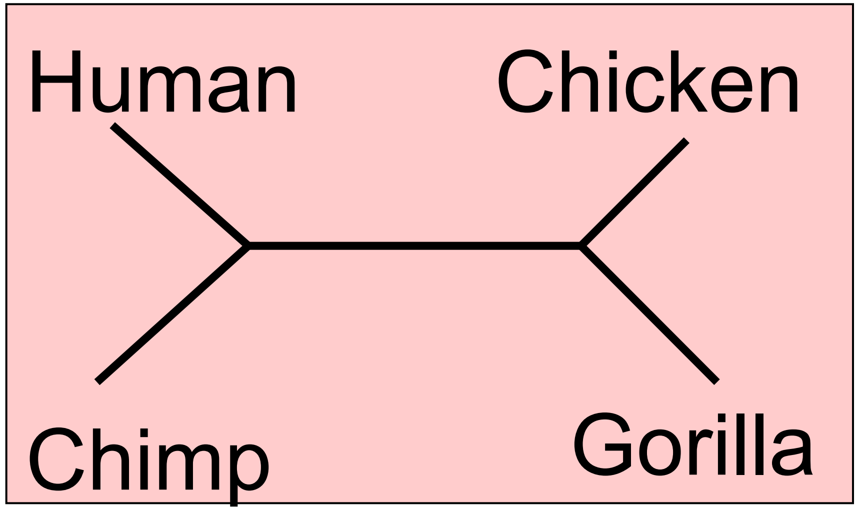
We can infer the ML tree


🟩 Freshwater annelid host
🟦 Marine annelid host
🟧 Brackish annelid host
0.1
MN037504 Sphaeractinomyxon sp. type 3 (Baltidrilus sp.)
MH017877 Sphaeractinomyxon sp. type 2 (Tubificoides insularis)
KU569318 Sphaeractinomyxon sp. type 9 (Tubificoides pseudogaster)
MK203073 Myxobolus pharyngobranchialis (Chelon ramada)
DQ473515 Triactinomyxon sp. (Psammoryctides albicola)
MK418447 Sphaeractinomyxon sp. type 2 (Psammoryctides barbatus)
MK418446 Sphaeractinomyxon sp. type 1 (Unidentified host)
MK418448 Sphaeractinomyxon sp. type 3 (Potamothrix sp.)
MK203076 Myxobolus adiposus (Chelon ramada)
MF118765 Myxobolus sp. IsraelMS-gipt (Mugil cephalus)
MK418449 Sphaeractinomyxon sp. (Tubificoides pseudogaster & Potamothrix sp)
MT437045 Myxobolus sp. k SAD-2020 (Mugil cephalus)
AY129314 [Myxobolus muelleri] (Chelon ramada)
AY129317 Myxobolus exiguus (Chelon ramada)
AY129316 Myxobolus exiguus (Chelon ramada)
AY129313 [Myxobolus muelleri] (Chelon ramada)
MH236070 Myxobolus exiguus (Chelon ramada)
OM065835 Myxobolus exiguus (Chelon ramada, this work)
OL604467 Myxobolus exiguus (Chelon ramada, this work)
KU569317 Sphaeractinomyxon sp. type 8 (Tubificoides pseudogaster)
DQ473516 Endocapsa sp. (Psammoryctides albicola)
MH017876 Sphaeractinomyxon sp. type 1 (Tubificoides insularis)
KU569319 Sphaeractinomyxon sp. type 10 (Tubificoides pseudogaster)
OL605966 Myxobolus pupkoi (Chelon ramada, this work)
MN037505 Sphaeractinomyxon sp. type 4 (Baltidrilus sp.)
MK203081 Myxobolus labrosus (Chelon labrosus)
MK203080 Myxobolus peritonaeum (Chelon labrosus)
KU569310 Sphaeractinomyxon sp. type 1 (Limnodriloides agnes)
KX242161 Myxobolus parvus (Chelon saliens)
KX242160 Myxobolus parvus (Chelon saliens)
MK203078 Myxobolus hepatobiliaris (Chelon ramada)
MK203077 Myxobolus renalis (Chelon ramada)
AF306791 Endocapsa rosulata (Heterodrilus cf. keenani)
MK203075 Myxobolus muscularis (Chelon ramada)
MK203079 Myxobolus cerveirensis (Chelon ramada)
KU569315 Sphaeractinomyxon sp. type 6 (Limnodriloides agnes)
KU569314 Sphaeractinomyxon sp. type 5 (Limnodriloides agnes)
MH017878 Sphaeractinomyxon sp. type 3 (Tubificoides insularis)
KU569312 Sphaeractinomyxon sp. type 3 (Limnodriloides agnes)
KU569316 Sphaeractinomyxon sp. type 7 (Limnodriloides agnes)
KU569313 Sphaeractinomyxon sp. type 4 (Limnodriloides agnes)
HQ166720 Myxobolus supamattayai (Valamugil seheli)
AF306793 Tetraspora discoidea (Doliodrilus diverticulatus)
MH183025 Myxobolus bragantinus (Mugil rubrioculus)
KC733437 Myxobolus episquamalis (Mugil cephalus)
JF810537 Myxobolus episquamalis (Mugil cephalus)
MK329248 Myxobolus episquamalis (Mugil cephalus &Chelon ramada)
AY129312 Myxobolus episquamalis (Mugil cephalus)
MN822014 Myxobolus episquamalis (Mugil liza)
MK012069 Myxobolus episquamalis (Mugil cephalus)
AY129318 Myxobolus bizerti (Mugil cephalus)
MH377061 Myxobolus spinacurvatura (Mugil cephalus)
AF378341 Myxobolus spinacurvatura (Mugil sp.)
KC733438 Myxobolus sp. WSK2013 (Mugil cephalus)
AF378337 Myxobolus ichkeulensis (Mugil cephalus)
AY129315 Myxobolus ichkeulensis (Mugil cephalus)
MH374629 Myxobolus ichkeulensis (Mugil cephalus)
AF306790 Sphaeractinomyxon ersei (Doliodrilus diverticulatus)
MF118764 Myxobolus sp. Spain6 tp (Mugil cephalus)
MF118772 Myxobolus sp. Spain5 it (Mugil cephalus)
MK203085 Myxobolus vesicularis (Mugil cephalus)
MF118770 Myxobolus sp. voucher Spain5 sk (Mugil cephalus)
MF118771 Myxobolus sp. voucher Spain5 k (Mugil cephalus)
MF118773 Myxobolus sp. voucher Spain5 m (Mugil cephalus)
MN037503 Sphaeractinomyxon sp. type 2 (Baltidrilus sp.)
MK203083 Myxobolus urinaris (Mugil cephalus)
MN037506 Sphaeractinomyxon sp. type 1 (Baltidrilus sp.)
MK203082 Myxobolus mugiliensis (Mugil cephalus)
MF118767 Myxobolus sp. Spain1 g (Mugil cephalus)
MH392320 Myxobolus sp. GZP2018 Samsun (Mugil cephalus)
KU569311 Sphaeractinomyxon sp. type 2 (Limnodriloides agnes)
MK203084 Myxobolus galaicoportucalensis (Mugil cephalus)
MF118766 Myxobolus sp. Spain1 t (Mugil cephalus)
MF118768 Myxobolus sp. Spain1 m (Mugil cephalus)
MF118763 Myxobolus sp. Spain7 s (Mugil cephalus)
MK203074 Myxobolus ramadus (Chelon ramada)
MT305666 Myxobolus cochinensis (Planiliza macrolepis)
MT305663 Myxobolus cochinensis (Planiliza macrolepis)
MT305662 Myxobolus cochinensis (Planiliza macrolepis)
KC711053 Myxobolus khaliji
JN017203 Henneguya cynoscioni
EU643630 Henneguya shaharini
AB447992 Henneguya rhinogobii
Myxobolid lineages infecting mainly Chelon
Myxobolid lineages infecting mainly Mugil
Mugiliform-infecting lineage
Perciform-infecting (Outgroup)
60/0.86
70/0.96
82/1.0
63/0.87
-/0.69
86/0.99
93/1.0
97/1.0
-/82
98/1.0
50/0.97
99/1.0
81/1.0
-/0.65
69/0.95
54/0.96
-/0.67
85/0.99
80/0.99
95/0.99
54/0.54
67/0.86
86/0.99
90/0.99
-/0.69
87/0.99
68/0.87
99/1.0
75/0.98
80/0.99
-/0.88
66/0.96
66/0.63
66/-
100/0.99
87/0.94
64/0.64
84/0.93
98/0.99
76/1.0
57/0.99
52/0.79
71/84
93/1.0
96/1.0
86/0.95
99/0.83
100/0.99
*

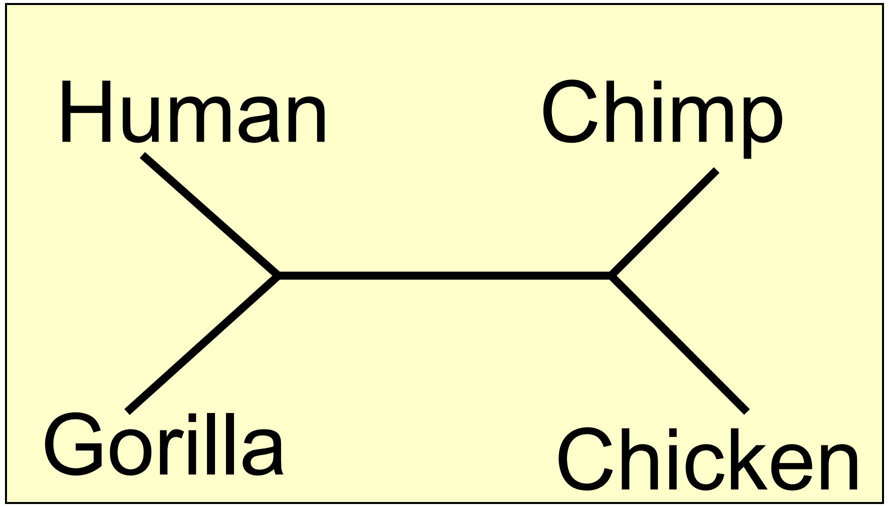
Different (unrooted) trees



LL score = -110



LL score = -107



LL score = -117

How confidence we are in the best tree?

The phylogenetic jackknife

Jackknife

- We create new data sets by sampling randomly half of the characters without replacement.

- We generate 100 pseudo-data sets.

- We do not change the number of sequences, just the number of positions!

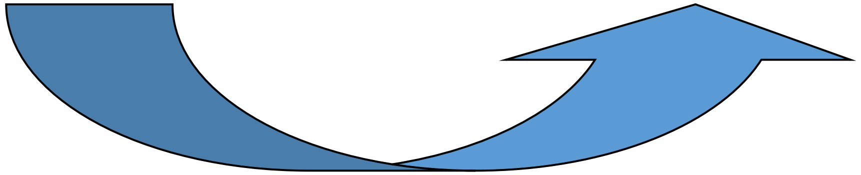
Jackknife

Original data

```
s1   AAGTAA
s2   CAAAAC
s3   CAGGAA
s4   AAATAC
```

Pseudo-data 1

```
s1   AGA
s2   AAA
s3   AGA
s4   AAA
```



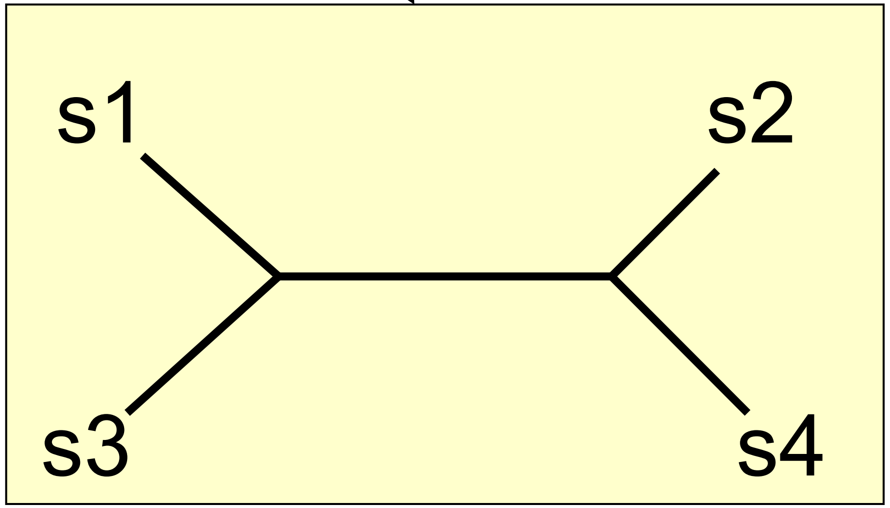
We removed positions 1,4, and 6

Jackknife

Pseudo-data 1

```
s1   AGA
s2   AAA
s3   AGA
s4   AAA
```

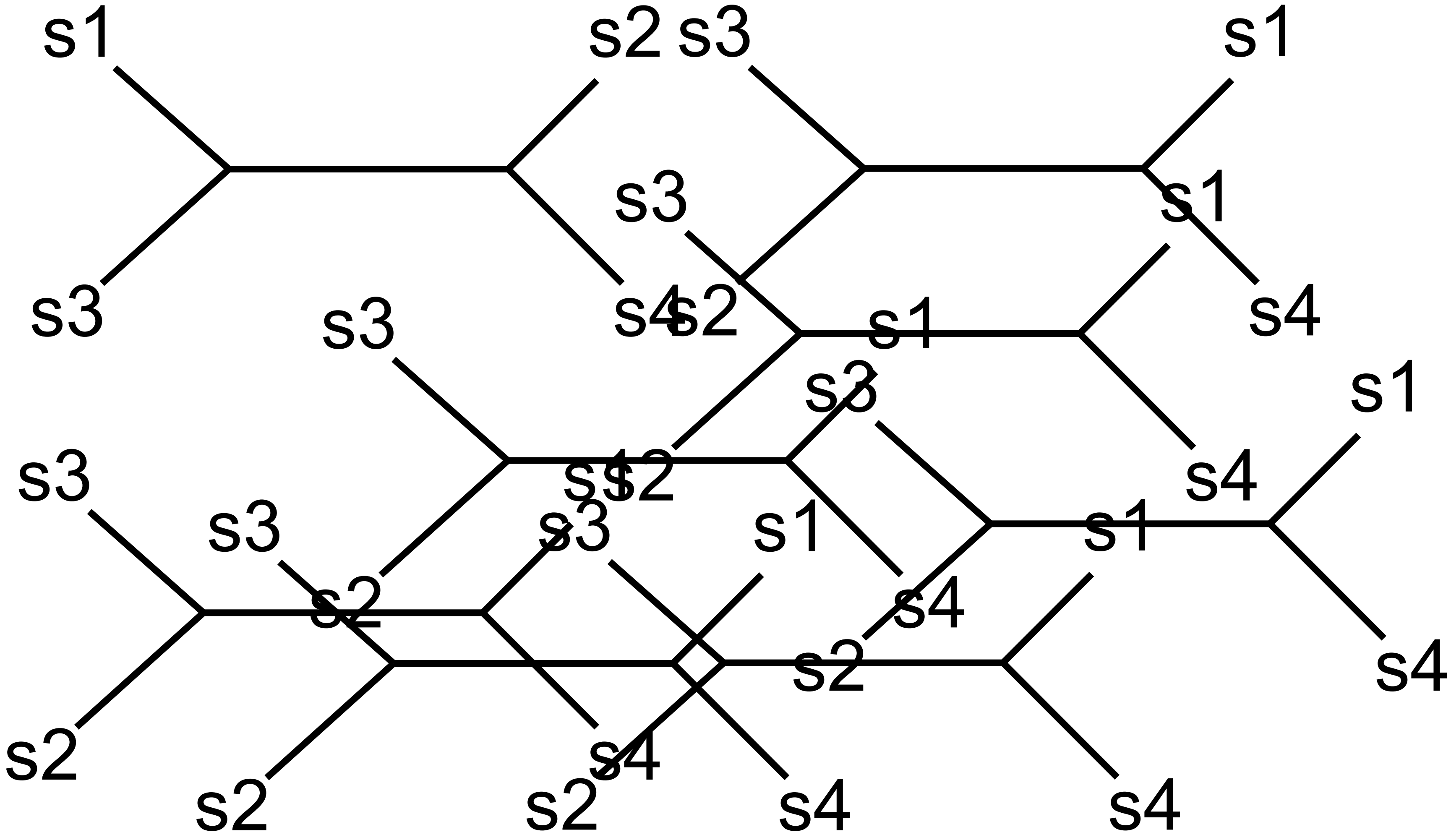
ML tree search



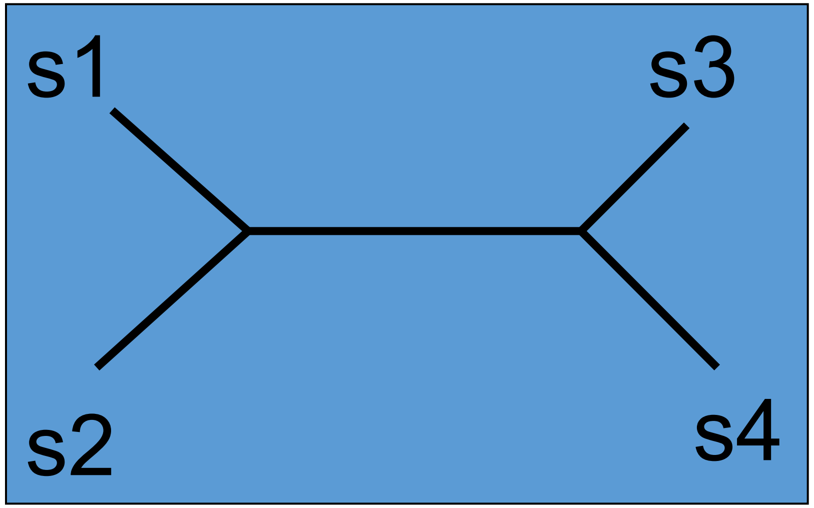
Best tree for pseudo-data 1

Jackknife

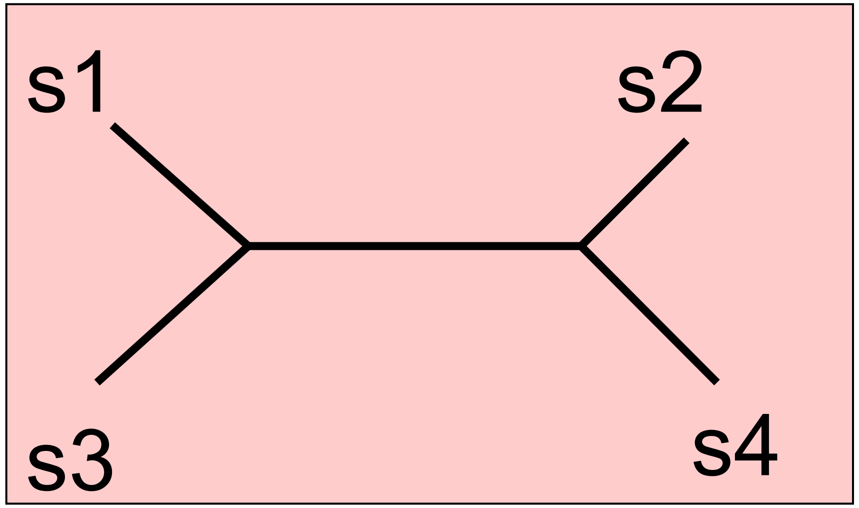
- We repeat the process 100 times and get 100 best trees



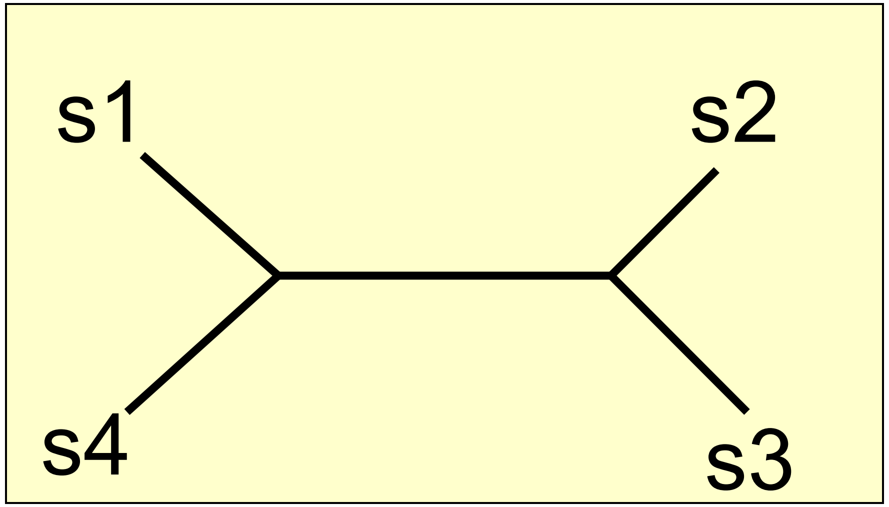
Jackknife percentage



30 times



50 times



20 times

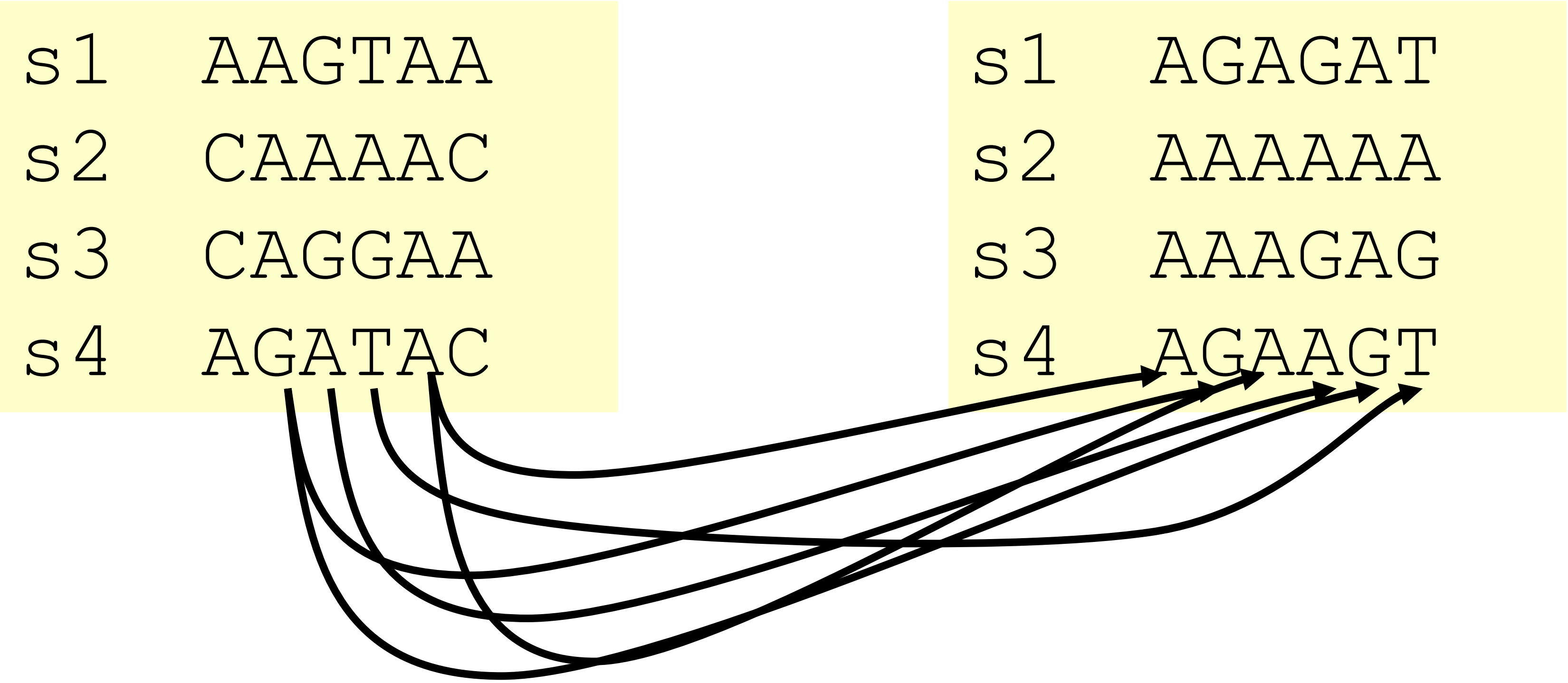
Our confidence in the pink tree in terms of jackknife support is 50%.

The phylogenetic bootstrap

Bootstrap

Original data

Pseudo-data 1



```
s1   AAGTAA        s1   AGAGAT
s2   CAAAAC        s2   AAAAAA
s3   CAGGAA        s3   AAAGAG
s4   AGATAC        s4   AGAAGT
```

Same idea as jackknife but we sample with repetition. In this example, we sample positions 2 twice and position 5 twice and zero for position 1 and 6.

Bootstrap

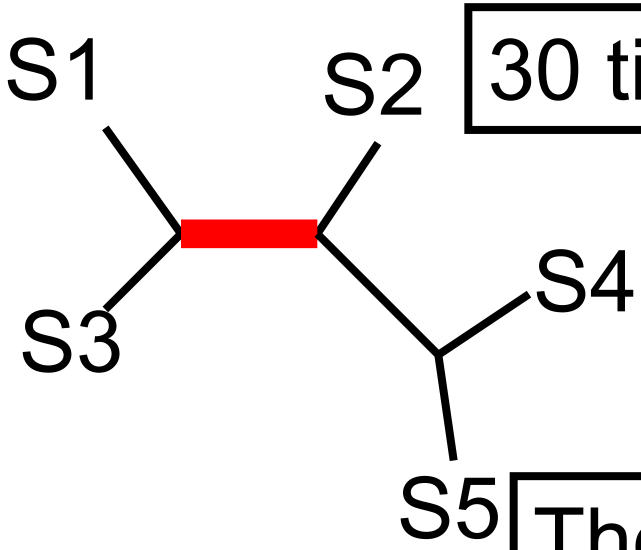
- Bootstrap is used more than jackknife in phylogeny, because it has the same data-size as the original data.

Bootstrap for splits

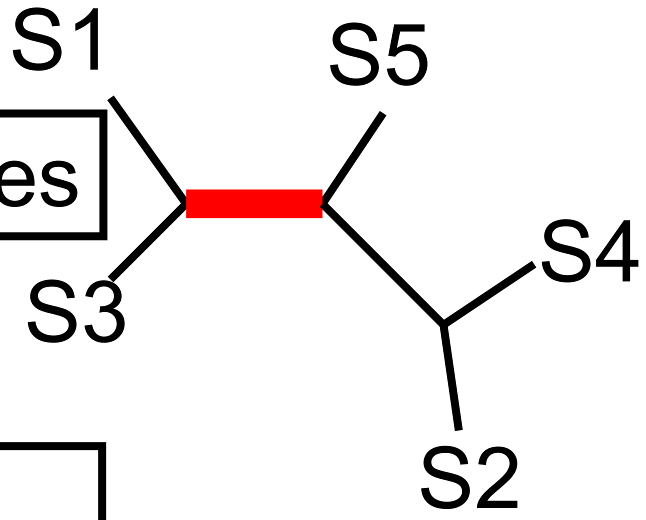
- Instead of getting the support for each tree, we can compute the support for a given split.

- The support for a given split is the percentage of pseudo-tree in which this split appears.

Bootstrap for splits

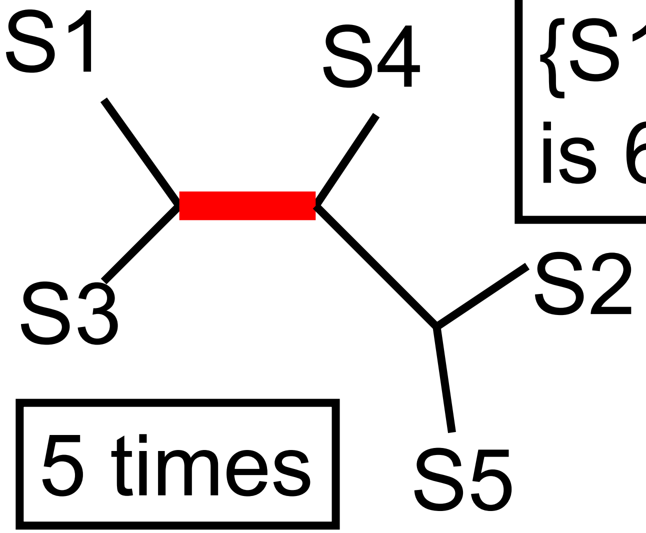


30 times

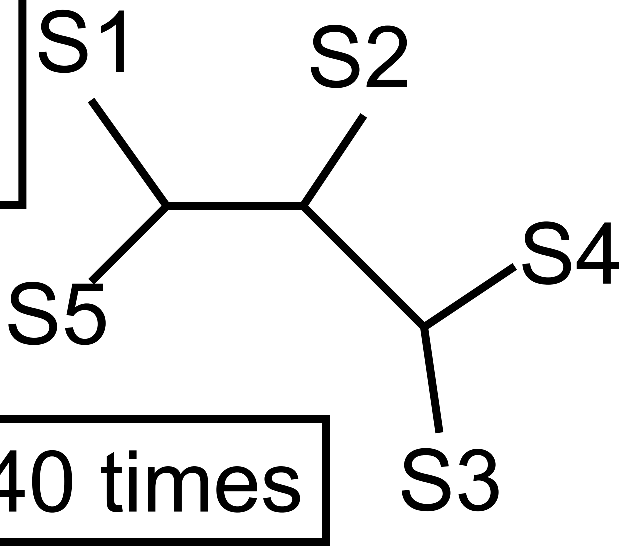


25 times

The support for the split {S1,S3}{S2,S4,S5} is 60%



5 times



40 times

Bootstrap is very slow

There’s a need for faster estimates

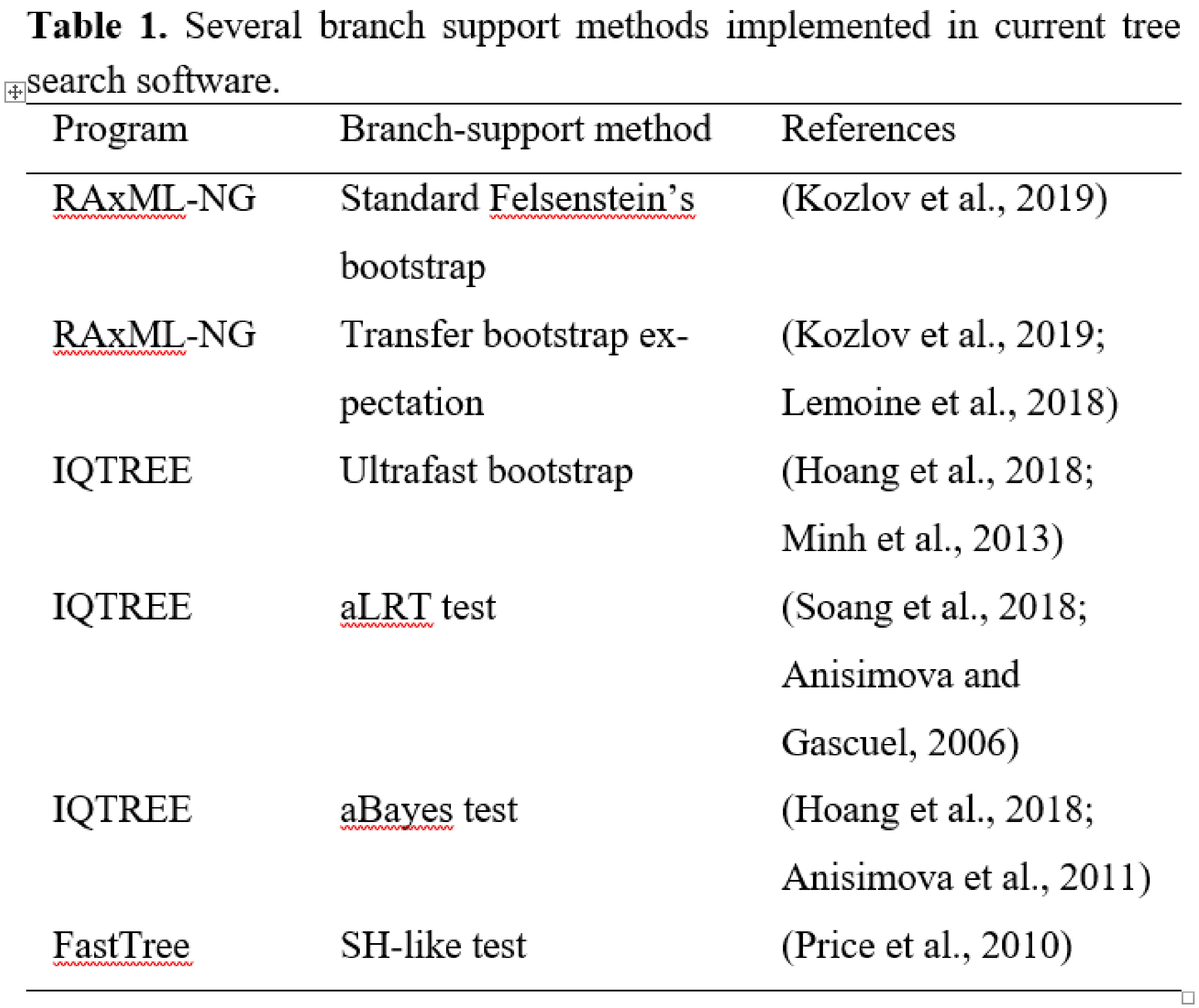
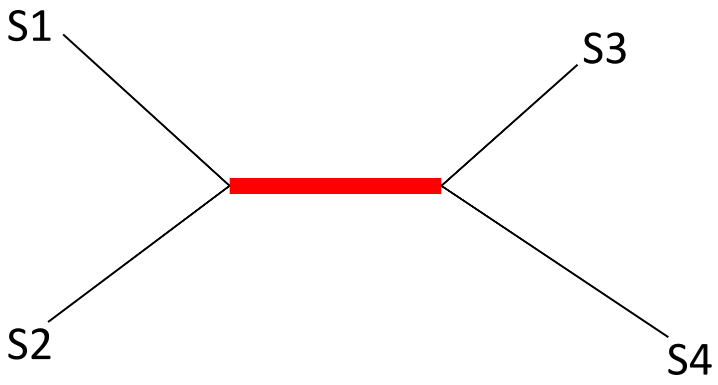
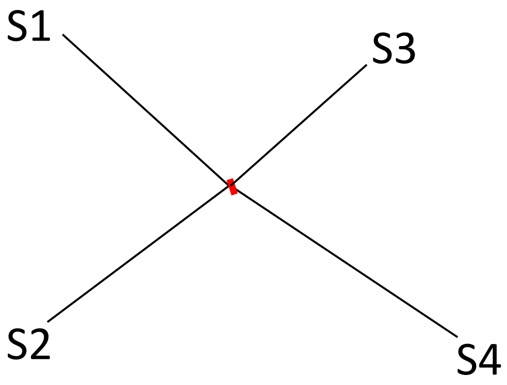


Table 1. Several branch support methods implemented in current tree search software.

Program	Branch-support method	References
RAxML-NG	Standard Felsenstein’s bootstrap	(Kozlov et al., 2019)
RAxML-NG	Transfer bootstrap expectation	(Kozlov et al., 2019; Lemoine et al., 2018)
IQTREE	Ultrafast bootstrap	(Hoang et al., 2018; Minh et al., 2013)
IQTREE	aLRT test	(Soang et al., 2018; Anisimova and Gascuel, 2006)
IQTREE	aBayes test	(Hoang et al., 2018; Anisimova et al., 2011)
FastTree	SH-like test	(Price et al., 2010)

Likelihood based methods

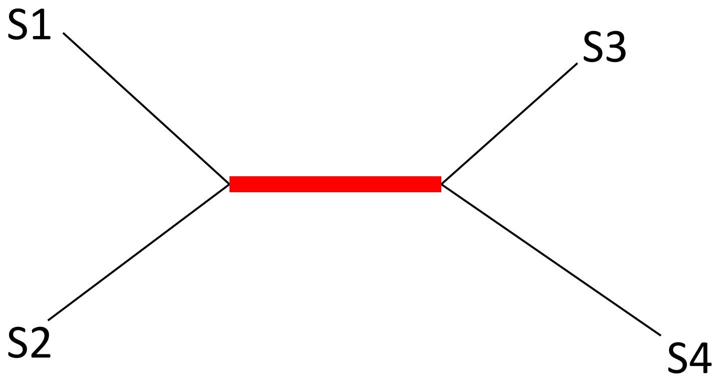




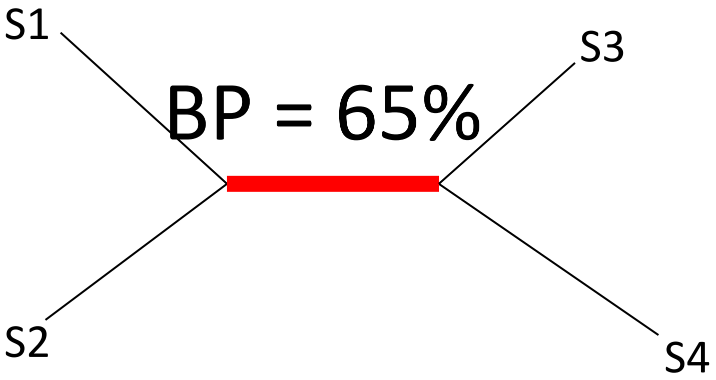
- Likelihood-based methods will estimate the support using differences in log-likelihood

How to evaluate different branch-support methods?

“TRUE TREE”

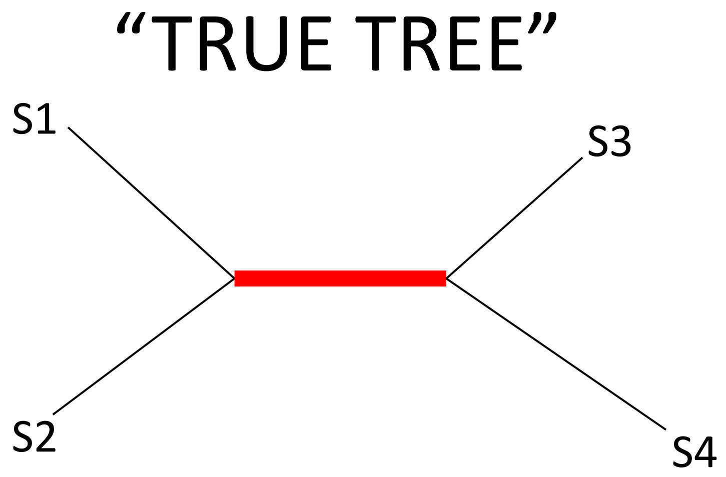


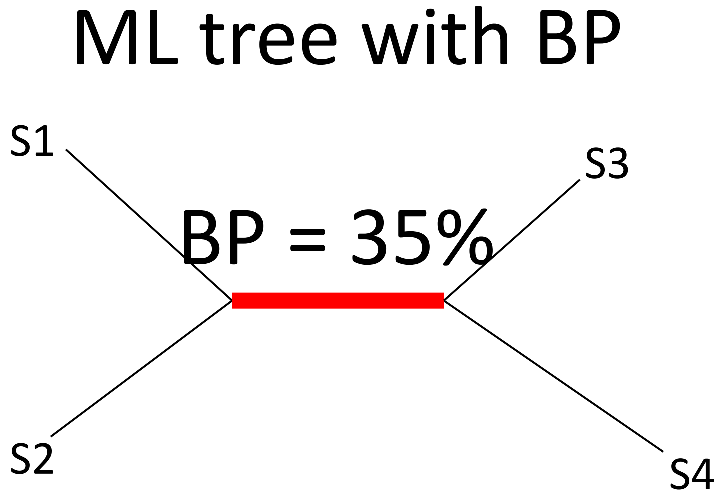
ML tree with BP



- In this case we have a true positive inference (TP)
- POSITIVE = the estimate is that the split exists
- TRUE = the estimate is correct

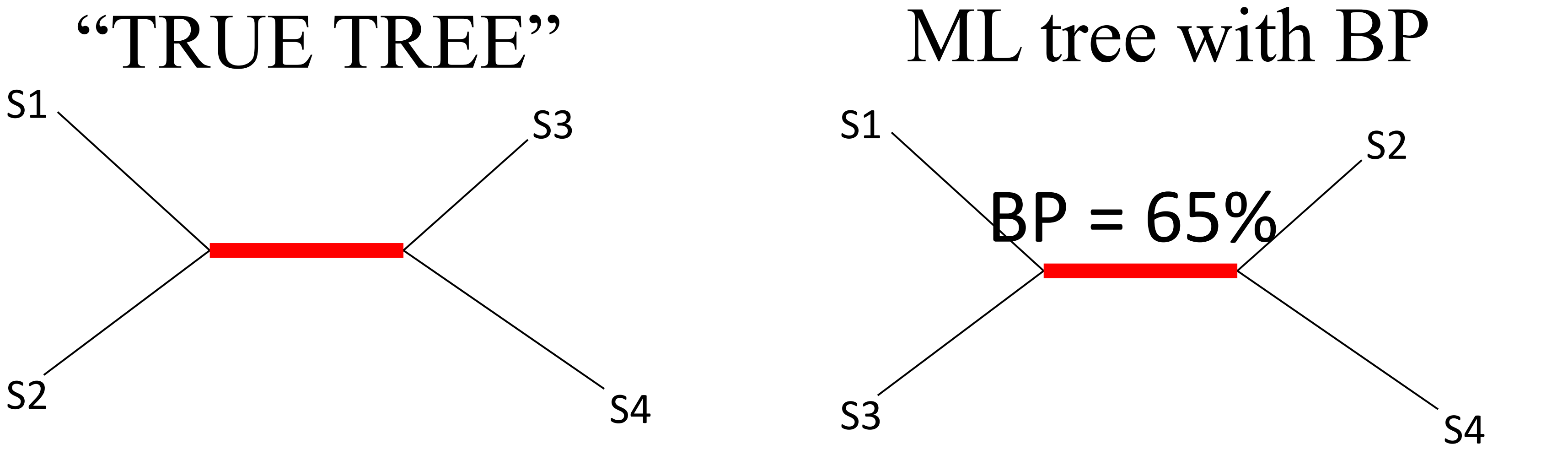
How to evaluate different branch-support methods?





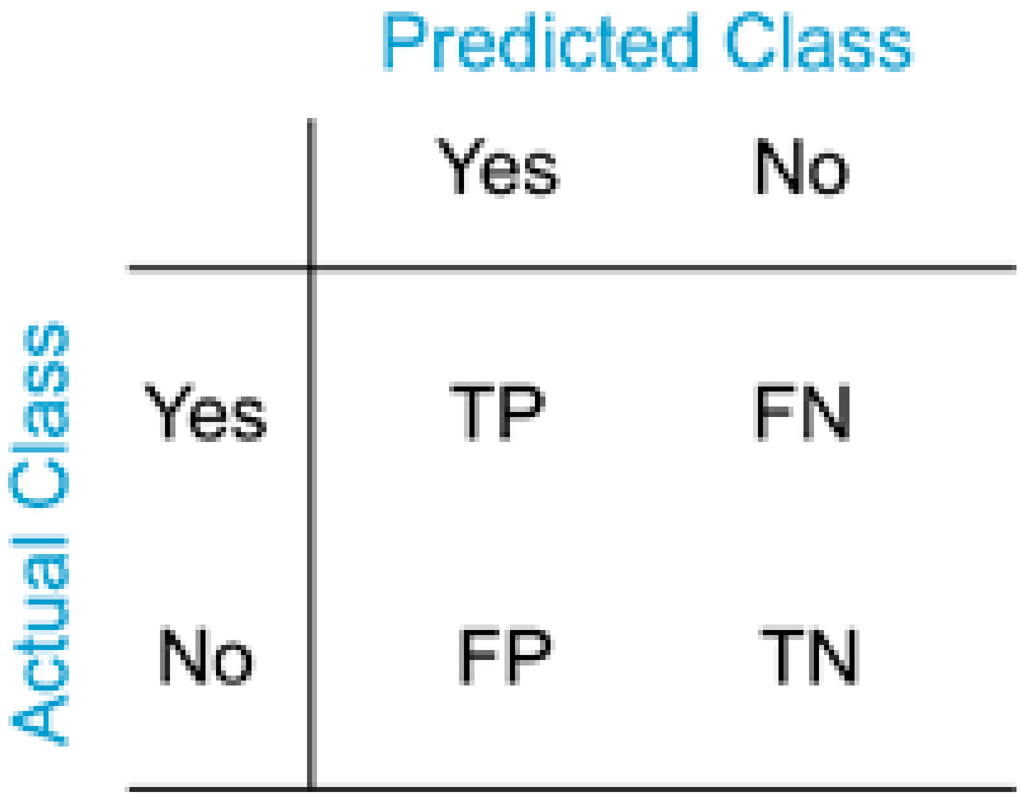
- In this case we have a false negative inference (FN)
- NEGATIVE = the estimate is that the split does not exist
- FALSE = the estimate is wrong

How to evaluate different branch-support methods?

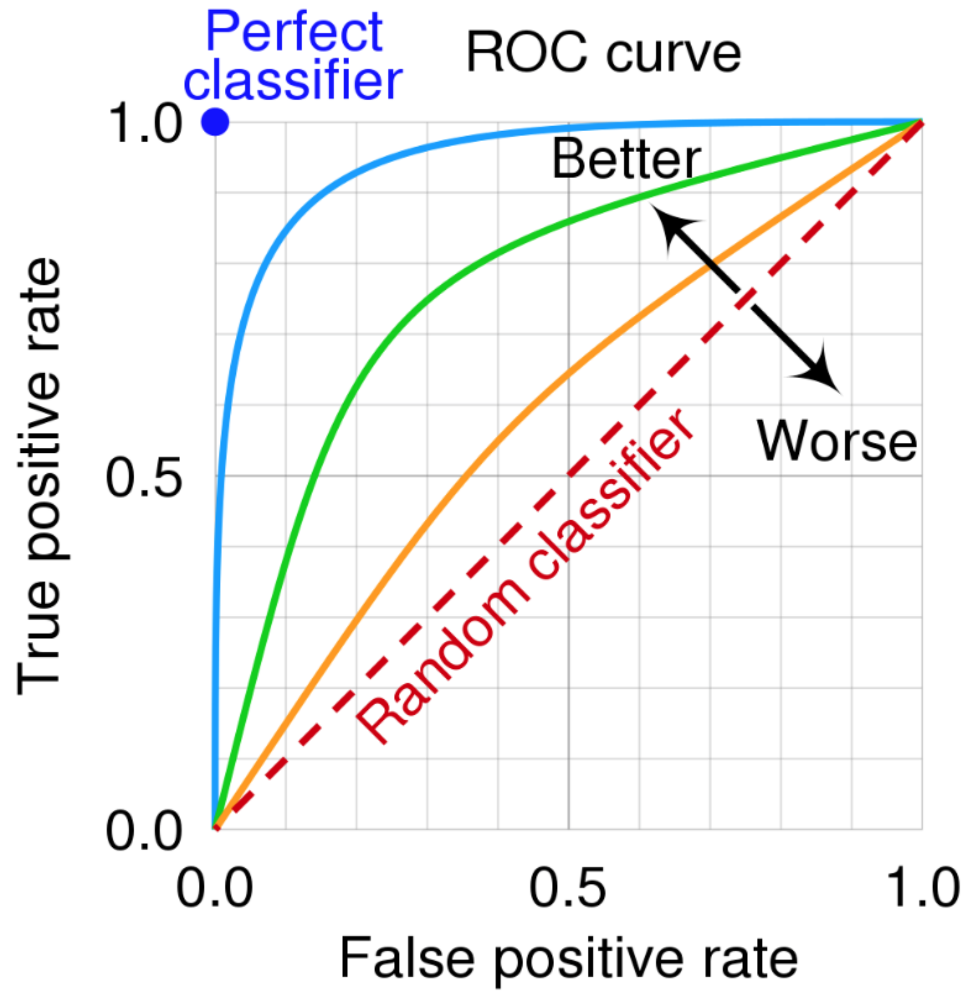


- In this case we have a false positive inference (FP)
- POSITIVE = the estimate is that the split exists
- FALSE = the estimate is wrong

We can compute confusion matrices and AUC scores



		Predicted Class	
		Yes	No
Actual Class	Yes	TP	FN
	No	FP	TN



TP/P

FP/N

ML (machine learning) for branch support values

Noa Ecker

ML (machine learning) for branch support values

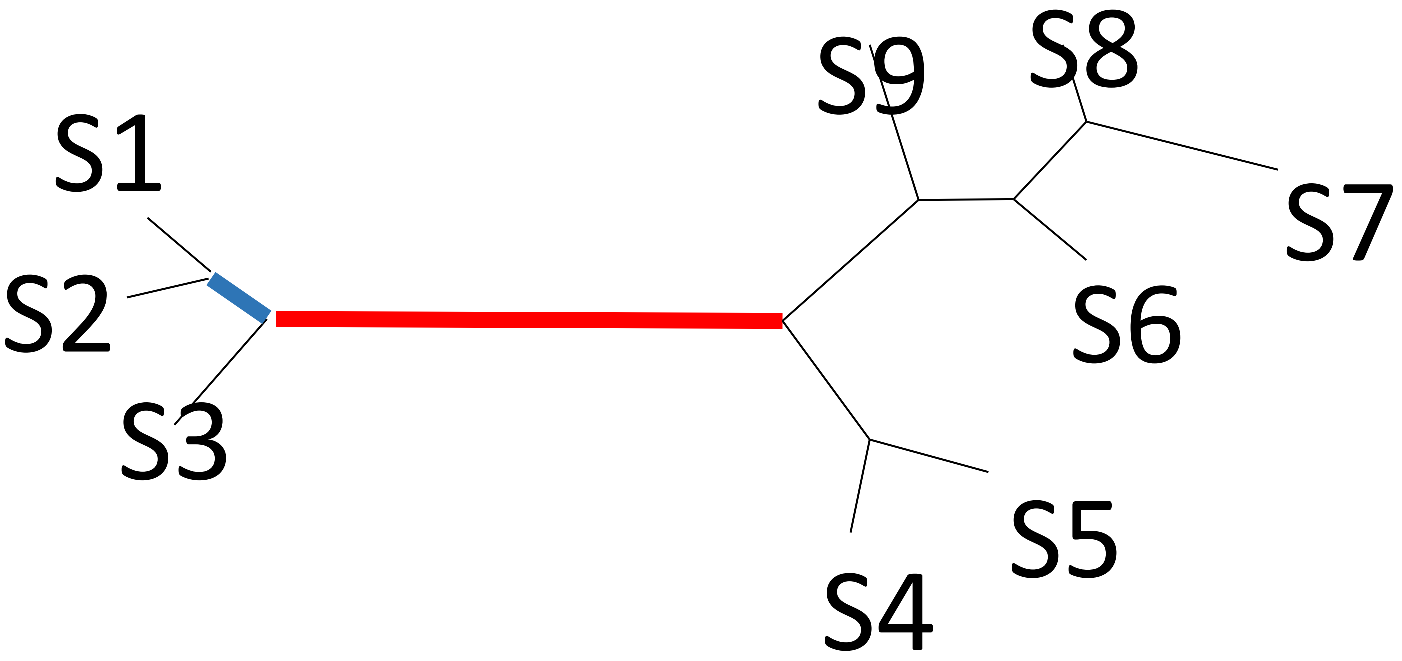
Prof. Yishay Mansour

Prof. Itay Mayrose

Prof. Dorothee Huchon

Intuition

- I would trust the red branch more than the blue one

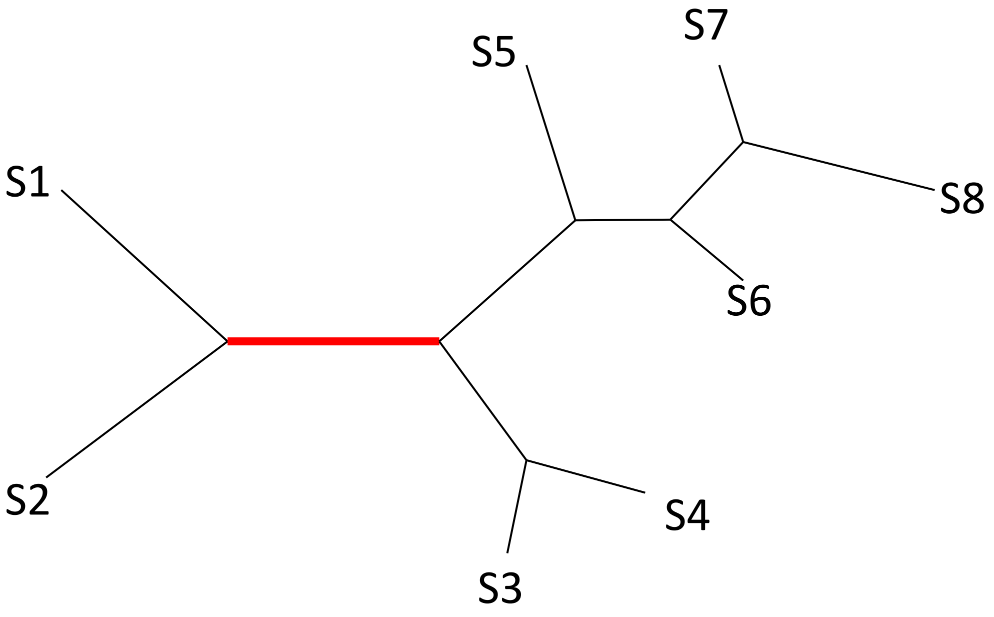


Features (out of 39)

- Branch length at the partition site
- Branch length divided by the mean branch length across the tree
- Branch length divided by the mean branch length among the four neighboring branches
- Number of MSA columns
- Number of unique MSA columns
- Percentage of constant sites
- The LL of NNI neighbors around the branch

Features (out of a 39)

- The count and proportion of taxa on the smaller or equal side of the bipartition



Proportion of taxa on the smaller side of the bipartition = 0.25

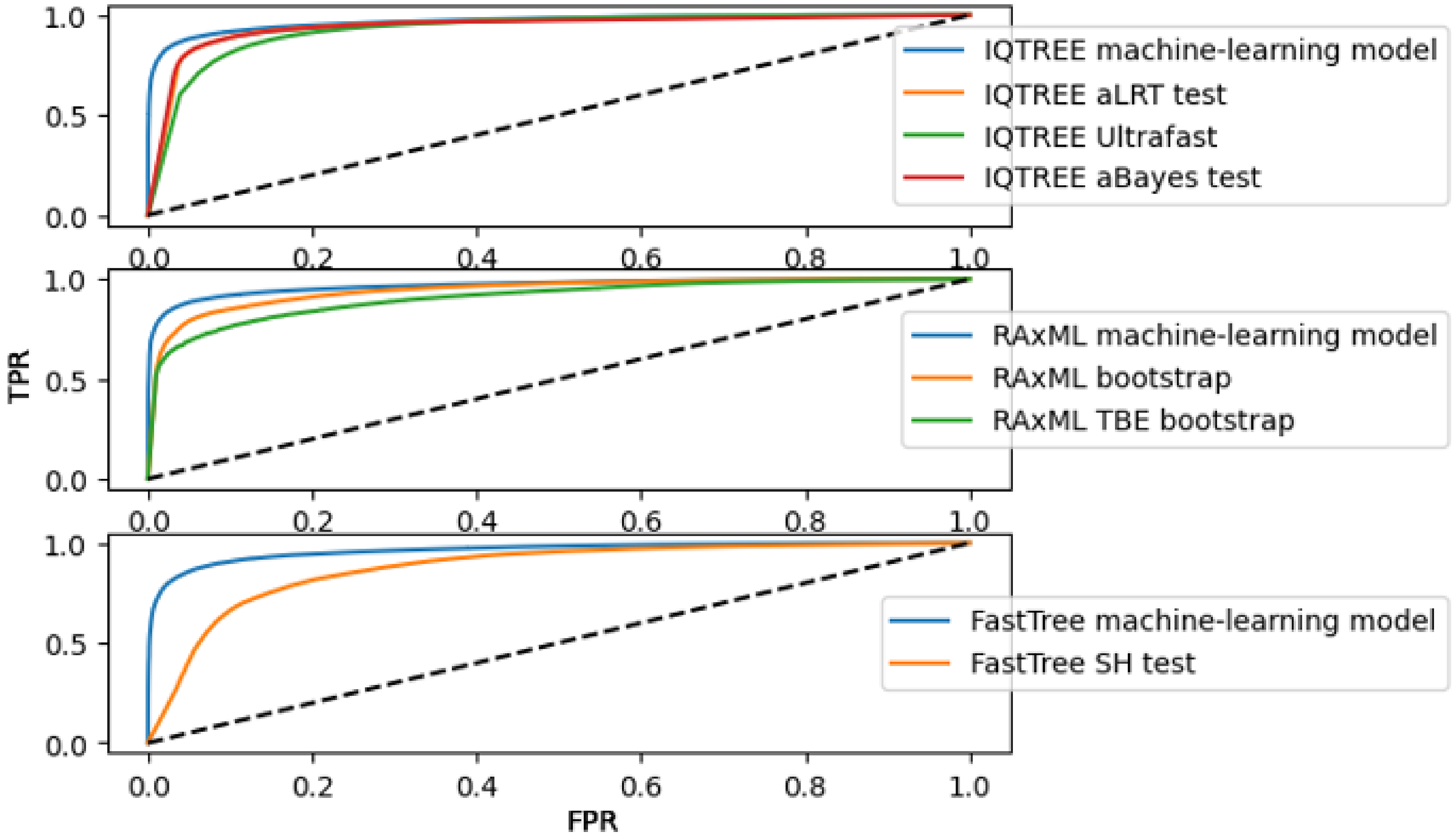
Number of taxa on the smaller side of the bipartition = 2

Train and test datasets

- 6,000 simulated MSAs with 100 to 10,000 sites and between 30 to 1,000 taxa.
- Each MSA was simulated along a different tree topology using Alisim (Ly-Trong et al., 2022), based on the script provided in the Github repository of RAxML-grove (Höhler et al., 2022).
- Each MSA was simulated using the DNA model associated with that tree in RAxML-Grove.
- Train set: 70% of the data; test set: the remaining 30%

Results, performance

- RAxML BP = 0.946 **Machine-learning = 0.968**
- RAxML transfer BP = 0.907

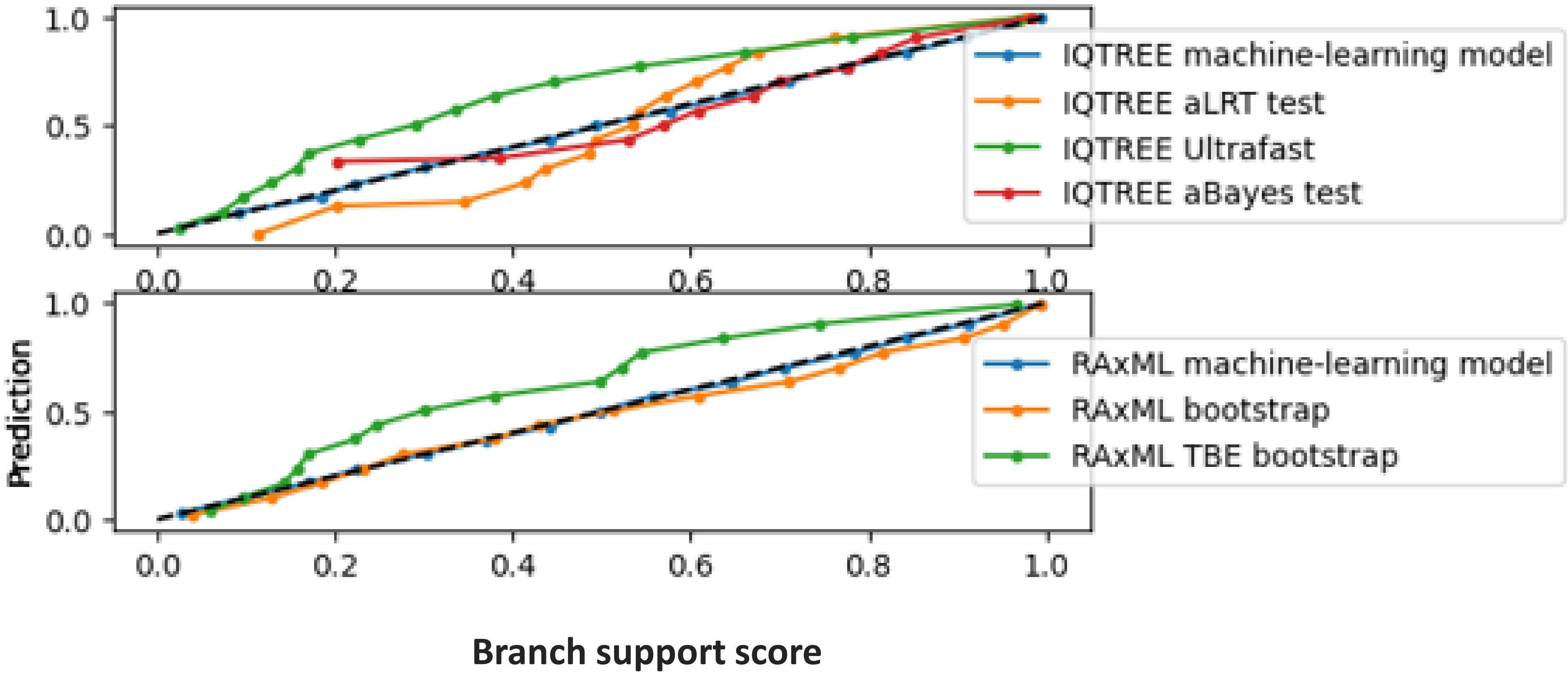


Results, running time

- **21 times faster** without optimizing the feature extraction algorithms:

The computation time of RAxML-NG standard bootstrap exhibited a median running time of 138 min on a single CPU. On the same data, our machine-learning model had a median running time of 6.5 min.

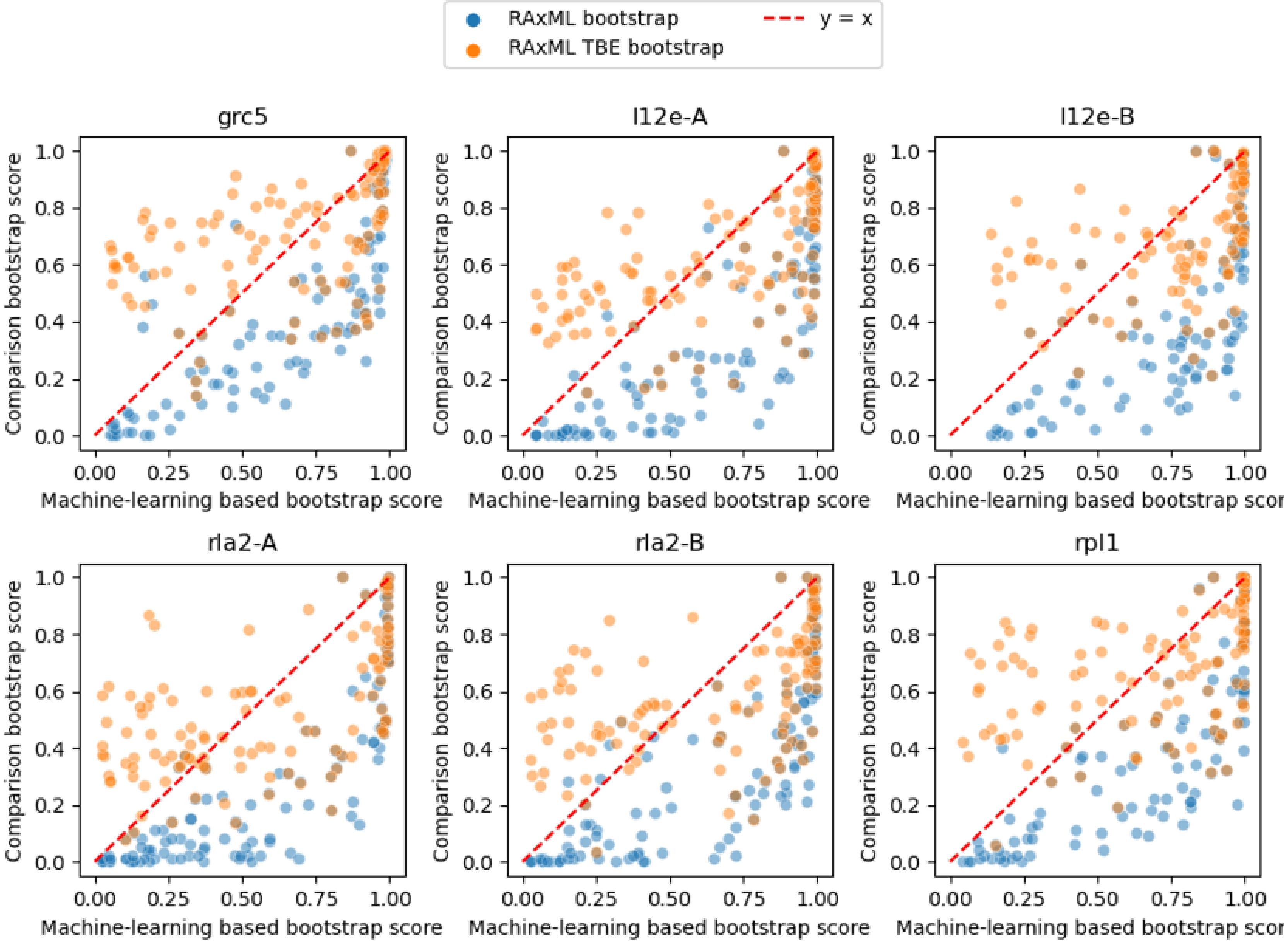
Results, calibration



Expected Calibration Error (ECE) of machine-learning method (IQTREE) = 0.002

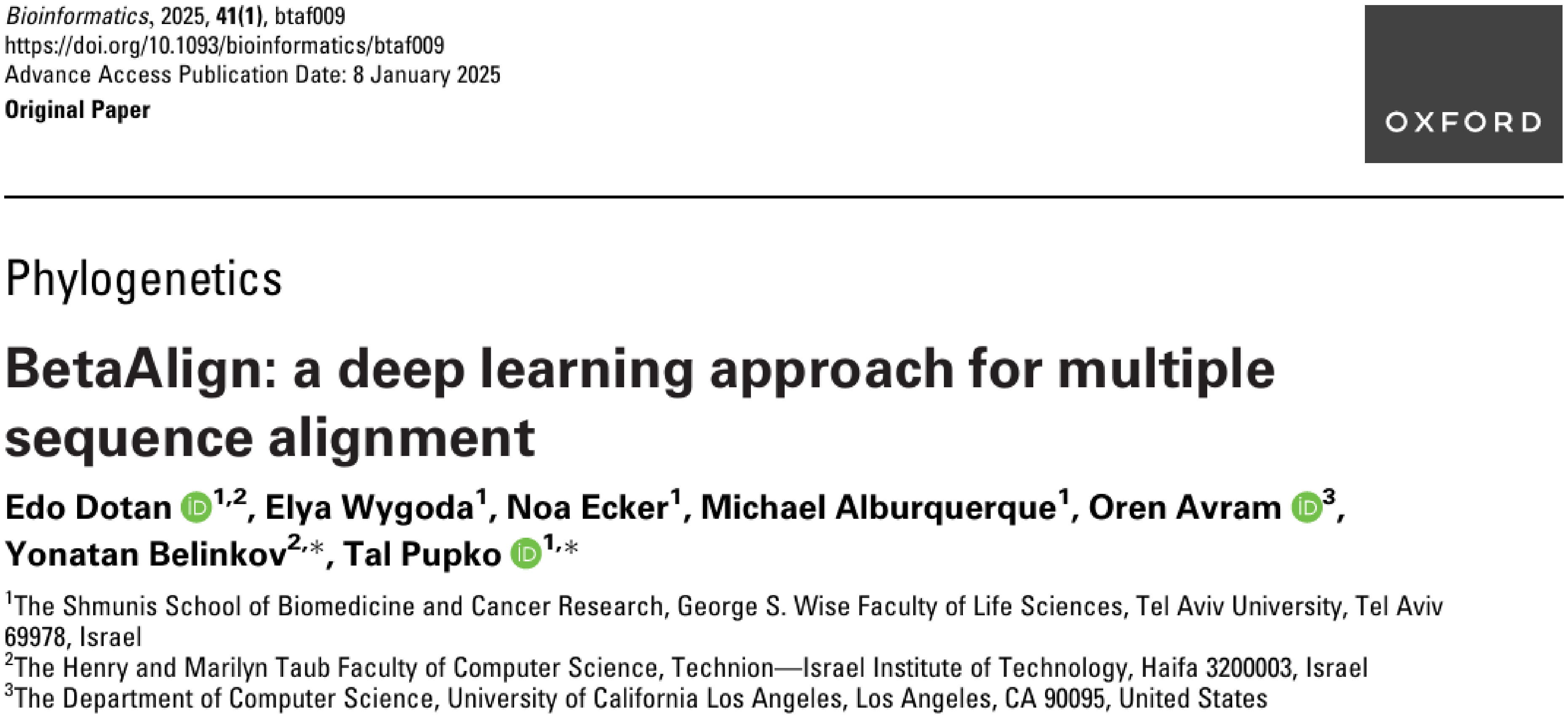
ECE for ultrafast bootstrap (IQTREE) = 0.043

Results, empirical data



AI phylogeny

PART 2:



Bioinformatics, 2025, **41(1)**, btaf009
https://doi.org/10.1093/bioinformatics/btaf009
Advance Access Publication Date: 8 January 2025
Original Paper

OXFORD

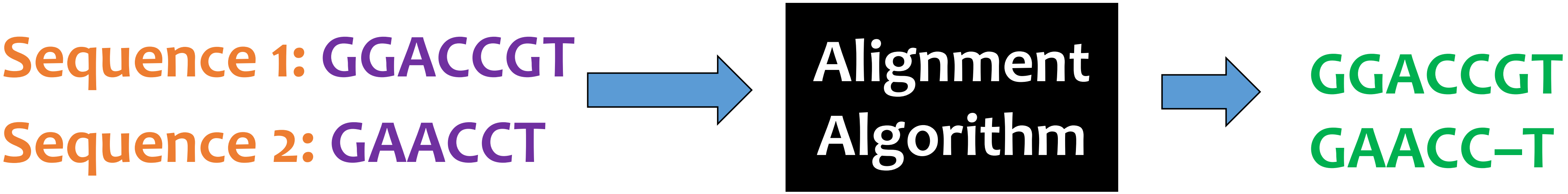
Phylogenetics

BetaAlign: a deep learning approach for multiple sequence alignment

Edo Dotan[1,2], Elya Wygoda[1], Noa Ecker[1], Michael Alburquerque[1], Oren Avram[3], Yonatan Belinkov[2,*], Tal Pupko[1,*]

[1]The Shmunis School of Biomedicine and Cancer Research, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 69978, Israel
[2]The Henry and Marilyn Taub Faculty of Computer Science, Technion—Israel Institute of Technology, Haifa 3200003, Israel
[3]The Department of Computer Science, University of California Los Angeles, Los Angeles, CA 90095, United States

NLP-based sequence alignment



Sequence 1: GGACCGT

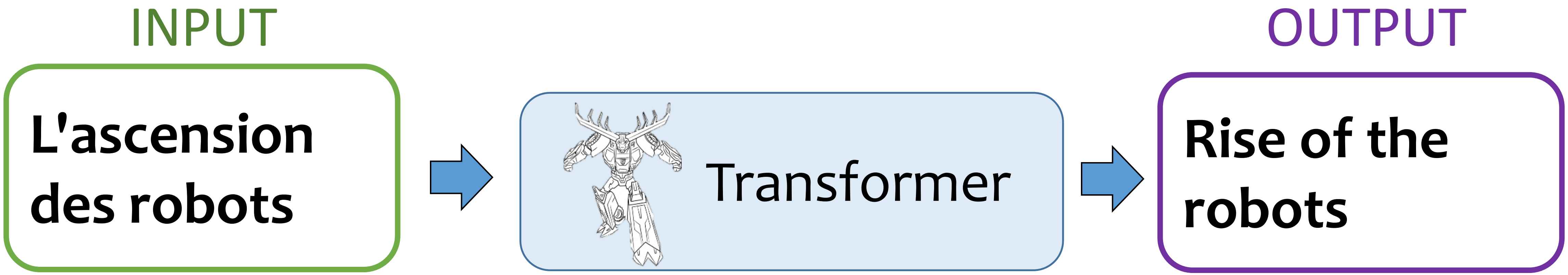
Sequence 2: GAACCT

Alignment Algorithm

GGACCGT

GAACC–T

The transformer



INPUT

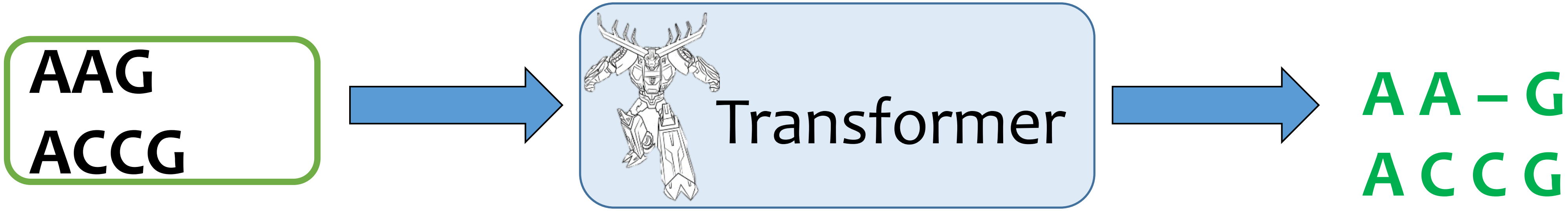
L'ascension des robots

Transformer

OUTPUT

Rise of the robots

BetaAlign



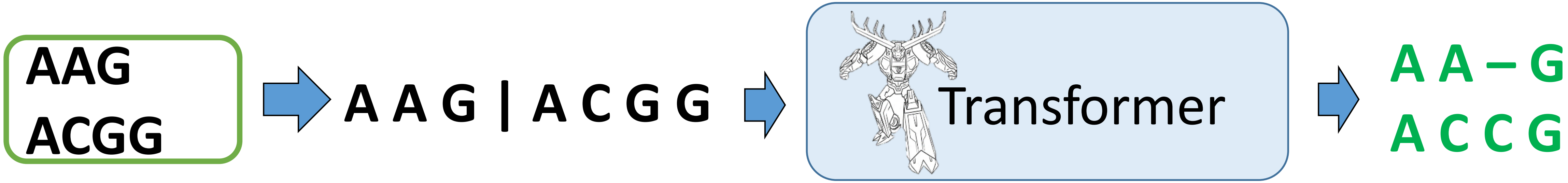
AAG
ACCG

Transformer

A A – G
A C C G

Encoding: the “concat” language

- Each letter in the “concat” language is a word, and the language has 5 different words (5 tokens)



AAG
ACGG

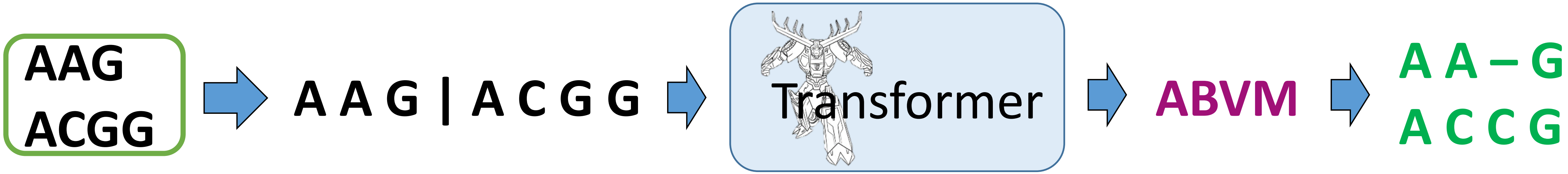
A A G | A C G G

Transformer

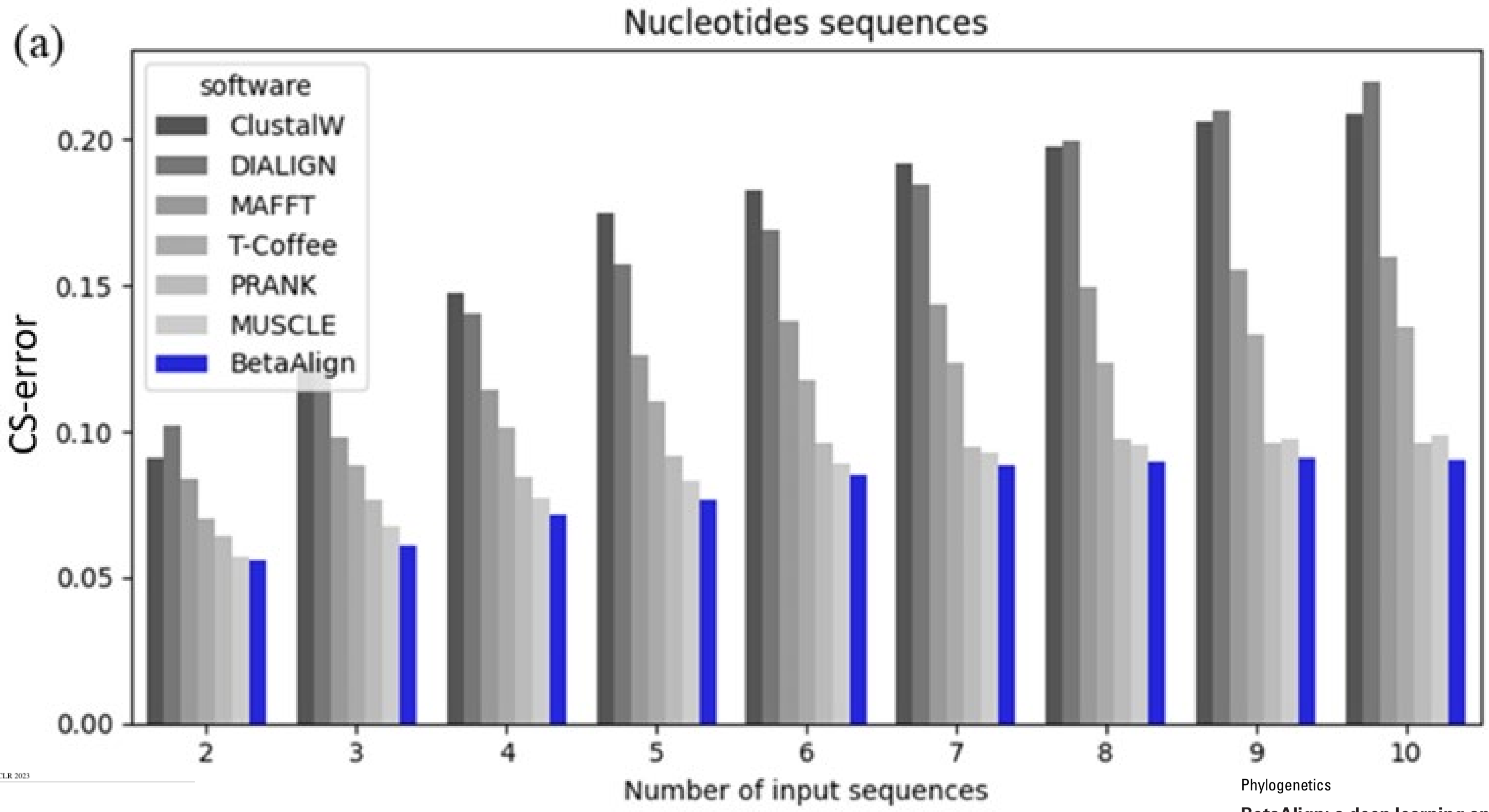
A A – G
A C C G

The output language

Upper letter	Lower letter	Symbol
A	A	A
A	C	B
A	G	C
A	T	D
A	-	E
…	…	…
-	T	X



Performance



Published as a conference paper at ICLR 2023

MULTIPLE SEQUENCE ALIGNMENT AS A SEQUENCE-TO-SEQUENCE LEARNING PROBLEM

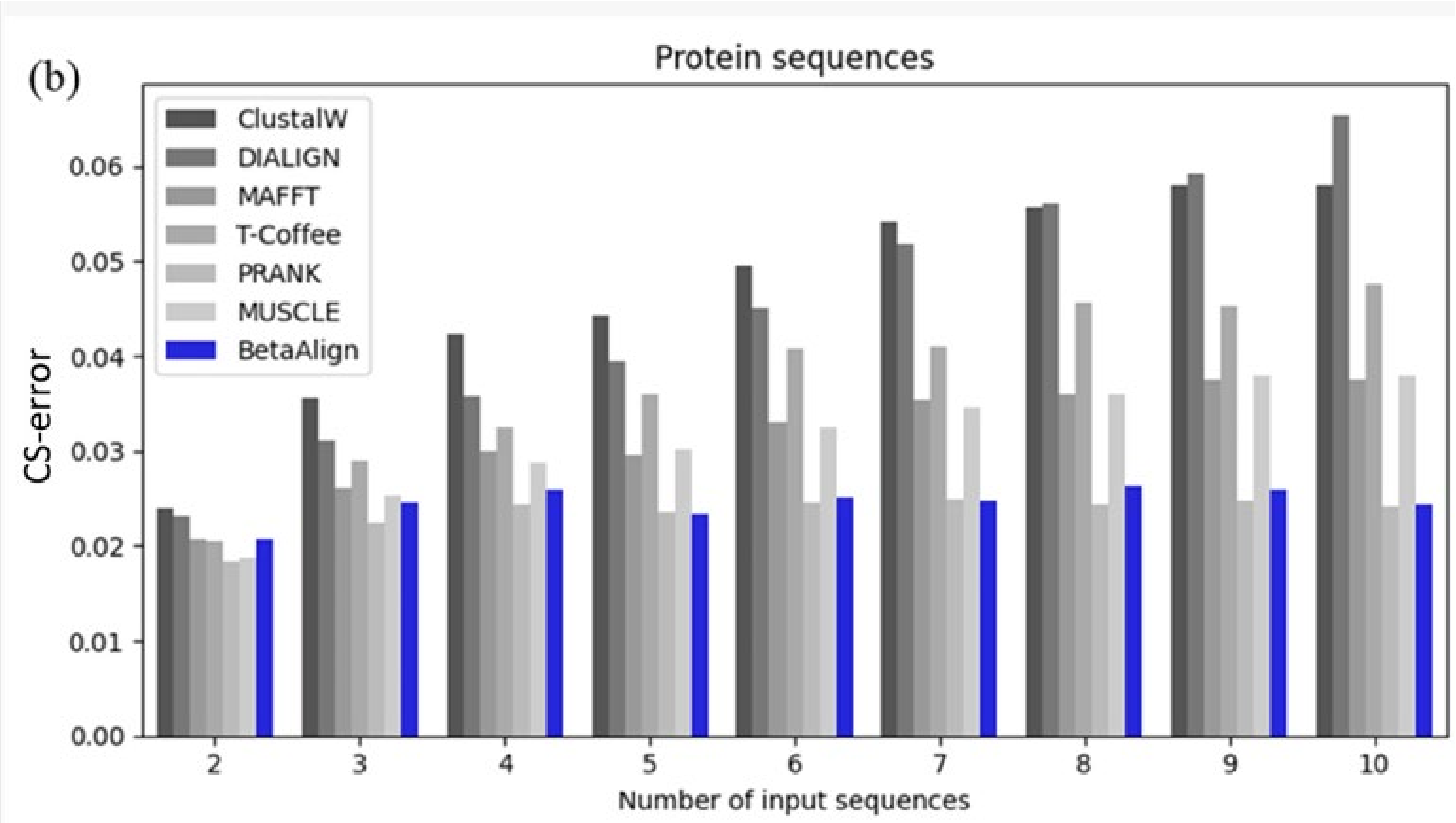
Edo Dotan[1‡], Yonatan Belinkov[2†*], Oren Avram[3], Elya Wygoda[1], Noa Ecker[1], Michael Alburquerque[1], Omri Keren[1], Gil Loewenthal[1], and Tal Pupko[1†]

Phylogenetics

BetaAlign: a deep learning approach for multiple sequence alignment

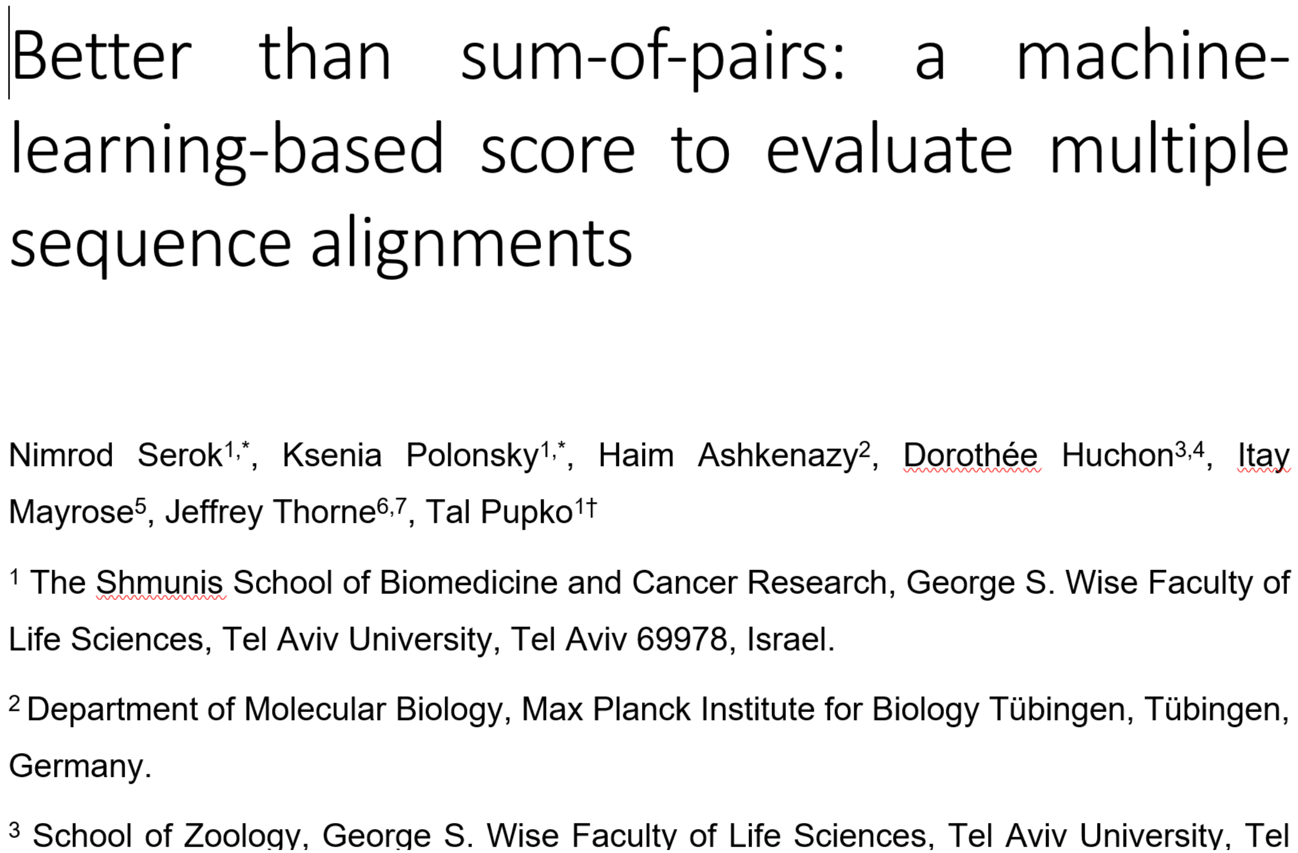
Edo Dotan[1,2], Elya Wygoda[1], Noa Ecker[1], Michael Alburquerque[1], Oren Avram[3], Yonatan Belinkov[2,*], Tal Pupko[1,*]

Performance



AI phylogeny

PART 3 (work in progress):



Better than sum-of-pairs: a machine-learning-based score to evaluate multiple sequence alignments

Nimrod Serok[1,*], Ksenia Polonsky[1,*], Haim Ashkenazy[2], Dorothée Huchon[3,4], Itay Mayrose[5], Jeffrey Thorne[6,7], Tal Pupko[1†]

[1] The Shmunis School of Biomedicine and Cancer Research, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 69978, Israel.

[2] Department of Molecular Biology, Max Planck Institute for Biology Tübingen, Tübingen, Germany.

[3] School of Zoology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel

An ML-based MSA objective function

- Inference of MSAs is a very difficult problem
  - The implicitly assumed indel evolutionary models are oversimplified
  - Which objective function should be optimized (the likelihood are very difficult to compute)
  - Optimizing the objective function is difficult
  - The MSA depends on the tree and vice versa

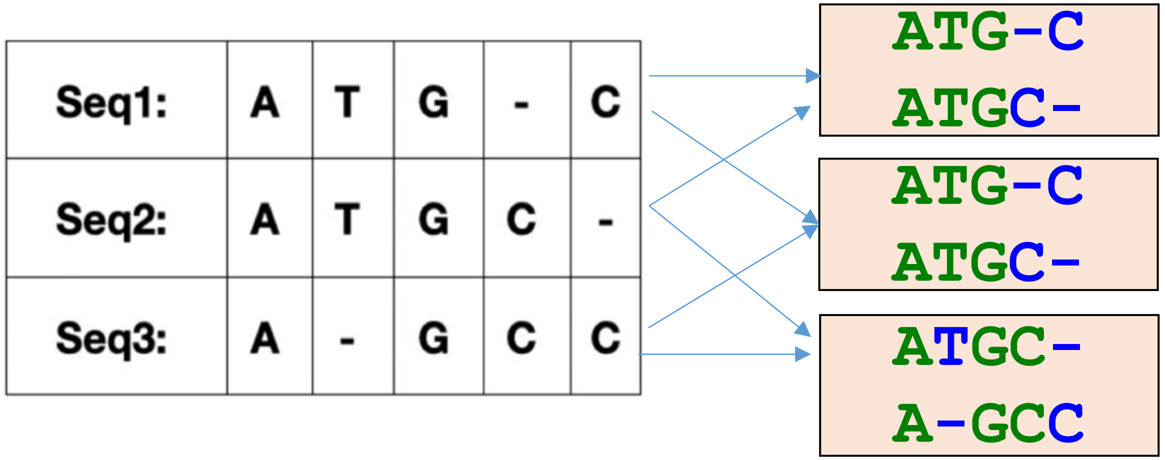
An ML-based MSA objective function

- Inference of MSAs is a very difficult problem
  - The implicitly assumed indel evolutionary models are oversimplified
  - **Which objective function should be optimized (the likelihood are very difficult to compute)**
  - Optimizing the objective function is difficult
  - The MSA depends on the tree and vice versa

Sum of pairs

- The sum-of-pairs (SoP) score is widely used for MSA scoring
- Higher SoP is supposed to be an indicator of a better MSA

Naïve scoring:
Mismatch: -1
Indel (gap): -1
Perfect match: +1



Seq1:	A	T	G	-	C
Seq2:	A	T	G	C	-
Seq3:	A	-	G	C	C

```
ATG-C
ATGC-
```

```
ATG-C
ATGC-
```

```
ATGC-
A-GCC
```

1 + 1 + 1 – 1 – 1 = 1

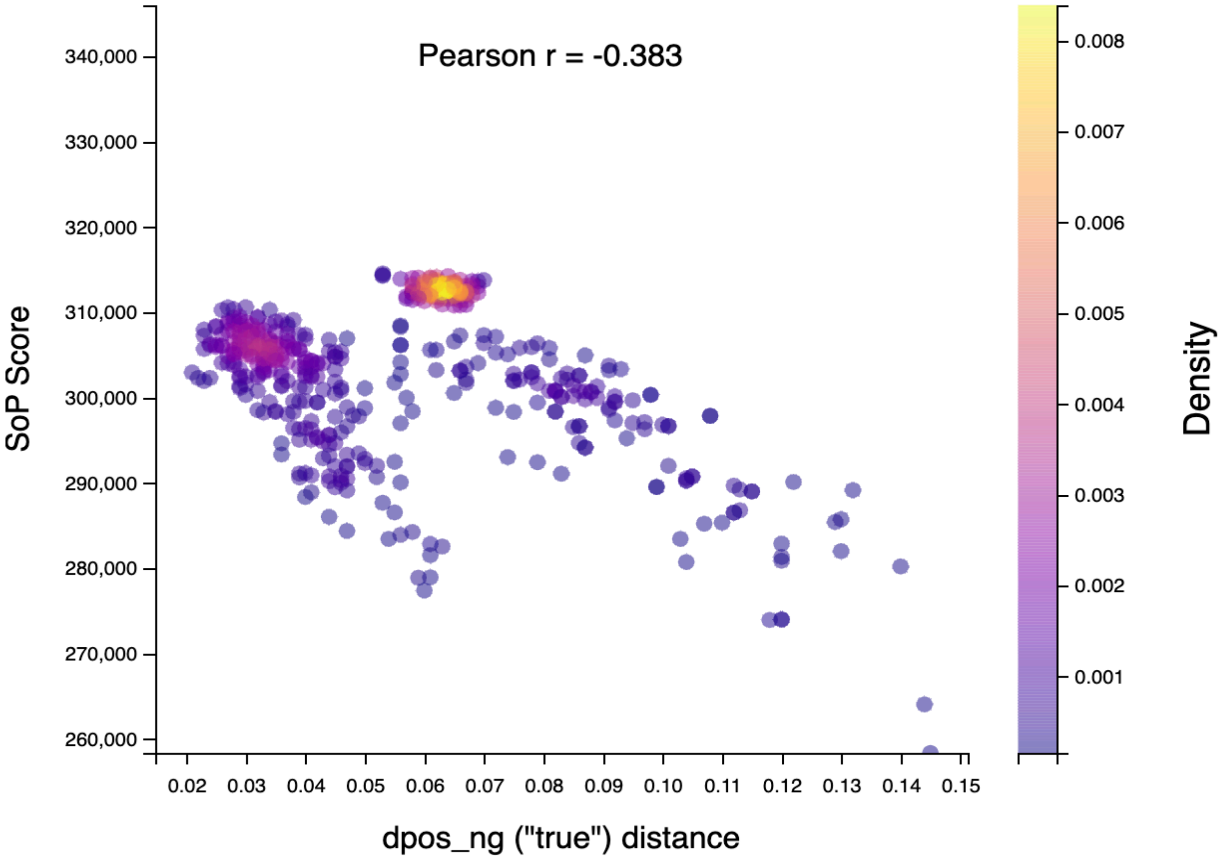
1 – 1 + 1 – 1 + 1 = 1

1 – 1 + 1 + 1 – 1 = 1

SoP = 1 + 1 + 1 = 3

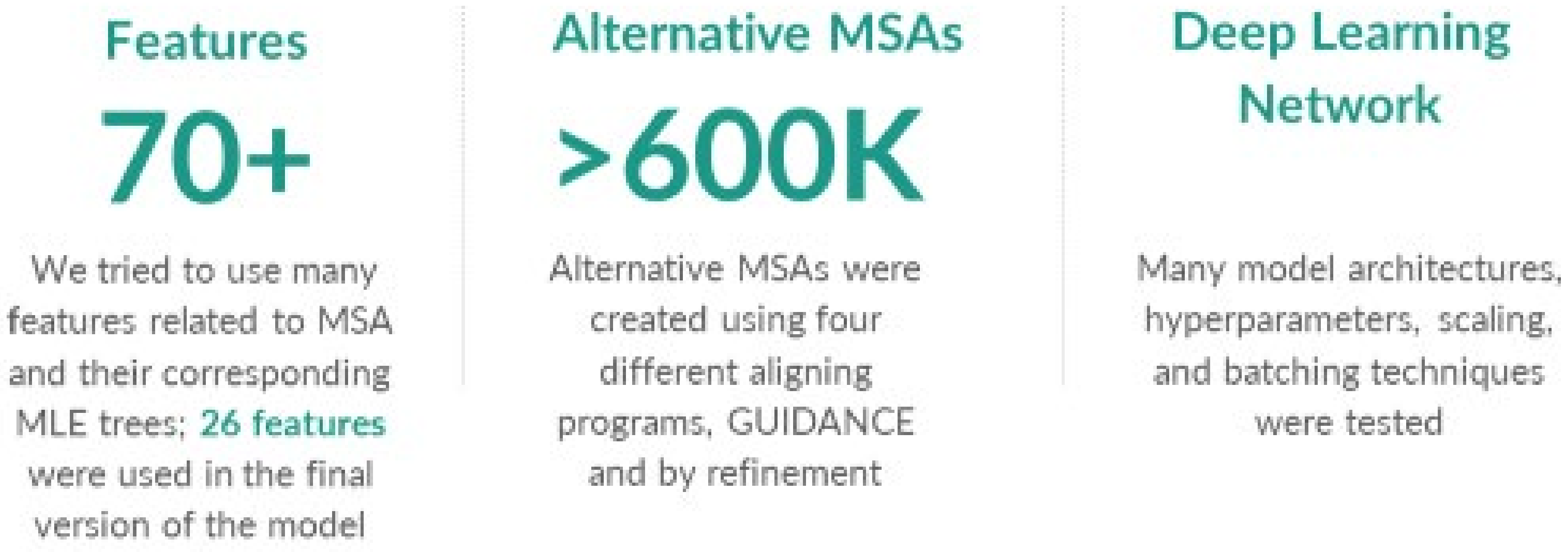
Sum of pairs does not corelate well with accuracy

- 500 alternative MSAs of a single dataset



Proposed solution

- Employing AI to develop novel scoring functions for MSAs that are better than the sum-of-pairs and to use our developed score to discriminate among MSAs



Features

70+

We tried to use many features related to MSA and their corresponding MLE trees; 26 features were used in the final version of the model

Alternative MSAs

>600K

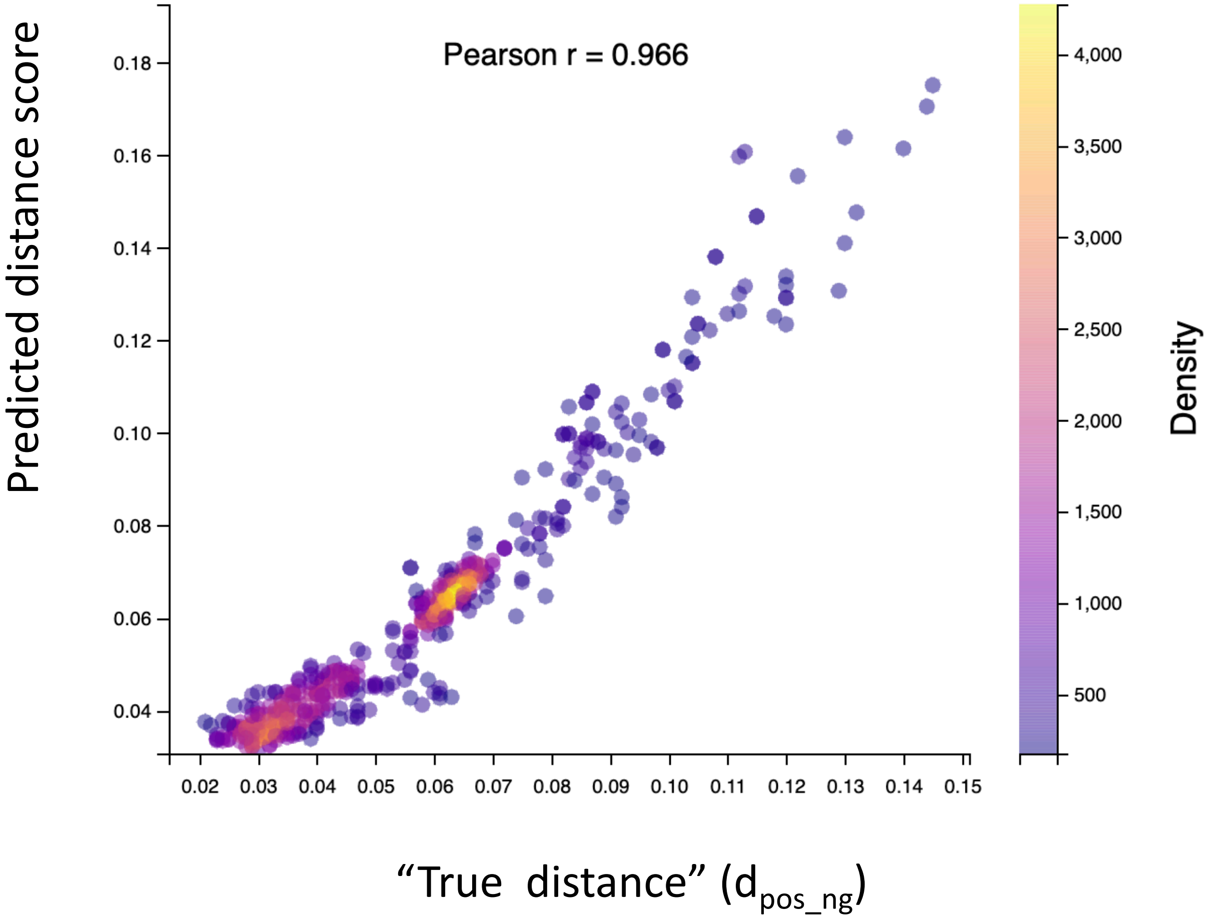
Alternative MSAs were created using four different aligning programs, GUIDANCE and by refinement

Deep Learning Network

Many model architectures, hyperparameters, scaling, and batching techniques were tested

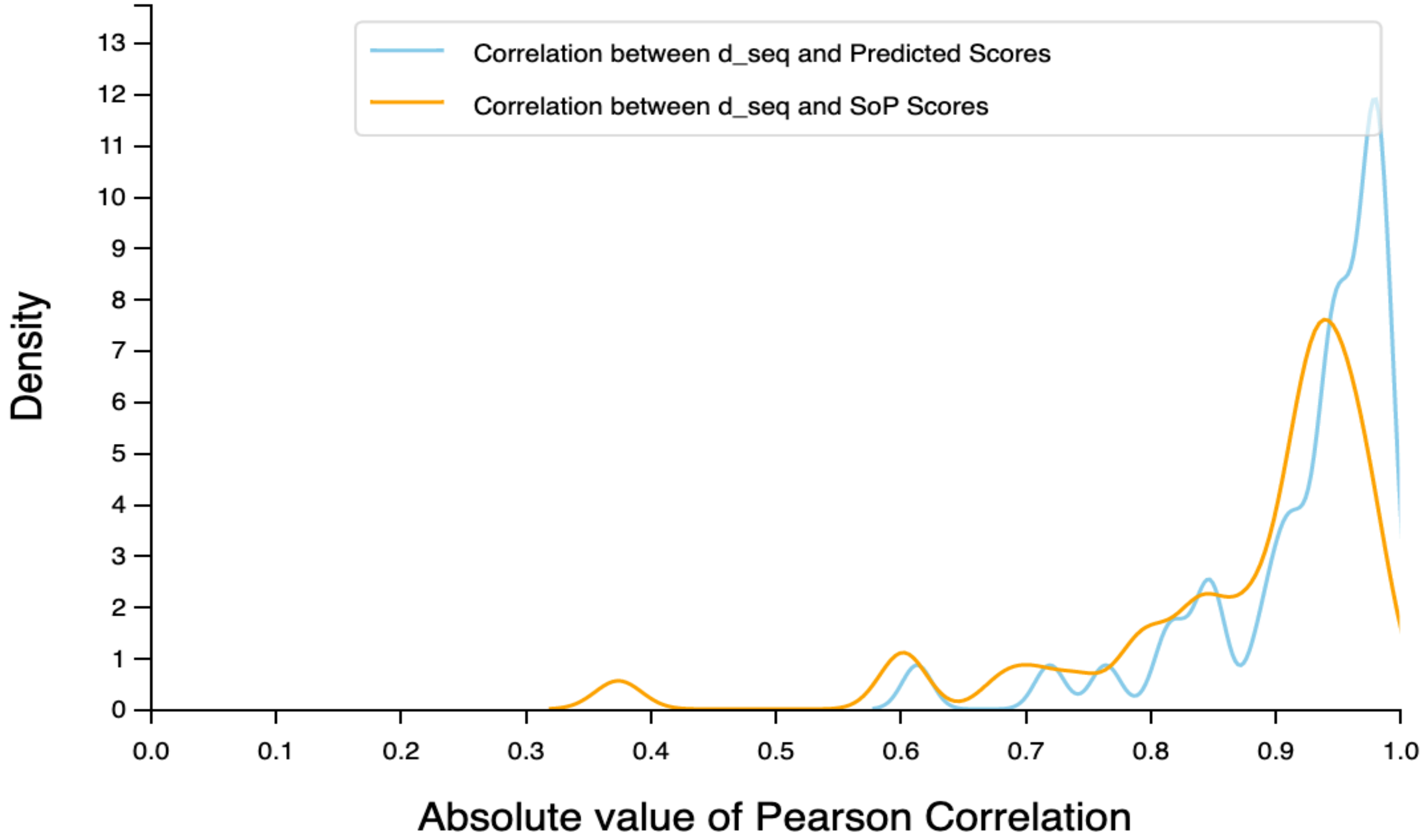
Our score well corelates with accuracy

- 500 alternative MSAs of a single dataset



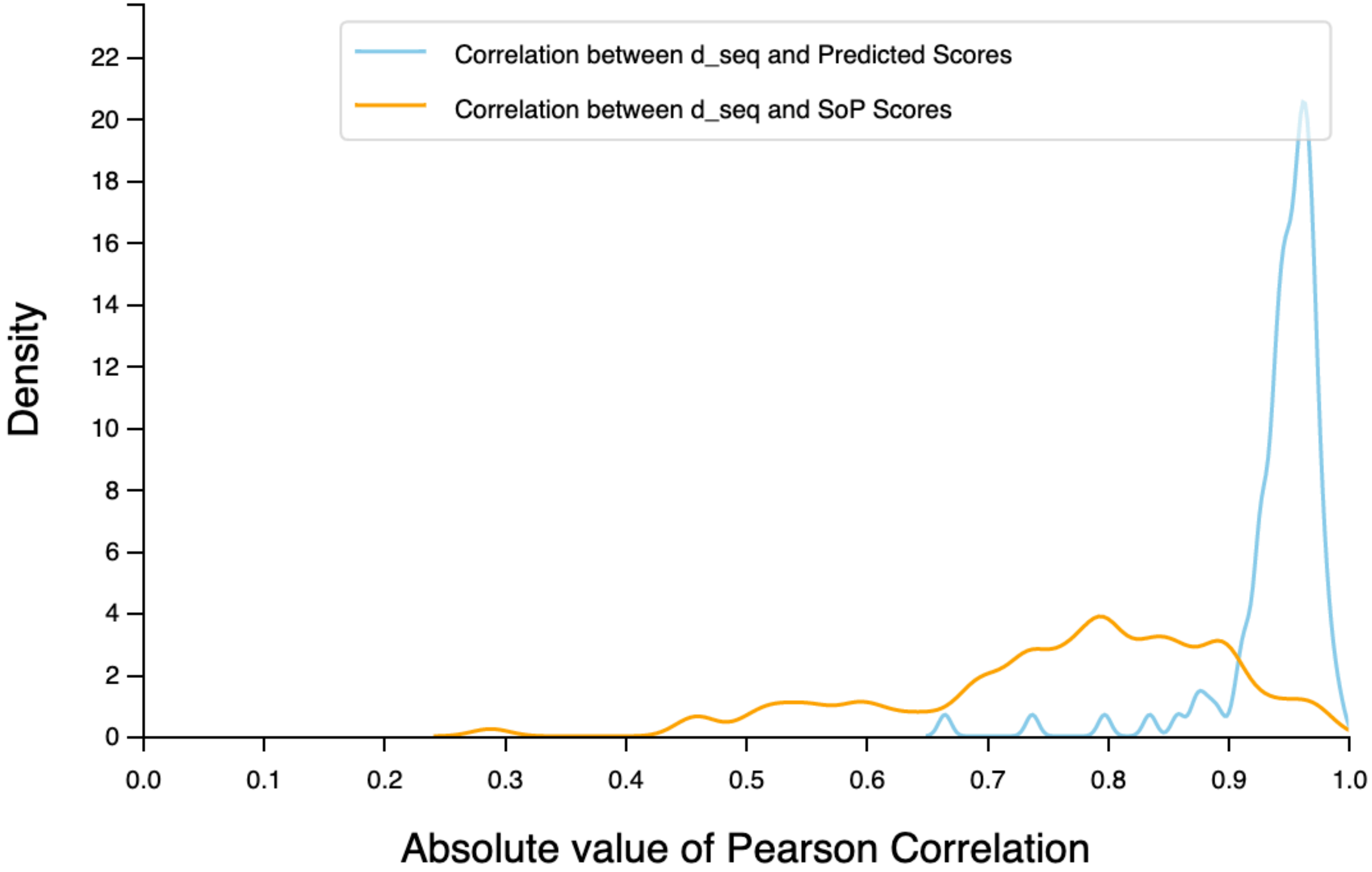
Our score well corelates with accuracy

- Simulated datasets



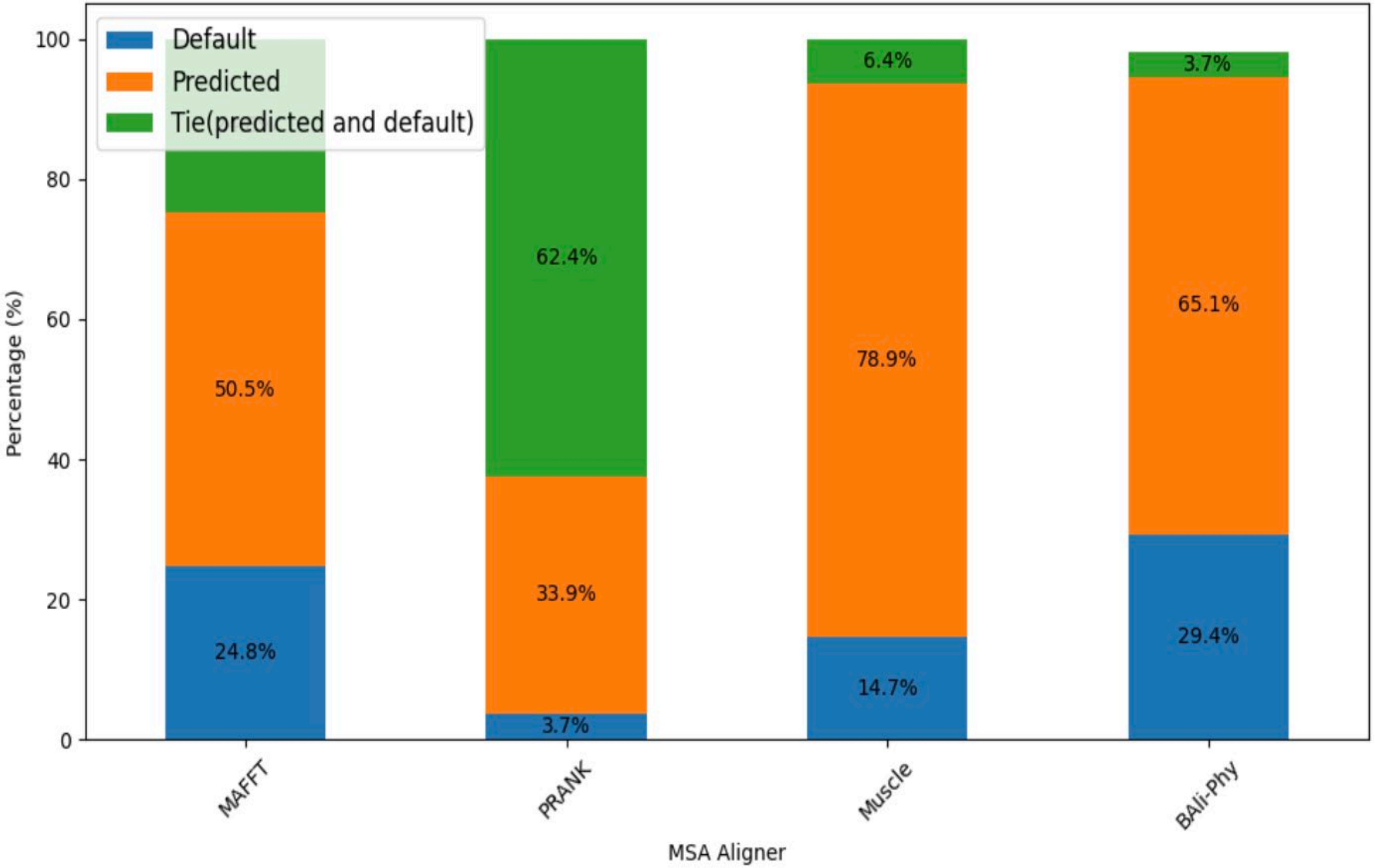
Our score well corelates with accuracy

- empirical datasets



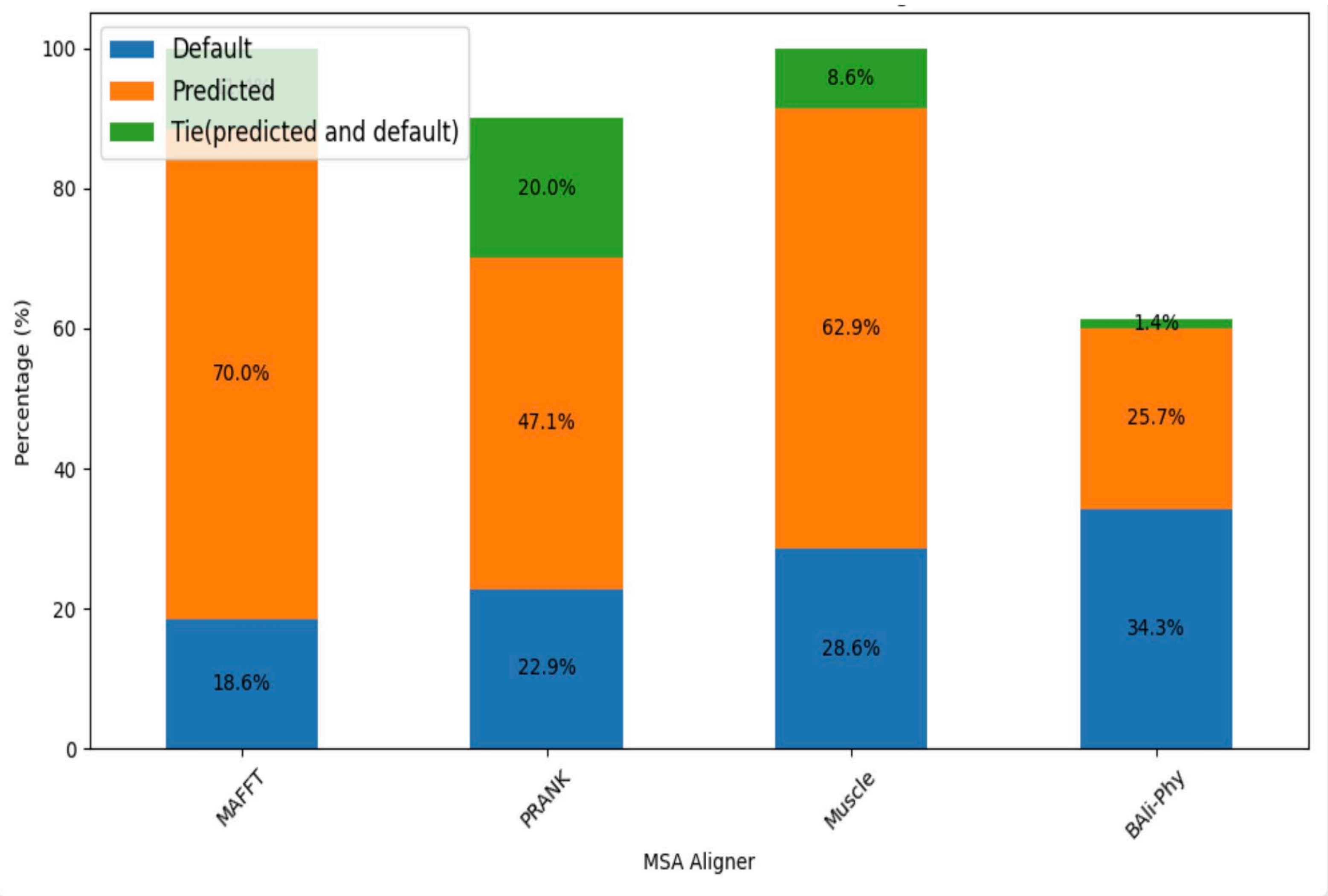
Our score well corelates with accuracy

- Pick me game (simulations)



Our score well corelates with accuracy

- Pick me game (empirical)



Joint work

Prof. Yishay Mansour

Prof. Itay Mayrose

Dr. Yonathan Belinkov

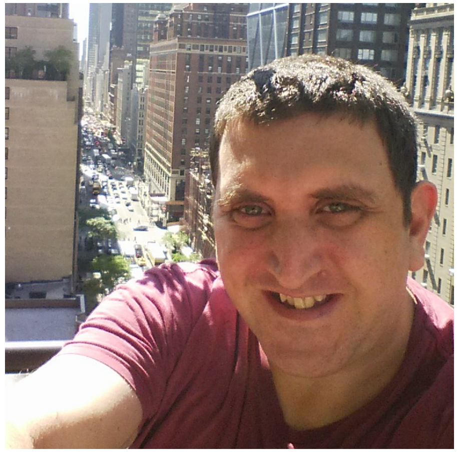
Prof. Dorothee Huchon

Prof. Jeffrey Thorne

Noa Ecker

Edo Dotan

Dr. Nimrod Serok



Dr. Haim Ashkenazy

Ksenia Polonsky

Acknowledgements

Faculty: Itay Mayrose, Dorothee Huchon, Yishay Mansur

Lab members: Naama Wagner, Nimrod Serok, Noa Ecker, Elya Wygoda, Edo Dotan, Ksenia Polonsky, Yair Shimony, Naiel Jabareen, Noam Bracha, and Boris Trayvas

הקרן הלאומית למדע
ISRAEL SCIENCE FOUNDATION

TAD
The AI and Data Science Center
Tel Aviv University

15 ANNIVERSARY
Edmond J. Safra
Center for Bioinformatics
TEL AVIV UNIVERSITY אוניברסיטת תל אביב