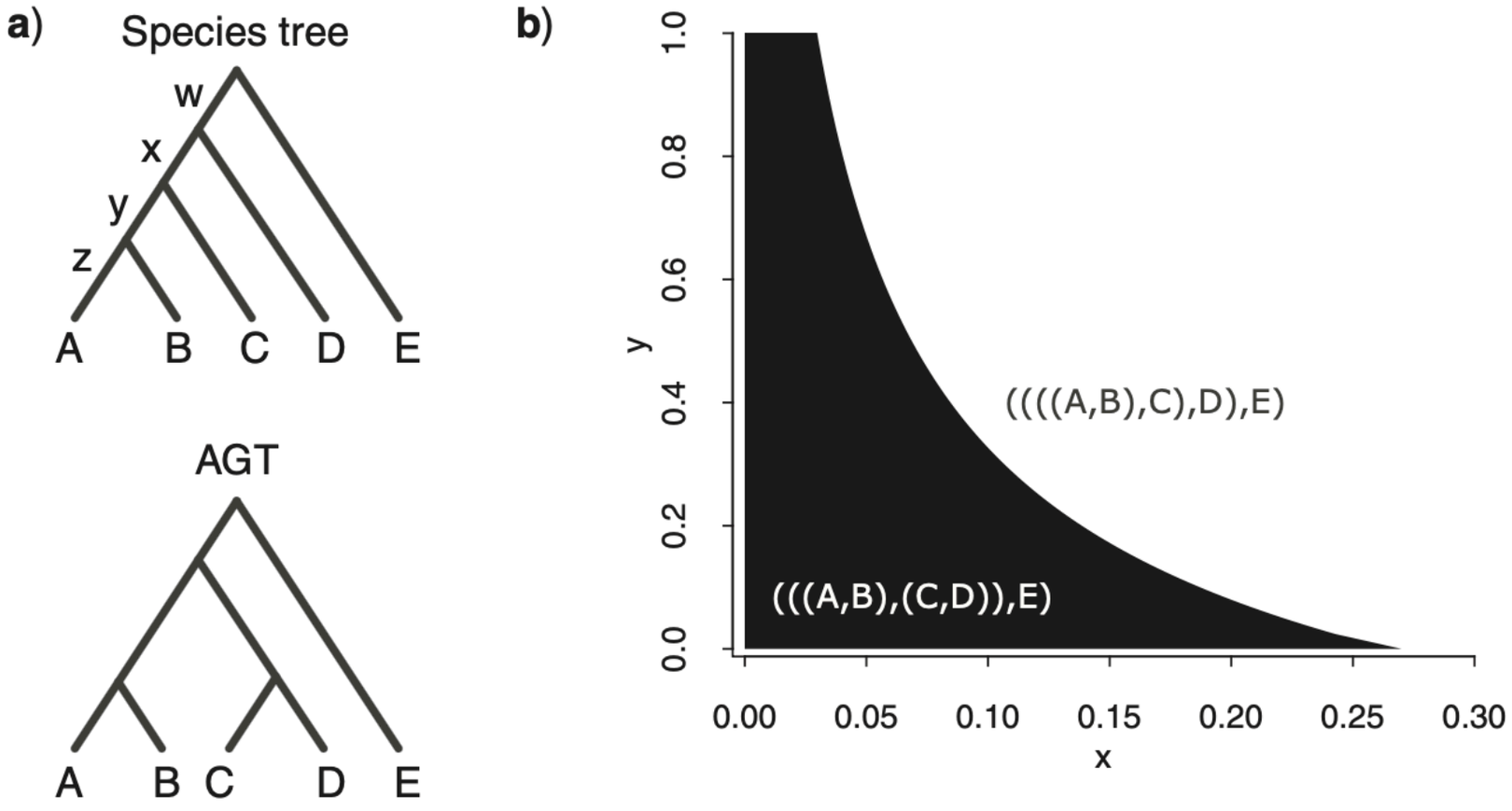
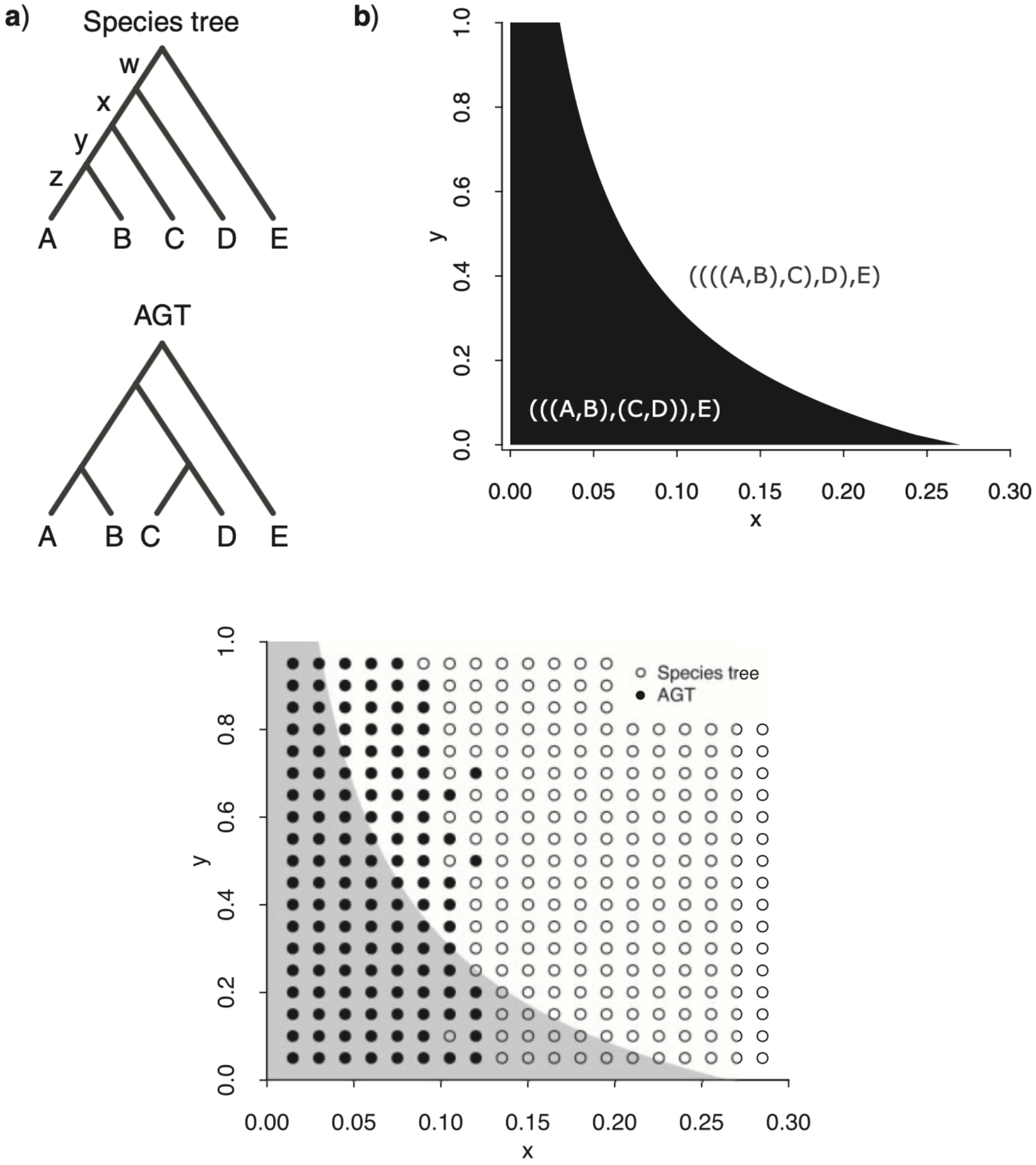

a)
Species tree
w
x
y
z
A
B
C
D
E
AGT
A
B
C
D
E
b)
1.0
0.8
0.6
0.4
0.2
0.0
y
((((A,B),C),D),E)
(((A,B),(C,D)),E)
0.00
0.05
0.10
0.15
0.20
0.25
0.30
x


a)
Species tree
w
x
y
z
A
B
C
D
E
AGT
A
B
C
D
E
b)
y
1.0
0.8
0.6
0.4
0.2
0.0
((((A,B),C),D),E)
(((A,B),(C,D)),E)
0.00
0.05
0.10
0.15
0.20
0.25
0.30
x
Species tree
AGT
y
1.0
0.8
0.6
0.4
0.2
0.0
0.00
0.05
0.10
0.15
0.20
0.25
0.30
x

JOURNAL ARTICLE

Evaluating Statistical Multiple Sequence Alignment in Comparison to Other Alignment Methods on Protein Data Sets

Michael Nute , Ehsan Saleh , Tandy Warnow

Systematic Biology, Volume 68, Issue 3, May 2019, Pages 396–411, https://doi.org/10.1093/sysbio/syy068

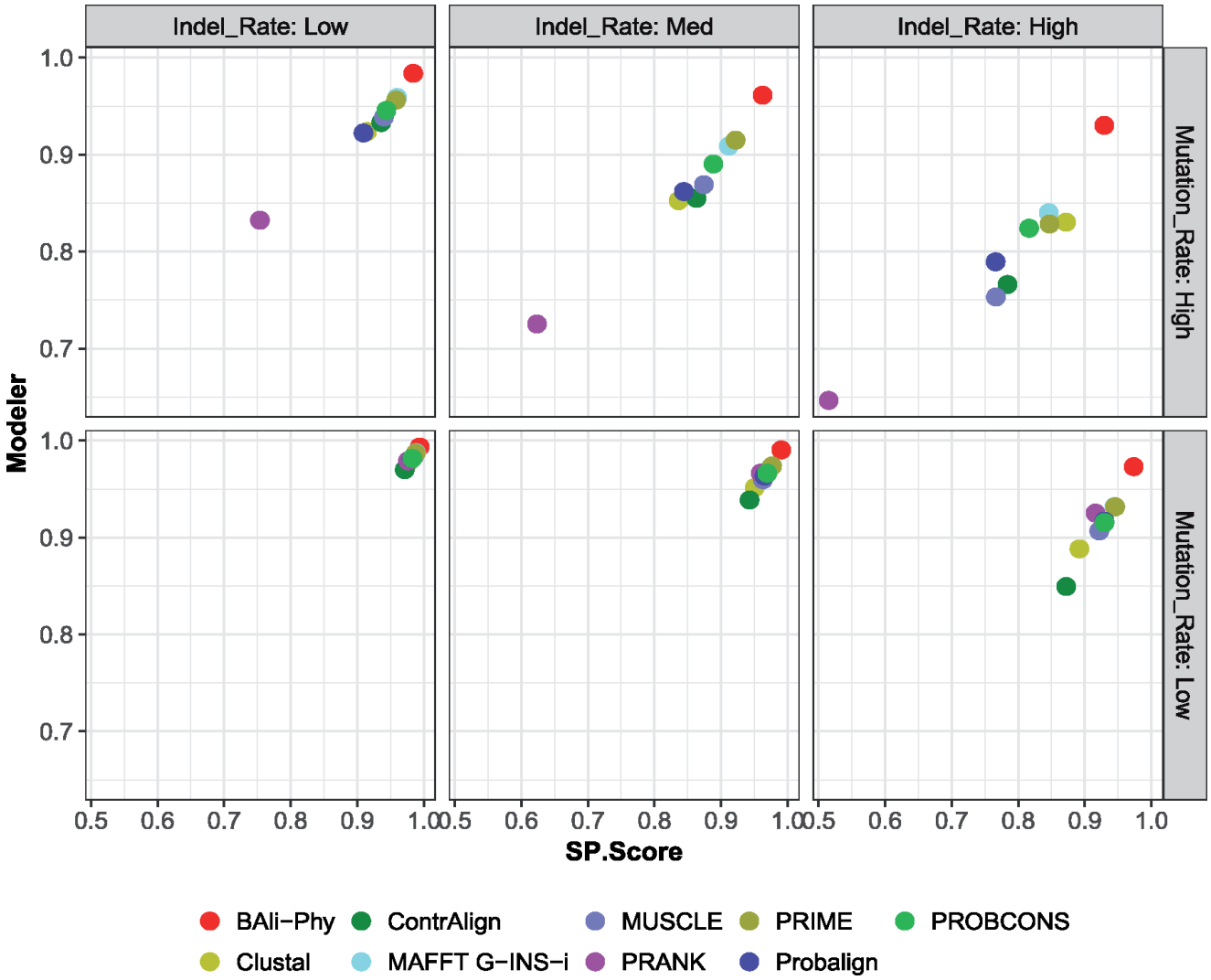
Published: 17 October 2018 **Article history** ▾

Compared BAli-Phy to other MSA methods on both 1192 protein biological benchmarks and 120 simulated datasets

Bali-Phy was run for 32 independent runs, each for 48 hours, to enable it to converge

All datasets at most 27 sequences

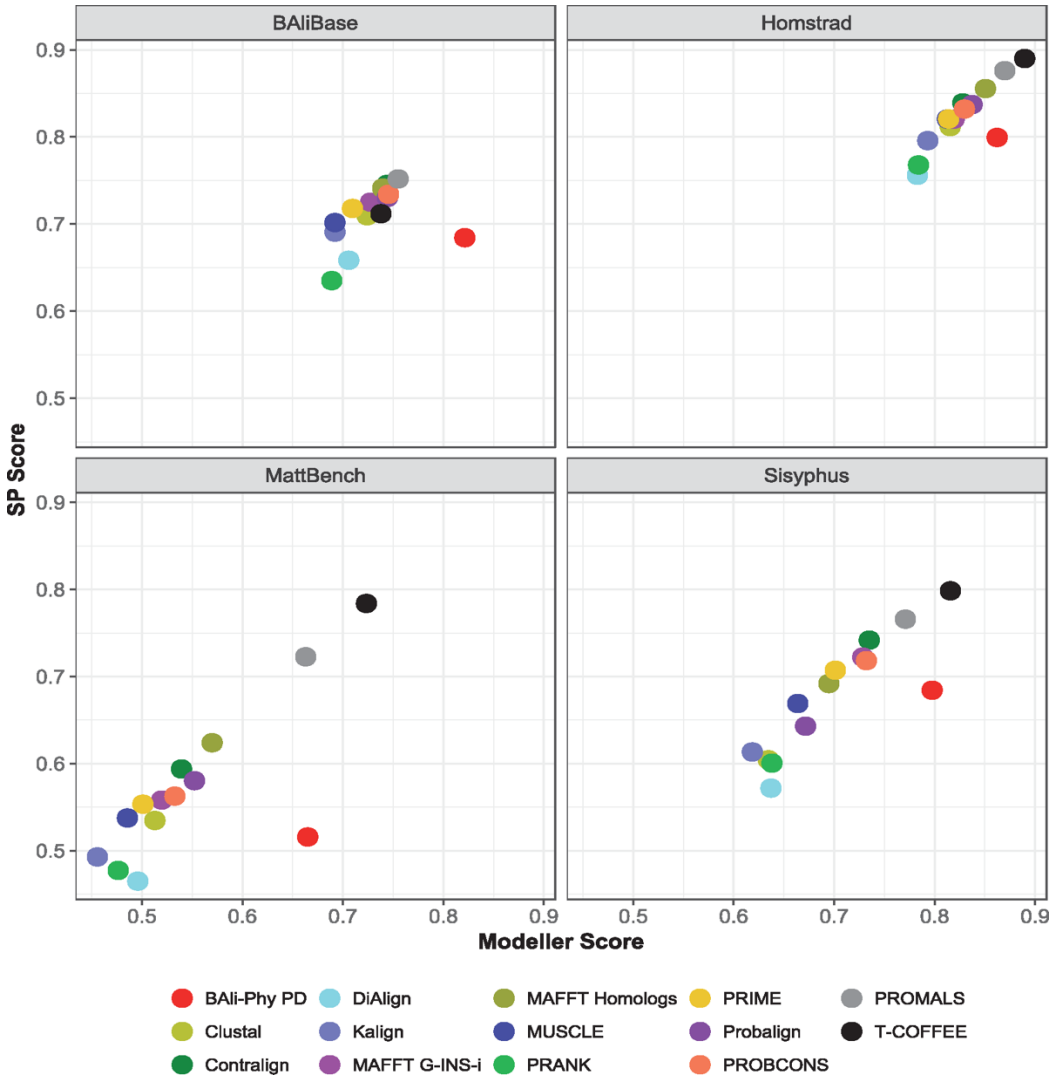
BAli-Phy is best on small simulated protein datasets!



BAli-Phy is best!

Systematic Biology, Volume 68, Issue 3, May 2019, Pages 396–411, https://doi.org/10.1093/sysbio/syy068

BAli-Phy not so great on on 1192 small biological protein datasets



T-Coffee and PROMALS are best!

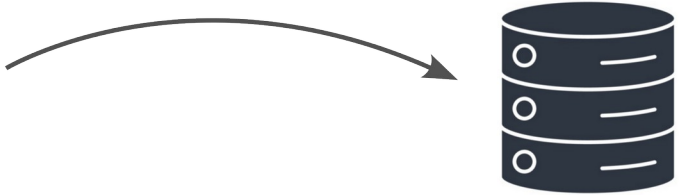
BAli-Phy good for Modeler score, but not so good for SP-Score (e.g., MAFFT better)

Systematic Biology, Volume 68, Issue 3, May 2019, Pages 396–411, https://doi.org/10.1093/sysbio/syy068

Taxonomic assignment and mapping of ancient environmental DNA at scale

Very short reads

Many mismatches



Whole tree of life

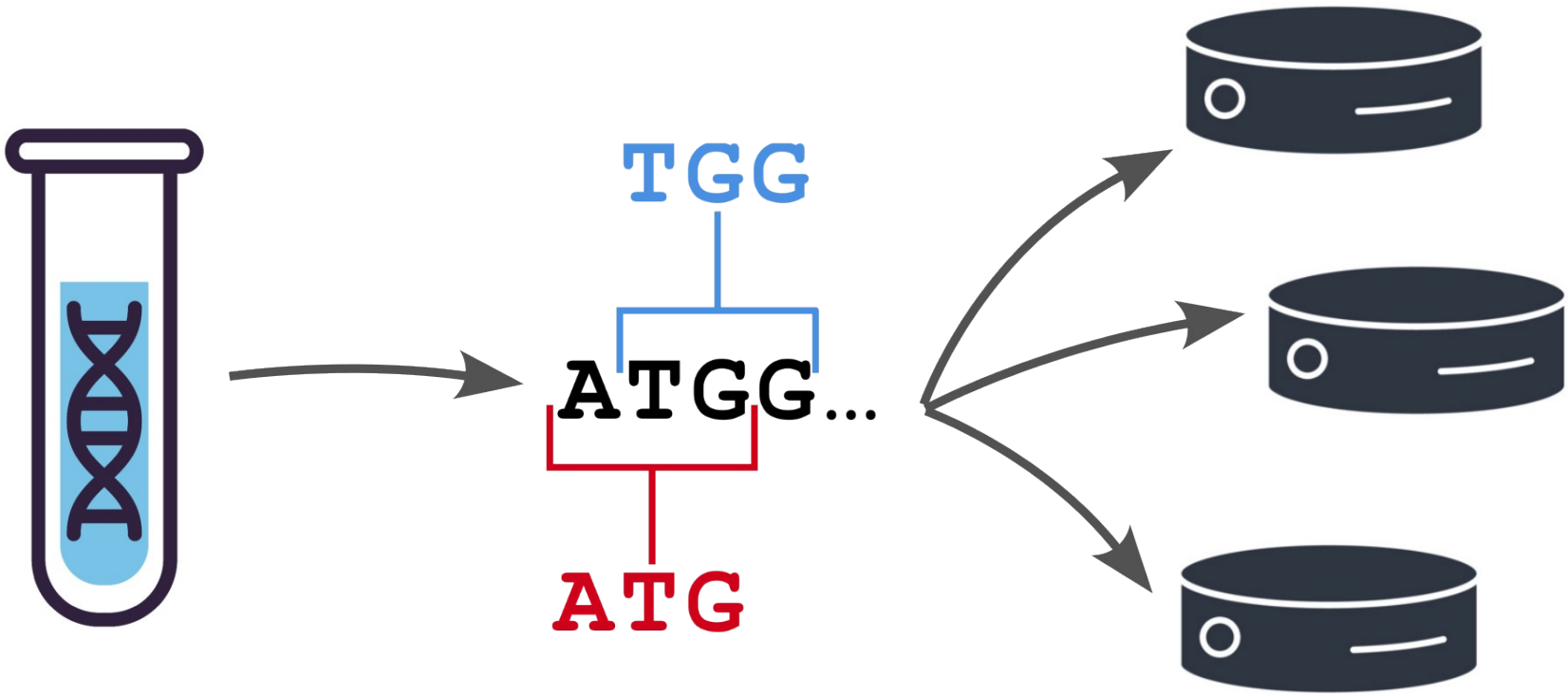
Whole genomes

We want mapping

– not just taxonomic assignment –

to estimate damage

Pre-select reference shards via k-mers,
then map reads to selected shards



Problems with established approaches

Storing the k-mer values:

→ Too much to keep all occurring k-mers

→ Subset the kmers (e.g., minimizers)

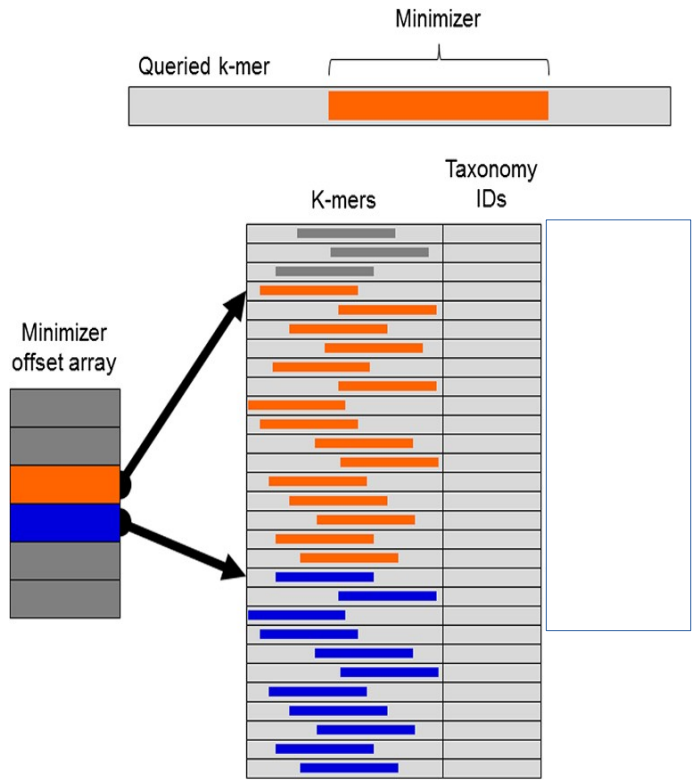
→ Loss of sensitivity

For aDNA, we need to keep

more k-mers with shorter k

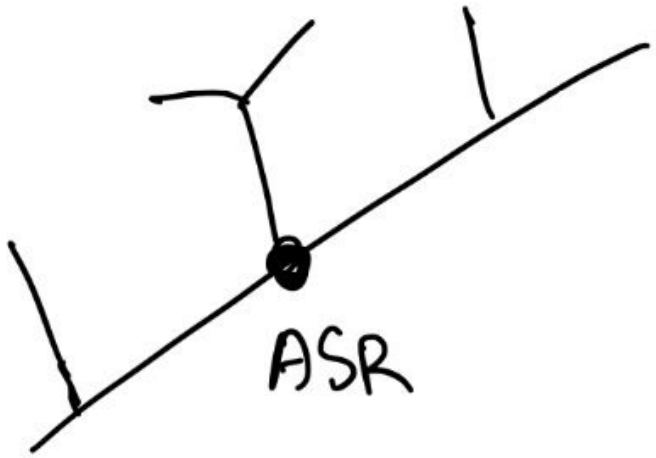
(in the order of the mapping seed length)

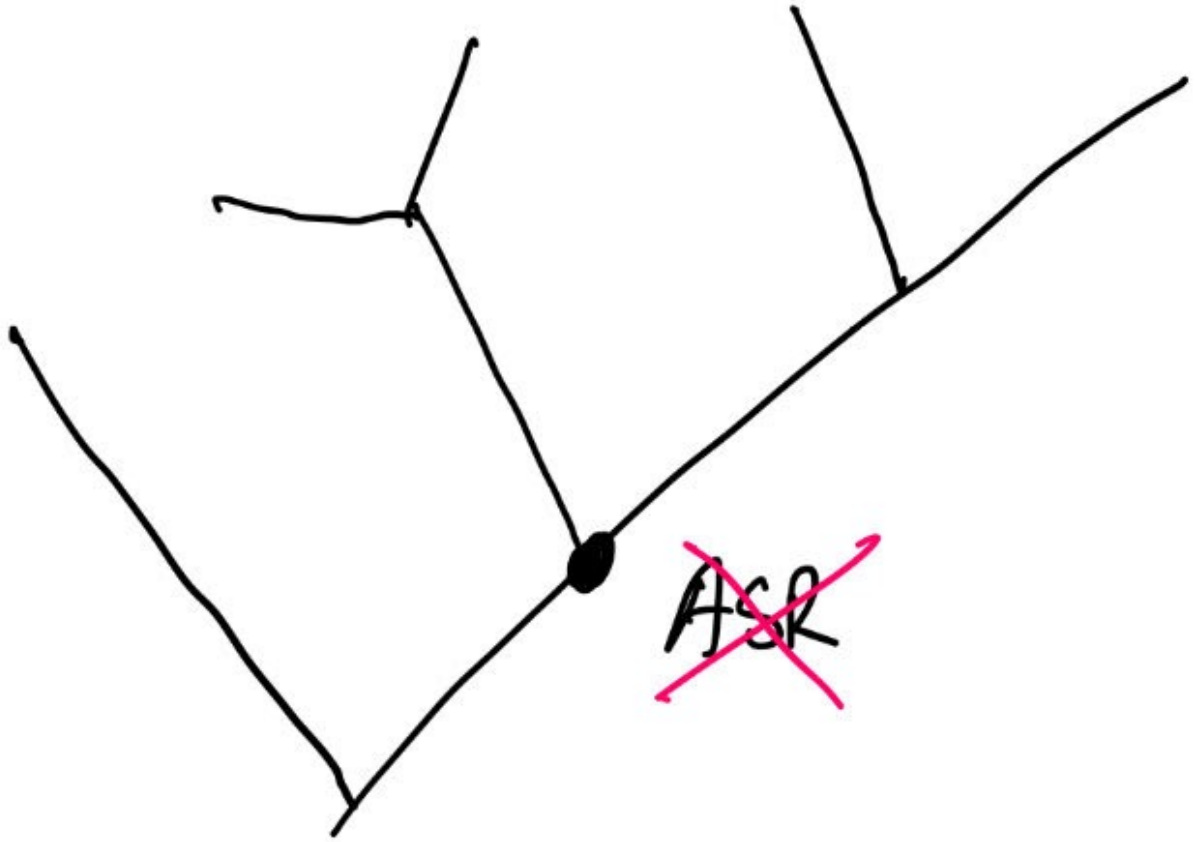
Kraken

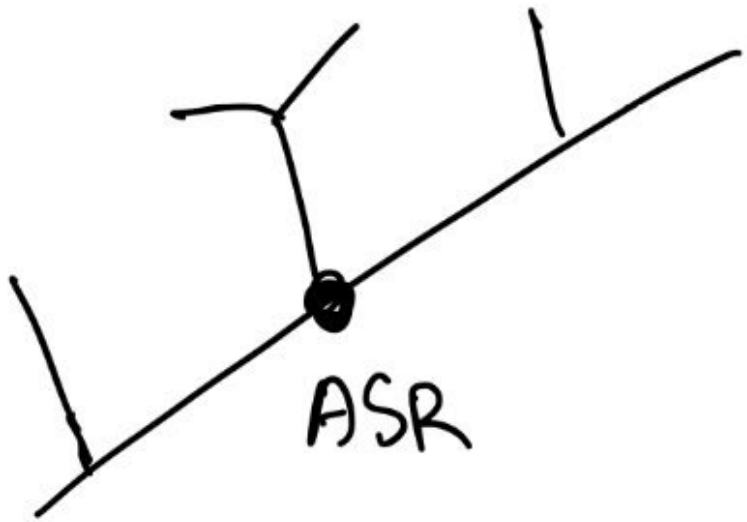


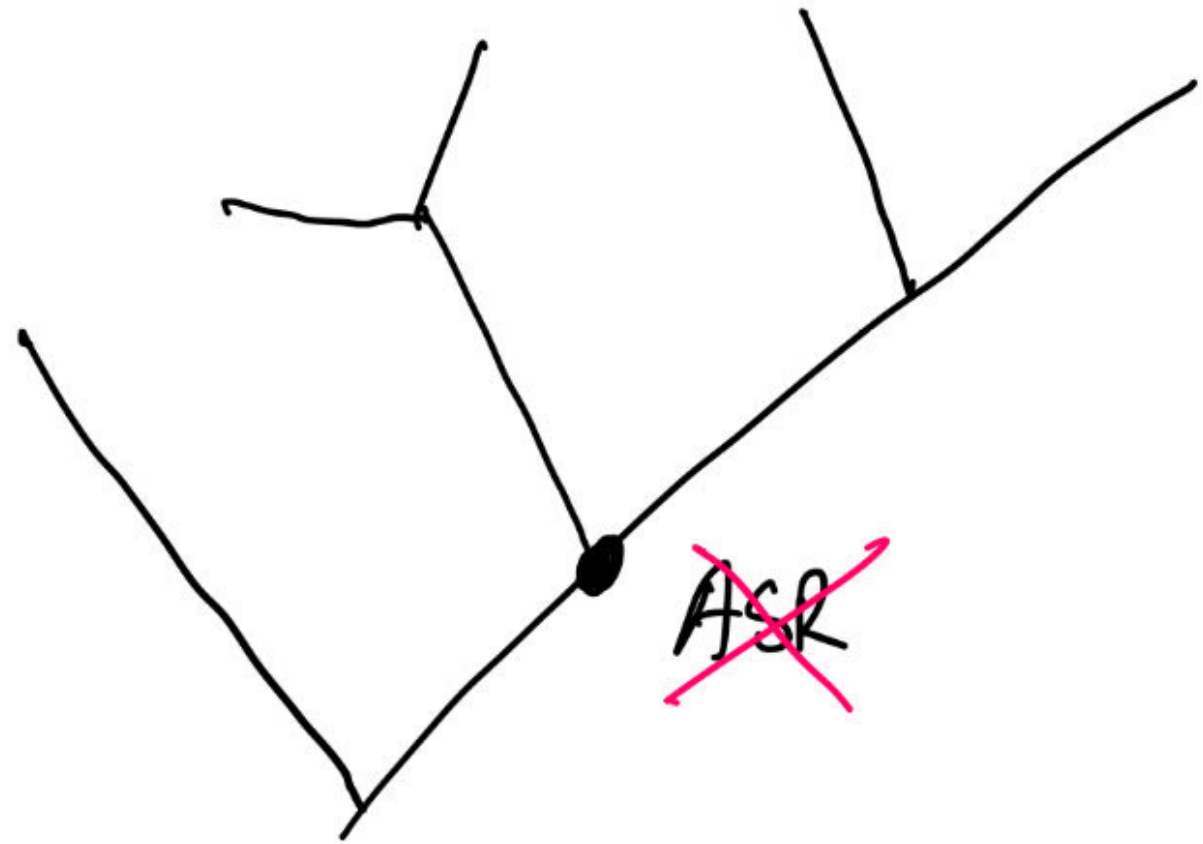
Wood and Salzberg, Genome Biology, 2014

Ancestral sequence reconstruction beyond the twilight zone?





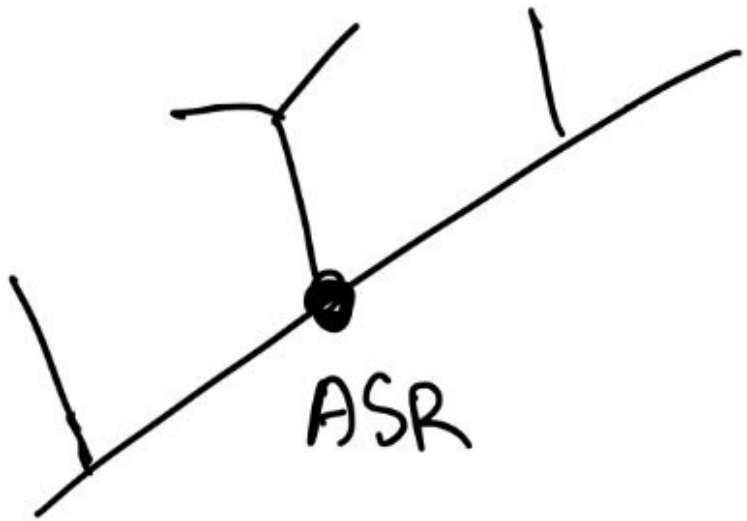




Structure

Sequence

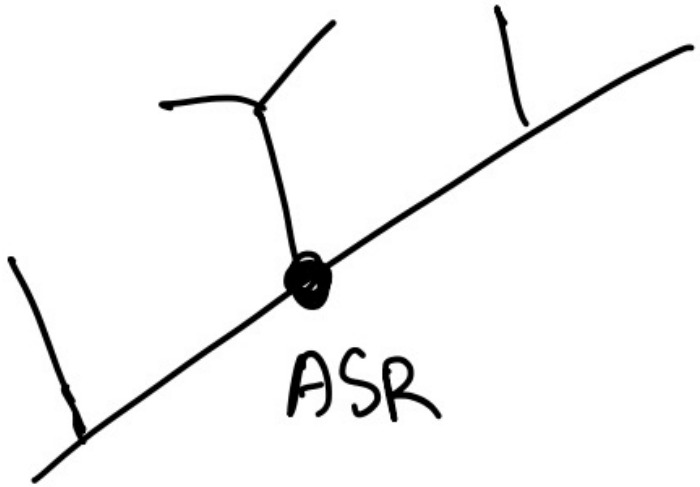
MSEAKF...

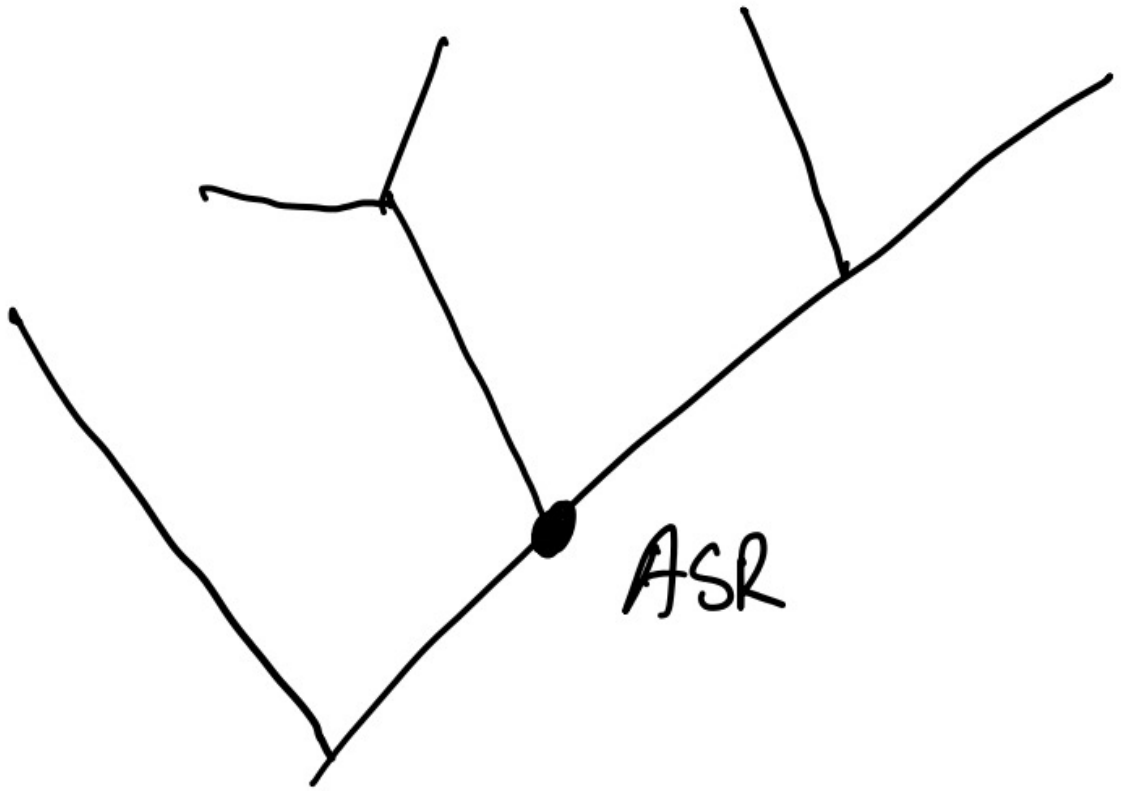


Structure
(3Di)

AKDVRC...
(3Di)

The problem:
going back to amino acid sequence from 3Di sequence





Weighting?

Certain 3Di sequences → Uncertain AA sequences

Joint probability?

- Let T be a tree with branch length on the node set V. Let sequences $\mathcal{S} = S_1 \ldots S_n$ evolve on T using some CTMC like JC or GTR (note: $S_i \in \{A, C, G, T\}^L$). Let $T_{i,j}$ be distances between S_i and S_j on T.
- Vanila: Find a function $\Phi(s) : \{A, C, G, T\}^L \to \mathbb{R}^d$ such that

$$\forall i, j : \lim_{L \to \infty} d(\Phi(S_i), \Phi(S_j)) \to T_{i,j}$$

  for some measure of distance d such as ℓ_2. What is the smallest d where this is possible?
- Extended: Instead of $\mathbb{R}^d$, we can look for embedding in another geometric space.