Accounting for Sequence Uncertainty in Maximum Likelihood Phylogenetic Inference

Alexey Kozlov[1], **Alexandros Stamatakis**[1,2]

[1]The Exelixis Lab, Scientific Computing Group, Heidelberg Institute for Theoretical Studies, Heidelberg, Germany
[2]Institute for Theoretical Informatics, Karlsruhe Institute of Technology, Karlsruhe, Germany

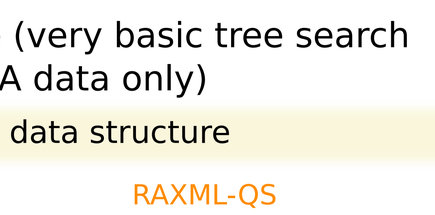
Background

Maximum Likelihood (ML) phylogenetic inference

Multiple sequence alignment (MSA)

```
A  AGCT-GCAGTACCC
B  AGCTTGCAGTACCG
C  AGCATGCA---CCG
D  AGCATGCA---CCT
```

Phylogenetic tree



LH(MSA | tree) ⇒ max

Problem:

Discrete characters cannot adequately represent MSA uncertainty (= errors in sequencing, assembly, alignment)

Traditional approaches to deal with uncertainty

- IUPAC ambiguity code (e.g., S = C or G)
- filtering (remove low-quality bases/columns)
- sequencing with high coverage (overlap multiple reads and call a consensus base)

Idea

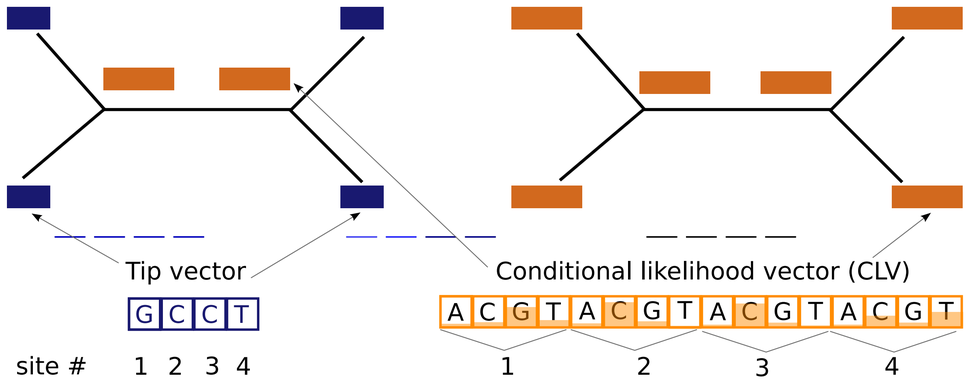
Explicitly use quantified sequence uncertainty in phylogenetic likelihood calculation

Implementation

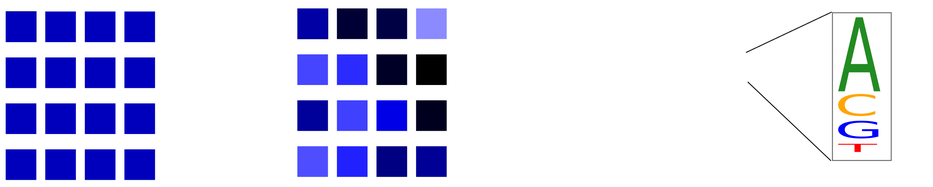
Experimental RAxML-QS code (very basic tree search functionality of RAxML[1], DNA data only)

Changes to the tree data structure

RAXML RAXML-QS



Available error models



uniform error ε (user-defined or estimated)

site-specific error ε_{ij} (FASTQ)

per-site base probabilities (CATG (pyRAD v3.0), FAST5 (Oxford Nanopore))

Computing base probabilities

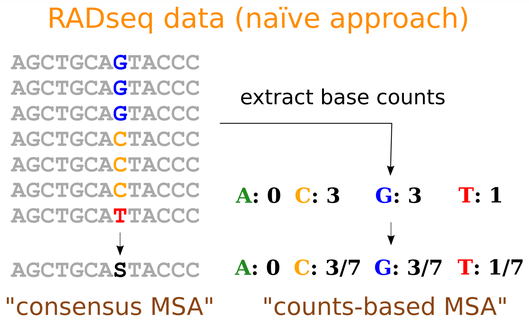
FASTQ quality scores

```
S  AGCTGCAGTACCC   sequence
Q  75:>E3/?8;BB@   phred+33
```

$Q = -10\log_{10}\varepsilon \rightarrow \varepsilon = 10^{-0.1Q}$

$P(A \mid S_i = \text{"A"}) = 1-\varepsilon$
$P(C \mid S_i = \text{"A"}) = \varepsilon/3$
$P(G \mid S_i = \text{"A"}) = \varepsilon/3$
$P(T \mid S_i = \text{"A"}) = \varepsilon/3$

RADseq data (naïve approach)

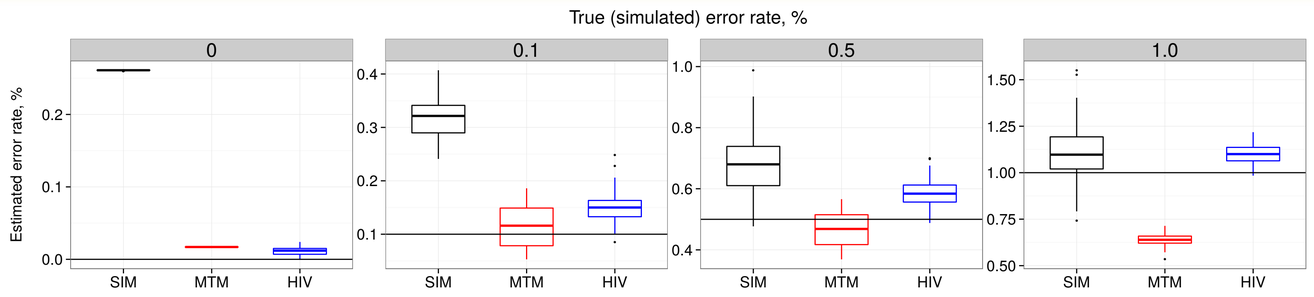


Results

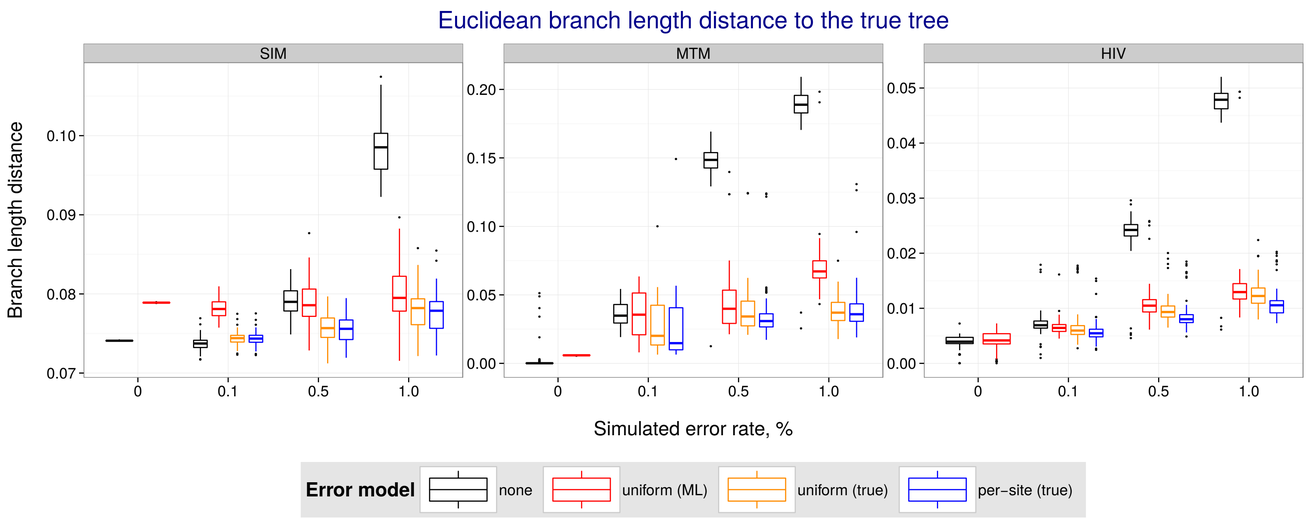
Test datasets

Designator	Organism group, gene	# taxa	# sites	Reference
SIM	(simulation)	50	2,000	
HIV	HIV-I virus, *pol*	23	2,841	[5]
MTM	Mammals, mtDNA	23	9,741	[6]
RADSEQ	Lizards, genome-wide	74	3,142 – 1,250,860	A. Leache (unpubl.)

ML estimation of uniform error rate

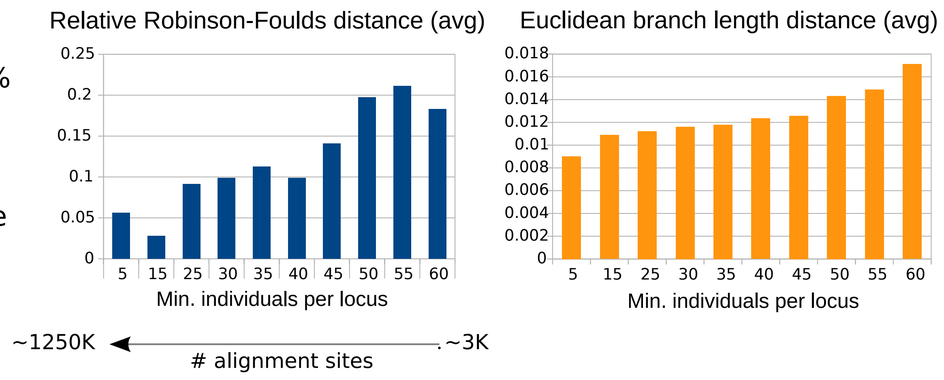


Tree inference on alignments with simulated error



Tree inference on empirical RADseq data: counts-based MSA vs. consensus MSA

- Relative RF distance 3-21%
- Difference is much more pronounced on shorter alignments (= high coverage threshold)



Conclusions & Outlook

- ML estimation of alignment error rate is possible, but its accuracy varies significanlty across datasets and experimental settings
- Accounting for sequence uncertainty improves branch length estimates on simulated data with moderately high error rates (> 0.5 %)
- On empirical RADseq data, using MSA built from raw base counts instead of consensus sequences yields consistently different trees
- Next steps:
  - Use quality scores provided by aligners (e.g., FSA[2])
  - Evaluate the proposed approach in the context of phylogenetic placement (RAxML-EPA)

References

[1] Stamatakis (2014) "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies" *Bioinformatics* 30(9)

[2] Bradley et al. (2009) Fast Statistical Alignment. *PLoS Comput Biol* 5(5)

[3] Kuhner and McGill (2014) "Correcting for sequencing error in maximum likelihood phylogeny inference" *G3 (Bethesda)* 4(12)

[4] Eaton (2014) "PyRAD: assembly of de novo RADseq loci for phylogenetic analyses" *Bioinformatics* 30(13)

[5] Arnason et al. (1997) "Phylogenetic analyses of mitochondrial DNA suggest a sister group relationship between Xenartha (Edentata) and Ferungulates." *Mol. Biol. Evol.* 14:762-768

[6] Yang et al. (2000) "Codon-substitution models for heterogeneous selection pressure at amino acid sites" *Genetics* 155:431-449

Acknowledgements

- Code availability: https://github.com/amkozlov/raxml-qs
- This work was funded by HITS gGmbH.
- We would like to thank Nick Goldman, Adam Leache and Deren Eaton for their invaluable assistance.

Heidelberg Institute for Theoretical Studies | HITS

Karlsruhe Institute of Technology