Family of HMMs

Nam Nguyen
University of Texas at Austin

Outline of Talk

- Background
- Family of HMMs
  - Model
  - Alignment algorithm
- Applications of fHMM
  - SEPP (Mirarab, Nguyen, and Warnow. PSB 2012)
  - TIPP (Nguyen, et al. Under review)
- Conclusions and future work

Phylogenetics

Courtesy of Tree of Life Project

- Study of evolutionary relationship between different species
- Applications to many fields such as drug discovery, agriculture, and biotechnology
- Critical are tools for alignment and phylogeny estimation.

Gather Sequences

```
S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA
```

Align Sequences

```
S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA
```

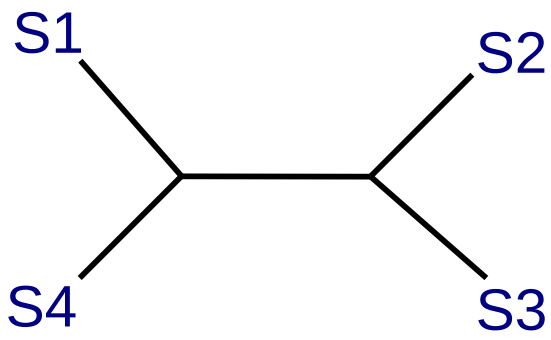
```
S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-CT-------GACCGC--
S4 = -------TCAC--GACCGACA
```

Estimate Tree

```
S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA
```

→

```
S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-CT-------GACCGC--
S4 = -------TCAC--GACCGACA
```



Multiple Sequence Alignment

- Fundamental step in bioinformatics pipelines
- Used in phylogeny estimation, prediction of 2D/3D protein structure, and detection of conserved regions
- Can be formulated as an NP-hard optimization problem
- Popular heuristics include progressive alignment methods and iterative methods
  - Heuristics do not scale linearly with the number of sequences
  - Not as accurate on large datasets or evolutionary divergent datasets

Profile Hidden Markov Model (HMM)

- Statistical model for representing an MSA
- Uses include
  - inserting sequences into an alignment
  - taxonomic identification
  - homology detection
  - functional annotation

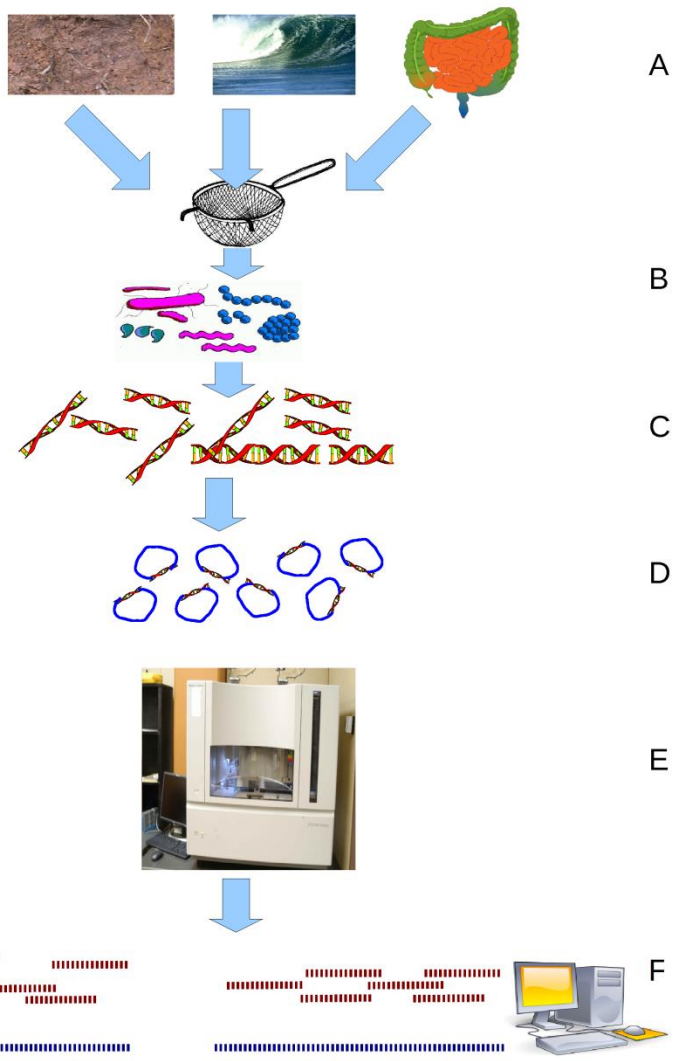
Profile Hidden Markov Model (HMM)

- Statistical model for representing an MSA
- Uses include
  - inserting sequences into an alignment
  - taxonomic identification
  - homology detection
  - functional annotation

Profile Hidden Markov Model (HMM)

- Statistical model for representing an MSA
- Uses include
  - inserting sequences into an alignment
  - taxonomic identification
  - homology detection
  - functional annotation

Metagenomics



Courtesy of Wikipedia

- Study of sequencing genetic material directly from the environment
- Applications to biofuel production, agriculture, human health
- Sequencing technology produces millions of short reads from unknown species
- Fundamental step in analysis is identifying taxa of read

Phylogenetic Placement

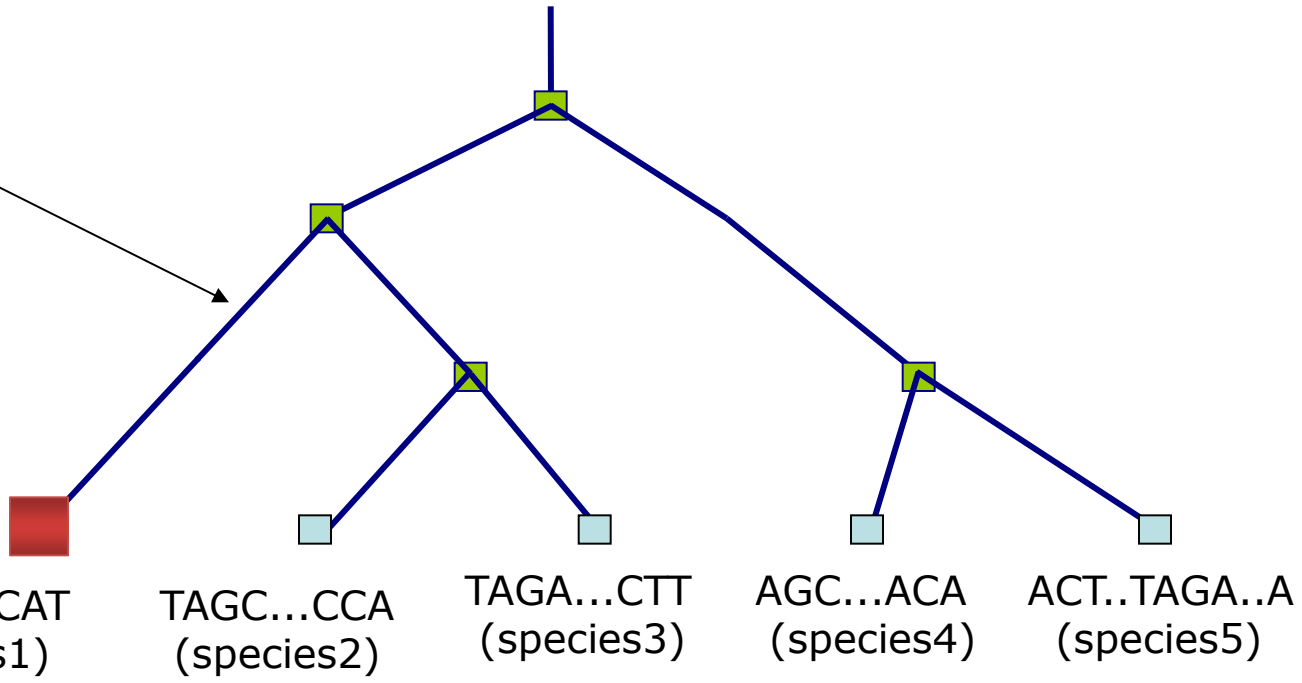
Fragmentary Reads:

(60-200 bp long)

- ACCG
- CGAG
- CGG
- GGCT
- TAGA
- GGGGG
- TCGAG
- GGCG
- GGG
- .
- .
- .
- ACCT

Known Full length Sequences,
and an alignment and a tree

(500-10,000 bp long)



Phylogenetic Placement

- Input: (Backbone) Alignment and tree on full-length sequences and a query sequence (short read)
- Output: Placement of the query sequence on the backbone tree
- Use placement to infer relationship between query sequence and full-length sequences in backbone tree
- Applications in metagenomic analysis
  - Millions of reads
  - Reads from different genomes mixed together
  - Use placement to identify read

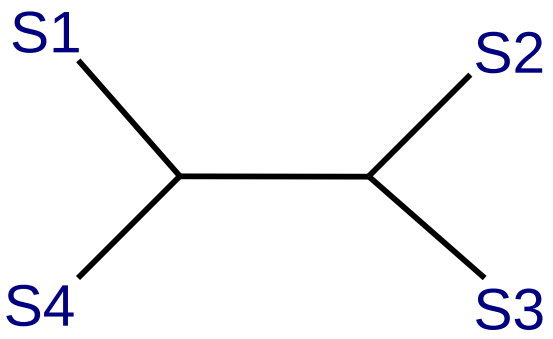
Phylogenetic Placement

- Align each query sequence to backbone alignment to produce an extended alignment

- Place each query sequence into the backbone tree using extended alignment

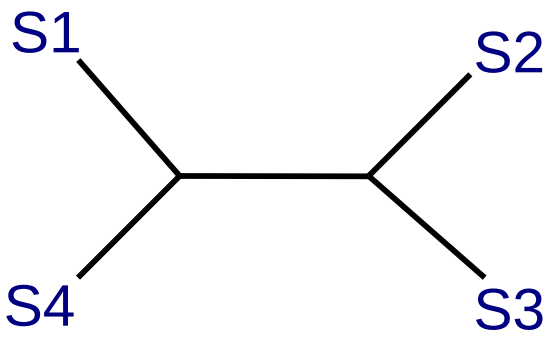
Align Sequence

```
S1  = -AGGCTATCACCTGACCTCCA-AA
S2  = TAG-CTATCAC--GACCGC--GCA
S3  = TAG-CT-------GACCGC--GCT
S4  = TAC----TCAC--GACCGACAGCT
Q1  = TAAAAC
```



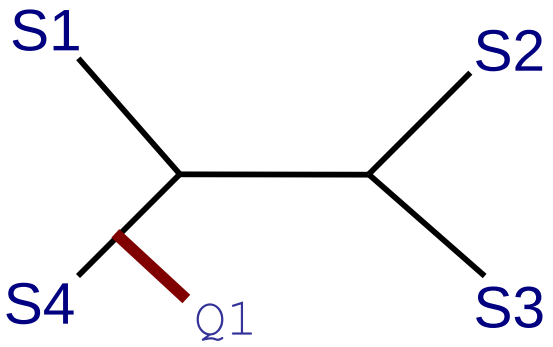
Align Sequence

```
S1  = -AGGCTATCACCTGACCTCCA-AA
S2  = TAG-CTATCAC--GACCGC--GCA
S3  = TAG-CT-------GACCGC--GCT
S4  = TAC----TCAC--GACCGACAGCT
Q1  = -------T-A--AAAC--------
```



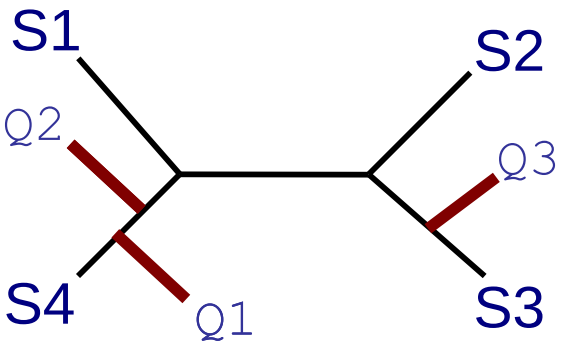
Place Sequence

```
S1  = -AGGCTATCACCTGACCTCCA-AA
S2  = TAG-CTATCAC--GACCGC--GCA
S3  = TAG-CT-------GACCGC--GCT
S4  = TAC----TCAC--GACCGACAGCT
Q1  = -------T-A--AAAC--------
```



Place Sequence

```
S1  = -AGGCTATCACCTGACCTCCA-AA
S2  = TAG-CTATCAC--GACCGC--GCA
S3  = TAG-CT-------GACCGC--GCT
S4  = TAC----TCAC--GACCGACAGCT
Q1  = -------T-A--AAAC--------
```



Query sequences are aligned and placed independently

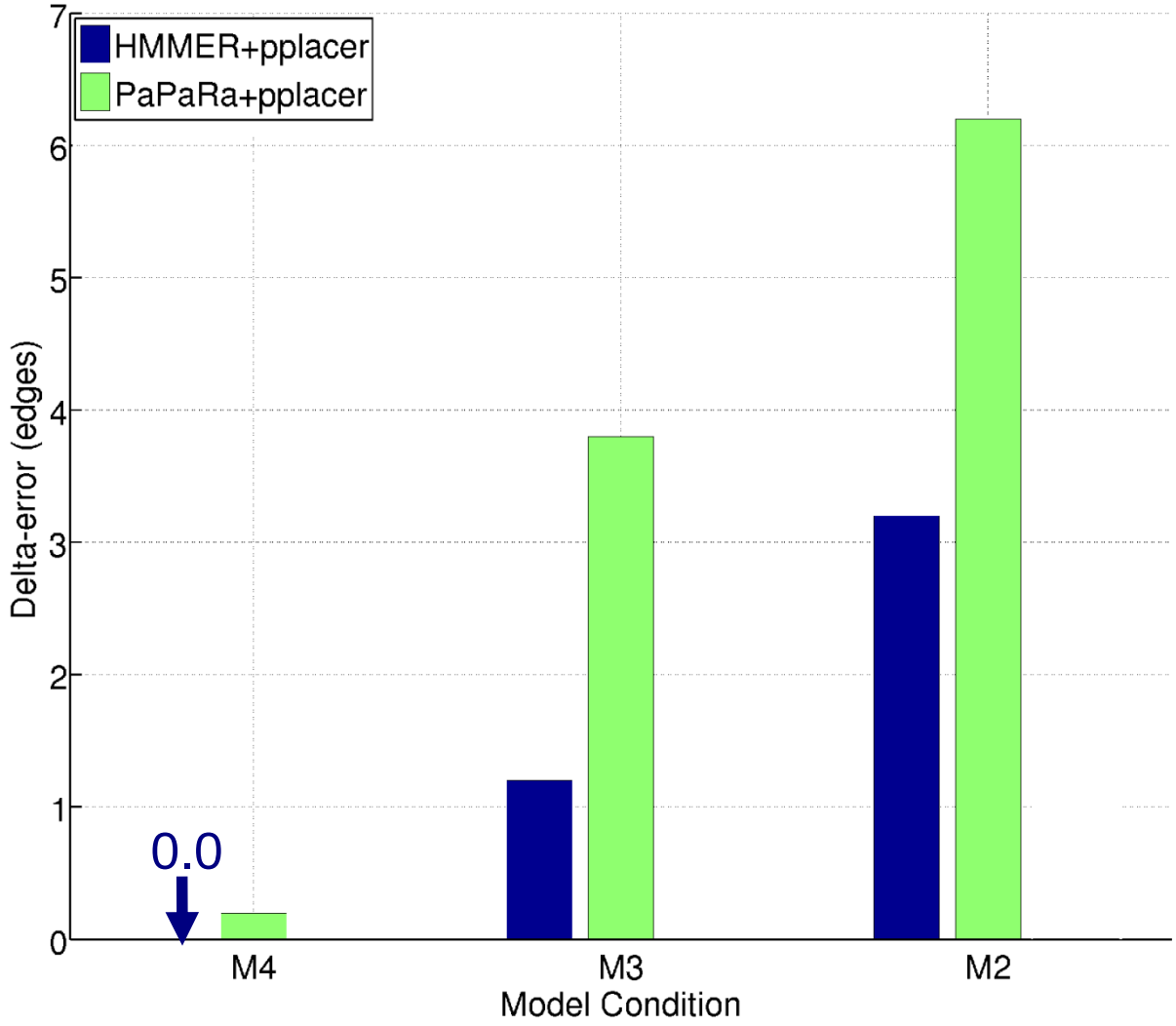
Phylogenetic Placement

- Align each query sequence to backbone alignment:
  - HMMALIGN (Eddy, Bioinformatics 1998)
  - PaPaRa (Berger and Stamatakis, Bioinformatics 2011)
- Place each query sequence into backbone tree, using extended alignment:
  - pplacer (Matsen et al., BMC Bioinformatics 2010)
  - EPA (Berger et al., Systematic Biology 2011)

Phylogenetic Placement

- Align each query sequence to backbone alignment:
  - HMMALIGN (Eddy, Bioinformatics 1998)
  - PaPaRa (Berger and Stamatakis, Bioinformatics 2011)
- Place each query sequence into backbone tree, using extended alignment:
  - pplacer (Matsen et al., BMC Bioinformatics 2010)
  - EPA (Berger et al., Systematic Biology 2011)

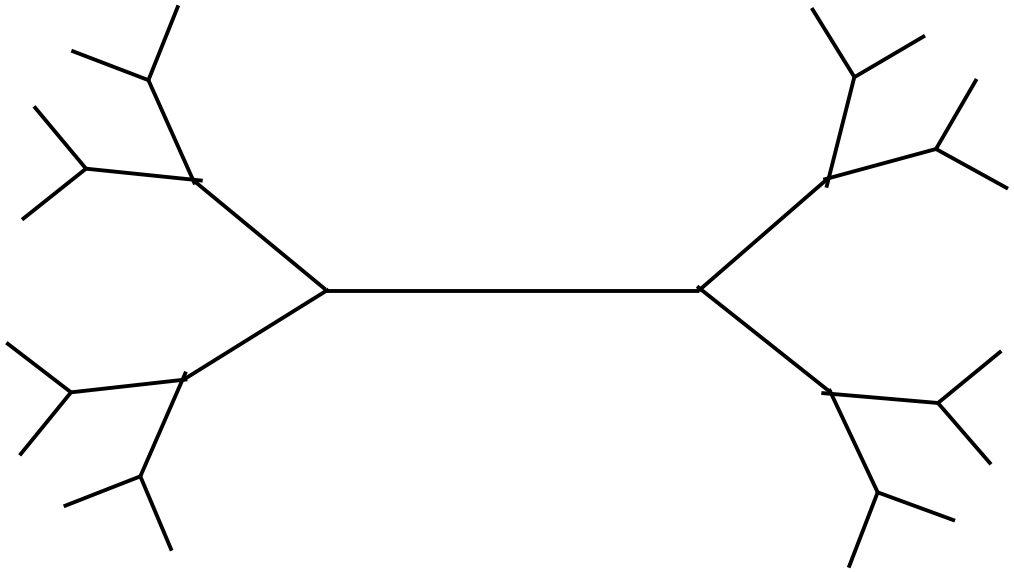
HMMER and PaPaRa results



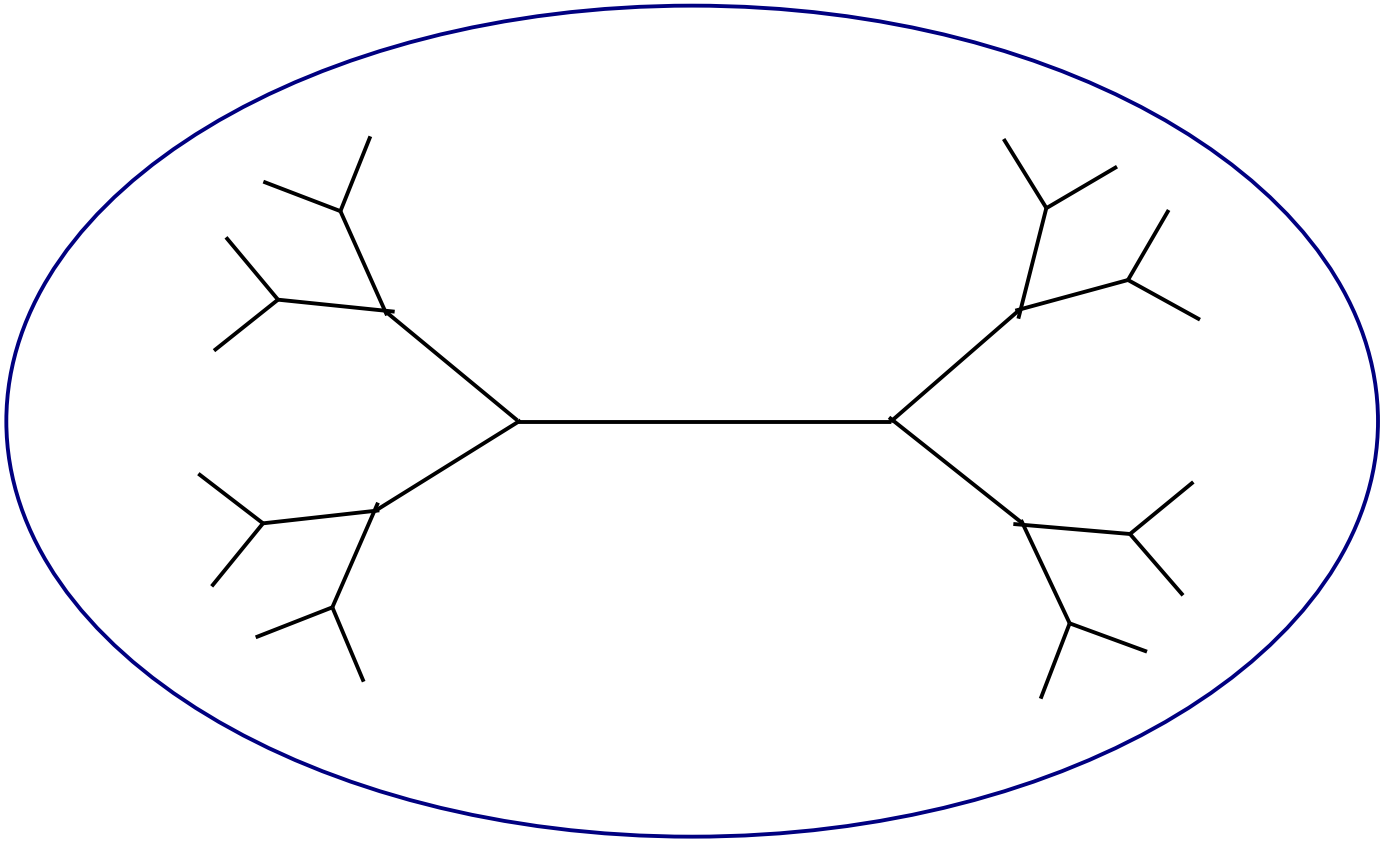
Increasing rate evolution

Backbone size: 500
5000 fragments
20 replicates

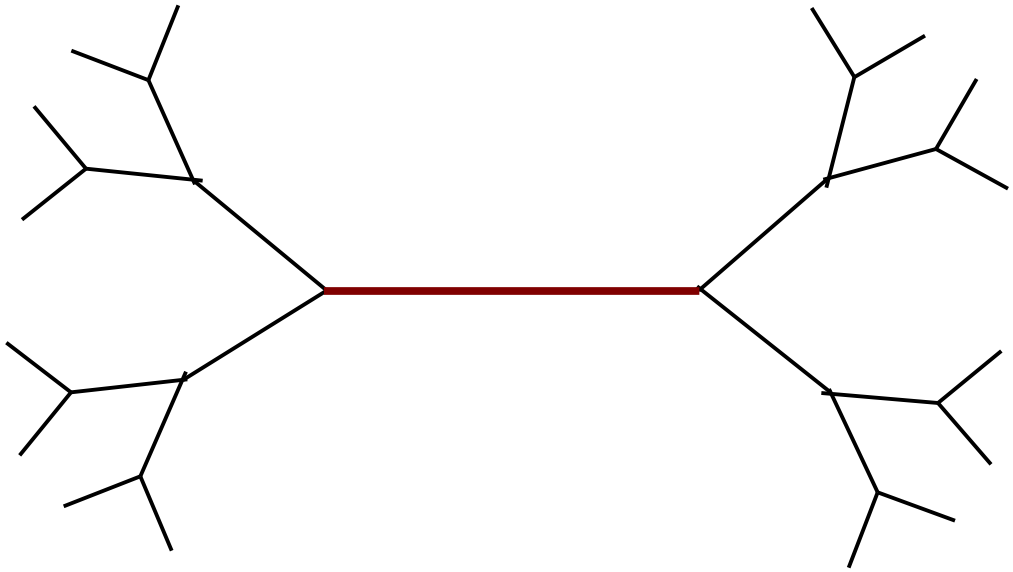
Reducing Evolutionary Distance



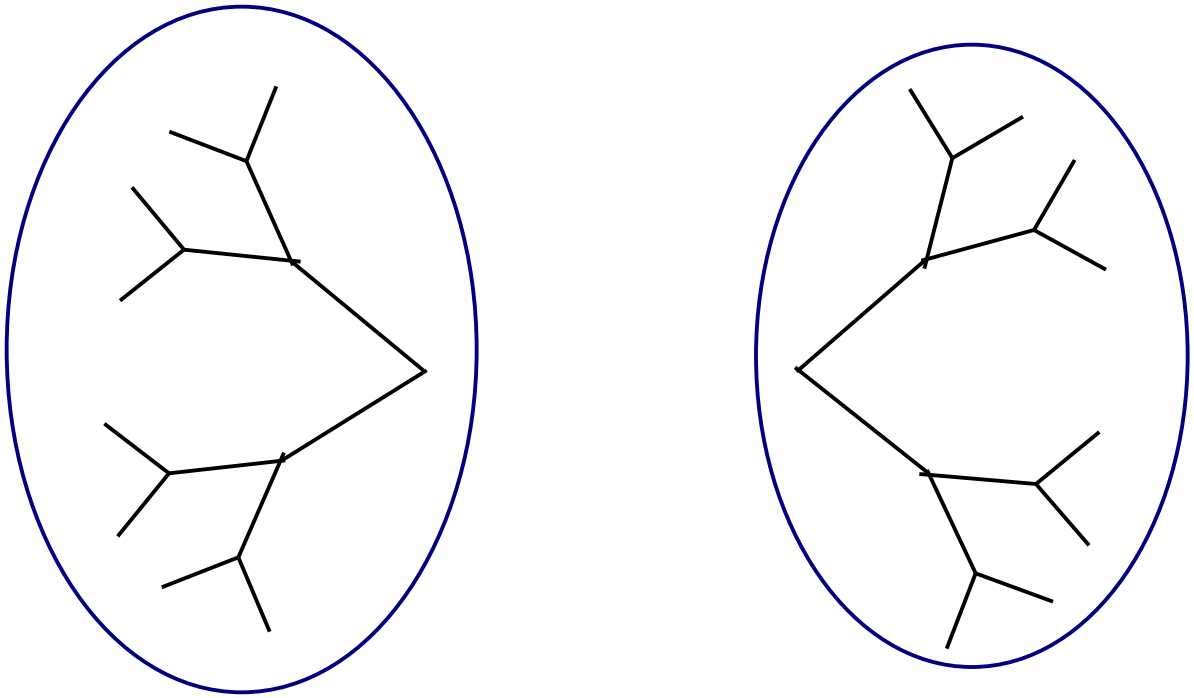
Reducing Evolutionary Distance



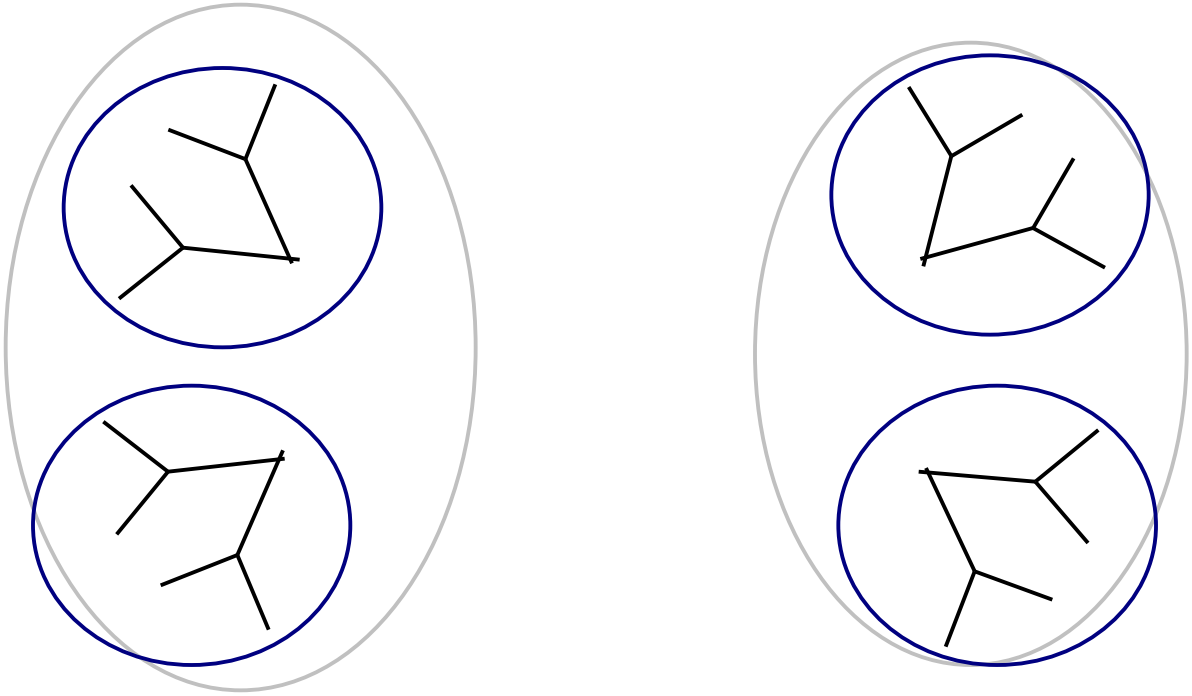
Reducing Evolutionary Distance



Reducing Evolutionary Distance



Reducing Evolutionary Distance



Family of HMMs (fHMM)

- Represents the MSA with multiple HMMs
- Input: backbone alignment and tree on full-length sequences S and max decomposition size N
- Two steps:
  - Decompose tree into subtrees of closely related sequences, with at most N leaves in each subtree
  - Build HMMs on subalignments induced by subtrees

Family of HMMs (fHMM)

- Represents the MSA with multiple HMMs
- Input: backbone alignment and tree on full-length sequences S and max decomposition size N
- Two steps:
  - Decompose tree into subtrees of closely related sequences, with at most N leaves in each subtree
  - Build HMMs on subalignments induced by subtrees

Family of HMMs (fHMM)

- Represents the MSA with multiple HMMs
- Input: backbone alignment and tree on full-length sequences S and max decomposition size N
- Two steps:
  - Decompose tree into subtrees of closely related sequences, with at most N leaves in each subtree
  - Build HMMs on subalignments induced by subtrees

Alignment using fHMM

- Score query sequence against every HMM and select HMM that yields best bit score
- Insert query sequence into subalignment, and by transitivity align query sequence to backbone alignment

Alignment using fHMM

- Score query sequence against every HMM and select HMM that yields best bit score
- Insert query sequence into subalignment, and by transitivity align query sequence to backbone alignment

Alignment using fHMM

- Score query sequence against every HMM and select HMM that yields best bit score
- Insert query sequence into subalignment, and by transitivity align query sequence to backbone alignment

SEPP

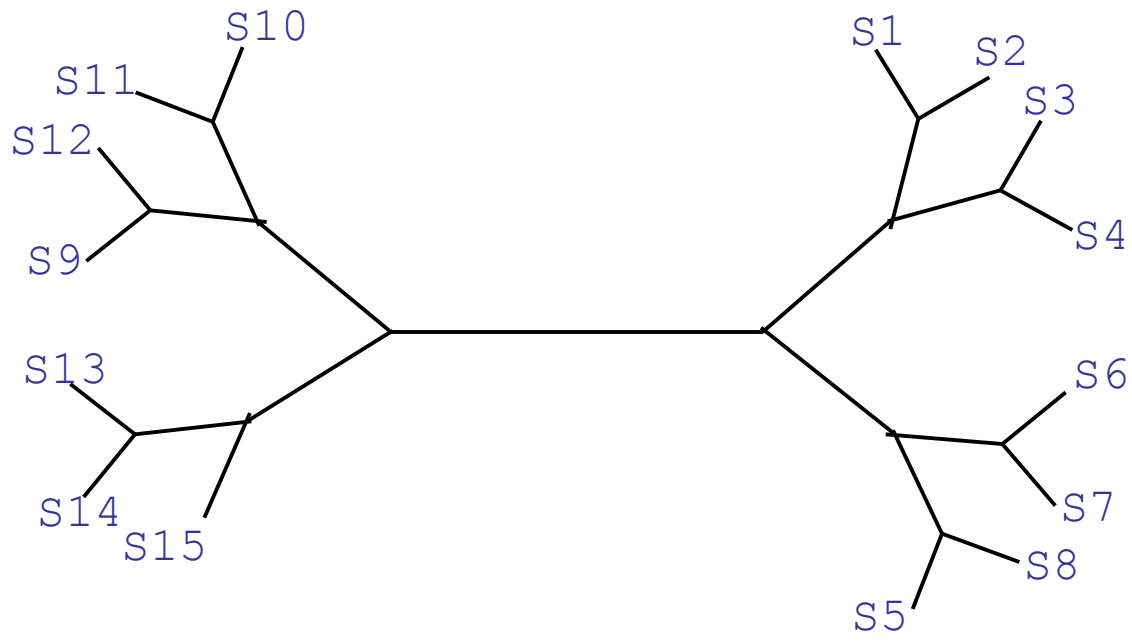
- SEPP = SATé-Enabled Phylogenetic Placement
- Developers: Mirarab, Nguyen, and Warnow
- Two stages of decomposition:
  - Placement decomposition
  - Alignment decomposition
- Parameterized by N and M
  - N: maximum size of alignment subsets
  - M: maximum size of placement subsets
  - $N \leq M$
- Published at Pacific Symposium on Biocomputing 2012

Stage 1: Placement decomposition

N=4, M=8

Decompose tree into placement sets of size ≤ 8

Decompose each placement set into alignment sets of size ≤ 4

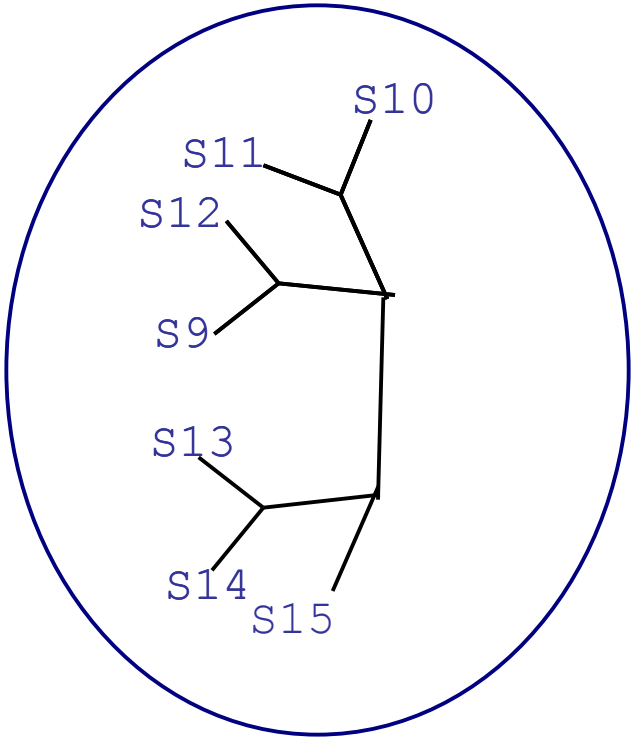


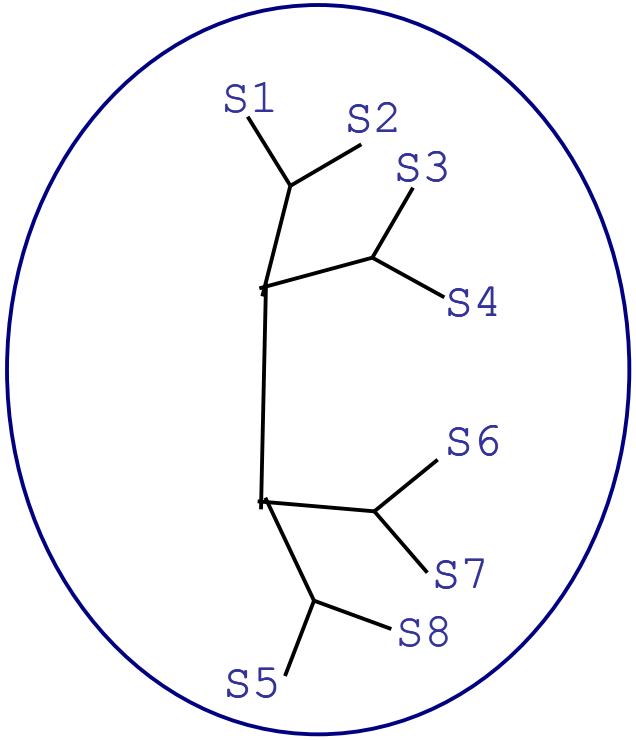
SEPP 4/8: Decompose Tree

N=4, M=8

Decompose tree into placement sets of size ≤ 8

Decompose each placement set into alignment sets of size ≤ 4



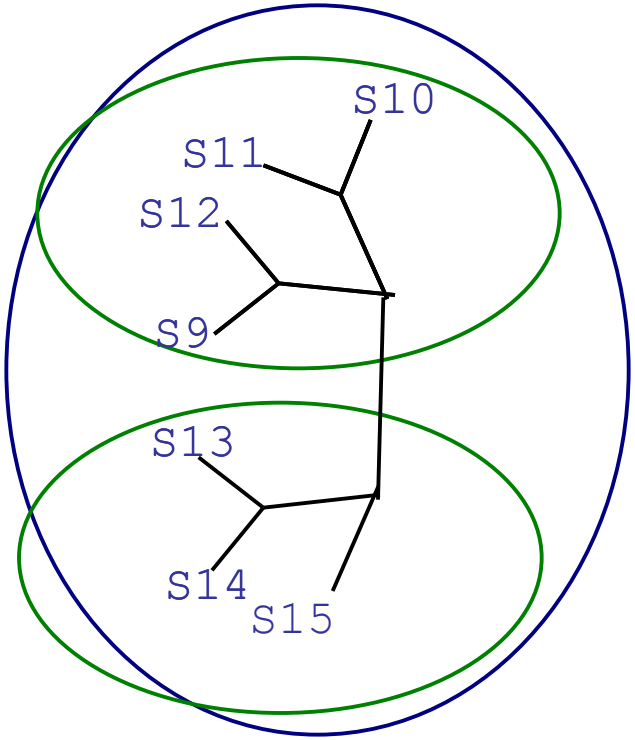


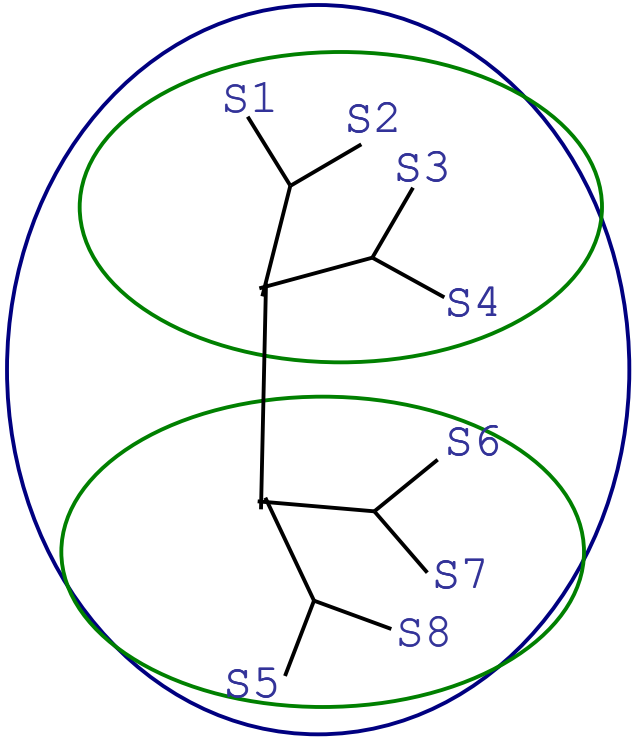
Stage 2: Alignment decomposition

N=4, M=8

Decompose tree into placement sets of size ≤ 8

Decompose each placement set into alignment sets of size ≤ 4

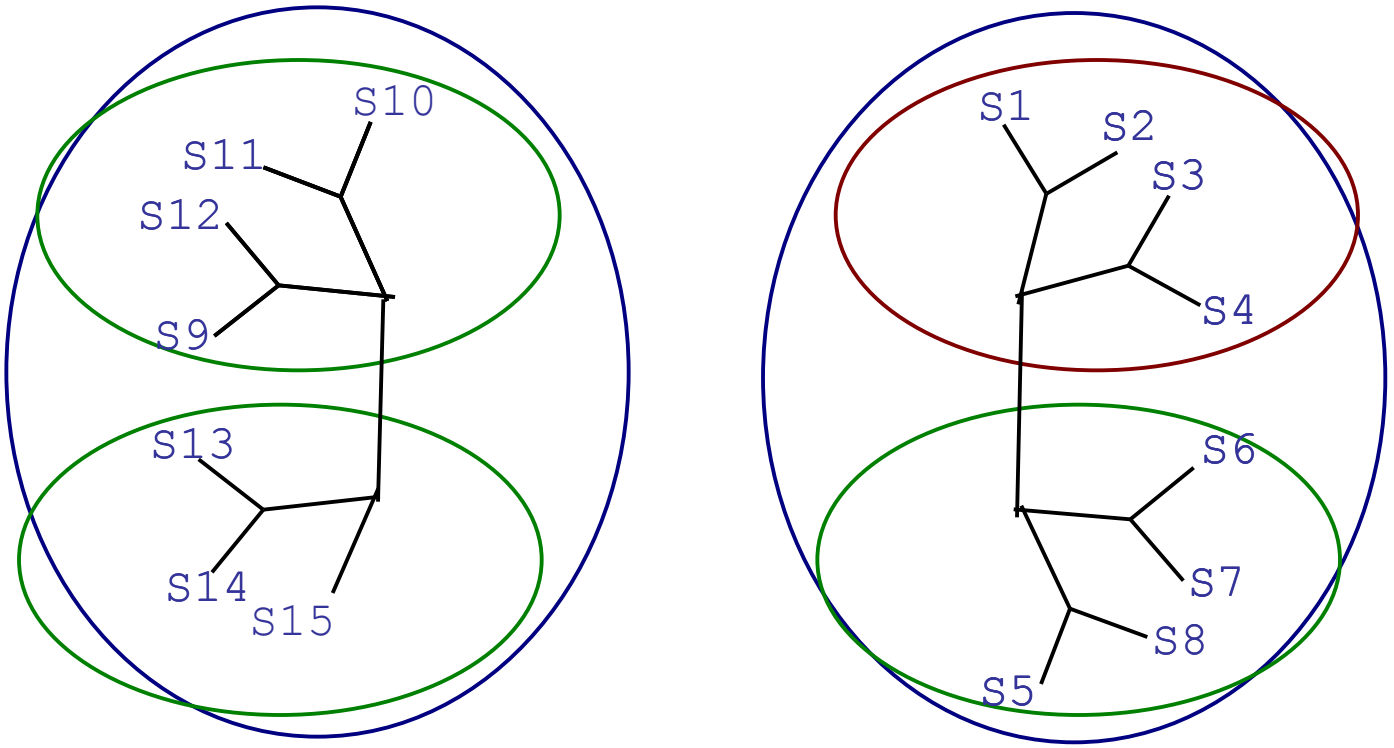




Align and Place Fragment

Align to best HMM

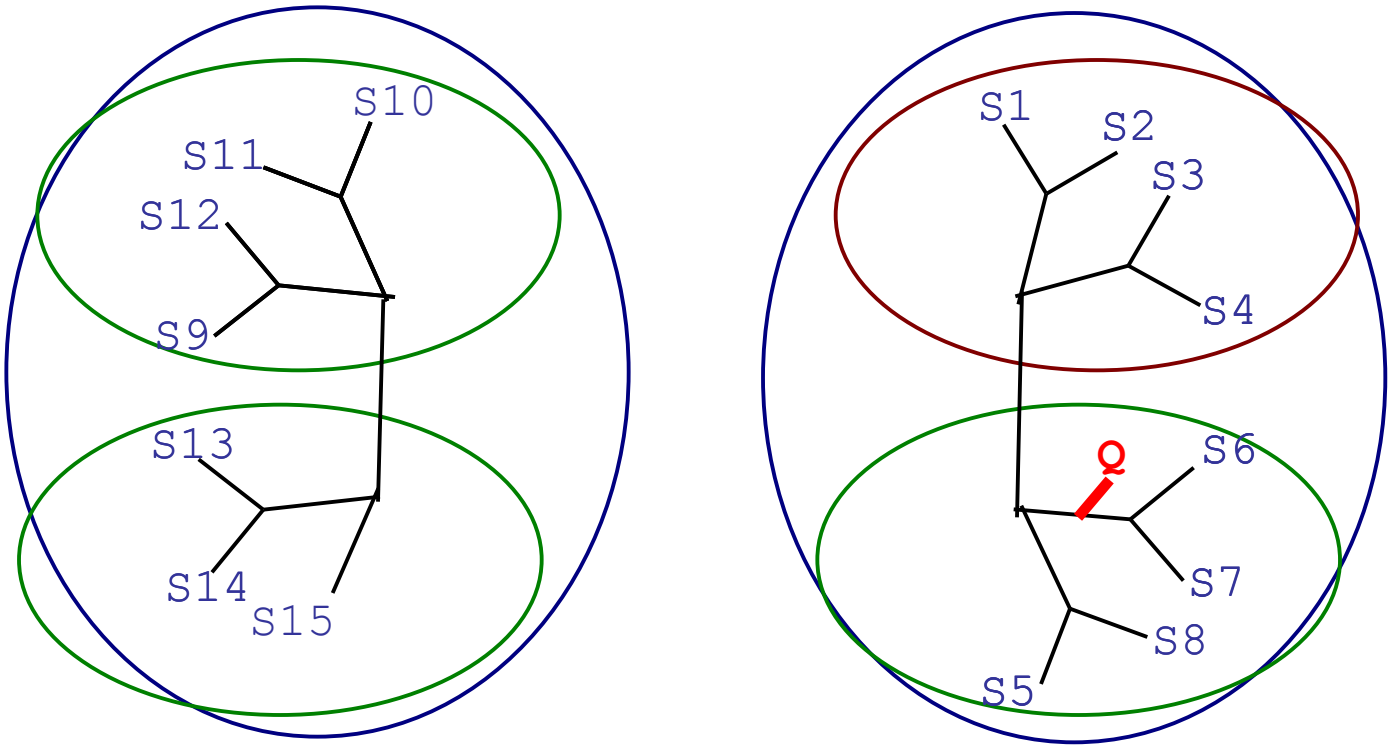
Place within placement subtree containing HMM



Align and Place Fragment

Align to best HMM

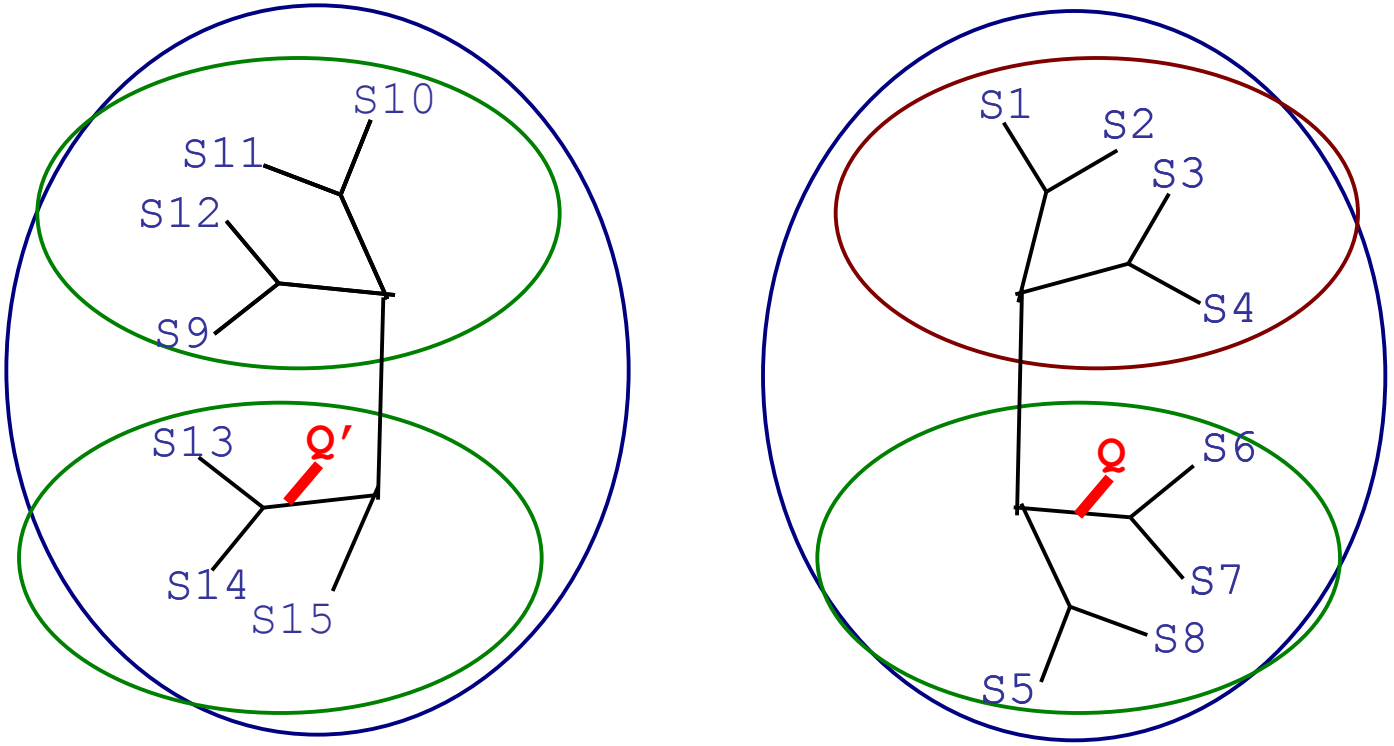
Place within placement subtree containing HMM



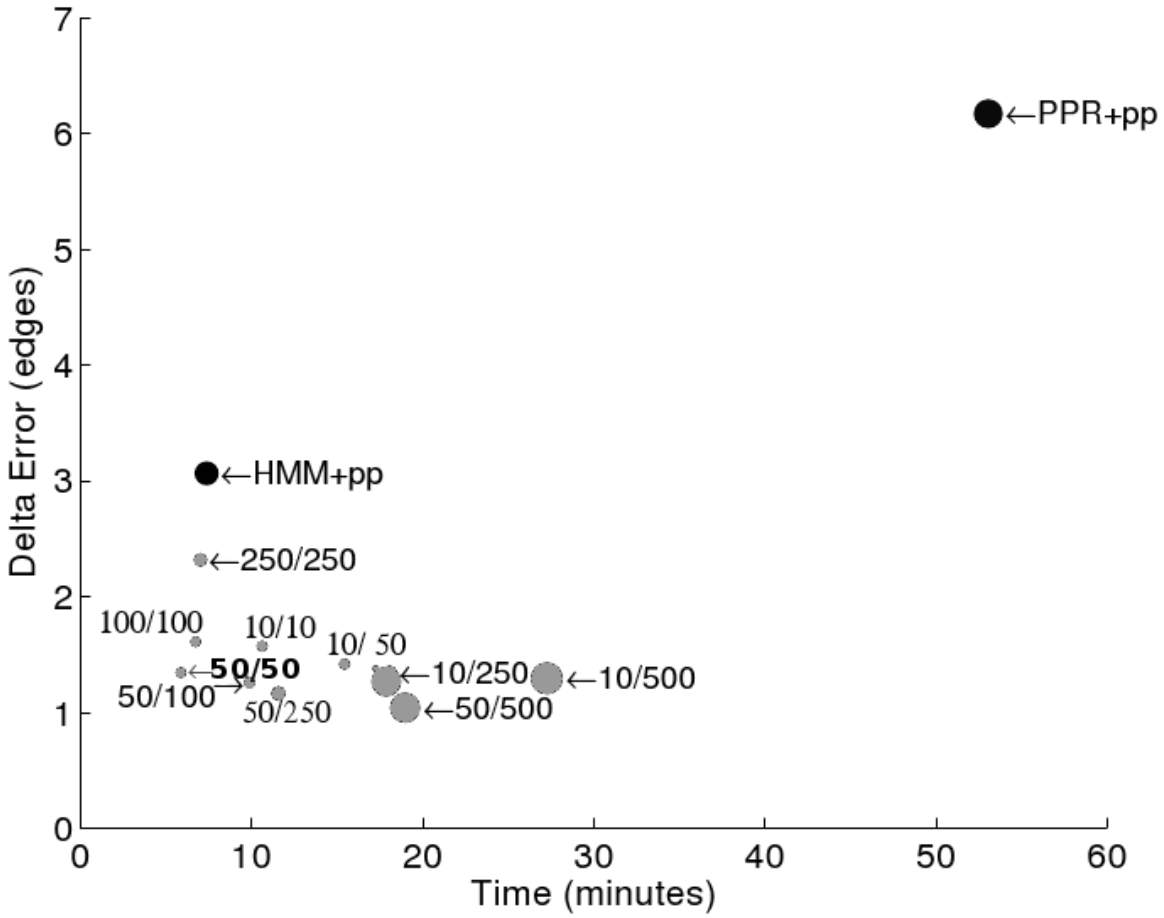
Align and Place Fragment

Align to best HMM

Place within placement subtree containing HMM



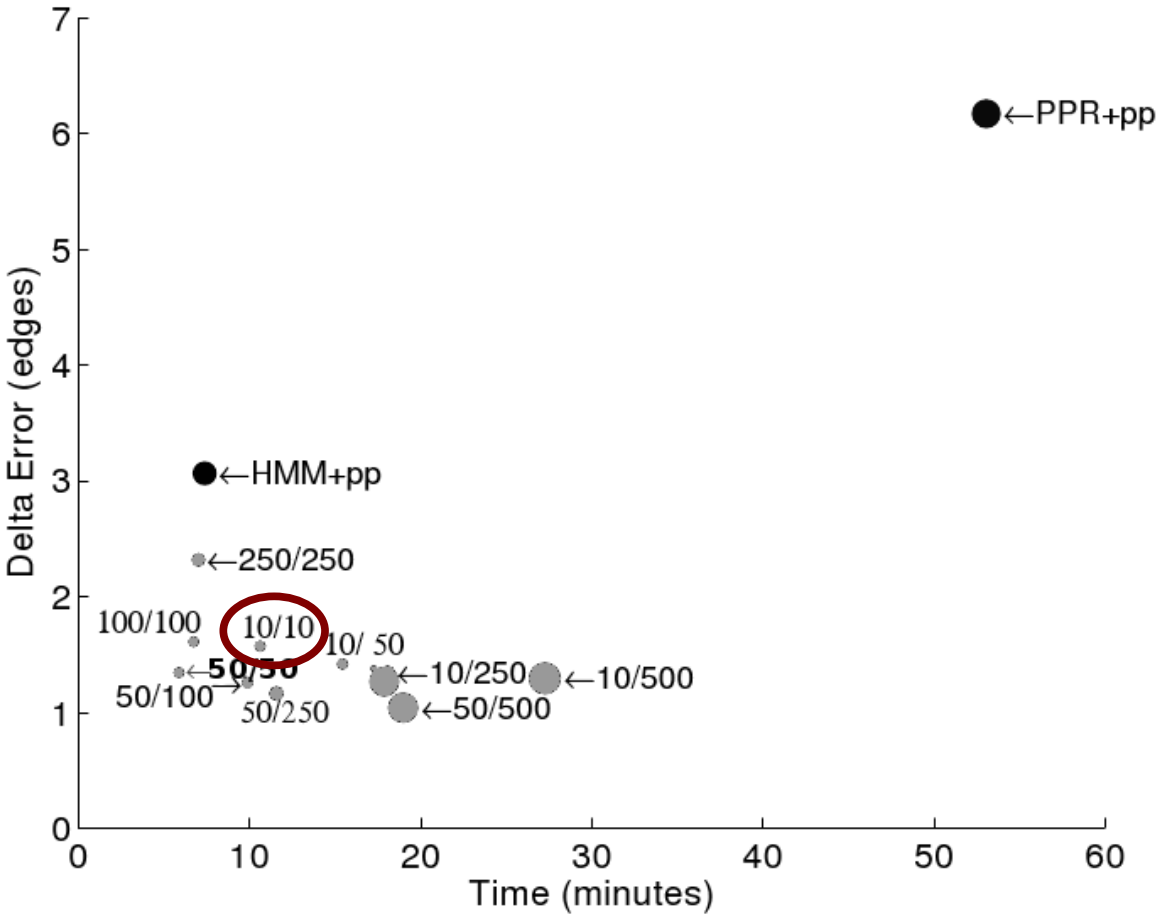
SEPP Parameters: Simulated



M2 model condition, 500 true backbone

SEPP Parameters: Simulated

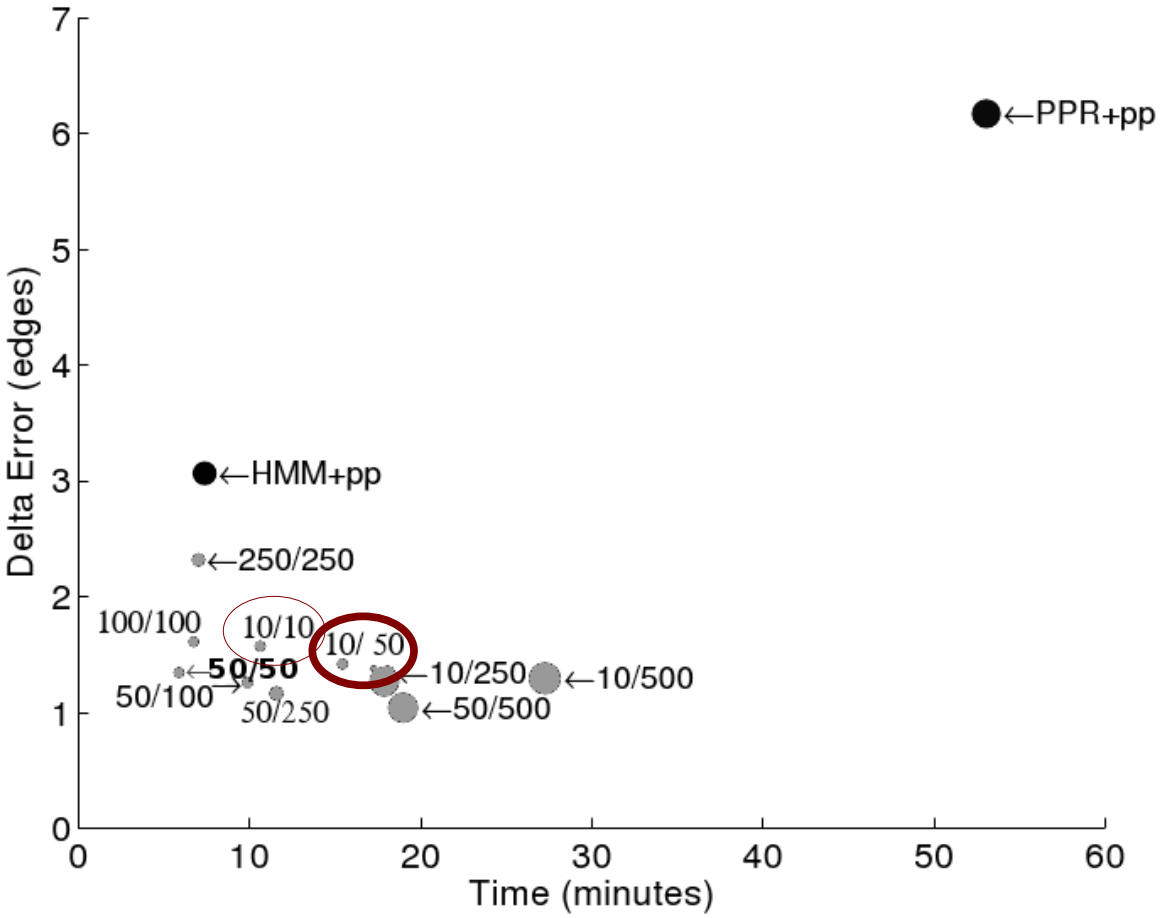
Increasing Placement Size:
10



M2 model condition, 500 true backbone

SEPP Parameters: Simulated

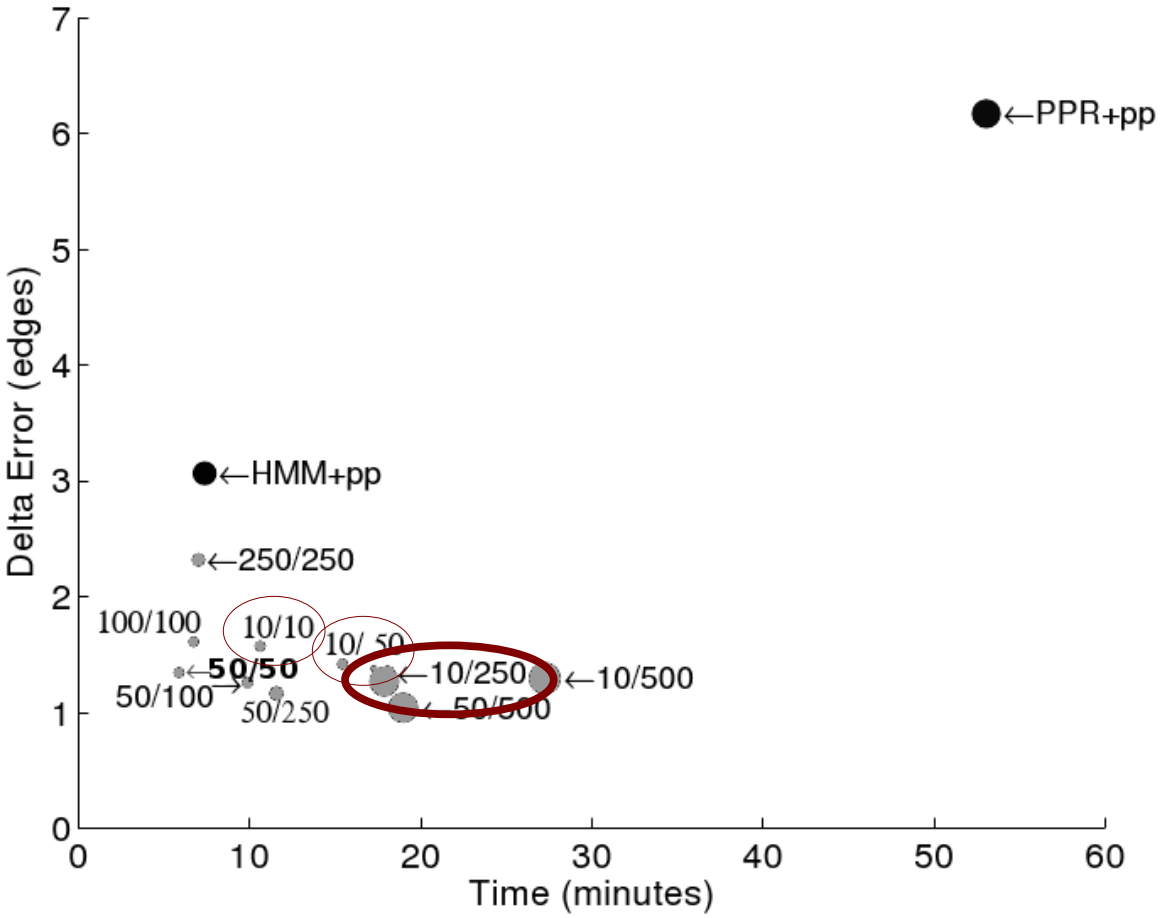
Increasing Placement Size: 10,50



M2 model condition, 500 true backbone

SEPP Parameters: Simulated

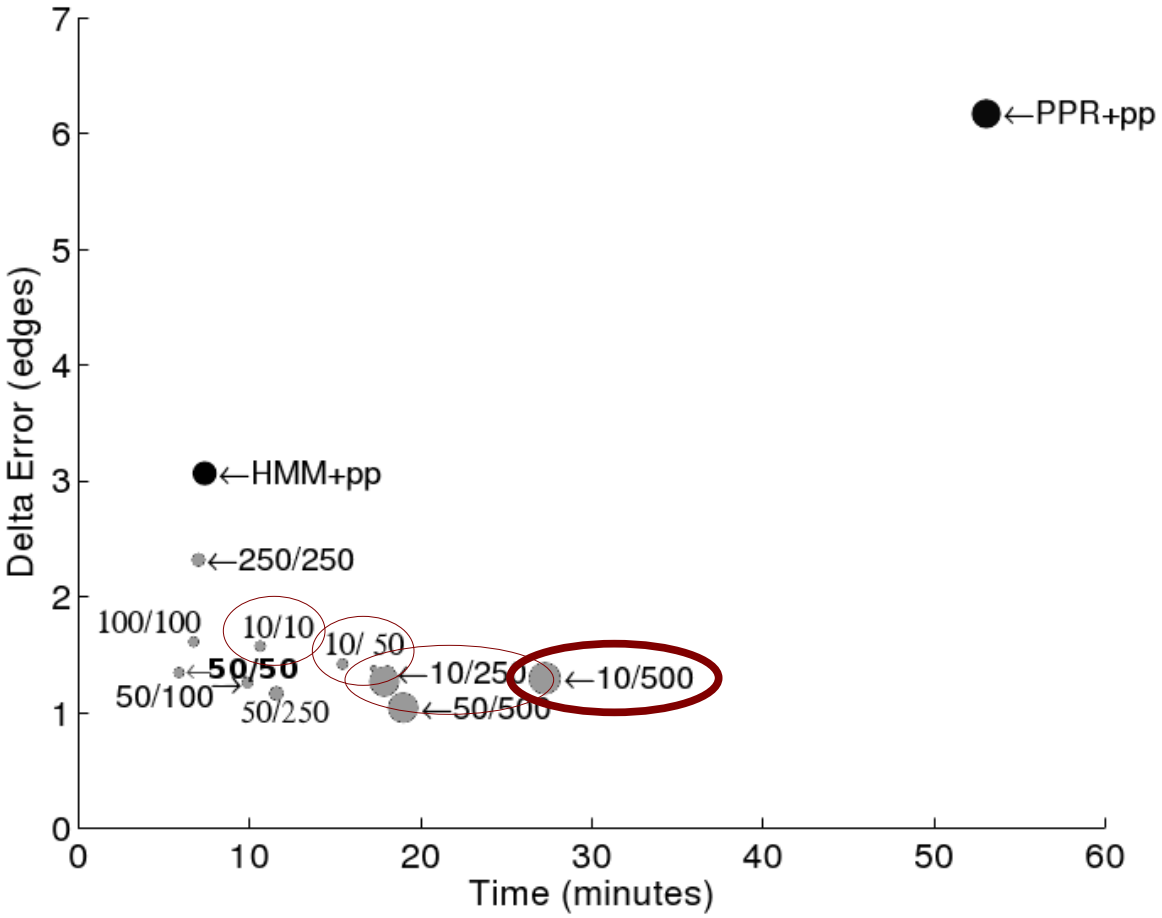
Increasing Placement Size: 10,50,250



M2 model condition, 500 true backbone

SEPP Parameters: Simulated

Increasing Placement Size: 10,50,250,500

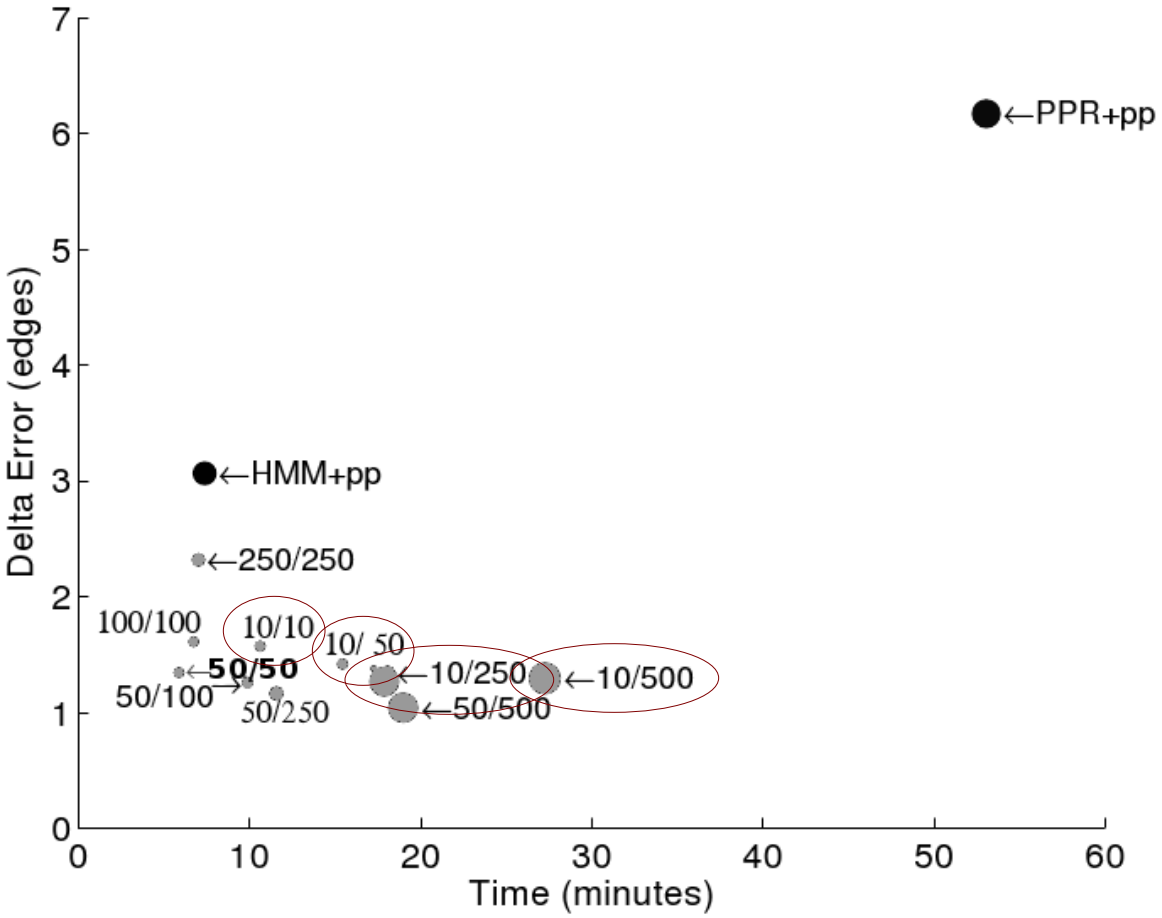


M2 model condition, 500 true backbone

SEPP Parameters: Simulated

Increasing Placement Size: 10,50,250,500

Increases: Accuracy

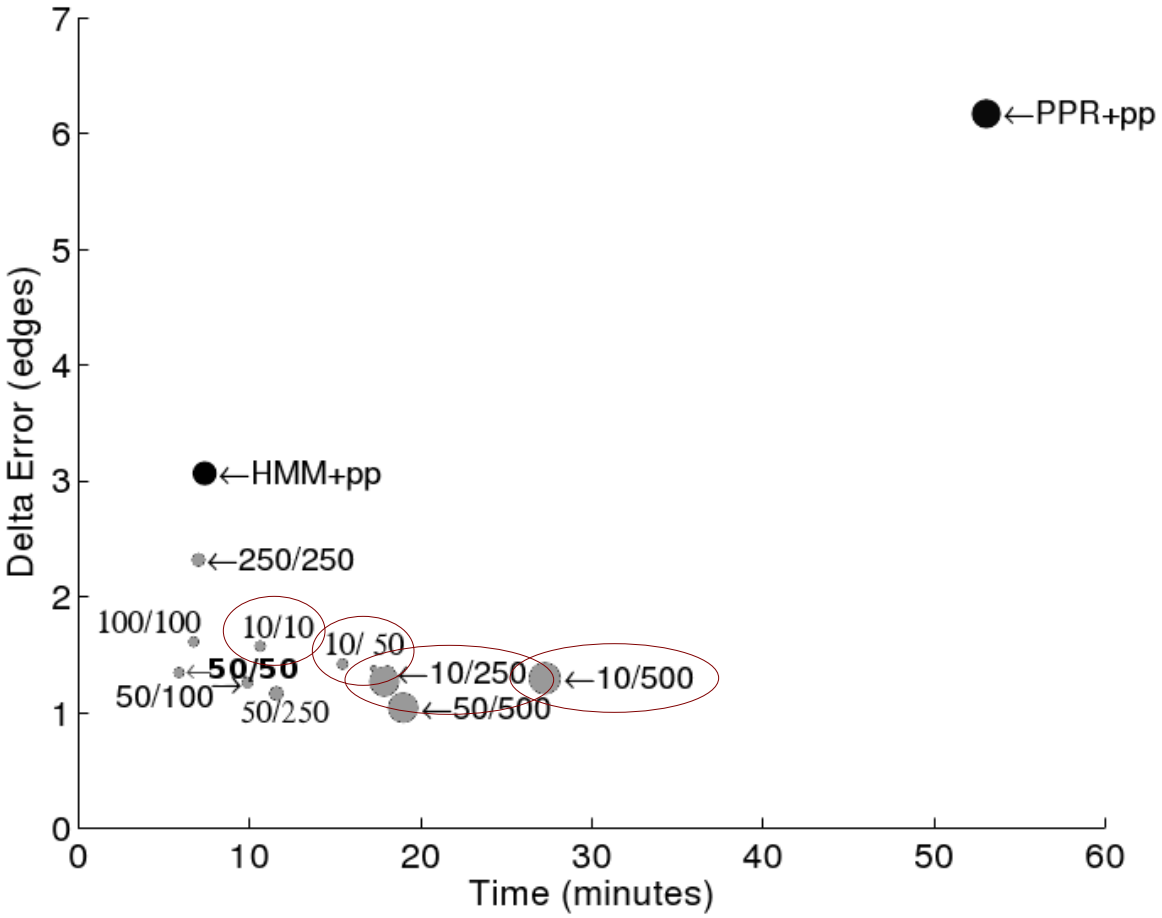


M2 model condition, 500 true backbone

SEPP Parameters: Simulated

Increasing Placement Size: 10,50,250,500

Increases: Accuracy Memory

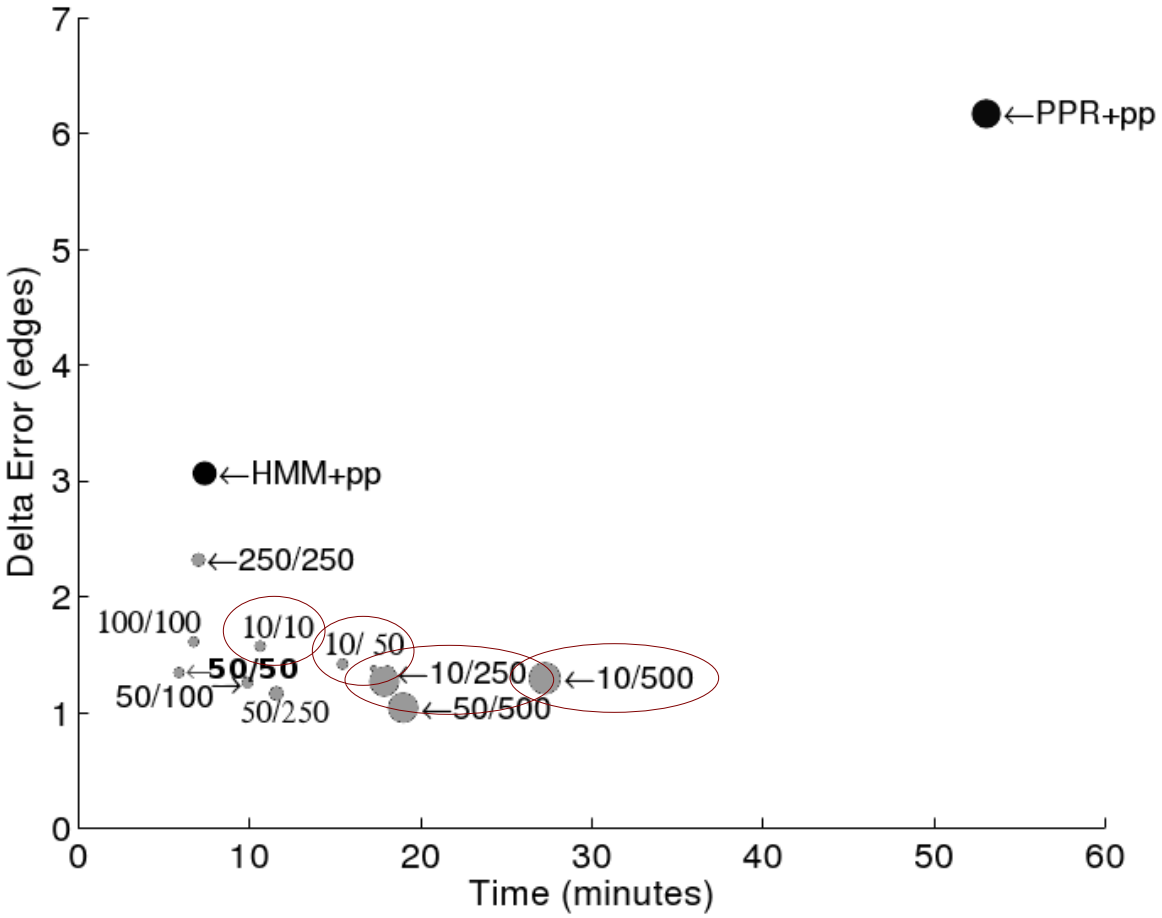


M2 model condition, 500 true backbone

SEPP Parameters: Simulated

Increasing Placement Size: 10,50,250,500

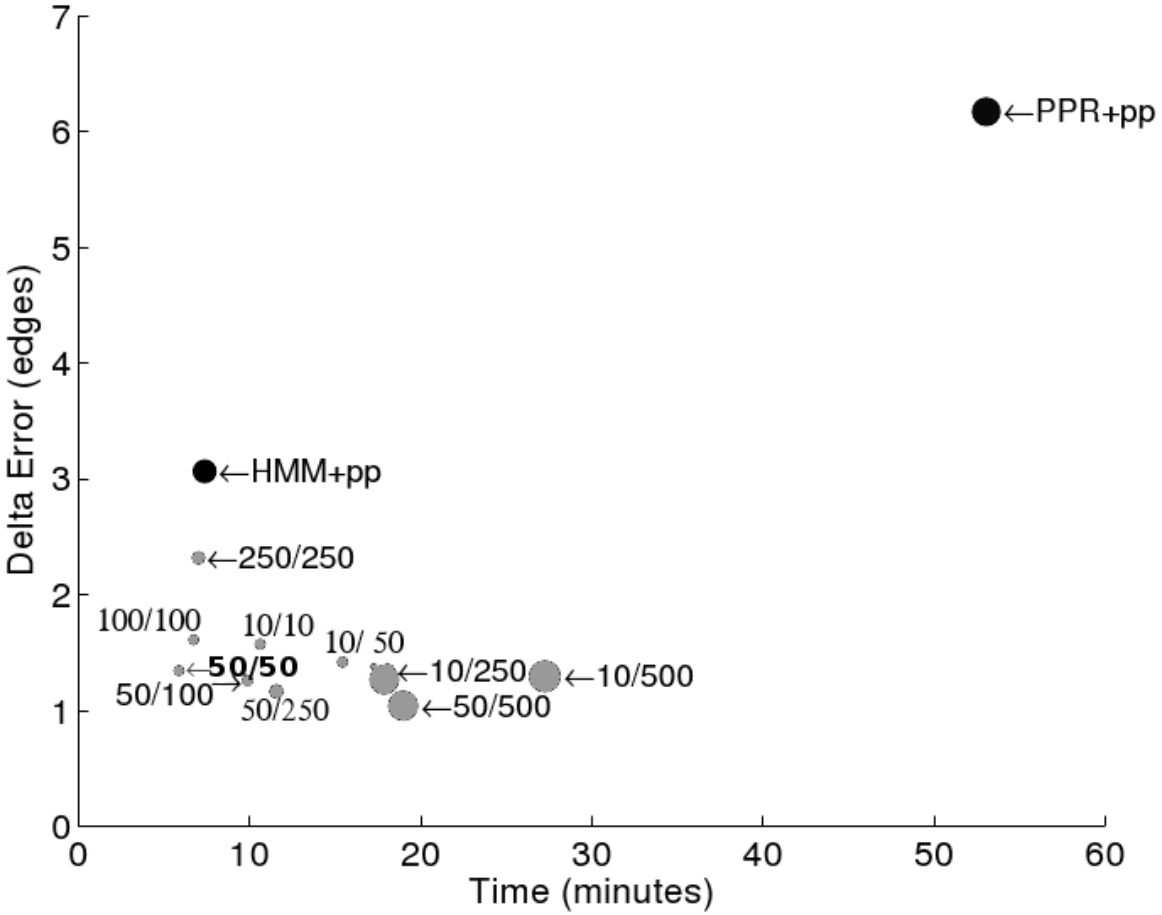
Increases:
Accuracy
Memory
Running time



M2 model condition, 500 true backbone

SEPP Parameters: Simulated

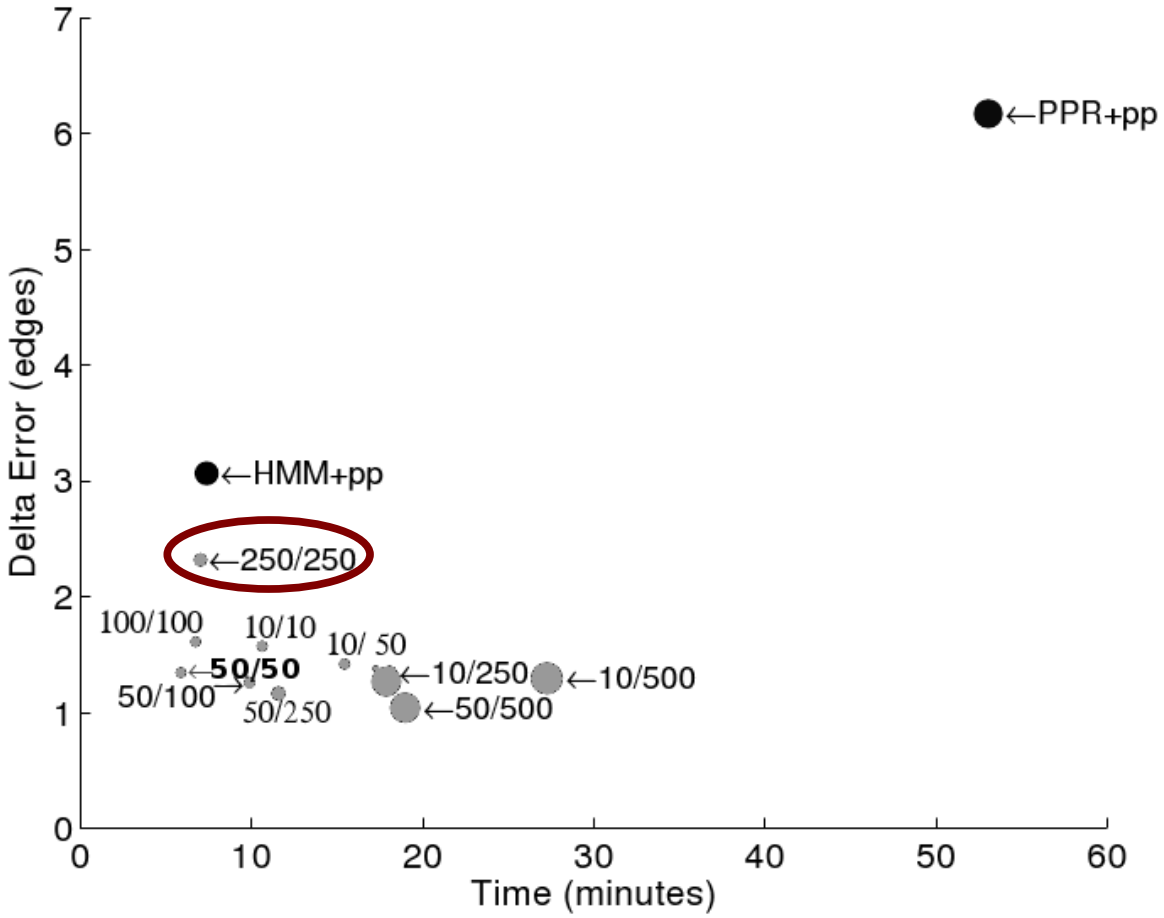
Decreasing Alignment Size:



M2 model condition, 500 true backbone

SEPP Parameters: Simulated

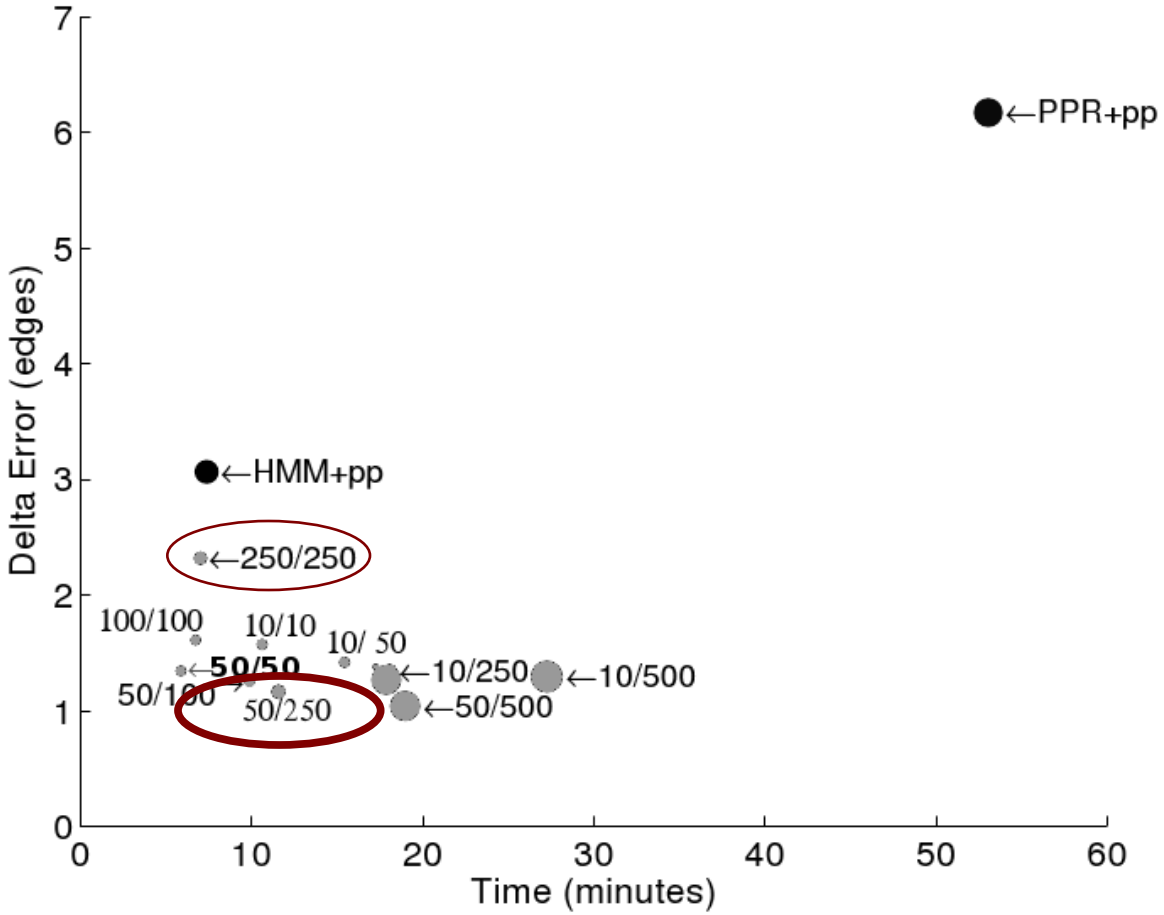
Decreasing Alignment Size:
250



M2 model condition, 500 true backbone

SEPP Parameters: Simulated

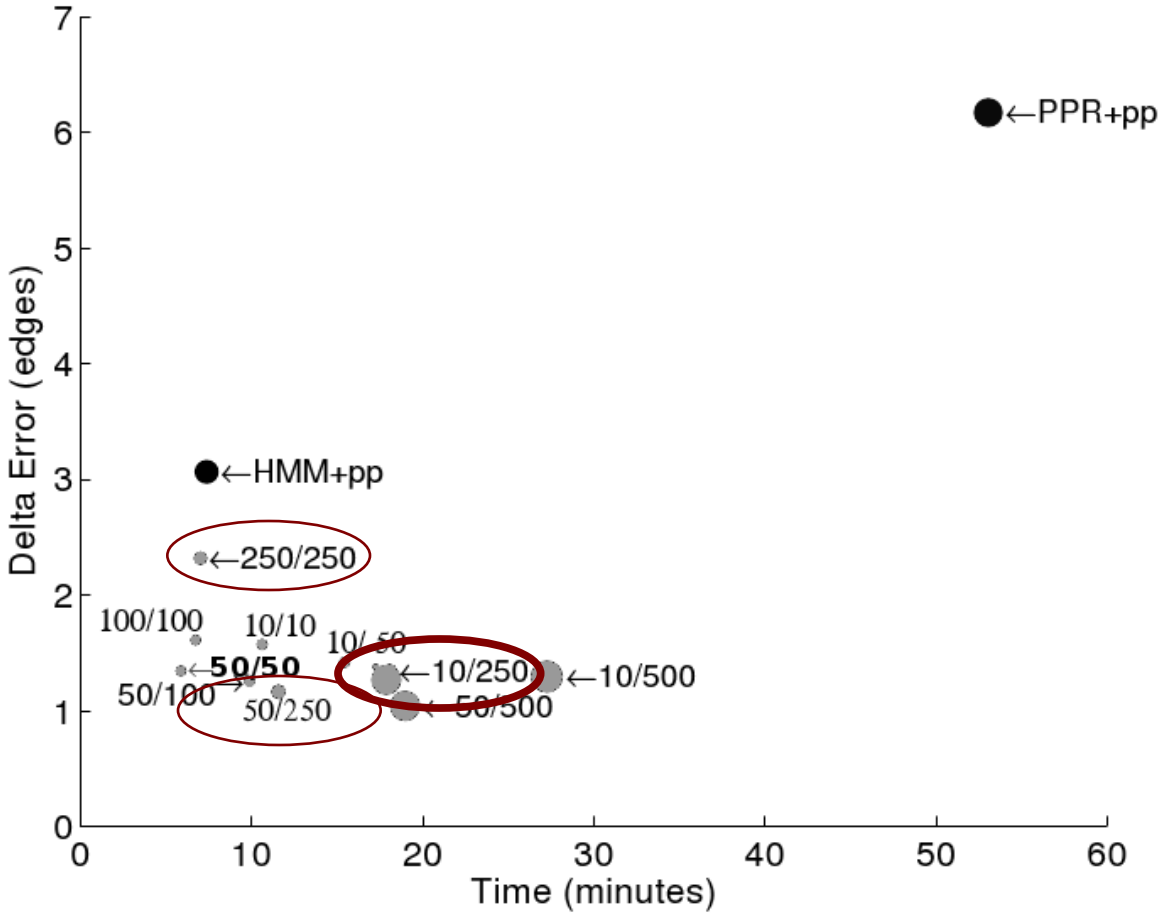
Decreasing Alignment Size: 250,50



M2 model condition, 500 true backbone

SEPP Parameters: Simulated

Decreasing Alignment Size: 250,50,10

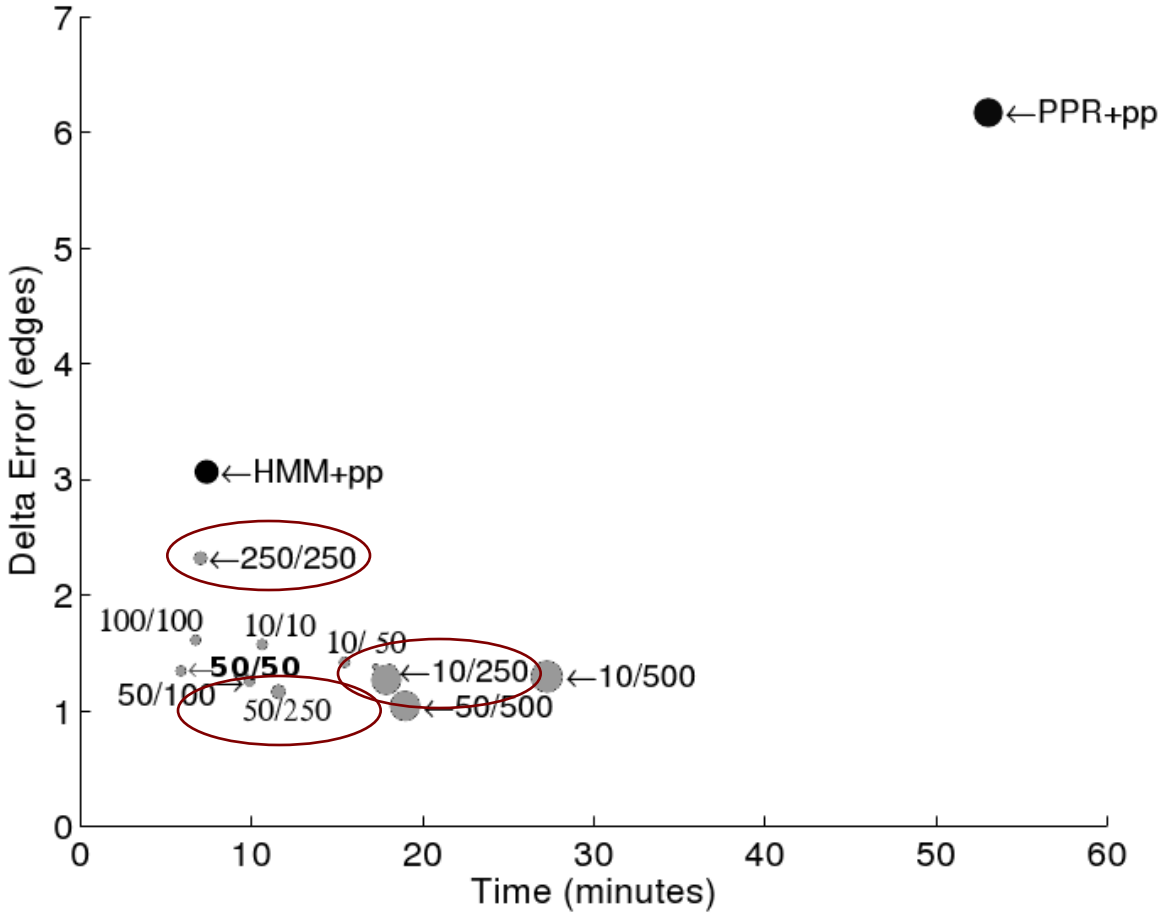


M2 model condition, 500 true backbone

SEPP Parameters: Simulated

Decreasing Alignment Size: 250,50,10

Increases:

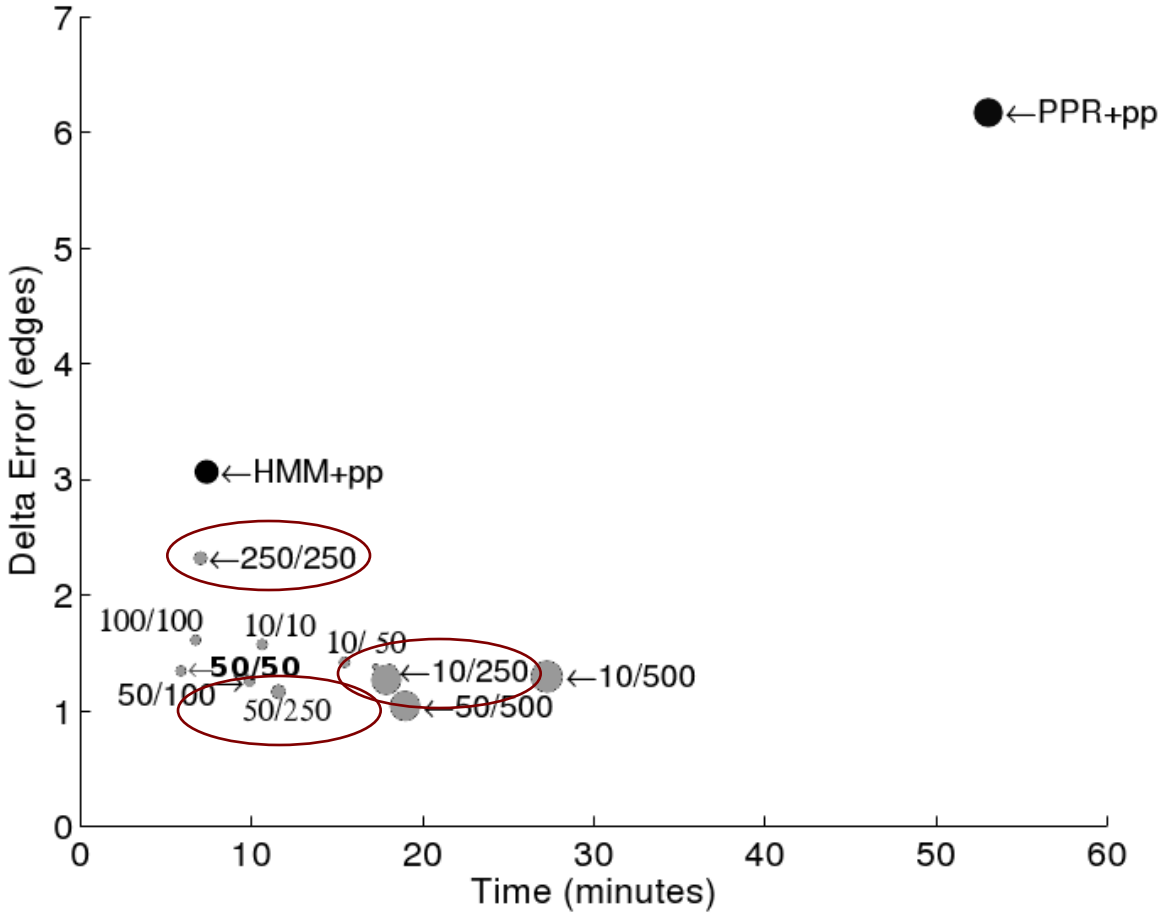


M2 model condition, 500 true backbone

SEPP Parameters: Simulated

Decreasing Alignment Size: 250,50,10

Increases: Running time

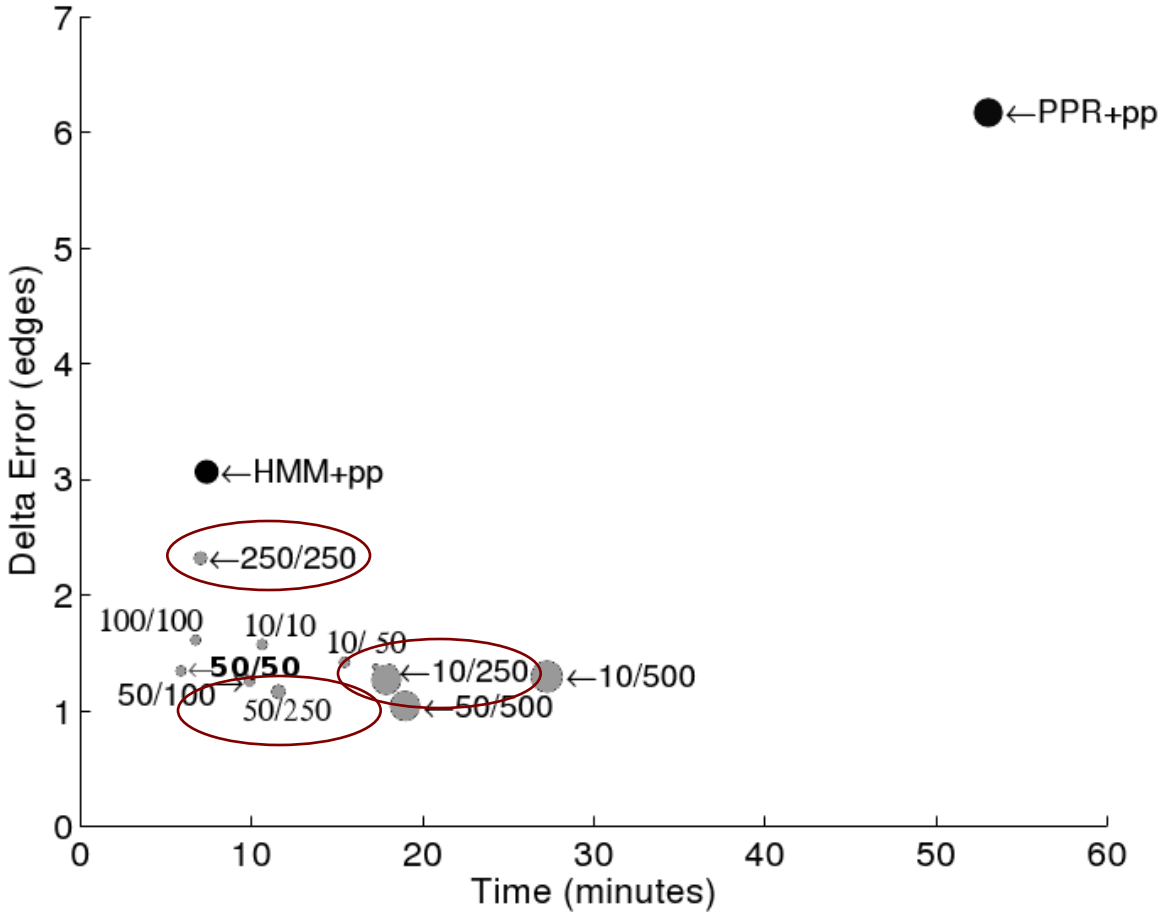


M2 model condition, 500 true backbone

SEPP Parameters: Simulated

Decreasing Alignment Size: 250,50,10

Increases: Running time
Accuracy?

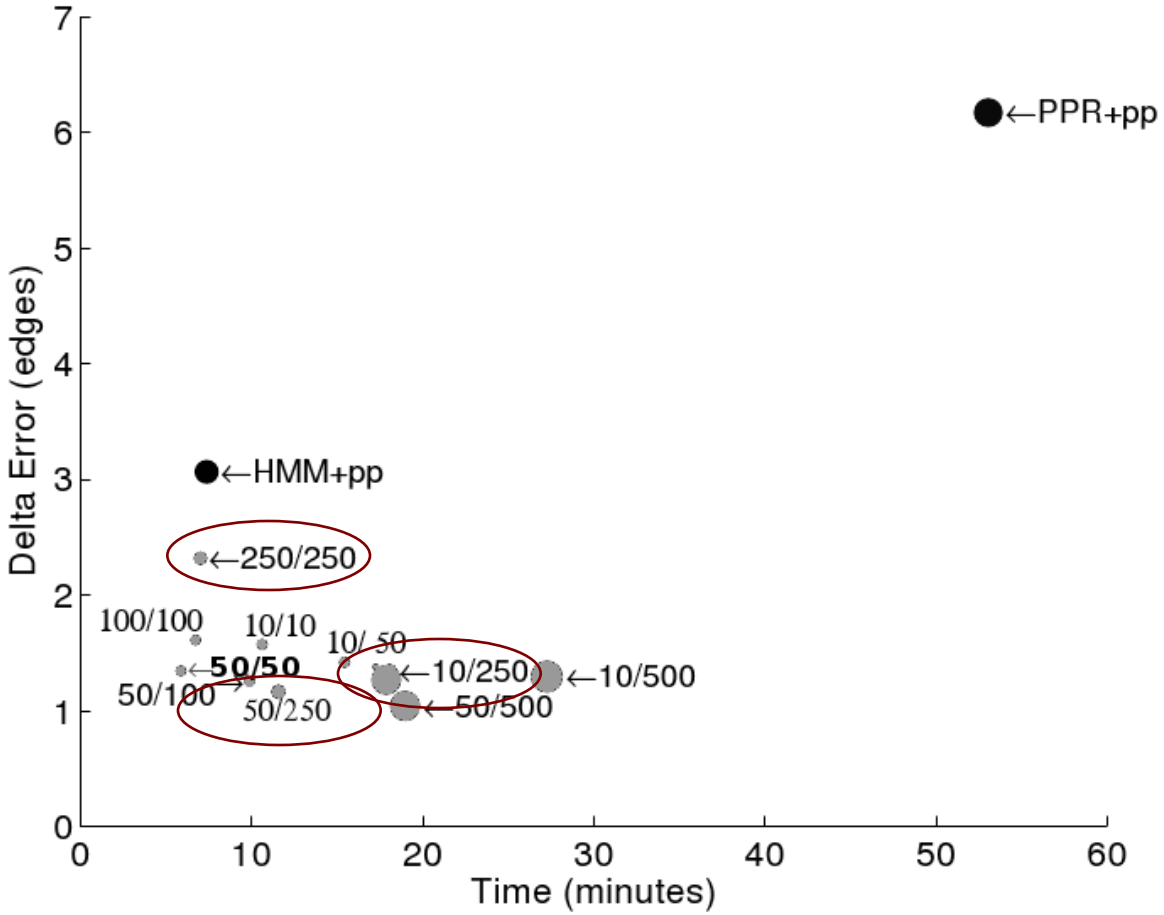


M2 model condition, 500 true backbone

SEPP Parameters: Simulated

Decreasing Alignment Size: 250,50,10

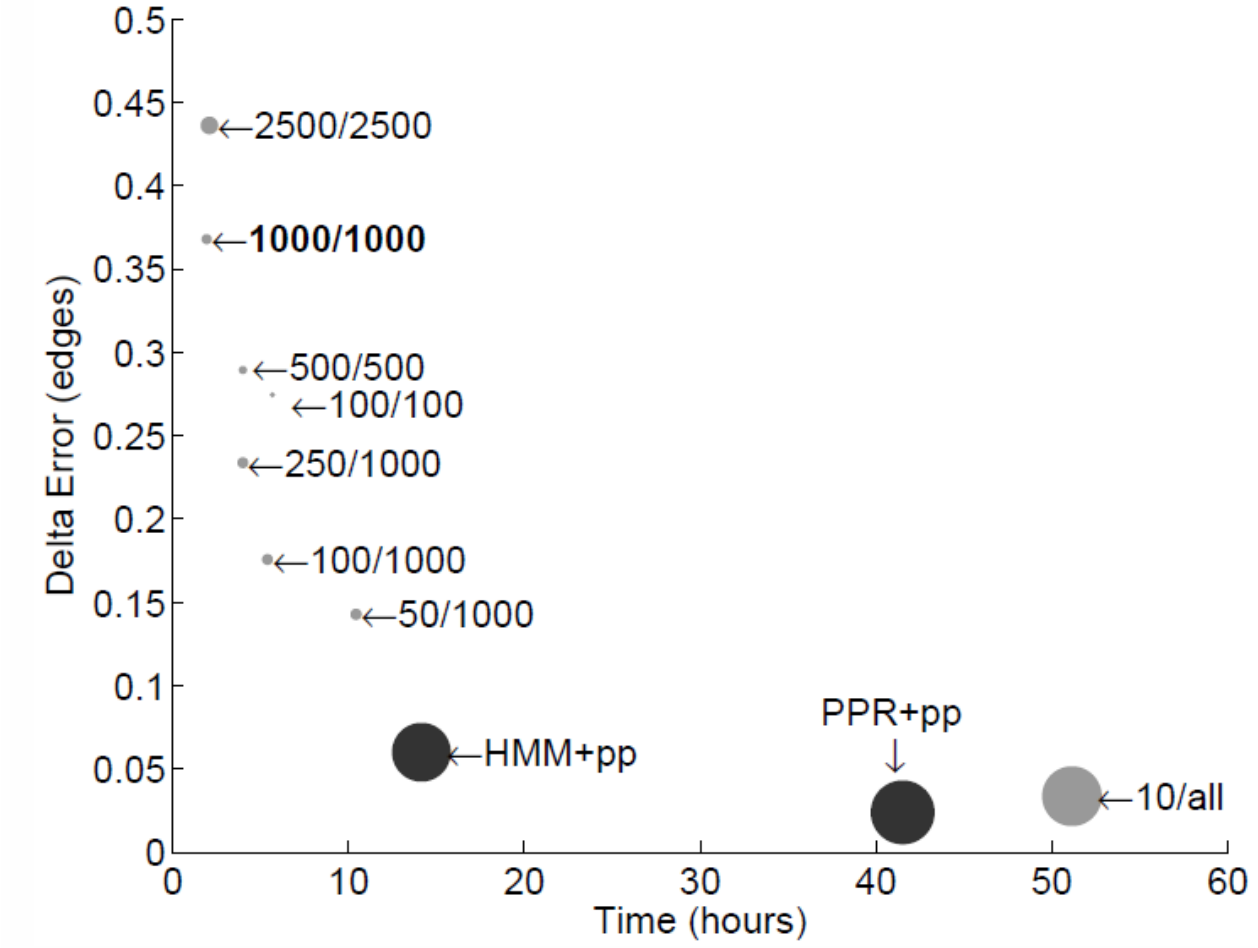
Increases: Running time Accuracy?



M2 model condition, 500 true backbone

SEPP Parameters: Biological

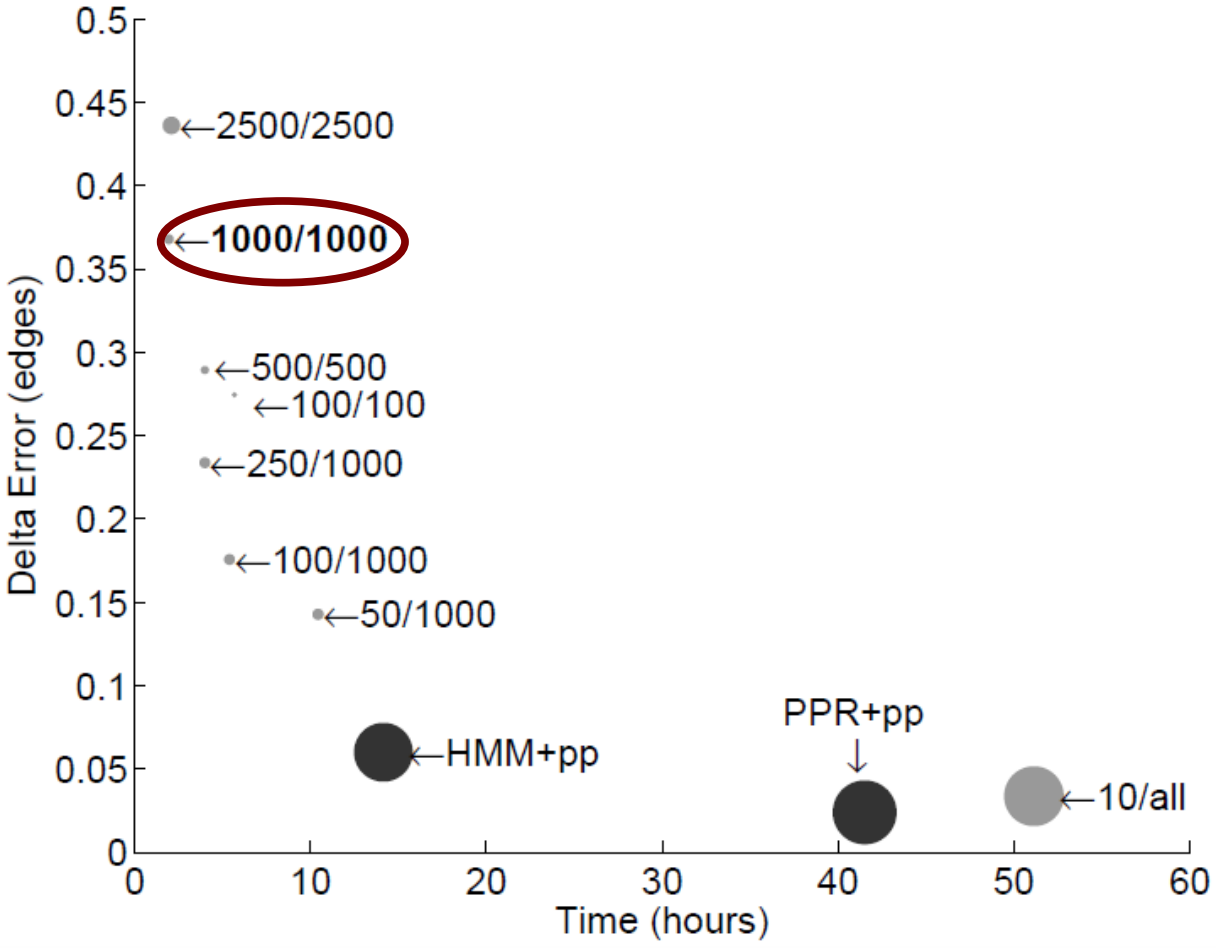
Decreasing Alignment Size:



16S.B.ALL dataset, 13k curated backbone

SEPP Parameters: Biological

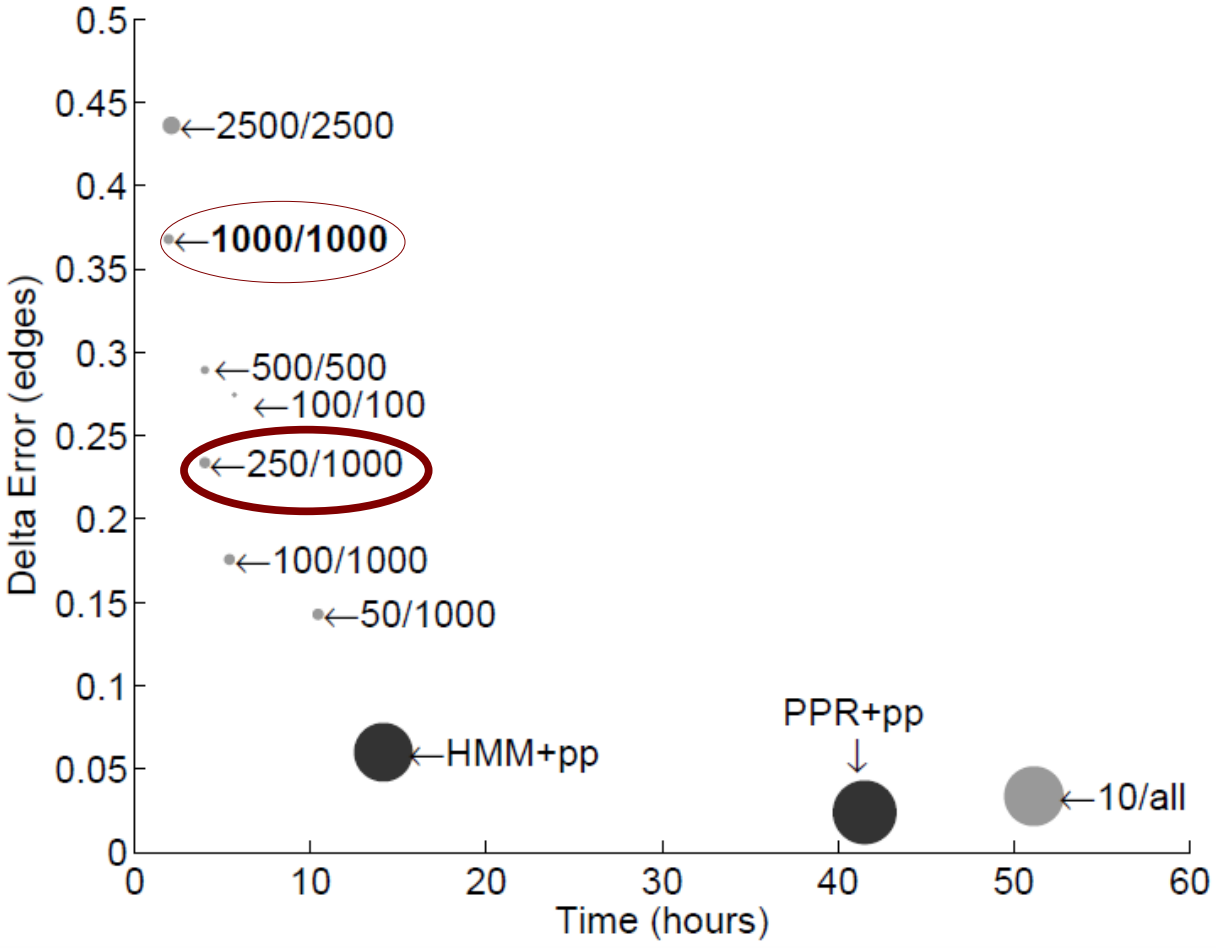
Decreasing Alignment Size: 1000



16S.B.ALL dataset, 13k curated backbone

SEPP Parameters: Biological

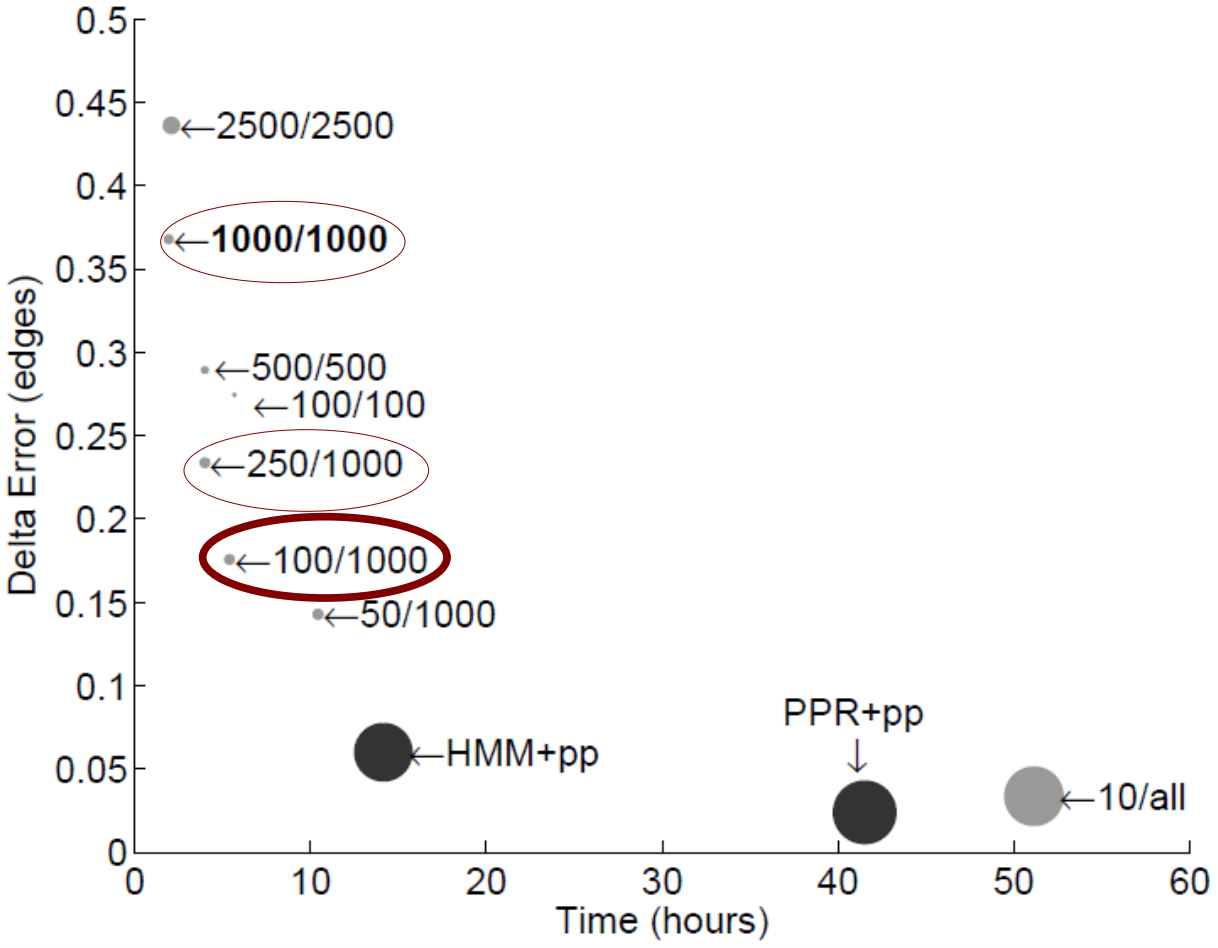
Decreasing Alignment Size: 1000,250



16S.B.ALL dataset, 13k curated backbone

SEPP Parameters: Biological

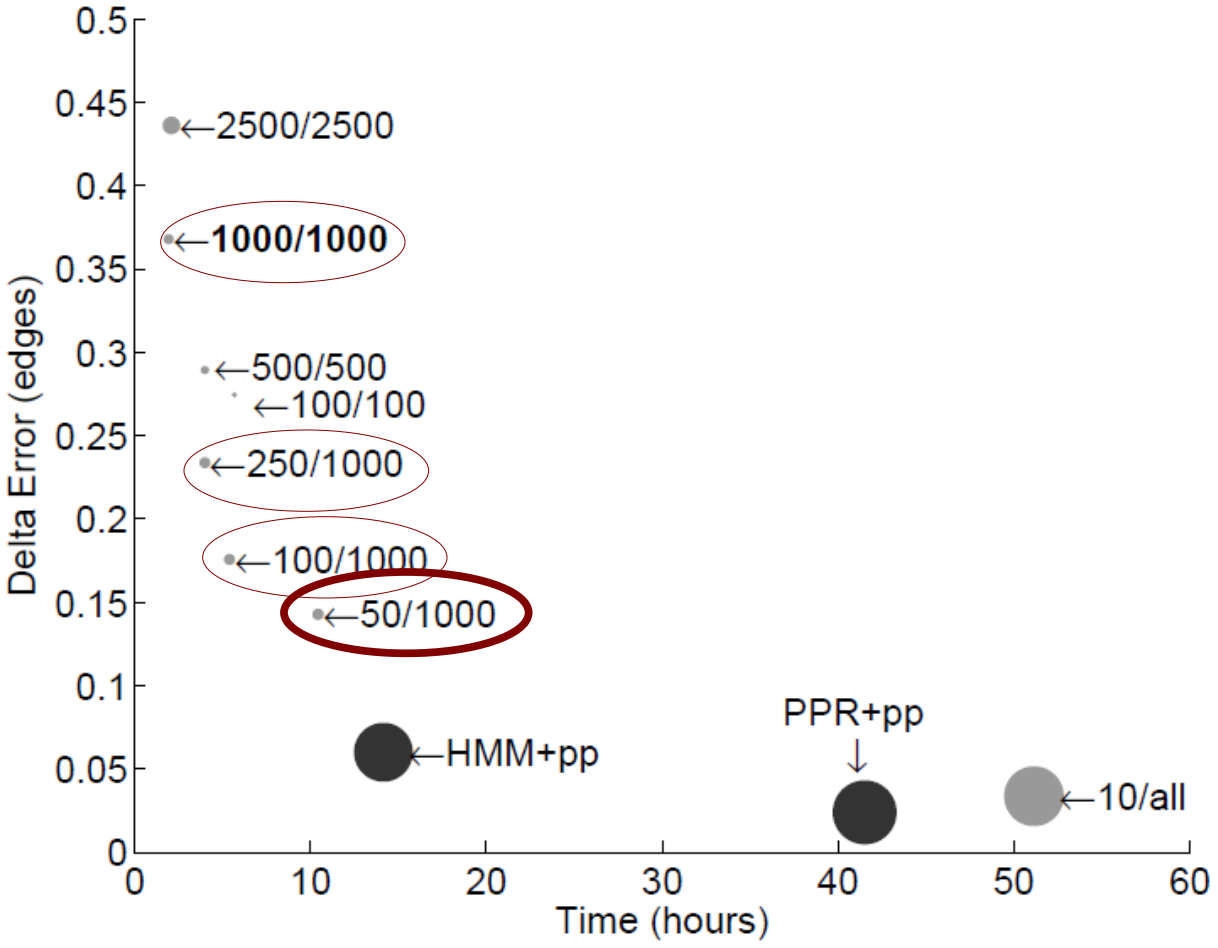
Decreasing Alignment Size: 1000,250, 100



16S.B.ALL dataset, 13k curated backbone

SEPP Parameters: Biological

Decreasing Alignment Size: 1000,250, 100,50

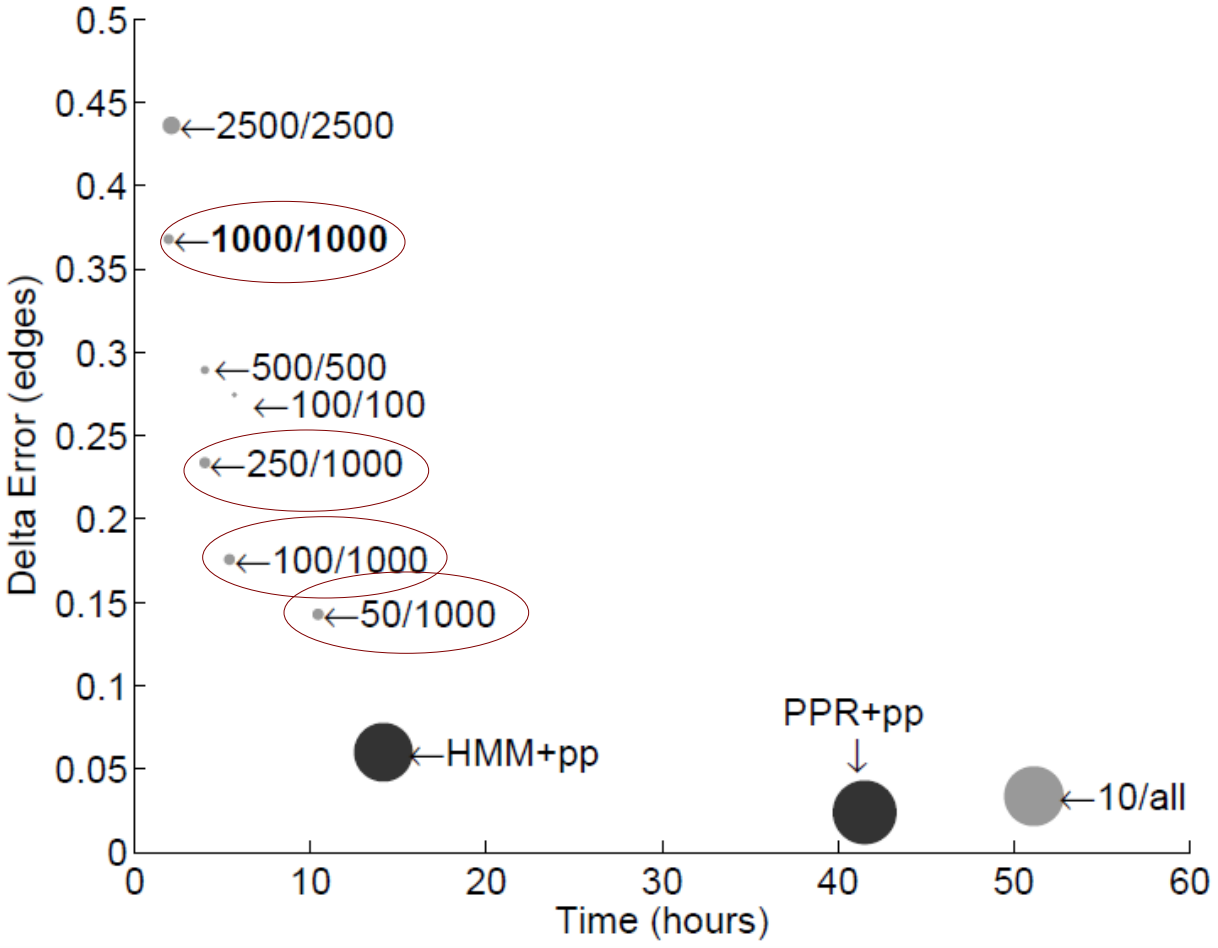


16S.B.ALL dataset, 13k curated backbone

SEPP Parameters: Biological

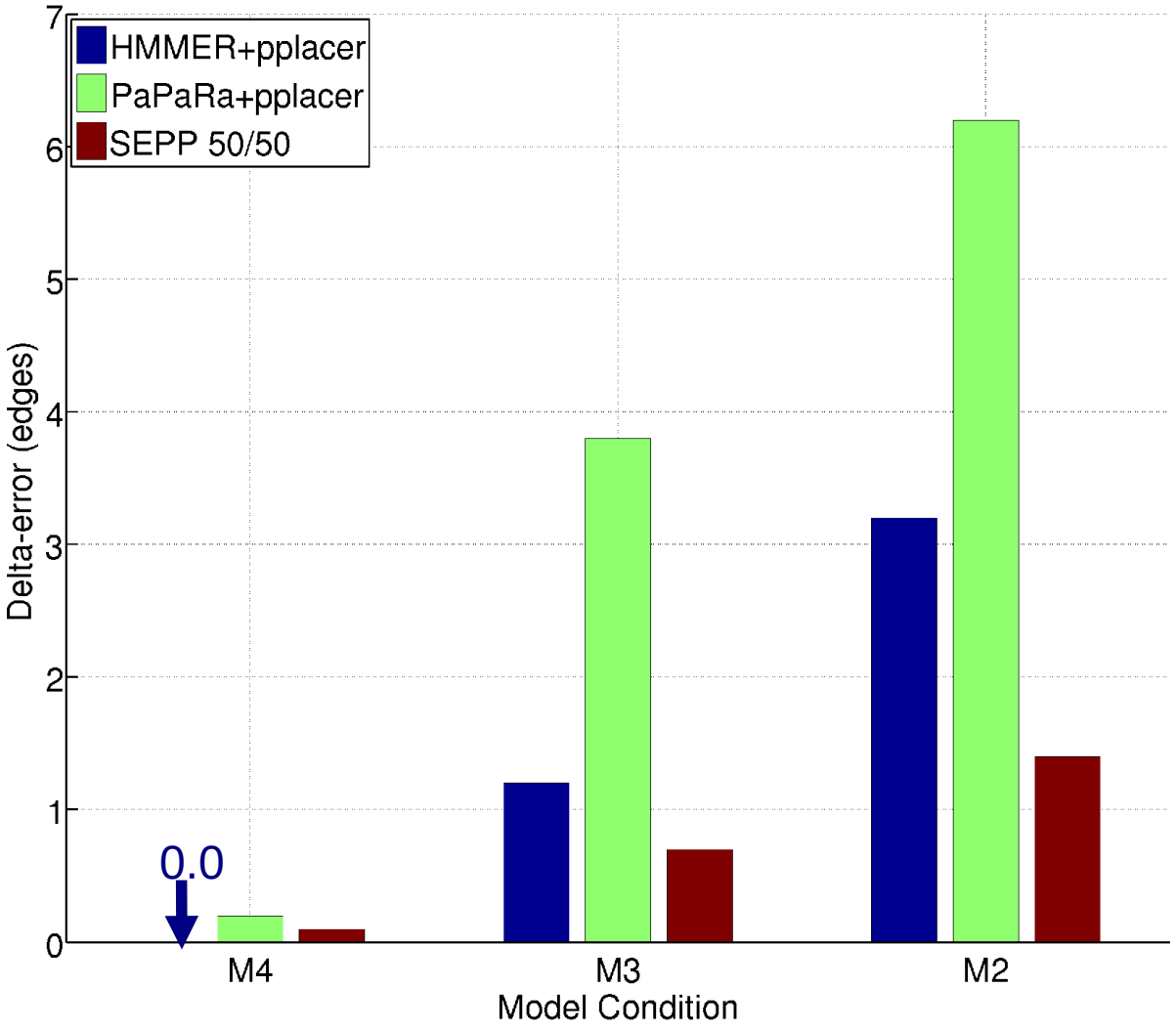
Decreasing Alignment Size: 1000,250, 100,50

Increases: Running time Accuracy



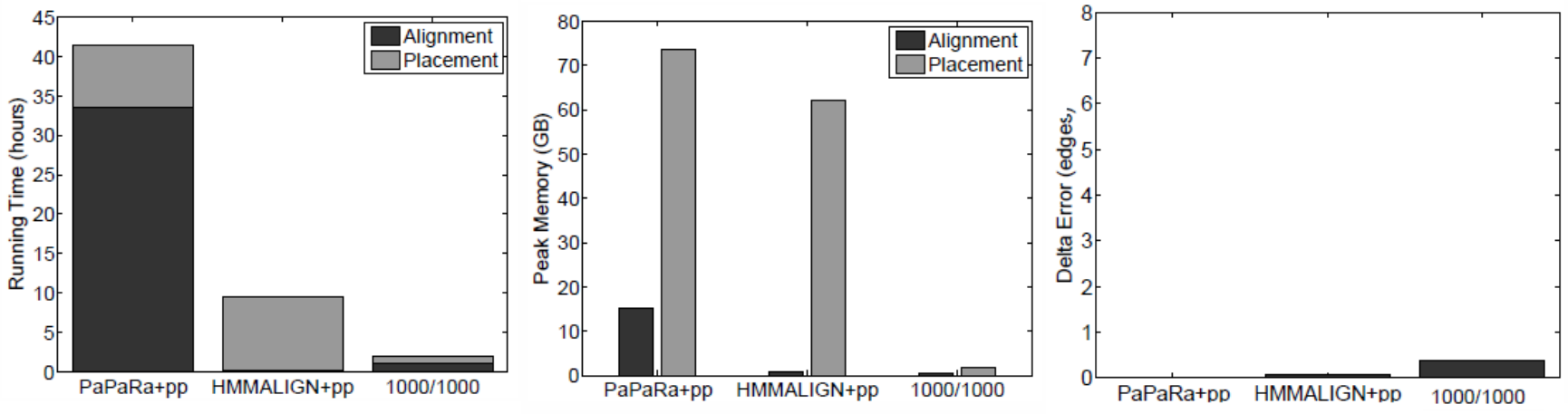
16S.B.ALL dataset, 13k curated backbone

SEPP (10% rule) Simulated Results



Backbone size: 500
5000 fragments
20 replicates

SEPP Biological Results



16S.B.ALL dataset, curated alignment/tree, 13k backbone, 13k total fragments

For 1 million fragments:

PaPaRa+pplacer: ~133 days

HMMER+pplacer: ~30 days

SEPP 1000/1000: ~6 days

SEPP summary

- Two stages of decomposition
  - Placement decomposition to form placement sets
  - Alignment decomposition to form fHMM
- Results in 40% lower placement error than HMMER+pplacer on divergent datasets
- 1/5 running time of HMMER+pplacer on large backbones
- Local placement uses less than 2 GB peak memory compared to 60-70 GB peak memory for global placement

Outline of Talk

- Background
- Family of HMMs
  - Model
  - Alignment algorithm
- Applications of fHMM
  - SEPP (Mirarab, Nguyen, and Warnow. PSB 2012)
  - TIPP (Nguyen, et al. Under review)
- Conclusions and future work

Taxonomic Identification and Profiling

- Taxonomic identification
  - Objective: Given a query sequence, identify the taxon (species, genus, family, etc...) of the sequence
  - Classification problem
  - Methods include Megan, PhymmBL, Metaphyler, and MetaPhylAn
- Taxonomic profiling
  - Objective: Given a set of query sequences collected from a sample, estimate the population profile of the sample
  - Estimation problem
  - Can be solved via taxonomic identification

Using SEPP

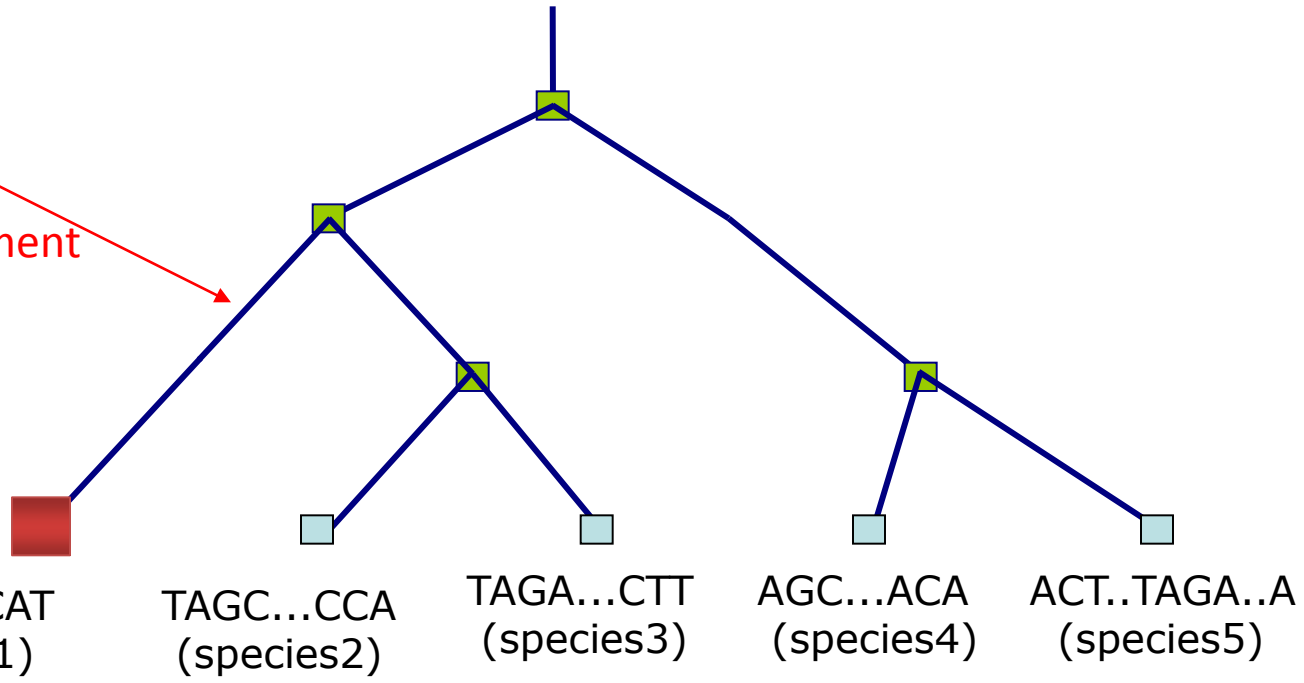
Fragmentary Unknown Reads:

(60-200 bp long)

- ACCG
- CGAG
- CGG
- GGCT
- TAGA
- GGGGG
- TCGAG
- GGCG
- GGG
- .
- .
- .
- ACCT

Known Full length Sequences,
and an alignment and a tree

(500-10,000 bp long)



Adding uncertainty

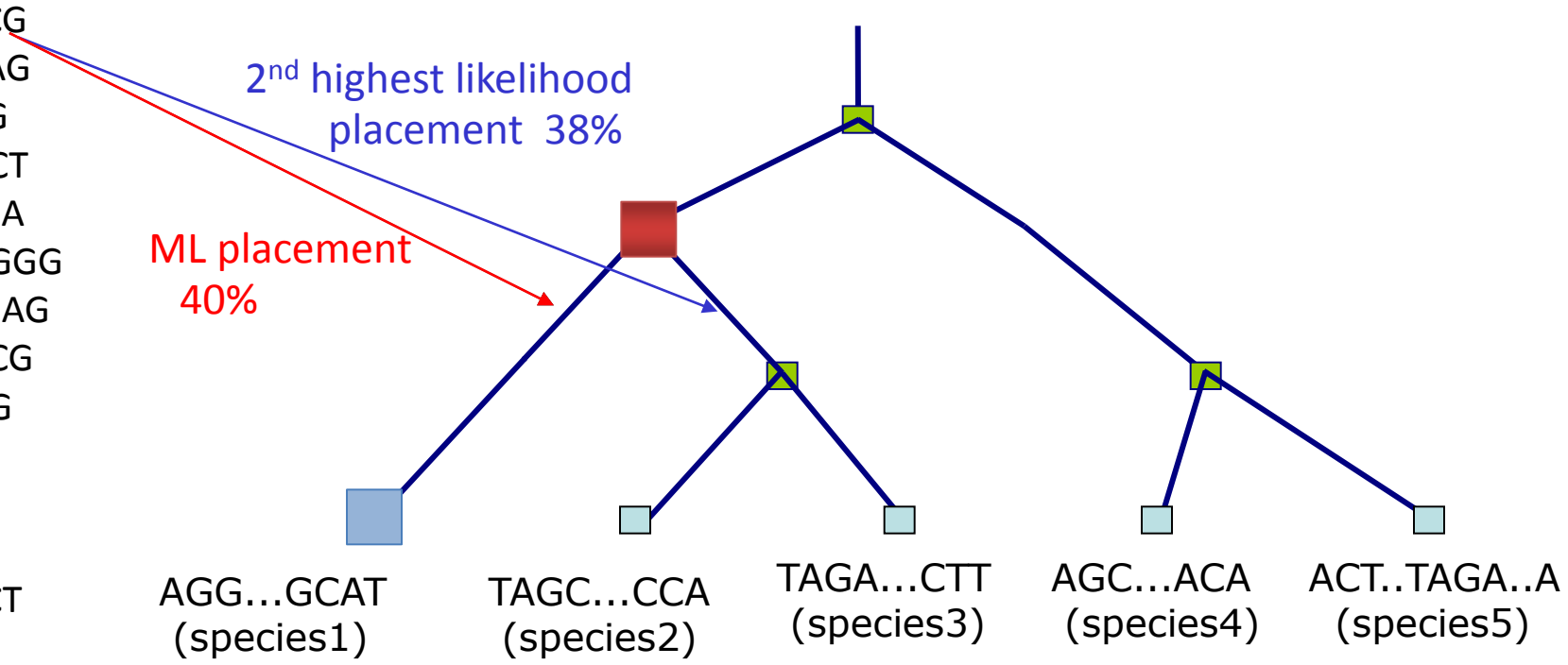
Fragmentary Unknown Reads:

(60-200 bp long)

- ACCG
- CGAG
- CGG
- GGCT
- TAGA
- GGGGG
- TCGAG
- GGCG
- GGG
- .
- .
- .
- ACCT

Known Full length Sequences,
and an alignment and a tree

(500-10,000 bp long)



TIPP: Taxonomic identification and phylogenetic profiling

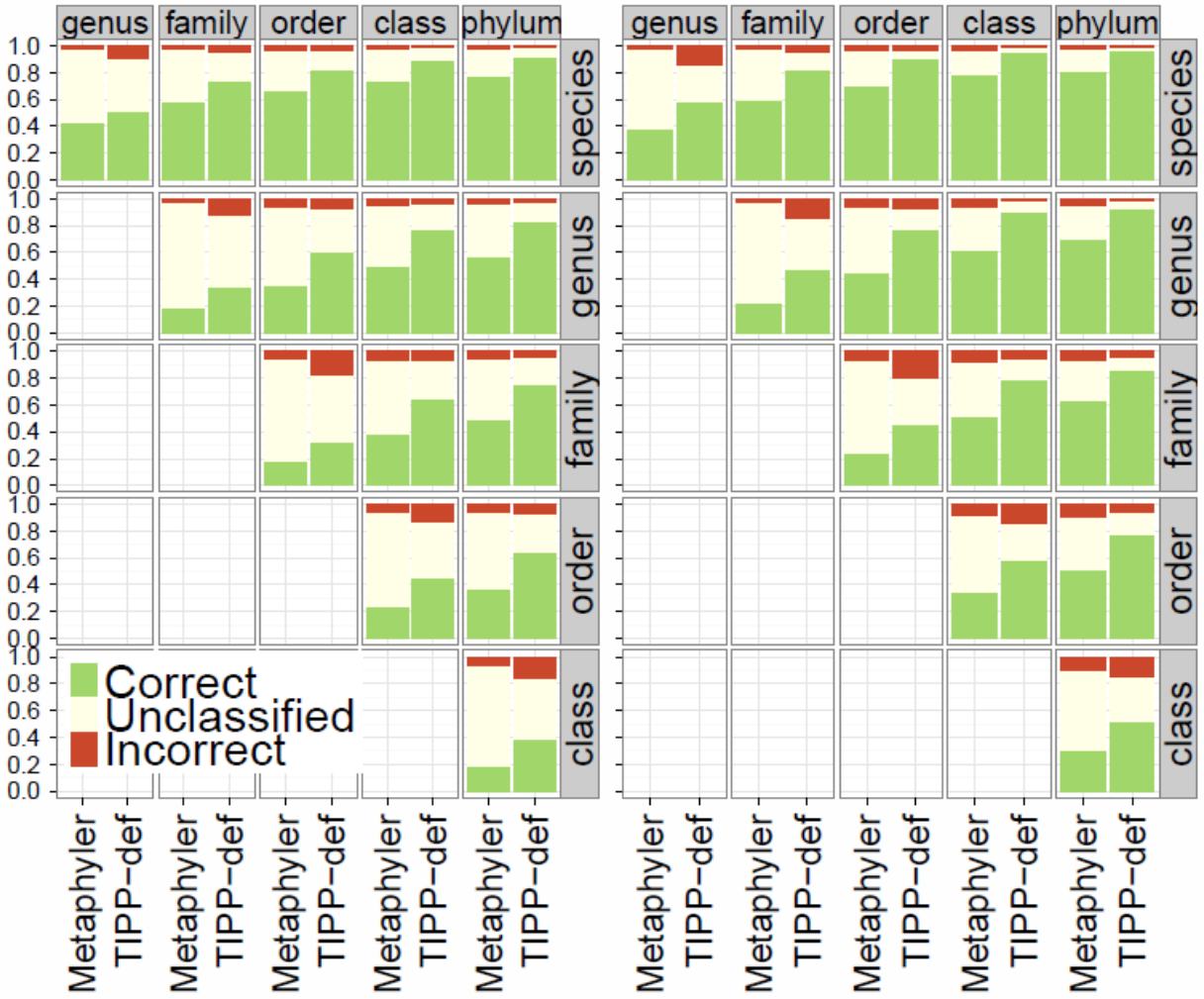
- Developers: Nguyen, Mirarab, Pop, and Warnow
- SEPP takes the best extended alignment and finds the ML placement.
- We modify SEPP to use uncertainty:
  - Find many extended alignments of fragments to each reference alignment to reach support alignment threshold
  - Find many placements of fragments for each extended alignment to reach placement support threshold

- Takes alignment and placement support values
- Classify each fragment at the Lowest Common Ancestor of all placements obtained for the fragment
- Under review

Experimental Design

- Taxonomic identification
  - Used leave-one-out experiments to examine classification accuracy on classifying novel taxa
  - Used non-leave-one-out experiments with fragments simulated under different error models to examine robustness
  - Fragments simulated under Illumina-like and 454-like error models
- Taxonomic profiling
  - Collected simulated datasets from various studies
  - Estimated profiles on simulated samples
  - Computed Root Mean Squared Error for each profile

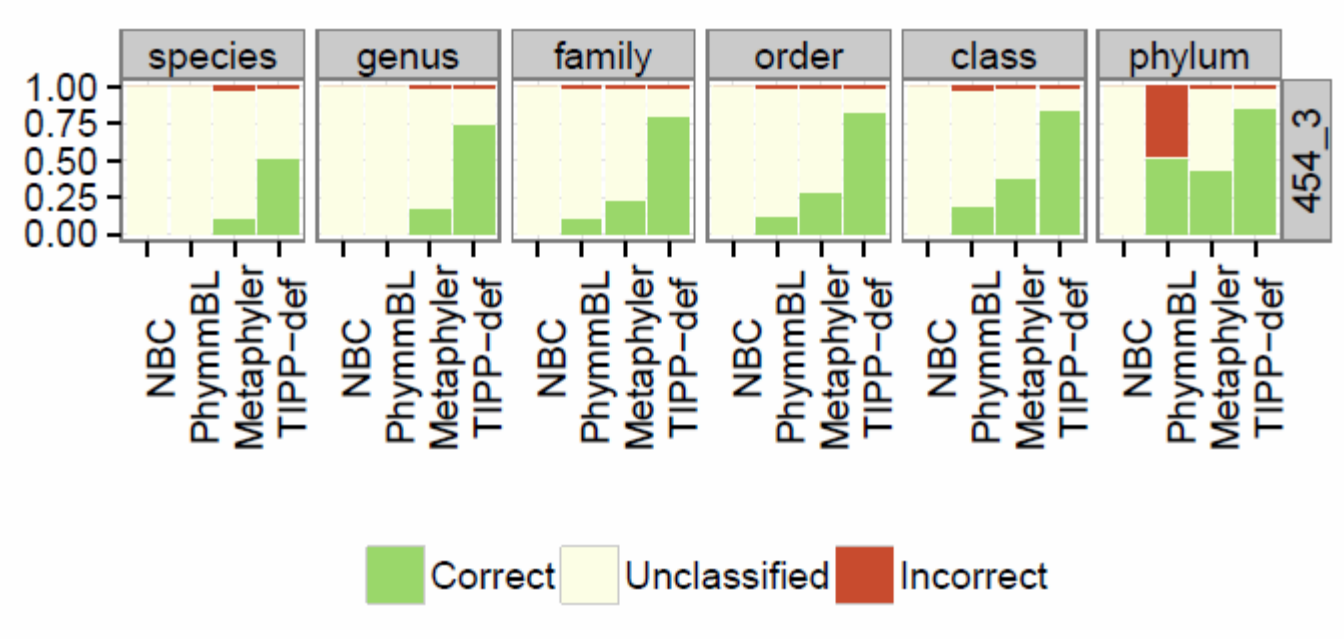
Leave-one-out comparison



(a) Illumina error model

(b) 454 error model

Robustness to sequencing error



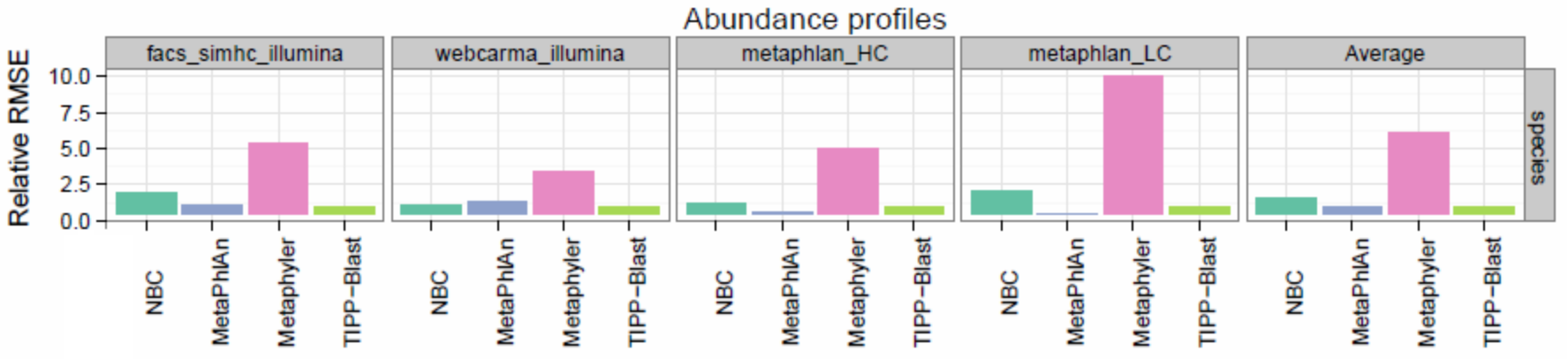
454_3 error model has reads with average length of 285 bps, with **60 indels** per read

Taxonomic profiling

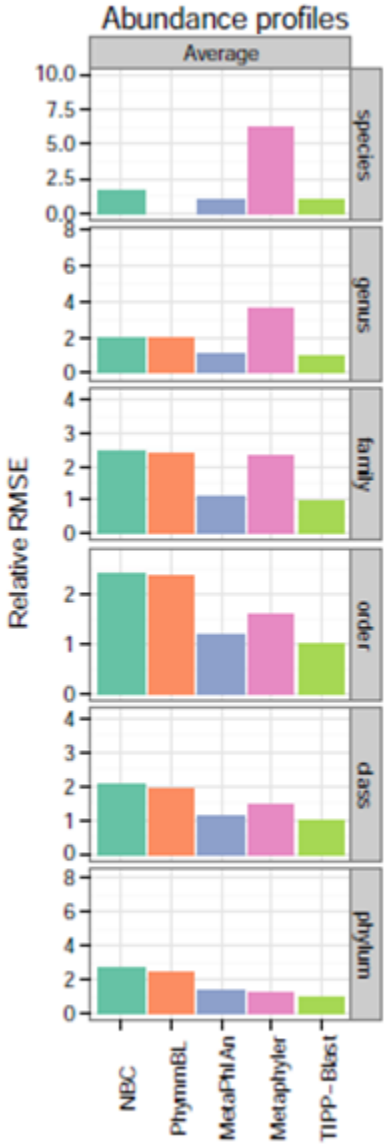
- Selected 9 different simulated metagenomic model conditions
- Divided datasets into two groups:
  - short fragments (<= 100 bps)
  - long fragments (>= 100 bps).
- Report RMSE relative to TIPP's RMSE

Dataset	# Genomes	Complex.	Seq. Model	Reads	Length
MetaPhlAn HC	100	High	NA	1000000	88
MetaPhlAn LC	25	Low	NA	240000	88
FAMeS HC	113	High	DOE-JGI	116771	949
FAMeS MC	113	Medium	DOE-JGI	114457	969
FAMeS LC	113	Low	DOE-JGI	97495	951
FACS HC	19	High	454	26984	268
FACS HC Illumina	19	High	Illumina	300000	100
WebCarma	25	High	454	25000	265
WebCarma Illumina	25	High	Illumina	300000	100

Profiling: Short Fragments

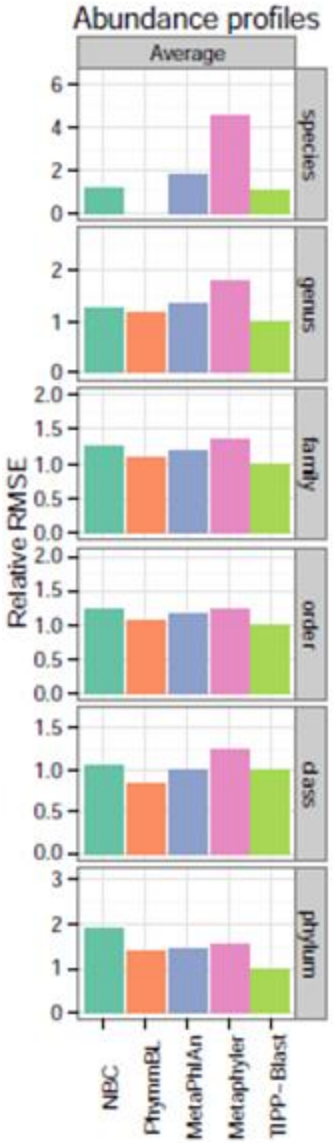


Profiling: Short Fragments



Note: PhymmBL does not report species level classification

Profiling: Long Fragments



Note: PhymmBL does not report species level classification

TIPP Summary

- Combines SEPP with statistical support threshold to increase precision with minor reduction in sensitivity
- Better sensitivity for classifying novel reads compared to MetaPhyler
- Very robust to sequencing errors
- Results in overall more accurate profiles (lowest average error in 10 of 12 conditions)
- Can be parameterized for precision or sensitivity

Summary

- fHMM as a statistical model for MSA
- Algorithm for alignment using fHMM
  - Computes HMMs on closely related subsets
  - Aligns query sequence to fHMM
- fHMM improves sequence alignment to an existing alignment

Future work

- Use fHMM as a replacement for profile HMM in other domains
  - Homology detection
  - Functional annotation
- Use different alignment methods within fHMM framework
- TIPP
  - Statistical models for combining profiles on different markers
  - Expand marker sets to include more genes

Acknowledgements

Siavash Mirarab

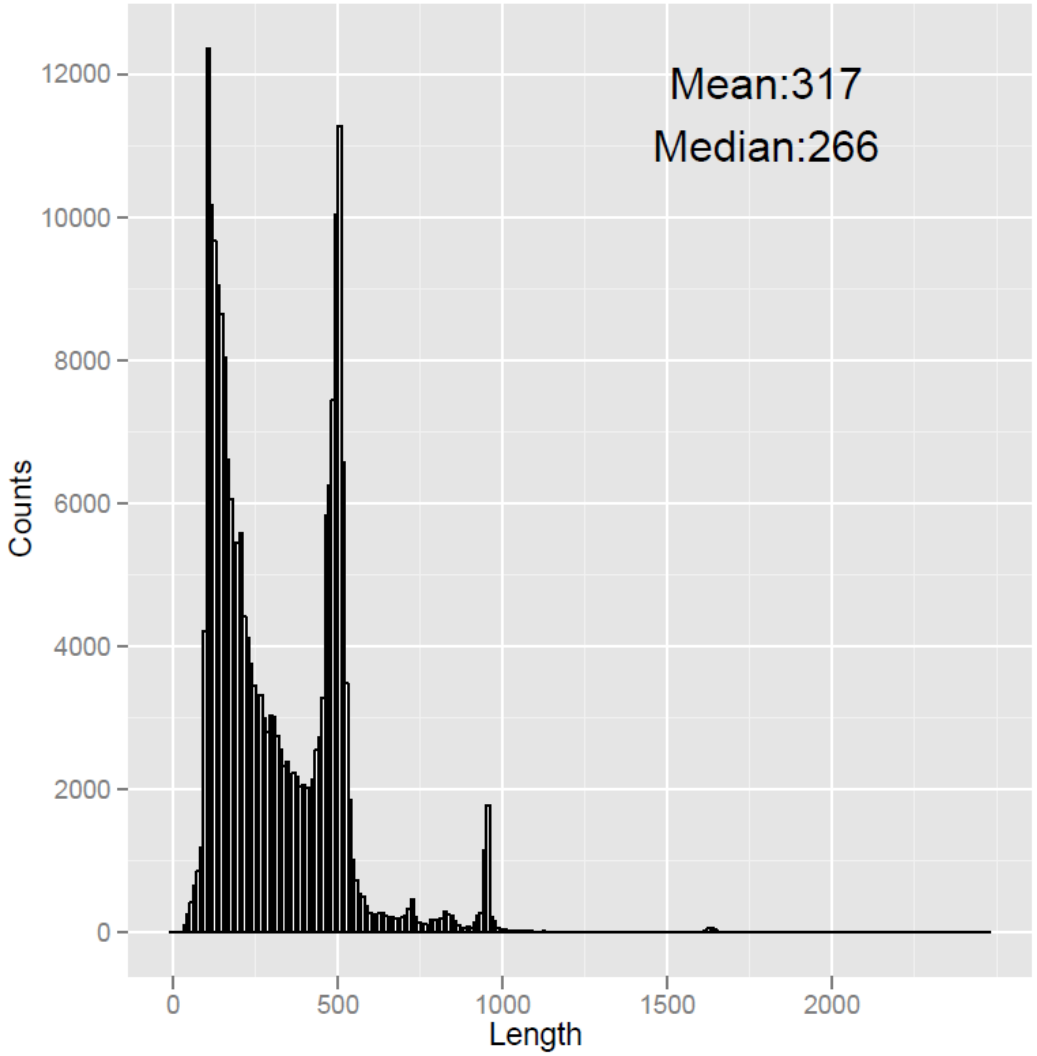
Mihai Pop

Bo Liu

Tandy Warnow

Supported by
NSF DEB 0733029
University of Alberta

1KP P450 transcriptome dataset



Full-length P450 gene ~500 AA

Total sequences before filtering ~225K

Ultra-large sequence alignment

- Most MSA techniques do not grow linearly with number of sequences
- Alignments are needed on very large datasets
  - Pfam contains families with more than 100,000 sequences
  - More than 1 million 16S sequences in Green Genes DB
- Datasets can contain fragmentary and full-length sequences

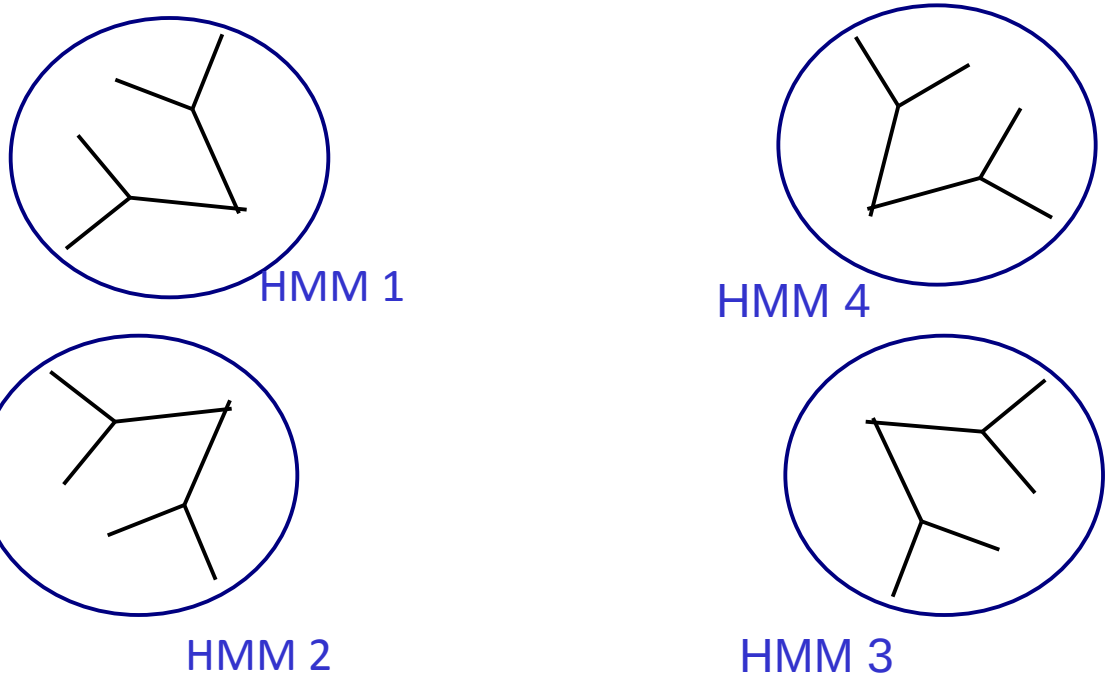
HMMs for MSA

- Given seed alignment (e.g., in PFAM) and a collection of sequences for the protein family:
  - Represent seed alignment using HMM
  - Align each additional sequence to the HMM
  - Use transitivity to obtain MSA
- Can we do something like this without a seed alignment?

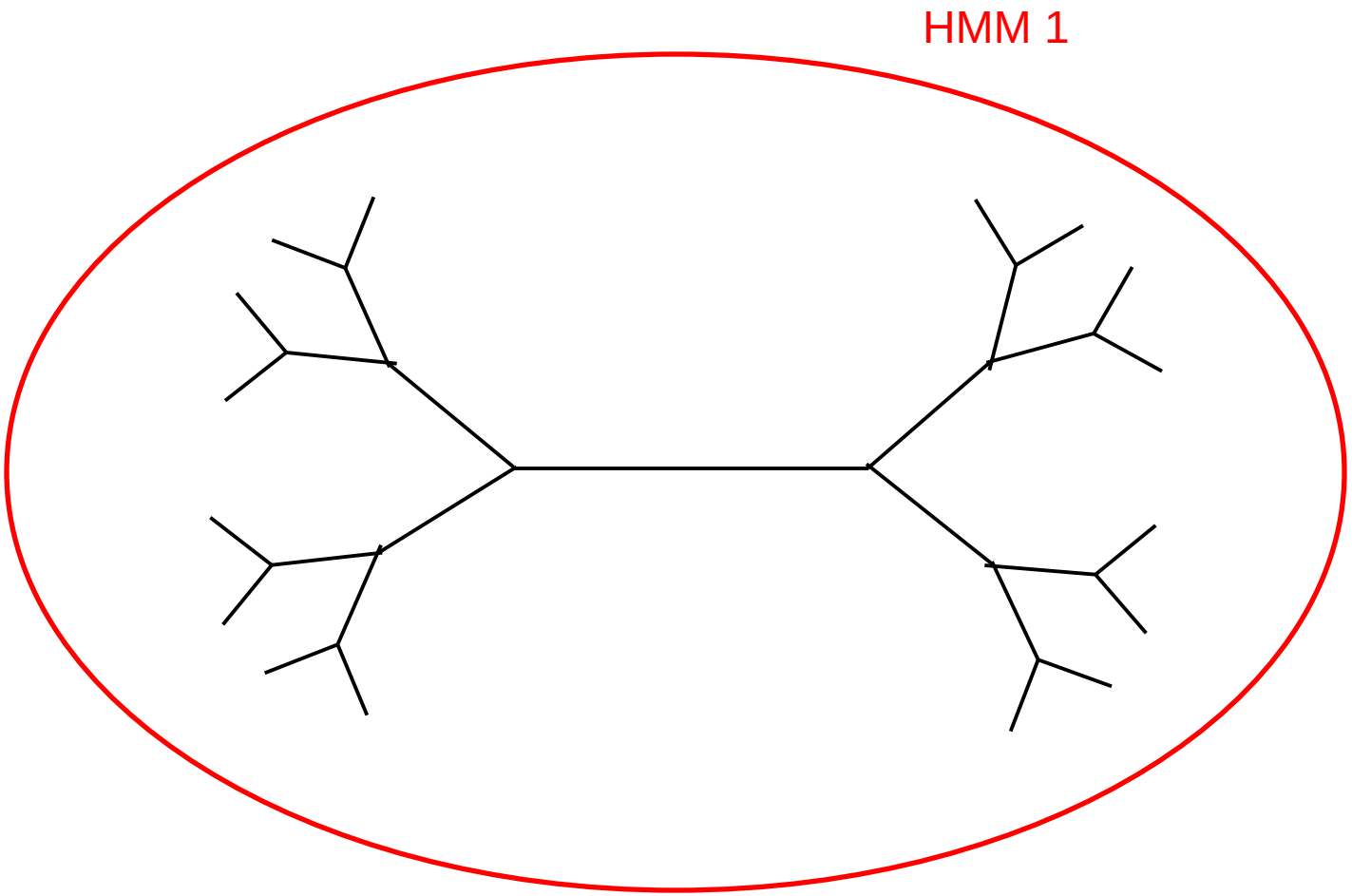
UPP: Ultra-large alignment using SEPP

- Developers: Nguyen, Mirarab, and Warnow
- Input: set of sequences S, backbone size B, and alignment subset size A
- Output: MSA on S
- Algorithm
  - Select B random full-length sequences (backbone set) from S
  - Estimate backbone alignment and backbone tree on backbone set
  - Align remaining sequences to backbone alignment
- Uses nested hierarchical fHMM
- In preparation

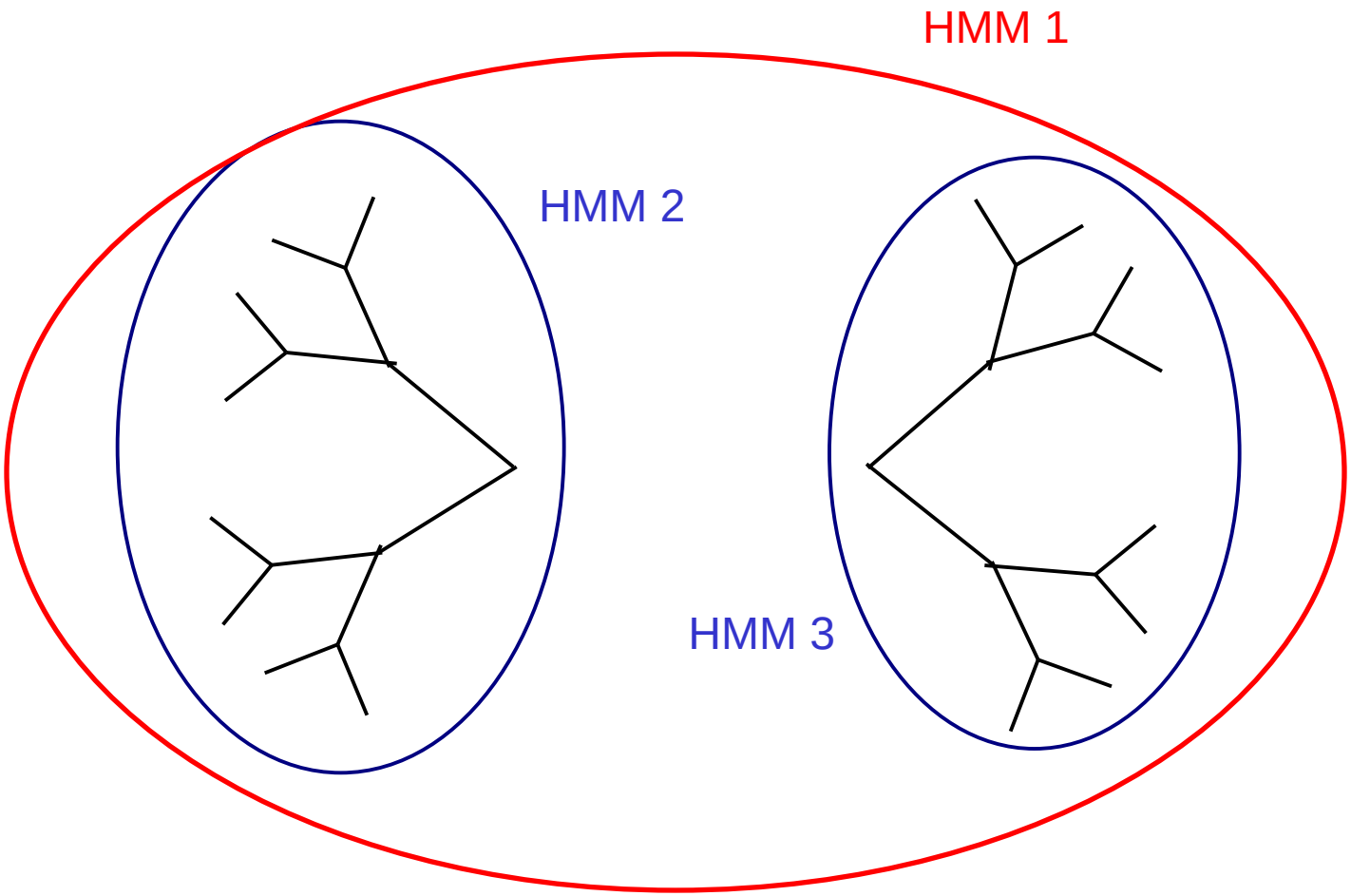
Disjoint HMMs



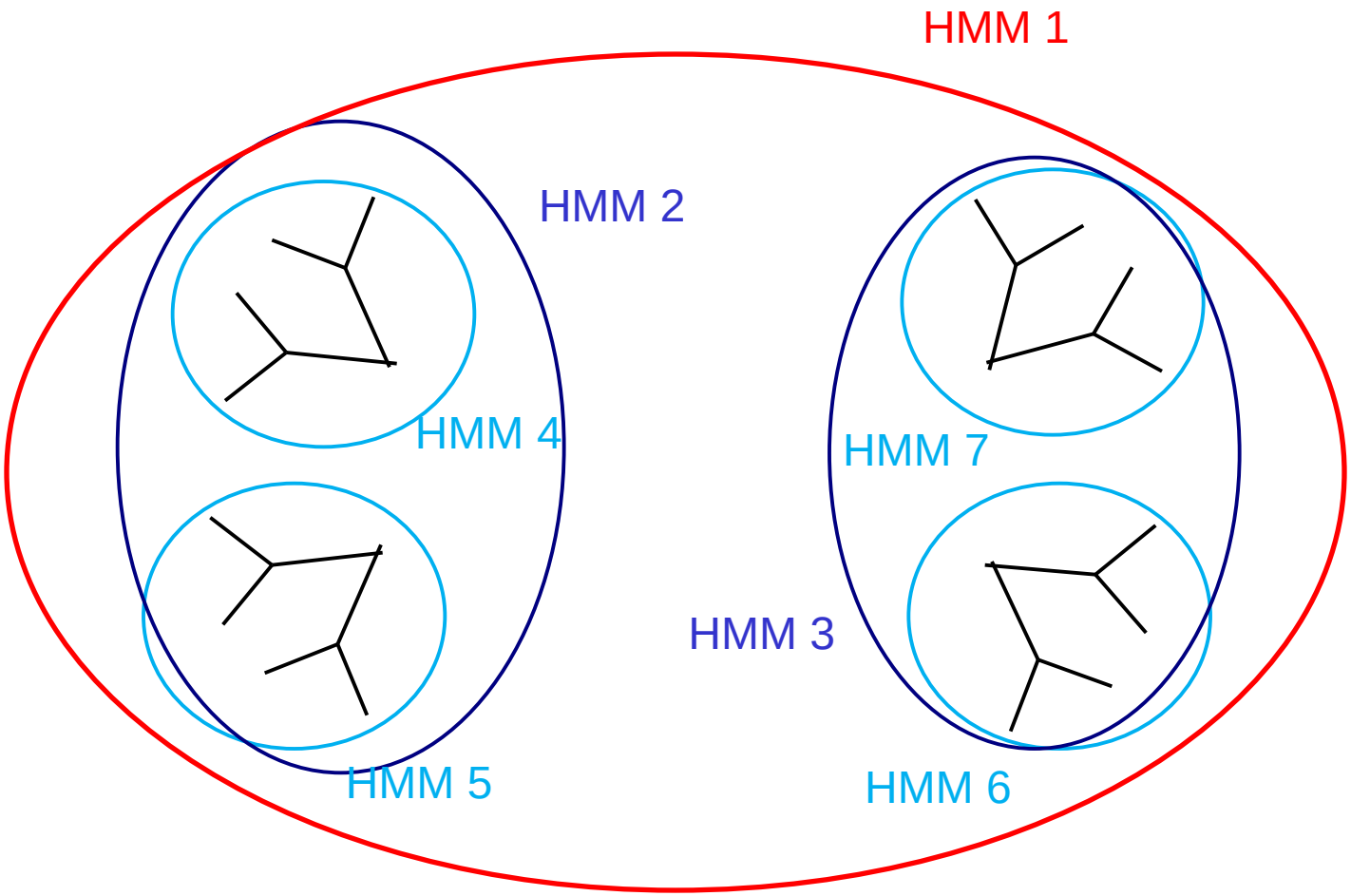
Nested HMMs



Nested HMMs



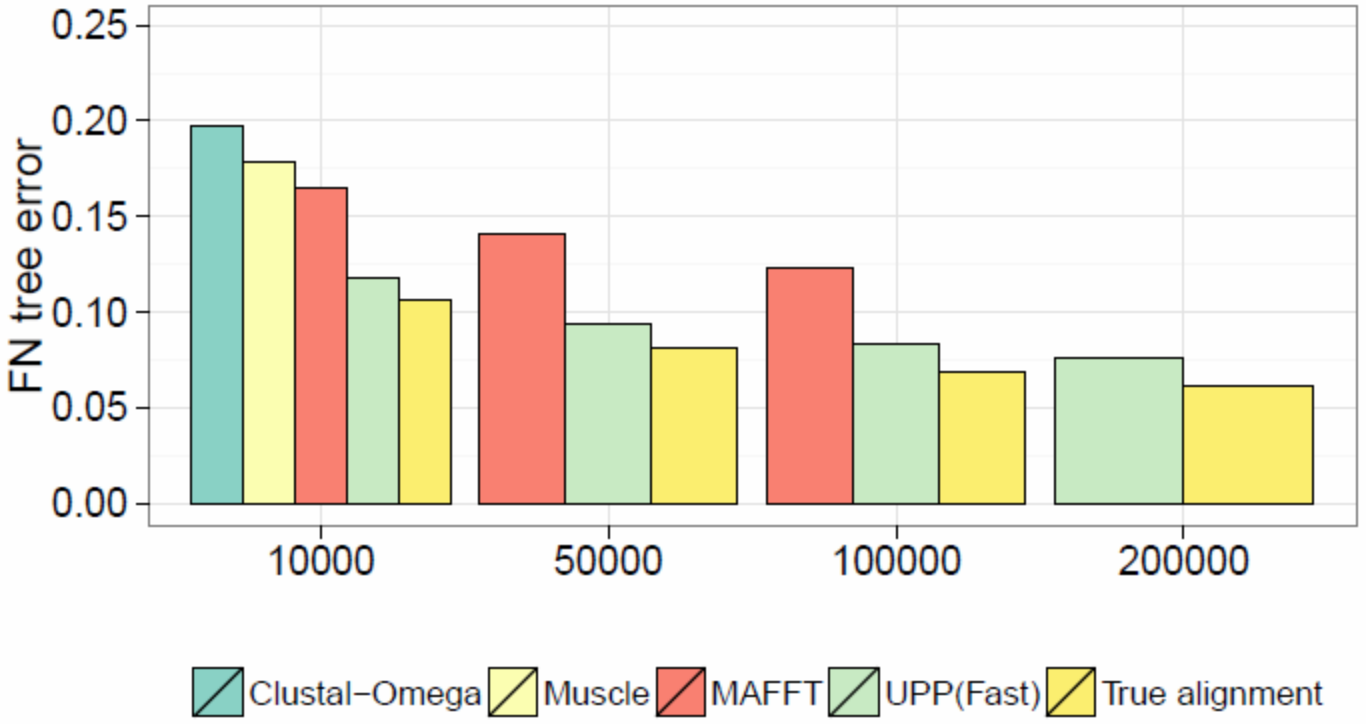
Nested HMMs



Experimental Design

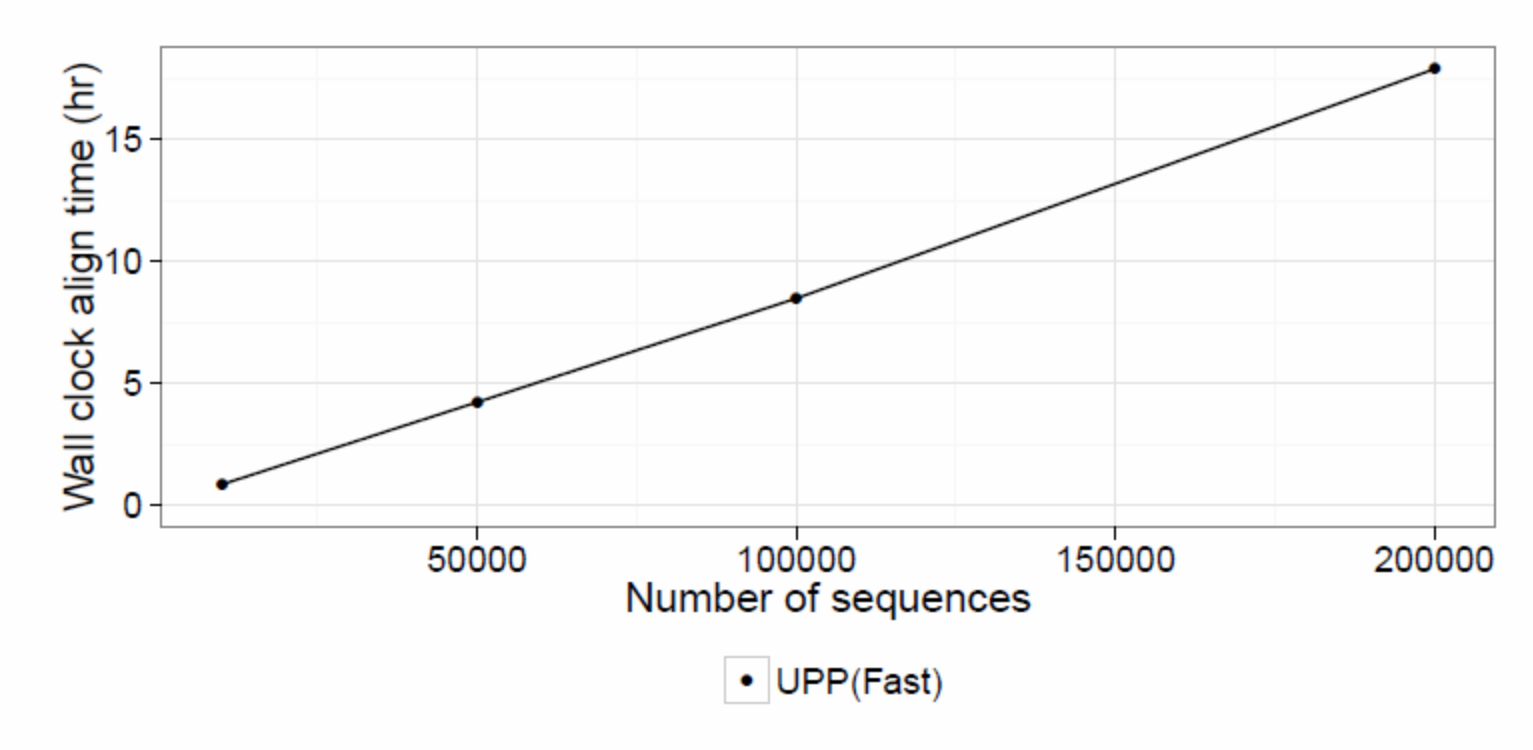
- Examined both simulated and biological DNA, RNA, and AA datasets
- Generated fragmentary datasets from the full-length datasets
- Compared Clustal-Omega, Mafft, Muscle, and UPP
- ML trees estimated on alignments using FastTree
- Scored alignment and tree error
  - Tree error measured in FN rate or Delta FN rate

Tree error on simulated RNA datasets



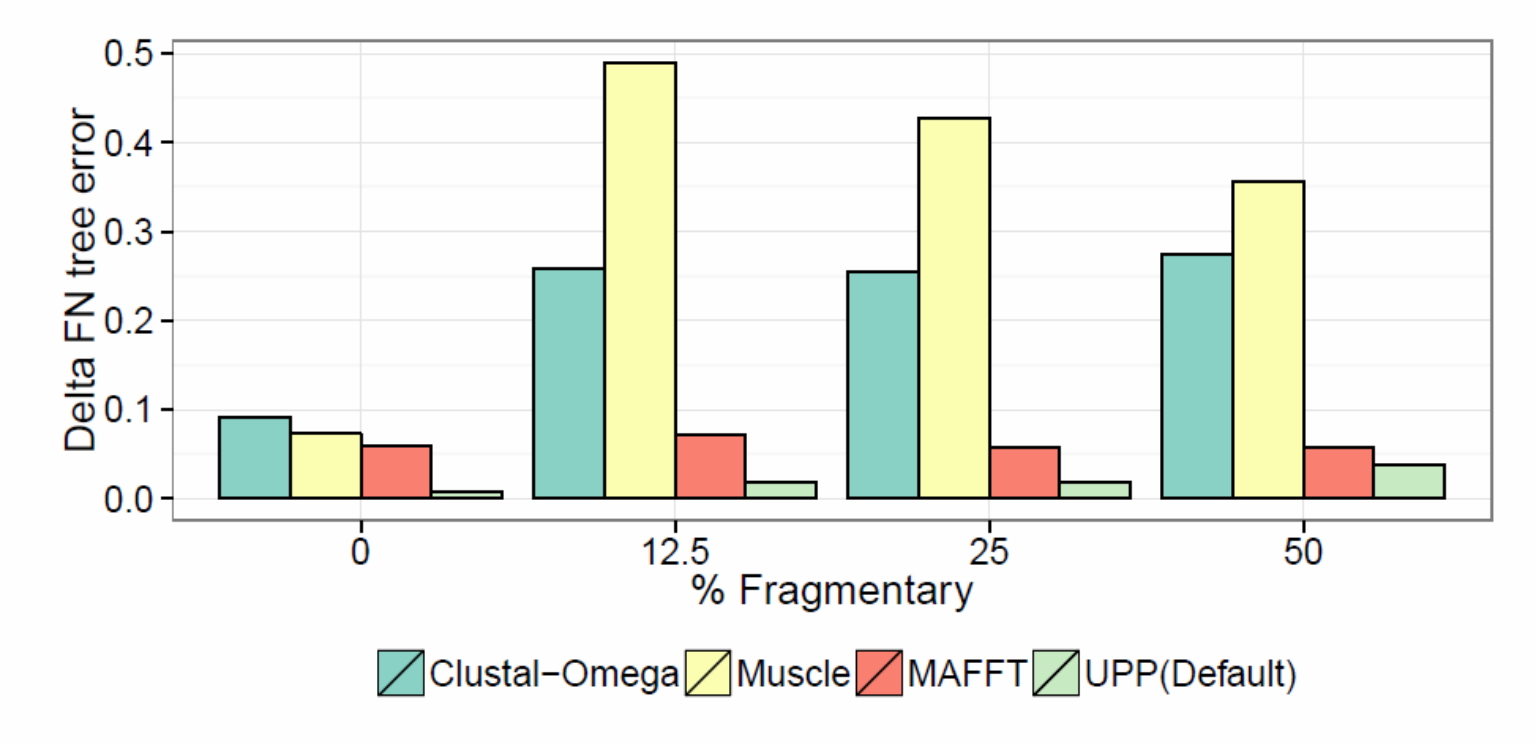
UPP(Fast): Backbone size=100, Alignment size=10
Average full-length sequence size 1500 bps
Only UPP completes on all datasets within 24 hours on a 12 core machine with 24 GB

Running time on simulated RNA datasets



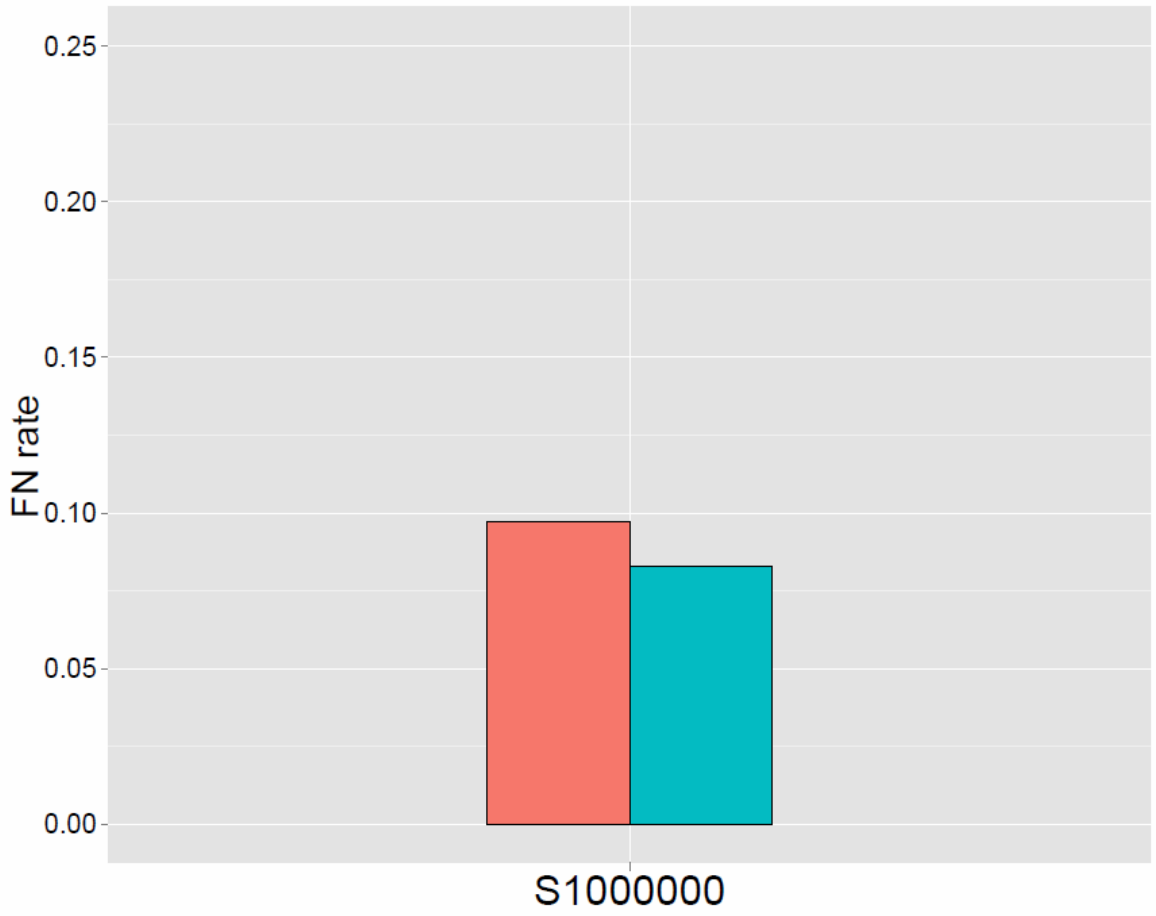
UPP has close to linear scaling

Tree Error on fragmentary RNASim 10K dataset



UPP(Default): Backbone size=1000, Alignment size=10
Average fragment length of 500 bps
Average full-length sequence size 1500 bps
Delta FN error: ML(Estimated)-ML(True)

One Million Sequences: Tree Error



UPP(100,100): 1.6 days using 12 processors

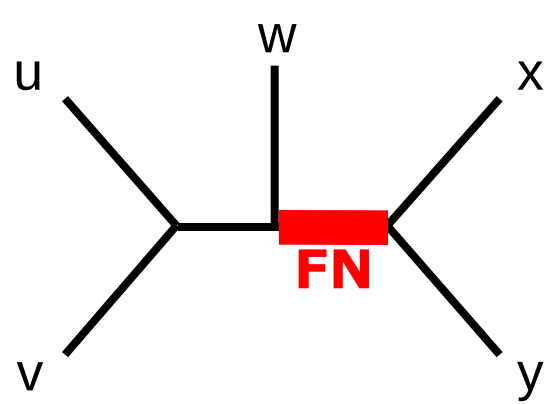
UPP(100,10): 7 days using 12 processors

Note: UPP Decomposition improves accuracy

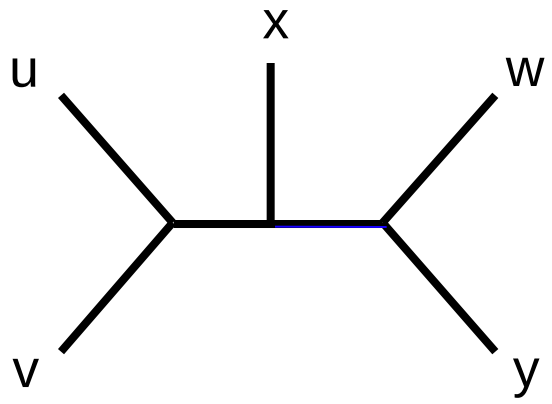
UPP summary

- Uses nested hierarchical fHMM for sequence alignment
- Overall, results in the most accurate alignments (not shown) and trees on full-length simulated datasets
    - Larger differences on highly divergent datasets
- Results in comparable or more accurate alignments and trees on biological datasets (not shown)
- Yields most accurate trees on both full-length and mixed datasets
- Only method that can complete within 24 hours on datasets with up to 200K sequences, 1M in less than 2 days

Tree Error



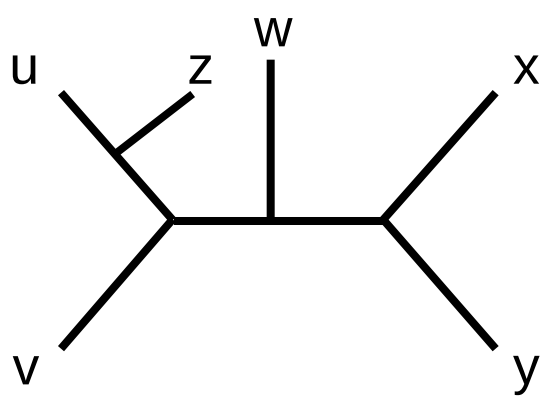
True Tree



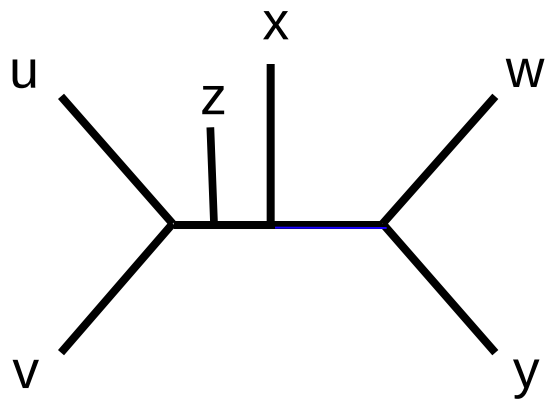
Estimated Tree

- **False Negative (FN)**: an edge in the true tree that is missing from the estimated tree
- **Delta Error**: the difference in FN of the backbone tree+placement and the backbone tree

Tree Error



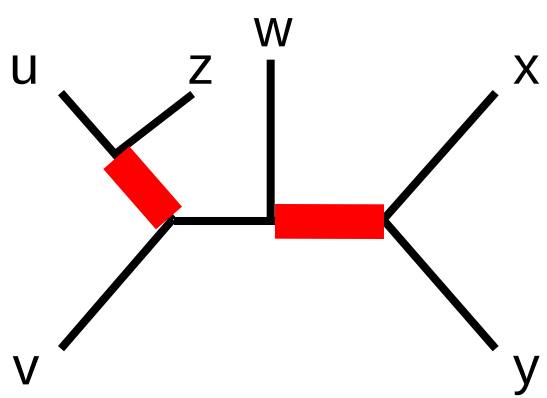
True Tree



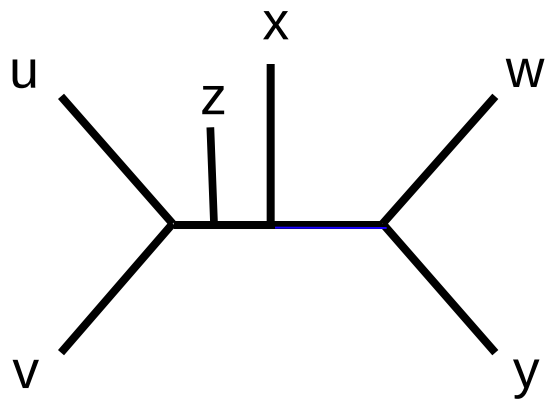
Estimated Tree

- **False Negative (FN)**: an edge in the true tree that is missing from the estimated tree
- **Delta Error**: the difference in FN of the backbone tree+placement and the backbone tree

Tree Error



True Tree

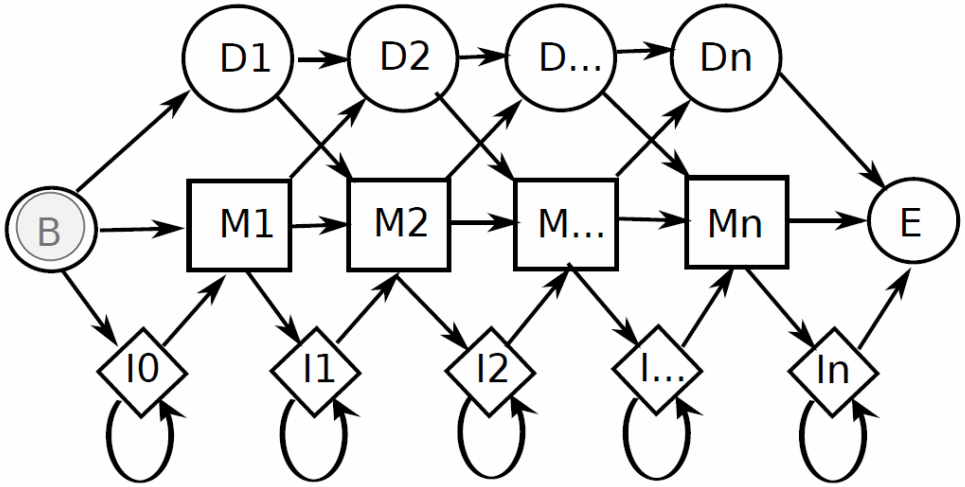


Estimated Tree

- **False Negative (FN)**: an edge in the true tree that is missing from the estimated tree
- **Delta Error**: the difference in FN of the backbone tree+placement and the backbone tree

Profile HMM

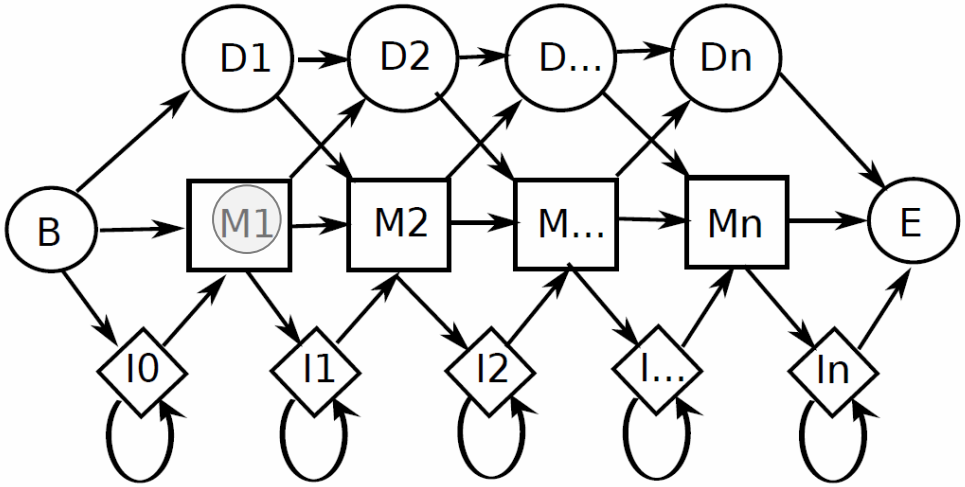
```
S1 A----ATC--TG---A
S2 A----GTT--TG---A
S3 AC--C-TT-A-AA-GA
S4 AC-AC-TCCA-GATGA
S5 TC-TCGT--T-CTTTA
...
Sn A-TTC-GC-A-GA--A
```



Q:

Profile HMM

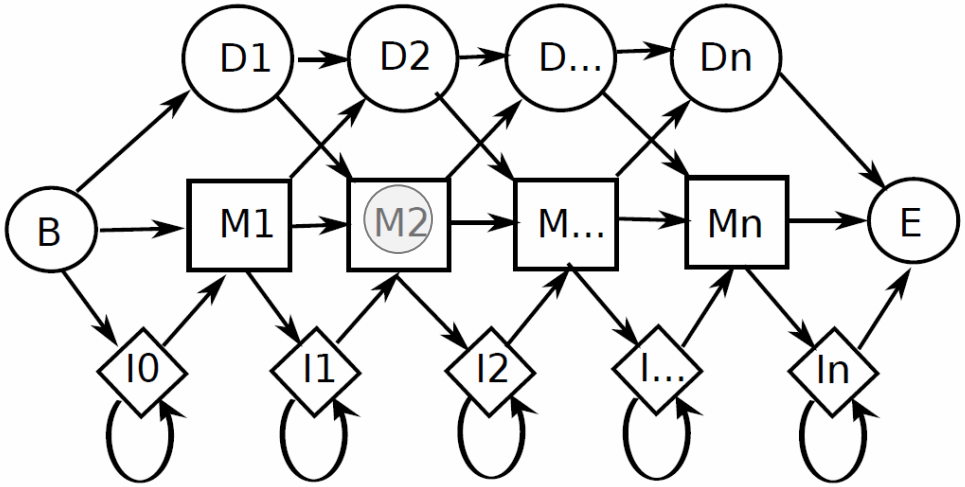
```
S1  A----ATC--TG---A
S2  A----GTT--TG---A
S3  AC--C-TT-A-AA-GA
S4  AC-AC-TCCA-GATGA
S5  TC-TCGT--T-CTTTA
...
Sn  A-TTC-GC-A-GA--A
```



Q:

Profile HMM

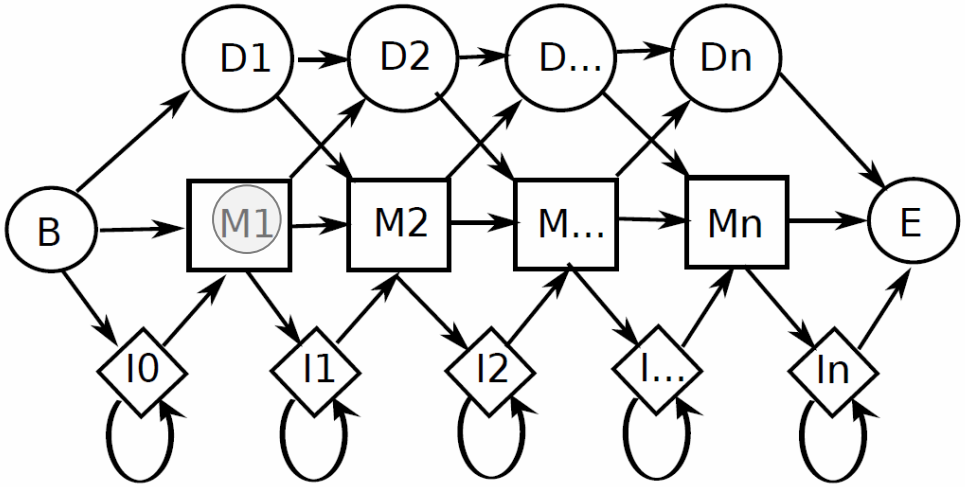
```
S1 A----ATC--TG---A
S2 A----GTT--TG---A
S3 AC--C-TT-A-AA-GA
S4 AC-AC-TCCA-GATGA
S5 TC-TCGT--T-CTTTA
...
Sn A-TTC-GC-A-GA--A
```



Q:**A**

Profile HMM

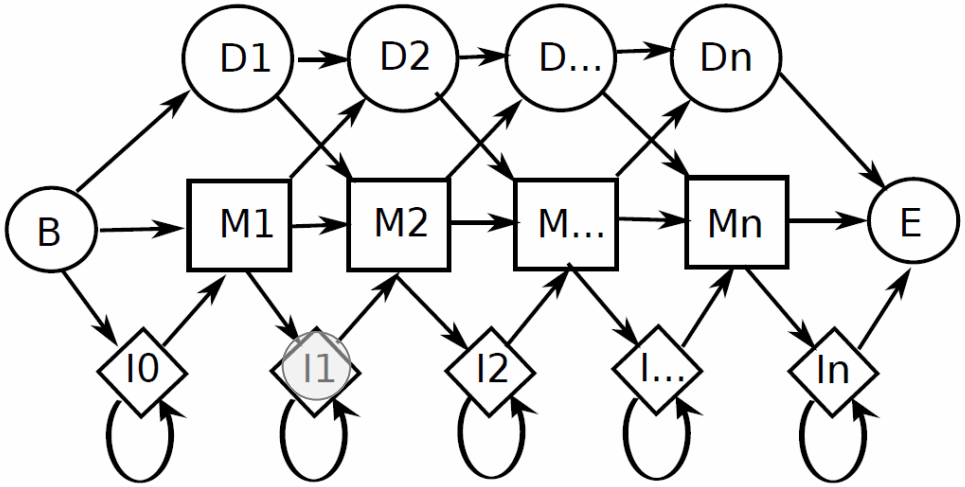
```
S1  A----ATC--TG---A
S2  A----GTT--TG---A
S3  AC--C-TT-A-AA-GA
S4  AC-AC-TCCA-GATGA
S5  TC-TCGT--T-CTTTA
...
Sn  A-TTC-GC-A-GA--A
```



Q:

Profile HMM

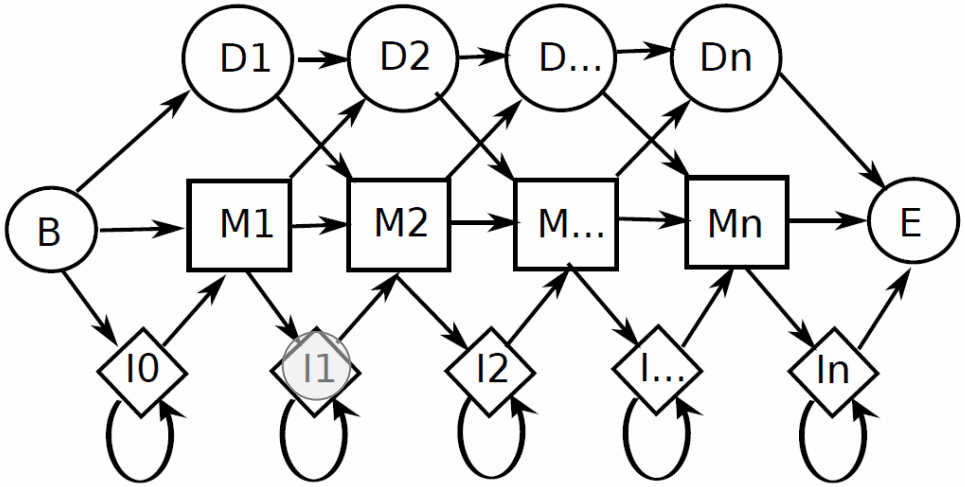
```
S1  A----ATC--TG---A
S2  A----GTT--TG---A
S3  AC--C-TT-A-AA-GA
S4  AC-AC-TCCA-GATGA
S5  TC-TCGT--T-CTTTA
...
Sn  A-TTC-GC-A-GA--A
```



Q:A

Profile HMM

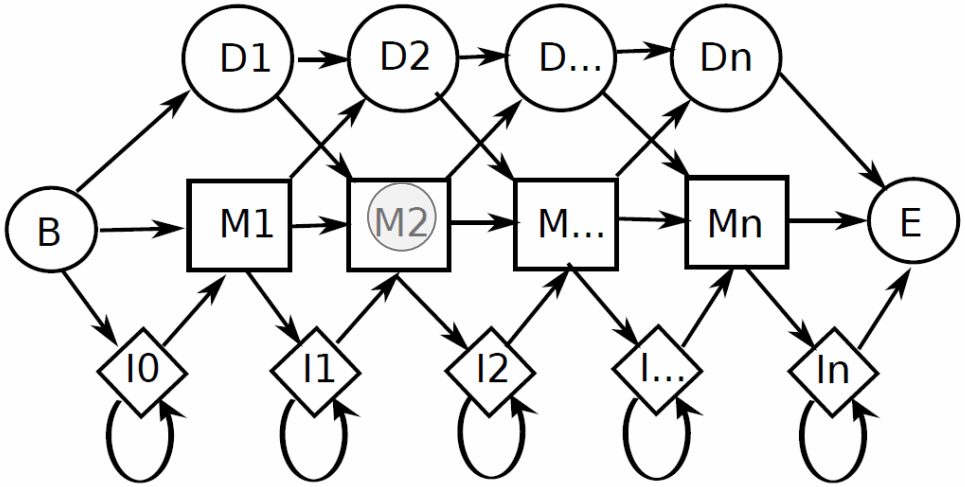
```
S1  A----ATC--TG---A
S2  A----GTT--TG---A
S3  AC--C-TT-A-AA-GA
S4  AC-AC-TCCA-GATGA
S5  TC-TCGT--T-CTTTA
...
Sn  A-TTC-GC-A-GA--A
```



Q:At

Profile HMM

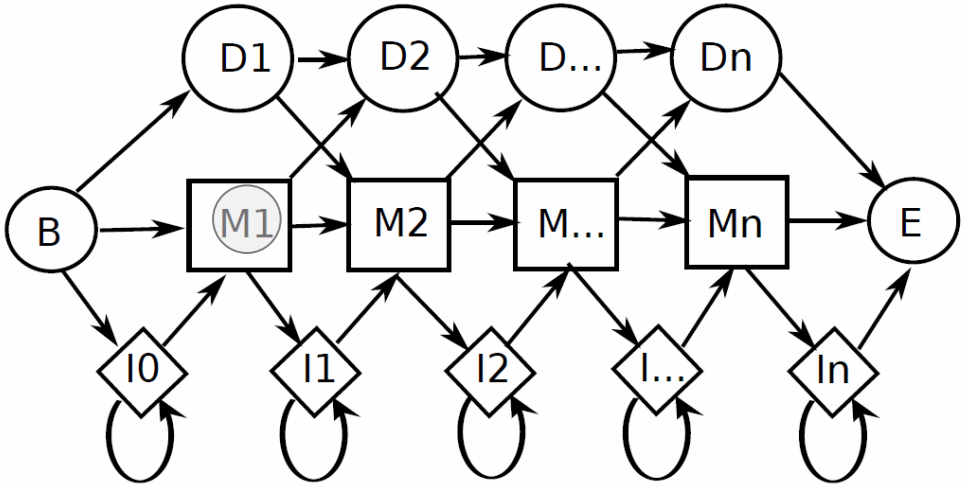
```
S1  A----ATC--TG---A
S2  A----GTT--TG---A
S3  AC--C-TT-A-AA-GA
S4  AC-AC-TCCA-GATGA
S5  TC-TCGT--T-CTTTA
...
Sn  A-TTC-GC-A-GA--A
```



Q:Atc

Profile HMM

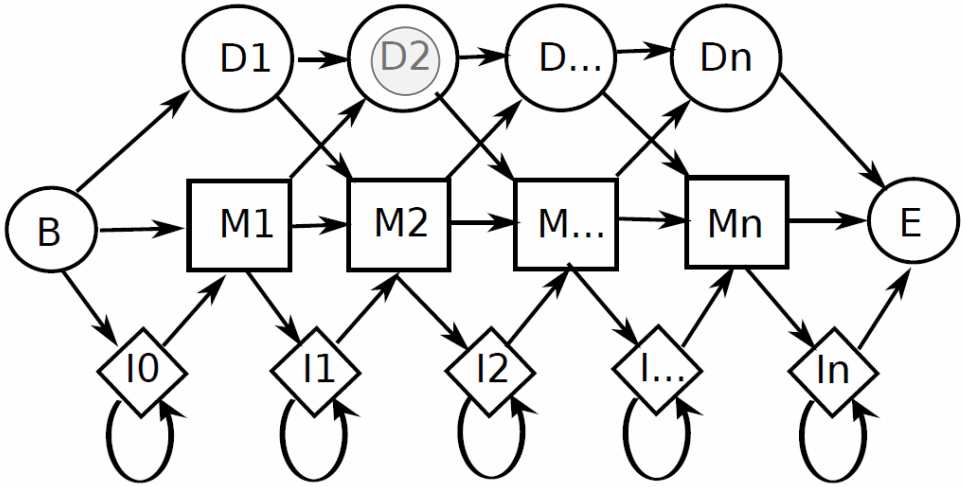
```
S1  A----ATC--TG---A
S2  A----GTT--TG---A
S3  AC--C-TT-A-AA-GA
S4  AC-AC-TCCA-GATGA
S5  TC-TCGT--T-CTTTA
...
Sn  A-TTC-GC-A-GA--A
```



Q:

Profile HMM

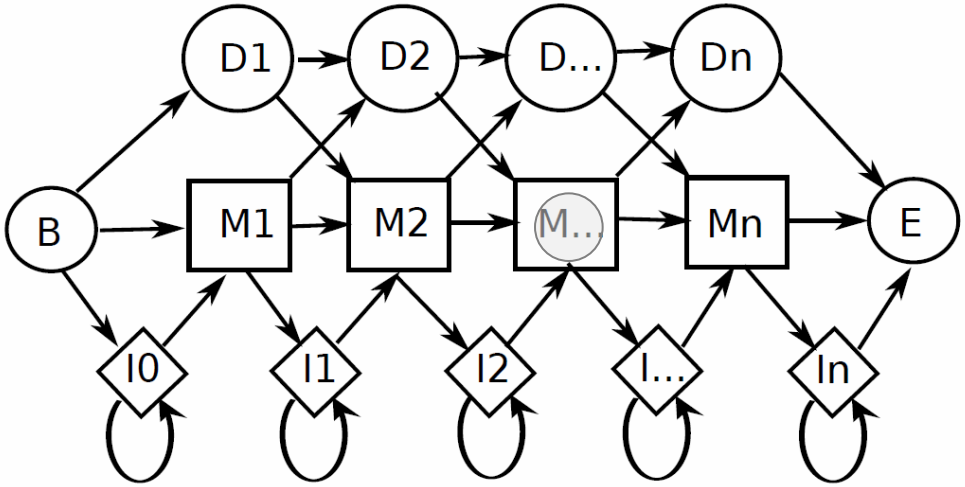
```
S1 A----ATC--TG---A
S2 A----GTT--TG---A
S3 AC--C-TT-A-AA-GA
S4 AC-AC-TCCA-GATGA
S5 TC-TCGT--T-CTTTA
...
Sn A-TTC-GC-A-GA--A
```



Q:A

Profile HMM

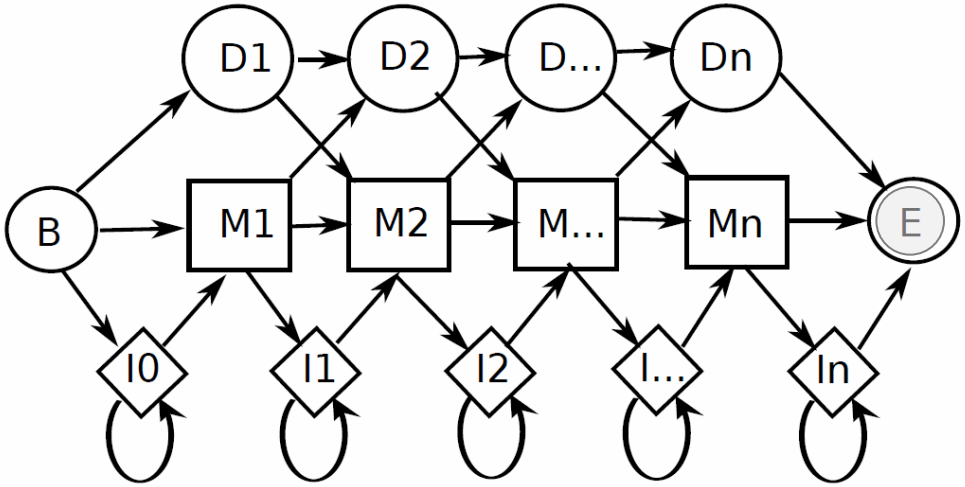
```
S1 A-----ATC--TG---A
S2 A-----GTT--TG---A
S3 AC---C-TT-A-AA-GA
S4 AC-AC-TCCA-GATGA
S5 TC-TCGT--T-CTTTA
...
Sn A-TTC-GC-A-GA--A
```



Q:A-

Profile HMM

```
S1 A----ATC--TG---A
S2 A----GTT--TG---A
S3 AC--C-TT-A-AA-GA
S4 AC-AC-TCCA-GATGA
S5 TC-TCGT--T-CTTTA
...
Sn A-TTC-GC-A-GA--A
```



Q:A-tcTCA-tATG

Metagenomic data analysis

NGS data produce fragmentary sequence data

Metagenomic analyses include unknown species

Taxon identification: given short sequences, identify the species for each fragment

Applications: Human Microbiome and other metagenomic projects

Issues: accuracy and speed

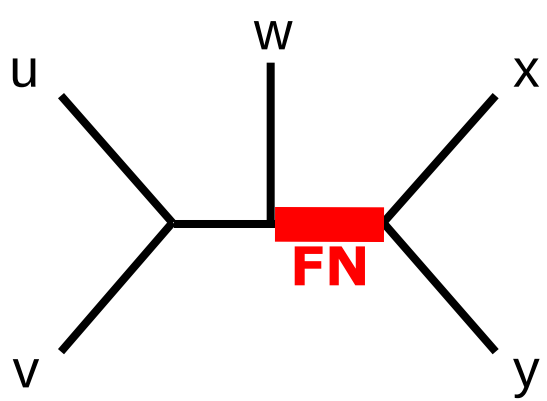
Contributions

- MRL and SuperFine+MRL: new supertree methods. Nguyen, Mirarab, and Warnow. AMB 2012
- SEPP: SATé-Enabled phylogenetic placement. Mirarab, Nguyen, and Warnow. PSB 2012
- TIPP: Taxonomic identification and phylogenetic profiling. Nguyen, Mirarab, Pop, and Warnow. Under review.
- UPP: Ultra-large alignment using SEPP. Nguyen, Mirarab, Kumar, Guo, Wang, and Warnow. In preparation.
- Comparison of different methods for masking alignments. Nguyen, Linder, and Warnow. In preparation.

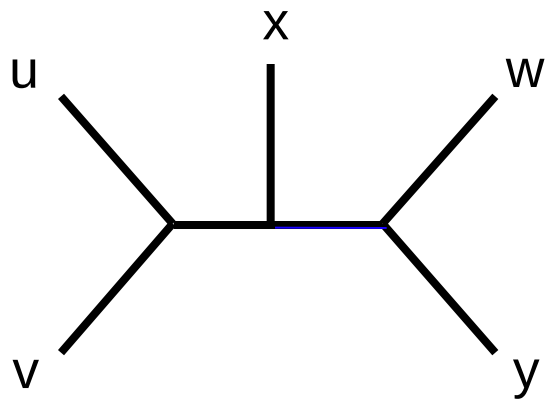
Contributions

- MRL and SuperFine+MRL: new supertree methods. Nguyen, Mirarab, and Warnow. AMB 2012
- SEPP: SATé-Enabled phylogenetic placement. Mirarab, Nguyen, and Warnow. PSB 2012
- TIPP: Taxonomic identification and phylogenetic profiling. Nguyen, Mirarab, Pop, and Warnow. Under review.
- UPP: Ultra-large alignment using SEPP. Nguyen, Mirarab, Kumar, Guo, Wang, and Warnow. In preparation.
- Comparison of different methods for masking alignments. Nguyen, Linder, and Warnow. In preparation.

Tree Error



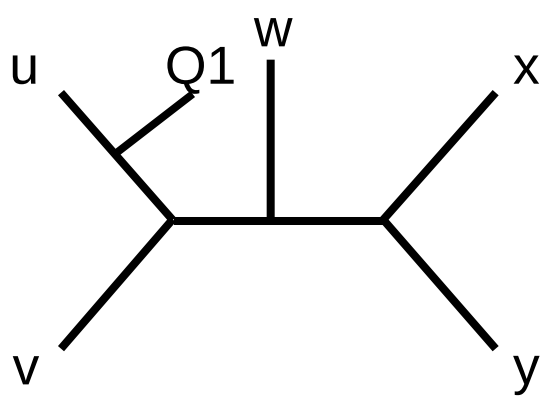
True Tree



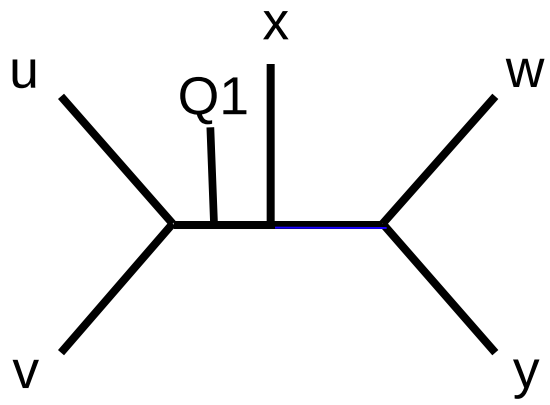
Estimated Tree

- **False Negative (FN)**: an edge in the true tree that is missing from the estimated tree

Placement Error



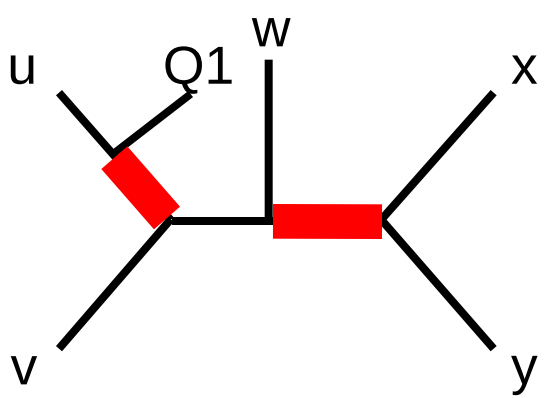
True Tree



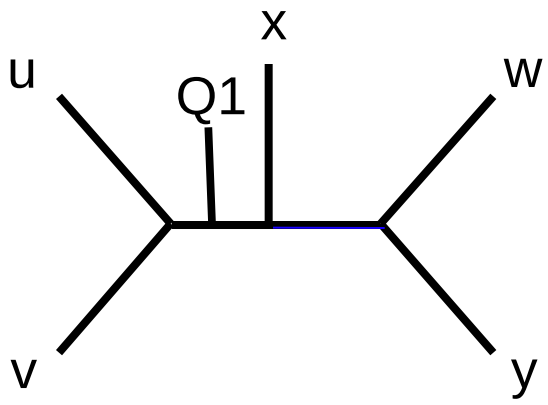
Estimated Tree

- **Delta Error**: the difference in FN of the extended tree and the backbone tree

Placement Error



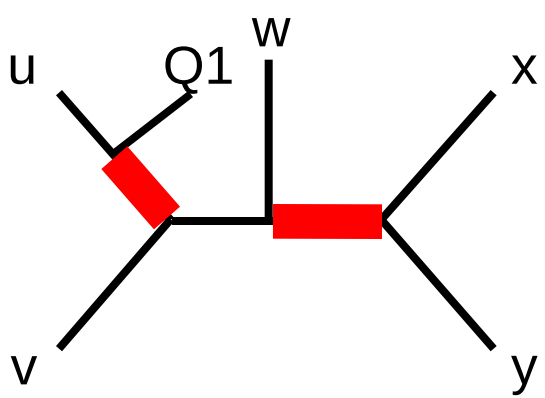
True Tree



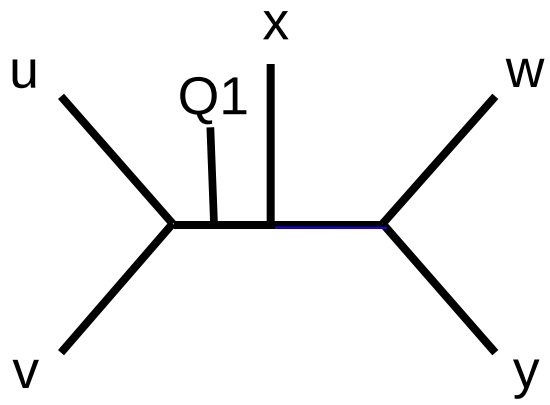
Estimated Tree

- **Delta Error**: the difference in FN of the extended tree and the backbone tree

Placement Error



True Tree



Estimated Tree

- **Delta Error**: the difference in FN of the extended tree and the backbone tree

Delta Error = 2 – 1 = 1

Alignment using fHMM