Novel approaches for large-scale multiple sequence alignment and phylogenetic estimation

Tandy Warnow
Department of Computer Science
The University of Texas at Austin
Input: unaligned sequences

S1 = AGGCTATCACCTGACCTCCCA
S2 = TAGCTATCAGCAGCCG
S3 = TAGCTGACCGC
S4 = TCACGACCGACA
Phase 1: Alignment

\begin{align*}
S1 &= \text{AGGCTATCACCTGACCTCCA} & S1 &= \text{--AGGCTATCACCTGACCTCCA} \\
S2 &= \text{TAGCTATCAGACCGGC} & S2 &= \text{TAG-CTATCAC--GACCGG--} \\
S3 &= \text{TAGCTGACCGGC} & S3 &= \text{TAG-CT---------GACCGG--} \\
S4 &= \text{TCACGACCGACA} & S4 &= \text{---------TCAC--GACCGACAA}
\end{align*}
Phase 2: Construct tree

\[ S1 = \text{AGGCTATCACCTGACCTCCA} \]
\[ S2 = \text{TAGCTATCACGACCGC} \]
\[ S3 = \text{TAGCTGACCGC} \]
\[ S4 = \text{TCACGACCGACA} \]
Simulation Studies

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

Unaligned Sequences

S1 = -AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC-
S3 = TAGCTGACCGC-
S4 = T---C-A-CGACCGA----CA

True tree and alignment

Compare

Estimated tree and alignment
1000 taxon models, ordered by difficulty (Liu et al., 2009)
Major Challenges: large datasets, fragmentary sequences

- **Phylogenetic analyses**: standard methods have *poor accuracy* on even moderately large datasets, and the most accurate methods are enormously *computationally intensive* (weeks or months, high memory requirements).

- **Multiple sequence alignment**: Few methods can run on large datasets, and alignment accuracy is generally poor for large datasets with high rates of evolution.

Both phylogenetic estimation and multiple sequence alignment are also impacted by *fragmentary data.*
This Talk

- **SATé** - co-estimating trees and alignments
- **DACTAL** - trees almost without alignments
- **SEPP** - phylogenetic placement of fragmentary sequence data (e.g., short reads)
Part I: SATé

Simultaneous Alignment and Tree Estimation

Liu et al., Systematic Biology 2012

Public software distribution (open source) through Mark Holder’s group at the University of Kansas
SATé Algorithm

Obtain initial alignment and estimated ML tree
SATé Algorithm

Obtain initial alignment and estimated ML tree

Tree

Use tree to compute new alignment

Alignment
SATé Algorithm

Obtain initial alignment and estimated ML tree

Estimate ML tree on new alignment

Use tree to compute new alignment
Re-aligning on a tree

1. Decompose dataset
2. Align subproblems
3. Merge sub-alignments
4. Estimate ML tree on merged alignment

Diagram:
- Initial tree with nodes A, B, C, D
- Decompose dataset into subproblems
- Align subproblems
- Merge sub-alignments
- Final aligned tree ABCD
1000 taxon models, ordered by difficulty

24 hour SATé analysis, on desktop machines
(Similar improvements for biological datasets)
1000 taxon models ranked by difficulty
Limitations

- Estimate ML tree on merged alignment
- Merge sub-alignments
- Align subproblems
- Decompose dataset

ABCD
Limitations

Decompose dataset

Align subproblems

Estimate ML tree on merged alignment

Merge sub-alignments
Part II: DACTAL
Divide-And-Conquer Trees without Alignments*

- Input: set S of unaligned sequences
- Output: tree on S (but no alignment)

Nelesen, Liu, Wang, Linder, and Warnow, ISMB 2012 and Bioinformatics 2012

*(almost)
DACTAL

Existing Method: RAxML(MAFFT)

New supertree method: SuperFine

Unaligned Sequences

pRecDCM3

BLAST-based

Overlapping subsets

A tree for the entire dataset

A tree for each subset
Average of 3 Largest CRW Datasets

CRW: Comparative RNA database,

Three 16S datasets with 6,323 to 27,643 sequences

Reference alignments based on secondary structure

Reference trees are 75% RAxML bootstrap trees

DACTAL (shown in red) run for 5 iterations starting from FT(Part)

FastTree (FT) and RAxML are ML methods
DACTAL and SATé

- DACTAL and SATé estimate large trees from full-length sequences for one or several genes

- DACTAL can be used with other types of data (not just sequences)

- But neither handles fragmentary data (e.g., short reads)
Phylogenetic Placement

Fragmentary Sequences from some gene

Full-length sequences for same gene, and an alignment and a tree

ACCG
CGAG
CGG
GGCT
TAGA
GGGGG
TCGAG
GGCG
GGG
ACCT
AGG...GCAT
TAGC...CCA
TAGA...CTT
AGC...ACA
ACT..TAGA..A
Part III: SEPP

• SEPP: SATé-enabled Phylogenetic Placement, by Mirarab, Nguyen, and Warnow

• Pacific Symposium on Biocomputing, 2012 (special session on the Human Microbiome)
Phylogenetic Placement

Step 1: Align each query sequence to backbone alignment

Step 2: Place each query sequence into 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-CTATCACA--GACCGC--GCA
S3  = TAG-CT-------GACCGC--GCT
S4  = TAC----TCAC--GACCGACAGCT
Q1  = --------T-A--AAAC--------
Place Sequence

S1 = -AGGCTATCACCTGACCTCCA-AA
S2 = TAG-CTATCAGACCGC--GCA
S3 = TAG-CT---------GACCGC--GCT
S4 = TAC----TCAC--GACCGACAGCT
Q1 = --------T-A--AAAC--------
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
  - Pplacer (Matsen et al., BMC Bioinformatics, 2011)
  - EPA (Berger and Stamatakis, Systematic Biology 2011)

Note: pplacer and EPA use maximum likelihood, and are reported to have the same accuracy.
HMMER vs. PaPaRa

Increasing rate of evolution

Model Condition

Delta-error (edges)

HMMER+pplacer
PaPaRa+pplacer

0.0
HMMER+pplacer:
1) build one HMM for the entire alignment
2) Align fragment to the HMM, and insert into alignment
3) Insert fragment into tree to optimize likelihood
One Hidden Markov Model for the entire alignment?
Or 2 HMMs?
Or 4 HMMs?
SEPP(10%), based on ~10 HMMs

Increasing rate of evolution
SEPP (10%) on Biological Data

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

For 1 million fragments:

PaPaRa+pplacer: ~133 days
HMMALIGN+pplacer: ~30 days
SEPP 1000/1000: ~6 days
Three “Boosters”

- **SATé**: co-estimation of alignments and trees
- **DACTAL**: large trees without full alignments
- **SEPP**: phylogenetic placement of short reads

Algorithmic strategies: divide-and-conquer and iteration to improve the accuracy and scalability of a base method
Applications of SEPP

• **UPP**: Ultra-large alignment using SEPP

• **TIPP**: taxon identification of fragmentary data
UPP: Ultra-large alignments using SEPP

Compared to Clustal-Omega and MAFFT on simulated datasets with 10,000 to 200,000 sequences
TIPP: Taxon Identification using SEPP (highly robust to sequencing error)
Using these methods

• SATé is being used in several large-scale projects (e.g., Avian and 1KP)
• SEPP, SuperFine, and DACTAL are available as command line
• UPP and TIPP are under development

We would be very happy to discuss potential collaborations!

Contact me by email, tandy@cs.utexas.edu
Acknowledgments

• Funding: Guggenheim Foundation Fellowship, NSF: ATOL, ITR, and IGERT grants, and David Bruton Jr. Professorship

• Collaborators:
  – SATé: Kevin Liu, Serita Nelesen, Sindhu Raghavan, and Randy Linder (and Mark Holder’s lab at Kansas for public distribution)
  – DACTAL: Serita Nelesen, Kevin Liu, Li-San Wang, and Randy Linder
  – SEPP and UPP: Siavash Mirarab and Nam Nguyen
  – TIPP: Nam Nguyen, Siavash Mirarab, Mihai Pop, and Bo Liu

• See http://www.cs.utexas.edu/users/ATOL-MSA.html for publications and downloadable software