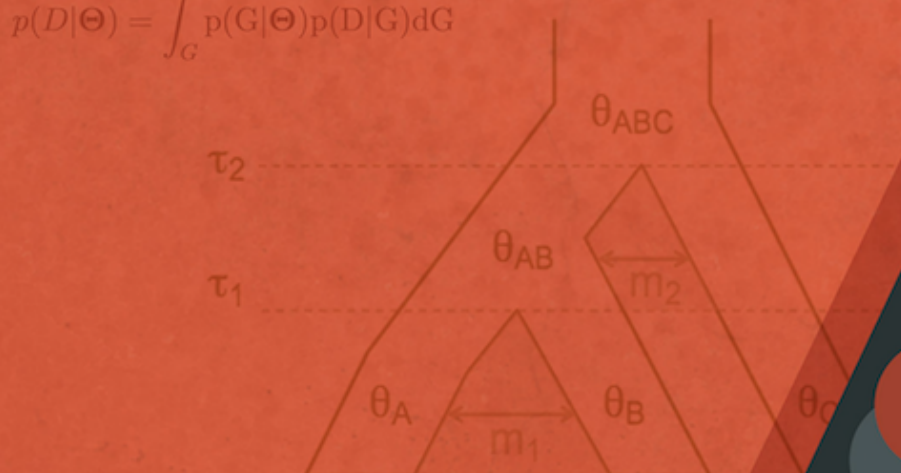
Species Tree Estimation: Theoretical Challenges of Today and Tomorrow

Laura Kubatko
Departments of Statistics and Evolution, Ecology, and Organismal Biology, OSU

Twitter: Laura_Kubatko



Theoretical challenges of today and tomorrow

- **Question:** From the perspective of a developer of methods/models, what are the challenges we currently face, and what do we expect the future challenges to be?
  - **Computational and modeling challenges**
  - **Challenges in maximizing information gained**
  - **User adoption and training**
  - **Challenge of training next generation of developers**

Computational and modeling challenges

- **Issue 1: "Local" vs. "Global" models**

- **Example:** For a single gene, we might use model selection to estimate the most appropriate evolutionary model
  - Estimate the phylogeny accurately
  - Estimate parameters, variances, learn about evolutionary process in that gene

- **Scaling up:** What if we have 100 genes? 1,000 genes?
  - Model selection for each gene would involve lots of tests
  - Fitting a general model to all genes involves lots of parameter estimates

- **Question:** What about machine learning approaches?

Computational and modeling challenges

- **Example:** Kubatko et al. (2011)

Locus[a]	Aligned length	PI sitesb (ingroup only)	Substitution model	Average divergence
A	296	31	K80 + *I*	0.02527
1	220	12	K80	0.01297
4	267	5	K80 +	0.00726
11	420	14	K80 + *I*	0.01199
25	262	14	TVMef + *I*	0.06160
31	256	8	F81	0.01071
41	274	7	HKY	0.00625
51	260	10	K80	0.02073
61	194	3	HKY	0.00819
63	471	8	HKY + *I*	0.01019
TBP	796	26	HKY + *I*	0.01444
CBA	525	9	HKY + *I*	0.08465
OD	522	16	K81uf + *I*	0.01337
ETS	840	20	HKY + *I*	0.01107

Computational and modeling challenges

- **Issue 2: "Dig deep" to understand model/method performance**

- **Example:** Statistical efficiency of invariants-based methods as we move to genome-scale data

  - Invariants approaches were widely viewed to lack statistical efficiency **for gene tree estimation**

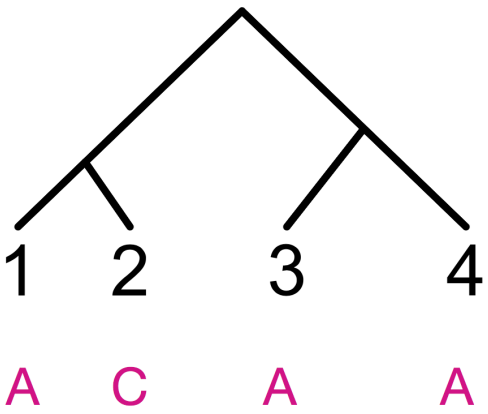
  - But, species tree estimation is fundamentally different

  - *"Invariants are worth attention, not for what they can do for us now, but what they might lead to in the future."*
    – Joseph Felsenstein (2006)

  - Current invariants-based methods: ABBA-BABA, HyDe, and (in some sense) SVDQuartets

What is a phylogenetic invariant?

- **Site pattern probability:**



Define p_{ACAA} to be the probability that

Taxon 1 has nucleotide A,

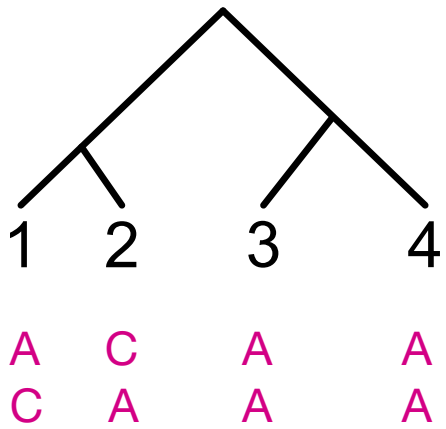
Taxon 2 has nucleotide C,

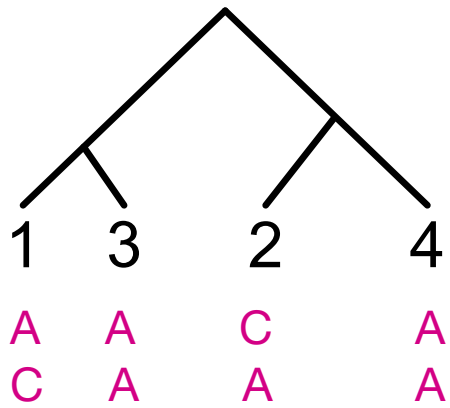
Taxon 3 has nucleotide A,

Taxon 4 has nucleotide A

What is a phylogenetic invariant?

- **Consider two four taxon trees and assume the Jukes-Cantor (JC69) model**



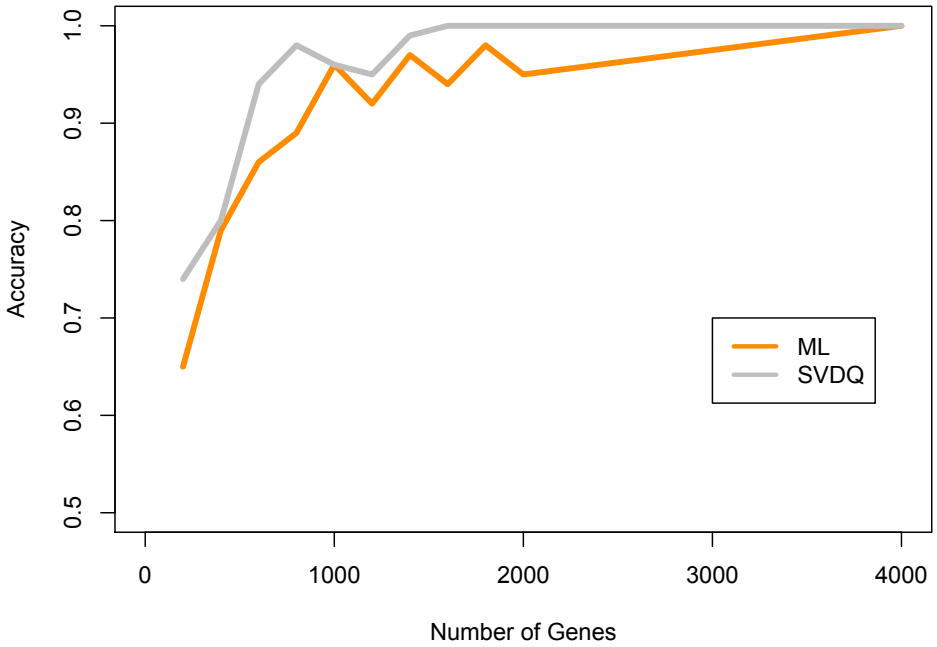


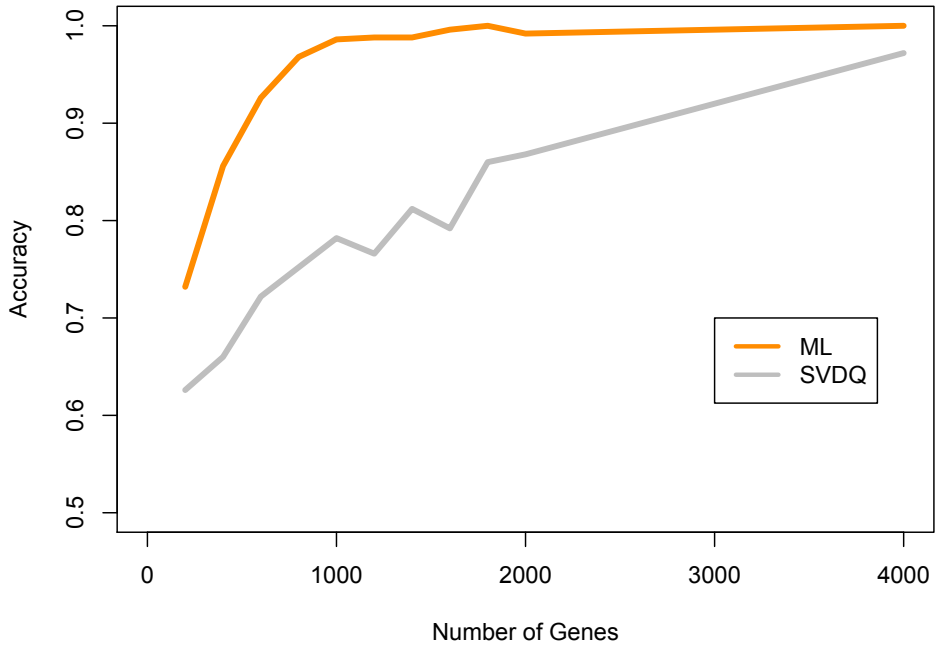
- **Topology invariant:**

$$p_{ACAA} - p_{CAAA}$$

Computational and modeling challenges

- **Relative statistical efficiency of maximum likelihood vs. SVDQuartets for 4 taxa**





Maximizing information

- **How do we select data to sequence/include in an analysis?**
- **Example 1:** What is the effect of missing data in an analysis?

Impact of Missing Data on Phylogenies Inferred from Empirical Phylogenomic Data Sets

Béatrice Roure,[1] Denis Baurain,[‡,2] and Hervé Philippe*,[1]

[1]Département de Biochimie, Centre Robert-Cedergren, Université de Montréal, Montréal, Québec, Canada

[2]Unit of Animal Genomics, GIGA-R and Faculty of Veterinary Medicine, Université de Liège, Liège, Belgium

[‡]Present address: Eukaryotic Phylogenomics, Department of Life Sciences, Université de Liège, Liège, Belgium

***Corresponding author:** E-mail: herve.philippe@umontreal.ca.

Associate editor: Barbara Holland

sequence evolution. First, we note that parsimony-uninformative incomplete characters are actually informative in a probabilistic framework. A reanalysis of Lemmon's data set with this in mind gives a very different interpretation of their results and shows that some of their conclusions may be unfounded. Second, we investigate the effect of the progressive introduction of missing data in a complete supermatrix (126 genes × 39 species) capable of resolving animal relationships. These analyses demonstrate that missing data perturb phylogenetic inference slightly beyond the expected decrease in resolving power. In particular, they

MBE 30(1):197–214, 2012

Maximizing information

- **How do we select data to sequence/include in an analysis?**
- **Example 1:** What is the effect of missing data in an analysis?

Impact of Missing Data on Phylogenies Inferred from Empirical Phylogenomic Data Sets

Béatrice Roure,[1] Denis Baurain,[‡,2] and Hervé Philippe[*,1]

[1]Département de Biochimie, Centre Robert-Cedergren, Université de Montréal, Montréal, Québec, Canada

[2]Unit of Animal Genomics, GIGA-R and Faculty of Veterinary Medicine, Université de Liège, Liège, Belgium

[‡]Present address: Eukaryotic Phylogenomics, Department of Life Sciences, Université de Liège, Liège, Belgium

***Corresponding author:** E-mail: herve.philippe@umontreal.ca.

Associate editor: Barbara Holland

sequence evolution. First, we note that parsimony-uninformative incomplete characters are actually informative in a probabilistic framework. A reanalysis of Lemmon's data set with this in mind gives a very different interpretation of their results and shows that some of their conclusions may be unfounded. Second, we investigate the effect of the progressive introduction of missing data in a complete supermatrix (126 genes × 39 species) capable of resolving animal relationships. These analyses demonstrate that missing data perturb phylogenetic inference slightly beyond the expected decrease in resolving power. In particular, they

MBE 30(1):197–214, 2012

Maximizing information

- **How do we select data to sequence/include in an analysis?**
- **Example 2:** Should we "filter" genes/sites/taxa?
- From yesterday's talks:
  - **Robert Literman:** SISRS software – will filter by percent missing and by amount of variation in loci
  - **Andy Anderson:** "Phylogenomic analysis of a putative missing link sparks reinterpretation of leech evolution"

User adoption and training

- **I have a great new idea! How do I get people to try it?**

  OR

  That seems like a great idea! Should I invest X+ hours to try to run my data through the software?

- **Example issues:**
  - Little incentive for making user-friendly software (funding, career progression)
  - "Teaching" the method: requires travel, substantial time preparing
  - These things are more or less difficult depending on career stage

- **Big worry** that important methods may be overlooked because they lack a good implementation, while less good methods may become popular because they are easy to run.

User adoption and training

- **How can we best reach people?**
  - **Workshops:** Woods Hole, Bodega Bay, etc.
    if feasible to attend
  - **Traditional classroom instruction**
    if at an institution with an appropriate instructor/resources
  - **Online materials**
    if internet connection and other infrastructure allow
  - **Books**
    keep costs low, keep book accessible

Training the next generation of developers

- **Broad-based experience seems necessary for development of effective methodology**
- **How do we achieve this for our students and post-docs?**
  - What combination of coursework and project-based work is appropriate?
  - Pre-requisite set of skills?
  - How do we adapt to changing demands/trends (technological, analytic, etc.)?

Tell us what you think!

- **Now, it's your turn!**
- Please compete the surveys to provide valuable information on your perceptions of this field.
- Please also communicate with us throughout the day and afterwards.