The statistical and informatics challenges posed by ascertainment biases in phylogenetic data collection

Mark T. Holder and Jordan M. Koch

Department of Ecology and Evolutionary Biology, University of Kansas.

David Swofford, Tracy Heath, David Bryant and Paul Lewis are collaborators on the second part of the talk.

Thanks to NSF and KU's IMSD for funding.

iEvoBio - 2013

Ascertainment bias

a bias in parameter estimation or testing caused by non-random sampling of the data.

This talk will focus on analyses of filtered data.

Filtered data → Some types of data will never be sampled.

Correcting for filtered data in tree estimation

Use:

$$\mathbb{P}(\text{Data} \mid \text{Tree}, \text{not excluded})$$

as the likelihood instead of:

$$\mathbb{P}(\text{Data} \mid \text{Tree})$$

$$\mathbb{P}(\text{Data} \mid \text{Tree}, \text{not excluded}) = \frac{\mathbb{P}(\text{Data} \mid \text{Tree})}{\mathbb{P}(\text{not excluded} \mid \text{Tree})}$$

Felsenstein (1992) and Lewis (2001)

Conclusions

- Analyzing variable-only data with Lewis' Mk_v model is consistent
- Inferring trees from parsimony-informative-only data:
    - can be consistent if the tree is not tiny.
    - can be feasible for multi-state data using new algorithms
- Treating gaps as missing data:
    - does not lead to inconsistency if indel process is independent of the substitution process.
    - can be positively misleading under mild violations of this independence assumption.
- "rules" for character encoding need to be linked to the data. $? \neq -$ in molecular data.

Filtering data: retain variable patterns

	Character					
Taxon	1	2	3	4	5	6
t1	0	1	1	0	0	0
t2	0	1	0	0	1	0
t3	0	1	0	1	1	1
t4	0	0	0	1	0	1

→

	Character				
Taxon	1	2	3	4	5
t1	1	1	0	0	0
t2	1	0	0	1	0
t3	1	0	1	1	1
t4	0	0	1	0	1

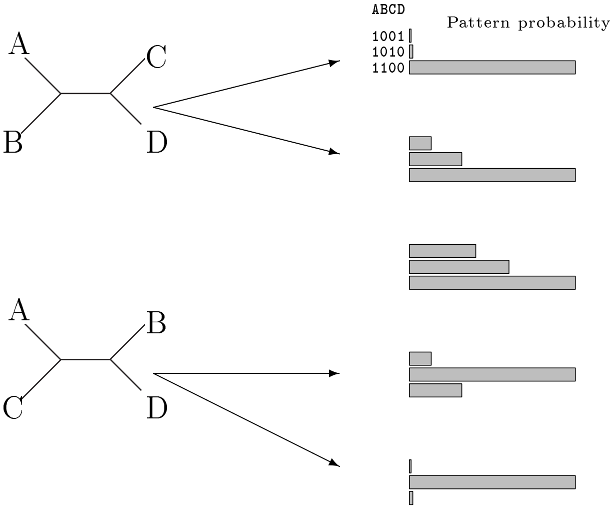
Filtering data: retain parsimony-informative patterns

	Character					
Taxon	1	2	3	4	5	6
t1	0	1	1	0	0	0
t2	0	1	0	0	1	0
t3	0	1	0	1	1	1
t4	0	0	0	1	0	1

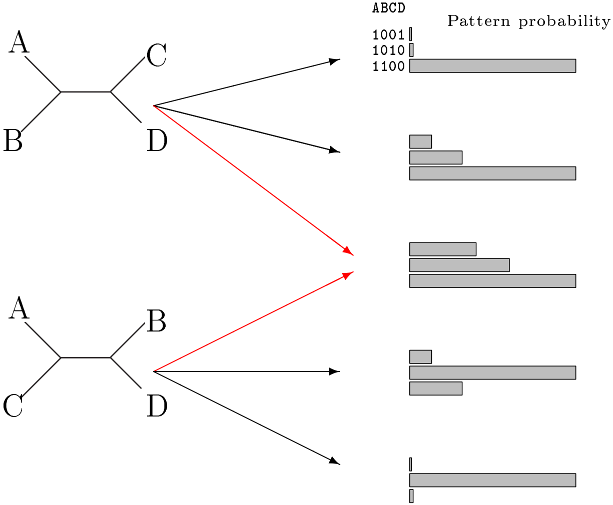
→

	Char.		
Taxon	1	2	3
t1	0	0	0
t2	0	1	0
t3	1	1	1
t4	1	0	1

Identifiability of the tree



Partial identifiability of the tree

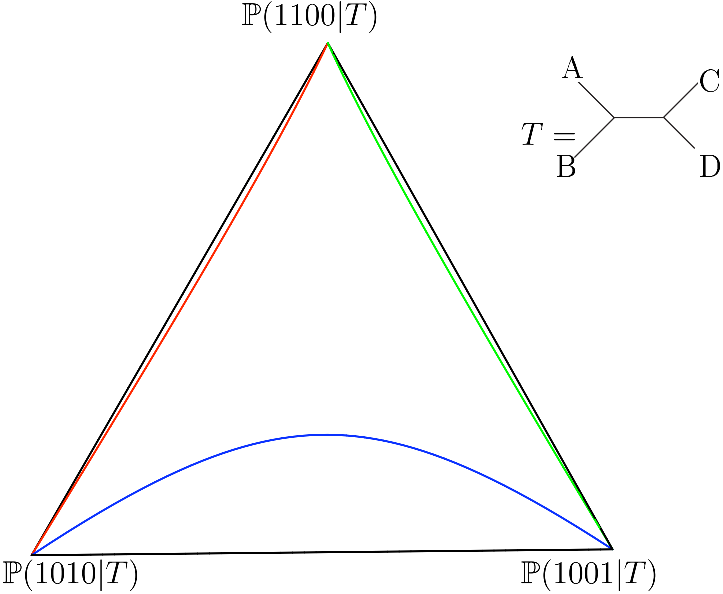


A Markov model for character evolution

$$0 \;\underset{q_{10}}{\overset{q_{01}}{\rightleftarrows}}\; 1$$

Mk: $q_{10} = q_{01}$

GMk: $q_{10} \neq q_{01}$


$\mathbb{P}(1100|T)$
$T =$
A
B
C
D
$\mathbb{P}(1010|T)$
$\mathbb{P}(1001|T)$

Extending identifiability results

Filtering	**Model**	**Identifiable?**
None	GMk	Yes. (Steel, 1994)
Variable	GMk_v	Yes
Pars-inf	Mk_{p-i}	Part. (Steel et al., 1993)
	GMk_{p-i}	$N = 4$ No $N \geq 8$ Yes $5 < N \leq 7$?

results in red: (Allman et al., 2010), extending (Allman and Rhodes, 2008) for GMk_v

Analyzing filtered data

	Filtering	
Analysis	Variable	Pars-inf.
$\mathbf{M}k$	Pos. Misleading	Pos. Misleading
$\mathbf{M}k_v$	Consistent	Pos. Misleading
$\mathbf{M}k_{p-i}$	-	Consistent $N \geq 8$

Calculating the probability of not being excluded

$$\mathbb{P}(\text{not excluded} \mid \text{Tree}) = 1 - \mathbb{P}(\text{excluded} \mid \text{Tree})$$

For variable-only, binary data:

$$\mathbb{P}(\text{Var. pat} \mid \text{Tree}) = 1 - \mathbb{P}(\text{all 0 pat.} \mid \text{Tree}) - \mathbb{P}(\text{all 1 pat.} \mid \text{Tree})$$

Calculating the probability of a parsimony-uninformative pattern

For multi-state character ($k > 2$), and parsimony-informative-only data, there are lots of patterns:

$$\mathcal{O}\left(2^{k-1}\binom{N}{k-1}\right)$$

JMK and MTH have implemented parsimony-uninformative specializations of an algorithm for calculating the prob. of classes of patterns.

Koch and Holder (2012)
`https://github.com/mtholder/PhyPatClassProb`

Informatics implications

To correct for ascertainment bias we need to know what form of data filtering was used.

If a character was chosen because it was variable in a related group, it is difficult to correct of the ascertainment bias.

```
A   PEEKSAVTALWGKVNVDEVGG
B   GEEKAAVLALWDKVNEEEVGG
C   PADKTNVKAAWGKVGAHAGEYGA
D   AADKTNVKAAWSKVGGHAGEYGA
E   EHEWQLVLHVWAKVEADVAGHGQ
```

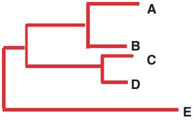
pairwise alignment

```
A  -
B  .17  -
C  .59  .60  -
D  .59  .59  .13  -
E  .77  .77  .75  .75  -
```

tree inference

```
A   PEEKSAVTALWGKVNVDEVGG
B   GEEKAAVLALWDKVNEEEVGG
C   PADKTNVKAAWGKVGAHAGEYGA
D   AADKTNVKAAWSKVGGHAGEYGA
E   EHEWQLVLHVWAKVEADVAGHGQ
```

+



alignment stage

```
A   PEEKSAVTALWGKVN--VDEVGG
B   GEEKAAVLALWDKVN--EEEVGG
C   PADKTNVKAAWGKVGAHAGEYGA
D   AADKTNVKAAWSKVGGHAGEYGA
E   EHEWQLVLHVWAKVEADVAGHGQ
```

gaps to missing

```
A   PEEKSAVTALWGKVN??VDEVGG
B   GEEKAAVLALWDKVN??EEEVGG
C   PADKTNVKAAWGKVGAHAGEYGA
D   AADKTNVKAAWSKVGGHAGEYGA
E   EHEWQLVLHVWAKVEADVAGHGQ
```

Gaps-as-missing-data ML (SML) on the correct alignment

- Clearly ignores information from indels,
- Warnow (2012) argues that the method is inconsistent - but her proof only works when there are no substitutions on the tree.

Holder, Heath, Lewis, Swofford, and Bryant (*in prep*):

1. Proof of consistency if indel process is independent of substitution process,
2. Example of the method being positively misleading under a +I model for indels and substitutions.

Theorem 1

The tree and parameter pair, $\hat{T}_M$, $\hat{\boldsymbol{\theta}}_M$, estimated via SML will yield a consistent estimator of the tree, T, if:

(a) the time-reversible substitution model, $\boldsymbol{\theta}$, results in consistent estimation of the T in the absence of indels;

(b) the indel process, $\boldsymbol{\phi}$, acts independently of substitution process and the sequence states;

(c) the probability distribution for newly inserted states is identical to the the equilibrium state frequency of the substitution process; and

(d) there is non-zero probability of generating a site without gaps under $\boldsymbol{\phi}$.

What if rates of indels are correlated with rates of substitutions?

- Substitution: Jukes-Cantor + a proportion of invariant sites
- Indel: invariant sites will not experience indels. All sites that are free to have substitutions are also free to experience indels.

The result:
If we calculate the expected pattern frequency spectra for extreme "Felsenstein-zone" trees, and mimic infinite character sampling the software using "gaps-as-missing-data" approach (ML and Bayesian) prefers the wrong tree.
We still need to verify that this is not an artifact of local optima being found in software.

Positively misleading behavior from treating gaps as missing data

- Long-branch attraction if the rate of the indel process is correlated with the rate of the substitution process.
- Non-random filtering of data → long branches underestimated.

The result could have implications for the (long-standing) debates in systematics about the effect of missing data and inapplicable character states.

Informatics implications

- Gaps aren't missing data, we really should be using models of the indel process.
- Terminal gaps in alignments often *are* the result of missing data. Software should not use the same symbol for gaps caused by indels and gaps caused by incomplete sequencing.

References I

E Allman and J Rhodes. Identifying evolutionary trees and substitution parameters for the general markov model with invariable sites. *Mathematical biosciences*, 211(1):18–33, Jan 2008. doi: 10.1016/j.mbs.2007.09.001. URL `http://linkinghub.elsevier.com/retrieve/pii/S0025556407001897`.

Elizabeth S. Allman, Mark T. Holder, and John A. Rhodes. Estimating trees from filtered data: Identifiability of models for morphological phylogenetics. *Journal of Theoretical Biology*, 263(1):108–119, 2010. ISSN 0022-5193. doi: DOI:10.1016/j.jtbi.2009.12.001. URL `http://www.sciencedirect.com/science/article/B6WMD-4XX160T-2/2/5adf8b8af77dd551890d7cb5b0e62dba`.

References II

Joseph Felsenstein. Phylogenies from restriction sites: a maximum-likelihood approach. *Evolution*, 46:159–173, Jan 1992.

Jordan M. Koch and Mark T. Holder. An algorithm for calculating the probability of classes of data patterns on a genealogy. *PLOS Currents Tree of Life*, Dec 14 [last modified: 2012 Dec 14](1), 2012. doi: 10.1371/4fd1286980c08. URL `http://currents.plos.org/treeoflife/article/an-algorithm-for-calculating-the-probability-of-classes-c`

P. O. Lewis. A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic Biology*, 50(6):913–925, 2001.

Mike Steel. Recovering a tree from the leaf colourations it generates under a markov model. *Appl. Math. Letters*, 7(2): 19–23, Jan 1994. URL `http://www.math.canterbury.ac.nz/~mathmas/research/markov3.pdf`.

References III

Mike Steel, Michael D. Hendy, and David Penny. Parsimony can be consistent! *Syst Biol*, 42(4):581–587, 1993.

Tandy J. Warnow. Standard maximum likelihood analyses of alignments with gaps can be statistically inconsistent. *PLOS Currents Tree of Life*, Mar 12:[last modified: 2012 Apr 3] Edition 1, 2012. doi: 10.1371/currents.RRN1308. URL `http://currents.plos.org/treeoflife/article/standard-maximum-likelihood-analyses-of-alignments-with-g`