A simulation study comparing phylogeny reconstruction methods for linguistics

Tandy Warnow The University of Texas at Austin The Newton Institute for Mathematical Research

Collaborators: Francois Barbancon, Don Ringe, Luay Nakhleh, Steve Evans

## Possible Indo-European tree (Ringe, Warnow and Taylor 2000)



### Phylogenetic Network for IE Nakhleh *et al.*, Language 2005



## **Controversies for IE history**

- Subgrouping: Other than the 10 major subgroups, what is likely to be true? In particular, what about
  - Italo-Celtic
  - Greco-Armenian
  - Anatolian + Tocharian
  - Satem Core (Indo-Iranian and Balto-Slavic)
  - Location of Germanic
- Dates?
- How tree-like is IE?

### **Controversies for IE history**

Note: many reconstructions of IE have been done, but produce different histories which differ in significant ways (e.g., the location of Germanic)

Possible issues:

Dataset (modern vs. ancient data, errors in the cognancy judgments, lexical vs. all types of characters, screened vs. unscreened)

Translation of multi-state data to binary data

Reconstruction method

The performance of methods on an IE data set (Transactions of the Philological Society, Nakhleh et al. 2005)

**Observation:** Different datasets (not just different methods) can give different reconstructed phylogenies.

**Objective:** Explore the differences in reconstructions as a function of data (lexical alone versus lexical, morphological, and phonological), screening (to remove obviously homoplastic characters), and methods. However, use a *better basic dataset* (where cognancy judgments are more reliable).

### Better datasets

- Ringe & Taylor
  - The screened full dataset of 294 characters (259 lexical, 13 morphological, 22 phonological)
  - The unscreened full dataset of 336 characters (297 lexical, 17 morphological, 22 phonological)
  - The screened lexical dataset of 259 characters.
  - The unscreened lexical dataset of 297 characters.

### Differences between different characters

- Lexical: most easily borrowed (most borrowings detectable), and homoplasy relatively frequent (we estimate about 25-30% overall for our wordlist, but a much smaller percentage for basic vocabulary).
- Phonological: can still be borrowed but much less likely than lexical. Complex phonological characters are infrequently (if ever) homoplastic, although simple phonological characters very often homoplastic.
- Morphological: least easily borrowed, least likely to be homoplastic.

### Lexical Characters

- For each basic meaning, assign two languages the same state if they contain cognates
- Example: basic meaning 'hand'
- French: main, Spanish: mano, Italian: mano, German: hand, English hand, Russian ruká
- Mathematically: Fr: 1, Sp: 1, It: 1, Ger: 2, Eng: 2, Rus: 3.

Table 1: The 24 IE languages analyzed.

Language	Abbreviation	Language	Abbreviation
Hittite	HI	Old English	OE
Luvian	LU	Old High German	OG
Lycian	LY	Classical Armenian	AR
Vedic	VE	Tocharian A	TA
Avestan	AV	Tocharian B	TB
Old Persian	PE	Old Irish	OI
Ancient Greek	GK	Welsh	WE
$\operatorname{Latin}$	LA	Old Church Slavonic	OC
Oscan	OS	Old Prussian	$\mathbf{PR}$
Umbrian	UM	Lithuanian	$\operatorname{LI}$
Gothic	GO	Latvian	LT
Old Norse	ON	Albanian	AL

### Phylogeny reconstruction methods

- Distance-based polynomial time methods: Neighbor joining and UPGMA (technique in glottochronology)
- Character-based methods:
  - Maximum parsimony (minimize number of changes),
  - Maximum compatibility (minimize number of incompatible characters)
  - Gray and Atkinson (Bayesian estimation based upon presence/absence of cognates, published in Nature 2003 with lots of controversy)

### Some observations

- UPGMA (i.e., the tree-building technique for glottochronology) does the worst (e.g. splits Italic and Iranian groups).
- Other than UPGMA, all methods reconstruct the ten major subgroups, as well as Anatolian + Tocharian and Greco-Armenian.
- The Satem Core (Indo-Iranian plus Balto-Slavic) is not always reconstructed.
- Almost all analyses put Italic, Celtic, and Germanic together. (The only exception is weighted maximum compatibility on datasets that include morphological characters.)





GA = Gray+Atkinson Bayesian MCMC method

WMC = weighted maximum compatibility

MC = maximum compatibility (identical to maximum parsimony on this dataset)

NJ = neighbor joining (distance-based method, based upon corrected distance)

UPGMA = agglomerative clustering technique used in glottochronology.

Different methods/data give different answers. We don't know which answer is correct. Which method(s)/data should we use?

### Simulation study (cartoon)



TRUE TREE



- S<sub>2</sub> ACCCTTAGAAC
- $S_3$  ACCATTCCAAC
- $S_4$  Accagaccaac
- $S_5 \qquad \text{ACCAGACCGGA}$

#### DNA SEQUENCES



INFERRED TREE

### Simulation study (cartoon)



TRUE TREE



#### DNA SEQUENCES

- FN: false negative (missing edge)
- FP: false positive (incorrect edge)

50% error rate



INFERRED TREE

### **Phylogenetic Network Evolution**



# Modelling borrowing: Networks and Trees within Networks



### Some useful terminology: homoplasy



### Some useful terminology: lexical clock





no lexical clock

edge lengths represent expected numbers of substitutions

### Heterotachy = departure from rates-across-sites



The underlying tree is fixed, but there are no constraints on edge length variations between characters.

### **Previous simulations**

- Most previous simulations of linguistic evolution had evolved characters without any borrowing or homoplasy, all under an i.i.d. assumption, and many have assumed a strong lexical clock.
- Some (notably McMahon and McMahon) had evolved characters with small amounts of borrowing but no homoplasy, on small networks (with one contact edge)

### Our datasets

- Unscreened (moderate homoplasy):
  - Morphology: 24% homoplastic, no borrowing
  - Lexical: 13% homoplastic, 7% borrowing
- Screened (low homoplasy):
  - Morphology: no homoplasy, no borrowing
  - Lexical: 1% homoplasy, 6% borrowing

## Our simulation study (in press)

- Model phylogenetic networks: each had 30 leaves and up to three contact edges, and varied in the *deviation from a lexical clock*.
- Characters: we had up to 360 lexical characters and 60 morphological characters, and varied probability of homoplasy and borrowing.
- Performance metric: We compared estimated trees to the "genetic tree" wrt the missing edge rate.

### **Observations**

- 1. Choice of reconstruction method does matter.
- 2. Relative performance between methods is quite stable (distance-based methods worse than character-based methods).
- 3. Choice of data does matter (good idea to add morphological characters).
- Accuracy only slightly lessened with small increases in homoplasy, borrowing, or deviation from the lexical clock.
- 5. Some amount of heterotachy helps!



Relative performance of methods on moderate homoplasy datasets under various model conditions:

- (i) varying the deviation from the lexical clock,
- (ii) varying heterotachy, and
- (iii) varying the number of contact edges.





(i)

Relative performance of methods for low homoplasy datasets under various model conditions:

- (i) Varying the deviation from the lexical clock,
- (ii) Varying the heterotachy, and
- (iii) Varying the number of contact edges.





Impact of homoplasy for characters evolved down a network with three contact edges under a moderate deviation from the lexical clock and moderate heterotachy.



Impact of homoplasy for characters evolved down a tree under a moderate deviation from a lexical clock and moderate heterotachy. (Our weighting is inappropriate for "unscreened" data.)



Impact of the number of contact edges for characters evolved under low homoplasy, moderate deviation from a lexical clock, and moderate heterotachy.



Impact of the deviation from a lexical clock (from low to moderate) for characters evolved down a network with three contact edges under low levels of homoplasy and with moderate heterotachy.



Impact of heterotachy for characters evolved down a network with three contact edges, with low homoplasy, and with moderate deviation from a lexical clock. Heterotachy increases with the parameter.



Impact of data selection for characters evolved down a network with three contact edges, under low homoplasy ("screened data"), moderate deviation from a lexical clock, and moderate heterotachy.

### **Observations**

- 1. Choice of reconstruction method does matter.
- 2. Relative performance between methods is quite stable (distance-based methods worse than character-based methods).
- 3. Choice of data does matter (good idea to add morphological characters).
- Accuracy only slightly lessened with small increases in homoplasy, borrowing, or deviation from the lexical clock.
- 5. Some amount of heterotachy helps!

### Future research

 We need more investigation of methods based on stochastic models (Bayesian beyond G+A, maximum likelihood, NJ with better distance corrections), as these are now the methods of choice in biology. This requires better models of linguistic evolution and hence input from linguists!

### Future research (continued)

- Should we screen? The simulation uses low homoplasy as a proxy for screening, but real screening throws away data and may introduce bias.
- How do we detect/reconstruct borrowing?
- How do we handle missing data in methods based on stochastic models?
- How do we handle polymorphism?

### Acknowledgements

- Funding: NSF, the David and Lucile Packard Foundation, the Radcliffe Institute for Advanced Studies, The Program for Evolutionary Dynamics at Harvard, and the Institute for Cellular and Molecular Biology at UT-Austin.
- Collaborators: Don Ringe, Steve Evans, Luay Nakhleh, and Francois Barbancon.

### For more information

 Please see the Computational Phylogenetics for Historical Linguistics web site for papers, data, and additional material

http://www.cs.rice.edu/~nakhleh/CPHL