Detecting language contact in Indo-European

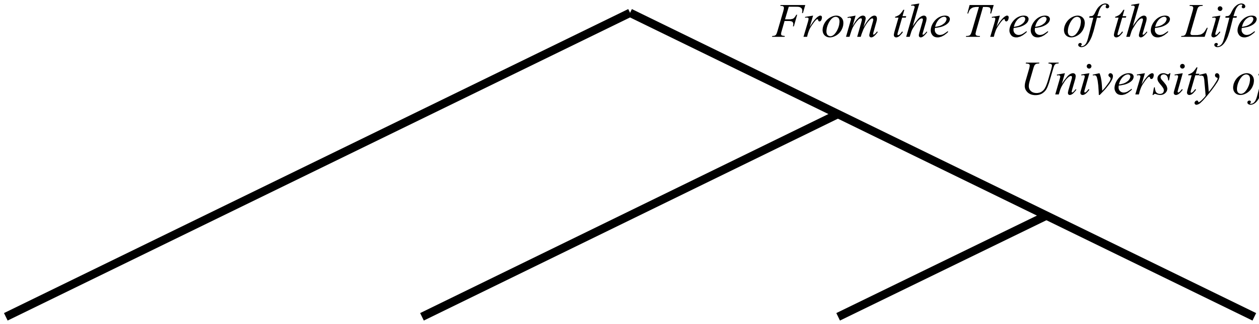
Tandy Warnow

The Program for Evolutionary Dynamics at Harvard

The University of Texas at Austin

(Joint work with Don Ringe, Steve Evans, and Luay Nakhleh)

Species phylogeny

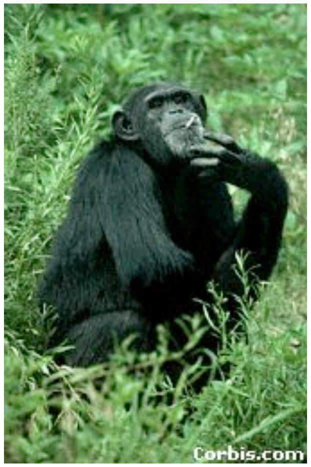


From the Tree of the Life Website,
University of Arizona

Orangutan

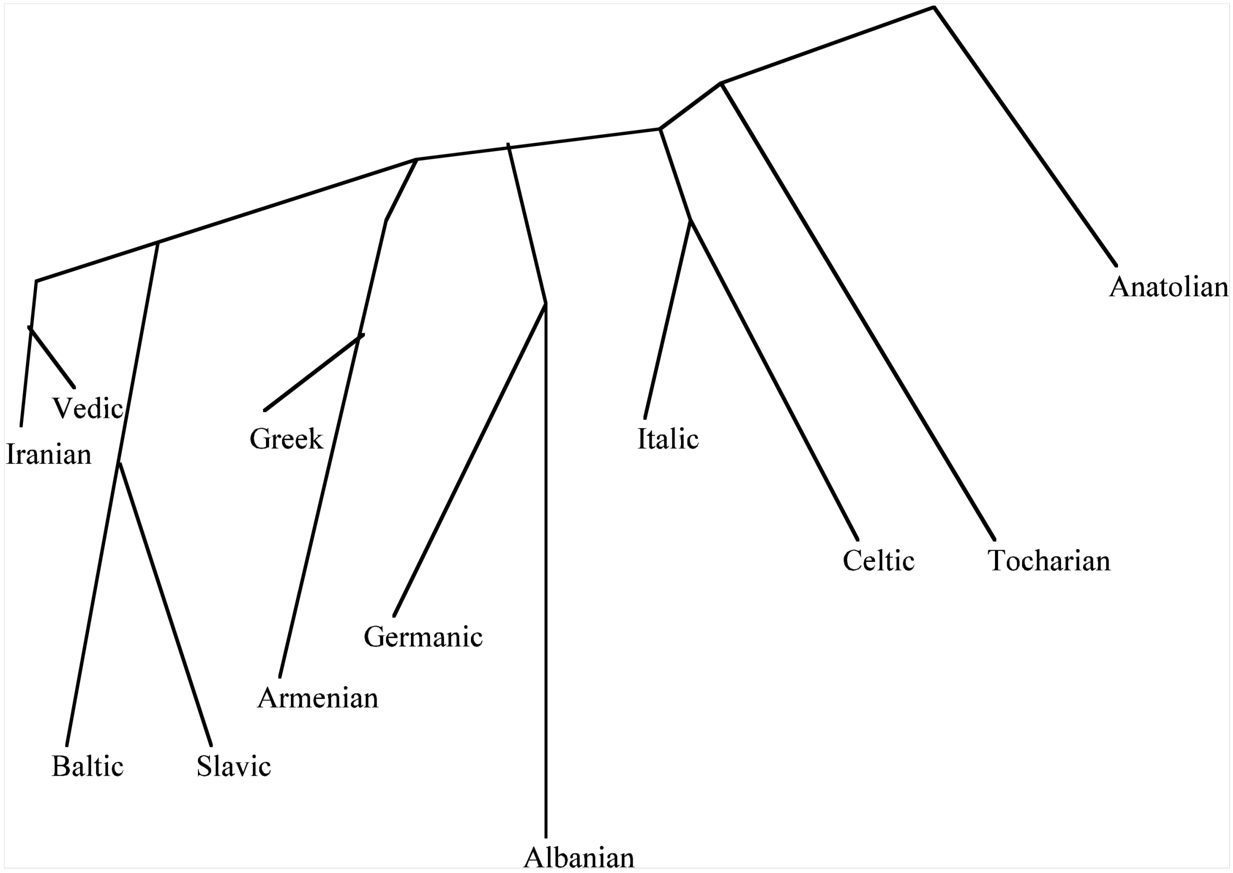
Gorilla

Chimpanzee

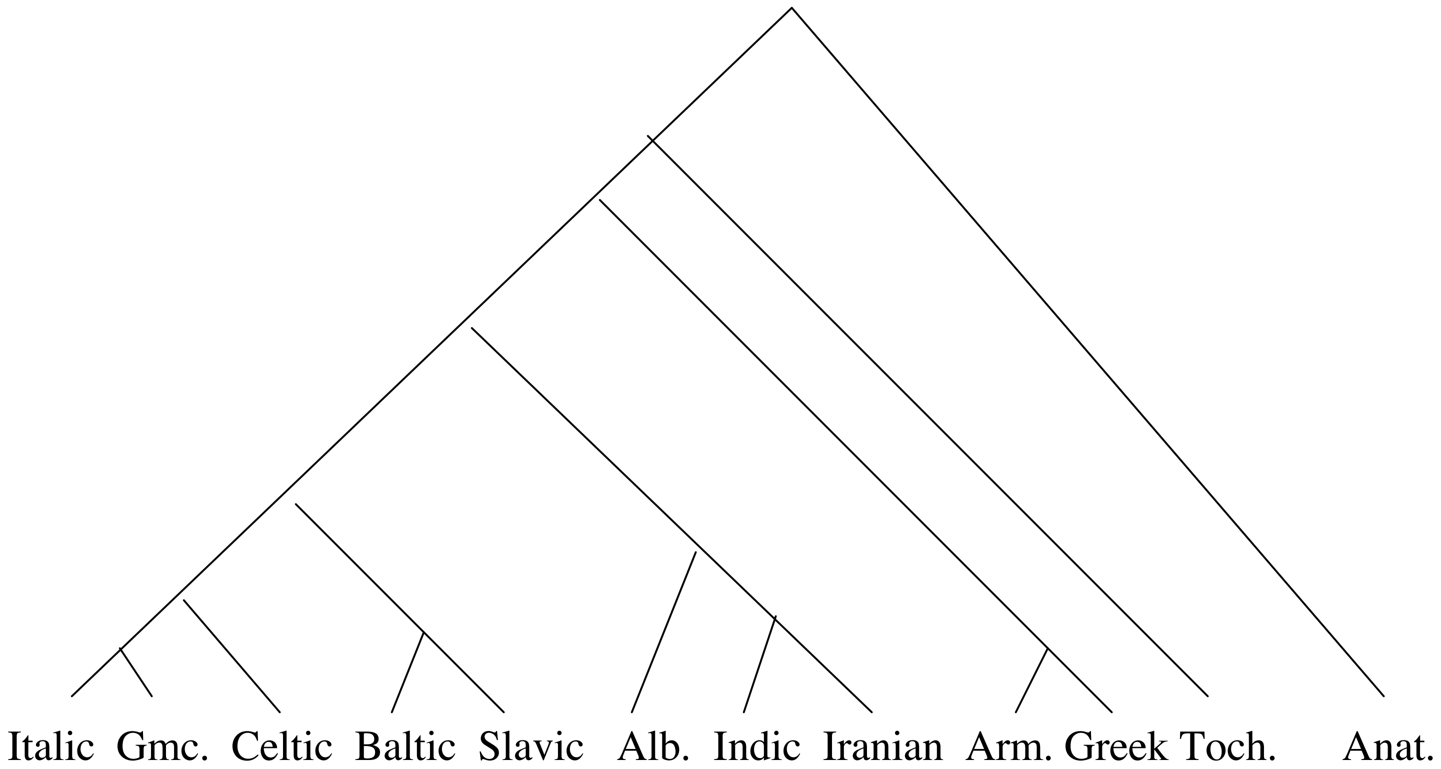


Human

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



Another possible Indo-European tree (Gray & Atkinson, 2004)



Controversies for Indo-European 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?
- Dates? A reconstruction of IE by biologists Gray & Atkinson (Nature, 2004) proposes that the origins of IE are 10,000 years ago, at least 2,000 years earlier than what historical linguists believe.

Why do biologists want to use their tools in historical linguistics?

- There are similarities in the issues involved in estimating evolutionary histories in both linguistics and in biology.
- Statistical estimation approaches (based upon stochastic models of evolution) have greatly impacted molecular phylogenetics.
- Hence, biologists may hope/expect/believe that similar approaches could yield significant contributions to Historical Linguistics.

Our main points

- Biomolecular data evolve differently from linguistic data, and linguistic models and methods should *not* be based upon biological models.

Our main points

- Biomolecular data evolve differently from linguistic data, and linguistic models and methods should *not* be based upon biological models.
- Better (more accurate) phylogenies can be obtained by formulating models and methods based upon linguistic scholarship, and using good data.

Our main points

- Biomolecular data evolve differently from linguistic data, and linguistic models and methods should *not* be based upon biological models.
- Better (more accurate) phylogenies can be obtained by formulating models and methods based upon linguistic scholarship, and using good data.
- Estimating dates at internal nodes requires better models than we have. All current approaches make strong model assumptions that probably do not apply to IE (or other language families).

Our main points

- Biomolecular data evolve differently from linguistic data, and linguistic models and methods should *not* be based upon biological models.
- Better (more accurate) phylogenies can be obtained by formulating models and methods based upon linguistic scholarship, and using good data.
- Estimating dates at internal nodes requires better models than we have. All current approaches make strong model assumptions that probably do not apply to IE (or other language families).
- All methods (whether explicitly based upon statistical models or not) need to be tested (preferably in simulation).

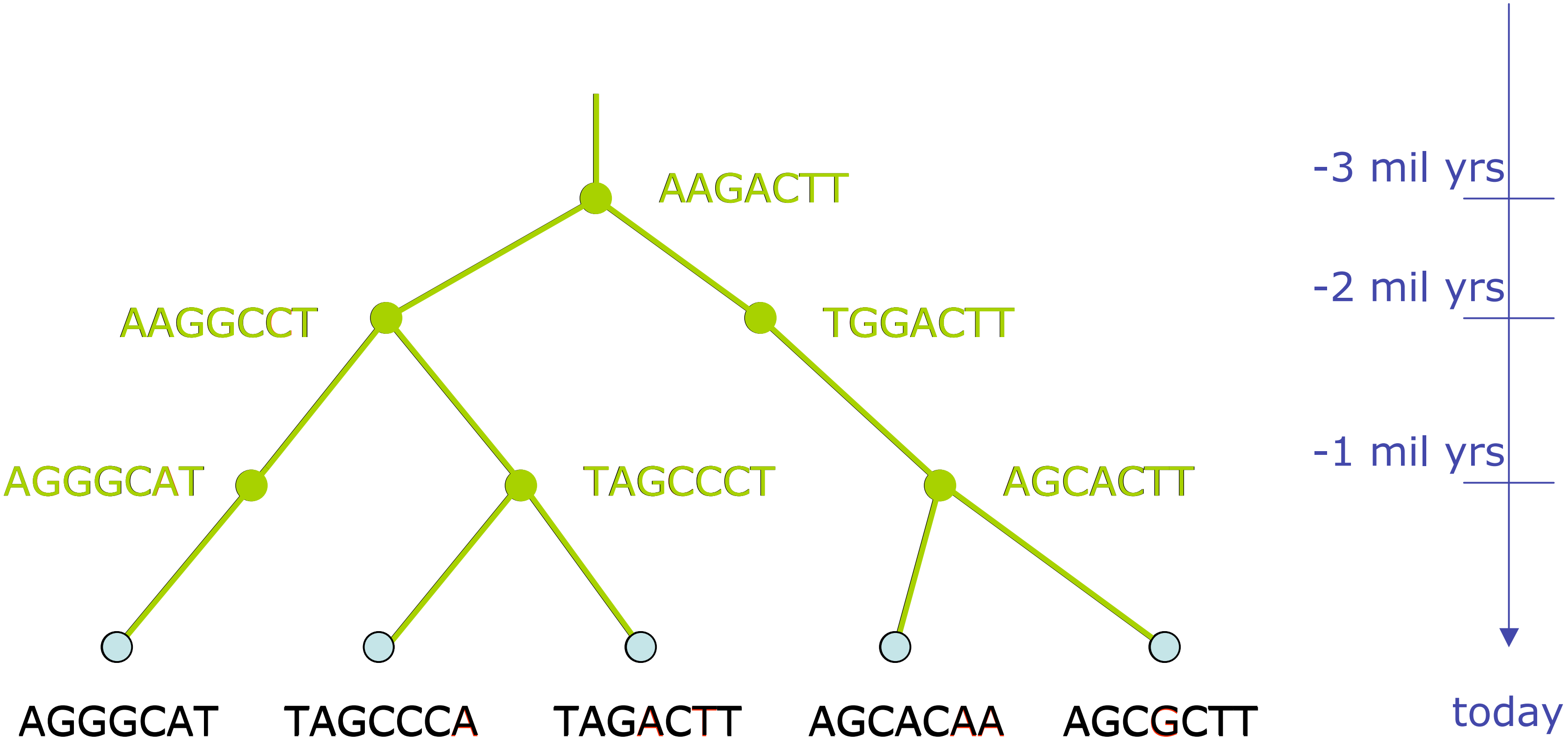
This talk

- General introduction to stochastic models of evolution, statistical estimation of phylogenies, and issues about dating internal nodes
- Differences between models in biology and in linguistics
- New models of language evolution incorporating borrowing and/or “homoplasy”, and a reconstruction of Indo-European
- Comparison to other methods
- Future work

Steps in phylogeny reconstruction

1. Gather **data**
2. Select/design a **model** for the evolutionary process
3. Apply a reconstruction **method** to find phylogenies (evolutionary histories) that best fit the model and the data

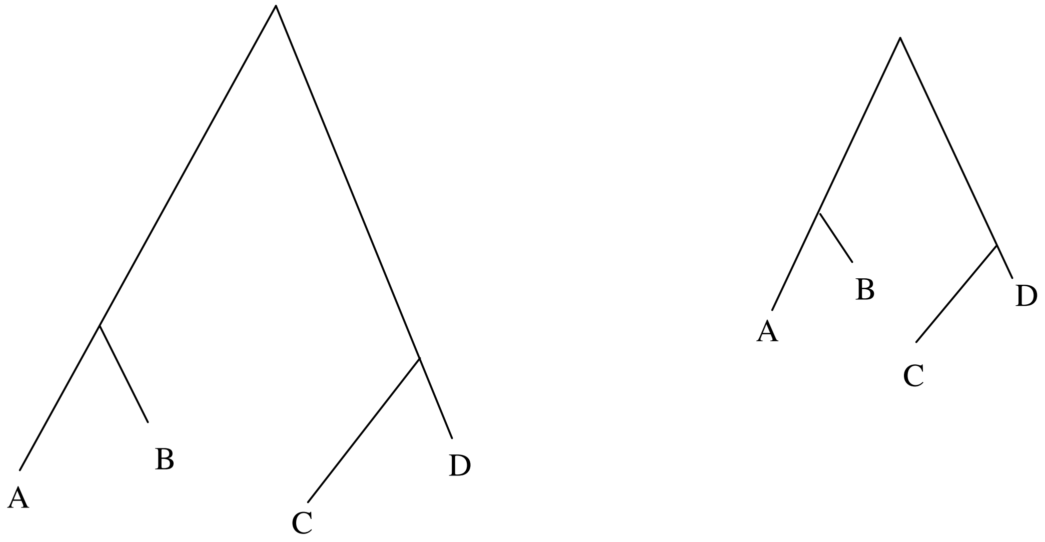
DNA Sequence Evolution



Standard assumptions about single site evolution

- There is a fixed and *finite* set of states (e.g., {A,C,T,G}).
- Each edge has a *length*, which is the number of times the site is expected to change state.
- There is one common *4x4 substitution matrix*.

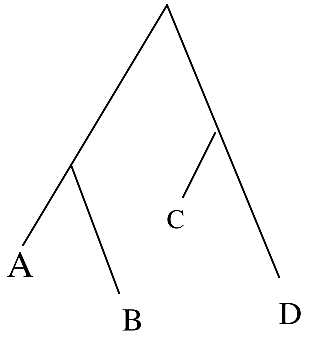
Rates-across-sites

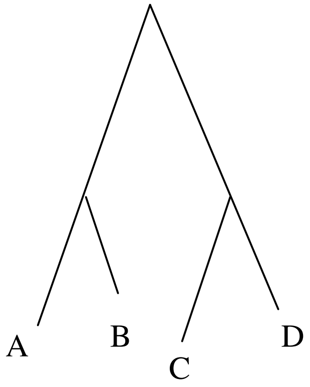


- If a site (i.e., character) is twice as fast as another on one edge, it is twice as fast everywhere.

The no-common-mechanism model

- In this model, there is a separate random variable for every combination of site and edge - the underlying tree is fixed, but otherwise there are no constraints on variation between sites.
- Including this assumption in the usual molecular evolution models makes the tree and dates unidentifiable.





Standard assumptions about variation between sites

- Sites evolve *independently* of each other.
- Each site has a rate-of-evolution, which scales its expected number of changes up or down relative to some fixed character: this is the *rates-across-sites* assumption.
- The site-specific rates of evolution are drawn from a *known distribution* (or one with a small number of parameters which can be estimated from the data).

Summary of molecular sequence phylogeny

- Data: lots of *homoplasy* (parallel evolution, and/or character reversal)
- Models: the models for single character evolution are quite complex, but the properties relating how different characters evolve are heavily constrained and unrealistic.

Biological models include questionable assumptions for theoretical tractability (in particular to ensure identifiability of the model). These assumptions may make phylogenetic reconstruction easier, but *not necessarily more accurate*.

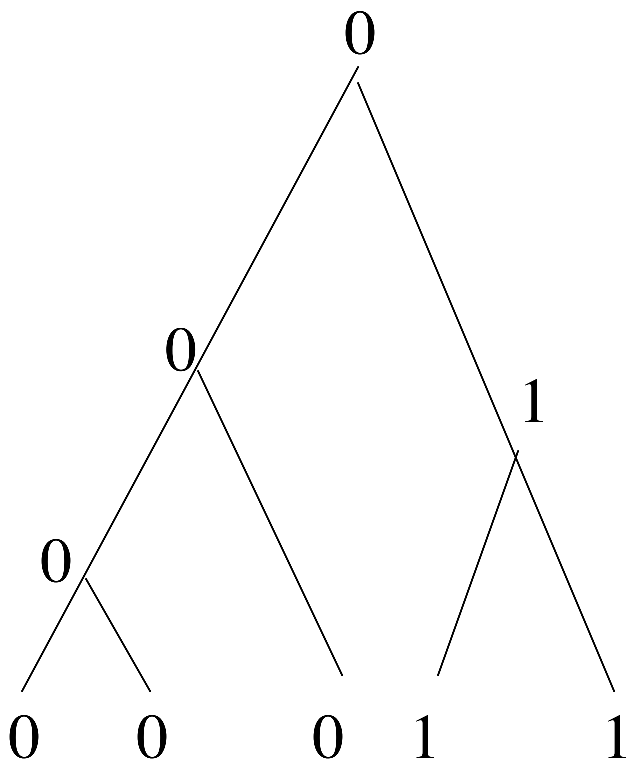
Historical Linguistic Data

- A character is a function that maps a set of languages, *L*, to a set of states.

- Three kinds of characters:
  - Phonological (sound changes)
  - Lexical (meanings based on a wordlist)
  - Morphological (especially inflectional)

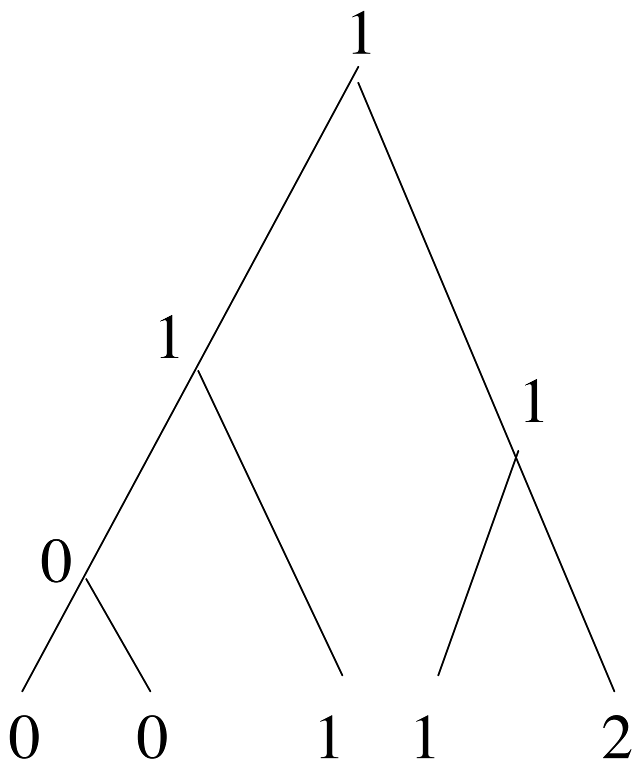
Homoplasy-free evolution

- When a character changes state, it changes to a new state not in the tree
- In other words, there is no homoplasy (character reversal or parallel evolution)
- First inferred for *weird innovations* in phonological characters and morphological characters in the 19th century.

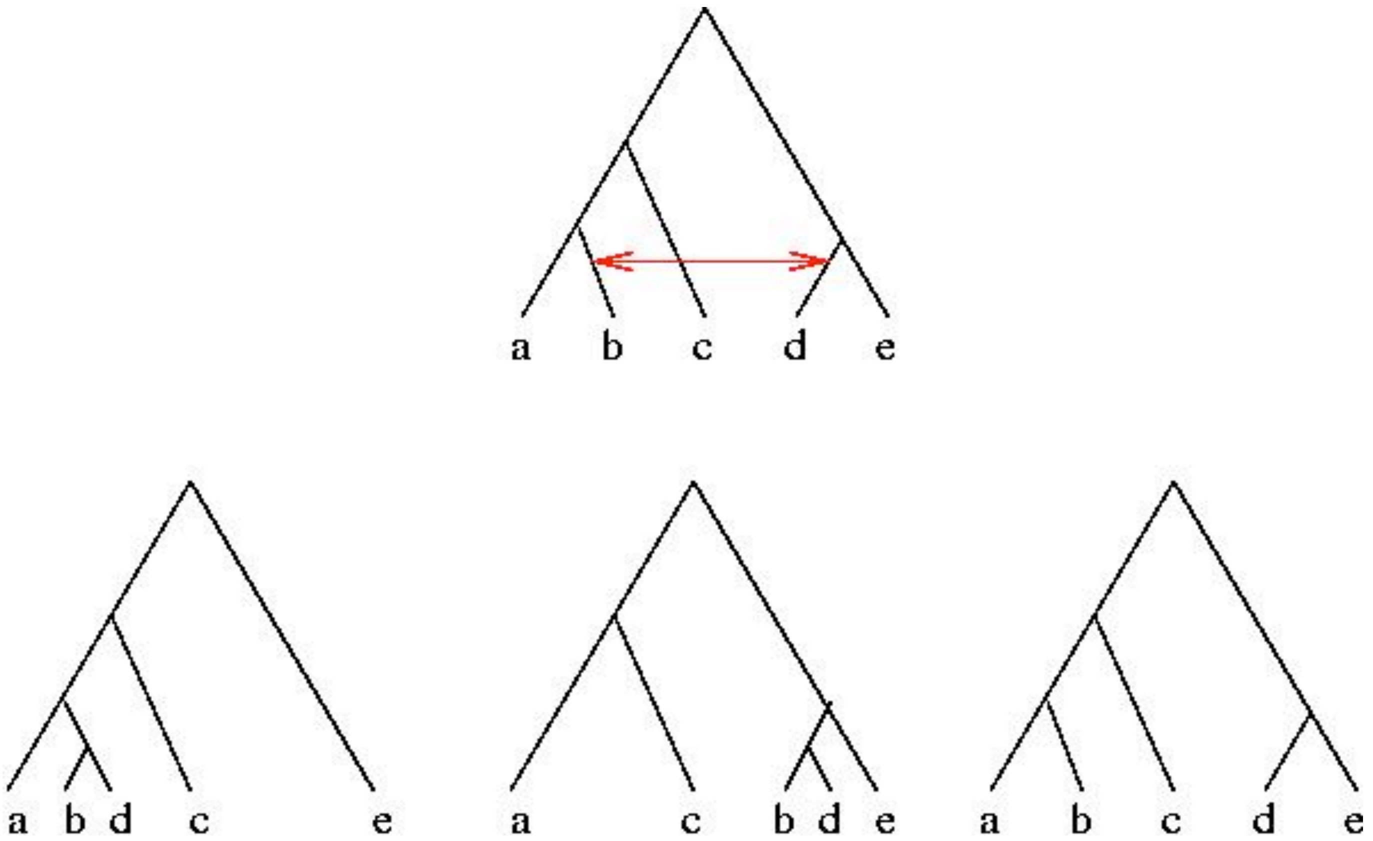


Lexical characters can also evolve without homoplasy

- For every cognate class, the nodes of the tree in that class should form a connected subset - *as long as there is no undetected borrowing nor parallel semantic shift.*
- However, in practice, lexical characters are more likely to evolve homoplastically than complex phonological or morphological characters.



Modelling borrowing: Networks and Trees within Networks



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.

Linguistic character evolution

- Characters are lexical, phonological, and morphological.
- Homoplasy is much less frequent: most changes result in a new state (and hence there is an unbounded number of possible states).
- There is even less basis for the assumption that the characters evolve under a rates-across-sites model.
- Borrowing between languages occurs, but can often be detected.
- NOTE: these properties are very different from models for molecular sequence evolution. Therefore, ***using models from molecular phylogenetics is problematic.***

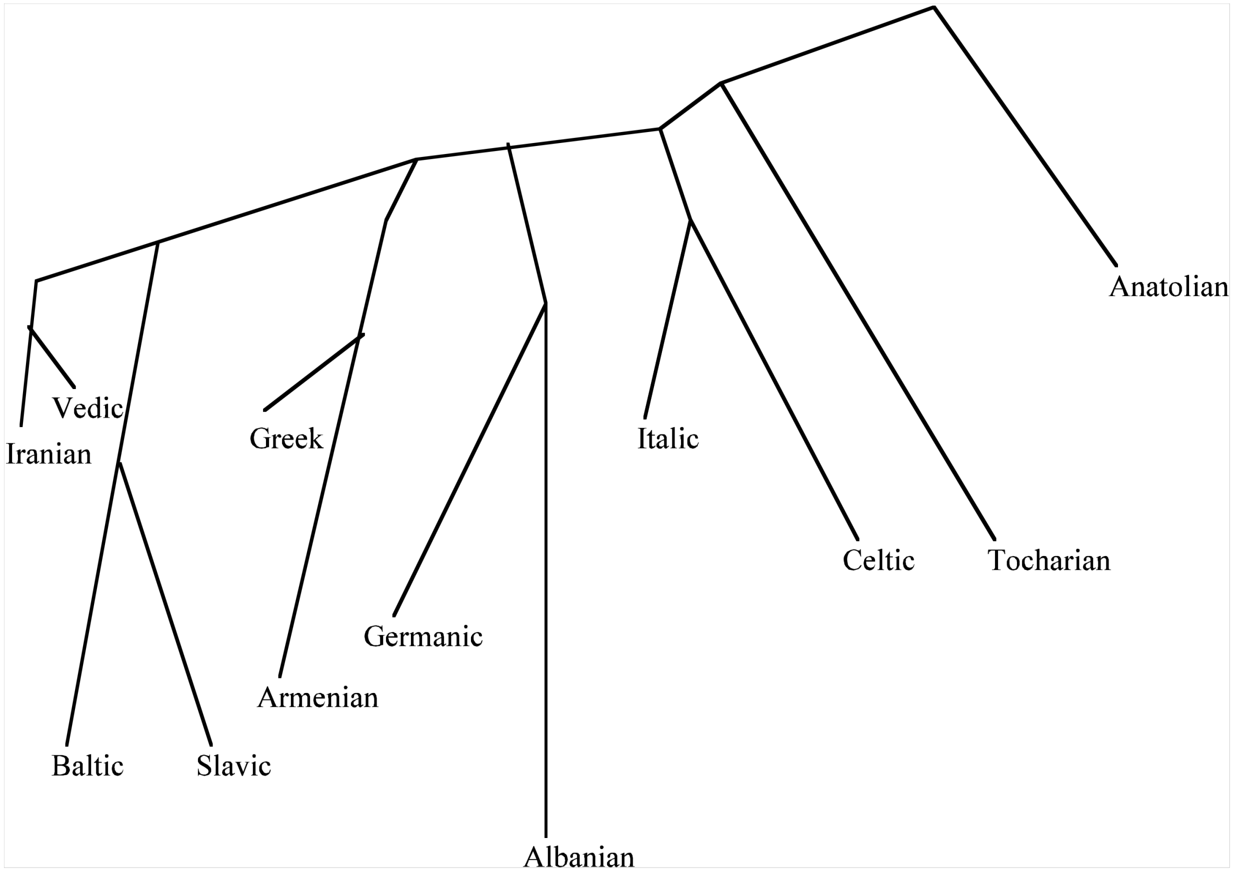
Our methods/models

- Ringe & Warnow “Almost Perfect Phylogeny”: most characters evolve without homoplasy under a no-common-mechanism assumption (various publications since 1995)
- Ringe, Warnow, & Nakhleh “Perfect Phylogenetic Network”: extends APP model to allow for borrowing, but assumes homoplasy-free evolution for all characters (to appear, Language, 2005)
- Warnow, Evans, Ringe & Nakhleh “Extended Markov model”: parameterizes PPN and allows for homoplasy provided that homoplastic states can be identified from the data (to appear in Cambridge University Press)
- Ongoing work: incorporating unidentified homoplasy.

First analysis: Almost Perfect Phylogeny

- The original dataset contained 375 characters (336 lexical, 17 morphological, and 22 phonological).
- We *screened* the dataset to eliminate characters likely to evolve homoplastically or by borrowing.
- On this reduced dataset (259 lexical, 13 morphological, 22 phonological), we attempted to maximize the number of compatible characters while *requiring that certain of the morphological and phonological characters be compatible*. (Computational problem NP-hard.)

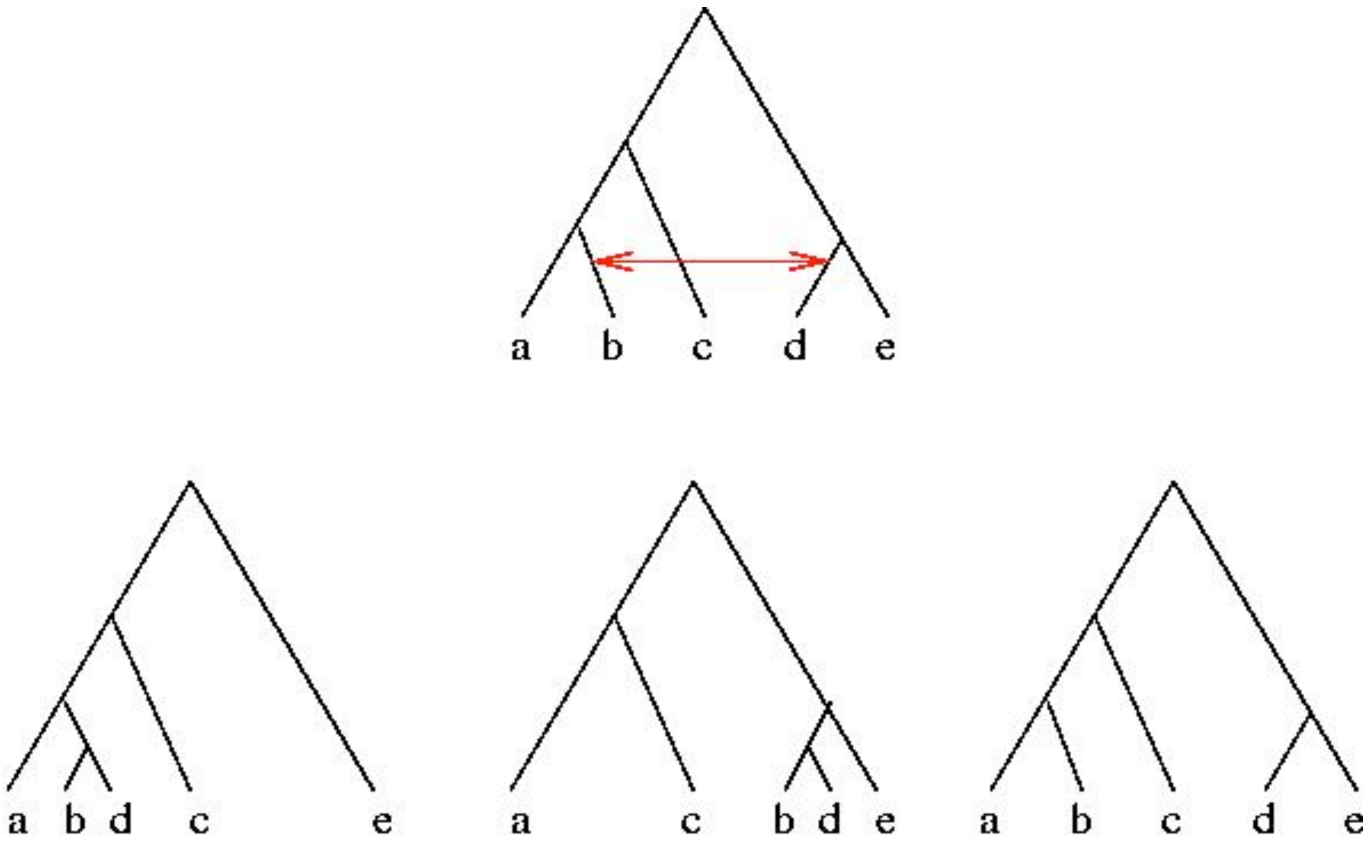
Indo-European Tree
(95% of the characters compatible)



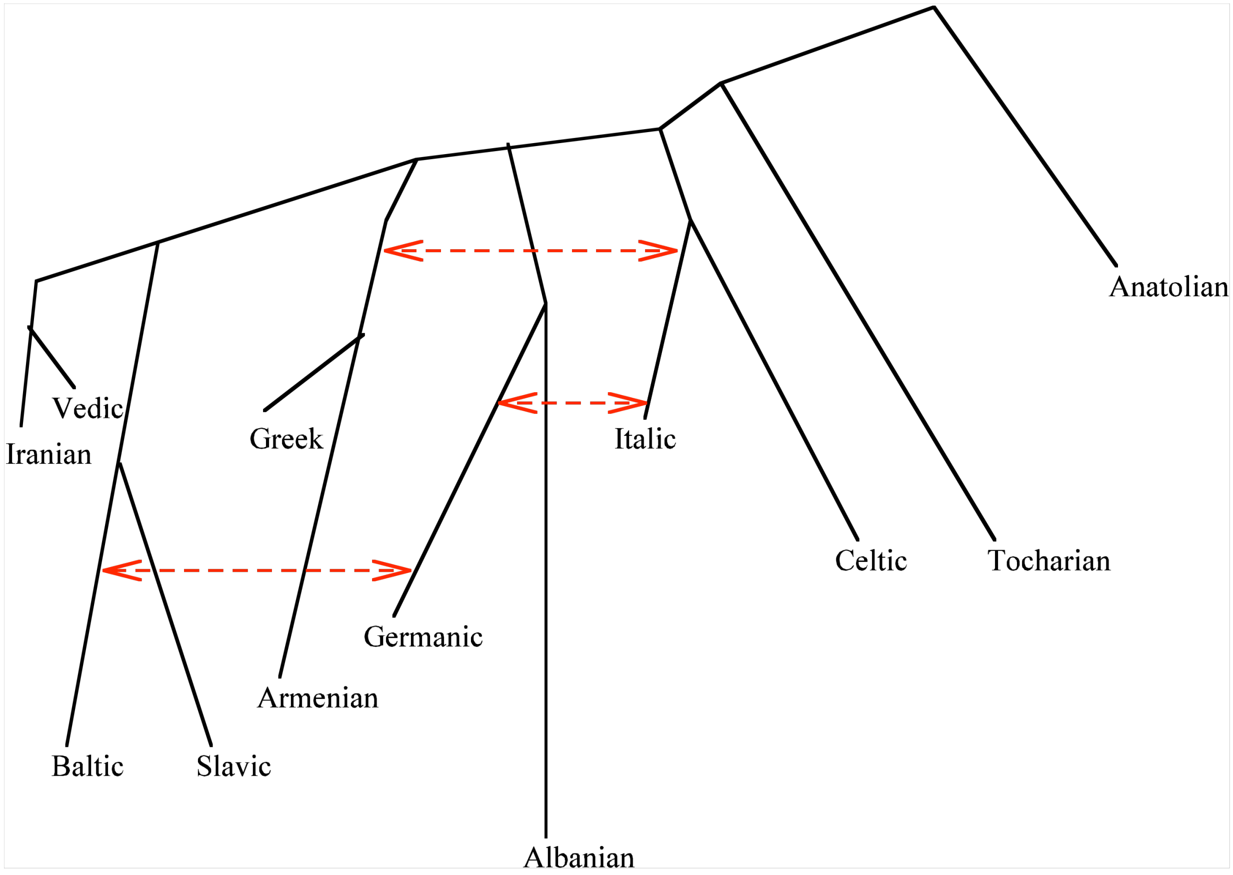
Second attempt: PPN

- We explain the remaining incompatible characters by inferring previously ***undetected*** "borrowing".
- We attempted to find a PPN (perfect phylogenetic network) with the smallest number of contact edges, borrowing events, and with maximal feasibility with respect to the historical record. (Computational problems NP-hard).
- Our analysis produced one solution with only three contact edges that optimized each of the criteria. Two of the contact edges are well-supported.

Networks and Trees



“Perfect Phylogenetic Network” (all characters compatible)



Issue: modelling homoplasy

- We observed that of the three contact edges, only two are well-supported. If we eliminate that weakly supported edge, then we must explain the incompatibility of some characters through homoplasy instead of borrowing.

- Challenge: How to model homoplasy, borrowing, and genetic transmission, appropriately?

Extended Markov model

- There are two types of states: those that can arise more than once, but others can arise only once, and for each state of each character we know which type it is. (This information is not inferred by the estimation procedure.)
- There are two types of substitutions: homoplastic and non-homoplastic.
- Parameters: each character has its own 2x2 substitution matrix, and a relative probability of being borrowed. Each "contact edge" has a relative probability of transmitting character states.
- Each character evolves down a tree contained within the network. The characters evolve independently under this *no-common-mechanism* model.

Initial results

- The model tree is identifiable under very mild conditions (where the substitution probabilities are bounded away from 0 and 1).
- Statistically consistent and efficient methods exist for reconstructing trees (as well as some networks).
- Maximum Likelihood and Bayesian analyses are also feasible, since likelihood calculations can be done in linear time.

Ongoing model development

- Not all homoplastic states are identifiable! Therefore, our ongoing work is seeking to develop improved models of language evolution which permit unidentified homoplasy. Such models are not likely to be identifiable, making inference of evolution more difficult
- Polymorphism (i.e., two or more states of a character present in a language) remains insufficiently characterized, and therefore cannot yet be used rigorously in a phylogenetic analysis. Our earlier work provided an initial model when evolution is tree-like, but we need to extend the model in the presence of borrowing.

Comparison to other work

- Gray and Atkinson (Nature, 2004) used a very different technique (MrBayes analysis of binary-encoding of lexical characters, assuming rates-across-sites and a relaxed molecular clock).
- Maximum Parsimony (minimizes number of changes) and Maximum Compatibility
- Lexico-statistics (distance-based approach, assumes molecular clock)
- Weighted Maximum Compatibility (requires linguist to assess reliability of input characters)

Our study

- Use better datasets, some of which includes morphological and phonological characters in addition to lexical characters, and some of which are only lexical.
- Compare various phylogenetic reconstruction methods on these better datasets.
- Evaluate the consequences of using different datasets.

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.

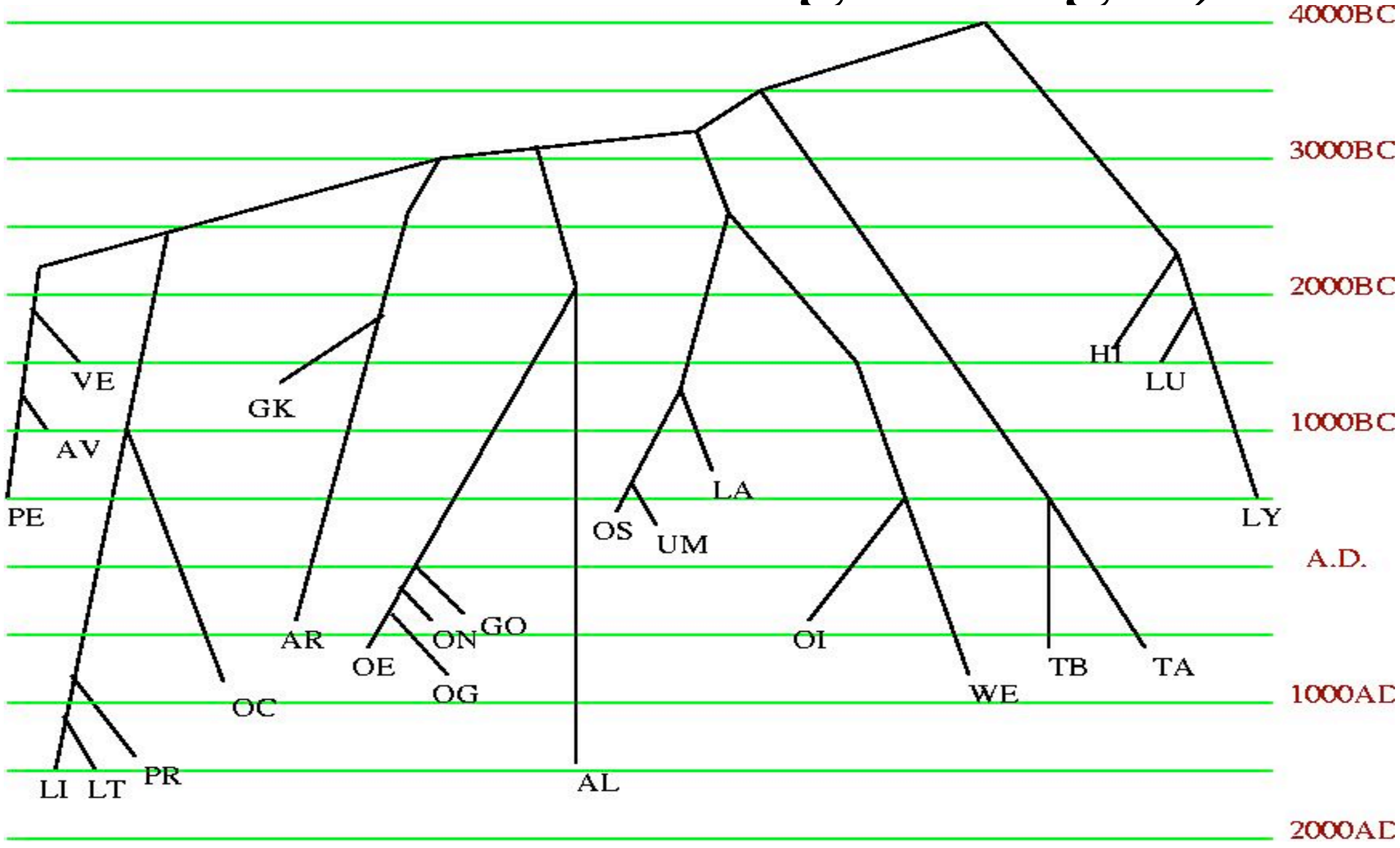
General observations

- UPGMA (lexico-statistics) does the worst.
- Other than UPGMA, all methods reconstruct the ten major subgroups, Anatolian + Tocharian, and Greco-Armenian.
- The Satem Core is not always reconstructed.
- The only analyses that do *not* put Italic and Celtic with Germanic are weighted maximum compatibility on the full datasets with high weights on morphological characters.
- When using lexical data only, all methods group Italic, Celtic, and Germanic together.
- ***Methods differ significantly on the datasets - and have different sets of incompatible characters***

Some characters

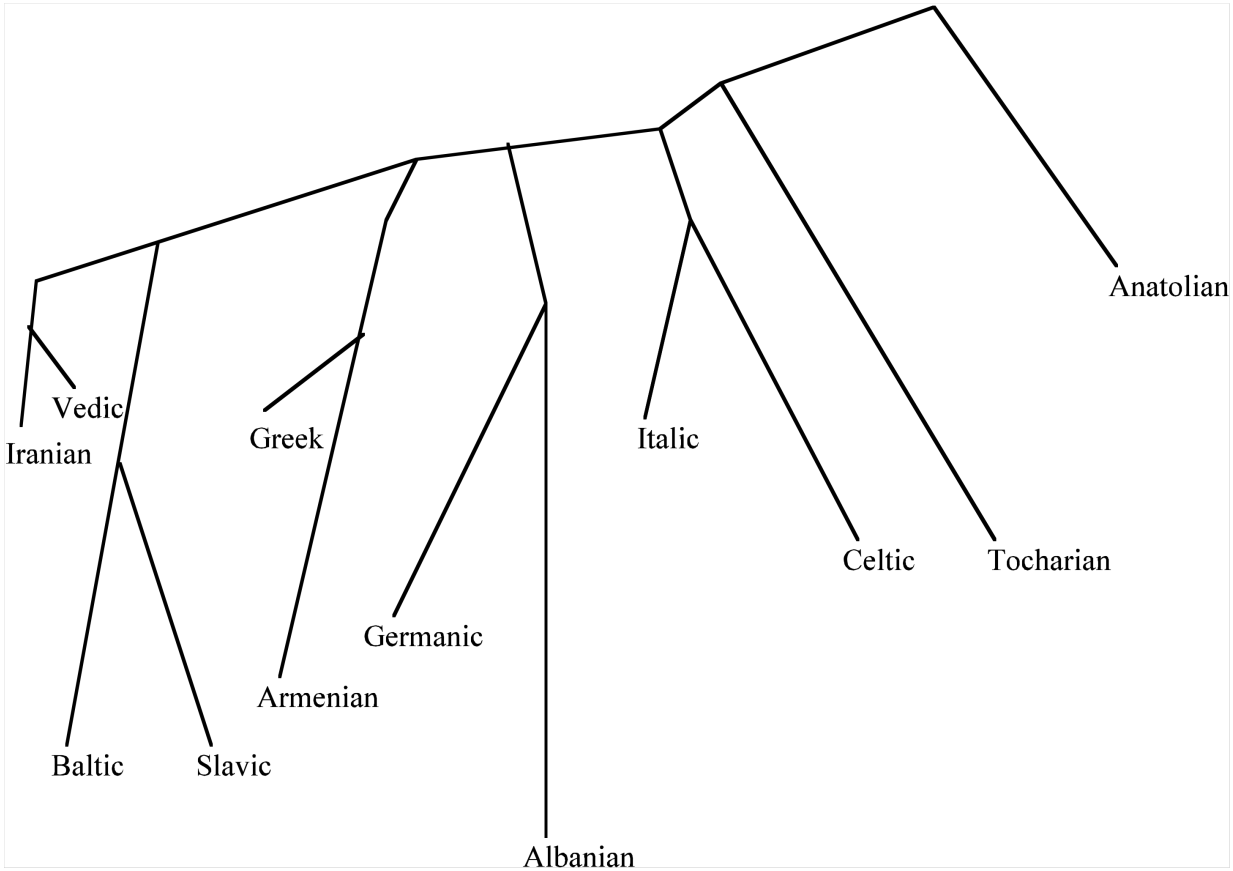
- M9a (one coding of the athematic dative plural ending)
- P2 (the “satem” development of dorsals)
- P3 (the “ruki” rule)
- M11 (*-ti- followed by (*-Hen-)
- P1 (*p…k^w > *k^w…k^w)
- M6 (the thematic optative suffix)
- M8 (the most archaic superlative suffix)
- M5 (the shape of the mediopassive primary person and number endings)

First WMC tree (all morphological and phonological characters have high weight)



First WMC analysis:

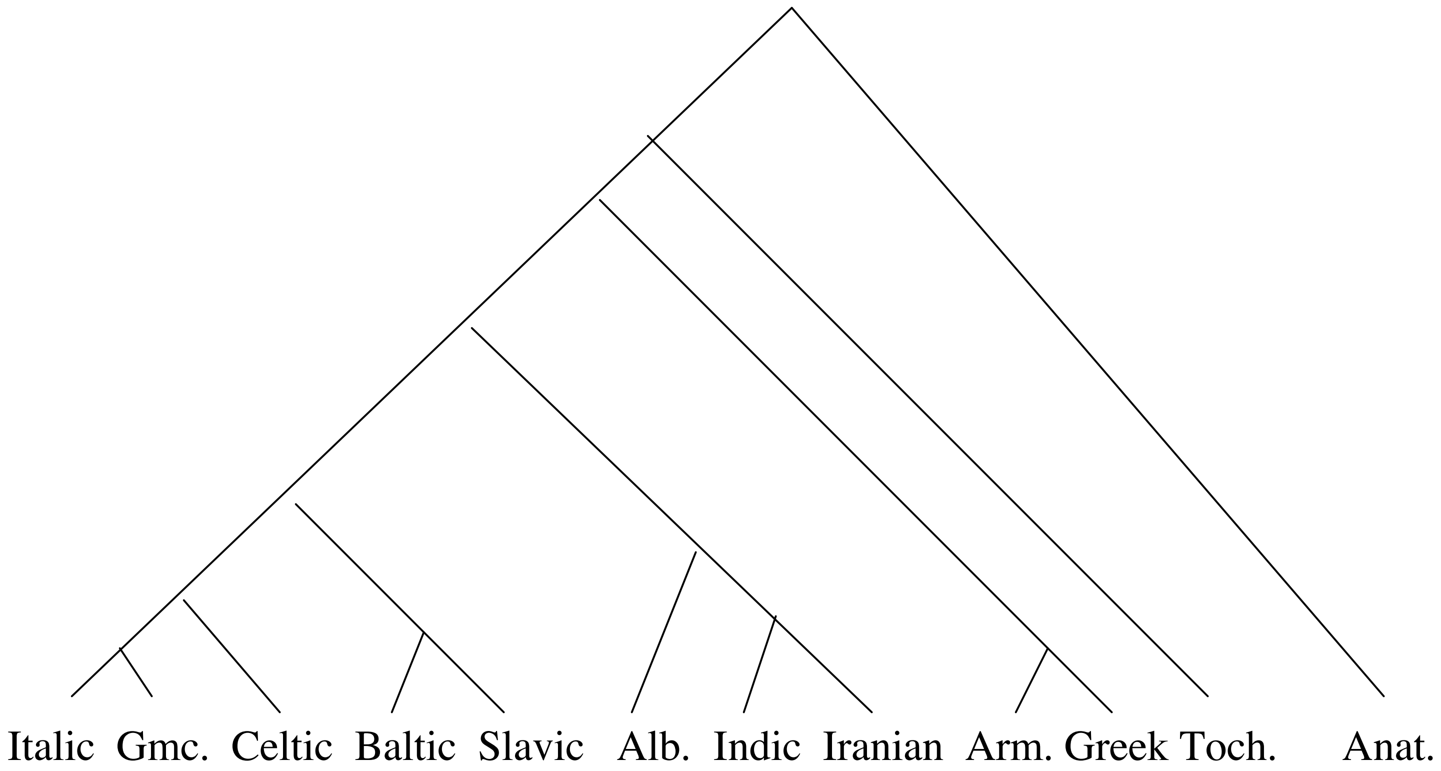
M9a incompatible



Second WMC analysis

- Weighting: all characters either have infinite weight (and are required to be compatible) or have unit weight.
- No change on the screened full dataset
- The tree changes on the unscreened full dataset! Greco-Armenian moves to be the sister group to Indo-Iranian, and the tree has 53 incompatible characters: 2 phonological characters (P2, P3), one morphological character (M9a), and 50 lexical characters.
- The trees obtained using WMC change when restricted to lexical characters alone - Italic and Celtic group with Germanic (as in the analyses obtained using other methods).

Original Gray & Atkinson Tree, Nature 2004)



New Gray & Atkinson analyses

- The trees obtained using G&A are very similar to the G&A Nature tree: Germanic, Italic, and Celtic always group together, the Satem Core is sometimes present, and Albanian moves around.
- G&A trees have about the same number of incompatible characters as WMC trees, but all the trees are incompatible with M5 (the mediopassive marker), and some are incompatible with other morphological and phonological characters (such as P1, P2, P3, M6, M8, M9a, and M11)

General comments

- Including high quality characters (both complex phonological and morphological characters) and giving them high weight has a big impact on the resultant reconstructions.
- Trained IEists will not necessarily agree on the selection of characters and/or their encodings, and so WMC is really best seen as a tool for IEists to explore the phylogenetic implications of their scholarship.

For more information

- Please see **http://www.cs.rice.edu/~nakhleh/CPHL**

  (the Computational Phylogenetics for Historical Linguistics web site) for data and papers
- This talk based upon papers to appear in *Language*, *Transactions of the Philological Society*, and Cambridge University Press

Acknowledgements

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