Challenges in Computational Linguistic Phylogenetics

Tandy Warnow
Department of Computer Science
University of Illinois

Indo-European languages

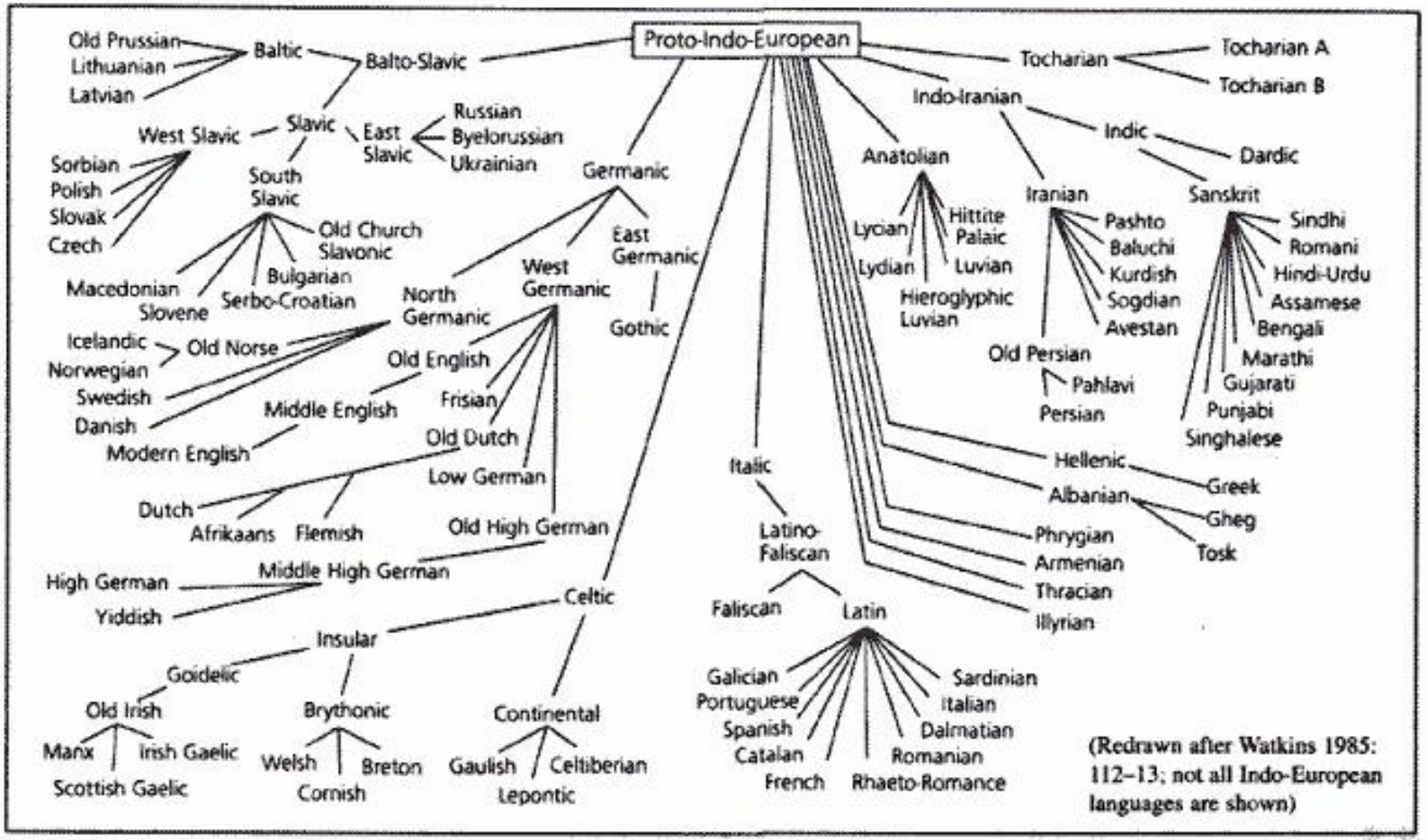


FIGURE 6.1: The Indo-European family tree

From linguistica.tribe.net

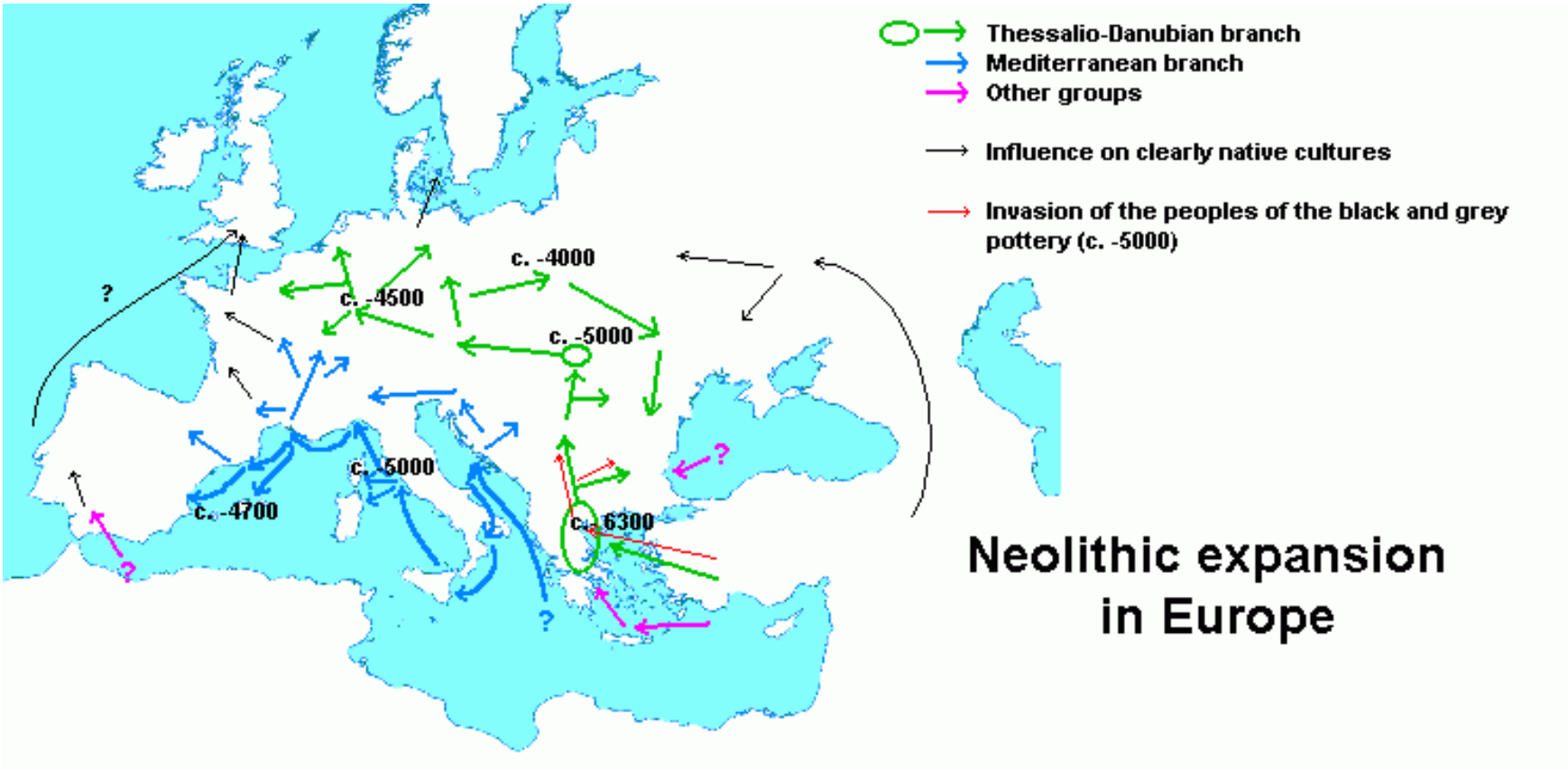
Controversies for IE history

- 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

Other questions about IE

- Where is the IE homeland?
- When did Proto-IE “end”?
- What was life like for the speakers of proto-Indo-European (PIE)?

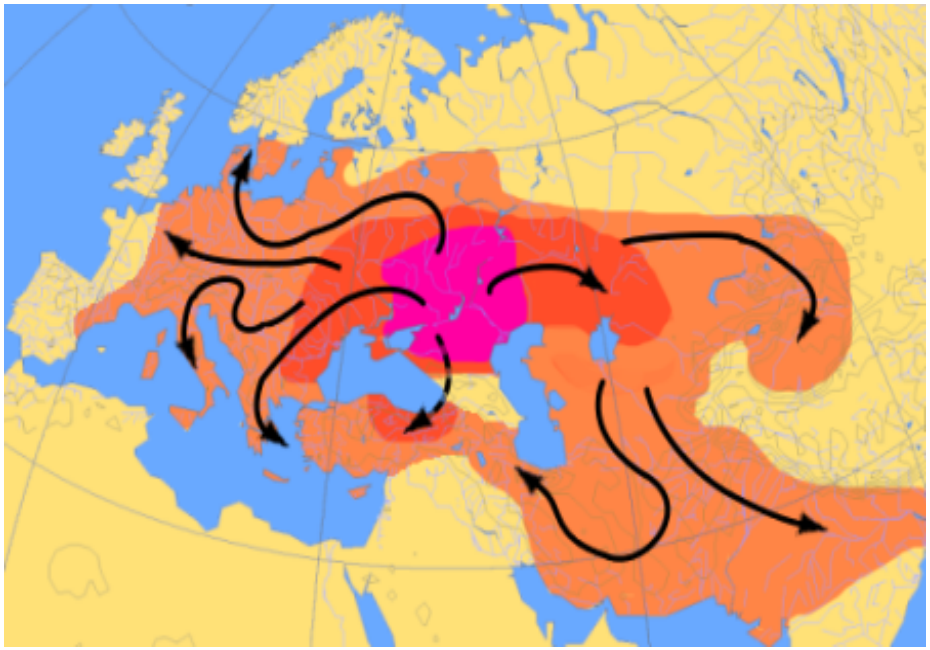
The Anatolian hypothesis (from wikipedia.org)



Date for PIE ~7000 BCE

The Kurgan Expansion

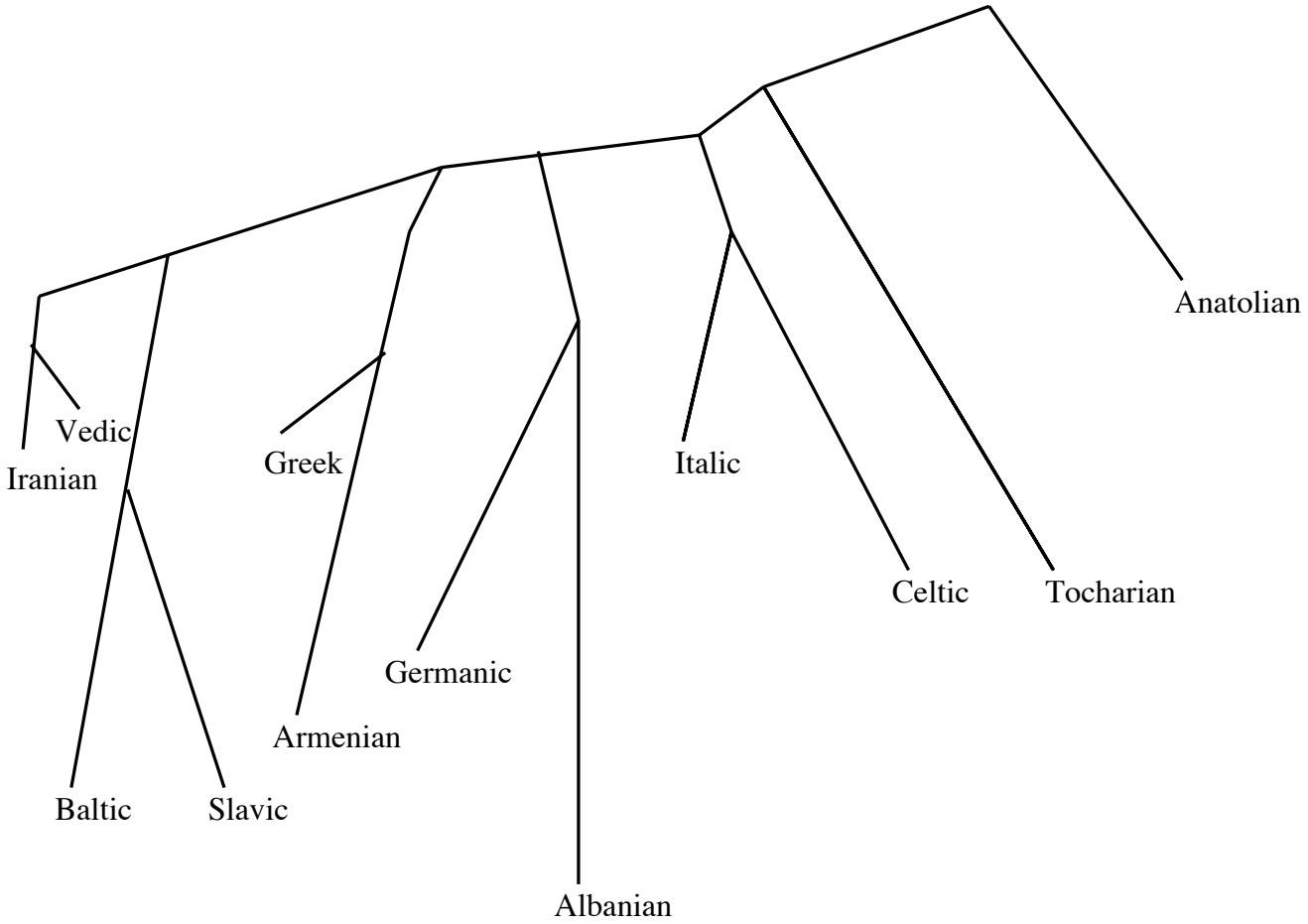
- Date of PIE ~4000 BCE.
- Map of Indo-European migrations from ca. 4000 to 1000 BC according to the Kurgan model
- From http://indo-european.eu/wiki



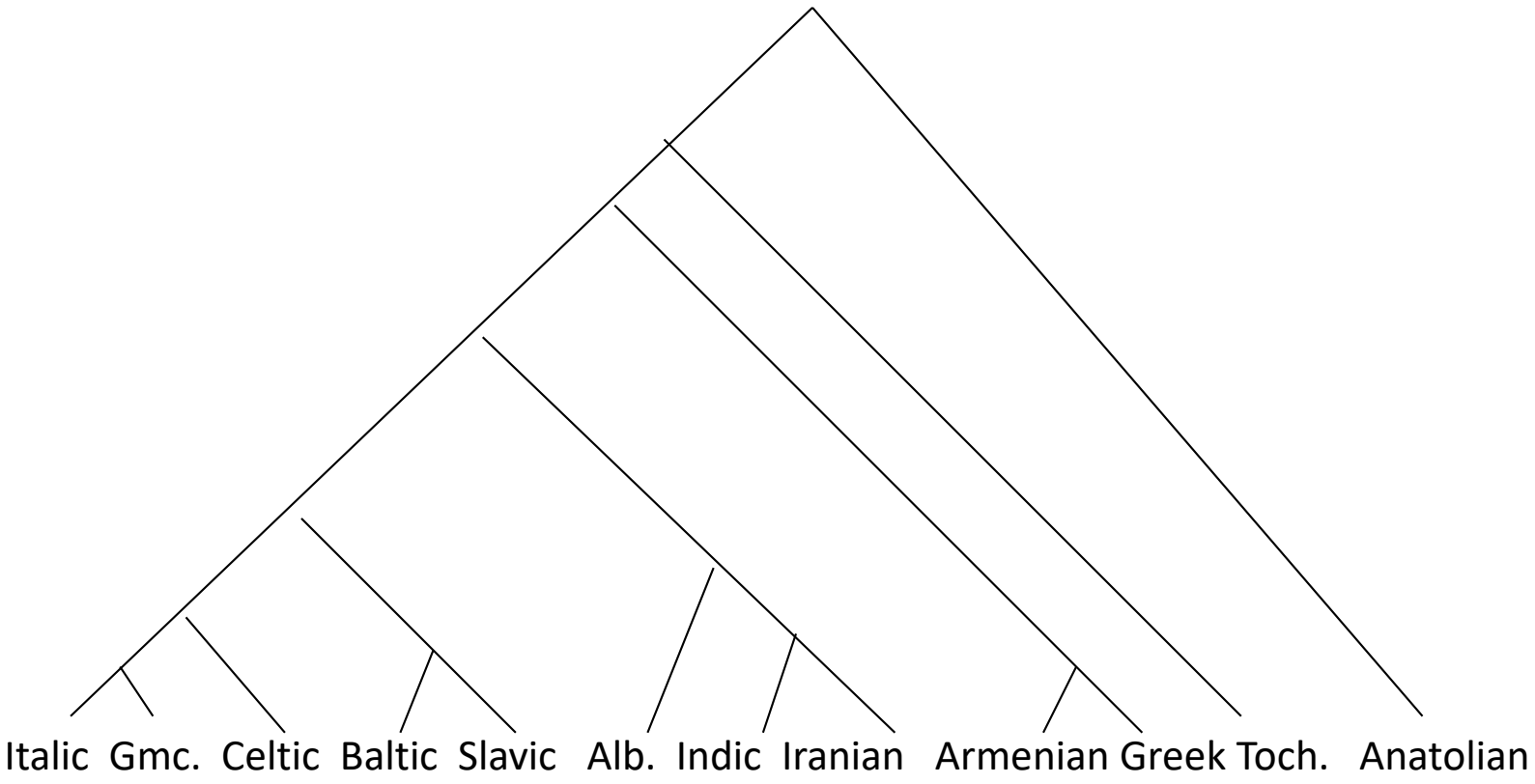
Estimating the date and homeland of the proto-Indo-Europeans

- Step 1: Estimate the phylogeny
- Step 2: Reconstruct words for proto-Indo-European (and for intermediate proto-languages)
- Step 3: Use archaeological evidence to constrain dates and geographic locations of the proto-languages

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



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



This talk

- Linguistic data and the Ringe-Warnow analyses of the Indo-European language family
- Comparison of different phylogenetic methods on Indo-European datasets (Nakhleh et al., Transactions of the Philological Society 2005)
- Perfect Phylogenetic Networks (Nakhleh et al., Language 2005)
- Simulation study evaluating different phylogenetic methods (Barbançon et al., Diachronica 2013)
- Discussion and Future work

The Computational Historical Linguistics Project

Collaboration with Don Ringe began in 1994; 17 papers since then, and two NSF grants.

Dataset generation by Ringe and Ann Taylor (then a postdoc with Ringe, now Senior Lecturer at York University).

Method development with Luay Nakhleh (then my student, now Associate Professor at Rice University), Steve Evans (Prof. Statistics, Berkeley). Simulation study with Francois Barbanson (then my postdoc).

Ongoing work in IE with Ringe.

Don Ringe

http://web.engr.illinois.edu/~warnow/histling.html

Indo-European languages

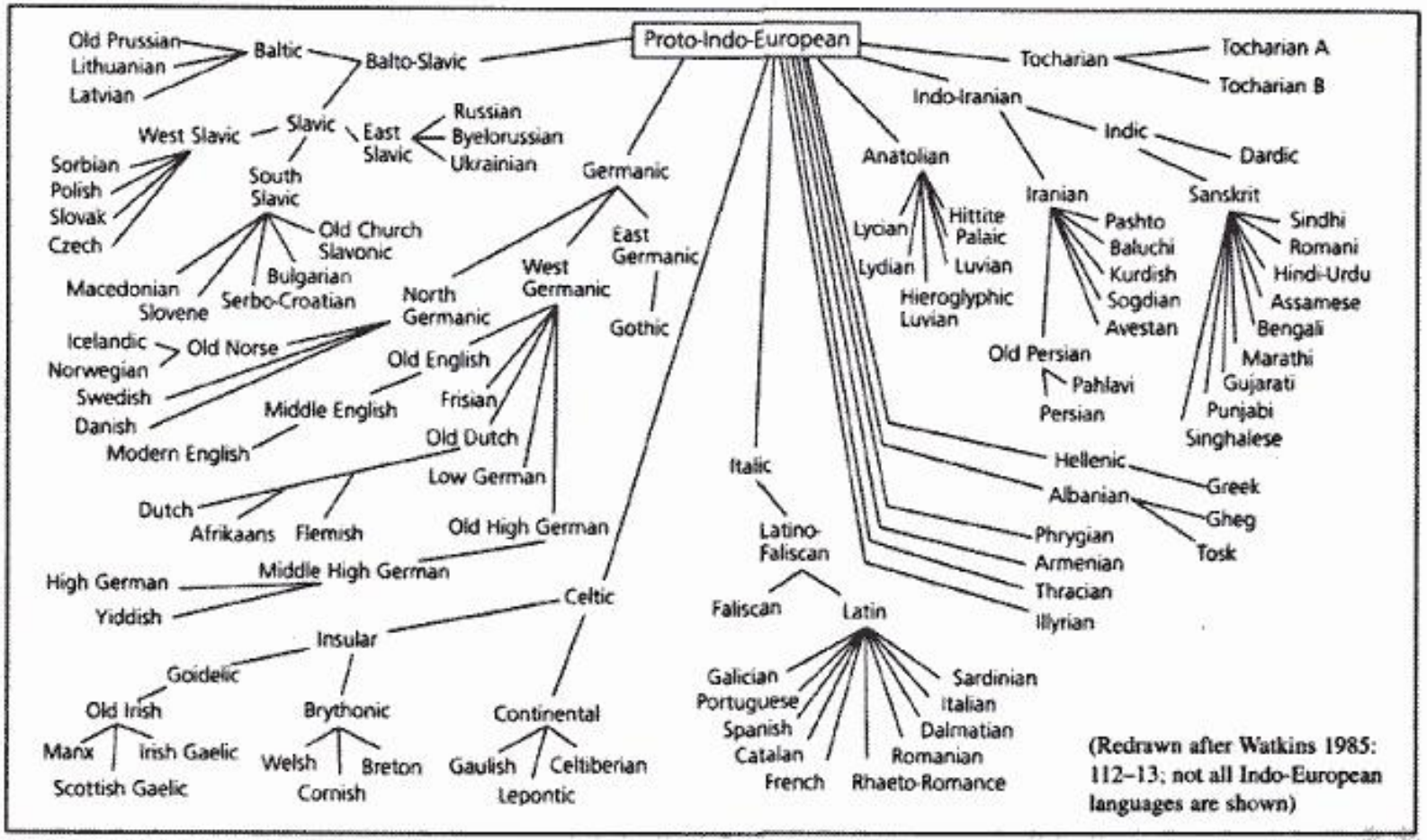


FIGURE 6.1: The Indo-European family tree

From linguistica.tribe.net

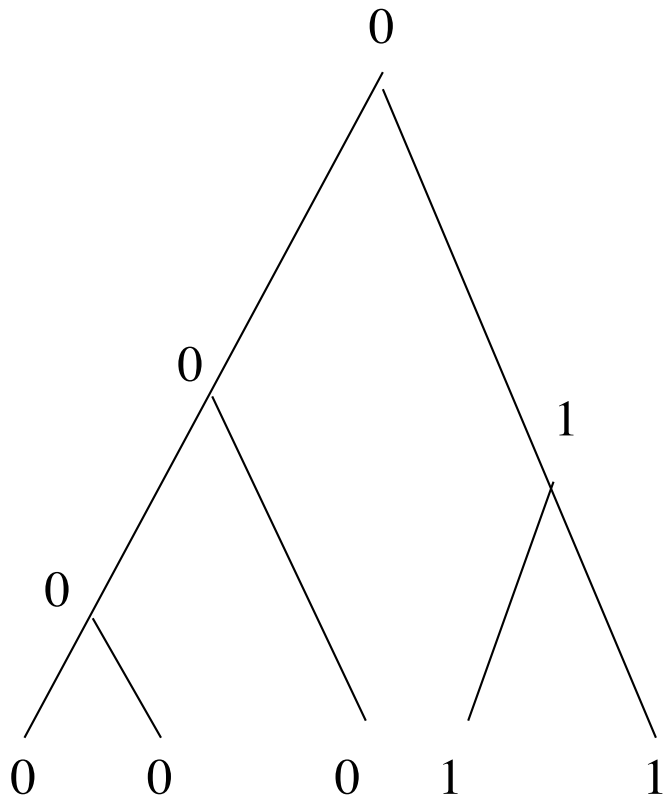
Historical Linguistic Data

- 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; i.e., there is no homoplasy (character reversal or parallel evolution)
- First inferred for weird innovations in phonological characters and morphological characters in the 19th century, and used to establish all the major subgroups within IE



Sound changes

- Many sound changes are natural, and should not be used for phylogenetic reconstruction.
- Others are bizarre, or are composed of a sequence of simple sound changes. These are useful for subgrouping purposes. Example: Grimm's Law.
    1. Proto-Indo-European voiceless stops change into voiceless fricatives.
    2. Proto-Indo-European voiced stops become voiceless stops.
    3. Proto-Indo-European voiced aspirated stops become voiced fricatives.

An Indo-European lexical character: 'hand'.

Data.

Hittite	kissar	Lithuanian	rankà	Old Prussian	rānkan (acc.)
Armenian	jeŕn	Old English	hand	Latvian	ròka
Greek	χείρ /kʰé:r/	Old Irish	lám	Gothic	handus
Albanian	dorë	Latin	manus	Old Norse	hǫnd
Tocharian B	ṣar	Luvian	īssaris	OHG	hant
Vedic	hástas	Lycian	izredi (instr.)	Welsh	llaw
Avestan	zastō	Tocharian A	tsar	Oscan	manim (acc.)
OCS	rǫka	Old Persian	dasta	Umbrian	manf (acc. pl.)

Semantic slot for hand – coded (Partitioned into cognate classes)

Coding.

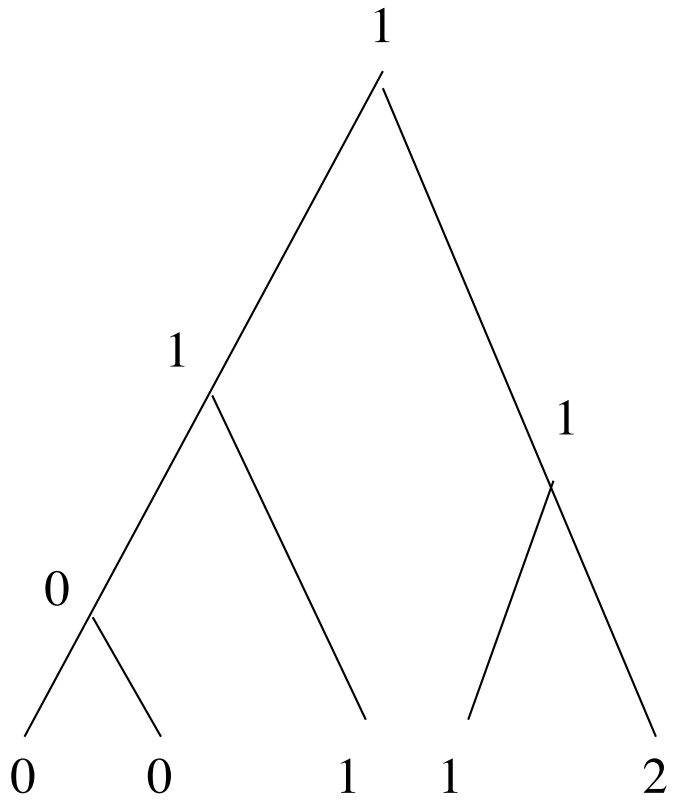
Hittite	1	Lithuanian	2	Old Prussian	2
Armenian	1	Old English	3	Latvian	2
Greek	1	Old Irish	4	Gothic	3
Albanian	1	Latin	5	Old Norse	3
Tocharian B	1	Luvian	1	OHG	3
Vedic	1a	Lycian	1	Welsh	4
Avestan	1a	Tocharian A	1	Oscan	5
OCS	2	Old Persian	1a	Umbrian	5

Justification of coding.
Note that ">" means "developed by regular sound change into"; this is important, because developments by regular sound change are mathematically demonstrable. On the other hand, "→" means "developed by process(es) other than regular sound change"; a hypothesis of such a development is not mathematically demonstrable, but it can be highly probable, since many changes are of known types with dozens of well-understood examples.

Proto-Indo-European *pĺ̥h$_2$meh$_2$ 'flat hand' (cf. Homeric Greek *paláme:*)
> Proto-Celtic *lāmā 'hand'
> Old Irish *lám*
> Welch *llaw*
Proto-Germanic *handuz 'hand'
> Gothic *handus*
>→ Runic Norse *handu (ending influenced by a different noun class)
> Old Norse *hǫnd*
> Proto-West Germanic *handu
> Old English *hand*
> Old High German *hant*
Proto-Italic *man- 'hand'
> Latin *manus* (transferred into the u-stems)
>→ Proto-Sabellian *man-
> Oscan *manis > *mans, accusative *manim* (transf. into the i-stems)
> Umbrian *man-, accusative plural *manf*

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.*



Our (RWT) Data

- Ringe & Taylor (2002)
  - 259 lexical
  - 13 morphological
  - 22 phonological
- These data have cognate judgments estimated by Ringe and Taylor, and vetted by other Indo-Europeanists. (Alternate encodings were tested, and mostly did not change the reconstruction.)
- Polymorphic characters, and characters known to evolve in parallel, were removed.

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.

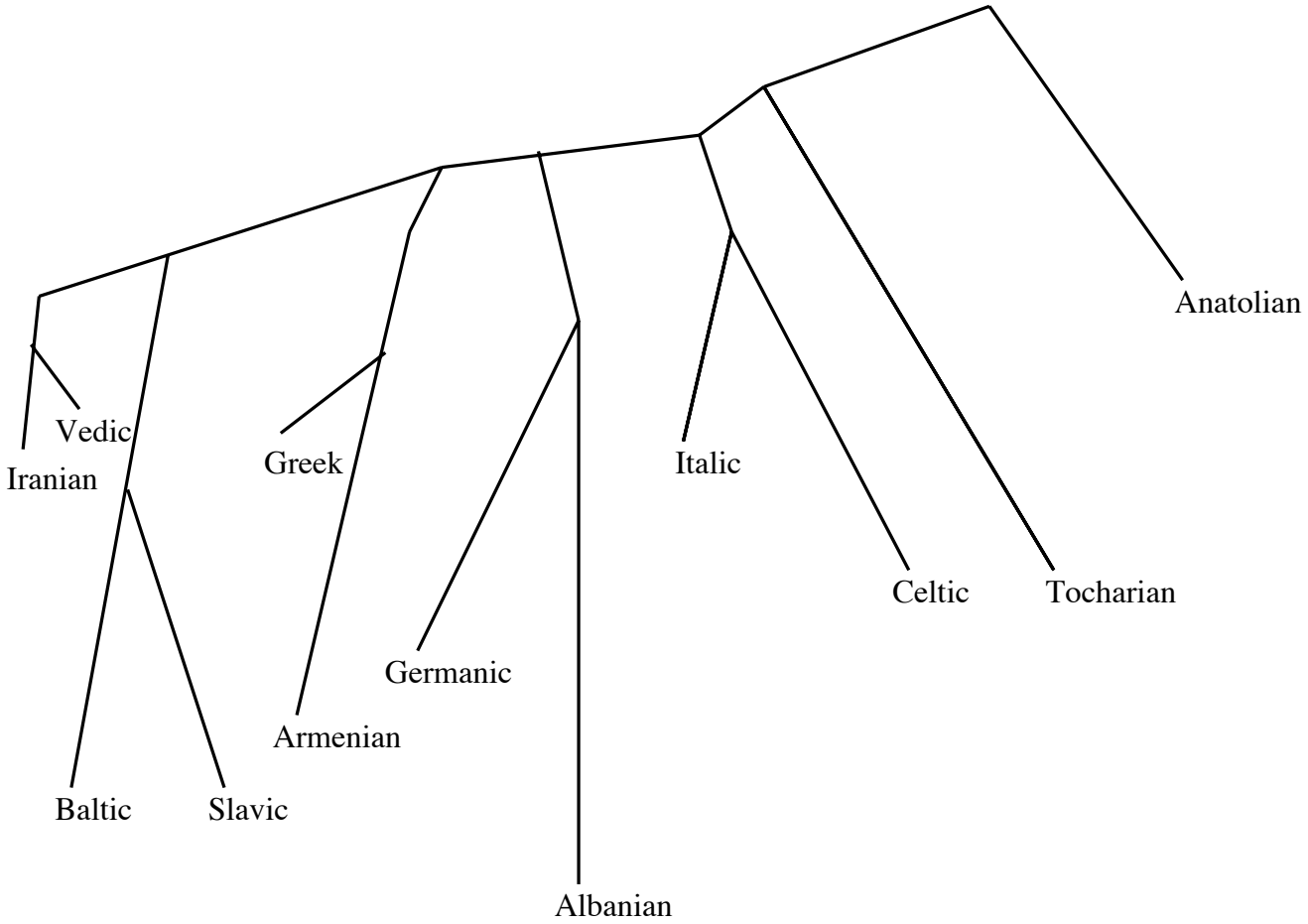
Our methods/models

- 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 (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 (Cambridge University Press)

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.)

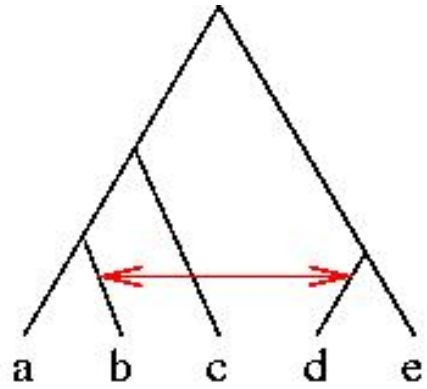
Indo-European Tree
(95% of the characters compatible)

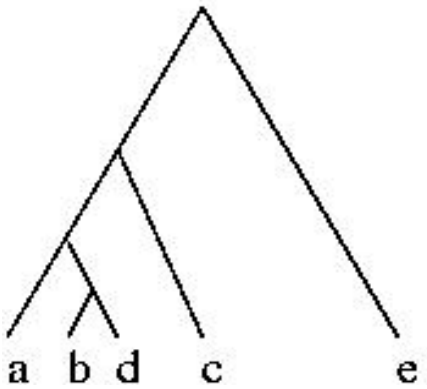


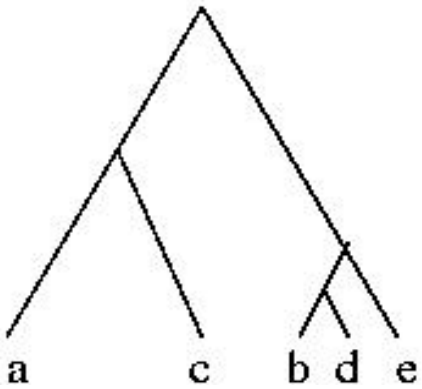
Second attempt: PPN

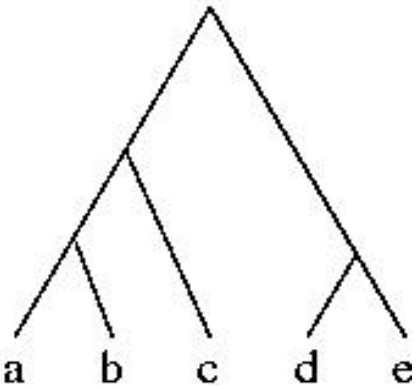
- We explain the remaining incompatible characters by inferring previously ***undetected*** "borrowing".
- We attempted to find a PPN (perfect phylogenetic network) with the smallest number of contact edges, borrowing events, and with maximal feasibility with respect to the historical record. (Computational 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.

Modelling borrowing: Networks and Trees within Networks









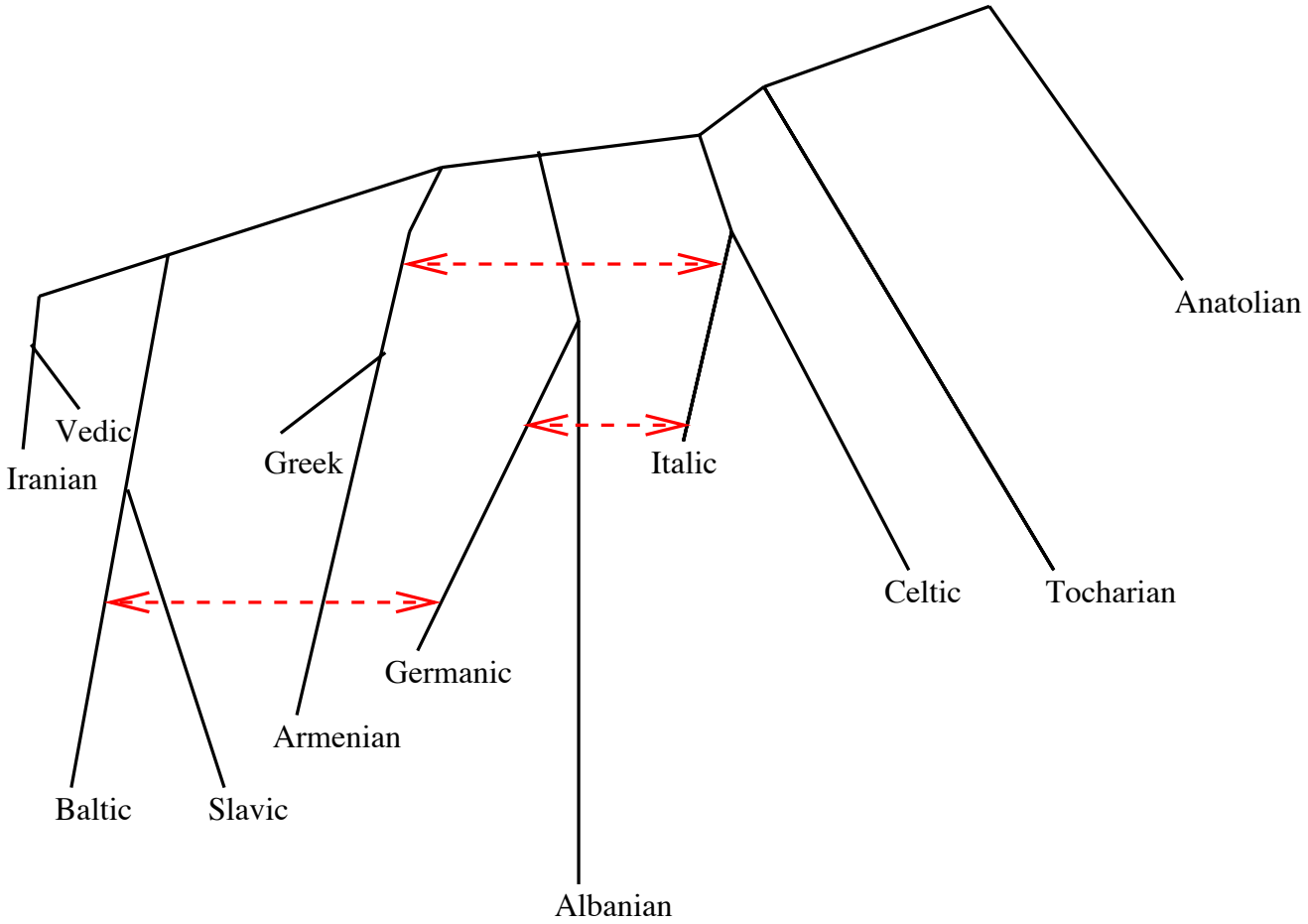
Perfect Phylogenetic Networks

Problem formulation

- Input: set of languages described by characters
- Output: Network on which all characters evolve without homoplasy, but can be borrowed

Nakhleh, Ringe, and Warnow, 2005. Language.

“Perfect Phylogenetic Network” (all characters compatible)



L. Nakhleh, D. Ringe, and T. Warnow, LANGUAGE, 2005

Comments

- This network is very “tree-like” (only three contact edges needed to explain the data.
- Two of the three contact edges are strongly supported by the data (many characters are borrowed).
- If the third contact edge is removed, then the evolution of the remaining (two) incompatible characters needs to be explained. *Probably this is parallel semantic shift.*

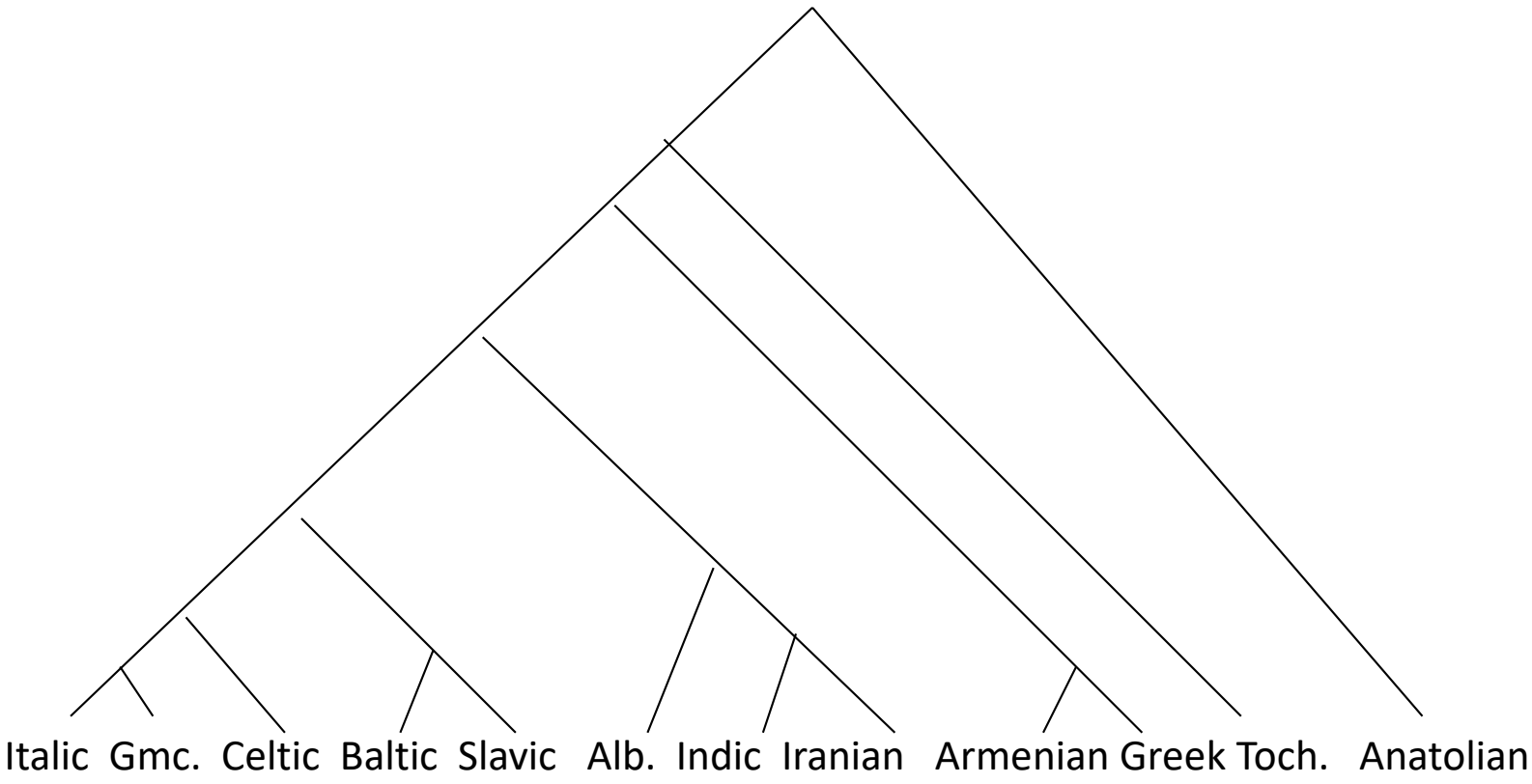
Other IE analyses

Note: many reconstructions of IE have been done, but produce different histories which differ in significant ways

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

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



Based only on lexical characters – with “binary encoding”

The performance of methods on an IE data set,
Transactions of the Philological Society,
L. Nakhleh, T. Warnow, D. Ringe, and S.N. Evans, 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, we use a ***better basic dataset*** (where cognancy judgments are more reliable).

Phylogeny reconstruction methods

- Perfect Phylogenetic Networks (Ringe, Warnow, and Nakhleh)
- Other network methods
- Neighbor joining (distance-based method)
- UPGMA (distance-based method, same as glottochronology)
- Maximum parsimony (minimize number of changes)
- Maximum compatibility (weighted and unweighted)
- Gray and Atkinson (Bayesian estimation based upon presence/absence of cognates, as described in Nature 2003)

Phylogeny reconstruction methods

- Perfect Phylogenetic Networks (Ringe, Warnow, and Nakhleh)
- Other network methods
- Neighbor joining (distance-based method)
- UPGMA (distance-based method, same as glottochronology)
- Maximum parsimony (minimize number of changes)
- Maximum compatibility (weighted and unweighted)
- Gray and Atkinson (Bayesian estimation based upon presence/absence of cognates, as described in Nature 2003)

IE Languages used in the study

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
Latin	LA	Old Church Slavonic	OC
Oscan	OS	Old Prussian	PR
Umbrian	UM	Lithuanian	LI
Gothic	GO	Latvian	LT
Old Norse	ON	Albanian	AL

Four IE 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.

Results: Likely Subgroups

Other than UPGMA, all methods reconstruct

- the ten major subgroups
- **Anatolian + Tocharian** (that under the assumption that Anatolian is the first daughter, then Tocharian is the second daughter)
- **Greco-Armenian** (that Greek and Armenian are sisters)

Nothing else is consistently reconstructed.

In particular, the choice of data (lexical only, or also morphology and phonological) has an impact on the final tree.

The choice of method also has an impact!

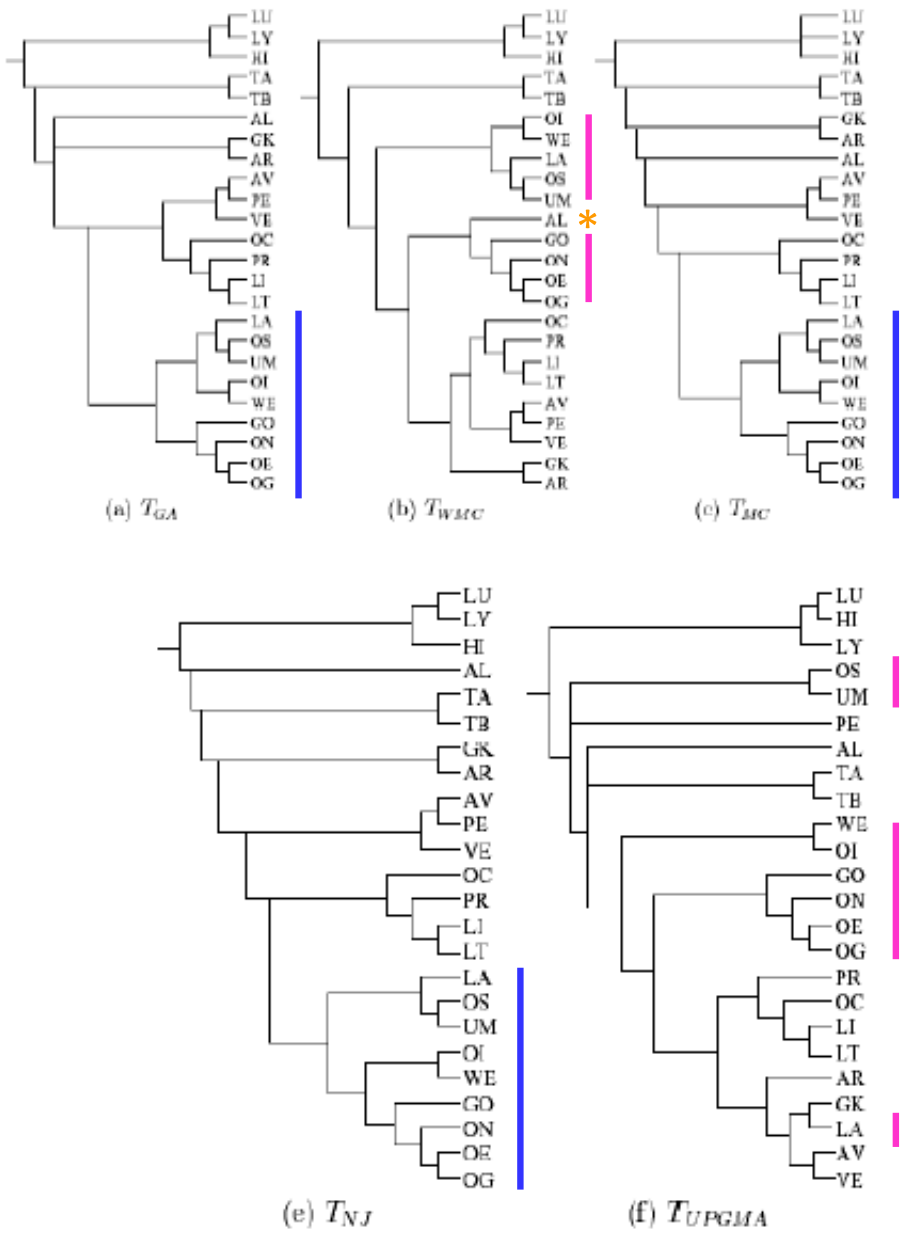


Figure 1. Five trees inferred on the screened full dataset

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.

Other observations

- UPGMA (i.e., the tree-building technique for glottochronology) does the worst (e.g. splits Italic and Iranian groups).
- 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 highly weighted morphological characters.

Different methods/data
give different answers.

We don’t know
which answer is correct.

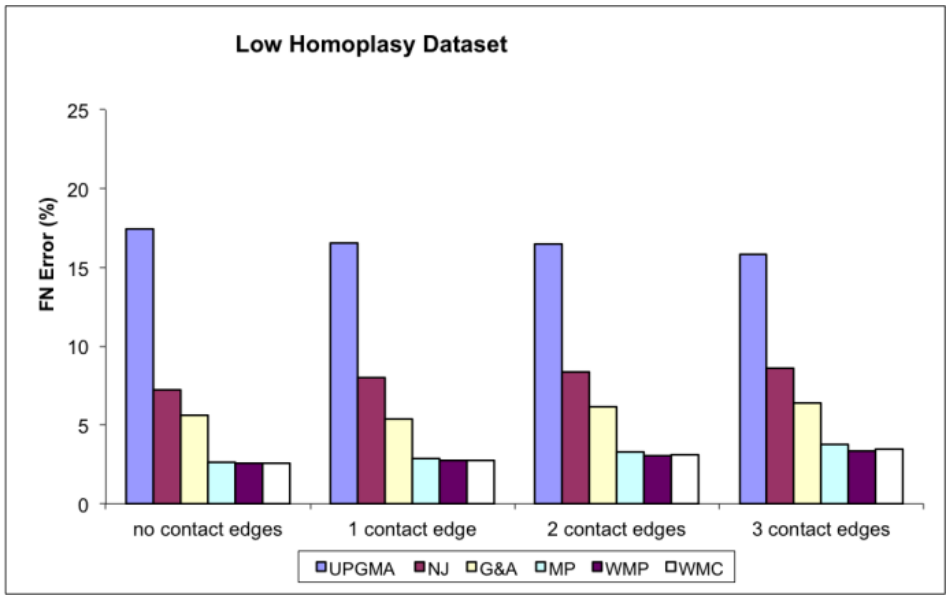
Which method(s)/data
should we use?

F. Barbancon, S.N. Evans, L. Nakhleh, D. Ringe, and T. Warnow, Diachronica 2013

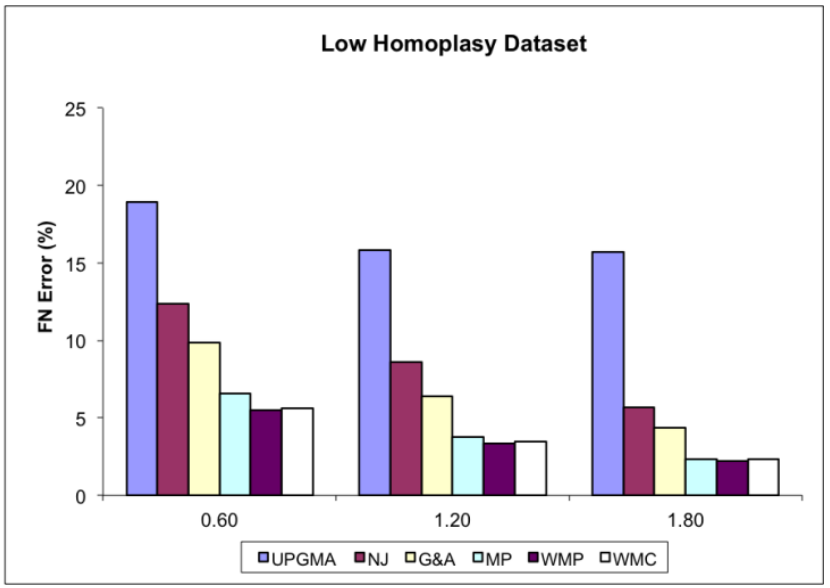
Simulation study based on stochastic model of language evolution (Warnow, Evans, Ringe, and Nakhleh, Cambridge University Press 2004)

- Lexical and morphological characters
- Networks with 1-3 contact edges, and also trees
- "Moderate homoplasy":
  - morphology: 24% homoplastic, no borrowing
  - lexical: 13% homoplastic, 7% borrowing
- "Low homoplasy":
  - morphology: no borrowing, no homoplasy;
  - lexical: 1% homoplastic, 6% borrowing

Simulation study – sample of results



Varying number of contact edges



Varying deviation from i.i.d. character evolution

Observations

1. Choice of data does matter (good idea to add morphological characters, and to screen well).
2. Accuracy only slightly lessened with small increases in homoplasy, borrowing, or deviation from the lexical clock. Some amount of heterotachy (deviation from i.i.d.) improves accuracy.
3. Relative performance between methods consistently shows:

    Distance-based methods least accurate

    Gray and Atkinson's method middle accuracy

    Parsimony and Compatibility methods most accurate

Critique of the Gray and Atkinson model

- Gray and Atkinson's model is for binary characters (presence/absence), not for multi-state characters.
- To use their method on multi-state data, they do a "binary encoding" – and so treat a single cognate class as a separate character, and all cognate classes for a single semantic slot are assumed to evolve identically and independently.
- This assumption is clearly violated by how languages evolve.

Critique of the Gray and Atkinson model

- Gray and Atkinson's model is for binary characters (presence/absence), not for multi-state characters.
- To use their method on multi-state data, they do a "binary encoding" – and so treat a single cognate class as a separate character, and all cognate classes for a single semantic slot are assumed to evolve identically and independently.
- This assumption is clearly violated by how languages evolve.
- Note: no rigorous biologist would perform the equivalent treatment on biological data. So this is not about linguistics vs. biology.

Estimating the date and homeland of the proto-Indo-Europeans

- Step 1: Estimate the phylogeny
- Step 2: Reconstruct words for proto-Indo-European (and for intermediate proto-languages)
- Step 3: Use archaeological evidence to constrain dates and geographic locations of the proto-languages

Implications regarding PIE homeland and date

- Linguists have “reconstructed” words for ‘wool’, ‘horse’, ‘thill’ (harness pole), and ‘yoke’, for Proto-Indo-European, for ‘wheel’ for the ancestor of IE minus Anatolian, and for `axle" to the ancestor of IE minus Anatolian and Tocharian.
- Archaeological evidence (positive and negative) for these objects used to constrain the date and location for proto-IE to be *after* the “secondary products revolution”, and somewhere with horses (wild or domesticated).
- Combination of evidence supports the date for PIE within 3000-5500 BCE (some would say 3500-4500 BCE), and location *not* Anatolia, thus ruling out the Anatolian hypothesis.

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 carefully tested.

Future research

- We need more investigation of statistical methods based on good stochastic models, as these are now the methods of choice in biology.
- This requires *realistic parametric models of linguistic evolution* and *method development under these parametric models*!

Modelling issues

- What are the units?
- Polymorphism
- Homoplasy
- Non-treelike evolution
- Non-*i.i.d.* evolution and violations of the rates-across-sites assumption (heterotachy)
- Deviation from the lexical clock (is dating even really possible?)

Note:

- *The statistical model of Warnow, Evans, Ringe, and Nakhleh has homoplasy and reticulation, but no polymorphism.*
- *The bag-of-words model of Nicholls and Gray (J R. Statist. Soc.B 2008) allows for polymorphism but no reticulation, and has homoplasy in the form of parallel back-muation.*
- *The model of Gray and Atkinson (binary-encoding) has unlimited polymorphism and homoplasy, but no reticulation.*

Acknowledgements

- Financial Support: The David and Lucile Packard Foundation, The National Science Foundation, The Program for Evolutionary Dynamics at Harvard, and The Radcliffe Institute for Advanced Studies

- Collaborators: Don Ringe, Steve Evans, Luay Nakhleh, and Francois Barbançon

- Please see **http://tandy.cs.illinois.edu/histling.html**