Phylogenetic Networks of Languages

Tandy Warnow
Department of Computer Science

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).

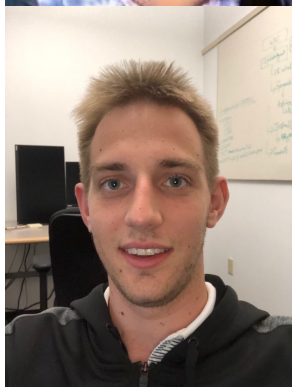
Method development with Luay Nakhleh (then my student, now Dean of Engineering at Rice University), Steve Evans (Prof. Statistics, Berkeley). Simulation study with Francois Barbançon (then my postdoc).

New work with Marc Canby (PhD student of Julia Hockenmaier at UIUC).

Ongoing work in IE with Ringe.

http://tandy.cs.illinois.edu/histling.html

Don Ringe



Indo-European languages

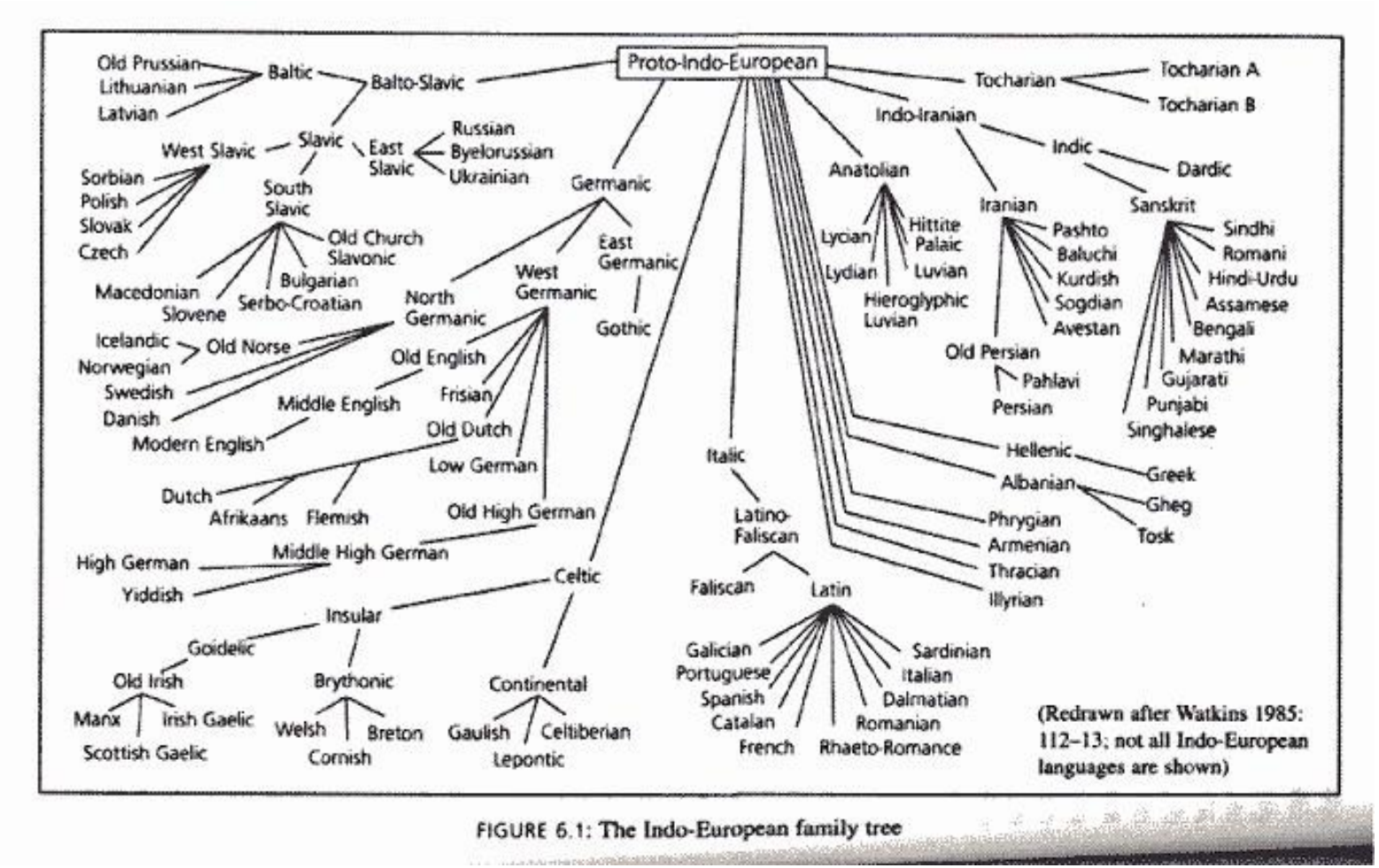


FIGURE 6.1: The Indo-European family tree

From linguistica.tribe.net

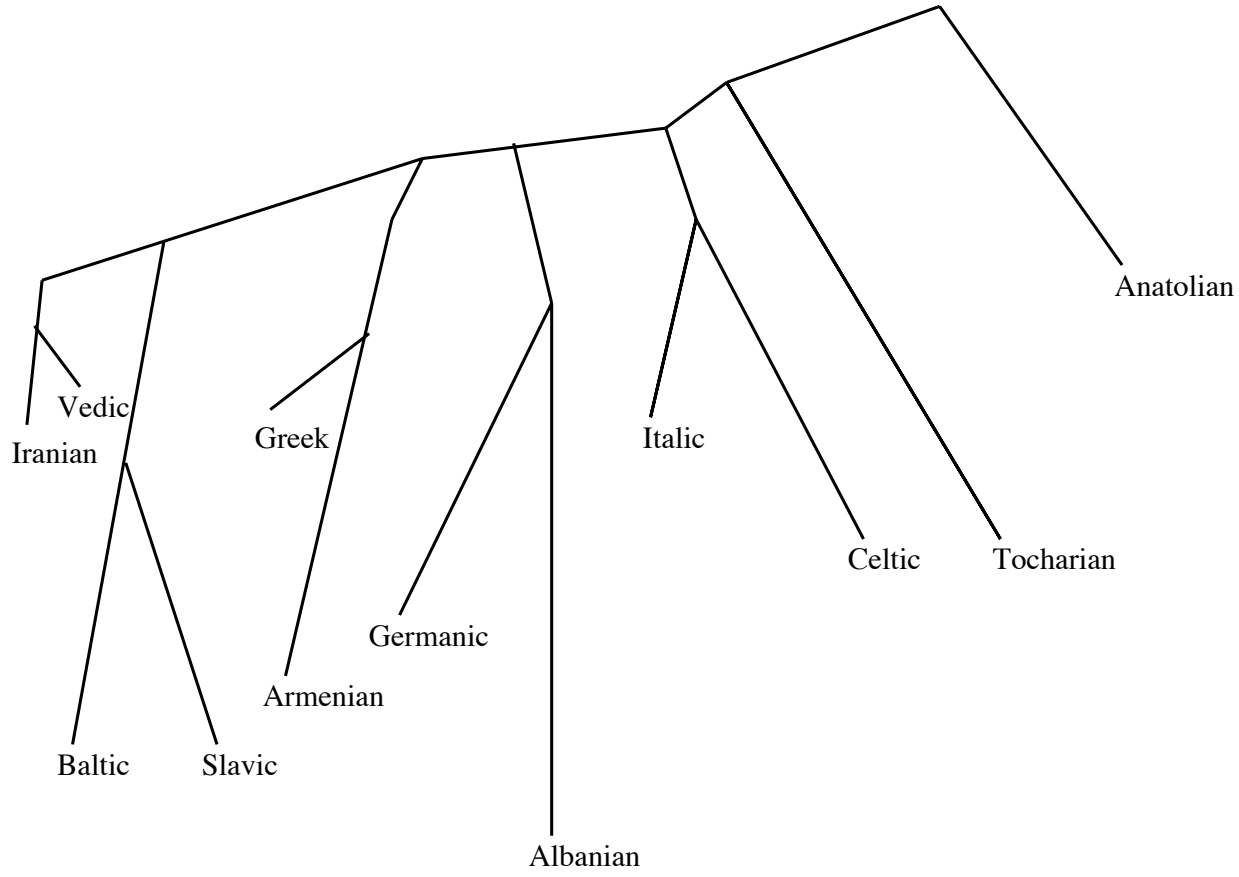
Controversies for IE history

- 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
- What is the homeland of the Indo-Europeans?

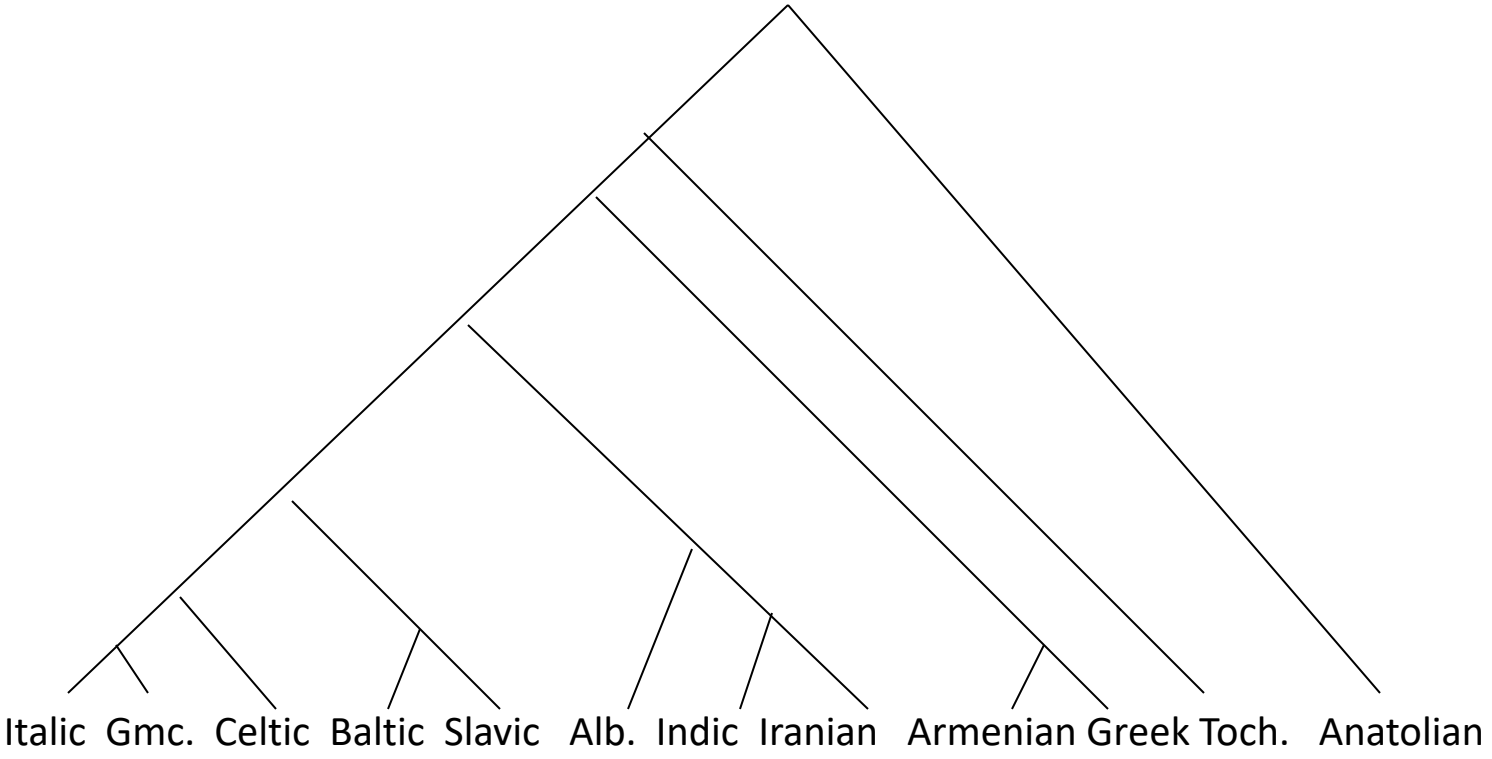
Estimating the date and homeland of the proto-Indo-Europeans

- 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

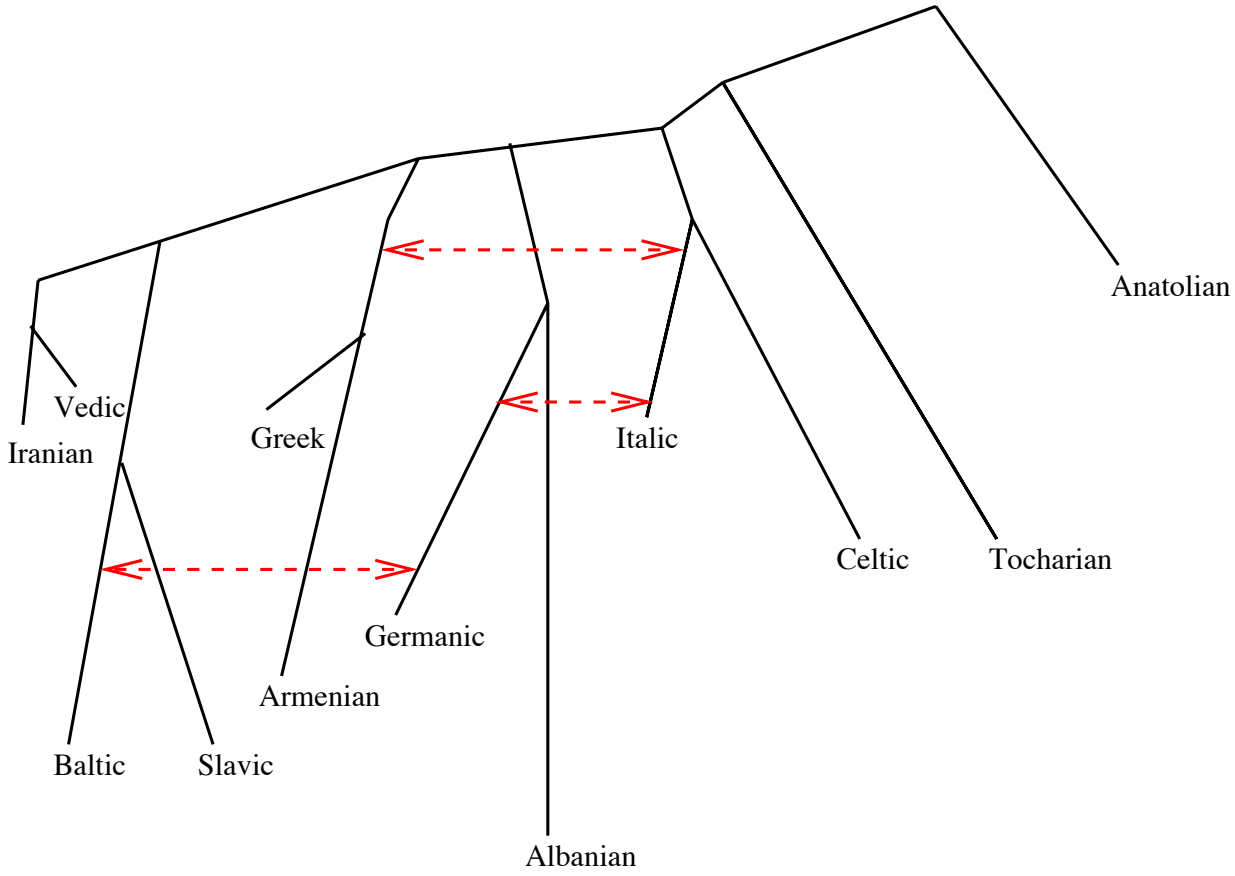
Possible Indo-European tree (Ringe, Warnow and Taylor 2000)



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



“Perfect Phylogenetic Network” (all characters compatible)



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

Preview

- Indo-European Phylogeny (tree or network) is highly debated, but relevant to research in archaeology, anthropology, and early origins of humans
- Phylogenetic networks are needed for modeling evolution, both in linguistics and in biology.
- Methods for biological network estimation are mainly likelihood-based and very computationally intensive.
- Discrete methods, such as the ones described, could advance discovery.
- BUT – very little theory so far establishing identifiability of phylogenetic networks, and statistically consistent methods are (so far) limited (and do not have any performance guarantees on finite data, especially if not generated by the model)
- New methods and new theory are both needed.

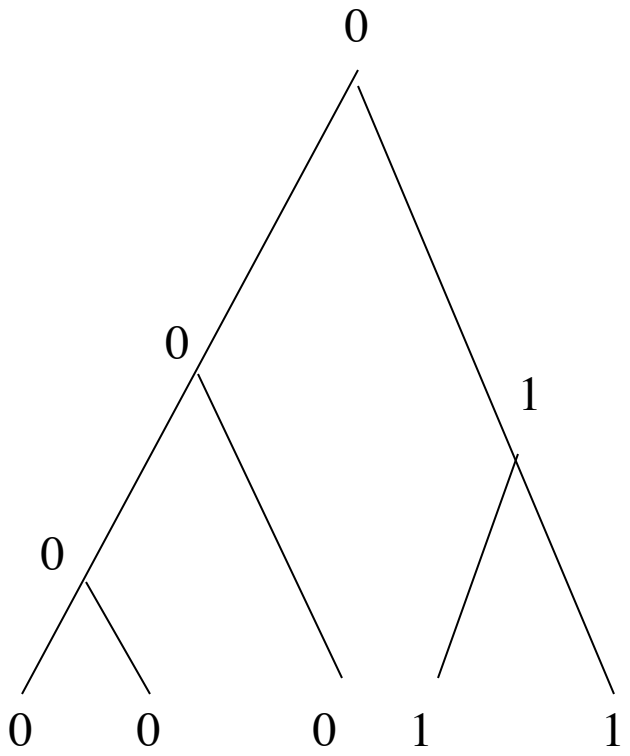
Historical Linguistic Data

- 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 (cognate classes, based on meanings from 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



Our methods/models

- 1995: Ringe & Warnow "Almost Perfect Phylogeny (APP)": most characters evolve without homoplasy under a no-common-mechanism assumption (various publications since 1995)
- 2005: Ringe, Warnow, & Nakhleh "Perfect Phylogenetic Network (PPN)": extends APP model to allow for borrowing, but assumes homoplasy-free evolution for all characters (Language, 2005)
- 2006: 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)

Our methods/models

- 1995: Ringe & Warnow "Almost Perfect Phylogeny (APP)": most characters evolve without homoplasy under a no-common-mechanism assumption (various publications since 1995)
- 2005: Ringe, Warnow, & Nakhleh "Perfect Phylogenetic Network (PPN)": extends APP model to allow for borrowing, but assumes homoplasy-free evolution for all characters (Language, 2005)
- 2006: 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)

WERN 2006 theorem and conjecture

- In WERN 2006, we proved that the phylogenetic tree was identifiable from the probability distribution on character states.

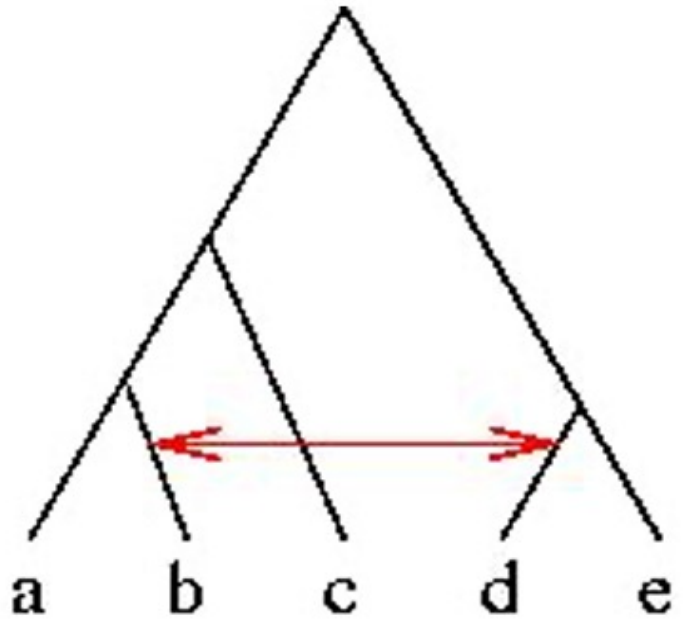
- We also conjectured that the phylogenetic network might be identifiable under some conditions (e.g., if there was only one borrowing edge).

- However, we did not provide any proof, and we did not make any progress on this.

This talk

- Introduction to Phylogenetic Networks, including Gambette et al., 2012 algorithm for reconstructing level-1 networks from quartet trees
- Models of evolution for linguistic characters (published)
- Identifiability of the Linguistic "Genetic" Tree (published)
- Identifiability of the Linguistic Phylogenetic Network Topology (unpublished)
- Open problems for phylogenetic network estimation

Tree-based network



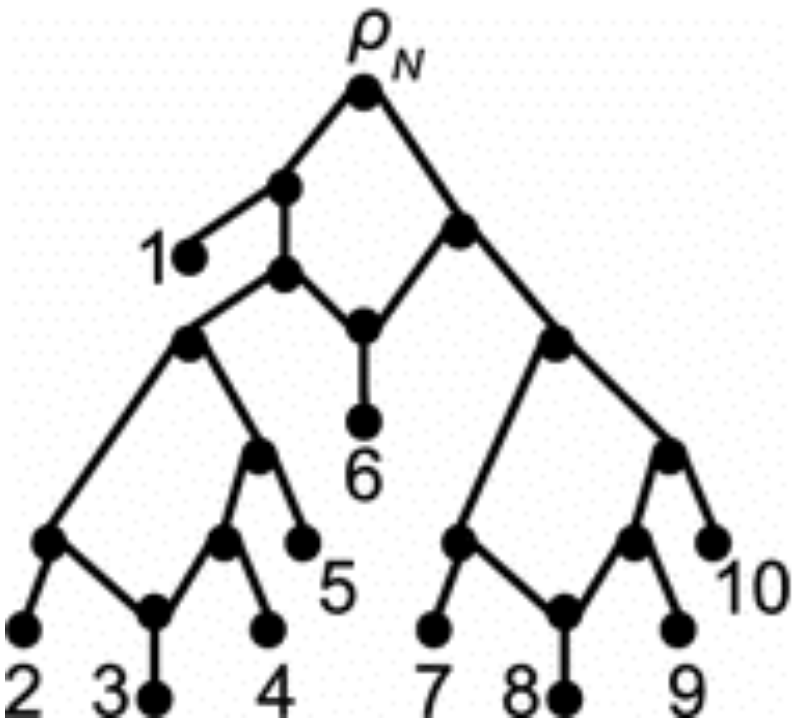
The network is rooted.

There is an underlying tree, on top of which there are extra edges.

This network has only one extra edge.

For linguistics, think of the red edge as indicating borrowing between two language communities.

Hybridization networks



Note: the network is rooted – imagine directing edges away from the root.

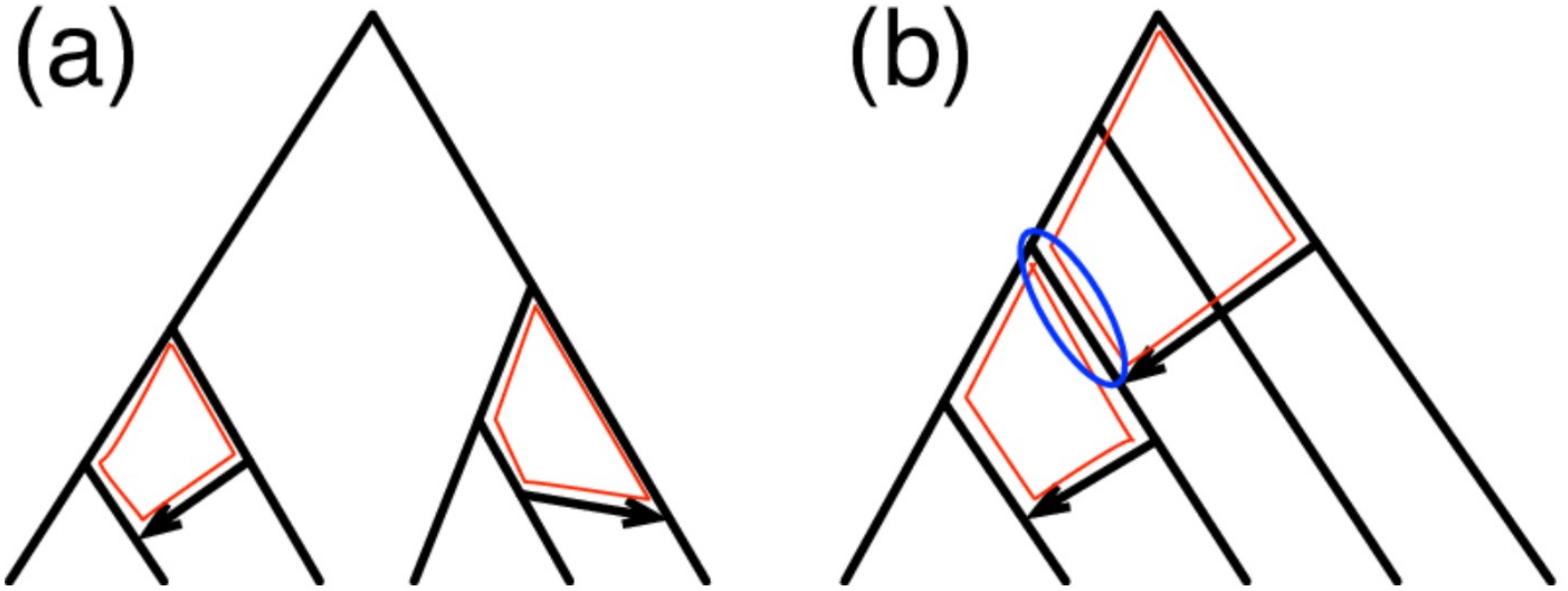
Every node (other than the root) has indegree that is either 1 or 2.

A node with indegree 2 is a hybridization node.

Note the cycles!
In this network, the cycles are node-disjoint.

From: Gambette,doi: 0.1007/s00285-016-1068-3

Level-1 networks (and others)

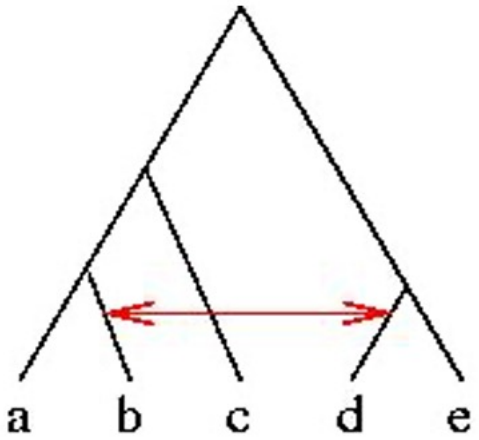


Level-1 network definition. Reticulation nodes induce cycles in the (undirected graphs underlying the) phylogenetic networks. The edges of the cycles are highlighted with red lines. (a) A level-1 network is one where no edge of the network is shared by two or more cycles. (b) A non-level-1 network is one where at least one edge is shared by at least two cycles (the shared edge in this case is the one inside the blue circle).

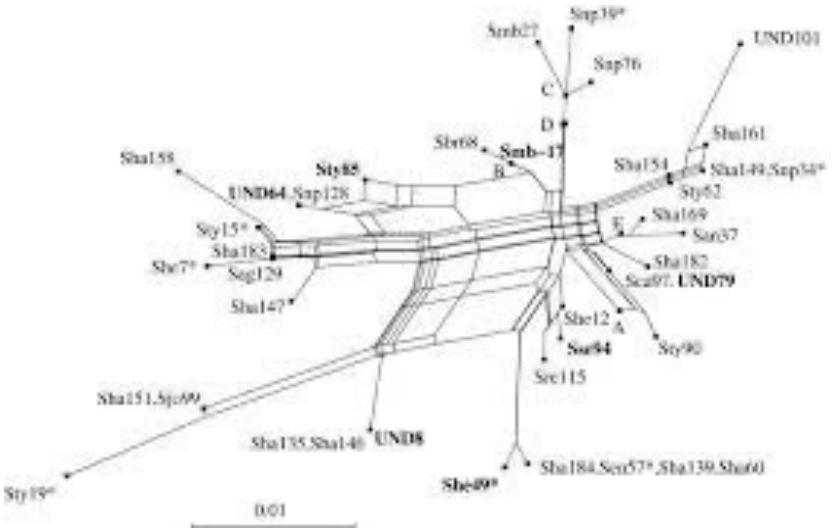
From: Elworth, RA Leo, et al. arXiv arXiv:1808.08662.

Different kinds of networks

- Explicit networks

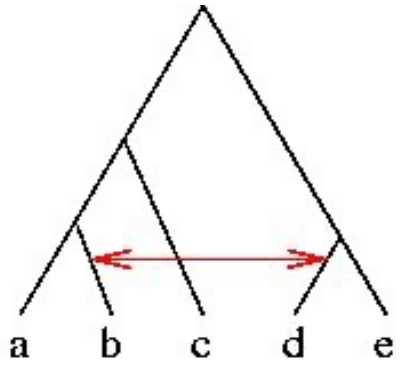


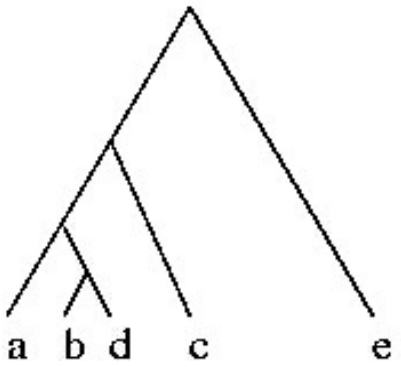
- Implicit networks

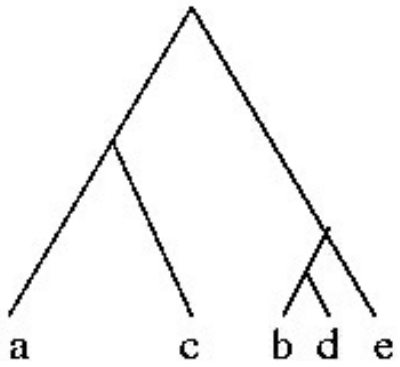


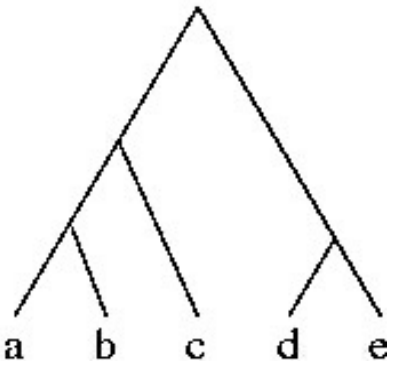
From Mol. Biol. Evol. 21(2):255–265. 2004
DOI: 10.1093/molbev/msh018

Phylogenetic Networks and the Trees they contain



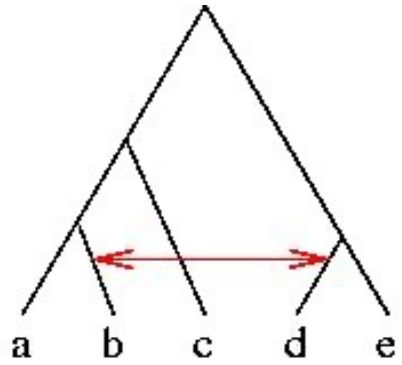


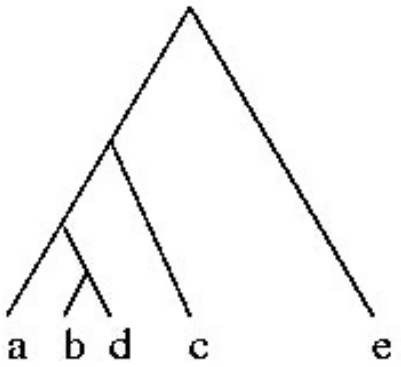




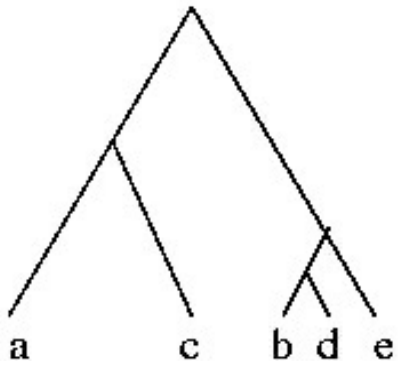
Q(N): the set of quartet trees in network N

Note, there can be more than one tree on a given set of 4 leaves!

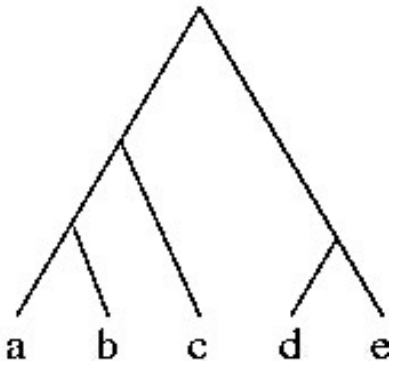




bd|ae, ab|ce



ac|be



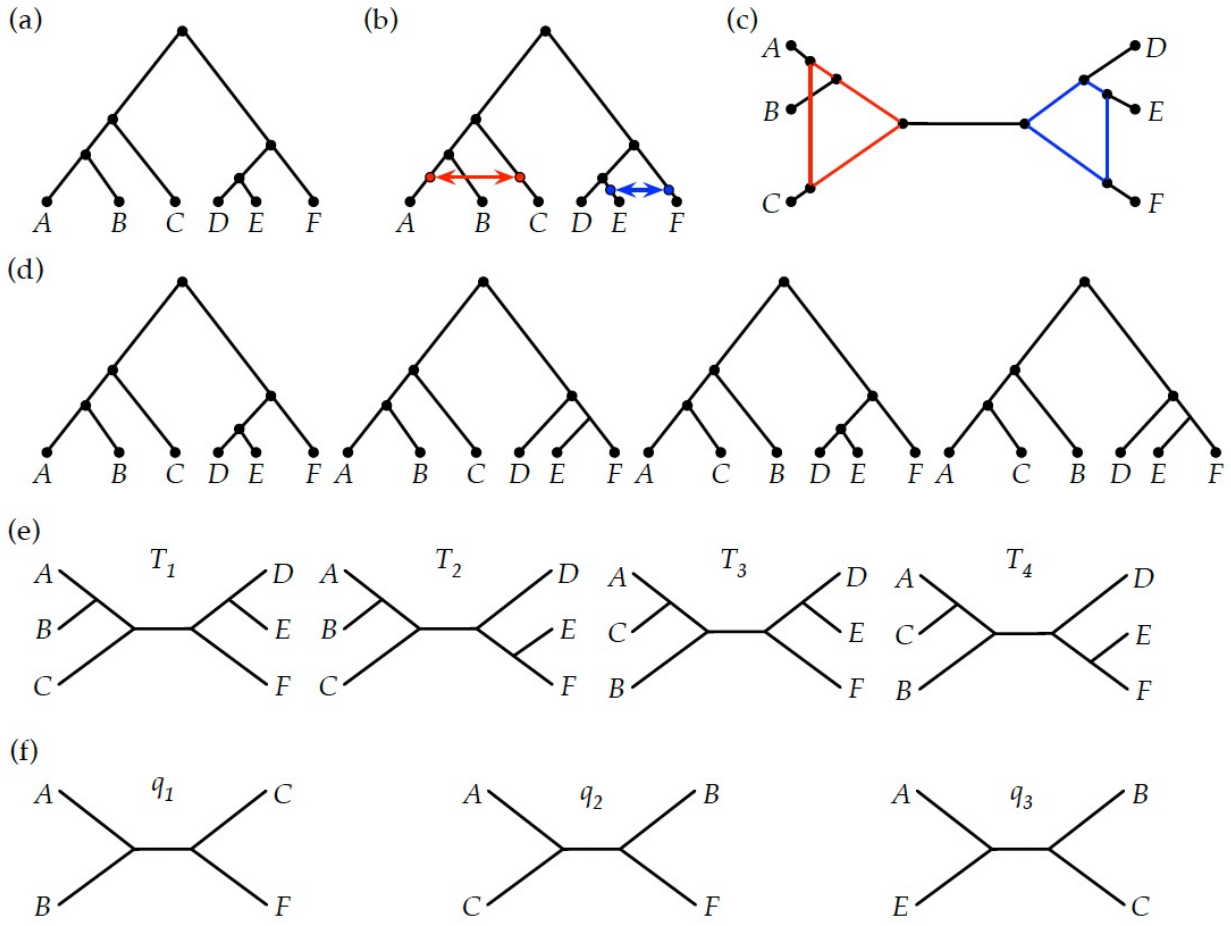


Figure 1 from Warnow et al., 2023.
Examine quartet trees (bottom row): which ones are displayed in the trees in the network?

Constructing Trees from Quartet Trees

- Quartet trees: the induced homeomorphic subtree on four leaves.
- Q(T): the set of all quartet trees of a tree T.
- Theorem: Q(T) uniquely determines T (and T is reconstructable in polynomial time).

In WERN 2006, we used this theorem to establish that if languages evolve down a binary tree T under the WERN 2006 model, then we can estimate T in polynomial time, and in a statistically consistent manner.

Constructing phylogenetic networks

Basic techniques:

- From the trees they contain (rooted or unrooted)
- From clades (rooted subtrees) or (rooted) triplet trees
- From characters that evolve down the network
- From (unrooted) quartet trees

Constructing Networks from Quartet Trees

- Quartet trees: the induced homeomorphic subtree on four leaves.
- Q(N): the set of all quartet trees of a network N.
- If N is a tree, then Q(N) uniquely determines N (and reconstructable in polynomial time).
- Which networks can we construct from sets of quartet trees?

Problems

- Problem 1: Given Q(N), can we reconstruct unrooted topology of N?
- Problem 2: Given a proper subset of Q(N), can we reconstruct unrooted topology of N?
- Problem 3: Can a set Q of quartet trees, can we find a phylogenetic network N such that Q(N) and Q are close?

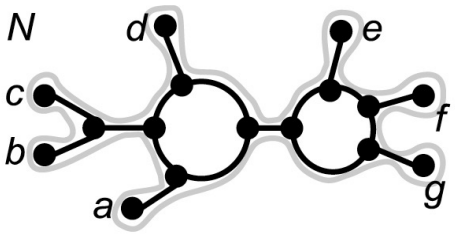
Answers require constraining the network topology.

Gambette et al., 2012

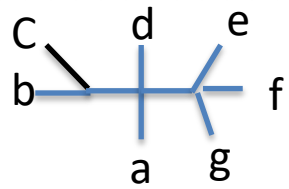
- Problem 1: Given Q(N), can we reconstruct unrooted topology of N?
- Answers:
  - if N is a level-1 or level-2 phylogenetic network, then YES (and in polynomial time).
  - Otherwise, NP-hard to determine
- Gambette et al., 2012 “Quartets and unrooted phylogenetic networks”, J. Bioinformatics and Computational Biology. They give an $O(n^4)$ algorithm to construct a level-1 network from Q(N), as well as other results.

Constructing level-1 network from Q(N) in $O(n^4)$ time.

- Step 1: Let Q* be the set of quartet trees ab|cd such that Q(N) does not contain ac|bd or ad|bc.
- Step 2: Construct a maximally resolved tree T* such that Q(T) contains all quartet trees in Q* . (Note: T* may not be binary.)
- Step 3: Replace high degree nodes by cycles, to produce Q(N).



N



T*

Step 1: Q* contains exactly those quartets for the splits defined by the cut edges.

Step 2: T* is tree formed by collapsing the cycles to single nodes.

Problem 2

Problem 2: Given a proper subset Q of Q(N), can we reconstruct N?

Answer: If Q is dense (at least one tree for every four leaves) and N is level-1, then solvable in polynomial time. Otherwise, not only NP-hard, but there can be exponentially many networks compatible with the set Q.

$O(n^6)$ algorithm in

- Keijsper and RA Pendavingh. Reconstructing a phylogenetic level-1 network from quartets. Bulletin of Mathematical Biology, 76(10):2517–2541, 2014.

Problem 3

Problem 3: Can a set Q of quartet trees, can we find a phylogenetic network N such that Q(N) and Q are close?

Answers (only for case of trees):

- NP-hard
- PTAS (for case of trees) if Q is dense (has a tree on every four leaves)

See Jiang et al. SICOMP 2001

This talk

- Introduction to Phylogenetic Networks
- Models of evolution for linguistic characters (published)
- Identifiability of the Linguistic "Genetic" Tree (published)
- Identifiability of the Linguistic Phylogenetic Network Topology (unpublished)

- Discussion and Future work

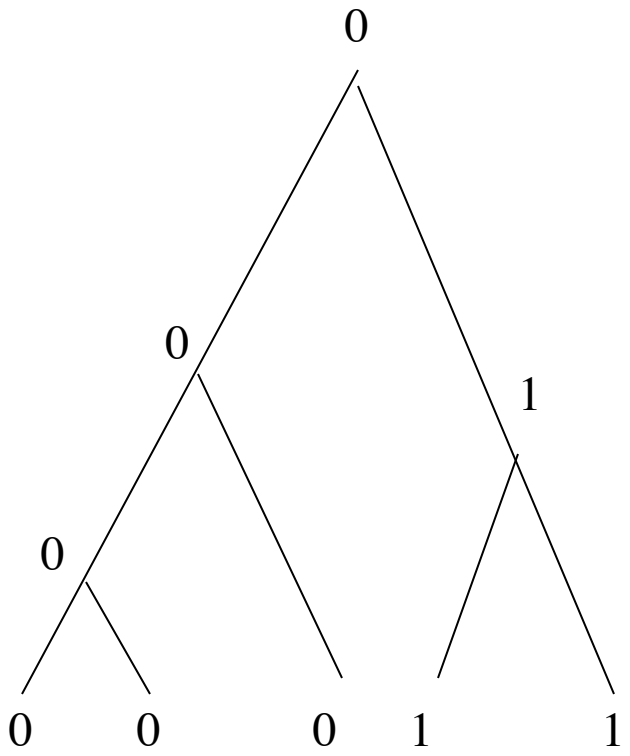
Historical Linguistic Data

- 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 (cognate classes, based on meanings from 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 /kʰéːr/	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.)

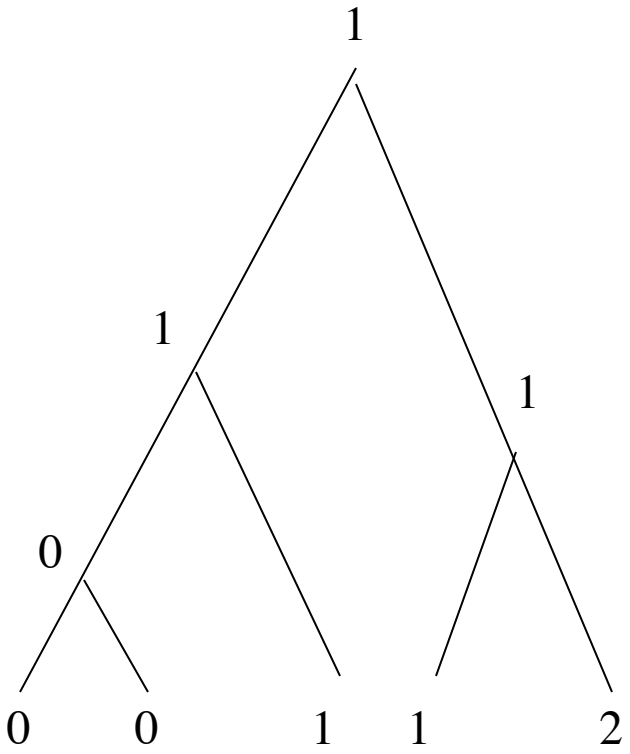
Semantic slot for hand – coded (Partitioned into cognate classes)

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

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

- 1995: Ringe & Warnow "Almost Perfect Phylogeny": most characters evolve without homoplasy under a no-common-mechanism assumption (various publications since 1995)
- 2005: Ringe, Warnow, & Nakhleh "Perfect Phylogenetic Network": extends APP model to allow for borrowing, but assumes homoplasy-free evolution for all characters (Language, 2005)
- 2006: 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)

WERN 2006 model

- Warnow, Evans, Ringe & Nakhleh (WERN) "Extended Markov model": parameterizes PPN and allows for homoplasy provided that homoplastic states can be identified from the data (Cambridge University Press)
- Basic idea:
  - Characters can evolve with homoplasy, but we know the homoplastic states
  - Characters evolve independently but not identically – can have a "no-common-mechanism model"
- The paper proves that if there is no borrowing (and so the evolution is down a tree), then the topology of the tree is identifiable (relatively easy proof from getting quartet trees)

WERN 2006 conjecture

- In WERN 2006, we conjectured that the phylogenetic network might be identifiable under some conditions (e.g., if there was only one borrowing edge).
- However, we did not provide any proof, and we did not make any progress on this.

Work in 2022-2023

- Work with Marc Canby (submitted): modeling polymorphism (e.g., two words for same meaning) – new analysis of Indo-European, plus simulation study.
- Work with Steve Evans and Luay Nakhleh: proving that a level-1 phylogenetic network is identifiable (and providing polynomial time methods to construct the network) under the Warnow, Evans, Ringe, and Nakhleh model (with mild constraints). This will appear as a book chapter.

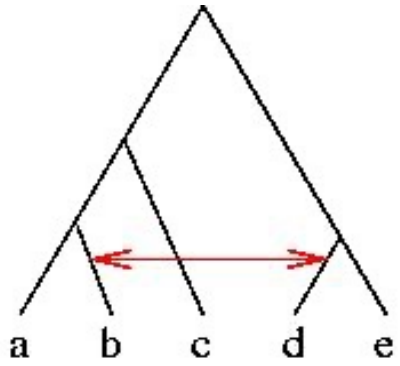
WERN 2023 model

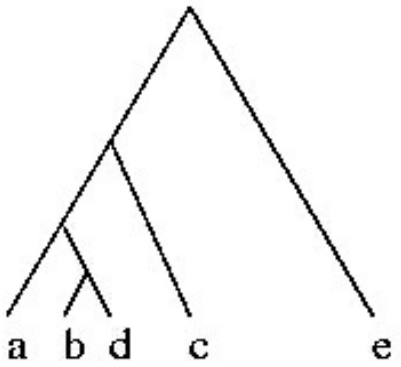
- WERN 2006: Warnow, Evans, Ringe & Nakhleh (WERN) "Extended Markov model": parameterizes PPN and allows for homoplasy provided that homoplastic states can be identified from the data (Cambridge University Press)
- WERN 2006 Basic idea:
  - Characters can evolve with homoplasy, but we know the homoplastic states
  - Characters evolve independently but not identically – can have a "no-common-mechanism model"
- WERN 2023:
  - Add the constraint that the probability of homoplastic state at the root is less than 1

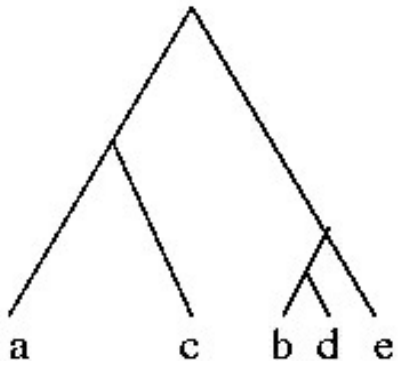
Warnow et al., 2023

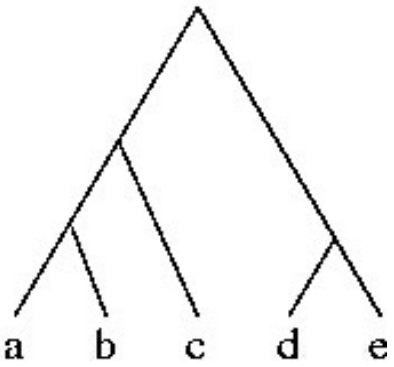
- Warnow, Evans, and Nakhleh 2023 (to appear) proves that:
  - The unrooted phylogenetic network topology is identifiable under the WERN 2023 model, as long as the unrooted network is level-1 (with a mild assumption on root state not being always homoplastic)
  - The rooted topology is also identifiable if the probability of homoplasy-free binary characters is strictly positive

Reminder: Phylogenetic Networks and the Trees they contain

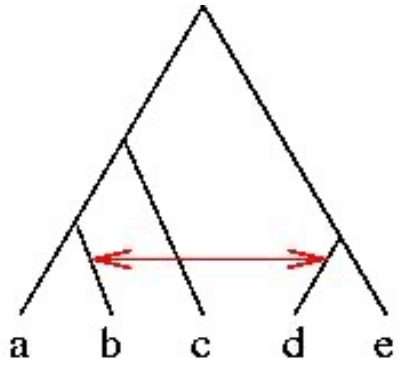


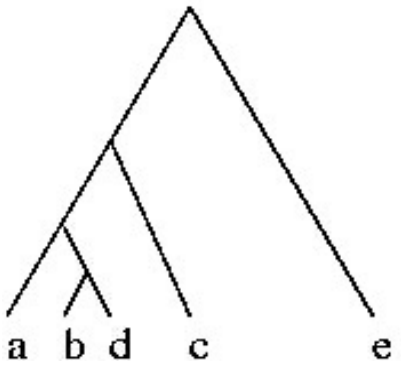


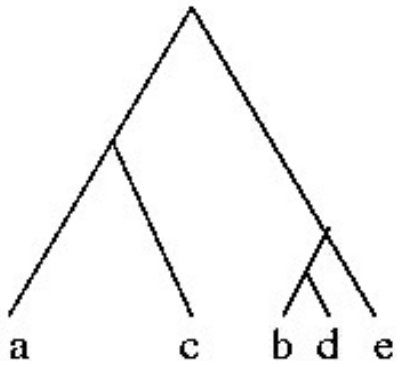


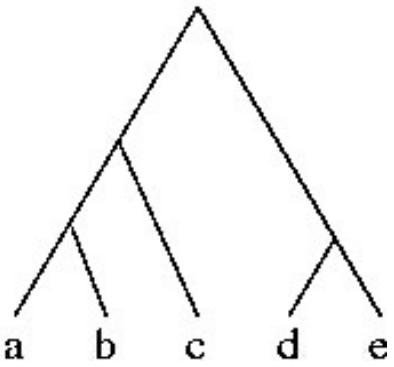


Q(N): The set of all quartet trees of a network N









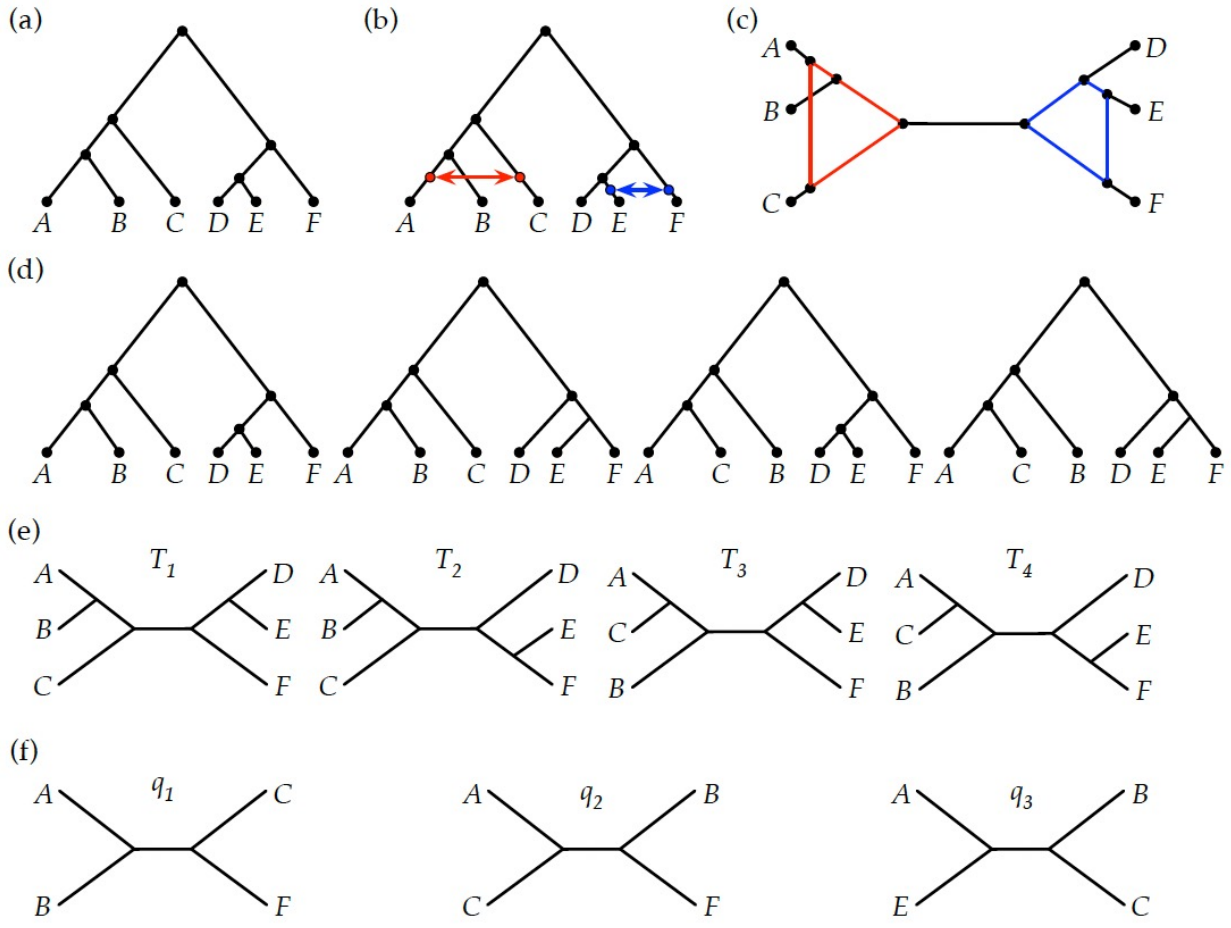


Figure 1 from Warnow et al., 2023. Examine quartet trees (bottom row): which ones are displayed in the trees in the network? (Which ones are in Q(N)?)

Quartet Tree Calculator (QTC)

- Collect quartet trees (uv|wx) satisfying:
  - Any character c with non-homoplastic states 1 and 2, where c(u)=c(v)=1 and c(w)=c(x)=2 defines a quartet
  - Note that such a quartet tree uv|wx is in Q(N)
- Theorem 1: As the number of characters goes to infinity, with probability converging to 1, QTC produces Q(N)

Constructing a network from Q(N)

- Problem: Given Q(N), can we construct the unrooted topology of N?
- Answer: If N is a level-1 network, then the unrooted topology for N can be constructed from Q(N) in polynomial time.
- Algorithms to do this are in Gambette et al. (2012) and Keijsper and Pendavingh (2014). These are not simple methods.

QBTE (Quartet-Based Topology Estimator)

Given the characters (evolving down a network under the the WERN 2023 model)

- Apply QTC to obtain quartet tree set Q
- Apply Gambette et al. (or some other method) to construct the unrooted topology

Theorem 2: If N is level-1 and characters evolve down N under the WERN 2023 model, then QBTE is statistically consistent for estimating the unrooted topology of N.

Limitations to QBTE

- Note that the algorithm we described has two steps:
  - Apply QBTE to obtain set Q (an estimate of Q(N))
  - Apply Gambette et al. (or some other method) to construct the unrooted network
- The algorithm is statistically consistent under the WERN 2006 model if N is a level-1 network, BUT:
  - It doesn't root the network
  - No guarantees if Q is not equal to Q(N)

Rooting the network

- If QBTE succeeds in constructing a level-1 network, we would like to root it. How do we do this?
- If there are homoplasy-free binary characters (with ancestral state known), we can at least partially identify the location of the root. But given enough of them, we can uniquely identify the location! (This is “Root-Network”). It’s not that simple a method .

What if Q is not Q(N)?

- QTC produces a set Q of quartet trees, but on “real” data (which are finite and may not be generated by the WERN model), it’s easily the case that Q is not Q(N).
- When Q is not Q(N), or a dense subset, the algorithms to construct networks fail to return anything. That is not satisfactory.
- What can we do then?

Theoretical questions

- Under what types of network models can the underlying tree be inferred in a statistically consistent manner?
- What can be said about network models with random borrowing?
- Determine approximability of optimization problems (e.g., given Q, find level-1 network N such that Q(N) is close to Q)
  - Recall PTAS from Tao Jiang (SICOMP) when Q is dense and N is required to be a tree

Practical Research Projects

- Design methods and test them for:
  - Estimating a phylogenetic network from Q, even when Q is not Q(N).
  - Estimating the underlying “genetic tree” from Q, when Q is not Q(N).
  - Adding contact edges to a rooted tree, under the WERN 2023 model.

Summary

- Phylogenetic networks are needed for modeling evolution, both in linguistics and in biology.
- Methods for biological network estimation are mainly likelihood-based and very computationally intensive.
- Discrete methods, such as the ones described, could advance discovery
- BUT – very little theory so far establishing identifiability of phylogenetic networks, and statistically consistent methods are (so far) limited (and do not have any performance guarantees on finite data, especially if not generated by the model)
- New methods and new theory are both needed.