Growing Trees from Big Data

Bayesian Phyologeny for Historical Linguistics

Gereon Kaiping

2017-07-18

Problem

Using the comparative method is hard and limited, because

- it is a lot of painstaking work,
- we don't know how to weigh the evidence,
- is a mix of hypothesis generation and validation,
- loan words and chance resemblances make our lives more difficult,
- cognates may have changed meanings (but how far?).

And then it doesn't even give us dates, just "not before" or "not after" if we are lucky.

Problem

Using the comparative method is hard and limited, because

- it is a lot of painstaking work,
- we don't know how to weigh the evidence,
- is a mix of hypothesis generation and validation,
- loan words and chance resemblances make our lives more difficult,
- cognates may have changed meanings (but how far?).

And then it doesn't even give us dates, just "not before" or "not after" if we are lucky.

Solution

Tree reconstruction methods from **Bioinformatics**

```
cctccacgcc aacggagcct cattcttctt
cctccacgcc aacaaagcct cattctt---
cctccacgcc aacaaagcct cattcttctt →
cctccacgcc aacggagcct caggtgtctt
cctccactcc aacggagcct caggtgtctt
```

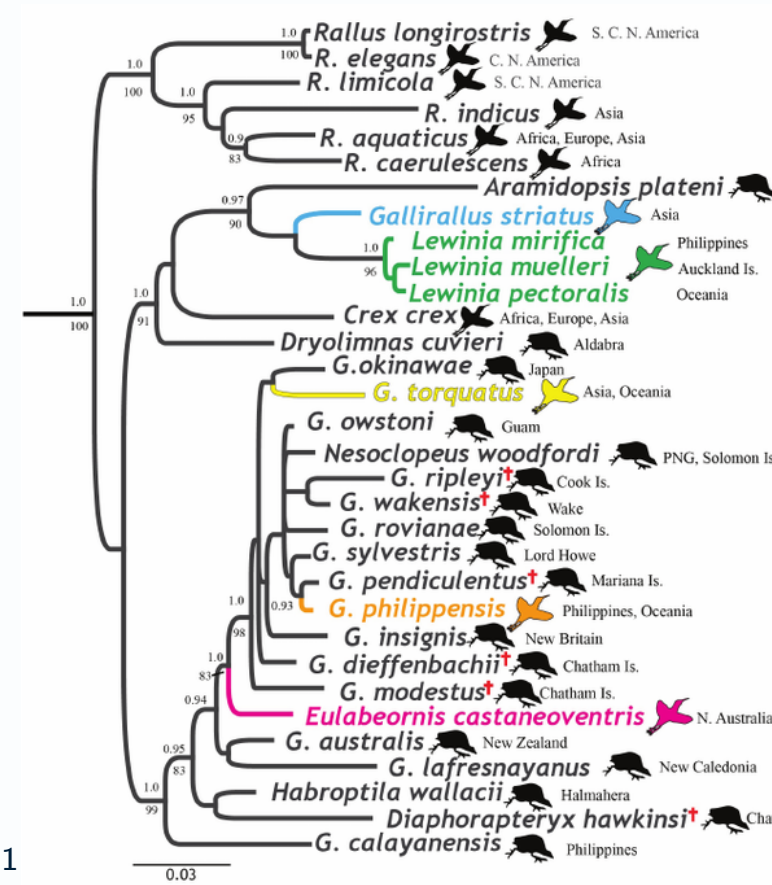
[1] Garcia-Ramirez et al. 2015

Solution

Tree reconstruction methods from **Bioinformatics**

```
cctccacgcc aacggagcct cattcttctt
cctccacgcc aacaaagcct cattctt---
cctccacgcc aacaaagcct cattcttctt
cctccacgcc aacggagcct caggtgtctt
cctccactcc aacggagcct caggtgtctt
```

→



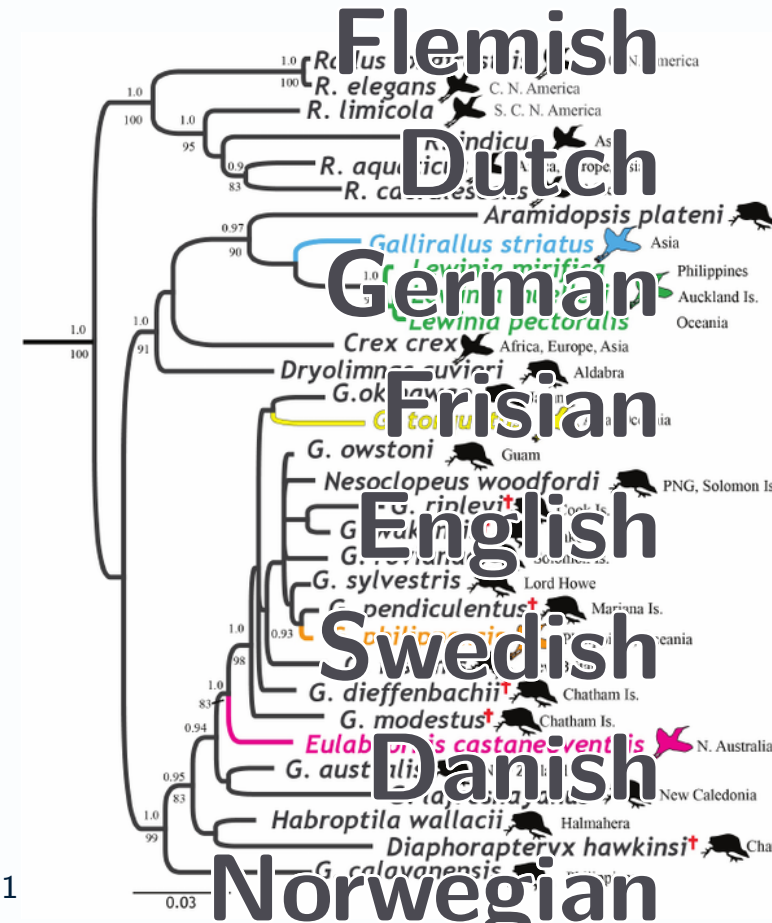
[1]

[1] Garcia-Ramirez et al. 2015

Solution

Tree reconstruction methods from **Bioinformatics**

```
h ʊ n - ə ɹ - t _ v ɔ  ɹ - t
h ʊ n d ɐ - - t _ v ɔː - - t
h ʊ n d ə - - t _ β œ  - - t →
h ɐ n d - ɹ ə d _ w ɜː - - d
h ʊ n d - ɾ a - _ - oː - - d
```



[1] Garcia-Ramirez et al. 2015

Evolution as a random process

Idea Evolution = a random process on a tree[2].

Generate a tree and some language history using dice rolls.

Example

- Start with TATATA
- Roll a dice: On a 5 or 6, split the tree here.
- Roll a dice: Replace that letter in the word with E.

"Likelihood": $P(\text{Data} \mid \text{Model})$

[2] or a network or a population, as long as we can formalize it.

Evolution as a random process

Idea Evolution = a random process on a tree[2].

Generate a tree and some language history using dice rolls.

Example

- Start with TATATA
- Roll a dice: On a 5 or 6, split the tree here.
- Roll a dice: Replace that letter in the word with E.

"Likelihood": $P(\text{Data} \mid \text{Model})$

[2] or a network or a population, as long as we can formalize it.

Evolution as a random process

Idea Evolution = a random process on a tree[2].

Generate a tree and some language history using dice rolls.

Example

- Start with TATATA
- Roll a dice: On a 5 or 6, split the tree here.
- Roll a dice: Replace that letter in the word with E.

"Likelihood": $P(\text{Data} \mid \text{Model})$

[2] or a network or a population, as long as we can formalize it.

Going back from data

Is given data compatible with this model?

- TATETA
- TATATA
- TATUTA
- TATETA, TAEATA
- TAEETA, TETEEA, EATEEA

How compatible is this data with this or that model?

Probabilities = confidence of belief. Not: repeatable random experiment.

"Posterior probability": $P(\text{Model} \mid \text{Data})$

Going back from data

Is given data compatible with this model?

- TATETA
- TATATA
- TATUTA
- TATETA, TAEATA
- TAEETA, TETEEA, EATEEA

How compatible is this data with this or that model?

Probabilities = confidence of belief. Not: repeatable random experiment.

"Posterior probability": $P(\text{Model} \mid \text{Data})$

Going back from data

Is given data compatible with this model?

- TATETA
- TATATA
- TATUTA
- TATETA, TAEATA
- TAEETA, TETEEA, EATEEA

How compatible is this data with this or that model?

Probabilities = confidence of belief. Not: repeatable random experiment.

"Posterior probability": $P(\text{Model} \mid \text{Data})$

Going back from data

Is given data compatible with this model?

- TATETA
- TATATA
- TATUTA
- TATETA, TAEATA
- TAEETA, TETEEA, EATEEA

How compatible is this data with this or that model?

Probabilities = confidence of belief. Not: repeatable random experiment.

"Posterior probability": $P(\text{Model} \mid \text{Data})$

Going back from data

Is given data compatible with this model?

- TATETA
- TATATA
- TATUTA
- TATETA, TAEATA
- TAEETA, TETEEA, EATEEA

How compatible is this data with this or that model?

Probabilities = confidence of belief. Not: repeatable random experiment.

"Posterior probability": $P(\text{Model} \mid \text{Data})$

Going back from data

Is given data compatible with this model?

- TATETA
- TATATA
- TATUTA
- TATETA, TAEATA
- TAEETA, TETEEA, EATEEA

How compatible is this data with this or that model?

Probabilities = confidence of belief. Not: repeatable random experiment.

"Posterior probability": $P(\text{Model} \mid \text{Data})$

Going back from data

Is given data compatible with this model?

- TATETA
- TATATA
- TATUTA
- TATETA, TAEATA
- TAEETA, TETEEA, EATEEA

How compatible is this data with this or that model?

Probabilities = confidence of belief. Not: repeatable random experiment.

"Posterior probability": $P(\text{Model} \mid \text{Data})$

Bayes' Theorem

$$P(\text{Model} \mid \text{Data}) \propto P(\text{Data} \mid \text{Model}) \times P(\text{Model}) \qquad (3)$$

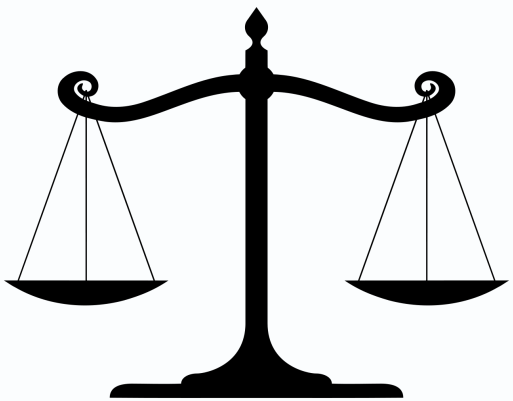
"What did the language history look like?"

=

"What trees are compatible with the data and my idea of language change?"

=

"Weighted by how 'strange' they are, how well does each tree explain my data?"



Bayesian inference may look complicated, but it is

- model-based
- can incorporate prior knowledge
- outputs result uncertainty
- gives implicit weights from first principles

[3] Michael et al. 2015

Bayes' Theorem

$$P(\text{Model} \mid \text{Data}) \propto P(\text{Data} \mid \text{Model}) \times P(\text{Model}) \qquad (3)$$

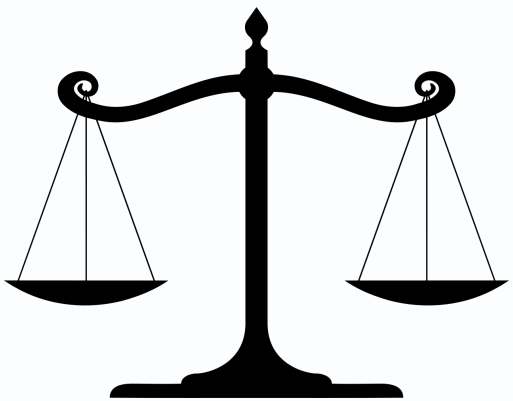
"What did the language history look like?"

=

"What trees are compatible with the data and my idea of language change?"

=

"Weighted by how 'strange' they are, how well does each tree explain my data?"



Bayesian inference may look complicated, but it is

- model-based
- can incorporate prior knowledge
- outputs result uncertainty
- gives implicit weights from first principles

[3] Michael et al. 2015

Computational Phylogenetics

"Roll dice to generate trees, but only keep the good ones"

Need:

- simple stochastic model(s) of language evolution, with parameters
- intuition ("prior") of how parameters look like
- large dataset of model-compatible data

Data:

- Swadesh lists: models based on semantic change (like Glottochronology – but without its problems)
- Geography: various models
- Phonetic alignments: still in infancy
- Typological data: Some approaches, problems with universals/pathways
- Morphosyntax: ??????

Computational Phylogenetics

"Roll dice to generate trees, but only keep the good ones"

Need:

- simple stochastic model(s) of language evolution, with parameters
- intuition ("prior") of how parameters look like
- large dataset of model-compatible data

Data:

- Swadesh lists: models based on semantic change (like Glottochronology – but without its problems)
- Geography: various models
- Phonetic alignments: still in infancy
- Typological data: Some approaches, problems with universals/pathways
- Morphosyntax: ??????

Computational Phylogenetics

"Roll dice to generate trees, but only keep the good ones"

Need:

- simple stochastic model(s) of language evolution, with parameters
- intuition ("prior") of how parameters look like
- large dataset of model-compatible data

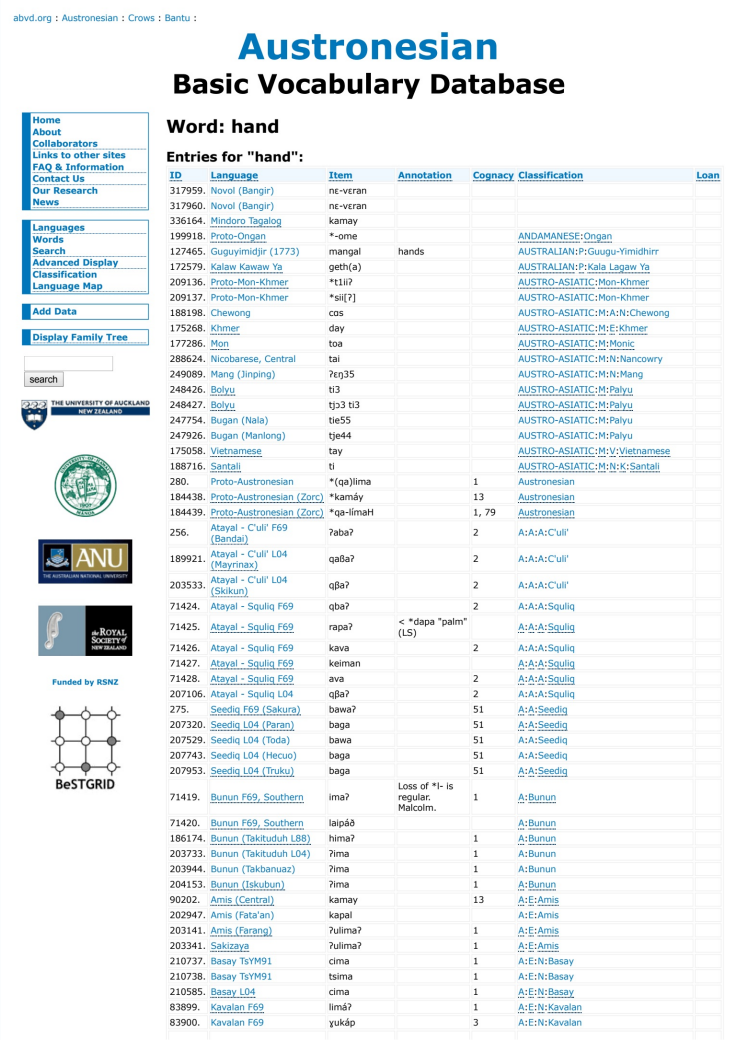
Data:

- Swadesh lists: models based on semantic change (like Glottochronology – but without its problems)
- Geography: various models
- Phonetic alignments: still in infancy
- Typological data: Some approaches, problems with universals/pathways
- Morphosyntax: ??????

So much the theory.

- Austronesian: Branches and times
- Bantu: Phylogeography
- Indo-European: Ancient written sources

Example 1: Austronesian



abvd.org : Austronesian : Crows : Bantu :

Austronesian
Basic Vocabulary Database

Home
About
Collaborators
Links to other sites
FAQ & Information
Contact Us
Our Research
News

Languages
Words
Search
Advanced Display
Classification
Language Map

Add Data

Display Family Tree

search

Funded by RSNZ

BeSTGRID

Word: hand

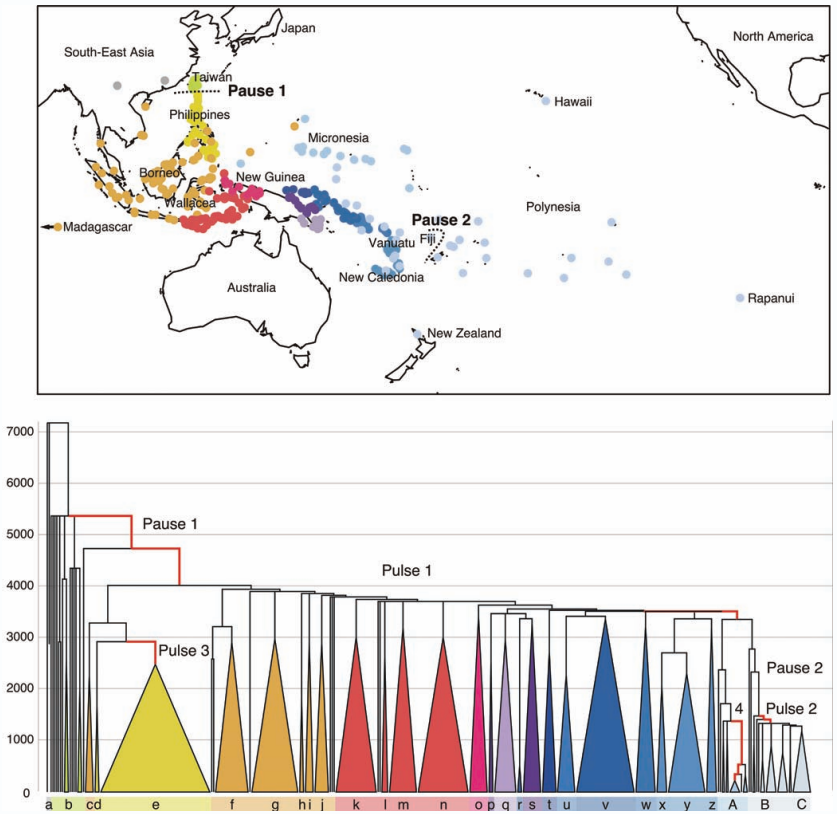
Entries for "hand":

ID	Language	Item	Annotation	Cognacy	Classification	Loan
317959.	Novol (Bangir)	nɛ-vɛran				
317960.	Novol (Bangir)	nɛ-vɛran				
336164.	Mindoro Tagalog	kamay				
199918.	Proto-Ongan	*-ome			ANDAMANESE:Ongan	
127465.	Guguyimidjir (1773)	mangal	hands		AUSTRALIAN:P:Guugu-Yimidhirr	
172579.	Kalaw Kawaw Ya	geth(a)			AUSTRALIAN:P:Kala Lagaw Ya	
209136.	Proto-Mon-Khmer	*t1ii?			AUSTRO-ASIATIC:Mon-Khmer	
209137.	Proto-Mon-Khmer	*sii[?]			AUSTRO-ASIATIC:Mon-Khmer	
188198.	Chewong	cɑs			AUSTRO-ASIATIC:M:A:N:Chewong	
175268.	Khmer	day			AUSTRO-ASIATIC:M:E:Khmer	
177286.	Mon	toa			AUSTRO-ASIATIC:M:Monic	
288624.	Nicobarese, Central	tai			AUSTRO-ASIATIC:M:N:Nancowry	
249089.	Mang (Jinping)	ʔɛŋ35			AUSTRO-ASIATIC:M:N:Mang	
248426.	Bolyu	ti3			AUSTRO-ASIATIC:M:Palyu	
248427.	Bolyu	tjɔ3 ti3			AUSTRO-ASIATIC:M:Palyu	
247754.	Bugan (Nala)	tie55			AUSTRO-ASIATIC:M:Palyu	
247926.	Bugan (Manlong)	tje44			AUSTRO-ASIATIC:M:Palyu	
175058.	Vietnamese	tay			AUSTRO-ASIATIC:M:V:Vietnamese	
188716.	Santali	ti			AUSTRO-ASIATIC:M:N:K:Santali	
280.	Proto-Austronesian	*(qa)lima		1	Austronesian	
184438.	Proto-Austronesian (Zorc)	*kamáy		13	Austronesian	
184439.	Proto-Austronesian (Zorc)	*qa-límaH		1, 79	Austronesian	
256.	Atayal - C'uli' F69 (Bandai)	ʔaba?		2	A:A:A:C'uli'	
189921.	Atayal - C'uli' L04 (Mayrinax)	qaβa?		2	A:A:A:C'uli'	
203533.	Atayal - C'uli' L04 (Skikun)	qβa?		2	A:A:A:C'uli'	
71424.	Atayal - Squliq F69	qba?		2	A:A:A:Squliq	
71425.	Atayal - Squliq F69	rapa?	< *dapa "palm" (LS)		A:A:A:Squliq	
71426.	Atayal - Squliq F69	kava		2	A:A:A:Squliq	
71427.	Atayal - Squliq F69	keiman			A:A:A:Squliq	
71428.	Atayal - Squliq F69	ava		2	A:A:A:Squliq	
207106.	Atayal - Squliq L04	qβa?		2	A:A:A:Squliq	
275.	Seediq F69 (Sakura)	bawa?		51	A:A:Seediq	
207320.	Seediq L04 (Paran)	baga		51	A:A:Seediq	
207529.	Seediq L04 (Toda)	bawa		51	A:A:Seediq	
207743.	Seediq L04 (Hecuo)	baga		51	A:A:Seediq	
207953.	Seediq L04 (Truku)	baga		51	A:A:Seediq	
71419.	Bunun F69, Southern	ima?	Loss of *l- is regular. Malcolm.	1	A:Bunun	
71420.	Bunun F69, Southern	laipâð			A:Bunun	
186174.	Bunun (Takituduh L88)	hima?		1	A:Bunun	
203733.	Bunun (Takituduh L04)	ʔima		1	A:Bunun	
203944.	Bunun (Takbanuaz)	ʔima		1	A:Bunun	
204153.	Bunun (Iskubun)	ʔima		1	A:Bunun	
90202.	Amis (Central)	kamay		13	A:E:Amis	
202947.	Amis (Fata'an)	kapal			A:E:Amis	
203141.	Amis (Farang)	ʔulima?		1	A:E:Amis	
203341.	Sakizaya	ʔulima?		1	A:E:Amis	
210737.	Basay TsYM91	cima		1	A:E:N:Basay	
210738.	Basay TsYM91	tsima		1	A:E:N:Basay	
210585.	Basay L04	cima		1	A:E:N:Basay	
83899.	Kavalan F69	limáʔ		1	A:E:N:Kavalan	
83900.	Kavalan F69	ɣukáp		3	A:E:N:Kavalan	

- Austronesian Basic Vocabulary Database: several 1000 cognate classes for 210 meanings in 400 langs
- Plus two "outgroup" langs, minus borrowings
- Binary covarion model
- Calibrations and variable replacement rates

[5] Greenhill, Blust & Gray 2008

Example 1: Austronesian



"The invention of the outrigger canoe and its sail may have enabled the Austronesians to move across this channel before spreading rapidly over the 7000 km from the Philippines to Polynesia (4). This is supported by linguistic reconstructions showing that the terminology associated with the outrigger canoe complex can only be traced back to Proto-Malayo-Polynesian and not Proto-Austronesian (41)."

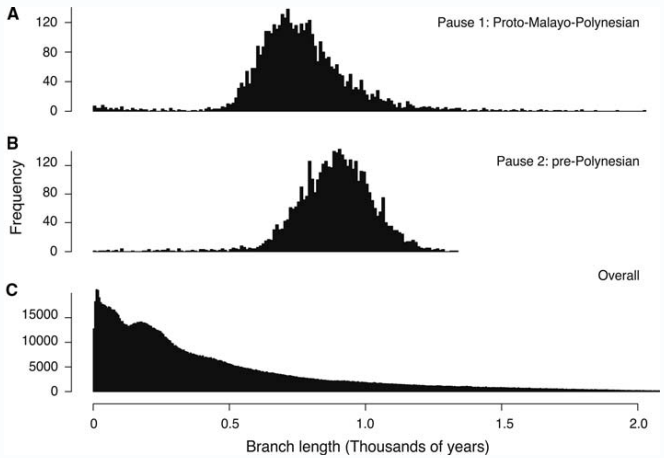


Fig. 3. Histograms of the branch length distributions. (**A**) The distribution of the Proto-Malayo-Polynesian pause, (**B**) the distribution of the pre-Polynesian pause, and (**C**) the overall branch-length distribution.

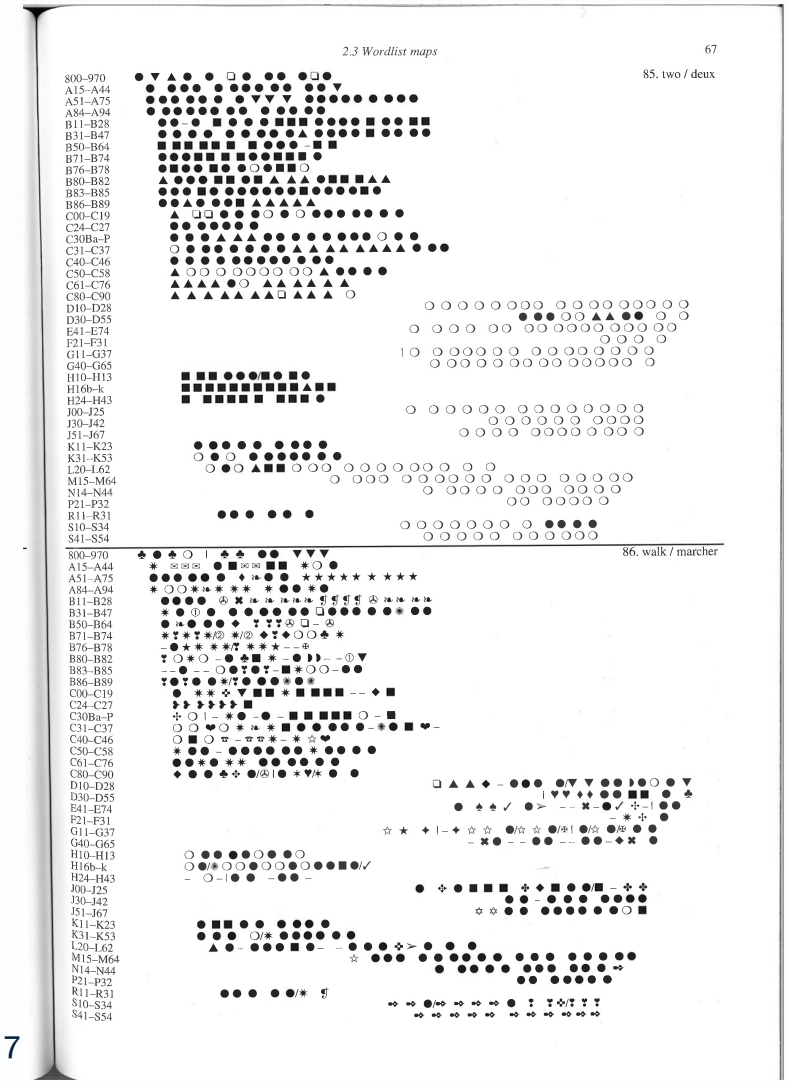
[6] Gray, Drummond & Greenhill 2009

Example 1: Austronesian – Critique

- Pauses and pulses appear with high posterior probability
- Prior? Do the results follow from data or original guess?
- Some subgroupings not linguistically supported – Data contains sociogeography
- How realistic is binary covarion?

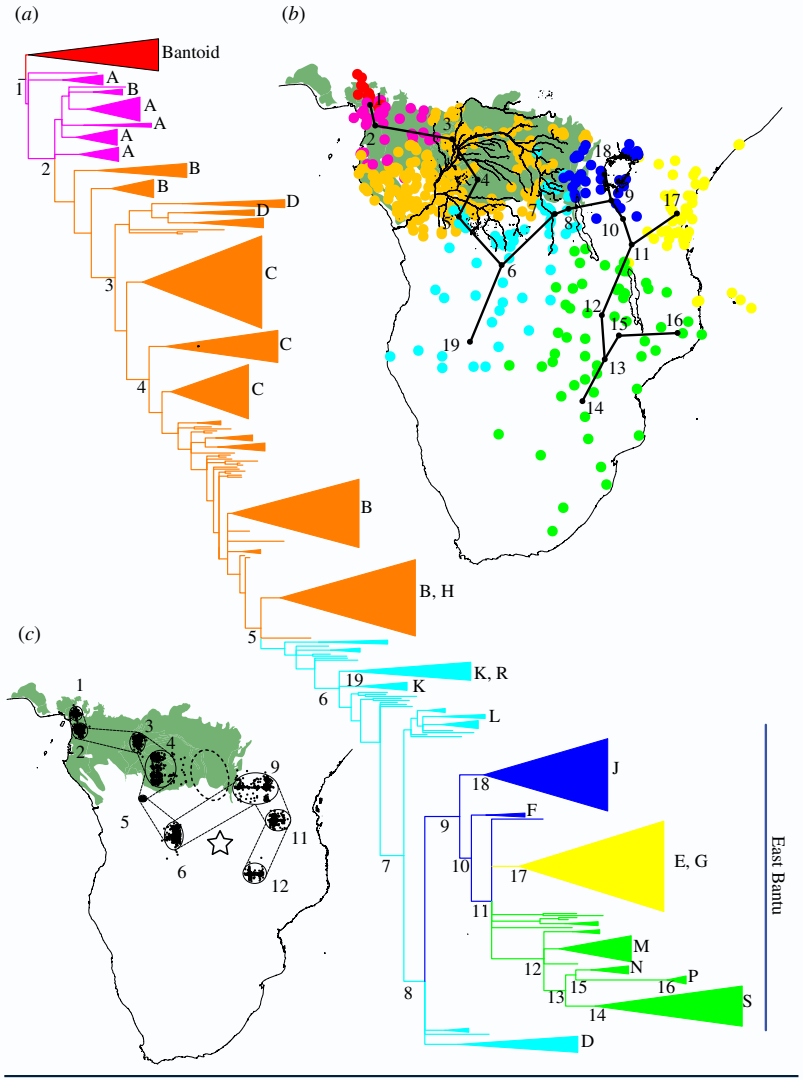
Example 2: Bantu

- 2908 cognate classes for 90 meanings in 542 varieties of Bantu/Bantoid, with geographical point-coordinates
- Binary covarion with 6 (empirical) rate categories
- Brownian motion ancestral state reconstruction of latitudes and longitudes on 500 best trees
- Branch-dependent speed of movement and lexical change
- Other statistical analyses



[7] Bastin, Coupez & Mann 1999

Example 2: Bantu



"These debates have implications regarding the origin and spread of important cultural innovations, such as metallurgy and cattle-keeping."
"[...] explicit mapping of ancestral locations to make inferences about the specific route taken during the dispersal of Bantu languages. The results clearly support the 'pathway through the rainforest' scenario for the expansion of Bantu through much of sub-Saharan Africa. There is no support in these analyses for an early, deep split between East and West Bantu languages and a movement by one branch north of the rainforest."

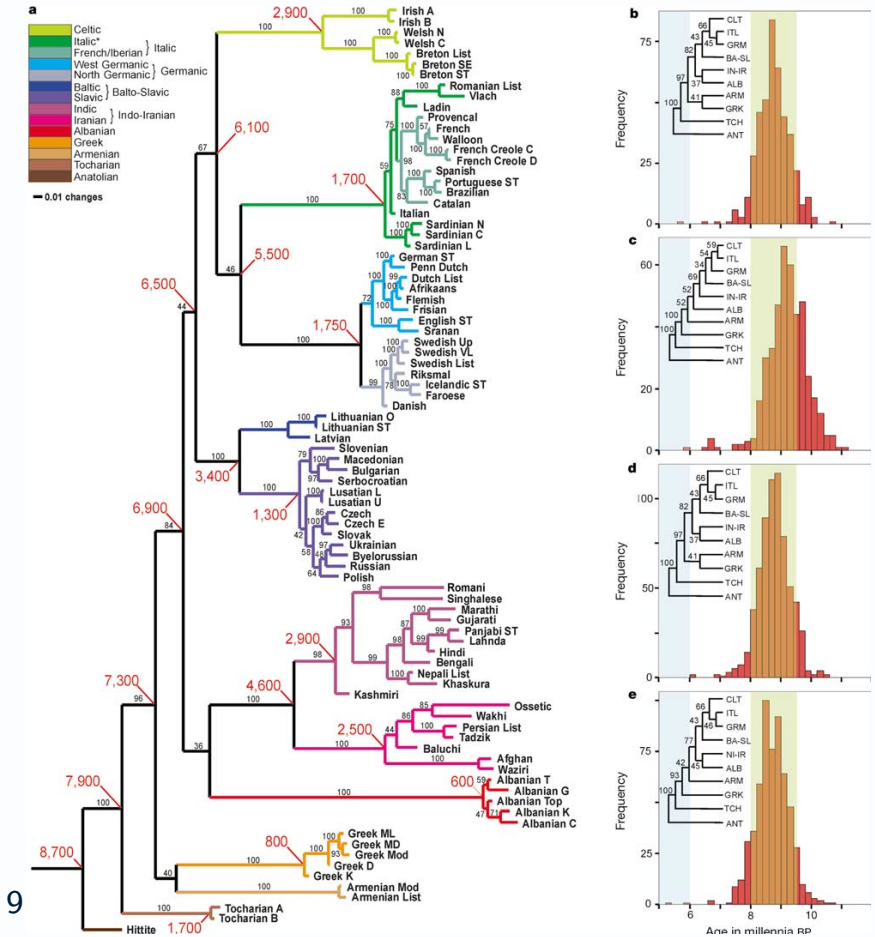
[8] Currie et al. 2013

Example 2: Bantu – Critique

- Several robustness checks of parameters
- Prior?
- How good is Brownian motion as model for language spread?
  Language shift and post-split contact might affect geographic inference.
  (Though the fundamental results look robust.)

Example 3: Indo-European – Data and Prologue

```
a 066 HAND
b                          001
  066 73 Ossetic             K"YX
  066 59 Gujarati            NATH
b                          002
  066 17 Sardinian N         MANU
  066 18 Sardinian L         MANU
  066 09 Vlach               MYNE
  066 22 Brazilian           MAO
  066 21 Portuguese ST       MAO
  066 15 French Creole C     LAME
  066 13 French              MAIN
  066 16 French Creole D     LAME
  066 14 Walloon             MIN
  066 12 Provencal           MAIN
  066 20 Spanish             MANO
  066 23 Catalan             MA
  066 10 Italian             MANO
  066 19 Sardinian C         MANU
  066 11 Ladin               MAUN
  066 08 Rumanian List       MINA
```



[8] Dyen 1997

[9] Gray & Atkinson 2003

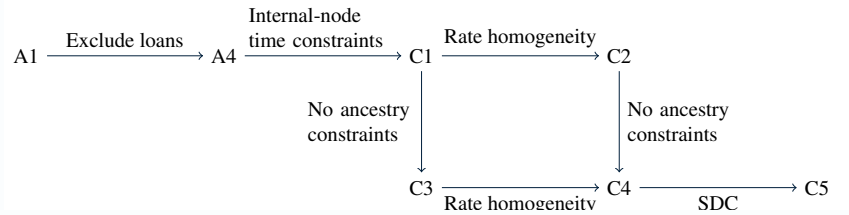
Example 3: Indo-European

IELex

login

hand

ID	Language	Source Form	Phonological Form	Notes	Cognate Class
114	Proto-Indo-European	*mon-u-			E
114	Proto-Indo-European	*ĝʰes-r(o)-, *ĝʰes-t(o)-,			C
114	Proto-Indo-European	*mar-			E
80	Hittite	keššar			C
133	Luvian	īssaris			C
134	Lycian	izre			C
81	Tocharian A	tsar			C
82	Tocharian B	ṣar			C
88	Albanian	dorë		A singularised neut. plural PAlb ...	C
143	Standard Albanian	dorë			C
2	Albanian Sicily	dorë		A singularised neut. plural PAlb ...	C
4	Albanian Corinth	dorë		A singularised neut. plural PAlb ...	C
3	Albanian Gheg	dorë		A singularised neut. plural PAlb ...	C
6	Albanian Tosk	dorë		A singularised neut. plural PAlb ...	C
173	Mycenaean Greek	ke-ᵽ	kʰer-	Attested as an element in ...	C
110	Ancient Greek	χείρ	kʰéːr	G.sg. χειρός	C
152	Tsakonian	χερα			C
32	Greek	χερί	'çeri		C
31	Greek Lesbos	CHERI			C
129	Classical Armenian	ձեռն	jeṙn		C
8	Armenian Eastern	ձեռք	d͡zɛrkʰ		C
7	Armenian Western	ձեռք	t͡sʰɛrkʰ		C
128	Avestan	zastō			C



Starting from a replication of previous work[11], improve

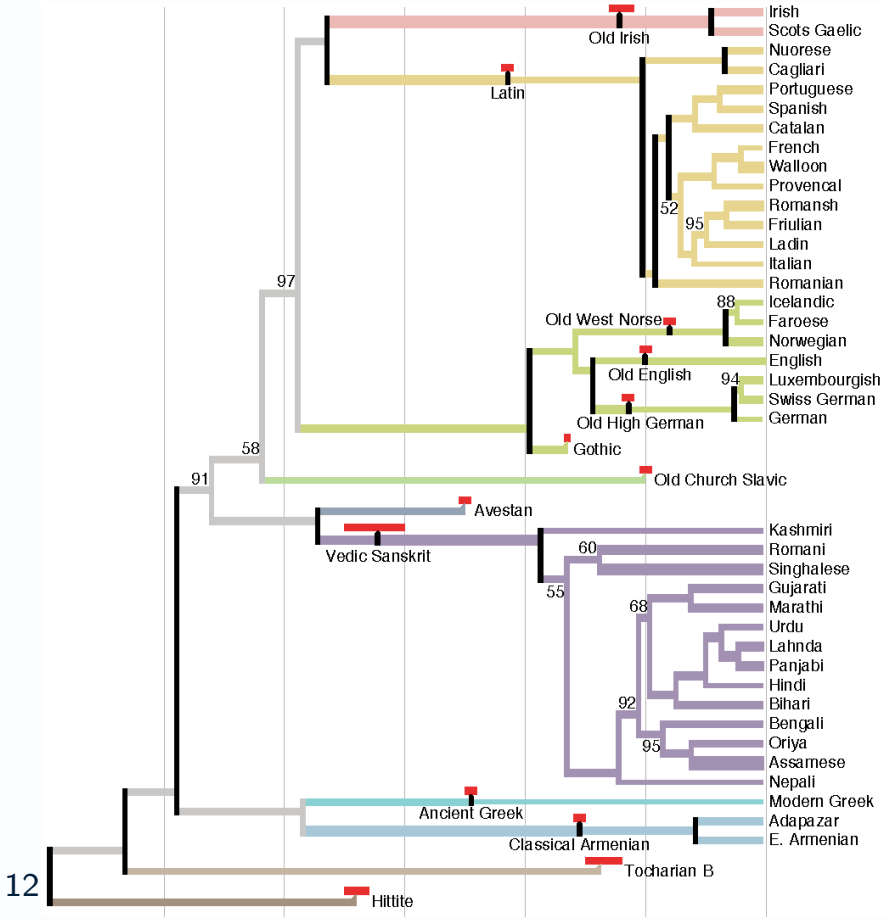
- data
- methodology
- tree prior
- post-processing

comparing each step.

[10] Dunn 2015

[11] Bouckaert et al. 2012

Example 3: Indo-European



"Here we present a phylogenetic analysis in which ancestry constraints permit more accurate inference of rates of change, based on observed changes between ancient or medieval languages and their modern descendants, and we show that the result strongly supports the steppe hypothesis."
"Because previous statistical phylogenetic research supported the Anatolian hypothesis, linguists who find that hypothesis implausible for other reasons may dismiss statistical analyses that purport to determine ancestral chronology. [...] statistical phylogenetic analysis can yield reliable information about pre-historic chronology, at least where all of the available data is taken into consideration."

[12] Chang et al. 2015

Example 3: Indo-European – Critique

- Very explicit about methodology (small steps, driver files available)
- Careful description of data coding
- Would someone have been this careful if the original results *had* matched the linguists' expectations?
- Ancestral constraints are very strong, and somewhat artifical in the model.

[12] Verkerk 2017

It will not solve all problems

Some papers

- disregard prior knowledge
- use models that don't fit their data
- don't show their priors

Not even in a better world.

State of the art models

- can only build trees, no language contact
- only support cognate data (baby steps towards phonetic data and typology), no distinction between innovation and retention
- are not realistic / not calibrated / have biases
- can't actually decide high-level relationships

I think we will always

- need domain experts
- have problems modelling morphosyntax
- have qualitative data that are hard to integrate

Not even in a better world.

State of the art models

- can only build trees, no language contact
- only support cognate data (baby steps towards phonetic data and typology), no distinction between innovation and retention
- are not realistic / not calibrated / have biases
- can't actually decide high-level relationships

I think we will always

- need domain experts
- have problems modelling morphosyntax
- have qualitative data that are hard to integrate

Not even in a better world.

State of the art models

- can only build trees, no language contact
- only support cognate data (baby steps towards phonetic data and typology), no distinction between innovation and retention
- are not realistic / not calibrated / have biases
- can't actually decide high-level relationships

I think we will always

- need domain experts
- have problems modelling morphosyntax
- have qualitative data that are hard to integrate

Not even in a better world.

State of the art models

- can only build trees, no language contact
- only support cognate data (baby steps towards phonetic data and typology), no distinction between innovation and retention
- are not realistic / not calibrated / have biases
- can't actually decide high-level relationships

I think we will always

- need domain experts
- have problems modelling morphosyntax
- have qualitative data that are hard to integrate

Not even in a better world.

State of the art models

- can only build trees, no language contact
- only support cognate data (baby steps towards phonetic data and typology), no distinction between innovation and retention
- are not realistic / not calibrated / have biases
- can't actually decide high-level relationships

I think we will always

- need domain experts
- have problems modelling morphosyntax
- have qualitative data that are hard to integrate

Not even in a better world.

State of the art models

- can only build trees, no language contact
- only support cognate data (baby steps towards phonetic data and typology), no distinction between innovation and retention
- are not realistic / not calibrated / have biases
- can't actually decide high-level relationships

I think we will always

- need domain experts
- have problems modelling morphosyntax
- have qualitative data that are hard to integrate

Conclusions

- Computer models can help make sense of large data sets
- The computer only tests consistency or helps build intuition, it does not replace expertise
- Very few language-appropriate models so far
- Building a *good* inference is hard!
- Mathematical models can handle and combine new types of data for new *types* of results

If you disagree with results, *what parameters or choices do you disagree with?*

Sources and Further Reading I

Bastin, Yvonne, André Coupez & Michael Mann. 1999. *Continuity and divergence in the bantu languages: perspectives from a lexicostatistic study*. Musée royal de l'Afrique centrale.

Bouckaert, Remco, Philippe Lemey, Michael Dunn, Simon J. Greenhill, Alexander V. Alekseyenko, Alexei J. Drummond, Russell D. Gray, Marc A. Suchard & Quentin D. Atkinson. 2012. Mapping the Origins and Expansion of the Indo-European Language Family. *Science* 337(6097). 957–960. https://doi.org/10.1126/science.1219669. http://www.sciencemag.org/content/337/6097/957 (3 December, 2014).

Chang, Will, Chundra Cathcart, David Hall & Andrew Garrett. 2015. Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language* 91(1). 194–244. https://doi.org/10.1353/lan.2015.0005. http://www.linguisticsociety.org/files/news/ChangEtAlPreprint.pdf (27 February, 2015).

Sources and Further Reading II


Currie, Thomas E., Andrew Meade, Myrtille Guillon & Ruth Mace. 2013. Cultural phylogeography of the bantu languages of sub-saharan africa. *Proceedings of the Royal Society of London B: Biological Sciences* 280(1762). https://doi.org/10.1098/rspb.2013.0695. http://rspb.royalsocietypublishing.org/content/280/1762/20130695.


Dunn, Michael. 2015. *IELex – Indo-European Lexical Cognacy Database*. http://ielex.mpi.nl/ (12 March, 2015).


Dyen, Isidore. 1997. *COMPARATIVE INDOEUROPEAN DATABASE COLLECTED BY ISIDORE DYEN*. http://www.wordgumbo.com/ie/cmp/iedata.txt.


Garcia-Ramirez, Juan C, Graeme Elliott, Kath Walker, Isabel Castro & Steven A Trewick. 2015. Trans-equatorial range of a land bird lineage (aves: rallidae) from tropical forests to subantarctic grasslands. *Journal of Avian Biology*.

Sources and Further Reading III


Gray, R. D., A. J. Drummond & S. J. Greenhill. 2009. Language phylogenies reveal expansion pulses and pauses in pacific settlement. *Science* 323(5913). 479–483. https://doi.org/10.1126/science.1166858. http://science.sciencemag.org/content/323/5913/479.


Gray, Russell D. & Quentin D. Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426(6965). 435–439. https://doi.org/10.1038/nature02029. http://www.nature.com/nature/journal/v426/n6965/abs/nature02029.html (27 November, 2014).


Greenhill, S.J., R Blust & R.D. Gray. 2008. The austronesian basic vocabulary database: from bioinformatics to lexomics. *Evolutionary Bioinformatics*. 271–283.


McMahon, April & Robert McMahon. 2005. *Language classification by numbers*. Oxford University Press. 285 pp.

Sources and Further Reading IV


Michael, Lev, Natalia Chousou-Polydouri, Keith Bartolomei, Erin Donnelly, Sérgio Meira, Vivian Wauters & Zachary O'Hagan. 2015. A Bayesian Phylogenetic Classification of Tupí-Guaraní. *LIAMES: Línguas Indígenas Americanas* 15(2). 193–221. https://doi.org/10.20396/liames.v15i2.8642301. https://periodicos.sbu.unicamp.br/ojs/index.php/liames/article/view/8642301 (29 August, 2017).


Verkerk, Annemarie. 2017. Phylogenies: Future, not fallacy. *Language Dynamics and Change* 7(1). 127–140. https://doi.org/10.1163/22105832-00601013. http://booksandjournals.brillonline.com/content/journals/10.1163/22105832-00601013 (9 October, 2017).