Understanding Bayesian phylogenetic inference

SSOL 2024, České Budějovice

Jessica Nieder & Kellen Parker van Dam
Lehrstuhl für Multilinguale Computerlinguistik
Universität Passau, Germany

23 August 2024

Structure for today

- What are Bayesian phylogenies
- How to read them

What are Bayesian phylogenies

The phylogenetic trees discussed today are probabilistic tree models based on Bayesian inference. They are quantitative and, importantly reproducible.

They are ‘Bayesian’ as they are based on Bayes’ theorem...

maybe Bayes, maybe not, who knows.

Bayes' theorem

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

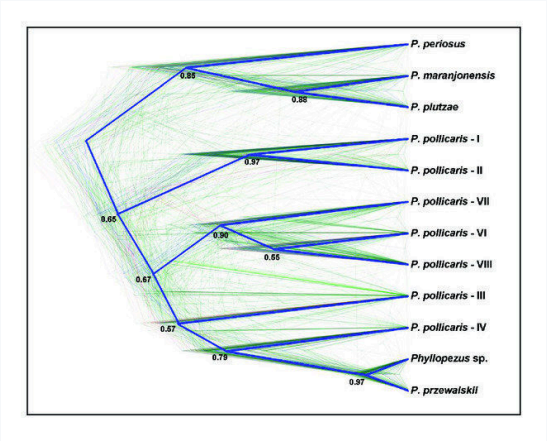
The probability that A is true given that B is true equals: the probability that B is true given that A is true multiplied by the probability that A is true, all divided by the probability that B is true.

$P(A \mid B)$ is our posterior probability

We can use it for phylogenetic (family tree) analysis to determine the likelihood of a given genealogical tree for a given data set.

Consensus trees

Bayesian phylogenies come in difference shapes. They can be consensus trees, i.e. showing the final maximum likelihood as a single clear tree, known as the consensus tree, or they can be shown as density trees.



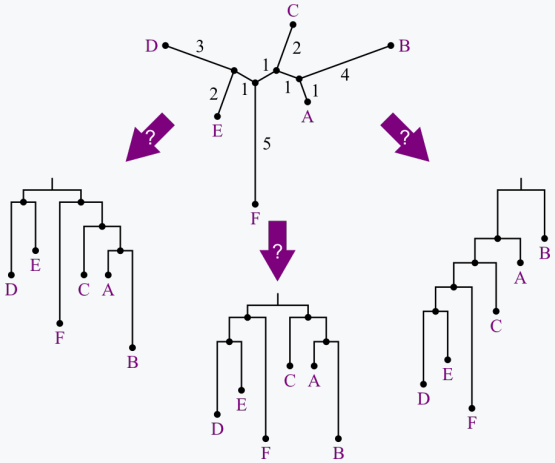
Cacciali, Pier, et al. "Cryptic diversity in the Neotropical gecko genus Phyllopezus Peters, 1878 (Reptilia: Squamata: Phyllodactylidae): A new species from Paraguay." International journal of zoology 2018.1 (2018): 3958327.

Rooting

Trees will be either rooted or unrooted. Rooting the tree is something the researcher does, not something the algorithm does.

In linguistics, you root the tree based on a languages which is related, but known to be more distantly related than all the rest. This is called the outgroup.

For **Romance** languages we may use German as the outgroup. Why?

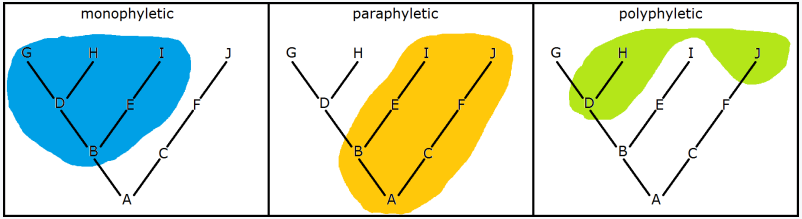


from https://pages.cs.wisc.edu/~aasmith/

Clades

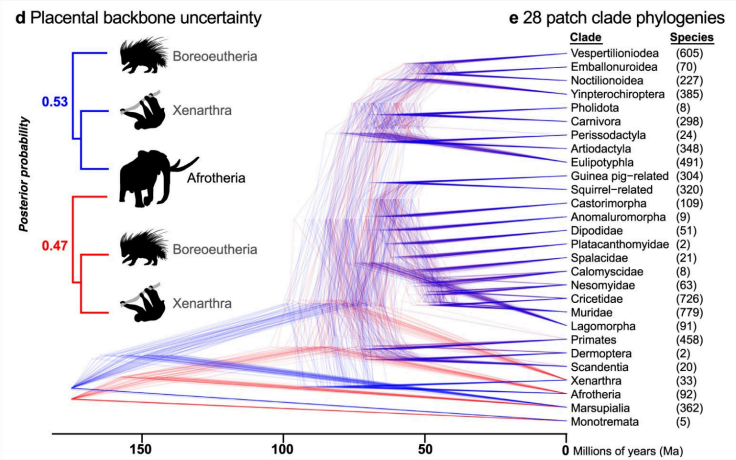
Subgroups (offshoots, branches...) are called clades.

these can be monophyletic, paraphyletic or polyphyletic.



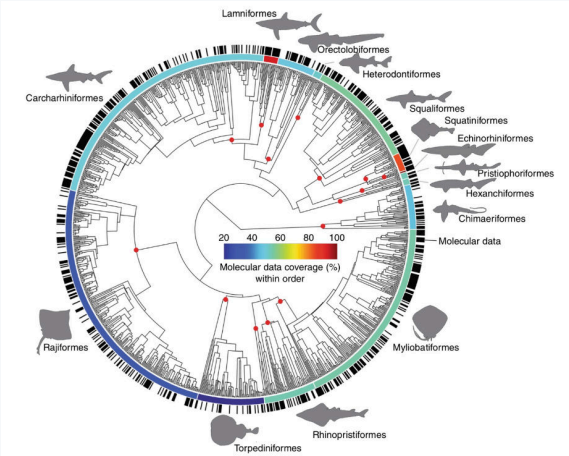
from https://www.sporcle.com/games/Scuadrado/taxon-taxoff

Posterior probabilities

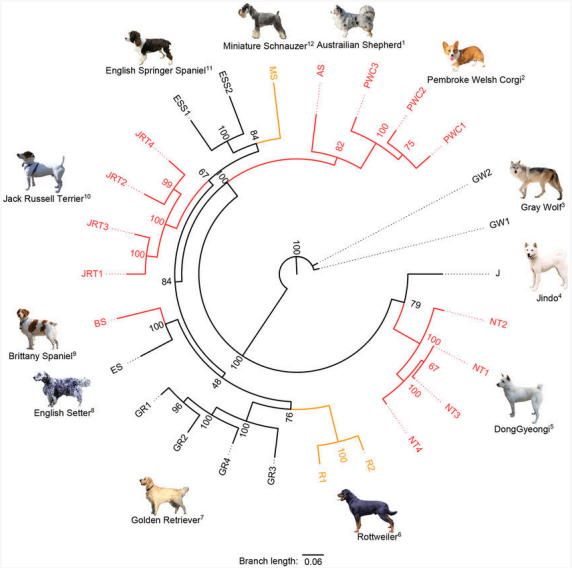


Upham, Nathan S., Jacob A. Esselstyn, and Walter Jetz. "Inferring the mammal tree: species-level sets of phylogenies for questions in ecology, evolution, and conservation." PLoS biology 17.12 (2019): e3000494.

Origins in genetics

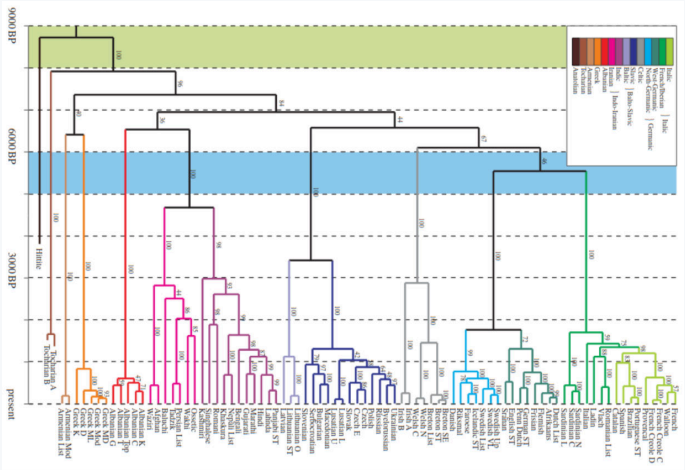


Stein RW, Mull CG, Kuhn TS, Aschliman NC, Davidson LNK, Joy JB, Smith GJ, Dulvy NK, and Mooers AO. Global priorities for conserving the evolutionary history of sharks, rays and chimaeras. Nat Ecol Evol. 2018;2: 288–298. http://dx.doi.org/10.1038/s41559-017-0448-4



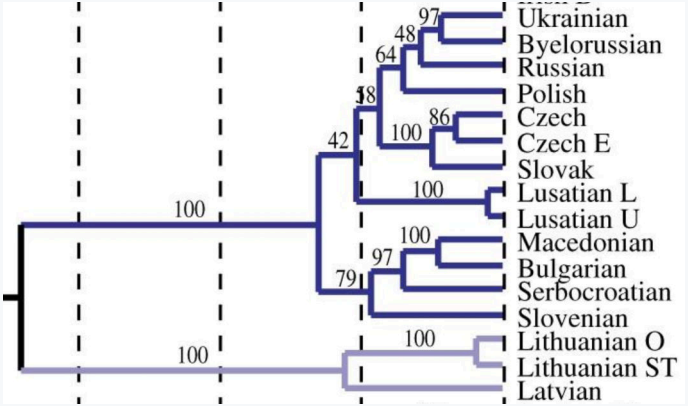
Yoo, DongAhn, et al. "The genetic origin of short tail in endangered Korean dog, DongGyeongi." Scientific reports 7.1 (2017): 10048.

Applications in Linguistics



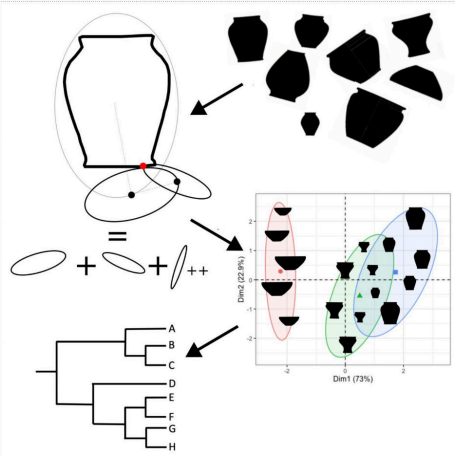
Gray, R.D., Atkinson, Q.D. and Greenhill, S.J., 2011. Language evolution and human history: what a difference a date makes. Philosophical Transactions of the Royal Society B: Biological Sciences, 366(1567), pp.1090-1100.

Applications in Linguistics



Gray, R.D., Atkinson, Q.D. and Greenhill, S.J., 2011. Language evolution and human history: what a difference a date makes. Philosophical Transactions of the Royal Society B: Biological Sciences, 366(1567), pp.1090-1100.

It's not just genetics or linguistics



Marwick, Ben, David N. Matzig, and Felix Riede. "Bayesian inference of material culture phylogenies using continuous traits: A birth–death model for Late Neolithic/Early Bronze Age arrowheads from Northwestern Europe." (2023).

Linguistic phylogenies

Most linguistic phylogenies coming out today are lexical in nature, i.e. the **cognate sets** are what determined the branches. But they do not need to be lexical

In the same way that we can code pottery features instead of words, we could also code grammatical features, or phonological features. However...

Would this be a good data set for a phylogenetic tree?

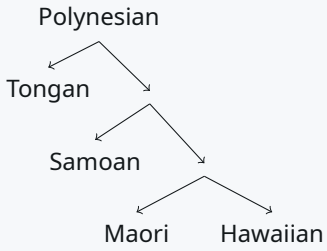
	Hawaiian	Maori	Samoan	Tongan	
1	manu	manu	manu	manu	‘bird’
2	awa	awa	awa	awa	‘channel’
3	niu	niu	niu	niu	‘coconut’
4	pua	pua	pua	pua	‘flower’
5	peʔa	peka	peʔa	peka	‘bat’
6	muli	muri	muli	mui	‘behind’
7	kani	taŋi	taŋi	taŋi	‘cry’
8	au	au	au	ʔau	‘current’
9	kuna	tuna	tuna	tuna	‘eel species’
10	walu	waru	walu	walu	‘eight’
11	iʔa	ika	iʔa	ika	‘fish’
12	kae	tae	tae	taʔe	‘excrement’
13	lau	rau	lau	lau	‘leaf’
14	ʔuku	kutu	ʔutu	kutu	‘louse’
15	umu	umu	umu	ʔumu	‘oven, earthen’
16	walu	waru	walu	wau	‘scratch’
17	kapu	tapu	tapu	tapu	‘taboo’
18	ako	ato	ato	ʔato	‘thatch, roof’
19	lua	rua	lua	ua	‘two’
20	lua	rua	lua	lua	‘vomit

Branching events

From the Polynesian example we can come up with a rough tree for the languages that shows how they may relate to each other, assuming phonology as informative in this case.

We could group them based on sound changes, or lexical changes, or whatever else we think might be informative.

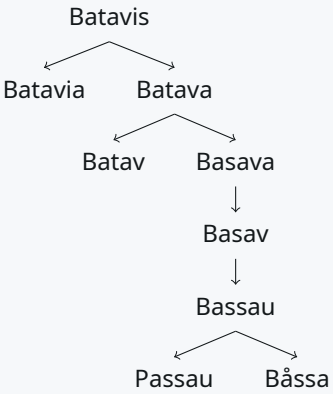
The wordlist we looked at only had cognates, so we'd have to go by phonology, but this is not usually genealogically informative. Why not?



Branching events

Linguistic data can also be used to determine other things, such as topics within forensic linguistics.

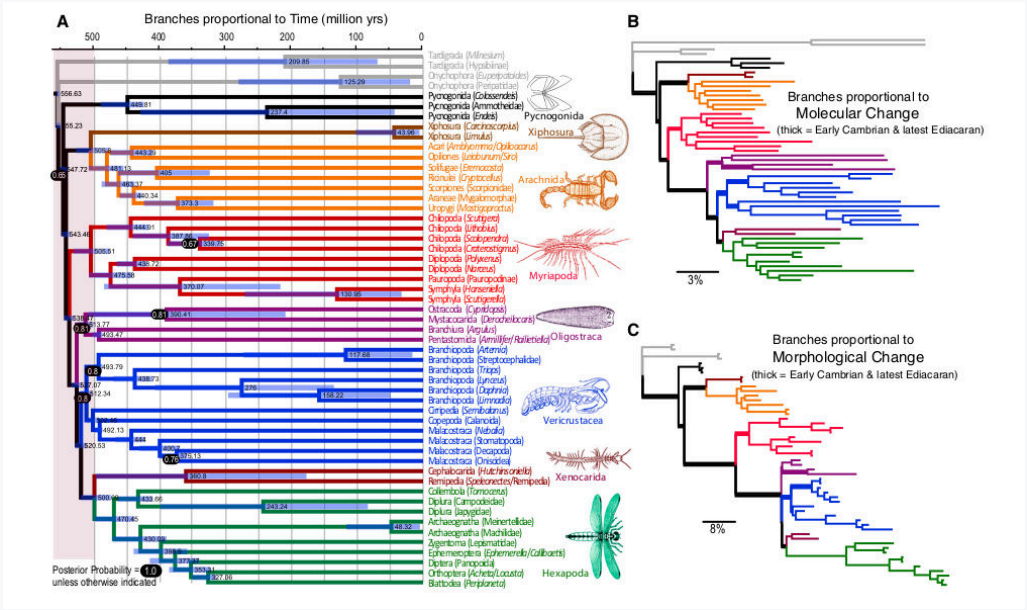
Imagine a series of old hand-copied versions of the same core document, perhaps a founding document from the diocese of Passau. The name of Passau has changed over time, and we might encode the spellings to determine the origin and order of the copies.



Software

There are a number of software tools available. The main ones are **BEAST2** and **MrBayes**.

The main difference, for practical purposes, is that BEAST relies on a clock, and MrBayes does not. This matters based
on what we want our branch lengths to represent, or if we even have something to calibrate a clock to.


A
Branches proportional to Time (million yrs)
500
400
300
200
100
0
Pycnogonida
Xiphosura
Arachnida
Myriapoda
Oligostraca
Vericrustacea
Xenocarida
Hexapoda
Posterior Probability = 1.0
unless otherwise indicated
B
Branches proportional to
Molecular Change
(thick = Early Cambrian & latest Ediacaran)
3%
C
Branches proportional to
Morphological Change
(thick = Early Cambrian & latest Ediacaran)
8%

Clock calibration

The idea of calculating age of linguistic branching events based on some steady rate of change (glottochonology) is an old idea, and one which has been rightfully rejected.

Languages don't change at anything like a steady rate.
For genes, it's a bit more stable (but still not fully regular).

However, if we have language varieties which we can assign to dates, this does let us at least place them in time within the tree and BEAST can then make some educated guesses on the timing of events otherwise.

Data preparation

Data starts out as the flat data we've discussed previously. Here's some of mine at an early stage.

concept	orthographic	phonetic	phonemic	unified	full_segments	ipa	tokens	language_id	branch
person		mə səŋ˧	mə səŋ˧	mə səŋ	m ə + s ə ŋ	miʔ xan	m i ʔ + x a n	Aasen	Patkaian
person		meʔ˧		meʔ	m e ʔ	miʔ	m i ʔ	Bote	Patkaian
person		miʔ˧		miʔ	m i ʔ	miʔ	m i ʔ	Chamkok	Patkaian
person		mit sɑ˧		mit sɑ	m i t + s a	mit sa	m i t + s a	Champhang	Patkaian
person	mət			mət	m ə t	mət	m i ʔ	ChangC	Patkaian
person		miːʔ˧		miʔ	m i ʔ	miʔ	m i ʔ	CholimHulawng	Patkaian
person			miʔ	miʔ	m i ʔ	miʔ	m i ʔ	CholimJotinKai	Patkaian
person		kʰau [ɲak]		kʰau ɲak	kʰ a u + ɲ a k	kʰau ɲa	kʰ i u + ɲ a ʔ	Chuyo	Patkaian
person		miʔ		miʔ	m i ʔ	miʔ	m i ʔ	DungiNS	Patkaian
person		miʔ˧		miʔ	m i ʔ	miʔ	m i ʔ	Gaji	Patkaian
person		nuk ŋɑ˨		nuk	n u k	nuk	n u k	Gaqha	Patkaian
person		hau [ɲak]		hau ɲak	h a u + ɲ a k	kʰau ɲa	kʰ i u + ɲ a ʔ	Gaqkat	Patkaian
person			miʔ$_4$	miʔ	m i ʔ	miʔ	m i ʔ	Gaqlun	Patkaian
person		miʔ˧		miʔ	m i ʔ	miʔ	m i ʔ	Gawkchung	Patkaian
person		mai		**mai**	m a i	miʔ	m i ʔ	Gongwan	Patkaian
person		maiʔ		**maiʔ**	m a i ʔ	miʔ	m i ʔ	Gongwan	Patkaian
person			maih	maiʔ	m a i ʔ	miʔ	m i ʔ	HahchengMulong	Patkaian
person		xu˧ ɲak		xu ɲak	x u + ɲ a k	kʰau ɲa	kʰ i u + ɲ a ʔ	Haqsik	Patkaian
person		kʰon˧ɲək		kʰon ɲək	kʰ o n + ɲ ə k	kʰau ɲa	kʰ i u + ɲ a ʔ	Karyaw	Patkaian
person		maiʔ		maiʔ	m a i ʔ	miʔ	m i ʔ	HahchengNS	Patkaian
person	miʔ			miʔ	m i ʔ	miʔ	m i ʔ	HakhunKB	Patkaian
person	miʔ			miʔ	m i ʔ	miʔ	m i ʔ	HakhunKB	Patkaian

Matrices

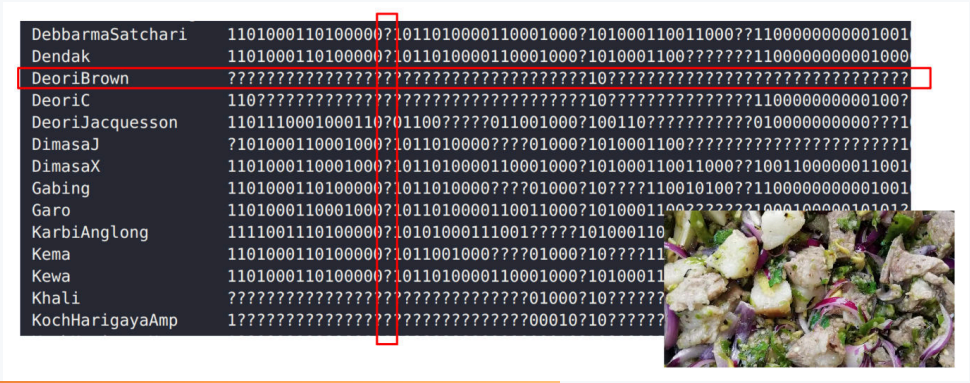
In many cases, data are coded essentially in CLDF or other flat tabular data. Then, with software such as LingPy, a matrix can be created, which looks like this.

```
DebbarmaSatchari    1101000110100000?10110100001100010007101000110011000??1100000000001001
Dendak              1101000110100000?10110100001100010007101000110000???????1100000000001000
DeoriBrown          ????????????????????????????????????????107????????????????????????????
DeoriC              110????????????????????????????????????107????????????????1100000000100?
DeoriJacquesson     1101110001000110?01100?????011001000?100110??????????0100000000???1
DimasaJ             ?1010001100010007101101000???701000?1010001100?????????????????????1
DimasaX             1101000110001000?10110100001100010007101000110011000??1001100000011001
Gabing              1101000110100000?10110100001???701000?107???1100101007?1100000000001001
Garo                1101000110001000?10110100001100110007101000110000???????1000100000101?
KarbiAnglong        1111001110100000?10101000011100?????101000110010001011?????????????1000
Kema                1101000110100000?10110010007???701000?107???1100110007?1100000000001007
Kewa                1101000110100000?10110100001100010007101000110011000??1100000000001001
Khali               ????????????????????????????????701000?107????????????????1100000000100?
KochHarigayaAmp     1?????????????????????????????????00010?107????????????????1100000000100?
```

Matricies

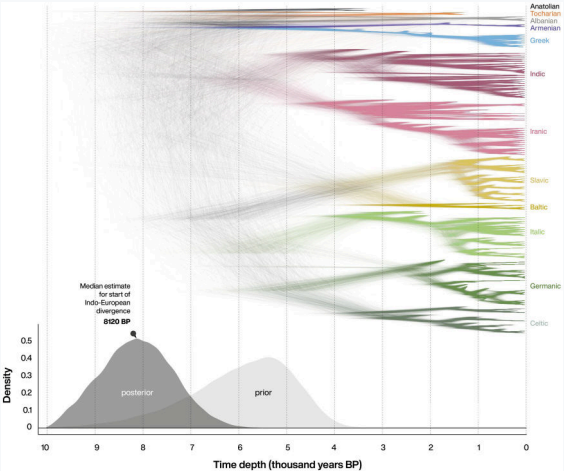
Having good coverage of concepts per language and languages per concept is important. Here, ? means a form simply wasn't given for that language/concept pair.

Ideally, every language would have every concept, but with outside sources this is often not possible while maintaining a large number of concepts.



Let’s see some trees

Indo-European



Heggarty, Paul, et al. "Language trees with sampled ancestors support a hybrid model for the origin of Indo-European languages." Science 381.6656 (2023): eabg0818.

Sino-Tibetan

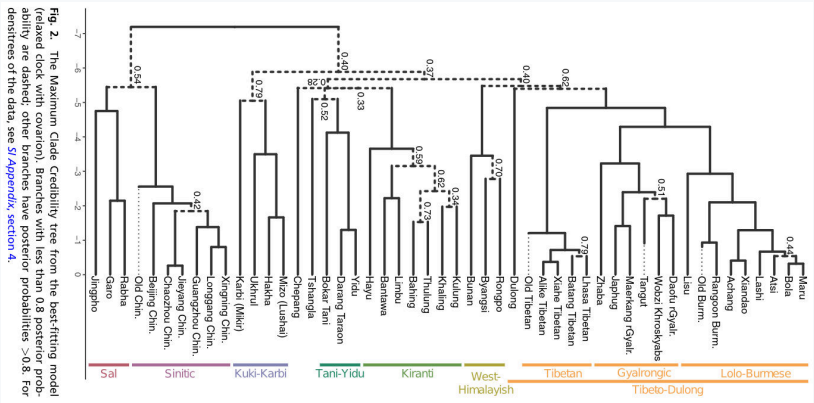
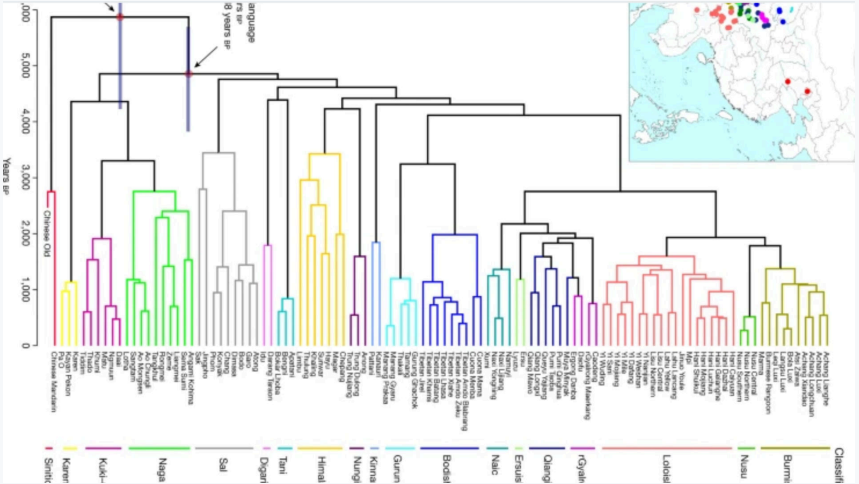


Fig. 2. The Maximum Clade Credibility tree from the best-fitting model (relaxed clock with covarion). Branches with less than 0.8 posterior probability are dashed; other branches have posterior probabilities >0.8. For denstrees of the data, see *SI Appendix*, section 4.

Sagart, Laurent, et al. "Dated language phylogenies shed light on the ancestry of Sino-Tibetan." Proceedings of the National Academy of Sciences 116.21 (2019): 10317-10322.

Sino-Tibetan, but different



Zhang, Menghan, et al. "Phylogenetic evidence for Sino-Tibetan origin in northern China in the Late Neolithic." Nature 569.7754 (2019): 112-115.

Bantu

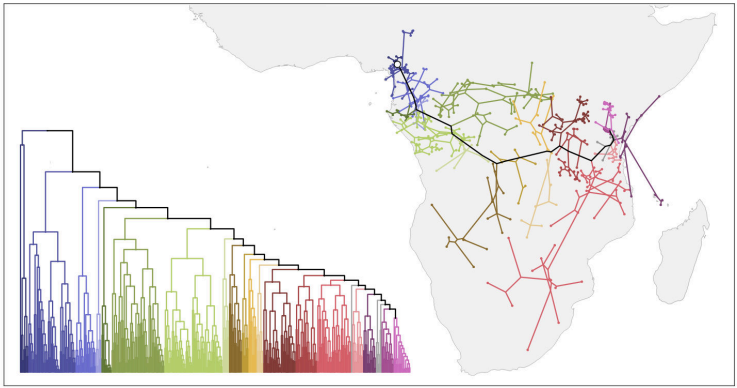


Figure 8. The phylogeny and spatial spread of the Bantu languages according to [12]. The colours mark the clades, splitting off one after another from the backbone of the expansion.

Neureiter, Nico, et al. "Can Bayesian phylogeography reconstruct migrations and expansions in linguistic evolution?." Royal Society open science 8.1 (2021): 201079.

Issues

Biases?

If you work with a language a lot, you probably have an idea anyway of what the likely branching events were.
However, through computational approaches, we can take the same type of data we'd develop such intuitions on, but on a much larger scale, either by including more languages, more data points, or both.

Rather than us having to keep all the details in mind at once and mentally work out likelihood of relatedness, by using computational tools we can introduce scientific reproducible into our analyses.

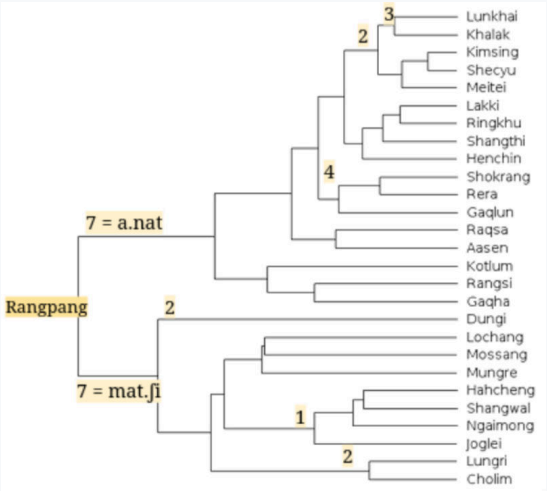
It also helps limit **some** of our biases (but of course never all).

Biases

Here numbers 1-4 represent previously proposed subgbroups of the Rangpang languages. A lexical analysis of around 50 words came up with different groupings. The main branching event also corresponds closely to which word for SEVEN they use in each language.

With 20 words, you may pick SEVEN since you recognise signal in it. You might miss "fiddle-head fern" dismissed as less basic.

When choosing only a few words, we often choose based on what we already expect, thus possibly missing important additional information.

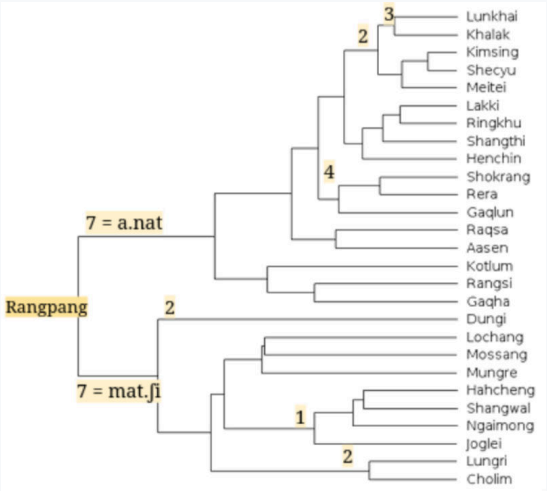


Biases

One major benefit of computational methods is that by opening up a much larger potential data set, our biases around such supposedly important items can be greatly reduced.

Instead of picking 20 words, maybe we pick 200. Instead of 4 languages, 40.

There are practical limits (your time and sanity for example) but we can massively expand our effectiveness with computer assistance.



Ensuring data quality

It's also important to remember that **the output is only as good as the data you feed it.**

You still need the expertise to know what you're looking at, and how to code the data.

As an example, how many etymological roots are in the data to the right?

Garbage in garbage out

It's also important to remember that **the output is only as good as the data you feed it.**

You still need the expertise to know what you're looking at, and how to code the data.

As an example, how many etymological roots are in the data to the right?

language	form
Chamkok	hu ki
Jiingi	ko ko
Rangsi	ge
Joglei	hi xeŋ
Kimsing	ku ku
Lainong	zai
Wancho	ki
Muishaung	ɣɨ he

Garbage in garbage out

These methods are not a magic bullet. Nor are they a replacement for linguistic expert iese.

Jiingi & Kimsing are borrowed from Assamese /kukur/ কুকুৰ

/ki~hi~ge/ are from **ɣɯi* (cf Chinese 狗 *gǒu*)

xeŋ < **hen*, a formerly productive plural marker that's been fossilised in Muishaung and Joglei.

A proper analysis requires that
all of these are coded accordingly.

language	form	
Chamkok	hu ki	4 1
Jiingi	ko ko	6 6
Rangsi	ge	1
Joglei	hi xeŋ	1 2
Kimsing	ku ku	6 6
Lainong	zai	1
Wancho	ki	1
Muishaung	ɣɨ he	1 2

Fine, now what?

Moving forward

Once we have flat data with **cognates identified**, either via EDICTOR, LingPy or on our own, we can start preparing the nexus file. Nexus files are matricies of the data that can be read and processed by Bayesian tools, either MrBayes or BEAST (or others).

Taking Eastern Polynesian as an example, in the end we may end up with something like the following, with some additional work such as linking to CONCEPTICON...

Eastern Polynesian (from https://github.com/lingpy/)

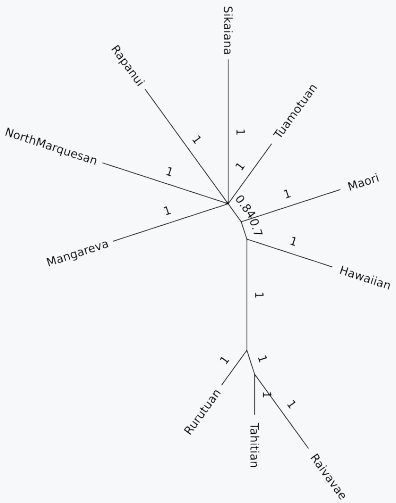
ID	DOCULECT	GLOTTOCODE	CONCEPTICON_ID	CONCEPT	FORM	SOURCE	COGID
#							
725	Maori	maor1246	1705	Eight	waru	Biggs-85-2005	663
1169	Tahitian	tahi1242	1705	Eight	va'u	Clark-173-2005	663
1595	Rapanui	rapa1244	1705	Eight	va'u	POLLEX	663
1853	Mangareva	mang1401	1705	Eight	varu	POLLEX	663
3076	Sikaiana	sika1261	1705	Eight	valu	POLLEX	663
3297	North_Marquesan	nort2845	1705	Eight	va'u	POLLEX	663
4395	Ra'ivavae	aust1304	1705	Eight	vaGu	Tamaititahio-1213-2015	663
4592	Tuamotuan	tuam1242	1705	Eight	varu	POLLEX	663
5101	Rurutuan	aust1304	1705	Eight	va?u	Meyer-128-2005	663
5614	Hawaiian	hawa1245	1705	Eight	walu	71458	663
#							
728	Maori	maor1246	493	Five	rima	Biggs-85-2005	2
1172	Tahitian	tahi1242	493	Five	pae	Clark-173-2005	1381
1173	Tahitian	tahi1242	493	Five	rima	Clark-173-2005	2
1598	Rapanui	rapa1244	493	Five	rima	POLLEX	2
1856	Mangareva	mang1401	493	Five	rima	POLLEX	2
3079	Sikaiana	sika1261	493	Five	lima	POLLEX	2
3300	North_Marquesan	nort2845	493	Five	'ima	POLLEX	2
4398	Ra'ivavae	aust1304	493	Five	pae	Tamaititahio-1213-2015	1381
4595	Tuamotuan	tuam1242	493	Five	rima	POLLEX	2
5104	Rurutuan	aust1304	493	Five	pae	Meyer-128-2005	1381
5617	Hawaiian	hawa1245	493	Five	lima	71458	2

```
#NEXUS

BEGIN DATA;
    DIMENSIONS NTAX=10 NCHAR=779;
    FORMAT DATATYPE=RESTRICTION SYMBOLS=01 GAP=- MISSING=?;
MATRIX
Hawaiian        10100101101101010011000111000110100001010000110110010100011000101000101101000000010100000000100000
Mangareva       11000101101100100111000101001010100100000011110010010111011010101010110010000100010000000101010000
Maori           10010101101100100111001101010010110000010100111010010000111000100001101011100000000110000000100000
NorthMarquesan  1????101111100l????11000101001010100010010000110011010100011001100100101101010000001100000001010010l
Rapanui         1????101101100101111001101????10101000111010110010010101011100101000101011000010000000110000001000
Raivavae        10001011101010110111000101100001110000000001011001001010001010001010100001000010000010001001000
Rurutuan        10001011101101010011000101100011110000000010110010110100010100101000101010001100101000000010001101
Sikaiana        1????101101100100111010001????10?100000100001100101001000101100001001001110100000000100000010100000
Tahitian        1????111101111110111000101101001110000000010110010010001010001010010100001000010100000000100100
Tuamotuan       1????101111100100110100101????10110000?????011101001111101100010111010110000010000110100000101100
;
END;

BEGIN MRBAYES;
    charset Eight = 1-1;
    charset Fifty = 2-5;
    charset Five = 6-7;
    charset Four = 8-8;
    charset I = 9-10;
    charset Nine = 11-11;
    charset One = 12-13;
    charset One_Hundred = 14-15;
    charset One_Thousand = 16-18;
    charset Seven = 19-19;
    charset Six = 20-21;
    charset Ten = 22-25;
    charset Three = 26-26;
```

Phylogenies

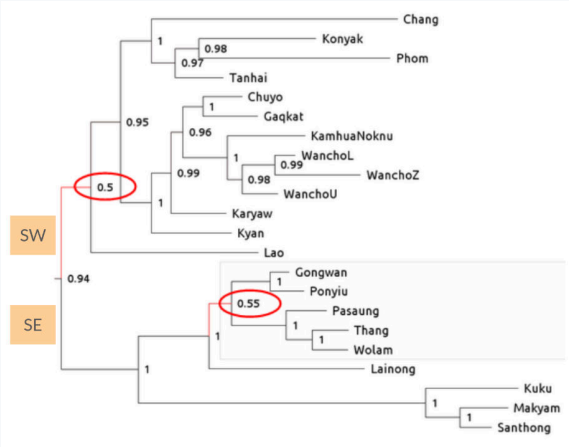
Running that through `MrBayes`, we may get something like this.

DOCULECT	GLOTTOCODE	CONCEPT	FORM	COGID
Maori	maor1246	Five	rima	2
Tahitian	tahi1242	Five	pae	1381
Tahitian	tahi1242	Five	rima	2
Rapanui	rapa1244	Five	rima	2
Mangareva	mang1401	Five	rima	2
Sikaiana	sika1261	Five	lima	2
North_Marquesan	nort2845	Five	’ima	2
Ra’ivavae	aust1304	Five	pae	1381
Tuamotuan	tuam1242	Five	rima	2
Rurutuan	aust1304	Five	pae	1381
Hawaiian	hawa1245	Five	lima	2



AI is not stealing our jobs

A puzzle

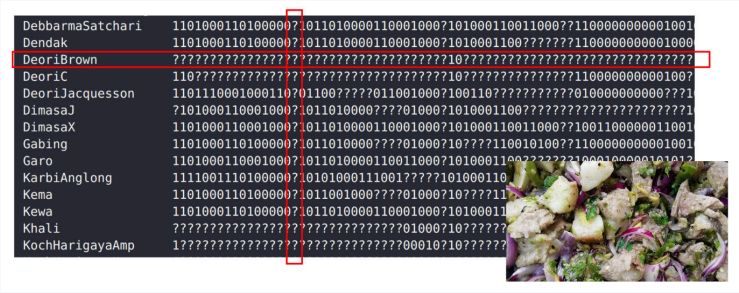


Knowing what we do about posterior probabilities, what might these circled numbers mean?

Low probabilities

Low posterior probabilities can come from a number of sources:

- unaccounted-for language contact
- incorrect cognate identification
- very week coverage of data for a given language or languages
- data which doesn't actually carry much phylogenetic signal



Low probabilities

Low posterior probabilities can come from a number of sources:

- unaccounted-for language contact
- incorrect cognate identification
- very week coverage of data for a given language or languages
- data which doesn't actually carry much phylogenetic signal (next slide)

```
Overlay plot for both runs:
(1 = Run number 1; 2 = Run number 2; * = Both runs)

+-------------------------------------------------------------+ -3428.71
|1           1   2                                            |
|                                                          2   |
|   1          *2    2          1    12       1  *   1    2  |
|     2  1                    2  22 2 2          1     11     |
|  2          1    1           2 1          2     2     12    2|
| * 2    2   22       1   2         2   1   2       1 1       |
|  1 1   1  1    11 2  12  1        1   12  12  1  22         |
|      2   2     2   2  1     1        2  1 1      1222  1 2  |
|     *            11    111 122    2    1 2  2              |
|2     1  1   1    2 1 2            1   1    *            1 1|
|      2  22                  11            2               |
|                                                             |
|                        2                              2    |
|                                                         1   |
|                          2                                  |
+------+-----+-----+-----+-----+-----+-----+-----+-----+-----+ -3431.87
^                                                             ^
170000                                                        1700000


Estimated marginal likelihoods for runs sampled in files
   "ep.nex.run1.p" and "ep.nex.run2.p":
   (Use the harmonic mean for Bayes factor comparisons of models)

   (Values are saved to the file ep.nex.lstat)

Run   Arithmetic mean   Harmonic mean
---------------------------------------
  1      -3426.74         -3438.31
  2      -3426.43         -3437.68
---------------------------------------
TOTAL    -3426.57         -3438.04
---------------------------------------


Model parameter summaries over the runs sampled in files
   "ep.nex.run1.p" and "ep.nex.run2.p":
   Summaries are based on a total of 3062 samples from 2 runs.
   Each run produced 1701 samples of which 1531 samples were included.
   Parameter summaries saved to file "ep.nex.pstat".

                                  95% HPD Interval
```