The Bayesian phylogenetics of grammar

Gerhard Jäger

Tübingen University

University of Amsterdam, April 4, 2018

WORDS BONES GENES TOOLS
Tracking Linguistic, Cultural, and Biological Trajectories of the Human Past

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

DFG

erc
European Research Council
Established by the European Commission

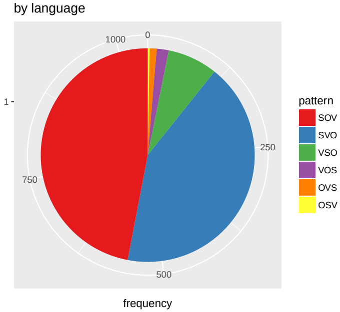
Major word orders

Statistics of major word order distribution

- data: WALS intersected with ASJP
- 1,045 languages, 211 lineages, 32 families with at least 5 languages

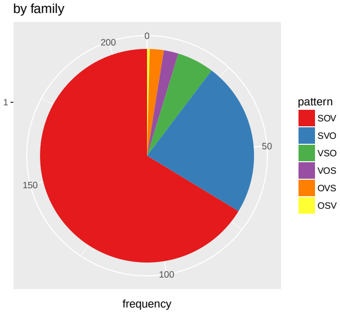
Raw numbers

SOV	SVO	VSO	VOS	OVS	OSV
491	442	79	19	11	3
47.0%	42.3%	7.6%	1.8%	1.1%	0.3%



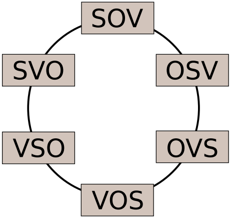
Weighted by lineages

SOV	SVO	VSO	VOS	OVS	OSV
139.1	49.3	11.8	4.7	4.5	0.8
66.3%	23.4%	5.6%	2.2%	2.1%	0.4%



Previous approaches

- Gell-Mann and Ruhlen (2011):
  - Proto-world was SOV
  - general pathway: SOV $\rightarrow$ SVO $\leftrightarrow$ VSO/VOS
  - minor pathway: SOV $\rightarrow$ OVS/OSV
  - exceptions due to diffusion
- Ferrer-i-Cancho (2015):



  - permutation circle
  - transition probability inversely related to path length

Previous approaches

- Maurits and Griffiths (2014):
  - Bayesian rate estimation, based on five families and NJ-trees

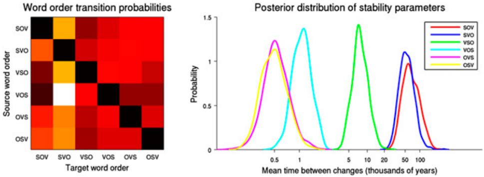
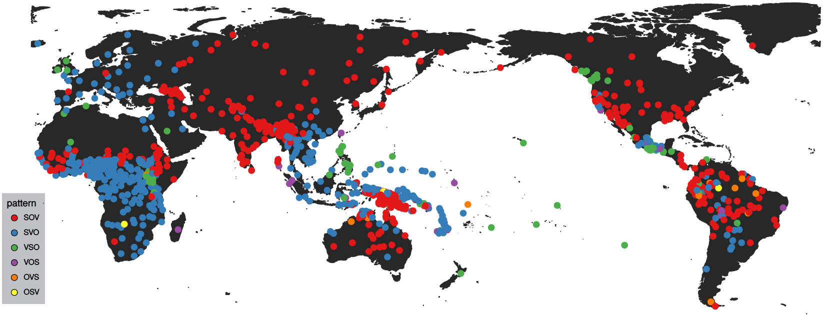


Fig. 1. Results of inferring a single mutation matrix Q for all six language families. (*Left*) Heat map showing the transition probabilities between word orders. Higher intensity (white, yellow) indicates more-probable transitions compared with lower intensity (red, brown), so SOV is most likely to transition to SVO and SVO to SOV. VSO is much more likely to transition to SVO than to SOV. (*Right*) Inferred posterior distributions of stability parameters for each word order. The horizontal axis shows the stability parameter, expressed as the mean time between transitions; i.e., higher values indicate a more stable word order.

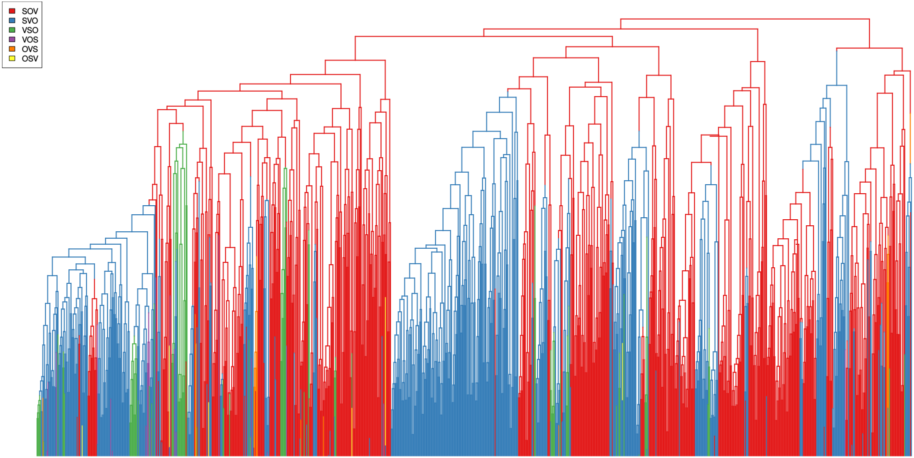
Phylogenetic non-independence

- languages are phylogenetically structured
- if two closely related languages display the same pattern, these are not two independent data points

⇒ we need to control for phylogenetic dependencies



Phylogenetic non-independence



Phylogenetic non-independence

Maslova (2000):

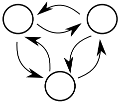
"If the A-distribution for a given typology cannot be assumed to be stationary, a distributional universal cannot be discovered on the basis of purely synchronic statistical data."

*"In this case, the only way to discover a distributional universal is to **estimate transition probabilities** and as it were to 'predict' the stationary distribution on the basis of the equations in (1)."*

The phylogenetic comparative method

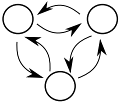
Modeling language change

Markov process



Modeling language change

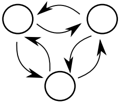
Markov process



Phylogeny

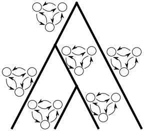
Modeling language change

Markov process



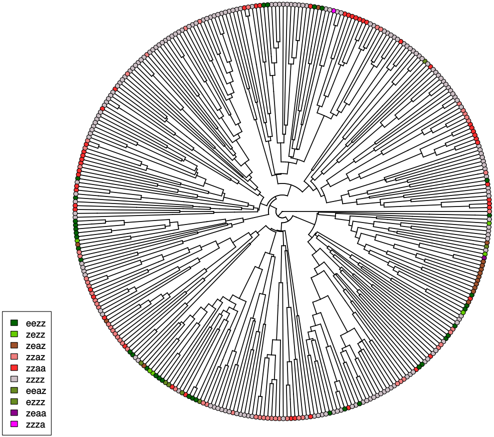
Phylogeny

Branching process



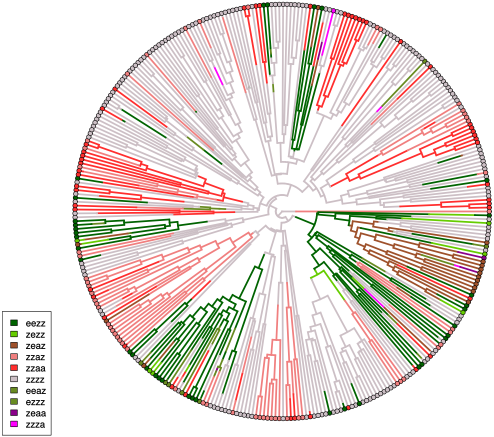
Estimating rates of change

- if phylogeny and states of extant languages are known...



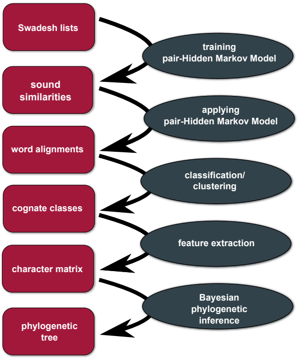
Estimating rates of change

- if phylogeny and states of extant languages are known...
- ... transition rates and ancestral states can be estimated based on Markov model

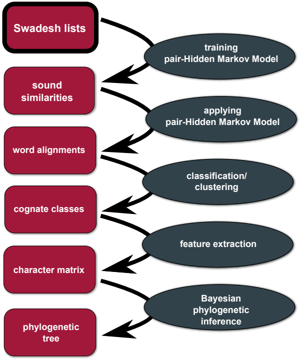


Inferring a world tree of languages

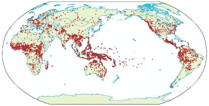
From words to trees



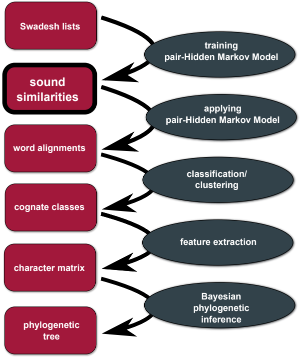
From words to trees

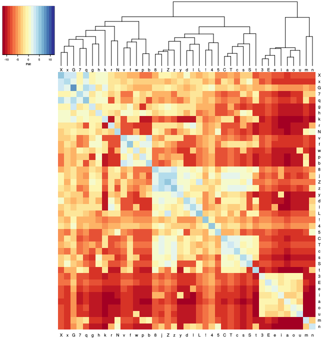


concept	Latin	English
I	ego	Ei
you	tu	yu
we	nos	wi
one	unus	w3n
two	duo	tu
person	persona, homo	pers3n
fish	piskis	fiS
dog	kanis	dag
louse	pedikulus	laus
tree	arbor	tri
leaf	foly∼u*	lif
skin	kutis	skin
blood	saNgw∼is	bl3d
bone	os	bon
horn	kornu	horn
ear	auris	ir
eye	okulus	Ei

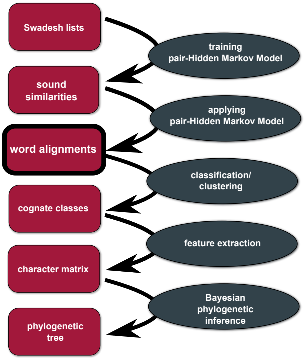


From words to trees



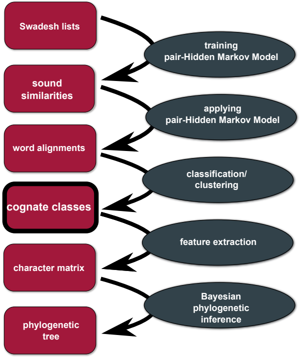


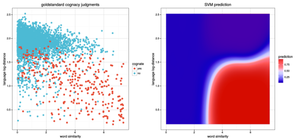
From words to trees



Language	*fish:z*	*tongue:1*	*smoke:1*
Abui-Atangmelang	-af-u		
Abui-Fuimelang	-af-u	tal-i-fi--	
Adang	aab--	tal-E-b---	awai--b-a-n-o-7o-
Blagar-Bakalang	-ab--	--j-e-bur-	--ad--b-a-n-aNka-
Blagar-Bama	aab--	teg-e-bur-	------b-e-n-a-xa-
Blagar-Kulijahi	-ab--	tej-e-bur-	------b-e-n-aNka-
Blagar-Nule	aab--	tej-e-bur-	--ad--b-e-n-aNka-
Blagar-Tuntuli	aab--	tej-e-bur-	a-adgeb-a-n-a-q--
Blagar-Warsalelang	-ab--	tel-e-bur-	a-ad--b-a-n-a-x--
Bunaq			------b-o-t-o-h--
Deing	haf--		------buu-n------
Hamap	7ab--	nar-ø-buN-	------b-a-n-o-7--
Kabola	hab--	tal-e-b---	awal--b-e-n-e-7o-
Kaera-Padangsul	-ab--	talee-b---	a-ad--b-e-naa-x--
Kafoa	-afUi	tal-i-p---	------f-o-n-a----
Kamang	-ap-i	nal---pu--	------p-u-n----a-
Kiraman	-Eb--	nal-i-bar-	--ar--b-a-n-o-kan
Klon	-eb-i	gel-E-b---	--ed-ab-o-n------
Kui	-eb--	tal-i-ber-	--ar--b-o-n-o-k--
Kula	-ap-i	-il-I-p---	------p---n-ekka-
Nedebang	aaf-i	gel-e-fu--	--ar-ab-u-n------
Reta	aab--	nal-e-bul-	a-ad--b-o-n-a----
Sar-Adiabang	haf--	--p-e-fal-	--ar--buu-n------
Sar-Nule	haf--	nal-e-faj-	
Sawila	-ap-i	gal-impuru	------p-u-n-a-ka-
Teiwa-Madar	xaf--	gel-i-vi--	------buu-n------
Wersing	-ap-i	nej-e-bur-	--ad-ap-u-n-a-k--
Wpantar	hap--	nal-e-bu--	------b-unn-a----

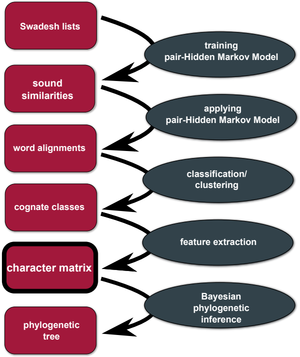
From words to trees





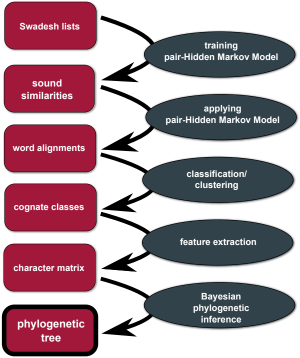
	English	Spanish	Modern Greek	Standard German
I	Ei:A	yo:B	exo:C	iX:D
you	yu:A	ustet:B, tu:C	esi:D	du:E
we	wi:A	nosotros:B	emis:C	vir:A
one	w3n:A	uno:B	enas:C, ena:C	ains:D
two	tu:A	dos:B	8y~o:C, 8io:D	cvai:E
person	pers3n:A	persona:A	an8~ropos:B	mEnS:C
fish	fiS:A	peskado:A, pes:A	psari:B	fiS:A
dog	dag:A	pero:B	sTili:C, sTilos:C	hunt:D
come	k3m:A	veni:B	erx~o:C	kh~om3n:A
sun	s3n:A	sol:B	ily~os:C, iLos:C	zon3:A
star	star:A	estreya:A	asteri:A, astro:A	StErn:A
water	wat3r:A	agw~a:B	nero:C	vas3r:A
stone	ston:A	piedra:B	petra:B	Stain:A
fire	fEir:A	fuego:B	foty~a:C	foia:D
path	pE8:A	senda:B	8romos:C	pf~at:A, vek:D
mountain	maunt3n:A	sero:B, monta5a:A	vuno:C, oros:D	bErk:E
full	ful:A	yeno:B	yematos:C, pliris:D	fol:A
new	nu:A	nuevo:A	neos:A, Tenury~os:B	noi:A
name	nem:A	nombre:A	onoma:A	nam3:A

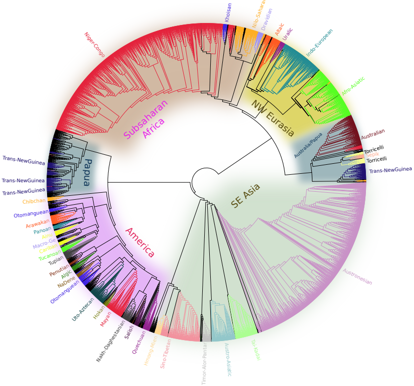
From words to trees




```
TNG.ENGAN.MAIBI                  10000000000000000000000000000000000000000000
TNG.ENGAN.POLE                   00000000000000000000000000000000000010000000
TNG.ENGAN.SAU                    00000000000000000000000000000000000000000000
TNG.ENGAN.YARIBA                 10000000000000000000000000000000000000000000
TNG.FASU.FASU                    00000000000000000000000000000000000010000000
TNG.FASU.NAMUMI                  00000000000000000000000000000000000000001000
TNG.FINISTERRE-HUON.AWARA        00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.BORONG       00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.BURUM        00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.BURUM_MIND   00000000000000000000000000000000001010000000
TNG.FINISTERRE-HUON.DEDUA        00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.HUBE         00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.KATE         00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.KOMBA        00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.KOSORONG     00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.MAPE         00000000000000000000000000000000001000000000
TNG.FINISTERRE-HUON.MAPE_2       00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.MIGABAC      00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.MINDIK       00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.MOMOLILI     00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.NABAK        00000000000000000000000000000000001000000000
TNG.FINISTERRE-HUON.NANKINA      00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.NEK          00000000000000000000000000000000000001000000
TNG.FINISTERRE-HUON.NUKNA        00000000000000000000000000000000000001000000
TNG.FINISTERRE-HUON.ONO          00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.SELEPET      00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.TIMBE        00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.TOBO         00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.WANTOAT      00000000000000000000000000000000000110000000
TNG.FINISTERRE-HUON.YOPNO        00000000000000000000000000000000000001000000
TNG.GOILALAN.AFOA                00000000000000000000000000000000000110000000
TNG.GOILALAN.KUNIMAIPA           00000000000000000000000000000000000010000000
TNG.GOILALAN.MAFULU              00000000000000000000000000000000000010000000
```

From words to trees





Estimating word-order transition patterns

Workflow

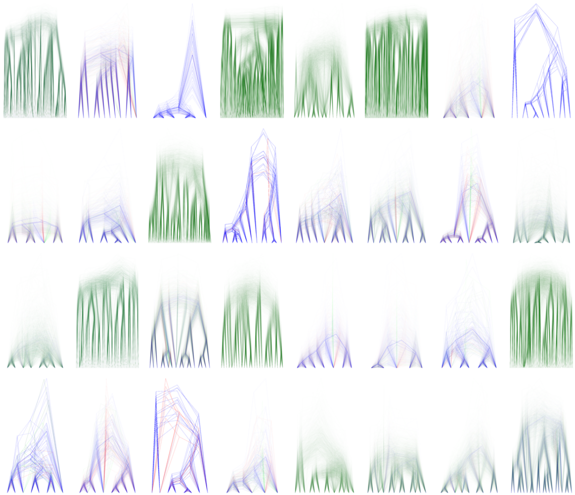
(data from all 32 families with $\geq$ 5 languages in data base; 778 languages in total)

- estimate posterior tree distributions with MrBayes for each family, using Glottolog as constraint tree
- test whether universal or lineage-specific model gives a better fit
- estimate transition rates with best model
- estimate stationary distribution of major word order categories
- apply *stochastic character mapping* (SIMMAP; Bollback 2006)
- estimate expected number of mutations for each transition type

Estimating posterior tree distributions

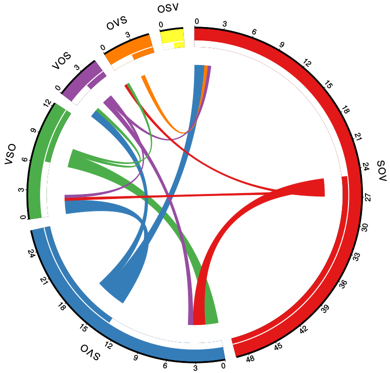
- using characters extracted from ASJP data (Jäger 2018)
- Glottolog as constraint tree
- Γ-distributed rates
- ascertainment bias correction
- relaxed molecular clock (IGR)
- uniform tree prior
- stop rule: 0.01, samplefreq=1000
- if convergence later than after 1,000,000 steps, sample 1,000 trees from posterior

Phylogenetic tree sample



Estimating transition rates

- totally unrestricted model, all 30 transition rates are estimed independently
- implementation using RevBayes (Höhna et al., 2016)



Reconstruction history with SIMMAP

- estimated frequency of mutations within the 32 families under consideration (posterior mean, 100 iterations)

	SOV	SVO	VSO	VOS	OVS	OSV
SOV	–	20.2	3.2	0.5	3.3	0.4
SVO	17.6	–	23.9	14.5	1.5	1.1
VSO	1.5	19.9	–	2.5	1.8	0.4
VOS	1.0	5.4	2.3	–	0.9	0.3
OVS	2.8	0.9	0.6	0.4	–	0.2
OSV	0.5	0.5	0.4	0.3	0.5	–

Refining the model with Reversibly Jump MCMC

- Estimating 30 transition rates is a tall order, given that the data possibly only reflect about 130 transition events
- hand-crafted sub-model construction: time consuming, subjective and error prone
- solution: posterior sampling over sub-models using *Reversible Jump Markov Chain Monte Carlo* (RJMCMC, Green 1995)

RJMCMC

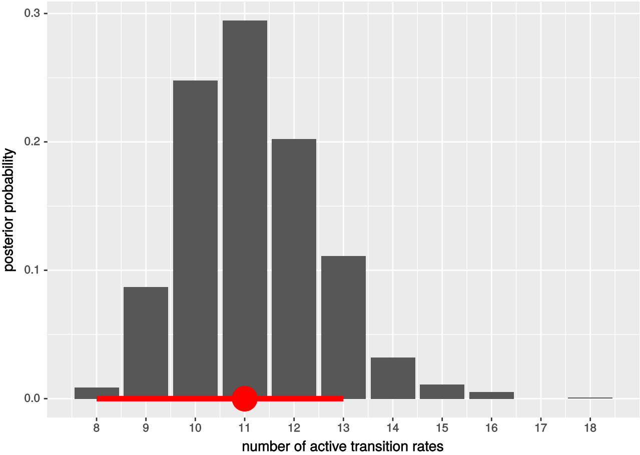
RJMCMC assumes a prior distribution over sub-models (where some transition rates are set to 0) and simultaneously samples from the set of sub-models and the parameter spaces of the sub-models.

Model comparison

model	marginal likelihood	AICM
lineage-specific	-423.0 ± 0.08	926.4 ± 0.5
circular GTR	-420.0 ± 1.72	851.7 ± 1.6
circular	-414.2 ± 0.72	851.6 ± 2.1
RJ/GTR	-413.4 ± 2.96	855.9 ± 4.7
unrestricted	-406.7 ± 0.78	846.4 ± 2.5
unrestricted GTR	-404.4 ± 0.89	843.5 ± 3.6
RJ	-398.0 ± 0.57	827.2 ± 2.1

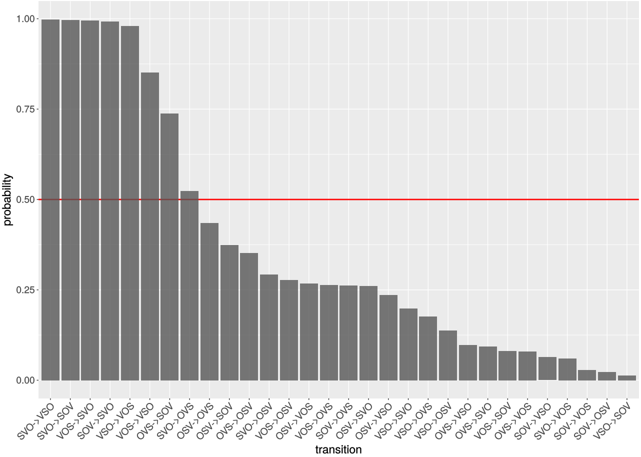
Refining the model with Reversibly Jump MCMC

Number of active transition rates: posterior distribution



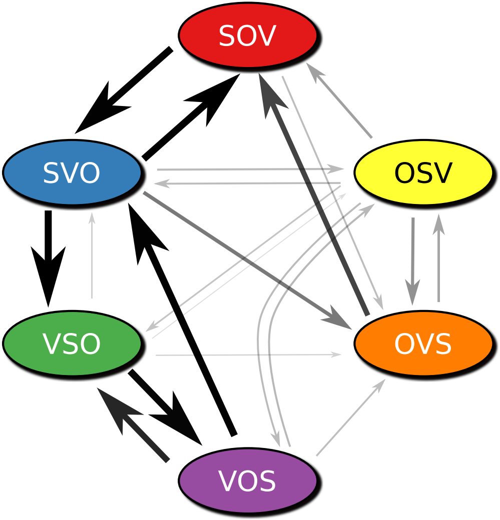
Refining the model with Reversibly Jump MCMC

Probabilities of active transition rates: posterior distribution



Refining the model with Reversibly Jump MCMC

Probabilities of active transition rates: posterior distribution



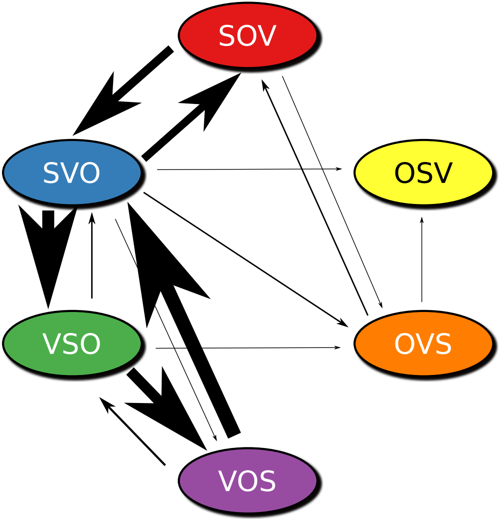
Reconstruction history with SIMMAP

- estimated frequency of mutations within the 32 families under consideration (posterior mean, 99 iterations)

	SOV		SVO		VSO		VOS		OVS		OSV	
SOV	–		23.1	[14; 30]	0.5	[0; 6]	0.1	[0; 0]	1.9	[0; 9]	0.1	[0; 0]
SVO	20.3	[16; 28]	–		33.0	[20; 45]	2.2	[0; 29]	3.4	[0; 11]	1.2	[0; 7]
VSO	0.0	[0; 0]	3.8	[0; 25]	–		29.7	[0; 46]	1.5	[0; 9]	0.5	[0; 4]
VOS	0.1	[0; 0]	38.3	[19; 54]	6.2	[0; 13]	–		0.9	[0; 5]	0.4	[0; 2]
OVS	4.0	[0; 10]	0.5	[0; 3]	0.9	[0; 6]	0.2	[0; 1]	–		1.1	[0; 6]
OSV	0.7	[0; 6]	0.3	[0; 3]	0.4	[0; 3]	0.6	[0; 5]	0.9	[0; 7]	–	

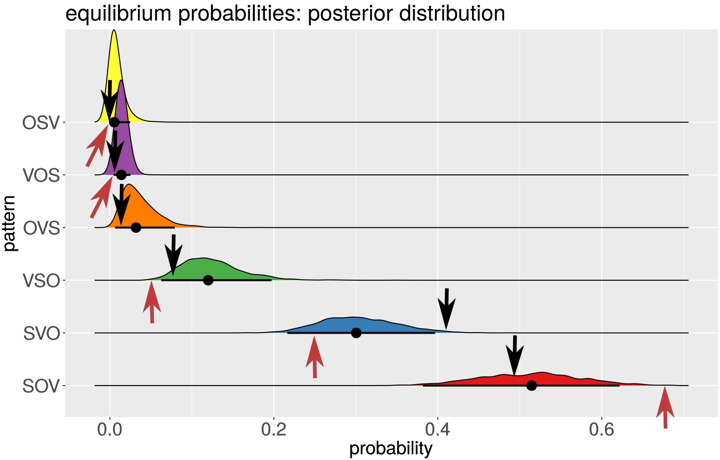
Reconstruction history with SIMMAP

Expected frequencies of transitions: posterior mean



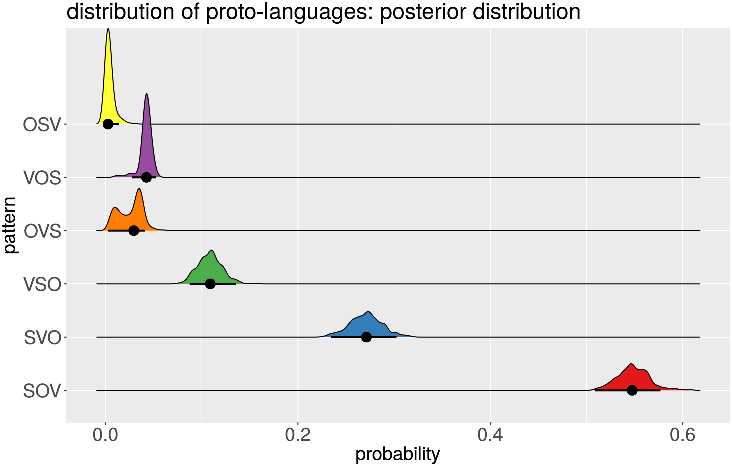
Posterior distributions

Empirical vs. estimated distribution



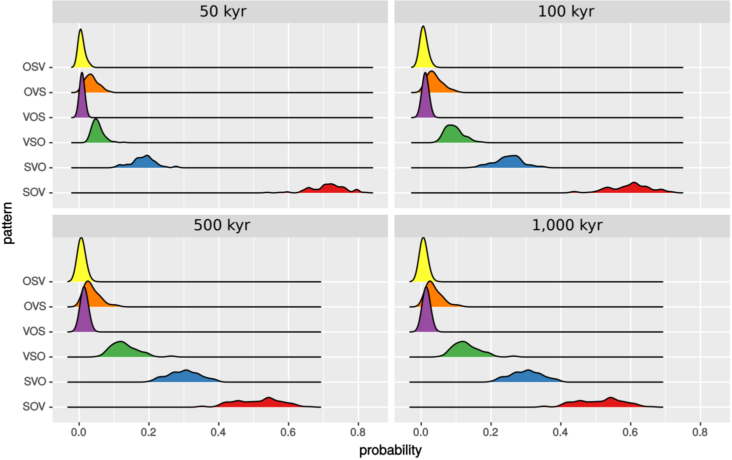
Posterior distributions

Expected distribution of Proto-languages



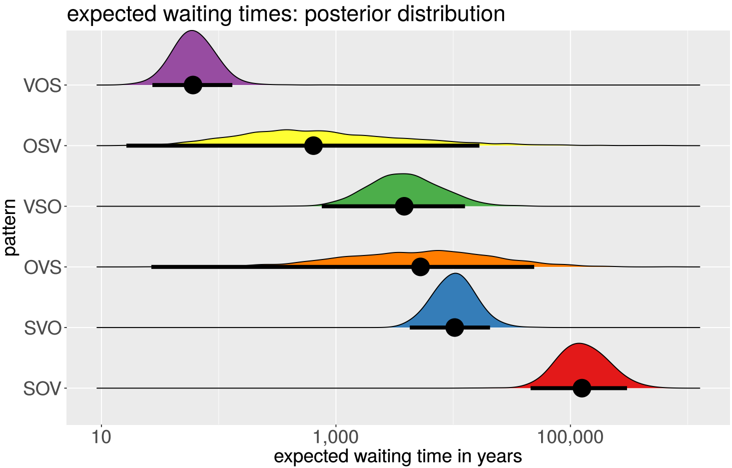
Posterior distributions

Expected probabilities of Proto-World, given that we can demonstrate SOV for all proto-languages



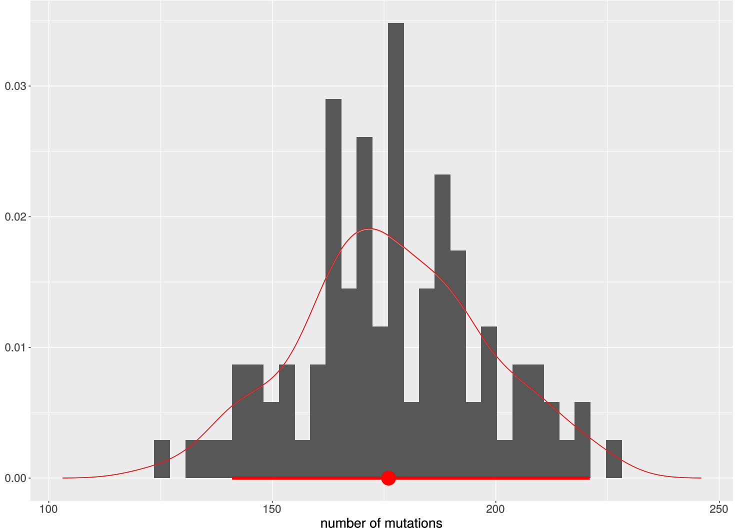
Posterior distributions

Waiting times

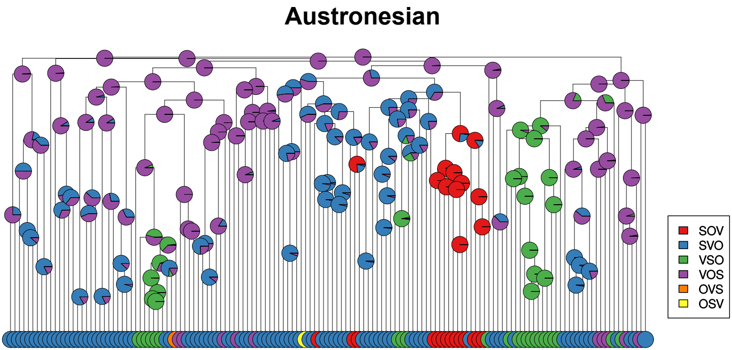


Posterior distributions

Number of state changes

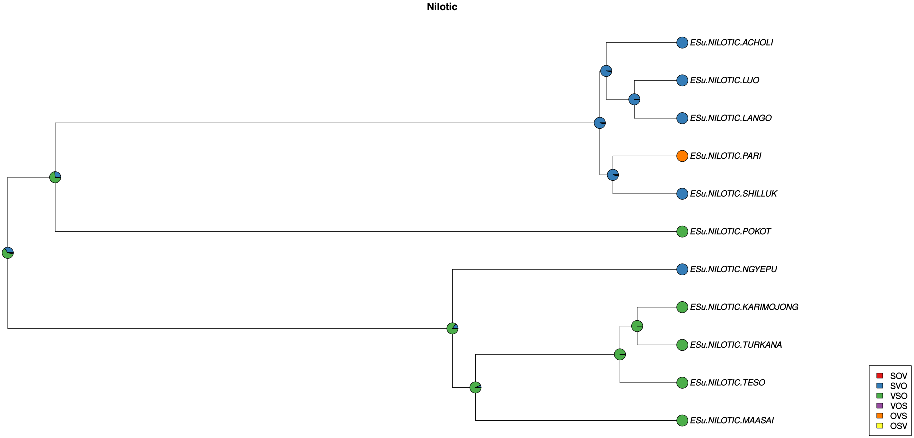


Ancestral state reconstruction



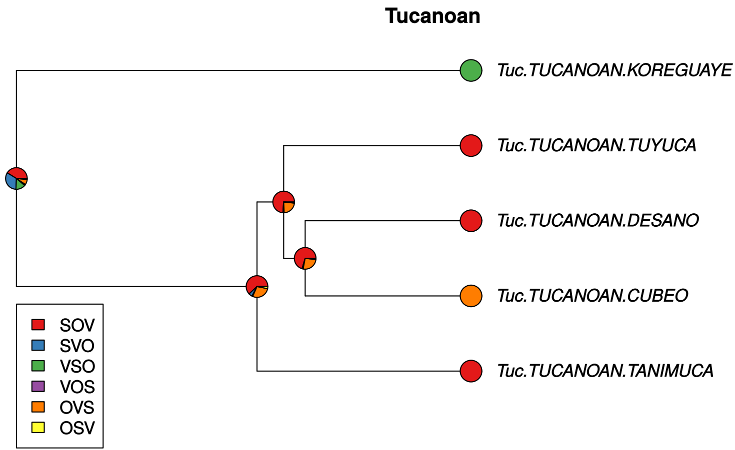
Examples for unexpected transitions

SVO → OVS



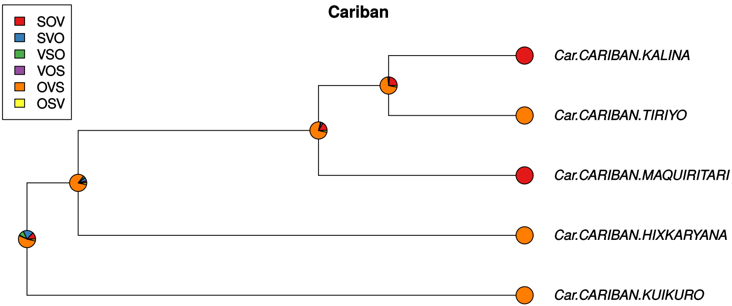
Examples for unexpected transitions

OVS → SOV



Examples for unexpected transitions

OVS → SOV



Summary

- no evidence for general preference of SOV $\rightarrow$ SVO over the reverse
- SVO is currently over-represented due to recent spread of Austronesian and Atlantic-Congo, but not excessively so
- multiple counter-evidence to Ramon-i-Ferrer's and Gell-Mann & Ruhlen's models

Jonathan P. Bollback. SIMMAP: stochastic character mapping of discrete traits on phylogenies. *BMC Bioinformatics*, 7(1):88, 2006.

Ramon Ferrer-i-Cancho. Kauffman's adjacent possible in word order evolution. arXiv preprint arXiv:1512.05582, 2015.

Murray Gell-Mann and Merritt Ruhlen. The origin and evolution of word order. *Proceedings of the National Academy of Sciences*, 108(42):17290–17295, 2011.

Peter J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4): 711–732, 1995.

Sebastian Höhna, Michael J. Landis, Tracy A. Heath, Bastien Boussau, Nicolas Lartillot, Brian R. Moore, John P. Huelsenbeck, and Frederik Ronquist. Revbayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Systematic biology*, 65(4):726–736, 2016.

Gerhard Jäger. Phylogenetic inference from word lists using weighted alignment with empirically determined weights. *Language Dynamics and Change*, 3(2):245–291, 2013.

Gerhard Jäger. Support for linguistic macrofamilies from weighted sequence alignment. *Proceedings of the National Academy of Sciences*, 112(41):12752–12757, 2015. doi: 10.1073/pnas.1500331112.

Gerhard Jäger. Global-scale phylogenetic linguistic inference from lexical resources. arXiv:1802.06079, 2018.

Gerhard Jäger and Søren Wichmann. Inferring the world tree of languages from word lists. In S. G. Roberts, C. Cuskley, L. McCrohon, L. Barceló-Coblijn, O. Feher, and T. Verhoef, editors, *The Evolution of Language: Proceedings of the 11th International Conference (EVOLANG11)*, 2016. Available online: http://evolang.org/neworleans/papers/147.html.

Elena Maslova. A dynamic approach to the verification of distributional universals. *Linguistic Typology*, 4(3):307–333, 2000.

Luke Maurits and Thomas L. Griffiths. Tracing the roots of syntax with Bayesian phylogenetics. *Proceedings of the National Academy of Sciences*, 111(37):13576–13581, 2014.

Søren Wichmann, Eric W. Holman, and Cecil H. Brown. The ASJP database (version 17). http://asjp.clld.org/, 2016.