

A stochastic model for binary trait evolution, and its application to
estimating a phylogeny for Semitic from lexical trait data

Data and Registration *Kitchen et al 2009*

25 Semitic languages (Hebrew, Akkadian, Amharic,...), 674 lexical traits

	all	bite	kill	knee	we
Akkadian	kalu	naʃaku	daku	birku	ninu
Aramaic	kul	nkhat	k"t'al	burka	ħnan
Ugaritic	kl	ntk	hrg	brk	?
Ogaden Arabic	kull	9ajj	qatal	rukba	niħin
Jibbali	kal	ʃaPar	leteγ	eerk	inħan
Tigre	k ^w illu	naksa	katla	-	ħina:
Amharic	hullu	nεkkεε	gεddεε	-	inna

	all	bite	kill	knee	we					
Akkadian	A	B	D	A	A	1	0100	000100	10	1
Aramaic	A	A	A	A	A	1	1000	100000	10	1
Ugaritic	A	C	C	A	?	1	0010	001000	10	?
Ogaden Arabic	A	A	F	B	A	1	1000	000001	01	1
Jibbali	A	D	E	A	A	1	0001	000010	10	1
Tigre	A	A	A	-	A	1	1000	100000	--	1
Amharic	A	B	B	-	A	1	0100	010000	--	1

Data and Registration *Kitchen et al 2009*

25 Semitic languages (Hebrew, Akkadian, Amharic,...), 674 lexical traits

	all	bite	kill	knee	we
Akkadian	kalu	naʃaku	daku	birku	ninu
Aramaic	kul	nkhat	k" t'al	burka	ħnan
Ugaritic	kl	ntk	hrg	brk	?
Ogaden Arabic	kull	9ajj	qatal	rukba	niħin
Jibbali	kal	ʔaPar	leteγ	eerk	inħan
Tigre	k ^w illu	naksa	katla	-	ħina:
Amharic	hullu	nεkkεε	gεddεε	-	inna

	all	bite	kill	knee	we					
Akkadian	A	B	D	A	A	1	01	0	1	1
Aramaic	A	A	A	A	A	1	10	1	1	1
Ugaritic	A	C	C	A	?	1	00	0	1	?
Ogaden Arabic	A	A	F	B	A	1	10	0	0	1
Jibbali	A	D	E	A	A	1	00	0	1	1
Tigre	A	A	A	-	A	1	10	1	?	1
Amharic	A	B	B	-	A	1	01	0	?	1

Generic trait-observation model

$$g = (E, t, k) \quad t = (t_1, \dots, t_{2L-1})$$

$$z = (z_1, \dots, z_N) \quad z_a = (\tau_a, i_a) \in [g]$$

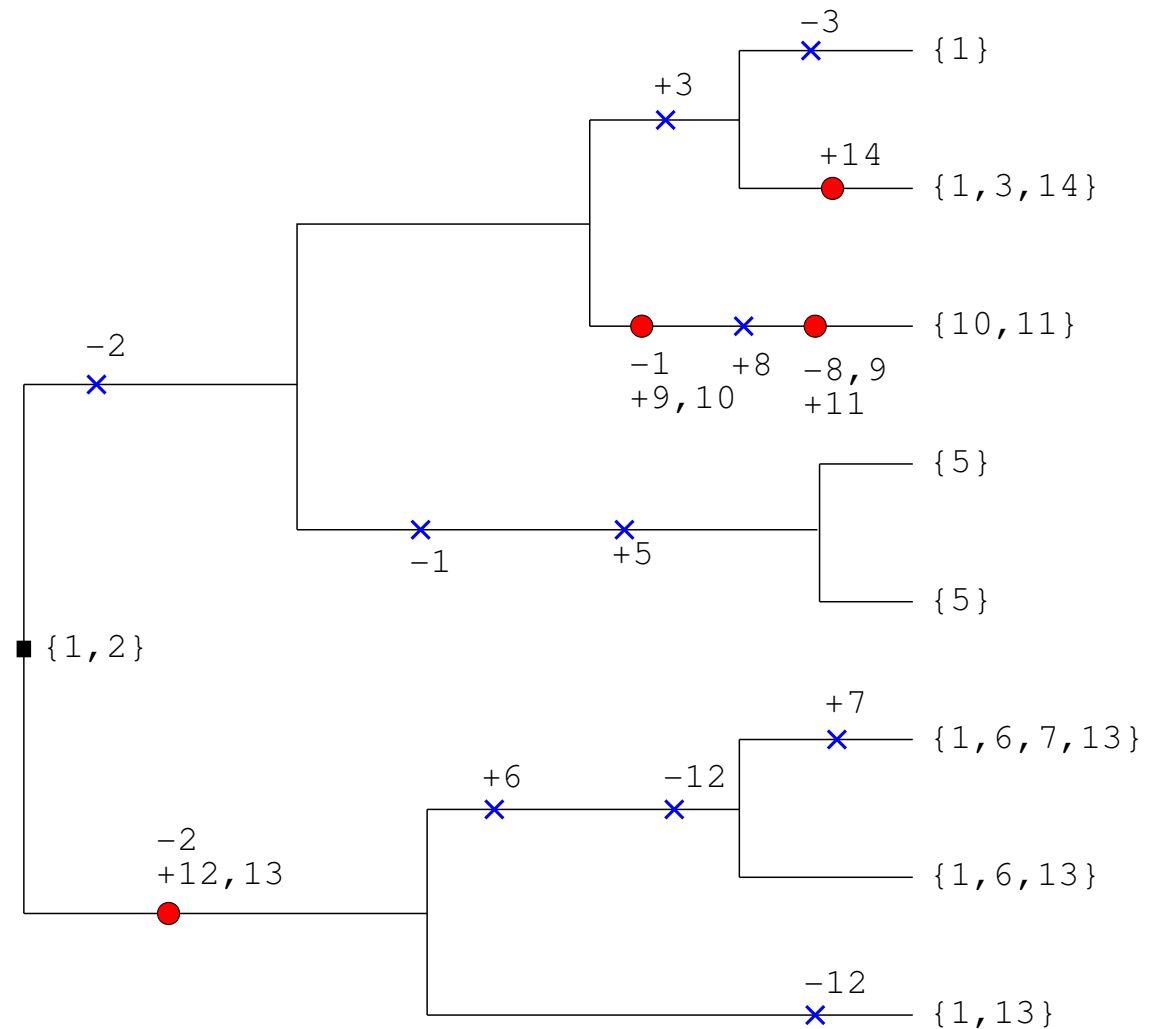
λ : set-element birth rate;

μ : set element death rate;

ρ : catastrophe rate, $k = (k_1, \dots, k_{2L-2})$

κ , death prob, $\nu = \lambda\kappa/\mu$ mean births

ξ_i : probability element a visible at leaf i

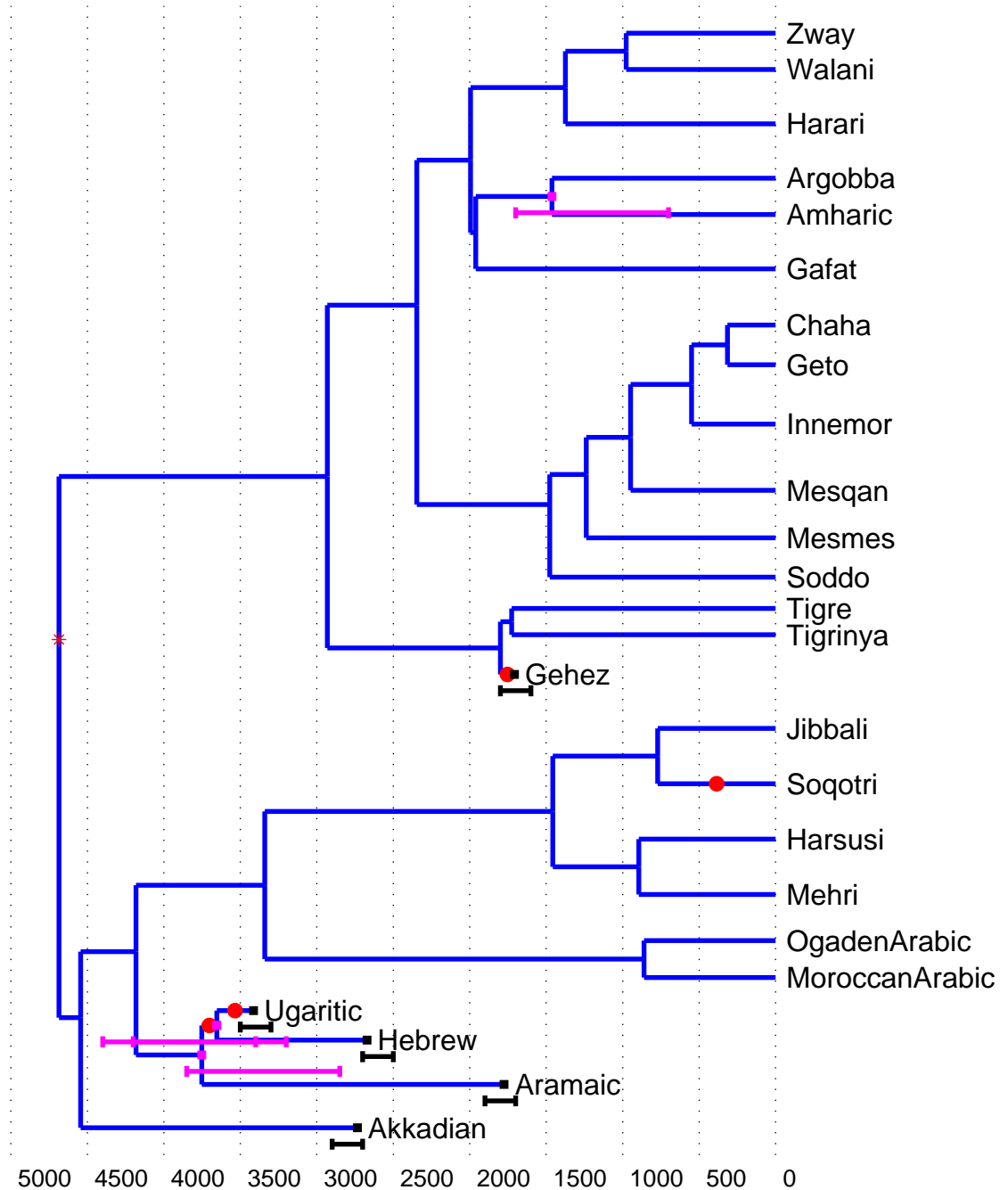


Calibration Data

For sources see citations in
Kitchen et al 2009

	Sampled	Branched
Semitic	-	4350-8000(?)
Akkadian	2700-2900	-
Biblical Aramaic	1700-1900	2850-3850
Ancient Hebrew	2500-2700	3200-4200
Ugaritic	3300-3500	3400-4400
Ge'ez	1600-1800	-
Modern Amharic	-	700-1700

Model goodness of fit greatly improved
by dropping Ugaritic and Ge'ez



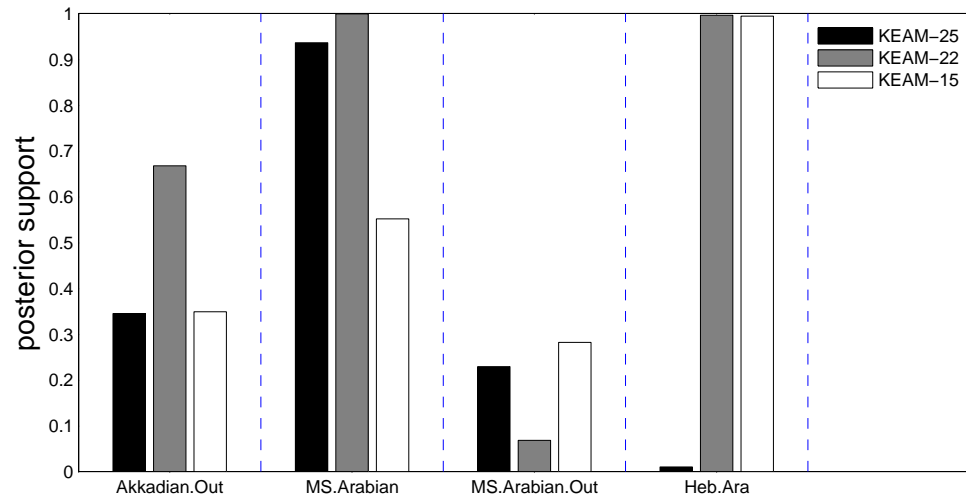
Posterior probability distribution

$$p(g, \mu, \lambda, \kappa, \rho, \xi | \mathbf{D} = D) = \frac{1}{\mu\lambda} f_G(g|T) p_R(\rho) \frac{e^{-\rho|g|} (\rho|g|)^{k_T}}{k_T!} \\ \times P[\mathbf{D} = D | g, \mu, \lambda, \xi, \mathbf{D} = R(\tilde{\mathbf{D}})]$$

Priors and estimation

Tree g	f_G	uniform marginal root age, uniform over topologies	MCMC
Death rate μ	$1/\mu$	improper scale invariant rate	MCMC
Birth rate λ	$1/\lambda$	improper scale invariant	integrate
Word birth times Z	PPP	thinned Poisson point process	integrate (pruning alg.)
Catastrophe event times	PPP	Poisson point process	integrate times to counts; MCMC
Catastrophe rate ρ	f_R, Γ	95% CI for number is $1/\text{tree} - 1/\text{edge}$	MCMC
Thinning probability κ	$U(0, 1)$		MCMC
Missing data probability ξ	$U(0, 1)^L$	Very well informed by data	MCMC

Results



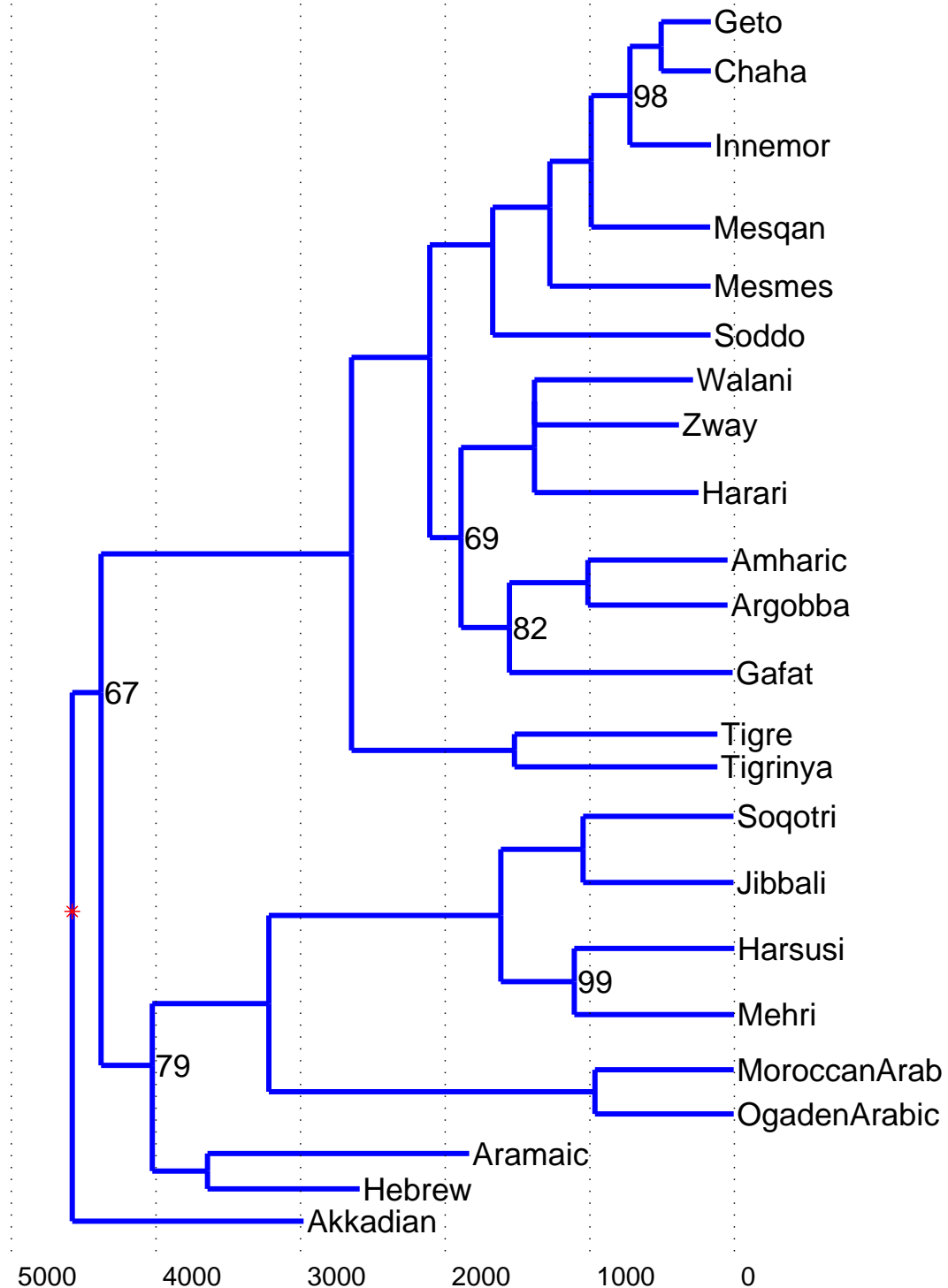
Root age 4400 – 5100YBP (95% HPD)

posterior/prior probabilities:

Akkadian outgroup, 0.67/0.04

Zero catastrophes, 0.33/0.01

Fair agreement with conclusions of Kitchen et al 2009 - clearer evidence for Akkadian outgroup, different MSA and CI for root age (was [4400, 7400]YBP).



Model Error 1: Predict Calibrations

Predict ages of nodes for which calibration data is available - would we reject the truth?

Γ tree space, all calibration constraints.

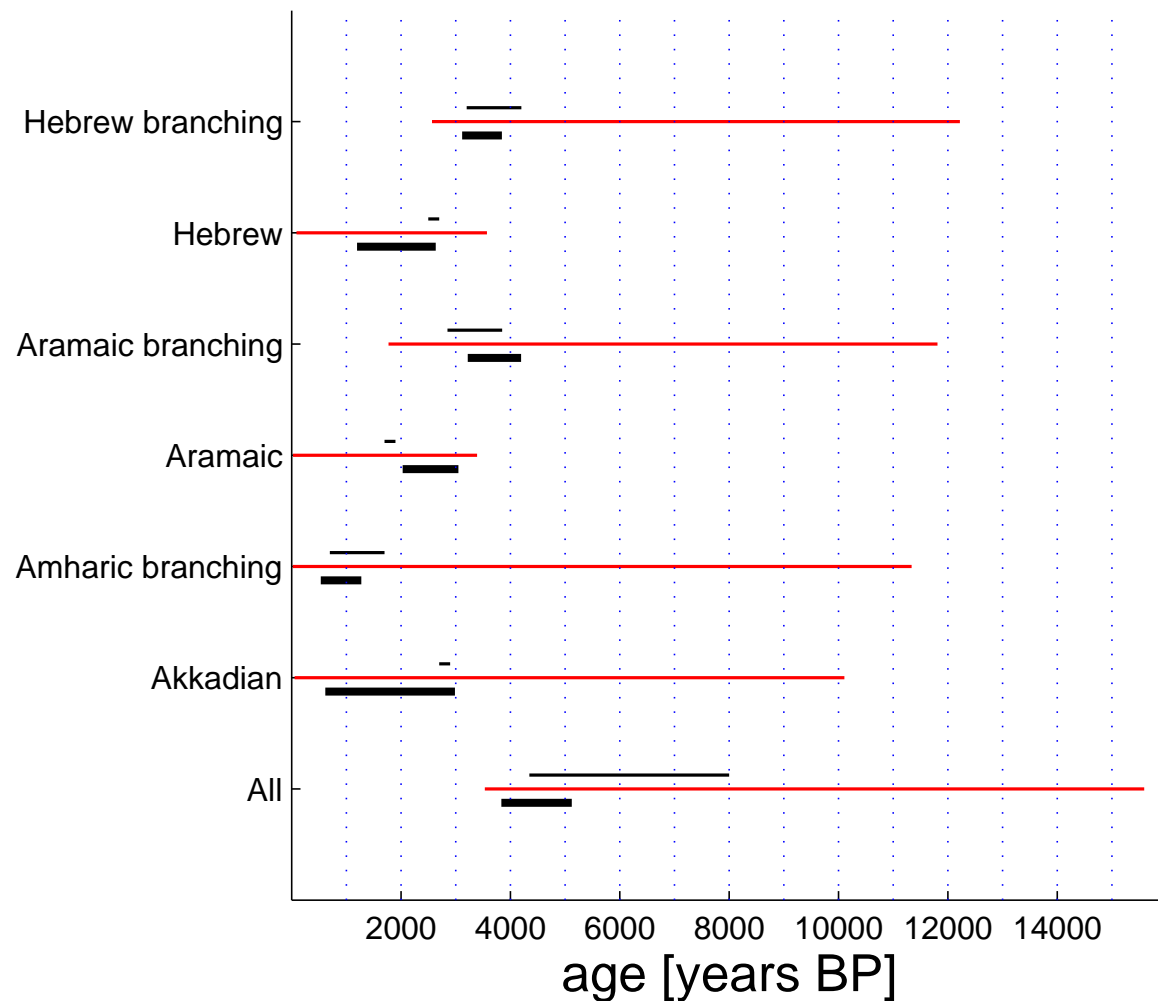
Γ^{-c} has $c = 1..7$ 'th constraint removed.

$M_0 : g \in \Gamma$, $M_1 : g \in \Gamma^{-c}$. Bayes factor,

$$B^{(c)} = \frac{P(g \in \Gamma | D, g \in \Gamma^{-c})}{P(g \in \Gamma | g \in \Gamma^{-c})}$$

Model conflicts constraints if $B^{(c)} \lesssim 0.08$.

	Sampled	Branched
Semitic	-	3.9
Akkadian	0.5	-
Aramaic	0.3	6
Hebrew	1.8	1.8
Amharic	-	2



Bayes factors measuring evidence for constraint.