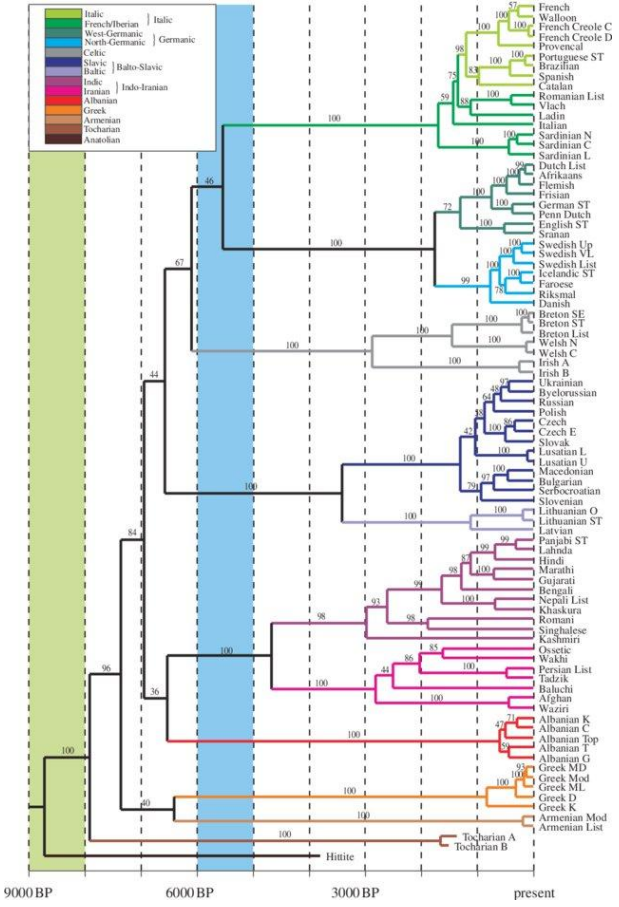
Illustrating Bayesian inference for stemmatology

Tiago Tresoldi
CEoT - Uppsala universitet

Studia Stemmatologica
2022/06/30

Introduction

- Advances in computational historical linguistics with evolutionary methods from biology
- Very different from NJ and UPGMA, a bit more similar to Maximum Likelihood
- A certain skepticism in stemmatology, which is right: the *tools* are not ready



What it does

- Given the *data* and *evolutionary model*, the tools collect thousands of trees, constantly looking for the best one
- The “goodness” of a tree is given by the probability of obtaining the *data* we observed when assuming an *underlying model*
- Everything is in terms of probability: not how much to “trust” a single result considered the best, but how *likely* that result is
- The collection of trees is reduced to a single tree with different strategies

Bayesian inference - I

Priors: original probability for the model parameters

All parameters have priors!

$$P(\text{model} \mid \text{data}) = \frac{P(\text{data} \mid \text{model})P(\text{model})}{P(\text{data})}$$

Source: Valenzuela (2021), TTB Online first steps

Bayesian inference - II

Likelihood: probability of data given the parameters

$$P(\text{model} \mid \text{data}) = \frac{P(\text{data} \mid \text{model})\,P(\text{model})}{P(\text{data})}$$

Source: Valenzuela (2021), TTB Online first steps

Bayesian inference - III

Evidence: probability in any combination of parameters

$$P(\text{model} \mid \text{data}) = \frac{P(\text{data} \mid \text{model})\, P(\text{model})}{P(\text{data})}$$

Source: Valenzuela (2021), TTB Online first steps

Bayesian inference - IV

Posterior: updated probability

$$P(\text{model} \mid \text{data}) = \frac{P(\text{data} \mid \text{model})\, P(\text{model})}{P(\text{data})}$$

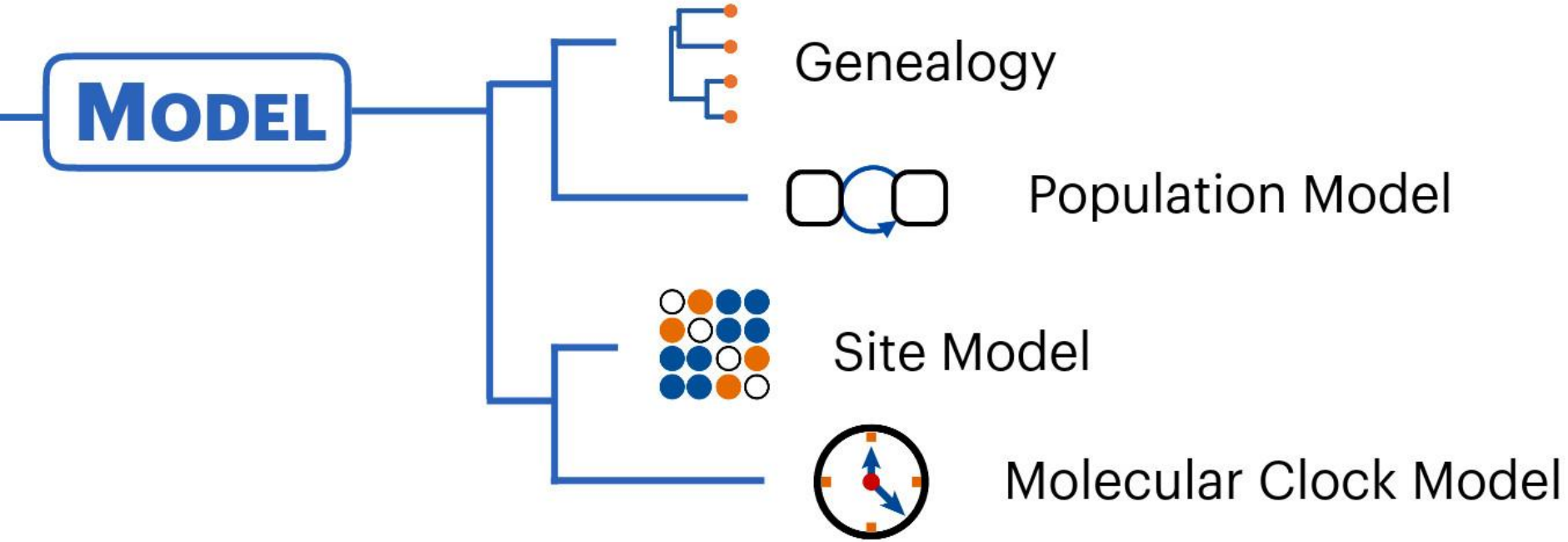
Source: Valenzuela (2021), TTB Online first steps

Data

$$P(\text{model} \mid \begin{matrix}ACAC\ldots\\TCAC\ldots\\ACAG\ldots\end{matrix}) = \frac{P(\begin{matrix}ACAC\ldots\\TCAC\ldots\\ACAG\ldots\end{matrix} \mid \text{model})\, P(\text{model})}{P(\begin{matrix}ACAC\ldots\\TCAC\ldots\\ACAG\ldots\end{matrix})}$$

Source: Valenzuela (2021), TTB Online first steps

Model



Source: Valenzuela (2021), TTB Online first steps

Model - II

Tree prior: How likely is a tree, given a demographic model

TREE
Realisation of a stochastic process

POPULATION MODEL
Describes the population dynamics (growth of the tree)

Source: Valenzuela (2021), TTB Online first steps

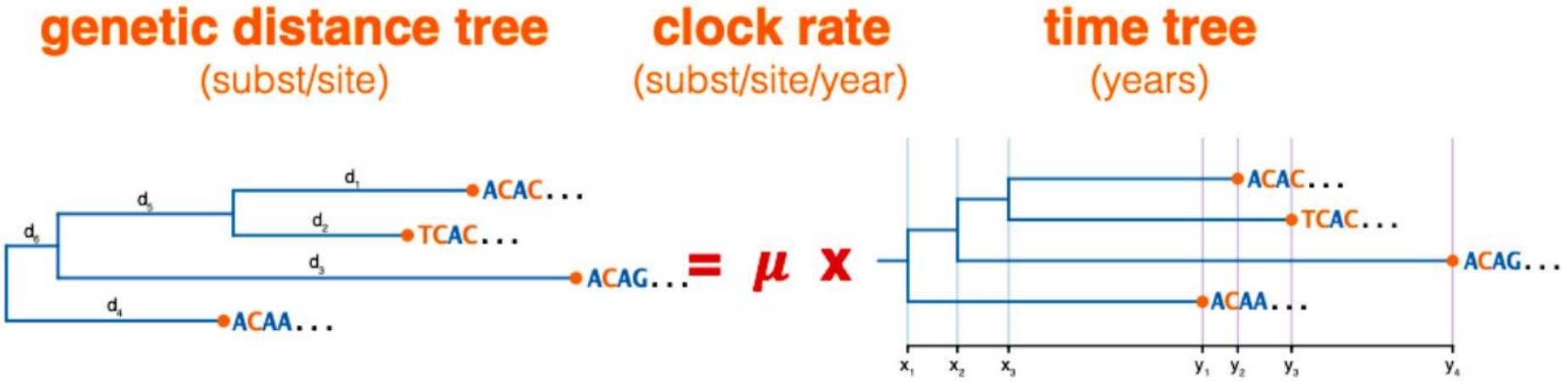
Model - III

Substitution model: tree likelihood sums all substitution histories leading to the ovserved sequences

Source: Valenzuela (2021), TTB Online first steps

Model - IV

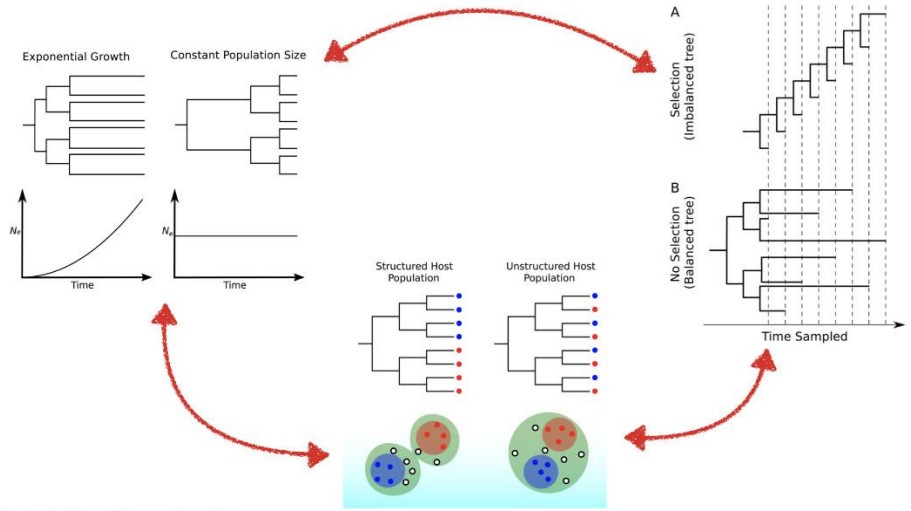
Molecular clock, strict/relaxed/random



Source: Valenzuela (2021), TTB Online first steps

Model - V

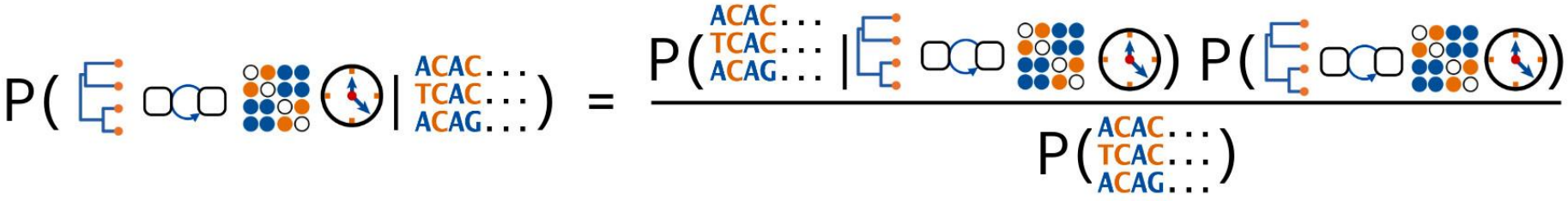
Different population dynamics give *very* different trees!



Volz *et al.* **PLoS Comp Biol 2013**
Grenfell *et al.* **Science 2004**

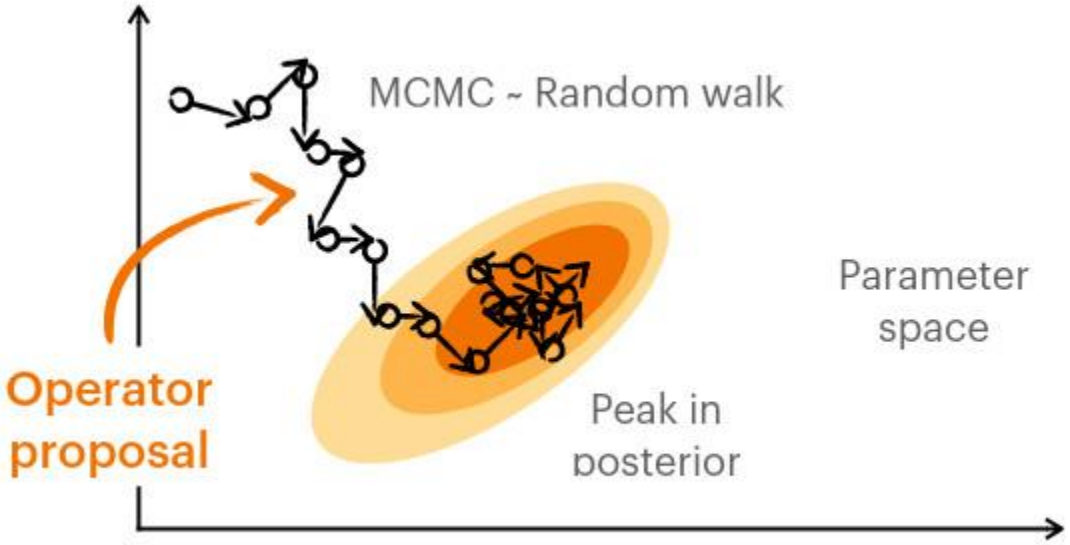
Source: Valenzuela (2021), TTB Online first steps

Model - VI



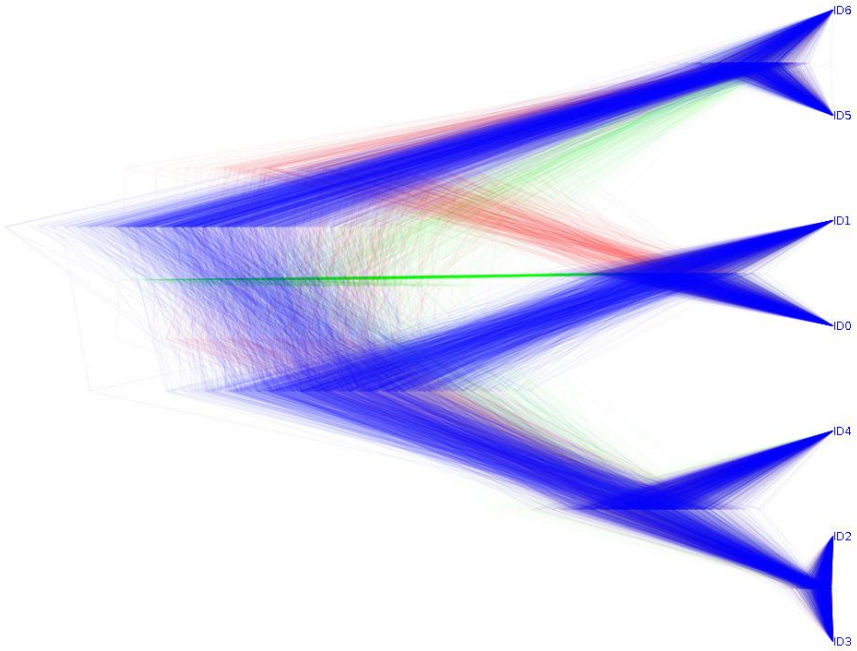
Source: Valenzuela (2021), TTB Online first steps

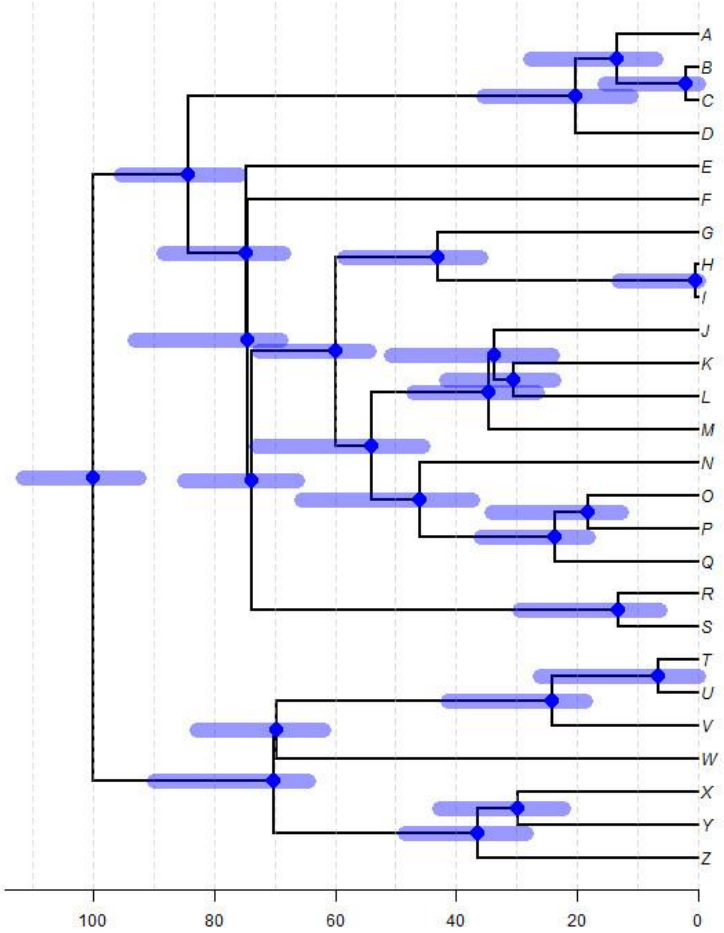
MCMC



Source: Valenzuela (2021), TTB Online first steps

Tree summary



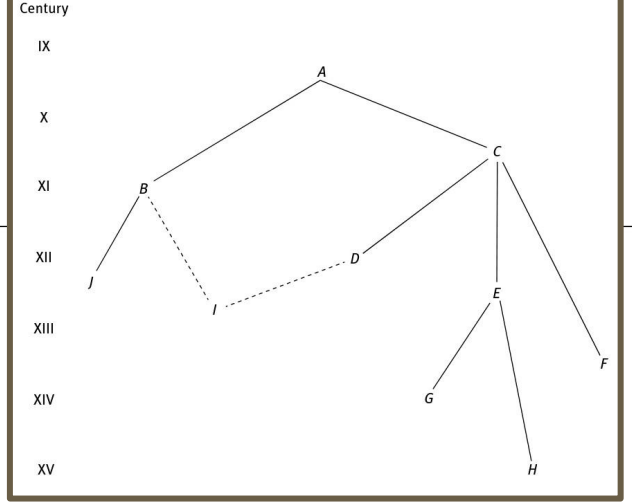


A	Eminentiora	prolixarum	arborum	culmina	perindeque	distenta	acuto	sonitu	resultabant
B	Eminentiora	prolixarum	abietum	cacumina	perindeque	distantia	acuto	sonitu	resultabant
C	Altiora	prolixarum	arborum	culmina	perindeque	distenta	acuto	sonitu	resonabant
D	Altiora	promissarum	arborum	culmina	perindeque	distenta	acutissimo	sonitu	resonabant
E	Altiora	prolixarum	arborum	culmina	perindeque	distenta	acuto	sono	resonabant
F	Altiora	prolixarum	arborum	fulmina	perindeque	et distenta	acuto	tinnitu	resonabant
G	Altiora		arborum	culmina	perindeque	discreta	acuto	sono	resonabant
H	Altiora	prolixarum	arborum	culmina	proptereaque	distenta	acuto	sono	resonabant
I	Eminentiora	promissarum	abietum	culmina	perindeque	distenta	acutissimo	sonitu	resultabant
J	Eminentiora	prolixarum	abietum	cacumina	per insignem	distantiam	acuto	sonitu	resultabant

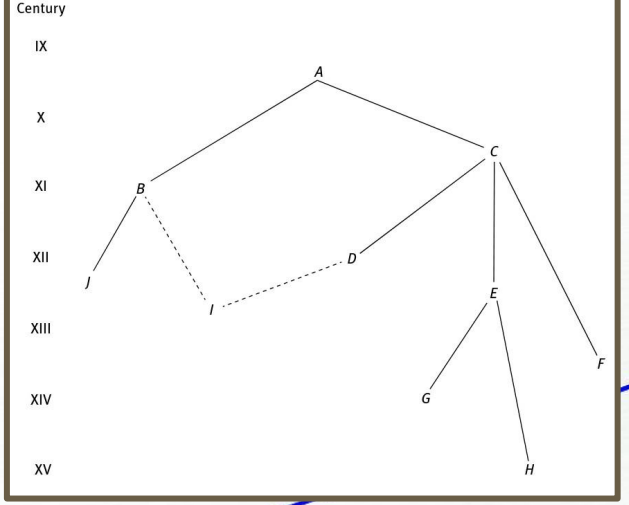
A	Eminentiora	prolixarum	arborum	culmina	perindeque	distenta	acuto	sonitu	resultabant
B	Eminentiora	prolixarum	abietum	cacumina	perindeque	distantia	acuto	sonitu	resultabant
C	Altiora	prolixarum	arborum	culmina	perindeque	distenta	acuto	sonitu	resonabant
D	Altiora	promissarum	arborum	culmina	perindeque	distenta	acutissimo	sonitu	resonabant
E	Altiora	prolixarum	arborum	culmina	perindeque	distenta	acuto	sono	resonabant
F	Altiora	prolixarum	arborum	fulmina	perindeque	et distenta	acuto	tinnitu	resonabant
G	Altiora		arborum	culmina	perindeque	discreta	acuto	sono	resonabant
H	Altiora	prolixarum	arborum	culmina	proptereaque	distenta	acuto	sono	resonabant
I	Eminentiora	promissarum	abietum	culmina	perindeque	distenta	acutissimo	sonitu	resultabant
J	Eminentiora	prolixarum	abietum	cacumina	per insignem	distantiam	acuto	sonitu	resultabant

A	A	C	F	H	K	N	S	U	X
B	A	C	G	I	K	O	S	U	X
C	B	C	F	H	K	N	S	U	Y
D	B	D	F	H	K	N	T	U	Y
E	B	C	F	H	K	N	S	V	Y
F	B	C	F	J	K	P	S	W	Y
G	B	E	F	H	K	Q	S	V	Y
H	B	C	F	H	L	N	S	V	Y
I	A	D	G	H	K	N	T	U	X
J	A	C	G	I	M	R	S	U	X

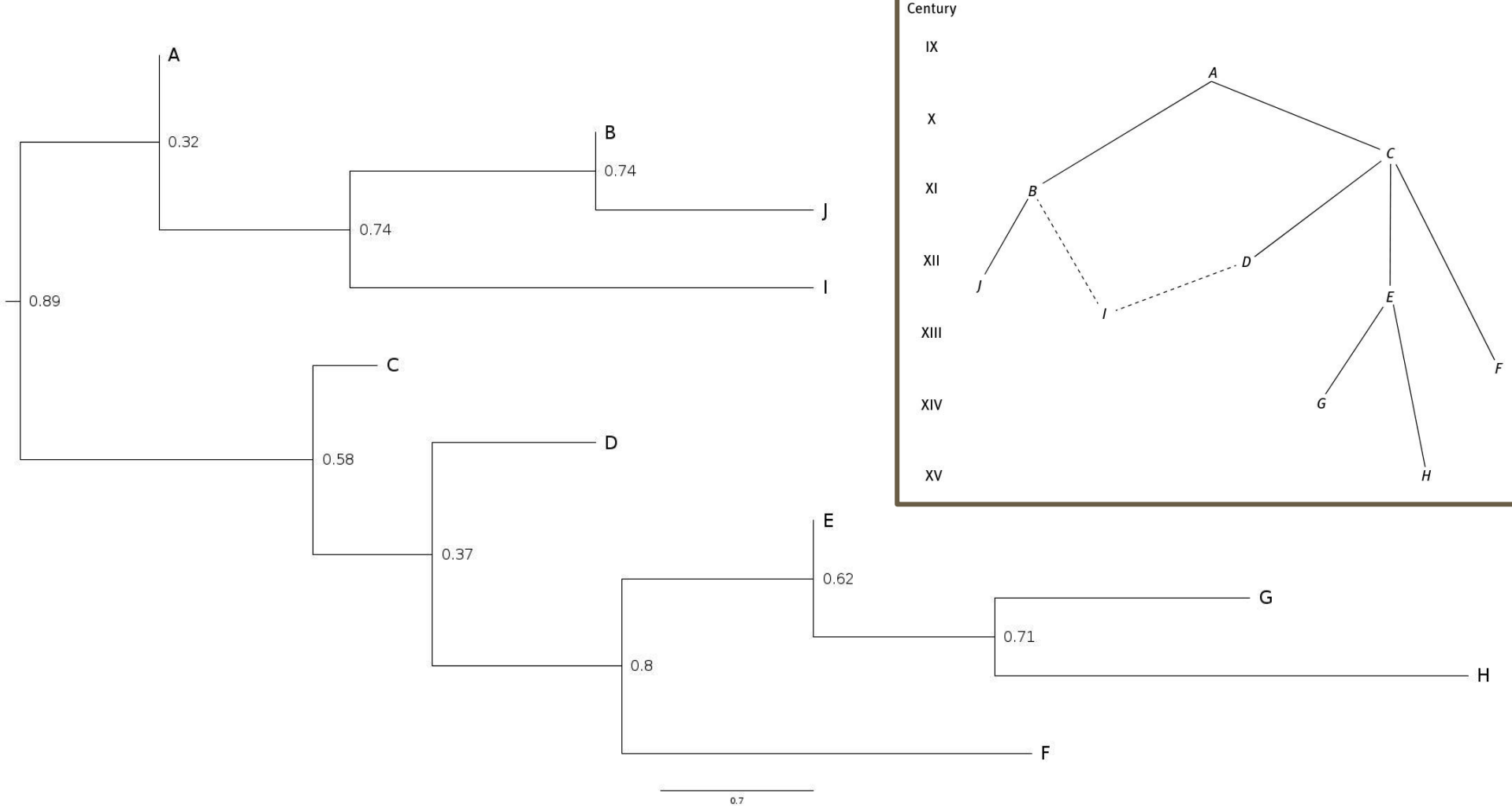
```
#NEXUS

BEGIN DATA;
    DIMENSIONS NTAX=10 NCHAR=34;
    FORMAT DATATYPE=STANDARD MISSING=? GAP=-  SYMBOLS="01";
    CHARSTATELABELS
        1 CHAR_1_ASCERT,
        2 CHAR_1_Altiora,
        3 CHAR_1_Eminentiora,
        (...)
        34 CHAR_9_resultabant
;
MATRIX
A 0010010001001000100000100010100001
B 0010010010010000100010000010100001
C 0100010001001000100000100010100010
D 0100001001001000100000100100100010
E 0100010001001000100000100010010010
F 0100010001000100100000010010001010
G 0100100001001000100100000010010010
H 0100010001001000010000100010010010
I 0010001010001000100000100100100001
J 0010010010010001000001000010100001
;
begin assumptions;
       charset CHAR_1 = 1-3;
       charset CHAR_2 = 4-7;
       charset CHAR_3 = 8-10;
       charset CHAR_4 = 11-14;
       charset CHAR_5 = 15-18;
```

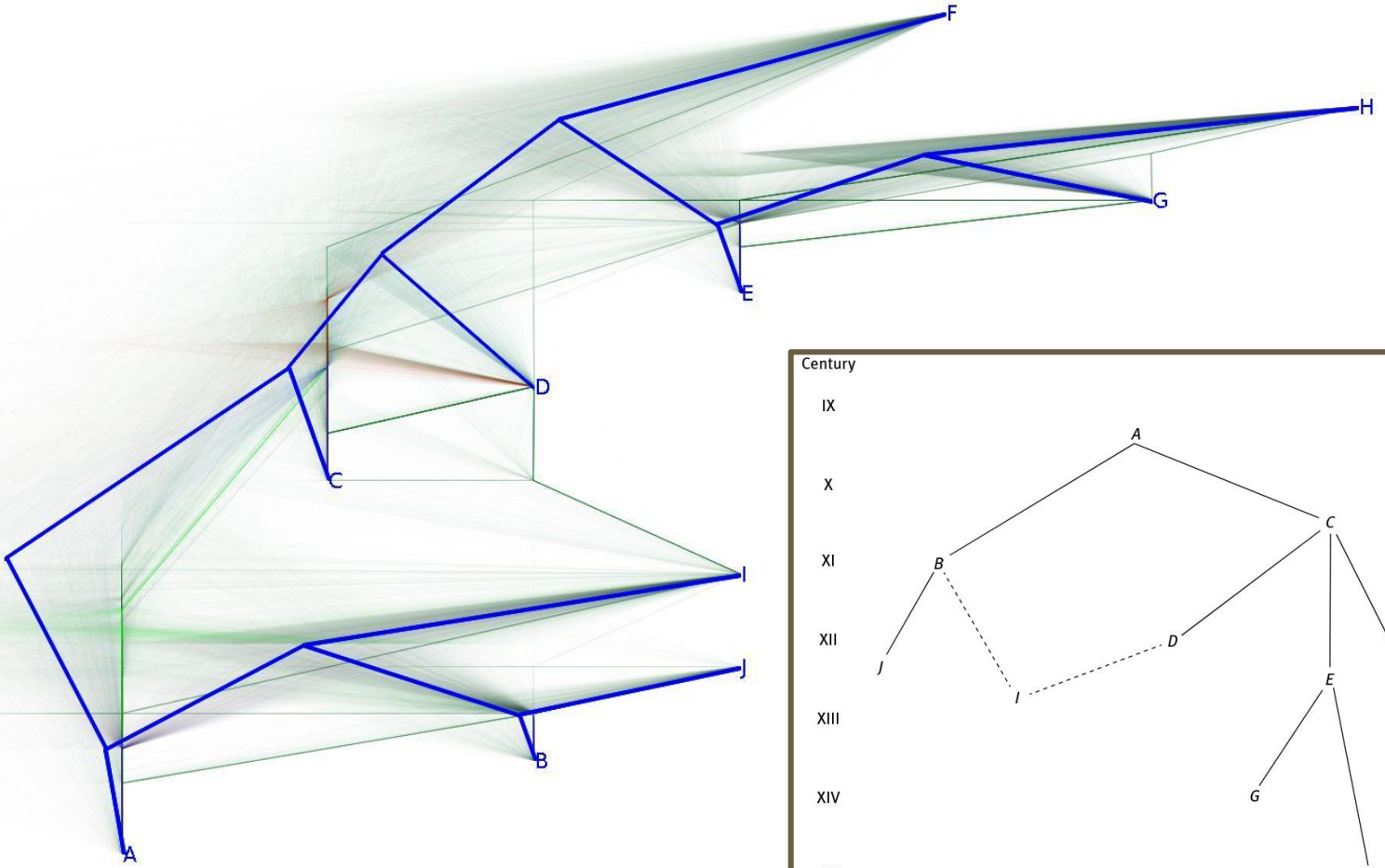

Century
IX
X
XI
XII
XIII
XIV
XV
A
B
C
D
E
F
G
H
I
J

1
0.8278
0.7636
0.3809
0.2547
0.4155
0.4969
0.5843
0.9498
A
I
B
J
C
D
E
H
G
F



F
G
H
E
D
C
J
B
I
A


A
0.32
B
0.74
J
0.74
I
0.89
C
0.58
D
0.37
E
0.62
G
0.71
H
0.8
F
0.7
Century
IX
X
XI
XII
XIII
XIV
XV
A
B
C
J
D
I
E
F
G
H


F
H
G
E
D
C
I
J
B
A

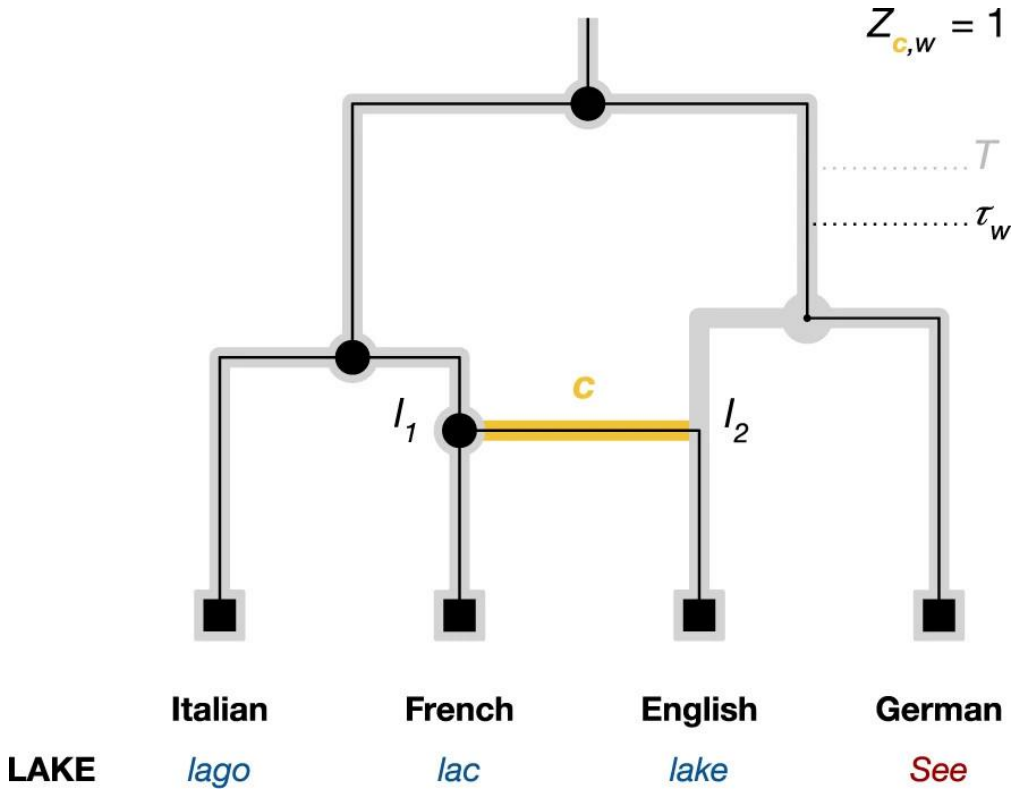
Century
IX
X
XI
XII
XIII
XIV
XV
A
B
C
D
E
F
G
H
I
J

Tools are not ready: some issues

- Tree prior: the biological models don’t reflect manuscript evolution (e.g. binary division)
- Molecular clock: time works differently when making copies from texts
- Site models: transitions tend to be very asymmetric, it is hard to encode expert knowledge
- We don’t really know what most parameters should be like

Where we are

- BEAST2 and MrBayes are limited in their offerings
- BEAST2 can be expanded, and RevBayes could be a solution
- We are currently experimenting with the detection on synthetic data (of which we know the “real” stemma) and the Divine Comedy (with data from Shaw 2011)



Source: Neureiter et al. (2022)

UPPSALA
UNIVERSITET

Thank you!

tiago.tresoldi@lingfil.uu.se