Bayesian Phylogenetic Cognate Prediction

Gerhard Jäger

Tübingen University

2022 Annual Conference of the North American Chapter of the Association for Computational Linguistics

SIGTYP 2022: The 4th Workshop on Research in Computational Typology and Multilingual NLP

`Shared Task: Prediction of Cognate Reflexes`

July 14, 2022

erc
European Research Council
Established by the European Commission

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

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

DFG

- **given:**
  - parallel word lists from related languages, including cognate classification

COGID	Albanian	English	French	German	Latin
920		h ɑ r t	k œ r	h e r ts ə n	k o r d
1083		h ɔ r n	k ɔ r n	h o r n	k o r n u:
1150	ʃ k u r t ə r	ʃ ɔ r t	k u r t	k u r ts	

Table: Example training data

  - reflexes of one (unseen) cognate class with one entry missing

COGID	Albanian	English	French	German	Latin
353-3	p e ʃ k	f ɪ ʃ	?	f i ʃ	p i s k i

Table: Example test data

SIGTYP 2022 Shared Task

- **task:**
  - predict the missing entry
  - in our case: *pʃ*
- more details on `https://github.com/sigtyp/ST2022`

General approach

- three-step procedure
    1. compute **multiple sequence alignment** of known words from test cognate class

Albanian	p	e	ʃ	k	-
English	f	ɪ	ʃ	-	-
French	?	?	?	?	?
German	f	i	ʃ	-	-
Latin	p	i	s	k	i

Table: MSA for cognate prediction

General approach

- three-step procedure
  1. compute **multiple sequence alignment** of known words from test cognate class
  2. infer family tree of the languages involved via **Bayesian phylogenetic inference**

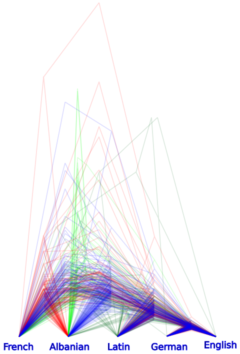


Figure: Posterior tree distribution

General approach

- three-step procedure
  1. compute **multiple sequence alignment** of known words from test cognate class
  2. infer family tree of the languages involved via **Bayesian phylogenetic inference**
  3. for each column, impute missing symbol based on tree

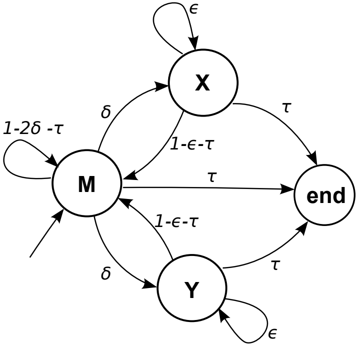
Albanian	p	e	ʃ	k	-
English	f	ɪ	ʃ	-	-
French	**p**	**i**	**ʃ**	**k**	**–**
German	f	i	ʃ	-	-
Latin	p	i	s	k	i

Table: MSA for cognate prediction

Pairwise sequence alignment

- **multiple sequence alignment** (MSA) relies on **pairwise sequence alignment**
- this project: **Pair-Hidden Markov Model** (pHMM; cf. Durbin et al. 1989)

- HMM with two output tapes
- state **M** emits one symbol on each tape
- states **X** (**Y**) emits only on first (second) tape
- fitted model defined probability distribution over pairs of strings
- training via EM (Baum-Welch) algorithm on all cognate pairs from training set
- Viterbi algorithm outputs pairwise alignment

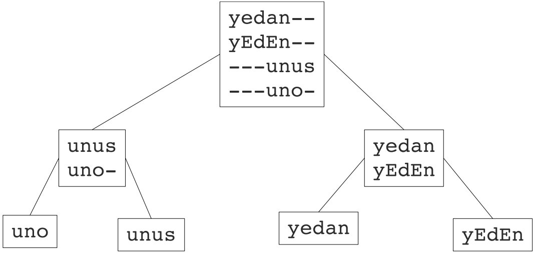


ϵ
X
δ
τ
$1\text{-}2\delta\ \text{-}\tau$
$1\text{-}\epsilon\text{-}\tau$
M
τ
end
$1\text{-}\epsilon\text{-}\tau$
δ
τ
Y
ϵ

Figure: Pair Hidden Markov Model

Progressive sequence alignment

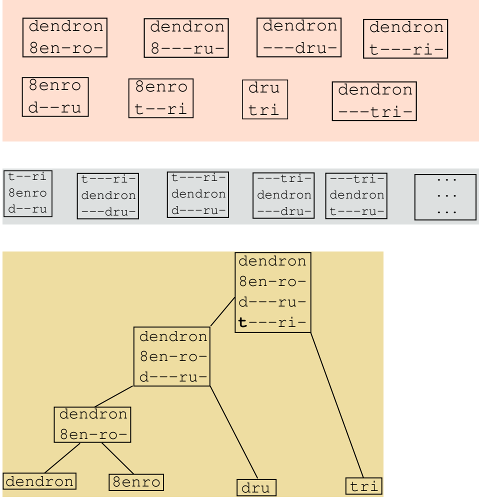
- start with a **guide tree** (using some heuristics such as mean LDN + UPGMA)
- working bottom-up, at each internal node, do pairwise alignment of the block alignments at the daughter node
- complexity is $\mathcal{O}(n^2k^3) \Rightarrow$ computationally feasible



T-Coffee

(Notredame et al., 2000)

1. pairwise alignment for all word pairs, using Viterbi/pHMM
2. ternary alignments via relation composition
3. indirect alignment scores between sound **occurrences**
4. progressive alignment using those scores



```
dendron   dendron   dendron   dendron
8en-ro-   8---ru-   ---dru-   t---ri-

8enro     8enro     dru       dendron
d--ru     t--ri     tri       ---tri-
```

```
t--ri   t---ri-   t---ri-   ---tri-   ---tri-   ...
8enro   dendron   dendron   dendron   dendron   ...
d--ru   ---dru-   d---ru-   ---dru-   t---ru-   ...
```

```
dendron
8en-ro-
d---ru-
t---ri-

dendron
8en-ro-
d---ru-

dendron
8en-ro-

dendron   8enro   dru   tri
```

- T-Coffee is used to produce MSAs of all cognate sets in the training data
- resulting MSAs are concatenated to one big **character matrix**

Albanian	ʃ	k	u	r	t	ə	r	t	-	⋯	.	.	.	.	.	.	.	.	.	.
English	ʃ	ɔ	-	r	t	-	-	.	.	⋯	.	.	.	.	.	.	.	.	.	.
French	k	u	-	r	t	-	-	t	y	⋯	.	.	.	.	.	.	.	.	.	.
German	k	u	-	r	ts	-	-	d	uː	⋯	n	eː	b	ə	l	-	h	au	t	-
Latin	.	.	.	.	.	.	.	t	uː	⋯	n	e	b	u	l	a	k	u	t	i

- each column is assumed to result from a *continuous time Markov process* on a phylogenetic tree

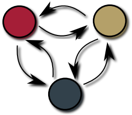
Continuous time Markov chains on a tree

- continuous time Markov process (CTMC) with discrete state space
- characterized by Q-matrix, e.g.

$$Q = r \begin{pmatrix} -2 & 1 & 1 \\ 1 & -2 & 1 \\ 1 & 1 & -2 \end{pmatrix}$$

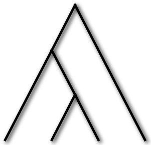
- Here: **Jukes-Cantor model** (originally developed for DNA evolution)
  - all rates are equal
  - global rate r (expected number of mutations per unit of time) as parameter to be estimated

Markov process



Continuous time Markov chains on a tree

Phylogeny



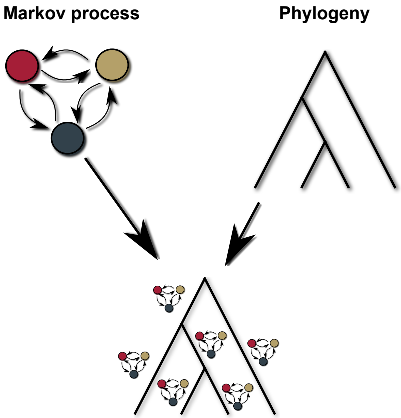
- (unordered) tree $\mathcal{T}$ with branch lengths
- Here:
  - inferred from lexical data
  - branch lengths represent amount of lexical change

Continuous time Markov chains on a tree

- phylogenetic CTMC
- independent copies of CTMC on each branch of the tree
- likelihood of a branch of length t with rate r, states a and b and top and bottom and n possible states:

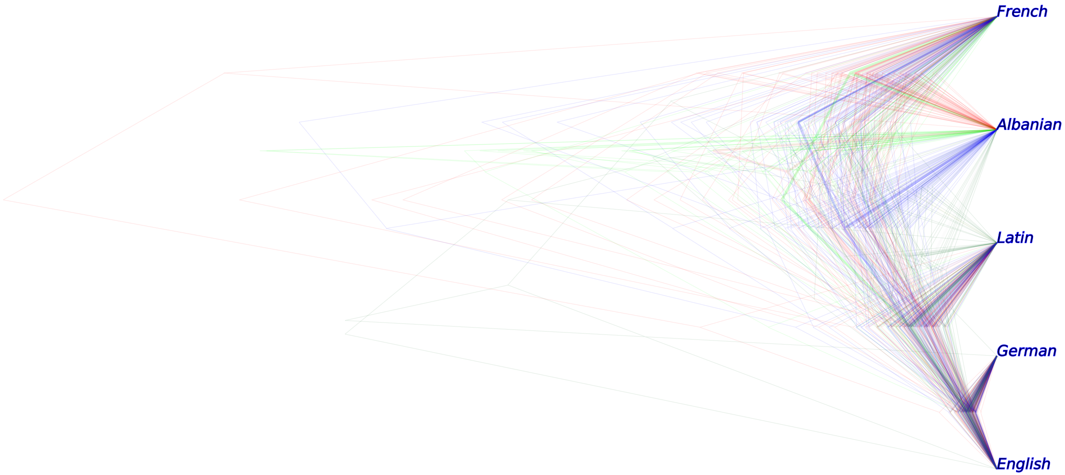
$$P(b|a;t,r) = \frac{1}{n}\begin{cases} 1+(n-1)e^{-tr} & \text{if } a=b \\ 1-e^{-tr} & \text{else} \end{cases}$$

- likelihood of entire tree is product of branch likelihoods
- unkown states are marginalized out
- marginal likelihood can be efficiently computed via dynamic programming (bottom-up recursion through the tree, cf. Felsenstein, 2004)

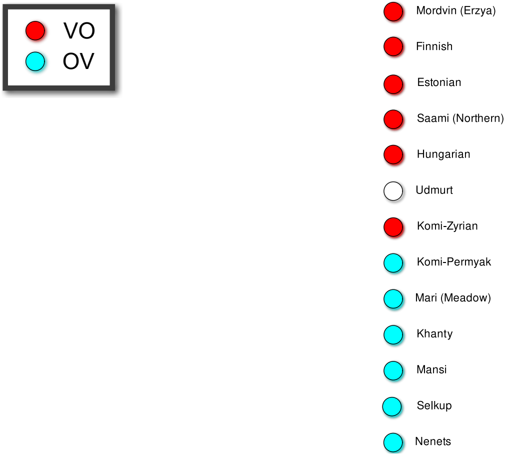


Phylogenetic inference

- via Bayesian inference *posterior distribution over phylogenies* and *mutation rates*



Phylogenetic value imputation

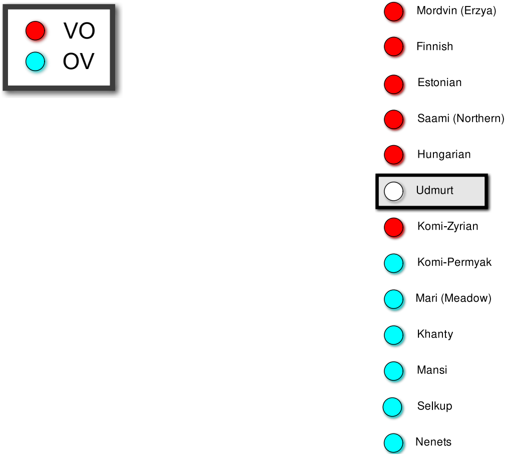


🔴 VO
🔵 OV

🔴 Mordvin (Erzya)
🔴 Finnish
🔴 Estonian
🔴 Saami (Northern)
🔴 Hungarian
○ Udmurt
🔴 Komi-Zyrian
🔵 Komi-Permyak
🔵 Mari (Meadow)
🔵 Khanty
🔵 Mansi
🔵 Selkup
🔵 Nenets

- missing value for a certain feature in a certain language (typological, symbol in an alignment, whatever)
- feature value in genetically related languages is known

Phylogenetic value imputation

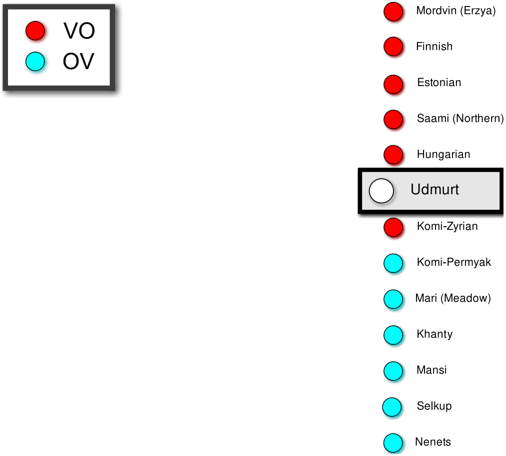


🔴 VO
🔵 OV

🔴 Mordvin (Erzya)
🔴 Finnish
🔴 Estonian
🔴 Saami (Northern)
🔴 Hungarian
○ Udmurt
🔴 Komi-Zyrian
🔵 Komi-Permyak
🔵 Mari (Meadow)
🔵 Khanty
🔵 Mansi
🔵 Selkup
🔵 Nenets

- missing value for a certain feature in a certain language (typological, symbol in an alignment, whatever)
- feature value in genetically related languages is known

Phylogenetic value imputation

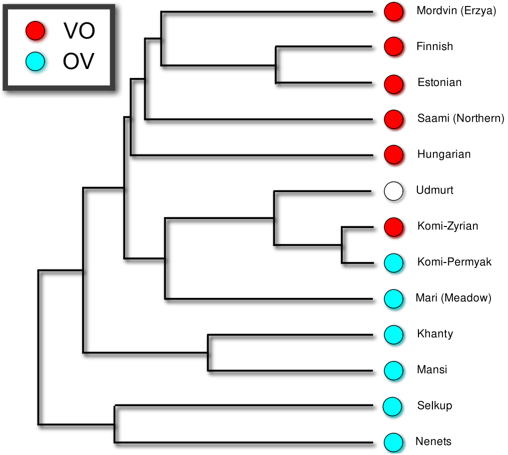


🔴 VO
🔵 OV

🔴 Mordvin (Erzya)
🔴 Finnish
🔴 Estonian
🔴 Saami (Northern)
🔴 Hungarian
○ Udmurt
🔴 Komi-Zyrian
🔵 Komi-Permyak
🔵 Mari (Meadow)
🔵 Khanty
🔵 Mansi
🔵 Selkup
🔵 Nenets

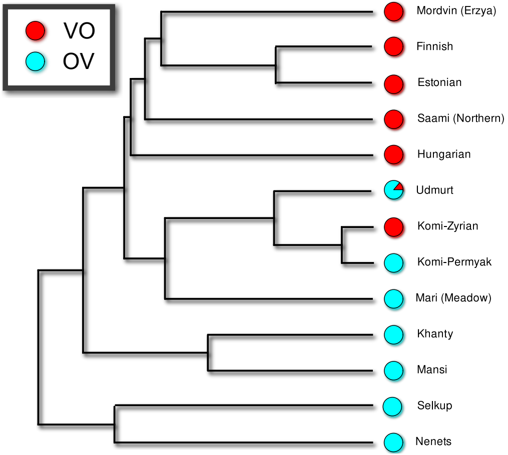
- missing value for a certain feature in a certain language (typological, symbol in an alignment, whatever)
- feature value in genetically related languages is known

Phylogenetic value imputation



- missing value for a certain feature in a certain language (typological, symbol in an alignment, whatever)
- feature value in genetically related languages is known
- use family tree

Phylogenetic value imputation



- missing value for a certain feature in a certain language (typological, symbol in an alignment, whatever)
- feature value in genetically related languages is known
- use family tree
- interpolate feature value from related languages as *Maximum A Posteriori* estimate of phylogenetic CTMC
- posterior distributions from training set are used as prior distributions for test set imputation

Phylogenetic value imputation

Albanian	p	e	ʃ	k	-
English	f	ɪ	ʃ	-	-
French					
German	f	i	ʃ	-	-
Latin	p	i	s	k	i

- gap symbols are treated as normal character states for phylogenetic inference
- for final prediction, gap states are removed

Phylogenetic value imputation

Albanian	p	e	ʃ	k	-
English	f	ɪ	ʃ	-	-
French	**p**				
German	f	i	ʃ	-	-
Latin	p	i	s	k	i

- gap symbols are treated as normal character states for phylogenetic inference
- for final prediction, gap states are removed

Phylogenetic value imputation

Albanian	p	e	ʃ	k	-
English	f	ɪ	ʃ	-	-
French	**p**	**i**			
German	f	i	ʃ	-	-
Latin	p	i	s	k	i

- gap symbols are treated as normal character states for phylogenetic inference
- for final prediction, gap states are removed

Phylogenetic value imputation

Albanian	p	e	ʃ	k	-
English	f	ɪ	ʃ	-	-
French	**p**	**i**	**ʃ**		
German	f	i	ʃ	-	-
Latin	p	i	s	k	i

- gap symbols are treated as normal character states for phylogenetic inference
- for final prediction, gap states are removed

Phylogenetic value imputation

Albanian	p	e	ʃ	k	-
English	f	ɪ	ʃ	-	-
French	**p**	**i**	**ʃ**	**k**	
German	f	i	ʃ	-	-
Latin	p	i	s	k	i

- gap symbols are treated as normal character states for phylogenetic inference
- for final prediction, gap states are removed

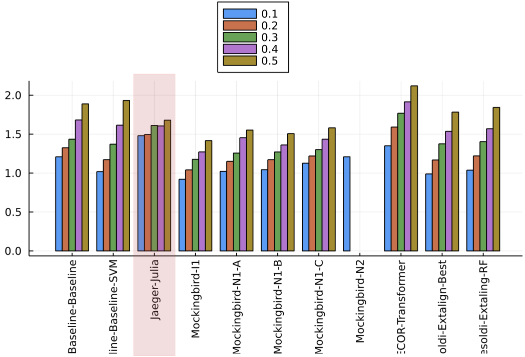
Phylogenetic value imputation

Albanian	p	e	ʃ	k	-
English	f	ɪ	ʃ	-	-
French	**p**	**i**	**ʃ**	**k**	**–**
German	f	i	ʃ	-	-
Latin	p	i	s	k	i

- gap symbols are treated as normal character states for phylogenetic inference
- for final prediction, gap states are removed

Evaluation

- performance is somewhere in the middle among the submitted systems
- possible advantage: system is relatively robust in the face of data sparseness



Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. *Biological Sequence Analysis*. Cambridge University Press, Cambridge, UK, 1989.

Joseph Felsenstein. *Inferring Phylogenies*. Sinauer Inc. Publishers, Sunderland, 2004.

Cédric Notredame, Desmond G. Higgins, and Jaap Heringa. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology*, 302(1):205–217, 2000.