Are Sounds Sound for Phylogenetic Reconstruction?

SIGTYP 2024 Workshops
March 22, 2024

Luise Häuser, Gerhard Jäger, Taraka Rama, Mattis List, Alexandros Stamatakis

Background (1)

- **Computational language phylogenies** applied and accepted in **comparative linguistics**
- Skepticism against language phylogenies based on **cognate sets** reflecting **lexical data** only
- Data based on **shared innovations** such as **sound correspondences** assessed to be superior by classical linguists

Background (2)

- Phylogenetic reconstruction using **Bayesian phylogenetic inference**
- Cognate sets encoded as **binary vectors**
- Assumption, that cognate sets evolve along a phylogenetic tree via a **gain and loss processes**
- Binary state data evolution modeled via a **time-reversible binary state Continuous Time Markov Chain model**

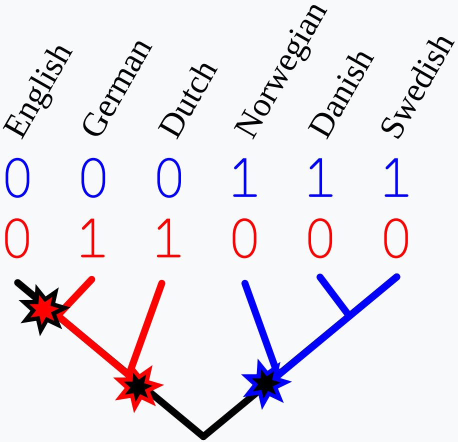
Background (3)

Language	Concept	Form	Cog-Set
English	"big"	big	1
German	"big"	groß	2
Dutch	"big"	groot	2
Norwegian	"big"	stor	3
Danish	"big"	stor	3
Swedish	"big"	stor	3

(A) multi-state matrix

Concept		" big"		
Cog-Set		1	2	3
English	big	0	0	0
German	groß	0	1	0
Dutch	groot	0	1	0
Norweg.	stor	0	0	1
Danish	stor	0	0	1
Swedish	stor	0	0	1

(B) binary-state matrix

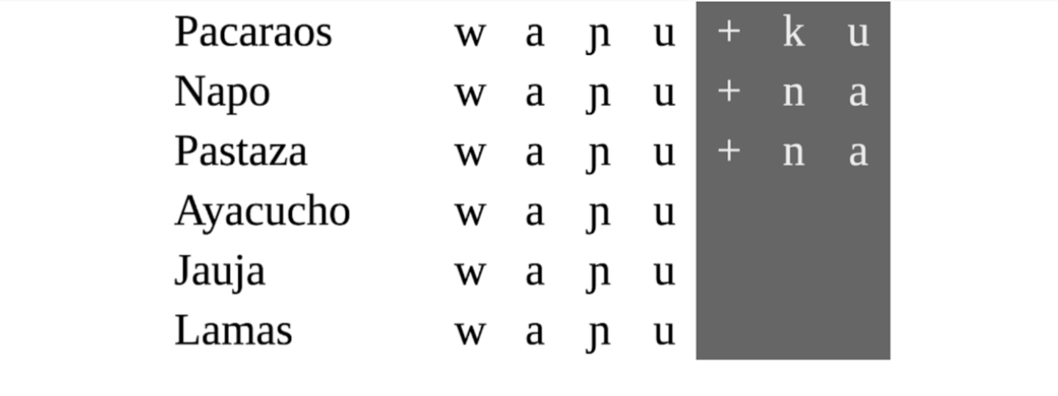


(C) evolutionary scenario (binary-state)

Contribution

- **Automated workflow** comprising:
  - Novel approaches for **inferring sound correspondence patterns**
  - State-of-the-art methods for **phonetic alignment**
  - Analysis with **Bayesian phylogenetic inference**
- Triangulation of the results from Bayesian inference via **Maximum Likelihood (ML)** tree reconstructions
- Comparison of the quality of phylogenetic reconstruction based on sound correspondences an based on lexical data
- Reassessment of the usefulness of sound-based as opposed to cognate-based phylogenies

Phonetic alignments



Pacaraos	w	a	ɲ	u	+	k	u
Napo	w	a	ɲ	u	+	n	a
Pastaza	w	a	ɲ	u	+	n	a
Ayacucho	w	a	ɲ	u			
Jauja	w	a	ɲ	u			
Lamas	w	a	ɲ	u			

Figure 2: Trimming morphemes in Quechua. The root is combined with different morphemes in some varieties.

Materials

- Ten Datasets from the **Lexibank** repository
- Languages linked to **Glottolog** for expert phylogenies
- Sounds provided in the phonetic transcription underlying the **Cross-Linguistic Transcription Systems** initiative
- Preprocessing:
    - **Phonetic alignment** of all cognate sets
    - Alignment **trimming**
    - Computation of **correspondence patterns**
    - Construction of **binary presence-absence matrices**

Methods

- **Character matrix types**:
  - Cognate matrix
  - Sound Correspondence matrix
  - Combined matrix
- **Hypotheses**:
  1. Phylogenetic inference is more accurate on cognate sets than on sound correspondence patterns
  2. Phylogenetic inference is more accurate on sound correspondence patterns than on cognate sets patterns
  3. Both character types do not differ substantially regarding their phylogenetic signal

Bayesian Inference using MrBayes

- **Priors**:
  - Base frequencies:Dirichlet(1.0, 1.0)
  - Tree topologies: Uniform
  - Branch lengths: Strict clock model
  - α shape parameter for Γ distributed rates: Uniform(0.01, 100)
- **Chains**: 2 cold chains
- **Stop criterion**: ASDSF$<$ 0.01
- **Burn-in**: 25%
- **Sampling**: Every 1,000th generation, 1,000 sampled trees drawn at random for further evaluation

Prior Bias

- Approximation of the Γ **model of rate heterogeneity** via four discrete rates
- Includes estimate of α **shape parameter** $\in [0.0201, 100]$
- The smaller α, the higher the rate heterogeneity
- Different priors yielding different distributions of α:
  - Exponential(1.0) prior: $\alpha \leqslant 10$ for almost all datasets
  - Uniform(0.01, 100) prior: $\alpha \geqslant 50$ for several datasets
- **Default exponential prior** developed for molecular datasets exhibiting a high rate heterogeneity
- **Prior bias** when applied for language datasets

ML Tree Inference using RAxML-NG

- **20 independent ML tree searches** for each dataset and each character matrix type
- **BIN+G model** of binary character substitution
- Extreme **bi-modal distribution** of α shape parameter estimates
- Reasons remain unclear

Rate Heterogeneity

Dataset	Cognates	Sound Correspondences	Combined
ConstenlaChibchan	0.592	**99.871**	4.178
CrossAndean	1.243	6.334	1.154
Dravlex	0.702	4.301	2.234
FelekeSemitic	1.062	7.430	2.693
Hattorijaponic	**99.848**	**99.897**	**99.890**
HouChinese	2.357	6.120	4.195
LeeKoreanic	8.316	8.420	3.284
RobinsonAP	**99.869**	15.269	3.486
WalworthPolynesian	1.333	4.233	1.624
ZhivlovObugrian	**99.850**	4.244	3.134

ML estimates of the alpha shape value of the Gamma model for among site rate heterogeneity

Rate Heterogeneity

Dataset	Cognates	Sound Correspondences	Combined
ConstenlaChibchan	1.758	**53.115**	1.138
CrossAndean	1.620	19.558	0.400
Dravlex	0.749	23.613	0.814
FelekeSemitic	0.932	41.669	0.727
HattoriJaponic	**58.012**	**60.602**	0.268
HouChinese	3.011	27.476	0.933
LeeKoreanic	**52.045**	39.354	0.058
RobinsonAP	**56.928**	**51.818**	0.373
WalworthPolynesian	1.480	4.348	0.800
ZhivlovObugrian	**58.652**	**51.280**	0.507

Median Bayesian estimates of the alpha shape value of the Gamma model for among site rate heterogeneity

Generalized Quartet Distance

- **Generalized quartet distance** (GQD): Number of quartets that are not shared between the two trees, divided by the number of all possible quartets
- GQD $\in [0, 1]$ - 0: identical, 1: completely different
- Used to measure **topological distances** of inferred phylogenies to the classification from Glottolog

Bayesian Inference

Generalized quartet distances (posterior medians):

Dataset	**Cog.**	**Sound C.**	**Conc.**
ConstenlaChibchan	0.245	0.414	**0.212**
CrossAndean	**0.148**	0.523	0.189
Dravlex	0.336	0.351	**0.320**
FelekeSemitic	**0.083**	0.146	0.113
HattoriJaponic	0.585	0.431	**0.362**
HouChinese	**0.240**	0.494	0.377
LeeKoreanic	0.224	0.358	**0.157**
RobinsonAP	0.424	0.281	**0.259**
WalworthPolynesian	0.179	0.252	**0.146**
ZhivlovObugrian	0.330	0.356	**0.316**
median	0.251	0.358	**0.240**

Bayesian Inference

- Results on cognate class data and on concatenated data about equally good, with a slight advantage for concatenated data
- Sound correspondences alone yield clearly worse results that are clearly worse
- Clear evidence in favor of Hypothesis 1 and against Hypothesis 2, Hypothesis 3 equivocal

Maximum Likelihood

Generalized quartet distances (best-scoring tree):

Dataset	**Cog.**	**Sound C.**	**Conc.**
ConstenlaChibchan	0.335	0.360	**0.283**
CrossAndean	0.246	0.470	**0.088**
Dravlex	0.358	0.472	**0.307**
FelekeSemitic	0.126	**0.103**	0.126
HattoriJaponic	**0.532**	0.681	0.559
HouChinese	0.224	0.529	**0.186**
LeeKoreanic	**0.178**	0.386	0.204
RobinsonAP	0.355	**0.321**	0.348
WalworthPolynesian	**0.139**	0.188	0.192
ZhivlovObugrian	**0.322**	0.356	0.360
median	0.251	0.358	**0.240**

Maximum Likelihood

- Tree inferred on the sound correspondences is never substantially better but clearly worse for three datasets
- Inferences on the cognate and combined datasets yield comparable results
- Consistent with the Bayesian inference results

Conclusion

- Cognate-based phylogenies are topologically closer to the gold standard than those inferred for sound correspondence patterns
- Unclear whether combined data leads to better results
- Re-assessment of priors required when applying Bayesian inference to language data
- Extreme bi-modal distribution of α values observed for unclear reasons

Thank you for your attention!
Questions?