An event-based model for linguistic phylogenetics

David Goldstein[*1], Shawn McCreight[2], Éva Buchi[3], and John Huelsenbeck[4]

[*]Corresponding Author: goldsteindm@gmail.com
[1]Department of Linguistics, University of California, Los Angeles, Los Angeles, USA
[2]Nytril LLC, Gig Harbor, USA
[3]Centre national de la recherche scientifique, Paris, France
[3]Université de Lorraine, Nancy, France
[4]Department of Integrative Biology, University of California, Berkeley, Berkeley, USA

Introduction. In linguistic phylogenetics inferences are standardly drawn from lexical cognate relationships, which are represented with abstract discrete values such as 0 and 1 in the case of binary characters (e.g., Bouckaert et al., 2012; Greenhill & Gray, 2012; Chang, Cathcart, Hall, & Garrett, 2015). Despite the prevalence of this approach, it suffers from well-known flaws.

Table 1. Cognate word-forms in Romance for 'stone'

Language	Aligned cognate word-forms						
Latin	p		e	t	r	a	m
Portuguese	p		ɛ	ð	ɾ	a	
Spanish	p	j	e	d	ɾ	a	
Catalan	p		e	d	ɾ	ə	
French	p	j	ɛ		ʁ		
Italian	p	j	ɛ	t	r	a	
Romanian	p	j	a	t	r	ə	

First, it discards a massive amount of information. Consider the Romance word-forms in Table 1, which all descend from a common ancestor. Under the conventional approach, they would all be assigned to the same cognate class. Although identical in this respect, they have diverged segmentally. It is precisely this segmental divergence that the standard practice ignores. Second, the representation of cognate relationships relies on arbitrary values, which lack consistent reference across cognate sets (Wright, Lloyd, & Hillis, 2016, 602). As a result, the standard approach does not model events of lexical change directly and estimated transition rates are not linguistically meaningful.

Incorporating segmental information. The TKF91 model overcomes these problems by modeling segmental changes among cognate word-forms (Thorne, Kishino, & Felsenstein, 1991; Lunter, Miklós, Song, & Hein, 2003). Under this model, one of three events is possible in an instant of time: an insertion of a single segment, a deletion of a single segment, or a transition from one segment to another. These are the very processes that give rise to the Romance word-forms in Table 1. Insertions and deletions are modeled as continuous-time birth-death processes, while substitution models such as JC69 or GTR are used for transitions between segments. This talk presents the first application of the TKF91 model to linguistic data.

Data and methods. Parameters are estimated in a Bayesian-MCMC framework, with estimates based on aligned phonemic sequences of 2,628 cognate word-forms from 9 Romance languages and Latin. Concepts for the cognate sets are selected from the Swadesh 207-word list. The model is provided with initial alignments, but they are marginalized over, so posterior distributions are not conditioned on any particular one. Tree topologies and branch lengths can also be estimated in this framework, but here I focus on transition rates.

Results and Discussion. Estimates of segmental volatility are presented in Figure 1. Vowels are on the whole more volatile than consonants, with long vowels and diphthongs being particularly unstable. Transition rates within each segmental class are remarkably similar.

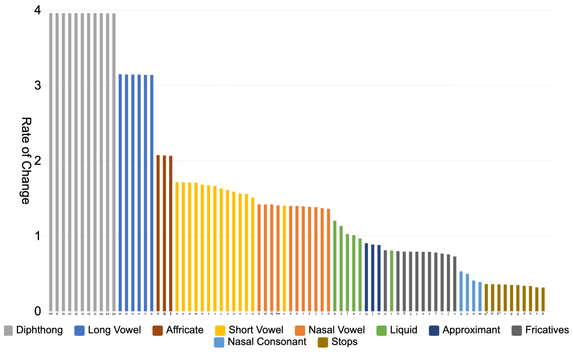

Figure 1. Segmental volatility

The event-based approach of the TKF91 model offers significant benefits. First, it allows scholars to take advantage of the rich information in words when drawing phylogenetic inferences. Second, it has enormous potential for phonology, since it provides the first phylogenetically based method for estimating the evolutionary stability of phonemes and phonetic segments. More broadly, the TKF91 model brings linguistic phylogenetics closer to the study of molecular phylogenetics, in as much as segmental sequences parallel those of nucleotides.

References

Bouckaert, R. R., Lemey, P., Dunn, M., Greenhill, S. J., Alekseyenko, A. V., Drummond, A. J., Gray, R. D., Suchard, M. A., & Atkinson, Q. D. (2012). Mapping the origins and expansion of the Indo-European language family. *Science*, *337*(6097), 957–960.

Chang, W., Cathcart, C. A., Hall, D. P., & Garrett, A. J. (2015). Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language*, *91*(1), 194–244.

Greenhill, S. J., & Gray, R. D. (2012). Basic vocabulary and Bayesian phylolinguistics. *Diachronica*, *29*(4), 523–537.

Lunter, G. A., Miklós, I., Song, Y. S., & Hein, J. (2003). An efficient algorithm for statistical multiple alignment on arbitrary phylogenetic trees. *Journal of Computational Biology*, *10*(6), 869–889.

Thorne, J. L., Kishino, H., & Felsenstein, J. (1991). An evolutionary model for maximum likelihood alignment of DNA sequences. *Journal of Molecular Evolution*, *33*(2), 114–124.

Wright, A. M., Lloyd, G. T., & Hillis, D. M. (2016). Modeling character change heterogeneity in phylogenetic analyses of morphology through the use of priors. *Systematic Biology*, *65*, 602–611.