Phonotactic variation in Pama-Nyungan tree inferece

Summary: This paper evaluates whether phylogenetic tree inference in linguistics is strengthened by the inclusion of phonotactic information. We take ~2k binary phonotactic variables and several hundred frequency variables and combine them with lexical cognate data from 111 Pama-Nyungan languages. The first part of the study explores the evolutionary dynamics of the phonotactic data. This is necessary to ascertain the best evolutionary model with which to infer a tree, since no one has used this kind of data in linguistic tree inference before. The second part of the study compares two models for inferring a Pama-Nyungan phylogeny using Bayesian methods. In one, a phonotactic data partition and lexical cognate partition are used jointly to infer trees. In the other these partitions are kept separate for the purpose of tree inference. Bayes factors for these two models are compared. We find that the combination of phonotactic data with lexical data **does/does not** significantly strengthen tree inference.

Notes:

Bayes factors are robust (Brown & Lemmon 2007)

1 Introduction

This section clearly needs fleshing out, but actually not too much. Aim is to keep it sharp and concise.

Background:

Phylogenies in linguistics are a big deal.

Lots of tree building been happening.

Phylogenies are also crucial for advances in comparative langauge sciences, studies of human history generally.

Data mainly limited to cognates. Some use of structural characters, but these tend to suffer from restricted state space.

In biology, Parins-Fukuchi (2018) find that combining continuous morphological characters to more traditional, categorical data can strengthen tree inference. An example of integration of continuous morphological data and genomic data Dömel et al. (2019).

Prev. study (Macklin-Cordes, Bowern & Round 2020) found phylogenetic signal in phonotactics. The hypothesis was that phonotactic systems are likely to evolve in an historically conservative way, reflect linguistic phylogenies and therefore be useful for tree inference. That finding was encouraging support for this hypothesis but not definitive proof by any means. Just because something has phylogenetic signal does not mean, by itself, that you can infer phylogenetic trees from it. For example, geography often has a pretty strong phylogenetic signal. In this study, we put the hypothesis to the test by attempting to infer a linguistic phylogeny with the aid of phonotactics.

2 Data and methods

Cognate data comes from (Bouckaert, Bowern & Atkinson 2018). Phonotactic data comes from Ausphonlex database of Australian language lexicons (Round 2017), which extends the Chirila database (Bowern 2016) by providing phonemicised wordforms and various parameters for phonemic normalisation choices between wordlists. In this study, we restrict attention to wordlists which i) represent Pama-Nyungan language varieties that are also included in Bouckaert, Bowern & Atkinson (2018), ii) have been published or are publicly accessible in some way, iii) have been compiled by trained linguists and iv) were compiled using some degree of in-person elicitation or audio recordings (reconstitutions using exclusively archival written records were not included). 111 Ausphonlex wordlists meet these criteria. Original wordlist sources and phonemic normalisation choices are listed in the Supplementary Materials.

Insert map of languages around here. Centroids colour-coded by subgroup.

From each wordlist, we extract data on the presence and frequencies of *biphones*, sequences of two segments (where each segment is either a phoneme or a word boundary). We extract two datasests. The first is a binary dataset marking the presence or absence of a given biphone in a language. A biphone is marked '1' if it is present in a language's wordlist (even if only once). If the biphone consists of two segments that are part of the language's phonemic inventory (and therefore the biphone could, in principle, occur in the language) but the biphone never occurs, it is marked '0' for absent. If one or both segments in the biphone are not part of the language's phonemic inventory, then it is marked as a gap '-' in the data. The second dataset A language's phonotactic system consists of rules governing how phonemic segments may combine into larger syllables and words. To represent phonotactics, we extract data on the presence and frequencies of *biphones*, two-segment sequences, from language wordlists. The second dataset extracts frequencies of transitions between segments. We extract forward transition frequencies—that is, the frequency of segment y following segment x, normalised over all instances of x. We also extract backward frequency transitions—the frequency of segment x preceding segment y, normalised over all instances of y.

We are motivated to extract these frequency datasets for a couple of reasons. Firstly, it allows us to capture a finer grained level of information than binary data would allow. Binary data is more similar to the kind of phonotactic information one might find in a published language grammar, where a description of phonotactics that one would typically encounter involves a series of statements on the (binary) permissibility or otherwise of certain combination of segments. This information does not, however, account for quantitative differences between common, high frequency sequences of segments versus dispreferred sequences that rarely arise in a language's lexicon. There is considerable evidence to suggest that speakers are psychologically attuned to these kinds of phonological frequencies (Coleman & Pierrehumbert 1997, Zuraw 2000, Ernestus & Baayen 2003, Albright & Hayes 2003, Eddington 2004, Hayes & Londe 2006, Gordon 2016). The second reason is that the relatively rapid, semi-automated extraction of transition frequencies from wordlists captures structural variation between languages at a scale and degree of precision that would be difficult to attain from manual data coding methods (as preferred for the coding of lexical cognate data and grammatical data used in previous linguistic phylogenetic work). Macklin-Cordes, Bowern & Round (2020) show that this transition frequency dataset contains stronger phylogenetic signal than its binary equivalent. There is one limitation of the frequency transition data, which is that presently we require positive values to use for tree inference (more on evolutionary models and tree inference below). Biphones of zero frequency (recorded as '0' in the binary dataset) get transformed to gaps in the dataset. By including the binary dataset in this study, we retain a distinction between biphones that are impossible in a language (because one or both of the segments are absent from the language's phonemic inventory) and biphones that are possible in principle but are never observed. Our phonotactic data captures information on which phonemic segments may combine immediately adjacent to one another and the frequencies at which they do so. This is phonotactics in the simplest sense, and does not directly capture phonotactic restrictions that depend on sequences beyond two segments, syllable structure or morpheme boundaries. Nevertheless, Macklin-Cordes, Bowern & Round (2020) confirm that this simple level of phonotactic data is sufficieent to detect strong phylogenetic signal.

Another argument might be that our method avoids *observer bias*. We don't have to rely on an expert picking and choosing which parts of a grammatical or lexical system are interesting and worth coding. This is described as an advantage of large-scale extraction of continuous morphological characters in biology too (Wright 2019). Another advantage of encoding structural variation with continuous characters over categorical ones: "phylogenetic error is very high for characters with ... very high rates of evolution (due to homoplasy of changes). Continuous characters do not display this relationship as strongly due to their large state space, though more research is needed to demonstrate this effect empirically." (Wright & Hillis 2014). Applicable to grammatical variables in linguistic phylogenetic tree inference, which show high rates of evolution and lots of homoplasy, due at least in part to tightly contrained state space (Greenhill et al. 2017). We don't have to worry about correcting for acquisition bias since the datasets reflect the full range of logically possible biphones in every language. We can include invariant sites (where all values are the same. These don't matter much for topology but are important for dating/branch lengths) and we don't need to correct for ascertainment bias (Leaché et al. 2015).

We use a Bayesian computational approach to infer linguistic phylogenies using BEAST phylogenetic software (v1.10.4) (Suchard et al. 2018). This is similar to earlier work on the Pama-Nyungan phylogeny (Bowern & Atkinson 2012, Bouckaert, Bowern & Atkinson 2018) which used BEAST2 (Bouckaert et al. 2019). We selected BEAST over BEAST2 because it offers the ability to infer trees with continuous characters. Throughout, we generally try to follow Bouckaert, Bowern & Atkinson (2018) as closely as possible. We follow Bouckaert, Bowern & Atkinson (2018) in constraining the tree topology using clade priors for well-established and commonly accepted Pama-Nyungan subgroups, as established by O'Grady, Voegelin & Voegelin (1966), Mühlhäusler, Tryon & Wurm (1996) and Koch & Nordlinger (2014) and subsequently recovered in computational phylogenetic analysis by Bowern & Atkinson (2012). Dating the Pama-Nyungan tree is a central focus of Bouckaert, Bowern & Atkinson (2018), combining lexical cognate data with geographical data and archaeological calibration points to give a best-available estimate of the geographic and temporal point of origin of the family. Accordingly, we retain their calibration prior on the Wati subgroup, which places a 95% probability of the subgroup's origin dating between 3,000-5,000 years, with most of the probability density skewing towards the younger end of that range (a gamma distribution of $\alpha = 2$, $\beta = 359$, with 3,000 year offset) based on a synthesis of archaeological evidence (see Bouckaert, Bowern & Atkinson 2018: p. 746). We place a prior on the root age of the Pama-Nyungan family centred on a mean of 5,791 years B.P., following the findings of Bouckaert, Bowern & Atkinson (2018). 5,791 years is the mean root age of the posterior for their best supported hypothesis on Pama-Nyungan's origins. We model this as a normal distribution (SD = 730) approximating the 95% range of posterior root age estimates. One aspect in which we differ from Bouckaert, Bowern & Atkinson (2018) is tip dates. Bouckaert, Bowern & Atkinson (2018) use a birth-death skyline tree model which allows for tip dates to differ and includes a parameter corresponding to the proportion of total taxa sampled at a given point in time. This is reasonable since they use language sources span over 200 years. In contrast, we assume all tips are contemporaneous. In our case, since we restrict attention to relatively modern sources, any extra precision to be gained from including tip dates is not worth the reduced tree model choice in BEAST and extra computational expence.

3 Results

3.1 Phonotactic evolutionary model

There have been a bunch of studies using lexical cognate data and some standards are beginning to emerge, for example covarion model seems widely preferred. [check state of the art language comparison article]. However, this is to the best of our knowledge the first attempt at tree inference with binary biphone characters as we use here [unless Gerhart tried it?] so we need to do a good deal of preliminary exploration testing various prior settings to get the best supported evolutionary model and set of sensible priors.

For each model specification, 2 independent chains of 25,000,000 iterations, with parameters logged every 10,000 iterations. Log marginal likelihood is calculated using BEAST's path sampling/stepping stone sampling procedure (Baele et al. 2012, 2013) consisting of 50 path steps of 500,000 iterations, with parameters logged every 10,000 iterations, conducted on each chain then combined to get an overall marginal likelihood. We conducted autocorrelation and convergence checks using Tracer v1.7.1 software (Rambaut et al. 2018). Note that the results here are a preliminary exploration of model parameters to determine the best parameter settings for the tree inference presented in Section 3.2 below. We do not anticipate that binary biphone characters will produce especially high quality or realistic language phylogenies on their own. The goal is to get a handle on how best to model the evolutionary dynamics of this dataset when used in combination with other sources of evidence.

3.1.1 Site model

We start by evaluating different site models that describe how binary biphone characters evolve through time. For this stage of evaluation, we fix the clock model to a strict clock (no variation in evolutionary rates between branches) and fix the tree model to a simple calibrated Yule tree model with a uniform birth rate

Table 1: Bayes factors for different site models. Each Bayes Factor represents the support for one model (listed left) against another (listed top). A positive value indicates the first model (left) is supported, and conversely, a negative value indicates the second model (top) is supported. A value over 100 is considered decisive.

Site model	SMH-SY	SSH-SY	SMG-SY	SSG-SY	CMH-SY	CSH-SY	CMG-SY	CSG-SY
SMH-SY	–	6	-1,277	-1,313	-47,211	-98,162	-94,904	-168,465
SSH-SY	-6	–	-1,283	-1,319	-47,217	-98,168	-94,910	-168,471
SMG-SY	1,277	1,283	–	-36	-45,934	-96,885	-93,627	-167,188
SSG-SY	1,313	1,319	36	–	-45,898	-96,849	-93,591	-167,152
CMH-SY	47,211	47,217	45,934	45,898	–	-50,951	-47,693	-121,254
CSH-SY	98,162	98,168	96,885	96,849	50,951	–	3,258	-70,303
CMG-SY	94,904	94,910	93,627	93,591	47,693	-3,258	–	-73,561
CSG-SY	168,465	168,471	167,188	167,152	121,254	70,303	73,561	–

prior (Yule tree models do not allow for extinction events). We then test all eight combinations of three site model parameters:

- A simple continuous time Markov chain (CTMC) model (which contains a single estimated parameter that specifies the frequencies with which biphones are gained and lost) versus a covarion model (which allows sites to switch between fast and slow states). The covarion model is the preferred model of lexical cognate evolution in Bouckaert et al. (2012), Bouckaert, Bowern & Atkinson (2018) and Kolipakam et al. (2018), although Chang et al. (2015: p. 219) find little difference between them and opt for the increased simplicity of the former model.
- Empirical character state frequencies versus estimated character state frequencies.
- Site homogeneity (fixed evolutionary rates across all character sites) versus heterogeneity (estimated using four gamma distributed categories, following Kolipakam et al. (2018)). For cognate data, Bouckaert, Bowern & Atkinson (2018) find a better fit with homgenous rates but Kolipakam et al. (2018) find a better fit with heterogenous ones.

We use Bayes factors to determine the best supported site model. Bayes factors give an indication of the support for one model over another and are calculated by calculating the ratio of the log marginal likelihoods of each model. A Bayes factor of 5 to 20 is taken as substantial support, greater than 20 as strong support, and greater than 100 as decisive (Kass & Raftery 1995). We table Bayes factors comparing each combination of site model settings in Table 2. The names of each model indicate site settings as follows: (S)imple CTMC versus (C)ovarion model, e(M)pirical versus e(S)timated character frequencies, (H)omogenous rates versus (G)amma-distributed heterogenous rates. All models contain the suffix "-SY" since they all contain a (S)trict clock and calibrated (Y)ule tree prior. So, for example, the model termed "CMH" consists of a covarion model with empirical frequencies and homogenous rates across all sites.

Table 2: Bayes factors for different site models. Each Bayes Factor represents the support for one model (listed left) against another (listed top). A positive value indicates the first model (left) is supported, and conversely, a negative value indicates the second model (top) is supported. A value over 100 is considered decisive.

Site model	CMG-R	CMG-S	CMH-R	CMH-S	CSG-R	CSG-S	CSH-R	CSH-S	SMG-R	SMG-H	SMH-R	SMH-S	SSG-R	SSG-S	SSH-R	SSH-S
CMG-R	–	20,274	44,531	20,746	-33,840	-16,924	42,417	22,034	59,797	59,966	61,049	61,250	59,793	59,932	61,036	61,246
CMG-S	-20,274	–	24,257	472	-54,114	-37,198	22,143	1,760	39,523	39,692	40,775	40,976	39,519	39,658	40,762	40,972
CMH-R	-44,531	-24,257	–	-23,785	-78,371	-61,455	-2,114	-22,497	15,266	15,435	16,518	16,719	15,262	15,401	16,505	16,715
CMH-S	-20,746	-472	23,785	–	-54,586	-37,670	21,671	1,288	39,051	39,220	40,303	40,504	39,047	39,186	40,290	40,500
CSG-R	33,840	54,114	78,371	54,586	–	16,916	76,257	55,874	93,637	93,806	94,889	95,090	93,633	93,772	94,876	95,086
CSG-S	16,924	37,198	61,455	37,670	-16,916	–	59,341	38,958	76,721	76,890	77,973	78,174	76,717	76,856	77,960	78,170
CSH-R	-42,417	-22,143	2,114	-21,671	-76,257	-59,341	–	-20,383	17,380	17,549	18,632	18,833	17,376	17,515	18,619	18,829
CSH-S	-22,034	-1,760	22,497	-1,288	-55,874	-38,958	20,383	–	37,763	37,932	39,015	39,216	37,759	37,898	39,002	39,212
SMG-R	-59,797	-39,523	-15,266	-39,051	-93,637	-76,721	-17,380	-37,763	–	169	1,252	1,453	-4	135	1,239	1,449
SMG-H	-59,966	-39,692	-15,435	-39,220	-93,806	-76,890	-17,549	-37,932	-169	–	1,083	1,284	-173	-34	1,070	1,280
SMH-R	-61,049	-40,775	-16,518	-40,303	-94,889	-77,973	-18,632	-39,015	-1,252	-1,083	–	201	-1,256	-1,117	-13	197
SMH-S	-61,250	-40,976	-16,719	-40,504	-95,090	-78,174	-18,833	-39,216	-1,453	-1,284	-201	–	-1,457	-1,318	-214	-4
SSG-R	-59,793	-39,519	-15,262	-39,047	-93,633	-76,717	-17,376	-37,759	4	173	1,256	1,457	–	139	1,243	1,453
SSG-S	-59,932	-39,658	-15,401	-39,186	-93,772	-76,856	-17,515	-37,898	-135	34	1,117	1,318	-139	–	1,104	1,314
SSH-R	-61,036	-40,762	-16,505	-40,290	-94,876	-77,960	-18,619	-39,002	-1,239	-1,070	13	214	-1,243	-1,104	–	210
SSH-S	-61,246	-40,972	-16,715	-40,500	-95,086	-78,170	-18,829	-39,212	-1,449	-1,280	-197	4	-1,453	-1,314	-210	–

The covarion model overwhelmingly outperforms the CTMC model in all instances. Furthermore, there is support for allowing evolutionary rates to vary across character sites. Unfortunately, this great increase in parameters results in a corresponding increase in computational demand. Models with heterogenous rates require 3–4 times as long as equivalent models with fixed rates. Lastly, there is decisive support for estimating character state frequencies rather than simply taking the observed frequencies when the covarion model is used, although the opposite is true with a CTMC model. A covarion model with estimated frequencies and homogenous evolutionary rates will beat a model where rates are allowed to vary but empirical frequencies are used. All up, we determine the best site model to be a covarion model with estimated frequencies and rate heterogeneity.

Note to self: Improper [0,Inf] uniform prior would have been okay for Yule birth rate since we have node calibrations (https://groups.google.com/forum/#!topic/beast-users/H_PjNgiZMe8). But I think [0,1] uniform prior is okay because the birth rate never gets anywhere near upper bound of 1 in the posteriors anyway. Kolipakam et al. (2018) uses bounded [0,1] birth rate while Bouckaert, Bowern & Atkinson (2018) uses [0,Inf].

Optional extension to this: Would be nice to test stochastic Dollo model, which has been implemented with some success for cognate data in linguistics (although covarion model seems to be winning these days). Stochastic Dollo only allows characters to spring into existance once and any losses are permanent. I wasn't too worried about SD because I figured it's a bit more realistic for cognates, since the state space of possible words is practically infinite (i.e. the chance of different people inventing the same word for the same thing independently is very low, although of course it does happen sometimes)[1]. By contrast, there are only so many possible biphone combinations, many unrelated/distantly related languages share biphones (consider, for example, shared biphones between English and Pama-Nyungan languages) and it seems fairly unreasonable to assume a single common point of origin for all of them. Nevertheless, it would nice to test to be sure. Unfortunately, I was getting some nasty errors in BEAST that seem difficult to resolve when the SD model is selected. I wasn't really worried about this because I figured covarion is likely more realistic anyway.

3.1.2 Clock and tree model

We take the best performing site model and compare it to the same model with a lognormally-distributed uncorrelated relaxed clock and a birth-death tree prior. This relaxed clock model generally has been found to outperform a strict clock when modelling lexical cognate evolution (Bouckaert, Bowern & Atkinson 2018, Kolipakam et al. 2018). The birth-death speciation model allows for extinction events and more closely approximates the birth-death skyline model favoured in Bouckaert, Bowern & Atkinson (2018), although a Yule speciation model was preferred in Bowern & Atkinson (2012) and Kolipakam et al. (2018).

For the relaxed clock, we used an uncorrelated lognormal setting with a uniform prior [0,1] following Kolipakam et al. (2018). Bouckaert, Bowern & Atkinson (2018) constrain the upper bound to 1.0E-4 to reduce burn-in time since, in practice, the mean clock value never approaches even that level. We chose the less informative upper bound given the uncertainty of working with a novel data type (but it doesn't matter too much).

Bayes factors are presented in Table 3. The model naming convention is as above. The suffix reflects the clock and tree prior settings: (S)trict clock versus (R)elaxed clock and calibrated (Y)ule versus (B)irth-death speciation. We find that the marginal likelihoods of relaxed clock models are decisively stronger than strict clock models, regardless of which tree prior is used. Of the two relaxed clock models, the calibrated Yule prior beats the birth-death model.

[1] That said, SD isn't super realistic for cognates either since it doesn't allow for borrowing, which appears as two independent origin points when plotted on a phylogenetic tree. This is likely why covarion tends to work better. As an aside, a dream phylogeographic model of cognate evolution would allow for independent points of origin with very low probability (the likelihood of chance resemblances) plus a relatively high probability of an independent point of origin springing up when a language is geographically adjacent to another where the cognate is already present (this wouldn't really reflect an independent point of origin but rather a borrowing). Computationally expensive though.

Table 3: Comparison of models with different clock and tree settings.

Clock/tree model	CSG-SY	CSG-RY	CSG-SB	CSG-RB
CSG-SY	–	-75,114	-9,571	-16,306
CSG-RY	75,114	–	65,543	58,808
CSG-SB	9,571	-65,543	–	-6,735
CSG-RB	16,306	-58,808	6,735	–

3.2 Combined cognate and phonotactics tree inference

Evolutionary model for phonotactic frequency dataset is more straightforward. We take a standard, lightweight Brownian motion model in which frequency values can wander up or down with equal probability through time. We are limited to this model by software constraints, but that is not a major limitation at this point. Firstly, Brownian motion is a standard starting point in comparable biological studies that jointly infer trees with continuous data. Secondly, it is the same model used in Macklin-Cordes, Bowern & Round (2020). One difference between Macklin-Cordes, Bowern & Round (2020) and this study is that Macklin-Cordes, Bowern & Round (2020) use raw frequency values whereas we use log-transformed frequency values. We observe that biphone transition frequencies tend to be skewed such that lexicons tend to contain relatively few high frequency biphone transitions and many low frequency transitions. It follows then that these biphone transition frequencies are more likely the outcome of an evolutionary process where characters wander along a skewed, lognormal scale than one in which they wander along a normal distribution (although, in practice, it may not matter too much. Macklin-Cordes, Bowern & Round (2020) find no significant difference in phylogenetic signal using raw values versus log-transformed values). These skewed distributions echo the skewed distributions of single segments observed by Macklin-Cordes & Round (2020). As Macklin-Cordes & Round (2020) makes clear, this does not mean that biphone transition frequencies are necessarily drawn from a lognormal distribution and a more sophisticated maximum likelihood test would be needed to distinguish between the lognormal and several other similarly skewed distribution types. Nevertheless, the lognormal distribution is a sufficient approximation of the skewed distribution of biphone transition frequencies for our purposes in this study.

Thinking briefly about what would be a realistic model of evolution for biphone transition frequencies. We would expect there to be two main forces impacting these frequencies. The first is the introduction of new vocabulary to a language via lexical innovation or borrowing. Each new word entering a lexicon will alter minutely the frequencies of biphone transitions in the language (similarly, transition frequencies will decline as words are replaced or fall out of usage). This is the kind of gradual accumulation of changes that we might expect to follow a Brownian motion-like pattern of evolution (although maybe the rates of going up and down are not equal). Further, since speakers show a preference for high frequency phonotactic sequences over low frequency sequences when coining new words, we might expect this accumulation of changes to follow a kind of 'rich get richer' process which would result in the kind of skewed frequency distributions that we observe. Also, when languages borrow vocabulary, the trend is for foreign words with dispreferred phonotactic sequences to shift towards more natively preferred patterns (sometimes gradually over a long period of time, i.e. look at various French words in English, stress has shifted to English pattern in some but not yet in others), which would strengthen this kind of 'rich get richer' process and also keep phonotactic frequency data historically conservative. The second major force on biphone frequencies is sound change. We would expect sound changes to result in sudden jumps in the frequencies of affected biphones, sometimes to 0 or 1. Our binary characters capture some of these effects to a limited extent. For example, perhaps a language has some frequency value for sequences of a nasal followed by a stop with a different place of articulation. If that nasal undergoes place assimilation, the biphone frequency will drop to 0 and thus disappears as a gap in the frequency dataset since evolutionary model requires non-zero values. On the other hand, this assimilation will be recorded in the binary data as a shift from '1' to '0'. In other instances, biphone characters may shift from missing to present and vice versa in both the frequency and binary datasets. For example, if a contrastive vowel length distinction emerges, certain biphones (namely those with long vowels) will go from being a gap in a language's biphone transition frequency data to some positive, non-missing value. In the case of a merger between short and long vowels, the opposite will be true. Our model, at present, simply does not account well for sound change. In this respect, there is an advantage to studying Australian languages, since Australian languages show uniquely constrained variation in phonological inventories [REFS] (easier to match biphones between languages, less dataset sparsity) and less history of identified sound changes relative to other parts of the world (historical linguists have long turned to sources of historical evidence in other parts of language like morphology etc. [REFS]). We return to this subject in Section 4.

For the cognate data partition, we approximate as much as possible the best supported priors from Bouckaert, Bowern & Atkinson (2018). We use a covarion model with a relaxed clock and fixed rates across cognate classes.

4 Discussion

Discussion in biology regarding combination of morphological and genomic datasets. "Simultaneous" approach where both morphological and genomic data are used jointly to infer the tree versus "scaffolding" approach where only genomic data is used to infer tree topology, then morphological data is used to assess e.g. dating (using fossil record) while being constrained to genomic tree topology (Lee & Palci 2015). Must be aware of the potential circularity of tracing the evolution of characters on a phylogeny which was itself partly based on those characters (de Queiroz 1996).

Limitations:

- Logical dependencies between variables (because of sound changes, phonotactic restrictions affecting natural classes)
- Logical dependencies between binary/continuous partitions (non-gap in freq data = 1 in binary data. 0 in binary data = gap in freq data)
- Didn't account for sound change
- Limitations of Brownian motion model

If we get a negative result (no significant difference between trees inferred with/without phonotactic data partition) then I would speculate that it's probably got a lot to do with the inability of our Brownian motion evolutionary model to capture the effects of sound change, which would manifest as sudden jumps in frequencies.

If we get a positive result, then we would advocate for the use of phonotactic data in combination with other sources of evidence, such as cognate data, to infer linguistic phylogenies.

- Could be used to help resolve phylogenetic conflicts in places where there is more phylogenetic uncertainty. Could be used to help with dating and branch lengths in places where otherwise the topology is quite well understood.
- Could help in under-resourced places that don't have as much lexical data. Studies of Pama-Nyungan phylogeny have benefitted from reasonably extensive cognate coding over nearly 300 meaning classes, but a lot of places will be limited to the scale of Swadesh lists or even less. (The opposite is true in biology, where morphological datasets make up ever shrinking proportion of total combined dataset when combined with genomic datasets that keep getting bigger)
- Could be used for quick and dirty tree inference where some phylogenetic information is required/better than nothing (for example, using phylogenetic comparative methods) but doesn't necessarily have to be perfect. e.g. could combine with very small lexical datasets/automatic cognate identification. Perhaps could be combined with, e.g. glottolog classifications to get something consistent with glottolog tree but fully resolved.

References

Albright, Adam & Bruce Hayes. 2003. Rules vs. analogy in english past tenses: a computational/experimental study. *Cognition* 90(2). 119–161. https://doi.org/10.1016/S0010-0277(03)00146-X.

Baele, Guy et al. 2012. Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Molecular Biology and Evolution* 29(9). 2157–2167. https://doi.org/10.1093/molbev/mss084.

Baele, Guy et al. 2013. Accurate model selection of relaxed molecular clocks in Bayesian phylogenetics. *Molecular Biology and Evolution* 30(2). 239–243. https://doi.org/10.1093/molbev/mss243.

Bouckaert, Remco et al. 2012. Corrections and clarifications. *Science* 342(6165). 1446. https://doi.org/10.1126/science.342.6165.1446-a.

Bouckaert, Remco et al. 2019. BEAST 2.5: an advanced software platform for bayesian evolutionary analysis. *PLOS Computational Biology* 15(4). e1006650. https://doi.org/10.1371/journal.pcbi.1006650.

Bouckaert, Remco R., Claire Bowern & Quentin D. Atkinson. 2018. The origin and expansion of Pama-Nyungan languages across Australia. *Nature Ecology & Evolution* 2(4). 741–749. https://doi.org/10.1038/s41559-018-0489-3.

Bowern, Claire. 2016. Chirila: contemporary and historical resources for the indigenous languages of Australia. *Language Documentation and Conservation* 10. http://hdl.handle.net/10125/24685.

Bowern, Claire & Quentin D. Atkinson. 2012. Computational phylogenetics and the internal structure of Pama-Nyungan. *Language* 88(4). 817–845. https://doi.org/10.1353/lan.2012.0081.

Brown, Jeremy M. & Alan R. Lemmon. 2007. The importance of data partitioning and the utility of bayes factors in bayesian phylogenetics. *Systematic Biology* 56(4). Publisher: Oxford Academic, 643–655. https://doi.org/10.1080/10635150701546249. http://academic.oup.com/sysbio/article/56/4/643/1684114 (26 September, 2020).

Chang, Will et al. 2015. Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language* 91(1). 194–244. https://doi.org/10.1353/lan.2015.0005.

Coleman, John & Janet Pierrehumbert. 1997. Stochastic phonological grammars and acceptability. In *Computational phonology: the third meeting of the ACL special interest group in computational phonology*, 49–56. Somerset, NJ: Association for Computational Linguistics. http://arxiv.org/abs/cmp-lg/9707017 (8 March, 2018).

Dömel, Jana S. et al. 2019. Combining morphological and genomic evidence to resolve species diversity and study speciation processes of the Pallenopsis patagonica (Pycnogonida) species complex. *Frontiers in Zoology* 16(1). 36. https://doi.org/10.1186/s12983-019-0316-y.

Eddington, David. 2004. *Spanish phonology and morphology: experimental and quantitative perspectives*. Amsterdam: John Benjamins.

Ernestus, Mirjam Theresia Constantia & R. Harald Baayen. 2003. Predicting the unpredictable: Interpreting neutralized segments in dutch. *Language* 79(1). 5–38. https://doi.org/10.1353/lan.2003.0076.

Gordon, Matthew K. 2016. *Phonological typology* (Oxford Surveys in Phonology and Phonetics 1). Oxford: Oxford University Press.

Greenhill, Simon J. et al. 2017. Evolutionary dynamics of language systems. *Proceedings of the National Academy of Sciences* 114(42). E8822–E8829. https://doi.org/10.1073/pnas.1700388114.

Hayes, Bruce & Zsuzsa Cziráky Londe. 2006. Stochastic phonological knowledge: the case of Hungarian vowel harmony. *Phonology* 23(1). 59–104. https://doi.org/10.1017/S0952675706000765.

Kass, Robert E. & Adrian E. Raftery. 1995. Bayes factors. *Journal of the American Statistical Association* 90(430). 773–795. https://doi.org/10.1080/01621459.1995.10476572.

Koch, Harold & Rachel Nordlinger. 2014. *The languages and linguistics of australia: a comprehensive guide*. Walter de Gruyter GmbH & Co KG.

Kolipakam, Vishnupriya et al. 2018. A Bayesian phylogenetic study of the Dravidian language family. *Royal Society Open Science* 5(3). 171504. https://doi.org/10.1098/rsos.171504.

Leaché, Adam D. et al. 2015. Short tree, long tree, right tree, wrong tree: new acquisition bias corrections for inferring SNP phylogenies. *Systematic Biology* 64(6). 1032–1047. https://doi.org/10.1093/sysbio/syv053.

Lee, Michael S. Y. & Alessandro Palci. 2015. Morphological phylogenetics in the genomic age. *Current Biology* 25(19). R922–R929. https://doi.org/10.1016/j.cub.2015.07.009. http://www.sciencedirect.com/science/article/pii/S096098221500812X.

Macklin-Cordes, Jayden L., Claire Bowern & Erich R. Round. 2020. Phylogenetic signal in phonotactics [submitted]. author-submitted preprint. arXiv:2002.00527.

Macklin-Cordes, Jayden L. & Erich R. Round. 2020. Re-evaluating phoneme frequencies. *arXiv:2006.05206 [physics, stat]*. http://arxiv.org/abs/2006.05206.

Mühlhäusler, Peter, Darrell T Tryon & Stephen A Wurm. 1996. *Atlas of languages of intercultural communication in the pacific, asia, and the americas: vol i: maps. vol II: texts.* Berlin: De Gruyter.

O'Grady, Geoffrey N., Charles Frederick Voegelin & Florence M. Voegelin. 1966. Languages of the world: Indo-Pacific fascicle six. *Anthropological Linguistics*. 1–197.

Parins-Fukuchi, Caroline. 2018. Use of continuous traits can improve morphological phylogenetics. *Systematic Biology* 67(2). 328–339. https://doi.org/10.1093/sysbio/syx072.

de Queiroz, Kevin. 1996. Including the characters of interest during tree reconstruction and the problems of circularity and bias in studies of character evolution. *The American Naturalist* 148(4). 700–708.

Rambaut, Andrew et al. 2018. Posterior summarization in bayesian phylogenetics using tracer 1.7. *Systematic Biology* 67(5). 901–904. https://doi.org/10.1093/sysbio/syy032.

Round, Erich R. 2017. The AusPhon-Lexicon project: 2 million normalized segments across 300 Australian languages. In *47th poznań linguistic meeting*. Poznań, Poland. http://wa.amu.edu.pl/plm_old/2017/files/abstracts/PLM2017_Abstract_Round.pdf.

Suchard, Marc A. et al. 2018. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution* 4(1). https://doi.org/10.1093/ve/vey016.

Wright, April M. 2019. A systematist's guide to estimating bayesian phylogenies from morphological data. *Insect Systematics and Diversity* 3(3). https://doi.org/10.1093/isd/ixz006.

Wright, April M. & David M. Hillis. 2014. Bayesian analysis using a simple likelihood model outperforms parsimony for estimation of phylogeny from discrete morphological data. *PLoS ONE* 9(10). https://doi.org/10.1371/journal.pone.0109210.

Zuraw, Kie Ross. 2000. *Patterned exceptions in phonology*. Los Angeles: University of California dissertation.