Dyr, a Program for Bayesian Inference of Language Phylogenies

Student Name: J.G. Byrne
Supervisor Name: Professsor M.J.R. Bordewich
Submitted as part of the degree of MEng Computer Science to the
Board of Examiners in the Department of Computer Sciences, Durham University

Abstract—Bayesian statistics provide a powerful framework for inferring the evolutionary history of language families from lexical data, a problem which is made computationally tractable by Markov Chain Monte Carlo (MCMC) algorithms. We present Dyr, a simple yet capable new system for Bayesian inference, and use it to replicate state-of-the-art results in the field of Indo-European linguistics.

Index Terms—Bayesian Inference, Markov Processes, Linguistics, Indo-European

1 Introduction

Languages evolve over time. Some of these changes are phonetic or grammatical, but many are lexical (changes in the vocabulary). Words are forgotten and replaced, sometimes by other words that have changed meaning, and sometimes by borrowings from other languages.[1] Occasionally, some speakers of a language will come to speak so differently to the others that the groups can no longer understand each other, and the language splits in two.

Scholars have noticed similarities between languages for millenia. In the 18th century, the field of comparative linguistics was born, which sought to systematically reconstruct the historical relationships between languages. The early comparative linguists observed that while languages can evolve, sunder, and go extinct, they rarely merge, and are never created anew. Therefore, from the 19th century onwards it became *de rigeur* to describe language descent with evolutionary trees, or 'phylogenies', and though not without criticism, this model remains predominant in linguistics to this day.

Historically, constructing phylogenies has been a human labour, relying on the judgement of expert scholars with deep knowledge of languages ancient and modern. Although fruitful, this approach is naturally susceptible to human biases and oversights. It also provides no way to determine the age of unattested ancestor languages other than the (admittedly finely-honed) intuition of learned intellectuals.

The problem of ancestral dating is of particular relevance to Indo-European, the largest and most studied language family in the world. Encompassing nearly all of the languages of Europe and a great many in Western Asia and India, the challenge of reconstructing the history of this vast grouping has captivated comparative linguists since the advent of the discipline. And yet, one central question is yet to be conclusively answered: where and when was Proto-Indo-European (PIE; the ancestor of all Indo-European languages) originally spoken? Two main theories abound. The first, known as the 'Kurgan' hypothesis, postulates an *Urheimat* (original homeland) in the Pontic-Caspian steppe north of the Black Sea.[2] Conversely, the second proposes an origin in Anatolia.[3] The crucial difference between the two theories is that while the former ascribes an age of 6000 years to PIE, the latter hinges on it being considerably more ancient, at approximately 8500 years old. A reliable estimate for the age of PIE - that is, the root age of the Indo-European language tree - could therefore be potent evidence for one theory over the other.

Since the millenium, researchers have found a new method to analyse language families in general and Indo-European in particular – Bayesian inference. By reappropriating software designed to analyse genetic data and construct biological evolutionary trees, they have successfully inferred plausible dated phylogenies from large vocabulary databases. Such research feels tantalisingly close to an objective solution to linguistic enigmas such as that of Indo-European provenance.

However, in reality, the human factor still plays a role. Bayesian inference requires the specification of prior distributions, the choice of which can radically affect the results of the inference. For research to be rigorous and reliable, priors should be chosen carefully and with good justification.

Yet much study in the relatively small field of 'Bayesian Phylolinguistics' falls short of this mark. Priors are often chosen seemingly out of habit or convention. Such laxness poses particular danger when these conventions have their origins in bioinformatics; a model which is sensible in the context of biological evolution may not be so logical when applied to languages.

1. For example, the Old English word *dēor* was displaced by the French 'animal', and survives only in the narrower sense 'deer'. However, the equivalent word in Danish, *dyr*, whence this project name, retains the older meaning. Both words derive from Proto-Germanic **deuzą*, ultimately from Proto-Indo-European **d^hwes*, meaning 'breath'.

2. A *kurgan* is a tumulus or burial mound. This theory is associates the spread of Indo-European with warlike tumulus-building charioteers.

3. This rather more sedate theory suggests a gradual proliferation of Indo-European in tandem with the spread of agriculture

We suggest that one reason for this tendency is the reliance on large, featureful bioinformatics software packages like BEAST2 and MrBayes. Any dedicated support for linguistics in these programs tends to be something of an afterthought, and their intimidating size makes them hard to understand in their totality. Therefore, we submit that to step out of the shadow of the older, larger discipline, it is desirable for Bayesian Phylolinguistics to have its own dedicated software package.

We therefore present *Dyr*, a new program for Bayesian Inference of linguistic phylogenies. Small enough that its source code may be read and understood in a day, and offering only the features required for linguistic analysis, *Dyr* is nonetheless built to be flexible and extensible. Capable of replicating state-of-the-art results in Bayesian Phylolinguistics, our hope and intention is that *Dyr* can serve as a worthy platform for trialling novel methods with sound linguistic justification.

2 Related Work

This section presents a survey of existing work on the problems that this project addresses. It should be be-tween 2 to 4 pages in length. The rest of this section shows the formats of subsections as well as some general formatting information for tables, figures, references and equations.

3 Methods

The aim of phylogenetic inference is to learn which phylogenies (dated trees) are most likely given a particular evolutionary model and a set of known data. This is known as the 'posterior likelihood'. In a linguistics context, the known data is typically 'lexical trait data' – that is, information about the vocabulary of the languages in question. The evolutionary model, which in Bayesian terms provides our 'prior likelihood', encapsulates our beliefs about how likely languages are to diverge and how rapidly their vocabulary changes. To allow our inferred phylogeny to best fit the signal present in the data, we also infer some parameters of our evolutionary model, though these too are subject to their own prior likelihood distributions.

The space of possible paramaterisations is very large and it is not typically feasible to derive a closed-form expression for the posterior likelihood distribution. However, Bayes' theorem allows us to assess the posterior likelihood of any specific choice of paramaterisation given the data. We therefore use a stochastic process called Markov Chain Monte Carlo (MCMC) to iteratively step-through the space of possible paramaterisations, with a preference for augmentations to the paramaterisation that improve the posterior likelihood. It is provable that the long-run outcome of this process will be to simulate the desired posterior distribution.

3.1 Lexical Trait Data

The primary evidence used to infer linguistic phylogenies is lexical trait data. In principle, many different linguistic features could be appropriated to inform Bayesian inference, but for both principled and pragmatic reasons the vast majority of analyses use lexical data. In particular, the class of traits used in our datasets is what Chang et al. termed Root Meaning traits. These traits can be described as tuples in the form (*root* , `semantic`). A Root Meaning trait is binary; either present or absent for a given language. A given trait is present in a language if that language has a common word for the `semantic` that derives from the *root*. For example, the Irish word for a `fish` is 'iasc', which like the English 'fish', comes from the Indo-European root **peysk-*. So the trait (**peysk-* , `fish`) is present in both Irish and English. However the Greek word for `fish` is 'ikhthus', which is believed to derive from the root **d^hg^hu-*. Therefore the trait is absent in Greek.

From a great many such traits is constructed the IELEX database, the gold-standard source of lexical data for Indo-European, upon which we shall found our analyses. In particular, we use the subsets of IELEX constructed by Chang et al., termed Narrow, Medium, and Broad, containing 52, 82, and 94 languages respectively. In our inferences, these languages will correspond to leaves in our phylogenies.

3.2 Defining the Posterior Distribution

Bayes' theorem is stated as follows:

$$Pr(A \mid B) = \frac{Pr(B \mid A) \cdot Pr(A)}{Pr(B)} \tag{1}$$

In the context of phylogenetic inference, we seek to infer the probability distribution of possible parameterisations given the observed data. Therefore $Pr(B)$ corresponds to $Pr(x)$, the probability of the trait data, which is by definition equal to 1 since it has been observed. Meanwhile, $Pr(A)$ corresponds to $Pr(\Gamma)$, the prior likelihood of a given parameterisation. We define Γ more thoroughly below:

$\Gamma = (\psi, \omega, \lambda)$

$\psi =$ parameters describing a dated tree

$\omega =$ parameters of the prior model for dated trees

$\lambda =$ parameters of the prior model for trait evolution

We thus derive equation (2). The posterior likelihood is proportional to the likelihood of the data given the paramaterisation multiplied by the prior likelihood of the paramaterisation. This is a proportionality because we do not require that our prior distributions sum to 1.

$$Pr(\psi, \omega, \lambda \mid x) \propto Pr(x \mid \psi, \omega, \lambda) \cdot f(\psi \mid \omega) \cdot f(\omega) \cdot f(\lambda) \tag{2}$$

We define $\psi = (\tau, \delta)$, where τ is the topology of the tree and δ is the branch lengths. We consider the node or nodes with the longest path from the root to be at $t = 0$ and therefore in the present day, while all other nodes are understood to be at some $t > 0$, corresponding to their distance from the present in years. Naturally, any interior node is required to be older than its children.

3.3 Calculating Likelihood

Although Bayes' theorem absolves us of the need to directly calculate the posterior, we must still calculate the likelihood of the data given the paramaterisation. Conceptually speaking, this is the likelihood across all possible trait assignments (that is, a binary array of absences and presences) across all possible nodes of the tree of the observed trait data being

evolved at the tips, under a given process of trait evolution we call the substitution model.

At first glance, this calculation sounds intractable. However, it can in fact be computed, with a divide-and-conquer method called Felsenstein's algorithm. The procedure is fairly inutitive: for each trait, we first define the trait's likelihood at the tips, based on the observed data. Then we work our way up the tree, deriving the 'partial' likelihoods at each interior node based on its two children, until we reach the root, where we can sum up a final likelihood for the trait across all assignments. The overall likelihood of the tree given the data is therefore simply the product across all traits.

Equation (3) below gives a formal definition of Felsenstein's algorithm. We denote the likelihood of a node x having state t for the trait i as $L^i_x(t)$, and it is defined separately for leaf nodes ℓ and for interior nodes a. It's worth remembering that for our purposes the set of states is simply $t \in \{0, 1\}$, but the algorithm is best understood in its full generality. When a leaf node ℓ has recorded data for trait i, the value $x^i_\ell[\,t\,]$ is 1.0 for the observed state t and 0.0 otherwise. When instead, the data is missing, the likelihood is split evenly between all values of t. For an interior node a, the likelihood of having state t for the trait i is calculated based on the product of the sum likelihood across all possible ways t can evolve along the branch to the left child a and the equivalent calculation along the branch to the right child b. The evolution probability is the core calculation performed by the substitution model and is predicated on the parameters λ. Note how the likelihood is thus recursively 'rolled-up' towards r, the root node of the tree.

$$
\begin{aligned}
L^i_\ell(t) &= x^i_\ell[\,t\,] \\
L^i_a(t) &= \left(\sum_u Pr^i_{a,b}(u \mid t)\, L^i_b(u)\right)\left(\sum_v Pr^i_{a,c}(v \mid t)\, L^i_c(v)\right) \\
\mathcal{L}^i &= \sum_u \pi_u \cdot L^i_r(u) \\
\mathcal{L} &= \prod_i \mathcal{L}^i
\end{aligned}
\tag{3}
$$

We denote the root likelihood across all states of trait i as $\mathcal{L}^i$. The calculation required is a weighted average, with the weights being the stationary frequencies π_u, which will be discussed shortly along with the other parameters in λ.

Calculating the final root likelihood is then a simple matter of taking the product of all of the per-trait root likelihoods. This value, $\mathcal{L}$, is the value of the term $Pr(x \mid \psi, \omega, \lambda)$ in equation (2).

We now seek to explain how the substitution model is used to calculate the probability of a state t evolving to a state u along a given edge of a tree. Let us state upfront the parameters of the model:

$\lambda = (\mu,\ \pi,\ \alpha,\ \beta,\ \phi)$
$\mu =$ Substitution base rate
$\pi =$ Stationary frequencies: $\pi_0, \pi_1 \in (0, 1) : \pi_0 + \pi_1 = 1.0$
$\alpha =$ Shape for Among Site Rate Variation (ASRV)
$\beta =$ Shape for Among Branch Rate Variation (ABRV)
$\phi =$ ABRV Rate Assignments

The core of the substitution model is the rate matrix, which describes the rates at which states mutate into each other. These matrices can be quite complex, but we opt for the simplest choice, the Generalised Time Reversible (GTR) model.[4] We have already seen how the stationary frequencies π_0 and π_1 influence the likelihood calculation, but they are also at the heart of the binary GTR rate matrix.

$$
\mathbf{Q} = \frac{1}{2\pi_0\pi_1}\begin{bmatrix} -\pi_1 & \pi_1 \\ \pi_0 & -\pi_0 \end{bmatrix} \text{ given that } \pi_0 + \pi_1 = 1.0 \tag{4}
$$

As the length of time over which a trait evolves increases, the probability of it being in a state u asymptopically approaches π_u, and the state it was initially in becomes irrelevant. To be more precise, for every branch (a, b) in the tree ψ, for every trait i, the transition probabilities $Pr^i_{a,b}(u \mid t)$ are defined by the transition matrix $\mathbf{P}^i_{a,b}$, where:

$$
\mathbf{P}^i_{a,b} = \exp\left(\mathbf{Q} \cdot \delta_{a,b} \cdot \eta^i_{a,b}\right) \text{ where } \eta^i_{a,b} = \mu \cdot \gamma_i \cdot \rho_{a,b} \tag{5}
$$

As previously stated, the value $\delta_{a,b}$ is the length (in years) of the branch (a, b). The value $\eta^i_{a,b}$ is the rate for the trait i on the branch (a, b), calculated as the product of three rate parameters. The first rate, μ, is the base rate – a global parameter that cancels out the units of the branch lengths and controls the overall rate of evolution. The second, γ_i, is the site rate,[5] which is specific to this trait i. It is drawn from a Gamma distribution $\Gamma(\alpha, \frac{1}{\alpha})$. The third rate, $\rho_{a,b}$, is the branch rate, which is specific to this branch. It is drawn from a log-normal distribution $\log\mathcal{N}(-\frac{\beta^2}{2}, \beta^2)$. Both of these distributions have a mean of 1, so regardless of their shapes the overall average rate is equal to μ. The choices of distributions are partly informed by implementation pragamatics and partly by the flexibility in shape that can be attained by modfications to α and β. [6]

At this point the reader may be questioning the necessity of allowing rates to vary both across sites and and across branches. However there is a strong justification for both laxities. Variation in site rate is necessary because words vary greatly in their volatility. Certain common words, in particular numerals, change incredibly rarely. Others are considerably more susceptible to replacement. Equally, branch rates must be variable because languages evolve at dramatically different rates. A failure to account for this fact was the downfall of much early research into quantitiative historical linguistics. The classic (though by no means sole) exemplar is the case of the Nordic languages; while Norwegians find Old Norse virtually incomprehensible, Icelanders can read it as easily as an Englishman can read Jane Austen.

4. Chang et al. calls this the Restriction Site Character (RSC) model; the two are equivalent in the case of binary traits.

5. The term 'site' is a relic from bioinformatics, where traits typically correspond to specific sites on the genome

6. This paragraph abstracts quite significantly over the specifics of how these rates are actually chosen, which shall be discussed later.