

Inferring sound changes using Bayesian MCMC

Abstract

In this paper we analyze dialect phonetic data using Bayesian Monte Carlo Markov Chain inference (MCMC), in recent years one of the most powerful and the most successful methods in molecular phylogeny for inferring the relationships between species. This method enables us to infer the historic relationships between the language varieties, but also to explore historical accounts of the sound correspondences in the data set. The dialect divisions obtained by applying Bayesian MCMC inference are compared to the divisions described in traditional scholarship. In this experiment we also test three different models of vowel evolution and show which changes are most likely within the vowel space.

1 Introduction

In the last decade there has been an increasing interest in the application of the methods taken from phylogenetics to language data. This line of research starts from the premise that there is a genuine similarity between the evolution of species and the evolution of languages. Although there are some important differences in their evolution, the mechanisms of change of species and languages are the same: they split into new species/languages, mutate, borrow material from neighboring species/languages (although this is rare in advanced species), and normally, once populations are independent, they innovate independently. They both document evolutionary history, species in molecules and various morphological characteristics, languages in phonetics/phonology, morphology, syntax. The evolution and relatedness of both biological species and languages can be described using family trees.

Phylogenetics is a branch of biology that studies the evolutionary relatedness among various species. The relatedness can be inferred at the molecular level by examining the differences between DNA or protein sequences of the organisms. Closely related organisms have similar DNA (protein) sequences, in particular similar orders of the nucleotides (amino acids) in their DNA (protein) sequences. More distantly related organisms show more dissimilarity if we compare their DNA (protein) sequences. Another approach to phylogenetic inference is to compare various morphological characteristics of the organisms. In this chapter we focus on molecular phylogenetics and try to use some of the models developed for the evolution of DNA and protein sequences on language data, focusing on sound correspondences.

Methods taken from computational phylogenetics have been applied to lexical (Gray and Jordan 2000; Gray and Atkinson 2003) and phonetic data (Warnow 1997; Nakhleh et al. 2005) to study evolutionary relationships between languages or dialects (Hamed 2005; Hamed and Wang 2006; McMahon et al. 2007). They have been used to address the problems of the origins of Indo-European (Gray and Jordan 2000) and Bantu languages (Holden 2002; Holden and Gray 2006). They were also applied to the problems of the subgrouping of Indo-European (Ringe et al. 2002; Nakhleh et al. 2005), as well as to test various hypotheses about human prehistory (Dunn et al. 2005; Greenhill and Gray 2005; Gray et al. 2009). As pointed out in Greenhill and Gray (2009), computational phylogenetic methods are seen as ‘a powerful supplement to the comparative method used in historical linguistics’. They are not a replacement for well-established

methods of comparative reconstruction in linguistics, but they can help in resolving some rather old questions in the history of languages. Although developed to work with different types of data, the use of the techniques developed for phylogenetic inference opens new perspectives in the field of historical linguistics. However, the use of the techniques does not come without its problems and concerns. Although they share the same mechanisms of change, species and languages differ in many ways. Languages change much faster than species. Borrowing between neighboring languages, regardless of their genetic relatedness, is much more common than between species. The two most important preconditions for analyzing languages using methods from phylogenetics are adequate linguistic data coding and the choice of an appropriate model of language change. If the data employed in the analyses is not well analyzed and coded, one will obtain wrong results. The same holds for a poor choice of evolutionary models. All models implemented in the computational phylogenetic software are naturally designed to cover various aspects of the evolution of species. Many models cannot be applied to the linguistic data since the assumptions behind those models violate the known facts of the linguistic change. But those that fit linguistic data well are a good start for the systematic exploration of language data, as they enable researchers to analyze larger bodies of data while consistently controlling many aspects of analysis.

This paper reports on work that is innovative in three respects. First, in automatically extracting sound correspondences. Earlier researchers have manually extracted selected sound correspondences and included them in their phylogenetic inference, but we present a technique for extracting large numbers of correspondences automatically. Second, in applying phylogenetic inference to a substantial amount of phonetic material. In this paper we apply Bayesian methods used to infer phylogenies from the Bulgarian phonetic data. It is, to our knowledge, the first time that methods borrowed from phylogenetics are directly applied to the phonetic transcriptions of a large number of words. The third innovation presented in this work is the testing of various hypotheses about sound changes. We present several models of vowel evolution that might be appropriate for our dialect data and apply them to (a simplified version of) the data. As a result we obtain dialect divisions that are historically motivated and compare them to the divisions described in traditional scholarship. This technique also enables us to see, given our data, which vowel changes are the most likely ones.

2 Data

The data used in this paper is part of the project *Buldialect—Measuring linguistic unity and diversity in Europe*. It consists of the pronunciations of 157 words collected at 197 places equally distributed all over Bulgaria, with the exception of the northeastern part inhabited by non-Bulgarian populations where the concentration of the sites is much smaller. During the data collection, mainly in the period 1950-1985, only villages that were dialectologically homogeneous were included. For example, villages with mixed Turkish-Bulgarian, or predominantly Turkish population were excluded. The informants were chosen among the oldest female inhabitants who were born locally. The material consists of frequently used words related to customs, religion, agricultural work, and surrounding nature. In this paper we use a subset of 152 words.¹ Phonetic transcriptions include various diacritics and suprasegmentals, making the total number

¹Words with a lot of missing entries and words where human experts could not agree how to align them phonetically were removed.

of unique phones in the data set 98: 43 vowels and 55 consonants.² In this experiment the sign for primary stress is moved to a corresponding vowel, so that there is a distinction between stressed and unstressed vowels. For some villages there were multiple pronunciations of the same word. For technical reasons we have randomly picked only one per village in this research. Detailed description of the data set can be found in Prokić et al. (2009).

3 Traditional scholarship

Since we shall derive a division of sites phylogenetically, we present here the traditional division of the sites in the data we examine. We shall compare the two below. Here we give a brief description of the traditional dialect divisions in Bulgaria, as given by Prof. Stoyko Stoykov (Stoykov 2002). In *Българска диалектология* [Bulgarian dialectology], Stoykov (2002) described the main dialect areas in Bulgaria. This division was based on the variation of different phonetic features, and neither lexical nor syntactic variation was taken into account.

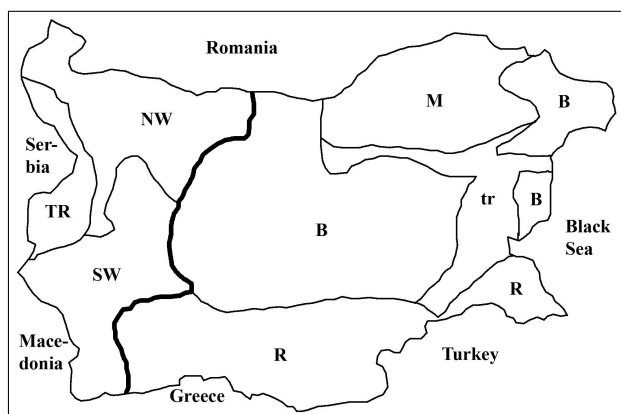


Figure 1: Traditional division of Bulgarian dialects. NW - northwest, SW - southwest, TR - transitional (between Bulgarian and Serbian), B - Balkan dialects, M - Moesian, R - Rupian, tr - transitional (between Balkan, Moesian and Rupian dialects). Taken from Houtzagers et al. (2010).

According to Stoykov, the main division of Bulgarian dialects is into western and eastern. The border between these two areas is the so-called *yat* border that reflects different pronunciations of the Old Bulgarian vowel *yat*. It goes from Nikopol in the north, near Plevna and Teteven down to Petrich in the south, represented by the thick black line in Figure 1. This is the oldest dialect border, and it is still very well preserved. Before a syllable that does not contain post-alveolar consonant, palatalized consonant or a front vowel, in the west the Old Bulgarian vowel **ǎ* (*yat*) is always pronounced as [e], while in the east it is pronounced either as [a] or a low variant of [e]. If the reflex of *yat* is [a] or a very low variant of [e], a preceding consonant is usually palatalized. For example [beɪ] vs. [bʲaɪ], [bʲæɪ] or [bɛɪ]. This isogloss divides Bulgarian language area into west and east. According to Stoykov (2002), east of the *yat* line there is a

²<http://www.bultreebank.org/BulDialects/index.html>

division into northeastern and southeastern areas based on the pronunciation of the old vowel *yat* in a palatal environment, i.e. if there is a post-alveolar consonant, palatalized consonant or a front vowel in the following syllable. In the northeast *yat* is pronounced as [e], while in the southeast it is pronounced as [a], [æ] or [ɛ]. For example [beli] vs. [b^jali], [b^jæli] or [bɛli].

Taking into account various phonetic features, the western part of the country can further be divided into northwestern and southwestern dialects, and a small zone at the border with Serbia that is a transitional zone between Bulgarian and Serbian. In the east, Stoykov also distinguishes three dialect areas, namely Balkan, Moesian and Rupian dialects, but these areas are not distinguished robustly in our data (Houtzagers et al. 2010). Balkan dialects cover the central area of present Bulgaria and represent the most extensive group of dialects of the Bulgarian language. Moesian dialects are situated in the northeastern part of Bulgaria, where the Rupian dialects occupy the southeastern part of Bulgaria. More information on Bulgarian traditional dialectology can be found in Houtzagers et al. (2010).

In this paper we analyze Bulgarian pronunciation data using a Bayesian inference of phylogeny and compare our results to the dialect division suggested by Stoykov. A brief introduction to the Bayesian inference of phylogeny is given in the next section, followed by the experimental set up and the description of the results.

4 Bayesian inference of phylogeny

Methods for phylogenetic inference with respect to biological organisms proceed from aligned DNA or protein sequences and from the aligned and compared sequences infer an evolutionary tree that represents genetic relatedness among the species. A tree consists of nodes connected by branches. There are three types of nodes: terminal nodes that represent organisms (sequences), internal nodes that represent hypothetical ancestors and a root node that is the ancestor of all organisms (Figure 2).

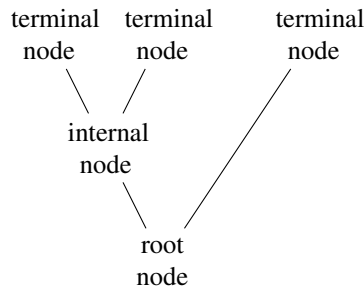


Figure 2: An example of a phylogenetic tree.

Bayesian inference of phylogeny, based on Bayes theorem, was independently proposed by several authors in 1996 (Rannala and Yang 1996; Mau 1996; Li 1996). In probability theory, Bayes theorem specifies how the conditional probability of event A given event B is related to the conditional probability of B given A :

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \quad (1)$$

where $P(A)$ is the prior probability of A , $P(B)$ is the prior probability of B , $P(A|B)$ is the conditional probability of A given B , also called posterior probability, and $P(B|A)$ is the conditional probability of B given A , also called likelihood.

Bayesian inference of phylogeny is based on finding a large number of trees with a high posterior probability. It is normally applied in a so-called character-based way, which means that the inferences about relatedness among species are based on the information found at each of the ‘genetic sites’, i.e. positions in the aligned sequences, separately. These will be phonetic segment positions in our application. This means that the posterior probability of a tree τ has to be calculated for each of the positions separately. For the i th position in the aligned sequences, the posterior probability of a tree τ can be expressed as:

$$P(\tau_i|D) = \frac{P(\tau_i)P(D|\tau_i)}{\sum_{j=1}^{B(s)} P(D|\tau_j)P(\tau_j)} \quad (2)$$

where $P(\tau_i|D)$ is the posterior probability of the i th tree, D is data, in particular the distribution of genetic categories at site i , $P(\tau_i)$ is the prior probability of the i th tree, $P(D|\tau_i)$ is the likelihood of the i th tree, and $\sum_{j=1}^{B(s)} P(D|\tau_j)P(\tau_j)$ is the probability of the data, i.e. $P(D)$.

The use of prior probability of a tree $P(\tau_i)$ is considered the strongest and at the same time the weakest point of the Bayesian inference. If we have reliable information on the prior, it can help us get better posterior estimates, and it can be a very powerful tool. But, in reality it is very hard to find realistic estimates for the priors. In the case of phylogenetic inference, usually all trees are considered equally probable and they are assigned the so-called flat priors where $P(\tau_i) = \frac{1}{|B(s)|}$, and $|B(s)|$ being the number of all possible trees for s species (see below).

To be able to calculate the likelihood of the i th tree $P(D|\tau_i)$, we need a tree with branch lengths and a model of character change. We will illustrate how the likelihood of a tree is calculated on a very simple example presented in Table 1 where we give aligned transcriptions for word *bɛ.nu /'beli/* ‘white - pl.’ for three villages:³

Table 1: A scheme of the aligned transcriptions for word ‘white’ for 3 villages.

	position1	position2	position3	position4
Lobosh:	b	'e	l	i
Mihaltsi:	b ^j	'e	l	i
Slaveino:	b	'ɛ	l	i

The number of all possible trees $B(s)$, i.e. possible ways to group species, depends on the number of species s , which will be language varieties in our application. For rooted binary branching trees it is calculated as:

$$B(s) = \frac{(2s-3)!}{2^{s-2}(s-2)!} \quad (3)$$

For example, for three species, there are three possible trees, i.e. different ways in

³For technical reasons the sign for primary stress is moved to the corresponding vowel.

which they can be grouped. In Figure 3 we present three possible trees for three villages: Lobosh, Mihaltsi and Slaveino.

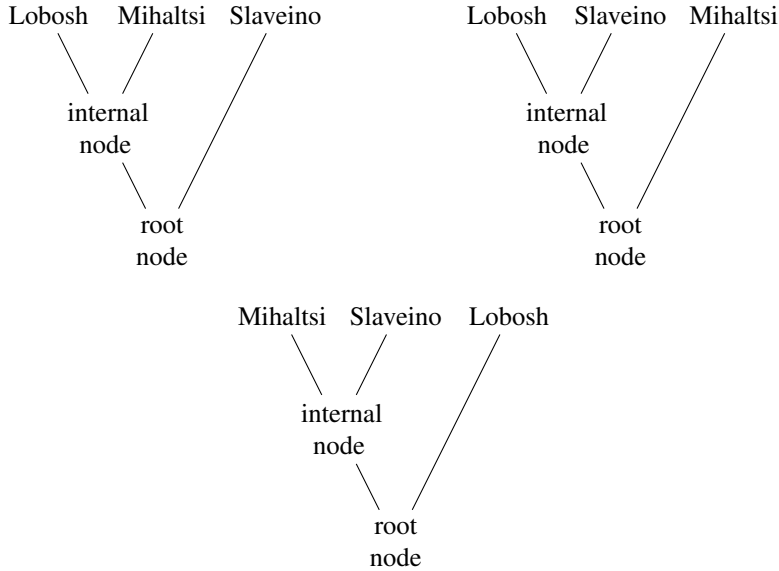


Figure 3: The 3 possible trees for 3 villages.

We will look closely at the topology presented in the upper left corner to see how to calculate its likelihood for the 2nd position in our alignment presented in Table 1. In Figure 4 we repeat that tree topology and additionally label all its nodes and branches. There are three terminal nodes 'e', 'e', 'e', one internal ('e) and a root node ('e). The branches are labeled from v_1 to v_4 . An internal node and a root node can have any state possible for the second position in our alignment. It could be any of the 43 tokens that we use for various vowels in our data set, since vowels can align only with other vowels and consonants only with the consonants. For the two nodes we get $43 \times 43 = 1849$ possible combinations for state assignments. In the tree in Figure 4 we present one of the possible assignments of the states for the internal node and a root node. We note that there is only one change of states on the tree: 'e \rightarrow 'e. We mark it with a dashed horizontal line on branch v_4 . Branch lengths in a tree represent the number of changes that have occurred in a certain branch. For example, in Figure 4 there is one change on branch v_4 , meaning that this branch has length 1.

To be able to calculate the likelihood of a tree, we need not only the tree τ_i itself with branch lengths v_i , but also a substitution model θ . The substitution model θ is a model of how probably one state changes into the other, i.e. a model that specifies the probability of one state changing into the other. In fact, our paper focuses on this aspect of the models. θ operates both to produce the leaf nodes but also to produce internal nodes. In our example, the model would need to specify the probability of one phone changing into any other phone present in the aligned sequences. In the simplest model, a character (phonetic segment) can change from any state into any other state. The probability of going from one state into the other is then equal for all pairs of states. This is realistic neither for most of the data in biology, nor for language data. We know that phones are not equally likely to change into just any other phone, but prefer some

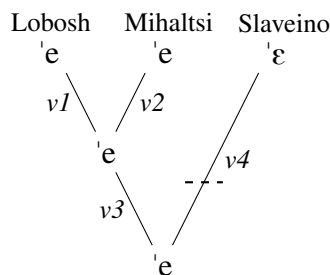


Figure 4: Labeled tree.

changes. More complex phylogenetic models allow different rates of change between the states, which we believe suits our data better. In more complex substitution models it is also possible to specify the directionality of a change. The substitutions may have different values for $'e \rightarrow '\epsilon$ and for $'\epsilon \rightarrow 'e$.

Let us note that the substitution model is a convenient point at which the degree of regularity of sound change is reflected. To the degree that the Neogrammarian doctrine of regular sound change is correct, θ ought to concentrate the probability mass of the mapping $\hat{a}\ddot{Y}e \rightarrow x$ onto a small number of sounds x (ideally one) in a given variety (and phonetic context, which we must ignore here to keep things simple).

Another parameter that we can add to the phylogenetic model is the ‘site heterogeneity rate’. It allows us to specify whether different positions in the aligned sequences evolve at the same or different rates. In our linguistic example, it is more likely that some characters evolve faster since changes are more frequent, for example, at the beginnings and ends of words than in the middle. In phylogenetic models this is modeled by having a distribution of character rates instead of a uniform rate. It is usually done by estimating a so-called gamma distribution of rate changes from the data (Yang 1994).

Once we have the tree τ_i with branch lengths v_i and a substitution model θ , the likelihood of a single tree is normally calculated under a Markov model of character evolution—the probability of every node depends only on the preceding node and the branch length between these two nodes. This assumes that all positions (sound segments) and all lineages (villages in our case) evolve independently. The likelihood of the tree presented in Figure 4 is the product of the probabilities of every node in the tree:

$$L = P('e)P('e \rightarrow 'e|v3)P('e \rightarrow 'e|v1)P('e \rightarrow 'e|v2)P('e \rightarrow '\epsilon|v4) \quad (4)$$

The probability of one state changing into the other, $'e \rightarrow 'e$ or $'e \rightarrow '\epsilon$ in our example, given a certain branch length ($v1-v4$), is defined by the substitution model θ . The likelihood of a tree τ_i for the position i is the product of all possible ancestral states combinations for that position (combinations of all possible assignments for the internal node a root node given a certain branch length v_i).

To be able to calculate the posterior probability of a tree, we need to do calculations for all possible trees and for each tree to integrate over all possible combinations of branch lengths and parameter values of the substitution model.⁴ Since the num-

⁴In genetics final grouping of leaves is regarded as defining a tree, regardless of the configure of the

ber of trees grows very fast as the number of species increases, it is computationally extremely expensive, if feasible at all. Bayesian inference of phylogeny solves that problem by using Markov Chain Monte Carlo (MCMC) modeling and sampling from a posterior probability distribution on trees. MCMC involves three steps: a) pick a tree randomly or one that is a good description of the data; b) propose a new tree by stochastically perturbing the current tree; and c) accept or reject new tree with a probability described by Metropolis-Hastings algorithm (Metropolis et al. 1953; Hastings 1970). The number of generations that the MCMC algorithm will execute is set by the user. It depends on the size of the data set and the complexity of the model. The chain length should be long enough to obtain a good approximation of the posterior probabilities of trees and the parameters. As a result of Bayesian inference we do not get a single tree, as in other character-based methods, but a sample of trees chosen according to their posterior probability. Information from sample trees can be summarized in a single tree using different methods, such as the 'maximum clade credibility tree', 'majority rule consensus tree' or simply a single tree that seems most probable. A tree where the information is summarized, contains the information on the posterior probabilities of the nodes and particular clades, i.e. branches in a phylogenetic tree.

5 Experiment

This section summarizes work from Prokić, Wieling, and Nerbonne (2009).

5.1 Multisequence alignment

In our experiment we proceed from automatically multi-aligning all phonetic transcriptions. In multiple string alignment all strings are aligned and compared at the same time, making it a good technique for discovering patterns, especially those that are weakly preserved and cannot be detected easily from sets of pairwise alignments. Multiple string comparison is not new in linguistic research. In the 19th century the comparative method became the most widely used technique for simultaneous comparison of different languages i.e. lists of cognate terms from the related languages. When applied to the phoneme level, cognate sets from various languages are manually multi-aligned in order to discover the re-occurring patterns of sound correspondences. In this study we present a technique that allows us to automatize this process, which is especially important today, when large amounts of digitalized data are becoming available. Using these techniques the data can be processed much faster and the re-occurring patterns can be detected automatically.

In order to multi-align phonetic transcriptions, we apply an iterative pairwise alignment program ALPHAMALIG (Alonso et al. 2004), to the phonetic transcriptions of words used in dialectological research. In Table 2 we give an example of the automatically multi-aligned strings from our data set and point out some important features of the simultaneous comparison of more than two strings.

In Table 2 we have multi-aligned pronunciations of the word *a3 /az/* 'I' automatically generated by ALPHAMALIG. In the alignment procedure we have used the constraint that vowels can be aligned only with the vowels and consonants only with the consonants. The advantages of this kind of alignment over pairwise alignment are twofold:

internal. The calculations must also take account of this.

Table 2: Example of multiple string alignment for six villages. The sign for primary stress is moved to the first vowel in the stressed syllable.

Aldomirovtsi:	j	'a	-	-	-	-
Beglezh:	-	'a	s	-	-	-
Belene:	-	'a	s	-	-	-
Chukovets:	j	'a	z	e	k	a
Dinevo:	j	'a	-	-	-	-
Dobroselets:	-	'a	s	-	-	-

- First, it is easier to detect and process corresponding phones in words and their alternations (like [s] and [z] in the third position in the alignment in Table 2).
- Second, the distances/similarities between strings can be different in pairwise comparison as opposed to multiple comparison. This is so because multi-aligned strings, unlike pairwise aligned strings, contain information on the positions where phones were inserted or deleted in both strings. For example, in the pairwise alignment the distance between the pronunciations for villages Aldomirovtsi and Beglezh would be $1/3 = 0.33$, while the similarity between these two strings based on the multi-aligned strings in Table 2 would be $4/6 = 0.66$.
- Third, this format conforms to the input format of the software used for Bayesian phylogenetic inference and some other techniques taken from biology.

Results of the evaluation of the automatically acquired multiple string alignments, produced by ALPHAMALIG, against the manually aligned strings were described in Prokić, Wieling, and Nerbonne (2009). The comparison was done on the Buldialect data set, also used in this research. Two novel evaluation techniques were applied in order to check the performance of the ALPHAMALIG algorithm. Using the technique that checks the quality of each column, regardless of whether the columns are in order, has shown that content-wise, alignments produced by ALPHAMALIG are 98.20 per cent correct. Since in Bayesian inference every column is treated separately, this measure is a good indicator of the quality of the alignments produced automatically. The results have shown that the alignments produced by ALPHAMALIG are of good quality and can be used directly for further processing by the software for Bayesian phylogenetic inference. Error analysis has shown that the main source of errors produced by ALPHAMALIG comes from the constraint that vowels and consonants cannot be aligned. Although true in most of the cases, there are still exceptions where vowels and consonants should be aligned.

5.2 Data preprocessing

After multi-aligning transcriptions for every word separately, we concatenate all aligned transcriptions into a single set of multi-aligned strings, where each string contains transcriptions of all 152 pronunciations collected at a certain village. Bayesian MCMC inference infers the relationships between language varieties by processing multiple alignments position by position. This allows us to merge the transcriptions of all words into a single set of multi-aligned strings, since our calculations do not take into account any information related to the word level (e.g. lexical identity, lexical semantics,

specific contexts in which certain phone occurs). We do not use any information concerning where one word begins or ends. Merging all multi-aligned transcriptions in our data set resulted in 620 columns that contain either consonants or vowels. For the missing words in our data set we use symbol ‘?’ to mark each of the positions where the corresponding phones would have been placed if the pronunciation for that village had been available. For the phones that were deleted in a certain pronunciation, we use symbol ‘-’ in order to keep these two types of missing tokens separate and prevent an overestimation of the correspondences between the sites where both contain missing values at certain positions in the alignments. In Table 3 we present multi-aligned pronunciations for words *вечер* /vetʃer/ ‘evening’, *дно* /dɫno/ ‘bottom’ and *лесно* /lesno/ ‘easily’ merged into a single set of multi-aligned strings.

Table 3: Pronunciation of different words merged into a single string.

Aldomirovtsi:	v	'e	tʃ	e	r	d	-	n	'o	l	'ɤ	s	n	o
Asparuhovo-Lom:	v	'e	tʃ	e	r	d	-	n	'o	l	'e	s	n	o
Asparuhovo-Prov:	vʲ	'e	tʃ	ə	r	d	'ɤ	n	u	lʲ	'e	s	n	u
Babyak:	v	'e	tʃ	e	r	d	-	n	'o	?	?	?	?	?
Bachkovo:	v	'e	ts	e	r	d	'a	n	u	lʲ	'e	s	n	u

It is evident that these multi-aligned sequences are very different from the sequences used in biology. Our linguistic alignment contains a large number of sites, 197, and relatively short strings comprising only 620 positions in total. At the same time alignments in biology would normally contain longer sequences for a much smaller number of species. On the one hand, the data had to be simplified so that we would be able to use software developed for biological data. On the other hand, the coding was linguistically informed so that we could preserve enough relevant information to allow us to address specific issues related to language change. We try to find a good compromise between the two. The other difference is in the number of unique tokens/character states: in protein sequences there are 20 different proteins, while in our linguistic alignments the number of unique phonetic segments was 97: 55 for consonants and 43 for vowels. It is computationally infeasible to explore the space of all phylogenetic trees modeling 620 sound segment positions varying nearly 100 ways, corresponding to 100 different phonemes, at each position. So we needed to simplify. We also wished to simplify in order to examine the linguistic plausibility of the sound changes more effectively. So we restrict our attention to vowels which we regard as belonging to one of the eight classes presented in Figure 5.

Representation of the pronunciation dialect data with only 8 symbols leads to information loss. We have completely discarded consonant changes, and, additionally, we have merged all 43 vowels into only 8 groups. However, we believe that this type of data representation still contains enough information for the exploration of some aspects of dialect variation and change. Consonant changes in our data set are less frequent and less various if compared to vowel changes. For that reason we choose to focus on vowels. We group all vowels in our data set based on their articulatory features, so that each of them can be defined based on the front/back and close/open opposition. For example, we can describe group 6 as a group comprising of close front vowels. By grouping vowels in such a way, we hope to be able to discover some of the general principles of substantial vowel changes within the vowel chart.

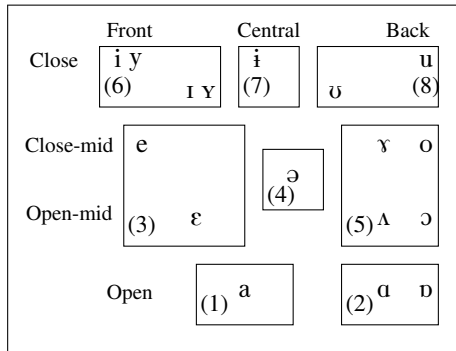


Figure 5: All vowels in the data set were placed into one of the 8 groups.

After putting our data into the format described, our next step was to choose suitable models of sound changes. We tested three models of evolution on our data set and they will be explained in more detail in the next subsection.

5.3 Models of evolution

All calculations related to the Bayesian MCMC inference were done using the BEAST software (Drummond and Rambaut 2007). All models implemented in this software that we have used in our experiment for Bayesian inference were originally developed to analyze molecular sequences. Among various possibilities, we have chosen to test three settings that can be applied to our phonetic data. In each of the settings we specify the following categories: a) a substitution model (s) and b) a position heterogeneity model (h).

Substitution models for biological data describe the process of one nucleotide or amino acid being replaced by another. In our case, they describe the process of one vowel, or more precisely one of our 8 groups, being substituted for another. In Table 4, we present the final, reduced, format of the alignment presented in Table 3 and additionally mark it with ‘s’ and ‘h’ to show which model applies to which part of the alignments. In our example the substitution model, marked with ‘s’ (in Table 4), calculates the probability of group 5 being substituted for group 8, or the other way around. In this model we were not able to specify the directionality of the change. As a result we get only one probability of change for each pair of phones.

Table 4: Different models of evolution apply to the parts of the alignments marked with ‘s’ (substitution model), and ‘h’ (rate heterogeneity model).

	h(1)	h(2)	h(3)	h(4)	h(5)	h(6)
Aldomirovtsi:	3 [e]	3 [e]	-	5 [o]	5 [ʏ]	5 [o]
Asparuhovo-Lom:	3 [e]	3 [e]	-	5 [o]	3 [e]	5 [o]
Asparuhovo-Prov:	3 [e]	4 [e]	5 [ʏ]	8 [u]	3 [e]	8 [u]
Babyak:	3 [e]	3 [e]	-	5 [o]	?	?
					↕s	
Bachkovo:	3 [e]	3 [e]	2 [ʌ]	8 [u]	3 [e]	8 [u]

The site (position) heterogeneity model allows us to specify whether the rate of variation in different positions, marked with 'h(1)', 'h(2)', ..., 'h(6)' in our example in Table 4, is the same or whether it varies from column to column. For our data it would mean that we can specify whether vowel changes occur more frequently in some positions in words than in others. We do not specify where the substitutions are more or less frequent, but some settings allow different columns to vary at different rates.

For all three settings we set the molecular clock option to the strict molecular clock. This setting specifies that different branches in a tree have the same rate of variation, i.e. that different species change constantly over time. This is the basic, and the simplest molecular clock model implemented in BEAST. In the future, we would certainly like to test the relaxed molecular clock options that assume independent rates on different branches.

Our **Setting 1** is the simplest one, with the following values for the two parameters:

- Substitution model: any state, i.e. phone, is equally likely to change into any other state. For example, vowel [a] (group 1) can change into a vowel in any other group and the probability of, for example, [a] changing into [ə] is the same as [a] changing into [u].
- Site (position) heterogeneity model was set to 'None', meaning that all sounds in all positions in words evolve at the same rate.

In **Setting 2** we have the following options:

- Substitution model: General Time Reversible (GTR) model. Under a GTR model any state, i.e. phone, can change into any other, but the probability of change differs depending on the phones involved. The rate of change is not set in advance, but calculated from the data. In this setting the probability of, e.g., [a] changing into [ə] is not the same as the probability of [a] changing into [u]. This allows us to calculate which phone changes are more likely than some others.
- Site (position) heterogeneity model was set to 'None'. The same as in the Setting 1, i.e. all sounds in all positions in words are assumed to evolve at the same rate.

Setting 3 comprises the following options:

- Substitution model: General Time Reversible (GTR) model. The same as in the Setting 2, any phone can change into any other. The probability of one phone changing into the other may vary depending on the phones involved. The directionality of the change is not specified.
- Site heterogeneity model was set to allow various substitution rates between different positions i.e. to allow the phones in different positions within the words to evolve differently. In contrast to the previous two settings, we assume that, for example, position h(1) might evolve more slowly or faster than position h(6).

The length of the Markov chain, i.e. the number of generations that the MCMC algorithm ran for, was 4×10^8 for all three settings. The trees were sampled after every 8000 generations, which gave us a final sample of 5000 trees. This number of generations was sufficient in all three runs to get a representative sample of trees.

Some assumptions made by the various models might seem more or less plausible depending on one’s linguistic tenets, for example that some sounds are more likely to change than others, or that sounds in certain positions are particularly subject to change, that some directions of change are more likely than others. By using rigorous quantitative methods, we want to test the validity of different hypotheses and try to answer some questions about language evolution and change in a more exact manner. In the next section we present the results for each of the settings tested.

6 Results

We use the TreeAnnotator program from the BEAST package to summarize the information from the sample trees produced by BEAST into a single tree. We select the option ‘maximum clade probability tree’ in order to get a tree where the node height⁵ and rate statistics are summarized on the tree in the posterior sample that has the maximum sum of posterior probabilities on its $n - 2$ internal nodes.

We first inspect the grouping of sites implicit in the various models of phylogenetic development and then turn to an analysis of the vowel substitution model. We do not present dendrograms which contain too many irrelevant details, but we do show the most important dialect divisions retrieved from dendrograms. In Figure 6 we present dialect divisions resulting from Setting 1, the two-way division on the left hand side and the three-way division on the right hand side. On the map that shows the two-way split of the data, sites are marked with red dots (eastern varieties) and blue dots (western and southern varieties). This two-way division of sites has a maximum posterior probability of 1 in the Monte Carlo estimation. It is geographically coherent and divides the Bulgarian language area in a such way that eastern varieties, in traditional literature referred to as Balkan and Moesian dialects, and on our map marked with red symbol are put in one group, while western and Rupian dialects, marked with blue, are put in an other group. On the right hand side map in Figure 6 we present a three-way division of sites. With a posterior probability of 0.531 the southern group of varieties is put into a separate group, while the classification of the western varieties into a single group is supported with maximum posterior probability. Although according to the posterior probability it is not highly certain that southern sites form a distinct group, they largely occupy a geographically coherent area in the south of the country. Since in Setting 1 the probability of any phone changing into any other phone was set to be equal we could not get any interesting information on vowel changes from this setting.

In Figure 7 we present the two-way and three-way divisions of the sites resulting from the Bayesian inference performed once we adopted the General Time Reversible (GTR) model. The two-way split has maximum posterior probability, and shows a split of the sites into western and eastern. The southern group of varieties is classified with the eastern dialects. This division also corresponds well with the traditional division of the sites (see Figure 1) since the split follows approximately the *yat* line and groups all the sites into eastern and western. Unlike in the Setting 1, varieties in the south are grouped with the eastern dialects. However, the support for this grouping is relatively low (0.505) and cannot be taken with any great confidence. Groupings of both southern and eastern varieties into separate varieties have low posterior probabilities, namely 0.134 and 0.526 respectively. The former has little basis in the model. Unlike the eastern division of the sites, the western varieties are grouped under the node with the

⁵The height of a node is the length of the longest downward path from that node to a leaf.

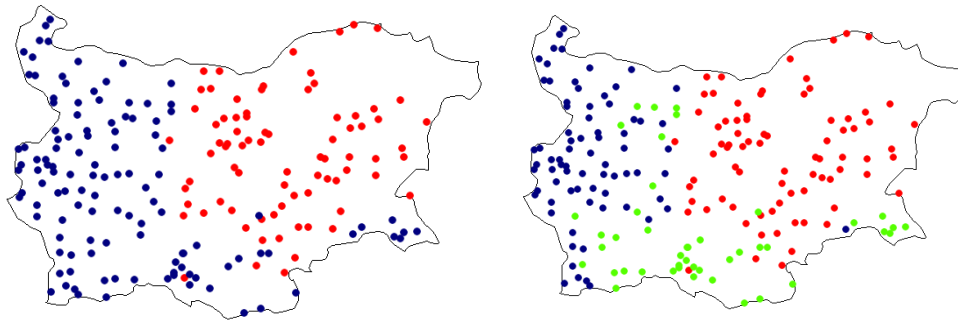


Figure 6: Distribution of the two (left) and three (right) group of sites using a free substitution model and no positional heterogeneity model (Setting 1).

high posterior probability and can be taken with great confidence to form a coherent group. We conclude from this examination of the divisions of sites (varieties) that both of the models inspected agree with the scholarly consensus on Bulgarian dialects. There is no reason to prefer one over the other.

Apart from reconstructing phylogenies, i.e. grouping of the varieties, Setting 2 also allows us to investigate how probable certain sound changes are. In Setting 2, we used a General Time Reversible Model to model sound changes. As a reminder, we recall that any group of sounds was allowed to change into any other group but the changes did not receive equal probability as in the Setting 1. One of the outputs of the Bayesian inference analysis were the probabilities of change between each pair of sounds calculated from the data. The results can be seen in Figure 8, where on the left hand side we present sound changes with the highest probabilities (connected with blue lines) and on the right hand side where we show changes that have the lowest probability (connected with red lines). For clarity, we put both numbers and sounds in the charts. Since all our sounds in the data are put into one of the eight groups, we can naturally talk only about the probability of a change of a vowel in one group into a vowel in another. In Figures 8 and 10 groups 1 and 7, which stand for [a] and [i] sounds are put in dashed squares since we could not get any reliable estimations for them. The reason for this is their very low frequency in the data set. The sound [a] appears only 147 times in our multiple alignment, while the sound [i] is present only 40 times. Vowels from the third group [ɛ, e], which is the most frequent group in the data set, appear 14663 times. As marked with the blue lines in the left-hand side vowel chart in Figure 8, changes that received the highest probability are between the following groups: 5 [ʌ, ɔ, ɤ, o] and 8 [ʊ, u], 3 [ɛ, e] and 6 [ɪ, y, i], 4 [ə] and 6 [ɪ, y, i], and 2 [ɑ, ɒ] and 4 [ə]. We can see in the chart that those changes involve moving only one step within the vowel chart. Unfortunately it was not possible to infer the directions of the changes and see whether, for example, it is more probable that vowels from group 3 would change into vowels from group 6 (3 → 6) or the other way around (6 → 3). However, our findings correspond well with the findings reported in the literature on the traditional analyses of the vowel reduction in Bulgarian (Wood and Pettersson 1988; Barnes 2006). According to them the most common vowel change in Bulgarian dialects is rise of unstressed midvowels [e] and [o] to neutralize with the high vowels [i] and [u]. The low unstressed vowel [a] rises to neutralize with [ə].

In the right-hand chart in Figure 8 we mark the changes between the groups with

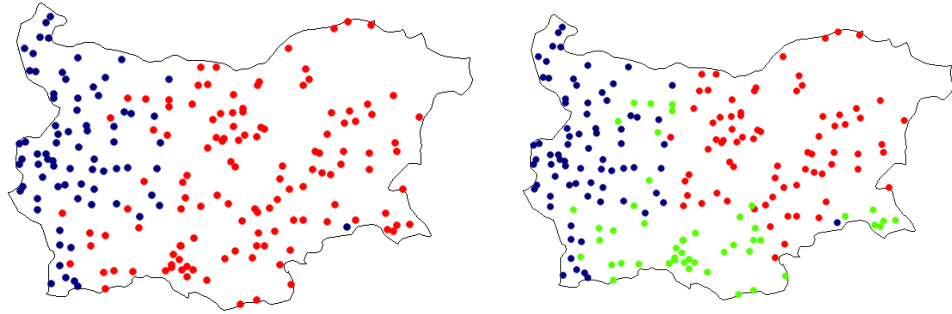


Figure 7: Distribution of the two (left) and three (right) group of sites using a General Time Reversible substitution model and no positional heterogeneity (Setting 2).

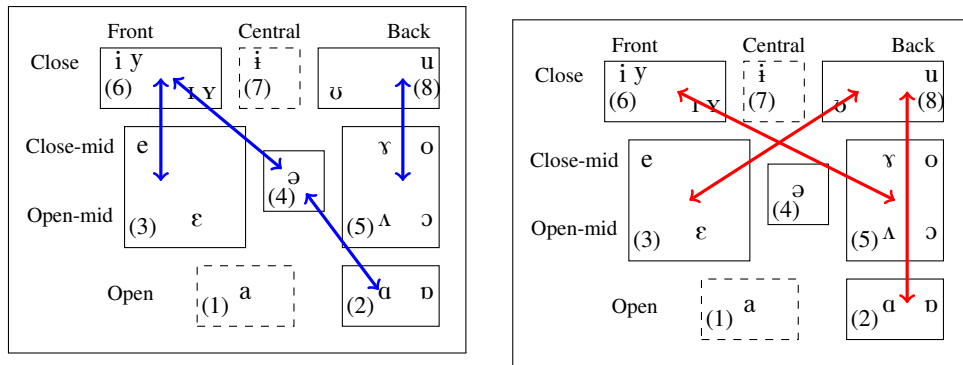


Figure 8: The most probable (left) and the least probable (right) vowel transitions under the GTR model and no positional heterogeneity. Boxes (1) and (7) are shown in dashed line to indicate that our estimations concerning them are based on too little data.

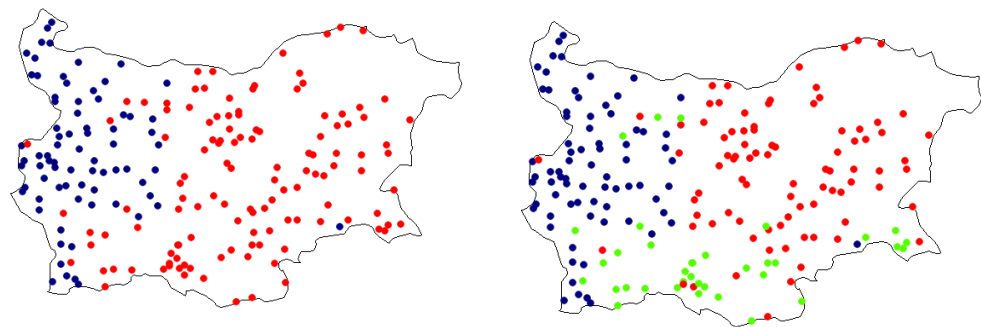


Figure 9: Distribution of the two (left) and three (right) group of sites using a GTR model with gamma positional heterogeneity (Setting 3).

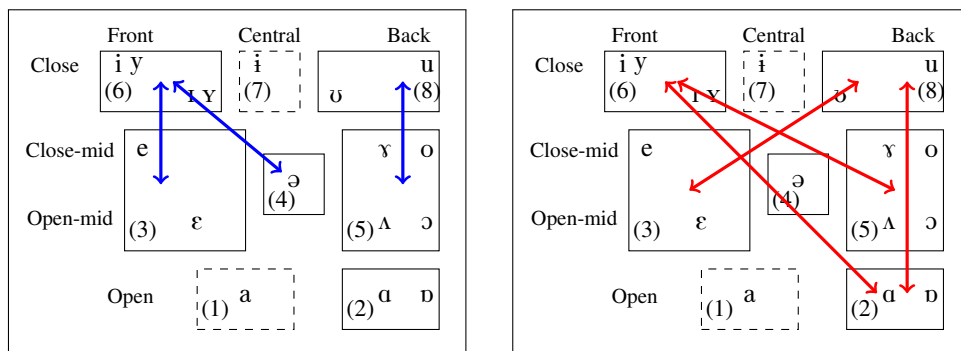


Figure 10: The most probable (left) and the least probable (right) vowel transitions using GTR and gamma site heterogeneity model (Setting 3).

the lowest probabilities using red lines: 2 [ɑ, ʊ] and 8 [ʊ, u], 3 [ɛ, e] and 8 [ʊ, u], and 5 [ʌ, ɔ, ɤ, o] and 6 [ɪ, ʏ, i]. In contrast to the alternations with the highest probabilities, they do not involve changes between the adjacent groups but rather between the groups separated by at least one group within the vowel chart.

In the Setting 3 under the General Time Reversible model, just as in the Setting 2, every phone was allowed to change into any other phone with the transition probabilities being inferred from the data. It was again not possible to calculate the directionality of the changes. The difference between the two settings is that in the Setting 3 the positions in the alignments were allowed to vary at different rates. We extracted the two-way division of the sites and represented it on the left-hand side map in Figure 9. On the right-hand side map in Figure 9 we mark the three groups extracted. Both the two-way and the three-way divisions of the sites are almost identical to the divisions for Setting 2: the first one goes along the *yat* line, while the second additionally distinguishes the southern area as separate. Division into western and eastern dialects gets the highest posterior probability, while other major splits were supported with much smaller posterior probabilities.

In Figure 10 we present vowel charts with the changes that are the most and the least probable. The sound changes with the highest probabilities are those between the groups 5 [ʌ, ɔ, ɤ, o] and 8 [ʊ, u], 3 [ɛ, e] and 6 [ɪ, ʏ, i], and 4 [ə] and 6 [ɪ, ʏ, i]. Just as in the previous analysis, sound correspondences that involve two adjacent groups within the vowel chart are the most probable. The least probable sound correspondences include alternations between the sounds that are more than one step apart within the vowel chart.

Settings 2 and 3 gave very similar results, both with the respect to the classification of villages and to the vowel transition probabilities. Although the results were similar, the two settings contain two different hypotheses about sound changes. In Setting 2 we assume that in all positions in words sounds change at the same rate. In Setting 3 we allowed that sound changes are more likely at some positions in words than at others. In order to check which of the two hypotheses is more probable, we calculated Bayes factor (K) for the two settings, which is a Bayesian alternative to a classical hypothesis testing in statistics. The Bayes factor was calculated using the following formula which examines the probability ratio between two hypotheses $H1$ and $H2$ given the data D :

$$K = \frac{P(D|H1)}{P(D|H2)}$$

where $P(D|H)$ expresses the marginal likelihood of a hypothesis H . For a more detailed explanation see Kass and Raftery (1995) or MacKay (2003). For our two settings we calculated the Bayes factor using the Tracer software.⁶ In Table 5 we present the values of the Bayes Factor in log 10 scale obtained after pairwise comparing all three settings.

Table 5: Values of the Bayes factor in log 10 scale. There is a strong support for Setting3 when compared to both Setting 1 and Setting 2.

	Setting 1	Setting 2	Setting 3
Setting 1	-	-573.369	-938.271
Setting 2	573.369	-	-364.902
Setting 3	938.271	364.90	-

All values of $K > 2$ for the log 10 scale indicate strong support for a favored model. All values for comparing our three settings are much bigger. It shows that there is very strong evidence in favor of Setting 3. Setting 2 is much more strongly supported than the Setting 1, while the Setting 3 is much more strongly supported than the Setting 1 and 2. Explanation of the scale for K can be found in Jeffreys (1961) and Kass and Raftery (1995). These results show that there is a strong evidence in our data that different vowel changes are not equally probable. Some changes are much more likely to occur than others. The data also strongly supports the hypothesis that vowel changes occur at different rates in various positions in words.

7 Discussion

In this research we have tried to automate the process of character selection in phylogenetic linguistic inference by automatically multi-aligning phonetic transcriptions. We have restricted ourselves to the investigation of the vowel changes, since vowels are more likely than consonants to contain sufficient information on dialect change. By classifying all the vowels into eight groups we have tried to attend to the main articulatory characteristics (open/close and front/back opposition) of the vowels in our analysis. This multi-state character encoding enabled us to test the probability of sound changes within the vowel chart. The coding of the characters can naturally be done differently, but we leave this to future research. We hope that in future it will become computationally feasible to process the data using a larger set of states.

The application of phylogenetic inference allows us to test various models of evolution and to investigate how related certain species are. By applying this method to the phonetic data, we were able to test various hypothesis about the mechanisms of sound change. Each model of evolution contains its own explicit assumptions. Relying on the models of evolution created for biological data, we were forced to draw parallels between the evolution of species and the evolution of languages. But very often models developed for the evolution of species contain assumptions that are not very realistic for the language data. For example, all character-based methods, including the Bayesian inference of phylogeny, assume that each position in the alignments evolves independently. For our phonetic data, it would mean that the phones are not influenced

⁶<http://tree.bio.ed.ac.uk/software/tracer/>

by the changes of the preceding or following sounds. Although this is not true for the mechanism of a sound change, it is one of the simplifications that we had to introduce in our analyses.

One of the models that is being heavily debated in linguistics is the lexical clock (see for example Eska and Ringe (2004)). All the trees produced in our experiments have shown an expected topology (structure), which suggests that the assumption of a constant molecular clock is not too extreme a simplification in the models examined here. These were, however, initial experiments and in the future, we would like to apply other molecular clock models, and statistically test whether other molecular clock hypotheses fit our data better.

By initially choosing simple models of evolution to be tested on our language data, we have tried to justify more complicated assumptions step by step. None of the models developed for the biological data can cover all aspects of language evolution and change. The possibility to test various hypotheses separately makes Bayesian inference a potentially very useful technique in exploration of languages. But its true potential in linguistics can be achieved only if models are developed specifically for language data.

The results of applying Bayesian phylogenetic inference to Bulgarian dialect data have shown that three dialect areas appear as the most prominent under various models of evolution: western, eastern, and southern. This three-way division also conforms to the traditional scholarship on Bulgarian dialectology (Stoykov 2002).

We have also shown that for the Bulgarian language the most probable vowel changes are those that involve neighboring vowels within the vowel chart. Most of the highly probable changes involve vowel height. The probability of vowels changing into vowels that are far apart in the vowel chart is very low. We were not able to include the directionality of vowel changes into our analysis, to see if, for example, [e] is more likely to change into [i] or the other way around. We hope to achieve this in future. The testing of different models of evolution has also shown that vowels change faster in some positions within the words. In future we would like to investigate changes of various positions in multi-aligned sequences in more detail and try to discover patterns of variation and how regular certain sound changes are.

References

- Alonso, L., I. Castellon, J. Escribano, X. Messeguer, and L. Padro (2004). Multiple Sequence Alignment for characterizing the linear structure of revision. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*.
- Barnes, J. (2006). *Strength and Weakness at the Interface: Positional Neutralization in Phonetics and Phonology*. Walter de Gruyter GmbH, Berlin.
- Drummond, A. J. and A. Rambaut (2007). Beast: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology* 7(214).
- Dunn, M., A. Terrill, G. Reesnik, R. A. Foley, and S. C. Levinson (2005). Structural phylogenetics and the reconstruction of ancient language history. *Science* (309).
- Eska, J. F. and D. Ringe (2004). Recent work in computational linguistic phylogeny. *Language* (80).
- Gray, R., A. Drummond, and S. J. Greenhill (2009). Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* (323).

- Gray, R. D. and Q. D. Atkinson (2003). Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426, 435–438.
- Gray, R. D. and F. M. Jordan (2000). Language trees support the express-train sequence of Austronesian expansion. *Nature* 405, 1052–1055.
- Greenhill, S. J. and R. D. Gray (2005). Testing population dispersal hypothesis: Pacific settlement, phylogenetic trees and Austronesian languages. In R. Mace, C. Holden, and S. Shennan (Eds.), *The evolution of cultural diversity: phylogenetic approaches*, pp. 31–52. UCL Press.
- Greenhill, S. J. and R. D. Gray (2009). Austronesian language phylogenies: myths and misconceptions about Bayesian computational methods. *Austronesian historical linguistics and culture history: a festschrift for Robert Blust*.
- Hamed, M. B. (2005). Neighbour-nets portray the Chinese dialect continuum and the linguistic legacy of China’s demic history. *Proceedings of the Royal Society B* 272(1567), 1015–1022.
- Hamed, M. B. and F. Wang (2006). Stuck in the forest: Trees, networks and Chinese dialects. *Diachronica* 23, 29–60(32).
- Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57(1), 97–109.
- Holden, C. J. (2002). Bantu language trees reflect the spread of farming across sub-Saharan Africa: a maximum parsimony approach. Volume 269, pp. 793–799.
- Holden, C. J. and R. D. Gray (2006). Rapid radiation, borrowing and dialect continua in the Bantu languages. In P. Foster and C. Renfrew (Eds.), *Phylogenetic methods and the prehistory of languages*, pp. 19–31. McDonald Institute for Archeological Research.
- Houtzagers, P., J. Nerbonne, and J. Prokić (2010). Quantitative and traditional classifications of Bulgarian dialects compared. *Scando Slavica* 56(2), 29–54.
- Jeffreys, H. (1961). *The Theory of Probability* (3rd ed.). Oxford University Press.
- Kass, R. E. and A. E. Raftery (1995). Bayes factors. *Journal of the American Statistical Association* 90, 773–795.
- Li, S. (1996). *Phylogenetic tree construction using Markov chain Monte Carlo*. Phd thesis, Ohio State University, Columbus.
- MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. CUP.
- Mau, B. (1996). *Bayesian phylogenetic inference via Markov chain Monte Carlo methods*. Phd thesis, University of Wisconsin, Madison.
- McMahon, A., P. Heggarty, R. McMahon, and W. Maguire (2007). The sound patterns of Englishes: representing phonetic similarity. *English Language and Linguistics* 11.1, 113–142.
- Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, and A. H. Teller (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics* 21(6), 1087–1092.
- Nakhleh, L., D. Ringe, and T. Warnow (2005). Perfect phylogenetic networks: a new methodology for reconstructing the evolutionary history of natural languages. *Language* 81, 382–420.

- Prokić, J., J. Nerbonne, V. Zhobov, P. Osenova, K. Simov, T. Zastrow, and E. Hinrichs (2009). The computational analysis of Bulgarian dialect pronunciation. *Serdica Journal of Computing* 3, 269–298.
- Prokić, J., M. Wieling, and J. Nerbonne (2009). Multiple string alignments in linguistics. In L. Borin and P. Landvai (Eds.), *Proceedings of Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (LaTeCH - SHELTER 2009) EACL Workshop*.
- Rannala, B. and Z. Yang (1996). Probability distribution of molecular evolutionary trees. *Journal of Molecular Evolution* (43), 304–311.
- Ringe, D., T. Warnow, and A. Taylor (2002). Indo-European and computational cladistics. *Transactions of the Philological Society*.
- Stoykov, S. (2002). *Българска диалектология*. [Bulgarian Dialectology]. Sofia, 4th ed.
- Warnow, T. (1997). Mathematical approaches to comparative linguistics. In *Proceedings of the National Academy of Science of the USA*, Volume 94, pp. 6585–6590.
- Wood, S. A. J. and T. Petterson (1988). Vowel reduction in Bulgarian: the phonetic data and model experiments. *Folia Linguistica* 22(3-4), 239–262.
- Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution* 39, 306–314.