Chapter 7

Bayesian phylogenetic inference

In this chapter we use automatically multi-aligned phonetic transcriptions to infer the historic relationships between the language varieties, but also to explore the relationship between the various phones in the data set. Multi-aligned transcriptions are analyzed using Bayesian Monte Carlo Markov Chain inference (MCMC), in recent years one of the most popular and the most powerful methods in molecular phylogeny for inferring the relationships between species. Bayesian MCMC inference belongs to the so-called character based methods, together with some other popular methods like maximum parsimony and maximum likelihood methods for phylogenetic inference. In the next section we briefly introduce molecular phylogenetics based on Page and Holmes (2006), followed by an introduction to character-based methods in Section 7.2. Section 7.3 gives an overview of the application of the methods taken from phylogenetics in linguistics. We then give introduction to Bayesian phylogenetic inference in Section 7.4. In Section 7.5 we present our experiment, followed by the results that we report in Section 7.6. We conclude the chapter with the discussion presented in Section 7.7.

7.1 Phylogenetic inference

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. DNA sequences are composed of four nucleotides (A, C, T, and G), while protein sequences comprise 20 different amino acids. Closely related organisms have similar structure of DNA (protein) sequences, i.e. similar order 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 mo-
lecular phylogenetics and try to use some of the models developed for the evolution of DNA and protein sequences on language data.

One of the most important events in the development of molecular evolution was the discovery of the molecular structure of DNA in 1953 by James Watson and Francis Crick (Page and Holmes, 2006, 4). The first comparison of amino acid sequences came in 1955, when Fred Sanger and his colleagues used it to compare protein insulin from cattle, pigs and sheep.

In order to recover evolutionary information from DNA and protein sequences, it is necessary to formalize the process of sequence change over time. In 1960s different models of molecular evolution started being developed. The comparison of sequences proceeds from their alignment— either pair-wise or multiple sequence alignment depending on the approach. The distances between the aligned sequences can be, in the simplest case, expressed as the number of the segments in which two sequences differ. Since there are usually multiple changes at each position within a sequence, distances inferred in this way are actually underestimating the amount of evolutionary change. To correct for this, different evolutionary models based on the frequency of the nucleotides and the probability of a nucleotide substitution have been developed. The most simple model, the so-called Jukes-Cantor model, assumes that the four nucleotides have equal frequencies and that all substitutions are equally likely. The most general model, general reversible model, allows each possible nucleotide substitution to have its own probability.

Information from the aligned and compared sequences is turned into an evolutionary tree that is used to represent 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 7.1).

![Figure 7.1: An example of a phylogenetic tree.](image-url)

There are numerous methods that are used to convert aligned sequences into trees. Based
7.2 CHARACTER-BASED METHODS

Character-based methods and distance-based methods. In distance-based methods, the distance between each two strings is represented as a single number and stored into a matrix. This matrix is used by various tree building methods to construct an evolutionary tree. This kind of approach is used in Chapter 3 of this thesis to infer the relatedness among the dialect varieties in Bulgaria. Character-based methods analyze each position in the aligned sequences separately. In the next section we describe character-based methods in more detail.

7.2 Character-based methods

Character-based methods (CBM) comprise various methods used in phylogenetics to study the evolutionary relatedness among species. In CBM each species is described in the terms of the states of certain characters. The term ‘character’ is used to refer to a different position in a DNA or a protein sequence. For every species, a character is in one of the states inherent for that character. Some states only vary between present/absent states, while the so-called multi-state characters can have multiple states. CBM proceed from the simultaneously aligned sequences of various species and perform the analysis based on each of the characters separately. A scheme of the aligned sequences for 3 species and 5 characters would look like this:

```
<table>
<thead>
<tr>
<th>character1</th>
<th>character2</th>
<th>character3</th>
<th>character4</th>
<th>character5</th>
</tr>
</thead>
<tbody>
<tr>
<td>species1</td>
<td>state1</td>
<td>state1</td>
<td>state1</td>
<td>state1</td>
</tr>
<tr>
<td>species2</td>
<td>state2</td>
<td>state1</td>
<td>state1</td>
<td>state2</td>
</tr>
<tr>
<td>species3</td>
<td>state2</td>
<td>state2</td>
<td>state1</td>
<td>state3</td>
</tr>
</tbody>
</table>
```

Figure 7.2: A scheme of the aligned sequences for 3 species and 5 characters.

Unlike the distance methods described in Chapter 3 that aggregate all the differences between each two strings into a single distance, character-based methods infer relatedness between the species separately on each character and later combine those analyses into a single tree. In that way the information provided by each character is retained and information loss that results from converting the sequence data into distance scores is avoided (Penny, 1982). Two well known CBM are maximum parsimony and maximum likelihood.

Parsimony methods were among the first methods to be used to infer phylogenies. They are based on the idea of the ‘minimum evolution’. While reconstructing the phylogenetic tree, the algorithm seeks the tree with the smallest number of events, i.e. for the smallest tree that would explain the data. This method has often been criticized because it does not rely on any model of evolution, but seeks for the most simple explanation of the data. The other problem related to this approach is the so-called long branch attrac-
phenomenon where species that evolve rapidly are grouped together in the phylogenetic analysis regardless of their true genetic relationship (Page and Holmes, 2006).

Unlike parsimony, probabilistic methods for phylogenetic inference, like maximum likelihood and Bayesian inference, are based on a specific model of evolution. The maximum likelihood method, as the name suggests, is based on the concept of likelihood, the probability of observing the data given a particular model or hypothesis. Given some data \( D \) and a hypothesis \( H \) the likelihood \( L \) can be expressed as:

\[
L = P(D|H)
\]

In phylogenetics, \( D \) is a set of aligned sequences, and \( H \) is a phylogenetic tree. The tree that makes our data the most probable is the maximum likelihood tree. Detailed explanation on the parsimony and likelihood methods in phylogenetics can be found in Felsenstein (2004). In our experiments we have used Bayesian inference, in recent years one of the most popular character-based methods for inferring phylogenies. In Section 7.4 we present it in more detail. Before that, we will look into the usage of phylogenetic methods in linguistics.

### 7.3 Phylogenetic inference in linguistics

In the last decade there has been an increasing interest in the application of the methods taken from phylogenetics to the 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 the change of species and languages are the same: they split into new species/languages, mutate, borrow material from neighboring species/languages, and innovations in both languages and species appear independently in unrelated elements. They both document evolutionary history, species in molecules and various morphological characteristics, languages in phonetics/phonology, morphology, syntax. Evolution and relatedness of both the species and languages can be described using family trees.

Methods taken from computational phylogenetics have been applied to lexical (Gray and Jordan, 2000; Gray and Atkinson, 2003) and phonetic data (Warnow, 1997; Nakhleh, Ringe, and Warnow, 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, Warnow, and Taylor, 2002; Nakhleh, Ringe, and Warnow, 2005), as well as to test various hypotheses about human prehistory (Dunn et al., 2005; Greenhill and Gray, 2005; Gray, Drummond, and Greenhill, 2009). As pointed out in Greenhill and Gray (2009), computational phylogenetic methods are seen as 'a powerful supplement to the comparative method used
7.4 Bayesian inference of phylogeny

In probability theory, Bayes theorem dates back to the 18th century. It gives a mathematical representation of how a conditional probability of event $A$ given event $B$ is related to the conditional probability of $B$ given $A$:

\[
P(A|B) = \frac{P(B|A)P(A)}{P(B)}
\]
\[ P(A|B) = \frac{P(B|A)P(A)}{P(B)} \] (7.2)

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
- \( P(B|A) \) is the conditional probability of \( B \) given \( A \), also called likelihood

Bayesian inference of phylogeny, based on Bayes theorem, was independently proposed by several authors in 1996 (Rannala and Yang, 1996; Mau, 1996; Li, 1996). Just as the maximum likelihood method, it is based on the likelihood function, i.e. the probability of observing the data given a tree. In addition to the maximum likelihood method, it includes the prior probability of a phylogeny, i.e. tree, in the testing of a hypothesis (Huelsenbeck et al., 2002). In phylogenetic inference, Bayes theorem (Equation 7.2) can be expressed as:

\[ P(\tau|D) = \frac{P(D|\tau)P(\tau)}{P(D)} \] (7.3)

where

- \( P(\tau) \) is the prior probability of a tree
- \( P(\tau|D) \) is the posterior probability of a tree
- \( P(D|\tau) \) is the likelihood of a tree
- \( P(D) \) is the probability of data, which is an aligned sequence of characters

Unlike maximum likelihood that searches for the most likely tree, Bayesian inference of phylogeny is based upon finding a large number of trees with a high posterior probability. The number of all possible trees \( B(s) \) for \( s \) species depends on the number of species \( s \). For rooted bifurcating trees

\[ B(s) = \frac{(2s-3)!}{2^{s-2}(s-2)!} \] (7.4)

\[ \begin{align*}
B(2) &= 1 \\
B(3) &= 3 \\
B(4) &= 15 \\
B(5) &= 105 \\
B(6) &= 945 \\
B(7) &= 10395
\end{align*} \]
B(8) = 135135
B(9) = 2027025
B(10) = 34459420
B(20) = 8.200795 \times 10^{21}
B(50) = 2.752921 \times 10^{76}

while for unrooted bifurcating trees

\[ B(s) = \frac{(2s-3)!}{2s-3(s-3)!} \]  

(7.5)

B(2) = 1
B(3) = 1
B(4) = 3
B(5) = 15
B(6) = 105
B(7) = 945
B(8) = 10395
B(9) = 135135
B(10) = 2027025
B(20) = 2.22 \times 10^{20}
B(50) = 2.84 \times 10^{74}

It is clear that for both rooted and unrooted trees the number of trees grows very fast as the number of species increases. It should be noted that term ‘tree’ refers to the way in which terminal nodes are grouped, regardless of the assignments to the internal nodes. For example, for three species (or in our case villages) we can have three different rooted trees (tree topologies). In Figure 7.3 we present three possible trees for villages Lobosh, Mihaitsi and Slaveino.

As in other character-based methods, all calculations in Bayesian inference are based on each of the sites, i.e. positions in the aligned sequences separately.\(^1\) To calculate the posterior probability of a tree (tree topology), we need a prior probability of a tree (tree topology) and a likelihood of a tree which is based on the observed data in each of the positions in the alignments separately. The posterior probability of a phylogenetic tree \(\tau_i\) for the \(i\)th position can be calculated using the following formula:

\[ P(\tau_i|D) = \frac{P(\tau_i)P(D|\tau_i)}{\sum_{j=1}^{\sum_{B(s)}} P(D|\tau_j)P(\tau_j)} \]  

(7.6)

\(^1\)In molecular biology term ‘site’ is used to refer to a position in a DNA or protein sequence. In dialectometry, and throughout this thesis, we use term ‘site’ to refer to a location where the data comes from. In order to avoid misunderstanding, in this chapter we will refer to a specific position in a sequence as a ‘position’ or simply try to give a descriptive explanation.
P(τ_i) is a prior probability of the \( i \)th tree. The use of prior probability sets Bayesian inference apart from the maximum likelihood method. This is considered the strongest and at the same time the weakest point of the Bayesian inference. If we have reliable information on the priors, it can help us get better posterior estimates, and it can be 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 \( P(τ_i) = \frac{1}{|B(3)|} \). In this case, Bayes inference and maximum likelihood do not differ in the trees they prefer. However, final result in a maximum likelihood approach is a single tree, while Bayesian approach provides the whole distribution of trees. This enables us to sample a large number of high probability trees from the posterior.

\( P(D|τ_i) \) is the likelihood of the \( i \)th tree, i.e. the probability of observing the data at the \( i \)th site. To be able to calculate the likelihoods we need the phylogenetic model that consists of a tree \( τ_i \), branch lengths on the tree \( v_i \) and the substitution model \( θ \). To illustrate how the likelihoods are calculated we use aligned transcriptions of the word бели /beli/ ‘white - pl.’ for three villages as our observed data (Figure 7.4). In this example the states of the characters are the phones themselves. In our example we will focus on the second position in our aligned data (character2). In Figure 7.5 we present one of the trees for the second position in the alignment given in Figure 7.4.
7.4. BAYESIAN INFERENCE OF PHYLOGENY

<table>
<thead>
<tr>
<th>character1</th>
<th>character2</th>
<th>character3</th>
<th>character4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lobosh:</td>
<td>b</td>
<td>'e'</td>
<td>l</td>
</tr>
<tr>
<td>Mihaltsi:</td>
<td>b</td>
<td>'e'</td>
<td>l</td>
</tr>
<tr>
<td>Slaveino:</td>
<td>b</td>
<td>'e'</td>
<td>l</td>
</tr>
</tbody>
</table>

Figure 7.4: A scheme of the aligned transcriptions for word ‘white’ for 3 villages.

There are three terminal nodes ‘e’, ‘e’, ‘ɛ’, one internal (y) and a root node (x). The branches are labeled from v1 to v4. An internal node and a root node can have any state inherent for the second character. In our case, 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 Figure 7.6 we present one of the possible assignments of the states for the nodes x and y. We note that there is only one change of states on the tree in Figure 7.6: ‘e’ → ‘ɛ’. We mark it with a dashed horizontal line on branch v4. Branch lengths in a tree represent the number of changes that have occurred in a certain branch. For example, in Figure 7.6 there is one change on branch v4, meaning that this branch has length 1.

To be able to calculate the likelihood of the ith tree $P(D|\tau_i)$, apart from a tree $\tau_i$ with branch lengths $v_i$, we need a substitution model $\theta$. The substitution model $\theta$ is a model of how one state changes into the other, i.e. a model that specifies the probability of one state changing into the other. $\theta$ operates both on the leaf nodes such as [‘e’] and [‘ɛ’], but also on internal nodes such as x and y. In our example, we would need to know the probability of one phone changing into any other phone present in the aligned sequences. In the simplest model, a character can go from any state into any other state. The probability of going from one into the other state is equal for all pairs of states. This is neither very realistic for most of the data in biology, nor for the language data. We know that phones are not equally likely to change into all other phones, but prefer some
Figure 7.6: An example of a possible state assignments for the internal and a root node.

changes. More complex phylogenetic models allow different rates of change between the states, which 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 → ‘e and for ‘e → ‘e. Another parameter that we can add to the phylogenetic model is the ‘site heterogeneity rate’. It allows us to specify if different characters, i.e. positions, evolve at the same or different rate. In our linguistic example, it is more likely that some characters evolve faster since changes are more frequent at the beginning and at the end of words, than in the middle. In phylogenetic models this is set 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).

The likelihood of a tree $P(D|\tau_i)$ is calculated by integrating over all possible combinations of branch lengths ($v_i$) and substitution model parameters (Huelsenbeck et al., 2002):

$$P(D|\tau_i) = \int_{v_i,\theta} P(D|\tau_i, v_i, \theta) P(v_i, \theta) dv_i d\theta$$

(7.7)

where $P(v_i, \theta)$ is the prior probability density of the branch lengths and substitution model parameters, and $d\theta$ is an infinitesimal interval. The likelihood $P(D|\tau_i, v_i, \theta)$ is normally calculated under a Markov model of character evolution—the probability of every node is dependent only on the preceding node and the branch length between these two nodes. This assumes that all positions and all lineages (villages in our case) evolve independently. The likelihood of the tree in Figure 7.6 is the product of the probabilities of every node in the tree:

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

Probability of one state changing into the other, ‘e → ‘e or ‘e → ‘e in our example, given a certain branch length ($v1-v4$), is defined by the substitution model $\theta$. The likelihood of a tree $\tau_i$ 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 $y$ a root node $x$ given a certain branch length $v_j$).

We now go back to Formula 7.6 (we repeat it for convenience) used to calculate the posterior probability of a phylogenetic tree $\tau_i$:

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

where $P(\tau_i)$ is a prior probability of the $i$th tree, and $P(D|\tau_i)$ the likelihood of the $i$th tree. The remaining element is a denominator $\sum_{j=1}^{B(s)} P(D|\tau_j)P(\tau_j)$ used as a normalizing constant. It denotes marginal probability of the data, obtained by summing the probability of the data under the assumption of all the different trees. $B(s)$ is a number of all possible trees for $s$ species. For both rooted and unrooted trees the number of trees grows very fast as the number of species increases. It is computationally extremely expensive to calculate the denominator in Equation 7.4 (repeated in 7.8) and in the general case not feasible at all.

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

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. In order to sample from a posterior probability distribution on trees, Bayesian inference in phylogeny uses Markov Chain Monte Carlo (MCMC) modeling. 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 run 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.

In the past decade Bayesian MCMC inference has become a very popular method in molecular phylogenetics. The possibility of including priors in the analysis makes it a potentially very powerful technique that sets it apart from similar statistical methods. Thanks to Monte Carlo sampling, it is also faster than the maximum likelihood method, which requires heavy computation, an issue with both of these methods (Archibald,
Mort, and Crawford, 2003). Recently there have been several attempts to apply this method to language data (Gray and Atkinson, 2003; Pagel, Atkinson, and Meade, 2007; Greenhill and Gray, 2009). However, they were used either on cognate sets or on lexical data from various languages. The present chapter is, to our knowledge, the first attempt to apply it directly on dialect phonetic data.

In this section we have tried to give a general overview of how Bayesian phylogenetic method works. For more technical and detailed explanation on Bayesian inference we refer an interested reader to Huelsenbeck et al. (2001) and Huelsenbeck et al. (2002). A very good, less technical, description of Bayesian inference can be found in Greenhill and Gray (2009).

7.5 Experiment

In the research described in this section, we apply Bayesian inference to the dialect phonetic data in order to discover the relationships between various sites, but also between the phones found in our data set. All calculations related to the Bayesian MCMC inference were done using the BEAST software (Drummond and Rambaut, 2007). The experiment was set as follows.

- We proceed by automatically multi-aligning 152 word transcriptions in the data set. We use the ALPHAMALIG algorithm described in Section 6.2. The algorithm is given a constraint that vowels can be aligned only with vowels and consonants only with consonants. The evaluation of the multiple sequence alignments produced by the ALPHAMALIG algorithm, when this constraint is used, has shown that they correspond well with the gold standard alignments and can be used in our experiment for further analyses. For example, in Figure 7.7 we present multi-aligned pronunciations for words вечер /vetʃər/ ‘evening’, дно /dɨnə/ ‘bottom’ and лесно /lesno/ ‘easily’ for five villages.

- If there are multiple pronunciation of a certain word in some villages, we randomly chose only one pronunciation per site in order to conform to the format that can be handled by the software used for Bayesian inference.

- After multi-aligning transcriptions for every word separately, we merge 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

---

2This experiment was conducted during the research visit to the University of Auckland. We would like to thank Prof. Russell Gray and Prof. Alexei Drummond for their help with setting this experiment and using the BEAST software.
7.5. EXPERIMENT

Aldomirovtsi: \( v \ 'e \ \tilde{t} \ e \ \tilde{r} \)  
Aldomirovtsi: \( d \ - \ n \ 'o \)
Asparuhovo-Lom: \( v \ 'e \ \tilde{t} \ e \ \tilde{r} \)  
Asparuhovo-Lom: \( d \ - \ n \ 'o \)
Asparuhovo-Prov: \( v^{|j} \ 'e \ \tilde{t} \ \tilde{r} \)  
Asparuhovo-Prov: \( d \ '\chi \ n \ u \)
Babyak: \( v \ 'e \ \tilde{t} \ e \ \tilde{r} \)  
Babyak: \( d \ - \ n \ 'o \)
Bachkovo: \( v \ 'e \ \tilde{t} \ e \ \tilde{r} \)  
Bachkovo: \( d \ 'a \ n \ u \)

Figure 7.7: Multiple alignments for three words and five villages.

do not take into account any information related to the word level (e.g. lexical identity, lexical semantics, specific context in which certain phone occurs). In Figure 7.8 all pronunciations of the three words presented in step 1 are merged into a single set of multi-aligned strings.

Aldomirovtsi: \( l \ '\gamma \ s \ n \ o \)  
Asparuhovo-Lom: \( l \ 'e \ s \ n \ o \)  
Asparuhovo-Prov: \( l^{|j} \ 'e \ s \ n \ u \)  
Babyak: \( ? \ ? \ ? \ ? \)  
Bachkovo: \( l^{|j} \ 'e \ s \ n \ u \)

Figure 7.8: Pronunciation of different words merged into a single string.

We do not use any information on where one words 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.

- 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 620 positions in total. At the same time alignments in biology would normally contain longer sequences for a much smaller number of species. The other difference is in the number of unique
tokens: for 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. Having a large number of different symbols in some columns on one hand, and such a small number of columns on the other, makes it impossible for the algorithm to reach convergence and obtain the desired analyses correctly. For that reason the data set was reduced to only the columns that contain vowels. As we have seen in Chapter 5 vowel changes are more frequent and more diverse. Consonant changes occur much less frequently and in most of the cases involve palatalization. We argue that on dialect level, most of the information on the language change and variation can be inferred from the processes related to vowel changes. Since, for technical reasons, we are forced to reduce the number of analyzed phones, we chose to base our analyses on the vowel changes only. From the merged alignments we removed all columns that contain consonants, making the total number of columns 303. After removing all the consonants, our example presented in Figure 7.8 would look like this:

Aldomirovtsi: 'e e - 'o 'y o
Asparuhovo-Lom: 'e e - 'o 'e o
Asparuhovo-Prov: 'e ø 'y u 'e u
Babyak: 'e e - 'o ? ?
Bachkovo: 'e e 'a u 'e u

Figure 7.9: Only columns with vowels are kept in the merged multiple string alignment.

- After reducing our data set only to vowels, there were still 43 different phonetic segment symbols, including various diacritics and suprasegmentals. It is still a much larger number of segments that any software made to process biological data is able to handle. In order to get smaller number of symbols, we have removed all diacritics and suprasegmentals and reduced our set of symbols to 16. In Table 7.1 we list the reduced set of symbols on the left hand side, and the full, unreduced, set on the right hand side.

- Taking into consideration the short length of strings (303 positions), 16 different symbols was still too large a number to be processed successfully. In the final reduction step, all 16 symbols were put into one of the 8 groups based on their position in the vowel chart (Figure 7.10). Finally, the data set is transformed into the format shown in the example in Figure 7.11.
7.5. EXPERIMENT

Table 7.1: Vowel inventory after removing all diacritics and suprasegmentals.

<table>
<thead>
<tr>
<th>reduced set</th>
<th>full set</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>e</td>
<td>e</td>
</tr>
<tr>
<td>e</td>
<td>e:</td>
</tr>
<tr>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>y</td>
<td>y:</td>
</tr>
<tr>
<td>d</td>
<td>d</td>
</tr>
<tr>
<td>i</td>
<td>i</td>
</tr>
<tr>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>u</td>
<td>u</td>
</tr>
<tr>
<td>o</td>
<td>u:</td>
</tr>
<tr>
<td>o</td>
<td>u:</td>
</tr>
<tr>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>a</td>
<td>a:</td>
</tr>
<tr>
<td>i</td>
<td>i</td>
</tr>
<tr>
<td>i</td>
<td>i:</td>
</tr>
<tr>
<td>e</td>
<td>e</td>
</tr>
<tr>
<td>e</td>
<td>e</td>
</tr>
<tr>
<td>i</td>
<td>i</td>
</tr>
<tr>
<td>y</td>
<td>y</td>
</tr>
</tbody>
</table>
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); b) a position heterogeneity model (h) and c) a molecular clock (m).

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 Figure 7.12, we repeat the alignment presented in Figure 7.11 but mark it with ‘s’, ‘h’ and ‘m’ to show which model applies to which part of the alignments. In our example substitution model, marked with ‘s’, 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.

The site (position) heterogeneity model allows us to specify whether the rate of variation in different position, marked with ‘h1’, ‘h2’, ..., ‘h6’ in our example in Figure 7.12, 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 in which positions the substitutions are more or less frequent, but some settings allow different columns to vary at different rates.
### 7.5. EXPERIMENT

<table>
<thead>
<tr>
<th>Setting</th>
<th>Location</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
<th>Model 5</th>
<th>Model 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>m(1)</td>
<td>Aldomirovtsi:</td>
<td>3 [e]</td>
<td>3 [e]</td>
<td>-</td>
<td>5 [o]</td>
<td>5 [\v]</td>
<td>5 [o]</td>
</tr>
<tr>
<td>m(2)</td>
<td>Asparuhovo-Lom:</td>
<td>3 [e]</td>
<td>3 [e]</td>
<td>-</td>
<td>5 [o]</td>
<td>3 [e]</td>
<td>5 [o]</td>
</tr>
<tr>
<td>m(3)</td>
<td>Asparuhovo-Prov:</td>
<td>3 [e]</td>
<td>4 [e]</td>
<td>5 [\v]</td>
<td>8 [u]</td>
<td>3 [e]</td>
<td>8 [u]</td>
</tr>
<tr>
<td>m(4)</td>
<td>Babyak:</td>
<td>3 [e]</td>
<td>3 [e]</td>
<td>-</td>
<td>5 [o]</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>m(5)</td>
<td>Bachkovo:</td>
<td>3 [e]</td>
<td>3 [e]</td>
<td>2 [\u]</td>
<td>8 [u]</td>
<td>3 [e]</td>
<td>8 [u]</td>
</tr>
</tbody>
</table>

Figure 7.12: Three models of evolution apply to the parts of the alignments marked with ‘s’ (substitution model), ‘h’ (rate heterogeneity model), and ‘m’ (molecular clock model).

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, in our case language varieties marked with ‘m1’, ‘m2’, ..., ‘m5’, change constantly over time. This is the basic, and the simplest molecular clock model implemented in BEAST. Since in this experiment our data is rather limited, we tried to build simple models and get reliable estimates of our parameters. 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 models:

- 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 from any other group and the probability of, for example, [a] changing into [\v] 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, for example, [a] changing into [\v] 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 Gamma. This setting allows various substitution rates between different positions, i.e. it allows for the phones in different positions within the words to evolve differently. Unlike in the previous two settings, we assume that, for example, position h(1) might evolve slower or faster than position h(6).

The length of the chain, i.e. the number of generations that the MCMC algorithm ran for, was $4 \times 10^7$ 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 intuition. 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.

### 7.6 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 height 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.

In Figure 7.13 we present the dendrogram where the trees produced using Setting 1 are summarized. On all dendrograms in this section we present the posterior probabilities of nodes. Due to the large number of sites in our data set, node labels were not readable. We have removed them from all dendrograms. In the dendrogram in Figure 7.13 we can see that on the highest level the split at the root node has maximum posterior probability 1. We mark two-way split with red and blue, where red represents eastern varieties and blue western and southern. In order to see the geographical distribution of the two groups of sites, we present this two-way split of Bulgarian dialect varieties on a map (see Figure 7.15). Two groups of sites are marked with red dots (eastern varieties) and blue dots (western and southern varieties). The two-way division of

---

The height of a node is the length of the longest downward path from that node to a leaf.
sites 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. Unlike in the aggregate analyses presented in Chapter 3, Rupian dialects are grouped together with the western, rather than with the eastern varieties. One step lower in the dendogram, there is a split with posterior probability of 0.898. According to the analysis performed, we can assume this split with a high confidence. It divides southern varieties from the western. On dendrogram in Figure 7.14 we mark the southern varieties with green and western with blue. Classification of the western varieties into a single group is supported with maximum posterior probability, while the grouping of southern varieties is much less certain since the node that is on the top of this group has posterior probability of 0.531. Although according to the posterior probability it is not highly certain that these sites form a group, they largely occupy a geographically coherent area in the south of the country. Some of the varieties placed in this group are found along the yat border. We present the three-way classification produced using Setting 1 in Figure 7.16. Based on the branch lengths in the dendrogram, groups presented on this map form three distinct varieties. Since in Setting 1 the probability of any state, i.e. any phone changing into any other state was set to be equal we could not get any interesting information on vowel changes from this setting.

In Figure 7.17 we present the tree that summarizes the trees resulting from the Bayesian inference performed once we adopted the General Time Reversible (GTR) model. The two-way split at the root node that has maximum posterior probability, shows a split of the sites into western and eastern. The southern group of varieties is classified with the eastern dialects (Figure 7.17). Just as with the previous dendrogram, we show this split on the map of Bulgaria (Figure 7.19). This division corresponds well with the division of the sites based on the aggregate analysis (Chapter 3) 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 (see map in Figure 7.20). 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 have low posterior probabilities, namely 0.134 and 0.526. 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 high posterior probability and can be taken with great confidence to form a coherent group. 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 two groups of sounds calculated from the data.
Figure 7.13: Dendrogram that summarizes the trees produced using Setting 1: free substitution model and no positional heterogeneity. Two-way division of the sites.
Figure 7.14: Dendrogram that summarizes the trees produced using Setting 1: free substitution model and no positional heterogeneity. Three-way division of the sites. This is a more detailed view of the dendrogram in Figure 7.13.
Figure 7.15: Distribution of the two group of sites using a free substitution model and no positional heterogeneity model (Setting 1).

Figure 7.16: Distribution of the three group of sites (Setting 1).
Figure 7.17: Dendrogram that summarizes the trees produced using Setting 2: GTR model with no positional heterogeneity. Two-way division of the sites.
Figure 7.18: Dendrogram that summarizes the trees produced using Setting 2: GTR model and no positional heterogeneity. Three-way division of the sites.
Figure 7.19: Distribution of the two group of sites using a GTR substitution model and no positional heterogeneity (Setting 2).

Figure 7.20: Distribution of the three group of sites (Setting 2).
Figure 7.21: The most probable vowel transitions, marked with blue, under the GTR model no positional heterogeneity. Groups (1) and (7) are put in dashed boxes to indicate that our estimations concerning these groups are unreliable.

Figure 7.22: The least probable vowel transitions, marked with red, under the GTR model with no positional heterogeneity.
Figure 7.23: Dendrogram that summarizes the trees produced using Setting 3: GTR model with gamma positional heterogeneity. Two-way division of the sites.
Figure 7.24: Dendrogram that summarizes the trees produced using Setting 3: GTR model and gamma positional heterogeneity. Three-way division of the sites.
7.6. RESULTS

The results can be seen in Figure 7.21, where we present sound changes with the highest probabilities (connected with blue lines) and in Figure 7.22 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 how probable the change of a vowel in one group into a vowel in another is. In Figures 7.21, 7.22, 7.27 and 7.28 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 [e, e], which is the most frequent group in the data set, appear 14663 times. As marked with the blue lines in the vowel chart in Figure 7.21, changes that received the highest probability are between the following groups: 5 [ε, ə, ɨ, ə] and 8 [ʊ, u], 3 [ɛ, e] and 6 [ɪ, ɨ, ɪ], 4 [ɔ] and 6 [ɪ, ɨ, ɪ], and 2 [a, ʊ] and 4 [ɔ]. 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 [ɔ].
In the chart in Figure 7.22 we mark the changes between the groups with the lowest probabilities using red lines: 2 [a, u] and 8 [o, u], 3 [e, e] and 8 [o, u], and 5 [a, o, y, o] and 6 [i, y, 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 state was allowed to change into any other state 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. From the dendrogram in Figure 7.23 we also extracted the two-way division of the sites and represented it on the map in Figure 7.25. In Figure 7.24 we mark three groups extracted and show that division in Figure 7.26. 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 7.27 and Figure 7.28 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 [a, o, y, o] and 8 [o, u], 3 [e, e] and 6 [i, y, i], and 4 [o] and 6 [i, y, 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.

In Table 7.6 we give the values of the modified Rand index (MRI) presented in Sec-
7.6. RESULTS

Figure 7.27: The most probable vowel transitions, marked with blue, using GTR and gamma site heterogeneity model.

Section 3.6.1 for the pairwise comparison of the classifications produced in all three settings, classification done by weighted pair group method using arithmetic averages (WPGMA) clustering algorithm and the traditional division of the sites according to Stoykov (2002). We note very high agreement between the 2-way divisions produced using Setting 2 and Setting 3: 0.939. There is also very high agreement between the 3-way divisions produced by Setting 1 and Setting 2: 0.945. Agreement on the 2-way and 3-way divisions produced by Setting 2 and Setting 3 in Bayesian inference experiment and WPGMA clustering algorithm is lower, but still high, ranging from 0.686 to 0.763. The 2-way division produced by the Setting 1 has lower values for MRI since, unlike WPGMA, it groups southern varieties with the western and not with the eastern dialects. Comparison of the divisions resulting from Setting 2 and Setting 3 to the traditional divisions as suggested by Stoykov (2002), shows that they give similar values of MRI that we get by comparing the divisions produced by WPGMA and traditional classification.

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 at some positions in words some sound changes are more likely than in some 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 ratio of the marginal likelihoods of the two models:

$$K = \frac{P(D|H1)}{P(D|H2)}$$
where $P(D|H)$ expresses the marginal likelihood of a hypothesis $H$. For a more detail explanation see Kass and Raftery (1995) or MacKay (2003). For our two settings we calculated the Bayes factor using the Tracer software.\footnote{http://tree.bio.ed.ac.uk/software/tracer} In Table 7.3 we present the values of the Bayes Factor in log 10 scale obtained after pairwise comparing all three settings.

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 a 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.

### 7.7 Discussion

In recent years there has been an increasing number of studies that apply methods taken from phylogenetics to the research of language change and evolution. However, only very few of them apply those techniques on the phonetic or phonological data (Nakhleh, Ringe, and Warnow, 2005; Warnow et al., 2006; McMahon et al., 2007). In previous work phonological and phonetic data was very carefully manually selected and coded, based on the substantial linguistic knowledge. We do not argue against the linguistic
Table 7.2: The modified Rand index (MRI) for the 2-fold and 3-fold divisions of sites produced by three various Bayesian inference settings (‘s1’, ‘s2’, ‘s3’), WPGMA (‘WA’) and traditional scholarship (‘trad.’).

<table>
<thead>
<tr>
<th></th>
<th>s1 (2)</th>
<th>s2 (2)</th>
<th>s3 (2)</th>
<th>WA (2)</th>
<th>trad. (2)</th>
<th>s1 (3)</th>
<th>s2 (3)</th>
<th>s3 (3)</th>
<th>WA (3)</th>
<th>trad. (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>s1 (2)</td>
<td>-</td>
<td>0.301</td>
<td>0.312</td>
<td>0.467</td>
<td>0.290</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>s2 (2)</td>
<td>-</td>
<td></td>
<td>0.939</td>
<td>0.717</td>
<td>0.716</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>s3 (2)</td>
<td>-</td>
<td>-</td>
<td></td>
<td>0.734</td>
<td>0.665</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>wa (2)</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td>0.700</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>trad. (2)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>s1 (3)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.945</td>
<td>0.854</td>
<td>0.727</td>
<td>0.601</td>
<td>-</td>
</tr>
<tr>
<td>s2 (3)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.829</td>
<td>0.763</td>
<td>0.597</td>
<td>-</td>
</tr>
<tr>
<td>s3 (3)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.686</td>
<td>0.543</td>
<td>-</td>
</tr>
<tr>
<td>wa (3)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.626</td>
<td>-</td>
</tr>
<tr>
<td>trad. (3)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 7.3: Values of the Bayes factor in log 10 scale. There is a strong support for Setting3 when compared to both Setting1 and Setting2.

<table>
<thead>
<tr>
<th>Setting</th>
<th>Setting 1</th>
<th>Setting 2</th>
<th>Setting 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Setting 1</td>
<td>-</td>
<td>-573.369</td>
<td>-938.271</td>
</tr>
<tr>
<td>Setting 2</td>
<td>573.369</td>
<td>-</td>
<td>-364.902</td>
</tr>
<tr>
<td>Setting 3</td>
<td>938.271</td>
<td>364.90</td>
<td>-</td>
</tr>
</tbody>
</table>

data coding, but we do try to apply a more robust and language-independent approach. In this research we have tried to automate the process of character selection by automatically multi-aligning phonetic transcriptions and using them as input to software for phylogenetic inference. However, we were not able to explore sound correspondences in all their varieties, since the number of phones in our data set was too large to be successfully processed by any software developed for the computational phylogenetics. We have restricted ourselves to the investigation of the vowel changes, since the analysis of the sounds presented in Chapter 5 has shown that most of the variation in our dialect data is between the vowels. In comparison to the consonants, they are more likely to contain sufficient information on dialect change. By putting all the vowels into eight groups we have tried to keep in our analyses at least the main articulatory characteristics (open/close and front/back opposition) of the vowels. 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 Bayesian 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 changes of the phones are not influenced by the changes of the proceeding or the 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. In future we hope to implement a model that would relax the assumption of independence, at the cost of substantial complexity.

One of the models that is being heavily debated in linguistics is the lexical clock. While some of the authors used this assumption in their attempts to date Proto Indo-European (Forster and Toth, 2003), others heavily criticize the usage of a uniform lexical clock (Eska and Ringe, 2004). A strict molecular clock model assumes that all lineages (language varieties) evolve at a constant rate. We have used this assumption in our experiments since it is the basic molecular clock model in the software for phylogenetic inference, and it makes the estimation of the other parameters easier, especially with such a small data set as ours. 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 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 obtained the same division of Bulgarian dialect area using the Levenshstein method that is based on the similarity between the pronunciation strings without any assumptions on the genetic relatedness of the compared varieties. Two alternative approaches gave very similar picture of the Bulgarian dialect variation. However, these two approaches are very hard to separate in the case of dialect data where we a priori test varieties that are genetically
very closely related.

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, and see if, for example, [e] is more likely to change into [i] or the other way around. We hope to achieve this in future. 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.