BAYESIAN CLASSIFICATION OF DNA BARCODES

by

MICHAEL P. ANDERSON

B.A., Utah State University, 2003
M.S., Kansas State University, 2006

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas
2009

Abstract

DNA barcodes are short strands of nucleotide bases taken from the cytochrome c oxidase subunit 1 (COI) of the mitochondrial DNA (mtDNA). A single barcode may have the form C C G G C A T A G T A G G C A C T G . . . and typically ranges in length from 255 to around 700 nucleotide bases. Unlike nuclear DNA (nDNA), mtDNA remains largely unchanged as it is passed from mother to offspring. It has been proposed that these barcodes may be used as a method of differentiating between biological species (Hebert, Ratnasingham, and deWaard 2003). While this proposal is sharply debated among some taxonomists (Will and Rubinoff 2004), it has gained momentum and attention from biologists. One issue at the heart of the controversy is the use of genetic distance measures as a tool for species differentiation. Current methods of species classification utilize these distance measures that are heavily dependent on both evolutionary model assumptions as well as a clearly defined "gap" between intra- and interspecies variation (Meyer and Paulay 2005). We point out the limitations of such distance measures and propose a character-based method of species classification which utilizes an application of Bayes' rule to overcome these deficiencies. The proposed method is shown to provide accurate species-level classification. The proposed methods also provide answers to important questions not addressable with current methods.

BAYESIAN CLASSIFICATION OF DNA BARCODES

by

MICHAEL P. ANDERSON

B.A., Utah State University, 2003
M.S., Kansas State University, 2006

A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas
2009

Approved by:

Major Professor
Suzanne Dubnicka

Copyright

Abstract

DNA barcodes are short strands of nucleotide bases taken from the cytochrome c oxidase subunit 1 (COI) of the mitochondrial DNA (mtDNA). A single barcode may have the form C C G G C A T A G T A G G C A C T G . . . and typically ranges in length from 255 to around 700 nucleotide bases. Unlike nuclear DNA (nDNA), mtDNA remains largely unchanged as it is passed from mother to offspring. It has been proposed that these barcodes may be used as a method of differentiating between biological species (Hebert, Ratnasingham, and deWaard 2003). While this proposal is sharply debated among some taxonomists (Will and Rubinoff 2004), it has gained momentum and attention from biologists. One issue at the heart of the controversy is the use of genetic distance measures as a tool for species differentiation. Current methods of species classification utilize these distance measures that are heavily dependent on both evolutionary model assumptions as well as a clearly defined "gap" between intra- and interspecies variation (Meyer and Paulay 2005). We point out the limitations of such distance measures and propose a character-based method of species classification which utilizes an application of Bayes' rule to overcome these deficiencies. The proposed method is shown to provide accurate species-level classification. The proposed methods also provide answers to important questions not addressable with current methods.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to acknowledge the unfailing support of my wife Anne and our three boys William, Isaac and Jacob. I love them with all of my heart.

I would like to acknowledge my major professor Dr. Suzanne Dubnicka for her superb guidance. She has gone over several versions of this manuscript and made improvements, comments that have contributed to the content, and questions that have spurred on further research.

I must give a special thanks to Dr. Haiyan Wang for not only serving on my committee, but for introducing me to DNA barcoding. A great deal of this work was conceptualized as a result of our initial discussions about the topic.

I would like to acknowledge Dr. Paul Nelson for serving on my committee and for his helpful comments and suggestions for research when I was first embarking on this journey. I also thank Dr. Karen Garrett for her willingness to serve on my committee and for providing insight into the biological aspects of this work.

A special thanks goes to my brothers of statistics, Dr. George Von Borries, Robert Poulson, and Wijith Munasinghe. Steinbeck once pronounced a blessing that I wish to invoke upon these good men - "May they live a thousand years and people the earth with their offspring."

Finally, I thank the rest of the Kansas State Statistics Department faculty, staff, and graduate students for their support and kindness.

Chapter 1

Introduction

1.1 Challenges in Taxonomy

Taxonomists face great challenges regarding the classification and discovery of congeneric, or closely related, species. In order to determine an organism's species, taxonomy relies mainly upon inspection of an organisms easily observed and described morphologic features, such as shapes, colors, sizes, and behaviors. These morphologic features are then compared against what has been previously observed about a species, and classification follows.

Reliance upon these physical characteristics, or morphological features, to determine species is an enormous challenge for several reasons. First, physical characteristics between two congeneric species may be so similar that they would typically be classified as the same species using these morphological diagnoses. On the other hand, physical characteristics between two organisms of the same species may appear quite different, resulting in the classification of two separate species. An example of this latter situation is discussed by Cooke, Rockwell, and Lank (1995) where the blue and white morphs of Snow Goose, *Chen caerulescens*, were thought to be two distinct species until only recently. Also, males and females of the same species often have different physical characteristics. This too can make correct species classification challenging. This situation can also arise when morphologic features for a particular species develop as the organism ages. For example, some frog species are easily distinguished at maturation due to color features (such as spots or stripes), but

as tadpoles, these features are absent making classification at that stage very difficult and much less certain.

Second, because these morphologic features can only be compared to what has already been observed, the process of discovery of new species can be very slow. According to Hebert, Ratnasingham, and deWaard (2003), an individual taxonomist can rarely identify more than 1000-1500 different species. This means that when observed features do not match those of any species known to a taxonomist, there must be a vast amount of collaboration with others before it can be decided a new species has been discovered. We can get a feel for the magnitude of the task at hand by noting that, in the millennia of recorded history, of the estimated 10-15 million species on earth (excluding bacteria and archaea) (Hammond 1992), taxonomists have classified roughly 1.7 million of them (Stoeckle 2003).

Lastly, it is sometimes necessary to make species classifications from organism fragments. These fragments may not include enough morphologic detail to assign the organism to a species with any amount of certainty. Fragment identification may be necessary, for example, when identifying the remnants in the stomach of a predator; identifying the type of bird that strikes an aircraft (Dove 2000); or identifying a carcass as belonging to a protected or regulated species (Guglich, Wilson, and White 1994). Identification based on morphologic features in these cases can be virtually impossible. Natural history museums often contain repositories of organism fragments that account for a large amount of the biodiversity on earth. A systematic method of identification for these archival organism fragments could prove to be an important step in the direction of classifying all of the species on earth.

These difficulties in classification and discovery of species necessitate research for a more precise and speedy discrimination among species that can complement the shortcomings of classification based solely on morphological features. It would be desirable to develop a method of species identification that would prove effective at every stage of the organism's life, provide quick and efficient comparisons of organisms to discover new species, and allow for proper grouping of species based on less than the complete organism. It seems that

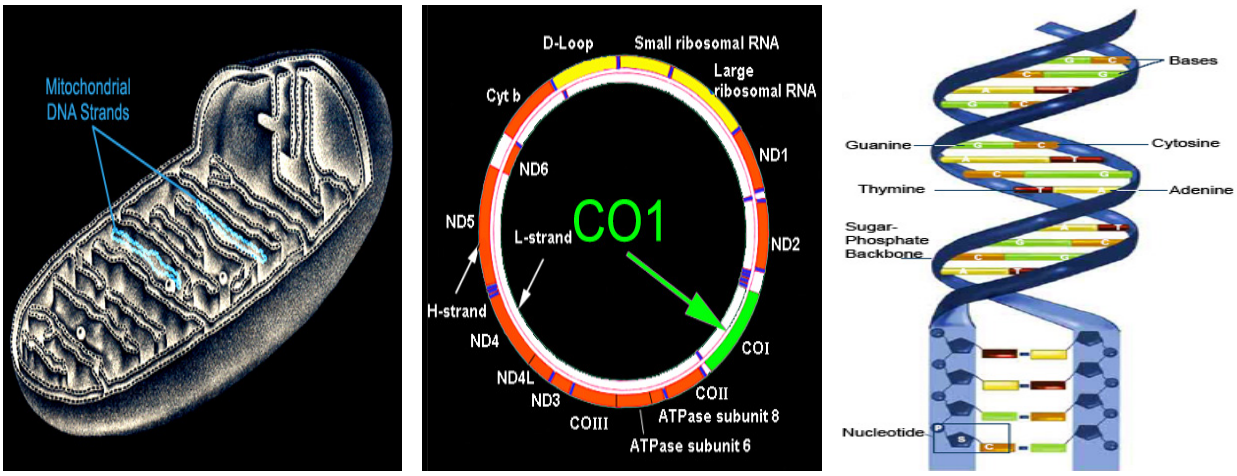

Figure 1.1: *(a) Mitochondria, (b) Location of the cytochrome c oxidase subunit 1 (COI) on a strand of mtDNA, (c) Nucleotide bases in a strand of mtDNA. Images provided by Hebert Laboratory and the National Institute of Medical Sciences.*

such a method capable of all this might be difficult to find, but recent advances in DNA (deoxyribonucleic acid) research have opened doors that allow us to consider such a method.

1.2 DNA Barcoding

Developments in genetic research indicate that a short DNA sequence known as a barcode, taken from the cytochrome c oxidase subunit 1 (COI) location of mitochondrial DNA (mtDNA), is an effective marker for identifying species in the animal kingdom (Hebert, Cywinska, Ball, and deWaard 2003). See Figure 1.1 (a) and (b). This barcode contains a sequence of the nucleotide bases: adenine (A), thymine (T), cytosine (C), and guanine (G). A single barcode may have the form C C G G C A T A G T A G G C A C T G ... and typically ranges in length from 255 to around 700 nucleotide bases. See Figure 1.1 (c).

While barcodes for the same species may not be identical, they will rarely have more than 2% divergence and will often have less than 1% divergence (Johns and Avice 1998). Hebert et al. (2003) found that between species that are closely related, sequence divergence averaged around 6.8%, with 99.98% having sequence divergence greater than 3%. It should

also be noted that they found sequence divergence to be higher between species that were not closely related. This discrepency between within-species variability and among species variability has come to be known as the genetic "gap".

With the initial studies above having provided some validation to the effectiveness of using these barcodes to discriminate between species, it is hopeful that the challenges facing taxonomy mentioned in Section 1.1 can be addressed. To be sure, a comprehensive catalogue of barcodes representing the earth's biodiversity could be used to search and match new barcodes to known species, or perhaps more importantly, identify them as not belonging to a species in the catalogue, potentially representing a new species. Also, barcodes can be retrieved from a very small amount of an organism's tissue (1-3mm^3), at any stage of life (Ivanova, deWaard, Hajibabaei, and Hebert 2007). Thus, organism fragments and development or change of morphological features over time do not represent significant obstacles for DNA barcoding. Figure 1.2 shows how typical samples for DNA barcode extraction compare to a pencil. Object (a) is the leg of a lepidoptera (moth or butterfly), object (b) is a small planktonic crustacean known as a Daphnia, object (c) is a feather sample, and object (d) is a small piece of muscle tissue. Obtaining these barcodes is a relatively quick and inexpensive procedure costing around $\$3 - \$5$ per barcode (Hajibabaei et al. 2005).

1.3 Challenges in DNA Barcoding

Due to the discrete, ordered nature of the data obtained from a DNA barcode, typical methods of classification and clustering based on Euclidean distances are not sufficient. Other distance measures have emerged by comparing the bases at each position and recording the number of differences between pairs of barcodes. These distance methods, such as p-distance (Hebert et al. 2003), typically employ thresholds derived from the afformentioned genetic "gap" in Section 1.2, while others, such as Kimura's Two Parameter (K2P) model, use assumption-rich, evolution-based models in addition to these thresholds (Kimura 1980).

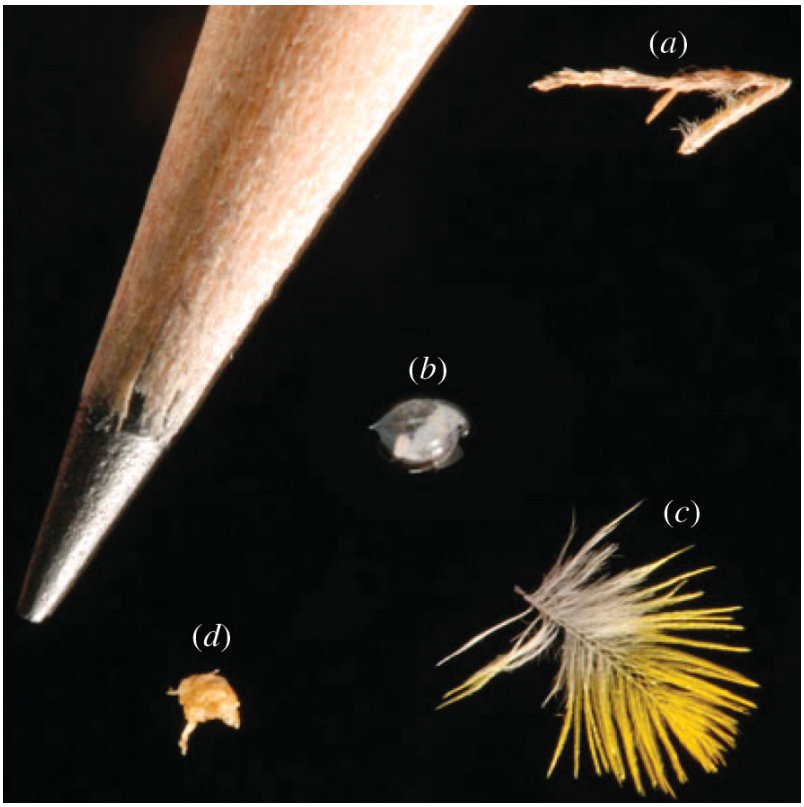

Figure 1.2: *Typical tissue samples used to obtain DNA barcodes. (a) a moth or butterfly leg, (b) a planktonic crustacean, (c) a feather fragment, (d) muscle tissue. Images provided by Philosophical Transactions of the Royal Society.*

Current methods of DNA barcode classification use these distance measures together with the Neighbor-Joining method (Saitou and Nei 1987) to reconstruct a phylogenetic tree to examine the relationship between an unknown barcode to be classified and several known barcodes. These distance methods, and their drawbacks, will be discussed in Section 2.

Large databases have begun to emerge as a result of efforts to catalogue these barcodes for all known species. The Barcode of Life Data System (BOLD) provides access to hundreds of thousands of these barcodes for a variety of species. Very large data sets are easily extracted from these data banks and can be used in the evaluation of classification methods.

The Center for Discrete Mathematics and Theoretical Computer Science (DIMACS 2007) has compiled a list of challenges such a barcoding analysis method should be capable of

addressing. Among the most intriguing and critical of the issues discussed by DIMACS is the desire for a character-based analysis method that could address questions regarding how much of the barcode is needed for proper classification. Hajibabaei et al. (2005) point out that large sequences from fresh mtDNA are easily obtained, but often, for archival mtDNA more than a decade old, sequences of more than 300-400 base positions long are rare, and sequences of 100 base positions are much more common. Also, modern sequencing methods can sequence up to 400 million nucleotide bases in about 10 hours (Roche 2009) and often require a trade-off between fewer, longer sequences and many shorter sequences. If shorter sequences can be used for proper classification, then the high through-put technologies of these sequencing methods can be more appropriately focused on obtaining more sequences of a shorter length. By comparing the nucleotide bases in corresponding positions for different barcodes sequentially, it may be possible to dramatically reduce the length of the barcode needed for classification and clustering. This dimension reduction will not only have a large impact on analyses that are highly computational, which is often the case for Bayesian methods, but also allow for older DNA specimen to be classified.

Another pressing issue is that of sample size. Current methods for DNA barcode analysis rely heavily on the well-defined "gap" between intra- and interspecies variation. If a species is represented in a data set by only one or two organisms, the intraspecies variation can be severely underestimated leading to overestimation of the accuracy of the classification method (Meyer and Paulay 2005). This is a difficulty in that DNA barcoding is relatively new and comprehensive databases that have more than a few organisms per species are rare. In some cases, a species may be represented by just a single organism. These may be rare species that are hard to locate, and thus difficult to catalogue, or they may be the result of a practitioner failing to understand the need for multiple observations within a species to accurately estimate intraspecies variability. It would be ideal to construct a method that is not so dependent upon genetic "gap" thresholds and provides accurate species-level classification even when a species is represented by only one or two organisms.

The main goal of DNA barcode analysis is to classify a set of unknown barcodes T to known species in a reference data set R, or to recognize that the unknown barcodes do not belong to any of those in the reference data set.

Understanding how classification is to take place will be important to the computational methods selected for the analysis. Because these barcode data sets can consist of many observations, we will need to carefully pursue computational methods that will not get excessively slowed down by the amount of data that will need to be processed. For example, we may seek to classify a few hundred barcodes to their proper taxonomic species. Knowing that this classification will utilize a reference data set, possibly consisting of millions of barcodes with up to 700 nucleotides each, one must carefully choose computational methods for such highly dimensional data. In this case, highly computational methods must be carefully implemented or avoided altogether. However, during the classification process, there may be a few barcodes that remain unclassified because they do not belong to any of the species in the reference data set. Among these, one may be interested in finding clusters of similar barcodes. Because this clustering would be done on a much smaller subset of T and without a reference data set, the data is not of such high dimension and there may be room for more computationally intensive methods. The focus of this dissertation with be on the former situation where we seek to classify many barcodes utilizing large reference data sets in high dimensions.

Chapter 2

Current Methods for Classification

2.1 Traditional Classification Methods

Traditional classification methods are not well suited to classifying DNA barcodes so we review some of these traditional methods and highlight deficiencies as they relate to DNA barcoding.

Discriminant functions were first proposed by Fisher (1936). For these discriminant functions, the squared Mahalanobis distance between the vector of predictor variables for a new observation $\mathbf{y}$, and the group averages of those predictor variables $\overline{\mathbf{y_i}}$ obtained from a reference data set is computed. The new observation is then classified as belonging to the group that produces the smallest squared Mahalanobis distance. More explicitly, we compute

$$D_i^2(\mathbf{y}) = (\mathbf{y} - \overline{\mathbf{y_i}})'\mathbf{S_{pl}^{-1}}(\mathbf{y} - \overline{\mathbf{y_i}}) \tag{2.1}$$

where $S_{pl}^{-1} = 1/(N-k)\sum_{i=1}^{k}(n_i - 1)S_i$ is the pooled sample covariance matrix with n_i and S_i equal to the sample size and sample covariance matrix of the i^{th} group, $N = \sum n_i$, and $k =$ the number of groups. This is used when equal variances are assumed and reduces to Fisher's linear discriminant function. For unequal variances, we replace the pooled sample covariance matrix with the i^{th} sample covariance matrix S_i which produces Fisher's quadratic discriminant function. We then assign the observation $\mathbf{y}$ to the group for which $D_i^2(\mathbf{y})$ is the smallest. One problem with using discriminant functions with DNA barcodes is that the

responses are discrete rather than continuous, thus measuring the distance between them in Euclidean distance is not possible. Also, the quadratic discriminant function requires each n_i to be greater than the number of predictors in order for S_i^{-1} to exist (Rencher 2002). This is clearly not feasible with barcode data in that we have upwards of 700 predictor variables and typically 1-7 observations per species. Lastly, these discriminant functions only provide yes/no classifications rather than measure the uncertainty associated with the classification through a probability.

Nearest Neighbors, proposed by Fix and Hodges (1951), uses a similar idea as discriminant functions but computes the squared Mahalanobis distance between the vector of predictor variables for a new observation $\mathbf{y_i}$ and the vector of predictor variables for every other observation in the reference data set. The new observation $\mathbf{y_i}$ is then classified as belonging to the group having a majority of the k-nearest neighbors. Specifically we compute

$$(\mathbf{y_i} - \mathbf{y_j})' S_{pl}^{-1} (\mathbf{y_i} - \mathbf{y_j}) \tag{2.2}$$

where S_{pl}^{-1} is the pooled sample covariance matrix discussed above. The drawbacks to this method are similar to those of discriminant functions, namely, continuous responses are required and no probability is provided to summarize classification uncertainty.

Logistic Regression can be used to compute probabilities of belonging to any one of s groups (Johnson 1998). This is done by creating $s-1$ logits and modeling group assignments based on estimated coefficients for each predictor variable. Here, discrete predictor variables can easily be incorporated into the model and a probability of group assignment can be obtained. The large number of parameters that must be estimated due to the large number of predictor variables, however, makes this classification method for DNA barcoding infeasible.

2.1.1 Naive Bayes Classifier

Another method of classification is that of the Naive Bayes (NB) classifier. As the name implies, Bayes' rule will be used to estimate the posterior probability that an observation belongs to one of s groups. "Naive" in this sense means that this classifier will use strong

independence assumptions that are often not true. Specifically, the NB classifier assumes conditional independence among all of the predictor variables of an observation. In the case of DNA barcoding, this amounts to assuming that the nucleotide positions of a barcode are conditionally independent given species. More succinctly, if the random variable group or class is denoted by S and $x^{(1)}, \ldots, x^{(p)}$ are p predictor variables, then the NB classifier is

$$class(x^{(1)}, \ldots, x^{(p)}) = \arg\max_{S} P(S = s) \prod_{i=1}^{p} P(x(i)|S = s) \tag{2.3}$$

It can be shown that this classifier is proportional to the highest posterior probability when the predictor variables are conditionally independent. One attractive aspect of the NB classifier is that the independence assumptions allow for estimating one dimensional distributions for each predictor variable which eliminates problems that arise due to the *curse of dimensionality*. This means the NB classifier can be used effectively in cases where the number of predictor variables is large while the number of observations is small. One reason for NB classifier's strong performance is that it will assign an observation to the correct group as long as the correct group is the most probable, even if the group probabilities are not well estimated. In cases where the dependent structure of the predictor variable is ignored, the group probabilities may in fact be poorly estimated, but the correct group may still be the most probable and hence, the correct classification occurs. Zhang (2004) discusses the implications of these independence assumptions and the NB classifier's surprisingly effective performance even when such assumptions are unwarranted. McCallum and Nigam (1998) provide a nice overview of Naive Bayes classifiers, particularly as they apply to text classification.

2.2 Distance and Similarity Measures

Initial attempts to extract information on the relatedness of DNA barcodes have resulted in a vast array of distance measures based on the pairwise differences between two barcodes. In keeping with the theory advanced by Hebert et al. (2003), that barcodes for the same

species should be similar, these distances, in some form or another, measure the percentage of dissimilarities between the two barcodes. If the dissimilarity percentage is high, this would be evidence favoring the notion that the two barcodes come from different species. Whereas a low dissimilarity percentage favors the possibility that the two barcodes belong to the same species.

2.2.1 p-distance

A direct measure of the proportion of pairwise dissimilarities is known as the p-distance. If n_d is the number of positions that differ between two barcodes, and n is the total number of positions being compared, then

$$p - distance = \frac{n_d}{n} \tag{2.4}$$

This is the distance measure used in Hebert et al. (2003) to assess the relatedness of congeneric species. From these distance measures, Hebert, Stoeckle, Zemlak, and Francis (2004) put forth thresholds for differentiating between closely related species.

2.2.2 Kimura's Two Parameter Model (K2P)

Another common measure of genetic distance between two sequences of DNA is Kimura's Two Parameter model (K2P). This model utilizes genetic model assumptions as well as the type of difference between the two sequences. For example, the base adenine (A) bonds with thymine (T) in the double helix structure of the DNA molecule. If, on a particular position, barcode i contains an A and barcode j contains a T, the K2P model would give a different distance than it would if barcode j contained a G. The first difference is know as a transversion, and the latter is know as a transition. A transversion is a substitution of a purine (A or G) for a pyrimidine (C or T) or the substitution of a pyrimidine for a purine. Possible transversions are A⇔C, A⇔T, G⇔C, and G⇔T. A transition is the substitution of a purine for another purine (A⇔G) or of a pyrimidine for another pyrimidine (C⇔T).

The K2P distance between barcode i and barcode j is computed by

$$d(i,j) = -\frac{1}{2}ln[(1-2P-Q)\sqrt{1-2Q}] = \frac{1}{2}ln\left[\frac{1}{1-2P-Q}\right] + \frac{1}{4}ln\left[\frac{1}{1-2Q}\right] \tag{2.5}$$

where P and Q are the proportion of differences between the two barcodes due to transitions and transversions, respectively (Kimura 1980).

2.2.3 Maximum Likelihood

The method of neighbor-joining (to be discussed in Section 2.3), which reconstructs a tree of the genetic relatedness of the barcodes, is statistically consistent, meaning the method reconstructs a tree of the true genetic relatedness of the barcodes, if the true pairwise distances are used (Saitou and Nei 1987 amended by Studier and Keppler 1988). As a practical matter, however, all distance estimates are subject to error. In the two methods mentioned above, pairwise distances are independently estimated and subject to error, which gets accentuated when estimating all $n(n-1)/2$ pairwise distances between n barcode sequences independently. Tamura, Nei, and Kumar (2004) point out that the standard errors of the estimates obtained by independent estimation tend to be rather large unless very long sequences are used. They propose a method of simultaneous estimation based on the maximum likelihood principle that dramatically reduces the standard errors and improves the accuracy of the neighbor-joining tree.

Using the distance

$$d(i,j) = 4(g_A g_G k_1 + g_T g_C k_2 + g_R g_Y)b_{ij} \tag{2.6}$$

where g_A, g_T, g_C, g_G each represent the frequencies of the nucleotides A, T, C, and G. Maximum likelihood estimates of $k_1 = a_{1ij}/b_{ij}$, $k_2 = a_{2ij}/b_{ij}$, and b_{ij} are sought for a specified likelihood function (Tamura et al. 2004), where a_{1ij} is the number of transitions between purines for sequences i and j, a_{2ij} is the number of transitions between pyrimidines for sequences i and j, and b_{ij} is the number of transversions for sequence i and j. The quantities $g_R = g_A + g_G$ and $g_Y = g_C + g_T$ just provide total purine and pyrimidine frequencies,

respectively. The form of the log likelihood function is given by

$$L_{ij} = \widehat{P}_{1ij}\ln(P_{1ij}) + \widehat{P}_{2ij}\ln(P_{2ij}) + \widehat{Q}_{ij}\ln(Q_{ij}) + (1 - \widehat{P}_{1ij} - \widehat{P}_{2ij} - \widehat{Q}_{ij})\ln(1 - P_{1ij} - P_{2ij} - Q_{ij}) \tag{2.7}$$

where $\widehat{P}_{1ij}$ is the observed proportion of transitional differences of purines, $\widehat{P}_{2ij}$ is the observed proportion of transitional differences of pyrimidines, and $\widehat{Q}_{ij}$ is the observed proportion of transversions. P_{1ij}, P_{2ij}, and Q_{ij} are the theoretical values of $\widehat{P}_{1ij}$, $\widehat{P}_{2ij}$, and $\widehat{Q}_{ij}$ which are given by

$$P_{1ij} = \frac{2g_A g_G}{g_R} g_R - exp[-2(g_R k_1 + g_Y)b_{ij}] + g_Y exp(-2b_{ij}) \tag{2.8}$$

$$P_{2ij} = \frac{2g_T g_C}{g_Y} g_Y - exp[-2(g_Y k_2 + g_R)b_{ij}] + g_R exp(-2b_{ij}) \tag{2.9}$$

$$Q_{ij} = 2g_R g_Y[1 - exp(-2b_{ij})] \tag{2.10}$$

The goal here is to find maximum likelihood estimates of k_1, k_2, and b_{ij}. The estimates are shown to be asymptotically unbiased (Tamura et al. 2004), but the computation time of equation (2.6) is much longer than the computation time for equations (2.4) and (2.5) (Frézal and Leblois 2008).

2.3 Neighbor-Joining Trees

One of the most widely used methods of classification of novel barcode sequences based on a database of reference barcode sequences is that of the neighbor-joining method (Saitou and Nei 1987 amended by Studier and Keppler 1988). This method can use a large variety of distance measures, such as those discussed in Section 2.2, to construct a phylogenetic tree depicting the evolutionary relatedness of the novel barcode with the reference barcodes. This is an agglomerative, or bottom-up, clustering method that starts with the branches of the tree equal to the number of reference barcodes, plus the novel barcode and uses Algorithm

1 to join the most similar pairs (Gascuel and Steel 2006). The algorithm is repeated until all of the barcodes have been joined and the branches have been reduced to a single node. The resulting tree is a diagram reflecting the relatedness of all the barcodes. In a barcode reference data set, the species information is known for each barcode. Neighbor-joining will ignore speices information in the reference data set and group barcodes based entirely on the distance measures used.

Algorithm 1 Neighbor-Joining Tree

Let the distance between any taxon pair i and j be denoted as $d(i,j)$. For current distance matrix D with elements $d(i,j)$.

1: Calculate the matrix Q where

$$Q(i,j) = (r-1)d(i,j) - \sum_{k=1}^{r} d(i,k) - \sum_{k=1}^{r} d(j,k) \tag{2.11}$$

and where r is the current number of taxa.
2: Find the pair of taxa in Q with the lowest value and create a node on the tree that joins these two taxa.
3: Calculate the distance of the two paired taxa f and g to the new node u using

$$d(f,u) = \frac{1}{2}d(f,g) + \frac{2}{2(r-2)}\left[\sum_{k=1}^{r} d(f,k) - \sum_{k=1}^{r} d(g,k)\right]. \tag{2.12}$$

4: Calculate the distance of all other taxa k to the new node u using

$$d(u,k) = \frac{1}{2}[d(f,k) - d(f,u)] + \frac{1}{2}[d(g,k) - d(g,u)]. \tag{2.13}$$

5: Repeat steps (1)-(4) treating the pair of joined neighbors as a new taxon and using the matrix of distances computed in step (4) as the original distance matrix in step (1).

2.4 Basic Local Alignment Search Tool (BLAST)

Because DNA sequences are often stored in very large databases, possibly containing hundreds of thousands of catalogued sequences, quick and efficient searches of these data banks

are useful. Altschul, Gish, Miller, Myers, and Lipman 1990 propose a basic local alignment search tool known as BLAST to return the DNA sequences from a data bank that are the most similar to a query sequence. This is done by calculating a measure of local similarity for any two sequences known as the maximal segment pair (MSP) score. Local similarity measures use only highly conserved regions of the DNA sequences rather than the entire sequence as with global similarity measures. Highly conserved regions of DNA are regions in a DNA sequence that experience almost no mutation and remain largely unchanged as the DNA passes from parent to offspring. It is widely believed that mutation in highly conserved regions may lead to a non-viable life form or have severe consequences for the organism, but the reasons for being conserved are not currently well known, and there are some exceptions to this widely held notion (Ahituv, Zhu, Visel, Holt, Afzal, Pennacchio, and Rubin 2007). While various choices of scores are possible, typically the MSP is computed by giving a +5 score to identical base-pairs and a -4 score to mismatches. For two aligned contiguous segments of base-pairs of equal length in a highly conserved region, the similarity score is the sum of the similarity values for each position compared in the segment. The MSP then is just the highest scoring pair of identical length segments from 2 sequences: one of which is the query sequence and the other is a sequence from the data bank.

Results from Karlin and Altschul 1990 allow for estimation of the highest MSP score, say T, for which chance similarities are likely. The search can then minimize the time that it spends on segments that are not likely to exceed this score in terms of their similarity to the query sequence. In practice, a "word," or w-mer, is a segment with fixed length w, and the BLAST search can find segments that contain words with MSP scores of at least T. Focus is then limited to these segments in that they represent significant biological relationships as opposed to chance similarities.

An interesting feature of the BLAST search is how DNA data is stored. From a test barcode, a list of all the contiguous w-mers, or word list, is compiled. Thus, a query sequence of length n will have a $n - w + 1$ words for its word list. For these data, 4 nucleotides are

compressed into a single byte, and Altschul, Gish, Miller, Myers, and Lipman 1990 point out that, if $w \geq 11$, each "hit", a word with MSP $\geq T$, must contain an 8-mer hit that lies on a byte boundary. Thus, the database can be scanned byte-wise which dramatically increases the speed of the search.

Gains in computational time are enhanced with a filter that removes noisy words from the query sequence. This is done by computing, *a priori*, frequencies of all 8-tuple words ($w = 8$) from the DNA data bank. The words that occurr much more frequently than is expected by chance are stored and used to filter out the remaining "noisy" words produced by the query sequence. A BLAST search performed with $w = 12$ can scan about 2×10^6 bases/sec (Altschul, Gish, Miller, Myers, and Lipman 1990).

2.5 Classification with Current Methods

The Barcode Of Life Data System (BOLD 2009) utilizes distance measures together with the neighbor-joining method in order to provide rapid classifications of new barcodes to the species in the BOLD data bank. Ratnasingham and Hebert (2007) provide a detailed overview of the BOLD system from how barcodes are stored and accessed to the classification of new barcodes. See also Kelly, Sarkar, Eernisse, and DeSalle 2006 and Frézal and Leblois 2008. BOLD uses a basic local alignment search tool (BLAST; Altschul, Gish, Miller, Myers, and Lipman 1990) of the barcodes in the data bank in order to begin the process of classification. By default, the search uses a reference data bank which consists of only "verified" barcodes. In order for a species to qualify for inclusion into the verified data bank, it must be represented by at least 3 organisms, each with barcode lengths of at least 500 positions. Also, it must have less than 2% within-species variation. The verified databank consists of around 166,000 barcodes representing about 14,000 unique species (October 2008). The user may request a larger BLAST search to include unverified barcodes as well. The unverified barcodes are barcodes that have at least 500 positions, but may consist of only one representative from a given species and may have more than 2% within-species

variation. By including the unverified barcodes (that contain species specific identification), the databank consists of about 412,000 barcodes representing roughly 38,000 species.

Classification of a new barcode using the BOLD system proceeds as follows. First, the BLAST linear search of the chosen data bank is implemented. This search returns the top 100 matching barcodes in terms of similar "features" between the new barcode and the barcodes in the BOLD system data bank. These features are typically smaller subsequences of the sequence, also called known as "words." The BLAST search will examine the frequency of words in the DNA sequence to be classified and then return the 100 barcodes from the data bank with the most similar word frequencies. For example, in a barcode consisting of the nucleotides A, T, C, and G, there are $4^8 = 65536$ possible 8-digit words. The BLAST search would find the barcodes in the data bank that have the most similar word counts as the barcode to be classified. The pairwise K2P distances are then computed for all 101 DNA barcodes. Next, the relationship between the new barcode and the top matches is assessed by using the neighbor-joining method to reconstruct a phylogenetic tree made up of the top 100 matches together with the new barcode. The new barcode is then classified as belonging to the species of its closest neighbor in the tree, regardless of the distance between them (Frézal and Leblois 2008).

Koski and Goulding (2001) point out that, while this process of classification is fast, it is prone to high rates of false matches because it will classify the new barcode to its closest match regardless of the genetic distance between the two. Also, the probability that the barcode actually belongs to the species to which it was classified is not provided. Rather the percent of bases that match the new barcode is provided for the top matches. This percentage of similarity lacks solid probabilistic interpretation at the species level and, as Ferguson (2002) points out, is somewhat unreliable.

Regarding the distance measures discussed in Section 2.2, DeSalle (2006) points out that character information is lost because, when these distances are computed, all character-based information is erased. Even more troubling is the work of Meyer and Paulay (2005) which

demonstrates that the supposedly well-separated "gap" between within-species variation and between-species variation, upon which the efficacy of distance measures is predicated, may not be so well-separated when comprehensive data sets are considered. They argue that using a reference data set that contains just a few observations per species (1-2 individuals) severely underestimates the within-species variation that would otherwise be seen were a larger number of observations per species to be used. Their conclusion is that there is more overlap in the two types of variation than was previously supposed by Hebert et al. (2003), which leaves the future of distance measures, based on thresholds from this gap, in question. Frézal and Leblois (2008) go on to point out that the accuracy of these distance measures heavily depends on the number of organisms per species present in the reference data set.

2.6 Unanswered Questions

There are several other nagging questions that current methodologies do not address. First, none of them allow for reasonable expressions of how likely the new barcode is to belong to the species to which it is classified. From the neighbor-joining tree method, the new barcode is simply assigned to the species of its nearest neighbor. Such dichotomous (yes/no) classifications leave no room for truly assessing proper classification. If a probability of belonging to that particular species were to be reported, it would aid in species discovery as well as provide an indication of a possible false positive classification.

Next, current methods do not satisfactorily address the issue of species discovery. As mentioned in Section 2.3, the barcode in question is simply classified to the species of its nearest neighbor, regardless of the distance between them. This ensures that if the barcode to be classified belongs to a species not contained in the training data set, it will go undetected (Kelly et al. 2006). This is not only a problem for discovering new species but also for classifying known rare species that may not be represented in the reference data.

Also, questions regarding how much of the barcode is really necessary for proper classification go unanswered. A somewhat arbitrary minimum of 500 base positions per sequence

is required for inclusion into the BOLD data base while the sequence to be classified is required to have at least 300 base positions (Ratnasingham and Hebert 2007), but little justification as to these particular lengths is provided. For a sequence 500 positions long, there are $4^{500} \approx 1.07 \times 10^{301}$ unique barcodes. Surely, for the estimated 15 million unique species on earth (Hammond 1992), there is room for reduction. Consider a barcode with only twelve positions. The possible number of unique barcodes is $4^{12} > 16$ million. While many closely related species will have strong similarities in their barcodes, and such a severe reduction is not likely, it does seem likely that large reductions should be possible. Such reductions have the potential to decrease computation time in analyses and decrease the amount of work that goes in to extracting and storing these DNA barcodes. More important, however, are the archived DNA samples that would be open to classification if shorter sequences could be properly identified. Another related concern is that modern sequencing methods that provide high-throughput sequences, such as Solexa and 454 sequencers, often require a trade-off between many sequences of short length or fewer longer sequences. These sequencers have the capacity to churn out millions of high-quality base readings in about 10 hours (Roche 2009). Establishing necessary barcode length could mean huge gains in terms of how many barcodes could be produced and catalogued.

Finally, comprehensive databases that contain at least three organisms from all of the species within a particular genus are relatively rare. The requirement for a species to be represented by at least three organisms in the reference data is in place to ensure proper representation of within-species variation, something the distance measure approaches heavily depend upon due to their threshold based assumptions. While understanding how within-species variation compares to between-species variation is important, methods that rely so heavily upon them will be difficult to fully assess and exploit until there are many more comprehensive databases available. Until then, methods that are not as influenced by the overlap between these two types of variation may prove advantageous on the currently available, less comprehensive data sets that may have several species represented by only one or

two organisms.

2.7 Preliminary Discussion

Chapter 3 defines a newly proposed method of classifying unknown barcodes that will allow the issues discussed in the previous section to be addressed. Namely, those issues are: calculating probabilities for species assignments, determining necessary barcode length, aiding in species discovery, and minimizing the reliance of the method on the genetic "gap" and genetic model assumptions.

At the heart of the proposed method is the sequential application of Bayes' rule to compute posterior probabilities of a new barcode belonging to any of the species in the reference data set. Equation 3.1 gives the form of this sequential calculation, and Section 3.1 describes its construction and implementation. These probabilities are crucial for assessing the resulting classification and, unlike current methods, have a direct species-level interpretation.

The sequential nature of the proposed method allows posterior probabilities to be calculated at each position of the barcode. Once the posterior probability gets sufficiently close to 1 for any species in the reference data set, the calculations may be terminated and the species assignment made. By noting the number of positions required for the classification to be made, one may begin to assess the issue of necessary barcode length. Section 3.7 discusses in more detail rules for terminating the calculation, and Section 4.2 presents the average number of positions required for classification of barcodes from five different data sets. We find that, in most cases, the classification can reliably be made using less than the entire barcode.

Because the important issue of species discovery is something lacking with current methods, Section 3.8 provides some discussion on how the proposed method can be used to facilitate species discovery, and an example is given in Section 3.10. We point out that a plot of the posterior probabilities versus barcode position can indicate whether the new barcode

to be classified belongs to any of the species in the reference data set. A very "noisy plot" of the posteriors that frequently changes species with the highest posterior probability indicates the new barcode does not belong to any of the species in the reference data set. These "noisy" plots may be of particular interest to a biologist in that they not only indicate possible new species but also which species in the reference data set the new barcodes are most similar to and at which positions.

It is worth remarking that, while the proposed method was developed with DNA barcoding in mind, it is quite flexible and can be extended to many other applications. Section 6.2.1 provides some discussion of other areas that may benefit from such a method. This is made possible in part by the proposed methods lack of dependence upon the genetic "gap" or upon genetic model assumptions. To be sure, the proposed method can be viewed as a general approach to classifying high dimensional data to various groups. The success of this method relies less upon underlying assumptions of the data, and more upon how many independent pieces of information are available for each observation and how discriminatory those pieces of information are.

Chapter 3

Proposed Method for Classification

The method proposed here is aimed at answering the questions left open by current methods that were discussed in Sections 2.6 and 2.7. Specifically, this method will compute the assignment probabilities for each species in the reference data set (Section 3.2), provide some indication as to when the barcode to be classified does not belong to any species in the reference data set aiding in species discovery (Section 3.8), allow the question of necessary barcode length to be addressed by implementing simple stopping rules (Section 3.7), and reduce the dependence on genetic "gap" and genetic model assumptions. It should also be noted that, while the proposed method is presented in the context of DNA barcoding, it is general enough to be applied to other situations in which an object is to be classified based on high-dimensional data as discussed in Section 6.2.1.

3.1 Constructing the Conditional Probabilities

A truncated data set containing DNA barcodes for 20 organisms from four different species (1, 2, 3, and 4) is shown in Table 3.1. If a new barcode is to be classified as belonging to one of these four species, it is clear that the classification should account for the ordered nature of the data, in that various orderings of the values A, T, C, and G lead to different species.

The proposed method of classification incorporates the ordering of the barcode and involves an application of Bayes' rule at each of the barcode's positions. This method computes

Species	Truncated Barcode
1	- - - - - - - - - - - - - - - - N C T G
1	C C G G C A T A G T A G G C A C T G
1	C C G G C A T A G T T G G N A C T G
1	C T G G C A T A G T A G G T A N T N
1	C C G G C A T A G T A G G C A C T G
1	- - - - - - - - - - - - - - - - - -
1	C C G G C A T A G T T G G C A C T G
1	C T G G C A T A G T A G G T A C T G
2	C T G G C A T A G T C G G A A C C G
2	C T G G C A T A G T C G G A N C C G
2	C T G G C A T A G T C G G A N C C G
3	C C G G C A T A G T A G G A A C A G
3	C T G G C A T A G T A G G A A C A G
3	- - G G C A T A G T A G G A A C A G
3	- - - - - - - - - - - - - - - - - -
3	C C G G C A T A G T A G G A A C A G
4	- - G G A A T A G T A G G T A C C G
4	C C G G A A T A G T A G G T A C C G
4	C C G G A A T A G T A G G T A C C G
4	C C G G A A T A G T A G G T A C C G

Table 3.1: *Truncated Barcode Data. Only the first 18 positions of the barcode are shown here. Typical barcodes range in length from 255 positions, to 700 positions. When a barcode contains a - or an N, the observed nucleotide base is missing at that position*

the probability that the new barcode belongs to each of the species at each position. If S_l represents the event that the barcode belongs to species l, $l = 1, \ldots, s$ and s is the number of species in the reference data set, prior probabilities can be specified for each species, $P(S_l)$. Let $x^{(1)}, \ldots, x^{(p)}$ represent the observed nucleotides along a barcode sequence with p positions, and S_l is the event the barcode belongs to species l. Then estimates of the true conditional probabilities of the nucleotides A, T, C, and G for each species, $P(x^{(j)}|S_l)$, at each position j can be computed from the observed proportions of nucleotides for each species at each position in a reference data set. For example, Table 3.2 gives the conditional probabilities of the first three positions for the values A, T, C, and G for species 2. Once all the conditional probabilities are constructed and prior probabilities are specified

Position 1	Position 2	Position 3	...
$P(A\|S_2) = 0$	$P(A\|S_2) = 0$	$P(A\|S_2) = 0$	...
$P(T\|S_2) = 0$	$P(T\|S_2) = \frac{3}{3}$	$P(T\|S_2) = 0$	...
$P(C\|S_2) = \frac{3}{3}$	$P(C\|S_2) = 0$	$P(C\|S_2) = 0$	...
$P(G\|S_2) = 0$	$P(G\|S_2) = 0$	$P(G\|S_2) = \frac{3}{3}$	...

Table 3.2: *Conditional probabilities for the first 3 positions of species S_2.*

for each species, posterior calculations are done sequentially at each position according to the following equation:

$$P(S_l|x^{(j)}) = \frac{P(S_l|x^{(j-1)})P(x^{(j)}|S_l)}{\sum_{l=1}^{s} P(S_l|x^{(j-1)})P(x^{(j)}|S_l)} \tag{3.1}$$

for $j = 1, \ldots, p$, where p is the number of positions on the barcode, $P(S_l|x^{(j)})$ is the posterior probability that the barcode belongs to species l after observing the nucleotide at position j, and $P(S_l|x^{(j-1)})$ is the posterior probability that the barcode belongs to species l after observing the nucleotide at position $j-1$ which also serves as the prior probability in the posterior calculation for position j. The initial value of $P(S_l|x^{(0)})$ is equal to the specified prior probability of belonging to species l, $P(S_l)$. The posterior probability that the barcode belongs to species l then becomes the prior probability for the calculation in the next position. Hence, this method provides a sequential calculation that the barcode belongs to any of the s species in the data set which is outlined in Algorithm 2. The effect of using the posterior probability from the previous position for the prior probability on the next position is discussed in Section 3.2.

The method proposed here can be thougth of as a sequential version of the Naive Bayes classifier discussed in Section 2.1.1. The major differences are that we will compute full posterior probabilities at each nucleotide position while traditional Naive Bayes classifiers compute a probaility proportional to the posterior over all of the positions. Reporting the actual posterior probability is useful in assesing the uncertainty of each classification while computing it at each nucleotide position, as we propose, yields a picture of the "movement" of the posterior probabilities of each species over all the nucleotide postions. This aids in

Algorithm 2 Sequential Posterior Calculations

Starting with the first position:

1: Specify prior probabilities $P(S_l)$ for all $l = 1 \ldots, s$, where s is the number of species in the reference data set.

2: Compute the conditional probabilities $P(x|S_l)$ of the bases A, T, C, and G, for all $l = 1 \ldots, s$, at the current position from the reference data set.

3: Compute the posterior probability that the new barcode belongs to species l for all $l = 1 \ldots, s$ using $P(S_l|x) = P(S_l|x)P(x|S_l)/\sum_{l=1}^{s} P(S_l|x)P(x|S_l)$ where $P(x|S_l)$ is determined by the base observed in the current position of the new barcode.

4: Using the posterior probability computed in Step 3 as the new prior probability for Step 1, repeat steps 1-3 for each position until the end of the barcode is reached or a stopping rule (discussed in Section 3.7) is invoked.

species discovery and will be discussed in more detail in Sections 3.10 and 4.1.2.

The assumption of conditional independence among the nucleotide positions given species is probably incorrect but in the spirit of Zhang (2004), we expect the proposed method will still provide accurate species-level classifications if the dependencies among nucleotides are evenly distributed over speceis.

The sequential calculations can run until the the end of the barcode is reached or may be terminated early via some kind of stopping rule. The benefit of implementing a stopping rule as well as possible stopping rules to consider for this calculation will be discussed in Section 3.7. Upon reaching the end of the barcode or the stopping rule, the new barcode is then classified as belonging to the species with the highest posterior probability, given that it is not flagged as a potentially new species not appearing in reference data set. The ideas behind species discovery with the proposed method will be discussed in Section 3.8.

3.2 Today's Posterior is Tomorrow's Prior

Lindley 1970 states, "If two pieces of data, x_1, and x_2, arise in sequence, then the distribution posterior to x_1 is prior to x_2 - today's posterior is tomorrow's prior." To be sure, Gelman, Carlin, Stern, and Rubin 2004 remark that one of the advantages to a Bayesian analysis is the ease with which sequential analyses can be performed. They point out that if a

posterior probability has been computed based on previous data, then when a new data point arises "the entire calculation does not need to be redone; rather we use the previous posterior distribution as the new prior distribution." This is the perspective we take with the proposed method of classification of DNA barcode sequences.

Initial prior probabilities can be selected in various ways discussed in Section 3.5. In the classical sense, these prior probabilities represent our belief that the barcode belongs to the various species in the reference data set prior to observing the data. After observing the nucleotide in the first position of the barcode, we may wish to update those prior beliefs by computing posterior probabilities. These reflect our updated belief that the barcode belongs to any of the species in the reference data set given the nucleotide observed in the first position. These updated beliefs are posterior to observing the data in the first position, but they are prior to observing the data in the second position. Thus, we move sequentially through the barcode treating the posterior probabilities calculated at each position as the prior probabilities for the calculation in the next position.

If $x^{(1)}, \ldots, x^{(p)}$ represent the observed nucleotides along a barcode sequence with p positions, and S_l is the event the barcode belongs to species l, then the goal of the proposed method's calculation is to computed $P(S_l|x^{(1)}, \ldots, x^{(p)})$.

Theorem 1. *Let $x^{(1)}, \ldots, x^{(p)}$ be p independent observations that arise in sequence. Suppose that $P(S_l)$ represents the prior probability that the sequence of observations belongs to group l and that the sequence of observations are also conditionaly independent, given group l . Suppose further that the conditional probabilities $P_1(x^{(1)}|S_l), \ldots, P_p(x^{(p)}|S_l)$ are known.*

Then, using the posterior probability from position j, $P(S_l|x^{(1)}, \ldots, x^{(j)})$ as the prior for computing the posterior at position $j+1$, $P(S_l|x^{(1)}, \ldots, x^{(j+1)})$, for $j = 1, \ldots, p$, in equation (3.1) results in computing $P(S_l|x^{(1)}, \ldots, x^{(p)})$.

Proof. First notice that because the observations $x^{(1)} \ldots x^{(p)}$ are marginally independent

and conditionally independent given group l, we have

$$P(S_l|x^{(1)},\ldots,x^{(p)}) = \frac{P(S_l)P_1(x^{(1)}|S_l)P_2(x^{(2)}|S_l)\cdots P_p(x^{(p)}|S_l)}{P_1(x^{(1)})P_2(x^{(2)})\cdots P_p(x^{(p)})}. \tag{3.2}$$

Now, using the prior probability $P(S_l)$, the posterior probability for position 1 is

$$P(S_l|x^{(1)}) = \frac{P(S_l)P_1(x^{(1)}|S_l)}{P_1(x^{(1)})}. \tag{3.3}$$

Using the RHS of equation (3.3) as the prior for calculating the posterior in position 2 gives

$$\frac{\frac{P(S_l)P_1(x^{(1)}|S_l)}{P_1(x^{(1)})}P_2(x^{(2)}|S_l)}{P_2(x^{(2)})} = \frac{P(S_l)P_1(x^{(1)}|S_l)P_2(x^{(2)}|S_l)}{P_1(x^{(1)})P_2(x^{(2)})}. \tag{3.4}$$

Using the RHS of equation (3.4) as the prior for calculating the posterior in position 3 gives

$$\frac{\frac{P(S_l)P_1(x^{(1)}|S_l)P_2(x^{(2)}|S_l)}{P_1(x^{(1)})P_2(x^{(2)})}P_3(x^{(3)}|S_l)}{P_3(x^{(3)})} = \frac{P(S_l)P_1(x^{(1)}|S_l)P_2(x^{(2)}|S_l)P_3(x^{(3)}|S_l)}{P_1(x^{(1)})P_2(x^{(2)})P_3(x^{(3)})}. \tag{3.5}$$

Continuing on in this fashion through the p^{th} position yields

$$\frac{P(S_l)P_1(x^{(1)}|S_l)P_2(x^{(2)}|S_l)\cdots P_p(x^{(p)}|S_l)}{P_1(x^{(1)})P_2(x^{(2)})\cdots P_p(x^{(p)})} = P(S_l|x^{(1)},\ldots,x^{(p)}) \tag{3.6}$$

which is the desired result. ■

Theorem 1 states that, by using the computed posterior probability at each position as the new prior probability for the next position, the proposed method is in fact computing the posterior probability $P(S_l|x^{(1)},\ldots,x^{(p)})$ assuming the positions are independent. Notice that, in Theorem 1, the conditional probabilities at each position are subscripted, indicating that the conditional probability distributions for $P_1(x^{(1)}|S_l),\ldots,P_p(x^{(p)}|S_l)$ need not be the same. Indeed, $x^{(1)},\ldots,x^{(p)}$ could be a combination of discrete and continuous variables. The assumption of independence seems to be adequate and will be discussed further in Section 4.3.

3.3 Monotonicity of the Posteriors

What is the asymptotic behavior of the posterior probabilities computed using equation (3.1)? The answer to that question will depend, in large measure, on the varying amounts

of within- and amoung-species variability present in the data set being examined. However, under ideal circumstances put forth in Theorem 2, we prove that the posterior probability for species l will be monotone increasing if, and only if, the new barcode to be classified belongs to species l.

Theorem 2. *Let $x^{(1)}, \ldots, x^{(p)}$ be p independent nucleotides that, in sequence, constitute a DNA barcode. Let $P(S_l)$ represent the prior probability that the barcode belongs to species l. Suppose that the nucleotides are conditionaly independent, given species l and that the conditional probabilities, $P(x^{(1)}|S_l), \ldots, P(x^{(p)}|S_l)$, are known. Suppose further that the barcodes within each species are identical, and the species in the reference data set, R, represent all possible species and have unique barcodes.*

Then the posterior probability of belonging to species l, $P(S_l|x^{(1)}, \ldots, x^{(p)})$, will be monotone increasing iff the new barcode T to be classified belongs to species l.

Proof. First suppose that the new barcode belongs to species l and that $P(S_l|x^{(1)}, \ldots, x^{(p)})$ is strictly decreasing. This means

$$P(S_l|x^{(1)}, \ldots, x^{(p)}) < P(S_l|x^{(1)}, \ldots, x^{(p-1)}) \tag{3.7}$$

which can be rewritten as

$$P(S_l|x^{(1)}, \ldots, x^{(p)})/P(S_l|x^{(1)}, \ldots, x^{(p-1)}) < 1. \tag{3.8}$$

Because $x^{(1)}, \ldots, x^{(p)}$ are both margianlly and conditionally independent given species l, we can expand both numerator and denominator on the LHS as in equation (3.2). Thus we obtain

$$\frac{P(S_l)P(x^{(1)}|S_l)\ldots P(x^{(p-1)}|S_l)P(x^{(p)}|S_l)}{P(x^{(1)})P(x^{(2)})\ldots P(x^{(p-1)})P(x^{(p)})} \Big/ \frac{P(S_l)P(x^{(1)}|S_l)\ldots P(x^{(p-1)}|S_l)}{P(x^{(1)})P(x^{(2)})\ldots P(x^{(p-1)})} < 1. \tag{3.9}$$

On the LHS, the denominator cancels all but the pth terms in the numerator leaving

$$P(x^{(p)}|S_l)/P(x^{(p)}) < 1 \tag{3.10}$$

which can be rewritten as

$$P(x^{(p)}|S_l) < P(x^{(p)}). \tag{3.11}$$

Finally, the RHS can be expanded using the Law of Total Probability to give

$$P(x^{(p)}|S_l) < P(S_1)P(x^{(p)}|S_1) + \cdots + P(S_s)P(x^{(p)}|S_s). \tag{3.12}$$

But $P(x^{(p)}|S_l)$ contains all the mass at position p for species l because the barcodes within a species are identical by assumption. This means the LHS of equation (3.12) is 1 making the RHS impossible because it violates an axiom of probability. By contradiction we conclude that $P(S_l|x^{(1)}, \ldots, x^{(p)})$ is monotone increasing.

Next, suppose that the new barcode does not belong to species l. Because the species in R have unique barcodes by assumption, and the barcodes within a speices are identical, also by assumption, there exists a position p, for which $P(x^{(p)}|S_l) = 0$ which implies $P(S_l|x^{(1)}, \ldots, x^{(p)}) < P(S_l|x^{(1)}, \ldots, x^{(p-1)})$. Therefore the posterior probability of species l, over all of the positions, is not monotone increasing.

■

3.4 Adjusting the Conditional Probabilities

By constructing the conditional probabilities of observing a particular nucleotide at position j given the barcode belongs to species l, as in Section 3.1, it is clear that, if the nucleotide in position j of the new barcode does not occur in the reference data set for any of the observations in a species at position j, the resulting posterior probability of the barcode belonging to that species computed by equation (3.1) will be zero. Furthermore, all subsequent posterior probabilities calculated for that species will also be zero. For example, suppose a new barcode came from species 1 and looked like $ACGGCATAGTTGGNACTG$. Compare this to the other barcodes from species 1 in Table 3.1. It is identical to several of the barcodes for species 1 with the exception of the leading A. In the data set, species 1 only had a C or $-$ in position 1. Therefore, the conditional probability used in the calculation

of equation (3.1) is $P(A^{(1)}|S_1) = 0$, and the resulting posterior probability of species 1 is zero. This becomes the prior for the next calculation leading to a posterior of zero on the next position and so on, regardless of the conditional probabilities at subsequent positions. This implies that even though the remaining positions match identically to those of species 1, it will have essentially been removed from consideration, and the new barcode cannot be classified to that species.

In Section 1.2, it was observed that, while barcodes for the same species will be very similar, there may be some small amount of variation making this strict specification of the conditional probability too rigid. A single position containing a value not observed in the data set for a particular species will provide a penalty so severe that the posterior will never recover. To avoid this potential problem, it is recommended that the conditional probabilities be slightly adjusted by assigning all of the conditional probabilities that would be zero some small bit of mass, δ, where $0 < \delta \ll 1$. While the calculation of the posterior probability would certainly reflect that the new barcode contained a value not observed in the data set, it would nevertheless, allow the probability to "recover" if subsequent matches are made or, on the other hand, drive the calculation in equation (3.1) to zero if subsequent matches are not made.

3.4.1 Choosing δ

It is clear from equation (3.1) that the calculated posterior probability will depend on the choice of δ and will lead to different misclassification rates for different choices of δ. Figure 3.1 shows the misclassification rates of a barcode data set for various choices of δ where 10% of the observations were held out and then classified using the remaining 90% of the observations as the reference data set. Vertical lines are given in the plot at $a = 1.5 \times 10^{-6}$ and $b = 3.6 \times 10^{-6}$. It appears that δ between between these two values will provide the smallest misclassification rate of 0.077. It should be noted that these misclassification rates were computed using prior probabilities equal to the observed proportion of each species in

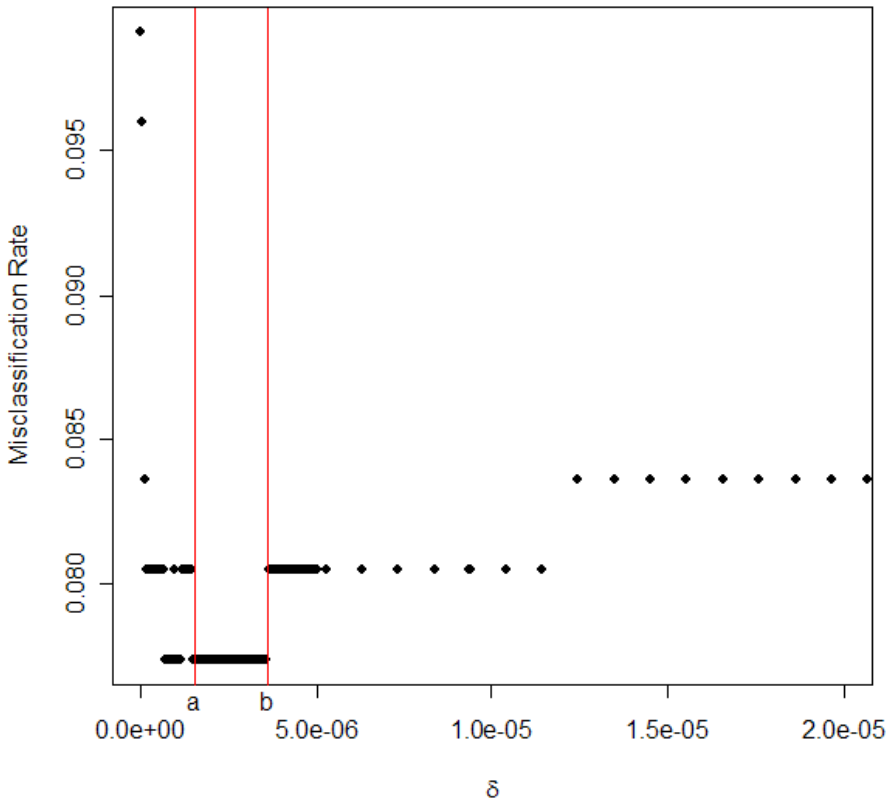

Figure 3.1: *Plotted misclassification rates for various choices of δ on a barcode data set using proportional priors. $a = 1.5 \times 10^{-6}$ and $b = 3.6 \times 10^{-6}$.*

the reference data set. This gives rise to the questions "What exactly does δ represent?" and "How should δ be chosen so that the misclassification rates are, in some sense, optimal?" First, we discuss some initial perspectives on δ as well as answers to both of these questions. We then provide a recommendation for choosing this quantity.

Initial Perspectives on δ.

The δ value that provides an optimum misclassification rate will likely depend on some (if not all or more) of the following characteristics. First, the choice of δ is likely to depend on the amount of within-species variation in the barcodes. For example, if a particular species

has very small within-species variation, then it is less likely for a new barcode to contain a value not accounted for in the conditional probabilities. In this case, a value of δ close to 0 would be appropriate in that this would provide a heavy penalty for discrepancies between the new barcode and those for a given species in the data set. Here, such discrepancies are more likely to be indicative of the barcode belonging to a different species as opposed to random variation in the barcode. On the other hand, if the amount of within-species variation is high for a particular species, then it is more probable for a new barcode to contain values that do not match those of the data set. It might be ideal here to choose a somewhat larger value of δ so that divergences between the new barcode and those of the data set for a particular species are not so heavily penalized. In this situation, the larger δ value reflects our belief that discrepancies could be due more to random variation and less to the notion that the barcode belongs to some other species.

Second, the choice of δ will probably depend on the number of observations per species in the data set. Fewer observations per species lead to more rigid, less established conditional probabilities. In some cases, several species in a reference data set are represented by only one or two observations. In these cases, a more tolerant (larger) value of δ may be appropriate. On the other hand, a smaller value of δ would be more appropriate when there are more observations per species, and the conditional probabilities are well established.

Here we are considering a single value of δ for the entire reference data set. An item for future research, to be discussed in more detail in Section 6.2.1, may be to use several δ values within the reference data set. If, for example, misclassification rates could be improved upon by selecting a species-specific value for δ, each species would then have a unique δ value used in adjusting the zero-valued conditional probabilities.

What does δ represent?

While the initial perspectives of δ discussed above shed some light upon its predicted behavior in various settings, it is the answer to this question that perhaps provides the greatest insight as to appropriately choosing a δ value. Consider for a moment the process of DNA

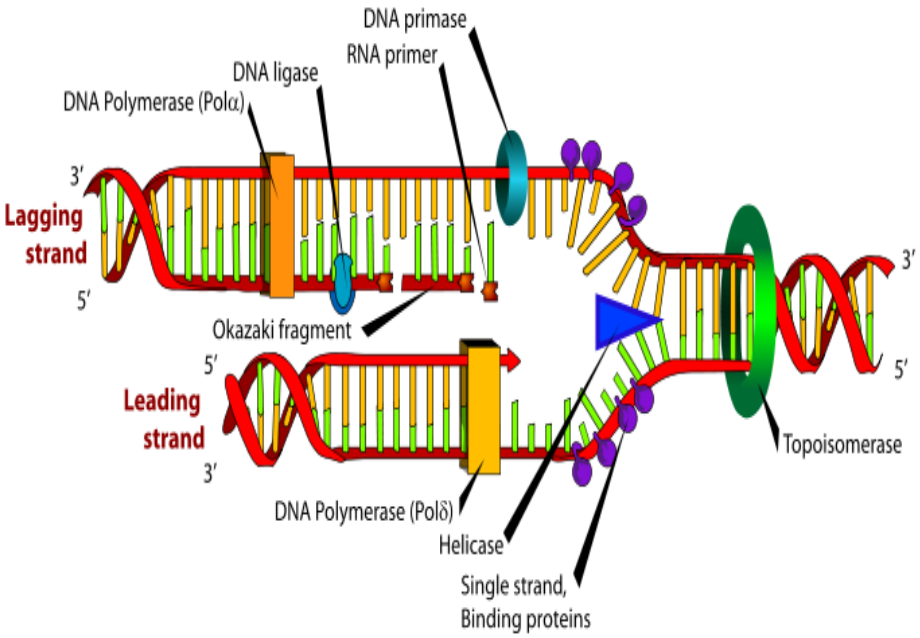

Figure 3.2: *The process of DNA replication.*

replication by which the double-helical strand of DNA can copy itself. This happens when an enzyme known as "helicase" moves along the double-stranded DNA structure breaking the hydrogen bond between the two nucleotide bases. This essentially unzips the DNA structure as depicted in Figure 3.2. Single-stranded binding proteins attach to the unzipped portion of the DNA structure to prevent them from annealing, or bonding back together. This results in two single-stranded DNA structures. Another enzyme, DNA polymerase, then moves along these single-stranded structures removing the binding proteins and pairing the nucleotide base in the single-strand with its complementary base (typically, complementary bases are A⇔T and C⇔G, but other pairings, while rare, are possible). When the process is complete, the two new strands of DNA are "semi-conserved," meaning they are nearly exact copies of the original. For a comprehensive treatment on the topic, see Kornberg and Baker (1992).

Errors in this process, while rare, are known to exist and can lead to mutations where the copied DNA strands are not identical to the original. Substitution of an incorrect

base, for example, is possible and represents a mutation of the original DNA structure. If the process were infallible, then the rigid specification of the conditional probabilities as discussed in Section 3.1 would be reasonable. However, understanding the potential for error in the process of DNA replication indicates the zero-valued conditional probabilities specified in that manner are not really zero. To more appropriately specify those conditional probabilities, we need to assign the zero-valued conditional probabilities some small amount of mass that represents this potential for error. This leads us to a logical interpretation for δ. The value of δ we seek is a measure of the rate at which mutations of the original DNA structure occur. In other words, δ is easily interpreted as the probability of a mutation at any given position.

How Should δ be chosen?

Thinking of δ in this way not only gives it biological relevance but also points to how an appropriate value of δ should be chosen. Because DNA barcodes come from mitochondrial cells, it seems that an estimate of the mutation rate within the mitochondrial genome would prove to be an ideal choice for δ. Denver, Morris, Lynch, Vassilieva, and Thomas (2000) provide an estimate of the probability of mutation at any position within the mitochondrial genome of 9.7×10^{-8}. As technology allows for more sophisticated methods of estimating this probability of mutation, we recommend the value chosen for δ evolve with it. In Table 3.9, we compare misclassification rates based on this mutation rate for using $\delta = 9.7 \times 10^{-8}$ to an arbitrarily chosen δ value of 1.0×10^{-4}. The misclassification rates are very similar for the two choices of δ. Figure 3.3 shows the misclassification rates for five animal data sets where 10% of the observations were randomly selected to be held out and the remaing 90% of the observations were used to constuct the conditional probabilities used in the proposed method for various choices of δ. The plot indicates that, if δ is zero, the misclassification rates increase dramatically, while choices of δ between 0 and 0.01 provide the smallest misclassification rates. As δ increases beyond this interval, the misclassification rates begin to rise. These results also indicated that a mutation rate of 9.7×10^{-8} is within the range

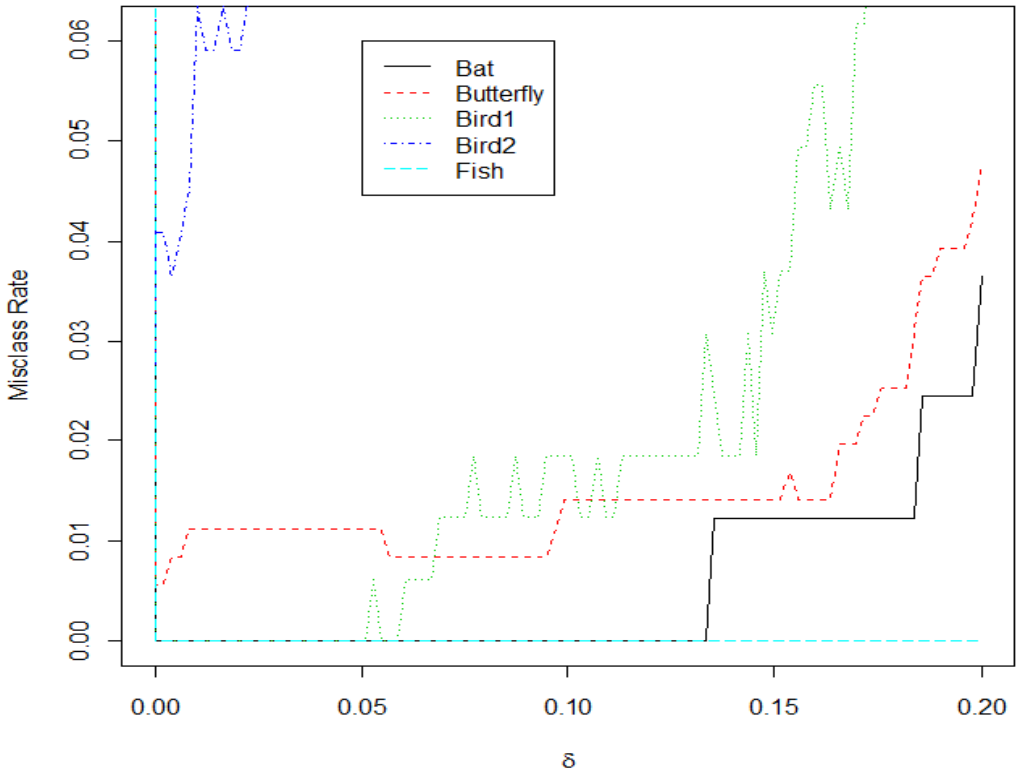

Figure 3.3: *Plotted misclassification rates for various δ values for five animal data sets. Notice that the misclassification rate depends on the choice of δ. Extremely small (close to zero) δ values and large δ values yield the larger misclassification rates while δ values between* 0 *and* 0.01 *yield the smaller misclassification rates.*

of acceptable δ values. Appendix A contains the tabulated δ values together with the corresponding misclassification rates for the five data sets. It is our recommendation that $\delta = 9.7 \times 10^{-8}$ be used to adjust the conditional probabilities. Because the choice of δ directly depends on the technology available to estimate the mutation rate of the mitochondrial genome, it is also our recommendation that, as technology advances in this direction, that the choice of δ advance with it.

3.5 Prior Specification

The prior probability, $P(S_l)$, is the probability the the barcode belongs to species l prior to observing any data. How these prior probabilities are specified is typically a subjective issue, but in the case of DNA barcodes, and in other cases that involve high-dimensional data, one can hope that the data will eventually dominate reasonable priors. Due to the sequential calculation illustrated in Section 3.1, the specified prior probabilities are constantly updated by the data, causing reasonable priors to stabilize over the process and yield similar posterior probabilities near the end of the calculations.

To establish the claim that the misclassification rates will be somewhat robust to the choice of priors, we used the proposed classification method on five different data sets with the following prior probabilities selected:

1. Arbitrary priors generated from a non-informative Dirichlet distribution.
2. Ascending Arbitrary priors. These are the same priors in (1) sorted in ascending order.
3. Descending Arbitrary priors. These are the same priors in (1) sorted in descending order.
4. Equal priors.
5. Data based proportional priors.

For the arbitrary priors in (1)-(3) above, we randomly generated the vector of probabilities $(P(S_1), \ldots, (S_s))$ from a non-informative $Dirichlet(1, 1, \ldots, 1)$ distribution. This is equivalent to randomly picking s values between 0 and 1 such that they sum to 1 and then assigning them at random to the s species. By sorting these as in (2) and (3) above, we essentially end up with different arbitrary priors assigned to each species. Arbitrary priors sorted in ascending order simply sorts the priors obtained by (1) from smallest to largest and assigns them systematically to the s species that are ordered alphabetically. Arbitrary priors sorted in descending order sorts the priors obtained by (1) from largest to smallest

Data Set	s	p	R	T	M_a	M_{a^*}	$M_{a^{**}}$	M_p	M_e
Bat	96	659	756	84	0	0	0	0	0
Bird1	150	690	1300	323	0.002	0.002	0.002	0.002	0.002
Bird2	656	255	2330	259	0.029	0.024	0.026	0.026	0.027
Butterfly	559	255	3839	427	0.007	0.005	0.007	0.006	0.007
Fish	211	255	678	76	0.008	0.008	0.006	0.008	0.006

Table 3.3: *Misclassification Rates for arbitrary priors (M_a), ascending arbitrary priors (M_{a^*}), descending arbitrary priors ($M_{a^{**}}$), data-based proportional priors (M_p), and equal priors (M_e). R and T are approximately the number of barcodes in the reference and test data sets, respectively. The values s and p are the number of species in R and number of positions on the barcode, respectively. In each case $\delta = 1.0 \times 10^{-4}$.*

and assigns them systematically to the s species that are ordered alphabetically. So if the species are sorted alphabetically, with arbitrary priors sorted in ascending order, the first species will be assigned the smallest prior probability while using arbitrary priors sorted in descending order, the same species will be assigned the largest prior probability.

Equal priors simply assigned the value $1/s$ to each of the s species. Priors selected in this fashion indicate that, based on the current information, the barcode is equally likely to belong to any of the s species in the reference data set. This discrete uniform distribution of the priors does not favor any species but relies on the data alone to distinguish between the possible species.

Data-based proportional priors are calculated by finding the prevalence of each species in the reference data set and using that quantity as the prior probability for each species in s. This choice of priors will favor the species that are most abundant in the reference data set while down-weighting species in the reference data set that are represented by just a few organisms.

Tables 3.3 and 3.4 show the 10-fold cross-validated average misclassification rates for the Bat, Bird1, Bird2, Butterfly, and Fish data sets. In all cases the average misclassification rates are very similar regardless of the choice of priors. In fact, for the Bat and Bird1 data sets in Table 3.3 that uses the arbitrary δ value of 1.0×10^{-4}, the average misclassification

Data Set	s	p	R	T	M_a	M_{a^*}	$M_{a^{**}}$	M_p	M_e
Bat	96	659	756	84	0.001	0.001	0.001	0.001	0.001
Bird1	150	690	1300	323	0.003	0.003	0.003	0.002	0.003
Bird2	656	255	2330	259	0.025	0.020	0.023	0.021	0.023
Butterfly	559	255	3839	427	0.005	0.004	0.006	0.005	0.005
Fish	211	255	678	76	0.008	0.008	0.006	0.008	0.006

Table 3.4: *Misclassification Rates for arbitrary priors (M_a), ascending arbitrary priors (M_{a^*}), descending arbitrary priors ($M_{a^{**}}$), Data based proportional priors (M_p), and equal priors (M_e). R and T are approximately the number of barcodes in the reference and test data sets, respectively. The values s and p are the number of species in R and number of positions on the barcode, respectively. In each case $\delta = 9.7 \times 10^{-8}$.*

rates are the same regardless of the choice of priors. While the Fish, Bird2, and Butterfly data sets do not yield identical misclassification rates, they do yield misclassification rates that are very similar.

In Table 3.4, which uses that mutation rate of 9.7×10^{-8} for δ, the misclassification rates for the various choices of priors are identical for only the Bat data set and yet, for the other data sets, the misclassification rates are very similar for the various choices of priors.

Because the misclassification rates are similar for the various choices of priors examined above, it is our recommendation that equal priors be used unless there is strong *a priori* evidence indicating that the barcode is not likely to belong to one or more species in the reference data set. In cases were there is strong *a priori* evidence that the barcode is not likely to belong to one or more species in the reference data set, a practitioner may subjectively choose the prior probabilities. If the prior probabilities are chosen correctly, the proposed method of classification may make the proper species assignment using fewer barcode positions. Based on the results in Table 3.3, if the prior probability assignments are incorrect, they should, nevertheless, eventually be dominated by the data, and the correct classification may occur at a latter barcode position. An example of this is given in Section 3.9.

3.6 Missing Data

The methods used to retrieve DNA barcodes, while very good, are not infallible. Occasionally, the process will not be able to identify the base at a particular location, or more frequently, the process of aligning the barcodes, so that their positions match, yields positions for which no bases have been observed. Each of these situations leads to missing data at various positions along the barcode and can be seen in the truncated barcodes given in Table 3.1. Notice that the first barcode in species 1 contains a series of "-"s. These dashes are a result of software that is commonly used to align the barcodes. This means that the first position of the first organism in species 1 really started at position 15 when viewed with the other organisms in species 1. This software induced "-" is seen in another organism in species 1 as well as in two organisms in species 3 and one organism in species 4.

A quick scan over the rest of the barcodes reveals an occasional character "N" which is clearly not one of the four nucleotide bases we have previously mentioned. In fact, this character is not a nucleotide base at all but, rather, standard notation to indicate that, during the process of DNA extraction, it was unclear which of the four nucleotide bases (A, T, C, or G) should be in that position.

Whether the data is missing due to software induced alignments or to the presence of an ambiguous base during extraction, it will have an impact on whether the method can properly classify the barcode. Note that data missing due to these causes can show up in the reference data sets we use to create the conditional probabilities discussed in Section 3.1, as well as in the barcodes to be classified via equation (3.1). Each case will need to be examined separately.

3.6.1 Missing Data in the Reference Data Set

The reference data set contains barcodes as well as information about the species the barcodes belong to. This means that, if a base is missing at a particular position, it may be possible to use the barcodes from other organisms within that species to tell us something

about what base should have been observed. With this in mind, let us consider a few alternatives to dealing with missing data within the reference data set.

Impute the Missing Data Using a Majority Rule

One approach to dealing with the missing data is to allow the missing data to be imputed by the observed data using a "majority rule." Because the species to which the barcodes in the reference data belong is known, each position for a particular species is easily examined. In keeping with the theory that the barcodes will be very similar within a species, it is possible to fill in a missing data value with the most frequently occurring base in that position for other organisms of the same species. In this way, most of the missing data values can be accounted for as illustrated by comparing the original data set in Table 3.1 with the majority rule imputed data set in Table 3.5. The advantages to this approach are that it can be easily implemented, and it is in keeping with the overall theory of DNA barcoding. Disadvantages include altering the observed proportions of bases for a given species at a given position after the imputation and reducing the within-species variation.

Impute the Missing Data Using a Proportional Allocation

An alternative to imputing the missing data with the majority rule is to impute the data by randomly selecting the bases A, T, C, or G with probabilities equal to the observed proportions of those bases for a given species at a particular location. Like the majority rule, this will use what is known about the other organisms within a species to fill in the missing data, but it will attempt to keep the proportion of bases in the imputed data roughly close to those of the observed data, something that is not guaranteed by the majority rule strategy. Table 3.6 gives an idea of what the procedure might cause the truncated barcode data set in Table 3.1 to look like.

Table 3.6 looks very similar to Table 3.5 except for position 11 for the first organism in species 1 and position 2 for the third organism in species 3. Because the value to be replaced using the majority rule approach is also the most likely value to be selected for replacement

Species	Truncated Barcode
1	C C G G C A T A G T A G G C A C T G
1	C C G G C A T A G T A G G C A C T G
1	C C G G C A T A G T T G G C A C T G
1	C T G G C A T A G T A G G T A C T G
1	C C G G C A T A G T A G G C A C T G
1	C C G G C A T A G T A G G C A C T G
1	C C G G C A T A G T T G G C A C T G
1	C T G G C A T A G T A G G T A C T G
2	C T G G C A T A G T C G G A A C C G
2	C T G G C A T A G T C G G A A C C G
2	C T G G C A T A G T C G G A A C C G
3	C C G G C A T A G T A G G A A C A G
3	C T G G C A T A G T A G G A A C A G
3	C C G G C A T A G T A G G A A C A G
3	C C G G C A T A G T A G G A A C A G
3	C C G G C A T A G T A G G A A C A G
4	C C G G A A T A G T A G G T A C C G
4	C C G G A A T A G T A G G T A C C G
4	C C G G A A T A G T A G G T A C C G
4	C C G G A A T A G T A G G T A C C G

Table 3.5: *Truncated Barcode Data with Majority Rule Imputation. The imputed values are in red. Only the first 18 positions of the barcode are shown here. Typical barcodes range in length from 255 positions, to 690 positions.*

in the proportional allocation approach, it is anticipated that these two alternatives will give similar results. The advantages to this approach are that observed proportions of bases for a given species at a given position after the imputation are theoretically maintained, and the conditional probabilities computed from this imputation may not reduce variation in a new barcode with the same severity of the majority rule approach leading in some cases to better misclassification rates. Because variation in a new barcode is not so severely reduced, the proposed method may require more positions to drive the posterior probabilities of the incorrect species to zero, the posterior probability of the correct species to unity, and trigger the stopping rule discussed in Section 3.7. Thus, disadvantages include more positions required to classify the species than for the majority rule approach.

Species	Truncated Barcode
1	C C G G C A T A G T T G G C A C T G
1	C C G G C A T A G T A G G C A C T G
1	C C G G C A T A G T T G G C A C T G
1	C T G G C A T A G T A G G T A C T G
1	C C G G C A T A G T A G G C A C T G
1	C C G G C A T A G T A G G C A C T G
1	C C G G C A T A G T T G G C A C T G
1	C T G G C A T A G T A G G T A C T G
2	C T G G C A T A G T C G G A A C C G
2	C T G G C A T A G T C G G A A C C G
2	C T G G C A T A G T C G G A A C C G
3	C C G G C A T A G T A G G A A C A G
3	C T G G C A T A G T A G G A A C A G
3	C T G G C A T A G T A G G A A C A G
3	C C G G C A T A G T A G G A A C A G
3	C C G G C A T A G T A G G A A C A G
4	C C G G A A T A G T A G G T A C C G
4	C C G G A A T A G T A G G T A C C G
4	C C G G A A T A G T A G G T A C C G
4	C C G G A A T A G T A G G T A C C G

Table 3.6: *Truncated Barcode Data with Proportional Allocation Imputation. The imputed values are in red and blue. The blue indicates an imputed value using proportional allocation that is different from the imputed value obtained via the majority rule imputation. Only the first 18 positions of the barcode are shown here. Typical barcodes range in length from 255 positions, to 690 positions.*

Majority Rule versus Proportional Allocation

Table 3.7 gives the 10-fold cross-validated misclassification rates for each group of test data for the five data sets we examined using both the majority rule and the proportional allocation approaches. In 10-fold cross-validation, each of the five data sets is randomly split into ten groups where each group was 10% of the entire data set, while the remaining 90% of the data is used as the reference data set. While this type of cross-validation does not provide an unbiased estimate of the true misclassification rates, as opposed to leave-one-out cross-validataion, it does provide a greater challenge to the capabilities of the proposed method.

Imputation	Individual Misclassification Rates									
Method	1	2	3	4	5	6	7	8	9	10
Bat data set										
Majority	0	0	0	0.012	0	0	0	0	0	0
Proportional	0	0	0	0.012	0	0	0	0	0	0
Bird1 data set										
Majority	0	0	0	0.006	0	0.006	0	0	0.012	0
Proportional	0	0	0	0.006	0	0.006	0	0	0.012	0
Bird2 data set										
Majority	0.045	0.023	0.032	0.023	0.005	0.032	0	0.018	0.014	0.028
Proportional	0.041	0.023	0.037	0.023	0.005	0.032	0	0.018	0.014	0.028
Butterfly data set										
Majority	0.006	0	0.003	0.003	0	0.003	0.008	0.011	0.011	0.009
Proportional	0.006	0	0.003	0.003	0	0.003	0.008	0.011	0.011	0.009
Fish data set										
Majority	0	0	0.015	0.015	0	0.015	0.015	0	0	0
Proportional	0	0	0.015	0.015	0	0.015	0.015	0	0	0

Table 3.7: *10-fold Cross-validated Misclassification rates for the five data sets using the Majority Rule, and Proportional Allocation imputation methods.*

All of the observations in each group are then classified and the overall misclassification rate for each group is reported in the table. Care was taken to ensure that each species in a test group had at least one representative in the reference data set. We see that the misclassification rates for the two approaches differ slightly for test data sets 1 and 3 in the Bird2 data set and yield identical misclassifaction rates in all other cases. In test data set 1 of Bird2, the misclassification rate using proportional allocation is slightly lower while the misclassification rate in test data set 3 of Bird2 is slightly higher for this approach. The conclusion we draw from Table 3.7 is that the imputation methods are nearly identical in terms of misclassification rates.

To further compare these two imputation methods, we examined overall position measures in classification. Table 3.8 gives the average, standard deviation, minimum, median,

Imputation Method	Overall Position Classification Measures				
	Mean	Std Dev	Min	Median	Max
Bat data set					
Majority	104.419	109.372	32	44	572
Proportional	104.392	109.383	32	44	572
Bird1 data set					
Majority	92.171	71.692	25	79	690
Proportional	92.152	71.685	25	79	690
Bird2 data set					
Majority	158.91	62.513	71	137	255
Proportional	158.972	62.557	71	137	255
Butterfly data set					
Majority	170.479	55.239	75	167	255
Proportional	170.481	55.242	75	167	255
Fish data set					
Majority	138.018	62.033	70	111	255
Proportional	138.018	62.033	70	111	255

Table 3.8: *10-fold Cross-validated overall position measures for the five data sets using the majority rule, and proportional allocation imputation methods.*

and maximum number of positions required to classify the barcodes in the five data sets for the two imputation methods. The average number of positions required for classification and the standard deviation are very similar for the two imputation methods in each case, while the minimum, median, and maximum number of positions required for classification were identical for the two imputation methods in each case.

While these two approaches to imputation yield nearly identical misclassification rates and number of positions required for classification, we recommend the proportional allocation approach to imputing the missing data because it strives to maintain the proportions observed in the reference data set.

As a variation to proportional allocation approach, one could randomly impute the

missing data with probabilities that are adjusted by the chosen δ value. Choosing an appropriate value of δ is discussed in Section 3.4. This would allow any of the four bases to be selected for imputation, with preference going to those that are the most frequent at that position for that species.

Special Cases of Missing Data

Imputing the missing data in the reference data set as illustrated above gives rise the the question of how imputation is to be carried out in the event that all of the bases are missing for a particular species at a given position. In this instance, nothing can be deduced as to what base should be used. One idea is to use the bases in that location for other species in the reference data set. This however could be dangerous in that the species with the largest number of organisms in the reference data will have a strong influence on the direction the classification takes, possibly pushing the classification in the direction of that species instead of where it otherwise might have gone. If genus-level information is known for each barcode, it may be advantageous to replace the missing values for a species by the values observed for other species within the same genus. However, this too could be dangerous because the among-species variability for even closely related species is usually larger than 6 or 7%. The danger of imputing values in this manner is that it could make it more difficult for the proposed method to distinguish between the species within the genus used to supply the missing nucleotides. This could result in incorrectly identifying the species as belonging to another species within the same genus. It seems then that there is little to gain by substituting bases present in the other species for the missing bases of the species in question, and we recommend against it.

Another idea is to simply skip over these positions. Using this approach can greatly reduce the amount of information utilized by the reference data set in constructing the conditional probabilities discussed in Section 3.1. This is especially true if several species in the reference data set have missing bases at different positions. For example, suppose a reference data set consists of 100 barcodes all with 700 positions and that the barcodes

are from 10 different species such that each species is represented by 10 barcodes. Suppose further that positions $1-10$ of species 1, positions $11-20$ of species 2, positions $21-30$ of species 3, positions $31-40$ of species 4, positions $41-50$ of species 5, positions $51-60$ of species 6, positions $61-70$ of species 7, positions $71-80$ of species 8, positions $81-90$ of species 9, and positions $91-100$ of species 10 are all missing. If we were to skip over the positions with missing values, we would eliminate a total of 100 positions, or 1/7 of the total data. While no species had more than 10 positions missing, the positions do not overlap among the species causing us to discard 90 viable positions for each species. A less contrived example comes from one data set we examined where there were 4266 barcodes all with 255 positions. One species in the reference data set was represented by a single barcode that had missing bases at all but three positions. While inclusion of such a barcode in the reference data set is questionable, this approach would eliminate all but those three positions assuming they are not missing for one or more of the other species in the reference data set. In this case, the classification would be based on three of the available 255 positions! Clearly, this is not ideal in that the possibility of such a poor classification scenario exists and we do not recommend this approach.

Another idea would be just carry the posteriors for the species with a completely missing position forward to the next position while the other posteriors update based on the data observed. This approach is equivalent to removing the species with the completely missing position from the sample space for that position. If the posterior is carried forward to the next position, the posterior probabilities no longer sum to 1. This violates the theoretical structure we draw upon to give meaning to the computed posterior probabilities. For these reasons, we do not recommend this approach.

It is recommended that when a species is completely missing a base at a position such that the imputation methods discussed in this Section are not possible, the conditional probabilities for each base at that position be set to 0.25. This implies that any of the four bases were equally likely to have been observed at that position. The posterior probability

for such species neither increases nor decreases, because the conditional probabilities used in equation (3.1) will all be the same. This essentially carries the posterior for that species forward to the next position while keeping it in the sample space. With this approach, the posteriors can still be estimated, the conditional probabilities carry intuitive meaning, and the use of the available data is maximized.

It should be noted here that Theorem 1 holds in the ideal situation where we have no missing data. From a practical view however, where we use the convention recommended above for a particular species and position, our estimate of the probability of any of the four nucleotides, given the barcode belongs to that species, is no longer consistent. While substituting 0.25 for that probability allows the process to continue and a posterior probability to be estimated, we recognize that this estimated posterior probability computed for a species at a position where this substitution has been made is likely to be different from the posterior that would have been computed were the data to have been observed at that position. As data sets grow and species with missing positions gain additional observations from which one may use the imputation methods discussed above, it is anticipated that the estimated posterior probabilities will be less biased in those positions.

3.6.2 Missing data in the Test Data Set

Missing values in the test data set give rise to a separate challenge than those encountered by missing data in the reference data set. Namely, the test data set will not have information about the barcode's species. This means that there can be no majority rule or proportional allocation imputation. While there may be several barcodes to classify, and some of them may appear to be similar, there is no way of knowing to which species they belong prior to their classification. Therefore, we recommend that no imputation should be done on the test data set, but to allow classification to continue, the positions with missing data should be skipped over so as not to contribute to the posterior probability.

3.6.3 Results with Imputations Based on Proportional Allocation

Table 3.9 has the misclassification rates using arbitrary priors (M_a), arbitrary priors using $\delta = 9,7 \times 10^{-8}$ (M_{a^m}), ascending arbitrary priors (M_{a^*}), ascending arbitrary priors using $\delta = 9,7 \times 10^{-8}$ ($M_{a^{*m}}$), descending arbitrary priors ($M_{a^{**}}$), descending arbitrary priors using $\delta = 9,7 \times 10^{-8}$ ($M_{a^{**m}}$), data-based proportional priors (M_p), data-based proportional priors using $\delta = 9,7 \times 10^{-8}$ (M_{p^m}), equal priors (M_e), and equal priors using using $\delta = 9,7 \times 10^{-8}$ (M_{e^m}) for five data sets using the proportional allocation method. If the δ value is not specified as 9.7×10^{-8}, then an arbitrary δ value of 1×10^{-4} was used. These misclassification rates are the average misclassification rates for a 10-fold cross-validated classification where each data set was split at random in to 10 groups with each group consisting of around 10% of the total observations in the data set. For a single group, the 10% selected at random served as the test data, and the remaining 90% served as the reference data. Using the reference data, conditional probabilities were constructed as outlined in Section 3.4, and the barcodes were then classified using the proposed method. Care was taken to ensure that each species selected for the test data set had at least one representative in the reference data set. This was to ensure that during the randomization to the two groups, the test group did not end up with all of the observations for any one species. This table also shows the average number of positions, $\overline{p}$, required by the proposed method in order to classify the new barcodes using arbitray priors with $\delta = 1 \times 10^{-4}$. When the posterior probability that the barcode belonged to a species was computed to be 1, the calculations were stopped, the barcode classified, and the number of barcode positions used in the calculation was recorded. The standard deviation for the number of positions used for the Bat data set is rather large. This is because there were several barcodes that required a large number of positions to be examined before the stopping rule was triggered while the rest of the observations were properly classified within the first sixty-five positions. To be specific, the median number of positions required for proper classification for the Bat data set is 57.

We can see that the average number of positions required for classification with the

M_a (Min,Max)	$M_a m$ (Min,Max)	M_{a*} (Min,Max)	M_{a*m} (Min,Max)	M_{a**} (Min,Max)	M_{a**m} (Min,Max)	M_p (Min,Max)	$M_p m$ (Min,Max)	M_e (Min,Max)	$M_e m$ (Min,Max)	$\overline{p}$(SD)
Bat data set s=96, p=659, R=756, T=84										
0	0.001	0	0.001	0	0.001	0	0.001	0	0.001	145.647 (150.886)†
(0,0)	(0,0.012)	(0,0)	(0,0.012)	(0,0)	(0,0.012)	(0,0)	(0,0.012)	(0,0)	(0,0.012)	98.868 (103.712)‡
Bird1 data set s=150, p=690, R=1461, T=162										
0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	118.499 (93.609)†
(0,0.006)	(0,0.012)	(0,0.006)	(0,0.012)	(0,0.006)	(0,0.012)	(0,0.006)	(0,0.012)	(0,0.006)	(0,0.012)	84.327 (65.790)‡
Bird2 data set s=656, p=255, R=2330, T=259										
0.029	0.025	0.024	0.020	0.026	0.023	0.026	0.021	0.027	0.023	181.950 (61.041)†
(0.005,0.046)	(0,0.041)	(0.005,0.05)	(0,0.05)	(0.005,0.041)	(0,0.041)	(0.005,0.05)	(0,0.046)	(0.005,0.046)	(0,0.041)	156.195 (63.055)‡
Butterfly data set s=559, p=255, R=3839, T=427										
0.008	0.005	0.005	0.004	0.007	0.006	0.006	0.005	0.007	0.005	201.279 (50.390)†
(0.003,0.017)	(0,0.011)	(0,0.014)	(0,0.008)	(0.003,0.017)	(0,0.014)	(0,0.014)	(0,0.011)	(0.003,0.0167)	(0,0.011)	168.839 (55.216)‡
Fish data set s=211, p=255, R=678, T=76										
0.00776	0.008	0.008	0.008	0.006	0.006	0.008	0.008	0.006	0.006	159.555 (61.387)†
(0,0.031)	(0,0.031)	(0,0.031)	(0,0.031)	(0,0.015)	(0,0.015)	(0,0.031)	(0,0.031)	(0,0.015)	(0,0.015)	137.748 (62.189)‡

Table 3.9: *Average misclassification rates for arbitrary priors (M_a), arbitrary priors using $\delta = 9,7 \times 10^{-8}$ (M_{a^m}), ascending arbitrary priors (M_{a^*}), ascending arbitrary priors using $\delta = 9,7 \times 10^{-8}$ ($M_{a^{*m}}$), descending arbitrary priors ($M_{a^{**}}$), descending arbitrary priors using $\delta = 9,7 \times 10^{-8}$ ($M_{a^{**m}}$), data-based proportional priors (M_p), data-based proportional priors using $\delta = 9,7 \times 10^{-8}$ (M_{p^m}), equal priors (M_e), and equal priors using using $\delta = 9,7 \times 10^{-8}$ (M_{e^m}). (Min,Max) give the smallest and largest misclassification rates for the 10-fold cross-validation. R and T are approximately the number of barcodes in the reference and test data sets, respectively. The values s and p are the number of species in R and number of positions on the barcode, respectively. $\overline{p}$ is the average number of positions required to assign all of the barcodes holding 10% out in a 10-fold cross-validation using arbitrary priors where † means $\delta = 1.0 \times 10^{-4}$ was used and ‡ means $\delta = 9.7 \times 10^{-8}$ was used. An arbitrary δ value of 1×10^{-4} was used in cases where the mutation rate of $9.7 \times 10{-}8$ was not used. See Section 4 for additional results.*

proposed method indicate significant reductions can be made in terms of barcode length. If we use the empirical rule, which says about 95% of the data will fall within two standard devaitions of the mean, together with the two data sets that require the largest number of postistions on average (Bird2 and Butterfly), we deduce that barcodes with lengths of $300-320$ nucleotide positions would be an upper bound on the number of positions required for proper classification using the proposed method.

3.7 Stopping Rules

A question posed by DIMACS is one that pertains to how much of the barcode is really necessary for proper classification. If the classification can be accurately done with a shorter barcode, the speed of calculations can be increased while and costs associated with DNA extraction and processing can be decreased. This question could be entertained in the proposed method by implementing a simple stopping rule. Once the posterior probability gets sufficiently large for a particular species, little is to be gained by continuing calculations until the end of the barcode. Thus, once the posterior probability gets sufficiently close to one, the barcode is classified to the species yielding that posterior probability and the calculation stops. By noting the position upon which the calculation stopped, one might begin to assess necessary barcode lengths for proper classification.

What if two or more closely related species have identical barcodes for the first few hundred positions? Will implementing a stopping rule like the one suggested above terminate the calculations prematurely and possibly yield the wrong classification? The answer to this question depends on whether the species with identical barcodes for the first few hundred positions are represented in the reference data set. As you will see, this is an issue only in the case of species discovery.

If two or more species represented in the reference data set have identical barcodes for the first few hundred positions, then, assuming equal priors are selected initially, the proposed method of classification of a barcode belonging to one of those species will result in identical

posterior probabilities for the contiguous section(s) of the barcodes that are identical. This means that none of the posteriors for the species involved will trigger the stopping rule while the positions are the same. To demonstrate this, let us consider a reference data set for fish in which the species *Thunnus alalunga*, *Thunnus maccoyii*, *Thunnus obesus*, *Thunnus orientalis*, and *Thunnus thynnus* all have identical barcodes for the first 255 positions. Let us further consider classification of a barcode belonging to one of those species, in this case, *Thunnus obesus*, using the proposed method with equal prior probabilities. Figure 3.4 gives a plot of the posterior probabilities for all of the 211 species in the reference data set. Notice that the five species mentioned above all have identical posterior probabilities for the 255 positions. We offset the posterior probabilities in this plot to make it easier to read. Notice that all five species have the same posterior probability of 0.199992 and the stopping rule is not triggered. We see in this case, that the posterior probabilities will only start to diverge when the barcodes for the various species involved start to diverge and the early stopping rule does not present an issue.

The early stopping rule can be an issue in the event that the new barcode to be classified belongs to a species that is not represented in the reference data set and is identical to the barcode(s) of one species in the reference data set for the first few hundred positions. In this case, the stopping rule will likely be triggered within the first few hundred positions, and the barcode will be classified to the species that is represented in the reference data set before there is opportunity to notice the differences that happen later on in the barcode. This means that the early stopping rule could incorrectly identify the barcode as belonging to the species in the reference data set when in fact it does not. It is recommended then that when this stopping rule is triggered, the classification be rerun starting at various, possibly random, barcode positions. Table 4.16 gives the 10-fold cross-validated misclassification rates for the fish data set. The table gives measures for the misclassification rates for the proposed method under various settings. One of these settings is given in the first column of the table labeled "Starting Position." The cross-validation was carried out starting from

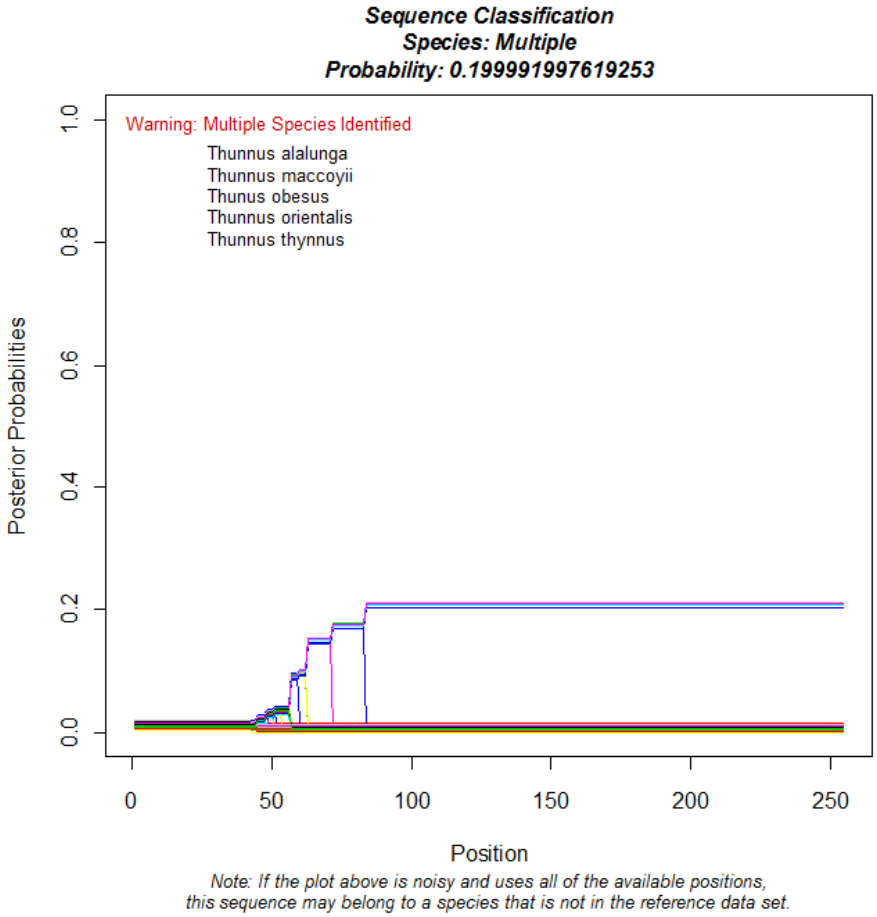

Figure 3.4: *Plotted posterior probabilities for classification of a barcode belonging to the Thunnus obesus in which the reference data set contains identical barcodes for the first* 255 *positions of the five species Thunnus alalunga, Thunnus maccoyii, Thunnus obesus, Thunnus orientalis, and Thunnus thynnus. The five species have identical posterior probabilities using the current method together with equal prior probabilities. For plotting purposes, a small bit of noise was added to the posterior probabilities for each species in the reference data set to insure visibility in the plot.*

position one of the barcode as well as selecting a random ordering for the positions and then carrying out the proposed method on this new ordering. For this data set, we see that the average misclassification rate is slightly lower for the random ordering than for simply starting with the first position. Missing data at the beginning of the data set has a lot to do with this and by randomizing the position sequence, we encounter more data

at the beginning of the sequence that would not have otherwise been there. More on this will be discussed in Sections 4.2.1 and 4.3, but here we see that the classification is still accurate when selecting a random starting position and carrying out the proposed method from that point. It is upon this footing that the above recommendation is substantiated. If the multiple runs, with different starting positions, all point to the same species contained in the reference data set, then one can feel confident that the barcode does in fact belong to the assigned species. If the multiple runs point to different species or have somewhat noisy posterior probability plots, which will be discussed in Section 3.8.3, then one may suspect that the barcode in question belongs to a species not represented in the reference data set.

3.8 Adjusting the Posteriors for Species Discovery

Another important feature that could be built into the proposed method of classification is the process of identifying new or rare species. If a barcode for a species not contained in the data set is classified, it is reasonable to think that the posterior probabilities of the species in the reference data set should all be around $1/s$ where s is the number of species in the reference data set. Convergence of the posterior probabilities to $1/s$ could then serve as a basis for attempting to discover new or rare species. If the posterior probabilities get sufficiently close to $1/s$, calculation stops, and the barcode is classified as not belonging to any of the species in the reference data set.

Equation (3.1) can classify an unknown barcode to one of the species in the reference data set, but what about the important issue of identifying a barcode that does not belong to a species in the reference data set? This question refers to the ability the method has to identify new or rare species. One would expect that, if a barcode does not belong to a species in the reference data set, then each of the posterior probabilities of the species in the reference data set should be approximately $1/s$. Equation (3.1) does not do this for the most general case, but a simple adjustment to the posterior probabilities in such cases can give the desired result. To see how this might be done, there are three cases to consider for

the situation in which the barcode from a species is not contained in the reference data set.

3.8.1 Case 1: Barcode with no matching nucleotide bases

In the first case, let us consider a barcode that is neither sequentially the same as those in the reference data set, nor does it share a common nucleotide base at any position with those in the reference data. For example, consider the barcode $AGCCTTAGCCGCCGTTGC$ as it relates to those in Table 3.1. Notice that for the first position, none of the species in the reference data set report an A. Likewise for the second position, none of the species in the reference data set report a G, and so on. If we take a closer look at equation (3.1) we see that the conditional probabilities used in the posterior calculation for each species will simply be δ. In this case, the numerator and denominator will provide some cancellation resulting in a posterior probability that is equal to the prior probability. Ideally, for this situation, the posterior probabilities would change in such a way that they move toward $1/s$, indicating that the barcode does not belong to any of those in the reference data set.

To achieve this, consider the following simple adjustment to the vector of posterior probabilities, $P(\underline{S}|x)$, when the base at position j in the barcode to be classified does not match any of the bases in position j of the barcodes in the reference data set: if the base at position j in the barcode to be classified does not match any of the bases in position j of the barcodes in the reference data set, then set

$$P(\underline{S}|x^{(j)}) = P(\underline{S}|x^{(j-1)}) + (1/s - P(\underline{S}|x^{(j-1)})) \cdot \epsilon \tag{3.13}$$

where s is the number of species in the reference data set, ϵ is a value between 0 and 1 that reflects the rate of convergence of the posterior probabilities to $1/s$, and $P(\underline{S}|x^{(j-1)})$ is the vector of prior probabilities for position j. Note that because of the sequential calculation of equation (3.1), $P(\underline{S}|x^{(j-1)})$ is also the vector of posterior probabilities calculated for the $j-1$ position. Note also that larger values of ϵ would cause $P(\underline{S}|x^{(j)})$ to converge to $1/s$ at a faster rate while smaller values of ϵ would cause $P(\underline{S}|x^{(j)})$ to converge to $1/s$ at a slower rate.

The convergence of equation 3.13 to $1/s$ is easily seen by rewriting the right-hand side as

$$P(\underline{S}|x^{(j-1)}) + (1/s - P(\underline{S}|x^{(j-1)})) \cdot \epsilon = P(\underline{S}|x^{(j-1)}) \cdot (1-\epsilon) + (1/s) \cdot \epsilon \tag{3.14}$$

One way to view equation (3.14) is as a weighted sum of the prior probabilities and $1/s$. If a large value of ϵ is chosen, the contribution of the prior probabilities to the posterior probabilities gets down-weighted while the contribution of $1/s$ to the posterior probabilities is increased. The value of ϵ used here can be thought of as something similar to the smoothing constant used in single exponential smoothing.

Figure 3.5 (a) illustrates how this adjustment would effect the posterior probabilities for the barcodes of the four species given in Table 3.1 using equation (3.13) with $\epsilon = 0.2$. Arbitrary initial prior probabilities selected for each of the four species are $P(S_1) = 0.287$, $P(S_2) = 0.629$, $P(S_3) = 0.058$, and $P(S_4) = 0.025$, respectively. If the base in the first position of the barcode in question does not match any of the those in the first position for the barcodes of the four species, and $\epsilon = 0.2$ were used, then the posterior probabilities for position 1 become $P(S_1|x^{(1)}) = 0.279$, $P(S_2|x^{(1)}) = 0.553$, $P(S_3|x^{(1)}) = 0.097$, and $P(S_4|x^{(1)}) = 0.067$. In each case, the probability has shifted slightly in the direction of $1/4 = 0.25$. Notice that if the barcode in question continues to provide bases that do not match those of the reference data for a given position, this adjustment continues to move the calculated posteriors in the direction of $1/s$. It is interesting to note how quickly the convergence takes place even for a relatively small value of ϵ. For $\epsilon = 0.2$, is seems that the adjustment causes the calculated posterior probabilities to converge to 0.25 around the 20th position. Figure 3.5 (b), (c), and (d) show how the posterior probabilities using the same prior probabilities would converge to $1/4$ setting ϵ equal to 0.4, 0.6, and 0.8 respectively. Convergence in these cases happens earlier around the 10th, 5th and 3rd positions, respectively.

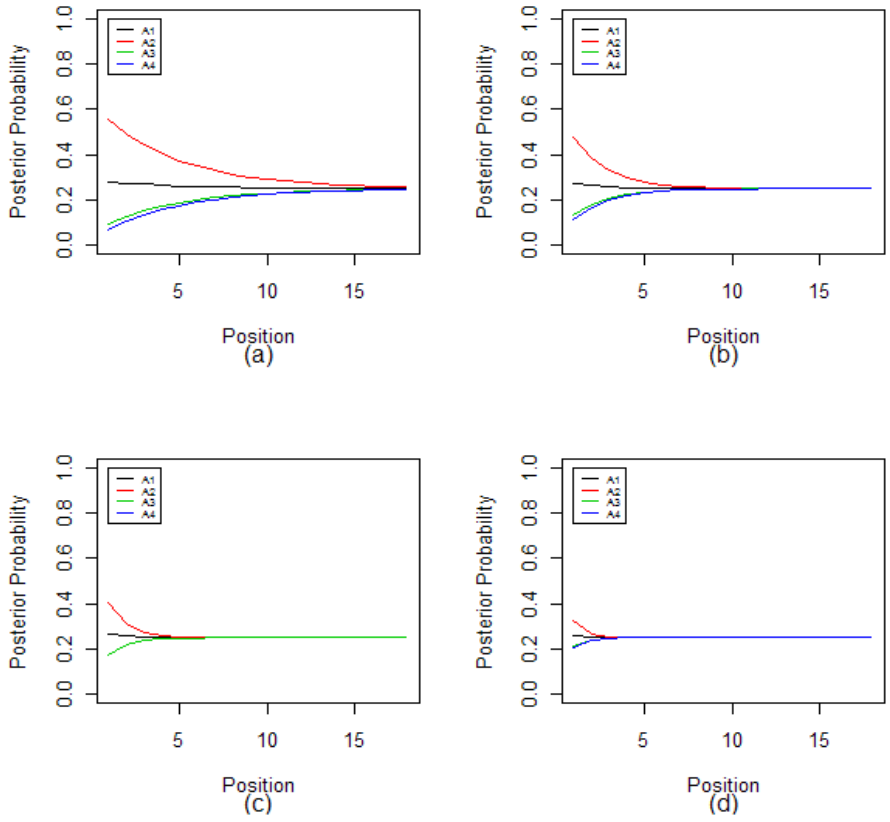

Figure 3.5: *Plotted posterior probability adjustments for each of four species at 18 positions having non informative Dirichlet priors of* $P(S_1) = 0.287$, $P(S_2) = 0.629$, $P(S_3) = 0.058$, *and* $P(S_4) = 0.025$. *(a)* ϵ *is 0.2 (b)* ϵ *is 0.4 (c)* ϵ *is 0.6 (d)* ϵ *is 0.8. Priors do not sum to one because of rounding.*

3.8.2 Case 2: Randomly generated barcode

In the second case, we consider a barcode created by randomly selecting from the values A, T, C, and G with equal probability. Some of the nucleotide bases might match those in the reference data set, but the generated barcode clearly does not belong to any of the species in the reference data set. With the adjustment in equation (3.13), we would expect to see the posterior probabilities converging to $1/s$ for bases in the generated barcode that do not match any of the bases in the reference data set at a given position. However, if the

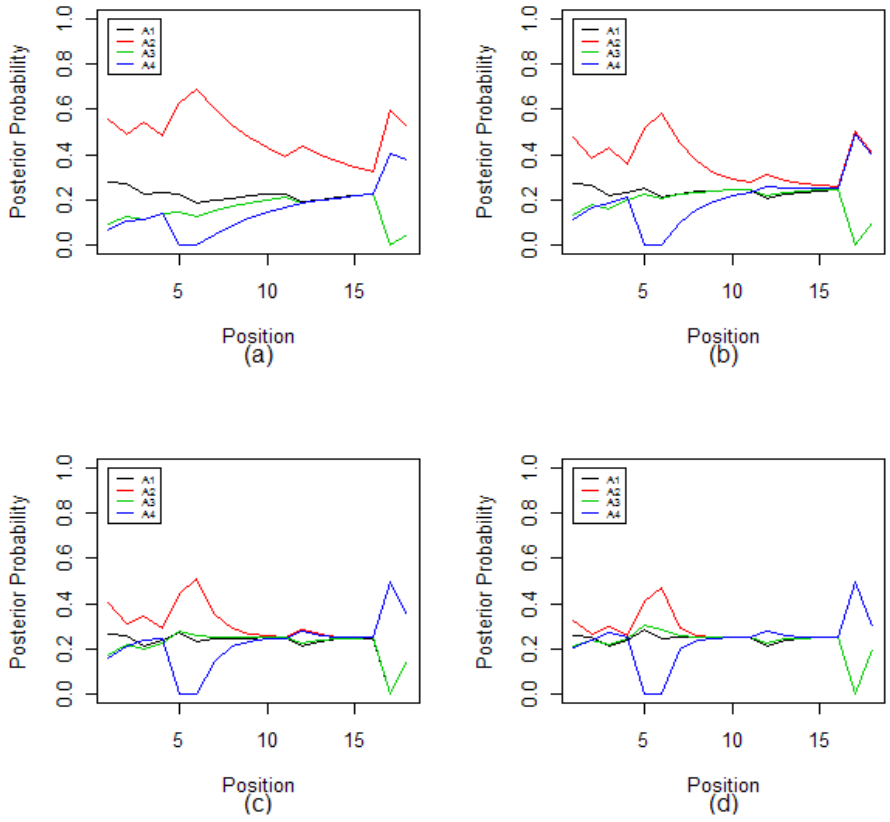

Figure 3.6: *Plotted posterior probability adjustments for each of four species at 18 randomly generated positions having non informative Dirichlet priors of* $P(S_1) = 0.287$, $P(S_2) = 0.629$, $P(S_3) = 0.058$, *and* $P(S_4) = 0.025$. *(a)* ϵ *is 0.2 (b)* ϵ *is 0.4 (c)* ϵ *is 0.6 (d)* ϵ *is 0.8. Priors do not sum to one because of rounding.*

bases match, the posterior probabilities will increase for the species that it matches, but decrease for the others. Figure 3.6 (a) illustrates this process for the randomly generated barcode $TAGACAGCTGGGCGGACA$ with $\epsilon = 0.2$. Notice the movement of the plotted posteriors to 1/4 in cases where the bases do not match and as well as their movement toward 0 or 1 when they do match.

Figure 3.6 (b), (c), and (d) show this same process setting ϵ equal to 0.4, 0.6, and 0.8, respectively. For larger values of ϵ, the posterior probabilities move toward 1/4 more

quickly and the notion of a stopping rule that terminates the calculation when the posterior probabilities get within a neighborhood of $1/s$ seems to be supported.

For large data sets, cases 1 and 2 do not seem likely to be encountered. While this adjustment to the posteriors serves as a reasonable "safety net" in the event a wildly different barcode is encountered, a case more likely to be encountered in practice is discussed in Section 3.8.3

3.8.3 Case 3: Real barcode from species not in the reference data set

In the third case, we consider the situation in which the sequence of the barcode does not match any sequence in the reference data but has matching nucleotide bases with one or more of them at many positions. Because barcodes of congeneric, or closely-related, species tend to be the most similar, this can be achieved simply by using a real barcode for one of the species that was excluded from the reference data in Table 3.1. The reference data consists of the first four species taken from a much larger data set containing 150 species. Let us consider the classification of a real barcode that comes from the sixth species of the larger data set. While, the sequence should have some divergence from the four in our reference data set, there may be several positions with matching nucleotide bases. The first 18 positions of this barcode are $CCGGAATAATTGGCACAG$. Because this barcode is not a species in the reference data set, and the nucleotide bases will match some of those in the reference data set, we might expect the maximum calculated posterior probability to bounce back and forth between the species in the reference data set, with the maximum posterior probability going to the species that happens to have a matching base at the current position. It is possible that there will be some positions in this barcode that do not have matching bases in the reference data set for a given position. In those cases, the posterior will be adjusted accordingly and should tend toward $1/s$. It is easier to see what happens in this case if we consider the entire barcode, which is 255 positions in length, and too long to reproduce entirely here, than just the previously considered 18 positions.

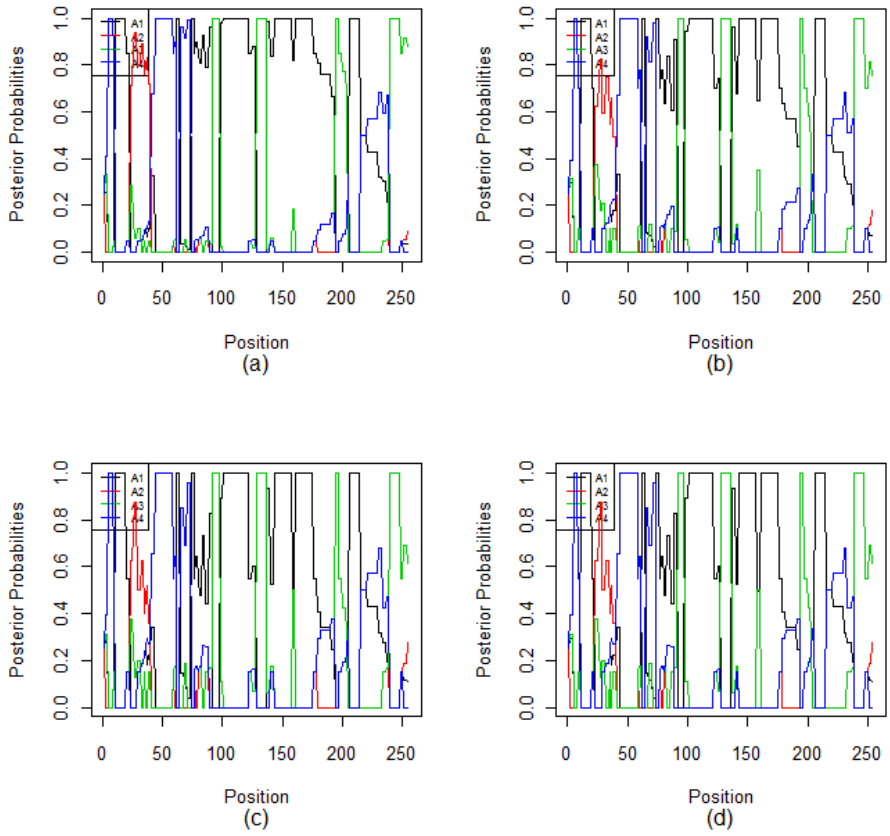

Figure 3.7: *Plotted posterior probability adjustments for each of four species at 255 positions having arbitrary priors of* $P(S_1) = P(S_2) = P(S_3) = P(S_4) = 0.25$. *(a)* ϵ *is 0.2 (b)* ϵ *is 0.4 (c)* ϵ *is 0.6 (d)* ϵ *is 0.8.*

Looking at the process over the entire barcode will give us a feel for how the posterior probability does not seem to favor any one particular species for an extended period of time.

Figures 3.7 and 3.8 show the classification using ϵ equal to 0.2, 0.4, 0.6, and 0.8 in (a), (b), (c), and (d), respectively. In Figure 3.7, equal priors of $P(S_1) = P(S_2) = P(S_3) = P(S_4) = 0.25$ were used and for Figure 3.8, arbitrary priors of $P(S_1) = 0.287$, $P(S_2) = 0.629$, $P(S_3) = 0.058$, and $P(S_4) = 0.025$ were generated from a non-informative Dirichlet distribution. It is interesting to see how frequently the species with the highest posterior

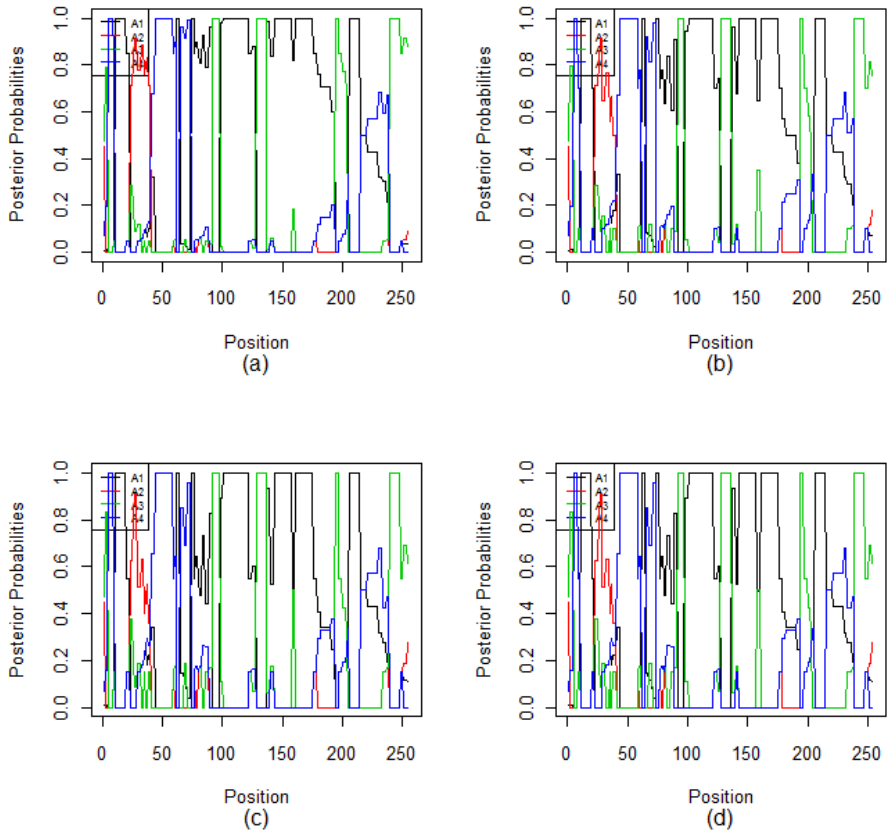

Figure 3.8: *Plotted posterior probability adjustments for each of four species at 255 positions having arbitrary priors of* $P(S_1) = 0.287$, $P(S_2) = 0.629$, $P(S_3) = 0.058$, *and* $P(S_4) = 0.025$. *(a)* ϵ *is 0.2 (b)* ϵ *is 0.4 (c)* ϵ *is 0.6 (d)* ϵ *is 0.8. Priors do not sum to one because of rounding.*

probability changes. While some of the posterior probabilities do get close to one, they do not stay there very long, and it is interesting to note that none of the posterior probabilities gets close enough to unity to trigger the stopping rule mentioned in Section 3.7. Plots that exhibit this fluctuation between species with the highest posterior probability are a very strong indication that the barcode in question does not belong to any of those in the reference data set. Stretches of the new barcode's sequence that do not match those of the reference data set become apparent as we compare the use of a moderate ϵ value of 0.2,

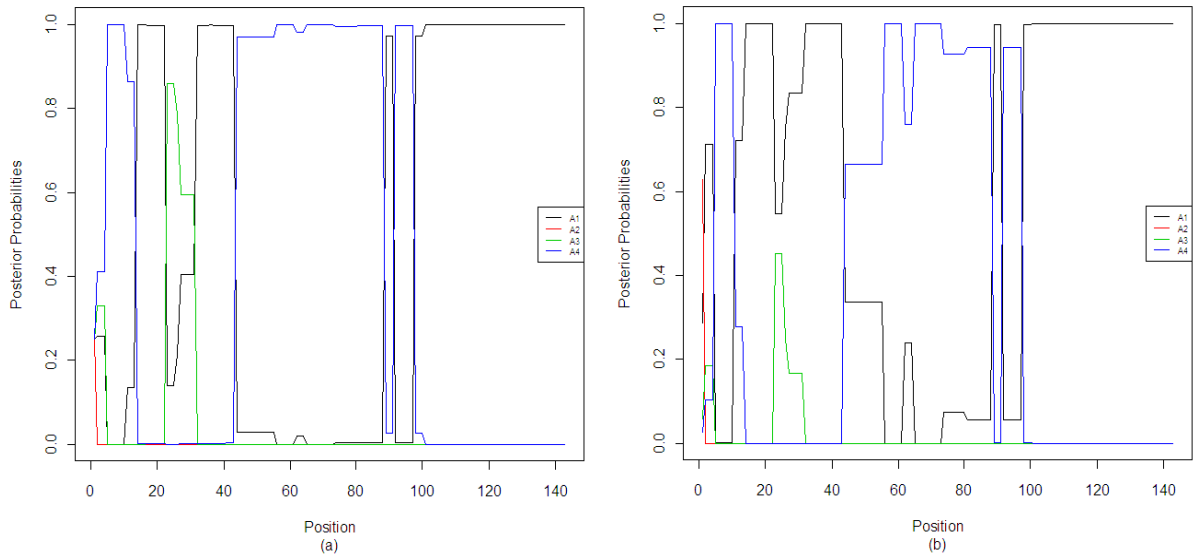

Figure 3.9: *Plotted posterior probability with* $\epsilon = 0$ *for each of four species at 143 positions having (a) equal priors, and (b) arbitrary priors of* $P(S_1) = 0.287$, $P(S_2) = 0.629$, $P(S_3) = 0.058$, *and* $P(S_4) = 0.025$. *Priors do not sum to one because of rounding.*

seen in plot (a) of both Figures 3.7 and 3.8, to a larger ϵ value of 0.8, seen in plot (d) of both Figures 3.7 and 3.8. For the (d) plots, we see a dramatic tendency toward 1/4 around positions 20-40, 65-90, and 175-190 while in the (a) plots, the posteriors tend to 1/4 in those same positions, but the movement is much more conservative. This tells us that the new barcode to be classified had few if any nucleotides in those positions that matched the collective nucleotides of the species in the reference data set at those positions. The plots of the posterior probabilities in Figures 3.7 and 3.8 look identical except for the early positions. This is an indication that the priors probabilities are quickly dominated by the data.

If $\epsilon = 0$ were to be used, the posterior adjustment would not pull the posterior probabilities toward $1/s$. Figure 3.9 (a) shows the plotted posterior probabilities using equal priors and $\epsilon = 0$. Likewise, Figure 3.9 (b) shows the plotted posterior probabilities using arbitrary priors of 0.287, 0.629, 0.058, and 0.025 for species 1, 2, 3, and 4, respectively.

In each case the stopping rule is triggered around the 143rd position and we see that the posterior probabilities still fluctuate between the four species. The posterior probabilities in these plots are not adjusted toward 1/4 when nucleotides do not match, unlike those in Figures 3.7 and 3.8. When the nucleotides in position j of the barcode to be classified do not match any of those in position j of the reference data set, the conditional probabilities in both the numerator and denominator in equation (3.1) are all be δ, and they cancel. In essence, this would not update the posterior probabilities, and they would remain the same until the next position is encountered in which the new barcode has a base that matches one in the reference data set. Graphically, this results in plotted posteriors that have horizontal bars over the regions in which there are no common bases between the new barcode and the barcodes of the reference data which is what we see when comparing Figure 3.9 to Figures 3.7 and 3.8. In the extreme case where none of the bases of the new barcode match the bases of the reference data set, the plotted posteriors result in horizontal lines (results not shown) for all species across all positions with the posterior probability of each species at the end of the calculation equal to the initial prior probability assigned to each species.

Such transition between species with the highest posterior probability is much less common in cases where the barcode being classified belongs to a species within the reference data set as illustrated by Figures 3.10 and 3.11. These figures show the plots of the calculated posterior probabilities that a barcode belongs to any of the four species when the barcode is in fact that of species 1 using all 255 positions and ϵ equal to 0.2, 0.4, 0.6, 0.8 in plots (a), (b), (c), and (d), respectively. For Figure 3.10 equal priors of $P(S_1) = P(S_2) = P(S_3) = P(S_4) = 0.25$ were used and for Figure 3.11 arbitrary priors of $P(S_1) = 0.287$, $P(S_2) = 0.629$, $P(S_3) = 0.058$, and $P(S_4) = 0.025$ were used. If the rule to stop the calculations when the highest posterior probability reaches one were to have been implemented, the calculations would have terminated around the 35^{th} position for both choices of prior probabilities. This provides a sharp contrast to the plots in Figures 3.7 and 3.8 and demonstrates the proposed method's ability to not only classify a barcode

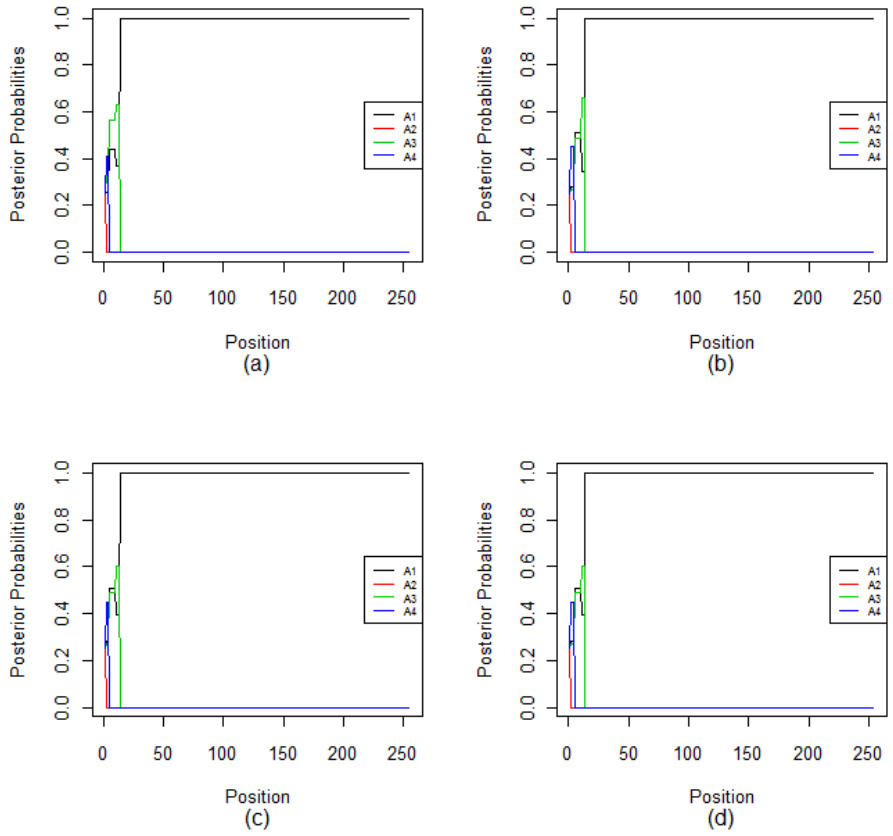

Figure 3.10: *Plotted posterior probability adjustments for each of four species at 255 positions having the priors* $P(S_1) = P(S_2) = P(S_3) = P(S_4) = 0.25$. *(a)* ϵ *is 0.2 (b)* ϵ *is 0.4 (c)* ϵ *is 0.6 (d)* ϵ *is 0.8.*

as belonging to one of the species in the reference data set, but also the method's ability to recognize when a barcode does not belong to the reference data set.

How necessary is this adjustment to the posteriors in real data sets? To address this question, we performed 10-fold cross-validation on the same five data sets examined in Section 3.6.1 using an ϵ value of 0 and 0.2. Table 3.7 gives the misclassification rates using both imputation methods discussed in Section 3.6.1 for the case in which $\epsilon = 0.2$. When $\epsilon = 0$, identical results were obtained, and no table is given here to avoid redundancy. We conclude that the misclassification rates for using a moderated value of ϵ will be similar to

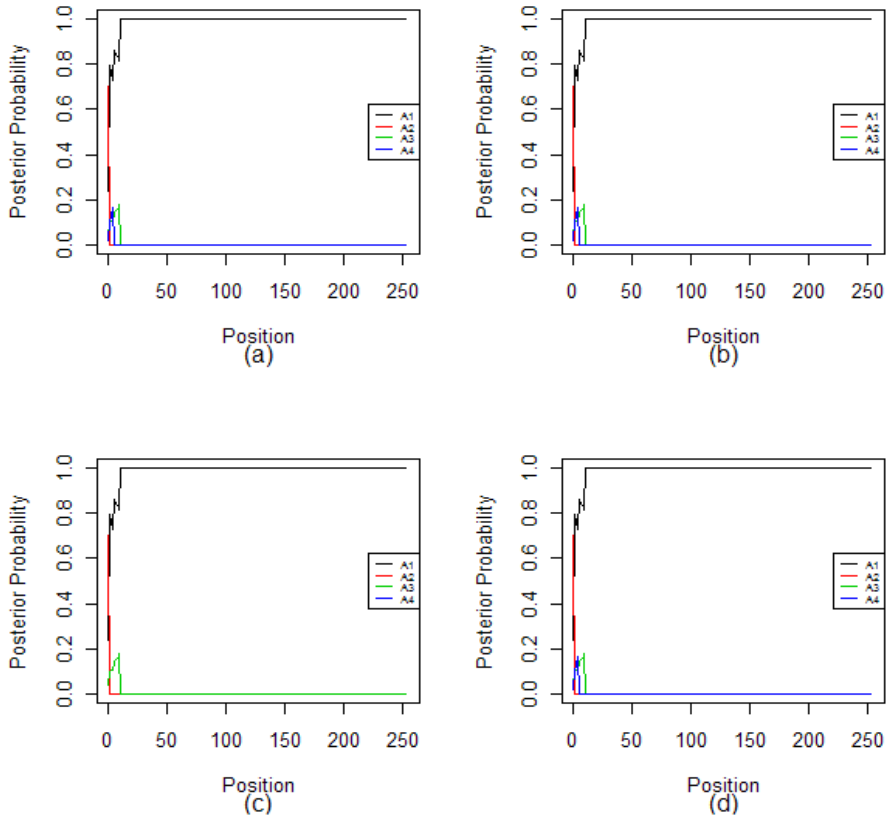

Figure 3.11: *Plotted posterior probability adjustments for each of four species at 255 positions having the arbitrary priors* $P(S_1) = 0.287$, $P(S_2) = 0.629$, $P(S_3) = 0.058$, *and* $P(S_4) = 0.025$. *(a)* ϵ *is 0.2 (b)* ϵ *is 0.4 (c)* ϵ *is 0.6 (d)* ϵ *is 0.8. Priors do not sum to one because of rounding.*

those where no ϵ adjustment is made for large data sets.

Overall measures of how many positions were required for classification in each for the five data sets were also compared using ϵ values of 0 and 0.2. Table 3.8 gives the average, standard deviation, minimum, median, and maximum number of positions required for classification using both imputation methods discussed in Section 3.6.1 and $\epsilon = 0.2$. Comparing these measures to those in Table 3.10 which uses an adjustment of $\epsilon = 0$, we see that the average number of positions required for classification for the two ϵ adjustment values of 0

Imputation Method	Overall Position Classification Measures				
	Mean	Std Dev	Min	Median	Max
Bat data set					
Majority	104.419	109.372	32	44	572
Proportional	104.392	109.383	32	44	572
Bird1 data set					
Majority	92.171	71.692	25	79	690
Proportional	92.152	71.685	25	79	690
Bird2 data set					
Majority	158.910	62.513	71	137	255
Proportional	158.972	62.557	71	137	255
Butterfly data set					
Majority	170.479	55.239	75	167	255
Proportional	170.481	55.242	75	167	255
Fish data set					
Majority	138.018	62.033	70	111	255
Proportional	138.018	62.033	70	111	255

Table 3.10: *10-fold cross-validated overall position measures for the five data sets using the Majority Rule, and Proportional Allocation imputation methods with* $\epsilon = 0$.

and 0.2 are very similar. It is interesting to note that the minimum, median, and maximum number of positions required are identical for the two different ϵ values. We conclude that the number of positions required for classification using moderate ϵ values will be similar to the number of positions required when no ϵ adjustment is made for large data sets.

While the classification results are very similar in terms of misclassification rates and number of positions required for classification, we recommend including a moderate ϵ adjustment. This will provide the posterior correction necessary in smaller data sets where a position in the reference data set may not have any nucleotides that match the nucleotide in the same position in the barcode to be classified. Using the adjustment to the posteriors in this case will be critical for species discovery.

Algorithm 3 outlines the process of the proposed method of classification.

Algorithm 3 Proposed Method of Classification

Based on a reference data set of barcodes R and a test barcode T, do the following.

1: Impute the missing data in R as discussed in Section 3.6.
2: Using R, compute the conditional probability of the bases A, T, C, and G for every species at every position.
3: Adjust the conditional probabilities above by assigning δ to all zero valued conditional probabilities while adjusting the nonzero conditional probabilities so that they will still sum to 1.
4: Use equal prior probabilities $P(S_1) = P(S_2) = \ldots = P(S_s) = 1/s$ for each species.
5: If the base in position j of T is missing, skip to position $j+1$. Otherwise, continue to the next step.
6: If the base in position j of T does not match any of the bases in position j of R, use equation (3.13) to calculate the posterior probabilities for each species. Otherwise, use equation (3.1) to calculate the posterior probabilities for each species.
7: Repeat (5) and (6) until any of the following occur

- If $P(S_l|x^{(j)}) = 1$, for any $l = 1, \ldots, s$, stop and classify barcode to species S_l. The classification should be rerun using a different starting position to confirm the classification of the barcode.
- If $P(S_l|x^{(j)}) = 1/s$ for all $l = 1, \ldots, s$, stop and conclude the new barcode does not belong to any species in R. Note: if equal priors are being used, this stopping rule will be true on the first position and should therefore not be considered for the first position.
- The end of the barcode is reached. At this point, posterior probabilities at each position should be plotted and examined. If the plot shows the species with the highest posterior probability frequently changing like the "noisy" plots in Figures 3.7 and 3.8 (b), conclude the new barcode does not belong to any species in R. If, however, the plot clearly favors one of the species over the rest like the plots in Figures 3.10 and 3.11 (a), classify the barcode as belonging to the species S_l with the highest computed posterior probability.

3.9 Classification Example

The proposed method of classification in this section will now be demonstrated on the truncated data set in Table 3.1. Here we will see how the method classifies the new barcode

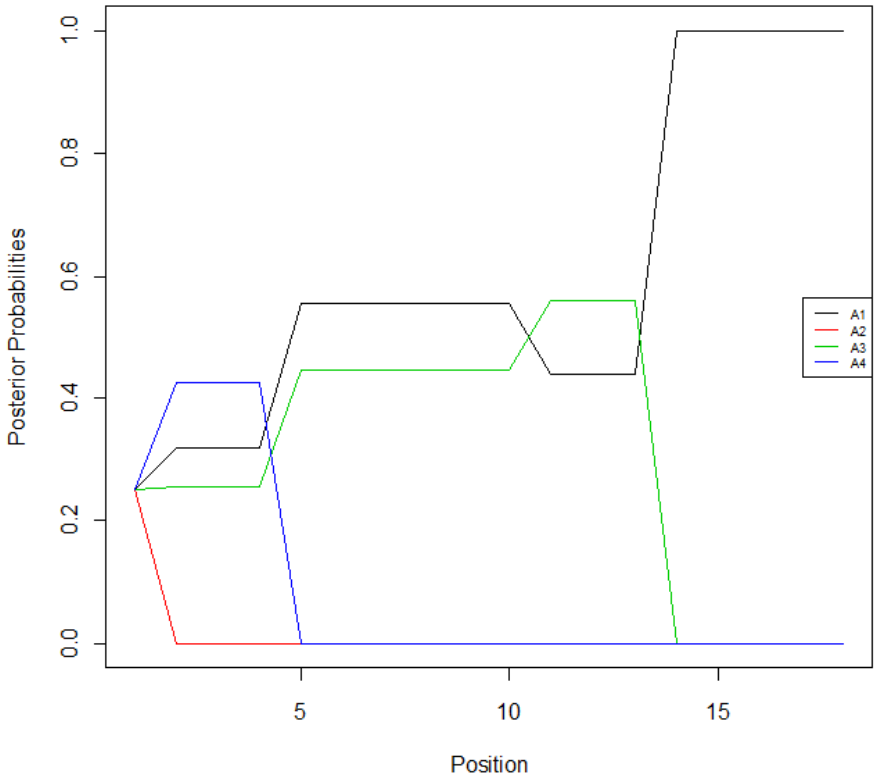

Figure 3.12: *Plotted posterior calculations for each of the four species* S_1, S_2, S_3, *and* S_4 *at each of the eighteen positions. (a) Prior probabilities were* $P(S_1) = P(S_2) = P(S_3) = P(S_4) = 0.25$.

$ACGGCATAGTTGGNACTG$ which comes from species 1 but, unlike the observations for species S_1 in the data set, the leading value is A rather than C or -. For this classification, we will use equal prior probabilities together with the recommended δ value of 9.7×10^{-8} from Section 3.4. After imputing the missing data using the proportional allocation method, constructing the conditional probabilities, and adjusting the conditional probabilities appropriately by δ, the posterior probabilities were computed for each position using equation (3.1). Figure 3.12 is a plot of the posterior probabilities for each position. Notice that the proposed method quickly begins to identify the correct species as that of species 1 around

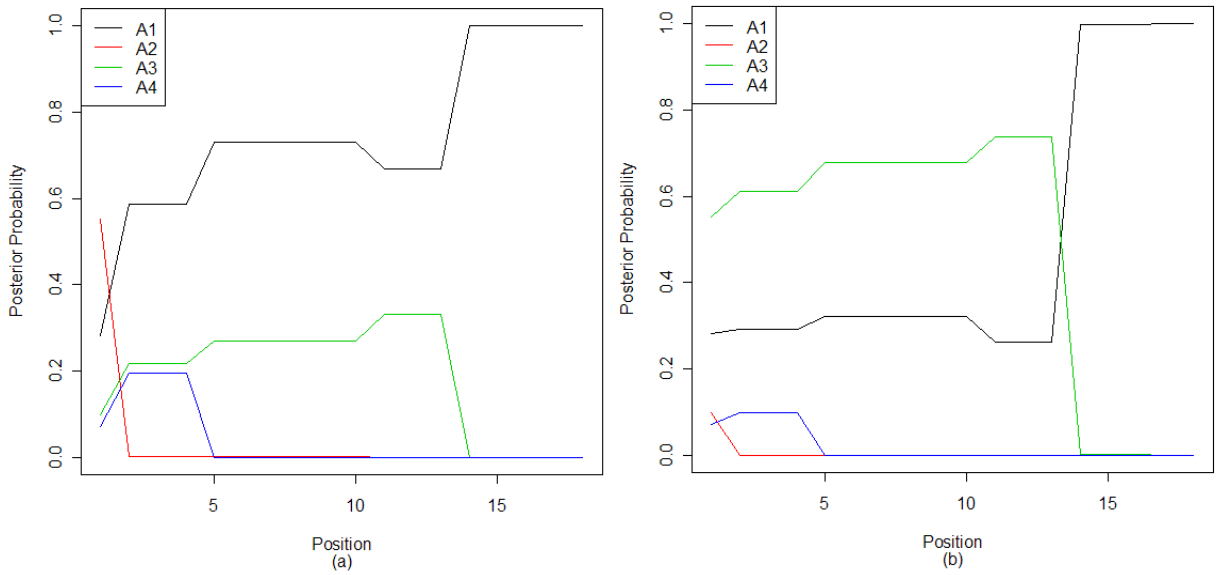

Figure 3.13: *Plotted posterior calculations for each of the four species 1, 2, 3, and 4 at each of the eighteen positions. (a) Arbitrary priors were* $P(S_1) = 0.287$, $P(S_2) = 0.629$, $P(S_3) = 0.058$, *and* $P(S_4) = 0.025$. *(b) Arbitrary priors were* $P(S_1) = 0.287$, $P(S_2) = 0.058$, $P(S_3) = 0.629$, *and* $P(S_4) = 0.025$. *Priors do not sum to one because of rounding.*

the 5th position and achieves a posterior probability close to one by the 15th position.

If we had just arbitrarily set $\delta = 1.0 \times 10^{-4}$ and used arbitrary prior probabilities of 0.28746055, 0.62927930, 0.05842476, and 0.02483539 for species 1, 2, 3, and 4, respectively, then Figure 3.13 (a) illustrates the resulting posterior probabilities at each position. It is interesting to note that the calculated posterior probability for the first position reflects the adjustment discussed in Section 3.8. This is because the value A does not appear in first position for any of the four species in the reference data set. It is also interesting to note that while species 2 was assigned the highest prior probability, it was quickly overcome by the data and forced down to zero as were the posterior probabilities for species 3 and 4.

Upon inspecting the data set, it seems that species 3 is the most similar to species 1 and the question arises, "How would this method have performed if S_3 had been assigned a stronger prior?" Figure 3.13 (b) shows the computed posteriors having swapped the prior

probabilities of species 2 and 3 such that the prior probabilities of species 1, 2, 3, and 4 are now 0.28746055, 0.05842476, 0.62927930 and 0.02483539 respectively. While the method supports the possibility that the barcode belongs to species 3 for a longer period of time in this case, it eventually makes the correct classification.

3.10 Discovery Example

To examine the ability of the proposed method to indicate that a new barcode does not belong to any of the species in the reference data set, we first look at an example of classification when the barcode does belong to a species in the reference data set, and then contrast that with an example of classification when the barcode does not belong to a species in the reference data set. In these examples, we use the Bat barcode data set which has 826 barcodes, each containing 659 nucleotides, belonging to 96 unique species. We randomly selected the *Uroderma bilobatum* species, more commonly known as the "Yellow-eared bat," and use it here for demonstrating both species classification and species discovery.

To demonstrate typical species classification, we randomly selected one of the four barcodes belonging to this species and held it out of the reference data set. Using equal prior probabilities with a δ value of 9.7×10^{-8}, the proposed method calculated the posterior probabilities at each position plotted in Figure 3.14 (a). This plot contains the posterior probabilities computed for each of the 659 positions in the barcodes belonging to the 96 species in this Bat data set. From plot (a), we see that the proposed method hones in on the correct species from about the 20th position on. If the stopping rule had been used, the calculations would have terminated around the 32nd position, and the correct classification to the species *Uroderma bilobatum* would have been made at that point. Plots like this one are typical when the barcode to be classified belongs to a species in the reference data set.

To demonstrate species discovery, all four observations belonging to this species were removed from the reference data set, and the randomly selected barcode above was used in the proposed classification method. Again, equal prior probabilities were used together with

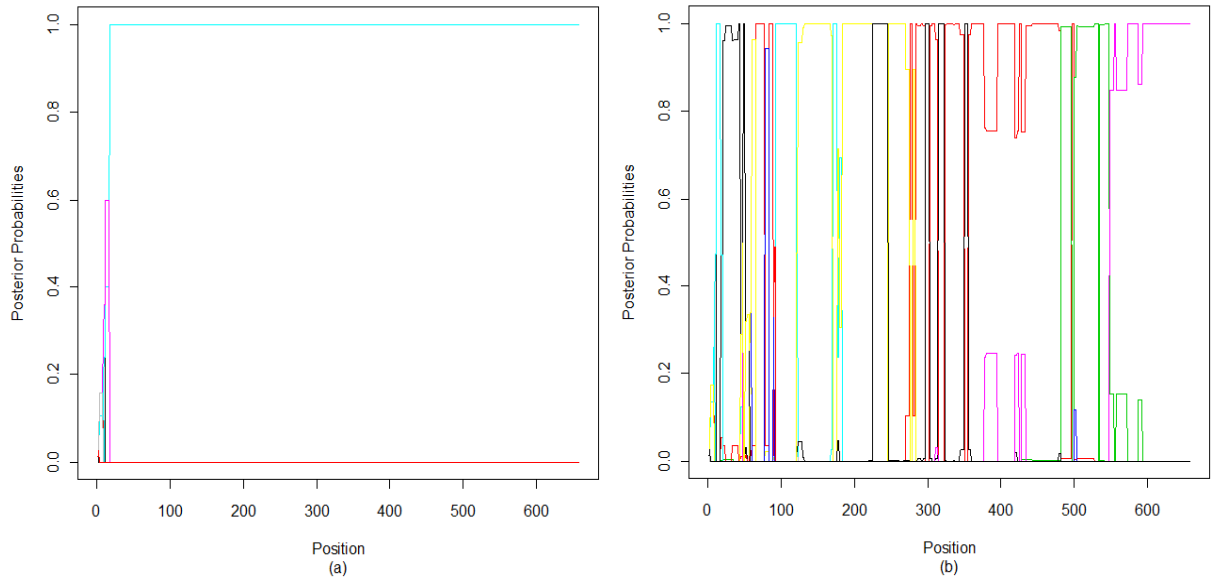

Figure 3.14: *Plotted posterior calculations for the Bat data set where (a) all of the 96 species were used in the reference data set with one barcode from the species Uroderma bilobatum held out of the reference data set and classified; (b) the species Uroderma bilobatum was completely removed from the reference data set and a barcode from that species was classified. For both plots, equal priors with a* δ *value of* 9.7×10^{-8} *was used.*

a δ value of 9.7×10^{-8}. Plot (b) of Figure 3.14 gives the computed posterior probabilities at each position for the 95 species in the reference data set. This plot never really hones in on one particular species but rather fluctuates back and forth between several species with the highest posterior probability. The posterior probabilities for several of these species get close to one, but none get close enough to trigger the stopping rule. A "noisy" plot of the posterior probabilities like this one is a strong indication that the barcode to be classified does not belong to any species in the reference data set. Plots like these will be especially interesting to biologists in that they not only give a clear indication the the new barcode does not belong to any species in the reference data set, possibly indicating a newly discovered species, but they also indicate which species in the reference data set the new barcode is most similar to and at what positions along the COI region they are similar.

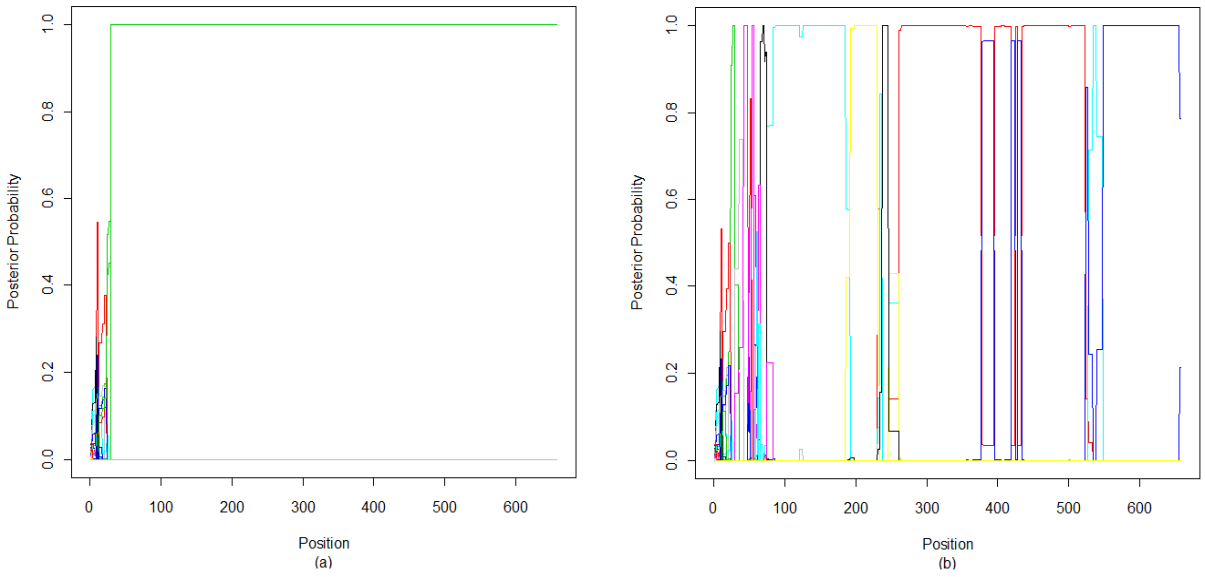

Figure 3.15: *Plotted posterior calculations for the Bat data set where (a) all of the 96 species were used in the reference data set with one barcode from the species Uroderma bilobatum held out of the reference data set and classified; (b) the species Uroderma bilobatum was completely removed from the reference data set and a barcode from that species was classified. For both plots, arbitrary priors with an arbitrary δ value of 1×10^{-4} was used.*

For example, from plot (b) of Figure 3.14, the species represented by the black, yellow, red, green and pink lines seem to keep matching short sequences of the new barcode. The species represented by the black line matches early on from about positions 40-60 and shows up several times later. The species represented by the yellow, red, green, and pink lines have somewhat larger sections of similar sequences in positions 125-225, 375-475, 475-550, and 550-659, respectively. This could possibly help a biologist determine the functionality of sequence sections in the COI region.

To see how changing the posterior probabilities and the δ value would effect the plots in Figure 3.14, we carried out the classification exactly as above with the exception of using arbitrary prior probabilities generated from a non-informative Dirichlet distribution with δ arbitrarily set to 1.0×10^{-4}.

Figure 3.15 (a) is a plot of the computed posterior probabilities where the species *Uroderma bilobatum* was in the reference data set, and the new barcode belonged to that species. This plot shows the posterior probabilities clearly favoring a single species from around the 30^{th} position on. Using the stopping rule, this classification would have terminated and made the correct classification at the 57^{th} position. Compare this plot to Figure 3.15 (b) where the species *Uroderma bilobatum* was completely removed, as before, from the reference data set. The species with the highest posterior probability fluctuates between several species, and none of them get close enough to unity to trigger the stopping rule discussed in Section 3.7. It is interesting to note that in this plot, we do not see the posteriors converging to 1/95, the number of species in the reference data set having removed the species above. This indicates that there are not long runs with bases that don't match any of those in the reference data set, but it does not find commonalities with any one species for long runs of bases either. This is a clear indication that the new barcode represents a new species not contained in the reference data set.

3.11 Genus-level classifications

One interesting question is how well the proposed method might work at higher taxonomic levels. To demonstrate the proposed methods ability at the genus level, we use four of the five data sets explored in Section 3.6.3 that had genus-level information and provide classifications of the barcodes to their respective genera. The results of these classifications are given in Table 3.11.

While the proposed method still yields better misclassification rates than the current method, the misclassification rate for the proposed method at the genus-level has increased slightly for the Bat data set and increased significantly for the Butterfly data set when compared to the classifications at the species level. Also, the misclassification rates for the Bird2 and Fish data sets have decreased slightly. The reason for the large increase in the misclassification rate of the Butterfly data set (around 0.0065 at the species level and

Data Set	s	p	R	T	M_a	M_e	M_p	M_c
Bat	50	659	756	84	0.0073	0.0073	0.0073	0.0040
Bird2	289	255	2330	259	0.0191	0.0191	0.0191	0.0540
Butterfly	205	255	3839	427	0.0666	0.0666	0.0652	0.1240
Fish	112	255	678	76	0.0054	0.0054	0.0040	0.0200

Table 3.11: *Genus-level Misclassification Rates for Arbitrary and Proportional Priors for Proportional Allocation Imputation. R and T are the number of barcodes in the reference and test data sets, respectively. M_a, M_e, M_p, M_c represent the misclassification rates for arbitrary priors, equal priors, and data based proportional priors, respectively. M_c represents the misclassification rates for the current method. In each case $\delta = 1.0 \times 10^{-4}$.*

0.0666 at the genus level) may be due to that data set having around 9% variability among genera while the among genus variability of the other data sets was around 20%. This would certainly provide a challenge to discriminate among the various species, and it is interesting to note that the misclassification rate of the current method has increased from 0.0962 at the species level to 0.124 at the genus level. This reduced variability among genera for the Butterfly data set may be attributed to how simple that organism is compared to the more complex organisms of the other three data sets.

Discovery at the genus level with the proposed method works in much the same fashion as discovery at the species level. Figure 3.16 (a) shows the classification of a barcode which belongs to the *Uroderma* genus which is represented in the reference data set. It seems that the proposed method is making the correct genus-level classification from about the 30th position on. In fact, the stopping rule would have been triggered at the 57th position, but we allowed the calculation to continue for all positions. In contrast, (b) of the same figure shows how classification of the same barcode would effect the posterior probabilities if the *Uroderma* genus were to be completely removed from the reference data set. In this case, the stopping rule was never triggered, and the genus with the highest posterior probability shifts between several genera, much like we saw in Section 3.10 at the species level. Again, this would provide a clear indication that the barcode to classified does not belong to any genus in the reference data set.

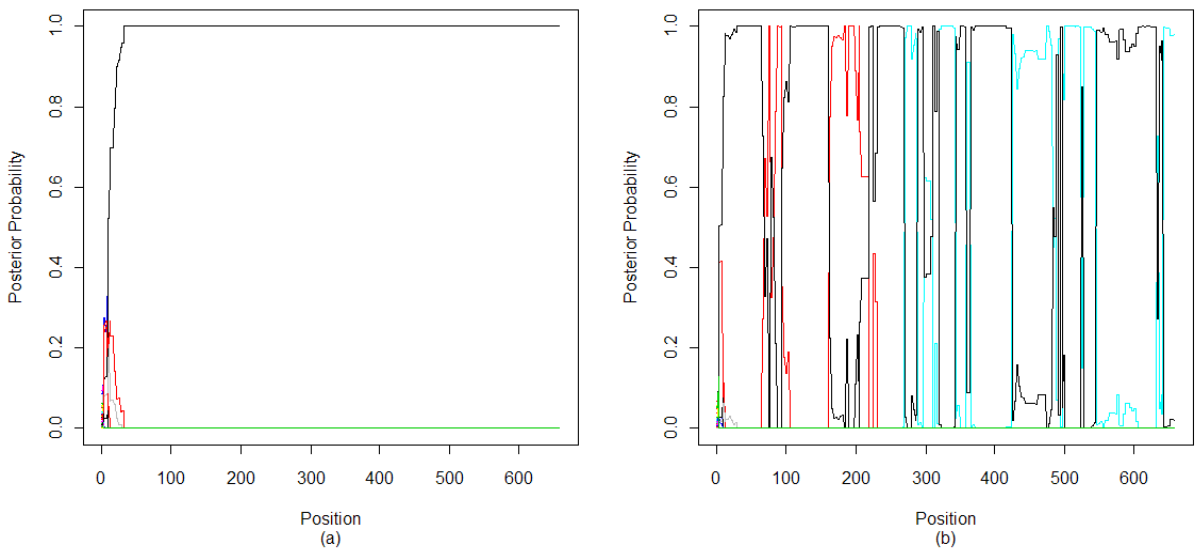

Figure 3.16: *Plotted posterior calculations for the Bat data set where (a) all of the 50 genera were used in the reference data set with one barcode from the genus Uroderma held out of the reference data set and classified; (b) the genus Uroderma was completely removed from the reference data set and a barcode from that genus was classified.*

3.12 Summary

The proposed sequential character-based method of classification overcomes some of the serious shortcomings of current distance-based methods. Namely, it will allow barcode classification based on maximum posterior calculations which provides a ready assessment of how likely the barcode is to belong to the species to which it was classified. With current methods, there is no probabilistic species-level interpretation of the results. At best, they will provide a percentage of identical positions between a barcode to be classified and its nearest neighbor in the neighbor-joining tree. This percentage is somewhat misleading in that a practitioner may easily mistake it for the probability the barcode belongs to the species of its nearest neighbor. With the proposed method, this is not an issue. As demonstrated in Theorem 1, the resulting calculation does in fact carry that interpretation. To be sure, using

the proposed method a practitioner can interpret the resulting quantity as the probability the barcode belongs to the assigned species.

Also, the proposed method implements a stopping rule that classifies the barcode once the posterior probability for a particular species achieves unity. This provides useful information on how much of the barcode is necessary for proper classification. In Section 3.6.3, we give the average number of positions required for classification of barcodes from five different data sets. For the data sets that required the largest number of positions, we showed using the empirical rule, that the average number of barcodes will often be less that 320 positions. In the case of the Bat data set, for which the misclassification rates were extremely small, we saw that proper classification did not require a large number of positions. The median number of positions required for proper classification was 57, and the minimum number of positions required was 30. These results indicate that, the proposed method can be effectively used on much shorter DNA sequences.

The proposed method creates conditional probabilities for each species at each position for the four bases A, T, C , and G to be used in the calculation of equation (3.1). This allows classification to take place even when there is only one observation for an organism in the reference data set. This is critical to properly evaluating current barcode data sets that may have several species represented by only one or two organisms. This takes some of the burden off of explicitly relying upon the genetic "gap" discussed in Section 1.2 that will be difficult validate with non-comprehensive data sets as pointed out by Meyer and Paulay (2005). It should also be pointed out that the proposed method also removes the need to make any genetic model assumptions beyond the mutation rate of the mitochondrial genome to obtain a value for δ. This makes it very flexible and easy to apply to other high dimensional data settings.

Finally, the proposed method can identify new species that are not present in the reference data set, which can assist in species discovery. As technology advances, practitioners anticipate in the not too distant future, the creation of a hand-held device that can easily

extract barcodes "in the field." If this device is equipped with an up to date DNA barcode reference data set and the proposed method of classification, then it is possible that the rate at which new species are discovered could be greatly increased. Within minutes of obtaining a barcode sample, the organism could be either classified or, perhaps more importantly, be identified as not belonging to any species in the reference data set and flagged as a potentially new species. As these devices become widely available, there is potential to bring the capabilities of species classification and discovery to even novice naturalists who can play an important role in exploring the biodiversity of larger areas than were otherwise possible with only a few experts in species identification.

Chapter 4

Results

The results in this chapter are separated into two types. The results in Section 4.1 demonstrate the proposed methods abilities on randomly generated DNA barcodes, whereas the results in Section 4.2 present an evaluation of the proposed method as it was applied to five real barcode data sets.

4.1 Results of the Proposed Method via Simulation

The aim of this simulation study is to challenge the proposed method in terms of classification and species discovery with barcodes that have specific amounts within- and among-species variability. To do this, we randomly selected a real DNA barcode data set and computed the prevalence of each nucleotide base in the data set. The data set randomly selected was the Fish data set with 750 barcodes each having 255 positions for 211 unique species. We then selected, with replacement, 700 nucleotide bases at random with each base having probability of selection equal to its observed proportion in the real barcode. This yielded a new "seed" barcode with 700 positions that had length and nucleotide prevalence about equal to what would be observed in a genuine barcode. This barcode served as a generating sequence, from which the entire data set was to be constructed. Table 4.1 contains the observed prevalences of each nucleotide base in: the real Fish data set, the seed barcode from which the entire simulated data comes, and the entire simulated data.

We sought to generate four barcodes per species for each of 12 species having within-

Data Set	A	T	C	G
Real Data	0.256	0.215	0.209	0.321
Barcode	0.283	0.194	0.210	0.313
Simulated Data	0.280	0.198	0.214	0.308

Table 4.1: *Nucleotide base prevalence in: the real Fish data set which contains* 750 *barcodes of length* 255 *for* 211 *unique species, the seed barcode generated from these prevalences, and the entire simulated data.*

species variability equal to 2% and among-species variability around 6-8% in accordance with the suggested within- and among-species variabilities presented in Johns and Avice (1998). This was achieved by first making 12 identical copies of the seed barcode. We desired these barcodes to have 8% dissimilarity among them so we selected $700 \times 0.04 = 28$ positions on each barcode and made them unique to the rest of the barcodes. Now when comparing the barcodes pairwise, there will be exactly, $28 \times 2 = 56$ differences between them. This resulted in altering 672 of the 700 positions across the 12 species.

Next we wanted to use these 12 barcodes as "seeds" for generating four barcodes within each species. We desired the within species variability to be exactly 2%. Again, we made four identical copies of each of the seeds and then selected $700 \times 0.01 = 7$ positions on each barcode and made them unique to the rest of the barcodes within that species. It would have been ideal to make them unique across all barcodes, not just within a species, to preserve the among species variability of 8%. This, however, presented a significant challenge, and it was not feasible to control both within- and between-species variability. It was determined that maintaining the within-species variablity exactly was more vital than controlling the between-species variability. This resulted in the between-species variability ranging from 6.7-10%, but the within-species variability was controlled exactly to be 2%. When comparing the barcodes within a species, there will be exaclty $7 \times 2 = 14$ pairwise differences among them. The simulated data set contains 48 barcodes, four from each of 12 species, with 700 nucleotide positions with exactly 2% within-species variability and among species variability between 6.7 and 10%. We repeated this process to obtain additional

data sets having within-species variablity of 4, 6, and 8%, while maintaining among-species variability of 6-10%.

4.1.1 Simulation: Classification

Test and reference data sets were created by randomly selecting 4 barcodes, or 12% of the data, at a time as the test data set. The reference data set consisted of the remaining 88% of the barcodes from which conditional probabilities for the proposed method were calculated. The barcodes in the test data set were then classified for various prior probabilities and δ values. Tables B.1 and B.2 in Appendix B and Tables 4.2 and 4.3 give the 10-fold cross-validated average misclassification rates with 2, 4, 6, and 8% within-species variablility, respectively. These tables give the results for the proposed method when using: arbitrary prior probabilities generated from a non-informative Dirichlet distribution, ascending arbitrary prior probabilities, descending arbitrary prior probabilities, data-based proportional prior probabilities determined from the prevalence of each species in the reference data set, and equal prior probabilities. Because the the reference data set will have unequal numbers of observations per species after randomly selecting the observations for the test data, the data-based proportional prior probabilities and the equal prior probabilities will not be identical so we include both in our study. We also examine the effect of using an arbitrarily selected $\delta = 1.0 \times 10^{-4}$ versus $\delta = 9.7 \times 10^{-8}$ which is based on the mutation rate of the mitochondrial genome. The proposed method makes the correct classification in every case for all of the chosen prior probabilities and for both δ values for the data simulated to have 2% within-species variability and 6-10% among-species variability. It is interesting to note that by increasing the within-species variability to 4% and holding the among-species variability between 6-10%, the proposed method still makes the correct classification in every case for every combination of prior probability and δ value. Tables B.1 and B.2 in Appendix B contain these results. These tables also contain the results for the current method, which uses the neighbor-joining method together with the Kimura's Two Parameter (K2P) model.

We see that this method also makes the correct classification for 2 and 4% within-species variability.

To challenge the proposed method further, we increased the within-species variability to 6% while keeping the between species variability around 6-10%. These results are found in Table 4.2. We found that, for the arbitrarily selected $\delta = 1.0 \times 10^{-4}$, the proposed method made the correct classification for every choice of prior probabilities. Using the mutation rate of 9.7×10^{-8}, however, resulted in an increased average misclassification for every choice of prior probabilities. To be sure, one barcode was misclassified for this choice of δ in the 9th test data set as well as the 11th test data set. The average misclassification rate using this value of δ was 0.042 for each choice of prior probabilities. The neighbor-joining method with the K2P model misclassified one of the observations in the eigth test data set achieving an average misclassification rate of 0.021. This misclassification rate is better than that of the proposed method when $\delta = 9.7 \times 10^{-8}$ is used, but is worse than that of the proposed method when $\delta = 1.0 \times 10^{-4}$ is used.

We would expect the average misclassification rates to continue to increase as the within-species variability increases. This can be seen in Table 4.3 in which the within-species variability was increased to 8% while holding the among-species variability at 6-10%. We see that, using $\delta = 1.0 \times 10^{-4}$, the proposed method correctly classified every observation in test data sets 5, 6, 7, 9, and 10, while using $\delta = 9.7 \times 10^{-8}$ yielded correct classifications for every observation in groups 2 and 6 only. The choice of prior probabilities did not seem to play a large roll in these misclassification rates, but the arbitrary prior probabilities with $\delta = 1.0 \times 10^{-4}$ made one fewer misclassifications in test data set 3 than the other prior probabilities. This choice of prior probabilities with $\delta = 9.7 \times 10^{-8}$ also made one fewer misclassifications in test data set 10 than the other prior probabilities. The neighbor-joining method with the K2P model is especially effected by this increase of within-species variability. This method has an average misclassification rate of 0.313 which is nearly double that of the proposed method using $\delta = 1.0 \times 10^{-4}$.

δ	Overall Misclassification Measures					Individual Misclassification Rates											
	Mean	S.D.	Min	Median	Max	1	2	3	4	5	6	7	8	9	10	11	12
Arbitrary Priors																	
1.0×10^{-4}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9.7×10^{-8}	0.042	0.097	0	0	0.25	0	0	0	0	0	0	0	0	0.25	0	0.25	0
Ascending Arbitrary Priors																	
1.0×10^{-4}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9.7×10^{-8}	0.042	0.097	0	0	0.25	0	0	0	0	0	0	0	0	0.25	0	0.25	0
Descending Arbitrary Priors																	
1.0×10^{-4}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9.7×10^{-8}	0.042	0.097	0	0	0.25	0	0	0	0	0	0	0	0	0.25	0	0.25	0
Data-Based Proportional Priors																	
1.0×10^{-4}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9.7×10^{-8}	0.042	0.097	0	0	0.25	0	0	0	0	0	0	0	0	0.25	0	0.25	0
Equal Priors																	
1.0×10^{-4}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9.7×10^{-8}	0.042	0.097	0	0	0.25	0	0	0	0	0	0	0	0	0.25	0	0.25	0
Neighbor-joining with K2P model																	
	0.021	0.072	0	0	0.25	0	0	0	0	0	0	0	0	0	0.25	0	0

Table 4.2: *Cross-validated Misclassification rates for the Simulated data set with 6% within-species variability. For the individual misclassification rates for groups 1-10; the number of observations in the reference data set R were 44, and the number of observations in the test data set T were 4.*

	Overall Misclassification Measures					Individual Misclassification Rates											
δ	Mean	S.D.	Min	Median	Max	1	2	3	4	5	6	7	8	9	10	11	12
Arbitrary Priors																	
1.0×10^{-4}	0.146	0.129	0	0.25	0.25	0.25	0.25	0.25	0.25	0	0	0	0.25	0	0	0.25	0.25
9.7×10^{-8}	0.229	0.167	0	0.25	0.5	0.25	0	0.5	0.25	0.25	0	0.25	0.25	0.25	0	0.5	0.25
Ascending Arbitrary Priors																	
1.0×10^{-4}	0.167	0.163	0	0.25	0.5	0.25	0.25	0.5	0.25	0	0	0	0.25	0	0	0.25	0.25
9.7×10^{-8}	0.25	0.151	0	0.25	0.5	0.25	0	0.5	0.25	0.25	0	0.25	0.25	0.25	0.25	0.5	0.25
Descending Arbitrary Priors																	
1.0×10^{-4}	0.167	0.163	0	0.25	0.5	0.25	0.25	0.5	0.25	0	0	0	0.25	0	0	0.25	0.25
9.7×10^{-8}	0.25	0.151	0	0.25	0.5	0.25	0	0.5	0.25	0.25	0	0.25	0.25	0.25	0.25	0.5	0.25
Data-Based Proportional Priors																	
1.0×10^{-4}	0.167	0.163	0	0.25	0.5	0.25	0.25	0.5	0.25	0	0	0	0.25	0	0	0.25	0.25
9.7×10^{-8}	0.25	0.151	0	0.25	0.5	0.25	0	0.5	0.25	0.25	0	0.25	0.25	0.25	0.25	0.5	0.25
Equal Priors																	
1.0×10^{-4}	0.167	0.163	0	0.25	0.5	0.25	0.25	0.5	0.25	0	0	0	0.25	0	0	0.25	0.25
9.7×10^{-8}	0.25	0.151	0	0.25	0.5	0.25	0	0.5	0.25	0.25	0	0.25	0.25	0.25	0.25	0.5	0.25
Neighbor-joining with K2P model																	
	0.333	0.222	0	0.25	0.75	0.75	0.5	0.5	0.25	0.25	0.25	0	0.5	0	0.25	0.5	0.25

Table 4.3: *Cross-validated Misclassification rates for the Simulated data set with 8% within-species variability. For the individual misclassification rates for groups 1-10; the number of observations in the reference data set R were 44, and the number of observations in the test data set T were 4.*

These results indicate that as the within-species variability increases, the misclassification rate of the proposed method with a carefully chosen δ value, increases at a slower rate than that of the neighbor-joining method with the K2P model. It also seems that the average misclassification rate has less to do with the initial prior probabilities and more to do with the choice of δ.

The misclassification rates for the simulated data set above were one exploration of how well the proposed method works. Another indicator of how well the method is working involves measures of the number of positions the proposed method used in the classification process. For the same 12 test data sets above, for each classification, the total number of positions required before encountering the stopping rule and making the classification was recorded. Table 4.4 gives the overall measures for the number of positions required for classification for the same combinations of prior probabilities and δ values where the within-species variability is 2% and the among-species variability is around 6-10%. The number of positions required for classification is very similar for all priors, holding the δ value fixed. For example, by choosing $\delta = 1.0 \times 10^{-4}$, the average number of positions required for classification ranges from 92.444 in the case with equal prior probabilities to 94.167 in the case using arbitrary prior probabilities generated from a non-informative Diriclet distribution. Likewise, using $\delta = 9.7 \times 10^{-8}$, the average number of positions required ranges from 60.611 in the cases of arbitrary priors, descending arbitrary priors, data-based proportional priors, and equal priors to 60.917 in the case of ascending arbitrary prior probabilities. Across the various choices of prior probabilities, there is less variability in the average number of positions required for classification using $\delta = 9.7 \times 10^{-8}$ than there is when using $\delta = 1.0 \times 10^{-4}$. Using $\delta = 9.7 \times 10^{-8}$ requires 32.539 fewer positions, on average, than the arbitrarily selected $\delta = 1.0 \times 10^{-4}$.

Table 4.5 gives the average number of positions required for classification when the within-species variability was increased to 4%. We see that the average number of positions required for classification is fairly consistent across the choice of prior probabilities for a fixed

δ	Overall Position Classification Measures				
	Mean	S.D.	Min	Median	Max
Arbitrary Priors					
1.0×10^{-4}	94.167	22.99	43	103	121
9.7×10^{-8}	60.611	24.126	33	54	107
Ascending Arbitrary Priors					
1.0×10^{-4}	93.972	24.241	43	104	121
9.7×10^{-8}	60.917	24.044	33	56	107
Descending Arbitrary Priors					
1.0×10^{-4}	92.806	23.833	42	103	122
9.7×10^{-8}	60.611	24.126	33	54	107
Data-Based Proportional Priors					
1.0×10^{-4}	92.667	24.299	41	103	121
9.7×10^{-8}	60.611	24.126	33	54	107
Equal Priors					
1.0×10^{-4}	92.444	24.305	41	103	121
9.7×10^{-8}	60.611	24.126	33	54	107

Table 4.4: *10-fold Cross-validated Classification positions for the Simulated data set with 2% within-species variability.*

δ value. The average number of positions required for classification using $\delta = 1.0 \times 10^{-4}$ ranges from 129.083 in the case of arbitrary priors and 133.5 in the case of descending arbitrary priors. The average number of positions required for classification using $\delta = 9.7 \times 10^{-8}$ is 87.521 across all choices of prior probabilities. The average number of positions required for classification using $\delta = 9.7 \times 10^{-8}$ is less than the average number required using $\delta = 1.0 \times 10^{-4}$ as in the case with 2% within-species variability, but the average difference in the number of positions required for classification between the two δ values has increased about 31% to 42.792.

Table 4.6 gives the average number of positions required for classification when the within-species variability is 6%. The number of positions required is stable over the choice of priors ranging from 190.75 in the case of equal priors to 193.875 in the case of descending

	Overall Position Classification Measures				
δ	Mean	S.D.	Min	Median	Max
Arbitrary Priors					
1.0×10^{-4}	129.083	62.123	43	118.5	301
9.7×10^{-8}	87.521	39.828	33	84.5	176
Ascending Arbitrary Priors					
1.0×10^{-4}	129.688	62.599	43	120	301
9.7×10^{-8}	87.521	39.828	33	84.5	176
Descending Arbitrary Priors					
1.0×10^{-4}	133.5	64.894	43	121	301
9.7×10^{-8}	87.521	39.828	33	84.5	176
Data-Based Proportional Priors					
1.0×10^{-4}	129.896	62.735	43	120	301
9.7×10^{-8}	87.521	39.828	33	84.5	176
Equal Priors					
1.0×10^{-4}	129.396	62.707	43	118.5	3012
9.7×10^{-8}	87.521	39.828	33	84.5	176

Table 4.5: *10-fold Cross-validated Classification positions for the Simulated data set with* 4% *within-species variability.*

arbitrary priors for $\delta = 1.0 \times 10^{-4}$. For $\delta = 9.7 \times 10^{-8}$, the average number of positions required for classification ranges from 112.05 in the case of equal priors, and 114.979 in the case of descending arbitrary priors. On average, the number of positions required for classification is 77.623 positions fewer using $\delta = 9.7 \times 10^{-8}$ than for the arbitrarily selected $\delta = 1.0 \times 10^{-4}$.

Table 4.7 gives the overall measures for the number of positions required for classification 8% within-species variablility. Again we see that across the various choices of prior probabilities, the average number of positions required for classification remains somewhat stable ranging from 284.146 in the case of arbitrary priors to 299.479 in the case of ascending arbitrary priors for $\delta = 1.0 \times 10^{-4}$. For $\delta = 9.7 \times 10^{-8}$, the average number of positions required for classfication is also very stable ranging from 139.063 in the case of equal priors

	Overall Position Classification Measures				
δ	Mean	S.D.	Min	Median	Max
Arbitrary Priors					
1.0×10^{-4}	192.063	121.621	78	152.5	601
9.7×10^{-8}	114.958	101.297	39	92	588
Ascending Arbitrary Priors					
1.0×10^{-4}	191.271	122.155	78	152.5	601
9.7×10^{-8}	114.771	101.441	34	92	588
Descending Arbitrary Priors					
1.0×10^{-4}	193.875	122.863	78	149.5	601
9.7×10^{-8}	114.979	101.387	34	92	588
Data-Based Proportional Priors					
1.0×10^{-4}	191.854	121.588	78	148	601
9.7×10^{-8}	114.938	101.39	34	92	588
Equal Priors					
1.0×10^{-4}	190.75	122.386	78	148	601
9.7×10^{-8}	112.05	93.137	34	98.5	588

Table 4.6: *10-fold Cross-validated Classification positions for the Simulated data set with* 6% *within-species variability.*

to 143.833 in the case of arbitrary priors. On average, choosing $\delta = 9.7 \times 10^{-8}$ results in 153.404 fewer positions used than that of the arbitrarily selected $\delta = 1.0 \times 10^{-4}$.

The average number of positions required for classification appears to be somewhat robust to the choice of initial prior probabilities, but it appears to depend on the amount of within-species variability as well as the choice of δ. It is not surprising that the within-species variability has such an effect on the average number of positions required for classification. As the barcodes within a species have more variation, distinguishing between barcodes will become more difficult, a point we observed in the misclassification rates that is also reflected here in terms of the number of positions used. It is also interesting to note that, not only do the average number of positions required for classification increase as the amount of within-species variability increases, but the difference in the average number of positions used for

	Overall Position Classification Measures				
δ	Mean	S.D.	Min	Median	Max
Arbitrary Priors					
1.0×10^{-4}	284.146	186.314	78	222	700
9.7×10^{-8}	143.833	87.075	47	116	363
Ascending Arbitrary Priors					
1.0×10^{-4}	299.479	203.877	61	222	700
9.7×10^{-8}	139.542	83.537	47	113.5	363
Descending Arbitrary Priors					
1.0×10^{-4}	297.833	192.665	78	226	700
9.7×10^{-8}	139.229	83.672	47	113.5	363
Data-Based Proportional Priors					
1.0×10^{-4}	294.125	194.042	78	222	700
9.7×10^{-8}	139.083	83.321	47	113.5	363
Equal Priors					
1.0×10^{-4}	292.188	193.715	78	222	700
9.7×10^{-8}	139.063	83.33	47	113.5	363

Table 4.7: *10-fold Cross-validated Classification positions for the Simulated data set with* 8% *within-species variability.*

the two δ values also increases. This means that, as the within-species variability increases, the average number of positions required for the two δ values do not increase at the same rate. Using the value for δ based on the mutation rate increases the average number of positions at a slower rate than is observed for the arbitrarily selected δ value.

4.1.2 Simulation: Discovery

To evaluate species discovery with the proposed method, we removed each species one at a time from the reference data set and then sought a classification of the four barcodes from the species that was removed. The proposed method was implemented using the recommended equal prior probabilities together with the recommended $\delta = 9.7 \times 10^{-8}$. Plots of the posterior probabilities were created and examined for each classification. Figure 4.1 gives

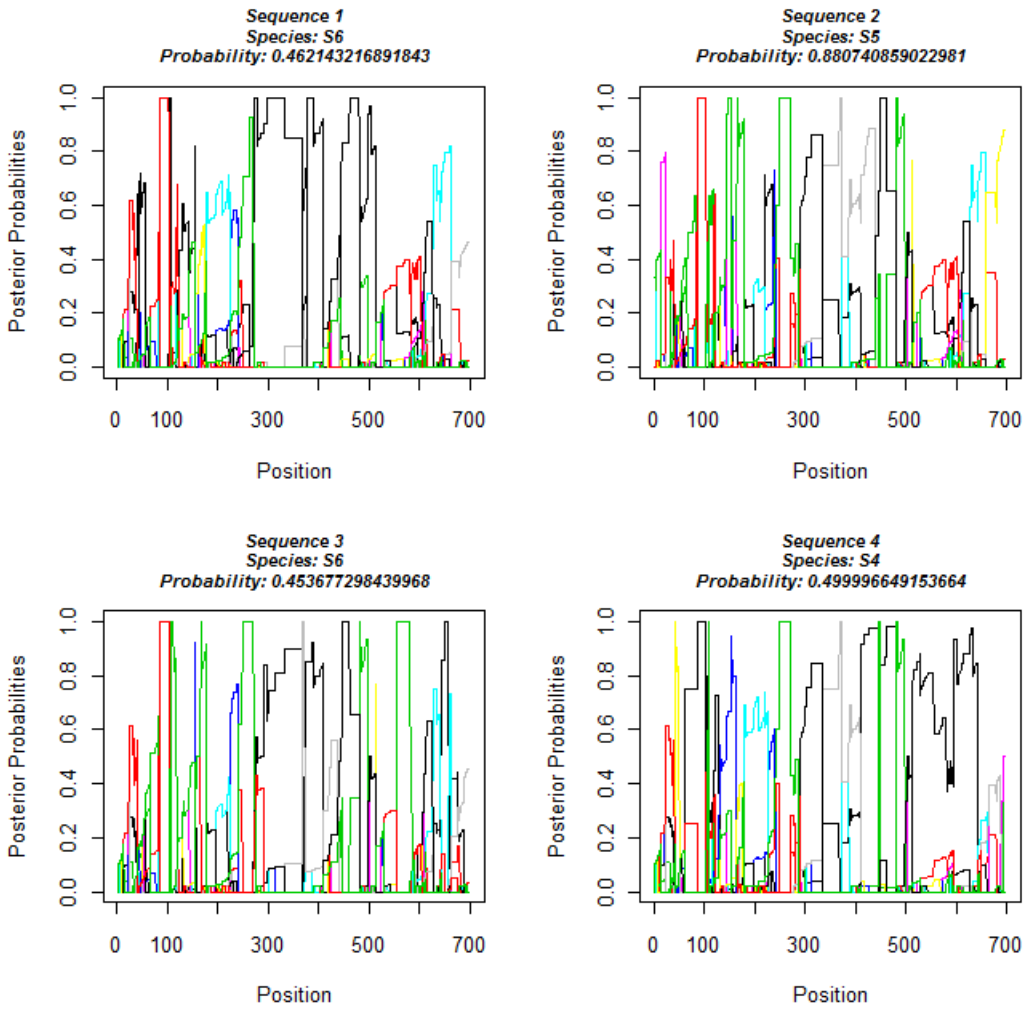

Figure 4.1: *Plotted posterior probabilities having removed species 3 from the reference data set and seeking a classification of the four barcodes belong to species 3. Equal prior proabilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 2% *were used.*

posterior probability plots of the four sequences in species 3 having within-species variability of 2% and having completely removed that species from the reference data set. The main title of these plots gives the sequence within species 3 being classified, the species in the reference data set with the highest posterior probability at the conclusion of the calculations, and the posterior probability at the termination of the calculations. We see that in these plots, no one species in the reference data set was favored for an extended period of time,

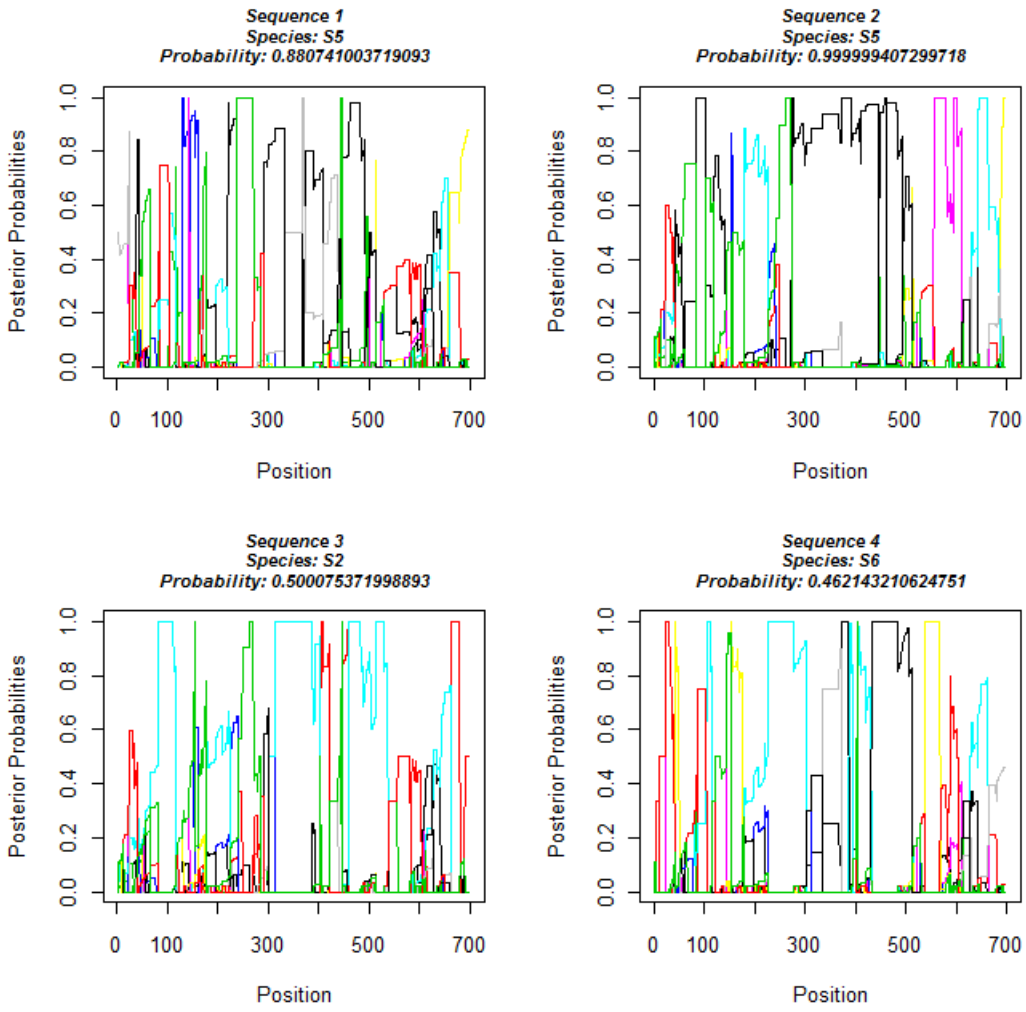

Figure 4.2: *Plotted posterior probabilities having removed species 3 from the reference data set and seeking a classification of the four barcodes belong to species 3. Equal prior proabilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 4% *were used.*

giving the plot a very noisy look. We also see that the stopping rule was not triggered in any of these plots, meaning that the posterior probability of belonging to any one species in the reference data set, while high at times, never got close enough to one to terminate the calculations.

Figure 4.2 contains the posterior probability plots for the four sequences belonging to species 3 when the within-species variability was increased to 4%. The plots are still very

noisy but with somewhat less fluctuation between species than was observed with 2% within-species variability. This decrease in fluctuation between species with the highest posterior probability is an indication that it is more challenging for the proposed method to distinguish between the species in the reference data set, as would be expected with an increase in the within-species variability. Also, the stopping rule was not engaged for any of the four sequences being classified.

When we increased the within-species variability to 6%, the plots in Figure 4.3 were obtained. Here we see even less fluctuation between species with the highest posterior probability than in the case with 4% within-species variability, indicating as before that the species in the reference data set are looking more similar. We also point out that at this level of within-species variability, the proposed method is starting to encounter the stopping rule. For example, in sequence 1 of species 3, the calculations stop around the 400th position and the barcode was assigned at that point to species 1. This means that around the 400th position, the posterior probability of the barcode belonging to species 1 was one. The stopping rule was encountered for sequence 3 around the 220th position and for sequence 4 around the 140th position.

By increasing the within-species variability further to 8%, the posterior probability plots in Figure 4.4 show a dramatic reduction in the fluctuation between species with the highest posterior probability, and the stopping rule was encountered for each of the four sequences. Clearly, it is more challenging for the proposed method to distinguish among species when the within-species variability is approximately equal to the among-species variability. Plots of posterior probabilities for every observation in each of the 12 species at the 2, 4, 6, and 8% levels of within-species variability can be found in Figures C.1-C.48 in Appendix C.

From this experiment, we conclude that the plotted posterior probabilities are vital to species discovery. We see that when a barcode to be classified does not belong to a species in the reference data set, these noisy posterior probability plots are typical. Obvious exceptions to this include sequences 2 and 3 in Figure C.2 with 4% within-species variability, sequences

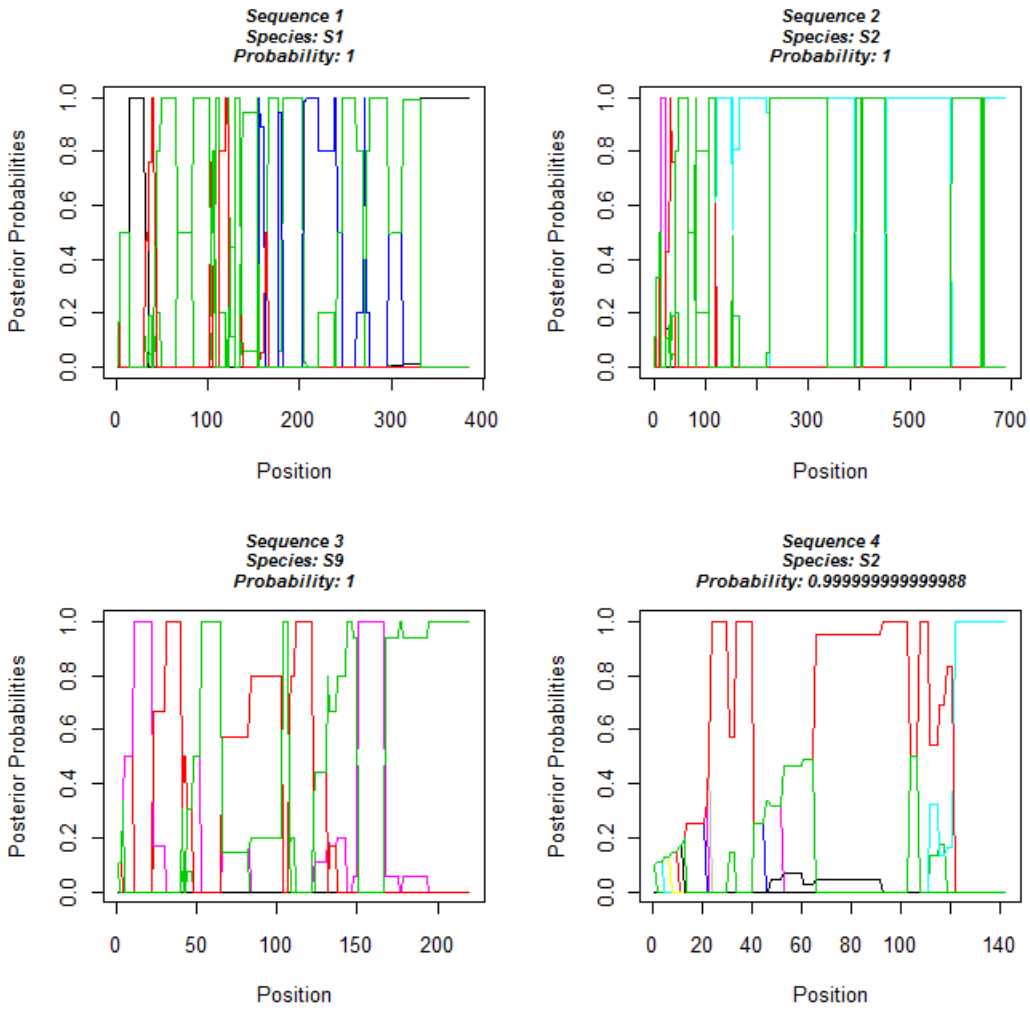

Figure 4.3: *Plotted posterior probabilities having removed species 3 from the reference data set and seeking a classification of the four barcodes belong to species 3. Equal prior proabilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 6% *were used.*

1 and 2 in Figure C.6 with 4% within-species variability, sequence 1 in Figure C.7 with 6% within-species variability, sequences 1, 3, and 4 in Figure C.8 with 8% within-species variability, and sequence 4 in Figure C.35 with 6% within-species variability. In these rare cases, the posterior probability plots did not exhibit the noise typical to these plots when the barcode to be classified does not belong to a species in the reference data set. One reason for this could be the higher than typical levels of within-species variability in each

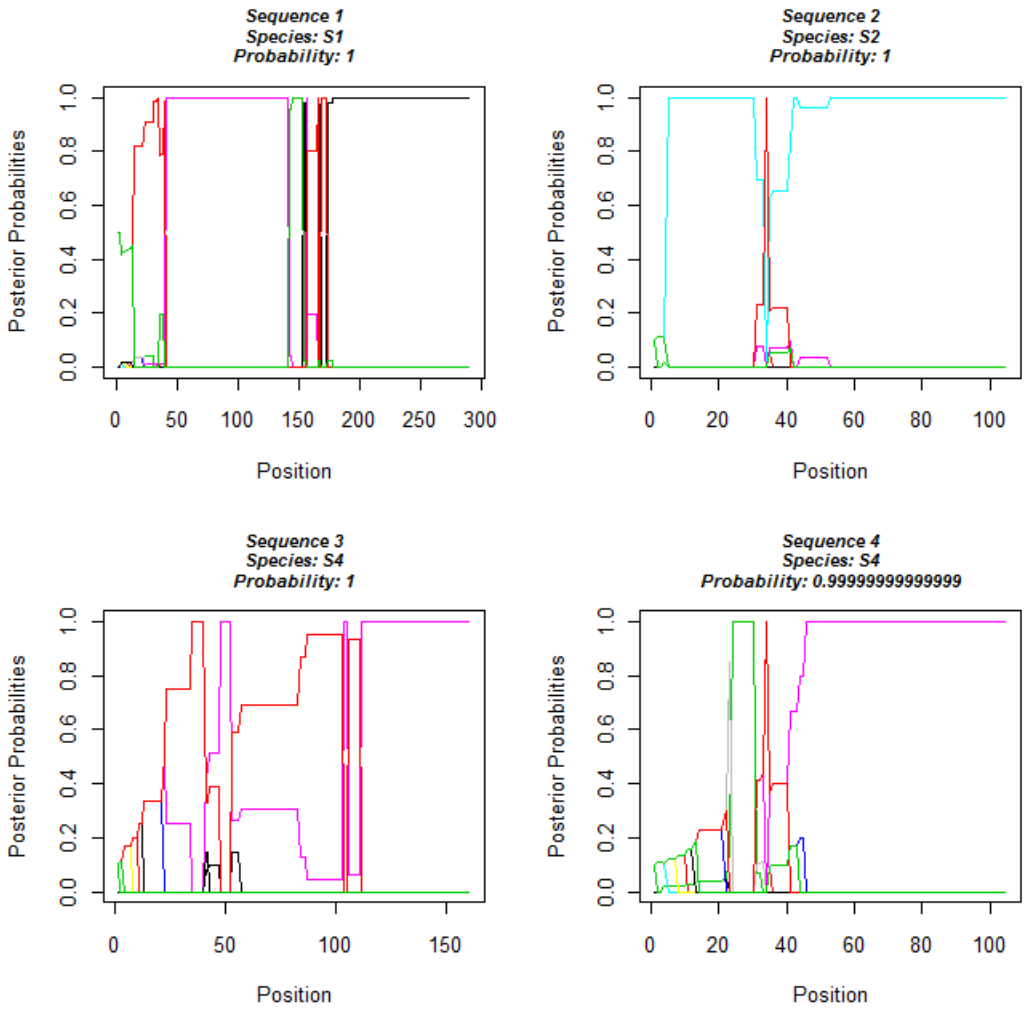

Figure 4.4: *Plotted posterior probabilities having removed species 3 from the reference data set and seeking a classification of the four barcodes belong to species 3. Equal prior proabilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 8% *were used.*

case.

In contrast to the posterior probability plots above, we carried out the current method using the simulated data set with 2% within-species variability in order to classify the four barcodes belonging to Species 3. Figure 4.5 gives the phylogenetic trees obtained by using K2P distances together with the neighbor joining tree for each of the sequences belonging to Species 3 when Species 3 is not represented in the reference data set. The sequences we seek

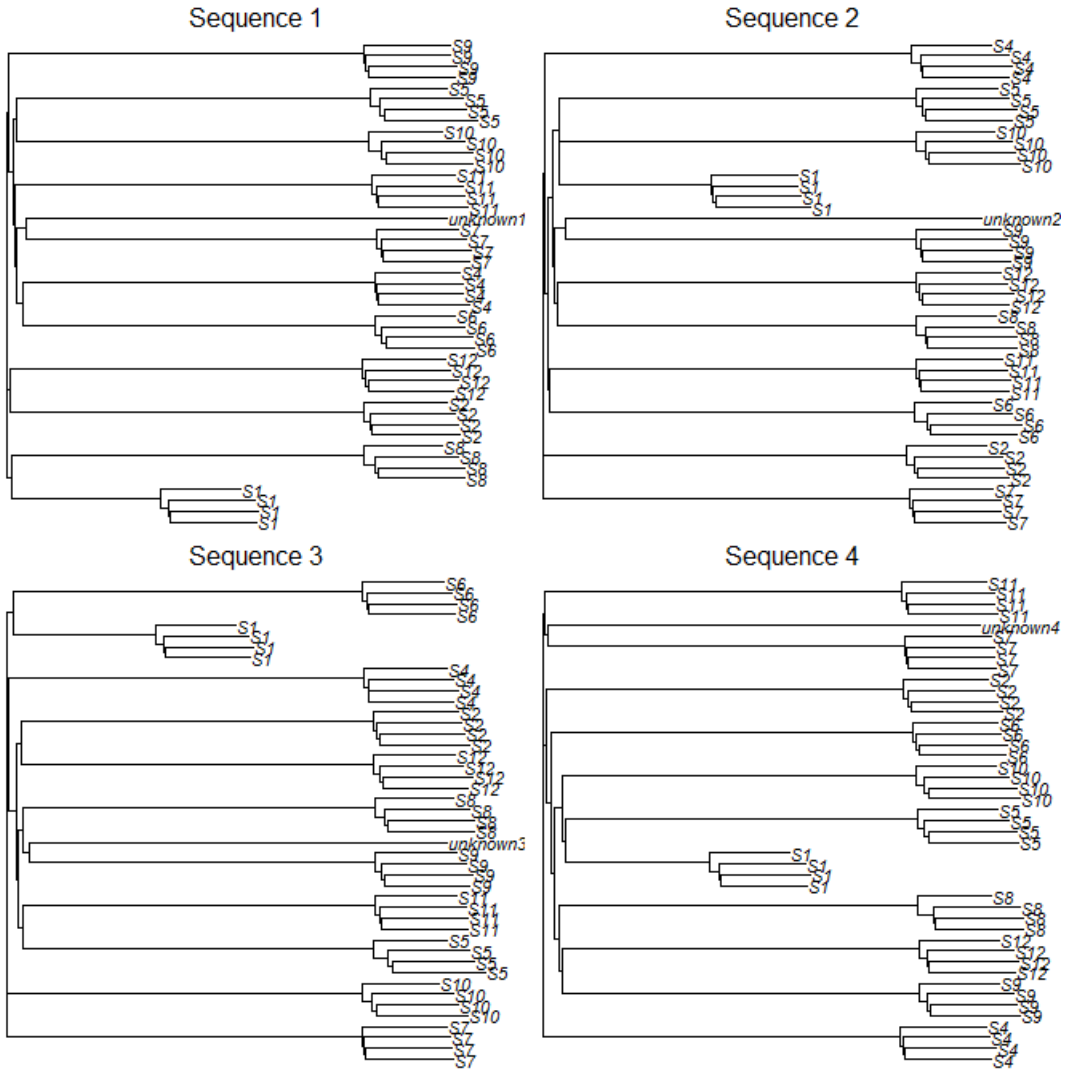

Figure 4.5: *Phylogenetic trees having removed Species 3 (S_3) from the reference data set and seeking a classification of the four barcodes belong to S_3. The K2P distances together with the neighbor-joining method were used.*

to classify appear in the trees labeled *unknown 1*, *unknown 2*, *unknown 3*, and *unknown 4*. We see that the phylogenetic trees provide no indication that the barcode to be classified may not belong to any of the species in the reference data set. Because the current methods classify the barcode as belonging to the species of its nearest neighbor in the tree, regardless of the distance between them (Frézal and Leblois 2008), sequences 1 and 4 are mistakenly classified as belonging to Species 7, while sequences 2 and 3 are mistakenly classified as

belonging to Species 9. The proposed method classifies these four barcodes to Species 6, 5, 6, and 4, respectively, but the plotted posterior probabilities provide a clear indication that these four barcodes do not, in fact, belong to any of the species in the reference data set. As seen in Figure 4.5, current methods lack the ability address species discovery while Figure 4.1 shows that, the proposed method readily identifies potentially new species.

4.2 Results of Proposed Method with Real Data

Five data sets, one representing bats (Bat), two representing birds (Bird1 and Bird2), one representing butterflies (Butterfly), and one representing fish (Fish), containing barcode data were extracted from BOLD and analyzed using the proposed method. Each data set was partitioned into ten "test" data sets with each one accounting for about 10% of the data in the original data set. The effectiveness of the proposed method was then analyzed by using the remaining 90% of the observations as the reference data set R, from which the conditional probabilities were obtained, and then by classifying all of the observations in the test data set T. The average misclassification rates for all ten test data sets were then recorded. The splitting of each data set consisted of randomly selecting the desired percentage of observations to make up the test data set. After this randomization, care was taken to ensure that each species had at least one representative in the reference data set R. The barcodes in the test data set were then classified using the proposed method with the missing data being imputed via the recommended proportional allocation approach. We selected for our priors:

1. Arbitrary prior probabilities randomly generated from a non-informative Dirichlet distribution.

2. Ascending arbitrary prior probabilities which are the priors from (1) sorted in ascending order.

3. Descending arbitrary prior probabilities which are the priors from (1) sorted in de-

scending order.

4. Data-based proportional prior probabilities based on the prevalence of each species in the reference data set.

5. Equal prior probabilities.

We also evaluated the performance of the proposed method on two different choices of δ. One is the arbitrarily selected value of $\delta = 1.0 \times 10^{-4}$, and the other is the estimated mutation rate of the mitochondrial genome of $\delta = 9.7 \times 10^{-8}$. We also chose to evaluate the proposed method by obtaining the misclassification rates when classification of the barcodes starts from the first given position, as would typically be done, and comparing those to the misclassification rates obtained when classification of the barcodes was done by randomizing the order of the positions. This is important for two reasons. First, it will allow us to determine if the proposed method will work if classification starts somewhere other than the first position. Second, it will allow us to address the assumption of independence among the nucleotide positions. By randomizing the order in which the barcode is to be analyzed, we break the dependence structure that may have existed among the nucleotides. More on this will be discussed in Section 4.3. In addition to evaluating how well the proposed method performs in terms of misclassification rates in Tables 4.8, 4.10, 4.12, 4.14, and 4.16, we also evaluate how well it performs in terms of time and number of positions required for classification in Tables 4.9, 4.11, 4.13, 4.15, and 4.17.

4.2.1 Misclassification Rates

Generally speaking, the average misclassification rates were robust to the choice of priors, δ values, and starting positions. In an ANOVA of the misclassification rates, treating the test data sets as blocks in a randomized complete block design, only the Bird2 and Butterfly data sets gave significant F tests (both had p-values$<$0.0001). For these two data sets, the misclassification rates tended to be lower for $\delta = 9.7 \times 10^{-8}$ and for the

Table 4.8: *10-fold Cross-validated Misclassification rates for the Bat data set. For the individual misclassification rates for groups 1,2,3,5,6,7,8,9, group 4, and group 10; the number of observations in the reference data set R were 758, 759, and 751, respectively and the number of observations in the test data set T were 82, 81, and 89, respectively.*

Starting		Overall Misclassification Measures					Individual Misclassification Rates									
Position	δ	Mean	S.D.	Min	Median	Max	1	2	3	4	5	6	7	8	9	10
Arbitrary Priors																
First	1.0×10^{-4}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
First	9.7×10^{-8}	0.001	0.004	0	0	0.012	0	0	0	0.012	0	0	0	0	0	0
Random	1.0×10^{-4}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Random	9.7×10^{-8}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Ascending Arbitrary Priors																
First	1.0×10^{-4}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
First	9.7×10^{-8}	0.001	0.004	0	0	0.012	0	0	0	0.012	0	0	0	0	0	0
Random	1.0×10^{-4}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Random	9.7×10^{-8}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Descending Arbitrary Priors																
First	1.0×10^{-4}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
First	9.7×10^{-8}	0.001	0.004	0	0	0.012	0	0	0	0.012	0	0	0	0	0	0
Random	1.0×10^{-4}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Random	9.7×10^{-8}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Data-Based Proportional Priors																
First	1.0×10^{-4}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
First	9.7×10^{-8}	0.001	0.004	0	0	0.012	0	0	0	0.012	0	0	0	0	0	0
Random	1.0×10^{-4}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Random	9.7×10^{-8}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Equal Priors																
First	1.0×10^{-4}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
First	9.7×10^{-8}	0.001	0.004	0	0	0.012	0	0	0	0.012	0	0	0	0	0	0
Random	1.0×10^{-4}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Random	9.7×10^{-8}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Neighbor-joining with K2P model																
		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 4.9: *10-fold Cross-validated Classification times and positions for the Bat data set.*

Starting Position	δ	Overall Time Classification Measures (min)					Overall Position Classification Measures				
		Mean	S.D.	Min	Median	Max	Mean	S.D.	Min	Median	Max
Arbitrary Priors											
First	1.0×10^{-4}	3.627	0.163	3.527	3.584	4.087	145.647	150.886	26	56	599
First	9.7×10^{-8}	3.164	0.079	3.019	3.154	3.279	98.868	103.712	25	44	554
Random	1.0×10^{-4}	3.103	0.011	3.031	3.060	3.326	116.144	117.362	27	54	659
Random	9.7×10^{-8}	2.986	0.087	2.895	2.956	3.132	81.929	83.225	15	46	503
Ascending Arbitrary Priors											
First	1.0×10^{-4}	2.927	0.043	2.885	2.908	3.030	145.871	153.896	28	56	572
First	9.7×10^{-8}	2.934	0.074	2.834	2.919	3.087	98.944	103.289	26	44	554
Random	1.0×10^{-4}	2.912	0.021	2.879	2.915	2.940	116.760	117.646	27	54	659
Random	9.7×10^{-8}	2.843	0.047	2.774	2.836	2.920	82.455	86.957	15	46	503
Descending Arbitrary Priors											
First	1.0×10^{-4}	3.001	0.075	2.936	2.975	3.140	148.121	154.048	26	56	572
First	9.7×10^{-8}	2.928	0.036	2.879	2.925	3.018	98.602	102.896	25	44	554
Random	1.0×10^{-4}	2.806	0.035	2.765	2.796	2.880	117.644	117.790	26	54	659
Random	9.7×10^{-8}	2.813	0.026	2.774	2.812	2.844	81.777	85.710	15	46	503
Data-Based Proportional Priors											
First	1.0×10^{-4}	2.985	0.067	2.886	3.002	3.092	144.306	153.292	25	56	572
First	9.7×10^{-8}	2.877	0.065	2.766	2.885	2.983	98.678	103.428	25	44	554
Random	1.0×10^{-4}	2.805	0.034	2.736	2.817	2.843	115.200	117.594	26	53	659
Random	9.7×10^{-8}	2.803	0.016	2.772	2.806	2.820	81.168	83.440	15	46	503
Equal Priors†											
First	1.0×10^{-4}	3.034	0.078	2.906	3.044	3.172	147.634^{d}	153.321	27	56	572
First	9.7×10^{-8}	3.299	0.067	3.174	3.308	3.421	98.744^{b}	103.122	27	44	554
Random	1.0×10^{-4}	2.810	0.023	2.782	2.804	2.848	117.954^{c}	117.535	31	54	659
Random	9.7×10^{-8}	2.803	0.024	2.773	2.802	2.841	82.473^{a}	86.924	15	46	503

† position averages with different superscripts represent significant differences at the 0.05 level.

Table 4.10: *10-fold Cross-validated Misclassification rates for the Bird1 data set. For the individual misclassification rates for groups 1-9, and group 10; the number of observations in the reference data set R were 1461, and 1458, respectively and the number of observations in the test data set T were 162, and 165, respectively.*

Starting		Overall Misclassification Measures					Individual Misclassification Rates									
Position	δ	Mean	S.D.	Min	Median	Max	1	2	3	4	5	6	7	8	9	10
Arbitrary Priors																
First	1.0×10^{-4}	0.002	0.003	0	0	0.006	0	0	0	0.006	0	0.006	0	0	0.006	0
First	9.7×10^{-8}	0.003	0.004	0	0	0.012	0	0	0	0.006	0	0.006	0	0	0.012	0
Random	1.0×10^{-4}	0.002	0.003	0	0	0.006	0	0	0	0.006	0	0.006	0	0	0.006	0
Random	9.7×10^{-8}	0.003	0.004	0	0	0.012	0	0	0	0.006	0	0.012	0	0	0.006	0
Ascending Arbitrary Priors																
First	1.0×10^{-4}	0.002	0.003	0	0	0.006	0	0	0	0.006	0	0.006	0	0	0.006	0
First	9.7×10^{-8}	0.003	0.004	0	0	0.012	0	0	0	0.006	0	0.006	0	0	0.012	0
Random	1.0×10^{-4}	0.002	0.003	0	0	0.006	0	0	0	0.006	0	0.006	0	0	0.006	0
Random	9.7×10^{-8}	0.004	0.004	0	0.003	0.012	0.006	0	0	0.006	0	0.012	0	0.006	0.006	0
Descending Arbitrary Priors																
First	1.0×10^{-4}	0.002	0.003	0	0	0.006	0	0	0	0.006	0	0.006	0	0	0.006	0
First	9.7×10^{-8}	0.003	0.004	0	0	0.012	0	0	0	0.006	0	0.006	0	0	0.012	0
Random	1.0×10^{-4}	0.002	0.003	0	0	0.006	0	0	0	0.006	0	0.006	0	0	0.006	0
Random	9.7×10^{-8}	0.003	0.004	0	0	0.012	0	0	0	0.006	0	0.012	0	0.006	0.006	0
Data-Based Proportional Priors																
First	1.0×10^{-4}	0.002	0.003	0	0	0.006	0	0	0	0.006	0	0.006	0	0	0.006	0
First	9.7×10^{-8}	0.003	0.004	0	0	0.012	0	0	0	0.006	0	0.006	0	0	0.012	0
Random	1.0×10^{-4}	0.002	0.003	0	0	0.006	0	0	0	0.006	0	0.006	0	0	0.006	0
Random	9.7×10^{-8}	0.003	0.004	0	0	0.012	0	0	0	0.006	0	0.012	0	0.006	0.006	0
Equal Priors																
First	1.0×10^{-4}	0.002	0.003	0	0	0.006	0	0	0	0.006	0	0.006	0	0	0.006	0
First	9.7×10^{-8}	0.003	0.004	0	0	0.012	0	0	0	0.006	0	0.006	0	0	0.012	0
Random	1.0×10^{-4}	0.002	0.003	0	0	0.006	0	0	0	0.006	0	0.006	0	0	0.006	0
Random	9.7×10^{-8}	0.003	0.004	0	0	0.012	0	0	0	0.006	0	0.012	0	0.006	0.006	0
Neighbor-joining with K2P model																
		0.001	0.002	0	0	0.006	0	0	0	0	0	0	0	0	0.006	0

Table 4.11: *10-fold Cross-validated Classification times and positions for the Bird1 data set.*

Starting Position	δ	Overall Time Classification Measures (min)					Overall Position Classification Measures				
		Mean	S.D.	Min	Median	Max	Mean	S.D.	Min	Median	Max
Arbitrary Priors											
First	1.0×10^{-4}	4.879	0.096	4.800	4.841	5.087	118.499	93.610	34	94	690
First	9.7×10^{-8}	4.410	0.091	4.329	4.394	4.655	84.327	65.790	22	70	690
Random	1.0×10^{-4}	5.012	0.088	4.901	4.993	5.239	156.283	131.559	39	105	690
Random	9.7×10^{-8}	4.953	0.049	4.867	4.953	5.027	111.140	102.151	23	71	690
Ascending Arbitrary Priors											
First	1.0×10^{-4}	4.651	0.173	4.416	4.616	5.006	117.797	91.298	34	94	690
First	9.7×10^{-8}	4.346	0.086	4.268	4.305	4.536	83.890	62.669	22	70	690
Random	1.0×10^{-4}	4.969	0.070	4.872	4.961	5.124	156.526	130.776	40	104	690
Random	9.7×10^{-8}	5.327	0.355	4.934	5.355	5.756	111.472	102.057	23	71	690
Descending Arbitrary Priors											
First	1.0×10^{-4}	5.066	0.442	4.616	5.054	5.572	118.001	90.498	34	94	690
First	9.7×10^{-8}	4.570	0.169	4.265	4.582	4.811	84.614	66.161	16	70	690
Random	1.0×10^{-4}	5.632	0.181	5.206	5.646	5.861	156.903	131.845	39	104	690
Random	9.7×10^{-8}	5.667	0.085	5.600	5.644	5.889	112.147	102.621	23	71	690
Data-Based Proportional Priors											
First	1.0×10^{-4}	4.982	0.097	4.893	4.944	5.179	117.482	91.280	34	94	690
First	9.7×10^{-8}	4.349	0.090	4.269	4.316	4.510	83.988	62.494	19	73	690
Random	1.0×10^{-4}	5.135	0.063	5.050	5.114	5.257	155.619	130.902	39	105	690
Random	9.7×10^{-8}	5.150	0.074	5.037	5.183	5.230	112.207	102.779	23	71	690
Equal Priors†											
First	1.0×10^{-4}	4.589	0.212	4.297	4.576	4.966	116.986^{b}	88.181	34	94	690
First	9.7×10^{-8}	4.848	0.112	4.749	4.818	5.119	84.490^{a}	64.340	22	73	690
Random	1.0×10^{-4}	4.726	0.193	4.470	4.703	5.232	156.512^{c}	131.051	40	104	690
Random	9.7×10^{-8}	4.472	0.070	4.392	4.453	4.604	111.925^{b}	102.001	23	71	690

† position averages with different superscripts represent significant differences at the 0.05 level.

Table 4.12: *10-fold Cross-validated Misclassification rates for the Bird2 data set. For the individual misclassification rates for groups 1-9, and group 10 the number of observations in the reference data set R were 2344, and 2350, respectively while the number of observations in the test data set T were 220, 217, 216, 261, 221, 220, 217, 219, 219, and 213 for groups 1-10, respectively.*

Starting		Overall Misclassification Measures					Individual Misclassification Rates									
Position	δ	Mean†	S.D.	Min	Median	Max	1	2	3	4	5	6	7	8	9	10
Arbitrary Priors																
First	1.0×10^{-4}	0.029^{c}	0.015	0.005	0.035	0.046	0.041	0.046	0.047	0.033	0.014	0.041	0.005	0.018	0.018	0.038
First	9.7×10^{-8}	0.025^{bc}	0.016	0	0.028	0.041	0.041	0.041	0.037	0.023	0	0.041	0	0.018	0.018	0.033
Random	1.0×10^{-4}	0.031^{c}	0.015	0.005	0.037	0.046	0.041	0.046	0.046	0.037	0.018	0.041	0.005	0.018	0.018	0.038
Random	9.7×10^{-8}	0.026^{bc}	0.015	0	0.030	0.047	0.041	0.037	0.042	0.028	0.005	0.036	0	0.018	0.018	0.033
Ascending Arbitrary Priors																
First	1.0×10^{-4}	0.024^{ab}	0.015	0.005	0.023	0.050	0.050	0.028	0.032	0.019	0.014	0.036	0.005	0.014	0.005	0.038
First	9.7×10^{-8}	0.020^{a}	0.017	0	0.021	0.050	0.050	0.028	0.028	0.009	0	0.036	0	0.014	0.005	0.033
Random	1.0×10^{-4}	0.025^{b}	0.015	0.005	0.025	0.050	0.050	0.032	0.032	0.019	0.018	0.036	0.005	0.014	0.005	0.038
Random	9.7×10^{-8}	0.020^{a}	0.016	0	0.018	0.050	0.050	0.023	0.028	0.009	0.005	0.032	0	0.014	0.005	0.033
Descending Arbitrary Priors																
First	1.0×10^{-4}	0.026^{bc}	0.012	0.005	0.032	0.041	0.041	0.032	0.037	0.033	0.018	0.032	0.005	0.018	0.014	0.033
First	9.7×10^{-8}	0.023^{ab}	0.013	0	0.026	0.041	0.041	0.032	0.032	0.023	0.005	0.032	0	0.018	0.014	0.028
Random	1.0×10^{-4}	0.028^{bc}	0.012	0.005	0.032	0.042	0.041	0.037	0.042	0.033	0.023	0.032	0.005	0.018	0.014	0.033
Random	9.7×10^{-8}	0.023^{ab}	0.013	0	0.025	0.041	0.041	0.028	0.037	0.023	0.009	0.027	0	0.018	0.018	0.028
Data-Based Proportional Priors																
First	1.0×10^{-4}	0.026^{bc}	0.013	0.005	0.023	0.050	0.050	0.018	0.037	0.028	0.018	0.032	0.005	0.018	0.018	0.033
First	9.7×10^{-8}	0.021^{ab}	0.013	0	0.018	0.046	0.046	0.014	0.028	0.019	0.005	0.032	0	0.018	0.018	0.029
Random	1.0×10^{-4}	0.026^{bc}	0.013	0.005	0.025	0.050	0.050	0.018	0.037	0.028	0.023	0.032	0.005	0.018	0.018	0.033
Random	9.7×10^{-8}	0.020^{a}	0.013	0	0.018	0.046	0.046	0.009	0.032	0.019	0.009	0.027	0	0.018	0.018	0.028
Equal Priors																
First	1.0×10^{-4}	0.027^{bc}	0.012	0.005	0.032	0.046	0.046	0.032	0.037	0.033	0.018	0.032	0.005	0.018	0.018	0.033
First	9.7×10^{-8}	0.023^{ab}	0.014	0	0.026	0.041	0.041	0.032	0.037	0.023	0.005	0.032	0	0.018	0.018	0.028
Random	1.0×10^{-4}	0.028^{bc}	0.012	0.005	0.032	0.042	0.041	0.037	0.042	0.033	0.023	0.032	0.005	0.018	0.018	0.033
Random	9.7×10^{-8}	0.022^{ab}	0.012	0	0.025	0.041	0.041	0.028	0.032	0.023	0.009	0.027	0	0.018	0.018	0.028
Neighbor-joining with K2P model‡																
		0.019	0.010	0.005	0.016	0.041	0.041	0.023	0.028	0.023	0.014	0.009	0.005	0.018	0.014	0.014

† misclassification averages with different superscripts represent significant differences at the 0.05 level.

‡ 1% of the observations were removed so K2P model could estimate distances.

Table 4.13: *10-fold Cross-validated Classification times and positions for the Bird2 data set.*

Starting Position	δ	Overall Time Classification Measures (min)					Overall Position Classification Measures				
		Mean	S.D.	Min	Median	Max	Mean	S.D.	Min	Median	Max
Arbitrary Priors											
First	1.0×10^{-4}	7.439	0.319	7.185	7.300	8.203	181.950	61.041	72	170	255
First	9.7×10^{-8}	7.213	0.384	6.840	7.094	8.103	156.195	63.055	66	134	255
Random	1.0×10^{-4}	8.594	0.192	8.450	8.490	9.016	175.175	68.234	49	178	255
Random	9.7×10^{-8}	7.489	0.641	6.815	7.147	8.465	138.393	72.224	35	102	255
Ascending Arbitrary Priors											
First	1.0×10^{-4}	6.948	0.245	6.665	6.876	7.497	183.151	60.697	72	173	255
First	9.7×10^{-8}	6.920	0.413	6.411	6.955	7.415	156.438	63.117	68	137	255
Random	1.0×10^{-4}	6.673	0.111	6.525	6.648	6.856	175.987	67.549	46	178	255
Random	9.7×10^{-8}	6.643	0.043	6.569	6.648	6.714	138.584	72.376	35	102	255
Descending Arbitrary Priors											
First	1.0×10^{-4}	8.882	1.634	7.874	8.024	12.859	182.480	61.030	72	170	255
First	9.7×10^{-8}	7.817	0.269	7.295	7.926	8.019	156.323	63.141	68	134	255
Random	1.0×10^{-4}	6.655	0.068	6.515	6.661	6.739	175.7560	68.136	50	180	255
Random	9.7×10^{-8}	6.630	0.049	6.533	6.647	6.686	138.608	72.420	35	102	255
Data-Based Proportional Priors											
First	1.0×10^{-4}	7.445	0.245	7.099	7.486	7.946	182.628	60.584	72	173	255
First	9.7×10^{-8}	6.597	0.127	6.479	6.573	6.920	156.359	63.112	68	135	255
Random	1.0×10^{-4}	6.661	0.045	6.594	6.665	6.722	175.529	67.591	49	178	255
Random	9.7×10^{-8}	6.628	0.048	6.550	6.636	6.683	138.535	72.406	35	102	255
Equal Priors†											
First	1.0×10^{-4}	7.184	0.396	6.852	6.942	7.906	183.229^{d}	60.560	72	173	255
First	9.7×10^{-8}	7.216	0.262	6.990	7.125	7.838	156.376^{b}	63.090	68	135	255
Random	1.0×10^{-4}	6.710	0.121	6.604	6.674	7.033	176.837^{c}	67.215	50	180	255
Random	9.7×10^{-8}	6.672	0.059	6.580	6.696	6.753	138.523^{a}	72.390	35	102	255

† position averages with different superscripts represent significant differences at the 0.05 level.

Table 4.14: *10-fold Cross-validated Misclassification rates for the Butterfly data set. For the individual misclassification rates for groups 1-9, and group 10 the number of observations in the reference data set R were 3868, and 3865, respectively while the number of observations in the test data set T were 356, 351, 359, 353, 359, 349, 353, 356, 355, and 351 for groups 1-10, respectively.*

Starting		Overall Misclassification Measures					Individual Misclassification Rates									
Position	δ	Mean†	S.D.	Min	Median	Max	1	2	3	4	5	6	7	8	9	10
Arbitrary Priors																
First	1.0×10^{-4}	0.007^{c}	0.004	0.003	0.007	0.017	0.006	0.006	0.003	0.009	0.008	0.003	0.009	0.011	0.017	0.006
First	9.7×10^{-8}	0.005^{b}	0.039	0	0.004	0.011	0.006	0.003	0.003	0.003	0	0.003	0.009	0.011	0.011	0.006
Random	1.0×10^{-4}	0.008^{c}	0.043	0.003	0.007	0.017	0.006	0.006	0.003	0.009	0.011	0.003	0.009	0.018	0.017	0.006
Random	9.7×10^{-8}	0.005^{ab}	0.003	0	0.004	0.011	0.006	0.003	0.003	0.003	0	0.003	0.009	0.008	0.011	0.006
Ascending Arbitrary Priors																
First	1.0×10^{-4}	0.005^{b}	0.004	0	0.006	0.014	0.006	0.006	0.003	0.003	0.008	0	0.006	0.006	0.014	0.003
First	9.7×10^{-8}	0.004^{ab}	0.003	0	0.003	0.008	0.006	0.003	0.003	0.003	0	0	0.006	0.006	0.009	0.003
Random	1.0×10^{-4}	0.006^{b}	0.004	0	0.006	0.014	0.006	0.006	0.003	0.003	0.011	0	0.006	0.006	0.014	0.003
Random	9.7×10^{-8}	0.003^{a}	0.003	0	0.003	0.009	0.006	0.003	0.003	0.003	0	0	0.006	0.003	0.009	0.003
Descending Arbitrary Priors																
First	1.0×10^{-4}	0.007^{bc}	0.005	0.003	0.007	0.017	0.006	0.003	0.003	0.003	0.008	0.003	0.009	0.011	0.017	0.009
First	9.7×10^{-8}	0.006^{b}	0.005	0	0.004	0.014	0.006	0	0.003	0.003	0	0.003	0	0.011	0.014	0.009
Random	1.0×10^{-4}	0.007^{c}	0.005	0.003	0.007	0.017	0.006	0.003	0.003	0.003	0.011	0.003	0.009	0.011	0.017	0.009
Random	9.7×10^{-8}	0.005^{ab}	0.004	0	0.004	0.011	0.006	0	0.003	0.003	0	0.003	0.009	0.008	0.011	0.009
Data-Based Proportional Priors																
First	1.0×10^{-4}	0.006^{bc}	0.004	0	0.006	0.014	0.006	0.003	0.003	0.003	0.008	0	0.009	0.011	0.014	0.006
First	9.7×10^{-8}	0.005^{ab}	0.004	0	0.004	0.011	0.006	0	0.003	0.003	0	0	0.009	0.011	0.009	0.006
Random	1.0×10^{-4}	0.007^{bc}	0.005	0	0.006	0.014	0.006	0.003	0.003	0.003	0.011	0	0.009	0.011	0.014	0.006
Random	9.7×10^{-8}	0.004^{ab}	0.004	0	0.004	0.009	0.006	0	0.003	0.003	0	0	0.009	0.008	0.009	0.006
Equal Priors																
First	1.0×10^{-4}	0.007^{bc}	0.005	0.003	0.007	0.0167	0.006	0.003	0.003	0.003	0.008	0.003	0.009	0.011	0.017	0.009
First	9.7×10^{-8}	0.005^{b}	0.004	0	0.004	0.011	0.006	0	0.003	0.003	0	0.003	0.009	0.011	0.011	0.009
Random	1.0×10^{-4}	0.007^{bc}	0.005	0.003	0.006	0.017	0.006	0.003	0.003	0.003	0.011	0.003	0.006	0.011	0.017	0.009
Random	9.7×10^{-8}	0.005^{ab}	0.004	0	0.004	0.011	0.006	0	0.003	0.003	0	0.003	0.009	0.008	0.011	0.009
Neighbor-joining with K2P model‡																
		0.002	0.002	0	0.001	0.006	0.003	0	0	0.003	0	0.003	0	0	0.006	0.006

† misclassification averages with different superscripts represent significant differences at the 0.05 level.

‡ 5.4% of the observations were removed so K2P model could estimate distances.

Table 4.15: *10-fold Cross-validated Classification times and positions for the Butterfly data set.*

Starting Position	δ	Overall Time Classification Measures (min) Mean	S.D.	Min	Median	Max	Overall Position Classification Measures Mean	S.D.	Min	Median	Max
Arbitrary Priors											
First	1.0×10^{-4}	6.359	0.114	6.229	6.351	6.644	201.279	50.390	75	206	255
First	9.7×10^{-8}	6.730	0.338	6.432	6.587	7.461	169.839	55.216	67	167	255
Random	1.0×10^{-4}	7.247	0.161	7.091	7.207	7.682	192.968	52.173	56	185	255
Random	9.7×10^{-8}	7.222	0.058	7.146	7.221	7.315	154.700	59.156	32	141.5	255
Ascending Arbitrary Priors											
First	1.0×10^{-4}	7.678	0.329	6.882	7.770	8.017	205.468	48.745	75	212	255
First	9.7×10^{-8}	6.594	0.169	6.440	6.546	7.051	169.585	55.133	68	167	255
Random	1.0×10^{-4}	7.215	0.083	7.051	7.212	7.343	193.640	52.299	56	185	255
Random	9.7×10^{-8}	7.227	0.060	7.124	7.226	7.324	154.609	59.253	32	142	255
Descending Arbitrary Priors											
First	1.0×10^{-4}	7.155	0.645	6.446	7.041	7.854	205.043	50.128	75	221	255
First	9.7×10^{-8}	6.866	0.459	6.490	6.657	7.716	169.703	55.172	68	167	255
Random	1.0×10^{-4}	7.309	0.049	7.220	7.314	7.385	194.661	52.614	56	195	255
Random	9.7×10^{-8}	7.271	0.112	7.120	7.272	7.432	154.854	59.298	32	142	255
Data-Based Proportional Priors											
First	1.0×10^{-4}	7.102	0.224	6.777	7.102	7.510	199.690	50.439	75	203	255
First	9.7×10^{-8}	6.709	0.364	6.244	6.796	7.110	169.743	55.205	68	167	255
Random	1.0×10^{-4}	7.052	0.094	6.877	7.062	7.191	188.989	53.273	56	177	255
Random	9.7×10^{-8}	7.028	0.093	6.915	7.043	7.149	154.813	59.352	32	142	255
Equal Priors†											
First	1.0×10^{-4}	6.536	0.148	6.350	6.531	6.887	206.666^{d}	48.817	75	224	255
First	9.7×10^{-8}	7.603	0.282	7.261	7.628	7.993	169.886^{b}	55.198	68	167	255
Random	1.0×10^{-4}	6.265	0.115	6.152	6.258	6.541	195.618^{c}	51.679	56	195	255
Random	9.7×10^{-8}	6.241	0.067	6.136	6.249	6.325	154.884^{a}	59.327	32	142	255

† position averages with different superscripts represent significant differences at the 0.05 level.

Table 4.16: *10-fold Cross-validated Misclassification rates for the Fish data set. For the individual misclassification rates for groups 1-9, and group 10 the number of observations in the reference data set R were 683, and 677, respectively while the number of observations in the test data set T were 67, 65, 66, 68, 68, 65, 66, 67, 66, and 74 for groups 1-10, respectively.*

Starting		Overall Misclassification Measures					Individual Misclassification Rates									
Position	δ	Mean	S.D.	Min	Median	Max	1	2	3	4	5	6	7	8	9	10
Arbitrary Priors																
First	1.0×10^{-4}	0.008	0.011	0	0	0.031	0	0	0.015	0.015	0	0.031	0.015	0	0	0
First	9.7×10^{-8}	0.008	0.011	0	0	0.031	0	0	0.015	0.015	0	0.031	0.015	0	0	0
Random	1.0×10^{-4}	0.006	0.011	0	0	0.031	0	0	0	0.015	0	0.031	0.015	0	0	0
Random	9.7×10^{-8}	0.006	0.011	0	0	0.031	0	0	0	0.015	0	0.031	0.015	0	0	0
Ascending Arbitrary Priors																
First	1.0×10^{-4}	0.008	0.011	0	0	0.031	0	0	0.015	0.015	0	0.031	0.015	0	0	0
First	9.7×10^{-8}	0.008	0.011	0	0	0.031	0	0	0.015	0.015	0	0.031	0.015	0	0	0
Random	1.0×10^{-4}	0.006	0.011	0	0	0.031	0	0	0	0.015	0	0.031	0.015	0	0	0
Random	9.7×10^{-8}	0.006	0.011	0	0	0.031	0	0	0	0.015	0	0.031	0.015	0	0	0
Descending Arbitrary Priors																
First	1.0×10^{-4}	0.006	0.008	0	0	0.015	0	0	0.015	0.015	0	0.015	0.015	0	0	0
First	9.7×10^{-8}	0.006	0.008	0	0	0.015	0	0	0.015	0.015	0	0.015	0.015	0	0	0
Random	1.0×10^{-4}	0.005	0.007	0	0	0.015	0	0	0	0.015	0	0.015	0.015	0	0	0
Random	9.7×10^{-8}	0.005	0.007	0	0	0.015	0	0	0	0.015	0	0.015	0.015	0	0	0
Data-Based Proportional Priors																
First	1.0×10^{-4}	0.008	0.011	0	0	0.031	0	0	0.015	0.015	0	0.031	0.015	0	0	0
First	9.7×10^{-8}	0.008	0.011	0	0	0.031	0	0	0.015	0.015	0	0.031	0.015	0	0	0
Random	1.0×10^{-4}	0.006	0.011	0	0	0.031	0	0	0	0.015	0	0.031	0.015	0	0	0
Random	9.7×10^{-8}	0.006	0.011	0	0	0.031	0	0	0	0.015	0	0.031	0.015	0	0	0
Equal Priors																
First	1.0×10^{-4}	0.006	0.008	0	0	0.015	0	0	0.015	0.015	0	0.015	0.015	0	0	0
First	9.7×10^{-8}	0.006	0.008	0	0	0.015	0	0	0.015	0.015	0	0.015	0.015	0	0	0
Random	1.0×10^{-4}	0.005	0.007	0	0	0.015	0	0	0	0.015	0	0.015	0.015	0	0	0
Random	9.7×10^{-8}	0.005	0.007	0	0	0.015	0	0	0	0.015	0	0.015	0.015	0	0	0
Neighbor-joining with K2P model																
		0.015	0.018	0	0.015	0.046	0.045	0	0.015	0.015	0.015	0.046	0.015	0	0	0

Table 4.17: *10-fold Cross-validated Classification times and positions for the Fish data set.*

Starting Position	δ	Overall Time Classification Measures (min) Mean	S.D.	Min	Median	Max	Overall Position Classification Measures Mean	S.D.	Min	Median	Max
Arbitrary Priors											
First	1.0×10^{-4}	2.319	0.059	2.238	2.336	2.409	159.555	61.388	68	134	255
First	9.7×10^{-8}	2.326	0.048	2.281	2.310	2.435	137.748	562.189	68	111	255
Random	1.0×10^{-4}	2.525	0.068	2.192	2.237	2.410	130.515	74.007	32	101	255
Random	9.7×10^{-8}	2.206	0.024	2.174	2.202	2.253	104.176	75.081	22	71	255
Ascending Arbitrary Priors											
First	1.0×10^{-4}	2.246	0.027	2.217	2.235	2.304	162.089	61.359	69	135.5	255
First	9.7×10^{-8}	2.077	0.061	2.004	2.081	2.218	137.853	62.247	69	111	255
Random	1.0×10^{-4}	2.207	0.026	2.168	2.204	2.263	132.167	74.052	32	108	255
Random	9.7×10^{-8}	2.204	0.029	2.176	2.201	2.271	104.167	75.087	22	71	255
Descending Arbitrary Priors											
First	1.0×10^{-4}	2.057	0.039	2.000	2.053	2.125	161.141	61.441	69	134	255
First	9.7×10^{-8}	2.027	0.025	1.997	2.027	2.079	137.877	62.267	68	111	255
Random	1.0×10^{-4}	2.201	0.023	2.174	2.201	2.254	132.231	73.579	32	108	255
Random	9.7×10^{-8}	2.200	0.022	2.169	2.202	2.233	104.179	75.089	22	71	255
Data-Based Proportional Priors											
First	1.0×10^{-4}	2.203	0.048	2.160	2.182	2.316	161.015	61.347	69	137	255
First	9.7×10^{-8}	2.115	0.043	2.080	2.097	2.219	137.760	62.176	68	111	255
Random	1.0×10^{-4}	2.220	0.040	2.182	2.206	2.300	130.515	73.177	32	108	255
Random	9.7×10^{-8}	2.203	0.025	2.175	2.120	2.240	104.192	75.081	22	71	255
Equal Priors†											
First	1.0×10^{-4}	2.248	0.053	2.193	2.230	2.357	162.712^{d}	61.263	69	137	255
First	9.7×10^{-8}	2.274	0.049	2.228	2.259	2.400	137.760^{c}	62.176	69	111	255
Random	1.0×10^{-4}	2.240	0.103	2.166	2.196	2.486	133.497^{b}	73.559	32	108	255
Random	9.7×10^{-8}	2.185	0.025	2.144	2.185	2.235	104.176^{a}	75.087	22	71	255

† position averages with different superscripts represent significant differences at the 0.05 level.

random starting position. It may be the case the mutation rate in the COI region for these organisms is close to 9.7×10^{-8} giving the proposed method a greater ability to make correct classifications. Also, both of these data sets had missing data at the beginning of many of the barcodes. This would explain the increased ability for classification using the random starting position and perhaps suggests and alternative strategy for classification in the presenece of larg amounts of missing data at the begining of the barcode. In these cases, it may be advantageous to shift starting positions with missing data to the end of the barcode. For example, if a reference data set has a large amount of missing data for the first 50 positions of the barcodes, one may move those 50 positions to the end of the barcode so that the proposed method begins on the 51st position. Tables 4.12 and 4.14 illustrate the significant differences in average misclassification rates for the various choices of priors, δ values, and starting positions. We see that there are not many significant differences but that the smallest average misclassification rates for both of these data sets are produced when $\delta = 9.7 \times 10^{-8}$ is used with the proposed method, while the choice of priors does not seem to have a substantial impact in the average misclassification rates. This indicates that the misclassification rates are robust to the choice of priors, but correctly selecting the δ value can lead to improvements in the quality of the barcode classifications.

Starting position did not have a large effect on misclassification rates either. For the Fish data set, the misclassification rates for the random starting position are slightly lower but the difference is not significant.

The δ value selected seemed to have the largest impact on misclassification. Using $\delta = 9.7 \times 10^{-8}$ caused the proposed method to give better misclassification rates in the Bird2 and Butterfly data sets as discussed above. In the Bat and Bird1 data sets, $\delta = 1.0 \times 10^{-4}$ generally gave better misclassification rates, but the difference was not significant. One reason for this is that, perhaps, the organisms in the Bat and Bird1 data sets experience a different rate of mutation in the COI region than our estimate of 9.7×10^{-8}. This is another example of why it is important to adapt the δ value as better estimates of the mutation rate

become available. Section 6.2.6 also discusses this issue.

To compare the proposed method to the current method which uses the neighbor-joining algorithm with the K2P model discussed in Sections 2.2.2 and 2.3, we performed classification with neighbor-joining on the same reference and test data sets that were used in evaluating the proposed method. The average misclassification rates obtained using the neighbor-joining method with the K2P model for the Bat and Bird1 data sets are given in Tables 4.8 and 4.10, respectively. We see that they are similar to those achieved using the proposed method. The proposed method typically misclassified two more observations for the Bird1 data set. The proposed method, however, handily outperforms the neighbor-joining method for the Fish data set. The average misclassification rates for the proposed method are about half that of the neighbor-joining approach.

Removing the early stopping rule slightly improves the average misclassification rates when $\delta = 9.7 \times 10{-}8$ is used with the proposed method, as seen in Table 4.19, but not significantly.

The Bird2 and Butterfly reference data sets had several barcodes that were missing data in many positions. This proved problematic for the K2P model in that it did not allow for estimating several pairwise distances. Without these distances, the neighbor-joining algorithm could not be implemented and classification was not possible. In general, the K2P model could not estimate pairwise distances for barcodes that contained fewer than 36 nucleotide positions. The Bird2 and Butterfly data sets had 22, and 209 barcodes, respectively, that contained fewer than 36 nucleotide positions. This accounted for about 1% of the barcodes in the Bird2 data set and about 5.4% of the barcodes in the Butterfly data set. Removing these barcodes allowed estimation of the pairwise distances using the K2P model and the average miscalssification rates for the Bird2 and Butterfly data sets are given in Tables 4.12 and 4.14, respectively. We see that the average misclassification rates for the K2P model are somewhat lower than those of the proposed method. However, by removing barcodes with fewer than 36 nucleotide positions, we are removing some of the more challenging cases for

classification. These challenging cases led to somewhat higher misclassification rates for the proposed method but caused the current method to fail entirely. When comparing the proposed method to the current method for these two data sets, it is important to keep in mind that differences in the misclassification rates are due not only to the method of classification but also to differences in reference and test data sets because of the observations which were removed. We emphasize the fact that inclusion of barcodes with a small number of positions did not cause the proposed method to fail because of its ability to impute missing data from other barcodes in the reference data set of the same species.

What happens to the number of barcodes correctly classified if a greater proportion of barcodes is selected for the hold-out group? To address this question, we selected the butterfly data set and held out 20, 40, and 60% of the observations for classification. This was done by combining the barcodes from the previously selected hold-out groups consisting of 10% of the total barcodes to achieve the desired proportion for the hold-out group. For example, the 20% hold-out group was created by combining the first two hold-out groups of the butterfly data set, each with 10% of the total butterfly barcodes. The 40 and 60% hold-out groups were likewise created by combing the first four and six hold-out groups respectively. Holding out 20% of the data resulted in the misclassification of two barcodes, while holding out 40 and 60 % of the barcodes resulted in misclassifying 7 and 29 barcodes, respectively. From the 10-fold cross-validation of the butterfly data set given in Table 4.14, we see that the first two hold-out groups misclassify two observations, while the first four misclassifiy four barcodes and the first six misclassify five barcodes. Here we see that using a smaller number of barcodes in the refernce data set, from which the conditional probabilities are to be computed, results in a greater number of barcodes that are misclassified. This is to be expected in that, less data is used to compute conditional probabilities, but it is surprising that holding out 20-40% of the observations still results in very low misclassification rates. It should be noted that it became difficult to insure that each species is represented by at least one barcode in the reference data set with hold-out groups of 40% and higher.

Data Set	Minimum	Median	Mean	Maximum
Bat	2.803	2.927	2.973	3.627
Bird1	4.346	4.916	4.887	5.667
Bird2	6.597	6.934	7.151	8.882
Butterfly	6.241	7.077	6.970	7.678
Fish	2.027	2.205	2.201	2.326

Table 4.18: *Computation time statistics for the 5 real data sets across all combinations of priors, δ values, and starting positions. Times are in minutes.*

4.2.2 Computation Time for Classification

The time required for classification was explored via ANOVA to see if there were significant differences among the combinations of priors, δ values, and starting positions. Again, treating test data sets as blocks, we carried out an ANOVA for a randomized complete block design, and the overall F test was significant (p-value$<$0.0001) for each of the 5 data sets. We should note, however, that many different processors were used in carrying out the proposed method of classification. These processors did not all have the same specifications and so differences in average time is confounded with processor. As a practical matter, the times were are very similar within each data set. Our primary concern was to ensure that a particular combination of priors, δ values, and starting positions did not grossly increase the computation time which turned out to be the case. Table 4.18 has the minimum, median, and maximum time in minutes for each of the five data sets.

4.2.3 Number of Positions

The average number of positions used before the barcodes were classified seemed to be very similar across priors, but appeared to depend upon the δ value and the starting position. Unfortunately, only the overall statistics were saved for the number of positions, and the actual number of positions for each barcode classification were not available. Ideally, we would carry out an ANOVA of the number of positions used treating the test data sets as blocks in a randomized complete block design. Since this, however, was not feasible without

the raw data, we carried out two-sample t-tests for the average number of positions for the equal priors case in each data set. Because, the samples are not really independent in that, the same test data sets were used for each combination of δ values and starting positions, a paired t-test would be preferable, but again, without the raw data, this type of test was not feasible. In this case, significant differences found with the two-sample t-test will also be significant using the paired t-test. Tables 4.9, 4.11, 4.13, 4.15, and 4.17 show where the significant differences occur. In each case, using a random starting position used significantly fewer positions, on average, than the first starting position except in the Bird1 data set, where using the random starting position significantly increased the number of positions used. This is likely due to the absence of missing data at the beginning of the barcodes and to the presence of missing data at the end of the barcodes. By randomizing the order of the positions used for this data set, we effectively moved some of the information the sequential calculation would have drawn upon from the beginning of the barcode to the end of the barcode. This resulted in more positions being used to make the classification. In every case, using $\delta = 9.7 \times 10^{-8}$ significantly reduced the average number of positions required for classification.

What if we remove the stopping rule and let the proposed method use the entire barcode? Which δ value would be better and how would the misclassification rates be affected? To answer these questions, we selected the Bat, Bird1, and Bird2 reference and test data sets and used the proposed method to classify the test barcodes without the early stopping rule. The Bat and Bird1 data sets were selected because using the arbitrarily selected δ value with the proposed method led to a slightly better misclassification rate than using the δ value based on the mutation rate within the COI region. We wanted to see how using these two δ values with the proposed method influenced misclassification rates when we use the entire barcode. Likewise, the Bird2 data set was selected because using the arbitrarily selected δ value with the proposed method led to slightly higher average misclassification rates than using the δ values based on the mutation rate and we wanted to see if that relationship remained the

same when removing the early stopping rule. We selected equal prior probabilities with the proportional allocation method of imputing the missing values and performed the proposed method of classification using both the arbitrarily selected $\delta = 1.0 \times 10^{-4}$ and the mutation rate $\delta = 9.7 \times 10^{-8}$. Table 4.19 contains the average misclassification rates for these three data sets with and without the early stopping rule. When the early stopping rule is removed for the Bat and Bird1 data sets, the average misclassification rates for the two δ values are identical and we see that the average misclassification rate using $\delta = 9.7 \times 10^{-8}$ is slightly better using the early stopping rule. For the Bird2 data set, using $\delta = 9.7 \times 10^{-8}$ with the proposed method gives a smaller misclassification rate with and without the early stopping rule. Again we see that removing the early stopping rule gives a slightly better misclassification rate using this choice of δ with the proposed method. These results indicate that using the entire barcode can lead to slight improvements in the average misclassification rates but that these improvements will be modest. They also provide some evidence in favor of using the biologically relevant $\delta = 9.7 \times 10^{-8}$. We are of the opinion that the slight improvements in the average misclassification rates do not justify the increase in the amount of barcode used or the amount of computation time required to classify the barcodes. We conclude that there is little to be gained by using the entire barcode once the posterior probability for a species in the reference data set equals 1 and recommend using the early stopping rule.

4.3 Assuming Independence Among Nucleotides

A necessary condition for Theorem 1 to hold is that the values $x^{(1)}, \ldots, x^{(p)}$, the nucleotides at each of the p positions, are independent. Is this assumption appropriate in the case of DNA barcoding? To answer this question, let us consider a small sequence of DNA made up of the nucleotide bases $ATGACGAAC$. Triplets of nucleotide bases, called codons, code for one of the twenty-two amino acids that serve as the basic building blocks of proteins. It is important to know where to start combining nucleotides to make the triplets. For instance, is the DNA sequence above made up of the triplets ATG, ACG and AAC or do the triplets

Table 4.19: *10-fold Cross-validated Misclassification rates for the Bat, Bird1, and Bird2 data sets with and without the early stopping rule. Here, equal priors were used with the proportional allocation imputation method starting from the first position.*

	Overall Misclassification Measures					Individual Misclassification Rates									
δ	Mean	S.D.	Min	Median	Max	1	2	3	4	5	6	7	8	9	10
Bat Equal Priors with stopping rule															
1.0×10^{-4}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9.7×10^{-8}	0.001	0.004	0	0	0.012	0	0	0	0.012	0	0	0	0	0	0
Bat Equal Priors without stopping rule															
1.0×10^{-4}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9.7×10^{-8}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Bird1 Equal Priors with stopping rule															
1.0×10^{-4}	0.002	0.003	0	0	0.006	0	0	0	0.006	0	0.006	0	0	0.006	0
9.7×10^{-8}	0.003	0.004	0	0	0.012	0	0	0	0.006	0	0.006	0	0	0.012	0
Bird1 Equal Priors without stopping rule															
1.0×10^{-4}	0.002	0.003	0	0	0.006	0	0	0	0.006	0	0.006	0	0	0.006	0
9.7×10^{-8}	0.002	0.003	0	0	0.006	0	0	0	0.006	0	0.006	0	0	0.006	0
Bird2 Equal Priors with stopping rule															
1.0×10^{-4}	0.027	0.012	0.005	0.032	0.046	0.046	0.032	0.037	0.033	0.018	0.032	0.005	0.018	0.018	0.033
9.7×10^{-8}	0.023	0.014	0	0.026	0.041	0.041	0.032	0.037	0.023	0.005	0.032	0	0.018	0.018	0.028
Bird2 Equal Priors without stopping rule															
1.0×10^{-4}	0.027	0.012	0.005	0.032	0.041	0.041	0.037	0.032	0.037	0.018	0.032	0.005	0.018	0.014	0.033
9.7×10^{-8}	0.022	0.013	0	0.025	0.041	0.041	0.023	0.032	0.028	0.005	0.027	0	0.018	0.014	0.028

	T			C			A			G		
T	TTT	}	Phenylalanine	TCT	}	Serine	TAT	}	Tyrosine	TGT	}	Cysteine
	TTC			TCC			TAC			TGC		
	TTA	}	Leucine	TCA			TAA		Stop	TGA		Stop
	TTG			TCG			TAG		Stop	TGG		Tryptophan
C	CTT	}	Leucine	CCT	}	Proline	CAT	}	Histidine	CGT	}	Arginine
	CTC			CCC			CAC			CGC		
	CTA			CCA			CAA	}	Glutamine	CGA		
	CTG			CCG			CAG			CGG		
A	ATT	}	Isoleucine	ACT	}	Threonine	AAT	}	Asparagine	AGT	}	Serine
	ATC			ACC			AAC			AGC		
	ATA			ACA			AAA	}	Lysine	AGA	}	Arginine
	ATG		Methionine	ACG			AAG			AGG		
G	GTT	}	Valine	GCT	}	Alanine	GAT	}	Aspartic acid	GGT	}	Glycine
	GTC			GCC			GAC			GGC		
	GTA			GCA			GAA	}	Glutamic acid	GGA		
	GTG			GCG			GAG			GGG		

Table 4.20: *Amino Acids and thier nucleotide triplet, or codon, combinations*

start with the second position to give TGA, CGA with the leading A and trailing AC associated with other nucleotides? What we are trying to establish is know as the "frame" of the sequence. The "start" codon, ATG, is the amino acid Methionine (Met) indicates where the nucleotides begin to be read in triplets. Thus, the above DNA sequence is in fact made up of the triplets ATG, ACG and AAC and not some other shifted combination of triples. Table 4.20 gives the Amino Acids together with their codons. The triple ACG codes for the amino acid Threonine (Thr) while the triplet AAC codes for the amino acid Asparagine (Asn). We could then re-express the DNA sequence above in terms of amino acids as Met-Thr-Asn. Knowing that the DNA sequence is comprised of these triplets may indicate some underlying correlation structure between the nucleotide bases. For example, knowing that a codon begins with AA may give us some information about what the third nucleotide is likely to be. See Griffiths, Miller, Suzuki, Lewontin, and Gelbart (2000) for a nice overview of the composition and role of Amino Acids.

To examine how treating the nucleotides as independent observations, ignoring the possible dependent structure among them, affects classification, we carried out the proposed method by reordering the positions of the training and test data sets at random. This re-

ordering of the barcode positions at random should break the dependent structure among the observations.

The results of the classifications on the training and test data sets with reordered positions are given in Tables 4.8-4.16 in Section 4.2 where the starting position is labeled "Random." In these tables, we see that the average misclassification rates for the randomized ordering are on par with the average misclassification rates of the original ordering and are actually better in some cases. This indicates that the assumption of independence among the nucleotide bases is reasonable. It may be, however, advantageous to explore the dependent correlation structure of the amino acids in future work which is discussed in Section 6.2.5.

One reason for the improved misclassification rates with the randomized ordering that the original data sets tend to have missing data at the beginning of the barcode because of the software induced alignments. By randomly reordering the positions, we bring some information up into earlier positions that may not have originally been there. Therefore, the sequential calculations get off to a better start in this case and occasionally return the correct classification where the proposed method with the original data would have returned the incorrect classification.

Chapter 5

R Package

This chapter provides a detailed description of the R-package "Bayesian Discrete Ordered Classification," or `bdoc` package, which was created to carry out classification of DNA barcodes. The `bdoc` package can be downloaded and installed from the Comprehensive R Archive Network, CRAN, (http://cran.at.r-project.org/). This package contains a function to perform barcode classification using the proposed method as well as several data sets. The data sets contained in the package will be discussed in Section 5.1, and the function usage is given in Section 5.2. Required input fields for `bdoc()` are given in Section 5.2, and output produced is discussed in Section 5.3. Some discussion about the development of the package and computational issues encountered is provided in Section 5.5. Its usage is demonstrated in Example 5.4

5.1 `bdoc` Package: Data Sets

The `bdoc` package contains 20 data sets consisting of reference and test barcodes of Neotropical bats within Guyana (Clare, Lim, Engstrom, Eger, and Hebert 2006). These data sets are subsets of the publicly available DNA barcode data sets found at the DIMACS (2007) website and are included with this package in order for users to perform/verify 10-fold cross-validated classification of the data sets.

These reference and test data sets were obtained by dividing the Bat data set at random into 10 data sets with each of the 10 consisting of about 10% of the observations. These 10

Data Type	Data Set Name
Test	`battestdata1 - battestdata10`
Reference	`battraindata1 - battraindata10`

Table 5.1: *Data sets available with the* `bdoc` *package. They can be loaded to the current workspace in R using* `data(battraindata1)`*, and so on.*

data sets consist of test data that may be used to crossvalidate the proposed method. For example, the data set `battestdata1` consists of 10% of the Bat data randomly selected and held out for classification, while `battraindata1` consists of the remaing observations to be used as the reference data set. Likewise, `battestdata2` and `battraindata2` are the test and reference data for the second hold out group, and so on. Table 5.1 gives an overview of the available data sets. These data sets may be added to the current workspace in R using the command: `data()`. For example, `data(battraindata1)` will make the first reference data set available in the current R session, while `data(battestdata1)` will make the first test data set available in the current R session. The observations in `battestdata1` are not included in the reference data set `battraindata1` and therefore represent a hold-out group that, when classified, allows us to crossvalidate the proposed method. See Section 5.4 for an example.

5.2 `bdoc()` Input Values

The function `bdoc()` accepts eight input arguments, two of which consist of the test and reference data sets. The formal entry of these inputs is

```
bdoc(traindata, testdata, delta = 9.7e-08, epsilon = 0.2, priors = ‘‘equal’’,
stoppingrule = TRUE, impute = 1, plot.file = ‘‘pdf’’)
```

Notice that all but two of the input arguments have default values. The default arguments are strongly recommended, but alternative specifications are discussed below.

1. `traindata`

   This is a reference data set of type data.frame or matrix. This data set will provide the conditional probabilities of observing a particular nucleotide at any position, given the barcode belongs to a particular species. The first two columns of this data set have organism identification information. It is assumed that the second column has the species-level information and should be named `species`. The remainder of the columns contain the nucleotide sequences of the DNA barcodes.

2. `testdata`

   This is a test data set of type data.frame, matrix, or vector of the DNA barcode(s) to be classified. Because the species-level information for the test data will be unknown, column 1 should contain the first nucleotide, position 2 the second, and so on.

3. `delta`

   This is a scalar value between 0 and 0.1 used to adjust the conditional probabilities. The default value is 9.7e-8 which is the estimated mutation rate of the mitochondrial genome. It should be noted that if δ is not strictly greater than zero or exceeds 0.1, an error message is produced, and the procedure is terminated.

4. `epsilon`

   This is a scalar value between 0 and 1 used to adjust the posterior probability calculations. The default value is a moderate 0.2 and reflects the speed at which the posterior probabilities of the non-matching DNA barcodes should converge to $1/s$. If the specified ϵ value is not greater than or equal to 0 and less than or equal to 1, an error message is produced, and the procedure is terminated.

5. `priors`

   This argument specifies the prior probabilities to be used. This can be a vector of probabilities for each species in the reference data set (which should sum to 1) or any of the following options: `''equal''` - to use prior probabilities all equal to 1/s if s is

the number of species in the reference data set; ‘‘data’’ - to use prior probabilities equal to the prevalence of each species in the reference data set; ‘‘dir’’ - to use unequal, arbitrary probabilities generated from a Dirichlet(1,1,...,1) distribution. Any other specification results in equal prior probabilities being used. If a vector of prior probabilities supplied by the user does not sum to 1, a warning message is given, and the priors are rescaled by dividing each prior by the sum of the priors. This ensures the priors sum to 1.

6. `stoppingrule`

   This is a logical argument. By default `stoppingrule=TRUE`. This will terminate the sequential calculation when the posterior probability for a species in the reference data set equals 1. If `stoppingrule=FALSE`, the calculation continues until the end of the barcode is reached. If any other option is specified, the stopping rule will be disengaged, and the calculation will run until the end of the barcode is reached.

7. `impute`

   This argument specifies the imputation method to be used. The possible values are 1 and 2. If `impute=1`, the proportional allocation method will be used. If `impute=2`, the majority rule imputation will be performed. By default, this argument implements the proportional allocation method. If any other argument other than 1 or 2 is supplied, the proportional allocation method will be used.

8. `plot.file`

   This specifies the file type to be used when the posterior probability plots are constructed and saved. The plots will be saved to the current working directory in R, which can be checked by the command: `getwd()`. By default `plot.file=‘‘pdf’’` which will save the plot(s) in a PDF file. Other possible file types include: ‘‘jpg’’ which will save a JPEG file of the plot(s); ‘‘png’’ which will save a PNG file of the plot(s); ‘‘wmf’’ which will save a Windows Meta File file of the plot(s); and ‘‘ps’’

which will save a Post Script file of the plot(s). Any other specification will produce the plots in PDF format. If the default value is used, plots for each barcode in the test data set will be saved to the current directory of R with the name `seq1.pdf`, `seq2.pdf`, and so on.

5.3 `bdoc()` Output Values

The `bdoc()` function returns several important values which will be discussed below.

1. `k`

   This is the total number of barcodes in the test data set. This will allow the user to verify that all of the barcodes in the test data set have in fact been used in the classification.

2. `totaltime`

   This is the total time required to classify all of the barcodes in the test data set from start to finish.

3. `delta`

   This returns the δ value used to adjust the conditional probabilities.

4. `species.class`

   This returns a matrix of the species-level assignments as well as the probability of assignment for each barcode in the test data set. The first row of this matrix is labeled `Species` and contains the species name, the second row is labeled `Prob` and contains the posterior probability the barcode belongs to that species. Each column corresponds to a barcode from the test data set and is labeled with the integers indicating which barcode from the test data set was being classified. For example, the second column would have the column name `2`, and contain the species assignment and posterior probability of the second barcode in the test data set.

5. `priors`

   This is a vector of the initial prior probabilities.

6. `posteriors`

   This is a list containing: (1) the species-level assignment for each barcode in the test data set; (2) the matrix of posterior probabilities at each position for each barcode in the test data set. See the example in Section 5.4 for proper usage.

7. *Posterior Probability Plots*

   Posterior probability plots are constructed and saved in the format of plot.file to the current R directory and will be named `seq1`, `seq2`, and so on. The main title of these plots will have the sequence number indicating the barcode from the test data set to which the plot belongs, the species to which the barcode has been assigned, and the probability of that assignment. The horizontal axis of the plot ranges from 1 to the number of positions used in the classification and the vertical axis ranges from 0 to 1. Each species in the reference data set will be represented by a colored line in the plot and the posterior probabilities for each species will be plotted on the vertical axis at each position along the horizontal axis. In the rare event that the barcode in the test data set is assigned to one of a few species in the reference data set having identical barcodes, the plot will produce a legend containing a warning message as well as the names of the species with identical barcodes. In this event, the title will contain the word `Multiple` in place of a species name.

5.4 `bdoc()` Example

The following is an example of how the `bdoc()` function could be used to classify the observations in the test data set `battestdata1` using the `battraindata1` as the reference data set. The plot produced for the classification of the fourth barcode in `battestdata1` by the `bdoc()` function is given in Figure 5.1.

After downloading and installing the `bdoc` package from CRAN , the package is loaded to the current work session with the following command.

```
> library(bdoc)
```

The data sets that accompany the `bdoc` package can be loaded for use by typing:

```
> data(battraindata1)
> data(battestdata1)
```

There are many data sets in the `bdoc` package. This example illustrates classification of the 82 barcodes in the test data set `battestdata1` using `battraindata1` for the reference data set.

```
> traindata<-battraindata1
```

The reference data set `battraindata1` contains the genus (column 1) and species (column 2) barcode information for 758 bats representing 96 unique species. The length of each barcode is 659 nucleotides long.

```
> testdata<-battestdata1
```

The test data set `battestdata1` contains the genus (column 1) and species (column 2) barcode information for 82 bats that were held out of battraindata1. The length of each barcode is 659 nucleotides long and to classify, the first two columns need to be removed as these will usually not be known.

```
> result<-bdoc(traindata,testdata[,-c(1:2)])
```

After the above statement executes, 82 plots of type `plot.file` named `seq1.pdf`, `seq2.pdf`, and so on can be found in the folder identified by `getwd()`.

The initial prior probabilities used by the proposed method in the sequential calculation can be inspected by issuing the following command:

```
> result$priors
```

Notice that the default ‘‘`equal`’’ priors was used so each of the 96 species is given the prior probability of $1/96 \approx 0.01041667$. If the user specifies their own vector of prior probabilities, a check is made to ensure that they sum to one. If they do not sum to one, a warning is given, the priors are rescaled, and the procedure continues. Part of the output produced by the above command is given below.

```
names(priors): Ametrida.centurio
[1] 0.01041667
------------------------------------------------------------
names(priors): Anoura.caudifer
[1] 0.01041667
------------------------------------------------------------
names(priors): Anoura.geoffroyi
[1] 0.01041667
------------------------------------------------------------
.
.
.
------------------------------------------------------------
names(priors): Vampyrodes.caraccioli
[1] 0.01041667
------------------------------------------------------------
names(priors): Vampyrum.spectrum
[1] 0.01041667
```

The matrix of species assignments and posterior probabilities can be veiwed by issuing the command

```
> result$species.class
```

This matrix gives the species assignment in row 1 and the posterior probability of the assignment in row 2 for each barcode in the test data set (columns). A portion of this classification matrix is given below.

```
        1                          ... 82
Species "Carollia.perspicillata" ... "Lonchophylla.thomasi"
Prob    "1"                        ... "1"
```

The computed posterior probabilities for the first barcode in the test data set can be output using

```
> result$posteriors[[1]]$post
```

The above statement will output the matrix of posterior probabilities for each species in the reference data set (rows) at each position in the barcode (columns) until the stopping rule was reached for the first barcode in the test data set. Posterior probabilities for the second barcode in the test data set can be output by using `result$posteriors[[2]]$post` and so on. Portions of the output are given below.

```
                       Position 1 Position 2 . . . Position 554
Ametrida.centurio      0.01041667 0.01041667 . . . 2.477806e-144
Anoura.caudifer        0.01041667 0.01041667 . . . 1.464438e-164
.
.
.
Carollia.brevicauda PS2 0.01041667 0.01041667 . . .  4.547419e-23
Carollia.perspicillata  0.01041667 0.01041667 . . .  1.000000e+00
Chiroderma.trinitatum   0.01041667 0.01041667 . . . 6.636076e-199
.
```

```
.
.
Vampyrodes.caraccioli   0.01041667 0.01041667 . . . 3.186693e-156
Vampyrum.spectrum       0.01041667 0.01041667 . . . 2.497555e-212
```

Notice in the example code above we specify

```
bdoc(traindata,testdata[,-c(1:2)])
```

which uses all 82 barcodes from the `battestdata1` data set but removes the first two columns of that data set. This is done because the first two columns of this data set contain identification information about the barcode, which we will typically not have. Removing those two columns allows the proposed method of classification to proceed utilizing just the barcode information as will typically be the case.

If we investigate the classification of the fourth barcode from the `battestdata1` data set, we may start by examining the plot produced by the `bdoc()` function given in Figure 5.1. We see that the proposed method classified the barcode as belonging to the "Lonchophylla thomasi" species with probability 1. The posterior probability of the barcode belonging to that species was close enough to 1 to trigger the stopping rule on the 32nd position. Because columns 1 and 2 of `battestdata1` have identification information, namely column 1 gives the barcode's genus and column 2 gives the barcode's species, we can compare the classification to the true species identity.

```
> testdata[4,1:2]
              ID               species
412 Lonchophylla Lonchophylla.thomasi
```

The above R syntax shows that the fourth barcode in `battestdata1` does in fact belong to the species "Lonchophylla thomasi" and the proposed method made the correct classification.

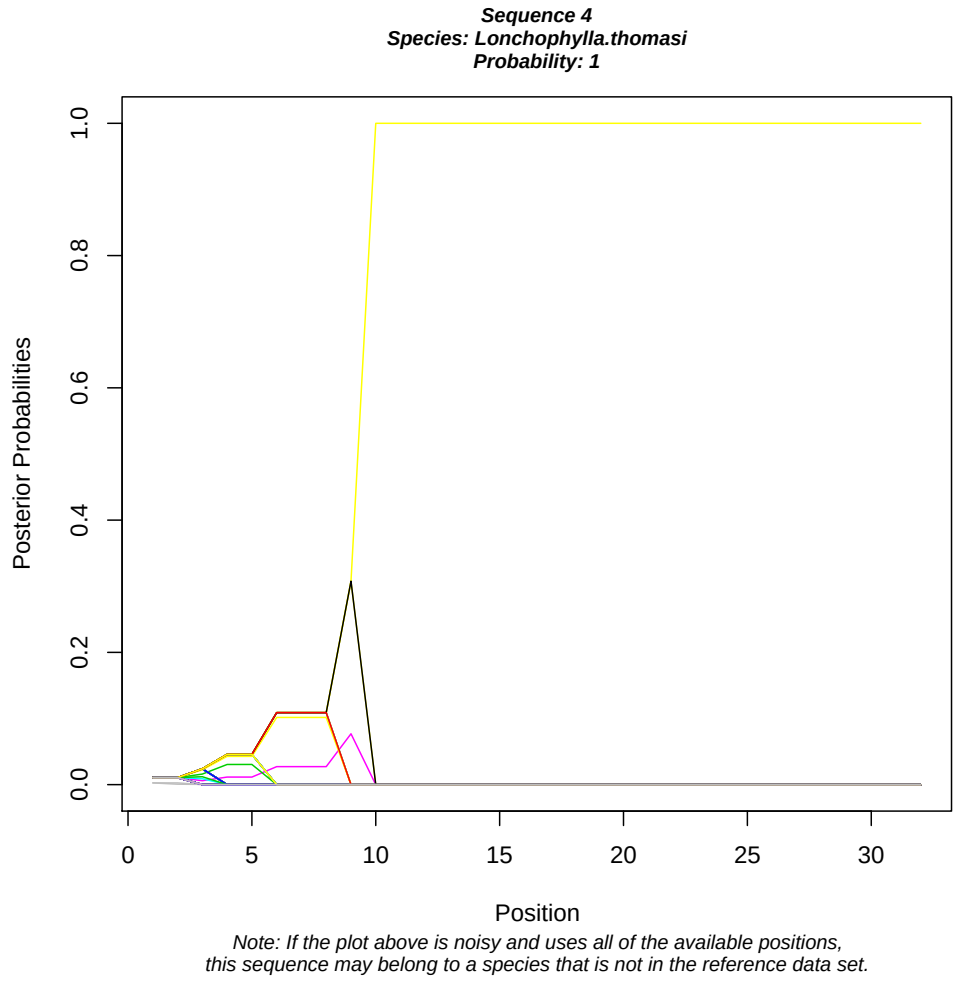

Figure 5.1: *Plot of the posterior probabilities produced using the* `bdoc()` *function with the* `battestdata1` *and* `battraindata1` *data sets. The fourth barcode in the* `battestdata1` *data set was classified to the species "Lonchophylla thomasi" with probability 1 on the 32nd position.*

5.5 Discussion of the Package

Because DNA barcode data sets often consist of many observations of long DNA sequences, care must be taken in how the sequential method is to be carried out so that the classification results can be obtained in a reasonable amount of time. Initially, the `bdoc()` function was written entirely in R code, and the classification time of one of the test data sets containing 220 barcodes with a reference data set containing 2343 barcodes from the larger Bird2 data

set took around 10 hours. The lion's share of the time was used computing the posterior probabilities, where imputing the missing data, constructing and adjusting the conditional probabilities, and computing the posterior probabilities required 4 minutes, 6 minutes, and 590 minutes, respectively. To increase computational speed, we rewrote the portion of the code that computes the posterior probabilities in C++ and used R's capabilities of calling C using the `.C()` function. See Lenarcic (2007), Rossi (2006), and Blay (2004) for some introductory tutorials on the topic. By using the complied language of C, rather than the interpreted R language for that portion of the computation, computing the posterior probabilities for the same test and references data sets now takes 1 minute. The total time for classifying the barcodes in this test data set is now 11 minutes.

With these significant gains in terms of computation time, it may be advantageous in the future to also call C from R to process the imputation of missing data as well as the construction and adjustment of the conditional probabilities.

Chapter 6

Conclusion

We have evaluated the proposed method as it applies to DNA barcoding in several ways. The first was to show that it is capable of providing accurate species-level classifications for DNA barcodes. In Sections 4.1.1 and 4.2, it was shown that the misclassification rates for both simulated data with 2-4% within-species variability and real data sets are very low. Some of the misclassifications in the real data sets appear to come from misidentified organisms.

Second, we sought to determine the effect of using the posterior probability computed for a position as the prior probability in the following position. Theorem 1, and its proof, demonstrate that this in fact estimates the probability that the barcode belongs to any species in the reference data set, given the observed data. The major assumption of Theorem 1 is that the nucleotide positions are independent. This assumption was investigated in Section 4.3 and appears to be reasonable.

Third, we wanted to generate DNA barcodes with specific amounts of within and among-species variability to evaluate the proposed method's performance in these controlled settings. We accomplished this by examining a real barcode data set and determining the prevalence of the four nucleotides. We then generated a barcode 700 positions in length by randomly selecting one of the four nucleotide bases for each position with selection probability equal to the observed prevalence in the real data set. This barcode served as a "seed" from which we generated 12 barcodes having 8% variability among them. From

each one of these, we generated four barcodes having 2, 4, 6, and 8 % variability among them. This produced four data sets, one for each level of within-species variability, having 48 barcodes, four for each of 12 species, and among-species variability of about 8%. We were then able to evaluate the proposed method under these controlled variabilities, and in Sections 4.1.1 and 4.1.2, it was shown that the proposed method had no miclassifications for 2 and 4 % within-species variability and slightly higher misclassification rates for 6 and 8% within-species variability. Johns and Avice (1998) point out that within-species variability is often less than 2%, but Meyer and Paulay (2005) point out that these estimates are typically based on non-comprehensive data sets and have, therefore, underestimated the within-species variability which could be as large as 4%. It is reassuring to know that the proposed method performs well (no misclassifications) on the generated data set with 2 and 4% within species variability and that the data sets with higher within-species variability are not likely to be encountered.

Fourth was to find an optimal δ value to be used in constructing the conditional probabilities for the reference data set. Section 3.4.1 provides some insight regarding the initial views we had about optimizing δ as well as a discussion of what this quantity represents. We learned that off-setting the zero-valued conditional probabilities by some small amount was really seeking to account for the possibility of a mutation at any of the positions along the barcode, and if chosen properly, it improved the classification rates. This led to a recommended δ value of 9.7×10^{-8} which is the estimated mutation rate of the mitochondrial genome (Denver, Morris, Lynch, Vassilieva, and Thomas 2000). Clearly, as better estimates emerge, they should be used and represent the optimal δ value we seek.

Fifth was to explore the effect of different prior probabilities on the misclassification rate. Sections 4.1.1 and 4.2 have the results that show the misclassification rates are somewhat robust to the choice of priors. For our evaluations, we used arbitrary priors generated from a non-informative Dirichlet distribution, sorted arbitrary priors (both ascending and descending), data-based proportional priors based on the prevalence of each species in the

reference data set, and equal priors. We found that the choice of priors had little to no effect on the average misclassification rates or the average number of positions used for the classification.

Sixth was to see if the proposed method could aid in species discovery by determining when a barcode does not belong to any species in the reference data set. By examining plots of the posterior probabilities, as discussed in Section 4.1.2, it is possible for one to determine that the barcode represents a new species if the plot contains excessive noise. That is to say, the highest posterior probability fluctuates among several species in the reference data set.

Lastly, we sought to produce an R package for the proposed method that the methodology might be widely available and easily implemented. We have written the "bdoc" package as well as documentation for the usage of the package which can be downloaded from CRAN . This package contains the bdoc() function as well as several data sets. Usage of the function is covered in Sections 5.2 and 5.3.

6.1 Summary

The method of classification of DNA barcodes proposed in this dissertation utilizes Bayes' Theorem applied sequentially at each nucleotide position. This is done by selecting prior probabilities for each species in the reference data set and constructing conditional probabilities from the observed barcodes in the reference data set. The posterior probability of a test barcode belonging to any of the species in the reference data set is computed for the first position and then used as the prior probability for computing the posterior probability at the second position. This sequential calculation proceeds until the posterior probability of belonging to one of the species in the reference data set equals one or the end of the barcode is reached. Either the barcode is assigned to the species in the reference data set with the highest posterior probability or it is determined to be a new species not contained in the reference data set.

This sequential approach to DNA barcode classification provides answers to some of the research challenges currently faced by DNA barcoding. Specifically, the proposed method of classification allows for assessing the uncertainty of each barcode assignment by computing posterior probabilities, addresses the issue of how much of the barcode is required for proper classification utilizing a stopping rule in the sequential calculations of equation (3.1), reduces dependence upon genetic model assumptions, and aids in the area of species discovery. More generally, this proposed method can be extended to classification in other types of high dimensional data.

6.2 Future Work

6.2.1 Extending the Proposed Method

The success of the proposed method depends, in large measure, on having several pieces of information for each observation which will be used to discriminate among groups in the data. In the context of DNA barcoding, the nucleotides are the distinct pieces of information for each barcode, and there are typically a few hundred of these. This means that the proposed method of classification is suitable for high dimensional data. We discuss one possible extension of this work presently.

Individuals with tumors had tissue samples taken of both the tumorous tissue and healthy tissue (Alon, Barkai, Notterman, Gish, Ybarra, Mack, and Levine 1999). Gene expressions were then measured for each of 2000 genes for each of the tissue samples via microarray. The measured intensities for the first five patient's tumorous and healthy tissue samples for the first ten genes are given in Table 6.1. The complete data set for all 62 patients' tumorous and healthy tissue samples for all 2000 genes is publicly available. If transcription is taking place, meaning the gene is active, the microarray expressed intensities will be larger than if the gene is not active. One important question is, "Do the tumorous and healthy tissue samples have different gene expressions?" In other words, are some genes active in the tumor tissue samples that are not active in the healthy tissue samples and vice versa. Answering

Tissue	Microarray Intensities				
Type	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5
Tumor	8589.416	5468.241	4263.408	4064.936	1997.893
Tumor	3825.705	6970.361	5369.969	4705.650	1166.554
Tumor	3230.329	3694.450	3400.740	3463.586	2181.420
Tumor	7126.599	3779.068	3705.554	6594.514	2460.905
Tumor	9330.679	7017.230	4723.783	9491.534	5346.542
Healthy	9164.254	6719.530	4883.449	3718.159	2015.221
Healthy	6246.449	7823.534	5955.835	3975.564	2002.613
Healthy	2510.325	1960.655	1566.315	3072.816	1810.205
Healthy	4028.710	3156.159	2870.255	4417.591	1854.106
Healthy	5271.518	4740.768	3318.514	6792.348	2632.889

Table 6.1: *The first 5 gene expressions for 5 individuals from the Alon data set.*

these questions can help researchers focus on genes that are the major players in tumor development. If conditional distributions of the intensities of each tumor type at each gene region can be determined, then it may be possible to use the proposed method to classify an unknown tissue sample to either of the two tissue types. This represents a significant extension of the proposed method from discrete conditional probability distributions to continuous conditional probability distributions. Further extensions could be assigning an unknown tissue sample to one of several classes using microarray data. This could happen, for example, when trying to classify a tissue sample to one of the known breast cancer cell lines. This methodology could be expanded so as to include both discrete and continuous measures for each observation. For example, there may be discrete demographic information that could be combined with the microarray data for the classification.

6.2.2 Clustering

The classification methods described in Chapter 3 were based on constructing conditional probabilities from a reference data set where the species category was known for each barcode. Clearly, we will need to identify an alternative method for conditional probability construction when dealing with a group of barcodes for which, not only the species for each

barcode is unknown, but also the number of distinct species in the data set is unknown. We may wish to identify how many distinct species are present as well as assign the barcodes to each of those species with some measure of probability.

Finite mixture models have become increasingly popular for clustering because they provide an intuitive setting for grouping observations (McLachlan and Basford 1988). Mixture models can be used to model data where each observation is assumed to have been taken from one of s groups with each group being modeled by some parametric density (component) having a mixing proportion (weight) equal to the prevalence of the group in the population. For example, consider the following traditional mixture model

$$p(x|\pi_1, \ldots, \pi_s, \phi_1, \ldots, \phi_s, \eta) = \pi_1 f(x|\phi_1, \eta) + \cdots + \pi_s f(x|\phi_s, \eta) \tag{6.1}$$

where $\pi_1, \ldots, \pi_s$ are mixing proportions, $\phi_1, \ldots, \phi_s$ are the parameters specific to each of the s components, and η is a parameter (possibly a vector of parameters) common to all components.

Using the finite mixture model in equation (6.1), and assuming the data $\boldsymbol{x}= (x^{(1)}, \ldots, x^{(p)})$ are independent observations, we can construct the likelihood function as:

$$L(s, \pi_1, \ldots \pi_s, \phi_1, \ldots, \phi_s, \eta) =$$
$$p(\boldsymbol{x}|s, \pi_1, \ldots \pi_s, \phi_1, \ldots, \phi_s, \eta) = \prod_{j=1}^{p} [\pi_1 f(x^{(j)}|\phi_1, \eta) + \cdots + \pi_s f(x^{(j)}|\phi_s, \eta)]. \tag{6.2}$$

The likelihood computed by equation (6.2) is invariant to switching the labels of the components. This means the likelihood will be multimodal and presents a challenge known as "label-switching" which will need to be addressed possibly using a similar approach to that of Stephens (2000b).

Because s is unknown, we will essentially select from set of mixture models with possibly different dimensions, a single model leading to an estimate of the number of species s. To accomplish this, Stephens (2000a) proposes constructing ergodic Markov chains with appropriate stationary distributions. The method is based on the construction of a continuous

time Markov birth-death process. This birth-death process (BDMCMC) allows movements between competing models by allowing new components to be born, which would move the current model into a higher dimension, and by allowing existing components to die, moving it into a lower dimension. In BDMCMC, a birth-death process is used to identify the appropriate number of components for the mixture model. We hope to use the BDMCMC approach to identify the appropriate number of components in the mixture model and hence, the number of distinct species present in barcode data set. Combining this birth-death process with Monte Carlo Markov Chain updates, we hope to be able to estimate the parameters in equation (6.2) which will in turn, allow us to classify the unknown barcodes to their respective species and assign probabilities to those classifications.

6.2.3 Correcting Database Errors

In Section 4.2, some of the misclassified barcodes in the real data sets appear to come from incorrectly identified organisms. For example, in the Fish test data set 7, barcode 61 is given the species id of *Pristiophorus nudipinnis*, or more commonly, the "shortnose sawshark," but the proposed method classified it as *Pristiophorus cirratus*, or the "longnose sawshark." Upon investigating the reason for the misclassification, we noticed that the test barcode had some positions differing from others of its own species. We also noticed that, when the test barcode differed from those of its own species, it contained a nucleotide that matched identically to the species *Pristiophorus cirratus*. In other words, the test barcode matched exactly the barcodes from the species *Pristiophorus cirratus* and did not match exactly the barcodes from the species *Pristiophorus nudipinnis* to which is was supposed to belong.

Table 6.2 contains the the barcodes for the organisms that belong to these two species. For brevity, we only list the non-identical positions between the two species. The barcode of the organism that the proposed method misclassified is given in the last row of the table. It seems that this organism was probably belonged to *Pristiophorus cirratus* and not to *Pristiophorus nudipinnis*. These two species are distinguished by the lengths of

species	Barcode Position																
	78	99	102	120	126	135	138	150	163	180	204	210	213	219	222	225	234
cirratus	G	C	C	A	T	G	T	A	A	T	T	A	T	T	C	A	C
cirratus	G	C	C	A	T	G	T	A	A	T	T	A	T	C	C	A	C
cirratus	G	C	C	A	T	G	T	A	A	T	T	A	T	T	C	A	C
cirratus	G	C	C	A	T	G	T	A	G	T	T	A	T	T	C	A	C
cirratus	G	C	C	A	T	G	T	A	G	T	T	A	T	T	C	A	C
nudipinnis	T	T	C	G	G	T	A	G	A	C	G	C	C	T	T	T	T
nudipinnis	T	T	T	G	G	T	A	G	A	C	G	C	C	T	T	T	T
nudipinnis	T	T	C	G	G	T	A	G	A	C	G	C	C	T	T	T	T
nudipinnis	T	T	C	G	G	T	A	G	A	C	G	C	C	T	T	T	T
nudipinnis	G	C	C	A	T	G	T	A	G	T	T	A	T	T	C	A	C

Table 6.2: *All of the barcodes in the Fish data set belonging to the two species Pristiophorus cirratus and Pristiophorus nudipinnis. For brevity, only the non-identical barcode positions are listed. It is possible that the barcode in the last row is incorrectly assigned to the species Pristiophorus nudipinnis and should be assigned to the species Pristiophorus cirratus.*

their nose and are so similar in terms of physical characteristics that telling a short-nosed longnose sawshark apart from a long-nosed shortnose sawshark would be difficult. Because morphological species identification can be difficult when two closely related species look very similar, it is likely that barcode data bases will contain these kind of errors. The proposed method can be useful in such cases. If the proposed method is applied to these databases by holding out one observation at a time to be classified, and a misclassification occurs, the barcode to be classified could be compared to the barcodes of its own species and the barcodes of the species to which it was classified. If it looks like the barcode in fact belongs to the species determined from the proposed method, the organism from which the test barcode was retrieved could be investigated further to see if perhaps it was incorrectly identified. The proposed method could, therefore, be used to clean barcode databases to remove these type of labeling errors.

6.2.4 Mismeasurement of DNA Sequence Flowgram

With modern sequencing methods, DNA fragments are attached to a synthetic bead and amplified using *polymerase chain reaction* (PCR) so that the bead will have approximately 10 million copies of the DNA fragment. The bead is then deposited into a cell where reagents cyclically flow over the DNA. On each flow cycle, the DNA strand could remain

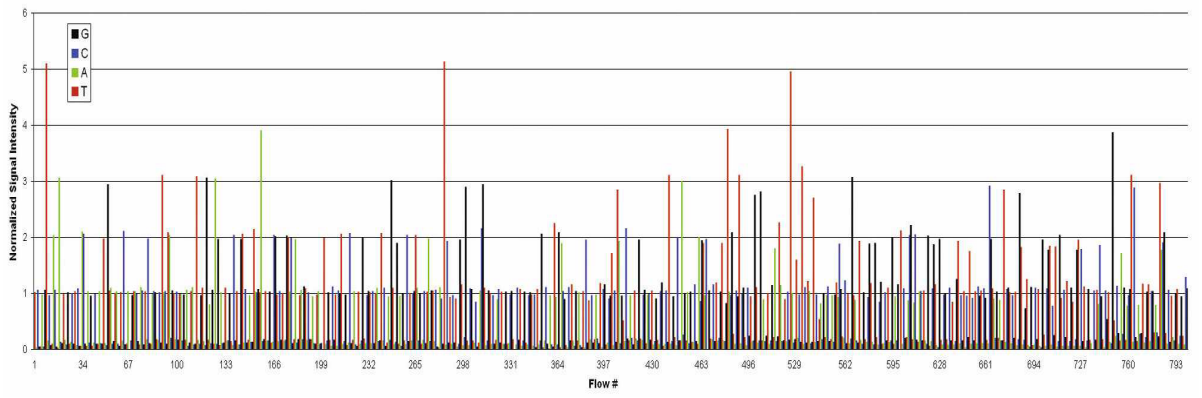

Figure 6.1: *Flowgram showing the signal intensities at each flow for sequence reading of* E. coli. *Signal intensities around one indicate a single nucleotide was added, while higher intensities indicate that multiple nucleotides were added. Image provided by 454 Sequencing* ©*2009 Roche Diagnostics*

unchanged, or be extended by one or more nucleotides. When the strand is extended, *pyrophosphate* is released and can be detected by a light sensitive laser. Each nucleotide produces a somewhat unique light emission. A *flowgram* is then produced of the observed light signals, from which, one can probabilistically determine which nucleotide what most likely to have produced the observed emission (Vacic, Jin, Zhu, and Lonardi 2008). Figure 6.1 contains a flowgram produced by sequencing 458 nucleotides from the bacteria *E. coli*. While these "high-throughput" sequencing methods are rapid and cost effective, they are prone to errors in nucleotide reads on the order of about 1 error for every 100 nucleotide bases. A few avenues for future research in this area would be; investigating this probabilistic method used for nucleotide identification to see if new models could produce better error, identifying how sequencing error of this kind effects the proposed method's ability to classify barcodes, and identifying how the proposed method might be used in detecting these errors.

6.2.5 Amino Acids versus Nucleotides

The discussion in Section 4.3 about nucleotide triplets coding for amino acids gives rise to the question, "How might the dependent structure of the nucleotides be accounted for?"

One possible approach could be changing how the conditional probabilities are constructed. Instead of specifying the conditional probabilities at each nucleotide position, we may specify them for each triplet of nucleotides. This could prove advantageous in terms of classification in that two barcodes of the same species may have different nucleotides at a position, but the nucleotides in that region may still code for the same amino acid. Thus, the two barcodes may look different when examining the individual nucleotides, but they may look identical when examining the amino acids. Consider Table 4.20 that gives all of the amino acids together with their defining nucleotide triplets. If we are to compare the following two DNA sequences, $ATGAACAAG$ and $ATGAATAAA$, we notice that nucleotide positions 6 and 9 are different. Consideration of the amino acids, however, tells us that both DNA sequences contain the codons for Methionine (ATG), Asparagine (AAC and AAT) and Lysine (AAG and AAA). Therefore, the two sequences are identical in terms of amino acid order and composition. The biological relevance of this is that using the amino acids, rather than the nucleotides, in a barcode could help account for the dependent structure of the nucleotides and possibly lead to better misclassification rates at the species-level.

6.2.6 Species-specific δ

The δ adjustment made to the conditional probabilities in equation (3.1) is based on the mutation rate of the mitochondrial genome. If estimates of the mutation rates for each species in the reference data could be obtained, they could be used for this adjustment of the conditional probabilities. This could improve the proposed method's ability to properly classify unknown barcodes by making it more sensitive to the prevalence of mutations for each species.

6.2.7 Computational Issues

As technology advances in the direction of a hand-held barcode reader that may be used to obtain barcodes "in the field," the need for the proposed method to be portable increases. A complete transition of programming languages from R to C will be necessary. This transition

will not only make the proposed method available on many platforms, but it will greatly increase the speed at which classifications can be made. Currently, only the computation of the posterior probabilities is done in C. This modification to the original R program greatly increased the speed with which the classifications can be made, but imputing the missing data and constructing the conditional probabilities can take several minutes, especially in cases where the reference data set is large. It is anticipated that the next release of the bdoc() function will use C for these routines as well, reducing overall classification from a few minutes to a few seconds.

Bibliography

Ahituv, N., Y. Zhu, A. Visel, A. Holt, V. Afzal, L. Pennacchio, and E. Rubin (2007). Deletion of ultraconserved elements yields viable mice. *PLoS Biol 5*, (9): e234. doi:10.1371/journal.pbio.0050234.

Alon, U., N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine (1999). Broad patterns of gene expression revealed by clustering of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. 96(12)*, 6745–6750.

Altschul, S., W. Gish, W. Miller, E. Myers, and D. Lipman (1990). Basic local alignment search tool. *J. Mol. Biol. 215*, 403–410.

Blay, S. (2004). Calling C code from R an introduction. http://www.sfu.ca/~sblay/R/index.html#c.

BOLD (2009). Barcode of life data (BOLD) systems. http://www.barcodinglife.org.

Clare, E. L., B. K. Lim, M. D. Engstrom, J. L. Eger, and P. D. N. Hebert (2006). DNA barcoding of Neotropical bats: species identification and discovery within Guyana. http://www.barcodeoflife.org/barcode/batsbirds/literature/MEN1657_final.pdf.

Cooke, F., R. Rockwell, and D. Lank (1995). *The Snow Geese of La Perouse Bay: natural selection in the wild.* Oxford: Oxford University Press.

CRAN. Comprehensive R Archive Network. http://www.r-project.org/.

Denver, D., K. Morris, M. Lynch, L. Vassilieva, and W. Thomas (2000). High direct estimate of the mutation rate in the mitochondrial genome of caenorhabditis elegans. *Science 289*, 2342–2344.

DeSalle, R. (2006). Species discovery versus species identification in DNA barcoding efforts: response to rubinoff. *Cons. Biol. 20*, 1545–1547.

DIMACS (2007). Center for discrete mathematics and theoretical computer science. http://dimacs.rutgers.edu/Workshops/BarcodeResearchChallenges2007.

Dove, C. (2000). A descriptive and phylogenetic analysis of plumulaceous feather chatacters in *Charadriiformes*. *Ornith. Monogr. 51*, 1–163.

Ferguson, J. (2002). On the use of genetic divergence for identifying species. *Biol. J. Linn. Soc. 75*, C509–C516.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugen. 7*, 179–188.

Fix, E. and J. Hodges (1951). Discriminatory analysis. nonparametric discrimination; consistency properties. Technical Report Report Number 4, Project Number 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas. (Reprinted as pp 261-279 of Agrawala, 1977).

Frézal, L. and R. Leblois (2008). Four years of DNA barcoding: Current advances and prospects. *Infect. Genet. Evol. 8(5)*, 727–736.

Gascuel, O. and M. Steel (2006). Neighbor-joining revealed. *Mol. Biol. Evol. 23*, 1997–2000.

Gelman, A., J. Carlin, H. Stern, and D. Rubin (2004). *Bayesian Data Analysis*, Volume 2nd ed. Boca Raton, Florida: Chapman & Hall.

Griffiths, A., J. Miller, D. Suzuki, R. Lewontin, and W. Gelbart (2000). *An Introduction to Genetic Analysis*, Volume 7th ed. W. H. Freeman and Company.

Guglich, E., P. Wilson, and B. White (1994). Forensic application of repetitive DNA markers to the species identification of animal tissues. *J. Forensic Sci. 39*, 353–361.

Hajibabaei, M., J. DeWaard, N. Ivanova, S. Ratnasingham, R. Dooh, S. Kirk, P. Mackie,

and P. Hebert (2005). Critical factors for assembling a high volume of DNA barcodes. *Phil. Trans. R. Soc. B 360*, 1959–1967.

Hammond, P. (1992). *Global biodiversity: status of the Earth's living resources.* London: Chapman & Hall.

Hebert, P., A. Cywinska, S. Ball, and J. deWaard (2003). Biological identifications through DNA barcodes. *Proc. R. Soc. Lond. (B) 270*, 313–322.

Hebert, P., S. Ratnasingham, and J. deWaard (2003). Barcoding animal life: Cytochome c oxidase subunit 1 divedrgences among closely related species. *Proceedings Biological Sciences 270*, S96–S99.

Hebert, P., M. Stoeckle, T. Zemlak, and C. Francis (2004). Identification of birds through DNA barcodes. *Plos. Biol. 2*, 1657–1663.

Ivanova, N., J. deWaard, M. Hajibabaei, and P. Hebert (2007). Protocols for high-volume DNA barcode analysis. http://www.barcoding.si.edu/PDF/Protocols_for_High_Volume_DNA_Barcode_Analysis.pdf.

Johns, G. and J. Avice (1998). A comparative summary of genetic distances in the vertbrates from the mitochondrial cytochrome b gene. *Mol. Biol. Evol. 15*, 1481–1490.

Johnson, D. (1998). *Applied Multivariate Methods for Data Analysis*. Pacific Grove, CA: Brooks/Cole.

Karlin, S. and S. Altschul (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Nat. Acad. Sci.*, 2264–2268.

Kelly, R., I. Sarkar, D. Eernisse, and R. DeSalle (2006). DNA barcoding using chitons (genus Mopalia). *Mol. Ecol. Notes 7*, 177–183.

Kimura, M. (1980). A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol. 16*, 111–120.

Kornberg, A. and T. Baker (1992). *DNA Replication, 2nd ed.* University Science Books.

Koski, L. and G. Goulding (2001). The closest BLAST hit is often not the nearest neighbor. *J. Mol. Ecol. 52*, 540–542.

Lenarcic, A. (2007). R package writing tutorial. http://www.stat.columbia.edu/~gelman/stuff_for_blog/AlanRPackageTutorial.pdf.

Lindley, D. (1970). Bayesian statistics, a review. In *Regional conference series in applied mathematics*, Society for Industrial and Applied Mathematics, Philadelphia.

McCallum, A. and K. Nigam (1998). A comparison of event models for naive Bayes text classification. Technical Report WS-98-05, AAAI-98 Workshop on Learning for Text Categorization.

McLachlan, G. and K. Basford (1988). *Mixture Models: Inference and Applications to Clustering*. New York: Dekker.

Meyer, C. and G. Paulay (2005). DNA barcoding: Error rates based on comprehensive sampling. *Plos. Biol. 3*, 2229–2238.

Ratnasingham, R. and P. Hebert (2007). BOLD: The barcode of life data system. *Mol. Ecol. Notes 7*, 355–364.

Rencher, A. (2002). *Methods of Multivariate Analysis*, Volume 2nd ed. New York: Johm Wiley & Sons.

Roche (2009). Gs flx titanium series 454 sequencer. http://www.454.com/about-454/index.asp.

Rossi, P. (2006). Making R packages under windows: A tutorial. http://faculty.chicagobooth.edu/peter.rossi/research/bayes%20book/bayesm/Making%20R%20Packages%20Under%20Windows.pdf.

Saitou, N. and M. Nei (1987). The neighbor-joining method: A new method for reconstruction phylogenetic trees. *Mol. Biol. Evol. 4*, 406–425.

Stephens, M. (2000a). Bayesian analysis of mixture models with an unknown number of components-an alternative to reversible jump methods. *Annals of Statistics 28(1)*, 40–74.

Stephens, M. (2000b). Dealing with label switching in mixture models. *J. Roy. Statist. Soc. Ser. B 62(4)*, 795–809.

Stoeckle, M. (2003). Taxonomy, DNA, and the barcode of life. *Bioscience 53(9)*, 2–3.

Studier, J. and K. Keppler (1988). A note on the neighbor-joining algorithm of Saitou and Nei. *Mol. Biol. Evol. 5*, 729–731.

Tamura, K., M. Nei, and S. Kumar (2004). Prospects for inferring very large phylogenies by using the neigbor-joining method. *Proc. Natl. Acad. Sci. 101*, 11030–11035.

Vacic, V., H. Jin, J. Zhu, and S. Lonardi (2008). A probabilistic method for small RNA flowgram matching. *Pacific Symposium on Biocomputing 13*, 75–86.

Will, K. and D. Rubinoff (2004). Myth of the molecule: DNA barcodes for species cannot replace morphology for identification and classification. *Cladistics 20*, 47–55.

Zhang, H. (2004). The optimality of naive Bayes. In *Proceedings of the Seventeenth Florida Artificial Intelligence Research Society Conference*, Miami Beach, FL, pp. 562–567. The AAAI Press.

Appendix A

Misclassfication Rates vs. δ

Table A.1: *Tabulated misclassification rates for various δ values. Notice that the misclassification rate depends on the choice of δ with extremely small (close to zero) δ values and large δ values yield the larger misclassification rates while δ values between 0 and 0.01 yeild the smaller misclassification rates.*

δ	Bat	Butterfly	Bird1	Bird2	Fish
1e-21	0.061	0.452	0.302	0.259	0.433
1e-11	0	0.006	0	0.027	0
9.7e-08	0	0.006	0	0.041	0
1e-04	0	0.006	0	0.041	0
1.5e-04	0	0.006	0	0.041	0
0.002	0	0.006	0	0.041	0
0.004	0	0.008	0	0.036	0
0.006	0	0.008	0	0.041	0
0.008	0	0.011	0	0.045	0
0.01	0	0.011	0	0.064	0
0.012	0	0.011	0	0.059	0
0.014	0	0.011	0	0.059	0
0.016	0	0.011	0	0.064	0
0.018	0	0.011	0	0.059	0
0.02	0	0.011	0	0.059	0
0.022	0	0.011	0	0.064	0
0.024	0	0.011	0	0.068	0
0.026	0	0.011	0	0.068	0
0.028	0	0.011	0	0.064	0
0.03	0	0.011	0	0.064	0
0.032	0	0.011	0	0.073	0
0.034	0	0.011	0	0.068	0
0.036	0	0.011	0	0.073	0

δ	Bat	Butterfly	Bird1	Bird2	Fish
0.039	0	0.011	0	0.068	0
0.041	0	0.011	0	0.073	0
0.043	0	0.011	0	0.068	0
0.045	0	0.011	0	0.068	0
0.047	0	0.011	0	0.068	0
0.049	0	0.011	0	0.073	0
0.051	0	0.011	0	0.073	0
0.053	0	0.011	0.006	0.073	0
0.055	0	0.011	0	0.068	0
0.057	0	0.008	0	0.073	0
0.059	0	0.008	0	0.073	0
0.061	0	0.008	0.006	0.073	0
0.063	0	0.008	0.006	0.073	0
0.065	0	0.008	0.006	0.073	0
0.067	0	0.008	0.006	0.077	0
0.069	0	0.008	0.012	0.073	0
0.071	0	0.008	0.012	0.068	0
0.073	0	0.008	0.012	0.073	0
0.075	0	0.008	0.012	0.073	0
0.077	0	0.008	0.019	0.068	0
0.079	0	0.008	0.012	0.077	0
0.081	0	0.008	0.012	0.077	0
0.083	0	0.008	0.012	0.068	0
0.085	0	0.008	0.012	0.077	0
0.087	0	0.008	0.019	0.077	0
0.089	0	0.008	0.012	0.082	0
0.091	0	0.008	0.012	0.082	0
0.093	0	0.008	0.012	0.077	0
0.095	0	0.008	0.019	0.077	0
0.097	0	0.011	0.019	0.077	0
0.099	0	0.014	0.019	0.077	0
0.101	0	0.014	0.019	0.082	0
0.103	0	0.014	0.012	0.077	0
0.105	0	0.014	0.012	0.077	0
0.107	0	0.014	0.019	0.077	0
0.109	0	0.014	0.012	0.077	0
0.111	0	0.014	0.012	0.082	0
0.113	0	0.014	0.019	0.082	0
0.115	0	0.014	0.019	0.077	0
0.117	0	0.014	0.019	0.077	0

δ	Bat	Butterfly	Bird1	Bird2	Fish
0.119	0	0.014	0.019	0.086	0
0.121	0	0.014	0.019	0.077	0
0.123	0	0.014	0.019	0.082	0
0.125	0	0.014	0.019	0.082	0
0.127	0	0.014	0.019	0.082	0
0.129	0	0.014	0.019	0.073	0
0.131	0	0.014	0.019	0.086	0
0.133	0	0.014	0.031	0.095	0
0.135	0.012	0.014	0.025	0.082	0
0.137	0.012	0.014	0.019	0.086	0
0.139	0.012	0.014	0.019	0.082	0
0.141	0.012	0.014	0.019	0.082	0
0.143	0.012	0.014	0.031	0.082	0
0.145	0.012	0.014	0.019	0.082	0
0.148	0.012	0.014	0.037	0.082	0
0.15	0.012	0.014	0.031	0.082	0
0.152	0.012	0.014	0.037	0.082	0
0.154	0.012	0.017	0.037	0.082	0
0.156	0.012	0.014	0.049	0.086	0
0.158	0.012	0.014	0.049	0.086	0
0.16	0.012	0.014	0.056	0.082	0
0.162	0.012	0.014	0.056	0.082	0
0.164	0.012	0.014	0.043	0.086	0
0.166	0.012	0.02	0.049	0.086	0
0.168	0.012	0.02	0.043	0.082	0
0.17	0.012	0.02	0.062	0.086	0
0.172	0.012	0.022	0.062	0.086	0
0.174	0.012	0.022	0.074	0.082	0
0.176	0.012	0.025	0.074	0.082	0
0.178	0.012	0.025	0.074	0.086	0
0.18	0.012	0.025	0.074	0.082	0
0.182	0.012	0.025	0.086	0.082	0
0.184	0.012	0.031	0.086	0.082	0
0.186	0.024	0.037	0.093	0.082	0
0.188	0.024	0.037	0.105	0.082	0
0.19	0.024	0.039	0.105	0.082	0
0.192	0.024	0.039	0.136	0.082	0
0.194	0.024	0.039	0.136	0.086	0
0.196	0.024	0.039	0.148	0.086	0
0.198	0.024	0.042	0.191	0.091	0
0.2	0.037	0.048	0.21	0.086	0

Appendix B

Miclassification Rates for Simulated data with 2 and 4% Within-Species Variability

	Overall Misclassification Measures					Individual Misclassification Rates											
δ	Mean	S.D.	Min	Median	Max	1	2	3	4	5	6	7	8	9	10	11	12
Unequal Unsorted Priors																	
1.0×10^{-4}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9.7×10^{-8}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Unequal Sorted in Ascending Order Priors																	
1.0×10^{-4}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9.7×10^{-8}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Unequal Sorted in Descending Order Priors																	
1.0×10^{-4}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9.7×10^{-8}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Unequal Data Based Proportional Priors																	
1.0×10^{-4}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9.7×10^{-8}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Equal Priors																	
1.0×10^{-4}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9.7×10^{-8}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table B.1: *Cross-validated Misclassification rates for the Simulated data set with 2% within-species variability. For the individual misclassification rates for groups 1-10; the number of observations in the reference data set R were 44, and the number of observations in the test data set T were 4.*

	Overall Misclassification Measures					Individual Misclassification Rates											
δ	Mean	S.D.	Min	Median	Max	1	2	3	4	5	6	7	8	9	10	11	12
Unequal Unsorted Priors																	
1.0×10^{-4}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9.7×10^{-8}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Unequal Sorted in Ascending Order Priors																	
1.0×10^{-4}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9.7×10^{-8}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Unequal Sorted in Descending Order Priors																	
1.0×10^{-4}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9.7×10^{-8}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Unequal Data Based Proportional Priors																	
1.0×10^{-4}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9.7×10^{-8}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Equal Priors																	
1.0×10^{-4}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9.7×10^{-8}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table B.2: *Cross-validated Misclassification rates for the Simulated data set with 4% within-species variability. For the individual misclassification rates for groups 1-10; the number of observations in the reference data set R were 44, and the number of observations in the test data set T were 4.*

Appendix C

Plotted Posteriors for Species Discovery

The plots that follow show the computed posterior probabilities versus barcode position for the simulated DNA barcodes discussed in Section 4.1. A single figure contains the plots for the four barcodes within each of the 12 species and the proposed method was carried out in cases of 2%, 4%, 6%, and 8% levels of within species variability. Ideally, the plots would look very noisy to indicate that the the barcode to be classified does not belong to and species in the reference data set. With a few obvious exceptions of sequences 2 and 3 for species 1 at 4%, Figure C.2, sequences 1 and 2 for species 2 at 4%, Figure C.6, sequence 1 for species 2 at 6%, Figure C.7, sequences 1, 3, and 4 for species 2 at 8%, Figure C.8, and sequence 4 for species 9 at 6%, Figure C.35, this is generally the case.

Notice that as the within-species variability increases, the plots become less noisy and have greater difficulty indicating the barcode to be classified does not belong to any species in the reference data set. Not only is the stopping rule triggered more often with larger amounts of within species variability, but the plots indicate that there is less fluctuation between species with the highest posterior probability. Expectedly, this means that it is more difficult to discriminate among species at the higher levels of within-species variability than at the lower levels.

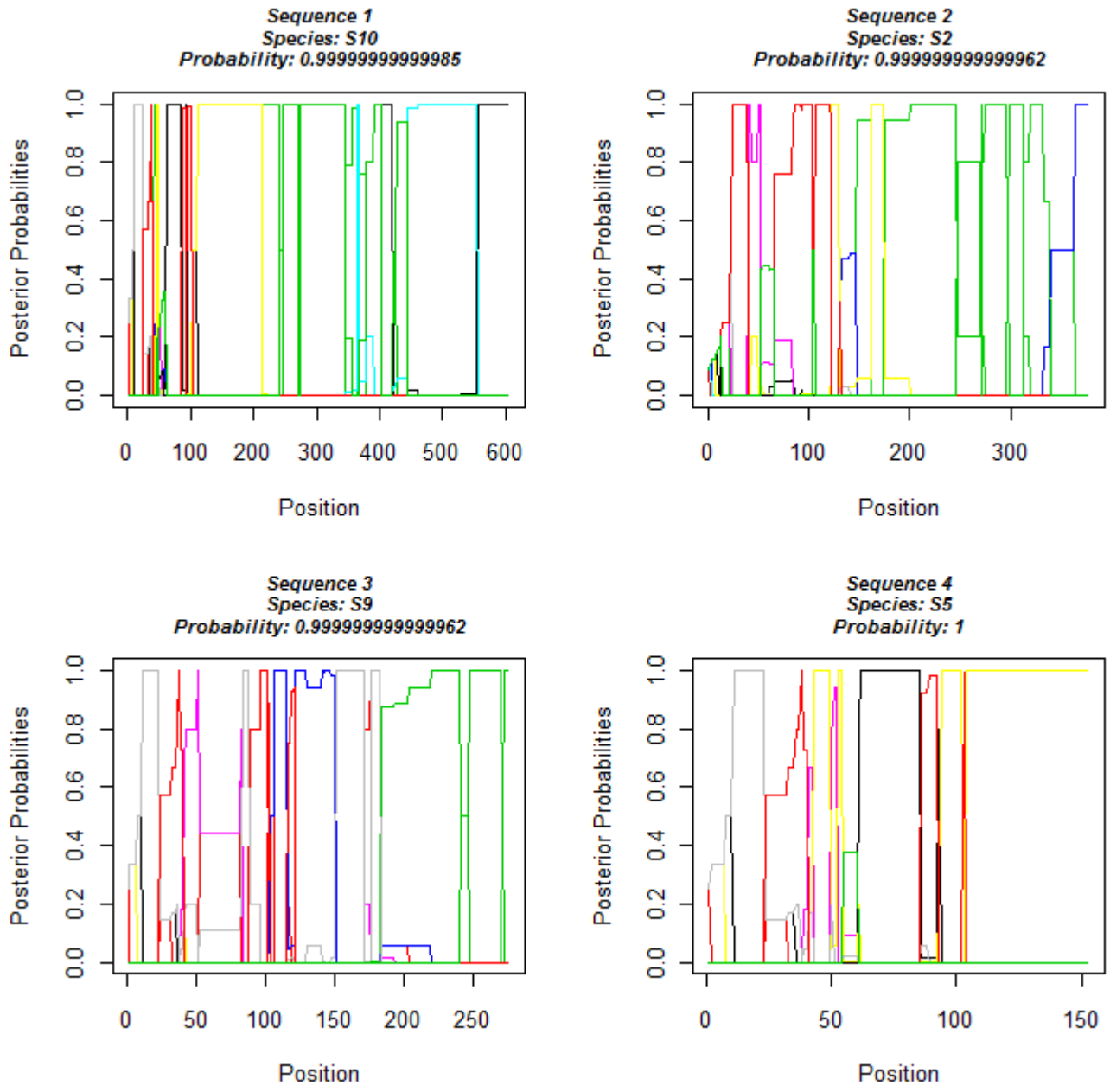

Figure C.1: *Plotted posterior probabilities having removed species 1 from the reference data set and seeking a classification of the four barcodes belong to species 1. Equal prior proabilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 2% *were used.*

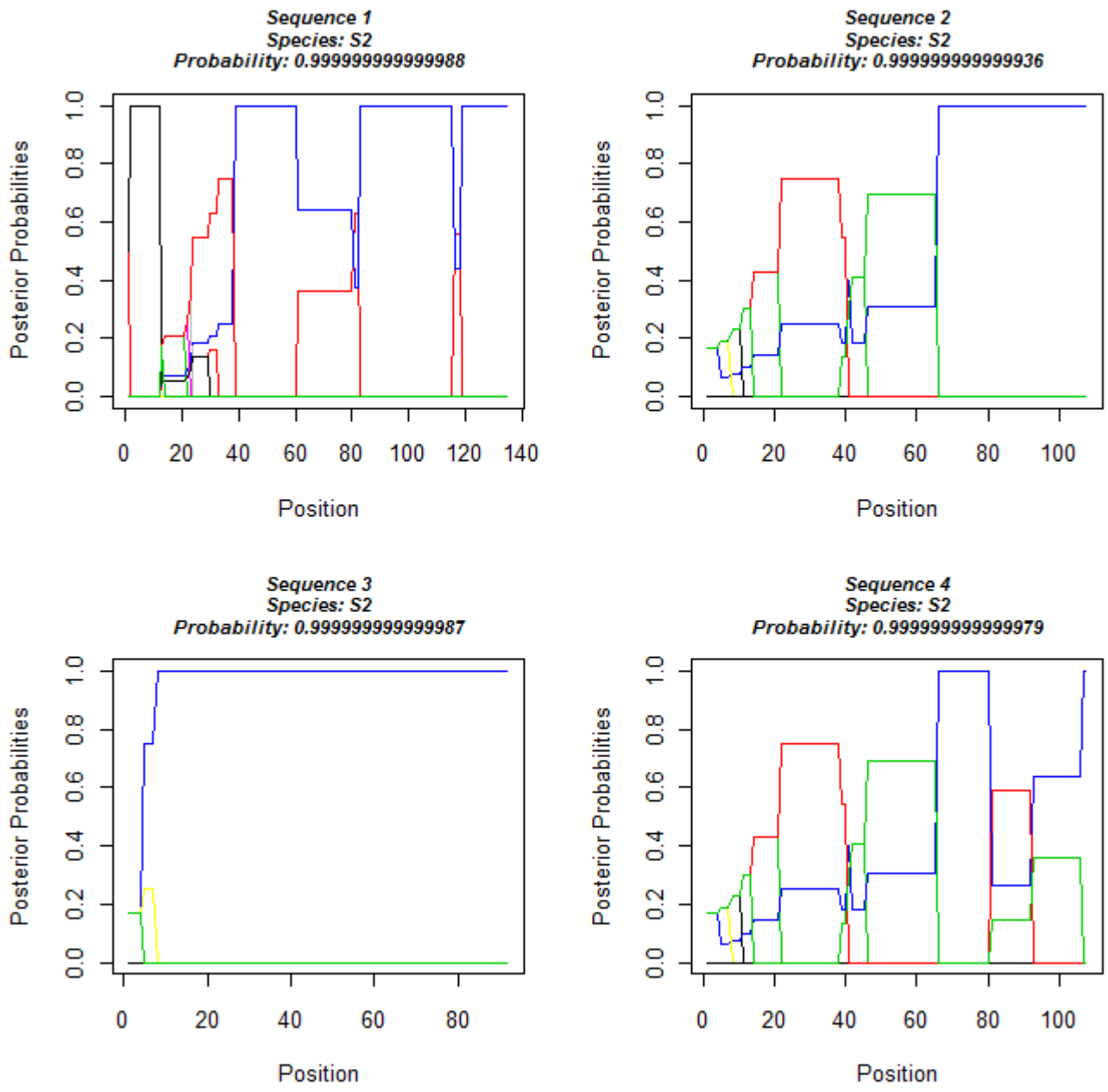

Figure C.2: *Plotted posterior probabilities having removed species 1 from the reference data set and seeking a classification of the four barcodes belong to species 1. Equal prior proabilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 4% *were used.*

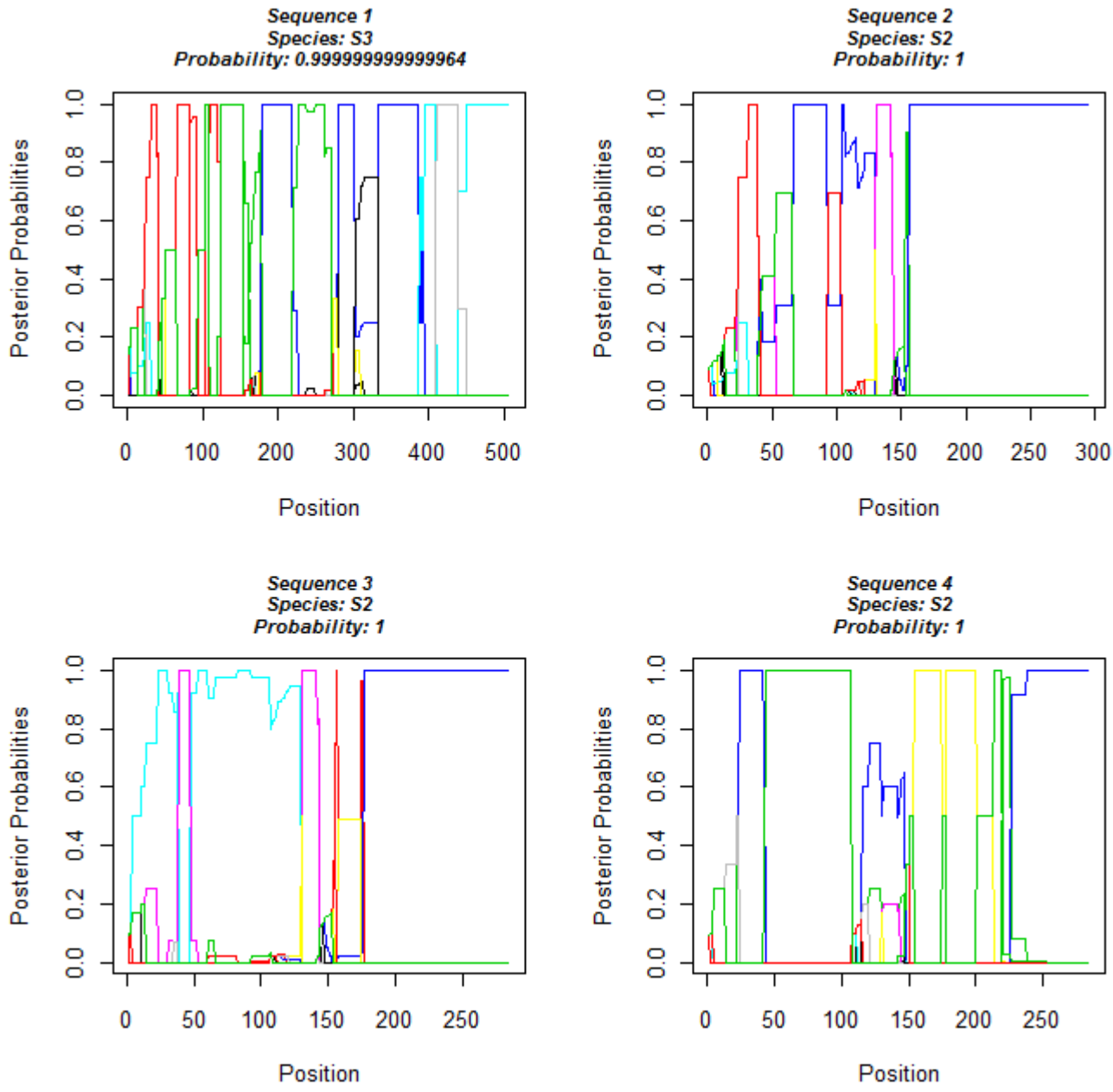

Figure C.3: *Plotted posterior probabilities having removed species 1 from the reference data set and seeking a classification of the four barcodes belong to species 1. Equal prior proabilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 6% *were used.*

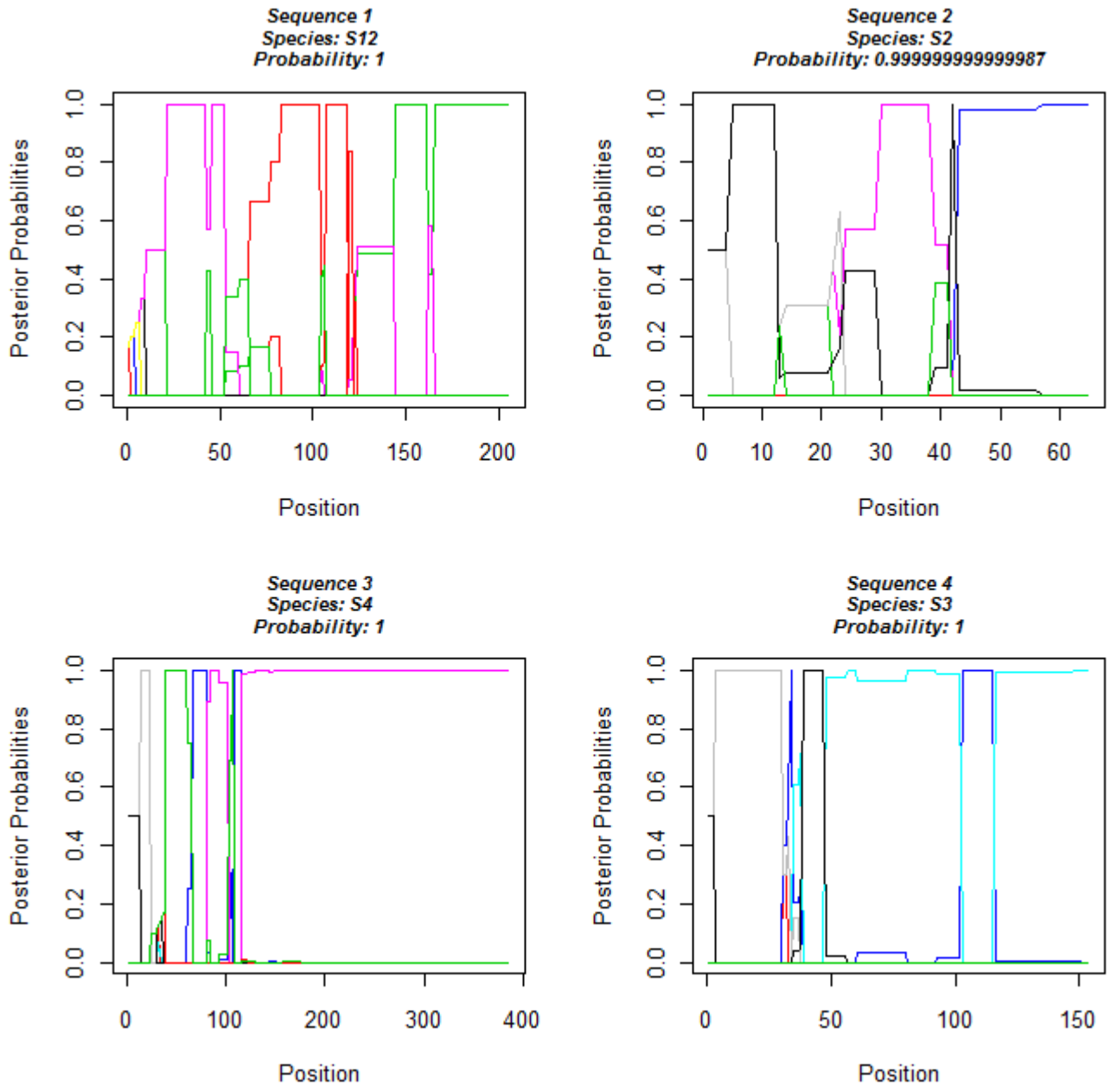

Figure C.4: *Plotted posterior probabilities having removed species 1 from the reference data set and seeking a classification of the four barcodes belong to species 1. Equal prior proababilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 8% *were used.*

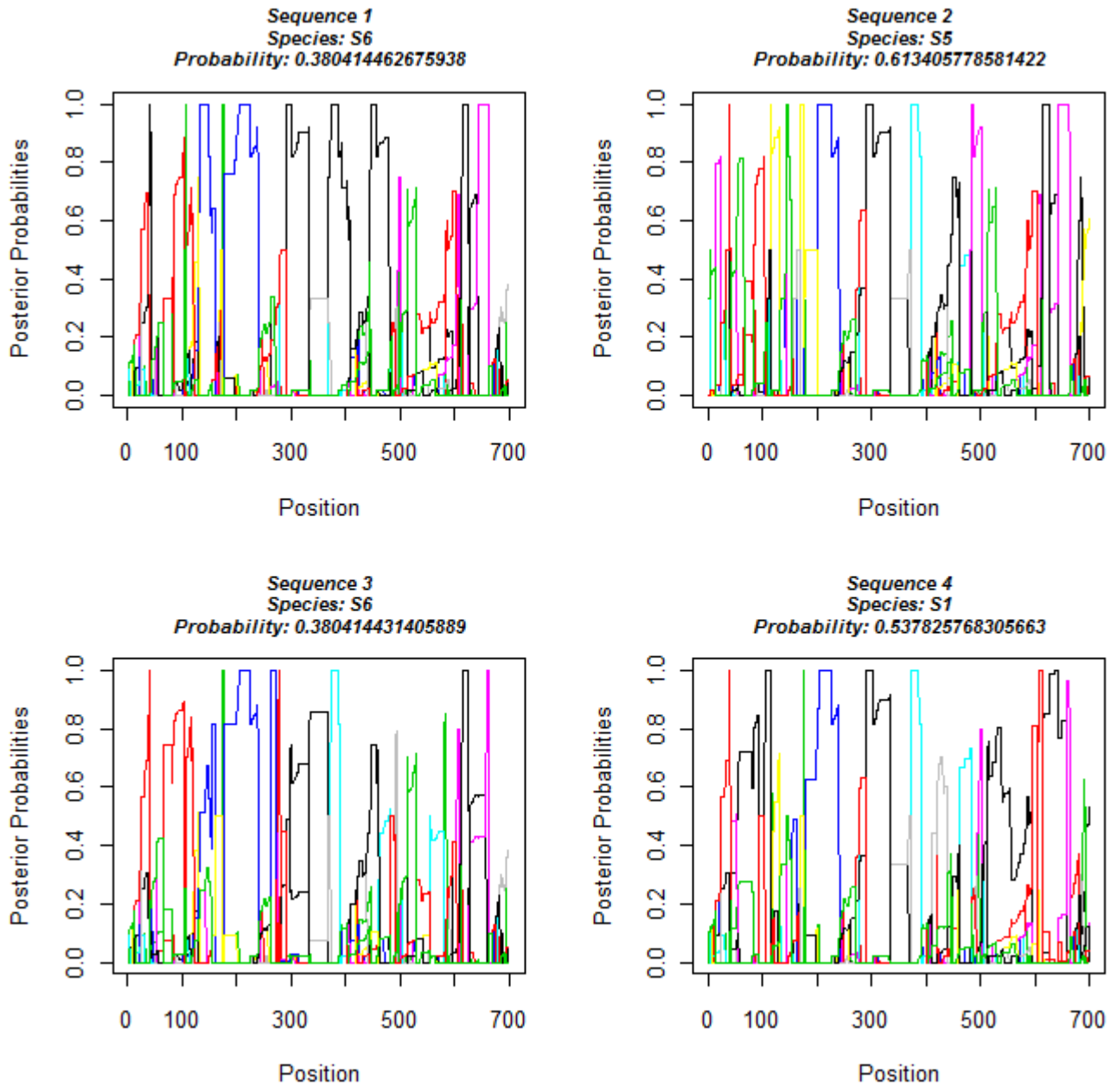

Figure C.5: *Plotted posterior probabilities having removed species 2 from the reference data set and seeking a classification of the four barcodes belong to species 2. Equal prior proabilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 2% *were used.*

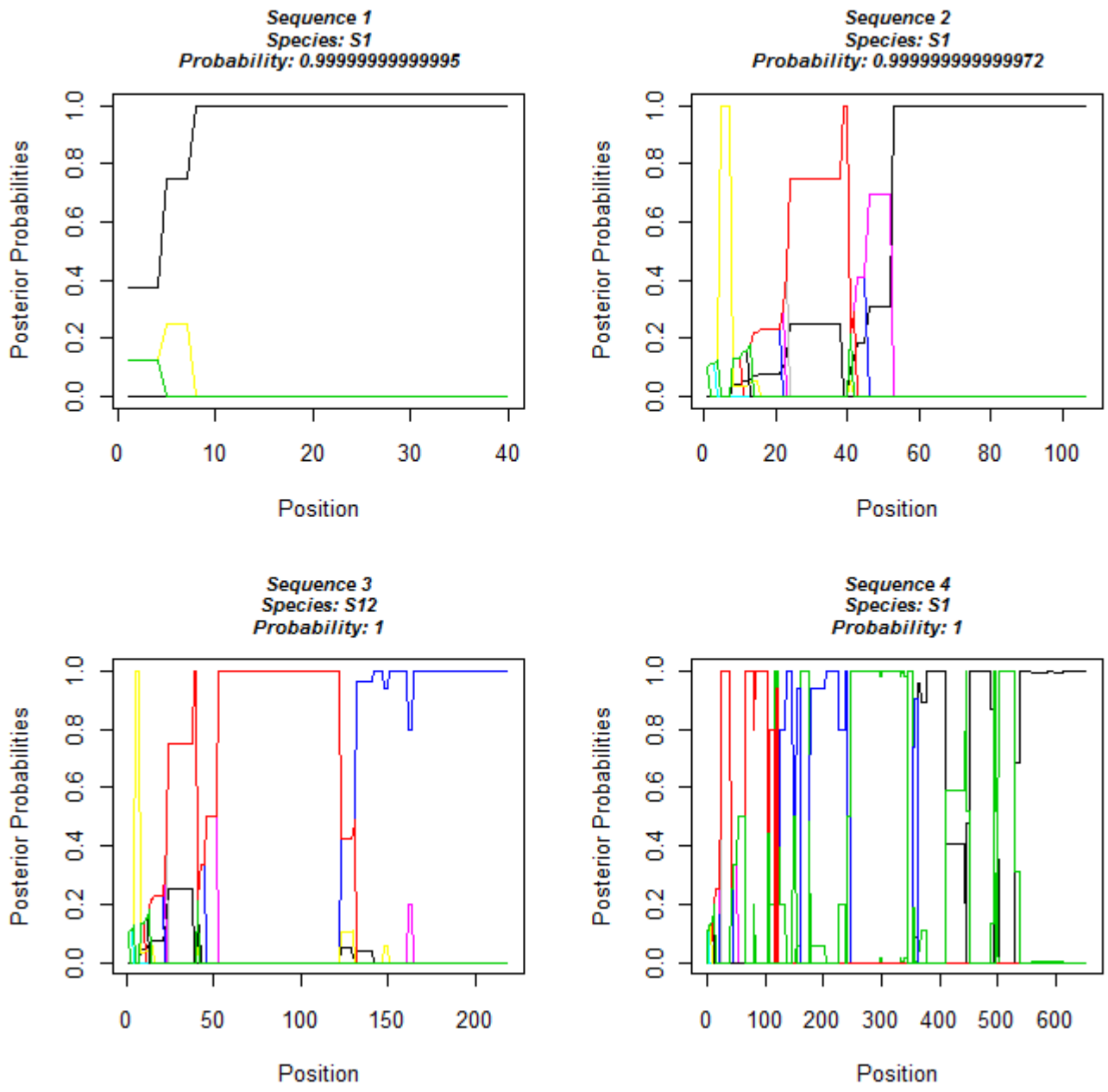

Figure C.6: *Plotted posterior probabilities having removed species 2 from the reference data set and seeking a classification of the four barcodes belong to species 2. Equal prior proabilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 4% *were used.*

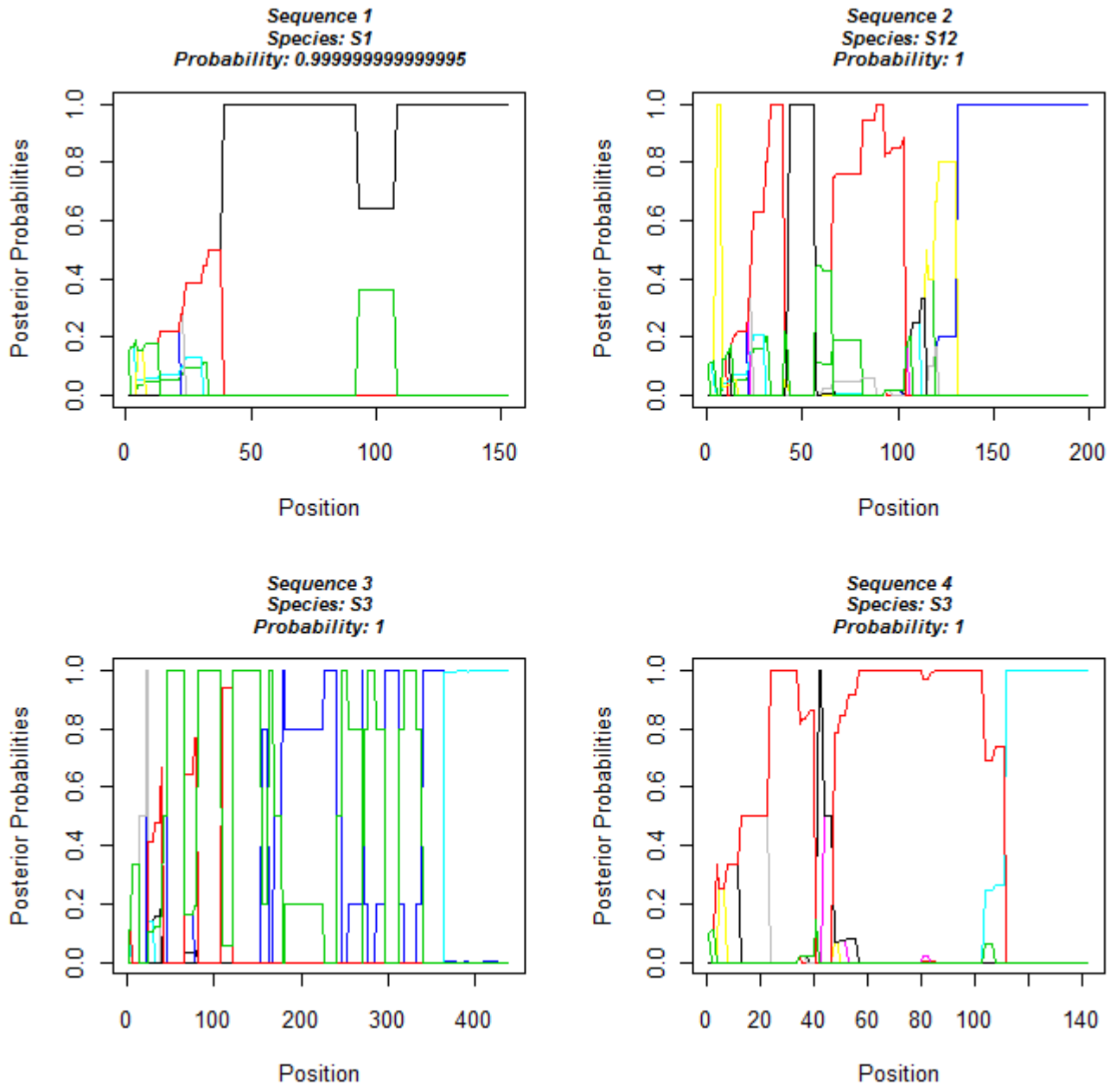

Figure C.7: *Plotted posterior probabilities having removed species 2 from the reference data set and seeking a classification of the four barcodes belong to species 2. Equal prior proabilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 6% *were used.*

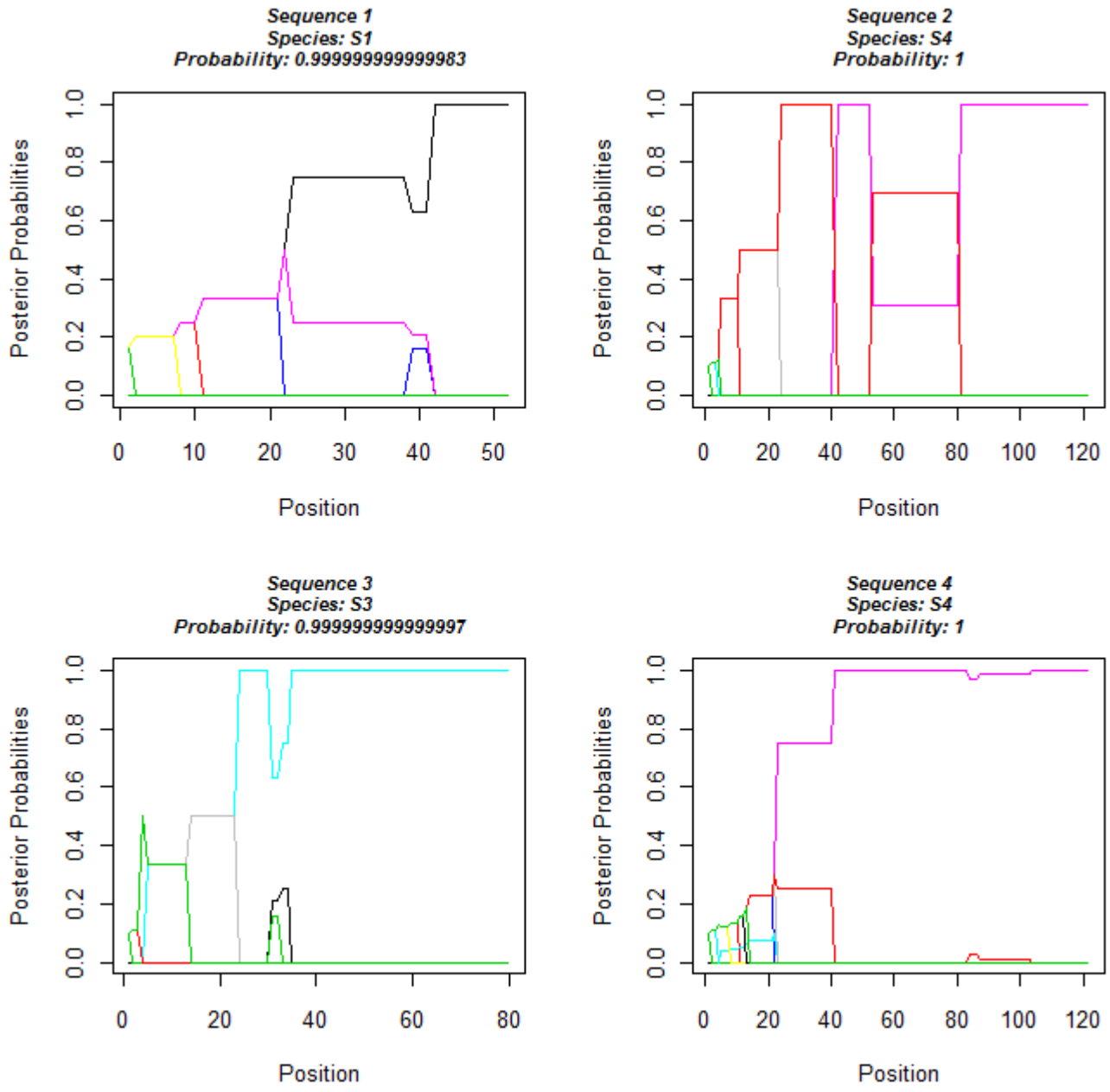

Figure C.8: *Plotted posterior probabilities having removed species 2 from the reference data set and seeking a classification of the four barcodes belong to species 2. Equal prior proababilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 8% *were used.*

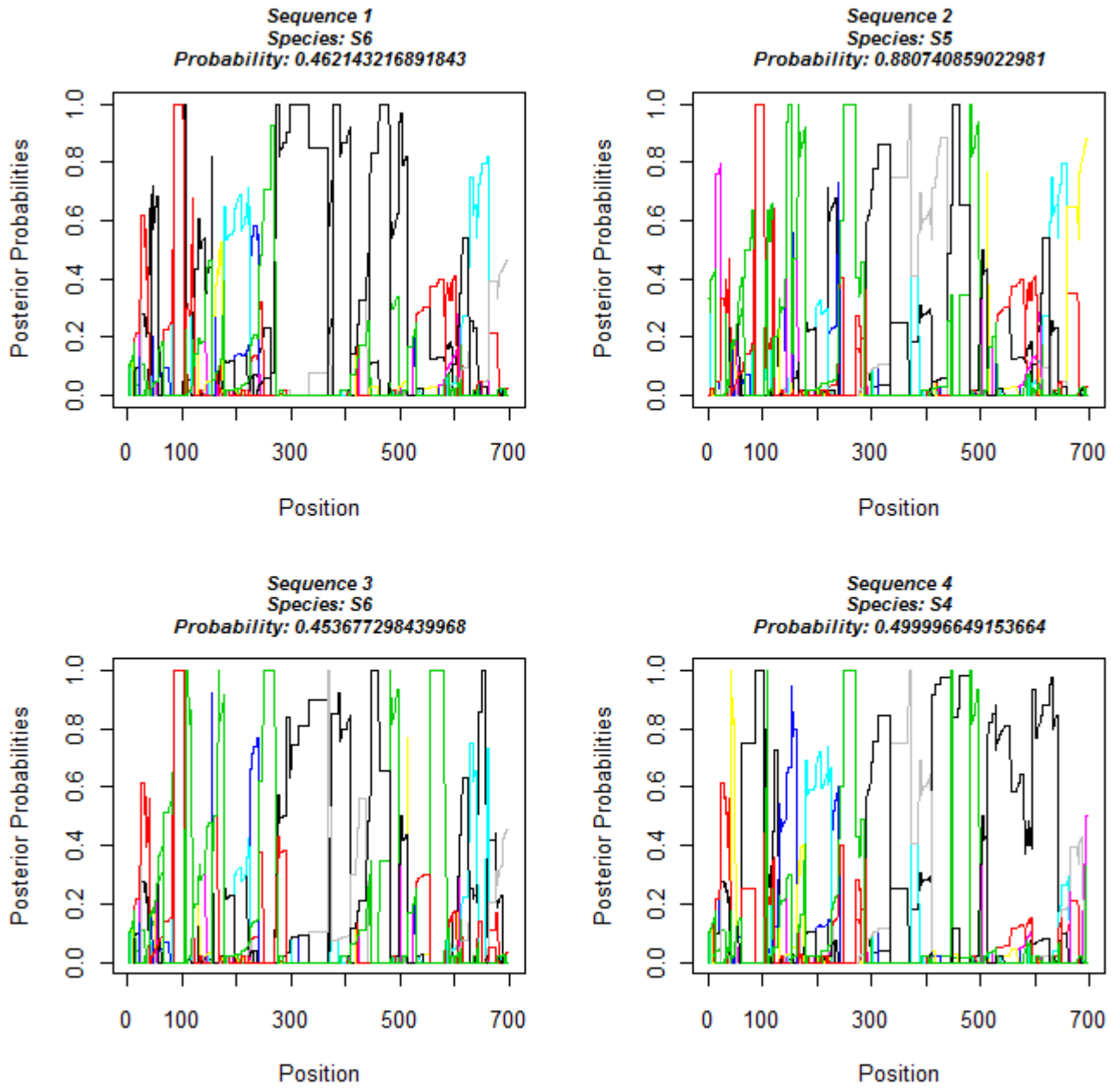

Figure C.9: *Plotted posterior probabilities having removed species 3 from the reference data set and seeking a classification of the four barcodes belong to species 3. Equal prior proabilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 2% *were used.*

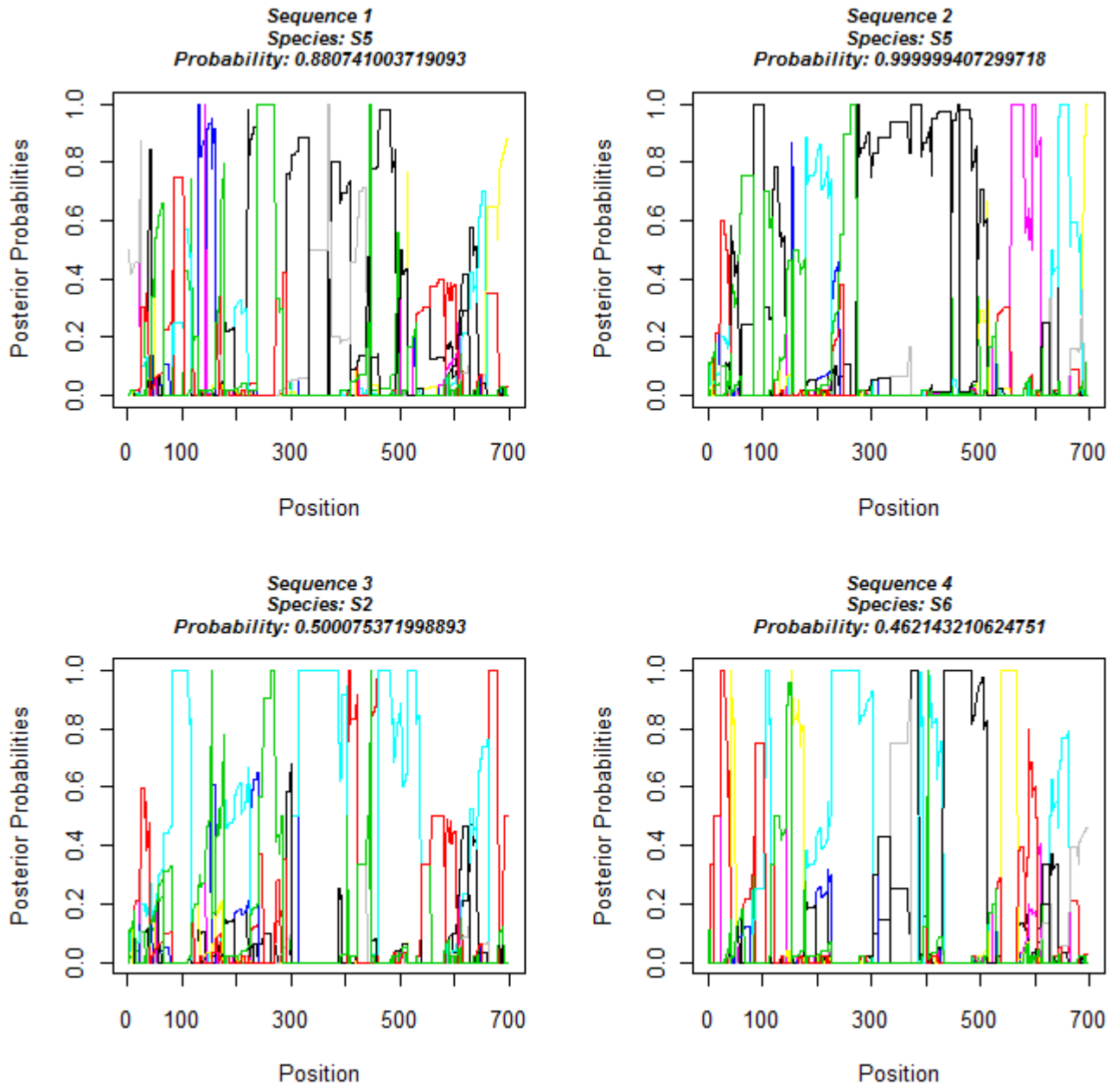

Figure C.10: *Plotted posterior probabilities having removed species 3 from the reference data set and seeking a classification of the four barcodes belong to species 3. Equal prior proabilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 4% *were used.*

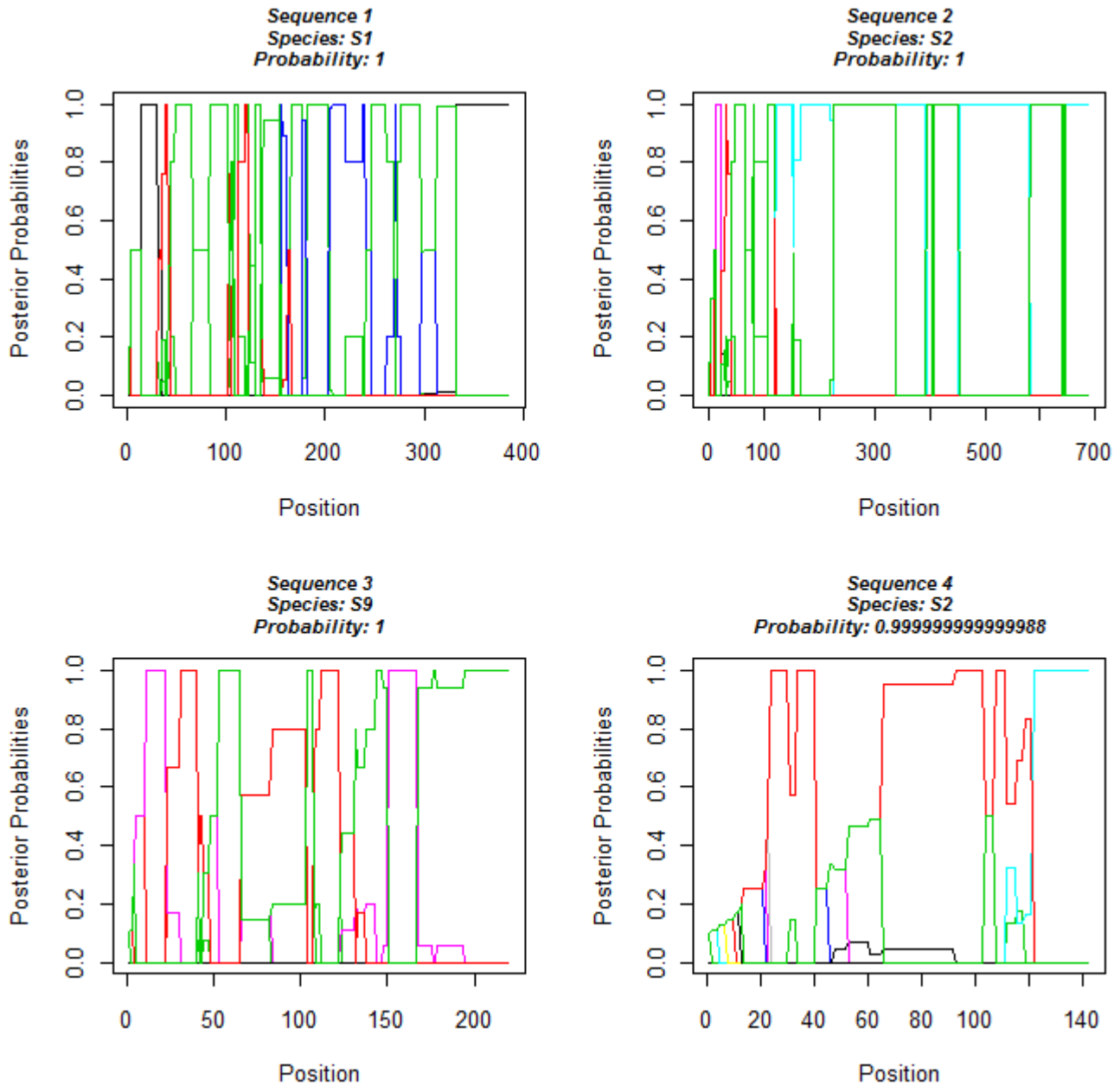

Figure C.11: *Plotted posterior probabilities having removed species 3 from the reference data set and seeking a classification of the four barcodes belong to species 3. Equal prior proabilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 6% *were used.*

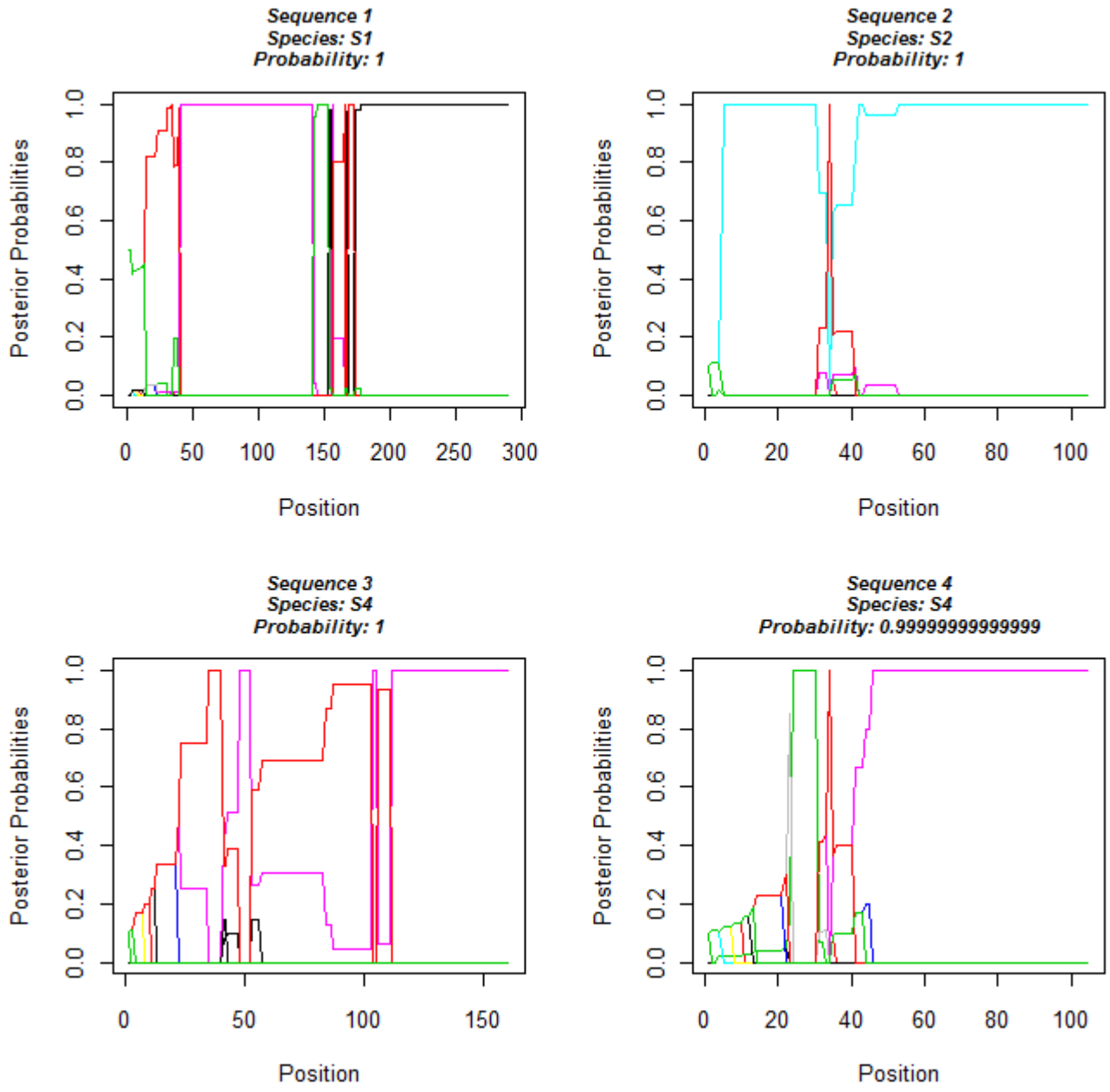

Figure C.12: *Plotted posterior probabilities having removed species 3 from the reference data set and seeking a classification of the four barcodes belong to species 3. Equal prior proabilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 8% *were used.*

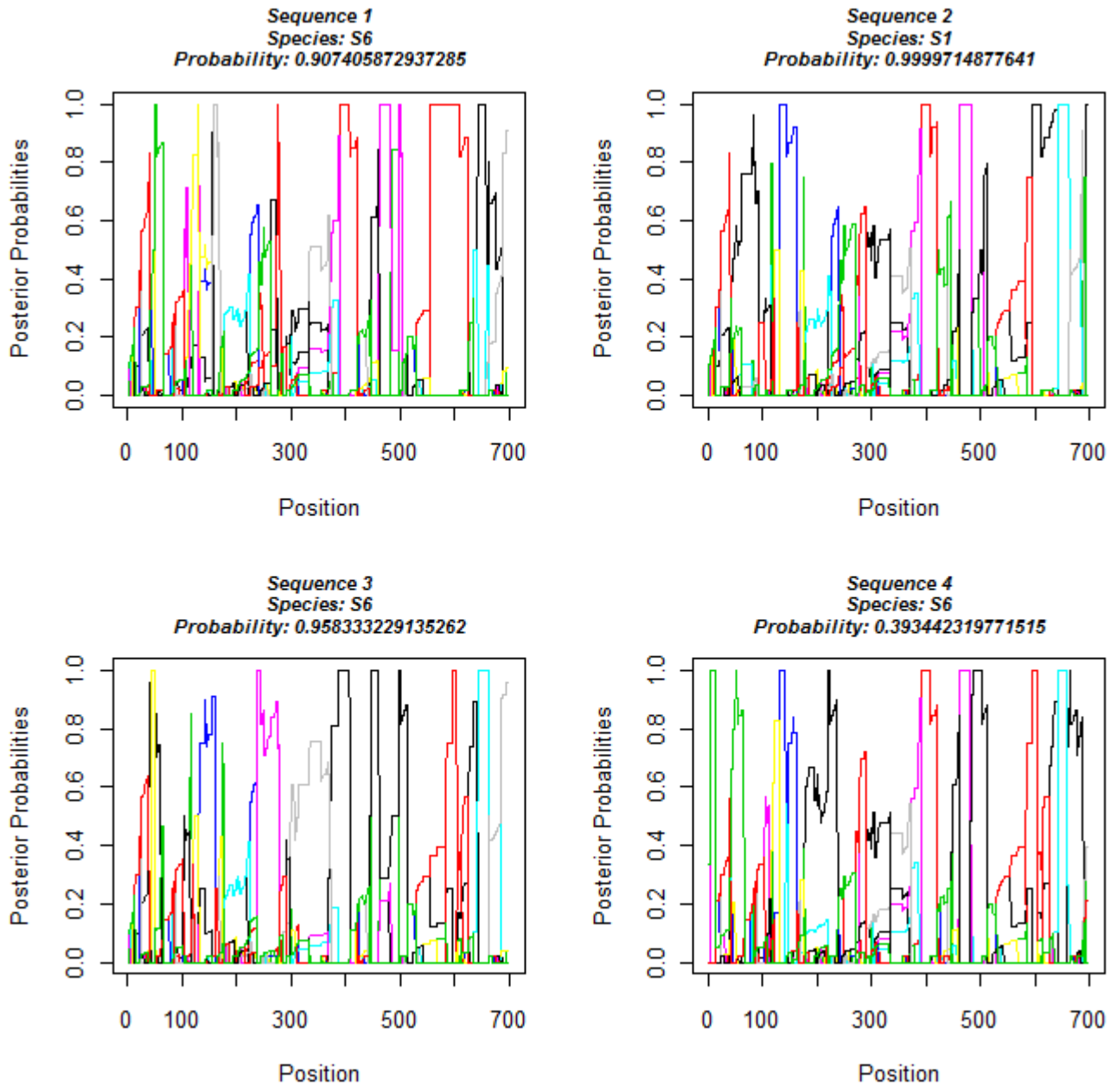

Figure C.13: *Plotted posterior probabilities having removed species 4 from the reference data set and seeking a classification of the four barcodes belong to species 4. Equal prior proababilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 2% *were used.*

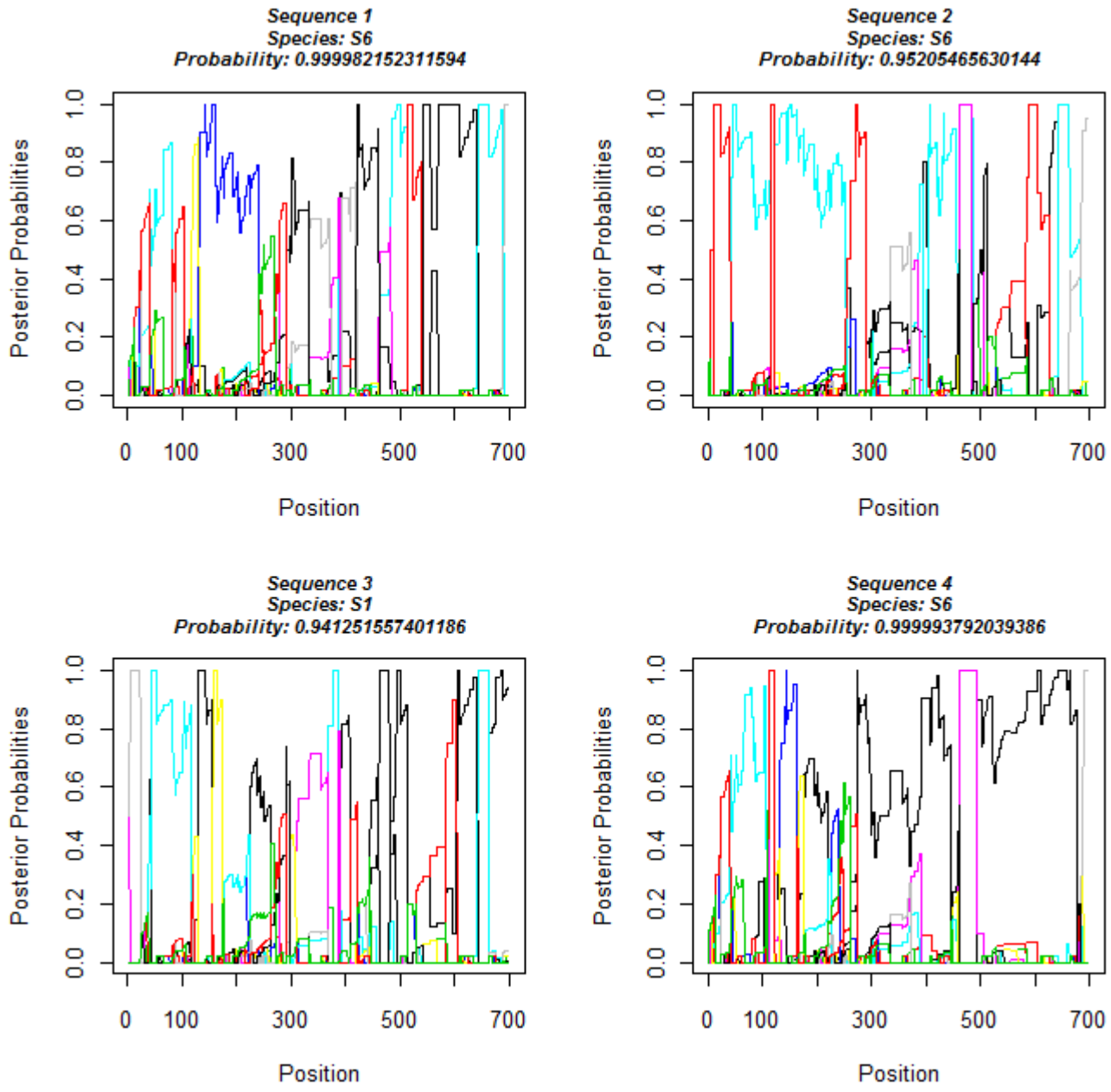

Figure C.14: *Plotted posterior probabilities having removed species 4 from the reference data set and seeking a classification of the four barcodes belong to species 4. Equal prior proabilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 4% *were used.*

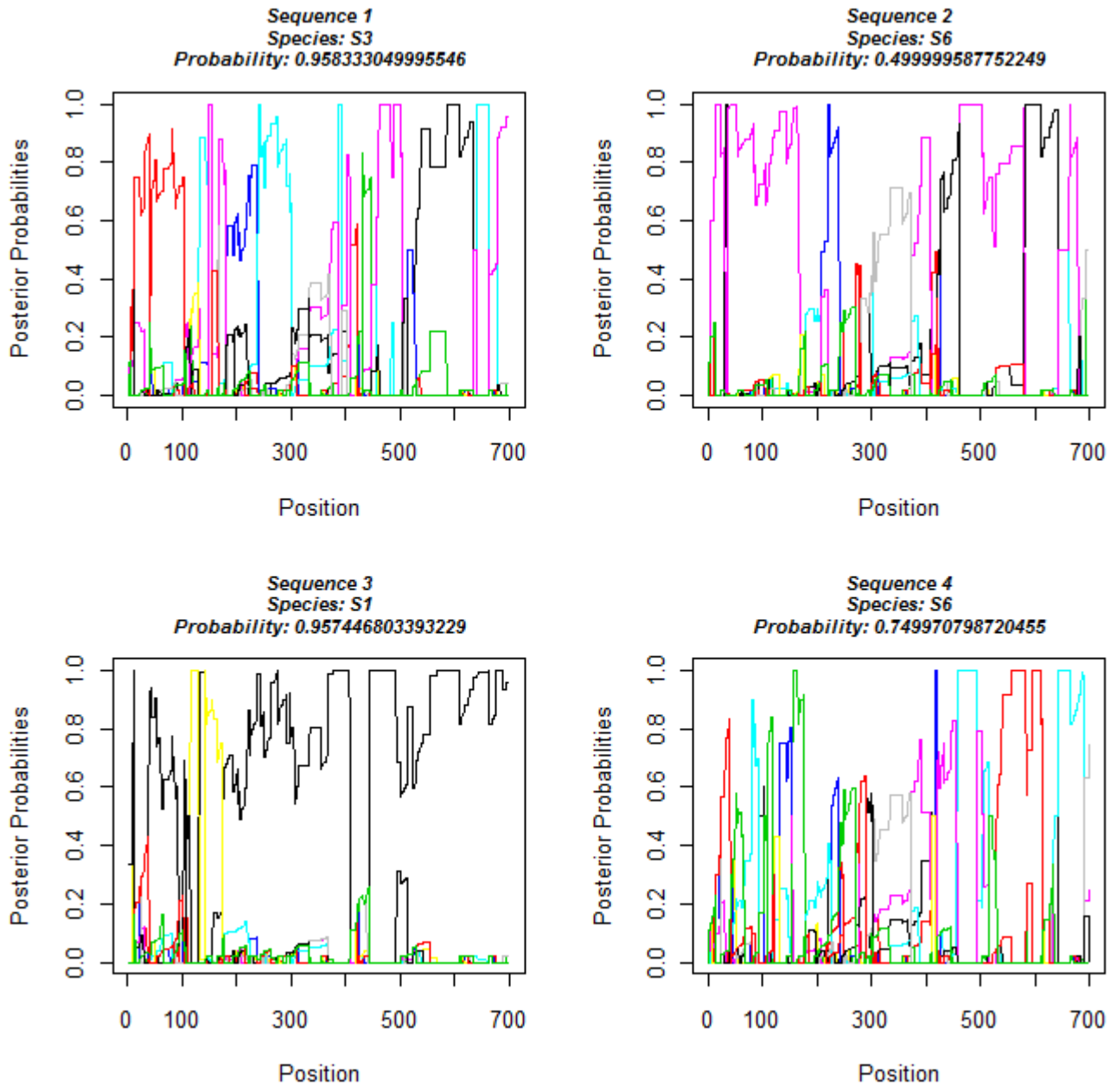

Figure C.15: *Plotted posterior probabilities having removed species 4 from the reference data set and seeking a classification of the four barcodes belong to species 4. Equal prior proabilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 6% *were used.*

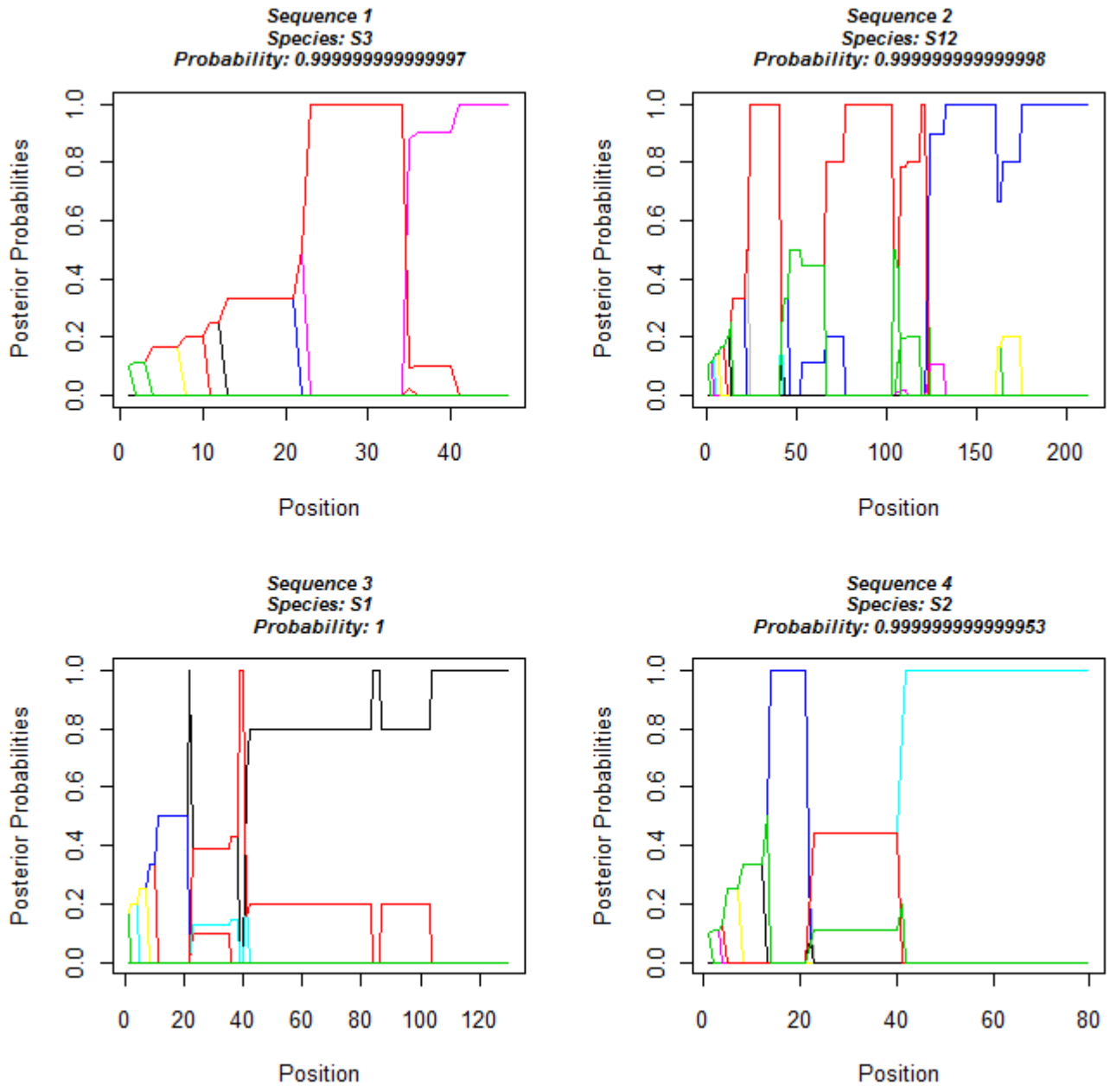

Figure C.16: *Plotted posterior probabilities having removed species 4 from the reference data set and seeking a classification of the four barcodes belong to species 4. Equal prior proabilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 8% *were used.*

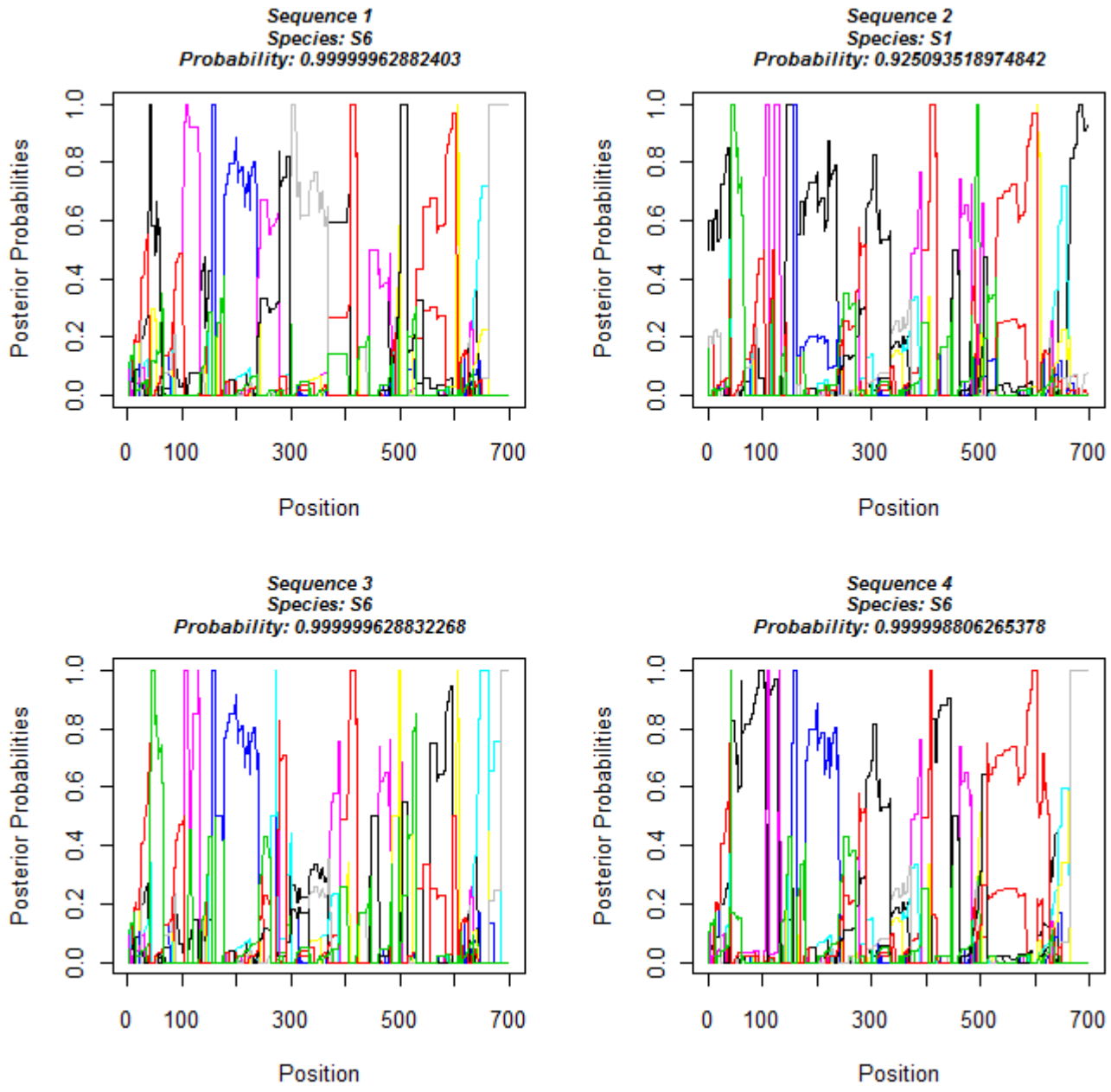

Figure C.17: *Plotted posterior probabilities having removed species 5 from the reference data set and seeking a classification of the four barcodes belong to species 5. Equal prior proababilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 2% *were used.*

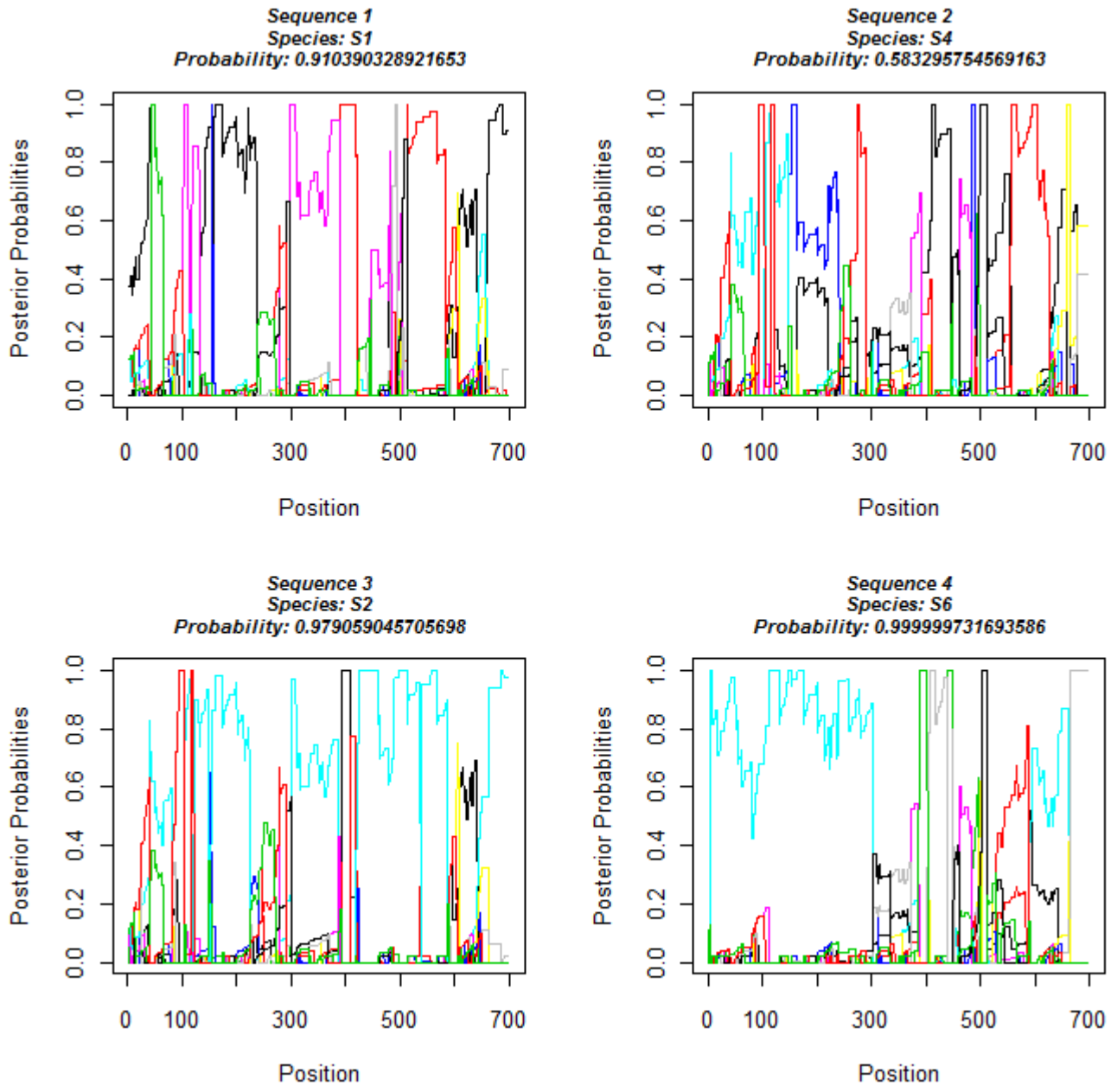

Figure C.18: *Plotted posterior probabilities having removed species 5 from the reference data set and seeking a classification of the four barcodes belong to species 5. Equal prior proabilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 4% *were used.*

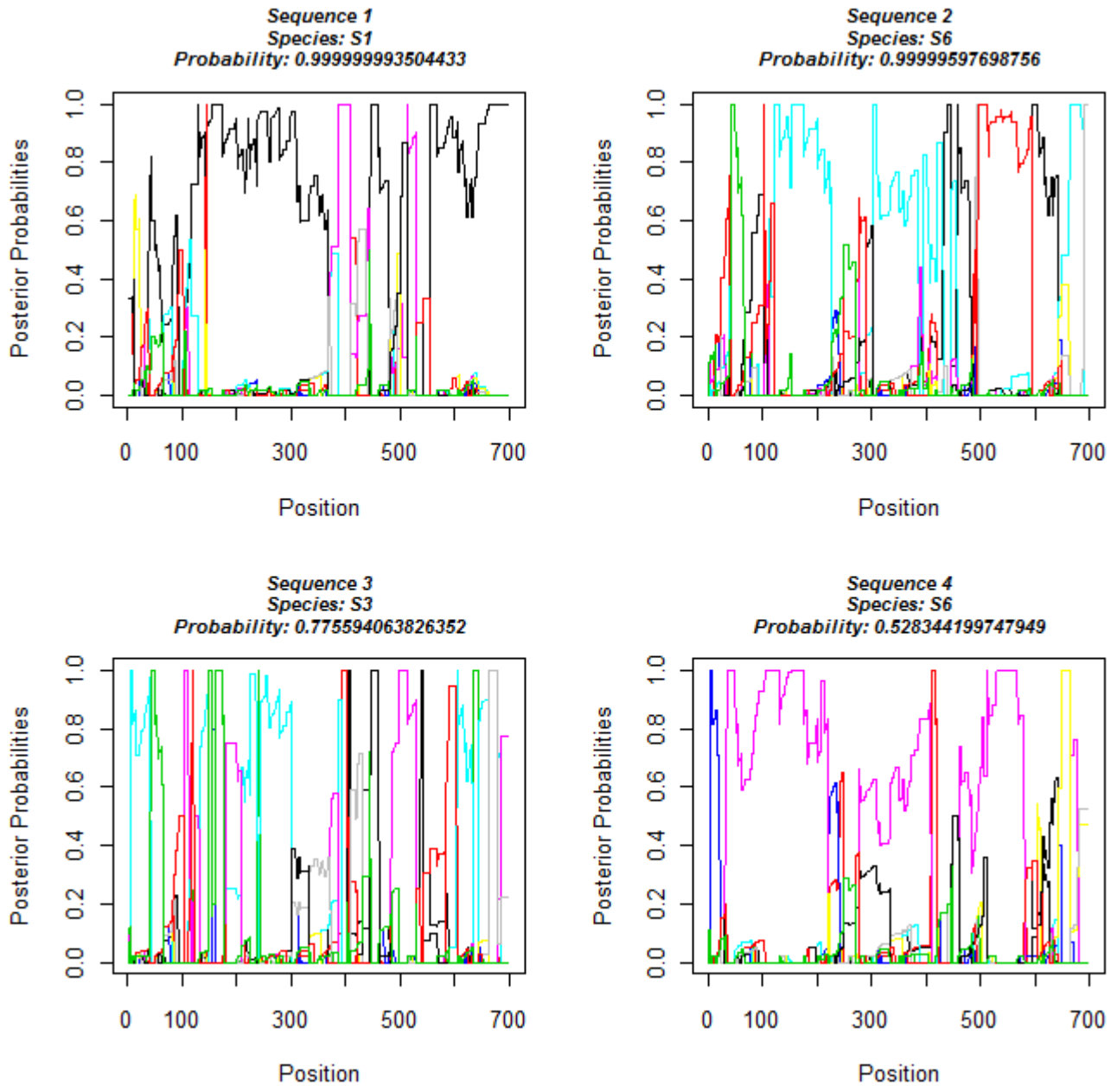

Figure C.19: *Plotted posterior probabilities having removed species 5 from the reference data set and seeking a classification of the four barcodes belong to species 5. Equal prior proabilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 6% *were used.*

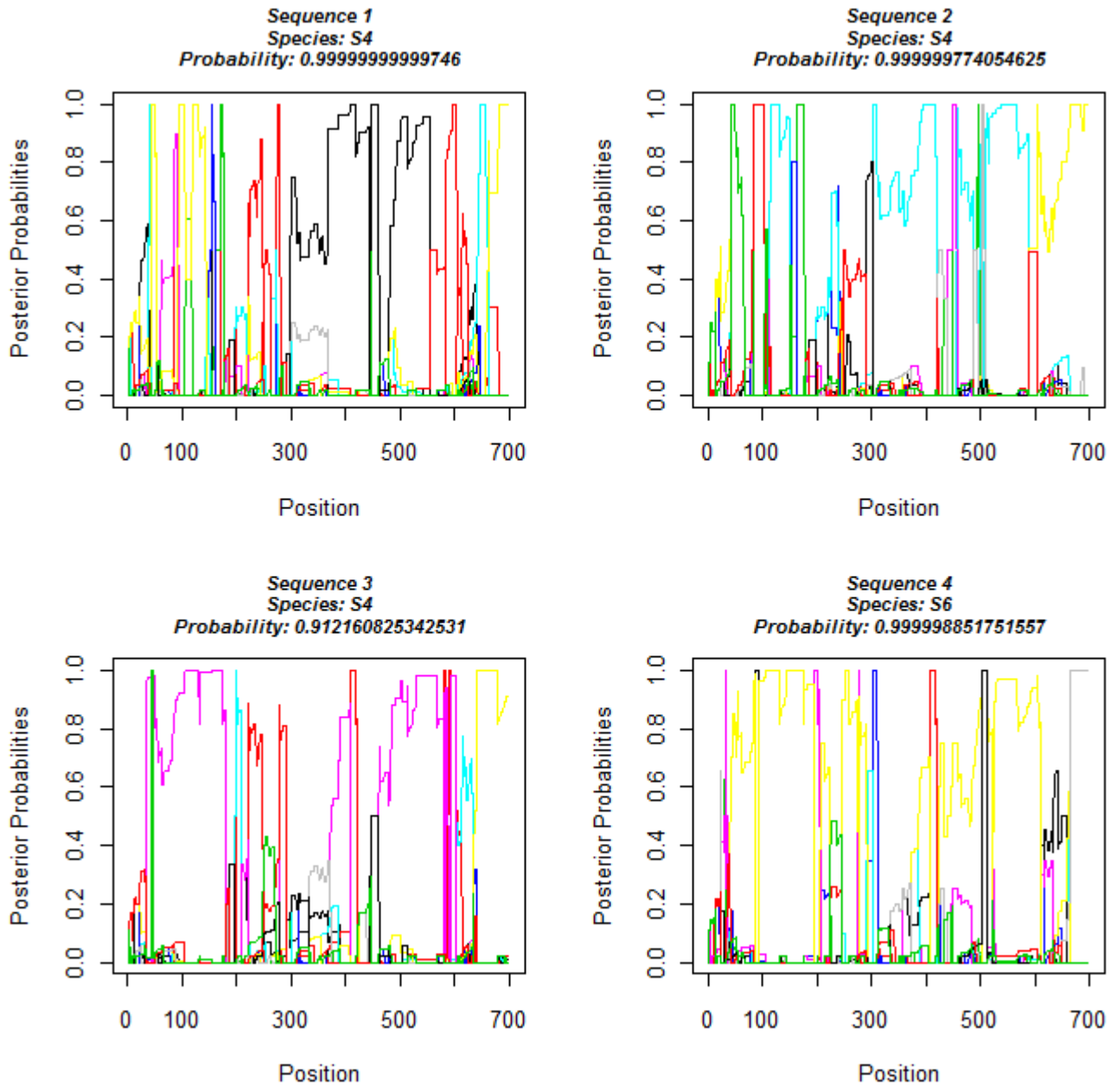

Figure C.20: *Plotted posterior probabilities having removed species 5 from the reference data set and seeking a classification of the four barcodes belong to species 5. Equal prior proabilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 8% *were used.*

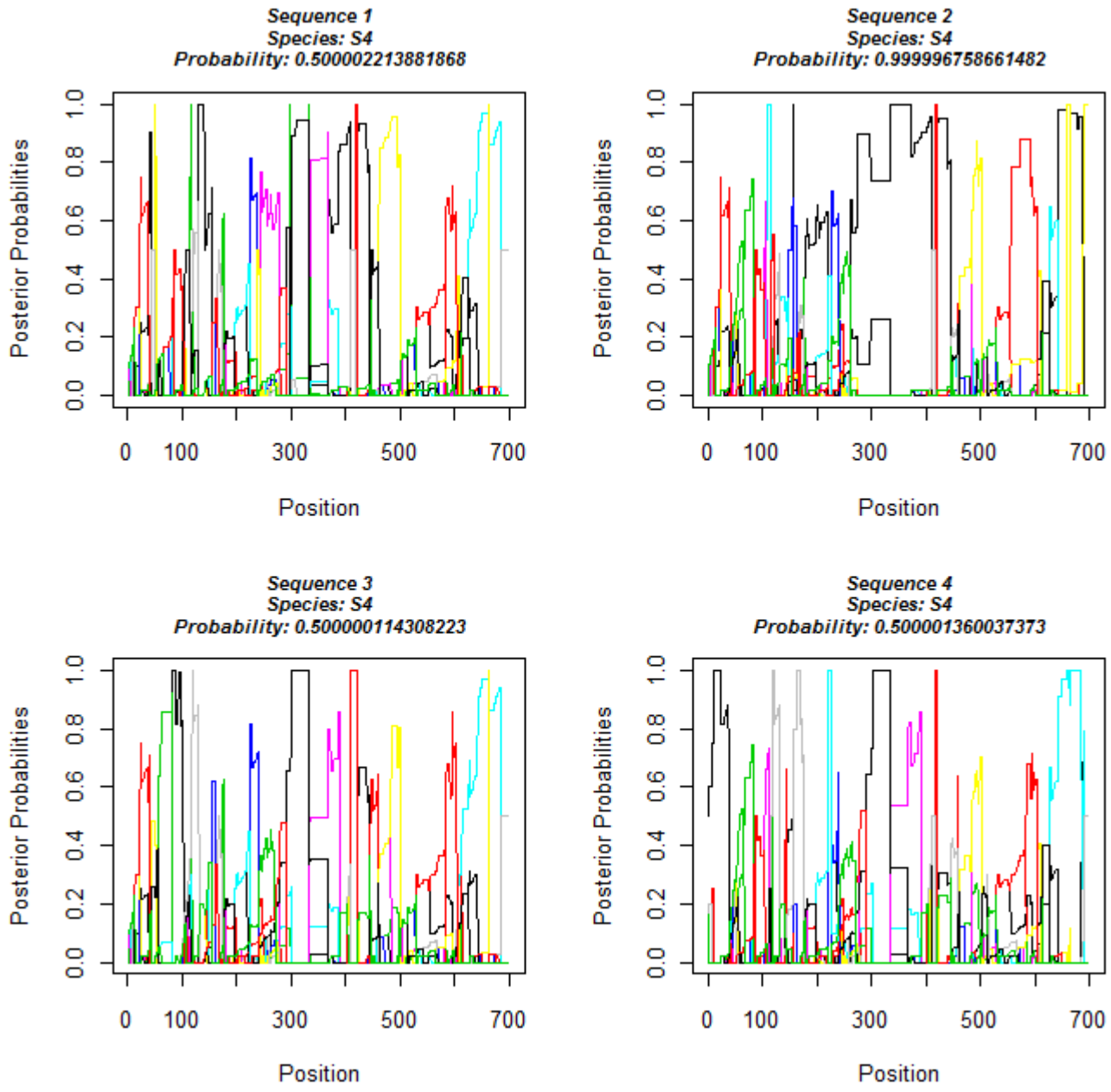

Figure C.21: *Plotted posterior probabilities having removed species 6 from the reference data set and seeking a classification of the four barcodes belong to species 6. Equal prior proabilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 2% *were used.*

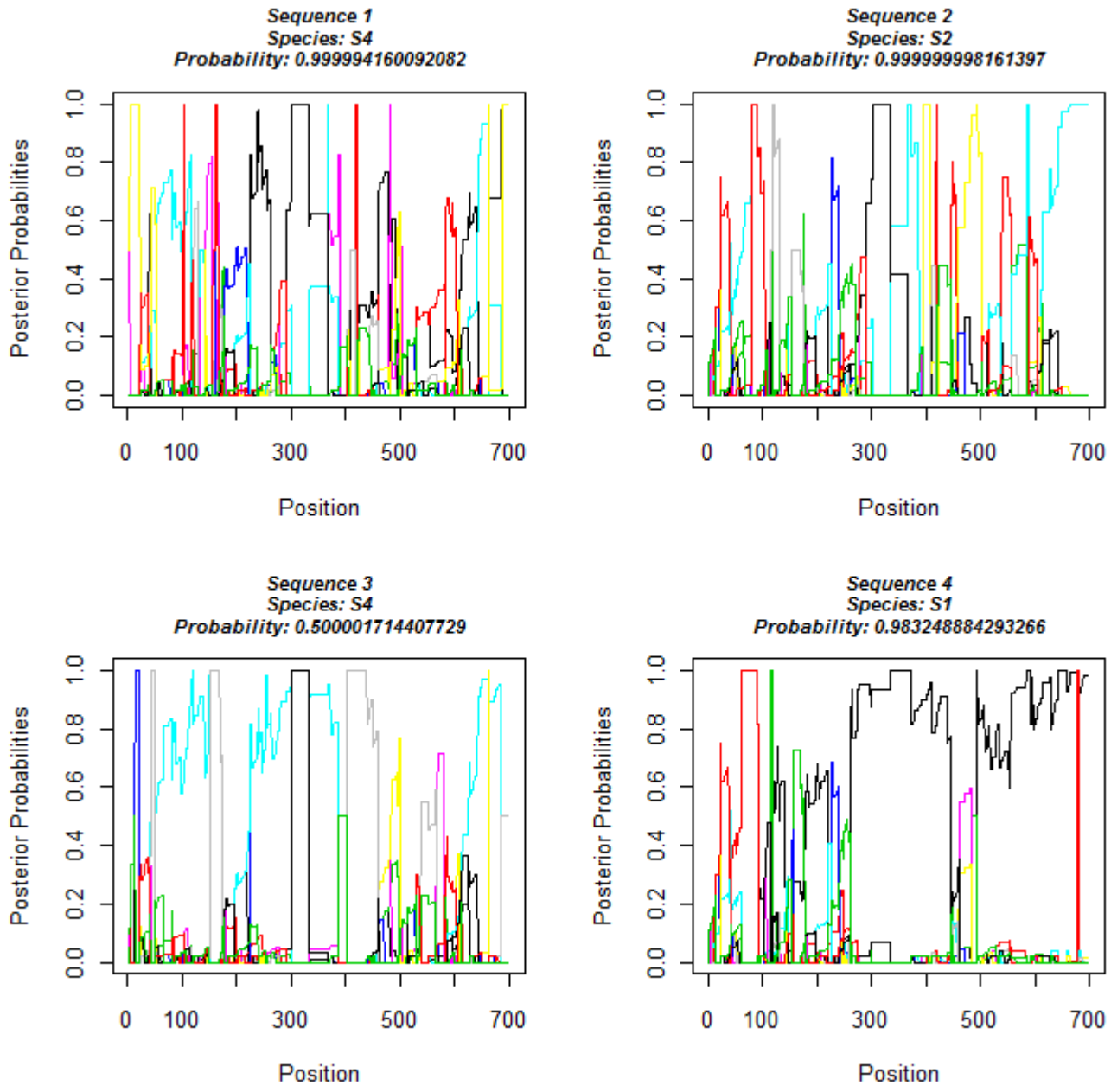

Figure C.22: *Plotted posterior probabilities having removed species 6 from the reference data set and seeking a classification of the four barcodes belong to species 6. Equal prior proabilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 4% *were used.*

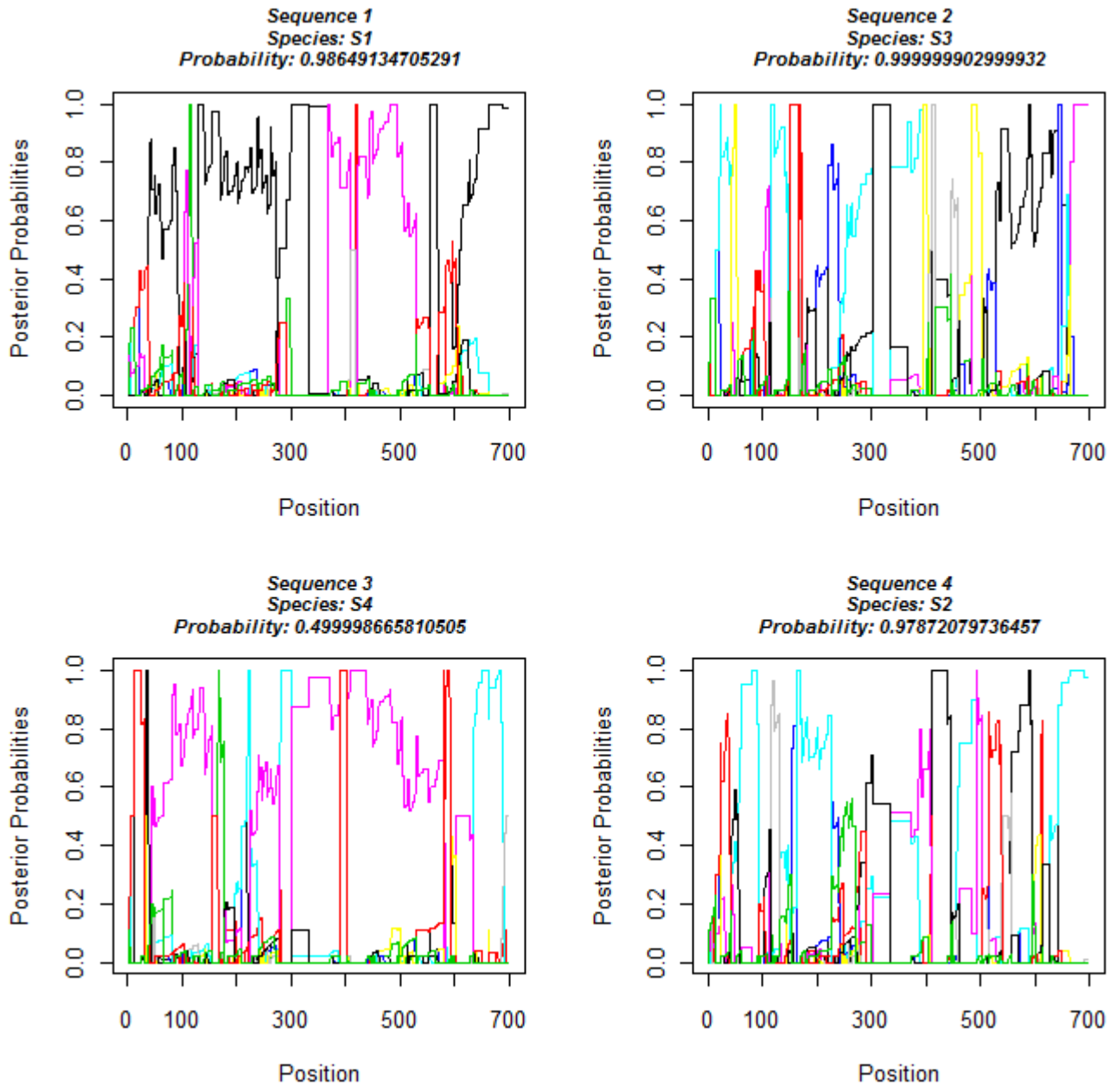

Figure C.23: *Plotted posterior probabilities having removed species 6 from the reference data set and seeking a classification of the four barcodes belong to species 6. Equal prior proabilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 6% *were used.*

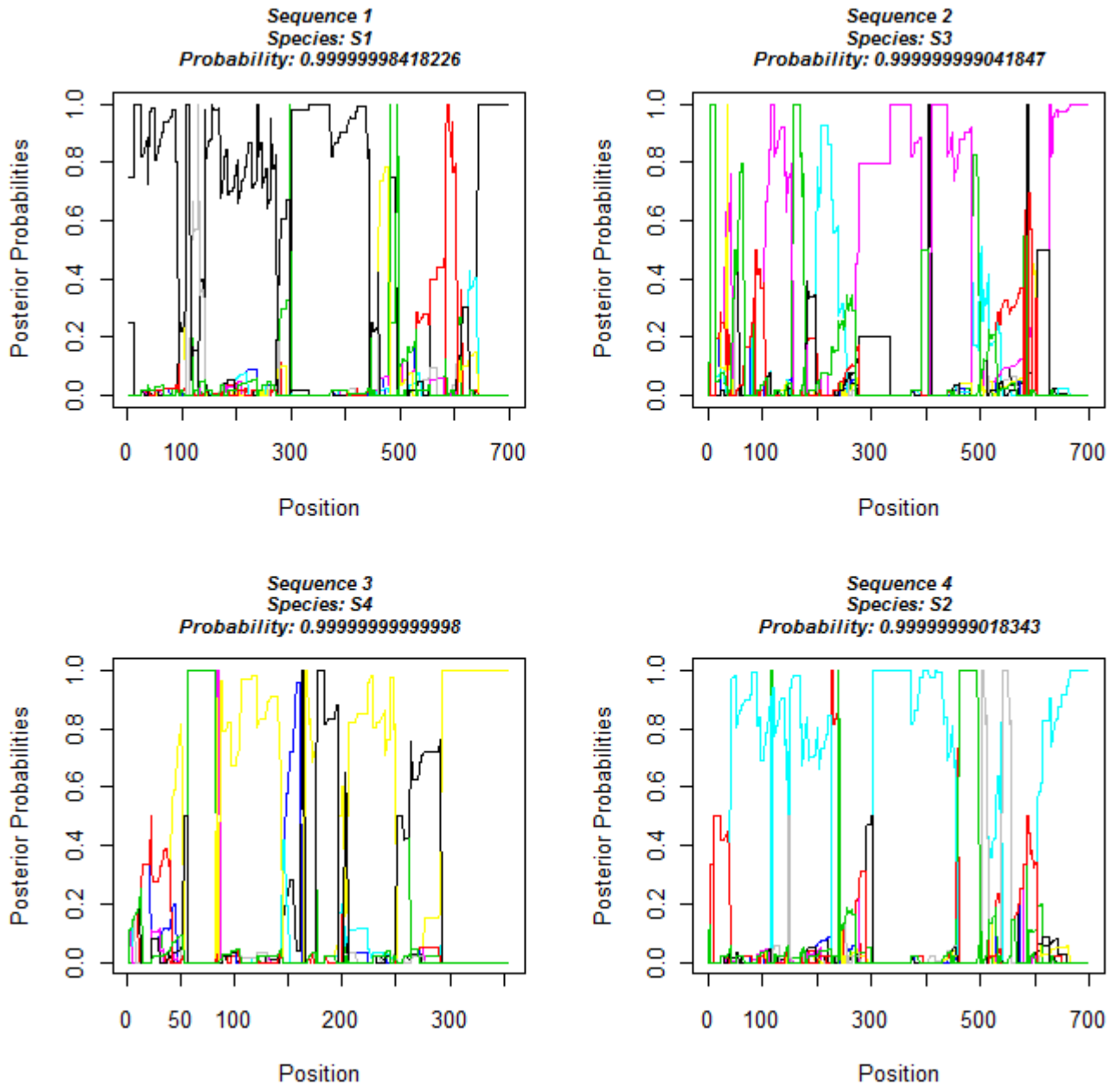

Figure C.24: *Plotted posterior probabilities having removed species 6 from the reference data set and seeking a classification of the four barcodes belong to species 6. Equal prior proabilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 8% *were used.*

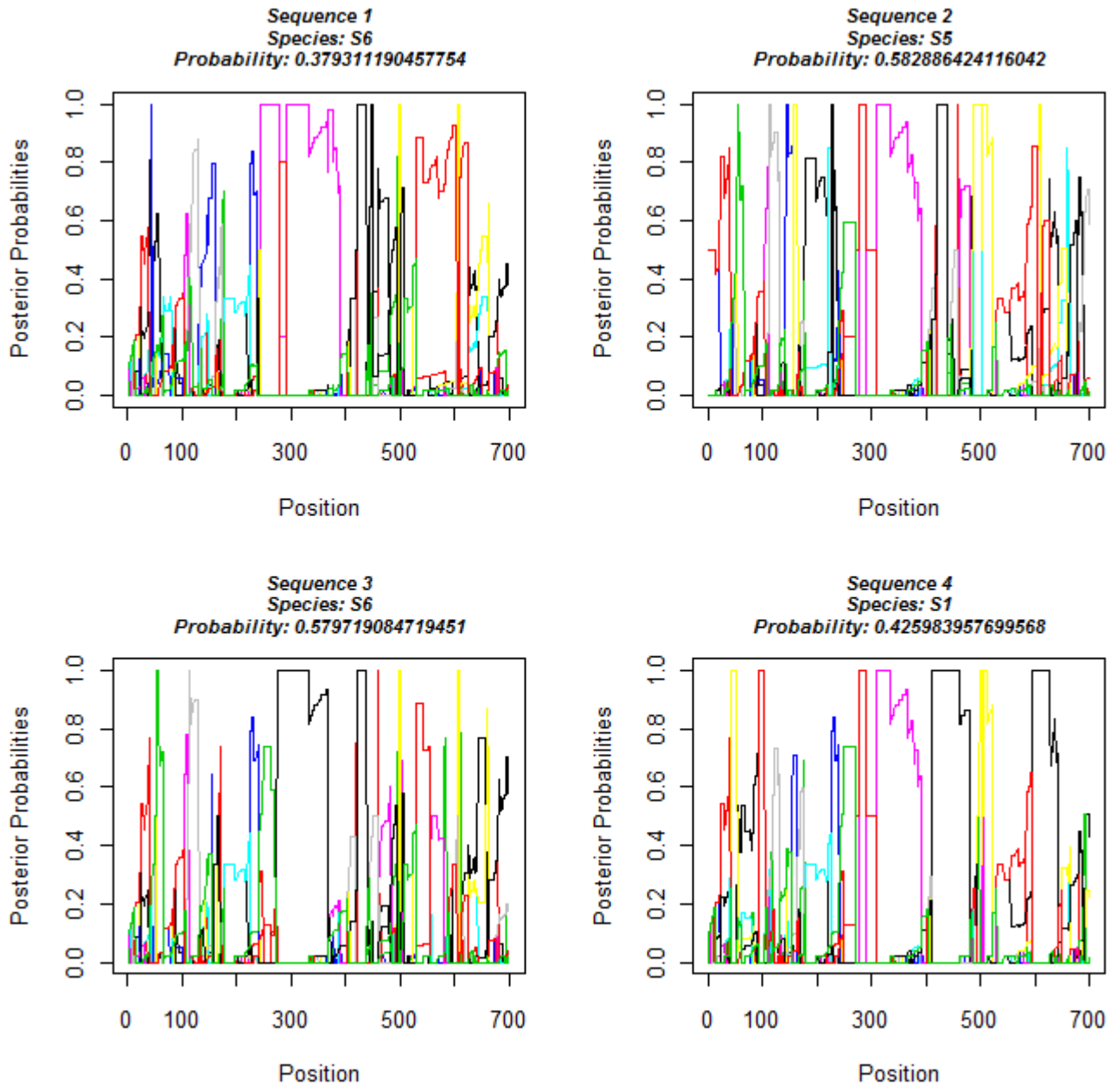

Figure C.25: *Plotted posterior probabilities having removed species 7 from the reference data set and seeking a classification of the four barcodes belong to species 7. Equal prior proabilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 2% *were used.*

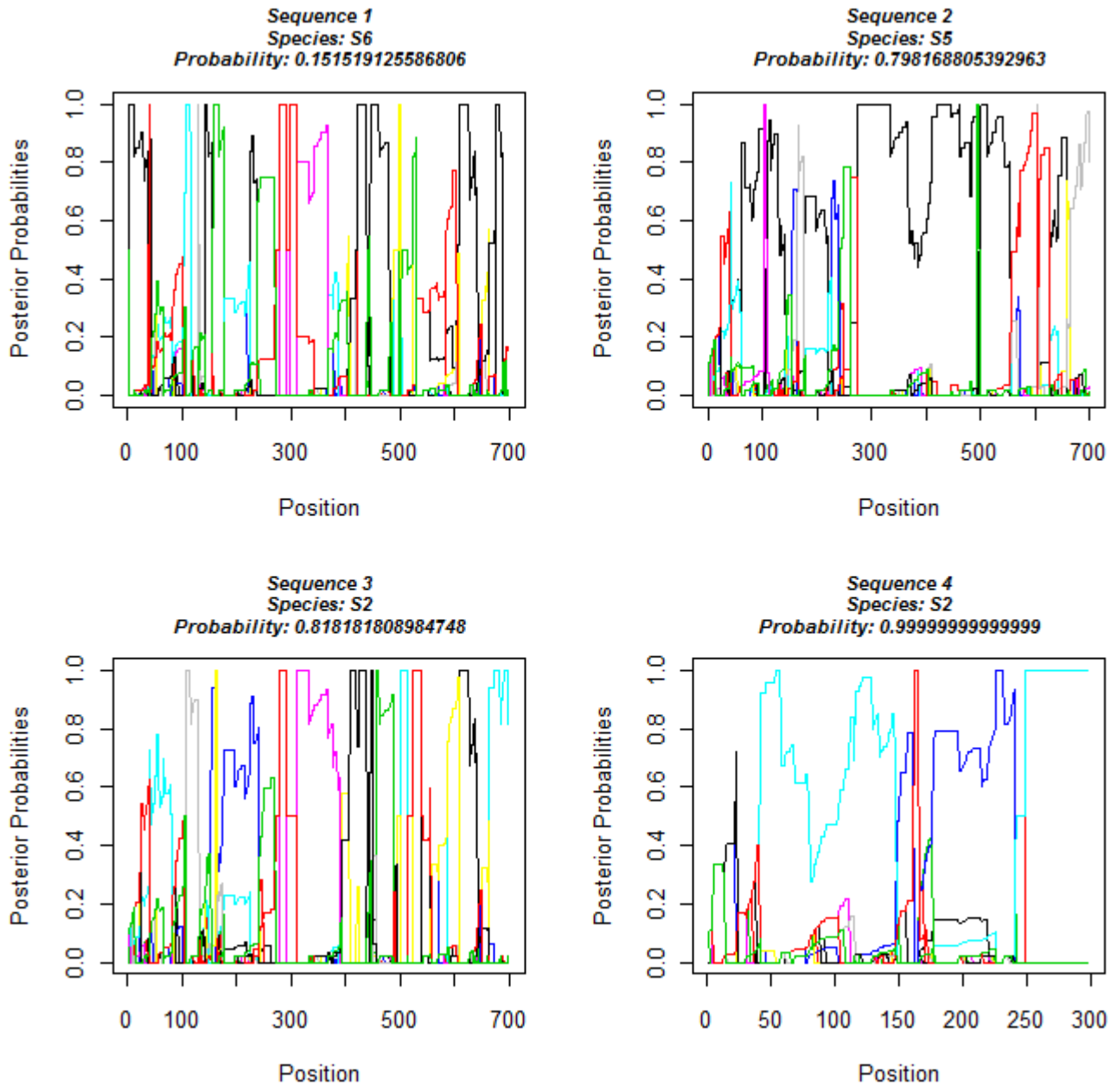

Figure C.26: *Plotted posterior probabilities having removed species 7 from the reference data set and seeking a classification of the four barcodes belong to species 7. Equal prior proabilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 4% *were used.*

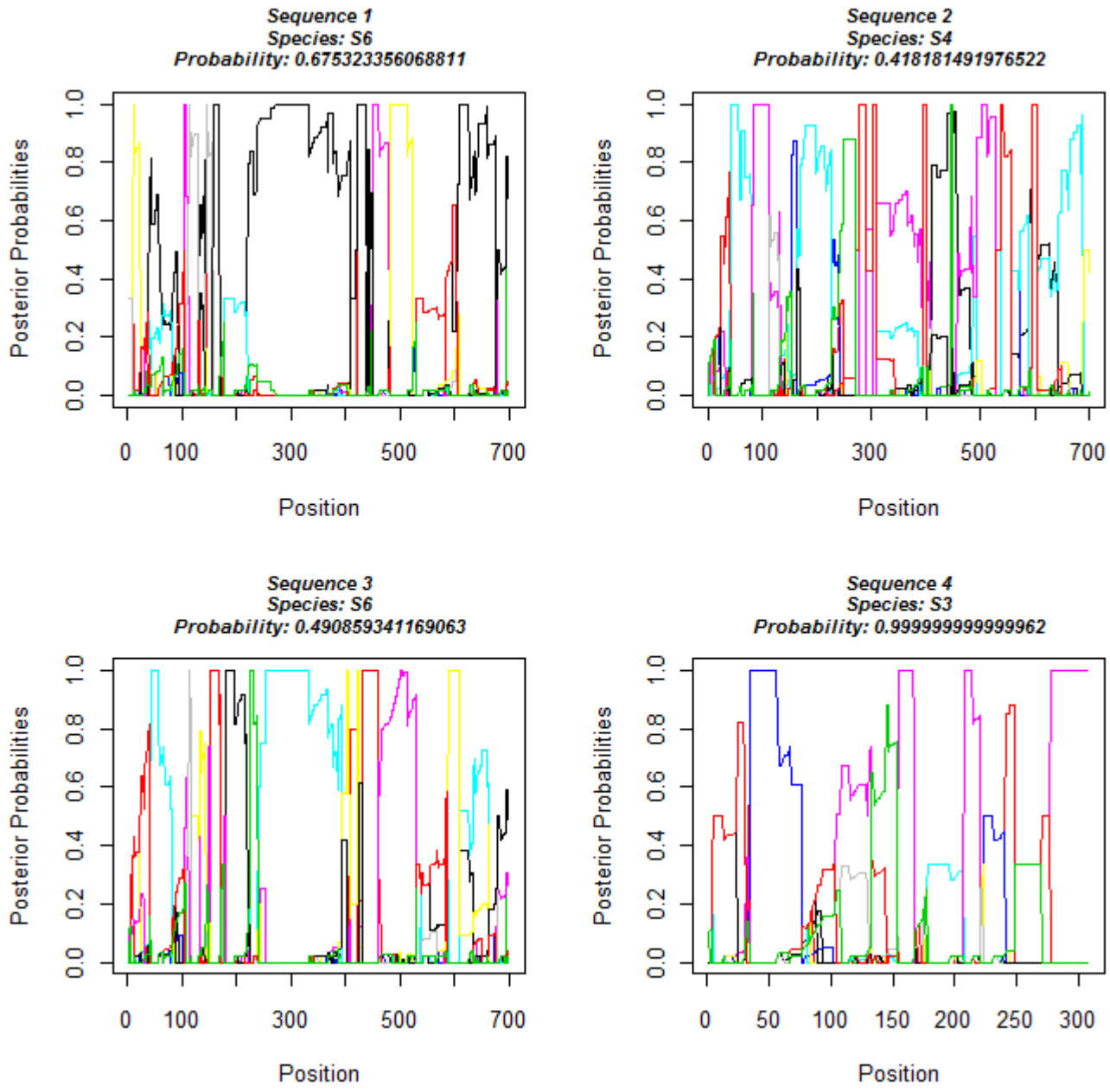

Figure C.27: *Plotted posterior probabilities having removed species 7 from the reference data set and seeking a classification of the four barcodes belong to species 7. Equal prior proabilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 6% *were used.*

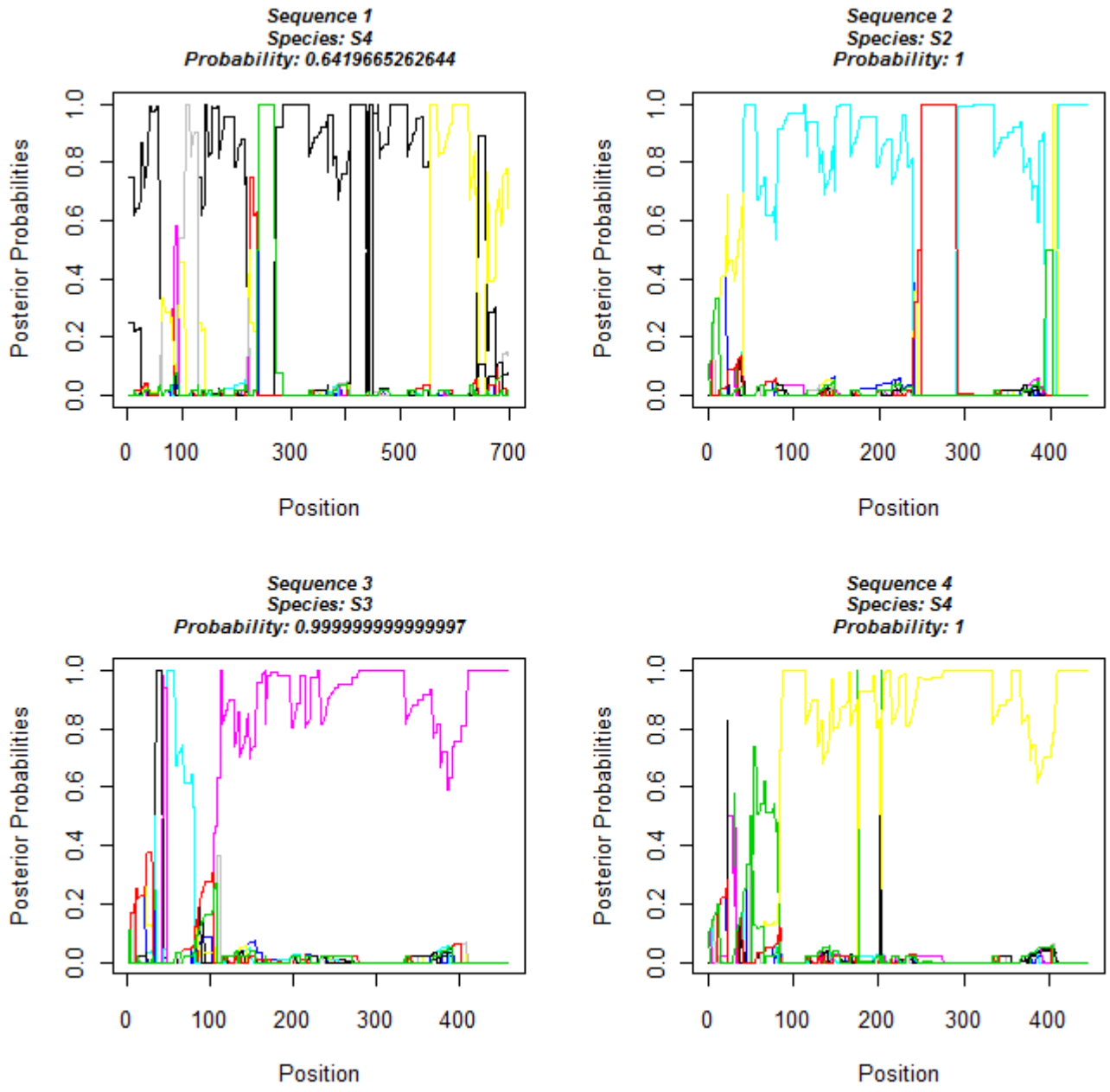

Figure C.28: *Plotted posterior probabilities having removed species 7 from the reference data set and seeking a classification of the four barcodes belong to species 7. Equal prior proabilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 8% *were used.*

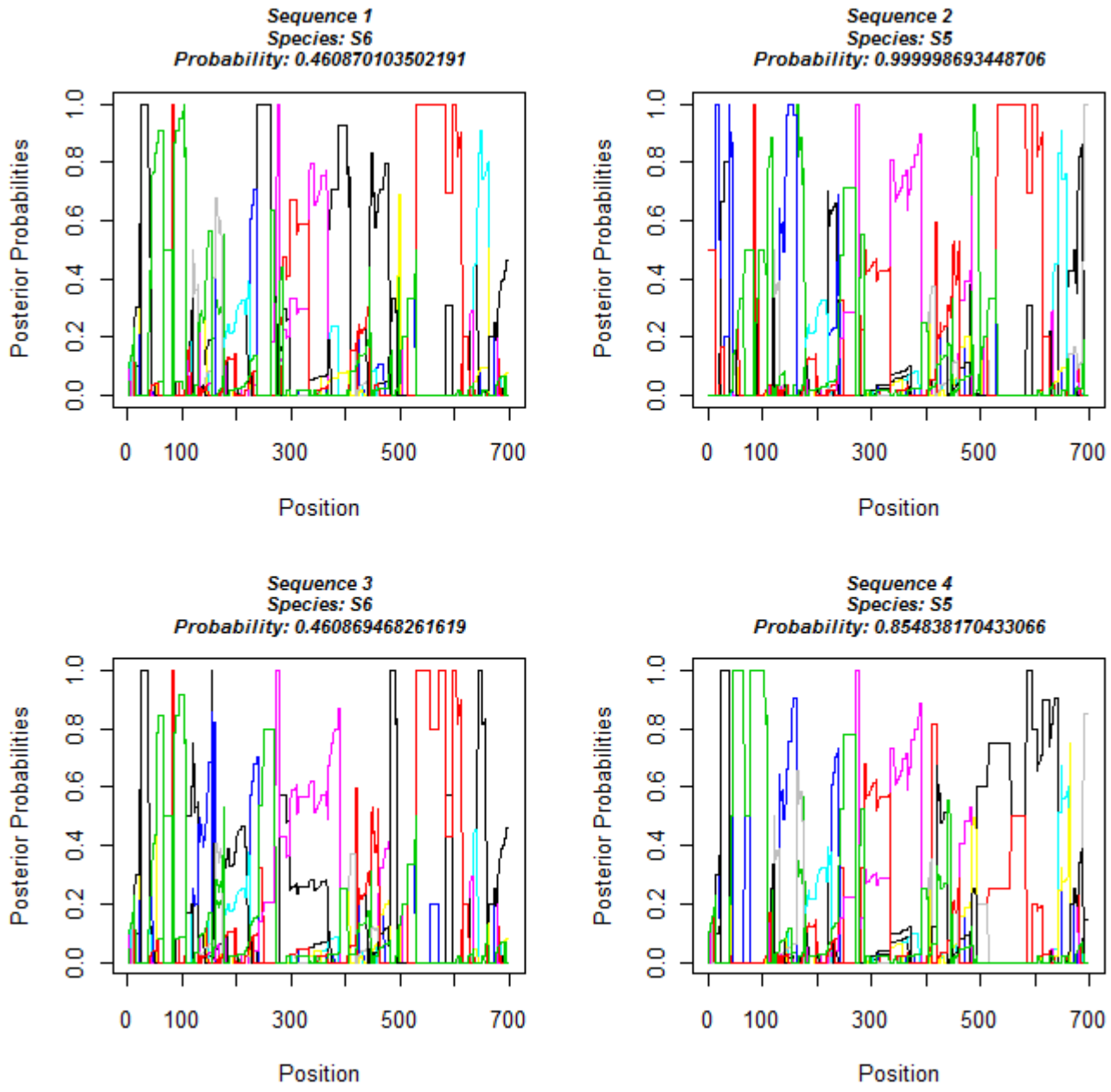

Figure C.29: *Plotted posterior probabilities having removed species 8 from the reference data set and seeking a classification of the four barcodes belong to species 8. Equal prior proabilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 2% *were used.*

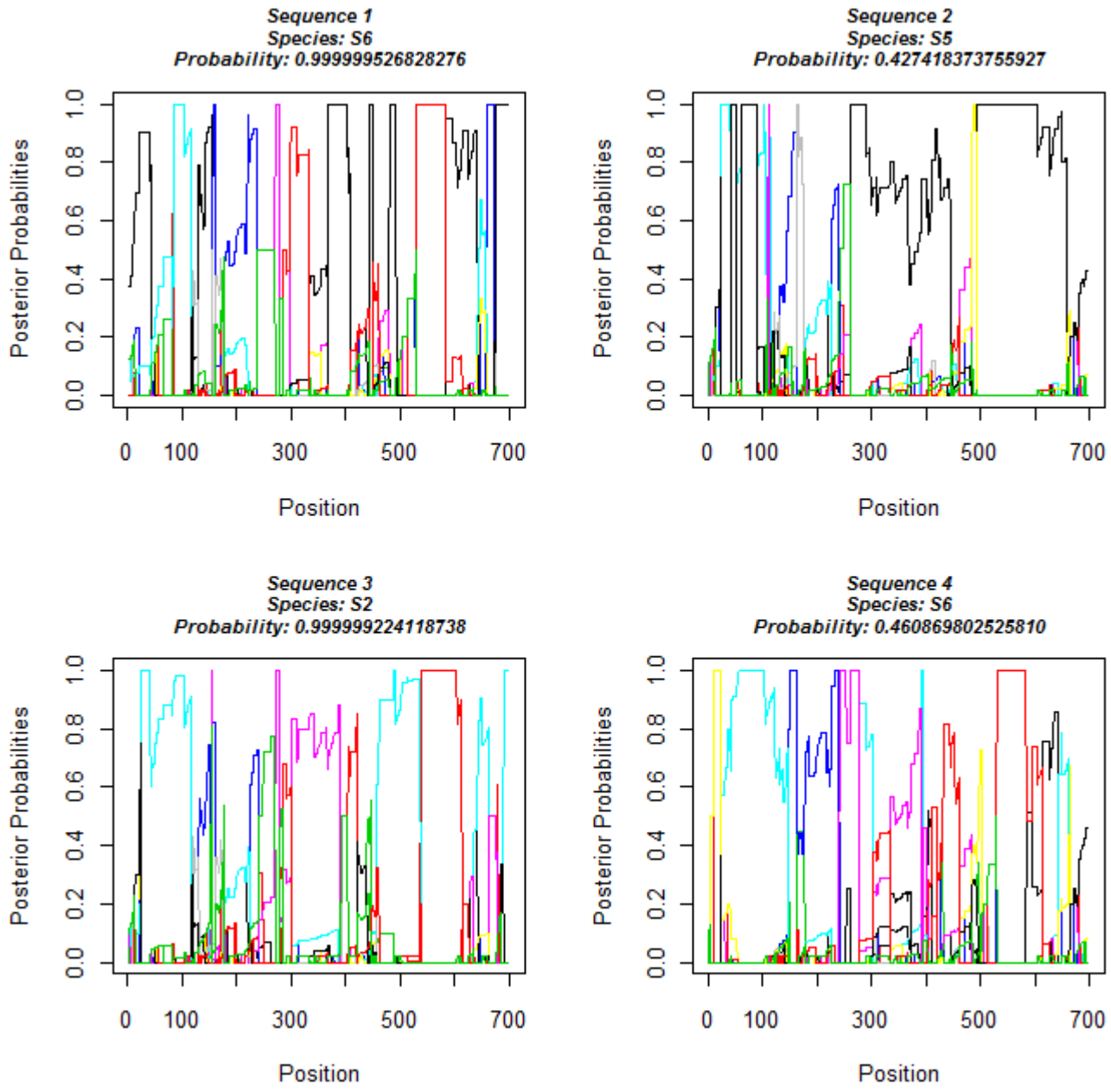

Figure C.30: *Plotted posterior probabilities having removed species 8 from the reference data set and seeking a classification of the four barcodes belong to species 8. Equal prior proabilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 4% *were used.*

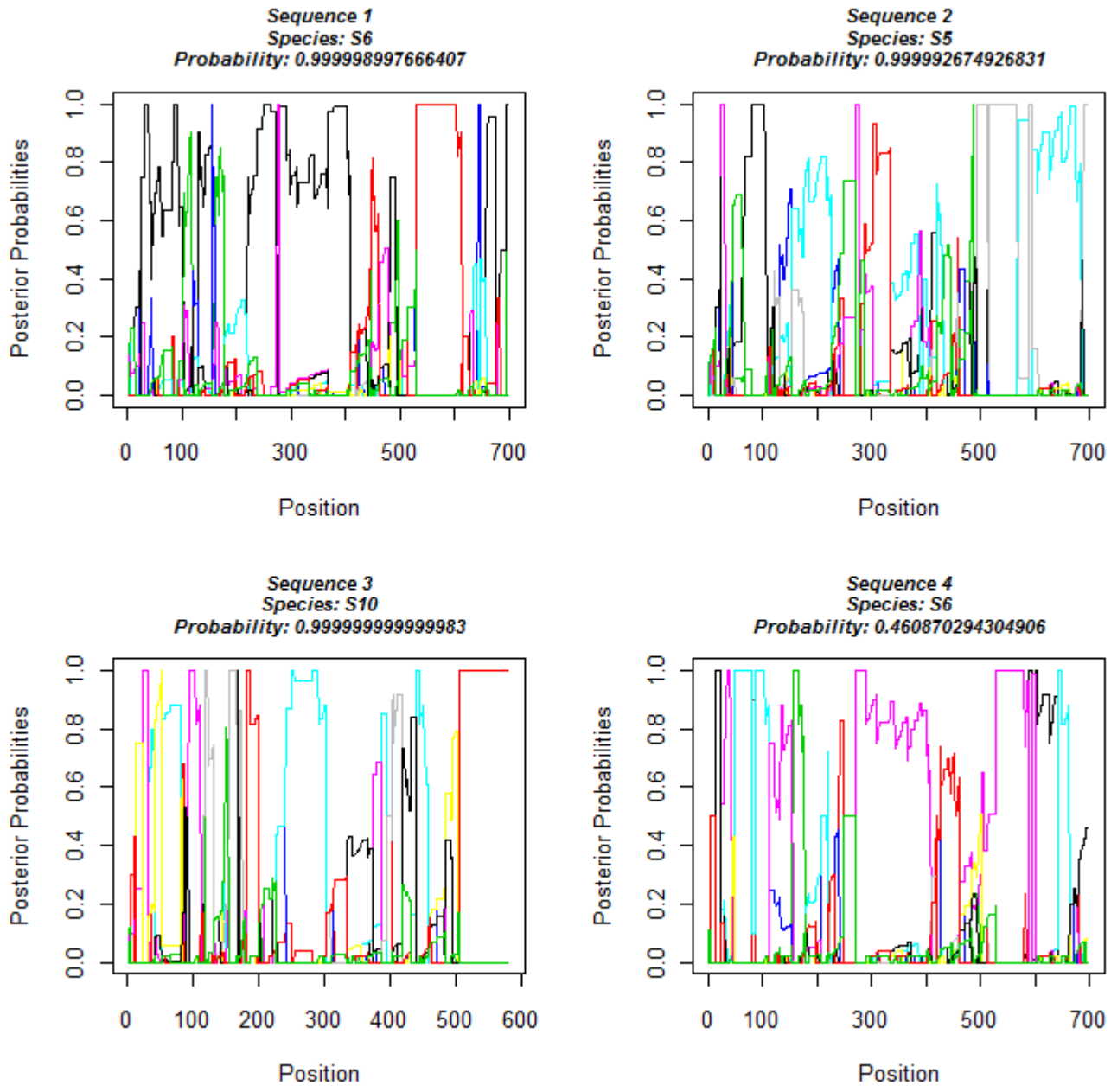

Figure C.31: *Plotted posterior probabilities having removed species 8 from the reference data set and seeking a classification of the four barcodes belong to species 8. Equal prior proababilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 6% *were used.*

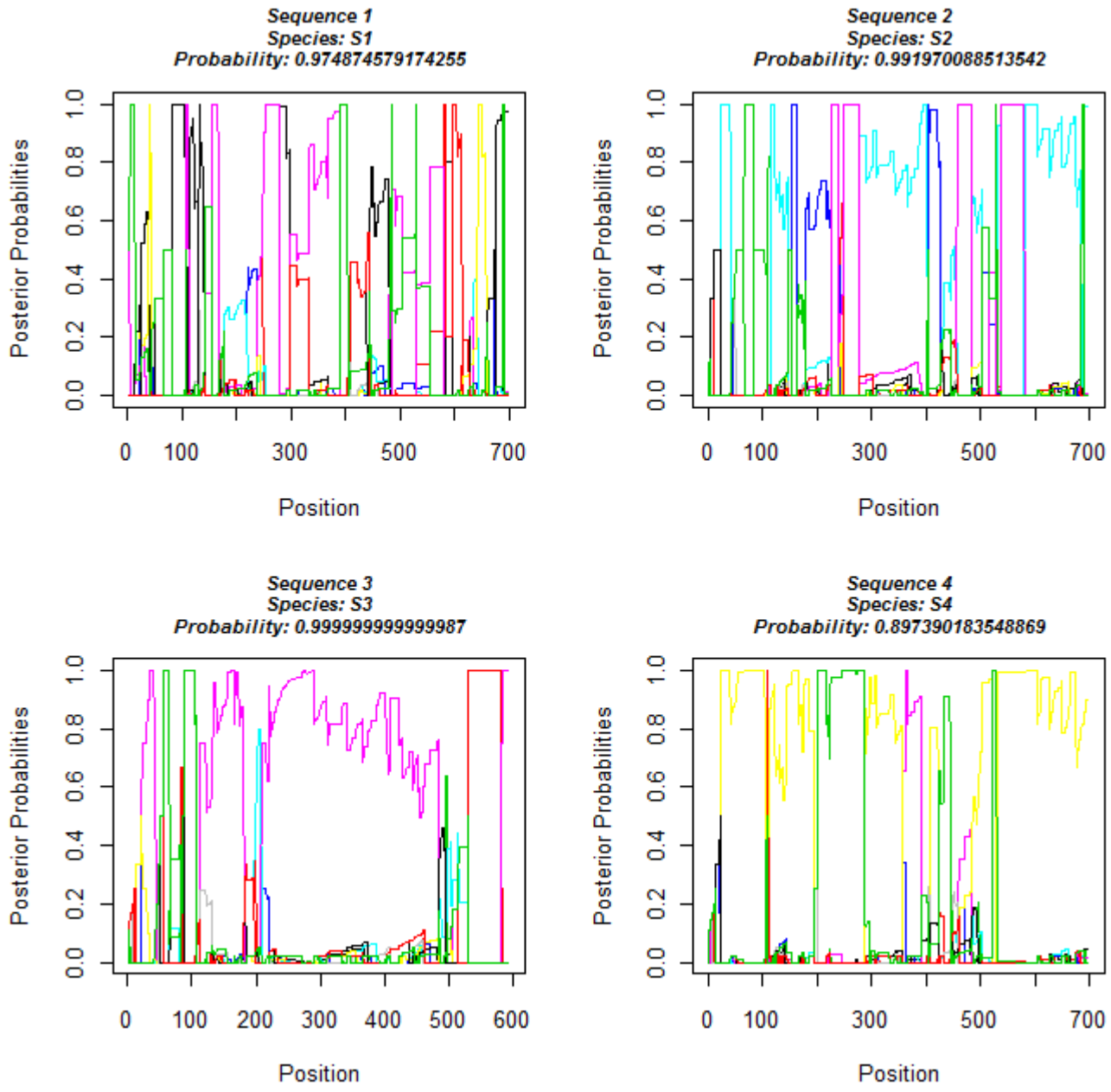

Figure C.32: *Plotted posterior probabilities having removed species 8 from the reference data set and seeking a classification of the four barcodes belong to species 8. Equal prior proabilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 8% *were used.*

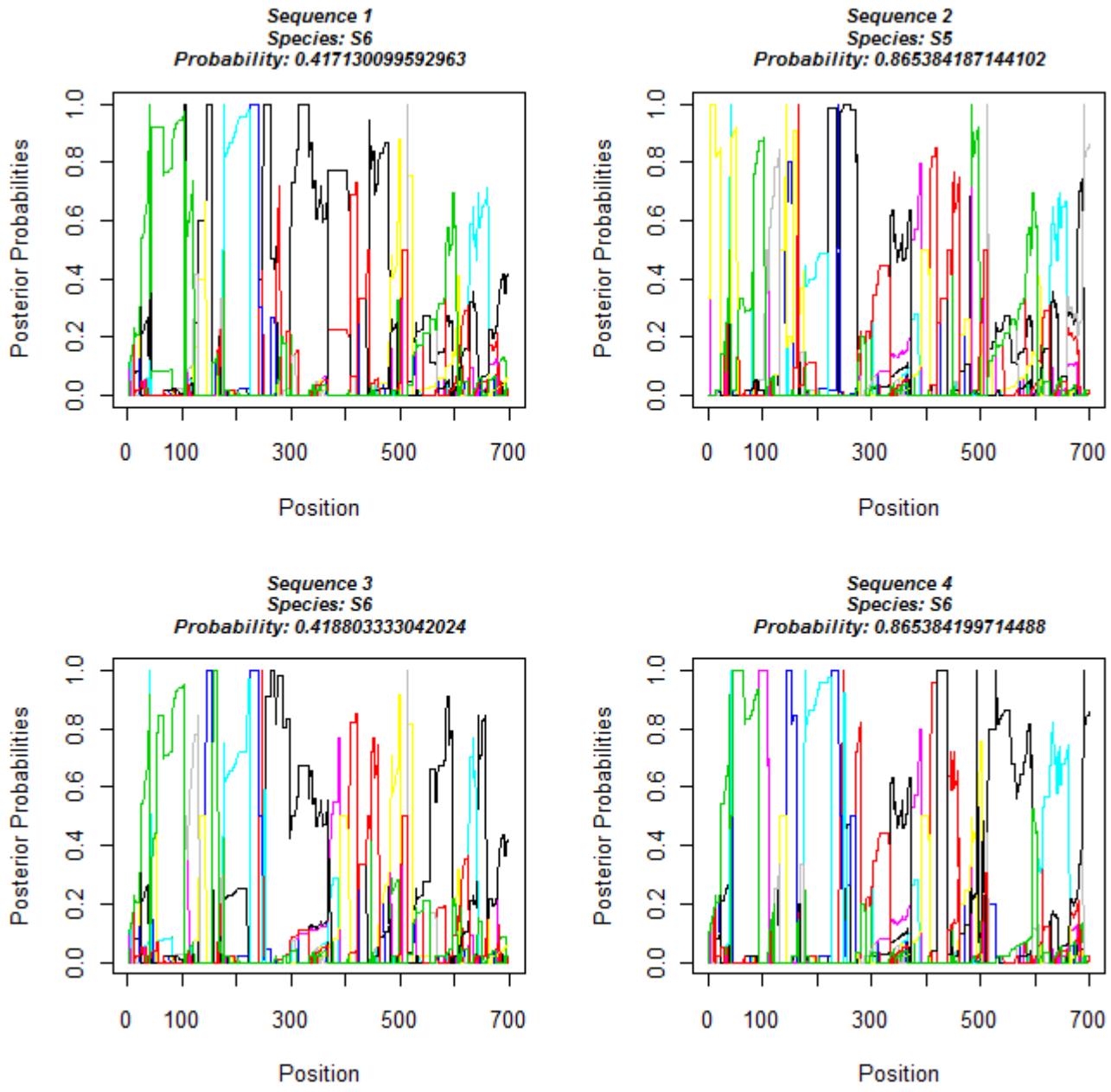

Figure C.33: *Plotted posterior probabilities having removed species 9 from the reference data set and seeking a classification of the four barcodes belong to species 9. Equal prior proabilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 2% *were used.*

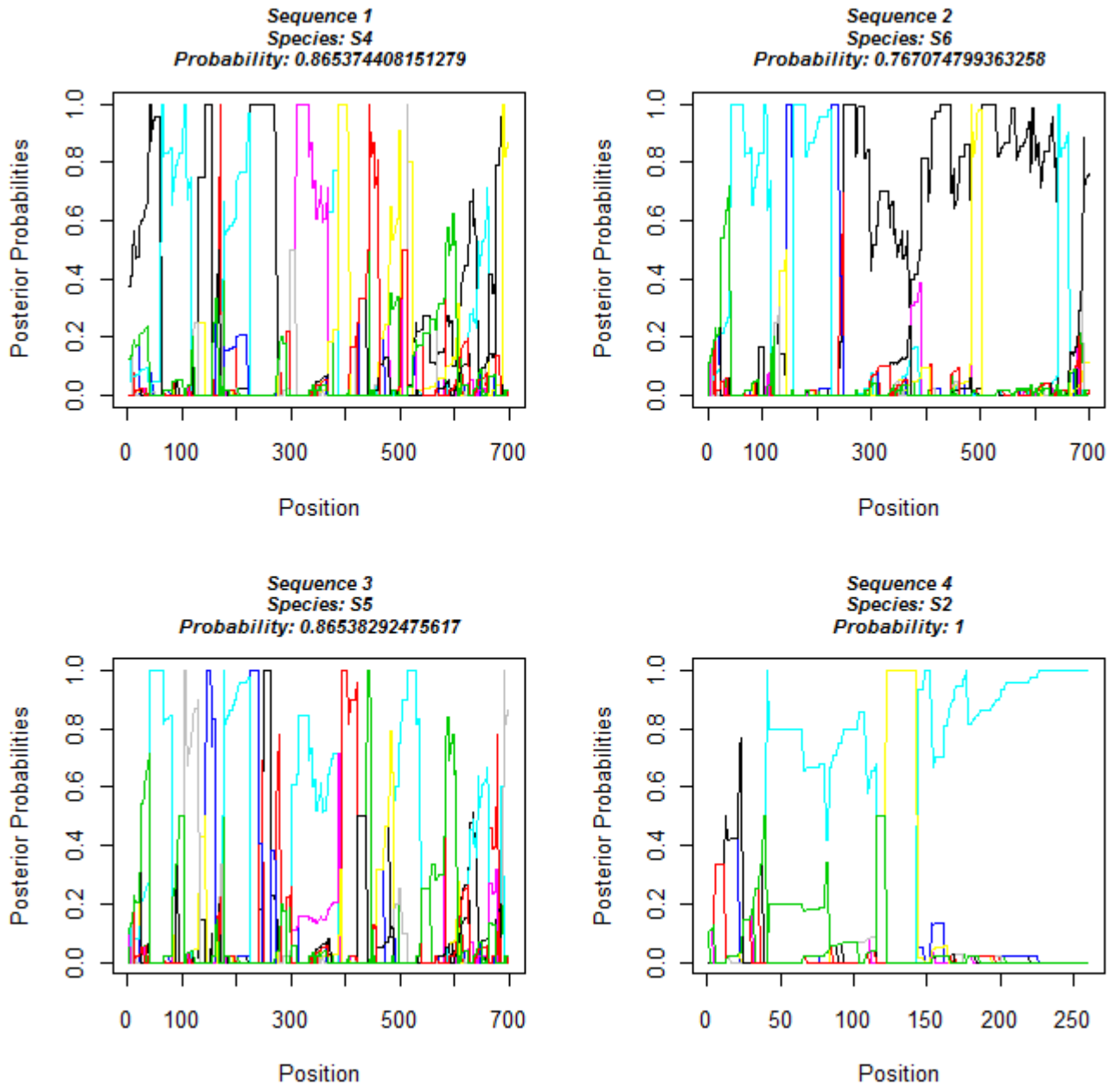

Figure C.34: *Plotted posterior probabilities having removed species 9 from the reference data set and seeking a classification of the four barcodes belong to species 9. Equal prior proabilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 4% *were used.*

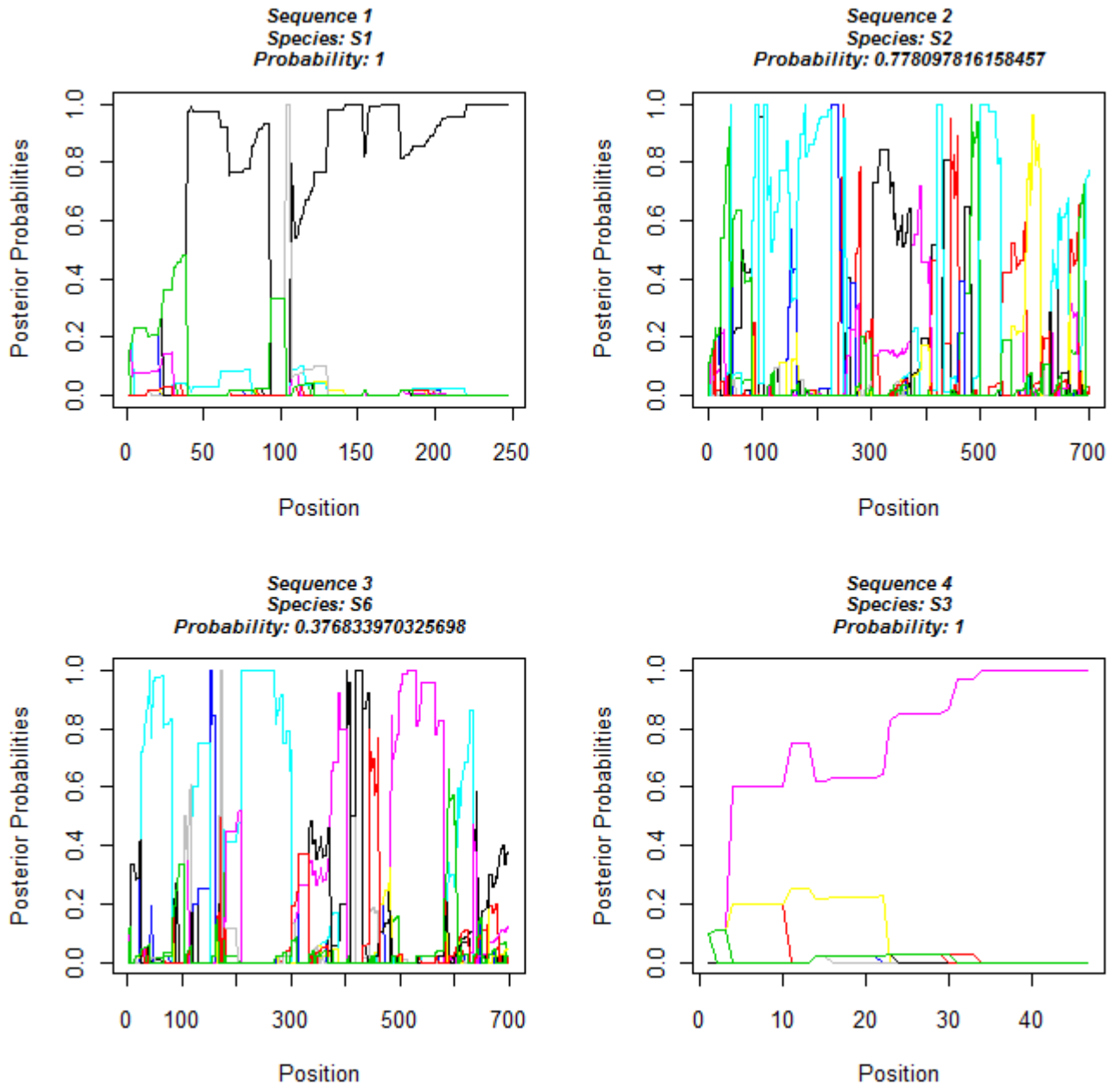

Figure C.35: *Plotted posterior probabilities having removed species 9 from the reference data set and seeking a classification of the four barcodes belong to species 9. Equal prior proabilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 6% *were used.*

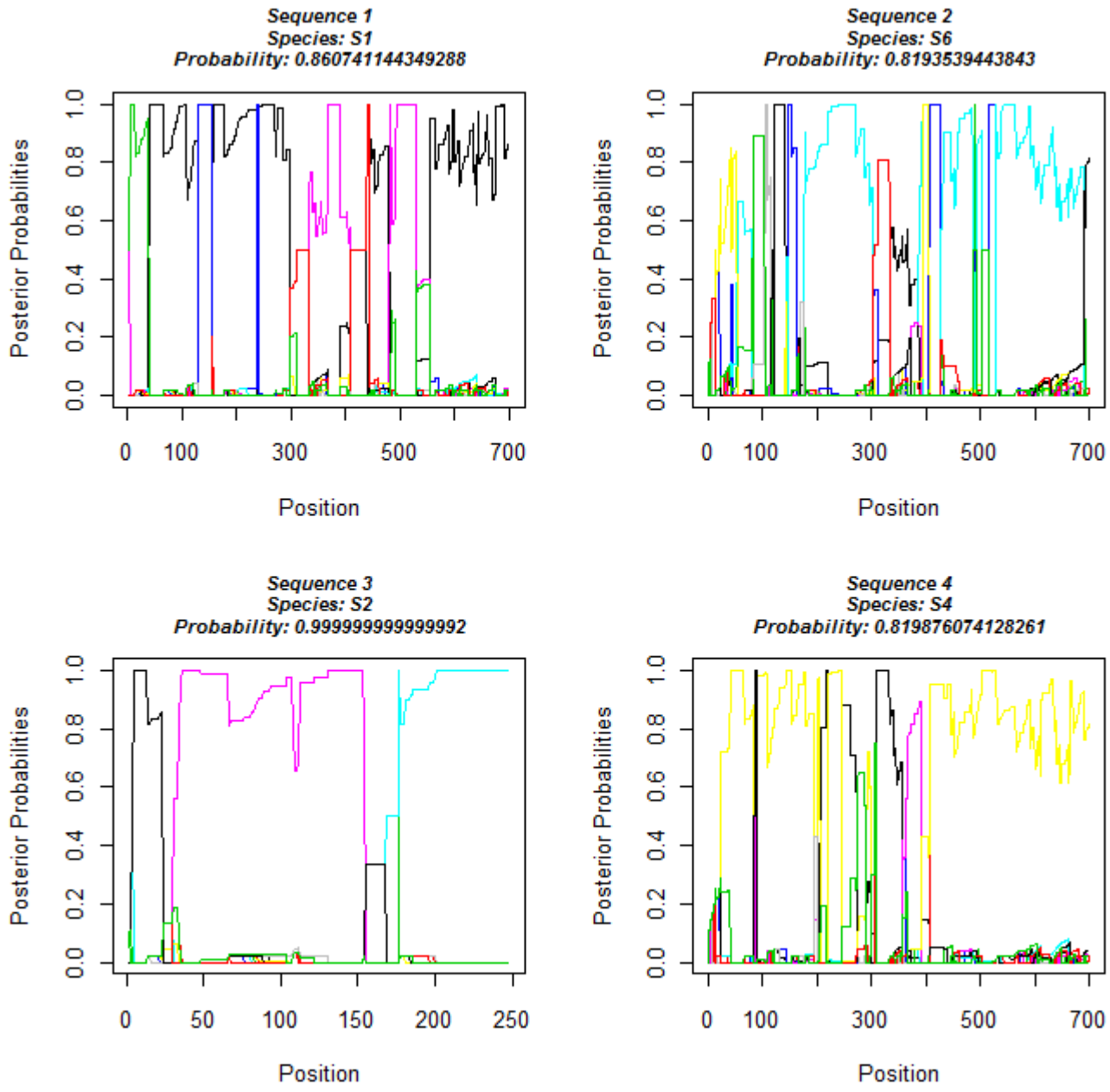

Figure C.36: *Plotted posterior probabilities having removed species 9 from the reference data set and seeking a classification of the four barcodes belong to species 9. Equal prior proabilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 8% *were used.*

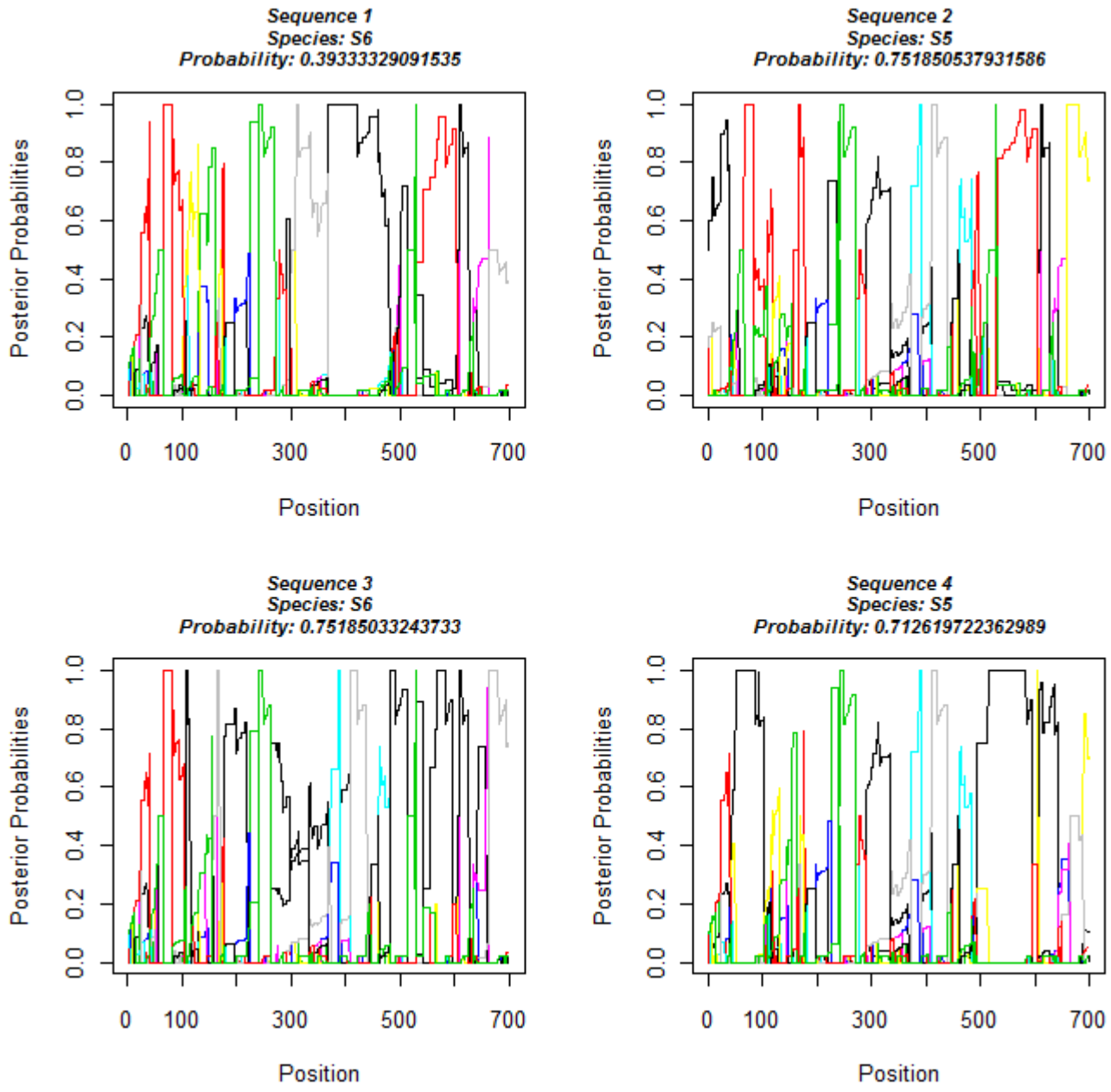

Figure C.37: *Plotted posterior probabilities having removed species 10 from the reference data set and seeking a classification of the four barcodes belong to species 10. Equal prior proabilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 2% *were used.*

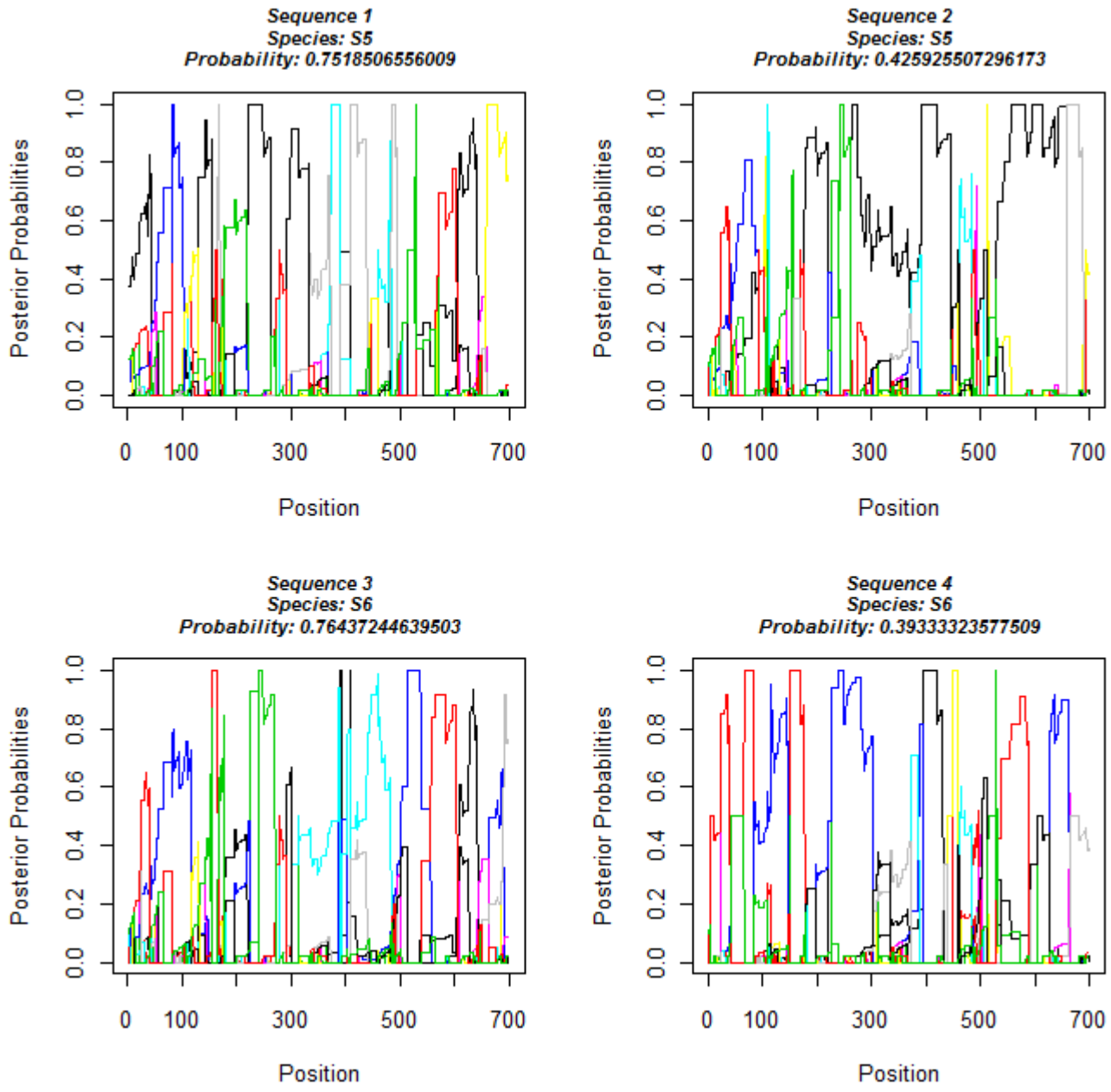

Figure C.38: *Plotted posterior probabilities having removed species 10 from the reference data set and seeking a classification of the four barcodes belong to species 10. Equal prior proabilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 4% *were used.*

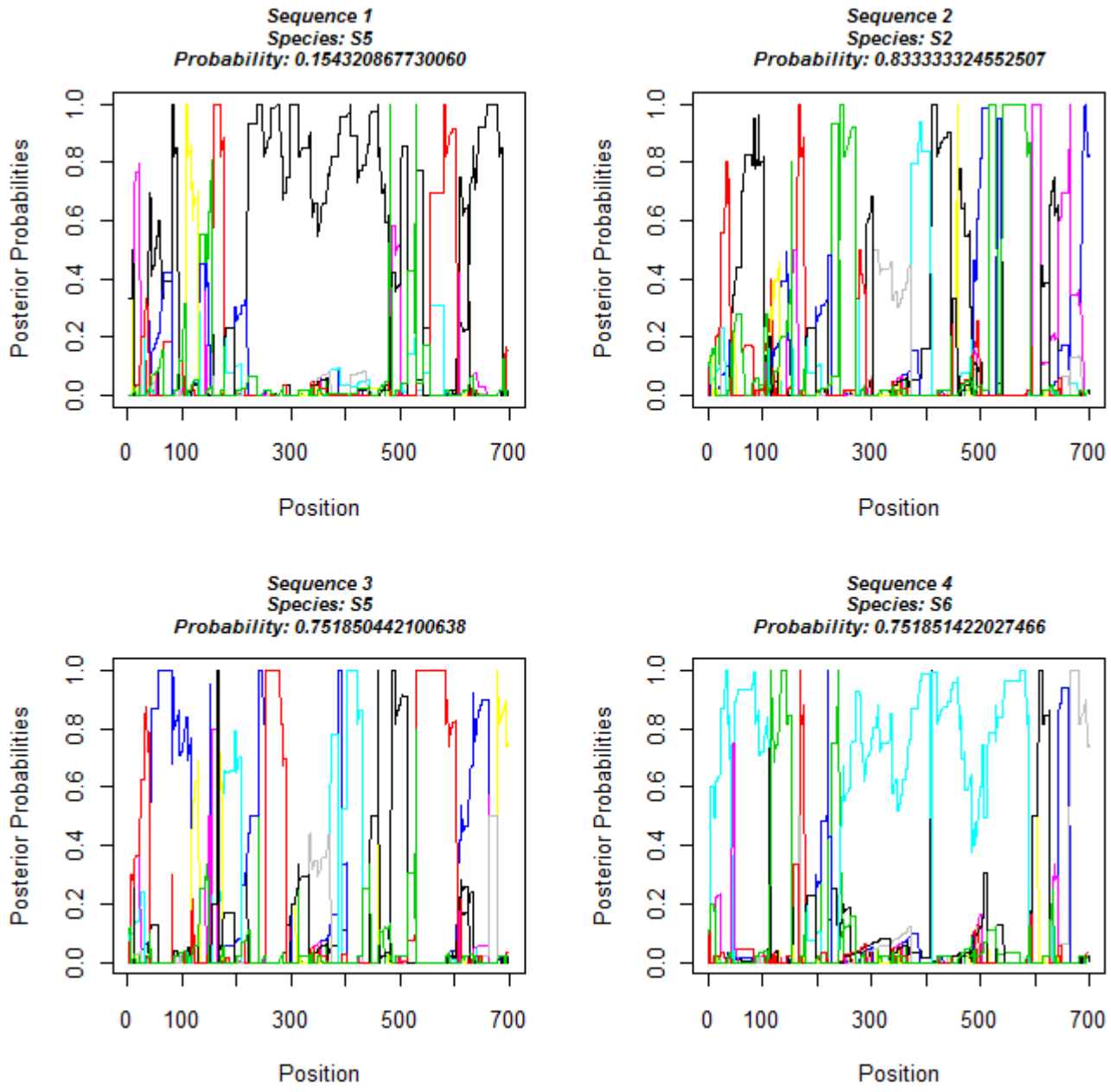

Figure C.39: *Plotted posterior probabilities having removed species 10 from the reference data set and seeking a classification of the four barcodes belong to species 10. Equal prior proabilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 6% *were used.*

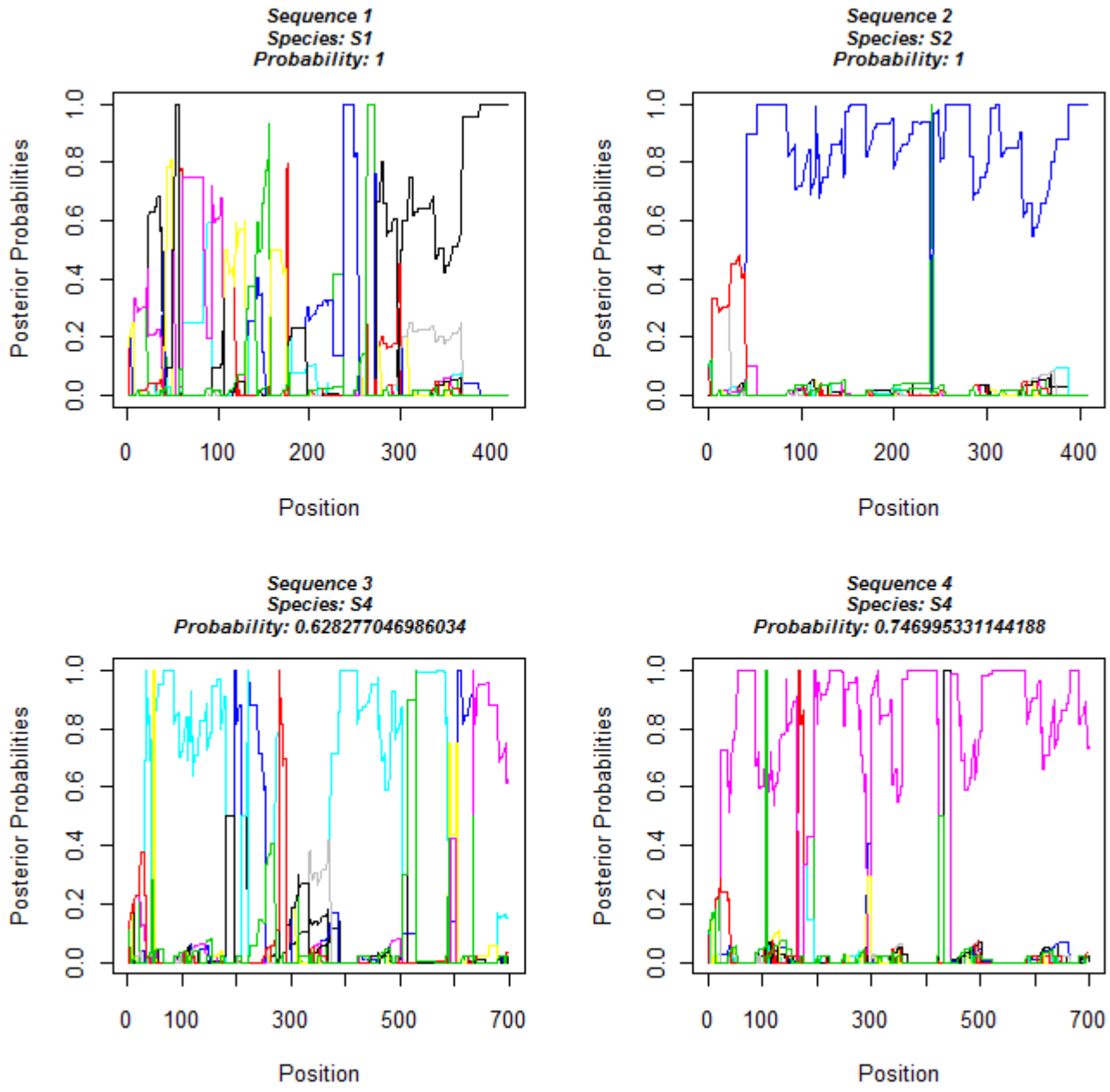

Figure C.40: *Plotted posterior probabilities having removed species 10 from the reference data set and seeking a classification of the four barcodes belong to species 10. Equal prior proabilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 8% *were used.*

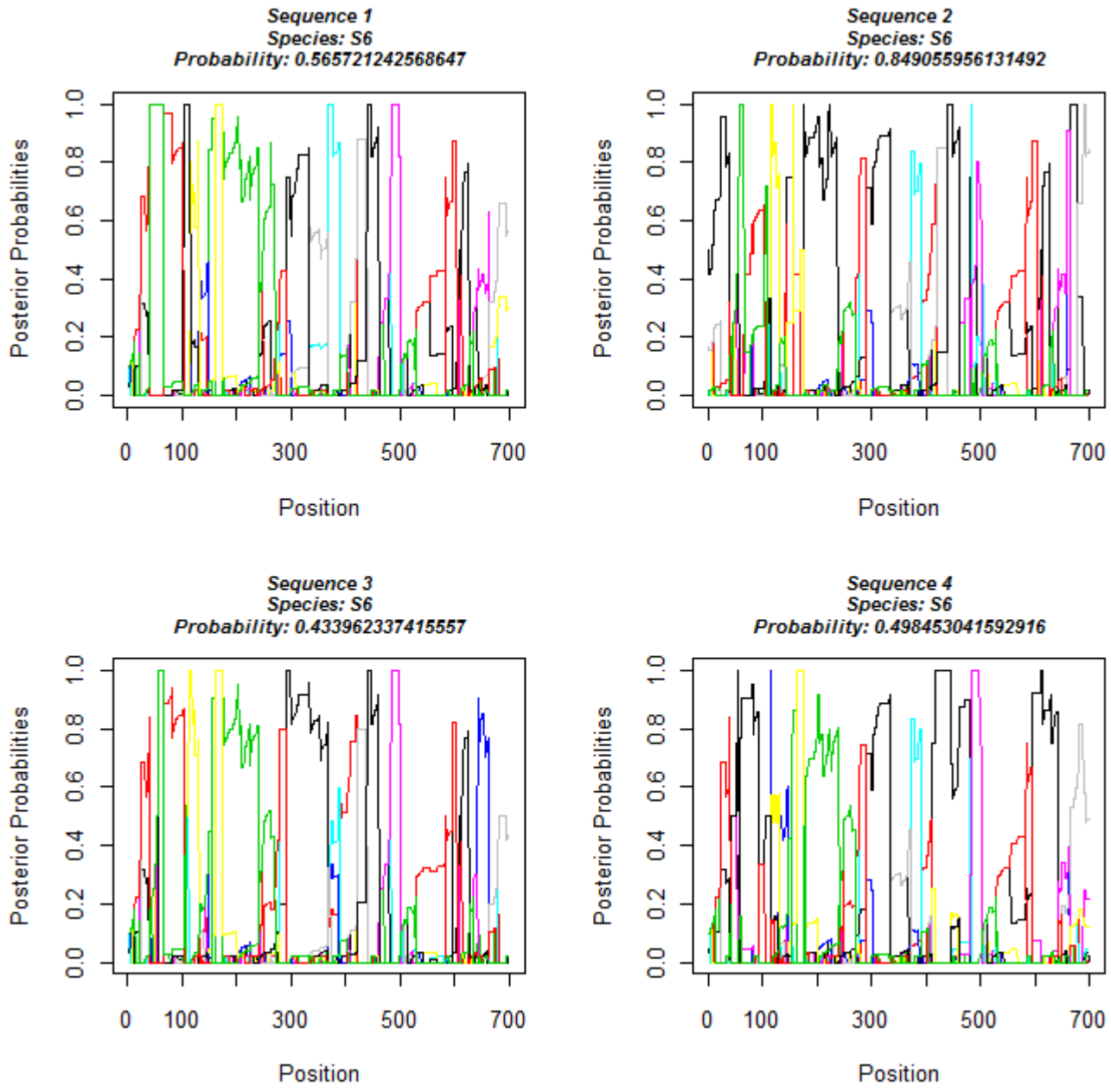

Figure C.41: *Plotted posterior probabilities having removed species 11 from the reference data set and seeking a classification of the four barcodes belong to species 11. Equal prior proababilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 2% *were used.*

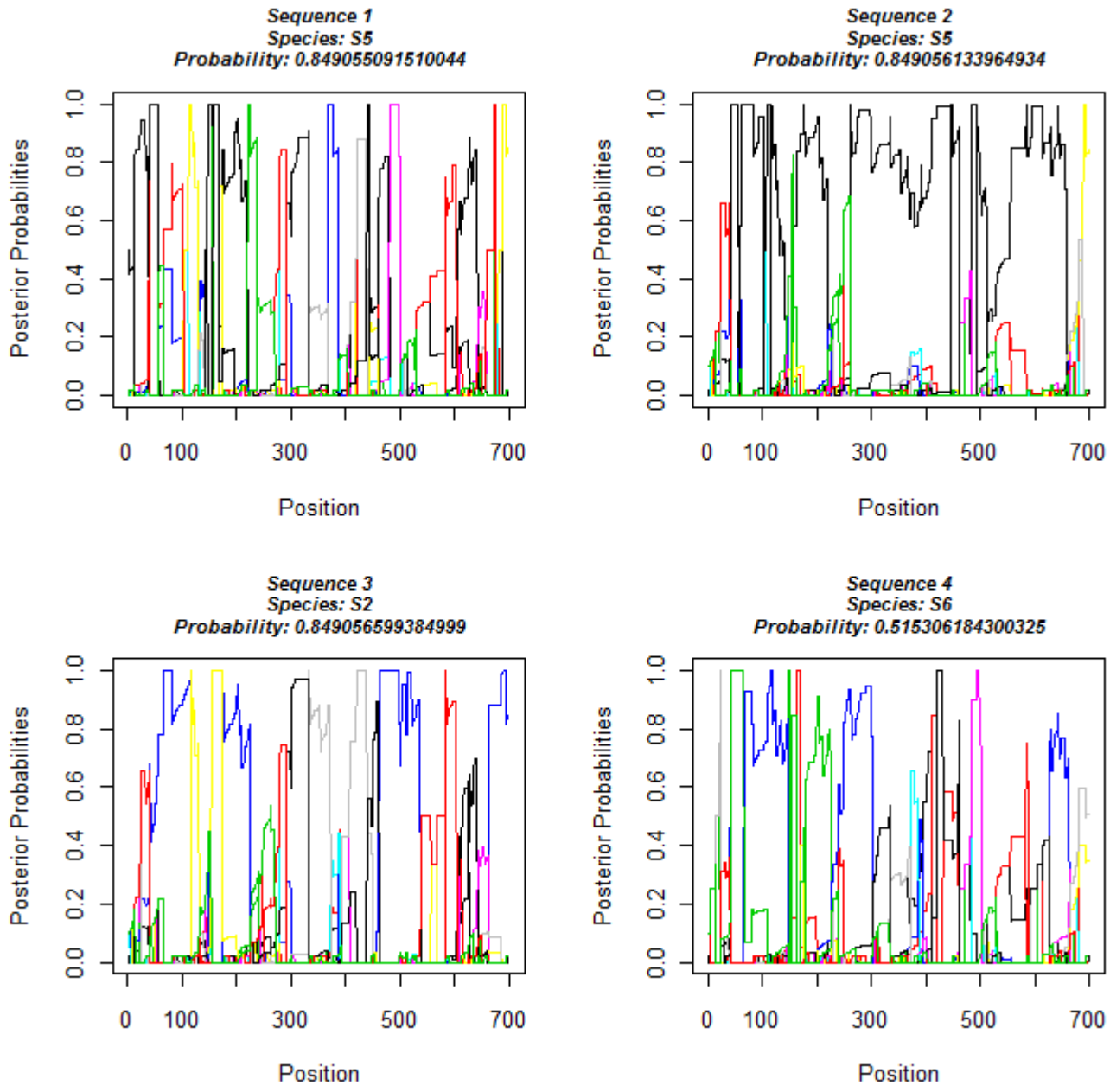

Figure C.42: *Plotted posterior probabilities having removed species 11 from the reference data set and seeking a classification of the four barcodes belong to species 11. Equal prior proabilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 4% *were used.*

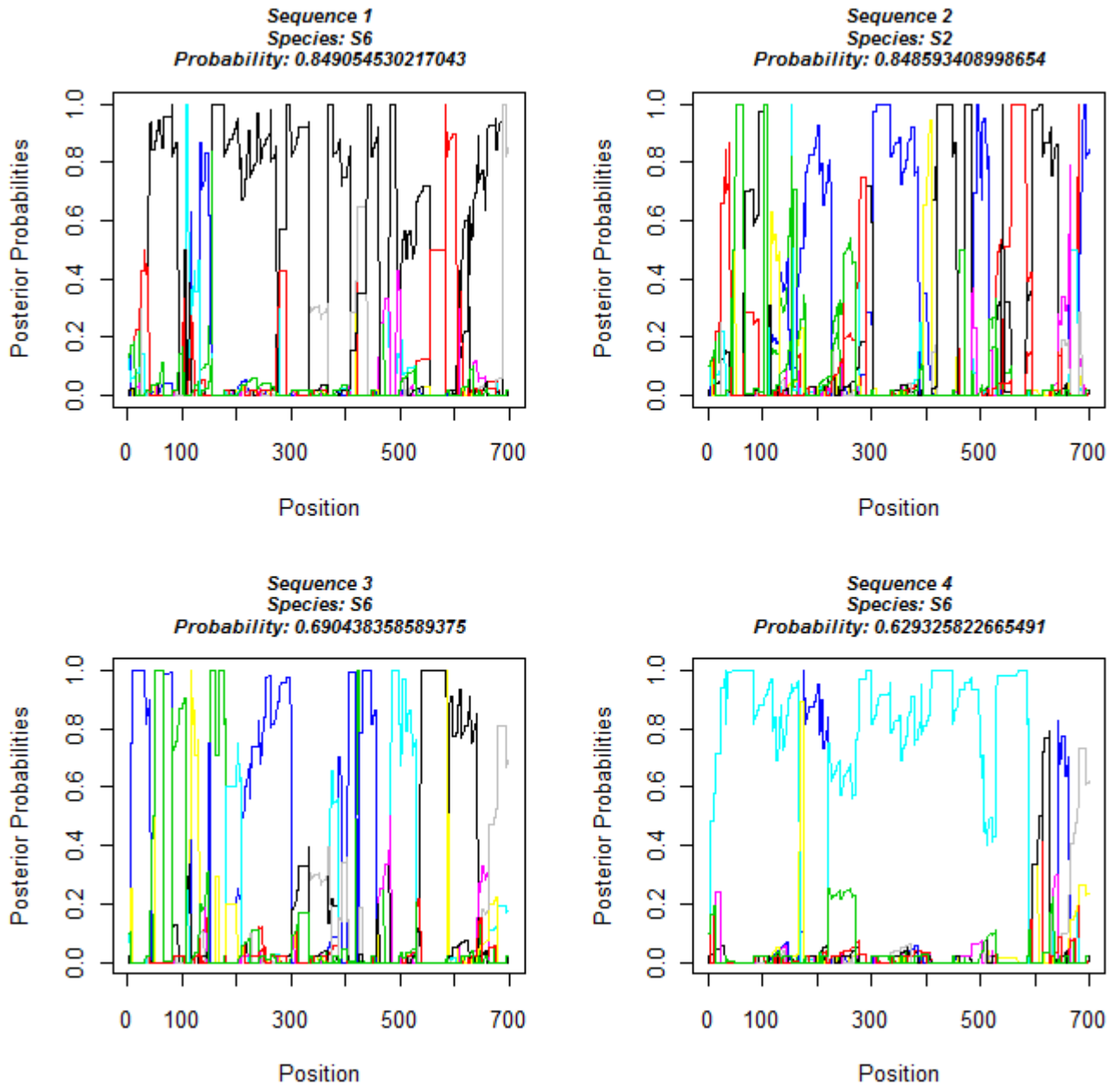

Figure C.43: *Plotted posterior probabilities having removed species 11 from the reference data set and seeking a classification of the four barcodes belong to species 11. Equal prior proabilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 6% *were used.*

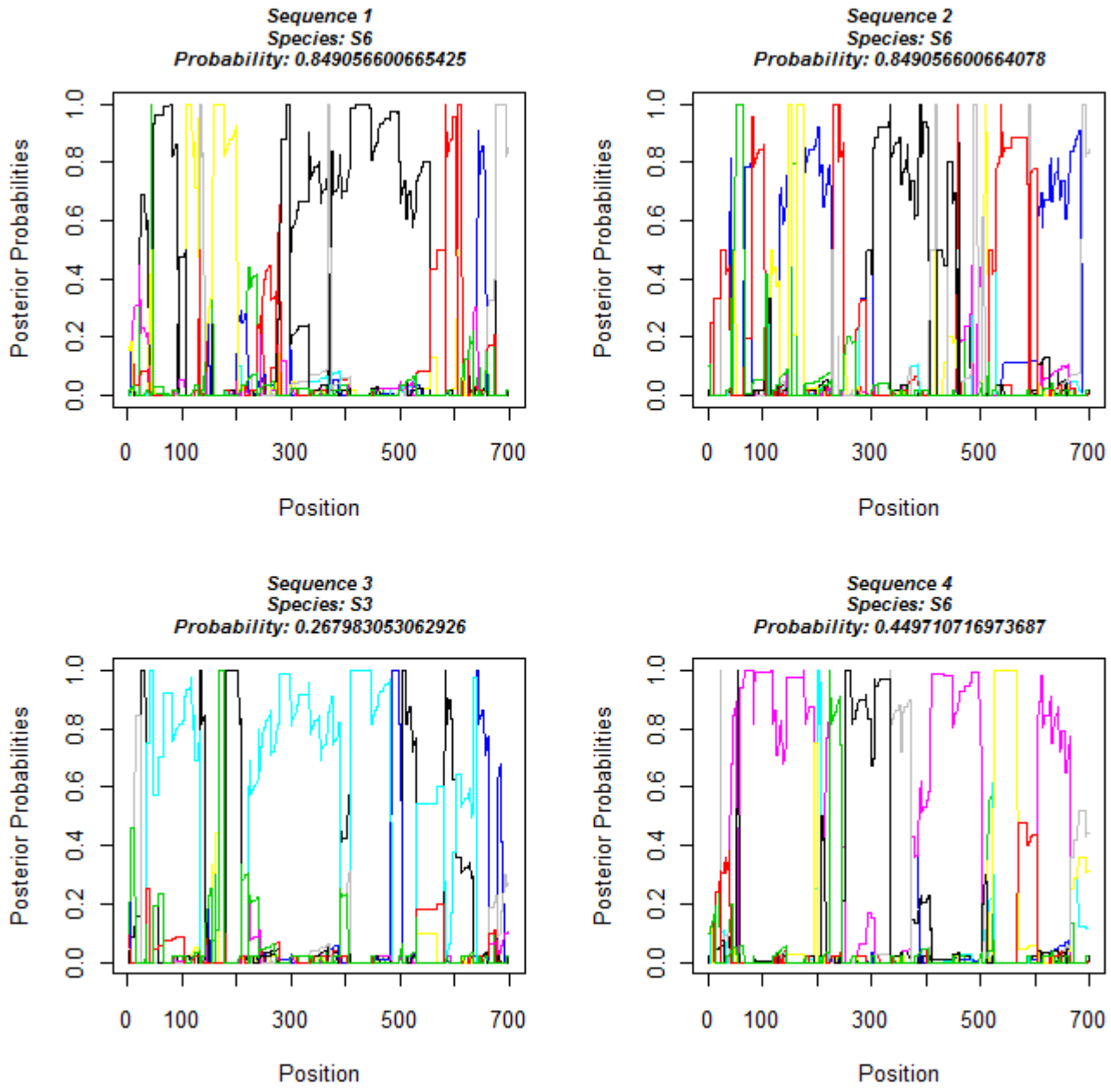

Figure C.44: *Plotted posterior probabilities having removed species 11 from the reference data set and seeking a classification of the four barcodes belong to species 11. Equal prior proabilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 8% *were used.*

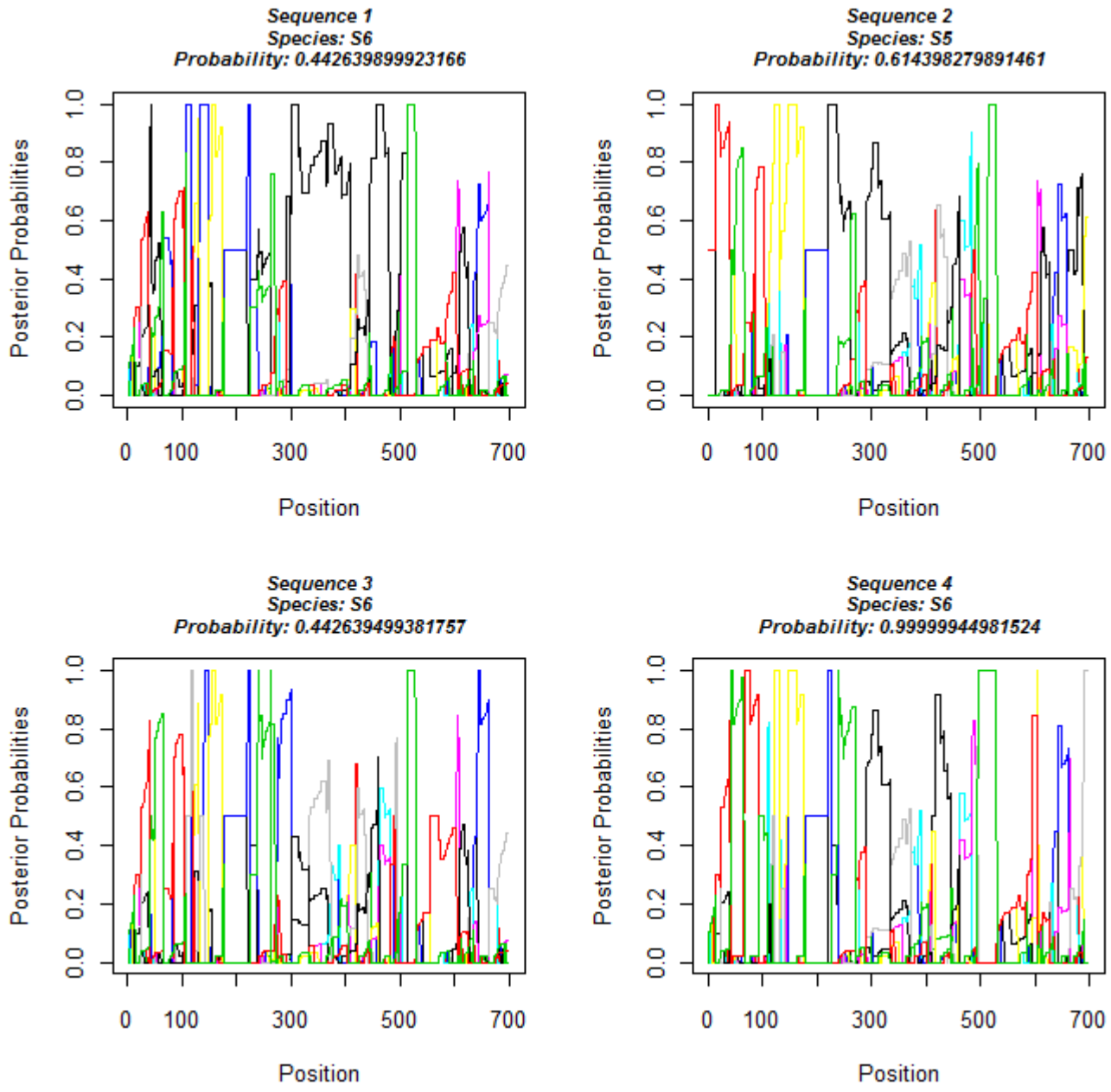

Figure C.45: *Plotted posterior probabilities having removed species 12 from the reference data set and seeking a classification of the four barcodes belong to species 12. Equal prior proabilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 2% *were used.*

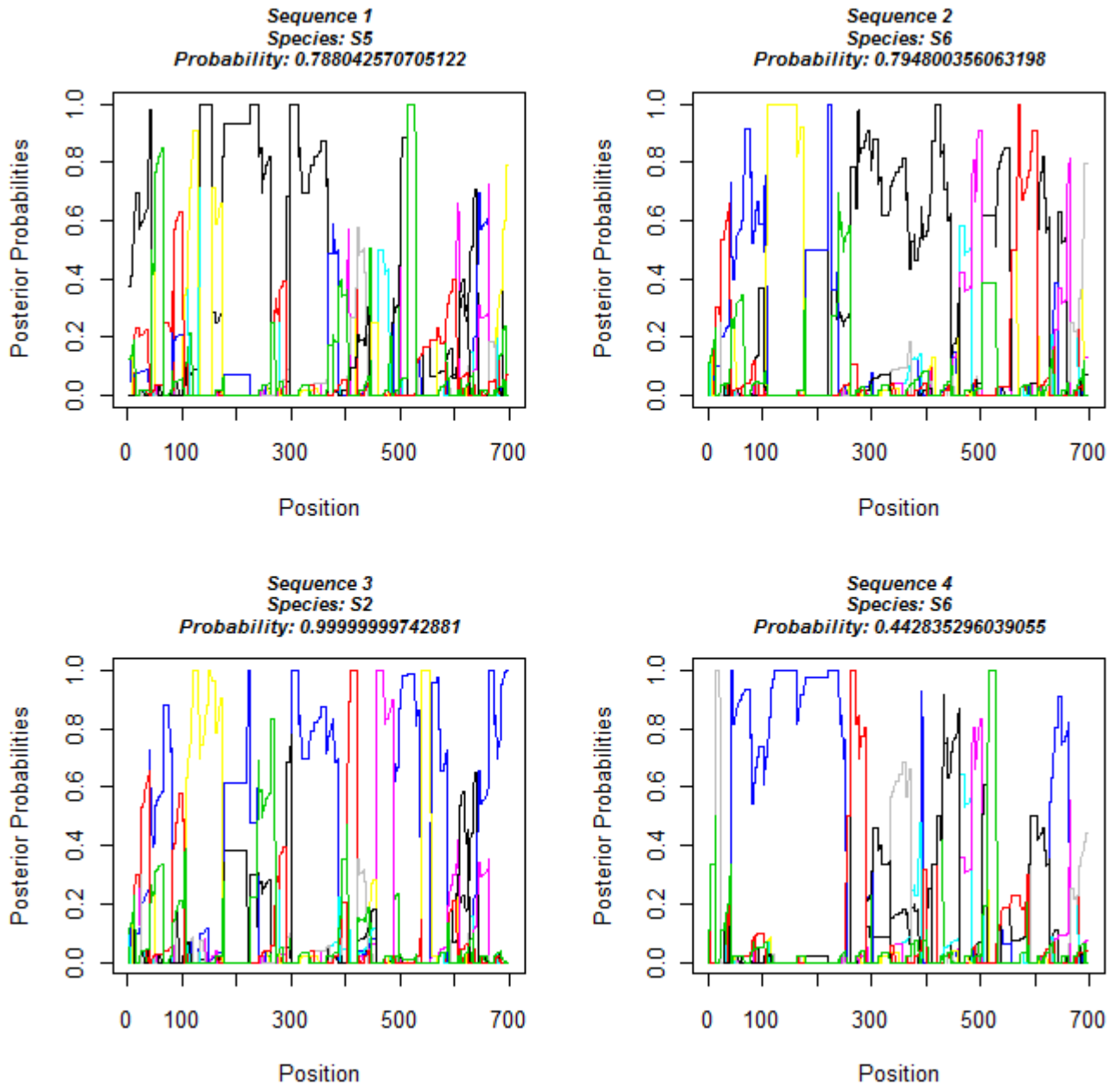

Figure C.46: *Plotted posterior probabilities having removed species 12 from the reference data set and seeking a classification of the four barcodes belong to species 12. Equal prior proabilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 4% *were used.*

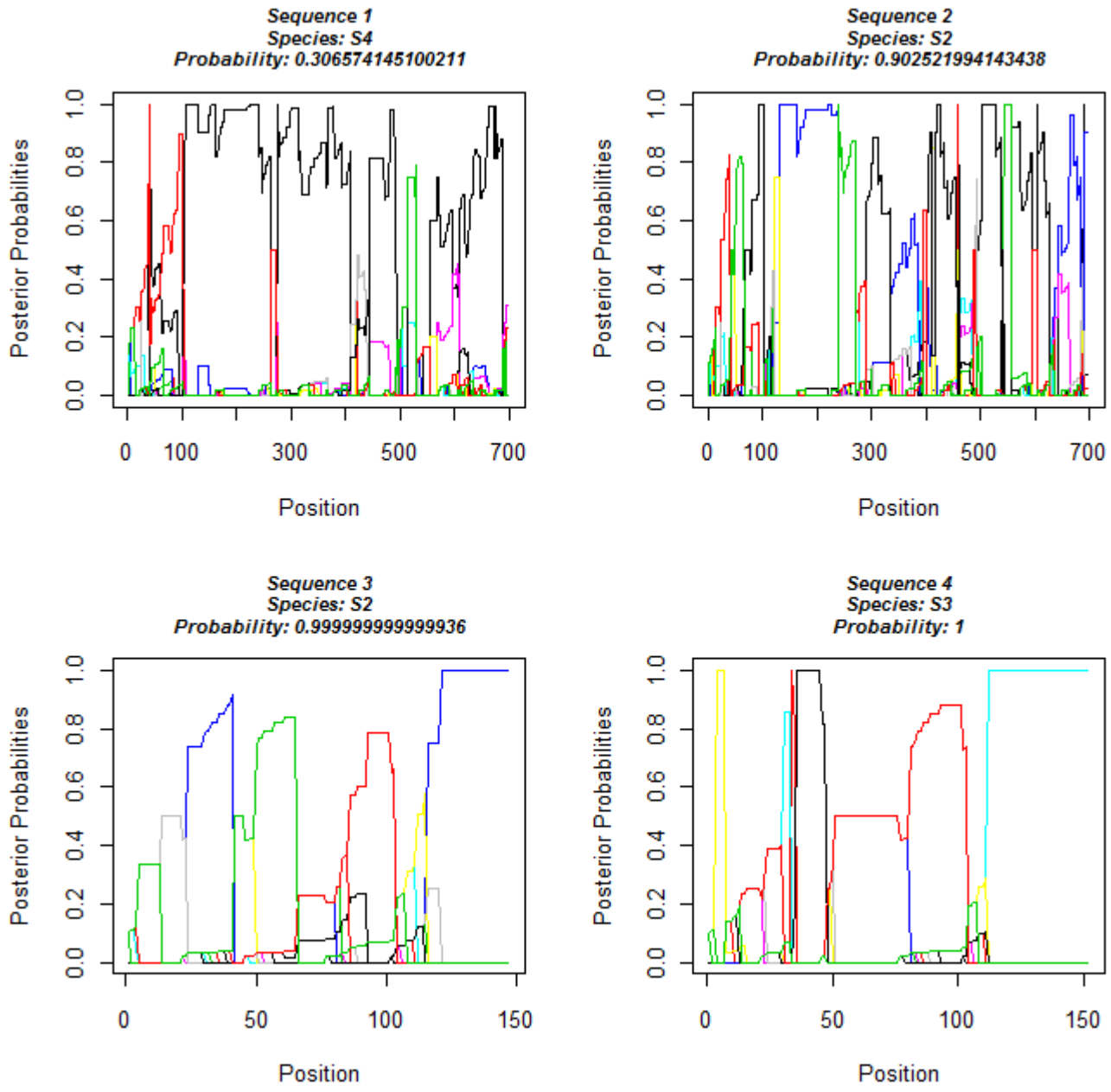

Figure C.47: *Plotted posterior probabilities having removed species 12 from the reference data set and seeking a classification of the four barcodes belong to species 12. Equal prior proabilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 6% *were used.*

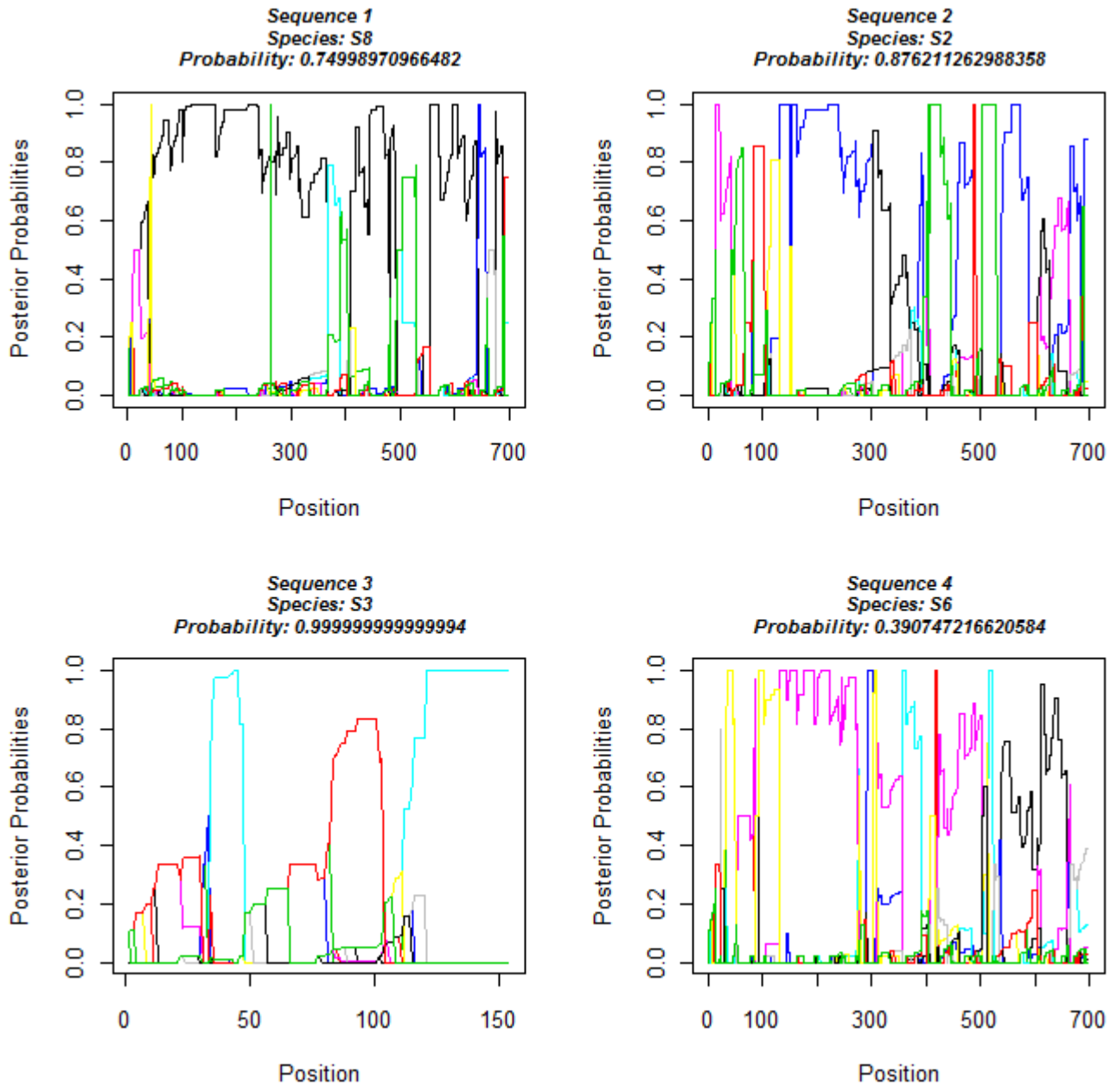

Figure C.48: *Plotted posterior probabilities having removed species 12 from the reference data set and seeking a classification of the four barcodes belong to species 12. Equal prior proababilities with* $\delta = 9.7 \times 10^{-8}$ *and within-species variability of* 8% *were used.*