Phylogenetic lineage of humans and primates

Julia Bohlmeyer, Ben Harp, Blair Heptig, Kyle Kinkade,
Christian Rayson, Joshua Strohm, Christine Wacker,
Robert Fraga, Denis Popel

Computer Science Department,
Baker University
Baldwin City, KS
denis.popel@bakeru.edu

Abstract

This work describes our understanding of the evolutionary lineage of species or genes, one of the major problems in bioinformatics, through (i) studying the genome project; (ii) analyzing genetic information; (iii) programming classification methods and alignment techniques; and (iv) experimenting with existing programs. Using several pre-existing programs and a pairwise global alignment program of our own, WSRalign, we compared 101 mitochondrial DNA sequences to create a set of phylogenetic trees.

1 Introduction

The human genome is composed of approximately one billion nucleotides [7; 8]. Thus, comparing entire genomes is a computationally expensive problem. By "comparing," we mean a set of actions which usually includes pairwise or multiple alignment, distance measuring, and phylogenetic tree design. There are around 16,000 nucleotides in a strand of *mitochondrial DNA* (mtDNA) which makes comparison more manageable than a full genome comparison. The smaller size of mtDNA allows it to evolve 5 to 10 times faster than a normal strand of nuclear DNA [9]. Most of the genetic history of humans can be traced through mtDNA which is only inherited maternally. Thus, the process of genetic recombination (the exchange of fragments of DNA between maternal and paternal sources) does not occur. This lack of genetic recombination leaves mutation as the only source of variation for a strand of mtDNA. All human mtDNA is very closely related with a relatively small amount of variation among the samples of mtDNA [3].

In this study, 101 mtDNA sequences of 90 different human ethnicities and 11 primates are compared. Using several pre-existing programs and a pairwise global alignment program of our own, we compared these sequences to create a set of phylogenetic trees [5]. The pre-existing programs that we used had options for both slow-and-accurate as well as fast-and-less-accurate alignments [1]. Our goal is to design a program for computing pairwise distances between each set of the sequences, thus enabling us to form a phylogenetic tree for humans and primates.

The paper is structured as follows: Section 2 gives the outline of programs and methods used in this study; the analysis of experimental results is outlined in Section 3; and Section 4 concludes the paper.

2 Programs and Methods

2.1 Input Data

All of the mtDNA sequences included in our study are obtained from the *Bank of Genetic Information* (GenBank) [2] and named by combining their origin and an identification number. All of these sequences came from individuals, and any anomalous position in the phylogenetic tree is attributed in this study to population migration [6].

2.2 Programs Used

In our study, we use a set of programs. The approach is depicted in Figure 1. We start with the raw sequences of mtDNA (ethnic taxa) and then create pairwise alignment for sets of taxa. A distance matrix is then created (the NEXUS format is chosen), in which the (i, j) entry is the distance between the genomes of the i-th and the j-th taxa. Using this matrix as input, a phylogenetic tree design program called PAUP [4] is used to create a graphical representation of the matrix. A program TreeView enables us to view the tree graphically and to save it as a graphical file (the vector format is selected). To visualize the alignment and to demonstrate the similarities between two sequences, we used the program Dotter.

2.2.1 ClustalX

Our interest in creating a phylogenetic tree of the 101 species requires an alignment of those species to compare the individual genetic codes of the mtDNA. Using an existing

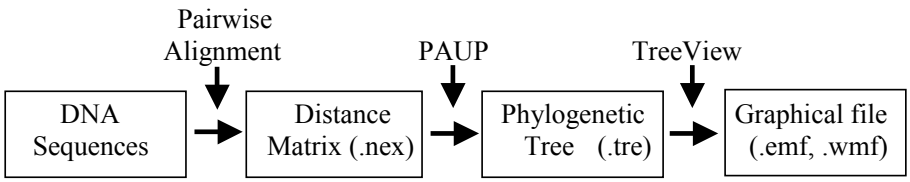


Figure 1: Steps of our experimental approach

program, ClustalX [1], we were able to align the nucleotides of each species to find the greatest similarities/differences relative to the other species. There are several possible matrix options for ClustalX, and each option yields different results. We used IUB DNA weight matrix in the alignment of the 101 nucleotide strands.

ClustalX aligns nucleotide sequences by two methods: a slow and accurate method, and a faster and less accurate method. The slow and accurate method is executed by comparing individual nucleotides to one another whereas the faster method compares series of multiple nucleotides. The faster method is considerably less accurate, but it saves computer run-time. Finally ClustalX yields results in various formats. These different formats can be used in additional programs and purposes. The NEXUS format allowed the aligned sequences to be formatted so that they could be input for the PAUP program. This allowed us to make more inferences, including the creation of a phylogenetic tree.

2.2.2 PAUP

After a pairwise or multiple alignment has been completed, it is possible to create a phylogenetic tree by putting the information into a program called PAUP [4]. The input of this program is either a distance matrix or pre-aligned sequences. To run the program, a file with the extension .nex is executed inside PAUP. Through the sequence of commands like UPGMA and SaveTrees, a phylogenetic tree is created and saved with the extension .tre. This file can then be input to the program Treeview in order to view a phylogenetic tree.

2.2.3 WSRalign

There were several difficulties to overcome in a pairwise alignment, one of which was how to accomplish

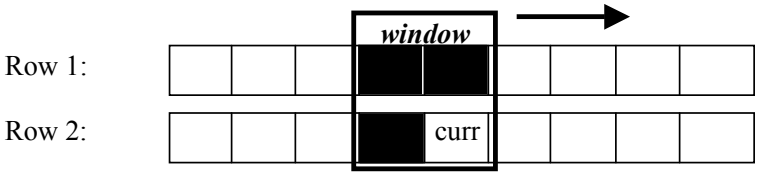
$$C(n, 2) = \frac{n(n-1)}{2} \tag{1}$$

repetitions of the comparison process required for forming the phylogenetic tree, where n is the number of mtDNA sequences being compared. With $n = 101$ test species, there are 5,050 mtDNA sequence pairs to be compared. The process of entering the file names by hand and then recording the results by hand would be prohibitively time consuming. Thus, we determined that in order to complete this process within a reasonable amount of time, we would need to automate the process. For this reason, we designed our own program to handle the pairwise alignment process.

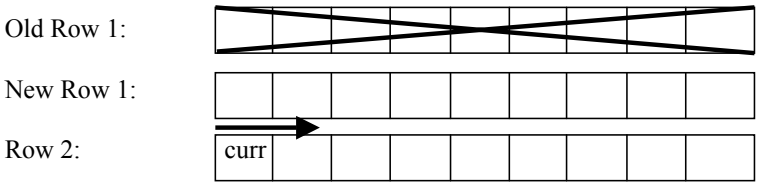
Our pairwise-alignment program, WSRalign, was created using C++, within the framework of the Microsoft Visual Studio Professional .NET programming environment. The computers used were equipped with a single 2.0 GHz Pentium IV processor with 256 Mb of memory, running the Microsoft Windows XP Professional operating system.

The distance-scoring section of the code presented a number of difficulties. Our method of calculating distances involves a global pairwise alignment. This creates a matrix that finds the optimal alignment of two mtDNA sequences. By allotting bonuses and penalties for matches, mismatches, and insertion gaps, it assigns a distance between the two sequences. However, since each of these sequences is approximately 16,000 base pairs long, comparing two of them according to the traditional approach requires a $16,000 \times 16,000$ matrix. This results in 256 Mb of space allocation. An early version of WSRalign created and attempted to manipulate the entire distance matrix. Upon execution of this code, however, we encountered a system error due to lack of system resources. In order to streamline our code, we determined that at any given point in the alignment, we needed only the row in the matrix currently being filled and the one directly above it. Thus, we could conserve system resources by allocating only two rows of memory space instead of storing the complete matrix. The following explains how the two rows are manipulated.

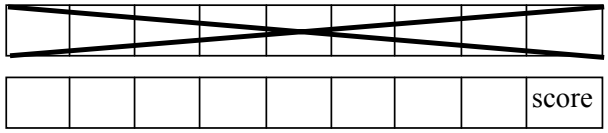
Step 1: Process each cell in Row 2 using the previous cell in Row 2, the corresponding cell in Row 1, and the cell previous to the corresponding cell in Row 1 (black cells).



Step 2: When the end of Row 2 is reached, delete the contents of Row 1 and place the contents of Row 2 in Row 1. This step is repeated according to the number of rows.



Step 3: When the last row has been analyzed, the final cell contains the distance score for the comparison.



Step 4: `For` loops are constructed to allow the program to repeat Steps 1-3 for each pair of mtDNA sets.

Our program then generates an output file "data.nex" in NEXUS file format. The contents of data.nex includes the calculated distance matrix as well as appropriate formatting to generate a NEXUS file that can be used by the phylogenetic tree-design program, PAUP. PAUP uses the distance data to generate a phylogenetic tree. A separate program called TreeView allows the tree to be viewed and saved as a graphic file (.emf or .wmf).

Using this method, we were able to calculate distances between all 101 test sequences in approximately ten hours. The program requires nearly 100 percent of the processor resources while running but requires only between 0.5 and 1.0 Mb of system memory. The overall computational complexity of WSRalign is in $O(n^2)$.

3 Experimental Results

3.1 ClustalX and PAUP

When the fast alignment (Figure 2) and slow alignment (Figure 3) options of the ClustalX program are compared, we find some differences. In the case of the accurate alignment using the ClustalX program, we see that the 101 mtDNA sequences can be put into one of three major groupings. The largest grouping is called SLOWA, the next largest SLOWB, and the smallest consisting of only four sequences, SLOWC. The fast alignment in ClustalX also has three groupings: a large group FAST1, a smaller group FAST2, and a group consisting of only one sequence, FAST3. By looking at the differences in the two phylogenetic trees, we conclude that the fast alignment is reasonably accurate. For example, all of the sequences in FAST2, except for JordanAF381999 and AustriaAF346964, are found in either SLOWB or SLOWC. All four sequences in SLOWC are found in FAST2. There is, however, one significant difference between the fast alignment and the slow alignment. All but two sequences in FAST2 are in either SLOWB or SLOWC, but there are several sequences scattered throughout the FAST1, FAST2, and FAST3 taxa which are also in SLOWB. In addition to this, there is no real consistency between the taxa inside FASTB and the taxa inside SLOW2. This means that the sequences in SLOW2 are grouped together differently than they are in FASTB.

Some sequences are seen to fit together. For example, AfricanD38112 and SweedishX93334[1] are closely related in both alignments. However, according to our results, we would prefer to exclude AfricanD38112 and SweedishX93334 in order for our results to be accurate. BaboonY180011 is consistently grouped with WesternTarsierAF348159 and WesternTarsierNC002811. The two Gorilla sequences are grouped with the Westernlowlandgorillax933471 sequence in both of the alignments, as well as the BonoboMonkeyD381161, PygmyChimpanzeeNC001644, ChimpD381131, ChineseAF346972, Chimpx933351, and the ChimpNC001643 sequences. In both alignments, the monkeys are grouped closely together. This can be seen in FASTA and in SLOW1. One anomaly in the two alignments is that ChineseAF346972, SweedishX93334, and AfricanD38112 are grouped closely to the primate sequences. These can be found in FASTA and SLOW1. The most consistent grouping was with the KikuyuAF346992, MauritaniaAF381992, IboAF346987, BiakaAF346968, MbenzeleAF346997, BiakaAF346969, MbenzeleAF346996, IboAF346986, and MauritaniaAF381994 sequences. In both of the alignments, the groupings which link these sequences are virtually identical; the only difference is that in the slow alignment, KikuyuAF346992 and MauritaniaAF381992 are switched due to the distance measures.

Other sequences that are consistently grouped in both of the alignments are as follows:

- CanaryAF382009, WaraoAF347012, WaraoAF347013;
- HausaAF346985, MoroccoAF381988, MbutiAF346998, MbutiAF346999, SanAF347008, SanAF347009;
- MandenkaAF346995, MauritaniaAF381981;
- EffikAF346976, and Effik346977;
- MauritaniaAF381991, JordanAF381998, and YorubaAF347014;
- FilipinoAF382012 and IndiaAF382013;
- BuriatAF346970 and EvenkiAF346979;

[1]The misspelled name appears in the data set, the correct spelling is 'Swedish.'

- JordanAF381996, and MoroccoAF381984;
- JapanAF346990, the JapanAF346989, GuaraniAF346984, and SiblunuitAF347010;
- EwondoAF346980, LisongoAF346984, and YorubaAF347015;
- GeorgianAF346982, LeonAF382006, and MoroccoAF381985;
- PimanAF347001, UzbekAF347011, SamoaAF347007, KhirgizAF346991, and KoreaAF346993.

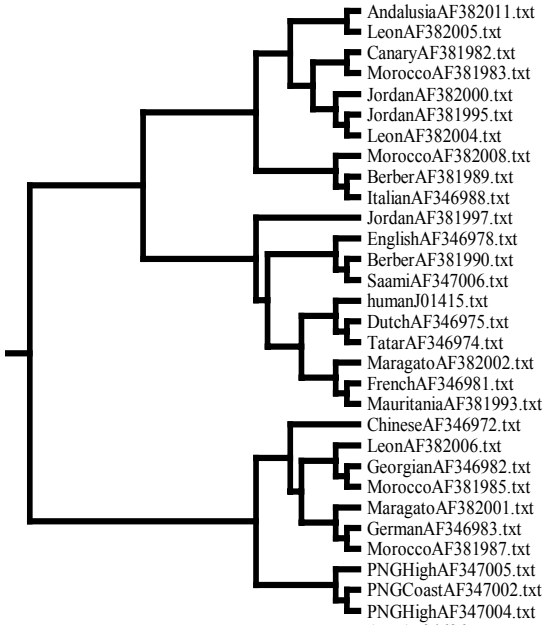
After comparing the two alignments, we conclude that the fast alignment is fairly accurate at grouping a sequence with a few other sequences consistently, but not at grouping a large set of sequences with another set of sequences. In order to get a very accurate phylogenetic tree, the slow alignment algorithm must be implemented. Fast alignment is a good first-order estimate of relations between sequences, but slow alignment fine-tunes the process.

3.2 WSRalign and PAUP

WSRalign allows a dynamic scoring system (determined by user) to be implemented. We ran WSRalign with five different scoring systems (see Figures 4-8):

- 2 for match, -1 for mismatch, and -2 for gap;
- 2 for match, 1 for mismatch, and 0 for gap;
- 3 for match, 0 for mismatch, and -3 for gap;
- 2 for match, -2 for mismatch, and -4 for gap;
- 1 for match, 0 for mismatch, and 0 for gap.

We next attempted to determine if there were any relations that were robust enough to be present in several scoring systems. The largest robust series of relations was thirty ethnicities long.



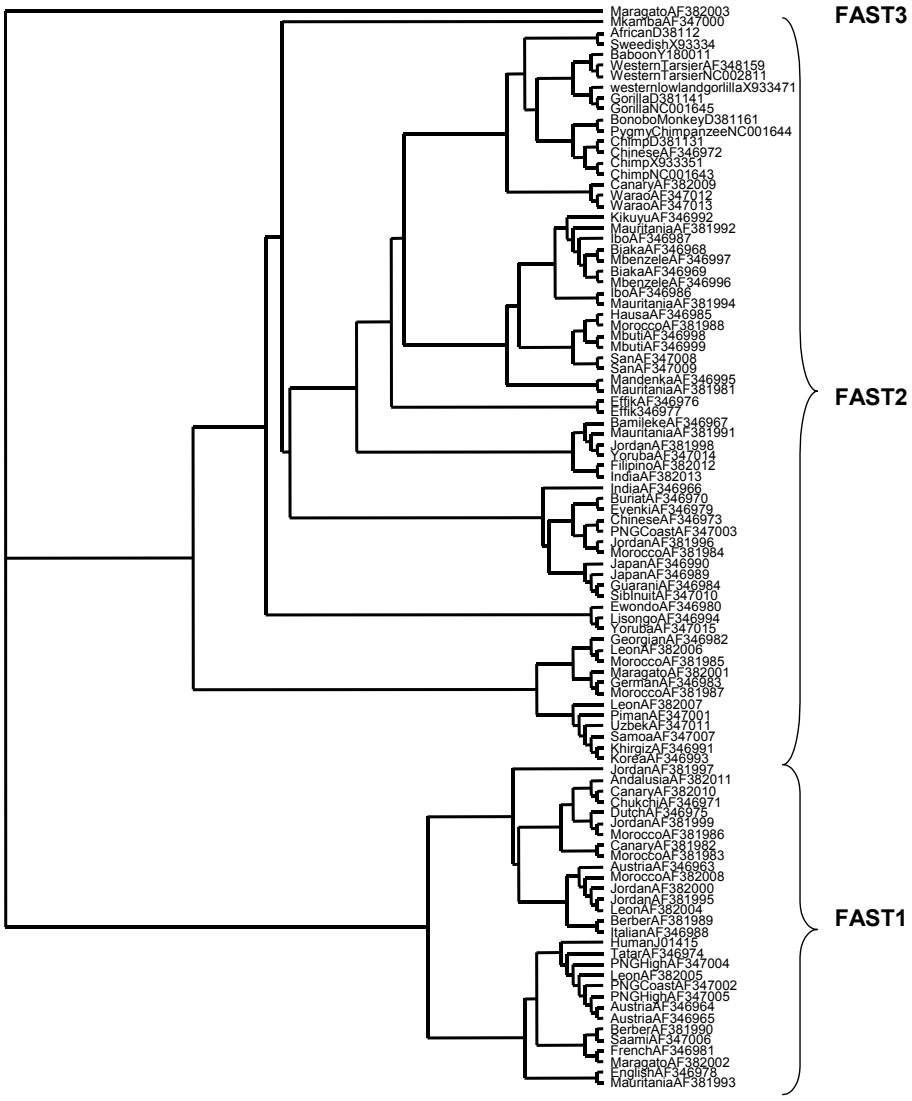


Figure 2: A phylogenetic tree for the fast alignment of sequences using the ClustalX program

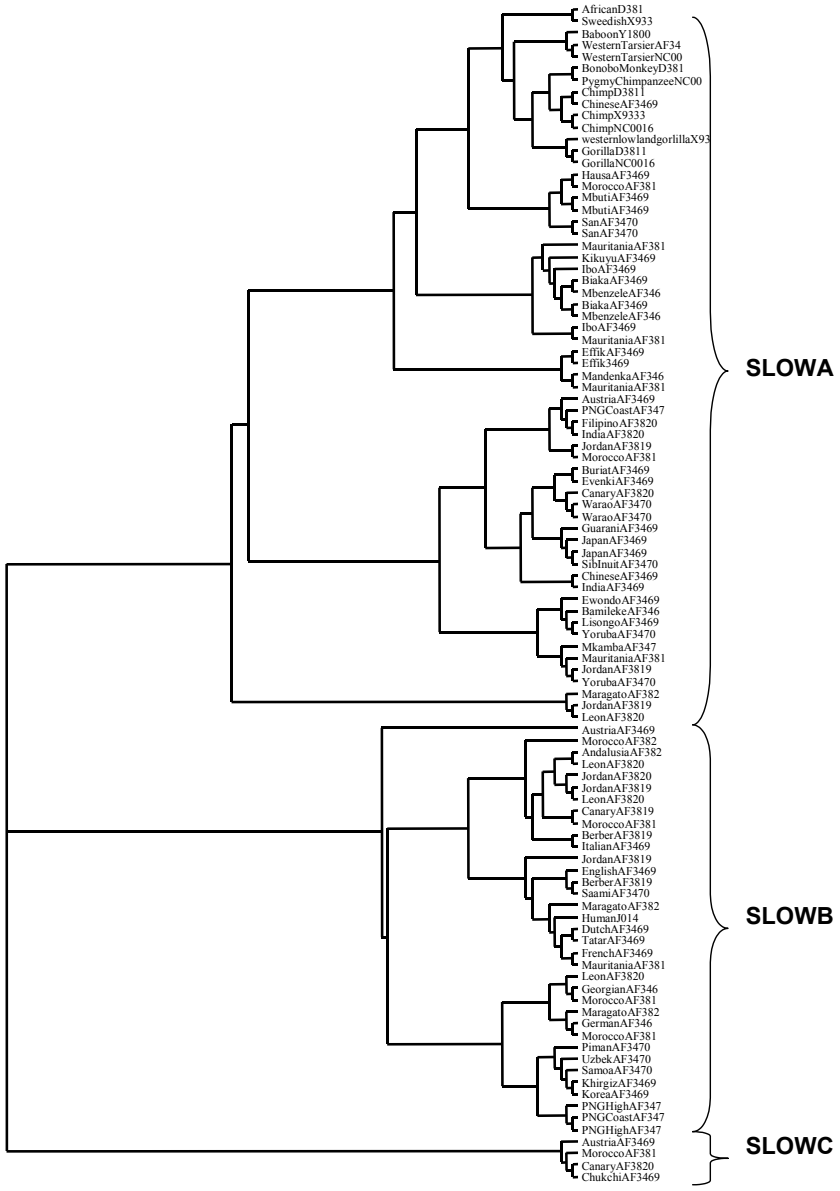
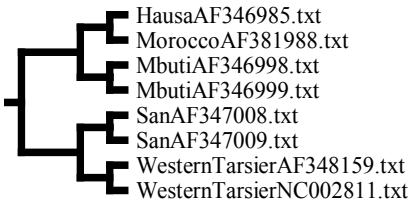


Figure 3: A phylogenetic tree for the slow and accurate alignment of sequences using the ClustalX program

One sequence that was present with great frequency is the one presented below. The Western Tarsier monkey that was introduced with the other ten varieties of primates for points of reference always was scored as being closer to humans than to their fellow primates, and in all but two cases showed up in the following sequence:

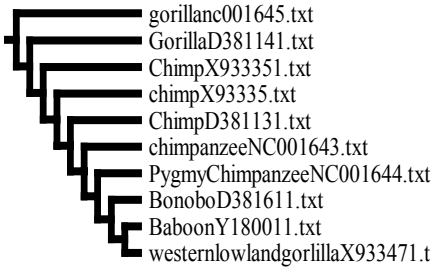


Another sequence that was present in all but one case (in which they appeared in a different order, however still together), was the following:

One segment seems to point towards a relationship between some Native American groups and some Asian groups.

The relation above is invariant; however, it shows up in various locations (The Guarani are a Native American group). This last grouping also tends to be close to the Asian groups (Buriat, Evenki, Filipino, and India), as well as the two Warao Native American data sets.

In all of the scorings done by WSRalign, all of the primates with the exception of the Western Tarsier, are grouped together and completely separate from *Homo Sapiens*.



4 Conclusions

Several immediate conclusions can be drawn from the results obtained in this project. Some of the groupings created by the program WSRalign were robust. The ClustalX and WSRalign programs produced many similarities and some differences in the phylogenetic groupings.

The WSRalign program produced certain groupings within ethnicities despite the changes in the scorning matrix. Native American groups and several Asian groups including the Guarani were consistently associated with Indian, Chinese, Japanese, and Siberian Inuit groups. This result appears to provide evidence for the theory that Native Americans migrated from Asia.

An immediate difference between the ClustalX program and the WSRalign program is the fact that ClustalX grouped the primates' genetic sets within the range of human taxa, including one human data set among the chimpanzees, whereas WSRalign had only one interaction between primates and humans. That was the Western Tarsiers, where previous studies have shown to be similar to humans in the area of mtDNA [3].

Several groupings were universal to both programs. The LeonAF382006, GeorigianAF346982, MorrocoAF381985, MaragatoAF382001, GermanAF346983, and MorrocoAF381987 were grouped identically using both alignment programs.

Despite these similarities, we cannot reach any definitive conclusion about the accuracy of these experimental data. Biological and historical facts need to be taken into consideration to make more accurate comparisons and to adjust the output of the phylogenetic programs.

References

[1] *ClustalX, http://www.ebi.ac.uk/clustalw/.*

[2] *GenBank, http://www.ncbi.nlm.nih.gov/.*

[3] *Origins of Modern Humans in Africa, http://www.indiana.edu/ origins/teach/.*

[4] *PAUP, http://www.paup.csit.fsu.edu.*

[5] D. Gusfield. *Algorithms on Strings, Trees and Sequences. Computer Science and Computational Biology.* Cambridge University Press, 1997.

[6] H. Kaessmann and S. Pablo. *The Genetical History of Humans and the Great Apes.* Blackwell Science Ltd, 2002.

[7] D. Krane and M. Raymer. *Fundamental Concepts of Bioinformatics.* Pearson Education, 2003.

[8] P. Pevzner. *Computational Molecular Biology. An Algorithmic Approach.* The MIT Press, 2000.

[9] P. Wenink. Mitochondrial dna sequence evolution in shorebird populations. *WAU*, 1752, 1994.

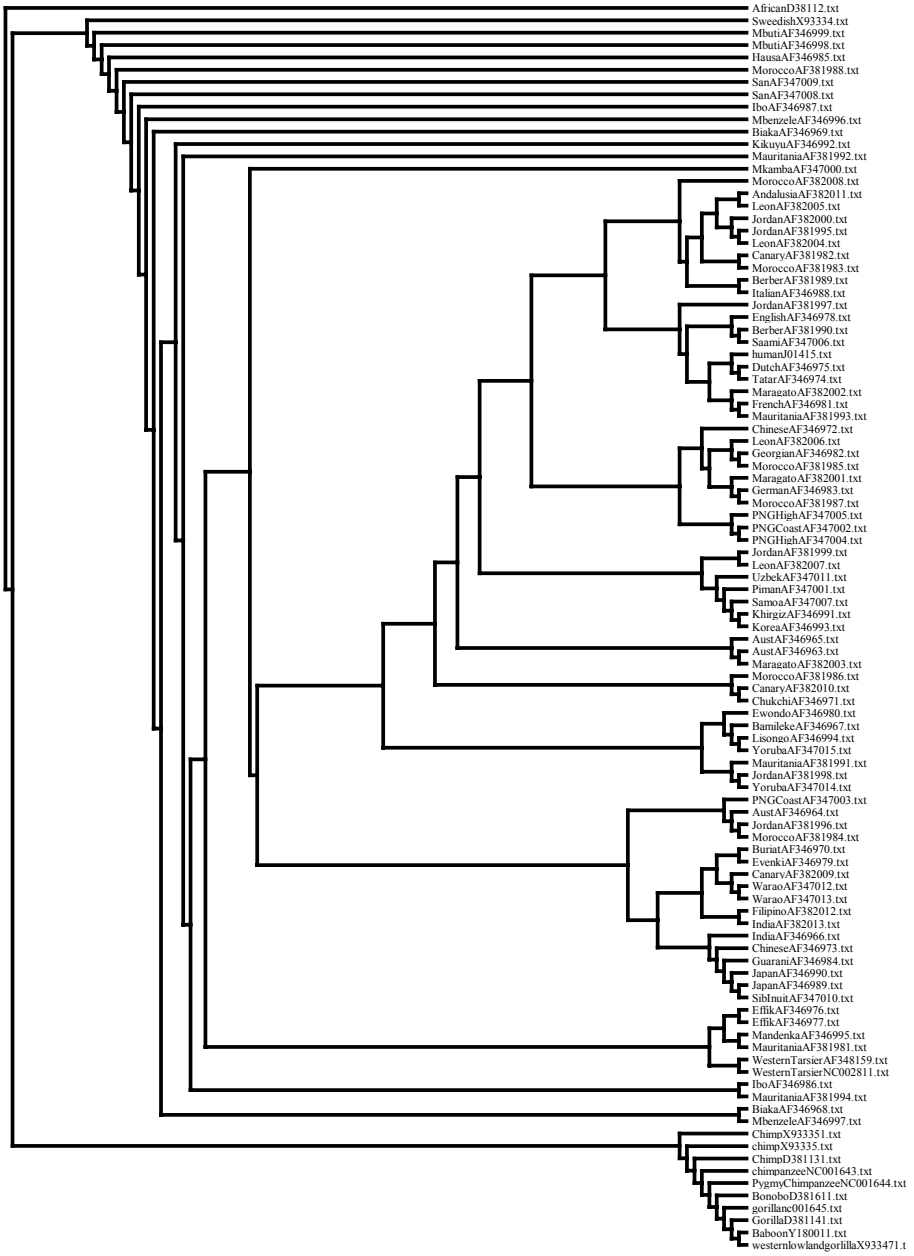


Figure 4: A phylogenetic tree for the alignment of sequences using the WSRalign program: match=1, mismatch=0, gap=0

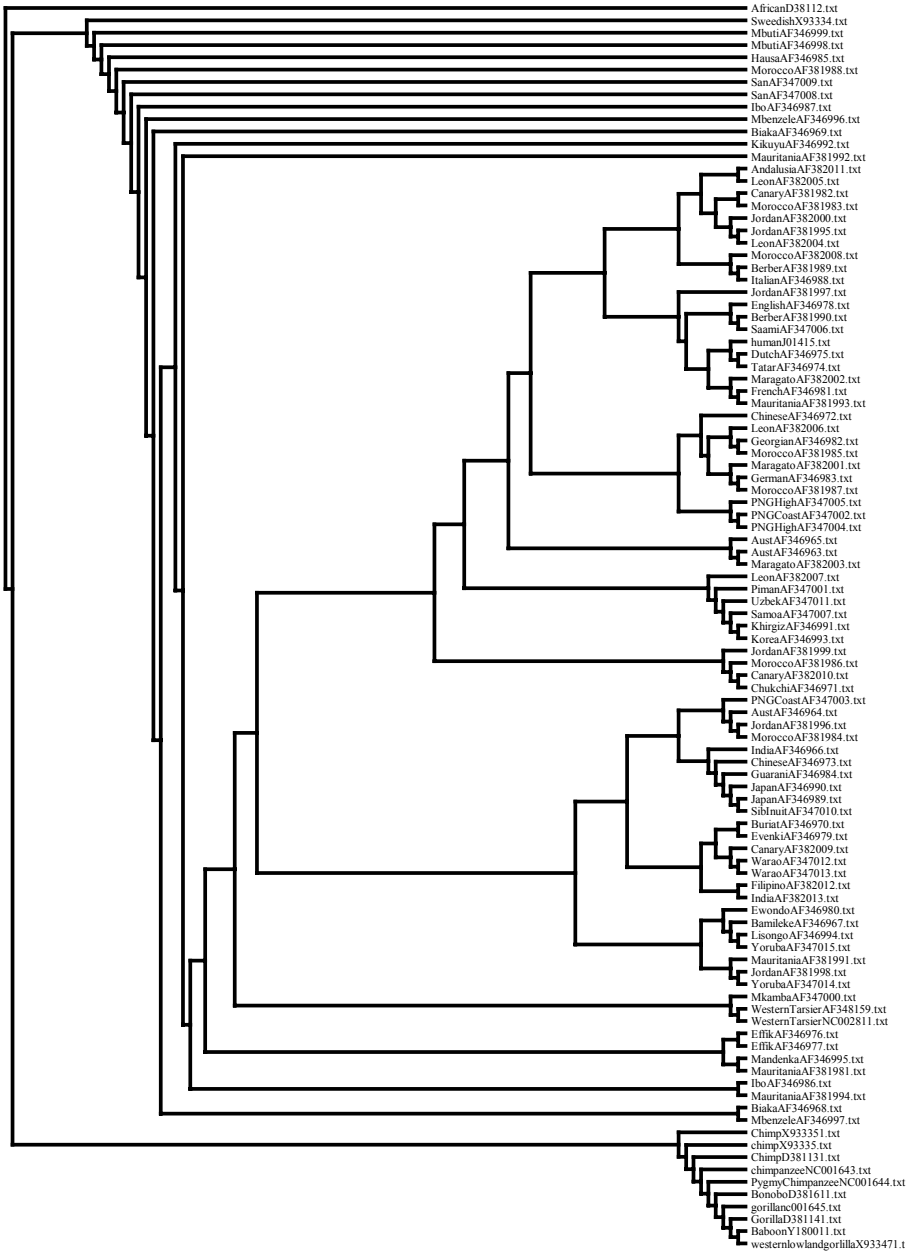


Figure 5: A phylogenetic tree for the alignment of sequences using the WSRalign program: match=2, mismatch=1, gap=0

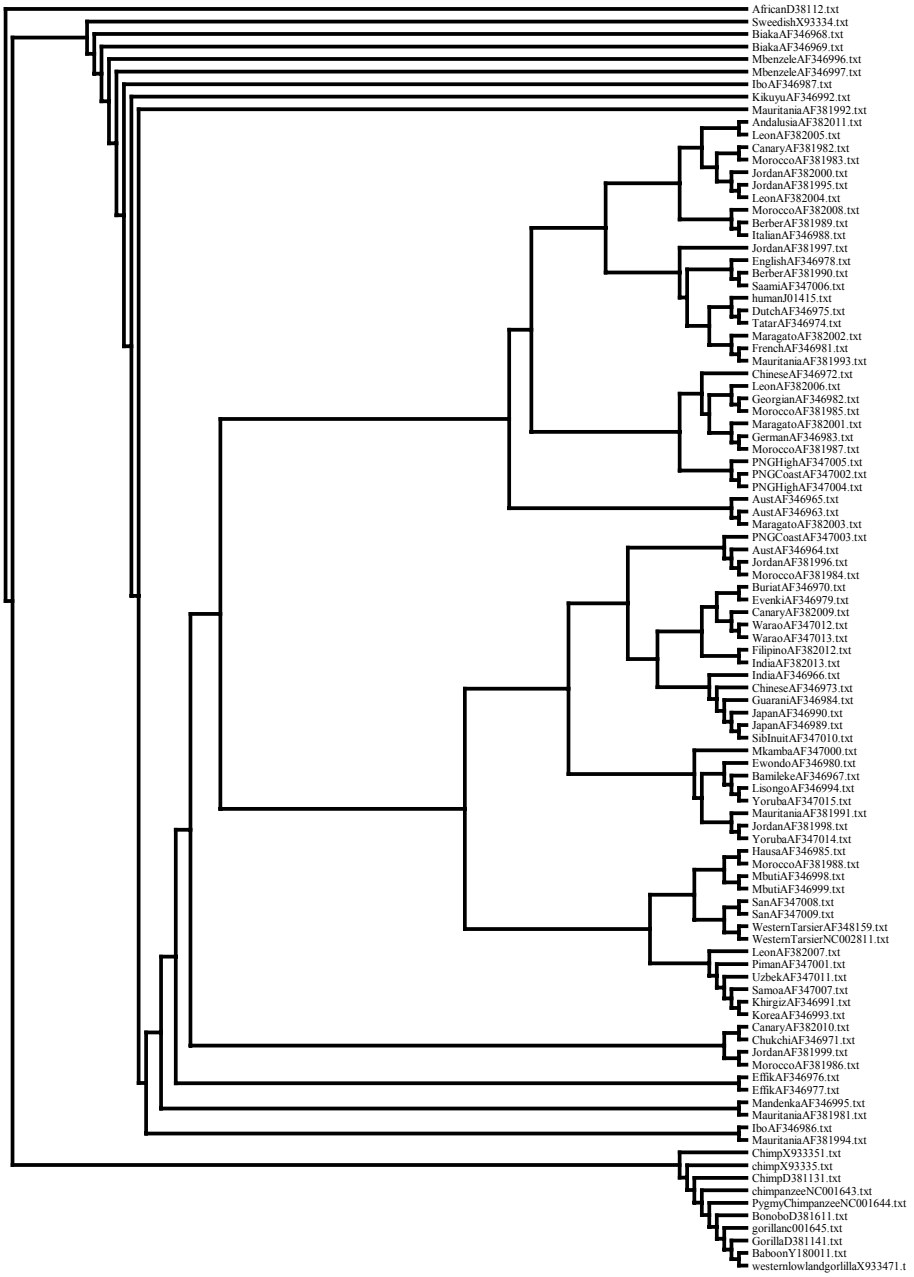


Figure 6: A phylogenetic tree for the alignment of sequences using the WSRalign program: match=2, mismatch=-1, gap=-2

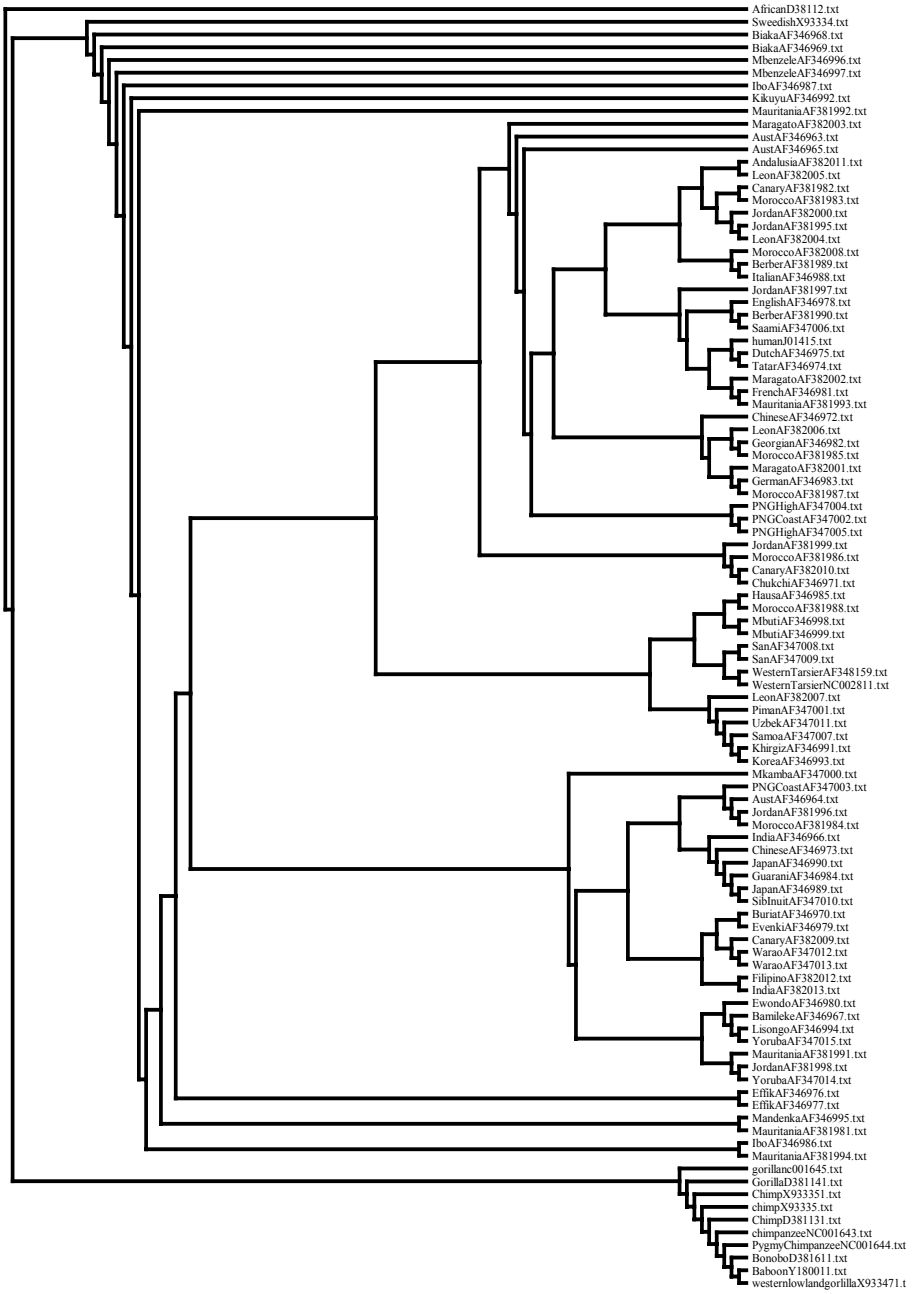


Figure 7: A phylogenetic tree for the alignment of sequences using the WSRalign program: match=2, mismatch=-2, gap=-4

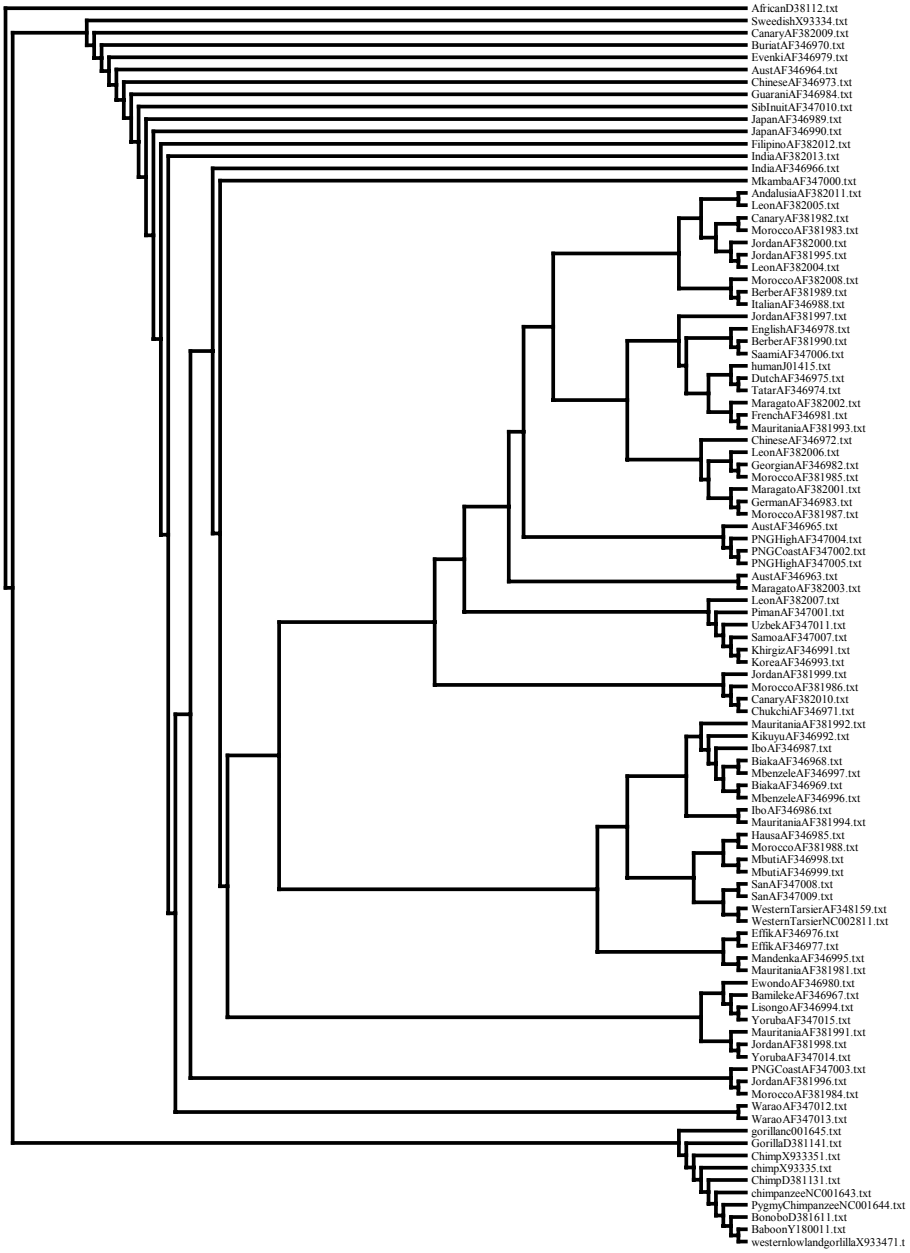


Figure 8: A phylogenetic tree for the alignment of sequences using the WSRalign program: match=3, mismatch=0, gap=-3