Bioinformatics Project II - Clustering Phylogeny

Paulina Kulczyk

January 7, 2024

Abstract

Clustering techniques are used to organize genes or proteins in some natural way, i.e. to organize them into groups of similar structure. We can also represent this relationship using phylogenetic trees. The above methods have applications in finding the evolutionary relatedness of proteins/genes as well as organisms containing the same proteins/genes. In this project, we will focus on tracing the affinity of proteins of the blood coagulation cascade. We will consider 8 organisms: H. sapiens and animal organisms that had high similarity in protein structure to human proteins, according to BLAST NCBI: S. syndactylus, P. abelii, R. roxellana, P. anubis, M. fascicularis, M. mulatta, F. catus.

1 Description of database

The process of blood clotting is one of the most important processes in our bodies. Thanks to it, we do not bleed out when there is a break in the continuity of our vascularized tissues (including, among others, the simplest skin cuts). This process can occur due to the participation of several important proteins - the so-called blood clotting proteins. It turns out that an analogous process also occurs in animals, in mammals it looks very similar and involves mostly the same proteins as in humans (this means that clotting is highly conserved throughout biology). We decided to exploit this fact and look at the structural similarity of these important proteins between organisms. Based on the resulting phylogenetic trees and clusters, we also hope to find evolutionary relatedness between the organisms. In our project, we considered the following blood coagulation proteins: prothrombin (factor II), factor VII, factor IX, factor X, factor XI, factor XII, protein C, thrombomodulin. The organism we considered are: H. sapiens, S. syndactylus, P. abelii, R. roxellana, P. anubis, M. fascicularis, M. mulatta, F. catus. The protein sequences used in the project were sourced from the NCBI database. Organisms that were selected as having proteins similar to those from humans were chosen using the BLAST program launched by the NCBI website. The entire database, including the protein ID number from the NCBI database, is shown in the table 1:

Table 1: Table with proteins and organism

Index	ID	Organism	Protein	NCBI Protein Description
0	AAA51983.1	H. sapiens	factor VII	factor VII
1	AAA52421.1	H. sapiens	factor X	coagulation factor X
2	AAA60166.1	H. sapiens	protein C	protein C
3	AAB59490.1	H. sapiens	factor XII	coagulation factor XII
4	AAB59508.1	H. sapiens	thrombomodulin	thrombomodulin
5	AAC63054.1	H. sapiens	prothrombin	prothrombin
6	AAO15583.1	H. sapiens	factor XI	coagulation factor XI, partial
7	AAO15585.1	F. catus	factor IX	coagulation factor IX, partial
8	BAA07807.1	F. catus	protein C	protein C, partial
9	BAX39000.1	F. catus	factor XII	coagulation factor XII
10	CCA61112.1	H. sapiens	factor IX	coagulation factor IX
11	EHH22946.1	M. mulatta	prothrombin	prothrombin
12	EHH29132.1	M. mulatta	factor X	coagulation factor X
13	EHH56297.1	M. fascicularis	prothrombin	prothrombin
14	KAI4027918.1	H. sapiens	factor XI	coagulation factor XI
15	NP_001073605.1	M. mulatta	factor VII	coagulation factor VII precursor
16	NP_001126851.1	P. abelii	prothrombin	prothrombin precursor
17	NP_001162447.1	P. anubis	factor X	coagulation factor X precursor
18	NP_001252978.1	M. mulatta	thrombomodulin	thrombomodulin precursor
19	XP_001089771.2	M. mulatta	factor XII	coagulation factor XII isoform X1
20	XP_002815399.1	P. abelii	factor XI	coagulation factor XI
21	XP_002816321.2	P. abelii	factor XII	coagulation factor XII
22	XP_002824495.1	P. ableii	factor X	coagulation factor X isoform X1
23	XP_002832230.2	P. abelii	factor IX	coagulation factor IX isoform X1
24	XP_003899481.2	P. anubis	factor XI	coagulation factor XI isoform X1
25	XP_003900640.2	P. anubis	factor XII	coagulation factor XII isoform X3
26	XP_003905194.2	P. anubis	thrombomodulin	thrombomodulin
27	XP_003910009.2	P. anubis	prothrombin	prothrombin
28	XP_003918402.1	P. anubis	factor IX	coagulation factor IX isoform X1
29	XP_003980582.1	F. catus	factor VII	coagulation factor VII
30	XP_003980607.2	F. catus	factor X	coagulation factor X
31	XP_003993267.2	F. catus	prothrombin	prothrombin
32	XP_005556540.2	M. fascicularis	factor XI	coagulation factor XI isoform X1
33	XP_005568247.2	M. fascicularis	thrombomodulin	thrombomodulin
34	XP_005572933.2	M. fascicularis	protein C	vitamin K-dependent protein C isoform X3
35	XP_005586353.1	M. fascicularis	factor X	coagulation factor X
36	XP_005594774.1	M. fascicularis	factor IX	coagulation factor IX isoform X1
37	XP_006930061.3	F. catus	thrombomodulin	thrombomodulin
38	XP_009183375.3	P. anubis	protein C	vitamin K-dependent protein C isoform X2
39	XP_009190519.2	P. anubis	factor VII	coagulation factor VII isoform X2
40	XP_010358374.2	R. roxellana	factor X	coagulation factor X
41	XP_010358380.2	R. roxellana	factor VII	coagulation factor VII isoform X5
42	XP_010361751.1	R. roxellana	protein C	vitamin K-dependent protein C isoform X1
43	XP_010367395.2	R. roxellana	thrombomodulin	thrombomodulin
44	XP_010369387.1	R. roxellana	prothrombin	prothrombin isoform X1
45	XP_010376291.1	R. roxellana	factor XI	coagulation factor XI isoform X2
46	XP_010384732.2	R. roxellana	factor XII	coagulation factor XII
47	XP_010387043.1	R. roxellana	factor IX	coagulation factor IX

48	XP_014965335.1	M. mulatta	protein C	vitamin K-dependent protein C isoform X1
49	XP_014995139.1	M. mulatta	factor XI	coagulation factor XI isoform X2
50	XP_015307796.2	M. fascicularis	factor XII	coagulation factor XII isoform X1
51	XP_024094817.2	P. abelii	thrombomodulin	thrombomodulin
52	XP_024099271.2	P. abelii	protein C	vitamin K-dependent protein C isoform X2
53	XP_028697499.1	M. mulatta	factor IX	coagulation factor IX isoform X1
54	XP_045233471.1	M. fascicularis	factor VII	LOW QUALITY PROTEIN: coagulation factor VII
55	XP_054385267.1	P. abelii	factor VII	coagulation factor VII isoform X2
56	XP_055101117.1	S. syndactylus	factor X	coagulation factor X isoform X2
57	XP_055101306.1	S. syndactylus	factor VII	LOW QUALITY PROTEIN: coagulation factor VII-like
58	XP_055118090.1	S. syndactylus	protein C	vitamin K-dependent protein C isoform X4
59	XP_055122104.1	S. syndactylus	thrombomodulin	thrombomodulin
60	XP_055123354.1	S. syndactylus	factor IX	coagulation factor IX isoform X1
61	XP_055132633.1	S. syndactylus	factor XI	coagulation factor XI
62	XP_055137660.1	S. syndactylus	prothrombin	prothrombin
63	XP_055141125.1	S. syndactylus	factor XII	coagulation factor XII

2 Methods and algorithms

In our work we used clustering and phylogenetic tree building methods.

In the first part focused on clustering, we used the BLAST program and the blastp tool. Using it, we obtained a database, based on which we created a similarity matrix. This matrix contained information about the degree of identity between all the proteins analyzed. We then used the obtained matrix in the hierarchical clustering algorithm. This algorithm is a type of clustering algorithm that builds a hierarchy of clusters. It organizes the points from the matrix into a tree-like structure known as a dendrogram. We visualize the resulting dendrogram to show the relationships and similarities between data points at different levels of detail. The horizontal lines on the dendrogram represent the merging (agglomeration) of clusters. The height of the vertical line on the dendrogram represents the dissimilarity (or similarity) at which the clusters were merged or divided. In our work, we chose the Average Linkage method. This method measures the average distance between all pairs of members in two clusters.

In the second part of the work focused on phylogenetics, we used the clustalw2 program. In it, we selected Multiple Sequence Alignment options. Based on the resulting calculations on different groups of proteins (the entire protein base, proteins of one type on different organisms and proteins from one cluster), we created phylogenetic trees. To create these trees, we used the biopython tool, specifically the Bio.Phylo.TreeConstruction module, in which the most important method is DistanceCalculator, which calculates a distance matrix based on protein sequence matching. We obtained the tree structure using the Bio.Phylo.TreeConstruction.DistanceTreeConstructor method, choosing the UPGMA algorithm. This algorithm was discussed by us in class. In short, it involves iteratively counting the distance between clusters. This iteration occurs until all data points are part of a single cluster. We then used the Bio.Phylo.draw module to visualize the resulting structure.

Finally, we created consensus trees using the Bio.Phylo.consensus module.

3 Clustering

In this section, we will focus on clustering. We will check whether group of proteins searched using BLAST on the NCBI website (i.e., the original similarity groups for a given protein created by BLAST NCBI were 8 in size) join the same clusters as proteins matched using hierarchical clustering. We

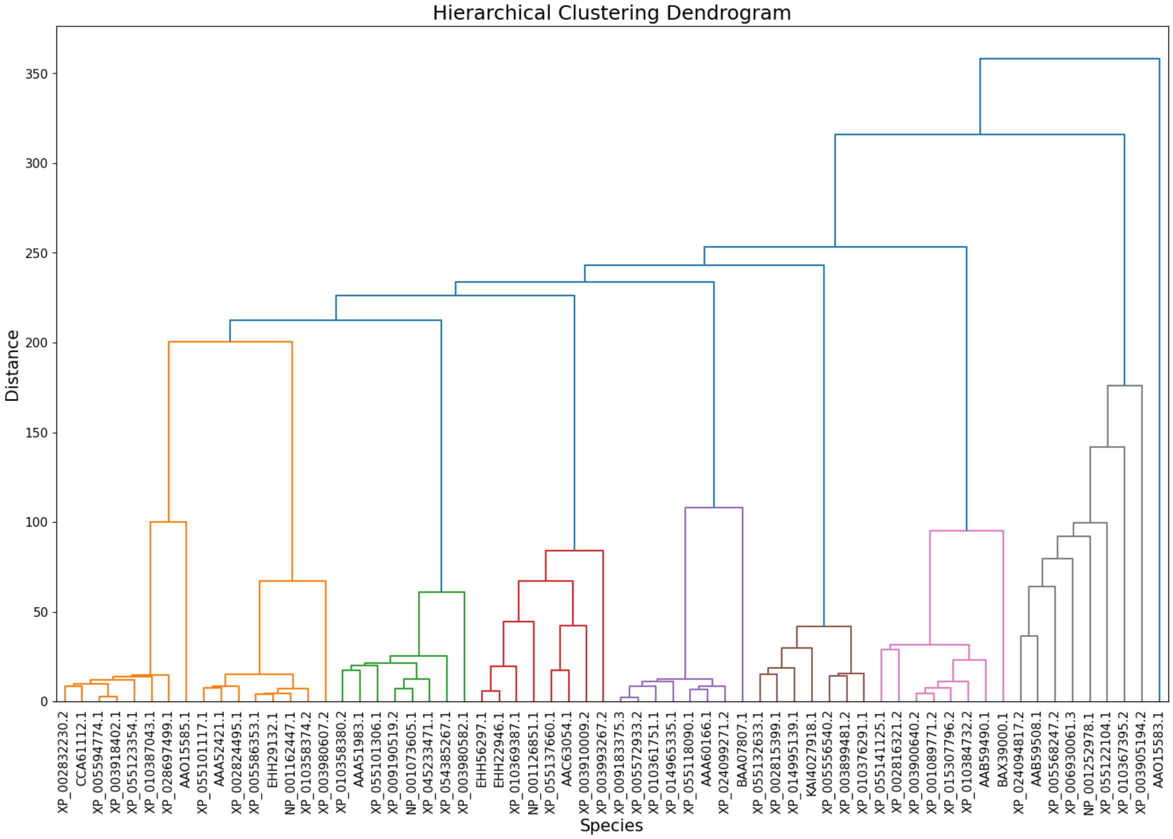
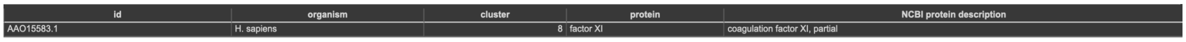


Figure 1: Hierarchical clustering of all proteins.



id	organism	cluster	protein	NCBI protein description
AAO15583.1	H. sapiens	8	factor XI	coagulation factor XI, partial

Figure 2: Hierarchical clustering of all proteins.

created the dendrogram so that 8 clusters would form, since we were considering 8 coagulation proteins. Below is the resulting dendogram of clusters using hierarchical clustering:

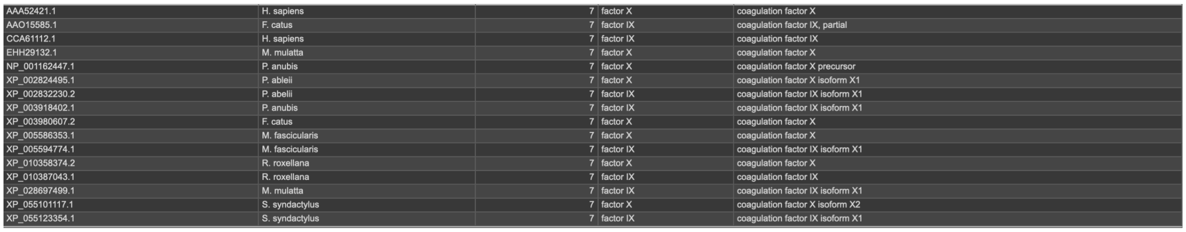
At a glance, we can see that not all clusters contain 8 observations each. In particular, the blue cluster, which consists of one protein, and the orange cluster, which consists of 16 proteins, stand out. Let's first take a closer look at these two clusters. In order to make it easier for us to look at each cluster group, we have created tables in relation to the cluster index.

The protein in the blue cluster is: coagulation factor XI partial, from the human body. This may mean that this protein is the farthest evolutionarily from the others and has reached a separate evolutionary path. However, it is also possible that it has not been assigned to any cluster because it is a partial sequence. Hence, we cannot draw too rash conclusions about this protein.

The orange cluster is composed of factor IX and factor XI proteins. Note, however, that two subgroups can be distinguished in it: the first to the left of the AAA protein (containing it), the second to the right of the XP protein (containing it). This means that as if we superimposed 9 clusters on our algorithm, this is where the breakdown would occur. However, we can conclude that there is the greatest structural similarity between these two groups of proteins. Which may indicate their close evolutionary relationship.

The remaining clusters consist of groups of identical proteins. Which may indicate an evolutionary leap between their formation and a significantly different function in the coagulation process, however, the same for all organisms (since structurally they are very similar between organisms).

Two pairs of proteins are still worth noting on the above dendrogram: (009183375.3 and 005572933.2) and (003918402.1 and 005594774.1). Both the first and second pair of proteins belong to the following organisms: (P. anubis and M. fascicularis). On the dendrogram, their similarity distance is very small. This means that these two organisms may be very related to each other.



AAA52421.1	H. sapiens	7	factor X	coagulation factor X
AAO15585.1	F. catus	7	factor IX	coagulation factor IX, partial
CCA61112.1	H. sapiens	7	factor IX	coagulation factor IX
EHH29132.1	M. mulatta	7	factor X	coagulation factor X
NP_001162447.1	P. anubis	7	factor X	coagulation factor X precursor
XP_002824495.1	P. ableii	7	factor X	coagulation factor X isoform X1
XP_002832230.2	P. abelii	7	factor IX	coagulation factor IX isoform X1
XP_003918402.1	P. anubis	7	factor IX	coagulation factor IX isoform X1
XP_003980607.2	F. catus	7	factor X	coagulation factor X
XP_005586353.1	M. fascicularis	7	factor X	coagulation factor X
XP_005594774.1	M. fascicularis	7	factor IX	coagulation factor IX isoform X1
XP_010358374.2	R. roxellana	7	factor X	coagulation factor X
XP_010387043.1	R. roxellana	7	factor IX	coagulation factor IX
XP_028697499.1	M. mulatta	7	factor IX	coagulation factor IX isoform X1
XP_055101117.1	S. syndactylus	7	factor X	coagulation factor X isoform X2
XP_055123354.1	S. syndactylus	7	factor IX	coagulation factor IX isoform X1

Figure 3: Hierarchical clustering of all proteins.

Name of protein	Group
thrombomodulin	1
factor IX	2
factor XII	3
factor X	4
factor XI	5
factor VII	6
protein C	7
prothrombin	8

Table 3: Groups of proteins.

The more detailed analysis of above dendrogram, however, is left to those willing to study it.

4 Phylogenetics

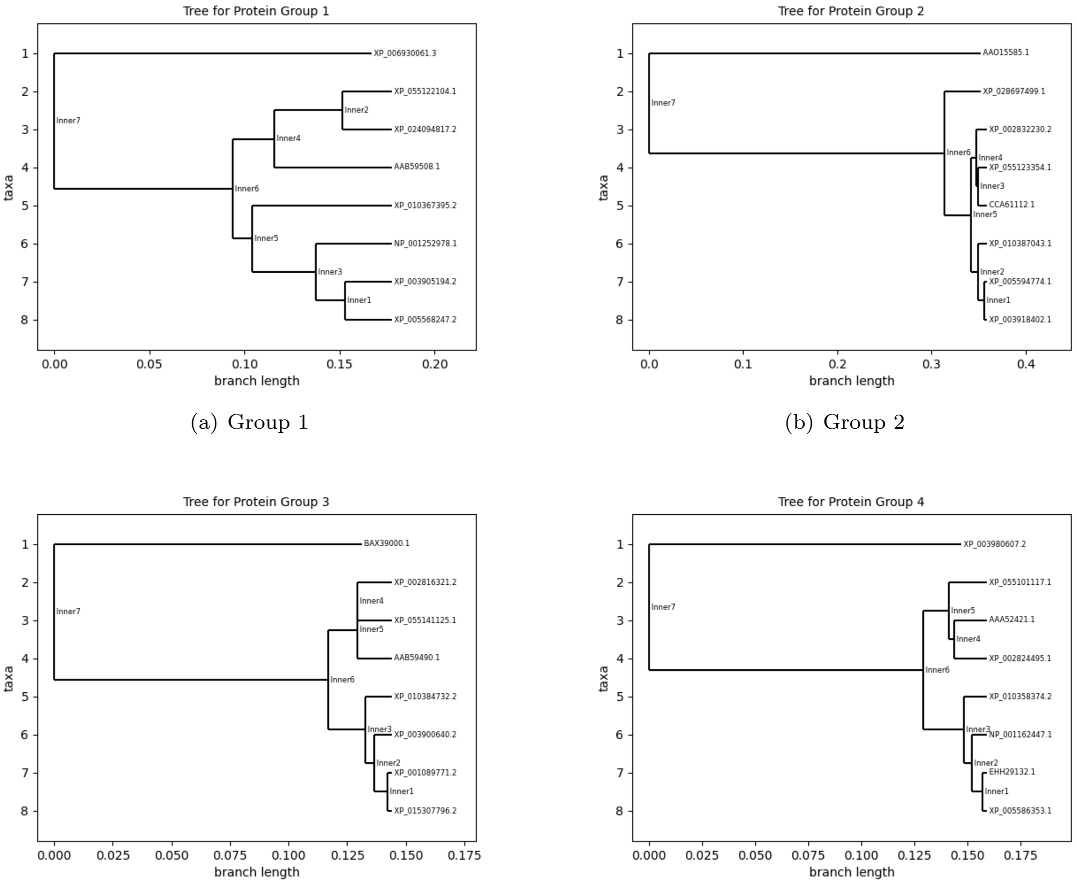
In this chapter, we will focus on analyzing phylogenetic trees for groups of proteins, clusters from the 3 branch and all proteins considered in this project, respectively. We will then apply colors to the tree for all proteins: first relative to the organism and then relative to the protein group. This will allow us to see how adding colors to a dendogram can facilitate its analysis. Aboveato the main conclusions we are going to draw from here is the possible evolutionary similarity between organisms and based on the whole tree between proteins. The colors will also help us distinguish clades.

4.1 Trees for each "group" of proteins

Recall that in our project we consider proteins of the blood coagulation cascade. The table 3 shows the assignment of proteins to their respective groups. We insert the results of our computations in the figure 4.

Analyzing the above trees, using our database, we can see the following relationships between proteins:

- for protein 1:
  - the closest relationship is between M. fascicularis and P. anubis, then these organisms have similar this protein to M. mulatta;
  - the second closest relationship is between P. abelii and S. syndactylus, then these organisms have similar this protein to H. sapiens;
  - the least concordant protein in this group relative to the others belongs to F. catus;
- for protein 2:
  - the closest relationship is between M. fascicularis and P. anubis, then these organisms have similar this protein to R. roxellana;
  - the second closest relationship is between H. sapiens and S. syndactylus, then these organisms have similar this protein to P. abelii;
  - the least concordant protein in this group relative to the others belongs to F. catus;



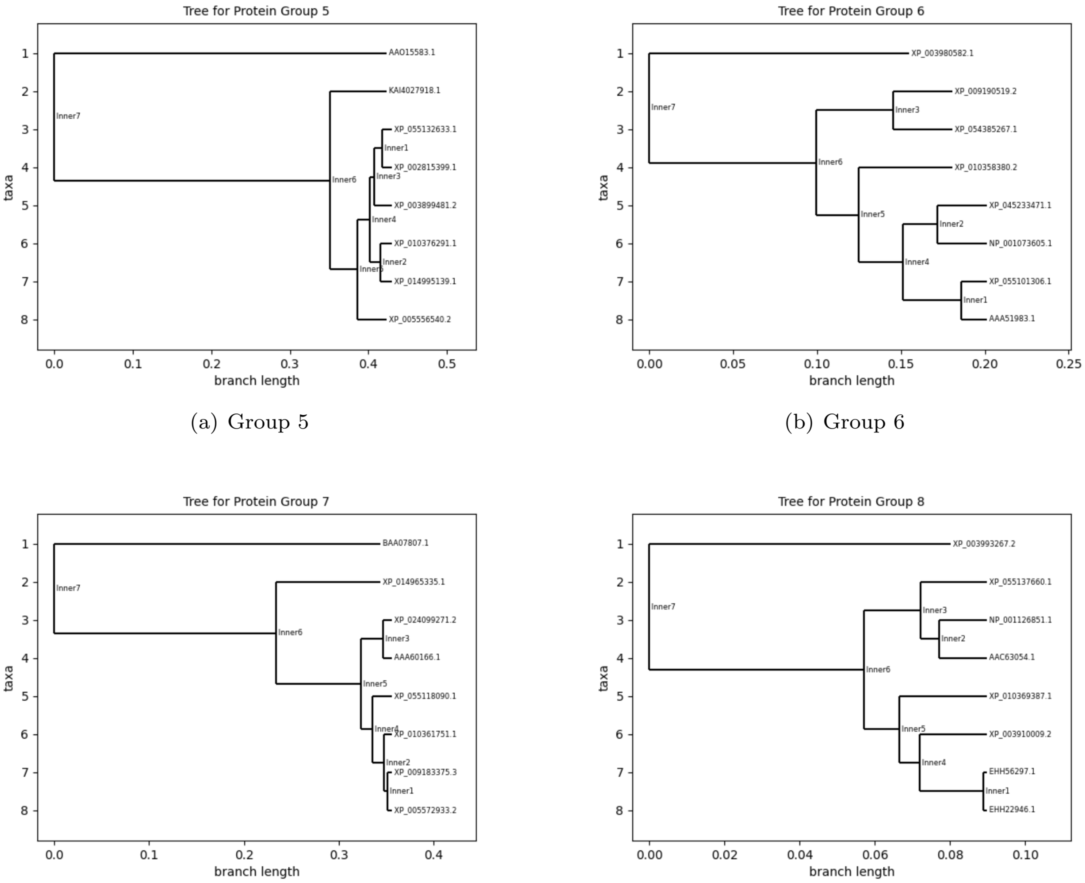
(a) Group 1

(b) Group 2

(c) Group 3

(d) Group 4

Figure 4: Phylogenetic trees for each groups of proteins



(a) Group 5

(b) Group 6

(c) Group 7

(d) Group 8

Figure 4: Phylogenetic trees for each groups of proteins

- for protein 3:
    - the closest relationship is between M. fascicularis and M. mulatta, then these organisms have similar this protein to P. anubis;
    - the least concordant protein in this group relative to the others belongs to F. catu;s
- for protein 4:
    - the closest relationship is between M. fascicularis and M. mulatta, then these organisms have similar this protein to P. anubis;
    - the least concordant protein in this group relative to the others belongs to F. catus;
- for protein 5:
    - the closest relationship is between S. syndactylus and P. abelii, then these organisms have similar this protein to P. anubis;
    - the second closest relationship is between R. roxellana and M. mulatta;
    - the least concordant protein in this group relative to the others belongs to H. sapiens;
- for protein 6:
    - the closest relationship is between S. syndactylus and H. sapiens;
    - the second closest relationship is between M. fascicularis and M. mulatta;
    - the least concordant protein in this group relative to the others belongs to F. catus;
- for protein 7:
    - the closest relationship is between M. fascicularis and P. anubis, then these organisms have similar this protein to R. roxellana;
    - the least concordant protein in this group relative to the others belongs to F. catus;
- for protein 8:
    - the closest relationship is between M. fascicularis and M. mulatta , then these organisms have similar this protein to P. anubis;
    - the second closest relationship is between H. sapiens and P. abelii, then these organisms have similar this protein to S. syndactylus;
    - the least concordant protein in this group relative to the others belongs to F. catus;

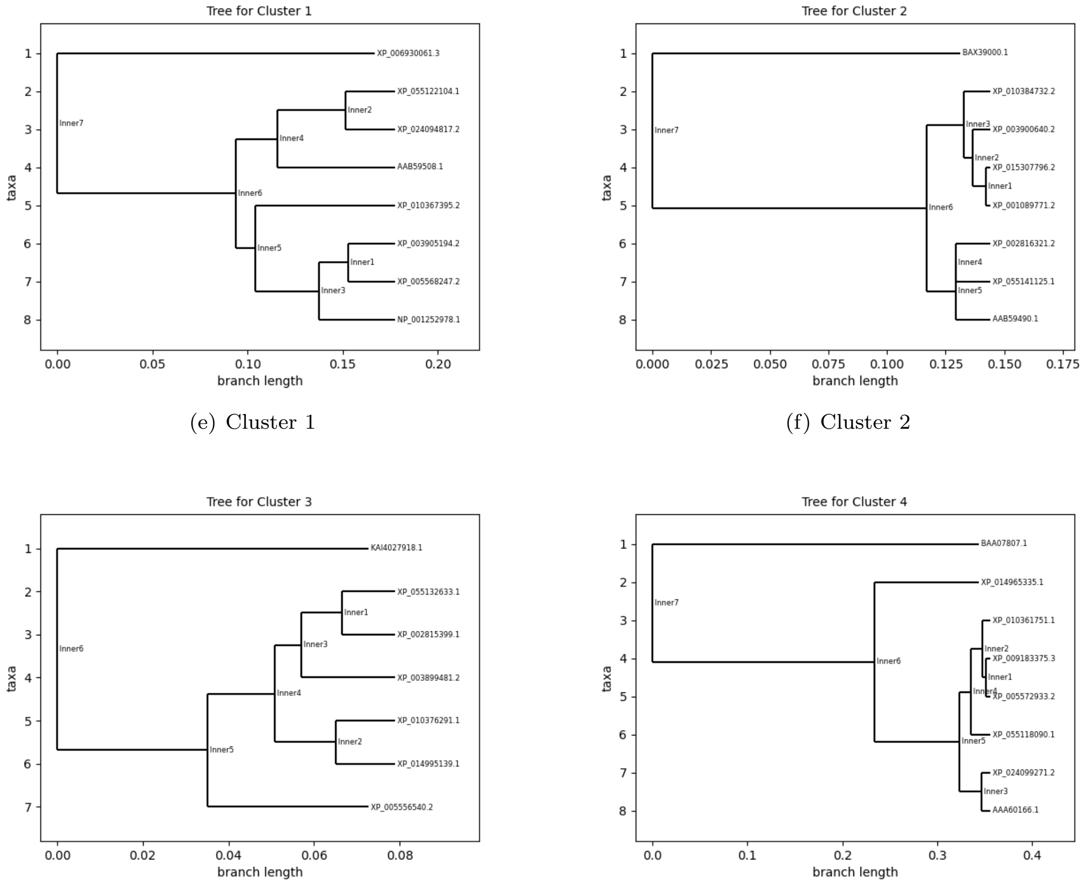
Based on the above summary, we see that probably evolutionarily (especially in terms of the coagulation system) the closest relatives are:

- M. fascicularis and M. muallata and P. anubis.

To a lesser extent:

- H. sapiens and S. syndactylus and P. abelii.

The furthest evolutionarily in terms of the coagulation system relative to the other organisms appears to be F. catus.



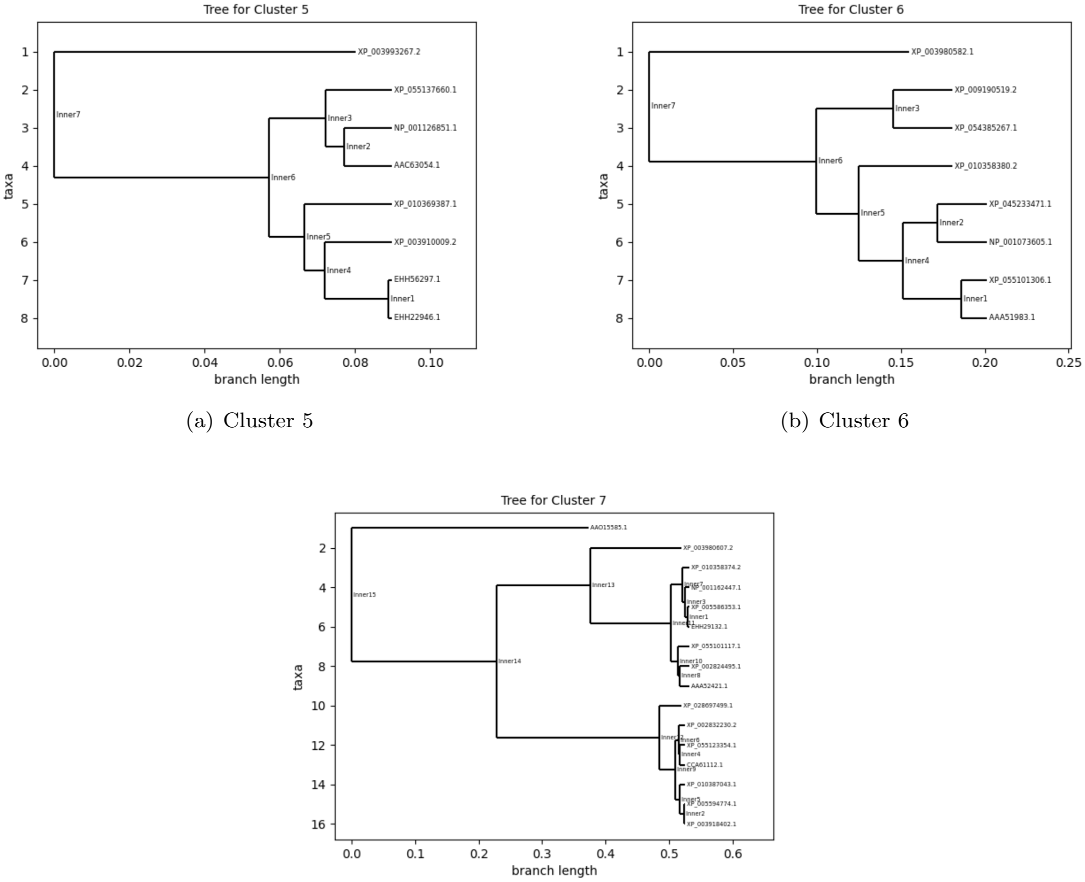
(e) Cluster 1

(f) Cluster 2

(g) Cluster 3

(h) Cluster 4

Figure 5: Phylogenetic trees for each clusters of proteins



(a) Cluster 5

(b) Cluster 6

(c) Cluster 7

Figure 5: Phylogenetic trees for each clusters of proteins

4.2 Trees for each clusters

In the previous chapter, we divided proteins into clusters. Phylogenetic trees corresponding to each cluster are presented in the figure 5. It is also worth mentioning here that the clusters are named as follows: gray cluster is cluster number 1, pink cluster is cluster number 2, brown cluster is cluster number 3, purple cluster is cluster number 4, red cluster is cluster number 5, green cluster is cluster number 6, orange cluster is cluster number 7, blue cluster would be cluster number 8 in this case, but for this cluster as it consisted of only one observation there is no point in building a tree.

Analyzing the phylogenetic trees for the corresponding clusters, we can see that most of them are the same as the phylogenetic trees for the corresponding protein groups. We should not be surprised by this fact, since except for the orange and blue clusters, the other clusters, as we noted, consisted of a single group of proteins. Thus, clustering tree 1 corresponds to the tree for protein group 1, clustering tree 2 corresponds to the tree for protein group 3, clustering tree 4 corresponds to the tree for protein group 7, clustering tree 5 corresponds to the tree for protein group 8, clustering tree 6 corresponds to the tree for protein group 6. Only the clustering trees 3 and 7 differ. The clustering tree 3 is a truncated version of the tree for protein 5, because one protein from the group of proteins 5 (coagulation factor XI partial, from the H. sapiens) was not included in it. The clustering tree 7 is a composite of trees for the group of proteins number 2 and 4. We can see that, in truth, clustering tree 7 is therefore a consensus tree for the trees of proteins groups number 2 and 4.

4.3 Tree for all coagulation proteins

In this subsection, we present the tree built for all the proteins considered in this project. We will present it in three ways - the first way will be visually no different from the previous trees, the next two will have colors added to allow an easier look at the corresponding group of proteins as well as the corresponding group of organisms. These trees are included in the Figures **??**, respectively. When considering trees for all proteins, we must keep in mind that the branches were colored as follows:

- organisms colors:
  - H. sapiens : red;
  - F. catus: green;
  - M. mulatta : blue;
  - M. fascicularis : lime
  - P. abelii : yellow;
  - P. anubis : orange;
  - R. roxellana : salmon;
  - S. syndactylus : aqua;
- proteins colors:
  - factor VII : red;
  - factor X: green;
  - protein C : blue;
  - factor XII : lime
  - factor IX : yellow;
  - prothrombin : orange;
  - thrombomodulin : salmon;
  - factor XIs : aqua;

We can see that the above trees almost all look like the combined phylogenetic trees of the 8 protein groups. So the conclusions we can draw from them are practically the same as in the previous chapters. However, we can see that coloring the branches of the trees relative to the organisms would make it much easier for us to analyze the trees for individual clusters and for individual proteins. The trees

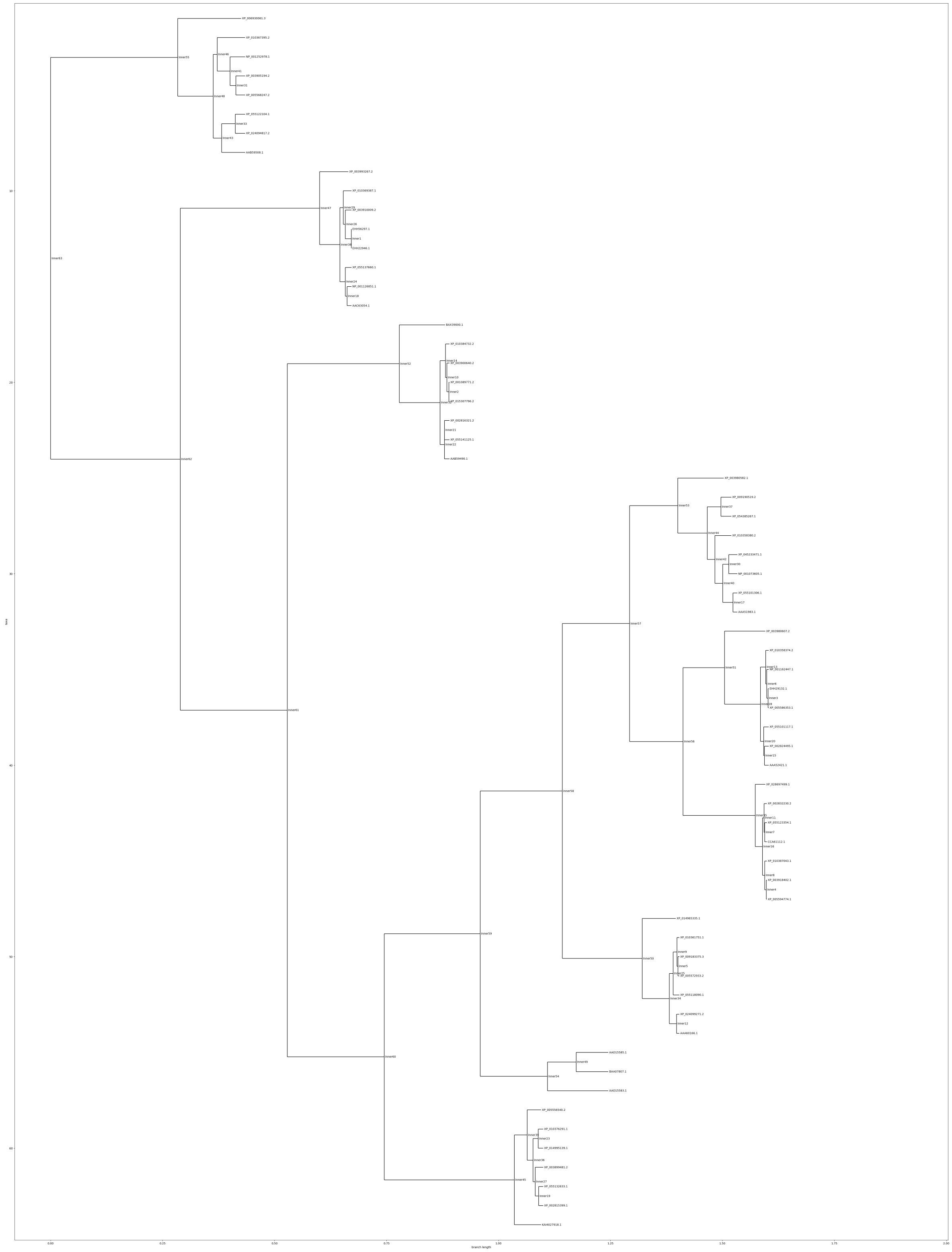


Figure 6: One common tree.

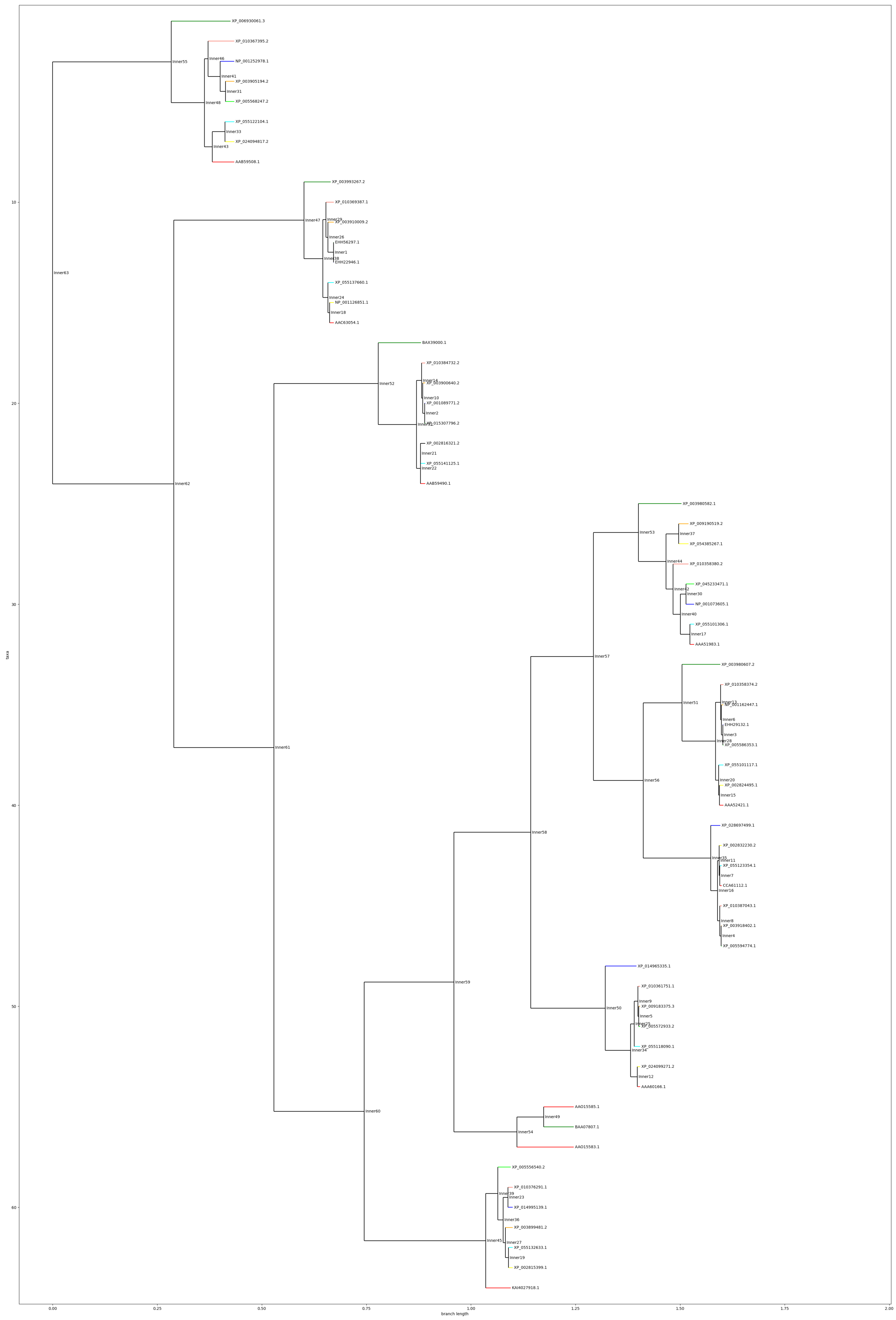


Figure 7: One common tree with color tree branches based on the organism.

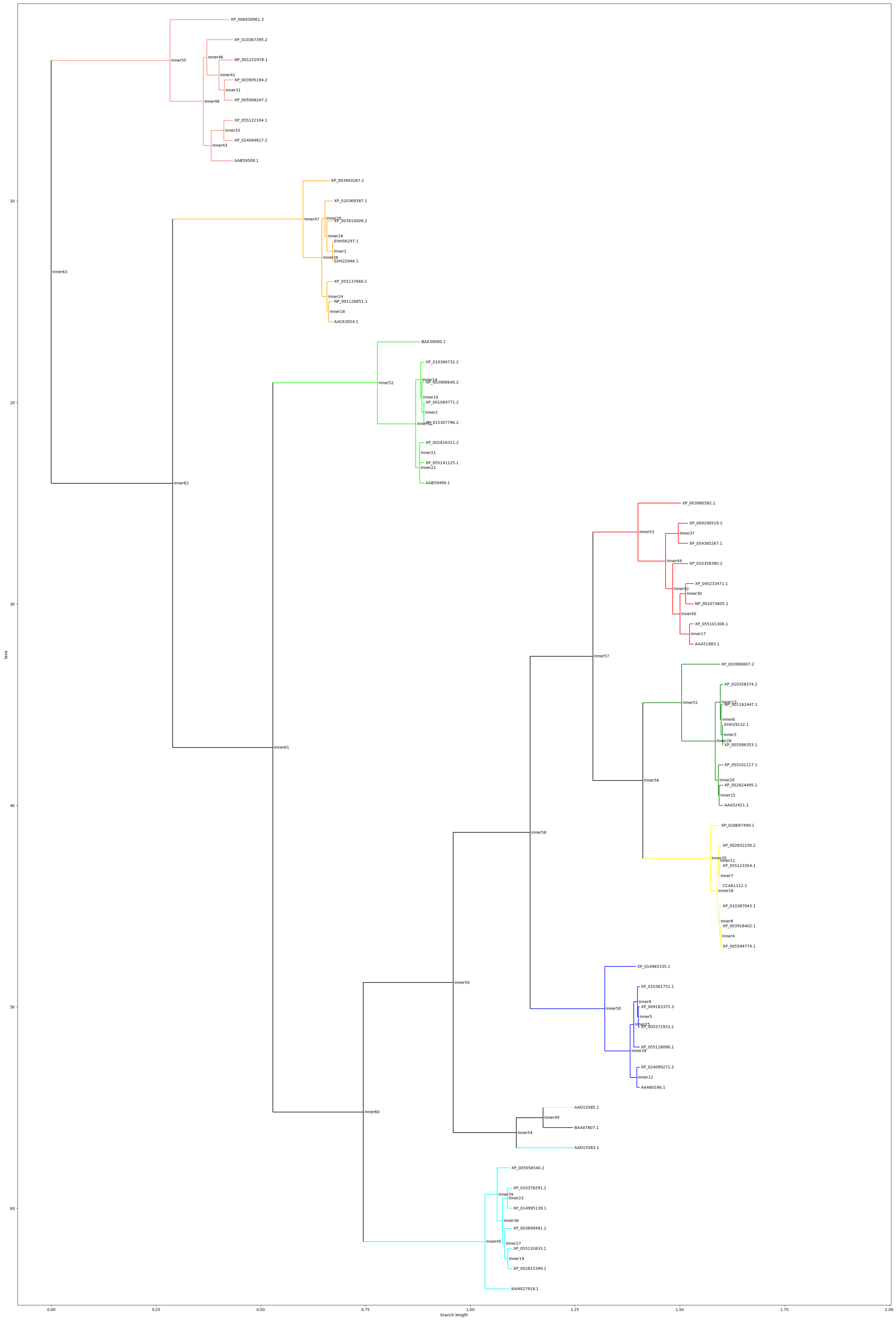


Figure 8: One common tree with olor tree branches based on the protein group.

colored based on groups of proteins draws our attention to the Inner 54 clade. In this tree, we see that this clade is not a clade of a single protein. In this branch we can see that there are proteins: factor IX of the organism F. catus, protein C of the organism F. catus and factor XI partial of the organism H. sapiens. We are somehow not very surprised to find the same organism in one clade. The fact that the H. sapiens organism was also found here may be due to the fact that just this protein was the ”problematic” protein, it was the one that ended up in the blue cluster consisting of only one observation The most interesting on the above trees is the Inner 54 branch.