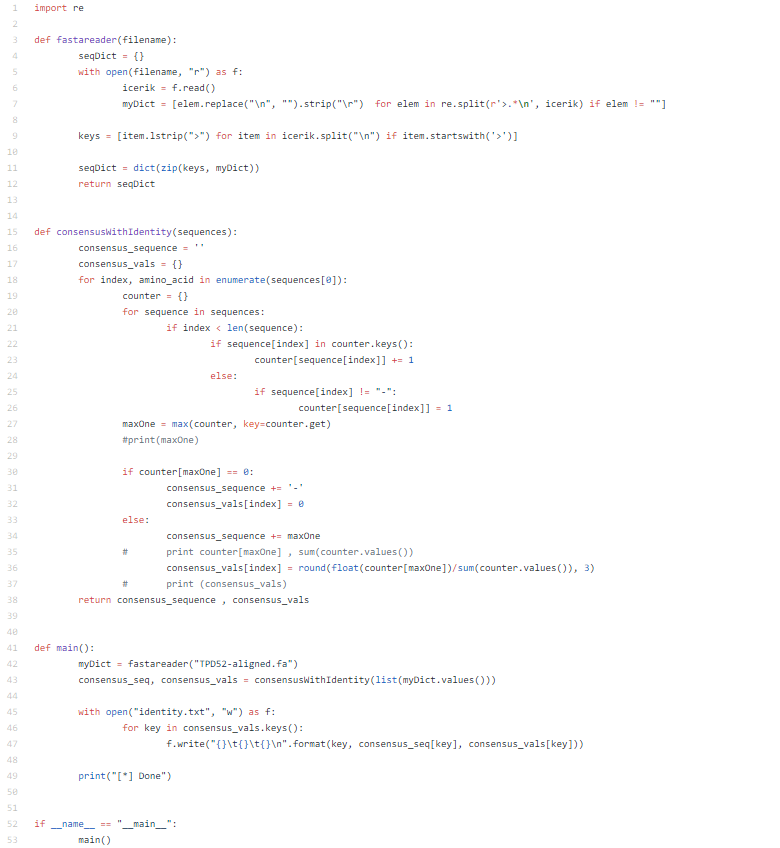
**Introduction**

A protein family is described as a group of proteins that share a common ancestor and evolutionary similar sequences and, therefore similar functional features. The ancestor at the root of hierarchy, is the common ancestor and as going down at the evolutionary tree, subfamilies of proteins are more closely related. Domains are the smallest functional and structural, independently folded units of the proteins. Domains may have evolved from events such as domain deletion, domain fusion etc. In that sense domain architecture might be useful to investigate evolution of a protein family. Moreover, domains are the conserved regions and since they represent the similar regions in different proteins, they can give an idea about protein homology. In this project, D52 protein family is chosen due to its overexpression in tumour cells, it is related especially with breast carcinoma. D52 protein functions are searched and mentioned, also functions are investigated to clarify which regions are conserved and what these conserved regions effects on disease such as cancer are. Furthermore, each subfamily of this protein family is observed to answer following questions: Which regions are conserved in each subfamily and what is the reason of a protein is conserved in a specific subfamily, but it is not conserved in other families? D52 domains’ sequences from 25 different organisms are used to acquire most conserved regions. For analysing these regions, conservation scores are calculated. While conservation scores are calculating, sequences are aligned and for each position of consensus sequence’s conservation proportion is found, for this step an algorithm that is written by us is used. At last, with using chosen domains phylogenetic tree is formed and information about homology of domains is obtained.

**Material Method**

Project has started with searching protein families on PFAM database. While determining protein family, domain organisations were examined and D52 protein has chosen. After choosing D52 protein family, in order to find domains of D52 family again PFAM database is used. However, since there are lots of information about many species, InterPro database which is provided by EMBL-EBI is used to find predicted domains and important sites of D52. Out of 2066 protein that are predicted to be part of D52 family, 60 tumour protein D52 (TPD52) are selected from different organisms. After selecting organisms, their sequences are obtained as FASTA format from InterPro and saved in “TPD52.fa” file. These sequences are aligned by using MEGA7 and saved in “TPD52-aligned.fa” file. Then for the first step their consensus sequence which represents the most common amino acid in each position is found and for each position conservation value is calculated. Following piece of code explains the algorithm of measurement of conservation values that have range between 0 and 1.



Amino acids on each position and calculation results are written in “identity.txt” file to be used draw conservation plot. By using RStudio, conservation plot is drawn. This plot will be used to analyse which regions are more conserved and to interpret if there is a mutation at well conserve region what the mutation’s effect might be.

For the second step, by using MEGA7 alignment, phylogenetic tree is constructed and Newick format of the phylogenetic tree is tried to be obtained. However, MEGA7 gave an error, therefore Simple Phylogeny tool which is offered by EMBL-EBI is used to get phylogenetic tree and Newick format of the tree. In order to be consistent about phylogenetic tree, by using Newick format, phylogenetic tree is plotted in Figtree.

At last, human TPD52 protein is chosen. By using CDART, similar domain architectures under TPD52 protein superfamily are searched. Proteins’ that are within this architecture, sequences are gathered and by using sequences their new phylogenetic tree is drawn. By using these phylogenetic trees and domain architecture, interpretations will be done about their homology. In addition, we used SMART and EggNOG tools to verify our results.

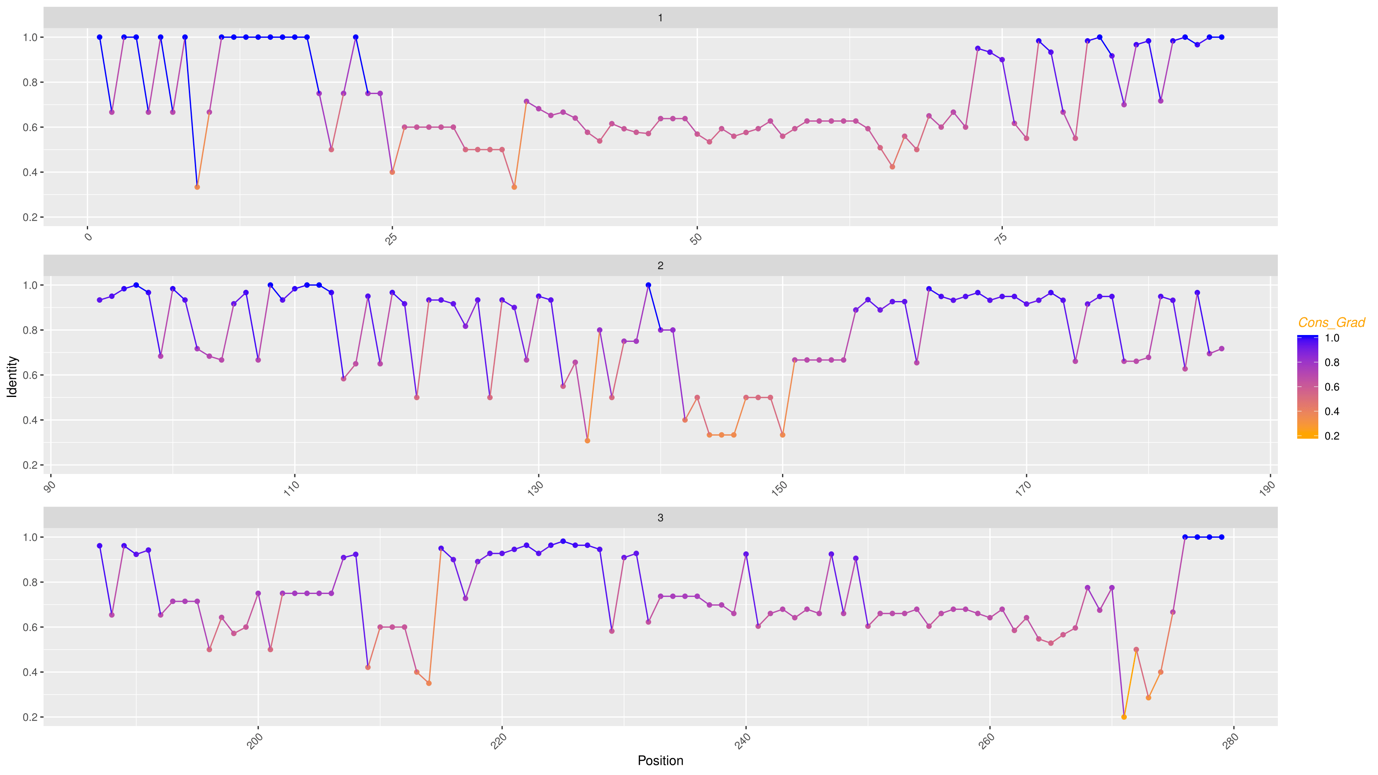
**Results**

TPD52 is tumour protein that placed in 8th chromosome in human. TPD52 has strong expression in liver, kidney and colon tissues and low-level expression of heart, lung and skeletal muscle. The studies which are related with D52 homologous proteins, indicate that D52 protein play a role on cell proliferation and calcium signalling. Studies reveal D52 proteins regulate their activities with D52-like proteins, work with them, this may conclude that D52 proteins have a potential effect on controlling cell division. In addition, D52 have some effects on B cell differentiation. During differentiation from B cells to plasma cells, TPD52 level is observed at maximum level. D52 protein is also thought to be as target gene that increasing copy number of 8th chromosome. As result TPD52 has significant relationship especially breast carcinoma as well as other type of carcinomas, it may it may contribute to tumour initiation and progression.

After searching expressions and functions of TPD52 firstly, TPD52 proteins from different organism are obtained and aligned. Aligned sequences are used to create phylogenetic tree. Aligned sequences is shown in Figure1.



Figure 1:Protein sequence alignment

Full version of alignment data is uploaded to GitHub. According to multiple sequence alignment, TPD52 proteins from different species are mostly conserved for many position. There are some gaps for few organisms however at most region amino acid’s properties remain similar even there is an amino acid change at specific position. To be able to interpret this alignment properly data consensus sequence is generated, then by using consensus sequence conservation score is calculated and plot of the conservation score (Figure 2) is drawn.

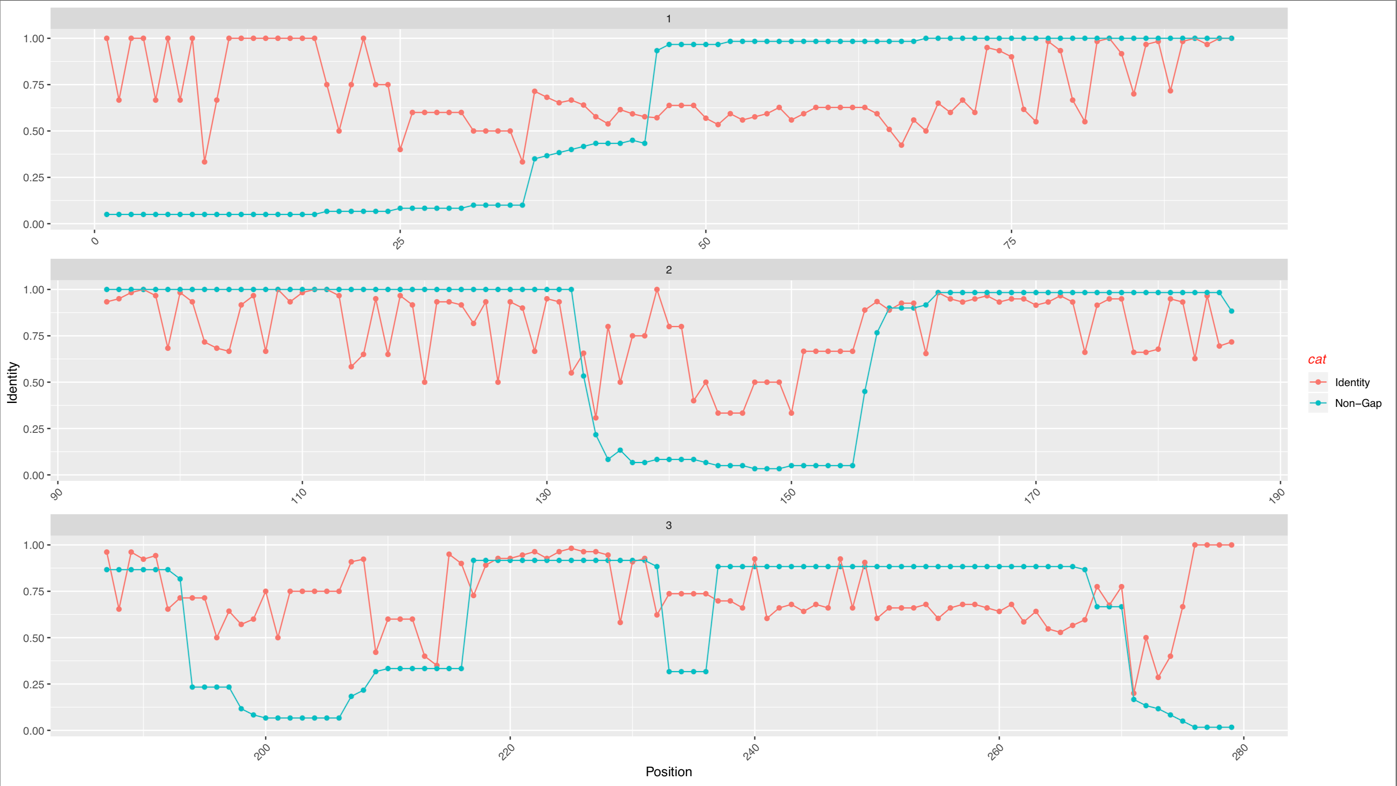
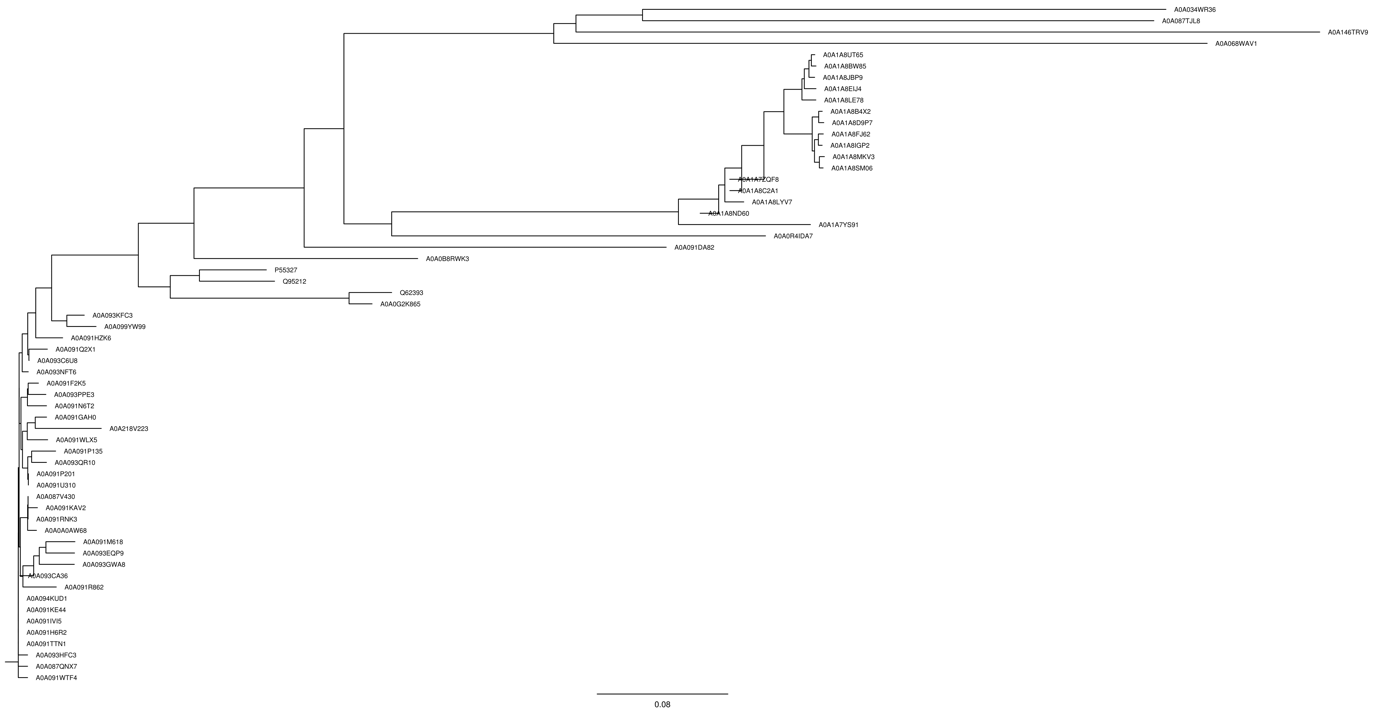


Figure 2:Conservation scores of each position

Figure 2 is also uploaded to GitHub. When the Figure 2 is examined, high conservation scores can be seen for most regions. There are some regions such as between position 10 and 20, conservation scores are 1. This result occurs because of amino acids in that region are presumably observed in a few organisms and therefore in the alignment there are many gaps. As a result, TPD52 protein change at that positions seems never changed in evolutionary scenario. However, this result is improper to interpret graph. That is why Figure 3 is plotted and it represents the number of gaps for each position. When number of gaps of positions that they have as conservation score 1, are examined, they are significantly high. On the other hand, there are also some regions that there are many changes such as amino acid substitution. For very conserved regions with significantly high conservation score, it can be said that if there is a mutation in that regions its effect is more significant than the regions that have lower conservation scores. Because when conservation score is low, many changes occurred evolutionary process without observing disease. As conclusion, if a mutation occurs at very conserved position, occurring a disease related with that mutation is probably observed on the phenotype of the organism.

For the next step, phylogenetic tree of these proteins is generated via both Simple Phylogeny and FigTree to pay attention to homogeny of these proteins. Figure 4 illustrates the phylogenetic tree which is created via FigTree. The list of organism names that have proteins shown in phylogenetic tree, is given under the tree and uploaded to GitHub.

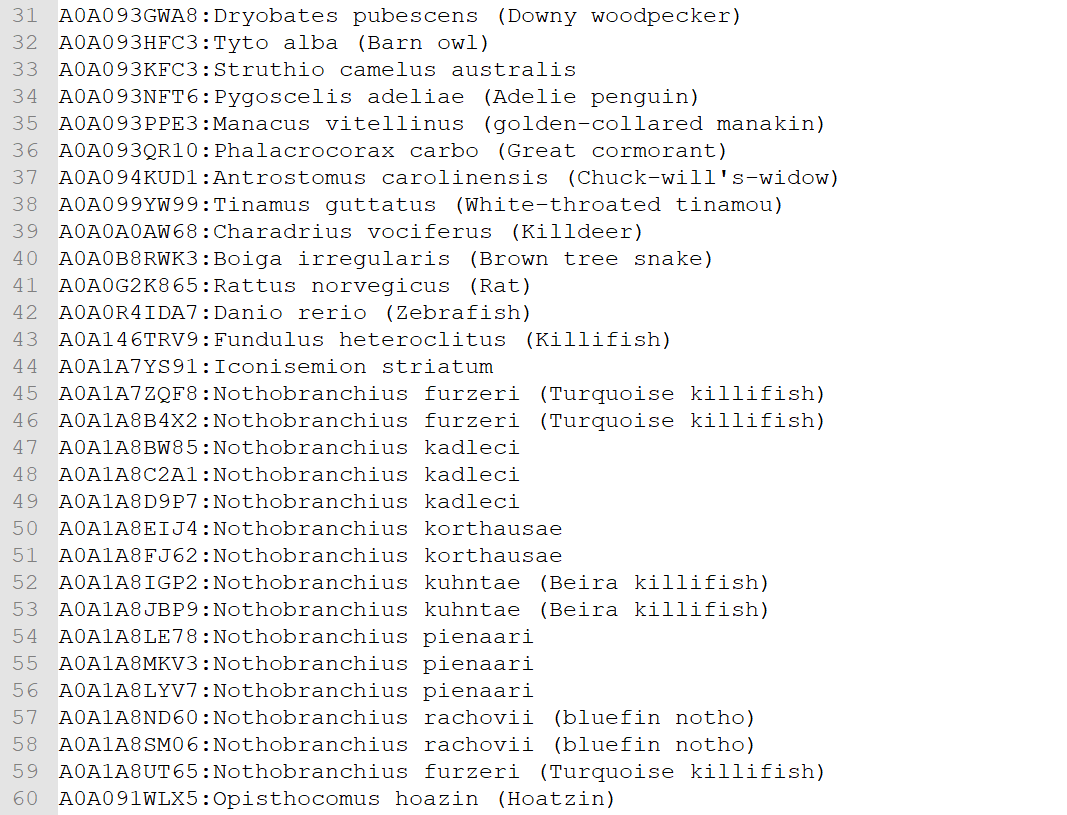


Figure 4: Phylogenetic tree of proteins

When the phylogenetic tree is examined, TPD52 protein seems to have some gene duplications and gene substitutions. For example, when Nothobranchius furzeri’s (Turquoise killifish) TPD52 proteins examined, phylogenetic tree gives the result that there is gene duplication at some point and these proteins are paralogous. Because, while A0A1A7ZQF8 protein is in a clade, A0A1A8UT65 and A0A1A8B4X2 proteins are in different clades, these proteins come from the same organism or common ancestor however at some point they gained different properties and clustered at different clades. When these proteins evolve and clustered different clades, one copy obtain a functionality while others obtain different functionalities. Additionally, if we look at the branches of the tree P55327 (Homo Sapiens-Human) and Q95212 (Ryctolagus cuniculus-Rabbit) are clustered at the same clade and these two species also clustered with Q62393 (Mus musculus-Mouse) and (Rattus norvegicus-Rat). These organisms’ proteins are evolved from a common ancestor. They might have similar functionalities beside of they might gain different functionalities. These specifications represent the orthologous proteins. These examinations and examples can be extended.

At last, domain architecture is examined and by using proteins’ that are within domain architecture a new phylogenetic tree is constructed with blast=100 (Figure 5).

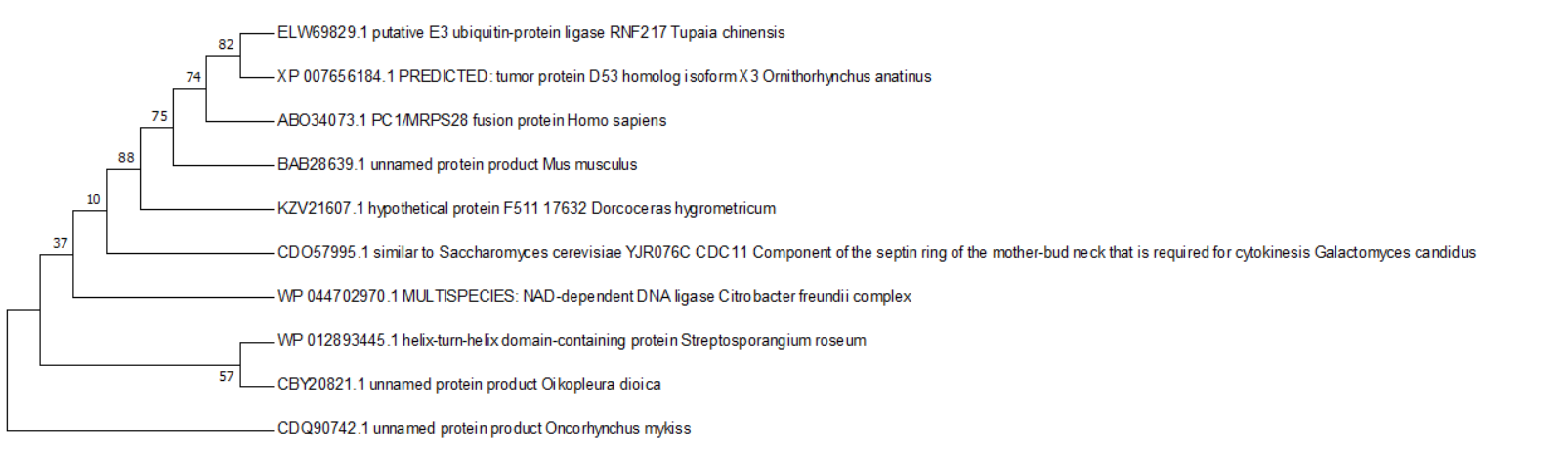


Figure 5: Proteins within domain architecture

This phylogenetic tree verifies the result of orthologous organisms in the multiple sequence alignment phylogenetic tree. For instance, as Figure 5 is examined, Homo Sapiens (Human) and Mus musculus (Mouse) are placed as they have orthologous proteins of TPD52. Then tree in Figure 5 is verified by using SMART and EggNOG. Figure 6 illustrates the part of tree which is generated by EggNOG, in that tree there are more organisms to compare each other and PFAM domain architecture can be seen beside the tree.

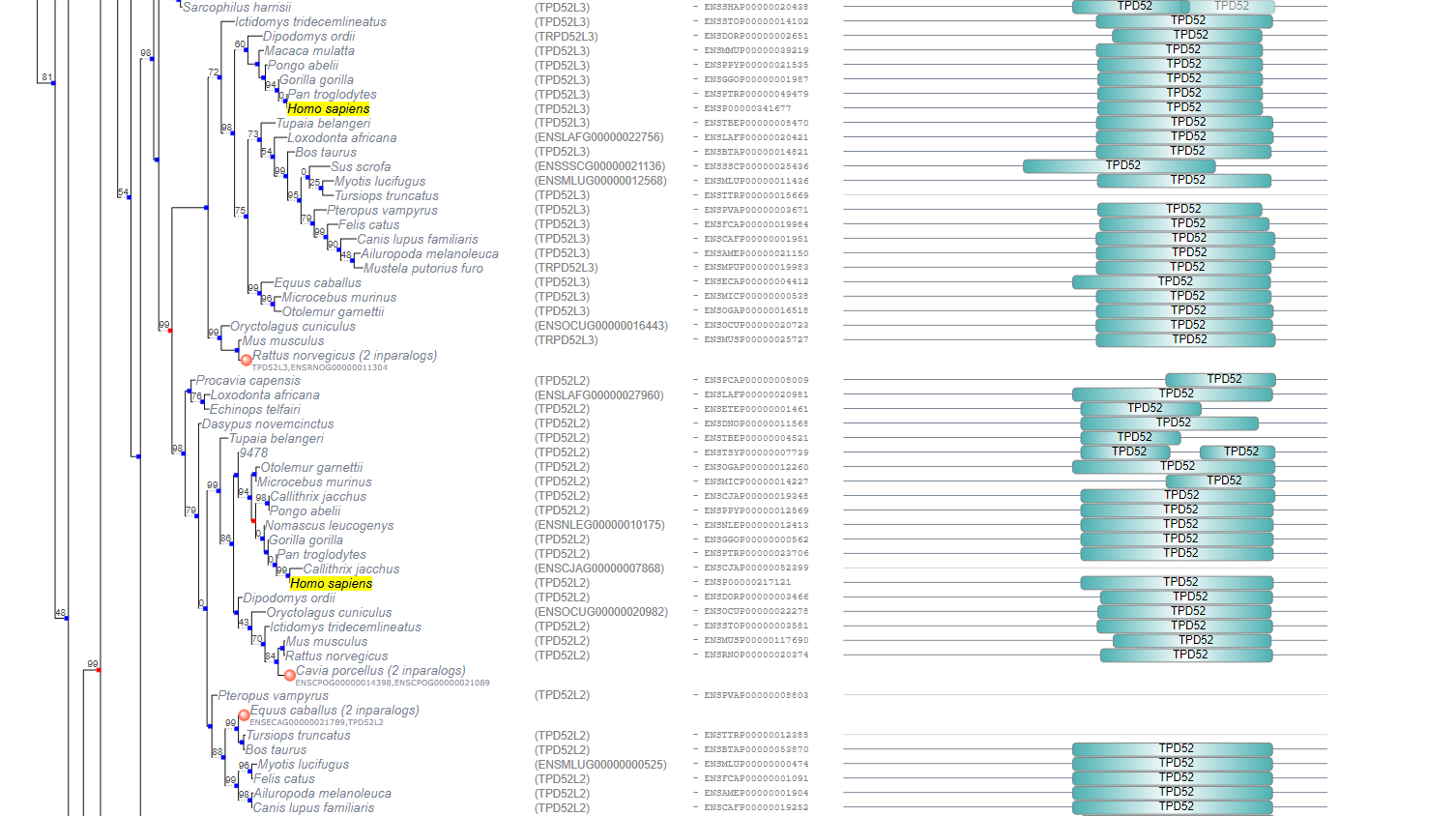


Figure 6: Domain alignment tree by EggNOG

Since there is no specific filtration in these tools number of organisms therefore homologous proteins are much more that we consider. When we look at Homo sapiens protein, phylogenetic tree is the evidence of gene duplication. In addition, while domain architectures are examined, it can be seen that, domains are very conserved, and these organisms have strong relations between each other. Figure 7 represents a closer look to domain architecture of these organisms. Figure 7 is also generated by EggNOG for aligned blocks.

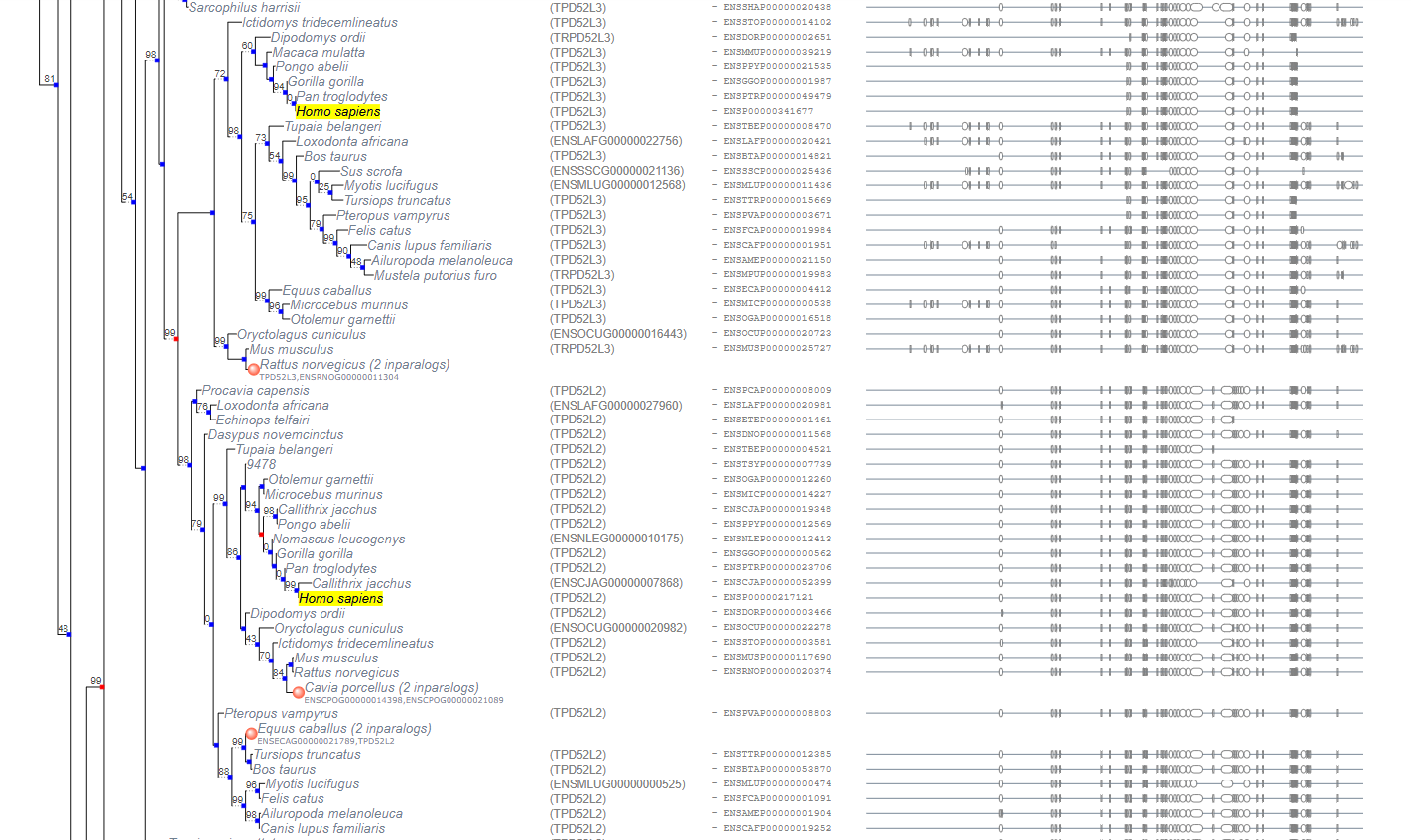


Figure 7: Closer look to aligned blocks

**Discussion**

Number of protein sequences which aligned can be increase by using phylogenetic tree of TPD52 protein. Superfamilies’ of TPD52 can be determine. Conservation positions and conservation amount between TPD52-53 can analyse and differences between those data can examine.

**References**

* European Bioinformatics Institute Protein Information Resource SIB Swiss Institute of Bioinformatics. (2018, December 05). Tumor protein D52. Retrieved from https://www.uniprot.org/uniprot/P55327#function
* Carugo, O., & Eisenhaber, F. (2016). Data Mining Techniques for the Life Sciences. Totowa: Humana Press.
* European Bioinformatics Institute Protein Information Resource SIB Swiss Institute of Bioinformatics. (n.d.). Tumor protein D52 (TPD52): A novel B-cell/plasma-cell molecule with unique expression pattern and Ca(2 )-dependent association with annexin VI. Retrieved from https://www.uniprot.org/citations/15576473
* European Bioinformatics Institute Protein Information Resource SIB Swiss Institute of Bioinformatics. (n.d.). Identification of homo- and heteromeric interactions between members of the breast carcinoma-associated D52 protein family using the yeast two-hybrid system. Retrieved from https://www.uniprot.org/citations/9484778
* (n.d.). Retrieved from http://atlasgeneticsoncology.org/Genes/GC\_TPD52.html
* What are protein families? (2017, March 27). Retrieved from https://www.ebi.ac.uk/training/online/course/introduction-protein-classification-ebi/protein-classification/what-are-protein-families
* Koonin, E. V. (1970, January 01). Evolutionary Concept in Genetics and Genomics. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK20255/
* Family: TPD52 (PF04201). (n.d.). Retrieved from https://pfam.xfam.org/family/PF04201#tabview=tab0
* The tumor protein D52 family: Many pieces, many puzzles. (2004, November 06). Retrieved from https://www.sciencedirect.com/science/article/pii/S0006291X0402409X
* EMBL-EBI, I. (n.d.). InterPro. Retrieved from http://www.ebi.ac.uk/interpro/entry/IPR007327/proteins-matched?start=20
* Gene Tree Image. (n.d.). Retrieved from http://www.ensembl.org/Multi/GeneTree/Image?gt=ENSGT00940000155294