Examination of TPD52 Family with respect to its Evolutionary History and Domains

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 family is searched on PFAM and InterPro, InterPro contains 2066 different sequences. Since 2066 sequences is too much to align and draw phylogenetic tree, data from EggNOG which contains 230 proteins from 85 different species are gathered. 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, functions, expression of protein and domain organisations were examined and D52 protein has chosen. After choosing D52 protein family, in order to find domains of D52 family 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, is used. InterPro contains 2066 protein sequences from

different organisms, this database includes homologous protein sequences of TPD52 such as TPD53 and TPD54. However, distinguish proteins out of 2066 proteins is hard, therefore another database EggNOG which based on orthology predictions and phylogenetic data, is used to obtain sequences. Protein sequences are obtained as FASTA format from EggNOG and saved in “ENOG4111M9H.fa” file. These sequences are aligned by using MEGA7 and saved in “ENOG4111M9H-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. Conservation scores have range between 0 and 1.

```python
import re

def fastareader(filename):
	seqDict = {}
	with open(filename, "r") as f:
		icerik = f.read()
		myDict = [elem.replace("\n", "").strip("\r")  for elem in re.split(r'>.*\n', icerik) if elem != ""]

	keys = [item.lstrip(">") for item in icerik.split("\n") if item.startswith('>')]

	seqDict = dict(zip(keys, myDict))
	return seqDict


def consensusWithIdentity(sequences):
	consensus_sequence = ''
	consensus_vals = {}
	for index, amino_acid in enumerate(sequences[0]):
		counter = {}
		for sequence in sequences:
			if index < len(sequence):
				if sequence[index] in counter.keys():
					counter[sequence[index]] += 1
				else:
					if sequence[index] != "-":
						counter[sequence[index]] = 1
		maxOne = max(counter, key=counter.get)
		#print(maxOne)

		if counter[maxOne] == 0:
			consensus_sequence += '-'
			consensus_vals[index] = 0
		else:
			consensus_sequence += maxOne
		#	print counter[maxOne] , sum(counter.values())
			consensus_vals[index] = round(float(counter[maxOne])/sum(counter.values()), 3)
		#	print (consensus_vals)
	return consensus_sequence , consensus_vals


def main():
	myDict = fastareader("TPD52-aligned.fa")
	consensus_seq, consensus_vals = consensusWithIdentity(list(myDict.values()))

	with open("identity.txt", "w") as f:
		for key in consensus_vals.keys():
			f.write("{}\t{}\t{}\n".format(key, consensus_seq[key], consensus_vals[key]))

	print("[*] Done")


if __name__ == "__main__":
	main()
```

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 Newick format of the tree is obtained from EggNOG. Then, phylogenetic tree is drawn by using Figtree. In order to be consistent about phylogenetic tree, phylogenetic tree from EggNOG's itself is also taken account.

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

In this study we focused on human TPD52 protein. TPD52 is tumour protein that placed in 8^{th} chromosome in human. It has high-level 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 (Lewis et al, 2007). Other 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 8^{th} 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.

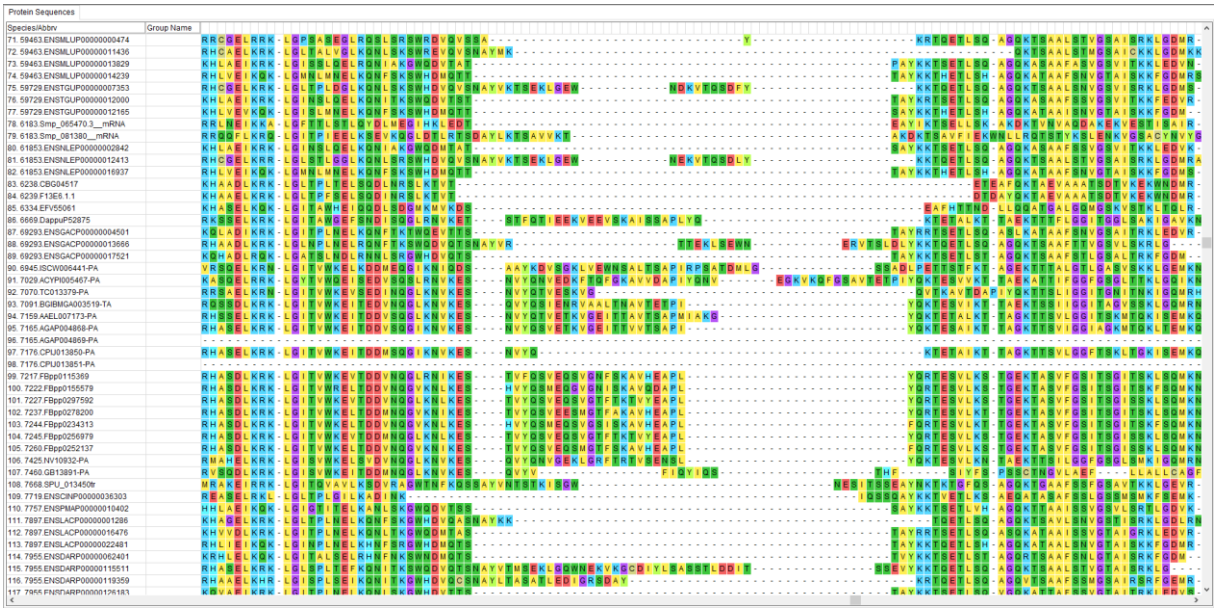

Figure 1:Protein sequence alignment

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.

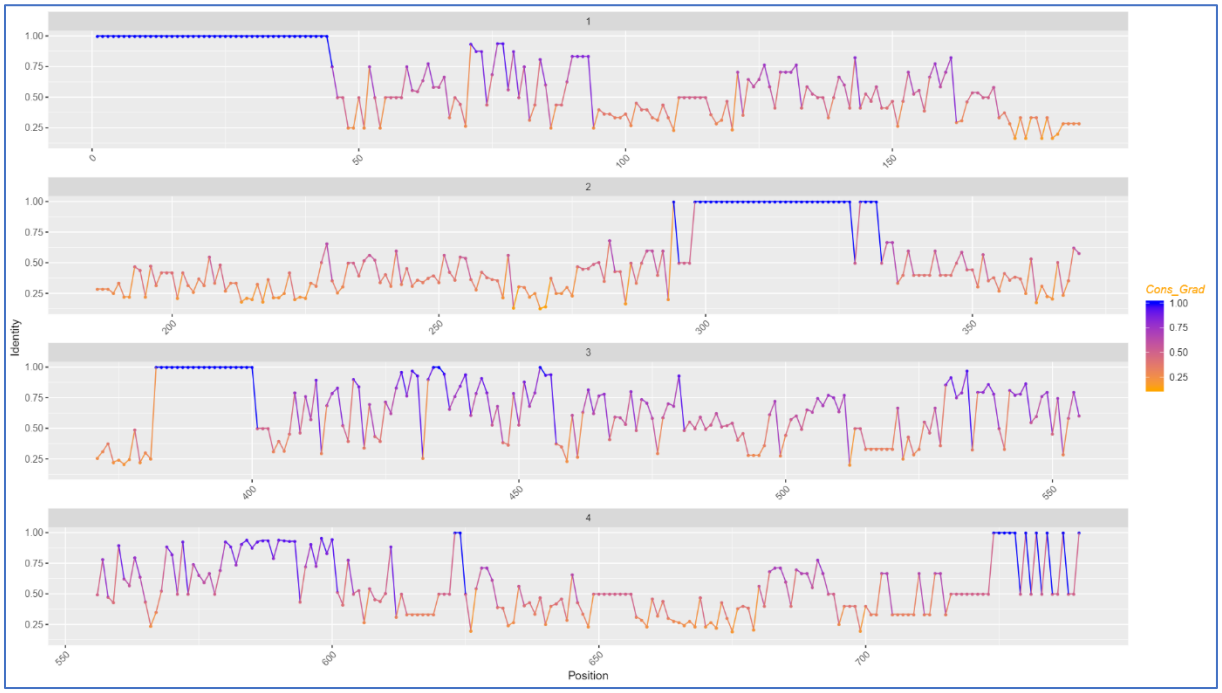

Figure 2:Conservation scores of each position

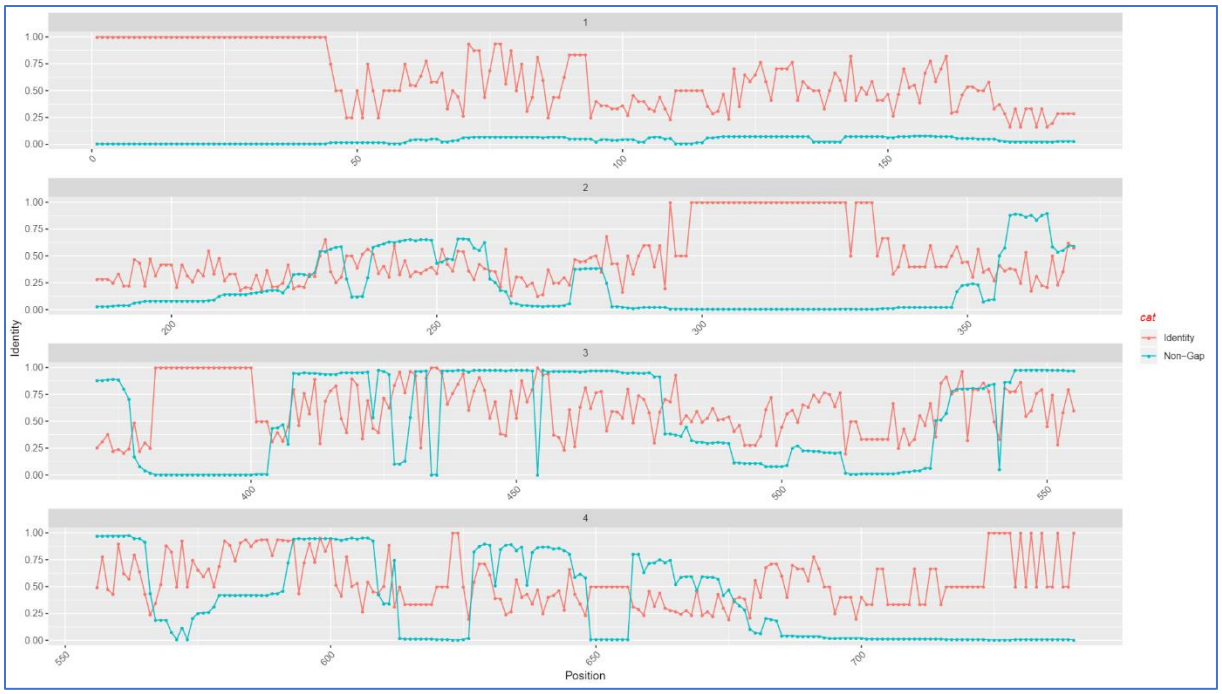

Figure 3:Conservation scores and non_gap region distribution

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 300 and 325, conservation scores are 1. As a result, TPD52 protein change at that positions seems never changed in evolutionary scenario. 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. 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 EggNOG 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.

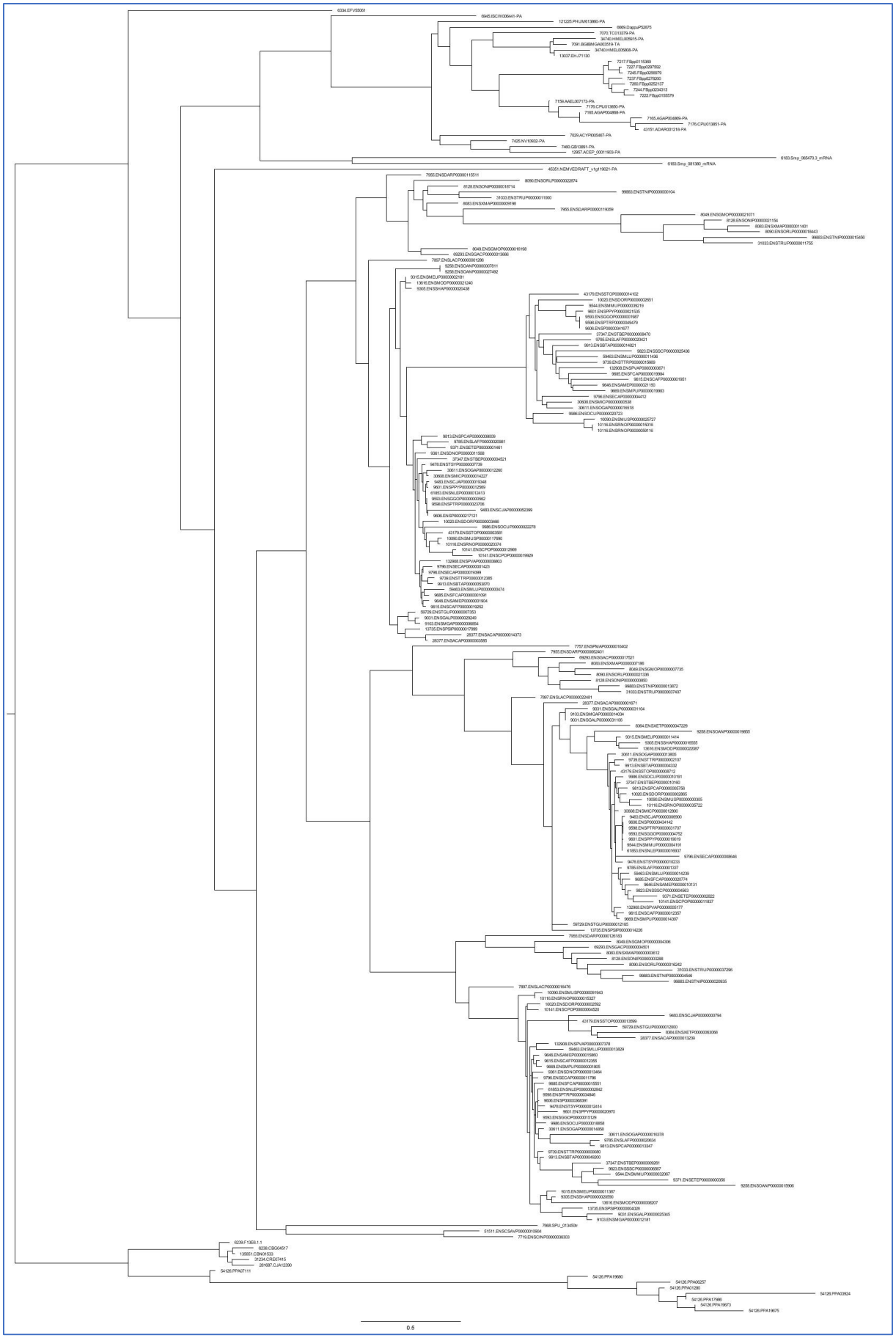

Figure 4: Phylogenetic tree of proteins

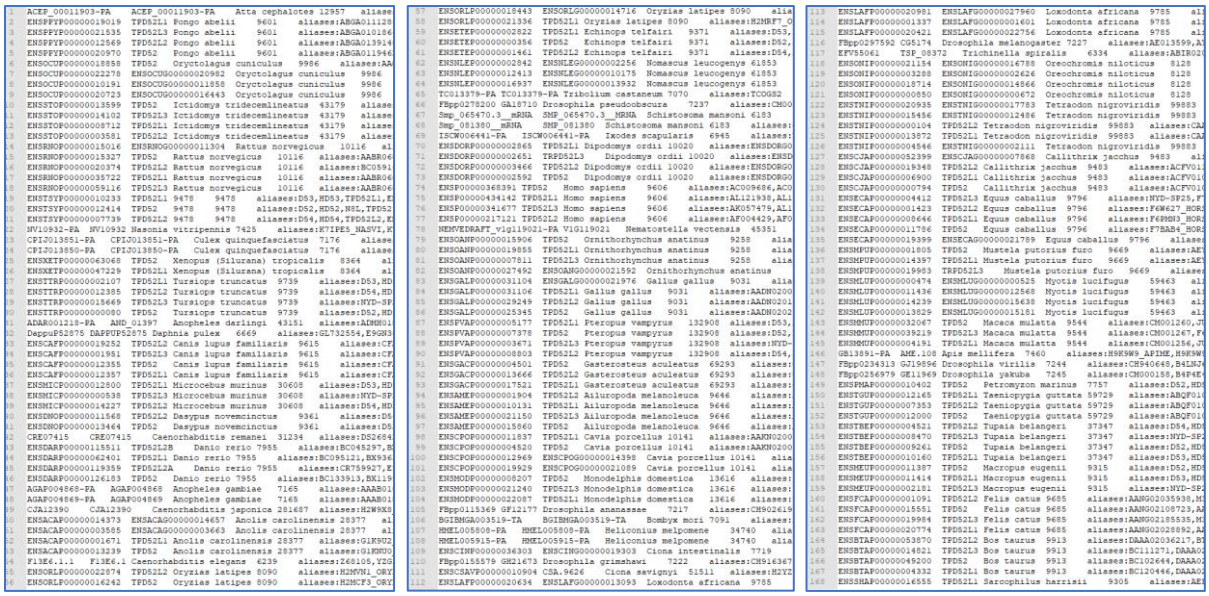

Figure 5: Name of the species

When the phylogenetic tree is examined, TPD52 protein seems to have some gene duplications and gene substitutions. For example, Homo sapiens's TPD52 gene examined, 4 different orthologous protein can be observed. While TPD52 and TPD52L1 genes are clustered at the same clade, TPD52L2 and TPD52L3 are clustered at the same clade. When common ancestors of TPD52L2 and TPD52L3 genes are examined, there is gene duplication at some point and these proteins are paralogous. In addition, TPD52 and TPD52L1 genes are paralogous pairs. These pairs of protein come from the same organism or a common ancestor, however at some point they gained different properties and clustered at different clades. These proteins are placed different clades because when these proteins evolve, one copy of the protein obtain a functionality while others obtain different functionalities. Example of paralogous genes can be extended, for instance Otolemur garnettii has gene duplication at the very bottom of the tree and gained different functions, separated as TPD52 and ENSOGAG00000034474. On the other hand, if we continue to examine TPD52 genes, TPD52 and one of from both TPD52L2 and TPD52L3 genes are orthologous pairs. These examples of pairs also share a common ancestor. However, in evolutionary scenario they might have similar functionalities beside of they might gain different functionalities. These genes are diverged after a speciation event. And if we look at ENSOGAG00000034474 (Otolemur garnettii) and ENSLAFG00000013093 (Loxodonta africana), these genes are also orthologous pairs. Speciation effect can be observed easily here, two different species have one of the homologous pairs of the gene.

At last, domain architecture of TPD52 human protein is examined. However, domain architecture itself cannot be explicated (Figure 5). Therefore, 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, also another example reveals that Ornithorhynchus anatinus(TPD52) share common region that is very similar to Homo Sapiens(TPD52) and they are orthologous proteins. 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.

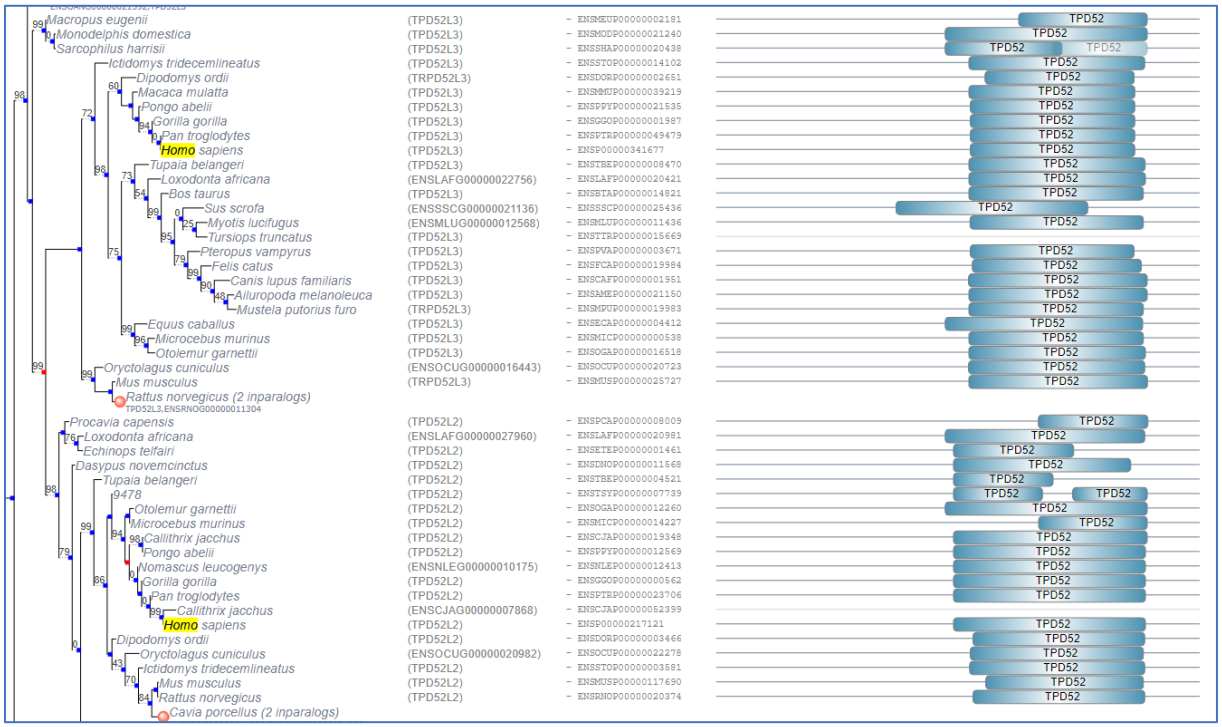

Figure 6: Domain alignment tree by EggNOG

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 and shows SMART domains of the species.

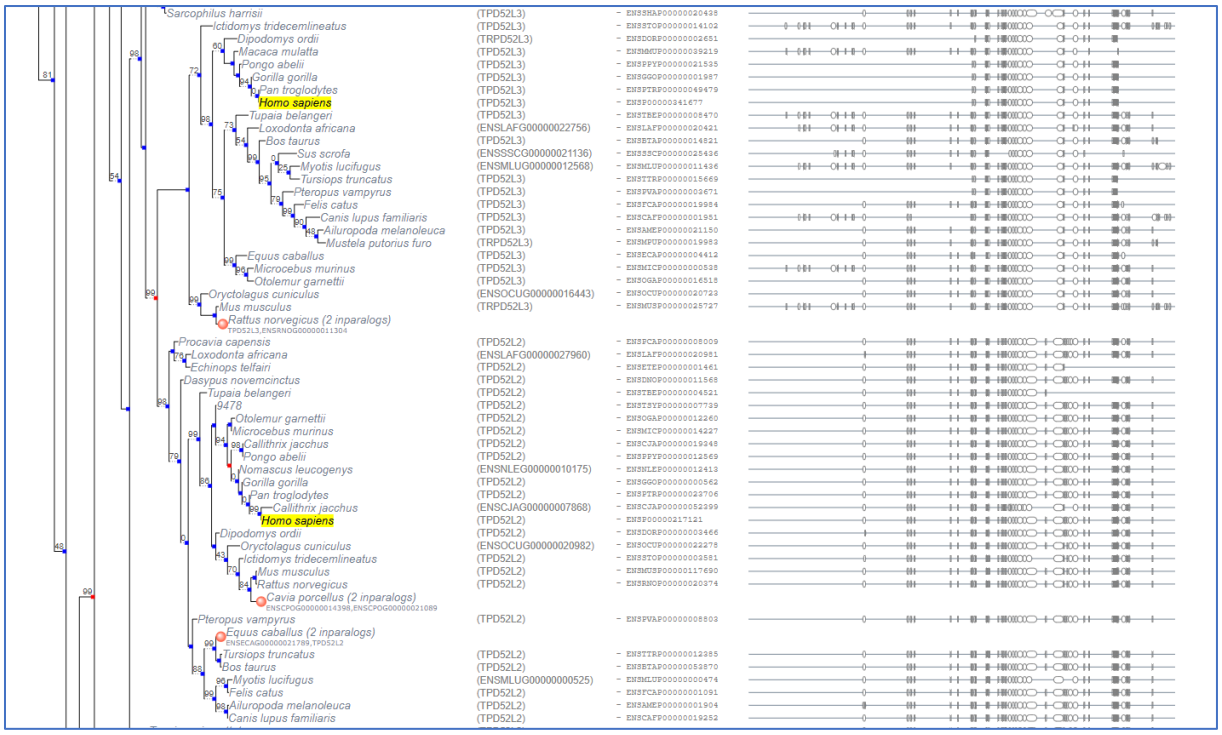

Figure 7: Closer look to aligned blocks

Discussion

This study focused on TPD52 protein in human while its domains were examined and interpreted. Other than domains, conservation scores were calculated and interpreted. Since there are lots of data, and very big phylogenetic tree, small proportion of phylogenetic tree is used while data was construing. In results part, homology of TPD52 protein was explained with few examples. For the future studies, big picture of phylogenetic tree can be studied and homology of the proteins can be proved by many examples.

At first study focused on only TPD52 proteins, then TPD52 family extended to TPD52-like and TPD53, TPD54 proteins. This project is planned to study with superfamily and subfamilies of TPD52 protein. However, we could not be able to distinguish subfamilies. Subfamilies are also can be subject of future studies. Which functions are conserved at which subfamilies and what is the reason of these specific conservations occurs in specific subfamilies subjects might be good for future studies.

Conservation values can be evaluated by not using consensus sequence but amino acid properties. Since some amino acids have similar properties such as positively charged groups or hydrophobic side chains, when there is an amino acid change, its effect might not be as harmful as amino acid property change.

References

- (n.d.). Retrieved from http://atlasgeneticsoncology.org/Genes/GC_TPD52.html
- A testis-specific and testis developmentally regulated tumor protein D52 (TPD52)-like protein TPD52L3/hD55 interacts with TPD52 family proteins. Cao Q, Chen J, Zhu L, Liu Y, Zhou Z, Sha J, Wang S, Li J. Biochem Biophys Res Commun. 2006 Jun 9;344(3):798-806. Epub 2006 Apr 19.
- Carugo, O., & Eisenhaber, F. (2016). Data Mining Techniques for the Life Sciences. Totowa: Humana Press.
- EMBL-EBI, I. (n.d.). InterPro. Retrieved from http://www.ebi.ac.uk/interpro/entry/IPR007327/proteins-matched?start=20
- 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
- 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
- Family: TPD52 (PF04201). (n.d.). Retrieved from https://pfam.xfam.org/family/PF04201#tabview=tab0
- Gene Tree Image. (n.d.). Retrieved from http://www.ensembl.org/Multi/GeneTree/Image?gt=ENSGT00940000155294
- Induction of tumorigenesis and metastasis by the murine orthologue of tumor protein D52. Lewis JD, Payton LA, Whitford JG, Byrne JA, Smith DI, Yang L, Bright RK. Mol Cancer Res. 2007 Feb;5(2):133-44.
- Isolation and characterization of a novel gene expressed in multiple cancers. Chen SL, Maroulakou IG, Green JE, Romano-Spica V, Modi W, Lautenberger J, Bhat NK. Oncogene. 1996 Feb 15;12(4):741-51.
- Koonin, E. V. (1970, January 01). Evolutionary Concept in Genetics and Genomics. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK20255/
- The tumor protein D52 family: Many pieces, many puzzles. (2004, November 06). Retrieved from https://www.sciencedirect.com/science/article/pii/S0006291X0402409X
- 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