Bioinformatics Project

Aim of this project: I have chosen the protein **INADL_HUMAN** (UniProt ID: Q8NI35) that is an initial PDZ domain containing protein. The aim of that project is to find similar proteins and find their PDZ domains with the help of the ones that we already know of.

Table of contents

Search in databases....2
Function....2
Domains....2
Publications....3
Novel Genetic Loci Identified for the Pathophysiology of Childhood Obesity in the Hispanic Population....3
The Multi-PDZ Domain Protein-1 (MUPP-1) Expression Regulates Cellular Levels of the PALS-1/PATJ Polarity Complex....3
NG2 Regulates Directional Migration of Oligodendrocyte Precursor Cells via Rho GTPases and Polarity Complex Proteins....3
A Structural Portrait of the PDZ Domain Family....3
A Tight Junction-Associated Merlin-Angiomotin Complex Mediates Merlin's Regulation of Mitogenic Signaling and Tumor Suppressive Functions....3
Structure....3
Swiss-Model....3
ModBase....3
Interactions....4
UniProt....4
BioGrid....4
String database....4
Another interesting fact....4
Finding sequences of domains....5
Aligning sequences....5
alignment....5
MSA Viewer....5
philogenetic tree....5
Compare domains (dotplot)....6
window size: 10, threshold: 23....6
window size: 30, threshold: 23....6
window size: 30, threshold: 50....6
window size: 40, threshold: 40....6
Finding similar proteins (BLAST)....7
Same domains as UniProtKB....7
Creating a database....7
HMM search....7
Creating profile HMM of PDZ domains....7
HMM search....7
Phylogenetic analysis....8
Finding animals....8
Downloaded both amino acid sequences and mRNA sequences....8
Multiple alignment....8
Created phylogenetic trees....8
Analysis....8

1. Search in databases: I searched in the following sites for the human INADL: UniProt[1], PubMed[2], RCSB PDB[3], and ExPaSy[4].
   a) Function: In the UniProt database I found the description of INADL gen's function: "*Scaffolding protein that may bring different proteins into adjacent positions at the cell membrane. May regulate protein targeting, cell polarity and integrity of tight junctions. May regulate the surface expression and/or function of ASIC3 in sensory neurons. May recruit ARHGEF18 to apical cell-cell boundaries*" [5]
   b) Domains: On the same page, I found the details of the domains of that gene. That gene has 11 domains, of which 3 has a known function: "*The L27 domain (also called Maguk recruitment domain) is required for interaction with MPP5 and CRB3, and MPP5 localization to tight junctions. The PDZ domain 6 mediates interaction with the C-terminus of TJP3 and is crucial for localization to the tight junctions. The PDZ domain 8 interacts with CLDN1 but is not required for proper localization.*" [6]

Domains and Repeats

Feature key	Position(s)	Length	Description	Graphical view
Domain	1 – 65	65	L27 PROSITE-ProRule annotation	
Domain	134 – 221	88	PDZ 1 PROSITE-ProRule annotation	
Domain	248 – 328	81	PDZ 2 PROSITE-ProRule annotation	
Domain	365 – 453	89	PDZ 3 PROSITE-ProRule annotation	
Domain	553 – 639	87	PDZ 4 PROSITE-ProRule annotation	
Domain	686 – 772	87	PDZ 5 PROSITE-ProRule annotation	
Domain	1068 – 1160	93	PDZ 6 PROSITE-ProRule annotation	
Domain	1239 – 1322	84	PDZ 7 PROSITE-ProRule annotation	
Domain	1437 – 1520	84	PDZ 8 PROSITE-ProRule annotation	
Domain	1533 – 1615	83	PDZ 9 PROSITE-ProRule annotation	
Domain	1676 – 1762	87	PDZ 10 PROSITE-ProRule annotation	

1 http://www.uniprot.org/uniprot/
2 https://www.ncbi.nlm.nih.gov/pubmed
3 http://www.rcsb.org/pdb/home/home.do
4 http://www.expasy.org/
5 http://www.uniprot.org/uniprot/Q8NI35
6 http://www.uniprot.org/uniprot/Q8NI35

c) Publications: On the page PubMed, there was 50 result for the search of “INADL”. Then I narrowed the search for human genome and for publications not older than 10 years, and I got only 19 publications. I sorted the results by relevance and chose the following papers:

i. Comuzzie, Anthony G., Shelley A. Cole, Sandra L. Laston, V. Saroja Voruganti, Karin Haack, Richard A. Gibbs, and Nancy F. Butte. "Novel Genetic Loci Identified for the Pathophysiology of Childhood Obesity in the Hispanic Population." *PLoS ONE* 7, no. 12 (2012). doi:10.1371/journal.pone.0051954. https://www.ncbi.nlm.nih.gov/pubmed/23251661
   - That article is cited by 66 publications (among the articles of PubMed) and mentions the INADL in its abstract.

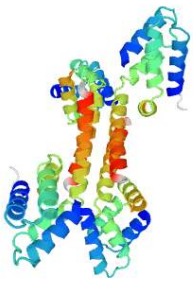
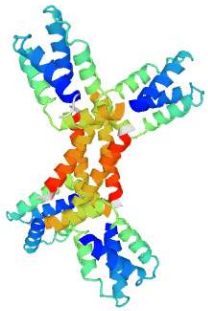
ii. Assémat, Emeline, Emmanuelle Crost, Marion Ponserre, Jan Wijnholds, Andre Le Bivic, and Dominique Massey-Harroche. "The Multi-PDZ Domain Protein-1 (MUPP-1) Expression Regulates Cellular Levels of the PALS-1/PATJ Polarity Complex." *Experimental Cell Research* 319, no. 17 (2013): 2514-525. doi:10.1016/j.yexcr.2013.07.011. https://www.ncbi.nlm.nih.gov/pubmed/23880463
   - That article has the “tight junction” keyword, the function of the INADL gene might be related with the integrity of tight junction.

iii. Biname, F., D. Sakry, L. Dimou, V. Jolivel, and J. Trotter. "NG2 Regulates Directional Migration of Oligodendrocyte Precursor Cells via Rho GTPases and Polarity Complex Proteins." *Journal of Neuroscience* 33, no. 26 (2013): 10858-0874. doi:10.1523/jneurosci.5010-12.2013. https://www.ncbi.nlm.nih.gov/pubmed/23804106
   - That publication also cited multiple times (17) and might give larger overview about how that protein influences the neural system on cellular level.

iv. Ernst, Andreas, Brent A. Appleton, Ylva Ivarsson, Yingnan Zhang, David Gfeller, Christian Wiesmann, and Sachdev S. Sidhu. "A Structural Portrait of the PDZ Domain Family." *Journal of Molecular Biology* 426, no. 21 (2014): 3509-519. doi:10.1016/j.jmb.2014.08.012. https://www.ncbi.nlm.nih.gov/pubmed/25158098
   - That publication describes the structure of the protein and role of the different domains according to the abstract.

v. Yi, Chunling, Scott Troutman, Daniela Fera, Anat Stemmer-Rachamimov, Jacqueline L. Avila, Neepa Christian, Nathalie Luna Persson, Akihiko Shimono, David W. Speicher, Ronen Marmorstein, Lars Holmgren, and Joseph L. Kissil. "A Tight Junction-Associated Merlin-Angiomotin Complex Mediates Merlin's Regulation of Mitogenic Signaling and Tumor Suppressive Functions." *Cancer Cell*19, no. 4 (2011): 527-40. doi:10.1016/j.ccr.2011.02.017. https://www.ncbi.nlm.nih.gov/pubmed/21481793
   - That article is cited 59 times and is interesting because it is about what are the effects if that gene is missing or malfunctioning.

d) Structure: On the Swiss-Model site of ExPASy, I found a nice interactive 3D model[7] and I took 3 “snapshots” from 3 different angles:

Also, on the page ModBase, I found another structure that is considered to be reliable:[8]

7https://swissmodel.expasy.org/repository/uniprot//Q8NI35
8https://modbase.compbio.ucsf.edu/modbase-cgi/model_details.cgi?searchmode=default&displaymode=moddetail&seq_id=795ca6555908cd1ccc8f8708c877f19bMPENMTAD&model_id=ff9d1089e6fea373afd940004f592188&queryfile=1479603265_748

e) Interactions: I found information about the interaction of human INADL gene in the UniProt database, and then I followed the links listed there and found 3 other interesting databases.

i. UniProt[9]: Found subunit and binary interactions with other genes.

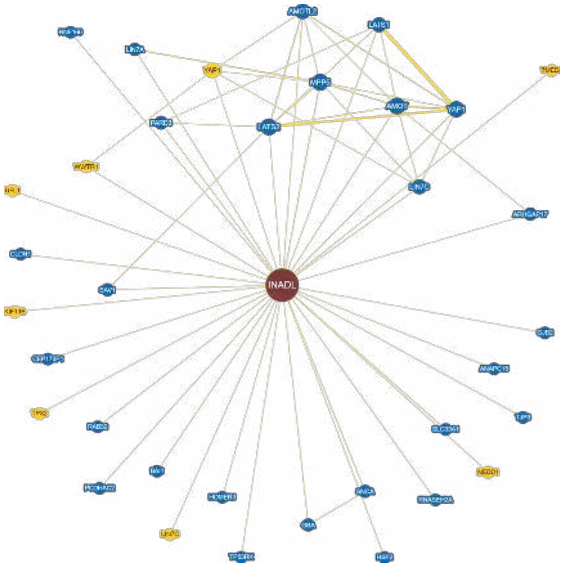
Subunit structure

Interacts with ASIC3, KCNJ10, KCNJ15, GRIN2A, GRIN2B, GRIN2C, GRIN2D, NLGN2, MPP7, HTR2A and SLC6A4 (By similarity). Forms a ternary complex with MPP5, CRB1 and CRB3. Interacts with TJP3/ZO-3 and CLDN1/claudin-1. Component of a complex whose core is composed of ARHGAP17, AMOT, MPP5/PALS1, INADL/PATJ and PARD3/PAR3. Directly interacts with HTR4 (By similarity). Interacts (via PDZ domain 8) with WWC1 (via the ADDV motif). Interacts (via C-terminus) with ARHGEF18 (PubMed:22006950). By similarity 6 Publications

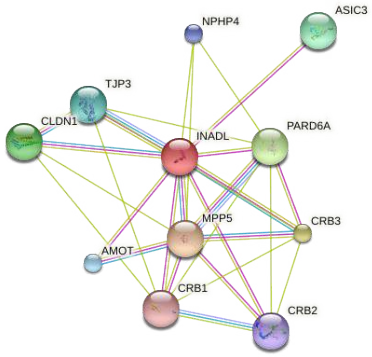
Binary interactions

With	Entry	#Exp.	IntAct	Notes
CRB1	P82279	2	EBI-724390,EBI-1048648	
MPP5	Q8N3R9	4	EBI-724390,EBI-2513978	
NF2	P35240	2	EBI-724390,EBI-1014472	

ii. BioGrid[10]: 44 physical interaction is found, of which 34 have a high throughput.



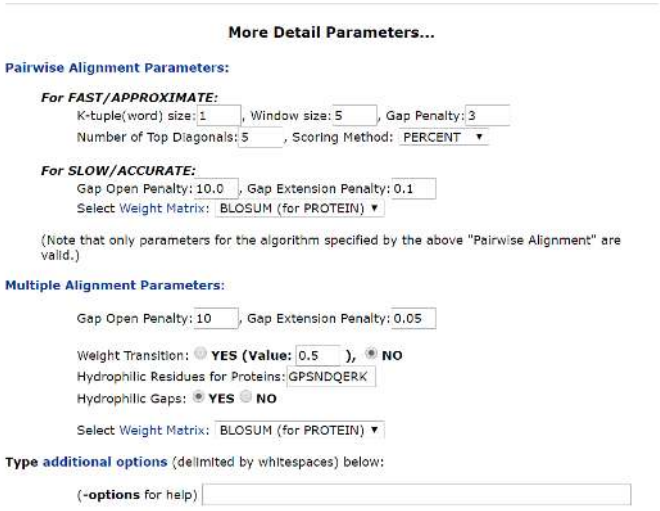
iii. String database[11]: Found fewer interactions, but has a nice visualization of type and strength of connections.



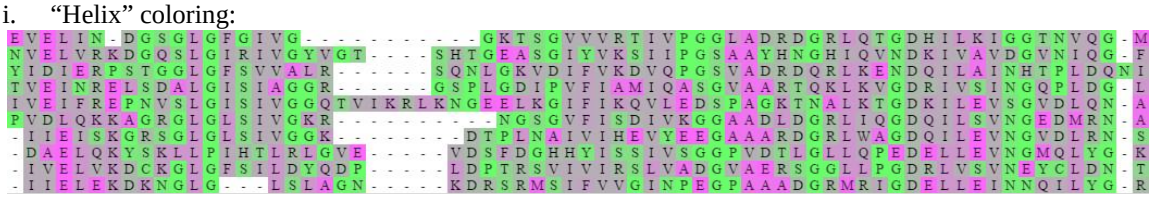
f) Another interesting fact: Siblings of that gene is found in every species at metazia level according to OrthoDB by Université de Genève[12] and phylogenetic tree can be drawn up to the bilateral animals shown by Ensembl[13].

9 http://www.uniprot.org/uniprot/Q8NI35#interaction
10 https://thebiogrid.org/115502
11 http://string-db.org/cgi/network.pl?taskId=iOj1nAjIlfUQ
12 http://www.orthodb.org/?query=INADL&universal=1&level=33208&species=33208
13 http://www.ensembl.org/Multi/GeneTree/Image?gt=ENSGT00760000119017

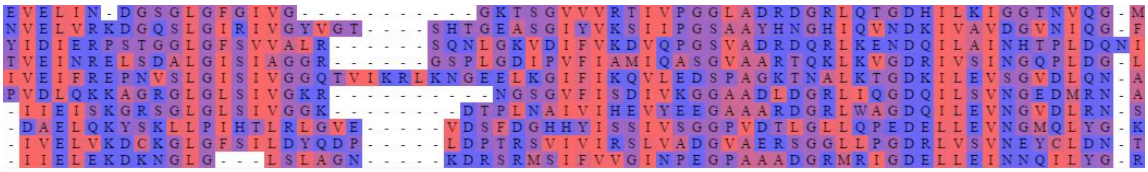
2) Finding sequences of domains: I saved the entire sequence of INADL gene to the file "2.Q8NI35.fasta" from the UniProtKB database[14]. Then I copied the sequences of domains also from UniProtKB (from "Family & Domain" section) and saved to the file "2.domains.fasta".
3) Aligning sequences: I made alignment with ClustalW[15] and created visualization by MSA Viewer[16].
   a) The alignment is done in slower and more accurate mode and by default parameters. Result of that alignment is saved to "3.clustalW.aln" and "3.clustalW.dnd" files.



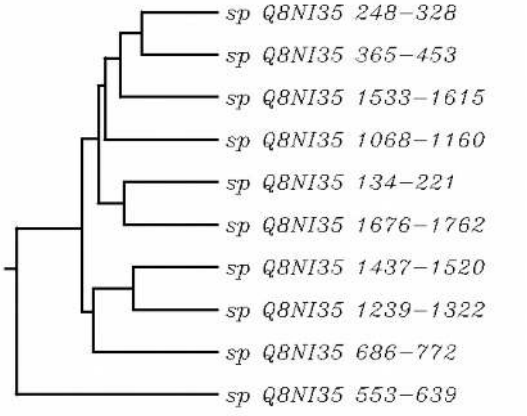
   b) In MSA Viewer I found 2 interesting coloring. These coloring shows that despite of differences the main characteristics of domains are similar. Moreover, these pictures show that PDZ 6 domain (5th row) has a sequence insertion between the 20th and 30th amino acids, and PDZ 1 domain (3rd row) has 1 acid insertion at the end of the sequence.
      i. "Helix" coloring:



      ii. "Hydrophobicity" coloring:

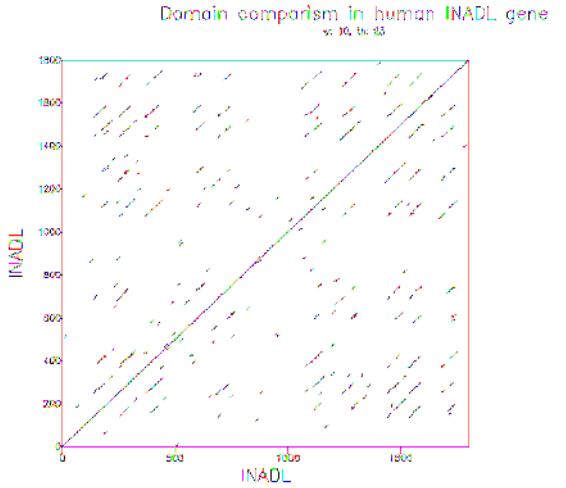


   c) I also generated philogenetic tree from ClustalW based on identity percentage by clicking on "select tree menu" on the top of the result page and choosing "Rooted philogenetic tree with branch length (UPGMA). That tree tells us that the PDZ 2 (248-328) and PDZ 3 (365-453), PDZ 7 (1239-1322) and PDZ 8 (1427-1520) domains are neighbors in the sequence and are very similar, so they might be result of a recent gene duplication. Also, other similarities can be product of older gene duplication. Only the PDZ 4 (553-639) domain is slightly different from others.

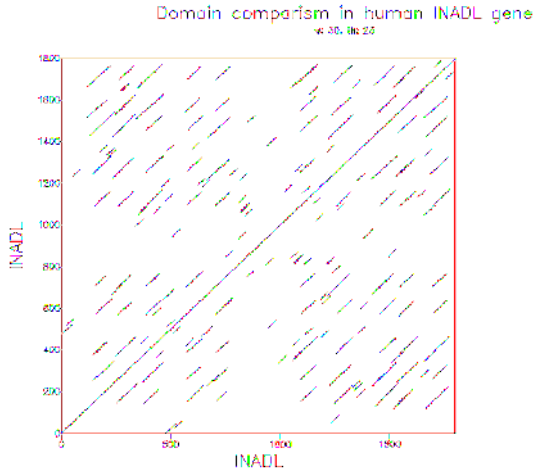


14 http://www.uniprot.org/uniprot/Q8NI35.fasta
15 http://www.genome.jp/tools/clustalw/
16 http://msa.biojs.net/app/

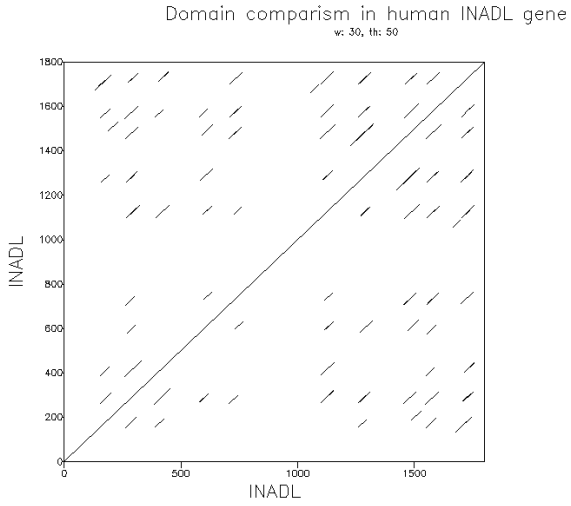
4) Compare domains (dotplot): I used the EMBOSS site[17] at bioinformatics.nl to generate dotplot. I pasted the whole sequence into both input sections, and created dotplots with different parameters (window size and threshold). I expected to get a dotplot that have a continuous, straight line along the main diagonal and that the plot is symmetric along that line. Furthermore, from the previous tasks I've already known that there is 10 similar domains in the gene, so I expected to have a grid-like arrangement of lines, where each row and column have approximately 10 small lines (including the line of main diagonal).
   a) First, I tried with default parameters: window size: 10, threshold: 23
      - That picture satisfies my expectation in that it has a straight line in the main diagonal and is symmetric, but the inner structure of the gene is rather noisy than meaningful.
   b) Second, I increased the window size because I knew that the length of domains is around 85 amino acids, so I expected longer lines: window size: 30, threshold: 23
      - That picture is much better, but still very noisy.
   c) Then, I increased threshold to filter noise: window size: 30, threshold: 50
      - Big improvement again, now the domains are recognizable, so now I just need to tune parameters to get the best result.
   d) I tried with multiple parameters, and finally found that the following parameters give the best plot: window size: 40, threshold: 40
      - I could even verify that this is a correct plot because it has approximately 80-90 amino acid long lines, the grid is almost complete, and the inter-domain region is shown between 800 and 1000.



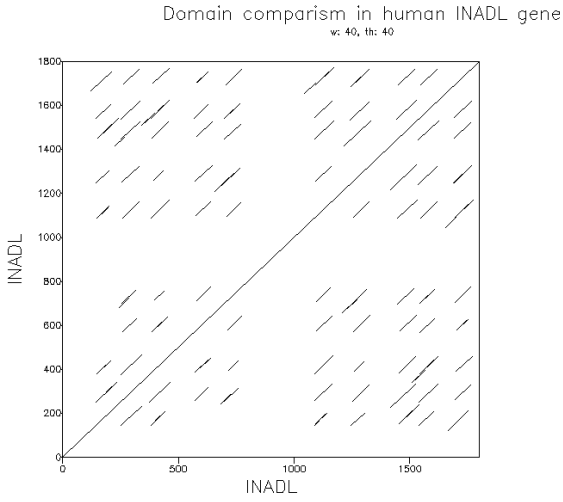
1. window size: 10, threshold: 23



2. window size: 30, threshold: 23



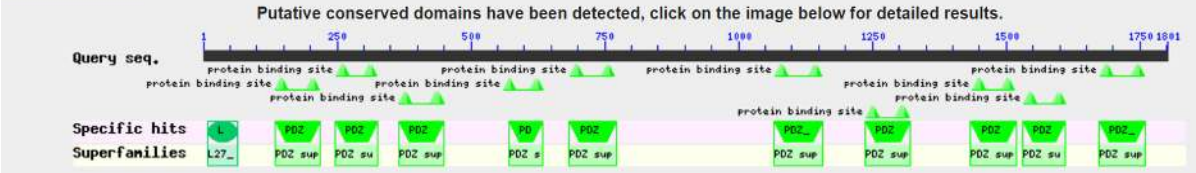
3. window size: 30, threshold: 50



4. window size: 40, threshold: 40

17 http://www.bioinformatics.nl/cgi-bin/emboss/dotmatcher

5) Finding similar proteins (BLAST): I used the NCBI BLAST[18] for creating a database of similar proteins. I had a protein and was looking for proteins, so I clicked on protein BLAST. Then I selected the "Q8NI35.fasta" file from my computer as query sequence, set the database to "Non-redundant protein sequences (nr)", selected "Homo sapiens" as organism, and used blastp algoritm.
   a) First, I checked whether it found the same domains as UniProtKB. And both of them found 11 domains, of which 10 is PDZ domain and the first is a L27 domain.

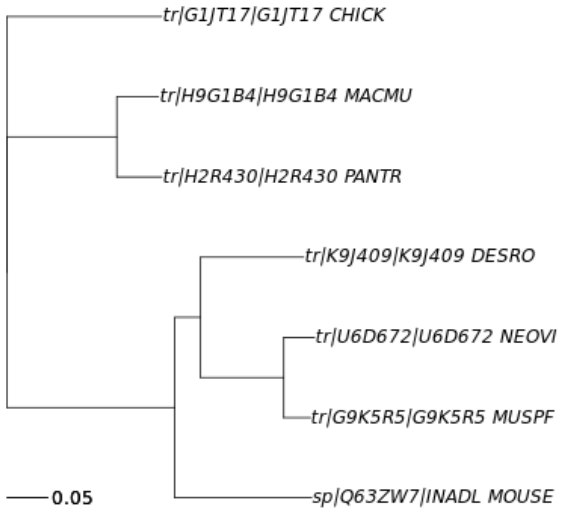


   b) Then, I have chosen the 15 most similar isoforms of that gene and created a database by saving it in different formats to the files "5.database.fasta" and "5.database.gb".

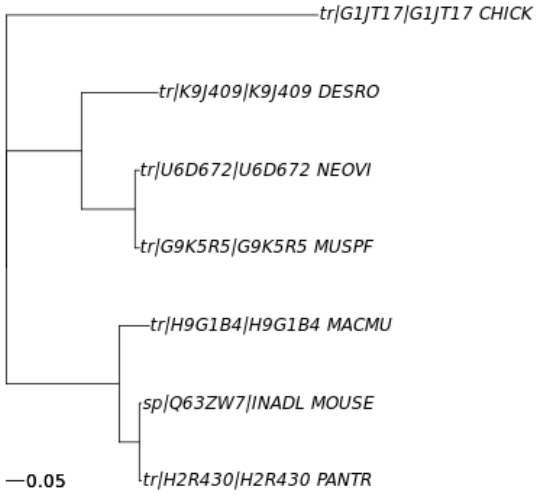
6) HMM search: My gene has 10 domains, so I only replaced one of the domains with one from the database.
   a) Creating profile HMM of PDZ domains:
      i. I looked up the gene "PREDICTED: inaD-like protein isoform X15 [Homo sapiens]" (that gene is the most different from my gene that is annotated) in the file "5.database.gb" (starting from line no. 2194), and cut out the region 135-218 (first PDZ domain) and formatted to fasta format using Notepad++ (I used the character counter, regex replacer, and lower-to-upper-case features for removing spaces and newline characters and finding the right region).
      ii. Then I replaced the PDZ 10 domain in the file "2.domains.fasta" and saved it as "6.database.fasta".
      iii. Installed the package clustalw (under the Linux on my computer) and aligned sequences in "6.database.fasta" to the files "6.database.aln" and "6.database.dnd" using the slow, but accurate method.
      iv. I converted the alignment file to stockholm format by a converter at bucago.com[19] and saved it to the file ("6.alignment.stockholm").
      v. I installed the package "hmmer" under Linux on my computer and built a hmm profile ("6.myhmm.hmm") using the command "hmmbuild 6.myhmm.hmm 6.alignment.stockholm > 6.hmm_profiler_output.txt".
   b) HMM search:
      i. I ran a search on my domains using the following command: "hmmsearch -o 6.result.txt 6.myhmm.hmm 6.database.fasta".
      ii. From the file 6.result.txt, I could read out that the PDZ sequence in found in every input sequences; however, in the isoforms X10-15 less than 10 matching domains is found and in the isoforms X1-9 more than 10 matches found that is surprising because these sequences have only 10 PDZ domains.
      iii. Though, the E-values of hits are very low (for the isoform X1-5 it is even equal to 0 on the full sequence meaning that (almost) total match is found), I set a threshold for E-value at 1E-200, and ran the search again (saving the result to "6.result2.txt"). As a result, I sorted out the hits with higher E-value than 1E-200, meaning that isoforms X11-15 in not listed in that file.
      iv. Comparing to the results of NCBI BLAST, the total and best scores are much lower (it is in the range of 3000-4000 in the results of BLAST, but is under 1000 in my results). Similarily, the E-value is different: in the BLAST search, all of the sequences have 0 E-value, but in my search it is slightly higher.

18 https://blast.ncbi.nlm.nih.gov/Blast.cgi
19 http://sequenceconversion.bugaco.com/converter/biology/sequences/index.html

7) Phylogenetic analysis:
   a) I have chosen the UniProt BLAST to find animals because its searching tool allows to filter by classes of animals and also have links to mRNA sequence if it is known. First I ran a BLAST for the query "INADL", then when I had results I filtered them. First I set the dropdown menu on the left to filter "Mammalia". I opened many genes and kept those, which had links to mRNA databases. Then I repeated it for "Neognathae" as well. At the end, I found these animals:
      i. K9J409_DESRO - Desmodus rotundus (Vampire bat) [20]
      ii. INADL_MOUSE - Mus musculus (Mouse) [21]
      iii. U6D672_NEOVI - Neovison vison (American mink) (Mustela vison) [22]
      iv. G9K5R5_MUSPF - Mustela putorius furo (European domestic ferret) (Mustela furo) [23]
      v. H2R430_PANTR - Pan troglodytes (Chimpanzee) [24]
      vi. H9G1B4_MACMU - Macaca mulatta (Rhesus macaque) [25]
      vii. G1JT17_CHICK - Gallus gallus (Chicken) [26]
   b) I downloaded both amino acid sequences and mRNA sequences (all of the mRNA sequence links from UniProt pointed to MBL-EBI[27], so I downloaded mRNA sequences from there). Then I concatenated the files using the 'cat' command under Linux ("7.animals.fasta" and "7.mRNA.animals.fasta").
   c) I made multiple alignment on both amino acid and mRNA sequences using clustalw program installed on my computer. ("7.animals.aln" and "7.mRNA.animals.aln").
   d) Finally, I created phylogenetic trees based on the alignments for both protein and mRNA sequences running the commands of "7.phylogenetic_tree.R" in R.



Phylogenetic tree of protein sequences



Phylogenetic tree of mRNA sequences

   e) Analysis: These plots are good examples for that phylogenetic trees are not necessarily equal because for instance it states correctly that American mink European domestic ferret are closely related species, but there is a difference between the two plots on that Mouse is related to Chimpanzee and Rhesus macaque, or not.

20 http://www.uniprot.org/uniprot/K9J409
21 http://www.uniprot.org/uniprot/Q63ZW7
22 http://www.uniprot.org/uniprot/U6D672
23 http://www.uniprot.org/uniprot/G9K5R5
24 http://www.uniprot.org/uniprot/H2R430
25 http://www.uniprot.org/uniprot/H9G1B4
26 http://www.uniprot.org/uniprot/G1JT17
27 http://www.ebi.ac.uk/