2016.11.16. Halász Viktor Olivér
B2SLFN

Introduction to bioinformatics – Project

Chosen protein: NHRF3_HUMAN

(Alternately Na(+)/H(+) exchange regulatory cofactor NHE-RF3)

1.

I searched the UniprotKB Database (http://www.uniprot.org) for information about the protein.

Function: "A scaffold protein that connects plasma membrane proteins and regulatory components, regulating their surface expression in epithelial cells apical domains."
Also plays role in multidrug resistance, maintaining the normal plasma cholesterol levels, localization and function of chloride-anion exchange. In general, proteins containing PDZ domains are highly important in anchoring receptor proteins in the membrane to cytoskeletal components. The protein is encoded by the PDZK1 gene.

According to Uniprot, the protein contains 4 domains: PDZ 1, 2, 3 and 4.

On Uniprot, there are 57 publications mentioning this protein. The first publication seems to be informative based on its title:
"Identification and partial characterization of PDZK1: a novel protein containing PDZ interaction domains." by Kocher O., Comella N., Tognazzi K., Brown L.F.

Under the interaction title you can find that the protein interacts with several other ones: *PDZK1IP1, ABCC2, AKAP2, BCR, CFTR, SLC22A12, SLC22A4, SLC22A5, SLC9A3R2* and *SLC17A1* are among them.

The secondary structure is determined but contains few orderly sections.

2.-3.

On Uniprot, I added all four domains to the basket and downloaded their sequences in fasta format. This stored the domains in a single file with their respectable headers. After that I used the online MSA Viewer (http://msa.biojs.net/app/) to visualize the protein sequence aligments. Based on the similarities I saw between the 1st - 2nd and the 3rd – 4th domains, I decided to align the two of them separately using EMBOSS's pairwise alignment tool (http://www.ebi.ac.uk/Tools/emboss/) with Needleman-Wunsch algorithm and default settings.

The similarity percentages came back at 55.4% and 60.5% which means that I might have been right. The 1st-2nd and the 3rd-4th domains likely share lineage but have mutated significantly since replication.

4.

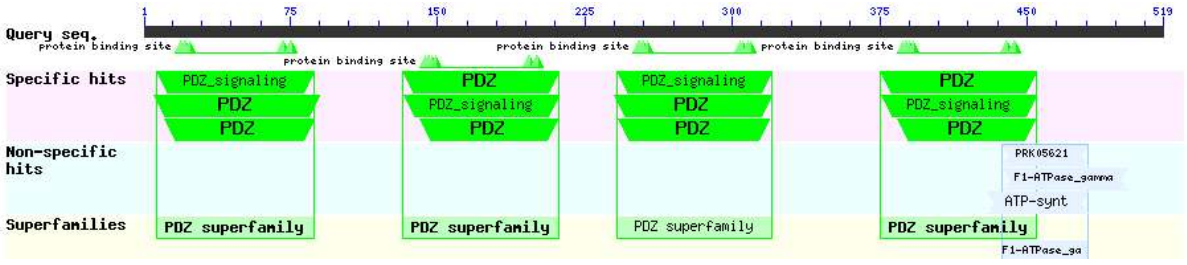
I compared the aforementioned domains using EMBOSS's dotplot. I set the window size a little higher and left the threshold unchanged (12 and 23, respectively), so the algorithm uses 12 amino-acid long segments and dots are displayed if the alignment score is at least 23.

In the first case, I got a quasi-straight line (breaking at several points but always remaining very close to the identity [x = y] line), that's missing a segment (about ¼ of it) in the middle.
In the second case, I got three lines. Two of them along the identity line (as in the prevoius case), with veryhort segments missing. The third (and somewhat short) line could be seen near the upper-left corner, meaning that two shorter sequence segment from different regions show similarities. From this we can deduct that the domains contain two very similar regions.
The results seem to support the previous theory, even furthering it in the case of the 3rd-4th domains.

5.

For this point, I used the NCBI Blast (https://blast.ncbi.nlm.nih.gov/Blast.cgi). I chose 'human' as species to search. (The sequence I used was the isoform-1 since isoform-2 is incomplete).

Blast found other members of the NHRFi family as similar proteins and also detected the PDZ protein binding sites in the sequence.



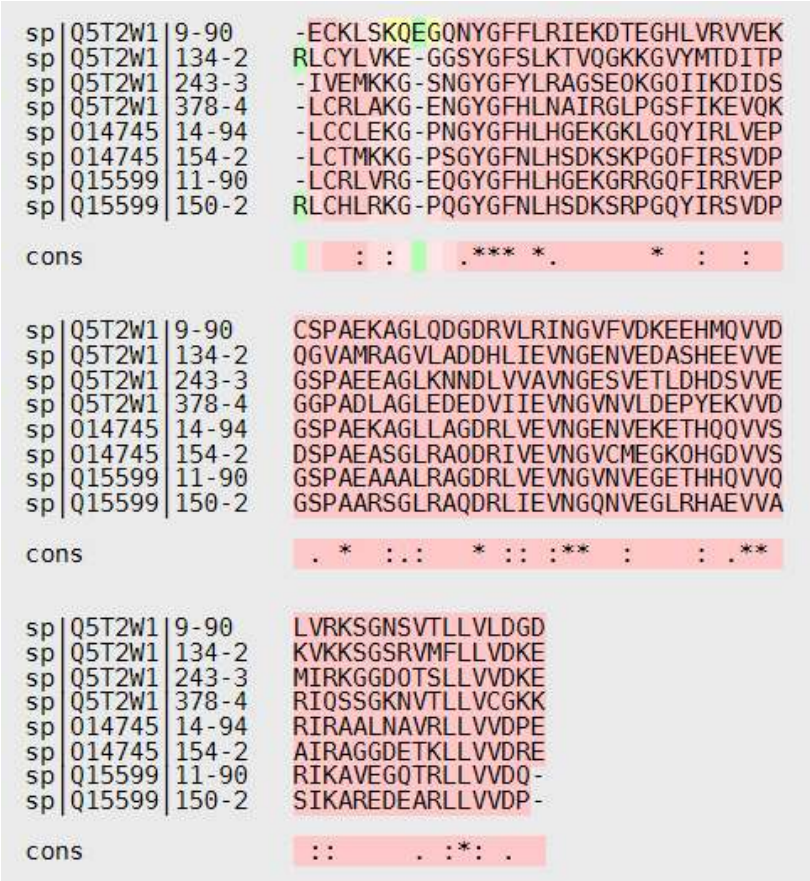
When hovered over the 'protein binding site' notation, an window containing info about the domain appears. The description says 'protein binding site on conserved domain PDZ_signaling' from which I infer that these 4 domains are highly conserved among PDZ signaling proteins and thus these are the same 4 domains annotated in the Uniprot database. Among the blast results, I could even see the domains separate by the alignment score.

When selecting proteins from the hits, I chose to not include the same protein in different isoforms, unnamed protein pruducts, putative pseudogenes and proteins with a total score under 200. This way the database is more relevant and prompt. I selected 15 proteins for the database and clicked on 'multiple aligment' and downloaded it in fasta format.

6.

Among the previous Blast results were the other members of the NHRFi_HUMAN family so I decided to add the PDZ domains of these proteins to my accumulated domain fasta file. (NHRF1 and NHRF2 to be exact). I searched the UniprotKB database for these proteins and downloaded the domain sequences. Hence, I had 8 domains in a single file (fasta format).

For the multiple alignment I used T-Coffee (More precisely: http://tcoffee.vital-it.ch/apps/tcoffee/result?rid=70801ad7).



The pink color signifies a "good" matching, while yellow is for "average" and green is for "bad".

The alignment score came back at 986 which is quite high (as it should be, since they're related).

I downloaded the alignment file and converted it to stockholm format (via http://sequenceconversion.bugaco.com/converter/biology/sequences/index.html).

I did this with the "database" file from the previus exercise, too, and continued to work with the protein sequence file from here on.
For the HMM build, I used the webserver (https://www.ebi.ac.uk/Tools/hmmer/).

For the first HMM searching, I used the default settings, with the target database being Reference Proteomes. There were 20221 Significant Query Matches. Among the results, 2960 posessed 1 PDZ domain, 921 posessed 2 PDZ domains, 466 posessed 3 PDZ domains and 406 posessed 4 PDZ domains. Going through the results, I didn't find more sequences containing solely the PDZ domains. On the other hand, a lot of them had several other kinds of domains in them, most frequently Trypsin-like peptidase, Peptidase S41, PID, Peptidase M50 and Protein Kinase domains. There were also a significant amount of hits with domains of unknown function in them. Since the PDZ domains take up ~63% of original NHRF3 protein sequence, and this was the second "similarity search" I've conducted, it was expected of the results to contain a PDZ domain.

Back in the score sheet, I could see that the top results were mostly NHRF3 itself, but isolated from different (but all mammalian) species. In the first 2 pages of hits, I could find the other members of the NHRFi family (yet again, from a diversity of species). There were a lot of uncharacterized proteins among the hits, too.

The results differed from that of the NCBI BLAST, because here I mostly found separate protein, whilst in the case of BLAST, many of the hits were single domains or protein complexes.

Searching with the target organism being restricted to 'human' results a significant drop in the amount of results (only 438 query matches).

	Target	Description	Species	E-value
>	NHRF3_HUMAN	Na(+)/H(+) exchange regulatory cofactor NHE-RF3	Homo sapiens	1.1e-214
>	NHRF4_HUMAN	Isoform 2 of Na(+)/H(+) exchange regulatory cofactor NHE-RF4	Homo sapiens	1.4e-178
>	NHRF4_HUMAN	Na(+)/H(+) exchange regulatory cofactor NHE-RF4	Homo sapiens	3.2e-178
>	E9PPZ1_HUMAN	Na(+)/H(+) exchange regulatory cofactor NHE-RF4	Homo sapiens	3.7e-178
>	NHRF4_HUMAN	Isoform 3 of Na(+)/H(+) exchange regulatory cofactor NHE-RF4	Homo sapiens	9.9e-173
>	NHRF3_HUMAN	Isoform 2 of Na(+)/H(+) exchange regulatory cofactor NHE-RF3	Homo sapiens	4.8e-148
>	NHRF2_HUMAN	Na(+)/H(+) exchange regulatory cofactor NHE-RF2	Homo sapiens	5.0e-148

This time around, there were many duplicates among the results, different isoforms of the same proteins, but also various multiple-PDZ-domain-proteins. Looking at their profiles, I found that some of them contained 10-15 PDZ domains.

I also ran a search with the significance E values changed: 0.1 for sequence and 0.3 for hit. (I didn't restrict by taxonomy.) As expected, this resulted in the increase of the number of hits (24213).

7.

I searched the UniProtKB database for the NHRF3 protein of different species (when searching, the results show their source organism). I chose 6 mammalian species: human, rat, mouse, rabbit, bovine and sumatran orangutan. I couldn't find any bird species, unforunately, even though I tried other databases as well (Ensembl, PDB, NCBI). The only non-mammalian sequence I could find is a slightly incomplete (and unreviewed) one from the blackstripe livebearer, which is a fish of the genus Poeciliopsis. I added them to the Uniprot Basket and aligned them.

For the mRNAs, I visited the uniprot pages of the proteins from above, and from the "Sequence databases" section I downloaded their mRNAs, creating a single fasta file of them. I aligned them using the nucleotide BLAST and downloaded the result. (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome)

I set up the windows version of R on my computer and used the command window for the phylogenetic exercises. I set the working directory to the folder I stored the files in (**setwd(…)**). I called the library seqinr and ape : **library(seqinr), library(ape).**
I read the alignments into the 'aln' variable:

```
> aln.prot<-read.alignment("protein.fasta",format="fasta")
> aln.mrna<-read.alignment("mrna.fasta",format="fasta")
```

In order to construct the phylogenetic tree, we must calculate the distance between the aligments:

```
> d.prot<-dist.alignment(aln.prot,matrix="similarity")
> d.mrna<-dist.alignment(aln.mrna,matrix="similarity")
```

Converting the alignment object:

```
> aln.mrna.b<-as.DNAbin(aln.mrna)
```

Using the Neighbour-Joining alhorithm:

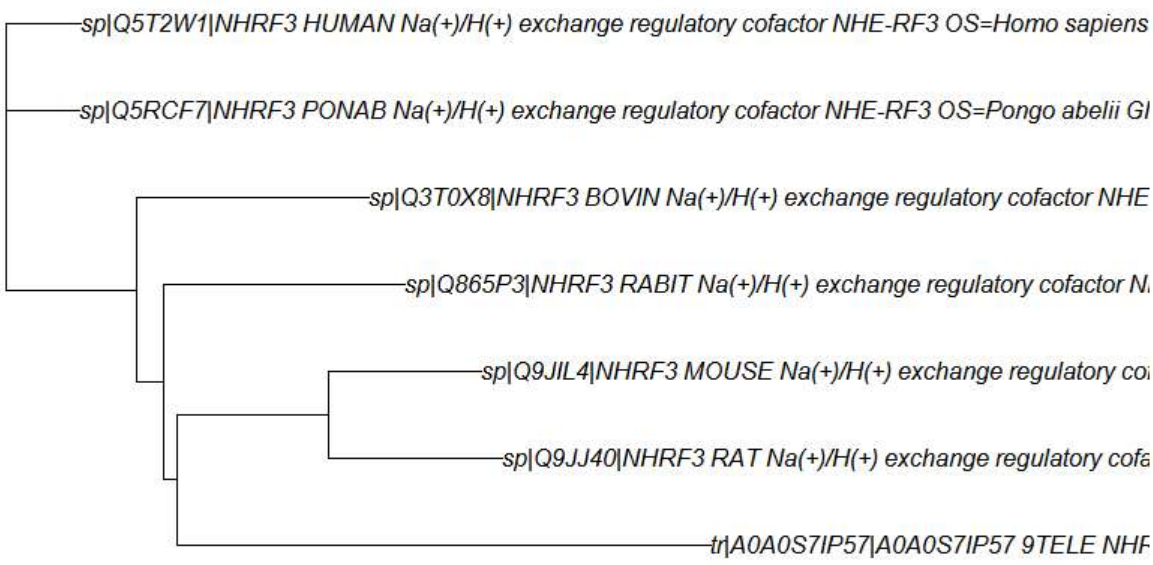
```
> t.prot<-nj(d.prot)
```

Checking whether or not it's a rooted tree:

```
> is.rooted(t.prot)
[1] FALSE
```

Since it's not, we have to use the outgroup method. Fish are clearly an outgroup compared to mammals, so we can use this to root the tree (fish are group 6 in the file):

```
> t.prot<-root(t.prot,outgroup=6,resolve.root=T)
```



The phylogenetic tree based on the proteins is ready. It's correct, with the only exception of the fish (the one at the bottom). This error may be due to the fact, that its protein sequence was incomplete and that's why the algorithm assigned the node incorrectly.