BD PROJECT

Data :

UniProt : P54315

PfamID : PF00151

Domain Position : 18-353

Organism : Homo sapiens (Human)

Pfam Name : Lipase/vitellogenin

Domain Sequence : KEVCYEDLGCFSDTEPWGGTAIRPLKILPWSPEKIGTRFLLYTNENPNNFQILLLSDPSTIEASNFQMDRKTRFIIHGFIDKGDESWVTDMCKKLFEVEEVNCICVDWKKGSQATYTQAANNVRVVGAQVAQMLDILLTEYSYPPSKVHLIGHSLGAHVAGEAGSKTPGLSRITGLDPVEASFESTPEEVRLDPSDADFVDVIHTDAAPLIPFLGFGTNQQMGHLDFFPNGGESMPGCKKNALSQIVDLDGIWAGTRDFVACNHLRSYKYYLESILNPDGFAAYPCTSYKSFESDKCFPCPDQGCPQMGHYADKFAGRTSEEQQKFFLNTGEASNF

Task 1

Use UniRef90 :

1. Domain Characteristics:

* PF00151 is a well-characterized domain family
* Lipases are found across many species
* Your sequence is from human (P54315), but you want to capture diversity

1. Why UniRef90 is particularly good here:

* It clusters sequences at 90% identity, which:
  + Removes redundant sequences (like multiple human variants)
  + Keeps enough diversity to capture functional variations in lipases
  + Will include homologs from different species
  + Reduces bias from overrepresented species

Blast Service :

<https://www.uniprot.org/blast>

Using these Parameters : (E-Thresh 0.001 to be more stringent and get high quality homologous , increased Hits to 500 (half of maximal possible, we want high quality homologous, but not too much which may result in not actual homologous and mess up our model)



444 Hits ; but can’t download FASTA from UniRef90 results directly :  
  
The reason you couldn't directly download from UniRef90 is due to how UniProt's BLAST results work. UniRef90 is a clustered database where sequences with 90% or more similarity are grouped together to reduce redundancy. When you BLAST against UniRef90, you're initially just getting cluster representatives.

To get the actual sequences for multiple sequence alignment, you need to:

1. Select the sequences first
2. Map them to UniProtKB entries (which you did)
3. Then download the FASTA files

This extra step is actually beneficial because:

* It helps you get the full, individual sequences rather than just cluster representatives
* You get access to more detailed annotation information

🡪 So first mapped to UniProtKB, resulting in

Ein Bild, das Text, Schrift, Screenshot, weiß enthält.

Automatisch generierte Beschreibung

Download FASTA from that

Redundancy (Rows) Analysis :

* First removed rows with 100% identity using Jalview (i.e. 100% threshold)
* From 839 to 73 rows

Conservation (Columns) Analysis :

* I’d say we use a code , first analyzing each column using multiple different ways (frequency amino acids, gap counts, etc.)
* We need to pay attention not to remove too harshly : we are supposed to model PSSM/HMM, thus if we remove too harshly we end up with only basically redundant alignments of little parts of the sequences, which will not help us in modeling at all
* How to prevent this ? (add stuff here) :
  + When we check for which amino acid sequences match in a column, we should pay attention to groups of amino acids that are similar to each other, i.e. if they are in the same group, they are still kind of similar and not a completely “wrong” alignment (groups found here : <https://en.wikipedia.org/wiki/Conservative_replacement#:~:text=There%20are%2020%20naturally%20occurring,both%20small%2C%20negatively%20charged%20residues>.)
* Until now we use :
  + Remove when (columns where we have more than 70% gaps OR columns where conservation based on amino acid groups is less than 30%)
  + Possibly entropy ?

PSSM CREATION

* Had to fix the parameters for removal of columns (i.e. conservation) such that we got like less than 100 residues in the alignment, else the ncbi+ blast algo didn’t work
* ncbi-blast-2.16.0+/bin/psiblast -subject data/protein\_family/trimmed\_alignment.fasta -in\_msa data/protein\_family/trimmed\_alignment.fasta -out\_ascii\_pssm data/protein\_family/trimmed\_alignment.pssm\_ascii -out\_pssm data/protein\_family/trimmed\_alignment.pssm

HMM CREATION

* hmmer-3.4/src/hmmbuild data/protein\_family/trimmed\_alignment\_HMM.hmm data/protein\_family/trimmed\_alignment.fasta