BD PROJECT

Data :

UniProt : P54315

PfamID : PF00151

Domain Position : 18-353

Organism : Homo sapiens (Human)

Pfam Name : Lipase/vitellogenin

Domain Sequence : KEVCYEDLGCFSDTEPWGGTAIRPLKILPWSPEKIGTRFLLYTNENPNNFQILLLSDPSTIEASNFQMDRKTRFIIHGFIDKGDESWVTDMCKKLFEVEEVNCICVDWKKGSQATYTQAANNVRVVGAQVAQMLDILLTEYSYPPSKVHLIGHSLGAHVAGEAGSKTPGLSRITGLDPVEASFESTPEEVRLDPSDADFVDVIHTDAAPLIPFLGFGTNQQMGHLDFFPNGGESMPGCKKNALSQIVDLDGIWAGTRDFVACNHLRSYKYYLESILNPDGFAAYPCTSYKSFESDKCFPCPDQGCPQMGHYADKFAGRTSEEQQKFFLNTGEASNF

Task 1

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

Use UniRef90 : (OPTIMIZE : against UniProt or UniRef50 (but I think since UniRef50 even less than UniRef90, only UniProt would be better)

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

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) OPTIMIZE : less stringent E-Thresh values (so higher E-value)



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

How to reduce entries (TO OPTIMIZE)

Using conservation.py

Redundancy (Rows) Analysis :

* First removed rows with 100% identity using Jalview (i.e. 100% threshold)
* From 839 to 73 rows (OPTIMIZE : 100% is already the most possible, so maybe if we have more hits than 839, we can change stuff up here , or maybe even more little than 100%, but I don’t know if taking less sequences would result in better results)

Conservation (Columns) Analysis :

* 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](https://en.wikipedia.org/wiki/Conservative_replacement#:~:text=There are 20 naturally occurring,both small%2C negatively charged residues).)
* 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 ?

CREATES trimmed\_alignment.fasta

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 (Here we have to see what the problem is, because possibly we need more than a 100 residues to get a good model ? OPTIMIZE)
* 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

Model evaluation

* 1. Generation of predictions

For PSIBLAST :

* When working on MAC : First have to change settings so we have access to use psiblast (in terminal, go to folder where psiblast/makeblastdb located and run this)

Ein Bild, das Text, Screenshot, Software, Multimedia-Software enthält.

Automatisch generierte Beschreibung

* Then we have to do the same also for makeblastdb, which we then use to create some kind of formatted swissprot database :
* ./ncbi-blast-2.16.0+/bin/makeblastdb -in uniprot\_sprot.fasta -dbtype prot -out swissprot

And then finally, to create the output such that we see where the pfam domains are in the sequence

./ncbi-blast-2.16.0+/bin/psiblast -in\_pssm trimmed\_alignment\_PSSM.pssm \

-db swissprot \

-out psiblast\_search\_output.txt \

-outfmt "6 qseqid sseqid qstart qend sstart send pident evalue" \

Columns meaning :  
  
qseqid: Query sequence identifier (your domain)

* sseqid: Subject sequence identifier (matched protein)
* qstart: Start position in your query domain
* qend: End position in your query domain
* sstart: Start position in the matched sequence
* send: End position in the matched sequence
* pident: Percentage of identical matches
* evalue: Expectation value (statistical signific

For HMMER :

./hmmer-3.4/src/hmmsearch trimmed\_alignment\_HMM.hmm uniprot\_sprot.fasta > hmmsearch\_output.txt

The matching positions are found in these outputs aswell as the found proteins

For PSIBLAST :

* use Rewriting\_helper.py to create more neat output for psiblast

For HMMER :

* to implement
  1. Defining the ground truth

Using InterPRO API :

Use code API\_search.py . Currently we use URL :

<https://www.ebi.ac.uk/interpro/api/protein/reviewed/entry/pfam/PF00151/>

maybe use URL :

<https://www.ebi.ac.uk/interpro/api/protein/unreviewed/entry/pfam/PF00151/> (unreviewed)

and then use json\_extractor.py (to turn .JSON into .CSV)

Model Comparison (Metrics) Step :

* Use metrics\_eval\_psiblast\_NEW.py for PSIblast
* PsiBlast against Swissprot using PSSM, we obtain 83 sequences. The results are downloaded as csv file.
* We extract from the csv file the name, alignement start and alignement end.
* From [here](https://www.ebi.ac.uk/interpro/entry/pfam/PF00151/logo/) it is possible to download the PF00151.hmm file