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>

NEW APPROACH (seems to make waaaay nicer sequence alignments)

First get alignment

* UniProtKB alignment (not UniRef90) at e-thresh 0.0001, 1000 hits, then ID matching to UniProtKB aswell : results in 1000 sequences ( I can show you what I did on the webpage), download FASTA

Reduce entries

* Reducing Rows : Use JalView at 90% identity 🡪 383 sequences (UNIPROTKB\_INITIAL.FASTA)
* In JalView you can also see that all sequences are directly aligned in the beginning, and then we mostly have just a whole bunch of gaps (🡪 so maybe conservation analysis just removing like all of the last columns where we basically have only gaps ?)
  + Reducing columns (using conservation.py) : For now I just did gap threshold 90% (so remove columns where more than 90% is just gaps) 🡪 we get around 800 columns (from initial 1400)
  + (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.)>) (we aren’t using this right now for the current solution, but maybe we can try to see if it is helpful

CREATES trimmed\_alignment.fasta

PSSM CREATION

* 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 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 \

-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 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