Make Datbase (BlastP)

./makeblastdb -in NveProt.fas -dbtype 'prot' -out NveProt -name -NveProt

Iedb (Autoimmune Disease 🡪 Epitope)

REQUIRES: <ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/>

Ncbi-blast+ latest

BLASTP (Epitope 🡪 UniProt Name)

<ftp://ftp.ncbi.nih.gov/blast/db/FASTA/>

INFO: <http://genetics.bwh.harvard.edu/msblast/iblast_databases.html>

blastp -query ../test.faa -db swissprot -evalue 1e-03 -outfmt 1 -max\_target\_seqs 1 | grep 'Full=' | cut -d '=' -f 2

To extract the name only (CHECKED AND SUCCESFUL):

blastp -query ./temp\_fasta.faa -db swissprot -outfmt 5 -max\_target\_seqs 50 | grep 'Full=' | awk -F 'Full=' '{ print $2 }' | awk -F '</Hit\_def>' '{ print $1 }' | cut -d ';' -f 1 | head -1

Need a way to keep sequence name

Removed -evalue because it was removing too many hits (still not catching all peptides)

blastp -query ./epitopes.faa -db swissprot -outfmt 5 -max\_target\_seqs 1 | grep 'Full=' | awk -F 'Full=' '{ print $2 }' | awk -F '</Hit\_def>' '{ print $1 }' | cut -d “;” -f1 | sort | uniq > uniq\_uniprot\_names.txt

UniProt Protein Name 🡪 GENE SYMBOL (Taxon: 9606)

\*Unique list of UniProt Name

<https://biodbnet-abcc.ncifcrf.gov/db/db2db.php>

Protein Atlas (GENE SYMBOL 🡪Tissue-Specificity)

\*where RELIABILITY==’Approved’ && LEVEL==’HIGH’

What should the threshold for level be?

GREP-STYLE

grep -m1 -B3 'LFALTLPIW' swissprot | grep '>' | cut -d '=' -f 2 | cut -d ";" -f 1

ACC-ID

grep -m1 -B3 ' LFALTLPIW' swissprot | grep ‘>’ | cut -d '>' -f 2 | cut -d '.' -f 1 | head -1

blastp -query ./test.fa -db swissprot -outfmt 5 -max\_target\_seqs 50 | grep 'Hit\_def>' | awk -F 'Hit\_def>' '{ print $2 }' | cut -d '.' -f 1 | head -1

blastp -query ./test2.faa -db swissprot -remote -entrez\_query "Homo sapiens [Organism]" -outfmt 1

LFALTLPIW

GENE NAME ALIAS

<https://www.biostars.org/p/1378/>

Documentation

1. Download ncbi-blast+ latest version
2. Add to path
3. Git clone repository
4. Cd ‘’/path/to/repository”
5. chmod -R 775 \*
6. Run build.sh

Directory structure set ahead of time in repository

Data/must contain tissue-system labels