ProteinPrediction II

PPII Ex3 - Week 3

Jonathan Boidol, Rene Schoeffel, Yann Spöri

Mapping IDs

1. `http://www.uniprot.org/mapping/` maps 2813 Uniprot ACs to 2801 Entrez Gene IDs
2. python script `map.py`
   - reads mapping and annotation files
   - maps Uniprot AC to HPO term #or root node, but we have better annotations for everything
3. output:

```
P00441   HP:0003394,HP:0002314,HP:0003202,HP:0010535
P31749   HP:0000400,HP:0004322,HP:0004325
P31213   HP:0000028,HP:0008736
...
```

Stuff done

1. Clean Header

Small python script uses a regular expression to get the uniprot ids

2. Prepare blast database

formatdb -i genes_UniProt.fasta

3. get the n nearest proteins

python __main__.py –seq “Sequence” -k <n>

Stuff still calculating

```
#!/bin/bash
multithread.pl 'pp2/*.seq' 'hhblits -i $file -d
/mnt/project/rost_db/hhblits/uniprot20 -oa3m $name.a3m' -cpu 4
multithread.pl 'pp2/*.a3m' $scriptdir/'addss.pl $file' -cpu 4
hhblitsdb.pl -o pp2/ -ia3m pp2/ -cpu 4
```

Interface

Each similar sequences search tool should return:

1. hit_id: The id of the hit
2. hit_value: A normalized score of the ident string
3. hit_oder: wether the hit_value is increasing or decreasing by quality
4. hit_from: the starting position of the hit
5. hit_to: the end position of the hit