Z2 protein enrichment report, Mateusz Bieganski mb385162

A. IMPLEMENTATION DETAILS

1 extend.py

BLAST queries done by NcbiblastpCommandline from Bio.Blast.Applications, saved to XML and parsed by NCBIXML. Full sequences obtained by batch request to Entrez, next written as list od sequences to 'blast_out.fa' file using Bio.SeqIO module. I chose that implementation because of it's simplicity (no strange caveats and hacking).

2 scan_pfam.py

Parsing HMMSCAN output is done manually (because of response format that changed recently). csv.writer 'likes' list of list to be dumped as a table, so to achieve it I keep maps seq2domains (seq -¿ [domain] type), dom2id and seq2id. Everything interesting is done by given lines:

```
1         for seq, dom_list in seq2domains.items():
2             for dom in dom_list:
3                 res[seq2id[seq]][dom2id[dom]] = str(1)
```

3 fisher.py

Really nothing special here, just fisher test implementation and it's usage with inputs csvs. I used numpy to easily add up by rows and columns. For 'clean code nad modularity' purpose, I dump results to csv file 'results.csv', with both fisher test implementation results contained (next script creates diagrams)

4 pipeline.py

Script that runs whole pipeline, draws charts and tables, nothing special here.

B. RESULTS

5 Fisher's test implementation benchmark

Let's start with Fisher test implementation. As shown in figure 1, custom, straightforward way leads to results deformed with row of size 1e-15, which is rather negligible.

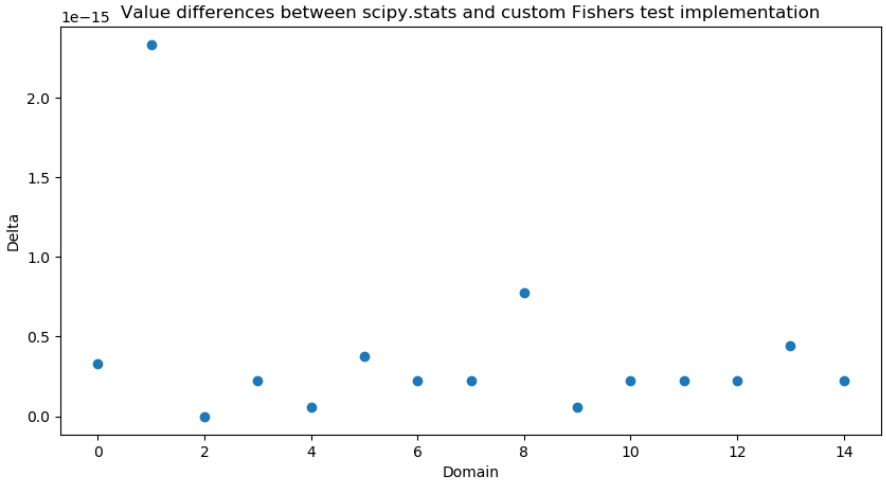

Figure 1: Fisher's deltas

6 Enrichment results

As a ground truth I used results from scipy.stats Fisher's test implementation from obvious reasons. Results of whole experiment clearly indicates enrichment in area of PF00271.31 domain, which represents *Helicase_C* family. There are no reasons to conclude enrichment in case of another examined domains, where results of Fisher's exact test were greater than 0.12 (I consider statistically significant with ¡= 0.05).

What is worth noticing, PF07717 and PF04408 approaches significant boundary with it's same result of 0.17. About PF07717 we can read "This family is found towards the C-terminus of the DEAD-box helicases (PF00270). In these helicases it is apparently always found in association with PF04408" which strongly suggest corelation with *Helicase_C* and PF04408. Another interesting family is PF00270.29 with it's 0.12. It's family of DEAD-/DEAH box helicases, as PF07717 and PF04408, which strengthens significance of whole observation.

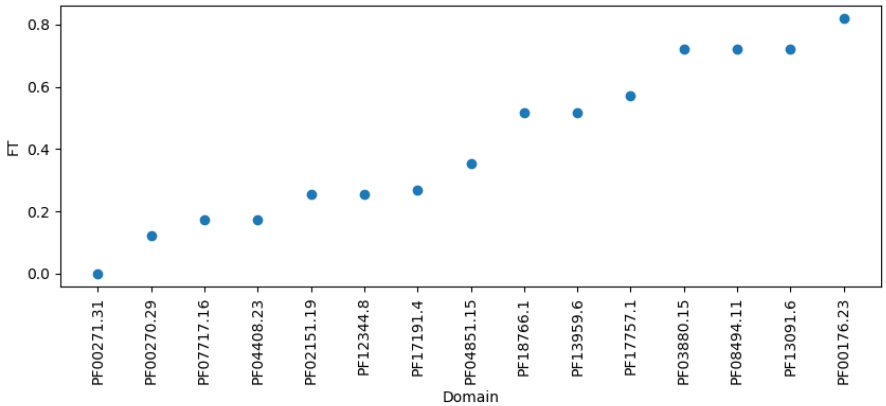

Figure 2: Sorted results of Fisher's test for each domain