Practical guide to BLASTp

Use this to find sequence similarity of your favorite protein in the proteins of another organism.

NCBI server

You can use NCBI BLAST to check for sequence similarity in NCBI's database of genomes, or to a set of sequences that you can upload. However, using BLAST for many sequences can be inconvenient to process as an output, and running BLAST locally would be the best approach

Locally with the command line

Create BLAST database in a folder with the same name as the database. As an example here, the database would be the proteome of Chlamydomonas.

```
makeblastdb -in chlamydomonas.fa -dbtype prot -out chlamydomonas
```

Run BLASTp from the database folder
(Note: adjust the parameters to your needs; i.e. evalue & max_target_seqs)
(Note II: there are different ways to run blast locally. See BLASTall, BLASTn, etc; the parameters are not the same as in BLASTp)

```
blastp -query ../secuencias_query.fasta -db chlamydomonas -out
../resultados_blastp.txt -evalue 0.05 -outfmt "6 std qcovs" -max_target_seqs
1
```

BLASTp output by column. Information of BLAST terms can be found in the glossary.

1. query
2. hit
3. identity
4. alignment length
5. #mismatch
6. #gaps
7. start query
8. end query
9. hit start
10. hit end
11. e-value → we want this close to zero
12. bit score
13. Coverage

Filter your results

Which BLAST hits do you actually keep?
→ The answer is always “It depends”. You need to know your data and there are different methods for filtering.

Identity and coverage thresholds

- Depending on your dataset you can keep the hits that meet certain percentage of identity and the percentage of length covered in the target sequence.
- For example, in a database of bacterial proteins, the suggested thresholds by PATRIC are 80% - 80% when using BLAST for bacteria, but 50% - 50% for finding human homologs.
- Another example: Sachli's threshold. She uses sequence identity > 90% and coverage > 95% to keep hits within bacteria strains of the same species.

E-value

- The e-value can be tricky to use, since it will change depending on the size of the database.

Bitscore

- The Bitscore is another indicator, but in contrast to the e-value, it is independent of sequence length and database size.