BLAST

Nociones Básicas de Bioinformática y Genómica
(Máster en Bioinformática, Universidad de Valencia)

David Montaner

2015-04-20

BLAST

BLAST: Basic *Local Alignment* Search Tool

- *pairwise* alignment
- *local* alignment (not equal size assumed for the sequences)

BLAST concepts

Alignment of pairs of sequences:

- **query**: our **input** nucleotide or protein sequences.
- **database**: the **reference** database of sequences.

We *blast a query to a database*

Local vs. Global alignment

```
Global  FTFTALILLAVAV
         F--TAL-LLA-AV


 Local  FTFTALILL-AVAV
         --FTAL-LLAAV--
```

From Wikipedia: Sequence alignment

BLAST variations

There are many different variations of BLAST available:

- **DNA** query to a **DNA** database: *blastn*
- **protein** query to a **protein** database: *blastp*
- **DNA** query (translated in all six reading frames) to a **protein** sequence database: *tblastx*

Many other variants

- PSI-BLAST: iterative protein sequence similarity searches
- RPS-BLAST: for protein domains
- MegaBLAST: optimized for nucleotide sequences which aligning almost perfectly

BLAST Scores and Statistics

Generally in an alignment we want to know:

- general description.
- how good is the alignment.
- how likely is it to happen just by random chance.

Main output parameters

- **E-value**: indication of the statistical significance
  - the lower thee better
  - dependent of the database size
- **bit score**: indication of how good the alignment is
  - the higher the better
  - is normalized: different alignments can be compared

BLAST output information

- query id
- subject ids
- % identity
- alignment length
- number of mismatches
- gaps
- query start
- query end
- sequence start
- sequence end
- **E-value**
- **bit score**

BLAST output formats

- Traditional Report: text or HTML with hyperlinks.
  - devised for human readability
  - not structured
- Hit Table: tabular text file
  - not all information is contained in it
  - easy to read in computers
  - not structured
- XML:
  - structured
  - good to be generally parsed (see `http://biopython.org/`)
  - contains the alignment
- ASN.1:
  - structured
  - to be internally used at the NCBI (see NCBI toolkit)
  - does contains the alignment

BLAST output formats
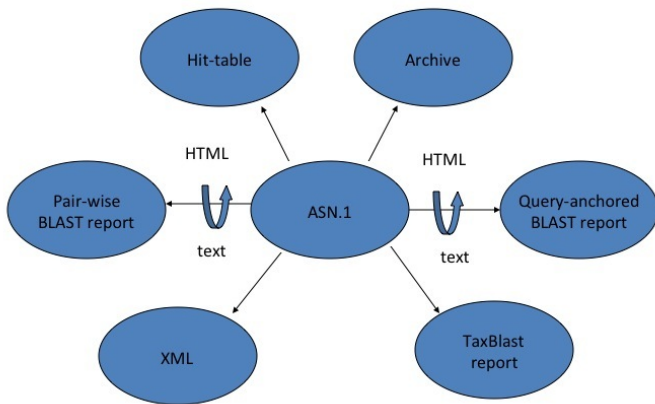
Practical

http://blast.ncbi.nlm.nih.gov/Blast.cgi

References

- NCBI Handbook: The BLAST Sequence Analysis Tool
- Wikipedia: Sequence alignment

Further reading:

- NCBI BlastDocs