

What is BLAST?

- Basic BLAST search
 - What is BLAST?
 - The framework of BLAST
 - Different BLAST programs
 - BLAST databases you can search
 - Where can I run BLAST?

1

What is BLAST?

- BLAST stands for Basic Local Alignment Search Tool
- Why BLAST is popular?
 - Good balance of sensitivity and speed
 - Reliable
 - Flexible
- Produce local alignments: short significant stretches of similarity, irrespective of where they are in the sequence

2

BLAST Programs

The most common BLAST search include five programs:

Program	Database (Subject)	Query
<i>blastn</i>	Nucleotide	Nucleotide
<i>blastp</i>	Protein	Protein
<i>blastx</i>	Protein	Nt. → Protein
<i>tblastn</i>	Nt. → Protein	Protein
<i>tblastx</i>	Nt. → Protein	Nt. → Protein

3

blastn

- *blastn*
 - The query is a nucleotide sequence
 - The database is a nucleotide database
 - No conversion is done on the query or database
- DNA :: DNA homology
 - Mapping oligos to a genome
 - Annotating genomic DNA with transcriptome data from ESTs and RNA-Seq
 - Annotating untranslated regions

4

blastp

- *blastp*
 - The query is an amino acid sequence
 - The database is an amino acid database
 - No conversion is done on the query or database
- Protein :: Protein homology
 - Protein function exploration
 - Novel gene → make parameters more sensitive

5

blastx

- *blastx*
 - The query is a nucleotide sequence
 - The database is an amino acid database
 - All six reading frames are translated on the query and used to search the database
- Coding nucleotide seq :: Protein homology
 - Gene finding in genomic DNA
 - Annotating ESTs and transcripts assembled from RNA-Seq data

6

tblastn

- *tblastn*
 - The query is an amino sequence
 - The database is a nucleotide database
 - All six frames are translated in the database and searched with the protein sequence
- Protein :: Coding nucleotide DB homology
 - Mapping a protein to a genome
 - Mining ESTs and RNA-Seq data for protein similarities

7

tblastx

- *tblastx*
 - The query is a nucleotide sequence
 - The database is a nucleotide database
 - All six frames are translated on the query and on the database
- Coding :: Coding homology
 - Searching distantly-related species
 - Sensitive but computationally expensive

8

BLAST output

1. List of sequences with scores
 - Raw score
 - Higher is better
 - Depends on aligned length
 - Expect Value (E-value)
 - Smaller is better
 - Independent of length and database size
2. List of alignments

9

The Databases (1)

- Nucleotide collection (nr/nt)
 - Nucleotide sequences from the GenBank, EMBL, DDBJ, PDB, and RefSeq databases
 - Non-redundant (compile and remove duplicates)
 - Low quality; names can be meaningless
- Core Nucleotide Database (core_nt)
 - Default nucleotide database as of August 2024
 - Includes all gene-related sequences in nr/nt
 - Excludes eukaryotic chromosome sequences

10

The Databases (2)

- Non-redundant protein sequences (nr)
 - Include GenBank CDS translations, and proteins from the PDB, Swiss-Prot, PIR, and PRF databases
- UniProt/Swiss-Prot
 - Curated from literature
 - **REAL** proteins; **REAL** functions; small
- Genomic databases
 - Human, Mouse, *Drosophila*, Arabidopsis, etc.
 - NCBI, species-specific web pages

11

Where Can I run BLAST?

1. NCBI BLAST web service
 - <https://blast.ncbi.nlm.nih.gov/Blast.cgi>
2. EBI BLAST web service
 - <https://www.ebi.ac.uk/jdispatcher/sss/ncbiblast>
3. FlyBase BLAST
 - <https://flybase.org/blast/>
 - *Drosophila melanogaster* only

12