Introduction to bioinformatics, databases and BLAST

Dr. Pankaj Singh

School of Agriculture and Environment

University of Western Australia

Introduction to Bioinformatics

- Definition:
- - Interdisciplinary field that develops methods and software tools for understanding biological data.
- - Combination of biology, computer science, and information technology.
- Importance:
- - Managing and analyzing large datasets generated by genomics and proteomics.
- - Deciphering genomic sequences to extract hidden biological information.

Importance for Plant and Environmental Scientists

- Agriculture:
- - Analyzing and improving crop traits.
- - Decoding genomic information.
- - Identifying micro-satellite markers.
- - Investigating molecular mechanisms of stress tolerance.
- Environmental Sciences:
- - Monitoring biodiversity.
- - Managing wildlife populations.
- - Preserving endangered species.
- - Assessing impacts of human activities and climate change.

Genomic Databases

- NCBI GenBank:
- - Public database of nucleotide sequences and annotations.
- - URL: https://www.ncbi.nlm.nih.gov/genbank/
- EMBL-EBI:
- - European Molecular Biology Laboratory’s database for nucleotide sequences.
- - URL: https://www.ebi.ac.uk/
- DDBJ:
- - DNA Data Bank of Japan, part of the International Nucleotide Sequence Database Collaboration (INSDC).
- - URL: https://www.ddbj.nig.ac.jp/index-e.html

Protein Databases

- UniProt:
- - Comprehensive resource for protein sequence and function annotation.
- - URL: https://www.uniprot.org/
- Protein Data Bank (PDB):
- - Archives 3D structural data of biological macromolecules.
- - URL: https://www.rcsb.org/

Metabolomics Databases

- KEGG:
- - Provides graphical representation of cellular processes and drug development pathways.
- - URL: https://www.genome.jp/kegg/
- PubChem:
- - Database of chemical molecules and their activities against biological assays.
- - URL: https://pubchem.ncbi.nlm.nih.gov/

Phenotypic Databases

- FlyBase:
- - Database for Drosophila genetics and molecular biology.
- - URL: https://flybase.org/
- WormBase:
- - Information on the genetics, genomics, and biology of C. elegans and other nematodes.
- - URL: https://www.wormbase.org/

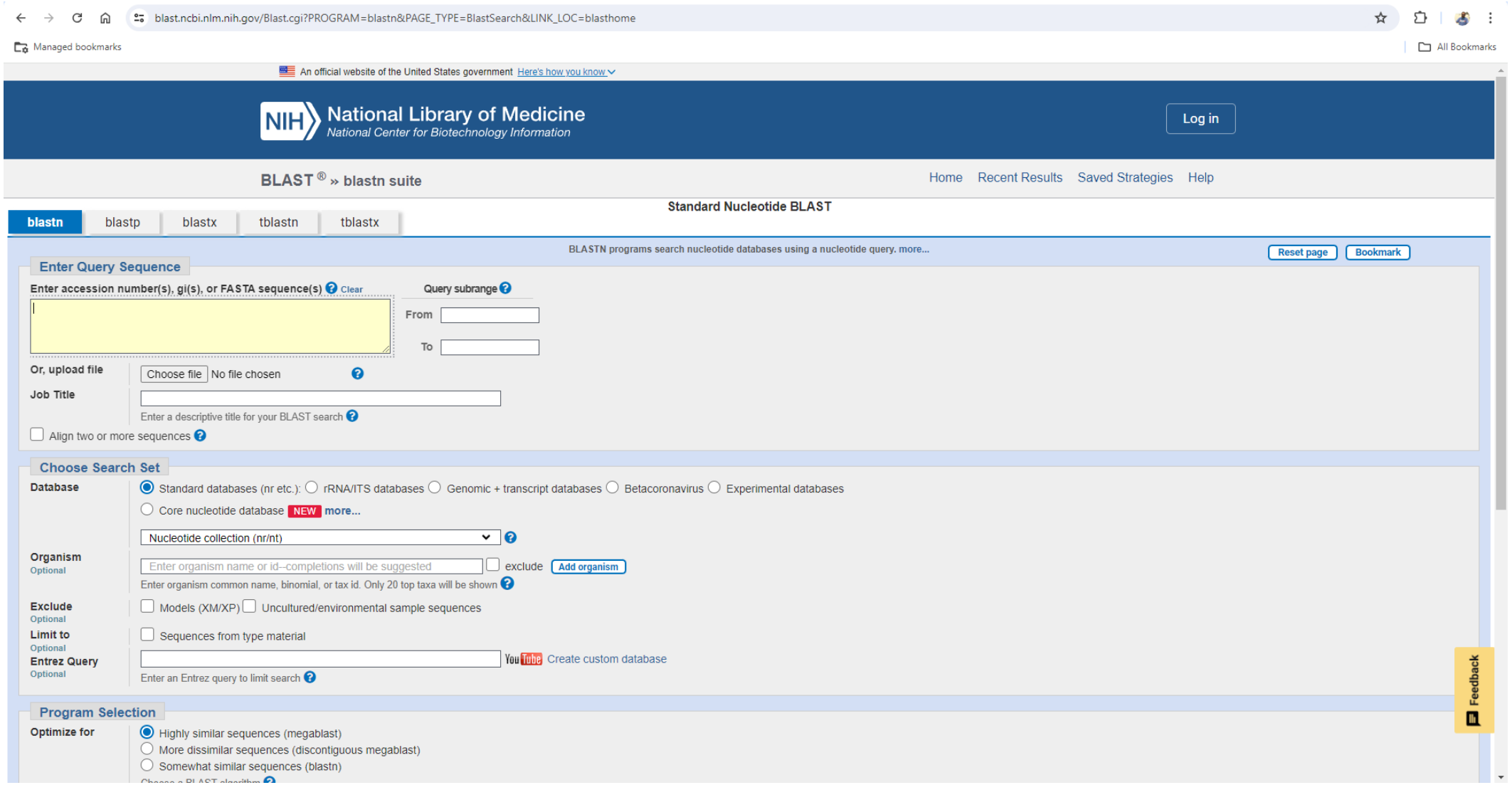
Ecological and Environmental Databases

- Global Biodiversity Information Facility (GBIF):
- - Global database providing data on biodiversity, supporting ecological research.
- - URL: https://www.gbif.org/

Microbiome Databases

- Earth Microbiome Project:
- - Analyzing microbial communities globally to understand microbial diversity and function.
- - URL: http://www.earthmicrobiome.org/
- SILVA NGS:
- - Resource for quality checked and aligned ribosomal RNA sequence data.
- - URL: https://www.arb-silva.de/
- Australian Microbiome Project:
- - Characterizing microbial diversity and ecosystem service provision in Australian ecosystems.
- - URL: https://www.australianmicrobiome.com/

NCBI BLAST


blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome
Managed bookmarks
All Bookmarks
An official website of the United States government Here's how you know
National Library of Medicine
National Center for Biotechnology Information
Log in
BLAST® » blastn suite
Home Recent Results Saved Strategies Help
Standard Nucleotide BLAST
blastn blastp blastx tblastn tblastx
BLASTN programs search nucleotide databases using a nucleotide query. more...
Reset page Bookmark
Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s) Clear
Query subrange
From
To
Or, upload file
Choose file No file chosen
Job Title
Enter a descriptive title for your BLAST search
Align two or more sequences
Choose Search Set
Database
Standard databases (nr etc.): rRNA/ITS databases Genomic + transcript databases Betacoronavirus Experimental databases
Core nucleotide database NEW more...
Nucleotide collection (nr/nt)
Organism
Optional
Enter organism name or id--completions will be suggested
exclude
Add organism
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown
Exclude
Optional
Models (XM/XP) Uncultured/environmental sample sequences
Limit to
Optional
Sequences from type material
Entrez Query
Optional
Create custom database
Enter an Entrez query to limit search
Program Selection
Optimize for
Highly similar sequences (megablast)
More dissimilar sequences (discontiguous megablast)
Somewhat similar sequences (blastn)
Feedback

Types of BLAST Programs

- • BLASTN (nucleotide-nucleotide BLAST)
- • BLASTP (protein-protein BLAST)
- • BLASTX (translated nucleotide-protein BLAST)
- • TBLASTN (protein-translated nucleotide BLAST)
- • TBLASTX (translated nucleotide-translated nucleotide BLAST)

How BLAST Works

- • The algorithm behind BLAST
- • Sequence alignment
- • Scoring systems (PAM, BLOSUM)
- • E-values and significance

Using NCBI BLAST

- • Accessing BLAST on the NCBI website
- • Inputting sequences
- • Choosing the right BLAST program
- • Setting parameters (e.g., database selection, filters)

Interpreting BLAST Results

- • Understanding the output format
- • Identifying high-scoring pairs (HSPs)
- • Reading alignment scores and E-values
- • Analyzing sequence similarities and differences

Homologous, paralogous and orthologous

- Homologous: Homologous sequences are sequences that share a common ancestor. This term broadly refers to sequences that are related by descent from a common ancestral DNA sequence. In simple words similar structure and function (proteins)

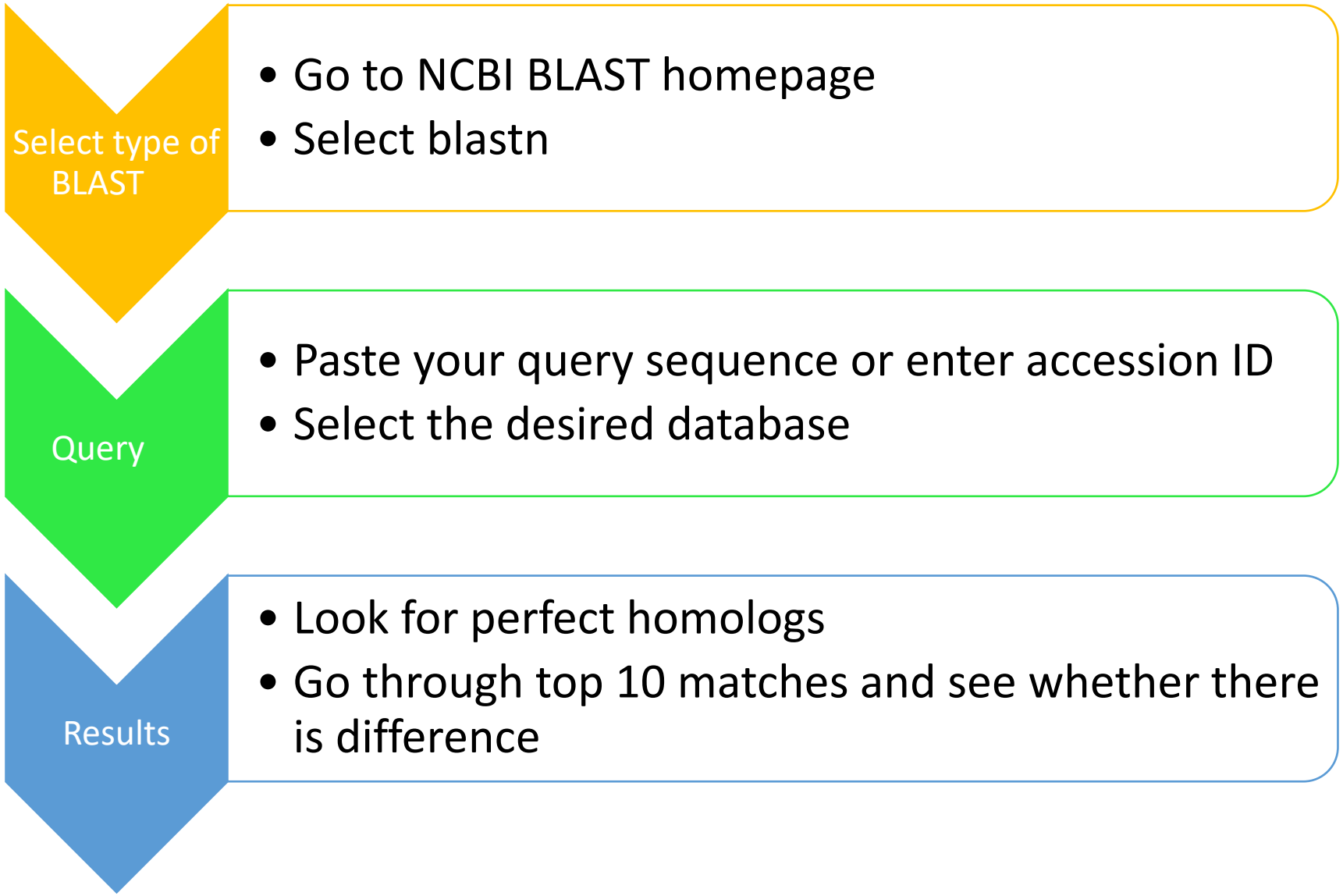
e.g. human haemoglobin and mouse haemoglobin

Orthologous: Orthologous sequences are homologous sequences that were separated by a speciation event. These sequences are found in different species and usually retain the same function.

E.g. human cytochrome c and yeast cytochrome c

Paralogous: Paralogous sequences are homologous sequences that were separated by a gene duplication event within the same species. These sequences can evolve new functions, even if related to the original one

Eg human haemoglobin and human myoglobin , both have resulted from gene duplication in a same invertebrate ancestor.


Select type of BLAST
• Go to NCBI BLAST homepage
• Select blastn
Query
• Paste your query sequence or enter accession ID
• Select the desired database
Results
• Look for perfect homologs
• Go through top 10 matches and see whether there is difference

Advanced BLAST Features

- • Customizing search parameters
- • Using specialized databases
- • BLAST+ and command-line usage
- • Batch BLAST

- Time for HANDS ON