NCBI patent sequences and BLAST

Peter Cooper
peter.cooper@nih.gov

Slides & Notes: https://go.usa.gov/xGRja
Today’s Topics

• NCBI introduction
• Sequences at NCBI
  • Patent Sequences
• Search systems (Entrez and BLAST)
• Entrez patent searches
• BLAST essentials
  • Standalone BLAST
  • Web BLAST
• Live demonstrations
What is NCBI?

Since 1988
- Establish public databases
- Research in computational biology
- Develop software tools for data analysis
- Disseminate biomedical information
Sources of sequences at NCBI

- Submitted nucleotide sequences and corresponding proteins
  - International Sequence Database Collaboration (INSDC) [ftp.ncbi.nlm.nih.gov/genbank/]
    - Assembled sequences
      - GenBank -- US Sequence Database at NCBI
      - European Nucleotide Archive at EBI
      - DNA Databank of Japan at NIG
    - Next-Gen sequencing reads
      - Sequence Read Archive (SRA)
      - 2.13 X 10^9 sequences, 9.854 X 10^{12} bases
      - Traditional, including Patent sequences, plus set based (WGS etc.)
      - 4.4 X 10^{16} bases (Yikes!)

- High quality curated DNA and protein records
  - NCBI Reference Sequences
  - Swiss-Prot
  - Proteins from PDB
Patent sequences at NCBI

**Scope**
- Granted data only, no application data
- Sequences from US patents submitted by USPTO
- Sequences from European and Japanese patents included through the INSDC collaboration (GenBank, ENA(EMBL), and DNA Database of Japan)
- Sequences from both **claims** and **exhibits**
- Often lacking features (genes, mRNA, coding regions, sources) due to format mismatch
  - New format standard coming that should improve this

47,109,909 nucleotide
7,067,455 protein
NCBI Search Systems

Entrez text search system
- 40-plus integrated databases
- Free text and database-specific fielded searches
- The PubMed search engine

BLAST -- Basic Local Alignment Search Tool
- Sequence similarity search tool
- Finds related nucleotide and protein sequences
  - Designed to find homologs
  - Used for other sequence analysis tasks
Entrez searching patent sequences

Example Entrez queries

gbdiv_pat[properties]  Nucleotide  Protein


gbdiv_pat[properties] AND genbank[filter]  Nucleotide  Protein

US 9260752[accession]  Nucleotide  Protein

cas9 AND gbdiv_pat[properties]  Nucleotide  Protein
What is BLAST?

- **Basic Local Alignment Search Tool**
- The most widely used sequence similarity search tool
- Finds high scoring **local alignments** between two sequences (protein or DNA)
- Includes a model of score distributions for random local alignments
- BLAST tells you about **non-chance similarities** between biological sequences.
  - If matches aren’t due to chance then they must be due to
    1. Homology
    2. Simple identification
- Web interface to databases at NCBI and other locations
- Standalone tool and BLAST-ready databases available for download

Standalone BLAST+ 2.10.1

Download BLAST programs and databases: ftp.ncbi.nlm.nih.gov/blast

- Local version of blast search programs (blastn, blastp, tblastn etc.)
- Utilities for working with and creating BLAST databases
- Flexible output
- Custom or NCBI databases
- Requires substantial investment in hardware to host large databases
- Dockerized version available that can be used on the cloud

Web BLAST Searching

- Query sequence or accession
- Database selection and limits
- Program flavor

NIH U.S. National Library of Medicine
National Center for Biotechnology Information

NCBI WORKSHOP
BLAST Patent Sequence DBs

- Patent sequences are not included in default BLAST databases (nt, nr)
- Databases are non-redundant
  - Identical sequences combined into one

Patent nucleotide database
38,653,273 sequences, Sept 13, 2020

Patent protein database
2,760,814 sequences, Sept 13, 2020
Live Demonstrations

- CRISPR/Cas9 protein patents
  - BLAST search to find patent proteins
  - Overview of BLAST results
  - Linking to protein
  - Finding all sequences for a patent number
  - Identical proteins
  - Finding homologs in unannotated sequences
BLAST Help

Help desk: blast-help@ncbi.nlm.nih.gov

Cloud docs: https://github.com/ncbi/blast_plus_docs