NCBI patent sequences and BLAST

Peter Cooper

peter.cooper@nih.gov

Slides & Notes: https://go.usa.gov/xGRja

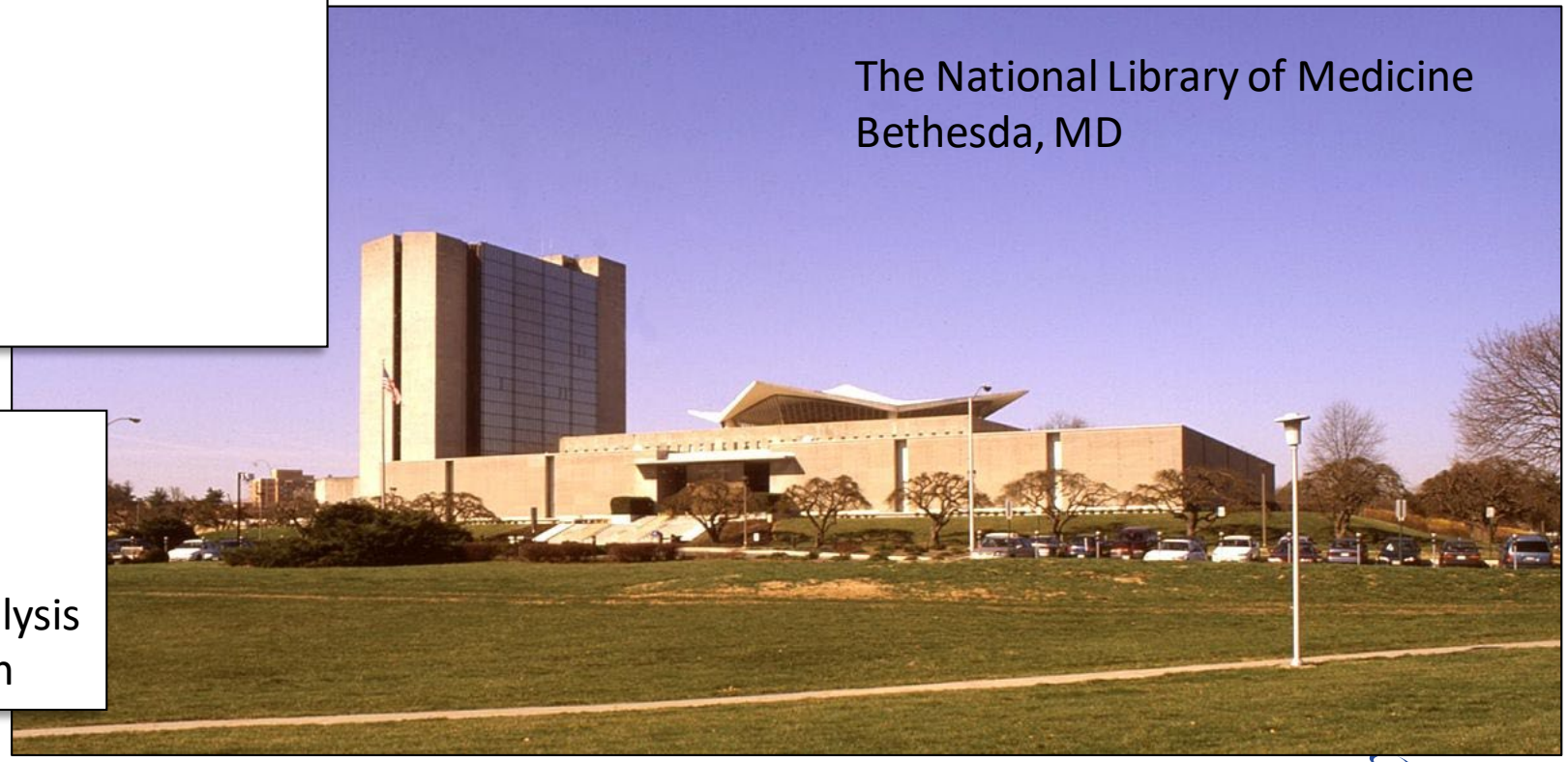
U.S. National Library of Medicine
National Center for Biotechnology Information

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?

www	
.	Dot
ncbi	National Center for Biotechnology Information
.	Dot
nlm	National Library of Medicine
.	Dot
nih	National Institutes of Health
.	Dot
gov	US Federal Government



The National Library of Medicine
Bethesda, MD

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 Colloration (**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)

2.13×10^9 sequences, 9.854×10^{12} bases
Traditional, including **Patent** sequences, plus set based (WGS etc.)

4.4×10^{16} bases (Yikes!)

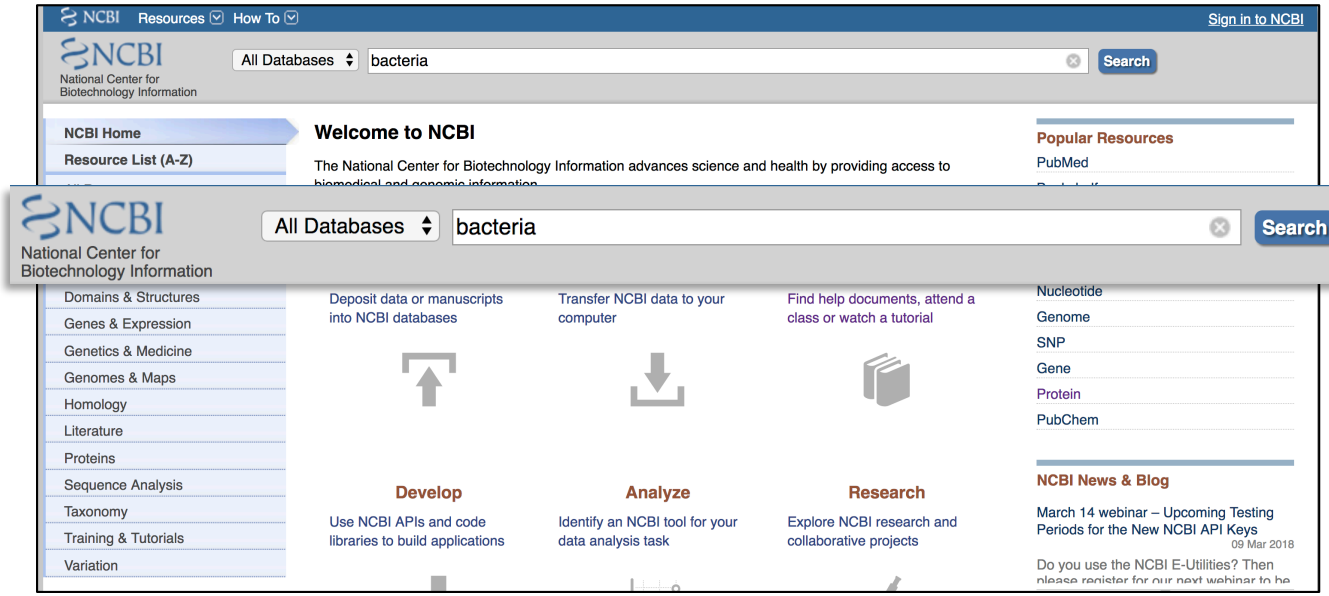
- High quality curated DNA and protein records
  - NCBI Reference Sequences
  - Swiss-Prot
  - Proteins from PDB

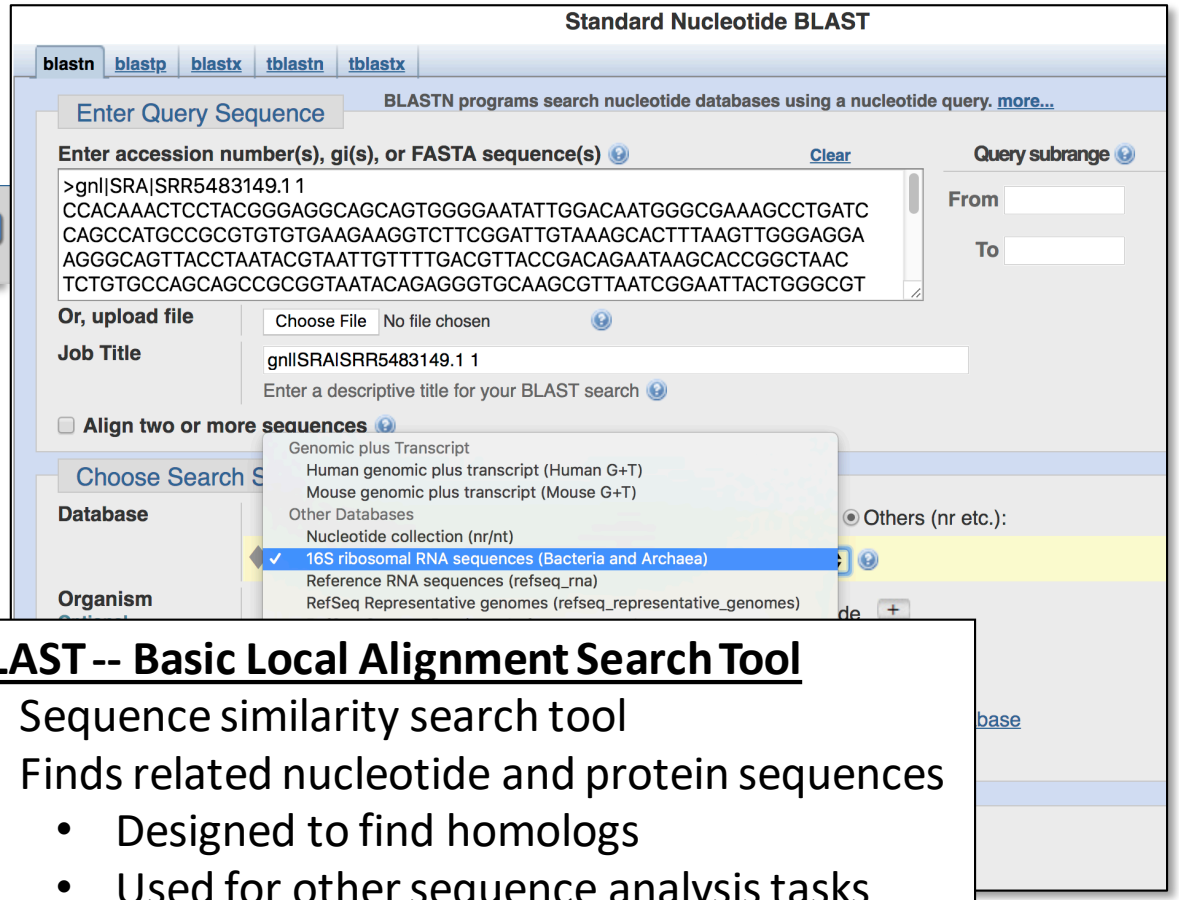
Patent sequences at NCBI

47,109,909 nucleotide
7,067,455 protein

- 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

NCBI Search Systems





Entrez text search system

40-plus integrated databases

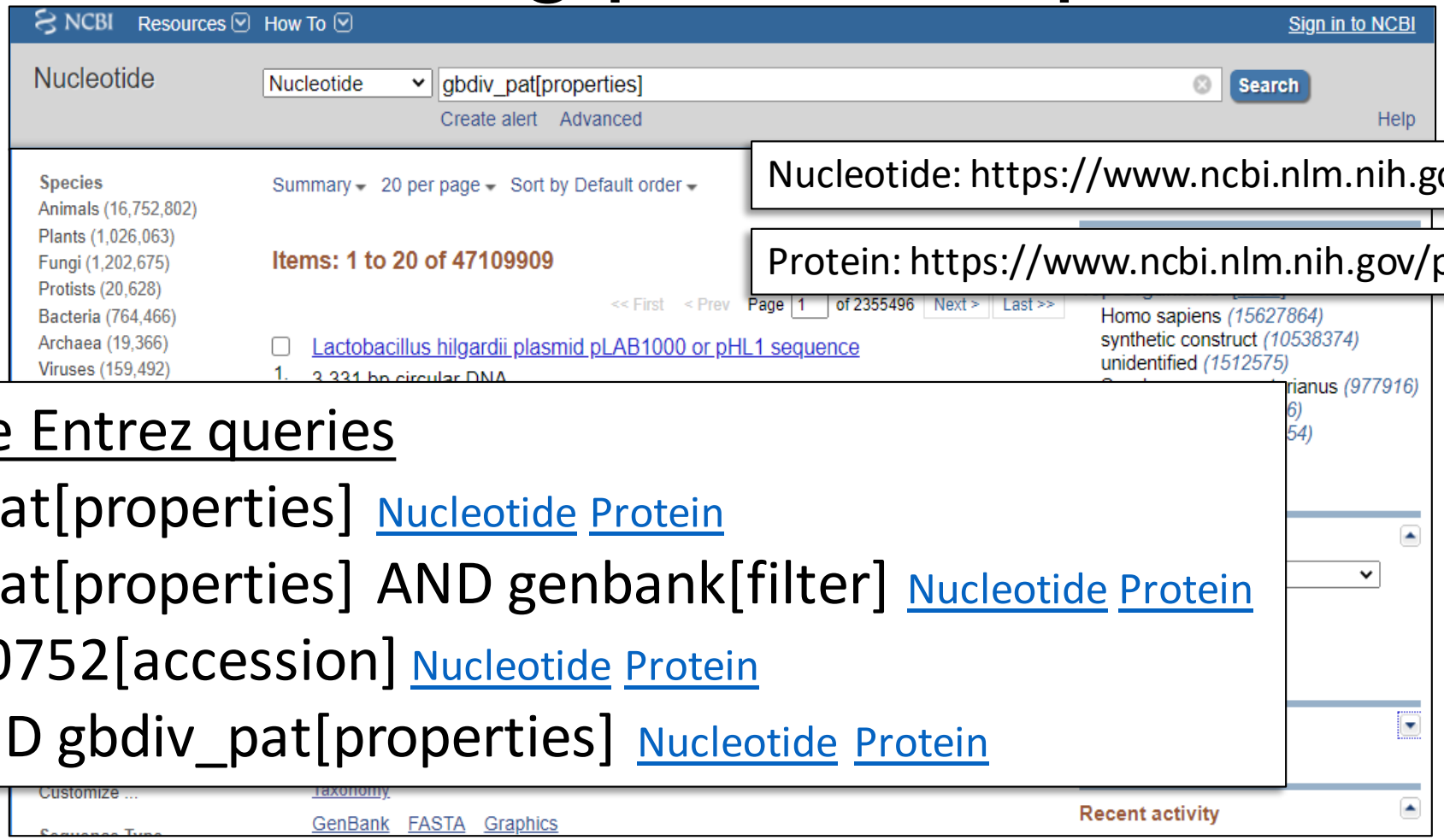
Free text and database-specific fielded searches

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



Nucleotide: https://www.ncbi.nlm.nih.gov/nuccore/

Protein: https://www.ncbi.nlm.nih.gov/protein/

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?

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215(3):403-410. PMID: 2231712

- 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

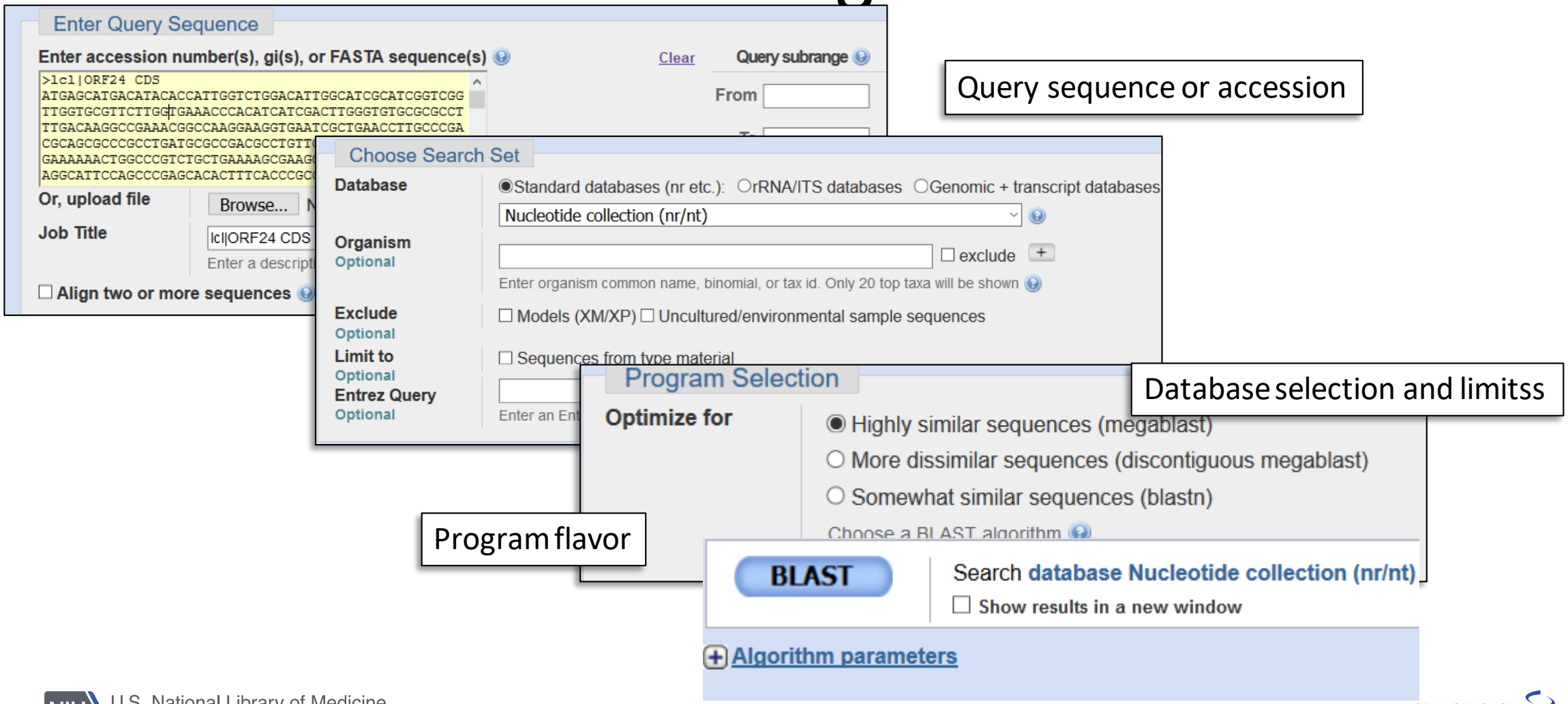
Download BLAST programs and databases: ftp.ncbi.nlm.nih.gov/blast

- 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

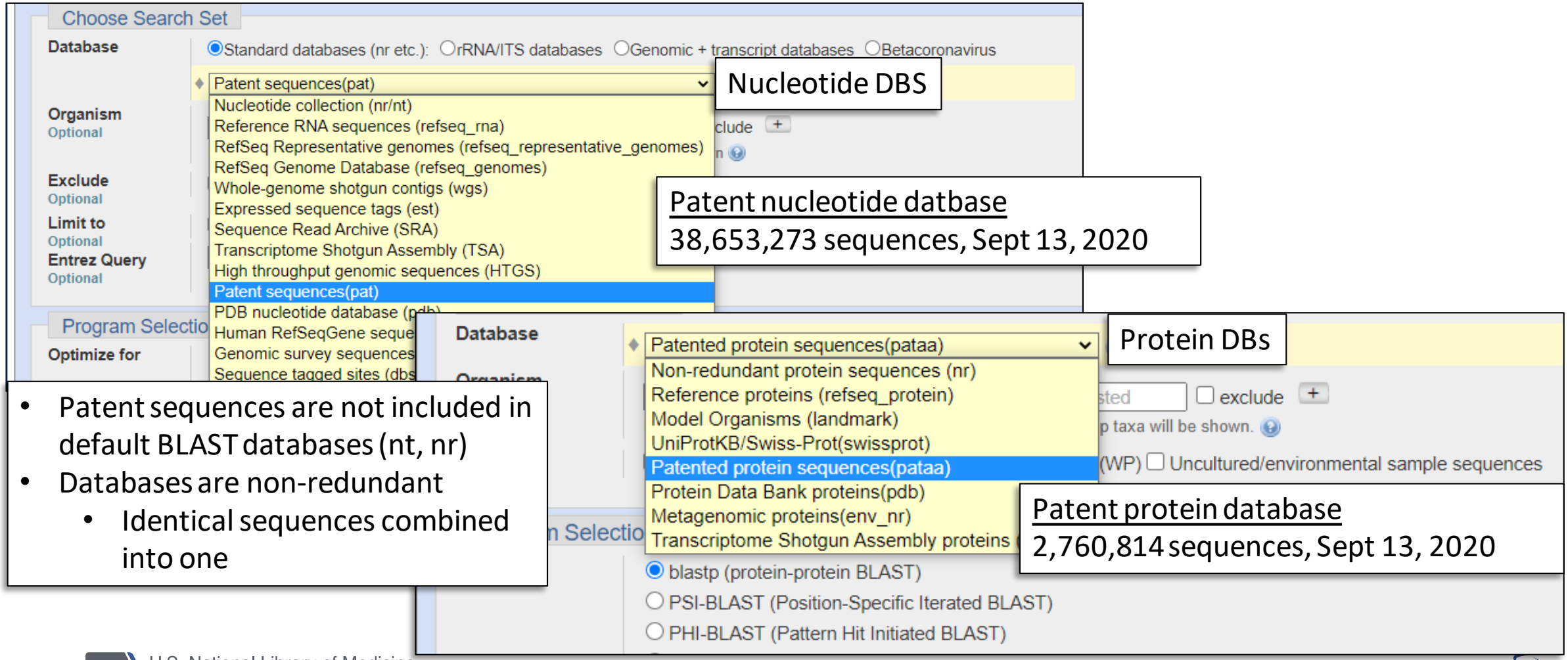
https://hub.docker.com/r/ncbi/blast

https://github.com/ncbi/blast_plus_docs

Web BLAST Searching



BLAST Patent Sequence DBs



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

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

BLAST documentation

Getting Started

- Guide to BLAST home and search pages
- Blast report description
- Blast topics

About BLAST

- Frequently Asked Questions
- NCBI Handbook: BLAST
- The Statistics of Sequence Similarity Scores
- BLAST glossary
- References
- Blast+ Command Line Applications User Manual
- BLAST News directory

Getting Help

- Write to the help desk
- Mailing list
- YouTube BLAST tutorials

Other BLAST information

- Download BLAST Software and Databases
- Developer information
- BLAST Searches at a Cloud Provider

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

https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs

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