

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?

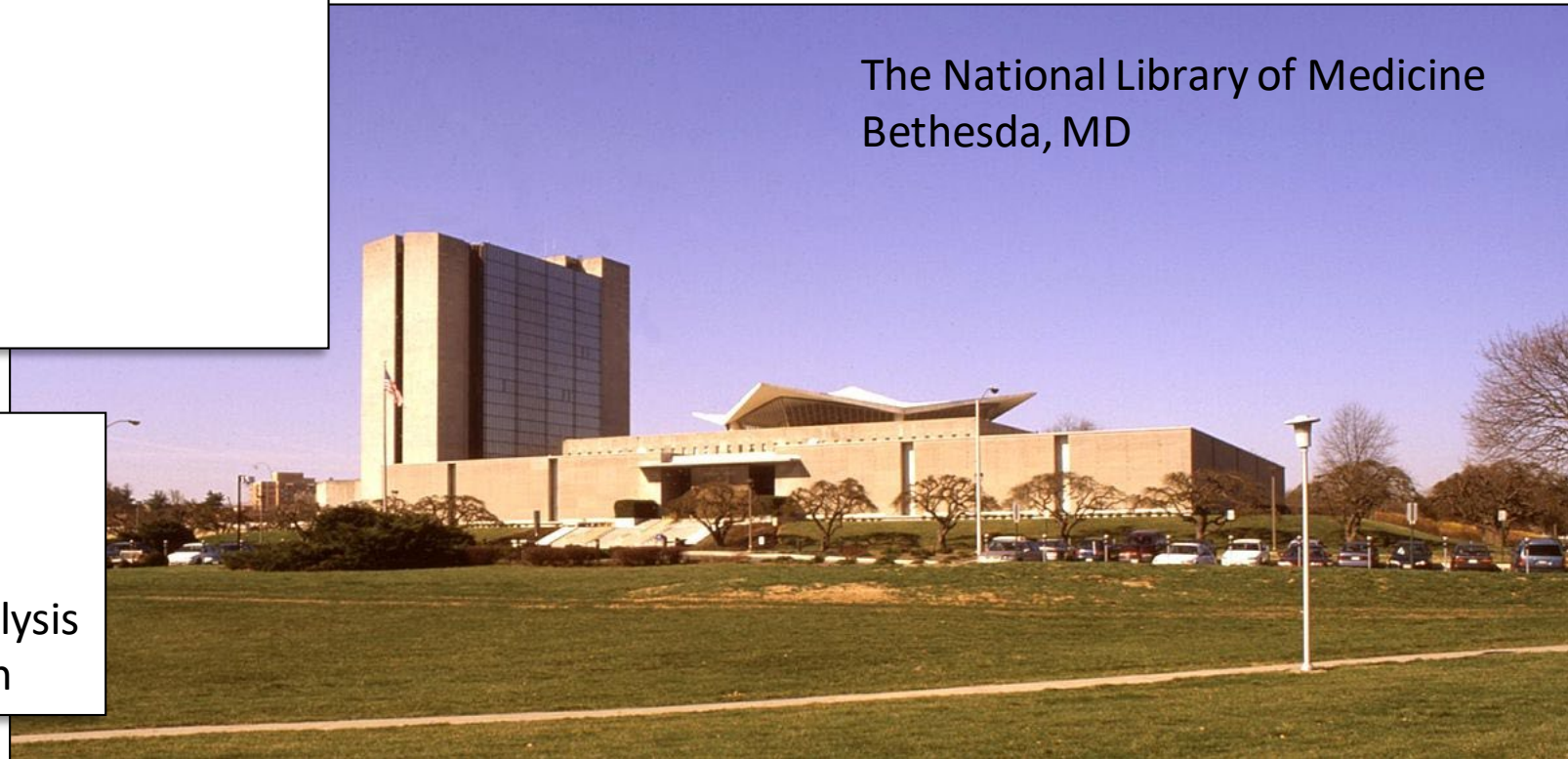
www

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

Since 1988

- Establish public databases
- Research in computational biology
- Develop software tools for data analysis
- Disseminate biomedical information

The National Library of Medicine
Bethesda, MD



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

2.13 X 10⁹ sequences, 9.854 X 10¹² bases
Traditional, including **Patent** sequences,
plus set based (WGS etc.)

4.4 X 10¹⁶ bases (Yikes!)

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

The screenshot shows the NCBI homepage with a search bar at the top containing the word "bacteria". Below the search bar, there are navigation links for "NCBI Home" and "Resource List (A-Z)". The main content area features a "Welcome to NCBI" message and several sections: "Popular Resources" (including PubMed), "Develop" (Deposit data or manuscripts into NCBI databases), "Analyze" (Transfer NCBI data to your computer), "Research" (Find help documents, attend a class or watch a tutorial), and "NCBI News & Blog" (March 14 webinar – Upcoming Testing Periods for the New NCBI API Keys).

Entrez text search system

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

The screenshot shows the "Standard Nucleotide BLAST" interface. The "Enter Query Sequence" section contains a text input field with the following sequence: `>gnl|SRA|SRR5483149.1.1
CCACAACTCCTACGGGAGGCAGCAGTGGGGAATATTGGACAATGGGCGAAAGCCTGATC
CAGCCATGCCGCGTGTGAAGAAGGTCTTCGGATTGTAAAGCACTTAAAGTTGGGAGGA
AGGGCAGTTACCTAATACGTAATTGTTTTGACGTTACCGACAGAATAAGCACCGGCTAAC
TCTGTGCCAGCAGCCGCGTAATACAGAGGGTGAAGCGTTAATCGGAATTACTGGGCGT`. The "Job Title" field contains "gnl|SRA|SRR5483149.1.1". The "Choose Search Set" dropdown menu is open, showing options like "Genomic plus Transcript", "Human genomic plus transcript (Human G+T)", "Mouse genomic plus transcript (Mouse G+T)", "Other Databases", "Nucleotide collection (nr/nt)", "16S ribosomal RNA sequences (Bacteria and Archaea)", "Reference RNA sequences (refseq_rna)", and "RefSeq Representative genomes (refseq_representative_genomes)". The "Database" field is set to "16S ribosomal RNA sequences (Bacteria and Archaea)".

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

NCBI Resources How To Sign in to NCBI

Nucleotide Nucleotide gbdiv_pat[properties] Search

Create alert Advanced Help

Species Summary 20 per page Sort by Default order

Animals (16,752,802)

Plants (1,026,063)

Fungi (1,202,675)

Protists (20,628)

Bacteria (764,466)

Archaea (19,366)

Viruses (159,492)

Items: 1 to 20 of 47109909

<< First < Prev Page 1 of 2355496 Next > Last >>

[Lactobacillus hilgardii plasmid pLAB1000 or pHL1 sequence](#)

1. 3,331 bp circular DNA

Homo sapiens (15627864)

synthetic construct (10538374)

unidentified (1512575)

rianius (977916)

6)

54)

Customize ... Taxonomy

GenBank FASTA Graphics

Recent activity

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](https://pubmed.ncbi.nlm.nih.gov/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

The image shows a screenshot of the NCBI BLAST web interface with several callout boxes highlighting key sections:

- Query sequence or accession:** Points to the "Enter Query Sequence" section where the accession number "lc1|ORF24 CDS" is entered.
- Database selection and limits:** Points to the "Choose Search Set" section, specifically the "Database" dropdown menu which is set to "Nucleotide collection (nr/nt)".
- Program flavor:** Points to the "Program Selection" section, where "Highly similar sequences (megablast)" is selected under "Optimize for".
- Algorithm parameters:** Points to the bottom section containing the "BLAST" button and the "Show results in a new window" checkbox.

BLAST Patent Sequence DBs

The screenshot shows the 'Choose Search Set' section of the BLAST web interface. It features a 'Database' dropdown menu with various options. Two callout boxes provide details about the 'Patent sequences(pat)' database: 'Nucleotide DBs' and 'Patent nucleotide database' with 38,653,273 sequences as of Sept 13, 2020. Another callout box identifies 'Patented protein sequences(pataa)' as 'Protein DBs' and 'Patent protein database' with 2,760,814 sequences as of Sept 13, 2020. A third callout box on the left lists key characteristics of these patent databases.

Choose Search Set

Database

- Standard databases (nr etc.):
- rRNA/ITS databases
- Genomic + transcript databases
- Betacoronavirus

Patent sequences(pat) (Nucleotide DBs)

- Nucleotide collection (nr/nt)
- Reference RNA sequences (refseq_rna)
- RefSeq Representative genomes (refseq_representative_genomes)
- RefSeq Genome Database (refseq_genomes)
- Whole-genome shotgun contigs (wgs)
- Expressed sequence tags (est)
- Sequence Read Archive (SRA)
- Transcriptome Shotgun Assembly (TSA)
- High throughput genomic sequences (HTGS)
- Patent sequences(pat)**
- PDB nucleotide database (pdb)

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

Protein DBs

Patented protein sequences(pataa)

- Non-redundant protein sequences (nr)
- Reference proteins (refseq_protein)
- Model Organisms (landmark)
- UniProtKB/Swiss-Prot (swissprot)
- Patented protein sequences(pataa)**
- Protein Data Bank proteins (pdb)
- Metagenomic proteins (env_nr)
- Transcriptome Shotgun Assembly proteins

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

Program Selection

- blastp (protein-protein BLAST)
- PSI-BLAST (Position-Specific Iterated BLAST)
- PHI-BLAST (Pattern Hit Initiated BLAST)

Optional filters:

- Organism: include exclude
- Taxonomy: taxa will be shown
- Uncultured/environmental sample sequences: (WP)

Optional filters:

- Exclude: Optional
- Limit to: Optional
- Entrez Query: Optional

Optimize for

- Human RefSeqGene sequences
- Genomic survey sequences
- Sequence tagged sites (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