

**NCBI resources II:**  
**web-based tools** and ftp  
resources

Yanbin Yin

Most materials are downloaded from <ftp://ftp.ncbi.nih.gov/pub/education/>

# Outline

- Tools
  - BLAST
  - Specialized BLAST
  - GEO
- ftp download
- Hands on exercise

# References

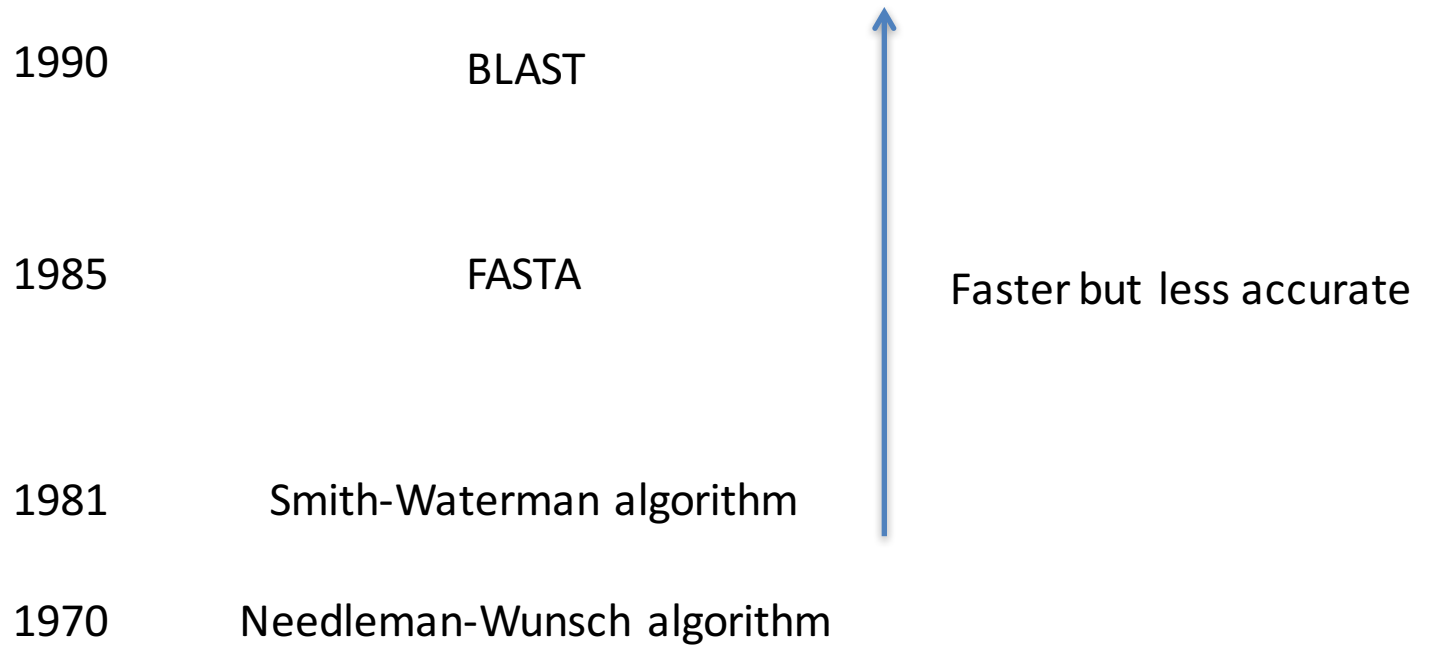
NCBI discovery workshops

[ftp://ftp.ncbi.nih.gov/pub/education/discovery\\_workshops/NLM/2012/Sept2012/](ftp://ftp.ncbi.nih.gov/pub/education/discovery_workshops/NLM/2012/Sept2012/)

[http://homepages.ulb.ac.be/~dgonze/TEACHING/stat\\_scores.pdf](http://homepages.ulb.ac.be/~dgonze/TEACHING/stat_scores.pdf)

[http://www.bioinformatics.wsu.edu/bioinfo\\_course/notes/lecture6.pdf](http://www.bioinformatics.wsu.edu/bioinfo_course/notes/lecture6.pdf)

# Evolution of pairwise alignment tools



# Basic Local Alignment Search Tool

- Widely used similarity search tool
- Heuristic approach based on Smith Waterman algorithm
- Finds best local alignments
- Provides statistical significance
- All combinations (DNA/Protein) query and database
  - DNA vs DNA
  - DNA translation vs Protein
  - Protein vs Protein
  - Protein vs DNA translation
  - DNA translation vs DNA translation
- www, standalone, and network client

# Why is BLAST Necessary?

§ Theoretically, one could perform a Global (or Local) Alignment between a query sequence and each protein or DNA sequence in a database

- Such an approach would be very computationally intensive and not practical for most purposes

§ BLAST approximates this methods in a heuristic

- BLAST is significantly faster than other heuristic methods
- BLAST is also more sensitive and selective than other heuristics

§ BLAST disadvantages:

- Misses some homology relations
- Does not guarantee optimal alignment

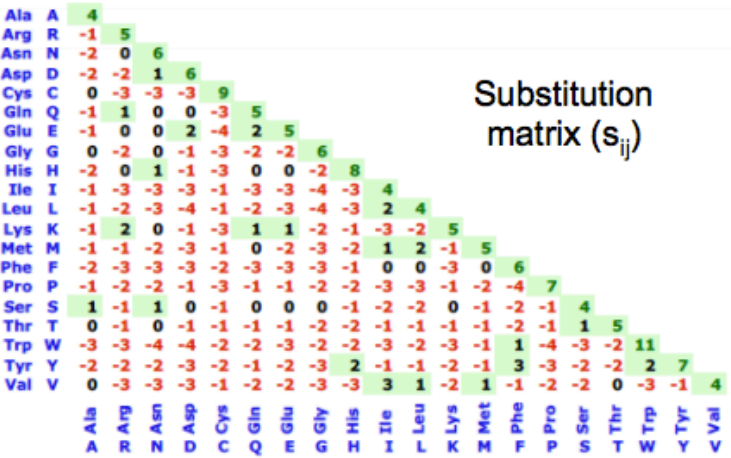
**Score:** A number used to assess the biological relevance of a finding.

In the context of sequence alignments, a score is a numerical value that describes the overall quality of an alignment. Higher numbers correspond to higher similarity. The score scale depends on the scoring system used (substitution matrix, gap penalty).

$$S = \sum_{i=1}^L s_{r_1,i,r_2,i}$$

Example:

<b>R</b>	<b>L</b>	<b>A</b>	<b>S</b>	<b>V</b>	<b>-</b>	<b>E</b>	<b>T</b>	<b>D</b>	<b>M</b>	<b>W</b>	<b>T</b>	<b>P</b>	<b>L</b>	<b>T</b>	<b>L</b>	<b>R</b>	<b>Q</b>	<b>H</b>
.		.		:		:		.	:			.		.	.			
<b>T</b>	<b>L</b>	<b>T</b>	<b>S</b>	<b>L</b>	<b>A</b>	<b>Q</b>	<b>T</b>	<b>T</b>	<b>L</b>	<b>-</b>	<b>-</b>	<b>K</b>	<b>A</b>	<b>H</b>	<b>L</b>	<b>G</b>	<b>T</b>	<b>H</b>
-1	+4	+0	+4	+1	-4	+2	+5	-1	+2	-4	-1	-1	-1	-2	+4	-2	-1	+8 = 12



gap penalty (s<sub>i,-</sub>)

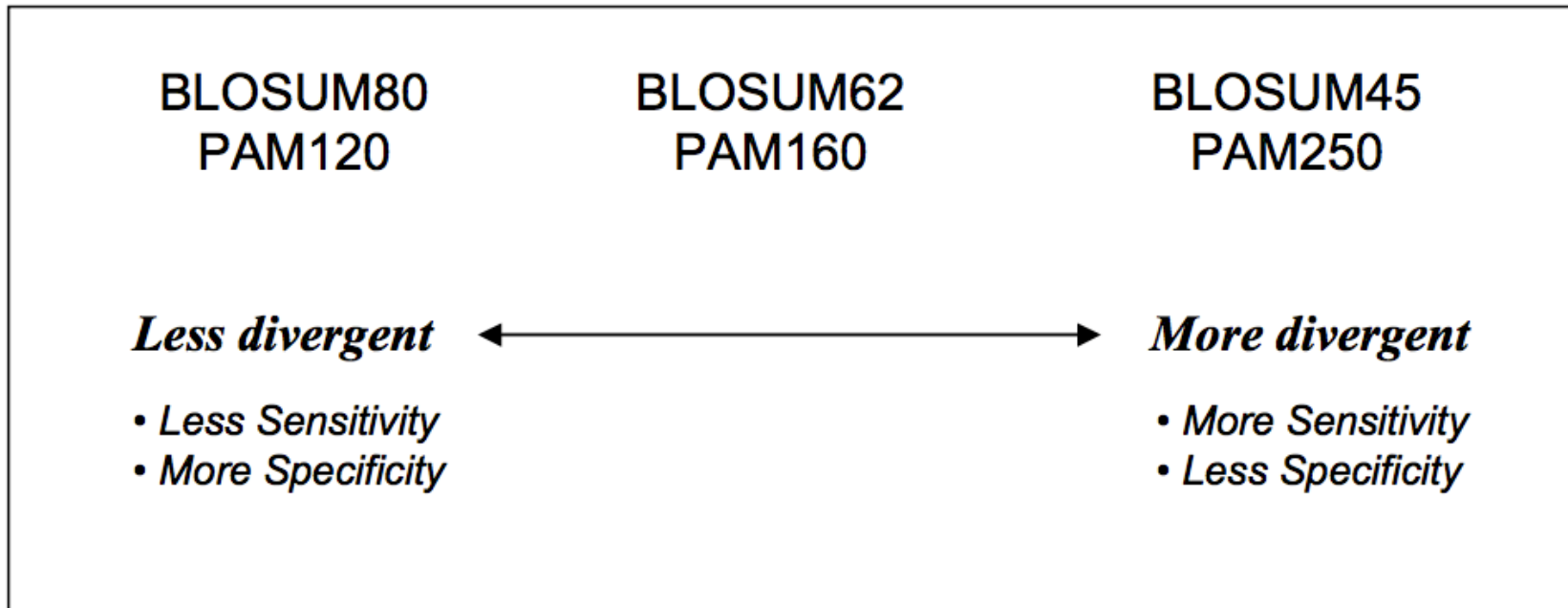
- gap opening -4
- gap extension -1
- end gap 0

# Matrix Choice

§ Good options include BLOSUM62 and PAM250

§ For PAM(X), higher X detects more divergent sequences

§ For BLOSUM(Y), lower Y detects more divergent sequences





# E-values

- § Scores are reported by BLAST for each high-scoring segment pair (HSP) as E-values
- § E-values approximate the number of HSPs with score **S** (or greater) that are expected by chance (i.e. not relevant)
- § E-values are calculated using the following formula:

$$E(S) = Kmne^{-\lambda S}$$

$K$  = estimated parameter

$m$  = total length of sequences in database

$n$  = length of query sequence

$\lambda$  = estimated parameter

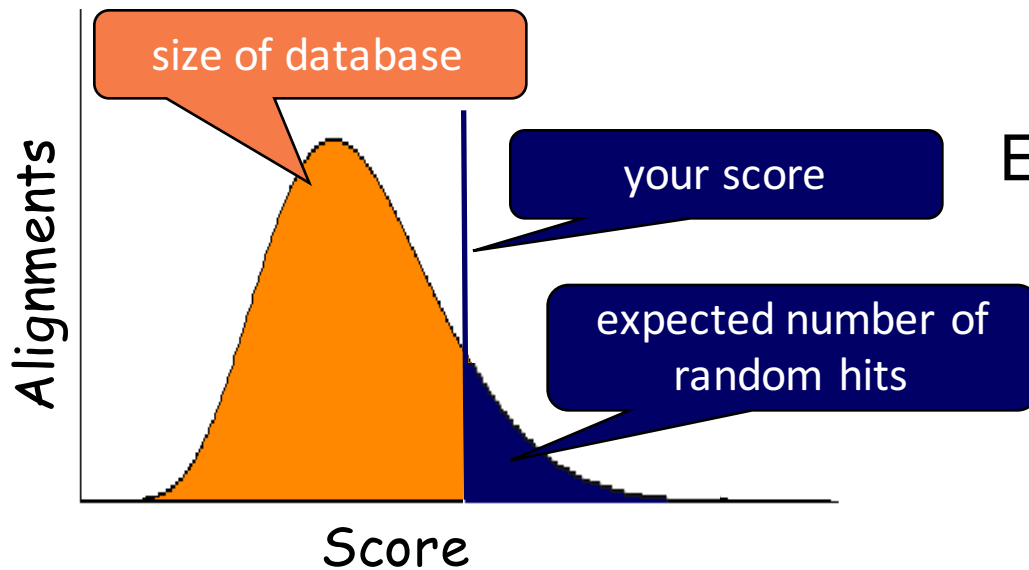
$S$  = score of the high-scoring segment pair (HSP)

# Local Alignment Statistics

High scores of local alignments between two random sequences follow the Extreme Value Distribution

## Expect Value

$E$  = number of database hits you expect to find by chance



$$E = Kmne^{-\lambda S} \quad \text{or} \quad E = mn2^{-S'}$$

$K$  = scale for search space  
 $\lambda$  = scale for scoring system  
 $S'$  = bitscore =  $(\lambda S - \ln K) / \ln 2$

(applies to ungapped alignments)

# Local Alignment Scoring: Protein

Number of Chance Alignments =  $4 \times 10^{-50}$

Score = 176 bits (447), Expect = 4e-50, Method: Compositional matrix adjust.  
 Identities = 98/232 (42%), Positives = 139/232 (60%), Gaps = 14/232 (6%)

```

Query   30   MAKVLTLELYKKLRDKETPSGFTVDDVIQTGV--DNP GHPFIMTVGCVAGDEESYEVFKE   87
          + K LT +L+++ +D+   GF+   I +G   N G       VG AG +SY F
Sbjct   26   LQKCLTKDLWEQCKDRRDKYGFSPKQAI FSGSKWTNSG-----VGVYAGSHDSYYAFAP   79

```

Query 80 ...  
 Sbjct 80 ...

K	K	Q	Gap
K +5	E +1	F -3	- (11 + 4 (1)) = -14

```

Query   138   +R ER+ +E L   AL   TGE KGKYY L++M++
Sbjct   138   AVTRKERKEIEHLVTSALGEFTGELKGKYY SLETMSD

```

```

Query   205   SGMARDWPDARGIWHNDNKSFLVWVNEEDHLRVISM EKGGNMKEVFRRFCVG   256
          +G+ RDWP+ARGI+HND K+FLVWVNEED LR+ISM+ G N+ EVF+R V
Sbjct   197   AGLERDWPEARGIFHND AKTFLVWVNEEDQLRIISMQAGSNILEVFKRLSVA   248

```

Scores from BLOSUM62, a position independent matrix

# Local Alignment Scoring: Nucleotide

Number of Chance Alignments =  $2 \times 10^{-73}$

Score = 288 bits (318), Expect = 2e-73  
 Identities = 262/325 (81%), Gaps = 8/325 (2%)  
 Strand=Plus/Plus

Query	1923	TCAGCCTACCATGAGAATAAGAGAAAGA-AAATGAAGATCAAAGCTTATTCATCTGTTT	1981
Sbjct	33774	TCAGACTACCCTGAGAATAAGAGAAAGAGAAATGAAGACCTAGA-CTTATCCATCTCTTT	33832
Query	1982	TTCTTTTTCGTTGGTGTAAAGCCAACACCCTGTCTAAAAAACATAAATTTCTTTAATCAT	2041
Sbjct		TGACAAAATTTCTTTAAATAT	33892
Query	2042	TTTGCCTCTTTTC	2100
Sbjct	33893	TTTGCCTCTTTTCTCTGTGCTACAATTAATAAAAAAATGAAAAGAATCTAATTTAATTGT	33952
Query	2101	ACAGCACTGTTA-T	2159
Sbjct	33953	CTATGACTGTTATT	34012
Query	2160	AAGTTCCAGTGTTCT	2219
Sbjct	34013	AAATTCCACTATTCTCTCTTTCCCTATTTCAATGGAGGACTTCTAGTTCCTTCTGGATTA	34072
Query	2220	AT----TAAATAAATCATTAACT	2240
Sbjct	34073	ATTGCATAAAAGAAACATTAATACT	34097

Match=+2

Mismatch=-3

Gap

$$-(5 + 4(2)) = -13$$

# BLAST and BLAST-like programs

- Traditional BLAST (formerly blastall) **nucleotide, protein, translations**
  - **blastn** nucleotide query vs. nucleotide database
  - **blastp** protein query vs. protein database
  - **blastx** nucleotide query vs. protein database
  - **tblastn** protein query vs. translated nucleotide database
  - **tblastx** translated query vs. translated database
- Megablast **nucleotide only**
  - **Contiguous megablast**
    - Nearly identical sequences
  - **Discontiguous megablast**
    - Cross-species comparison

# Position-specific BLAST Programs

(protein only)

- **Position Specific Iterative BLAST (PSI-BLAST)**

Automatically generates a position specific score matrix (PSSM)

- **Position-Hit Initiated BLAST (PHI-BLAST)**

Focuses search around pattern (motif)

- **Domain Enhanced Lookup Time Accelerated (DELTA) BLAST**

Uses domain PSSM in first round of search

- **Reverse PSI-BLAST (RPS-BLAST)**

Searches a database of PSI-BLAST PSSMs

Conserved Domain Database Search

# PSSM: frequencies

```

GHEGVGKVVKIG
GHEKKGYFEDRG
GHEGYGGRSRGG
GHEFEGPKGCGA
GHELRGTTFMPA
  
```



	1	2	3	4	5	6	7	8	9	10	11	12
A	0	0	0	0	0	0	0	0	0	0	0	2
C	0	0	0	0	0	0	0	0	0	0	1	0
D	0	0	0	0	0	0	0	0	0	1	0	0
E	0	0	5	0	1	0	0	0	1	0	0	0
F	0	0	0	1	0	0	0	1	1	0	0	0
G	5	0	0	2	0	5	1	0	1	0	2	3
H	0	5	0	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	0	0	1	0
K	0	0	0	1	1	0	1	1	0	1	0	0
L	0	0	0	1	0	0	0	0	0	0	0	0
M	0	0	0	0	0	0	0	0	0	1	0	0
N	0	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	1	0	0	0	1	0
Q	0	0	0	0	0	0	0	0	0	0	0	0
R	0	0	0	0	1	0	0	1	0	1	1	0
S	0	0	0	0	0	0	0	0	1	0	0	0
T	0	0	0	0	0	0	1	1	0	0	0	0
V	0	0	0	0	1	0	0	1	1	0	0	0
W	0	0	0	0	0	0	0	0	0	0	0	0
Y	0	0	0	0	1	0	1	0	0	0	0	0

$$f_{A,1} = \frac{0}{5} = 0, f_{G,1} = \frac{5}{5} = 1, \dots$$


$$f_{A,2} = \frac{0}{5} = 0, f_{H,2} = \frac{5}{5} = 1, \dots$$

...

$$f_{A,12} = \frac{2}{5} = 0.4, f_{G,12} = \frac{3}{5} = 0.6, \dots$$

# Non-redundant protein

Choose Search Set


Database: Non-redundant protein sequences (nr) 


Organism: Optional

Exclude: Optional

Entrez Query: Optional

Enter an Entrez query

suggested  Exclude 

tax id. Only 20 top taxa will be shown. 

environmental sample sequences

**nr** (non-redundant protein sequences)

- GenBank CDS translations
- NP\_, XP\_ **refseq\_protein**
- Outside Protein
  - PIR, **Swiss-Prot**, PRF
  - **PDB** (sequences from structures)

**pat** protein patents

**env\_nr** metagenomes

(environmental samples)

**Services**  
blastp  
blastx



# Nucleotide Databases: Traditional

The image shows a screenshot of the NCBI 'Choose Search Set' interface. On the left, there are several sections: 'Database', 'Organism Optional', 'Exclude Optional', and 'Entrez Query Optional'. The 'Database' section has three radio buttons: 'Human genomic + transcript', 'Mouse genomic + transcript', and 'Others (nr etc.):'. The 'Others (nr etc.):' option is selected. A dropdown menu is open under 'Others (nr etc.):', showing a list of 'Other Databases'. The first item is 'Nucleotide collection (nr/nt)', which is highlighted in blue. Below it are several other database options: 'Reference RNA sequences (refseq\_rna)', 'Reference genomic sequences (refseq\_genomic)', 'NCBI Genomes (chromosome)', 'Expressed sequence tags (est)', 'Genomic survey sequences (gss)', 'High throughput genomic sequences (HTGS)', 'Patent sequences(pat)', 'Protein Data Bank (pdb)', 'Human ALU repeat elements (alu\_repeats)', 'Sequence tagged sites (dbsts)', 'Whole-genome shotgun contigs (wgs)', 'Transcriptome Shotgun Assembly (TSA)', and '16S ribosomal RNA sequences (Bacteria and Archaea)'. To the right of the dropdown menu, there is a 'Services' box containing the text: 'Services', 'blastn', 'tblastn', and 'tblastx'.

**Choose Search Set**

**Database**  Human genomic + transcript  Mouse genomic + transcript  Others (nr etc.):

**Organism** Optional

**Exclude** Optional

**Entrez Query** Optional

**Other Databases**

- Nucleotide collection (nr/nt)
- Reference RNA sequences (refseq\_rna)
- Reference genomic sequences (refseq\_genomic)
- NCBI Genomes (chromosome)
- Expressed sequence tags (est)
- Genomic survey sequences (gss)
- High throughput genomic sequences (HTGS)
- Patent sequences(pat)
- Protein Data Bank (pdb)
- Human ALU repeat elements (alu\_repeats)
- Sequence tagged sites (dbsts)
- Whole-genome shotgun contigs (wgs)
- Transcriptome Shotgun Assembly (TSA)
- 16S ribosomal RNA sequences (Bacteria and Archaea)

**Services**

- blastn
- tblastn
- tblastx

# Nucleotide Databases: Traditional

Databases are mostly non-overlapping

- **nr (nt)**

- Traditional GenBank
- NM\_ and XM\_ RefSeqs
  - refseq\_rna

- **NCBI Genomes**

- NC\_ RefSeqs
- GenBank Chromosomes

- **dbest**

- EST Division
  - non-human, non-mouse ests

- **htgs**

- HTG division

- **gss**

- GSS division

- **wgs**

- whole genome shotgun contigs

- **tsa**

- transcriptome shotgun assembly

- **16S microbial**

- Selected 16S sequences (targeted loci)

# Specialized BLAST Pages

## Specialized BLAST

Choose a type of specialized search (or database name in parentheses.)

- ▣ Make specific primers with [Primer-BLAST](#)
- ▣ Search [trace archives](#)
- ▣ Find [conserved domains](#) in your sequence (cds)
- ▣ Find sequences with similar [conserved domain architecture](#) (cdart)
- ▣ Search sequences that have [gene expression profiles](#) (GEO)
- ▣ Search [immunoglobulins](#) (IgBLAST)
- ▣ Search for [SNPs](#) (snp)
- ▣ Screen sequence for [vector contamination](#) (vecscreen)
- ▣ [Align](#) two (or more) sequences using BLAST (bl2seq)
- ▣ Search [protein](#) or [nucleotide](#) targets in PubChem BioAssay
- ▣ Search SRA [transcript and genomic libraries](#)
- ▣ Constraint Based Protein [Multiple Alignment Tool](#)
- ▣ Needleman-Wunsch [Global Sequence Alignment Tool](#)
- ▣ Search [RefSeqGene](#)
- ▣ Search [WGS sequences](#) grouped by organism

# Hands on exercise 1

blastn and megablast

- NCBI Home
- Resource List (A-Z)
- All Resources
- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Homology
- Literature
- Proteins
- Sequence Analysis
- Taxonomy
- Training & Tutorials
- Variation

### Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [Research](#) | [RSS Feeds](#)

### Get Started

- [Tools](#): Analyze data using NCBI software
- [Downloads](#): Get NCBI data or software
- [How-To's](#): Learn how to accomplish specific tasks at NCBI
- [Submissions](#): Submit data to GenBank or other NCBI databases

NCBI YouTube channel

Learn how to get the most out of NCBI tools and databases with video tutorials on the NCBI YouTube Channel. [GO](#)



1 2 3 4 5 6 7 8

### Popular Resources

- PubMed
- Bookshelf
- PubMed Central
- PubMed Health
- BLAST**
- Nucleotide
- Genome
- SNP
- Gene
- Protein
- PubChem

### NCBI Announcements

- Come to the NCBI Discovery Workshops on February 4&5! 16 Jan 201
- Spaces are still available for the free, 2-day Discovery Workshops to be held on
- New version of Genome Workbench available 06 Sep 201
- An integrated, downloadable application

## Basic Local Alignment Search Tool

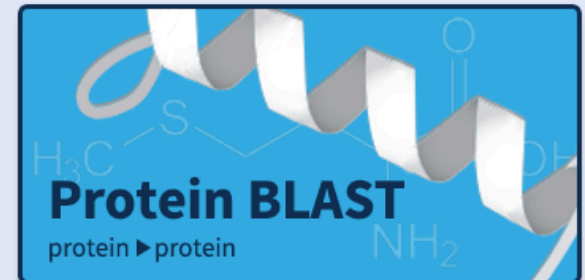
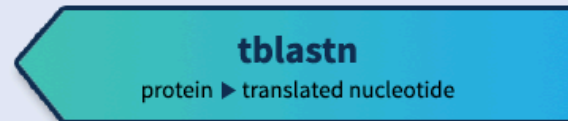
**BLAST** finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. [Learn more](#)

NEWS

### Introducing: Magic-BLAST

Magic-BLAST is a new tool for mapping large sets of next-generation RNA or DNA sequencing runs against a whole genome or transcriptome.  
Wed, 24 Aug 2016 11:00:00 EST [More BLAST news...](#)

## Web BLAST




### BLAST Genomes


Enter organism common name, scientific name, or tax id


Human  Mouse  Rat  Microbes

Search against human database

## Standalone and API BLAST

 **Download BLAST**  
Get BLAST databases and executables

 **Use BLAST API**  
Call BLAST from your application

 **Use BLAST in the cloud**  
Start an instance at a cloud provider

blastn blastp blastx tblastn tblastx

Enter Query Sequence

BLASTN programs search nucleotide databases using a nucleotide query. [more...](#)

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

Choose Search Set

Database

Genome (all assemblies top-level) (548 sequences)

Exclude

Optional

Models (XM/XP)

Entrez Query

Optional

Enter an Entrez query to limit search

Program Selection

Optimize for

- Highly similar sequences (megablast)
- More dissimilar sequences (discontiguous megablast)
- Somewhat similar sequences (blastn)

Choose a BLAST algorithm

BLAST

Search database Genome (all assemblies top-level) - Homo sapiens using Megablast (Optimize for highly similar sequences)

Show results in a new window

+ Algorithm parameters

Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign

A lot of things you may explore

Or, upload file

Choose File No file chosen

Upload a text file with human tp53 mRNA fasta sequence

Job Title

Enter a descriptive title for your BLAST search

Download from course webpage

Choose Search Set

Database

ESTs 8704778 sequences

Exclude

Models (XM/XP) Uncultured/environmental sample sequences

Entrez Query

Enter an Entrez query to limit search

Program Selection

Optimize for

- Highly similar sequences (megablast)
  - More dissimilar sequences (discontiguous megablast)
  - Somewhat similar sequences (blastn)
- Choose a BLAST algorithm

Question: how many ESTs match tp53 genes?

BLAST

Search database ESTs - Homo sapiens using Megablast (Optimize for highly similar sequences)

Show results in a new window

Algorithm parameters

Note: Parameter values that differ from the default are highlighted in yellow and marked with diamond sign

General Parameters

Max target sequences

100

Change here to 1000

Select the maximum number of aligned sequences to display

Short queries

Automatically adjust parameters for short input sequences

Expect threshold

10

Word size

Max matches in a query range

0

Scoring Parameters

Match/Mismatch Scores

2,-3

Gap Costs

Existence: 5 Extension: 2

Filters and Masking





It took ~1 minute to finish

Edit and Resubmit

Save Search Strategies > Formatting options > Download

Change the result display back to traditional format

YouTube Learn about the enhanced report Blast report description

### gi|371502114|ref|NM\_000546.5| Homo sapiens

**Query ID** lcl|39445  
**Description** gi|371502114|ref|NM\_000546.5| Homo sapiens tumor protein p53 (TP53), transcript variant 1, mRNA  
**Molecule type** nucleic acid  
**Query Length** 2591

**Database Name** ESTs  
**Description** Homo sapiens ESTs  
**Program** BLASTN 2.2.27+ > Citation

A lot of things you may explore!!!

Other reports: > Search Summary [Taxonomy reports] [Distance tree of results]

+ Graphic Summary

- Descriptions

Provide feedback on the new report

#### Sequences producing significant alignments:

Select: All None Selected:999

Alignments Download > GenBank Graphics Distance tree of results

	Description	Max score	Total score	Query cover	E value	Max ident	Accession
<input checked="" type="checkbox"/>	<a href="#">AL530477 Homo sapiens NEUROBLASTOMA COT 50-NORMALIZED Homo sapiens cDNA clone CS0DD007YN17 5-PRIME, r</a>	1766	1766	42%	0.0	96%	<a href="#">AL530477.3</a>
<input checked="" type="checkbox"/>	<a href="#">BX345594 Homo sapiens PLACENTA COT 25-NORMALIZED Homo sapiens cDNA clone CS0DI024YH07 5-PRIME, mRNA seq</a>	1678	1678	38%	0.0	97%	<a href="#">BX345594.2</a>
<input checked="" type="checkbox"/>	<a href="#">AGENCOURT_6437640 NIH_MGC_71 Homo sapiens cDNA clone IMAGE:5532699 5', mRNA sequence</a>	1640	1640	41%	0.0	94%	<a href="#">BM467806.1</a>
<input checked="" type="checkbox"/>	<a href="#">AGENCOURT_6871501 NIH_MGC_99 Homo sapiens cDNA clone IMAGE:5930620 5', mRNA sequence</a>	1577	1577	35%	0.0	98%	<a href="#">BQ066009.1</a>
<input checked="" type="checkbox"/>	<a href="#">AL521701 Homo sapiens NEUROBLASTOMA COT 10-NORMALIZED Homo sapiens cDNA clone CS0DB003YB13 3-PRIME, r</a>	1572	1572	38%	0.0	96%	<a href="#">AL521701.3</a>
<input checked="" type="checkbox"/>	<a href="#">AGENCOURT_10099789 NIH_MGC_71 Homo sapiens cDNA clone IMAGE:6502488 5', mRNA sequence</a>	1563	1563	34%	0.0	99%	<a href="#">BU508407.1</a>
<input checked="" type="checkbox"/>	<a href="#">AGENCOURT_7937273 NIH_MGC_92 Homo sapiens cDNA clone IMAGE:6012571 5', mRNA sequence</a>	1537	1537	35%	0.0	98%	<a href="#">BU163993.1</a>
<input checked="" type="checkbox"/>	<a href="#">AGENCOURT_8736418 NIH_MGC_47 Homo sapiens cDNA clone IMAGE:6338637 5', mRNA sequence</a>	1537	1537	35%	0.0	98%	<a href="#">BQ894209.1</a>
<input checked="" type="checkbox"/>	<a href="#">AGENCOURT_6925248 NIH_MGC_110 Homo sapiens cDNA clone IMAGE:5952587 5', mRNA sequence</a>	1528	1528	36%	0.0	97%	<a href="#">BU157354.1</a>
<input checked="" type="checkbox"/>	<a href="#">603062136F1 NIH_MGC_118 Homo sapiens cDNA clone IMAGE:5211321 5', mRNA sequence</a>	1526	1526	36%	0.0	97%	<a href="#">BI518429.5</a>

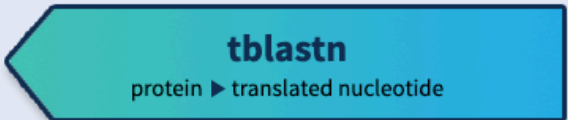
# Web BLAST



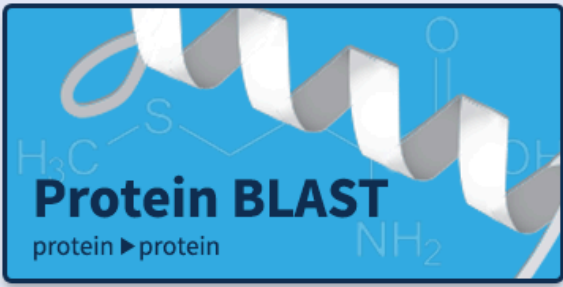
**Nucleotide BLAST**  
nucleotide ► nucleotide



**blastx**  
translated nucleotide ► protein



**tblastn**  
protein ► translated nucleotide



**Protein BLAST**  
protein ► protein

Search against other refseq genomes


### BLAST Genomes




asper Search

- Aspergillus oryza (taxid:5062)
- Aspergillus oryzae (taxid:5062)
- Aspergillus flavus (taxid:5059)
- Aspergillus flavus Link 1809 (taxid:5059)
- Aspergillus cf. flavus PWE1 (taxid:5067)
- Aspergillus cf. flavus PWE10 (taxid:5067)
- Aspergillus cf. flavus PWE2 (taxid:5067)
- Aspergillus cf. flavus PWE27 (taxid:5067)

### Standalone and API BLAS



**Download BLAST**  
Get BLAST databases and execu



**Use BLAST in the cloud**  
Start an instance at a cloud provider

# Hands on exercise 2

Protein blast (blastp and tblastn)

## Basic Local Alignment Search Tool

**BLAST** finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance.

[Learn more](#)

NEWS

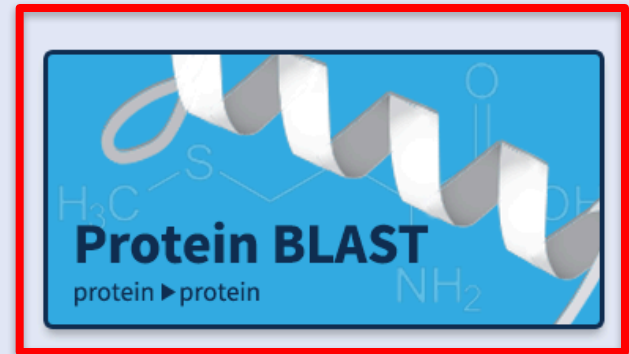
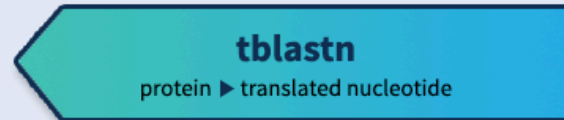
### Introducing: Magic-BLAST

Magic-BLAST is a new tool for mapping large sets of next-generation RNA or DNA sequencing runs against a whole genome or transcriptome.  
Wed, 24 Aug 2016 11:00:00 EST

[More BLAST news...](#)

## Web BLAST

If not select organisms ...



## BLAST Genomes

[Human](#)   [Mouse](#)   [Rat](#)   [Microbes](#)

[blastn](#) **[blastp](#)** [blastx](#) [tblastn](#) [tblastx](#)

BLASTP programs search protein databases using a protein query. [more...](#)

## 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

Non-redundant protein sequences (nr)

Organism   
 Optional

Enter organism name or id--completions will be suggested  Exclude

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude   
 Optional

Models (XM/XP)  Uncultured/environmental sample sequences

Entrez Query   
 Optional

Enter an Entrez query to limit search

You can still specify organisms ...

- Non-redundant protein sequences (nr)
- Reference proteins (refseq\_protein)
- UniProtKB/Swiss-Prot (swissprot)
- Patented protein sequences (pat)
- Protein Data Bank proteins (pdb)
- Metagenomic proteins (env\_nr)

## Program Selection

Algorithm

- blastp (protein-protein BLAST)
- PSI-BLAST (Position-Specific Iterated BLAST)
- PHI-BLAST (Pattern Hit Initiated BLAST)
- DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

- Genomic plus Transcript
  - Human genomic plus transcript (Human G+T)
  - Mouse genomic plus transcript (Mouse G+T)
- Other Databases
- Nucleotide collection (nr/nt)
- Reference RNA sequences (refseq\_rna)
- Reference genomic sequences (refseq\_genomic)
- NCBI Genomes (chromosome)
- Expressed sequence tags (est)
- Genomic survey sequences (gss)
- High throughput genomic sequences (HTGS)
- Patent sequences (pat)
- Protein Data Bank (pdb)
- Human ALU repeat elements (alu\_repeats)
- Sequence tagged sites (dbsts)
- Whole-genome shotgun contigs (wgs)
- Transcriptome Shotgun Assembly (TSA)
- 16S ribosomal RNA sequences (Bacteria and Archaea)

**BLAST**

Search **database Non-redundant protein sequences (nr)** using **Blastp (protein-prot**

Show results in a new window

[blastn](#) [blastp](#) [blastx](#) [tblastn](#) [tblastx](#)

BLASTP programs search protein databases using a protein query. [more...](#)

[Reset page](#) [Bookmarks](#)

### Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

[Clear](#)

Query subrange

From

To

Or, upload file

Choose File test2

Job Title

Enter a descriptive title for your BLAST search

Align two or more sequences

Upload a text file with two arabidopsis protein fasta sequence

Download from course webpage

### Choose Search Set

Database

Non-redundant protein sequences (nr)

Organism  
Optional

populus

- Populus (taxid:3689)
- Populus L. (taxid:3689)
- Populus balsamifera subsp. trichocarpa (t...
- Populus trichocarpa (taxid:3694)
- Populus balsamifera (taxid:73824)
- Populus balsamifera L. (taxid:73824)
- Populus alba x Populus tremula (taxid:808...
- Populus alba x tremula (taxid:80863)

Exclude

0 top taxa will be shown.

sample sequences

Type in populus to choose populus trichocarpa

Question: what are the homologs in poplar tree?

### Program Selection

Algorithm

DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

You may submit many sequences, but expect it takes time

**BLAST**

Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)

Show results in a new window

It took ~1 minute (smaller database)

Your search is limited to records matching entrez query: txid3694 [ORGN].

[Edit and Resubmit](#) [Save Search Strategies](#) [Formatting options](#) [Download](#)

[Change the result display back to traditional format](#)

You [Learn about the enhanced report](#) [Blast report description](#)

AT5G22740.1|AT5G22740.1|csIA (535 letters)

Click here to choose to view which query protein

Results for: 1:|cl|96604 AT5G22740.1|AT5G22740.1|csIA(535aa)

**Query ID** |cl|96604  
**Description** AT5G22740.1|AT5G22740.1|csIA  
**Molecule type** amino acid  
**Query Length** 535

**Database Name** nr  
**Description** All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects  
**Program** BLASTP 2.2.27+ [Citation](#)

Other reports: [Search Summary](#) [Taxonomy reports](#) [Distance tree of results](#) [Multiple alignment](#)

**New** DELTA-BLAST, a more sensitive protein-protein search

**+ Graphic Summary**

**- Descriptions**

[Provide feedback on the new report](#)

Sequences producing significant alignments:

Select: [All](#) [None](#) Selected:0

[Alignments](#) [Download](#) [GenPept](#) [Graphics](#) [Distance tree of results](#) [Multiple alignment](#)

	Description	Max score	Total score	Query cover	E value	Max ident	Accession
<input type="checkbox"/>	<a href="#">predicted protein [Populus trichocarpa] &gt;gb EEE76058.1</a> <a href="#">predicted protein [Populus trichocarpa]</a>	930	930	99%	0.0	82%	<a href="#">XP_002328178.1</a>
<input type="checkbox"/>	<a href="#">predicted protein [Populus trichocarpa] &gt;gb EEE86848.1</a> <a href="#">predicted protein [Populus trichocarpa]</a>	929	929	99%	0.0	83%	<a href="#">XP_002312893.1</a>
<input type="checkbox"/>	<a href="#">predicted protein [Populus trichocarpa] &gt;gb EEE71909.1</a> <a href="#">predicted protein [Populus trichocarpa]</a>	830	830	99%	0.0	73%	<a href="#">XP_002326239.1</a>
<input type="checkbox"/>	<a href="#">predicted protein [Populus trichocarpa] &gt;gb EEF01528.1</a> <a href="#">predicted protein [Populus trichocarpa]</a>	826	826	99%	0.0	72%	<a href="#">XP_002315357.1</a>
<input type="checkbox"/>	<a href="#">predicted protein [Populus trichocarpa] &gt;gb EEE89339.1</a> <a href="#">predicted protein [Populus trichocarpa]</a>	811	811	97%	0.0	74%	<a href="#">XP_002311972.1</a>
<input type="checkbox"/>	<a href="#">predicted protein [Populus trichocarpa] &gt;gb EEE92253.1</a> <a href="#">predicted protein [Populus trichocarpa]</a>	428	428	85%	2e-141	45%	<a href="#">XP_002308730.1</a>

How to determine what is a good e-value cutoff to select homologs?

<http://www.youtube.com/watch?v=nO0wJgZRZJs&list=PL8FD4CC12DABD6B39&index=6>



Enter Query Sequence

TBLASTN search translated nucleotide databases using a protein query. [more...](#)

Enter accession number(s), gi(s), or FASTA sequence(s)

Clear

Query subrange

From

To

Or, upload file

Choose File 1.txt

Job Title

Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Database

Transcriptome Shotgun Assembly (TSA)

Organism

Optional

charophy

Charophyta/Embryophyta group (taxid:131221)

charophyte/embryophyte group (taxid:131221)

charophytes (taxid:3146)

Charophyceae (taxid:304574)

Exclude

will be shown.

Type in charoph to choose charophytes

BLAST

using Tblastn (search translated nucleotide databases using a protein

Question: what are the EST homologs in charophytic algae?

**Your search is limited to records matching entrez query: txid3146 [ORGN].**

[Edit and Resubmit](#) [Save Search Strategies](#) [Formatting options](#) [Download](#)

[Change the result display back to traditional format](#)

[YouTube](#) [Learn about the enhanced report](#) [Blast report description](#)

### AT5G22740.1|AT5G22740.1|csIA (535 letters)

Results for:

**Query ID** lc|24042  
**Description** AT5G22740.1|AT5G22740.1|csIA  
**Molecule type** amino acid  
**Query Length** 535

**Database Name** tsa\_nt  
**Description** Transcriptome Shotgun Assembly (TSA) sequences  
**Program** TBLASTN 2.2.27+ [Citation](#)

Other reports: [Search Summary](#) [Taxonomy reports](#)

**+ Graphic Summary**

**- Descriptions**

[Provide feedback on the new report](#)

#### Sequences producing significant alignments:

Select: [All](#) [None](#) Selected:0

[Alignments](#) [Download](#) [GenBank](#) [Graphics](#)

	Description	Max score	Total score	Query cover	E value	Max ident	Accession
<input type="checkbox"/>	<a href="#">Nitella mirabilis comp297_c0_seq1 mRNA sequence</a>	392	392	96%	2e-123	41%	<a href="#">JV768755.1</a>
<input type="checkbox"/>	<a href="#">Nitella mirabilis comp1819_c0_seq1 mRNA sequence</a>	364	364	82%	2e-111	44%	<a href="#">JV767519.1</a>

Download ▾ GenBank Graphics

TSA: Nitella mirabilis comp297\_c0\_seq1 mRNA sequence

Sequence ID: [gb|JV768755.1](#) Length: 3278 Number of Matches: 1

Range 1: 305 to 1849 [GenBank](#) [Graphics](#)

▾ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
392 bits(1008)	2e-123	Compositional matrix adjust.	212/521(41%)	310/521(59%)	11/521(2%)	-2
Query 9		VLPEITFDGVRMEITGQLGMIWELVKAPVIVPLLQLAVYICLLMSVMLLCERVYMGIVIVL			68	
Sbjct 1849		VL TAPLAGV-AELFAGLTASFRSFRARHVAPVMQSVINVLIVVFAVQSMDTLGMTLISFY			1673	
Query 69		VKLF-WK--KPKDKRYKFEPIHDDDEELGSSN--FPVVLVQIPMFNEREVYKLSIGAACGLS			123	
Sbjct 1672		FSLTGWKARKVTPLVTHHPRKADNLSTKTEVYPRVLIQIPMFNERECYQISISACSQLD			1493	
Query 124		WPSDRLVIQVLLDDSDPTVKQMVVEVCQRWASKGINIRYQIRENRVGYKAGALKEGLKRS			183	
Sbjct 1492		WPRDKLVIQVLLDDSNNEEIKEMVKEEVSKWQSRGVNIDYRHRVDRITGYKGGSLKQGM LAP			1313	
Query 184		YVKHCEYVVFADADFQPEPDFLRRSIPFLMHNPNIALVQARWRVNSDECLLRMQEMSL			243	
Sbjct 1312		YVKDCDFVAVFDADFQPRADWLLQTVPYFKDDPKLGLVQTRWEYSNQFTNLLTRFQFINM			1133	
Query 244		DYHFTVEQEVGSSSTHAFFGFNGTAGIWRIAAINEAGGWKDRITVEDMDLAVRASLRGWKF			303	
Sbjct 1132		SYHFEVEQQVMGAIMNFFGFNGTGGIWRVAAVNDCCGGWDRITVEDMDIAVRAIHGWNF			953	
Query 304		LYLGDLDQVKSELPSTFRAFREFQQRHWSCGPANLFRKMVMEIVRNKKVRFWKKVYVIYSFF			363	
Sbjct 952		VFLNHVRVPCLEPQILEAYTRQQRHWAGPMNLFRLILAPKIVRSKSLTFASKFHLIVLFF			773	
Query 364		FVRKIIAHWVTFCFYCVVLPILVPEVKVPIWGSVYIPIIITILNSVGTIPRSIHLLFYW			423	
Sbjct 772		FVRRLLVPTVNFLLFVLLPLSLFVPEANIPIWVTYTFPMFLSLFRMCLCPTLFPYMPFY			593	
Query 424		ILFENVMSLHRTKATLIGLFEAGRANWVVTAKLGGQSAKGNTKGIKRFPRIFKLPDRL			483	
Sbjct 592		FFENTMVMTKLSANIQGLFQFGRVNEWVVTAKVGA--LAAKNPDAVNK-PK--RKPLKL			428	
Query 484		NTLELGFAAFLFVCGCYDFVHGKNNYFIYLFQTMSEFFISG	524			
Sbjct 427		FKRELLMSAFLLLAIIQSLAIEKGIHFYIFLQGLTFFSFG	305			

# Hands on exercise 3

PHI-BLAST

Query protein + short motif/pattern

&

PSI-BLAST (iterated BLAST)

Multi-round BLASTP

Example: plant glycosyltransferase family 8 (GT8) has signature motif

We want to search Arabidopsis GAUT1 protein (gi #: 86611465) and the HXXGXXKPW motif

GT8 Class	Clade	DxD	HxxGxxKPW
I	Position	between a5 and b6	after b10
	GAUT	KYYELDDD <sub>Y</sub> VVQ <sub>K</sub> DL	H <sub>Y</sub> NG <sub>NM</sub> KPWL
	GATL	RV <sub>I</sub> Y <sub>L</sub> LDSD <sub>L</sub> YYVDD <sub>I</sub>	HW <sub>S</sub> SG <sub>K</sub> GKPW <sub>L</sub>
	GATR	R <sub>F</sub> I <sub>I</sub> Y <sub>L</sub> LDSD <sub>T</sub> PL <sub>I</sub> V <sub>V</sub> K <sub>G</sub> NI	H <sub>F</sub> NG <sub>K</sub> EKPW <sub>K</sub>
	Metazoan-1	K <sub>O</sub> I <sub>I</sub> Y <sub>L</sub> DDD <sub>V</sub> I <sub>V</sub> Q <sub>A</sub> TSGWLN <sub>L</sub> LD <sub>I</sub>	H <sub>W</sub> NG <sub>H</sub> EKPW <sub>S</sub>
	Metazoan-2	IS <sub>V</sub> I <sub>V</sub> LD <sub>T</sub> D <sub>Y</sub> T <sub>F</sub> K <sub>S</sub> P <sub>I</sub>	H <sub>W</sub> N <sub>S</sub> P <sub>K</sub> K <sub>L</sub> I <sub>V</sub> K <sub>S</sub>
II	GoIS	K <sub>M</sub> I <sub>I</sub> Y <sub>L</sub> D <sub>S</sub> DIQVE <sub>R</sub> NI	H <sub>Y</sub> C <sub>A</sub> AG <sub>S</sub> KP <sub>W</sub> <sub>B</sub>
	PGSIP-A	K <sub>I</sub> <sub>I</sub> Y <sub>L</sub> D <sub>S</sub> DIQVE <sub>R</sub> NI	H <sub>Y</sub> L <sub>G</sub> <sub>L</sub> KP <sub>W</sub> <sub>L</sub> C
	PGSIP-B	K <sub>V</sub> V <sub>Y</sub> LDAD <sub>T</sub> I <sub>V</sub> V <sub>K</sub> S <sub>I</sub>	H <sub>Y</sub> TL <sub>G</sub> PL <sub>K</sub> P <sub>W</sub> <sub>D</sub> W
	PGSIP-C	R <sub>V</sub> V <sub>M</sub> LD <sub>S</sub> DN <sub>L</sub> FL <sub>S</sub> NT	F <sub>P</sub> S <sub>A</sub> P <sub>M</sub> L <sub>K</sub> P <sub>W</sub> <sub>Y</sub> WW

ProSite style pattern: H-x(2)-G-x(2)-K-P-W

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#) Clear [?](#) Query subrange [?](#)

From   
To

Or, upload file  1.txt [?](#)

Job Title   
Enter a descriptive title for your BLAST search [?](#)

Align two or more sequences [?](#)

### Choose Search Set

Database  [?](#)

Organism Optional   Exclude   
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown. [?](#)

Exclude Optional  Models (XM/XP)  Uncultured/environmental sample sequences

Entrez Query Optional   
Enter an Entrez query to limit search [?](#)

### Program Selection

Algorithm

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

[?](#)  
Enter a PHI pattern [?](#)

Enter a PHI pattern to start the search. PHI-BLAST may perform better than simple pattern searching because it is random and not indicative of homology).

[Edit and Resubmit](#) [Save Search Strategies](#) [Formatting options](#) [Download](#)

[Change the result display back to traditional format](#)

[You Tube Learn about the enhanced report](#) [Blast report description](#)

**PSI blast Iteration 1**

gi|86611465 (673 letters)

**Query ID** [gi|86611465|gb|ABD14404.1|](#)  
**Description** homogalacturonan alpha-1,4-galacturonosyltransferase [Arabidopsis thaliana]  
**Molecule type** amino acid  
**Query Length** 673

**Database Name** nr  
**Description** All non-redundant GenBank CDS translations+PDB+SwissProt+excluding environmental samples from WGS projects  
**Program** BLASTP 2.2.27+ [Citation](#)

Other reports: [Search Summary](#) [Taxonomy reports](#) [Distance tree of results](#) [Multiple alignment](#)

**New** DELTA-BLAST, a more sensitive protein-protein search

[+ Graphic Summary](#)

[- Descriptions](#)

[Provide feedback on the new report](#)

Run PSI-Blast iteration 2 with max   Pattern at position:

**- Sequences producing significant alignments with E-value BETTER than threshold**

Select: [All](#) [None](#) Selected:0

[Alignments](#) [Download](#) [GenPept](#) [Graphics](#) [Distance tree of results](#) [Multiple alignment](#)

	Description	Max score	Total score	Query cover	E value	Max ident	Accession	Select for PSI blast	Used to build PSSM
<input type="checkbox"/>	<a href="#">alpha-1,4-galacturonosyltransferase 1 [Arabidopsis thaliana] &gt;sp Q9LE59.1 GAUT1_ARATH RecName: Full=Polvgala</a>	1351	1351	100%	0.0	0%	<a href="#">NP_191672.1</a>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	<a href="#">CAUT1/CT1 [Arabidopsis lyrata subsp. lyrata] &gt;abi55452853.1 CAUT1/CT1 [Arabidopsis lyrata subsp. lyrata]</a>	1318	1318	100%	0.0	0%	<a href="#">XP_002876594.1</a>	<input checked="" type="checkbox"/>	<input type="checkbox"/>

PSI blast Iteration 2

gij86611465 (673 letters)

[Skip to the first new s](#)

**Query ID** [gij86611465|qb|ABD14404.1](#)  
**Description** homogalacturonan alpha-1,4-galacturonosyltransferase [Arabidopsis thaliana]  
**Molecule type** amino acid  
**Query Length** 673

**Database Name** nr  
**Description** All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects  
**Program** BLASTP 2.2.27+ [▶ Citation](#)

Other reports: [▶ Search Summary](#) [\[Taxonomy reports\]](#) [\[Distance tree of results\]](#) [\[Multiple alignment\]](#)

+ [Graphic Summary](#)

- [Descriptions](#)

[Provide feedback on the new report](#)

Run PSI-Blast iteration 3 with max

- **Sequences producing significant alignments with E-value BETTER than threshold**

Select: [All](#) [None](#) Selected:0 Yellow: sequences scoring below threshold on previous iteration

[↑](#) [Alignments](#) [Download](#) [GenPept](#) [Graphics](#) [Distance tree of results](#) [Multiple alignment](#)

	Description	Max score	Total score	Query cover	E value	Max ident	Accession	Select for PSI blast	Used to build PSSM
<input type="checkbox"/>	<a href="#">alpha-1,4-galacturonosyltransferase 1 [Arabidopsis thaliana] &gt;sp Q9LE59.1 GAUT1_ARATH RecName: Full=Polygal</a>	1093	1093	100%	0.0	100%	<a href="#">NP_191672.1</a>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input type="checkbox"/>	<a href="#">GAUT1/LGT1 [Arabidopsis lyrata subsp. lyrata] &gt;qb EFH52853.1 GAUT1/LGT1 [Arabidopsis lyrata subsp. lyrata]</a>	1085	1085	100%	0.0	98%	<a href="#">XP_002876594.1</a>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input type="checkbox"/>	<a href="#">unnamed protein product [Vitis vinifera]</a>	1022	1022	100%	0.0	75%	<a href="#">CBI38820.3</a>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input type="checkbox"/>	<a href="#">JHL05D22.8 [Jatropha curcas]</a>	1016	1016	100%	0.0	76%	<a href="#">BAJ53137.1</a>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input type="checkbox"/>	<a href="#">Os09q0531900 [Oryza sativa Japonica Group] &gt;dbj BAD46018.1 glycosyl transferase family 8 protein-like [Oryza sati</a>	1011	1011	99%	0.0	72%	<a href="#">NP_001063757.1</a>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input type="checkbox"/>	<a href="#">hypothetical protein Osl_32147 [Oryza sativa Indica Group]</a>	1008	1008	99%	0.0	72%	<a href="#">EEC84934.1</a>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>




# Hands on exercise 4

## RPS-BLAST

Given protein sequences, find conserved functional domains


# Specialized searches

**SmartBLAST**



Find proteins highly similar to your query

**Primer-BLAST**




Design primers specific to your PCR template

**Global Align**



Compare two sequences across their entire span (Needleman-Wunsch)

**CD-search**



Find conserved domains in your sequence

**GEO**



Find matches to gene expression profiles

**IgBLAST**



Search immunoglobulins and T cell receptor sequences

**VecScreen**




Search sequences for vector contamination

**CDART**



Find sequences with similar conserved domain architecture

**Targeted Loci**




Search markers for phylogenetic analysis

**Multiple Alignment**



Align sequences using domain and protein constraints

**BioAssay**

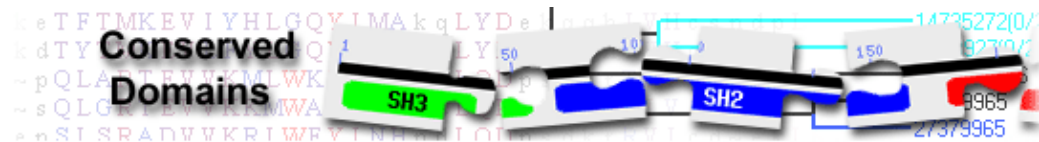


Search protein or nucleotide targets in PubChem BioAssay

**MOLE-BLAST**



Establish taxonomy for uncultured or environmental sequences



## Search for Conserved Domains within a protein or coding nucleotide sequence

**NEW!** Use [Batch CD-search](#) to submit multiple query proteins at once!

Enter **protein** or **nucleotide** query as accession, gi, or sequence in [FASTA format](#) ?

86611465

### OPTIONS

Search against database ? : CDD v3.08 - 43334 PSSMs ▼

Expect Value ? threshold: 0.01 ▼

Apply low-complexity filter ?

Force live search ?

Maximum number of hits ? 500

Result mode  Concise ?  Full ?




Submit

Reset

## Retrieve previous CD-search result

Request ID:  Retrieve ?

### References:

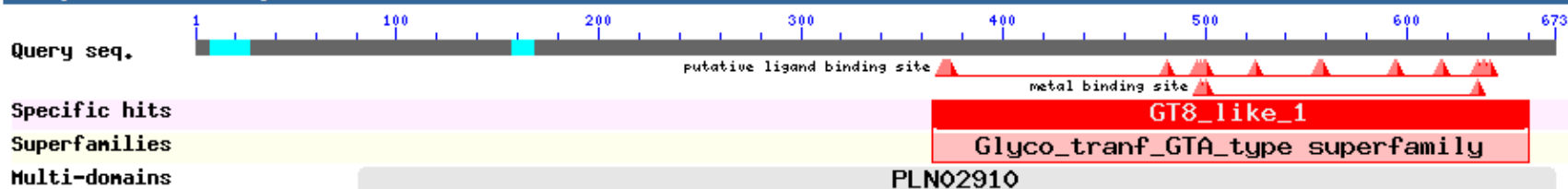
-  Marchler-Bauer A et al. (2013), "CDD: conserved domains and protein three-dimensional structure.", **Nucleic Acids Res.**41(D1)348-52.
-  Marchler-Bauer A et al. (2011), "CDD: a Conserved Domain Database for the functional annotation of proteins.", **Nucleic Acids Res.**39(D)225-9.
-  Marchler-Bauer A, Bryant SH (2004), "CD-Search: protein domain annotations on the fly.", **Nucleic Acids Res.**32(W)327-331.

## Conserved domains on [gi|86611465|gb|ABD14404|]

[View full result](#)

homogalacturonan alpha-1,4-galacturonosyltransferase [Arabidopsis thaliana]

### Graphical summary [show options >](#)

[Search for similar domain architectures](#)[Refine search](#)

### List of domain hits

	Description	Pssmid	Multi-dom	E-value
[+]GT8_like_1[cd06429]	GT8_like_1 represents a subfamily of GT8 with unknown function.; A subfamily of glycosyltransferase family 8 with unknown function.	133051	no	1.93e-124
[+]PLN02910[PLN02910]	polygalacturonate 4-alpha-galacturonosyltransferase	178498	yes	0e+00

### References:

- Marchler-Bauer A et al. (2013), "CDD: conserved domains and protein three-dimensional structure.", **Nucleic Acids Res.**41(D1)348-52.
- Marchler-Bauer A et al. (2011), "CDD: a Conserved Domain Database for the functional annotation of proteins.", **Nucleic Acids Res.**39(D)225-9.
- Marchler-Bauer A, Bryant SH (2004), "CD-Search: protein domain annotations on the fly.", **Nucleic Acids Res.**32(W)327-331.

[Help](#) | [Disclaimer](#) | [Write to the Help Desk](#)

Next class: NCBI GEO and ftp  
resource (with a little bit intro to  
Linux skills) and practice