

NCBI resources II: web-based tools and ftp resources

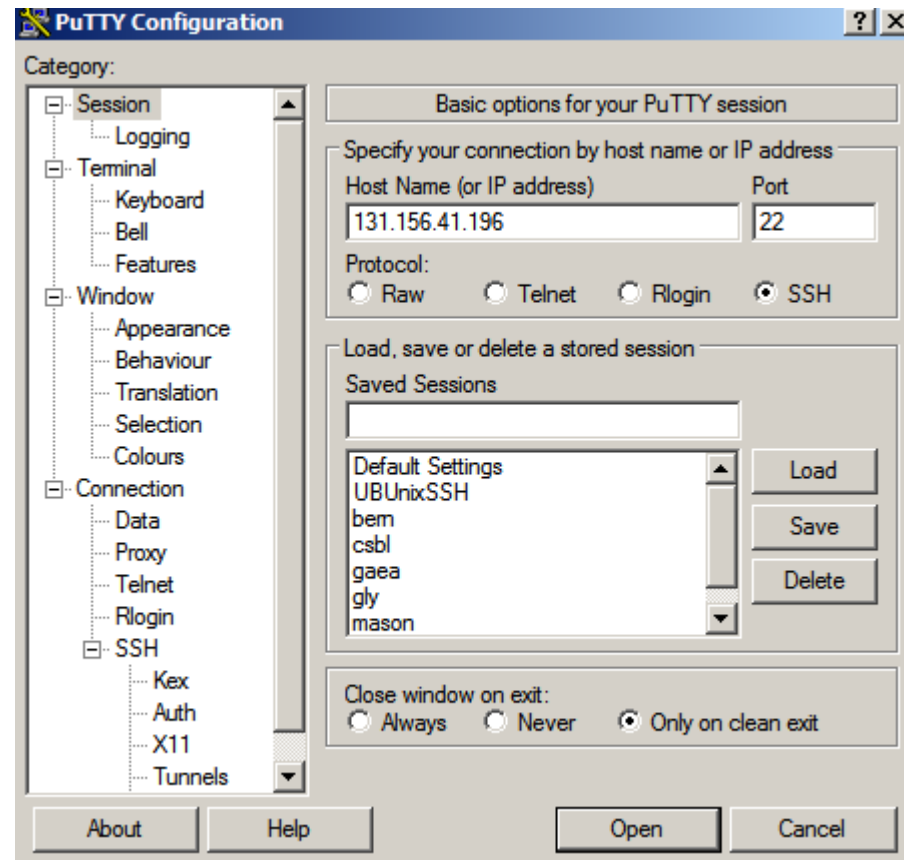
Yanbin Yin
Spring 2013

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

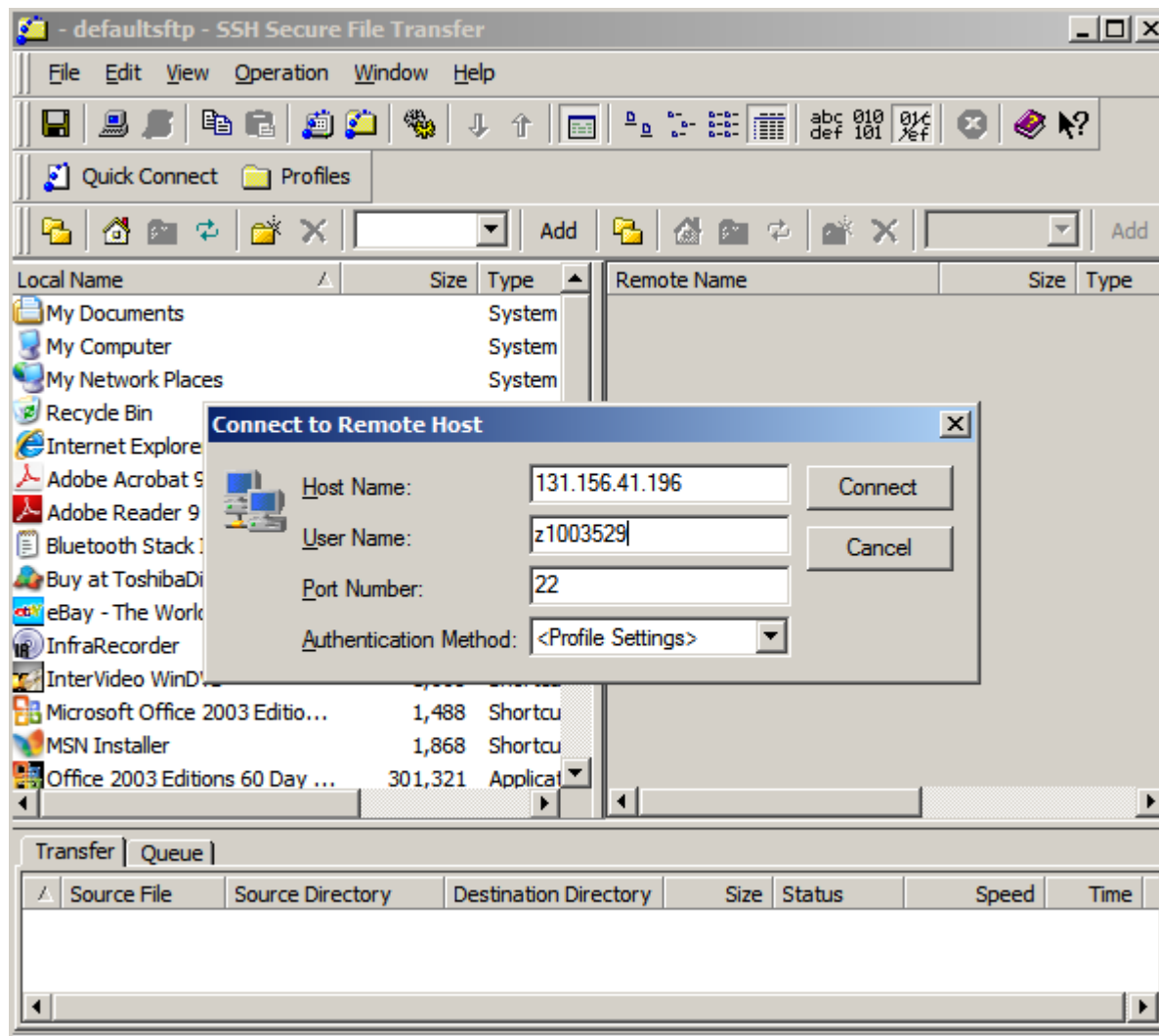
New login info

IP 131.156.41.196
Account student ID (e.g. z1003529)
Pswd student ID (e.g. z1003529)

z117479
z1245176
z1608050
z1576493
z1598039
z1559435
z1660438
z1003529
z1678230



PuTTY



SSH secure client

Outline

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

References

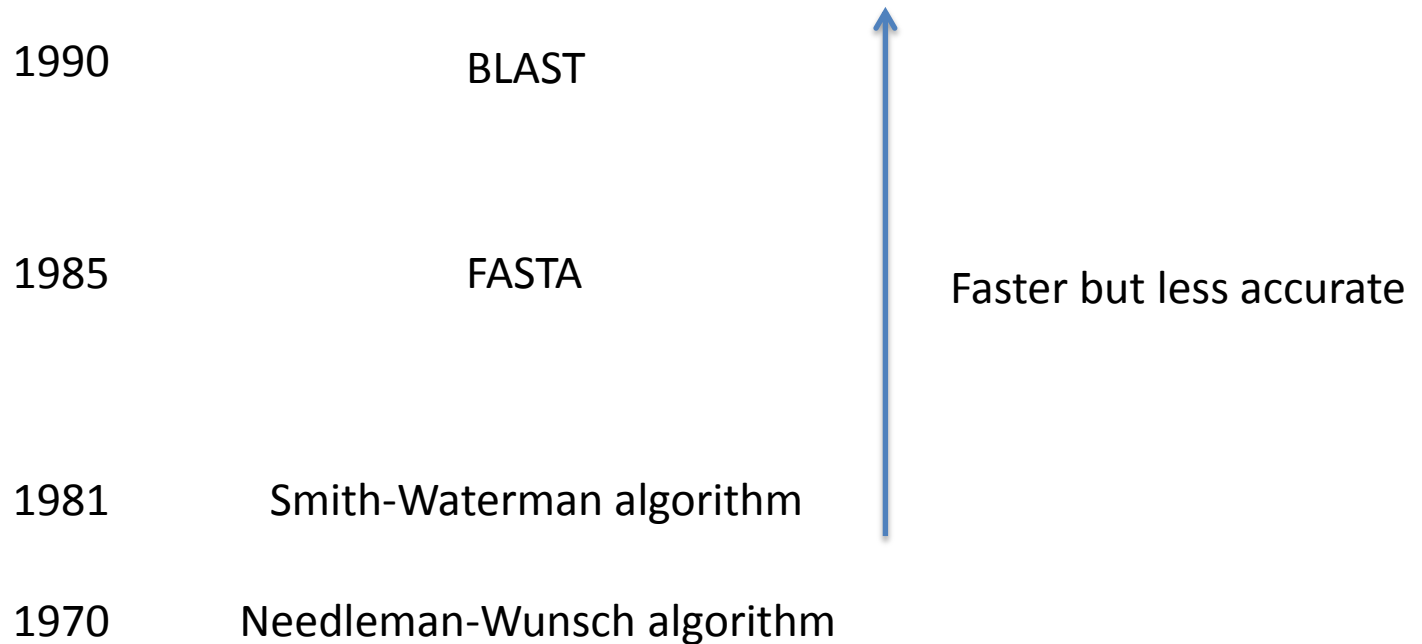
NCBI discovery workshops

ftp://ftp.ncbi.nih.gov/pub/education/discovery_workshops/NLM/2012/Sept2012/

http://homepages.ulb.ac.be/~dgonze/TEACHING/stat_scores.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

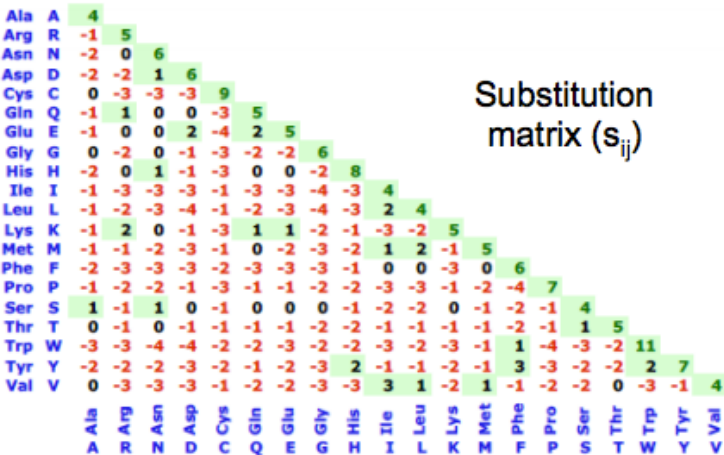
Score: A number used to assess the biological relevance of a finding.

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:

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



gap penalty (s_{i-})

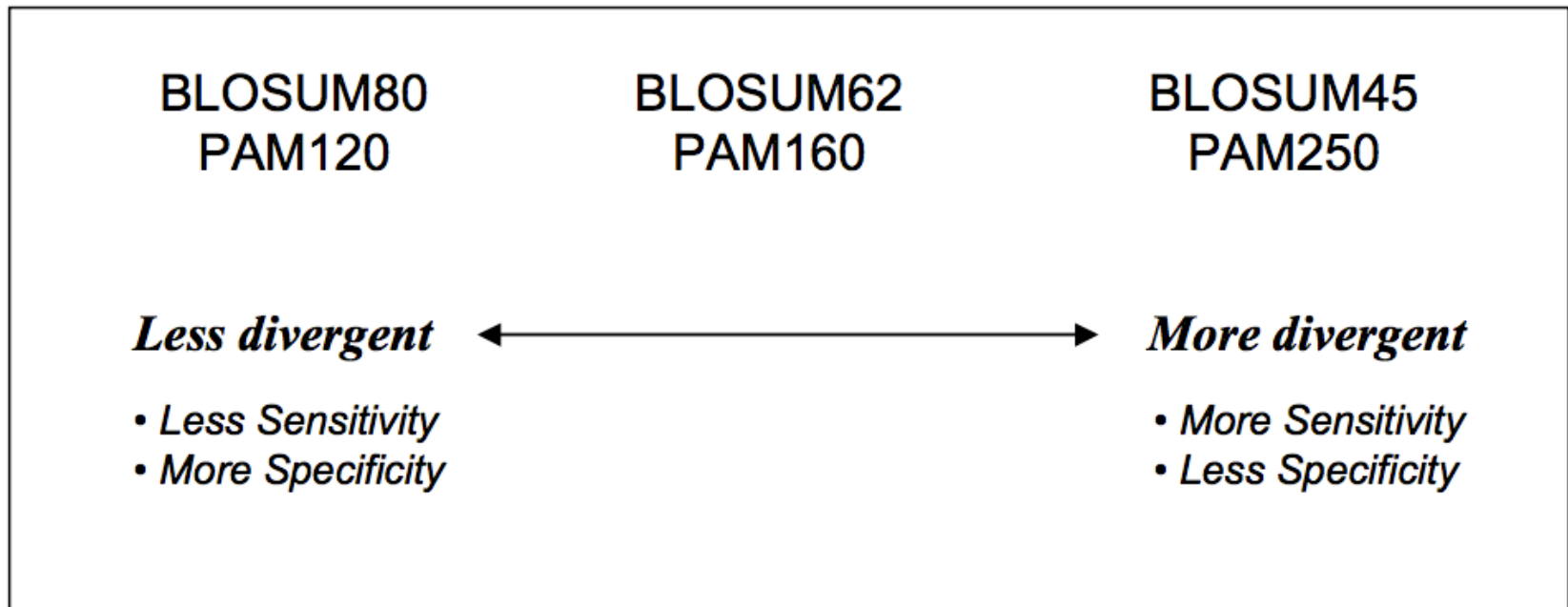
- gap opening -4
- gap extension -1
- end gap 0

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

λ = estimated parameter

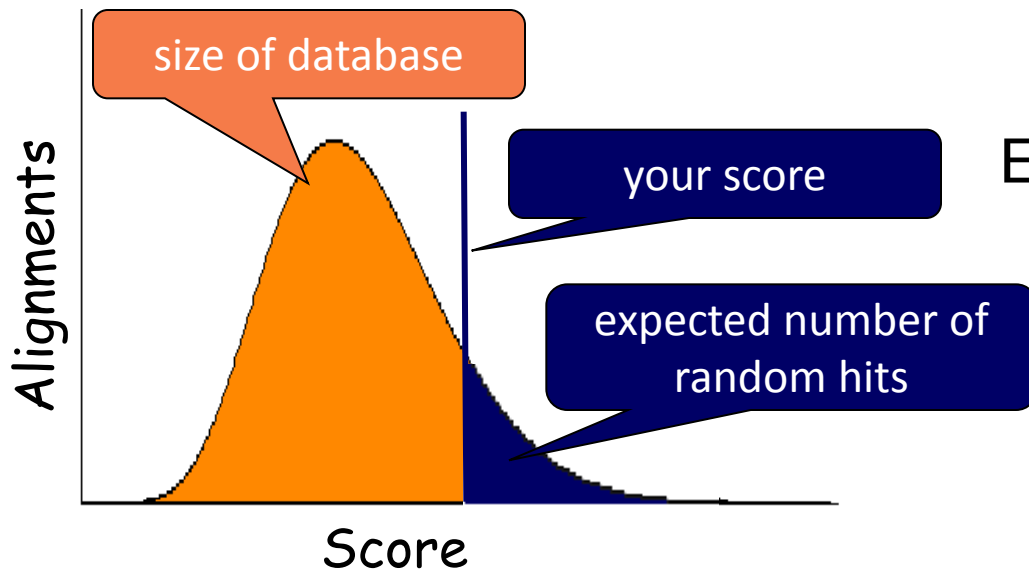
S = score of the high-scoring segment pair (HSP)

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

λ = scale for scoring system

S' = bitscore = $(\lambda S - \ln K) / \ln 2$

(applies to ungapped alignments)

Local Alignment Scoring: Protein

Number of Chance Alignments = 4×10^{-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--DNPGHPFIMTVGCVAGDEESYEVFKE 87
+ K LT +L+++ +D+ GF+ I +G N G VG AG +SY F
Sbjct 26 LQKCLTKDLWEQCKDRRDKEYGFSFKQAI FSGSKWTNSG-----VG VYAGSHDSYYAFAP 79

Query	80	MAK	VTLE	LYKK	LRDK	ETPS	GTVD	DDVI	QTGV	--DN	PGHP	FIMT	VGCV	AGDE	ESYE	VFKE	87	
Sbjct	8	LQK	CLTK	DLWE	QCKD	RRDK	EYGF	SFKQ	AI F	S	GSKW	TNSG	-----	VG	VYAG	SHDS	YYAF	AP 79

Query	1	K	+5	E	+1	F	-3	Gap	- (11 + 4 (1)) = -14	
Sbjct	138	AVTR	KERK	EIEH	LVTS	ALGE	FTGE	LKGY	YSLE	TMSD

Query 205 SGMARDWPDARGIWHNDNKSFLVWVNEEDHLRVISMKEKGGNMKEVFRRFCVG 256
+G+ RDWP+ARGI+HND K+FLVWVNEED LR+ISM+ G N+ EVF+R V
Sbjct 197 AGLERDWPEARGIFHNDAKTFLVWVNEEDQLRIISMQAGSNILEVFKRLSVA 248

Scores from BLOSUM62, a position independent matrix

Local Alignment Scoring: Nucleotide

Number of Chance Alignments = 2×10^{-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		TTCTTTTTCGTTGGTGTAAAGCCAACACCCTGTCTAAAAAACATAAATTTCTTTAATCAT	33892
Query	2042	TTTGCTCTTTTCTCTGTGCTACAATTAATAAAAAAATGAAAAGAATCTAATTTAATTGT	2100
Sbjct	33893	TTTGCTCTTTTCTCTGTGCTACAATTAATAAAAAAATGAAAAGAATCTAATTTAATTGT	33952
Query	2101	ACAGCACTGTTA-T	2159
Sbjct	33953	CTATGACTGTTATT	34012
Query	2160	AAGTTCCAGTGTTCT	2219
Sbjct	34013	AAATTCCACTATTCTCTCTTTCCCTATTTCATGGAGGACTTCTAGTTCCTTCTGGATTA	34072
Query	2220	AT----TAAATAAATCATTAATACT	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	1	0	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	0	1	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
Organism
Optional

Exclude
Optional

Entrez Query
Optional

Non-redundant protein sequences (nr)
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)

suggested ☐ Exclude +

tax id. Only 20 top taxa will be shown.

environmental sample sequences

Enter an Entrez query

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

Choose Search Set

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

Organism
Optional

Exclude
Optional

Entrez Query
Optional

Nucleotide collection (nr/nt)

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

- **tss**

- transcriptome shotgun assembly

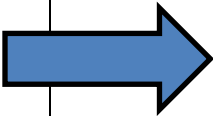
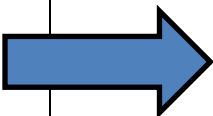
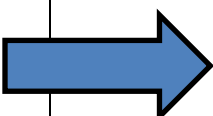
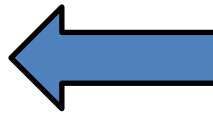
- **16S microbial**

- 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](#)

BLAST finds regions of similarity between biological sequences. [more...](#)

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

BLAST Assembled RefSeq Genomes

Choose a species genome to search, or [list all genomic BLAST databases](#).

- Human

Mouse

Rat

Arabidopsis thaliana

Oryza sativa

Bos taurus

Danio rerio

Drosophila melanogaster

Gallus gallus

Pan troglodytes

Microbes

Apis mellifera

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a nucleotide database using a nucleotide query <i>Algorithms:</i> blastn, megablast, discontinuous megablast
protein blast	Search protein database using a protein query <i>Algorithms:</i> blastp, psi-blast, phi-blast, delta-blast
blastx	Search protein database using a translated nucleotide query
tblastn	Search translated nucleotide database using a protein query
tblastx	Search translated nucleotide database using a translated nucleotide query

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 using [SNP flanks](#)
- Search using [SNP flanks](#)

News

[Improved BLAST statistics described in BMC Research Notes.](#)

BLAST calculates expect values that describe the significance of a match, with a lower expect value indicating a more significant match.

Fri, 11 Jan 2013 10:00:00 EST

[More BLAST news...](#)

Tip of the Day

[How to save custom search pages.](#)

So you have made a few BLAST searches and after adjusting the database, organism limits and maybe a few Algorithm Parameters you arrive at what you think is a good search strategy.

[More tips...](#)



blastn blastp blastx tblastn tblastx

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

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

Choose Search Set

Database

Genome (all assemblies scaffolds) 7154 sequences

Exclude

Optional

☐ Models (XM/XP) ☐ Uncultured/environmental sample sequences

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 scaffolds) - Homo sapiens using Megablast (Optimize for highly similar sequences)

☐ Show results in a new window

+ Algorithm parameters

A lot of things you may explore

[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 [?](#)

Choose Search Set

Database

RefSeq protein

31390 sequences [?](#)

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

Choose a BLAST algorithm [?](#)

Protein db search has different programs

BLAST

Search **database RefSeq protein - Homo sapiens** using **Blastp (protein-protein BLAST)**

☐ Show results in a new window

Or, upload file

Choose File No file chosen

Job Title

Enter a descriptive title for your BLAST search

Upload a text file with human tp53 mRNA fasta sequence
Download from course webpage

Choose Search Set

Database

ESTs 8704778 sequences

Exclude

Optional

Entrez Query

Optional

☐ Models (XM/XP) ☐ Uncultured/environmental sample sequences

Enter an Entrez query to limit search

Go back to blastn page

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 ♦ 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

27

Max matches in a query range

0

Scoring Parameters

Match/Mismatch Scores

2,-3

Gap Costs

Existence: 5 Extension: 2

Filters and Masking



BLAST®

Basic Local Alignment Search Tool

[Home](#)[Recent Results](#)[Saved Strategies](#)[Help](#)[My NCBI](#)[\[Sign In\]](#) [\[Register\]](#)

It took ~1 minute to finish

[NCBI/BLAST/blastn suite/](#) **Formatting Results - G0VRDCPW013**[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](#)

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
☒	AL530477 Homo sapiens NEUROBLASTOMA COT 50-NORMALIZED Homo sapiens cDNA clone CS0DD007YN17 5-PRIME, r	1766	1766	42%	0.0	96%	AL530477.3
☒	BX345594 Homo sapiens PLACENTA COT 25-NORMALIZED Homo sapiens cDNA clone CS0DI024YH07 5-PRIME, mRNA seq	1678	1678	38%	0.0	97%	BX345594.2
☒	AGENCOURT_6437640 NIH_MGC_71 Homo sapiens cDNA clone IMAGE:5532699 5', mRNA sequence	1640	1640	41%	0.0	94%	BM467806.1
☒	AGENCOURT_6871501 NIH_MGC_99 Homo sapiens cDNA clone IMAGE:5930620 5', mRNA sequence	1577	1577	35%	0.0	98%	BQ066009.1
☒	AL521701 Homo sapiens NEUROBLASTOMA COT 10-NORMALIZED Homo sapiens cDNA clone CS0DB003YB13 3-PRIME, r	1572	1572	38%	0.0	96%	AL521701.3
☒	AGENCOURT_10099789 NIH_MGC_71 Homo sapiens cDNA clone IMAGE:6502488 5', mRNA sequence	1563	1563	34%	0.0	99%	BU508407.1
☒	AGENCOURT_7937273 NIH_MGC_92 Homo sapiens cDNA clone IMAGE:6012571 5', mRNA sequence	1537	1537	35%	0.0	98%	BU163993.1
☒	AGENCOURT_8736418 NIH_MGC_47 Homo sapiens cDNA clone IMAGE:6338637 5', mRNA sequence	1537	1537	35%	0.0	98%	BQ894209.1
☒	AGENCOURT_6925248 NIH_MGC_110 Homo sapiens cDNA clone IMAGE:5952587 5', mRNA sequence	1528	1528	36%	0.0	97%	BU157354.1
☒	603062136F1 NIH_MGC_118 Homo sapiens cDNA clone IMAGE:5211321 5', mRNA sequence	1526	1526	36%	0.0	97%	BI518429.1

BLAST finds regions of similarity between biological sequences. [more...](#)

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

BLAST Assembled RefSeq Genomes

Choose a species genome to search, or [list all genomic BLAST databases.](#)

- Human

Mouse

Rat

Arabidopsis thaliana

Oryza sativa

Bos taurus

Danio rerio

Drosophila melanogaster

Gallus gallus

Pan troglodytes

Microbes

Apis mellifera

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a nucleotide database using a nucleotide query <i>Algorithms:</i> blastn, megablast, discontinuous megablast
protein blast	Search protein database using a protein query <i>Algorithms:</i> blastp, psi-blast, phi-blast, delta-blast
blastx	Search protein database using a translated nucleotide query
tblastn	Search translated nucleotide database using a protein query
tblastx	Search translated nucleotide database using a translated nucleotide query

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 using [SNP flanks](#)

[Conserved domain architecture](#)

News

[Improved BLAST statistics described in BMC Research Notes.](#)

BLAST calculates expect values that describe the significance of a match, with a lower expect value indicating a more significant match.
Fri, 11 Jan 2013 10:00:00 EST

[More BLAST news...](#)

Tip of the Day

[How to save custom search pages.](#)

So you have made a few BLAST searches and after adjusting the database, organism limits and maybe a few Algorithm Parameters you arrive at what you think is a good search strategy.

[More tips...](#)

Search against other refseq databases

Tools Legend

Q

 Search or Browse the Genome

B

 BLAST

Cf

 Clone Finder

R

 Go to region on a chromosome

G

 Genome Resources page

News

Nasonia vitripennis build 2.1 released Nov 1, 2011

Nasonia vitripennis build 2.1 has been released and is now a... [more](#)

Oreochromis niloticus build Orenil1.0 released Oct 31, 2011

Oreochromis niloticus build Orenil1.0 has been released and ... [more](#)

Cavia porcellus build 1.1 released Oct 31, 2011

Cavia porcellus build 1.1 has been released and is now avail... [more](#)

Canis familiaris build 2.2 released Oct 27, 2011

Canis familiaris build 2.2 has been released and is now avai... [more](#)

[Show all](#)

Related Resources

NCBI Home
 NCBI Web Search
 NCBI Site map
 Genome Browser agreement
 Genome Biology
 Taxonomy
 Entrez (Global Query)
 BLAST
 Map Viewer FTP

Small Genomes

Bacteria

Organelles

Viruses

<i>Homo sapiens</i>	human	Annotation Release 104	Q	B	Cf	G
		Build 36.3	Q	B	R	Cf
<i>Macaca mulatta</i>	rhesus macaque	Build 1.2	Q	B	R	G
<i>Nomascus leucogenys</i>	northern white-cheeked gibbon	Build 1.1	Q	B		
<i>Pan troglodytes</i>	chimpanzee	Build 3.1	Q	B	R	G
		Build 2.1	Q	B	R	
<i>Pongo abelii</i>	Sumatran orangutan	Build 1.2	Q	B	R	
▼ Rodents (4)						
Scientific name	Common name	Build	Tools			
<i>Cavia porcellus</i>	domestic guinea pig	Build 1.1	Q	B		
<i>Cricetulus griseus</i>	Chinese hamster	Build 1.1	Q	B		
<i>Mus musculus</i>	laboratory mouse	Build 38.1	Q	B	Cf	G
		Build 37.2	Q	B	R	Cf
<i>Rattus norvegicus</i>	rat	Build 5.1 (Annotation Release 103)	Q	B	Cf	G
		RGSC v3.4	Q	B	R	Cf
▶ Monotremes (1)						
▶ Marsupials (1)						
▶ Other Mammals (9)						
▶ Other Vertebrates (7)						
▶ Invertebrates (16)						
▶ Protozoa (B) (19)						
▶ Plants (Q) (118)						
▼ Fungi (Q, B) (17)						
Scientific name	Common name	Build	Tools			
<i>Aspergillus clavatus</i>		Build 1.1	Q	B		G
<i>Aspergillus fumigatus</i>		Build 2.1	Q	B	R	G
<i>Aspergillus niger</i>		Build 1.1	Q	B	R	G
<i>Candida glabrata</i>		Build 1.1	Q	B	R	
<i>Cryptococcus neoformans</i>		Build 2.1	Q	B	R	
<i>Debaryomyces hansenii</i>		Build 1.1	Q	B	R	
<i>Encephalitozoon cuniculi</i>		Build 1.1	Q	B	R	
<i>Eremothecium gossypii</i>		Build 3.1	Q	B	R	
<i>Gibberella zeae</i>		Build 1.2	Q	B	R	
<i>Kluyveromyces lactis</i>		Build 1.1	Q	B	R	
<i>Magnaporthe oryzae</i>	rice blast fungus	Build 3.1	Q	B	R	
<i>Neurospora crassa</i>		Build 2.1	Q	B	R	
<i>Saccharomyces cerevisiae</i>	baker's yeast	Build 3.1	Q	B		G
		Build 2.1	Q	B	R	
<i>Scheffersomyces stipitis</i>		Build 1.1	Q	B	R	
<i>Schizosaccharomyces pombe</i>	fission yeast	Build 1.1	Q	B	R	
<i>Ustilago maydis</i>		Build 1.1	Q	B	R	
<i>Yarrowia lipolytica</i>		Build 1.1	Q	B	R	

BLAST with microbial genomes (2428 bacterial/122 archaeal/426 eukaryotic genomes tree)

P - indicates the ability to search against protein sequences, - completed genomic sequence, - Whole Genome Shotgun, - add/remove from selection, See [\[Help\]](#) and [\[Article\]](#) for details.

Attention!
Because genome sequencing centers are now submitting unfinished genomes to the WGS division of DDBJ/EMBL/GenBank, the unfinished genomic sequences were removed from this page. We encourage any submitters of these sequences to submit their unfinished microbial genomes to WGS.

Enter your query sequence as Accession/GI or FASTA:

Select type of query and database or [BLAST-program](#)

Query: Database: Blast-program: MegaBlast: ☒

You may change BLAST options

[Expect:](#) [Filter](#) [Descriptions](#) [Alignments](#)

alphabetical menu

Check ☐ if you want to select only completed genomes

- [Archaea](#)
 - [Crenarchaeota](#)
 - [Desulfurococcales](#)
 - ☐ [Aeropyrum pernix K1](#)
 - ☐ [Desulfurococcus fermentans DSM 16532](#)
 - ☐ [Desulfurococcus kamchatkensis 1221n](#)
 - ☐ [Desulfurococcus mucosus DSM 2162](#)
 - ☐ [Hyperthermus butylicus DSM 5456](#)
 - ☐ [Ignicoccus hospitalis KIN4/I](#)
 - ☐ [Ignisphaera aggregans DSM 17230](#)
 - ☐ [Pyrolobus fumarii 1A](#)
 - ☐ [Staphylothermus hellenicus DSM 12710](#)

Try Neisseria

Ctrl+f and type neisseria

- P ☐ [Burkholderia sp. F1160](#)
- P ☐ [Burkholderia sp. KJ006](#)
- ☐ [Burkholderia sp. SJ98](#)
- P ☐ [Burkholderia sp. TJI49](#)
- P ☐ [Burkholderia sp. YI23](#)
- P ☐ [Burkholderia thailandensis](#)
- P ☐ [Burkholderia ubonensis Bu](#)
- P ☐ [Burkholderia vietnamiensis G4](#)
- P ☐ [Burkholderia xenovorans LB400](#)
- P ☐ [Candidatus Glomeribacter gigasporarum BEG34](#)
- ☐ [Cupriavidus basilensis OR16](#)
- P ☐ [Cupriavidus metallidurans CH34](#)
- P ☐ [Cupriavidus necator](#)
- P ☐ [Cupriavidus taiwanensis LMG 19424](#)
- P ☐ [Lautropia mirabilis ATCC 51599](#)
- P ☐ [Limnobacter sp. MED105](#)
- P ☐ [Polynucleobacter necessarius](#)
- P ☐ [Ralstonia eutropha JMP134](#)
- P ☐ [Ralstonia pickettii](#)
- P ☐ [Ralstonia solanacearum](#)
- P ☐ [Ralstonia sp. 5_2_56FAA](#)
- P ☐ [Ralstonia sp. 5_7_47FAA](#)
- ☒ **Neisseriaceae**
- P ☒ [Chromobacterium violaceum ATCC 12472](#)
- P ☒ [Eikenella corrodens ATCC 23834](#)

Neisseria 1 of 20 ^ v x

Click on + to select all Neisseria

BLAST with microbial genomes ([2428 bacterial](#)/[122 archaeal](#)/[426 eukaryotic genomes tree](#))

P - indicates the ability to search against protein sequences,  - completed genomic sequence,  - Whole Genome Shotgun,  - add/remove from selection, See [\[Help\]](#) and [\[Article\]](#) for details.

Attention!

Because genome sequencing centers are now submitting unfinished genomes to the WGS division of DDBJ/EMBL/GenBank, the unfinished genomic sequences were removed from this page. We encourage any submitters of these sequences to submit their unfinished microbial genomes to WGS.

Enter your query sequence as Accession/GI or FASTA:

Select type of query and database or [BLAST-program](#)

Query: Database:

Blast-program: MegaBlast: ☒

You may change BLAST options

Expect: [Filter](#)

[Descriptions](#) [Alignments](#)

[Select all](#) [Clear all](#)

26

genomes

[BLAST](#)

[Adv. BLAST](#)

[Show](#) alphabetical menu

Check ☐ if you want  to select only completed genomes

 [Archaea](#)

 [Crenarchaeota](#)

 [Desulfurococcales](#)

☐ [Aeropyrum pernix K1](#)

☐ [Desulfurococcus fermentans DSM 16532](#)

☐ [Desulfurococcus kamchatkensis 1221n](#)

☐ [Desulfurococcus mucosus DSM 2162](#)

☐ [Hyperthermus butylicus DSM 5456](#)

☐ [Thermoplasma acidophilum DSM 1728](#)

The Map Viewer provides a wide variety of genome mapping and sequencing data. [More..](#)

Search

Search:

for:

Go

Tools Legend

Q Search or Browse the Genome

B BLAST

Cf Clone Finder

R Go to region on a chromosome

G Genome Resources page

News

Nasonia vitripennis build 2.1 released Nov 1, 2011

Nasonia vitripennis build 2.1 has been released and is now a... [more](#)

Oreochromis niloticus build Orenil1.0 released Oct 31, 2011

Oreochromis niloticus build Orenil1.0 has been released and ... [more](#)

Cavia porcellus build 1.1 released Oct 31, 2011

Cavia porcellus build 1.1 has been released and is now avail... [more](#)

Canis familiaris build 2.2 released Oct 27, 2011

Canis familiaris build 2.2 has been released and is now avail... [more](#)

[Show all](#)

Related Resources

NCBI Home

NCBI Web Search

NCBI Site map

Genome Browser agreement

Genome Biology

Taxonomy

Entrez (Global Query)

BLAST

Map Viewer FTP

Small Genomes

▼ Vertebrates (28)			
▼ Mammals (21)			
▼ Primates (6)			
Scientific name	Common name	Build	Tools
<i>Callithrix jacchus</i>	white-tufted-ear marmoset	Build 1.1	Q B R
<i>Homo sapiens</i>	human	Annotation Release 104	Q B Cf G
		Build 36.3	Q B R Cf
<i>Macaca mulatta</i>	rhesus macaque	Build 1.2	Q B R G
<i>Nomascus leucogenys</i>	northern white-cheeked gibbon	Build 1.1	Q B
<i>Pan troglodytes</i>	chimpanzee	Build 3.1	Q B R G
		Build 2.1	Q B R
<i>Pongo abelii</i>	Sumatran orangutan	Build 1.2	Q B R
▼ Rodents (4)			
Scientific name	Common name	Build	Tools
<i>Cavia porcellus</i>	domestic guinea pig	Build 1.1	Q B
<i>Cricetulus griseus</i>	Chinese hamster	Build 1.1	Q B
<i>Mus musculus</i>	laboratory mouse	Build 38.1	Q B Cf G
		Build 37.2	Q B R Cf
<i>Rattus norvegicus</i>	rat	Build 5.1 (Annotation Release 103)	Q B Cf G
		RGSC v3.4	Q B R Cf
▶ Monotremes (1)			
▶ Marsupials (1)			
▶ Other Mammals (9)			
▶ Other Vertebrates (7)			
▶ Invertebrates (16)			
▶ Protozoa B (19)			
▶ Plants Q (118)			
▶ Fungi Q B (17)			

Other organisms: click on B

To explore genomes: <http://www.ncbi.nlm.nih.gov/genome>

Hands on exercise 2

Protein blast (blastp and tblastn)

BLAST finds regions of similarity between biological sequences. [more...](#)

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

BLAST Assembled RefSeq Genomes

Choose a species genome to search, or [list all genomic BLAST databases](#).

- ☐ [Human](#)

☐ [Mouse](#)

☐ [Rat](#)

☐ [Arabidopsis thaliana](#)
- ☐ [Oryza sativa](#)

☐ [Bos taurus](#)

☐ [Danio rerio](#)

☐ [Drosophila melanogaster](#)
- ☐ [Gallus gallus](#)

☐ [Pan troglodytes](#)

☐ [Microbes](#)

☐ [Apis mellifera](#)

Basic BLAST

Choose a BLAST program to run.

- nucleotide blast

Search a **nucleotide** database using a **nucleotide** query
Algorithms: blastn, megablast, discontinuous megablast
- protein blast

Search **protein** database using a **protein** query
Algorithms: blastp, psi-blast, phi-blast, delta-blast
- blastx

Search **protein** database using a **translated nucleotide** query
- tblastn

Search **translated nucleotide** database using a **protein** query
- tblastx

Search **translated nucleotide** database using a **translated nucleotide** query

If not select organisms ...

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 using [SNP flanks](#)
- ☐ Search using [SNP flanks](#)

News

[Improved BLAST statistics described in BMC Research Notes.](#)

BLAST calculates expect values that describe the significance of a match, with a lower expect value indicating a more significant match.

Fri, 11 Jan 2013 10:00:00 EST

 [More BLAST news...](#)

Tip of the Day

[How to save custom search pages.](#)

So you have made a few BLAST searches and after adjusting the database, organism limits and maybe a few Algorithm Parameters you arrive at what you think is a good search strategy.

 [More tips...](#)

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

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

BLAST

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

☐ Show results in a new window

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)

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)

[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

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

Exclude

Optional

Entrez Query

Optional

☐ 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)

► [NCBI/ BLAST/ blastp suite/ Formatting Results - G0Z5ATWW01R](#)

ⓘ 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
☐	predicted protein [Populus trichocarpa] >gb EEE76058.1 predicted protein [Populus trichocarpa]	930	930	99%	0.0	82%	XP_002328178.1
☐	predicted protein [Populus trichocarpa] >gb EEE86848.1 predicted protein [Populus trichocarpa]	929	929	99%	0.0	83%	XP_002312893.1
☐	predicted protein [Populus trichocarpa] >gb EEE71909.1 predicted protein [Populus trichocarpa]	830	830	99%	0.0	73%	XP_002326239.1
☐	predicted protein [Populus trichocarpa] >gb EEF01528.1 predicted protein [Populus trichocarpa]	826	826	99%	0.0	72%	XP_002315357.1
☐	predicted protein [Populus trichocarpa] >gb EEE89339.1 predicted protein [Populus trichocarpa]	811	811	97%	0.0	74%	XP_002311972.1
☐	predicted protein [Populus trichocarpa] >gb EEE92253.1 predicted protein [Populus trichocarpa]	428	428	85%	2e-141	45%	XP_002308730.1

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

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

blastn blastp blastx **tblastn** tblastx

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

Exclude

Optional

Entrez Query

Optional

☐ Exclude +

will be shown. ?

nces

charophy

Charophyta/Embryophyta group (taxid:131221)

charophyte/embryophyte group (taxid:131221)

charophytes (taxid:3146)

Charophyceae (taxid:304574)

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
☐	Nitella mirabilis comp297_c0_seq1 mRNA sequence	392	392	96%	2e-123	41%	JV768755.1
☐	Nitella mirabilis comp1819_c0_seq1 mRNA sequence	364	364	82%	2e-111	44%	JV767519.1

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	VLPEITFDGVRMEITGQLGMIWELVKAPVIVPPLLQAVYICLLMSVMLLCERVYMGIVIVL				68	
	VL GV E+ L + +A + P++Q + + +++ + + + M ++					
Sbjct 1849	VLTAPLAGV-AELFAGLTASFRSFRARHVAPVMQSVINVLIVVFAVQSMIDLGMTLISFY				1673	
Query 69	VKLF-WK--KPKDKRYKFEPIHDDEELGSSN--FPVVLVQIPMFNEREVYKLSIGAACGLS				123	
	L WK K P D + L + +P VL+QIPMFNERE Y++SI A L					
Sbjct 1672	FSLTGWKARKVTPLVTHHPRKDADNLSTKTEVYPRVLIQIPMFNERECYQISISACSQDL				1493	
Query 124	WPSDRLLVIQVLDSDTDPTVKQMVVEVCQRWASKGINIRYQIRENRVGYKAGALKEGLKRS				183	
	WP D+LVIQVLDSD + +K+MV+ E +W S+G+NI Y+ R +R GYK G+LK+G+					
Sbjct 1492	WPRDKLVIQVLDSDNNEEIKEMVKEEVSKWQSRGVNIDYRHRVDRITGYKGGSLKQGM LAP				1313	
Query 184	YVKHCEYVVFIDADFQPEPDFLRRSIPFLMHNPNIALVQARWRVNSDECLLRMQEMSL				243	
	YVK C++V +FDADFQ P D+L +++P+ +P + LVQ RW + N LLTR Q +++					
Sbjct 1312	YVKDCDFVAVFDADFQPRADWLLQTVPYFKDDPKLGLVQTRWEYSNQFTNLLTRFQFINM				1133	
Query 244	DYHFTVEQEVGSSTHAFFGFNGTAGIWRIAAINEAGGWKDRITVEDMDLAVRASLRGWKF				303	
	YHF VEQ+V +T FFGFNGT GIWR+AA+N+ GGW RITVEDMD+AVRA + GW F					
Sbjct 1132	SYHFEVEQQVMGATMNF FFGFNGTGGIWRVAAVNDCCGWDTRITVEDMDIAVRAHIHGNF				953	
Query 304	LYLGDLLQVKSELPSTFRAFRFQQHRWSCGPANLFRKMVMEIVRNKKVRFWKKVYVIYSFF				363	
	++L ++V ELP T A+ QQHRW GP NLFR + +IVR+K + F K ++I FF					
Sbjct 952	VFLNHVRVPCELPQILEAYTRQQHRWHAGPMNLFRILAPKIVRSKSLTFASKFHLIVLFF				773	
Query 364	FVRKIIAHWVTF CFYCVVLP LTI L VPEVKVPIWGSVYIPSIITILNSVGTIPRSIHLLFYW				423	
	FVR+++ V F + V+LPL++ VPE +PIW + P +++ P +F +					
Sbjct 772	FVRRLLVPTVNFLFVLLPLSLFVPEANIPIWVTYTFPMFLSLFRMCLCPTLFPMFPY				593	
Query 424	ILFENVMSLHRIKATLIGLFEAGRANWVVTAKLGGSGSAKGNKGIKRFPRIFKLPDRL				483	
	+ FEN M + + A + GLF+ GR NEWVVTAK+G+ A N + + P+ + P +L					
Sbjct 592	LFFENTMVMTKLSANIQGLFQFGRVNEWVVTAKVGA--LAAKNPDAVNK-PK--RKPLKL				428	
Query 484	NTLELGFAAFLFVCGCYDFVHGKNNYFIYFLQTMSTFFISG		524			
	EL +AFL + K +F Q ++FF G					
Sbjct 427	FKRELLMSAFLLLAAIQSLAIEKGIHFYIFLQGLTFFSFG		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	KYYELDDDYYVQKDL	H.Y.N.G.N.M.K.P.W.L
	GATL	R.V.I.Y.L.D.S.D.L.Y.Y.V.D.D.I	H.W.S.G.K.G.K.P.W.L
	GATR	R.F.I.Y.L.D.S.D.T.P.L.I.V.Y.K.G.N.I	H.F.N.G.K.E.K.P.W.K
	Metazoan-1	K.D.I.Y.L.D.D.D.V.I.V.Q.A.T.S.G.W.L.N.L.G.D.I	H.W.N.G.H.E.K.P.W.S
	Metazoan-2	E.V.I.V.L.D.T.D.Y.T.F.K.S.P.I	H.W.N.S.P.K.L.V.K
II	GoIS	K.M.I.Y.L.D.S.D.I.Q.V.E.N.I	H.Y.C.A.A.G.S.K.P.W.B
	PGSIP-A	K.I.I.F.I.D.A.D.L.L.Y.L.R.N.G	H.Y.L.G.L.K.P.W.C
	PGSIP-B	K.V.V.Y.L.D.A.D.T.I.V.Y.K.S.I	H.Y.T.L.G.P.L.K.P.W.D.W
	PGSIP-C	R.V.V.M.L.D.D.N.L.F.L.S.N.T	F.P.S.P.M.L.K.P.W.Y.W.W

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](#) [?](#)

86611465

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](#)

[YouTube](#) [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

Go

[+ Graphic Summary](#)

[- Descriptions](#)

[Provide feedback on the new report](#)

Run PSI-Blast iteration 2 with max 500

Go

Pattern at position: 635

[- 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
☐	alpha-1,4-galacturonosyltransferase 1 [Arabidopsis thaliana] >sp Q9LE59.1 GAUT1_ARATH RecName: Full=Polygalacturonate 4-epimerase	1351	1351	100%	0.0	0%	NP_191672.1	☒	
☐	GAUT1/L CT1 [Arabidopsis lyrata subsp. lyrata] >sp E5H52852.1 CAUT1/L CT1 [Arabidopsis lyrata subsp. lyrata]	1318	1318	100%	0.0	0%	XP_002876594.1	☒	

PSI blast Iteration 2

gi|86611465 (673 letters)

[Skip to the first new s](#)

Query ID [gi|86611465|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
☐	alpha-1,4-galacturonosyltransferase 1 [Arabidopsis thaliana] >sp Q9LE59.1 GAUT1_ARATH RecName: Full=Polvgal	1093	1093	100%	0.0	100%	NP_191672.1	☒	☒
☐	GAUT1/LGT1 [Arabidopsis lyrata subsp. lyrata] >qb EFH52853.1 GAUT1/LGT1 [Arabidopsis lyrata subsp. lyrata]	1085	1085	100%	0.0	98%	XP_002876594.1	☒	☒
☐	unnamed protein product [Vitis vinifera]	1022	1022	100%	0.0	75%	CBI38820.3	☒	☒
☐	JHL05D22.8 [Jatropha curcas]	1016	1016	100%	0.0	76%	BAJ53137.1	☒	☒
☐	Os09g0531900 [Oryza sativa Japonica Group] >dbj BAD46018.1 glycosyl transferase family 8 protein-like [Oryza sati	1011	1011	99%	0.0	72%	NP_001063757.1	☒	☒
☐	hypothetical protein Osl_32147 [Oryza sativa Indica Group]	1008	1008	99%	0.0	72%	EEC84934.1	☒	☒

Hands on exercise 4

RPS-BLAST

Given protein sequences, find
conserved functional domains

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 using [SNP flanks](#)
- ❑ 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](#)

Conserved Domains

SH3 SH2

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

86611465

Submit

Reset

Search against database : CDD v3.08 - 43334 PSSMs

Expect Value [?](#) threshold:

Apply low-complexity filter ☐ ☒

Force live search ☐

Maximum number of hits

Result mode ☒ Concise [?](#) ☐ Full [?](#)

[Retrieve previous CD-search result](#)

Request ID: [Retrieve](#) [?](#)

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.

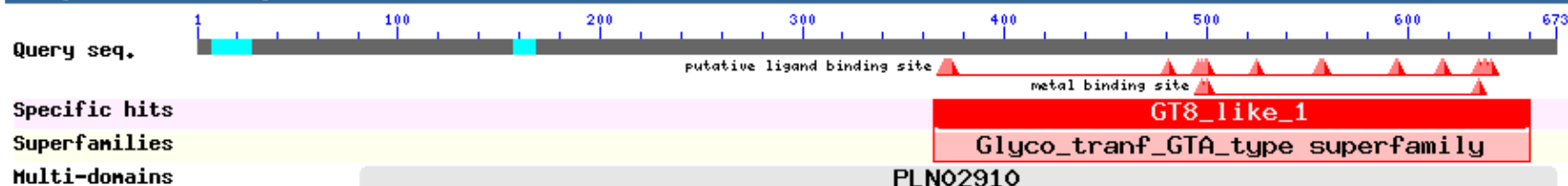
Marchler-Bauer A, Bryant SH (2004), "CD-Search: protein domain annotations on the fly.", **Nucleic Acids Res.**32(W)327-331.

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