Sequence Alignment Program: BLAST MIC 405c

Dr Shilpa Kaistha

Department of Microbiology

IBSBT, CSJM University, Kanpur

Terminology

- **Similarity**: Extent to which 2 sequences are related. Calculated using percent sequence identity/conservation. In BLAST program given as positive score matrix
- **Identity:** Extent to which 2 sequence are invariant (not different)
- **Homology:** Similarity attributed to common ancestor
  - Sequence can be similar and not homologous
  - Homologous sequence are not always highly similar
  - Low complexity regions can be highly similar without being homologous
  - 50% similarity over short sequence ofter occurs by chance
  - Generally, 2 sequences are highly similar over entire length, they are likely to be homologous

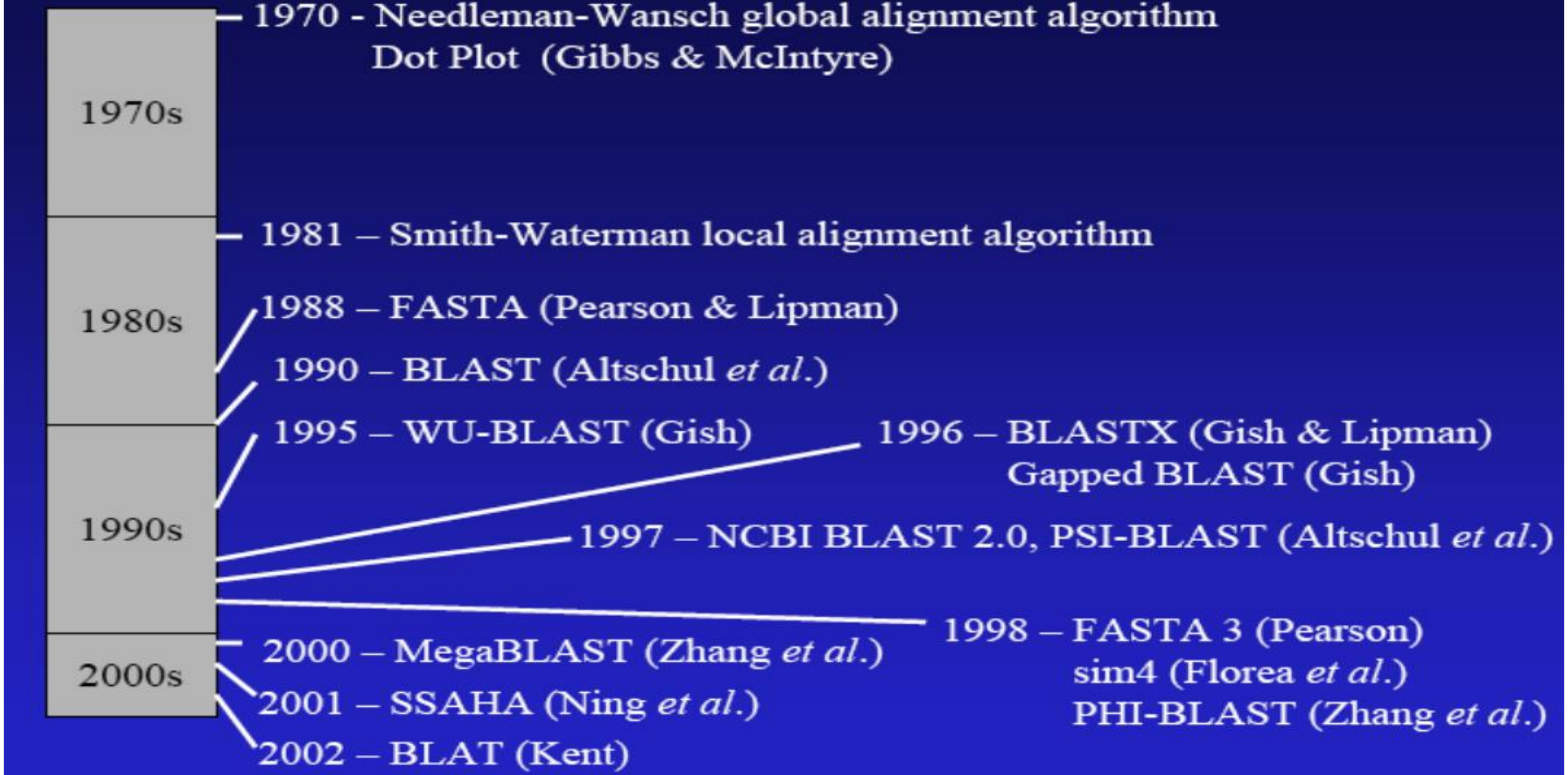
Sequence Similarity Search

- Comparing sequence with existing databases to determine function and relative information
- Identify homologs and eliminate false positives (Non homologs)
- Infer function, transfer annotation, structure- domain information
- Search large number of sequences efficiently
- Alignments can be global or local algorithm
    - Global alignment is an optimal alignment includes all characters from each sequence for alignment (CLUSTAL)
    - Local Alignment is an optimal alignment that includes only most similar regions (BLAST)

Local Alignment- Why?

- Compare short sequence to a large one
- Compare a single sequence to entire database
- Compare partial sequence to the whole sequence
- Results
- Identify newly determined sequence
- Compare new genes to previously known genes
- Guess functions (annotation) for genomes with ORF of unknown functions

Sequence Alignment Programs



1970 - Needleman-Wansch global alignment algorithm
Dot Plot (Gibbs & McIntyre)

1981 – Smith-Waterman local alignment algorithm

1988 – FASTA (Pearson & Lipman)

1990 – BLAST (Altschul *et al.*)

1995 – WU-BLAST (Gish)

1996 – BLASTX (Gish & Lipman)
Gapped BLAST (Gish)

1997 – NCBI BLAST 2.0, PSI-BLAST (Altschul *et al.*)

1998 – FASTA 3 (Pearson)
sim4 (Florea *et al.*)
PHI-BLAST (Zhang *et al.*)

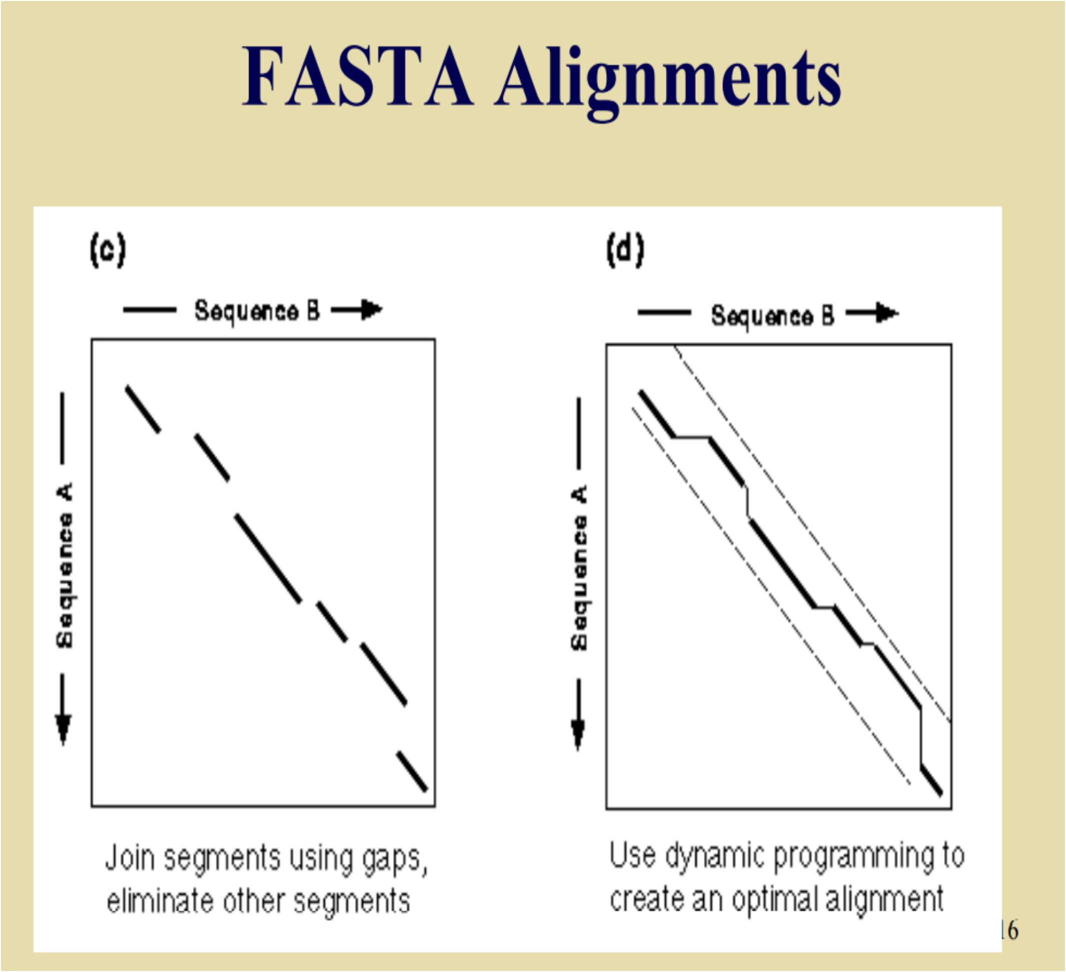
2000 – MegaBLAST (Zhang *et al.*)

2001 – SSAHA (Ning *et al.*)

2002 – BLAT (Kent)

FASTA

- First fast sequence searching algorithm for comparing a query sequence against a database
- BLAST: improvement of FASTA with speed, ease of search and statistical rigor
  - First, identify very short exact matches
  - Next, extend to longer regions of similarity
  - Best hits are optimized



FASTA Alignments

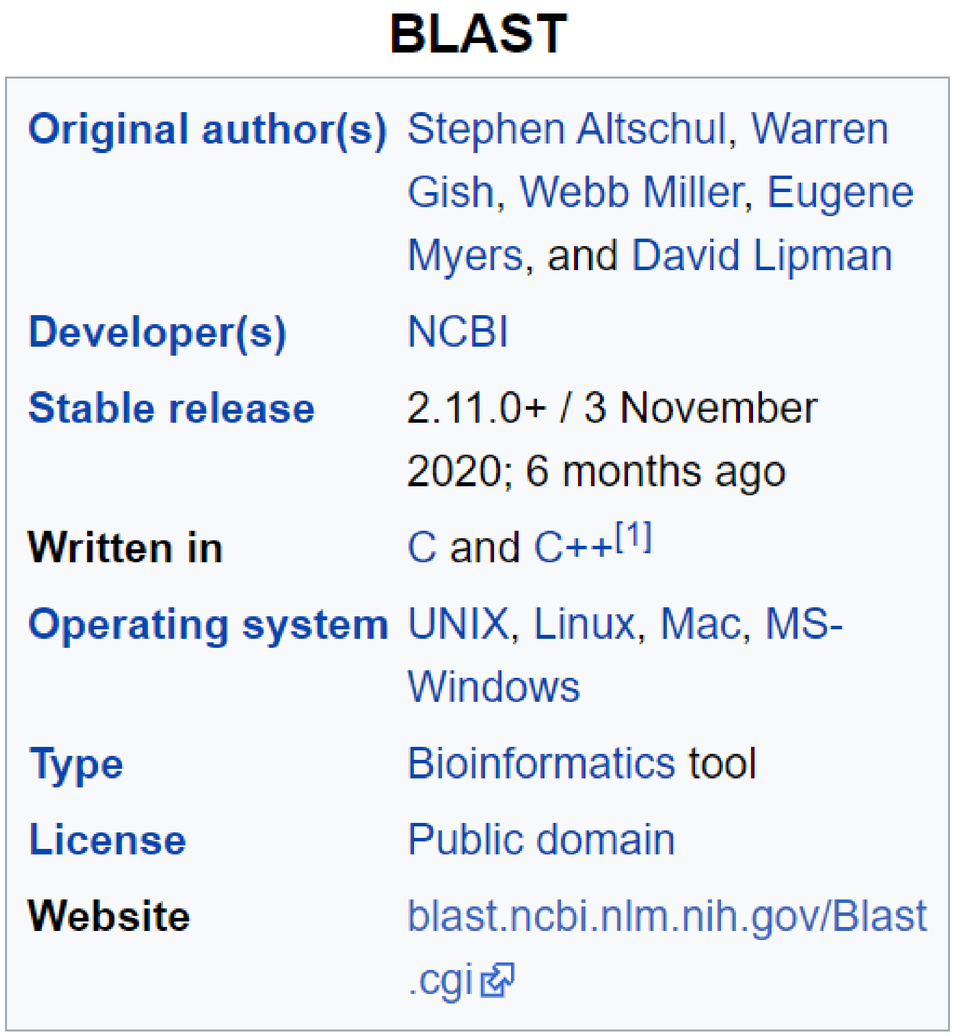
Join segments using gaps, eliminate other segments

Use dynamic programming to create an optimal alignment

16

BLAST

- Basic Local Alignment Search Tool
- Algorithm and program for comparing primary sequence information such as nucleotide or amino acid sequences
- Called “Google of biological research” NYTimes
- Freely available on NCBI and other sites. BLAST program can either be downloaded and run as a command-line utility "blastall" or accessed for free over the web
- Fast, accurate and with statistical rigor



BLAST

Original author(s)	Stephen Altschul, Warren Gish, Webb Miller, Eugene Myers, and David Lipman
Developer(s)	NCBI
Stable release	2.11.0+ / 3 November 2020; 6 months ago
Written in	C and C++[1]
Operating system	UNIX, Linux, Mac, MS-Windows
Type	Bioinformatics tool
License	Public domain
Website	blast.ncbi.nlm.nih.gov/Blast.cgi

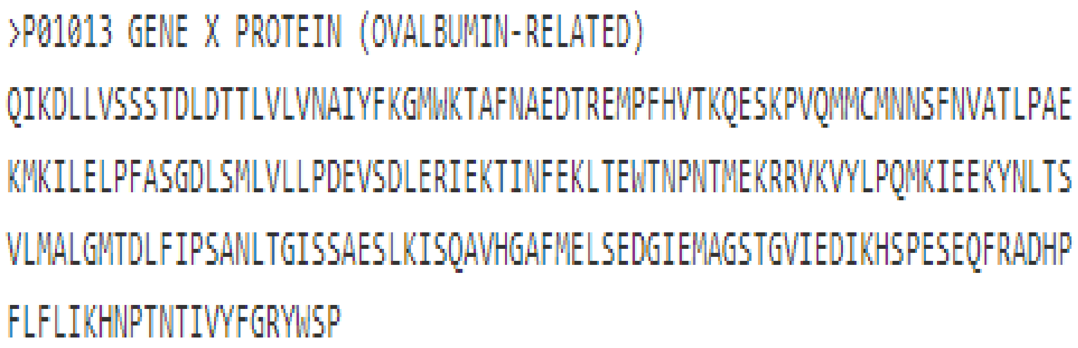
https://en.wikipedia.org/wiki/BLAST_(biotechnology)

BLAST Algorithm

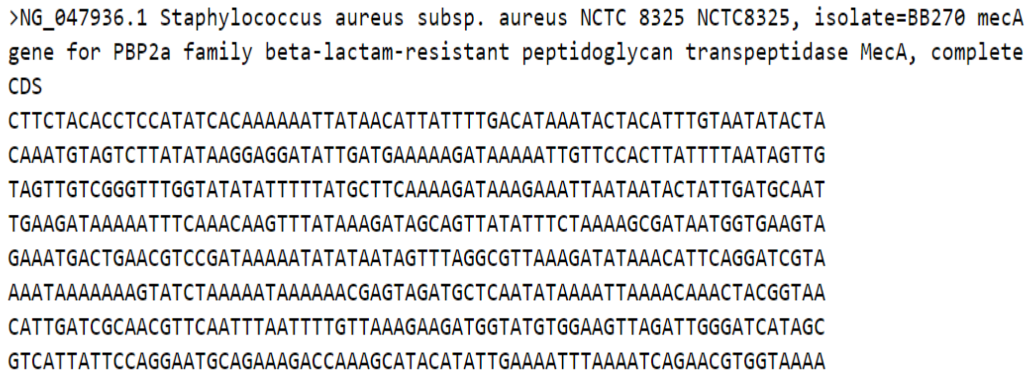
- Heuristic algorithm (practical/approximate/based on trial and error approach). Altschul et al., 1990
- Sequence is split into words , Default n=3 amino acids, 11 bases
- Assumption as FASTA that good alignment contain short lengths of exact matches
- Scoring of matches done using a scoring matrix (PAM/BLOSUM). Scoring matrices area used to calculate score of alignment base by base (Nucleotides) or amino acid by amino acid. Alignment score is sum of scores for each position.
- High Scoring Segment Pair (HSP): BLAST extend initial “seed” hit into HSP using local optimal alignment.
- Quality of each pair wise alignment is represented as a score and scores are ranked

Input: FASTA /Genbank format (query sequence)

- Simple format used by almost all programs
- 1st line: > header with a return/enter at the end
- 2nd line : Sequence no requirement of length or characters



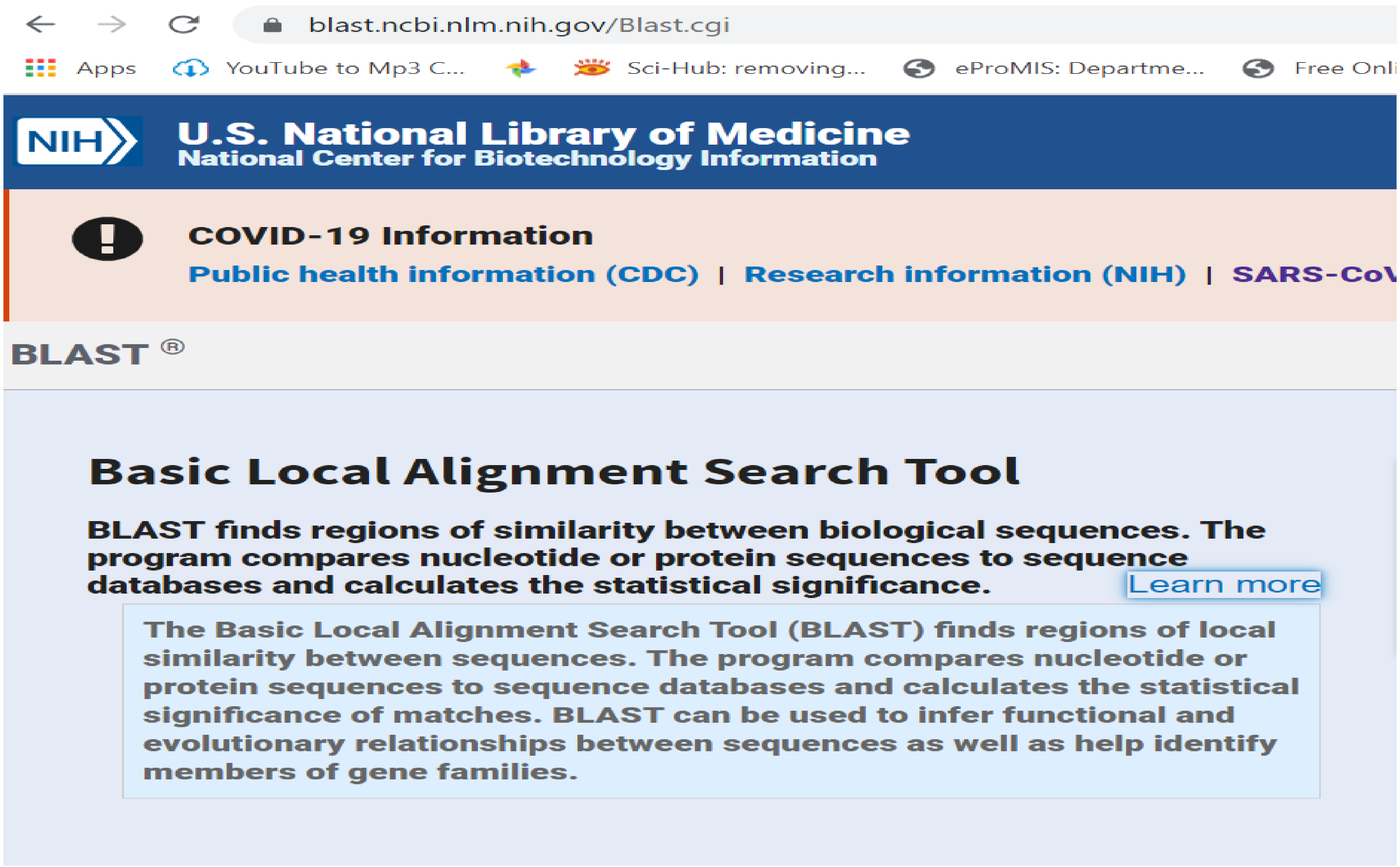
```
>P01013 GENE X PROTEIN (OVALBUMIN-RELATED)
QIKDLLVSSSTDLDTTLVLVNAIYFKGMWKTAFNAEDTREMPFHVTKQESKPVQMMCMNNSFNVATLPAE
KMKILELPFASGDLSMLVLLPDEVSDLERIEKTINFEKLTEWTNPNTMEKRRVKVYLPQMKIEEKYNLTS
VLMALGMTDLFIPSANLTGISSAESLKISQAVHGAFMELSEDGIEMAGSTGVIEDIKHSPESEQFRADHP
FLFLIKHNPTNTIVYFGRYWSP
```



```
>NG_047936.1 Staphylococcus aureus subsp. aureus NCTC 8325 NCTC8325, isolate=BB270 mecA
gene for PBP2a family beta-lactam-resistant peptidoglycan transpeptidase MecA, complete
CDS
CTTCTACACCTCCATATCACAAAAAATTATAACATTATTTTGACATAAATACTACATTTGTAATATACTA
CAAATGTAGTCTTATATAAGGAGGATATTGATGAAAAAGATAAAAATTGTTCCACTTATTTTAATAGTTG
TAGTTGTCGGGTTTGGTATATATTTTTATGCTTCAAAAGATAAAGAAATTAATAATACTATTGATGCAAT
TGAAGATAAAAATTTCAAACAAGTTTATAAAGATAGCAGTTATATTTCTAAAAGCGATAATGGTGAAGTA
GAAATGACTGAACGTCCGATAAAAATATATAATAGTTTAGGCGTTAAAGATATAAACATTCAGGATCGTA
AAATAAAAAAAGTATCTAAAAATAAAAAACGAGTAGATGCTCAATATAAAATTAAAACAAACTACGGTAA
CATTGATCGCAACGTTCAATTTAATTTTGTTAAAGAAGATGGTATGTGGAAGTTAGATTGGGATCATAGC
GTCATTATTCCAGGAATGCAGAAAGACCAAAGCATACATATTGAAAATTTAAAATCAGAACGTGGTAAAA
```

Quality of Alignment

- **Score**: Sum of substitutions and gap scores.
- **E value /Expectation Value**: The number of different alignments with score equivalent to or better that S that are expected in a database by chance. The lower the E value, more significant the score
- Low E values suggest sequence is homologous. Cannot show non-homology
- E value increases as database gets bigger
- E value decreases as alignment get longer
- E value below 10^{-6} are more statistically significant


blast.ncbi.nlm.nih.gov/Blast.cgi
Apps
YouTube to Mp3 C...
Sci-Hub: removing...
eProMIS: Departme...
Free Onli
NIH
U.S. National Library of Medicine
National Center for Biotechnology Information
Public health information (CDC) | Research information (NIH) | SARS-CoV
BLAST ®
Learn more

COVID-19 Information

Basic Local Alignment Search Tool

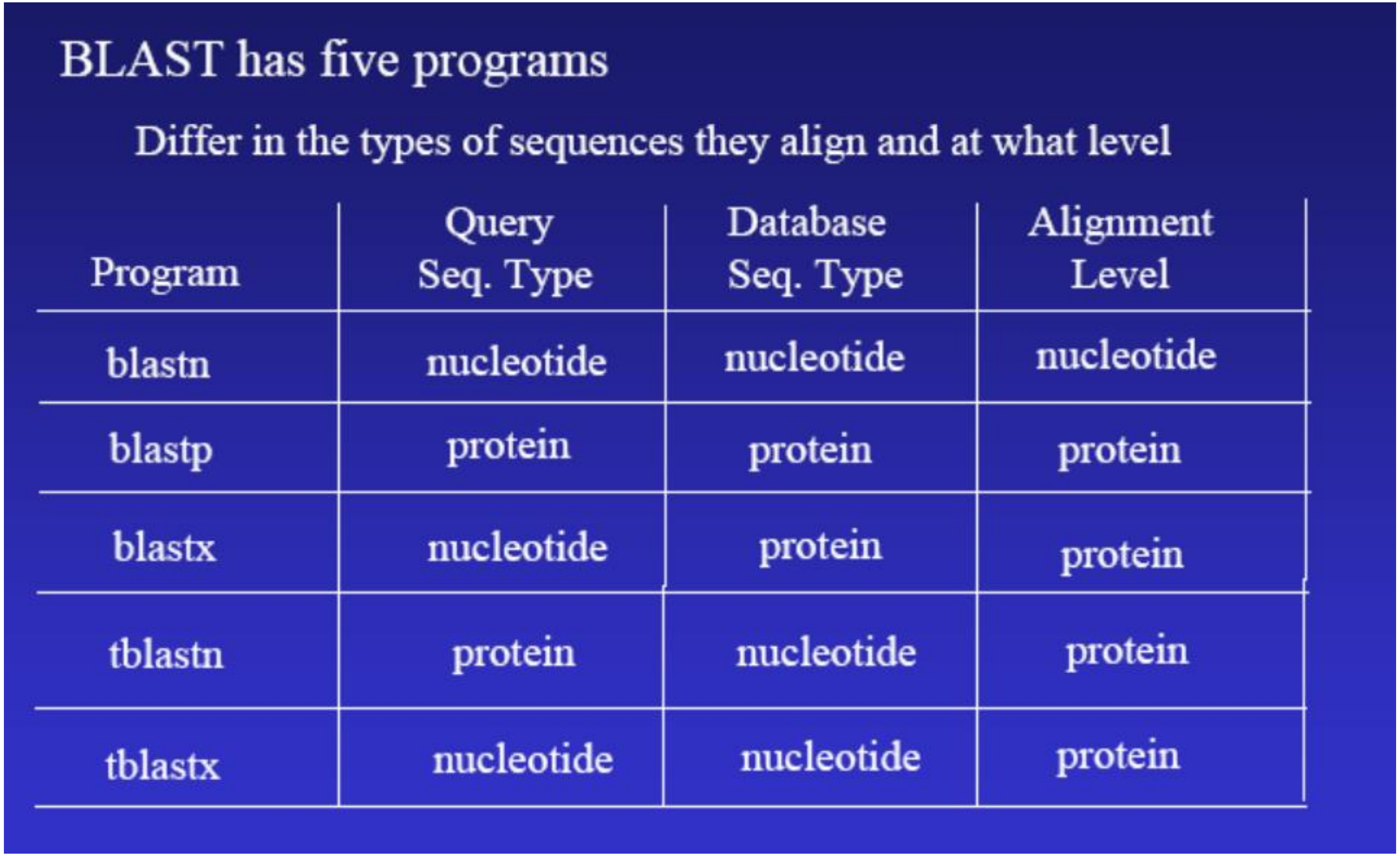
BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance.

The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

BLAST searches GenBank

- NCBI BLAST Webserver allows comparison of query sequence to various sections
- Limit by Organism
- Limit by Entrez
  - Nr: non redundant
  - Refseq_rna (RNA entries from NCBI reference sequence project
  - Refseq_genomic
  - EST
  - Taxon
  - Protein
  - Pdb (sequence derived from 3 D structure from Protein Data Bank

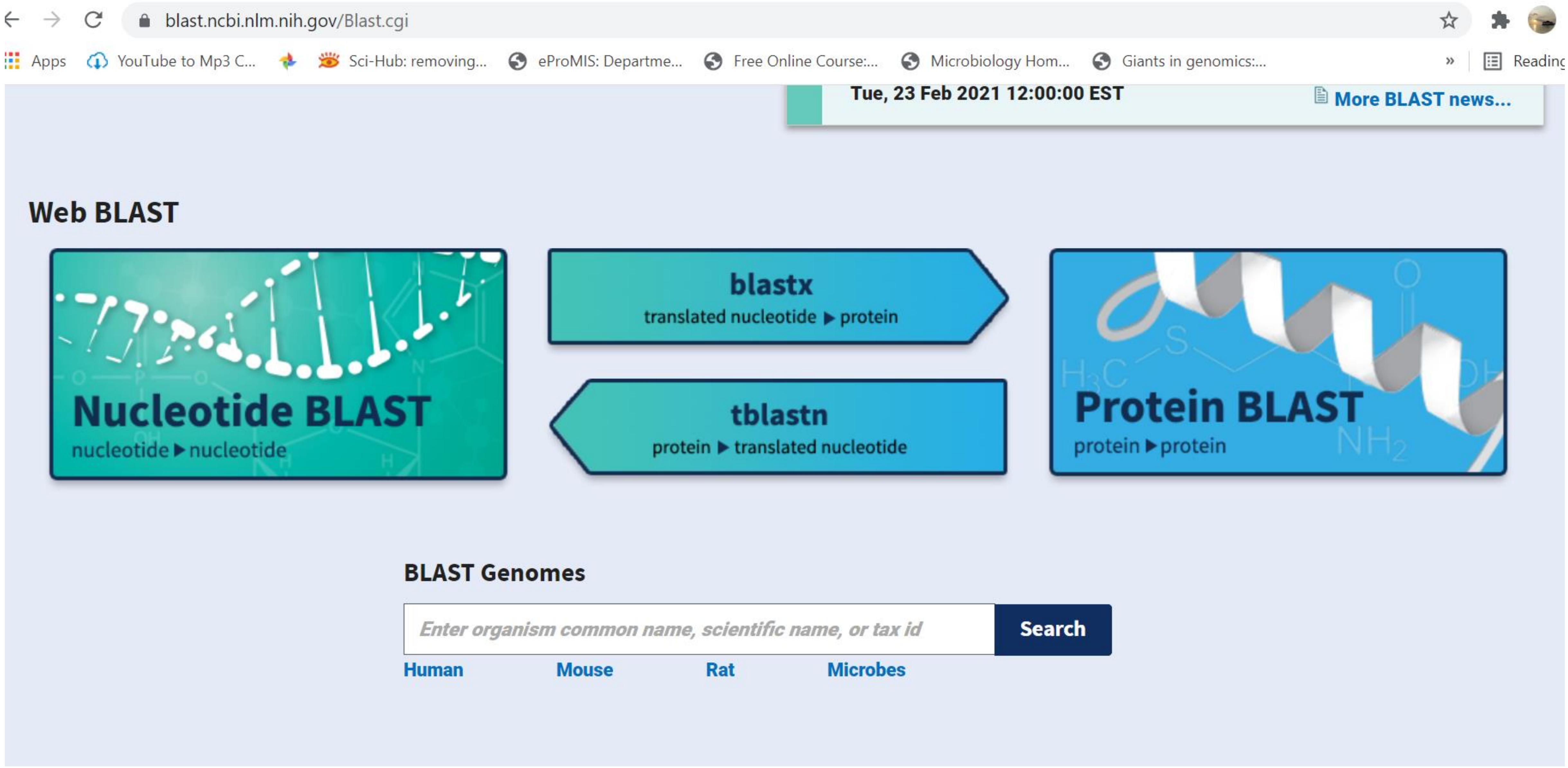
BLAST Variants



BLAST has five programs

Differ in the types of sequences they align and at what level

Program	Query Seq. Type	Database Seq. Type	Alignment Level
blastn	nucleotide	nucleotide	nucleotide
blastp	protein	protein	protein
blastx	nucleotide	protein	protein
tblastn	protein	nucleotide	protein
tblastx	nucleotide	nucleotide	protein



blast.ncbi.nlm.nih.gov/Blast.cgi

Tue, 23 Feb 2021 12:00:00 EST

More BLAST news...

Web BLAST

BLAST Genomes

BLAST for DNA sequences

- Blastn= Compares a nucleotide sequence with nucleotide database

Use for mapping oligonucleotides, cDNA, PCR products, screening repetitive DNA elements, comparative genomics, annotate gene sequence, map genes to genome

- tblastx= compare a DNA translated into protein sequence with a DNA database translated in to protein

Use for cross species gene prediction at genome or transcript level (Expressed Sequence Tags;EST), Searching for genes not yet in protein database

- Blastx= Compares a DNA translated into protein with protein dataset

Use for finding protein coding genes in genomic cDNA if it corresponds to a known protein

MEGABLAST

- Used for comparing large numbers of input sequences via the command-line BLAST
- Faster than running BLAST multiple times.
- It concatenates many input sequences together to form a large sequence before searching the BLAST database
- Analyzes the search results to obtain individual alignments and statistical values
- Discontiguous MEGABLAST: retrives some dissimilar sequences

BLAST for protein sequence

- Blastp= Compares a protein sequence with a protein database

- Use to compare your protein with other proteins in database to identify commone regions between proteins, or perform
- Phylogenetic analysis

- tblastn= compare a protein sequence with a nucleotide database

- Use to discover new genes encoding proteins (from multiple organisms) by comparing protein with DNA sequence
- Translated into their possible six reading frames or to map a protein to genomic DNA

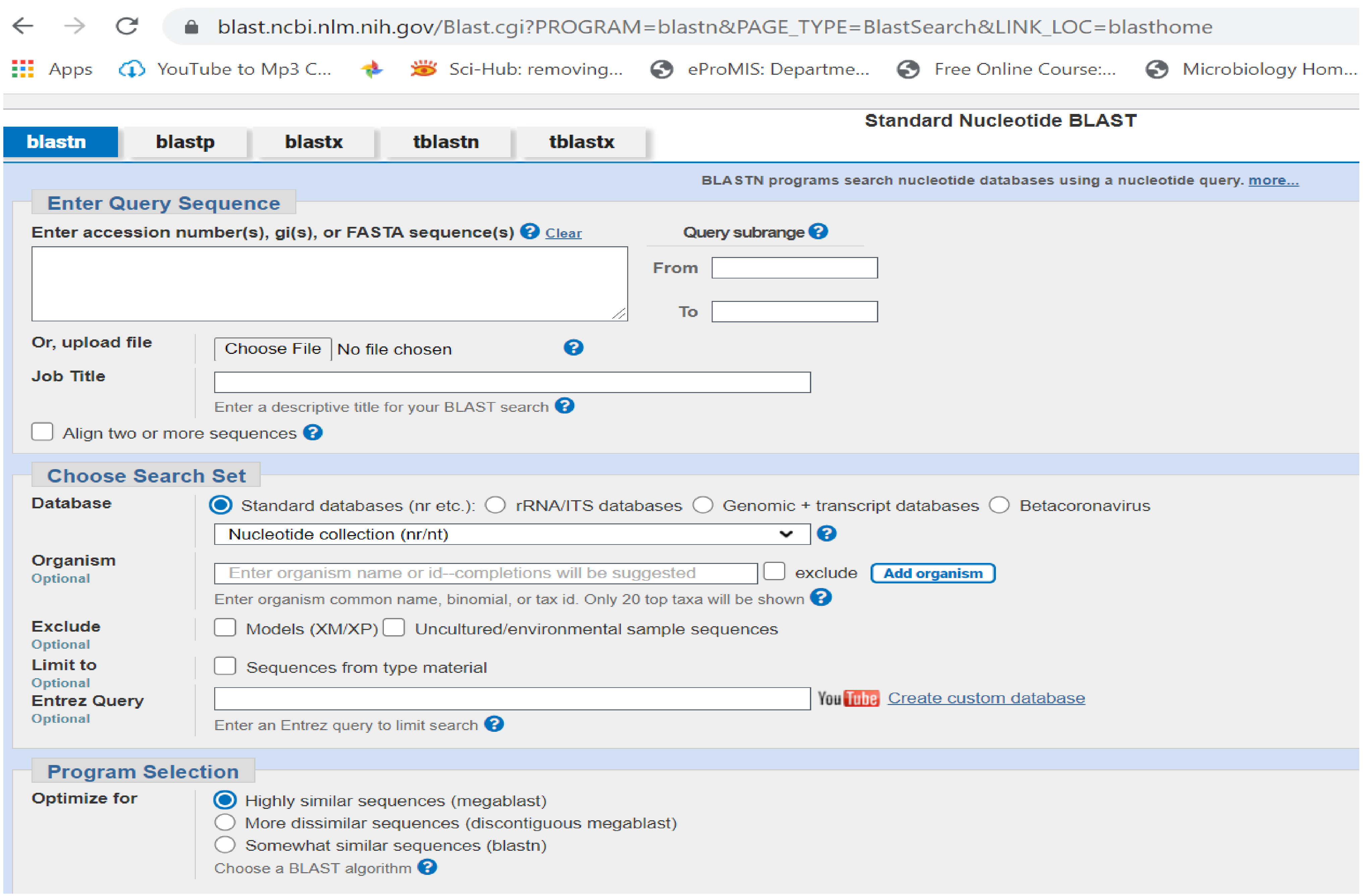
Protein BLAST: PSI- BLAST

- Find distant relatives of proteins compared to BLASTp
- The Position-Specific Iterated BLAST (PSI-BLAST) program performs iterative searches with a protein query, in which sequences found in one round of search are used to build a custom score model for the next round.
- PSI-BLAST first performs a BLASTP search to collect information that it then uses to produce a Position-Specific-Scoring-Matrix (PSSM).
- A PSSM for a query of length N is an N x 20 matrix. Each of the N columns corresponds to a letter in the query, and each column contains 20 rows. Each row corresponds to a specific residue and describes the probability of related sequences having that residue at that position.
- PSI-BLAST can then search a database of protein sequences with this PSSM.

PHI BLAST

- PHI-BLAST (Pattern-Hit Initiated BLAST) is a search program that combines matching of regular expressions with local alignments surrounding the match.
- Given a protein sequence S and a regular expression pattern P occurring in S, PHI-BLAST helps answer the question:

- What other protein sequences both contain an occurrence of P and are homologous to S in the vicinity of the pattern occurrences?
- PHI-BLAST may be preferable to just searching for pattern occurrences because it filters out those cases where the pattern occurrence is probably random and not indicative of homology.



blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome

Apps | YouTube to Mp3 C... | Sci-Hub: removing... | eProMIS: Departme... | Free Online Course:... | Microbiology Hom...

blastn | blastp | blastx | tblastn | tblastx

Standard Nucleotide BLAST

BLASTN programs search nucleotide databases using a nucleotide 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: ● Standard databases (nr etc.): ○ rRNA/ITS databases ○ Genomic + transcript databases ○ Betacoronavirus

Nucleotide collection (nr/nt) ?

Organism
Optional

Enter organism name or id--completions will be suggested □ exclude Add organism

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

Exclude
Optional

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

Limit to
Optional

□ Sequences from type material

Entrez Query
Optional

YouTube Create custom database

Enter an Entrez query to limit search ?

Program Selection

Optimize for

● Highly similar sequences (megablast)

○ More dissimilar sequences (discontiguous megablast)

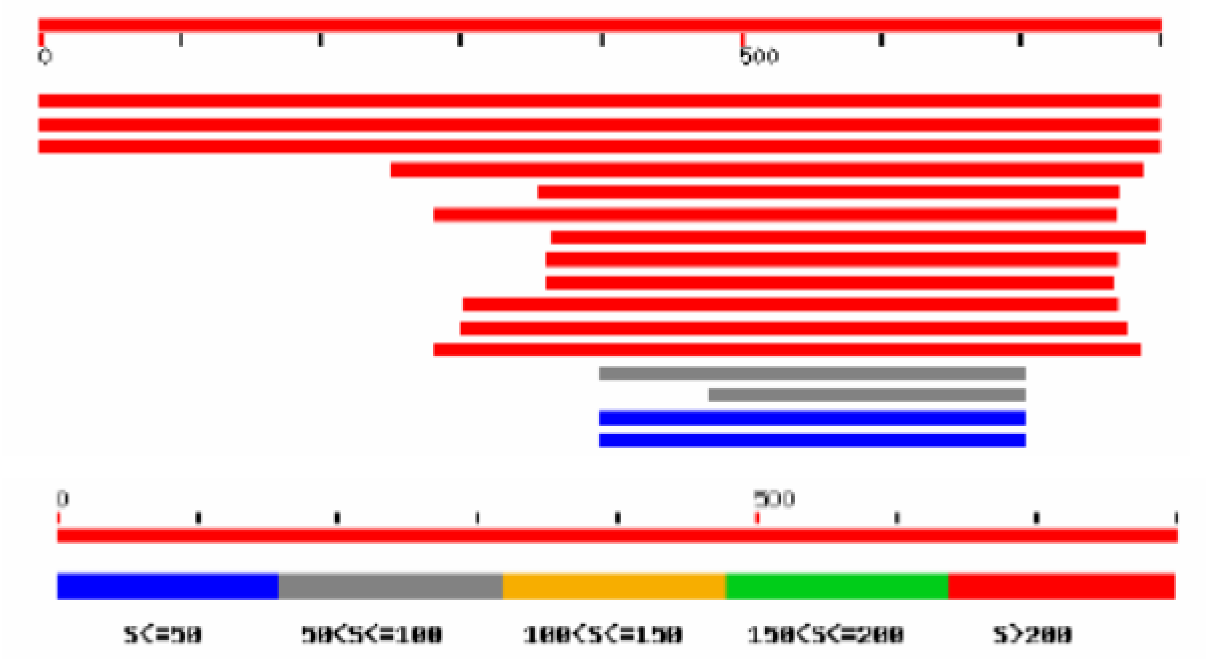
○ Somewhat similar sequences (blastn)

Choose a BLAST algorithm ?

BLAST output

- Graphic display: shows where your query is similar to other sequences
- Hit List: Name of sequence similar to your query, ranked by similarity
- Alignment: every alignment between query and reported hits
- Parameter: list of parameters used for the search

Graphic display

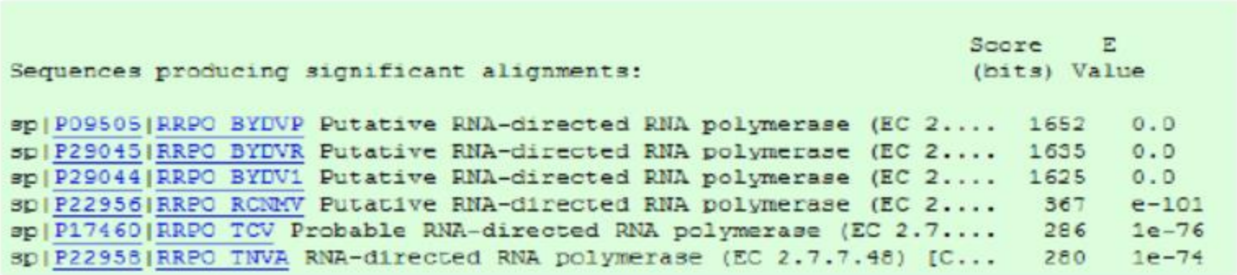


query sequence

Portion of another sequence similar to your query sequence:

red, green, ochre, matches: good
grey matches: intermediate
blue: bad, (twilight zone)

Hit List



```
                                                                    Score    E
Sequences producing significant alignments:                        (bits) Value

sp|P09505|RRPO_BYDVP Putative RNA-directed RNA polymerase (EC 2....  1652   0.0
sp|P29045|RRPO_BYDVR Putative RNA-directed RNA polymerase (EC 2....  1635   0.0
sp|P29044|RRPO_BYDV1 Putative RNA-directed RNA polymerase (EC 2....  1625   0.0
sp|P22956|RRPO_RCNMV Putative RNA-directed RNA polymerase (EC 2....   367   e-101
sp|P17460|RRPO_TCV Probable RNA-directed RNA polymerase (EC 2.7....   286   1e-76
sp|P22958|RRPO_TNVA RNA-directed RNA polymerase (EC 2.7.7.48) [C...   280   1e-74
```

Sequence ac number and name Description Bit score E-value

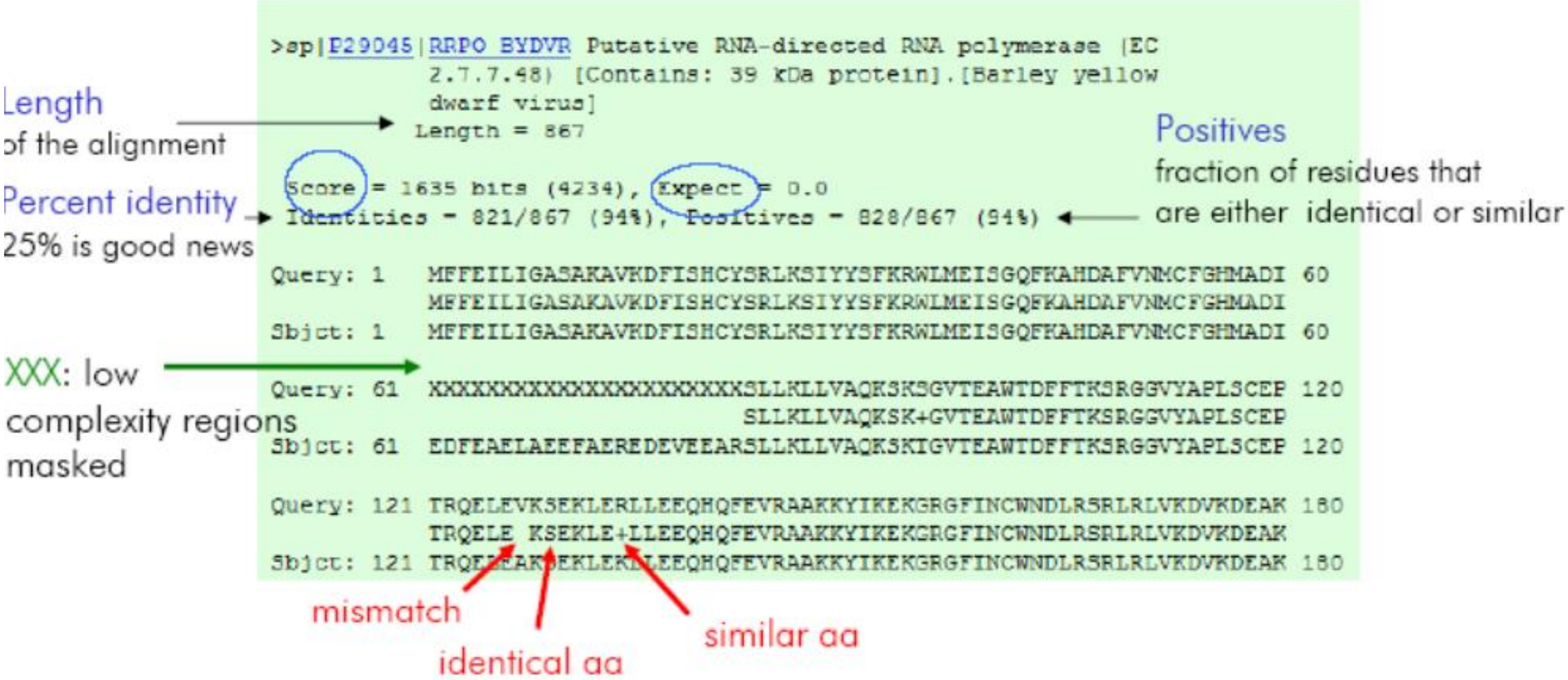
- Sequence ac number and name: Hyperlink to the database entry: useful annotations
- Description: better to check the full annotation

- Bit score (normalized score) : A measure of the similarity between the two sequences:
  the higher the better (matches below 50 bits are very unreliable)

- E-value: The lower the E-value, the better. Sequences identical to the query have an E-value of 0.

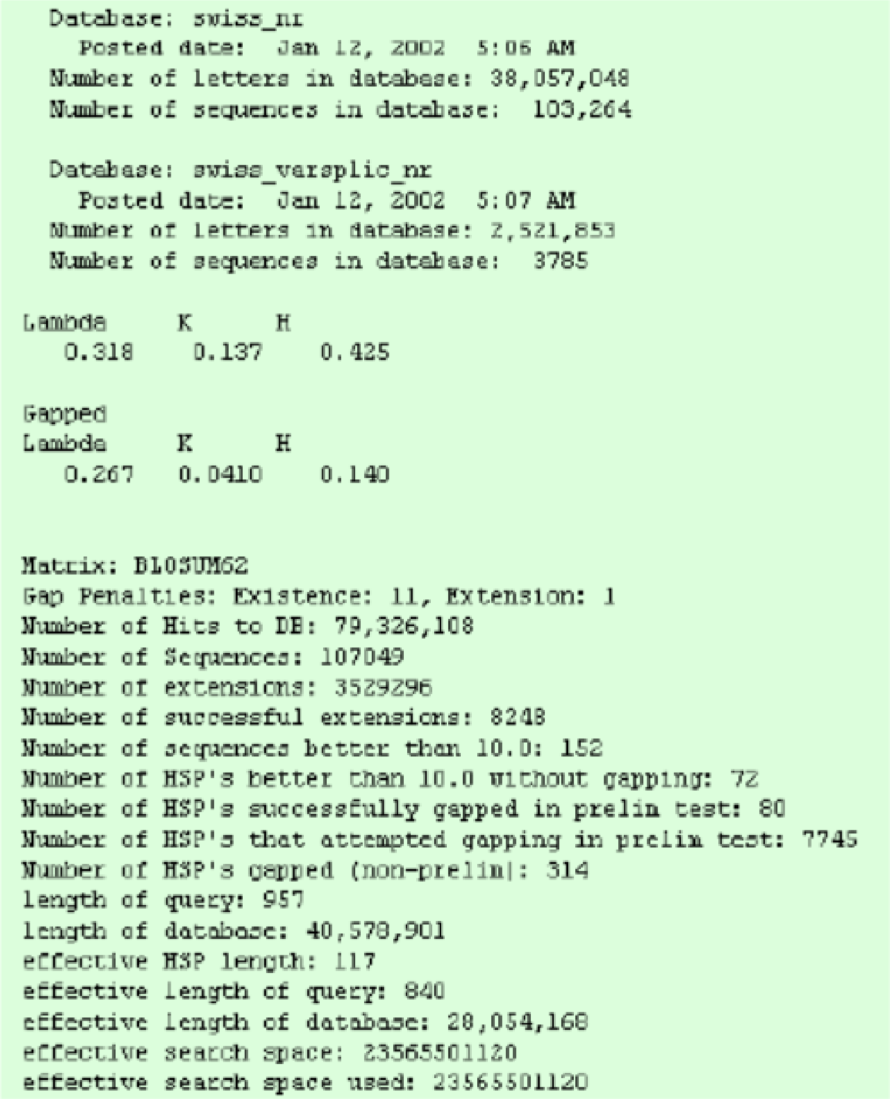
Matches above 0.001 are often close to the twilight zone. As a rule-of-thumb an E-value above 10-4 (0.0001) is not necessarily interesting. If you want to be certain of the homology, your E-value must be lower than 10^{-4}

Alignment



```
>sp|P29045|RRPO_BYDVR Putative RNA-directed RNA polymerase (EC
          2.7.7.48) [Contains: 39 kDa protein].[Barley yellow
          dwarf virus]
         Length = 867

 Score = 1635 bits (4234), Expect = 0.0
 Identities = 821/867 (94%), Positives = 828/867 (94%)

Query: 1   MFFEILIGASAKAVKDFISHCYSRLKSIYYSFKRWLMEISGQFKAHDAFVNMCFGHMADI 60
            MFFEILIGASAKAVKDFISHCYSRLKSIYYSFKRWLMEISGQFKAHDAFVNMCFGHMADI
Sbjct: 1   MFFEILIGASAKAVKDFISHCYSRLKSIYYSFKRWLMEISGQFKAHDAFVNMCFGHMADI 60

Query: 61  XXXXXXXXXXXXXXXXXXXXXXSLLKLLVAQKSKSGVTEAWTDFFTKSRGGVYAPLSCEP 120
                                  SLLKLLVAQKSK+GVTEAWTDFFTKSRGGVYAPLSCEP
Sbjct: 61  EDFEAELAEEFAEREDEVEEARSLLKLLVAQKSKTGVTEAWTDFFTKSRGGVYAPLSCEP 120

Query: 121 TRQELEVKSEKLERLLEEQHQFEVRAAKKYIKEKGRGFINCWNDLRSRLRLVKDVKDEAK 180
            TRQELE KSEKLE+LLEEQHQFEVRAAKKYIKEKGRGFINCWNDLRSRLRLVKDVKDEAK
Sbjct: 121 TRQELEAKSEKLEKLLEEQHQFEVRAAKKYIKEKGRGFINCWNDLRSRLRLVKDVKDEAK 180
```

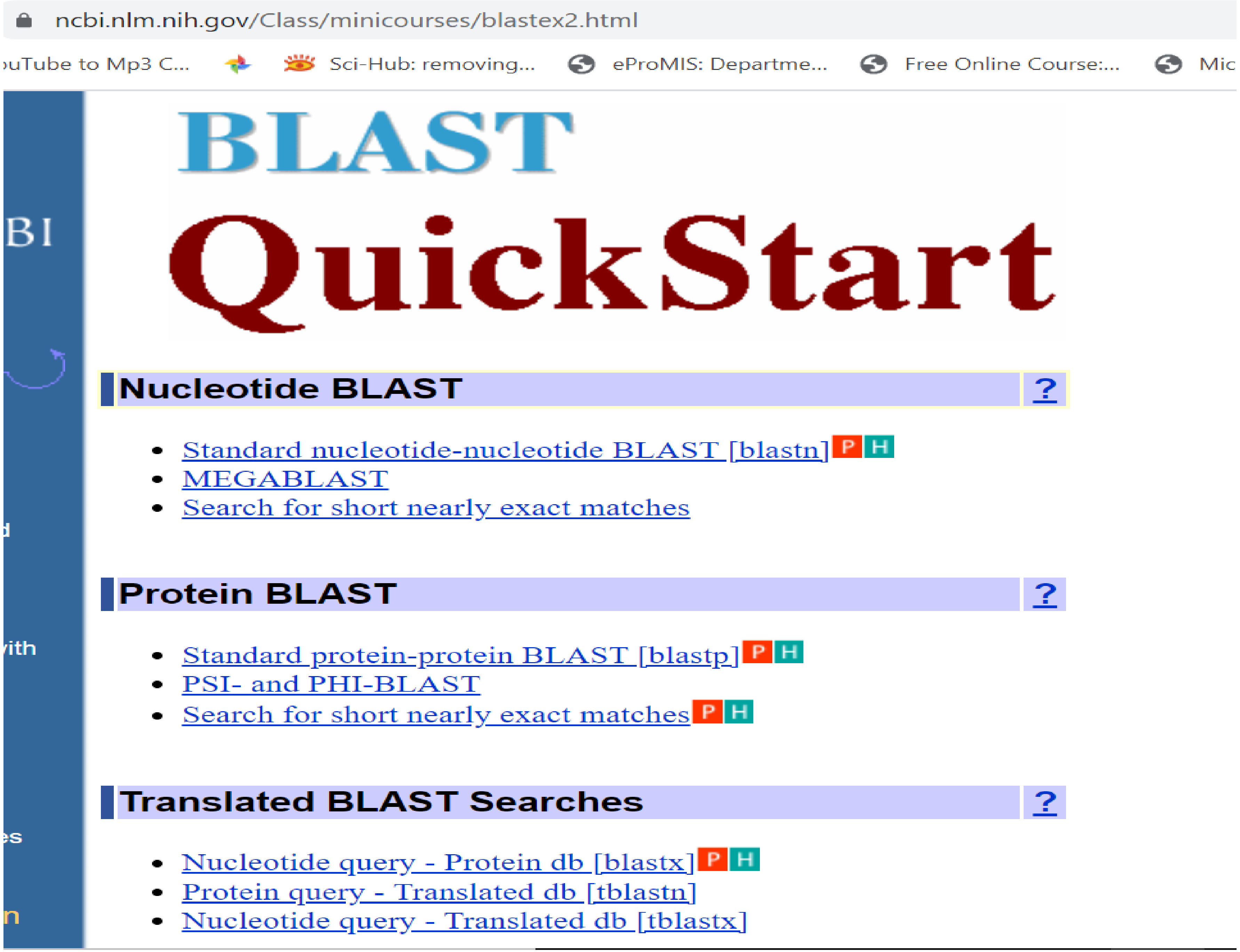
Parameters



```
  Database: swiss_nr
    Posted date:  Jan 12, 2002  5:06 AM
  Number of letters in database: 38,057,048
  Number of sequences in database:  103,264

  Database: swiss_varsplic_nr
    Posted date:  Jan 12, 2002  5:07 AM
  Number of letters in database: 2,521,853
  Number of sequences in database:  3785

Lambda     K      H
   0.318    0.137    0.425

Gapped
Lambda     K      H
   0.267   0.0410    0.140


Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Number of Hits to DB: 79,326,108
Number of Sequences: 107049
Number of extensions: 3529296
Number of successful extensions: 8248
Number of sequences better than 10.0: 152
Number of HSP's better than 10.0 without gapping: 72
Number of HSP's successfully gapped in prelim test: 80
Number of HSP's that attempted gapping in prelim test: 7745
Number of HSP's gapped (non-prelim): 314
length of query: 957
length of database: 40,578,901
effective HSP length: 117
effective length of query: 840
effective length of database: 28,054,168
effective search space: 23565501120
effective search space used: 23565501120
```

Search details (at the bottom of the results)

- Size of the database searched
- Scoring system parameters
- Details about the number of hits found



BLAST QuickStart

Nucleotide BLAST ?

- Standard nucleotide-nucleotide BLAST [blastn] P H
- MEGABLAST
- Search for short nearly exact matches

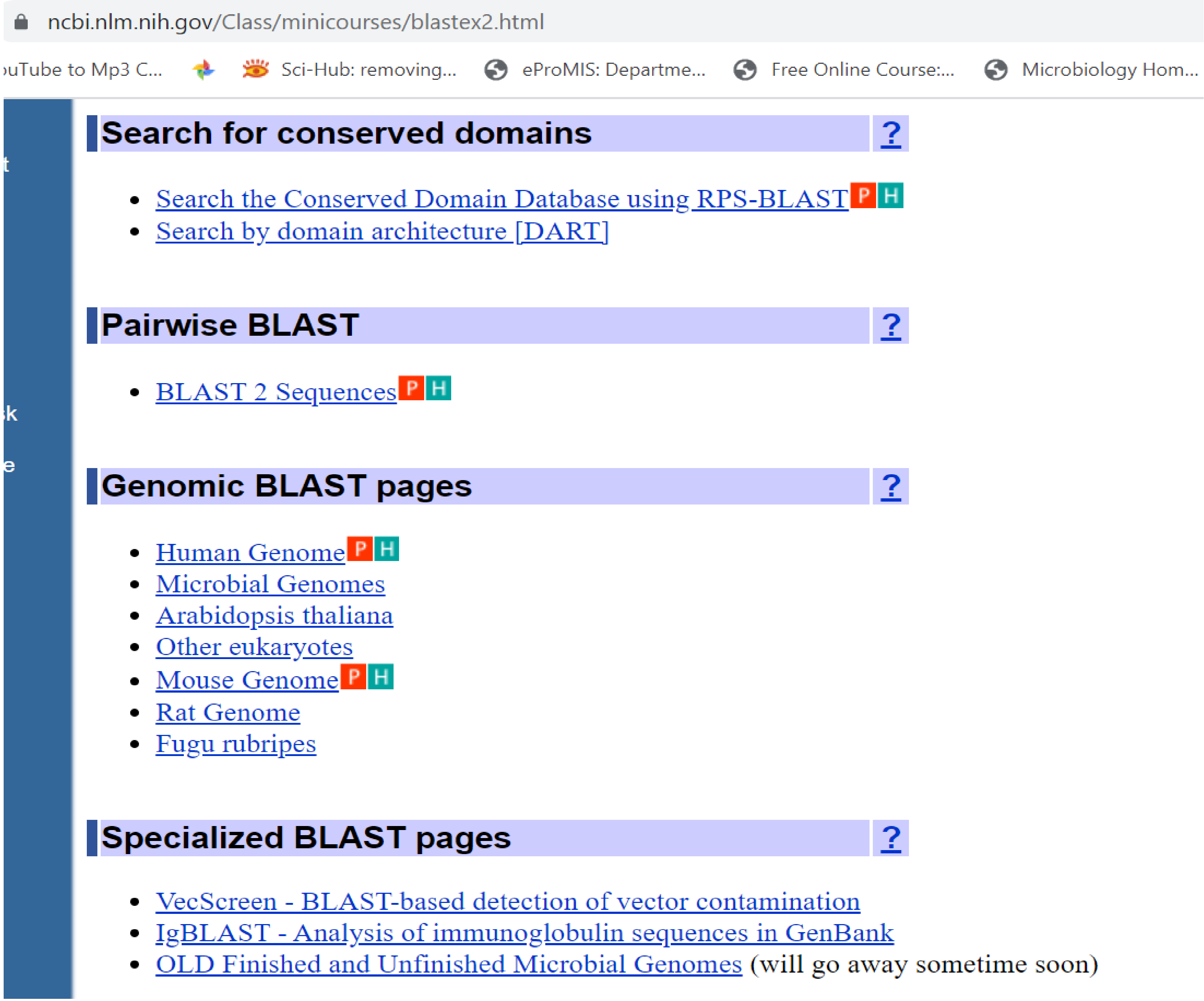
Protein BLAST ?

- Standard protein-protein BLAST [blastp] P H
- PSI- and PHI-BLAST
- Search for short nearly exact matches P H

Translated BLAST Searches ?

- Nucleotide query - Protein db [blastx] P H
- Protein query - Translated db [tblastn]
- Nucleotide query - Translated db [tblastx]

Advanced BLAST



Search for conserved domains

- Search the Conserved Domain Database using RPS-BLAST P H
- Search by domain architecture [DART]

Pairwise BLAST

- BLAST 2 Sequences P H

Genomic BLAST pages

- Human Genome P H
- Microbial Genomes
- Arabidopsis thaliana
- Other eukaryotes
- Mouse Genome P H
- Rat Genome
- Fugu rubripes

Specialized BLAST pages

- VecScreen - BLAST-based detection of vector contamination
- IgBLAST - Analysis of immunoglobulin sequences in GenBank
- OLD Finished and Unfinished Microbial Genomes (will go away sometime soon)

Uses

- Identifying species
- Locating domains
- Gene Annotation
- Comparative Genomics
- Establish phylogenetic relationship: create phylogenetic trees using BLAST. Less reliable than phylogenetic softwares
- DNA mapping: Magic BLAST compare the chromosomal position of the sequence of interest, to relevant sequences in the database

Assignment

- Use the following sequence to perform BLAST
  - MYDAGLYAAPWSCLKGMSWN
  - Ggccatgccatcaggaacgt
  - CAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTAC
- What information can you get about the above unknown sequences

References

https://blast.ncbi.nlm.nih.gov/