Universitas Esa Unggul

Smart, Creative and Entrepreneurial

www.esaunggul.ac.id

PENGANTAR BIOINFORMATIKA
IBT 431

By Seprianto S.Pi, M.Si

Pertemuan 9

Analisis BLAST

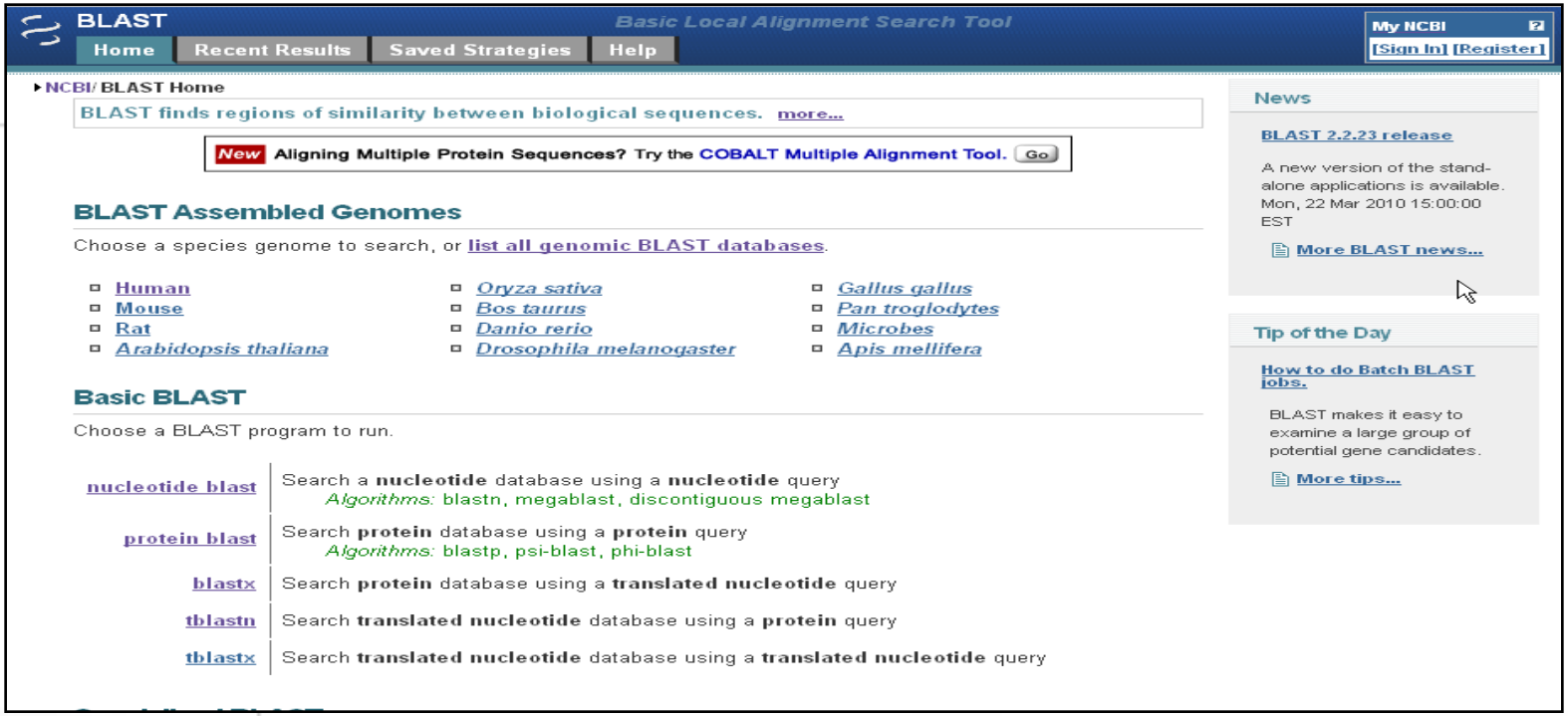
Universitas Esa Unggul

Sasaran Perkuliahan

- Mahasiswa mampu menganalisis hasil sekuens dengan Teknik BLAST (BLASTn, BLASTx, proteinBLAST)
- Mahasiswa Mampu menerjemahkan hasil penelusuran BLAST dari nilai Query Covarage, E-Value dan Maximum identity

What is BLAST?

Free, online service from National Center for Biotechnology Information (NCBI)



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

What is BLAST?

BLAST : Nucleotide/Protein Sequence Databases

Basic Local Alignment Search Tool

as

Google : Internet

Some Uses for BLAST

- Identify an unknown sequence
- Build a homology tree for a protein and Nucleotide
- Get clues about protein structure by finding similar proteins with known structures
- Map a sequence in a genome
- Etc., etc.

BLAST helps you to find homologous genes and proteins

Homologous Proteins (or genes)

- **Have a common ancestor (they're related)**
- **Have similar structures**
- **Have similar functions**

BLAST is Search Tool

- By aligning query sequence against all sequences in a database, alignment can be used to search database for similar sequences
- But alignment algorithms are slow

Why do we need local alignments?

- To compare a short sequence to a large one.
- To compare a single sequence to an entire database
- To compare a partial sequence to the whole.
- Identify newly determined sequences
- Compare new genes to known ones
- Guess functions for entire genomes full of ORFs of unknown function

Pairwise Alignment

Global

- Best score from among alignments of full-length sequences
- Needelman-Wunch algorithm

Local

- Best score from among alignments of partial sequences
- Smith-Waterman algorithm

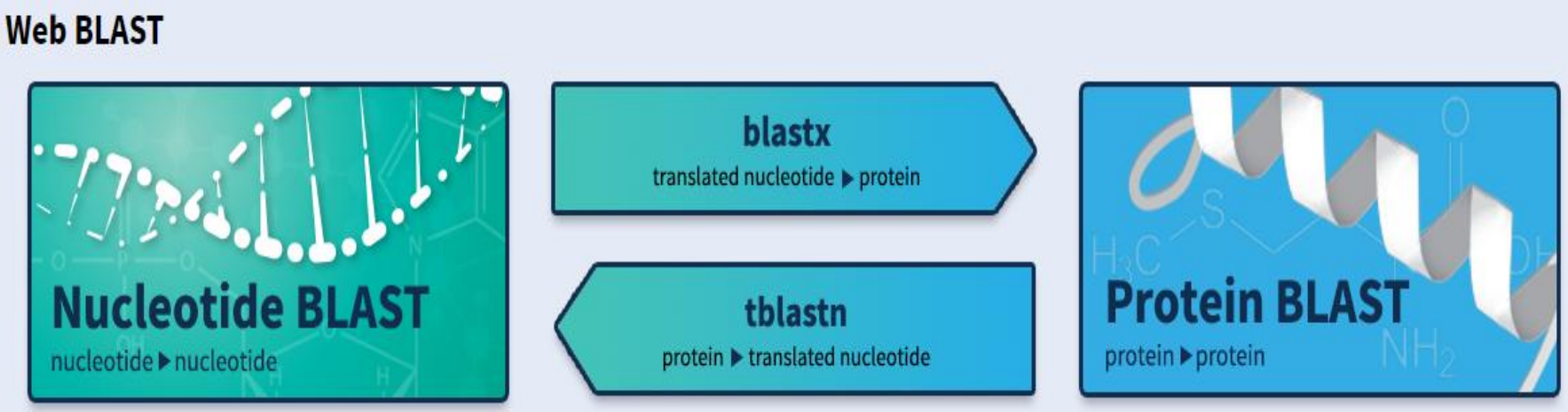
BLAST programs

Program	Description
blastp	Compares an amino acid query sequence against a protein sequence database.
blastn	Compares a nucleotide query sequence against a nucleotide sequence database.
blastx	Compares a nucleotide query sequence translated in all reading frames against a protein sequence database. You could use this option to find potential translation products of an unknown nucleotide sequence.
tblastn	Compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames.

How do I input a query into BLAST?

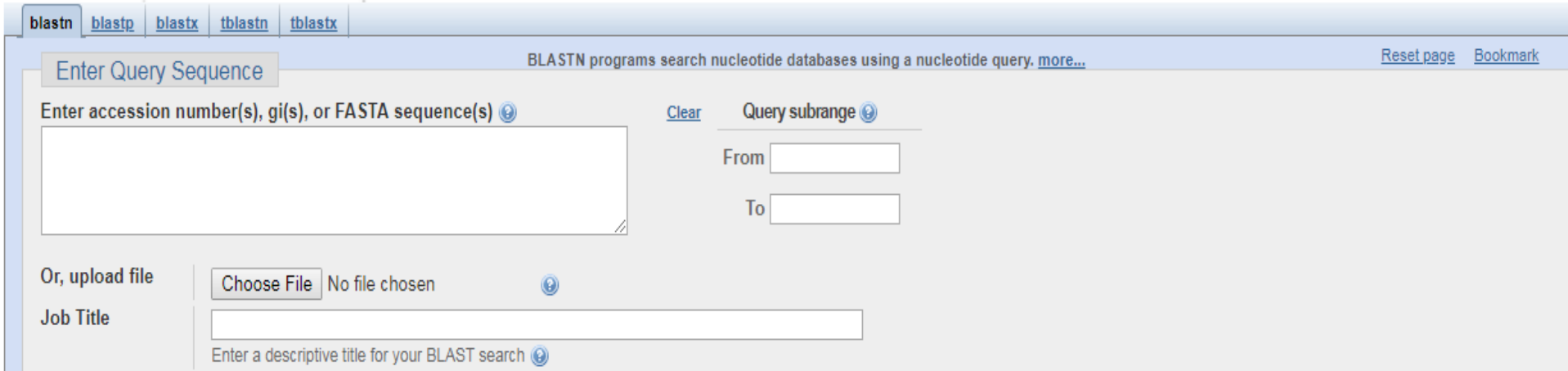
Choose which "flavor" of BLAST to use

- **BLAST comes in many "flavors"**
  - Protein BLAST (BLASTp)
    - Compares a protein query with sequences in GenBank protein database
  - Nucleotide BLAST (BLASTn)
  - BLASTx



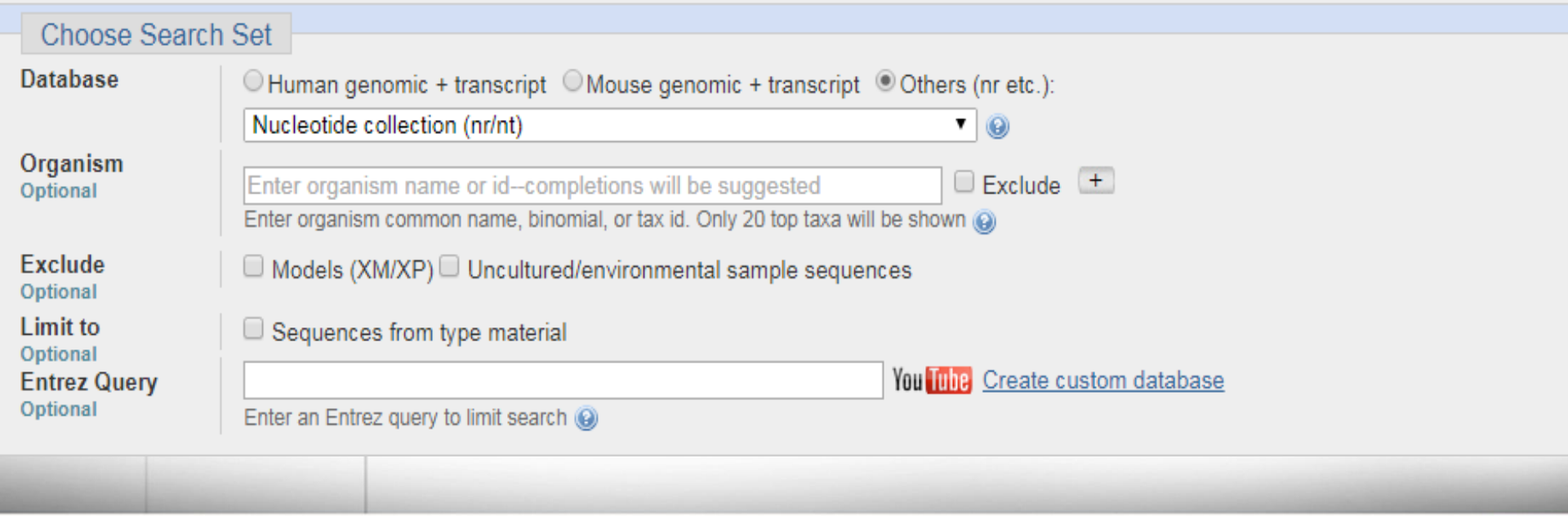
Enter your "query" sequence

- **A sequence can be input as a (an)**
  - **FASTA** format sequence
  - **Accession number**
  - **Choose file**
  - Protein blast can only accept amino acid sequences



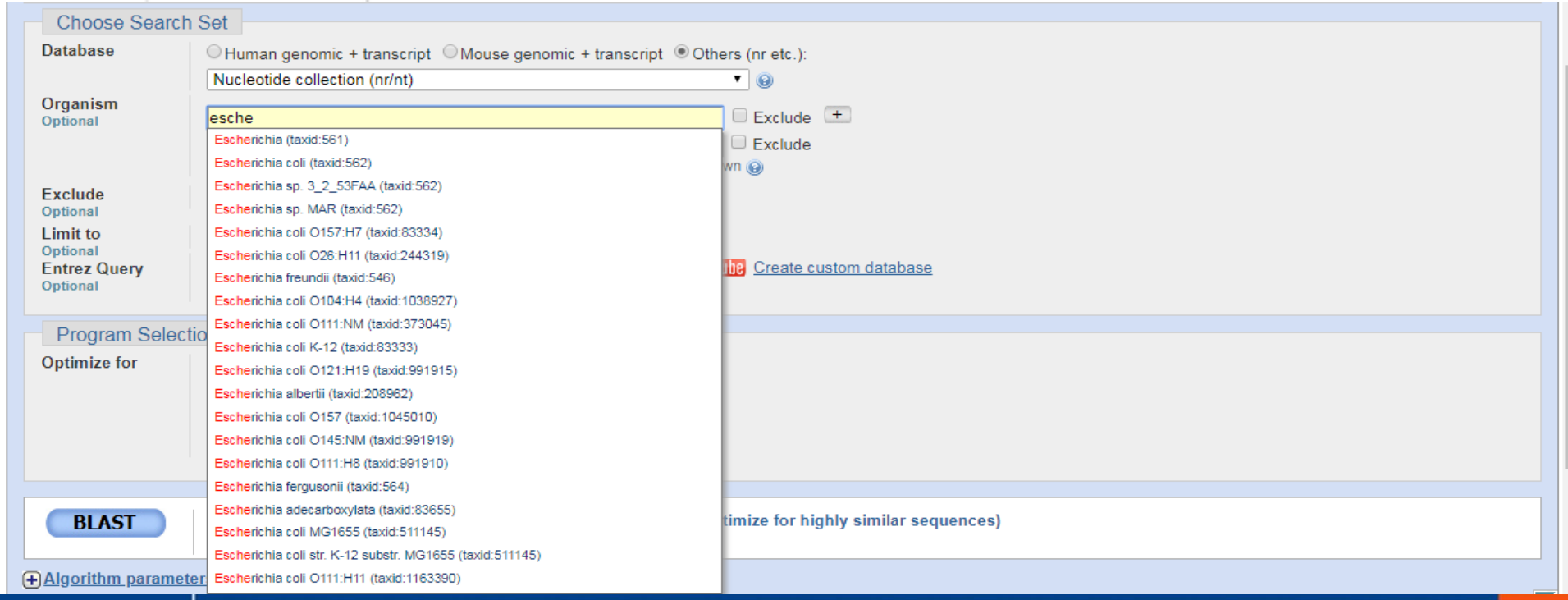
Choose search set

- **Choose which database to search**
  - Default is non-redundant protein sequences (nr)
    - Searches all databases that contain protein sequences



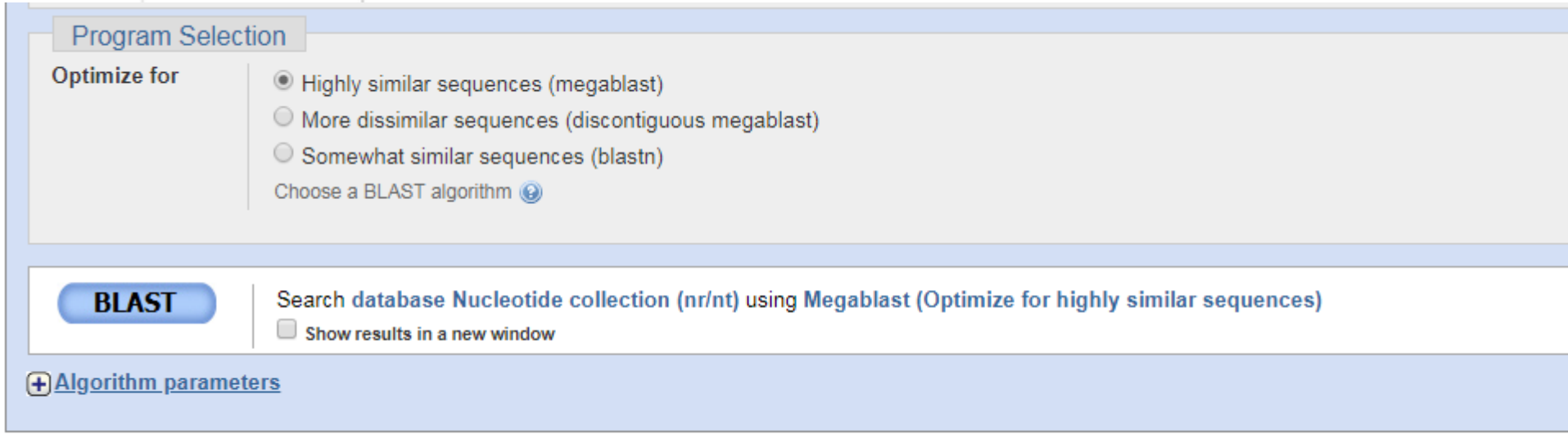
Choose organism

- Default is all organisms represented in databases
- Use this to limit your search to one organism (eg. Yeast)



BLAST off!!

- Choose Highly similar sequence (MegaBLAST)
- Click on the BLAST button at the bottom of the page!



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

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.

NEWS

IgBLAST 1.9.0 released

IgBLAST now supports AIRR rearrangement reports.

Fri, 18 May 2018 08:00:00 EST

More BLAST news...

Web BLAST

Nucleotide BLAST
nucleotide ▶ nucleotide

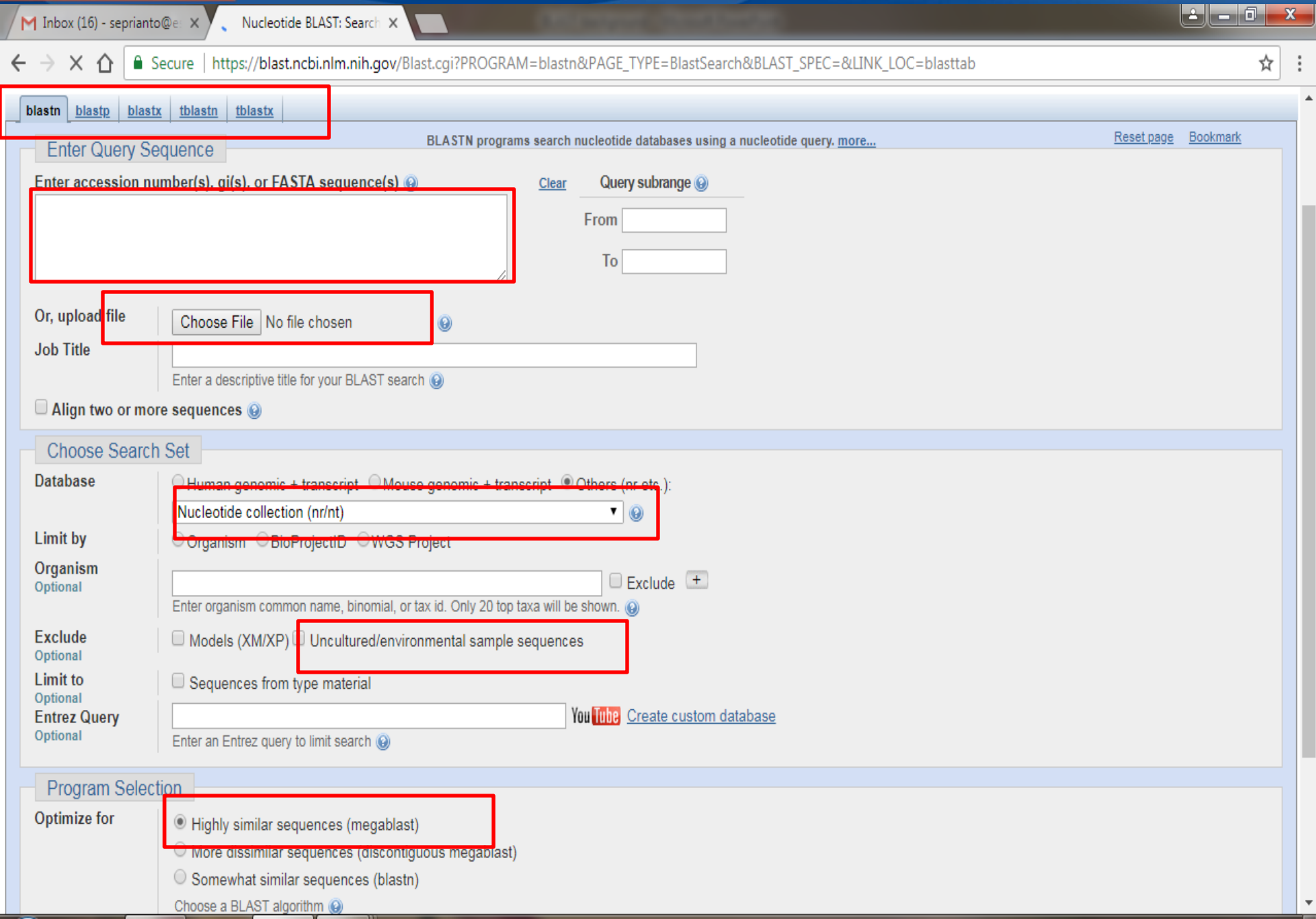
blastx
translated nucleotide ▶ protein

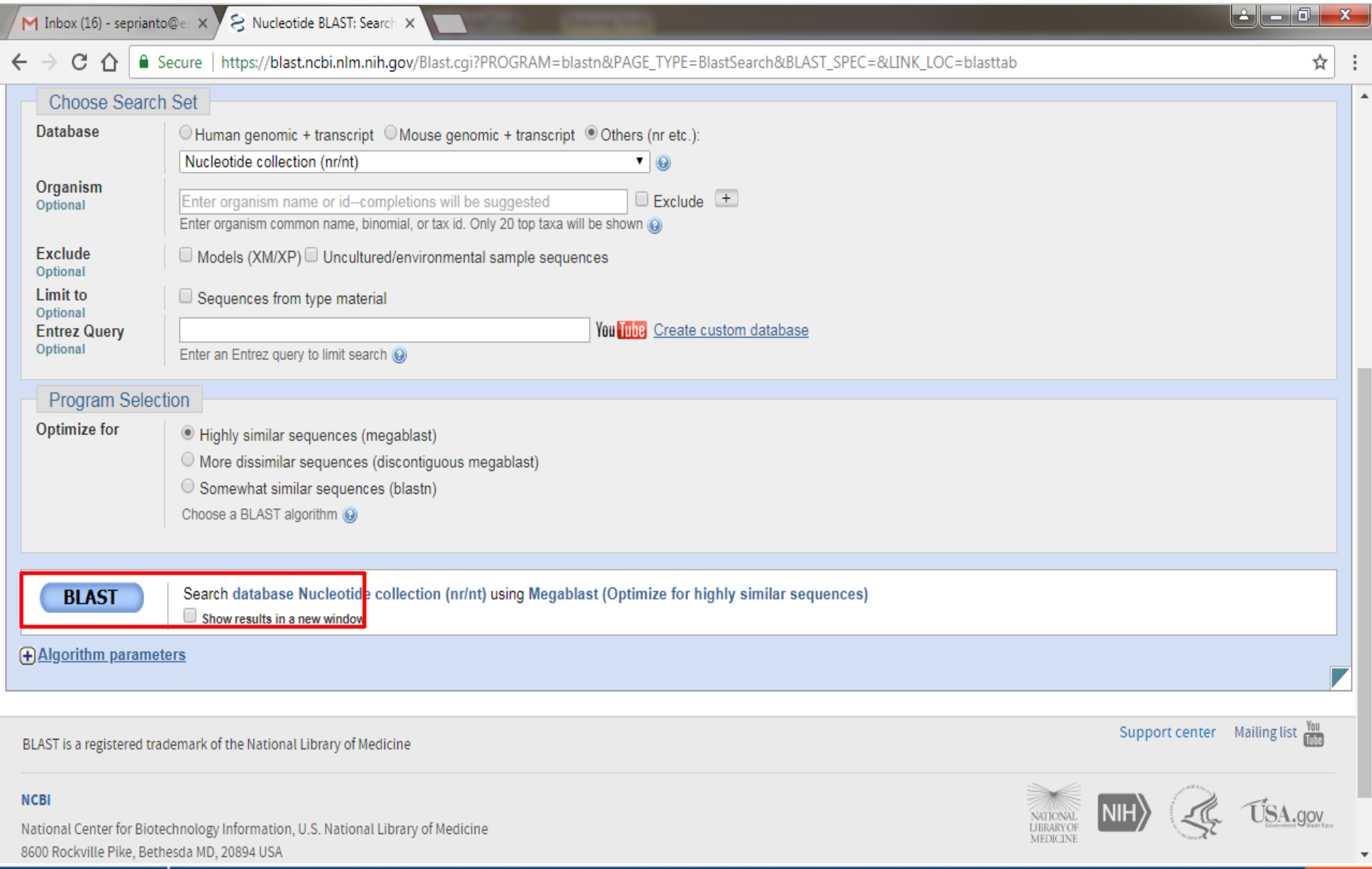
tblastn
protein ▶ translated nucleotide

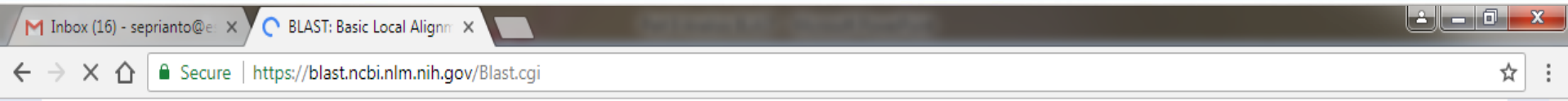
Protein BLAST
protein ▶ protein

BLAST Genomes

Search


Inbox (16) - seprianto@e
Nucleotide BLAST: Search
Secure
https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&BLAST_SPEC=&LINK_LOC=blasttab
blastn
blastp
blastx
tblastn
tblastx
BLASTN programs search nucleotide databases using a nucleotide query. more...
Reset page
Bookmark
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
Human genomic + transcript
Mouse genomic + transcript
Others (nr etc.):
Nucleotide collection (nr/nt)
Limit by
Organism
BioProject ID
WGS Project
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
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


Inbox (16) - seprianto@e
Nucleotide BLAST: Search
Secure
https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&BLAST_SPEC=&LINK_LOC=blasttab
Choose Search Set
Database
Human genomic + transcript
Mouse genomic + transcript
Others (nr etc.):
Nucleotide collection (nr/nt)
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
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
Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)
Show results in a new window
Algorithm parameters
BLAST is a registered trademark of the National Library of Medicine
Support center
Mailing list
NCBI
National Center for Biotechnology Information, U.S. National Library of Medicine
8600 Rockville Pike, Bethesda MD, 20894 USA
NATIONAL LIBRARY OF MEDICINE
NIH
USA.gov



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

How do I interpret the results of a BLAST search?

BLAST creates local alignments

- **What is a local alignment?**
  - **BLAST looks for similarities between regions of two sequences**

```
Global  FGFTALILLAVKV
        F--TAL-LLA--V

Local   FGFTALILL-AVKAV
        --FTAL-LLAAV---
```

The Graphic Display

1. How good is the match?

- **Red = excellent!**
- Pink = pretty good
- **Green = OK, but look at other factors**
- **Blue = bad**
- **Black = really bad!**

2. How long are the matched segments?

Longer = better

Secure | https://blast.ncbi.nlm.nih.gov/Blast.cgi

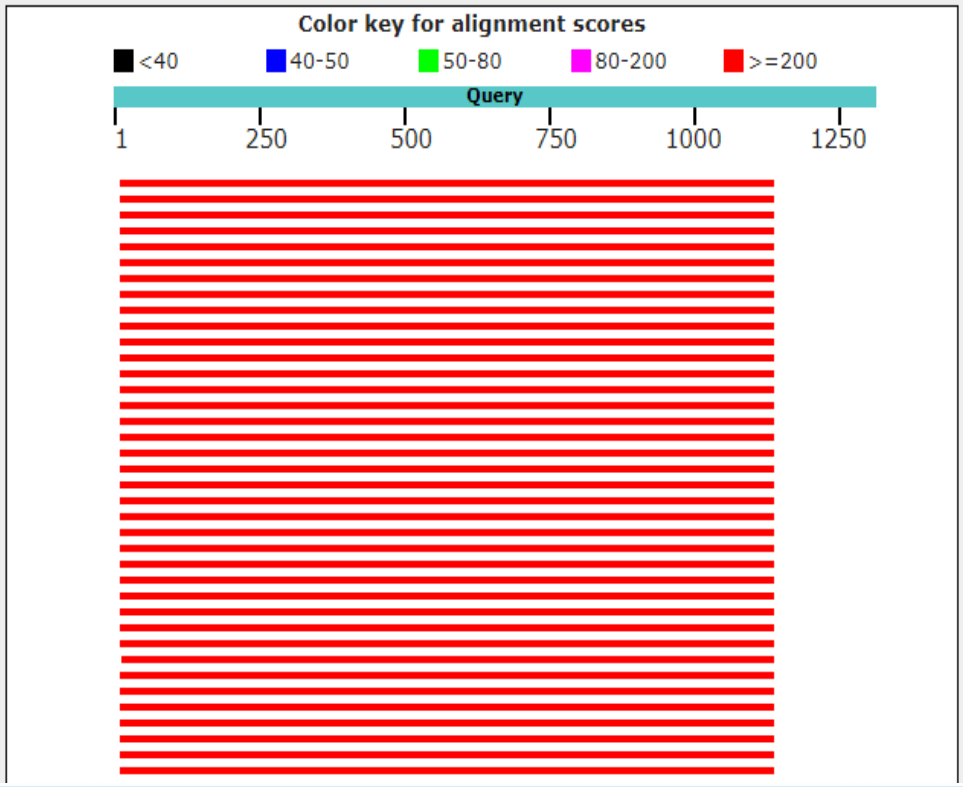
Query ID	lcl|Query_55165	**Database Name**	nr
Description	SPB	**Description**	Nucleotide collection (nt)
Molecule type	nucleic acid	**Program**	BLASTN 2.8.0+ ▷Citation
Query Length	1310		

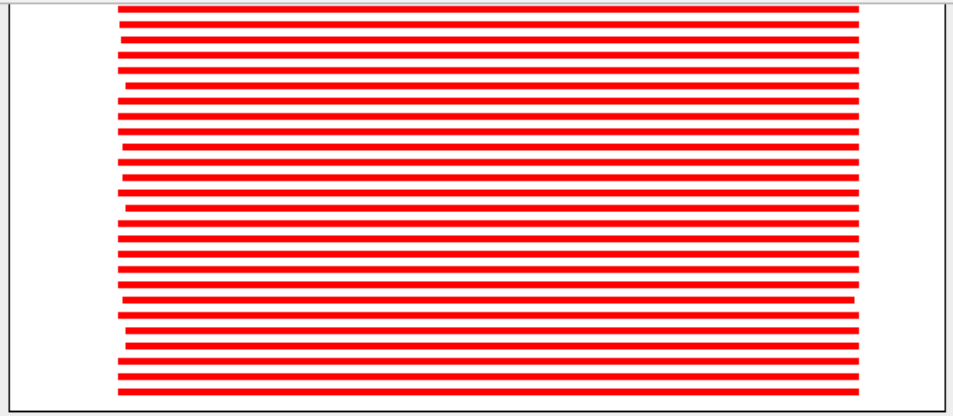
Other reports: ▷Search Summary [Taxonomy reports] [Distance tree of results] [MSA viewer]

Graphic Summary

Distribution of the top 119 Blast Hits on 100 subject sequences

Mouse over to see the title, click to show alignments





Descriptions

Sequences producing significant alignments:

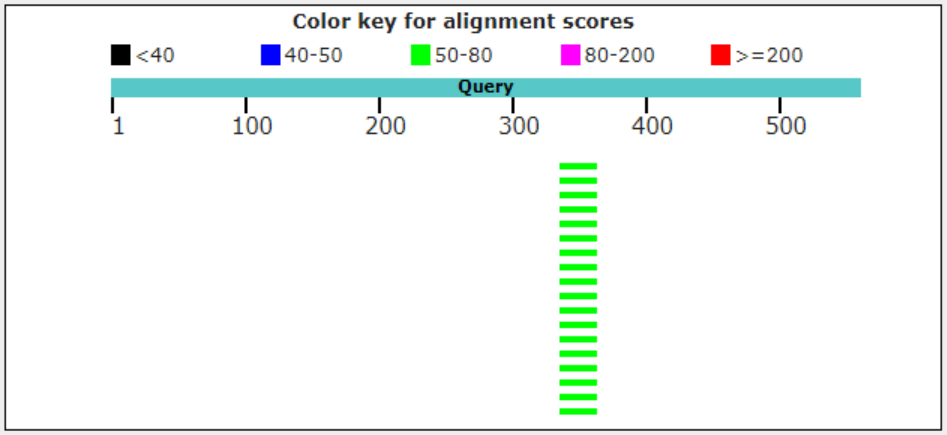
Select: All None Selected:0

Alignments Download GenBank Graphics Distance tree of results

Description	Max score	Total score	Query cover	E value	Ident	Accession
Streptomyces aurantiacogriseus gene for 16S rRNA, partial sequence, strain: NBRC 13668	2584	2584	98%	0.0	99%	AB184451.2
Streptomyces sp. B7 16S ribosomal RNA gene, partial sequence	2579	2579	98%	0.0	99%	AF332543.1
Streptomyces luridus strain YIM 131011 16S ribosomal RNA gene, partial sequence	2555	2555	96%	0.0	99%	KX502956.1
Streptomyces sp. W26 16S ribosomal RNA gene, partial sequence	2553	2553	99%	0.0	99%	EU596428.1
Streptomyces sp. CPE275 16S ribosomal RNA gene, partial sequence	2536	2536	98%	0.0	99%	JN969009.1
Streptomyces setonensis strain HBUM82838 16S ribosomal RNA gene, partial sequence	2536	2536	98%	0.0	99%	EU841570.1
Streptomyces cyaneogriseus strain 179 16S ribosomal RNA gene, partial sequence	2534	2534	98%	0.0	99%	EF063457.1
Streptomyces kurssanovii strain ICTA165 16S ribosomal RNA gene, partial sequence	2531	2531	98%	0		
Streptomyces sp. YH4 16S ribosomal RNA gene, partial sequence	2531	2531	98%	0		

Distribution of the top 18 Blast Hits on 18 subject sequences

Mouse over to see the title, click to show alignments



Descriptions

Sequences producing significant alignments:

Select: All None Selected:0

Alignments | Download | GenBank | Graphics | Distance tree of results

Description	Max score	Total score	Query cover	E value	Ident	Accession
Mycobacterium marinum ATCC 927 DNA, complete genome	52.8	52.8	5%	0.011	100%	AP018496.1
Mycobacterium pseudoshottsii JCM 15466 DNA, complete genome	52.8	52.8	5%	0.011	100%	AP018410.1
Actinosynnema pretiosum strain X47 chromosome, complete genome	52.8	52.8	5%	0.011	100%	CP023445.1
Mycobacterium aurum isolate liquid genome assembly, chromosome: I	52.8	52.8	5%	0.011	100%	LT899394.1
Cupriavidus sp. USMAA1020 chromosome 2, complete sequence	52.8	52.8	5%	0.0		
Cupriavidus sp. USMAHM13 chromosome 2, complete sequence	52.8	52.8	5%	0.0		

Interpreting Results of **BLAST**

- Score:
- Max score:
- Total score:
- Query coverage
- E Value:
- Max Identity:

Score

- Score: jumlah keselarasan semua segmen dari sekuens database yang cocok dengan sekuen nukleotida kita.
- Nilai skor menunjukkan keakuratan nilai penjajaran sekuens berupa nukleotida yang tidak diketahui dengan sekuens nukleotida yang terdapat di dalam GenBank.
- Semakin tinggi nilai skor yang diperoleh maka semakin tinggi tingkat homologi kedua sekuens
- Max score: Score of single best aligned sequence
- Total score: Sum of scores of all aligned sequences

Query coverage

- *Query coverage* adalah persentasi dari panjang nukleotida yang sesuai dengan sekuen database yang terdapat pada BLAST

Max identity

- *Max identity* adalah nilai tertinggi dari persentasi identitas atau kecocokan antara sekuen nukleotida dengan sekuen database yang tersejajarkan
- **Hagstrom *et al* (2000)** menyatakan bahwa bakteri yang mempunyai nilai max identity 16S rRNA lebih besar dari 97% adalah **spesies yang sama.**
- Sedangkan persamaan sekuen antara 93%-97%dapat mewakili identitas pada tingkat **genus** tetapi berbeda pada tingkat spesies

What is an E-value?

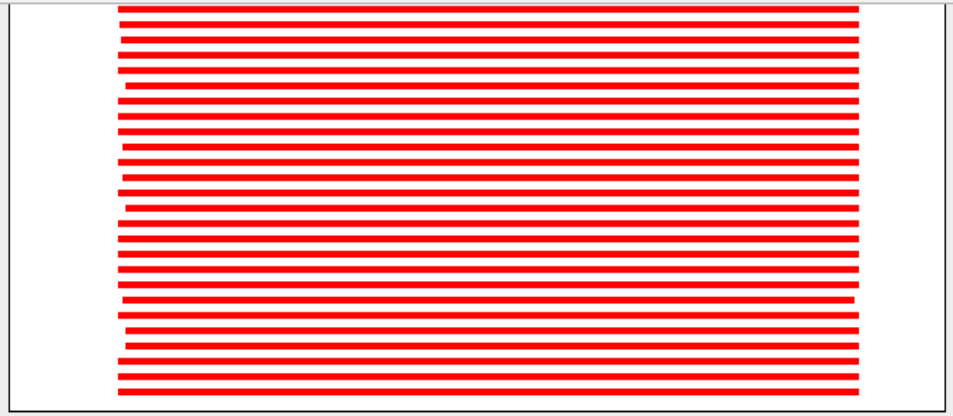
- **E-value**
  - **The chance that the match could be random**
  - **The lower the E-value, the more significant the match**
    - **E = 10^{-4}** is considered the **cutoff point**
    - **E = 0** means that the two sequences are statistically **identical**

E-value

- Nilai E-value merupakan nilai dugaan yang memberikan ukuran statistic yang signifikan terhadap kedua sekuen.
- Nilai E-value yang semakin tinggi menunjukkan tingkat homologi antara sekuen semakin rendah, sedangkan nilai E-value yang semakin rendah menunjukkan tingkat homologi antar sekuens semakin tinggi.
- Nilai E-value bernilai 0 (nol) menunjukkan bahwa kedua sekuen tersebut identik

Accession Number

- Accession number (bioinformatics), a unique identifier given to a biological polymer sequence (DNA, protein) when it is submitted to a sequence database.



Descriptions

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenBank Graphics Distance tree of results

Description	Max score	Total score	Query cover	E value	Ident	Accession
Streptomyces aurantiacogriseus gene for 16S rRNA, partial sequence, strain: NBRC 13668	2584	2584	98%	0.0	99%	AB184451.2
Streptomyces sp. B7 16S ribosomal RNA gene, partial sequence	2579	2579	98%	0.0	99%	AF332543.1
Streptomyces luridus strain YIM 131011 16S ribosomal RNA gene, partial sequence	2555	2555	96%	0.0	99%	KX502956.1
Streptomyces sp. W26 16S ribosomal RNA gene, partial sequence	2553	2553	99%	0.0	99%	EU596428.1
Streptomyces sp. CPE275 16S ribosomal RNA gene, partial sequence	2536	2536	98%	0.0	99%	JN969009.1
Streptomyces setonensis strain HBUM82838 16S ribosomal RNA gene, partial sequence	2536	2536	98%	0.0	99%	EU841570.1
Streptomyces cyaneogriseus strain 179 16S ribosomal RNA gene, partial sequence	2534	2534	98%	0.0	99%	EF063457.1
Streptomyces kurssanovii strain ICTA165 16S ribosomal RNA gene, partial sequence	2531	2531	98%	0		
Streptomyces sp. YH4 16S ribosomal RNA gene, partial sequence	2531	2531	98%	0		

TUGAS PRAKTIKUM ---- KUMPULKAN MINGGU DEPAN

1. Buat Lah langkah langkah dalam analisis sekuens dengan teknik BLAST
2. Interpretasikanlah hasil BLAST berdasarkan E-Value, Max score, Total Score, Max Identified and Query Coverage


THANK
YOU
Novi Jan's Blog