Basic bioinformatics for gene and protein sequences analysis

Thewarach Laha, PhD
Department of Parasitology
Faculty of Medicine
KKU

Contents

- Searching and retrieving DNA and protein sequences from public databases
- Sequence similarity search (Blast search)
- Sequence manipulation (using Bioedit sofware)

Data retrieving

Application:

Databases

- Pubmed- reference journal
- DNA
- Protein
- Genome

Web site

- NCBI - http://www.ncbi.nlm.nih.gov/
- EBI - http://www.ebi.ac.uk/
- Sanger centre
  http://www.sanger.ac.uk/

Practice

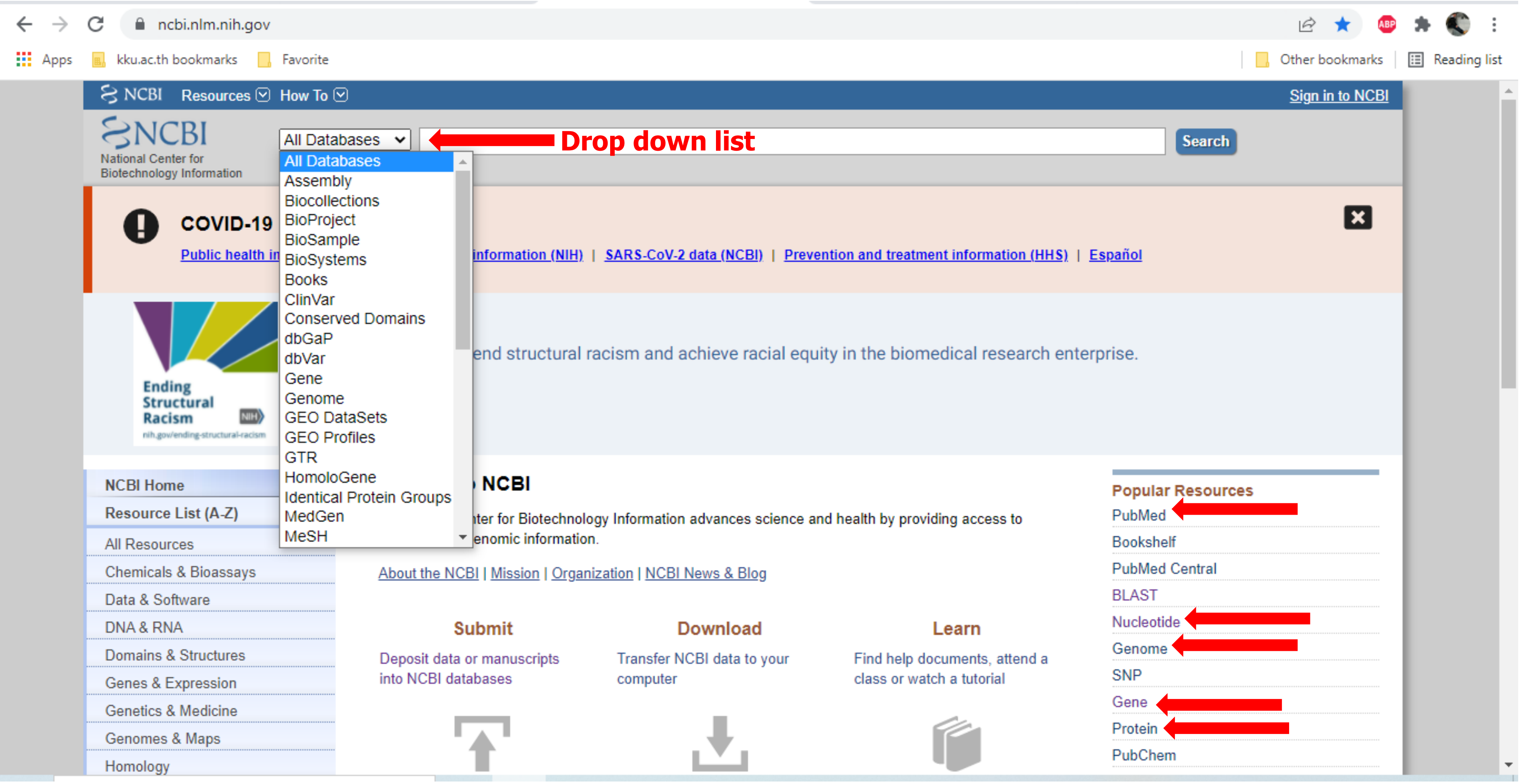
- https://www.ncbi.nlm.nih.gov/

Retrieving data

- PubMed
- Nucleotide
- Proteins

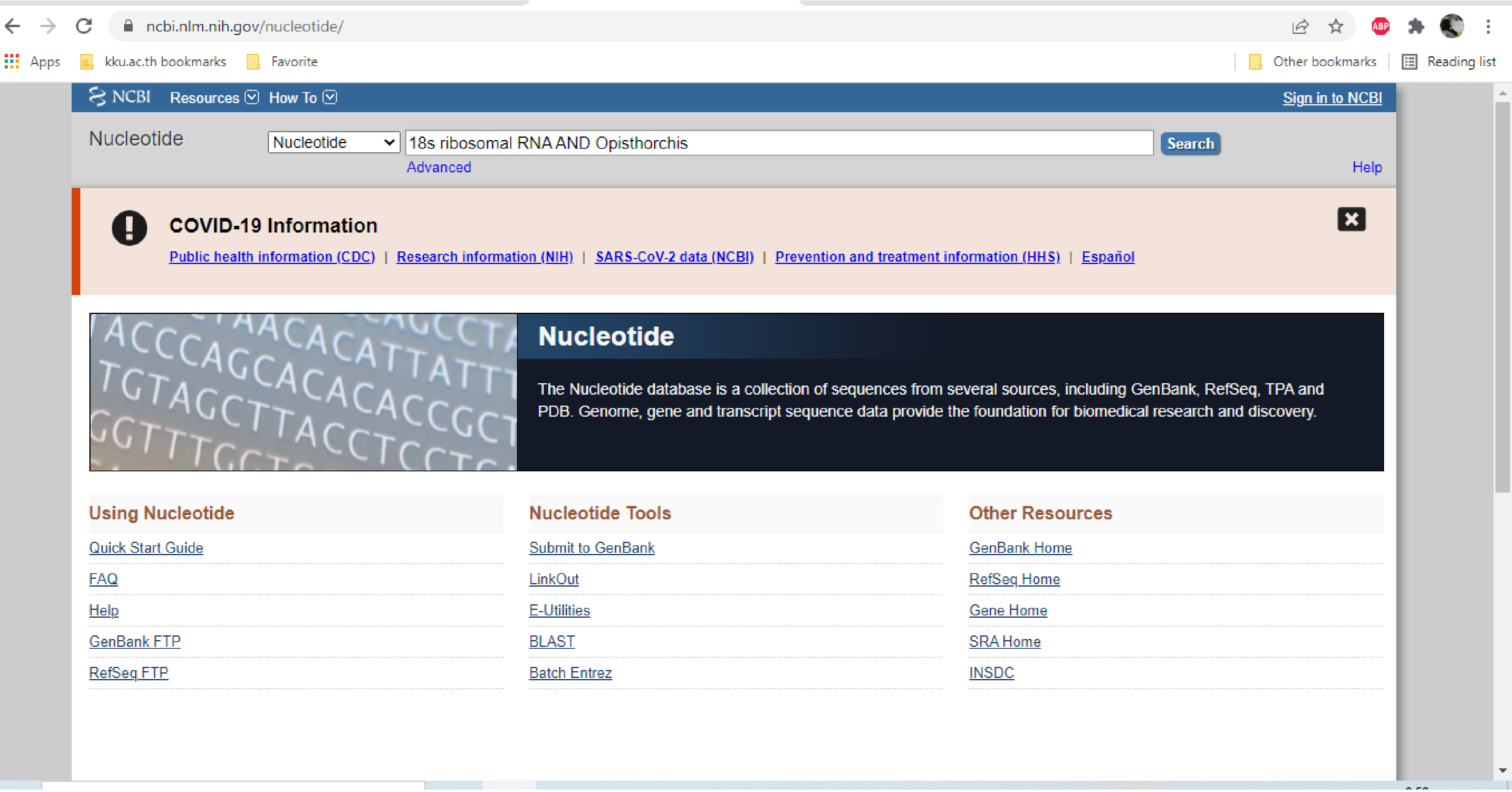
Search and retrieve data from NCBI

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


Drop down list

end structural racism and achieve racial equity in the biomedical research enterprise.

nter for Biotechnology Information advances science and health by providing access to enomic information.



ncbi.nlm.nih.gov/nucleotide/

Nucleotide: 18s ribosomal RNA AND Opisthorchis

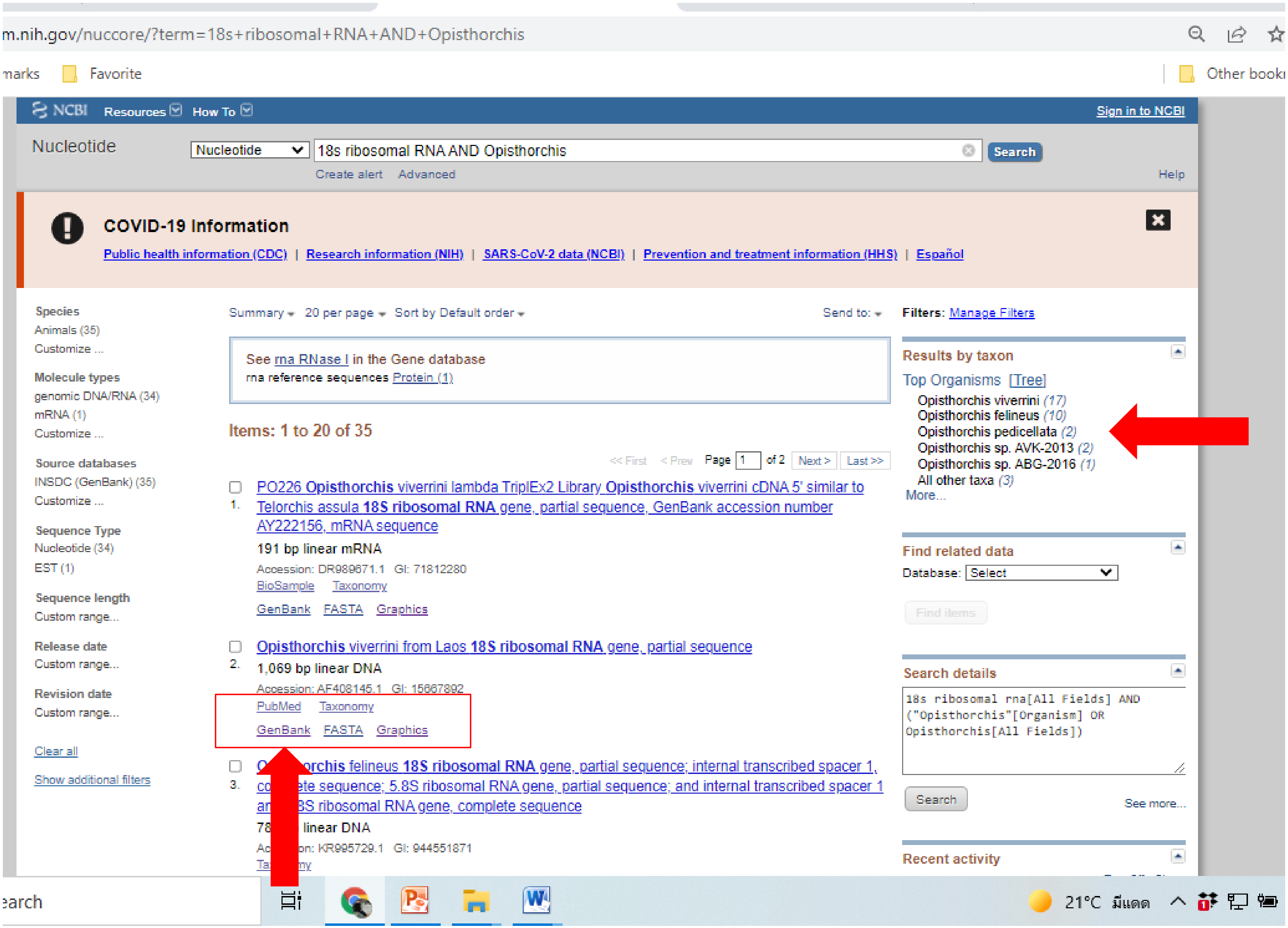
COVID-19 Information

Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS) | Español

Nucleotide

The Nucleotide database is a collection of sequences from several sources, including GenBank, RefSeq, TPA and PDB. Genome, gene and transcript sequence data provide the foundation for biomedical research and discovery.

Using Nucleotide	Nucleotide Tools	Other Resources
Quick Start Guide	Submit to GenBank	GenBank Home
FAQ	LinkOut	RefSeq Home
Help	E-Utilities	Gene Home
GenBank FTP	BLAST	SRA Home
RefSeq FTP	Batch Entrez	INSDC



Search: 18s ribosomal RNA AND Opisthorchis

COVID-19 Information

Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS) | Español

Species
Animals (35)
Customize ...

Molecule types
genomic DNA/RNA (34)
mRNA (1)
Customize ...

Source databases
INSDC (GenBank) (35)
Customize ...

Sequence Type
Nucleotide (34)
EST (1)

Sequence length
Custom range...

Release date
Custom range...

Revision date
Custom range...

Clear all

Show additional filters

See rna RNase I in the Gene database
rna reference sequences Protein (1)

Items: 1 to 20 of 35

☐ 1. PO226 **Opisthorchis** viverrini lambda TriplEx2 Library **Opisthorchis** viverrini cDNA 5' similar to Telorchis assula **18S ribosomal RNA** gene, partial sequence, GenBank accession number AY222156, mRNA sequence
191 bp linear mRNA
Accession: DR989671.1 GI: 71812280
BioSample Taxonomy
GenBank FASTA Graphics

☐ 2. **Opisthorchis** viverrini from Laos **18S ribosomal RNA** gene, partial sequence
1,069 bp linear DNA
Accession: AF408145.1 GI: 15667892
PubMed Taxonomy
GenBank FASTA Graphics

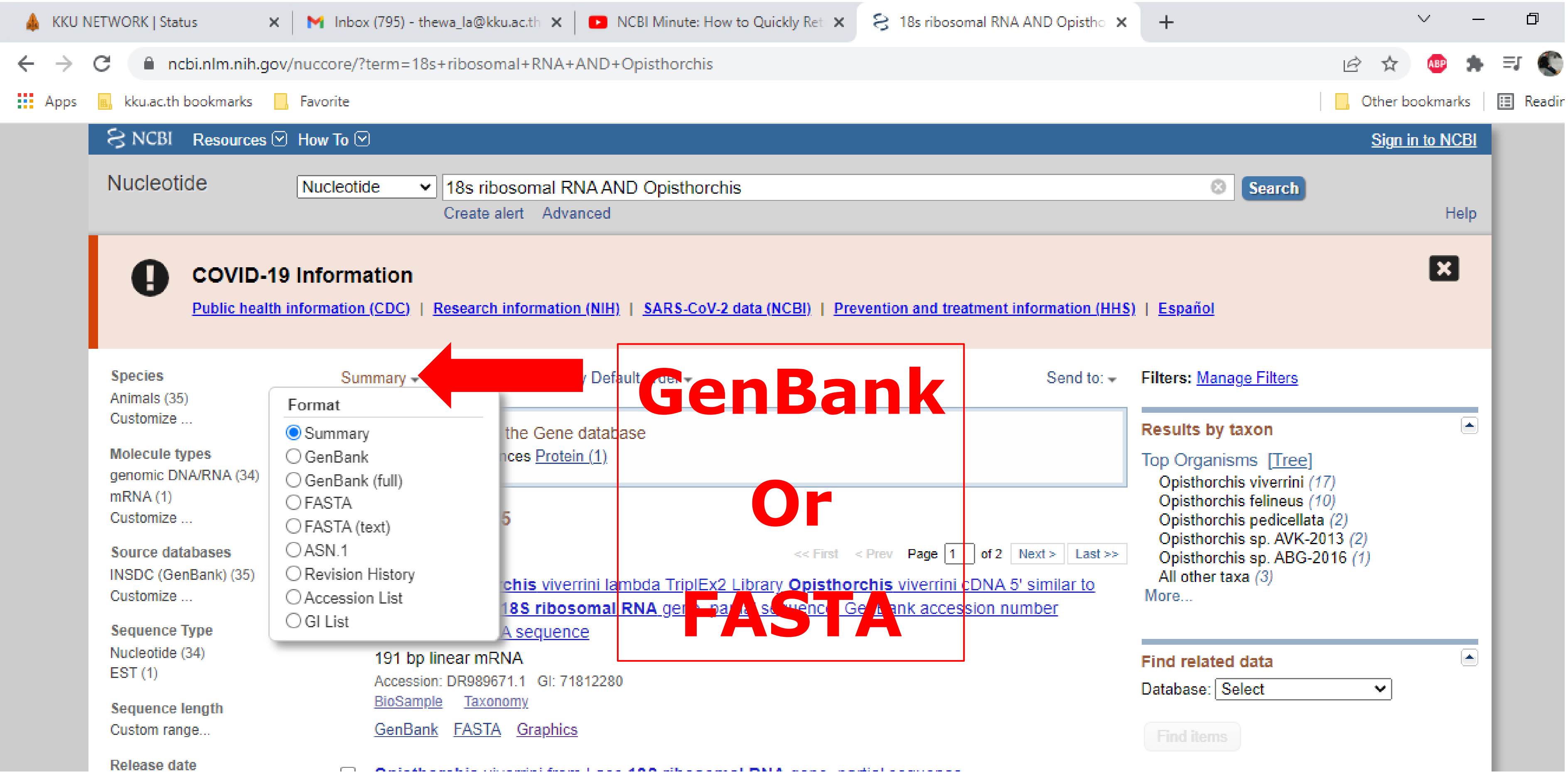
☐ 3. O orchis felineus **18S ribosomal RNA** gene, partial sequence; internal transcribed spacer 1, co ete sequence; 5.8S ribosomal RNA gene, partial sequence; and internal transcribed spacer 1 an 3S ribosomal RNA gene, complete sequence
78 linear DNA
Ac on: KR995729.1 GI: 944551871
Ta my

Results by taxon

Top Organisms [Tree]
Opisthorchis viverrini (17)
Opisthorchis felineus (10)
Opisthorchis pedicellata (2)
Opisthorchis sp. AVK-2013 (2)
Opisthorchis sp. ABG-2016 (1)
All other taxa (3)
More...

Search details

```
18s ribosomal rna[All Fields] AND
("Opisthorchis"[Organism] OR
Opisthorchis[All Fields])
```



KKU NETWORK | Status

Inbox (795) - thewa_la@kku.ac.th

NCBI Minute: How to Quickly Ret

18s ribosomal RNA AND Opistho

ncbi.nlm.nih.gov/nuccore/?term=18s+ribosomal+RNA+AND+Opisthorchis

Apps | kku.ac.th bookmarks | Favorite | Other bookmarks | Readir

NCBI Resources How To — Sign in to NCBI

Nucleotide

Nucleotide | 18s ribosomal RNA AND Opisthorchis | Search

Create alert Advanced — Help

COVID-19 Information

Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS) | Español

Species
Animals (35)
Customize ...

Molecule types
genomic DNA/RNA (34)
mRNA (1)
Customize ...

Source databases
INSDC (GenBank) (35)
Customize ...

Sequence Type
Nucleotide (34)
EST (1)

Sequence length
Custom range...

Release date

Summary ▾ — Default order ▾ — Send to: ▾

Format
- ● Summary
- ○ GenBank
- ○ GenBank (full)
- ○ FASTA
- ○ FASTA (text)
- ○ ASN.1
- ○ Revision History
- ○ Accession List
- ○ GI List

the Gene database
nces Protein (1)

<< First < Prev Page 1 of 2 Next > Last >>

chis viverrini lambda TriplEx2 Library Opisthorchis viverrini cDNA 5' similar to 18S ribosomal RNA gene partial sequence GenBank accession number A sequence

191 bp linear mRNA

Accession: DR989671.1 GI: 71812280

BioSample Taxonomy

GenBank FASTA Graphics

Filters: Manage Filters

Results by taxon

Top Organisms [Tree]
- Opisthorchis viverrini (17)
- Opisthorchis felineus (10)
- Opisthorchis pedicellata (2)
- Opisthorchis sp. AVK-2013 (2)
- Opisthorchis sp. ABG-2016 (1)
- All other taxa (3)

More...

Find related data

Database: Select

Find items

GenBank

Or

FASTA

Exercise 1

1. Practice- search and retrieve nucleotide and/or protein sequence from NCBI
2. Practice-

Bioinformatics for homology and structure prediction

- Blast homology search

Basic using of NCBI blast

What is Blast?

-The Basic Local Alignment Search Tool (BLAST) is a sequence similarity search tool (probably is the most widely used)

- Finds regions of similarity between 2 sequences (both protein and DNA).

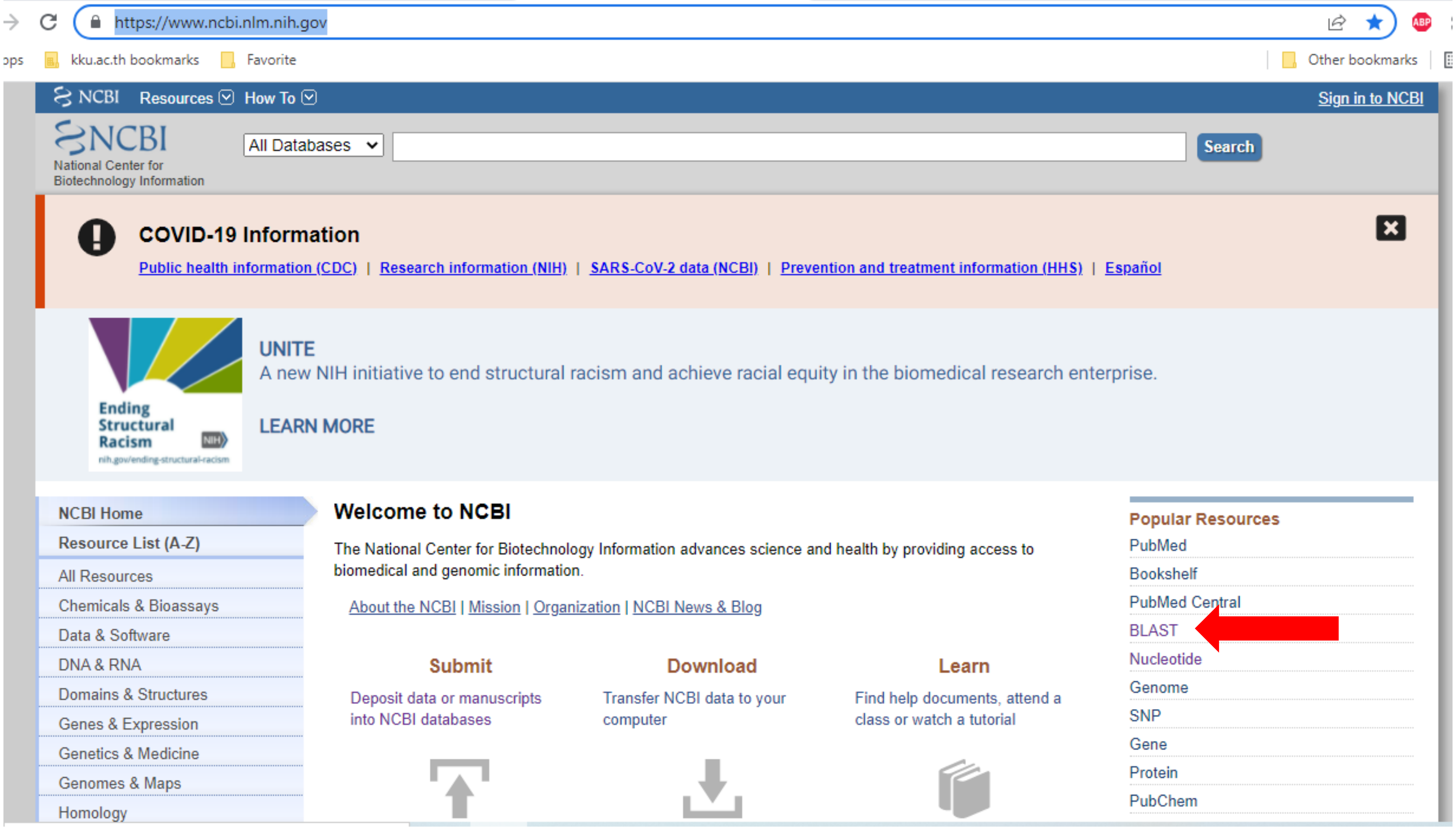
- The program compares DNA or protein sequences to sequence databases and calculates the statistical significance of matches.

What BLAST tells you?

Your query sequence ;

1. What it is related to?
    - Homology
    - Function
2. Is it already in the database?
    - Find the matching sequences in the database
    - Origin of organism
3. Where is it located?
    - In a genome
    - Annotation problem such as frame shifts

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



COVID-19 Information

Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS) | Español

UNITE

A new NIH initiative to end structural racism and achieve racial equity in the biomedical research enterprise.

LEARN MORE

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 | NCBI News & Blog

Submit

Deposit data or manuscripts into NCBI databases

Download

Transfer NCBI data to your computer

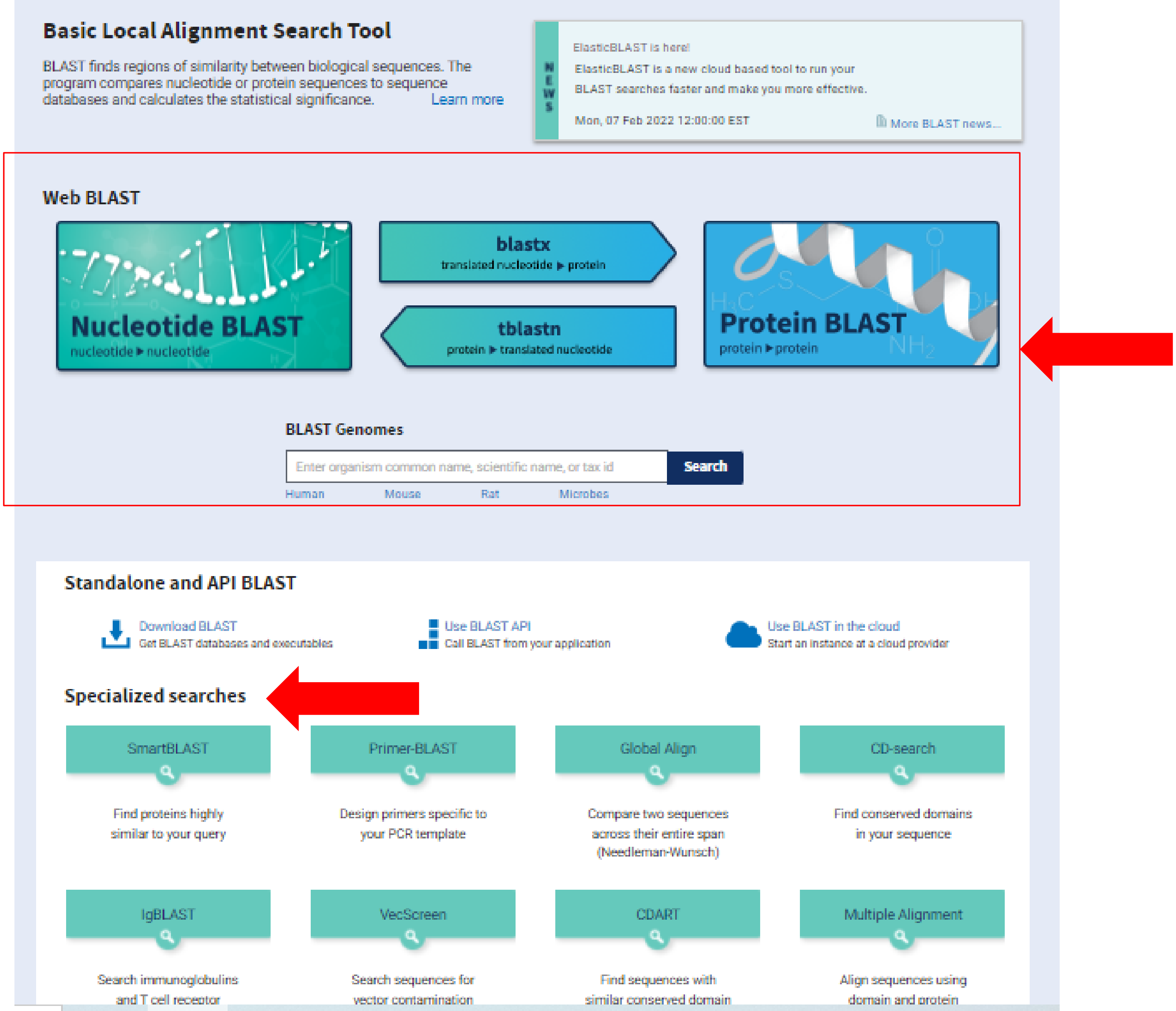
Learn

Find help documents, attend a class or watch a tutorial

- >AF408145.1 Opisthorchis viverrini from Laos 18S ribosomal RNA gene, partial sequence

  GGTGGATTTATTAGAACAGAACCAACCGGCGGTGACTTCGGTTGCTGTCGTTGCATTCTGTGATGACTCT
  GGATAACTTTACTGATCGCAGTCGGCCTTGTGTCGGCGACGGATCTTTCAAATGTCTGCCCTATCAATTT
  TCGATGGTAGGTGACCTGCCTACCATGGTGATAACGGGTAACGGGGAATCAGGGTTCGATTCCGGAGAGG
  GAGCCTGAGAAACGGCTACCACTTCCAAGGAAGGCAGCAGGCGCGCAAATTACCCAATCCCGGCACGGAG
  GTAGTGACGAAAAATACGGATACGGGACTCATTAGAGGCTCCGTAATTCGAATGAGTACAATTTAAATCC
  TTTAACGAGGATCAACTGGAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAACTCCAGCTCCAGAAGCGTAT
  ATTAAAGTTGTTGCAGTTAAAAAGCTCGTAGTTGGATCTGGGTCGCATGGCTACATGCCGTTGCTCGTAT
  TCCTGGCCTGGTTCACACCGGGACGGGTTTGTGAGTCGGTGTCGTGGTTGTGCAGCCTTTCTGCCGTGTC
  TGTTTCGACAGGTGTTAGCATGATTGGTGGGTTCGTCCTGCTGATCTGTTGACATGCTTCCAGGTGCCCT
  TAACCGGGTGTCGGGGGCGGACGGCACGTTTACTTTGAACAAATTTGAGTGCTCAAAGCAGGCCCGTGTG
  CCTGAAAATTCTTGCATGGAATAATGGAATAGGACTTCGGTTCTATTTTGTTGGTTTTCGGATCCGAAGT
  AATGGTTAAAAGGGACAGACGGGGGCATTTGTATGGCGGTGTTAGAGGTGAAATTCTTGGATCACCGCCA
  GACAAACTACAGCGAAAGCATTTGCCAAGGATGTTTTCATTGATCTGGAGCGAAAGTCAGAGGTTCGAAG
  ACGATCAGATACCGTCCTAGTTCTGACCATAAACGATGCCAACTGACGATCCGTGGTGGCGCGATTATTG
  GCCCCGCGGGCAGCCCCCGGGAAACCTTTAAGTCTTTGGGCTCCGGGGGAAGTATGGTTGCAAAGCTGAA
  ACTTAAATGAATTGACGGA

ATGGTCTCGCTAAGTTGTGGATATCGTTGTCTTCAAATTATACTTGTCATCCTTAATACTTTTGTCATCG
CATGTGGTGTTGGCCTCATAGTTGTCGGGAGTTTGGCCGAAGTTTCTTTGAAGACTTTCGGAGAATCTAA
CGAAACCAGCATCCAAATTATCGTTATTTTCATCATTTGTCTCGGATGTCTCACATTTTTAATAGGATTT
TTGGGGTTCTGTGGAGCATGCCTGAAAAATGTCTGTATGCTCATTACCTATGCTATACTGCTTGGAGTGA
CGGCAGTTGCACAAGTCGTATGTGGAATCATTGGCCTTGTTCTCCGCGATAAGATCCCCGGATTGGTAAA
TCACAATCTGGAGGTGCTGTATACAGAGTATAGTGCCAACAAAGATGTTCAAAAGTTGATTAACGTGATC
CAGAGTGAGCTGAAATGCTGTGGAGCGACTGGAACGTGGGCGAACCCGGGGTCTGAACCAGAATCTTGCC
GCAGCCCAGAAGGTGTCGTTTACAAGGATGGCTGTGTTCCGAAAGTCGAAGAGTTCATCCAGGAAAATAT
GGTTGCAATCGGCGTGTGCGTCTTCATTTTTGCACTAATTCAACTCATCTGTATGACGTTTGCCATCTGT
GTTGTGCAAGCTTTACGAAAAGGAGAAGGCGAGACGGTATAG



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

ElasticBLAST is here!

ElasticBLAST is a new cloud based tool to run your BLAST searches faster and make you more effective.

Mon, 07 Feb 2022 12:00:00 EST

More BLAST news...

Web BLAST

BLAST Genomes

Enter organism common name, scientific name, or tax id

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

Specialized searches

Find proteins highly similar to your query

Design primers specific to your PCR template

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

Find conserved domains in your sequence

Search immunoglobulins and T cell receptor

Search sequences for vector contamination

Find sequences with similar conserved domain

Align sequences using domain and protein

Functions of BLAST Search Pages

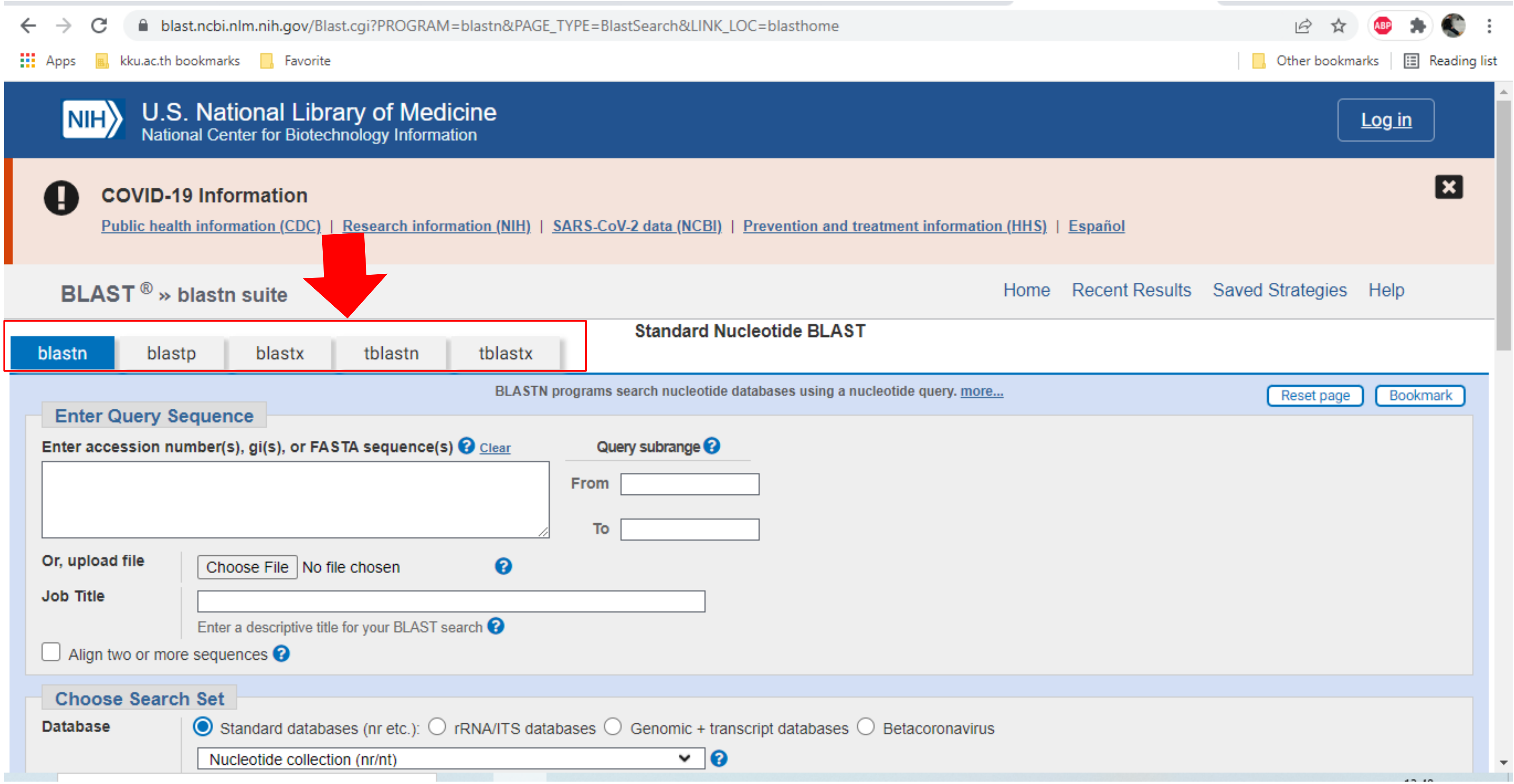
1. Nucleotide blast- **blastn suite**
   - megablast: for sequence identification, intra-species comparison
   - discontiguous megablast: for cross-species comparison, searching with coding sequences
   - blastn: for searching with shorter queries, cross-species comparison

Protein blast

- blastp: general sequence identification and similarity searches
- blastx: for identifying potential protein products encoded by a nucleotide query
- tblastn: for identifying database sequences encoding proteins similar to the query
- tblastx: for identifying nucleotide sequences similar to the query based on their coding potential

Table 1. Key features of the BLAST search pages in the "Basic BLAST" category

Search page	**Query & database combination**	**Alignment type**	**Programs & functions** (default program in bold)
nucleotide blast	nucleotide vs nucleotide	nucleotide vs nucleotide	**megablast**: for sequence identification, intra-species comparison discontiguous megablast: for cross-species comparison, searching with coding sequences blastn: for searching with shorter queries, cross-species comparison
protein blast	Protein vs protein	protein vs protein	**blastp**: general sequence identification and similarity searches DELTA-BLAST [2] : protein similarity search with higher sensitivity than blastp PSI-BLAST: iterative search for position-specific score matrix (PSSM) construction or identification of distant relatives for a protein family PHI-BLAST: protein alignment with input pattern as anchor/constraint
blastx	nucleotide (translated) vs protein	protein vs protein	blastx: for identifying potential protein products encoded by a nucleotide query
tblastn	protein vs nucleotide (translated)	protein vs protein	tblastn: for identifying database sequences encoding proteins similar to the query
tblastx	nucleotide (translated) vs nucleotide (translated)	protein vs protein	tblastx: for identifying nucleotide sequences similar to the query based on their coding potential



U.S. National Library of Medicine
National Center for Biotechnology Information

COVID-19 Information

Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS) | Español

BLAST ® » blastn suite

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)

Query subrange
From
To

Or, upload file: No file chosen

Job Title

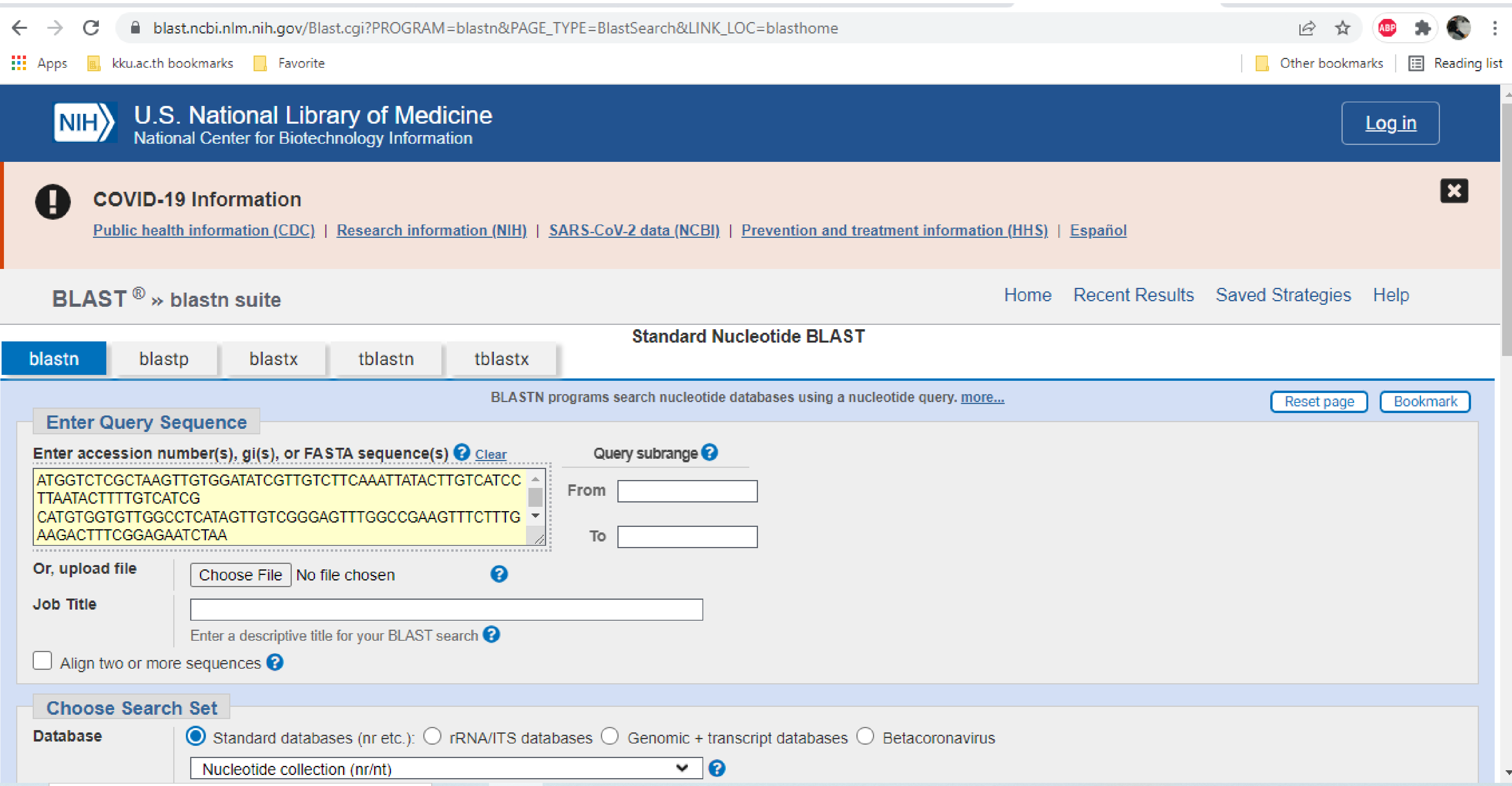
Enter a descriptive title for your BLAST search

☐ Align two or more sequences

Choose Search Set

Database: ● Standard databases (nr etc.): ○ rRNA/ITS databases ○ Genomic + transcript databases ○ Betacoronavirus

Nucleotide collection (nr/nt)



COVID-19 Information

Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS) | Español

BLAST ® » blastn suite

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)

```
ATGGTCTCGCTAAGTTGTGGATATCGTTGTCTTCAAATTATACTTGTCATCC
TTAATACTTTTGTCATCG
CATGTGGTGTTGGCCTCATAGTTGTCGGGAGTTTGGCCGAAGTTTCTTTG
AAGACTTTCGGAGAATCTAA
```

Query subrange

From

To

Or, upload file

Job Title

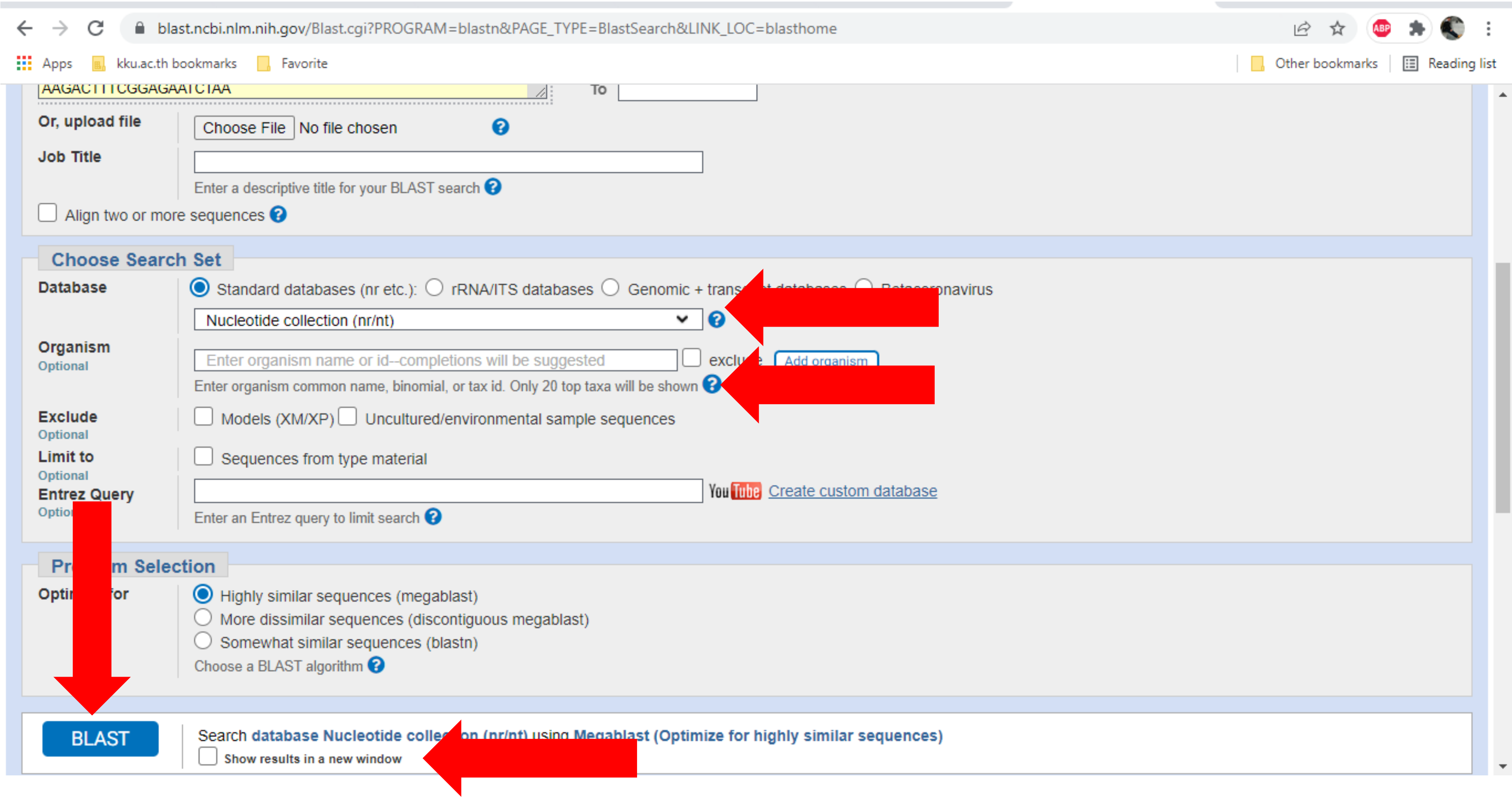
Enter a descriptive title for your BLAST search

☐ Align two or more sequences

Choose Search Set

Database ● Standard databases (nr etc.): ○ rRNA/ITS databases ○ Genomic + transcript databases ○ Betacoronavirus

Nucleotide collection (nr/nt)



AAGACTTTCGGAGAATCTAA

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 + trans ○ B navirus

Nucleotide collection (nr/nt)

Organism
Optional

Enter organism name or id--completions will be suggested ☐ exclu 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
Optio

You Tube Create custom database

Enter an Entrez query to limit search

Pr m Selection

Opti for

● Highly similar sequences (megablast)

○ More dissimilar sequences (discontiguous megablast)

○ Somewhat similar sequences (blastn)

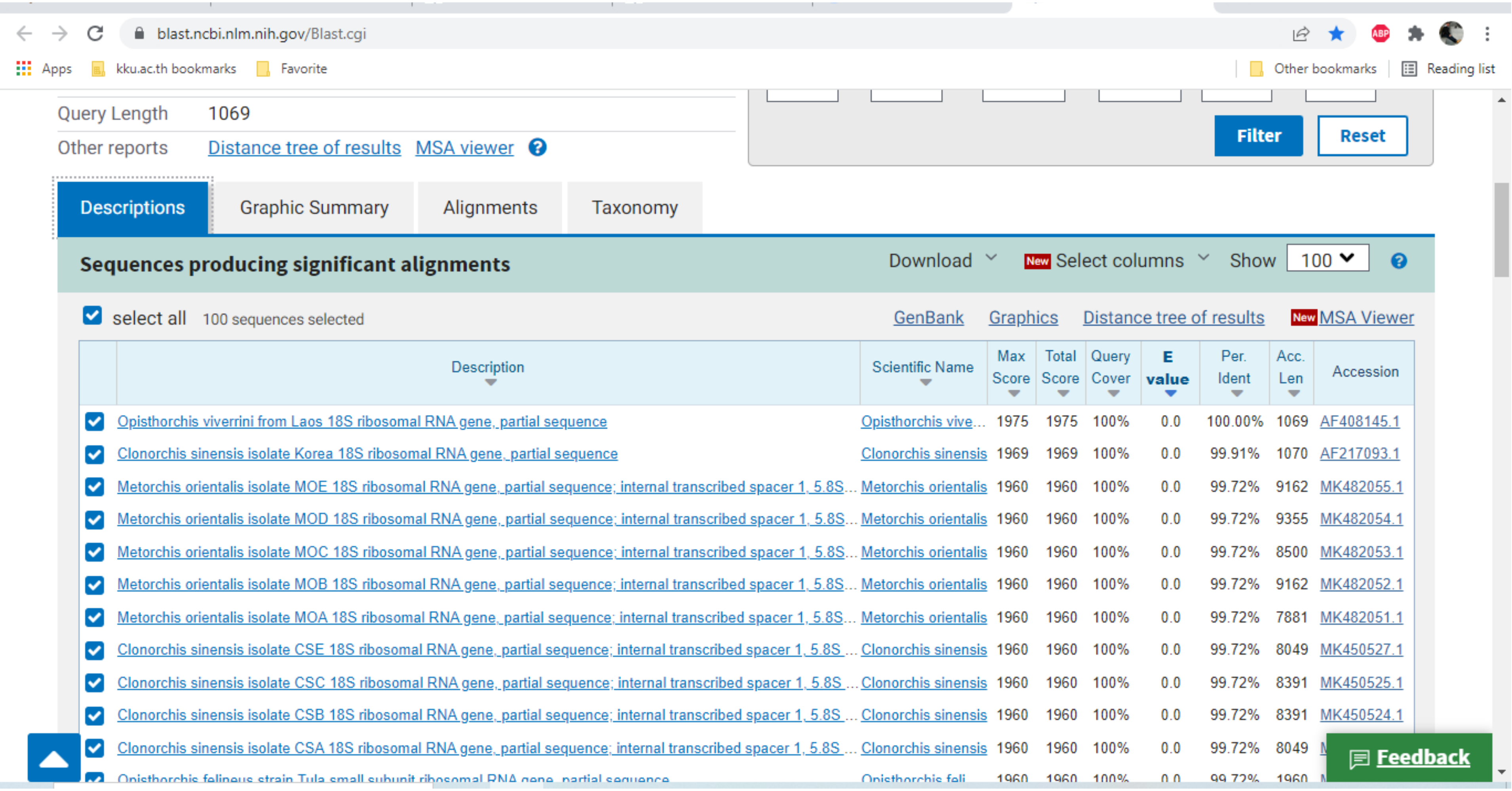
Choose a BLAST algorithm

BLAST

Search database Nucleotide colle on (nr/nt) using Megablast (Optimize for highly similar sequences)

☐ Show results in a new window

Result from BLAST search



Query Length 1069

Other reports Distance tree of results MSA viewer

Sequences producing significant alignments

☑ select all 100 sequences selected

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	Opisthorchis viverrini from Laos 18S ribosomal RNA gene, partial sequence	Opisthorchis vive...	1975	1975	100%	0.0	100.00%	1069	AF408145.1
☑	Clonorchis sinensis isolate Korea 18S ribosomal RNA gene, partial sequence	Clonorchis sinensis	1969	1969	100%	0.0	99.91%	1070	AF217093.1
☑	Metorchis orientalis isolate MOE 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S...	Metorchis orientalis	1960	1960	100%	0.0	99.72%	9162	MK482055.1
☑	Metorchis orientalis isolate MOD 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S...	Metorchis orientalis	1960	1960	100%	0.0	99.72%	9355	MK482054.1
☑	Metorchis orientalis isolate MOC 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S...	Metorchis orientalis	1960	1960	100%	0.0	99.72%	8500	MK482053.1
☑	Metorchis orientalis isolate MOB 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S...	Metorchis orientalis	1960	1960	100%	0.0	99.72%	9162	MK482052.1
☑	Metorchis orientalis isolate MOA 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S...	Metorchis orientalis	1960	1960	100%	0.0	99.72%	7881	MK482051.1
☑	Clonorchis sinensis isolate CSE 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ...	Clonorchis sinensis	1960	1960	100%	0.0	99.72%	8049	MK450527.1
☑	Clonorchis sinensis isolate CSC 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ...	Clonorchis sinensis	1960	1960	100%	0.0	99.72%	8391	MK450525.1
☑	Clonorchis sinensis isolate CSB 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ...	Clonorchis sinensis	1960	1960	100%	0.0	99.72%	8391	MK450524.1
☑	Clonorchis sinensis isolate CSA 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ...	Clonorchis sinensis	1960	1960	100%	0.0	99.72%	8049	M
☑	Opisthorchis felineus strain Tula small subunit ribosomal RNA gene, partial sequence	Opisthorchis feli...	1960	1960	100%	0.0	99.72%	1960	M

Practice

- Run BLAST search with your interested sequences.

Basic tools

BioEdit Software

- **Sequence editing**
  - View and edit DNA /protein sequence file
- **Sequence manipulation**
  - DNA or nucleic acid sequence
    - Base composition- %CG
    - Complement or reverse complement
    - Restriction enzyme map
    - Translation to protein (amino acid sequence)
  - Protein
    - Amino acid composition- prediction of MW
- **Sequence alignment**

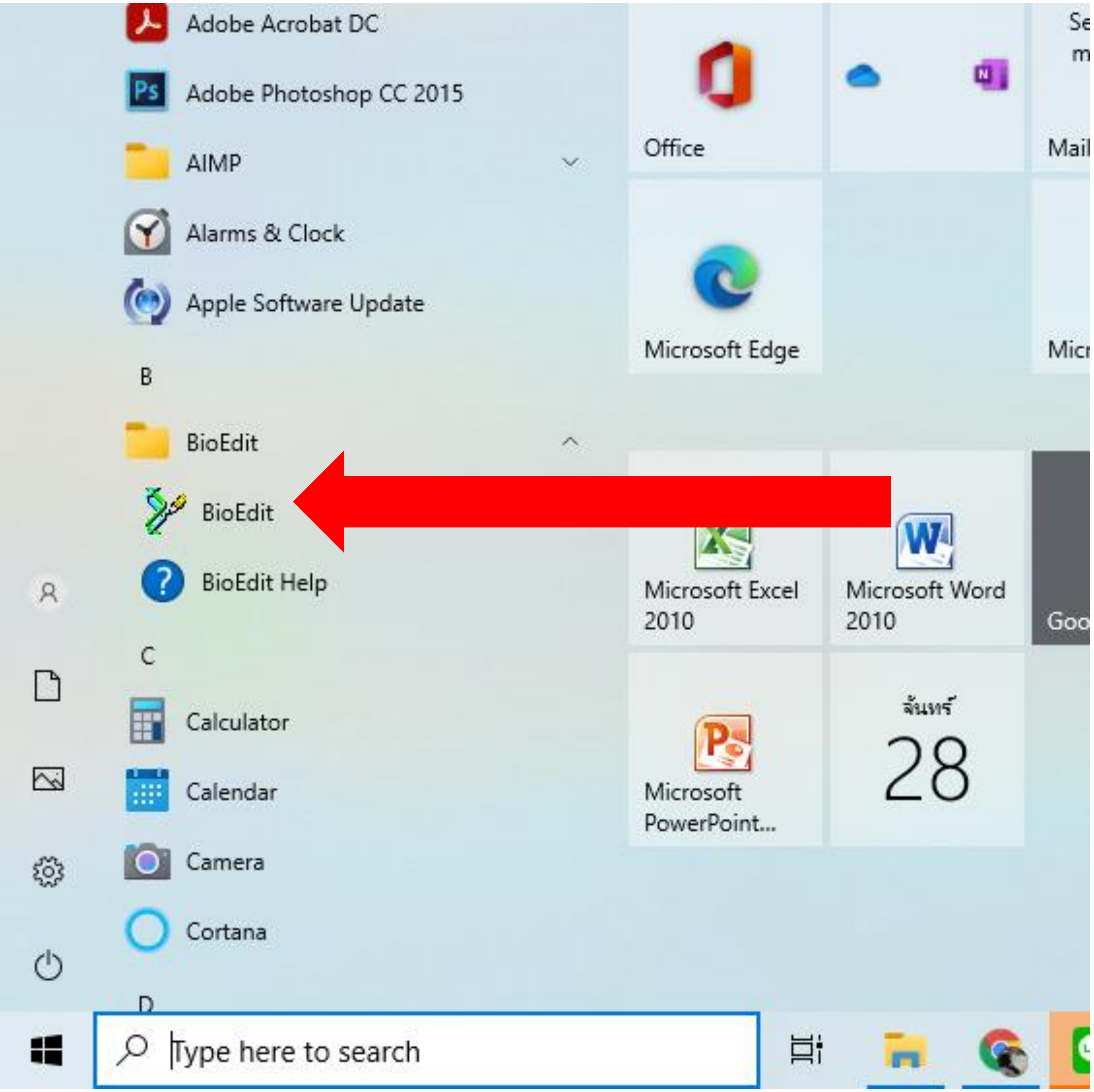
Practice

1. Install BioEdit software

	- Google search- BioEdit software download.

	- Download BioEdit (compatible version for your OS and computer).

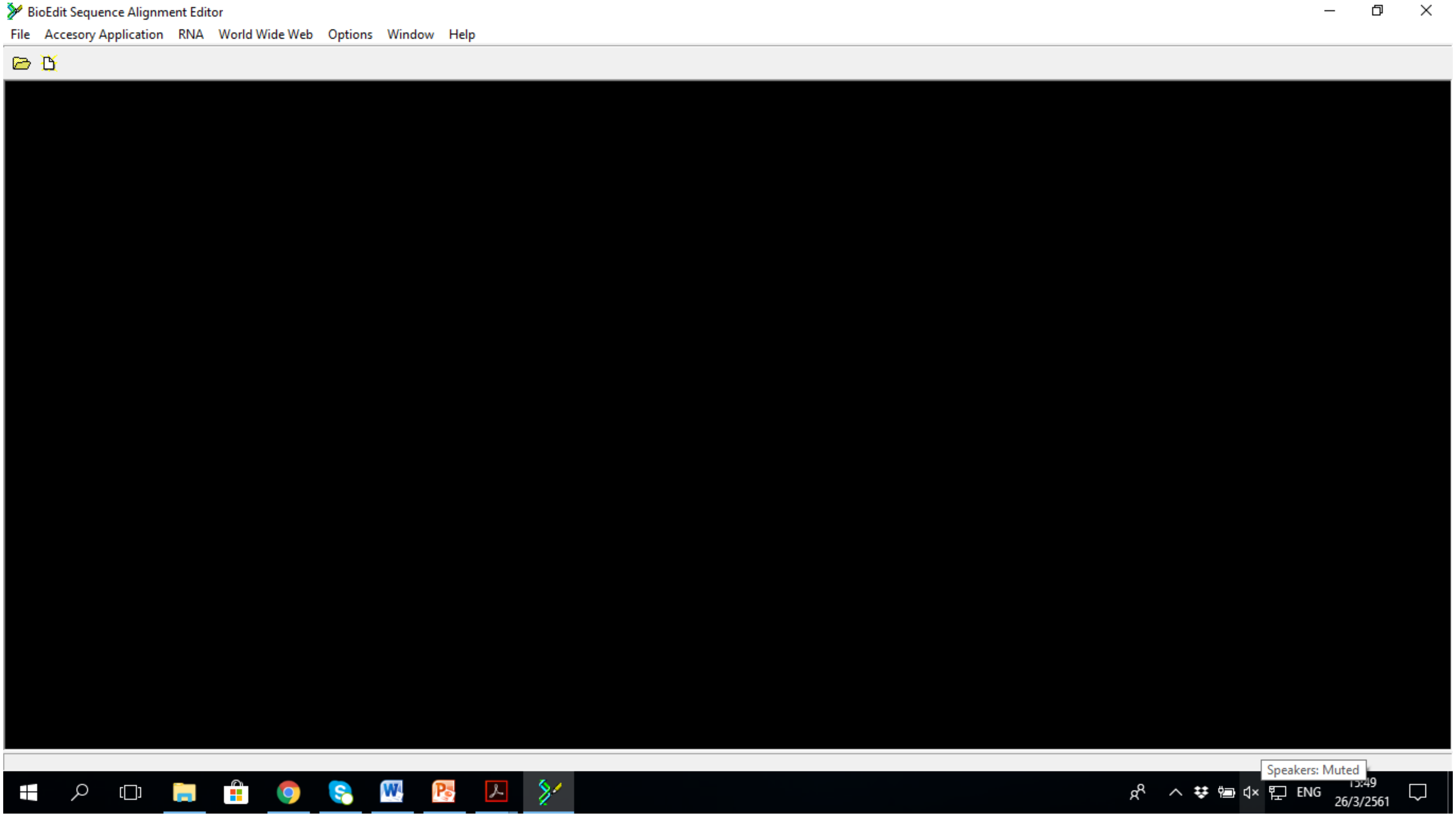
	- Unzip and install program.

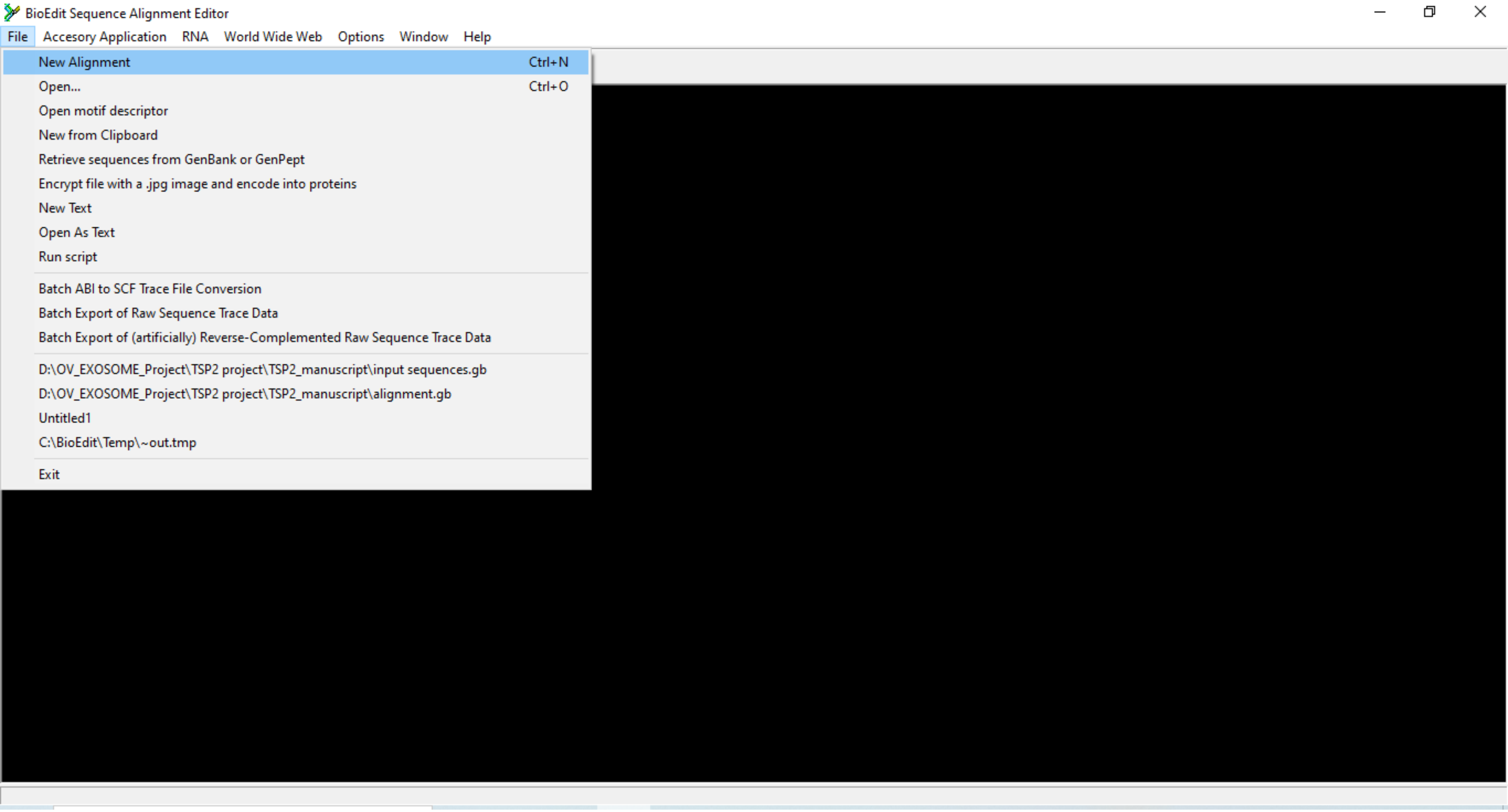

Adobe Acrobat DC
Adobe Photoshop CC 2015
AIMP
Alarms & Clock
Apple Software Update
B
BioEdit
BioEdit
BioEdit Help
C
Calculator
Calendar
Camera
Cortana
D
Type here to search
Office
Se
m
Mail
Microsoft Edge
Micr
Microsoft Excel 2010
Microsoft Word 2010
Goo
Microsoft PowerPoint...
จันทร์
28

BioEdit Software

- **Sequence editing**
  - **View and edit DNA /protein sequence file**

Open program



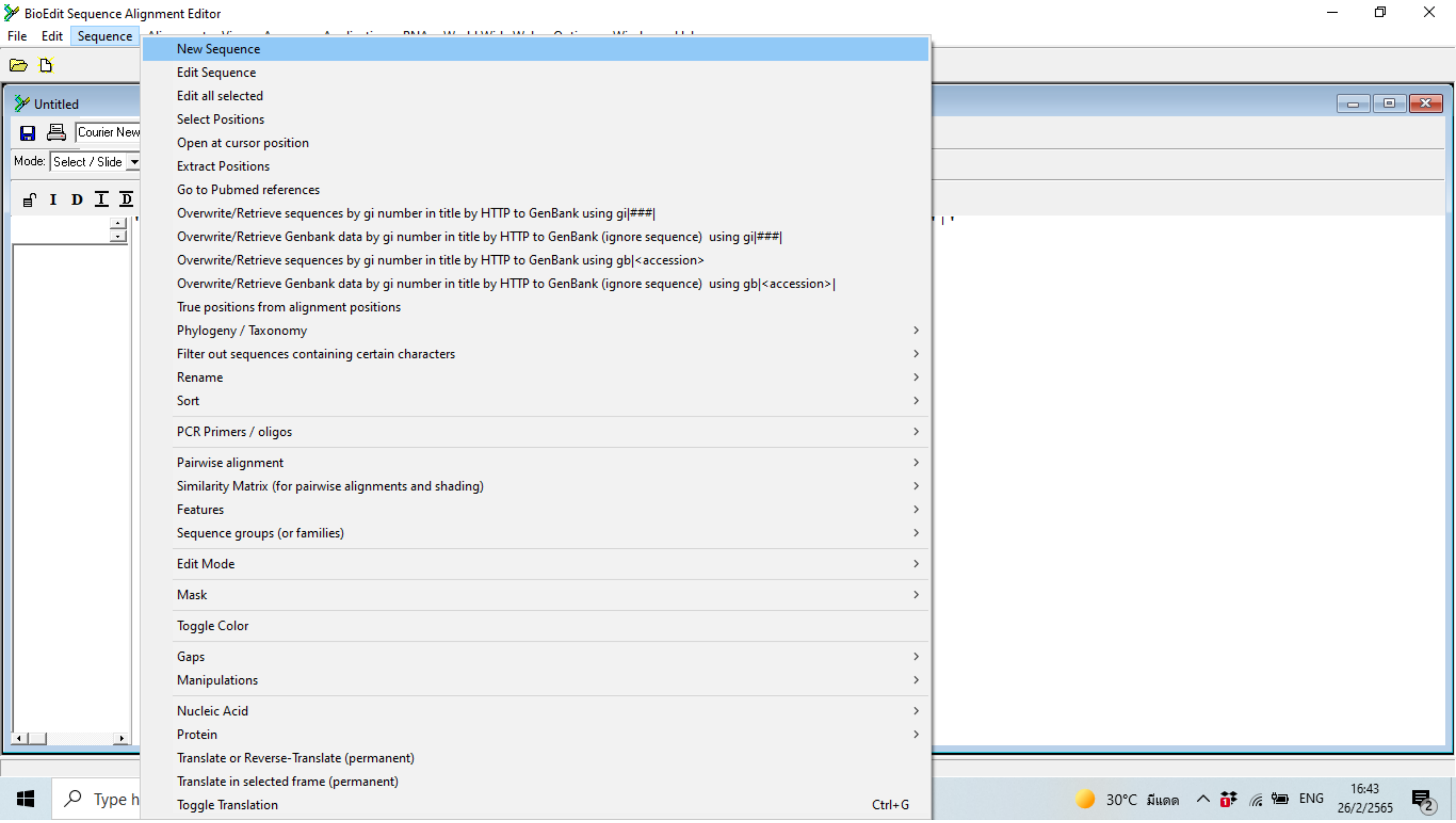

BioEdit Sequence Alignment Editor
File
Accesory Application
RNA
World Wide Web
Options
Window
Help
New Alignment Ctrl+N
Open... Ctrl+O
Open motif descriptor
New from Clipboard
Retrieve sequences from GenBank or GenPept
Encrypt file with a .jpg image and encode into proteins
New Text
Open As Text
Run script
Batch ABI to SCF Trace File Conversion
Batch Export of Raw Sequence Trace Data
Batch Export of (artificially) Reverse-Complemented Raw Sequence Trace Data
D:\OV_EXOSOME_Project\TSP2 project\TSP2_manuscript\input sequences.gb
D:\OV_EXOSOME_Project\TSP2 project\TSP2_manuscript\alignment.gb
Untitled1
C:\BioEdit\Temp\~out.tmp
Exit

Sequence editing

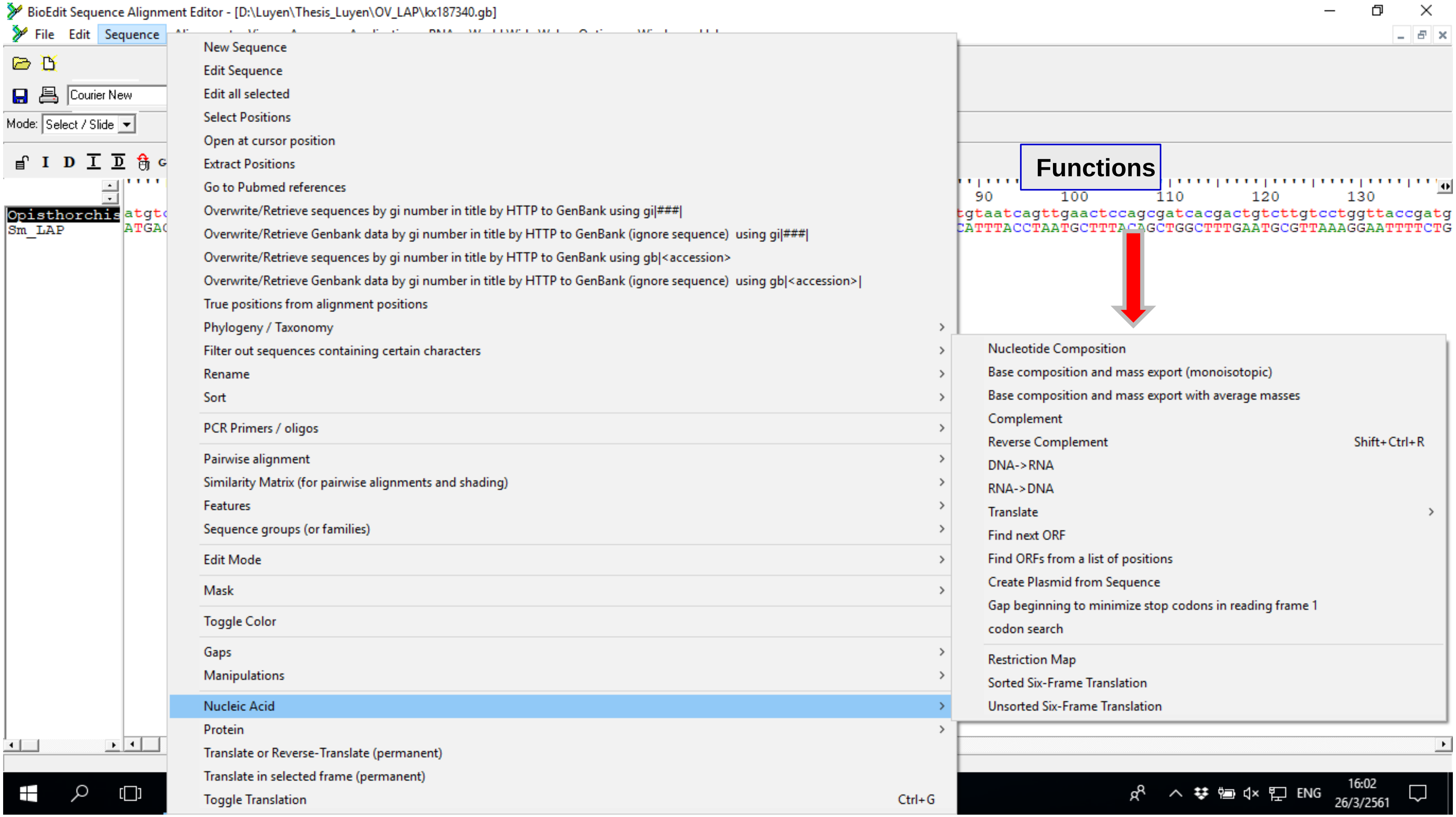
ผลไม่ชัดเจน

BioEdit Software

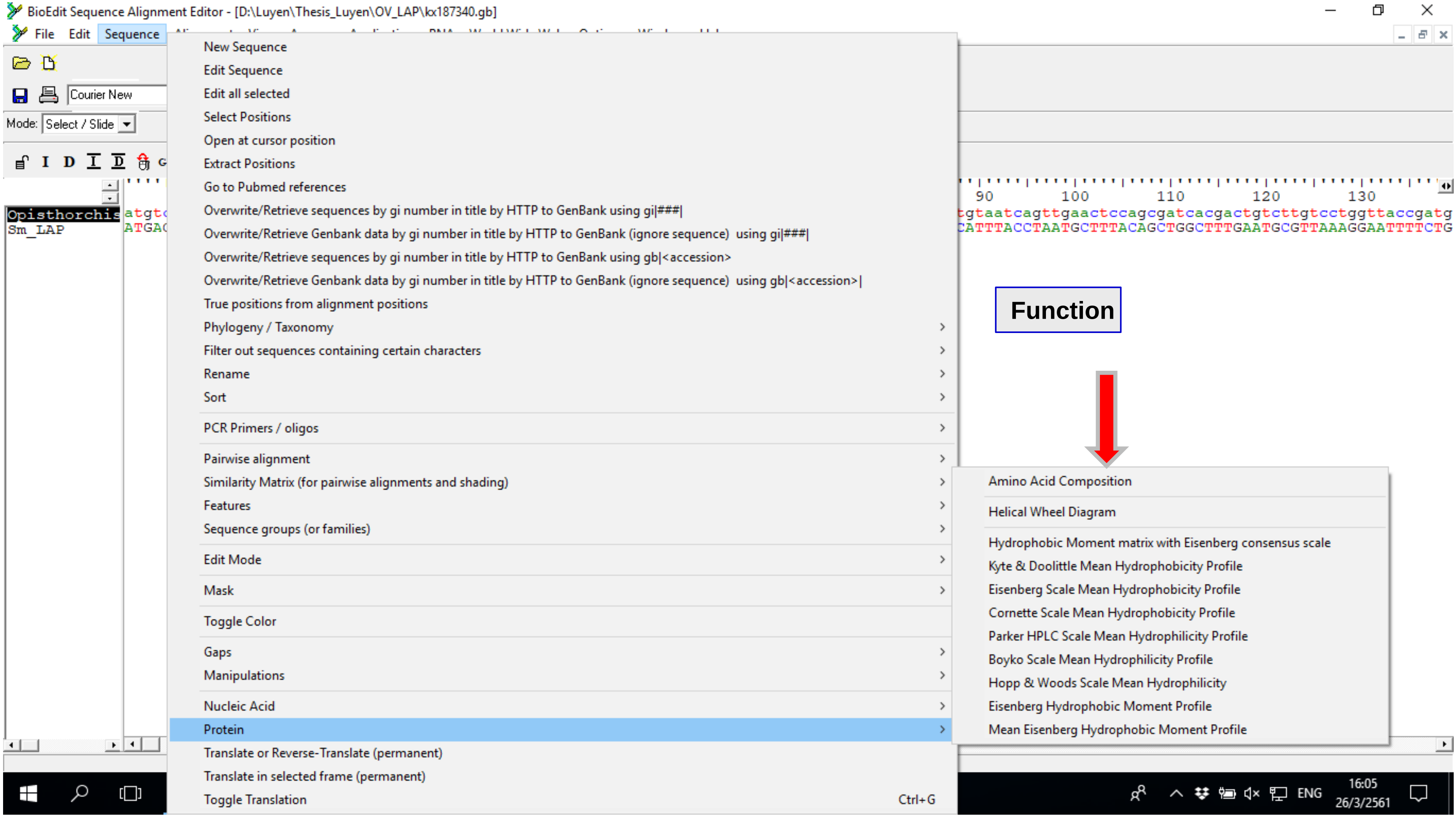
- Sequence manipulation
  - DNA or nucleic acid sequence
    - Base composition- %CG
    - Complement or reverse complement
    - Restriction enzyme map
    - Translation to protein (amino acid sequence)
  - Protein
    - Amino acid composition- prediction of MW


BioEdit Sequence Alignment Editor
File
Edit
Sequence
Untitled
Courier New
Mode:
Select / Slide
New Sequence
Edit Sequence
Edit all selected
Select Positions
Open at cursor position
Extract Positions
Go to Pubmed references
Overwrite/Retrieve sequences by gi number in title by HTTP to GenBank using gi|###|
Overwrite/Retrieve Genbank data by gi number in title by HTTP to GenBank (ignore sequence) using gi|###|
Overwrite/Retrieve sequences by gi number in title by HTTP to GenBank using gb|<accession>
Overwrite/Retrieve Genbank data by gi number in title by HTTP to GenBank (ignore sequence) using gb|<accession>|
True positions from alignment positions
Phylogeny / Taxonomy
Filter out sequences containing certain characters
Rename
Sort
PCR Primers / oligos
Pairwise alignment
Similarity Matrix (for pairwise alignments and shading)
Features
Sequence groups (or families)
Edit Mode
Mask
Toggle Color
Gaps
Manipulations
Nucleic Acid
Protein
Translate or Reverse-Translate (permanent)
Translate in selected frame (permanent)
Toggle Translation
Ctrl+G
Type h
30°C มีแดด
ENG
16:43
26/2/2565

Sequence manipulation (DNA)



Sequence manipulation (protein)



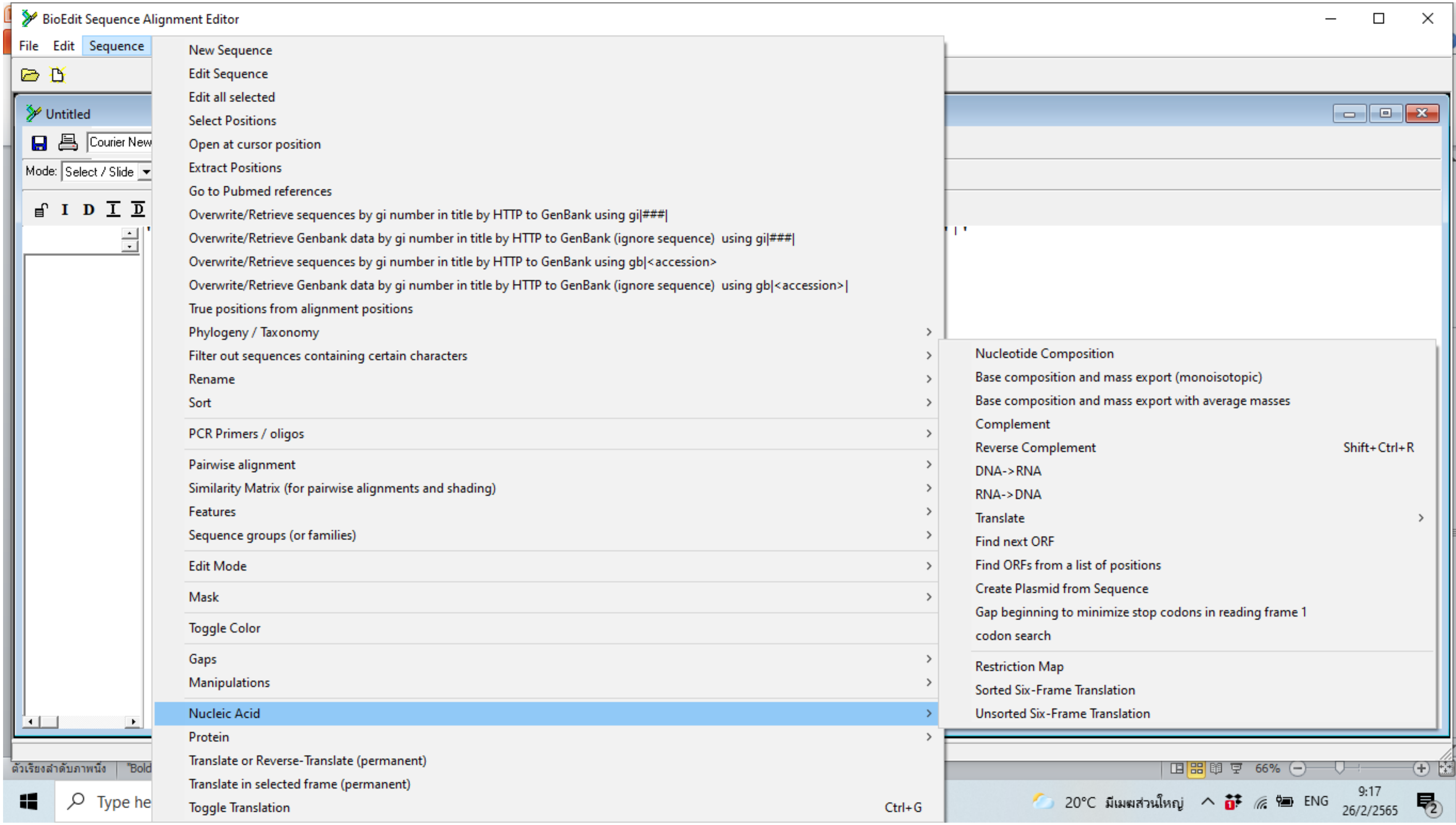
Function

Sequences manipulation

DNA sequence

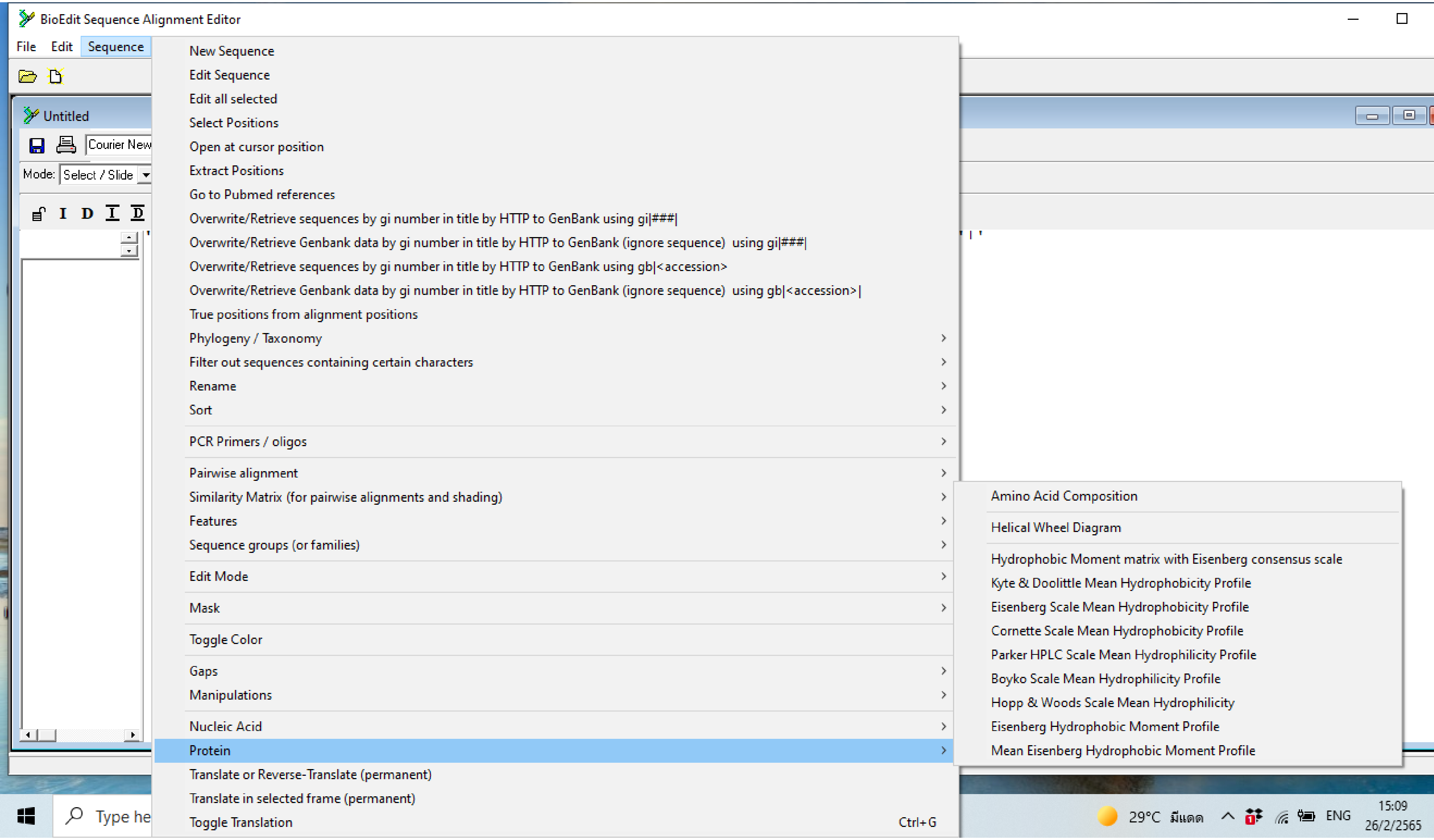
- Translation: translate DNA to protein (amino acid) sequence
  - BioEdit-sequence-nucleic acid-translate
  - Web http://bio.lundberg.gu.se/edu/translat.html

- Reverse complement
- Complement
- Contig assembly


BioEdit Sequence Alignment Editor
File
Edit
Sequence
Untitled
Courier New
Mode:
Select / Slide
New Sequence
Edit Sequence
Edit all selected
Select Positions
Open at cursor position
Extract Positions
Go to Pubmed references
Overwrite/Retrieve sequences by gi number in title by HTTP to GenBank using gi|###|
Overwrite/Retrieve Genbank data by gi number in title by HTTP to GenBank (ignore sequence) using gi|###|
Overwrite/Retrieve sequences by gi number in title by HTTP to GenBank using gb|<accession>
Overwrite/Retrieve Genbank data by gi number in title by HTTP to GenBank (ignore sequence) using gb|<accession>|
True positions from alignment positions
Phylogeny / Taxonomy
Filter out sequences containing certain characters
Rename
Sort
PCR Primers / oligos
Pairwise alignment
Similarity Matrix (for pairwise alignments and shading)
Features
Sequence groups (or families)
Edit Mode
Mask
Toggle Color
Gaps
Manipulations
Nucleic Acid
Protein
Translate or Reverse-Translate (permanent)
Translate in selected frame (permanent)
Toggle Translation Ctrl+G
Nucleotide Composition
Base composition and mass export (monoisotopic)
Base composition and mass export with average masses
Complement
Reverse Complement Shift+Ctrl+R
DNA->RNA
RNA->DNA
Translate
Find next ORF
Find ORFs from a list of positions
Create Plasmid from Sequence
Gap beginning to minimize stop codons in reading frame 1
codon search
Restriction Map
Sorted Six-Frame Translation
Unsorted Six-Frame Translation
Type he
20°C มีเมฆส่วนใหญ่
ENG
9:17
26/2/2565
66%

Sequence manipulation (cont.)

- Protein sequence

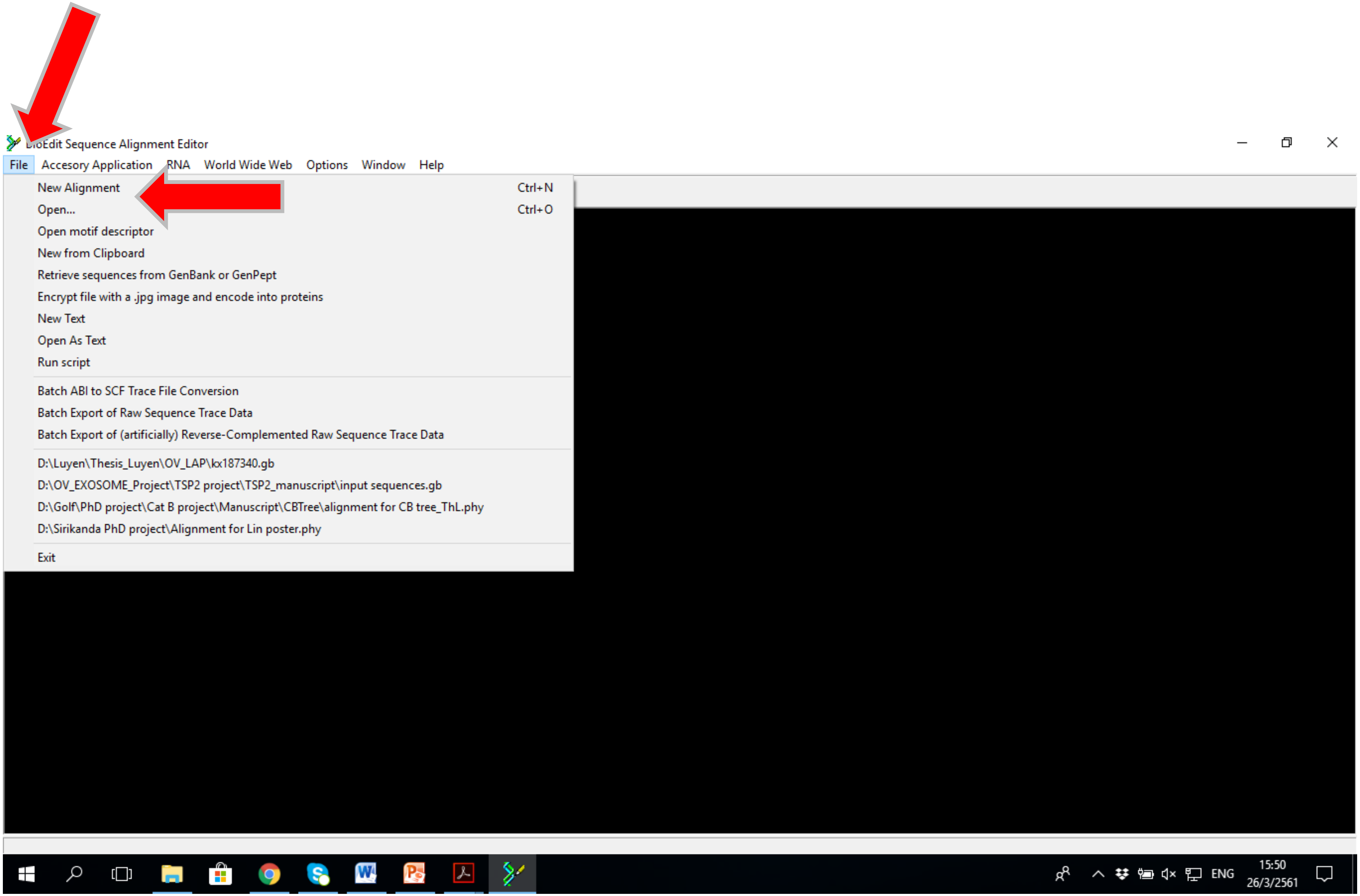

BioEdit Sequence Alignment Editor
File
Edit
Sequence
Untitled
Courier New
Mode:
Select / Slide
New Sequence
Edit Sequence
Edit all selected
Select Positions
Open at cursor position
Extract Positions
Go to Pubmed references
Overwrite/Retrieve sequences by gi number in title by HTTP to GenBank using gi|###|
Overwrite/Retrieve Genbank data by gi number in title by HTTP to GenBank (ignore sequence) using gi|###|
Overwrite/Retrieve sequences by gi number in title by HTTP to GenBank using gb|<accession>
Overwrite/Retrieve Genbank data by gi number in title by HTTP to GenBank (ignore sequence) using gb|<accession>|
True positions from alignment positions
Phylogeny / Taxonomy
Filter out sequences containing certain characters
Rename
Sort
PCR Primers / oligos
Pairwise alignment
Similarity Matrix (for pairwise alignments and shading)
Features
Sequence groups (or families)
Edit Mode
Mask
Toggle Color
Gaps
Manipulations
Nucleic Acid
Protein
Translate or Reverse-Translate (permanent)
Translate in selected frame (permanent)
Toggle Translation
Ctrl+G
Amino Acid Composition
Helical Wheel Diagram
Hydrophobic Moment matrix with Eisenberg consensus scale
Kyte & Doolittle Mean Hydrophobicity Profile
Eisenberg Scale Mean Hydrophobicity Profile
Cornette Scale Mean Hydrophobicity Profile
Parker HPLC Scale Mean Hydrophilicity Profile
Boyko Scale Mean Hydrophilicity Profile
Hopp & Woods Scale Mean Hydrophilicity
Eisenberg Hydrophobic Moment Profile
Mean Eisenberg Hydrophobic Moment Profile
Type he
29°C มีแดด
ENG
15:09
26/2/2565

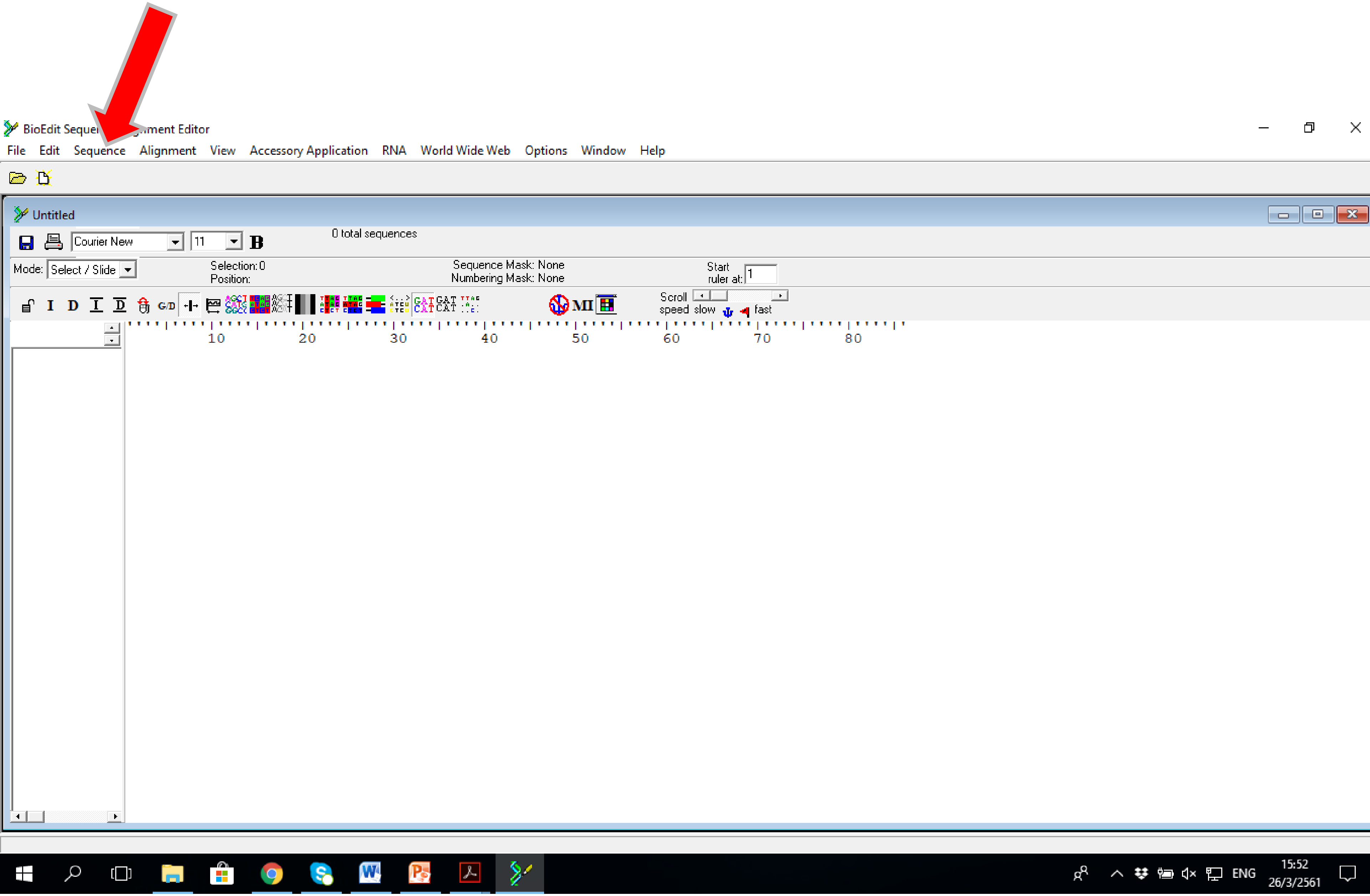
Multiple sequence alignment

1. Retrieve DNA or protein sequence from GenBank (more than 3 sequences).
2. Sequence should be high similarity (same gene or protein)
3. Save in BioEdit sequence file
4. Align-
   - Accessory application-ClustalW multiple alignment

Practice

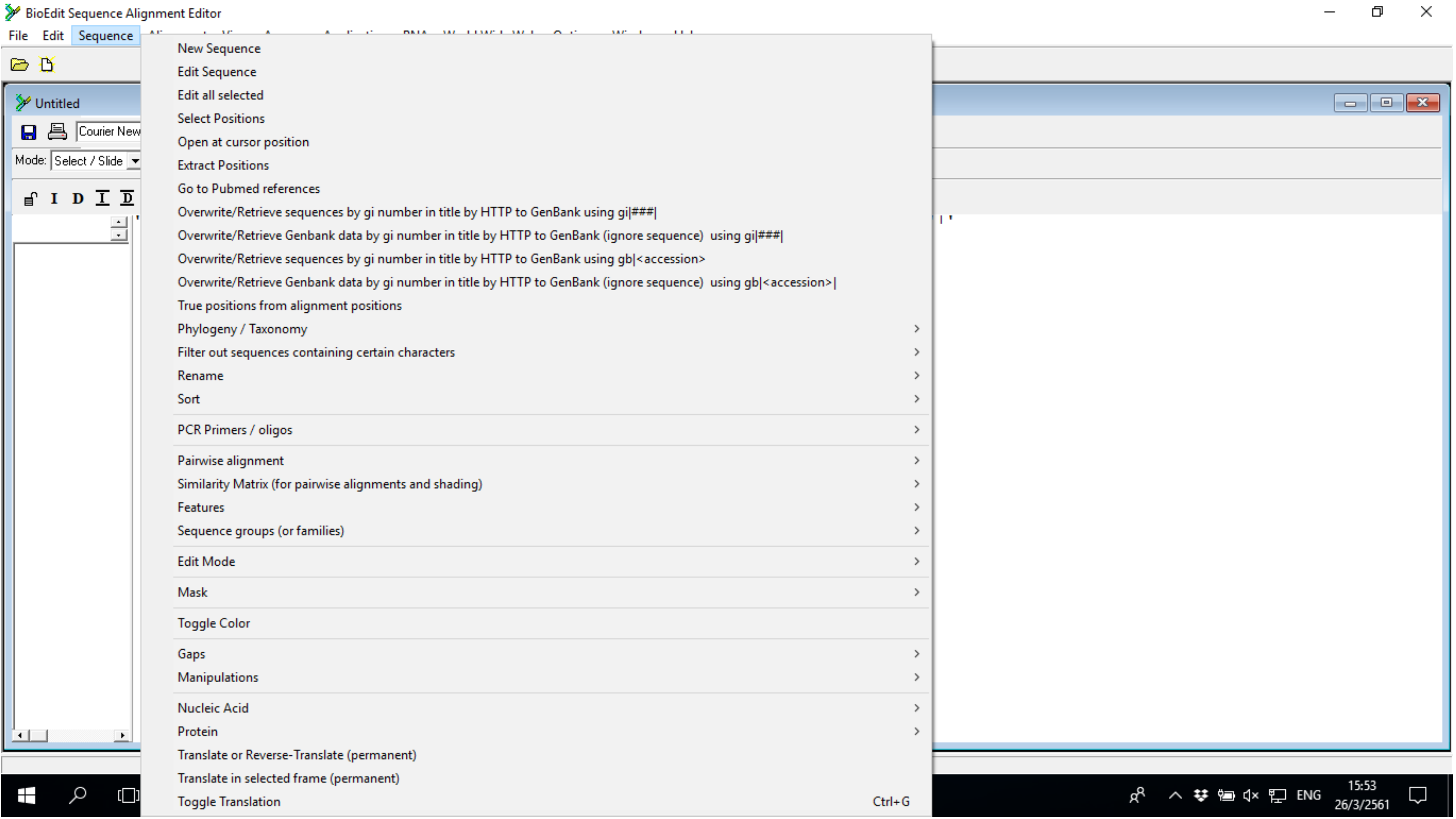
1. Retrieve DNA sequence from GenBank database (Select interested sequences)
2. Import new sequence into BioEdit
3. Practice
    - Manipulate sequence
    - Multiple sequences alignment


BioEdit Sequence Alignment Editor
File
Accesory Application
RNA
World Wide Web
Options
Window
Help
New Alignment Ctrl+N
Open... Ctrl+O
Open motif descriptor
New from Clipboard
Retrieve sequences from GenBank or GenPept
Encrypt file with a .jpg image and encode into proteins
New Text
Open As Text
Run script
Batch ABI to SCF Trace File Conversion
Batch Export of Raw Sequence Trace Data
Batch Export of (artificially) Reverse-Complemented Raw Sequence Trace Data
D:\Luyen\Thesis_Luyen\OV_LAP\kx187340.gb
D:\OV_EXOSOME_Project\TSP2 project\TSP2_manuscript\input sequences.gb
D:\Golf\PhD project\Cat B project\Manuscript\CBTree\alignment for CB tree_ThL.phy
D:\Sirikanda PhD project\Alignment for Lin poster.phy
Exit
ENG
15:50
26/3/2561

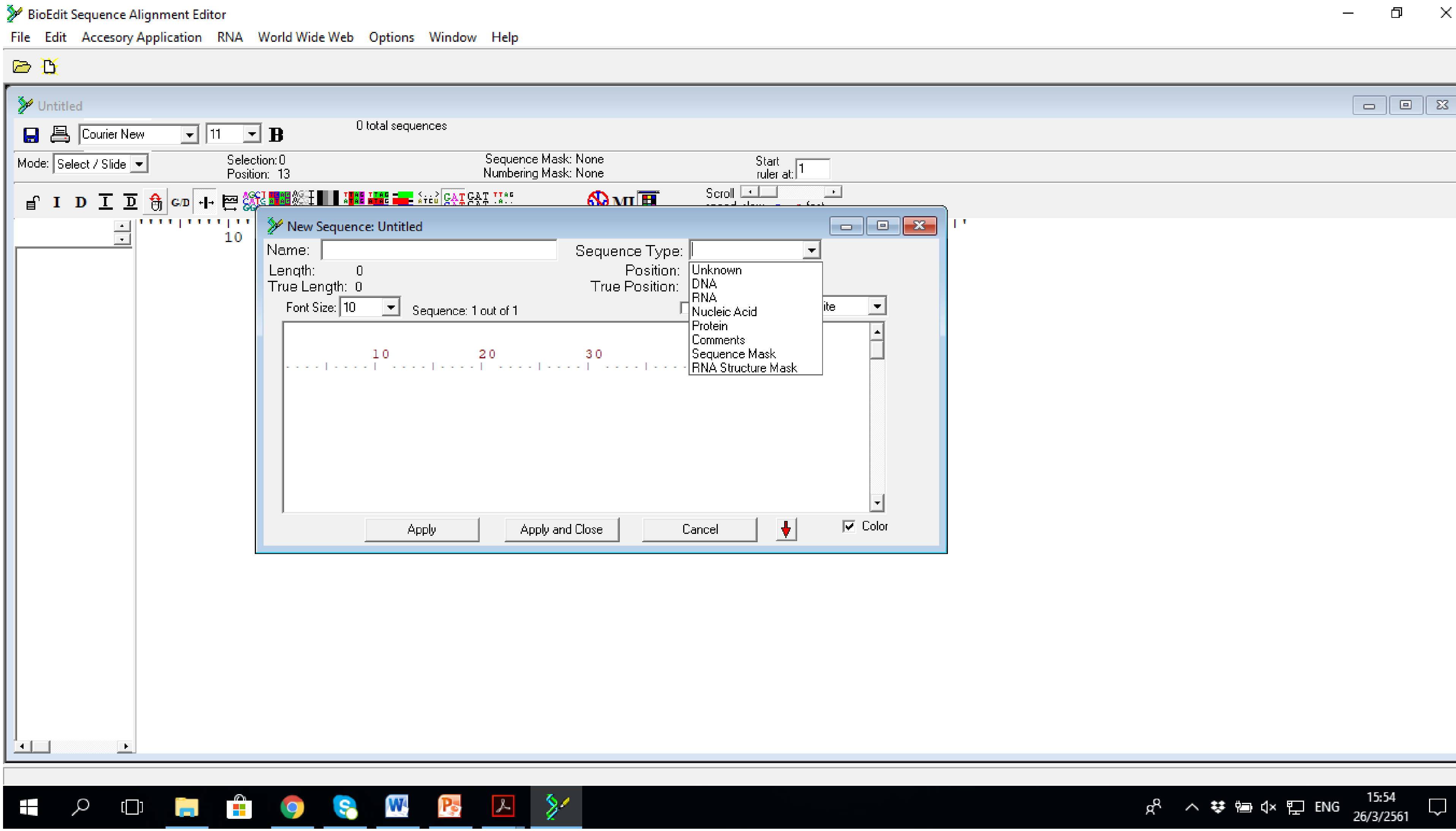


Courier New
11
0 total sequences
Mode: Select / Slide
Selection: 0
Position:
Sequence Mask: None
Numbering Mask: None
Start ruler at: 1

New sequence



Paste sequence in box, name sequence and indicate sequence type.



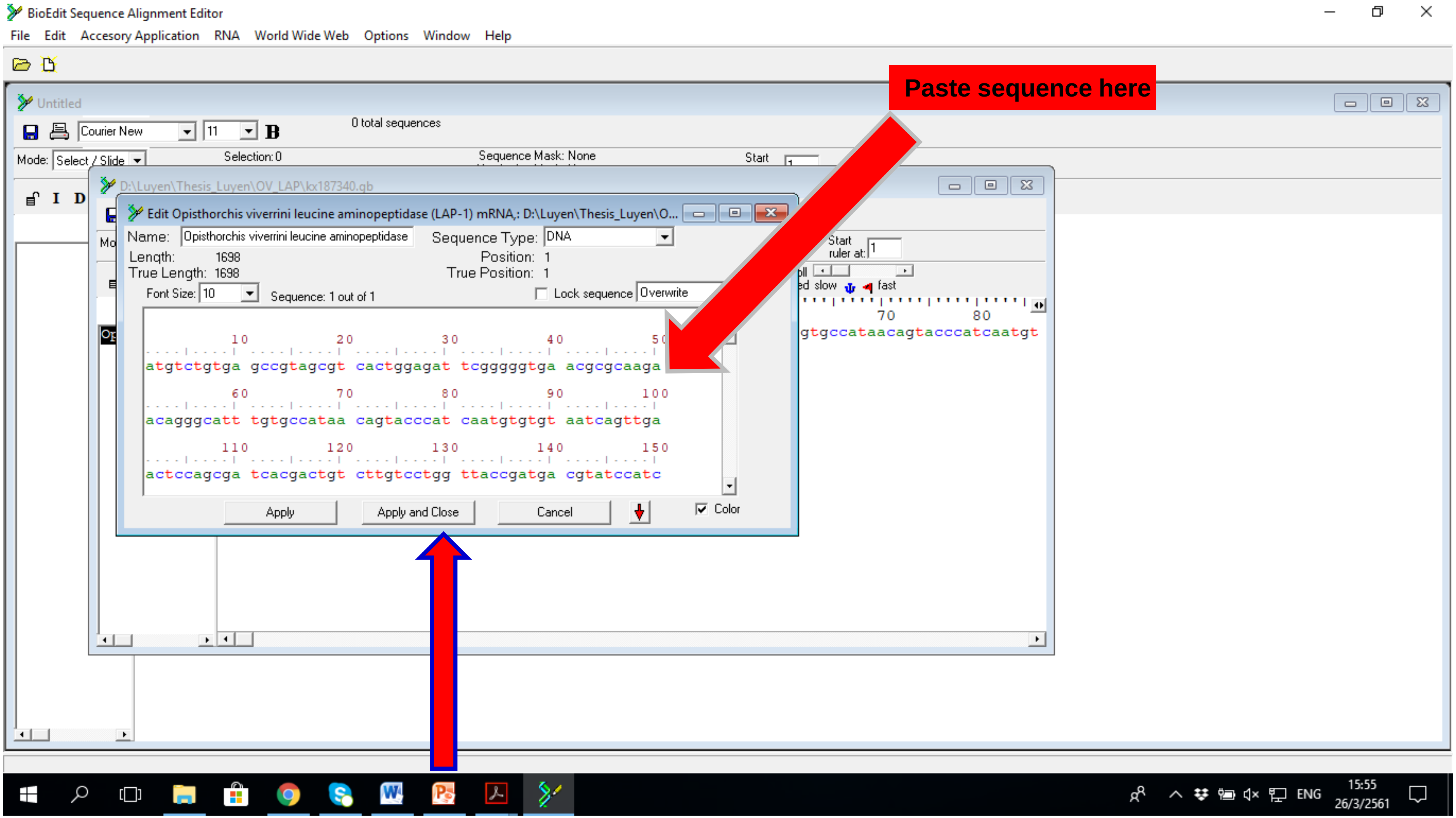
0 total sequences

Selection: 0
Position: 13

Sequence Mask: None
Numbering Mask: None

Length: 0
True Length: 0

Sequence: 1 out of 1



Paste sequence here

0 total sequences

Mode: Select / Slide

Selection: 0

Sequence Mask: None

D:\Luyen\Thesis_Luyen\OV_LAP\kx187340.gb

Edit Opisthorchis viverrini leucine aminopeptidase (LAP-1) mRNA,: D:\Luyen\Thesis_Luyen\O...

Name: Opisthorchis viverrini leucine aminopeptidase

Sequence Type: DNA

Length: 1698

True Length: 1698

Position: 1

True Position: 1

Font Size: 10

Sequence: 1 out of 1

☐ Lock sequence

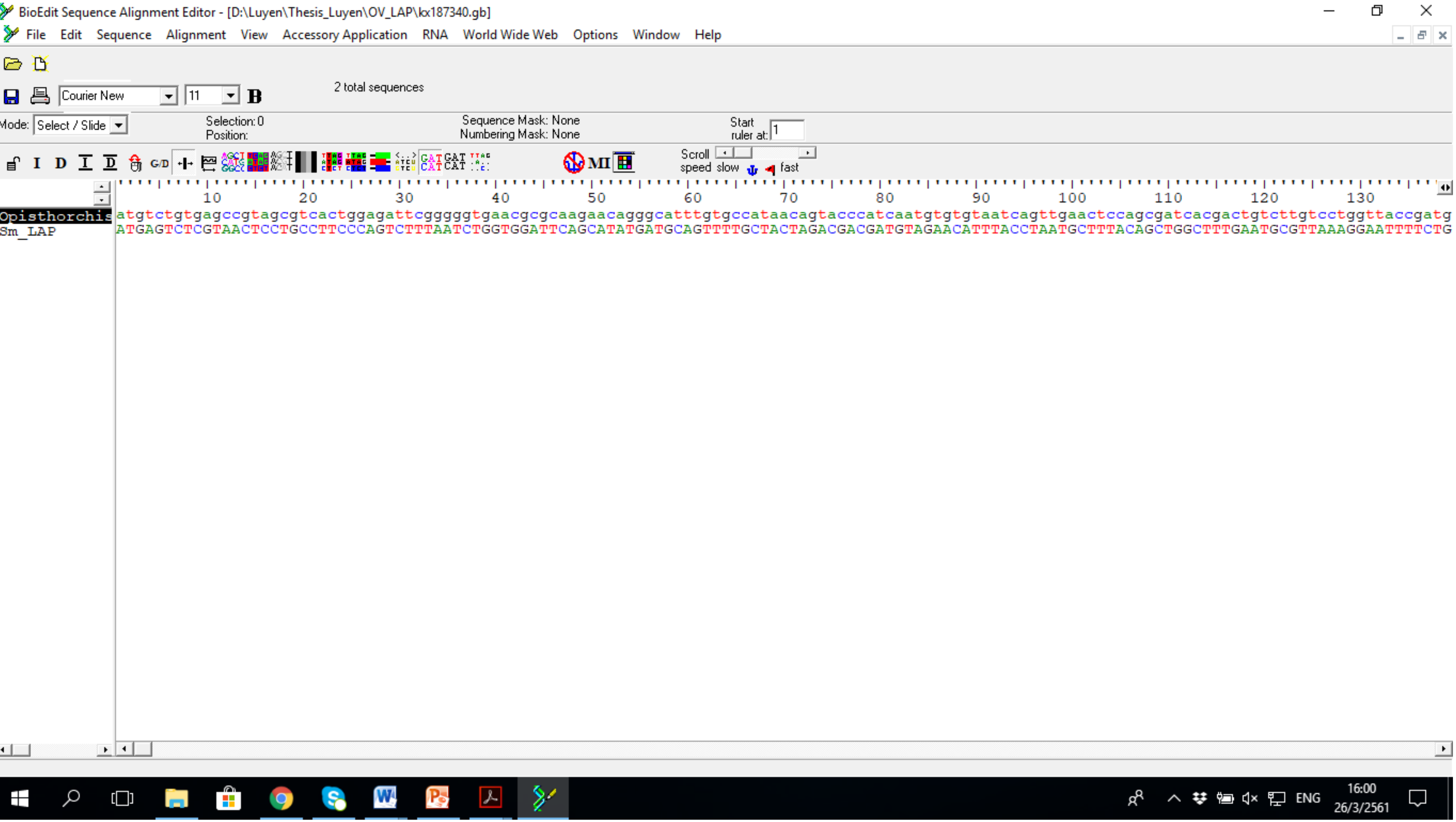
Overwrite

```
        10         20         30         40         50
atgtctgtga gccgtagcgt cactggagat tcgggggtga acgcgcaaga
        60         70         80         90        100
acagggcatt tgtgccataa cagtacccat caatgtgtgt aatcagttga
       110        120        130        140        150
actccagcga tcacgactgt cttgtcctgg ttaccgatga cgtatccatc
```

☑ Color

gtgccataacagtacccatcaatgt

Add more sequences



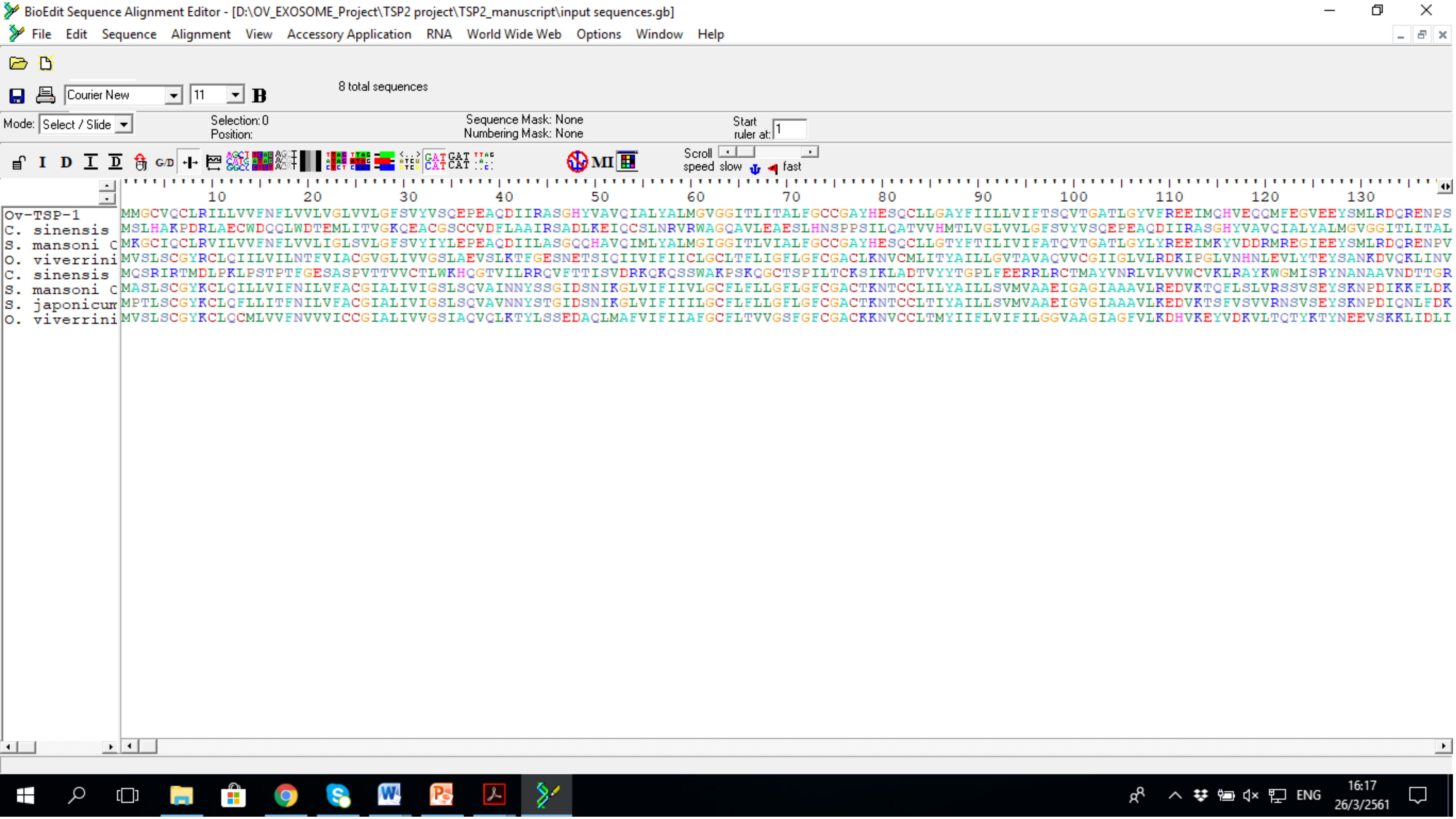
2 total sequences

```
Opisthorchis atgtctgtgagccgtagcgtcactggagattcgggggtgaacgcgcaagaacagggcatttgtgccataacagtacccatcaatgtgtgtaatcagttgaactccagcgatcacgactgtcttgtcctggttaccgatg
Sm_LAP       ATGAGTCTCGTAACTCCTGCCTTCCCAGTCTTTAATCTGGTGGATTCAGCATATGATGCAGTTTTGCTACTAGACGACGATGTAGAACATTTACCTAATGCTTTACAGCTGGCTTTGAATGCGTTAAAGGAATTTTCTG
```

Sequence alignment

- Pairwise alignment- เป็นการ align sequence 2 sequence เพื่อแสดง identities และ similarities

Add sequences



BioEdit Sequence Alignment Editor - [D:\OV_EXOSOME_Project\TSP2 project\TSP2_manuscript\input sequences.gb]

Courier New
11

8 total sequences

Mode: Select / Slide

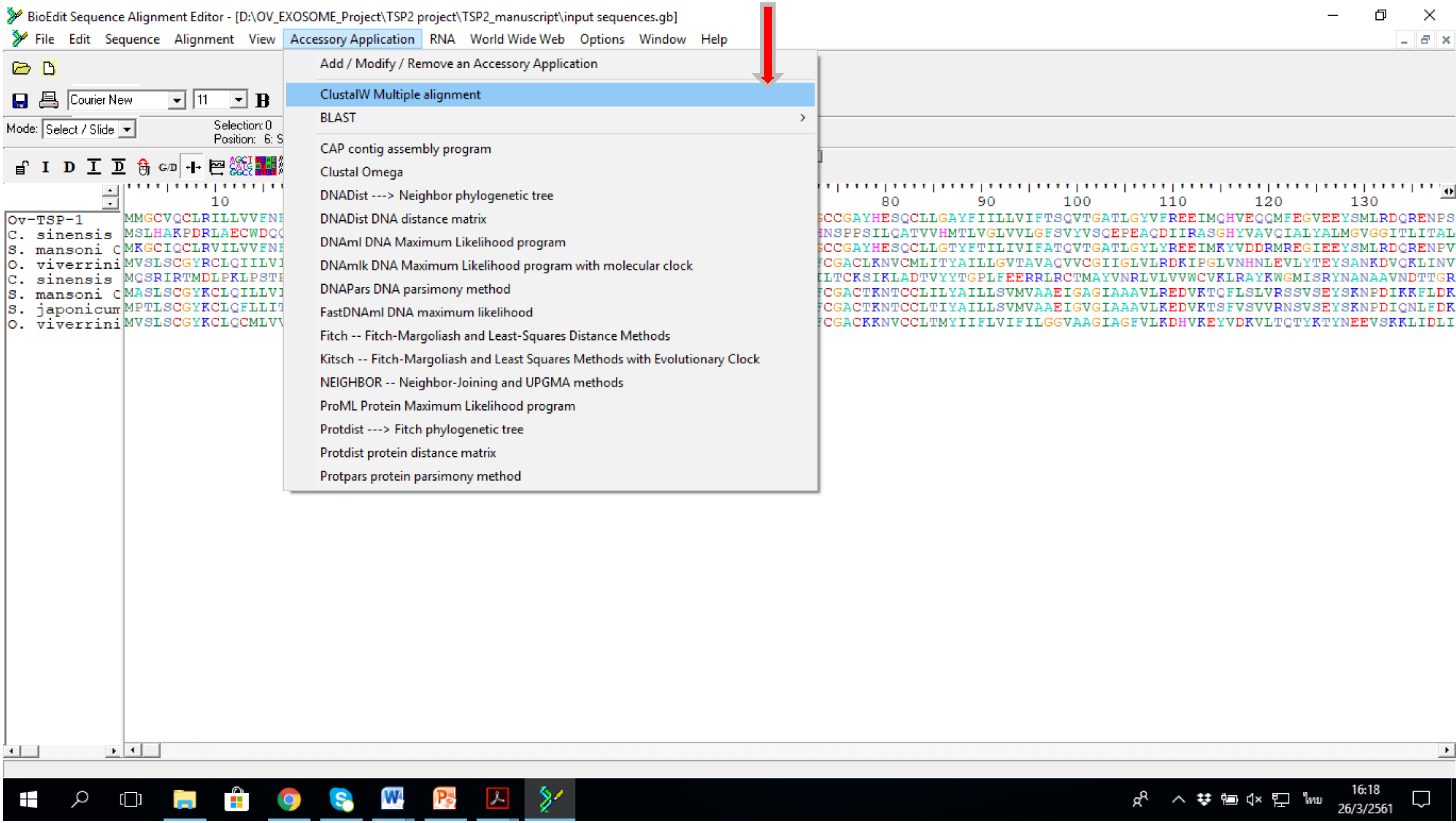
Selection: 0
Position:

Sequence Mask: None
Numbering Mask: None

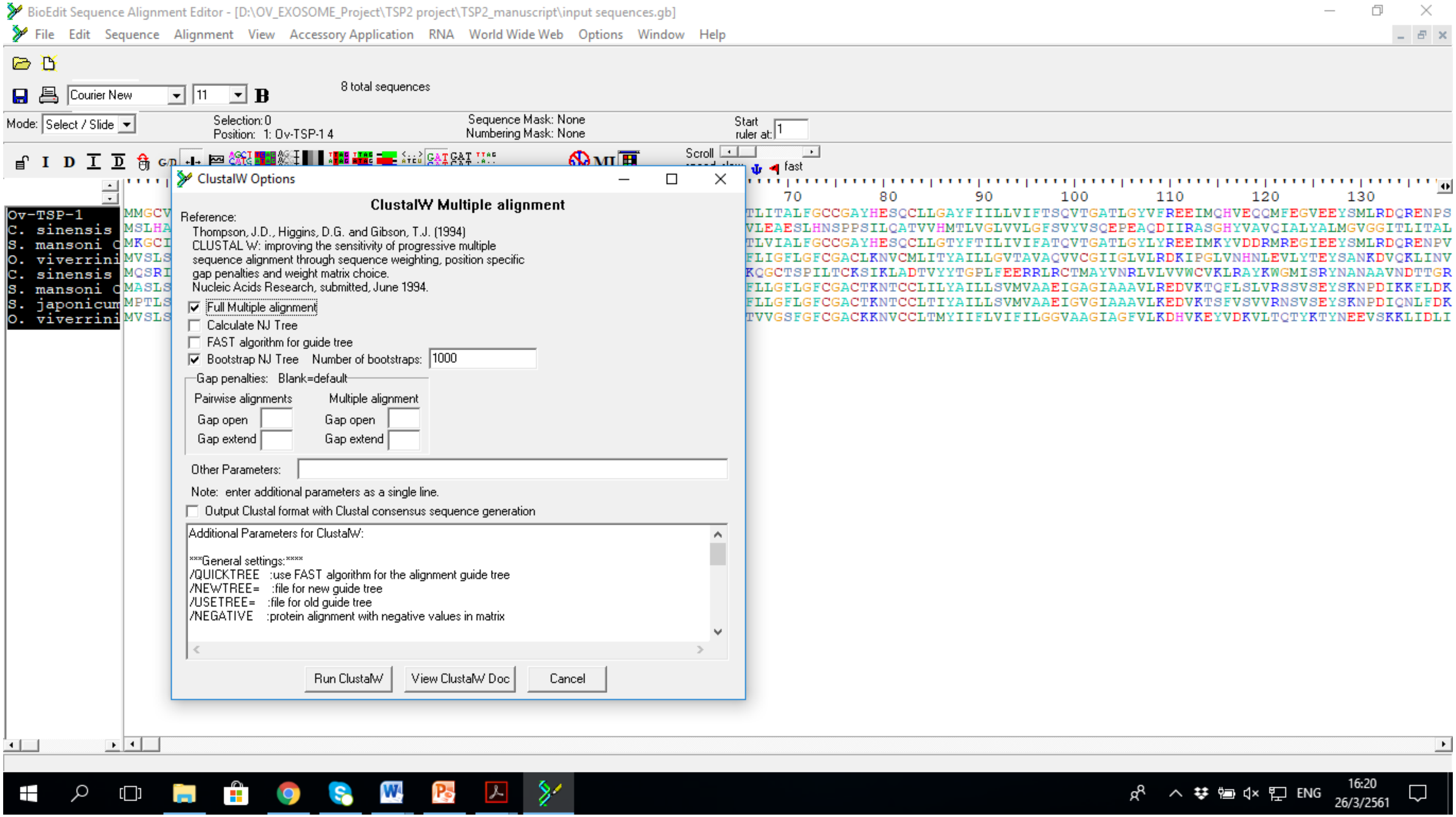
Start ruler at: 1

```
Ov-TSP-1     MMGCVQCLRILLVVFNFLVVLVGLVVLGFSVYVSQEPEAQDIIRASGHYVAVQIALYALMGVGGITLITALFGCCGAYHESQCLLGAYFIILLVIFTSQVTGATLGYVFREEIMQHVEQQMFEGVEEYSMLRDQRENPS
C. sinensis  MSLHAKPDRLAECWDQQLWDTEMLITVGKQEACGSCCVDFLAAIRSADLKEIQCSLNRVRWAGQAVLEAESLHNSPPSILQATVVHMTLVGLVVLGFSVYVSQEPEAQDIIRASGHYVAVQIALYALMGVGGITLITAL
S. mansoni C MKGCIQCLRVILVVFNFLVVLIGLSVLGFSVYIYLEPEAQDIILASGQQHAVQIMLYALMGIGGITLVIALFGCCGAYHESQCLLGTYFTILIVIFATQVTGATLGYLYREEIMKYVDDRMREGIEEYSMLRDQRENPV
O. viverrini MVSLSCGYRCLQIILVILNTFVIACGVGLIVVGSLAEVSLKTFGESNETSIQIIVIFIICLGCLTFLIGFLGFCGACLKNVCMLITYAILLGVTAVAQVVCGIIGLVLRDKIPGLVNHNLEVLYTEYSANKDVQKLINV
C. sinensis  MQSRIRTMDLPKLPSTPTFGESASPVTTVVCTLWKHQGTVILRRQVFTTISVDRKQKQSSWAKPSKQGCTSPILTCKSIKLADTVYYTGPLFEERRLRCTMAYVNRLVLVVWCVKLRAYKWGMISRYNANAAVNDTTGR
S. mansoni C MASLSCGYKCLQILLVIFNILVFACGIALIVIGSLSQVAINNYSSGIDSNIKGLVIFIIVLGCFLFLLGFLGFCGACTKNTCCLILYAILLSVMVAAEIGAGIAAAVLREDVKTQFLSLVRSSVSEYSKNPDIKKFLDK
S. japonicum MPTLSCGYKCLQFLLITFNILVFACGIALIVIGSLSQVAVNNYSTGIDSNIKGLVIFIIILGCFLFLLGFLGFCGACTKNTCCLTIYAILLSVMVAAEIGVGIAAAVLKEDVKTSFVSVVRNSVSEYSKNPDIQNLFDK
O. viverrini MVSLSCGYKCLQCMLVVFNVVVICCGIALIVVGSIAQVQLKTYLSSEDAQLMAFVIFIIAFGCFLTVVGSFGFCGACKKNVCCLTMYIIFLVIFILGGVAAGIAGFVLKDHVKEYVDKVLTQTYKTYNEEVSKKLIDLI
```

1. Accessory application
2. ClustalW Multiple alignment



3. Run ClustalW


BioEdit Sequence Alignment Editor - [D:\OV_EXOSOME_Project\TSP2 project\TSP2_manuscript\input sequences.gb]

Courier New 11

8 total sequences

Mode: Select / Slide

Selection: 0
Position: 1: Ov-TSP-1 4

Sequence Mask: None
Numbering Mask: None

Start ruler at: 1

Scroll speed: slow fast

```
Ov-TSP-1          MMGCV
C. sinensis       MSLHA
S. mansoni C      MKGCI
O. viverrini      MVSLS
C. sinensis       MQSRI
S. mansoni C      MASLS
S. japonicum      MPTLS
O. viverrini      MVSLS
```

```
 70        80        90        100       110       120       130
TLITALFGCCGAYHESQCLLGAYFIILLVIFTSQVTGATLGYVFREEIMQHVEQQMFEGVEEYSMLRDQRENPS
VLEAESLHNSPPSILQATVVHMTLVGLVVLGFSVYVSQEPEAQDIIRASGHYVAVQIALYALMGVGGITLITAL
TLVIALFGCCGAYHESQCLLGTYFTILIVIFATQVTGATLGYLYREEIMKYVDDRMREGIEEYSMLRDQRENPV
FLIGFLGFCGACLKNVCMLITYAILLGVTAVAQVVCGIIGLVLRDKIPGLVNHNLEVLYTEYSANKDVQKLINV
KQGCTSPILTCKSIKLADTVYYTGPLFEERRLRCTMAYVNRLVLVVWCVKLRAYKWGMISRYNANAAVNDTTGR
FLLGFLGFCGACTKNTCCLILYAILLSVMVAAEIGAGIAAAVLREDVKTQFLSLVRSSVSEYSKNPDIKKFLDK
FLLGFLGFCGACTKNTCCLTIYAILLSVMVAAEIGVGIAAAVLKEDVKTSFVSVVRNSVSEYSKNPDIQNLFDK
TVVGSFGFCGACKKNVCCLTMYIIFLVIFILGGVAAGIAGFVLKDHVKEYVDKVLTQTYKTYNEEVSKKLIDLI
```

ClustalW Options

ClustalW Multiple alignment

Reference:
Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994)
CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice.
Nucleic Acids Research, submitted, June 1994.

☑ Full Multiple alignment
☐ Calculate NJ Tree
☐ FAST algorithm for guide tree
☑ Bootstrap NJ Tree Number of bootstraps: 1000

Gap penalties: Blank=default

Pairwise alignments
Gap open
Gap extend

Multiple alignment
Gap open
Gap extend

Other Parameters:

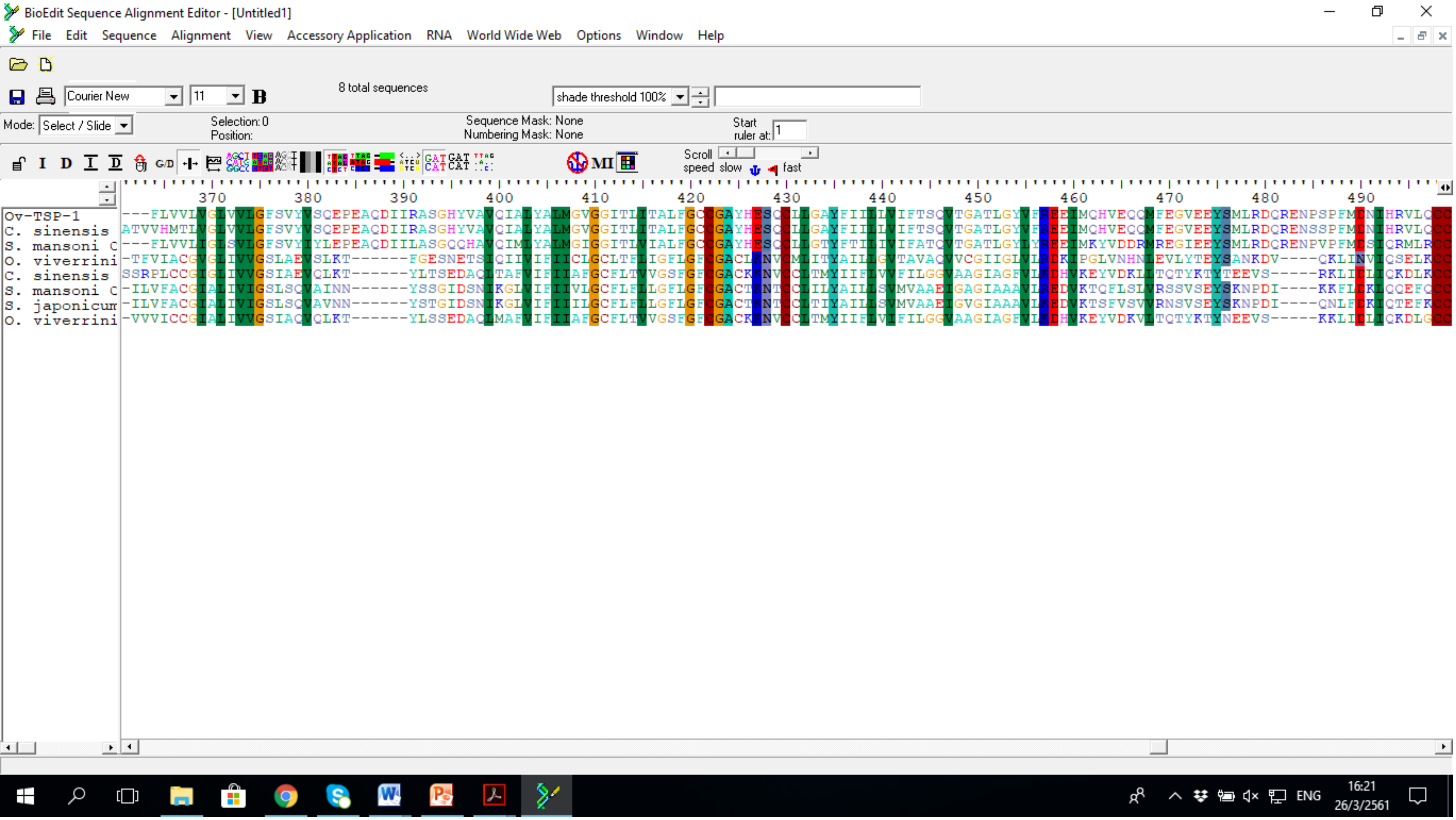
Note: enter additional parameters as a single line.

☐ Output Clustal format with Clustal consensus sequence generation

Additional Parameters for ClustalW:

General settings:*
/QUICKTREE :use FAST algorithm for the alignment guide tree
/NEWTREE= :file for new guide tree
/USETREE= :file for old guide tree
/NEGATIVE :protein alignment with negative values in matrix

Save as



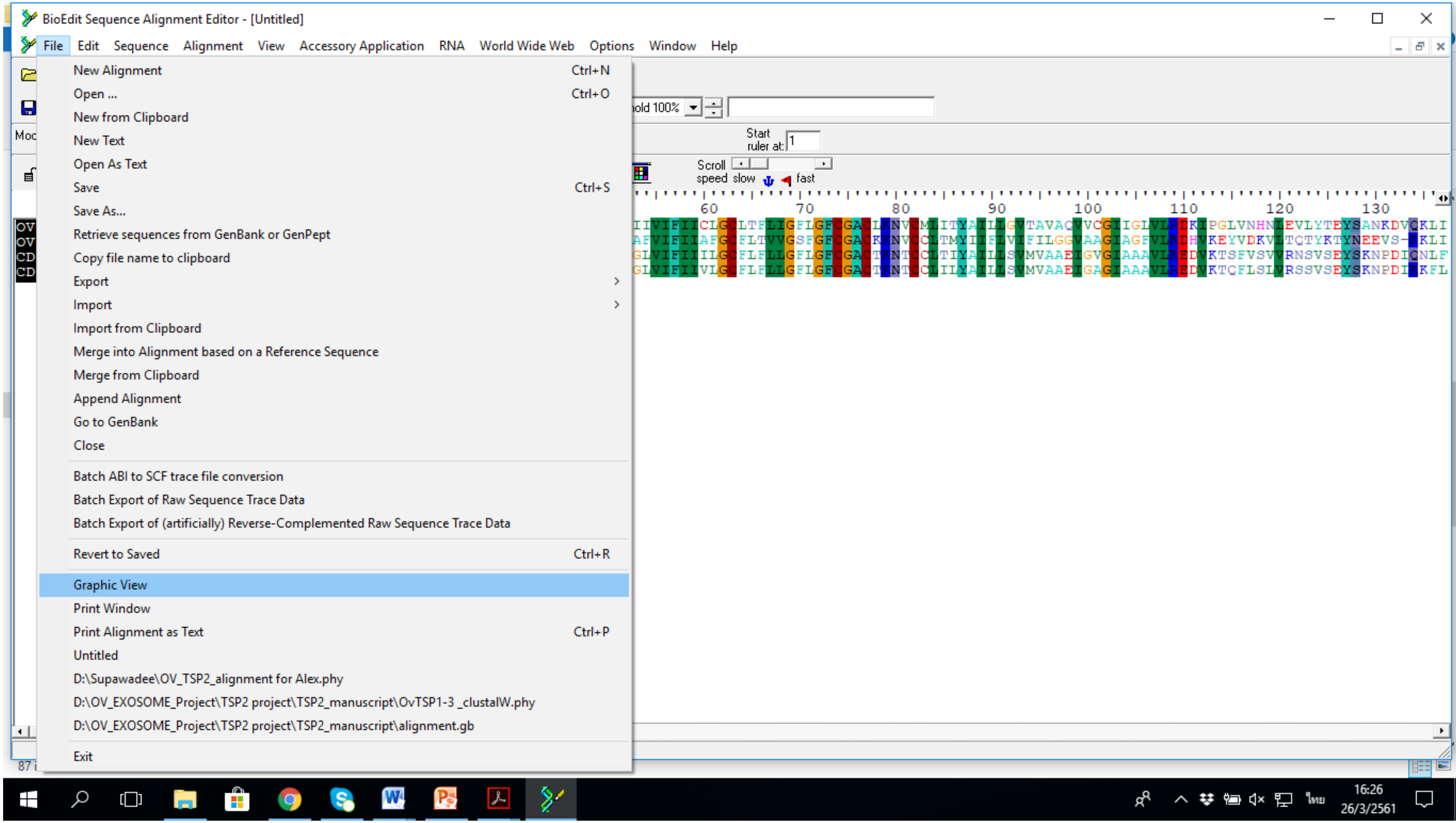
Courier New
11
8 total sequences
shade threshold 100%
Mode: Select / Slide
Selection: 0
Position:
Sequence Mask: None
Numbering Mask: None
Start ruler at: 1

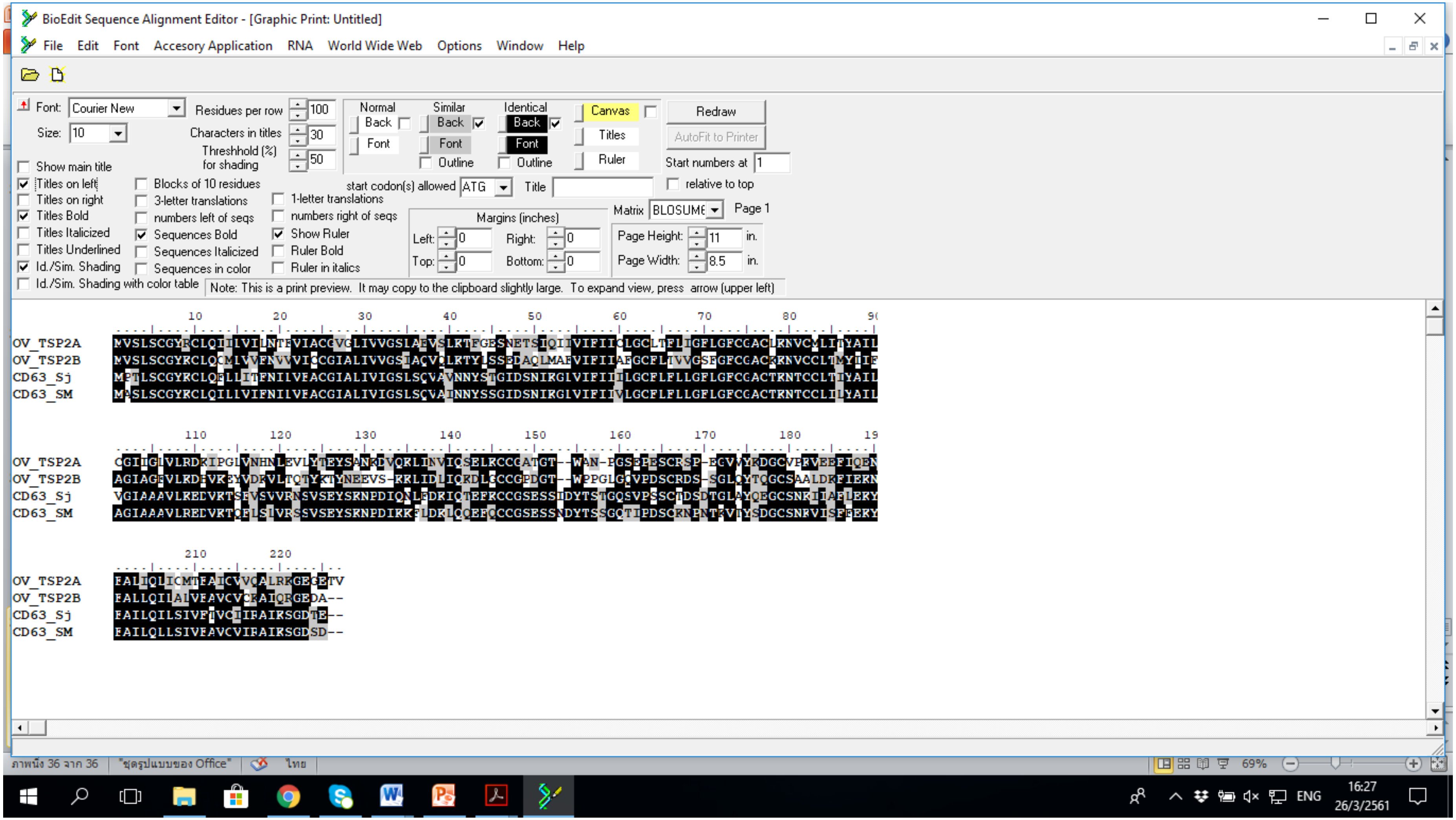
```
                   370       380       390       400       410       420       430       440       450       460       470       480       490
Ov-TSP-1     ---FLVVLVGLVVLGFSVYVSQEPEAQDIIRASGHYVAVQIALYALMGVGGITLITALFGCCGAYHESQCLLGAYFIILLVIFTSQVTGATLGYVFREEIMQHVEQQMFEGVEEYSMLRDQRENPSPFMDNIHRVLQCC
C. sinensis  ATVVHMTLVGLVVLGFSVYVSQEPEAQDIIRASGHYVAVQIALYALMGVGGITLITALFGCCGAYHESQCLLGAYFIILLVIFTSQVTGATLGYVFREEIMQHVEQQMFEGVEEYSMLRDQRENSSPFMDNIHRVLQCC
S. mansoni C ---FLVVLIGLSVLGFSVYIYLEPEAQDIILASGQQHAVQIMLYALMGIGGITLVIALFGCCGAYHESQCLLGTYFTILIVIFATQVTGATLGYLYREEIMKYVDDRMREGIEEYSMLRDQRENPVPFMDSIQRMLRCC
O. viverrini -TFVIACGVGLIVVGSLAEVSLKT------FGESNETSIQIIVIFIICLGCLTFLIGFLGFCGACLKNVCMLITYAILLGVTAVAQVVCGIIGLVLRDKIPGLVNHNLEVLYTEYSANKDV-----QKLINVIQSELKCC
C. sinensis  SSRPLCCGIGLIVVGSIAEVQLKT------YLTSEDAQLTAFVIFIIAFGCFLTVVGSFGFCGACKKNVCCLTMYIIFLVVFILGGVAAGIAGFVLKDHVKEYVDKLLTQTYKTYTEEVS------RKLIDLIQKDLKCC
S. mansoni C -ILVFACGIALIVIGSLSQVAINN------YSSGIDSNIKGLVIFIIVLGCFLFLLGFLGFCGACTKNTCCLILYAILLSVMVAAEIGAGIAAAVLREDVKTQFLSLVRSSVSEYSKNPDI-----KKFLDKLQQEFQCC
S. japonicum -ILVFACGIALIVIGSLSQVAVNN------YSTGIDSNIKGLVIFIIILGCFLFLLGFLGFCGACTKNTCCLTIYAILLSVMVAAEIGVGIAAAVLKEDVKTSFVSVVRNSVSEYSKNPDI-----QNLFDKIQTEFKCC
O. viverrini -VVVICCGIALIVVGSIAQVQLKT------YLSSEDAQLMAFVIFIIAFGCFLTVVGSFGFCGACKKNVCCLTMYIIFLVIFILGGVAAGIAGFVLKDHVKEYVDKVLTQTYKTYNEEVS------KKLIDLIQKDLGCC
```

16:21
26/3/2561

View alignments

1. File
2. Graphic view





Font: Courier New
Size: 10
Residues per row 100
Characters in titles 30
Threshhold (%) for shading 50
Start numbers at 1
☐ relative to top
start codon(s) allowed ATG
Title
Matrix BLOSUM€
☐ Show main title
☑ Titles on left
☐ Titles on right
☑ Titles Bold
☐ Titles Italicized
☐ Titles Underlined
☑ Id./Sim. Shading
☐ Id./Sim. Shading with color table
☐ Blocks of 10 residues
☐ 3-letter translations
☐ numbers left of seqs
☑ Sequences Bold
☐ Sequences Italicized
☐ Sequences in color
☐ 1-letter translations
☐ numbers right of seqs
☑ Show Ruler
☐ Ruler Bold
☐ Ruler in italics
Left: 0
Right: 0
Top: 0
Bottom: 0
Page Height: 11 in.
Page Width: 8.5 in.

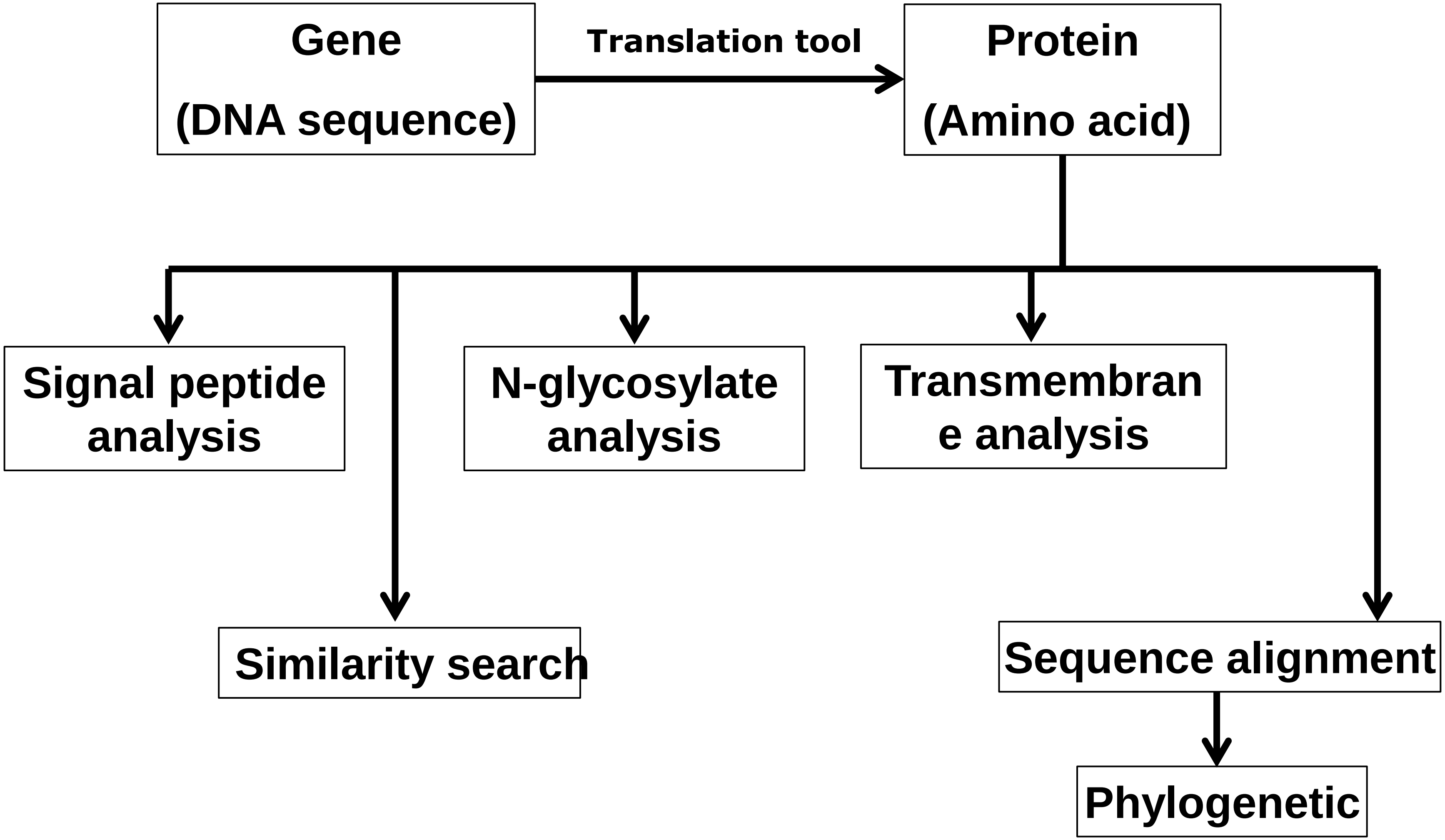
Note: This is a print preview. It may copy to the clipboard slightly large. To expand view, press arrow (upper left)

```
                   10        20        30        40        50        60        70        80        90
OV_TSP2A   MVSLSCGYRCLQIIIVILNTEVIACGVGLIVVGSLAEVSLKTFGESNETSIQIIVIFIICLGCLTFLIGFLGCCACLRNVCMLITYAIL
OV_TSP2B   MVSLSCGYKCLQCMIVVFNVVVICCGIALIVVGSIACVQLKTYLSSEDAQLMAFVIFIIAFGCFLIVVGSFGFCCACKKNVCCLTMYIIF
CD63_Sj    MPTLSCGYKCLQFLLITFNIIVFACGIALIVIGSLSCVAVNNYSTGIDSNIKGIVIFIIILGCFLFLLGFLGFCCACTKNTCCLTIYAIL
CD63_SM    MASLSCGYKCLQILIVIFNIIVFACGIALIVIGSLSCVAINNYSSGIDSNIKGIVIFIIVLGCFLFLLGFLGFCCACTKNTCCLILYAIL

                  110       120       130       140       150       160       170       180       19
OV_TSP2A   CGIIGLVLRDKIPGLVNHNLEVLYTEYSANKDVQKLINVIQSELKCCGATGT--WAN-PGSEPESCRSP-EGVWYKDGCVPKVEEFIQEN
OV_TSP2B   AGIAGFVLRDHVKEYVDKVLTQTYKTYNEEVS-KKLIDLIQKDLGCCGPDGT--WPPGLGQVPDSCRDS-SGLQYTQGCSAALDKFIERN
CD63_Sj    VGIAAAVLKEDVKTSEVSVVRNSVSEYSKNPDIQNLFDKIQTEFKCCGSESSIDYTSTGQSVPSSCTDSDTGLAYQEGCSNKIIAFLEKY
CD63_SM    AGIAAAVLREDVKTQELSIVRSSVSEYSKNPDIKKFLDKLQQEFQCCGSESSNDYTSSGQTIPDSCKNPNTKVTYSDGCSNKVISFFEKY

                  210       220
OV_TSP2A   EALIQLICMTEAICVVQALRKGEGETV
OV_TSP2B   EALLQILAIVEAVCVCKAIQRGEDA--
CD63_Sj    EAILQILSIVFTVCIIFAIKSGDTE--
CD63_SM    EAILQLLSIVEAVCVIFAIKSGDSD--
```

Flow chart for gene structure and function



Gene

(DNA sequence)

Translation tool

Protein

(Amino acid)

Signal peptide analysis

N-glycosylate analysis

Transmembrane analysis

Similarity search

Sequence alignment

Phylogenetic

Other tools on web

- Signal peptide analysis;

https://services.healthtech.dtu.dk/service.php?SignalP-5.0

- Transmembrane analysis

https://services.healthtech.dtu.dk/service.php?TMHMM-2.0

- EXPASY (Bioinformatics resource portal of the SIB Swiss Institute of Bioinformatics)

https://www.expasy.org/