

Molecular Biology Laboratory

Bioinformatics and Genomics Lab.

Week2. Protein Sequence Alignment (BLAST, Clustal Omega) & Domain Search

1. Practice how to use UniProt database

- Find information on GFP and get the amino acid sequence
 - Search "uniprot" in google and access UniProt.

A screenshot of a Google search for 'uniprot'. The search bar contains 'uniprot' and the search button is highlighted. Below the search bar, there are navigation options: '전체', '이미지', '동영상', '도서', '뉴스', and '더보기'. The search results show approximately 21,300,000 results in 0.31 seconds. The top result is 'UniProt', with a red arrow pointing to the URL 'https://www.uniprot.org'. Below the main result, there are links for 'UniProtKB', 'BLAST', 'Retrieve/ID mapping', and 'UniProtKB 227,339,950 results'. On the right side, there is a UniProt logo and a section titled 'UniProt' with a description in Korean and a '원래 설명 보기' link.

- Search "GFP" and select 1st one (P42212).

A screenshot of the UniProt website showing search results for 'GFP'. The top navigation bar includes 'UniProt', 'BLAST', 'Align', 'Peptide search', 'ID mapping', 'SPARQL', 'UniProtKB', 'GFP', 'Advanced', 'List', 'Search', and 'Help'. The main content area shows 'UniProtKB 6,839 results' for the search 'GFP'. Below this, there is a table with columns: 'Entry', 'Entry Name', 'Protein Names', 'Gene Names', 'Organism', and 'Length'. The first row is highlighted with a red arrow pointing to the entry 'P42212'. The second row is 'Q9SEU7'. On the left side, there is a 'Status' section with 'Reviewed (Swiss-Prot) (392)' and 'Unreviewed (TrEMBL) (6,447)'. Below that, there is a 'Popular organisms' section with a list of organisms and their counts: Zebrafish (57), A. thaliana (42), Human (25), Mouse (18), and E. coli K12 (12). A 'Taxonomy' section is also visible at the bottom left. A 'Feedback' button is located at the bottom right.

Entry	Entry Name	Protein Names	Gene Names	Organism	Length
<input type="checkbox"/> P42212	GFP_AEQVI	Green fluorescent protein	GFP	Aequorea victoria (Water jellyfish) (Mesonema victoria)	238 AA
<input type="checkbox"/> Q9SEU7	TRXM3_ARATH	Thioredoxin M3, chloroplastic[...]	GAT1, At2g15570, F9O13.12	Arabidopsis thaliana (Mouse-ear cress)	173 AA

- We can find information on GFP protein. To get the amino acid sequence, click "Sequence"- "Download".

UniProt BLAST Align Peptide search ID mapping SPARQL UniProtKB Advanced | List Search

P42212 · GFP_AEQVI

Green fluorescent protein · *Aequorea victoria* (Water jellyfish) (*Mesonema victoria*) · Gene: GFP · 238 amino acids · Evidence at protein level · Annotation score: 5/9

Entry Feature viewer Publications External links History

BLAST Align Download Add Add a publication Entry feedback

Functionⁱ

Energy-transfer acceptor. Its role is to transduce the blue chemiluminescence of the protein aequorin into green fluorescent light by energy transfer. Fluoresces in vivo upon receiving energy from the Ca²⁺-activated photoprotein aequorin.

Biotechnology

Green fluorescent protein has been engineered to produce a vast number of variously colored mutants, fusion proteins, and biosensors. Green fluorescent protein can be mutated to emit at different wavelengths such as blue for BFP (when Tyr-66 is replaced by His), cyan for CFP (when Tyr-66 is replaced by Trp), and yellow for YFP (when THR-203 is replaced by Tyr). Further generation of mutants led to more stable proteins (at 37 degrees Celsius for example) with brighter fluorescence and longer fluorescence lifetimes. Fluorescent proteins and their mutated allelic forms have become a useful

Function

Names & Taxonomy

Subcellular Location

Phenotypes & Variants

PTM/Processing

Expression

Interaction

Structure

Family & Domains

Sequence

Similar Proteins

Sequenceⁱ

Tools Download Add Highlight Copy sequence

Length 238 Last updated 1995-11-01 v1
Mass (Da) 26,886 Checksumⁱ EA5A6F21FBFB6E05

MSKGEELFTG VVPILVELDG DVNGHKFSVS GEGEGDATYG KLTCLKFICTT GKLPVWPPTL VTTFSYGVQC FSRYPDHMKQ
HDFKSAPE GYVQERTIFF KDDGNYKTRA EVKFEQDTLV NRIELKGI DFKEDGN LGHKLEYNYN SHNYY I MADKQKNGI KYNFKI RHN I EDGSVQLAD YIMADKQKNG
IKVNFKIRHN IEDGSVQLAD HYQQNTPIGD GPVLLPDNHV LSTQSALSKD PNEKRDHMLV LEFVTAAGIT HGMDELYK

Features

Showing features for sequence conflict¹.

1 20 40 60 80 100 120 140 160 180 200 220 238

```
>sp|P42212|GFP_AEQVI Green fluorescent protein OS=Aequorea victoria OX=6100 GN=GFP PE=1 SV=1
MSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGLTCLKFICTTGKLPVWPPTL
VTTFSYGVQCFSRYPDHMKQHDFFKSAPEGYVQERTIFFKDDGNYKTRA EVKFEQDTLV
NRIELKGI DFKEDGN LGHKLEYNYN SHNYY I MADKQKNGI KYNFKI RHN I EDGSVQLAD
HYQQNTPI GDGPVLLPDNHV LSTQSALSKDPNEKRDHMLV LEFVTAAGIT HGMDELYK
```

2. Practice how to use BLASTP

- Try sequence alignment of GFP
 - Search "blastp" in google and access BLASTP

A screenshot of a Google search for "blastp". The search bar shows "blastp" and the results list includes:

- https://blast.ncbi.nlm.nih.gov > Blast
Protein BLAST: search protein databases using a protein query
BLASTP simply compares a protein query to a protein database. PSI-BLAST allows the user to build a PSSM (position-specific scoring matrix) using the results of ...
- https://www.ncbi.nlm.nih.gov > BLAST
BLAST: Basic Local Alignment Search Tool - NCBI
The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to ...
- http://www.incodom.kr > BLAST
BLAST - 인코덤, 생물정보 전문위키
2019. 12. 1. — -p, n/a, BLAST program: blastn, **blastp**, blastx, tblastn, ... -i, -query, Input sequence file. -d, -db, BLAST database. -o, -out, Output file.

On the right side, there is a sidebar with a "Protein BLAST" logo and a "BLAST" section with the text "생명공학기술" and "BLAST는 생물정보학에서 단백질의 나 DNA/RNA 서열의 뉴클레오타이드와 비교하여 유사성을 찾는 프로그램이다."

- Copy and paste the GFP sequence to the query sequence box and set job title. Select program algorithm as "blastp" and click "BLAST" button to run the tool

A screenshot of the BLASTP suite web interface. The page title is "BLAST® » blastp suite" and the sub-header is "Standard Protein BLAST".

The interface is divided into several sections:

- blastp suite**: A navigation bar with tabs for "blastn", "blastp" (selected), "blastx", "tblastn", and "tblastx".
- Enter Query Sequence**: A section with a text input field containing a FASTA sequence: "NRIELKGIDFKEDGNILGHKLEYNNSHNHYIMADKQKNGIKVNFKIRHNIED GSVQLAD HYQQNTPIGDGPVLLPDNHYLSTQSALS KDPNEKRDHMLLEFVTAAGITHG MDELYK". A red arrow points to the "Clear" button next to the input field. Below the input field is a "Job Title" field containing "GFP". A red arrow points to the "Job Title" field. There is also a "Query subrange" section with "From" and "To" input fields.
- Choose Search Set**: A section with a "Databases" section where "Standard databases (nr etc.):" is selected (marked with a "New" badge) and "Experimental databases" is unselected. There is a button that says "Try experimental clustered nr database" with a magnifying glass icon. Below it, there is a "Compare" section with a checkbox for "Select to compare standard and experimental database".

Program Selection

Algorithm

- Quick BLASTP (Accelerated protein-protein BLAST)
- 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 nr** using **Blastp (protein-protein BLAST)**

Show results in a new window

- Select GFP of "*Aequorea victoria*" and check information (Query cover, Per. Ident, Graphic Summary etc.)

Job Title GFP

RID [PHBHJCYM01R](#) Search expires on 11-08 14:05 pm [Download All](#)

Program BLASTP [Citation](#)

Database nr [See details](#)

Query ID lc|Query_69491

Description sp|P42212|GFP_AEQVI Green fluorescent protein OS=Ae ...

Molecule type amino acid

Query Length 238

Other reports [Distance tree of results](#) [Multiple alignment](#) [MSA viewer](#)

Filter Results

Organism only top 20 will appear exclude

Type common name, binomial, taxid or group name

[+ Add organism](#)

Percent Identity to

E value to

Query Coverage to

[Filter](#) [Reset](#)

Compare these results against the new Clustered nr database [?](#) [BLAST](#)

Descriptions | [Graphic Summary](#) | [Alignments](#) | [Taxonomy](#)

Sequences producing significant alignments [Download](#) [Select columns](#) Show [?](#)

select all 1 sequences selected [GenPept](#) [Graphics](#) [Distance tree of results](#) [Multiple alignment](#) [MSA Viewer](#)

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input type="checkbox"/>	green fluorescent protein [synthetic construct]	synthetic construct	496	496	100%	5e-177	100.00%	247	AAB51347.1
<input type="checkbox"/>	linker-GFP [Cloning vector pGGD001]	Cloning vector p...	497	497	100%	6e-177	99.16%	272	AHE38523.1
<input type="checkbox"/>	unnamed protein product [Binary vector pZH2B-2ox3]	Binary vector pZ...	496	496	100%	6e-177	99.16%	266	BAJ22064.1
<input type="checkbox"/>	GFP-NLS [Cloning vector pGGC012]	Cloning vector p...	498	498	100%	1e-176	99.16%	323	AHE38509.1
<input checked="" type="checkbox"/>	RecName: Full=Green fluorescent protein [Aequorea victoria]	Aequorea victoria	494	494	100%	1e-176	100.00%	238	P42212.1
<input type="checkbox"/>	mitochondria-targeted synthetic green fluorescent protein [Gateway binary vector R4L1pGWB471]	Gateway binary ...	497	497	100%	1e-176	99.16%	296	BBG75473.1

3. Practice how to use Clustal Omega

- Try multiple sequence alignment (MSA) of GFP, CFP, YFP, and RFP
 - Copy and paste sequence to the given "FASTA" file (Week2_Fluorescence_Protein_Sequences_for_Upload.fa). Edit header part of GFP as ">GFP" for MSA.



```
Week2_Fluorescence_Protein_Sequences.fa - Windows 메모장
파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)
>GFP
MSKGEELFTGVVPIVVELDGDVNGHKFSVSGEGEGDATYGKLT LKFICTTGKLPVPWPTL
VTTFSYGVQCFSRYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFEGDTLV
NRIELKGIDFKEDGNILGHKLEYNNSHNVYIMADKQKNGIKVNFKIRHNIEDGSVQLAD
HYQQNTPIGDGPVLLPDNHYLSTQSALS KDPNEKRDHMLLEFVTAAGITHGMDELYK
>CFP
MSKGEELFTGVVPIVVELDGDVNGHKFSVSGEGEGDATYGKLT LKFICTTGKLPVPWPTL
VTTFSWGVQCFSRYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFEGDTLV
NRIELKGIDFKEDGNILGHKLEYNNSHNVYIMADKQKNGIKVNFKIRHNIEDGSVQLAD
HYQQNTPIGDGPVLLPDNHYLSTQSALS KDPNEKRDHMLLEFVTAAGITHGMDELYK
>YFP
MSKGEELFTGVVPIVVELDGDVNGHKFSVSGEGEGDATYGKLT LKLLCTTGKLPVPWPTL
VTTFGYGLQCFARYPDHMKRHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFEGDTLV
NRIELKGIDFKEDGNILGHKLEYNNSHNVYIMADKQKNGIKVNFKIRHNIEDGSVQLAD
HYQQNTPIGDGPVLLPDNHYLSYQSALFKDPNEKRDHMLLEFLTAAGITEGMNELYK
>RFP
MRSSKNVIKEFMRFKVRMEGTVNGHEFEIEGEGEGRPYEGHNTV KLVTKGGPLPFAWDI
LSPQFQYGSKVYVKHPADIPDYKLSFPEGFKWERVMN FEDGGVVTVDSSLDGCFIY
KVKFIGVNFPSDGPVMQKKTMGWEASTERLYPRDGV LKGEIHKALKLKDGGHYLVEFKSI
YMAKKPVQLPGYYYVDSKLDITSHNEDYTIVEQYERTEGRHHLFL
Ln 20, Col 46 100% Windows (CRLF) UTF-8
```

- Search "clustal omega" in google and access Clustal Omega.



- Upload "FASTA" file, go to below and click "Submit" for MSA

Multiple Sequence Alignment

Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between **three or more** sequences. For the alignment of two sequences please instead use our [pairwise sequence alignment tools](#).

Important note: This tool can align up to 4000 sequences or a maximum file size of 4 MB.

STEP 1 - Enter your input sequences

Enter or paste a set of

PROTEIN

sequences in any supported format:

Or, [upload a file:](#) Week2_Fluor...Sequences.fa [Use an example sequence](#) | [Clear sequence](#) | [See more example inputs](#)

- In the result, we can see how similar each fluorescence proteins.

Results for job clustalo-l20220907-021907-0571-83271431-p1m

Alignments
Result Summary
Guide Tree
Phylogenetic Tree
Results Viewers
Submission Details

Download Alignment File
Hide Colors

CLUSTAL O(1.2.4) multiple sequence alignment

```

RFP      MRSSKNV I KEFMRFKVRMEGTVNGHEFE I EGEGEGRPYEGHNTYKLVTKGGPLPFAWD I   60
YFP      MSKGEELFTGVVPI LVELDGDVNGHKFVS SGEEGDATYGKLTLLKLLCT-TGKLPVPIPT   59
GFP      MSKGEELFTGVVPI LVELDGDVNGHKFVS SGEEGDATYGKLTLLKLLCT-TGKLPVPIPT   59
CFP      MSKGEELFTGVVPI LVELDGDVNGHKFVS SGEEGDATYGKLTLLKLLCT-TGKLPVPIPT   59
* . . . . . . . . . . * . . . * * * * * . . . . . * * * * * * * * * *

RFP      LSPQFQVGSKVYVWHPADI --PDYKKL SFPEGFKWERVMNFEDGGVVTYTDQSSLQDGC F   118
YFP      LVTTFGYGLQCFARYPDHMKRHDFK SAMPEGVVQERT IFFKDDGNKYKTRAEVKFEGDTL   119
GFP      LVTTFSVGVQCFSRYPDHMKQHDFK SAMPEGVVQERT IFFKDDGNKYKTRAEVKFEGDTL   119
CFP      LVTTFSVGVQCFSRYPDHMKQHDFK SAMPEGVVQERT IFFKDDGNKYKTRAEVKFEGDTL   119
* * * * * . . . . . * * * * * * * * * * * * * * * * * * * * * *

RFP      IYKVKFI GVNFPSDGPMQKKTMTGWEASTERLYPRDGVKGEIHKALK----LKGGHYL   174
YFP      VNR I ELKGI DFKEDGNILGHKL-EVNYNSHNIVYIMADKQKNGIKVNFKIRHNI EDGVSQ L   178
GFP      VNR I ELKGI DFKEDGNILGHKL-EVNYNSHNIVYIMADKQKNGIKVNFKIRHNI EDGVSQ L   178
CFP      VNR I ELKGI DFKEDGNILGHKL-EVNYNSHNIVYIMADKQKNGIKVNFKIRHNI EDGVSQ L   178
. . . . . * * * * * * * * * * . . . . . * * * * * . . . . . *

RFP      VEF--KS IYMAKKPVQLPGVYVYVDSKLDI TSHNEOYT IVEQYERTEGRHHLFL----- 225
YFP      ADHYQQNTP IGDGPVLLPDNHVLSYQSALF-----KDPNEKRDHMVLLFVLTAA   227
GFP      ADHYQQNTP IGDGPVLLPDNHVLSYQSALS-----KDPNEKRDHMVLLFVLTAA   227
CFP      ADHYQQNTP IGDGPVLLPDNHVLSYQSALS-----KDPNEKRDHMVLLFVLTAA   227
. . . . . * * * * * * * * * * . . . . . * * * * * *

RFP      ----- 225
YFP      G I T G M D E L Y K 238
GFP      G I T G M D E L Y K 238
CFP      G I T G M D E L Y K 238
  
```

4. Practice how to find protein domain

- Find sequence and domains of TP53 using "UniProt"
 - Search "Human P53" and select 1st one (P04637).

UniProt Tools ▾ SPARQL UniProtKB ▾ Human P53 Advanced | List Search

Status

- Reviewed (Swiss-Prot) (1,018)
- Unreviewed (TrEMBL) (25,846)

Popular organisms

- Human (1,348)
- Zebrafish (298)
- Mouse (142)
- Rat (22)
- Fruit fly (11)

UniProtKB 26,864 results

BLAST Align Map IDs Download Add View: Cards Table Customize columns Share ▾

Entry	Entry Name	Protein Names	Gene Names	Organism	Length
<input type="checkbox"/> P04637	P53_HUMAN	Cellular tumor antigen p53[...]	TP53, P53	Homo sapiens (Human)	393 AA
<input type="checkbox"/> P02340	P53_MOUSE	Cellular tumor antigen p53[...]	Tp53, P53, Trp53	Mus musculus (Mouse)	390 AA
<input type="checkbox"/> P07193	P53_XENLA	Cellular tumor antigen p53[...]	tp53	Xenopus laevis (African clawed frog)	363 AA

- Get the protein sequence of TP53.

```
>sp|P04637|P53_HUMAN Cellular tumor antigen p53 OS=Homo sapiens OX=9606 GN=TP53 PE=1 SV=4
MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGP
DEAPRMPEAAPPVAPAPAAPTPAAPAPAPSWPLSSSVPSQKTYQGSYGFRLGFLHSGTAK
SVTCTYSPALNKMFCQLAKTQPVQLWVDSTPPPGTRVRAMA IYKQSQHMTTEVYRRCPHHE
RCSDSDGLAPPQHLI RVEGNLRVEYLDLRNTFRHSVYVPYEPPEVYSDCTT IHYNYMCNS
SCMGGMNRRPILTI I TLEDSSGNLLGRNSFEVRYCACPGRDRRT EEENLRKKGEPHHELP
PGSTKRALPNNTSSSPQPKKKPLDGEVFTLQ I RGRERFEMFRELNEALELKDAQAGKEPG
GSRAHSSHLKSKKGQSTSRHKKLMFKTEGPDSD
```

- Click "Family & Domains" to see the domain information of TP53.

Function

Names & Taxonomy

Subcellular Location

Disease & Variants

PTM/Processing

Expression

Interaction

Structure

Family & Domains

Sequence & Isoforms

Similar Proteins

Family & Domains¹

Features

Showing features for region¹, motif¹, compositional bias¹.

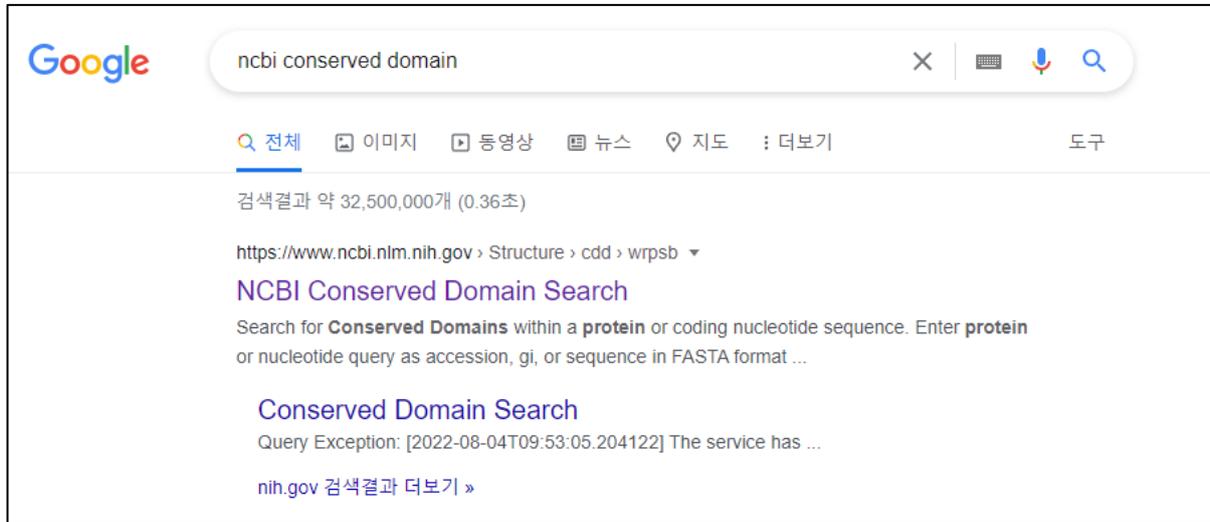
1 50 100 150 200 250 300 350 393

TYPE	ID	POSITION(S)	DESCRIPTION	BLAST
Region		1-44	Transcription activation (acidic)	BLAST Add
Region		1-83	Interaction with HRMT1L2 1 Publication	BLAST Add
Region		1-320	Interaction with CCAR2 1 Publication	BLAST Add
Motif		17-25	TADI	BLAST Add
Motif		48-56	TADII	BLAST Add

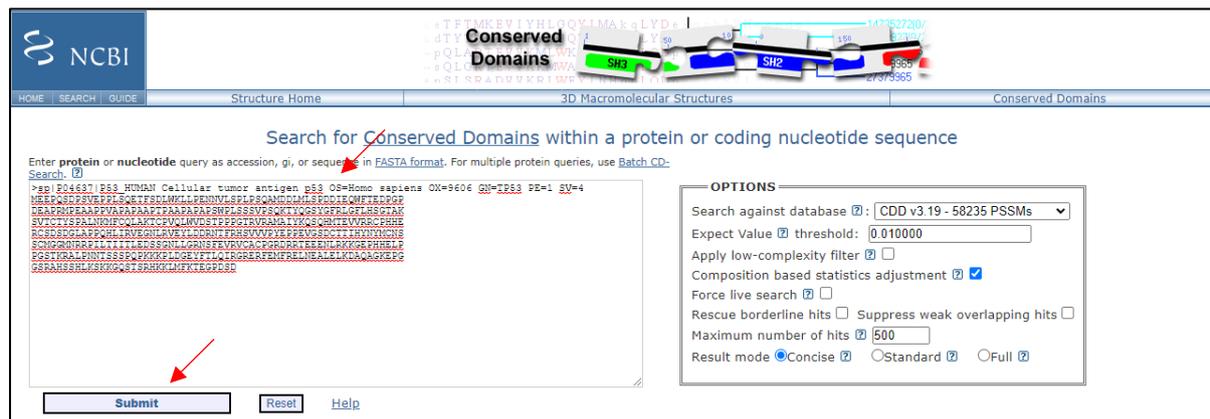
[Feedback](#) [Help](#)

- Find domains of TP53 using "NCBI Conserved Domains"

- Search "ncbi conserved domain" in google and access NCBI Conserved Domain Search.



- Copy and paste the sequence of TP53 in the box and click "Submit".



- In the result, we can see domains of TP53.

