

Molecular Biology Laboratory

Bioinformatics and Genomics Lab.

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

TA

Junseob Han, Hyunseok Song



Contact

Junseob Han

010.2113.6458

hljs502@gmail.com

Goal of This Week

1. To know how to get information of protein (UniProt, Pfam, RCSB PDB)
2. To know how to analyze the sequence data of protein (BLASTX, Clustal Omega)
3. To know how to find domain information of protein

Proteins

- Proteins account for the second largest proportion of the body after water and they have unique functions like body composition, hormone, immune response, and et cetera
- Proteins consist of 20 amino acids and these amino acids are connected with a peptide bond
- Many researchers are working hard to find proteins and their functions
 - Human Proteome Project (HPP) is in progress to discover all human proteins and their functions

19,750

PREDICTED PROTEINS ENCODED BY THE HUMAN GENOME
(neXtProt PE1+ PE2 + PE3 + PE4)



1,343

MISSING PROTEINS
(neXtProt PE2 + PE3 + PE4)



18,407

FOUND PROTEINS
(neXtProt PE1)



93.2%

PERCENT HUMAN PROTEOME DISCOVERED
(neXtProt PE1/(PE1 + PE2 + PE3 + PE4)) *100



HPP Progress to Date, HUPO, 2022

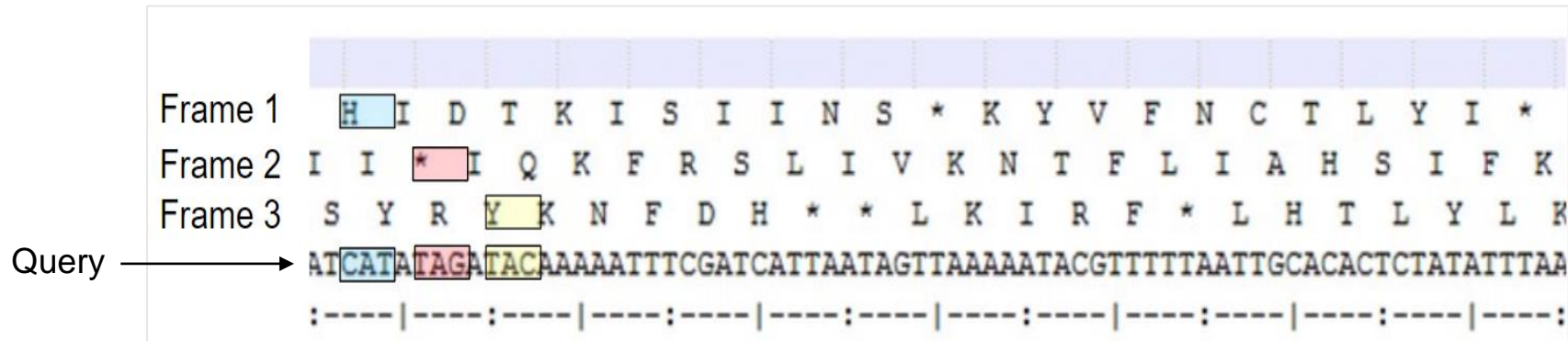
Protein Database

- There are three commonly used protein database
 - UniProt: The largest protein database (Verified: 568,002, Expected: 226,771,949)
 - Search "UniProt" in google or use hyperlink <https://www.uniprot.org/>
 - Pfam: The database which is based on UniProt and it is sorting proteins by protein family
 - Search "Pfam" in google or use hyperlink <https://pfam.xfam.org/>
 - RCSB PDB: The database which focuses on protein's structure
 - Search "RCSB PDB" in google or use hyperlink <https://www.rcsb.org/>



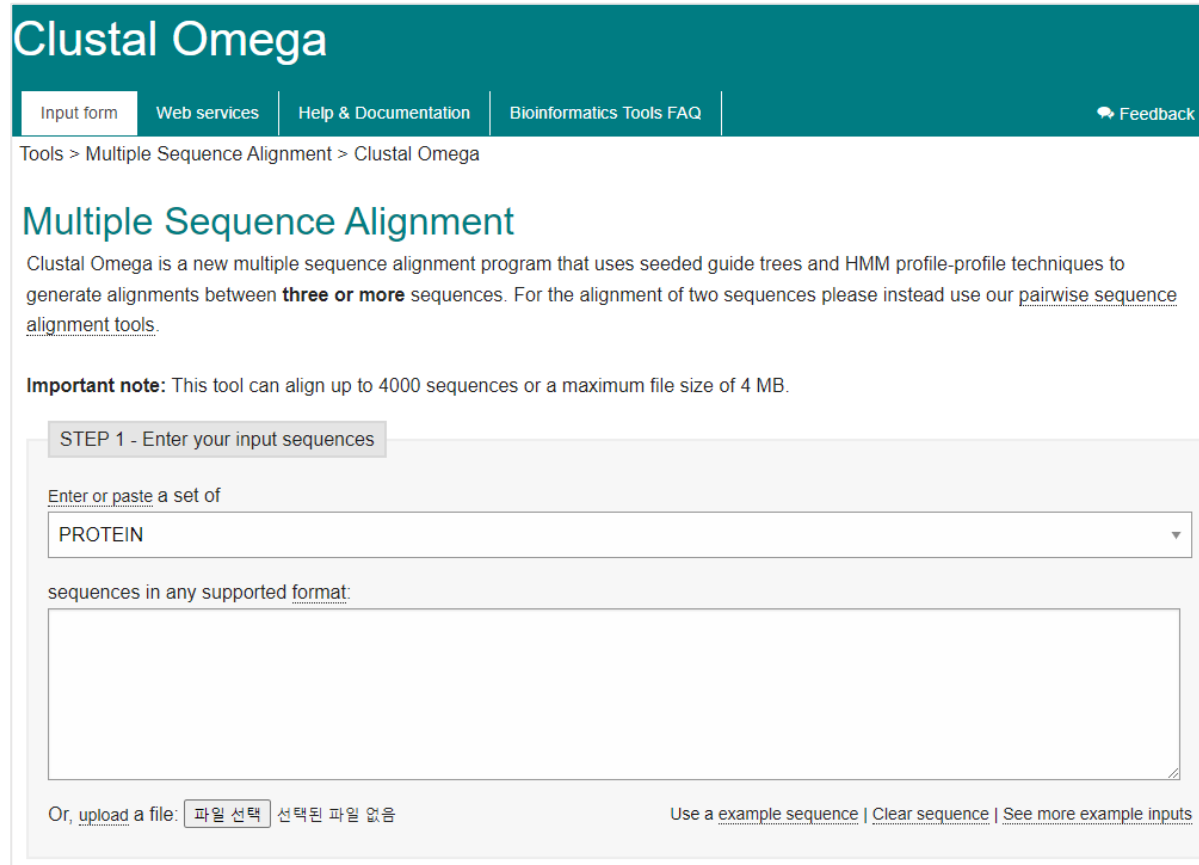
Protein Sequence Alignment - BLASTX

- "BLASTX" is one of the tools which is contained in "BLAST" program
- "BLASTX" is used for comparing nucleotide sequence (query) and amino acid sequence (subject)
- "BLASTX" changed nucleotide sequence to amino acid sequence with 6 frames and alignment to amino acid sequence
 - 3 "+" strand frame & 3 "-" strand frame



Protein Sequence Alignment - Clustal Omega

- "Clustal Omega" is used for multiple sequence alignment
- It is basically used for protein sequence alignment, but DNA and RNA sequence alignment is also possible



The screenshot shows the Clustal Omega web interface. At the top, there is a teal header with the text "Clustal Omega" and navigation links for "Input form", "Web services", "Help & Documentation", "Bioinformatics Tools FAQ", and "Feedback". Below the header, a breadcrumb trail reads "Tools > Multiple Sequence Alignment > Clustal Omega". The main heading is "Multiple Sequence Alignment". A paragraph explains that Clustal Omega is a new multiple sequence alignment program using seeded guide trees and HMM profile-profile techniques for three or more sequences, with a link to pairwise sequence alignment tools. An "Important note" states the tool can align up to 4000 sequences or a maximum file size of 4 MB. The "STEP 1 - Enter your input sequences" section includes a dropdown menu for "Enter or paste a set of" with "PROTEIN" selected, and a large text area for "sequences in any supported format:". At the bottom, there are links for "Or, upload a file:", "파일 선택", "선택된 파일 없음", "Use a example sequence", "Clear sequence", and "See more example inputs".

Protein Domain

- Domains are distinct functional structural units in proteins
- Each domain forms a compact 3D structure and they have a unique function
 - RNA binding domain, Zinc finger DNA binding domain, etc.
- Protein domain information can be found in "UniProt" database or "NCBI Conserved Domains"
 - "NCBI Conserved Domains" needs "FASTA" format sequence data of proteins
 - Search "NCBI conserved domain" in google or use the hyperlink

<https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi>

NCBI

Conserved Domains

HOME SEARCH GUIDE Structure Home 3D Macromolecular Structures Conserved Domains

Search for Conserved Domains within a protein or coding nucleotide sequence

Enter **protein** or **nucleotide** query as accession, gi, or sequence in [FASTA format](#). For multiple protein queries, use [Batch CD-Search](#).

[Help](#)

OPTIONS

Search against database

Expect Value threshold:

Apply low-complexity filter

Composition based statistics adjustment

Force live search

Rescue borderline hits Suppress weak overlapping hits

Maximum number of hits

Result mode Concise Standard Full

Practical Exercise

1. Practice how to use UniProt database
 - Find information on GFP and get a amino acid sequence
2. Practice how to use Clustal Omega
 - Try multiple sequence alignment (MSA) of GFP, CFP, YFP, and RFP
3. Practice how to find protein domain
 - Find sequence and domains of TP53 using "UniProt"
 - Find domains of TP53 using "NCBI Conserved Domain Search"