Molecular Biology Laboratory

Bioinformatics and Genomics Lab.

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

TA
Jihoon Kim, Junseob Han

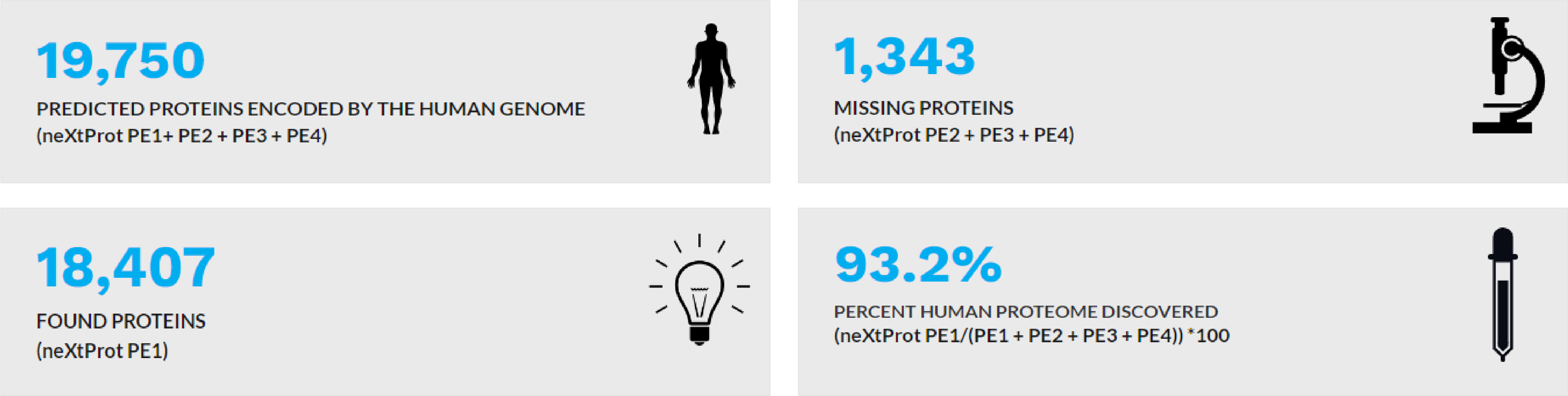
Contact
Jihoon Kim
010.6628.1102
cjscjs1102@hanyang.ac.kr

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

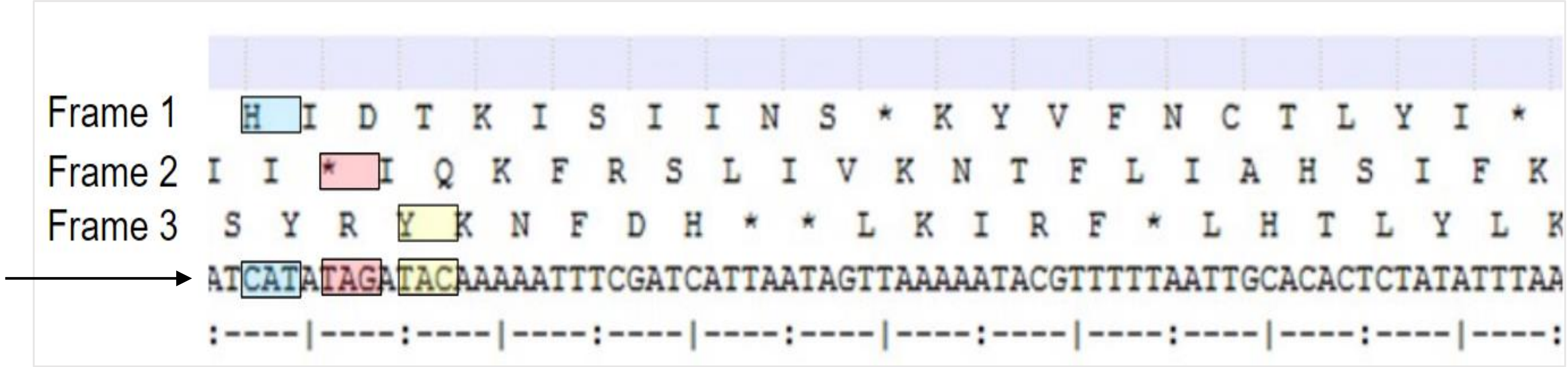
HPP Progress to Date, HUPO, 2022

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

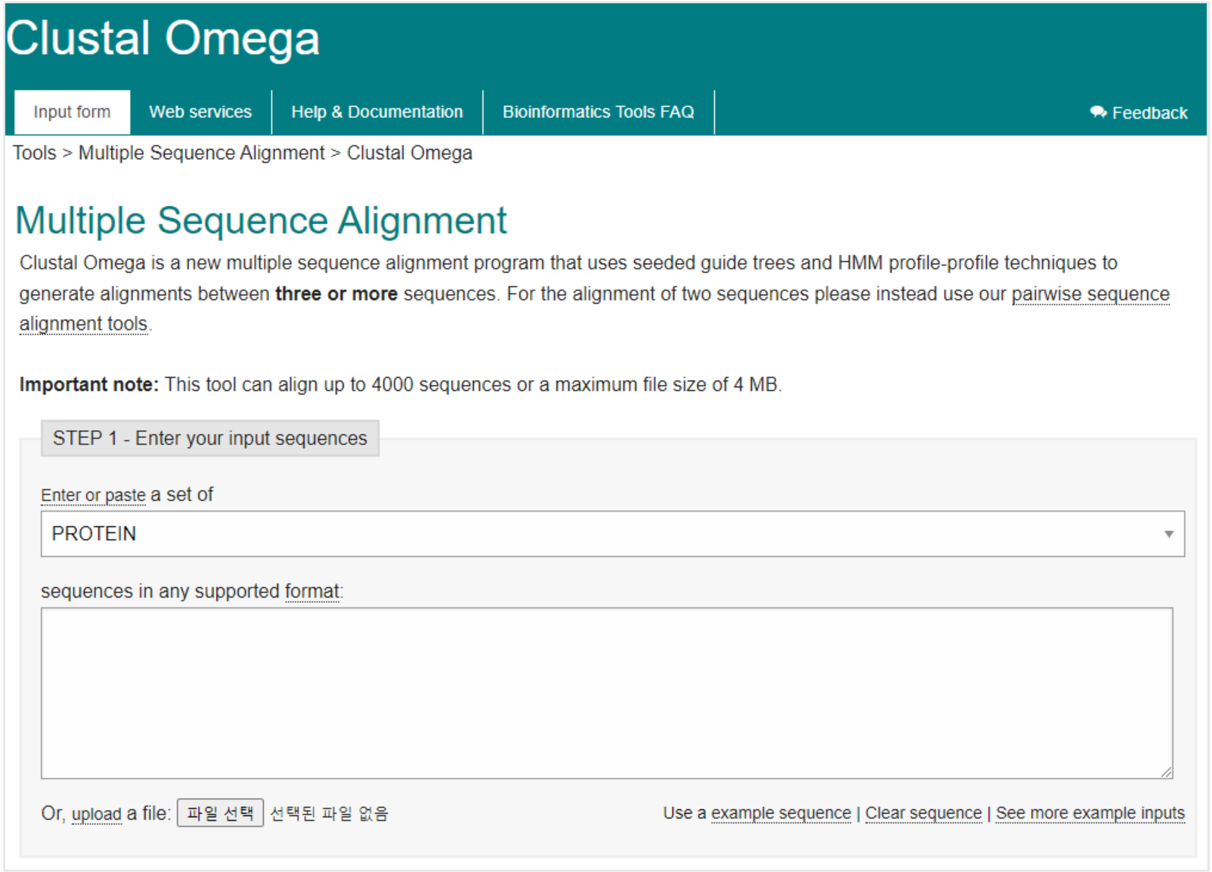


```
Frame 1   H  I  D  T  K  I  S  I  I  N  S  *  K  Y  V  F  N  C  T  L  Y  I  *
Frame 2 I  I  *  I  Q  K  F  R  S  L  I  V  K  N  T  F  L  I  A  H  S  I  F  K
Frame 3  S  Y  R  Y  K  N  F  D  H  *  *  L  K  I  R  F  *  L  H  T  L  Y  L  K
        ATCATATAGATACAAAAATTTCGATCATTAATAGTTAAAAATACGTTTTTAATTGCACACTCTATATTTAA
        :----|----:----|----:----|----:----|----:----|----:----|----:----|----:
```

Query

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



Clustal Omega

Tools > Multiple Sequence Alignment > Clustal Omega

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

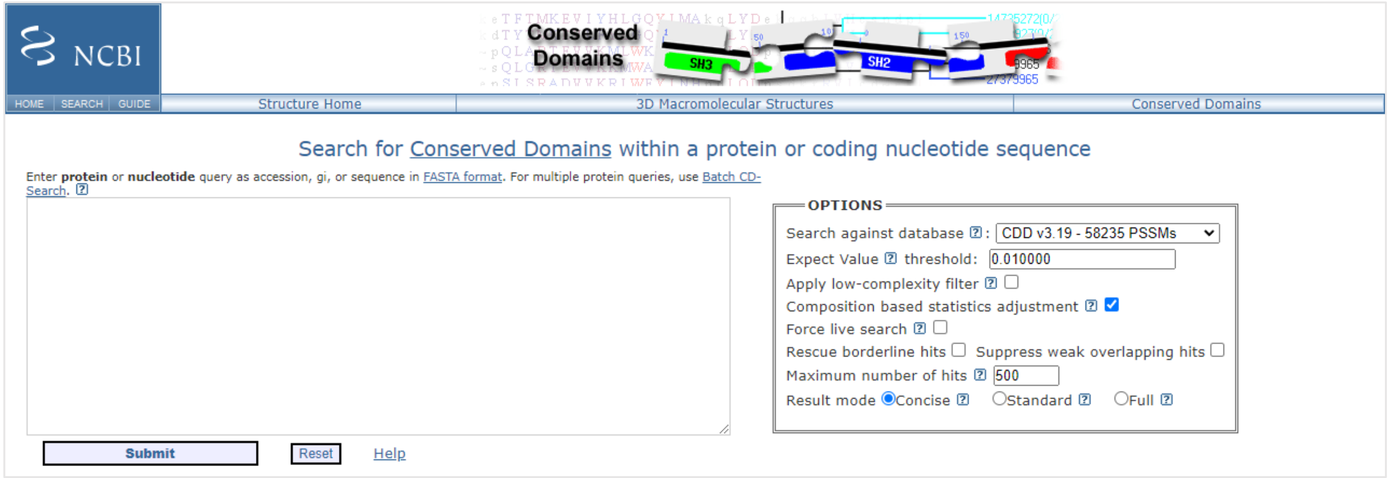
PROTEIN

sequences in any supported format:

Or, upload a file: 파일 선택 선택된 파일 없음

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



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.

OPTIONS

Search against database: CDD v3.19 - 58235 PSSMs

Expect Value threshold: 0.010000

Apply low-complexity filter ☐

Composition based statistics adjustment ☑

Force live search ☐

Rescue borderline hits ☐ Suppress weak overlapping hits ☐

Maximum number of hits 500

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"