# Bioinformatics

## CS300

## Prediction and Modeling Protein Structure

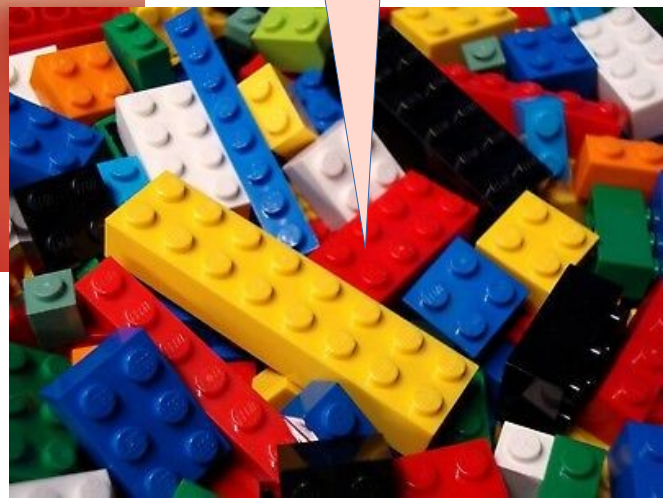
**Spring 2021**

**Oliver BONHAM-CARTER**

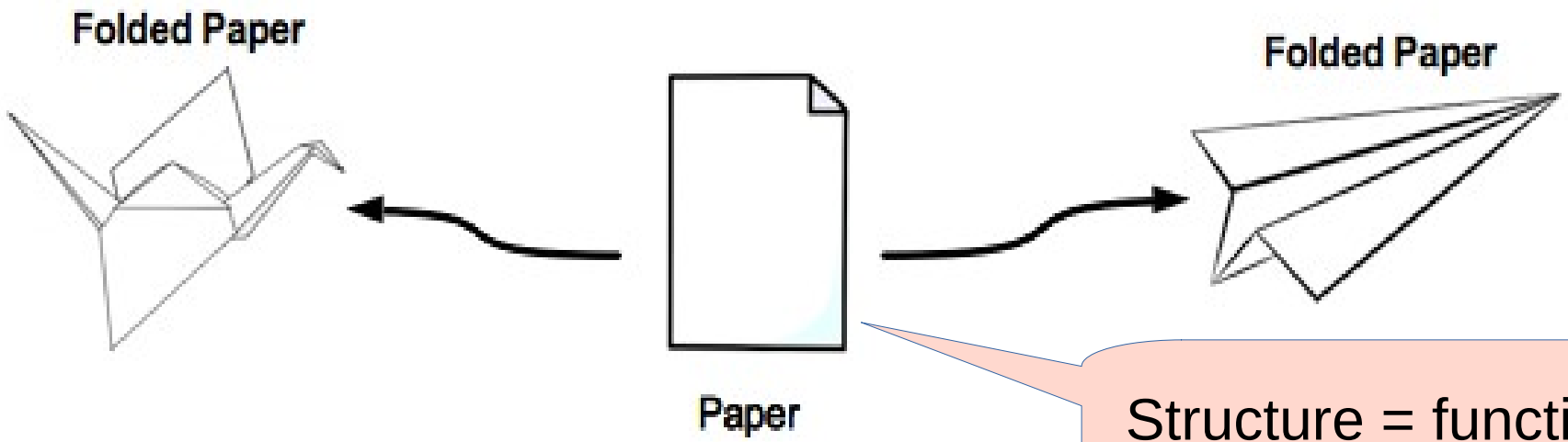
# Properties From Combining Pieces

A cool living room made from Lego pieces!

From these pieces?

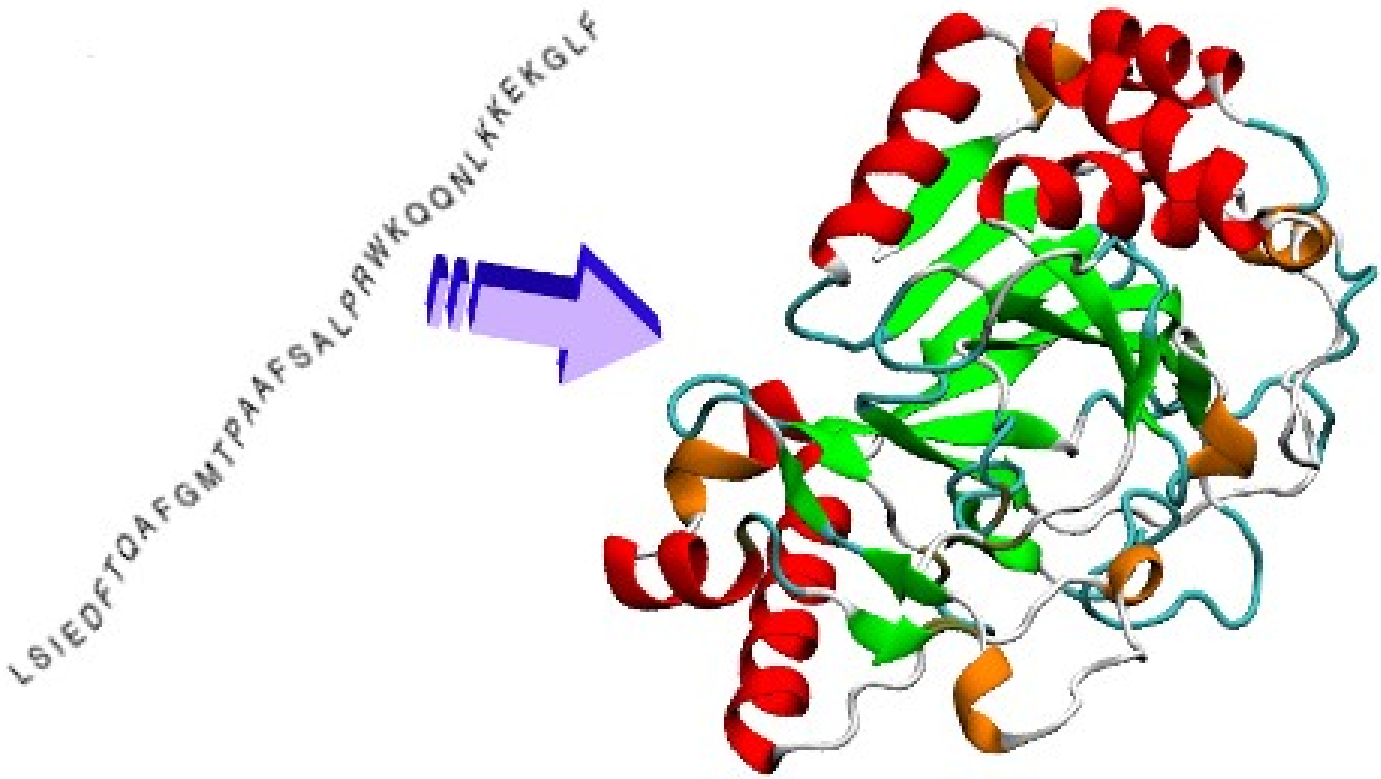


# Properties From Folding



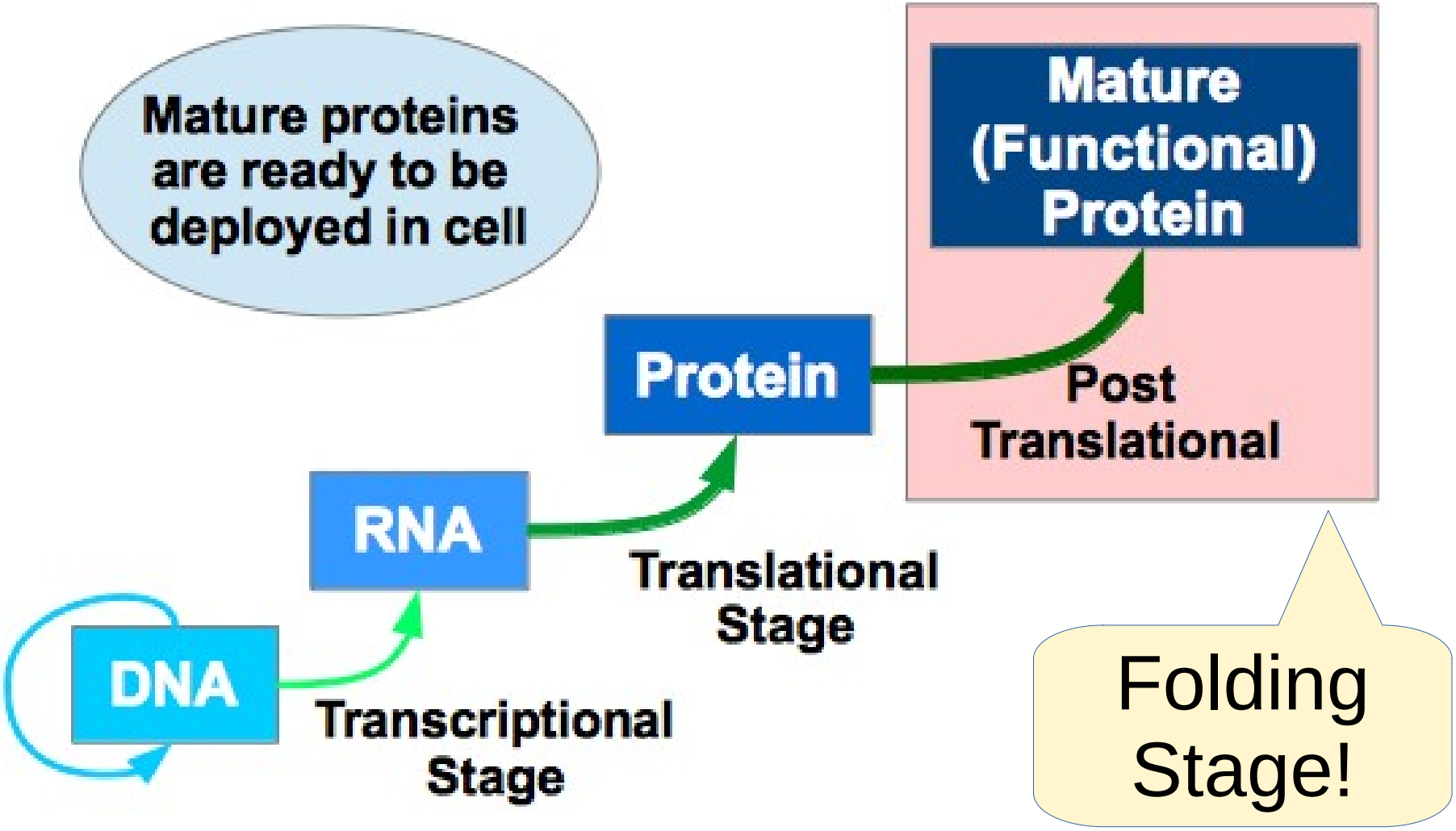

Structure = function

# Protein Folding

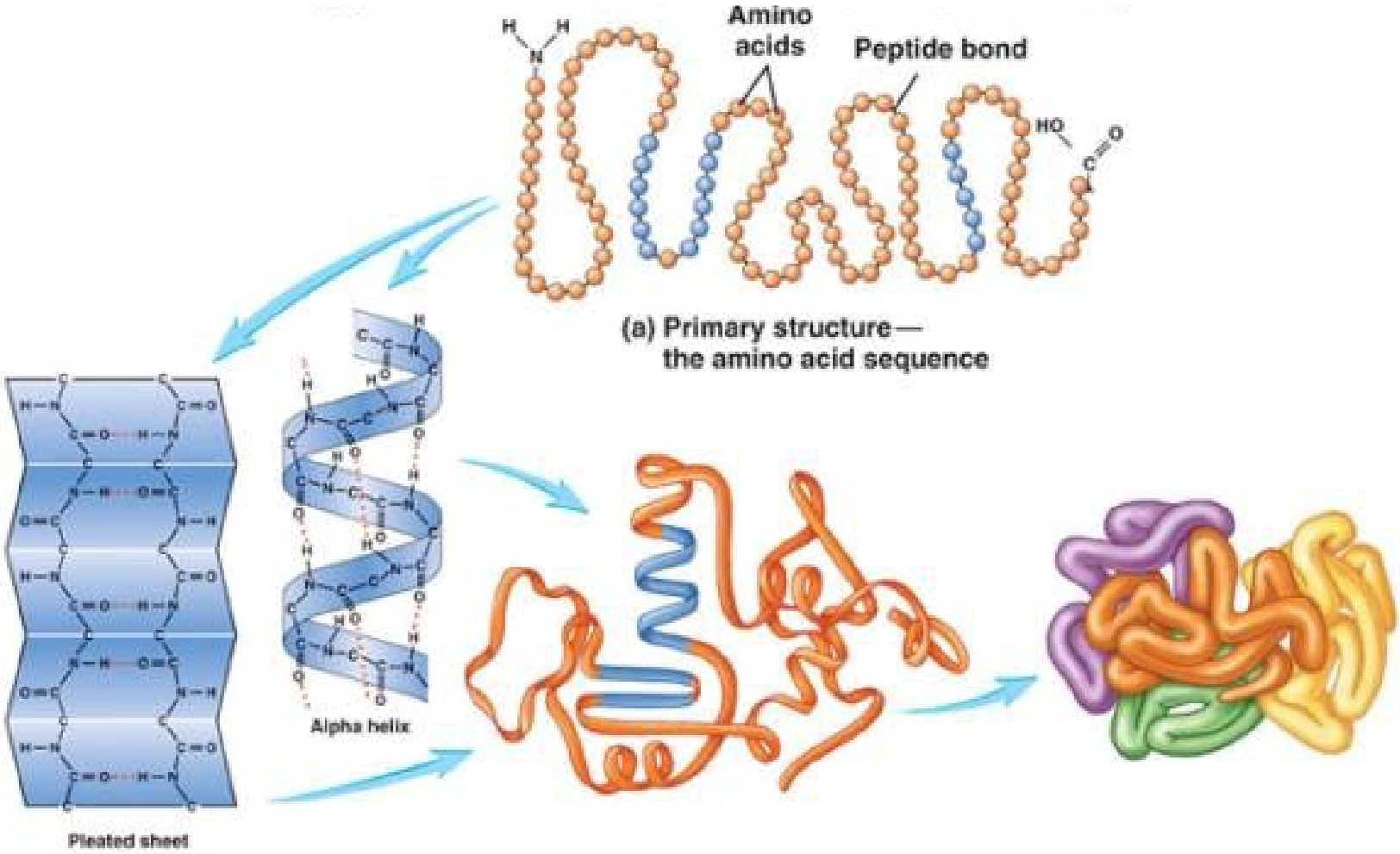



- A protein sequence is a linear chain of amino acids produced by ribosomes during translation
- A structure from folding, 3D state based on properties of amino acids and structure

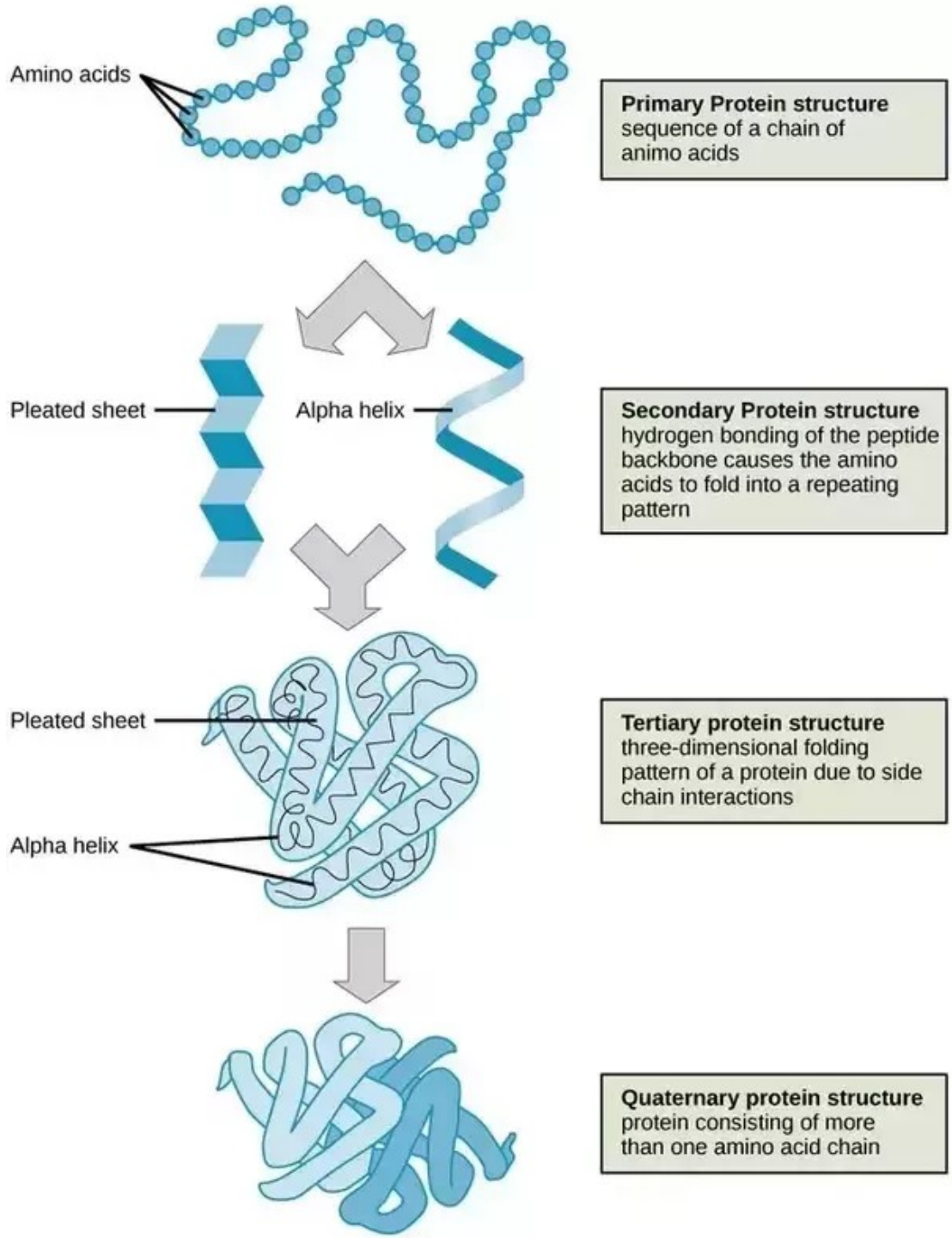
# Protein Folding and the Central Dogma of Biology



# Protein Folding: Four Stages

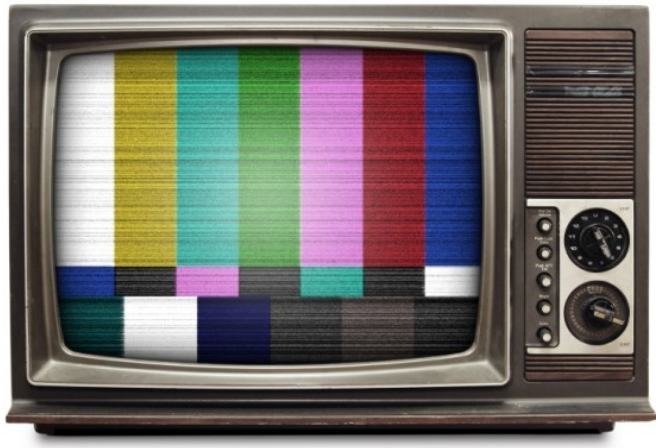


# Protein Folding: Another View



# Supporting Videos

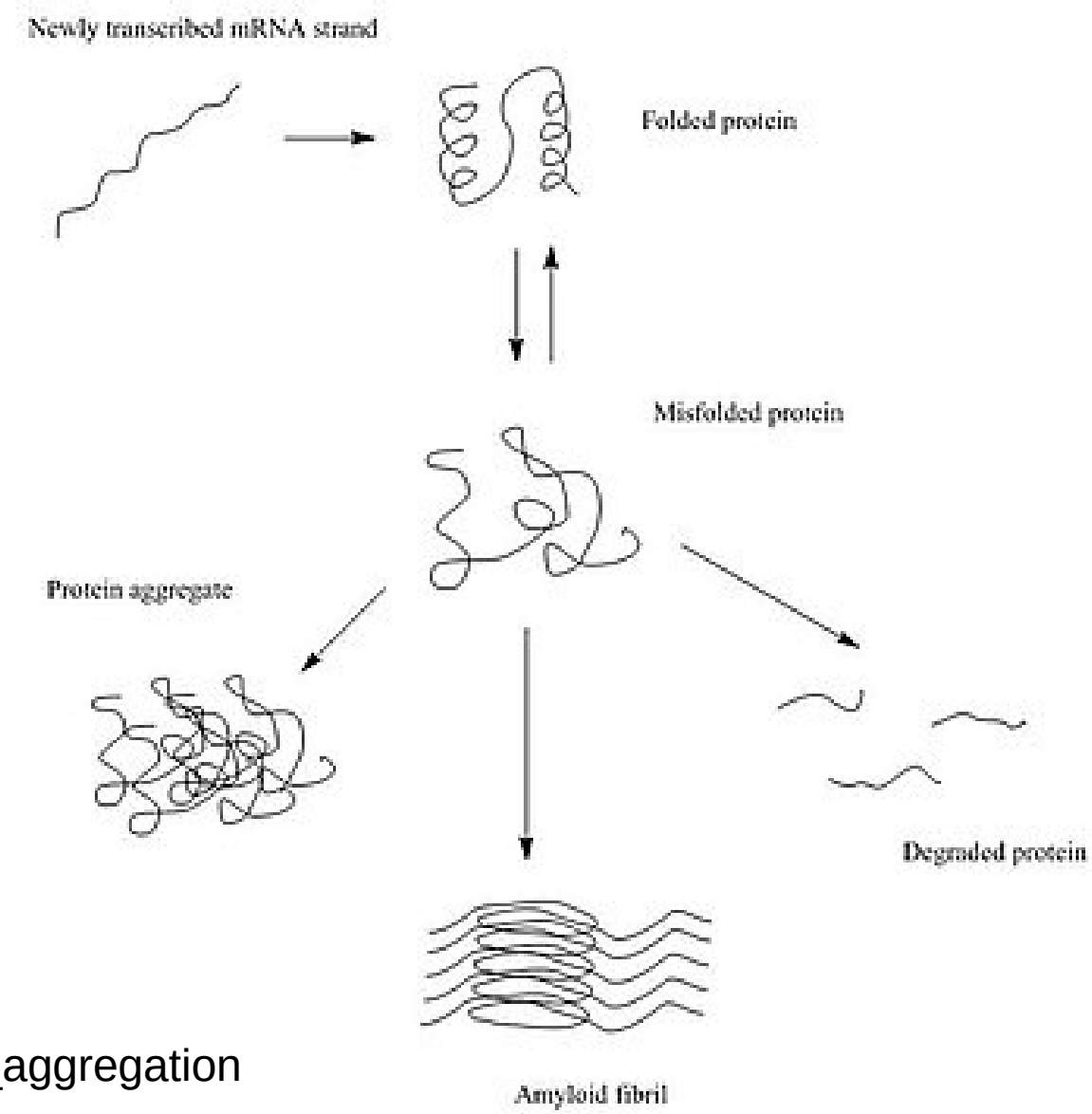
- Protein Folding (3 mins)
  - https://www.youtube.com/watch?v=yZ2aY5lxEGE
- What is a protein? (3D shape and function, 3 mins)
  - https://www.youtube.com/watch?v=qBRFIMcxZNM
- Protein folding simulation (3 mins)
  - https://www.youtube.com/watch?v=meNEUTn9Atg



# Protein Folding - Applications

- **Protein must fold *"correctly"* to function *"correctly"***

- Misfolded proteins
  - Accumulation (*clumping*) – Huntington's and Parkinson's disease
  - Tagged for degradation – emphysema, cystic fibrosis
    - Article: Pharmaceutical chaperones – therapies to fold mutated proteins to render them functional (placed in stabilized state)

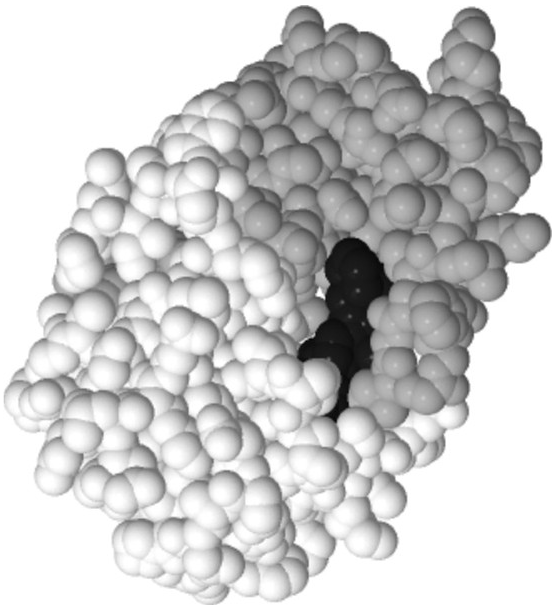


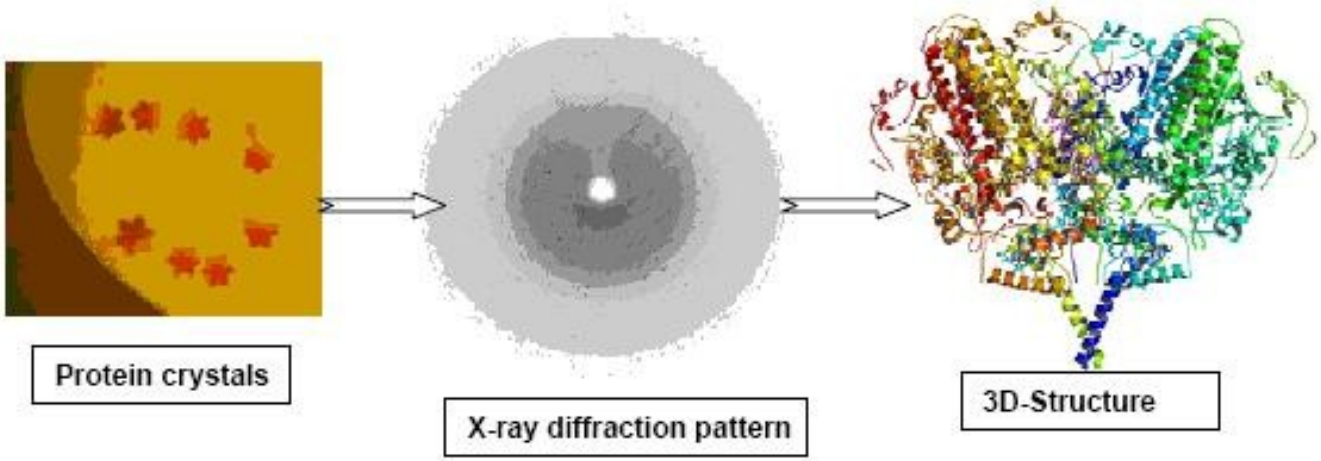

Ref: https://en.wikipedia.org/wiki/Protein_aggregation

# Protein Folding - Applications

- Development of Antimicrobial Drugs: help to...
  - Be effective against the disease-causing agents
  - Be selectively toxic
  - kill or inhibit the microbe without harming the host

- Drugs Structures
  - Study 3-D structure (and function) of viral proteins
  - Design drugs to fit (dock to) to proteins and block functions
- Laboratory – challenging to predict 3-D structure





# Genomics & Computational Structural Biology

**Genomics** (study)

- Determines the ordered sequence of nucleotides in a genome
- Determines/ assigns (predicted) functions to regions of nucleotides by annotations

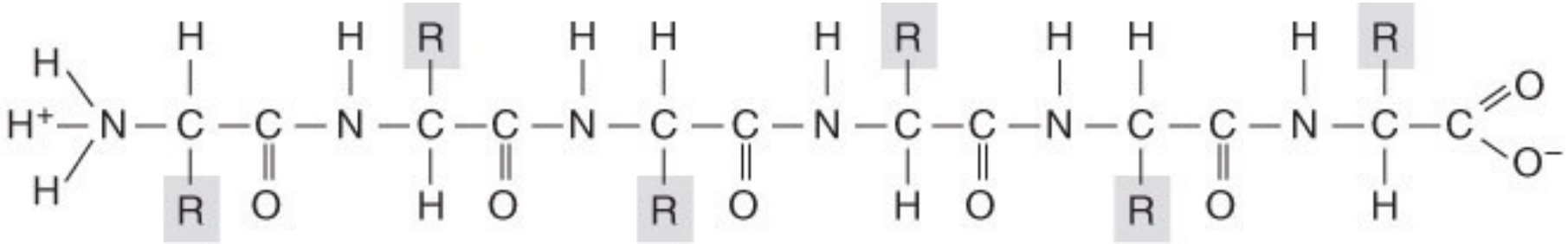
**Computational Structural Biology** (study)

- Takes predicted gene sequence for translation into primary amino acid sequence
- Predicts the 3-D protein structure based on the (primary) amino acid sequence
- Note: this step is very difficult because the number of possible outcomes to process and consider is enormous
- The study of structural rules and their contribution to the final mature protein.

# Structural Rules for Protein Folding

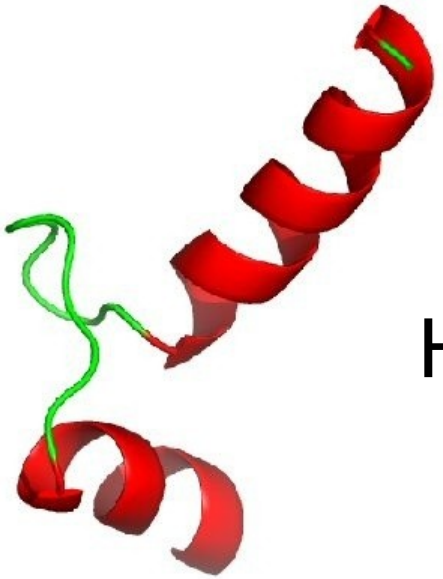
- Linus Pauling – Studied the limitations on protein folding
  - Nature of chemical bonds between amino acids
  - Bond angles
  - Rotation of atoms
  - Flexibility of side chains

Christian B. Anfisen – Studied the influence of thermodynamics of cellular environment



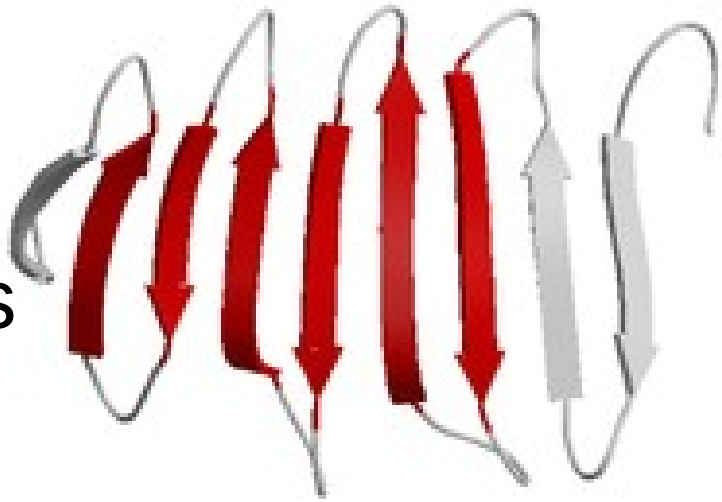
**(A) Primary (1°) structure**

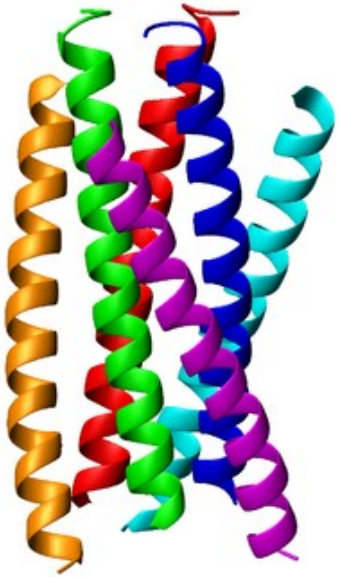
# Parts of Protein (Structures)



Helices

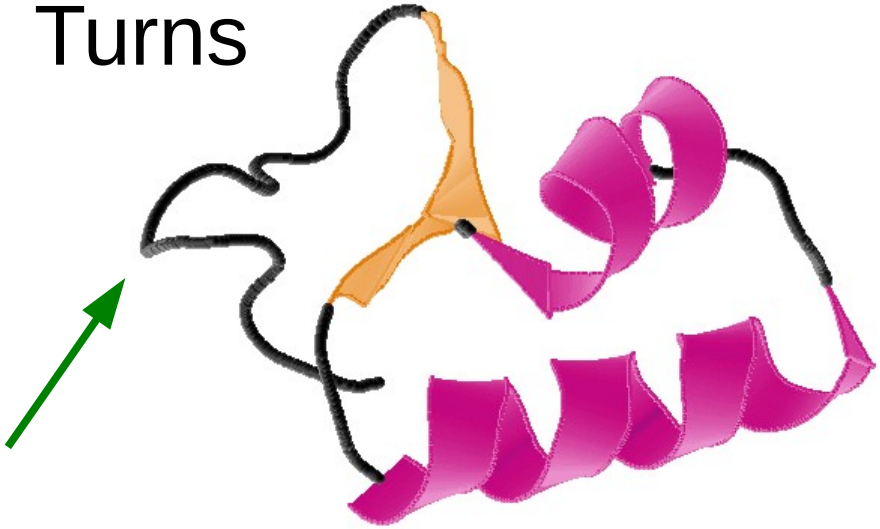
Sheets





Coils

Turns



# Protein Folding: An Idea of Structure

- **Garnier**: a text-based, command-line tool from EMBOSS
  - Input: protein sequence in fasta format
  - Output: a model of folding in text base
  - **Usage: garnier file.fasta**

```
          .   10    .   20    .   30    .   40    .   50
     MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQL
helix HH                 HHHHHHHHHHH        H
sheet   EEEE      EEEEE                      EEEEE
turns          T                    TTTT  TT
 coil       CCC CC     CCC              CC        CCCCC
          .   60    .   70
     EDGRTLSDYNIQKESVNHLVLRLRGG
helix            HHH HHH
sheet                    EEEEE
turns TTTT   TT              TTT
 coil     CCC  CC   C

#-----------------------------------
#
#  Residue totals: H: 20   E: 19   T: 16   C: 21
#         percent: H: 33.3 E: 31.7 T: 26.7 C: 35.0
#
#-----------------------------------
```

**H: Helices, E: Sheets**
**T: Turns, C: Coils**

Ref: http://emboss.open-bio.org/rel/rel6/apps/garnier.html

# Bring the Tool!

**Up Next!**

# Protein Folding: Quick Solutions

EMBOSS explorer

## garnier

Predict protein secondary structure using GOR method (read the manual)

Unshaded fields are optional and can safely be ignored. (hide optional fields)

This image is a link!

Input section

Select an input sequence. Use one of the following three fields:

1. To access a sequence from a database, enter the USA here:
2. To upload a sequence from your local computer, select it here: Choose File No file chosen
3. To enter the sequence data manually, type here:

```
>KX932045.1 Orchid fleck dichorhavirus isolate U1 phosphoprotein mRNA, complete cds
ATGTTCACTACCAAGGTAAATATGTACCCAGAGGTGCCCAGCTCATCCCAGGTGTCAGACGACATAGACA
ATGACACGCACATCGACGAGGTCGCTGCATTTGTGAGAAAGTGGTCGGCAGCCGGACTATCTCCCCCCAT
CACCCTTGCGAAGAACCTCAGAGCATGGATATCAAGCAACACCAGCCCTGGAAGCCCCCTAGTGTTGGAT
GACAGAATGCTGAGCCTTACAACCATGATATGGAACACAGCAGCAGAGCACTACACAATGATAGGCAAAT
CCCAGGTCAATCGTATGTCATCACTCATAGATCAGCTGGGGGAGATTTCCGGCCGCAAACCGCCGCAGGG
CCCAGCATTCGACATGCCTCCTCCCCCTCCTAAGAGAAAACATCCGGATTCACTAGACACTAATCCAATA
TTAGGCTTAATAGGTCAAGATTGGGACGACAATAAAGACAAGCACTGGAGAGAGAAACCAGCAGACAAGA
AGCTCCTCGTGCTCAACTGGGTGTTGCATGAGTATCTGGGGGTCCTCACAAAACCTGTCACCATCAAGTG
GATAACGGATAACCCCGCGTCTTTAGAGTTGGGAGCAGTGTCAGCTTATGCCCTGAAACATCAGGCCAGC
TTATCCGACTGCGACAAGGAAGCCCTCAGAGCGTTGGTGGTTCAAACAGTGAAAAACACCCCCAAAAGGC
CATGCCTGGACTAG
```

https://www.bioinformatics.nl/cgi-bin/emboss/garnier

# Garnier Output (text)

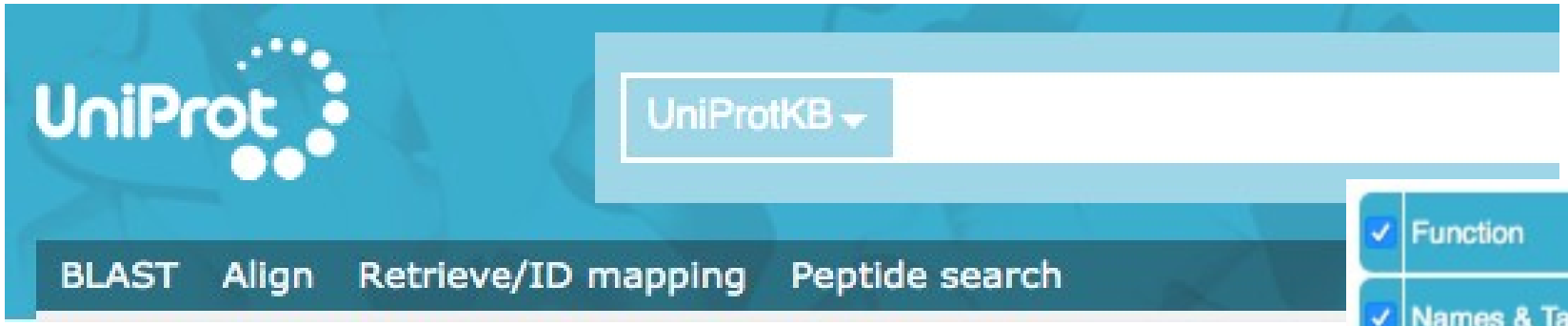
**H: Helices, E: Sheets**
**T: Turns, C: Coils**

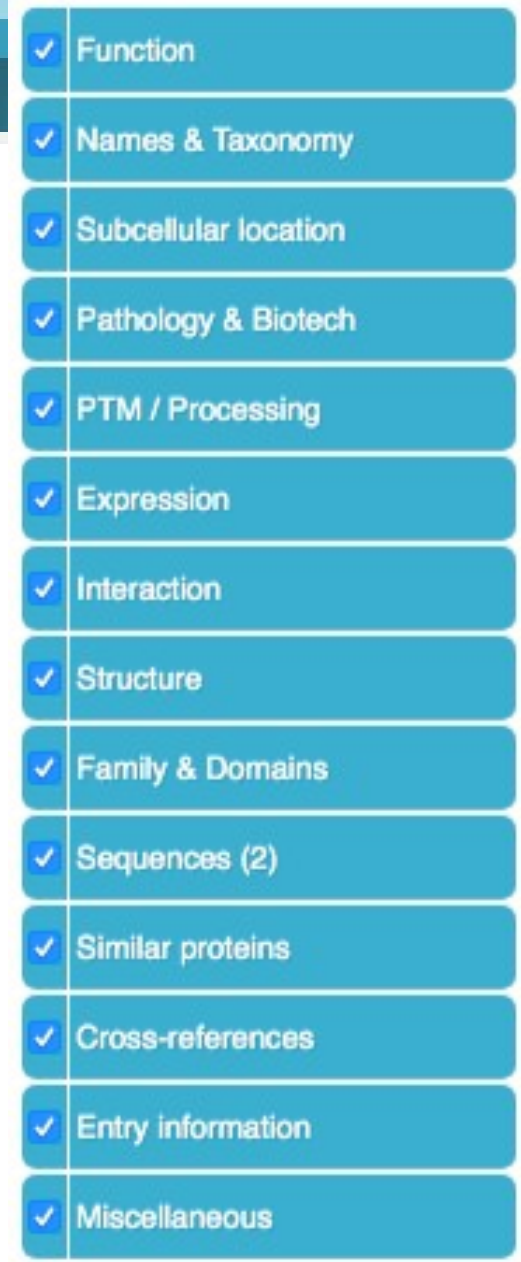
**OUTPUT FILE** outfile

```
#######################################
# Program: garnier
# Rundate: Mon 26 Apr 2021 05:31:19
# Commandline: garnier
#    -auto
#    -sequence /var/lib/emboss-explorer/output/626691/.sequence
#    -outfile outfile
#    -rformat2 tagseq
# Report_format: tagseq
# Report_file: outfile
#######################################

#=======================================
#
# Sequence: KX932045.1     from: 1   to: 714
# HitCount: 134
#
# DCH = 0, DCS = 0
#
#  Please cite:
#  Garnier, Osguthorpe and Robson (1978) J. Mol. Biol. 120:97-120
#
#
#=======================================

          .    10     .    20     .    30     .    40     .    50
      ATGTTCACTACCAAGGTAAATATGTACCCAGAGGTGCCCAGCTCATCCCA
helix              H   HHHHH
sheet         EEEE             EEEE                EEE
turns TTTTTTTT     TT      TTTT    TTTTTTTTTTTTTTTT   TT
 coil                 C
          .    60     .    70     .    80     .    90     .   100
      GGTGTCAGACGACATAGACAATGACACGCACATCGACGAGGTCGCTGCAT
helix            HHHHHHHHHHH
sheet      E                            E             EEE
turns TTTTT TTTTT           TTTTTTTTTT TTTTTTTTTTTTT
 coil
```

Not a Recent algorithm

https://www.bioinformatics.nl/cgi-bin/emboss/garnier

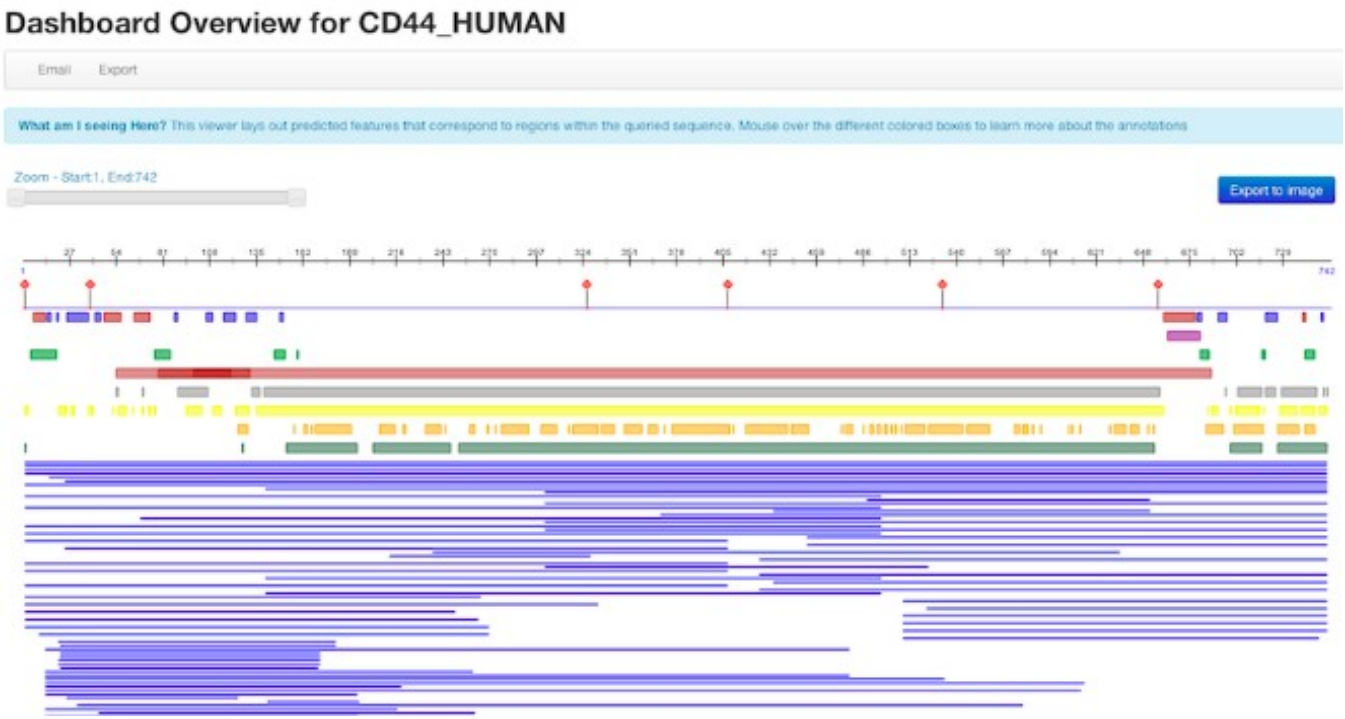



Protein Information: The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

- Some protein study databases require information from UniProt to find protein samples.
- https://www.uniprot.org/

# PREDICT PROTEIN



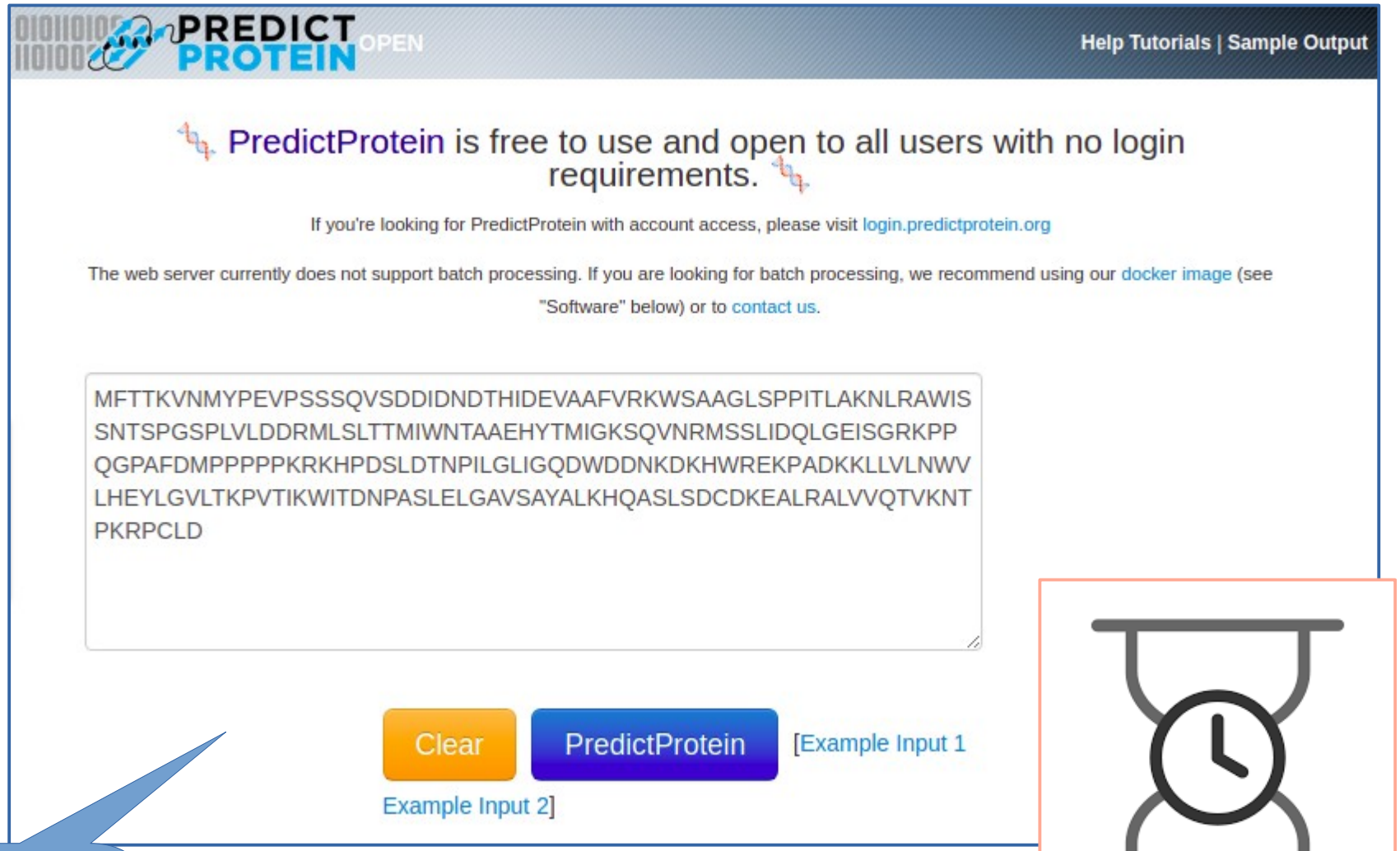

Structural Annotations of protein: prediction of protein function, e.g. assisting in the annotation of subcellular localization (LocTree, LocTree2, NLSpred), identifying protein-protein interaction sites (PPSites) and protein-DNA binding sites, and more.

- https://www.predictprotein.org/
- https://open.predictprotein.org/
- https://github.com/Rostlab/predictprotein-docker

# Bring the Tool!

**Up Next!**

# Protein Folding: Slower Solutions




This image is a link!

- https://www.predictprotein.org/

Can take some time ...

# Predict Protein output

## Input

```
>query
MFTTKVNMYP EVPSSSQVSD DIDNDTHIDE VAAFVRKWSA AGLSPPITLA
KNLRAWISSN TSPGSPLVLD DRMLSLTTMI WNTAAEHYTM IGKSQVNRMS
SLIDQLGEIS GRKPPQGPAF DMPPPPPKRK HPDSLDTNPI LGLIGQDWDD
NKDKHWREKP ADKKLLVLNW VLHEYLGVLT KPVTIKWITD NPASLELGAV
SAYALKHQAS LSDCDKEALR ALVVQTVKNT PKRPCLD
```

## Secondary Structure

### PROFsec summary

Protein can be classified as **mixed** given the following classes:

- 'all-alpha': %H > 45% AND %E < 5%
- 'all-beta': %H < 5% AND %E > 45%
- 'alpha-beta': %H > 30% AND %E > 20%
- 'mixed': all others

# Predict Protein output

**Predicted solvent accessibility composition (core/surface ratio) for your protein:**

Classes used:

- e: residues exposed with more than 16% of their surface
- b: all other residues.

| accessib type | b | e |
|---|---|---|
| % in protein | 40.08 | 59.92 |

**About your protein:**

| | |
|---|---|
| prot_nres | 237 |
| prot_nali | 4 |
| prot_nchn | 1 |
| prot_nfar | 3 |

**Residue composition for your protein:**

| | | | | |
|---|---|---|---|---|
| %A: 7.2 | %C: 0.8 | %D: 8.4 | %E: 3.4 | %F: 1.3 |
| %G: 4.2 | %H: 2.5 | %I: 5.1 | %K: 7.6 | %L: 10.6 |
| %M: 3.0 | %N: 4.6 | %P: 8.9 | %Q: 3.0 | %R: 3.8 |
| %S: 8.4 | %T: 6.3 | %V: 6.3 | %W: 3.0 | %Y: 1.7 |

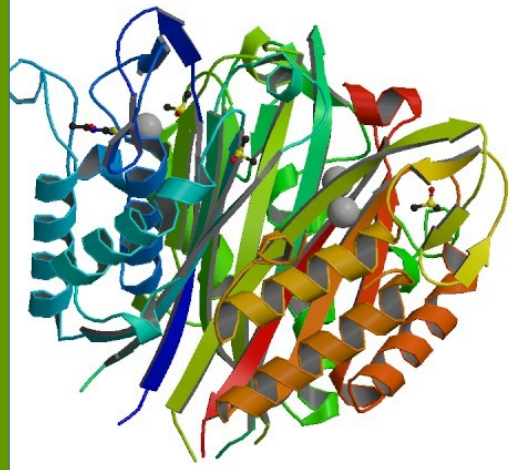
# Predict Protein Output

**Helix = H, Strand = S, Loop = L**

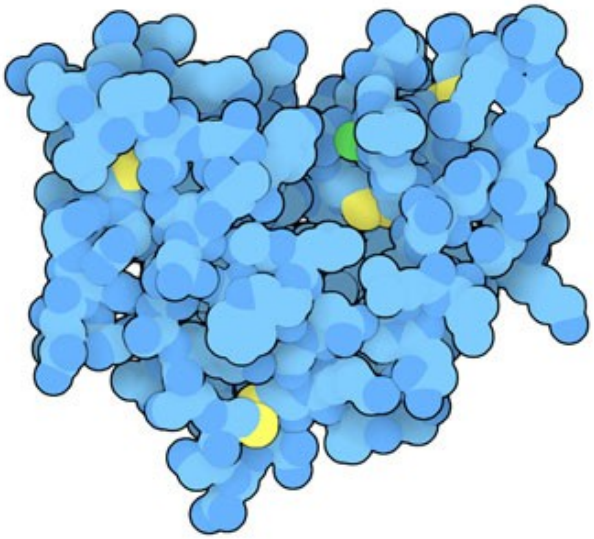
```
                ....,....1....,....2....,....3....,....4....,....5....,....6
AA              MFTTKVNMYPEVPSSSQVSDDIDNDTHIDEVAAFVRKWSAAGLSPPITLAKNLRAWISSN
OBS_sec         ??????????????????????????????????????????????????????????
PROF_sec                                        HHHHHHHHHHH      HHHHHHHHHHHHH
Rel_sec         954210246667861012100245456543678876411237777223778788888521
SUB_sec         LL......LLLLLL.........L.LLL..HHHHHH.....LLLL...HHHHHHHHHH..

O_3_acc         bbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbb
P_3_acc         eeeee  b  e eeeeebee beeee e  bbbbbeebee ebebbbebbe b e beee
Rel_acc         302131111010121210331132311022524553262413310102335484506332
SUB_acc         ..............................b.bbb..b.e..........eibie.b...


                ....,....7....,....8....,....9....,....10...,....11...,....12
AA              TSPGSPLVLDDRMLSLTTMIWNTAAEHYTMIGKSQVNRMSSLIDQLGEISGRKPPQGPAF
OBS_sec         ??????????????????????????????????????????????????????????
PROF_sec              EEEE     HHHHHH  HHHHHHHHHHHHHHHHHHHHHHHHHHH
Rel_sec         477861354011000000112002555543201003468888888876504777767666
SUB_sec         .LLLL..E................HHHH.........HHHHHHHHHHHH..LLLLLLLLL

O_3_acc         bbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbb
P_3_acc         eeee  bbbbe bb bbbbbbbbbb bbebb  bbbe bbebbebb eb eee ee  e
Rel_acc         115321347112102454693335212022322014237146840923024220121202
SUB_acc         ..e....bb......bbbbb...b...........b..b.ebbe.b....e.........
```

RCSB PDB
PROTEIN DATA BANK

Protein archives: This resource is powered by the Protein Data Bank archive-information about the 3D shapes of proteins, nucleic acids, and complex assemblies that helps students and researchers understand all aspects of biomedicine and agriculture, from protein synthesis to health and disease.

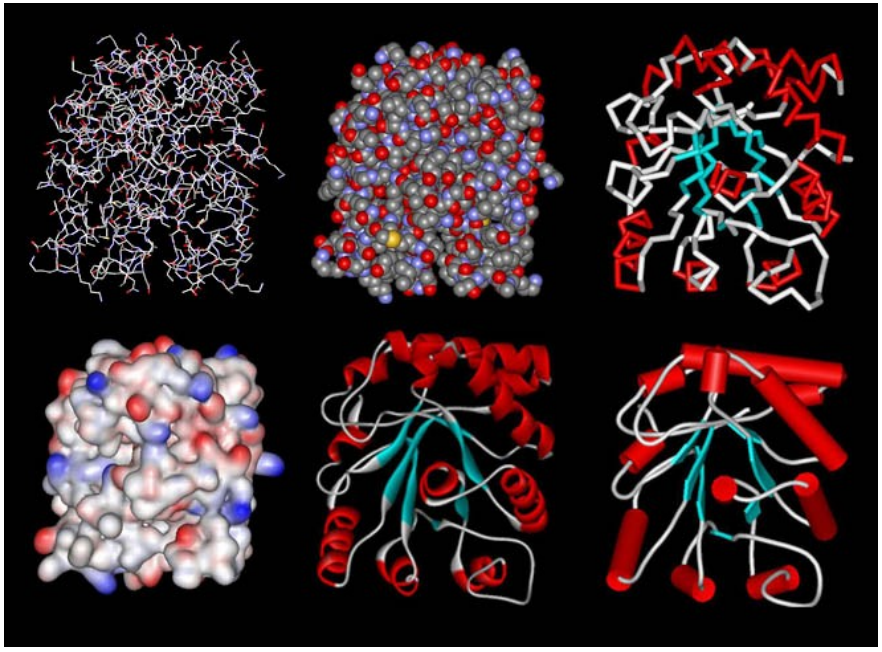


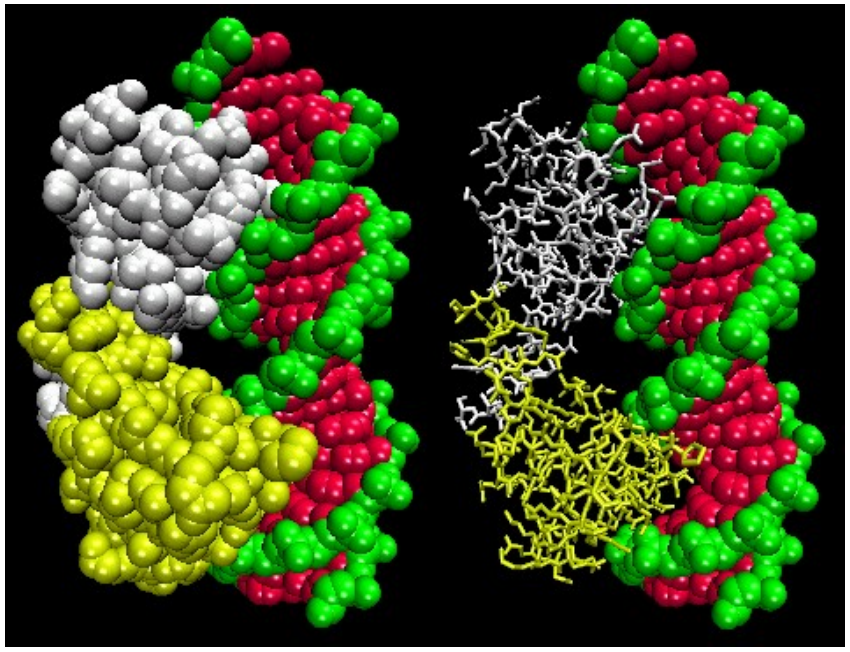
- http://www.rcsb.org/



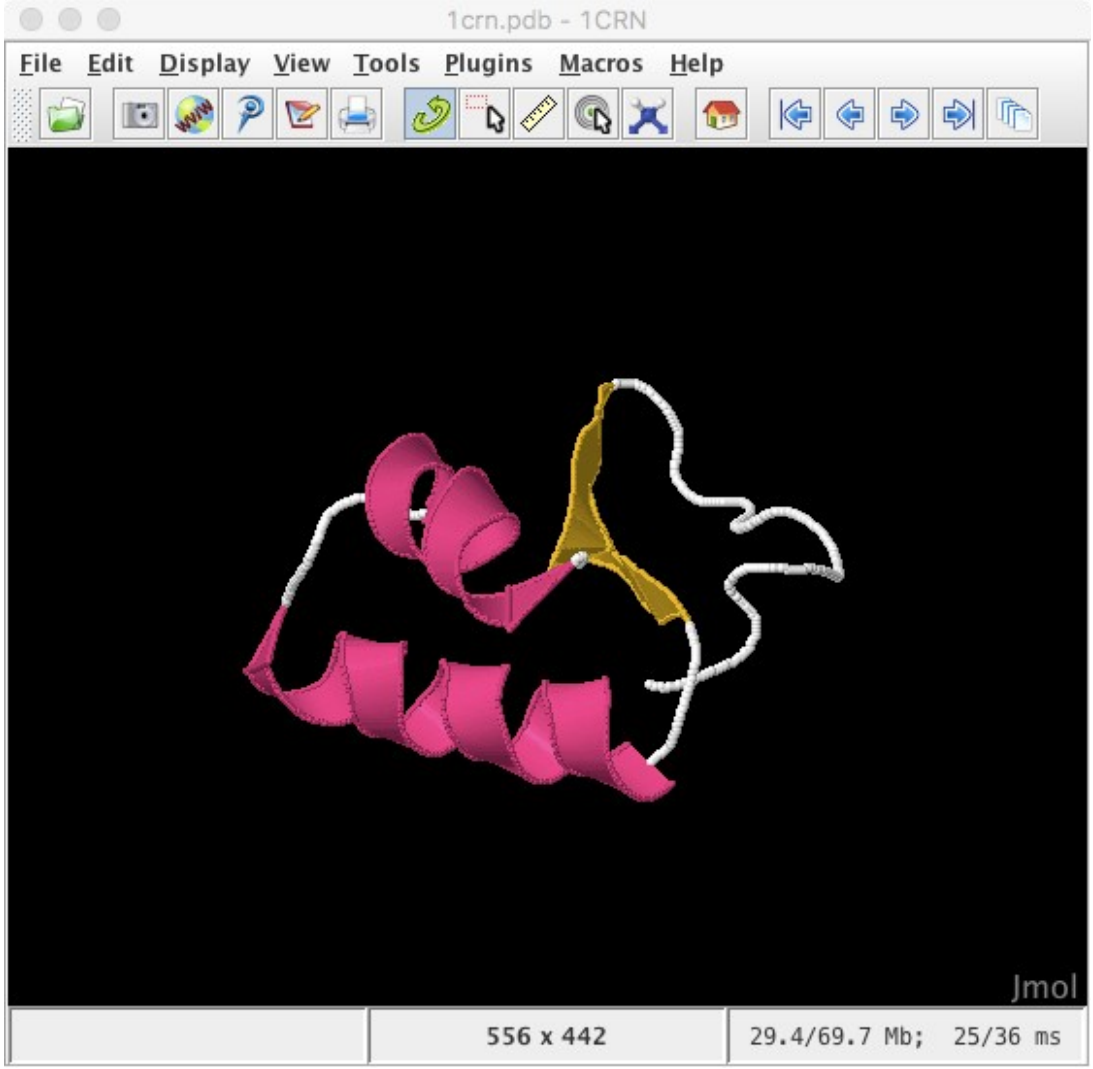
# Protein DataBase (PDB)

- Database for 3-D structural data of large biological molecules
- https://www.rcsb.org/
- Data is viewable using jmol (local use) and with online tools.





# Jmol: A (local) Graphical Viewer For Protein Sequences



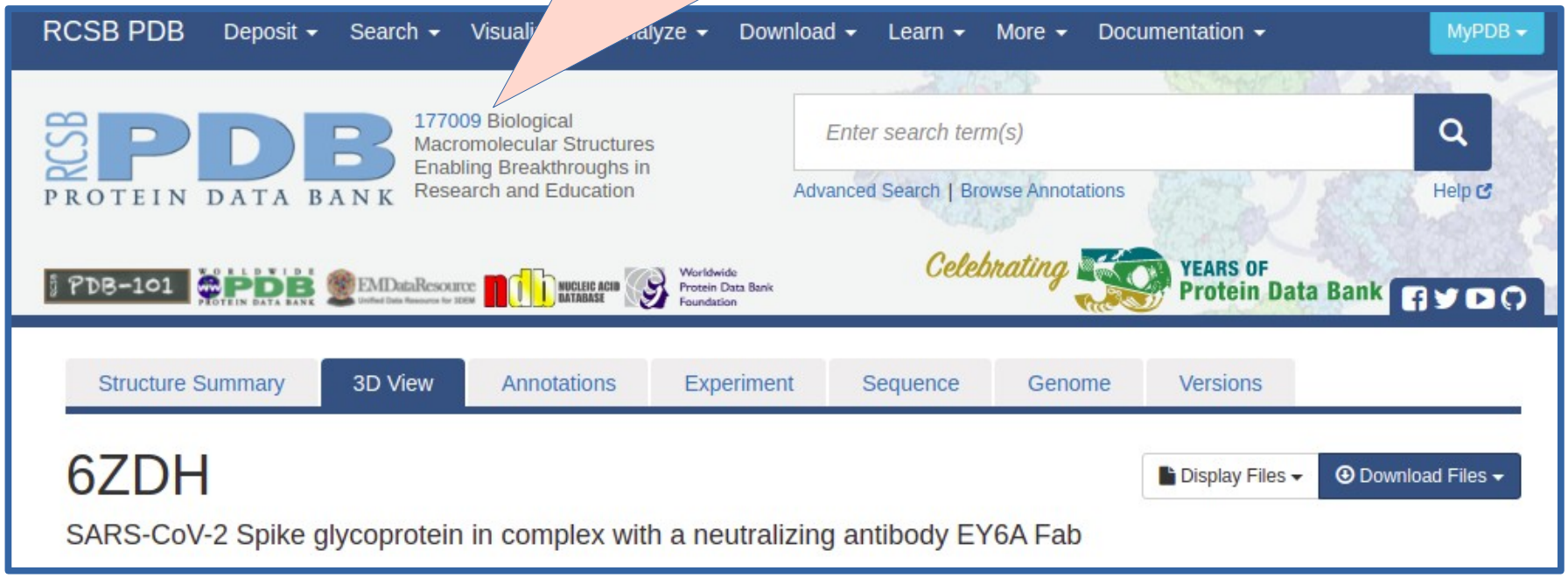

- Download:
  - http://jmol.sourceforge.net/
- Wiki:
  - http://wiki.jmol.org/index.php/Jmol_Application#Installing_Jmol_Application

# Bring the Tool!

**Up Next!**

# Protein Folding: Pre-Compiled Solutions

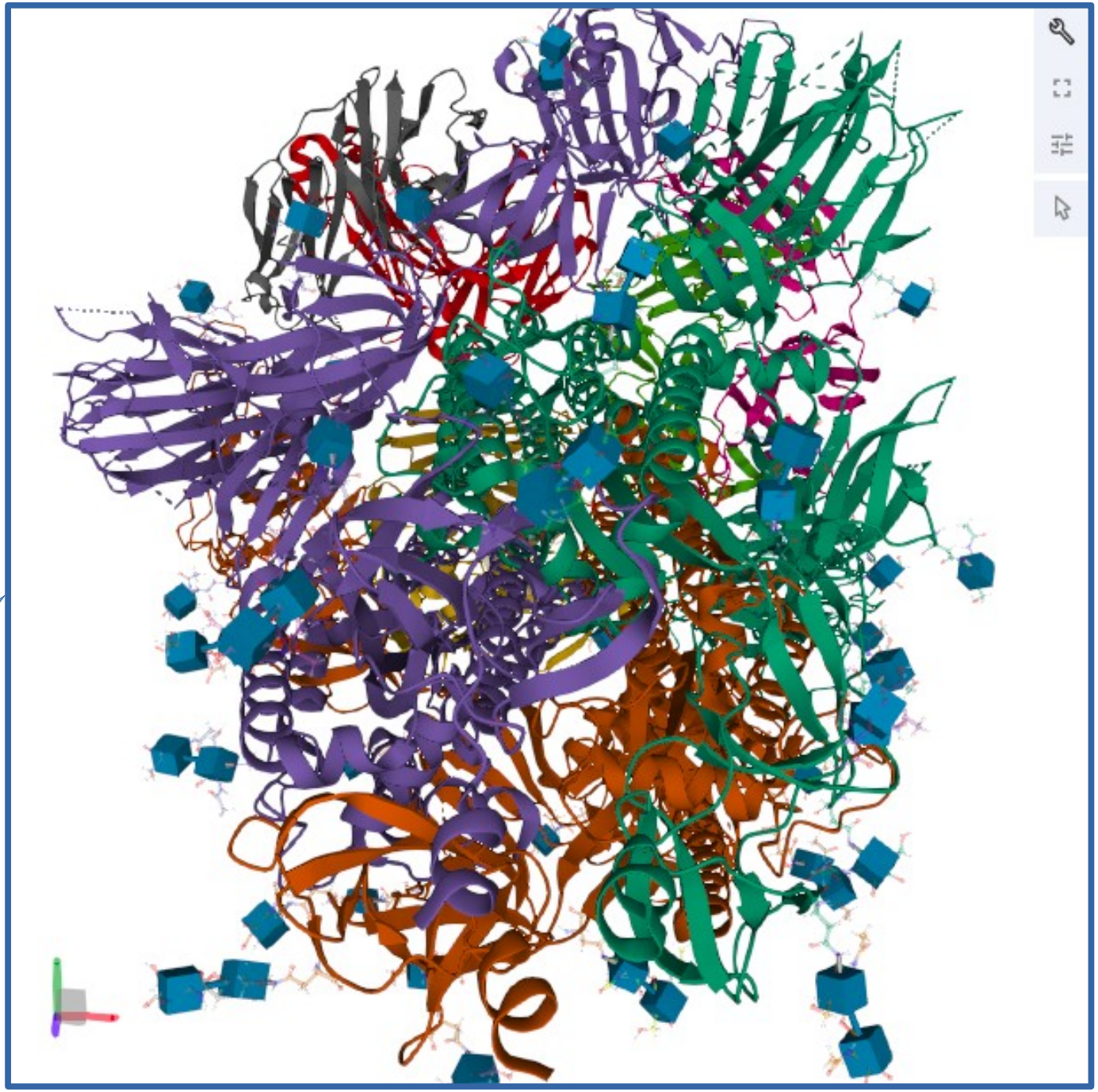
It takes a long time to virtually fold proteins. This data is already “folded” and you can view it as a folded protein structure.




- http://www.rcsb.org/

# RCSB Output

This image is a link!



# Viewing Options To Animate


Save animation to a file