

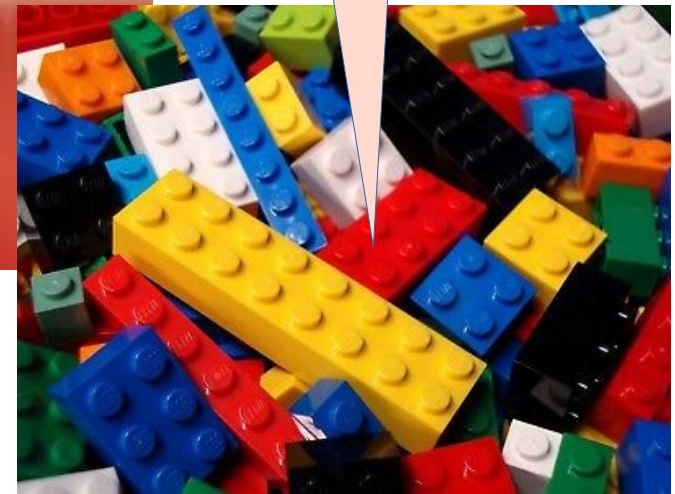
**Bioinformatics**  
**CS300**  
**Prediction and**  
**Modeling Protein Structure**

**Spring 2021**  
**Oliver BONHAM-CARTER**

# Properties From Combining Pieces

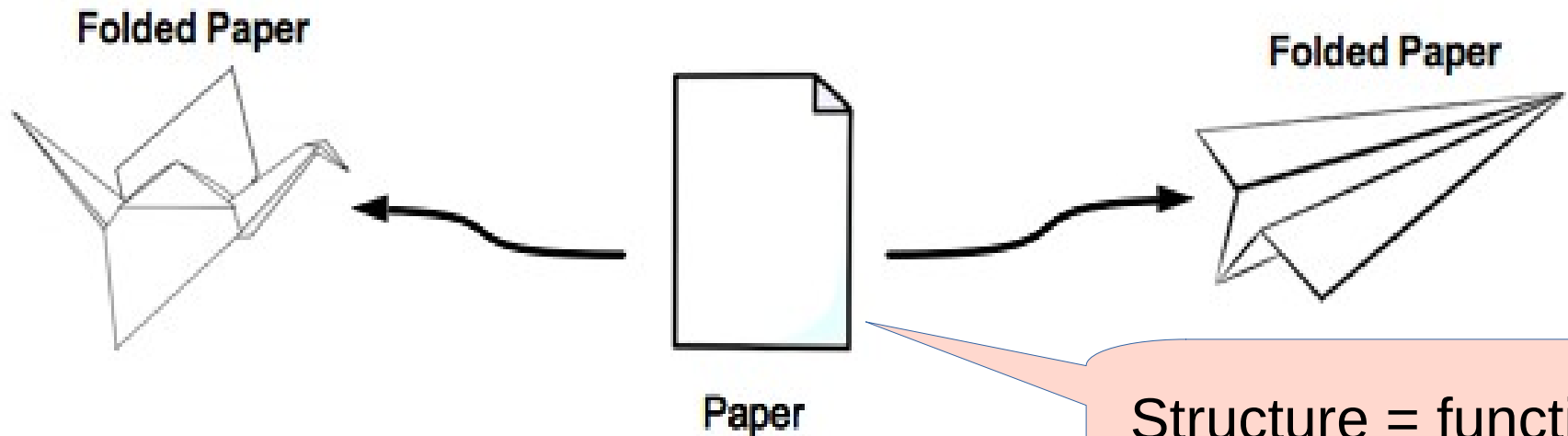


**From  
these  
pieces?**

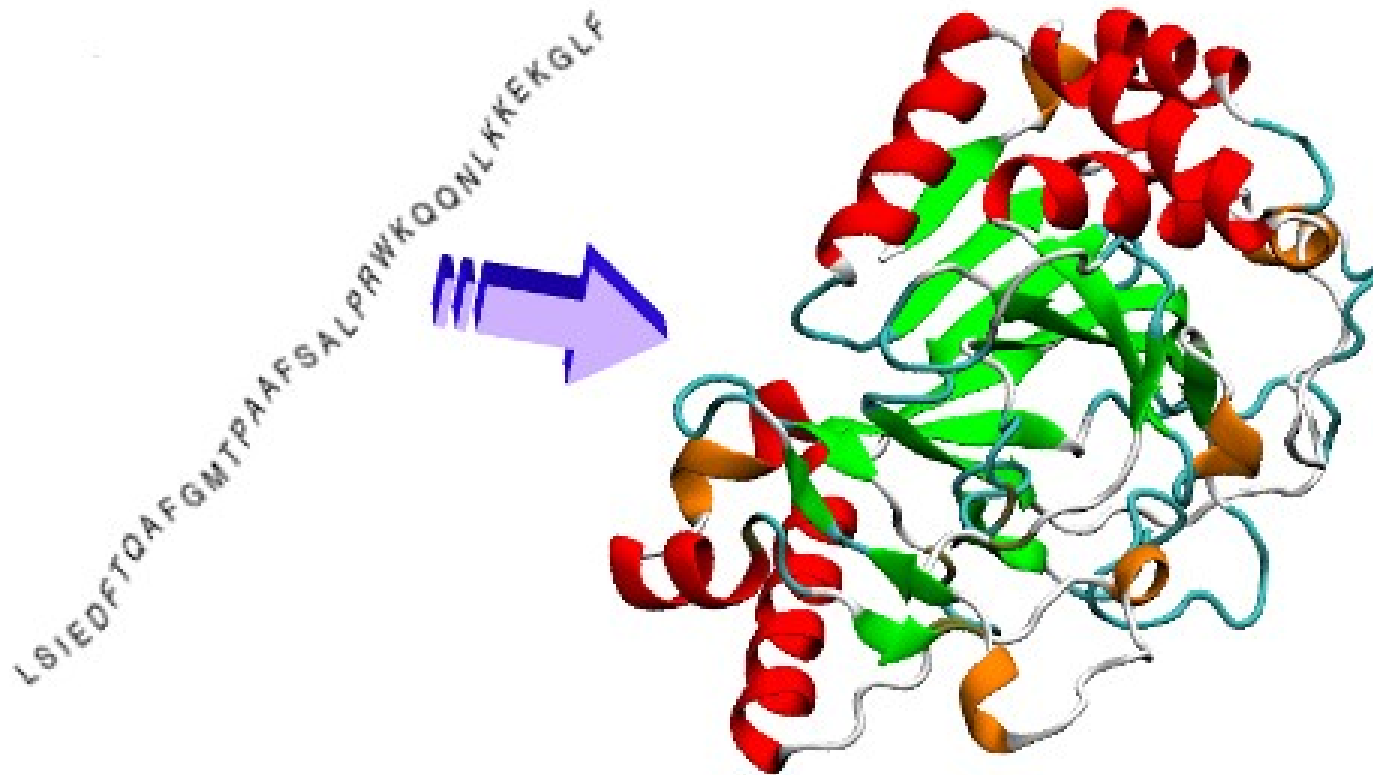


**A cool living room  
made from Lego pieces!**

# Properties From Folding



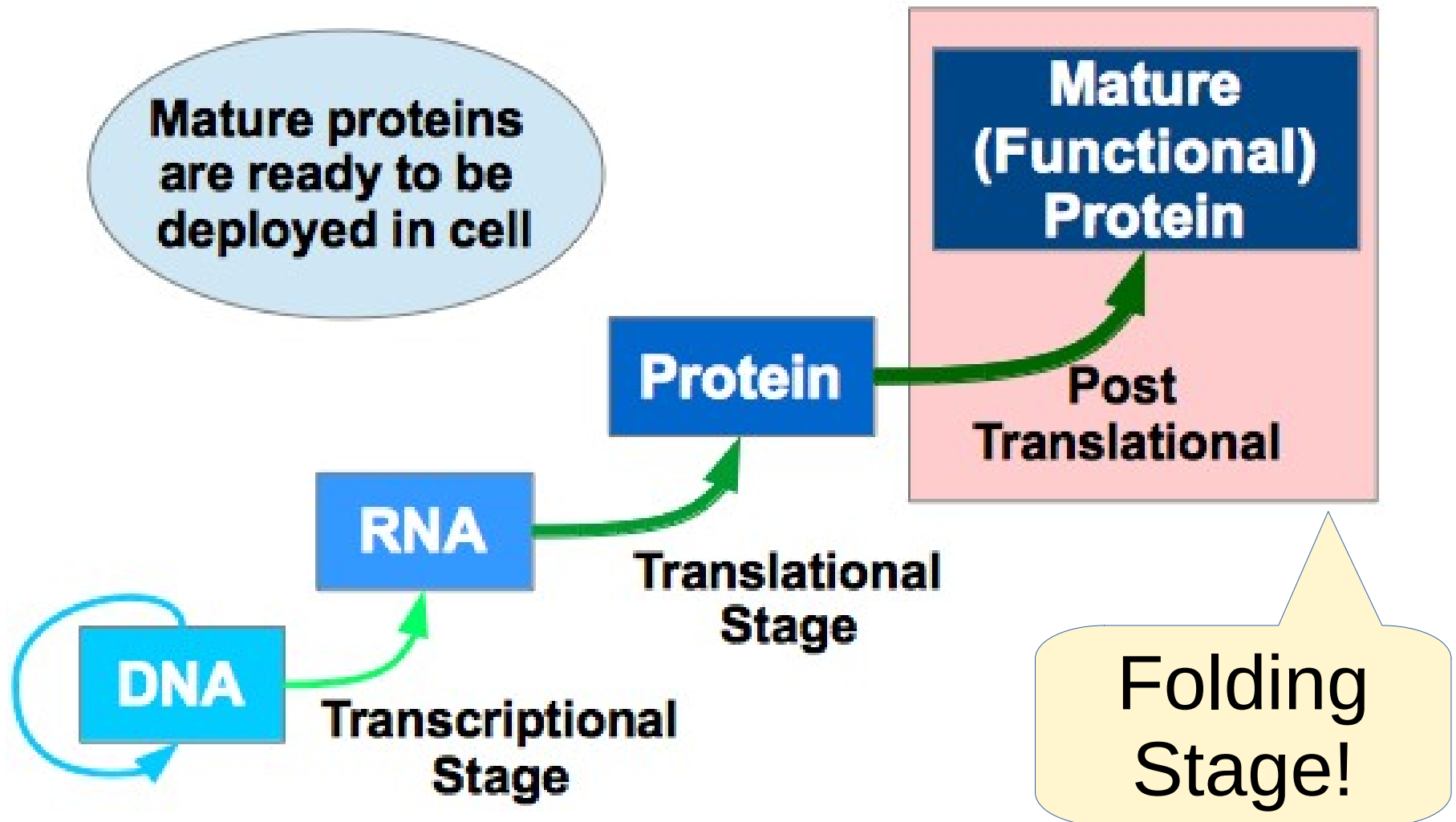
# Protein Folding



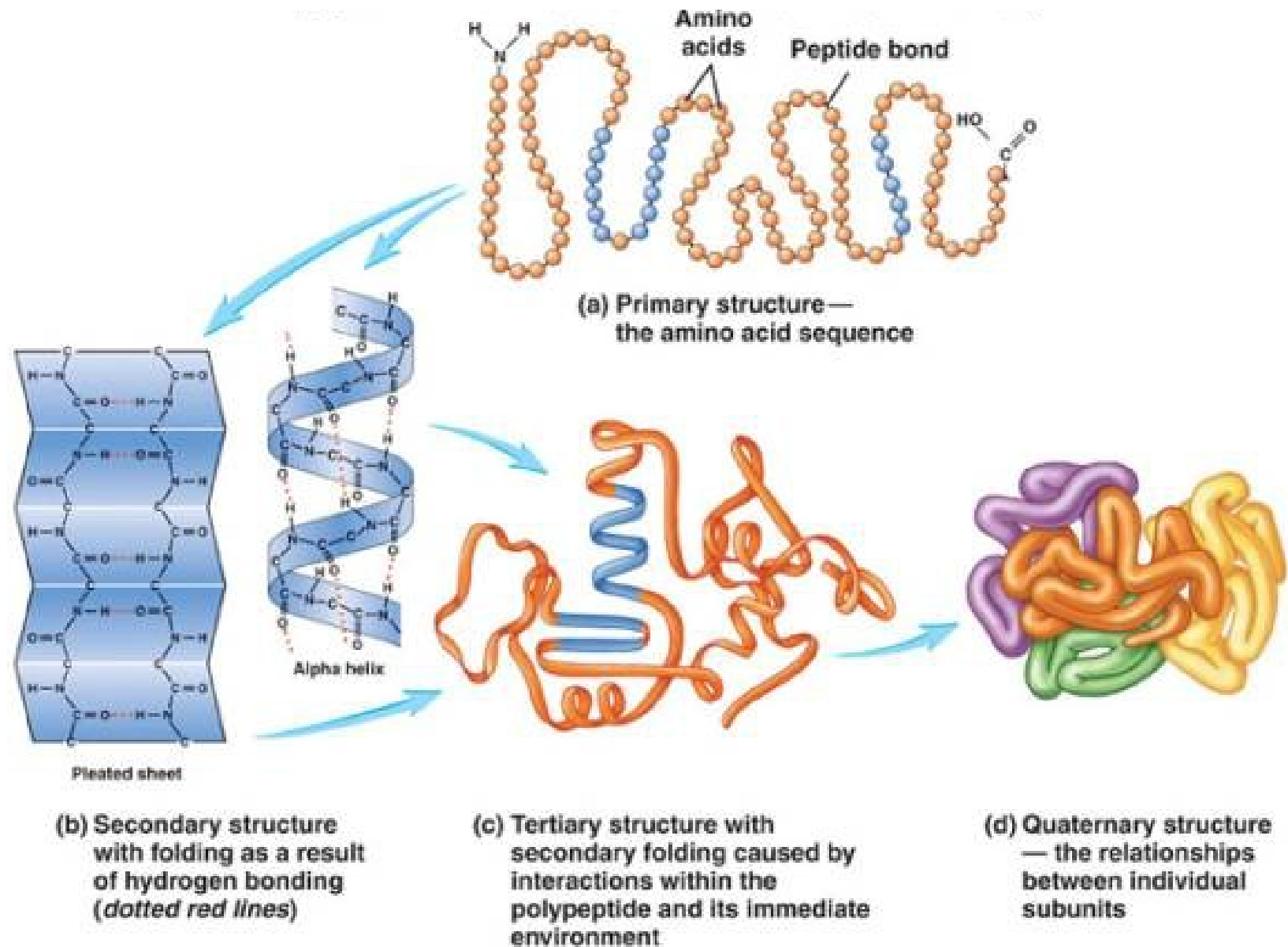
- A protein sequence is a linear chain of amino acids produced by ribosomes during translation
- A structure from folding, 3D state based on properties of amino acids and structure



# Protein Folding and the Central Dogma of Biology

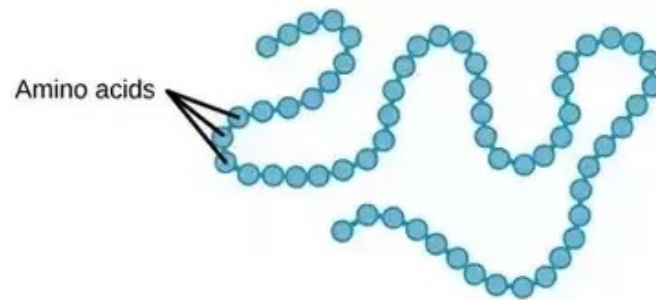


# Protein Folding: Four Stages

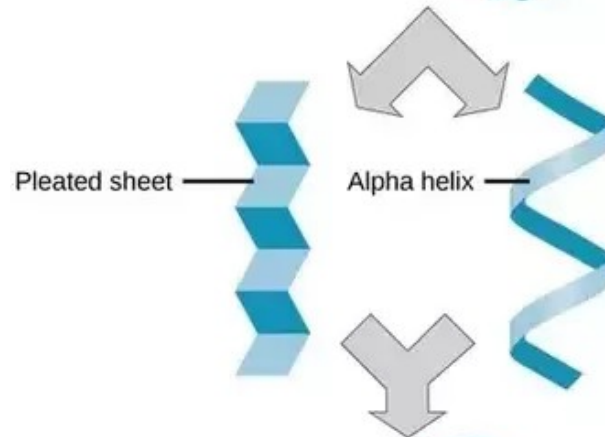




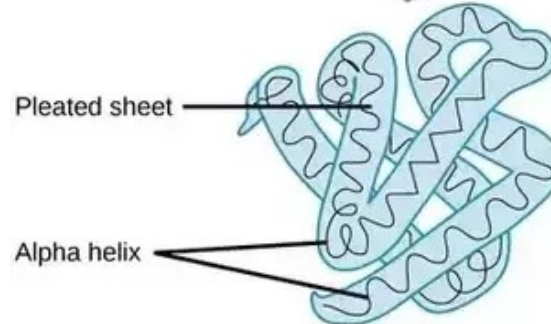
# Protein Folding: Another View



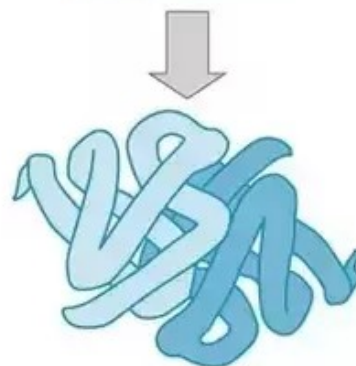
**Primary Protein structure**  
sequence of a chain of  
amino acids



**Secondary Protein structure**  
hydrogen bonding of the peptide  
backbone causes the amino  
acids to fold into a repeating  
pattern



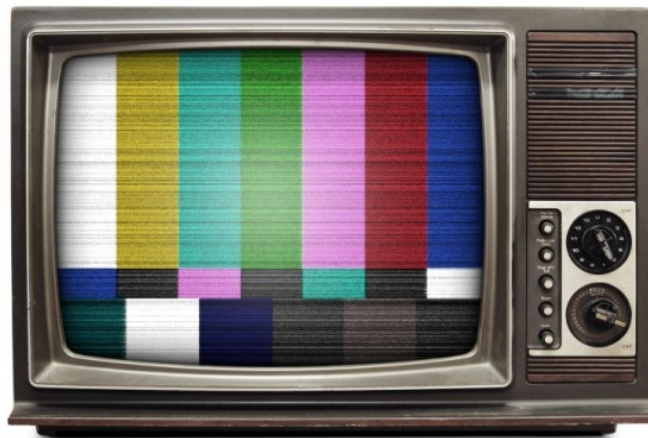
**Tertiary protein structure**  
three-dimensional folding  
pattern of a protein due to side  
chain interactions



**Quaternary protein structure**  
protein consisting of more  
than one amino acid chain

# Supporting Videos

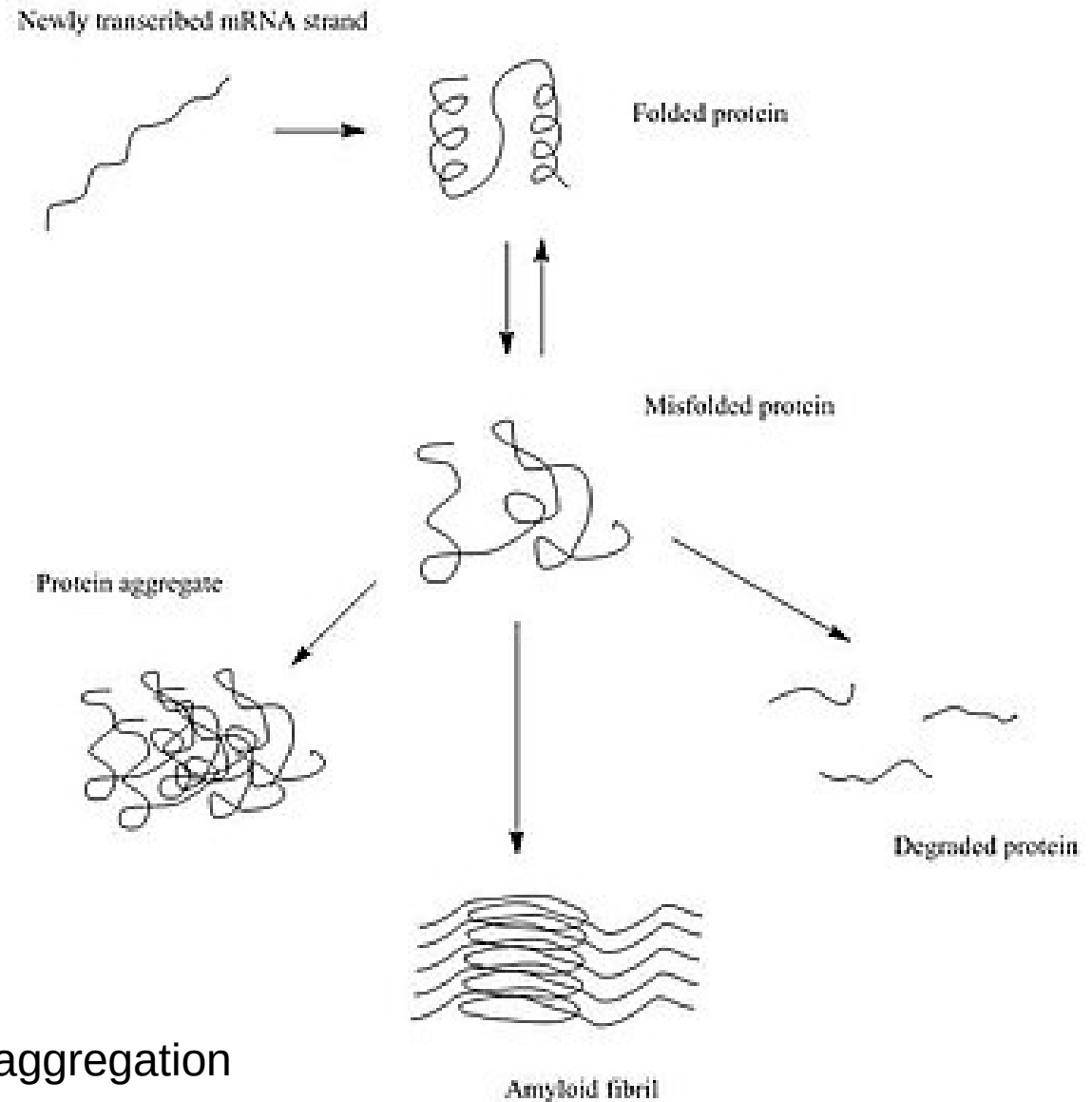
- **Protein Folding (3 mins)**
  - <https://www.youtube.com/watch?v=yZ2aY5lxEGE>
- **What is a protein? (3D shape and function, 3 mins)**
  - <https://www.youtube.com/watch?v=qBRFIMcxZNM>
- **Protein folding simulation (3 mins)**
  - <https://www.youtube.com/watch?v=meNEUTn9Atg>





# 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)

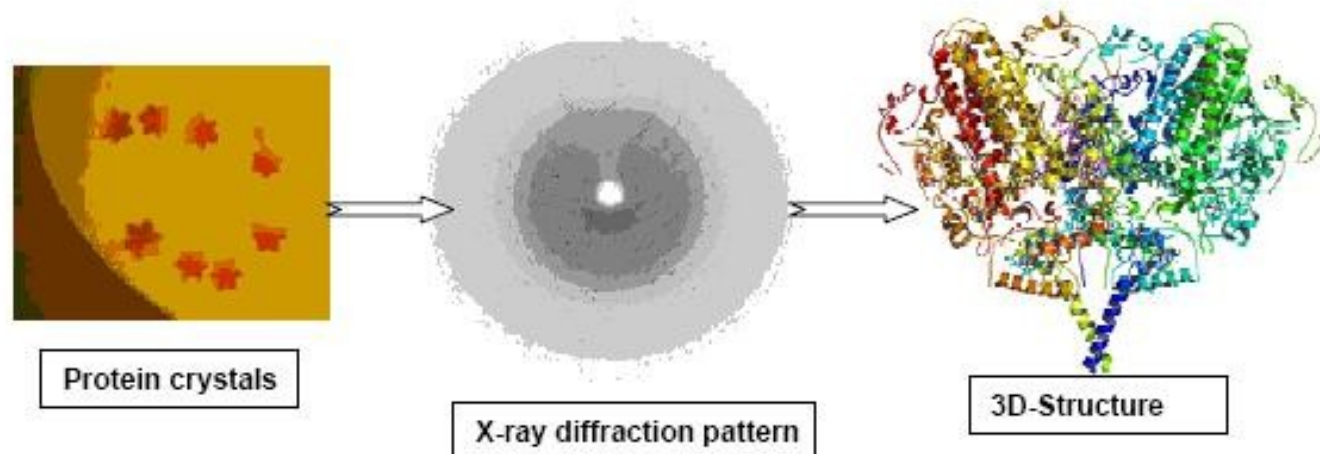
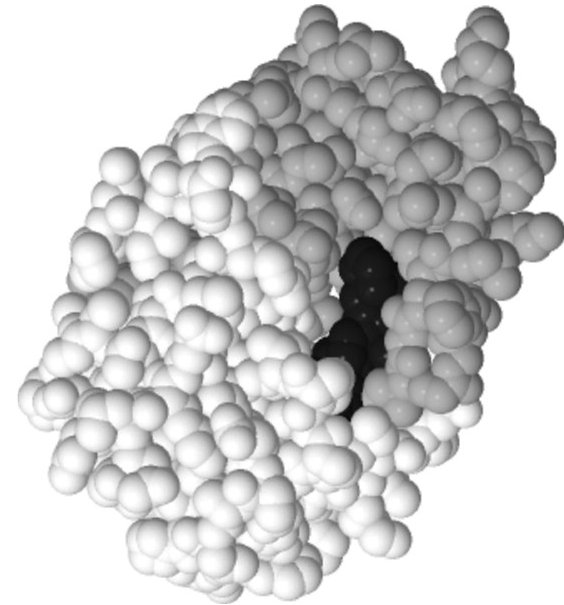


Ref:

[https://en.wikipedia.org/wiki/Protein\\_aggregation](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

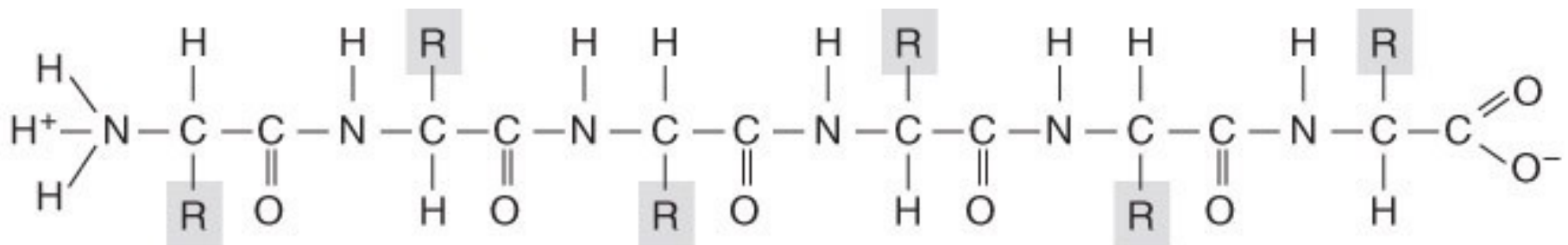
## **Computational Structural Biology** (study)

- 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

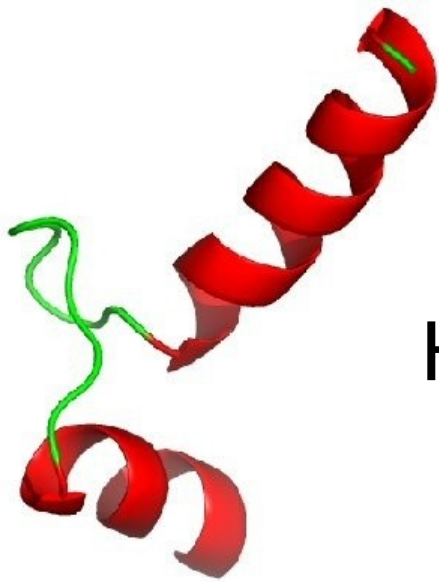
- Linus Pauling – Studied the limitations on protein folding
  - Nature of chemical bonds between amino acids
  - Bond angles
  - Rotation of atoms
  - Flexibility of side chains

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

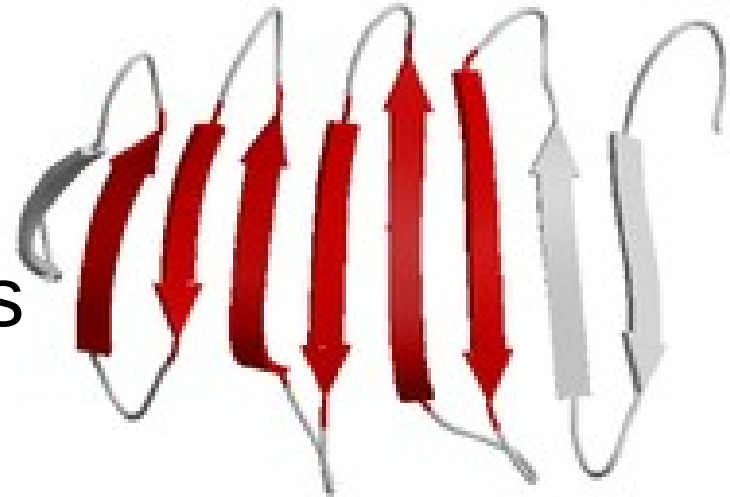


(A) Primary (1°) structure

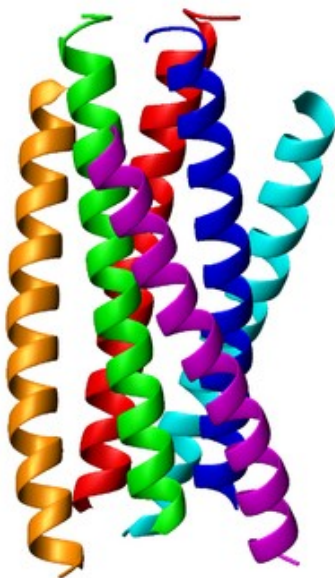
# Parts of Protein (Structures)



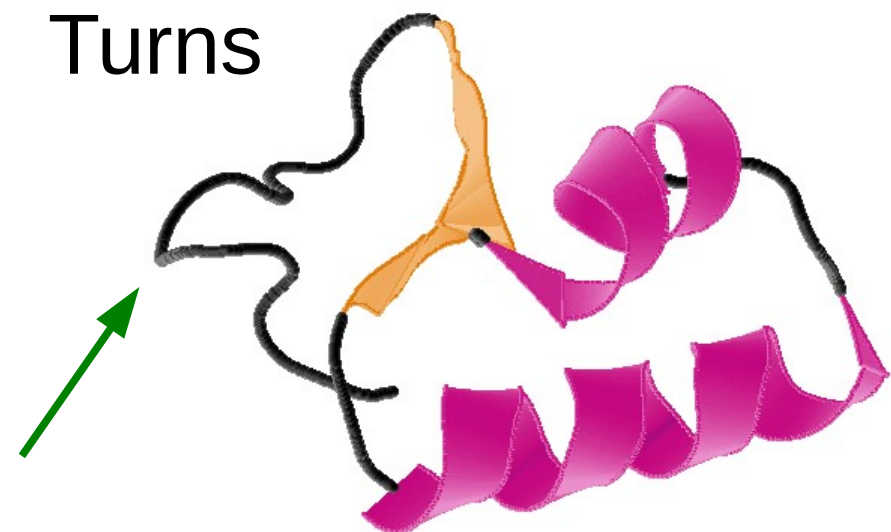
Helices



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
      MQIFVKLTGKTTITLEVEPSDTIENVKAKIQDKEGIPPDQQLIFAGKQL
helix  HH                      HHHHHHHHHHHH                      H
sheet  EEEE          EEEEE
turns  T
coil   CCC CC          CCC                      CC          CCCCC
      .   60   .   70
      EDGRTLSDYNIQKESVNHLVLRRLGG
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**





ALLEGHENY  
COLLEGE

# Bring the Tool!



**Up Next!**

# Protein Folding: Quick Solutions

This image  
is a link!

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](#))

### 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:  No file chosen
3. To enter the sequence data manually, type here:

```
>KX932045.1 Orchid fleck dichorhavirus isolate U1 phosphoprotein mRNA, complete cds
ATGTTCACTACCAAGGTAAATATGTACCCAGAGGTGCCAGCTCATCCAGGTGTCAGACGACATAGACA
ATGACACGCACATCGACGAGGTCGCTGCATTTGTGAGAAAGTGGTCGGCAGCCGGACTATCTCCCCCAT
CACCCTTGCGAAGAACCTCAGAGCATGGATATCAAGCAACACCAGCCCTGGAAGCCCCCTAGTGTGGAT
GACAGAATGCTGAGCCTTACAACCATGATATGGAACACAGCAGCAGAGCACTACACAATGATAGGCAAT
CCCAGGTCAATCGTATGTCATCACTCATAGATCAGCTGGGGGAGATTTCCGGCCGCAAACCGCCGAGGG
CCCAGCATTGACATGCCTCCTCCCCCTCCTAAGAGAAAACATCCGGATTCACTAGACACTAATCCAATA
TTAGGCTTAATAGGTCAAGATTGGGACGACAATAAAGACAAGCACTGGAGAGAGAAACCAGCAGACAAGA
AGCTCCTCGTGCTCAACTGGGTGTTGCATGAGTATCTGGGGGTCTCACAAAACCTGTCACCATCAAGTG
GATAACGGATAACCCCGCGTCTTTAGAGTTGGGAGCAGTGTGAGCTTATGCCCTGAAACATCAGGCCAGC
TTATCCGACTGCGACAAGGAAGCCCTCAGAGCGTTGGTGGTTCAAACAGTGAAAAACACCCCCAAAAGGC
CATGCCTGGACTAG
```



# Garnier Output (text)

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

## OUTPUT FILE [outfile](#)

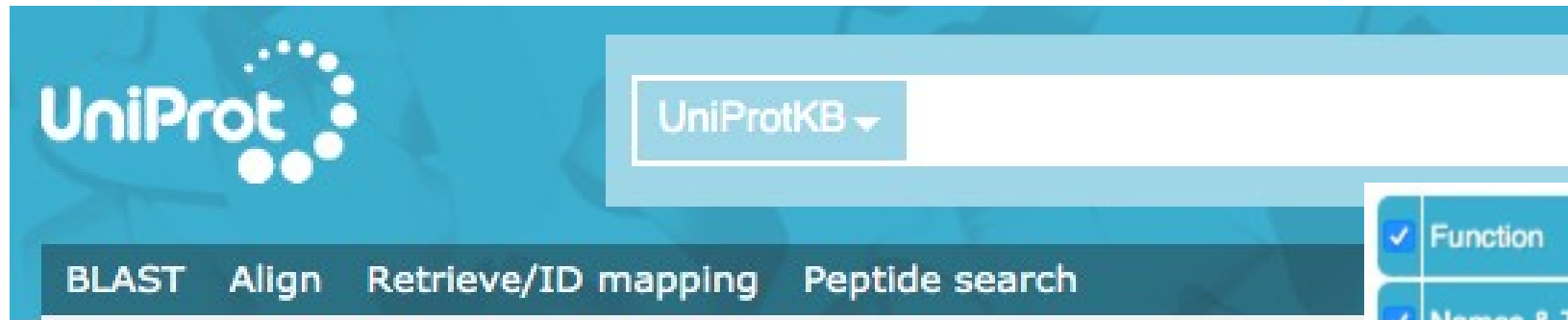
```
#####
# Program: garnier
# Rundate: Mon 26 Apr 2021 05:31:19
# Commandline: garnier
#   -auto
#   -sequence /var/lib/emboss-explorer/output/626691/.sequence
#   -outfile outfile
#   -rformat2 tagseq
# Report_format: tagseq
# Report_file: outfile
#####

#=====
#
# 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
      ATGTTCACTACCAAGGTAAATATGTACCCAGAGGTGCCAGCTCATCCCA
helix                H   HHHHH
sheet                EEEE          EEEE          EEE
turns TTTTTTTT      TT          TTTT  TTTTTTTTTTTTTTTT  TT
coil                C

      .   60   .   70   .   80   .   90   .  100
      GGTGTCAGACGACATAGACAATGACACGCACATCGACGAGGTCGCTGCAT
helix                HHHHHHHHHHHH
sheet                E                      E          EEE
turns TTTTTT TTTTT          TTTTTTTTTTTT TTTTTTTTTTTT
coil
```

Not a  
Recent  
algorithm



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/>

✓	Function
✓	Names & Taxonomy
✓	Subcellular location
✓	Pathology & Biotech
✓	PTM / Processing
✓	Expression
✓	Interaction
✓	Structure
✓	Family & Domains
✓	Sequences (2)
✓	Similar proteins
✓	Cross-references
✓	Entry information
✓	Miscellaneous

# PREDICT PROTEIN

## Dashboard Overview for CD44\_HUMAN



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>



ALLEGHENY  
COLLEGE

# Bring the Tool!




**Up Next!**







# Protein Folding: Slower Solutions

 **PREDICT**  
**PROTEIN** OPEN

[Help Tutorials](#) | [Sample Output](#)

 **PredictProtein** is free to use and open to all users with no login requirements. 

If you're looking for PredictProtein with account access, please visit [login.predictprotein.org](https://login.predictprotein.org)

The web server currently does not support batch processing. If you are looking for batch processing, we recommend using our [docker image](#) (see "Software" below) or to [contact us](#).

```
MFTTKVNMYPEVPSSSQVSDDIDNDTHIDEVAAFVRKWSAAGLSPPITLAKNLRRAWIS
SNTSPGSPLVLDDRMLSLTTMIWNTAAEHYTMIGKSQVNRMSSLIDQLGEISGRKPP
QGPAFDMPPPPPKRKHPDSLDTNPILGLIGQDWDDNKDKHWREKPADKLLVLNVV
LHEYLGVLTKPVTIKWITDNPASLELGAVSAYALKHQASLSDCDKEALRALVVQTVKNT
PKRPCLD
```

[Clear](#) [PredictProtein](#) [\[Example Input 1\]](#)  
[\[Example Input 2\]](#)

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  ADKLLVLNW  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



ALLEGHENY  
COLLEGE

# Predict Protein Output

Helix = H,  
Strand = S,  
Loop = L

	.....1.....2.....3.....4.....5.....6
AA	MFTTKVNMYP <b>EV</b> PSSSQVSD <b>DD</b> IDNDTHID <b>EV</b> AA <b>FVR</b> KWSAAGLSPPITLAK <b>NL</b> RAWISSN
OBS_sec	??
PROF_sec	HHHHHHHHHHH HHHHHHHHHHHH
Rel_sec	954210246667861012100245456543678876411237777223778788888521
SUB_sec	<b>LL</b> ..... <b>LLLLLL</b> ..... <b>L.LLL</b> .. <b>HHHHHH</b> ..... <b>LLLL</b> .. <b>HHHHHHHHHHH</b> ..
O_3_acc	bb
P_3_acc	eeeeee <b>b</b> <b>e</b> eeeee <b>bee</b> <b>beeee</b> <b>e</b> bbbbbb <b>eebee</b> <b>eb</b> bbbbb <b>ebbe</b> <b>b</b> <b>e</b> <b>beee</b>
Rel_acc	302131111010121210331132311022524553262413310102335484506332
SUB_acc	..... <b>b</b> .. <b>bbb</b> .. <b>b</b> .. <b>e</b> ..... <b>eibie</b> .. <b>b</b> ...

	.....7.....8.....9.....10.....11.....12
AA	TSPGSPLVL <b>DD</b> RMLSLTTMIWNTAA <b>E</b> HYTMIG <b>K</b> SQV <b>N</b> RMS <b>S</b> LIDQLGEISGR <b>K</b> PPQGP <b>A</b> F
OBS_sec	??
PROF_sec	EEEE HHHHHH HHH
Rel_sec	4778613540110000001120025555432010034688888888876504777767666
SUB_sec	.. <b>LLLL</b> .. <b>E</b> ..... <b>HHHH</b> ..... <b>HHHHHHHHHHHHHH</b> .. <b>LLLLLLLLLL</b>
O_3_acc	bb
P_3_acc	eeee bbbbe bb bbbbbbbbbbb bbebb bbbbe bbebb <b>ebb</b> <b>eb</b> eee ee <b>e</b>
Rel_acc	115321347112102454693335212022322014237146840923024220121202
SUB_acc	.. <b>e</b> ..... <b>bb</b> ..... <b>bbbbbb</b> .. <b>b</b> ..... <b>b</b> .. <b>b</b> .. <b>ebbe</b> .. <b>b</b> ..... <b>e</b> .....



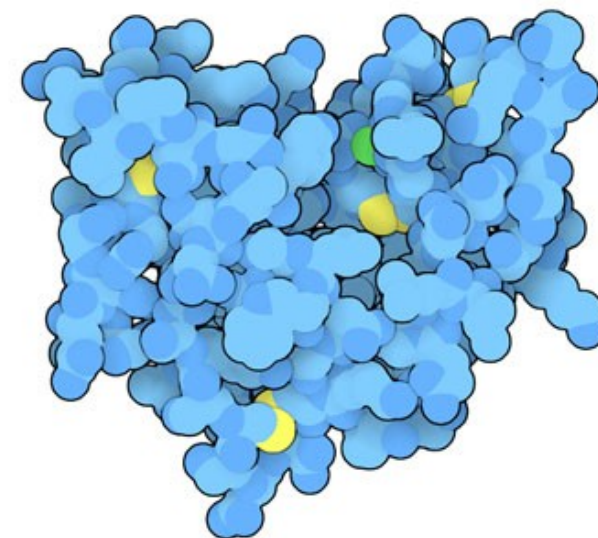
# RCSB PDB

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.



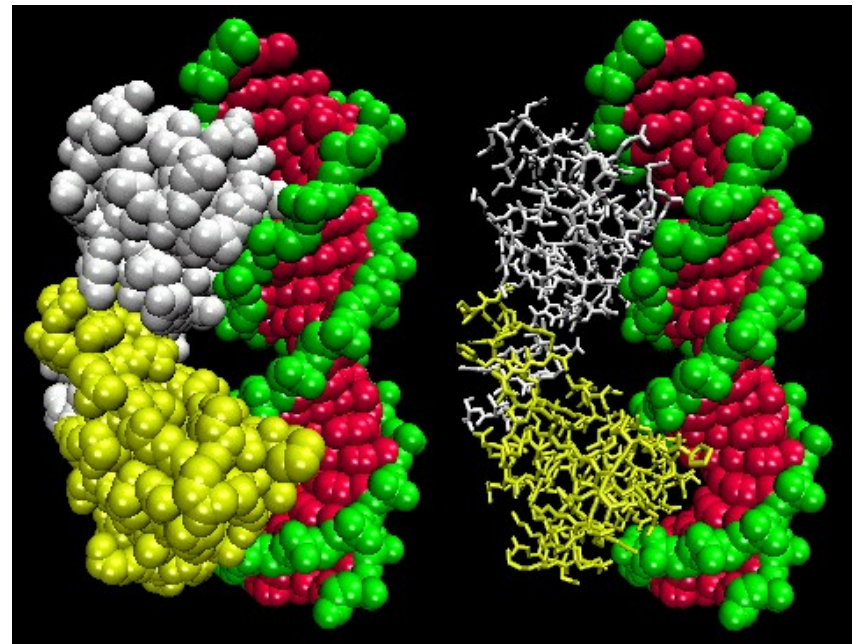
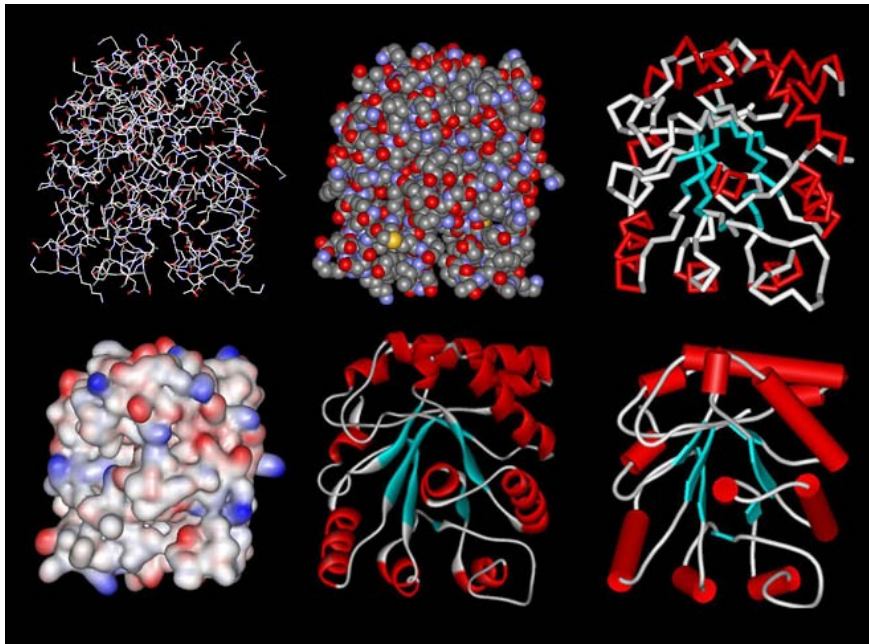
- <http://www.rcsb.org/>





# 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.

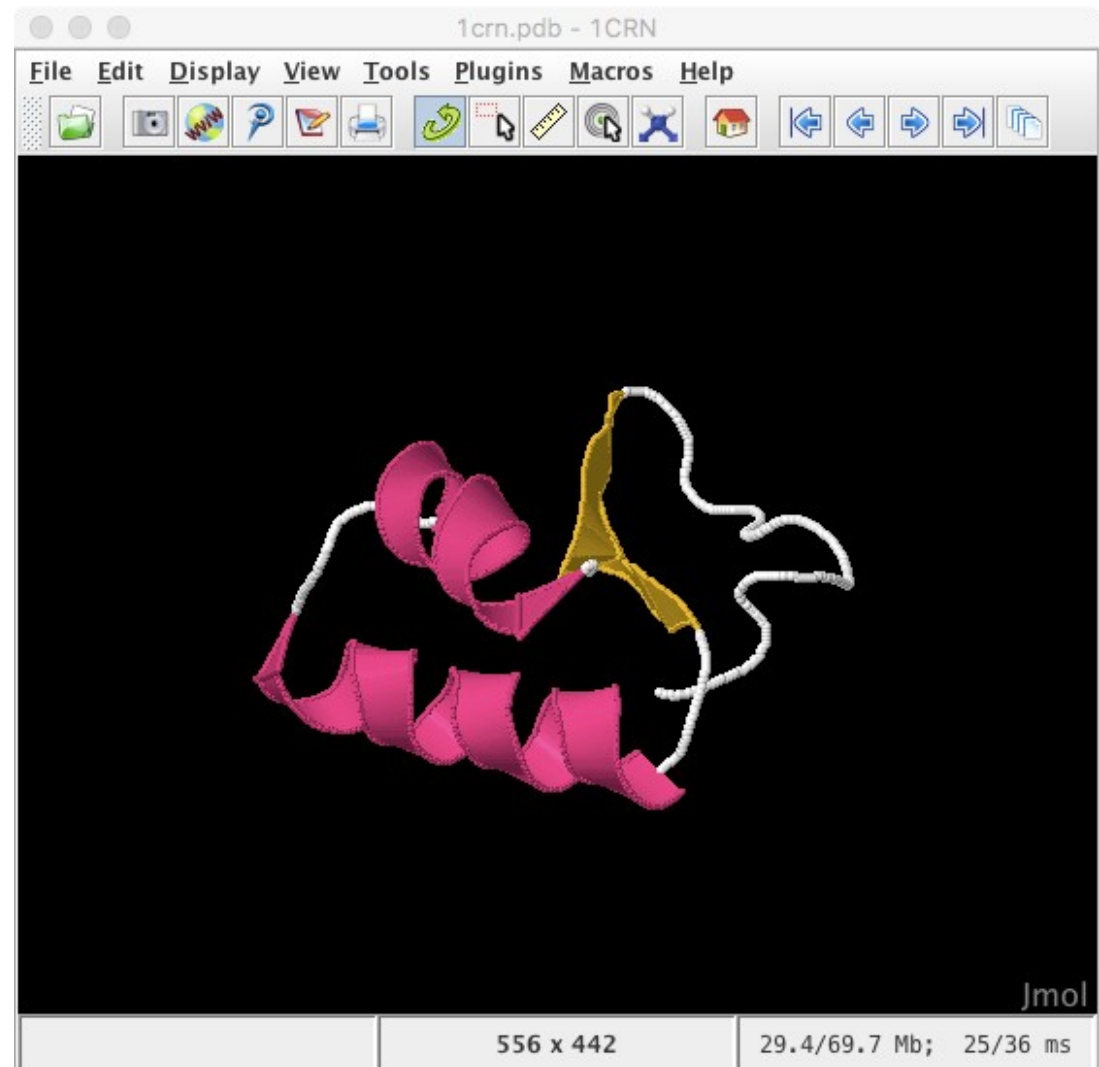




# Jmol: A (local) Graphical Viewer For Protein Sequences



- Download:
  - <http://jmol.sourceforge.net/>
- Wiki:
  - [http://wiki.jmol.org/index.php/Jmol\\_Application#Installing\\_Jmol\\_Application](http://wiki.jmol.org/index.php/Jmol_Application#Installing_Jmol_Application)





ALLEGHENY  
COLLEGE

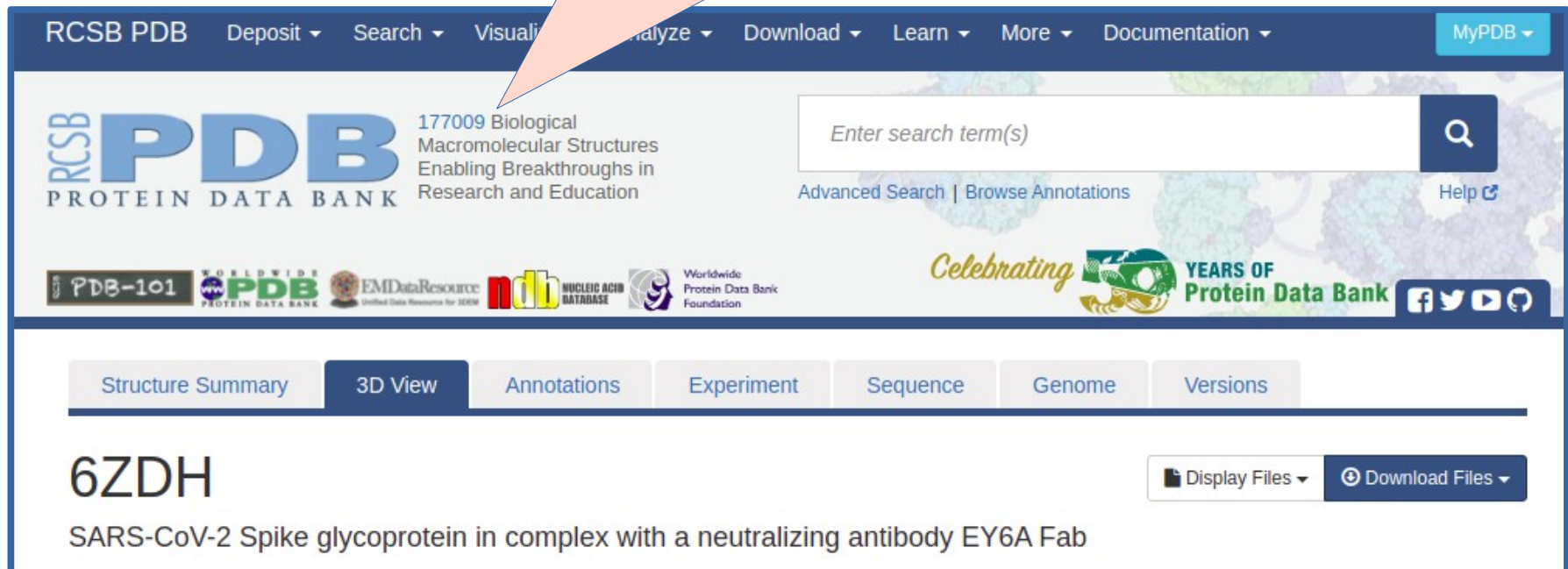
# Bring the Tool!



**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.

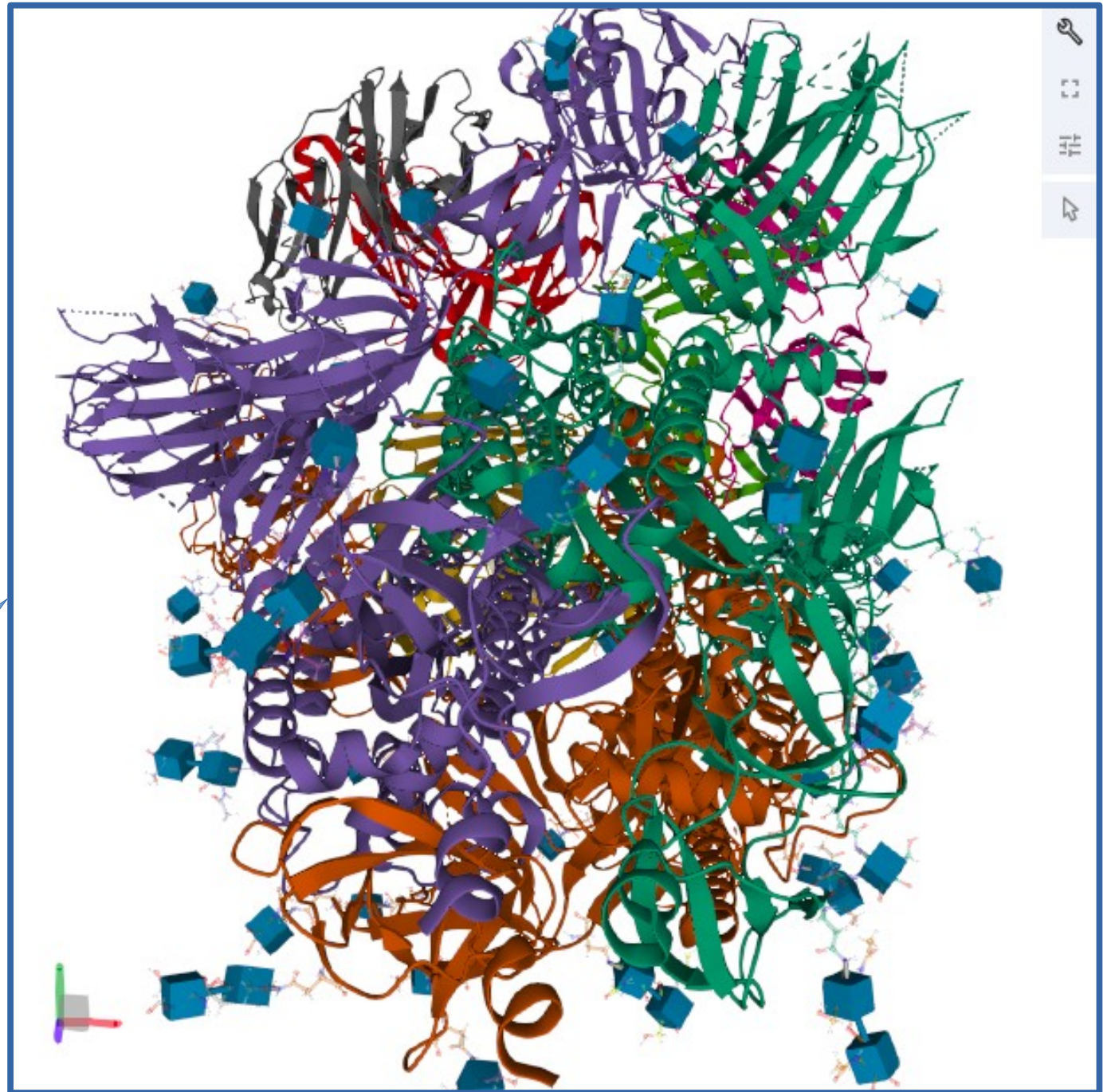


The screenshot displays the RCSB PDB website interface. The top navigation bar includes links for Deposit, Search, Visualize, Analyze, Download, Learn, More, and Documentation, along with a MyPDB button. The main header features the RCSB PDB logo, the text "177009 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education", a search bar with the placeholder "Enter search term(s)", and links for Advanced Search, Browse Annotations, and Help. Below the header, there are logos for PDB-101, Worldwide PDB, EMDatabank, Nucleic Acid Database, and the Worldwide Protein Data Bank Foundation. A banner celebrates "50 YEARS OF Protein Data Bank". The main content area shows tabs for Structure Summary, 3D View, Annotations, Experiment, Sequence, Genome, and Versions. The 3D View tab is selected, displaying the protein structure 6ZDH, identified as "SARS-CoV-2 Spike glycoprotein in complex with a neutralizing antibody EY6A Fab". There are buttons for Display Files and Download Files.

- <http://www.rcsb.org/>

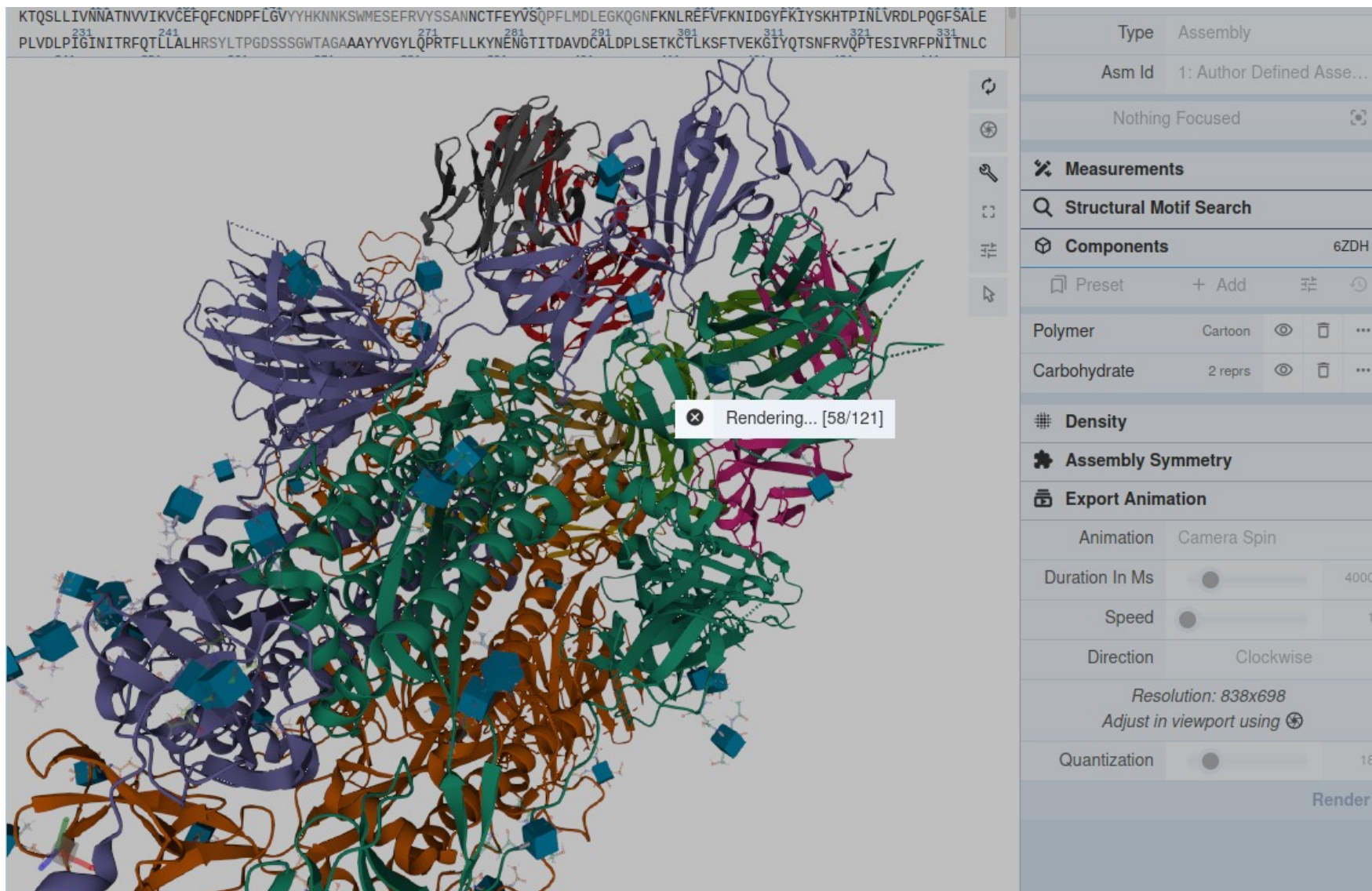
# RCSB Output

This image  
is a link!





# Viewing Options To Animate



The screenshot displays a 3D ribbon model of a protein complex, likely a dimeric enzyme, rendered in various colors (purple, green, orange, red, blue) to distinguish different subunits or regions. The protein is shown in a ribbon representation, with some regions highlighted in solid colors. A small white box with a red 'X' icon and the text "Rendering... [58/121]" is overlaid on the protein model.

The interface includes a top bar with the protein sequence: `KTQSLLIVNNATNVVVKVEFCFCNDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFNIDGYFKIYSKHTPINLVRDLPQGFSALE` and `PLVDLPIGINITRFQTLALHRSYLTGDSSSGWTAGAAAYYVGYLQPRTFLLKYNENGITIDAVDCALDPLSETKCTLKSTVEKGIYQTSNFRVQPTESIVRFPNITNLC`.

The right sidebar contains the following sections:

- Type**: Assembly
- Asm Id**: 1: Author Defined Asse...
- Nothing Focused**
- Measurements**
- Structural Motif Search**
- Components**: 6ZDH
- Preset**: + Add
- Polymer**: Cartoon (eye icon, trash icon, ...)
- Carbohydrate**: 2 reprs (eye icon, trash icon, ...)
- Density**
- Assembly Symmetry**
- Export Animation**
- Animation**: Camera Spin
- Duration In Ms**: 4000
- Speed**: 1
- Direction**: Clockwise
- Resolution**: 838x698
- Adjust in viewport using** (gear icon)
- Quantization**: 18
- Render**

Save animation to a file