Philipps Universität Marburg

UKGM UNIVERSITÄTSKLINIKUM GIESSEN UND MARBURG

Leveraging Molecular Databases for Variant Interpretation

Functional Genomics; understanding how pathological missense variants affect proteins

Instructor:

Pouria Dasmeh
Center for Human Genetics, Marburg University/UKGM medical center,
dasmeh@staff.uni-marburg.de
www.dasmehlab.com

Course material
lectures & tutorials)

https://github.com/dasmeh/UKGM_Humangenetik_PROTEINS

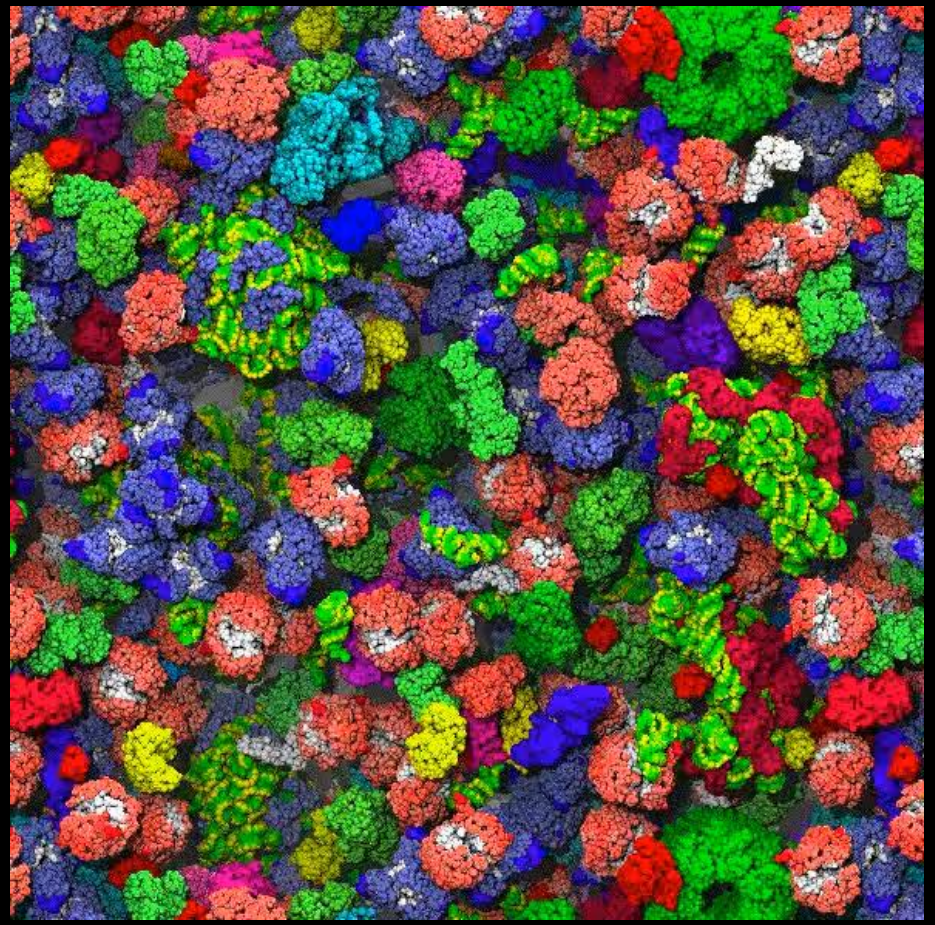
Course agenda and organizational notes

Theory lecture (04:30-05:00): Protein sequence, structure, function

Practical session 1 (05:00-05:30)

Short break

Practical session 2 (05:45:06:30)



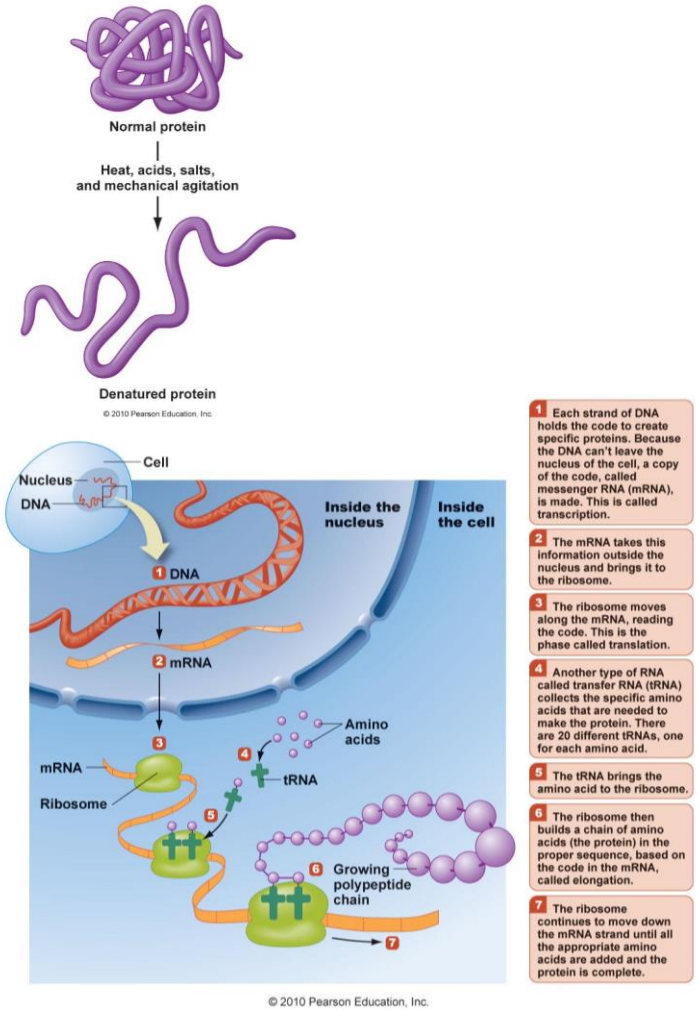
PLoS computational biology, *6*(3), e1000694.

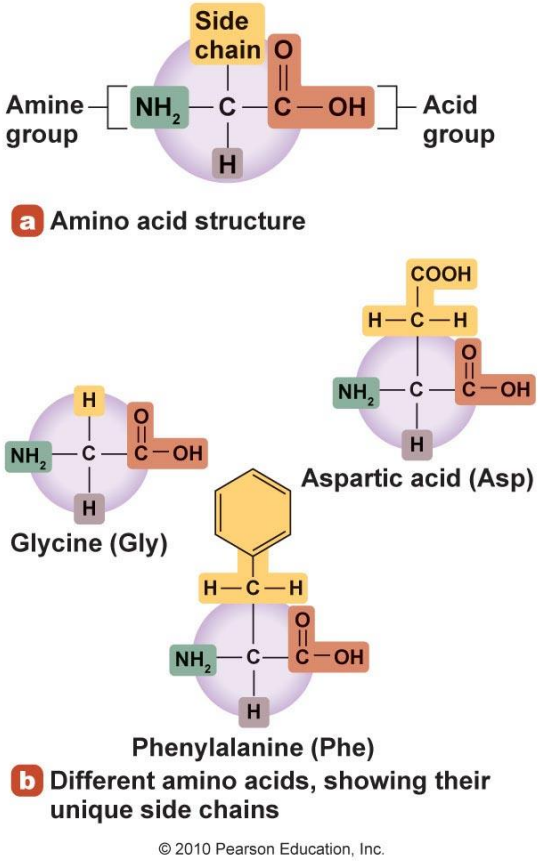
Cytoplsm’s viscosity > 100 mPa.s

Water ~ 1 mPa.s
Engine oil ~ 100 mPa.s

Part 1:
Protein Sequence, Structure, Function

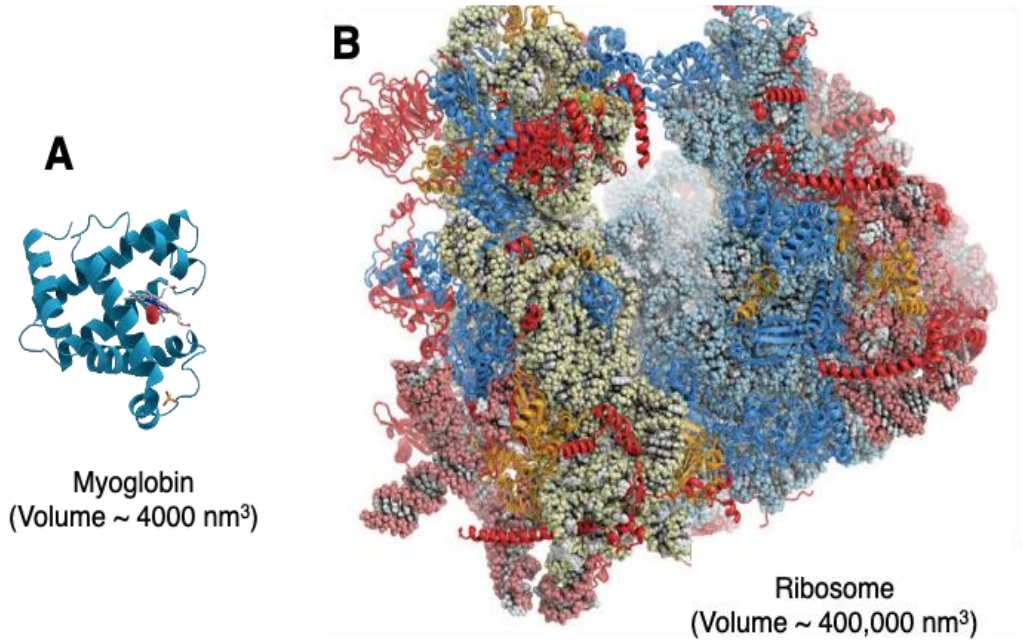
Proteins and Amino Acids





Proteins

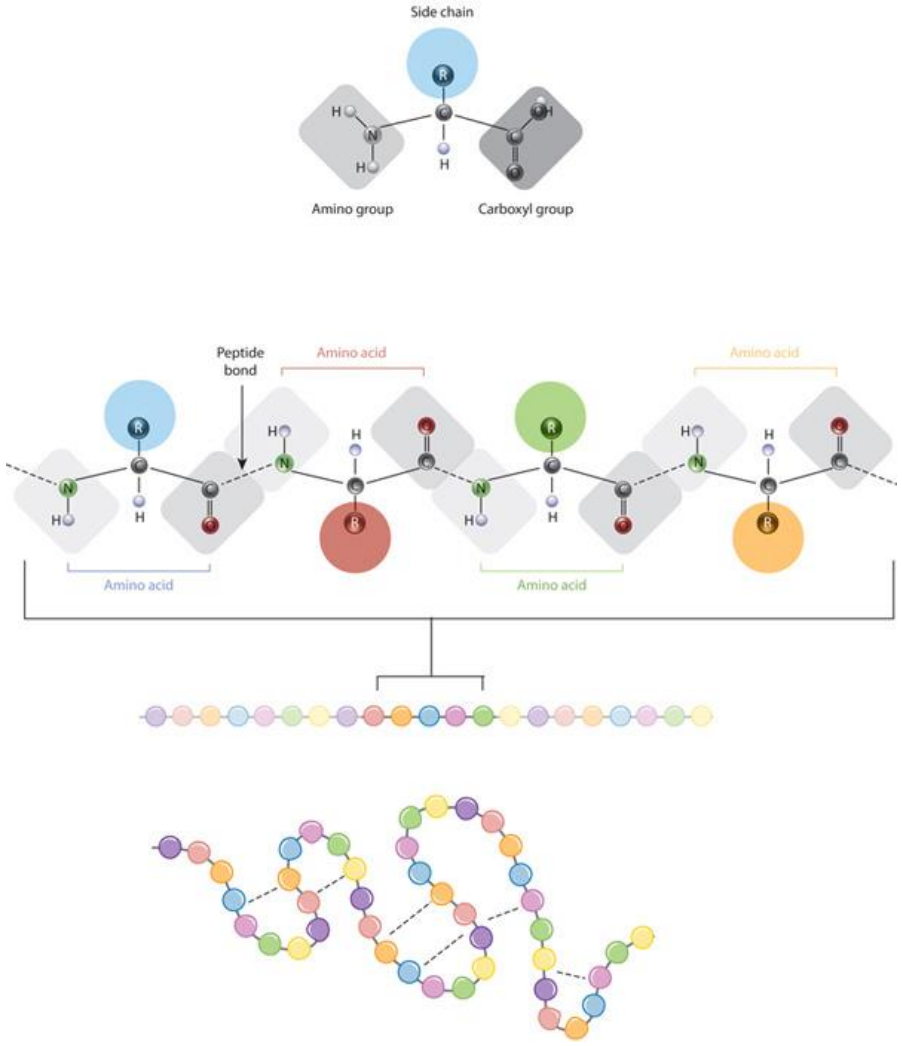
- Made up of chains of amino acids
- Are involved in most of the body's functions and life processes
- The sequence of amino acids is determined by DNA



Structure of Proteins

- Made up of chains of amino acids; classified by number of amino acids in a chain
  - Peptides: fewer than 50 amino acids
    - Dipeptides: 2 amino acids
    - Tripeptides: 3 amino acids
    - Polypeptides: more than 10 amino acids
  - Proteins: more than 50 amino acids
    - Typically 100 to 10,000 amino acids linked together
- Chains are synthesizes based on specific bodily DNA.
- Amino acids are composed of carbon, hydrogen, oxygen, and nitrogen.

An amino acid



Amino acids

TWENTY-ONE PROTEINOGENIC α-AMINO ACIDS

Side chain charge at physiological pH 7.4

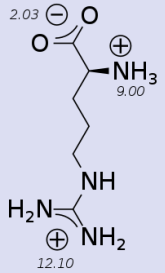
pK_a values shown italicized

⊕ Positive
⊖ Negative

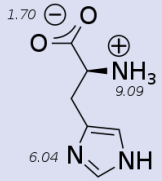
A. Amino Acids with Electrically Charged Side Chains

Positive

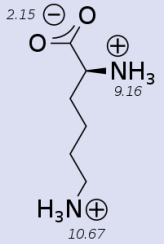
Arginine Arg R



Histidine His H

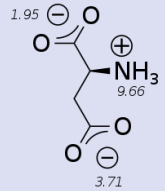


Lysine Lys K

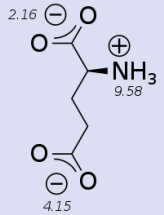


Negative

Aspartic Acid Asp D

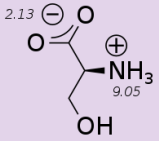


Glutamic Acid Glu E

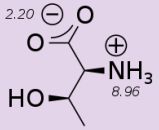


B. Amino Acids with Polar Uncharged Side Chains

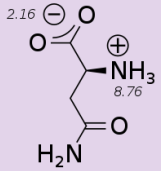
Serine Ser S



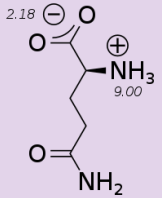
Threonine Thr T



Asparagine Asn N

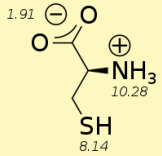


Glutamine Gln Q

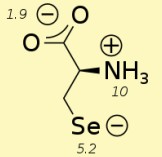


C. Special Cases

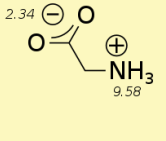
Cysteine Cys C



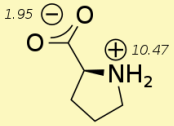
Selenocysteine Sec U



Glycine Gly G

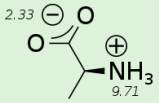


Proline Pro P

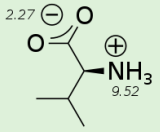


D. Amino Acids with Hydrophobic Side Chains

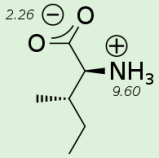
Alanine Ala A



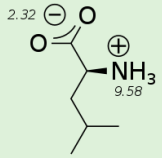
Valine Val V



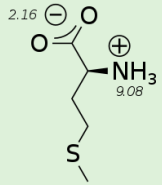
Isoleucine Ile I



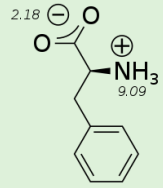
Leucine Leu L



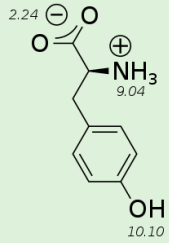
Methionine Met M



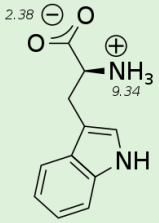
Phenylalanine Phe F



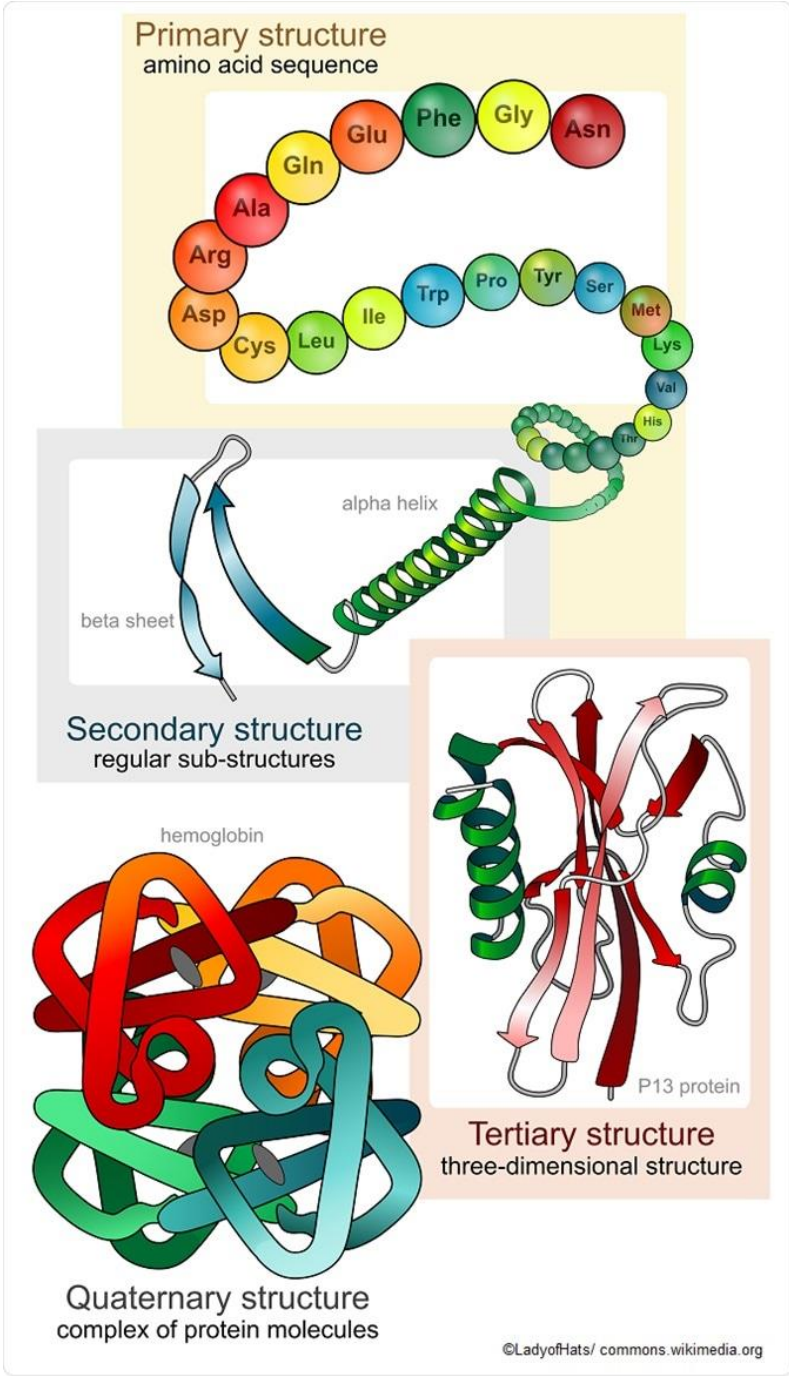
Tyrosine Tyr Y



Tryptophan Trp W



Structure of the Protein



Different proteins have different percentages of secondary structure elements

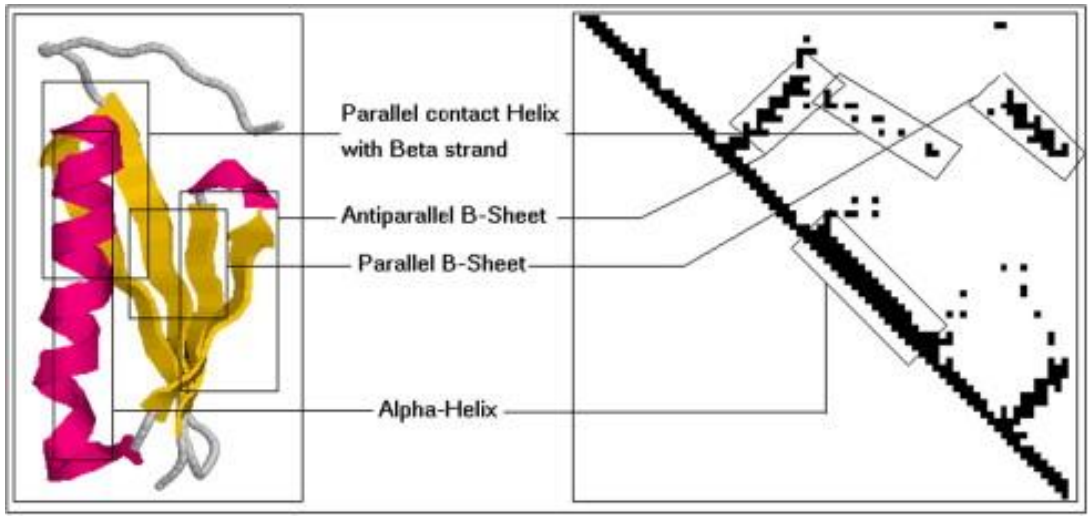
TABLE 4–4 **Approximate Proportion of α Helix and β Conformation in Some Single-Chain Proteins**

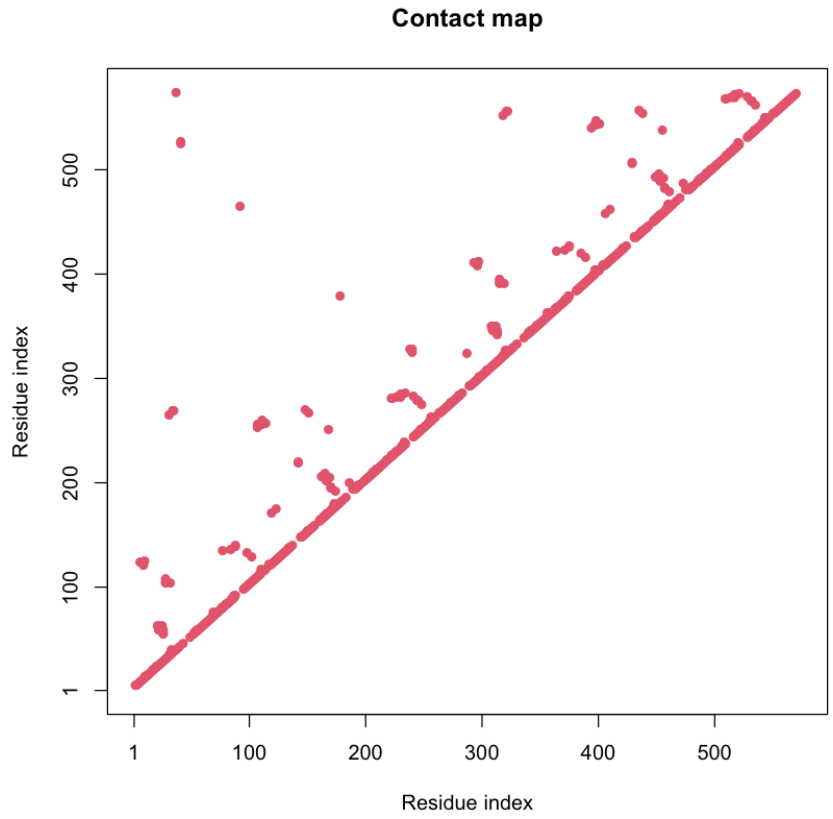
	Residuevs (%)*	
Protein (total residues)	**α Helix**	**β Conformation**
Chymotrypsin (247)	14	45
Ribonuclease (124)	26	35
Carboxypeptidase (307)	38	17
Cytochrome *c* (104)	39	0
Lysozyme (129)	40	12
Myoglobin (153)	78	0

Source: Data from Cantor, C.R. & Schimmel, P.R. (1980) *Biophysical Chemistry,* Part I: *The Conformation of Biological Macromolecules*, p. 100, W. H. Freeman and Company, New York.

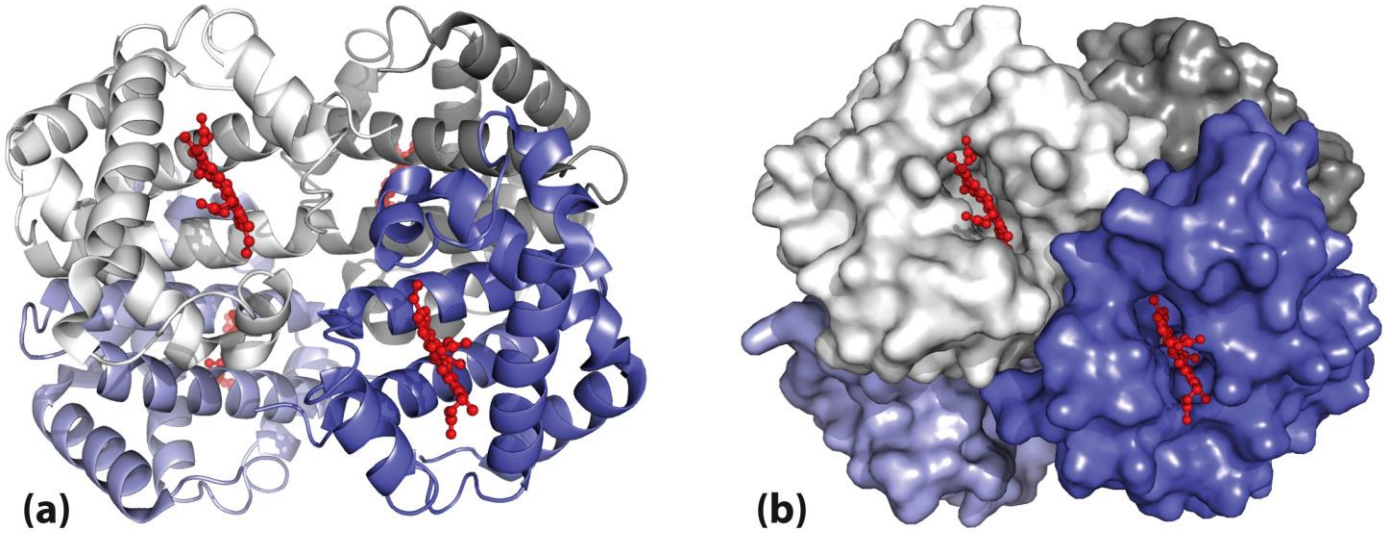
*Portions of the polypeptide chains not accounted for by α helix or β conformation consist of bends and irregularly coiled or extended stretches. Segments of α helix and β conformation sometimes deviate slightly from their normal dimensions and geometry.

Protein Contact Map





Quaternary Structure of Hemoglobin



- Hemoglobin (M_w 64,500) contains four polypeptide chains with one heme group each.
- The protein portion, globin, consists of two α chains (141 residues each) and two ß chains (146 residues each).
- The subunits of hemoglobin are arranged in symmetric pairs, each pair having one α and one ß subunit. Hemoglobin can therefore be described either as a tetramer or as a dimer of αß protomers.

Protein denaturation

- Desaturating agent/factors:
  - Heat
  - Acids
  - Bases
  - Salts
  - Mechanical agitation
  - Mutations

  (interference with folding)

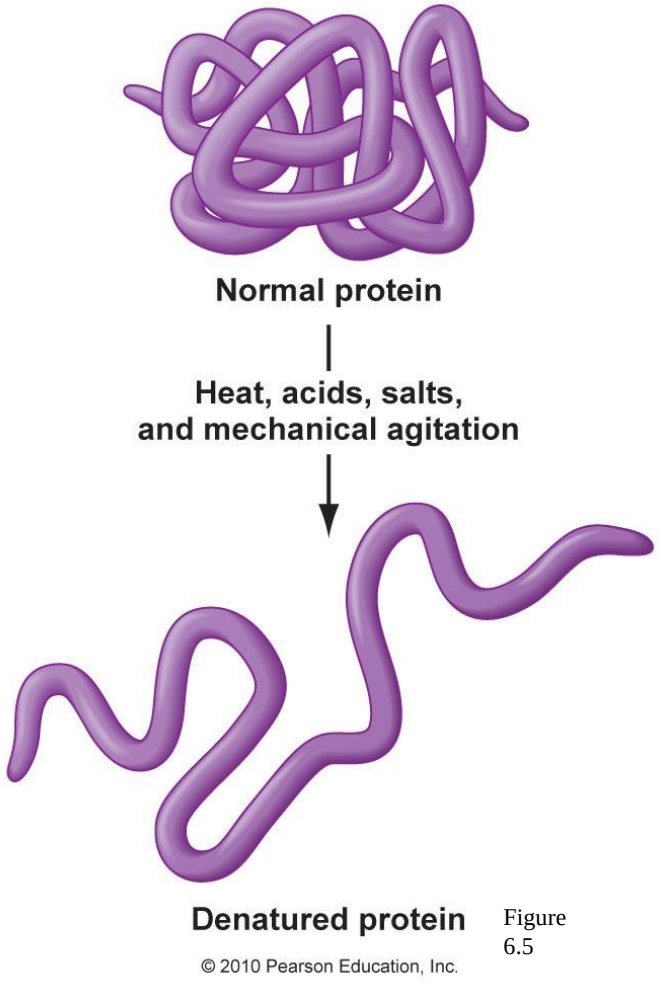
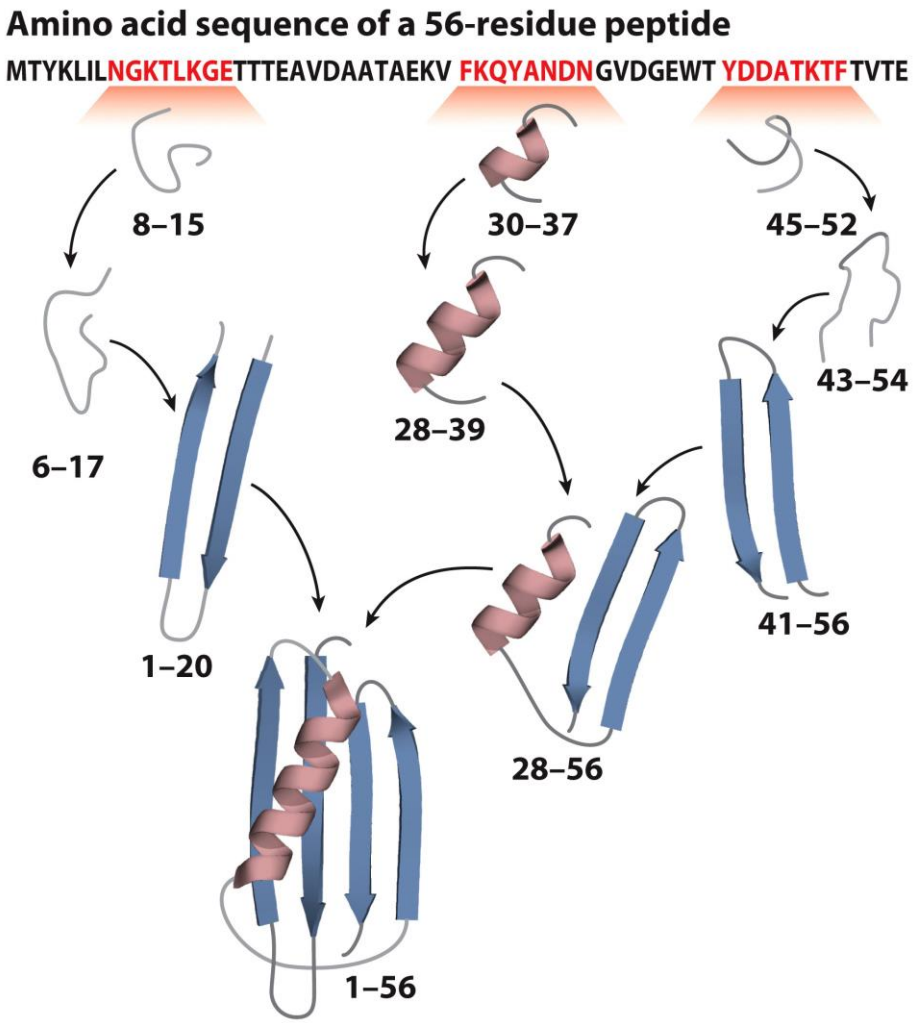
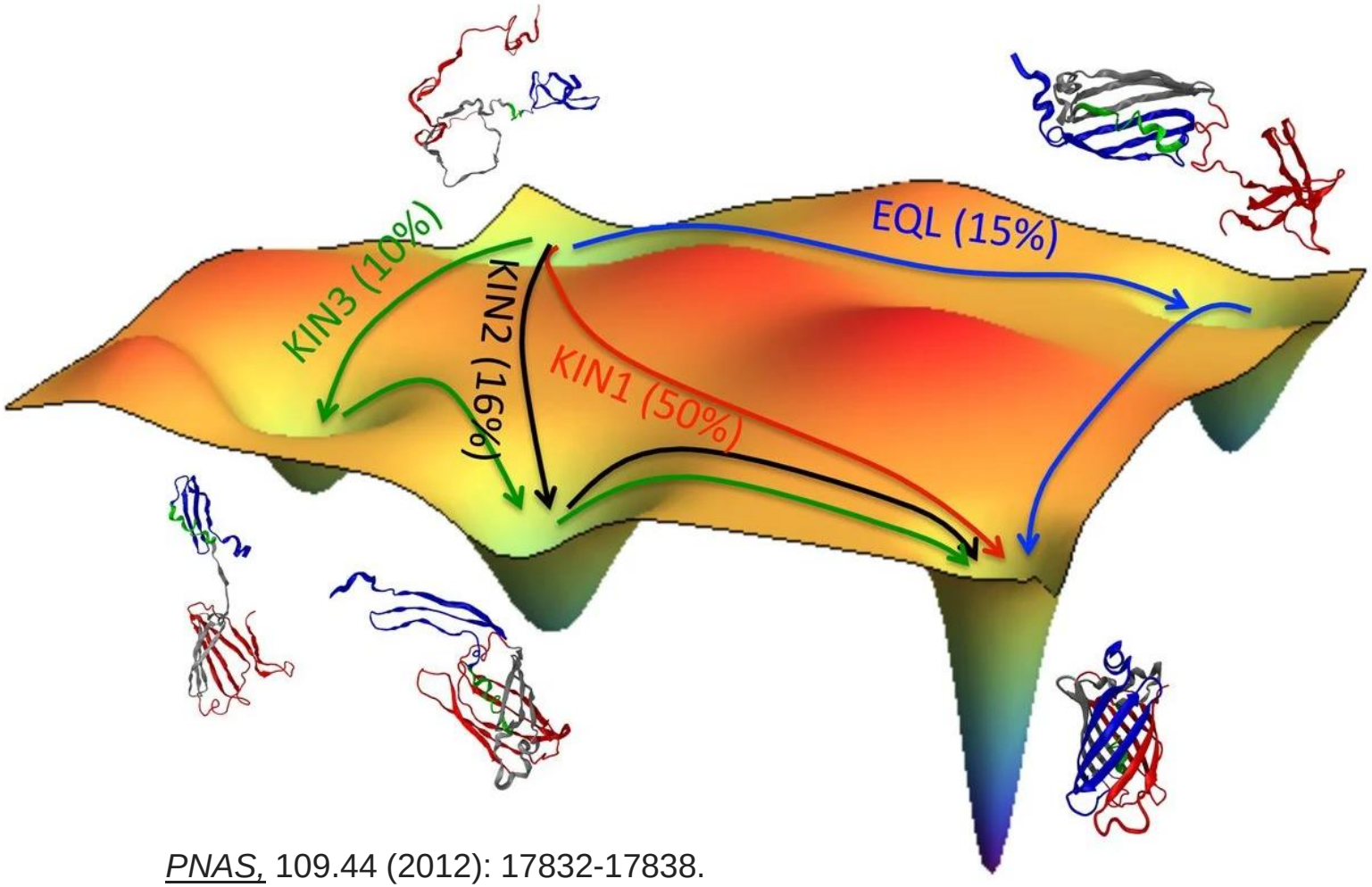


Figure 6.5

Overview of Protein Folding



Protein Folding Intermediates



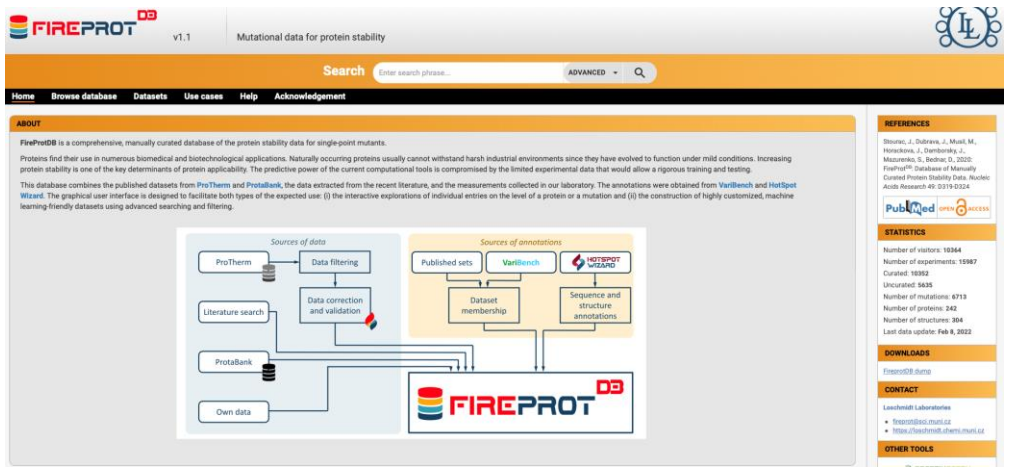
PNAS, 109.44 (2012): 17832-17838.

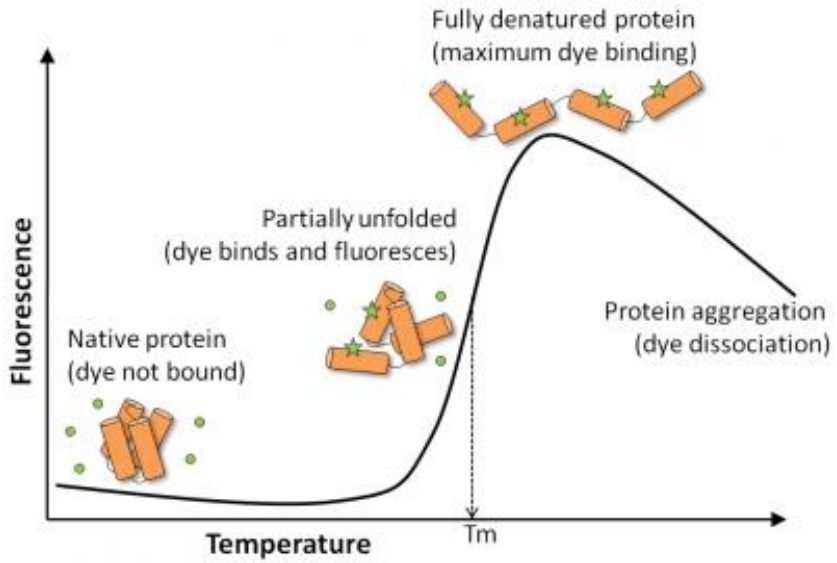
Mutational Effects on Protein Stability
(FireProt database)

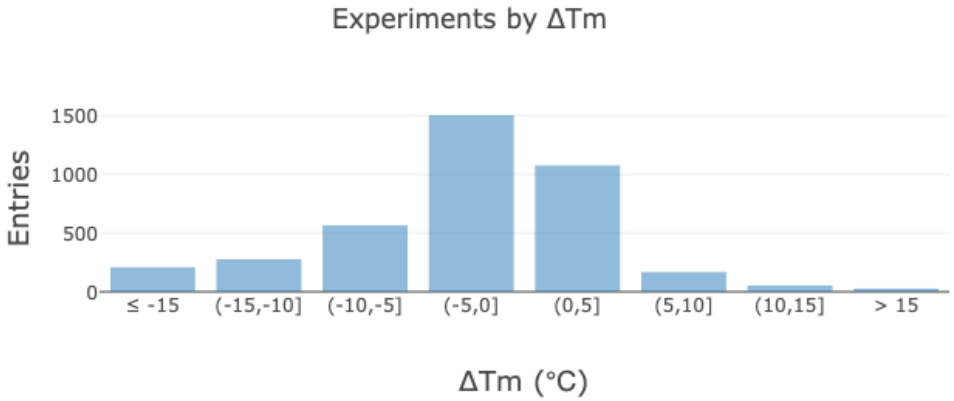
What does an average mutation do to a protein ?!

Mutational Effects on Protein Stability

(FireProt database)







What does an average mutation do to a protein ?!

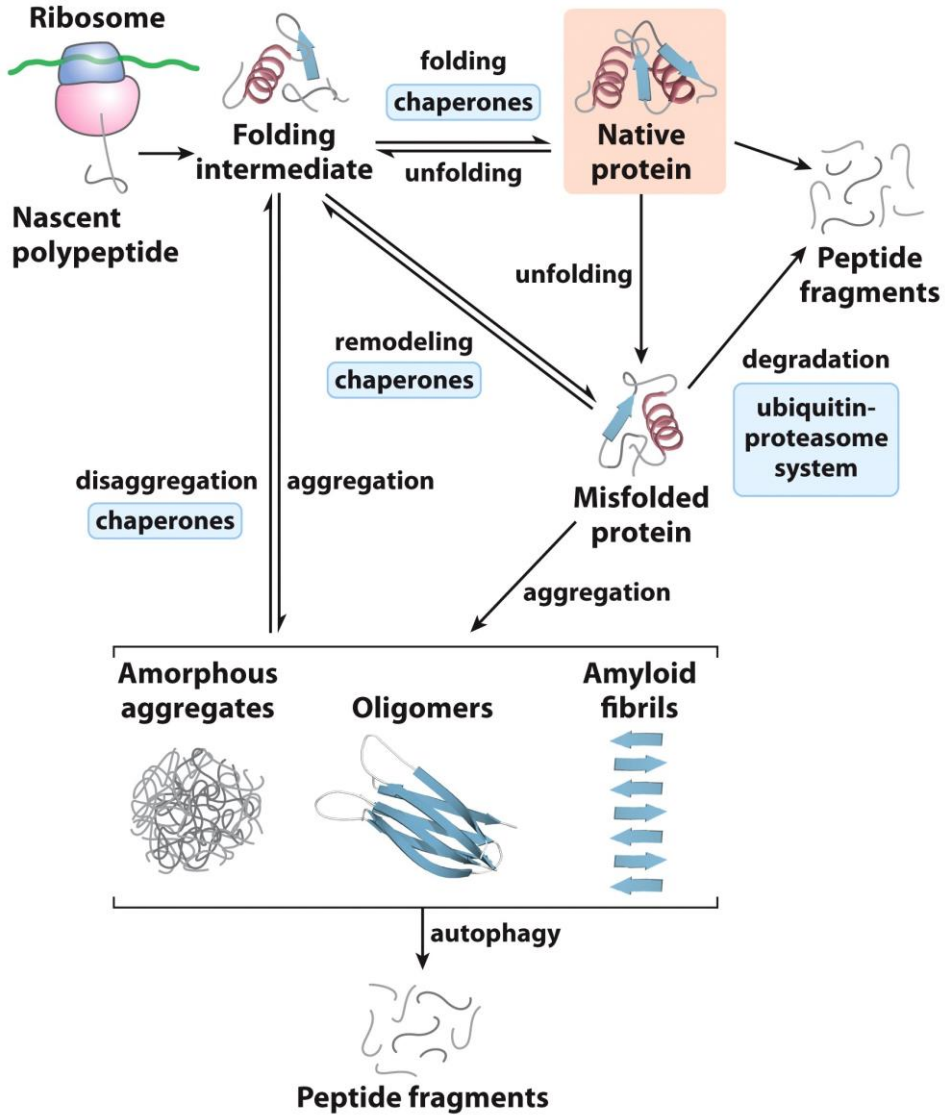
Proteostasis

The maintenance of an active set of cellular proteins.

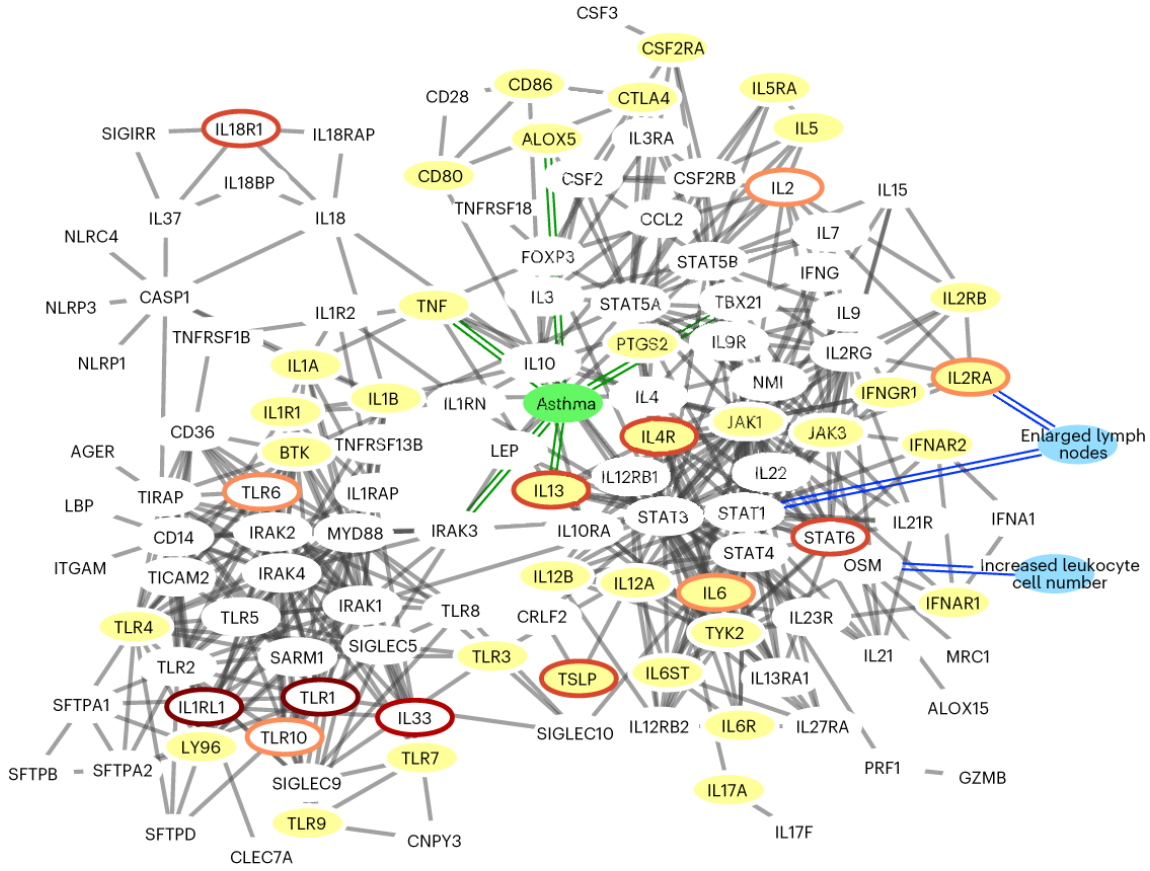
First, proteins are synthesized on a ribosome.

Second, multiple pathways contribute to protein folding, many of which involve the activity of complexes called chaperones. Chaperones (including chaperonins) also contribute to the refolding of proteins that are partially and transiently unfolded.

Third, proteins that are irreversibly unfolded are subject to sequestration and degradation by several additional pathways



Proteins work together!



Proteins “work together” forming multi complexes to carry out the specific functions

Nature genetics, *55*(3), 389-398.

STRING

Welcome to STRING

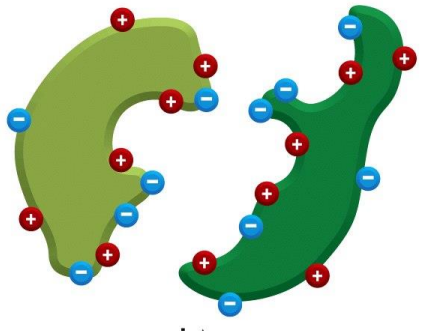
Protein-Protein Interaction Networks

Functional Enrichment Analysis

ORGANISMS	PROTEINS	INTERACTIONS
12535	59.3 mio	>20 bln

SEARCH

Protein structure affects protein-protein interactions



Protein structure affects protein-protein interactions

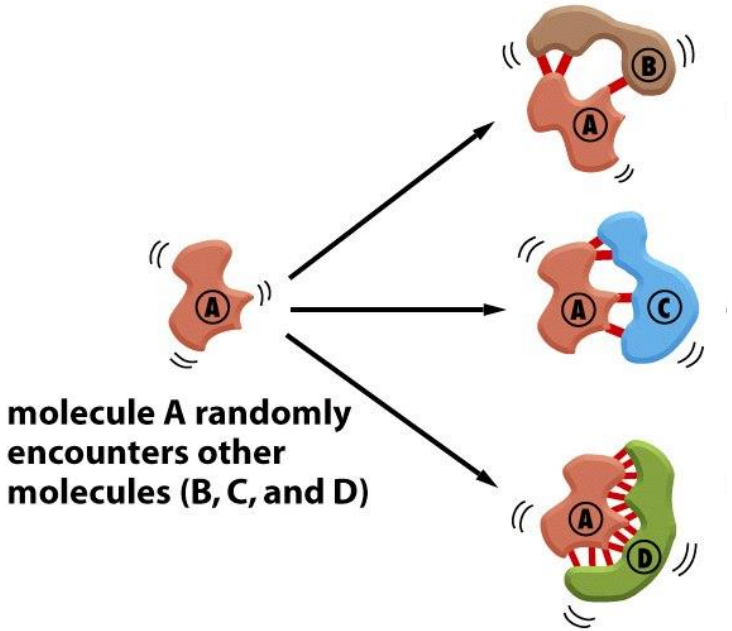
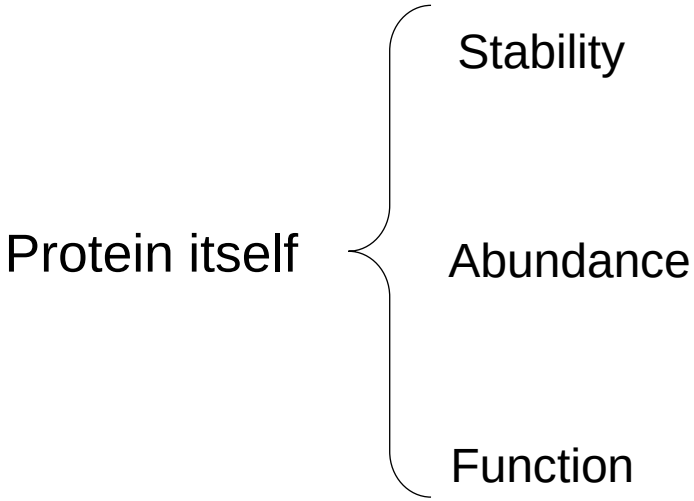


Figure 3-42 Molecular Biology of the Cell 5/e (© Garland Science 2008)

Disease variant's effect on proteins



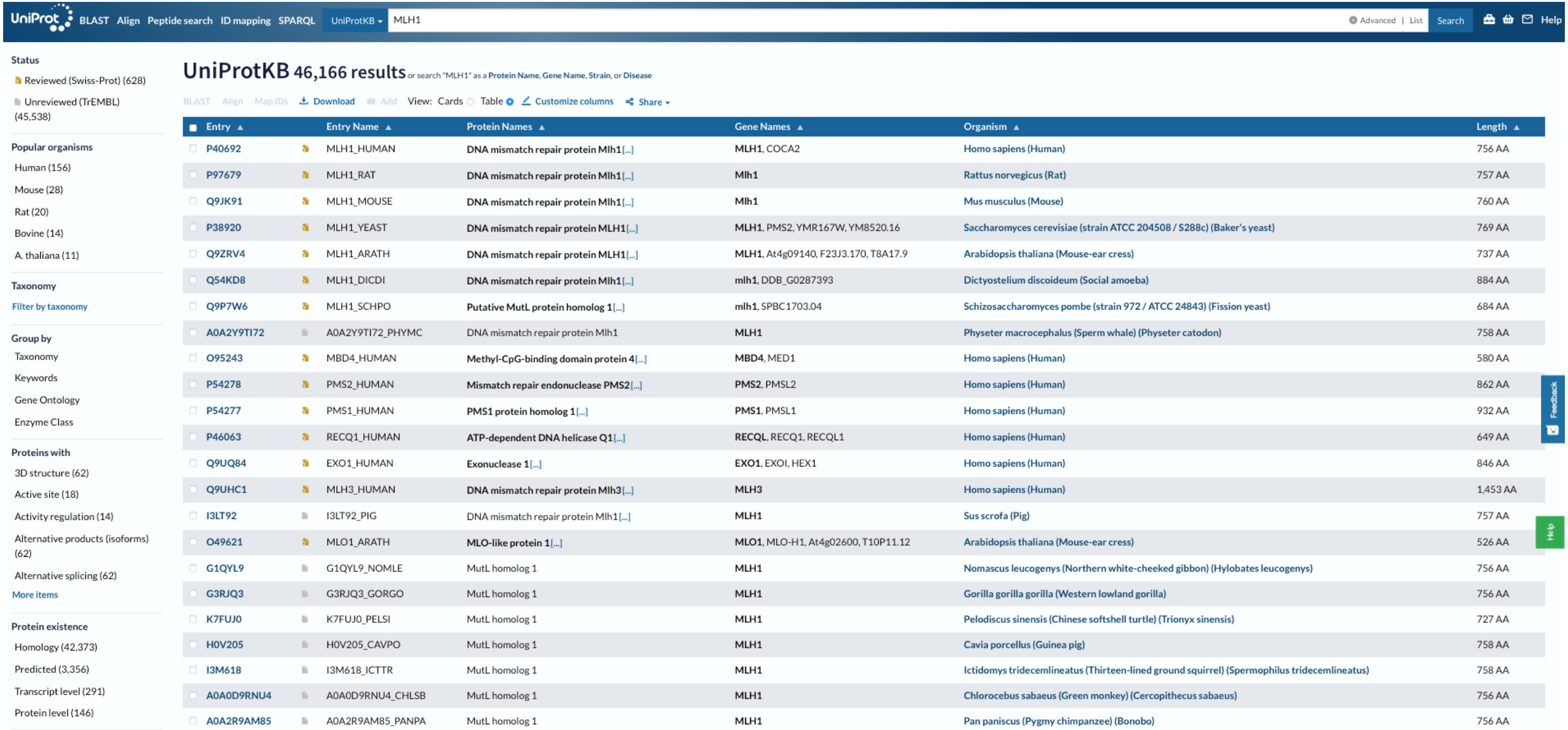
Through interactions
(including regulatory effects)

Part 2:

Disease variants of MLH1 (DNA mismatch repair)

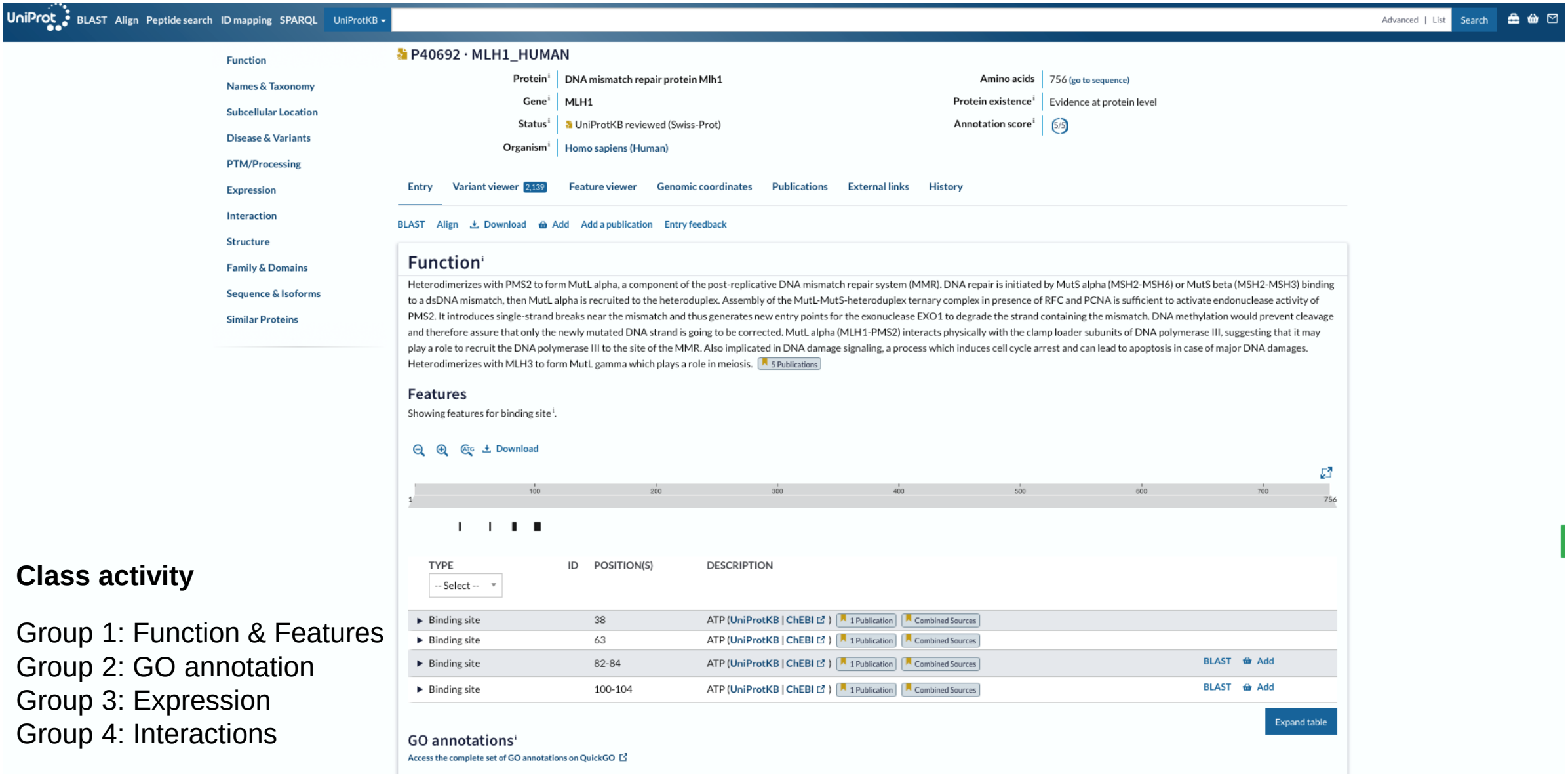
HHMI

Uniprot database



https://www.uniprot.org

Uniprot database



Class activity

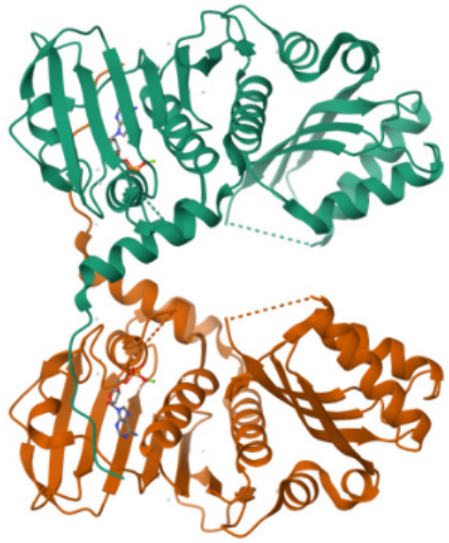
Group 1: Function & Features
Group 2: GO annotation
Group 3: Expression
Group 4: Interactions

Uniprot database: variant page

Class activity

Find the mutation P28L

Biological Assembly 1



Explore in 3D: Structure | Sequence Annotations | Electron Density | Validation Report | Ligand Interaction (ADP)

Global Symmetry: Cyclic - C2 (Explore in 3D)
Global Stoichiometry: Homo 2-mer - A2

Find Similar Assemblies

Biological assembly 1 generated by PISA (software)

Display Files | Download Files | Data API

4P7A

Crystal Structure of human MLH1

PDB DOI: https://doi.org/10.2210/pdb4P7A/pdb Entry: 4P7A **supersedes:** 3NA3

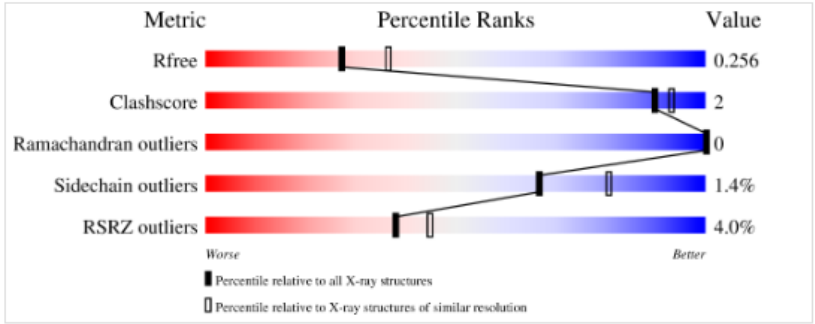
Classification: DNA BINDING PROTEIN
Organism(s): Homo sapiens
Expression System: Escherichia coli BL21(DE3)
Mutation(s): No

Deposited: 2014-03-26 **Released:** 2014-04-09
Deposition Author(s): Tempel, W., Lam, R., Zeng, H., Walker, J.R., Loppnau, P., Bountra, C., Arrowsmith, C.H., Edwards, A.M., Min, J., Wu, H., Structural Genomics Consortium (SGC)

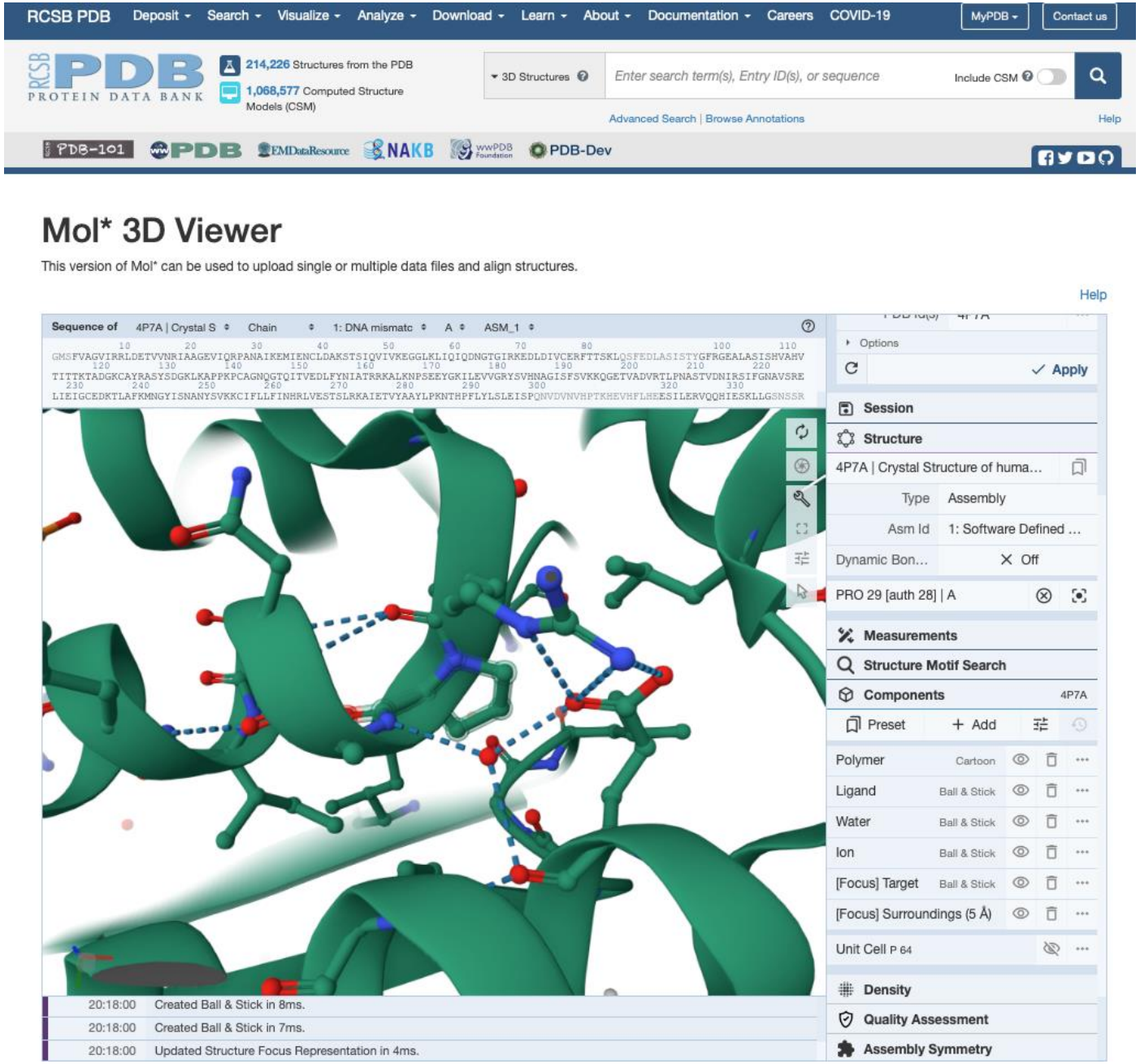
Experimental Data Snapshot

Method: X-RAY DIFFRACTION
Resolution: 2.30 Å
R-Value Free: 0.254
R-Value Work: 0.203
R-Value Observed: 0.205

wwPDB Validation 3D Report | Full Report



Ligand Structure Quality Assessment



Disease variant:
P28L

Uniprot database: variant page

Class activity

Group 1: 1-70
Group 2: 71-140
Group 3: 141-210
Group 4: 210-280
Group 5: > 280

Look at pathological/clinical variants that fall in these position bins

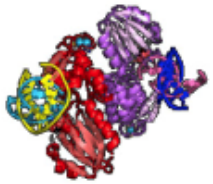
PDBsum database (EMBL-EBI)

EMBL-EBI | Services | Research | Training | About us

PDBsum

Pictorial database of 3D structures in the Protein Data Bank

- Browse options:
  - List of PDB codes
  - Het Groups
  - Ligands
  - Drugs
  - Enzymes
- Generate
- Figures from Papers
  - Gallery
  - Figure stats
- Documentation
- Downloads
- Acknowledgements
- Contact us



Databases > Structure Databases > PDBsum

PDBsum is a pictorial database that provides an at-a-glance overview of the contents of each 3D structure deposited in the Protein Data Bank (PDB). It shows the molecule(s) that make up the structure (ie protein chains, DNA, ligands and metal ions) and schematic diagrams of the interactions between them.
NEW Also included are the AlphaFold predicted models for all human protein for comparison with experimentally determined structures. Read more ...

PDB code *(4 chars)* Find Example: "1kfv"

NEW **Alpha Fold model** *(human proteins only)* Find see *Analyses*

Enter UniProt accession (or UniProt id), to find Alpha Fold model of given protein. Eg P00734 (THRB_HUMAN).

Text search Search

Scans all TITLE, HEADER, COMPND, SOURCE and AUTHOR records in the PDB (eg to find a given protein by name).

Search by sequence

Search

Perform FASTA search vs all sequences in the PDB to get a list of the closest matches.

Search by

UniProt id: *(eg P03023, LACI_ECOLI, etc)* Search

Pfam id: *(eg PF07992)* Search

Ensembl id: *(eg ENSG00000086205, ENST00000256999)* Search

Contents

PDBsum contains 206,449 entries, including 1,781 superseded
Last update: 10 April, 2023

In-house version

Download PDBsum *1* ro process your own structures in-house

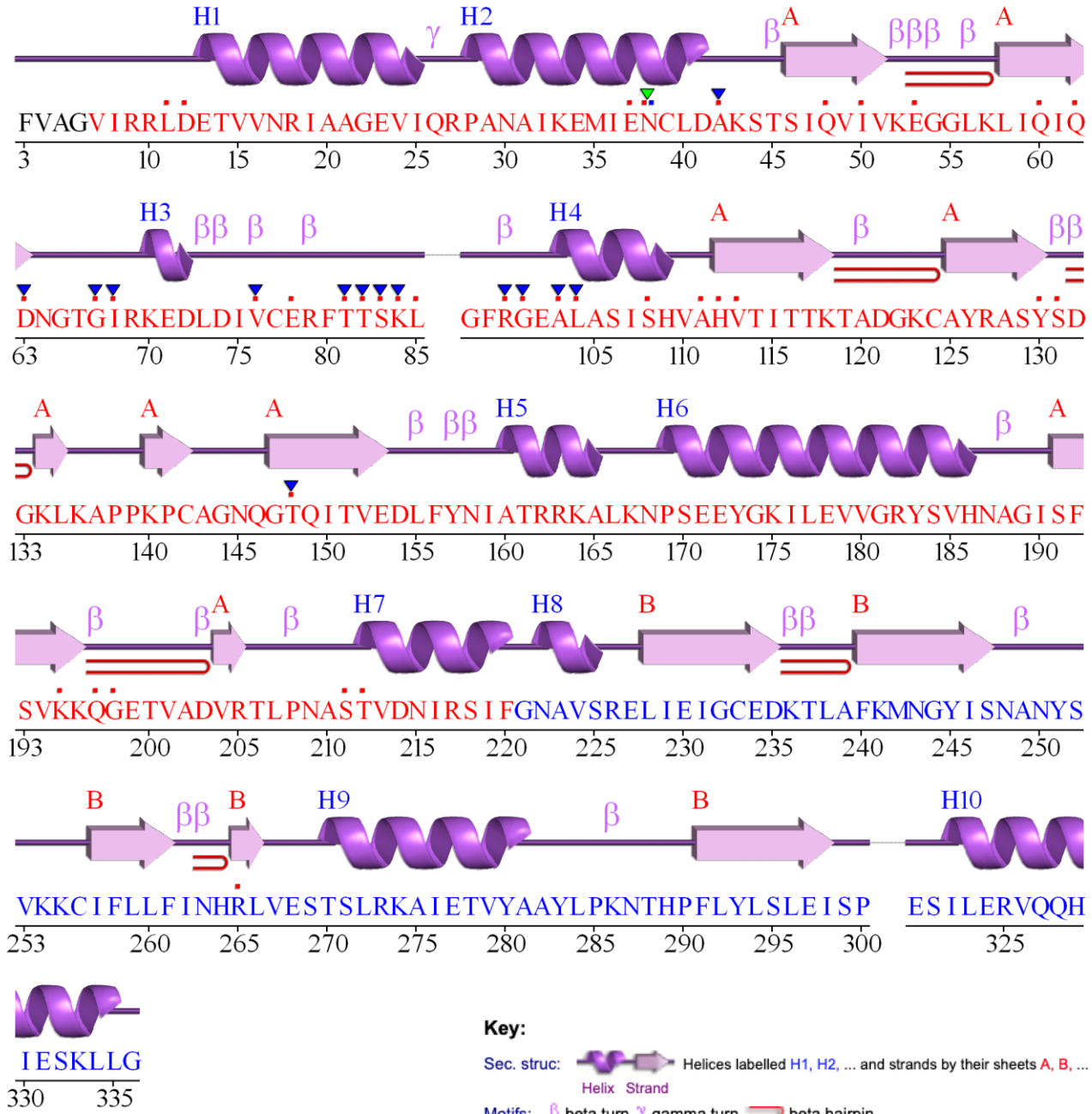
Related databases

EC-PDB

Enzyme 3D structures organized by the E.C. numbering hierarchy.

DrugPort

Structures of drugs and their target proteins in the PDB.


H1 γ H2 β A βββ β A
FVAGVIRRLDETVVNRIAAGEVIQRPANAIKEMIENCLDAKSTSIQVIVKEGGLKLIQIQ
3 10 15 20 25 30 35 40 45 50 55 60
H3 ββ β β β H4 A β A ββ
DNGTGIRKEDLDIVCERFTTSKL GFRGEALASISHVAHVTITTKTADGKCAYRASYSD
63 70 75 80 85 105 110 115 120 125 130
A A A β ββ H5 H6 β A
GKLKAPPKPCAGNQGTQITVEDLFYNIATRRKALKNPSEEYGKILEVVGRYSVHNAGISF
133 140 145 150 155 160 165 170 175 180 185 190
β β A β H7 H8 B ββ B β
SVKKQGETVADVRTLPNASTVDNIRSIFGNAVSRELIEIGCEDKTLAFKMNGYISNANYS
193 200 205 210 215 220 225 230 235 240 245 250
B ββ B H9 β B H10
VKKCIFLLFINHRLVESTSLRKAIETVYAAYLPKNTHPFLYLSLEISP ESILERVQQH
253 260 265 270 275 280 285 290 295 300 325
IESKLLG
330 335
Key:
Sec. struc: Helix Strand Helices labelled H1, H2, ... and strands by their sheets A, B, ...
Motifs: β beta turn γ gamma turn beta hairpin
Residue contacts: to ligand to metal
PDB SITE records: AC1 AC2

Uniprot database: variant page

Look at pathological/clinical variants that fall in these position bins

Did you notice any difference between the protein length (in structure) and the gene length?

Iupred database

AlphaFold (EMBL)

AlphaFold Protein Structure Database

Home About FAQs Downloads API

Search for protein, gene, UniProt accession or organism or sequence search BETA Search

Examples: MENFQKVEKIGEGTYGV... Free fatty acid receptor 2 At1g58602 Q5VSL9 E. coli See search help

DNA mismatch repair protein Mlh1

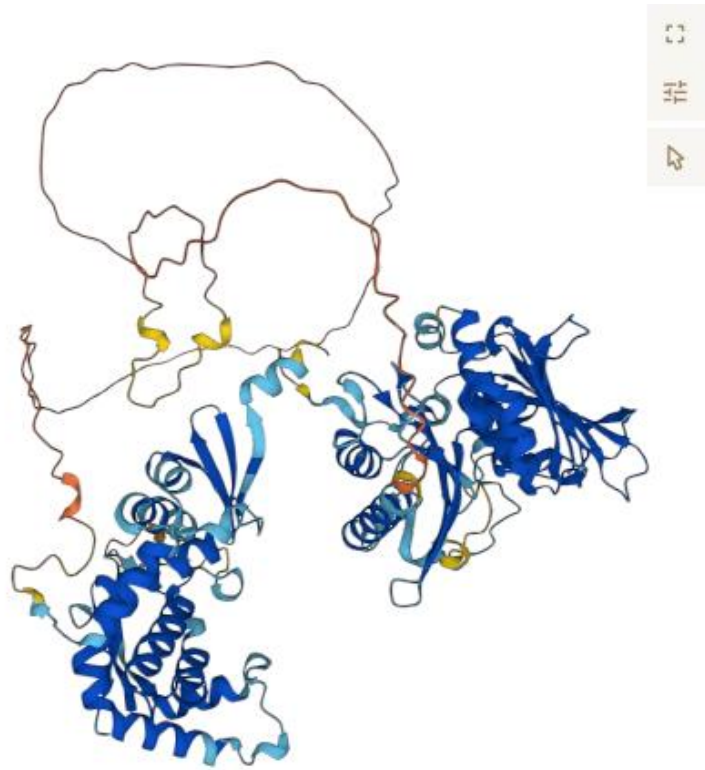
AlphaFold structure prediction

Download PDB file mmCIF file Predicted aligned error

Share your feedback on structure with DeepMind Looks great Could be improved

Information

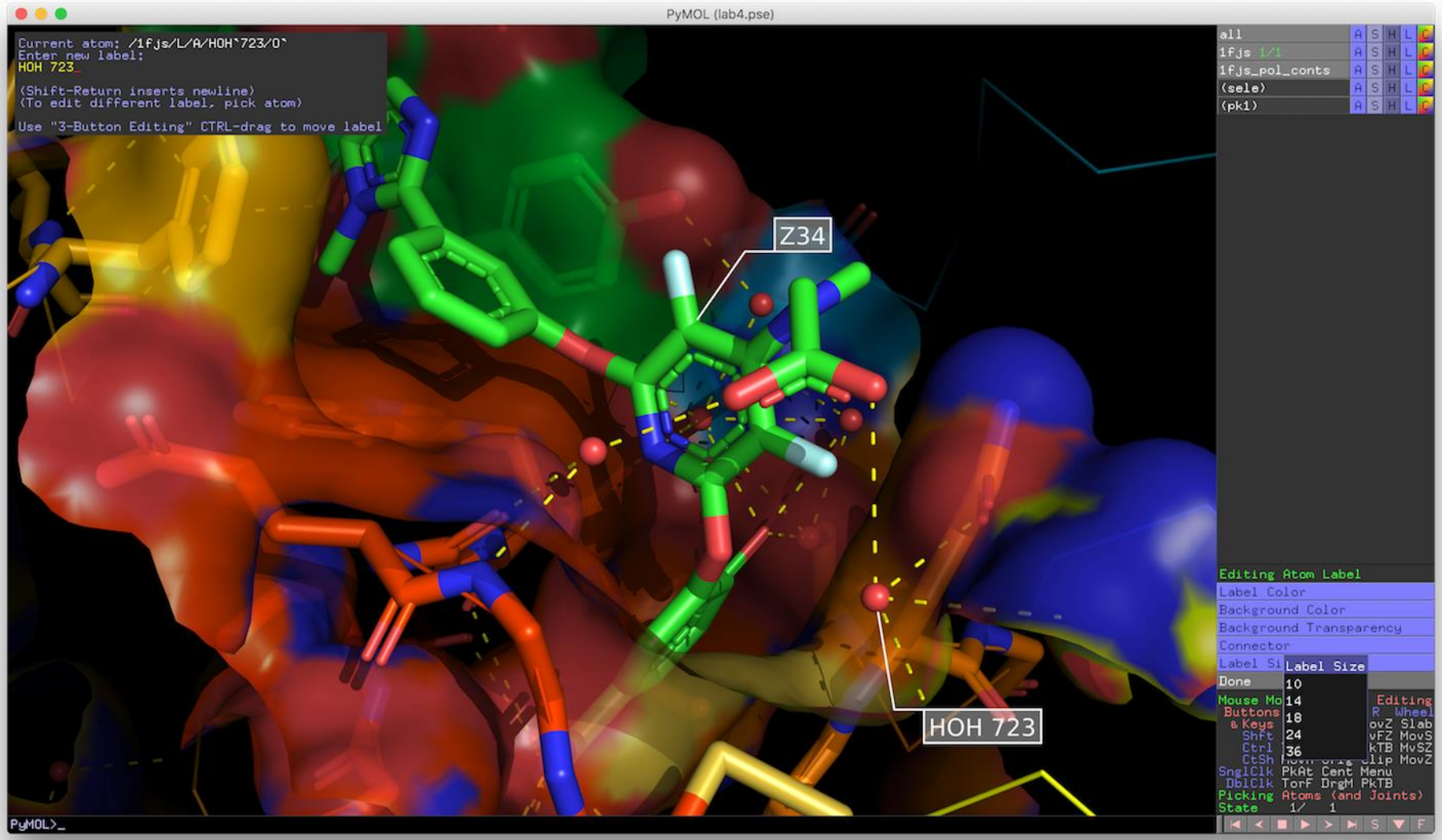
Protein	DNA mismatch repair protein Mlh1
Gene	MLH1
Source organism	Homo sapiens (Human) go to search
UniProt	P40692 go to UniProt
Experimental structures	7 structures in PDB for P40692 go to PDBe-KB
Biological function	Heterodimerizes with PMS2 to form MutL alpha, a component of the post-replicative DNA mismatch repair system (MMR). DNA repair is initiated by MutS alpha (MSH2-MSH6) or MutS beta (MSH2-MSH3) binding to a dsDNA mismatch, then MutL alpha is recruited to the heteroduplex. Assembly of the MutL-MutS-heteroduplex ternary complex in presence of RFC and PCNA is sufficient to activate endonuclease activity of PMS2. It introduces single-strand breaks near the mismatch and ... [show more] go to UniProt



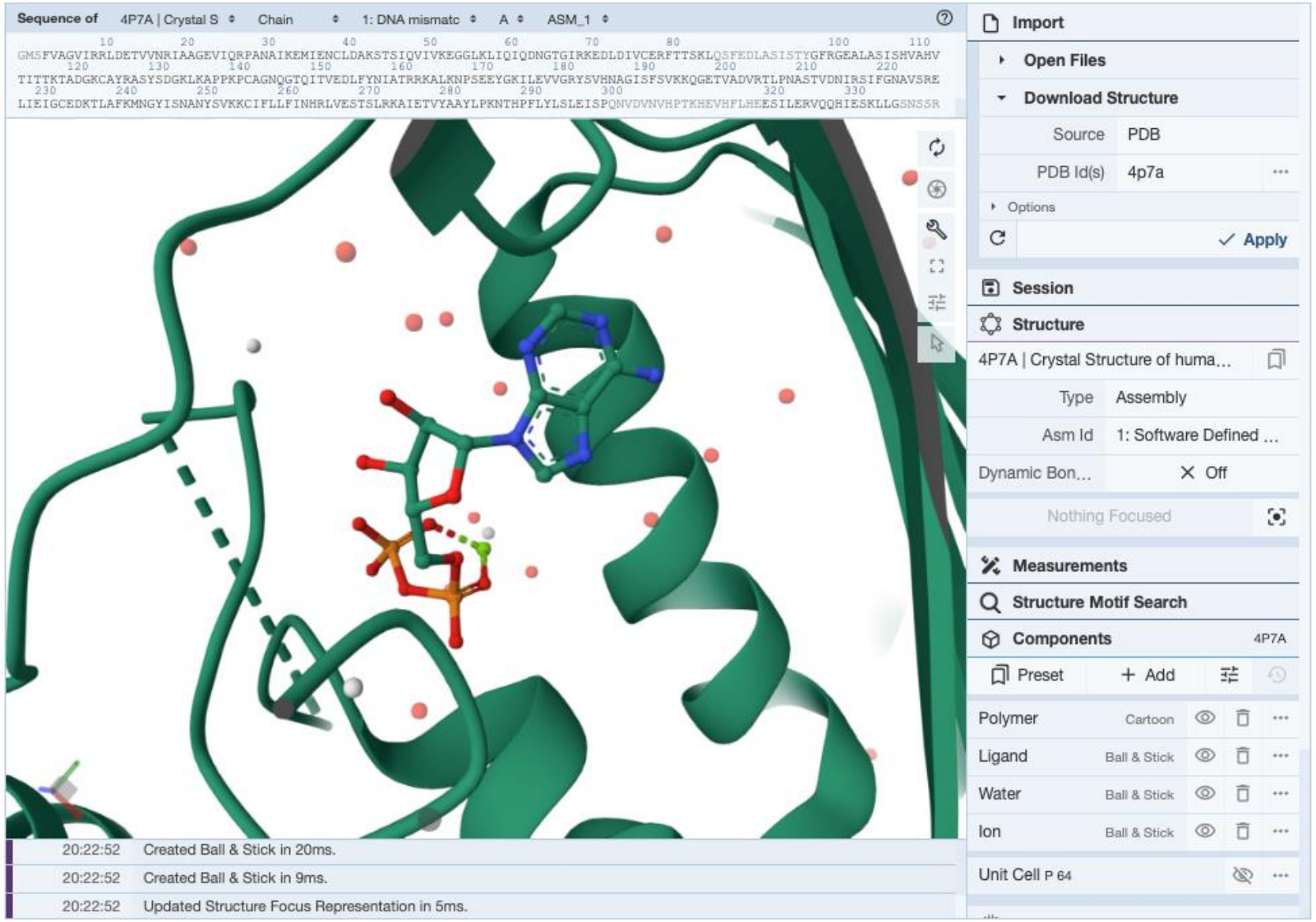
https://alphafold.ebi.ac.uk

What clinical variants fall in the disordered regions?

Pymol: Protein and Molecular Visualization



Which variants might affect the ATP binding affinity?



Cross validate your results with PDBsum (EMBL) database