

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)

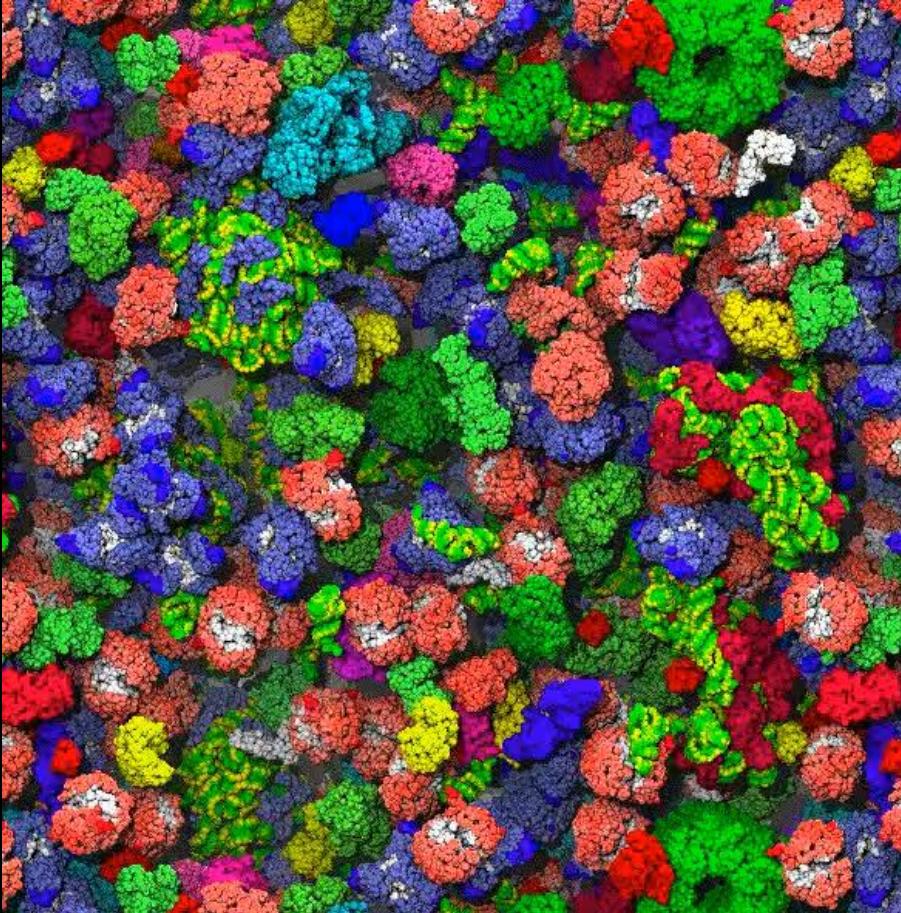
Course agenda and organizational notes

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)



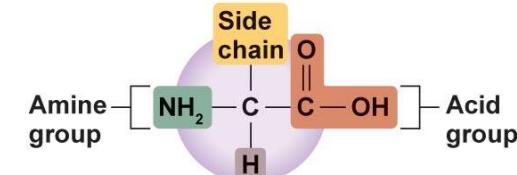
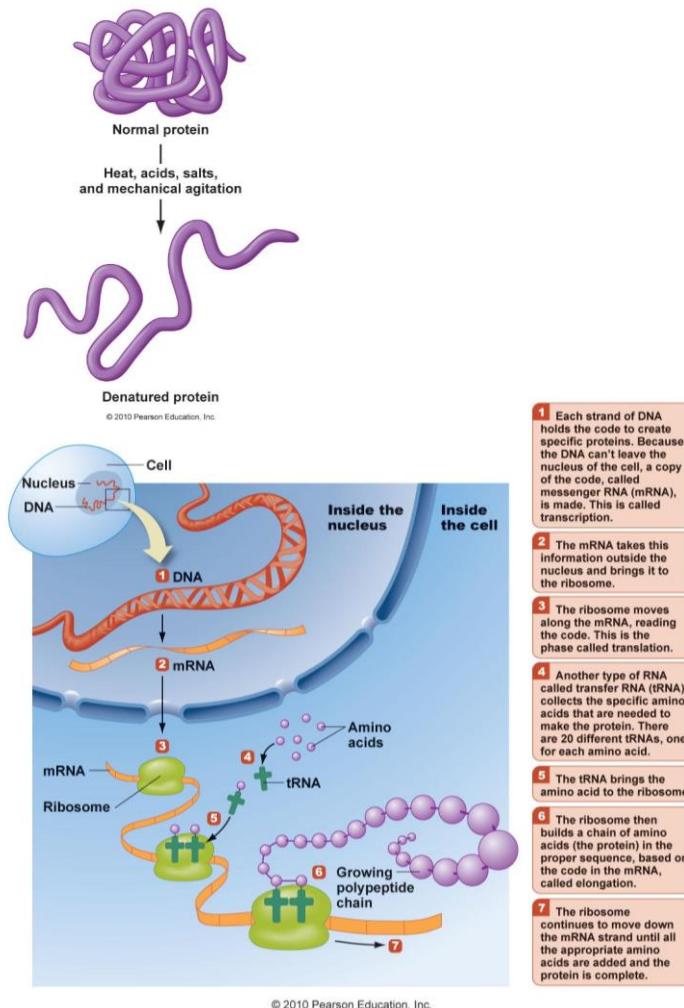
PLoS computational biology,
6(3), e1000694.

Cytoplasm's viscosity > 100 mPa.s

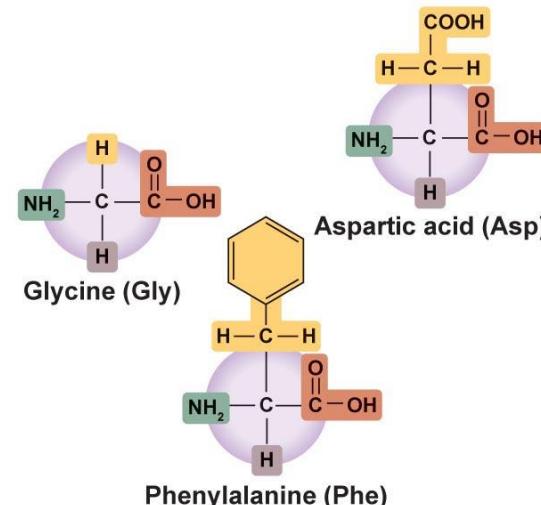
Water ~ 1 mPa.s
Engine oil ~ 100 mPa.s

Part 1: Protein Sequence, Structure, Function

Proteins and Amino Acids



a Amino acid structure

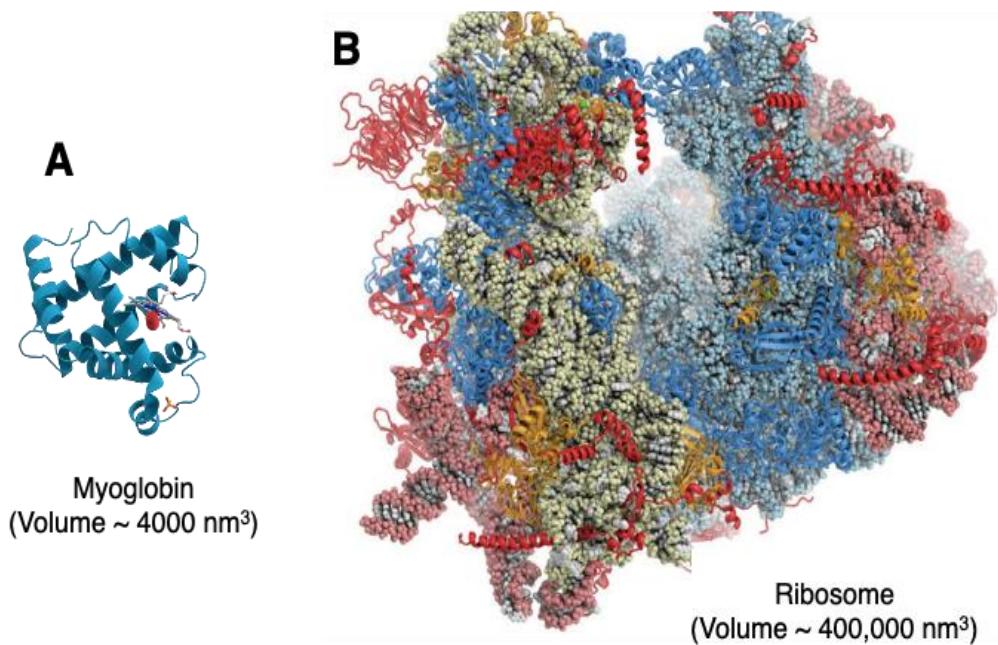


b Different amino acids, showing their unique side chains

© 2010 Pearson Education, Inc.

Proteins

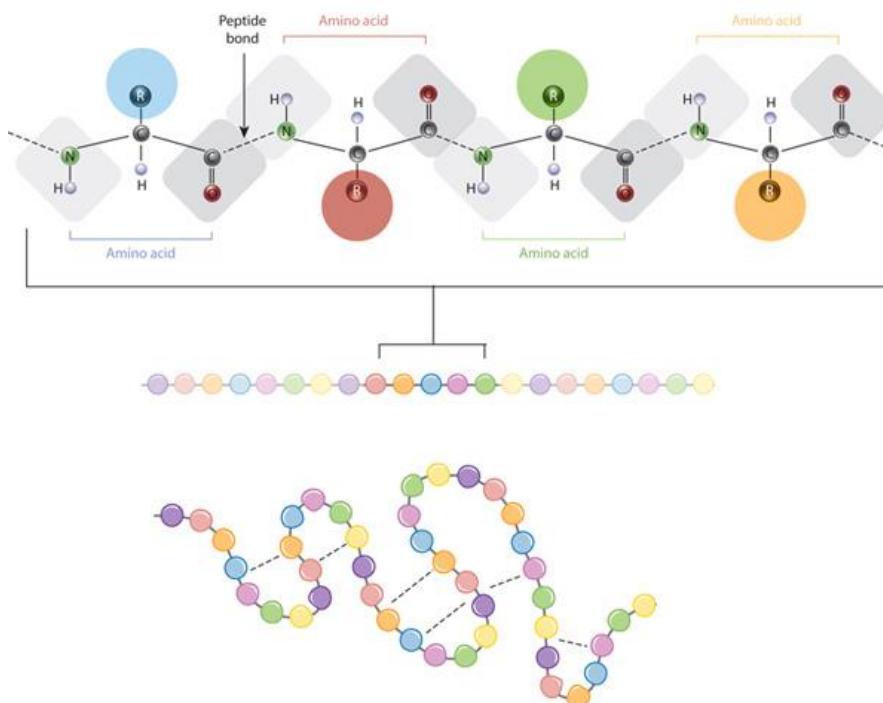
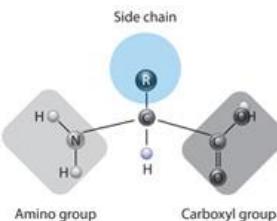
- Made up of chains of amino acids
- Are involved in most of the body's functions and life processes
- The sequence of amino acids is determined by DNA



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 synthesized 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

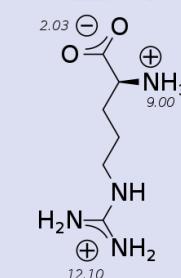
⊕ Positive
⊖ Negative

A. Amino Acids with Electrically Charged Side Chains

Positive

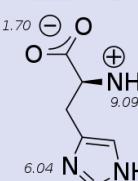
Arginine

Arg R



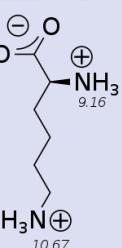
Histidine

His H



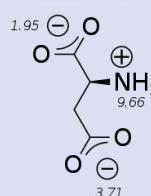
Lysine

Lys K



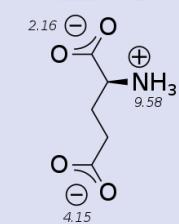
Aspartic Acid

Asp D



Glutamic Acid

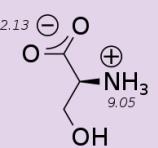
Glu E



B. Amino Acids with Polar Uncharged Side Chains

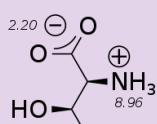
Serine

Ser S



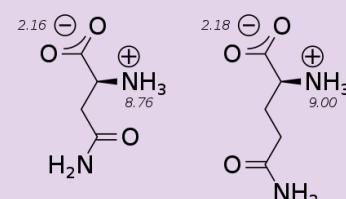
Threonine

Thr T



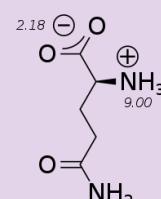
Asparagine

Asn N



Glutamine

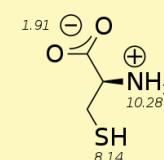
Gln Q



C. Special Cases

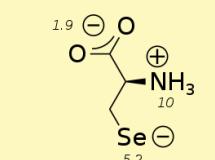
Cysteine

Cys C



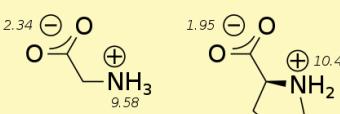
Selenocysteine

Sec U



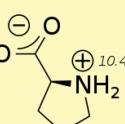
Glycine

Gly G



Proline

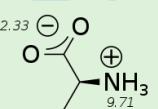
Pro P



D. Amino Acids with Hydrophobic Side Chains

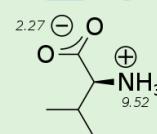
Alanine

Ala A



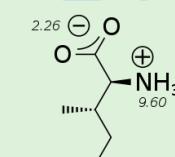
Valine

Val V



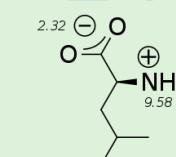
Isoleucine

Ile I



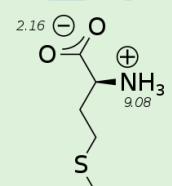
Leucine

Leu L



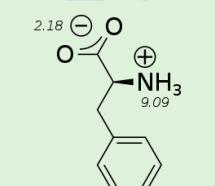
Methionine

Met M



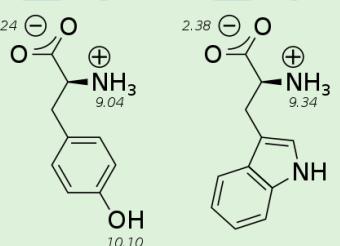
Phenylalanine

Phe F



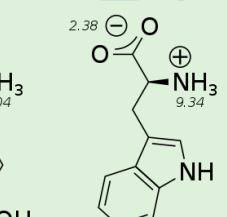
Tyrosine

Tyr Y

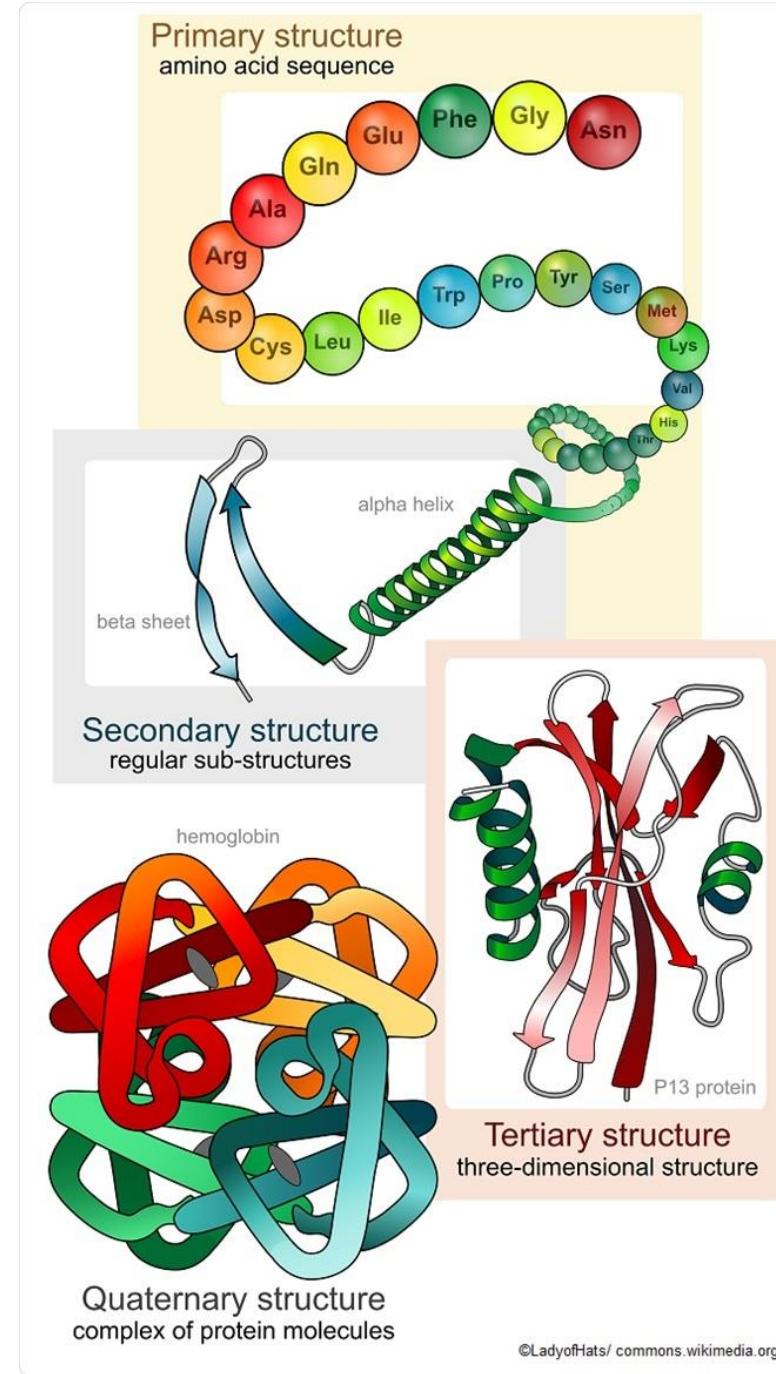


Tryptophan

Trp W



Structure of the Protein



Different proteins have different percentages of secondary structure elements

TABLE 4-4

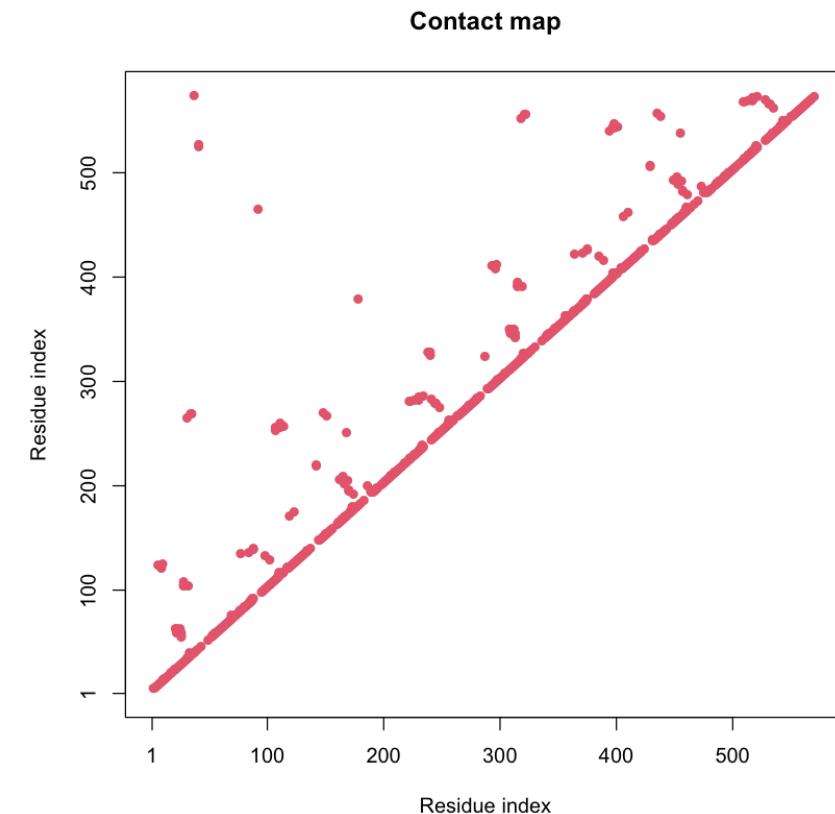
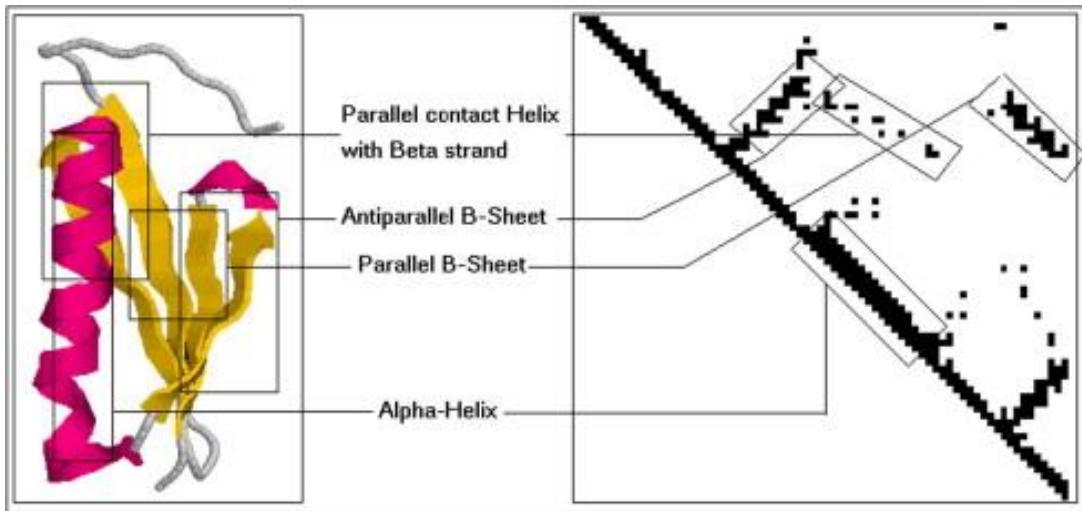
**Approximate Proportion of α Helix and
 β Conformation in Some Single-Chain
Proteins**

Protein (total residues)	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

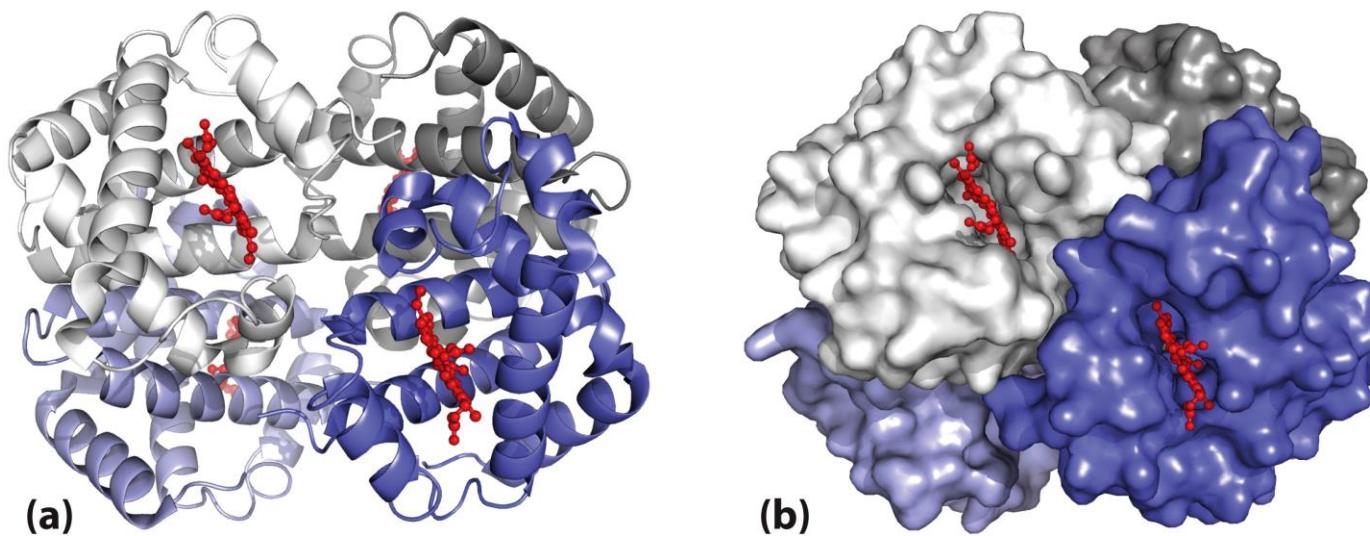
Source: Data from Cantor, C.R. & Schimmel, P.R. (1980) *Biophysical Chemistry, Part I: The Conformation of Biological Macromolecules*, p. 100, W. H. Freeman and Company, New York.

*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



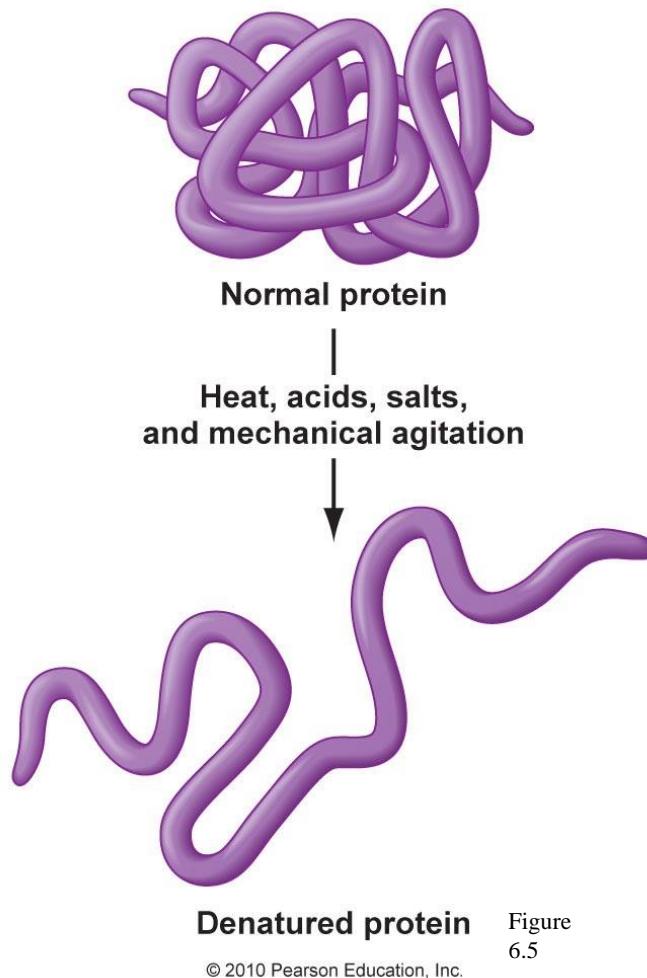
Quaternary Structure of Hemoglobin



- 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 β chains (146 residues each).
- The subunits of hemoglobin are arranged in symmetric pairs, each pair having one α and one β subunit. Hemoglobin can therefore be described either as a tetramer or as a dimer of $\alpha\beta$ protomers.

Protein denaturation

- Desaturating agent/factors:
 - Heat
 - Acids
 - Bases
 - Salts
 - Mechanical agitation
 - Mutations
(interference with folding)

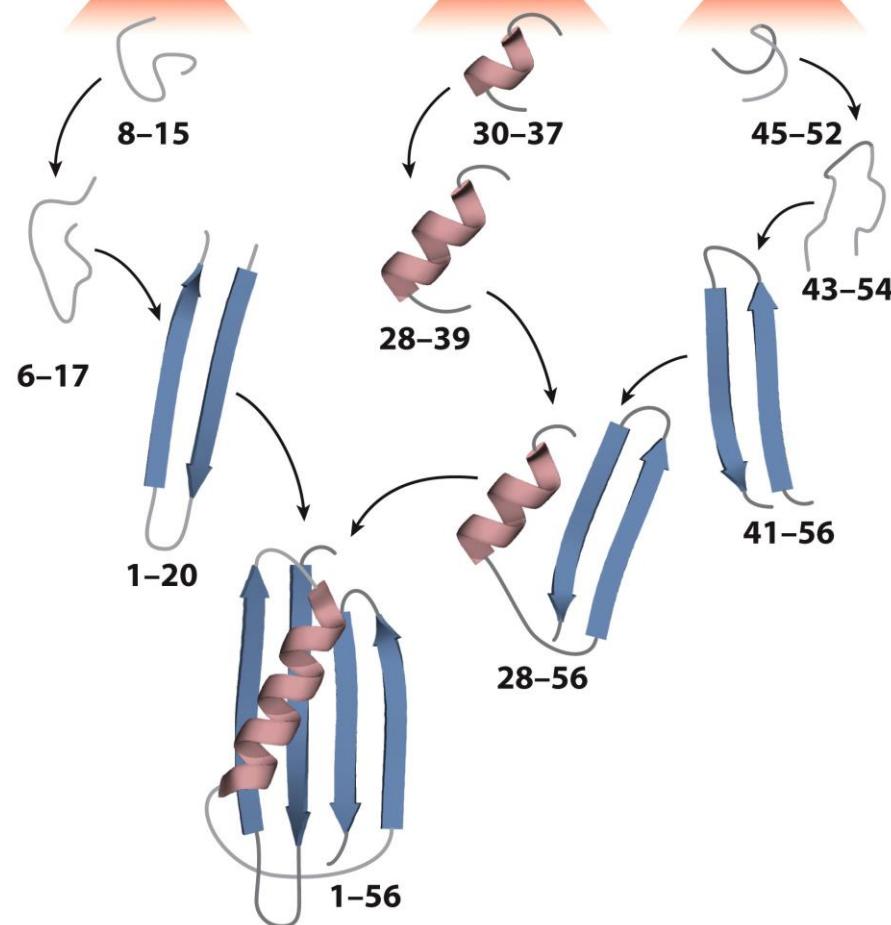


© 2010 Pearson Education, Inc.

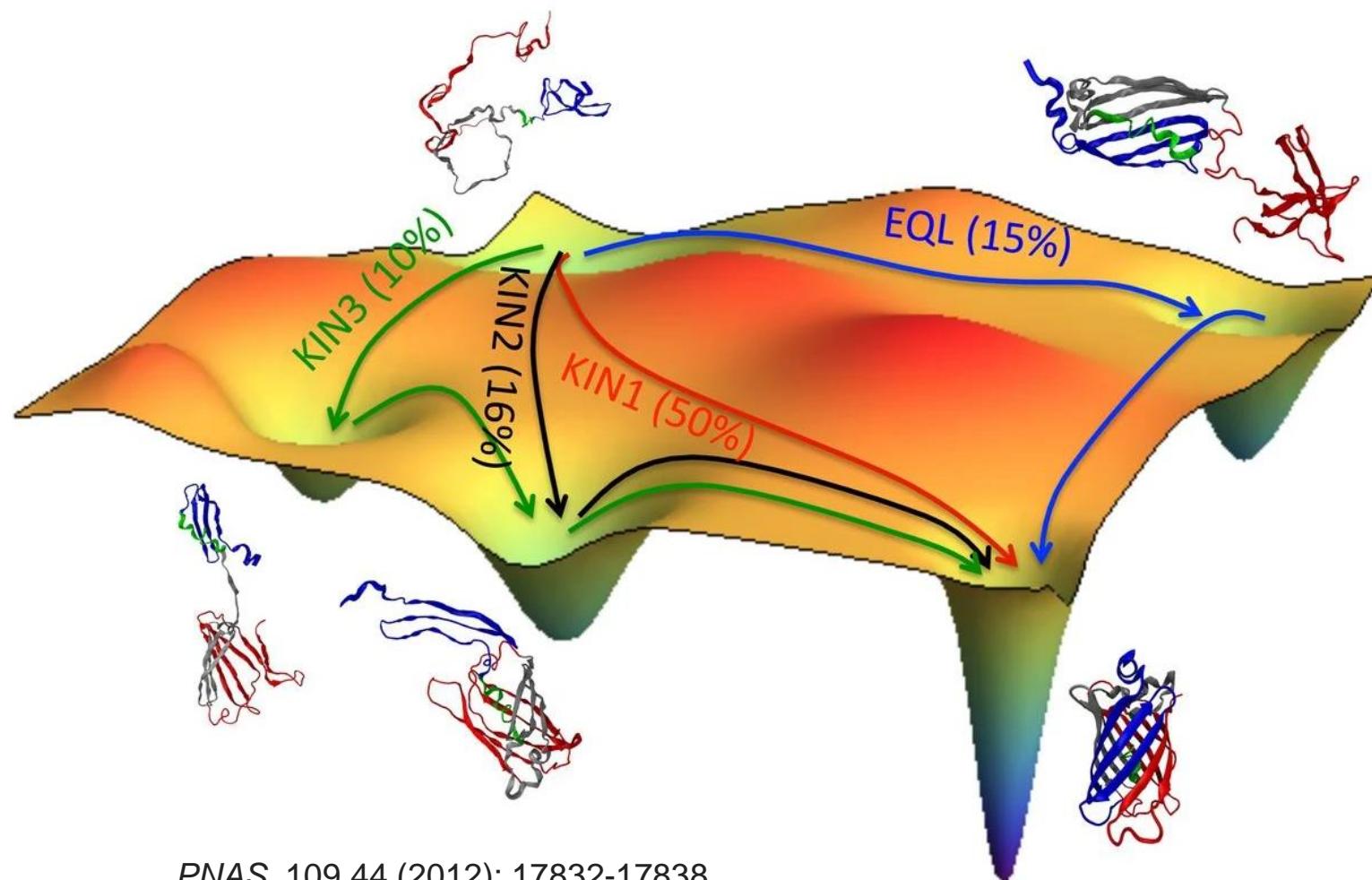
Figure
6.5

Overview of Protein Folding

Amino acid sequence of a 56-residue peptide
MTYKLIL**NGKTLKGE**TTTEAVDAATAEKV **FKQYANDN**GVDGEWT **YDDATKTF**TVTE



Protein Folding Intermediates



PNAS, 109.44 (2012): 17832-17838.

Mutational Effects on Protein Stability (FireProt database)

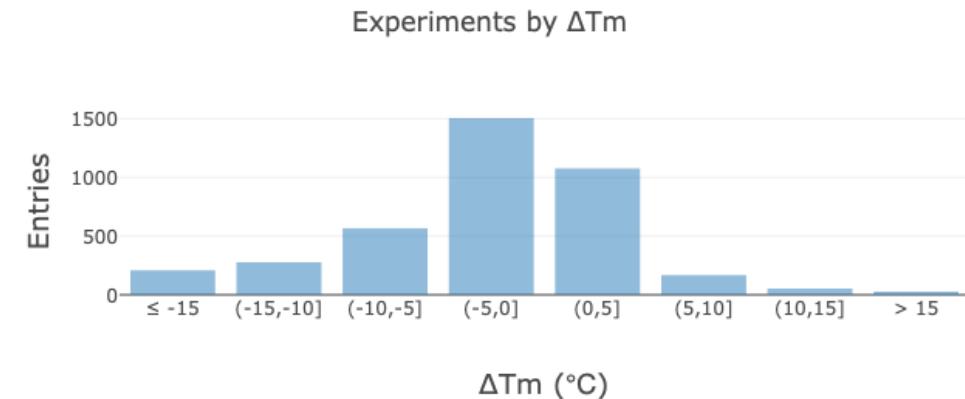
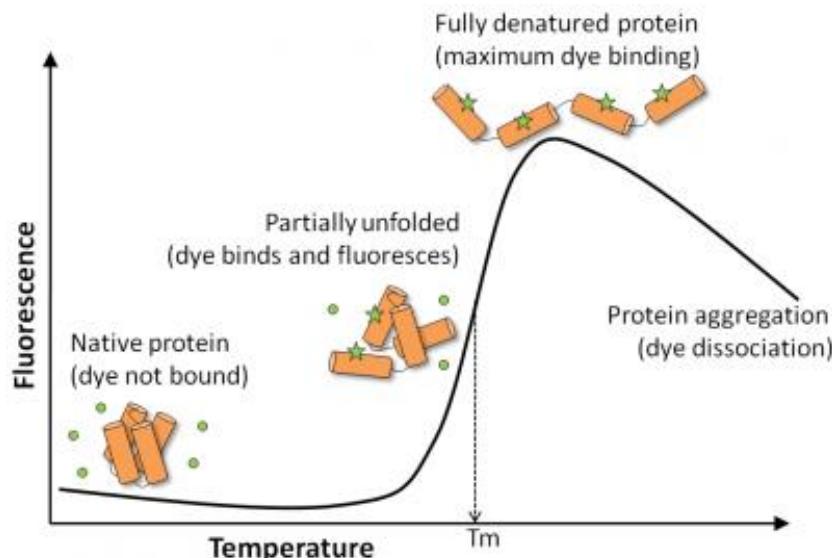
What does an average mutation do to a protein ?!

Mutational Effects on Protein Stability (FireProt database)

The screenshot shows the FireProtDB homepage. At the top, there's a search bar with 'Search' and 'Enter search phrase...' fields, and an 'ADVANCED' dropdown. Below the search bar are navigation links: Home, Browse database, Datasets, Use cases, Help, and Acknowledgement. To the right is a logo for the Institute of Biotechnology of the Czech Academy of Sciences.

The main content area has a 'ABOUT' section with a detailed description of the database's purpose and data sources. It also features a flowchart illustrating the data pipeline: Sources of data (ProTherm, Literature search, ProtBank, Own data) feed into Data filtering, which then leads to Data connection and validation. Sources of annotations (Published sets, VarBench, HotSpot Wizard) feed into Dataset membership and Sequence and structure annotations. Both paths converge into the central 'FIREPROT DB' box.

On the right side, there are sections for REFERENCES (listing authors like Strelak, Dobrevska, Masić, Horovcik, Damborsky, and Masařík), STATISTICS (number of visitors: 10344, number of experiments: 15987, curated: 10344, uncurated: 5453, number of mutations: 6713, number of proteins: 24, number of structures: 304, last data update: Feb 9, 2022), DOWNLOADS (FireProtDB data), CONTACT (Loděnice Laboratories), and OTHER TOOLS.



What does an average mutation do to a protein ?!

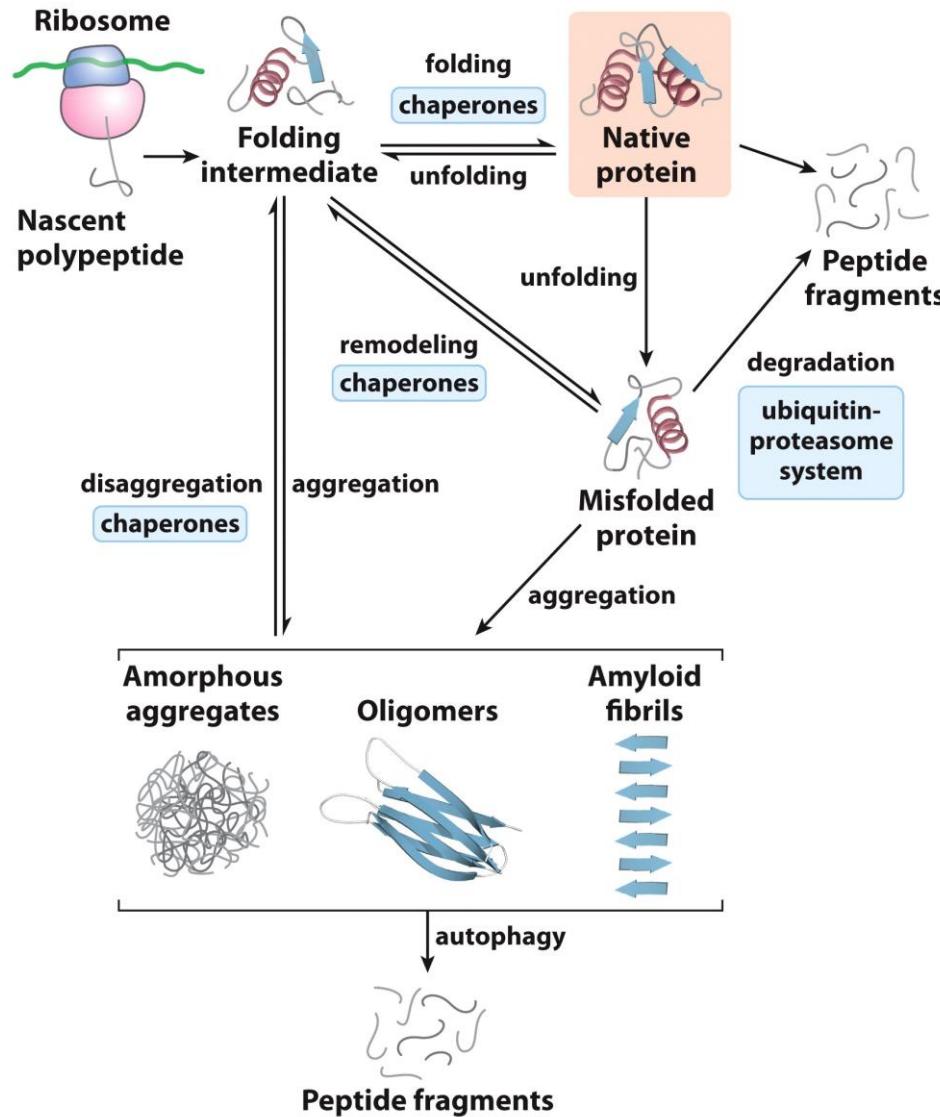
Proteostasis

The maintenance of an active set of cellular proteins.

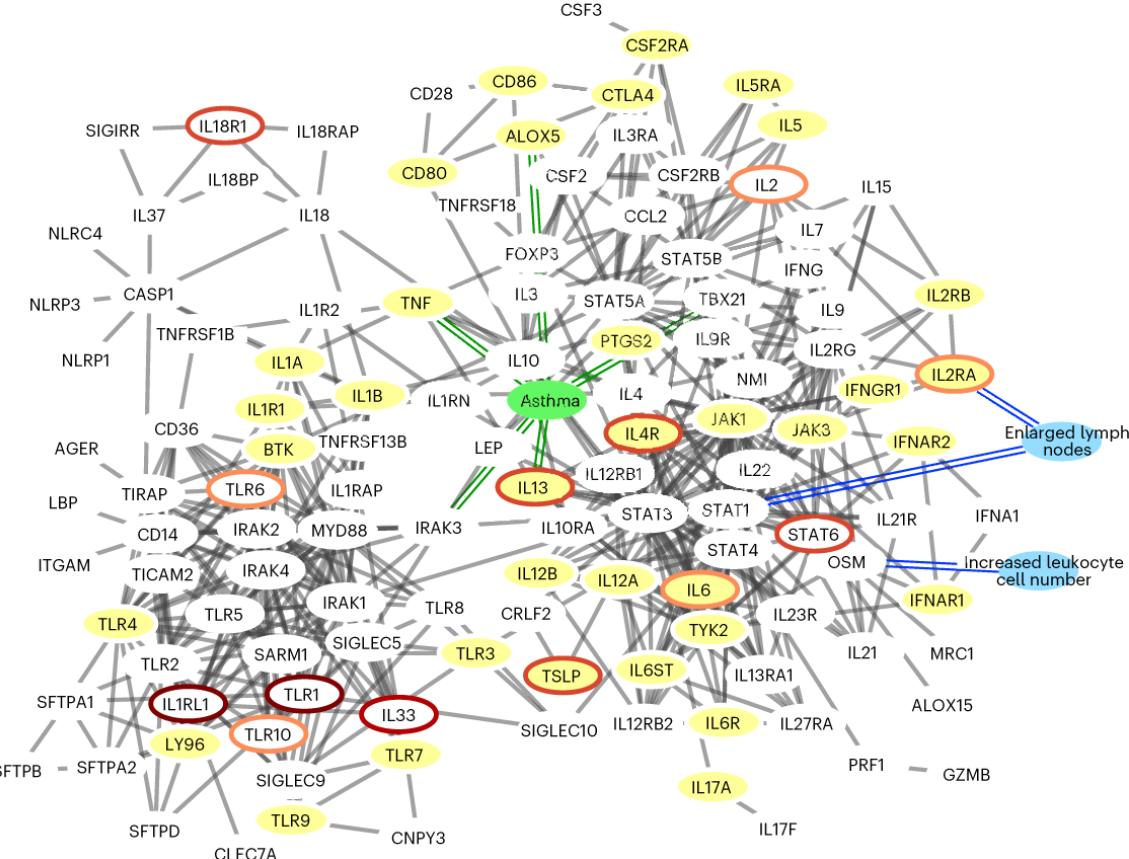
First, proteins are synthesized on a ribosome.

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.

[Search](#)[Download](#)[Help](#)[My Data](#)

Welcome to STRING

Protein-Protein Interaction Networks

Functional Enrichment Analysis

ORGANISMS

12535

PROTEINS

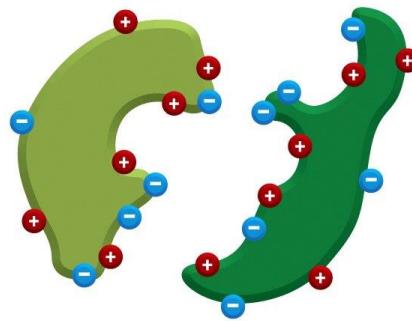
59.3 mio

INTERACTIONS

>20 bln

[SEARCH](#)

Protein structure affects protein-protein interactions



Protein structure affects protein-protein interactions

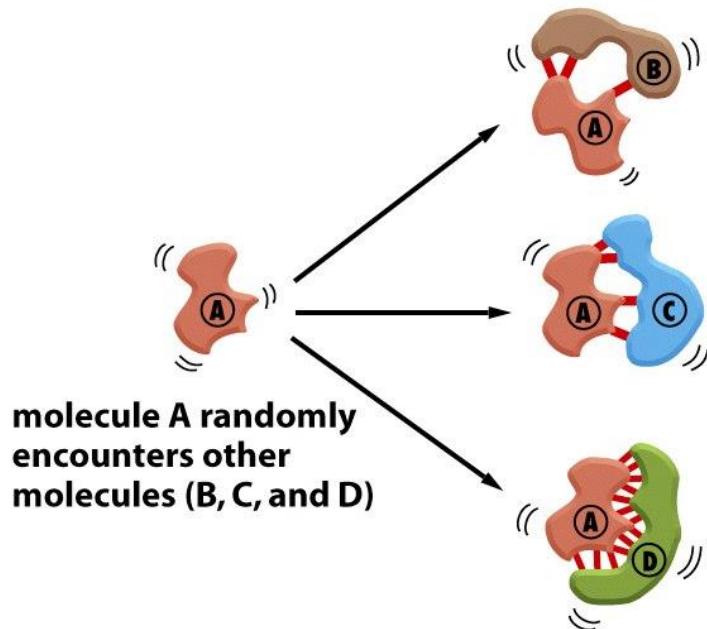
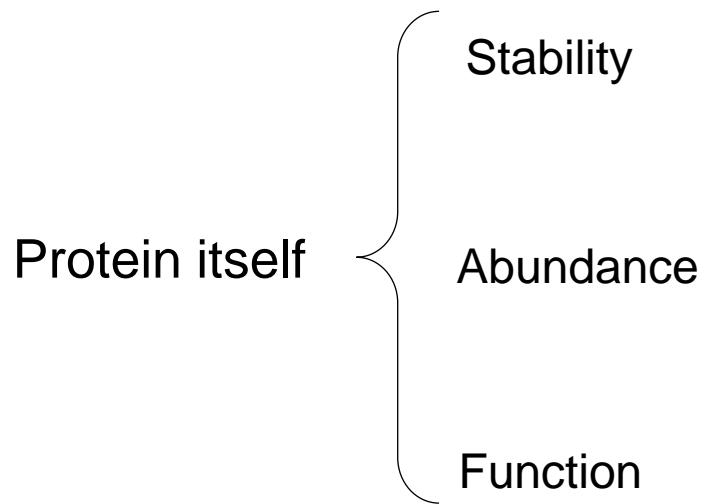


Figure 3-42 Molecular Biology of the Cell 5/e (© Garland Science 2008)

Disease variant's effect on proteins



Through interactions
(including regulatory effects)

Part 2:

Disease variants of MLH1 (DNA mismatch repair)

Uniprot database

UniProt BLAST Align Peptide search ID mapping SPARQL UniProtKB MLH1 Advanced | List Search Help

Status
Reviewed (Swiss-Prot) (628)
Unreviewed (TrEMBL) (45,538)

Popular organisms
Human (156)
Mouse (28)
Rat (20)
Bovine (14)
A. thaliana (11)

Taxonomy
Filter by taxonomy

Group by
Taxonomy
Keywords
Gene Ontology
Enzyme Class

Feedback

Proteins with
3D structure (62)
Active site (18)
Activity regulation (14)
Alternative products (isoforms) (62)
Alternative splicing (62)
More items

Help

Entry ▲ Entry Name ▲ Protein Names ▲ Gene Names ▲ Organism ▲ Length ▲

Entry	Entry Name	Protein Names	Gene Names	Organism	Length
P40692	MLH1_HUMAN	DNA mismatch repair protein Mlh1[...]	MLH1, COCA2	Homo sapiens (Human)	756 AA
P97679	MLH1_RAT	DNA mismatch repair protein Mlh1[...]	Mlh1	Rattus norvegicus (Rat)	757 AA
Q9JK91	MLH1_MOUSE	DNA mismatch repair protein Mlh1[...]	Mlh1	Mus musculus (Mouse)	760 AA
P38920	MLH1_YEAST	DNA mismatch repair protein MLH1[...]	MLH1, PMS2, YMR167W, YM8520.16	Saccharomyces cerevisiae (strain ATCC 204508 / S288c) (Baker's yeast)	769 AA
Q9ZRV4	MLH1_ARATH	DNA mismatch repair protein MLH1[...]	MLH1, At4g09140, F2J3.170, T8A17.9	Arabidopsis thaliana (Mouse-ear cress)	737 AA
Q54KD8	MLH1_DICDI	DNA mismatch repair protein Mlh1[...]	mlh1, DDB_G0287393	Dictyostelium discoideum (Social amoeba)	884 AA
Q9P7W6	MLH1_SCHPO	Putative MutL protein homolog 1[...]	mlh1, SPBC1703.04	Schizosaccharomyces pombe (strain 972 / ATCC 24843) (Fission yeast)	684 AA
A0A2Y9TI72	A0A2Y9TI72_PHYMC	DNA mismatch repair protein Mlh1	MLH1	Physeter macrocephalus (Sperm whale) (Physeter catodon)	758 AA
O95243	MBD4_HUMAN	Methyl-CpG-binding domain protein 4[...]	MBD4, MED1	Homo sapiens (Human)	580 AA
P54278	PMS2_HUMAN	Mismatch repair endonuclease PMS2[...]	PMS2, PMSL2	Homo sapiens (Human)	862 AA
P54277	PMS1_HUMAN	PMS1 protein homolog 1[...]	PMS1, PMSL1	Homo sapiens (Human)	932 AA
P46063	RECOQ1_HUMAN	ATP-dependent DNA helicase Q1[...]	RECQL, RECOQ1, RECQL1	Homo sapiens (Human)	649 AA
Q9UQ84	EXO1_HUMAN	Exonuclease 1[...]	EXO1, EXO1, HEX1	Homo sapiens (Human)	846 AA
Q9UHC1	MLH3_HUMAN	DNA mismatch repair protein Mlh3[...]	MLH3	Homo sapiens (Human)	1,453 AA
I3LT92	I3LT92_PIG	DNA mismatch repair protein Mlh1[...]	MLH1	Sus scrofa (Pig)	757 AA
O49621	MLO1_ARATH	MLO-like protein 1[...]	MLO1, MLO-H1, At4g02600, T10P11.12	Arabidopsis thaliana (Mouse-ear cress)	526 AA
G1QYL9	G1QYL9_NOMLE	MutL homolog 1	MLH1	Nomascus leucogenys (Northern white-cheeked gibbon) (Hylobates leucogenys)	756 AA
G3RJQ3	G3RJQ3_GORGO	MutL homolog 1	MLH1	Gorilla gorilla gorilla (Western lowland gorilla)	756 AA
K7FUJ0	K7FUJ0_PELSI	MutL homolog 1	MLH1	Pelodiscus sinensis (Chinese softshell turtle) (Trionyx sinensis)	727 AA
H0V205	H0V205_CAVPO	MutL homolog 1	MLH1	Cavia porcellus (Guinea pig)	758 AA
I3M618	I3M618_ICTTR	MutL homolog 1	MLH1	Ictidomys tridecemlineatus (Thirteen-lined ground squirrel) (Spermophilus tridecemlineatus)	758 AA
A0A0D9RNU4	A0A0D9RNU4_CHLSB	MutL homolog 1	MLH1	Chlorocebus sabaeus (Green monkey) (Cercopithecus sabaeus)	756 AA
A0A2R9AM85	A0A2R9AM85_PANPA	MutL homolog 1	MLH1	Pan paniscus (Pygmy chimpanzee) (Bonobo)	756 AA

<https://www.uniprot.org>

Uniprot database

UniProt BLAST Align Peptide search ID mapping SPARQL UniProtKB Advanced | List Search

P40692 · MLH1_HUMAN

Function

Names & Taxonomy

Subcellular Location

Disease & Variants

PTM/Processing

Expression

Interaction

Structure

Family & Domains

Sequence & Isoforms

Similar Proteins

Entry Variant viewer 2,139 Feature viewer Genomic coordinates Publications External links History

BLAST Align Download Add Add a publication Entry feedback

Functionⁱ

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 thus generates new entry points for the exonuclease EXO1 to degrade the strand containing the mismatch. DNA methylation would prevent cleavage and therefore assure that only the newly mutated DNA strand is going to be corrected. MutL alpha (MLH1-PMS2) interacts physically with the clamp loader subunits of DNA polymerase III, suggesting that it may play a role to recruit the DNA polymerase III to the site of the MMR. Also implicated in DNA damage signaling, a process which induces cell cycle arrest and can lead to apoptosis in case of major DNA damages.

Heterodimerizes with MLH3 to form MutL gamma which plays a role in meiosis. [5 Publications](#)

Features

Showing features for binding siteⁱ.

BLAST Align Download

TYPE ID POSITION(S) DESCRIPTION

-- Select --

▶ Binding site	38	ATP (UniProtKB ChEBI) 1 Publication Combined Sources
▶ Binding site	63	ATP (UniProtKB ChEBI) 1 Publication Combined Sources
▶ Binding site	82-84	ATP (UniProtKB ChEBI) 1 Publication Combined Sources
▶ Binding site	100-104	ATP (UniProtKB ChEBI) 1 Publication Combined Sources

GO annotationsⁱ

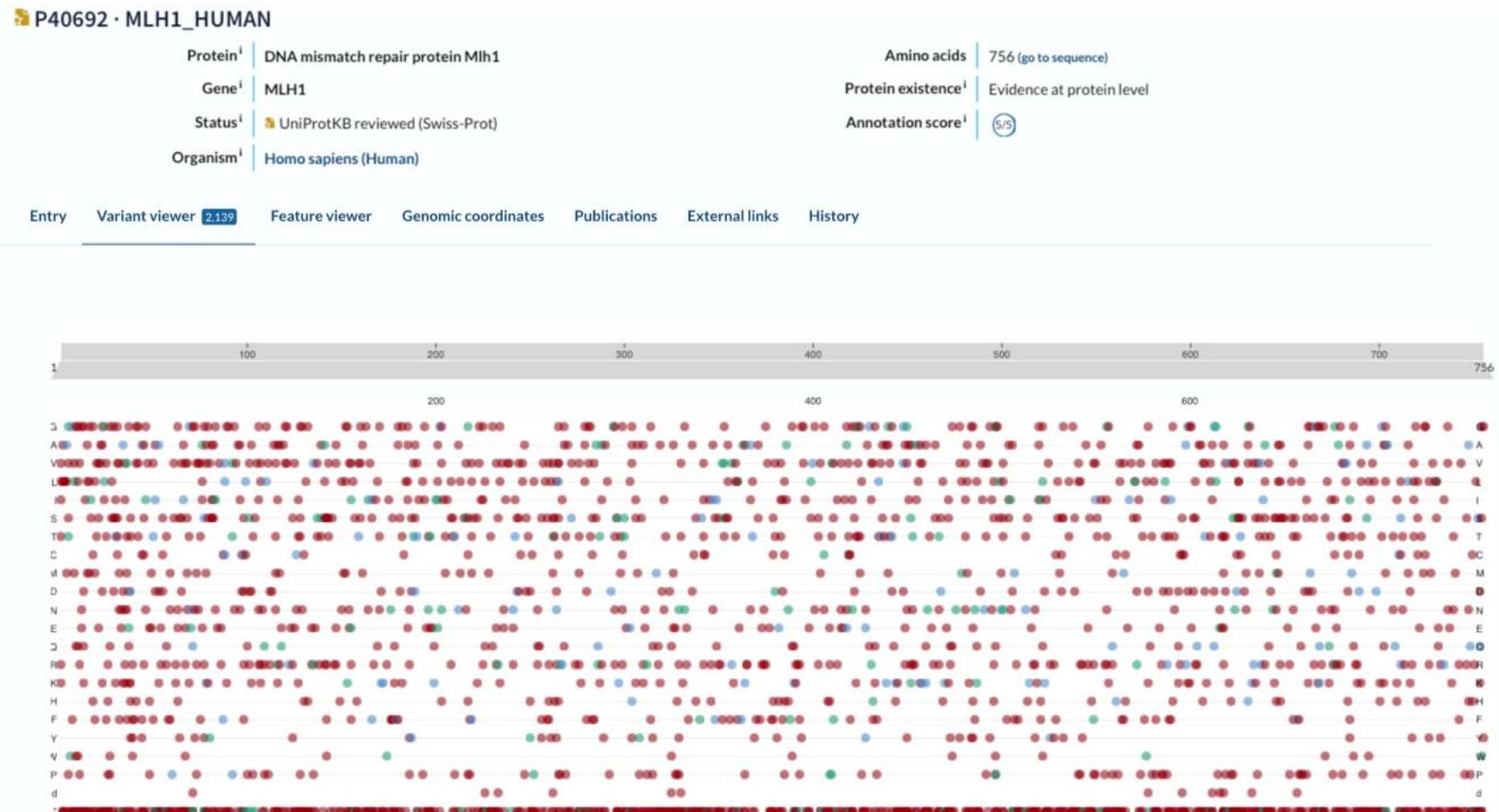
Access the complete set of GO annotations on QuickGO

Expand table

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

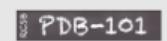
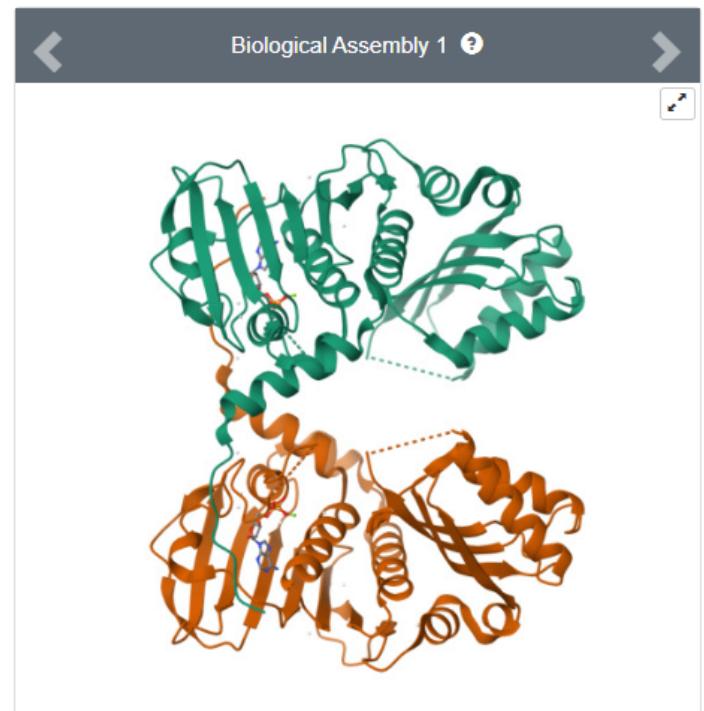


229,681 Structures from the PDB



1,068,577 Computed Structure Models (CSM)

Enter search term(s), Entry ID(s), or sequence

Include CSM [Advanced Search](#) | [Browse Annotations](#)[Help](#)[Structure Summary](#)[Structure](#)[Annotations](#)[Experiment](#)[Sequence](#)[Genome](#)[Ligands](#)[Versions](#)**Global Symmetry:** Cyclic - C2 **Global Stoichiometry:** Homo 2-mer - A2 [Find Similar Assemblies](#)

Biological assembly 1 generated by PISA (software)

4P7A

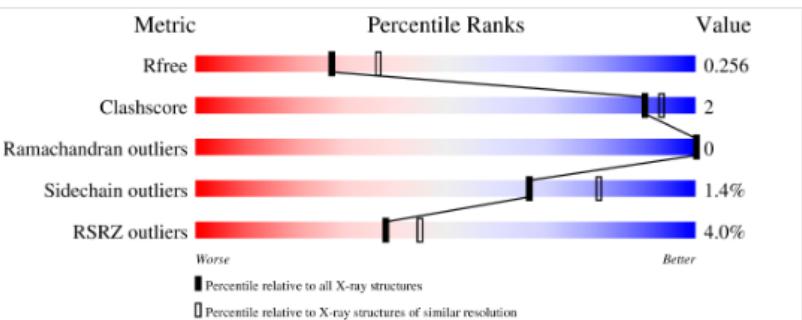
Crystal Structure of human MLH1

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](#)

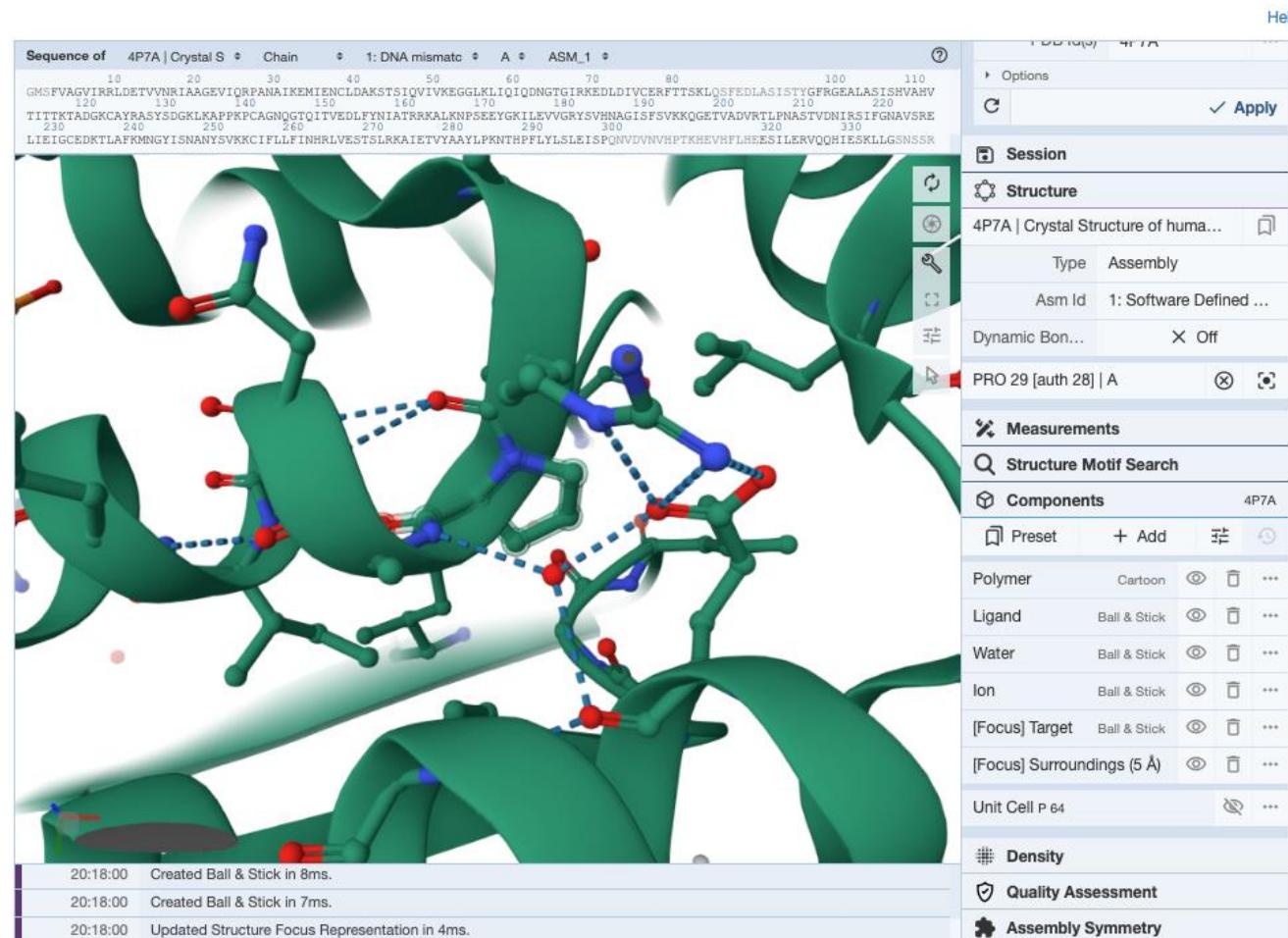
Ligand Structure Quality Assessment

Worse 0 1 Better

Ligand structure goodness of fit to experimental data

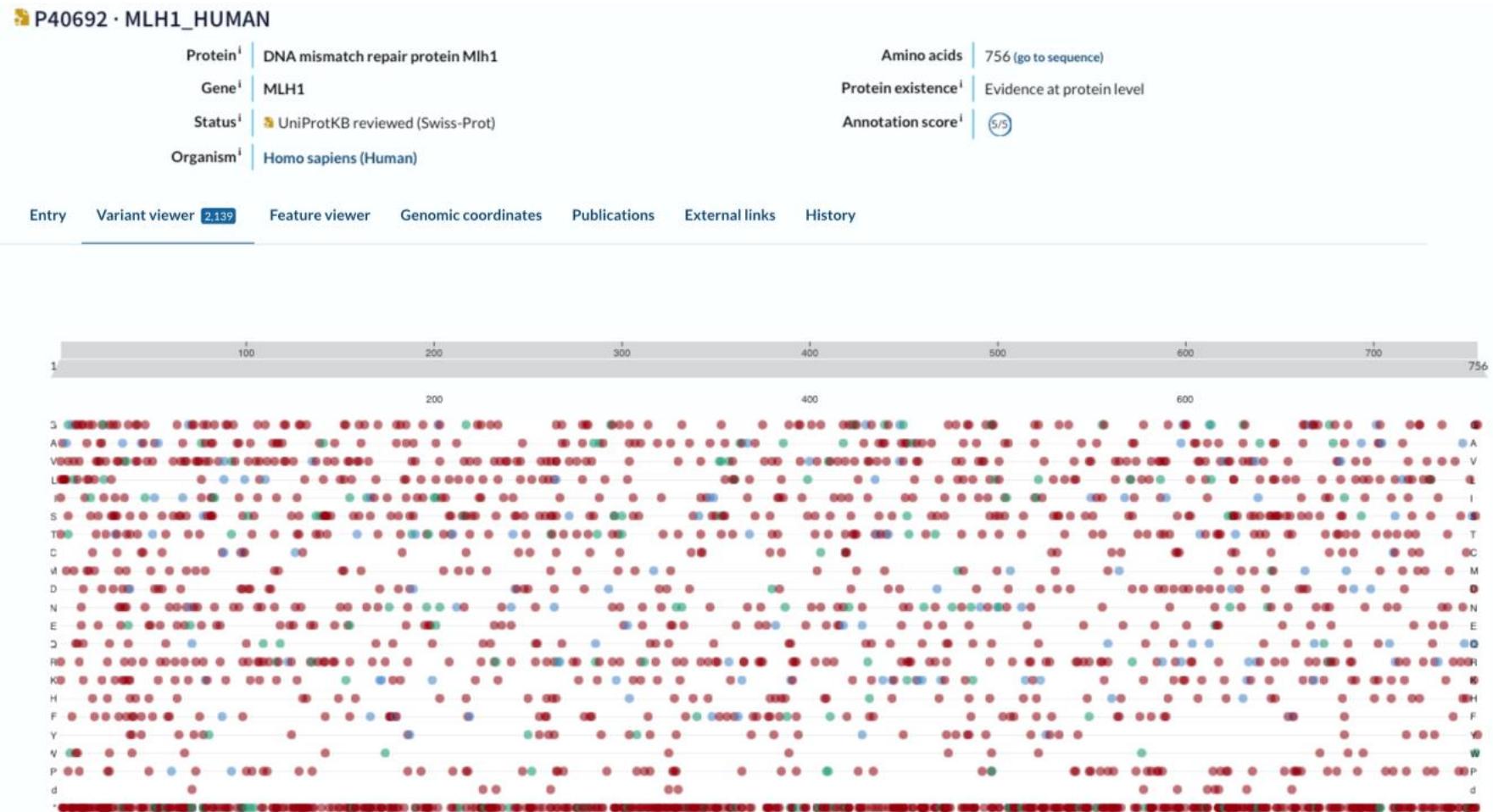
Mol* 3D Viewer

This version of Mol* can be used to upload single or multiple data files and align structures.



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

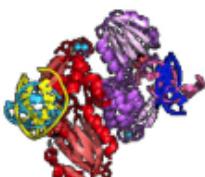
Look at pathological/clinical variants that fall in these position bins

PDBsum database (EMBL-EBI)

[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)

FindExample: "[1kfv](#)"**NEW Alpha Fold model** (human proteins only)Enter UniProt accession (or UniProt id), to find Alpha Fold model of given protein. Eg [P00734 \(THR8_HUMAN\)](#).**Find**see [Analyses](#)**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

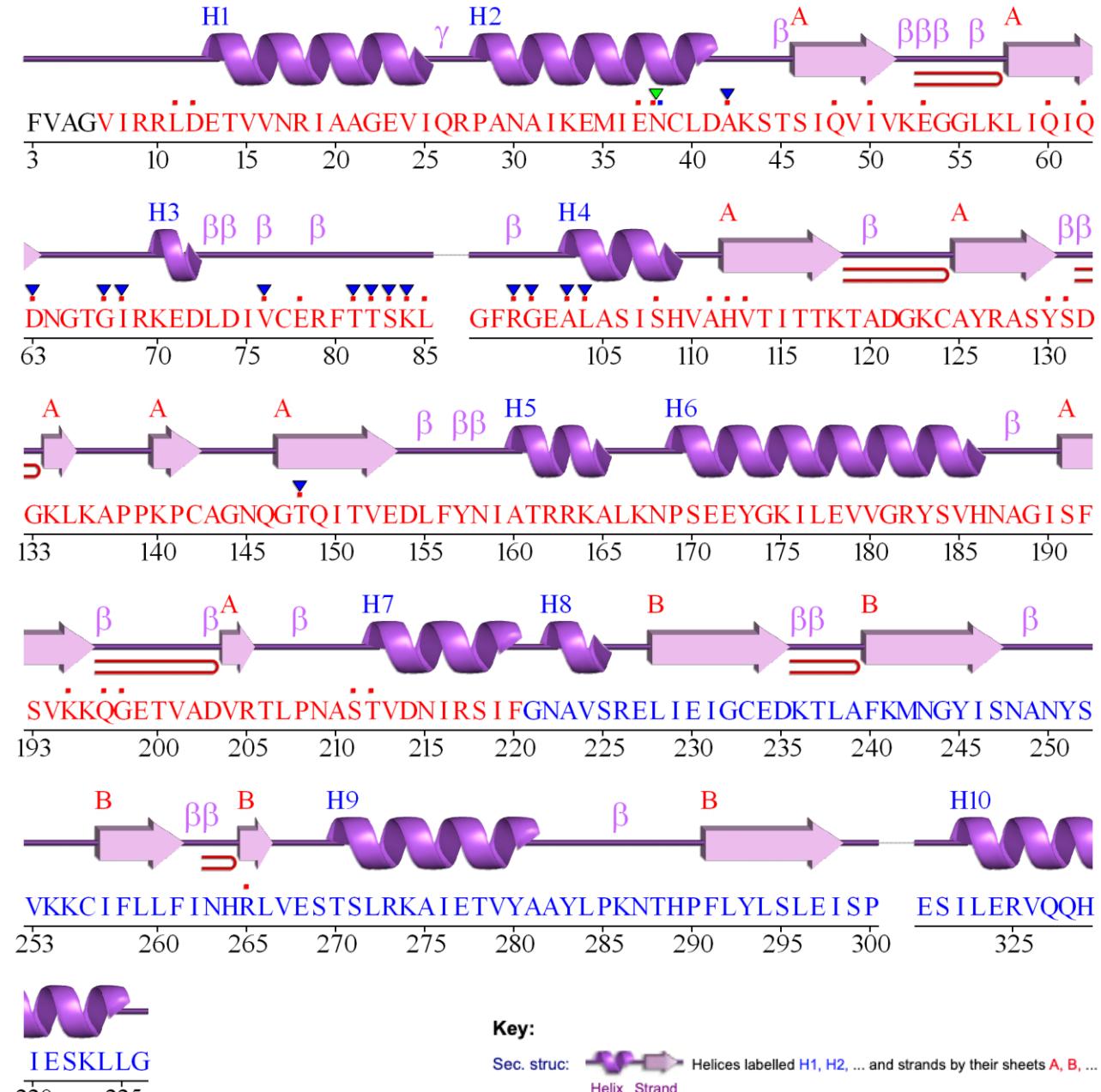
Download PDBsum 1 to process your own structures in-house

Related databases

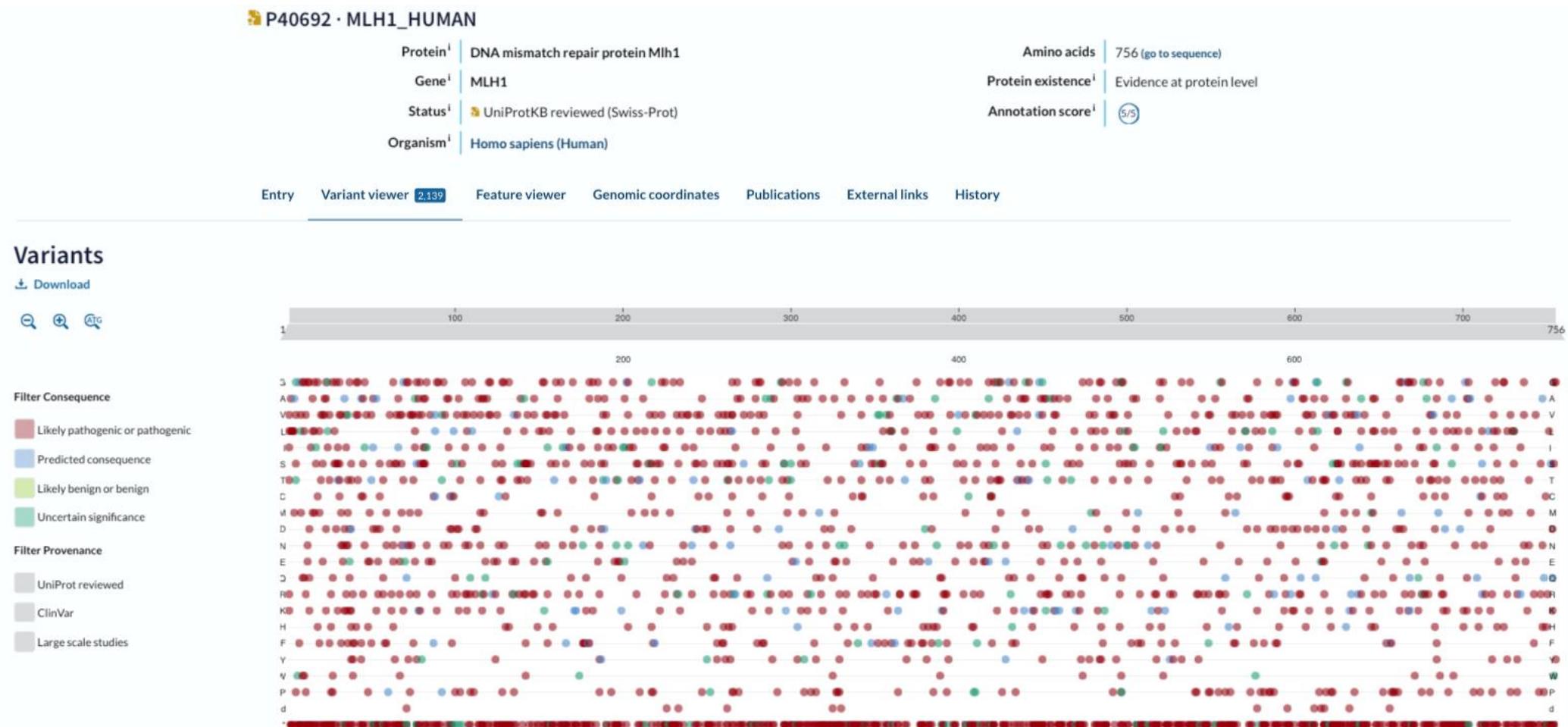
Enzyme 3D structures organized by the E.C. numbering hierarchy.



Structures of drugs and their target proteins in the PDB.



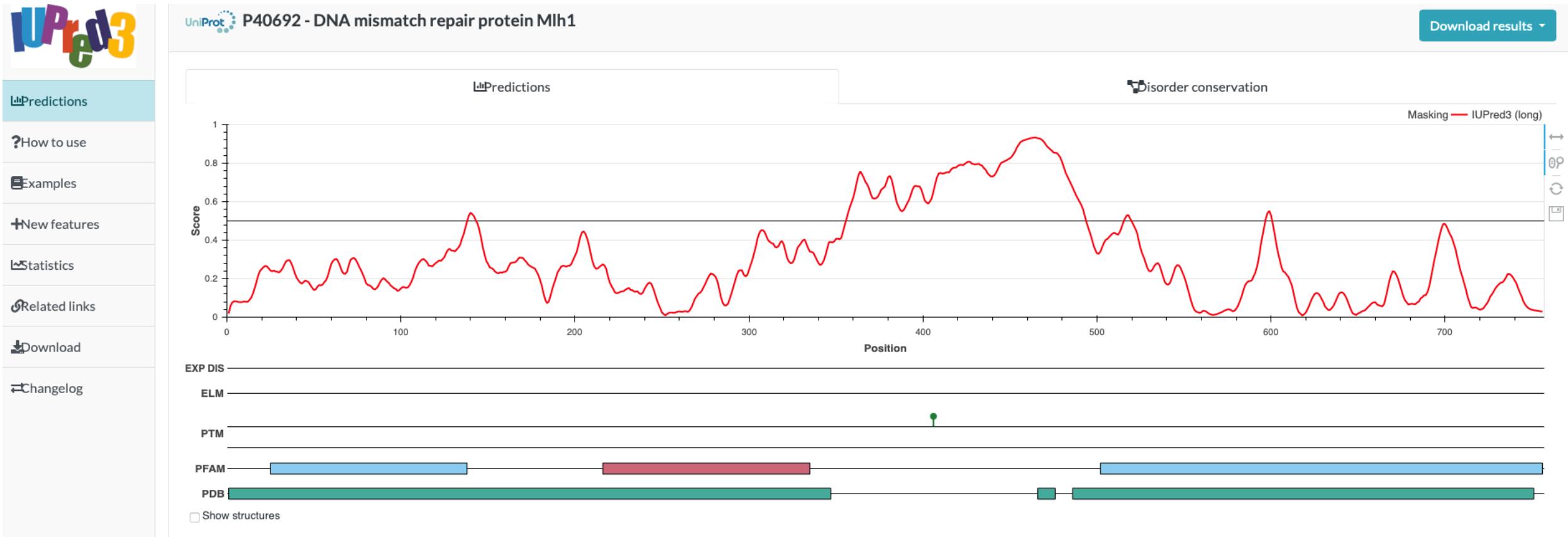
Uniprot database: variant page



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: MENFQKVKEKIGEGTYGV... 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

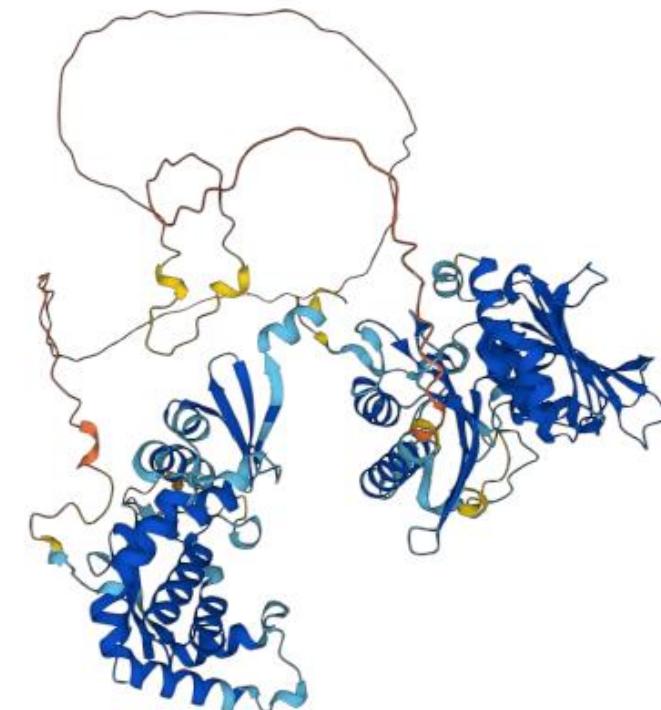
Gene MLH1

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](#) ↗

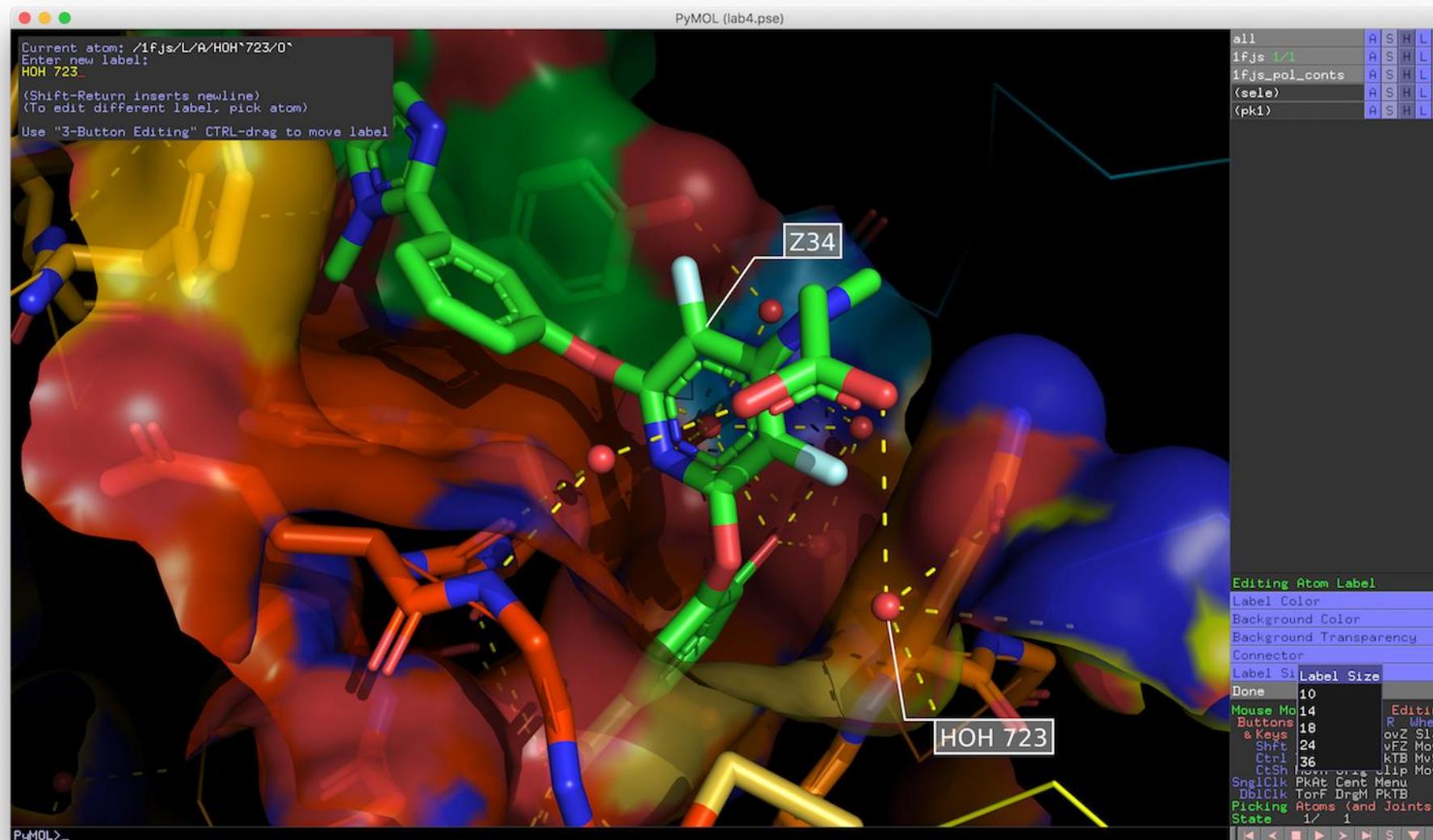


<https://alphafold.ebi.ac.uk>

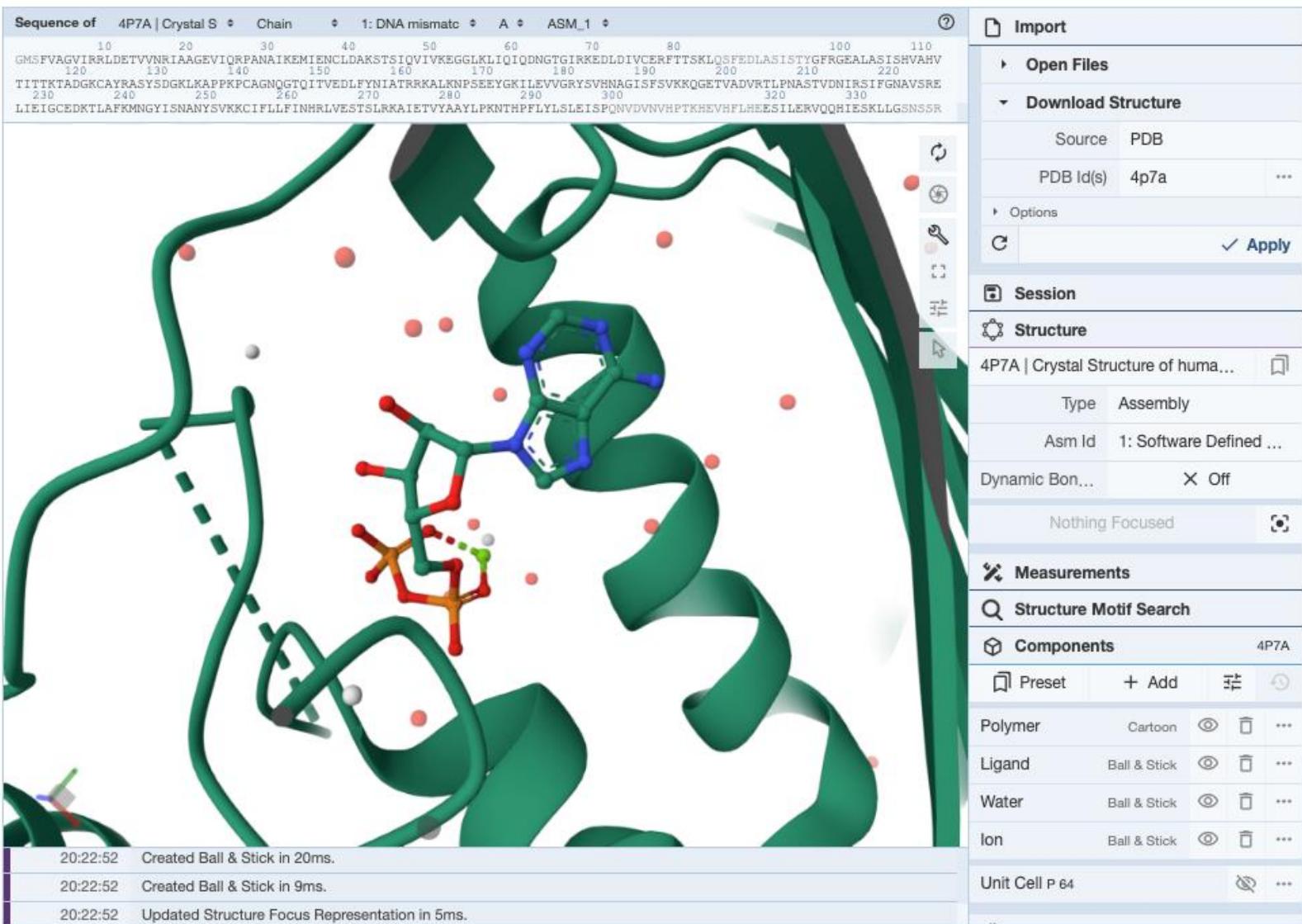


What clinical variants fall in the disordered regions?

Pymol: Protein and Molecular Visualization



Which variants might affect the ATP binding affinity?



Cross validate your results with PDBsum (EMBL) database