

Leveraging Molecular Databases for Variant Interpretation

Part 2: 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)

Introduction to proteins:

Proteins are one of the most important molecules in biology due to their critical roles in numerous cellular processes. Proteins serve as enzymes, hormones, receptors, and structural components of cells, tissues, and organs (Figure 1). They are involved in the transport of molecules across cell membranes, the regulation of gene expression, and the maintenance of the immune system. Any malfunction in protein structure or function can lead to diseases such as cystic fibrosis, Alzheimer's disease, Parkinson's disease, and many others. Therefore, understanding the structure and function of proteins is crucial for advancing our knowledge of biological processes and developing treatments for a wide range of diseases.

gi

The sequence of amino acids in a protein determines its three-dimensional structure, which in turn determines its function. The primary structure of a protein is the linear sequence of amino acids that make up the protein chain. The sequence of amino acids determines the secondary structure, such as alpha helices and beta sheets, which are stabilized by hydrogen bonds (Figure 2).

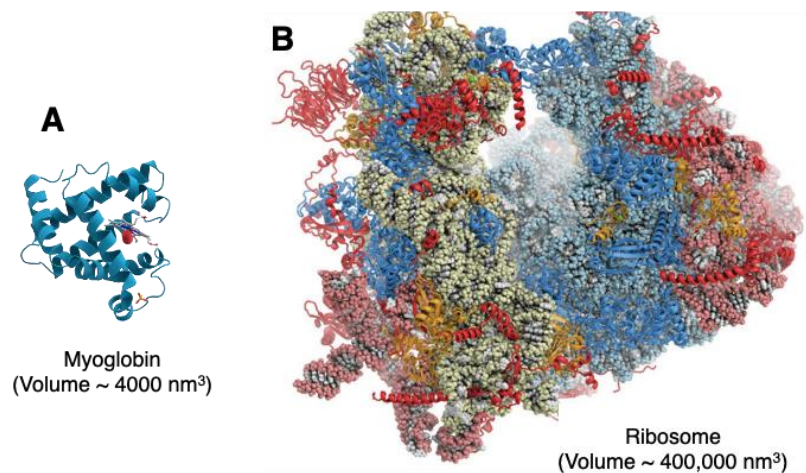


Figure 1. The 3D structure of A) Human Myoglobin, and B) Ribosome.

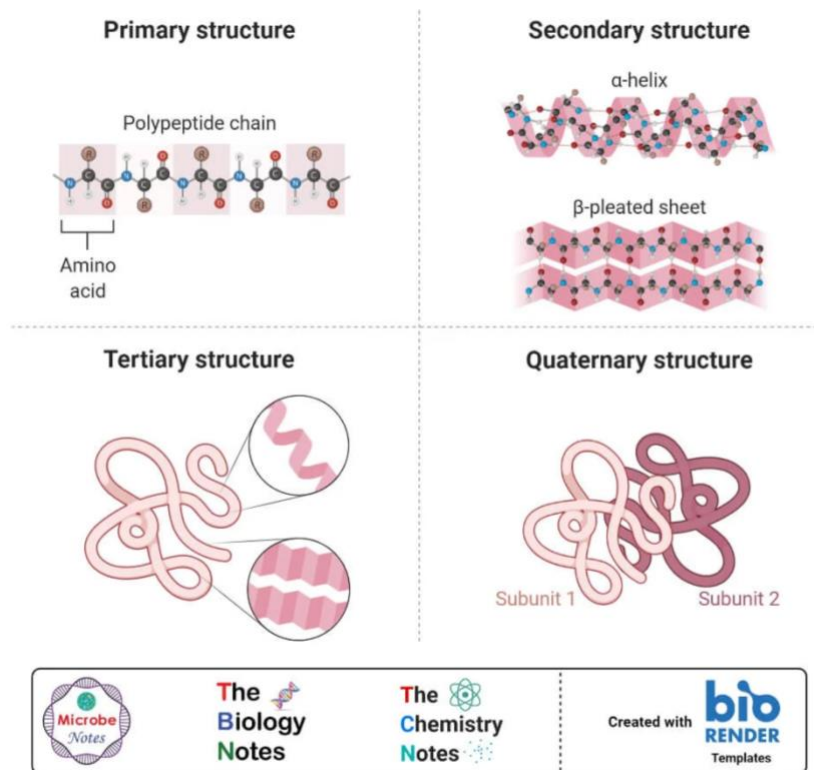


Figure 2. Different structures of proteins (credit: /microbenotes.com)

The secondary structure then folds into the tertiary structure, which is the overall 3D shape of the protein, stabilized by a variety of chemical bonds and interactions. The final quaternary structure is formed when multiple protein subunits come together to form a functional protein complex. The specific amino acid sequence and resulting protein structure determines the protein's function, such as catalyzing biochemical reactions, transporting molecules across membranes, and transmitting signals within cells. Therefore, understanding the relationship between protein sequence, structure, and function is crucial for developing treatments for various diseases (Figure 3).

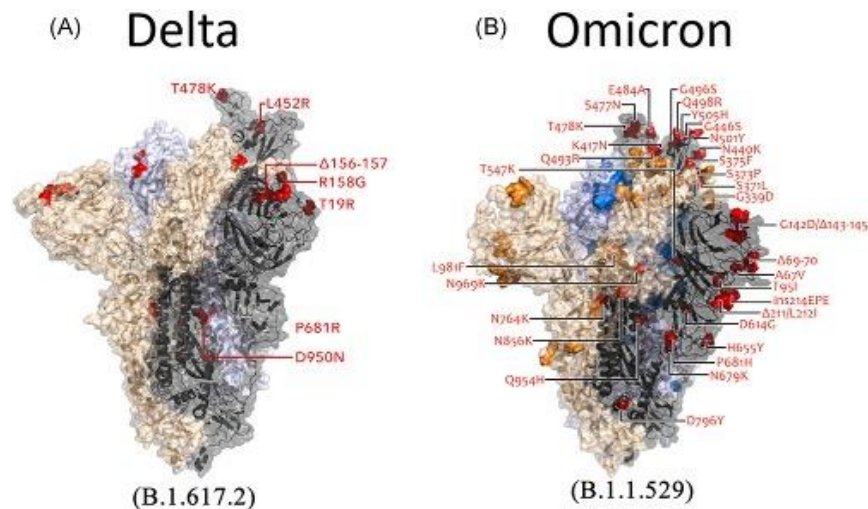


Figure 3. A comparison of (A) Delta and (B) Omicron variant spike mutation (Image source: Modified from COVID-19 Genomics UK Consortium; *Journal of Medical Virology* doi: 10.1002/jmv.27526).

Protein stability refers to the ability of a protein to maintain its native conformation, which is essential for its proper function. The stability of a protein is dependent on various factors, such as its primary sequence, solvent conditions, temperature, and pH. A protein that is unstable may undergo conformational changes, leading to aggregation or denaturation, which can cause loss of function or even cellular toxicity (Figure 4).

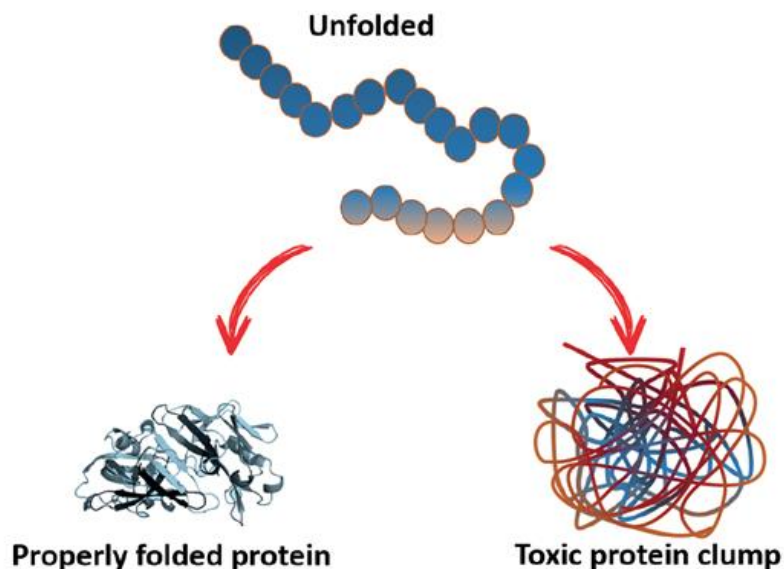


Figure 4. An unfolded protein can fold to properly folded or misfolded states.

Examples where changes in protein stability and aggregation are associated with diseases:

1. **Cystic Fibrosis:** Cystic fibrosis is a genetic disease caused by mutations in the *CFTR* gene. The CFTR protein is responsible for regulating the movement of salt and water in and out of cells, and mutations in the *CFTR* gene lead to a misfolded and unstable protein. The misfolded CFTR protein is degraded by the cell, leading to a decrease in function and symptoms of cystic fibrosis.
2. **Huntington's disease:** Huntington's disease is caused by an expansion of the CAG trinucleotide repeat in the huntingtin gene, *HTT*. The expanded huntingtin protein is prone to misfolding and aggregation, leading to the formation of toxic protein aggregates in the brain. These protein aggregates are thought to contribute to the degeneration of neurons and the symptoms of Huntington's disease.
3. **Alzheimer's disease:** Alzheimer's disease is characterized by the accumulation of amyloid-beta protein in the brain, which forms plaques that contribute to the degeneration of neurons. Mutations in the amyloid precursor protein (APP) and presenilin genes can lead to increased production of amyloid-beta protein, which can then misfold and aggregate.

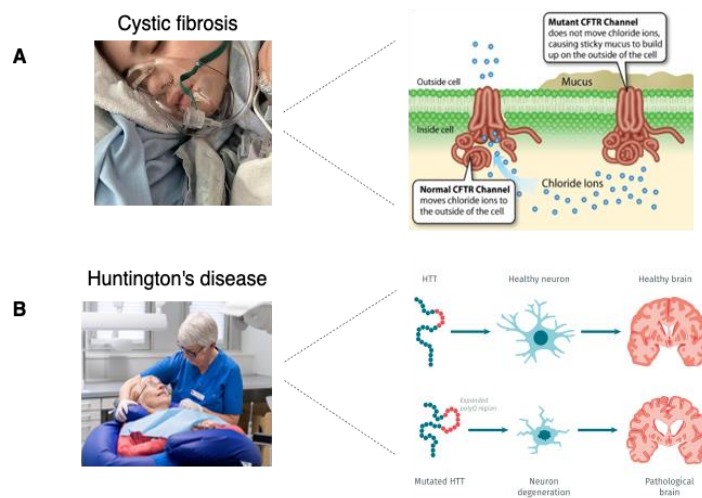


Figure 5. The molecular basis of the genetic diseases A) Cystic fibrosis, and B) Huntington's disease. The most common mutation (~ in 70% of cystic fibrosis patients) is a three-base pair deletion in the DNA sequence of CFTR, which causes the absence of a single amino acid in this protein. The mutated CFTR ion channel cannot move the chloride ions out of the cell leading to the accumulation of sticky mucus outside the cell. In the case of Huntington's disease (HD), the mutated Huntingtin protein (HTT) results in the production of an altered protein leading to dysfunction and neuronal death in the brain's striatum region.

The case of lynch syndrome and protein disease variants:

Lynch syndrome is a genetic disorder that increases an individual's risk of developing various types of cancers, most notably colorectal cancer and endometrial cancer. This syndrome is caused by inherited mutations in certain genes responsible for DNA repair, particularly the mismatch repair (MMR) genes. Lynch syndrome is an autosomal dominant genetic condition, which means that a person only needs to inherit one mutated copy of the responsible gene from either parent to be at risk. The MMR genes, such as *MLH1*, *MSH2*, *MSH6*, *PMS2*, and *EPCAM*, play a crucial role in ensuring the accuracy of DNA replication and repair. When mutations occur in these genes, the MMR system becomes impaired, leading to a higher likelihood of genetic mutations accumulating in an individual's cells over time. These accumulated mutations can increase the risk of cancer development, as they may affect the regulation of cell growth and division. One of the key aspects of Lynch syndrome is the increased risk of developing certain types of cancers at an earlier age than in the general population. Colorectal cancer and endometrial cancer are the most common malignancies associated with Lynch syndrome, but it can also elevate the risk of other cancers, including ovarian, stomach, small intestine, urinary tract, and brain cancers (Figure 6).

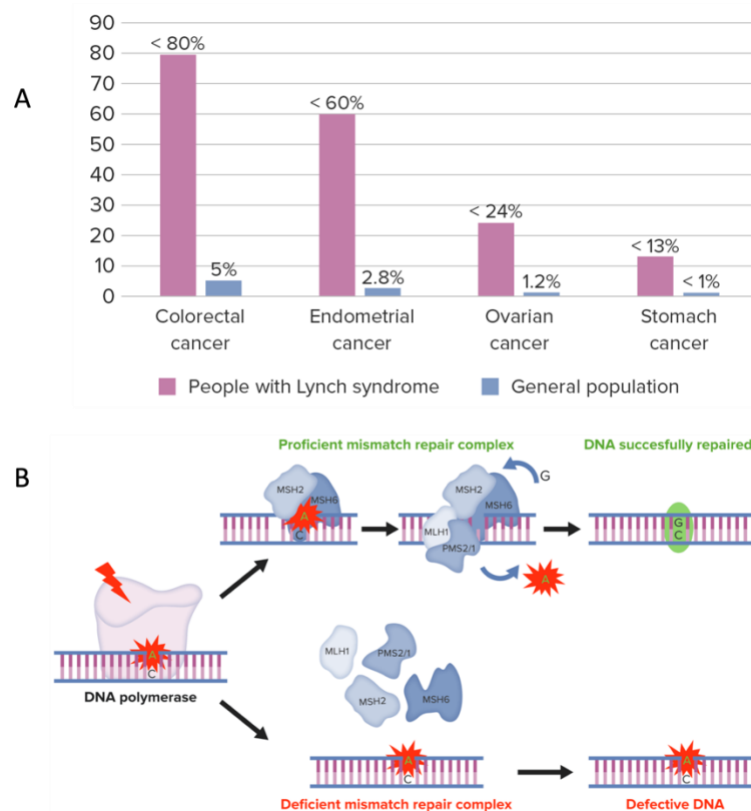


Figure 6. A) The lifetime cancer risk comparison between people with Lynch syndrome and the general population. B) DNA mismatch repair (MMR) recognizes and repairs genetic mismatches generated during DNA replication. In tumor cells the presence of a deficient MMR system results in faulty DNA repair and mutations. (Credit: lecturio medical: <https://app.lecturio.com/>)

In Lynch syndrome, the protein variants resulting from mutations in MMR genes disrupt the DNA repair process, ultimately contributing to the development of cancer. Understanding the genetic basis of Lynch syndrome and its protein variants is essential for early diagnosis, effective management, and potential prevention of associated cancers. We will particularly focus on two pathological variants c.83C>T (Pro28Leu) and c.464T>G (Leu155Arg) in the *MLH1* gene. To better understand how these mutations, we will query different biological databases.

Practical part:

We will start our practical session by using the Uniprot database. The UniProt database, short for Universal Protein Resource, is a comprehensive and widely used biological database that serves as a valuable resource for researchers in the fields of molecular biology, bioinformatics, and genomics. UniProt collects, annotates, and curates protein sequences and functional information from various sources, including experimental data and literature. The database is regularly updated to ensure the accuracy and relevance of its content, making it an indispensable tool for studying protein structure, function, and evolution. UniProt is a collaborative effort involving several organizations, and it offers multiple access points, including UniProtKB for protein sequences, UniRef for sequence clustering, and UniProtKB/Swiss-Prot for high-quality, manually curated protein entries. We will look up the information of the gene *MLH* in human (Figure 7)

Entry	Entry Name	Protein Names	Gene Names	Organism	Length
P40692	MLH1_HUMAN	DNA mismatch repair protein Mlh1	MLH1, COCA2	Homo sapiens (Human)	757 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

Figure 7. MLH1 query on the uniprot database.

For every gene and its encoded protein, Uniprot offers a wealth of information (Figure 8). The "Entry" tab typically presents a summary of the protein's name, gene, and primary sequence. The "Function" tab delves into details about the protein's role in biological processes, its molecular function, and interactions with other molecules. The "Expression" tab offers insights into where the protein is expressed in different tissues or cell types. The "Structure" tab provides information

on the protein's 3D structure if available. The "Interaction" tab elucidates protein-protein interactions and partners. Lastly, the "Family & Domains" tab outlines protein domains, motifs, and similarity to other proteins, while the "Sequence" tab offers the full amino acid sequence and related details. These tabs collectively offer a comprehensive view of a protein's attributes and functions, aiding researchers in various biological and biomedical investigations.

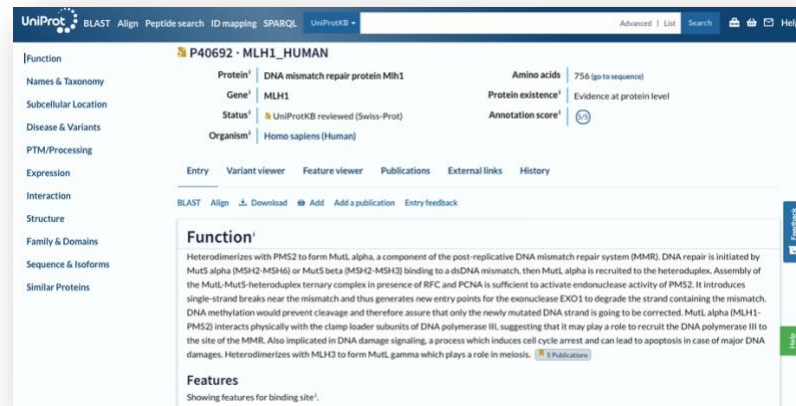


Figure 8. The “function” information tab for the gene *MLH1* in human.

Of particular importance for us is the "disease & variants" tab where we can find crucial information about the protein's involvement in various diseases, genetic variants associated with it, and their potential implications. This tab often includes details about mutations, polymorphisms, and their impact on the protein's structure or function. Researchers and clinicians can use this information to better understand the genetic basis of diseases, identify potential therapeutic targets, or assess the clinical significance of specific variants in a diagnostic or research context (Figure 9).

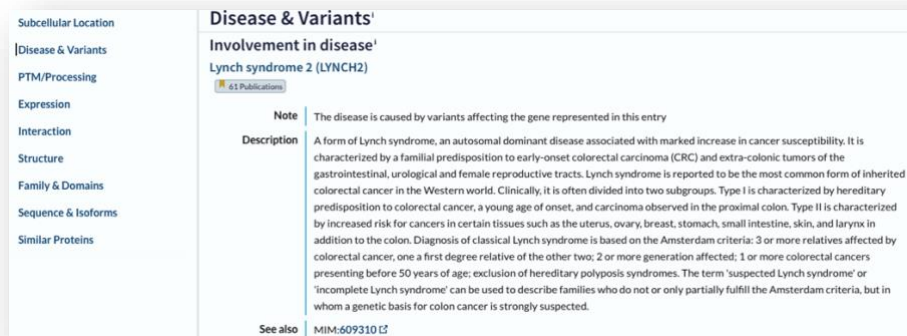


Figure 9. The “disease and variants” information tab for the gene *MLH1* in human.

Resource box 1: the history of Uniprot database

The UniProt database has a rich history that dates back to the early days of protein sequence data collection and management. Here's a brief overview of its historical development:

1. Early Protein Databases: The need for a comprehensive protein sequence database became evident as researchers began generating protein sequences using methods like DNA sequencing techniques. In the 1970s and 1980s, various protein databases and repositories emerged independently in different regions of the world, including PIR (Protein Information Resource) in the United States and SWISS-PROT in Switzerland.
2. Creation of SWISS-PROT: In 1986, the Swiss Institute of Bioinformatics (SIB) established SWISS-PROT, a manually curated protein sequence database characterized by high-quality annotations. SWISS-PROT aimed to provide accurate and reliable information about protein sequences and their functions.
3. Collaborative Efforts: Recognizing the need for global collaboration and integration of protein sequence data, the European Bioinformatics Institute (EBI), and the Protein Information Resource (PIR) joined forces with SWISS-PROT in the late 1990s. This collaboration led to the creation of UniProt in 2002.
4. UniProt Consortium: The UniProt Consortium was officially formed to oversee the UniProt database. It comprises several organizations, including the European Bioinformatics Institute (EBI), the Swiss Institute of Bioinformatics (SIB), and the Protein Information Resource (PIR). This consortium ensures the ongoing curation, maintenance, and expansion of UniProt to meet the growing needs of the scientific community.
5. UniProtKB and Expansion: UniProt initially combined the SWISS-PROT and TrEMBL (Translated EMBL Nucleotide Sequence Data Library) databases. SWISS-PROT continued to provide manually curated, high-quality entries, while TrEMBL included automatically generated annotations from translated DNA sequences. Over the years, UniProt has expanded to include additional resources like UniRef (sequence clustering) and specialized knowledge bases like UniProtKB/Swiss-Prot and UniProtKB/TrEMBL.
6. Regular Updates: UniProt is continuously updated to incorporate new protein sequences, annotations, and features, ensuring that researchers have access to the most current and accurate information.

We will now focus on the first clinical variant that we study in this practical session: Pro to Leu mutation at the amino acid position 28. As shown in Figure 10, Uniprot offers an external link to the ExPASy database that we can use to further gain information on this variant. Particularly, we can look at the sequence of different orthologs of this protein and see whether this amino acid position is variable or conserved in other species.

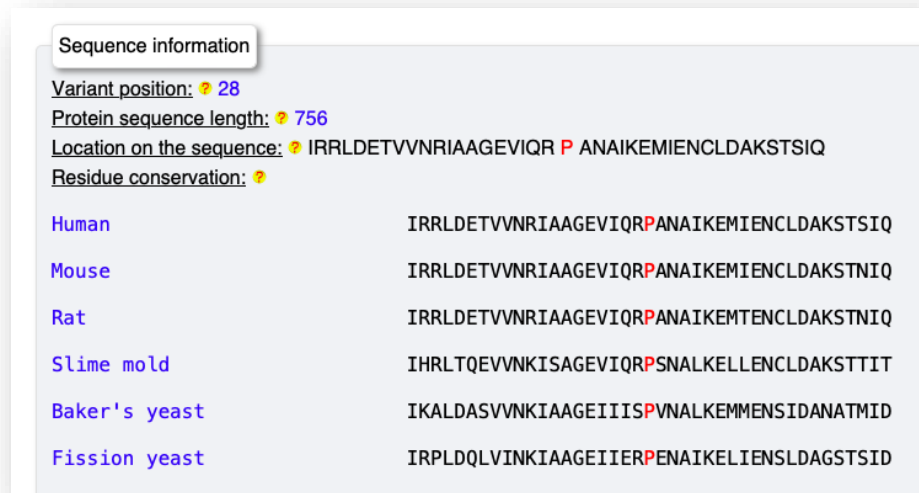


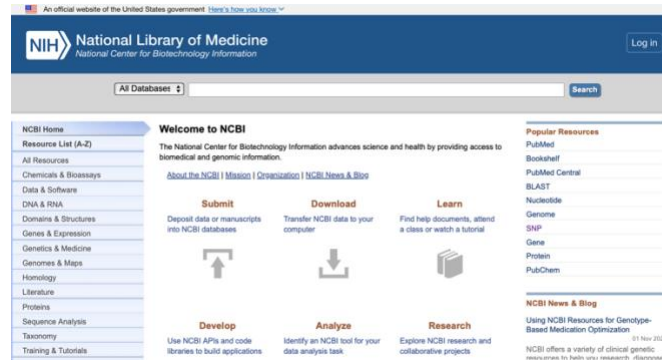
Figure 10. The protein sequence information and an overview the conservation of the amino acid position 28 in the *MLH1* gene in other species from Expaty database.

We see that the proline amino acid in position 28 is conserved in several species that are evolutionarily distant. Investigating the sequence conservation of protein amino acid positions can often help us gain information on the nature of mutations and whether they might cause deleterious changes that lead to disease. When a specific amino acid residue is highly conserved across diverse species, it suggests that this position plays a crucial role in the protein's structure or function. Any mutation that disrupts this conservation could potentially result in a loss of function or stability, affecting the protein's normal physiological role and potentially leading to diseases. Therefore, identifying such conserved residues is instrumental in finding potential hotspots for genetic and functional analysis in both basic research and clinical studies. To systematically retrieve information on the available orthologs of *MLH1* gene, we will query the NCBI database (Resource Box 2).

We will particularly look at the gene information tab of the NCBI database, query the gene *MLH1* and click on the “orthologs” tab. We will then particularly select the “primates” sequences and use the amino acid sequence alignment COBALT to build a multiple sequence alignment from the orthologous sequences of the *MLH1* gene (Figure 11).

Resource box 2: NCBI database and its history

The National Center for Biotechnology Information (NCBI) is a renowned and comprehensive resource for biological and biomedical data. It was established as a division of the National Library of Medicine (NLM), which is part of the National Institutes of Health (NIH) in the United States. NCBI was founded with the aim of organizing and providing access to a vast array of biological information, including DNA sequences, genomic data, protein sequences, scientific literature, and more.



Here's a brief overview of the history and evolution of the NCBI database:

1. **Early Origins:** The roots of NCBI can be traced back to the late 1980s when efforts to manage and share the growing volumes of biological data led to the development of various sequence databases and resources. NCBI's GenBank database, established in 1982, served as one of the earliest repositories for DNA sequences.
2. **Establishment of NCBI:** The National Center for Biotechnology Information was officially established in 1988, consolidating and expanding upon existing resources. Dr. David J. Lipman played a pivotal role as the first Director of NCBI.
3. **GenBank and Entrez:** NCBI introduced the Entrez system, a powerful search and retrieval platform, which allowed users to access multiple databases seamlessly. GenBank, part of Entrez, became a central repository for DNA sequences, and it continues to be one of the world's most significant and widely used genomic databases.
4. **Expansion of Databases:** Over the years, NCBI expanded its portfolio of databases to include resources like PubMed (a repository of biomedical literature), BLAST (a sequence alignment tool), RefSeq (a curated collection of reference sequences), and more. These resources cover a wide range of biological and biomedical data types.
5. **Genome Sequencing Projects:** NCBI played a pivotal role in coordinating and archiving data from major genome sequencing projects, including the Human Genome Project. This initiative culminated in the release of the first draft of the human genome sequence in 2000, marking a significant milestone in genomics.
6. **Continued Development:** NCBI has continued to evolve and expand its offerings, incorporating new technologies and data types such as functional genomics, structural biology, and high-throughput sequencing data. It remains a global hub for bioinformatics and computational biology research.

Today, the NCBI provides a suite of invaluable resources and tools that support research and discovery in the fields of genomics, genetics, and molecular biology. Its mission to advance science and health by providing access to a wealth of biological information has had a profound impact on scientific research and medical advancements worldwide.

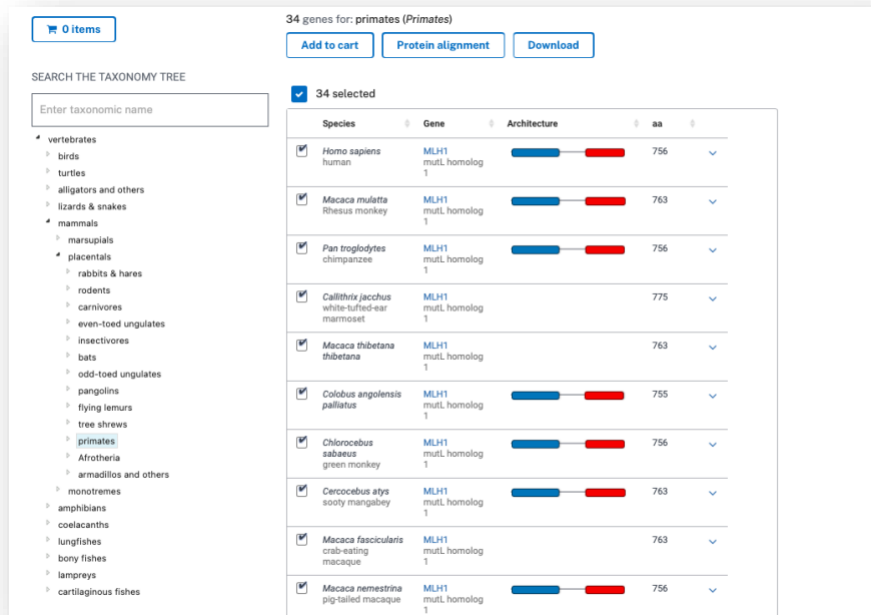


Figure 11. The ortholog tab of the NCBI database for the gene *MLH1* with selected sequences for primates.

Detecting pathological variants in genes is crucial for understanding genetic diseases and potential treatments. Various methods and tools have been developed to identify such variants. A traditional method is **SIFT (Sorting Intolerant From Tolerant)**. SIFT is a widely used bioinformatics tool designed to predict the impact of amino acid substitutions on protein function. It helps identify potentially pathological variants in protein-coding genes by assessing whether a specific amino acid change is likely to be tolerated (non-pathogenic) or intolerated (pathogenic). SIFT relies on sequence conservation across species; it assumes that conserved amino acids are more likely to be functionally important. The algorithm calculates a SIFT score for each variant, and a lower score suggests a higher likelihood of the variant being pathogenic. Researchers and clinicians often use a predetermined cut-off to classify variants as deleterious or benign (Figure 12).

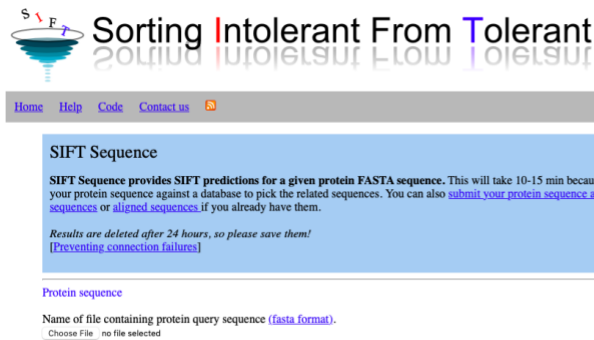


Figure 12. The SIFT webserver that predicts the impact of amino acid substitutions on protein function. We will next analyze the amino acid sequence of the protein MLH1 and cross-check our clinical variant (P28L) in the results (Figure 13). We will further expand the list of clinical variants from the uniprot database and see whether SIFT can accurately predict the impact of amino acid substitutions that lead to diseases.

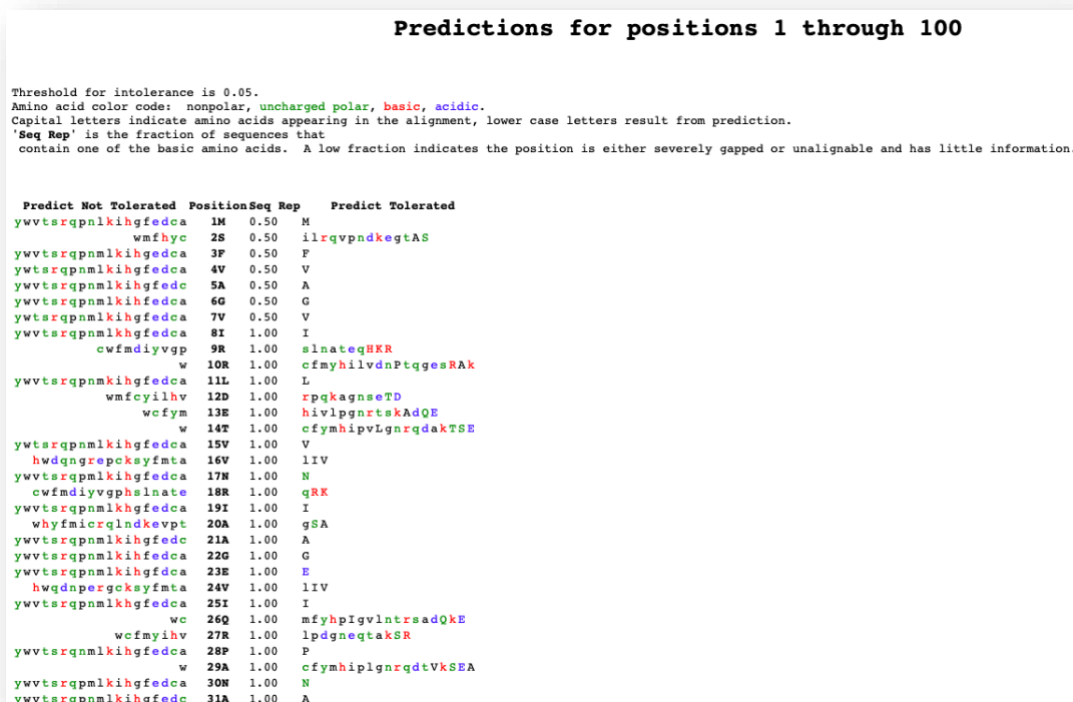
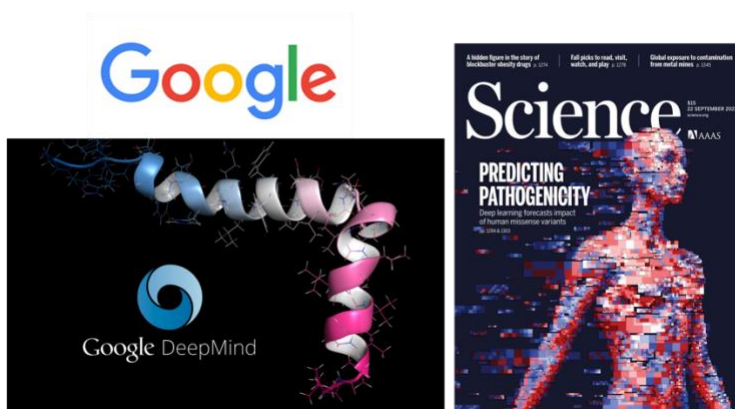


Figure 13. The predicted effect of amino acid mutations on the first 100 positions of the protein MLH1 using the SIFT method.

The traditional methods such as SIFT have long been employed for identifying disease-associated genetic variants by analyzing sequence conservation. However, recent advancements in structural

biology and artificial intelligence have begun a new era of variant prediction. Approaches like AlphaFold, developed by DeepMind, and Google's DeepMind have revolutionized the field by leveraging structural and evolutionary information. AlphaFold, for instance, predicts protein structures with unprecedented accuracy, allowing for a deeper understanding of how genetic variants can disrupt protein folding and function. These advanced methods use the power of machine learning and deep neural networks to integrate diverse data sources, including evolutionary information and protein structures, enabling researchers to make remarkably precise predictions about the functional consequences of genetic variations (Resource box 3).

Resource box 3: AI and Google's attempts at variant prediction



Identifying the pathological variants is at the forefront of medical bioinformatics and an intensive research area. This has also attracted the research focus on giant technology corporations such as Google. A team of researchers from Google DeepMind has developed AlphaMissense, an artificial intelligence-based approach to predict the pathogenicity of missense variants adapted from their previously described protein structure prediction tool AlphaFold. Given that only ~2% of the 4 million missense variants observed in the human genome have been clinically classified, AlphaMissense has enormous potential for informing diagnosis and, possibly, therapy of rare genetic disease. AlphaMissense was benchmarked against a diverse set of databases, including annotated missense variants in ClinVar and MAVE datasets (Credit: Minton, K. (2023). Predicting variant pathogenicity with AlphaMissense. Nature Reviews Genetics, 1-1.)

Our next attempt is to gain a deeper insight into disease variants by investigating how such variants may disrupt a protein's structure. The crystal structure of a molecule or biomolecule represents its three-dimensional arrangement of atoms in a crystalline lattice. It is a fundamental aspect of structural biology, providing crucial insights into the molecule's shape, interactions, and function. To convey this structural information, researchers often rely on Protein Data Bank (PDB) files. These files contain detailed atomic coordinates, bond lengths, and angles, allowing us to recreate and visualize the molecule's structure in computational software. PDB files are often built out of crystallography experiments that are used to determine the atomic and molecular structure of

crystalline materials. By directing X-rays or electrons at a crystal, we can observe the diffraction patterns produced, which provide information about the spatial arrangement of atoms within the crystal lattice.

The availability of PDB files has revolutionized the fields of molecular biology, chemistry, and drug discovery, facilitating our understanding of the molecular basis of biological processes and aiding in the development of novel therapeutics. To this aim we will open the pdb structure using a widely-used protein structural software called PyMOL. PyMOL reads a protein structure using a specific file known as a pdf file. A pdb file is often identified with an ID which can be directly queried from uniprot or PDB database. Note that the PDB ID as well as the resolution of the crystal structure can be found on the uniprot database (Figure 14).

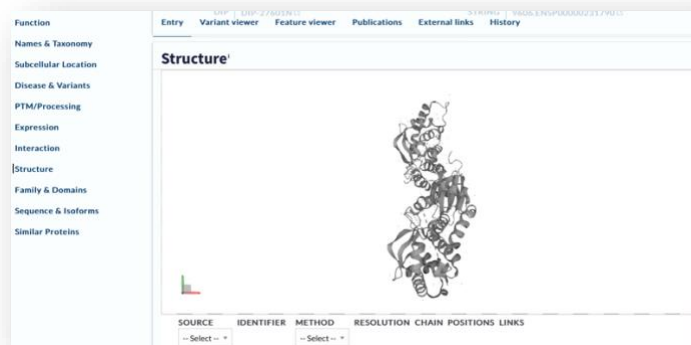


Figure 14. The “structure” information tab of the protein MLH1 in the uniprot database.








Although Uniprot database nowadays compiles the list of protein structures, this information is best retrieved from the Protein Data Bank (PDB) database. The PDB database is a globally recognized and indispensable resource in the fields of structural biology and bioinformatics. It serves as a centralized repository for the three-dimensional structures of biological macromolecules, including proteins, nucleic acids, and complex molecular assemblies. Researchers worldwide deposit, access, and analyze structural data stored in the PDB to unravel the intricate architectures of biomolecules. With over a hundred thousand structures and continuous updates, the PDB provides critical insights into the functions, interactions, and mechanisms of these molecules (Figure 15).




Figure 15. A snapshot of the Protein Data Bank (PDB) database.

In the PDB database, we can query our protein of interest either by searching its name/alias or by using the protein sequencing and the BLAST sequence similarity method. The latter approach allows us to identify structurally similar proteins within the database, which can be particularly valuable when seeking homologous structures or templates for comparative modeling and structural analysis. We can also specify the organisms which the protein belongs to, taxonomy, experimental method, or refinement resolution of the crystal structure in the query page (Figure 16).

Search Summary | This query matches 23 Structures.

Refinements     -- Tabular Report  All  Selected 

1 to 23 of 23 Structures Page 1 of 1 25 Sort by Score 

Structure Determination Methodology

☐ experimental (23)

Scientific Name of Source Organism

☐ Homo sapiens (12)

☐ Mus musculus (6)

☐ Saccharomyces cerevisiae (4)

☐ Saccharomyces cerevisiae S288C (4)

☐ Streptomyces sp. (2)

☐ synthetic construct (2)

☐ Thermus aquaticus (1)

Taxonomy

☐ Eukaryota (20)

☐ Bacteria (3)

☐ other sequences (2)

Experimental Method

☐ X-RAY DIFFRACTION (23)

Polymer Entity Type

☐ Protein (23)




☐ DNA (3)

Refinement Resolution (Å)

☐ 2.0 - 2.5 (14)

☐ 2.5 - 3.0 (7)

☐ 3.0 - 3.5 (2)

4E4W   

Structure of the C-terminal domain of the *Saccharomyces cerevisiae* MUTL alpha (MLH1/PMS1) heterodimer

Gueneau, E., Legrand, P., Charbonnier, J.B.

(2013) Nat Struct Mol Biol **20**: 461-468

Released 2013-02-20



Method X-RAY DIFFRACTION 2.5 Å




Organisms [Saccharomyces cerevisiae S288C](#)

Macromolecule [DNA mismatch repair protein MLH1](#) (protein)

[DNA mismatch repair protein PMS1](#) (protein)

Unique Ligands EDO, GOL, ZN

  Explore in 3D

6RMN   

DNA mismatch repair proteins MLH1 and MLH3

Dai, J., Chervy, P., Legrand, P., Ropars, V., Charbonnier, J.B.

(2021) Proc Natl Acad Sci U S A **118**:

Released 2021-05-19

Method X-RAY DIFFRACTION 2.2 Å

Organisms [Saccharomyces cerevisiae S288C](#)

Macromolecule [DNA mismatch repair protein MLH1](#) (protein)

[DNA mismatch repair protein MLH3](#) (protein)

Unique Ligands ZN



  Explore in 3D

Figure 16. A snapshot of the Protein Data Bank (PDB) database for the MLH1 protein.

We can particularly explore the protein structure using the "Explore in 3D" table below the image of each protein. This tab opens a molecular viewer that enables us to interactively investigate the protein's three-dimensional conformation. With this feature, we can rotate, zoom in, and analyze the intricate details of the protein's atomic arrangement, gaining a deeper understanding of its structural intricacies and potential functional sites (Figure 17). Each protein in the PDB database comes with additional information tabs that offer valuable insights beyond their structural details. These tabs cover a wide range of supplementary data, including information related to diseases, function, ligands, and experimental methods. The "Disease Associations" tab, for instance, provides details about any known associations between the protein and specific diseases.

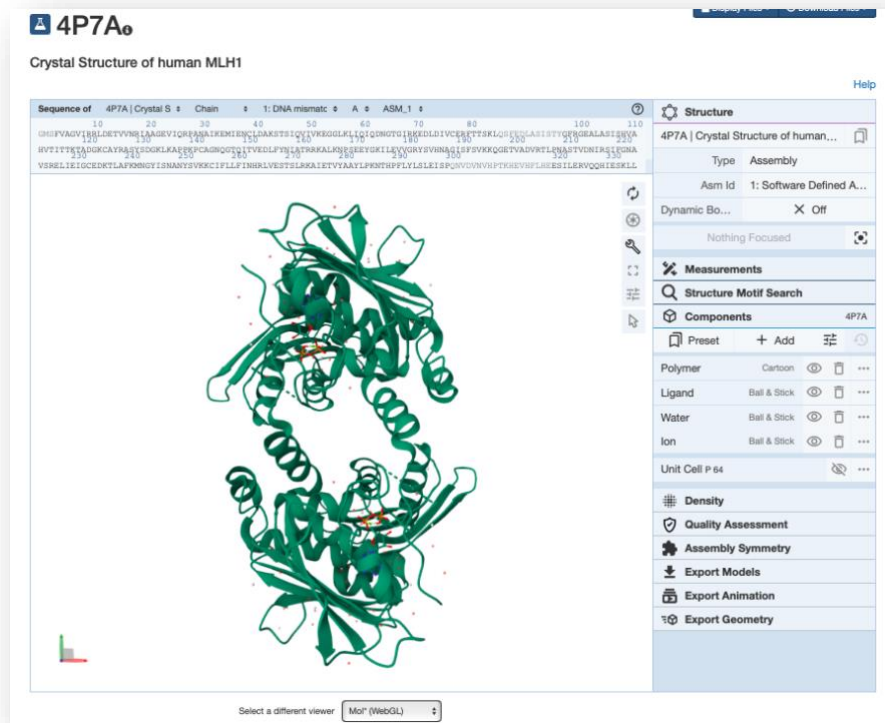


Figure 17. The molecular viewer of the PDB database for the structure of the MLH1 protein (pdb id: 4P7A)

While the molecular viewer tool within the PDB database is undoubtedly a valuable resource for examining protein structures, there are instances where we require more flexibility and advanced capabilities for in-depth structural analysis. It's for this reason that many scientists turn to specialized software like PyMOL. PyMOL is a widely used molecular visualization software tool that plays a crucial role in structural biology and bioinformatics research. Developed by Warren L. DeLano, PyMOL allows researchers to visualize and manipulate three-dimensional molecular structures with remarkable ease and sophistication. Its user-friendly interface and powerful rendering capabilities make it an invaluable resource for scientists studying protein structures, DNA, RNA, and complex molecular assemblies. PyMOL enables the creation of high-quality molecular graphics, facilitating the communication of complex structural information. It also offers a scripting language, Python-based customization, and a broad range of plugins, making it a versatile platform for various research applications, from drug discovery to protein engineering. PyMOL's ability to transform intricate structural data into intuitive visual representations has significantly contributed to advancements in structural biology and our understanding of molecular interactions. We will open the crystal structure of MLH1 protein (pdb id: 4P7A) and use the mutagenesis wizard tab of PyMOL to investigate the effect of Proline to Leucine substitutions in the protein (Figure 18).

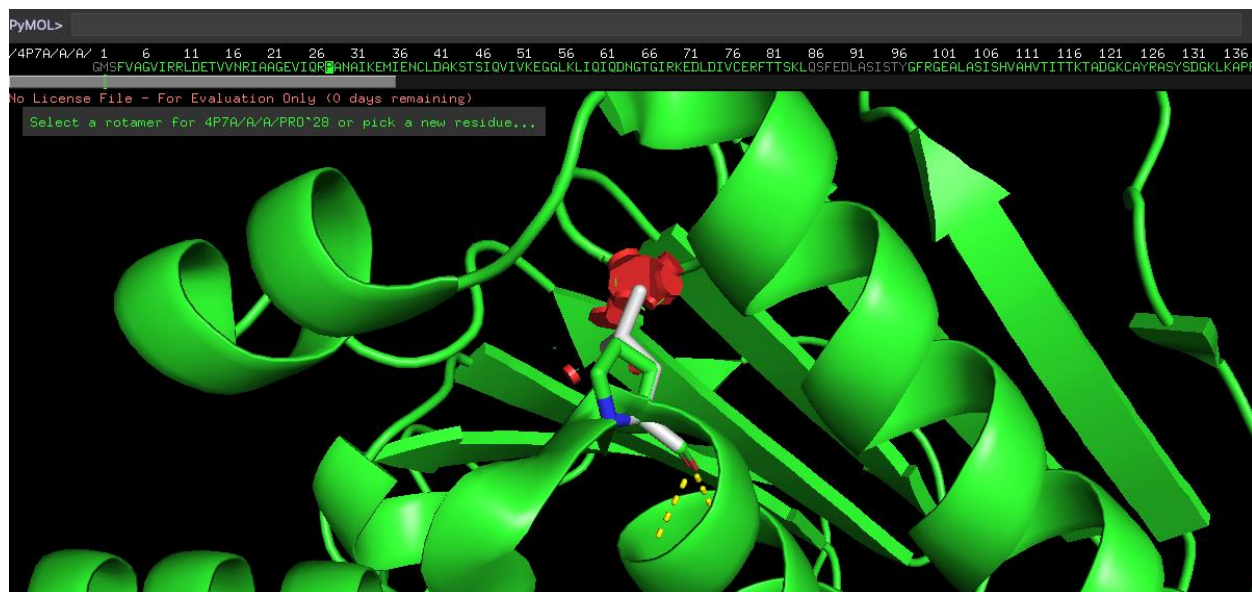


Figure 18. A PyMOL session with in-silico mutagenesis of the amino acid position 28 and the specific P28L mutation. The steric clashes are shown in red.

Structural and functional data can be employed to assess the pathogenicity of *MLH1* mutations detected during genetic testing for hereditary cancer syndromes. In this context, we highlight two such pathogenic variations, namely, the c.83C>T (p.Pro28Leu) and c.464T>G (p.Leu155Arg) mutations (Thompson et al., 2014). Using PyMol one can investigate the structural basis for the pathogenicity of *MLH1* mutations. For instance, Pro28 represents a deeply embedded amino acid residue located at the beginning of the α A helix within the ATPase domain, and it is entirely shielded from the surrounding solvent. The substitution of Pro28 with Leu in the p.Pro28Leu variant is expected to create significant steric conflicts due to the bulkier side chain of Leu. Even the most favorable rotamer conformation of Leu28 results in elevated van der Waals (vdW) strain and clashes involving Gly54, Gly55, Ile59, and Ile176, which are likely to disrupt the core protein fold (Figure 19A).

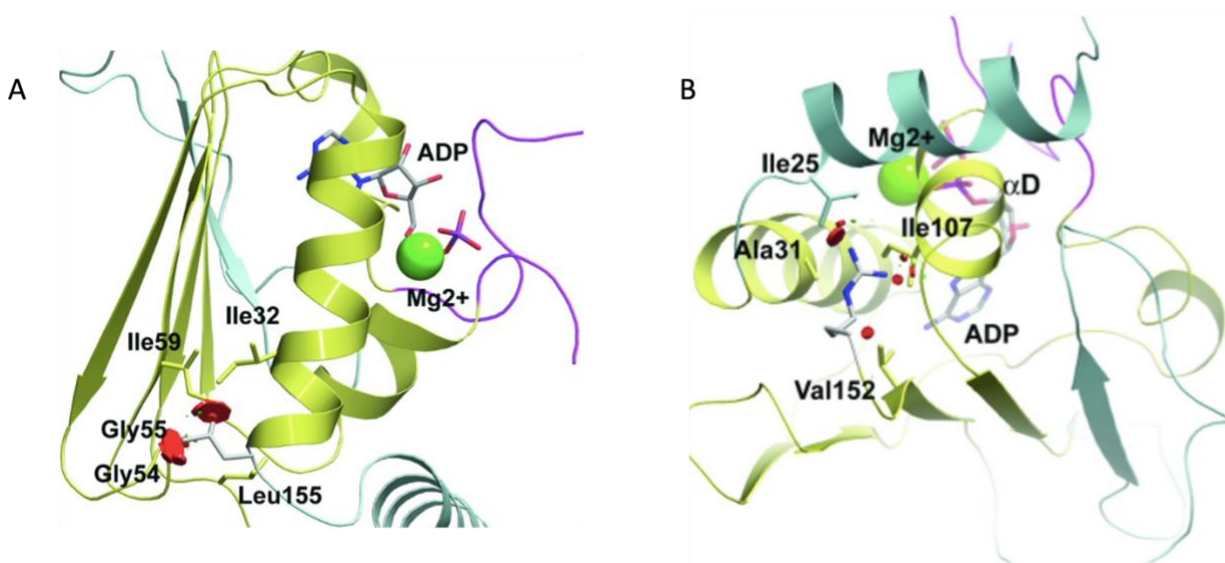


Figure 19. Structural basis for the pathogenicity of MLH1 missense variants. Ribbon diagrams showing the structural consequences of (A) c.83C>T (p.Pro28Leu) and (B) c.464T>G (p.Leu155Arg). Important amino acids around the mutation are represented as sticks. The mutation is colored grey. Red circles represent steric clashes with surrounding parts of the structure. (Credit: Wu et al., *Acta Crystallographica Section F: Structural Biology Communications* 71.8 (2015): 981-985.)

Leu155 is also found within the interior of the ATPase domain, between helix α B and the extended β -sheet, as illustrated in Figure 19B. The substitution of Leu with Arg at this position can introduce an unbalanced, buried charge, such as Arg, outside of an active site or regions stabilizing secondary structure, is often viewed as destabilizing to the overall protein structure. When the most favorable rotamer conformation of Arg at position 155 is considered, it becomes evident that this Arginine is surrounded by a cluster of nonpolar residues (Ala31, Ile25, Ile107, and Val152) and lacks the capacity to form hydrogen bonds with neighboring side-chain or main-chain atoms.

Additional tutorials

Tutorial 1: PDB database

1. Navigate to the PDB website: Go to the PDB website (www.rcsb.org) using your web browser.
2. Search for a protein structure: Enter the name or ID of the protein structure you want to search for in the search bar. You can also use the advanced search option to search for structures based on specific criteria such as organism, resolution, or molecule type.
3. View protein structure information: After finding the structure, you can view information such as the protein name, organism, resolution, and author. You can also view the 3D structure of the protein by clicking on the “3D View” button.
4. Download protein structure files: To download the protein structure files, click on the “Download Files” button. This will give you access to the coordinates file (PDB format), which contains the atomic coordinates of the protein structure, and the structure factors file (MTZ format), which contains the diffraction data used to solve the structure.
5. Analyze the protein structure: Once you have downloaded the protein structure files, you can analyze the protein structure using molecular visualization software such as PyMOL, Chimera, or VMD. These programs allow you to view and manipulate the 3D structure, perform structural alignments, and calculate various properties of the protein structure.
6. Learn more about the protein structure: You can also use the PDB database to learn more about the protein structure, such as the protein function, ligand binding sites, and interacting residues. This information can be found in the PDB header, which is included in the PDB file.

Tutorial 2: Pymol

PyMOL is a powerful software tool used to visualize and analyze molecular structures. It is widely used in the scientific community for tasks such as protein modeling, protein structure refinement, and molecular docking. Here is a brief tutorial on how to use PyMOL:

1. Installation: PyMOL can be downloaded and installed from its official website (<https://pymol.org/>). Once the installation is complete, launch the software.
2. Loading a molecular structure: To load a molecular structure, go to the "File" menu and select "Open." Navigate to the location where the molecular structure file is saved and select it. PyMOL can read various file formats, including PDB, CIF, and SDF.
3. Displaying the molecular structure: Once the molecular structure is loaded, it is displayed in the PyMOL viewer. The structure can be manipulated using the mouse or the

commands in the PyMOL interface. For example, to rotate the structure, click and drag the mouse while holding down the left button. To zoom in or out, hold down the Ctrl key while clicking and dragging the mouse up or down.

4. **Selection and visualization:** PyMOL allows users to select specific parts of a molecular structure for further analysis or visualization. Selection can be done by clicking on specific atoms or residues in the PyMOL viewer or by using PyMOL's selection language. Once a selection is made, PyMOL offers various visualization options, such as showing the selection as a ball-and-stick model, cartoon representation, or surface representation.
 5. **Analysis:** PyMOL also offers various analysis tools for the visualization of molecular structures. For example, users can measure distances between atoms or residues, calculate surface areas, and create electrostatic potential maps. These analysis tools can help researchers gain insights into the properties and behavior of molecular structures.
 6. **Saving the image:** Once the molecular structure and its visualizations are finalized, users can save the image as a file in various formats, including PNG, JPEG, and PDF.
- In conclusion, PyMOL is a versatile and powerful software tool for the visualization and analysis of molecular structures. Its user-friendly interface and wide range of functionalities make it an essential tool for researchers in fields such as biochemistry, molecular biology, and drug discovery.