

UniProt and the PDB: A frequent user's guide inside

UniProt

A scientist's first stop when trying to find out more information about a protein is typically the web database UniProt. UniProt directly provides key pieces of information about proteins (size, sequence, splicing isoforms (different versions of the protein made from the same gene due to processing differences), etc.). Rather than just showing the sequence, it draws in information from other databases in order to **annotate** the sequence (label what parts do what and/or are predicted to do what). And UniProt provides quick links to associated entries in the databases it draws from. UniProt also serves as an interface to facilitate easy use of tools for exploring the protein in more detail (e.g. ProtParam) and finding (BLAST) and comparing (Clustal/Align) similar proteins.

UniProt is largely sequence oriented. This is reflected by the core nature of UniProt entries. Although we often use simplified language and say that each entry corresponds to a single protein, it's a little more complicated than that... Technically speaking, each UniProt entry contains information about all the versions (isoforms) of a protein made from a single gene. A single UniProt entry thus might contain information for multiple isoforms that come from things like alternative splicing. Further complicating matters, some functional forms of proteins contain multiple subunits made from different genes. Each of those subunits would have its own UniProt entry, meaning that the functional protein will be associated with multiple UniProt entries. For example, adult hemoglobin consists of 2 copies of α -globin chains and 2 copies of β -globin chains. α -globin and β -globin each have their own entry in UniProt.

Data within UniProt is housed in 3 databases:

- **UniProtKB:** UniProt KnowledgeBase. It includes annotated entries-one per protein-encoding gene and is typically where you want to be if you're looking for information about a specific protein.
 - o Further split into:
 - **Swiss-Prot** (sp)- reviewed, manually curated (more reliable, more info, often something people have actually experimented with in a lab). Swiss-Prot entries are indicated by a gold paper icon.
 - **TrEMBL** (Translated EMBL Nucleotide Sequence Data Library)(tr)- unreviewed, autocurated (less known, be more cautious, most stuff from less-studied organisms is here). TrEMBL entries are indicated by a gray paper icon.
 - o The amount of annotation varies from entry to entry and is reflected in the annotation score, which can be found in the upper right of the entry.
- **UniRef:** UniProt Reference Clusters, which you search protein sequence data more efficiently. It includes data clustered by % sequence identity (100, 90, or 50 for UniRef100, UniRef90, and UniRef50, respectively), then represented by a reference entry. Clusters are giving accession codes containing "UniRef100", "UniRef90_", or "UniRef100_", followed by the code of the reference entry.
- **UniParc:** UniProt Archives. It includes ALL non-redundant protein sequence data. Data in UniParc is the least reliable, but it ensures everything is stored permanently somewhere.
- More here: <https://youtu.be/uXbwNHnhVoA> & <https://www.ebi.ac.uk/training/online/courses/uniprot-quick-tour/the-uniprot-databases/>

Each entry is given a 6-10 character UniProt accession number/ID (e.g., A2BC19, P12345, A0A023GPI8). Don't confuse this with the gene name that is often listed next to it, often in the format GENE_SPECIES.

You can add entries and/or sequences (including sequence variants) to your **basket** by clicking the little "Add to basket" basket icon by the entry or sequence. You can then quickly re-find and work with them by clicking on the basket icon () in the upper right corner of the page. If you click on items in your basket, you can visit their entries. Or, if you click the selection boxes next to items, you can use tools such as Align and BLAST on them. *Note: items in the basket are divided up into tabs indicating which of the sub-databases of UniProt they come from (i.e., UniProtKB, UniRef, or UniParc).*

Just to the left of the basket icon, you'll find the **toolbox** (). This is where results of your jobs (alignments, BLASTs, and ID mappings) are saved. Tool results are saved for 7 days, with input parameters stored longer for quick modifications &/or reruns.

You can click on the rerun icon () next to an entry title to modify the parameters and rerun and/or rerun after the data has been discarded.

First, an overview of some of the main features of UniProt

Key information provided directly in UniProt entries

- **Protein size(s)**
 - # of amino acids
 - Molecular weight is given in Daltons (Da), where a Da = a g/mol. We often talk about protein sizes in kDa = 1000 g/mol
- **Protein sequence(s)**
- **Annotations pulled in from other databases**
 - Known (and/or predicted) features of the protein such as binding sites and enzyme active sites
 - Known information about the protein's function, localization within or outside of cells, and interactions.
 - Known information about variants of the protein.

Key tools you can access through UniProt

- **ExPasy ProtParam**
 - Information provided includes:
 - Predicted isoelectric point (pI) – pH at which a protein is electrically neutral overall
 - Conversion factors to calculate concentration from UV280 absorbance
 - How many times each amino acid occurs
 - You can get to it by clicking on the "Tools" drop-down menu above a sequence
- **BLAST (Basic Local Alignment Search Tool)**
 - Allows you to find proteins with similar sequences (e.g., different versions of it in the same organism or versions of it in other organisms)
 - You can get to it from "Tools" – accessed from an entry or from your basket.
 - Note: there are options to choose which databases to search against (e.g. you can narrow down your search UniProt KB-Swiss Prot (which only contains manually reviewed entries) or one of the "UniRef" options (which groups together highly similar entries).

- **Clustal Omega (“Align” feature)**

- Align and compare the sequences of multiple proteins
- You can get to it from “Tools” – accessed from an entry or from your basket

Now, here’s some more detailed information . . .

Choosing an entry: Sometimes, there might be multiple similar-looking entries and you aren’t sure which is the one you want. Things get trickier if you’re looking for a minimally-studied protein. But, if you’re looking for a well-characterized one (especially a human and/or mouse one), a few things to look for are:

- 1) **Species!** (you can restrict this when you set up your search, or filter your results, or just look at the hits)
- 2) **Review status** – if you’re looking for a human protein and/or a well-studied protein, the entry will likely be in the manually reviewed, UniProt KB – Swiss Prot, side of UniProt, and thus be indicated with a gold paper icon, as opposed to being on the unreviewed, TrEMBL side, as indicated by a gray paper icon.
- 3) **Annotation score** – if you’re looking for a well-studied protein, there will likely be lots of information discovered about it that was subsequently mapped onto the sequence in the form of annotations telling you what parts do and/or are predicted to do what. The more annotations there are, the higher the annotation score (out of 5), which you can see at the top right of an entry page.
- 4) **Number of publications** – again, the higher this number, the more likely you’re looking at the “popular” protein you’re looking for.
- 5) **Warning messages** – sometimes, there are red or yellow boxes on an entry page warning you that there are potential concerns with an entry, etc.

Navigating a UniProt entry

The main Entry page: There’s a ton in a UniProt entry, and I’m not going to go into it all, but you can jump to different sections to explore them from the left-hand menu. A few key ones:

- **Function:** Here’s a lot of key information as well as a “Features” section showing the locations of binding sites, active sites, etc. in the sequence. Some of these features have experimental evidence, others are predicted, and that info is all indicated, with links to learn more
- **Structure:** Be careful here . . . UniProt entries display and direct you to ALL PDB entries (i.e., all 3D structures) containing part or all of that UniProt entry’s chain (in reverse alphabetical order). The structures displayed in UniProt may contain more or less than you want and it’s up to you to check them out individually to see (or save yourself some time by starting in the PDB).
 - Some structures might only contain small structured regions of proteins (such as individual domains) & some structures might contain the protein you’re interested in bound to another molecule as part of a complex.
 - Look to see which parts of the chain are - physically included in the PDB entry (but note that some of these might not be included in the structural model due to insufficient data) & go to the PDB for more details.
 - Compare position numbers to the # of amino acids in the chain to get a sense of coverage.

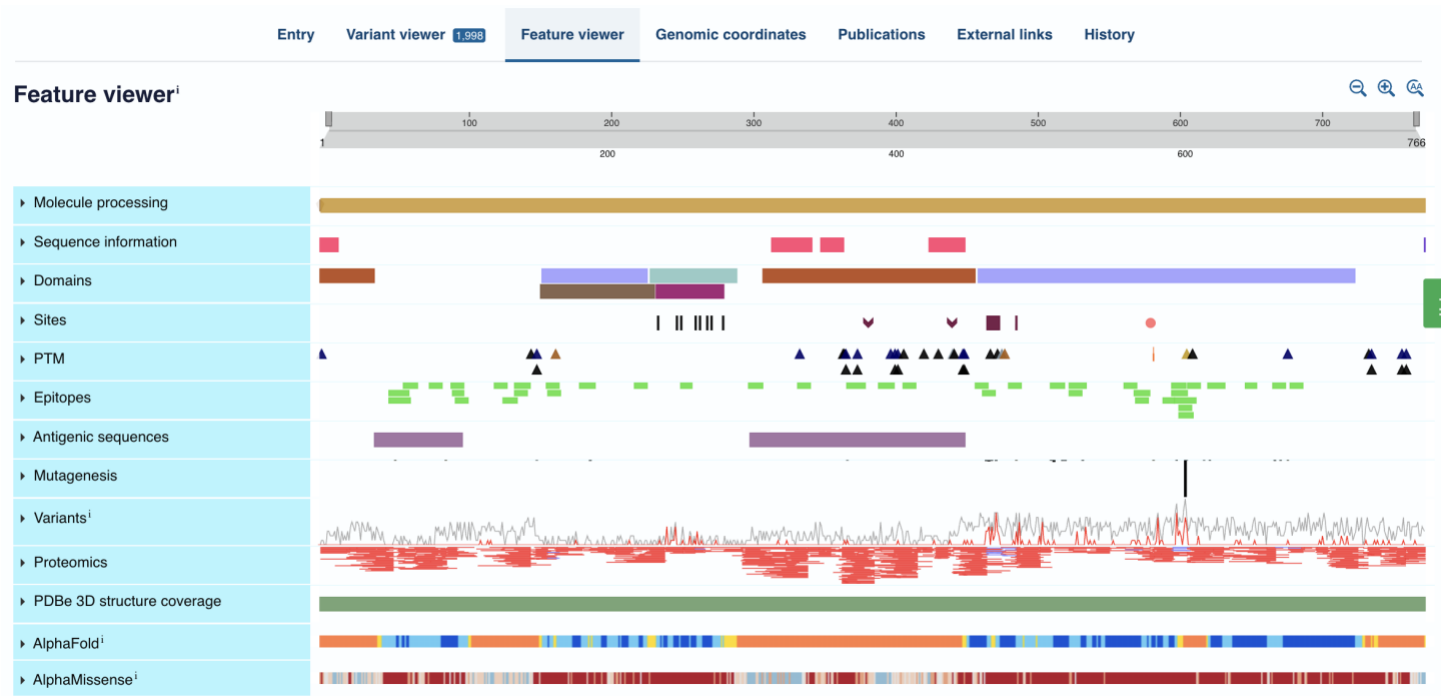
- **Sequence:** In the “Sequence” section of a UniProt entry page, you will see information about each of the different isoforms of the protein (splicing variants, alternative start site, etc.). The main (canonical) isoform is shown at the top and its sequence is displayed. Any additional isoforms are listed below, with information about how they differ from the canonical sequence. You can click on them to show the sequence.
 - Although the canonical isoform is used by default if you add a UniProt entry to your basket and/or run a tool on it, you can also/alternatively work with any of the isoforms. To do this, click on “Show sequence” and you’ll see an “Add to basket” option as well as various tools.
 - Above each sequence, you’ll see options to download or copy the sequence
 - Note that if you click “Download” it will take you to a new tab displaying the sequence in FASTA format. Right-click and “Save as . . .” to save the file. More here: <https://thebumblingbiochemist.com/365-days-of-science/fasta/>
 - Above each sequence, you should also see a “Tools” drop down menu. This provides you with direct access to a variety of tools, my favorite being ProtParam, which lets you find, among other things: Predicted isoelectric point (pI) – pH at which a protein is electrically neutral overall; Conversion factors to calculate concentration from UV280 absorbance; How many times each amino acid occurs.
 - *Note: ProtParam will open in a new browser tab. It will give you the option to select a part of the sequence to analyze, but I typically just use the default option, which corresponds to the full sequence. To run on the default, just click “Submit” (no need to select anything first).*

That was a lot of information it gave you – and that’s just on the main Entry page! Now let’s look at some of the other tabs . . .

From sequencing the genes of lots and lots of people, scientists know that the sequence of proteins can differ slightly in some individuals. Often this variation involves point mutations that lead to substitution of one amino acid for another. These and other variants are mapped onto the protein sequence in the “**Variant viewer**” tab, along with any known information about whether the variant has any known functional consequences and/or is associated with any conditions. If desired, you can use the checkboxes on the left-hand panel to filter results and/or command/control + “F” to search the page for a term of interest. You can also click on individual variants to find more information about them.



The “**Feature viewer**” tab lets you investigate annotated features of the protein. Here, you will find different types of features (binding sites, active sites, domains, post-translational modifications (PTMs), etc.) displayed in different rows/tracks.

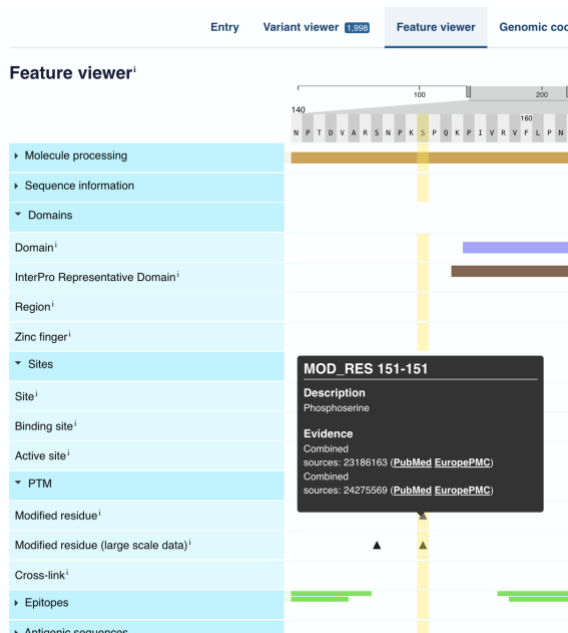


The bumbling biochemist (Bri Bibel), last updated 9/19/26

Clicking on the arrow next to a track name will open up additional tracks (e.g., “Sites” may further split into tracks for binding sites and active sites).



By default, it shows a zoomed-out view, but you can zoom in using your trackpad or the magnifying glass icon on the upper right of the display. You can also use the magnifying glass with “AA” () to quickly zoom in on the sequence. You can click on individual feature icons (dots, bars, etc.) to get more information about them.



Below the sequence annotations, you’ll also see a structure viewer window, where you can see where features map onto various structural models of the protein.

Other tabs include “**Genomic coordinates**,” where you can find information about exons encoding it, etc., with links to their entries in DNA databases; “**Publications**,” where you kind find papers about it and sources of the annotations, etc.; “**External links**,” which will link you out to entries associated with the protein in other databases (the sources of a lot of the annotations!); and “**History**,” which just tells you about when changes were made to the entry.

Now, let’s check out some of the tools you can run through UniProt. There are 3 main ones accessible from your basket: BLAST, Align, and Map IDs (which I’m not going to go into because it’s more “boring” – it “just” lets you find associated entries in other databases).

BLAST (Basic Local Alignment Search Tool)

BLAST (specifically blastp) searches protein databases for similar protein sequences and takes into account both sequence identity (are the amino acids in the sequences identical at a given location) and similarity (are the amino acids in the sequences biochemically similar at a given location). You can access it either from the main entry page, or from your basket (with an entry selected). Tools → BLAST.

There are a variety of options you can choose from to help restrict your search. For example, you can restrict to just search UniprotKB Swiss-Prot (reviewed entries), one of the UniRefs, or a specific species (under Restrict Taxonomy). You can also limit the number of hits to make it run faster.

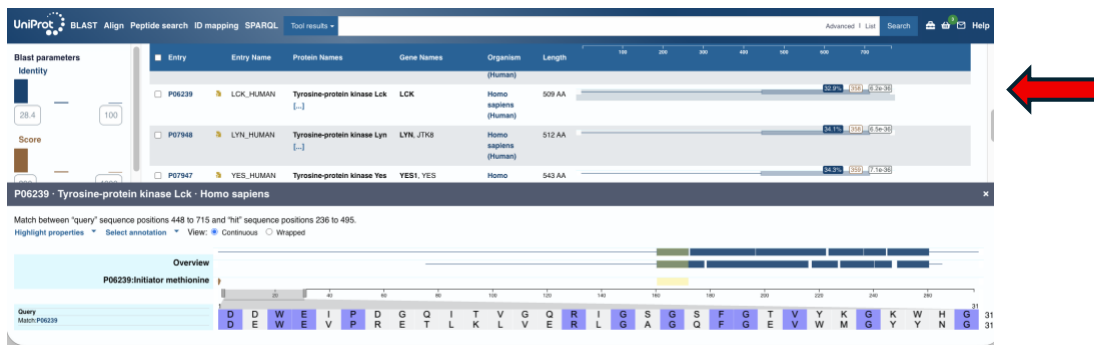
A parameter called “HSPs per hit” lets you limit how many (if any) different aligned portions (High-Scoring Segment Pairs) to allow for the same entry. For example, if a sequence can be aligned to the same entry in 2 different places, do you want each of those alignments to show in your hits list, or just the best?

You can also change the name of the run if you want, then press “Run BLAST.” Wait for the page to show “Complete” and then click on “Complete” to view the results. *(Note: you can navigate away and find the results at any point from the toolbox icon on the top right of the page. A green dot will appear when there are new job results.)*

On the results page, you will see a list of UniProt entries for proteins with similar sequences to the search query, along with (on the right) graphical bar indicators of the degree of similarity to the search query. The 3 main scores it provides are:

- **Identity:** The percentage of amino acids that are identical in the aligned sequences.
- **Score:** An overall measure of how similar the sequences are, taking into account gaps, coverage level, etc. The higher, the better.
- **E-value:** A statistical measure . . . Long story short, (unlike “Identity” and “Score” values) the lower the E-value number, the less likely it is to come up as a “hit” due to chance alone (as opposed to being biologically relevant). You typically want it to be less than 0.1 or so. You can, if desired, limit the E-value when setting up a job (it defaults to 10).

If desired, you can click on one of the graphical bars to view the alignment to the search query.



You can check the boxes next to the entries of interest to add them to your basket for easy reference and/or tool-running.

One thing you might want to do with the sequences you find from BLAST is align them using the tool Clustal Omega. . .

Align (Clustal Omega)

Note: Clustal Omega is a Multiple Sequence Alignment (MSA) tool, so it tries to find the best alignment for the group of sequences, as opposed to BLAST, which is a pairwise alignment tool that aligns individual sequences “one at a time” to a single search query.

UniProt makes it easy to align protein sequences. Just click on them in your basket or on a results page and use Tools → Align.

Note: you can add other sequences here as well and/or modify the names for easier interpretation of the results. To change the name, change the text in the row(s) starting with a carat >. These are the FASTA header rows and, as long as you keep the carat, you can change the rest of the header to what you want to call the sequence. More here: <https://thebumblingbiochemist.com/365-days-of-science/fasta/>

Wait for the page to show “Complete” next to the job submission and then click on “Complete” to view the results.

You should now see an alignment of the sequences.

Above the alignment on the top left, you should see a “Select annotation” drop down menu. You can use it to map annotated features (e.g., binding sites, active sites) from one of the entries onto the sequence alignment. This can help you, for example, quickly see whether a feature is conserved. *Note that the number positions in the sequences may differ between the proteins but you can find the corresponding numbers by clicking on a feature in the alignment.*

In addition to showing features, you can color based on various properties using the “Highlight properties” drop down menu (next to the “Select annotations” menu). One useful one is to color the alignment by “Clustal.” This will color the amino acids by biophysical properties such as charge and polarity (see below), which can be useful for predicting how consequential differences in the amino acid sequence at a position may or may not be.

CLUSTAL coloring colors each conserved or semi-conserved* amino acid residue in a sequence based on its biophysical properties:

- Light blue: hydrophobic (nonpolar)
- Red: positively-charged (basic) (yes, that’s weird and confusing because blue is typically used for this but it is what it is!)
- Purple: negatively-charged (acidic)
- Green: polar, uncharged
- Salmon: cysteine
- Orange: glycine
- Yellow: proline
- Teal: aromatic
- White: unconserved

*Whether or not a residue is colored depends on the percentage of amino acids at that position with similar biophysical properties. Prolines, cysteines, and glycines, are colored even if not conserved. *Note: Some of the thresholds are more complicated than others and allow for some leniency in classification if stricter limits are*

The bumbling biochemist (Bri Bibel), last updated 9/19/26

surpassed (e.g., a K or an R is considered conserved enough to color red if >60% of the residues at that position are K or R, OR if > 85% are K, R, or Q. <https://www.jalview.org/help/html/colourSchemes/clustal.html>

More on CLUSTAL formatting here: <https://youtube.com/shorts/RUgtc378oqE> & <https://thebumblingbiochemist.com/365-days-of-science/proteinalignments/>

Another helpful feature of the Align results is the percent identity matrix. It (in a pairwise fashion) tells you how identical each sequence is to each other sequence.

The Protein Data Bank (PDB)

Scientists use various methods (x-ray crystallography/x-ray diffraction (XRD), cryo-electron microscopy (cryo-EM), and nuclear magnetic resonance (NMR) being the main three) to determine (“solve”) the 3D atomic structures of proteins and other macromolecules. Data for these structures are housed in the Protein Data Bank (PDB) – an international effort with several hosting sites with different interfaces. I typically use the [RCSB PDB](#), which is hosted in the US. In addition to storing structures, the PDB is itself a great tool for exploring and comparing structures.

Key features of individual PDB entries

- Downloadable and fetchable data for structural models
- Interactive model viewer
- Information about what molecules are included in the model
- Experimental details
- Quality indicators (“validation scores”) for both the structural model and the experimental data used to build the model

Before going into how to use the PDB, it’s important to understand a key difference about the nature of UniProt entries (sequence centered) and PDB entries (structure centered).

Fundamental differences of UniProt & PDB entries

- UniProt entries correspond to different proteins (more specifically, different protein chains/subunits)
 - Each UniProt entry can be associated with multiple PDB entries
- PDB entries correspond to different structural models
 - A single PDB entry might contain multiple different proteins, thus be associated with multiple UniProt entries

A key concept to grasp is that when we talk about experimental “structures,” we are typically actually referring to **structural models**. These models are built based on experimental maps/“density” showing a sort of fuzzy/sausagey outline of the atomic positions. The higher the resolution of the data, the crisper the maps, and the easier it is for scientists to more confidently model in the precise atomic positions.

Resolution refers to how close together two things can be before you stop being able to tell that they’re two separate things. Thus, smaller numbers correspond to higher resolution. Resolution is typically reported in units of Ångstroms (Å), where $1 \text{ Å} = 0.1 \text{ nm}$. Because the average C-C bond length is ~ 1.09 , “atomic-level” resolution requires a resolution of $\sim 1.5 \text{ Å}$ or so. More commonly, resolutions for structures are lower (2-3 Å is common).

All else being equal, higher resolution (corresponding to lower numbers) is “better” because it means the underlying data used to build the data was stronger overall. But there are a couple caveats: 1) the model could still be built poorly and 2) regions of the data might be weak. Parts of the structure, although physically present, might be too “floppy” to be able to collect sufficient data to model in atomic positions – thus, these are left as “unmodeled regions” and appear as grayed out letters in the sequence and dashed lines in the structural model.

Similar to UniProt entries, PDB entries also have unique accession codes/IDs. These IDs have historically been 4 characters, but are transitioning to 12 characters because they're running out of 4 character ones. You can adapt old codes by adding "pdb_0000" to the beginning of them. For example, 9cn3 becomes pdb_00009cn3. More here: <https://thebumblingbiochemist.com/365-days-of-science/pdbids/> ; YouTube: <https://youtu.be/U-MEtWaFNdl>

The PDB also hosts predicted structures, which it refers to as Computed Structure Models (CSMs), mainly AlphaFold (AF) structures. These structures come from AI programs trained on data from experimentally solved structures. They typically predict the structure of the full-length protein, whereas experimentally solved structures might only contain parts of proteins, such as isolated domains (which are easier to work with). AlphaFold predicted structures are often colored by confidence with orange indicating low confidence and dark blue indicating high confidence. The accession codes of AlphaFold-predicted models start with "AF" to indicate AlphaFold. More here: <https://thebumblingbiochemist.com/365-days-of-science/interpreting-alphafold-predictions/>

With this context in mind, let's see what there is to see in a PDB entry (which remember, corresponds to a single structure as opposed to a single protein/protein chain)!

You can find a PDB entry a couple main ways:

- 1) By entering the PDB ID (accession code), such as can be found at the bottom of the paper reporting the structure (frequently in the "Data Availability" section)
- 2) By using search features of the PDB. You can do a simple search by protein name, UniProt ID, etc. from the top of the page or use the advanced search feature to specify things further.

When you go to a PDB entry in the RCSB PDB (the US-hosted portal to data hosted by the worldwide PDB (wwPDB)), you will see a number of tabs, with the "**Structure Summary**" page showing by default. This page provides an overview of a structure - measurements of the quality of the structural model and underlying data, a description of what the structure contains, information about the source of the sample, information about the experimental method, and links to any accompanying paper and other related structures.

A couple **features for exploring individual structures** that you can get to from the main page (or the other tabs) are:

- **Validation Reports**, including 3D view - points out potential problems with the structural model's geometry (bond angles, etc.) and clashes (atoms too close in space) as well as places the model doesn't match the data well. More here: <https://thebumblingbiochemist.com/365-days-of-science/structurequality/> & <https://youtu.be/0AswlZBLpKw>
- **3D Sequence Viewer** - maps annotations like functional sites to the 3D structure. More here: <https://youtu.be/6zKBBkzl-4A>

Additionally, since the PDB houses lots and lots of structures (over a quarter million), it also offers **tools to help you find and compare structures**. A few key tools:

- **Advanced Search Query Builder:** Search for proteins based on similar sequences, structures, motifs, experimental methods, ligands, etc., narrowing things down using logic operators (AND/OR/NOT) etc. to find exactly what you want. More here: <https://youtu.be/iSM6orTve-Y>
- **Pairwise Structure Alignment:** Compare structures based on their shapes - Unlike Clustal, which aligns protein sequences based on their sequences, the PDB Pairwise Structure Alignment tool aligns one or more structures (experimentally determined or predicted) and their sequences to a reference structure and sequence based on their shapes. More here: <https://youtu.be/MOBMBuAIAKo>
- **Structure Grouping:** Group structures by % sequence similarity or Uniprot accession (if you want grouped by exact same protein). You can then further filter the results (Click on a bar to filter based on a property (ligand, etc.), shift+click to see the individual structures in the group with that property). And view &/or export sequence and structure alignments. This helps you quickly explore structures within & between groups of closely or more distantly related proteins. Or find the "best" structures of a specific protein for your needs. You can get to it from advanced search or directly from an entry. More here: <https://youtu.be/07gRZXI5uCU>

The PDB has some great **educational resources** as well:

- **RCSB Help Guide:** Super detailed guide to all things PDB, with hyperlinked examples, etc. <https://www.rcsb.org/docs/tools/pairwise-structure-alignment>
- **PDB 101:** Great educational and teaching resources, as well as a fun molecule of the month feature: If you go to this page in the "learn" tab, you'll find this Introduction to PDB Data: <http://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/introduction>

Sometimes, atomic structural models don't look like what you expect

- More than you expect?
 - The protein might be part of a complex
 - You could be seeing a sort of artifact of the experimental method (e.g., the asymmetric unit in a crystal structure (smallest unique part of the crystal) might be different from the functional form (biological unit/assembly).
- Less than you expect?
 - The scientists might have worked with a truncated (shortened) version of a protein, such as an isolated domain or even just a small peptide from it
 - Parts of the structure, although physically present, might be too "floppy" to be able to collect sufficient data to model in atomic positions – thus, these are left as "unmodeled regions" and appear as grayed out letters in the sequence and dashed lines in the structural model
 - For predicted structures (e.g., AlphaFold structures), you might see less than expected if it's a multi-chain protein & the predicted structure is only for one of the chains.

Choosing PDB entries

- Make sure the model includes the region you're interested in
 - And, if there's a region you really care about, carefully inspect that region of the model (look in 3D at the underlying experimental maps and the fit of the model into the map).
- Check to see if there are other molecules included in the structure you do or don't want included
- Check to see if there are introduced mutations you do or don't want
- All else being equal, a structure with a higher resolution (corresponding to lower numbers) will typically be more reliable, because the underlying data used to build the model is "crisper"
- Look to the validation slider bars to see how well the model matches the underlying data and how physically/geometrically plausible the model is.
 - For crystal structures, Rfree tells you about how well the model matches the data. The rest of the bars tell you how physically & biochemically "reasonable" the structure is.
 - The sliders compare this structure to other structures in the PDB.

Sometimes, a good place to start if you're looking for a structure to use the Group Structures feature of the PDB, which allows you to group structures by % sequence similarity or Uniprot accession (if you want grouped by exact same protein). You can then further filter the results (Click on a bar to filter based on a property (ligand, etc.), shift+click to see the individual structures in the group with that property). And view &/or export sequence and structure alignments. This helps you quickly explore structures within & between groups of closely or more distantly related proteins. Or find the "best" structures of a specific protein for your needs. You can get to it from advanced search or directly from an entry. More on this feature here:

Navigating PDB entries

On the main Structure Summary page, you'll see a number of features. . .

At the top you get a quick "snapshot" structure-wise and experiment-wise. On the top left you can click through any biological assemblies, asymmetric units, etc. (see the "Crystallographic quirks section of the guide).

Next to that, you get info about the experimental method used to solve the structure, the resolution, the protein source, etc.

On the right, you'll see PDB validation sliders telling you about the quality of the structural model (more on these below).

If there is a publication associated with the structure, it will be listed in the "Literature" section. If no paper is associated, it will say, "To be published."

In the "Macromolecules" section, you'll see information about the proteins/protein chains, DNA, RNA, etc. contained in the structure, as well as what chain(s) (A, B, etc.) correspond to them. If you click on the UniProt ID next to "Explore," you can access and use the "Group Sequences" feature. It will take you to a page containing all the structures of that UniProt accession aligned in 3D with feature annotations. You get a quick overview of the annotations, which is especially helpful for seeing what regions are unmodeled (gray bars in the unmodeled track).. Note that differences from the UniProt canonical sequence are indicated

The bumbling biochemist (Bri Bibel), last updated 9/19/26

by pink dots in the UniProt track, but these might be because it's a different isoform, rather than being actual mutations.

In the "Small molecules" section, you may find things like intentional ligands (inhibitors etc.) as well as "artifacts" like buffer components that were included in the structural model.

The "Structure" tab will allow you to view the 3D structure and play around (making use of the web-based molecular visualization tool Mol*). You can click on parts of the sequence (other than grayed-out, unmodeled, regions) to see where they are in the structure and vice versa. You can select which chain's sequence to show on the top left and which assembly or model to show on the top right. To show the maps, click on "Density" and enable from the right-hand menu. If desired, you can take snapshots using the lens icon in the right-hand menu.

The "Sequence" tab features a variety of annotations (functional sites, secondary structure, disorder, etc.) in different rows, similarly to you see in UniProt. Here, however, you can click on "Explore Sequence Annotations in 3D" to see those annotations mapped onto the structure. You can click on individual annotations to zoom in on a site. You can also view annotations regarding structural quality. *Note: be sure to select the chain corresponding to the molecule you want to assess.*

The "Experiment" tab provides details about things like crystallization conditions, data collection parameters, and how the model was built and refined ("Table 1" type info. More here: <https://youtu.be/ewGz-O7dAS0>).

PDB validation

The better the experimental data collected, the better the chances the scientists can build a well-fitting, geometrically-sound atomic model. But it's not guaranteed that they'll succeed! So it's important to evaluate models for their soundness and there are parameters we can use to evaluate the quality of both the model and the underlying data. And the PDB (Protein Data Bank) collates it all for you in a "PDB Validation Report" and even lets you see problem areas in 3D.

At least on the RCSB website, an overview of the validation report is shown in a quick visual form as slider bars. And then you can click to see the full report. These bars contain some global statistics that tell you about the overall fit and the molecular feasibility of the model, shown as red (worse than average) to blue (better than average) slider bars in their wwPDB validation report. "Average" here is compared to all other structures solved using that same experimental technique (e.g., x-ray crystallography or cryo-EM) (black bars) as well as to all other structures solved using that same experimental technique with similar resolution (unshaded bars).

Note: it's known from really high-res structures that it's normal for actual proteins to physically have some outliers and deviations from ideal, so it's also normal for models to have some. If the outliers are too low, it could mean that the model was being refined too much to be "ideal" based on geometry instead of being refined to actually match the data. Refinement should include a mixture of the two, and you should end up with a low, but not too low, level of weirdos.

Above the sliders, you'll see options to download a full, detailed, report or to view a "3D report" which can help you see where potential problems are in the structure. When you click on "3D report" you will be directed to the "Structure" tab, where you can toggle the types of report data you want to map onto the structure using the

“Quality Assessment” drop-down menu on the right-hand side of the page. “Validation Report” will show clashes and other geometry-type problems whereas “Experimental Support Confidence” will provide color indications of how well the model matches the data in specific locations.



This can be very helpful if you're really interested in a specific region. You can also look at something called a B-factor which can tell you about how reliable the pinpointed position in the model is likely to be. You can set the PDB to color based on uncertainty/disorder if you want to see that.

Much more on validation scores here: <https://thebumblingbiochemist.com/365-days-of-science/structurequality/> & <https://youtu.be/0AswIZBLpKw>

Crystallographic quirks . . .

A crystal is made up of lots and lots and lots of copies of the molecule(s) you're trying to look at arranged in a repeating 3D pattern. Each copy is too tiny to see by itself, but if you put a ton of them together, their signals can contribute to each other such that you get a strong enough spots on the diffraction pattern to work back from.

That protein or complex you're interested is looking at is usually the “biological assembly” which is the form of the molecule that is thought to be the active form. But, if you look at the “Model” in the 3D viewer, you might see something different. This is because it's showing you the “asymmetric unit” which is the smallest repeating part of the crystal. If you know the structure of the asymmetric unit and the dimensions of something called the unit cell, you can recreate the whole crystal just using symmetry operations (rotate, move left, move up, etc.). Confusingly, that asymmetric unit might itself contain more than one copy of the biological assembly.

The copies of the biological assembly within the asymmetric unit are identical in their sequence but they might be slightly different in their shape, and during the structure-solving they are therefore treated separately. In the PDB, these “different copies” within the asymmetric unit are called “instances” so if you're confused by that, that's what it refers to and you can choose which one you want to look at.

In sum:

- Crystal lattice: the “whole” crystal - it contains lots and lots (billions typically!) of copies of the thing you care about
- Unit cell: the basic repeating unit of a crystal, like the bricks in a brick wall
- Asymmetric unit: The smallest “unique” part of the crystal
- Biological assembly: The biologically relevant form - might be different than the asymmetric unit.
- Each unit cell has one or more full or partial copies of the asymmetric unit
- Copies of the asymmetric unit are related to one another symmetrically through space group transformations to build up the unit cell. Unit cells are positionally related through lattice transformations to build the entire crystal.

More here: <https://thebumblingbiochemist.com/365-days-of-science/using-the-pdb-to-investigate-x-ray-crystallography-structures-resolution-r-factor-outliers-more/>

Pairwise Structure Alignment of protein structures through the PDB

Unlike Clustal, which aligns protein sequences based on their sequences, the PDB Pairwise Structure Alignment tool aligns one or more structures (experimentally-determined or predicted) and their sequences to a reference structure and sequence based on their shapes. *Note: Although you can see >2 structures overlaid, the alignments themselves are pairwise – each structure is aligned to the first structure in the list (so be sure to put your desired comparator first!).*

You can give it structures to align from (among other options) ...

- PDB ID (for specific structure)
- Uniprot ID (for specific protein)
- AlphaFold Predicted structure
- Uploaded structure file

You can choose from a variety of alignment methods based on how similar the proteins &/or their structures are. Choose a rigid method to directly superimpose the structures without altering them (good if very similar) or a flexible method (which breaks the structure up and aligns parts) for ones with large structural differences.

More on the different options here: <https://www.rcsb.org/docs/tools/pairwise-structure-alignment>

In terms of how to use it . . .

1. Navigate to the “**Pairwise structure alignment**” tool from the “Analyze” drop down menu from the main menu at the top of the page.
2. Use the dropdown menu on the left (the downward “v” to the left of the entry box) to change the search criteria as desired (by default it’s set to PDB ID).
 - a. Do this *before* entering the ID, as it will reset the box. By default it only shows 2 rows, but you can click the + button to add additional ones.
 - b. Be sure to enter the one you want to compare to first.
3. Enter the search criteria (e.g. PDB ID). It will then allow you to select a specific structure and/or chain.
4. Select a desired alignment method from the bottom left dropdown menu.

You will now see an overlay of the fetched structures along with a structure-based alignment of their sequences (as opposed to the sequence-based alignment provided by Clustal Omega in UniProt).

If you want to save the alignment, you can use the “Copy Link” button above the graphics window. You could also download the aligned structural model files and/or the sequence alignment with the “Export” button.

Looking at the *Sequence Alignment* in 3D output, you can use the boxes next to the individual entries () to show or hide them in the 3D structure viewer window on the right side of the page, which shows the *structures* aligned with one another. The first box shows/hides the protein chain associated with the UniProt ID you entered. The second box shows/hides additional chains present (such as other proteins bound to that main

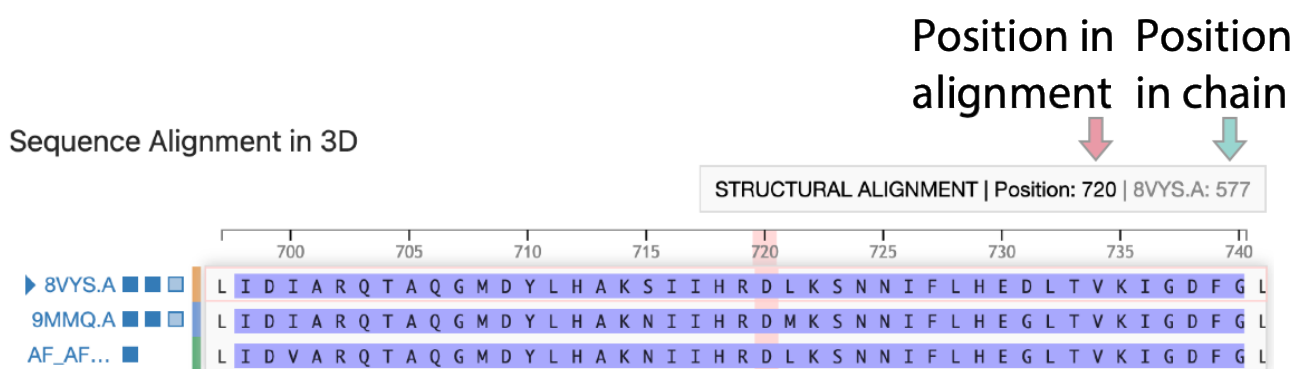
The bumbling biochemist (Bri Bibel), last updated 9/19/26

protein you searched for) – click to hide them for now. And the third box shows/hides any small molecule ligands present in the structure.

You can click on the information icons (i) in the sequence alignment get more quick information about the individual entries, including the technique used to determine/solve or predict that structure.

You can investigate the *sequence* alignment on the left side of the page (you can zoom in for a better view).

Note: If looking for a specific amino acid location in a sequence, be sure to look to the number in the chain (far right), not the position in the alignment. Also note that the numbering in the alignment might be slightly different from the numbering in UniProt. It's helpful to zoom in and compare the sequence to that found in UniProt to be sure you're looking at the right place.



Clicking on a specific site in the sequence alignment will zoom into that location in the 3D structure viewer portion of the page. You can take a snapshot of the aligned structures of the active site by clicking the lens icon (🔍) in the upper right of the 3D structure viewer.

RCSB PDB Pairwise Structure Alignment: <https://www.rcsb.org/alignment>

RCSB Help Guide: <https://www.rcsb.org/docs/tools/pairwise-structure-alignment>

Accompanying paper: Bittrich, S.; Segura, J.; Duarte, J. M.; Burley, S. K.; Rose, Y. RCSB Protein Data Bank: Exploring Protein 3D Similarities via Comprehensive Structural Alignments. *Bioinformatics* **2024**, *40* (6), btae370. <https://doi.org/10.1093/bioinformatics/btae370>.

More here: <https://youtu.be/MOBMBuAlAKo>

Additional resources

You can find and download this guide here: <https://bit.ly/uniprotprotparam>

More on the PDB: <http://bit.ly/thepdb> & <https://bit.ly/pdbstructures> & <https://youtu.be/kIB4T0X-sVU>

PDB YouTube playlist: <https://youtube.com/playlist?list=PLU1mvOopp1Ik&si=LLBBOZkQfvOX-YIe>

More on Uniprot: <https://bit.ly/uniprotprotparam>

More on other tools for working with proteins: <https://thebumblingbiochemist.com/365-days-of-science/proteintools/>

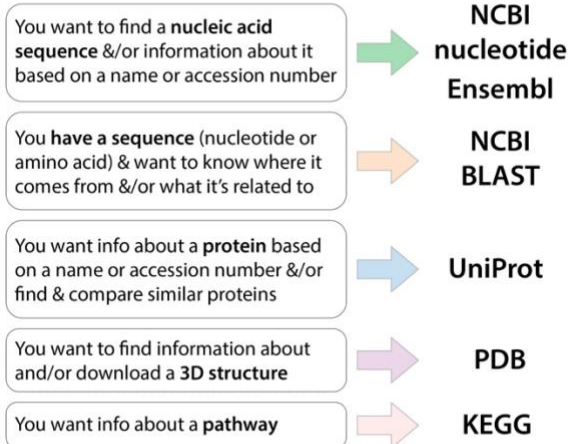
Structural biology resources guide: https://bit.ly/structural_biology

My structural biology YouTube playlist: <https://youtube.com/playlist?list=PLUWsCDtjESrGhwVxsRbTJdL-BEsN60RCs>

More on crystallography: <http://bit.ly/xraycrystallography2>

My x-ray crystallography YouTube playlist:
https://www.youtube.com/playlist?list=PLUWsCDtjESrHIGzgUctYRWSWm-lw9_VhX

Where to go to find biomolecular info if . . .



Each of these can serve as a first stop that will link you to directly to entries in other, more specialized, databases.

Annotations provide context & allow you to make sense of sequence-based data

They tell you where genes are, what parts of proteins do what, etc.

As scientists collect data, it gets added to **central databases** so you can see all that info mapped onto the sequence.

NCBI Nucleotide or Ensembl, UniProt, PDB, etc. genes, mRNA transcripts etc. proteins structures

Annotations can be manually curated (added by hand by scientists - slower but more trustworthy) &/or through automated processes that look for telltale, conserved signs of specific types of features.

More well-studied sequences will have more annotations than less well-studied sequences, which is why some database entries are more detailed than others.

UniProt provides an **annotation score** to help you quickly see how much info is known.



Protein? First stop, UniProt!

Screenshot of the UniProt website showing search and protein entry details. The search bar is at the top, and the protein entry for P40926-MDHM_HUMAN is shown below. The entry page includes a summary, function, and a table of domain annotations.

Search for a **protein** based on a name or accession number. You can also access BLAST & align directly here.

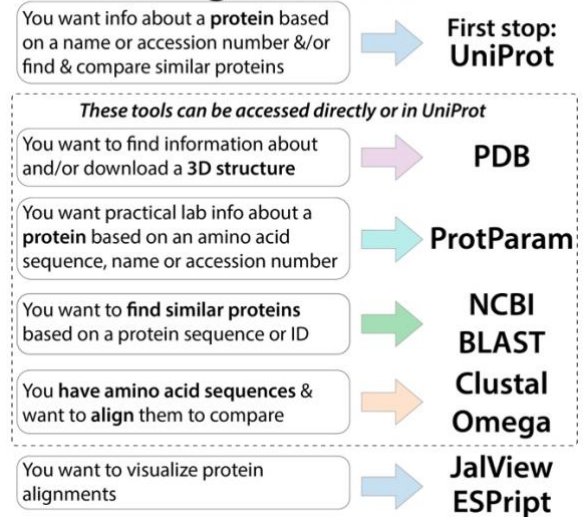
Add proteins of interest to your basket, then align them and/or run BLAST to find similar ones.

Find your protein

On each **entry page**, you'll find lots of information about the **protein**: downloadable sequences, including of isoforms, molecular weight(s), key residues, etc. as well as **direct links to other databases** with more information.

Above each isoform on the entry page is a **Tools** link that can take you to ProtParam, which has more great info (pI, extinction coefficient, etc.).

Where to go: Protein edition



Each UniProt accession code & entry is associated with an **individual protein chain** (or multiple versions of chains made from a single gene, such as through alternative splicing).

Each PDB accession code & entry is associated with a **single 3D structure**, which might contain multiple protein chains, parts of chains &/or additional molecules.

UniProt entries display and direct you to **ALL PDB entries** (i.e., **all 3D structures**) containing part or all of that UniProt entry's chain (in reverse alphabetical order).

The structures displayed in UniProt may contain more or less than you want and it's up to you to check them out individually to see (or save yourself some time by starting in the PDB).

Some structures might only contain small structured regions of proteins (such as individual domains) & some structures might contain the protein you're interested in bound to another molecule as part of a complex.

Look to see which parts of the chain are physically included in the PDB entry (but note that some of these might not be included in the structural model due to insufficient data) & go to the PDB for more details.



View the AlphaFold model for the predicted full-length structure.

UniProt Basket and Toolbox

The basket helps you quickly re-find and work with sequences/entries you're interested in and the toolbox helps you quickly view, modify, and/or rerun alignments, BLASTs, and ID mappings.

Screenshot of the UniProt Basket and Toolbox interfaces. The Basket shows a list of proteins, and the Toolbox shows the results of a BLAST search.

Each entry has these options: Tools, Download, Add

Search for a **protein** based on a name or accession number. You can also access BLAST & align directly here.

Add sequences of interest to your **basket**, then align them &/or run BLAST to find similar ones &/or Map IDs to find their accession codes & entries in other databases.

View results from tools (BLAST, Clustal sequence alignments, & ID mapping) in the **toolbox**.

Tool results are saved for 7 days, with input parameters stored longer for quick modifications &/or reruns.

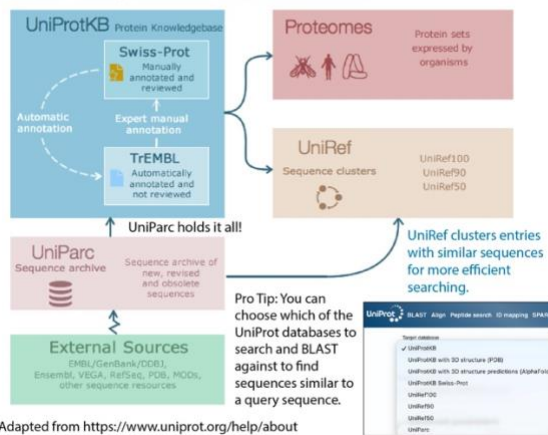
Click "Completed" to view results. Click this to modify/rerun.

Entries within the basket are split up by database.

UniProtKB: annotated
UniRef: reference entries for groups of entries clustered by % sequence identity
UniParc: all (including deprecated)

UniProt curates lots of protein sequence data from many sources.

The average user mostly just sees and interacts with UniProtKB, which has annotated entries.

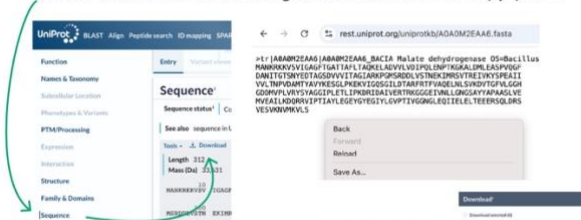


Adapted from <https://www.uniprot.org/help/about>

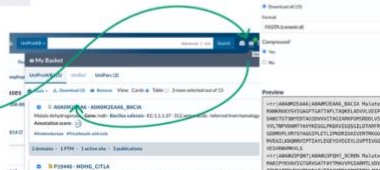
Downloading protein sequences from UniProt

Single sequence from an entry page ...

Sequence -> "Download" takes you to a screen with the FASTA-formatted text. Just right click and save. Or copy-paste.

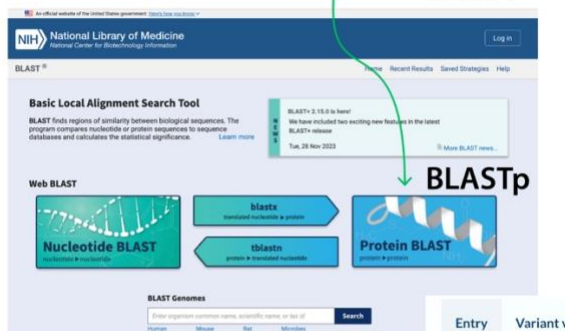


Multiple sequences from your basket ...



You want to find similar proteins based on a protein sequence or ID

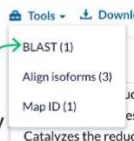
NCBI BLAST



It's also accessible via UniProt



From the top bar or the "Tools" menu for an entry



UniRef UniProt Reference Clusters

Helps you search protein sequence data more efficiently. Protein sequence data is clustered by % sequence identity, then represented by a reference entry.

Pro Tip: You can BLAST against these databases to quickly find a wider variety of sequences similar to a query sequence.

UniRef100



Identity percentage	Number of UniRef clusters	For example, the "same" protein in very closely-related species.
100%	220,919,788 clusters	

UniRef90



Identity percentage	Number of UniRef clusters	For example, the "same" protein in similar species.
90%	121,389,642 clusters	

UniRef50



Identity percentage	Number of UniRef clusters	For example, the "same" protein in distantly-related species.
50%	38,794,121 clusters	

Clusters are giving accession codes containing "UniRef100_", "UniRef90_", or "UniRef100_", followed by the code of the reference entry.

UniProt FASTA headers

All FASTA header lines start with a carat (^) (no matter what type of sequence)

Then there's the accession numbers/codes for database(s)

And more information

>db[UniRefIdentifier]EntryName ProteinName OS=OrganismName
OX=OrganismIdentifier [GN=GeneName](optional) PE=ProteinExistence
SV=SequenceVersion

db: UniProt database—UniProtKB (UniProt Knowledgebase) consists of 2 databases:

sp: Swiss-Prot - manually curated (more reliable, more info)

tr: TrEMBL - auto curated (less known often because they come from less well studied organisms, be more cautious)

PE = Protein Existence (# from 1 (most) to 5 (least) evidence it exists)

1. Experimental evidence at protein level
2. Experimental evidence at transcript level
3. Protein inferred from homology
4. Protein predicted
5. Protein uncertain

>trP12345MDH_BACSU Malate dehydrogenase OS=Escherichia coli (strain K12) GN=mdh PE=1 SV=1
MANKRKKVSVIGAGTGTATFLAQKELADVLDVDPQLENPTKGGKALDLEASPVQGF
DANITGTSNYEDTAGSDVVITAGIARPKMSRDDLVSTNEKIMRSVTEIVKYSPEAR
VLTNPVDAMTYAYKESGPKKEVIGQSGVLTARFTFAVELNLVSKVDYTFVFLGG
GDDMPVLYRYAGGIPLETLPKIDIVERTBKGGGEIVLLNGSAYAPASLITE
MVEALKQDRVPTIAYLEGEYGYEGYLVPTVVGNGLEQIELTETTERSQLRS
VESYKVMKVL

trEMBL database & inferred from homology

Swiss-Prot database and protein-level evidence of existence

You have amino acid sequences & want to align them to compare

Clustal Omega

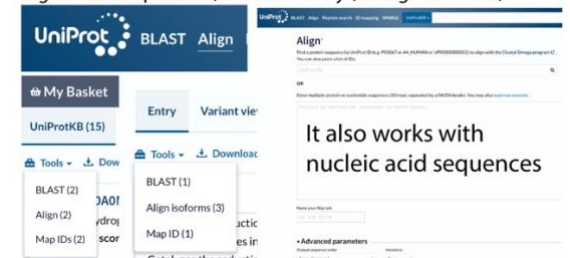


Input sequence ID

Sequence Type

It's also accessible via UniProt

From the top bar or from the "Tools" menu from the basket (to align different proteins) or for an entry (to align isoforms)



It also works with nucleic acid sequences

A grossly over-simplified comparison of Single particle cryo-EM X-ray crystallography

It's kinda like this, but a lot more complicated, with more lenses.

Electron waves come in...

Particles randomly oriented

... interact with sample

Get focused & magnified by electromagnetic lenses

Hit a camera detector

The detector records which pixels get hit & how many times & sends that data to a computer.

X-ray waves come in...

Particles are orderly arranged into a repeating 3D lattice

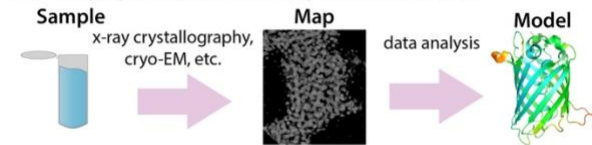
... interact with sample

You can't focus x-rays with lenses so you have to work with the scattered x-rays.

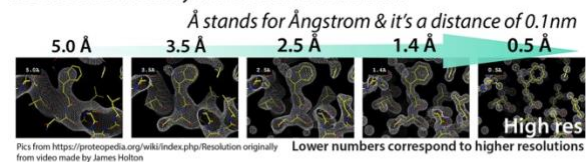
They add together with interference

to give a diffraction pattern that's recorded on a detector

Atomic models of proteins and other molecules are built into maps produced from experimental data



The better the resolution of the maps (more crisp, less saugagey), the more confidently the models can be built.

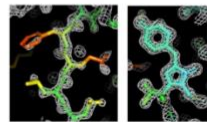


But beware, signal quality isn't the same everywhere in the map!

The figures to the right are from the same structure (3b3p)

Regions might even have such weak data they can't even be modeled

Such "unmodeled regions" are often shown as dashed lines in the structure and indicated by grayed out portions of the sequence.



The PDB is running out of 4-character IDs

So, PDB IDs (accession codes) are growing to 12 characters

4 characters → 12 characters starting with pdb_

How to adapt old codes: Add 4 zeros to the beginning: **pdb_0000**

pdb_00000000 + old PDB ID = extended PDB ID

For example, **9cn3** becomes **pdb_00009cn3**

The legacy PDB file format is incompatible with the longer codes

Need to start switching to **PDBx/mmCIF** file format

These work basically the same as PDB files from the user standpoint (compatible with PyMol, ChimeraX, etc.)

PDBx/mmCIF: Protein Data Bank Exchange macromolecular Crystallographic Information Framework; files end in .cif (not .pdb)

The extended PDBs already work - so start switching!



Components of crystal structures

crystal lattice

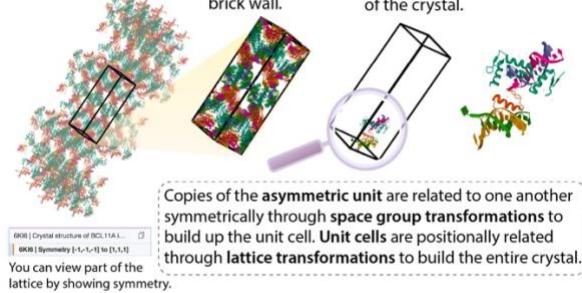
The whole crystal - it contains lots and lots of copies of the thing you care about.

unit cell

The basic repeating unit of the crystal, like the bricks in a brick wall.

Each unit cell has one or more full or partial copies of the **asymmetric unit**

The smallest "unique" part of the crystal.



biological assembly

The biologically relevant form - might be different than the asymmetric unit.

This example: "Crystal structure of BCL11A in complex with gamma-globin - 115 HPFH region"

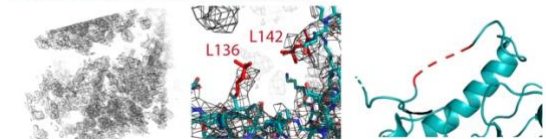
PDB ID: 6ki6

The brakes on the fetal hemoglobin gene.



Beware parts of the model may not be there!

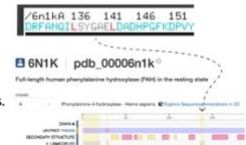
Scientists build structural models into experimental maps. Sometimes the experimental data is not strong enough in places, so the model has gaps!



Unmodeled regions were physically there when the data were collected (i.e. that region of the protein or other macromolecule was not missing), but the atoms in the region haven't left sufficient evidence for scientists to know exactly where they are. Often, these regions (frequently at termini & internal loops) are highly flexible so their signals kinda blur each other out.

If a letter is grayed out in the sequence and you can't select it (or see it!), this is because it is in an unmodeled region.

You can also see this information in the PDB entry's "Structure Summary" & "Sequence" Tabs. Look for an "Unmodeled" track. Gray regions indicate that residues there are not modeled.



Looking for something (structure-wise)?

PDB (Protein Data Bank) Advanced Search Query Builder

Some features you'll find in the PDB Sequence Viewer

It has sequence-level information about both molecular features (functional sites, etc.) & the quality of the structural model. There's an abbreviated version of this on the main entry page and an expanded version under the "sequence" tab.

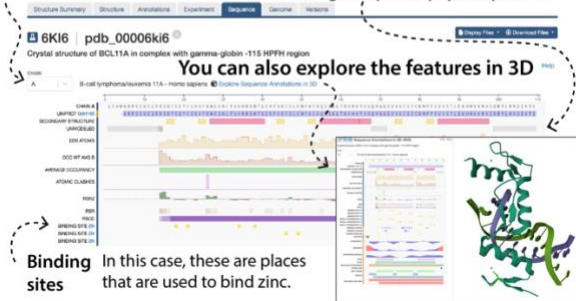


Each row has different, structure-dependent info. For example...

Unmodeled regions

These gray bars are places where the data wasn't strong enough to model in the atoms, even though they were physically there.

Be sure to select the chain you're interested in.



You can also explore the features in 3D

Binding sites In this case, these are places that are used to bind zinc.

PDB Pairwise Structure Alignment

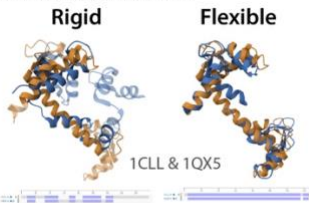
Aligns one or more structures (experimentally-determined or predicted) & their sequences to a reference structure & sequence based on their shapes.

You can give it structures to align from (among other options) ...

You can choose from a variety of alignment methods based on how similar the proteins &/or their structures are.

Choose a rigid method to directly superimpose the structures without altering them (good if very similar) or a flexible method (which breaks the structure up and aligns parts) for ones with large structural differences.

You can visualize in 3D and export the aligned structures and their aligned sequences.



Choosing PDB structures

Make sure the model includes the region you're interested in. And, if there's a region you really care about, carefully inspect that region of the model (look in 3D at the underlying experimental maps and the fit of the model into the map).

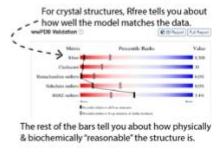
Check to see if there are introduced mutations you do or don't want.

Check to see if there are other molecules included in the structure you do or don't want.

All else being equal, a structure with a **higher resolution** (corresponding to lower numbers) will typically be more reliable, because the underlying data used to build the model is "crisper."

Look to the **validation** slider bars to see how well the model matches the underlying data and how physically and geometrically plausible the model is. The sliders compare this structure to other structures in the PDB.

Key info's conveniently located on the Structure Summary page



The PDB allows you to **group structures** by % sequence similarity or Uniprot accession (if you want grouped by exact same protein)



You can then further filter the results

And view sequence and structure alignments

Click on a bar to filter based on a property (ligand, etc.), shift+click to see the individual structures in the group with that property

This helps you quickly explore structures within & between groups of closely or more distantly related proteins. Or find the "best" structures of a specific protein for your needs.

You can get to it from advanced search or from an individual entry

Sometimes, molecular structure models don't look like what you expect ...

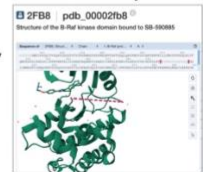
Seeing more than you expected?

- The molecule might be part of a complex
- You could be seeing a sort of artifact of the experimental method (e.g., the asymmetric unit in a crystal structure (smallest unique part of the crystal) might be different from the functional form (biological unit/assembly)).



Seeing less than you expected?

- The scientists might have worked with a truncated (shortened) version of a protein, such as an isolated domain or even just a small peptide from it.
- Parts of the structure, although physically present, might be too "floppy" to collect sufficient data to model in atomic positions. These are left as "unmodeled regions" & appear as grayed out letters in the sequence & dashed lines in the structural model.



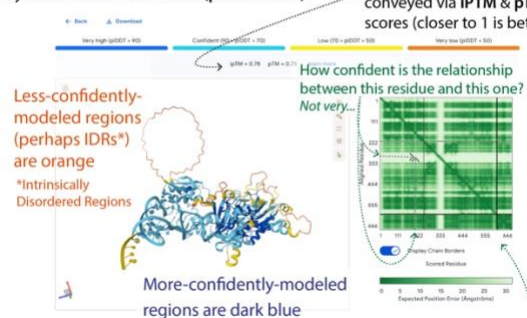
Note: For predicted structures (e.g., AlphaFold structures), you might see less than expected if it's a multi-chain protein & the predicted structure is only for one of the chains.

AlphaFold predicts protein structures based on their amino acid sequences

Using AI trained on the experimentally-determined structures of lots of proteins, physics principles, etc.

It returns a set of "best guesses" & measures of confidence for each. The top-scoring structure is displayed, colored by **residue-level confidence** (pLDDT score).

Overall confidence is conveyed via iPTM & pTM scores (closer to 1 is better)



Predicted aligned error (PAE) is plotted, giving a residue-level pairwise indicator of whether "packing" & relative position of domains is accurate (dark green's better).