Software tools and databases for working with proteins

Bri Bibel (the bumbling biochemist); updated 8/4/26
Blog version: https://thebumblingbiochemist.com/365-days-of-science/proteintools/

- **UniProt:** Basics (functions, length, sequence(s), isoforms, structures, interactions, etc. - lots of links to more detailed databases for each) https://www.uniprot.org/
    - Consists of: UniProtKB (annotated entries, with reviewed entries in UniProtKB/Swiss-Prot and unreviewed entries in UniProtKB/TrEMBL); UniRef (data is clustered by % sequence identity (100, 90, or 50), then each cluster is represented by a reference entry for efficient searching, etc.); and UniParc (archives for all non-redundant protein sequence data)
    - More about it here: https://bit.ly/uniprotprotparam & YouTube: https://youtu.be/f75f6QCe1gA
- **PDB (Protein Data Bank)**: Protein (and other macromolecule) structures https://www.rcsb.org/ (US-maintained site), https://www.ebi.ac.uk/pdbe/ (European-maintained site)
    - More about it here: http://bit.ly/thepdb & https://bit.ly/pdbstructures (emphasis on x-ray crystallography entries) & YouTube: https://youtu.be/kIB4T0X-sVU (overview) & https://youtube.com/playlist?list=PLU1mvOopp1Ik&si=LLBBOZkQfvOX-YIe (playlist)
    - Features include **3D sequence annotation viewer**, **PDB validation reports**, **advanced search query builder** (including searching by structure or structural motif), **structure grouping**, and **structural alignment**.
- **EMDB (Electron Microscopy Data Bank):** Electron microscopy structure data, validation information, 3D viewing: https://www.ebi.ac.uk/emdb/
    - More: http://bit.ly/cryoemxray & https://youtu.be/G57aBLsBgOk
- **AlphaFold Protein Structure Database (AFDB):** Contains pre-computed predicted structures of single chains of proteins https://alphafold.ebi.ac.uk/
- **AlphaFold:** Predicts the structure of proteins, nucleic acids, and complexes. The server has more limited options than the stand-alone (GitHub) and ColabFold (Python notebook-based) versions, but uses the latest version of AlphaFold https://alphafoldserver.com/
    - More here: https://thebumblingbiochemist.com/365-days-of-science/interpreting-alphafold-predictions/
- **Viewing protein structures (molecular visualization software):** There are 2 main tools, ChimeraX & PyMOL, as well as other, web-based, tools such as Mol* and iCN3D. ChimeraX is fully free for academic use. PyMOL has a free version with slightly limited functionality and academic labs often have a license for the full thing.
    - ChimeraX home page: https://www.rbvi.ucsf.edu/chimerax/ & tutorials: https://www.rbvi.ucsf.edu/chimerax/tutorials.html
    - PyMOL: https://www.PyMOL.org/
        - More about it here: https://bit.ly/PyMOLintro & YouTube: https://www.youtube.com/watch?v=CeTTmm5DI_k (overview/fundamentals) & https://www.youtube.com/playlist?list=PLUWsCDtjESrG-46VDvPZfYPQyyXhZbGCN (playlist)

    - Less powerful, but convenient, browser-based alternatives include iCN3D and Mol*
        - **iCN3D** https://www.ncbi.nlm.nih.gov/Structure/icn3d/full.html
        - **Mol*** https://molstar.org/
- **Expasy ProtParam**: Protein stats (amino acid composition (e.g. # of each amino acid, isoelectric point (pI) (for predicting overall charge at different pHs), extinction coefficient (for getting concentration from UV), etc.) can also use for custom sequences you copy and paste: great for recombinant expression constructs https://web.expasy.org/protparam/
    - More here: https://bit.ly/uniprotprotparam
- **PDBSum**: Easy way to automatically generate protein topology & domain diagrams & investigate interfaces https://www.ebi.ac.uk/thornton-srv/databases/pdbsum/
    - Just upload or fetch a PDB file and it'll generate a bunch of things including: domain architecture diagram, secondary structure diagrams, topology map, diagram of and information about ligand-protein and protein-protein interactions (distances, types, etc.), information about pockets, tunnels, and channels.
    - *Note: Their server is down as of this writing, but there's a stand-alone version you can download from GitHub.*
    - More here: https://youtu.be/AFmN4CKU1uE
- **HawkDock:** Models protein-protein interactions and analyzes strength of interactions at a residue level basis (for either predicted structures or already-prepared PDB files) https://cadd.zju.edu.cn/hawkdock/
- **3D bio notes:** You put in the PDB code for a structure, then this site integrates a ton of information about the protein from various sources and lets you see sites of mutations, etc. in 3D. Also good for a quick overview or on the go look when you don't want to open PyMOL. http://3dbionotes.cnb.csic.es/ws
- **H++:** Free web-based tool that takes a pdb file of a protein and predicts the protonation state of each amino acid residue at a pH of your choosing http://newbiophysics.cs.vt.edu/H++/index.php
    - More here: https://bit.ly/hplusplus
- **ExplorEnz:** Official IUBMB info about enzymes and their classification https://www.enzyme-database.org/
    - More here: https://bit.ly/ec_numbers & YouTube: https://youtu.be/8yO1XEzoVIE
- **BRENDA:** Additional information about enzymes w/links to other sites https://www.brenda-enzymes.org/
- **M-CSA** (Mechanism and Catalytic Site Atlas): Info about enzyme mechanisms and key catalytic residues https://www.ebi.ac.uk/thornton-srv/m-csa/
- **InterPro**: Information about protein families and conserved domains https://www.ebi.ac.uk/interpro/
- **Human Protein Atlas**: Mass spec, antibody evidence, mRNA-seq, etc. showing where different proteins are expressed in humans, and even within cells https://www.proteinatlas.org/
- **EMBL-EBI Expression Atlas**: Protein expression in multiple (but a limited number of) species https://www.ebi.ac.uk/gxa/home

- **AlUpred:** For predicting intrinsically disordered regions of proteins https://aiupred.elte.hu/
    - More here: https://thebumblingbiochemist.com/365-days-of-science/idrs/
- **Expasy Swiss Bioinformatics Resource Portal:** Tons of bioinformatics tools, many of which are for proteins https://www.expasy.org/
- **MPI Bioinformatics Toolkit:** this has lots of tools for structural prediction, homology finding, etc. https://toolkit.tuebingen.mpg.de/
- If you want to **find proteins that are similar**:
    - *In sequence:*
        - **BLAST (specifically blastp)**: This searches protein databases for 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 directly or through UniProt ("Tools" → "BLAST") https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome
    - *In structure:*
        - **Foldseek**: It compares (single-chain) structures against one another to find similar ones. What's even cooler is that it searches both experimentally determined structures (housed in the PDB) and AI-predicted (AlphaFold Protein Structure DataBase (AFDB)) structures.
            - You can access the web version of it here: https://search.foldseek.com/search
            - Or the open-access code to run it in command line is here: https://github.com/steineggerlab/foldseek
            - It's also integrated into the AlphaFold database so you can get results there too https://alphafold.ebi.ac.uk/
        - **Foldseek multimer:** Goes beyond a single chain – it searches for complexes with similar structures https://search.foldseek.com/multimer
        - **PDB Advanced Search Query Builder**: https://www.rcsb.org/search/advanced
            - More here: https://youtu.be/iSM6orTve-Y
        - **ENDscript** https://endscript.ibcp.fr/ESPript/ENDscript/index.php
            - This will find sequences to align and also get their PDB codes and structural information.
- If you're looking to create and/or visualize (interactively or for figures) **protein *sequence* alignments**:
    - For **creating sequence alignments:**
        - **Clustal Omega:** https://www.ebi.ac.uk/jdispatcher/msa/clustalo
            - You can also access it through the "Tools" → "Align" button in UniProt: https://www.uniprot.org/

- For **visualizing sequence alignments interactively**:
  - **JalView**JS (web version) or JalView desktop version: https://www.jalview.org/jalview-js/JalviewJS/
  - For quickly comparing sequences in a clustal alignment and finding what number in one sequence corresponds to what number in another sequence, etc. here's a simple Jupyter/GoogleColab notebook: https://github.com/bribibel/clustal_comparer/blob/ab4cd7391b9add2857e20798d2da60c2f5caacfd/clustal_alignment_comparer.ipynb
- For **making nice pretty sequence alignment figures**, including with overlying secondary structural information, etc.:
  - **ESPript**: https://espript.ibcp.fr/ESPript/ESPript/index.php
- More here: https://bit.ly/proteinalignments

- If you're looking to create and visualize **3D protein structural alignments**:
  - **RCSB PDB Pairwise Structure Alignment Tool**: https://www.rcsb.org/alignment
    - 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.
    - This will also allow you to download mmCIF files of them overlaid.
    - More here: https://youtu.be/MOBMBuAIAKo
  - **FoldMason MSTA:** Unlike a pairwise alignment tool, FoldMason MSTA does Multiple Structure Alignment, finding the best alignment for a group of sequences, rather than comparing each structure to the same reference structure. https://search.foldseek.com/foldmason

- Tools for structural biology: https://bit.ly/structural_biology