



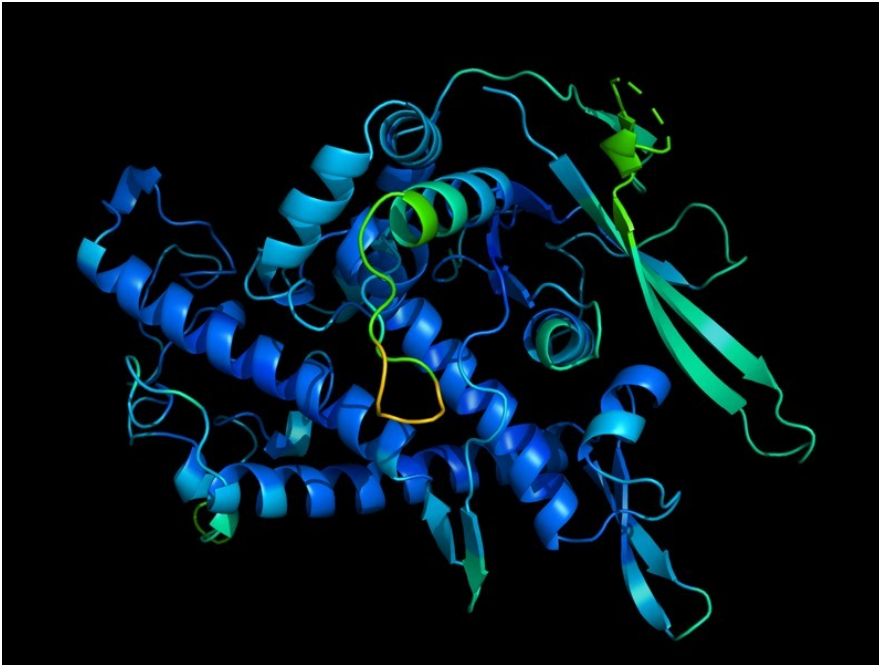
Intrinsically Disordered Proteins

Mariane Gonçalves-Kulik
Kristina Kastano
Miguel Andrade

Faculty of Biology,
Johannes Gutenberg University
Mainz, Germany
magoncal@uni-mainz.de

Protein structures

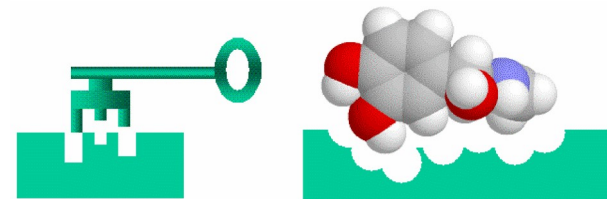
Representation of the 3D structure of a protein



Structure-function paradigm:

Structure → **function**

Emil Fisher's key-lock model (1894)



PDB was established in 1971 containing 7 protein structures.



New: More Computed Structure Models (CSM) available [Learn more](#)

Welcome

Deposit

Search

Visualize

Analyze

Download

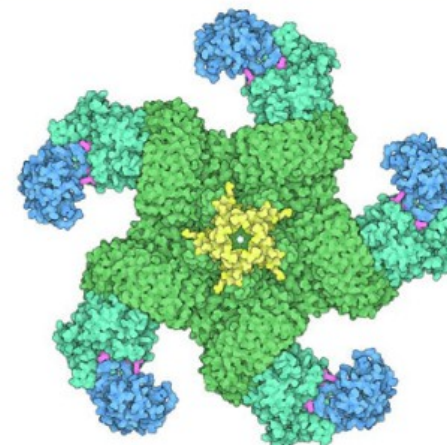
Learn

RCSB Protein Data Bank (RCSB PDB) enables breakthroughs in science and education by providing access and tools for exploration, visualization, and analysis of:

- Experimentally-determined 3D structures from the Protein Data Bank (PDB) archive
- Computed Structure Models (CSM) from AlphaFold DB and ModelArchive

These data can be explored in context of external annotations providing a structural view of biology.

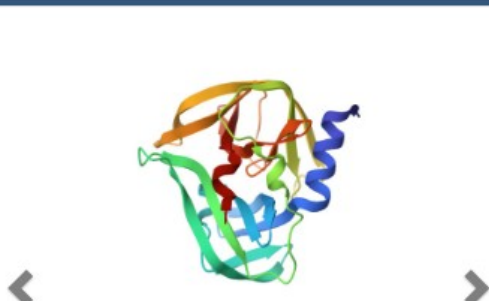
November Molecule of the Month



ZAR1 Resistosome

Latest Entries

As of Tue Nov 28 2023



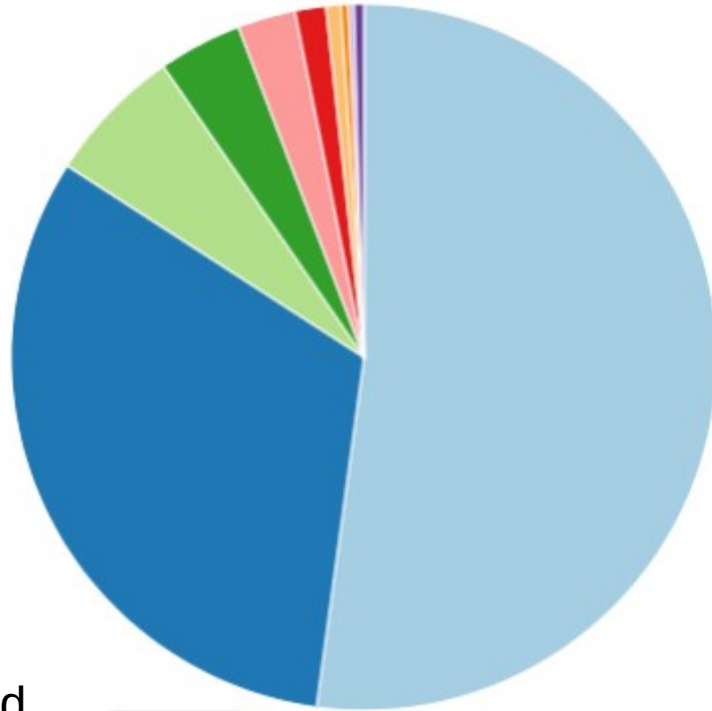
Features & Highlights

- Deprecation of FTP File Download Protocol in the PDB Archive
wwPDB plans to deprecate FTP download protocol on November 1st 2024
- Backbone Annotation and Standardization of Peptide Residues is Now Live
- Explore Antibiotic Resistance in 3D

News

Publications -

- Watch the Crash Course: RCSB PDB APIs
Learn about *Leveraging RCSB PDB APIs for Bioinformatics Analyses and Machine Learning.*
» 11/28/2023
- Papers Published in Special Issue of Journal of Molecular Biology
Read how the ModelCIF data framework and RCSB PDB APIs each support



Taxonomy counts

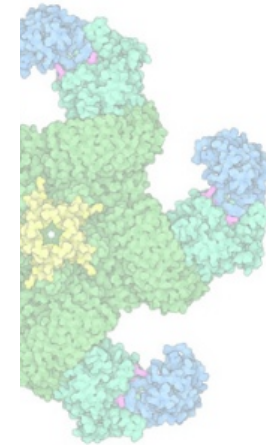
- Eukaryota (117,291)*
- Bacteria (71,899)
- Riboviria (14,096)
- other sequences (8,624)
- Archaea (5,962)
- Duplodnaviria (3,173)
- Eukaryota (eukaryotes) (1,521)
- Varidnaviria (739)
- Monodnaviria (618)
- Others (989)

* Entry counts (several experiments for the same proteins or complexes);



more

Structure of the Month



Resistosome

Expected structure

Latest Entries

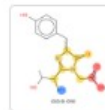
As of Tue Nov 28 2023



Features & Highlights



Deprecation of FTP File Download Protocol in the PDB Archive
wwPDB plans to deprecate FTP download protocol on November 1st 2024



Backbone Annotation and Standardization of Peptide Residues is Now Live



Explore Antibiotic Resistance in 3D

News



Watch the Crash Course: RCSB PDB APIs
Learn about *Leveraging RCSB PDB APIs for Bioinformatics Analyses and Machine Learning.*
» 11/28/2023

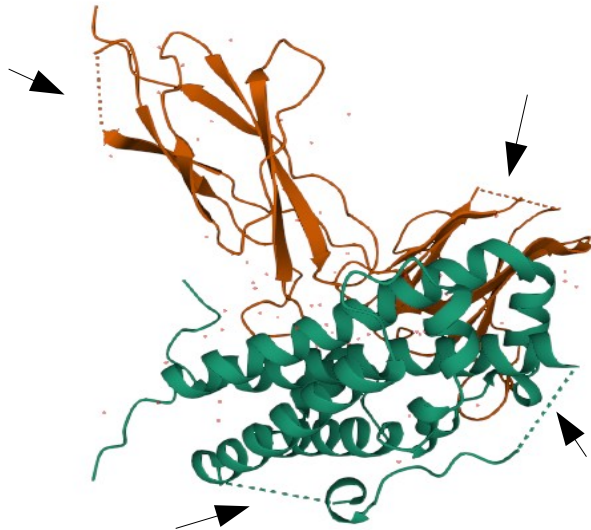


Papers Published in Special Issue of Journal of Molecular Biology
Read how the ModelCIF data framework and RCSB PDB APIs each support

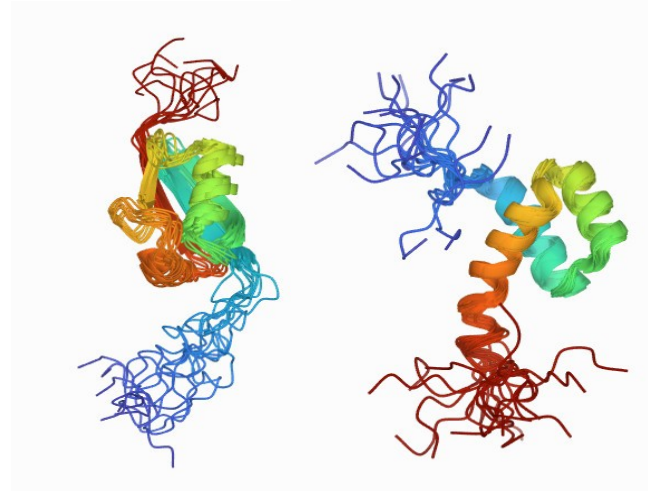
What is protein disorder?

Intrinsically disordered proteins (IDPs): proteins with regions that lack a single well-defined 3D structure in native conditions.

Missing electron densities in X-ray crystallography from PDB

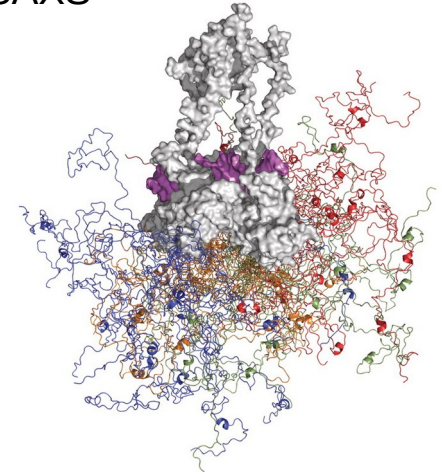


Human Growth hormone bound to receptor



NMR (nuclear magnetic resonance) ensembles of SUMO-1 and antennapedia from PDB

Model of tumor suppressor p53 using X-ray, NMR and SAXS



Disordered in 40% of its length!

~30% of PDB structures have such regions!

JMB



Intrinsically Unstructured Proteins: Re-assessing the Protein Structure-Function Paradigm

Peter E. Wright* and **H. Jane Dyson***

*Department of Molecular
Biology and Skaggs Institute of
Chemical Biology, The Scripps
Research Institute, 10550 North
Torrey Pines Road, La Jolla
CA 92037, USA*

A major challenge in the post-genome era will be determination of the functions of the encoded protein sequences. Since it is generally assumed that the function of a protein is closely linked to its three-dimensional structure, prediction or experimental determination of the library of protein structures is a matter of high priority. However, a large proportion of gene sequences appear to code not for folded, globular proteins, but for long stretches of amino acids that are likely to be either unfolded in solution or adopt non-globular structures of unknown conformation. Characterization of the conformational propensities and function of the non-globular protein sequences represents a major challenge. The high proportion of these sequences in the genomes of all organisms studied to date argues for important, as yet unknown functions, since there could be no other reason for their persistence throughout evolution. Clearly the assumption that a folded three-dimensional structure is necessary for function needs to be re-examined. Although the functions of many pro-

Disordered region functions

- Flexible linkers/spacers between domains
- Entropic chains (contribute to the structure energy)
- Molecular recognition:
 - binding to proteins, nucleic acid polymers, membrane, metal ions
 - As enzymes that undergo disorder-to-order transitions
 - Formation of multiprotein complexes
- Protein modifications and regulation (e.g. phosphorylation)

Why study IDRs?

- Their mutation is involved in diseases:
 - Cancer
 - Neurodegenerative diseases (Parkinson's, Dementia, Alzheimer's, Down's syndrome, SCA1)
 - Diabetes
 - Cardiovascular diseases
- Can be used in drug delivery (synthetic IDRs);
- Understanding of protein complexes interactions.

Revised lock-and-key model for IDPs



Disordered proteins can bind to many structured partners!

Disorder databases

Organism	Proteins	Regions
▶ Viruses	217	848
▶ Eukaryota	2,055	5,779
▶ Bacteria	356	905
▶ Archaea	21	54

- Disprot: 

database of experimentally verified IDPs

- IDEAL: database of experimentally verified 1110 IDPs

- MobiDB: 

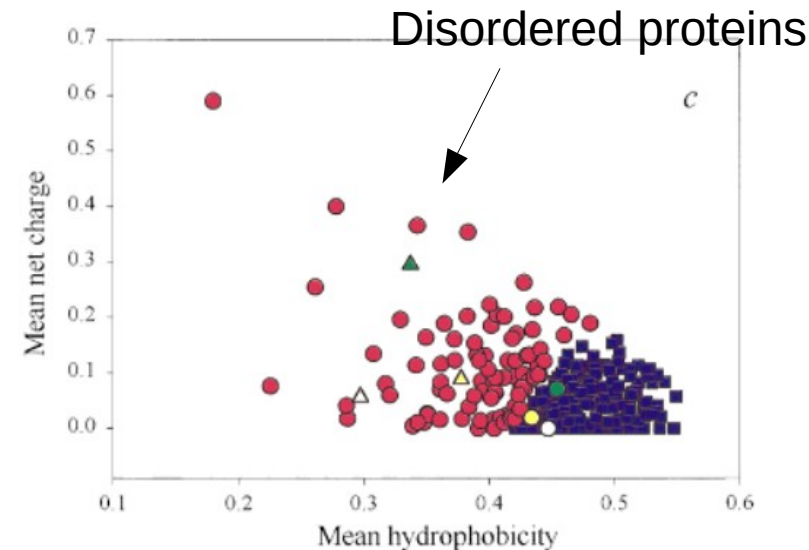
centralized resource that combines experimental and predicted data into a consensus annotation

- DIBS: Disordered Binding Sites (DIBS) with 1,576 complexes with curated interactions on the IDR region.

Disordered sequences

- Low content of bulky hydrophobic amino acids (Val, Leu, Ile, Phe, Trp, Tyr, Met)
- High content of charged and polar Ser, Pro, Glu, Lys)

Used in disorder predictor FoldIndex



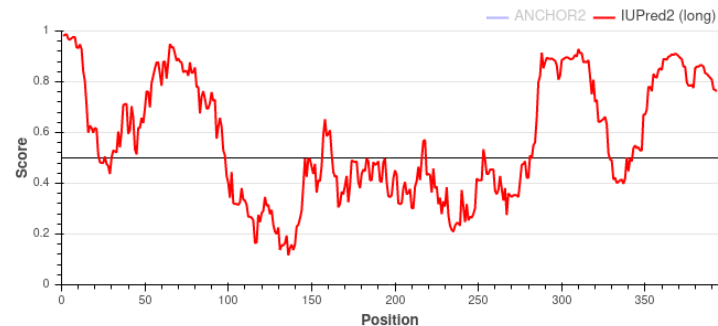
Uversky et al 2000

Disorder prediction

Prediction methods can be based on:

- Physical/chemical features (FoldIndex)
- Machine learning algorithms (DISOPRED2, Spritz, PONDR)
- Energy estimation (***IUPred2***):

Globular proteins form many favorable interactions to ensure the stability of the structure.



IUPred output for p53

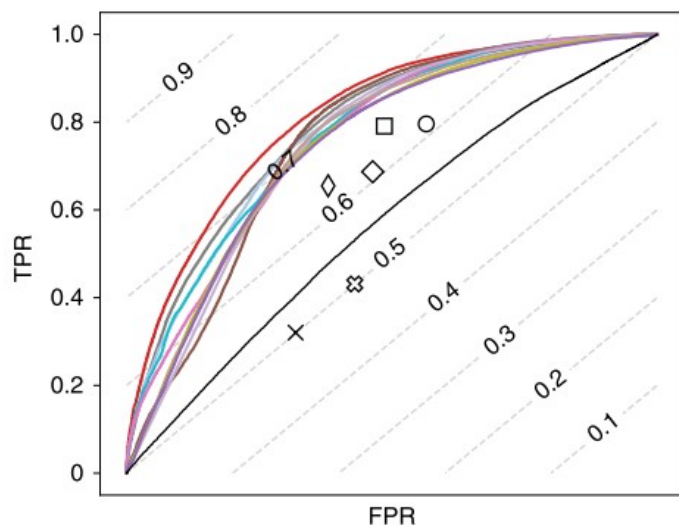
Disorder prediction

Critical Assessment of protein Intrinsic Disorder prediction (CAID)

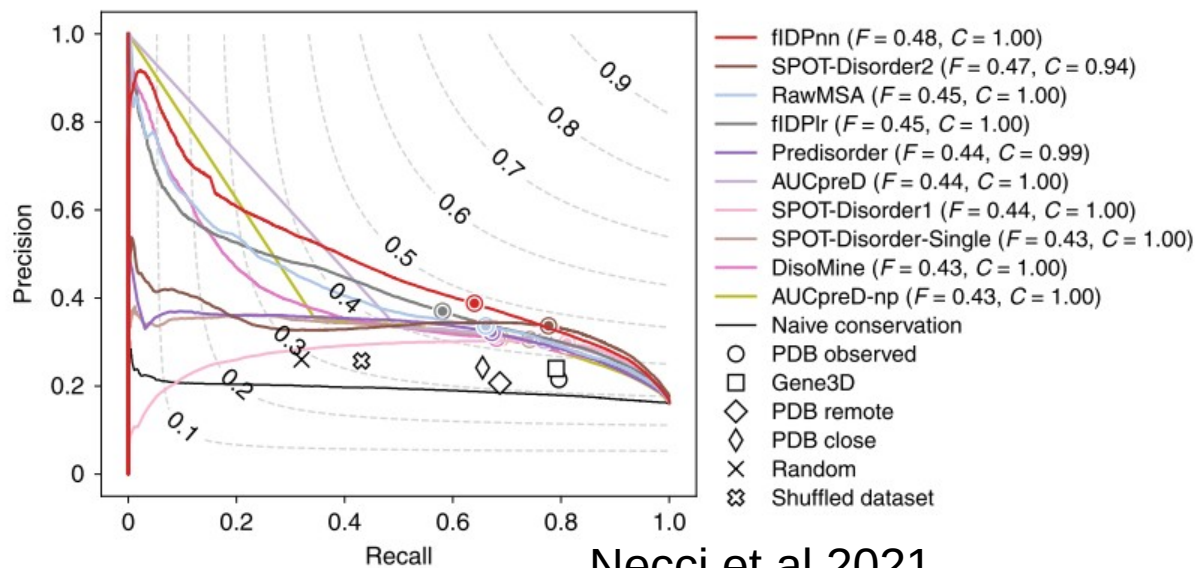
CAID

Biennial competition designed to assess the quality of new IDR/IDP and binding predictors.

AUC curve



Precision-Recall curve

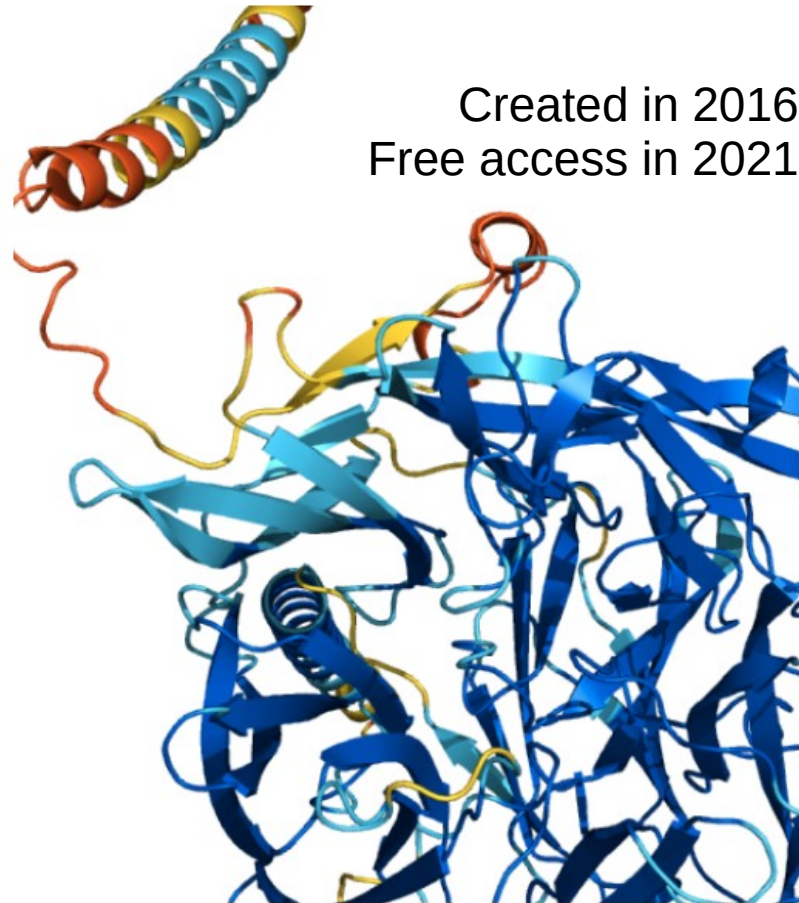


Necci et al 2021

What about order prediction?

AlphaFold is an AI system developed by **DeepMind** that predicts a protein's 3D structure from its amino acid sequence. It regularly achieves accuracy competitive with experiment.

DeepMind and EMBL's European Bioinformatics Institute (**EMBL-EBI**) have partnered to create AlphaFold DB to make these predictions freely available to the scientific community. The latest database release contains over 200 million entries, providing broad coverage of **UniProt** (the standard repository of protein sequences and annotations). We provide individual **downloads** for the human proteome and for the proteomes of 47 other key organisms important in research and global health. We also provide a download for the manually curated subset of UniProt (**Swiss-Prot**).



Created in 2016
Free access in 2021

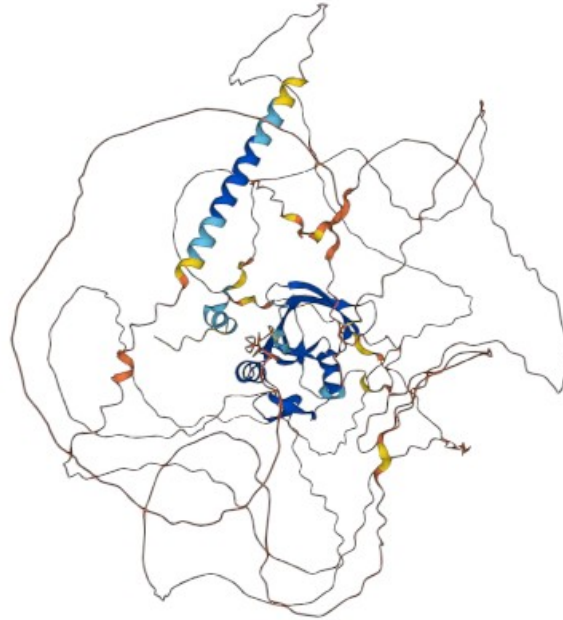
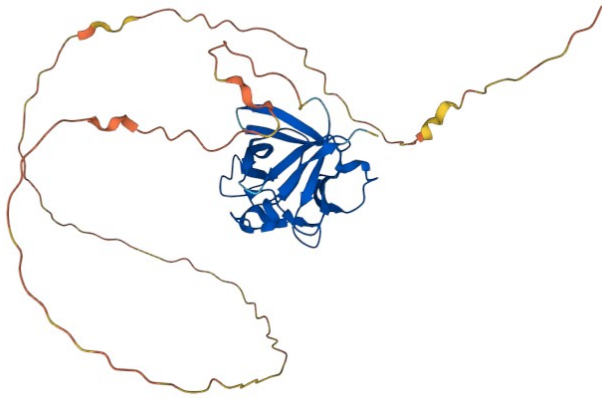
Q8I3H7: May protect the malaria parasite against attack by the immune system.
Mean pLDDT 85.57.

[View protein](#)

What about order prediction?

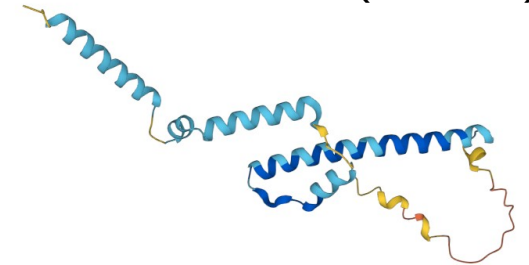
AlphaFold also “predicts” disorder

P09038 - Fibroblast growth Factor 2 (Human)



P54253 - Ataxin-1 (Human)

Q8WVH0 – Complexin-3 (Human)



Low pLDDT scores can be used as an indication of disorder.

Model Confidence ⓘ

- Very high (pLDDT > 90)
- High (90 > pLDDT > 70)
- Low (70 > pLDDT > 50)
- Very low (pLDDT < 50)

AlphaFold produces a per-residue model confidence score (pLDDT) between 0 and 100. Some regions below 50 pLDDT may be unstructured in isolation.

Research examples

Structure in disordered regions

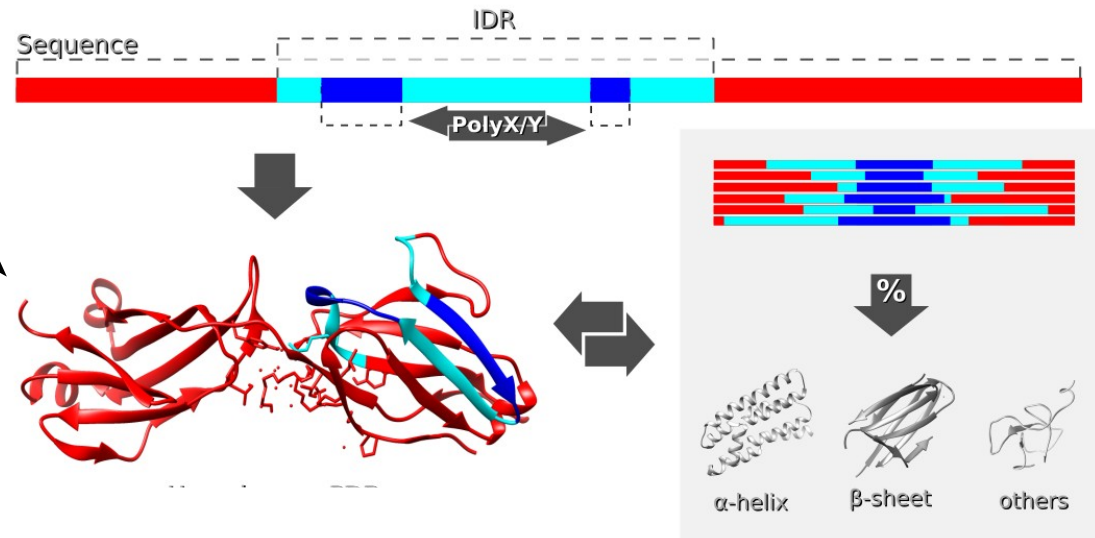
PolyXYs - QS, GS, RG...

- Filtered PolyX and PolyXYs within IDRs of the Human proteome;
- Extracted 100 residues surrounding the repeated region to analyse its structural content.

Goncalves-Kulik et al 2022:
Based on sequences of homologous PDB structures

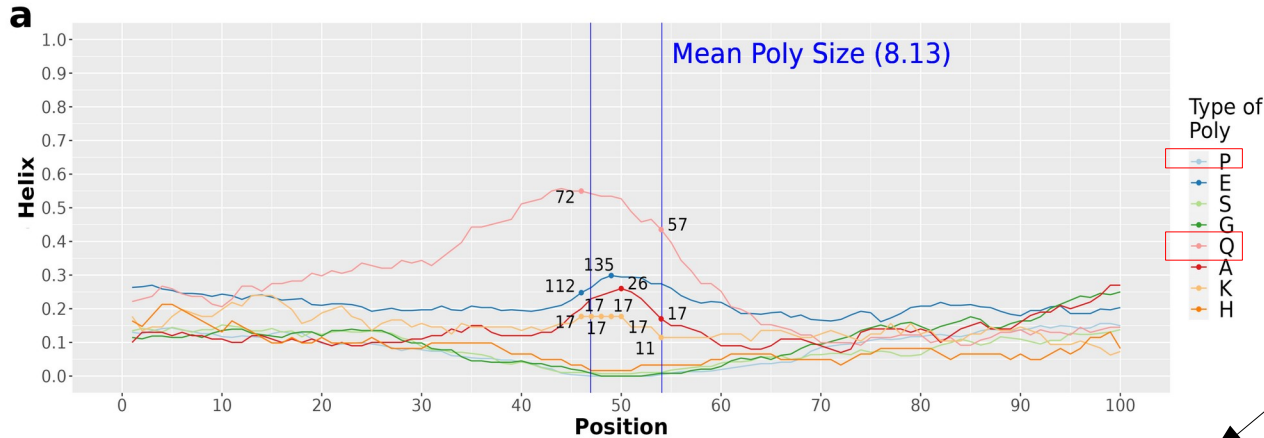
Goncalves-Kulik et al 2023:
Analyses based on AlphaFold predictions

```
>sp|P35637|FUS_HUMAN RNA-binding protein
MASNDYTQQATQSYGAYPTQPGQGYSQQSSQPYGQQSYSGYSQSTDTSGYGQSSYSSYGQSQNTGYGTQSTPQG
YGSTGGYSSQSSQSSYQQSSYPGYGQQPAPSSTSGSYGSSSQSSSYGQPQSGSYSQQPSYGGQQQSYGQQQS
YNPPQGYGQQNQYNSSSGGGGGGGGGGGNYGQDQSSSSGGGSGGGYGNQDQSGGGGSGGYGQQDRGGRGRGGSG
GGGGGGGGYNRSSGGYEPRGRGGGRGRRGGMGGSDRGGFNKFGGPRDQGSRHQSEQDQNSDNTIFVQQLGENV
TIESVADYFKQIGIIKTNKKTGQPMINLYTDRETGKLKGEATVSFDDPPSAKAAIDWFDGKEFSGNPIKVSFAT
RRADFNRRGGGNRGGRRGGPMGRGGYGGGGSGGGRRGGFPSSGGGGGGQQ RAGDWKCPNPTCENMNFSWRNEC
NQCKAPKPDGPGGGPGGSHMGGNYGDDRRGGRRGGYDRGGYRGRGGDRGGFRGGRRGGDRGGFGPGKMDSRGEHR
QDRRERPY
```

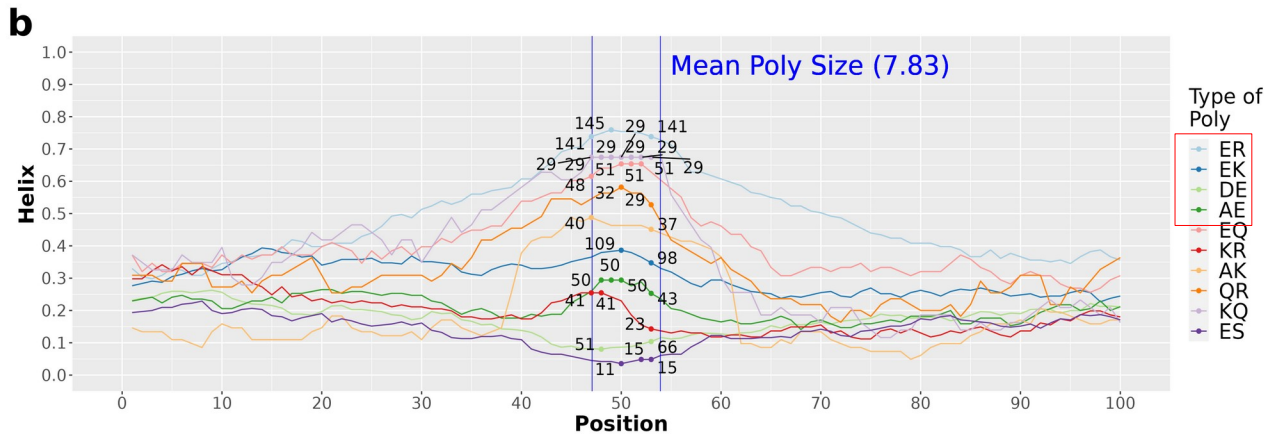


Research examples

Structure in disordered regions



Some polyX (Q, P) and polyXYs (ER, EK, DE, AE) show preference to adopt helical structures in the repeated regions.

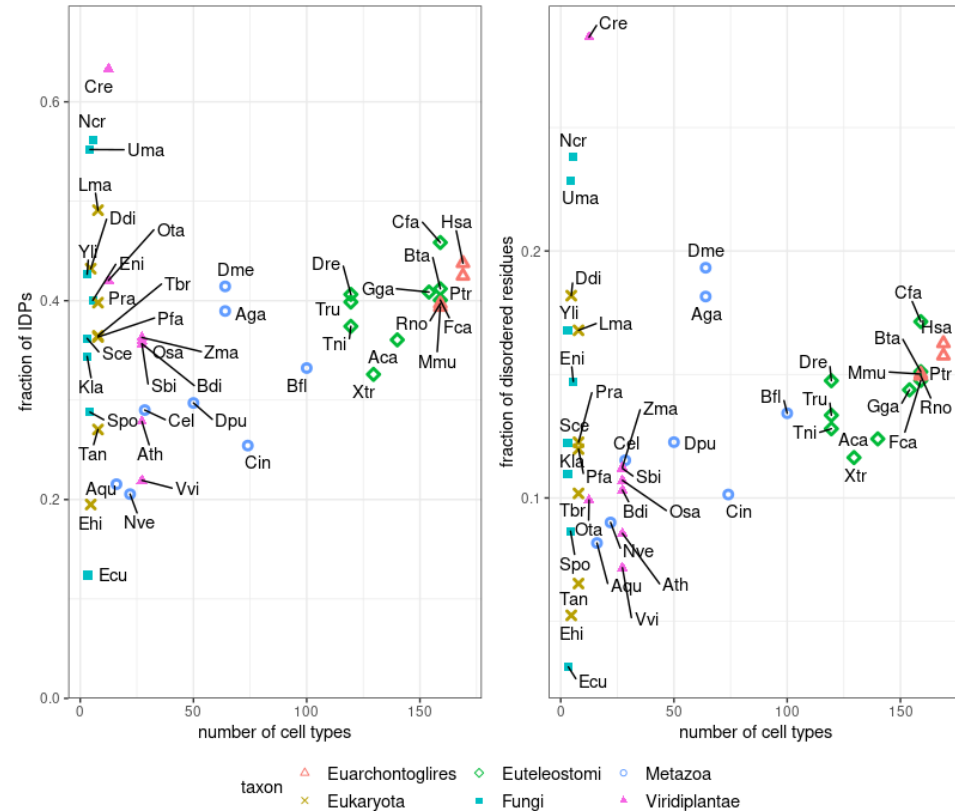


Research examples

Evolutionary study of disorder

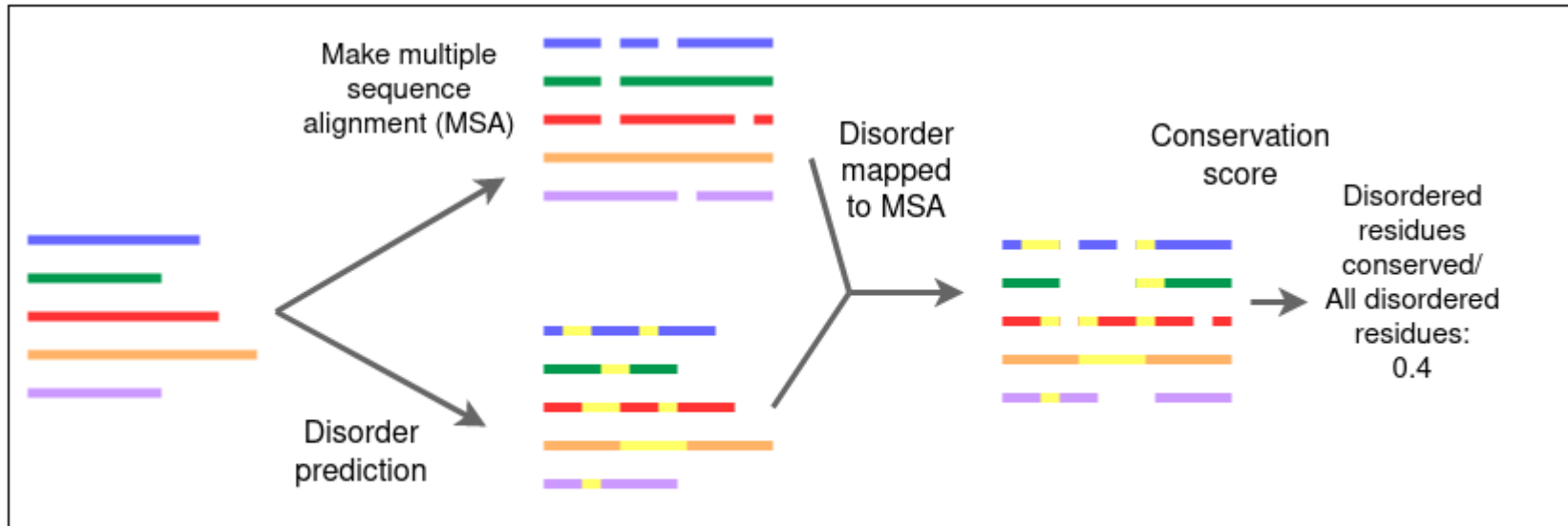
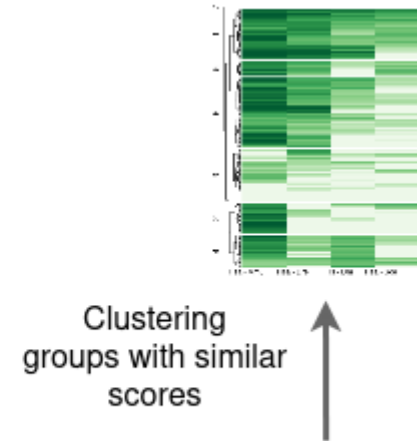
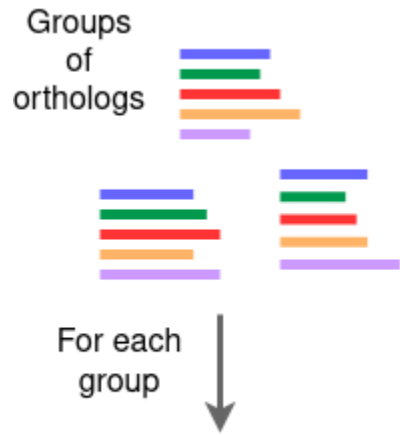
Natural abundance and phylogenetic distribution

- ~40% of human proteins predicted to be IDPs
- ~30% of Eukaryotic proteins predicted to be IDPs



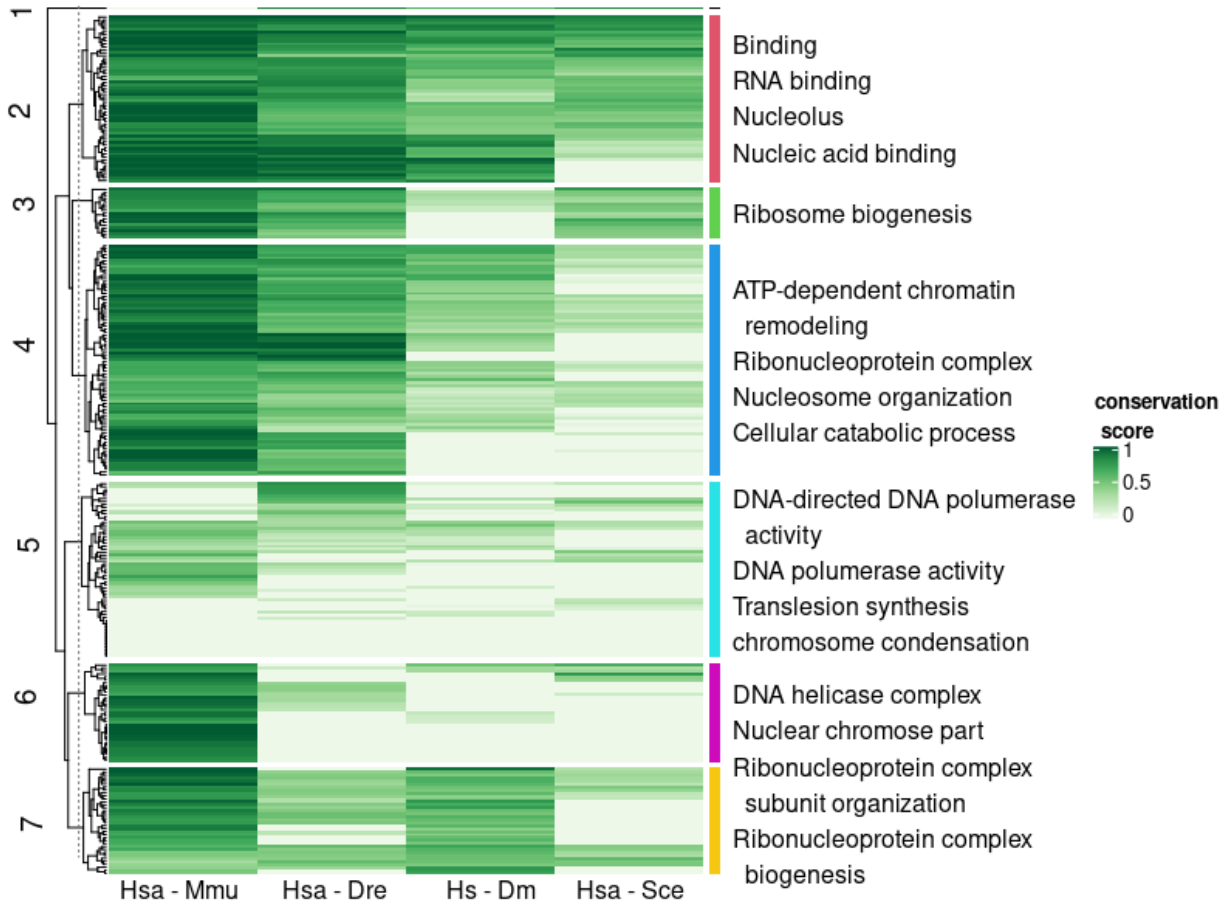
Research examples

Evolutionary study of disorder



Research examples

Evolutionary study of disorder



Correlating similar disorder conservation patterns with protein functions

Exercises

Exercise time!



Exercise: MobiDB and DisProt

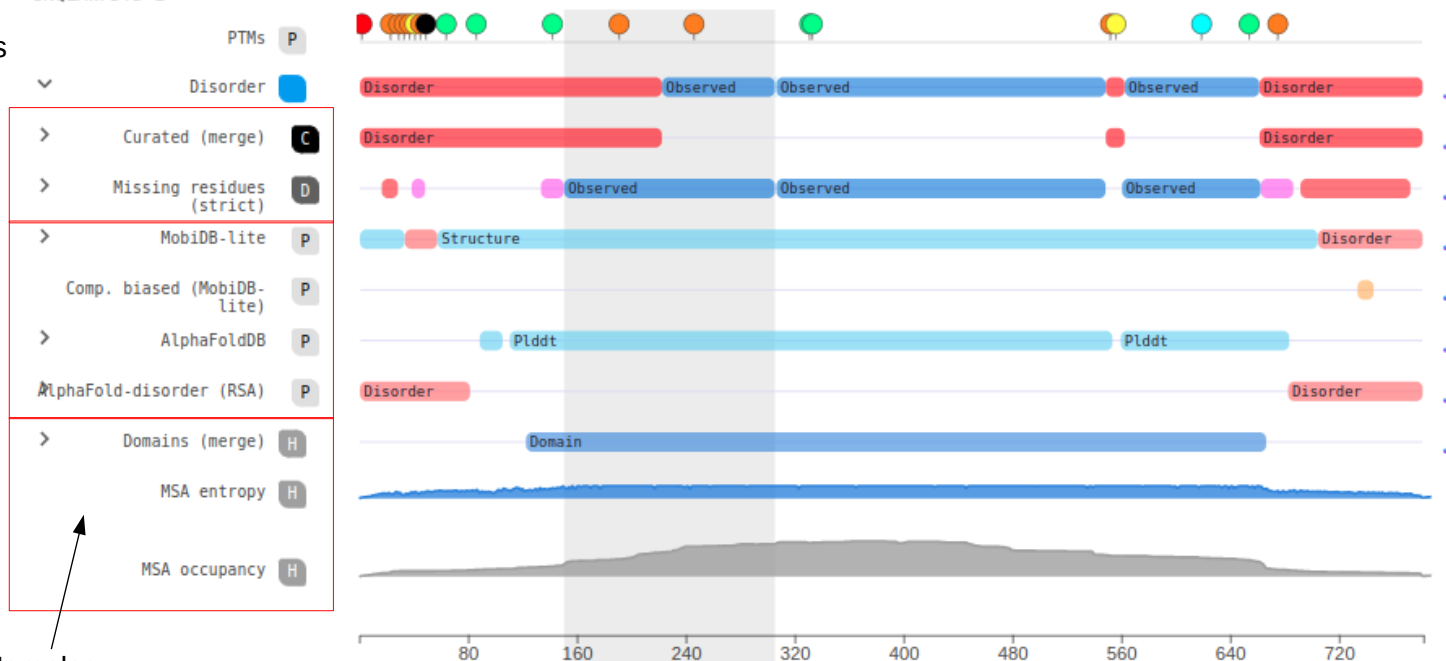
```
120      130      140      150      160      170      180      190      200      210      220
STQFDAAHPT NVQRLAEP SQ MLKHAVV NLI NYQDDAELAT RAIPELTKLL NDEDQVVVNK AAVMVHQLSK KEASRHAIMR SPQMVSAIVR TMQNTNDVET ARCTAGTLHN
230      240      250      260      270      280      290      300      310      320      330
LSHHREGLLA IFKSGGIPAL VKMLGSPVDS VLFYAITTLH NLLLHQEGAK MAVRLAGGLO KMVALLNKTN VKFLAITTDC LQILAYGNQE SKLILASGG PQALVNIMRT
340      350      360      370      380      390      400      410      420      430      440
YTYEKLLWTT SRVLKVLSVC SSNKPAIVEA GGMQALGLHL TDPSRLVQN CLWTLRNLSD AATKQEGMEG LLGTLVQLLG SDDINVVTCA AGILSNLTCN NYKNKMMVCQ
450      460      470      480      490      500      510      520      530      540      550
VGGIEALVRT VLRAGDREDI TEPAICALRH LTSRHQEAEM AQNAVRLHYG LPVVKLLHP PSHWPLIKAT VGLIRNLALC PANHAPLREQ GAIPRLVQLL VRAHQDTQRR
560      570      580      590      600      610      620      630      640      650      660
TSMGGTQQQF VEGVRMEEIV EGCTGALHIL ARDVHNRIVI RGLNTIPLFV QLLYSPIENI QRVAAGVLCE LAQDKEAEA EIEAEGATAPL TELLHSRNEG VATYAAAVLF
670      680      690      700      710      720      730      740      750      760      770
RMSEDKPQDY KKRLSVELTS SLFRTEPMAW NETADLGLDI GAQEPLGYR QDDPSYRSFH SGGYGQDALG MDPMEHEMG GHHPGADYPV DGLPDLGHAQ DLMDGLPPGD
780
SNQLAWFDTD L
```

Experimental results

Predictions

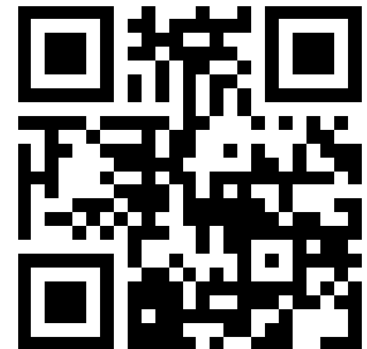
Search protein beta-catenin-1 (P35222)

Homology



Exercise: MobiDB and DisProt

- Search beta-catenin-1 (P35222) on MobiDB. **On top of the protein structure, click on the Disorder tab.**
 - 1. What kind of disorder annotations are there for this protein?
 - 2. The different annotations don't give exactly the same results. Which annotation gives the longest IDRs (Intrinsically Disordered Regions)?
- Go to the Disprot entry linked on top box with cross references.
 - 3. How many different Experimental techniques (indicated as "Evidence") were used to identify the presence of disorder?



Exercise: MobiDB and DisProt

- Search beta-catenin-1 (P35222) on MobiDB. **On top of the protein structure, click on the Disorder tab.**
 - 1. What kind of disorder annotations are there for this protein?
 - 2. The different annotations don't give exactly the same results. Which annotation gives the longest IDRs (Intrinsically Disordered Regions)?
- Go to the Disprot entry linked on top box with cross references.
 - 3. How many different Experimental techniques (indicated as "Evidence") were used to identify the presence of disorder?



Exercise: IDPs and cancer mutations

<https://pecan.stjude.cloud/variants/proteinpaint>

Or

<https://pecan.stjude.cloud/>

> Click in Variants

(bottom-left);

> Click in “Explore Variants”;

> Then you search the tab “PROTEINPAINT”.

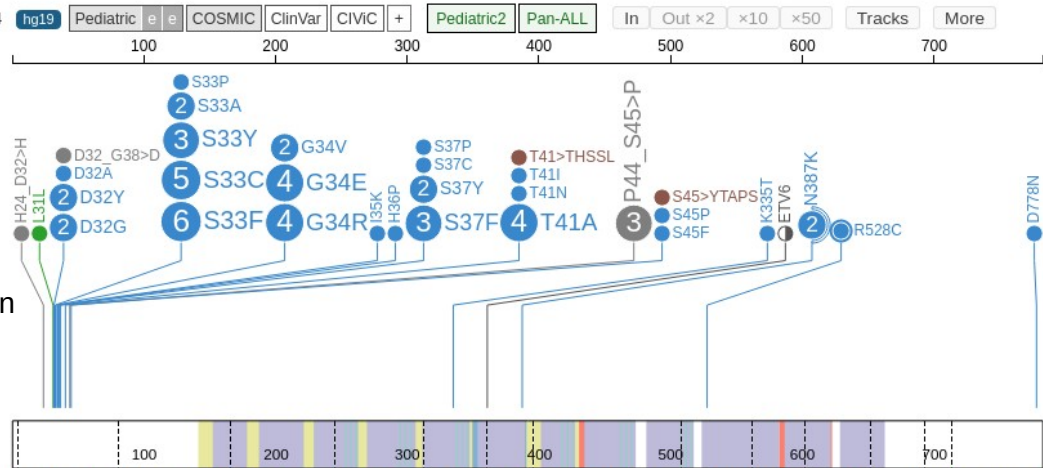
Search

CTNNB1 NM_001904

Protein length

Pediatric 63 of 64 mutations
12 cancer subtypes
7 datasets

Cancer type distribution
Click to select and visualize types



Proteinpaint a webtool to explore pediatric and adult cancer mutations.

You can click on these to filter the results

LEGEND
CLASS | 5589 MISSENSE | 270 PROTEINDEL | 52 SILENT | 42 NONSENSE | 40 Fusion transcript | 26 FRAMESHIFT | 13 SPLICE | 9 PROTEININS | 5 SPLICE_REGION | 1 INTRON

Exercise: IDPs and cancer mutations

- 4. Go on PECAN at <https://pecan.stjude.cloud/variants/proteinpaint> and search for beta-catenin (CTNNB1). Where in the sequence (residue number) are most of the mutations located in pediatric cancers?
- 5. In what type of cancer are the mutations common?
- 6. Turn on COSMIC (Catalogue Of Somatic Mutations In Cancer) mutations. In what location are mutations most common?
- 7. Go on DIBS (Database of Disordered Binding Sites) at <http://dibs.enzim.ttk.mta.hu/search.php> and search for beta-catenin (P35222). Which entry contains the mutated region?
- 8. Highlight the most mutated positions in the structure. Are they structured? What evidence is there for their status (ordered/disordered)?
- 9. What is the binding partner?



Exercise: IDPs and cancer mutations

<https://pecan.stjude.cloud/variants/proteinpaint>

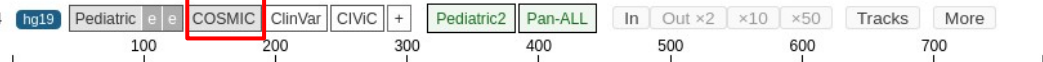
Or
<https://pecan.stjude.cloud/>

- > Click in Variants (bottom-left);
- > Click in “Explore Variants”;
- > Then you search the tab “PROTEINPAINT”.

Search

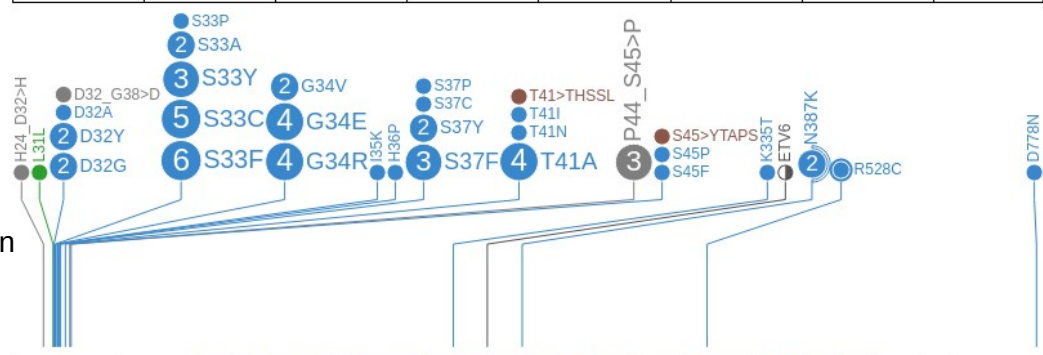
CTNNB1 NM_001904

Turning COSMIC mutations on



Protein length

Pediatric
 63 of 64 mutations
 12 cancer subtypes
 7 datasets

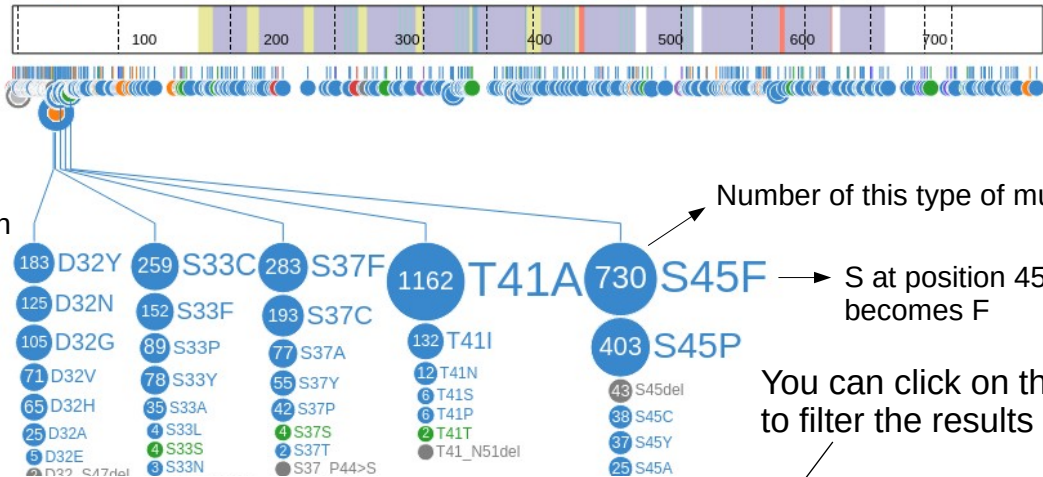


Cancer type distribution
 Click to select and visualize types

CTNNB1
 NM_001904

COSMIC
 5941 of 5983 mutations
 34 tissue types

Tissue distribution
 Click to select tissues



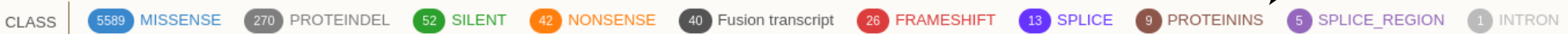
Number of this type of mutation

S at position 45 becomes F

You can click on these to filter the results

Proteinpaint a webtool to explore pediatric and adult cancer mutations.

LEGEND



Exercise: IDPs and cancer mutations

- 4. Go on PECAN at <https://pecan.stjude.cloud/variants/proteinpaint> and search for beta-catenin (CTNNB1). Where in the sequence (residue number) are most of the mutations located in pediatric cancers?
- 5. In what type of cancer are the mutations common?
- 6. Turn on COSMIC (Catalogue Of Somatic Mutations In Cancer) mutations. In what location are mutations most common?
- 7. Go on DIBS (Database of Disordered Binding Sites) at <http://dibs.enzim.ttk.mta.hu/search.php> and search for beta-catenin (P35222). Which entry contains the mutated region?
- 8. Highlight the most mutated positions in the structure. What evidence is there for their status (ordered/disordered)?
- 9. Click in Evidence in the left menu. Can you identify which is the binding partner? (If you visualize the PDB structure on the top right, you will see that only one of the partners actually interact with beta-catenin - “orange ribbon”)



Exercise: IDPs and cancer mutations

DIBs: Database of Disordered Binding Sites

<http://dibs.enzim.ttk.mta.hu/search.php>

>Go to PDB 1p22
>Find the chain corresponding to P35222.

Structure Summary

Entry contents: 2 distinct polypeptide molecules

Chains: A, B

Notes: No modifications of the original PDB file.

Chain A

Structural status

Name: Cellular tumor antigen p53 **Disordered** **Confirmed**

Source organism: *Homo sapiens*

Length: 20 residues

Sequence:  SHLKSKKGQSTSRHKLMFK

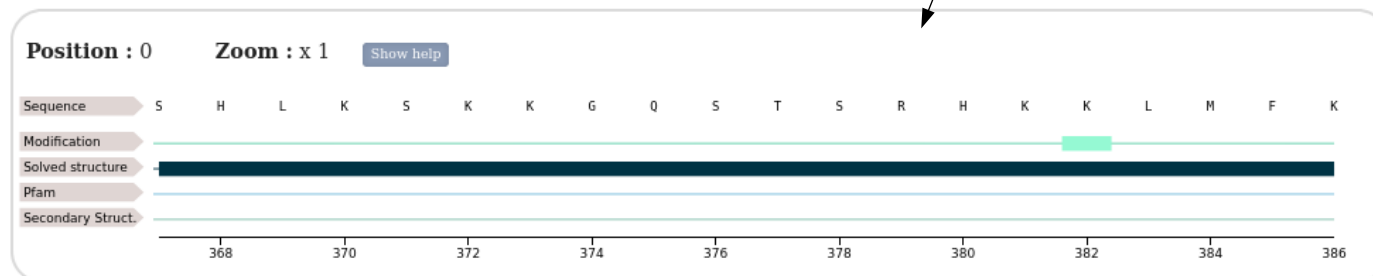
The sequence contains the following modified/non-standard residues:

- N(6)-acetyllysine (K) at position 382 (PDB position: 382)

UniProtKB AC: P04637 (positions: 367-386) **UniProt** Coverage: 5.1%

UniRef90 AC: UniRef90_P04637 (positions: 367-386) **UniRef90**

Residues of this chain involved in binding



Chain B

Name: CREB-binding protein **Ordered**

Source organism: *Homo sapiens*

Length: 121 residues

General Information

Function and Biology

Structure Summary

▪ Chain A

▪ Chain B

Evidence

Related Structure(s)

Click to go to evidence for structural status

Domain Type:

Bromodomain

Exercise: IDPs and cancer mutations

- 4. Go on PECAN at <https://pecan.stjude.cloud/variants/proteinpaint> and search for beta-catenin (CTNNB1). Where in the sequence (residue number) are most of the mutations located in pediatric cancers?
- 5. In what type of cancer are the mutations common?
- 6. Turn on COSMIC (Catalogue Of Somatic Mutations In Cancer) mutations. In what location are mutations most common?
- 7. Go on DIBS (Database of Disordered Binding Sites) at <http://dibs.enzim.ttk.mta.hu/search.php> and search for beta-catenin (P35222). Which entry contains the mutated region?
- 8. Highlight the most mutated positions in the structure. What evidence is there for their status (ordered/disordered)?
- 9. Click in Evidence in the left menu. Can you identify which is the binding partner? (If you visualize the PDB structure on the top right, you will see that only one of the partners actually interact with beta-catenin - “orange ribbon”)

References

Wright, P. E., & Dyson, H. J. (1999). Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *Journal of molecular biology*, 293(2), 321–331. <https://doi.org/10.1006/jmbi.1999.3110>.

Uversky, V. N., Gillespie, J. R., & Fink, A. L. (2000). Why are "natively unfolded" proteins unstructured under physiologic conditions?. *Proteins*, 41(3), 415–427. PMID: 11025552.

Uversky V. N. (2013). A decade and a half of protein intrinsic disorder: biology still waits for physics. *Protein science : a publication of the Protein Society*, 22(6), 693–724. <https://doi.org/10.1002/pro.2261>.

Wells, M., Tidow, H., Rutherford, T. J., Markwick, P., Jensen, M. R., Mylonas, E., Svergun, D. I., Blackledge, M., & Fersht, A. R. (2008). Structure of tumor suppressor p53 and its intrinsically disordered N-terminal transactivation domain. *Proceedings of the National Academy of Sciences of the United States of America*, 105(15), 5762–5767. <https://doi.org/10.1073/pnas.0801353105>.

Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M., & Obradović, Z. (2002). Intrinsic disorder and protein function. *Biochemistry*, 41(21), 6573–6582. <https://doi.org/10.1021/bi012159+>.

Kastano, K., Erdős, G., Mier, P., Alanis-Lobato, G., Promponas, V. J., Dosztányi, Z., & Andrade-Navarro, M. A. (2020). Evolutionary Study of Disorder in Protein Sequences. *Biomolecules*, 10(10), 1413. <https://doi.org/10.3390/biom10101413>.

Uversky, Vladimir N. *Intrinsically Disordered Proteins*. SpringerBriefs in Molecular Science. Cham: Springer International Publishing, 2014. <https://doi.org/10.1007/978-3-319-08921-8>.

Gonçalves-Kulik, M., Mier, P., Kastano, K., Cortés, J., Bernadó, P., Schmid, F., & Andrade-Navarro, M. A. (2022). Low Complexity Induces Structure in Protein Regions Predicted as Intrinsically Disordered. *Biomolecules*, 12(8), 1098. <https://doi.org/10.3390/biom12081098>.

Gonçalves-Kulik, M., Schmid, F., & Andrade-Navarro, M. A. (2023). One Step Closer to the Understanding of the Relationship IDR-LCR-Structure. *Genes*, 14(9), 1711. <https://doi.org/10.3390/genes14091711>