Structure Prediction Module-Guide

Introduction

Proteins are composed of chains of amino acids. The sequence of amino acids determines a protein's three-dimensional structure, which then determines a protein's function. A protein's structure can be broken down into a hierarchy of structures, primary, secondary, tertiary and for larger proteins quaternary. Secondary structure is the first level of folding of the amino acid chain, which includes either alpha helices or beta sheets. Within a protein there can be different regions with different structural and sequence motifs. Structural motifs are commonly found substructures, which can be associated with certain functions or properties. One example of a structural motif is a helix-turn-helix motif which is a conformation that allows for binding to DNA and this type of motif is commonly found in transcription factors.
Because there are way more possibilities for protein sequence than possible protein folds, many proteins with very different sequences can adopt the same protein fold/structure. This is useful when looking at evolutionary novel proteins, which do not have similar sequence to any other proteins but likely have some similar structure domains. However keep in mind there are examples of proteins with similar folds but different functions and examples of proteins with different folds but similar functions! As well as proteins or protein domains that do not form a stable structure but are intrinsically disordered instead.

There are methods to experimentally determine the structure of a protein, such as x-ray crystallography, however these methods are very time consuming and do not work for all proteins. Because of this many researchers have been trying to find ways to accurately predict the potential structure of a protein, which is termed 'the protein folding problem'. There is a yearly contest called CASP where researchers use their latest algorithms to try and win the best algorithm to predict protein structure. In 2020 the algorithm AlphaFold2 won the yearly CASP contest by a large margin, producing protein structures with way higher accuracy than anyone had seen before. AlphaFold2 was dubbed a solution to "One of the Biggest Problems" and news reports with headlines proclaiming the protein folding problem has been solved. One major limitation to AlphaFold2 is that it relies on sequence conservation information to predict structure, meaning it cannot predict newly evolved sequences with high confidence. In response to this other research teams have come out with algorithms such as ESMFold (in 2023) and Omega Fold (in 2022) which do not rely on sequence conservation information; however, the confidence level for evolutionary novel sequences is still lower than for conserved sequences.

In this module we will use one of these cutting edge algorithms, ESMFold, to predict the potential structure of your proto-gene. We will then compare the predicted structures to other known structures using Foldseek to identify structural motifs that could give us insights into its potential function.

Goals: In this module, participants will learn how to investigate ESMFold structure prediction, Foldseek structure search and CATH databases

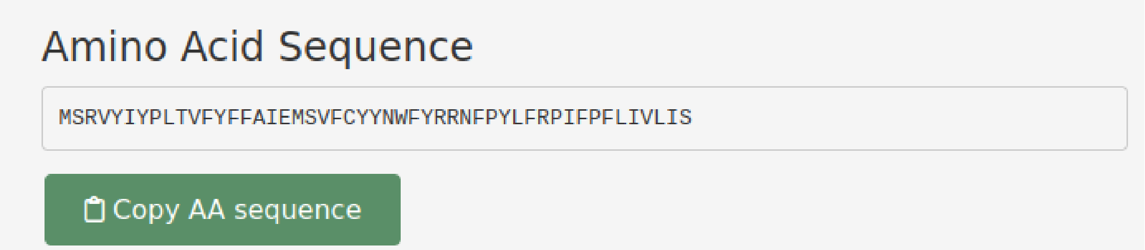
Objectives: After completing this module, participants will be able to:

1) Use ESMFold to investigate protein structure prediction from a primary amino acid sequence and assess the confidence of the prediction
2) Use Foldseek to search for proteins with similar structures
3) Explore CATH search results to determine potential GO terms and functions of proteins with similar structures

Activity

Predict 3D protein structure using ESMFold

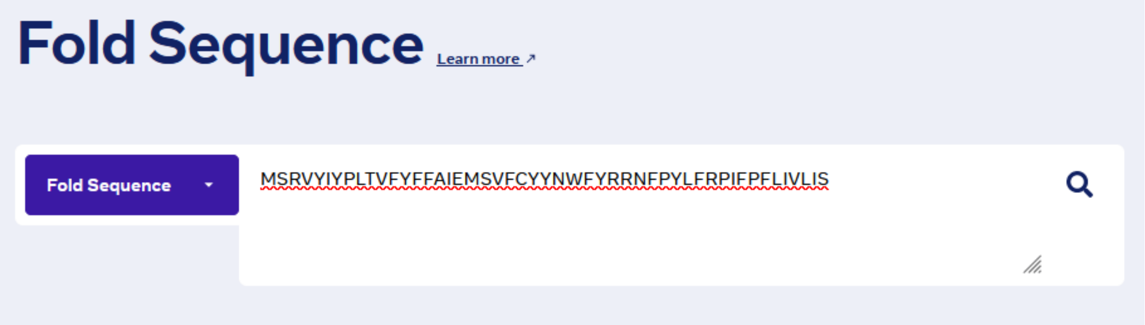
1. Get the amino acid sequence for your proto-gene from the ORF info app: https://carvunislab.csb.pitt.edu/shiny/coexpression/



Amino Acid Sequence

MSRVYIYPLTVFYFFAIEMSVFCYYNWFYRRNFPYLFRPIFPFLIVLIS

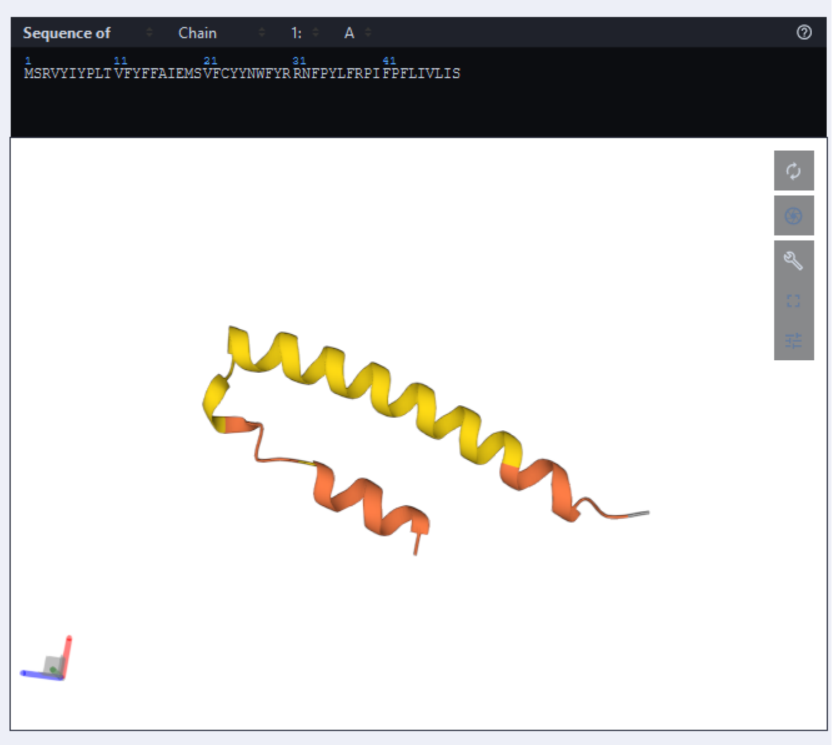
2. Go to the ESMFold website https://esmatlas.com/resources?action=fold
3. Paste the amino acid sequence into the text box and click the search icon. Make sure there are no spaces in your amino acid sequence or you will get an error).



Fold Sequence

MSRVYIYPLTVFYFFAIEMSVFCYYNWFYRRNFPYLFRPIFPFLIVLIS

4. The predicted structure will be output on the screen. The coloring scheme indicates how confident the algorithm is in its prediction, where red indicates very low confidence, orange means low confidence, yellow means medium confidence, light blue means high confidence and dark blue indicates very high confidence. You can rotate the structure by clicking and dragging.



Sequence of Chain 1: A

```
1         11        21        31        41
MSRVYIYPLTVFYFFAIEMSVFCYYNWFYRRNFPYLFRPIFPFLIVLIS
```

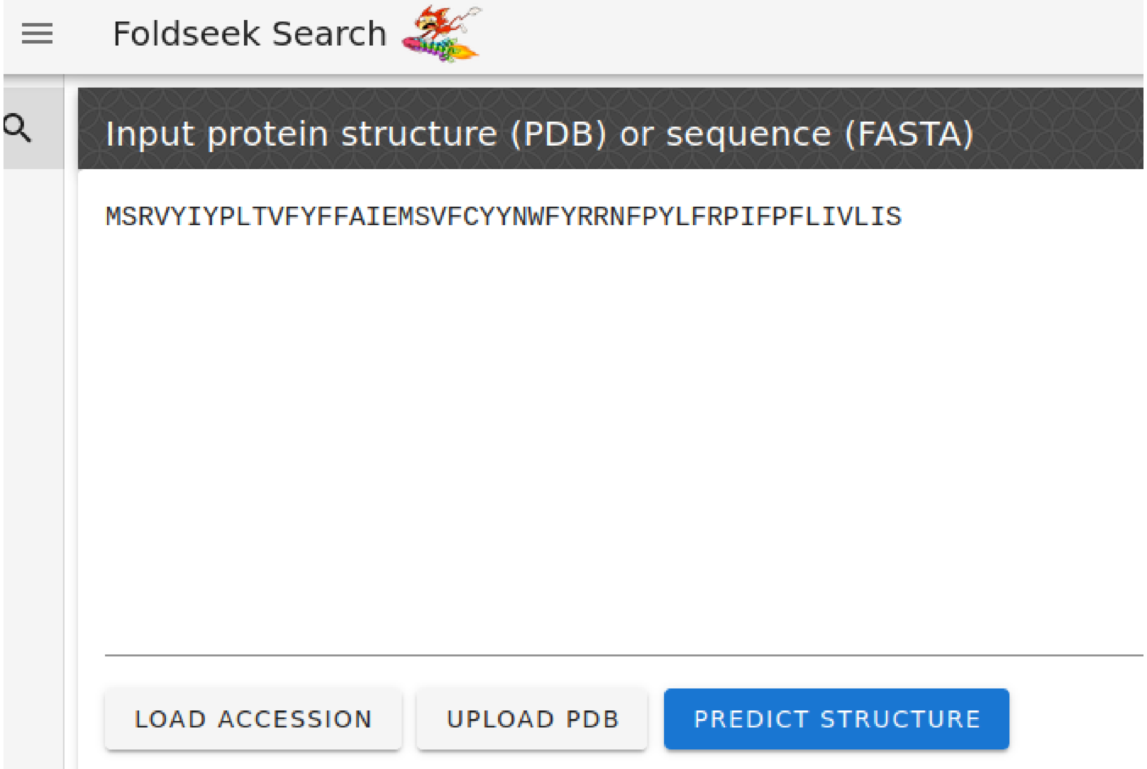
Answer question 1 in the worksheet: What is the predicted structure of your protein as determined by ESMFold? Is ESMFold confident in its prediction?

Foldseek: find proteins with similar structures

Foldseek is an algorithm that searches known protein structures to find ones that look similar to the query structure and then performs a structural alignment to calculate how similar the structures are. Foldseek then returns information about proteins from different databases. In this activity we will use Foldseek to search for proteins with similar structures to your proto-gene in the CATH database.

5. Next we will determine which proteins contain similar structures similar to the predicted structure of your proto-gene. Go to the Foldseek website: https://search.foldseek.com/search

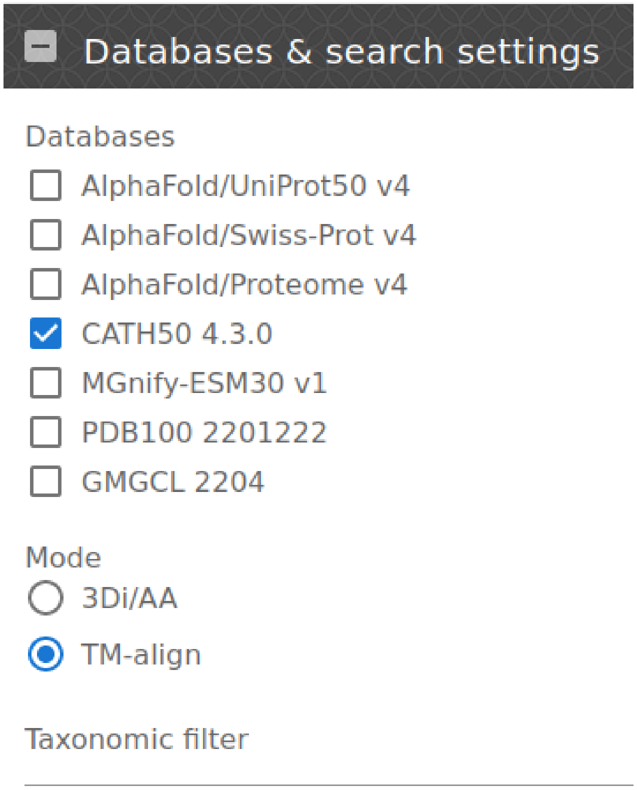
6. On the Foldseek website, paste your proto-gene's amino acid sequence into the box and click the 'Predict Structure' button. There will probably be text already in the box which you can press the 'ctrl'+'a' buttons on your keyboard to select the text inside and delete it. Foldseek uses ESMFold in the background to predict the structure from an amino acid sequence and converts it into the correct format for querying.



```
MSRVYIYPLTVFYFFAIEMSVFCYYNWFYRRNFPYLFRPIFPFLIVLIS
```

Once ESMFold is done predicting the structure, you will see text again in the box. This text contains the coordinate positions for the atoms in the predicted structure.

7. Next click the + button beside the Databases & search settings heading to specify which databases we want to search in and what type of structure alignment we want to use. In this activity we will only use the CATH50 database to search against, so unselect all other boxes. For the Mode setting choose TM-align. Then click the ‘search’ button.



Databases & search settings

Databases

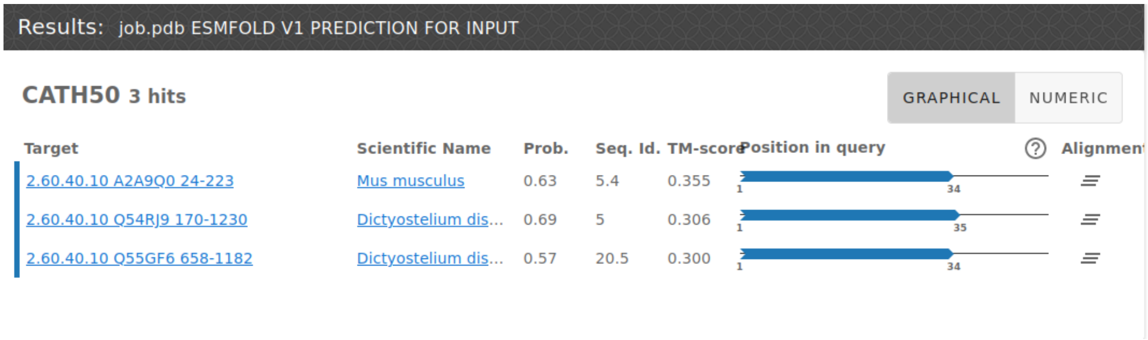
□ AlphaFold/UniProt50 v4
□ AlphaFold/Swiss-Prot v4
□ AlphaFold/Proteome v4
☑ CATH50 4.3.0
□ MGnify-ESM30 v1
□ PDB100 2201222
□ GMGCL 2204

Mode

○ 3Di/AA
● TM-align

Taxonomic filter

8. The output will look something like this: the *Target* column represents a protein/class of proteins that has a similar structure to your protein (the query). The *Prob.* column shows the probability that the target is a good match to your query protein, and has values that range from 0-1 where a probability of 0 means the target is unlikely to be a good match and a probability of 1 means very likely the target is a good match. The *Seq. Id.* column represents the sequence identity which is the percentage of your proto-gene's amino acid sequence that is shared with the target protein. Sequence identity scores range from 0-100%, where 100% means identical amino acid sequence. The *TM-Score* column shows the similarity between the target and query protein structure. TM-Scores range from 0-1, where a score of 0 means no similarity and a score of 1 means perfect similarity. The *Position in query* column shows what parts of the query protein align with the target. The results are ordered by decreasing TM-score.

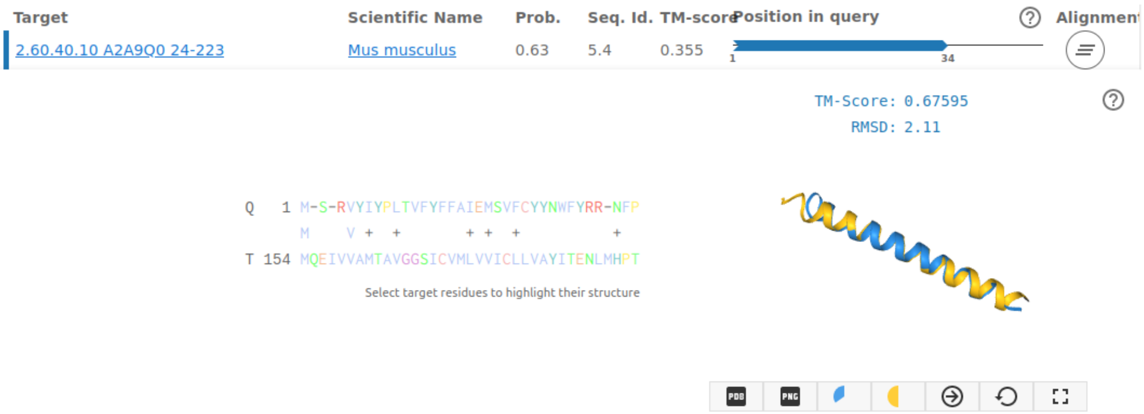


Results: job.pdb ESMFOLD V1 PREDICTION FOR INPUT

CATH50 3 hits

Target	Scientific Name	Prob.	Seq. Id.	TM-score	Position in query	Alignmen
2.60.40.10 A2A9Q0 24-223	Mus musculus	0.63	5.4	0.355	1 34	≡
2.60.40.10 Q54RJ9 170-1230	Dictyostelium dis...	0.69	5	0.306	1 35	≡
2.60.40.10 Q55GF6 658-1182	Dictyostelium dis...	0.57	20.5	0.300	1 34	≡

If you click on the 3 lines under the *Alignment* column it will show the structural alignment between your proto-gene's protein (in blue) and the target protein (in yellow). You can click and drag the structure to rotate the proteins.



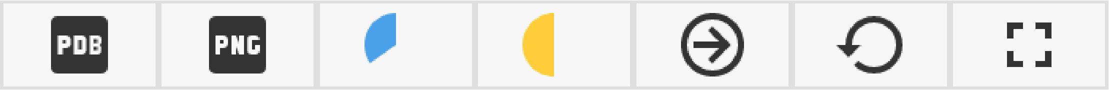
Target	Scientific Name	Prob.	Seq. Id.	TM-score	Position in query	Alignmen
2.60.40.10 A2A9Q0 24-223	Mus musculus	0.63	5.4	0.355	1 34	

TM-Score: 0.67595
RMSD: 2.11

```
Q   1 M-S-RVYIYPLTVFYFFAIEMSVFCYYNWFYRR-NFP
      M    V + +       + +  +          +
T 154 MQEIVVAMTAVGGSICVMLVVICLLVAYITENLMHPT
```

Select target residues to highlight their structure

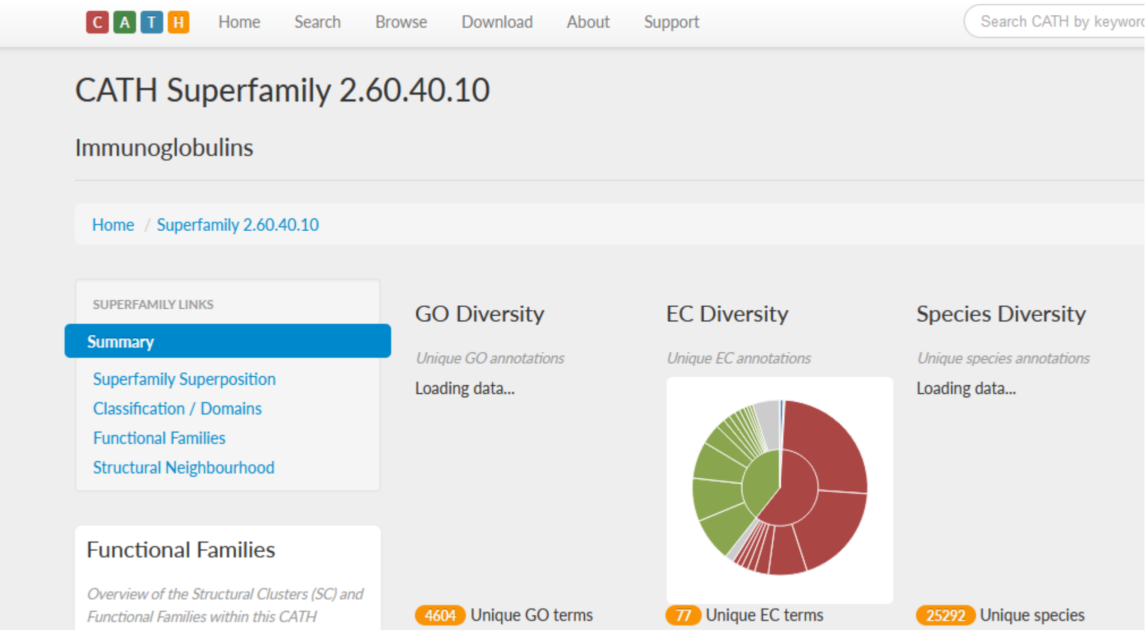
The button below the structure with the yellow circle allows you to toggle between visualizing only the structure of your proto-gene that aligns to the target (quarter or half blue circle) and visualizing all of your proto-gene (full blue circle). The button with the yellow circle does the same but for the target protein. You can click on the square outline button to view the structures in full screen as well as click the 'PNG' button to save a picture of the structure alignment.



Answer question 2-3 in the worksheet: What is the top scoring target? (i.e. the target with the highest TM-score) Record the probability, sequence identity, TM-score, and picture of the structure alignment in your worksheet. Do you find the structure alignment between your proto-gene and the top scoring target convincing? Why or why not?

CATH: investigate proteins with similar structures

9. To find out more about the proteins that have a similar structure to your proto-gene, click on one of the links in the *Target* column. This will take you to the CATH website. Click on the Classification/Domains tab on the left hand side



CATH Superfamily 2.60.40.10

Immunoglobulins

Home / Superfamily 2.60.40.10

Functional Families

Overview of the Structural Clusters (SC) and Functional Families within this CATH

GO Diversity

Unique GO annotations

Loading data...

4604 Unique GO terms

EC Diversity

Unique EC annotations

77 Unique EC terms

Species Diversity

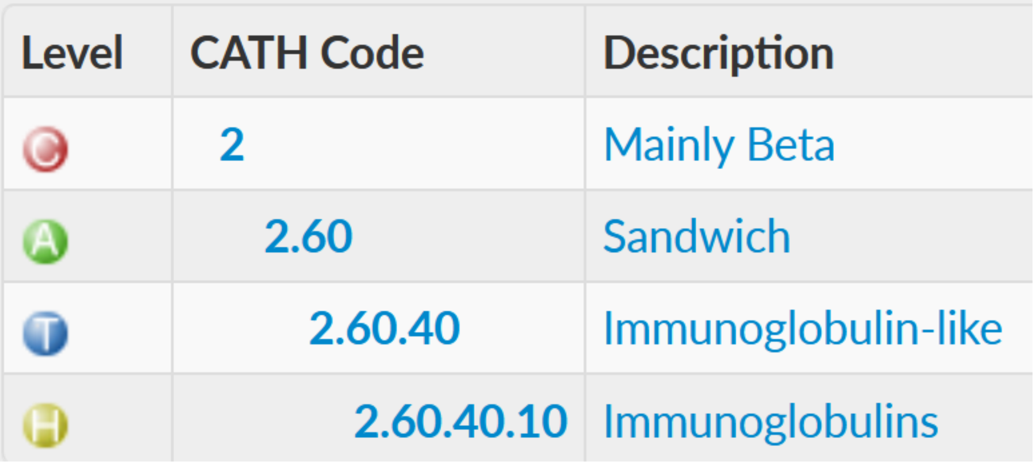
Unique species annotations

Loading data...

25292 Unique species

CATH database uses a hierarchy to categorize protein structures, which includes 4 levels:
Class (C), Architecture (A), Topology/fold (T), and Homologous superfamily (H) levels.
C level: describes the secondary structure of the protein.
A level: “A description of the gross arrangement of secondary structures, independent of connectivity”
T level: “groups proteins with the same overall fold, i.e. have a similar number and arrangement of secondary structures and connectivity linking their secondary structure. Within a given topology level, the structures are similar, but may have diverse functions.”
H level: “structures are grouped by their high structural similarity and similar functions.”

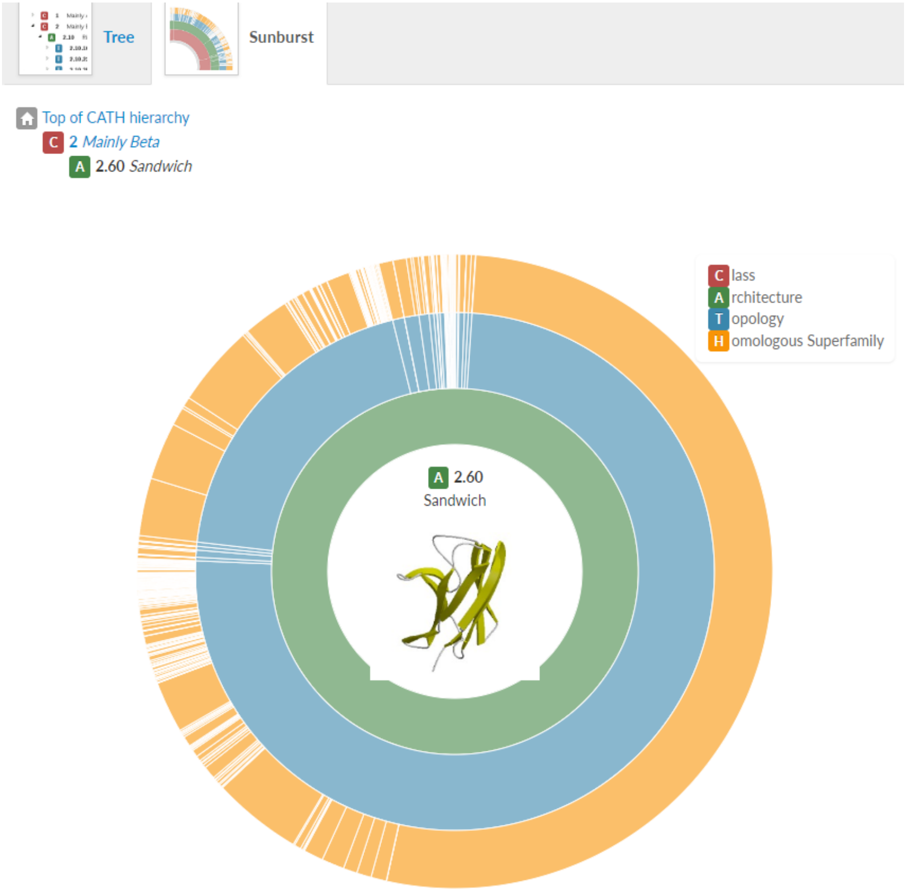
Here is the CATH classification for this example protein.



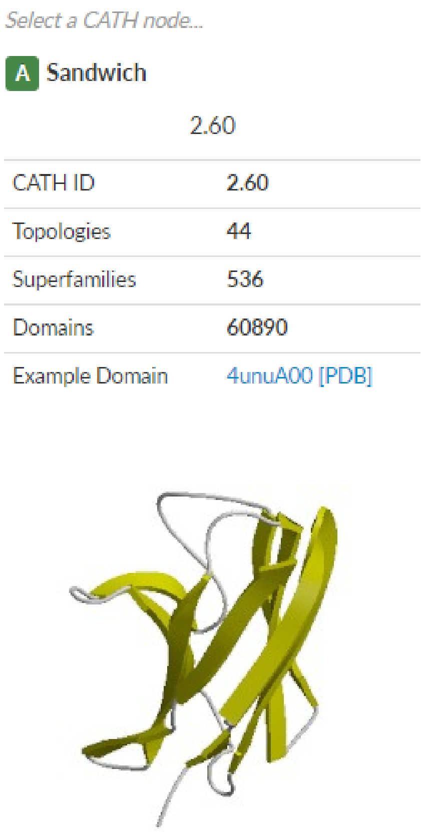
Level	CATH Code	Description
C	2	Mainly Beta
A	2.60	Sandwich
T	2.60.40	Immunoglobulin-like
H	2.60.40.10	Immunoglobulins

Answer question 4 on the worksheet: What is the CATH protein fold hierarchy for the proteins most similar to your proto-gene? (ie what is the CATH hierarchy for the target with the highest TM-score)

10. If you click on the links in the *Description* column, a sunburst plot of the hierarchy will show up. You can hover over the different categories within that level or subsequent ones to explore the different structures.



11. On the left hand side of the screen there is an example domain for each level. How do these example domains compare to your protein's predicted structure?



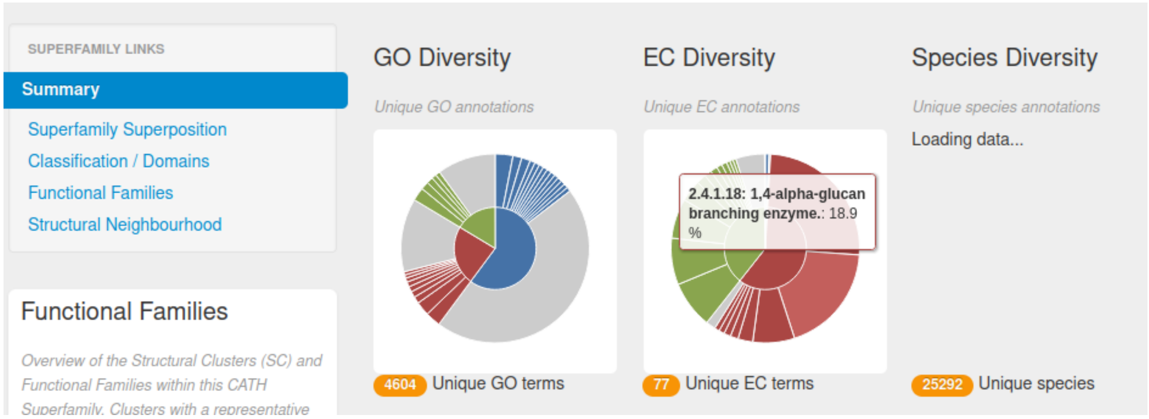
Select a CATH node...

A Sandwich

2.60

CATH ID	2.60
Topologies	44
Superfamilies	536
Domains	60890
Example Domain	4unuA00 [PDB]

12. Back on the summary tab, There is information about the functions and localizations of proteins with this type of structure under the GO diversity and EC diversity graphs. Hover over the GO diversity and EC diversity graphs to view the functions and localizations. Hover over the species Diversity graph to view the species these proteins are found in. These graphs can sometimes take a few minutes to load (~5 minutes).



SUPERFAMILY LINKS

Summary

Superfamily Superposition

Classification / Domains

Functional Families

Structural Neighbourhood

Functional Families

Overview of the Structural Clusters (SC) and Functional Families within this CATH Superfamily. Clusters with a representative

GO Diversity

Unique GO annotations

4604 Unique GO terms

EC Diversity

Unique EC annotations

2.4.1.18: 1,4-alpha-glucan branching enzyme.: 18.9 %

77 Unique EC terms

Species Diversity

Unique species annotations

Loading data...

25292 Unique species

Answer questions 5-8 on the worksheet: What is the largest percentage of species that this protein fold has been found in? What cellular component has the largest proportion of annotations for this protein fold? What molecular function has the largest proportion of annotations for this protein fold? What biological process has the largest proportion of annotations for this protein fold?

Sources

- Mirdita, M. et al. (2022). ColabFold: making protein folding accessible to all, *Nature Methods*. https://doi.org/10.1038/s41592-022-01488-1
- Jumper, J. et al. (2021). Highly accurate protein structure prediction with AlphaFold, *Nature*. https://doi.org/10.1038/s41586-021-03819-2
- Wu, R. et al. (2022). High-resolution de novo structure prediction from primary sequence, *BioRxiv*. https://doi.org/10.1101/2022.07.21.500999
- Lin, Z. et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model, *Science*. https://doi.org/10.1126/science.ade2574
- van Kempen, M. et al. (2023). Fast and accurate protein structure search with Foldseek, *Nature Biotechnology*. https://doi.org/10.1038/s41587-023-01773-0
- Orengo, C. et al. (1997). CATH – a hierarchic classification of protein domain structures, *Structure*. https://doi.org/10.1016/S0969-2126(97)00260-8