Date: 16 FEB 22

WEBLEM 2
INTRODUCTION TO PROTEIN CLASSIFICATION

Introduction:

One of the applications of protein structure comparison is structural classification. The ability to compare protein structure allows classification of the structure data and identification of relationships among structures. The reason to develop a protein structure classification system is to establish hierarchical relationships among protein structures and to provide a comprehensive and evolutionary view of known structure. Once a hierarchical classification system is established, a newly obtained protein structure can find its place in a proper category. As a result, its functions can be better understood based on association with other proteins.

CATH:

CATH (www.biochem.ucl.ac.uk/bsm/cath new/index.html) classifies proteins based on the automatic structural alignment program SSAP as well as manual comparison. Structural domain separation is carried out also as a combined effort of a human expert and computer programs. Individual domain structures are classified at five major levels: class, architecture, fold/topology, homologous superfamily, and homologous family. In the CATH release version 2.5.1 (January 2004), there are 4 classes, 37 architectures, 813 topologies, 1,467 homologous superfamilies, and 4,036 homologous families. The definition for class in CATH is similar to that in SCOP, and is based on secondary structure content. Architecture is a unique level in CATH, intermediate between fold and class. This level describes the overall packing and arrangement of secondary structures independent of connectivity between the elements. The topology level is equivalent to the fold level in SCOP, which describes overall orientation of secondary structures and takes into account the sequence connectivity between the secondary structure elements. The homologous superfamily and homologous family levels are equivalent to the superfamily and family levels in SCOP with similar evolutionary definitions, respectively.

SCOPe:

Structural Classification of Proteins-extended (SCOPe, http://scop.berkeley.edu) is a database of protein structural relationships that extends the SCOP database. SCOP is a manually curated ordering of domains from the majority of proteins of known structure in a hierarchy according to structural and evolutionary relationships. Development of the SCOP 1.x series concluded with SCOP 1.75. The ASTRAL compendium provides several databases and tools to aid in the analysis of the protein structures classified in SCOP, particularly through the use of their sequences. SCOPe extends version 1.75 of the SCOP database, using automated curation methods to classify many structures released since SCOP 1.75. We have rigorously benchmarked our automated methods to ensure that they are as accurate as manual curation, though there are many proteins to which our methods cannot be applied. SCOPe is also partially manually curated to correct some errors in SCOP. SCOPe aims to be backward compatible with SCOP, providing the same parseable files and a history of changes between all stable SCOP and SCOPe releases. SCOPe also incorporates and updates the ASTRAL database. The latest release of SCOPe, 2.03,

contains 59 514 Protein Data Bank (PDB) entries, increasing the number of structures classified in SCOP by 55% and including more than 65% of the protein structures in the PDB.

References:

1. Berman, H. M. (2000). The Protein Data Bank. *Nucleic Acids Research*, *28*(1), 235–242. https://doi.org/10.1093/nar/28.1.235
2. Murzin, A. G. (1995). *Journal of Molecular Biology*, *247*(4), 536–540. https://doi.org/10.1006/jmbi.1995.0159

DATE: 16 FEB 22

WEBLEM 2A
(URL:https://www.cathdb.info/)

Aim:

To study the structural classification of proteins (Leucine) using **CATH** and **SCOPe** database.

Introduction:

- **CATH:**
  - CATH (www.biochem.ucl.ac.uk/bsm/cath new/index.html) classifies proteins based on the automatic structural alignment program SSAP as well as manual comparison. Structural domain separation is carried out also as a combined effort of a human expert and computer programs. Individual domain structures are classified at five major levels: class, architecture, fold/topology, homologous superfamily, and homologous family. In the CATH release version 2.5.1 (January 2004), there are 4 classes, 37 architectures, 813 topologies, 1,467 homologous superfamilies, and 4,036 homologous families. The definition for class in CATH is similar to that in SCOP, and is based on secondary structure content. Architecture is a unique level in CATH, intermediate between fold and class.
- **SCOPe:**
  - Structural Classification of Proteins-extended (SCOPe, http://scop.berkeley.edu) is a database of protein structural relationships that extends the SCOP database. SCOP is a manually curated ordering of domains from the majority of proteins of known structure in a hierarchy according to structural and evolutionary relationships. Development of the SCOP 1.x series concluded with SCOP 1.75. The ASTRAL compendium provides several databases and tools to aid in the analysis of the protein structures classified in SCOP, particularly through the use of their sequences. SCOPe extends version 1.75 of the SCOP database, using automated curation methods to classify many structures released since SCOP 1.75.
- **Leucine:**
  - Leucine is one of nine essential amino acids in humans (provided by food), Leucine is important for protein synthesis and many metabolic functions. Leucine contributes to regulation of blood-sugar levels; growth and repair of muscle and bone tissue; growth hormone production; and wound healing. Leucine also prevents breakdown of muscle proteins after trauma or severe stress and may be beneficial for individuals with phenylketonuria. Leucine is available in many foods and deficiency is rare.

- **Methodology:**
  - Open Homepage of CATH and SCOPe
  - Enter the query Leucine in the search bar
  - Open the result page of each category
  - Interpret the results.

Observation:

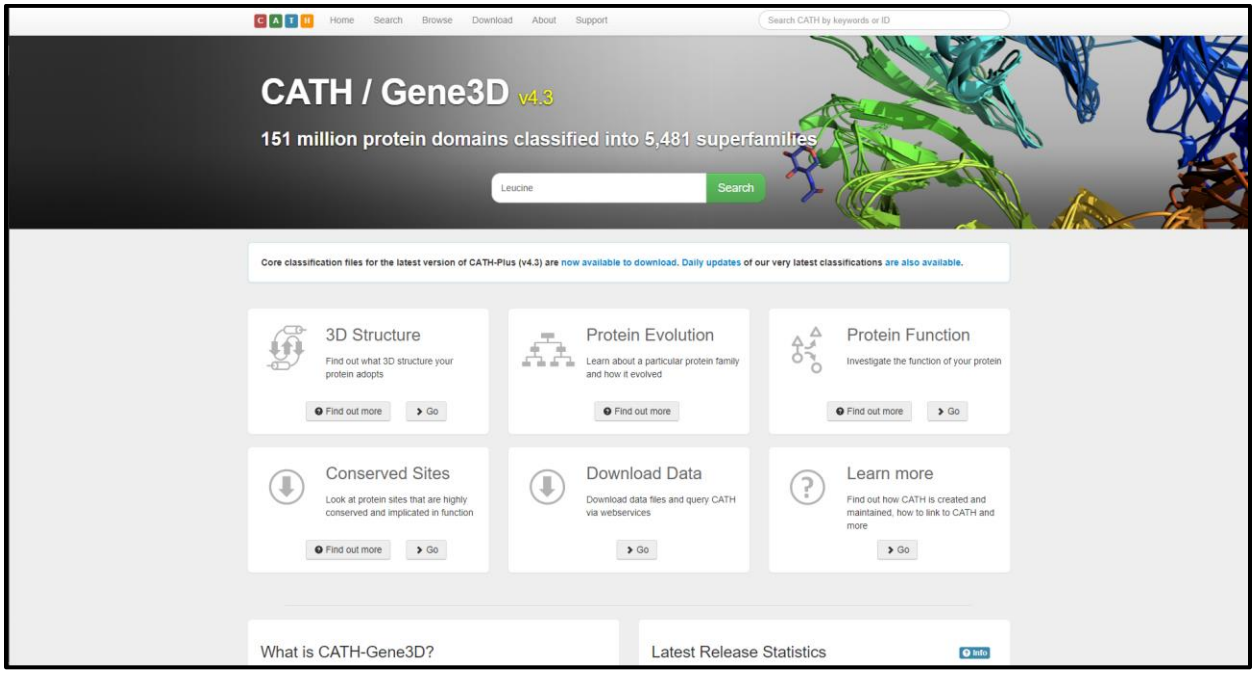

Fig1. Homepage of CATH database with query Leucine

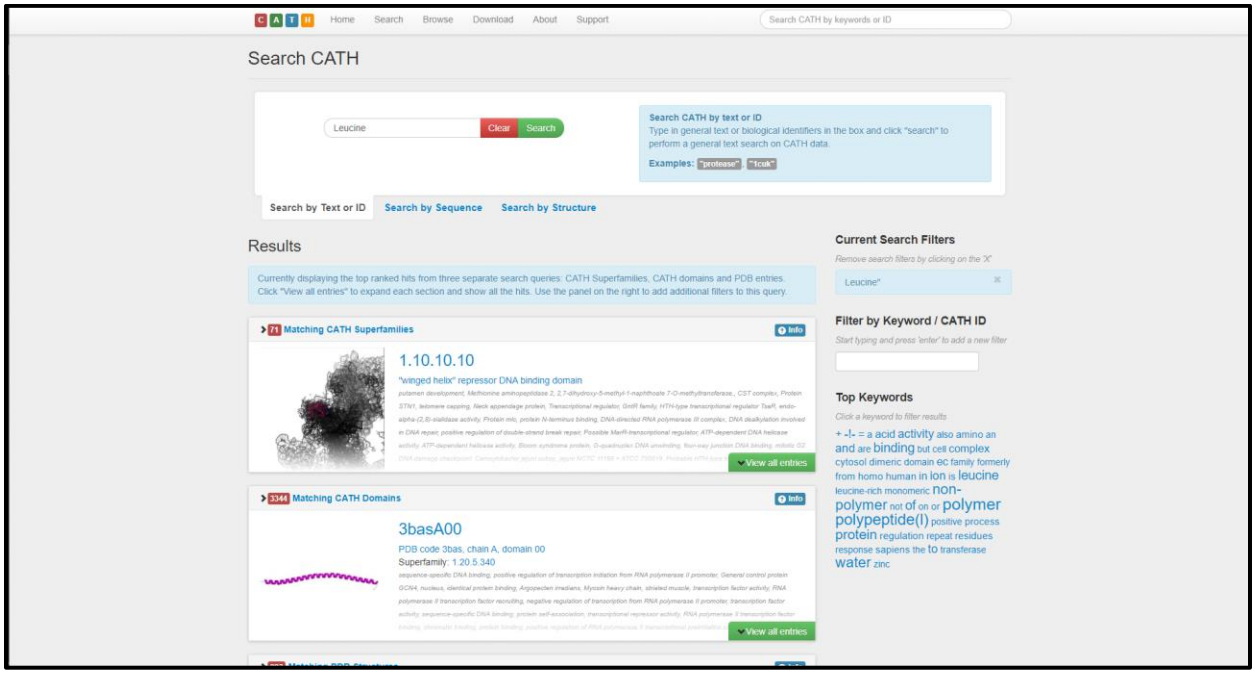

Fig2. Result page for leucine in CATH database

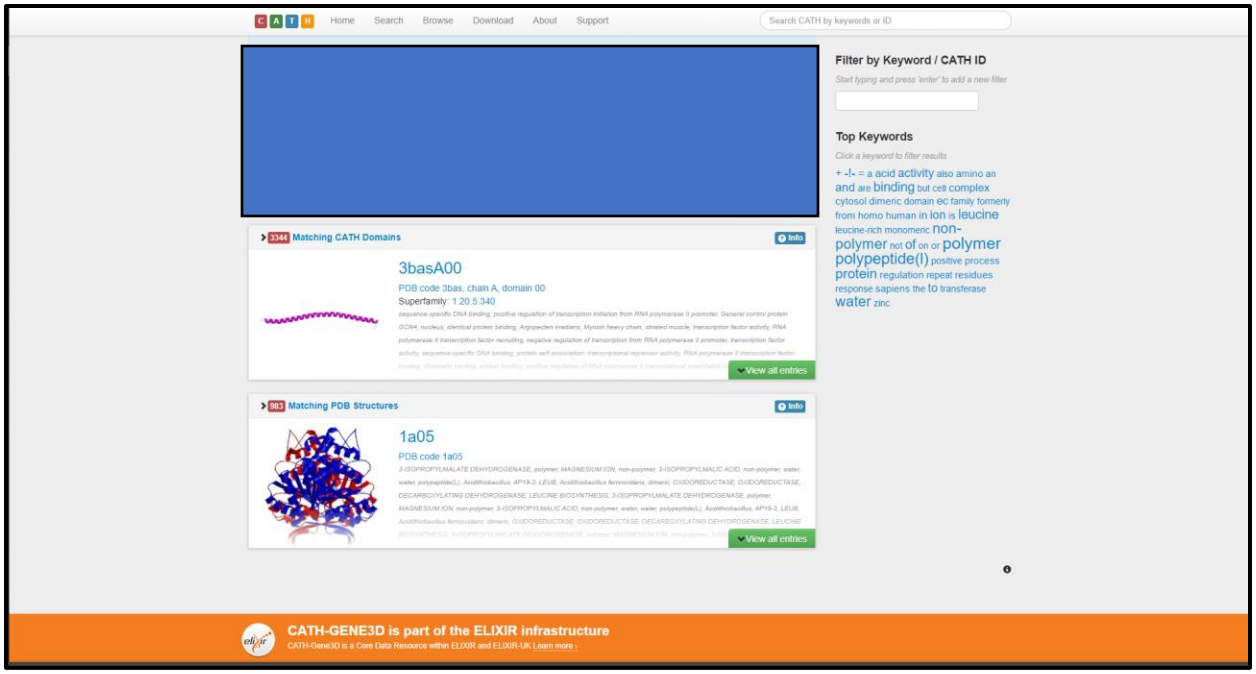

Fig3. Available CATH Superfamilies for leucine

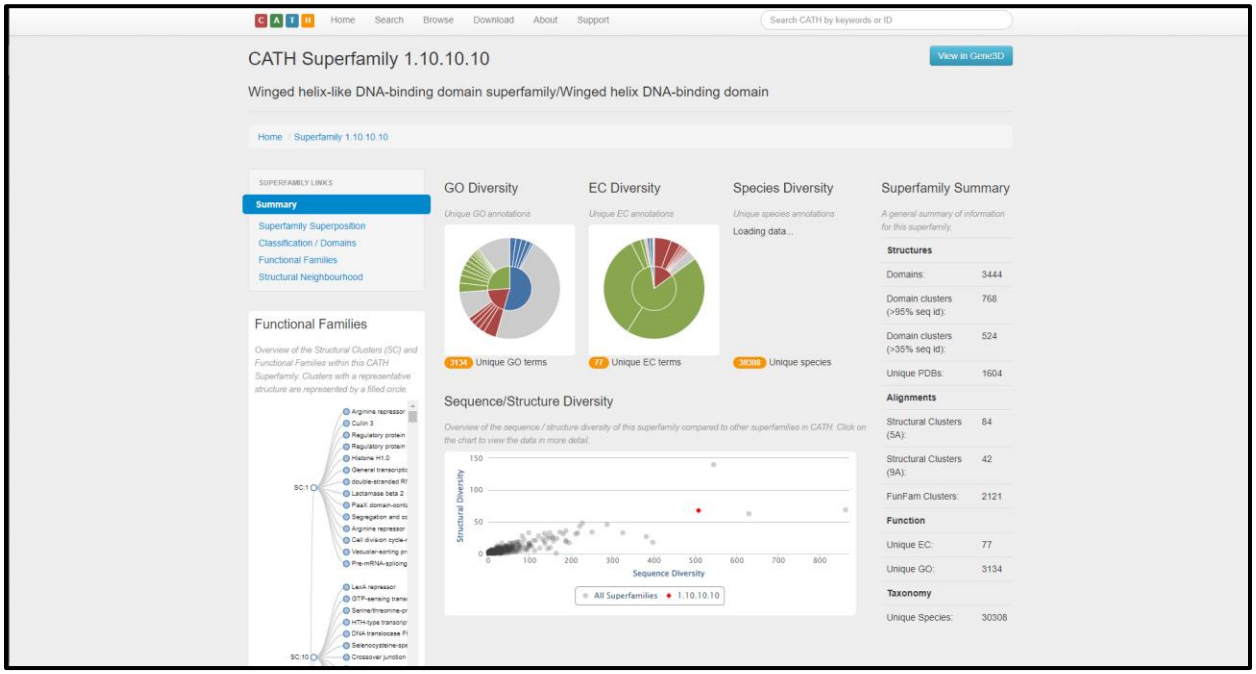

Fig3.1. Summary of CATH superfamily for Leucine

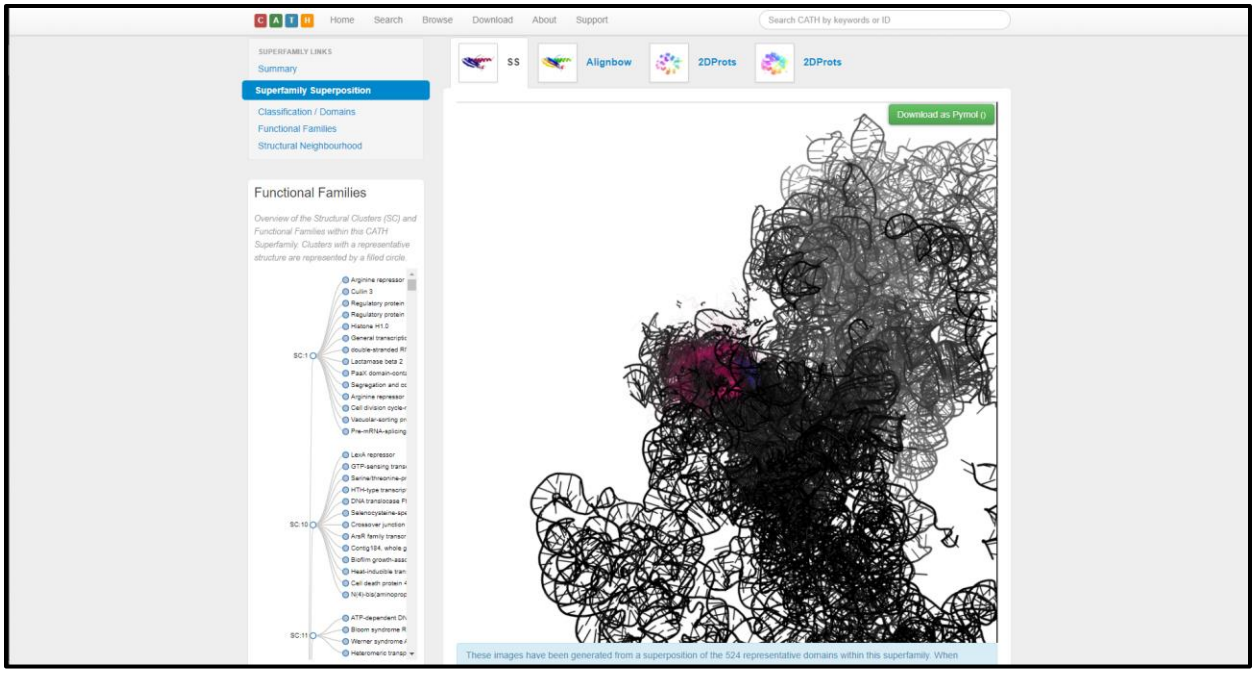

Fig3.2. Superfamily superposition under CATH superfamily for leucine

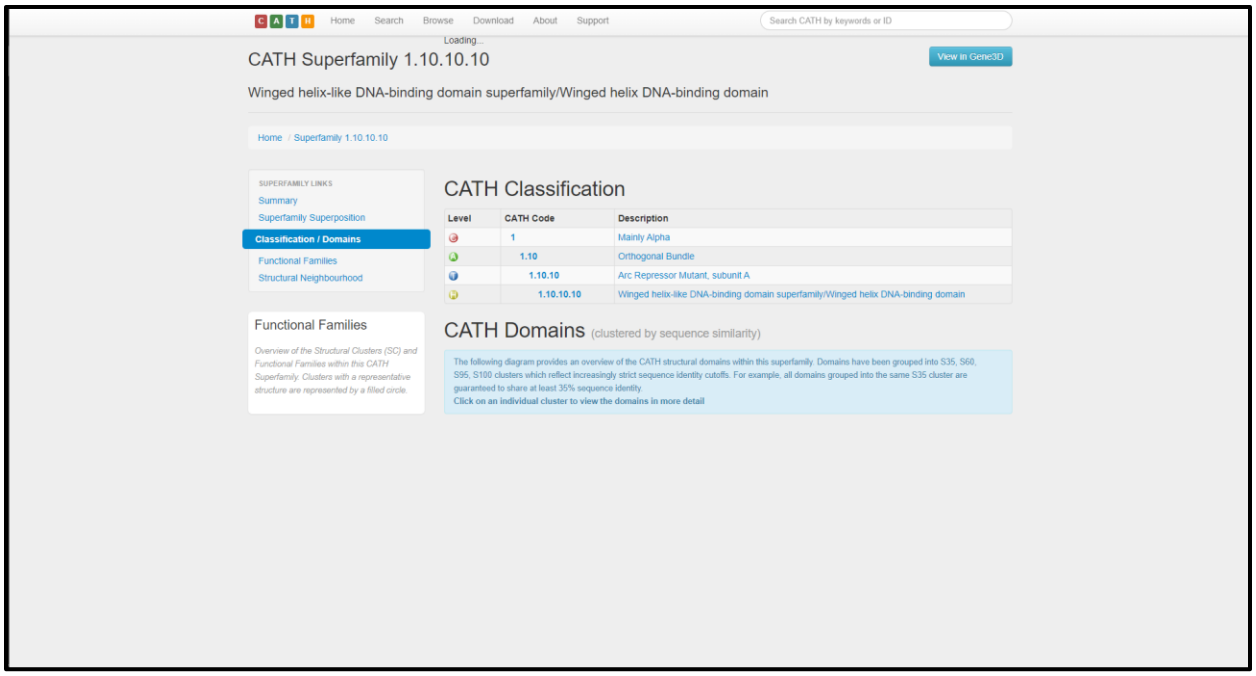

Fig3.3. Classification / Domain of CATH superfamilies for Leucine

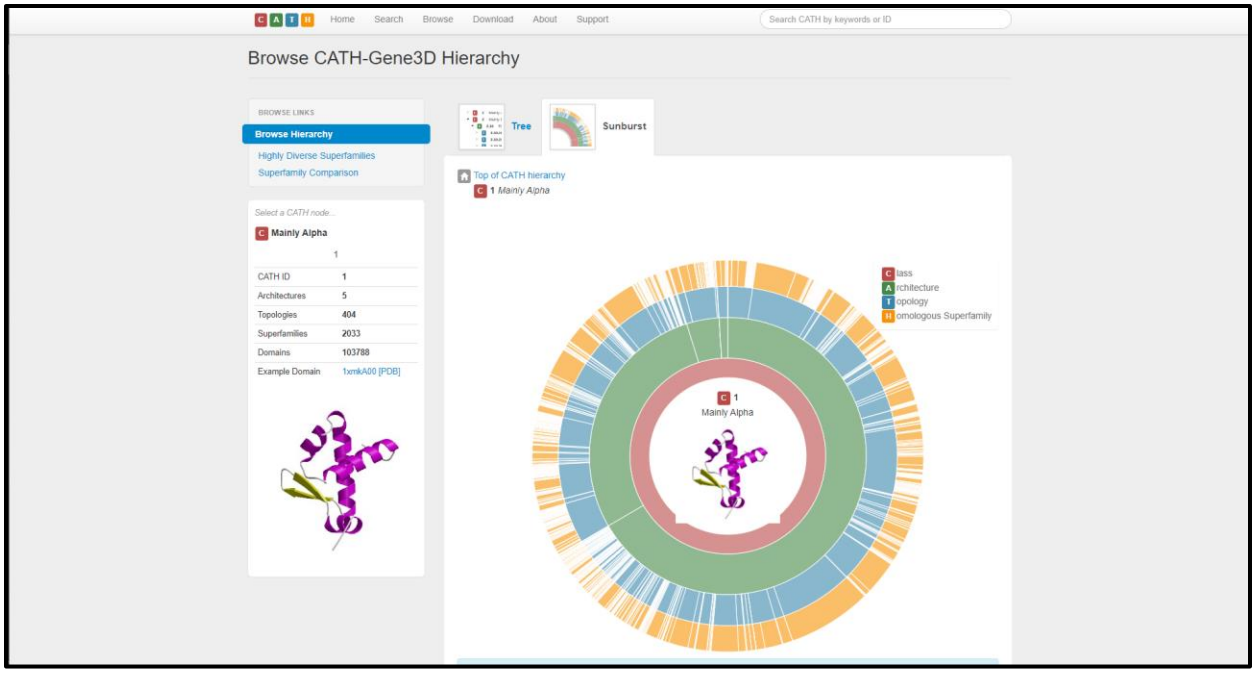

Fig3.4. Sunburst diagram of Alpha domain of CATH superfamilies for leucine

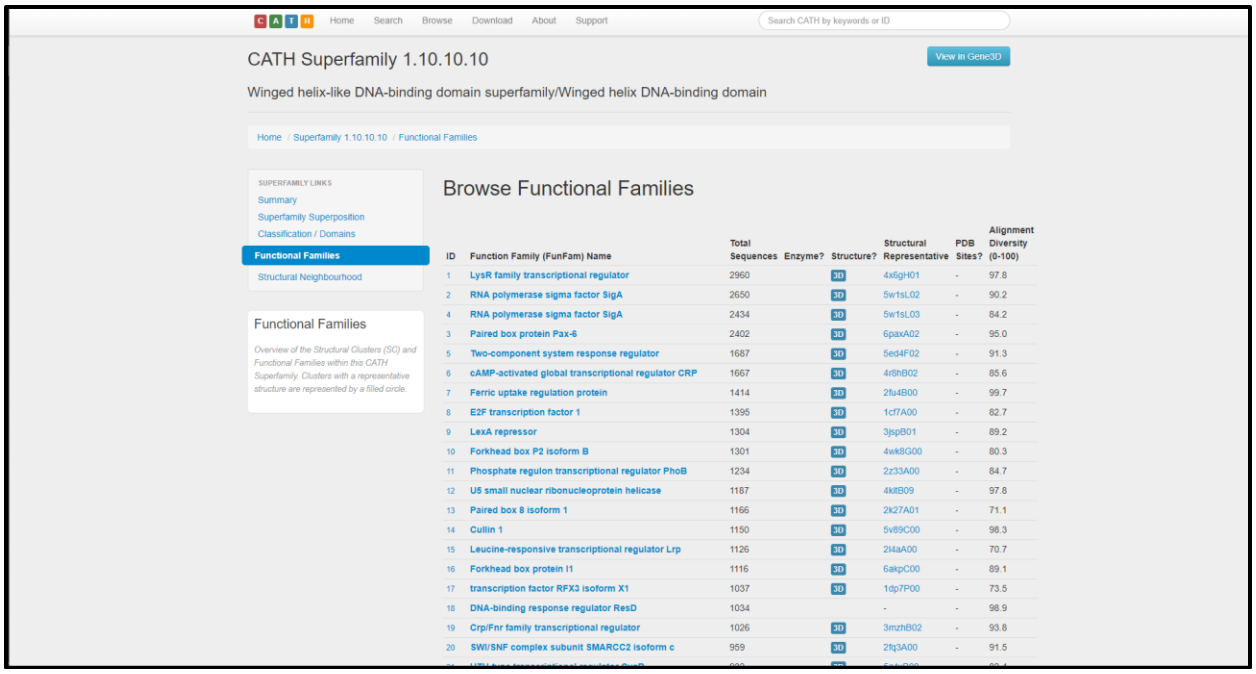

Fig3.5. Functional Families of CATH superfamilies for leucine

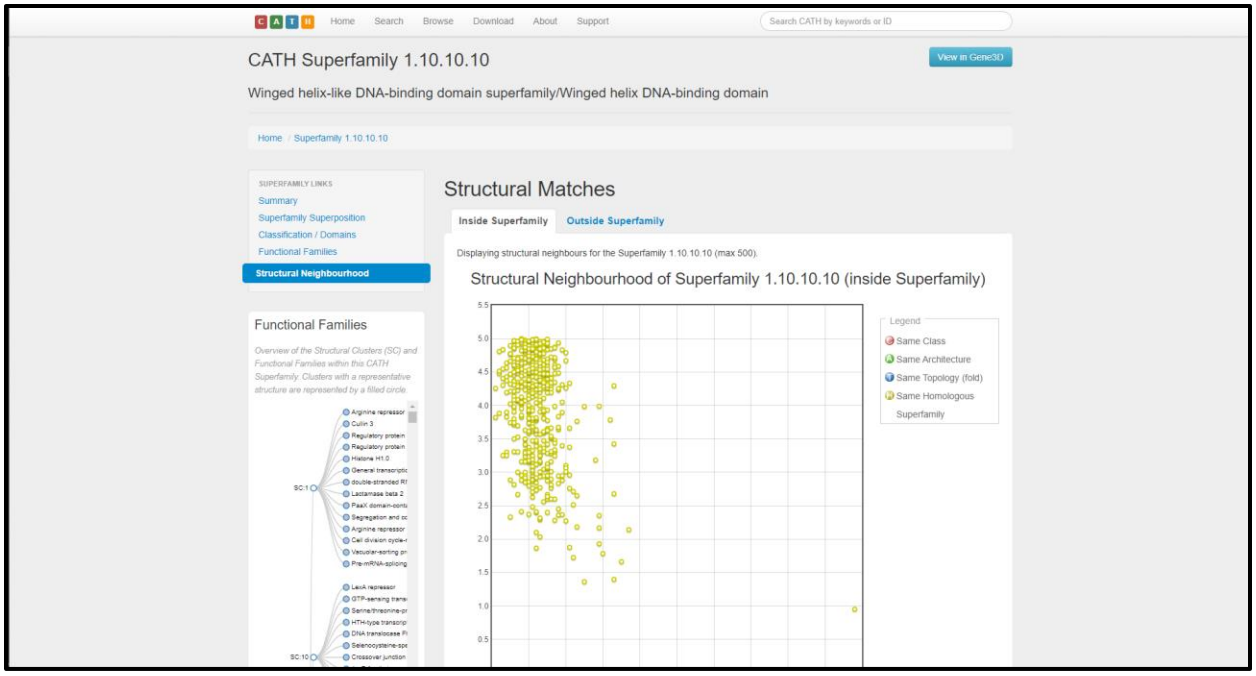

Fig3.6. Structural neighborhood of CATH superfamilies for leucine (Inside Superfamily)

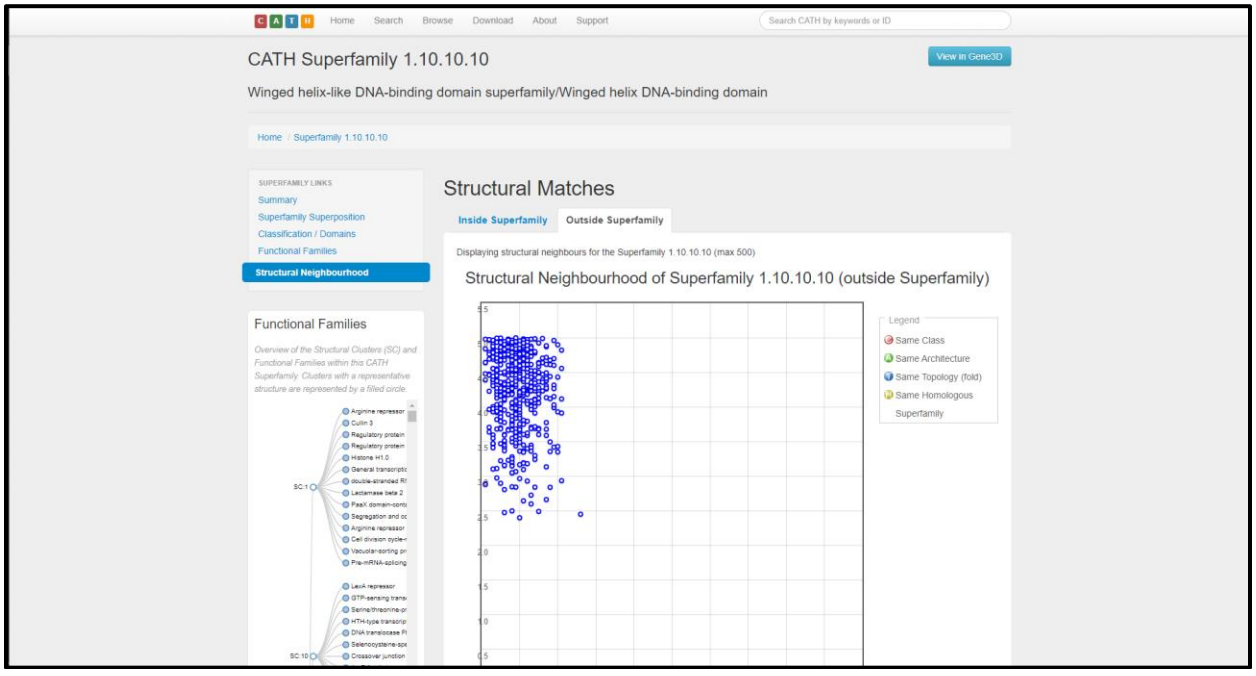

Fig3.7. Structural neighborhood of CATH superfamilies for leucine (Outside Superfamily)

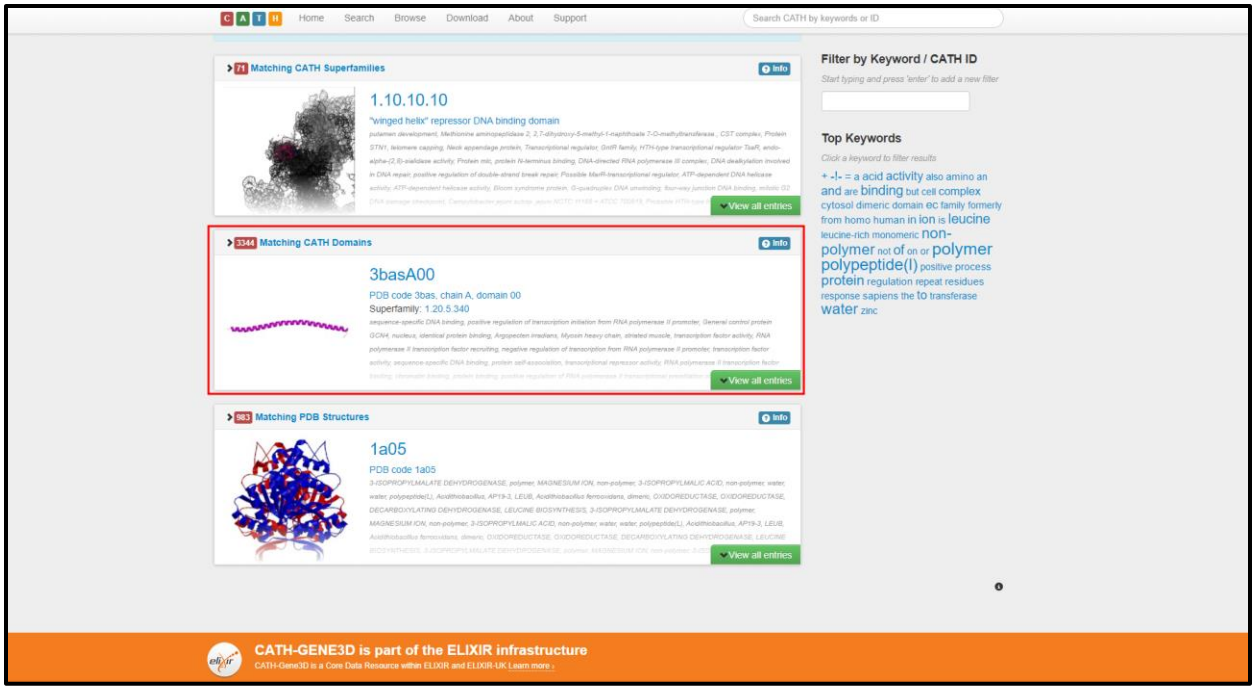

Fig4. Matching CATH domains for Leucine

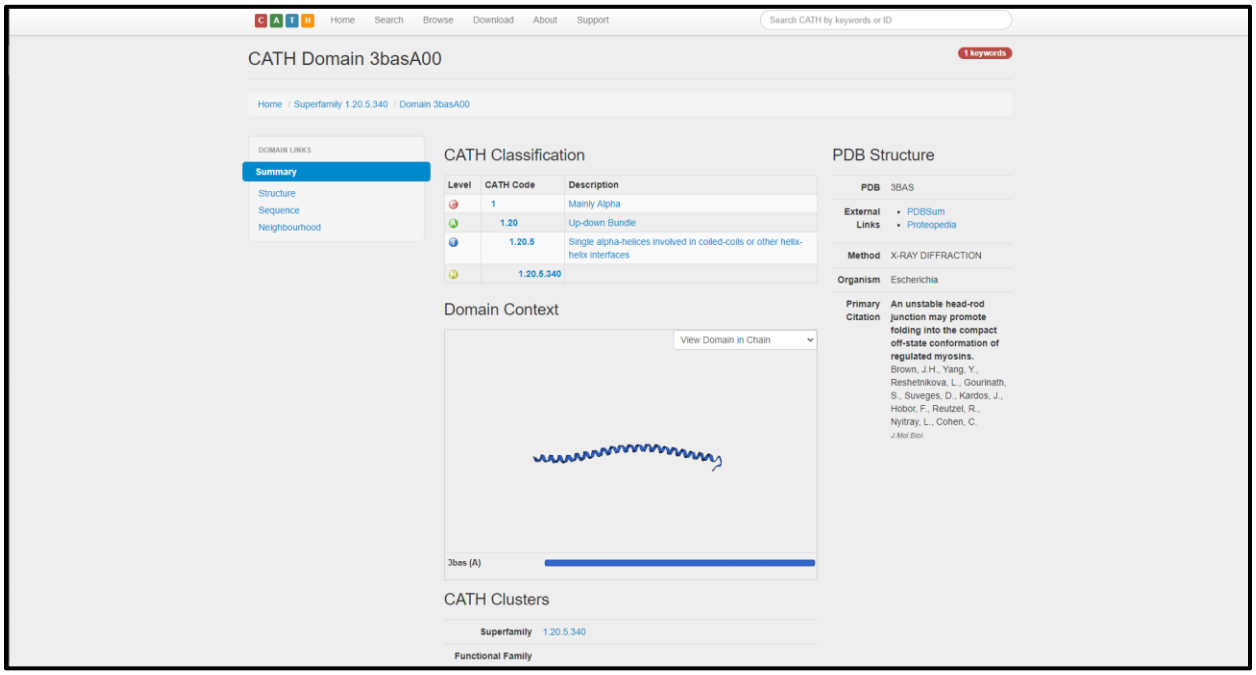

Fig4.1. Summary of matching CATH domains for Leucine

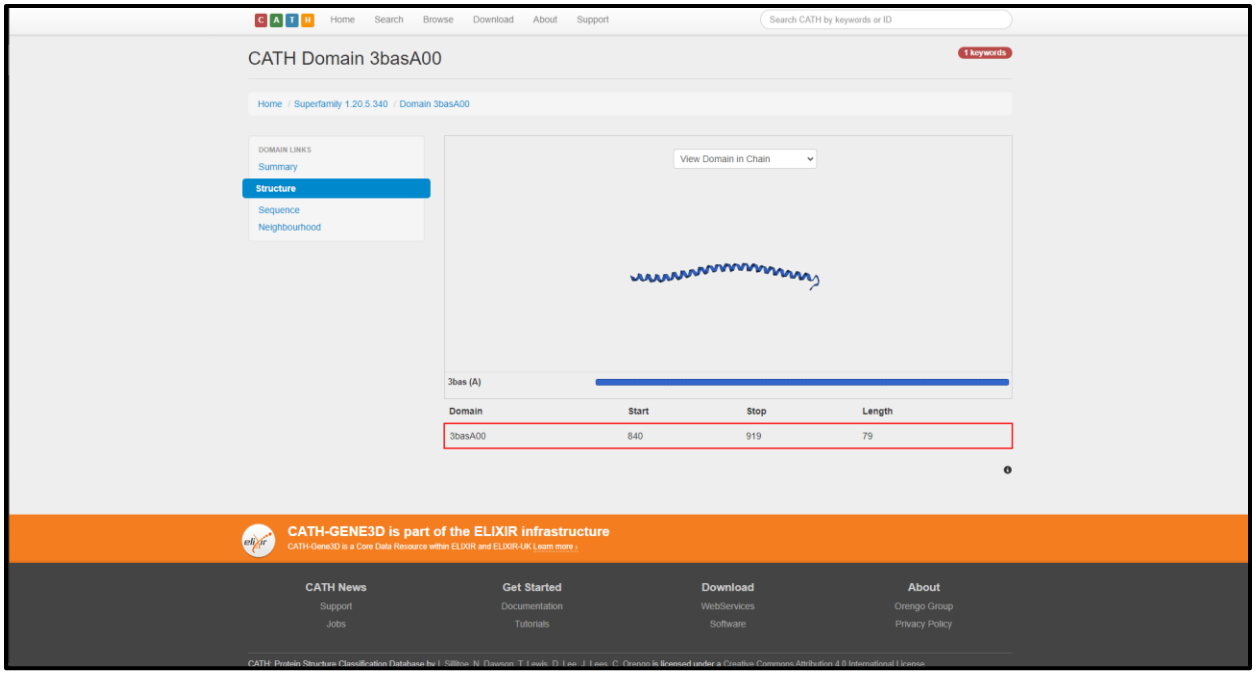

Fig4.2. Structure of matching CATH domains for Leucine

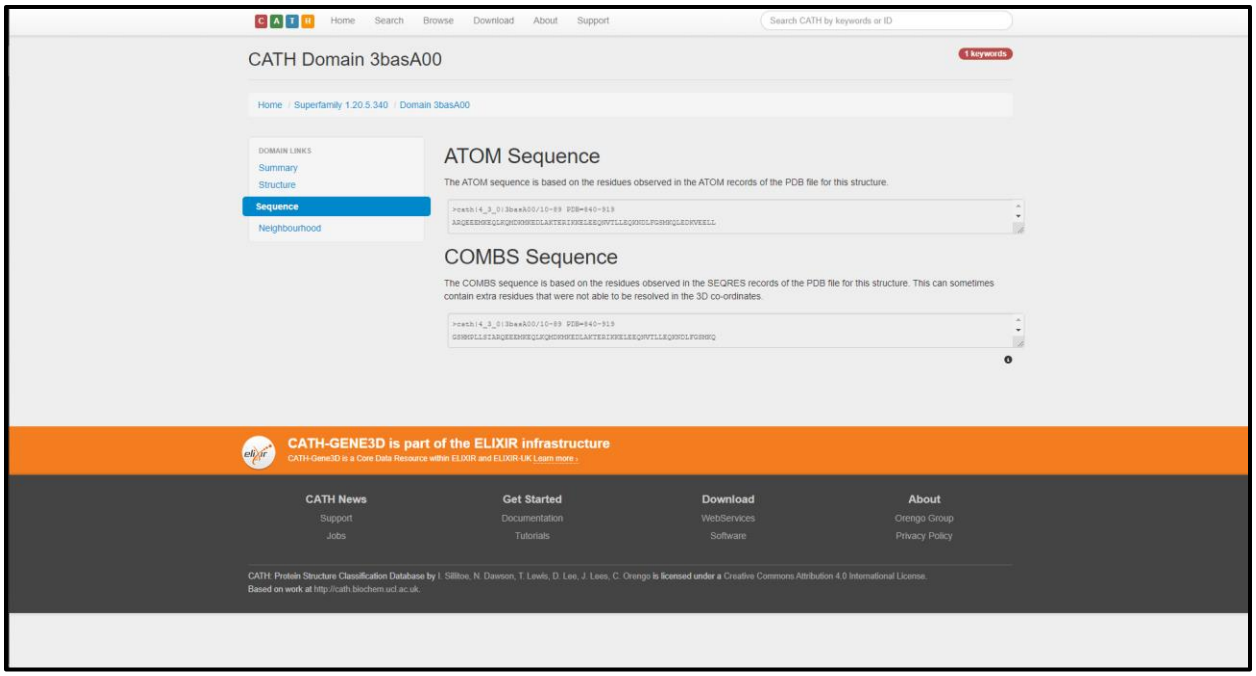

Fig4.3. ATOM and COMBS sequence of matching CATH domains for Leucine

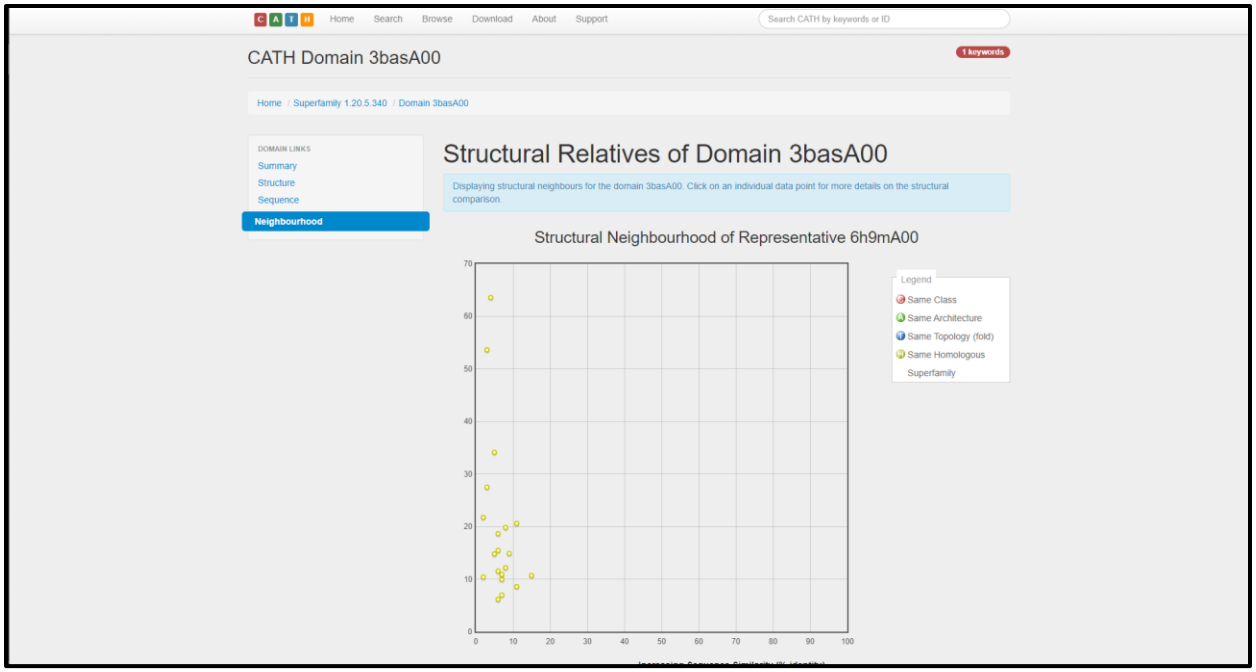

Fig4.4. Structural relatives of matching CATH domains for leucine

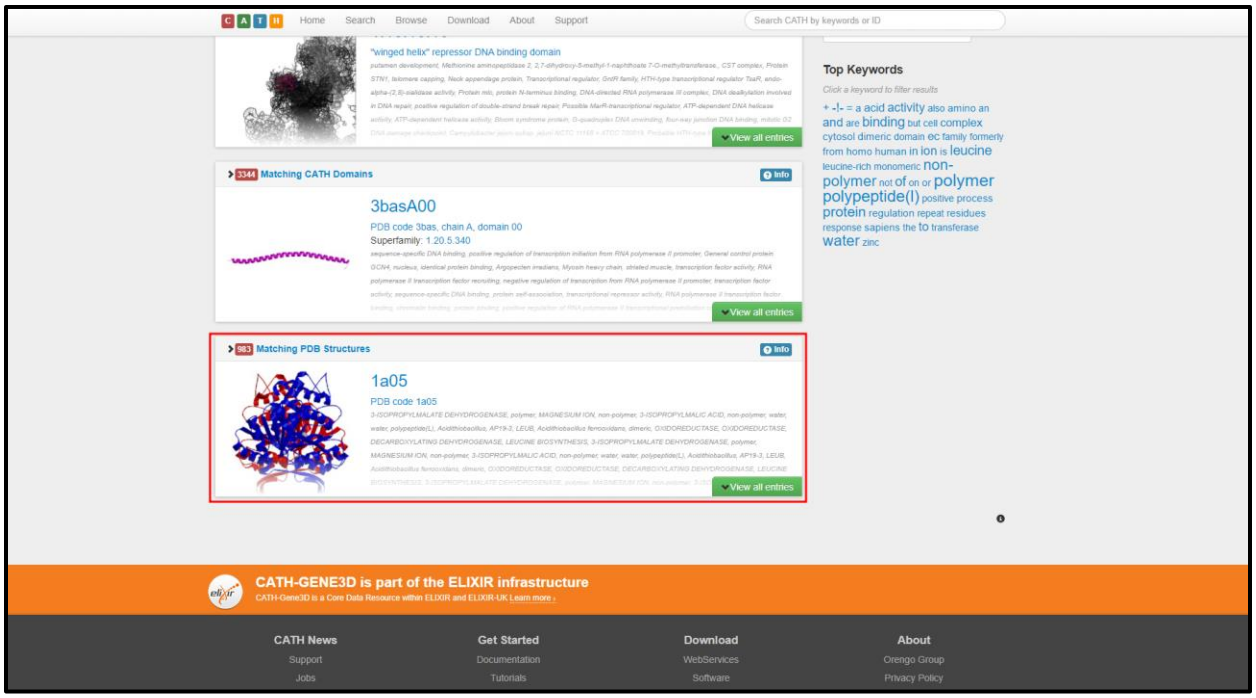

Fig5. Matching PDB structure for leucine

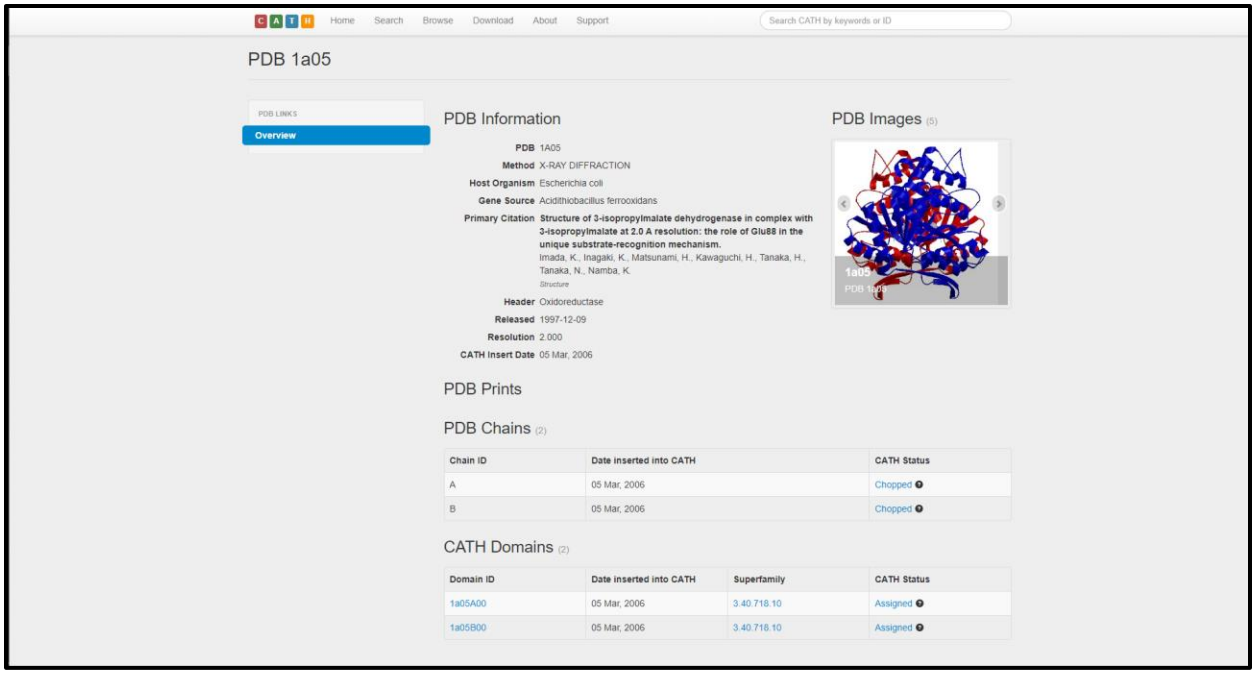

Fig5.1. Overview of matching PDB structures for leucine

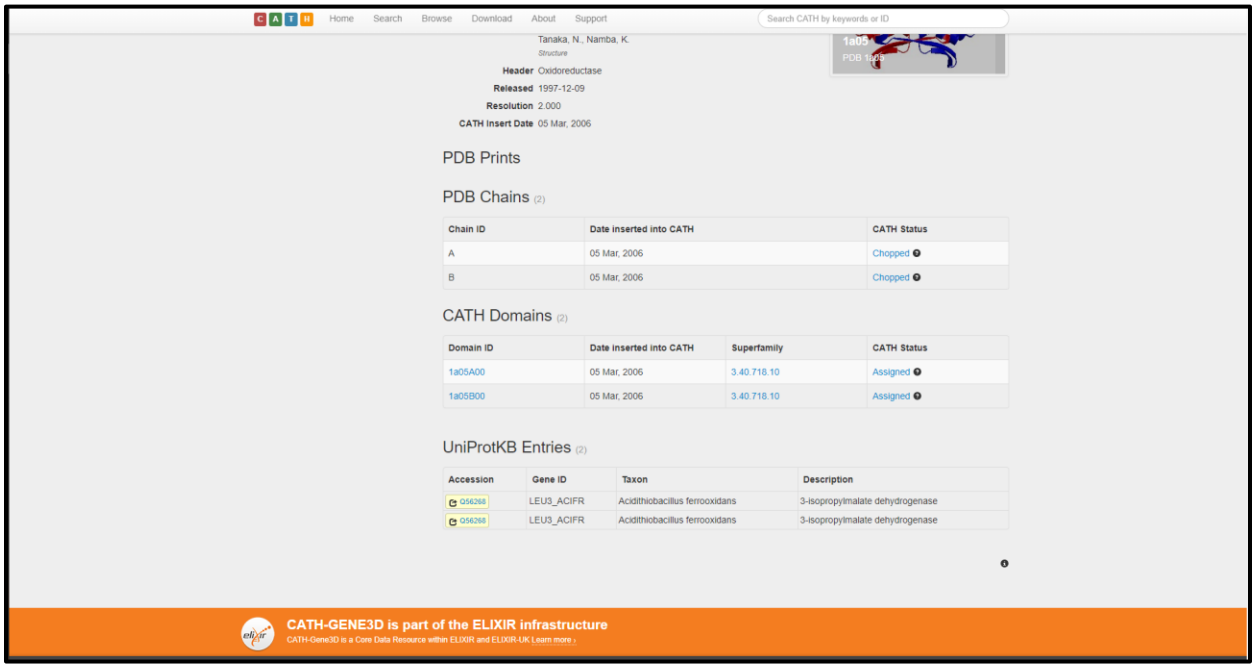

Fig5.2. Prints, Chains, CATH domains and UniProtKB entries for matching PDB structure for leucine

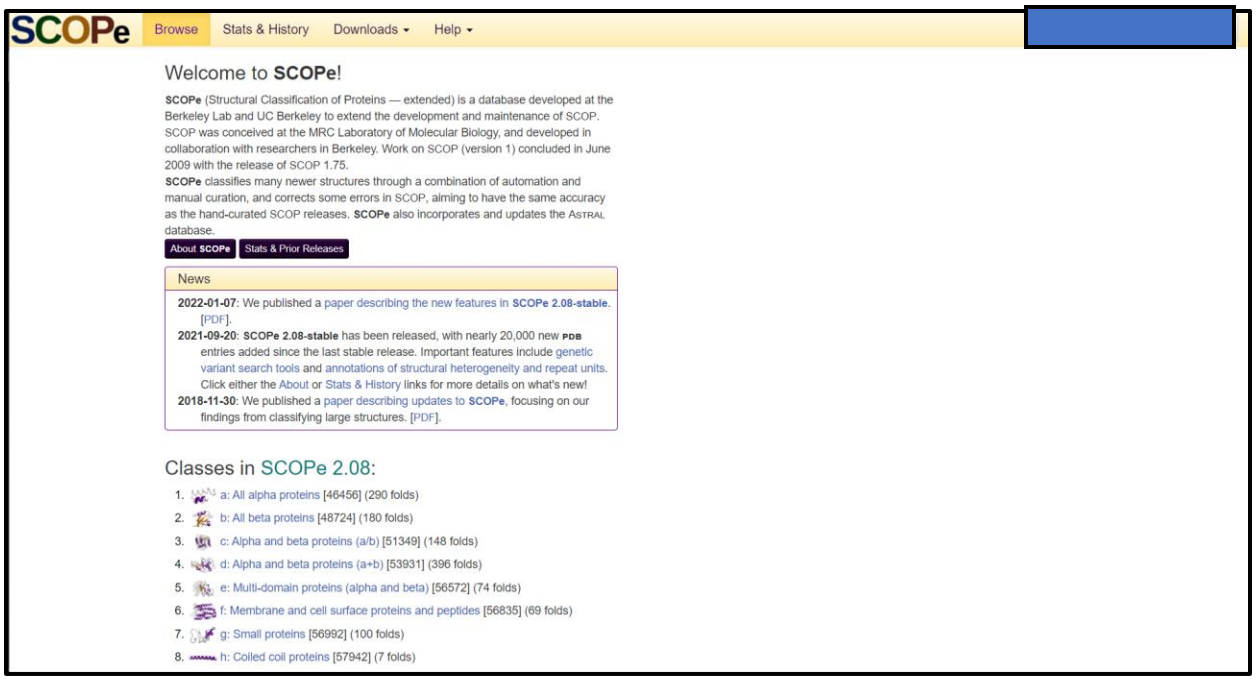

Fig6. Homepage of SCOPe Database with query leucine

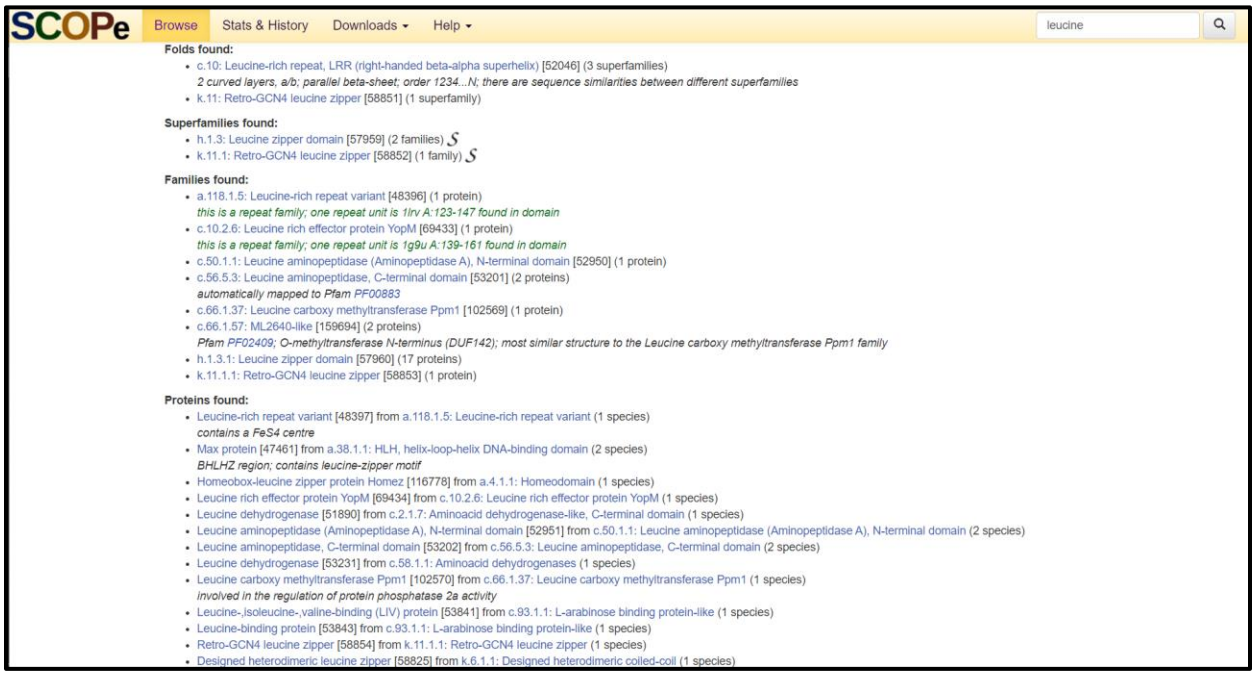

Fig7. Result page of SCOPe for query leucine

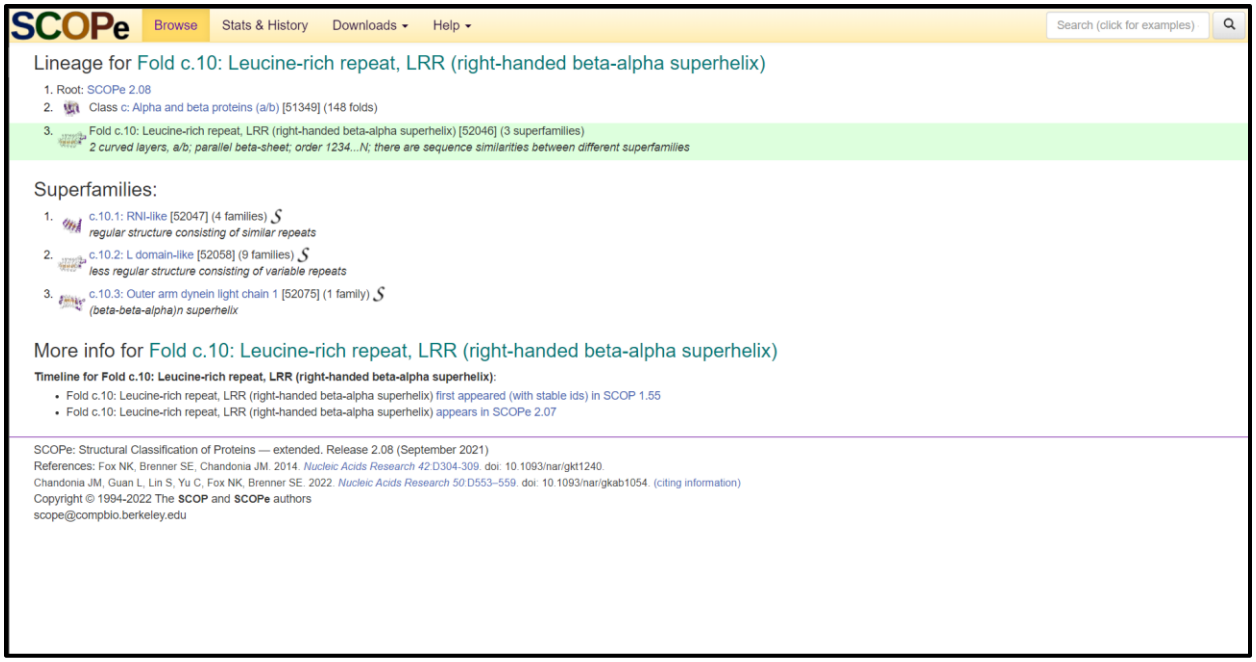

Fig8. Random result from "Folds found" of SCOPe database for leucine

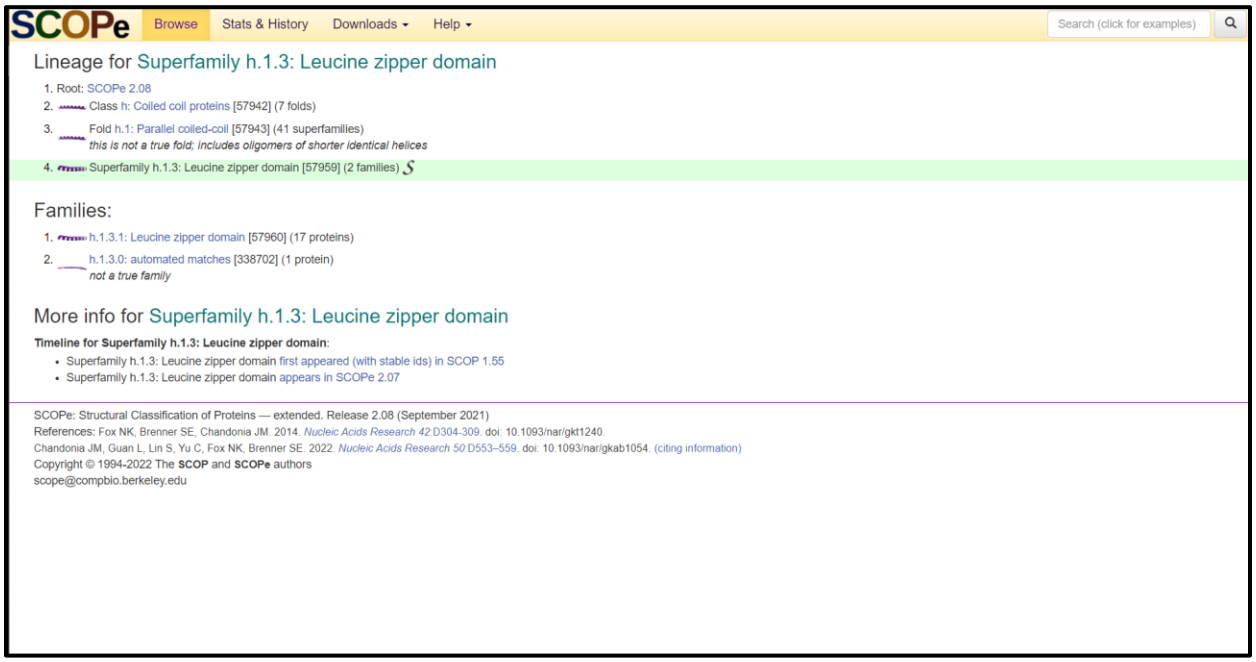

Fig8. Random result from "Superfamilies" of SCOPe database for leucine

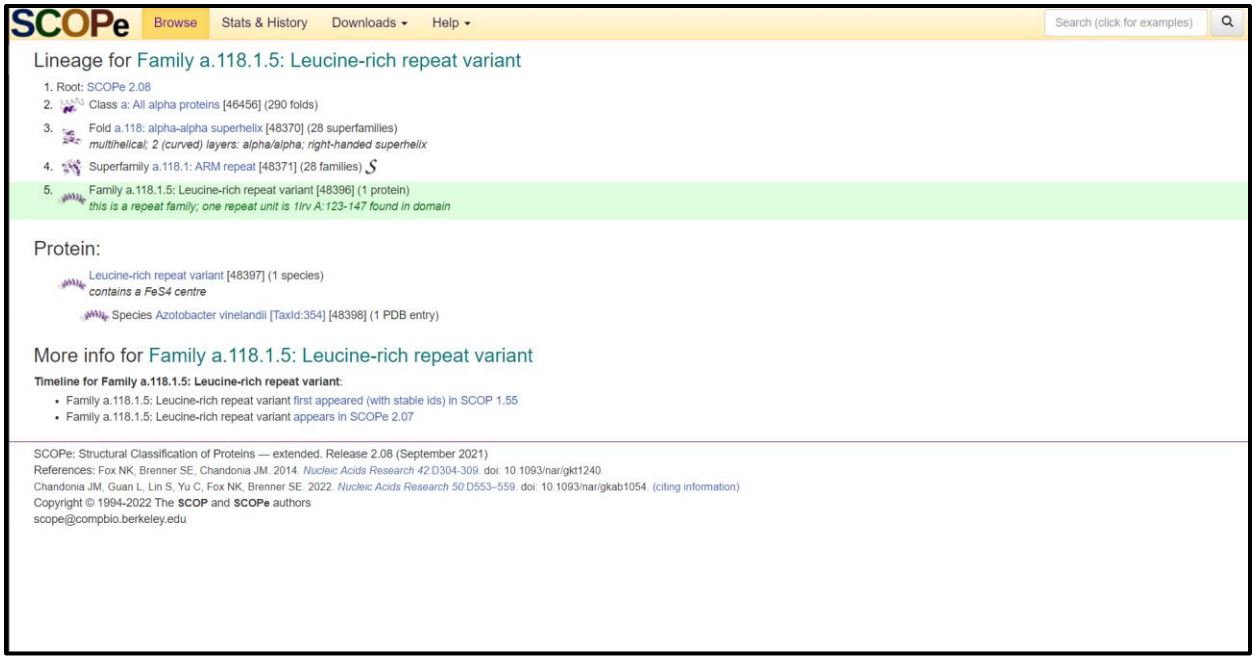

Fig9. Random result from "Families" of SCOPe database for leucine

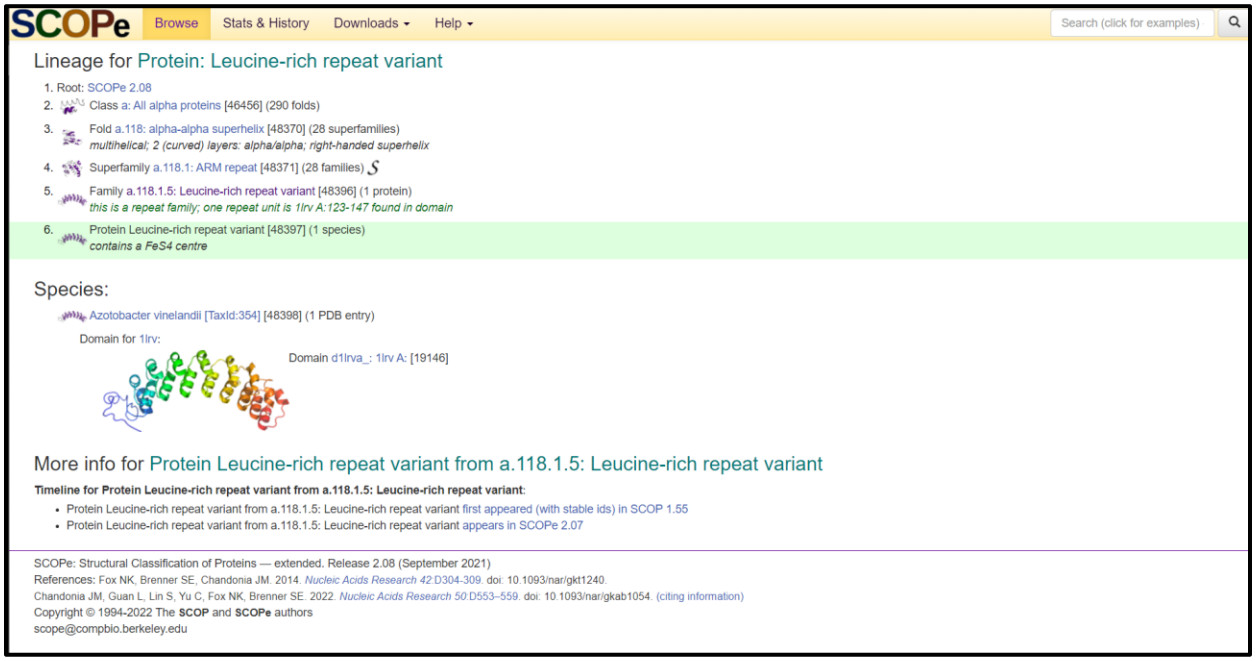

Fig10. Random result from "Proteins found" of SCOPe database for leucine

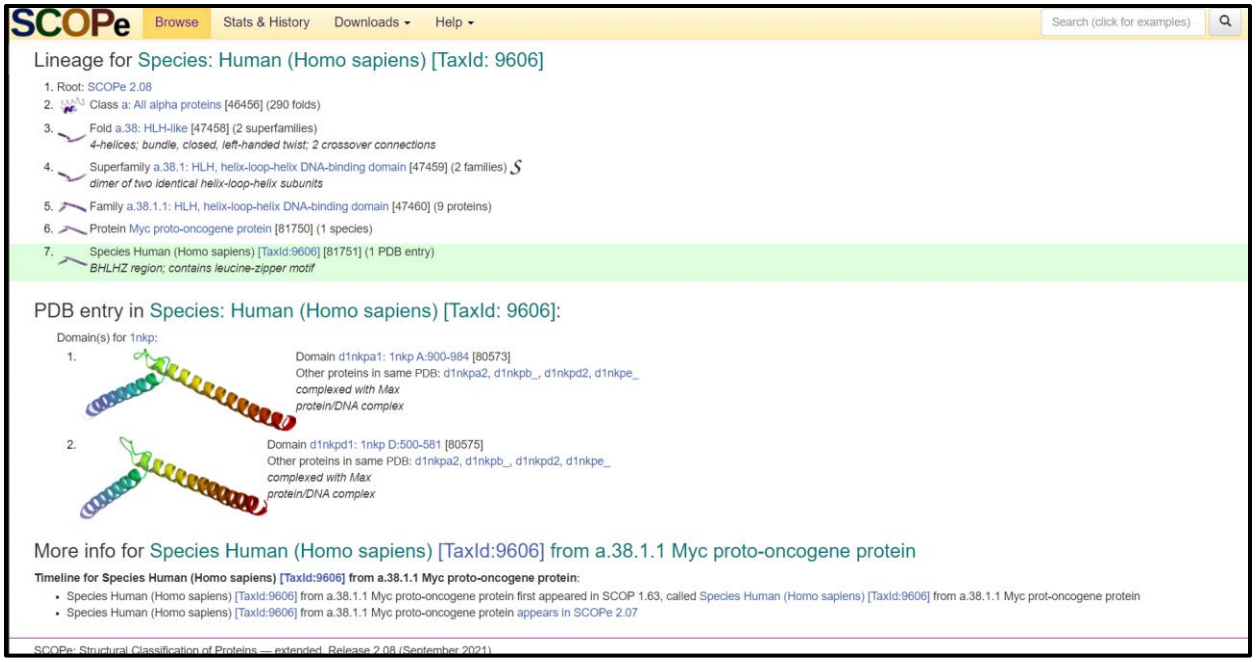

Fig11. Random result from "Species found" of SCOPe database for leucine

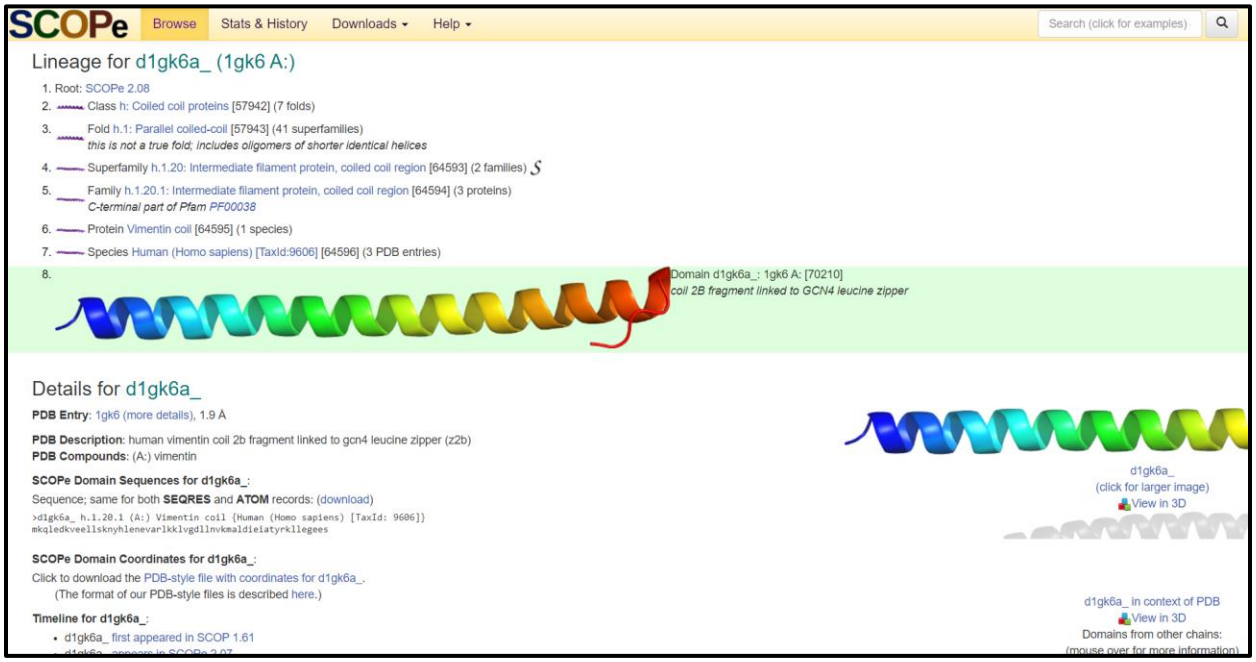

Fig11. Random result from "Domains found" of SCOPe database for leucine

Results:

CATH:

- ➔ In protein classification using CATH database for query leucine. It shows,
  - ➔ 71 matching CATH superfamilies.
  - ➔ 3344 matching CATH Domains
  - ➔ 983 matching PDB structures
- ➔ For CATH superfamilies we saw the Winged helix-like DNA-binding domain superfamily.
- ➔ For CATH Domains we saw CATH Domain 3basA00
- ➔ For PDB structures we saw PDB 1A05 from Escherichia Coli

SCOPe:

- ➔ In SCOPe database the result was divided into 6 sections:
  - ➔ Folds found [Fold c.10: Leucine-rich repeat, LRR (right-handed beta-alpha superhelix)]
  - ➔ Superfamilies found [Superfamily h.1.3: Leucine zipper domain]
  - ➔ Family found [Family a.118.1.5: Leucine-rich repeat variant]
  - ➔ Protein Found [Protein: Leucine-rich repeat variant]
  - ➔ Species found [Species: Human (Homo sapiens) [TaxId: 9606]]
  - ➔ Domains found [d1gk6a_ (1gk6 A:)]

Conclusions:

- ➔ The CATH database is valuable for biologists and bioinformaticians alike.
- ➔ For biologists with very specific tasks, browsing for individual domains is made easy by the user-friendly web interface.
- ➔ For bioinformaticians with a focus on large-scale analyses can find complete datasets available for downloading.
- ➔ Thus, working with CATH is remarkably uncomplicated.
- ➔ Updates are frequent, and, given the significant upcoming extension with horizontal layers complementary to the hierarchical structure, CATH is likely to become an even more valuable resource in the future.
- ➔ Since it was created, the development of SCOPe has been always guided by its user's feedback and needs.
- ➔ The automation in crystallography and advances in cryo-electron microscopy open a new era in structural biology and with it come new demands for data suitable for modelling of large proteins and protein complexes.
- ➔ In addition to a range of new annotations, SCOPe has introduced new functionalities that support relatively easy retrieval and assembly of independently determined, structurally characterized parts of proteins of interest.
- ➔ They will continue updating the database and providing regular releases while working on steadily increasing the coverage of structural data and adding new functionalities to the web interface.

References:

1. *Leucine*. Leucine - Health Encyclopedia - University of Rochester Medical Center. (n.d.). Retrieved February 27, 2022, from https://www.urmc.rochester.edu/encyclopedia/content.aspx?contenttypeid=19&contentid=Leucine#:~:text=Leucine%20is%20one%20of%20the,enough%20of%20these%20amino%20acids.
2. Berman, H. M. (2000). The Protein Data Bank. *Nucleic Acids Research*, *28*(1), 235–242. https://doi.org/10.1093/nar/28.1.235
3. Murzin, A. G. (1995). *Journal of Molecular Biology*, *247*(4), 536–540. https://doi.org/10.1006/jmbi.1995.0159
4. CATH Database. (2022b, February 21). CATH Database. Retrieved February 21, 2022, from https://www.cathdb.info/
5. *Search cath*. CATH Search: Browse. (n.d.). Retrieved February 26, 2022, from http://www.cathdb.info/search?q=Leucine
6. Cath superfamily 1.10.10.10. (n.d.). Retrieved February 26, 2022, from http://www.cathdb.info/version/latest/superfamily/1.10.10.10
7. Cath superfamily 1.10.10.10. (n.d.). Retrieved February 26, 2022, from http://www.cathdb.info/version/latest/superfamily/1.10.10.10/superposition
8. Cath superfamily 1.10.10.10. (n.d.). Retrieved February 26, 2022, from http://www.cathdb.info/version/latest/superfamily/1.10.10.10/classification
9. Browse cath-gene3d hierarchy. (n.d.). Retrieved February 26, 2022, from http://www.cathdb.info/browse/sunburst?from_cath_id=1
10. *Cath superfamily 1.10.10.10 - cathdb.info*. (n.d.). Retrieved February 26, 2022, from https://www.cathdb.info/version/latest/superfamily/1.10.10.10/alignments
11. Cath superfamily 1.10.10.10. (n.d.). Retrieved February 26, 2022, from http://www.cathdb.info/version/latest/superfamily/1.10.10.10/structure
12. *Cath domain 3basA00*. CATH. (n.d.). Retrieved February 26, 2022, from http://www.cathdb.info/version/latest/domain/3basA00
13. *Cath domain 3basA00*. CATH. (n.d.). Retrieved February 26, 2022, from http://www.cathdb.info/version/latest/domain/3basA00/structure
14. *Cath domain 3basA00*. CATH. (n.d.). Retrieved February 26, 2022, from http://www.cathdb.info/version/latest/domain/3basA00/neighbourhood
15. *Cath domain 3basA00*. CATH. (n.d.). Retrieved February 26, 2022, from http://www.cathdb.info/version/latest/domain/3basA00/sequence
16. PDB 1A05. (n.d.). Retrieved February 26, 2022, from http://www.cathdb.info/pdb/1a05
17. *Scop.berkeley.edu*. (n.d.). Retrieved February 26, 2022, from https://scop.berkeley.edu/search/?ver=2.08&key=leucine
18. *Scope 2.08: Domain d1gk6a_: 1GK6 A:* SCOPe. (n.d.). Retrieved February 26, 2022, from https://scop.berkeley.edu/sunid=70210
19. *Scope 2.08: Family A.118.1.5: Leucine-rich repeat variant*. SCOPe. (n.d.). Retrieved February 26, 2022, from https://scop.berkeley.edu/sunid=48396
20. *Scope 2.08: Fold c.10: Leucine-rich repeat, LRR (right-handed beta-alpha superhelix)*. SCOPe. (n.d.). Retrieved February 26, 2022, from https://scop.berkeley.edu/sunid=52046
21. *Scope 2.08: Protein: Leucine-rich repeat variant*. SCOPe. (n.d.). Retrieved February 26, 2022, from https://scop.berkeley.edu/sunid=48397

22. *Scope 2.08: Species: Human (homo sapiens) [taxid: 9606]*. SCOPe. (n.d.). Retrieved February 26, 2022, from https://scop.berkeley.edu/sunid=81751
23. *Scope 2.08: Superfamily H.1.3: Leucine zipper domain*. SCOPe. (n.d.). Retrieved February 26, 2022, from https://scop.berkeley.edu/sunid=57959
24. *Structural classification of proteins - extended. release 2.08 (September 2021)*. SCOPe. (n.d.). Retrieved February 26, 2022, from https://scop.berkeley.edu/