## **Exploring Levels of Protein Structure in the PDB**

**Authors**: Brian Gadd, Los Angeles Mission College, CA and Shuchismita Dutta, Rutgers University, NJ

**ASBMB Learning Objectives**

(<https://www.asbmb.org/education/core-concept-teaching-strategies/foundational-concepts/structure-function>)

1. Biological macromolecules are large and complex

* Students should be able to **describe the basic units of the macromolecules** and the types of linkages between them *(Introductory)*.

2. Structure is determined by several factors

* Students should be able to **recognize the repeating units in biological macromolecules** and be able to discuss the structural impacts of the covalent and noncovalent interactions involved *(Introductory)*.
* Students should be able to discuss the chemical and physical relationships between composition and structure of macromolecules *(Introductory)*.
* Students should be able to compare and contrast the primary, secondary, tertiary and quaternary structures of proteins and nucleic acids *(Intermediate)*.

### About Proteins

* Proteins are built by **covalently linking together a number of amino acids** in a specific order.
* **Protein synthesis** or peptide bond formation occurs in cells at the ribosome. Amino acids bound to their cognate tRNA molecules are transported to the ribosome and based on the order of amino acids defined by the mRNA, the amino acids are linked together to form proteins.
* Specific enzymes called **Proteases** can break the peptide bonds between amino acids. Some proteases act at specific sites in proteins (e.g., trypsin, chymotrypsin) while others remove amino acids from one end of the protein/peptide (e.g., carboxypeptidase).
* When the polymer is made of 2 to 25/30 amino acids, they are sometimes referred to as **peptides**, while longer polymers are called **proteins**. In general, proteins consist of one or more polypeptide chains.

Note: In some cases when two or more polymer chains interact together to perform a specific function, the assembly may also be referred to as a protein. (e.g., hemoglobin).

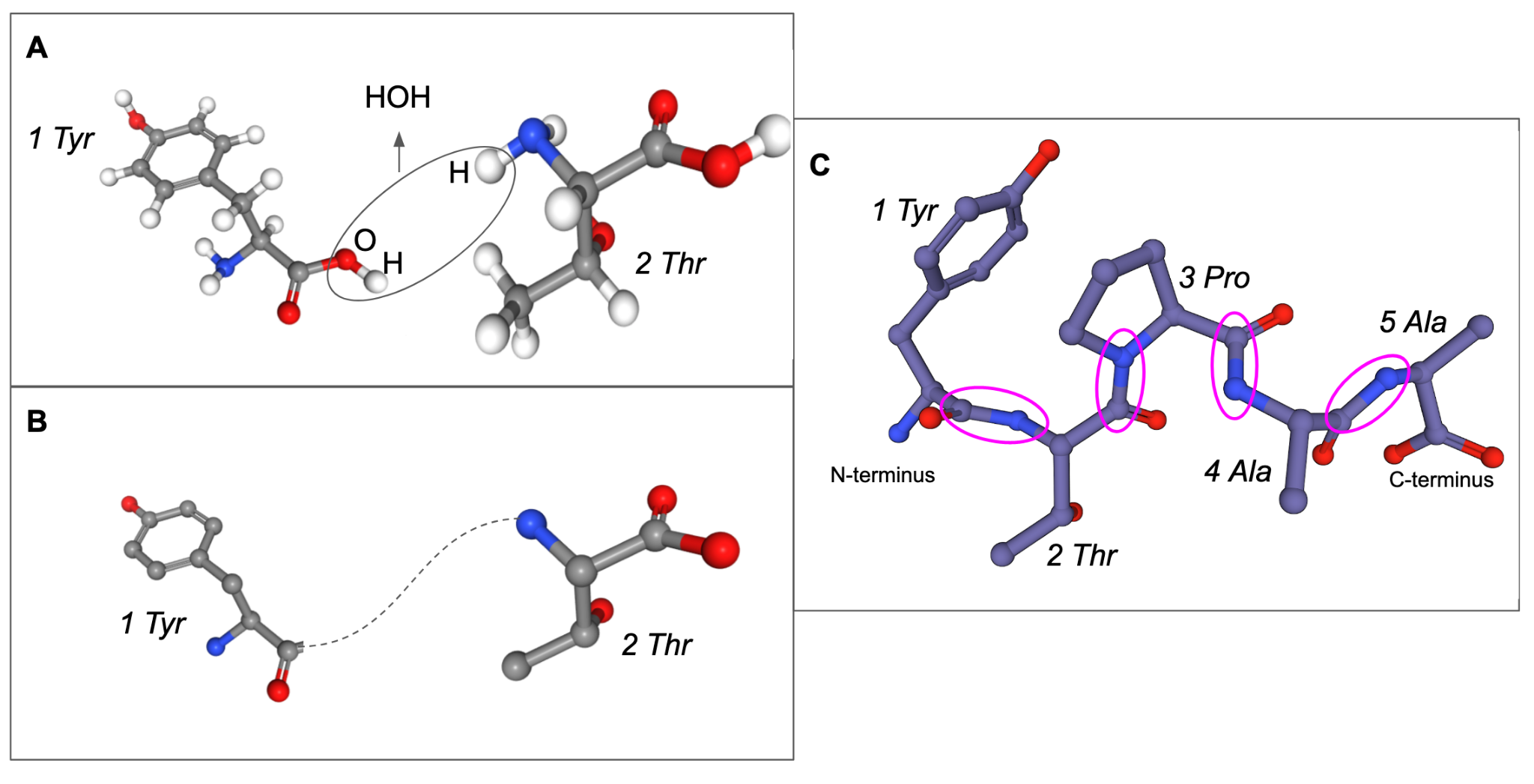
* The amino acid atoms in a polymer interact with each other in specific ways to form stable three dimensional (3D) shapes and assemblies that perform biological functions. Levels of protein structure (**primary, secondary, tertiary, and quaternary**), a few other terms (**motifs, domain**) commonly used in describing protein structure and function are included here. Guidelines for how to identify, explore, and use the knowledge of these levels of protein structure are also included here.

### Levels of Protein Structures:

#### Primary structure:

##### What is it?

* + This describes the order of amino acids that are covalently linked to each other to form the polymer. It is also referred to as the **protein sequence**. Learn [more about amino acids](https://cdn.rcsb.org/pdb101/teach/files/AminoAcids-in-PDB-091723.docx).
  + The covalent bond formed between amino acids as a result of a dehydration reaction is called the **peptide bond**. The OH group from the carboxyl end of the first amino acid (e.g., 1 Tyr) and one H from the amino end of the next amino acid (e.g., 2 Thr) form a water molecule that is released (see Figure 1A). The C=O group of the first amino acid (1 Tyr) and NH group of the next amino acid (2 Thr) are covalently linked together (see Figure 1B). This process is repeated for all other amino acids that form the polymer (see Figure 1C).



*Figure 1: Peptide bond formation: A. Two amino acids lose a water molecule; B. a covalent bond forms between the two amino acids (hydrogen atoms are not shown in this panel; C. the peptide bond formation is repeated to form a polymer where the peptide bonds are shown in magenta ovals.*

* + The primary structure or sequence of the polymer is often recorded as the one-letter abbreviation of amino acids, listed in the order of its appearance in the protein, from the N-terminus (beginning of protein) to C-terminus (end of protein), reading from left to right and called the FASTA format protein sequence. Learn [more about the FASTA format](https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=BlastHelp) of biomolecular sequences.

##### Where can you find it?

* + You can access the protein sequence for any polymer in any PDB structure by clicking on the Display or Download file options for the structure of interest. For example, in the case of the PDB entry 1trz, you can go to the structure summary page for this structure (<https://www.rcsb.org/structure/1TRZ>) and then view/download the sequence of each polymer as shown in (Figure 2a). The actual sequence can be accessed by selecting the FASTA sequence (Figure 2b).

|  |  |
| --- | --- |

*Figure 2: A. Display and Downloading protein sequences; B. FASTA format for the protein sequence on chain A in PDB entry 1trz.*

* + The protein sequence(s) available from the PDB is/are that of the protein(s) used in the experiment. If the PDB structure included only part of the complete protein or had additional amino acids added (e.g., for ease of purification or to increase stability or solubility of the protein fragment) the protein sequence will include these too. When available, the complete protein sequence can be accessed from UniProt (<https://www.uniprot.org/>).

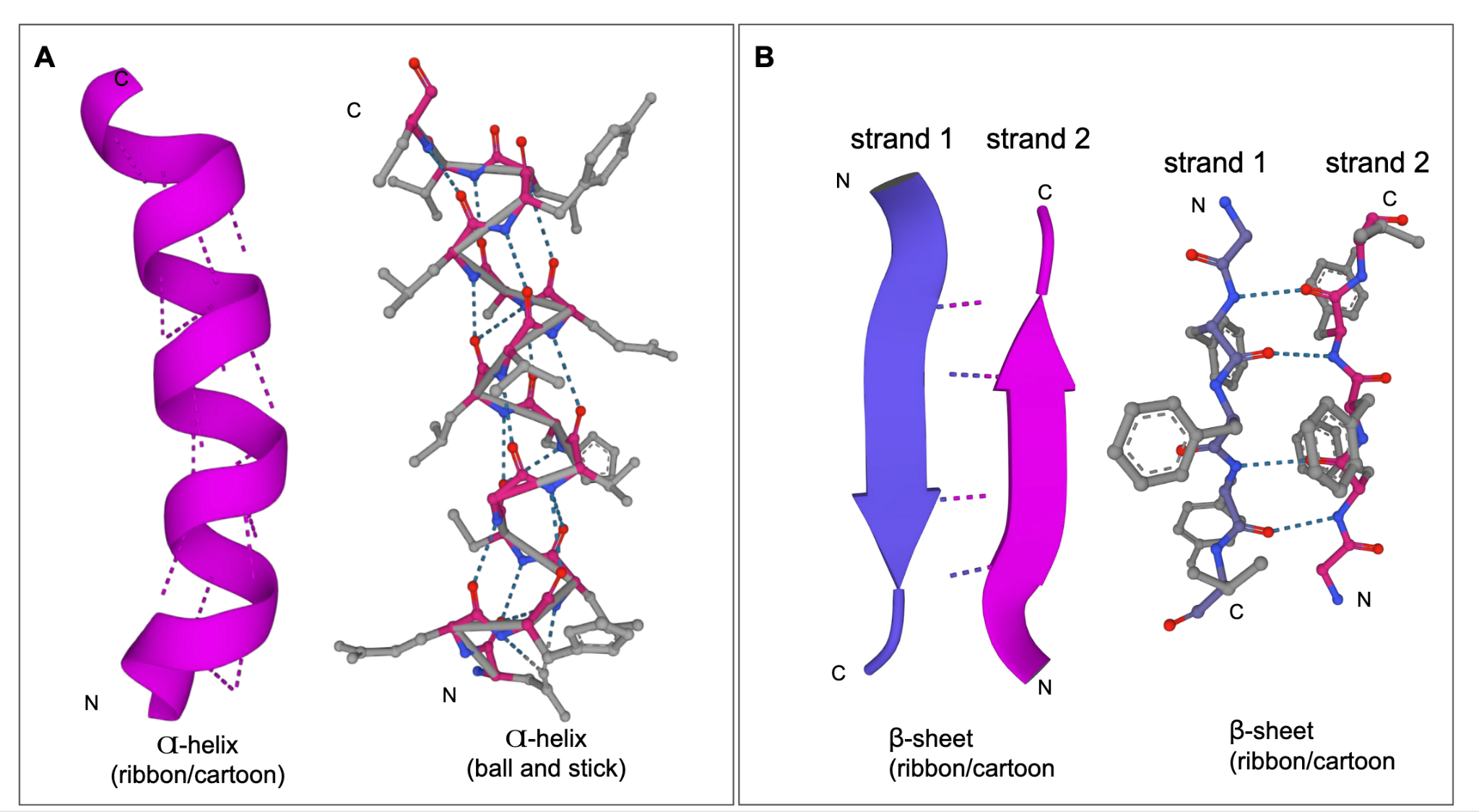
##### What can you do with this information?

* + You can use it to identify specific amino acids that play key roles in the structural stability or function of the protein.
  + You can use it to query the PDB archive to find other structures that have the same or similar sequences. These proteins may be related to each other through evolution.
  + You can map various chemical, structural, and functional properties of the protein onto the sequence and use this for analysis and comparison of various proteins.

#### Secondary structure:

##### What is it?

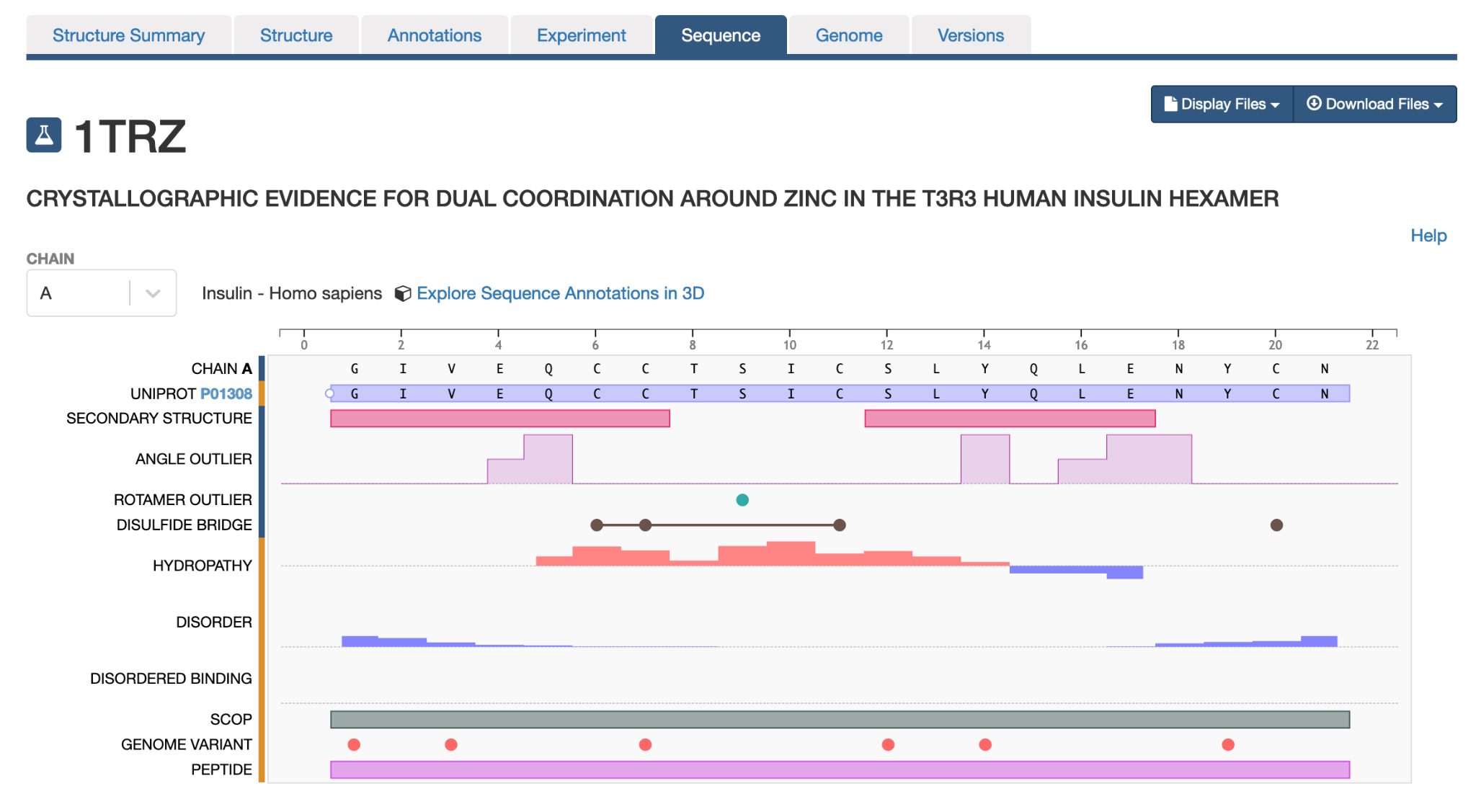
* + This level of protein structure is formed by folding the amino acid polymer so that the backbone atoms form local non-covalent interactions, e.g., hydrogen bonds formed between the C=O and NH groups in the protein backbone.
  + Regular hydrogen bonding patterns over stretches of amino acids in the protein yield helical structures (e.g., ɑ-helices, read alpha helices, Figure 3A), sheet like structures (e.g., ꞵ-sheets, read as beta-sheets, Figure 3B) or specific loops and turns (e.g., ꞵ-turns, read as beta-turns) (see Figure 3). Learn [more about the handedness and twist](https://cdn.rcsb.org/pdb101/teach/files/Handedness%20of%20Helices%20and%20Sheets.docx) in helices and sheets.



*Figure 3: Secondary structures in proteins shown in ribbons/cartoon representation or ball and stick. A. Hydrogen bonding (shown as dashed lines) patterns between protein backbone atoms in* ɑ*-helices. B. Hydrogen bonding (shown as dashed lines) patterns in a* ꞵ*-sheets (in PDB ID 1trz). The N- and C-termini are marked as N and C. Sidechain atoms are shown in gray, while backbone atoms are colored purple or magenta.*

##### Where can you find it?

* + You can visualize the 3D structures of helices and sheets/strands in Mol\* (e.g., <https://www.rcsb.org/3d-view/1TRZ>).
  + To see the secondary structure mapped on the protein sequence (primary structures) click on the sequence tab on the structure summary page for the structure of interest (Figure 4)
  + You can also view the secondary structural elements by sequence and its 3D structure side-by-side using the “Sequence Annotations” option from the structure summary page (e.g., <https://www.rcsb.org/3d-sequence/1TRZ?assemblyId=1>, see Figure 8)



*Figure 4: Sequence of chain A for the PDB entry 1trz showing the secondary structure mapping.*

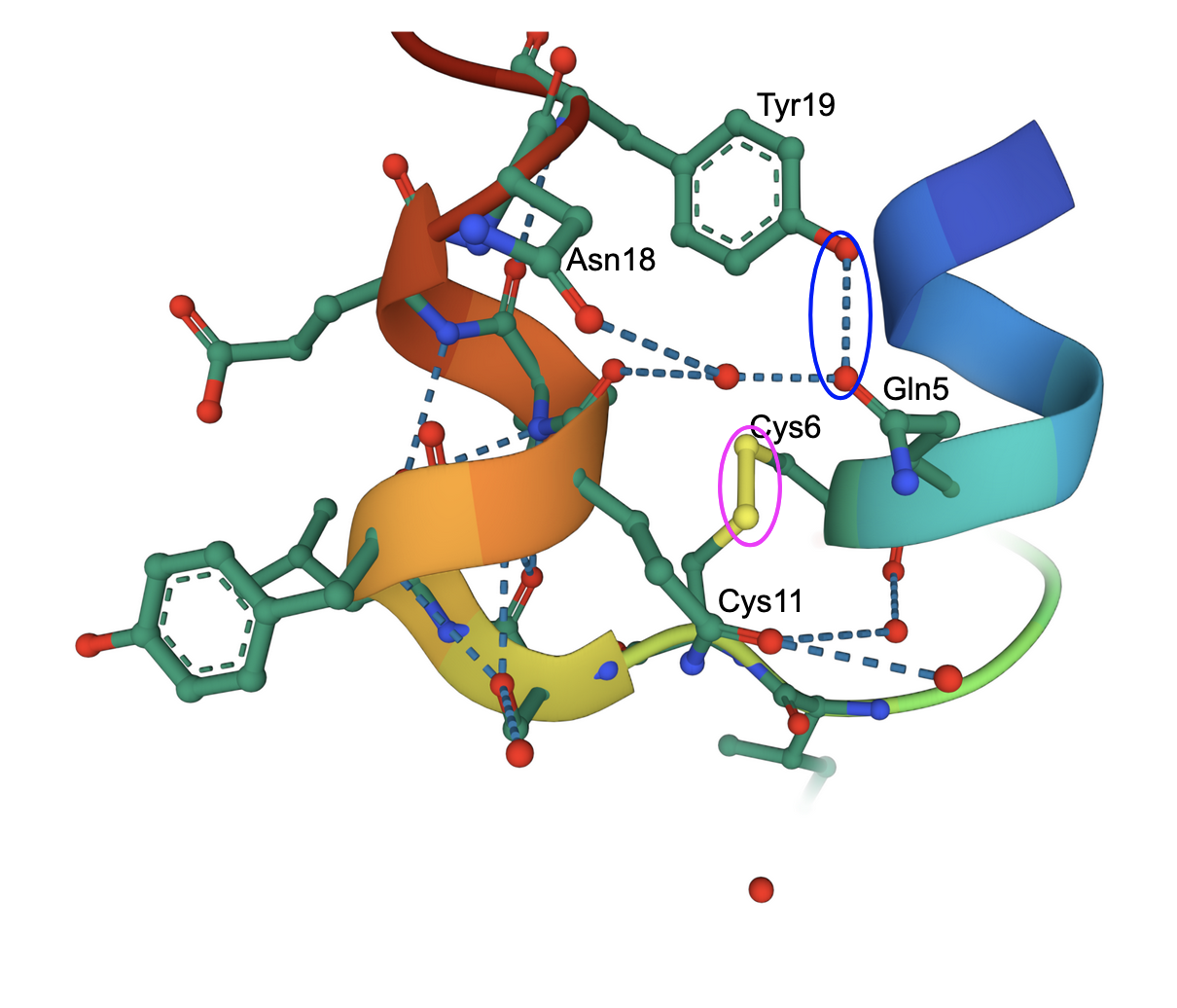
##### What can you do with this information?

* + Learning where the helices and sheets are in the structure can help with examining the interactions between these elements in the complete structure.
  + This may also be helpful in describing how some of these secondary structural elements participate in the function of the protein. For example, comparing secondary structural elements in proteins with and without interacting with ligands or partner proteins etc. may show how the structure changes in different contexts/conditions.

#### Tertiary structure:

##### What is it?

* + Tertiary structure is formed when several secondary structural elements interact with each other to form a stable structure.
  + This level of protein structure is stabilized by covalent and non-covalent interactions between the all amino acid atoms (backbone and side chains) in the polymer.
  + Types of covalent interactions may include disulfide bonds, while non-covalent interactions may include hydrogen bonds, hydrophobic interactions and more. These interactions may occur between amino acids that are far away in the primary sequence but located near each other in 3D space (see Figure 5).



*Figure 5: Tertiary structure of chain A in PDB entry 1trz showing a covalent bond (disulfide bridge, marked with a magenta oval) and some hydrogen bonds (e.g., between Gln5 and Tyr19, marked with a blue oval). The cartoon representation of the protein chain is colored by the rainbow color scheme (N-terminus is blue and C-terminus is red).*

* + Since several secondary structural elements come together to form tertiary structures, their relative positions and connectivity may provide unique insights into their interactions. Using these properties PDB structures have been organized and classified into groups by a variety of projects e.g., [CATH](http://www.cathdb.info/), [SCOPe](https://scop.berkeley.edu/), [ECOD](http://prodata.swmed.edu/ecod/).

##### Where can you find it?

* + To visualize the interactions in Mol\* (e.g., <https://www.rcsb.org/3d-view/1TRZ>) click on any amino acid in the structure of interest to zoom in, focus on it, and view the covalent/non-covalent interactions within its 5Å neighborhood.
  + The CATH, SCOP, ECOD etc. classifications for each PDB entry is available from its Annotations page (e.g., <https://www.rcsb.org/annotations/1trz>), and in the Sequence Annotations view (<https://www.rcsb.org/3d-sequence/1TRZ?assemblyId=1>).

##### What can you do with this information?

* + Exploring the tertiary structure shows how amino acids that are far apart in the primary structure may be located next to each other and may play key roles in the protein’s structural stability and function.
  + It provides insights into why the stability and functions of the protein of interest may change if closely located and interacting amino acids are changed (e.g., in natural variants or mutated proteins).
  + You can browse through the archive to explore, and compare related protein structures organized by the CATH, SCOP, etc. (<https://www.rcsb.org/search/browse>)
  + You can also search for other structures in the PDB with similar structures using the Structure similarity options in the Advanced Search Query Builder (e.g., [for macromolecule 1 (Chain A) in PDB entry 1trz](https://www.rcsb.org/search?request=%7B%22query%22%3A%7B%22type%22%3A%22group%22%2C%22nodes%22%3A%5B%7B%22type%22%3A%22terminal%22%2C%22service%22%3A%22structure%22%2C%22parameters%22%3A%7B%22value%22%3A%7B%22entry_id%22%3A%221TRZ%22%2C%22asym_id%22%3A%22A%22%7D%2C%22operator%22%3A%22strict_shape_match%22%7D%7D%5D%2C%22logical_operator%22%3A%22and%22%7D%2C%22return_type%22%3A%22polymer_entity%22%2C%22request_options%22%3A%7B%22paginate%22%3A%7B%22start%22%3A0%2C%22rows%22%3A25%7D%2C%22scoring_strategy%22%3A%22combined%22%2C%22sort%22%3A%5B%7B%22sort_by%22%3A%22score%22%2C%22direction%22%3A%22desc%22%7D%5D%7D%2C%22request_info%22%3A%7B%22query_id%22%3A%225de6f5d39c20358b20a15a4a1b8ff148%22%7D%7D))

#### Quaternary Structure:

##### What is it?

* + When multiple protein chains interact together to form an assembly it is called the quaternary structure.
  + The interactions between different polymers may be covalent (e.g., disulfide bonds) or non-covalent (e.g., hydrogen bonds),( see Figure 6).

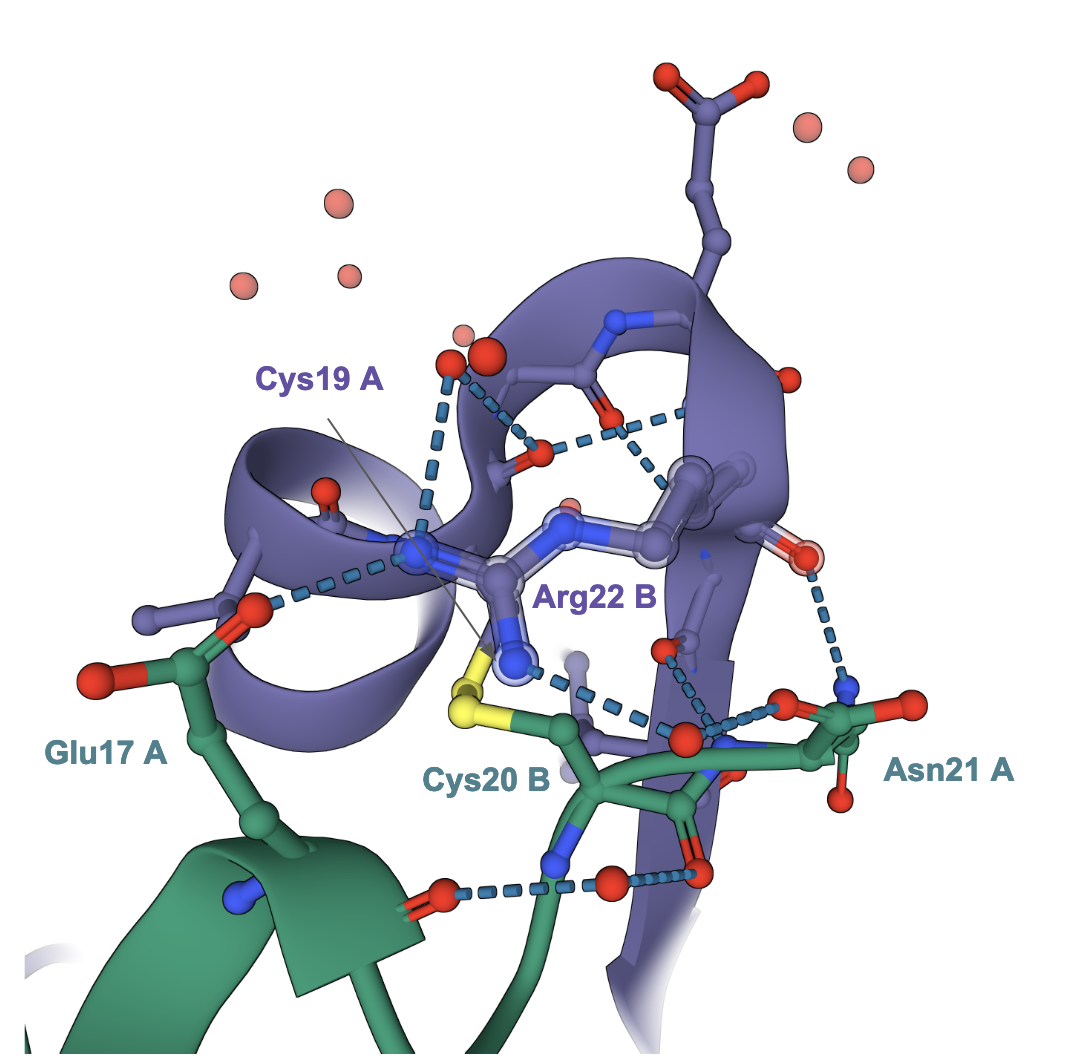
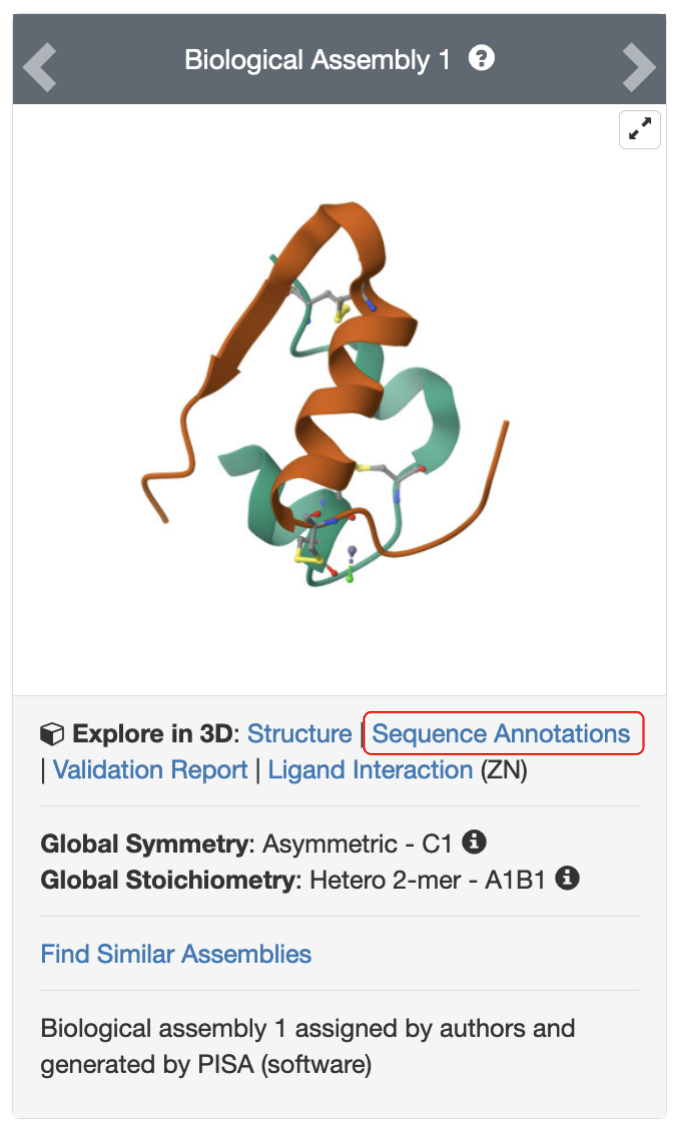


Figure 6: *Quaternary structure in PDB entry 1trz showing a covalent bond (disulfide bridge), and some hydrogen bonds between chains A (green) and B (violet).*

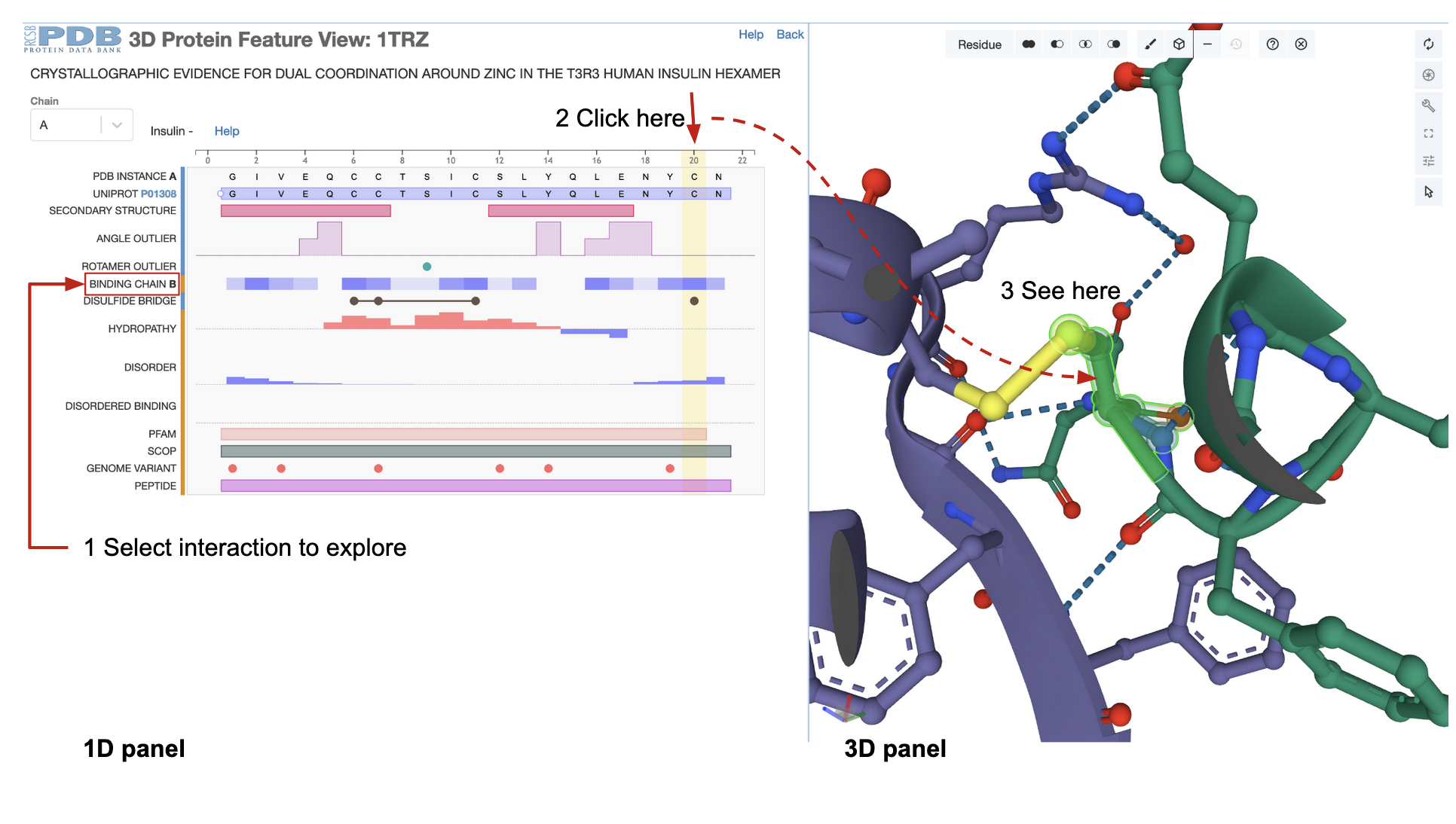
##### Where can you find it?

* + To visualize the interactions between polymer chains in Mol\* (e.g., <https://www.rcsb.org/3d-view/1TRZ>) click on any amino acid at the interface between polymer chains in the structure of interest to zoom in, and view the covalent/non-covalent interactions within its 5Å neighborhood.
  + You can also view the interactions between different polymers in a structure using the Sequence Annotations view. To access this view click on the link labeled Sequence Annotations below the snapshot of the structure on the structure summary page (see Figure 7).



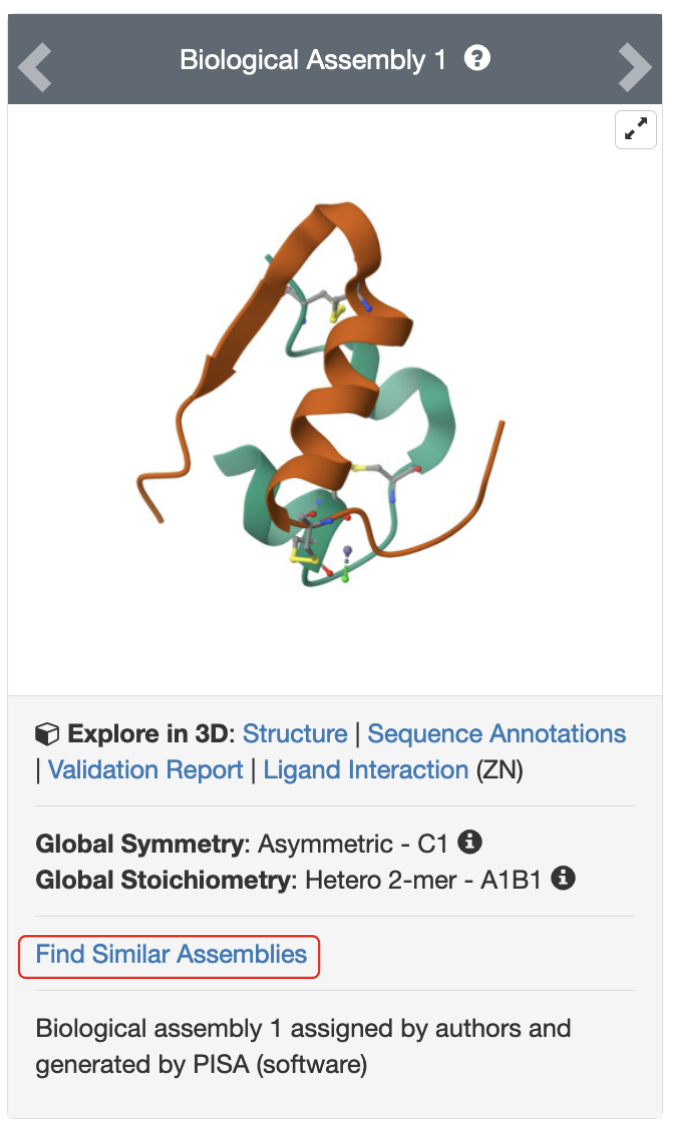
*Figure 7: Link to access Sequence Annotations of the PDB structure 1trz.*

* + Once the Sequence Annotations view opens, click on the residues of interest in the sequence panel to examine its location in the protein in the 3D structure panel (See Figure 8).



*Figure 8: Sequence Annotations view of the PDB entry 1trz showing quaternary interactions between the 2 polymer chains forming insulin.*

* + You can explore the PDB archive to see if the assembly has any local or global symmetry
  + You can explore the PDB archive to see if there are other structures with the similar assembly structure by clicking on the find similar assembly link (See Figure 9).



*Figure 9: Link to launch a search for PDB structures with similar assembly (quaternary) structure for one of the assemblies of the PDB structure 1trz.*

##### What can you do with this information?

* + Examining the interactions between proteins in an assembly can shed light on how the polymer chains in the protein interact with each other
  + This may help identify critical interactions that can be modified, or removed to alter the protein’s functions - e.g., in designing new therapeutic approaches and/or novel properties and functions in proteins.

### Other Protein Related Vocabulary:

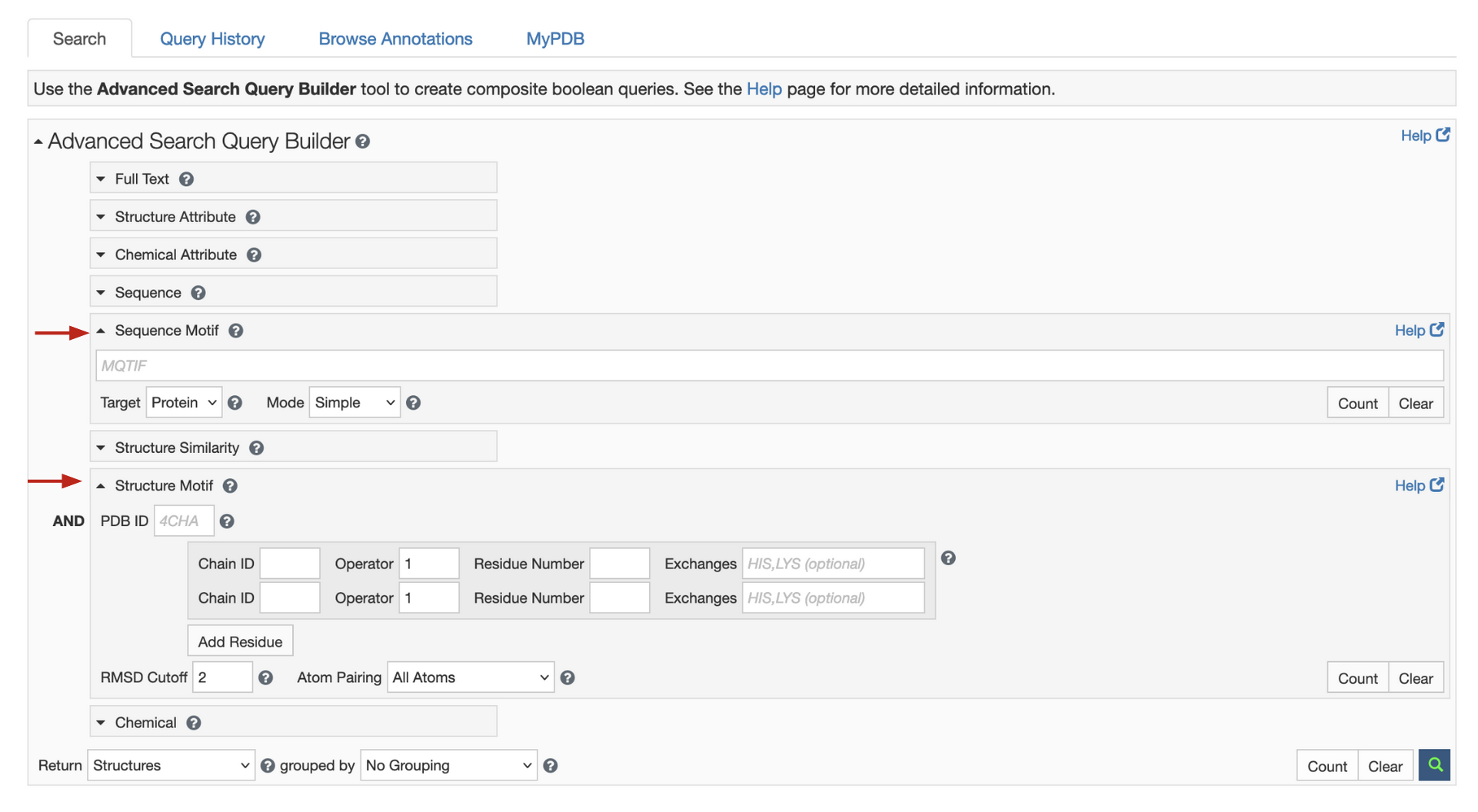
#### Protein Motifs:

##### What is it?

* + These short sequence or structure patterns are called sequence or structure motifs may be seen in many different proteins and may have the same or similar functions
  + Sequence motifs are short (<20 amino acids) stretches of sequence that may have a specific property or function (e.g., MQTIF).
  + Structure motifs have a conserved 3D structure, and defined by 3-10 amino acid locations in a structure (e.g., a C2H2, Zinc finger motif)

##### Where can you find it?

* + You can use the Advanced search query builder to specify the sequence or structure motif details (See Figure 10).



*Figure 10: Advanced search query builder options to search the PDB for sequence and structure motifs.*

##### What can you do with this information?

* + Identifying a specific sequence or structure motif in a structure indicates that it may have a specific function or property.
  + These motifs can provide insights into the functions of a protein and also its evolutionary history.

#### Protein Domains:

##### What is it?

* + Many large proteins are long and have regions of organized 3D structure separated by flexible and less structured regions.
  + The structured regions called domains can fold independently and may have specific functions.
  + Many examples of conserved domains have been identified in the scientific literature and even been classified by various resources (e.g., SCOP, CATH, Pfam, ECOD).

##### Where can you find it?

* + You can explore protein domains in the Annotations page of the structure of interest (e.g., <https://www.rcsb.org/annotations/1TRZ>).
  + You can visualize the domains using the Sequence Annotations view and clicking on the domains defined by CATH, SCOP etc. in the 1D panel to view it in the 3D panel (see Figure 11).

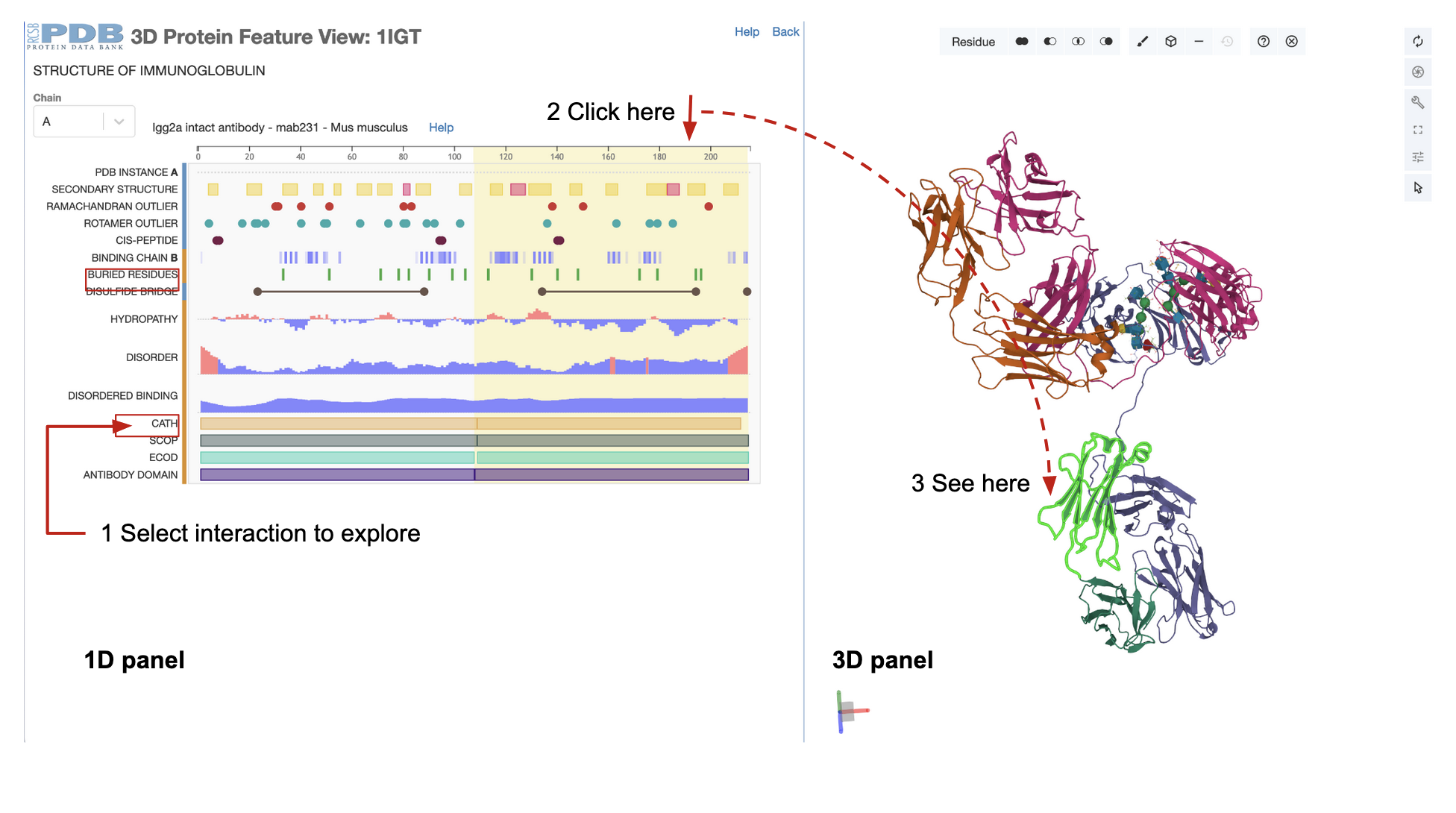


Figure 11: *Sequence Annotations view of the PDB entry 1igt showing (immunoglobulin) domains in the polymer chains forming immunoglobulin IgG2a.*

##### What can you do with this information?

* + Knowledge about the domains can help define functions of the entire protein.
  + This knowledge is often used in designing experiments - only the domains relevant to the function being studied are included in the experiment to make it easier to manipulate. e.g., single immunoglobulin domains may be included in an experiment instead of the entire protein).