## Protein Sequence (Primary Structure)

**Authors**: Alexandra Pettit, Carleton University, Ottawa, ON, Canada and Shuchismita Dutta, Rutgers University, NJ.

**Lesson Overview**:

This lesson reviews protein primary structure and introduces students to a series of bioinformatics tools that they can use to investigate and analyze proteins at this level of structure.

**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

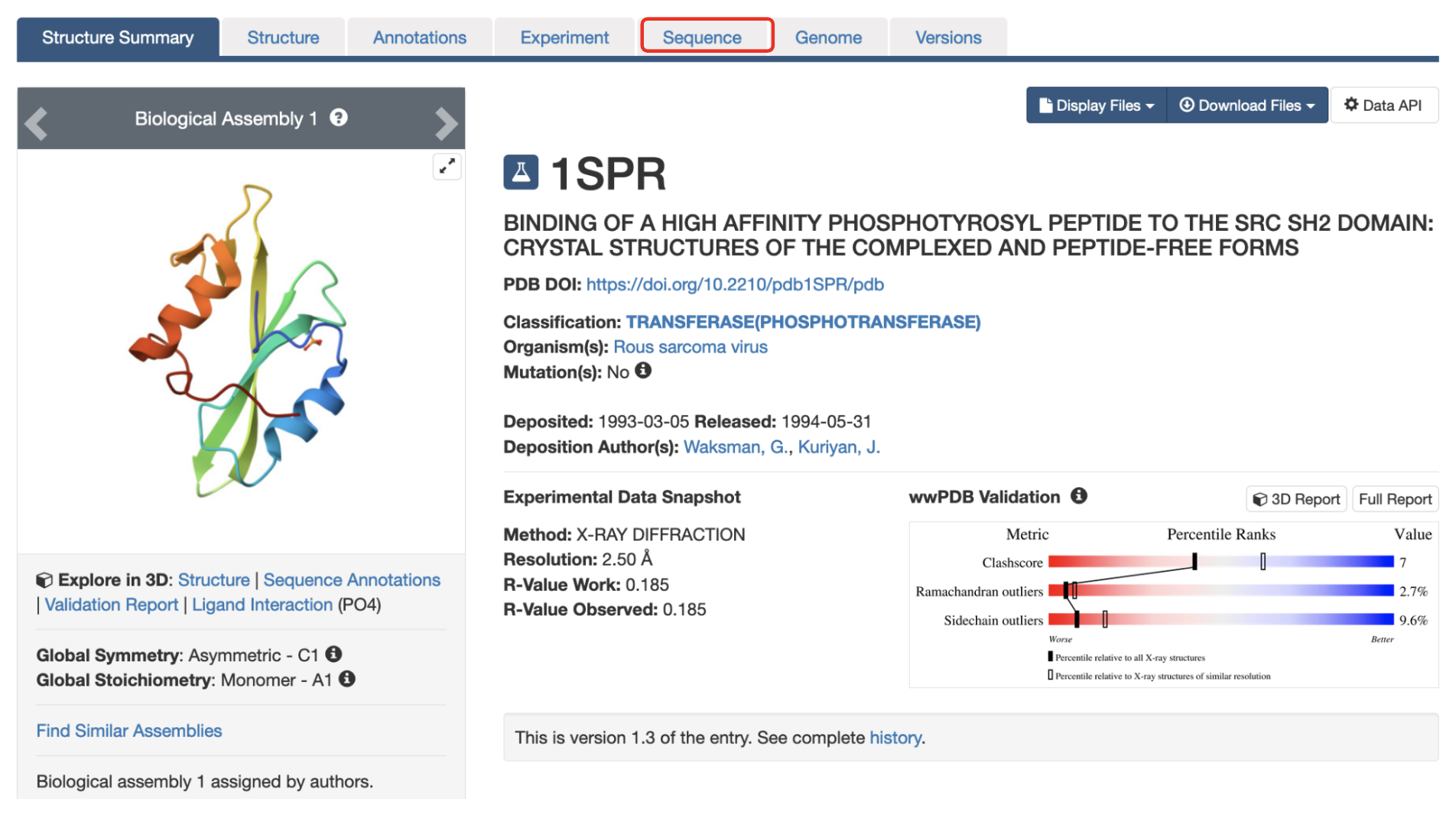
* Students should be able to **recognize the repeating units in biological macromolecules** and be able to discuss the structural impacts of the covalent and noncovalent interactions involved *(Introductory)*.
* Students should be able to **discuss the chemical and physical relationships between composition 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)*.
* Students should be able to use various **bioinformatics approaches to analyze macromolecular primary sequence and structure**. (*Introductory*)

**Activity**:

Primary structure refers to 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**. In this lesson, we will explore the primary sequence of a specific domain of a common tyrosine kinase (called SH2 domain).

### **Part I: Exploring Protein Sequences on the RCSB PDB website.**

Go to the [RCSB PDB home page](https://www.rcsb.org/) and enter the PDB code **1spr** in the top search box and click on it to open the Structure Summary page (figure 1 shows you the top of this page) for this PDB structure or go to the page (<https://www.rcsb.org/structure/1SPR>). The PDB entry you will be examining contains the crystal structure for an Src tyrosine kinase SH2 domain solved at 2.50 Å.



*Figure 1: Top portion of the RCSB PDB structure summary page for* [*1spr*](https://www.rcsb.org/structure/1SPR)*, the Src SH2 tyrosine kinase domain.*

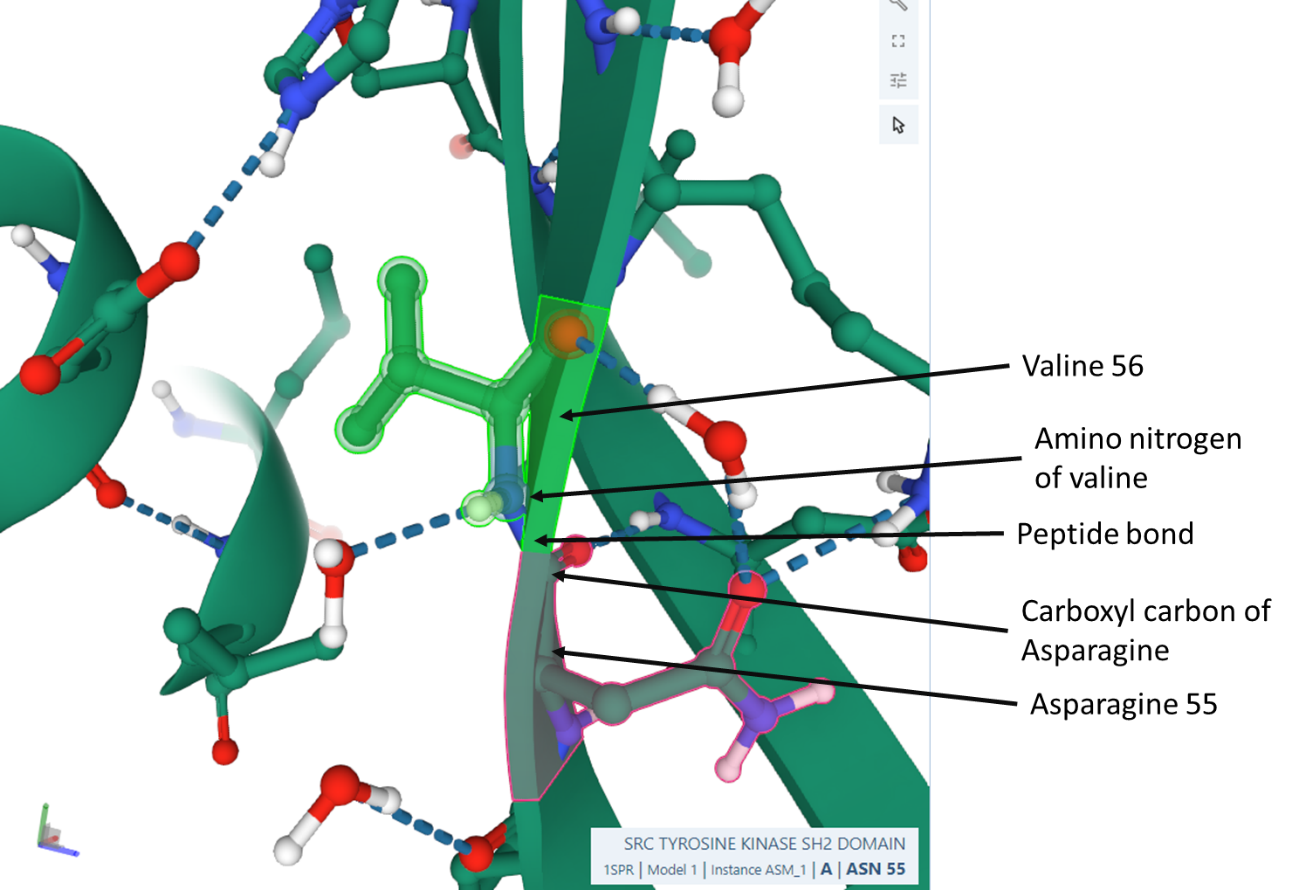
Scroll down to the “Macromolecules” section of the structure summary page.

1. Using the information provided about the ‘SRC TYROSINE KINASE SH2 DOMAIN’ answer the following questions:
   1. In what organism did this protein originate? \_\_\_\_\_
   2. How many amino acids are found in the modeled protein? \_\_\_\_\_\_
   3. Does the sequence of the modeled protein contain any mutations? \_\_\_\_
   4. What UniProt ID is this protein associated with? \_\_\_\_

Within the “Macromolecules” section of the structure summary page there is a “Protein Feature View” which lists out the primary sequence for the modeled protein, comparing it to the associated UniProt entry.

The graphic below the sequence in this view also shows a limited number of features related to the protein’s primary sequence. Additional features can be seen by clicking on the **Expand** button (in line with the protein feature view title) or the **Sequence** tab at the top of the structure summary page (see the red boxed area in figure 1). If you click on the ‘view features in 3D’ you will also open a panel that maps these sequence features onto the solved 3D model. Alternatively, you can also click on the “Sequence Annotations” below the image of the structure on the structure summary page (or go to <https://www.rcsb.org/3d-sequence/1SPR?assemblyId=1>).

1. Click on any amino acid in the primary sequence, this will highlight its features in the sequence view and zoom into its location within the 3D protein structure. Rotate the molecule so that the protein backbone is visible. Take a screenshot of the protein which shows the peptide bond between your chosen amino acid and one of its neighbours. Annotate your image, labeling both amino acids (location and identity), the bond itself, and the specific atoms which participate in the peptide bond. Use your favorite graphics manipulation software (e.g., PowerPoint, Photoshop) for adding labels. Paste your annotated screenshot below. A sample solution shown below.



Create a similar figure for any other amino acid of your choice and paste it here.

Answer:

1. Consult the primary sequence for 1spr and use it to answer the following questions:
   1. Name the amino acids (1 and 3 letter code) found in the following positions:

(Hint: This can be determined by clicking on the matching residue at the appropriate position in the sequence, using the number line for reference.)

N-terminus \_\_\_\_\_\_; 10 - \_\_\_\_\_\_\_ ; 45 - \_\_\_\_\_\_\_; 83 - \_\_\_\_\_\_\_; 104 - \_\_\_\_\_\_\_; C-terminus \_\_\_\_\_\_\_\_.

* 1. Does this protein contain any proline residues? If so, how many and in what position(s)?

Answer:

* 1. Referring to the hydropathy plot provided, what position(s) contain hydrophobic regions in this protein:

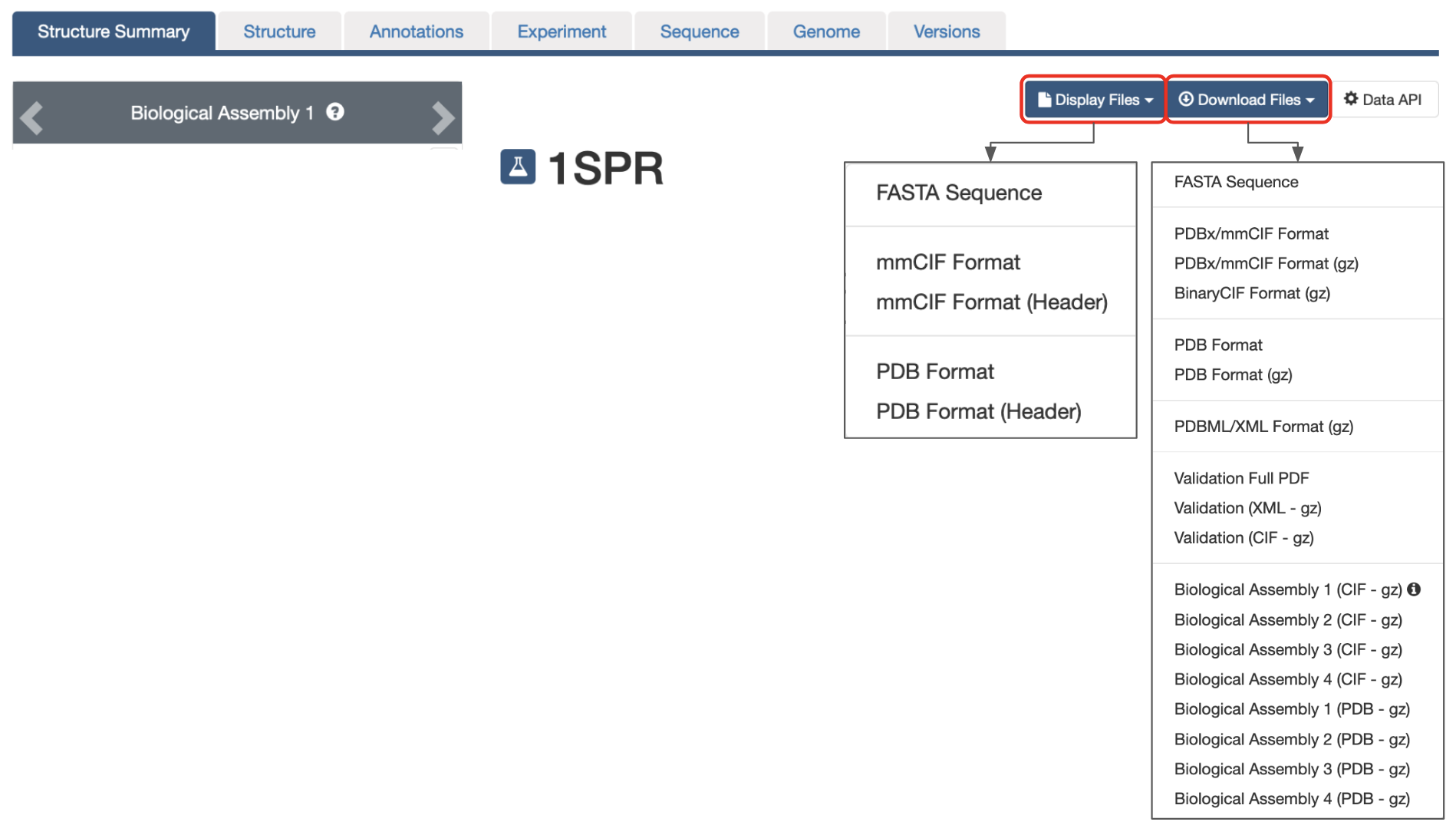
Answer:

* 1. Which amino acids are involved in forming the binding site for the phosphate (PO4) in this protein? What do these amino acids have in common?

Answer:

#### **Accessing and Downloading Protein Sequence**

In the top right-hand corner of any of a PDB entry’s pages you will see drop-down menus that allow you to display or download its associated files (see figure 2). The [FASTA sequence](https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=BlastHelp) is a text file that contains a description of the entry followed by the primary sequence information for a polymeric macromolecule (e.g. – a nucleic acid or protein). Sequences are displayed using the relevant one-letter code.

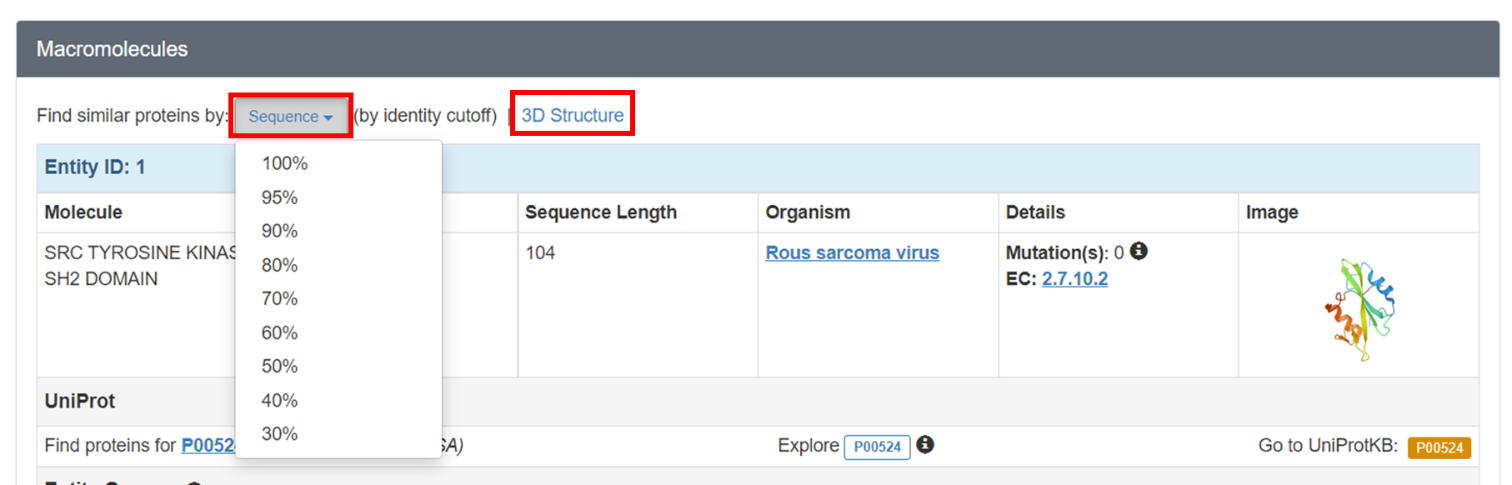


*Figure 2: Top portion of the RCSB PDB structure summary page for* [*1spr*](https://www.rcsb.org/structure/1SPR)*, with the file display and download menus highlighted (boxed in red). Insets below show the options found in both menus.*

Try displaying the FASTA sequence for 1spr. You can also download the file for use in later parts of this exercise.

#### **Exploring and PDB for other structures with the same/similar Protein Sequence**

For now, we will return to the “Structure Summary” page, “Macromolecules” section, to conduct searches for similar entries in the PDB using the sequence or the structure (see figure 3) and compare our finding between the two.



*Figure 3: View of the” Macromolecules” section of the RCSB PDB “Structure Summary” page for* [*1spr*](https://www.rcsb.org/structure/1SPR)*, with the search by sequence and search by 3D structure, links highlighted.*

Exploring the structures of other proteins that share sequence or structure similarity can provide insights about functional mechanisms and/or evolutionary relationships. Structures with high level of sequence identity usually share the same structure (and function), while structures with high structure similarity may or may not have high sequence similarity.

Note: Sequence identity between two proteins is a measure of the %age of amino acids that are exactly the same in them. These are the conserved residues and often play critical roles in the structure and functions of the protein. Sequence similarity, on the other hand measures the extent to which amino acids in the sequences being compared are similar in their size, chemical nature, etc. (e.g., Ile and Leu or Ser and Thr). Grouping protein sequences by similarity is used to determine homology and potential evolutionary relationships between them.

Conduct three separate searches by sequence using the identity cutoff set to 100, 90, and 40% respectively.

1. Complete the following table indicating how many results your search returned and the organism(s) they originated in.

| **% identity cutoff** | **# of results returned** | **Source Organism** |
| --- | --- | --- |
| **100%** | 11 | Rous sarcoma virus (wt and Schmidt-Ruppin strain), Gallus gallus |
| **90%** |  |  |
| **40%** |  |  |

\* note the results are for searches conducted July 14th 2022, results will vary with the ongoing evolution of PDB entries

1. Examine the results for the 90% sequence similarity search and locate the sequence comparison results for the PDB entries listed below. What % identity do these sequences share with 1spr? What specific differences do they display? Explain how you determined this.

Hint: % identity value is listed in the ‘sequence match’ line in the results returned

Answer:

Now conduct a search using the “3D Structure” option above (see figure 3). By default, this reach will use ‘strict’ criteria for comparison.

1. How many results does your ‘strict’ search by 3D structure return? What organisms do the results belong to?

Answer:

1. Compare the number of results returned for the 90% sequence identity search and the strict 3D structure search? Which returns more results? Does this make sense? Why or why not?

Answer:

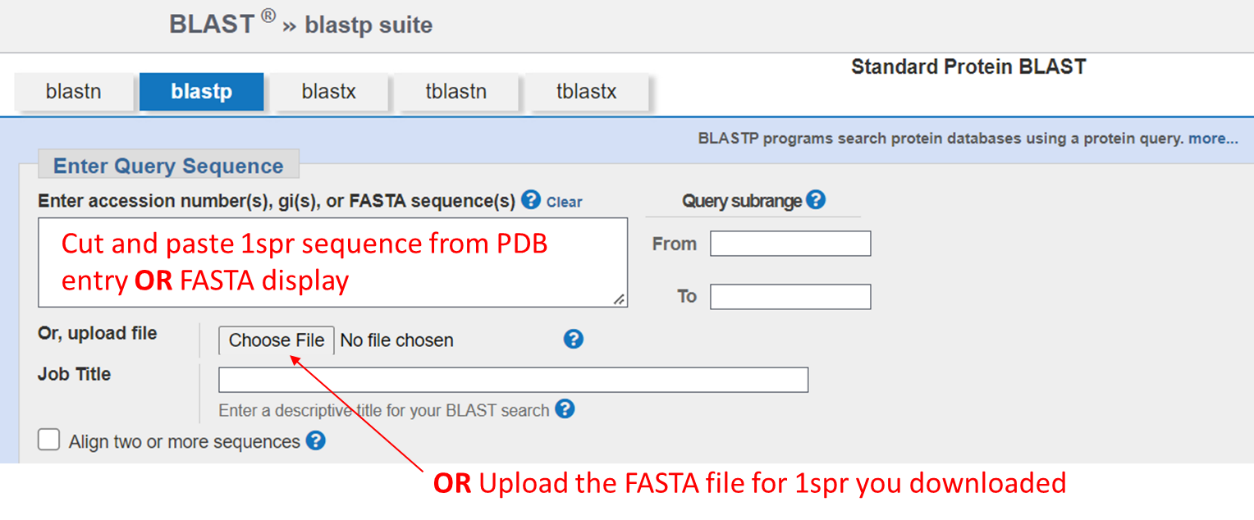
### **Part II: Exploring Protein Sequences in other Bioinformatics resources.**

The PDB is not the only database in which we can search for proteins by their sequence; in fact, there are many other possibilities. The [National Center for Biotechnology Information’s (NCBI) Protein](https://www.ncbi.nlm.nih.gov/protein/) is a repository of protein sequence information that is linked to many other repositories (e.g. - GenBank, RefSeq, SwissProt, PIR, PRF, and PDB) that contain both protein sequence information and more. Another key resource for protein sequences is UniProt.

#### **NCBI Protein**

Now you will search this repository for sequences similar to the 1spr protein using their [protein BLAST tool](https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome).

* Open the protein BLAST webpage and enter the sequence for 1spr



*Figure 4: BLASTp interface for searching protein’s that have the same or similar sequence compared to the query.*

*For your convenience the FASTA sequence for the protein in PDB entry 1spr is:*

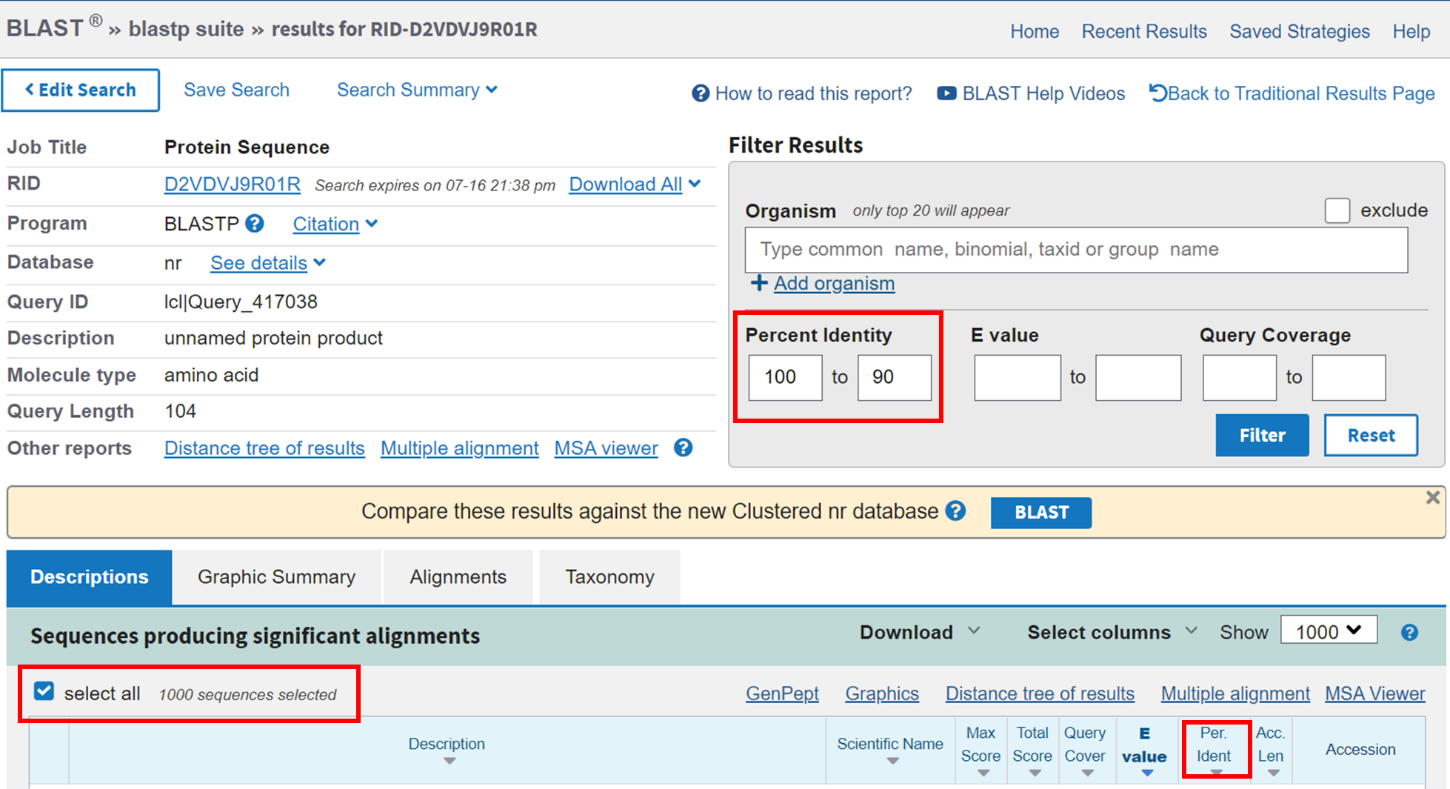
*QAEEWYFGKITRRESERLLLNPENPRGTFLVRESETTKGAYCLSVSDFDNAKGLNVKHYKIRKLDSGGFYITSRTQFSSLQQLVAYYSKHADGLCHRLTNVCPT*

* Scroll down and open (click the +) the Algorithm Parameters section at the bottom of the page
  + Adjust the first setting ‘Max target sequences’ to 1000
* Click BLAST to begin your search, wait as your search is completed

1. Scroll to the bottom of the results returned. What % identity does the 1000th entry share with 1spr?

Answer:

Using the filtering options at the top of the search results page, display only the results with 90% sequence identity or higher.



1. How many sequences are returned with 90% or greater sequence identity in your protein BLAST search? Is this more or less than your PDB search with the same cutoff?

Answer:

1. Do any of these belong to organisms not reflected in your earlier PDB search? If so, name one that did not appear in your PDB results.

Answer:

1. Why do the searches using the same query sequence produce such different results depending on if you use BLAST or the PDB?

Answer:

#### **UniProt**

[UniProt](https://www.uniprot.org/) or the Universal Protein Resource acts as a single, centralized, authoritative resource for protein sequences and functional information (1). Not only can you access annotated sequence information at UniProt, but it also contains a variety of integrated analysis tools (including the BLAST search you just completed) which allow you to explore these sequences in greater detail.

While you could search for your protein of interest directly at UniProt, we will access it via links provided in our PDB entry. Within the “Macromolecules” portion of the “Structure Summary” page, you will find a series of links related to the protein’s UniProt ID. The two left-most clickable links in the UniProt row of the table (see links in the left-hand boxed region of Fig.4) allow you to search within the PDB using that UniProt ID. The right-most clickable link (see link in the right-hand boxed region of Fig.4) will take you to the related UniProt entry directly.



*Figure 4: View of the” Macromolecules” section of the RCSB PDB “Structure Summary” page for* [*1spr*](https://www.rcsb.org/structure/1SPR)*, with the UniProt search options highlighted.*

Click the ‘Go to UniProtKB’ link on the 1spr structure summary page and take a few moments to explore the P00524 entry in UniProt.

1. What is the name of the protein and the gene listed in this UniProt entry?

Protein – \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Gene – \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. How many amino acids are listed as being part of the protein in this UniProt entry? Is this the same/different from what is reported in the PDB for the PDB entry 1spr (where you began)?

Answer:

Click on ‘Feature Viewer’ in the navigation menu at the top of the page. Use the information presented here to answer questions 13-15 below.

1. Using the information presented in the feature viewer, explain your observations with respect to sequence length of 1spr vs P00524 from question 12.

Answer:

1. What other domains does this protein contain?

Answer:

1. What posttranslational modifications (PTMs) does this protein undergo? Be sure to specify the type and location at which each one occurs.

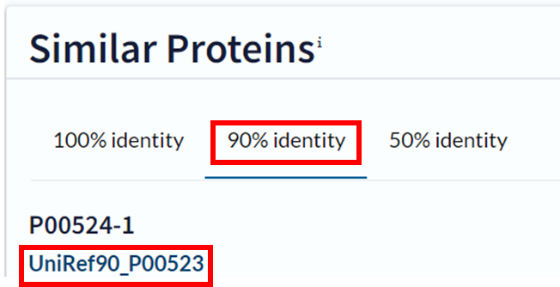
Answer:

Click ‘Entry’ in the top navigation bar to return to the main UniProt page for this protein. Select Sequence in the left-hand navigation menu.

1. What is the first amino acid listed in this protein’s sequence? Could you have predicted this? Why or why not?

Answer:

Scroll down within the sequence section to the ‘Similar Proteins’ section and click on the 90% identity tab to conduct a search. Once the results are returned, click on the search reference ID “UniRef90\_P00523’ to expand to the complete results



1. How many results were returned with 90% sequence identity to P00524?

Answer:

Using the checkbox at the start of each entry, select results P00523, P15054, P14085, and G1MWN3. Click align at the top of the results table to perform a multisequence alignment using the integrated Clustal Omega tool.

Click ‘align 4 sequences’ to perform the alignment. Once the alignment is complete, open the returned result by clicking on the available link in the results table.

1. Describe the key sequence differences identified.

Answer:

At the top of the alignment results page, select the “Phylogenetic Tree” tab.

1. Take a screenshot of the phylogenetic tree calculated for your 4 aligned sequences and include it below.

Answer:

References

1. The UniProt Consortium, UniProt: the universal protein knowledgebase in 2021, *Nucleic Acids Research*, Volume 49, Issue D1, 8 January 2021, Pages D480–D489,