

# CSCI-GA.3033-111: Protein Design HW1

Christine Wu (email: cw4459@nyu.edu)

Oct 6 2024

## 1 Learn how to download protein data from PDB. Become familiar with common protein tools.

### 1.1 Look up Trypsin on Uniprot, explore its function. Extract its PDB ID and explore on PDB.

**Solution:** According to the web information from [UniProt Trypsin Entry](#), the functionality of Trypsin can be considered as:

- Catalytic activity: preferential cleavage at Arg | Xaa, Lys | Xaa, meaning the enzyme cleaves peptide bonds on the C-terminal side of arginine (Arg) or lysine (Lys), regardless of the following amino acid (Xaa).
- Cofactor: the enzyme requires  $\text{Ca}^{2+}$  as a cofactor, binding one calcium ion per subunit. This is supported by databases like UniProtKB, Rhea, and ChEBI (ChEBI:29108).
- Activity regulation: it is inhibited by scorpion cyclotide trypsin inhibitor TopI1.
- Features: the enzyme has several active and binding sites. The active sites at positions 63, 107, and 200 are involved in the charge relay system, essential for catalysis. It binds  $\text{Ca}^{2+}$  at positions 75, 77, 80, and 85, while positions 194-195, 197-198, and 200 are responsible for substrate binding.
- Cellular Component: it is located in the extracellular space and part of the serine protease inhibitor complex.
- Molecular Function: it exhibits serine-type endopeptidase activity, endopeptidase activity, metal ion binding, and serpin family protein binding.
- Biological Process: it is involved in digestion and proteolysis.
- Keywords: Hydrolase, protease, serine protease, calcium binding, metal-binding.

Besides the general functionality of Trypsin, we also learned that Trypsin is secreted as an inactive precursor, trypsinogen, by the pancreas. It is activated in the small intestine by enteropeptidase, which cleaves trypsinogen to form active trypsin. Active trypsin also plays a role in activating other digestive enzymes, such as chymotrypsinogen and procarboxypeptidase, driving a cascade of enzyme activations that contribute to protein digestion. It also uses a serine residue in its active site to perform hydrolysis of peptide bonds. This mechanism is characteristic of serine proteases, etc.

For its PDB ID, we can move to the structure section of [UniProt Trypsin Entry](#), and we want to choose a structure with a high resolution (preferably below 2.5 Å, as higher-resolution structures are typically more accurate, as well as a common form. Hence we picked **2PTN**.

## 1.2 Using PDB's website built-in-search, download two sets of similar proteins. One found using a sequence similarity search, and one using a structure similarity search.

**Solution:** Now, we're using [PDB 2PTN](#) to do the sequence and structure searches.

- **Sequence similarity search:** For its sequence similarity, we found a set of PDB IDS that can be found from here: [2PTN sequence similarity search](#). This includes **1AQ7**, **1AUJ**, **1AZ8**, etc.
- **Structure similarity search:** For its structure similarity, we found a set of PDB IDS from here: [2PTN structure similarity search](#). This includes **3PTB**, **5MOR**, **1TNL**, etc.

## 1.3 Explore how the sequence search set's structures look with pymol.

**Solution:** Here, we choose the sequences of **2PTN**, **1AQ7**, and **1AUJ** to do the comparisons (Note: I'm using PyMol open source for this assignment).

### 1.3.1 Align structures on top of each other

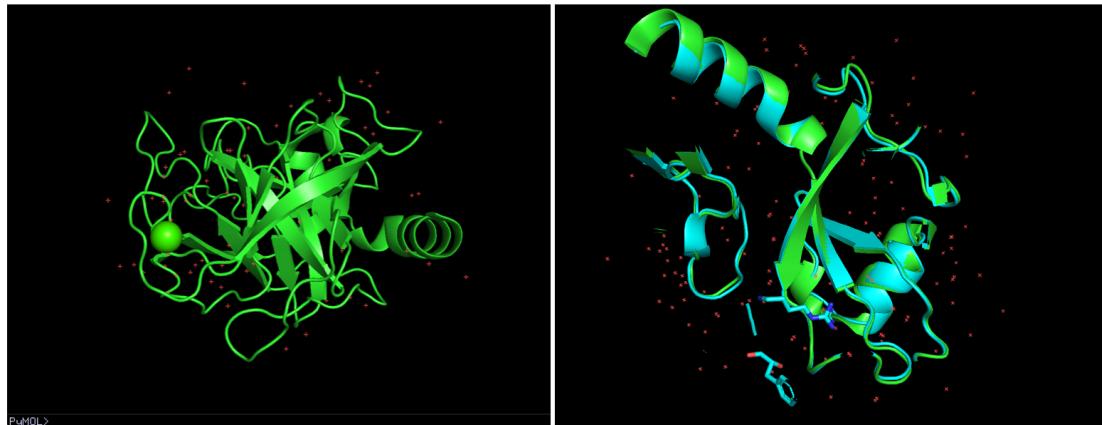


Figure 1: left: 2PTN; right: 2PTN and 1AQ7 aligning on top of each other, where 2PTN is the green one and 1AQ7 is the blue one.

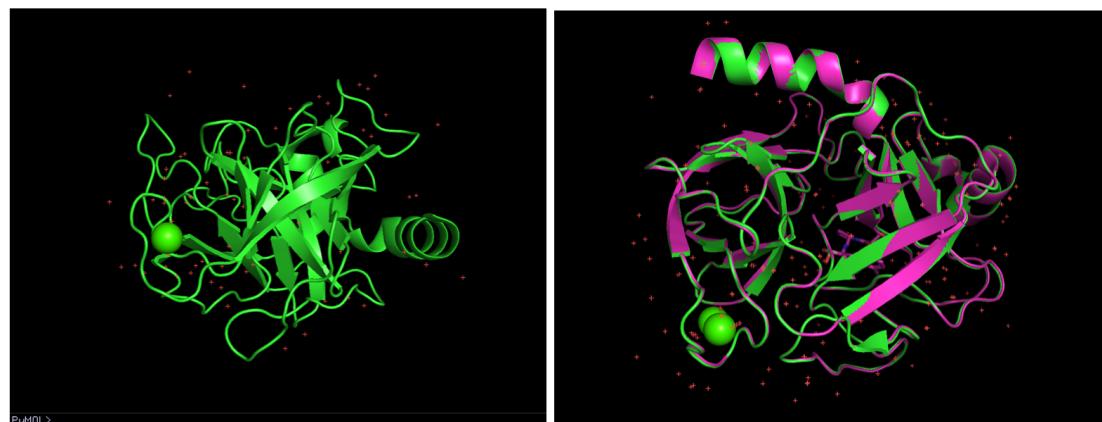


Figure 2: left: 2PTN; right: 2PTN and 1AUJ aligning on top of each other, where 2PTN is the green one and 1AUJ is the pink one.

### 1.3.2 Try different visualization methods (surface, cartoon, etc.)

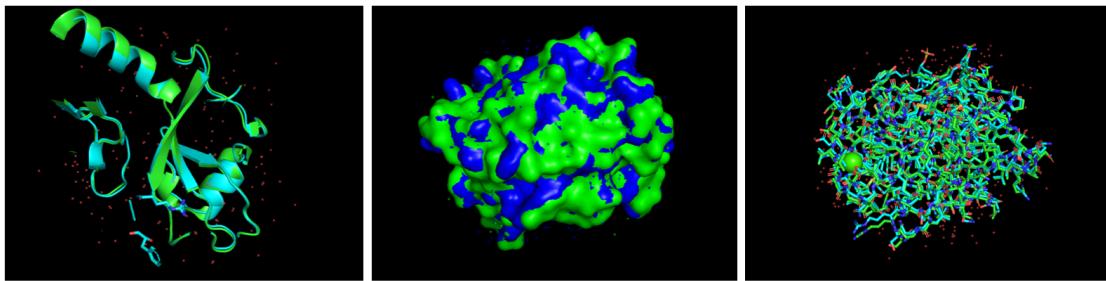


Figure 3: 2PTN and 1AQ7 aligning on top of each other. left: cartoon method; middle: surface method; right: stick method

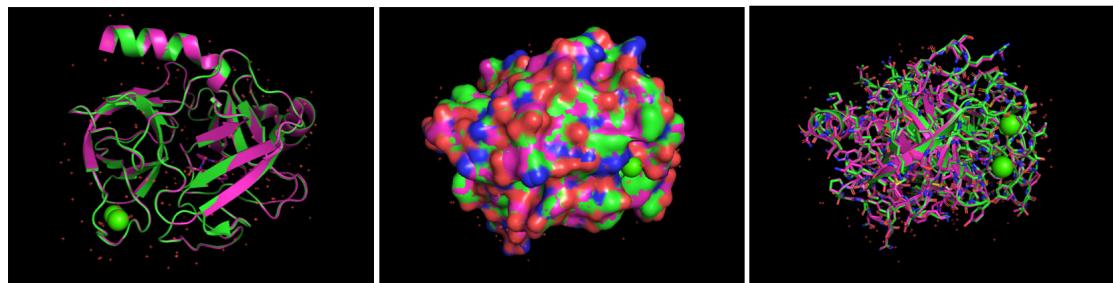


Figure 4: 2PTN and 1AUJ aligning on top of each other. left: cartoon method; middle: surface method; right: stick method

## 1.4 Using Blast and Align, create an MSA for Trypsin. Try to include proteins from at least 10 different organisms in the MSA.

**Solutions:** From the Blast, we included at least 10 different organisms as the following (with top 10 organisms selected):

### BLAST 250 results found in UniProtKB

Overview	Taxonomy	Hit Distribution	Text Output	Input Parameters	API Request
<a href="#">Tools</a> <a href="#">Download</a> <a href="#">Add</a> <a href="#">Customize columns</a> <a href="#">Resubmit</a>					
Entry	Entry Name	Protein Names	Gene Names	Organism	Length
<input checked="" type="checkbox"/> P00760	TRY1_BOVIN	Serine protease 1[...]	PRSS1, TRP1, TRY1, TRYPI	Bos taurus (Bovine)	246 AA
<input checked="" type="checkbox"/> A0A4W2FD21	A0A4W2FD21_BOBOX	Cationic trypsin	LOC113890875	Bos indicus x Bos taurus (Hybrid cattle)	246 AA
<input checked="" type="checkbox"/> A0A6P5BKV6	A0A6P5BKV6_BOSIN	Cationic trypsin	LOC109557710	Bos indicus (Zebu)	246 AA
<input checked="" type="checkbox"/> A0A452EML0	A0A452EML0_CAPHI	Cationic trypsin	LOC102178719, LOC102179184	Capra hircus (Goat)	246 AA
<input checked="" type="checkbox"/> A0A6P3IFR1	A0A6P3IFR1_BISBB	Cationic trypsin isoform X1	LOC104999401	Bison bison bison (North American plains bison)	246 AA
<input checked="" type="checkbox"/> A0A8B9W8U3	A0A8B9W8U3_BOSMU	Peptidase S1 domain-containing protein		Bos mutus grunniens (Wild yak) (Bos grunniens)	246 AA
<input checked="" type="checkbox"/> A0A452FA14	A0A452FA14_CAPHI	Peptidase S1 domain-containing protein	LOC102179184	Capra hircus (Goat)	246 AA
<input checked="" type="checkbox"/> A0A6P3IN88	A0A6P3IN88_BISBB	Cationic trypsin isoform X2	LOC104999401	Bison bison bison (North American plains bison)	246 AA
<input checked="" type="checkbox"/> W5Q1Y9	W5Q1Y9_SHEEP	Peptidase S1 domain-containing protein	LOC101112049	Ovis aries (Sheep)	246 AA
<input checked="" type="checkbox"/> A0A5N3XT99	A0A5N3XT99_MUNRE	Peptidase S1 domain-containing protein	FD755_012394	Muntiacus reevesi (Reeves' muntjac) (Cervus reevesi)	246 AA

Figure 5: BLAST Result

For the **Align** result, we have the following:

## Align results

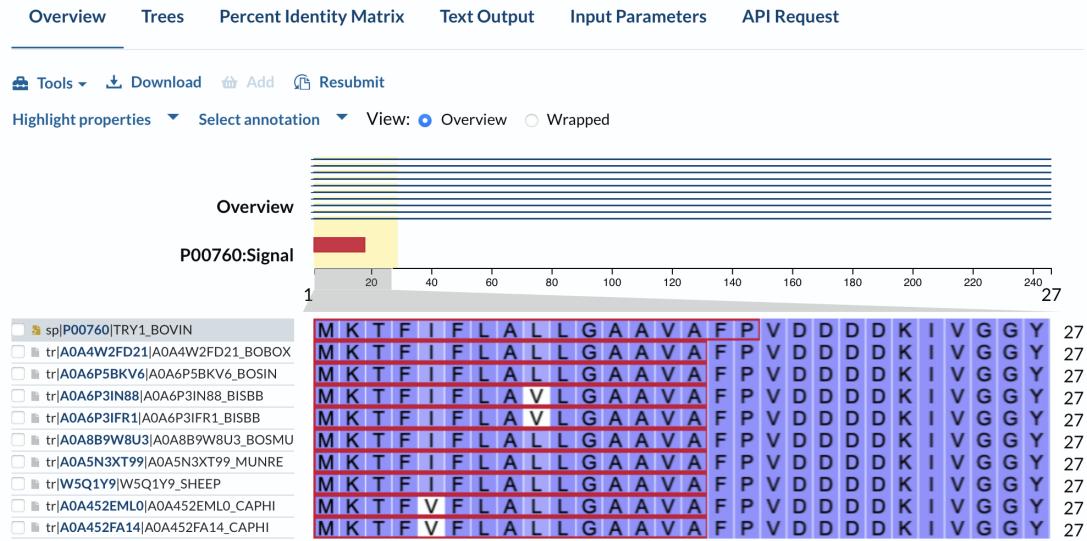


Figure 6: The overview of the aligned result for MSA

(**Note:** the more detailed version of the wrapped result and the output MSA are attached at the end of the document).

**1.5 Use the MSA to identify adjacent conserved residues within Trypsin. Using PyMol select these resides and use them to align to the other proteins you obtained in B).**

**Solution:** The adjacent conserved residues we will be using are: **Positions 1-4:** MKTF, **Positions 37-46:** GGYTCGANT, **Positions 65-72:** CGGSLINS, and **Positions 222-229:** KPGVYTKV. Then, the results of using PyMol are the following:

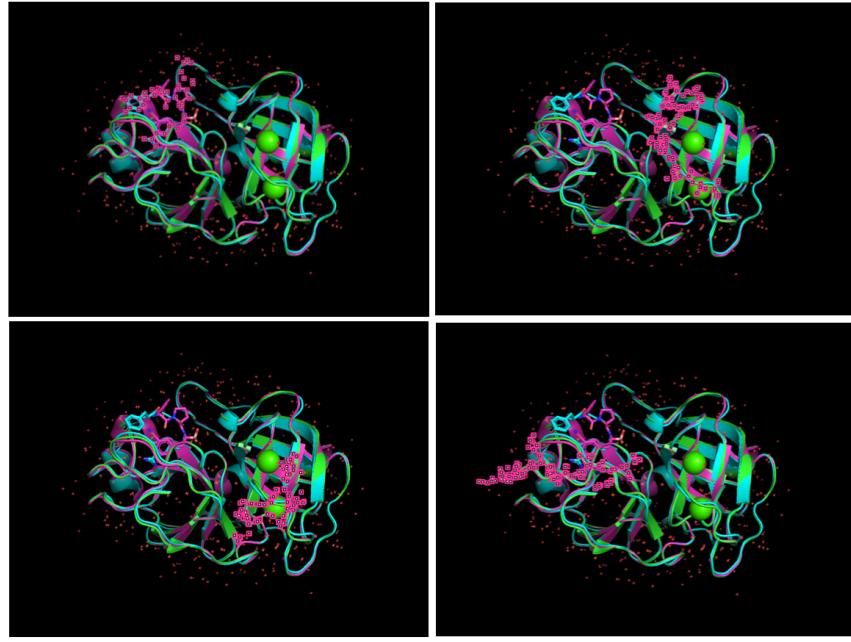


Figure 7: Sequence similarity proteins: green-2PTN, blue-1AQ7, pink-1AUJ, with adjacent conserved residues **Positions 1-4:** MKTF, **Positions 37-46:** GGYTCGANT, **Positions 65-72:** CGGSLINS, and **Positions 222-229:** KPGVYTKV (highlighted in red dots).

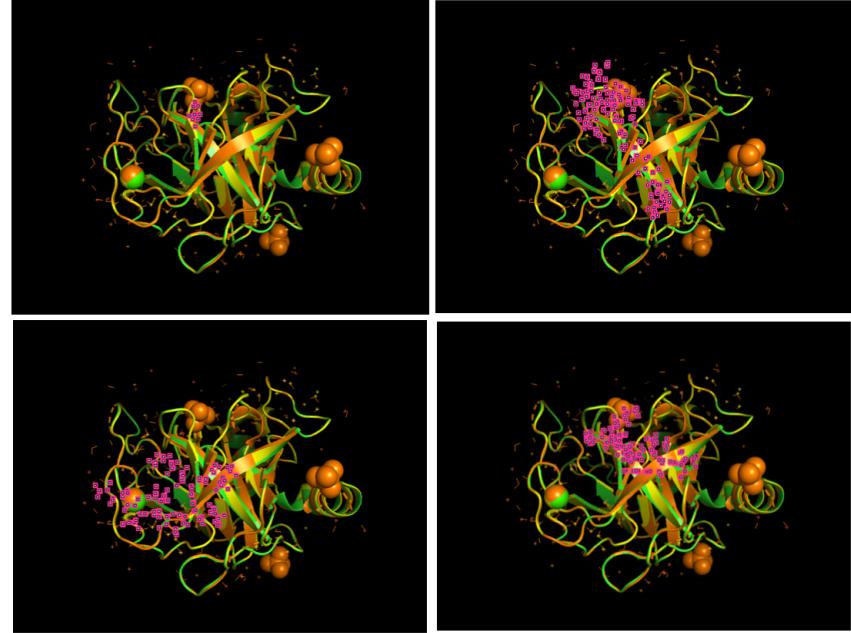


Figure 8: Structure similarity proteins: green-2PTN, yellow-3PTB, orange-5MOR, with adjacent conserved residues **Positions 1-4:** MKTF, **Positions 37-46:** GGYTCGANT, **Positions 65-72:** CGGSLINS, and **Positions 222-229:** KPGVYTKV (highlighted in red dots).

**1.6** In Python, using the examples obtained in B), implement two distance functions: one that is based on the proteins' sequence, and another that relies on its structure. You may use an existing library or create this yourself, depending on the level of challenge you want to attempt.

**Solution:** Here, we used the built-in libraries to implement the functions:

- a) The distance function that calculates based on protein's sequences:

```
1 import os
2 import warnings
3 from Bio import PDB
4 from Bio import pairwise2
5 from Bio.pairwise2 import format_alignment
6 from prody import *
7 import time

8 # Suppress PDB warnings generically
9 warnings.filterwarnings("ignore", message="WARNING:.*discontinuous.*")

10 #-----
11 def get_pdb_id(pdb_file):
12     """
13         Extracts and returns the PDB ID (filename without the extension) from a PDB file path.
14     """
15     return os.path.splitext(os.path.basename(pdb_file))[0]

16 #-----
17 def extract_sequence_from_pdb(pdb_file):
18     """
19         Extracts the sequence of the first chain from the given PDB file using PPBuilder.
20     """
21     parser = PDB.PDBParser()
22     structure = parser.get_structure("protein", pdb_file)

23     # Use the PPBuilder to extract the sequence
24     ppb = PDB.PPBuilder()

25     # Get the sequence from the first polypeptide found in the first model
26     for pp in ppb.build_peptides(structure):
27         sequence = pp.get_sequence()
28         return str(sequence)

29 #-----
30 def sequence_distance(pdb1, pdb2):
31     """
32         Calculates sequence distance (based on percentage of non-identical residues)
33         between two proteins based on their PDB files.
34     """
35     seq1 = extract_sequence_from_pdb(pdb1)
36     seq2 = extract_sequence_from_pdb(pdb2)

37     # Use global pairwise alignment (Needleman-Wunsch) to align sequences
38     alignments = pairwise2.align.globalxx(seq1, seq2)

39     # Get the alignment with the highest score (optimal alignment)
40     best_alignment = alignments[0]

41     # Calculate the percentage of identical residues in the alignment
42     identical_residues = sum(1 for a, b in zip(best_alignment[0], best_alignment[1]) if a == b)
43     alignment_length = len(best_alignment[0])

44     # Calculate percentage identity
45     percentage_identity = (identical_residues / alignment_length) * 100
46     return 100 - percentage_identity
```

- b) The distance function that calculates based on protein's structures:

```

1 #-----
2 def calculate_rmsd(pdb1, pdb2):
3     """
4     Calculates the RMSD (Root Mean Square Deviation) between two PDB files based on their backbone
5     alignment.
6     """
7     # Parse the PDB files
8     structure1 = parsePDB(pdb1)
9     structure2 = parsePDB(pdb2)
10
11    # Select backbone atoms for alignment
12    backbone1 = structure1.select('backbone')
13    backbone2 = structure2.select('backbone')
14
15    if backbone1 is None or backbone2 is None:
16        print(f"Error: Backbone atoms not found in {pdb1} or {pdb2}")
17        return None
18
19    # Check if both backbones have the same number of atoms
20    if backbone1.numAtoms() != backbone2.numAtoms():
21        print(f"Skipping RMSD calculation for {pdb1} and {pdb2} due to mismatched atom counts.")
22        return None
23
24    # Superimpose the two structures
25    transformation = calcTransformation(backbone2, backbone1)
26    transformation.apply(backbone2)
27
28    # Calculate RMSD
29    rmsd_value = calcRMSD(backbone1, backbone2)
30    return rmsd_value

```

- 1.7 Benchmark both the sequence-similar set and the structure-similar set with both criterions. Compare their respective values. Compare them to a set of randomly chosen proteins from PDB. Write a short conclusion on why those values make sense.**

**Solution:** We first have the benchmark function as the following:

```

1 #-----
2 def benchmark_proteins(pdb_set_1, pdb_set_2):
3     """
4     Benchmarks protein pairs for both sequence distance and structure distance (RMSD).
5     """
6     for pdb1 in pdb_set_1:
7         for pdb2 in pdb_set_2:
8             start_time = time.time()
9
10            try:
11                seq_distance = sequence_distance(pdb1, pdb2)
12                struct_distance = calculate_rmsd(pdb1, pdb2)
13
14                pdb_id1 = get_pdb_id(pdb1)
15                pdb_id2 = get_pdb_id(pdb2)
16
17                print(f"Comparing {pdb_id1} and {pdb_id2}:")
18                print(f" Sequence distance: {seq_distance}")
19
20                if struct_distance is not None:
21                    print(f" Structure distance (RMSD): {struct_distance}")
22                else:
23                    print(f" Structure distance (RMSD): Skipped due to mismatched atom counts or missing
24                         backbone.")
25
26            except Exception as e:

```

```

26     print(f"Error comparing {pdb1} and {pdb2}: {e}")
27
28     finally:
29         print(f"Time taken: {time.time() - start_time:.2f} seconds\n")
30
31
32 #-----
33 # load the files:
34 sequence_similar_set = [
35     "2ptn.pdb",
36     "1aq7.pdb",
37     "1auj.pdb"
38 ]
39
40 structure_similar_set = [
41     "2ptn.pdb",
42     "3ptb.pdb",
43     "5mor.pdb"
44 ]
45
46 random_pdb_ids = [
47     "1xyz.pdb",
48     "4hhb.pdb",
49     "3def.pdb"
50 ]
51
52 #-----
53 print("----- START -----")
54 # Perform the benchmark for sequence-similar, structure-similar, and random sets
55 print("Benchmarking Sequence-Similar Set:")
56 benchmark_proteins(sequence_similar_set, sequence_similar_set)
57 print("-----")
58
59 print("\nBenchmarking Structure-Similar Set:")
60 benchmark_proteins(structure_similar_set, structure_similar_set)
61 print("-----")
62
63 print("\nBenchmarking Random Proteins Set:")
64 benchmark_proteins(random_pdb_ids, random_pdb_ids)
65 print("----- END -----")

```

Now, we have the outputs as:

```

1 ----- START -----
2 Benchmarking Sequence-Similar Set:
3 @> 1712 atoms and 1 coordinate set(s) were parsed in 0.02s.
4 @> 1712 atoms and 1 coordinate set(s) were parsed in 0.02s.
5 Comparing 2ptn and 2ptn:
6     Sequence distance: 0.0
7     Structure distance (RMSD): 4.976419911094366e-15
8 Time taken: 0.08 seconds
9
10 @> 1712 atoms and 1 coordinate set(s) were parsed in 0.02s.
11 @> 1886 atoms and 1 coordinate set(s) were parsed in 0.02s.
12 Comparing 2ptn and 1aq7:
13     Sequence distance: 0.0
14     Structure distance (RMSD): 0.4472085649964476
15 Time taken: 0.09 seconds
16
17 @> 1712 atoms and 1 coordinate set(s) were parsed in 0.02s.
18 @> 1802 atoms and 1 coordinate set(s) were parsed in 0.02s.
19 Comparing 2ptn and 1auj:
20     Sequence distance: 0.0
21     Structure distance (RMSD): 0.18055592175961002
22 Time taken: 0.07 seconds
23
24 @> 1886 atoms and 1 coordinate set(s) were parsed in 0.02s.
25 @> 1712 atoms and 1 coordinate set(s) were parsed in 0.02s.

```

```

26 | Comparing 1aq7 and 2ptn:
27 |   Sequence distance: 0.0
28 |   Structure distance (RMSD): 0.44720856499644845
29 | Time taken: 0.07 seconds
30 |
31 | @> 1886 atoms and 1 coordinate set(s) were parsed in 0.02s.
32 | @> 1886 atoms and 1 coordinate set(s) were parsed in 0.02s.
33 | Comparing 1aq7 and 1aq7:
34 |   Sequence distance: 0.0
35 |   Structure distance (RMSD): 7.235314718896078e-15
36 | Time taken: 0.08 seconds
37 |
38 | @> 1886 atoms and 1 coordinate set(s) were parsed in 0.02s.
39 | @> 1802 atoms and 1 coordinate set(s) were parsed in 0.02s.
40 | Comparing 1aq7 and 1auj:
41 |   Sequence distance: 0.0
42 |   Structure distance (RMSD): 0.450915874320609
43 | Time taken: 0.10 seconds
44 |
45 | @> 1802 atoms and 1 coordinate set(s) were parsed in 0.02s.
46 | @> 1712 atoms and 1 coordinate set(s) were parsed in 0.02s.
47 | Comparing 1auj and 2ptn:
48 |   Sequence distance: 0.0
49 |   Structure distance (RMSD): 0.18055592175960897
50 | Time taken: 0.07 seconds
51 |
52 | @> 1802 atoms and 1 coordinate set(s) were parsed in 0.02s.
53 | @> 1886 atoms and 1 coordinate set(s) were parsed in 0.02s.
54 | Comparing 1auj and 1aq7:
55 |   Sequence distance: 0.0
56 |   Structure distance (RMSD): 0.4509158743206105
57 | Time taken: 0.07 seconds
58 |
59 | @> 1802 atoms and 1 coordinate set(s) were parsed in 0.02s.
60 | @> 1802 atoms and 1 coordinate set(s) were parsed in 0.02s.
61 | Comparing 1auj and 1auj:
62 |   Sequence distance: 0.0
63 |   Structure distance (RMSD): 7.535431622541843e-15
64 | Time taken: 0.07 seconds
65 |
66 -----
67 |
68 | Benchmarking Structure-Similar Set:
69 | @> 1712 atoms and 1 coordinate set(s) were parsed in 0.02s.
70 | @> 1712 atoms and 1 coordinate set(s) were parsed in 0.02s.
71 | Comparing 2ptn and 2ptn:
72 |   Sequence distance: 0.0
73 |   Structure distance (RMSD): 4.976419911094366e-15
74 | Time taken: 0.10 seconds
75 |
76 | @> 1712 atoms and 1 coordinate set(s) were parsed in 0.02s.
77 | @> 1701 atoms and 1 coordinate set(s) were parsed in 0.02s.
78 | Comparing 2ptn and 3ptb:
79 |   Sequence distance: 0.0
80 |   Structure distance (RMSD): 0.10683739877958595
81 | Time taken: 0.07 seconds
82 |
83 | @> 1712 atoms and 1 coordinate set(s) were parsed in 0.02s.
84 | @> 3367 atoms and 1 coordinate set(s) were parsed in 0.04s.
85 | Comparing 2ptn and 5mor:
86 |   Sequence distance: 0.0
87 |   Structure distance (RMSD): 0.13068966615174452
88 | Time taken: 0.11 seconds
89 |
90 | @> 1701 atoms and 1 coordinate set(s) were parsed in 0.02s.
91 | @> 1712 atoms and 1 coordinate set(s) were parsed in 0.02s.
92 | Comparing 3ptb and 2ptn:
93 |   Sequence distance: 0.0

```

```

94 Structure distance (RMSD): 0.10683739877958587
95 Time taken: 0.07 seconds
96
97 @> 1701 atoms and 1 coordinate set(s) were parsed in 0.02s.
98 @> 1701 atoms and 1 coordinate set(s) were parsed in 0.02s.
99 Comparing 3ptb and 3ptb:
100 Sequence distance: 0.0
101 Structure distance (RMSD): 1.0887956813044752e-14
102 Time taken: 0.10 seconds
103
104 @> 1701 atoms and 1 coordinate set(s) were parsed in 0.02s.
105 @> 3367 atoms and 1 coordinate set(s) were parsed in 0.04s.
106 Comparing 3ptb and 5mor:
107 Sequence distance: 0.0
108 Structure distance (RMSD): 0.12969343118978552
109 Time taken: 0.11 seconds
110
111 @> 3367 atoms and 1 coordinate set(s) were parsed in 0.04s.
112 @> 1712 atoms and 1 coordinate set(s) were parsed in 0.02s.
113 Comparing 5mor and 2ptn:
114 Sequence distance: 0.0
115 Structure distance (RMSD): 0.13068966615174468
116 Time taken: 0.11 seconds
117
118 @> 3367 atoms and 1 coordinate set(s) were parsed in 0.04s.
119 @> 1701 atoms and 1 coordinate set(s) were parsed in 0.02s.
120 Comparing 5mor and 3ptb:
121 Sequence distance: 0.0
122 Structure distance (RMSD): 0.12969343118978627
123 Time taken: 0.14 seconds
124
125 @> 3367 atoms and 1 coordinate set(s) were parsed in 0.04s.
126 @> 3367 atoms and 1 coordinate set(s) were parsed in 0.04s.
127 Comparing 5mor and 5mor:
128 Sequence distance: 0.0
129 Structure distance (RMSD): 5.6429828945950354e-14
130 Time taken: 0.15 seconds
131 -----
132
133
134 Benchmarking Random Proteins Set:
135 @> 5621 atoms and 1 coordinate set(s) were parsed in 0.06s.
136 @> 5621 atoms and 1 coordinate set(s) were parsed in 0.06s.
137 Comparing 1xyz and 1xyz:
138 Sequence distance: 0.0
139 Structure distance (RMSD): 1.947462900819195e-14
140 Time taken: 0.24 seconds
141
142 @> 5621 atoms and 1 coordinate set(s) were parsed in 0.06s.
143 @> 4779 atoms and 1 coordinate set(s) were parsed in 0.05s.
144 Skipping RMSD calculation for 1xyz.pdb and 4hhb.pdb due to mismatched atom counts.
145 Comparing 1xyz and 4hhb:
146 Sequence distance: 82.39795918367346
147 Structure distance (RMSD): Skipped due to mismatched atom counts or missing backbone.
148 Time taken: 0.23 seconds
149
150 @> 5621 atoms and 1 coordinate set(s) were parsed in 0.06s.
151 @> 2128 atoms and 1 coordinate set(s) were parsed in 0.02s.
152 Skipping RMSD calculation for 1xyz.pdb and 3def.pdb due to mismatched atom counts.
153 Comparing 1xyz and 3def:
154 Sequence distance: 87.05882352941177
155 Structure distance (RMSD): Skipped due to mismatched atom counts or missing backbone.
156 Time taken: 0.15 seconds
157
158 @> 4779 atoms and 1 coordinate set(s) were parsed in 0.05s.
159 @> 5621 atoms and 1 coordinate set(s) were parsed in 0.06s.
160 Skipping RMSD calculation for 4hhb.pdb and 1xyz.pdb due to mismatched atom counts.
161 Comparing 4hhb and 1xyz:

```

```

162 Sequence distance: 82.39795918367346
163 Structure distance (RMSD): Skipped due to mismatched atom counts or missing backbone.
164 Time taken: 0.23 seconds
165
166 @> 4779 atoms and 1 coordinate set(s) were parsed in 0.05s.
167 @> 4779 atoms and 1 coordinate set(s) were parsed in 0.05s.
168 Comparing 4hhb and 4hhb:
169     Sequence distance: 0.0
170     Structure distance (RMSD): 6.146794224938476e-15
171 Time taken: 0.21 seconds
172
173 @> 4779 atoms and 1 coordinate set(s) were parsed in 0.05s.
174 @> 2128 atoms and 1 coordinate set(s) were parsed in 0.02s.
175 Skipping RMSD calculation for 4hhb.pdb and 3def.pdb due to mismatched atom counts.
176 Comparing 4hhb and 3def:
177     Sequence distance: 82.183908045977
178     Structure distance (RMSD): Skipped due to mismatched atom counts or missing backbone.
179 Time taken: 0.14 seconds
180
181 @> 2128 atoms and 1 coordinate set(s) were parsed in 0.02s.
182 @> 5621 atoms and 1 coordinate set(s) were parsed in 0.06s.
183 Skipping RMSD calculation for 3def.pdb and 1xyz.pdb due to mismatched atom counts.
184 Comparing 3def and 1xyz:
185     Sequence distance: 87.05882352941177
186     Structure distance (RMSD): Skipped due to mismatched atom counts or missing backbone.
187 Time taken: 0.18 seconds
188
189 @> 2128 atoms and 1 coordinate set(s) were parsed in 0.02s.
190 @> 4779 atoms and 1 coordinate set(s) were parsed in 0.05s.
191 Skipping RMSD calculation for 3def.pdb and 4hhb.pdb due to mismatched atom counts.
192 Comparing 3def and 4hhb:
193     Sequence distance: 82.183908045977
194     Structure distance (RMSD): Skipped due to mismatched atom counts or missing backbone.
195 Time taken: 0.14 seconds
196
197 @> 2128 atoms and 1 coordinate set(s) were parsed in 0.02s.
198 @> 2128 atoms and 1 coordinate set(s) were parsed in 0.02s.
199 Comparing 3def and 3def:
200     Sequence distance: 0.0
201     Structure distance (RMSD): 6.093927363022499e-14
202 Time taken: 0.08 seconds
203
204 ----- END -----

```

### Summary of Benchmarking Results:

The benchmarking results of sequence-similar, structure-similar, and randomly selected proteins provide insights into the relationship between protein sequences and structures.

- Sequence-Similar Set:** The sequence distance for all protein pairs was consistently 0.0, indicating nearly identical sequences. The RMSD values were small but varied slightly, with 0.447 Å for 2ptn vs. 1aq7 and 0.181 Å for 2ptn vs. 1auj, which is expected due to minor structural deviations. Comparisons of identical proteins (e.g., 2ptn vs. 2ptn) produced RMSD values close to zero, confirming no structural differences.
- Structure-Similar Set:** Low RMSD values (0.1-0.13 Å) confirmed that these proteins share very similar structures. However, the sequence distance was also 0.0 for all pairs, suggesting that these proteins are also highly conserved in sequence, which may require re-evaluation to focus on structurally similar but sequence-divergent proteins.
- Random Proteins Set:** The random proteins showed high sequence divergence (82-87 percent non-identity) as expected. Many RMSD calculations were skipped due to mismatched atom counts, reflecting the structural dissimilarity between the proteins. RMSD values were close to zero for comparisons of identical proteins, as anticipated.

## 1.8 Compute an intra-protein residue-wise distance map for Trypsin.

**Solution:** We have the map as below:

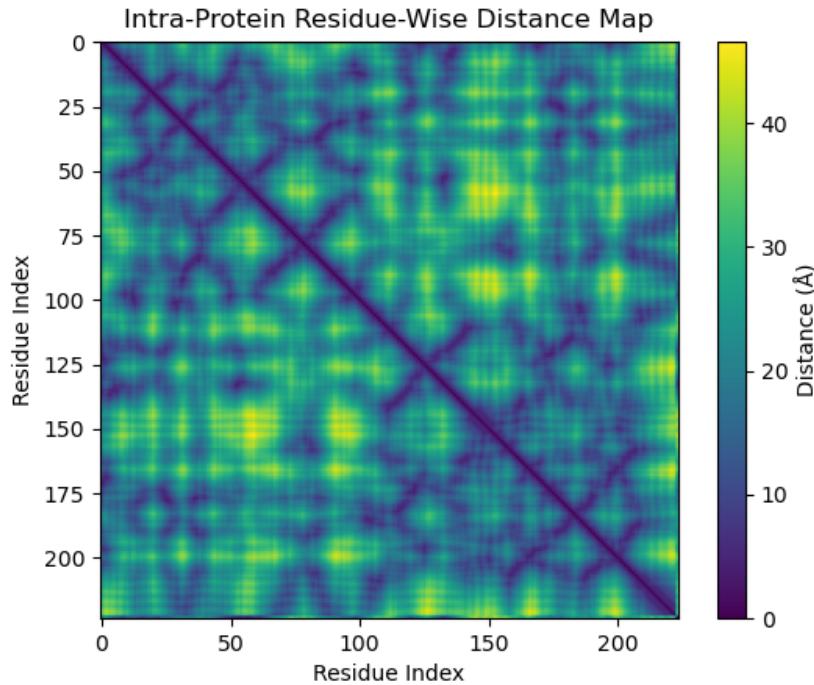


Figure 9: Intra-protein residue-wise distance map for Trypsin

The code is:

```
1-----  
2 def compute_distance_map(pdb_file):  
3     """  
4         Computes the intra-protein residue-wise distance map for the given PDB file.  
5         The distances are calculated between the alpha carbon atoms of all residues.  
6     """  
7     parser = PDBParser()  
8     structure = parser.get_structure("protein", pdb_file)  
9  
10    # Extract alpha carbon atoms  
11    ca_atoms = []  
12    for model in structure:  
13        for chain in model:  
14            for residue in chain:  
15                if 'CA' in residue: # Check if residue has an alpha carbon  
16                    ca_atoms.append(residue['CA'].get_coord())  
17  
18    # Convert the list of atom coordinates into a NumPy array  
19    ca_coords = np.array(ca_atoms)  
20  
21    # Initialize an empty distance matrix  
22    n = len(ca_coords)  
23    distance_map = np.zeros((n, n))  
24  
25    # Compute pairwise distances  
26    for i in range(n):  
27        for j in range(i, n):  
28            distance = np.linalg.norm(ca_coords[i] - ca_coords[j])
```

```
29         distance_map[i, j] = distance_map[j, i] = distance
30
31     return distance_map
32
33 def plot_distance_map(distance_map, title="Intra-Protein Residue-Wise Distance Map"):
34     """
35     Plots the distance map using Matplotlib.
36     """
37     plt.imshow(distance_map, cmap='viridis')
38     plt.colorbar(label='Distance ()')
39     plt.title(title)
40     plt.xlabel("Residue Index")
41     plt.ylabel("Residue Index")
42     plt.show()
43
44
45 pdb_file = "2ptn.pdb"
46 distance_map = compute_distance_map(pdb_file)
47 plot_distance_map(distance_map)
```

>sp|P00760|TRY1\_BOVIN  
MKTIFFLALLGAAVAFPVDDDKIVGGYTCGANTVPYQVSLNSGYHFCGGLINSQWVVS  
AAHCYKSGIQVRLGEDNINVVEGNEQFISASKSIVHPSYNSNTLNNNDIMLIKLKSAASLN  
SRVASISLPTSCASAGTQCLISGWGNTKSSGTSPDVLKCLKAPIILDSSCKSAYPGQIT  
SNMFCAGYLEGGKDSCQGDGGPVVCSDKLQGIVSWGSGCAQKNKPGVYTKVCNYVSWIK  
QTIASN  
>tr|A0A4W2FD21|A0A4W2FD21\_BOBOX  
MKTIFFLALLGAAVAFPVDDDKIVGGYTCGANTVPYQVSLNSGYHFCGGLINSQWVVS  
AAHCYKSGIQVRLGEDNINVVEGNEQFISASKSIVHPSYNSNTLNNNDIMLIKLKSAASLN  
SRVASISLPTSCASAGTQCLISGWGNTKSSGTSPDVLKCLKAPIILDSSCKSAYPGQIT  
SNMFCAGYLEGGKDSCQGDGGPVVCSDKLQGIVSWGSGCAQKNKPGVYTKVCNYVSWIK  
QTIASN  
>tr|A0A6P5BKV6|A0A6P5BKV6\_BOSIN  
MKTIFFLALLGAAVAFPVDDDKIVGGYTCGANTVPYQVSLNSGYHFCGGLINSQWVVS  
AAHCYKSGIQVRLGEDNINVVEGNEQFISASKSIVHPSYNSNTLNNNDIMLIKLKSAASLN  
SRVASISLPTSCASAGTQCLISGWGNTKSSGTSPDVLXCLKAPIILDSSCKSAYPGQIT  
SNMFCAGYLEGGKDSCQGDGGPVVCSDKLQGIVSWGSGCAQKNKPGVYTKVCNYVSWIK  
QTIASN  
>tr|A0A6P3IN88|A0A6P3IN88\_BISBB  
MKTIFFLAVLGAAVAFPVDDDKIVGGYTCGANTVPYQVSLNSGYHFCGGLINSQWVVS  
AAHCYKSGIQVRLGEDNINVVAEGNEQFISASKSIVHPSYNSNTLNNNDIMLIKLKSAASLN  
SRVASISLPTSCASAGTQCLISGWGNTKSSGTSPDVLQCLKAPIILDSSCKSAYPGQIT  
SNMFCAGYLEGGKDSCQGDGGPVVCSDKLQGIVSWGYGCAQKNKPGVYTKVCNYVSWIK  
QTIASN  
>tr|A0A6P3IFR1|A0A6P3IFR1\_BISBB  
MKTIFFLAVLGAAVAFPVDDDKIVGGYTCGANTVPYQVSLNSGYHFCGGLINSQWVVS  
AAHCYKSGIQVRLGEDNINVVAEGNEQFISASKSIVHPSYNSNTLNNNDIMLIKLKSAASLN  
SRVASISLPTSCASAGTQCLISGWGNTKSSGTSPDVLQCLKAPIILDSSCKSAYPGQIT  
SNMFCAGYLEGGKDSCQGDGGPVVCSDKLQGIVSWGYGCAQKNKPGVYTKVCNYVSWIK  
QTIASN  
>tr|A0A8B9W8U3|A0A8B9W8U3\_BOSMU  
MKTIFFLALLGAAVAFPVDDDKIVGGYTCGANTVPYQVSLNSGYHFCGGLINSQWVVS  
AAHCYKSGIQVRLGEDNINVVAEGNEQFISASKSIVHPSYNSNTLNNNDIMLIKLKSAASLN  
SRVASISLPTSCASAGTQCLISGWGNTKSSGTSPDVLQCLKAPIILDSSCKSAYPGQIT  
SNMFCAGYLEGGKDSCQGDGGPVVCSDKLQGIVSWGYGCAQKNKPGVYTKVCNYVSWIK  
QTIASN  
>tr|A0A5N3XT99|A0A5N3XT99\_MUNRE  
MKTIFFLALLGAAVAFPVDDDKIVGGYTCGANTVPYQVSLNSGYHFCGGLINSQWVVS  
AAHCYKSGIQVRLGEDNINVVAEGNEQFISASKSIVHPSYNQNTLNNNDIMLIKLKSAASLN  
SRVASISLPTSCASAGTQCLISGWGNTKSSGTSPDVLQCLKAPIILDSSCKSAYPGQIT  
SNMFCAGYLEGGKDSCQGDGGPVVCNGKLQGIVSWGYGCAQKNKPGVYTKVCNYVSWI  
QTIASN  
>tr|W5Q1Y9|W5Q1Y9\_SHEEP  
MKTIFFLALLGAAVAFPVDDDKIVGGYTCGANTVPYQVSLNSGYHFCGGLINSRWVVS  
AAHCYKSGIQVRLGEDNINVVAEGNEQFISASKSIVHPSYNSNTLNNNDIMLIKLKSAASLN  
SRVASISLPTSCASAGTQCLISGWGNTKSSGTSPDVLQCLKAPIILDSSCKSAYPGQIT  
SNMFCAGYLEGGKDSCQGDGGPVVCNGKLQGIVSWGYGCAQKNKPGVYTKVCNYVSWI  
QTIASN  
>tr|A0A452EML0|A0A452EML0\_CAPHI  
MKTIFFLALLGAAVAFPVDDDKIVGGYTCGANTVPYQVSLNSGYHFCGGLINSQWVVS  
AAHCYKSGIQVRLGEDNINVVAEGNEQFISASKSIVHPSYNSNTLNNNDIMLIKLKSAASLN  
SRVASISLPTSCASAGTQCLISGWGNTKSSGTSPDVLQCLKAPIILDSSCKSAYPGQIT  
SNMFCAGYLEGGKDSCQGDGGPVVCSDKLQGIVSWGYGCAQKNKPGVYTKVCNYVSWI  
QTIASN  
>tr|A0A452FA14|A0A452FA14\_CAPHI  
MKTIFFLALLGAAVAFPVDDDKIVGGYTCGANTVPYQVSLNSGYHFCGGLINSQWVVS  
AAHCYKSGIQVRLGEDNINVVAEGNEQFISASKSIVHPSYNSNTLNNNDIMLIKLKSAASLN  
SRVASISLPTSCASAGTQCLISGWGNTKSSGTSPDVLQCLKAPIILDSSCKSAYPGQIT  
SNMFCAGYLEGGKDSCQGDGGPVVCSDKLQGIVSWGYGCAQKNKPGVYTKVCNYVSWI  
QTIASN

Tools ▾

Tool results ▾

Advanced | List

Search

Help

# Align results

[Overview](#)   [Trees](#)   [Percent Identity Matrix](#)   [Text Output](#)   [Input Parameters](#)   [API Request](#)
[Tools](#) ▾   [Download](#)   [Add](#)   [Resubmit](#)
[Highlight properties](#) ▾   [Select annotation](#) ▾   View:  Overview  Wrapped

<input type="checkbox"/>	sp P00760 TRY1_BOVIN	M K T F I F L A L L G A A V A F P V D D D D K I V G G Y T C G A N T V P Y	37
<input type="checkbox"/>	tr AOA4W2FD21 AOA4W2FD21_BOBOX	M K T F I F L A L L G A A V A F P V D D D D K I V G G Y T C G A N T V P Y	37
<input type="checkbox"/>	tr AOA6P5BKV6 AOA6P5BKV6_BOSIN	M K T F I F L A L L G A A V A F P V D D D D K I V G G Y T C G A N T V P Y	37
<input type="checkbox"/>	tr AOA6P3IN88 AOA6P3IN88_BISBB	M K T F I F L A V L G A A V A F P V D D D D K I V G G Y T C G A N T V P Y	37
<input type="checkbox"/>	tr AOA6P3IFR1 AOA6P3IFR1_BISBB	M K T F I F L A V L G A A V A F P V D D D D K I V G G Y T C G A N T V P Y	37
<input type="checkbox"/>	tr AOA8B9W8U3 AOA8B9W8U3_BOSMU	M K T F I F L A L L G A A V A F P V D D D D K I V G G Y T C G A N T V P Y	37
<input type="checkbox"/>	tr AOA5N3XT99 AOA5N3XT99_MUNRE	M K T F I F L A L L G A A V A F P V D D D D K I V G G Y T C G A N T V P Y	37
<input type="checkbox"/>	tr W5Q1Y9 W5Q1Y9_SHEEP	M K T F I F L A L L G A A V A F P V D D D D K I V G G Y T C G A N T V P Y	37
<input type="checkbox"/>	tr AOA452EML0 AOA452EML0_CAPHI	M K T F V F L A L L G A A V A F P V D D D D K I V G G Y T C G A N T V P Y	37
<input type="checkbox"/>	tr AOA452FA14 AOA452FA14_CAPHI	M K T F V F L A L L G A A V A F P V D D D D K I V G G Y T C G A N T V P Y	37

P00760:Signal

<input type="checkbox"/>	sp P00760 TRY1_BOVIN	Q V S L N S G Y H F C G G S L I N S Q W V V S A A H C Y K S G I Q V R L G	74
<input type="checkbox"/>	tr AOA4W2FD21 AOA4W2FD21_BOBOX	Q V S L N S G Y H F C G G S L I N S Q W V V S A A H C Y K S G I Q V R L G	74
<input type="checkbox"/>	tr AOA6P5BKV6 AOA6P5BKV6_BOSIN	Q V S L N S G Y H F C G G S L I N S Q W V V S A A H C Y K S G I Q V R L G	74
<input type="checkbox"/>	tr AOA6P3IN88 AOA6P3IN88_BISBB	Q V S L N S G Y H F C G G S L I N S Q W V V S A A H C Y K S G I Q V R L G	74
<input type="checkbox"/>	tr AOA6P3IFR1 AOA6P3IFR1_BISBB	Q V S L N S G Y H F C G G S L I N S Q W V V S A A H C Y K S G I Q V R L G	74
<input type="checkbox"/>	tr AOA8B9W8U3 AOA8B9W8U3_BOSMU	Q V S L N S G Y H F C G G S L I N S Q W V V S A A H C Y K S G I Q V R L G	74
<input type="checkbox"/>	tr AOA5N3XT99 AOA5N3XT99_MUNRE	Q V S L N S G Y H F C G G S L I N S Q W V V S A A H C Y K S G I Q V R L G	74
<input type="checkbox"/>	tr W5Q1Y9 W5Q1Y9_SHEEP	Q V S L N S G Y H F C G G S L I N S R W V V S A A H C Y K S G I Q V R L G	74
<input type="checkbox"/>	tr AOA452EML0 AOA452EML0_CAPHI	Q V S L N S G Y H F C G G S L I N S Q W V V S A A H C Y K S G I Q V R L G	74
<input type="checkbox"/>	tr AOA452FA14 AOA452FA14_CAPHI	Q V S L N S G Y H F C G G S L I N S Q W V V S A A H C Y K S G I Q V R L G	74

P00760:Signal

<input type="checkbox"/>	sp P00760 TRY1_BOVIN	E D N I N V V E G N E Q F I S A S K S I V H P S Y N S N T L N N D I M L I	111
<input type="checkbox"/>	tr AOA4W2FD21 AOA4W2FD21_BOBOX	E D N I N V V E G N E Q F I S A S K S I V H P S Y N S N T L N N D I M L I	111
<input type="checkbox"/>	tr AOA6P5BKV6 AOA6P5BKV6_BOSIN	E D N I N V V E G N E Q F I S A S K S I V H P S Y N S N T L N N D I M L I	111
<input type="checkbox"/>	tr AOA6P3IN88 AOA6P3IN88_BISBB	E D N I N V V E G N E Q F I S A S K S I V H P S Y N S N T L N N D I M L I	111
<input type="checkbox"/>	tr AOA6P3IFR1 AOA6P3IFR1_BISBB	E D N I N V V E G N E Q F I S A S K S I V H P S Y N S N T L N N D I M L I	111
<input type="checkbox"/>	tr AOA8B9W8U3 AOA8B9W8U3_BOSMU	E D N I N V V E G N E Q F I S A S K S I V H P S Y N S N T L N N D I M L I	111
<input type="checkbox"/>	tr AOA5N3XT99 AOA5N3XT99_MUNRE	E D N I N V V E G N E Q F I S A S K S I V H P S Y N S N T L N N D I M L I	111
<input type="checkbox"/>	tr W5Q1Y9 W5Q1Y9_SHEEP	E D N I N V V E G N E Q F I S A S K S I V H P S Y N S N T L N N D I M L I	111
<input type="checkbox"/>	tr AOA452EML0 AOA452EML0_CAPHI	E D N I N V V E G N E Q F I S A S K S I V H P S Y N S N T L N N D I M L I	111
<input type="checkbox"/>	tr AOA452FA14 AOA452FA14_CAPHI	E D N I N V V E G N E Q F I S A S K S I V H P S Y N S N T L N N D I M L I	111

P00760:Signal

<input type="checkbox"/>	sp P00760 TRY1_BOVIN	K L K S A A S L N S R V A S I S L P T S C A S A G T Q C L I S G W G N T K	148
<input type="checkbox"/>	tr AOA4W2FD21 AOA4W2FD21_BOBOX	K L K S A A S L N S R V A S I S L P T S C A S A G T Q C L I S G W G N T K	148
<input type="checkbox"/>	tr AOA6P5BKV6 AOA6P5BKV6_BOSIN	K L K S A A S L N S R V A S I S L P T S C A S A G T Q C L I S G W G N T K	148
<input type="checkbox"/>	tr AOA6P3IN88 AOA6P3IN88_BISBB	K L K S A A S L N S R V A S I S L P T S C A S A G T Q C L I S G W G N T K	148
<input type="checkbox"/>	tr AOA6P3IFR1 AOA6P3IFR1_BISBB	K L K S A A S L N S R V A S I S L P T S C A S A G T Q C L I S G W G N T K	148
<input type="checkbox"/>	tr AOA8B9W8U3 AOA8B9W8U3_BOSMU	K L K S A A S L N S R V A S I S L P T S C A S A G T Q C L I S G W G N T K	148
<input type="checkbox"/>	tr AOA5N3XT99 AOA5N3XT99_MUNRE	K L K S A A S L N S R V A S V S L P T S C A S A G T Q C L I S G W G N T K	148
<input type="checkbox"/>	tr W5Q1Y9 W5Q1Y9_SHEEP	K L K S A A S L N S R V A S V S L P T S C A S A G T Q C L I S G W G N T K	148
<input type="checkbox"/>	tr AOA452EML0 AOA452EML0_CAPHI	K L K S A A S L N S R V A S V S L P T S C A S A G T Q C L I S G W G N T K	148
<input type="checkbox"/>	tr AOA452FA14 AOA452FA14_CAPHI	K L K S A A S L N S R V A S V S L P T S C A S A G T Q C L I S G W G N T K	148



**P00760:Signal**

<input type="checkbox"/> sp P00760 TRY1_BOVIN	S S G T S Y P D V L K C L K A P I L S D S S C K S A Y P G Q I T S N M F C	185
<input type="checkbox"/> tr AOA4W2FD21 AOA4W2FD21_BOBOX	S S G T S Y P D V L K C L K A P I L S D S S C K S A Y P G Q I T S N M F C	185
<input type="checkbox"/> tr AOA6P5BKV6 AOA6P5BKV6_BOSIN	S S G T S Y P D V L X C L K A P I L S D S S C K S A Y P G Q I T S N M F C	185
<input type="checkbox"/> tr AOA6P3IN88 AOA6P3IN88_BISBB	S S G S N Y P D V L Q C L K A P I L S D S S C K S A Y P G Q I T S N M F C	185
<input type="checkbox"/> tr AOA6P3IFR1 AOA6P3IFR1_BISBB	S S G T S Y P D V L Q C L K A P I L S D S S C K S A Y P G Q I T S N M F C	185
<input type="checkbox"/> tr AOA8B9W8U3 AOA8B9W8U3_BOSMU	S S G T S Y P N V L Q C L K A P I L S D S S C K S A Y P G Q I T S N M F C	185
<input type="checkbox"/> tr AOA5N3XT99 AOA5N3XT99_MUNRE	S S G T S Y P D V L Q C L K A P I L S D S S C S S A Y P G Q I T S N M F C	185
<input type="checkbox"/> tr W5Q1Y9 W5Q1Y9_SHEEP	S S G T S Y P D V L Q C L K A P I L S D S S C K S A Y P G Q I T S N M F C	185
<input type="checkbox"/> tr AOA452EMLO AOA452EMLO_CAPII	S S G T S Y P D V L Q C L K A P I L S D S S C K S A Y P G Q I T S N M F C	185
<input type="checkbox"/> tr AOA452FA14 AOA452FA14_CAPII	S S G S N Y P D V L Q C L K A P I L S D S S C K S A Y P G Q I T S N M F C	185

**P00760:Signal**

<input type="checkbox"/> sp P00760 TRY1_BOVIN	A G Y L E G G K D S C Q G D S G G P V V C S G K L Q G I V S W G S G C A Q	222
<input type="checkbox"/> tr AOA4W2FD21 AOA4W2FD21_BOBOX	A G Y L E G G K D S C Q G D S G G P V V C S G K L Q G I V S W G S G C A Q	222
<input type="checkbox"/> tr AOA6P5BKV6 AOA6P5BKV6_BOSIN	A G Y L E G G K D S C Q G D S G G P V V C S G K L Q G I V S W G S G C A Q	222
<input type="checkbox"/> tr AOA6P3IN88 AOA6P3IN88_BISBB	A G Y L E G G K D S C Q G D S G G P V V C S G K L Q G I V S W G Y G C A Q	222
<input type="checkbox"/> tr AOA6P3IFR1 AOA6P3IFR1_BISBB	A G Y L E G G K D S C Q G D S G G P V V C S G K L Q G I V S W G Y G C A Q	222
<input type="checkbox"/> tr AOA8B9W8U3 AOA8B9W8U3_BOSMU	A G Y L E G G K D S C Q G D S G G P V V C S G K L Q G I V S W G Y G C A Q	222
<input type="checkbox"/> tr AOA5N3XT99 AOA5N3XT99_MUNRE	A G Y L E G G K D S C Q G D S G G P V V C N G K L Q G I V S W G Y G C A Q	222
<input type="checkbox"/> tr W5Q1Y9 W5Q1Y9_SHEEP	A G Y L E G G K D S C Q G D S G G P V V C N G K L Q G I V S W G Y G C A Q	222
<input type="checkbox"/> tr AOA452EMLO AOA452EMLO_CAPII	A G Y L E G G K D S C Q G D S G G P V V C S G K L Q G I V S W G Y G C A Q	222
<input type="checkbox"/> tr AOA452FA14 AOA452FA14_CAPII	A G Y L E G G K D S C Q G D S G G P V V C S G K L Q G I V S W G Y G C A Q	222

**P00760:Signal**

<input type="checkbox"/> sp P00760 TRY1_BOVIN	K N K P G V Y T K V C N Y V S W I K Q T I A S N	246
<input type="checkbox"/> tr AOA4W2FD21 AOA4W2FD21_BOBOX	K N K P G V Y T K V C N Y V S W I K Q T I A S N	246
<input type="checkbox"/> tr AOA6P5BKV6 AOA6P5BKV6_BOSIN	K N K P G V Y T K V C N Y V S W I K Q T I A S N	246
<input type="checkbox"/> tr AOA6P3IN88 AOA6P3IN88_BISBB	K N K P G V Y T K V C N Y V S W I K Q T I A S N	246
<input type="checkbox"/> tr AOA6P3IFR1 AOA6P3IFR1_BISBB	K N K P G V Y T K V C N Y V S W I K Q T I A S N	246
<input type="checkbox"/> tr AOA8B9W8U3 AOA8B9W8U3_BOSMU	K N K P G V Y T K V C N Y V S W I K Q T I A S N	246
<input type="checkbox"/> tr AOA5N3XT99 AOA5N3XT99_MUNRE	K N K P G V Y T K V C N Y V S W I Q Q T I A S N	246
<input type="checkbox"/> tr W5Q1Y9 W5Q1Y9_SHEEP	K N K P G V Y T K V C N Y V S W I Q Q T I A S N	246
<input type="checkbox"/> tr AOA452EMLO AOA452EMLO_CAPII	K N K P G V Y T K V C N Y V S W I Q Q T I A S N	246
<input type="checkbox"/> tr AOA452FA14 AOA452FA14_CAPII	K N K P G V Y T K V C N Y V S W I Q Q T I A S N	246

**P00760:Signal**

EMBL-EBI PIR SIB

© 2002 – 2024 UniProt consortium

License &amp; Disclaimer | Privacy Notice

Core data	Supporting data	Tools	Information
Proteins (UniProtKB)	Literature citations	BLAST	Cite UniProt*
Species (Proteomes)	Taxonomy	Align	About & Help
Protein clusters (UniRef)	Keywords	Retrieve/ID mapping	UniProtKB manual
Sequence archive (UniParc)	Subcellular locations	Peptide search	Technical corner
	Cross-referenced databases	Tool results	Expert biocuration
	Diseases		Statistics

Get in touch

UniProt is an ELIXIR core

data resource



UniProt is a GBC global core biodata resource

