

Practical 6 Bioinformatics

Here is the link of all codes: [BS22B009_p6.ipynb](#)

Questions 1. Using AL2CO server, obtain the positional conservation scores from multiple sequence alignment (MSA) of given set of protein sequences (set1 and set2) using the methods given below:

- (i) Unweighted frequency and entropy-based measure
- (ii) Unweighted frequency and variance-based measure
- (iii) Unweighted frequency and sum of pairs measure
- (iv) Weighted frequency and variance-based measure
- (v) Normalize the scores obtained

with

(i) Sequences: Set 1 P69905, P01946, P01942, P01966, P01958, P01959, P01965, P06635, P60529, P80043 and P01980

Set 2 TPIS_HUMAN, TPIS_YEAST, TPIS_GRAGA, TPIS_TRYCR, TPIS_MAIZE, TPIS_MOUSE, TPIS_DROME, TPIS_RABIT, TPIS_CAEEL Instructions:













The given IDs in Set 1 were mapped using UniProt and the sequences were downloaded in FASTA format.

The screenshot shows the UniProt ID mapping web interface. The browser address bar displays 'uniprot.org/id-mapping'. The navigation bar includes links for BLAST, Align, Peptide search, ID mapping (selected), and SPARQL, along with a UniProtKB dropdown menu. The main heading is 'Retrieve/ID mapping'. Below this, instructions state: 'Enter one or more IDs (100,000 max). You may also load from a text file. Separate IDs by whitespace (space, tab, newline) or commas. This service can also be used programmatically. Alternatively, the underlying data can be downloaded.' A large text input field contains the following IDs: 'P69905, P01946, P01942, P01966, P01958, P01959, P01965, P06635, P60529, P80043, P01980'. Below the input field, a status bar indicates 'Your Input contains 11 unique IDs'. Under the 'From database' section, 'UniProtKB AC/ID' is selected. Under the 'To database' section, 'UniProtKB' is selected. At the bottom, there is a field to 'Name your ID Mapping job' with the text 'P69905 +10 UniProtKB_AC-ID → UniProtKB'.

ID mapping 11 results found for UniProtKB_AC-ID → UniProtKB

Tools ▾ Download (11) Add View: Cards ☐ Table ☒ Customize columns Resubmit Share ▾

11 IDs were mapped to 11 results

	From	Entry 	Entry Name 	Protein Names 	Gene Names 	Organism 	Length
<input type="checkbox"/>	P69905	P69905	 HBA_HUMAN	Hemoglobin subunit alpha [...]	HBA1, HBA2	Homo sapiens (Human)	142 A
<input type="checkbox"/>	P01946	P01946	 HBA_RAT	Hemoglobin subunit alpha-1/2[...]	Hba1, Hba-a1	Rattus norvegicus (Rat)	142 A
<input type="checkbox"/>	P01942	P01942	 HBA_MOUSE	Hemoglobin subunit alpha [...]	Hba, Hba-a1	Mus musculus (Mouse)	142 A
<input type="checkbox"/>	P01966	P01966	 HBA_BOVIN	Hemoglobin subunit alpha [...]	HBA	Bos taurus (Bovine)	142 A
<input type="checkbox"/>	P01958	P01958	 HBA_HORSE	Hemoglobin subunit alpha [...]	HBA	Equus caballus (Horse)	142 A
<input type="checkbox"/>	P01959	P01959	 HBA_EQUAS	Hemoglobin subunit alpha [...]	HBA1, HBA2	Equus asinus (Donkey) (Equus africanus asinus)	142 A

The FASTA file was fed into Clustal Omega to obtain the MSA. The result was downloaded in CLUSTAL format.

Results for Job ID `clustalo-I20250305-120338-0829-8683773-p1m` [Copy](#) [Resubmission](#) [Feedback](#)

Results for Job ID `clustalo-I20250305-120338-0829-8683773-p1m` [Copy](#) [Resubmission](#) [Feedback](#)

Copy Resubmission

Resubmission 

[Tool Output](#)
[Alignments](#)
[Guide Tree](#)
[Phylogenetic Tree](#)
[Results Viewers](#)
[Result Files](#)
[Submission Details](#)

[Tool Output](#)
[Alignments](#)
[Guide Tree](#)
[Phylogenetic Tree](#)
[Results Viewers](#)
[Result Files](#)
[Submission Details](#)

[Tool Output](#)
[Alignments](#)
[Guide Tree](#)
[Phylogenetic Tree](#)
[Results Viewers](#)
[Result Files](#)
[Submission Details](#)

[Tool Output](#)
[Alignments](#)
[Guide Tree](#)
[Phylogenetic Tree](#)
[Results Viewers](#)
[Result Files](#)
[Submission Details](#)

[Tool Output](#)
[Alignments](#)
[Guide Tree](#)
[Phylogenetic Tree](#)
[Results Viewers](#)
[Result Files](#)
[Submission Details](#)

[Tool Output](#)
[Alignments](#)
[Guide Tree](#)
[Phylogenetic Tree](#)
[Results Viewers](#)
[Result Files](#)
[Submission Details](#)

[Tool Output](#)
[Alignments](#)
[Guide Tree](#)
[Phylogenetic Tree](#)
[Results Viewers](#)
[Result Files](#)
[Submission Details](#)

Nightingale

COLOR SCHEME

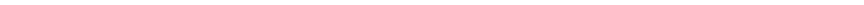
COLOR SCHEME

COLOR SCHEME



COLOR SCHEME

11 sequences



11 sequences



Sequence logo showing conservation across 11 sequences. The x-axis represents positions from 1 to 140. The y-axis represents the height of the sequence logo. The logo shows a high degree of conservation across all 11 sequences, with a peak at position 100.

Species	Sequence
SP100043HBA TREBE	-LSLDR AARALNIGKSDAINDALSIVVYFCVTFSPIDTPSPHIAHRI VGGIALVSKI DDLKTGVELSCHYAI DPA
SP101980HIBA APTFO	VILNADSNVIFSISSHAEAEATLENTFTYFCVTFPF DLHSGSADGHHKIAAALIEANHCDDIAGALKSLDLEHLOPI
SP106529HIBA CANLF	-VILPADTITSTDIGGAGHGGEALCTQSFSEVTFPF DLHSGSADGHHKIDALITVAHCHDIDGALKSLDHAHLOPI
SP101946HIBA RAT	-VILGADTITNCHGIGGGHGGEEALCNFAAFVTFPSHIDLHSGSADGHHKIDALAKAGAHVEDGALSTSLDHAHLOPI
RAT SP101965HIBA PIG	-VILGADTIAAAGVIGGGAGHGAFALENFLQFSEVTFPF DLHSGSADGHHKIDALAKAGAHVEDGALSTSLDHAHLOPI
SP101942HIBA MOUSE	VILGGETSKAAAGVIGGGAGHGAFALENFASFSEVTFPF DLHSGSADGHHKIDALASAGAHVEDGALSTSLDHAHLOPI
SP101958HIBA HORSE	VILGADTITKAAVIGGGAGHGAFALENFLQFSEVTFPF DLHSGSADGHHKIDALTLVGHCHDIDGALKSLDHAHLOPI
SP101959HIBA EQUAS	VILGADTITKAAVIGGGAGHGAFALENFLQFSEVTFPF DLHSGSADGHHKIDALTLVGHCHDIDGALKSLDHAHLOPI
SP101966HIBA BOVIN	VILGADTITKAAVIGGGAAEYGAFALENFLSFSEVTFPF DLHSGSADGHHKIDALTLKAVECHDIDGALKSLDHAHLOPI
SP106905HIBA HUMAN	VILGADTITKAAVIGGAHAGHGAFALENFLSFSEVTFPF DLHSGSADGHHKIDALTVAHCHDIDGALKSLDHAHLOPI
SP106635HIBA PONPY	VILPADTITNTAKVIGGAHAGHGAFALENFLSFSEVTFPF DLHSGSADGHHKIDALTVAHCHDIDGALKSLDHAHLOPI

Species	Sequence
SP100043HBA TREBE	-LSLDR AARALNIGKSDAINDALSIVVYFCVTFSPIDTPSPHIAHRI VGGIALVSKI DDLKTGVELSCHYAI DPA
SP101980HIBA APTFO	VILNADSNVIFSISSHAEAEATLENTFTYFCVTFPF DLHSGSADGHHKIAAALIEANHCDDIAGALKSLDLEHLOPI
SP106529HIBA CANLF	-VILPADTITSTDIGGAGHGGEALCTQSFSEVTFPF DLHSGSADGHHKIDALITVAHCHDIDGALKSLDHAHLOPI
SP101946HIBA RAT	-VILGADTITNCHGIGGGHGGEEALCNFAAFVTFPSHIDLHSGSADGHHKIDALAKAGAHVEDGALSTSLDHAHLOPI
RAT SP101965HIBA PIG	-VILGADTIAAAGVIGGGAGHGAFALENFLQFSEVTFPF DLHSGSADGHHKIDALAKAGAHVEDGALSTSLDHAHLOPI
SP101942HIBA MOUSE	VILGGETSKAAAGVIGGGAGHGAFALENFASFSEVTFPF DLHSGSADGHHKIDALASAGAHVEDGALSTSLDHAHLOPI
SP101958HIBA HORSE	VILGADTITKAAVIGGGAGHGAFALENFLQFSEVTFPF DLHSGSADGHHKIDALTLVGHCHDIDGALKSLDHAHLOPI
SP101959HIBA EQUAS	VILGADTITKAAVIGGGAGHGAFALENFLQFSEVTFPF DLHSGSADGHHKIDALTLVGHCHDIDGALKSLDHAHLOPI
SP101966HIBA BOVIN	VILGADTITKAAVIGGGAAEYGAFALENFLSFSEVTFPF DLHSGSADGHHKIDALTKAVECHDIDGALKSLDHAHLOPI
SP1069905HIBA HUMAN	VILGADTITKAAVIGGAHAGHGAFALENFLSFSEVTFPF DLHSGSADGHHKIDALTVAVHCHDIDGALKSLDHAHLOPI
SP106635HIBA PONPY	VILPADTITNTAKVIGGAHAGHGAFALENFLSFSEVTFPF DLHSGSADGHHKIDALTVAVHCHDIDGALKSLDHAHLOPI

The file in CLUSTAL format was used as input for the AL2CO server along with the necessary scoring parameters.

(i) Unweighted frequency and entropy-based measure

Used this command in terminal under my directory.

```
al2co -i bs22b009.aln-clustal_num -f 0 -c 0
```

```
al2co -i bs22b009.aln-clustal_num -f 0 -c 0 -o output_Q1_Set1_unweighted_entropy
```

```
bt3040@sudha-ThinkCe  ×  +  v
bs22b009.aln-clustal_num
bt3040@sudha-ThinkCentre-M82:~/bs22b009$ al2co -i bs22b009.aln-clustal_num -f 0 -c 0
1      -      0.943
2      S      0.253
3      L      0.943
4      S      0.943
5      D      -1.597
6      K      -1.685
7      D      0.943
8      K      0.943
9      A      -1.705
10     A      0.253
11     V      -0.385
12     R      0.253
13     A      -1.399
14     L      -1.685
15     W      0.253
16     S      -1.134
17     K      0.943
18     I      -0.618
19     G      0.253
20     K      -1.399
21     S      -1.064
22     A      -0.131
23     D      -1.399
24     A      -1.114
25     I      -1.064
26     G      0.943
27     N      -1.064
28     D      0.253
29     A      0.253
30     L      0.943
31     S      -1.064
32     R      0.943
33     M      0.253
34     I      0.253
35     V      -1.991
36     V      -2.154
37     Y      -0.131
38     P      0.943
39     Q      -0.131
40     T      0.943
41     K      0.943
42     T      0.943
43     Y      0.943
44     F      0.943
45     S      -0.131
46     H      0.943
47     W      -0.417
48     P      -1.000      *
49     D      0.253
50     V      -0.385
51     T      -0.417
52     P      -0.385
53     G      0.943
54     S      0.943
55     P      -0.417
56     H      0.253
```

```
bt3040@sudha-ThinkCe x + v
99 N 0.943
100 F 0.943
101 K 0.943
102 I -0.417
103 L 0.943
104 N -0.417
105 H 0.253
106 C 0.253
107 I -0.417
108 L 0.253
109 V -0.131
110 V -0.131
111 I -0.417
112 S 0.253
113 T -2.955
114 M 0.253
115 F -1.175
116 P 0.943
117 K -3.241
118 E -1.134
119 F 0.253
120 T 0.253
121 P 0.943
122 E -0.778
123 A -1.064
124 H 0.943
125 V 0.253
126 S 0.943
127 L 0.943
128 D 0.943
129 K 0.943
130 F 0.943
131 L 0.253
132 S -1.006
133 G -2.262
134 V 0.943
135 A -0.417
136 L -0.417
137 A 0.253
138 L 0.943
139 A 0.253
140 E 0.253
141 R 0.253
142 Y 0.943
143 R 0.943
* gap fraction no less than 0.50; conservation set to M-S
M: mean; S: standard deviation

al2co - The parameters are:
Input alignment file - bs22b009.aln-clustal_num
Output conservation - STDOUT
Weighting scheme - unweighted
Conservation calculation method - entropy-based
Window size - 1
Conservation normalized to zero mean and unity variance
Gap fraction to suppress calculation - 0.50
bt3040@sudha-ThinkCentre-M82:~/bs22b009$ |
```

(ii) Unweighted frequency and variance-based measure

```
al2co -i bs22b009.aln-clustal_num -f 0 -c 1
```

```
al2co -i bs22b009.aln-clustal_num -f 0 -c 1 -o output_Q1_Set1_unweighted_variance
```

```
bt3040@sudha-ThinkCe  ×  +  ▾
bt3040@sudha-ThinkCentre-M82:~/bs22b009$ al2co -i bs22b009.aln-clustal_num -f 0 -c 1
1      -      1.228
2      S      0.275
3      L      0.546
4      S      0.816
5      D     -1.795
6      K     -1.584
7      D      0.938
8      K      0.832
9      A     -1.707
10     A      0.592
11     V     -0.467
12     R      0.314
13     A     -1.550
14     L     -1.596
15     W      0.750
16     S     -1.273
17     K      0.832
18     I     -0.788
19     G      0.401
20     K     -1.266
21     S     -0.656
22     A     -0.493
23     D     -1.200
24     A     -0.948
25     I     -0.385
26     G      0.958
27     N     -1.040
28     D      0.623
29     A     -0.020
30     L      0.546
31     S     -0.435
32     R      1.195
33     M      0.690
34     I      0.516
35     V     -2.145
36     V     -2.129
37     Y      0.062
38     P      1.062
39     Q      0.006
40     T      0.958
41     K      0.832
42     T      0.958
43     Y      1.199
44     F      1.030
45     S      0.010
46     H      0.914
47     W      0.016
48     P     -1.000      *
49     D      0.417
50     V     -0.897
51     T     -0.272
52     P     -0.457
53     G      0.958
54     S      0.816
55     P     -0.552
56     H      0.713
```

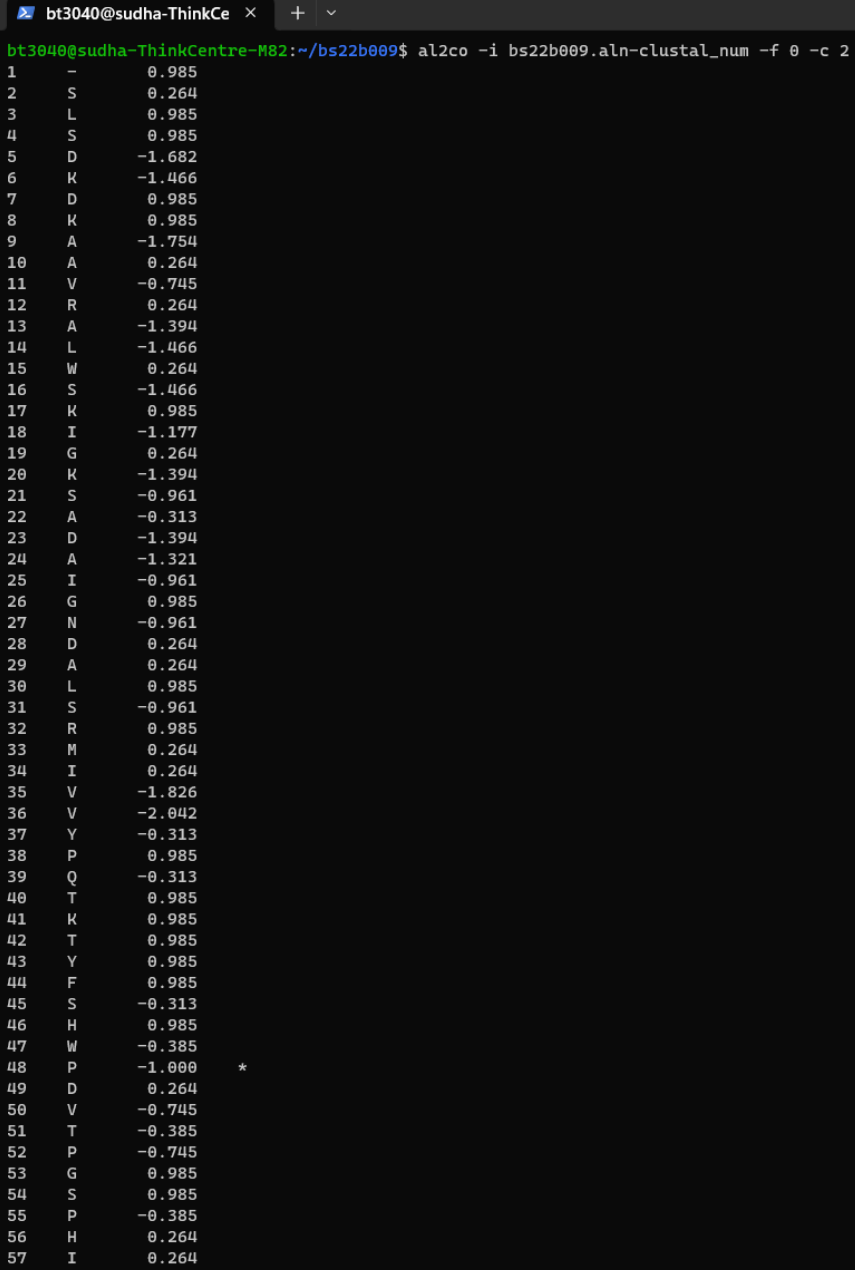
```
bt3040@sudha-ThinkCe × + v
99 N 1.176
100 F 1.030
101 K 0.832
102 I -0.496
103 L 0.546
104 N -0.242
105 H 0.402
106 C 0.719
107 I -0.496
108 L 0.032
109 V -0.218
110 V -0.089
111 I -0.520
112 S -0.036
113 T -2.908
114 M 0.399
115 F -1.596
116 P 1.062
117 K -3.173
118 E -1.145
119 F 0.448
120 T 0.437
121 P 1.062
122 E -0.983
123 A -0.767
124 H 0.914
125 V -0.034
126 S 0.816
127 L 0.546
128 D 0.938
129 K 0.832
130 F 1.030
131 L 0.009
132 S -1.381
133 G -2.209
134 V 0.832
135 A -0.315
136 L -0.144
137 A 0.246
138 L 0.546
139 A 0.373
140 E 0.295
141 R 0.314
142 Y 1.199
143 R 1.195
* gap fraction no less than 0.50; conservation set to M-S
M: mean; S: standard deviation

al2co - The parameters are:
Input alignment file - bs22b009.aln-clustal_num
Output conservation - STDOUT
Weighting scheme - unweighted
Conservation calculation method - variance-based
Window size - 1
Conservation normalized to zero mean and unity variance
Gap fraction to suppress calculation - 0.50
bt3040@sudha-ThinkCentre-M82:~/bs22b009$ |
```

(iii) Unweighted frequency and sum of pairs measure

```
al2co -i bs22b009.aln-clustal_num -f 0 -c 2
```

```
al2co -i bs22b009.aln-clustal_num -f 0 -c 2 -o output_Q1_Set1_unweighted_sum
```



The terminal window shows the command `al2co -i bs22b009.aln-clustal_num -f 0 -c 2` being executed. The output is a list of 57 items, each consisting of a number, a single-letter amino acid code, and a numerical value. The values range from -2.042 to 0.985. Item 48, 'P', has a value of -1.000 and is followed by an asterisk.

Item	Code	Value
1	-	0.985
2	S	0.264
3	L	0.985
4	S	0.985
5	D	-1.682
6	K	-1.466
7	D	0.985
8	K	0.985
9	A	-1.754
10	A	0.264
11	V	-0.745
12	R	0.264
13	A	-1.394
14	L	-1.466
15	W	0.264
16	S	-1.466
17	K	0.985
18	I	-1.177
19	G	0.264
20	K	-1.394
21	S	-0.961
22	A	-0.313
23	D	-1.394
24	A	-1.321
25	I	-0.961
26	G	0.985
27	N	-0.961
28	D	0.264
29	A	0.264
30	L	0.985
31	S	-0.961
32	R	0.985
33	M	0.264
34	I	0.264
35	V	-1.826
36	V	-2.042
37	Y	-0.313
38	P	0.985
39	Q	-0.313
40	T	0.985
41	K	0.985
42	T	0.985
43	Y	0.985
44	F	0.985
45	S	-0.313
46	H	0.985
47	W	-0.385
48	P	-1.000 *
49	D	0.264
50	V	-0.745
51	T	-0.385
52	P	-0.745
53	G	0.985
54	S	0.985
55	P	-0.385
56	H	0.264
57	I	0.264

bt3040@sudha-ThinkCe		
86	S	0.985
87	E	0.264
88	Q	0.264
89	H	0.985
90	A	0.985
91	Y	-0.889
92	K	0.985
93	L	0.985
94	R	0.985
95	V	0.985
96	D	0.985
97	P	0.985
98	A	0.264
99	N	0.985
100	F	0.985
101	K	0.985
102	I	-0.385
103	L	0.985
104	N	-0.385
105	H	0.264
106	C	0.264
107	I	-0.385
108	L	0.264
109	V	-0.313
110	V	-0.313
111	I	-0.385
112	S	0.264
113	T	-2.547
114	M	0.264
115	F	-1.538
116	P	0.985
117	K	-2.619
118	E	-1.466
119	F	0.264
120	T	0.264
121	P	0.985
122	E	-0.889
123	A	-0.961
124	H	0.985
125	V	0.264
126	S	0.985
127	L	0.985
128	D	0.985
129	K	0.985
130	F	0.985
131	L	0.264
132	S	-1.249
133	G	-2.114
134	V	0.985
135	A	-0.385
136	L	-0.385
137	A	0.264
138	L	0.985
139	A	0.264
140	E	0.264
141	R	0.264
142	Y	0.985
143	R	0.985

(iv) Weighted frequency and variance-based measure

al2co -i bs22b009.aln-clustal_num -f 1 -c 1

Or for file:

al2co -i bs22b009.aln-clustal_num -f 1 -c 1 -o output_Q1_Set1_Weighted_Variance


```
bt3040@sudha-ThinkCe  ×  +  ∨
bt3040@sudha-ThinkCentre-M82:~/bs22b009$ al2co -i bs22b009.aln-clustal_num -f 1 -c 1
1      -      1.323
2      S      -0.218
3      L      0.775
4      S      0.965
5      D      -1.478
6      K      -1.760
7      D      1.077
8      K      0.957
9      A      -1.623
10     A      -0.002
11     V      -0.036
12     R      -0.138
13     A      -1.085
14     L      -1.887
15     W      0.629
16     S      -0.881
17     K      0.957
18     I      -0.273
19     G      0.325
20     K      -1.425
21     S      -0.758
22     A      -0.016
23     D      -1.345
24     A      -0.939
25     I      -0.485
26     G      1.101
27     N      -1.111
28     D      0.077
29     A      -0.010
30     L      0.775
31     S      -0.754
32     R      1.284
33     M      0.928
34     I      0.055
35     V      -2.110
36     V      -2.043
37     Y      -0.277
38     P      1.186
39     Q      -0.290
40     T      1.113
41     K      0.957
42     T      1.113
43     Y      1.290
44     F      1.166
45     S      -0.309
46     H      1.076
47     W      -0.315
48     P      -1.000      *
49     D      0.696
50     V      -0.789
51     T      -0.878
52     P      -0.432
53     G      1.101
54     S      0.965
55     P      -0.753
56     H      0.163
57     I      -0.115
```

86	S	0.965
87	E	-0.035
88	Q	-0.290
89	H	1.076
90	A	0.739
91	Y	-0.809
92	K	0.957
93	L	0.775
94	R	1.284
95	V	0.987
96	D	1.077
97	P	1.186
98	A	-0.292
99	N	1.281
100	F	1.166
101	K	0.957
102	I	-0.732
103	L	0.775
104	N	-0.803
105	H	0.368
106	C	1.020
107	I	-0.966
108	L	0.069
109	V	0.395
110	V	-0.495
111	I	-1.011
112	S	-0.462
113	T	-2.205
114	M	-0.007
115	F	-1.282
116	P	1.186
117	K	-2.507
118	E	-0.989
119	F	0.358
120	T	0.732
121	P	1.186
122	E	-0.964
123	A	-1.336
124	H	1.076
125	V	-0.454
126	S	0.965
127	L	0.775
128	D	1.077
129	K	0.957
130	F	1.166
131	L	0.368
132	S	-1.284
133	G	-1.969
134	V	0.987
135	A	-1.024
136	L	-0.823
137	A	-0.292
138	L	0.775
139	A	-0.167
140	E	-0.141
141	R	-0.138
142	Y	1.290
143	R	1.284

(v) Normalize the scores obtained with (i)

```
al2co -i bs22b009.aln-clustal_num -f 0 -c 0 -n 1 -o output_Q1_Set1_Normalise_score
```

For Set2

Set 2 TPIS_HUMAN, TPIS_YEAST, TPIS_GRAGA, TPIS_TRYCR, TPIS_MAIZE,
TPIS_MOUSE, TPIS_DROME, TPIS_RABIT, TPIS_CAEEL

ID mapping 9 results found for UniProtKB_AC-ID → UniProtKB

Overview	Input Parameters	API Request
----------	------------------	-------------

Tools
Download (9)
Add
View:
Cards
Table
Customize columns
Resubmit
Share

9 IDs were mapped to 9 results

From	Entry	Entry Name	Protein Names	Gene Names	Organism	Length
<input type="checkbox"/> TPIS_HUMAN	P60174	TPIS_HUMAN	Triosephosphate isomerase [...]	TP1I, TPI	Homo sapiens (Human)	249 AA
<input type="checkbox"/> TPIS_YEAST	P00942	TPIS_YEAST	Triosephosphate isomerase [...]	TP1I, YDR050C, YD9609.05C	Saccharomyces cerevisiae (strain ATCC 204508 / S288c) (Baker's yeast)	248 AA
<input type="checkbox"/> TPIS_GRAGA	P48492	TPIS_GRAGA	Triosephosphate isomerase, cytosolic[...]	TP1I	Gracilaria gracilis (Red alga)	250 AA
<input type="checkbox"/> TPIS_TRYCR	P52270	TPIS_TRYCR	Triosephosphate isomerase, glycosomal[...]		Trypanosoma cruzi	251 AA
<input type="checkbox"/> TPIS_MAIZE	P12863	TPIS_MAIZE	Triosephosphate isomerase, cytosolic[...]		Zea mays (Maize)	253 AA
<input type="checkbox"/> TPIS_MOUSE	P17751	TPIS_MOUSE	Triosephosphate isomerase [...]	Tpi1, Tpi	Mus musculus (Mouse)	249 AA

Results for Job ID

clustalo-I20250307-185542-0592-95769104-p1m

Copy

Resubmission



[Tool Output](#)
[Alignments](#)
[Guide Tree](#)
[Phylogenetic Tree](#)
[Results Viewers](#)
[Result Files](#)
[Submission Details](#)

Nightingale

COLOR SCHEME

clustal2

LEGEND

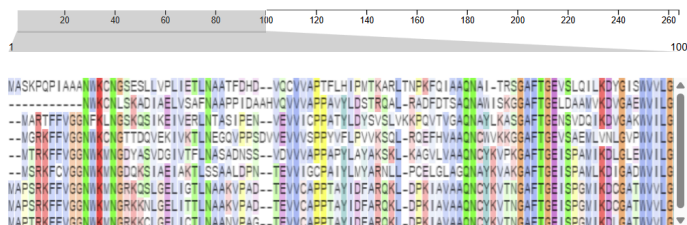
ARNDCEGHILKMFPSTWYVBXZ



9 sequences



SPIP52270|TPIS_TRYCR
SPIP48492|TPIS_GRAGA
SPIP00942|TPIS_YEAST
SPIP12863|TPIS_MAIZE
SPIQ10657|TPIS_CAEL
SPIP29613|TPIS_DROME
SPIP60174|TPIS_HUMAN
SPIP00939|TPIS_RABIT
SPIP17751|TPIS_MOUSE



```
(base) PS C:\Users\aryan> scp C:/Users/aryan/Downloads/set2.aln-clustal_num bt3040@10.21.49.164:bs22b009
bt3040@10.21.49.164's password:
set2.aln-clustal_num
(base) PS C:\Users\aryan> |
```

Remove the first line from the files:

```
GNU nano 2.9.3
sp|P52270|TPIS_TRYCR      MASKPQPIAAANWKCNGSESLVPLIETLNAATFDHD--VQCVVAPTFLHIPMTKARLTN  58
sp|P48492|TPIS_GRAGA      -----NWKCNLKADIAELVSFAFNAAPPIDAAHVQVVVAPPVLDSTRQAL-R  48
sp|P00942|TPIS_YEAST      --MARTFFVGGNFKLNGSKQSIKEIVERLNTASIPEN--VEVVICPPATYLDYSVSLVKK  56
sp|P12863|TPIS_MAIZE      --MGRKFFVGGNWKNGTDDQVEKIVKTLNEGQVPPSDVVEVVVSPPYVFLPVVKSQ-L  57
sp|Q10657|TPIS_CAEEL      --MTRKFFVGGNWKMGDYASVDGIVTFLNASADNSS--VDVVVAPPAPYLAYAKSK-L  55
sp|P29613|TPIS_DROME      --MSRKFCVGGNWKMGDQKSIKIAETLSSAALDPN--TEVVIGCPAIYLYARNLL-P  55
sp|P60174|TPIS_HUMAN      MAPSRKFFVGGNWKMGGRKQSLGELIGTLNAAKVPAD--TEVVCAPPTAYIDFARQKL-D  57
sp|P00939|TPIS_RABIT      MAPSRKFFVGGNWKMGGRKKNLDELITTLNAAKVPAD--TEVVCAPPTAYIDFARQKL-D  57
sp|P17751|TPIS_MOUSE      MAPTRKFFVGGNWKMGGRKKCLGELICTLNAANVPAG--TEVVCAPPTAYIDFARQKL-D  57
                                     *:* *   : :  . .   . : *   . :   :
sp|P52270|TPIS_TRYCR      PKFQIAAQNAI-TRSGAFTGEVSLQILKDYGISWVVLGHSERRLY--YGETNEIVAELKVA  115
sp|P48492|TPIS_GRAGA      ADFDTSQANAWISKGAFTGELDAAMVVDVGAEWVILGHSERRHIAQLKESDHTIAMKAA  108
sp|P00942|TPIS_YEAST      PQVTGVAQNAYLKASGAFTGENSDQIKDVGAKWVILGHSERRSY--FHEDDKFIADKTK  114
sp|P12863|TPIS_MAIZE      QEFHVAQAQNCWVKKGGAFTGEVSAEMLVNLGVPVILGHSERRAL--LGESNEFVGDKVA  115
sp|Q10657|TPIS_CAEEL      AGVLVAAQNCYKVPKGAFTGEISPAMIKDLGLEWVILGHSERRHV--FGESDALIAEKT  113
sp|P29613|TPIS_DROME      CELGLAGQNAVYKVAKGAFTGEISPAMLDKIDGADWVILGHSERRAI--FGESDALIAEK  113
sp|P60174|TPIS_HUMAN      PKIAVAAQNCYKVTNGAFTGEISPGMIKDCGATWVVLGHSERRHV--FGESDELIGQKVA  115
sp|P00939|TPIS_RABIT      PKIAVAAQNCYKVTNGAFTGEISPGMIKDCGATWVVLGHSERRHV--FGESDELIGQKVA  115
sp|P17751|TPIS_MOUSE      PKIAVAAQNCYKVTNGAFTGEISPGMIKDLGATWVVLGHSERRHV--FGESDELIGQKVS  115
                                     . . ** .   ***** .   : : *   ** :*****   * :   . . * .
sp|P52270|TPIS_TRYCR      QACA-AGFHVIVCVGETNEEREAGRTAAVLTQLAAVAQKLSKEAWSRVVIAYEPVWAIG  174
sp|P48492|TPIS_GRAGA      YALQHASLGVIYICIGELLEERESGQTIACVCEQLQALSDAI--SDWSDVVIAYEPVWAIG  166
sp|P00942|TPIS_YEAST      FALG-QGVGVILCIGETLEEKKAGKGLDVVERQLNAVLEEV--KDWTVNVVAYEPVWAIG  171
sp|P12863|TPIS_MAIZE      YALS-QGLKVIACVGETLEQREAGSTHDDVAAQTKAIAEKI--KDSNVVVAYEPVWAIG  172
sp|Q10657|TPIS_CAEEL      HALA-AGIKVVFICIGEKLEEREAGHTKDVNFRQLQAIVDKG--VSWENIVIAVYEPVWAIG  170
sp|P29613|TPIS_DROME      HALA-EGLKVIACIGETLEEREAGTNEVVARQMCAYAQKI--KDWKNVVVAYEPVWAIG  170
sp|P60174|TPIS_HUMAN      HALA-EGLGVIACIGEKLDEREAGITEKVVFEQTKVIADNV--KDWSKVVLAYEPVWAIG  172
sp|P00939|TPIS_RABIT      HALS-EGLGVIACIGEKLDEREAGITEKVVFEQTKVIADNV--KDWSKVVLAYEPVWAIG  172
sp|P17751|TPIS_MOUSE      HALA-EGLGVIACIGEKLDEREAGITEKVVFEQTKVIADNV--KDWSKVVLAYEPVWAIG  172
                                     *   . . * :***   : : : * *   *   .   :   * :*****
sp|P52270|TPIS_TRYCR      TGKVATPQQAQEVHELLRRWVRSKLGTDIAAQLRILYGGSVTAKNARTLYQMRDINGFLV  234
sp|P48492|TPIS_GRAGA      TGKVATPEQAQEVHEAVRAWLANNVSPQVAASTRILYGGSVSPANCESLAKQPNIDGFLV  226
sp|P00942|TPIS_YEAST      TGLAATPEDAQDIHASIRKFLASKLGDKAASELRILYGGSSANGSNAVTFKDKADVDFGLV  231
sp|P12863|TPIS_MAIZE      TGKVATPAQAQEVHASLRDWLKTNASPEVAESTRIIYGGSVTAANCKELAAQPDVDGFLV  232
sp|Q10657|TPIS_CAEEL      TGKTASGEQAQEVHEWIRAFLEKVKSPAVADATRIIYGGSVTADNAELGKKPDIDGFLV  230
sp|P29613|TPIS_DROME      TGQTATPDQAQEVHAFRLRWLSDNISKEVSASLRIQYGGSVTAANAKELAKKPDIDGFLV  230
sp|P60174|TPIS_HUMAN      TGKTATPQQAQEVHEKLRGWLKSNVSDAVAQSTRIIYGGSVTGATCKELASQPDVDGFLV  232
sp|P00939|TPIS_RABIT      TGKTATPQQAQEVHEKLRGWLKSNVSDAVAQSTRIIYGGSVTGATCKELASQPDVDGFLV  232
sp|P17751|TPIS_MOUSE      TGKTATPQQAQEVHEKLRGWLKSNVNDGVAQSTRIIYGGSVTGATCKELASQPDVDGFLV  232
                                     ** .* : :*** * * : :   :   :   ** ***** . .   :   :*****
```

(i) Unweighted frequency and entropy-based measure

al2co -i set2.aln-clustal_num -f 0 -c 0 -o output_Q1_Set2_unweighted_entropy

(ii) Unweighted frequency and variance-based measure

al2co -i set2.aln-clustal_num -f 0 -c 1 -o output_Q1_Set2_unweighted_varience

(iii) Unweighted frequency and sum of pairs measure

al2co -i set2.aln-clustal_num -f 0 -c 2 -o output_Q1_Set2_unweighted_sumpairs

(iv) Weighted frequency and variance-based measure

```
al2co -i set2.aln-clustal_num -f 1 -c 1 -o output_Q1_Set2_weighted_variance
```

(v) Normalize the scores obtained

```
al2co -i set2.aln-clustal_num -f 1 -c 1 -n 1 -o output_Q1_Set2_normalise
```

2. Tabulate the topmost 10 residues with the highest and lowest conservation scores (in both Set1 and Set 2) obtained with method (i).

Highest Conservation Scores: These refer to the **most positive values** (closer to 1 or above), meaning these residues are **highly conserved** (less variability).

Lowest Conservation Scores: These refer to the **most negative values** (closer to -3 or lower), meaning these residues are **least conserved** (high variability).

```
(base) PS C:\Users\aryan> scp bt3040@10.21.49.164:bs22b009/output_Q1_Set1_unweighted_entropy C:/Users/aryan/Downloads/  
bt3040@10.21.49.164's password:  
output_Q1_Set1_unweighted_entropy  
(base) PS C:\Users\aryan> |
```

Highest conservation scores Set1

```
al2co -i bs22b009.aln-clustal_num -f 0 -c 0
```

Rank	Residue	Position	Score
1	L	127	0.943
2	A	90	0.943
3	K	92	0.943
4	L	103	0.943
5	H	124	0.943
6	F	130	0.943
7	L	138	0.943
8	Y	142	0.943
9	R	143	0.943
10	P	97	0.943

Lowest conservation scores Set1

Rank	Residue	Position	Score
1	K	117	-3.241
2	T	113	-2.955
3	L	70	-2.848
4	S	73	-2.67
5	G	133	-2.262
6	E	84	-2.262
7	V	36	-2.154
8	V	35	-1.991
9	D	5	-1.597
10	L	14	-1.685

Highest conservation scores Set2

al2co -i set2.aln-clustal_num -f 0 -c 0

Position	Residue	Conservation Score
12	N	1.215
14	K	1.215
16	N	1.215
43	V	1.215
68	Q	1.215
69	N	1.215
76	G	1.215
77	A	1.215
78	F	1.215
79	T	1.215

Lowest conservation scores Set2

Position	Residue	Conservation Score
33	T	-2.227
64	Q	-2.121
147	A	-2.121
23	V	-1.841
36	H	-1.841
49	L	-1.841
93	S	-1.841
196	L	-1.841
199	R	-1.841
208	D	-1.841

Code:

```
BS22B009_p6.pybnc ☆ ☰
File Edit View Insert Runtime Tools Help
Q Commands | + Code + Text

import numpy as np
import pandas as pd

blosum62 = {
    ('W', 'F'): 1, ('L', 'R'): -2, ('S', 'P'): -1, ('V', 'T'): 0,
    ('Q', 'Q'): 5, ('N', 'A'): -2, ('Z', 'Y'): -2, ('W', 'R'): -3,
    ('Q', 'A'): -1, ('S', 'D'): 0, ('H', 'H'): 8, ('S', 'I'): -1,
    ('H', 'D'): -1, ('L', 'N'): -3, ('W', 'A'): -3, ('Y', 'M'): -1,
    ('G', 'R'): -2, ('Y', 'I'): -1, ('Y', 'E'): -2, ('B', 'V'): -3,
    ('Y', 'A'): -2, ('V', 'D'): -3, ('B', 'S'): 0, ('Y', 'Y'): 7,
    ('G', 'N'): 0, ('E', 'C'): -4, ('Y', 'Q'): -1, ('Z', 'Z'): 4,
    ('V', 'A'): 0, ('C', 'C'): 9, ('M', 'R'): -1, ('V', 'E'): -2,
    ('T', 'N'): 0, ('P', 'P'): 7, ('V', 'I'): 3, ('V', 'S'): -2,
    ('Z', 'P'): -1, ('V', 'M'): 1, ('T', 'F'): -2, ('V', 'Q'): -2,
    ('K', 'K'): 5, ('P', 'D'): -1, ('I', 'H'): -3, ('I', 'D'): -3,
    ('T', 'R'): -1, ('P', 'L'): -3, ('K', 'G'): -2, ('M', 'N'): -2,
    ('P', 'H'): -2, ('F', 'Q'): -3, ('Z', 'G'): -2, ('X', 'L'): -1,
    ('T', 'M'): -1, ('Z', 'C'): -3, ('X', 'H'): -1, ('D', 'R'): -2,
    ('B', 'W'): -4, ('X', 'D'): -1, ('Z', 'K'): 1, ('F', 'A'): -2,
    ('Z', 'W'): -3, ('F', 'E'): -3, ('D', 'N'): 1, ('B', 'K'): 0,
    ('X', 'X'): -1, ('F', 'I'): 0, ('B', 'G'): -1, ('X', 'T'): 0,
    ('F', 'H'): 0, ('B', 'C'): -3, ('Z', 'I'): -3, ('Z', 'V'): -2,
    ('S', 'S'): 4, ('L', 'Q'): -2, ('W', 'E'): -3, ('Q', 'R'): 1,
    ('N', 'N'): 6, ('W', 'M'): -1, ('Q', 'C'): -3, ('W', 'I'): -3,
    ('S', 'C'): -1, ('L', 'A'): -1, ('S', 'G'): 0, ('L', 'E'): -3,
    ('W', 'Q'): -2, ('H', 'G'): -2, ('S', 'K'): 0, ('Q', 'N'): 0,
    ('N', 'R'): 0, ('H', 'C'): -3, ('Y', 'N'): -2, ('G', 'Q'): -2,
    ('Y', 'F'): 3, ('C', 'A'): 0, ('V', 'L'): 1, ('G', 'E'): -2,
    ('G', 'A'): 0, ('K', 'R'): 2, ('E', 'D'): 2, ('Y', 'R'): -2,
    ('M', 'Q'): 0, ('T', 'I'): -1, ('C', 'D'): -3, ('V', 'F'): -1,
    ('T', 'A'): 0, ('T', 'P'): -1, ('B', 'P'): -2, ('T', 'E'): -1,
    ('V', 'W'): -3, ('P', 'G'): -2, ('M', 'A'): -1, ('K', 'H'): -1,
    ('V', 'R'): -3, ('P', 'C'): -3, ('M', 'E'): -2, ('K', 'L'): -2,
    ('V', 'V'): 4, ('M', 'I'): 1, ('T', 'Q'): -1, ('T', 'G'): -4,
    ('P', 'K'): -1, ('M', 'M'): 5, ('K', 'D'): -1, ('I', 'C'): -1,
    ('P', 'L'): -3, ('L', 'L'): 4, ('L', 'I'): -3, ('L', 'V'): -2,
    ('I', 'I'): 4, ('I', 'V'): -2, ('I', 'T'): -3, ('I', 'R'): -3,
    ('I', 'D'): -3, ('I', 'N'): -3, ('I', 'S'): -3, ('I', 'M'): -3,
    ('I', 'H'): -3, ('I', 'G'): -3, ('I', 'Q'): -3, ('I', 'K'): -3,
    ('I', 'W'): -3, ('I', 'X'): -3, ('I', 'Y'): -3, ('I', 'Z'): -3,
    ('I', 'A'): -3, ('I', 'C'): -3, ('I', 'E'): -3, ('I', 'G'): -3,
    ('I', 'H'): -3, ('I', 'I'): -3, ('I', 'K'): -3, ('I', 'L'): -3,
    ('I', 'M'): -3, ('I', 'N'): -3, ('I', 'O'): -3, ('I', 'P'): -3,
    ('I', 'Q'): -3, ('I', 'R'): -3, ('I', 'S'): -3, ('I', 'T'): -3,
    ('I', 'U'): -3, ('I', 'V'): -3, ('I', 'W'): -3, ('I', 'X'): -3,
    ('I', 'Y'): -3, ('I', 'Z'): -3, ('I', 'A'): -3, ('I', 'C'): -3,
    ('I', 'E'): -3, ('I', 'G'): -3, ('I', 'H'): -3, ('I', 'I'): -3,
    ('I', 'K'): -3, ('I', 'L'): -3, ('I', 'M'): -3, ('I', 'N'): -3,
    ('I', 'O'): -3, ('I', 'P'): -3, ('I', 'Q'): -3, ('I', 'R'): -3,
    ('I', 'S'): -3, ('I', 'T'): -3, ('I', 'U'): -3, ('I', 'V'): -3,
    ('I', 'W'): -3, ('I', 'X'): -3, ('I', 'Y'): -3, ('I', 'Z'): -3,
    ('I', 'A'): -3, ('I', 'C'): -3, ('I', 'E'): -3, ('I', 'G'): -3,
    ('I', 'H'): -3, ('I', 'I'): -3, ('I', 'K'): -3, ('I', 'L'): -3,
    ('I', 'M'): -3, ('I', 'N'): -3, ('I', 'O'): -3, ('I', 'P'): -3,
    ('I', 'Q'): -3, ('I', 'R'): -3, ('I', 'S'): -3, ('I', 'T'): -3,
    ('I', 'U'): -3, ('I', 'V'): -3, ('I', 'W'): -3, ('I', 'X'): -3,
    ('I', 'Y'): -3, ('I', 'Z'): -3, ('I', 'A'): -3, ('I', 'C'): -3,
    ('I', 'E'): -3, ('I', 'G'): -3, ('I', 'H'): -3, ('I', 'I'): -3,
    ('I', 'K'): -3, ('I', 'L'): -3, ('I', 'M'): -3, ('I', 'N'): -3,
    ('I', 'O'): -3, ('I', 'P'): -3, ('I', 'Q'): -3, ('I', 'R'): -3,
    ('I', 'S'): -3, ('I', 'T'): -3, ('I', 'U'): -3, ('I', 'V'): -3,
    ('I', 'W'): -3, ('I', 'X'): -3, ('I', 'Y'): -3, ('I', 'Z'): -3,
    ('I', 'A'): -3, ('I', 'C'): -3, ('I', 'E'): -3, ('I', 'G'): -3,
    ('I', 'H'): -3, ('I', 'I'): -3, ('I', 'K'): -3, ('I', 'L'): -3,
    ('I', 'M'): -3, ('I', 'N'): -3, ('I', 'O'): -3, ('I', 'P'): -3,
    ('I', 'Q'): -3, ('I', 'R'): -3, ('I', 'S'): -3, ('I', 'T'): -3,
    ('I', 'U'): -3, ('I', 'V'): -3, ('I', 'W'): -3, ('I', 'X'): -3,
    ('I', 'Y'): -3, ('I', 'Z'): -3, ('I', 'A'): -3, ('I', 'C'): -3,
    ('I', 'E'): -3, ('I', 'G'): -3, ('I', 'H'): -3, ('I', 'I'): -3,
    ('I', 'K'): -3, ('I', 'L'): -3, ('I', 'M'): -3, ('I', 'N'): -3,
    ('I', 'O'): -3, ('I', 'P'): -3, ('I', 'Q'): -3, ('I', 'R'): -3,
    ('I', 'S'): -3, ('I', 'T'): -3, ('I', 'U'): -3, ('I', 'V'): -3,
    ('I', 'W'): -3, ('I', 'X'): -3, ('I', 'Y'): -3, ('I', 'Z'): -3,
    ('I', 'A'): -3, ('I', 'C'): -3, ('I', 'E'): -3, ('I', 'G'): -3,
    ('I', 'H'): -3, ('I', 'I'): -3, ('I', 'K'): -3, ('I', 'L'): -3,
    ('I', 'M'): -3, ('I', 'N'): -3, ('I', 'O'): -3, ('I', 'P'): -3,
    ('I', 'Q'): -3, ('I', 'R'): -3, ('I', 'S'): -3, ('I', 'T'): -3,
    ('I', 'U'): -3, ('I', 'V'): -3, ('I', 'W'): -3, ('I', 'X'): -3,
    ('I', 'Y'): -3, ('I', 'Z'): -3, ('I', 'A'): -3, ('I', 'C'): -3,
    ('I', 'E'): -3, ('I', 'G'): -3, ('I', 'H'): -3, ('I', 'I'): -3,
    ('I', 'K'): -3, ('I', 'L'): -3, ('I', 'M'): -3, ('I', 'N'): -3,
    ('I', 'O'): -3, ('I', 'P'): -3, ('I', 'Q'): -3, ('I', 'R'): -3,
    ('I', 'S'): -3, ('I', 'T'): -3, ('I', 'U'): -3, ('I', 'V'): -3,
    ('I', 'W'): -3, ('I', 'X'): -3, ('I', 'Y'): -3, ('I', 'Z'): -3,
    ('I', 'A'): -3, ('I', 'C'): -3, ('I', 'E'): -3, ('I', 'G'): -3,
    ('I', 'H'): -3, ('I', 'I'): -3, ('I', 'K'): -3, ('I', 'L'): -3,
    ('I', 'M'): -3, ('I', 'N'): -3, ('I', 'O'): -3, ('I', 'P'): -3,
    ('I', 'Q'): -3, ('I', 'R'): -3, ('I', 'S'): -3, ('I', 'T'): -3,
    ('I', 'U'): -3, ('I', 'V'): -3, ('I', 'W'): -3, ('I', 'X'): -3,
    ('I', 'Y'): -3, ('I', 'Z'): -3, ('I', 'A'): -3, ('I', 'C'): -3,
    ('I', 'E'): -3, ('I', 'G'): -3, ('I', 'H'): -3, ('I', 'I'): -3,
    ('I', 'K'): -3, ('I', 'L'): -3, ('I', 'M'): -3, ('I', 'N'): -3,
    ('I', 'O'): -3, ('I', 'P'): -3, ('I', 'Q'): -3, ('I', 'R'): -3,
    ('I', 'S'): -3, ('I', 'T'): -3, ('I', 'U'): -3, ('I', 'V'): -3,
    ('I', 'W'): -3, ('I', 'X'): -3, ('I', 'Y'): -3, ('I', 'Z'): -3,
    ('I', 'A'): -3, ('I', 'C'): -3, ('I', 'E'): -3, ('I', 'G'): -3,
    ('I', 'H'): -3, ('I', 'I'): -3, ('I', 'K'): -3, ('I', 'L'): -3,
    ('I', 'M'): -3, ('I', 'N'): -3, ('I', 'O'): -3, ('I', 'P'): -3,
    ('I', 'Q'): -3, ('I', 'R'): -3, ('I', 'S'): -3, ('I', 'T'): -3,
    ('I', 'U'): -3, ('I', 'V'): -3, ('I', 'W'): -3, ('I', 'X'): -3,
    ('I', 'Y'): -3, ('I', 'Z'): -3, ('I', 'A'): -3, ('I', 'C'): -3,
    ('I', 'E'): -3, ('I', 'G'): -3, ('I', 'H'): -3, ('I', 'I'): -3,
    ('I', 'K'): -3, ('I', 'L'): -3, ('I', 'M'): -3, ('I', 'N'): -3,
    ('I', 'O'): -3, ('I', 'P'): -3, ('I', 'Q'): -3, ('I', 'R'): -3,
    ('I', 'S'): -3, ('I', 'T'): -3, ('I', 'U'): -3, ('I', 'V'): -3,
    ('I', 'W'): -3, ('I', 'X'): -3, ('I', 'Y'): -3, ('I', 'Z'): -3,
    ('I', 'A'): -3, ('I', 'C'): -3, ('I', 'E'): -3, ('I', 'G'): -3,
    ('I', 'H'): -3, ('I', 'I'): -3, ('I', 'K'): -3, ('I', 'L'): -3,
    ('I', 'M'): -3, ('I', 'N'): -3, ('I', 'O'): -3, ('I', 'P'): -3,
    ('I', 'Q'): -3, ('I', 'R'): -3, ('I', 'S'): -
```


Result:

	Residue	Entropy	Variance	Sum of Pairs
0	-	0.0	0.984	5.0
1	S	-0.305	0.829	3.008
2	L	0.0	0.876	4.0
3	S	0.0	0.917	4.0
4	D	-1.121	0.479	1.149
..
138	A	-0.305	0.827	3.504
139	E	-0.305	0.831	4.339
140	R	-0.305	0.89	5.496
141	Y	-0.305	0.889	3.86
142	R	0.0	0.978	5.0

[143 rows x 4 columns]

4. Using the program written in Q3 (unweighted frequency and entropy-based measure), compare the MSA from Clustal Omega, MAFFT, and MUSCLE. Identify the residues with (i) similar and (ii) Different conservation scores among the three alignment methods.

Code:

```
assign6bioinfo.ipynb ☆
File Edit View Insert Runtime Tools Help

Commands | + Code + Text

clustal_entropy = [
    "-SLSDKDKAAVRALWSKIGKSADAI GNDALSRMIVVPQTCTYFPHFDLHSGSAQVKAHGKKVADALTA AAVHDDLPGAL SALSDLHAHKL RVDPVNFKLLSHCLLVT LASHHPADFTPAVHASL DKFLASVSVTLTSKYR",
    "MVL SANDKSI NVKSI FSKTSSHAEYGAET LERMFTTYPQTCTYFPHFDLHSGSAQVKAHGKKVADALTA AAVHDDLPGAL SALSDLHAHKL RVDPVNFKLLSHCLLVT LASHHPADFTPAVHASL DKFLASVSVTLTSKYR",
    "-VLSPADKTI IKS TMDKIGGHAGDYGGEALDRTFQSFP TTKTYFPHFDLSPGSAQVKAHGKKVADALTA AAVHDDLPGAL SALSDLHAHKL RVDPVNFKLLSHCLLVT LASHHPADFTPAVHASL DKFLASVSVTLTSKYR",
    "MVL SADDKTI IKNCWGKIGGHGGEYGEALQRMFAAFTTKTYFPHFDLHSGSAQVKAHGKKVADALTA AAVHDDLPGAL SALSDLHAHKL RVDPVNFKLLSHCLLVT LASHHPADFTPAVHASL DKFLASVSVTLTSKYR",
    "-VL SAADKAI NVKAAWGVGGQAGHGA EALERMFLGFPTTKTYFPHFNL SHGSDQVKAHGKKVADALTA AAVHDDLPGAL SALSDLHAHKL RVDPVNFKLLSHCLLVT LASHHPADFTPAVHASL DKFLASVSVTLTSKYR",
    "MVL SGEKDSI NIKA AWGKIGGHGAEYGA EALERMFLGFPTTKTYFPHFDLHSGSAQVKAHGKKVADALTA AAVHDDLPGAL SALSDLHAHKL RVDPVNFKLLSHCLLVT LASHHPADFTPAVHASL DKFLASVSVTLTSKYR",
    "MVL SPADKTI NVKTA WGVGAHAGDYGA EALERMFLSFPTTKTYFPHFDLHSGSAQVKAHGKKVADALTA AAVHDDLPGAL SALSDLHAHKL RVDPVNFKLLSHCLLVT LASHHPADFTPAVHASL DKFLASVSVTLTSKYR",
    "MVL SPADKTI NVKTA WGVGAHAGDYGA EALERMFLSFPTTKTYFPHFDLHSGSAQVKAHGKKVADALTA AAVHDDLPGAL SALSDLHAHKL RVDPVNFKLLSHCLLVT LASHHPADFTPAVHASL DKFLASVSVTLTSKYR",
    "-VL SAADKAI NVKAAWGVGGQAGHGA EALERMFLGFPTTKTYFPHFNL SHGSDQVKAHGKKVADALTA AAVHDDLPGAL SALSDLHAHKL RVDPVNFKLLSHCLLVT LASHHPADFTPAVHASL DKFLASVSVTLTSKYR",
    "MVL SGEKDSI NIKA AWGKIGGHGAEYGA EALERMFLSFPTTKTYFPHFDLHSGSAQVKAHGKKVADALTA AAVHDDLPGAL SALSDLHAHKL RVDPVNFKLLSHCLLVT LASHHPADFTPAVHASL DKFLASVSVTLTSKYR",
    "MVL SADDKTI IKNCWGKIGGHGGEYGEALQRMFAAFTTKTYFPHFDLHSGSAQVKAHGKKVADALTA AAVHDDLPGAL SALSDLHAHKL RVDPVNFKLLSHCLLVT LASHHPADFTPAVHASL DKFLASVSVTLTSKYR",
    "-VL SPADKTI IKS TMDKIGGHAGDYGGEALDRTFQSFP TTKTYFPHFDLSPGSAQVKAHGKKVADALTA AAVHDDLPGAL SALSDLHAHKL RVDPVNFKLLSHCLLVT LASHHPADFTPAVHASL DKFLASVSVTLTSKYR",
    "MVL SANDKSI NVKSI FSKTSSHAEYGAET LERMFTTYPQTCTYFPHFDLHSGSAQVKAHGKKVADALTA AAVHDDLPGAL SALSDLHAHKL RVDPVNFKLLSHCLLVT LASHHPADFTPAVHASL DKFLASVSVTLTSKYR",
    "-SLSDKDKAAVRALWSKIGKSADAI GNDALSRMIVVPQTCTYFPHFDLHSGSAQVKAHGKKVADALTA AAVHDDLPGAL SALSDLHAHKL RVDPVNFKLLSHCLLVT LASHHPADFTPAVHASL DKFLASVSVTLTSKYR"
]

mafft_entropy = [
    "MVL SGEKDSI NIKA AWGKIGGHGAEYGA EALERMFLGFPTTKTYFPHFDLHSGSAQVKAHGKKVADALTA AAVHDDLPGAL SALSDLHAHKL RVDPVNFKLLSHCLLVT LASHHPADFTPAVHASL DKFLASVSVTLTSKYR",
    "MVL SAADKTI NVKAAWGVGGQAGHGA EALERMFLGFPTTKTYFPHFDLHSGSAQVKAHGKKVADALTA AAVHDDLPGAL SALSDLHAHKL RVDPVNFKLLSHCLLVT LASHHPADFTPAVHASL DKFLASVSVTLTSKYR",
    "MVL SPADKTI NVKTA WGVGAHAGDYGA EALERMFLSFPTTKTYFPHFDLHSGSAQVKAHGKKVADALTA AAVHDDLPGAL SALSDLHAHKL RVDPVNFKLLSHCLLVT LASHHPADFTPAVHASL DKFLASVSVTLTSKYR",
    "MVL SADDKTI IKNCWGKIGGHGGEYGEALQRMFAAFTTKTYFPHFDLHSGSAQVKAHGKKVADALTA AAVHDDLPGAL SALSDLHAHKL RVDPVNFKLLSHCLLVT LASHHPADFTPAVHASL DKFLASVSVTLTSKYR",
    "-VL SAADKAI NVKAAWGVGGQAGHGA EALERMFLGFPTTKTYFPHFNL SHGSDQVKAHGKKVADALTA AAVHDDLPGAL SALSDLHAHKL RVDPVNFKLLSHCLLVT LASHHPADFTPAVHASL DKFLASVSVTLTSKYR",
    "MVL SPADKTI NVKTA WGVGAHAGDYGA EALERMFLSFPTTKTYFPHFDLHSGSAQVKAHGKKVADALTA AAVHDDLPGAL SALSDLHAHKL RVDPVNFKLLSHCLLVT LASHHPADFTPAVHASL DKFLASVSVTLTSKYR",
    "MVL SPADKTI NVKTA WGVGAHAGDYGA EALERMFLSFPTTKTYFPHFDLHSGSAQVKAHGKKVADALTA AAVHDDLPGAL SALSDLHAHKL RVDPVNFKLLSHCLLVT LASHHPADFTPAVHASL DKFLASVSVTLTSKYR",
    "-VL SAADKAI NVKAAWGVGGQAGHGA EALERMFLGFPTTKTYFPHFNL SHGSDQVKAHGKKVADALTA AAVHDDLPGAL SALSDLHAHKL RVDPVNFKLLSHCLLVT LASHHPADFTPAVHASL DKFLASVSVTLTSKYR",
    "MVL SGEKDSI NIKA AWGKIGGHGAEYGA EALERMFLGFPTTKTYFPHFDLHSGSAQVKAHGKKVADALTA AAVHDDLPGAL SALSDLHAHKL RVDPVNFKLLSHCLLVT LASHHPADFTPAVHASL DKFLASVSVTLTSKYR",
    "MVL SADDKTI IKNCWGKIGGHGGEYGEALQRMFAAFTTKTYFPHFDLHSGSAQVKAHGKKVADALTA AAVHDDLPGAL SALSDLHAHKL RVDPVNFKLLSHCLLVT LASHHPADFTPAVHASL DKFLASVSVTLTSKYR",
    "-VL SPADKTI IKS TMDKIGGHAGDYGGEALDRTFQSFP TTKTYFPHFDLSPGSAQVKAHGKKVADALTA AAVHDDLPGAL SALSDLHAHKL RVDPVNFKLLSHCLLVT LASHHPADFTPAVHASL DKFLASVSVTLTSKYR",
    "MVL SANDKSI NVKSI FSKTSSHAEYGAET LERMFTTYPQTCTYFPHFDLHSGSAQVKAHGKKVADALTA AAVHDDLPGAL SALSDLHAHKL RVDPVNFKLLSHCLLVT LASHHPADFTPAVHASL DKFLASVSVTLTSKYR",
    "-SLSDKDKAAVRALWSKIGKSADAI GNDALSRMIVVPQTCTYFPHFDLHSGSAQVKAHGKKVADALTA AAVHDDLPGAL SALSDLHAHKL RVDPVNFKLLSHCLLVT LASHHPADFTPAVHASL DKFLASVSVTLTSKYR"
]

muscle_entropy = [
    "-SLSDKDKAAVRALWSKIGKSADAI GNDALSRMIVVPQTCTYFPHFDLHSGSAQVKAHGKKVADALTA AAVHDDLPGAL SALSDLHAHKL RVDPVNFKLLSHCLLVT LASHHPADFTPAVHASL DKFLASVSVTLTSKYR",
    "MVL SANDKSI NVKSI FSKTSSHAEYGAET LERMFTTYPQTCTYFPHFDLHSGSAQVKAHGKKVADALTA AAVHDDLPGAL SALSDLHAHKL RVDPVNFKLLSHCLLVT LASHHPADFTPAVHASL DKFLASVSVTLTSKYR",
    "-VLSPADKTI IKS TMDKIGGHAGDYGGEALDRTFQSFP TTKTYFPHFDLSPGSAQVKAHGKKVADALTA AAVHDDLPGAL SALSDLHAHKL RVDPVNFKLLSHCLLVT LASHHPADFTPAVHASL DKFLASVSVTLTSKYR",
    "MVL SADDKTI IKNCWGKIGGHGGEYGEALQRMFAAFTTKTYFPHFDLHSGSAQVKAHGKKVADALTA AAVHDDLPGAL SALSDLHAHKL RVDPVNFKLLSHCLLVT LASHHPADFTPAVHASL DKFLASVSVTLTSKYR",
    "-VL SAADKAI NVKAAWGVGGQAGHGA EALERMFLGFPTTKTYFPHFNL SHGSDQVKAHGKKVADALTA AAVHDDLPGAL SALSDLHAHKL RVDPVNFKLLSHCLLVT LASHHPADFTPAVHASL DKFLASVSVTLTSKYR",
    "MVL SGEKDSI NIKA AWGKIGGHGAEYGA EALERMFLGFPTTKTYFPHFDLHSGSAQVKAHGKKVADALTA AAVHDDLPGAL SALSDLHAHKL RVDPVNFKLLSHCLLVT LASHHPADFTPAVHASL DKFLASVSVTLTSKYR",
    "MVL SPADKTI NVKTA WGVGAHAGDYGA EALERMFLSFPTTKTYFPHFDLHSGSAQVKAHGKKVADALTA AAVHDDLPGAL SALSDLHAHKL RVDPVNFKLLSHCLLVT LASHHPADFTPAVHASL DKFLASVSVTLTSKYR",
    "MVL SPADKTI NVKTA WGVGAHAGDYGA EALERMFLSFPTTKTYFPHFDLHSGSAQVKAHGKKVADALTA AAVHDDLPGAL SALSDLHAHKL RVDPVNFKLLSHCLLVT LASHHPADFTPAVHASL DKFLASVSVTLTSKYR",
    "-VL SAADKAI NVKAAWGVGGQAGHGA EALERMFLGFPTTKTYFPHFNL SHGSDQVKAHGKKVADALTA AAVHDDLPGAL SALSDLHAHKL RVDPVNFKLLSHCLLVT LASHHPADFTPAVHASL DKFLASVSVTLTSKYR",
    "MVL SGEKDSI NIKA AWGKIGGHGAEYGA EALERMFLGFPTTKTYFPHFDLHSGSAQVKAHGKKVADALTA AAVHDDLPGAL SALSDLHAHKL RVDPVNFKLLSHCLLVT LASHHPADFTPAVHASL DKFLASVSVTLTSKYR",
    "MVL SADDKTI IKNCWGKIGGHGGEYGEALQRMFAAFTTKTYFPHFDLHSGSAQVKAHGKKVADALTA AAVHDDLPGAL SALSDLHAHKL RVDPVNFKLLSHCLLVT LASHHPADFTPAVHASL DKFLASVSVTLTSKYR",
    "-VL SPADKTI IKS TMDKIGGHAGDYGGEALDRTFQSFP TTKTYFPHFDLSPGSAQVKAHGKKVADALTA AAVHDDLPGAL SALSDLHAHKL RVDPVNFKLLSHCLLVT LASHHPADFTPAVHASL DKFLASVSVTLTSKYR",
    "MVL SANDKSI NVKSI FSKTSSHAEYGAET LERMFTTYPQTCTYFPHFDLHSGSAQVKAHGKKVADALTA AAVHDDLPGAL SALSDLHAHKL RVDPVNFKLLSHCLLVT LASHHPADFTPAVHASL DKFLASVSVTLTSKYR",
    "-SLSDKDKAAVRALWSKIGKSADAI GNDALSRMIVVPQTCTYFPHFDLHSGSAQVKAHGKKVADALTA AAVHDDLPGAL SALSDLHAHKL RVDPVNFKLLSHCLLVT LASHHPADFTPAVHASL DKFLASVSVTLTSKYR"
]
```

```

]
# Compute conservation scores for different alignments
clustal_entropy = compute_conservation_scores(clustal)["Entropy"]
mafft_entropy = compute_conservation_scores(mafft)["Entropy"]
muscle_entropy = compute_conservation_scores(muscle)["Entropy"]

# Create a DataFrame to store entropy scores
df = pd.DataFrame([clustal_entropy, mafft_entropy, muscle_entropy]).transpose()
df.columns = ["Clustal Omega", "MAFFT", "MUSCLE"]

# Print the DataFrame
print(df)

# Compute and display correlation between different alignment methods
print("\nCorrelation between conservation scores of different alignments:\n")
print(df.corr())
```

Result:

```
      Clustal Omega  MAFFT  MUSCLE
0      0.000  0.000  0.000
1     -0.305 -0.305 -0.305
2      0.000  0.000  0.000
3      0.000  0.000  0.000
4     -1.121 -1.121 -1.121
..      ...   ...   ...
138    -0.305 -0.305 -0.305
139    -0.305 -0.305 -0.305
140    -0.305 -0.305 -0.305
141    -0.305 -0.305 -0.305
142     0.000   NaN  0.000

[143 rows x 3 columns]

Correlation between conservation scores of different alignments:

      Clustal Omega  MAFFT  MUSCLE
Clustal Omega      1.0    1.0    1.0
MAFFT              1.0    1.0    1.0
MUSCLE             1.0    1.0    1.0
```

5. Check the scores manually at positions 9, 11, 20, 22, and 30 (use MSA from Clustal Omega)

Manual verification of unweighted frequency entropy scores:-

Pos 9: A, S, T, T, A, S, T, T, G, T, T

$$\begin{aligned} f(A) &= \frac{2}{11} & f(S) &= \frac{2}{11} & f(T) &= \frac{6}{11} & f(G) &= \frac{1}{11} \\ &= 0.182 & &= 0.182 & &= 0.545 & &= 0.091 \end{aligned}$$

$$\begin{aligned} \text{Entropy score} &= \sum f(i) \cdot (\ln(f(i))) \\ &= 0.182(-1.704) + 0.182(-1.704) + 0.545(-0.607) \\ &= -1.169 \quad \therefore \text{verified.} \end{aligned}$$

Pos 11: V, V, I, I, V, I, V, V, V, V, V

$$\begin{aligned} f(V) &= \frac{8}{11} & f(I) &= \frac{3}{11} \\ &= 0.727 & &= 0.273 \end{aligned}$$

$$\begin{aligned} \text{Entropy} &= 0.727(-0.319) + 0.273(-1.298) \\ &= -0.586 \Rightarrow \text{verified.} \end{aligned}$$

Pos 20: K, S, G, G, G, A, A, G, G, G

$$\begin{aligned} f(K) &= \frac{1}{11} & f(S) &= \frac{1}{11} & f(G) &= \frac{7}{11} & f(A) &= \frac{2}{11} \\ &= 0.091 & &= 0.091 & &= 0.636 & &= 0.182 \end{aligned}$$

$$\begin{aligned} \text{Entropy} &= 0.091(-2.397) + 0.091(-2.397) + 0.636(-0.453) \\ &\quad + 0.182(-1.704) \\ &= -1.034 \Rightarrow \text{verified.} \end{aligned}$$

Pos 22: A, A, A, G, A, G, A, A, A, A, A

$$\begin{aligned} f(A) &= \frac{9}{11} & f(G) &= \frac{2}{11} \\ &= 0.818 & &= 0.182 \end{aligned}$$

$$\begin{aligned} \text{Entropy} &= 0.818(-0.201) + 0.182(-1.704) \\ &= -0.472 \Rightarrow \text{verified} \end{aligned}$$

Pos 30: L, L, L, L, L, L, L, L, L, L, L

$$f(L) = \frac{11}{11} = 1$$

$$\begin{aligned} \text{Entropy score} &= 1(0) \\ &= 0 \Rightarrow \text{verified} \end{aligned}$$

6. Obtain the conservation score of 1BTM, A-chain using Consurf server (<https://consurf.tau.ac.il/>)

Step 1: Access the Consurf Server

1. Go to the [Consurf website](https://consurf.tau.ac.il/).

Step 2: Input the Protein Structure

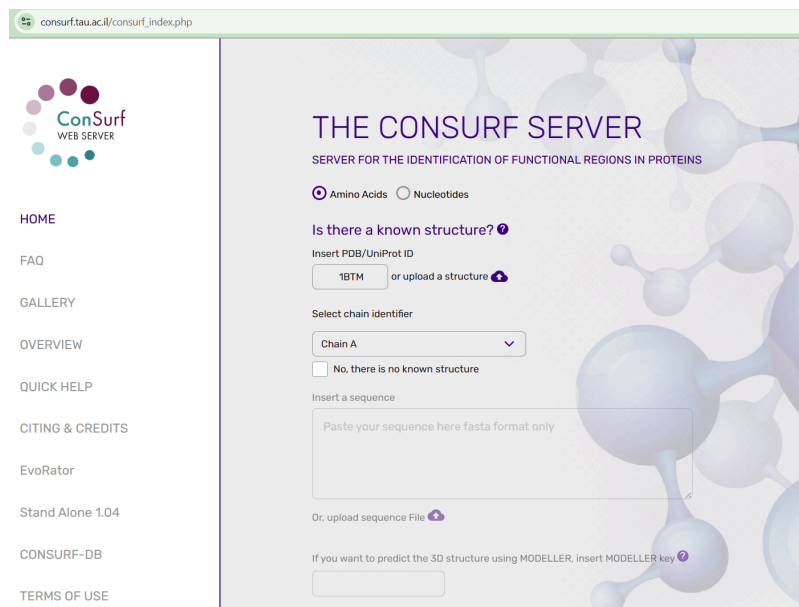
2. In the "Submit a New Job" section, select "Protein" as the macromolecule type.
3. Under "Choose Input Type", select "PDB ID" and enter 1BTM (this is the PDB code for the protein).
4. Click "Retrieve" to load the structure.

Step 3: Specify the Chain

5. In the "Select Chain" dropdown, choose "A" (A-chain of 1BTM).

Step 4: Configure Sequence Source

6. Choose "Run BLAST and build an alignment" (recommended).
7. Adjust parameters if needed, but default settings usually work well.



The screenshot displays the ConSurf web server interface. On the left is a navigation menu with links: HOME, FAQ, GALLERY, OVERVIEW, QUICK HELP, CITING & CREDITS, EvoRator, Stand Alone 1.04, CONSURF-DB, and TERMS OF USE. The main content area is titled "THE CONSURF SERVER" and "SERVER FOR THE IDENTIFICATION OF FUNCTIONAL REGIONS IN PROTEINS". It features radio buttons for "Amino Acids" (selected) and "Nucleotides". A section titled "Is there a known structure?" includes a field for "Insert PDB/UniProt ID" with "1BTM" entered and a button to "upload a structure". Below this is a "Select chain identifier" dropdown menu set to "Chain A", with an option to "No, there is no known structure". A text area for "Insert a sequence" is provided with the instruction "Paste your sequence here fasta format only". At the bottom, there is a button to "Or, upload sequence File" and a field for "If you want to predict the 3D structure using MODELLER, insert MODELLER key".

ConSurf-DB Analysis for: **1BTM chain A**

The following calculations are based on protein **1BTM chain B**, which is 100% sequence identical to the protein 1BTM chain A

Default display ☒ Color blind display



1 2 3 4 5 6 7 8 9
Variable Average Conserved

1 2 3 4 5 6 7 8 9
Variable Average Conserved

Insufficient Data

• [Download Scale](#)

1	11	21	31	41
RKPIIAGNWK	MHKTAEAVQ	FVEDVKGVP	PADEVISVVC	APFLFLDRLV
51	61	71	81	91
QAADGTDLKI	GAQTMHFADQ	GAYTGEVSPV	MLKDLGVITYV	ILGHSERRQM
101	111	121	131	141
FAETDETIVNK	KVLAAFTIRGL	IPITCCGESL	EEREAGQTNA	VVASQVEKAL
151	161	171	181	191
AGLTPEQVKQ	AVIAYEPIWA	IGTGKSSSTPE	DANSVCGHIR	SVVSRLFGPE
201	211	221	231	241
AAEAIRIQYG	GSVKPDNIRD	FLAQQQIDCP	LVGGASIEPA	SELQIVEAGR
251				
HE				

In the summary downloaded we got scores for all the SEQ:

consurfdb.tau.ac.il/DB_NEW/1BT1M/B/1BT1M_8_consurf_grads.txt

POS	SEQ	ATOM	SCORE	COLOR	CONFIDENCE	INTERVAL	B/E	F/S	MSA DATA	RESIDUE VARIETY
1	R	ARG:11:B	-1.236	9	-1.265, -1.224	9, 9	e	f	276/300	R 99%, K <1%
2	K	LYS:2:B	-0.120	5	-0.200, 0.010	6, 5	e	f	172/300	K 45%, R 27%, T 14%, Q 2%, V 2%, H 1%, I <1%, L <1%, N <1%, F <1%, M <1%, P <1%, Y <1%, S <1%
3	P	PRO:3:B	-0.060	5	-0.223, 0.010	6, 5	e	f	244/300	P 45%, K 31%, F 4%, R 3%, Q 3%, L 2%, I 2%, N 2%, T 1%, A 1%, M <1%, V <1%, G <1%, D <1%, H <1%, Y
4	I	ILE:4:B	-0.207	6	-0.408, -0.114	6, 5	e	f	273/300	L 47%, I 31%, F 10%, V 5%, Y 1%, M 1%, H <1%, T <1%, C <1%, W <1%
			-0.755	8	-0.851, -0.690	8, 7	e	f	279/300	V 35%, A 7%, M 4%, L 3%, F <1%, Y <1%
5	A	ALA:6:B	-0.857	8	-0.941, -0.801	8, 8	e	f	281/300	A 70%, G 9%, V 9%, I 6%, S 2%, L <1%, M <1%, C <1%, Y <1%, F <1%
7	G	GLY:7:B	-0.750	8	-0.851, -0.690	8, 7	e	f	282/300	G 85%, A 12%, T 1%, L <1%, S <1%
8	N	ASN:8:B	-1.247	9	-1.265, -1.244	9, 9	e	f	282/300	N 99%, S <1%
9	W	TRP:9:B	-1.102	9	-1.177, -1.055	9, 9	e	f	283/300	W 96%, F 2%, L <1%
10	K	LYS:10:B	-1.257	9	-1.265, -1.259	9, 9	e	f	283/300	K
11	M	PHE:11:B	-1.143	9	-1.201, -1.120	9, 9	e	f	287/300	M 92%, L 4%, C 4%, A <1%, S <1%, T <1%
12	H	HIS:12:B	-0.847	8	-0.941, -0.801	8, 8	e	f	287/300	W 73%, H 16%, V 6%, F 2%, T <1%, K <1%, K <1%
13	K	LYS:13:B	-0.307	6	-0.488, -0.223	7, 6	e	f	287/300	G 40%, L 24%, L 10%, M 4%, H 3%, T 3%, N 2%, P 2%, V 1%, Q 1%, A 1%, F 1%, R <1%, I <1%, V <1%, S
14	T	THR:14:B	-0.038	5	-0.223, 0.010	6, 5	e	f	286/300	T 41%, D 14%, S 14%, L 9%, N 8%, G 3%, V 1%, R 1%, H <1%, M <1%, Q <1%, A <1%, X <1%, I <1%
			0.943	4	0.525, 1.132	4, 3	e	f	289/300	L 21%, H 11%, P 10%, R 10%, V 8%, T 7%, K 7%, S 5%, A 3%, C <1%, T <1%, D <1%, F <1%, Y <1%, E 1%, M
15	A	ALA:16:B	0.960	4	0.525, 1.132	4, 3	e	f	289/300	A 32%, S 11%, E 9%, L 8%, K 5%, D 4%, Q 4%, V 4%, R 3%, T 3%, G 2%, N 2%, F 2%, P 1%, H 1%, I
16	Y	GLU:17:B	-0.324	6	-0.488, -0.223	7, 6	e	f	288/300	E 43%, S 14%, Q 11%, A 10%, D 6%, M 3%, T 2%, L 1%, F 1%, N 1%, G 1%, H 1%, V <1%, K <1%, R <1%, C
17	A	ALA:18:B	-0.436	7	-0.561, -0.408	7, 6	e	f	288/300	A 37%, L 12%, G 9%, N 9%, T 7%, S 7%, I 7%, V 5%, M <1%, K <1%, R <1%, Q <1%
19	V	VAL:19:B	1.675	3	1.132, 1.675	3, 3	e	f	289/300	A 13%, I 12%, V 12%, E 11%, R 8%, L 7%, T 6%, K 6%, D 4%, S 4%, Q 3%, G 2%, N 1%, P 1%, F
20	C	HIS:20:B	0.782	4	0.320, 0.781	5, 4	e	f	289/300	E 28%, A 21%, Q 8%, D 8%, T 8%, S 5%, K 5%, R 2%, H 2%, N 2%, V 1%, L 1%, I 1%, M <1%, Y
21	F	PHE:21:B	-0.124	5	-0.320, 0.010	6, 5	e	f	245/300	L 56%, F 26%, I 4%, Y 2%, M 2%, T 1%, W 1%, H 1%, V <1%, A <1%, S <1%, D <1%
22	V	VAL:22:B	0.660	4	0.320, 0.781	5, 4	e	f	290/300	V 26%, L 24%, A 18%, I 15%, F 8%, T 7%, C 3%, S 1%, M 1%
23	E	GLU:23:B	1.140	3	0.781, 1.132	4, 3	e	f	290/300	Q 14%, E 14%, A 13%, K 11%, R 8%, T 7%, S 7%, N 6%, T 6%, G 3%, V 1%, H 1%, L 1%, M 1%, I <1%, C
24	D	ASP:24:B	0.988	4	0.525, 1.132	4, 3	e	f	290/300	A 44%, E 17%, K 15%, G 11%, D 8%, T 4%, S 3%, Q 3%, R 2%, H 2%, N 2%, F 1%, L 1%, V 1%, I 1%, P
25	V	VAL:25:B	0.143	6	-0.320, -0.114	6, 5	e	f	290/300	L 55%, I 17%, V 12%, F 9%, M 2%, T <1%, C <1%, G <1%, Y <1%
26	K	LYS:26:B	1.349	3	0.781, 1.675	4, 3	e	f	290/300	A 18%, K 16%, L 12%, N 9%, V 8%, R 7%, I 5%, S 5%, G 4%, D 3%, T 2%, Q 1%, C 1%, M 1%, H 1%, W
27	G	GLY:27:B	1.517	3	0.781, 1.675	4, 3	e	f	287/300	A 22%, D 9%, P 8%, E 8%, S 6%, K 6%, R 6%, G 5%, Q 5%, M 4%, T 4%, N 3%, Y 2%, F 2%, H 1%, V
28	H	HIS:28:B	2.988	1	1.675, 3.138	3, 1	e	f	269/300	G 17%, A 15%, L 14%, E 12%, S 6%, K 5%, T 5%, X 4%, I 4%, W 4%, T 3%, N 1%, Y 1%, V 1%, D 1%, F 1%, P
29	N	VAL:29:B	1.576	3	1.132, 1.675	3, 3	e	f	258/300	L 38%, V 14%, I 9%, A 8%, T 5%, S 3%, Y 3%, P 3%, D 1%, K 1%, G 1%, E 1%, F 1%, N 1%, R 1%, Q
30	C	PRO:30:B	2.388	2	1.132, 3.138	3, 1	e	f	179/300	K 15%, D 11%, A 11%, S 10%, P 9%, G 8%, E 8%, R 6%, N 3%, T 3%, I 2%, H 2%, V 2%, L 1%, Q 1%, M
31	P	PRO:31:B	3.125	1	1.675, 3.138	3, 1	e	f	284/300	D 15%, L 14%, G 10%, A 8%, K 7%, P 5%, S 5%, T 5%, R 5%, V 4%, E 4%, I 3%, Q 3%, N 2%, Y 1%, M
32	A	ALA:32:B	2.834	1	1.675, 3.138	3, 1	e	f	284/300	P 14%, A 11%, D 10%, V 7%, T 6%, S 5%, N 4%, F 4%, L 4%, K 4%, G 4%, C 3%, E 3%, H 3%, R
33	D	ASP:33:B	3.072	1	1.675, 3.138	3, 1	e	f	290/300	P 14%, P 14%, D 12%, E 11%, S 10%, G 7%, T 5%, K 4%, L 3%, N 3%, Q 3%, R 2%, V 1%, H 1%, M <1%, C
34	E	GLU:34:B	1.654	3	1.132, 1.675	3, 3	e	f	291/300	G 15%, D 14%, A 12%, N 12%, K 11%, R 6%, S 6%, H 5%, E 4%, T 3%, P 2%, Q 1%, V 1%, L 1%, F <1%, I