Applied Computational Genomics (PHC 7736)

Spring 2023

Lecture 13: Study Protein from Its Sequence

Bi Zhao, PhD

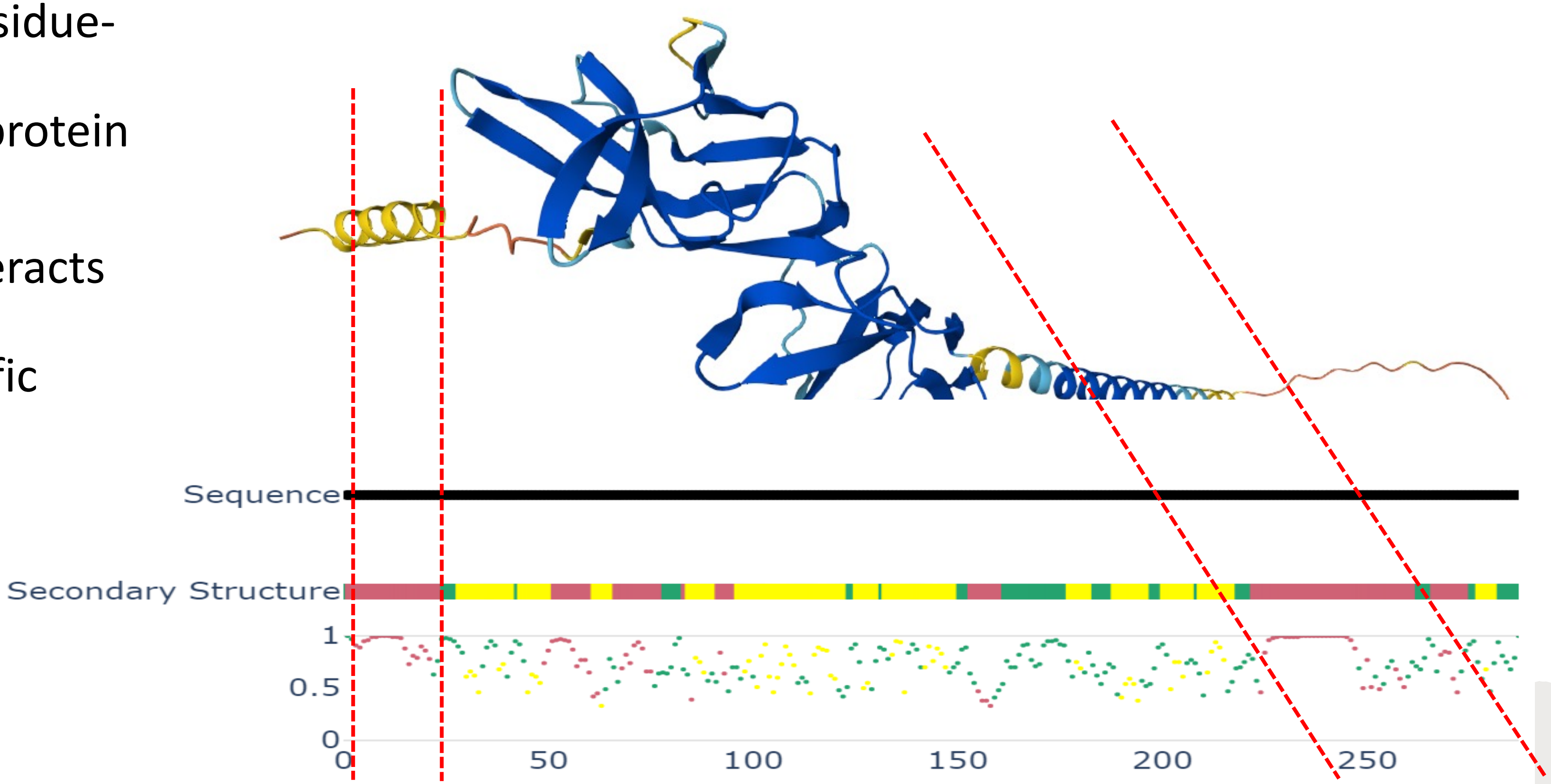
Overview

- Introduction to protein solvent accessibility and conservation
- Part 1: Obtain predictive results from ASAquick and MMseqs2
    - In-class work
        - Installation of tools
        - Get the predictive results
        - Short break (~10 mins)
    - Description of the Result format
- Part 2: Analyze the outputs from ASAquick and MMseqs2 to putative annotate protein residue-level solvent accessibility and conservation score
    - In-class work
        - Get the formatted putative annotations
    - Results intepretation

What can be learned from protein sequence

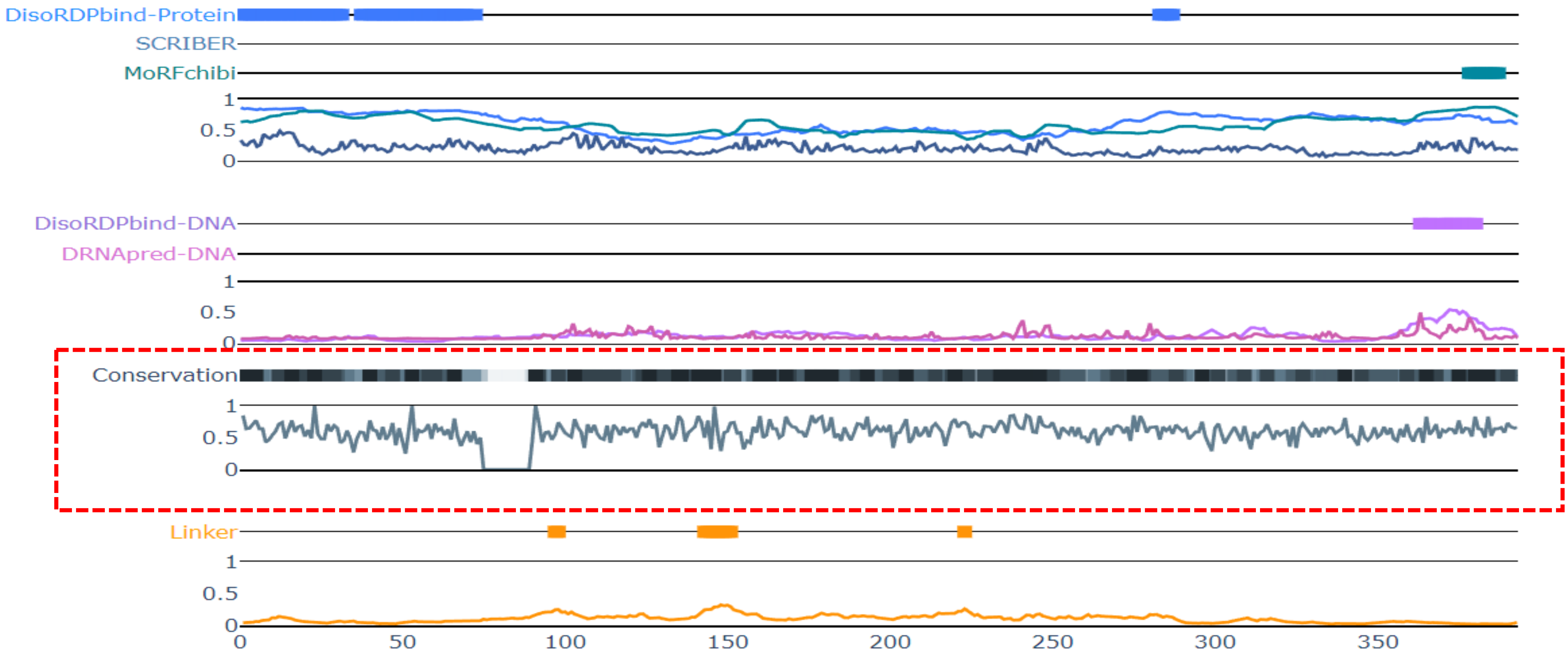
- **Structure**
- What can be predicted?
  - 3D structure (AlphaFold, Atom-level)
  - Secondary structure (alpha helix, beta sheet, etc., PSIPRED, residue-level)
  - View the overall fold of the protein
- How can this be used?
  - Understand how protein interacts with other
  - Design drugs targeting specific regions of a protein

- P24071
- Immunoglobulin alpha Fc receptor
- FCAR



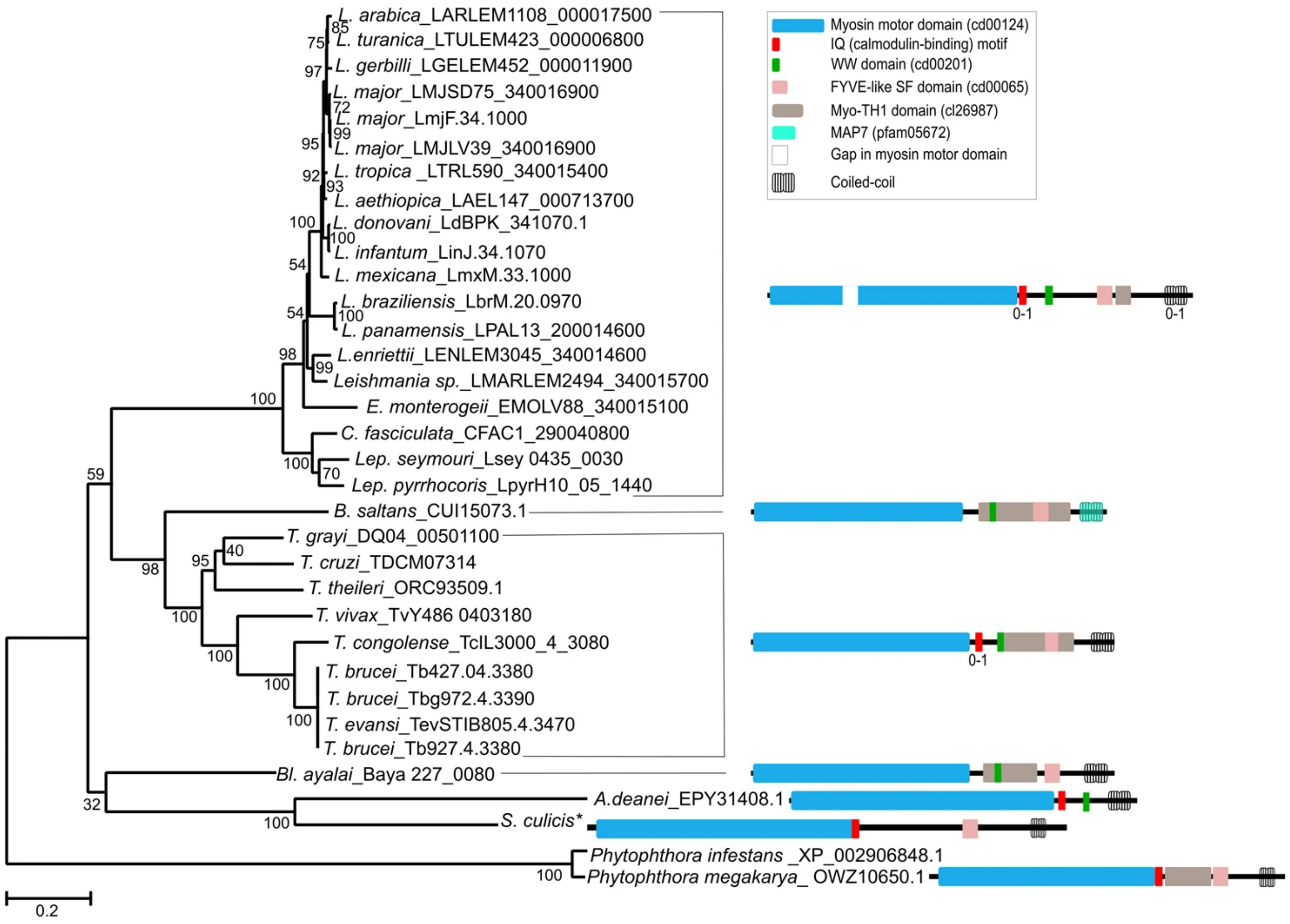
What can be learned from protein sequence (Cont’d)

- Function (P04637 Cellular_tumor_antigen_p53)
  - The highly conserved regions across different species.
  - Identify binding regions. E.g. DNA-binding regions for gene expression regulation.



What can be learned from protein sequence (Cont’d)

- Evolution
    - Reveal relationships
    - Similar sequences have a common ancestor
    - Different species comparing may indicate the evolutionary history of those species

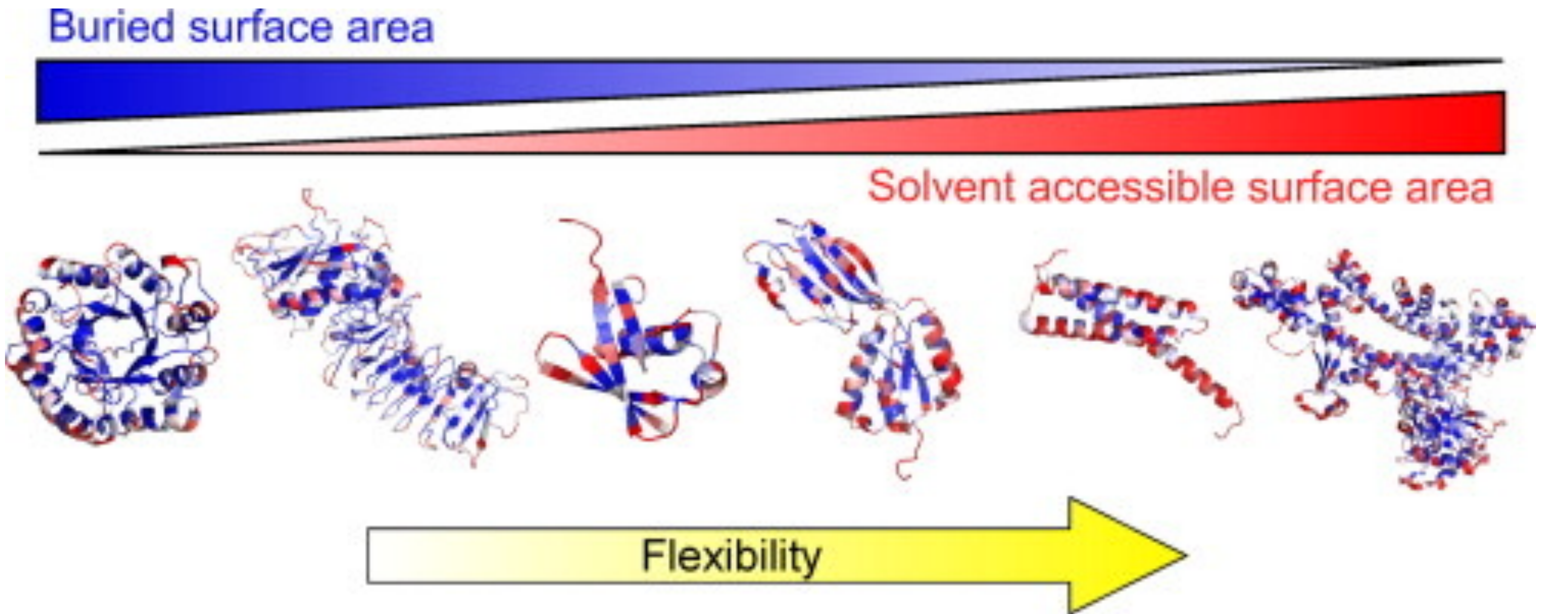


L. arabica_LARLEM1108_000017500
85
75
L. turanica_LTULEM423_000006800
97
L. gerbilli_LGELEM452_000011900
L. major_LMJSD75_340016900
72
L. major_LmjF.34.1000
99
95
L. major_LMJLV39_340016900
92
L. tropica _LTRL590_340015400
93
L. aethiopica_LAEL147_000713700
100
L. donovani_LdBPK_341070.1
100
L. infantum_LinJ.34.1070
54
L. mexicana_LmxM.33.1000
L. braziliensis_LbrM.20.0970
54
100
L. panamensis_LPAL13_200014600
98
L.enriettii_LENLEM3045_340014600
99
Leishmania sp._LMARLEM2494_340015700
100
E. monterogeii_EMOLV88_340015100
C. fasciculata_CFAC1_290040800
Lep. seymouri_Lsey 0435_0030
100
70
Lep. pyrrhocoris_LpyrH10_05_1440
59
B. saltans_CUI15073.1
T. grayi_DQ04_00501100
40
95
T. cruzi_TDCM07314
98
T. theileri_ORC93509.1
100
T. vivax_TvY486 0403180
T. congolense_TcIL3000_4_3080
100
T. brucei_Tb427.04.3380
100
T. brucei_Tbg972.4.3390
100
T. evansi_TevSTIB805.4.3470
T. brucei_Tb927.4.3380
Bl. ayalai_Baya 227_0080
A.deanei_EPY31408.1
32
S. culicis*
100
Phytophthora infestans _XP_002906848.1
100
Phytophthora megakarya_ OWZ10650.1
0.2
0-1
0-1
0-1
Myosin motor domain (cd00124)
IQ (calmodulin-binding) motif
WW domain (cd00201)
FYVE-like SF domain (cd00065)
Myo-TH1 domain (cl26987)
MAP7 (pfam05672)
Gap in myosin motor domain
Coiled-coil

Souza, Denise, et al. (2018). Scientific Reports. 8. 10.1038/s41598-017-18865-y.

Solvent Accessibility

- The propensity of an amino acid that is exposed to the surrounding solvent or buried within the protein core.
- Solvent accessibility can be used to predict the functional and structural properties of a protein.
- ASAquick that are used to annotate protein solvent accessibility
  - Physicochemical properties, 20 parameters related to residue mutation probabilities from BLOSUM62 matrix, the residue length of the protein, its one-residue composition, etc
  - Fast and relatively accurate
  - Outputs include raw solvent accessibility score, relative solvent accessibility, secondary structure, etc.



Protein sequence conservation

- Multiple sequence alignment
- Conservation score
  - Calculated to quantify the degree of sequence conservation at each amino acid residue
- MMseqs2
  - Many to many sequence searching
  - Optimized algorithm that is fast and accurate
  - Outputs: cluster of sequence, sequence-level consensus score, Position-specific scoring matrix

In-class work

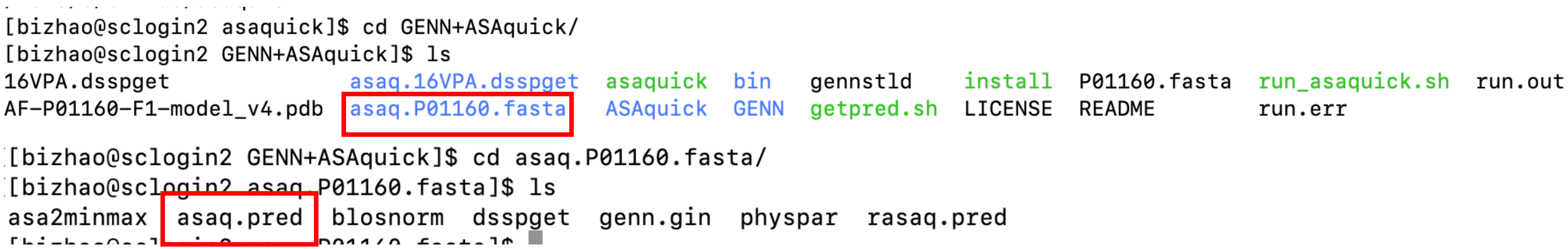
- Download installation tutorial
- Download the reference protein sequences and uploaded to the folder where installing MMseqs2

```
$ cd ~/mmseqs2
$ mkdir dataset
$ scp ref_protein_seq.fasta.gz bullrocky@sc.rc.usf.edu:/home/b/bullrochky/mmseq2/dataset
$ gzip –d ref_protein_seq.fasta.gz
```

- Download code run_asaquick.sh, create_mmseqsDB.sh and run_mmseqs.sh upload to each folder
- Run each script

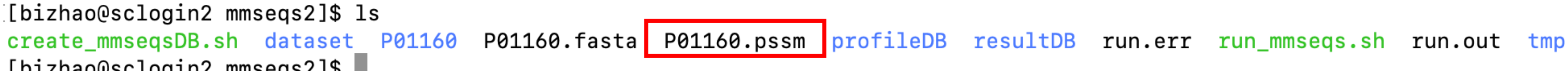
Results

ASAquick



```
[bizhao@sclogin2 asaquick]$ cd GENN+ASAquick/
[bizhao@sclogin2 GENN+ASAquick]$ ls
16VPA.dsspget              asaq.16VPA.dsspget  asaquick  bin   gennstld    install  P01160.fasta  run_asaquick.sh  run.out
AF-P01160-F1-model_v4.pdb  asaq.P01160.fasta   ASAquick  GENN  getpred.sh  LICENSE  README        run.err

[bizhao@sclogin2 GENN+ASAquick]$ cd asaq.P01160.fasta/
[bizhao@sclogin2 asaq.P01160.fasta]$ ls
asa2minmax  asaq.pred  blosnorm  dsspget  genn.gin  physpar  rasaq.pred
[bizhao@sclogin2 asaq.P01160.fasta]$
```

mmseqs2



```
[bizhao@sclogin2 mmseqs2]$ ls
create_mmseqsDB.sh  dataset  P01160  P01160.fasta  P01160.pssm  profileDB  resultDB  run.err  run_mmseqs.sh  run.out  tmp
[bizhao@sclogin2 mmseqs2]$
```

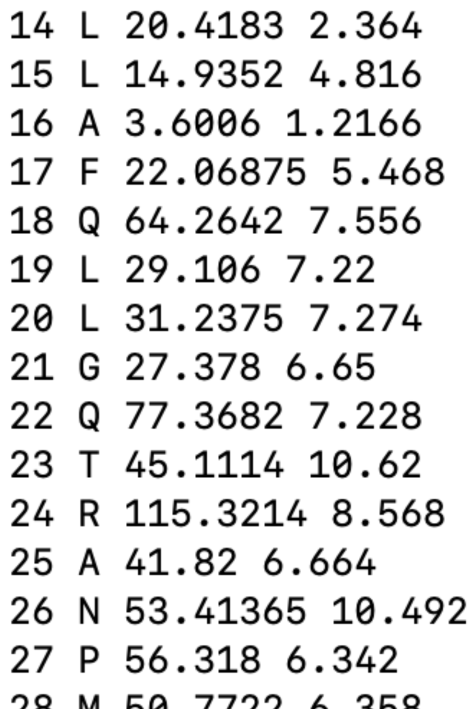
```
# create the pred_result folder under prot_analysis directory
$ cd ~/prot_study
$ mkdir pred_result
$ cp ./mmseqs2/P01160.pssm ./pred_result
$ cp ./asaquick/GENN+ASAquick/asaq.P01160.fasta/asaq.pred ./pred_result
```

Format ASAquick

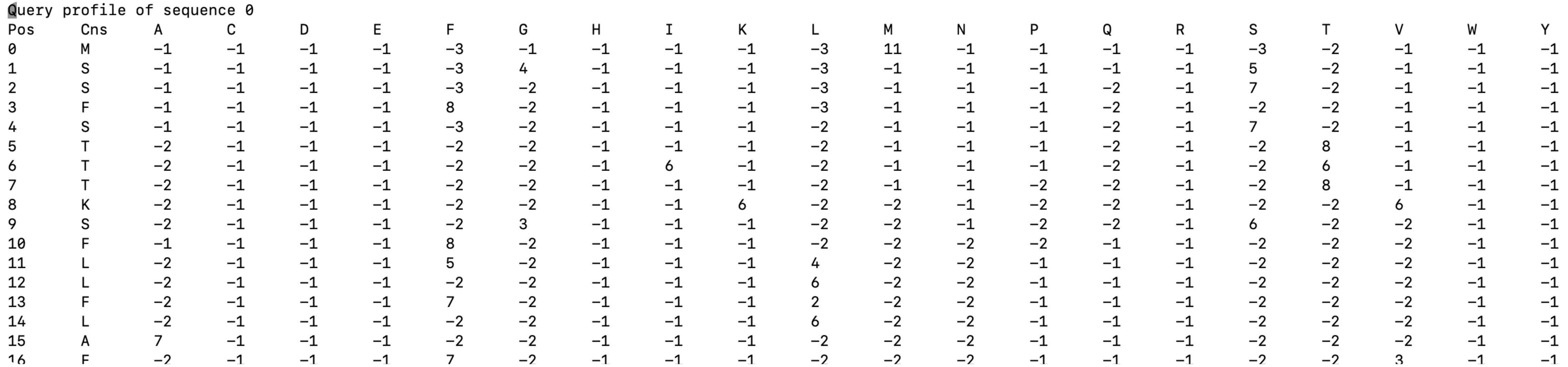
- The score represents the how much of the surrounding of a residue are occupied by other parts of the chain and how much is accessible to the solvent or to other interactions external to chain.
- To normalize the information, the experimental/theoretical score for each amino acid type are applied. These scores are from ref. Tien MZ. PLOS ONE 8(11):e80635.

  max_value = {"A":129,"R":274,"N":195,"D":193,"C":167, "E":223,"Q":225,"G":104,"H":224,"I":197,"L":201,"K":197, "M":224,"F":240,"P":159,"S":155,"T":172,"W":285,"Y":263, "V":174}

  Score = residue[“X”]/max_value[“X”]



```
14 L 20.4183 2.364
15 L 14.9352 4.816
16 A 3.6006 1.2166
17 F 22.06875 5.468
18 Q 64.2642 7.556
19 L 29.106 7.22
20 L 31.2375 7.274
21 G 27.378 6.65
22 Q 77.3682 7.228
23 T 45.1114 10.62
24 R 115.3214 8.568
25 A 41.82 6.664
26 N 53.41365 10.492
27 P 56.318 6.342
28 M 50.7722 6.358
```

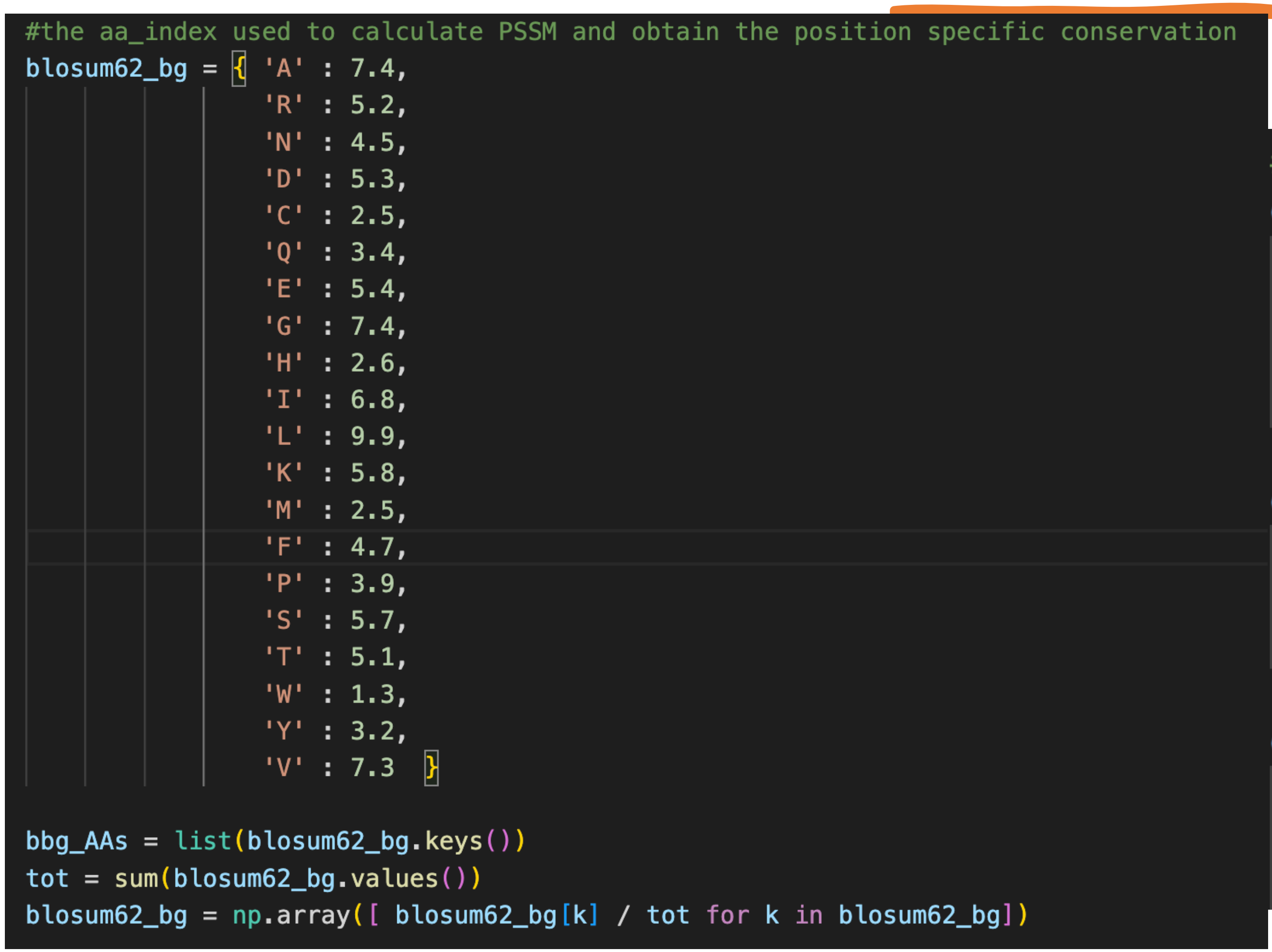


```
Query profile of sequence 0
Pos     Cns     A       C       D       E       F       G       H       I       K       L       M       N       P       Q       R       S       T       V       W       Y
0       M       -1      -1      -1      -1      -3      -1      -1      -1      -1      -3      11      -1      -1      -1      -1      -3      -2      -1      -1      -1
1       S       -1      -1      -1      -1      -3      4       -1      -1      -1      -3      -1      -1      -1      -1      -1      5       -2      -1      -1      -1
2       S       -1      -1      -1      -1      -3      -2      -1      -1      -1      -3      -1      -1      -1      -2      -1      7       -2      -1      -1      -1
3       F       -1      -1      -1      -1      8       -2      -1      -1      -1      -3      -1      -1      -1      -2      -1      -2      -2      -1      -1      -1
4       S       -1      -1      -1      -1      -3      -2      -1      -1      -1      -2      -1      -1      -1      -2      -1      7       -2      -1      -1      -1
5       T       -2      -1      -1      -1      -2      -2      -1      -1      -1      -2      -1      -1      -1      -2      -1      -2      8       -1      -1      -1
6       T       -2      -1      -1      -1      -2      -2      -1      6       -1      -2      -1      -1      -1      -2      -1      -2      6       -1      -1      -1
7       T       -2      -1      -1      -1      -2      -2      -1      -1      -1      -2      -1      -1      -2      -2      -1      -2      8       -1      -1      -1
8       K       -2      -1      -1      -1      -2      -2      -1      -1      6       -2      -2      -1      -2      -2      -1      -2      -2      6       -1      -1
9       S       -2      -1      -1      -1      -2      3       -1      -1      -1      -2      -2      -1      -2      -2      -1      6       -2      -2      -1      -1
10      F       -1      -1      -1      -1      8       -2      -1      -1      -1      -2      -2      -2      -2      -1      -1      -2      -2      -2      -1      -1
11      L       -2      -1      -1      -1      5       -2      -1      -1      -1      4       -2      -2      -1      -1      -1      -2      -2      -2      -1      -1
12      L       -2      -1      -1      -1      -2      -2      -1      -1      -1      6       -2      -2      -1      -1      -1      -2      -2      -2      -1      -1
13      F       -2      -1      -1      -1      7       -2      -1      -1      -1      2       -2      -2      -1      -1      -1      -2      -2      -2      -1      -1
14      L       -2      -1      -1      -1      -2      -2      -1      -1      -1      6       -2      -2      -1      -1      -1      -2      -2      -2      -1      -1
15      A       7       -1      -1      -1      -2      -2      -1      -1      -1      -2      -2      -2      -1      -1      -1      -2      -2      -2      -1      -1
16      F       -2      -1      -1      -1      7       -2      -1      -1      -1      -2      -2      -2      -1      -1      -1      -2      -2      3       -1      -1
```

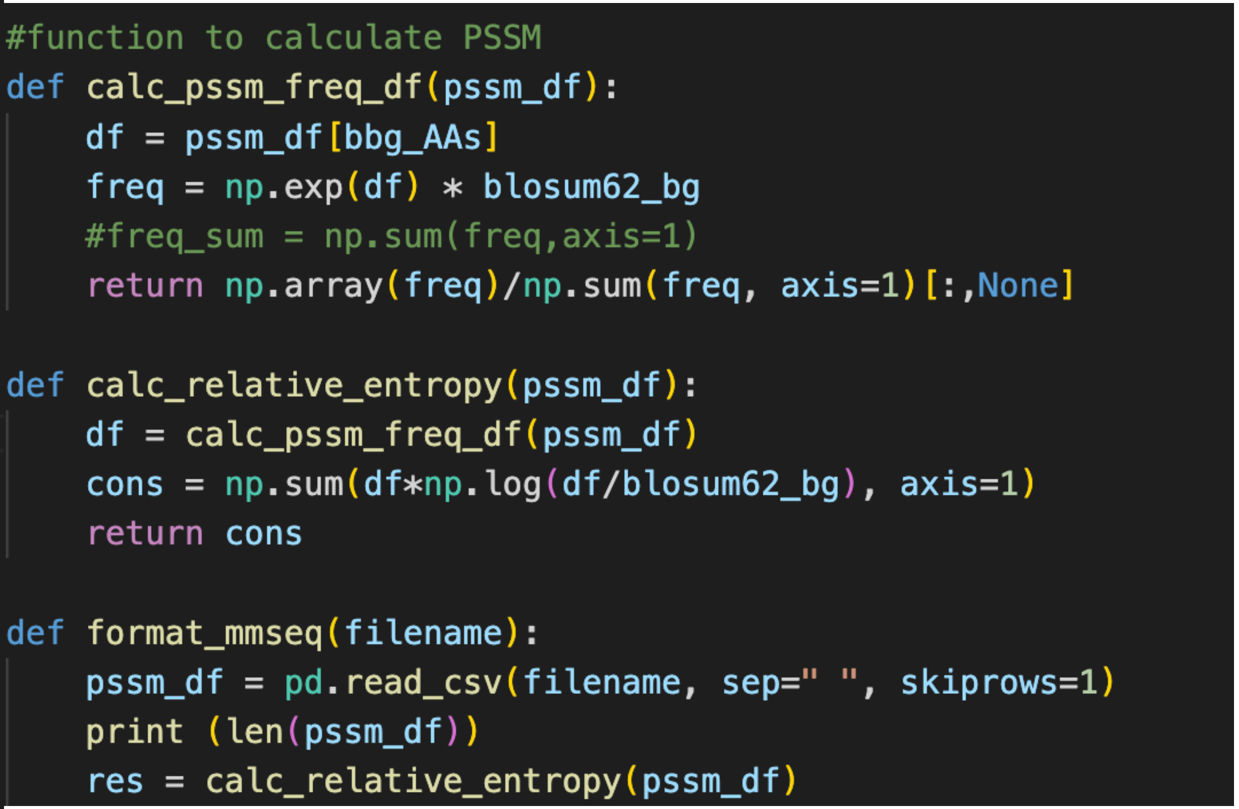
Format MMseqs2

- Position-Specific Scoring Matrix (PSSM)
- Reflect the probability of observing each amino acid at each position
- The conservation score for each position is calculated by comparing with background frequency of that amino acid in protein database

How the conservation score calculated?



```python
#the aa_index used to calculate PSSM and obtain the position specific conservation
blosum62_bg = { 'A' : 7.4,
                'R' : 5.2,
                'N' : 4.5,
                'D' : 5.3,
                'C' : 2.5,
                'Q' : 3.4,
                'E' : 5.4,
                'G' : 7.4,
                'H' : 2.6,
                'I' : 6.8,
                'L' : 9.9,
                'K' : 5.8,
                'M' : 2.5,
                'F' : 4.7,
                'P' : 3.9,
                'S' : 5.7,
                'T' : 5.1,
                'W' : 1.3,
                'Y' : 3.2,
                'V' : 7.3  }

bbg_AAs = list(blosum62_bg.keys())
tot = sum(blosum62_bg.values())
blosum62_bg = np.array([ blosum62_bg[k] / tot for k in blosum62_bg])
```

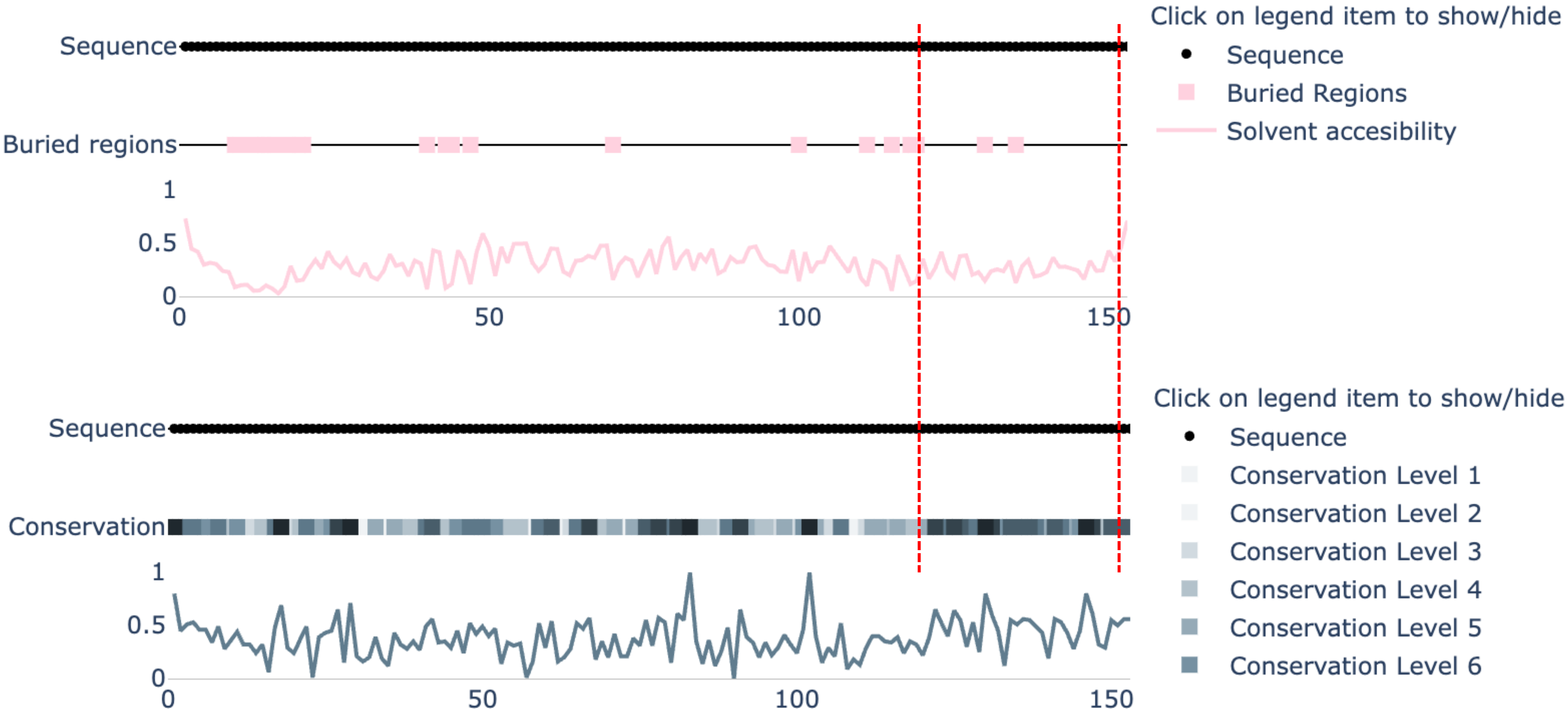


```python
#function to calculate PSSM
def calc_pssm_freq_df(pssm_df):
    df = pssm_df[bbg_AAs]
    freq = np.exp(df) * blosum62_bg
    #freq_sum = np.sum(freq,axis=1)
    return np.array(freq)/np.sum(freq, axis=1)[:,None]

def calc_relative_entropy(pssm_df):
    df = calc_pssm_freq_df(pssm_df)
    cons = np.sum(df*np.log(df/blosum62_bg), axis=1)
    return cons

def format_mmseq(filename):
    pssm_df = pd.read_csv(filename, sep=" ", skiprows=1)
    print (len(pssm_df))
    res = calc_relative_entropy(pssm_df)
```

In-class work

- Download python script and shell file, uploaded to sc cluster
- Run the shell file and obtain the result.
- Download and visualize the results.

Results


Sequence
Buried regions
1
0.5
0
0
50
100
150
Click on legend item to show/hide
● Sequence
Buried Regions
Solvent accesibility
Sequence
Conservation
1
0.5
0
0
50
100
150
Click on legend item to show/hide
● Sequence
Conservation Level 1
Conservation Level 2
Conservation Level 3
Conservation Level 4
Conservation Level 5
Conservation Level 6