

Name: Maradana Roshan

Roll No: BS21B019

Bioinformatics

Practical 8

1. Identify the pair of sequences which are close to each other using Hamming and Euclidean distance methods.

(i) AMENLNMDLLYMAAAVMMGLAAIGAAIGIGILGGKFLEGAARQPDLIPLLRTQFFIVMGLVD
AIPMIAVG LGLYVMFAVA

(ii) AADVSAAVGATGQSGMTYRLGLSWDWDKSWWQTSTGRLTGYWDAGYTYWEGGDEGA
GKHSLSFAP

VFVYEFAGDSIKPFIEAGIGVAAFSGTRVGDQNLGSSLNFEDRIGAGLKFANGQSVGVRAIHYS
NAGLKQPN DGIESYSLFYKIPI

(iii) MALLPAAPGAPARATPTRWPVGCNRPWTKWSYDEALDGIKAAGYAWTGLLTASKPSLH
HATATPEY LAALKQKSRHAA

```
import math

def calculate_composition(sequence):
    composition = {}
    seq_length = len(sequence)
    for base in set(sequence):
        composition[base] = sequence.count(base) / seq_length
    return composition

def hamming_distance(seq1, seq2):
    composition_seq1 = calculate_composition(seq1)
    composition_seq2 = calculate_composition(seq2)

    seq_union =
    set(composition_seq1.keys()).union(set(composition_seq2.keys()))

    distance = 0
    euclidean_distance = 0

    for base in seq_union:
        mod = abs(composition_seq1.get(base, 0) -
composition_seq2.get(base, 0))
        distance += mod
        euclidean_distance += mod ** 2

    euclidean_distance = math.sqrt(euclidean_distance) * 100
    return distance, euclidean_distance
```

```

seq_1 =
"AMENLNMDLLYMAAAVMMGLAAIGAAIGIGILGGKFLEGAARQPDLIPLLRTQFFIVMGLVDAIPMIAVGLGL
YVMFAVA"
seq_2 =
"AADVSAAVGATGQSGMTYRLGLSWDWDKSWWQTSTGRLTGYWDAGYTYWEGGDEGAGKHSLSFAPVFVYEFAG
DSIKPFIEAGIGVAAFSGTRVGDQNLGSSLNFEDRIGAGLKFANGQSVGVRAIHYSNAGLKQPNDGIESYSLFY
KIPI"
seq_3 =
"MALLPAAPGAPARATPTRWPVGCENRPWTKWSYDEALDGIKAAGYAWTGLLTASKPSLHHATATPEYLAALKQ
KSRHAA"

print(hamming_distance(seq_1, seq_2))
print(hamming_distance(seq_2, seq_3))
print(hamming_distance(seq_1, seq_3))

```

Output:

```

Run: P9Q1
"C:\Users\Maradana Roshan\PycharmProjects\pythonProject1\venv\Scripts\python.exe" "C:/Users/Maradana Roshan/PycharmProjects/pythonProject1/P9Q1
.py"
(0.6657284768211921, 20.1062168421535)
(0.726632576075111, 20.112952107271116)
(0.8433544303797469, 22.086816691389572)
Process finished with exit code 0

```

2. Get the non-redundant sequences of beta barrel membrane proteins with sequence identities of less than 40%, 50%, 75% and 90% using CD-HIT

- Downloaded the manually done CD HIT sequences using MobaXTerm & followed the instructions given in the question

Percentage identity	Total no.of clusters
40%	239
50%	264
75%	329
90%	369

3. Get the non-redundant sequences of the same type of proteins with sequence identities of less than 20%, 30%, 40% and 50% using PISCES
(<https://dunbrack.fccc.edu/piscs/>)

Ans)

- Ran PISCES and from the results obtained we can see the chains representing each cluster
- | Percentage identity | Total no.of clusters |
|---------------------|----------------------|
| 20% | 31 |
| 30% | 39 |
| 40% | 42 |
| 50% | 47 |

PISCES: A Protein Sequence Culling Server

Step 1: Input PDB list

Paste or type in your list of PDB chains in the following textbox [Help?](#)



Step 2: Choose your desired thresholds

Maximum pairwise percent sequence identity:	<input type="text" value="20"/>
Minimum resolution (X-ray and EM):	<input type="text" value="0.0"/>
Maximum resolution (X-ray and EM):	<input type="text" value="2.0"/>
Maximum R-value (X-ray only):	<input type="text" value="0.25"/>
Minimum chain length:	<input type="text" value="40"/>
Maximum chain length:	<input type="text" value="10000"/>
Include X-ray entries?	<input checked="" type="radio"/> Yes <input type="radio"/> No
Include cryo-EM entries?	<input type="radio"/> Yes <input checked="" type="radio"/> No
Include NMR entries?	<input type="radio"/> Yes <input checked="" type="radio"/> No
Include chains with chain breaks? (clicking "No" will reduce number of chains by ~20%)	<input checked="" type="radio"/> Yes <input type="radio"/> No
Include chains with missing residues due to disorder? (clicking "No" will reduce number of chains by 50-75%)	<input checked="" type="radio"/> Yes <input type="radio"/> No
Write file of pairwise sequence identities of returned list?	<input type="radio"/> Yes <input checked="" type="radio"/> No

Step 3: Submit your job

PISCES: A Protein Sequence Culling Server

Your representative PDB list will be generated based on the following criteria:

Sequence percentage identity	<= 20%
Sequence chain length	40 ~ 10000
Resolution	0.0 ~ 2.0
R-factor value	0.25
X-ray entries	include
EM entries	exclude
NMR entries	exclude
Allow chain breaks	yes
Allow disorder	yes
Print seqids	no

In order to send you the result, please fill out following information:

User Name:
Email address:
Institution:

- First get the codes of the redundant sequences of beta barrel membrane proteins and paste them in the PISCES website
- Set the sequence identity to 20%,30%,40% and 50% and obtain the results and count the no.of clusters

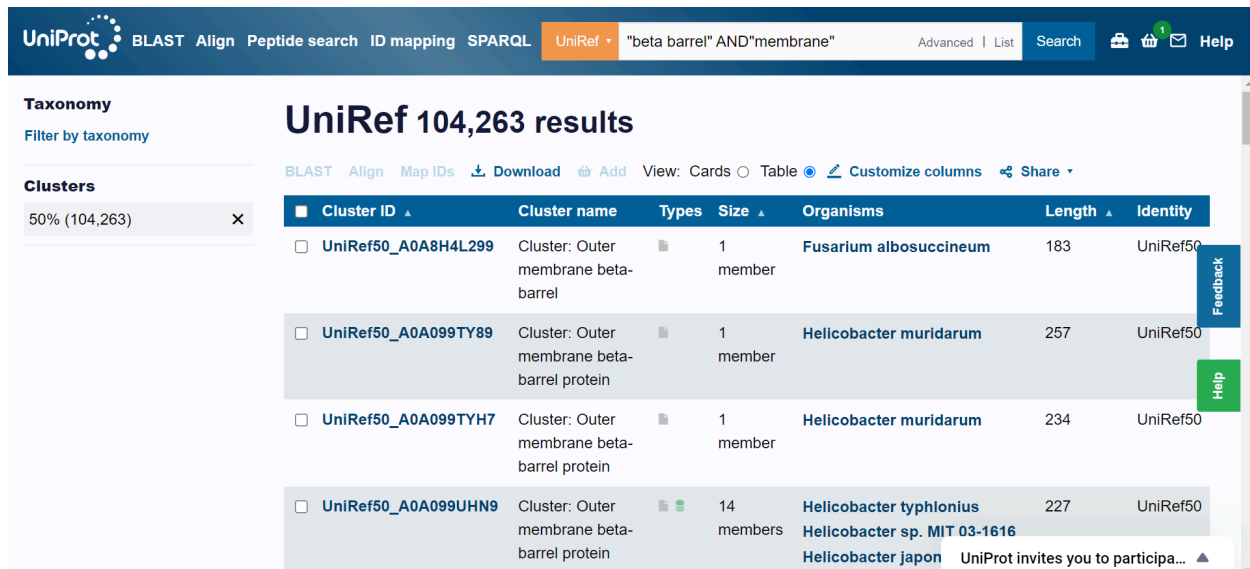
4. Compare the results obtained with the cut-offs 40% and 50%.

cullpdb_pc40.0_res0.0-2.0_len40-1						cullpdb_pc50.0_res0.0-2.0_len40-1					
File Edit View						File Edit View					
PDBchain	len	method	resol	rfac	freerfac	PDBchain	len	method	resol	rfac	freerfac
3GP6A	163	XRAY	1.4	0.173	0.208	3GP6A	163	XRAY	1.4	0.173	0.208
2FGQX	332	XRAY	1.45	0.15	0.171	2FGQX	332	XRAY	1.45	0.15	0.171
5FVNA	342	XRAY	1.45	0.156	0.183	5FVNA	342	XRAY	1.45	0.156	0.183
3SZVA	401	XRAY	1.45	0.184	0.198	3SZVA	401	XRAY	1.45	0.184	0.198
4RJWA	429	XRAY	1.52	0.151	0.168	4RJWA	429	XRAY	1.52	0.151	0.168
6EHBA	320	XRAY	1.55	0.172	0.199	6EHBA	320	XRAY	1.55	0.172	0.199
4RLCA	175	XRAY	1.6	0.191	0.229	4RLCA	175	XRAY	1.6	0.191	0.229
1QJPA	171	XRAY	1.65	0.155	0.198	1QJPA	171	XRAY	1.65	0.155	0.198
6EHDA	325	XRAY	1.66	0.188	0.214	6EHDA	325	XRAY	1.66	0.188	0.214
5AZPA	455	XRAY	1.69	0.161	0.18	5AZPA	455	XRAY	1.69	0.161	0.18
3PGUA	427	XRAY	1.7	0.168	0.18	3PGUA	427	XRAY	1.7	0.168	0.18
4D5BA	339	XRAY	1.702	0.168	0.19	4D5BA	339	XRAY	1.702	0.168	0.19
5DL7A	419	XRAY	1.75	0.169	0.191	5DL7A	419	XRAY	1.75	0.169	0.191
8PZ4A	479	XRAY	1.77	0.19	0.234	8PZ4A	479	XRAY	1.77	0.19	0.234
2PORA	301	XRAY	1.8	0.186	NA	2PORA	301	XRAY	1.8	0.186	NA
1YC9A	442	XRAY	1.8	0.19	0.221	1YC9A	442	XRAY	1.8	0.19	0.221
2WJRA	214	XRAY	1.8	0.195	0.23	2WJRA	214	XRAY	1.8	0.195	0.23
3QRAA	157	XRAY	1.801	0.2	0.228	3QRAA	157	XRAY	1.801	0.2	0.228
6HCPA	706	XRAY	1.83	0.157	0.177	6HCPA	706	XRAY	1.83	0.157	0.177
7VU2A	435	XRAY	1.85	0.174	0.197	7VU2A	435	XRAY	1.85	0.174	0.197
6I96A	677	XRAY	1.85	0.183	0.219	6I96A	677	XRAY	1.85	0.183	0.219
2X55A	293	XRAY	1.85	0.196	0.209	2X55A	293	XRAY	1.85	0.196	0.209
6TZKA	451	XRAY	1.852	0.165	0.183	6TZKA	451	XRAY	1.852	0.165	0.183
4E1SA	242	XRAY	1.855	0.178	0.233	4E1SA	242	XRAY	1.855	0.178	0.233
7AHLA	293	XRAY	1.89	0.199	0.257	7AHLA	293	XRAY	1.89	0.199	0.257
3PRNA	289	XRAY	1.9	0.169	0.19	3PRNA	289	XRAY	1.9	0.169	0.19
5MDRA	352	XRAY	1.9	0.172	0.183	5MDRA	352	XRAY	1.9	0.172	0.183
5FOKA	730	XRAY	1.9	0.19	0.211	5FOKA	730	XRAY	1.9	0.19	0.211
3FIDA	296	XRAY	1.9	0.193	0.23	3FIDA	296	XRAY	1.9	0.193	0.23
4FQEA	222	XRAY	1.93	0.215	0.248	4FQEA	222	XRAY	1.93	0.215	0.248
6QGWA	411	XRAY	1.938	0.191	0.23	6QGWA	411	XRAY	1.938	0.191	0.23
6QGWB	123	XRAY	1.938	0.191	0.23	6QGWB	123	XRAY	1.938	0.191	0.23
2GUFA	594	XRAY	1.95	0.191	0.225	2GUFA	594	XRAY	1.95	0.191	0.225
2VDFA	253	XRAY	1.95	0.227	0.272	2VDFA	253	XRAY	1.95	0.227	0.272
6FOKA	723	XRAY	1.97	0.211	0.232	6FOKA	723	XRAY	1.97	0.211	0.232
3AEHA	308	XRAY	2.0	0.177	0.222	3AEHA	308	XRAY	2.0	0.177	0.222
2GR8A	99	XRAY	2.0	0.193	0.215	2GR8A	99	XRAY	2.0	0.193	0.215
2ERVA	150	XRAY	2.0	0.2	0.233	2ERVA	150	XRAY	2.0	0.2	0.233
1KMOA	774	XRAY	2.0	0.207	0.245	1KMOA	774	XRAY	2.0	0.207	0.245
6E4VA	693	XRAY	2.0	0.208	0.235	6E4VA	693	XRAY	2.0	0.208	0.235
1XKWA	665	XRAY	2.0	0.222	0.242	1XKWA	665	XRAY	2.0	0.222	0.242
7PGEB	215	XRAY	2.0	0.222	0.253	7PGEB	215	XRAY	2.0	0.222	0.253

No. of chains with 40% identity are 42

No. of chains with 50% identity are 47

5. Extract the data with the cut-off of 50% from Uniprot and compare with CD-HIT and PISCES.



UniProt BLAST Align Peptide search ID mapping SPARQL UniRef "beta barrel" AND "membrane" Advanced | List Search

Taxonomy
Filter by taxonomy

Clusters
50% (104,263) X

UniRef 104,263 results

BLAST Align Map IDs Download Add View: Cards Table Customize columns Share

Cluster ID	Cluster name	Types	Size	Organisms	Length	Identity
<input type="checkbox"/> UniRef50_A0A8H4L299	Cluster: Outer membrane beta-barrel		1 member	<i>Fusarium albosuccineum</i>	183	UniRef50
<input type="checkbox"/> UniRef50_A0A099TY89	Cluster: Outer membrane beta-barrel protein		1 member	<i>Helicobacter muridarum</i>	257	UniRef50
<input type="checkbox"/> UniRef50_A0A099TYH7	Cluster: Outer membrane beta-barrel protein		1 member	<i>Helicobacter muridarum</i>	234	UniRef50
<input type="checkbox"/> UniRef50_A0A099UHN9	Cluster: Outer membrane beta-barrel protein		14 members	<i>Helicobacter typhionius</i> <i>Helicobacter sp. MIT 03-1616</i> <i>Helicobacter japon</i>	227	UniRef50

UniProt invites you to participa...

- First go to Uniprot
- Change the search setting to UniRef
- Search "beta barrel"AND"membrane"
- Select the 50%sequence identity

No.of clusters with cutoff of 50% in

CD-HIT : 264

PISCES: 47

UniProt: 104,263