session3_clustering_solution

September 23, 2019

1 [MIRACUM 2019][Session 3] Solution

- Kim Hee (Graduate research assistant)
- Universitätsmedizin Mannheim, Mannheim (UMM)

- This is prepared for a turorial `Data analysis tools (Datenanalysewerkzeuge)`

2 Depict a phylogenetic tree of five species

2.0.1 Protocol:

1. Import required libraries 2. Load protein of Hemoglobin subunit alpha data (download) 3. Sequence alignment - calculate the similarities of sequence 4. Visualize the result in dendrogram (phylogenetic tree)

[11]:
```
from Bio import SeqIO, Phylo
from Bio.Align.Applications import ClustalwCommandline
```

2.0.2 Protocol:

1. Import required libraries 2. Load protein of Hemoglobin subunit alpha data (download) 3. Sequence alignment - calculate the similarities of sequence 4. Visualize the result in dendrogram (phylogenetic tree)

[6]:
```
FILE_PATH = 'data/protein.fasta'
records = list(SeqIO.parse(FILE_PATH, "fasta"))
records
```

[6]:
```
[SeqRecord(seq=Seq('MVLSADDKTNIKNCWGKIGGHGGEYGEEALQRMFAAFPTTKTYFSHIDVSPGSA…KYR
', SingleLetterAlphabet()), id='rat', name='rat', description='rat Species_rat',
dbxrefs=[]),
 SeqRecord(seq=Seq('MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSA…KYR
', SingleLetterAlphabet()), id='human', name='human', description='human
Species_human', dbxrefs=[]),
 SeqRecord(seq=Seq('MSLSDTDKAVVKAIWAKISPKADEIGAEALARMLTVYPQTKTYFSHWADLSPGS…KYR
', SingleLetterAlphabet()), id='zebrafish', name='zebrafish',
```

```
description='zebrafish Species_zebrafish', dbxrefs=[]),
 SeqRecord(seq=Seq('MVLSAADKSNVKAAWGKVGGNAGAYGAEALERMFLSFPTTKTYFPHFDLSHGSA…KYR
', SingleLetterAlphabet()), id='sheep', name='sheep', description='sheep
Species_sheep', dbxrefs=[]),
 SeqRecord(seq=Seq('MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSA…KYR
', SingleLetterAlphabet()), id='chimpanzee', name='chimpanzee',
description='chimpanzee Species_chimpanzee', dbxrefs=[])]
```

2.0.3 Protocol:

1. Import required libraries 2. Load protein of Hemoglobin subunit alpha data (download) 3. Sequence alignment - calculate the similarities of sequence 4. Visualize the result in dendrogram (phylogenetic tree)

[7]:
```
clustalw_cline = ClustalwCommandline("clustalw2", infile=FILE_PATH)
stdout, stderr = clustalw_cline()
print(stdout)
```

```
 CLUSTAL 2.1 Multiple Sequence Alignments


Sequence format is Pearson
Sequence 1: rat           142 aa
Sequence 2: human         142 aa
Sequence 3: zebrafish     143 aa
Sequence 4: sheep         142 aa
Sequence 5: chimpanzee    142 aa
Start of Pairwise alignments
Aligning…

Sequences (1:2) Aligned. Score:  78
Sequences (1:3) Aligned. Score:  51
Sequences (1:4) Aligned. Score:  76
Sequences (1:5) Aligned. Score:  78
Sequences (2:3) Aligned. Score:  53
Sequences (2:4) Aligned. Score:  86
Sequences (2:5) Aligned. Score:  100
Sequences (3:4) Aligned. Score:  53
Sequences (3:5) Aligned. Score:  53
Sequences (4:5) Aligned. Score:  86
Guide tree file created:   [data/protein.dnd]

There are 4 groups
Start of Multiple Alignment
```

```
Aligning…
Group 1: Sequences:   2      Score:3061
Group 2: Sequences:   3      Score:2877
Group 3: Sequences:   4      Score:2764
Group 4: Sequences:   5      Score:2355
Alignment Score 6225

CLUSTAL-Alignment file created  [data/protein.aln]
```

2.0.4 Protocol:

1. Import required libraries 2. Load protein of Hemoglobin subunit alpha data (download) 3. Sequence alignment - calculate the similarities of sequence 4. Visualize the result in dendrogram (phylogenetic tree)

[12]:
```
newick_path = 'data/protein.dnd'
tree = Phylo.read(newick_path, "newick")
Phylo.draw_ascii(tree)
```

```
       ____________________ rat
  _____|
 |     |_____________________________________________________________ zebrafish
 |
_|          , human
 |__________|
 |          | chimpanzee
 |
 |___________ sheep
```