

End term project (Traccia B):

Uno zoo ha ampliato il suo repertorio in modo considerevole, ma non è riuscito ancora a catalogare e sistemare il tutto. Il direttore vuole creare, per una certa categoria di animali, dei percorsi di visita guidati, organizzandoli per gruppi filogeneticamente vicini, alla fine dei quali i visitatori raggiungono il capostipite (o l'antenato, se ancora esistente) del gruppo.

Delle specie ancora da catalogare sono note le corrispondenti sequenze in formato FASTA.

1. Recuperare ed annotare tutte le informazioni utili alla classificazione delle varie sequenze. Usando tutte le risorse disponibili, si tenti di trovare informazioni di base circa il genere e/o la specie; per esempio, l'habitat dove le specie comunemente vivono, ed altre caratteristiche interessanti.
2. Aiutare il direttore nella definizione dell'albero filogenetico per la visita guidata.
3. Stilare un report in cui si descrivano la metodologia utilizzata per lo svolgimento dell'analisi ed i risultati ottenuti, con particolare attenzione alla scelta degli strumenti ed ai parametri utilizzati indicando, se possibile, le motivazioni per cui l'utilizzo di particolari impostazioni (se modificate rispetto a quelle di default) hanno portato miglioramenti o peggioramenti nella ricerca.
4. (Bonus) utilizzando il linguaggio R, costruire un istogramma in cui si mostra, per ogni famiglia, il numero di elementi ad essa appartenenti.
5. (Bonus) scegliere una sequenza fra quelle proposte. Creare, usando il linguaggio R, un istogramma che mostri le occorrenze di ciascun carattere nella sequenza (suggerimento: si vedano i comandi *tables* e *as.data.frame*. Che differenza c'è fra i due?)
6. (Bonus) Calcolare, usando il linguaggio R, media, mediana, moda, max e min degli score (uno a scelta fra Total, Max score, Total score, Query cover, E value, Ident, giustificando la scelta) ottenuti durante la classificazione delle sequenze.

>Unknown1

```
SAIPYIGADLVEWIWGGFSVDKATLTRFFAFHFILPFIISALAAVHLLFLHETGSNNPSGMVSDSDKIPFHPYYTIKDILGLLVLILT  
LMLLVLFSPDLLGDPDNYTPANPLSTPPHIKPEWYFLFAYAILRSIPNKLGG  
VLALVLSILILAIIPALHTSKQRGMMFRPLSQCLFWFLVADLLTLTWIGGQPVEHPFITIGQLASILYFS  
ILLILMPISGIIENRLLKW
```

>Unknown2

```
PFIIGQLASILYFTILLVLMPIAGIENRLLKW
```

>Unknown3

```
AGMVGTAALLIRAE LGQPGTLLGDDQIYNVIVTAHAFVMIFFMVMPIMIGGFGNWLVLPLMIGAPDMAFPRMNMNSFWLLPPSF  
LLLLASSMVEAGAGTGWTVYPLAGNLAHAGASVDLTFISLHLAGVSSILGAINFITTIINMKPPAMSQYQTPLFVWSVLVTAVAL  
LLSLPVLAAGITMMLTDRNLNTTFDPAGGGDP
```

>Unknown4

```
METTLNLSMMLTLTIILVPILLSLLSTTHSIPPTITITIKTAFLTSVPMIAIFMYSGSESIISHWEW  
KFITNFKIPFSLKIDQYSMMFFPIALFVTWSILQFALWYMASEPHITKFFFLMLFIAMLTLTIANMMF  
LLFIGWEGVGIMSFLIGWWQGRAEANTAALQAVLYNRIGDIGLILSMAYLASTNTWEIQQALSPNQAPTLP LLGFILAATGKS  
AQFGLHPWLPAAMEGPTPVSA LLSSTMVVAGIFLLIRIHP LLSSTNQTALTLCLCLGALSTLFAAICALTQNDIKKIIAFSTSSQLG  
LMMVTIGLNLPLQAFFHISTHAFFKAMFLCSGSIIHALNGEQDIRKMGS LKQTLPTTTTCLTIGNLALMGTPFLAGFYSKDLIIES  
LNTSYLNAWALLLTLLATSFTATYSLRMTLLVQTSFPRMPTITPMNENTPTLINPITRLALGSIMAGLLITSYIPPTKTPMPTMPLT  
KTTAIIITTLGVILALELSNMHTLTQPKQTPPLNFSSMLGYNSLTHRLTSLTLLHSGQKIASHLIDLSWYKKMGPEGIADLQLTA  
AKTTTPLHTGLIKAYLGT FALSTFIILLSAH
```

>Unknown5

```
MMNFVLFMSLCFVLGGLAVASNPSPYYGVVGLVVASIAGCGWLVS MGVSFISLVLMVYLGGMLVVFVYSVSLAADPYPESWG  
DWRVVGYGAGIGLVVVVGVFMGAFSEVVMGGDTVNNGGLSWVRTDFSGVAMFYSSGSGLLLIAGWG LLLTLFVVLELVRGL  
SRGAIRAV
```

>Unknown6

```
PFMILALWGIVMTSSICMRQTDKLSLIAYSSVSHMGLVTTACLIQTPWSITGAMILMIAHGLTSSMLFCLANTNYERTHTRTLILA
```

RGLQTILPLMASWWLLANLTNMALPPTINLMGELMIISALFNWSTPTIILTGLGLTITAIYSLHMFLMTQRNKLPLHMTMMNPTH
TREHLIMALHMLPLTMLIMKPTLISSMLPC

>Unknown7

LPTPSNISIWVWFGSLLGLTLMIQILTGVFLMMHFSSSDATAFSSVAYTSREIWFGWLIRNLHTNGASLFFMFIFLHIGRGLYYA
SYLHENTWNIGVIMLFLLMATAFMGYVLPWGQMSFWGATVITNLLSAIPYIGDTIVPWIWGGPSVNNATLTRFTTLHFLLPFVLL
ALLITHLIFLHEQGSFNPTGLSKNADKIPFHPYYTMKDTLGAVLAASMLLTALYLPTMLGDPENFTPANPMTTPSHIKPEWYFLF
AYAILRSIPNKLGGVLAMFSSIFILFLMPALHTTNQQPMSVRPLSQLLFWILILDFLALTWIGGQPVNPPYVLLGQMASLLYFTI

>Unknown8

DECTPLTHDCTHDSCCRGPTEFKYKCDCLYPFDNSTSAWDQTELCFCVEPGVHHFLDEVMDKTIGIFG

>Unknown9

GTLYFLFGSWSGMVGTSLSLIRAEELGNPGSLIGDDQIYNVIVTAHAFIMIFFMVMPIIMIGGFGNWLVPMLGAPDMAFPRMNN
MSFWLLPPSLTLLLTSSMVENGAGTGWTVPPLSGNIAHSGASVDLAIFSLHLAGISSILGAVNFITTVINMRSTGMSFDRMPLF
AWSVMLTAIILLLSLPVLAGAITMLLTDRNINTSFFDPAGGGDPILYQHL

>Unknown10

MLLLAMAAVDSFYFLYREIARSCNCCYLEALALVGAWYTARKSITIIHDFYSLIRLHFIPRLVSKADLIKQ
YGRWAVVSGATDGIGKAYAEELASHGLNIIISRNKEKLEKVAENITEVYKVETDIIVADFSNGREIYSPIREALRDKDIGILVNN
VGVFYYPQYFTQVPEDKLWDIINVNIAAANFMIHIVLPGMVDRKKGAIVNISSGSCCKPTPQMTAYSASKAYLDHFSRALQYE
YASKGIFVQSLIPFYVATNMNTFSGFLHSNPWLVPSPKVYAQHAVSTLGISKRTTGYWSHSIQFIFAQYMPPEWLWVWGAILN
SSLRQEALSHRL