

End term project (Traccia A):

A botanic garden has expanded its collection of plants and it is classifying and catalog them. The director wants to create guided tours, arranging plants following their phylogenetic relations. A set of sequences in FASTA format of the species to classify is given:

1. Retrieve and annotate all the information useful for classifying the various sequences. Using all available resources, you try to find basic information about the gender and / or species; For example, habitat where species commonly live, and other interesting features.
2. Help the director in defining the phylogenetic tree for the guided tour.
3. Produce a report describing the methodology used to conduct the analysis and the results obtained with particular attention to the choice of instruments and parameters used, indicating, where possible, the reasons why the use of particular settings (if different from default) which produced improvements or variations in the search. (Bonus) Using the R language, construct a bar chart showing the number of elements belonging to each family.
4. (Bonus) Choose a sequence. Create, using the R language, a bar chart showing the occurrences of each character in the sequence (hint: see the commands `tabs` and `as.data.frame`. What is the difference between the two?)
5. (Bonus) Compute using the R, mean, median, trend, max and minimum score (one among Total, Max score, Total score, Query cover, E value, Ident, justifying the choice) obtained during the analysis.

```
>Unknown1
RTSISKKRIRKTIWKKKGYWVALKAFSLAKSLSTGNSKSFFVQQIQTLE
```

```
>Unknown2
LXRTXKSCRLRWLNYL
```

```
>Unknown3
MDKFKRDGEEYISYQRRFLYPLLFQEDLYAIAIDHYFNRSSCFEMENSSSNDRFSFLTVMKRLISRIHQNDLIISFLNCDQNP
VGHNSSFYSELVLEGPTVVLEVFPFLMRKSHSLEKKNEWKSFRSIHSIFSFMEDKFPHPNSIS
```

```
>Unknown4
IVVSGNDVAIKEGDLVKRTGSIVDVPGKALPGRVVDALGVPIDGKALQVNEGAERRRVEAKAPGIIERKSVHEPLQTGLKAV
DSLVPITGRGQRELIIGDRQTGKTAIAIDIISNQKRSRSHRSSGTSDSVYRVHVAIGQKRSTVAQLVQIPPEADALEYSIINAATASD
PAPPQFLAPYSGCAMGEYSRDNGMHALIIYDDPSKQAVAYRQMSLSPRRPPGREAFPGDVFYSHSRLSERAARKSDQTGAGSL
TALPVIETQAGDVSAYIPTNVIPITDGOICPETELFHRGIRPAINVGLPVS RVGSAAQLKAMKQVRGSSKPESAQYREVAASAQS
GSDLDAAT
```

```
>Unknown5
MASIDAQMRLLVPGKVEDDKLIEYDALLDRFLDILQDLHGEDIEMVQECYERSGEYEGKNDPHKLEELGNVLTSLDPGDSI
VVASSFSHMLNLANLAEVQIAYRRRNKIKRGGFADESNATTESDIEETFKRLVNQLGKSPAEVFDALKNQTVDLVLTAHPTQS
VRRSLLQKHARIRNCLSQLYGKDITPDEKQELDEALLREIQAAFRTDEIRRTPTPQDEM RAGMSYFHETIWKGVPKFLRRIDTA
LKSIGINERVYPNAPLIQFSSWMGGDRDGNPRVTPEVTRDVCLLARMMAANLYYSQIEDLMFELSMWRCSDEL RARALQLHSA
SKKDAKHIEFWKQIPNEPFRVILGDVRDKLYNTRERTRQLLSNGISDIPPEVFTFNIDEFLEPLELCYRSLCSTGDQPIADGSL
LDFMRQVSTFGLSFVKLDIRQESDRHSDVADAITRHLGIGSYKEWSEEQRQAWLLSELQGKRPLFGPDLPKTDEV RDVLDTFH
VISELPADNFGAYIISMATAASDVLVVELLQRECHVKKPLRVVPLFEKLADLEAAPAALARLFSINWYRNRIDGKQEVMI GYSDS
GKDAGRLSAGWALYKAQEDLIKVAKEFGIKLTMFHGRGGTVGRGGGPTHLAISQPPDTIHGSFRVTVQGEVIEQSFGE EHL
C FRTLQRFTAATLEHGM RPPVAPKPEWRELMDEMAVVATKEYRSIVFQDPRFVEYFRSATPELEYGRMNIGSRPSKRKPSGGIES
LRAIPWIFAWTQTRFHLPVWLGFGA AFKHVMEKDIRNLHMLQMYNEWPFRRVTIDLIEMVFAKGDPGIAALYDKLLVSDDLW
AIGEKL RANYGETKDLLLQVAGHKDLLEGDPYLKQRLRLRDSYITLNV CQAYTLKRIRDPNYHVNL RPHLSKESSTKPAAE LVK
LNPTSEYAPGLEDTLILTMKGIAAGMQNTG
```

```
>Unknown6
GVGFKAGVKDYRLTYFTPDYETKDDTILAAFRMTPQPGVPPPEEAGAAVAESSTGTWTTVWTDGLTSLDRYKGRCYNI EPVAG
EDNQFIAYVAYPLDLFEEGSVTNLFTSIVGNVFGFKALRALRLEDLRIPPAYSKTFIGPPHGIQVERDKLNKYGRPLL GCTIKPKLG
LSAKNYGRAVYECRLGGLDFTKDDENVNSQPFMRWRDRFLFVAESIFKSQAETGEIKGHYLNATAGTCEEMLKRAVFARELGA
PIIMHDYLTGGFTANTSLAFYCRDNGLLLIHHRAMHAVIDRQKNHGIHFRVLAKALRMSSGGDHIHAGTVVGKLEGERDLTLGFV
```

DLRDFIEKDRSRGIYFTQDWVSMGPVLPVASSGGIHWHPALTEIFGDDSVLQFGGGTLGHPWGNAPGGAANRVAVEAC
VQARNEGRDLATEGNEIIREAAK

>Unknown7

LGNNEVRAMASATDGLTRGMGAVDTGAPLSVPVGETTLGRISNVLGEPVDNLGPVRSSVTSPIHRSAPAFTQLDTKLSIFETG
IKVVDLLAPYRRGGKIGLFGGAGVGKTVPI TELINNIKAHGGVSVSGGVGERTREGNDLYMEMKESK VINEQNISES KVALVY
GQMNEPPGARMRVGSTALLTMAEYFRDVKQDVLLFIDNIFRFVQAGSEVSALLGRMPSAVGYQPTLGTMGSLQERITSTKEG
SITSIQAVYVPADDLTD PAPATTS AHDATTVLSRGLAAKGIYPAVDPLDSTSTMLQPWIVGEEHYETAQGVKQTLQRYKELQDI
IAIPGLDELSEEDRLTVARARKIERFLSQPFVAEVFTGSPGKYVSLSETIKGFQMILSGELDTLPEQAFYLVGNIDEATAKAAALQ
VEG

>Unknown8

MLQFPALSHLNL SGLILRLLLSIILLIPDSRIQLIRSIGPRFTLITFLYSPFFWIEFDNSTAKFQFVE
SIRWPPYSNINLYIGIDGIFLSSAVLTTF LIPRISVGVSSIKSYKKEYMIAFLIRESLMIASVSRMSDFL
LFHVFFESVSISMFRQRKIQAAYQSFLYTLSGFVFMLLAISSISFQTGTTDLRILLTTESSEWRQVLLRI
AFFASFAVKVPMVPHIRLPEAYVEAPTAGSVISAGILSKLGTYGSLRFSILMFPEVTFFFISFICTLSV
IAIITYPLTTVRQIDLKMIAYFSVAHMNFVTIGMFSLNIQIEGSISPMLSHGLVLLAPPLCVGALYDR
HKTRLVKYYGGLVSTMSIFCTISLFFTLANMSLLGSSSFIGEFLILIGAFERNSLVATPAALGMILGAAY
FLWLYNCVVFGNFKPNFLKKFSDLNRREVFI SLLFIVGVIWMGVYLEVFLECMYTCVSNLVQHGFDF

>Unknown9

MFEHDFLALFPEIFLINATIILLIYG VVFSTSKKYDYPPLVCNVSWLGLLSVLITILLVASSTPLTVANL
FYNNLIIDNFTYFCQIFLLISTASTIVMCLGYFKEESLNAFESIVLILLSTCSMLFMISAYDLIAMYLAIELQSLCFYVIAASKRDSEF
STEAGLKYFILGAFSSGILLFGCSMIYGFTGVTNFEELAKIFTGYEITLFGAQSSGIFMGILFIAVGFLFKITAVPFHMWAPDVYEG
SPTLVTAFFSIAPKISILANMVRVFIYSFYDPTWQQLFFFCSIAS MILGALAAMAQNKVKRLLAYSSIGHVGYLFIGFSCGTIEGIQ
SLLIGVFIYVLMTINVF AIVLALRQNRKYIADLGALAKTNPILAITLSITMFSYAGIPPLAGFCSKFYLFFAALGCGAYLLALIGVVT
SVISCFYYIRFVKIMYFDTPKKWILYKAMDREKSLLLAITLFLISFFFLYPSPLFLVSHQMALSCL

>Unknown10

MEEFQVYLELNRSRRHDFLYPLIFREYIYALAHDHGLNKSMIFLENQGYVVKFSSLIVKRLIIRMDQQNHLIISANDSNQNPFFG
HNNNLYSQMISAGFAVIVEIPFSLRLVSYSGEEVAKSHNLQSIHSIFPFLEDKFSHLNYVLDVLIHPHILEILVQALRYWVKDA
SSLHLLRFSLYEYCNLKSFITPKKSI SIFNPRLFFFLYN SHACEYESIFLFLRNQSSHLRSTSSGVFLERIYFYGKIEYLLLEVYNDFQ
NNLWLFKDPFIHFIRYQ GKAILASKDTSLLMNKWKYFVDLWQYHFYMWSQSGRVRINQLSKYSLDFLGYLSSVRLNPSVRS
QMLENSFIIDNAMKKLDTRIPISLIGSLSKAKFCNTLGHPI SKPTWADSSDSIIDRFVRCRNL SHYHSGSSKKKSLYRIKYIL
RFSCVKT LARKHKNTVRAFLKRLGSEFLEEFFTETEEHVFSLIFPRVFFTSRKLYRGRIWYLDIICINALVNHE