

PRACTICAL 8

Question 1. Obtain the consensus phylogenetic tree for the following two sets of sequences:

- (1) Set 1: tim.dat
- (2) Set 2: tim-hemo.dat

Hint:

- i Multiple sequence alignment using MAFFT
- ii Save in Phylip format
- iii Install Phylip (windows) / Phylip (Mac OS)
- iv Bootstrapping (Seqboot program)
- v Maximum likelihood method (proml program)
- vi Consensus tree (Consens program)
- vii Use TreeView / MEGA-X to view the tree
- viii NJ and UPGMA methods (protdist and neighbor programs)

Refer to practical 10.ppt for detailed steps.

Solution. A phylogenetic tree is a graphical representation of the evolutionary relationships between biological entities, usually sequences or species. Relationships between entities are captured by the topology (branching order) and amount of evolutionary change (branch lengths) between nodes. A consensus tree is a representation of a set of phylogenetic trees that attempts to summarise them in a way that captures their most frequently occurring characteristics. More specifically, a consensus tree usually represents the splits that occur in a majority of the trees; in a tree a split corresponds to an edge.

So, in the given question, we are being asked to find the consensus phylogenetic tree for the two sets of sequences given. The steps to be followed for the first set sequences is as follows:

i **Multiple sequence alignment using (MAFFT)**

The input file is uploaded to the MAFFT website given as a plain text dat file. The MAFFT is a server that gives the multiple sequence alignment of the given set of sequences. Other parameters seen here are left to default. The additional parameters include direction of nucleotide sequence, output order of sequences, title length in the CLUSTAL output, etc. Some additional settings include strategies mediating them, progressive methods, iterative refinement methods, parameters like scoring matrix, imposing penalties, etc. We have set all these parameters to default as seen in the images below.

Multiple sequence alignment and NJ / UPGMA phylogeny

Input:
 Paste protein or DNA sequences in fasta format. [Example](#)

```
>tr|E1K4V0|E1K4V0_9EURY Xylose isomerase domain protein TIM barre.
MRIGISSSIFLDSKDLKYALEYLEKKVRYVELNCDGNINMEKENMDIPNSYDLNNTLH
CPLTDLNLSSFRDKIRKVSLEDFEDILKTADKVNANLVHLHFGYCVFKYDYKSLNALIQ
SLKDLNNIQKESYIKITTIENMPSYSMFMPREFTDEIINNIGDLGITFDIGHSFLNKGK
FLNDELIIKISHIHHDNNGEFDEHLAIGKGRIDFEKYKGNIKKIRGKIKLVEMQKNSIN
DLDLCTIRLKNLLV
>tr|E7QXB8|E7QXB8_9EURY Xylose isomerase domain protein TIM barre.
MIVAGKCPPTADELRAASERGFDAVELHLITDLDLDAIEETTAACRAAPVDVSVHTPHVG
LDELAIVQRANDLCERLDATLVVHSTKIFLSNLGYVLDRIDITVPHGFENSTGHSRHFLT
NVLLDEGRPLVLDTAHLYTAEAEYRSILETLAADDISIFVHCCDGTIKITDGLAFGTGT
MDMERVITALEHNYDGIIVVLEVPDEQADALELWRDVIRG
>tr|D2S0D9|D2S0D9_HALTIV TIM-barrel signal transduction protein OS
MEFMREEAVRLEETVSNDEPIIGAGAGTGMASAKFAERGGVDLLIYNSGRYRMNGRGS
AGLLPYGDANEIVVMGRQVLPVVEDTFVLAVNGTDPFRQMDVFIEDLKRRGFSGVQNF
```

or upload a plain text file: No file chosen

☐ Use [DASH](#) to add homologous structures (protein only)

☒ Output original plus DASH sequences ☐ Output original sequences only

☐ Give structural alignment(s) externally prepared

☐ Allow unusual symbols (Selenocysteine "U", Inosine "I", non-alphabetical characters, etc.) [Help](#)

FIGURE 1. Loading the input set of sequences to MAFFT

UPPERCASE / lowercase:

☐ Same as input

☒ Amino acid → UPPERCASE / Nucleotide → lowercase

Direction of nucleotide sequences: [Help](#)

☒ Same as input

☐ Adjust direction according to the first sequence (accurate enough for most cases)

☐ Adjust direction according to the first sequence (only for highly divergent data; **extremely slow**)

Output order:

☐ Same as input

☒ Aligned

Title length in Clustal format (only first word is used as title):

(10 – 100)

Job name (optional; used as output file name and subject of emails):

(basic Latin alphabet, number and space only)

Notify when finished (optional; recommended when submitting large data):

Email address:

FIGURE 2. Extra factors required for MSA

Advanced settings

Strategy:

☒ Auto (FFT-NS-1, FFT-NS-2, FFT-NS-i or L-INS-i; depends on data size) [Updated](#)

Progressive methods

☐ FFT-NS-1 (Very fast; recommended for >2,000 sequences; progressive method)

☐ FFT-NS-2 (Fast; progressive method)

☐ G-INS-1 (Slow; progressive method with an accurate guide tree)

Iterative refinement methods

☐ FFT-NS-i (Slow; iterative refinement method)

☐ E-INS-i (Very slow; recommended for <200 sequences with multiple conserved domains and long gaps; **2 iterative cycles only**) [Help](#) [Updated](#) (2015/Jun)

☐ L-INS-i (Very slow; recommended for <200 sequences with one conserved domain and long gaps; **2 iterative cycles only**) [Help](#)

☐ G-INS-i (Very slow; recommended for <200 sequences with global homology; **2 iterative cycles only**) [Help](#)

☐ Q-INS-i (Extremely slow; secondary structure of RNA is considered; recommended for a global alignment of highly divergent ncRNAs with <200 sequences × <1,000 nucleotides; the number of iterative cycles is restricted to two, 2016/May) [Help](#)

Align unrelated segments, too? *in Alpha Testing (2014/Mar)*

If the input data is expected to be globally conserved but locally contaminated by unrelated segments, try 'Unalignlevel>0' and possibly 'Leave gappy regions'.

Unalignlevel:

☒ Default

This feature is available only when G-INS-1 or G-INS-i is selected in the Strategy section above.

☒ Try to align gappy regions anyway

☐ Leave gappy regions (Not recommended for >~1,000 sequences)

Parameters:

Scoring matrix for amino acid sequences:

Scoring matrix for nucleotide sequences:

! Switch it to '1PAM / k=2' when aligning closely related DNA sequences.

Gap opening penalty: (1.0 – 5.0)

Offset value: (0.0 – 1.0)

Score of π in nucleotide data: [Example](#)

☒ Long stretches of π s tend to be gapped (excluded from the alignment).

(nzero) π has no effect on the alignment score.

☐ (mwildcard) π is treated like a wildcard. **Experimental option** (2016/Apr/26)

! Try this if π s should be aligned with usual letters.

Guide tree:

☒ Default ☐ UPGMA

☒ Output guide tree

To display the tree, follow the "Refine dataset" link in the result page.

MAFFT-homologs (Collects homologs by PSI-BLAST and aligns homologs with input sequences; Protein only): [Help](#)

☐ On

☐ Show homologs (if any)

Number of homologs: (5 – 600)

Threshold: $E = 1e-10$ (1e-1 – 1e-40)

☐ Use SwissProt (less comprehensive and requires shorter search time; previous default)

☒ Use UniRef50 (more comprehensive and requires **longer search time**) 2019/Mar

Plot LAST hits (DNA only):

☒ The top sequence vs the others ☐ The longest sequence vs the others

☒ Plot and alignment ☐ Plot only ☐ Alignment only

Threshold:

FIGURE 3. Additional settings regulated based on use case for MSA

Below are the outputs. The outputs of MAFFT MSA are in CLUSTAL format.

```

CLUSTAL format alignment by MAFFT (v7.511)

tr|E1K4V0|M-----RIGISSSIFLDS--DKDLKVALEYLEK-KVRYVELNC-DG---NI
tr|E1K6G0|MSCFYLMFIGENMCELKIGCSTLFFW---EYSIEIICDIFKDMNLNCHMEFF---ENP
tr|E1K572|MLGVCMSRSLKGIKPCISIG-----NLVQNRVMTFKHIDWGL-----
tr|E1K541|M-----LNFGTAGIPINVR-PRITIGAFDFLKKIGLNAMEIEFVRG-----
tr|F2L4F0|M-----ARWMLGAGIPYAKARKDTVEGIRAVKELGLNAMEIEFVRG-----
tr|F4B9Y8|M-----PKYLGAGIPISAK-GKSTIEGVKVKELGLNAMEIEFVRG-----
tr|E7QXB8|M-----IVAGK-----CPTADELRAASERGFDAVELHL-----
tr|D2S1E6|M-----GVGYTTIMYDF--AEVLSEGLGDFACRYDGVGEIGL-----
tr|D2S1F8|M-----GTGYTTIMYG--AESIEDALGDIAACRYDGVGEISL-----
tr|E7QZB7|M-----ARPAVQLYSLRTV--DEPLAKLITRAGDAGFEVGEFA-----
tr|E7QZC9|M-----SNPSDYSISVQSVVFR--SRSLGDLDALESTDIDRLWLG-----
tr|E7QQB9|M-----DIGVHTFPLY--GESLEDALAYLNDIGVDAIEPGV--GGYPGDT
tr|D2S0D9|M-----EPMREEAVRRLLETVSND-----
tr|E7QVD7|M-----KFTREESLARLEXTIANG-----
*
tr|E1K4V0|NVMEKENMDIPNSYD-----LNVTLHCP--LTDNLSSFRDKIRKVSLDVEFD
tr|E1K6G0|EFW--ERRNDLDYICSLKELKH---FNLIHAP--TIELNPSSTNDVYHEASIKETLW
tr|E1K572|-----HYPKIKIKNK--SIIGFAP--IVNLG--DKDKSKTCLMTLKN
tr|E1K541|-----VNIKEKA-EELKDYSDN--IILSVHAP-Y-YINLN--AKEPEKVKSSMDRIIN
tr|F2L4F0|-----VRMGVEAA-ERAGEAARELG-VKLSVHAP-Y-FINLC--SDEAEKVEASIKRLWD
tr|F4B9Y8|-----VMSIEKA-KELGEVAKTIG-VKLSVHAP-Y-YINLC--SEEEKIEASKNRIE
tr|E7QXB8|-----TTDDLAIEETTAACRAAFVIVVSVHTP-H-----
tr|D2S1E6|-----GKVEYIGV-DSLQESLEE---HGLDIYC-----VMAWLNNEEDVDAVR
tr|D2S1F8|-----EKIRANDA-ETVDRWLE--YDLEFYL-----AMSEWIEDDAVRVID
tr|E7QZB7|-----NRITDTDA-DAVKAALDDGGIESVAHVHG--IDELE-----DDLDE
tr|E7QZC9|-----EHLSPEDD-EATIAAGRRII-EESVAVDGYGVIDIE-----DTGEARD
tr|E7QQB9|HLFPDEYLDDEEAQDELHDLAE--YDMRISAL--ATHNNPLHFDDEEAEADTELRE
tr|D2S0D9|-----EPIIAGAGTGMASAKFAERGVDLLI-----TNSGRYRNNGRSLAGLLPYGD
tr|E7QVD7|-----EPIIAGAGTGISAKFAERGVDMLI-----TNSGRYRNNGRSLAGLLPYGD
*
tr|E1K4V0|ILKTADRVNANL-----VVL-HFGYCVFKY-DYKKSINALIQLSKD
tr|E1K6G0|SINLAKLYGARY-----ITI-HPGKRPTKRPPKQEDRFYRYLDE
tr|E1K572|IIDEIKGYNYLT-----IHL-HNGKEPKD-----DTLNNLSE
tr|E1K541|SAKIISIFGKKS-----KKNKN-VVF-HPGYYLKQD-K-NRVYRIMVKNINI
tr|F2L4F0|SVDRARHMGAWI-----VVV-HAAYYKLG-P-DKCTDAVKERLGE
tr|F4B9Y8|TATRAEAMGADA-----IAT-HVAFYFKKS--KREELDEVSSLSE
tr|E7QXB8|-----VGL-----DELA-YVQR
tr|D2S1E6|GAGIAELDARF-----LGI-LPPPRGQTD-D-----ATFDEWLR
tr|D2S1F8|DIPVAADLGAEF-----VGI-LPPQARHD-G-----DIVERWLSR
tr|E7QZB7|TVSFYRSFGCDR-----LVV--PWLDPHF-ESEAAIDETAERLTS
tr|E7QZC9|SFAFAADLGAEY-----VTNYPFAR-----DDITE--E
tr|E7QQB9|AIRLADQLDVGTVCFSGLPAGGFNDPEYNWITAPWFSHEEAL-E-YQWEDVAIPWQE
tr|D2S0D9|ANEIVDMGRQV-----LP-----VVEDTPVLAVNGTDPFRQMDVFIEDLKR
tr|E7QVD7|ANKIVVMGHEV-----LP-----VVEDTPVLAVNGTDPFRQMDVFIEDLKR
*
tr|E1K4V0|LNNIQEYSI---KITIENMPYS---MFMFREPTD--ETINNLG-----
tr|E1K6G0|VLDYAKNNI---TLCLENLPKI---NYICYSPOEMGEVLNKNYNY--DDPNNDKY
tr|E1K572|ISDYAKNNI---KLCIENLR-----KGFSSNNPNNVIMVDIC-----
tr|E1K541|ILSYLTENKI-NAMLRPETTGK-----ISQGNVDELIRLSEEL-----
tr|F2L4F0|LTDRLDEGIDVDFIGVETIAR-----NNQFSGVEEAFGLAKELK-----
tr|F4B9Y8|ILDKSKELGIVNVKGIETMAK-----ETALGTDEVILSKELDR-----
tr|E7QXB8|ANDLCERLDA--TLVHSTKIEL-----SMLGYVLDRIDITVPHGFENSTGH
tr|D2S1E6|ICTAADEAGV--TPVHHHG-----ASHIESPEEIEEWLERAPD-----
tr|D2S1F8|ISEAAVDAGL--RPLVHHG-----ATSVEQPDEIRRYLDAVD-----
tr|E7QZB7|LARAFAEDV--ELGYHNDH-----EFETVGGRPAPERFAEASGD-----
tr|E7QZC9|LVDLAEAFDL--DVGHNHST--VHDDLSTVFSGIDDRSVRVDYDHP-----
tr|E7QQB9|LEAFADDDHV--DLAEMHF-----NMLVHEFAGMLELRNATGE-----
tr|D2S0D9|-----RGFSGVQNFET-----VGLIDEDSQFRNLEETGMGY--DKEVEMI
tr|E7QVD7|-----RGFSGVQNFET-----VGLIDEDSQFRNLEETGMGY--DKEVEMI
*
tr|E1K4V0|-----DLGITFDIGHSFLN-----KNMGK--FLDNDELI
tr|E1K6G0|KKNKNNDL-----NLKMTLDFAHAK-----EEAQN--FV--DRLN
tr|E1K572|-----NCNITFDIGH-----TDYKREDEFI--DIFS
tr|E1K541|-----NILECIDFAHYARS LGKI--NDYDSFYKIM--ENME
tr|F2L4F0|-----RVRFVDMWGLHARSNGT--IDYGS--VL--DLWR
tr|F4B9Y8|-----QIIPYIDMAHFAHQSGG--IDYGE--IL--DRLT
tr|E7QXB8|SRHFLTNVL--LDE-----GREVLDTAHLTYAE--AYRS--IL--ETLL
tr|D2S1E6|-----NLELLYDTAHQYQY-----GDVID--GI--HRFA
tr|D2S1F8|-----GLELLFDTAHYYPYDGNFDPGDVID--GL--ERFA
tr|E7QZB7|-----ALELEVDLGWATAAG-----ADPTA--FL--ERWS
tr|E7QZC9|-----RLGVCIDTGHFLVMD-----ESPAD--VI--STVG
tr|E7QQB9|-----RVGANFDPHLYWQG-----IDVPE--AI--RFLG
tr|D2S0D9|REAAEQEMLTCPIVFTEEQAREMTAGADIVSHMGLTTS--GDIGAE-TALDLDA
tr|E7QVD7|QEAEEQMLTCPIVFTEEQAREMTAGADIVSHMGLTTS--GDIGAE-TALDLDA
*
tr|E1K4V0|KK-----I--SHIHHDNNGEFDE-----HLAIGKRIDFEK
tr|E1K6G0|QY-----I--KNIHISGVVNGKD-----HYPLELSQIDFSK
tr|E1K572|NR-----I--YNVHYVELEKDKI-----GHIAPNNDNLRS
tr|E1K541|NILGKKAINDMHVHLSGIEFGKGGEKN-----HLPLDNSNFNYKD
tr|F2L4F0|TTFGEHM--HTHFTSVKRYNRGRIVDE-----REFISANMPFPEF
tr|F4B9Y8|KELGLTH--NSHPEGLEIRANKYVDI-----RSISVYMPFPEF
tr|E7QXB8|AA--DDISIPVHCCDGTKITD-----GLAFGTGMDMER
tr|D2S1E6|DD-----I--AYVHLKDIPTVDFQDHVDLTAGNVQYDLSFVLISFVLDGGVDFEA
tr|D2S1F8|DD-----I--GYVHLKDVDPVKDFAANRDALSDADFLDNVINYFRSFDLGEGLDFA
tr|E7QZB7|DR-----I--PLVHVSDADETRS-----PIEVGDGVLDVFA
tr|E7QZC9|DR-----I--VAVHLKDTSDDEI-----EDLPAGATLDLFT
tr|E7QQB9|EH--DAI--HHFHAKDTKVYSQARYGVLDTPYTEANRSLFRSVGYGHGEHGWKD
tr|D2S0D9|ER-----VQAHHDAAKVNDVVLVIC-----HGGPIAWFDDAEY
tr|E7QVD7|ER-----VQAHHDAAKVNDVVMVIC-----HGGPIAWFDDAEY
*
tr|E1K4V0|YKGNKK---IKGIKLV-----MQNKSINDLDCITRLKNLL-----
tr|E1K6G0|PVNDLVYKYNKCNFNLEIDDKNCKIRSKNEKIEFIEKEAYLENLINKN-----
tr|E1K572|VLDKLLDNK-----CDFWLIEIMKLEIIYTKNLEEDYLDNH-----
tr|E1K541|VLKILKDFN-TSGTVICE-----SPRMEYDALILKRVME-----
tr|F2L4F0|LAELSIRD-VAITLICE-----SFLLDQDALLMKEILEKHGVL-----
tr|F4B9Y8|LAELIKRD-TSTLICE-----SELENDALRMKLELDGYKFG-----
tr|E7QXB8|VITALHEN--YDGIVVLE-----VMEDEQDALELWDRVIRG-----
tr|D2S1E6|VDEALDEIG-YDGOITIE-----IENRDDRVLVHAKRNDIHWDTAID-----
tr|D2S1F8|IFETLSDAG-YEGHYTIE-----IENRTERPLVHAKRNDYWAARVN-----
tr|E7QZB7|CASAVRADG-VEWA-IYEHD-----EPEAPLESLSHGGVGLGR-----
tr|E7QZC9|VLGLFDHGAVDAPLVVEYE-----LPDDRVLPALEAEINVRTAVEGG-----
tr|E7QQB9|IVSALRLVD-YDGALSIEHE-----DSLTSREGLEKAVELLQRAVFRITPG
tr|D2S0D9|VLNNTGVAGFFGASSLERLPTEETATNQAREFKSIEFR-----
tr|E7QVD7|VLNNTGVAGFFGASSLERLPTEETATNQAREFKSIEFR-----
*
tr|E1K4V0|----V
tr|E1K6G0|----K
tr|E1K572|----K
tr|E1K541|----L
tr|F2L4F0|----V
tr|F4B9Y8|----E
tr|E7QXB8|----R
tr|D2S1E6|----A
tr|D2S1F8|----A
tr|E7QZB7|----F
tr|E7QZC9|----R
tr|E7QQB9|EAYWA
tr|D2S0D9|----P
tr|E7QVD7|-----

```

FIGURE 4. MAFFT MSA output in CLUSTAL format

ii Save in Phylip format

Now, we have the multiple sequence alignment of the set of sequences in CLUSTAL format. However, we are using the Phylip platform to compute the phylogenetic trees. Hence the MSA output should be compatible with that software. Hence, we convert the MSA output from the CLUSTAL to Phylip—Phylip4 format using the bio-sequence conversion tool.

The screenshot shows the 'Readseq -- biosequence conversion tool' interface. At the top, it says 'Sequence data' and 'Paste data or URL in box below'. A text box contains the URL 'http://localhost/alignment/server/spool/_out.2404121242555Yrjuh3ZFXKtIvblkM'. Below this are 'Submit' and 'Reset' buttons. A note says 'See [here](#) for help.' Below that, a red warning states: 'Format conversion will time out ~10 minutes after starting. If fails, please download [the zipped Fasta format file](#) and convert it locally.' The 'Options' section is divided into two columns. The left column has 'Output sequence format:' with a dropdown set to 'Phylip|Phylip4', 'Return biosequence data:' with radio buttons for 'Download to file' (selected) and 'View in browser', and 'Change sequence case to' with radio buttons for 'No change' (selected), 'lower', and 'UPPER'. The right column has checkboxes for 'Remove gap symbols:' (with a '-' input) and 'Calculate checksum of sequences', a section 'Select ☒ all, or ☐ sequences by number:' with an input field, and a checkbox for 'Translate bases (list as from-base:to-base pairs)' with an input field. At the bottom, it says 'Readseq by D.G. Gilbert, 2.1.26 (18-Oct-2007)' and 'Software at <http://iubio.bio.indiana.edu/soft/molbio/readseq/java/>'.

FIGURE 5. Bio-sequence conversion tool

iii Install **Phylip** (windows) / **Phylip** (Mac OS)

Now, the PHYLIP package needs to be installed. PHYLIP is a free package of programs for inferring phylogenies. It is distributed as source code, documentation files, and a number of different types of executables. Methods that are available in the package include parsimony, distance matrix, and likelihood methods, including bootstrapping and consensus trees. Data types that can be handled include molecular sequences, gene frequencies, restriction sites and fragments, distance matrices, and discrete characters.

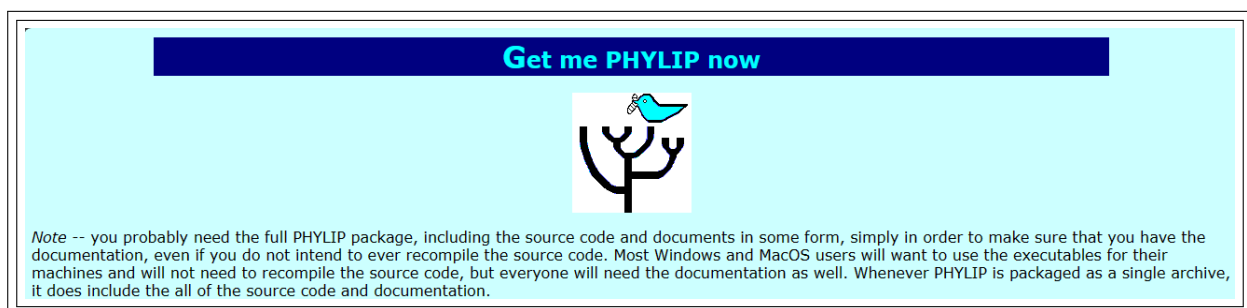
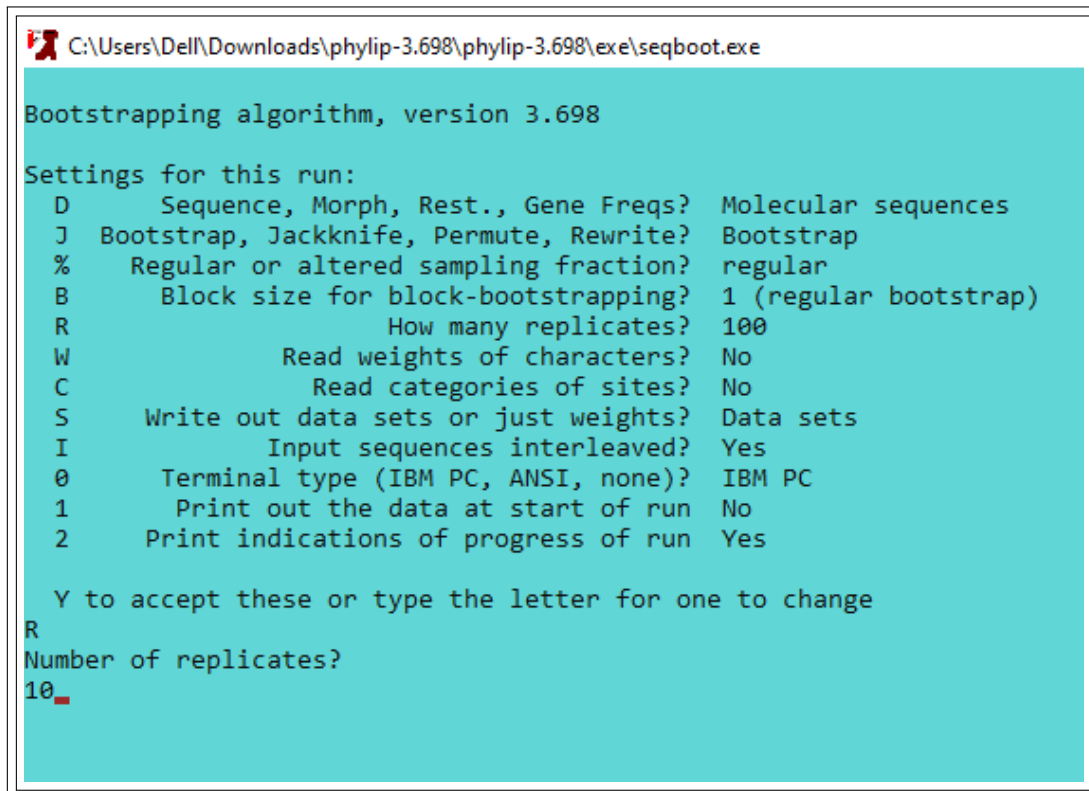


FIGURE 6. Phylip package installation

iv Bootstrapping (Seqboot program)

SEQBOOT is a general bootstrapping tool. It is intended to allow you to generate multiple data sets that are resampled versions of the input data set. Since almost all programs in the package can analyze these multiple data sets, this allows almost anything in this package to be bootstrapped, jackknifed, or permuted. SEQBOOT can handle molecular sequences, binary characters, restriction sites, or gene frequencies. The downloaded MSA (in phylip format) is given as input to the seqboot. Number of replicates is changed from 100 to 10 for ease. The output of the seqboot is stored to outfile. Below are the images of prompt windows of seqboot and the output example for one of the bootstrapped sets.



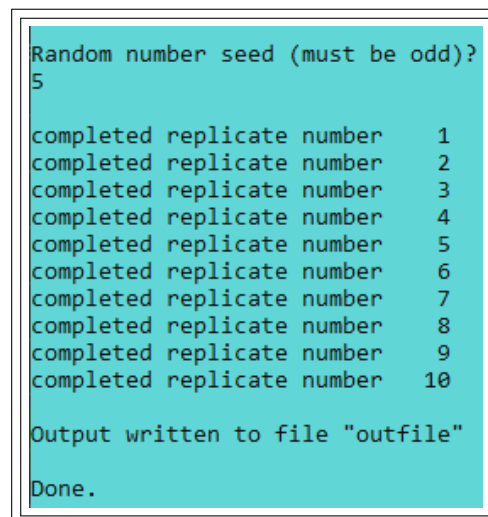
```
C:\Users\Dell\Downloads\phylip-3.698\phylip-3.698\exe\seqboot.exe

Bootstrapping algorithm, version 3.698

Settings for this run:
D      Sequence, Morph, Rest., Gene Freqs?  Molecular sequences
J      Bootstrap, Jackknife, Permute, Rewrite?  Bootstrap
%      Regular or altered sampling fraction?  regular
B      Block size for block-bootstrapping?  1 (regular bootstrap)
R      How many replicates?  100
W      Read weights of characters?  No
C      Read categories of sites?  No
S      Write out data sets or just weights?  Data sets
I      Input sequences interleaved?  Yes
0      Terminal type (IBM PC, ANSI, none)?  IBM PC
1      Print out the data at start of run  No
2      Print indications of progress of run  Yes

Y to accept these or type the letter for one to change
R
Number of replicates?
10_
```

FIGURE 7. Seqboot for bootstrapping



```
Random number seed (must be odd)?
5

completed replicate number 1
completed replicate number 2
completed replicate number 3
completed replicate number 4
completed replicate number 5
completed replicate number 6
completed replicate number 7
completed replicate number 8
completed replicate number 9
completed replicate number 10

Output written to file "outfile"

Done.
```

FIGURE 8. Seqboot post bootstrapping

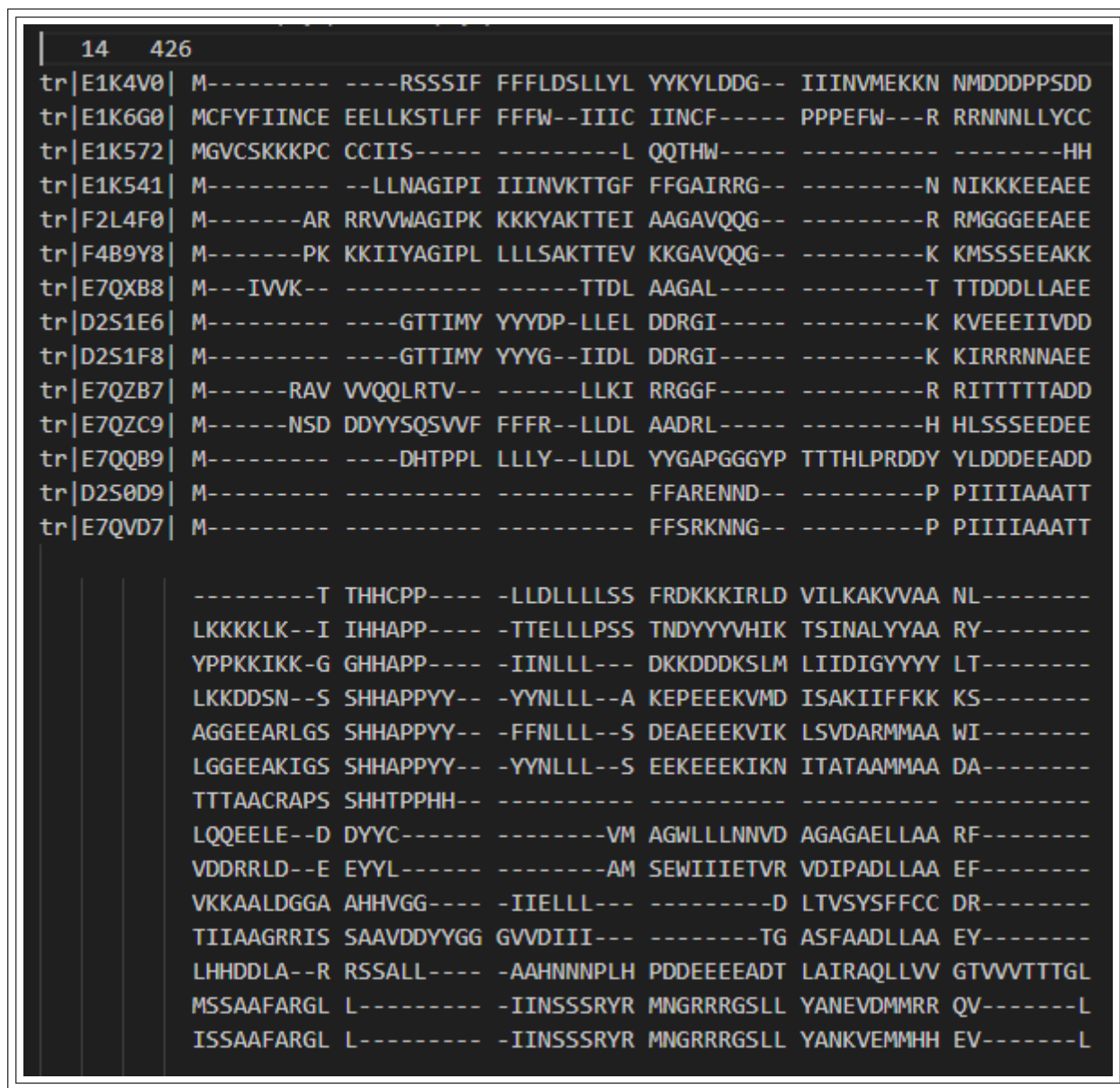


FIGURE 9. Seqboot bootstrapped single example

v Maximum likelihood method (proml program)

Maximum Likelihood (ML) on molecular sequence data uses a model of evolution, which is typically a family of trees with n taxa at their leaves, and a substitution model. The parameters of the substitution model describe probabilities of changes in character states (e.g., point mutations in DNA nucleotides). Given a set of n observed sequences, the goal is to find the best explanation for the data within the model space. In our context, this usually means a weighted tree (where the weights are parameters of the substitution model for each edge) which maximizes the likelihood (the conditional probability under the model of generating the observed sequences). There are two optimization problems related to ML in phylogenetics. The first is to optimize the branch (edge) lengths for a given tree or trees. The second is to find the ML tree by searching in the tree space.

Now, we have bootstrapped data for the set 1 of sequences stored in outfile. This is given as input to the proml. The output of the proml is saved in work1a. The parameter of "Analyze multiple datasets" is changed to 10 datasets. And the output also includes a phylogeny tree based on maximum likelihood function, which is saved in outtree.

C:\Users\Dell\Desktop\Courses\Sem_VI\BT3040\Assignment10\phylip-3.698\exe\proml.exe

```
proml.exe: can't find input file "infile"
Please enter a new file name> outfile

proml.exe: the file "outfile" that you wanted to
    use as output file already exists.
    Do you want to Replace it, Append to it,
    write to a new File, or Quit?
    (please type R, A, F, or Q)
F
Please enter a new file name> work1a_
```

FIGURE 10. proml for maximum likelihood based tree

C:\Users\Dell\Downloads\phylip-3.698\phylip-3.698\exe\proml.exe

```
Amino acid sequence Maximum Likelihood method, version 3.698

Settings for this run:
  U          Search for best tree?  Yes
  P      JTT, PMB or PAM probability model?  Jones-Taylor-Thornton
  C          One category of sites?  Yes
  R          Rate variation among sites?  constant rate of change
  W          Sites weighted?  No
  S      Speedier but rougher analysis?  Yes
  G          Global rearrangements?  No
  J  Randomize input order of sequences?  No. Use input order
  O          Outgroup root?  No, use as outgroup species  1
  M          Analyze multiple data sets?  No
  I      Input sequences interleaved?  Yes
  0  Terminal type (IBM PC, ANSI, none)?  IBM PC
  1      Print out the data at start of run  No
  2  Print indications of progress of run  Yes
  3          Print out tree  Yes
  4      Write out trees onto tree file?  Yes
  5  Reconstruct hypothetical sequences?  No

  Y to accept these or type the letter for one to change
M
Multiple data sets or multiple weights? (type D or W)
D
How many data sets?
10

Random number seed (must be odd)?
5
Number of times to jumble?
3
```

FIGURE 11. proml parametric change

```
Select C:\Users\Dell\Desktop\Courses\Sem_VI\BT3040\Assignment10\phylip-3.698\exe\proml.exe

10. tr|F4B9Y8|
11. tr|E7QXB8|
12. tr|E7QZC9|
13. tr|D2S1E6|
14. tr|D2S1F8|

Adding species:
 1. tr|E1K4V0|
 2. tr|E1K541|
 3. tr|E7QXB8|
 4. tr|F2L4F0|
 5. tr|E1K572|
 6. tr|E7QVD7|
 7. tr|E7QZC9|
 8. tr|F4B9Y8|
 9. tr|E7QQB9|
10. tr|D2S0D9|
11. tr|E7QZB7|
12. tr|D2S1F8|
13. tr|D2S1E6|
14. tr|E1K6G0|

Output written to file "work1a"

Tree also written onto file "outtree"
```

FIGURE 12. proml output on command window

vi Consensus tree (Consens program)

Consense reads a file of computer-readable trees and prints out (and may also write out onto a file) a consensus tree. At the moment it carries out a family of consensus tree methods called the MI (M-sub-L) methods (Margush and McMorris, 1981). These include strict consensus and majority rule consensus. Basically the consensus tree consists of monophyletic groups that occur as often as possible in the data. If a group occurs in more than 50% of all the input trees it will definitely appear in the consensus tree. The outtree from proml is given as input to consens program. The output file is work1cons and the output consensus tree is work1consttree.

```
C:\Users\Dell\Desktop\Courses\Sem_VI\BT3040\Assignment10\phylip-3.698\exe\consense.exe

consense.exe: can't find input tree file "intree"
Please enter a new file name> outtree

consense.exe: the file "outfile" that you wanted to
    use as output file already exists.
    Do you want to Replace it, Append to it,
    write to a new File, or Quit?
    (please type R, A, F, or Q)
F
Please enter a new file name> work1cons
```

FIGURE 13. consens on command window


```
C:\Users\Del\\Desktop\Courses\Sem_VI\BT3040\Assignment10\phylip-3.698\exe\consense.exe
Consensus tree program, version 3.698

Settings for this run:
C      Consensus type (MRe, strict, MR, Ml):  Majority rule (extended)
O      Outgroup root:  No, use as outgroup species  1
R      Trees to be treated as Rooted:  No
T      Terminal type (IBM PC, ANSI, none):  IBM PC
1      Print out the sets of species:  Yes
2      Print indications of progress of run:  Yes
3      Print out tree:  Yes
4      Write out trees onto tree file:  Yes

Are these settings correct? (type Y or the letter for one to change)
y

consense.exe: the file "outtree" that you wanted to
use as output tree file already exists.
Do you want to Replace it, Append to it,
write to a new File, or Quit?
(please type R, A, F, or Q)
F
Please enter a new file name> work1constree

Consensus tree written to file "work1constree"

Output written to file "work1cons"

Done.
```

FIGURE 14. consens output on command window

vii Use **TreeView** / **MEGA-X** to view the tree

TreeView is a program for displaying and printing phylogenies. The program reads most NEXUS tree files (such as those produced by PAUP and COMPONENT) and PHYLIP style tree files (including those produced by fastDNaml and CLUSTALW). The trees for the above methods and the ones below are enclosed after the next step.

viii **NJ and UPGMA methods (protdist and neighbor programs)**

PROTDIST:- This program uses protein sequences to compute a distance matrix, under three different models of amino acid replacement. The distance for each pair of species estimates the total branch length between the two species, and can be used in the distance matrix programs FITCH, KITSCH or NEIGHBOR. This is an alternative to use of the sequence data itself in the parsimony program PROTPARS.

The program reads in protein sequences and writes an output file containing the distance matrix. The three models of amino acid substitution are one which is based on the PAM matrixes of Margaret Dayhoff, one due to Kimura (1983) which approximates it based simply on the fraction of similar amino acids, and one based on a model in which the amino acids are divided up into groups, with change occurring based on the genetic code but with greater difficulty of changing between groups. The program correctly takes into account a variety of sequence ambiguities. The three methods are:

- The Dayhoff PAM matrix
- Kimura's distance
- The categorical distance

The outfile from step (iv) is given as input to the protdist. The output of the protdist is saved in work1-protdist. The parameter of "Analyze multiple datasets" is changed to 10 datasets. The output includes a distance matrix that is stored in above mentioned file.

```

C:\Users\De\l\Desktop\Courses\Sem_VI\BT3040\Assignment10\phylip-3.698\exe\protdist.exe

Protein distance algorithm, version 3.698

Settings for this run:
P Use JTT, PMB, PAM, Kimura, categories model? Jones-Taylor-Thornton matrix
G Gamma distribution of rates among positions? No
C One category of substitution rates? Yes
W Use weights for positions? No
M Analyze multiple data sets? No
I Input sequences interleaved? Yes
0 Terminal type (IBM PC, ANSI)? IBM PC
1 Print out the data at start of run No
2 Print indications of progress of run Yes

Are these settings correct? (type Y or the letter for one to change)
m
Multiple data sets or multiple weights? (type D or W)
d
How many data sets?
10

```

FIGURE 15. protdist input window

```

Output written to file "work1-protdist"

Data set # 10:

Computing distances:
tr|E1K4V0|
tr|E1K6G0| .
tr|E1K572| ..
tr|E1K541| ...
tr|F2L4F0| ....
tr|F4B9Y8| .....
tr|E7QXB8| .....
tr|D2S1E6| .....
tr|D2S1F8| .....
tr|E7QZB7| .....
tr|E7QZC9| .....
tr|E7QQB9| .....
tr|D2S0D9| .....
tr|E7QVD7| .....

Output written to file "work1-protdist"

Done.

```

FIGURE 16. protdist output window after changing parameter

NEIGHBOR:- It constructs a tree by successive clustering of lineages, setting branch lengths as the lineages join. The tree is not rearranged thereafter. The tree does not assume an evolutionary clock, so that it is in effect an unrooted tree. It should be somewhat similar to the tree obtained by FITCH. The program cannot evaluate a User tree, nor can it prevent branch lengths from becoming negative. However the algorithm is far faster than FITCH or KITSCH. This will make it particularly effective in their place for large studies or for bootstrap or jackknife resampling studies which require runs on multiple data sets.

The UPGMA option constructs a tree by successive (agglomerative) clustering using an average-linkage method of clustering. It has some relationship to KITSCH, in that when the tree topology turns out the same, the branch lengths with UPGMA will turn out to be the same as with the $P = 0$ option of KITSCH.

The protdist outputs a distance matrix that is given as input to the neighbor program. We need to perform two iterations of neighbor program. The first will be for the Neighbor Joining method and the second will be for the UPGMA method. The first method gives rise to an unrooted tree, while the second method gives rise to a rooted tree. Once, the trees in both the cases have been obtained, the trees are given as input to consens to get the consensus tree of both the methods.

```

C:\Users\Dell\Desktop\Courses\Sem_VI\BT3040\Assignment10\phylip-3.698\exe\neighbor.exe

Neighbor-Joining/UPGMA method version 3.698

Settings for this run:
N      Neighbor-joining or UPGMA tree?  Neighbor-joining
O      Outgroup root?  No, use as outgroup species  1
L      Lower-triangular data matrix?  No
R      Upper-triangular data matrix?  No
S      Subreplicates?  No
J      Randomize input order of species?  Yes (random number seed =      5)
M      Analyze multiple data sets?  Yes, 10 sets
0      Terminal type (IBM PC, ANSI, none)?  IBM PC
1      Print out the data at start of run  No
2      Print indications of progress of run  Yes
3      Print out tree  Yes
4      Write out trees onto tree file?  Yes

Y to accept these or type the letter for one to change

```

FIGURE 17. neighbour input for nj algorithm

```

C:\Users\Dell\Desktop\Courses\Sem_VI\BT3040\Assignment10\phylip-3.698\exe\neighbor.exe

node 9 ( 0.55286) joins species 11 ( 1.06212) joins node 12 ( 0.12418)

Output written on file "work1-nj"

Tree written on file "work1-nj-tree"

Data set # 10:

Cycle 11: species 13 ( 0.14871) joins species 14 ( -0.03773)
Cycle 10: species 6 ( 0.39240) joins species 5 ( 0.26952)
Cycle 9: node 6 ( 0.54957) joins species 4 ( 0.42293)
Cycle 8: species 9 ( 0.27676) joins species 8 ( 0.43397)
Cycle 7: node 6 ( 0.68663) joins species 2 ( 1.11078)
Cycle 6: species 1 ( 0.84261) joins species 3 ( 1.15061)
Cycle 5: node 6 ( 0.23998) joins node 1 ( 0.12780)
Cycle 4: species 12 ( 0.89028) joins node 13 ( 1.98139)
Cycle 3: node 6 ( 0.25547) joins node 12 ( 0.17915)
Cycle 2: node 9 ( 0.57501) joins species 10 ( 1.25682)
Cycle 1: node 6 ( 0.23221) joins species 7 ( 0.81354)
last cycle:
node 6 ( 0.04302) joins node 9 ( 0.07173) joins species 11 ( 0.82709)

Output written on file "work1-nj"

Tree written on file "work1-nj-tree"

Done.

```

FIGURE 18. neighbour output for nj on window

```
C:\Users\Dell\Desktop\Courses\Sem_VI\BT3040\Assignment10\phylip-3.698\exe\consense.exe

Settings for this run:
C      Consensus type (MRe, strict, MR, ML): Majority rule (extended)
O      Outgroup root: No, use as outgroup species 1
R      Trees to be treated as Rooted: No
T      Terminal type (IBM PC, ANSI, none): IBM PC
1      Print out the sets of species: Yes
2      Print indications of progress of run: Yes
3      Print out tree: Yes
4      Write out trees onto tree file: Yes

Are these settings correct? (type Y or the letter for one to change)
y

consense.exe: the file "outtree" that you wanted to
  use as output tree file already exists.
  Do you want to Replace it, Append to it,
  write to a new File, or Quit?
  (please type R, A, F, or Q)
F
Please enter a new file name> work1-nj-constree

Consensus tree written to file "work1-nj-constree"

Output written to file "work1-nj-cons"

Done.

Press enter to quit.
```

FIGURE 19. consensus of neighbour output nj-tree

Above shown images were for constructing a consensus phylogenetic tree following NJ method. Below are the images for constructing a consensus phylogenetic tree following UPGMA method.

```
C:\Users\Dell\Desktop\Courses\Sem_VI\BT3040\Assignment10\phylip-3.698\exe\neighbor.exe

Neighbor-Joining/UPGMA method version 3.698

Settings for this run:
N      Neighbor-joining or UPGMA tree? UPGMA
L      Lower-triangular data matrix? No
R      Upper-triangular data matrix? No
S      Subreplicates? No
J      Randomize input order of species? Yes (random number seed = 5)
M      Analyze multiple data sets? Yes, 10 sets
0      Terminal type (IBM PC, ANSI, none)? IBM PC
1      Print out the data at start of run No
2      Print indications of progress of run Yes
3      Print out tree Yes
4      Write out trees onto tree file? Yes

Y to accept these or type the letter for one to change
```

FIGURE 20. neighbour input for upgma algorithm

```

C:\Users\Del\\Desktop\Courses\Sem_V\BT3040\Assignment10\phylip-3.698\exe\neighbor.exe
Cycle 1: node 11 ( 0.18455) joins node 13 ( 1.62938)
Output written on file "work1-upgma"
Tree written on file "work1-upgma-tree"
Data set # 10:
Cycle 13: species 13 ( 0.05549) joins species 14 ( 0.05549)
Cycle 12: species 6 ( 0.33096) joins species 5 ( 0.33096)
Cycle 11: species 9 ( 0.35537) joins species 8 ( 0.35537)
Cycle 10: node 6 ( 0.32077) joins species 4 ( 0.65173)
Cycle 9: species 7 ( 0.85594) joins species 11 ( 0.85594)
Cycle 8: node 7 ( 0.04576) joins node 9 ( 0.54634)
Cycle 7: species 1 ( 0.99661) joins species 3 ( 0.99661)
Cycle 6: node 7 ( 0.19075) joins species 10 ( 1.09246)
Cycle 5: node 7 ( 0.08568) joins species 12 ( 1.17813)
Cycle 4: node 6 ( 0.59719) joins species 2 ( 1.24892)
Cycle 3: node 7 ( 0.15828) joins node 1 ( 0.33981)
Cycle 2: node 6 ( 0.21798) joins node 7 ( 0.13048)
Cycle 1: node 6 ( 0.39549) joins node 13 ( 1.80690)
Output written on file "work1-upgma"
Tree written on file "work1-upgma-tree"
Done.

```

FIGURE 21. neighbour output for upgma algorithm

```

C:\Users\Del\\Desktop\Courses\Sem_V\BT3040\Assignment10\phylip-3.698\exe\consense.exe
Settings for this run:
C      Consensus type (MRe, strict, MR, Ml): Majority rule (extended)
O      Outgroup root: No, use as outgroup species 1
R      Trees to be treated as Rooted: No
T      Terminal type (IBM PC, ANSI, none): IBM PC
1      Print out the sets of species: Yes
2      Print indications of progress of run: Yes
3      Print out tree: Yes
4      Write out trees onto tree file: Yes
Are these settings correct? (type Y or the letter for one to change)
y
consense.exe: the file "outtree" that you wanted to
use as output tree file already exists.
Do you want to Replace it, Append to it,
write to a new File, or Quit?
(please type R, A, F, or Q)
F
Please enter a new file name> work1-upgma-constree
Consensus tree written to file "work1-upgma-constree"
Output written to file "work1-upgma-cons"
Done.

```

FIGURE 22. consensus of neighbour output upgma-tree

All the steps are now over. We have six trees to be shown.

- Maximum likelihood method tree and its consensus tree
- Neighbour Joining (NJ) method and its consensus tree
- UPGMA method and its consensus tree

TREES FOR SET 1 of sequences (tim.dat)

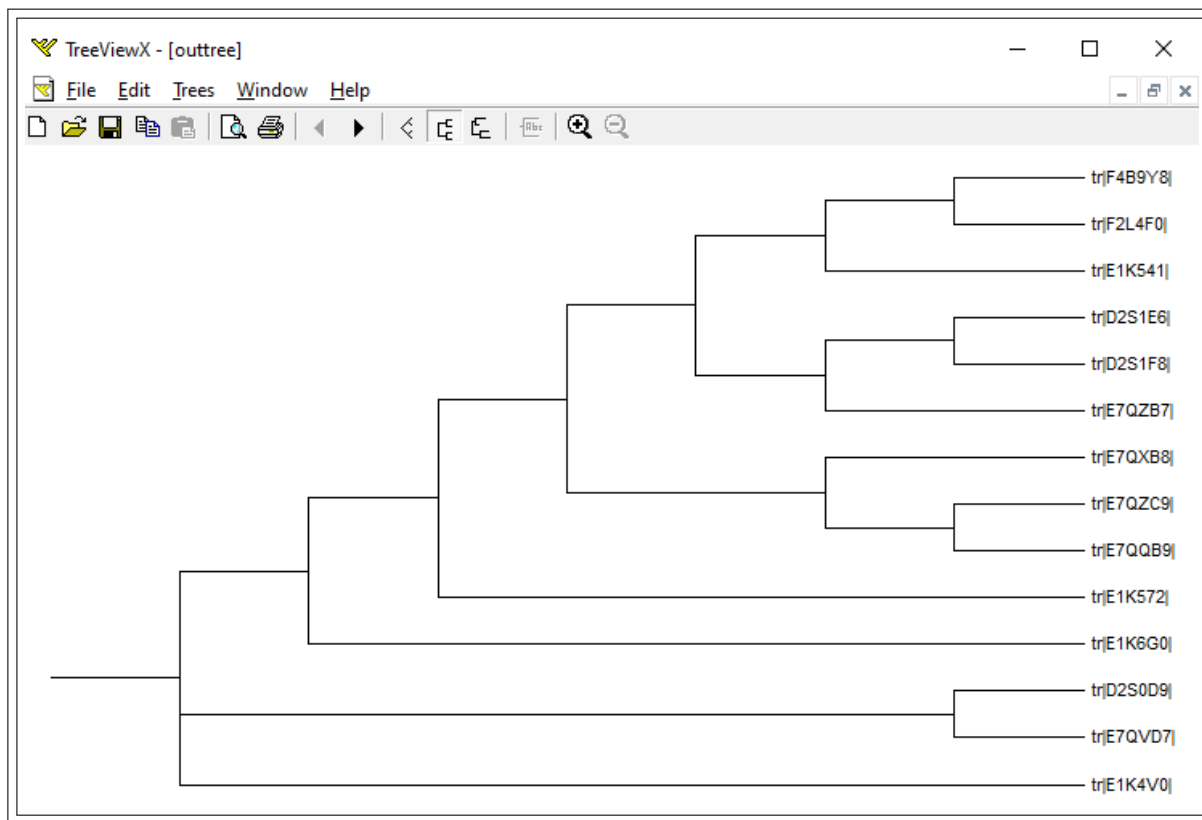


FIGURE 23. Tree obtained via proml program

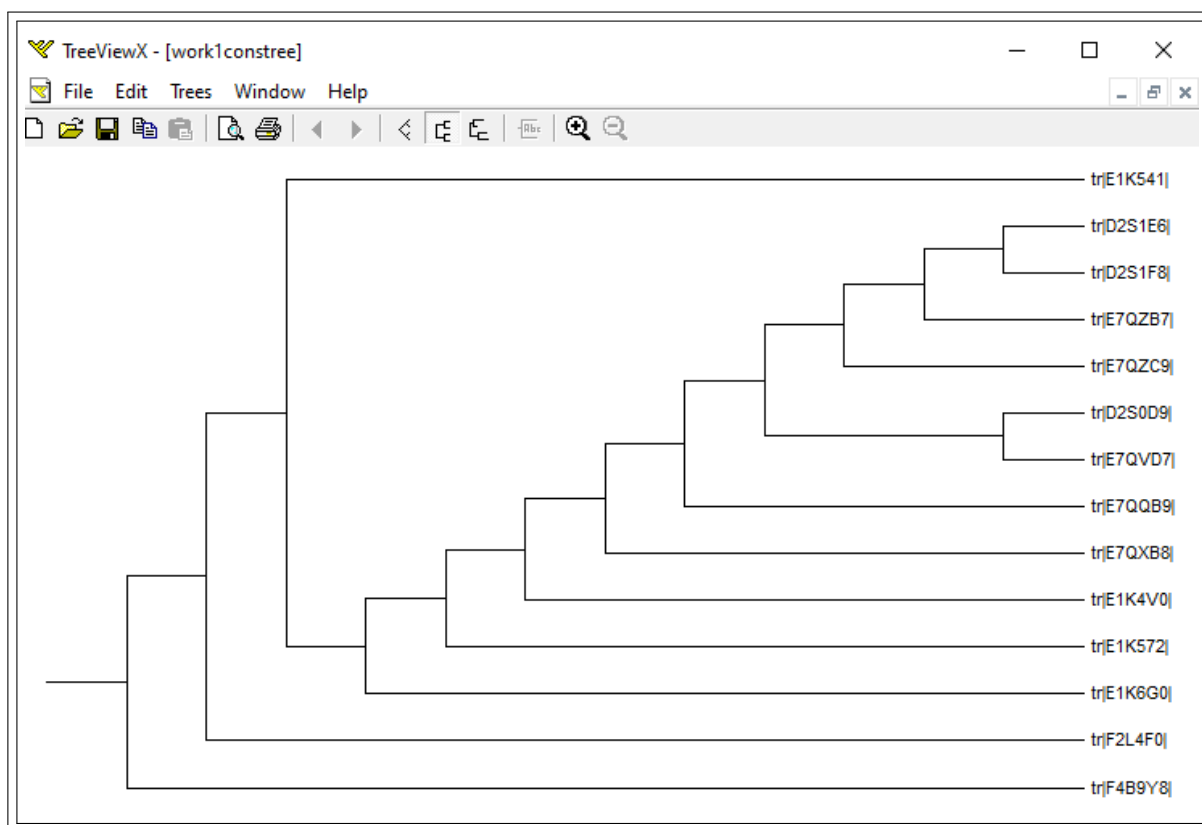


FIGURE 24. Tree obtained via consensus program on proml output

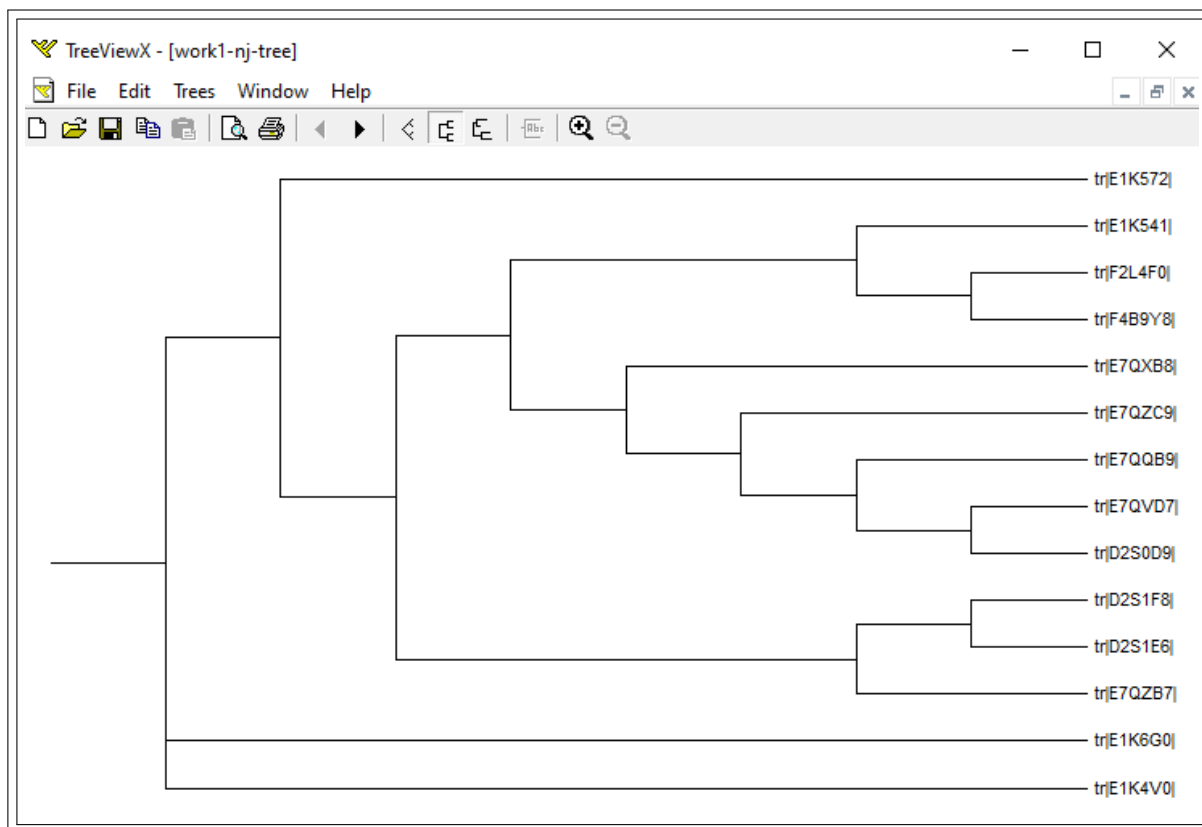


FIGURE 25. Tree obtained via neighbour program (nj method)

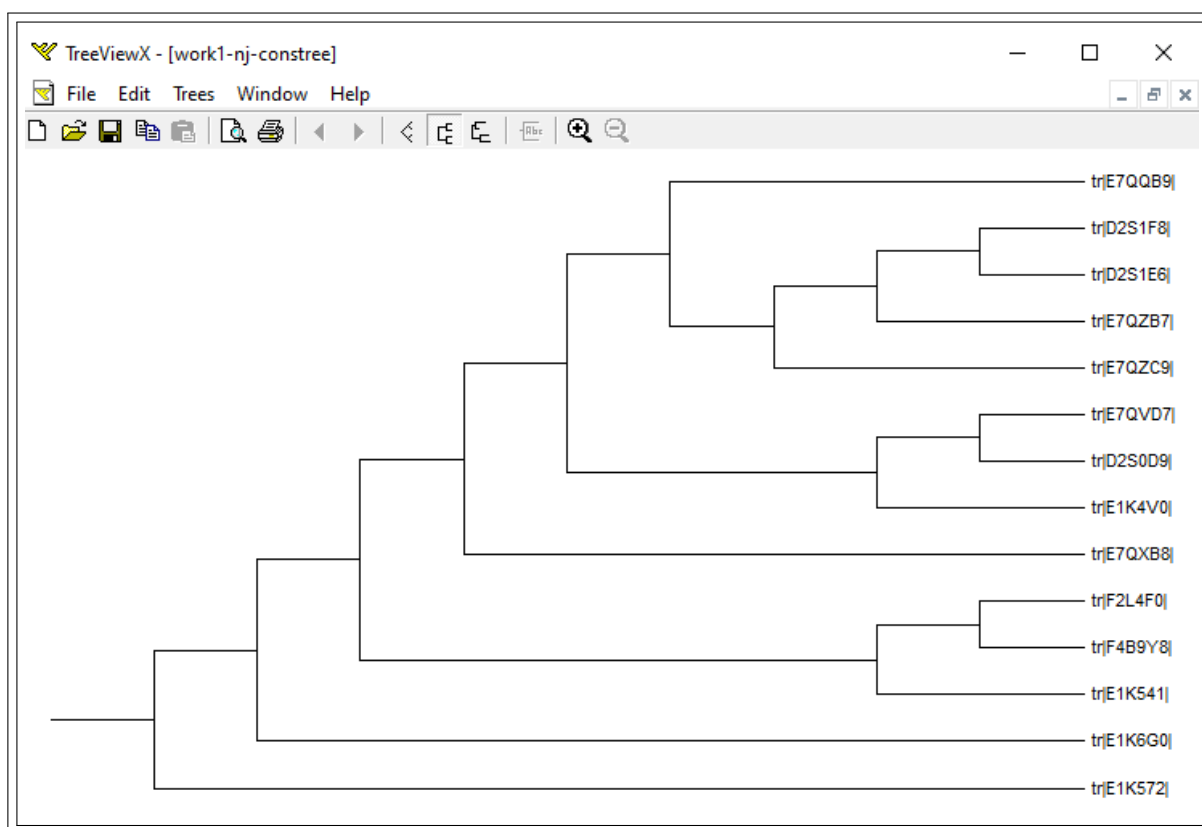


FIGURE 26. Tree obtained via consensus program on nj tree

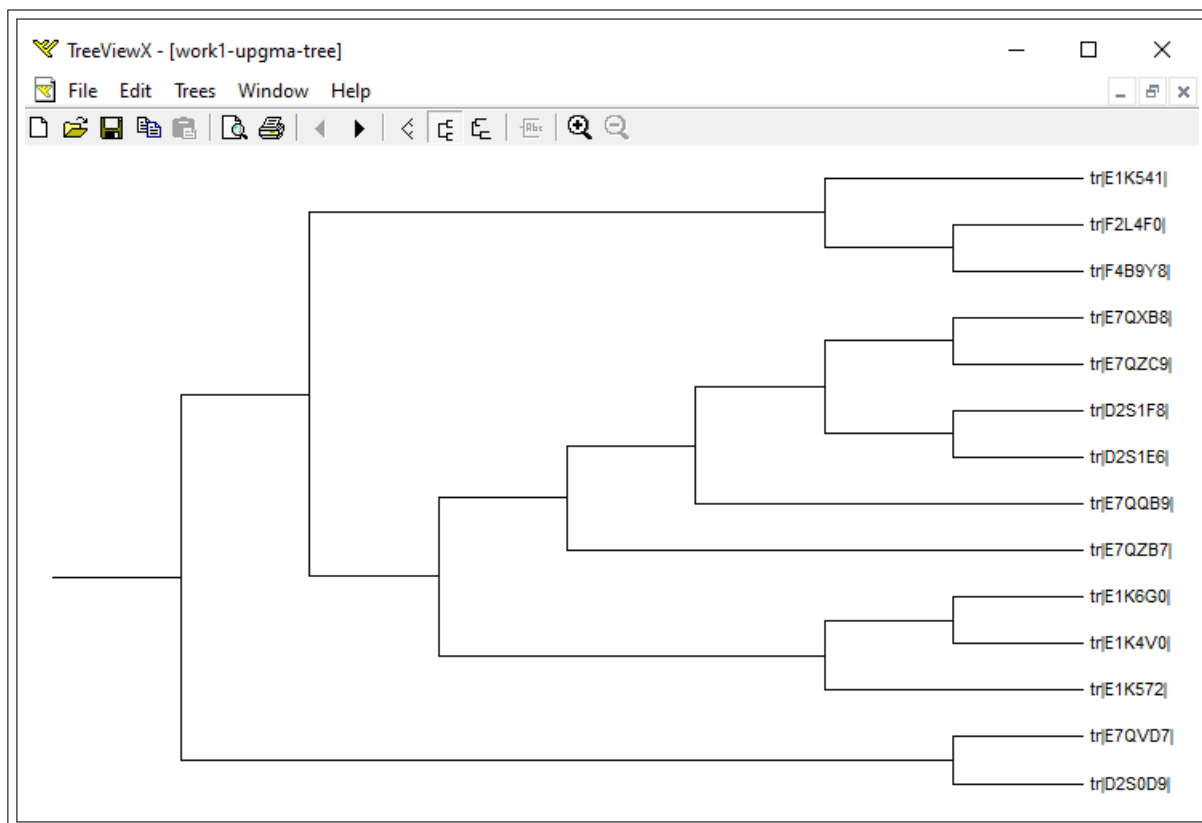


FIGURE 27. Tree obtained via neighbour program (upgma method)

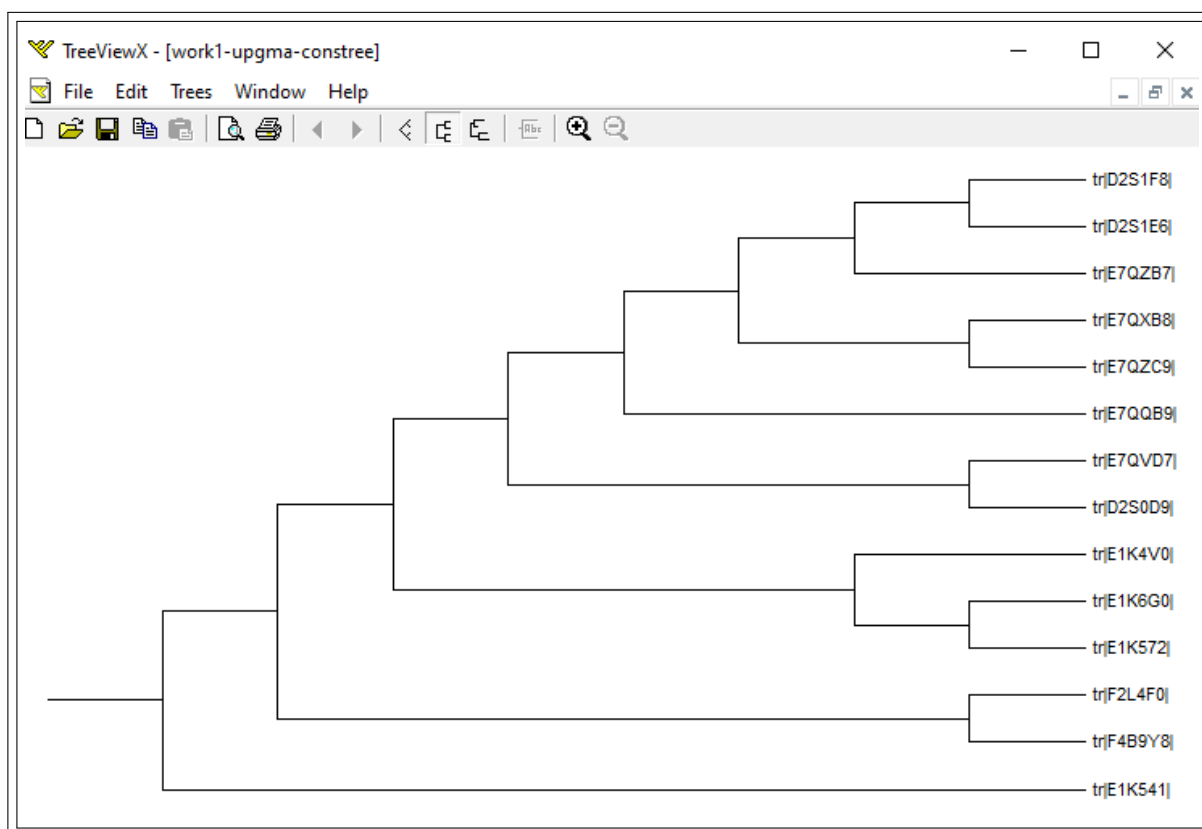


FIGURE 28. Tree obtained via consensus program on upgma tree

TREES FOR SET 2 of sequences (tim-hemo.dat)

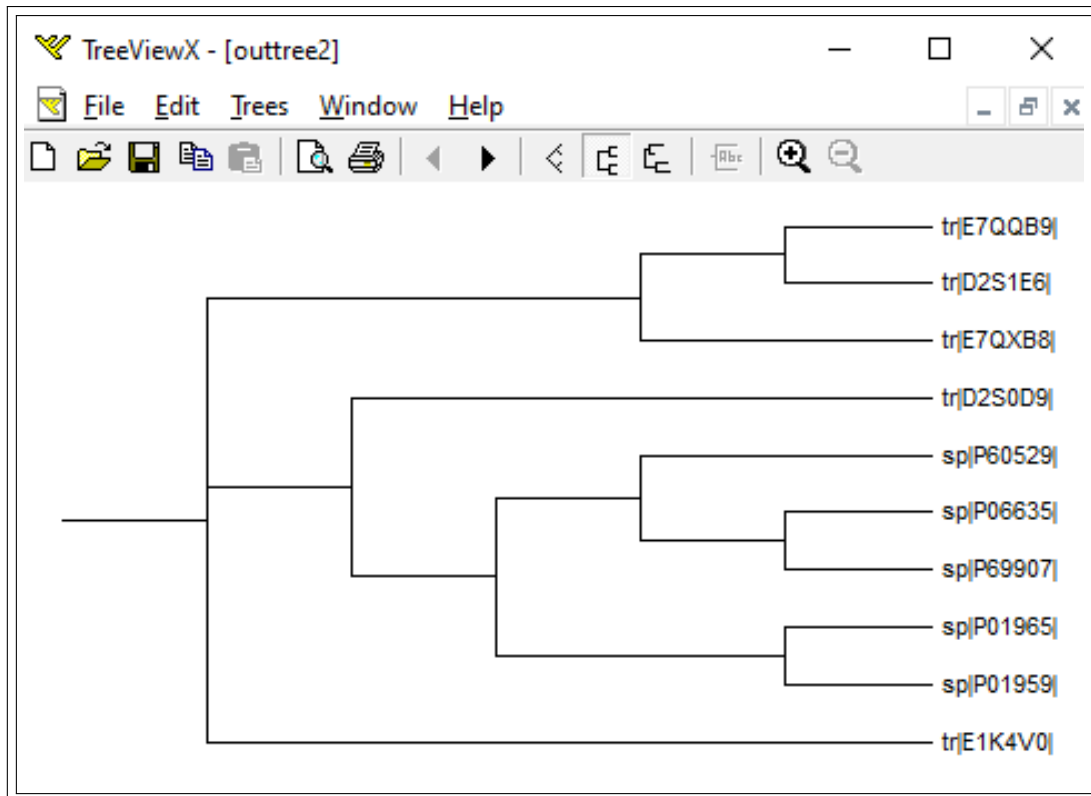


FIGURE 29. Tree obtained via proml program

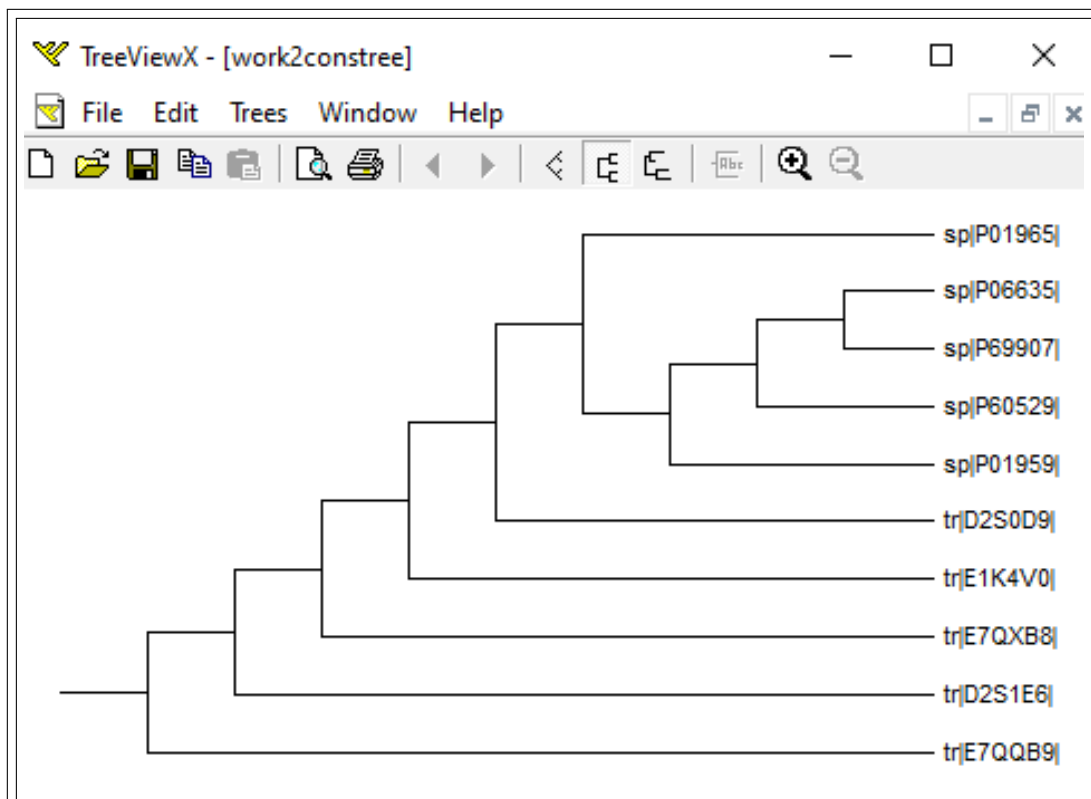


FIGURE 30. Tree obtained via consensus program on proml output

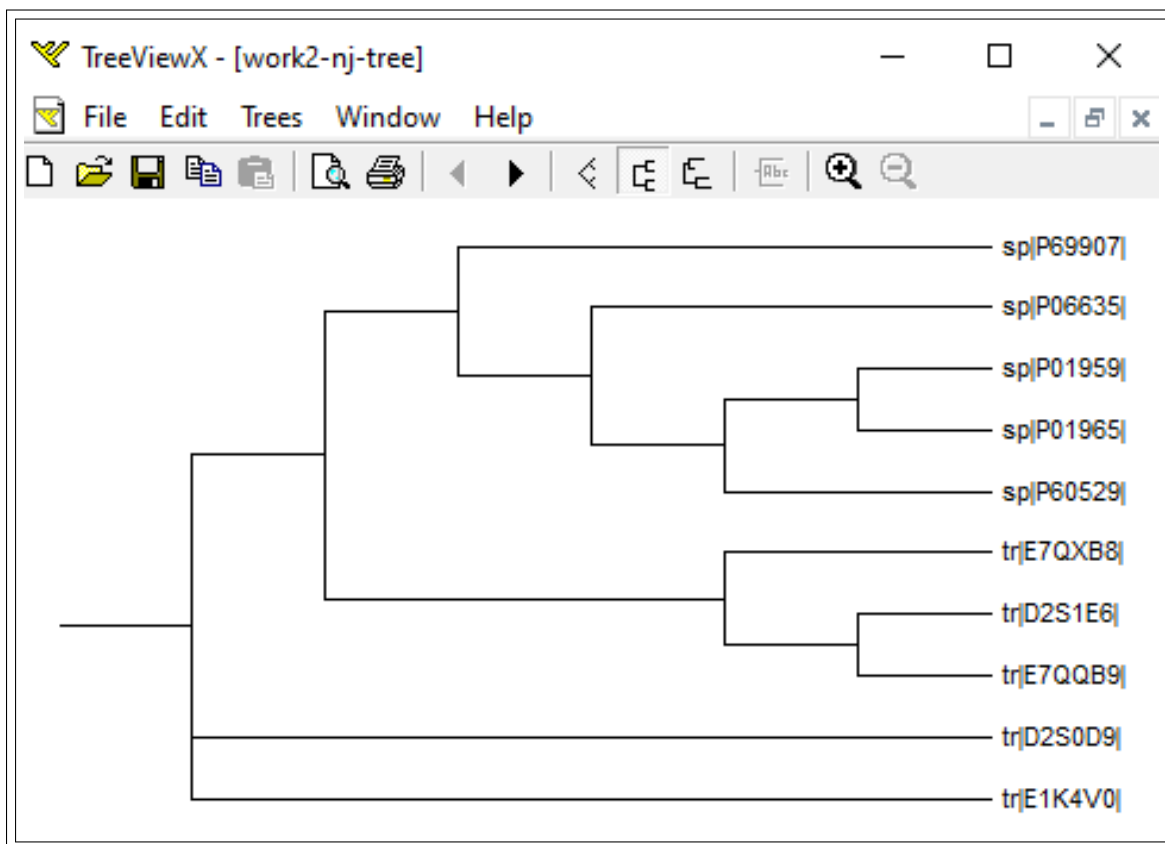


FIGURE 31. Tree obtained via neighbour program (nj method)

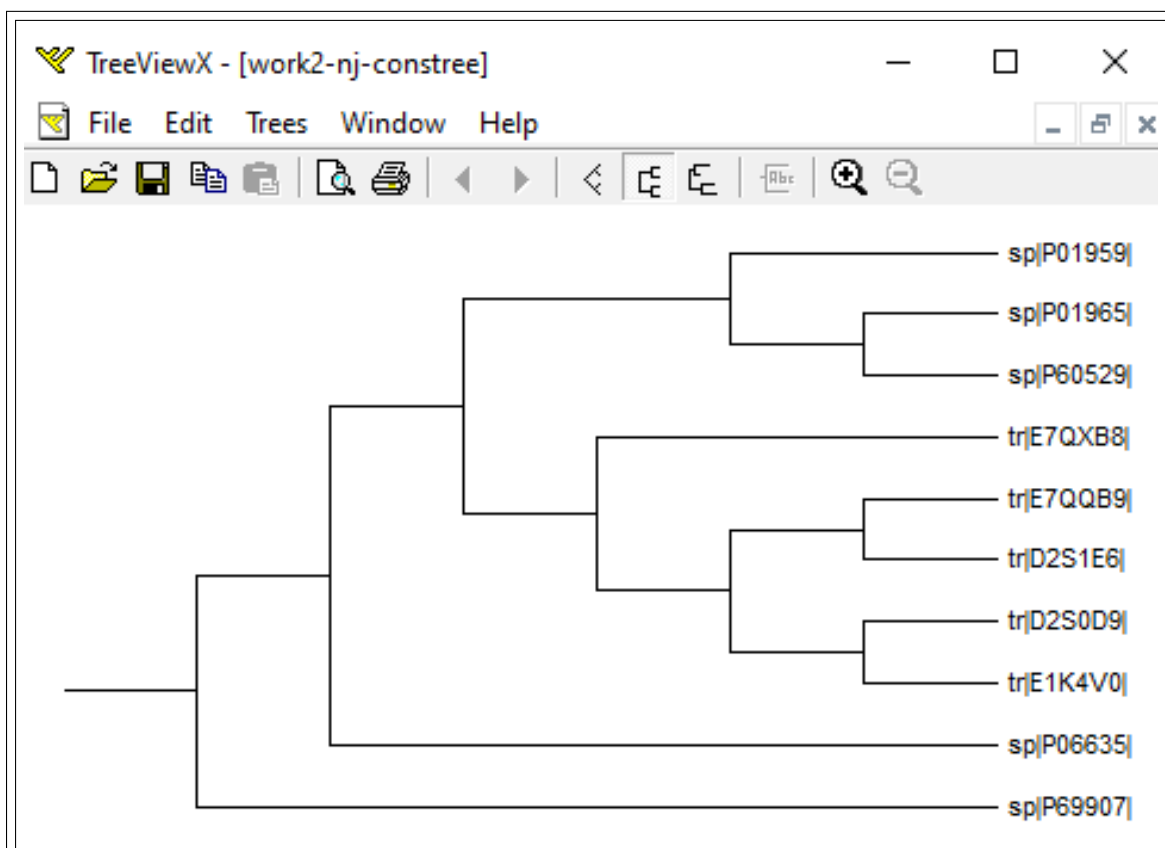


FIGURE 32. Tree obtained via consensus program on nj tree

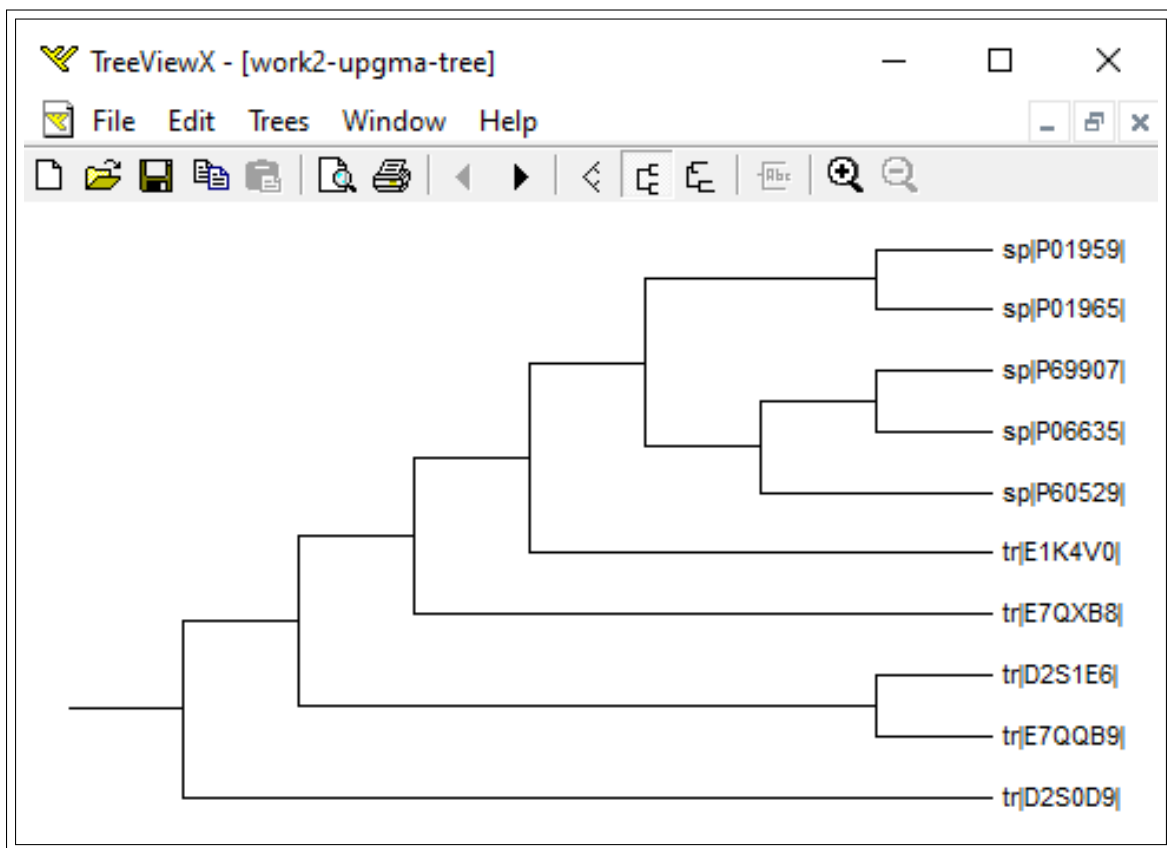


FIGURE 33. Tree obtained via neighbour program (upgma method)

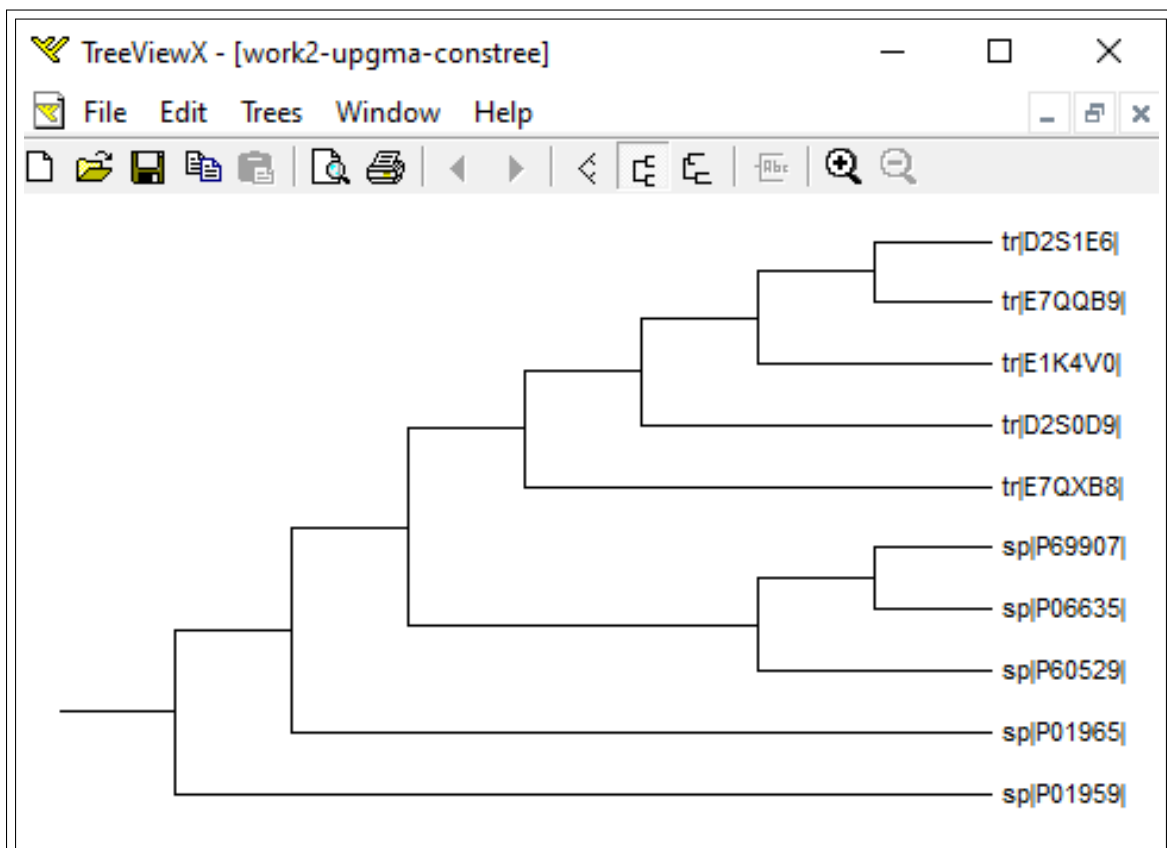


FIGURE 34. Tree obtained via consens program on upgma tree

Question 2. Obtain the weight matrix for the following sequences

Given 7 sequences of length 27 each

```
1 MVLSPADKTNVKGKVGAGHAGEYGAAAW
2 MKRLPADPPCVKGKVKAKAGDYGATTW
3 MALSAADKTNVKSQVGGHAGEYGAATS
4 MVLSAADKTNVKSAGGNAGEWAAAAW
5 MVLSAADKTNVKSQVLNAGEFGAAAW
6 ALLPIRTTYHKKCASGHIPEEKDLNNV
7 DEASSLKGGHHKKLEADALLIPLSASS
```

Solution. In order to find the weight matrix, we first find the frequency matrix of amino acids, which stores the frequency of occurrence of each amino acid at every position in the sequence. After that we use a formula to compute the weighted matrix from the sequence matrix.

$$(1) \quad w_{i,j} = \ln \frac{(n_{i,j} + p_i)/(N + 1)}{p_i}$$

$w_{i,j}$ = weight at (i,j)th position in the matrix

$n_{i,j}$ = number of times letter i has appeared at j position

N = total number of sequences (7 in this case)

p_i = priori probability of letter i (1/20 for all amino acids in this example)

The code to compute the weight matrix is given below:

```
1 import math
2 import seaborn as sns
3 import numpy as np
4 import matplotlib.pyplot as plt
5
6 seq = ["MVLSPADKTNVKGKVGAGHAGEYGAAAW",
7        "MKRLPADPPCVKGKVKAKAGDYGATTW",
8        "MALSAADKTNVKSQVGGHAGEYGAATS",
9        "MVLSAADKTNVKSAGGNAGEWAAAAW",
10       "MVLSAADKTNVKSQVLNAGEFGAAAW",
11       "ALLPIRTTYHKKCASGHIPEEKDLNNV",
12       "DEASSLKGGHHKKLEADALLIPLSASS"]
13 freq_matrix = [[0 for i in range(len(seq[0]))] for j in range(20)]
14 weight_matrix = [[0 for i in range(len(seq[0]))] for j in range(20)]
15 amino_acid_list =
16 {"A": 0, "R": 1, "N": 2, "D": 3, "C": 4, "Q": 5, "E": 6, "G": 7, "H": 8, "I": 9,
17  "L": 10, "K": 11, "M": 12, "F": 13, "P": 14, "S": 15, "T": 16, "W": 17, "Y": 18, "V": 19}
18 # Computing the frequency matrix
19 for i in range(len(seq[0])):
20     for j in range(len(seq)):
21         freq_matrix[amino_acid_list[seq[j][i]]][i] += 1
22 # Calculate the weight matrix
23 N = len(seq)
24 p = 0.05
25 for i in range(len(freq_matrix)):
26     for j in range(len(freq_matrix[0])):
27         numerator = (freq_matrix[i][j] + p) / (N + 1)
28         weight = math.log((numerator/p))
29         weight = round(weight, 2)
30         weight_matrix[i][j] = weight
```



```

31 # Frequency matrix plot
32 plt.figure(figsize=(13,10))
33 sns.heatmap(freq_matrix, annot=True, linewidth=0.5)
34 plt.xticks(ticks=np.arange(len(freq_matrix[0])) + 0.5, labels=[i for i in
    range(1,28)])
35 plt.yticks(ticks=np.arange(len(freq_matrix)) + 0.5, labels=["A","R","N","D","
    C","Q","E","G","H","I","L","K","M","F","P","S","T","W","Y","V"])
36 plt.show()
37
38 # Weight matrix plot
39 plt.figure(figsize=(13,10))
40 sns.heatmap(weight_matrix, annot=True, linewidth=0.5)
41 plt.xticks(ticks=np.arange(len(weight_matrix[0])) + 0.5, labels=[i for i in
    range(1,28)])
42 plt.yticks(ticks=np.arange(len(weight_matrix)) + 0.5, labels=["A","R","N","D"
    ,"C","Q","E","G","H","I","L","K","M","F","P","S","T","W","Y","V"])
43 plt.show()

```

LISTING 1. Code to compute the frequency matrix and weight matrix

The images of the frequency table and the weight matrix are given below:

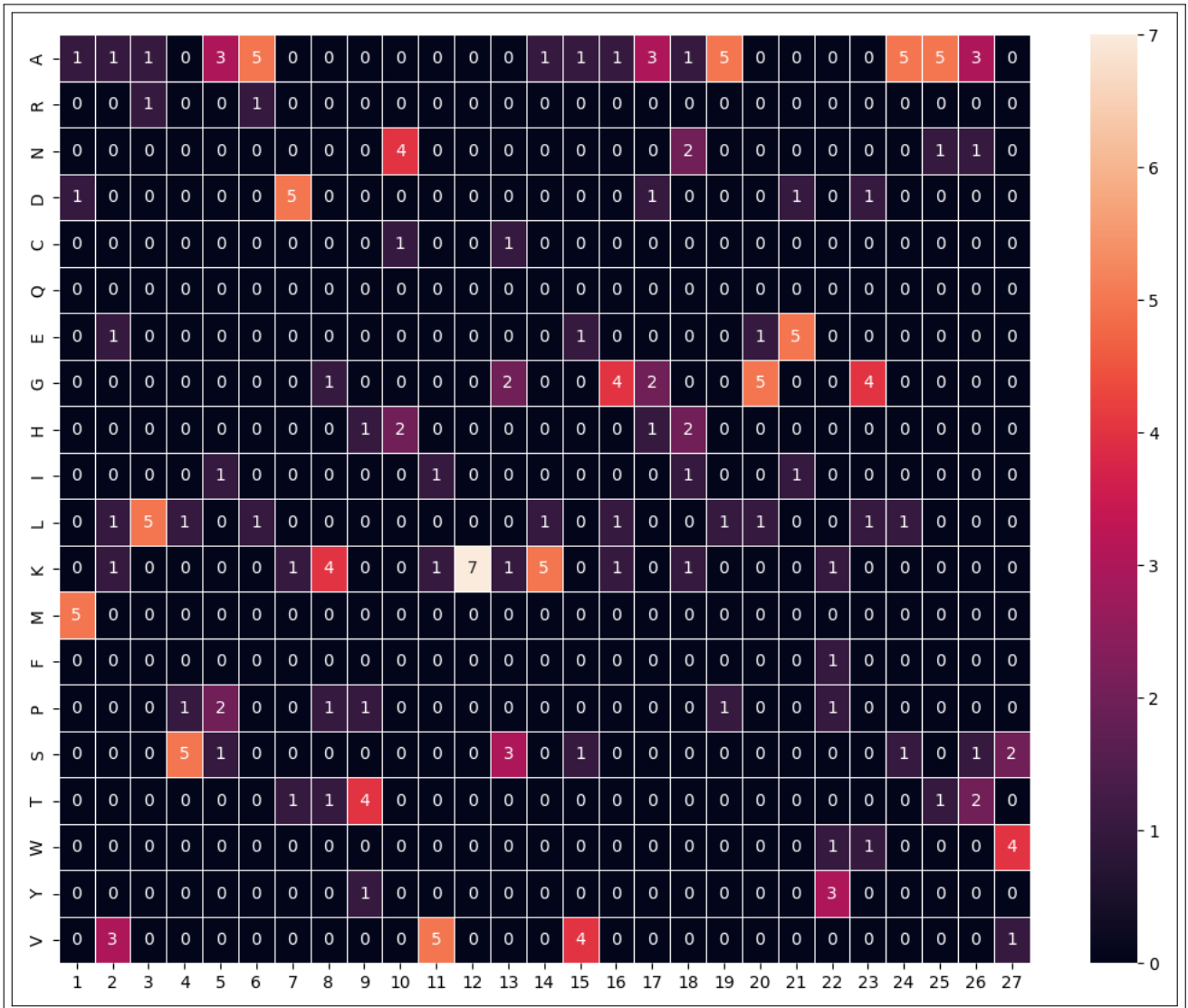


FIGURE 35. Frequency Matrix for given set of sequences

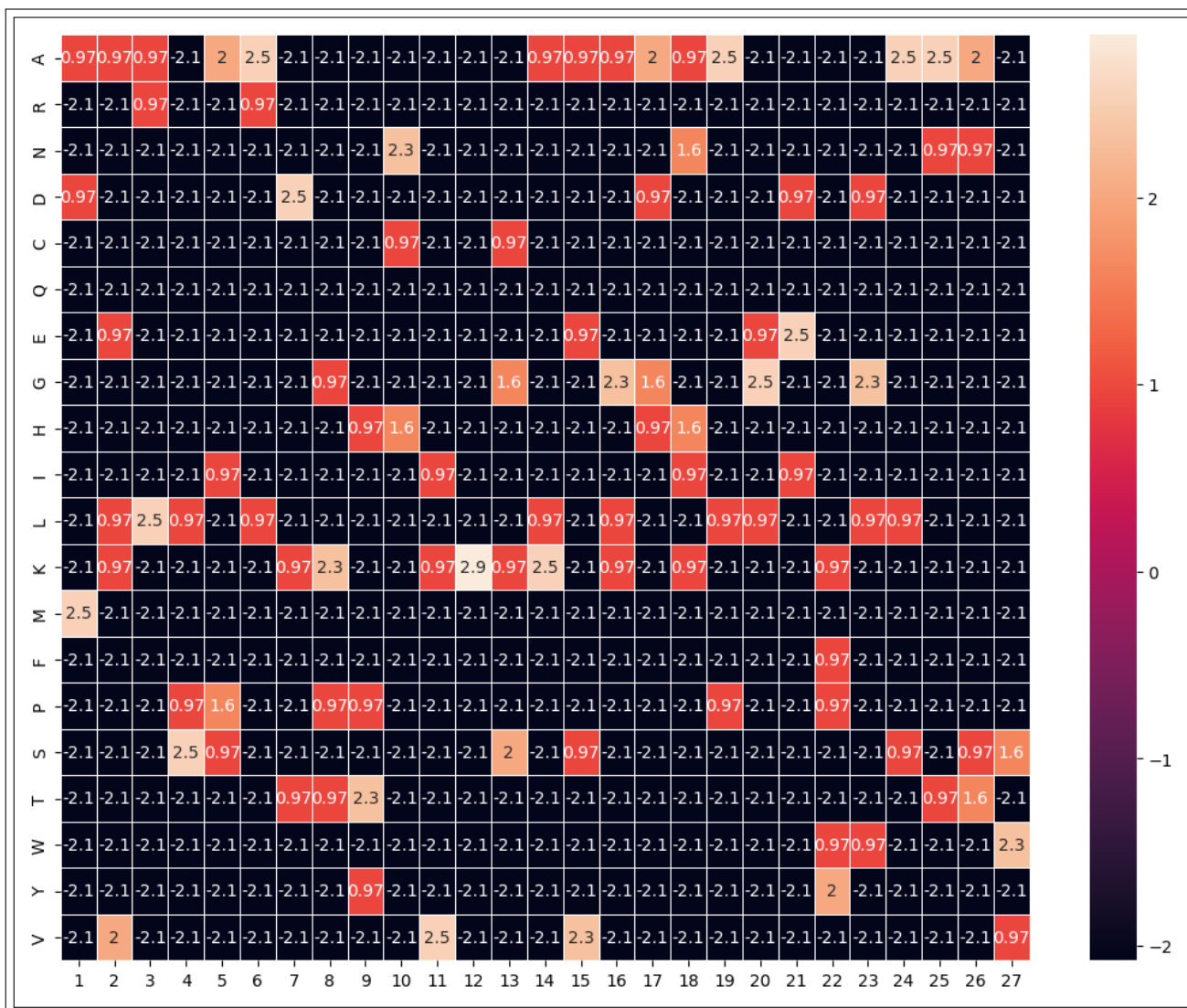


FIGURE 36. Weight Matrix for given set of sequences