Help Manual

Welcome! This program analyses protein sequences across different genus and species.

Overview

The program is divided into 5 parts:

1. Downloading sequences from NCBI
2. Modification of dataset
3. Multiple sequence alignment
4. Similarity and unique amino acid plots
5. Further analyses
    a. Motif scanning
    b. Pattern scanning
    c. Transmembrane segment plot
    d. Hydropathy plot

All example outputs in this manual are generated using *glucose-6-phosphatase* in *Aves*(birds).

Downloading sequences from NCBI

In this section, the program will download sequences from NCBI with search preferences set by you.

1. Specifying path

– First, please enter the full path to the folder you would like to save all outputs of this analysis. Check for spelling errors and letter cases. The program will ask for the path until the correct path is given. EG. /localdisk/home/s123456/folder1.

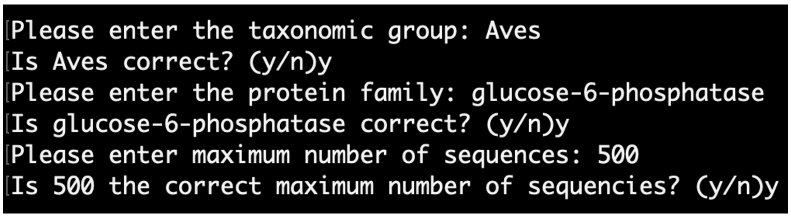
– NB. The program will remove any folders called ‘analysis’ in this specified space before starting. Please choose a space that is either empty or make sure that there is no important folder called ‘analysis’.

2. Specifying the taxonomic group, protein family and maximum number of sequences

– Please enter the taxonomic group, protein family and maximum number of sequences. The program allows you to double check these.

– NB. It is recommended to set the maximum number below 1,000 sequences. Larger than this may result in longer waiting time and potential crash.

Example:



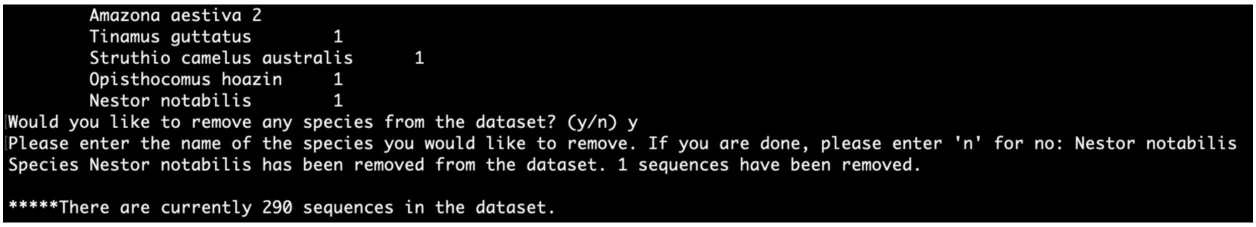
```
[Please enter the taxonomic group: Aves
[Is Aves correct? (y/n)y
[Please enter the protein family: glucose-6-phosphatase
[Is glucose-6-phosphatase correct? (y/n)y
[Please enter maximum number of sequences: 500
[Is 500 the correct maximum number of sequencies? (y/n)y
```

3. Specifying search parameters

- Please enter whether you would like to include: partial, low-quality, hypothetical, predicted and isoform of sequences. These are simple yes or no questions.
- NB. It is recommended to exclude uncertain or incomplete sequences for this analysis. Gaps introduced by these sequences during alignment may give rise to inaccurate analysis.
- If successful, the program prints the number of sequences in the current dataset. If the number of sequences is zero or it exceeds the maximum, the program will quit. Please restart and redefine the parameters.

Modification of dataset

In this section, you have the option to remove any species or genus from the dataset. Enter 'species' if you would like to display sequences per species or 'genus' for genus.



```
        Amazona aestiva 2
        Tinamus guttatus        1
        Struthio camelus australis      1
        Opisthocomus hoazin     1
        Nestor notabilis        1
Would you like to remove any species from the dataset? (y/n) y
Please enter the name of the species you would like to remove. If you are done, please enter 'n' for no: Nestor notabilis
Species Nestor notabilis has been removed from the dataset. 1 sequences have been removed.

*****There are currently 290 sequences in the dataset.
```

- Enter the name of species/genus you would like to remove. Once you are done editing the dataset, the program will ask whether you would like to continue.

Multiple sequence alignment

In this section, the program performs multiple sequence alignment on all sequences. Depending on the number of sequences, the program runs differently.

1. Sequence number > 250

- In this case, the program will first perform protein BLAST to identify 250 most similar sequences. For this purpose, please specify the reference genus/species you would like to blast against. The new dataset will contain 250 most similar sequences. Once this is complete, the program will run Clustal-Omega.

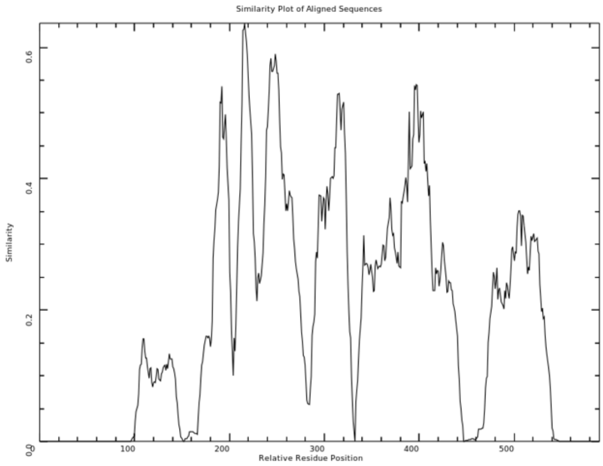
2. Sequence number < 250

- In this case, the program will run Clustal-Omega straight away.

Similarity and unique amino acid plots

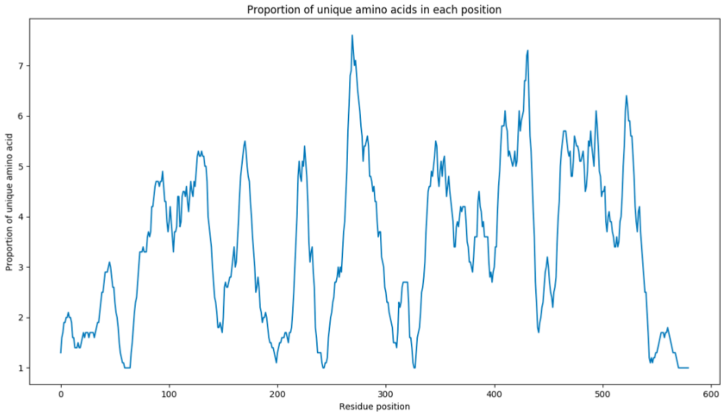
In this section, the program plots the similarity of aligned sequences and the proportion of unique amino acids in each position. For each plot, you have the option to save or remove the output.

1. Similarity of aligned sequences



- The plot shows the similarity of aligned sequences along its length.
- Peaks represent regions of high similarity in aligned sequences. This suggests that the residues in those regions are conserved among this set of sequences. In this example, residue 210 – 230 seem conserved.
- Conversely, troughs represent regions that are highly divergent. In this plot, amino acids in position 275 – 285 vary largely between sequences in this dataset.

2. Unique amino acids plot



- This plot shows the proportion of unique amino acids along its length.
- Opposite to the previous plot, peaks here represent regions that are highly divergent, with many different amino acids present.
- Conversely, troughs represent regions that are highly conserved with similar residues.

Further analyses

In this section, you have the option to select a subset of sequences from the dataset. The program will perform 4 different analyses on these sequences. There are three ways of selecting the subset of sequences:

1. Species

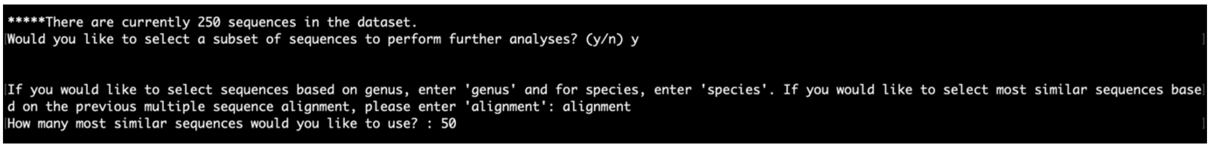
- Different species and sequences per species will be displayed. You can choose to remove any species from the dataset.

2. Genus

- Different genus and sequences per genus will be displayed. You can choose to remove any genus from the dataset.

3. Alignment

- You can choose the number of most similar sequences based on the previous multiple sequence alignment.



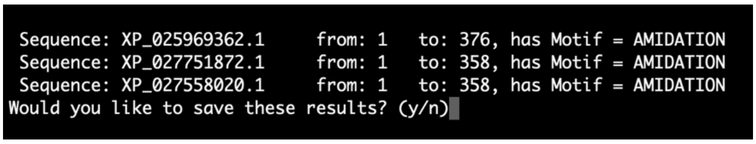
```
*****There are currently 250 sequences in the dataset.
Would you like to select a subset of sequences to perform further analyses? (y/n) y


If you would like to select sequences based on genus, enter 'genus' and for species, enter 'species'. If you would like to select most similar sequences base
d on the previous multiple sequence alignment, please enter 'alignment': alignment
How many most similar sequences would you like to use? : 50
```

For each analysis, you have the option to perform or skip to the next analysis. You also have the option to save or remove all outputs. All outputs will be saved in the folder you specified at the beginning.

Further analyses: 1. Motif scanning

In this analysis, the sequences are scanned for motifs from the PROSITE database. If motifs are found, the program outputs the sequence ID and the name of the motif.



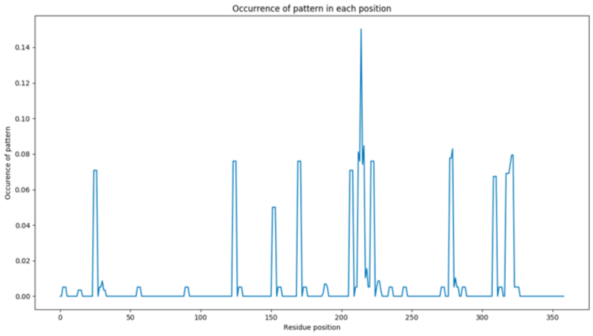
```
 Sequence: XP_025969362.1     from: 1   to: 376, has Motif = AMIDATION
 Sequence: XP_027751872.1     from: 1   to: 358, has Motif = AMIDATION
 Sequence: XP_027558020.1     from: 1   to: 358, has Motif = AMIDATION
Would you like to save these results? (y/n)
```

- In this example, sequence XP_025969362.1 with length of 376 has motif amidation.

Further analyses: 2. Pattern scanning

In this analysis, you can specify a search pattern and the program will scan sequences with this pattern. A plot is produced showing the occurrence of pattern along its length.

- This analysis is useful for identifying specific patterns such as binding sites or active sites of enzymes. It can be used to determine the conservation of this region of interest between species or genus.
- NB. The pattern must be written in PROSITE style. If you are unsure of this, please visit: https://prosite.expasy.org/scanprosite/scanprosite_doc.html . You can also access this manual through the program.

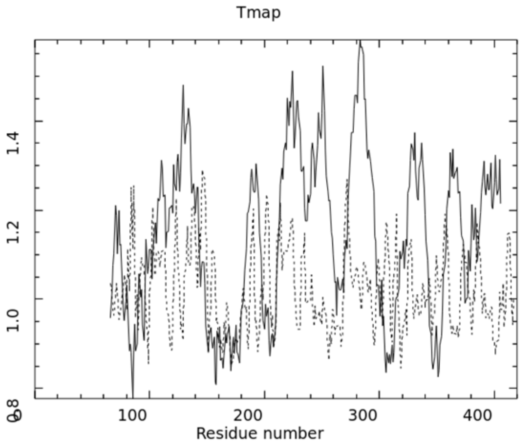


- High peaks indicate high occurrence of the pattern in that region across sequences in the dataset. It is highly likely that the pattern serves an important role in that region if the pattern is conserved in this group of sequences.
- In this example, the pattern occurred mostly between residue 200 – 225 in this group of sequences. This could be the active site of glucose-6-phosphatase.

Further analyses: 3. Transmembrane segment plot

In this analysis, transmembrane segments of aligned sequences are predicted and plotted.

– This tool is useful for proteins that span across the membrane. It can be used to determine the conservation of transmembrane segments across different species or genus.

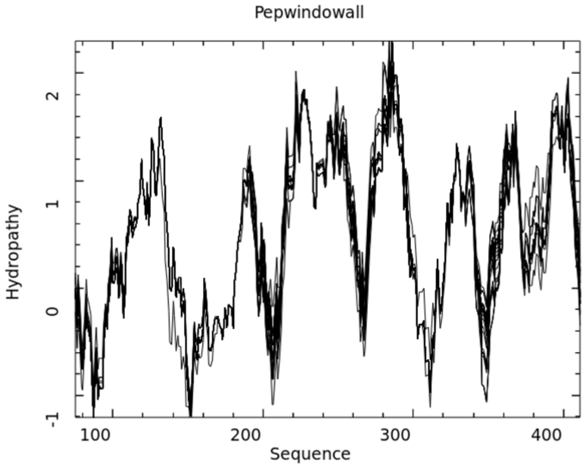


- Solid line represents propensity to form the middle region of the transmembrane protein. Hence, these residues are highly likely to be hydrophobic.
- Dotted line represents propensity to form the ends of the transmembrane protein. Therefore, these residues are highly likely to be hydrophilic.
- In this plot, no pattern is observed as glucose-6-phosphatase is not a membrane spanning protein.

Further analyses: 4. Hydropathy plot

In this analysis, the program plots hydropathy plot for aligned sequences.

– This tool is useful in visualising the hydrophobicity over the length of aligned sequences.
– Particularly useful in determining conservation of membrane proteins between species or genus.



- The peaks represent regions of high hydrophobicity, while troughs indicate regions of high hydrophilicity.
- In this example, most sequences in the dataset appear to have a hydrophobic core. This is because glucose-6-phosphatase is a globular protein with hydrophobic core surrounded by hydrophilic surfaces.

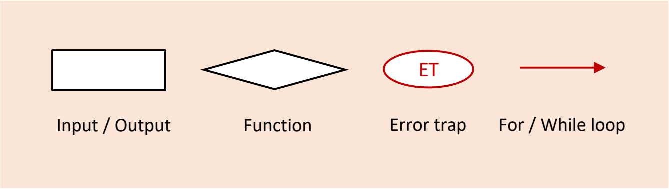
Maintenance manual

Overview

- The program is divided into 5 parts:

    1. Downloading sequences from NCBI
    2. Modification of dataset
    3. Multiple sequence alignment
    4. Similarity and unique amino acid plots
    5. Further analyses
        a. Motif scanning
        b. Pattern scanning
        c. Transmembrane segment plot
        d. Hydropathy plot

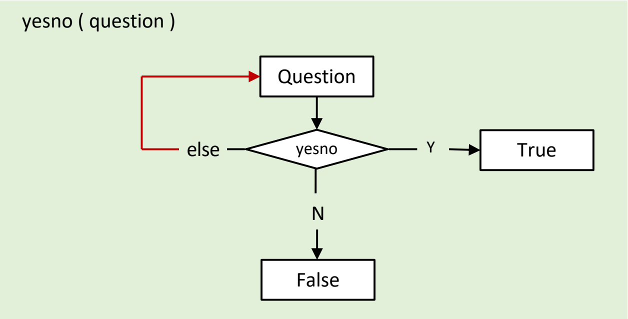
- The components of the program will be presented using flowcharts. The keys are:



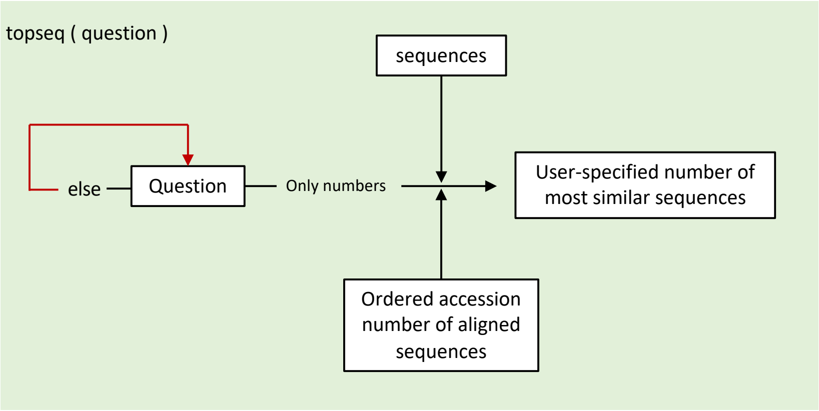
Functions

- This program uses 4 functions:

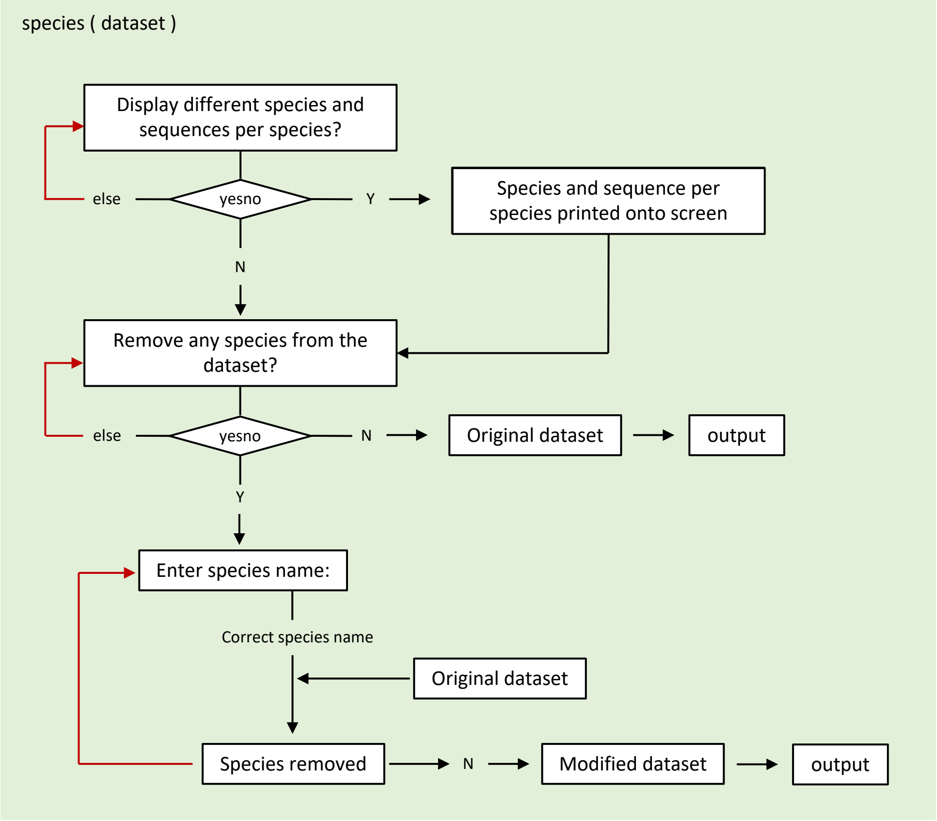
    1. yesno (question)
    2. topseq (question)
    3. species (dataset)
    4. genus (dataset)



- yesno( ) function asks the same question until the correct input is given. The while loop in the function is not broken until either ‘y’ or ‘n’ is given. If ‘y’, the function returns True and if ‘n’, the function returns False.


topseq (question)
sequences
else
Question
Only numbers
User-specified number of most similar sequences
Ordered accession number of aligned sequences

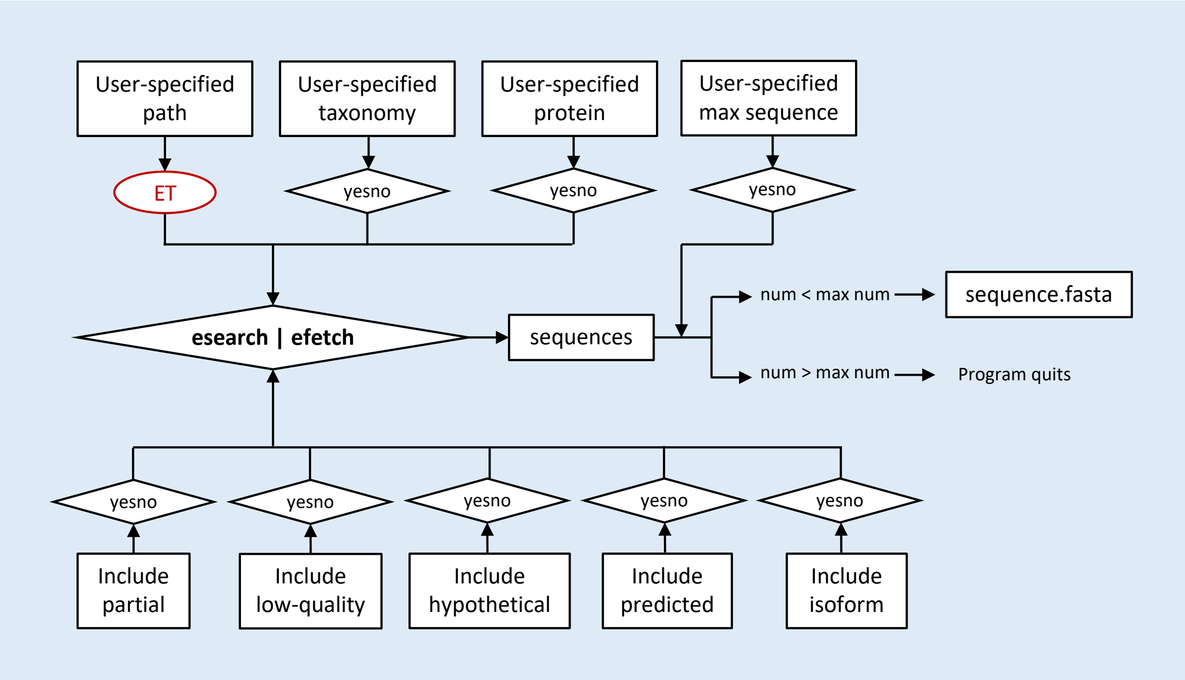
– topseq( ) function returns user-specified number of top most similar sequences. The while loop is not broken until a number is given.



– species ( ) function performs two roles :
  1. Displays different species and the number of sequences per species.
  2. Removes user-specified species from the dataset.
– The last while loop in the function iterates until the user inputs ‘n’.
– The function returns either modified dataset with species removed or the original dataset with no species removed.
– The fourth function, genus ( ) does the same as species ( ), but for genus.

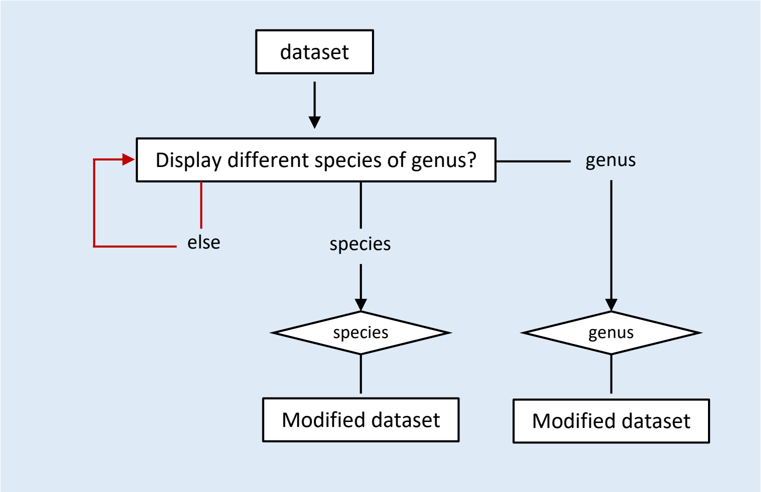
Program

1. Downloading sequences from NCBI



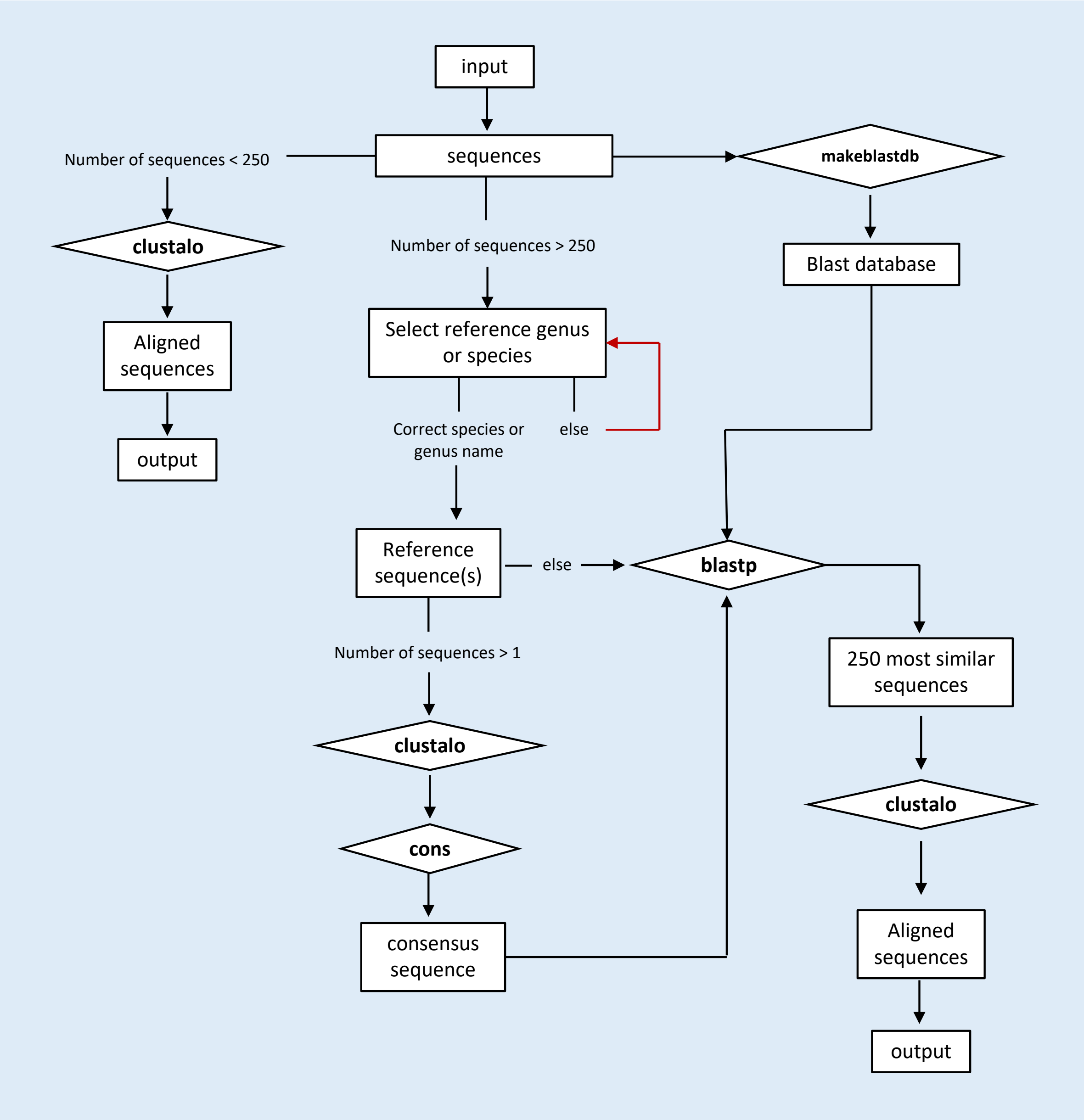
- User specifies the search parameters with series of yesno( ) function.
- If the number of downloaded sequences is greater than the user-specified maximum number, the program quits. Else, the program continues.
- Before continuing to the next step, the program asks user whether to continue or not.

2. Modification of dataset



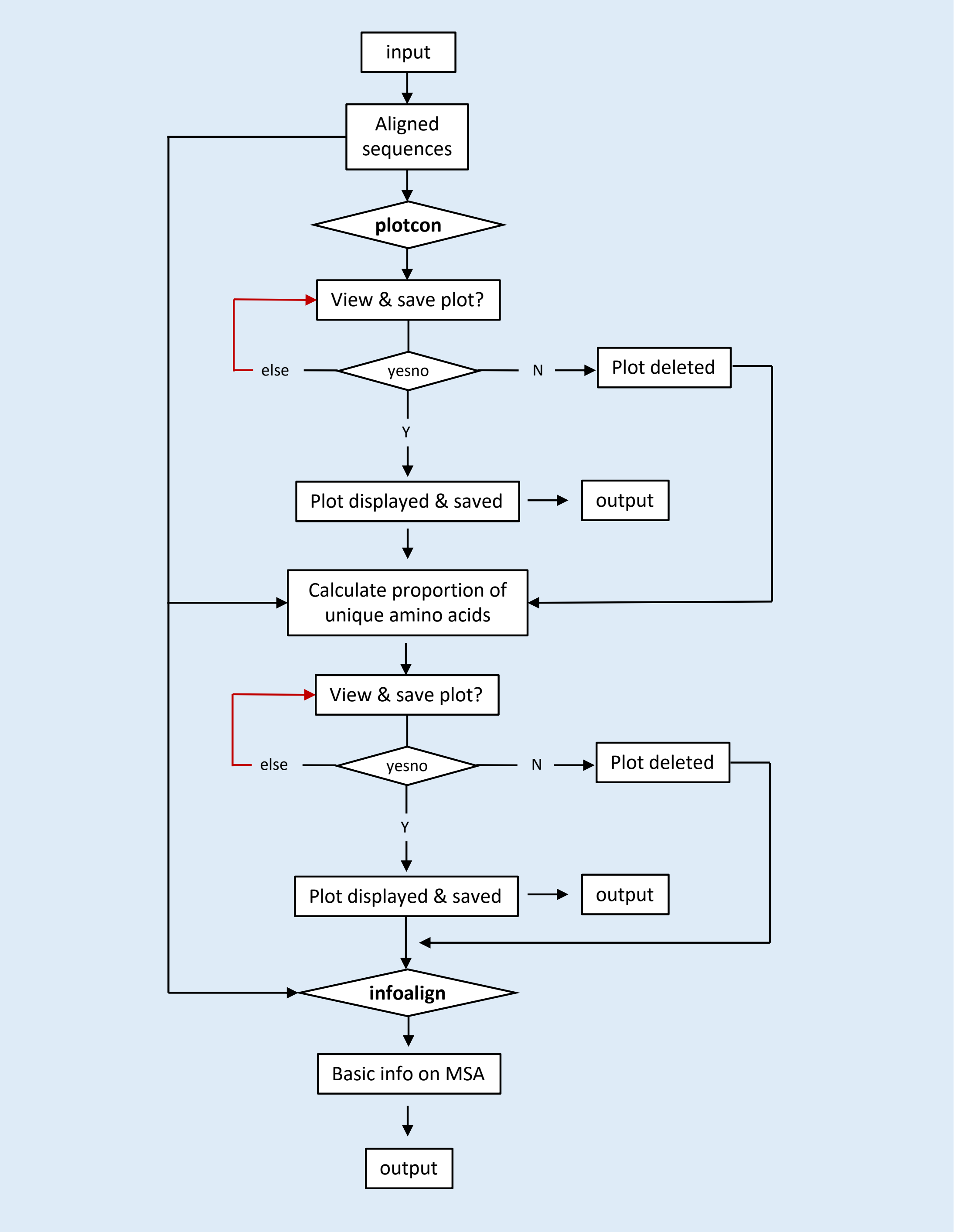
- The user chooses to edit the dataset based on species or genus.
- The species( ) and genus( ) functions display different species/genus and sequences per species/genus and removes user-specified species/genus from the dataset
- This part of the program outputs modified dataset with species/genus removed or the original dataset without any modifications.

3. Multiple sequence alignment



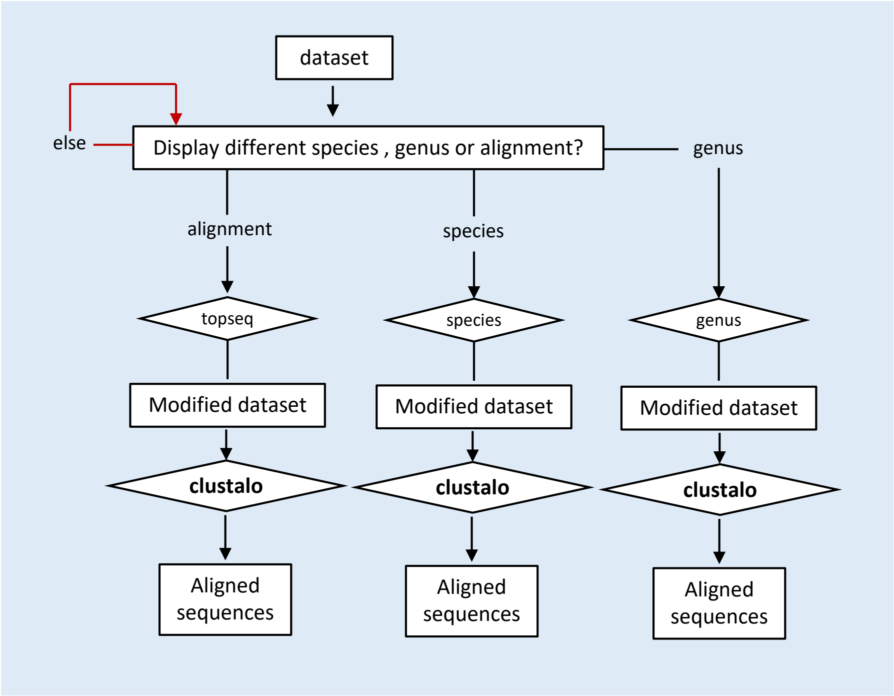
- Here, the output depends on the number of sequences in the input dataset.
- If the number is greater than 250, the user is asked to specify a genus or species to blast against. If there are more than one sequence for this reference genus/species, the program performs clustalo followed by cons to generate a consensus sequence. This consensus sequence is blasted against all sequences and top 250 most similar sequences are selected. These 250 sequences are aligned using clustalo.
- If the number is less than 250, these sequences are aligned using clustalo.
- Whether the initial number of sequences is greater or less than 250, this part of the program outputs a multiple sequence alignment result.

4. Similarity and unique amino acid plots



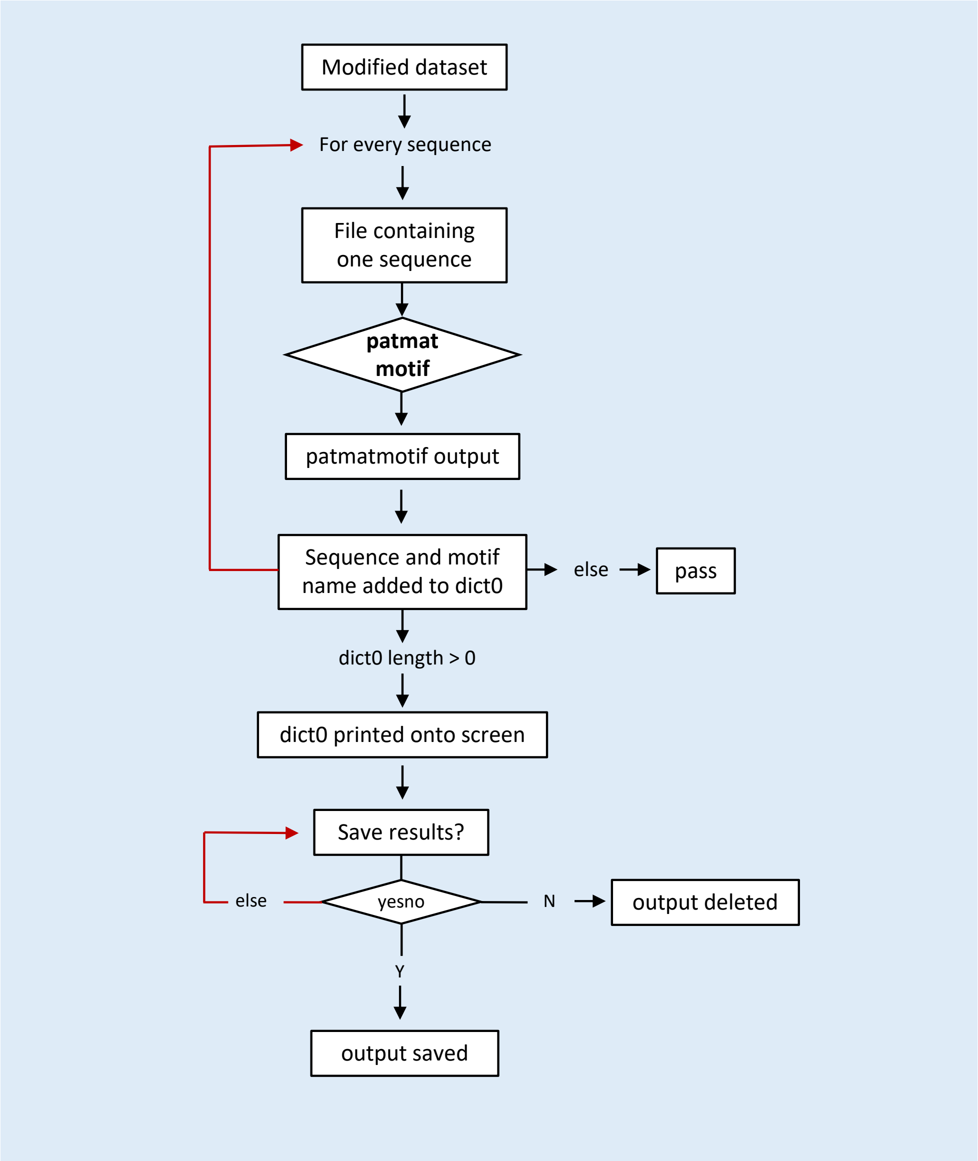
- This part of the program generates two plots: Similarity of aligned sequences and Proportion of unique amino acids in each position.
- Series of yesno( ) allows user to either save or remove the plot.
- The proportion of unique amino acids in each position is determined by scanning the sequences with a sliding window.
- Basic information on the multiple sequence alignment is also generated for the user.

5. Further analyses: Selecting subset of sequences



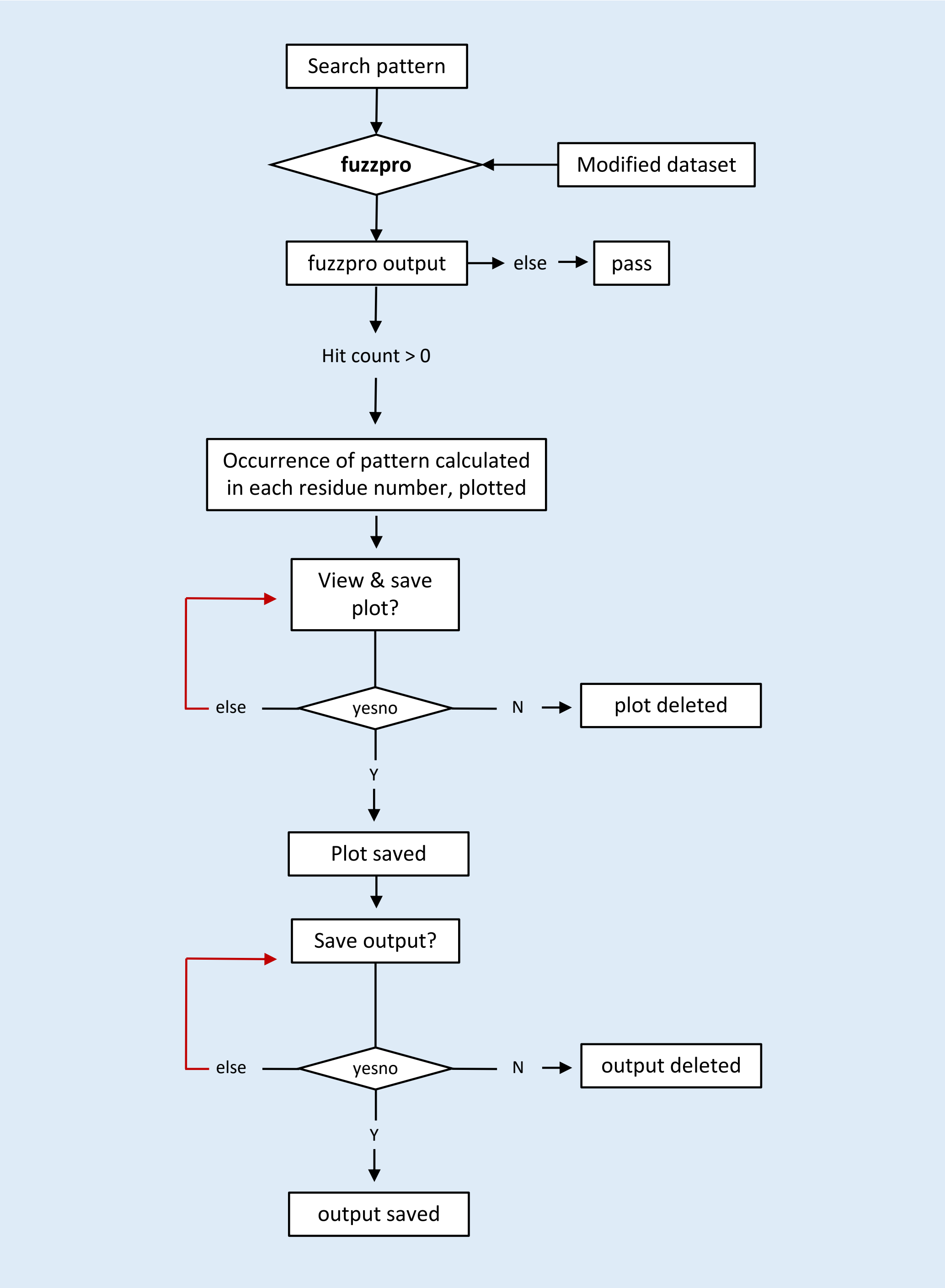
- Similar to the modification of dataset step, the user has the option to select a subset of sequences from the dataset to perform further analyses.
- Depending on the user's choice, the program calls topseq( ) or species( ) or genus( ) function, which outputs a modified dataset.
- This modified dataset containing subset of sequences is aligned using clustalo.
- This part of the program outputs multiple sequence alignment result of the subset of sequences.

Analysis 1: Motif scanning



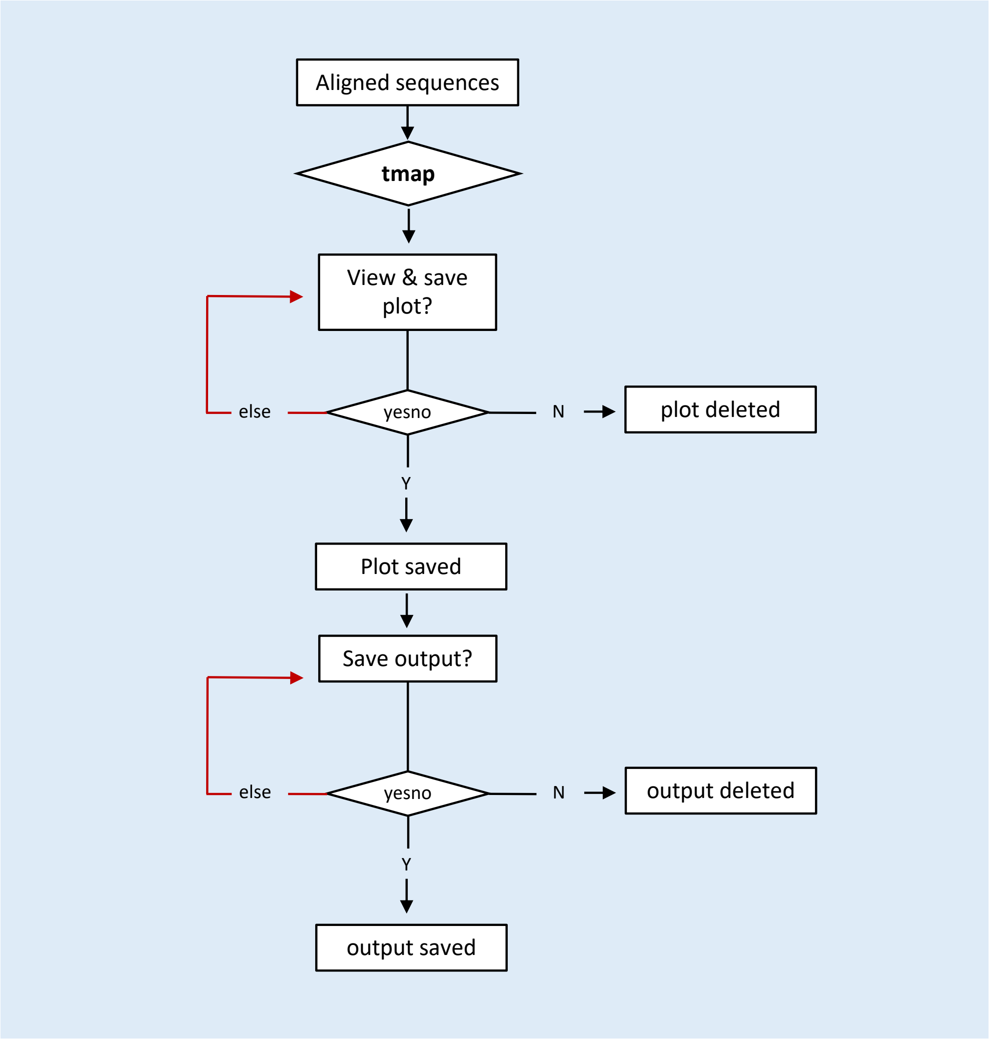
- The modified dataset generated in the previous step is scanned for any known motifs.
- Since patmatmotif function scans one protein sequence each time, the individual sequences are written to a file, scanned for motifs, and the output is saved in a dictionary. This for loop iterates for all sequences in the modified dataset.
- At the end of the loop, if the dictionary has elements, they are printed onto the user's screen.
- The user is given the option to save or remove the output.

Analysis 2: Pattern scanning



- The user inputs a search pattern and fuzzpro scans sequences in the modified dataset for this pattern.
- If there are matches, the residue positions of occurrences is counted and stored in a dictionary. The occurrence in each residue is divided by the total number of matches in the sequence. The program plots this proportion.
- Series of yesno( ) function gives user the option to save or remove the outputs of this analysis.

Analysis 3: Transmembrane segment plot



- The function tmap uses aligned sequences in the modified dataset to predict and plot transmembrane segments along its length.
- Series of yesno( ) function gives user the option to save or remove outputs of this analysis.

Analysis 4: Hydropathy plot

- Similar to the previous analysis, the function pepwindowall takes aligned sequences in the modified dataset and plots the degree of hydrophobicity along its length.
- Series of yesno( ) function gives user the option to save or remove outputs of this analysis.

End of program

- At the end, the user is given the option to remove all sequence files generated during the analysis.