Help Manual for Ordinary User

This Program can search, fetch, and check the protein sequences from NCBI, and then do some analysis as follows:

(1) Calculate and Plot the Similarity of Protein Sequences
(2) Blast Sequence Alignment
(3) Generate a Phylogenetic Tree
(4) Scan Protein Sequence of Interest with Motifs from the PROSITE Database

FYI: line of words in red are codes. Words following "#" will not be executed.

How to run the program?

ls code.py #First, check whether the program is in your current working directory.

```
bioinfmsc5:~/Assignment2$ ls code.py
code.py
```

Good! Find it!

python3 code.py #Run the program.

```
bioinfmsc5:~/Assignment2$ python3 code.py
--------------------------------------------------------------------------------------------------------------------------------
        This Program can search, fetch and check the protein sequences from NCBI, then do some analysis as follows:
                (1) Calculate and Plot the Similarity of Protein Sequences
                (2) Blast Sequence Alignment
                (3) Generate a Phylogenetic Tree
                (4) Scan Protein Sequence of Interest with Motifs from the PROSITE Database
--------------------------------------------------------------------------------------------------------------------------------
```

How to see the content of an outputfile with filename "protein.fasta"
less protein.fasta

Sequence Part:

```
What is the protein family(default value is glucose-6-phosphatase)
```

Now you can input the "protein family" and "subset of taxonomic tree" you want to use to fetch the data.
For example:
glucose-6-phosphatase # protein family
Aves # subset of taxonomic tree

```
What is the protein family(default value is glucose-6-phosphatase)
glucose-6-phosphatase
What is the subset of the taxonomic tree(default value is Aves)
Aves
```

Choose what kind of search to use.
Precise search: use specific protein name like glucose-6-phosphatase. If you input a larger family like kinases, maybe find no sequence in this search.
Open search: use protein family as search term, you may find thousands of sequences in the search results.

```
Please choose whether precise search or open search for protein do you want?
precise, tpye:p          open, type:o
```

p

```
Please choose whether precise search or open search for protein do you want?
precise, tpye:p          open, type:o
p
Precise search:
Your protein family is glucose-6-phosphatase and your taxonomic subset is Aves
```

Choose whether partial sequences are accepted.

```
CHECK-1-Do you want to get all available sequences or non-partial sequences(recommended)?
All data,type: all      Non-partial: just press 'Enter'
```

Recommend you choose "search only for non-partial sequence", which is more useful in sequence analysis in the further steps.

```
Precise non-partial search:
<ENTREZ_DIRECT>
  <Db>protein</Db>
  <WebEnv>MCID_61a3d4460fcffd16b7623613</WebEnv>
  <QueryKey>1</QueryKey>
  <Count>69</Count>
  <Step>1</Step>
</ENTREZ_DIRECT>
glucose-6-phosphatase[Protein Name] AND Aves[Organism] NOT PARTIAL

CHECK-2-Sequence Number:

In above search results,<Count> means sequence number you got.
If you got more than 1000 sequences, you will wait for a long time for sequence fetching.
If you want to continue,type:continue
If you want to reset the search information, type: reset
```

Check the sequence number:

- If you are not satisfied with the **sequence number** in the results, type:
  reset

  ```
  In above search results,<Count> means sequence number you got.
  If you got more than 1000 sequences, you will wait for a long time for sequence fetching.
  If you want to continue,type:continue
  If you want to reset the search information, type: reset
  reset
  Reset the search information...

  What is the protein family(default value is glucose-6-phosphatase)
  ```

  now, you can reset the input and do all we have done again.

- In this example, we got 69 sequences, no more than 1000, good! So, we can continue:
  continue

Fetch the search sequences and calculate species number in the starting dataset:

```
Please wait, fetching the sequences you want...

CHECK-3-Species Number:
There are 69 sequences of 63 species in your corrent dataset

Do you want to continue with the current dataset?
Type:continue    Type: reset
```

- If you are not satisfied with the **species number** in the results, type:
  reset

  ```
  Reset the search information...

  What is the protein family(default value is glucose-6-phosphatase)
  ```

- In this example, we have 63 species in the starting data set, the diversity is also good. So, we can continue.

continue

Give a limit to the sequence length:

```
CHECK-3-Species Number:
There are 69 sequences of 63 species in your corrent dataset

Do you want to continue with the current dataset?
Type:continue    Type: reset
continue
Good job!

This programm will automatically choose the longest sequence of each species in your starting data base for further analysis.

The median number of sequence length is  426.0
You can input a length limit to remove those short sequences from your data set.
 Or use the defaults.
```

- You can **input a number**, so that sequences with length longer than your input number will be chosen to do following analysis.
- Or just **press "Enter"**, the program will use the default number that is median number minus 30.

```
The median number of sequence length is  426.0
You can input a length limit to remove those short sequences from your data set.
 Or use the defaults[median-30].

Your check value is 396

The length of sequences of choice are as follows.

[439, 414, 430, 428, 423, 481, 435, 420, 427, 422, 496, 415, 436, 435, 416, 434, 423, 427, 492, 435, 429, 423, 426, 424, 4
72, 430, 419, 492, 419, 430, 419, 498, 419, 426, 419, 446, 431, 435, 483, 415, 417, 495, 423, 420, 435, 418, 422, 420, 433
, 435, 423, 436, 422, 435, 430, 474, 417, 402, 415]
There are 59 output sequences in 'pr_seq_choose.fasta'

Well done! Let's start analysis!
```

Analysis Parts:

In this part, you will be asked whether or not do a certain analysis?
if you don't want to do any of the analysis, type:

No

```
OK, sikp this step.
```

Analysis-1- Calculate and Plot the Similarity of Protein Sequences

```
Do you want to get the similarity and level of conservation of the protein sequences in the dataset?
Yes or No:
```

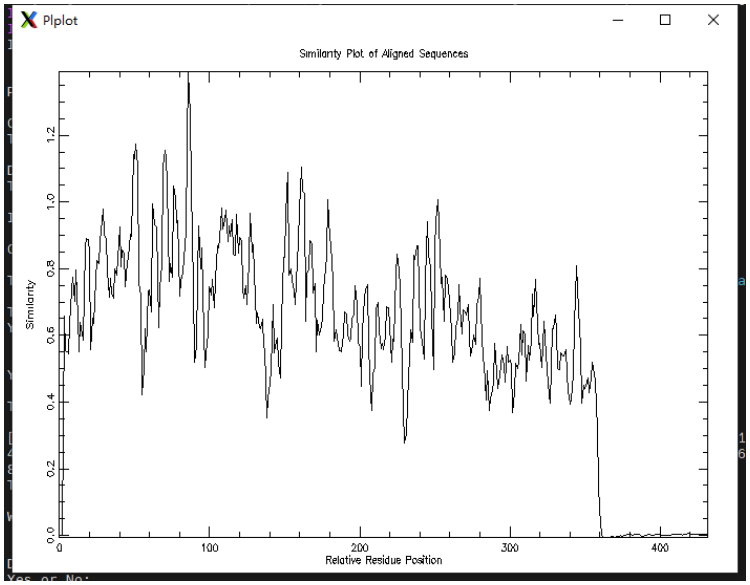
Yes

```
Analysis-1-Calculate and Plot the Similarity of Protein Sequences:
Plot conservation of a sequence alignment
Created plotcon1.dat
Plot conservation of a sequence alignment
Window size [4]:
```

Input a window size (default is 4), Eg.:

3

A window will be popped up showing "Similarity Plot of Aligned Sequences".
When you close the windows of Plplot, the program can continue.



In Analysis-2-Blast Aequence Alignment:

The output file of Blast within the sequences in our dataset is a tab separated table 'pr_seq_choose_blast.tsv', including columns of query acc.ver, subject acc.ver, % identity, alignment length, mismatches, gap opens, q. start, q. end, s. start, s. end, evalue, and bit score. The output file will go to screen automatically.

In Analysis-3-Generate a Phylogenetic Tree with Edialign:

The Sequence Length, Similarity, Phylogenetic Tree are in file 'pr_seq_choose.edialign'. The phylogenetic tree is constructed using UPGMA based on DIALIGN fragment weight scores.
The alignment sequence is in file 'edialign_pr_seq.fasta'.
File 'pr_seq_choose.edialign' will go to screen automatically.

In Analysis-4-Scan Protein Sequence of Interest with Motifs from the PROSITE Database:

```
You want to use all sequences in our dataset or choose some of the sequences of your interest(Default is partial)?
Type:all        Type:partial
```

- if you choose all sequences for analysis, type:

  all

```
Analysis-4-Scan Protein Sequence of Interest with Motifs from the PROSITE Database:
```

```
Output results for all the protein you scanned are in file 'pr_seq.patmatmotifs'
Name of motif(s) associated with the protein you interested in:
All kinds of motifs found are in the following list:
 ['Motif = AMIDATION']
```

  The output file "pr_seq.patmatmotifs" will go the screen automatically.
  Associated motifs found in the PROSITE Database will be show on the screen.

- if you want to use some sequences in the dataset for analysis, type:

  partial

```
Analysis-4-Scan Protein Sequence of Interest with Motifs from the PROSITE Database:
>KQK79067.1_glucose-6-phosphatase_[Amazona_aestiva]
>XP_030321679.1_glucose-6-phosphatase_[Calypte_anna]
>KAF4787459.1_Glucose-6-phosphatase_[Turdus_rufiventris]
>XP_014740911.1_PREDICTED:_glucose-6-phosphatase_[Sturnus_vulgaris]
>XP_021233644.1_glucose-6-phosphatase_[Numida_meleagris]
```

```
From all above sequence information, you can find the accession number.

Please choose one sequence that you're interested in and input the accession number.The default is XP_031456423.1
```

Copy from the reference above and input one of your interest protein sequences, Eg.

OPJ74548.1 #my pr sequences of interest.

```
Find sequence:
 OPJ74548.1_glucose-6-phosphatase_[Patagioenas_fasciata_monilis]
MESGMNVLHDSGIQATRWLQQHFQGSQDWFLFISFAADLRNAFFVLFPIWFHVSESVGVRLIWVAVIGDW
LNLVFKWILFGERPYWWVHETNYYSNTSAPEIQQFPLTCETGPGSPSGHAMGAAGVYYVMVTAILSAAAG
KKQSRTLKYRVLWTVLWTGFWAVQVCVCLSRVFIAAHFPHQVIAGVISGMAVAKTFQHVRCIYHASLRRY
LGITLFLFTFALGFYLLLRALGVDLLWTLEKAQRWCDRPEWVHMDTTPFASLLRNLGILFGLGLALNSHM
YLESCRGKQGQHLPFRLGCAVTSLLVLHLFDAFKPPAHMQLLFYVLSFCKSAAVPLATAGLIPYCVSQLL
ATQDKKGV

Scan a protein sequence with motifs from the PROSITE database

Do you have other proteins to scan (Default is Yes)?
Yes or No?
```

Yes #if you have other proteins to scan

```
Please choose one sequence that you're interested in and input the accession number.The default is XP_031456423.1
```

You can input your sequences of interest one by one in this way.

When you finish the input of all your sequences, type:

```
Do you have other proteins to scan (Default is Yes)?
Yes or No?
```

No

The program will print your input sequences and the output motif results.

```
Ok, protein sequence motif scanning finished.
Your input sequences are as follows:
OPJ74548.1_glucose-6-phosphatase_[Patagioenas_fasciata_monilis]
MESGMNVLHDSGIQATRWLQQHFQGSQDWFLFISFAADLRNAFFVLFPIWFHVSESVGVRLIWVAVIGDW
LNLVFKWILFGERPYWWVHETNYYSNTSAPEIQQFPLTCETGPGSPSGHAMGAAGVYYVMVTAILSAAAG
KKQSRTLKYRVLWTVLWTGFWAVQVCVCLSRVFIAAHFPHQVIAGVISGMAVAKTFQHVRCIYHASLRRY
LGITLFLFTFALGFYLLLRALGVDLLWTLEKAQRWCDRPEWVHMDTTPFASLLRNLGILFGLGLALNSHM
YLESCRGKQGQHLPFRLGCAVTSLLVLHLFDAFKPPAHMQLLFYVLSFCKSAAVPLATAGLIPYCVSQLL
ATQDKKGV

XP_031456423.1_glucose-6-phosphatase_[Phasianus_colchicus]
MPYATLQPVFPLFASCSKHVCNNPSVMHPSAKVVLGPKINAGASDLSAANTNPWGIKLRWSTGQPEPSRR
LRRRMEAPMNLLHDAGIQATHWLQEHFQGSQDWFLFISFAADLRNTFFVLFPIWFHLCEPVGIRLIWVAV
IGDWLNLVFKWILFGERPYWWVHETDYYSNTSAPTIQQFPLTCETGPGSPSGHAMGAAGVYYVMVTALLS
```

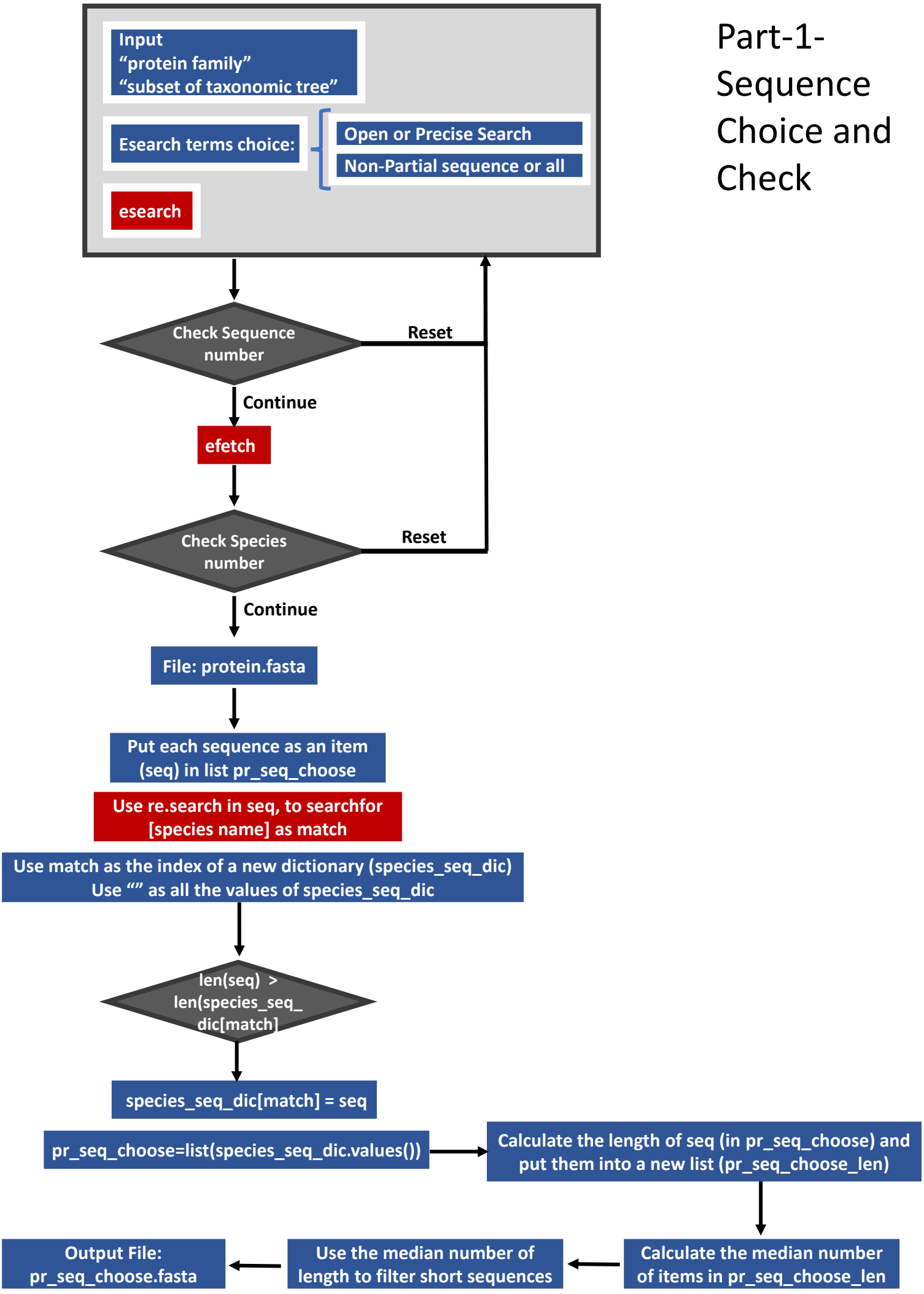
```
Output results for all the protein you scanned are in file 'pr_seq.patmatmotifs'
Name of motif(s) associated with the protein you interested in:
All kinds of motifs found are in the following list:
 ['Motif = AMIDATION']
```

The output file “pr_seq.patmatmotifs” will go the screen automatically.

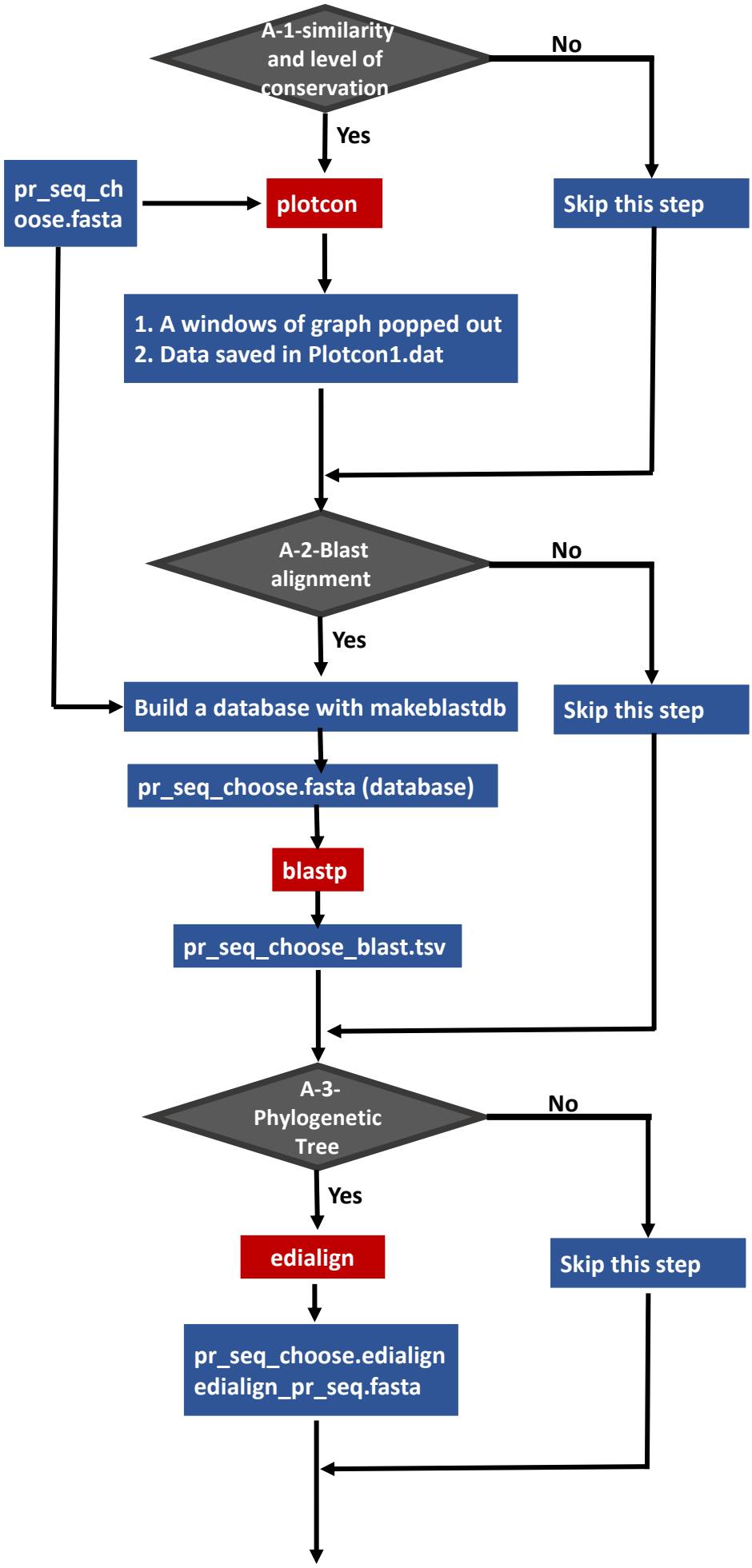
Some Other Useful Instructions:

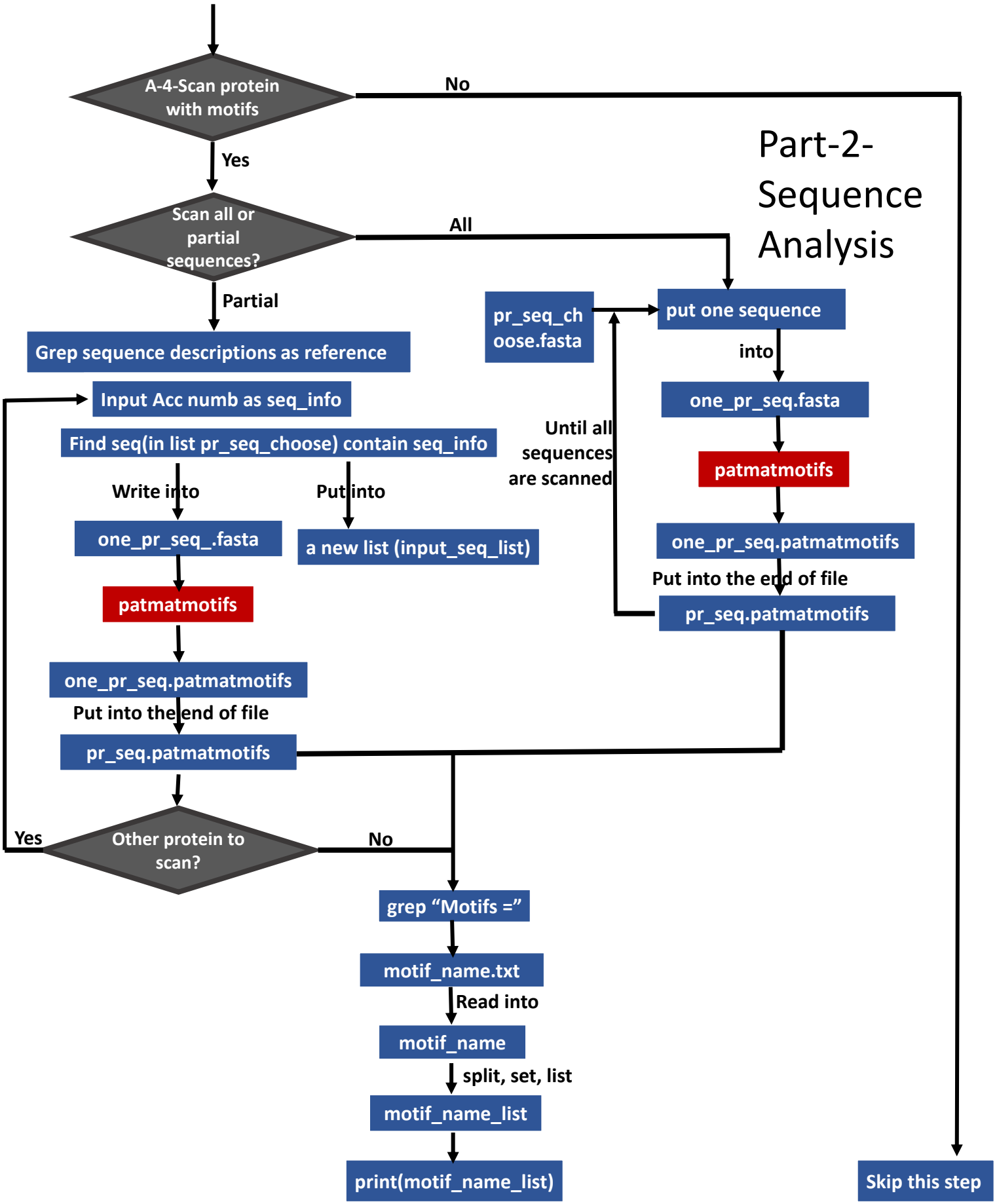
- To use the default number, you can just press “Enter” on your keyboard.
- For the analysis part, the content of some output file will go to screen automatically.
- you can scroll mouse up and down to view the contents.
- You can also press “space” on your keyboard to see the next page.
- You can press “q” on your keyboard to quit and go back to the code line.
- If you want to copy something, Do NOT press “Control+C”, please use Right-click of your mouse and choose “copy” instead!
- If you want to stop a running program, press “Control+C”.

Part-1- Sequence Choice and Check



Part-2- Sequence Analysis




A-4-Scan protein with motifs
No
Yes
Part-2-Sequence Analysis
Scan all or partial sequences?
All
Partial
pr_seq_choose.fasta
put one sequence
into
one_pr_seq.fasta
patmatmotifs
one_pr_seq.patmatmotifs
Put into the end of file
pr_seq.patmatmotifs
Until all sequences are scanned
Grep sequence descriptions as reference
Input Acc numb as seq_info
Find seq(in list pr_seq_choose) contain seq_info
Write into
Put into
one_pr_seq_.fasta
a new list (input_seq_list)
patmatmotifs
one_pr_seq.patmatmotifs
Put into the end of file
pr_seq.patmatmotifs
Yes
Other protein to scan?
No
grep "Motifs ="
motif_name.txt
Read into
motif_name
split, set, list
motif_name_list
print(motif_name_list)
Skip this step