Help manual

This program fetches protein sequences using protein family and taxonomic group specified by a user, produces Similarity Plot of Alignment Sequences and performs BlastP against the Swiss-Prot database [1] to find regions of local similarity between data bases. EXAMPLE_glucose-6-phosphatase_aves_report.pdf contains an example of generated report.

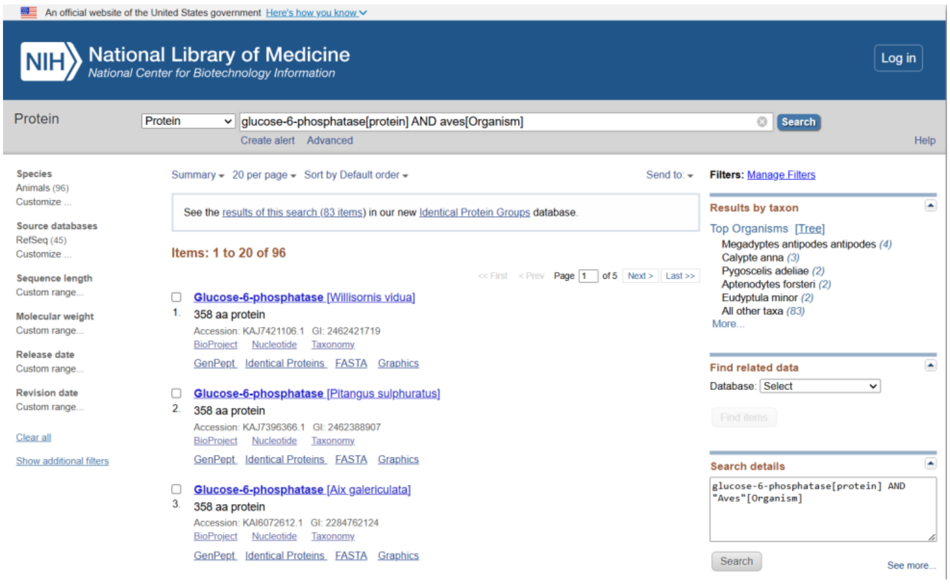
Program can be run in a terminal by typing:

```
./my_program.py
```

If not installed, get FPDF:

```
pip install fpdf
```

After program starts it prompts user to enter the protein family (e.g. glucose-6-phosphatase) and taxonomic group of interest (e.g. Aves). Then it searches NCBI database [2] for protein sequences from the specified subset of the taxonomic tree.



glucose-6-phosphatase[protein] AND aves[Organism]

Species
Animals (96)
Customize ...

Source databases
RefSeq (45)
Customize ...

Sequence length
Custom range...

Molecular weight
Custom range...

Release date
Custom range...

Revision date
Custom range...

See the results of this search (83 items) in our new Identical Protein Groups database.

Items: 1 to 20 of 96

Page 1 of 5

□ Glucose-6-phosphatase [Willisornis vidua]
1. 358 aa protein
Accession: KAJ7421106.1 GI: 2462421719
BioProject Nucleotide Taxonomy
GenPept Identical Proteins FASTA Graphics

□ Glucose-6-phosphatase [Pitangus sulphuratus]
2. 358 aa protein
Accession: KAJ7396366.1 GI: 2462388907
BioProject Nucleotide Taxonomy
GenPept Identical Proteins FASTA Graphics

□ Glucose-6-phosphatase [Aix galericulata]
3. 358 aa protein
Accession: KAI6072612.1 GI: 2284762124
BioProject Nucleotide Taxonomy
GenPept Identical Proteins FASTA Graphics

Filters: Manage Filters

Results by taxon
Top Organisms [Tree]
Megadyptes antipodes antipodes (4)
Calypte anna (3)
Pygoscelis adeliae (2)
Aptenodytes forsteri (2)
Eudyptula minor (2)
All other taxa (83)
More...

Find related data
Database: Select

Search details
glucose-6-phosphatase[protein] AND "Aves"[Organism]

After fetching the protein sequences user has an option to choose a different data set if unhappy with the results. There are four possible options to narrow down the dataset for further analysis.

Example:

```
There are 96 sequences in the current data set. Which sequences
would you like to use for conservation analysis?
A. all of them (not recommended)
B. choose a smaller data set at random (specify the size)
C. choose a new data set from the same Genus
D. choose organims from a current data set manually
```

Option A is not recommended as plotting conservation level between too many sequences might lead to information loss. ***Option B*** allows user to choose a smaller data set at random. After choosing Option C program counts genera and returns them to the user in order from the highest to lowest frequency, user can limit the number of displayed genera. From there, user can either type names of genera of interest or choose a number of genera with the highest species counts.

```
There are 62 Genera. Select the number of genrea you want to
display along with their counts. They will be presented in
order of highest to lowest frequency.
5
Eudyptes: 7
Corvus: 4
Pygoscelis: 4
Megadyptes: 4
Spheniscus: 4
```

Option D displays all species of interest and prompts the user to type the names manually.

Then, Similarity Plot of Alignment Sequences is generated and BlastP is performed on a new subset. You might have to wait for the Similarity Plot to load, be patient!

Important: typing species/genera is not caption sensitive but in case of any spelling errors given specie may not be added to the new data set for further analysis. If user's input is invalid and there are no matched species name the program will prompt user to re-enter the information.

Results folder contains:

{QUERY}_report.pdf – "QUERY" will be replaced with protein family and taxonomic group. It contains all files below as one PDF file.

blastp_con_an_deq.txt - BlastP results

con_an_seq_out.fasta – Subset used for Similarity Plot of Alignment Sequences and BlastP

my_aligned_seq.fasta – Aligned con_an_seq_out.fasta

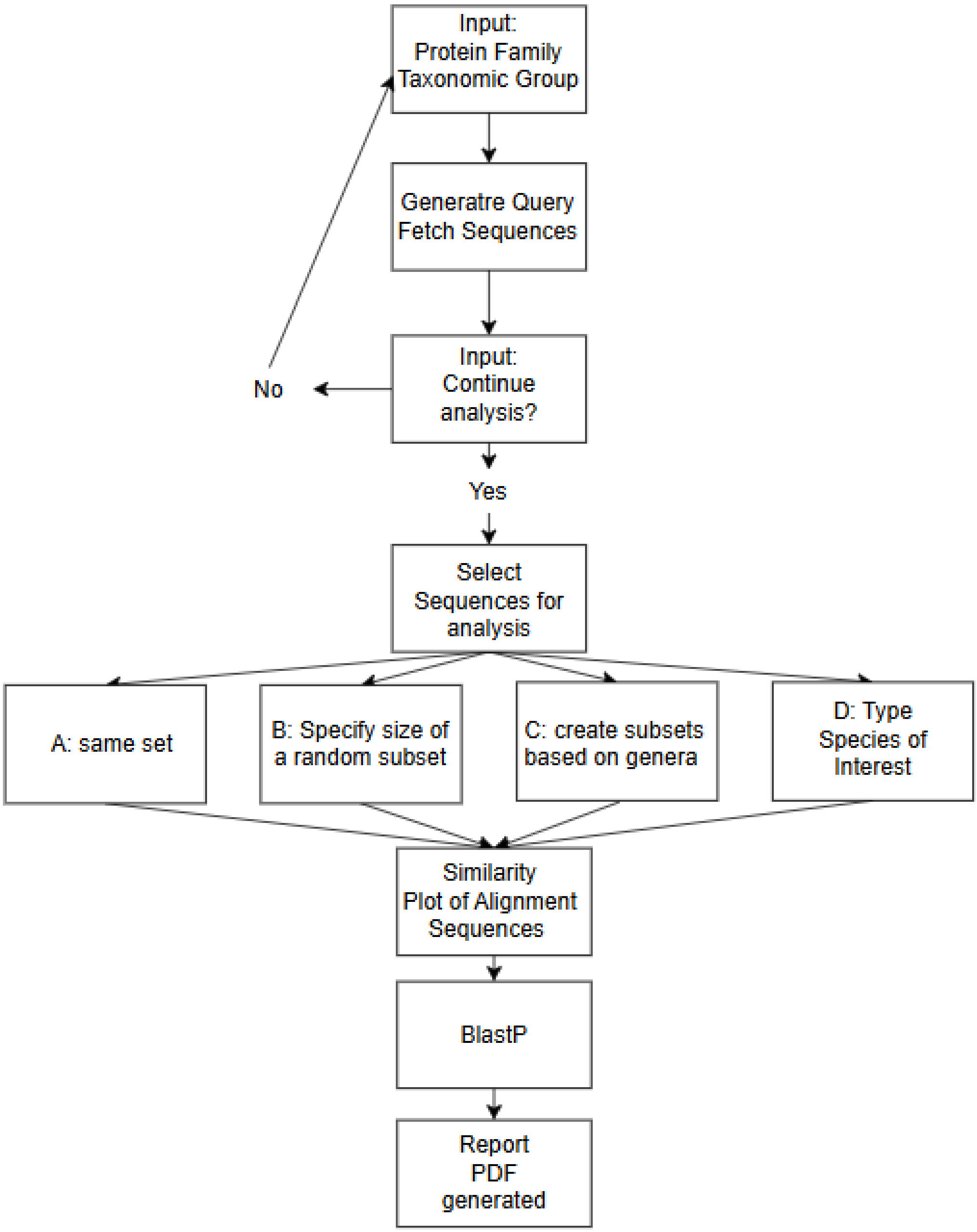
my_sequence.fasta – All initially fetched sequences

plotcon.1.png - Similarity Plot of Alignment Sequences

results.txt – query and number of sequences, if option C/D chosen also names of genra/species

Maintenance Manual

Flowchart illustrates the program's workflow:



Functions:

valid_number(top_num) – "Error Trap" makes sure user input is a number smaller than top_num

add_line(line) – adds line "line" to the {QUERY}_report.pdf

References:

1. **UniProt**:

UniProt Consortium. "UniProt: a worldwide hub of protein knowledge." *Nucleic Acids Research*, vol. 47, Database issue, 2019, pp. D506–D515. DOI: 10.1093/nar/gky1049. Access: https://www.uniprot.org/

2. **NCBI (National Center for Biotechnology Information)**:

Coordinators NR. "Database Resources of the National Center for Biotechnology Information." *Nucleic Acids Research*, vol. 50, Database issue, 2022, pp. D20–D26. DOI: 10.1093/nar/gkab1112.
Access: https://www.ncbi.nlm.nih.gov/