BPSM ICA2

B242415-2023

Github Repository Link:

https://github.com/B242415-2023/ICA2

ccrypt Encryption Key:

hial

Tools used:

EMBOSS (Rice, Longden & Bleasby, 2000), Entrez (Sayers et al., 2022), ClustalO (Sievers et al., 2011)

User Manual:

script.py is a python script where a user is able to define a protein family and taxonomic group to gather a set of protein sequences from the NCBI database that match the stated filters to align sequences, plot level of conservation, search for PROSITE motifs, and conduct other EMBOSS analysis such as: identifying signal sequence cleavage sites, plot protein charge, residue frequency, nucleic acid binding (helix-turn-helix) motifs, hydrophobic moments, isoelectric points, predicted coiled coil regions, and general protein statistics.

This tool can be used to find out more about the protein features of sequences in a chosen protein family and taxonomic group. sigcleave may provide insight to protein localization depending on the signal peptide. Protein charge, hydrophobic moments and isoelectric points give information about the protein features, while nucleic acid binding motifs, predicted coiled coil motifs give information about the protein structure.

Potential biological questions:

What features are shared amongst a protein family in a taxonomic group?

Which features are key to its function?

Which features are conserved amongst the queried protein sequences?

Requirements:

Packages installed: Entrez, EMBOSS, ClustalO

Python modules installed: os, sys, subprocess

How to use:

1. chmod 700 ./script.py - Gives permission to run script
2. ./script.py - Run the script
3. Input:
    a. Taxonomic group (will also search subgroups of taxonomic group)
    b. Protein family (Avoid using too vague or pluralized familes)
    c. Remove partial sequences? (y/n) (y to remove partial sequences)
    d. Remove sequences with rmkeywords? (y/n) (y to remove sequences with an element in rmkeywords in their headers)
        i. (rmkeywords default to : “associated”, “predicted”, “isoform”)
4. (Optional) If sequences violate any validity checks, will prompt user for next actions. Otherwise will, will continue automatically
5. Wait for analysis. Results are put into the ./results/ directory

How it works:

1. User defines taxonomic group, protein family, and additional filters for the NCBI database search.
2. Script use input to search against NCBI database to gather protein sequences in one fasta file.
3. Validity checks for gathered sequences.
4. Sequences get parsed to various analysis programs and EMBOSS packages. Results dir or file as shown:
    a. clustalo - sequence alignment
        i. ./results/aligned.fasta
    b. plotcon - plotting sequence conservation
        i. ./results/plotcon.1.png
    c. patmatmotifs - searching for PROSITE motifs
        i. ./results/patmatmotifs
    d. sigcleave - searching for signal peptide cleavage site motifs
        i. ./results/sigcleave
    e. charge - protein charge plot
        i. ./results/charge
    f. freak - residue frequency
        i. ./results/freak
    g. helixturnhelix - searching for nucleic acid binding site motifs (helix turn helix)
        i. ./results/helixturnhelix
    h. hmoment - plotting hydrophobic moments
        i. ./results/hmoment
    i. iep - calculating isoelectric points
        i. ./results/iep
    j. pepcoil - searching for coiled coil regions
        i. ./results/pepcoil
    k. pepstats - various statistics for the protein
        a. protein weight, length, avg residue weight, charge, isoelectric point, stats for each amino acid, stats for each physico chemical class of amino acid, molar extinction coefficient,

extinction coefficient at 1mg/ml, probability of protein expression in E. coli inclusion bodies

ii. ./results/pepstats

5. Outputs from analysis put into ./results/ directory

Example for glucose-6-phosphatase in Aves

Inputs:

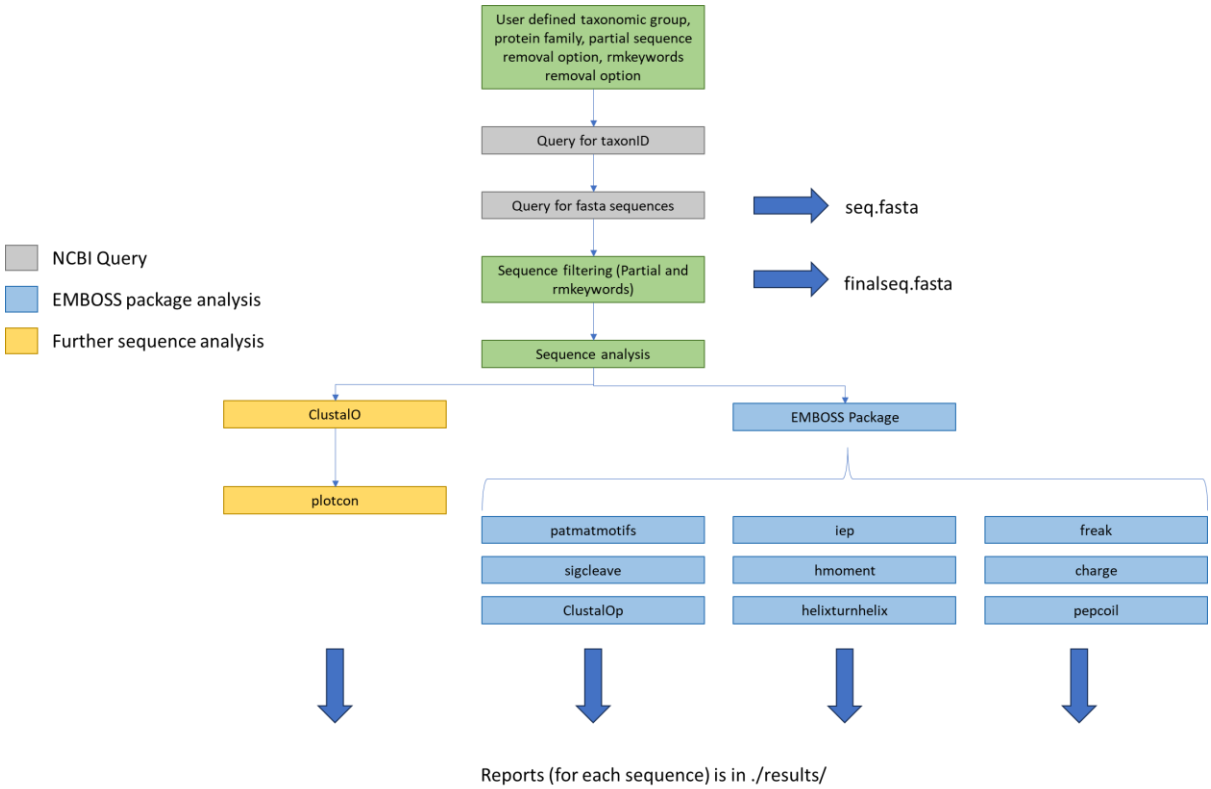
- Taxonomic group: Aves
- Protein family: glucose-6-phosphatase
- Remove partial sequences? n
- Remove sequences with rmkeywords? n

Outputs:

- seq.fasta .fasta file with all sequences gathered from NCBI database
- finalseq.fasta .fasta file with filtered sequences
- ./results/{dir} various EMBOSS sequence analysis on each sequence

Maintenance Manual:

Figure 1: Flowchart of script.py



1. OPTIONS
    a. Variables that can be changed easily that are parameters of several commands. Contains:
        i. max_number_of_sequences - variable that dictates the threshold where a sequence quantity check will be conducted by asking the user if they want to continue.
        ii. availthreads - variable that dictates the number of threads some commands use. Automatically gets max number of threads but can be manually changed by commenting out and setting variable to an int.
        iii. rmkeywords – list of keywords the script will search for in the fasta headers when filtering out sequences.
2. FUNCTIONS
    a. Contains defined functions.
        i. indivbash – a function that will iterate through every sequence in the seqdict dictionary to run a specific bash command. Arguments include: the bash command, an output file location, format, and output file format. Output file name is the header of the sequence with format altered for file name suitability.
3. Gathering user input
    a. Gathers user query taxonomic group (taxo), protein family (pfam), partial sequence removal option(pfampartialflag), keyword filter option from user (rmkeywordflag).
    b. Error catch by using:
        i. Checks if user input is y/n. Will retry if anything other than y/n

4. Gathering protein sequences
    a. Query for taxonID via esearch | efetch -format UID to get taxonID into esearchTaxoUID variable.
        i. Error catch by using:
            1. try: & except: to catch additional errors.
            2. esearchTaxoUID == “” to return an error message if taxonID comes back empty, which may occur in misspelt or invalid queries.
    b. Query for protein sequence using esearch | efetch -format fasta into seq.fasta.
        i. NOT PARTIAL filter depends on pfampartialflag variable.
        ii. Error catch by using:
            1. pfam[-1] == “s” before search to catch user inputted protein families which are plural. NCBI database does not list plural protein families. Gives user option to continue with plural or not.
            2. try: & except: to catch additional errors.
    c. Return statistics about gathered sequences.
        i. Counts “>” in seq.fasta to get number of sequences gathered.
            1. Error catch by using:
                a. seqcount == 0 for invalid query protein family (cannot be taxonID failure because of earlier empty taxonID error catch).
                b. seqcount > max_number_of_sequences for large datasets. Gives user option to continue or stop.
5. Data preparation
    a. Create allseq list which contains all individual sequences in seq.fasta
    b. Create seqheaders list which contains all headers of all sequences in seq.fasta
    c. Create header:fasta dictionary seqdict for easier downstream calling
    d. Removing sequences with rmkeywords in it
        i. Iterate through each rmkeyword while iterating through each sequence in the dictionary to remove any header:sequence entries in the dictionary containing any rmkeyword
        ii. Write final dictionary fasta into a finalseq.fasta file
        iii. Report number of sequences removed
    e. Gather number of unique species in dataset
        i. Find positions of “[“ and “]” in headers to extract species name
        ii. Calculate number of unique species
6. SEQUENCE ANALYSIS
    a. ClustalO and plotcon
        i. Calls bash command directly with os.system since command is able to take one fasta file with all the sequences (ClustalO) or one file for sequence alignment (plotcon)
    b. patmatmotifs, sigcleave, charge, freak, helixturnhelix, hmoment, iep, pepcoil, pepstats
        i. Uses indivbash function to iterate over each sequence in seq dictionary to output a report in ./results/.

References

Rice, P., Longden, I. & Bleasby, A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends in Genetics : TIG.* 16 (6), 276-277. 10.1016/s0168-9525(00)02024-2.

Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., Connor, R., Funk, K., Kelly, C., Kim, S., Madej, T., Marchler-Bauer, A., Lanczycki, C., Lathrop, S., Lu, Z., Thibaud-Nissen, F., Murphy, T., Phan, L., Skripchenko, Y., Tse, T., Wang, J., Williams, R., Trawick, B. W., Pruitt, K. D. & Sherry, S. T. (2022) Database resources of the national center for biotechnology information. *Nucleic Acids Research.* 50 (D1), D20-D26. 10.1093/nar/gkab1112.

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D. & Higgins, D. G. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology.* 7 (1), 539. 10.1038/msb.2011.75.