Protein Domain Finding With Hidden Markov Model

Read Me

Jiayi Shou

December 9th, 2020

Introduction

This program is built with an intent to investigate protein domain searching and model building using Hidden Markov Model. It does not rely on external packages besides the built in packages from golang. The program is built with an intent to have easy user experiences. To run the program, simply hit go run *.go and all the instruction would be given by the program. The program has mainly four functions:

1. Produce a profile HMM given multi-alignments
2. Check the likelihood of a sequence belonging to domain family
3. Find the most probable path of a sequence aligning to a HMM, as well as the probability of that path produced by the HMM.
4. Generate fictional domain sequences with profile HMM

Important notice: If the user does not have an alignment file that has the corresponding format, then Step 1 is a basis for the following functions.

One: Build Profile Hidden Markov Model

Before proceeding to step one, we need to gather alignment files. The program file has already provided two sample folders containing the alignment files. For more options, please download alignment files either from:

Pfam (http://pfam.xfam.org/) or BLAST(https://blast.ncbi.nlm.nih.gov/Blast.cgi).

1. **Using aligned protein domain family from Pfam**
   a. VIEW A PFAM ENTRY allows user to search for interested domain families.
   - After finding the domain family on Pfam website, choose alignments
   - Under the format alignment section, choose
     - Alignment: Seed
     - Format: Selex
     - Order: (either one would work)
     - Sequence: All Upper Case,
     - Gaps: Gaps as “-” (dashes)

   Download the File

2. **Use Blast to find similar sequences**
    - On Blast page, choose protein BLAST
    - Enter the interested query sequence
    - Under the Program Selection, algorithm, choose
        - DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)
        - BLAST it!
    - Once results appear, under the Alignments section, choose
        - Alignment view: Flat query-anchored with letters for identities
        - Download Text (aligned sequences)

Once the file is ready, you are ready to build the profile HMM. Start the program and choose option 1. The program will ask where you gather the file as well as the domain code (to label the output file). You shall find the output file as domaincodeTrMap.txt and domaincodeEmiMap.txt in the same directory as the program. Once you have these file, you can run other functions of the program.

Two: Likelihood of a Sequence Given Domain Family

First, check if you have the files ready. To run this function, you will need a sequence, a domaincodeTrMap.txt and domaincodeEmiMap.txt. The sequence should not have any dashes, dots or space in it. For example, a good sequence input would look like this:

VAKFDYVAQQEQELDIKKNERLWLLDDSKSWWRVRNSMNKTGFVPS

Once you choose option 2 in the program, simply paste the sequence, file names of transition map TrMap and emission map EmiMap each on its own line.

Important notice: Sequence should have no dashes! Please do not swap the sequence of the files.

Three: Most Probable Path of a Sequence Aligning to HMM

To run this function, you will need a sequence, a domaincodeTrMap.txt and domaincodeEmiMap.txt. The sequence should not have any dashes, dots or space in it as mentioned in Two. Important notice: The sequence actually has to be the exact same length as the matching states of the transition map… Otherwise it would have problems.

Once you choose option 3 in the program, simply paste the sequence, file names of transition map TrMap and emission map EmiMap each on its own line.

Four: Generate fictional domain sequences

To run this function, you will need a number indicating how many fictional domains you want, a domaincodeTrMap.txt and domaincodeEmiMap.txt. The generated sequences will appear on the screen. You can enter them to BLAST and see if you are lucky!! If you BLAST find alignments, then you got really really lucky.

Contact

Any further problems running the program: Please reach to Jiayi at jiayisho@andrew.cmu.edu