SYSB 3036 W06: Annotating a Complete Genome with HMMs

Frank Aylward[1]

[1]Virginia Tech

Mar 04, 2019

Working

dx.doi.org/10.17504/protocols.io.ytcfwiw

Frank Aylward
Virginia Tech

PROTOCOL STATUS

Working

We use this protocol in our group and it is working

1

First let's download some files from GitHub:

git clone https://github.com/faylward/hmm_tutorial

And move inside the new folder and see what is there:
cd hmm_tutorial

ls

2 In the new file you should see a gzipped file that contains proteins from a mystery genome. We can gunzip it now:

gunzip mystery_genome_proteins.faa.gz

For this analysis we want to annotate all of the proteins in this genome using a suite of HMMs. We will use the TIGRfam database today. I have uploaded this database onto a Virginia Tech library link, so we can download it with the followingn command.

wget -O tigrfams.tar.gz https://data.lib.vt.edu/downloads/pc289j19b

This should download a .tar.gz file, which is essentially a gzipped folder, also called a "tarball". This is a bit different then just a gzipped file, and we need a special command to unzip it.

tar xvfz tigrfams.tar.gz

If you look inside of the tigrfams file you should see two files- one HMM file and one .txt file.
The HMM file contains thousands of HMMs all merged together. This is the full TIGRfam database. Note that the file is quite large.
The .txt file contains descriptions of each HMM. When we want to annotate a protein we don't just want to know that the best hit was TIGR20034. We also want to know what the function of TIGR20034 is. That's what this file is for.

3

To run the initial HMM annotation we can use the hmmsearch command in HMMER3:

hmmsearch --tblout mystery_genome.hmmout --cut_tc tigrfams/all_tigrfam.hmm mystery_genome_proteins.faa > full_hmmout.txt

Note that in the command above we are using the --cut_tc option, which specifies that we want to use pre-specified bit score cutoffs to classify these proteins. All hits that have bit scores below the specifications for each HMM will be disregarded, regardless of their e-value.

This is an open access protocol distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

4 After we run the hmmsearch we can parse the results with a Python script that I wrote.
This script goes through the tabular hmmsearch output and finds only the best hits for each protein. It also removes the weird space-delimited format that HMMER3 insists on using and converts it into a nice tab-delimited format.
We can run it with:

python parse_hmmout.py tigrfams/profile_annotations.tigr.txt mystery_genome.hmmout > mystery_genome.parsed.hmmout.txt

Note that we need to specify the profile annotations file, since we want the final annotation file to contain the TIGRfam descriptions.

5 Another way to annotate genomes, and get some visualizations of pathway-level annotations, is to use the Kyoto Encyclopedia of Genes and Genomes (KEGG). They host an annotation server here:

https://www.genome.jp/kaas-bin/kaas_main

Upload the protein file, add in your email, and make sure to use the "For Prokaryotes" option at the bottom of the page so that the correct reference database is used. A link will be mailed to you, and you need to click the "submit" option in that email for the annotation to begin. After that you will need to wait ~10-45 minutes for the annotation to finish, and the link will be mailed to you again.

This is an open access protocol distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

This is an open access protocol distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited