SYSB 3036 W02: Parsing FASTA files

Frank Aylward[1]

[1]Virginia Tech

dx.doi.org/10.17504/protocols.io.vfwe3pe

Frank Aylward
Virginia Tech

ABSTRACT

Week 1
Introduction to parsing FASTA files.

Commands to be entered into the command line are in bold.

Here we will be using various base Unix commands such as head, tail, sort, wget, and others.

We will also be using the seqkit tool to process FASTA files. The main page for seqkit is here:
https://github.com/shenwei356/seqkit

PROTOCOL STATUS

Working
We use this protocol in our group and it is working

1 Today we will be looking at the genome of Yersinia pestis, which can be found on NCBI at this location
ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/009/065/GCA_000009065.1_ASM906v1

Copy this URL into your browser and take a look at the files. These are publicly-available files that are made available from the National Center for Biotechnology Information (NCBI).

Note that many of the files are in a compressed .gz format.

.fna files are Fasta Nucleic Acid (chromosome or gene sequences)
.faa files are Fasta Amino Acid (protein sequences)

Today we will be most interested in the gene and chromosome files.

2 To get started let's download the main genome FASTA file for Yersinia Pestis CO92

wget ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/009/065/GCA_000009065.1_ASM906v1/GCA_000009065.1_ASM906v1_genomic.fna.gz

This command uses the common Unix utility "wget", which will download a file directy to the folder in which you are located. After doing this you should see the .fna.gz file in your folder. You can check this with the "ls" command.

3 Because this file is compressed, we must first uncompress it with the "gunzip" command (another common Unix utility).

gunzip GCA_000009065.1_ASM906v1_genomic.fna.gz

After this you should see the exact same file, only without the .gz ending. You can check this with the "ls" command.

4 Now to start analyzing this FASTA file we first want to check on the formatting.
Sometimes genome files can be quite large, so we don't want to open the entire file with a text editor. Instead we can just check the first and last few lines to see what the format looks like. For this we can use the "head" and "tail" Unix commands.

head GCA_000009065.1_ASM906v1_genomic.fna

and

tail GCA_000009065.1_ASM906v1_genomic.fna

You should see a pretty typical FASTA format. Header lines start with a ">" and provide names and descriptions, and subsequent lines have the actual sequence information (in this case ATGCs since the sequence is DNA).

5 Now that we have confirmed this is a typical FASTA file, we can start analyzing it with the "seqkit" tool.

The home page for seqkit with instructions for use is here: https://github.com/shenwei356/seqkit

You can also get a list of instructions by typing:

seqkit --help

seqkit is a very versatile tools and it has a large number of sub-commands. We will primarily be using the "stats" and "fx2tab" commands, so check out the help menu for those with:

seqkit stats --help

and

seqkit fx2tab --help

6 Now let's use the "stats" subcommand in seqkit to get some genome statistics for Yersinia pestis:

seqkit stats GCA_000009065.1_ASM906v1_cds_from_genomic.fna

7 Now let's get some stats for the individual sequences in the file using the "fx2tab" sub-command:

seqkit fx2tab -g -l -n GCA_000009065.1_ASM906v1_genomic.fna

8 Now let's download and analyze the individual genes present in the Yersinia pestis genome:

We will download and unzip the "cds_from_genomic.fna.gz" file from the FTP site in Step 1:

wget
ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/009/065/GCA_000009065.1_ASM906v1/GCA_000009065.1_ASM906v1_cds_from_genomic.fna.gz
gunzip GCA_000009065.1_ASM906v1_cds_from_genomic.fna.gz

Don't forget to use the "ls" command afterwards to make sure the files are there.

9

Now let's get some basic stats about the genes with the "stats" command.

seqkit stats GCA_000009065.1_ASM906v1_cds_from_genomic.fna

How does the total length of the protein-coding genes compare to the total length of the whole genome?
What is the range of gene lengths?

10 Let's look at some stats from individual genes using the "fx2tab" command:

seqkit fx2tab -a -i -g -l -n GCA_000009065.1_ASM906v1_cds_from_genomic.fna | head

Note that we are piping the command to the "head" command here, so that only the first 10 lines are shown. Otherwise thousands of entries would flood our terminal, which is always difficult to interpret (and may cause it to crash).

11 Now let's try to sort the genes based on their length, so that we can find the names of the longest and shortest genes:

seqkit fx2tab -a -i -g -l -n GCA_000009065.1_ASM906v1_cds_from_genomic.fna | sort -r -n -k 2,2 | head

This should return the longest 10 genes. The "sort" command uses several flags.
-r indicates a reverse sort (default is from low to high.
-n indicates a numeric sort (default is alphabetical).
-k denotes the columns to sort by. The "2,2" means we are sorting only by the second column. Note that all whitespace counts as a single tab here, so the second column is the length (it would be the 4th if we exported the results to a file).

We can do the same with "tail" instead of head to find the names of the shortest genes.

12 We can use the same logic as above to find the genes with the highest and lowest %GC content. For this we need to sort by the third column.

seqkit fx2tab -H -a -i -g -l -n GCA_000009065.1_ASM906v1_cds_from_genomic.fna | sort -rn -k 3,3 | head

13 Now let's say we want to retrieve the actual DNA sequence of the gene with the highest %GC content. We can do this by using a "seqkit fx2tab" command and piping the results to a "grep" command.

seqkit fx2tab -H -a -i -g -l GCA_000009065.1_ASM906v1_cds_from_genomic.fna | grep "lcl|AL590842.1_cds_CAL20117.1_1427"

Note that we did not use the "-n" flag in the seqkit fx2tab command, since this time we wanted the sequence (before we just wanted the statistics).