Analysis Pipeline Demo

March 8, 2019

0.1 # Demonstration of an Analysis Pipeline

In general terms, we usually divide command line instruction into two parts. In the first part we focus on commands that are used to navigate the file system. In the second, we introduce commands that can be used to interact with files - creatings, reading and writing, deleting, etc. Or as an acronym - CRUD (__c__reate, __r__ead, __u__pdate, __d__elete).

We've looked at navigation commands, but before moving into the CRUD commands, we're going to step through an example workflow provided by one of your colleagues which demonstrates the efficiencies that can be gained from command line scripting.

The workflow requires us to install two applications, `ncbi-blast` and `wget`. First, we can check to see if they are already installed on our systems using the `-v` or `--version` flag.

```
jwheel01@UL-D8VH1S22 MINGW64 ~
wget --version
SYSTEM_WGETRC = c:/progra~1/wget/etc/wgetrc
syswgetrc = C:\Program Files (x86)\GnuWin32/etc/wgetrc
GNU Wget 1.11.4

Copyright (C) 2008 Free Software Foundation, Inc.
License GPLv3+: GNU GPL version 3 or later
<http://www.gnu.org/licenses/gpl.html>.
This is free software: you are free to change and redistribute it.
There is NO WARRANTY, to the extent permitted by law.

Originally written by Hrvoje Niksic <hniksic@xemacs.org>.
Currently maintained by Micah Cowan <micah@cowan.name>.
```

If the application is not installed, we will get an error:

```
jwheel01@UL-D8VH1S22 MINGW64 ~
blastn --version
bash: blastn: command not found

jwheel01@UL-D8VH1S22 MINGW64 ~
blastn -v
bash: blastn: command not found
```

On an Ubuntu or other Debian system, we can use the `apt` or `apt-get` command to install the applications as needed.

```
sudo apt-get --help

sudo apt-get install wget

sudo apt-get install ncbi-blast+
```

Note that you can install multiple applications with a single command:

```
sudo apt-get install wget, ncbi-blast+
```

For other platforms, download documentation is available:

- wget: https://www.gnu.org/software/wget/faq.html#download
- ncbi-blast: https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=Download

Depending on the system permissions of the user running the installation, in order to run either application it may be necessary in Windows systems to update the PATH environment variable to point at the downloaded executable.

A short term solution used here is to copy the executables to our desktop or another easily accessed location and use the relative or absolute path when running commands. Whichever solution we use, we can verify successful installations using the `-v` or `--version` commands as above:

```
# From a Windows Desktop:

jwheel01@UL-D8VH1S22 MINGW64 ~/Desktop/
./blast-2.8.1+/bin/blastn -version
blastn: 2.8.1+
Package: blast 2.8.1, build Nov 26 2018 11:53:19

wget --version
SYSTEM_WGETRC = c:/progra~1/wget/etc/wgetrc
syswgetrc = C:\Program Files (x86)\GnuWin32/etc/wgetrc
GNU Wget 1.11.4
```

Once we have `wget` and `ncbi-blast` installed, we use `wget` to download a dataset:

```
wget ftp://ftp.ensembl.org/pub/release-95/fasta/taeniopygia_guttata/dna/Taeniopygia_guttata.ta
SYSTEM_WGETRC = c:/progra~1/wget/etc/wgetrc
syswgetrc = C:\Program Files (x86)\GnuWin32/etc/wgetrc
--2019-03-07 10:48:16--  ftp://ftp.ensembl.org/pub/release-95/fasta/taeniopygia_guttata/dna/Ta
           => `Taeniopygia_guttata.taeGut3.2.4.dna.chromosome.Z.fa.gz'
Resolving ftp.ensembl.org... 193.62.193.8
Connecting to ftp.ensembl.org|193.62.193.8|:21... connected.
Logging in as anonymous ... Logged in!
==> SYST ... done.    ==> PWD ... done.
==> TYPE I ... done.  ==> CWD /pub/release-95/fasta/taeniopygia_guttata/dna ... done.
==> SIZE Taeniopygia_guttata.taeGut3.2.4.dna.chromosome.Z.fa.gz ... 21934154
```

```
==> PASV ... done.    ==> RETR Taeniopygia_guttata.taeGut3.2.4.dna.chromosome.Z.fa.gz ... done
Length: 21934154 (21M)

100%[=======================================================================================

2019-03-07 10:48:20 (8.43 MB/s) - `Taeniopygia_guttata.taeGut3.2.4.dna.chromosome.Z.fa.gz' save
```

Unzip the file - note we can use a wildcard ('*') so that we don't have to type the full name:

```
gunzip *.fa.gz
```

Run a query using `blastn`:

```
# Note I am using a relative path to the executable
# You may be able to just use 'blastn' without the path

./blast-2.8.1+/bin/blastn -query horornis.fasta -subject *Z.fa -evalue 1e-30 -outfmt 6 -out z_l
```

The above command may not print any output, but we can check that the 'z_loci.txt' file was created using the `ls` command:

```
jwheel01@UL-D8VH1S22 MINGW64 ~/Desktop
ls -l z_loci.txt
-rw-r--r-- 1 jwheel01 1049089 37514 Mar  7 10:53 z_loci.txt
```

Finally, we can search the output file for specific lines:

```
sed -i 's/|.*//g' z_loci.txt
```

The whole process can be added to a shell script for sharing or reuse:

```
#!/bin/bash
sudo apt-get install ncbi-blast+ &&
wget ftp://ftp.ensembl.org/pub/release-95/fasta/taeniopygia_guttata/dna/Taeniopygia_guttata.tae
gunzip *.fa.gz &&
blastn -query horornis.fasta -subject *Z.fa -evalue 1e-30 -outfmt 6 -out z_loci.txt &&
sed -i 's/|.*//g' z_loci.txt
```