TD Bioinfo

Pierre Lindenbaum - Institut du Thorax. Nantes. France

August 9, 2013

Abstract

Here are few questions to prepare the courses **"Advanced NCBI"** and **"Next-Generation-Sequencing"**.
The aim of those exercices is to assure that you have a basic knowledge of the tools and vocabulary that will be used during the courses.
Only basic linux commands are required here, no external program should be installed.

1 Linux

Briefly describe the usage of the following linux commands :

- man
- cd
- ls
- pwd
- ls
- cp
- rm
- mv
- cat
- more
- grep
- sort
- uniq
- paste

- join
- tr
- head
- tail
- mkdir
- curl
- awk
- make

What is **stdout** ?
What is **stderr** ?
What is **stdin** ?
How do you **redirect** the ouput of a program to a file ?
How do you **append** the ouput of a program to a file ?

In the HTTP protocol, what are the POST and GET methods ?
What is the JSON format ?
What is a XML well-formed document ?
What is a XML DTD ?
What is a XML Schema ?
What is a XML valid document ?
What is Xpath ?

2 General Bioinformatics

- What is a 'SNP' ?
- What is a 'Genome build' ?
- What's the approximate size of the human genome ?
- How many chromosomes is there in the human genome ?
- What the approximate size of the longest human chromosome ?
- What's the difference between a binary file and flat file ?
- How many bytes do you need to store the **length** of the human genome ?
- In a sequence containing only 'A', 'T', 'G' and 'C' , how many bases can you store in one byte ?

- Cite some advantages/inconvenients of storing data in a tab delimited format vs using a structured format (e.g: XML, JSON, ASN.1)
- Cite some advantages/inconvenients of handling data using a graphical interface (like Miscrosoft Excel) vs using the command line.
- Find the following nucleotide entry in the NCBI: *a Sequence of "Nicotiana tabacum" for the gene 'trnH' with a length comprised between 100 and 150 nucleotides and having a feature "variation".*
- Find the following protein entry in the NCBI: *a sequence for the gene COL1A1 published by Dr. Asara in the "Science" journal, that is not a sequence of "Mammut americanum" nor "Brachylophosaurus canadensis".*

3 Bash

What's the purpose of the following bash command line ? (`https://gist.github.com/lindenb/65e98e5752ea26eb9868`)

```
1  $ curl  "http://hgdownload.cse.ucsc.edu/goldenPath/
       hg19/chromosomes/chrM.fa.gz" |\
2  gunzip -c |\
3  tail -n +2 |\
4  tr "[:lower:]" "[:upper:]" |\
5  tr -d '\n' |\
6  tr "AC" "TG" |\
7  sed 's/\(.\)/\1#/g' /\
8  tr "#" "\n" |\
9  LC_ALL=C sort |\
10 uniq -c |\
11 sed 's/^[ ]*//' |\
12 cut -d ' ' -f1 |\
13 tr "\n" "␣" |\
14 awk '{printf("%f␣%%\n",$1/($1+$2));}' > result.txt
```

4 Makefile

(wikipedia:) *"In software development, Make is a utility that automatically builds executable programs and libraries from source code by reading files called makefiles which specify how to derive the target program".*
You'll find many simple tutorials for **make** on the web.

Here is a file named **Makefile** (`https://gist.github.com/lindenb/65e98e5752ea26eb9868`).

```
.PHONY:all clean
.SECONDARY=
SEQUENCES=alpha beta gamma

%.rna:%.dna
	tr "Tt" "Uu" < $< > $@
%.fa:%.rna
	echo ">$(basename␣$(notdir␣$<))" > $@ && cat $
	   < >> $@


all: sequences.fa

sequences.fa : $(foreach S,${SEQUENCES}, $(addsuffix .
   fa,${S}))
	cat $^ > $@

alpha.dna:
	echo  "ATCGATCGCATCGATATAGC" > $@

beta.dna:
	echo  "ATCCGGCTAAGCTATATAGCT" > $@

gamma.dna:
	echo  "CCTTGACTGAGCGATCGGG" > $@

clean:
	rm -f *.fa *.dna *.rna
```

Answer the following questions (shell should be '`bash`')

- type '`make clean && make`'. What happens ?
- what is a target ?
- what is a dependency ?
- what is the symbol '`$@`' ?
- what is the symbol '`$<`' ?
- what is the symbol '`$^`' ?
- where are the '`tab (\t)`' characters in the Makefile ?
- what is the meaning of the line

```
%.rna:%.dna
	tr "Tt" "Uu" < $< > $@
```

- type '`make clean && make all`'. What happens ?
- why '`all`' was placed at the top ?
- what is the default target ?
- type '`make clean && make && make && make && make && make`'. What happens ?
- type '`make clean && make gamma.fa`'. What happens ?
- type '`make clean && make delta.fa`'. What happens ?
- type '`make clean && make && make -B gamma.fa`'. What happens ?
- type '`make clean && make -n`'. What happens ?
- type '`make clean && make && touch alpha.fa && make` '. What happens ?
- type '`make clean && make && rm alpha.fa && make` '. What happens ?
- why '`clean`' and '`all`' were declared as '`.PHONY`'.
- type '`make clean && make -j 3 all`'. What happens ?
- remove the line '`.SECONDARY=`' and type '`make clean && make all`'. What happens ?
- what is the benefit of using a makefile rather than a shell script ?

5 The Human Genome

The chromosomes for the latest Human build are available at: `http://hgdownload.cse.ucsc.edu/goldenPath/hg19/chromosomes/`. Answer the following question using a linux command line:

- What's the length of the chr22 ?
- Look at the 100000 first lines of the chr22. Explain what you see.
- Get a count of each base in the chr22
- Using the linux commands 'curl', 'gunzip', 'tr' and 'rev', get the reverse complement of the chromosome chrM.

6 Using XSLT

Many NCBI web-services produce a XML document. For example, the following URL is a XML-based list to the databases available from the **NCBI**: `http://eutils.ncbi.nlm.nih.gov/entrez/eutils/einfo.fcgi`.

```
<!DOCTYPE eInfoResult PUBLIC "-//NLM//DTD␣eInfoResult,
   ␣11␣May␣2002//EN" "http://www.ncbi.nlm.nih.gov/
   entrez/query/DTD/eInfo_020511.dtd">
<eInfoResult>
  <DbList>
    <DbName>pubmed</DbName>
    <DbName>protein</DbName>
    <DbName>nuccore</DbName>
    <DbName>nucleotide</DbName>
    <DbName>nucgss</DbName>
    <DbName>nucest</DbName>
    <DbName>structure</DbName>
    (...)
```

XSLT is a XML specification used to transform a XML to another type of document (XML, HTML or text). As an example, the XSLT stylesheet **einfo2html.xsl** available at `https://gist.github.com/lindenb/65e98e5752ea26eb9868` transforms `http://eutils.ncbi.nlm.nih.gov/entrez/eutils/einfo.fcgi` to a HTML document. The ouput would start with:

```
<html><body><ul>
<li><a>pubmed</a></li>
<li><a>protein</a></li>
<li><a>nuccore</a></li>
<li><a>nucleotide</a></li>
<li><a>nucgss</a></li>
<li><a>nucest</a></li>
<li><a>structure</a></li>
<li><a>genome</a></li>
<li><a>assembly</a></li>
    (...)
```

Using the command line XSLT processor **xsltproc**, generate this HTML and visualize it in a web browser.
The **href** attribute is missing in the <a/>anchors. For example, the href for 'pubmed' would be :

```
<html><body><ul>
<li><a href="http://eutils.ncbi.nlm.nih.gov/entrez/
   eutils/einfo.fcgi?db=pubmed">pubmed</a></li>
```

3
```
    (...)
```

Modify the XSLT stylesheet: using the XSLT element <xsl:attribute/>add the missing XML attributes. For example see: `http://stackoverflow.com/questions/3321119`.