

Bioinformatics course 1

Regular expressions (1)

Virginie Orgogozo

14 Oct 2011

- download the example files from

http://practicalcomputing.org/files/pcfb_examples.zip

- download a powerful texteditor

- for Linux: **jEdit** (in Ubuntu: use the Synaptic Package Manager to install jedit. Once the installation is complete, jEdit can be launched by selecting the Programming submenu from the Applications menu)

- for Mac OS X: **TextWrangler** (you may need to drop files onto the program's icon or use the Open Dialog box within textWrangler to open files).

- for Windows: **jEdit** (click on the download link on the jEdit web site and select the Windows Installer for the stable version. Follow the instructions when you launch the installer. You will need to have Java Runtime Environment (JRE) 1.4 or greater installed. Once you've installed jEdit, select Find... from the Search menu and make sure that Regular Expressions is checked).

What is a computer file?

a block of arbitrary information = a linear sequence of binary numbers
available to a computer program
it remains available for programs to use after the current program has finished

What is a text file?

a special type of computer file
where binary numbers correspond to human-readable characters (digits, letters,
space, punctuations). ASCII, UTF-8, UTF-16

What is usually called a binary file?

a computer file that is not a text file
where binary numbers do not correspond to human-readable characters
more compact, faster

Exercise 1

Which ones are text files?

.fasta

.txt

.pdf

.doc

.rtf

.html

.jpeg

.xls

.csv (comma-separated value)

Genbank format file

Example of a Genbank file

```
LOCUS       AF068625                200 bp    mRNA    linear   ROD 06-DEC-1999
DEFINITION Mus musculus DNA cytosine-5 methyltransferase 3A (Dnmt3a) mRNA,
            complete cds.
ACCESSION   AF068625 REGION: 1..200
VERSION     AF068625.2  GI:6449467
KEYWORDS    .
SOURCE      Mus musculus (house mouse)
  ORGANISM  Mus musculus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia;
            Sciurognathi; Muroidea; Muridae; Murinae; Mus.
REFERENCE   1 (bases 1 to 200)
  AUTHORS   Okano,M., Xie,S. and Li,E.
  TITLE     Cloning and characterization of a family of novel mammalian DNA
            (cytosine-5) methyltransferases
  JOURNAL   Nat. Genet. 19 (3), 219-220 (1998)
  PUBMED   9662389
REFERENCE   2 (bases 1 to 200)
  AUTHORS   Xie,S., Okano,M. and Li,E.
  TITLE     Direct Submission
  JOURNAL   Submitted (28-MAY-1998) CVRC, Mass. Gen. Hospital, 149 13th Street,
            Charlestown, MA 02129, USA
REFERENCE   3 (bases 1 to 200)
  AUTHORS   Okano,M., Chijiwa,T., Sasaki,H. and Li,E.
  TITLE     Direct Submission
  JOURNAL   Submitted (04-NOV-1999) CVRC, Mass. Gen. Hospital, 149 13th Street,
            Charlestown, MA 02129, USA
REMARK      Sequence update by submitter
COMMENT     On Nov 18, 1999 this sequence version replaced gi:3327977.
FEATURES             Location/Qualifiers
     source           1..200
                     /organism="Mus musculus"
                     /mol_type="mRNA"
                     /db_xref="taxon:10090"
                     /chromosome="12"
                     /map="4.0 cM"
     gene             1..>200
                     /gene="Dnmt3a"
ORIGIN
  1 gaattccggc ctgctgccgg gccgccgac ccgccggcc acacggcaga gccgcctgaa
  61 gccagcgct gaggctgcac ttttccgagg gcttgacatc agggctctatg ttttaagtctt
 121 agctcttgct tacaagacc acggcaattc cttctctgaa gccctgcag cccacagcg
 181 ccctgcgagc cccagcctgc

//
```

Setting up the text editor

1) use a fixed-width font

(iiii and OOOO should have the same length. `iiii` versus `OOOO`)

2) display line numbers

in Textwrangler (Windows): open a blank document, >Text Options > Show Line Numbers

In jedit (Mac, Linux) : >View > Line numbers

3) use « line feed » for the line ending character (already set up in jedit)

Unix systems including OS X use newlines (`\n`) (LF) to mark line endings in text files.

The old MacOS uses carriage-returns (`\r`) (CR).

Windows uses a carriage-return followed by a newline (`\r\n`).

jEdit can read and write files in all three formats.

In TextWrangler: change to LF by scrolling down in the drop-down menu and saving the file

4) view invisible characters

In Textwrangler: drop-down menu : Show invisibles and Show spaces

Exercise 2

Explore the example files

Open `ctd.rtf` and `ctd.txt` by double-clicking

Open `ctd.rtf` and `ctd.txt` in jedit/Textwrangler

What are the differences?

What is a regular expression?

a concise and flexible means for "matching" (specifying and recognizing) strings of text

very useful to modify text files (replace/delete)

also referred to as regexp, regex or grep

use **wildcards** = special characters that can match more than one particular character

```
Mus musculus  
Agalma elegans  
Frillagalma vityazi  
Cordagalma tottoni
```



```
M. musculus  
A. elegans  
F. vityazi  
C. tottoni
```


What is a regular expression?

a concise and flexible means for "matching" (specifying and recognizing) strings of text

very useful to modify text files (replace/delete)

also referred to as regexp, regex or grep

use **wildcards** = special characters that can match more than one particular character

Setting up the text editor for regular expression searches

in Textwrangler (Windows): > Search > Find : make sure that Grep is checked

in jedit : > Search > Find : make sure that Regular Expressions is checked

```
+40 46 'N +014 15 'E  
+21 17 'N -157 52 'W  
...
```



```
+40 46 ' +014 15 '  
+21 17 ' -157 52 '  
...
```

\w = any letter (A-Z) or digit (0-1) or _

```
+40 46 'N +014 15 'E  
+21 17 'N -157 52 'W  
...
```



```
+40 46 ' +014 15 '  
+21 17 ' -157 52 '  
...
```

\w = any letter (A-Z) or digit (0-1) or _

```
Search for:  
'\w  
Replace by :  
'
```

' or '

Curly quote mark = won't

Data tick mark = 5' -ATGC- 3'

Best = copy from the document and paste into the search/replace

```
5th
3rd
2nd
1rst
..
```



```
5
3
2
1
..
```

`\w` = any letter (A-Z) or digit (0-1) or `_`

Regular expressions are non-overlapping
(`\w\w` would match 5t and then 3r, etc.)

Use () to capture part of the text and put it into the replacement term

```
Search for:
(\w)\w\w
Replace by :
$1
```

in jedit

```
Search for:
(\w)\w\w
Replace by :
\1
```

in Textwrangler

5th
3rd
2nd
1rst
..



Position: 5
Position: 3
Position: 2
Position: 1t
..

```
5th
3rd
2nd
1rst
..
```



```
Position: 5
Position: 3
Position: 2
Position: 1t
..
```

```
Search for:
(\w)\w\w
Replace by :
Position: $1
```

in jedit

```
Search for:
(\w)\w\w
Replace by :
Position: \1
```

in Textwrangler

5th
3rd
2nd
1rst
..



5
3
2
1
..

Use a quantifier: + to match one or more entities

`\w+` = a string composed of (any letter (A-Z) or digit (0-1) or `_`)
up until the next non `/w` character


```
5th
3rd
2nd
1rst
..
```



```
5
3
2
1
..
```

Use a quantifier: + to match one or more entities

`\w+` = a string composed of (any letter (A-Z) or digit (0-1) or `_`) up until the next non `/w` character

```
Search for:
(\w)\w+
Replace by :
$1
```

in jedit

```
Search for:
(\w)\w+
Replace by :
\1
```

in Textwrangler

Mus musculus
Agalma elegans
Frillagalma vityazi
Cordagalma tottoni



M. musculus
A. elegans
F. vityazi
C. tottoni

```
Mus musculus
Agalma elegans
Frillagalma vityazi
Cordagalma tottoni
```



```
M. musculus
A. elegans
F. vityazi
C. tottoni
```

```
Search for:
(\w)\w+ (\w+)
Replace by :
$1. $2
```

in jedit

\$1\$2 \$3 \$1_\$3

```
Search for:
(\w)\w+ (\w+)
Replace by :
\1. \2
```

in Textwrangler

```
Mus musculus
Agalma elegans
Frillagalma vityazi
Cordagalma tottoni
```



```
Mus musculus M_musculus
Agalma elegans A_elegans
Frillagalma vityazi F_vityazi
Cordagalma tottoni C_tottoni
```

```
Mus musculus
Agalma elegans
Frillagalma vityazi
Cordagalma tottoni
```



```
M. musculus
A. elegans
F. vityazi
C. tottoni
```

```
Search for:
(\w)\w+ (\w+)
Replace by :
$1. $2
```

in jedit

```
Search for:
(\w)\w+ (\w+)
Replace by :
\1. \2
```

in Textwrangler

```
Mus musculus
Agalma elegans
Frillagalma vityazi
Cordagalma tottoni
```



```
Mus musculus M_musculus
Agalma elegans A_elegans
Frillagalma vityazi F_vityazi
Cordagalma tottoni C_tottoni
```

```
Search for:
(\w)\w+ (\w+)
Replace by :
$1$2 $3 $1_$3
```

in jedit

```
Search for:
(\w)\w+ (\w+)
Replace by :
\1\2 \3 \1_\3
```

in Textwrangler

Escaping punctuation characters: \

```
Mus musculus (Y456)  
Agalma elegans (AB34)  
Frillagalma vityazi  
Cordagalma tottoni
```



```
M.musculus_Y456  
A.elegans_AB34  
F.vityazi  
C.tottoni
```

```
Search for:  
(\w)\w+ (\w+) \((\w+)\)  
Replace by :  
$1. $2_$3
```

in jedit

```
Search for:  
(\w)\w+ (\w+) \((\w+)\)  
Replace by :  
\1. \2_\3
```

in Textwrangler

\t = a tab character

\s = a white space character (space, tabs, end-of-line, etc.)

\n or \r = end of line

\d = a digit, from 0 to 9

. = any letter, number or symbol except end-of-line character