Text editors

Bioinformatics Applications (PLPTH813)

Sanzhen Liu

1/19/2017

Outline

Goal: to understand how to organize data in a proper format and efficiently input and edit data.

- Formats of text data files
- Excel to generate a text file and tips in Excel
- TextWrangler (Mac) Notepad++ (PC): text editor
- Regular expression
- *vi*: another text editor

Text file – flat file

- **Flat file**

1. Simple format, consisting of readable characters

- ASCII (American Standard Code for Information Interchange, 128 characters)
- No rich format control (e.g. bold or Italics, etc)

2. Easy for sharing

- **The organization of data in a text file**

1. Most popular formats for tabular data: space or tab separated data file (.txt) and comma-separated values (.csv)
2. Most popular format for DNA sequences: FASTA format (.fa, .fas, .fasta)

File formats

- Tab separated file (.txt)

```
name    age >30?    gender
Josh    23  FALSE   male
Rose    35  TRUE    female
```

- Comma-separated file (.csv)

```
name,age,>30?,gender
Josh,23,FALSE,male
Rose,35,TRUE,female
```

- FASTA (.fa, .fas, .fasta)

```
>Aa1
CCATCTCATCCCTGCGTGTCTCCGACTCAG
>Aa2
CTGAGTCGGAGACACGCAGGGATGAGATGGTT
```

Text editors

- Notepad or Notepad++ (PC)
- TextEdit (Mac)
- TextWrangler (Mac)
- vi (Unix and Linux)
- Emacs

- *Excel (PC and Mac): save as …*
- etc

Newline – end of line (EOL)

Two types of EOL: line feed (LF) and carriage return (CR):

LF: \n
CR: \r

- LF: Unix, Linux, OS X
- CR: Mac OS up to version 9 and OS-9
- CR+LF: Microsoft Windows

http://en.wikipedia.org/wiki/Newline

Excel to generate a text file

name	age
Josh	23
Rose	35
Jone	18
Molly	21
Lisa	36

- copy and paste to a text editor (e.g. vi)
- save as …

Excel function - examples

Q1: =**AVERAGE**(B3:B7)

Q2: =**COUNTIF**(B3:B7, ">20")

Q3: =B3>30

Q4: search information at Table 2

1. define the Table 2: gender (control + l)

2. =**VLOOKUP**(A3, gender, 2, FALSE)

	A	B	C	D
1	Table 1			
2	**name**	**age**	**>30?**	**gender**
3	Josh	23		
4	Rose	35	Q3	Q4
5	Jone	18		
6	Molly	21		
7	Lisa	36		
8				
9	Table 2			
10	**name**	**gender**		
11	Josh	male		
12	Rose	female		
13	Jone	male		
14	Molly	female		
15	Lisa	female		
16				
17	Question:			
18	average age	Q1		
19	# of persons >20	Q2		

Table 1

name	**age**	**>30?**	**gender**
Josh	23	FALSE	male
Rose	35	TRUE	female
Jone	18	FALSE	male
Molly	21	FALSE	female
Lisa	36	TRUE	female

Table 2

name	**gender**
Josh	male
Rose	female
Jone	male
Molly	female
Lisa	female

Question:

average age	26.6
# of persons >20	4

Useful functions in Excel

- max/min/average/sum
- len/left/right
- if/countif
- >, <, =
- & (concatenate)
- vlookup

- LEFT function Returns the leftmost characters from a text value

	A	B
1	this class is boring	=LEFT(A1, 14)&"great!"

Functions can be combined.

Replacement

Replace the words containing “genome” with “XXX” regardless of letter case.

Genome old and new charted the emergence of agriculture. Contemporary Europeans carry DNA inherited from light-skinned, brown-eyed farmers who migrated from the Middle East beginning 7,000–8,000 years ago, in addition to more-ancient ancestry. The achievements of these early farmers — domestication of crops such as wheat and barley — are also being understood through genome sequencing.

Which software and what trick will you use?

Problem

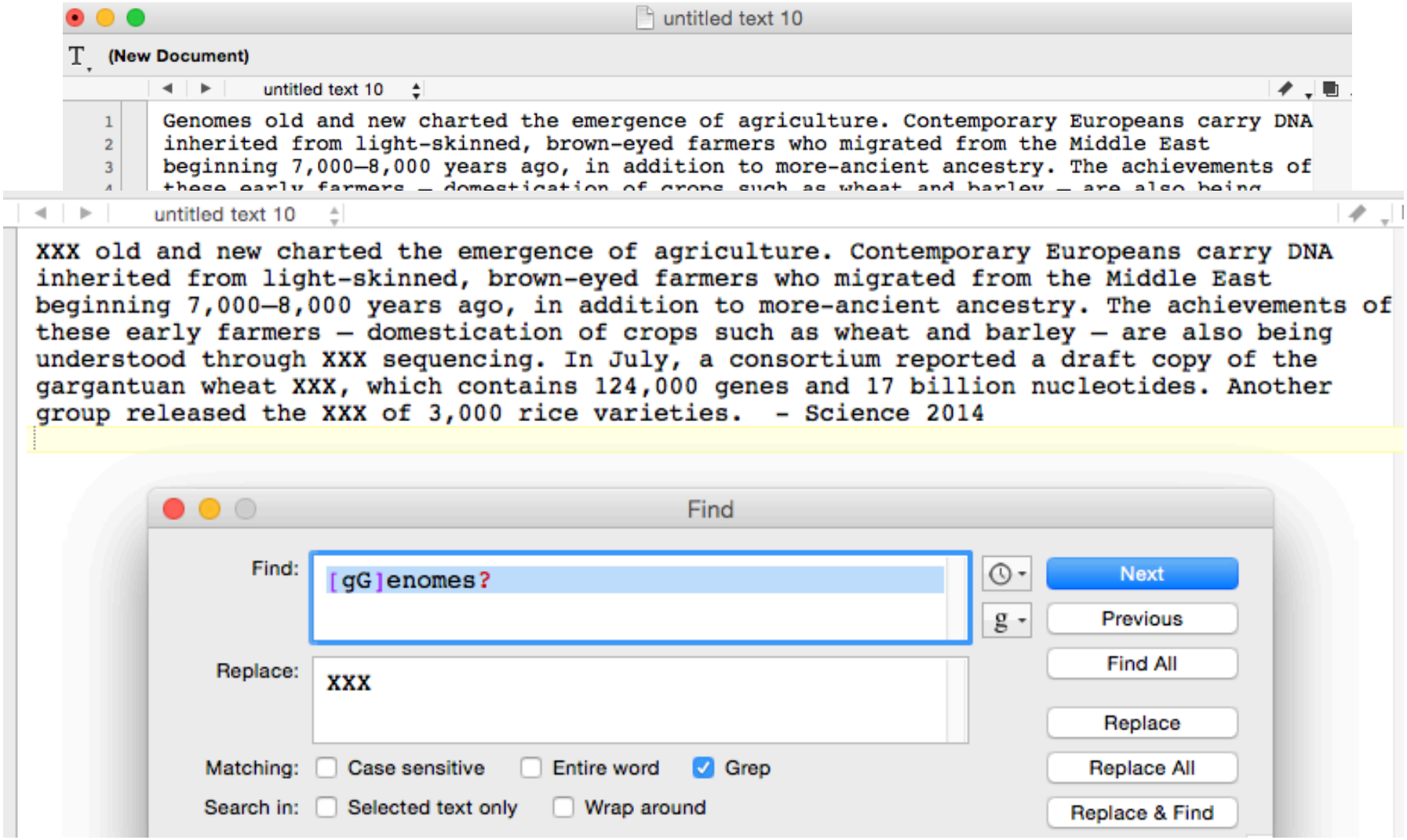
Replace the words containing “genome” with “XXX” regardless of letter case (e.g., Genome = genome = genomes = Genomes).

Genomes old and new charted the emergence of agriculture. Contemporary Europeans carry DNA inherited from light-skinned, brown-eyed farmers who migrated from the Middle East beginning 7,000–8,000 years ago, in addition to more-ancient ancestry. The achievements of these early farmers — domestication of crops such as wheat and barley — are also being understood through genome sequencing. In July, a consortium reported a draft copy of the gargantuan wheat genome, which contains 124,000 genes and 17 billion nucleotides. Another group released the genomes of 3,000 rice varieties. - Science 2014

Which software and what trick will you use?

TextWrangler

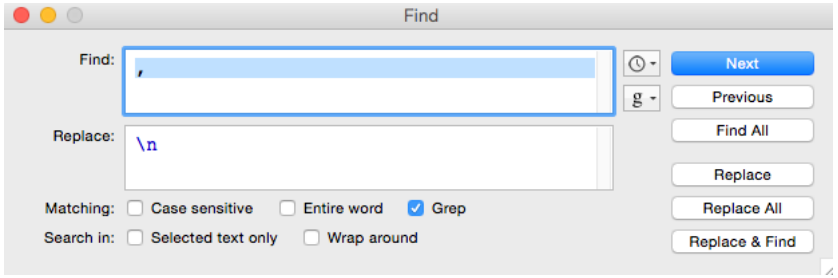
A flexible text editor with powerful functions of searching and editing.



TextWrangler – more examples

Class participation 15%, Homework 15%, Midterm Exam 20%, Project 20%, Final Exam 30%

Class participation 15%
Homework 15%
Midterm Exam 20%
Project 20%
Final Exam 30%

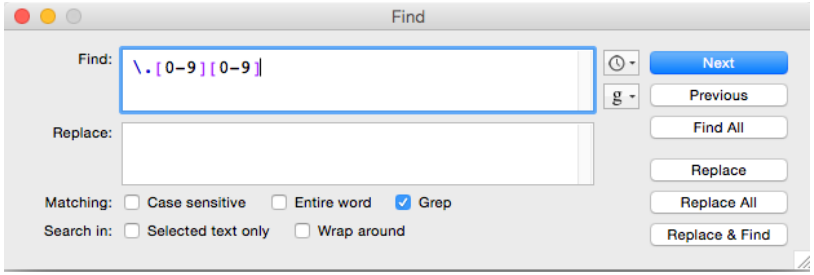


\n: end of line character (line separator)

Class participation 15.01%, Homework 15.03%, Midterm Exam 20.10%, Project 20.10%, Final Exam 30.01%

\.[0-9][0-9]

\.: the character of "."
. : any character



Regular expression

- **Regular expression** (regex or regexp) is a sequence of characters that forms a search pattern.

Search genome or genomes:

[gG]enomes?

[] : a single character of a range indicated in the square brackets
?: no matches or just one match

More regex characters

Wildcards	
`\w`	Letters, numbers and _
`.`	Any character except `\n` `\r`
`\d`	Numerical digits
`\t`	Tab
`\r`	Return character. Also used as the generic end-of-line character in TextWrangler
`\n`	Line-feed character. Also used as the generic end-of-line character in Notepad++
`\s`	Space, tab, or end of line
`[A-Z]`	A single character of the ranges indicated in square brackets
`[^A-Z]`	A single character including all characters *not* in the brackets. Note that this will include `\n` unless otherwise specified, and may cause you to match across lines
`\`	Used to escape punctuation characters so they are searched for as themselves, not interpreted as wildcards or special symbols
`\\`	The \ symbol itself, escaped
Boundaries	
`^`	Match the start of the line, i.e., the position before the first character
`$`	Match the last position before the end-of-line character

Regular expression (I)

\t : a tab character

\r (or \n): end-of-line

Potato,apple,orange

Regexp	Replace
,	\t

Potato apple orange

Potato apple orange

Regexp	Replace
\t	\n

Potato
apple
orange

Regular expression (II)

- **^** beginnings
- **$** endings

Potato
apple
orange

Regexp	Replace
^	-

-Potato
-apple
-orange

Potato
apple
orange

Regexp	Replace
$	s

Potatos
apples
oranges

Regular expression (II)

- **\w** a word character, including letters, numbers and underscore
- **\d** : numerical digits

I have 5 apples.

Regexp	Replace
^\w	We

We have 5 apples.

I have 5 apples.

Regexp	Replace
\d	a lot of

I have a lot of apples.

Regular expression (III)

+ : 1 or more previous regular expression

? : 0 or 1 previous regular expression

. : any character except \n \r

potato,apple,orange

Regexp	Replace
p+	_

-otato,a-le,orange

potato,apple,orange

Regexp	Replace
p?	_

--o-t-a-t-o-,-a---l-e-,-o-r-a-n-g-e

potato,apple,orange

Regexp	Replace
p.	_

-tato,a-le,orange

Regular expression (IV)

[A-Z] : any single letter

NspI

[AG]CATG[CT]

select 2012, 2013, 2014 201[2-4]

{} : specify a range of numbers to repeat the match of the immediately preceding character.

Poly A (12 A in a row) A{12}

Poly A (10-12 A in a row) A{10,12}

Poly A (>=10 A in a row) A{10,}

vi

- *vi* is a text editor created for the Unix operating system.

- *vi* has two modes:
  1. insert mode (edit as other text editors)
  2. command mode (commands that control the edit session).

  switch modes by using “i” and “ESC” key

Your keyboard controls “everything”.

Goal of today's lab

- Familiar to Excel functions
- Try *vi* at Beocat
- Practice using regular expression in TextWrangler