Regular expressions in egrep

pattern	**matches lines that…**	**sample strings**
“ab”	have *ab*	r<u>ab</u>bit
“**^**ab”	**begin** with *ab*	<u>ab</u>out
“ab**$**”	**end** with *ab*	c<u>ab</u>
“**(**ab**|**ba**)**”	have *ab* **or** *ba*	<u>ba</u>t, j<u>ab</u>
“**(**aa|hh|xx**)**”	have *aa*, *hh* **or** *xx*	baz<u>aa</u>r, E<u>xx</u>on
“**[**XxY**]**”	have *X*, *x* or *Y*	LaTe<u>X</u>, ta<u>x</u>, <u>Y</u>ork
“**^[^**aeiou**]**o**$**”	2-letter words that start with a consonant end with 'o'	do, go, so

Regular expressions in egrep

pattern	matches lines that…	sample strings
"a**+**"	have a **sequence** of *a*'s	cat, aardvark
"a**{2}**"	have a **sequence** of 2 *a*'s	aardvark
"**(**ba**)+**" h	have a **sequence** of *ba*'s	bat, alibaba
"**.**"	have **any** character	
"**^**a**.***a**$**"	**begins** and **ends** with *a*	aa, ana, alpha
"**\.**"	have a **literal** dot	uplb**.**edu**.**ph

Note on "quantifier" meta-symbols

***** 0 or more

+ 1 or more

? 0 or 1 (i.e., optional)

{*n***}** exactly n

{*n***,}** n or more

{*n***,***m***}** between n and m

Word games using egrep

Find all words in /usr/share/dict/words that:

- have a dash or digits in them
- start with a capital vowel
- are all-cap acronyms/words
- start with ‘a’ and end with ‘x’ and length>2
- have exactly 20 letters
- have a double-‘x’ or a double-‘u’
- have a ‘Q’ or ‘q’, but not followed by a ‘u’
- have all the 5 vowels in alphabetical order
- have 2 or more occurrences of ‘ab’
- have a sequence of 4 or more vowels
- have 2 or more double-vowels (aa, ee, ii, oo, uu)

Some uses of egrep

- Find all lines that begin a loop in a C program

  egrep **“while|do|for”** myprogram.c

- Find all lines that contain an email address

  egrep **“[A-Za-z]@[A-Za-z]”** textfile

- Find all lines that contain a floating point number

  egrep **“[0-9]+\.[0-9]+”** myprogram.c

- Find all occurences of some unusual code in several programs for purposes of code-plagiarism-detection

  egrep **“foobar\(.*\)”** cs11progassign*.c

- Find all occurrences of some DNA/protein subsequence pattern that may be biologically significant (e.g., “sequence motifs” that may be indicators of a particular gene)

  egrep **“ATCG(A|C){2,3}AT”** seqdata.*

Programming project

- Write a program that implements substring search, incorporating the most basic features of egrep

  egrep “pattern” file

  – basic sequence, e.g., egrep “abc” file
  – allow alternates, e.g., egrep “ab|ba” file
  – allow Kleene plus, e.g., egrep “(ab)+” file

Links and further reading

- egrep is a standard utility in UNIX/Linux. See the man page (e.g., http://www.mediacollege.com/cgi-bin/man/page.cgi?topic=egrep). For Windows, try installing **cygwin-** **www.cygwin.com**
- Algorithms for string matching (focus on the fast Boyer-Moore algorithm) for **fgrep** where patterns are fixed strings

  http://en.wikipedia.org/wiki/Boyer%E2%80%93Moore_string_search_algorithm

- Other approaches to plagiarism-detection

  http://www.cs.berkeley.edu/~benr/publications/auscc04/papers/burrows-auscc04.pdf

- Sequence motifs in bioinformatics

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

- jmsamaniego@uplb.edu.ph (revised 2008)