上海交通大学 2015-2016 学年第二学期

《Linux 操作系统》小测 2：正则表达式

题号	一	二	三	总分	阅卷人
得分					

学号:_____ 姓名:_____ 班级:_____ 系:_____

装 订 线 内 不 要 答 题

一、填空题： (共 20 小题，每小题 2 分，共 30 分)

1. 正则表达式中. 表示的是_____，而? 表示的则是_____。

2. 正则表达式主要分为基本正则表达式、_____、_____。

3. 用于表示锚定 (anchoring) 的元字符 (metacharacter) 中，^ 表示的是_____，$ 表示的是_____，而\b 表示的是_____。

4. 如果要匹配中国的一个固定电话号码，正则表达式应该写成_____，而匹配中国移动手机号码的正则表达式应该为_____。

5. 如果将元字符写成对应的字符类(character set)，\w 对应的是_____，而\d 对应的是_____。

6. 一个 IPv4 的 IP 地址用正则表达式可以表示为_____。

7. 测序结果中知道某个位置为嘌呤，用正则表达式表示为_____，如果是嘧啶，则可以表示为_____。

8. 一个碱性氨基酸用正则表达式表示为_____，而极性氨基酸用正则表达式表示为_____。

9. grep 命令默认使用的是基本正则表达式，如果需要使用扩展正则表达式，需用到选项_____；而对于 Perl 类的正则表达式，需用到选项_____。

10. awk 中 $0 表示的是_____，而 $NF 表示的是_____。

11. awk 的语法中包含两个特殊的模式，分别为_____和_____。

12. 如 sed 需要使用到扩展正则表达式，需使用选项_____；sed 能否使用 Perl 正则表达式？_____。

13. 在网站注册的时候，常常需要你选择你个人的密码，如果密码需要大小写字母、数值、特殊符号的组合，且密码长度最小为 8，这时候如果用正则表达式匹配，应该写成_____。

14. 匹配一个整数（正负皆可）的正则表达式是_____。

15. awk 的 match() 函数返回的是_____，此外 RSTART 和 RLENGTH 分别表示_____和_____。

二、选择题： (共 20 小题，每小题 2 分，共 40 分)

1. 下面哪个正则表达式匹配的是不含 DEC 的单词？

A. `\b((?!DEC)\w)+\b`

B. `\b((?^DEC)\w+)\b`

C. `\b((?>DEC)\w)+\b`

D. `\b((?=DEC)\w)+\b`

2. sed 命令进行全局替换命令的选项是哪个？

A. a　　B. A　　C. g　　D. G

3. sed 命令需要关闭默认输出的选项是哪个？

A. -e　　B. -r　　C. -n　　D. -i

4. 下面哪些正则表达式可以用来匹配偶数个数的连续"x" 字符串？

A. `^(xx*)\1$`

B. `^((xx{1,})\1+$)`

C. `^(xx?)\1+?`

D. `^(x+)\1$`

5. 下面哪些正则表达式可以正确匹配非质数个数的连续"x" 字串？

A. `^(xx+)\1+$`

B. `^(xx*)\1+$`

C. `^(xx?)\1+$`

D. `^(xx+)\1$`

6. sed 是什么的缩写？

A. stream editor　　B. scripting editor　　C. scientific editor　　D. super editor

7. 下面哪个不是 awk 的内置函数？

A. print　　B. printf　　C. write　　D. length

装 订 线 内 不 要 答 题

8. 下面哪个变量指定了 awk 的列分隔符？

A. FS　　B. RS　　C. NR　　D. FNR

9. 下面关于 awk 的陈述，哪些是正确的？

A. awk 一般用来处理分析表格类型的文件　　B. awk 只能处理单个文件

C. awk 不能调用 bash 的环境变量　　D. awk 不支持多维数组

10. 下面哪些 sed 命令可以将文件按行倒序输出？(假设采用了 -n 选项)

A. `1!G;$!h;$p`

B. `1!g;$!h;$p`

C. `1!G;h;$p`

D. `G；h;$p`

11. 下面哪些 sed 命令只输出文件的奇数行？(假设采用了 -n 选项)

A. `N;p`

B. `n;p`

C. `N;P`

D. `n;P`

12. 下面哪个命令可以输出文件的偶数行？

A. `'NR%2{print}'`

B. `'NR%2==0{print}'`

C. `'NR%2==0{print $0}'`

D. `'NR/2==0{print}'`

13. 下面哪个 bash 命令可以输出当前系统下每个用户的所有进程的内存消耗？

A. `ps aux | awk 'NR>1{a[$1]+=$6} END{for (i in a){print i, a[i];}}'`

B. `ps aux | awk 'NR<>1{a[$1]+=$6} END{for (i in a){print i, a[i];}}'`

C. `ps aux | awk 'NR>1{a[$1]+=$6} END{print a;}}'`

D. `ps aux | awk 'NR>1{a[$1]+=$6} END{for (i in a){print i;}}'`

14. 下面哪个命令可以统计当前系统下默认使用不同 shell 的用户数量？

A. `cat /etc/passwd | awk '{usern[$7]+=1} END{for (i in used) {print i, used[i];}}'`

B. `cat /etc/passwd | awk 'BEGIN{FS=":"} {usern[$7]+=1} END{for (i in used) {print i, used[`

C. `cat /etc/passwd | awk 'BEGIN{IFS=":"} {usern[$7]+=1} END{for (i in used) {print i, used`

学号:_____ 姓名:_____ 班级:_____ 系:_____

装 订 线 内 不 要 答 题

D. `cat /etc/passwd | cut -d: -f7 | uniq -c`

15. 正则表达式`^\d*\*[^\d]*[\w]{6}$`, 下面的字符串中哪个能正确匹配？

A. ***abcABCD_89　B. abc*abcABCDEF　C. 123*abcABCD_89　D. 123*ABCabcd-89

16. 正则表达式 (01|10|1001|0110)* 与下列哪个表达式一样？

A. (0|1)*　B. (01|01)*　C. (01|10)*　D. (11|01)*

17. 正则表达式 A*B 可以匹配哪些选项？

A. A　B. ACB　C. B　D. AB

18. 下面哪些选项与正则表达式 x|(yx+) 不匹配？

A. x　B. xyxx　C. yx　D. yxxxx

19. 以下哪个字符串不能被正则表达式 a(bc?)d 匹配？

A. abcd　B. abd　C. abc　D. acd

20. 下列正则表达式不可以匹配网址 www.huawei-inc.com 的是？

A. `^\w+\.\w+\-\w+\.\w+$`

B. `[w]{0,3}.[a-z\-]*.[a-z]+`

C. `[c-w.]{3,10}[.][c-w.][.][a]`

D. `[w][w][w][Huawei-inc]+[com]+^\w.*com$`

E. `[w]{3}.[a-z\-]{11}.[a-z]{3}`

三、解答题：　(共 10 小题, 每小题 3 分, 共 30 分)

1. 什么是正则表达式？举例说明，并谈谈你对正则表达式的看法。

2. 正则表达式都有哪些元字符（metacharacter）？这些元字符分别的作用是什么？有些元字符与 bash 命令行下的通配符相同，请问是哪些？其意义有何区别？

3. 你能否用正则表达式写出匹配一个基因的模式？假设该基因是连续的，不存在真核基因的外显子 -内含子结构。

4. 如何从 GENBANK 文件中抽取所有的 CDS 序列？写出你的思路，目前不需要你去实现，但正确的正则表达式是必须的。

5. 命名捕获组（naming captured group）与非命名捕获组相比，有何优缺点？

6. 说说什么是零宽断言（zero-length assertion）。你觉得零宽断言有哪些分类，其主要的应

用有哪些？

7. 什么是贪婪匹配？其和懒惰匹配有什么区别，你觉得其在生物信息学分析中有何应用？

8. 写一个 sed 脚本，将 fastq 文件转化为 fasta 文件。

9. 说说 sed 的模式空间（pattern space）和保留空间（hold space）的作用以及相关的命令如 G/g、H/h、D/d 的作用，并举例说明。

10. 谈谈 sed 的 b、t/T 的用法及其应用。

11. 写一个正则表达式，匹配下表中左边的所有字符串，但不匹配右边的所有字符串。你可以将其拆分到两个单独的文件，看看你写的正则表达式是否能够匹配其中一个文件的所有行，而不能匹配另一个文件的所有行。

abac	beam
accede	buoy
adead	canjac
babe	chymia
bead	corah
bebed	cupula
bedad	griece
bedded	hafter
bedead	idic
bedeaf	lucy
caba	martyr
caffa	matron
dace	messrs
dade	mucose
daff	relose
dead	sonly
deed	tegua
deface	threap
faded	towned

faff	widish
feed	yite

12. 同样，用最简洁的正则表达式匹配下表中左边所有的字串，但同时不能匹配右边所有的字串。

Mick	Kickapoo
Rick	Nickneven
allocochick	Rickettsiales
backtrick	billsticker
bestick	borickite
candlestick	chickell
counterprick	fickleness
heartsick	finickily
lampwick	kilbrickenite
lick	lickpenny
lungsick	mispickel
potstick	quickfoot
quick	quickhatch
rampick	ricksha
rebrick	rollicking
relick	slapsticky
seasick	snickdrawing
slick	sunstricken
tick	tricklingly
unsick	unlicked
upstick	unnickeled