

实用生物信息技术期中考试（一班 B卷）

北京大学生命科学学院 2008年4月21日

姓名罗静初 座位号PKU08S1_____ 得分_____

1. 写出实现以下操作的UNIX命令: (10分)

- 1) 统计当前目录下所有以HBA_为起始文件名的文件数

```
ls -l HBA_*. * | wc
```

- 2) 删除子目录test及其所有子目录和文件

```
rm -rf test
```

- 3) 在当前目录下创建子目录hba, 并在该子目录下创建子目录tree

```
mkdir hba; mkdir hba/tree
```

- 4) 将当前目录下文件7hba.fasta移到子目录hba下的子目录tree中

```
mv 7hba.fasta hba/tree
```

- 5) 提取棉花转录因子mRNA序列文件ghtf-mrna.fasta中所有注释行, 并保存到文件ghtf-mrna.lis中

```
grep ">" ghtf-mrna.fasta >ghtf-mrna.lis
```

- 6) 显示文件ghtf-mrna.lis中第11-20行内容, 并按字母表顺序排列

```
head -n 20 ghtf-mrna.lis | tail -n 10 | sort; sed -n '11-20p' ghtf-mrna.fasta | sort
```

- 7) 用Phylip程序包邻接法构建个15个SBP转录因子DNA结合结构域蛋白质序列(15atsbpd.fasta)系统发育树

```
clustalw; seqboot; protdist; neighbor; consense
```

- 8) 比较当前目录下两个蛋白质序列hba-ansan.fas和hba-ansin.fas的差异残基个数

```
diffseq hba-ansan.fas hba-ansin.fas
```

- 9) 构建棉花氨基酸序列 (ghtf-pep.fasta) BLAST数据库, 以拟南芥转录因子编码序列atsbp3-cds.fasta 搜索期望值E<0.001的相似序列

```
formatdb -i ghtf-pep.fasta
```

```
blast -p blastx -d ghtf-pep.fasta -i atsbp3-cds.fasta -o atsbp3-cds.out -e 0.001
```

- 10) 将子目录seq下所有以.fas结尾的文件设置为只读

```
chmod -w *.fas
```

2. 举例说明以下EMBOSS程序的用途和具体用法。

(10分)

1) `seqretsplit 7hba.fasta`

将多序列文件7hba.fasta分为单个序列文件，并用注释信息中“>”后序列名作为文件名

2) `wordcount ay274119.fasta -word 8`

以8bp为单位，统计第一个被鉴定的SARS冠状病毒基因组TOR2序列(ay274119.fasta)中不同碱基组分

3) `sixpack y12618.fasta y12618.sixpack -outseq y12618.orf`

显示豌豆开花相关基因mRNA 序列ay12618.fasta可能的6个读码框，结果保存于文件ay12618.sixpack中，读码框翻译后得到的氨基酸序列保存于文件ay12618.orf中

4) `seealso dottup`

检索EMBOSS软件包中与dottup相关的其它程序名

5) `cpGREport af164138.fasta`

按滑动窗口计算河豚鱼基因组序列片段af164138.fasta中CG含量，推测是否有CpG岛

6) `pepinfo ppf1-pea.fasta`

以图形方式豌豆开花相关基因蛋白质序列ppf1-pea.fasta序列特征，包括残基侧链大小、亲水疏水性、极性、电荷性等

7) `extractalign 15sbpd.aln`

从15个拟南芥SBP转录因子DNA结合结构域多序列比对结果文件15sbpd.aln中提取相关信息，以表格方式输出

8) `etandem hh7tetra.fas -min 6 -max 14`

寻找人7型疱疹病毒端粒重复区串联重复序列片段，最小重复序列长度设定为6BP，最大重复序列长度设定为14bp

9) `dotmatcher slit-drome.fas slit-drome.fas -windowsize 25 -threshold 25`

以点阵图方式显示果蝇体节发育基因蛋白质序列slit-drome.fas中的重复片段

10) `hmoment ppf1-pea.fas -plot`

以图形方式显示豌豆开花相关基因序列ppf1-pea.fas疏水性特征

3. 从植物转录因子数据库网站PlantTFDB (<http://planttfdb.cbi.pku.edu.cn/>) 下载棉花转录因子氨基酸序列数据 (Gossypium_hirsutum(upland_cotton).TF.pep), 构建BLAST数据库, 以拟南芥转录因子 (AT2G33810) 编码序列搜索期望值E<0.001的相似序列, 将结果填入表中。 (10分)

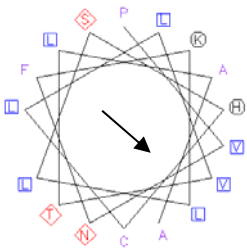
Subject id	Id (%)	AL	Mis	Qs	Qe	Ss	Se	E-value	Score
PTGh01181.1	68	93	26	127	393	40	132	3.00E-34	135
PTGh01189.1	63	93	30	127	393	31	123	2.00E-31	125
PTGh01192.1	63	87	32	133	393	124	210	4.00E-31	125
PTGh01183.1	62	92	34	121	393	175	266	9.00E-31	124
PTGh01188.1	68	78	25	160	393	42	119	3.00E-30	122
PTGh01182.1	67	78	26	160	393	35	112	1.00E-29	120
PTGh01180.1	72	74	21	160	381	91	164	2.00E-29	119
PTGh01193.1	69	74	23	160	381	25	98	2.00E-29	119
PTGh01184.1	70	74	22	160	381	81	154	6.00E-29	117
PTGh01186.1	68	74	24	160	381	82	155	2.00E-28	115
PTGh01191.1	63	80	30	142	381	59	138	1.00E-27	113
PTGh01185.1	61	79	31	157	393	94	172	5.00E-27	111
PTGh01187.1	61	56	22	160	327	204	259	1.00E-18	84
PTGh01190.1	59	49	20	160	306	71	119	2.00E-15	73

Id: % Identity, AL: Alignment Length, Mis: Mismatches, Qs: Query start, Qe: Query end, Ss: Subject start, Se: Subject end

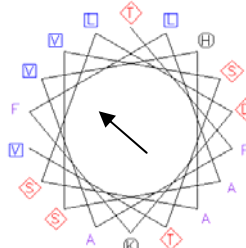
4. 用EMBOSS软件包中蛋白质二级结构预测程序garnier预测人血红蛋白 α -亚基(HBA_HUMAN)中 α -螺旋含量, 与该序列条目注释比较, 将结果填入下表, 并说明预测精度; 用 α -螺旋论显示程序pepwheel显示C'端两个 α -螺旋侧链残基分布并画图表示, 推测它们可能的空间堆积方式。 (10分)

No	1	2	3	4	5	6	7
Garnier	5-17	24-35	53-56	64-72	75-91	96-113	115-134
Annotation	4-35	37-42	53-71	73-79	81-89	96-112	119-136

PEPWHEEL of HBA-HUMAN from 96 to 112



PEPWHEEL of HBA-HUMAN from 119 to 136



如图中箭头所示, 这两个螺旋均有疏水面, 可能在空间结构中形成疏水内核。

5. 用EMBOSS程序包TMAP程序和ExPASy服务器TMHMM和TMPred分别预测豌豆开花相关基因（PPF1_PEA）可能的跨膜螺旋，说明三种方法预测结果差异。（10分）

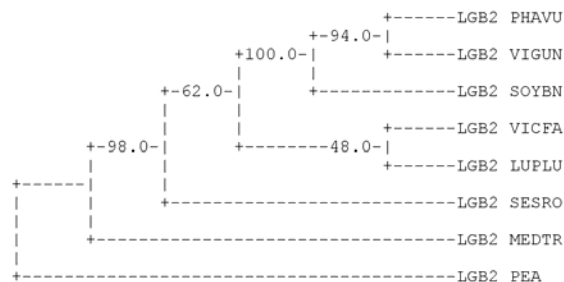
No	0A	1	2	2A	3	4
TMAP		111-135	173-201		251-279	290-315
TMHMM		112-134	179-201		250-272	292-314
TMPred	1-19	114-132	185-201	211-231	256-272	296-316

结果表明，三种方法均预测到四个跨膜螺旋(1-4)，TMAP和TMHMM预测到的位置和长度接近，而TMPred预测位置和长度差别较大。TMPred预测到另外2个跨膜螺旋，其可信度分值均较低。

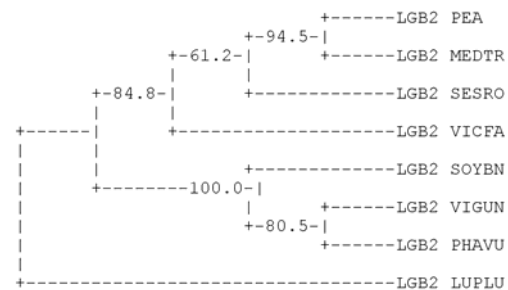
6. 以豆血红蛋白(P02240)为检测序列进行BLAST搜索，找出Swiss-Prot数据库中8个植物2型血红蛋白（Leghemoglobin 2），将搜索结果填入下表。（10分）

Name	Length	Score	E-Value	Identity(%)	Positive(%)	Gap(%)
LGB2_SESRO	148	165	3e-46	58	73	2
LGB2_MEDTR	146	158	3e-44	59	75	4
LGB2_LUPLU	154	302	2e-87	100	100	0
LGB2_PEA	147	149	2e-41	55	72	3
LGB2_SOYBN	145	159	2e-44	58	71	3
LGB2_VICFA	148	154	8e-43	55	67	1
LGB2_PHAVU	146	146	1e-40	54	68	3
LGB2_VIGUN	145	137	6e-38	52	68	3

7. 分别用Phylip软件包中邻接法（Neighbor-joining）和最大简约法（Maximum-parsimony）对上表中8个豆血红蛋白序列构建系统发育树，画出两种树的拓扑结构。（10分）



Neighbor-joining tree



Maximum-parsimony tree

8. 简述你正在研究或最感兴趣的基因或蛋白质并从NCBI、EBI和ExPASy中获取有关信息。

(10分)

基因/蛋白质名称	
研究目的	
进展情况	
物种来源和分类学地位	
生物学功能和表达组织特异性	
PubMed编号(1-5篇)	
PMC编号(1-3篇)	
综述编号(1-2篇)	
GenBank/EMBL编号	
Swiss/TrEMBL编号	
交叉链接数据库	
序列特征	

9. 期中小结

1) 简述已经基本掌握的常用方法和软件工具

2) 简述希望掌握的常用方法和软件工具

3) 学习本课程的收获、体会和遇到的困难

4) 对改进本课程教学的意见和建议