

# 实用生物信息技术期中考试（一班 A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Linux/EMBOSS命令:

(10分)

1) 在当前目录下创建子目录pku08s1, 并在该子目录下创建子目录hba

```
mkdir pku08s1; mkdir pku08s1/hba
```

2) 将当前目录下文件hba\_human.fasta移到子目录pku08s1下的子目录hba中

```
mv hba_human.fasta pku08s1/hba
```

3) 统计当前目录下所有以“.fasta”为文件名结尾的文件数

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

4) 提取玉米转录因子蛋白质序列文件zmtf-pep.fasta中所有注释行, 并保存为zmtf-pep.list

```
grep ">" zmtf-pep.fasta > zmtf-pep.list
```

5) 显示文件zmtf-pep.list中第101-120行的内容, 并按字母表顺序排列

```
head -n 120 zmtf-pep.list | tail -n 20 | sort
```

```
sed -n '101,120p' zmtf-pep.list | sort
```

6) 删除子目录temp下所有子目录和文件, 保留该目录

```
rm temp/*.*
```

7) 将当前目录下ppf1.fas设置为只读文件

```
chmod -w ppf1.fas
```

8) 比较当前目录下文件seq1和seq2的差异

```
diff seq1 seq2
```

9) 构建棉花mRNA序列(ghtf-mrna.fasta)BLAST数据库, 以拟南芥转录因子DNA结合结构域atsbpd3.fasta蛋白质序列搜索期望值 $E < 0.001$ 的相似序列

```
formatdb -I ghtf-mrna.fasta
```

```
blast -p tblastn -d ghtf-mrna.fasta -i atsbpd3.fasta -o atsbpd3-tblastn.out -e 0.001
```

10) 用Phylip程序包邻接法构建20个SRAS冠状病毒特征序列(20sars.fasta)系统发育树

```
Clustwal; seqboot; protdist; neighbor; consense
```

2. 举例说明以下EMBOSS程序的用途和具体用法。

(10分)

1) `seqret hba-human.sw hba-human.fasta`

从Swiss-Prot格式序列条目hba-human.sw中提取序列并以FASTA格式保存到文件hba-human.fasta中

2) `coderet y12618.gb`

从GenBank格式序列条目y12618.gb中根据序列特征表注释信息提取编码区和苏编码氨基酸序列,并以FASTA格式保存

3) `getorf y12618.fasta -min 1000`

从FASTA格式序列条目y12618.fasta中提取开发读码框,并翻译成氨基酸序列,最小读码框长度为1000bp

4) `wossname protein`

以关键词protein检索EMBOSS软件包中所有可用于蛋白质序列分析的程序

5) `compseq af164138_cds_1.fasta -word 3 -frame 1`

计算河豚鱼多药耐药基因编码序列af164138\_cds\_1.fasta密码子使用情况

6) `pepwheel p69905.fas -sbegin 96 -send 112`

用alpha-螺旋论显示人血红蛋白alpha亚基第96-112位残基侧链在alpha螺旋上分布情况

7) `showalign 7hba.aln`

以各种不同格式输出多序列比对结果,以便用于论文图表

8) `remap y12618.fasta`

显示DNA序列y12618.fasta的限制性内切酶图谱

9) `einverted af164138.fasta`

显示DNA序列af164138.fasta反向互补序列片段

10) `diffseq hba-ansan.fasta hba-ansin.fasta`

显示斑头雁和灰雁血红蛋白alpha亚基序列差异

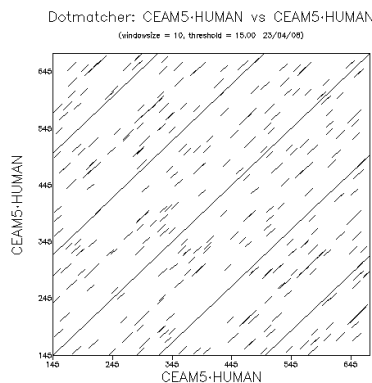
3. 从植物转录因子数据库网站（<http://planttfdb.cbi.pku.edu.cn/>）下载棉花转录因子 mRNA 序列数据（Gossypium\_hirsutum(upland\_cotton).TF.mRNA），构建BLAST数据库，以拟南芥转录因子DNA结合结构域atsbpd3.fasta蛋白质序列搜索相似序列，将结果填入表中。（10分）

Subject id	Id (%)	AL	Mis	Qs	Qe	Ss	Se	E-value	Score
PTGh01181.1	76	78	19	2	79	241	474	8E-34	133
PTGh01189.1	71	78	23	2	79	137	370	6E-31	124
PTGh01183.1	68	78	25	2	79	614	847	1E-30	122
PTGh01188.1	68	78	25	2	79	122	355	2E-30	122
PTGh01192.1	68	78	25	2	79	399	632	2E-30	122
PTGh01182.1	67	78	26	2	79	115	348	7E-30	120
PTGh01180.1	72	74	21	2	75	342	563	2E-29	119
PTGh01193.1	69	74	23	2	75	424	645	2E-29	119
PTGh01184.1	70	74	22	2	75	442	663	5E-29	117
PTGh01187.1	68	74	24	2	75	718	939	6E-29	117
PTGh01186.1	68	74	24	2	75	258	479	2E-28	115
PTGh01191.1	66	74	25	2	75	708	929	2E-27	112
PTGh01185.1	61	79	31	1	79	484	720	3E-27	111
PTGh01190.1	59	49	20	2	50	210	356	1E-15	73

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

4. 用needle程序分析人癌胚抗原5型粘连蛋白（CEAM5\_HUMAN）结构域Ig-like 2和Ig-like 4，结构域Ig-like 1和小鼠癌胚抗原1型粘连蛋白（CEAM1\_MOUSE）结构域Ig-like V type序列相似性，将结果填入下表。用dotmatcher程序分析CEAM5\_HUMAN第146到677位残基六个免疫球蛋白结构域之间的序列相似性，说明分析结果。（10分）

Domains	Length	Score	Identity (%)	Similarity (%)	Gap (%)
HUMAN Ig-like 2 / HUMAN Ig-like 4	92	365	75.0	81.5	0.0
HUMAN Ig-like 1 / Mouse Ig-like V-type	110	228	44.5	58.2	1.8



上图为利用dotmatcher程序运行结果，运行命令为：

dotmatcher ceam5\_human.fasta -sbegin 146 -send 677 ceam5\_human.fasta -sbegin 146 -send 677

从图中可以看出，人癌胚抗原5型粘连蛋白（CEAM5\_HUMAN）146-677位残基有三个相似性较高的重复结构域

5. 用EMBOSS程序包TMAP程序和ExPASy服务器TMHMM和TMPred分别预测拟南芥白化基因（ALB3\_ARATH）可能的跨膜螺旋，说明三种方法预测结果差异。（10分）

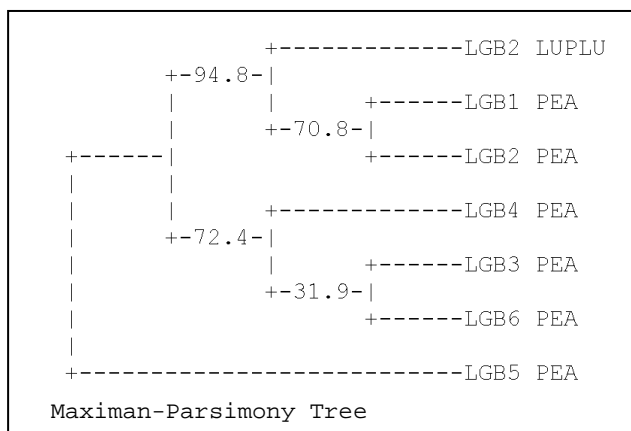
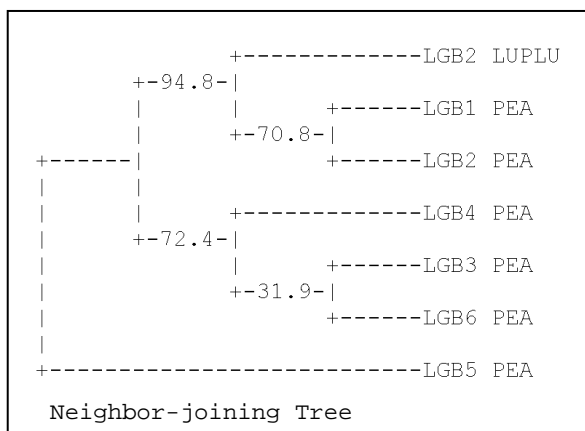
No	0A	0B	1	2	2A	3	4
TMAP			133—157	196—223		273—301	311—337
TMHMM			134-156	201-223		273-295	316-338
TMPred	4—24	80—101	136—154	207—223	233—253	275—294	318—338

结果表明，三种方法都预测到了4个跨膜螺旋，螺旋长度略有出入；TMPred预测到另外三个跨膜螺旋，分值较低，可信度不高。综上所述，拟南芥白化基因（ALB3\_ARATH）含4个跨膜螺旋的可能性较大。

6. 以豆血红蛋白(P02240)为检测序列进行BLAST搜索，找出SwissProt数据库中六个豌豆血红蛋白，将搜索结果填入下表。（10分）

Name	Accession	Length	Score	E-Value	Identity(%)	Positive(%)	Gap(%)
LGB2_LUPLU	P02240	154	-	-	100	100	0
LGB1_PEA	P02233	148	148	1e-38	55	71	4
LGB2_PEA	O48668	147	153	5e-40	54	72	3
LGB3_PEA	O80405	146	157	2e-41	57	75	4
LGB4_PEA	Q9SAZ1	146	154	2e-40	56	73	4
LGB5_PEA	O48665	146	150	5e-39	53	74	4
LGB6_PEA	Q9SAZ0	146	157	2e-41	56	75	4

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7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）



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