实训一 数据库的认识与检索

三大信息中心：

NCBI： http://www.ncbi.nlm.nih.gov/

EBI：http://www.ebi.ac.uk/

DDBJ：http://www.ddbj.nig.ac.jp/

数据库的认识：

GenBank
PIR
PDB
pubMed

请认识下列序列文件：

NCBI:BBQG01000088（序列类型，序列长度，物种，CDS）

NCBI:GQ871748（序列类型，序列长度，物种，CDS）

PDB:6MZ2（结构获取方法，物种）

PubMed:31760684（杂志，日期）

PIR:A0A0A0MS15(序列类型，序列长度，物种)

1：请检索Deng在2018年1月-2019年5月发表的关于链霉菌相关科研论文的数量

2：请检索在2015年1月-2019年7月提交芽胞杆菌基因组的数目

3：请检索苏云金芽胞杆菌Cry基因相关的2018-2019年的文章的数目

4：请检索关于拟南芥在207年1月-2019年12月关于转录调控因子的发表在Plant cell杂志上文章的数目

5：请检索2015年至今关于biosynthesis 相关的书籍

1、查找与水稻抗病基因*Xa21*有关的资料:

（1）有多少条序列具有全长CDS，分别由多少碱基构成？编码多少个氨基酸？

（2）选择修改时间最早的一条序列，指出该基因exon和intron的位置。

2、检索注册号在AF123456－AF123478之间并且序列长度在1500到1800 bp之间的核苷酸数据，共有多少条？如何批量下载？

3、查找秀丽小杆线虫基因组的资料：

（1）chromosome I的测序是否已完成？

（2）已知的chromosome I的序列有多少碱基？序列发表在哪份杂志上？期号和页码？

4、查看拟南芥的系谱关系(lineage)。

5、在PubMed中检索清华大学在2013年1月发表的科研论文。

6、2013年3月底，在上海和安徽两地率先发现了一种能感染人类的H7N9型禽流感病毒（Avian-Origin Influenza A）。中国科学家迅速分离了该病毒并进行了初步研究，首篇正式的论文4月发表在医学领域权威期刊《The New England Journal of Medicine》。目前，NCBI GenBank中已收录该病毒分离自不同病人的多个毒株的序列，以下问题如提到“新H7N9”特指名为“A/Hangzhou/1/2013”的毒株。请根据该背景资料回答以下问题。

(1)请找出这篇文献，列出其在PubMed中的PMID号。

(2)该病毒属于H7N9亚型，其中的“H”代表血凝素（Hemagglutinin），“N”代表神经氨酸酶(Neuraminidase)，分别是病毒外膜上的两种蛋白。H是病毒吸咐于细胞表面的工具，N则是病毒复制完成后脱离细胞表面的工具。请在NCBI核酸数据库（Nucleotide）中找出该毒株编码这两种蛋白的基因的序列，列出Accession号并简要写明过程。

(3)列出该毒株在NCBI物种分类数据库（Taxonomy）中的ID号。NCBI蛋白质数据库（Protein）目前收录了多少条该毒株的蛋白质序列？

蛋白质分析内容

- 蛋白质基本理化性质分析
- 蛋白质跨膜区分析
- 蛋白质信号肽及其剪切位点

1：蛋白质理化性质分析

ExPASy

↓

ProtParam

↓

粘贴序列进行分析

↓

蛋白质的 pI、Mw、氨基酸组成等

Example：P05130

2：蛋白质跨膜区分析

ExPASy

↓

TMHMM

↓

粘贴序列进行分析

↓

分析结果

Example：P05130

ExPASy - Google 搜索 | TMHMM result | ExPASy - AACo

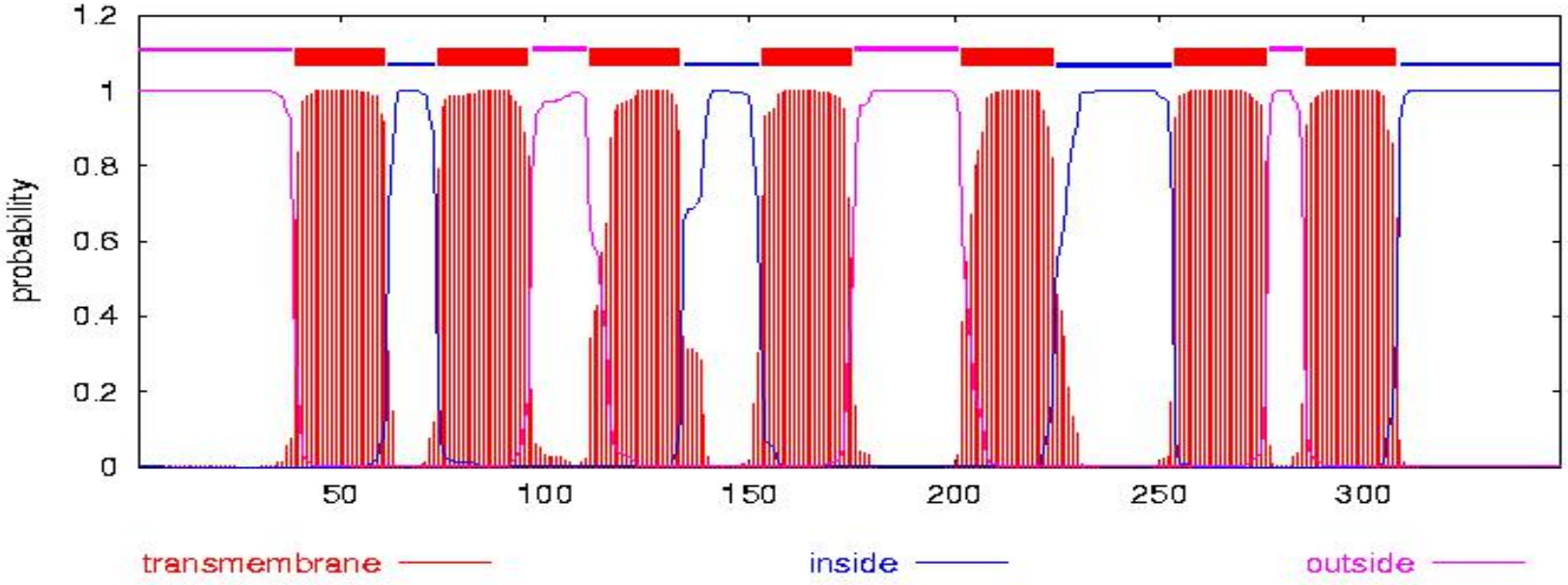
www.cbs.dtu.dk/cgi-bin/nph-webface?jobid=TMHMM

百度 | Google Reader (545) | National Center f··· | BLAST: Basic Loca·

TMHMM result

HELP with output formats

```
# Sequence Length: 5491
# Sequence Number of predicted TMHs:  3
# Sequence Exp number of AAs in TMHs: 72.2963
# Sequence Exp number, first 60 AAs:  1.7974
# Sequence Total prob of N-in:        0.08654
Sequence          TMHMM2.0          outside       1   5349
Sequence          TMHMM2.0          TMhelix    5350   5372
Sequence          TMHMM2.0          inside     5373   5402
Sequence          TMHMM2.0          TMhelix    5403   5425
Sequence          TMHMM2.0          outside    5426   5460
Sequence          TMHMM2.0          TMhelix    5461   5483
Sequence          TMHMM2.0          inside     5484   5491
```



plot in postscript, script for making the plot in gnuplot, data for plot

3：信号肽及其剪切位点的预测

ExPASy

↓

SignalIP

↓

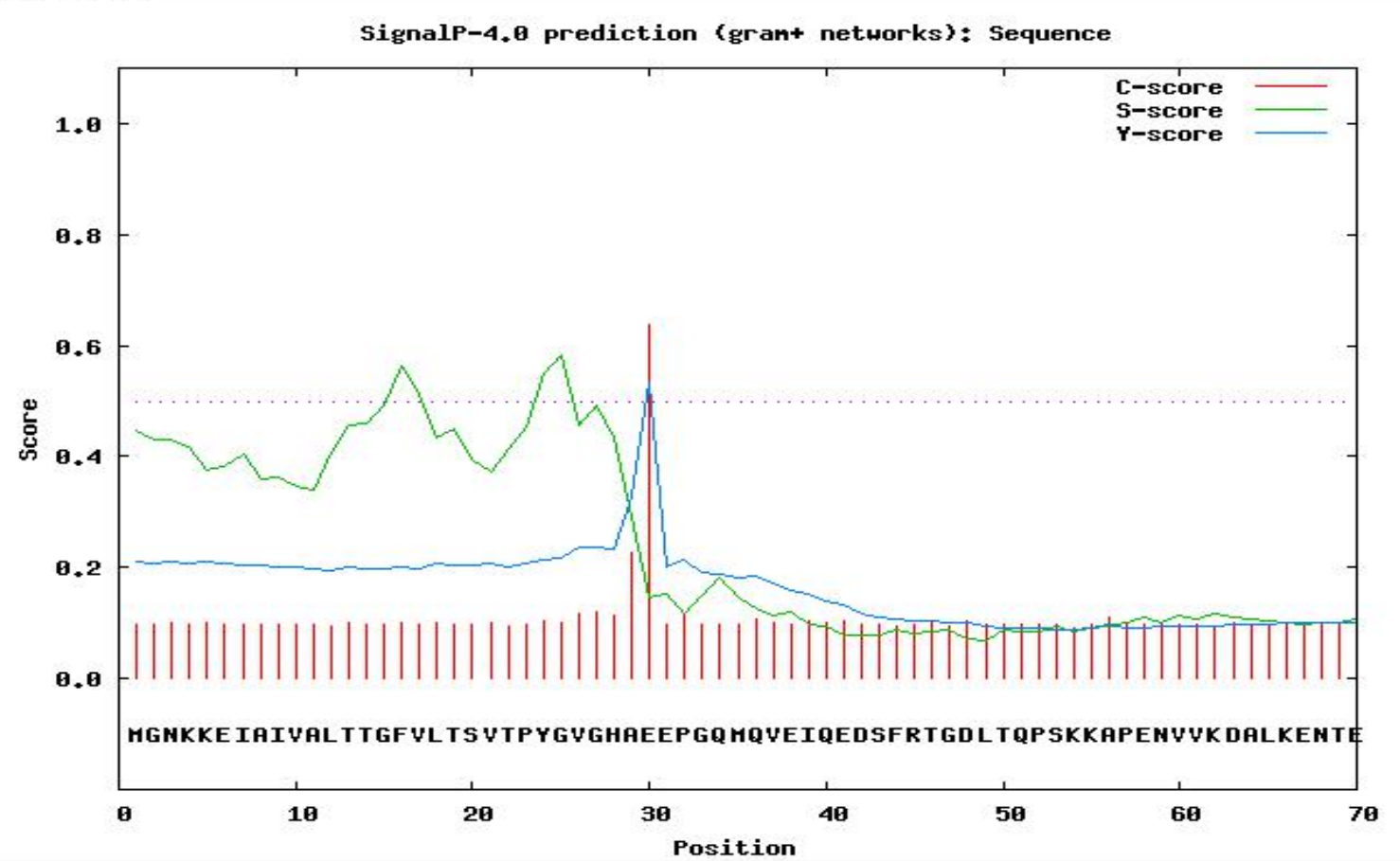
粘贴序列

↓

分析结果

Example：P02699

>Sequence




```
# Measure  Position  Value    Cutoff    signal peptide?
  max. C     30        0.636
  max. Y     30        0.535
  max. S     25        0.583
  mean S      1-29     0.432
       D      1-29     0.495    0.450    YES
Name=Sequence    SP='YES' Cleavage site between pos. 29 and 30: GHA-EE D=0.495 D-cutoff=0.450 Networks=SignalP-TM
# data
# gnuplot script
```

请分析下列序列的基本理化性质，及跨膜螺旋域及信号肽切割位点