Chapter-04. 生物分子序列分析

本章内容提要

- 4.1 引言
- 4.2 DNA特征的序列分析
- 4.3 Protein蛋白质特征的序列分析
- 4.4 生物序列分子的综合分析
- 4.5 其他重要分析资源
- 4.6 数据批量分析方法

重要知识点提示

- ✓ Database and Biological database
- ✓ Database language and tool
- ✓ Primary and secondary databases
- ✓ NCBI, EMBL, DDBJ
- ✓ Software, tools

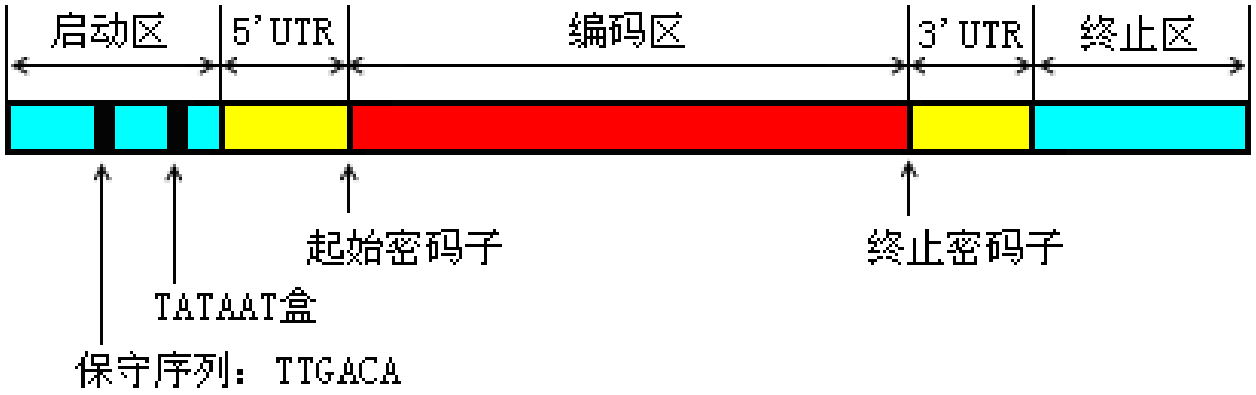
❖ 1.1 几个重要概念

- 基因（gene）指负载特定生物遗传信息的DNA分子片段，在一定的条件下能够表达这种遗传信息，产生特定的生理功能。

- 基因的结构（gene structure）从结构上来讲，包括启动子、编码区及其他调控区等。
  - 原核生物基因及其结构
  - 真核生物基因及其结构

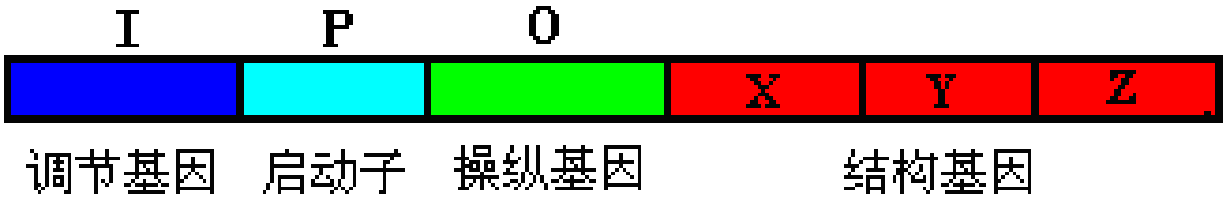
- 蛋白质的结构（protein structure）使通常包括一级结构、二级结构、三级结构甚至四级结构等。

原核生物基因结构：



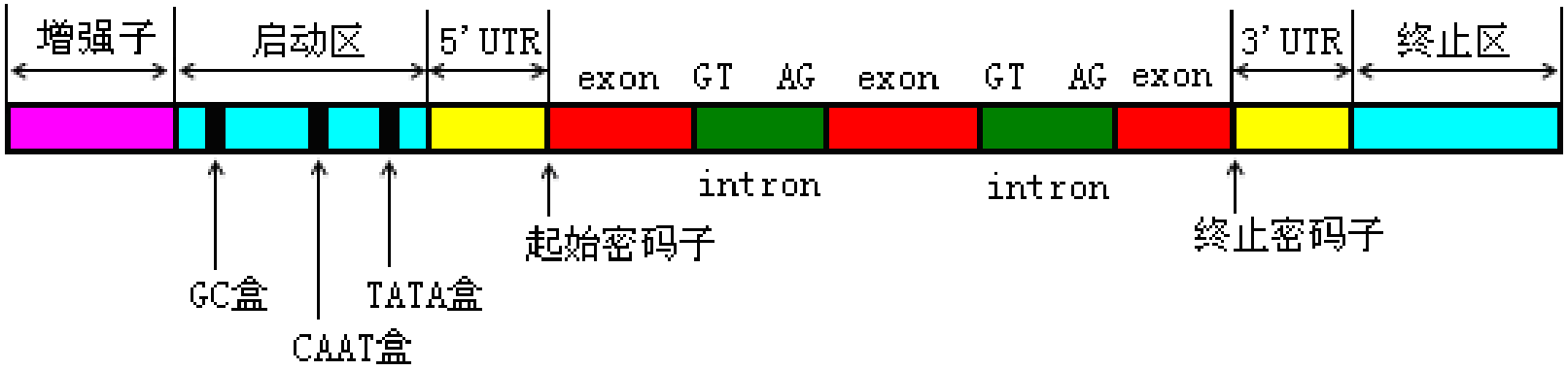
一个完整的原核基因结构是从基因的5’端启动子区域开始，到3’端终止区域结束。基因的转录开始位置由转录起始位点确定，转录过程直至遇到转录终止位点结束，转录的内容包括5’端非翻译区、开放阅读框及3’端非翻译区。基因翻译的准确起止位置由起始密码子和终止密码子决定，翻译的对象即为介于这两者之间的开放阅读框ORF。

操纵子模型结构



原核生物大多数基因表达调控是通过操纵子机制实现的。所谓操纵子通常由调节基因、启动子、操纵基因以及2个以上的编码序列（结构基因）在原核生物基因组中成簇串联组成。其中结构基因的表达受到操纵基因的调控。调节基因能产生作用于操纵基因的阻遏物（一种蛋白质），操纵基因靠近它所控制的结构基因，阻遏物与操纵基因的结合能阻止结构基因的转录。

真核生物基因结构：

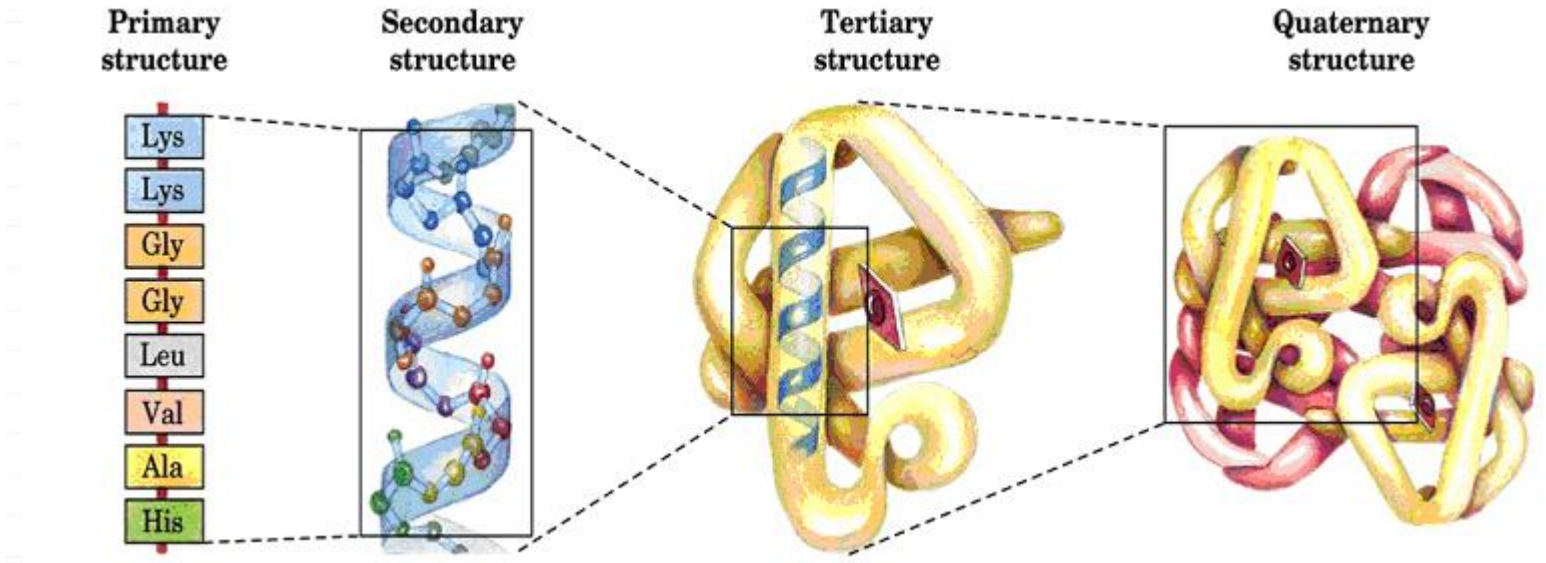


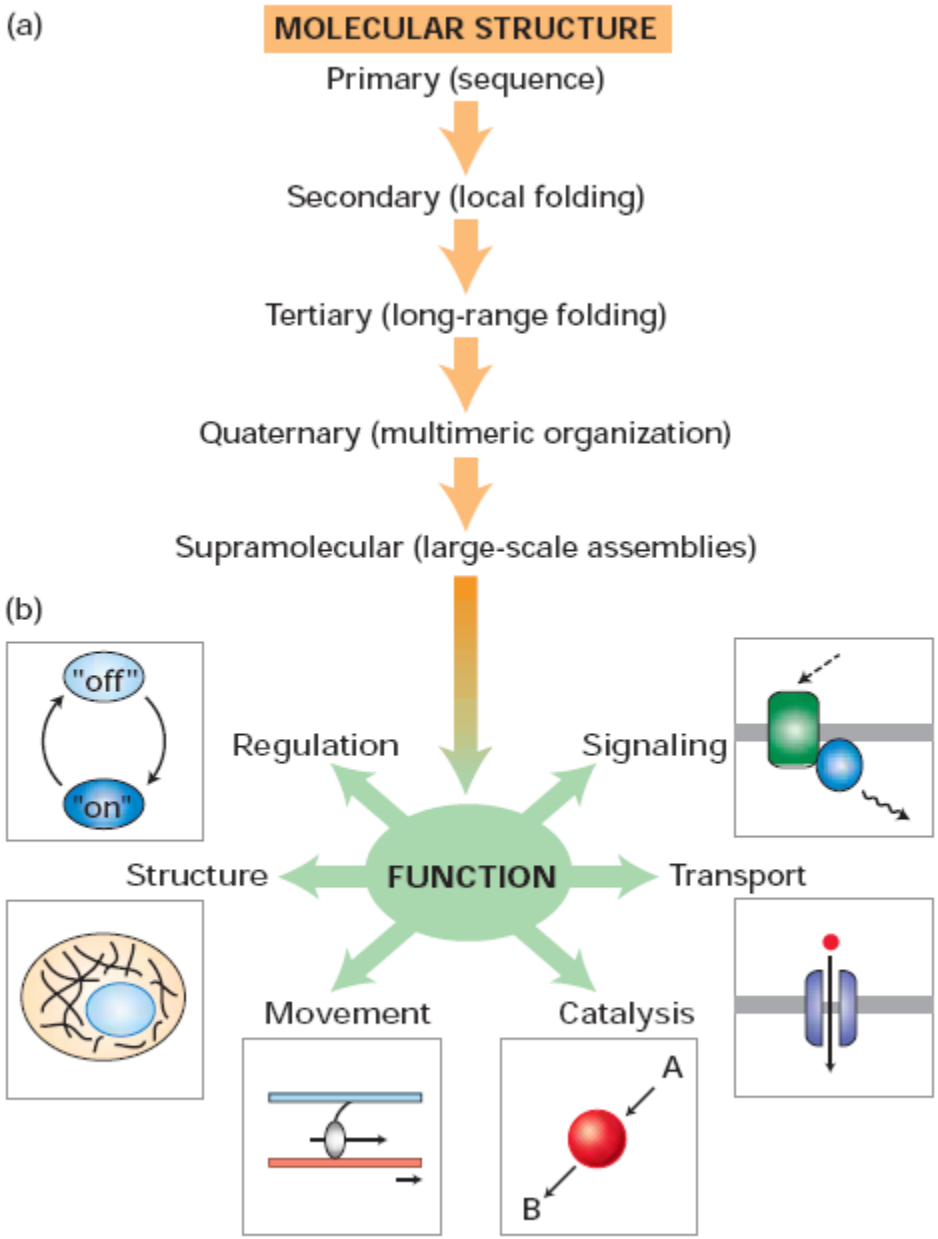
一个完整的真核生物基因，不但包括编码区域，还包括5’端和3’端两侧长度不等的特异性序列，虽然这些序列不编码氨基酸，却在基因表达的过程中起着重要的作用。所以，严格的“基因”这一术语的分子生物学定义是：产生一条多肽链或功能RNA所必须的全部核苷酸序列。

❖ 1.3 蛋白质的结构及其层级

- 蛋白质是一种生物大分子，蛋白质中相邻的氨基酸通过肽键形成一条伸展的肽链，这条链称为蛋白质的一级结构。

- 肽链上的氨基酸残基形成局部的二级结构，各种二级结构在空间卷曲折叠形成特定的三维空间结构。有的蛋白质由多条肽链组成，每条肽链称为亚基，亚基之间又有特定的空间关系，称为蛋白质的四级结构。

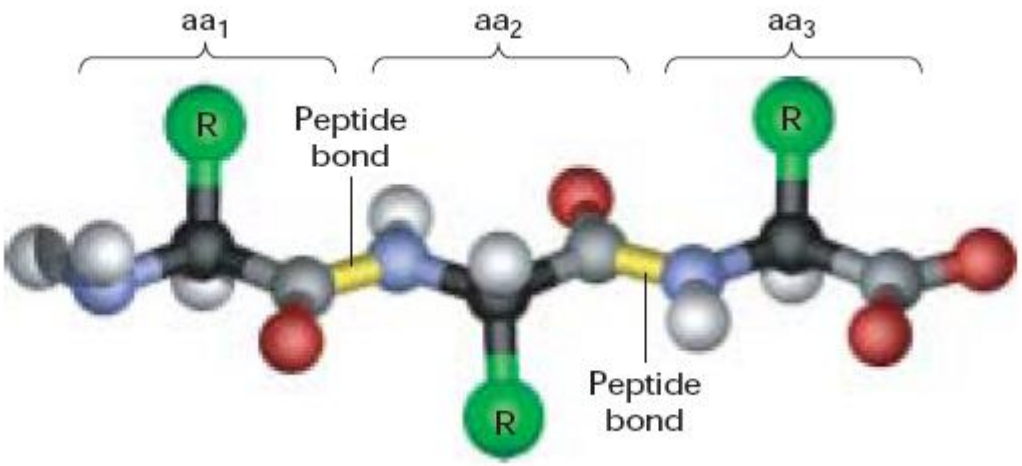




▲ **FIGURE 3-1** Overview of protein structure and function.

The Primary Structure of a Protein Is Its Linear Arrangement of Amino Acids

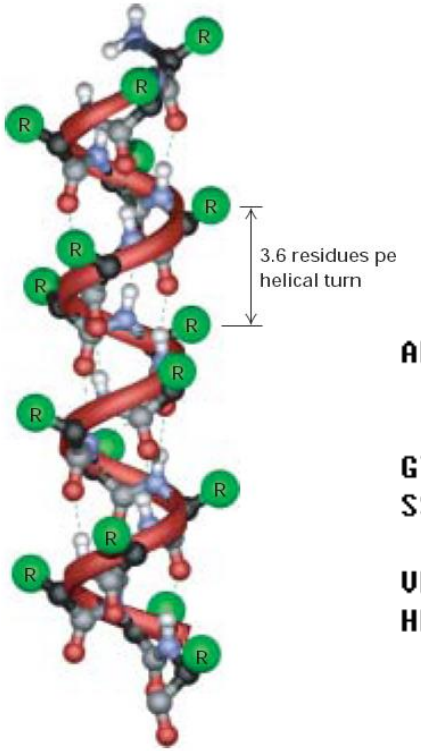
```
>1GGZ:A|PDBID|CHAIN|SEQUENCE
ADQLTEEQVTEFKEAFSLFDKDGDGCITTRELGTVMRSLGQNPTEAELRDMMSEIDRDGNGTVDF
PEFLGMMARKMKDTDNEEEIREAFRVFDKDGNGFVSAAELRHVMTRLGEKLSDEEVDEMIRAADT
DGDGQVNYEEFVRVLVSK
```



蛋白质的一级结构决定二级结构

蛋白质的二级结构决定三级结构

Secondary Structures Are the Core Elements of Protein Architecture



H表示螺旋　　E表示折叠　　B表示β桥

G表示3-螺旋　　I表示π螺旋　　T表示氢键转角

S代表转向

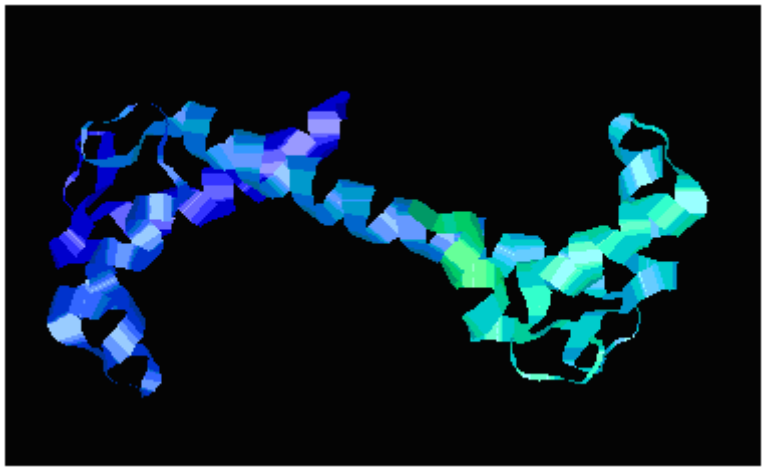
```
ADQLTEEQVT EFKEAFSLFD KDGDGCITTR ELGTVMRSLG QNPTEAELRD MMSEIDRDGN
     HHHHH HHHHHHHHHH  TT SSEE HH HHHHHHHHTT     HHHHHH HHHTT TT S

GTVDFPEFLG MMARKMKDTD NEEEIREAFR VFDKDGNGFV SAAELRHVMT RLGEKLSDEE
SSEEHHHHHH HHHHHHHHHH HHHHHHHHHH HH TT SSEE HHHHHHHHHH HH     HHH

VDEMIRAADT DGDGQVNYEE FVRVLVSK
HHHHHHHH T TSSSSEEHHH HHHHHH
```

H = alpha helix; B = residue in isolated beta-bridge; T = hydrogen bonded turn;
E = extended strand, participates in beta ladder; G = 3-helix; I = 5 helix; S = bend

TABLE 4–2 Secondary Structures and Properties of Some Fibrous Proteins

Structure	Characteristics	Examples of occurrence
α Helix, cross-linked by disulfide bonds	Tough, insoluble protective structures of varying hardness and flexibility	α-Keratin of hair, feathers, and nails
β Conformation	Soft, flexible filaments	Silk fibroin
Collagen triple helix	High tensile strength, without stretch	Collagen of tendons, bone matrix

蛋白质空间结构



蛋白质的生物学功能在很大程度上取决于蛋白质的空间结构，但蛋白质的空间结构又取决于蛋白质一级结构中的氨基酸组成和排列顺序，蛋白质结构构象多样性导致了不同的生物学功能。蛋白质分子只有处于它自己特定的空间结构情况下，才能获得它特定的生物活性，空间结构稍有破坏，就很可能会导致蛋白质生物活性的降低甚至丧失，因为它们的特定的结构允许它们结合特定的配体分子。知道了基因密码，科学家们可以推演出组成某种蛋白质的氨基酸序列，却无法绘制蛋白质空间结构。

因而，揭示人类每一种蛋白质的空间结构，已成为后基因组时代的制高点，这也是结构基因组学的基本任务。

对DNA序列和蛋白质序列进行序列特征分析，能够使我们从分子层次上了解基因的结构特点，了解与基因表达调控相关的信息，了解DNA序列与蛋白质序列之间的编码，了解蛋白质序列与蛋白质空间结构之间的关系和规律，为进一步研究了解蛋白质功能与蛋白质结构之间的关系提供理论依据。

第2节：DNA序列特征分析

重要知识点提示：

- 开放阅读框（ORF）：
  - 外显子
  - 内含子
- CG岛
- 转录终止信号
- 启动子
- 密码子偏好性

- 除了进行序列比对之外，**DNA序列分析中更重要的工作是从序列中找到基因及其表达调控信息。**
- 寻找基因的两层含义：
  - 一是识别与基因相关的特殊序列信号，如启动子、起始密码子，通过信号识别大致确定基因所在的区域。
  - 二是预测基因的编码区域，或预测外显子所在的区域。
- **在此基础上，结合两个方面的结果确定基因的位置和结构。**
- 绝大部分基因表达调控信息隐藏在基因序列的上游区域，在组成上具有一定的特征，**可以通过序列分析识别这些特征。**

❖ 2.1 开放阅读框

- 开放阅读框指的是从5‘端起始密码子（ATG）开始到终止密码子的一段用于编码蛋白质的核酸序列。
- 每个序列都有6个可能的开放阅读框，其中3个开始于第1、2、3个碱基位点并沿着给定序列的5‘→ 3’的方向进行延伸，而另外的3个开始于第1、2、3个碱基位点但沿着互补序列的5‘→ 3’的方向进行延伸。起我们并不知道DNA双链中哪一条单链是编码链，也不知道准确的翻译起始点在何处，由于每条链都有3种可能的开发阅读框，2条链共计6种可能的开放读框。
- 目标：从这6个可能的开放阅读框中找出一个正确的开放阅读框。根据这个开放阅读框翻译得到的氨基酸序列才是真正表达的蛋白质产物。

真核生物的开放阅读框

真核生物的开放阅读框不仅含有编码蛋白的外显子（exon），而且还有内含子（intron），并且内含子将开放阅读框分割为若干个小片段。开放阅读框的长度变化范围非常大，因此真核生物的基因预测远比原核生物困难。

但是，在真核生物的开放阅读框中，外显子与内含子之间的连接绝大部分情况下满足GT-AG规律：内含子序列 5’ 端的起始两个核苷酸总是GT，并且其3’端的最后两个核苷酸总是AG，即：5’-GT .....AG-3’，这个规律有助于真核生物开放阅读框的识别。

❖ 2.2 基于同源比较和模型预测的基因预测

- **同源比较算法：Smith-waterman算法，FASTA算法**

  基于同源性来预测基因

- **隐马尔科夫模型（及广义隐马尔科夫模型）**

  把DNA序列看做随机过程，根据在核苷酸选用频率上的不同来自动寻找其内部蕴藏的规律，进而预测基因

- **动态规划方法**

  用来将预测的可能的外显子和内含子拼接成完整的基因，将各种可能的拼接打分，从而得出最有可能的基因结构

- **神经网络预测方法**

  用来使用一群训练数据集来训练出神经网络模型（经过参数优化），用该模型去预测未知基因

GENE Prediction

Gene prediction methods

- *ab initio*
  - *Gene signals*
    - ✓ *Start/stop codons*
    - ✓ *Intron splice signals*
    - ✓ *Transcription factor binding sites*
    - ✓ *Ribosomal binding sites*
    - ✓ *Poly-A sites*
  - *Gene content*
    - ✓ *Statistical description of coding regions*
    - ✓ *Difference between coding and non-coding regions*
- *Homology*
  - ✓ *Translated DNA matches known protein sequences*
  - ✓ *Exons of genomic DNA match a sequenced cDNA*

Statistical approaches

- *Exploit statistical characteristics of coding regions and non-coding regions and other knowledge about gens*
- *Can be potentially detect new genes*
- *May not be reliable*

Similarity approaches

- *Exploit fact that many genes are conserved across species*
- *Can be highly reliable*
- *Good at finding known genes*

目前最为流行的GHMM方法

- 第一代的软件：如GENMARK, GeneID和GRAIL II等

  基于ANN与HMM；假定序列中正好有一个完整的基因存在；预测准确性不高

- 第二代基因识别软件：如GenScan, HMMGene，FFG和GeneMark.html等

  一般不需要关于存在完整在序列中的假定；基本采用GHMMM模型的方法；预测准确率大幅提升

- GenScan是一种广义上的目的基因预测软件，用来分析多个物种的DNA序列。

利用GENSCAN识别基因开放阅读框

GENSCAN是美国麻省理工学院的Chris Burge于1997年开发成功的人类（或脊椎动物）基因预测软件，它是根据基因组DNA序列来预测开放阅读框及基因结构信息的开放式在线资源，尤其适用于脊椎动物、拟南芥和玉米等真核生物。

GENSCAN的网址为：

http：//genes.Mit.edu/GENSCAN.html

Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268,** 78-94.

The GENSCAN Web Server at MIT

Identification of complete gene structures in genomic DNA

For information about Genscan, click here

Server update, November, 2009: We've been recently upgrading the GENSCAN webserver hardware, which resulted in some problems in the output of GENSCAN. We apologize for the inconvenience. These output errors were resolved.

This server provides access to the program Genscan for predicting the locations and exon-intron structures of genes in genomic sequences from a variety of organisms.

This server can accept sequences up to 1 million base pairs (1 Mbp) in length. If you have trouble with the web server or if you have a large number of sequences to process, request a local copy of the program (see instructions at the bottom of this page).

Organism: Vertebrate ▾ Suboptimal exon cutoff (optional): 1.00 ▾

Sequence name (optional):

Print options: Predicted peptides only ▾

Upload your DNA sequence file (upper or lower case, spaces/numbers ignored): 选择文件 未选择任何文件

Or paste your DNA sequence here (upper or lower case, spaces/numbers ignored):

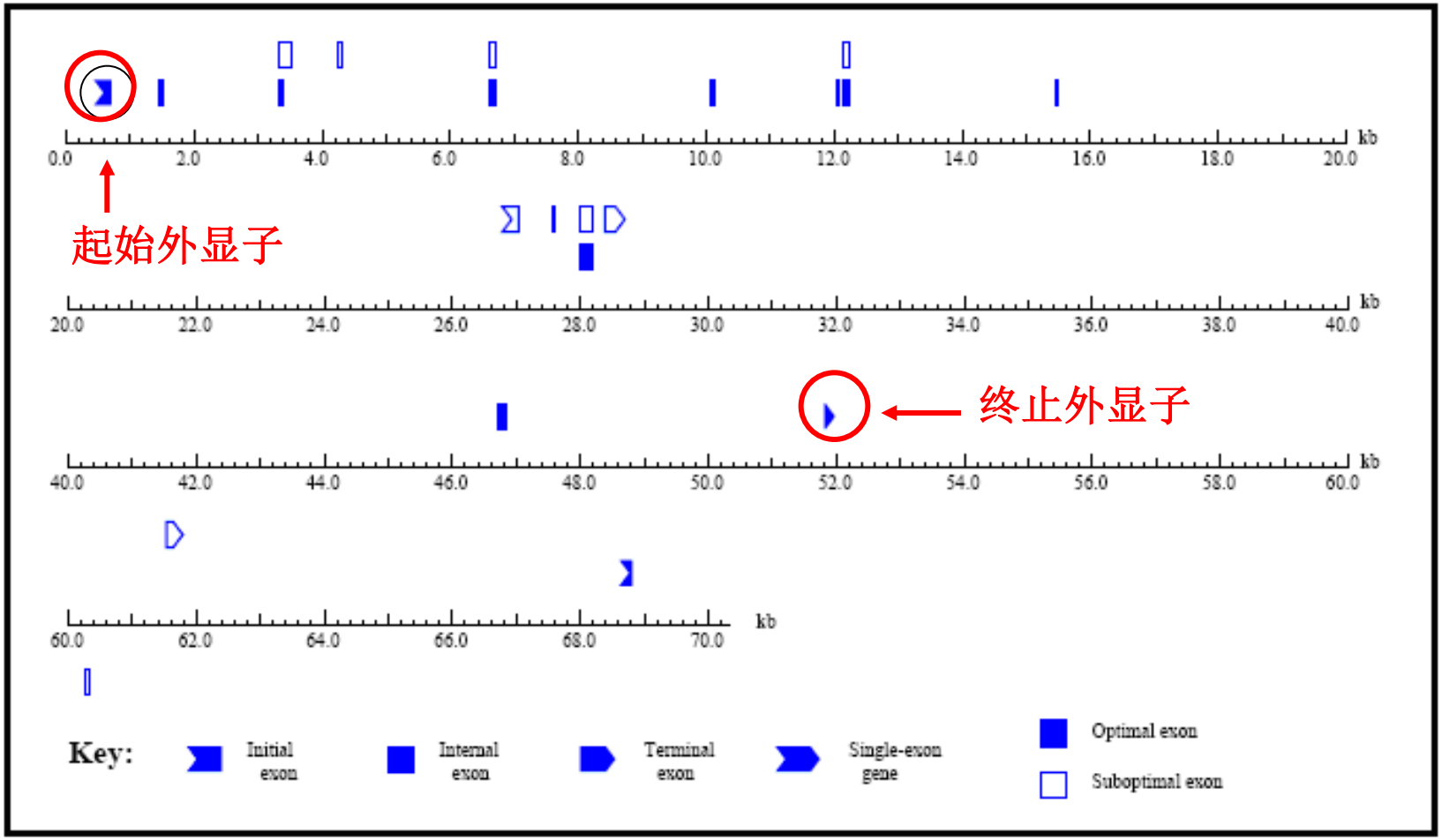
Run GENSCAN Clear Input

Back to the top

用GENSCAN预测AC002390序列的基因/外显子

Gn.Ex	Type	S	.Begin	...End	.Len	Fr	Ph	I/Ac	Do/T	CodRg	P....	Tscr..
1.01	Init	+	532	657	126	0	0	66	105	46	0.633	2.88
1.02	Intr	+	1399	1459	61	0	1	90	94	20	0.688	1.11
1.03	Intr	+	3269	3349	81	0	0	118	94	76	0.606	10.81
1.04	Intr	+	6557	6649	93	0	0	42	80	77	0.503	2.24
1.05	Intr	+	10004	10093	90	0	0	66	53	84	0.861	2.67
1.06	Intr	+	11990	12019	30	0	0	135	115	37	0.954	9.20
1.07	Intr	+	12099	12173	75	1	0	128	44	90	0.339	8.09
1.08	Intr	+	15414	15459	46	1	1	130	109	21	0.433	5.87
1.09	Intr	+	27955	28151	197	1	2	77	98	122	0.487	11.16
1.10	Intr	+	46659	46791	133	1	1	112	38	68	0.244	3.90
1.11	Term	+	51762	51783	22	0	1	101	38	8	0.025	-5.12
1.12	PlyA	+	52398	52403	6							1.05
2.00	Prom	+	59901	59940	40							-2.16
2.01	Init	+	68711	68764	54	1	0	89	92	66	0.282	8.38

用GENSCAN预测AC002390序列的基因/外显子的位置图



GenomeScan

webserver at MIT

This server provides access to the program GenomeScan for predicting the locations and exon-intron structures of genes in genomic sequences from a variety of organisms.

GenomeScan incorporates protein homology information when predicting genes. This server allows you to input proteins suspected to be similar to regions of your DNA sequence. You can find such proteins by doing a BLASTX comparison of your sequence to all known proteins, or by running GENSCAN and then comparing the results to known proteins using BLASTP. Please input the proteins in FastA format; the file may contain multiple proteins so long as each is separated by a header on its own line. Files should contain less than one million bases.

If you would like to test the program, feel free to use this DNA testfile and the corresponding protein file.

More information on GenomeScan: GenomeScan homepage

You may also wish to use or read about the GENSCAN server, GenomeScan's predecessor.

Run GenomeScan:

Organism: Vertebrate

Sequence name (optional):

Print options: Predicted peptides only

Upload your DNA sequence file (one-letter code, upper or lower case, spaces/numbers ignored):

选择文件 未选择任何文件

Or paste your DNA sequence here (one-letter code, upper or lower case, spaces/numbers ignored):

Run GenomeScan Clear Input

❖ 2.3 CpG岛及其预测

CpG岛是指DNA序列上的一个区域，此区域含有大量相联的胞嘧啶（C）、鸟嘌呤（G），以及使两者相连的磷酸酯键（p）。

CpG岛的概念是Gardiner-garden和Fromner于1987年提出的，基因中平均每100 Kb即可出现。CpG岛位于基因的启动子和第一个外显子区，约有60%~80%的人类基因的启动子和起始外显子含有CpG岛，其中GC含量大于50%，长度超过200bp。因此搜索CpG岛可以为基因及其启动子预测提供重要线索。

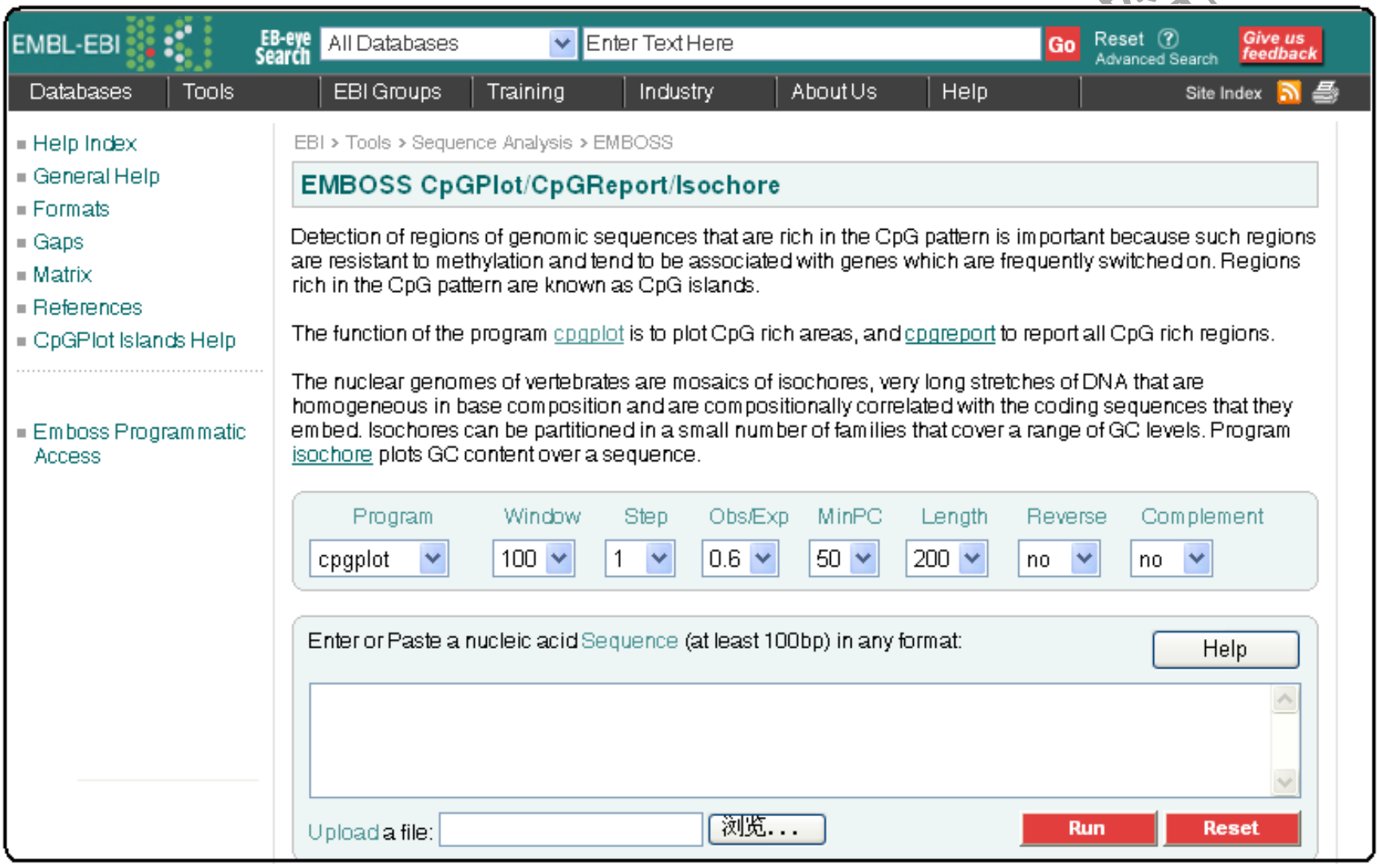
利用CpGPlot预测分析CpG岛

CpGPlot是预测CpG岛的在线工具，它是由欧洲分子生物学实验室EMBL —— European Molecular Biology Laboratory提供的。

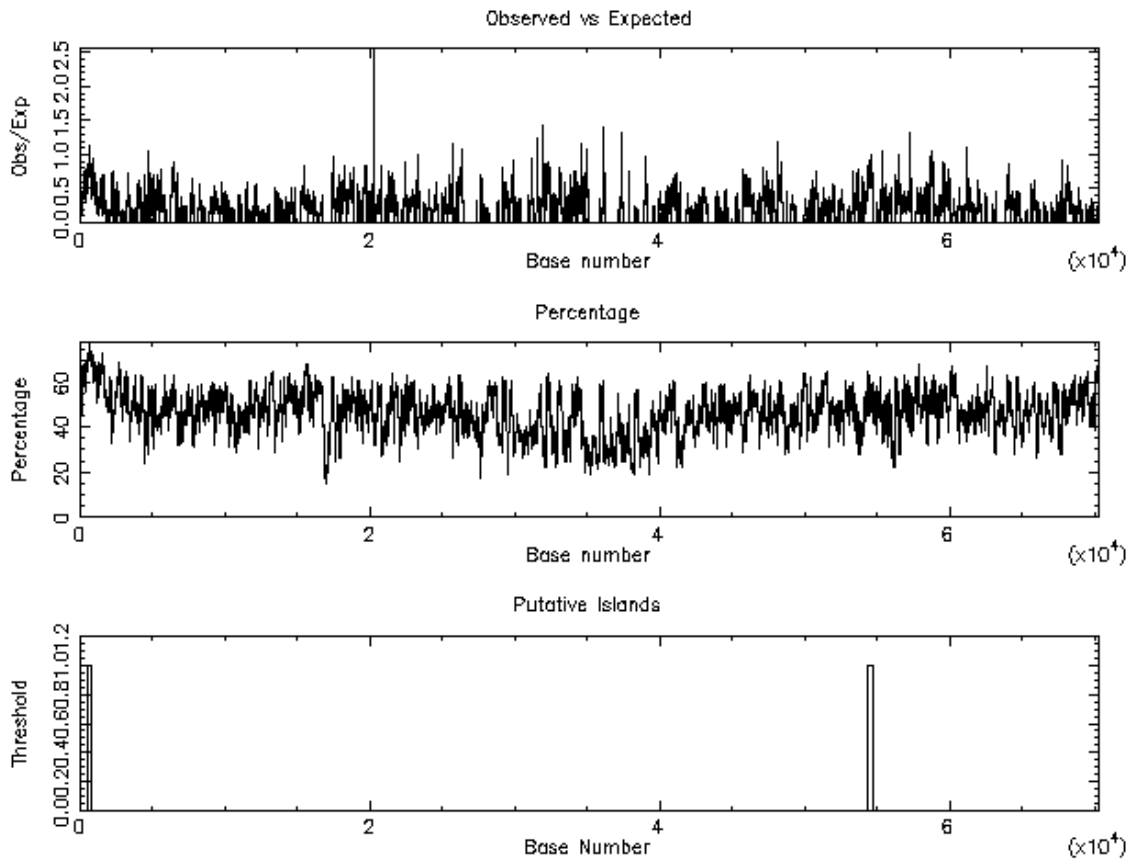
其网址为：

http://www.ebi.ac.uk/Tools/emboss/cpgplot/index.html

CpGPlot在线操作页面



用CpGplot预测AC002390序列的CpG岛的结果




```
CPGPLOT islands of unusual CG composition
AC002390,Human from 1 to 70311

     Observed/Expected ratio > 0.60
     Percent C + Percent G > 50.00
     Length > 200

 Length 227 (501..727)

 Length 312 (54380..54691)
```

用CpGReport预测AC002390序列的CpG岛的结果

```
ID   AC002390,Human  70311 BP.
XX
DE   CpG Island report.
XX
CC   Obs/Exp ratio > 0.60.
CC   % C + % G > 50.00.
CC   Length > 200.
XX
FH   Key                 Location/Qualifiers
FT   CpG island          501..727
FT                       /size=227
FT                       /Sum C+G=161
FT                       /Percent CG=70.93
FT                       /ObsExp=0.73
FT   CpG island          54380..54691
FT                       /size=312
FT                       /Sum C+G=164
FT                       /Percent CG=52.56
FT                       /ObsExp=0.66
FT   numislands          2
//
```

❖ 2.4 转录终止信号及其预测

转录终止信号是在mRNA序列的3'端终止密码子下游位置上的加尾信号（tailing signal）。

前体mRNA 3‘端多聚腺苷酸化是真核细胞内mRNA转录后处理的三个最主要步骤之一，这三个步骤包括：5’帽子结构的形成、内含子的剪切及3‘端的多聚腺苷酸化。

因此，前体mRNA 3'端多聚腺苷酸化与mRNA稳定性的调节、mRNA的细胞内转运、翻译的起始以及一些其他的细胞机制和疾病机制有着重要关系。

真核生物前体mRNA3'端的多聚腺苷酸化包括两个步骤：

1. **特异性的核苷酸内切酶在PolyA位点处进行断裂；**
2. **腺苷酸聚合酶在断裂位点处添加PolyA尾巴，其主要标志为AATAAA或ATTAAA两种序列，称为多聚腺苷酸信号（polyadenylation signal），简称PolyA信号序列，也称为转录终止信号。**

在3'UTR区存在多个潜在PolyA位点，因此对PolyA位点的准确识别，对于预测基因结构、理解mRNA的形成机制及某些疾病的分子机制具有巨大的作用。

利用POLYAH预测分析转录终止信号

SoftBerry网站的POLYAH软件是识别3'端剪切和PolyA区域的在线工具。

其网址为：

http://linux1.softberry.com/berry.phtml?topic=polyah&group=programs&subgroup=promoter

POLYAH在线页面


SoftBerry
HOME
PRODUCTS
NEW PRODUCTS
SERVICES
MANAGEMENT TEAM
CORPORATE PROFILE
CONTACT
TEST ON LINE
GENE FINDING in Eukaryota
GENE FINDING WITH SIMILARITY
OPERON AND GENE FINDING IN BACTERIA
GENE FINDING IN VIRUSES
ALIGNMENT /Sequences&genomes
GENOME EXPLORER /Infogene
SEARCH FOR MOTIFS /promoters&functional
PROTEIN LOCATION /patterns/Epitops
RNA STRUCTURE COMPUTING
SoftBerry
POLYAH
POLYAH / Recognition of 3'-end cleavage and polyadenilation region
Paste nucleotide sequence here:
Alternatively, load a local file with sequence in Fasta format:
Local file name:
浏览...
PROCESS
RESET
[Help]
[Example]

用POLYAH预测AC002390序列的转录终止信号的结果

```
> test sequence
 Length of sequence-      70313
    50 potential polyA sites were predicted
 Pos.:     122 LDF-   3.38
 Pos.:    5057 LDF-   6.18
 Pos.:    6060 LDF-   3.62
 Pos.:    6064 LDF-   4.24
 Pos.:    6076 LDF-   6.17
   .        .
   .        .
   .        .
 Pos.:   44580 LDF-   6.78
 Pos.:   50627 LDF-   4.15
 Pos.:   50635 LDF-   2.84
 Pos.:   52398 LDF-   2.54
 Pos.:   56541 LDF-   5.73
 Pos.:   56546 LDF-   5.64
 Pos.:   56551 LDF-   2.71
   .        .
   .        .
   .        .
```

❖ 2.5 基因启动子及其预测

启动子是基因的一个组成部分，是位于结构基因5'端上游区的DNA序列，控制基因表达（转录）的起始时间和表达的程度。启动子本身并不控制基因活动，而是通过与称为转录因子的蛋白质结合而控制基因活动的。转录因子就像一面“旗子”，指挥RNA聚合酶的活动。如果基因的启动子部分发生突变，则会导致基因表达的调节障碍。这种突变常见于恶性肿瘤。

利用PromoterScan预测分析启动子区域

BioInformatics and Molecular Analysis Section网站的**PromoterScan软件**是预测分析启动子区域的在线工具。

其网址为：

http://www-bimas.cit.nih.gov/molbio/proscan/

PromoterScan在线网页

WWW Promoter Scan

Function: Predicts Promoter regions based on scoring homologies with putative eukaryotic Pol II promoter sequences.
The analysis is done using the PROSCAN Version 1.7 suite of programs developed by Dr. Dan Prestridge, Information on PROSCAN, including details on obtaining a copy, is maintained at the Advanced Biosciences Computing Center, University of Minnesota.

A DNA sequence is all that needs to be supplied. There are no optional parameters for PROSCAN.

Please enter or paste a Nucleic Acid sequence to analyze (most formats accepted):

☑ Echo input sequence (generally recommended)

Be Forewarned!

Patience is a virtue: Analysis for a 10Kbp sequence may take as long as 5 minutes (or more)!

submit reset *Credits: WWW implementation by BIMAS staff*

用PromoterScan预测AC002390序列的启动子区域的结果

Promoter region predicted on forward strand in 47985 to 48235
Promoter Score: 57.71 (Promoter Cutoff = 53.000000)

Significant Signals:

Name	TFD #	Strand	Location	Weight
PEA1	S01595	+	48087	1.539000
AP-1	S01426	-	48093	1.513000
TFIID	S01540	+	48111	1.971000
TFIID	S00087	+	48111	2.618000
AABS_CS2	S01612	+	48199	1.012000
Sp1	S00952	+	48224	50.000000
Sp1	S01542	-	48233	3.608000

Promoter region predicted on forward strand in 55226 to 55476
Promoter Score: 60.49 (Promoter Cutoff = 53.000000)
TATA found at 55449, Est.TSS = 55479
Significant Signals:

Name	TFD #	Strand	Location	Weight
Sp1	S00802	+	55260	3.292000
Sp1	S00978	-	55265	3.361000
UCE.2	S00437	+	55315	1.278000
NFI	S00281	+	55377	1.948000
CTF	S00780	-	55383	1.448000
JCV_repeated_sequenc	S01193	-	55408	1.658000
TFIID	S01540	+	55450	1.971000
TFIID	S00087	+	55450	2.618000
TFIID	S00615	+	55450	2.920000

❖ 2.6 密码子偏好性及其预测

密码子使用偏性是指生物体中编码同一种氨基酸的同义密码子的非均匀使用现象。

这一现象的产生与诸多因素有关，如基因的表达水平、翻译起始效应、基因的碱基组分、某些二核苷酸的出现频率、G+C含量、基因的长度、tRNA的丰度、蛋白质的结构及密码子—反密码子间结合能的大小等。

所以对密码子使用偏好性的分析具有重要的生物学意义。

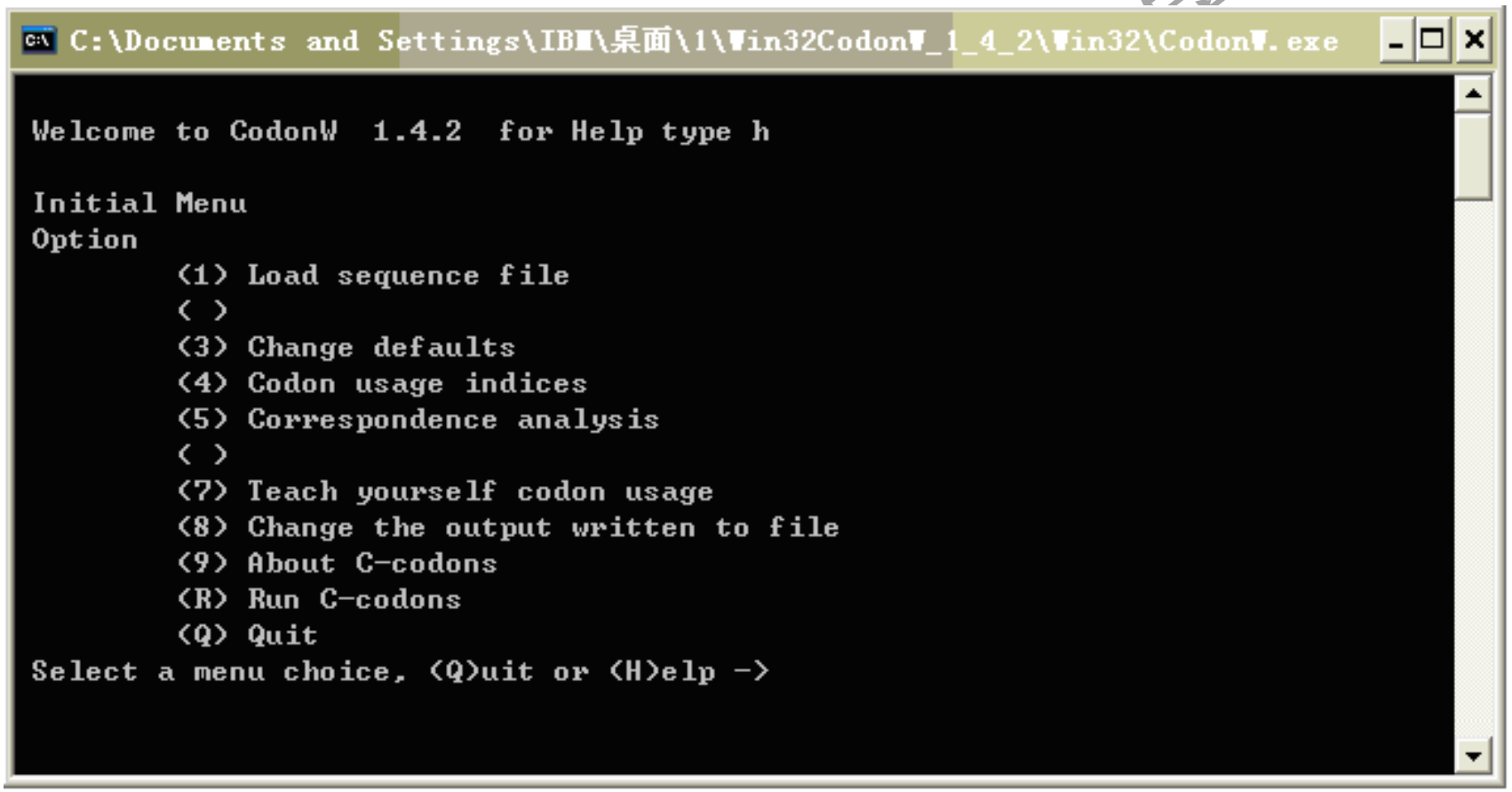
利用CodonW分析密码子偏好性

CodonW是美国DEC公司开发的对密码子的使用进行分析的免费的软件工具。

此软件是建立在大量的统计学分析的基础上，为了简化在线分析的复杂性而开发的，它可以在Windows环境下运行，并且可以同时处理2000条以上的序列。通过对DNA或RNA序列的分析，CodonW会产生关于密码子使用的相关指标的统计学分析的数据，我们可以利用这些数据对我们所要了解的序列进行分析。

其下载网址为： ftp://molbiol.ox.ac.uk/cu/codonW.tar.Z

CodonW1.4主菜单的操作页面



11个密码子使用的指标

序 号	全 称	缩 写
1	**Codon Adaptation Index**	**CAI**
2	**Frequency of Optimal Codons**	**Fop**
3	**Codon Bias Index**	**CBI**
4	**The effective number of codons**	**ENc**
5	**G+C content of the gene**	**G+C**
6	**G+C content at 3rd position of synonymous codons**	**GC3s**
7	**Silent base composition**	**LSil**
8	**Number of silent sites**	**LAA**
9	**and amino acids**	**GRAVY**
10	**Hydrophobicity of protein**	**Aromo**
11	**Aromaticity score**	

waxy基因的序列

序号	Genebank登陆号	物 种	基因功能
1	AY094405	Arabidopsis haliana	granule bound starch synthase I mRNA
2	AF486514	Hordeum vulgare	granule bound starch synthase I mRNA
3	X03935	Zea mays	glucosyl transferase
4	X62134	O.sativa	granule bound starch synthase I mRNA
5	X88789	P.sativum	mRNA for starch synthase
6	U23945	Sorghum bicolor	granule-bound starch synthase precursor (Wx)mRNA
7	X57233	Wheat	waxy mRNA for granule-bound starch synthase

用CodonW分析waxy基因所得的RSCU值 和 个数

		High RSCU	Bias CU	Low RSCU	Bias CU			High RSCU	Bias CU	Low RSCU	Bias CU
Phe	UUU	0.00	(0)	1.33	(2)	Ser	UCU	0.00	(0)	0.41	(2)
	UUC	0.00	(0)	1.67	(20)		UCC	0.00	(0)	2.07	(10)
Leu	UUA	0.00	(0)	0.00	(0)		UCA	0.00	(0)	0.00	(0)
	UUG	0.00	(0)	0.24	(2)		UCG	0.00	(0)	0.41	(2)
	CUU	0.00	(0)	0.71	(6)	Pro	CCU	0.00	(0)	0.13	(1)
	CUC	0.00	(0)	3.06	(26)		CCC	0.00	(0)	1.81	(14)
	CUA	0.00	(0)	0.12	(1)		CCA	0.00	(0)	0.13	(1)
	CUG	0.00	(0)	1.88	(16)		CCG	0.00	(0)	1.94	(15)
Ile	AUU	0.00	(0)	0.21	(2)	Thr	ACU	0.00	(0)	0.15	(1)
	AUC	0.00	(0)	2.79	(26)		ACC	0.00	(0)	2.37	(16)
	AUA	0.00	(0)	0.00	(0)		ACA	0.00	(0)	0.15	(1)
Met	AUG	0.00	(0)	1.00	(18)		ACG	0.00	(0)	1.33	(9)
Val	GUU	0.00	(0)	0.00	(0)	Ala	GCU	0.00	(0)	0.43	(6)
	GUC	0.00	(0)	1.74	(23)		GCC	0.00	(0)	2.14	(30)
	GUA	0.00	(0)	0.00	(0)		GCA	0.00	(0)	0.29	(4)
	GUG	0.00	(0)	2.26	(30)		GCG	0.00	(0)	1.14	(16)
Tyr	UAU	0.00	(0)	0.22	(2)	Cys	UGU	0.00	(0)	0.00	(0)
	UAC	0.00	(0)	1.78	(16)		UGC	0.00	(0)	2.00	(12)
TER	UAA	0.00	(0)	0.00	(0)	TER	UGA	0.00	(0)	3.00	(1)
	UAG	0.00	(0)	0.00	(0)	Trp	UGG	0.00	(0)	1.00	(8)
His	CAU	0.00	(0)	0.00	(0)	Arg	CGU	0.00	(0)	0.18	(1)
	CAC	0.00	(0)	2.00	(8)		CGC	0.00	(0)	2.00	(11)
Gln	CAA	0.00	(0)	0.11	(1)		CGA	0.00	(0)	0.00	(0)
	CAG	0.00	(0)	1.89	(18)		CGG	0.00	(0)	1.45	(8)
Asn	AAU	0.00	(0)	0.09	(1)	Ser	AGU	0.00	(0)	0.21	(1)
	AAC	0.00	(0)	1.91	(22)		AGC	0.00	(0)	2.90	(14)
Lys	AAA	0.00	(0)	0.05	(1)	Arg	AGA	0.00	(0)	0.18	(1)
	AAG	0.00	(0)	1.95	(38)		AGG	0.00	(0)	2.18	(12)
Asp	GAU	0.00	(0)	0.23	(4)	Gly	GGU	0.00	(0)	0.30	(4)
	GAC	0.00	(0)	1.77	(31)		GGC	0.00	(0)	2.22	(30)
Glu	GAA	0.00	(0)	0.16	(3)		GGA	0.00	(0)	0.52	(7)
	GAG	0.00	(0)	1.84	(34)		GGG	0.00	(0)	0.96	(13)

第3节：蛋白质序列特征分析

- 蛋白质是组成生物体的基本物质，是生命活动的主要承担者，一切生命活动都与蛋白质有关。
- 虽然遗传信息的携带者是核酸，但遗传信息的传递和表达不仅要在酶的催化之下，并且也是在各种蛋白质的调节控制下进行的。
- 因此，分析处理蛋白质序列数据的重要性并不亚于分析DNA序列数据。
- 蛋白质的生物功能由蛋白质的结构所决定，因此在研究蛋白质的功能时需要了解蛋白质的空间结构。

目前一种基本认可的假设是：

蛋白质的空间结构由蛋白质序列所决定，即我们可以根据蛋白质序列预测蛋白质结构，这是第二遗传密码的问题，也是一个更为复杂的问题，因为蛋白质序列和蛋白质空间结构之间的关系要比DNA序列和蛋白质序列之间的关系复杂得多。

因此我们需要分析大量的数据，从中找出蛋白质序列和蛋白质结构间存在的关系和规律。

一、蛋白质的理化性质

蛋白质是由氨基酸组成的大分子化合物，对组成蛋白质的氨基酸进行理化性质的统计分析是对一个未知蛋白质进行分析的基础。

蛋白质的理化性质包括蛋白质的分子量、氨基酸的组成、等电点、消光系数、亲水性和疏水性、跨膜区、信号肽、翻译后修饰位点等。

利用ProtParam分析蛋白质的理化性质

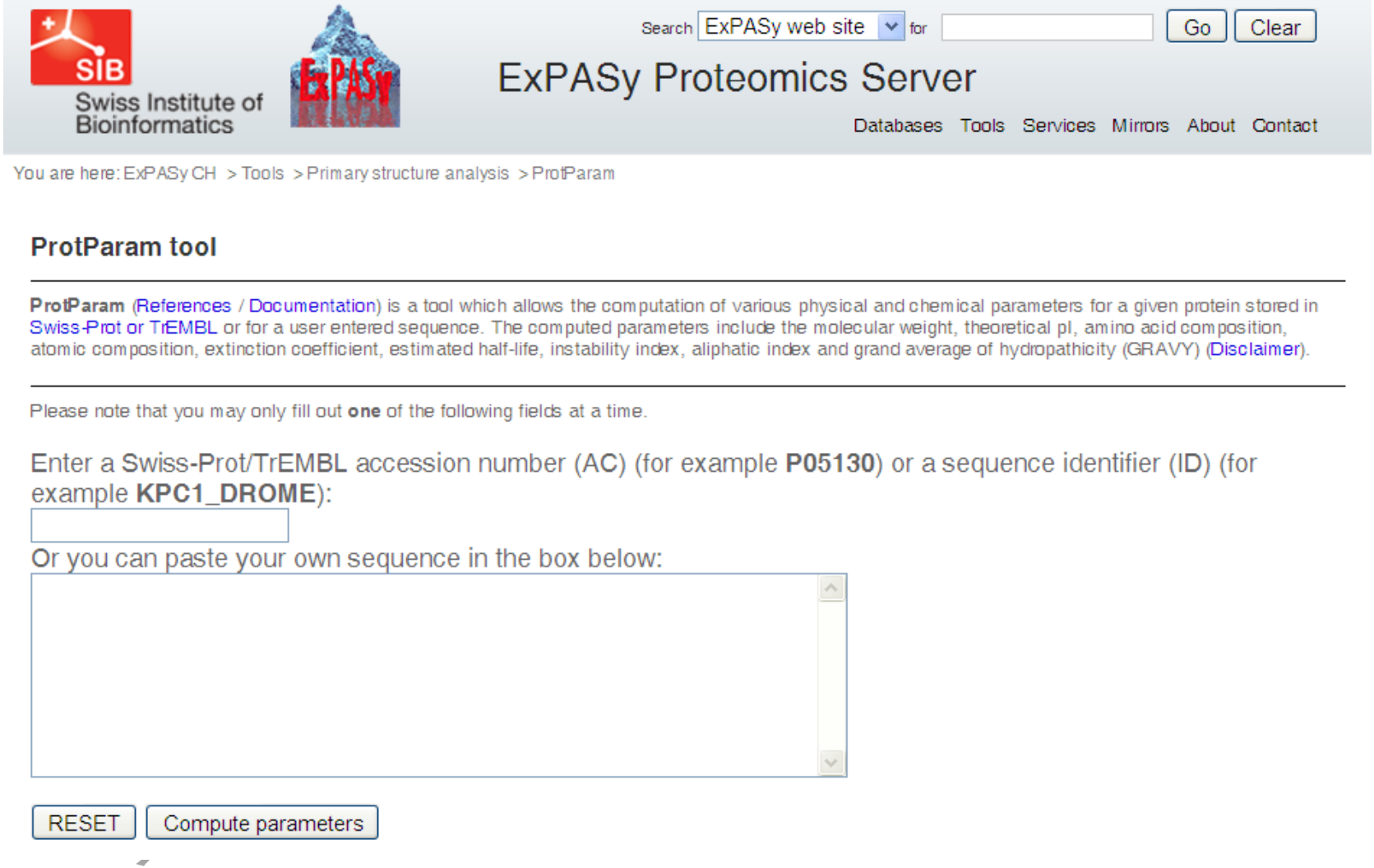
ExPASy（Expert Protein Analysis System）是由瑞士生物信息学中心维护，并与欧洲生物信息学中心（EBI）及蛋白质信息资源（protein in formation resource，PIR）组成Universal Protein Knowledgebase联盟。

ExPASy数据库提供了一系列蛋白质理化分析工具，以便于检索未知蛋白质的理化性质，并基于这些理化性质鉴别未知蛋白质的类别，为后续实验提供帮助。其中ProtParam（physico-chemical parameters of a protein sequence ）就是计算氨基酸理化参数常用的在线工具。

其网址为：

http://expasy.org/tools/protparam.html

ProtParam在线页面



用ProtParam分析G00016序列理化性质的结果

Number of amino acids: 157 ← 氨基酸残基数

Molecular weight: 18191.9

Theoretical pI: 8.43 ← 理论等电点

Amino acid composition: CSV format

Ala (A) 12 7.6%
Arg (R) 11 7.0%
⋮
Val (V) 11 7.0%

Total number of negatively charged residues (Asp + Glu): 19 ← 负电荷氨基酸残基总数
Total number of positively charged residues (Arg + Lys): 21 ← 正电荷氨基酸残基总数

Atomic composition:

Carbon	C	807
Hydrogen	H	1269
Nitrogen	N	223
Oxygen	O	234
Sulfur	S	11

Formula: $C_{807}H_{1269}N_{223}O_{234}S_{11}$

Total number of atoms: 2544

Extinction coefficients: ← 消光系数

Extinction coefficients are in units of M^{-1} cm^{-1}, at 280 nm measured in water.

Ext. coefficient 26025
Abs 0.1% (=1 g/l) 1.431, assuming ALL Cys residues appear as half cystines
Ext. coefficient 25900
Abs 0.1% (=1 g/l) 1.424, assuming NO Cys residues appear as half cystines

Estimated half-life:

The N-terminal of the sequence considered is E (Glu).

The estimated half-life is: 1 hours (mammalian reticulocytes, in vitro).
30 min (yeast, in vivo).
>10 hours (Escherichia coli, in vivo).

Instability index: ← 不稳定系数
The instability index (II) is computed to be 52.82
This classifies the protein as unstable.

Aliphatic index: 82.61 ← 脂肪系数

Grand average of hydropathicity (GRAVY): -0.400 ← 总平均疏水性

❖ 3.2 蛋白质的亲水性和疏水性

蛋白质的基本组成单元是氨基酸。

氨基酸通常被分为三类：

1. 疏水氨基酸（hydrophobic amino acid），其侧链大部分或者全部由碳原子和氢原子组成，因此这类氨基酸不太可能与水分子形成氢键；
2. 极性氨基酸（polar amino acid），其测链通常由氧原子或氮原子组成，它们比较容易与水分子形成氢键，因此也称为亲水氨基酸；
3. 带电氨基酸（charged amino acids），这类氨基酸在生物pH环境中带有正电或负电。

蛋白质的亲水性或疏水性

- 具有大的正的ΔGt的AA侧链
  - 疏水性强
  - 优先选择处在有机相
  - 倾向于配置在蛋白质分子的内部
- 具有负的ΔGt的AA侧链
  - 亲水性的
  - 配置在蛋白质分子的表面

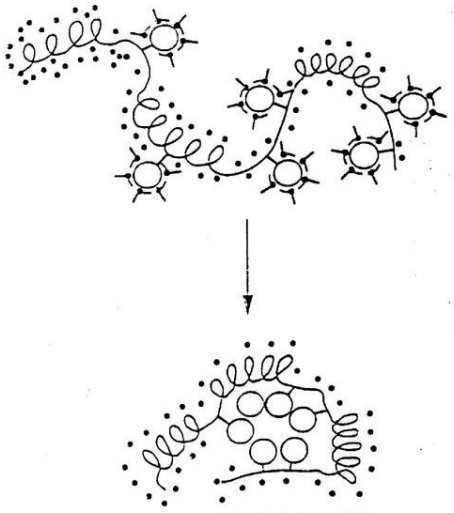


图 2-11 球状蛋白质的疏水相互作用

空心圆圈代表疏水基团，围绕着空心圆圈的"L-状"分子是疏水表面定向的水分子，小黑点代表与极性基团缔合的水分子

氨基酸的亲疏水性是构成蛋白质折叠的主要驱动力，一般通过亲水性分布图（hydropathy profile）反映蛋白质的折叠情况。蛋白质折叠时会形成疏水内核和亲水表面，同时在潜在跨膜区出现高疏水值区域，据此可以测定跨膜螺旋等二级结构和蛋白质表面氨基酸分布。

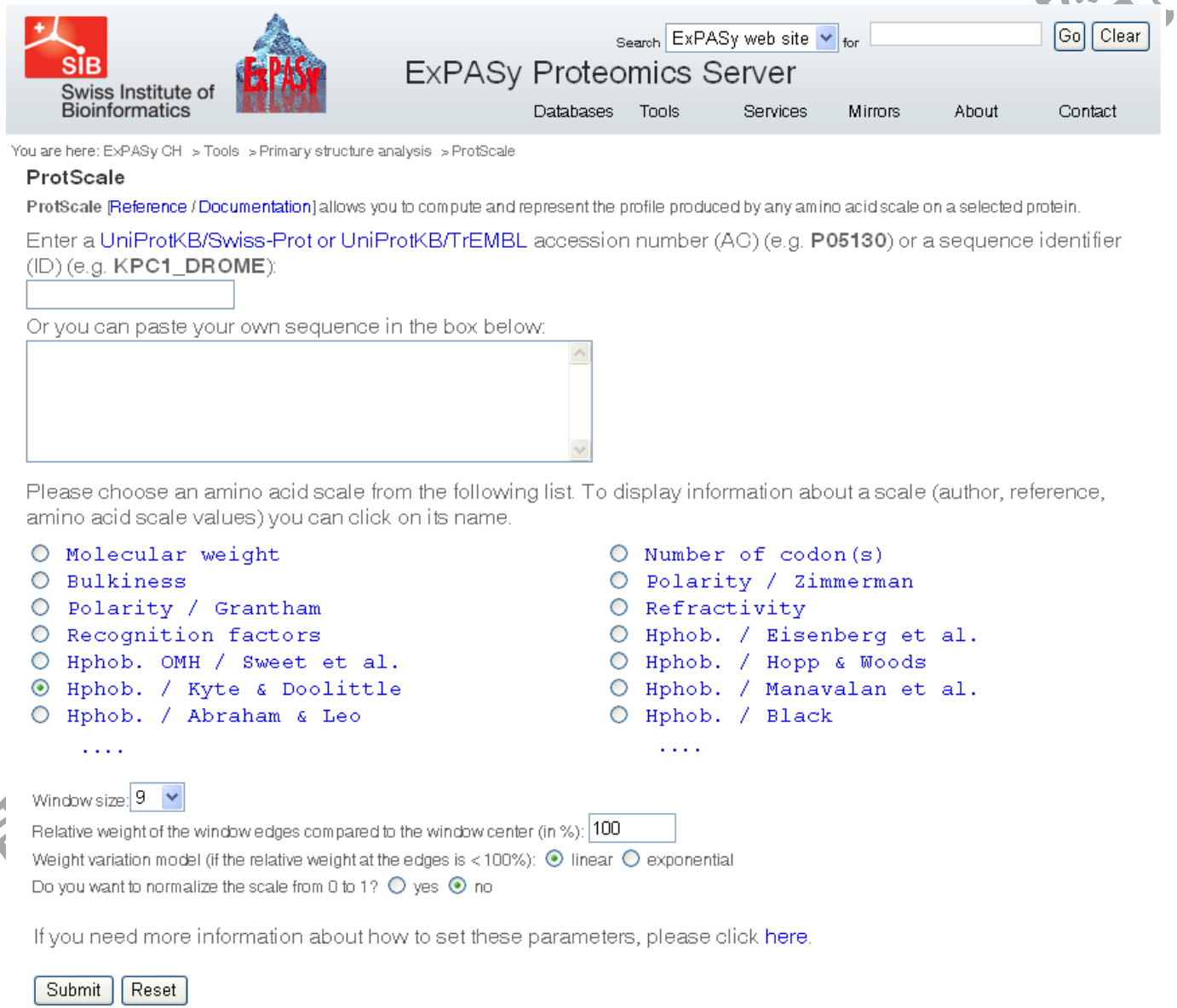
利用ProtScale分析蛋白质的亲水性或疏水性

ExPASy的ProtScale程序是计算蛋白质亲疏水性分析的在线工具。

其网址为：

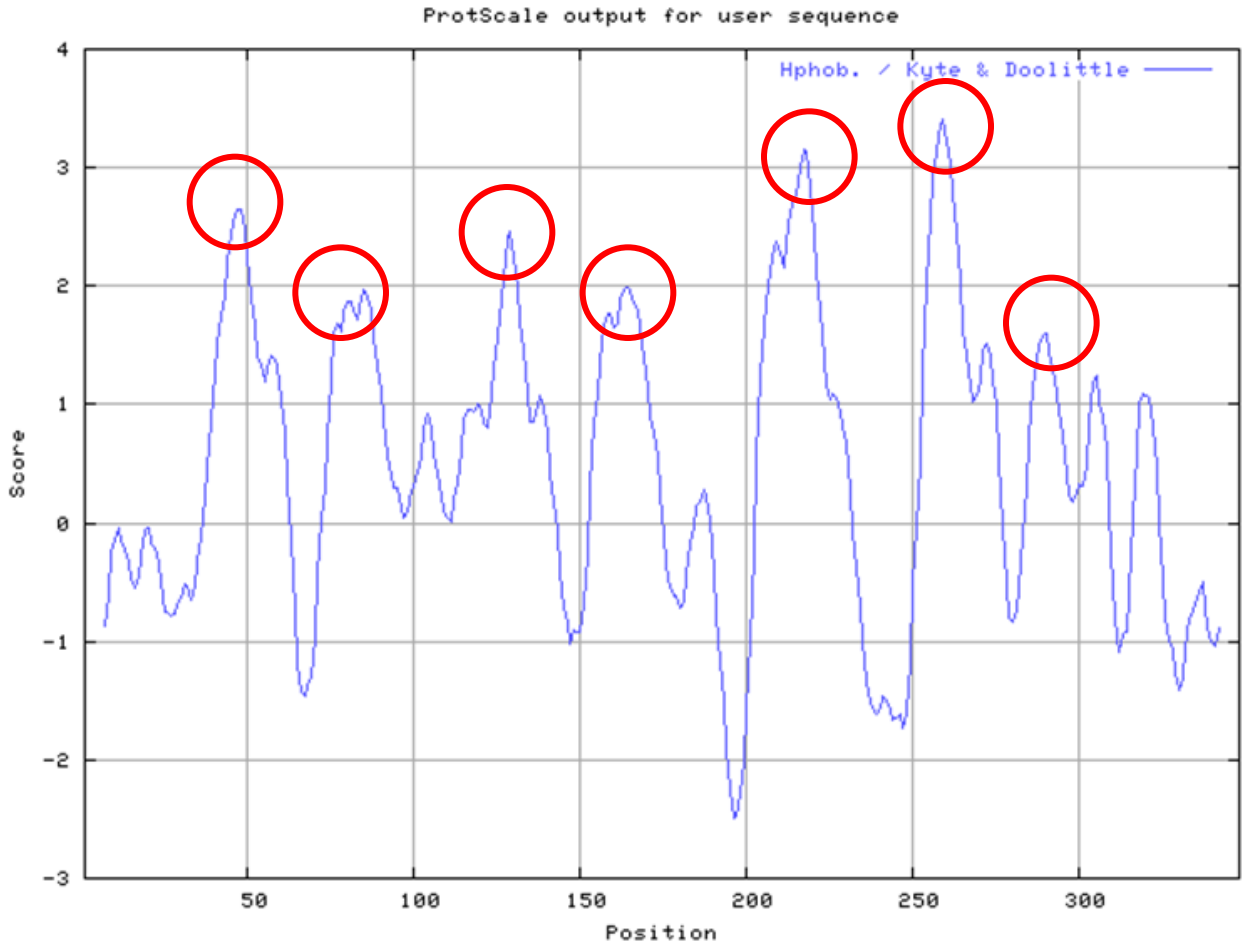
http://expasy.org/tools/protscale.html

ProtScale在线页面



用ProtScale分析P02699序列疏水性结果的图形显示

MIN: -2.487
MAX: 3.407



The results of your ProtScale query are available in the following formats:

- Image in GIF-format
- Image in Postscript-format
- Numerical format (verbose)
- Numerical format (minimal, to be exported into an external application)

其他格式的结果

Hohob./Kyte & Doolittle标度

Using the scale **Hphob. / Kyte & Doolittle**, the individual values for the 20 amino acids are:
(The values in parentheses are the original values, the normalized values have been used in the computation.)

Ala: 0.700 (1.800)	Arg: 0.000 (-4.500)	Asn: 0.111 (-3.500)
Asp: 0.111 (-3.500)	Cys: 0.778 (2.500)	Gln: 0.111 (-3.500)
Glu: 0.111 (-3.500)	Gly: 0.456 (-0.400)	His: 0.144 (-3.200)
Ile: 1.000 (4.500)	Leu: 0.922 (3.800)	Lys: 0.067 (-3.900)
Met: 0.711 (1.900)	Phe: 0.811 (2.800)	Pro: 0.322 (-1.600)
Ser: 0.411 (-0.800)	Thr: 0.422 (-0.700)	Trp: 0.400 (-0.900)
Tyr: 0.356 (-1.300)	Val: 0.967 (4.200)	: 0.111 (-3.500)
: 0.111 (-3.500)	: 0.446 (-0.490)	

用Window size=13时计算窗口内每个位置上氨基酸的标度权值

Weights for window positions 1,..,13, using **linear weight variation model**:

1	2	3	4	5	6	7	8	9	10	11	12	13
0.10	0.25	0.40	0.55	0.70	0.85	1.00	0.85	0.70	0.55	0.40	0.25	0.10
edge						center						edge

❖ 3.3 蛋白质的跨膜区

生物膜所含的蛋白质叫膜蛋白，是生物膜功能的主要承担者。根据蛋白质分离的难易及在膜中分布的位置，膜蛋白基本可分为两大类：外在膜蛋白和内在膜蛋白。

外在膜蛋白约占膜蛋白的20%～30%，分布在膜的内外表面，主要在内表面，为水溶性蛋白，它通过离子键、氢键与膜脂分子的极性头部相结合，或通过与内在蛋白质的相互作用间接与膜结合；

内在膜蛋白约占膜蛋白的70%～80%，是双亲媒性分子，可不同程度的嵌入脂双层分子中。有的贯穿整个脂双层，两端暴露于膜的内外表面，这种类型的膜蛋白又称跨膜蛋白。

蛋白质的跨膜区

内在膜蛋白露出膜外的部分含较多的极性氨基酸，属亲水性，与磷脂分子的亲水头部邻近；嵌入脂双层内部的膜蛋白由一些非极性的氨基酸组成，与脂质分子的疏水尾部相互结合，因此与膜结合非常紧密。所以，对膜蛋白的跨膜区进行预测是生物信息学的重要应用。

利用TMpred分析蛋白质的跨膜区

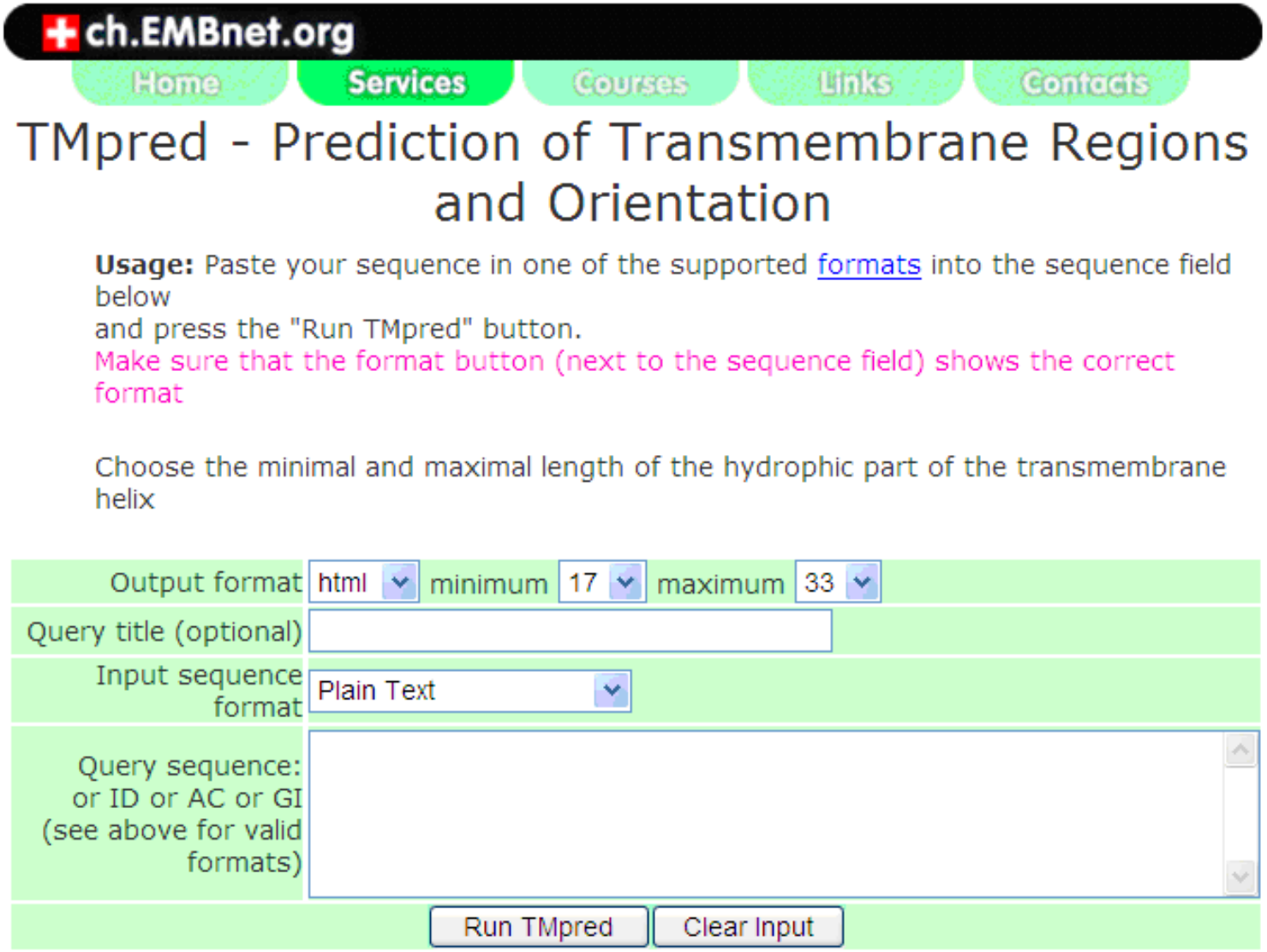
TMpred是EMBnet开发的一个分析蛋白质跨膜区的在线工具，TMpred基于对TMbase数据库的统计分析来预测蛋白质跨膜区和跨膜方向。Tmbase数据库来源于Swiss-Prot库，并包含了每个序列的一些附加信息，如：跨膜结构区域的数量、跨膜结构域的位置及其侧翼序列的情况。

TMpred利用这些信息并与若干加权矩阵结合来进行预测。

其网址为：

http://www.ch.embnet.org/software/TMPRED_form.html

TMpred在线网页


ch.EMBnet.org
Home
Services
Courses
Links
Contacts
TMpred - Prediction of Transmembrane Regions and Orientation
Usage: Paste your sequence in one of the supported formats into the sequence field below
and press the "Run TMpred" button.
Make sure that the format button (next to the sequence field) shows the correct format
Choose the minimal and maximal length of the hydrophic part of the transmembrane helix
Output format
html
minimum
17
maximum
33
Query title (optional)
Input sequence format
Plain Text
Query sequence: or ID or AC or GI (see above for valid formats)
Run TMpred
Clear Input

用TMpred分析P51684序列所得到的可能的7个跨膜螺旋区

1.) Possible transmembrane helices

The sequence positions in brackets denominate the core region.
Only scores above 500 are considered significant.

Inside to outside helices : 7 found

from		to		score	center
47	(51)	69	(69)	2494	61
83	(86)	104	(104)	1914	94
123	(123)	141	(139)	1352	131
166	(168)	184	(184)	2170	176
219	(219)	236	(236)	2453	227
255	(255)	276	(273)	2140	265
300	(300)	319	(319)	915	309

Outside to inside helices : 7 found

from		to		score	center
55	(55)	74	(71)	2707	63
84	(86)	104	(104)	1470	94
120	(123)	141	(139)	1451	131
166	(166)	185	(185)	1934	176
212	(214)	235	(232)	2530	224
252	(258)	274	(274)	1386	266
299	(299)	319	(319)	1299	309

用TMpred分析P51684序列所得到的7个可能的跨膜螺旋区的相关性列表

2.) Table of correspondences

Here is shown, which of the inside->outside helices correspond to which of the outside->inside helices.

Helices shown in brackets are considered insignificant.
A "+"-symbol indicates a preference of this orientation.
A "++"-symbol indicates a strong preference of this orientation.

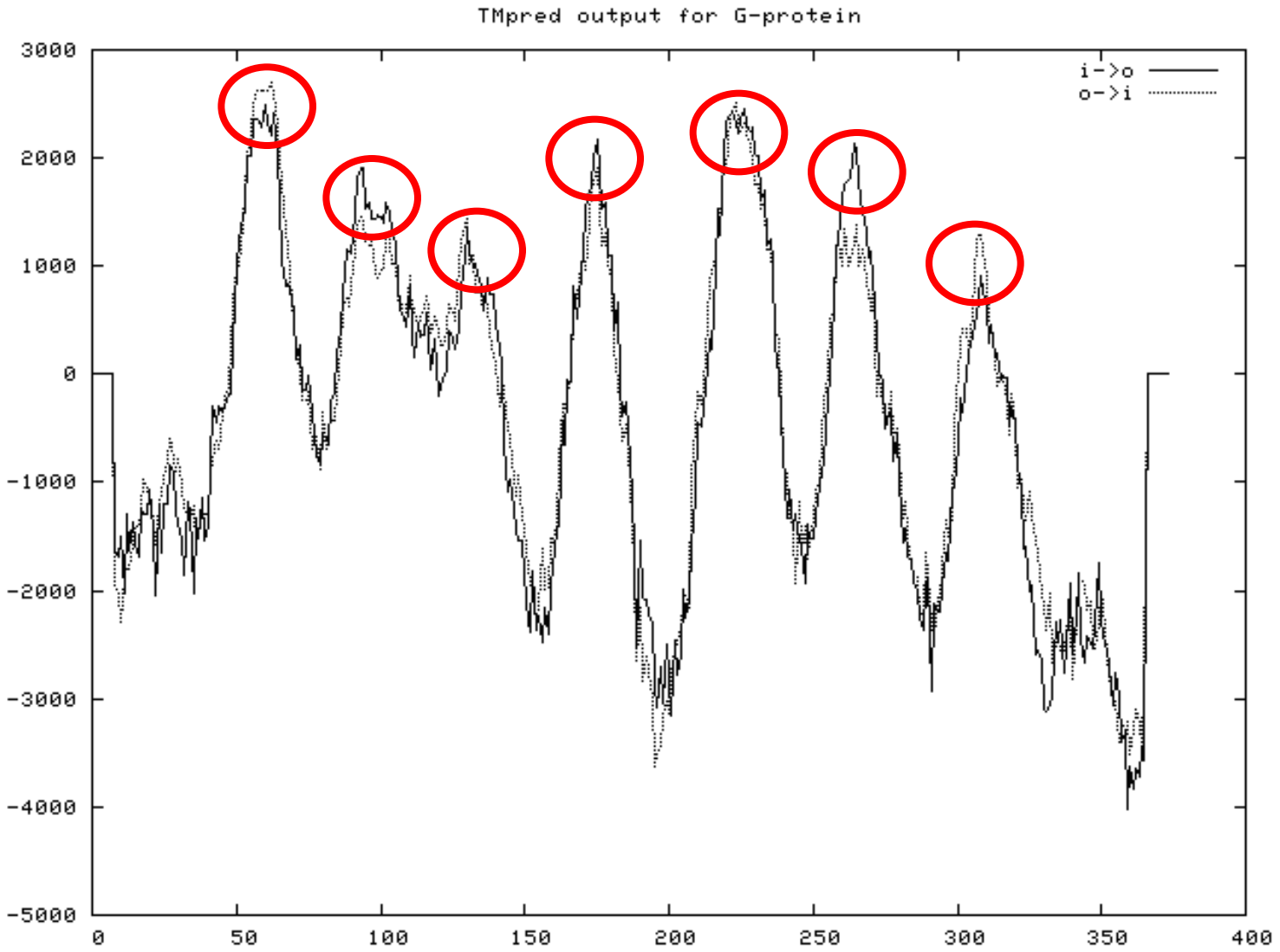
inside->outside				outside->inside			
47- 69	(23)	2494		55- 74	(20)	2707	++
83- 104	(22)	1914	++	84- 104	(21)	1470	
123- 141	(19)	1352		120- 141	(22)	1451	+
166- 184	(19)	2170	++	166- 185	(20)	1934	
219- 236	(18)	2453		212- 235	(24)	2530	
255- 276	(22)	2140	++	252- 274	(23)	1386	
300- 319	(20)	915		299- 319	(21)	1299	++

用TMpred分析P51684序列所得到的7个可能的跨膜螺旋区的建议的跨膜拓扑模型

3.) Suggested models for transmembrane topology

```
2 possible models considered, only significant TM-segments used

-----> STRONGLY prefered model: N-terminus outside
 7 strong transmembrane helices, total score : 14211
 # from   to length score orientation
 1   55   74 (20)    2707 o-i
 2   83  104 (22)    1914 i-o
 3  120  141 (22)    1451 o-i
 4  166  184 (19)    2170 i-o
 5  212  235 (24)    2530 o-i
 6  255  276 (22)    2140 i-o
 7  299  319 (21)    1299 o-i

------> alternative model
 7 strong transmembrane helices, total score : 12004
 # from   to length score orientation
 1   47   69 (23)    2494 i-o
 2   84  104 (21)    1470 o-i
 3  123  141 (19)    1352 i-o
 4  166  185 (20)    1934 o-i
 5  219  236 (18)    2453 i-o
 6  252  274 (23)    1386 o-i
 7  300  319 (20)     915 i-o
```

用TMpred分析P51684序列所得到的7个可能的跨膜螺旋区的图形显示结果



❖ 3.4 蛋白质序列中的信号肽

信号肽是指新合成多肽链中用于指导蛋白质跨膜转移的末端（通常为N末端）的氨基酸序列。信号肽中至少含有一个带正电荷的氨基酸，中部有一个高度疏水区以通过细胞膜。

信号肽假说认为，编码分泌蛋白的mRNA在翻译时首先合成的是N末端带有疏水氨基酸残基的信号肽，它被内质网膜上的受体识别并与之相结合。信号肽经由膜中蛋白质形成的孔道到达内质网内腔，随机被位于腔表面的信号肽酶水解，由于它的引导，新生的多肽就能够通过内质网膜进入腔内，最终被分泌到胞外。

蛋白质的前导肽—leader Peptide

前导肽是信号肽的一种。在线粒体蛋白质的跨膜转运过程中，通过线粒体膜的蛋白质在转运之前大多数以前体形式存在，它由成熟蛋白质和N端延伸出的一段前导肽共同组成。迄今已有40多种线粒体蛋白质前导肽的一级结构被阐明，它们约含20~80个氨基酸残基，当前体蛋白跨模时，前导肽被一种或两种多肽酶所水解转变成为成熟蛋白质，同时失去继续跨膜的能力。

前导肽一般具有以下特性：（1）带正电荷的碱性氨基酸（特别是精氨酸）含量较为丰富，它们分散于不带电荷的氨基酸序列之间；（2）缺失带负电荷的酸性氨基酸；（3）羟基氨基酸（特别是丝氨酸）含量较高；（4）有形成两亲（既有亲水又有疏水部分）α-螺旋结构的能力。

利用SignalP分析蛋白质的前导肽

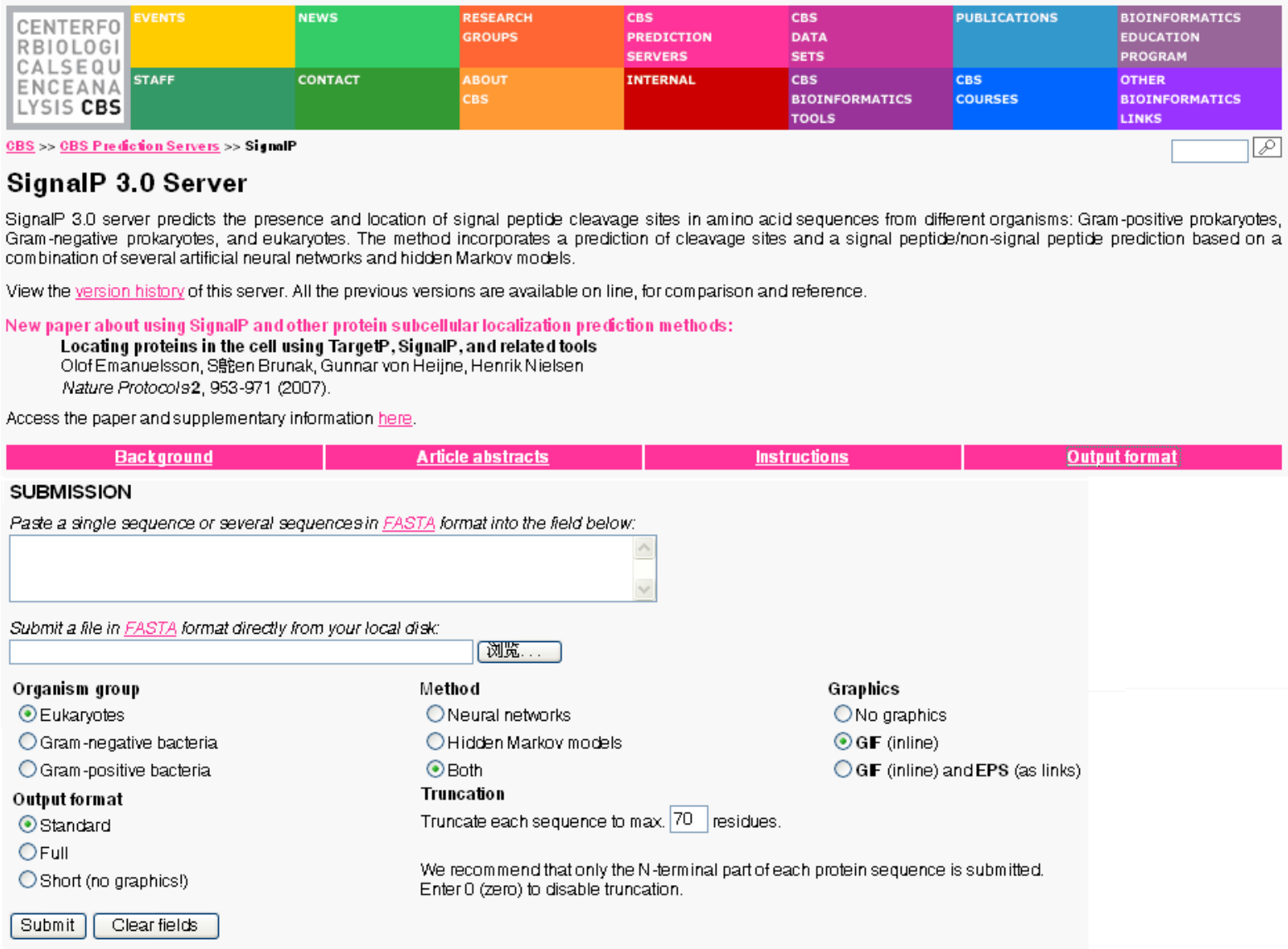
SignalP是丹麦技术大学的生物序列分析中心开发的信号肽及其剪切位点检测的在线工具，该软件基于神经网络方法，用已知信号序列的革兰氏阴性原核生物、革兰氏阳性原核生物及真核生物的序列分别作为训练集。

SignalP预测的是分泌型信号肽，而不是那些参与细胞内信号传递的蛋白。

其网址为：

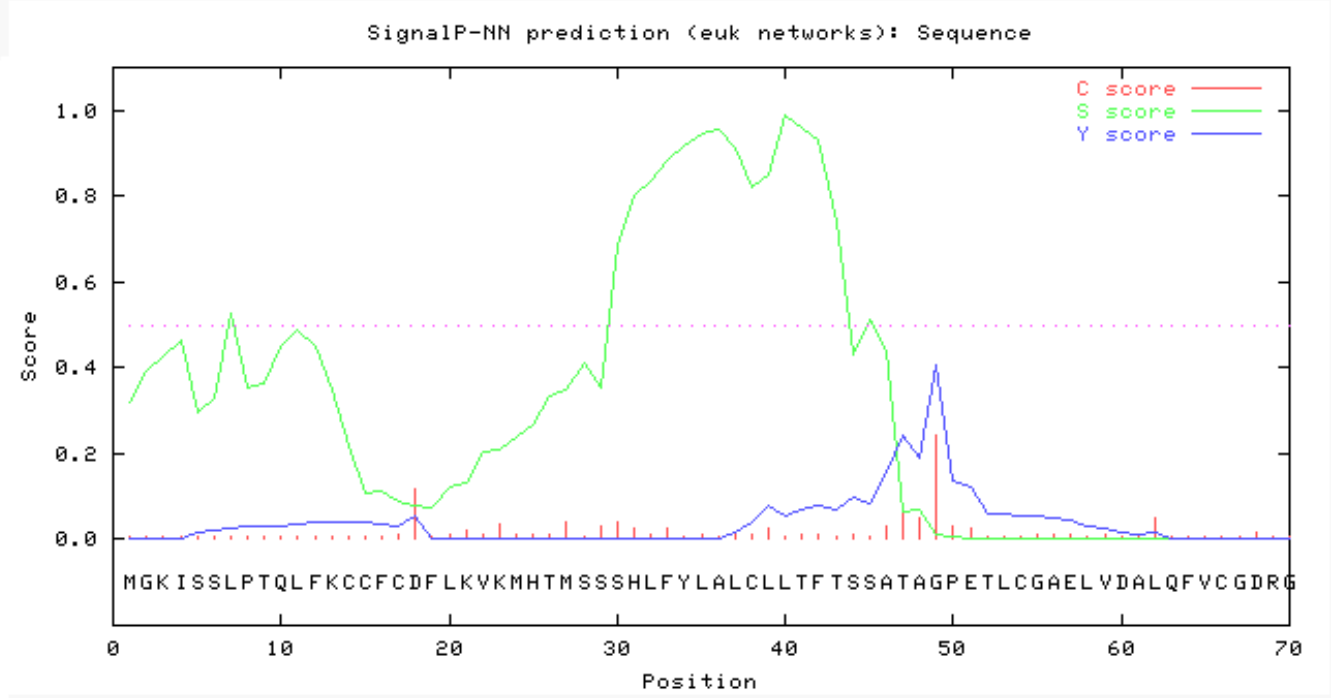
http://genome.cbs.dtu.dk/services/SignalP/

SignalP在线网页



用SignalP（神经网络方法）分析P05019序列前导肽的结果

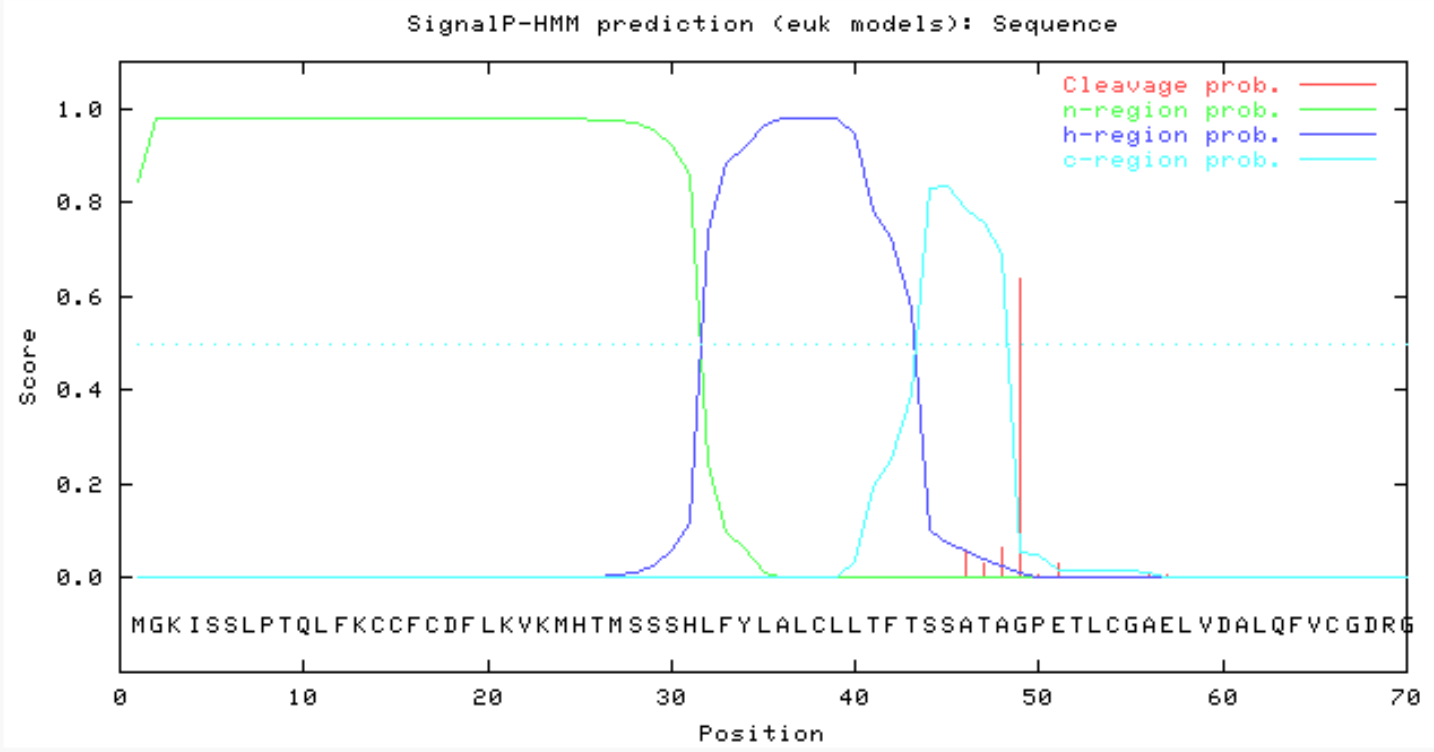
SignalP-NN result:




```
# data

>Sequence                  length = 70
# Measure  Position  Value  Cutoff  signal peptide?
  max. C     49       0.241   0.32   NO
  max. Y     49       0.406   0.33   YES
  max. S     40       0.989   0.87   YES
  mean S      1-48    0.464   0.48   NO
       D      1-48    0.435   0.43   YES
# Most likely cleavage site between pos. 48 and 49: ATA-GP
```

用SignalP（隐马尔可夫方法）分析P05019序列前导肽的结果

SignalP-HMM result:




```
# data

>Sequence
Prediction: Signal peptide
Signal peptide probability: 0.844
Signal anchor probability: 0.138
Max cleavage site probability: 0.637 between pos. 48 and 49
```

❖ 3.5 蛋白质的卷曲螺旋

卷曲螺旋是蛋白质空间结构中的一种，它是由2～7个α螺旋相互缠绕而形成超螺旋结构的总称。

卷曲螺旋区域一般由7个氨基酸残基为单位组成，以a、b、c、d、e、f、g位置表示，其中a和d位置为疏水性氨基酸，而其他位置的氨基酸残基为亲水性。许多含有卷曲螺旋结构的蛋白质具有重要的生物学功能，例如基因表达调控中的转录因子。

含有卷曲螺旋结构最知名的蛋白质有原癌蛋白（oncoprotein）c-fos和jun，以及原肌球蛋（tropomyosin）。

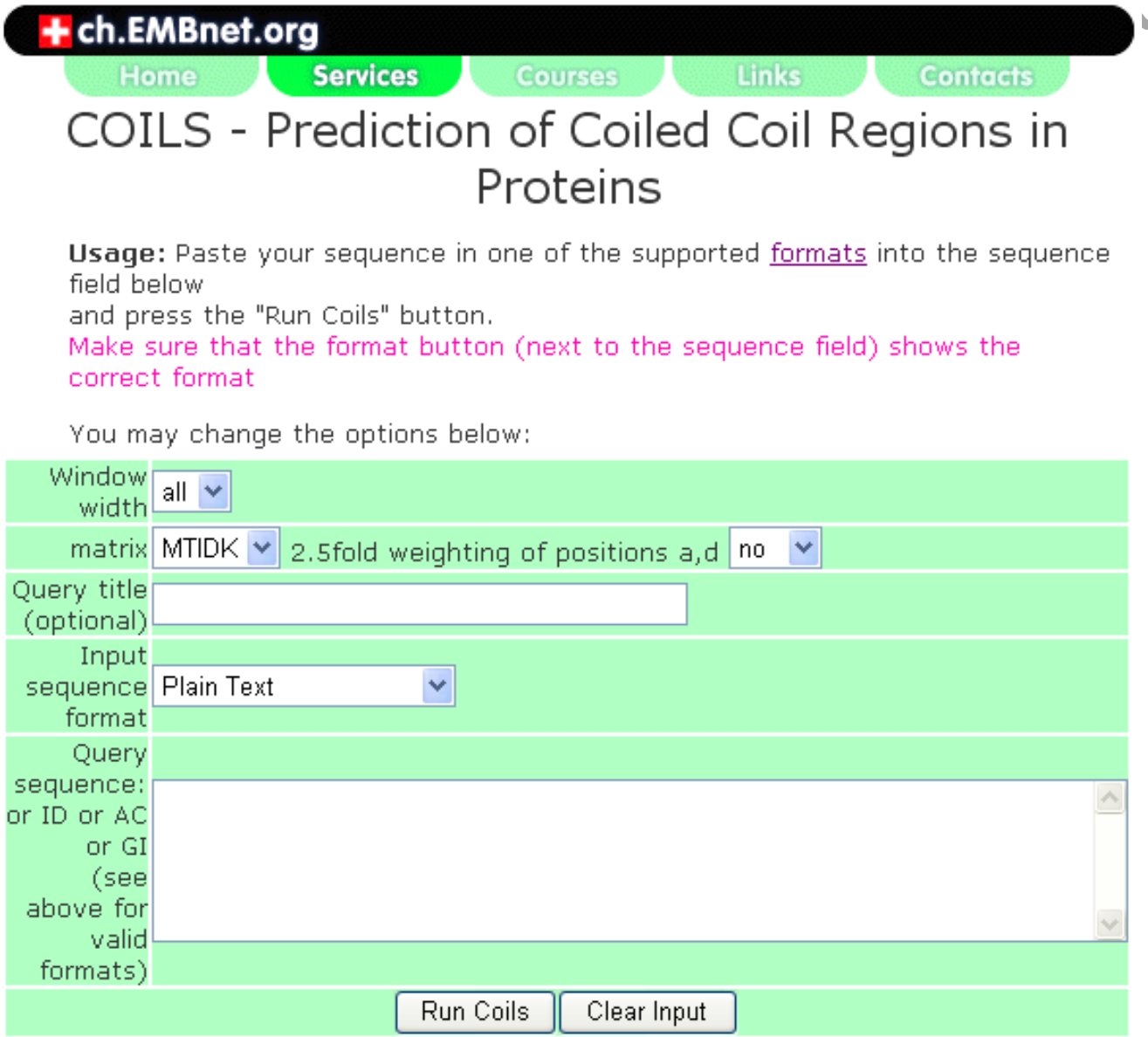
利用COILS分析蛋白质的卷曲螺旋

COILS是由Swiss EMBNet维护的预测卷曲螺旋的在线工具，该软件是基于Lupas算法，将查询序列在一个由已知包含卷曲螺旋蛋白结构的数据库中进行搜索，同时也将查询序列与包含球状蛋白序列的PDB次级库进行比较，并根据两个库搜索得分决定查询序列形成卷曲螺旋的概率。COILS也可以下载到本地进行运算。

其网址为：

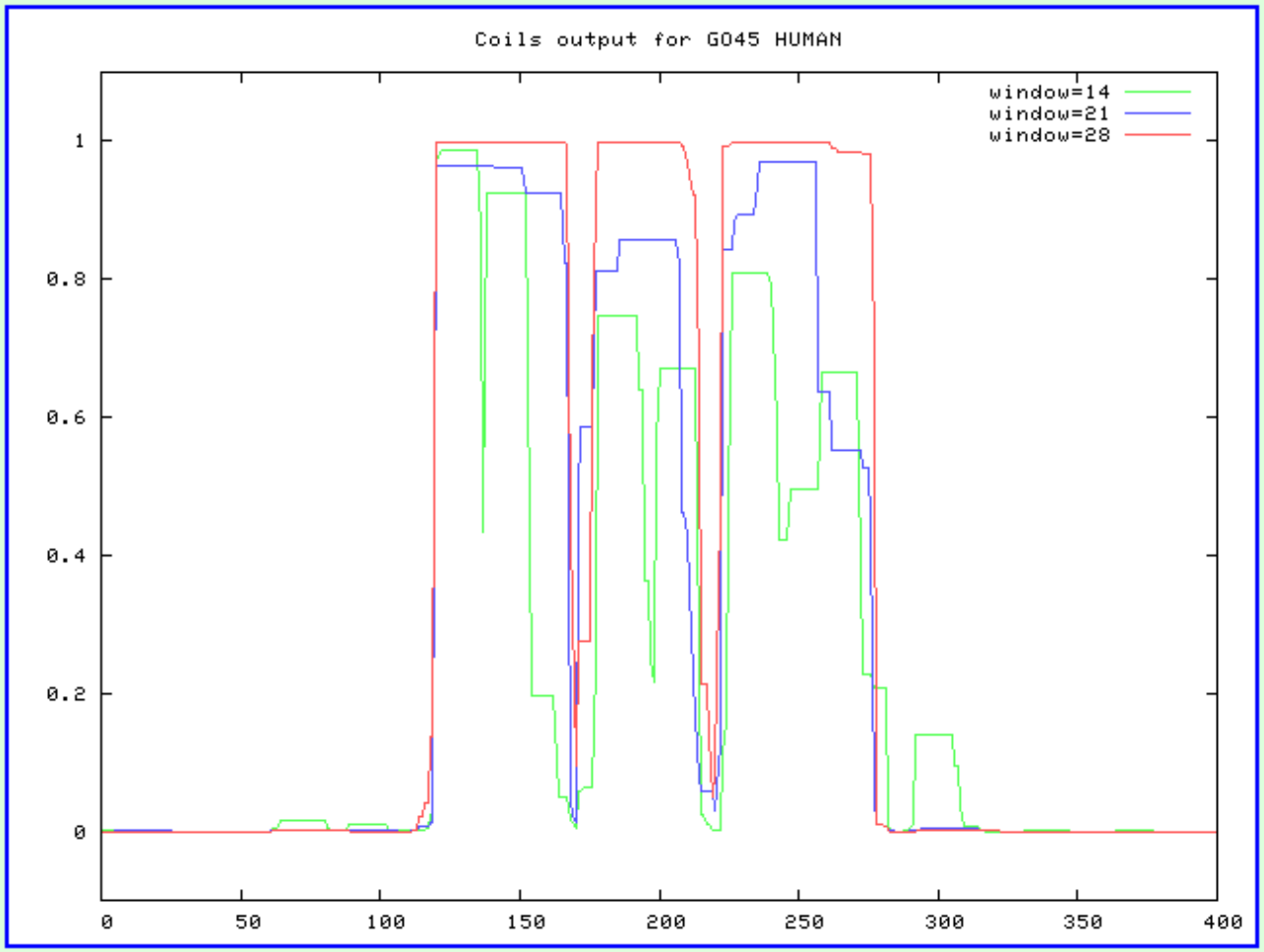
http://www.ch.embnet.org/software/COILS_form.html

COILS在线网页


ch.EMBnet.org
Home
Services
Courses
Links
Contacts
COILS - Prediction of Coiled Coil Regions in Proteins
Usage: Paste your sequence in one of the supported formats into the sequence field below
and press the "Run Coils" button.
Make sure that the format button (next to the sequence field) shows the correct format
You may change the options below:
Window width
all
matrix
MTIDK
2.5fold weighting of positions a,d
no
Query title (optional)
Input sequence format
Plain Text
Query sequence: or ID or AC or GI (see above for valid formats)
Run Coils
Clear Input

用COILS分析GO45_HUMAN卷曲螺旋的图形显示结果

```
# NCOILS version 1.0
# using MTIDK matrix
# weights: a,d=2.5 and b,c,e,f,g=1.0
# Input file is ../wwwtmp/.COILS.18722.7019.seq
#
```



用COILS分析GO45_HUMAN卷曲螺旋的文本显示结果

```
# NCOILS version 1.0
# using MTIDK matrix
# weights: a,d=2.5 and b,c,e,f,g=1.0
# Input file is ../wwwtmp/.COILS.18722.7019.seq
#
```

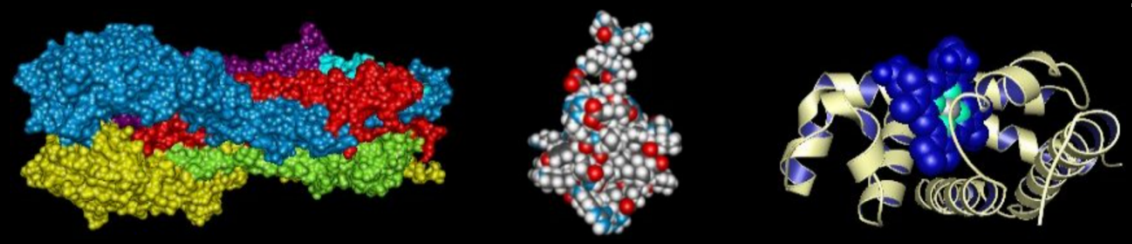
		Window=14	Window=21	Window=28
1	M	d 0.001	d 0.000	d 0.000
2	T	g 0.001	g 0.000	g 0.000
...		...	...	...
118	E	f 0.023	f 0.013	f 0.136
119	P	g 0.023	g 0.013	g 0.136
120	N	e 0.971	e 0.964	e 0.999
121	K	f 0.978	f 0.964	f 0.999
...		...	...	...
151	L	a 0.924	a 0.963	a 0.999
152	R	b 0.924	b 0.926	b 0.999
153	V	c 0.677	c 0.926	c 0.999
154	Q	d 0.196	d 0.926	d 0.999
...		...	...	...
164	L	g 0.049	g 0.926	g 0.999
165	L	a 0.049	a 0.926	a 0.999
166	V	b 0.049	b 0.822	b 0.997
167	A	c 0.049	c 0.822	c 0.994
168	S	d 0.016	d 0.043	d 0.436
169	V	b 0.012	b 0.022	b 0.317
...		...	...	...

❖ 3.6 蛋白质序列的综合分析

对蛋白质的研究是生物化学领域一个非常重要的部分。随着人类基因组计划的实施和完成，得到了大量的蛋白质序列数据。

但是，面对如此众多的蛋白质序列数据，其分析工作是一个非常困难的工作。用人工的方法是不可能完成如此大量的分析工作的。运用计算机，利用一定的运算规则，进行蛋白序列分析是唯一的方法。

蛋白序列分析软件包**Antheprot**正是这样的一个程序。



ANTHEPROT 6.9.3

- Home
- Snapshots
- User manual
- Videos

Downloads

- ANTHEPROT
- ANTHEPROT 3D

Contact

- Address
- Send a mail
- Author web site

About

- Citations
- Acknowledgements

Links

- ANTHEPROT 3D gallery
- Dicroprot
- IBCP

ANTHEPROT is a PC software for protein analysis.

2010-05-12 **Actualité** ANTHEPROT 3D
ANTHEPROT 3D is a molecular graphics program intended for the visualisation of proteins, nucleic acids from RCSB archive. The program is aimed at display, teaching and generation of publication quality images.

The 3D module of ANTHEPROT 3D has been completely written in OPENGL for maximal graphic speed. It is now available as a standalone program but can be still associated with the general ANTHEPROT plateform of protein sequence analysis. An image gallery and video are available which demonstrate the software graphic capabilities. Antheprot documentation file is being rewritten so as to fit with ANTHEPROT last version.

3D zalmann

PRABI-IBCP 7, passage du Vercors 69367 Lyon, FRANCE © G.Deléage (2010- 2022) 313392 visits Sun 16 Oct 2022

Homepage of Antheprot software (v6.9.3)

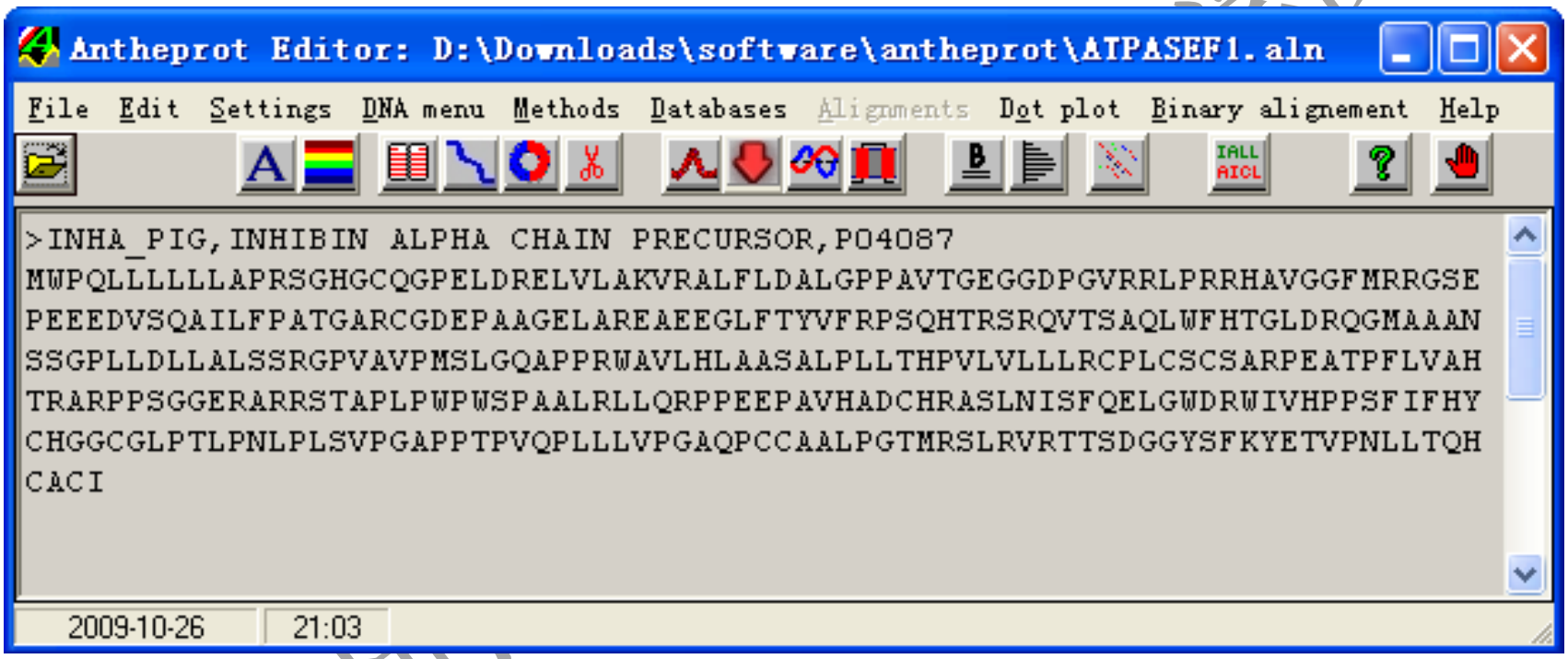
http://antheprot-pbil.ibcp.fr/

蛋白质序列分析软件包：Antheprot

Antheprot是位于法国的蛋白质生物与化学研究院用十多年时间开发出的蛋白质研究软件包，它包括了蛋白质研究领域所包括的大多数内容，功能非常强大。

Antheprot的原始网站：http://antheprot-pbil.ibcp.fr/，我们可以到这个网站上下载软件包，软件包为一个自解压执行文件，文件名为**Antheprot.exe**，大小为**51.3M**。执行此文件，输入解压后存放的目录名，便可将所有文件解压在此目录下。主程序名为**Anthepro**，双击主程序名就可以打开**Antheprot_2000**的主窗口。通过主程序，我们可以输入蛋白序列，对序列进行编辑、打印、拷贝、改变设置等操作，更重要的是，我们可以在此调用各种所需的分析工具，对蛋白序列进行分析。

Antheprot主窗口



Antheprot主窗口中各按键的含义

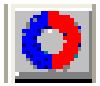
Open file，打开文件；

Change text font，更改字体、字型和大小；

Change text color，更改选定区域内字的颜色；

Sequence information，序列信息，计算蛋白质序列的分子量、比溶、各氨基酸残基的百分比组成；

Titration curve，滴定曲线，计算蛋白质序列滴定曲线与等电点；

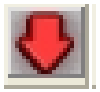


Helical wheel projection，选定序列的一个片段后，绘制Helical wheel图；

Antheprot主窗口中各按键的含义

Prediction of cleavage site for signal peptide，
预测信号肽的剪切位点；

Secondary structure prediction by all，
预测蛋白质序列的二级结构；



PROSITE site / signature detection，在蛋白质序列中查找符合PROSITE数据库的特征序列；

Physico-chemical profiles，
绘制蛋白质序列的理化特性曲线；

Pridict transmembrane region，预测跨膜区；

Similarity search with Blast，用Blast方法在选择的数据库中查找相似序列；

Antheprot主窗口中各按键的含义

Similarity search with Fasta，用Fasta方法在选择的数据库中查找相似序列；

Dot Matrix Plot，进行点阵图分析；

Multiple alignment，多序列比对；

Binary alignment（BINALIGN），在当前蛋白质序列中查找符合Prosites数据库的特征序列；

Help，打开一个简单的帮助文件；

Quit，推出程序。

Antheprot基本功能

1. 编辑（edit）
2. 参数设置（ setting）
3. 方法选择（methods）
4. 数据库（ database）

第4节：生物信息分子序列的综合分析

几个重要软件：

- Blast
- EMBOSS
- Clustal Omiga
- DNAStar
- Vector NTI
- R/Bioconductor

❖ 4.1 EMBOSS软件包

EMBOSS—European Molecular Biology Open Software Suite

EMBOSS软件包是一个开源的序列分析软件包，该软件包源于1988年开始开发的EGCG系统，整合了目前可以获得的大部分序列分析软件，并有一套专门设计的C语言库函数。

该软件包含160多个小型程序，能够完成自动识别处理不同格式存储的数据，可以通过互联网提取数据，能很好地进行序列模体（motif）、关键词同源性数据库搜索，进行序列比较、进化分析、序列两级结构分析、限制性酶切图谱分析、引物设计、序列模式识别与翻译、片段拼接等工作。EMBOSS遵照GPL协议，打破了商业软件包发展的传统模式，使科研工作者在自由、免费的软件世界享受功能强大的分析工具。

EMBOSS的主页网址为：http://emboss.sourceforge.net/

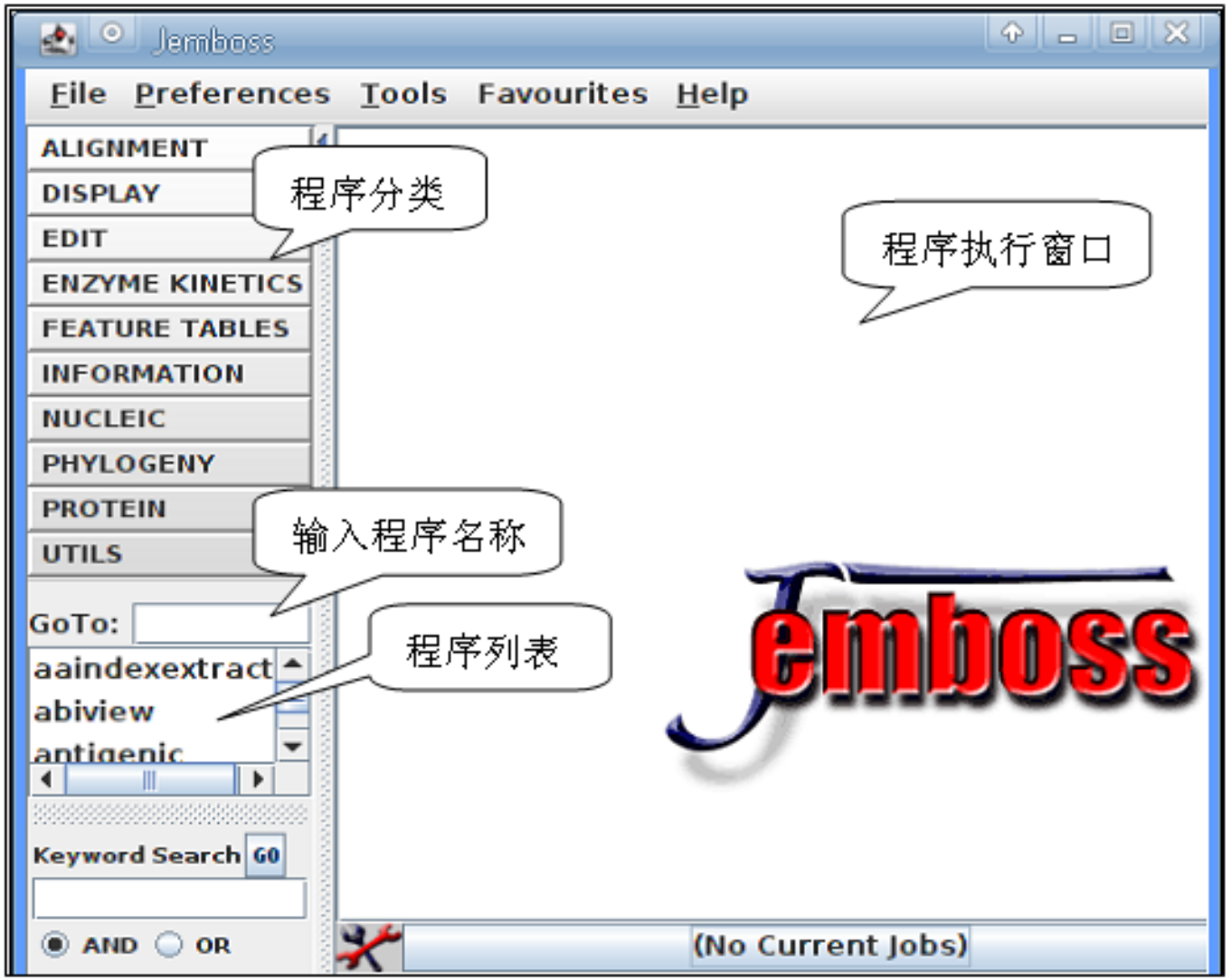
EMBOSS的运行环境

EMBOSS软件包主要运行于linux操作系统和Mac操作系统。现在基于Windows操作系统的EMBOSS也是能自由免费使用的。需要说明的是基于windows操作系统时，主要采用staden进入EMBOSS，在使用的同时，需要安装Embosswin软件。

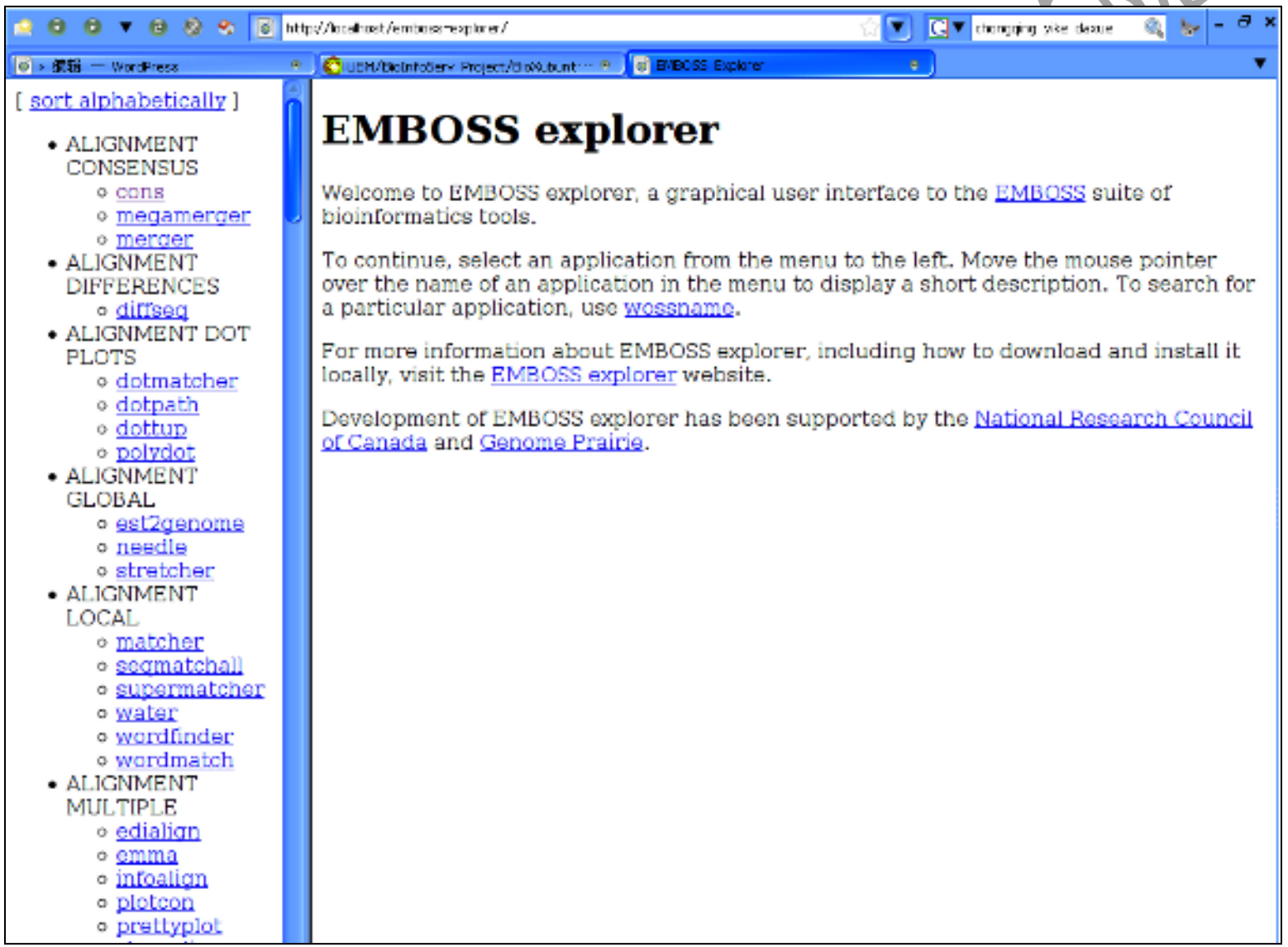
Embosswin的下载网址是：

ftp://emboss.open-bio.org/pub/EMBOSS/wEMBOSS Explorerindows/

JEMBOSS 使用界面



EMBOSS Explorer 使用界面



❖ 4.2 DNAstar软件包

DNAstar软件包可进行分子生物学中的DNA和蛋白小规模序列分析和多序列比对。

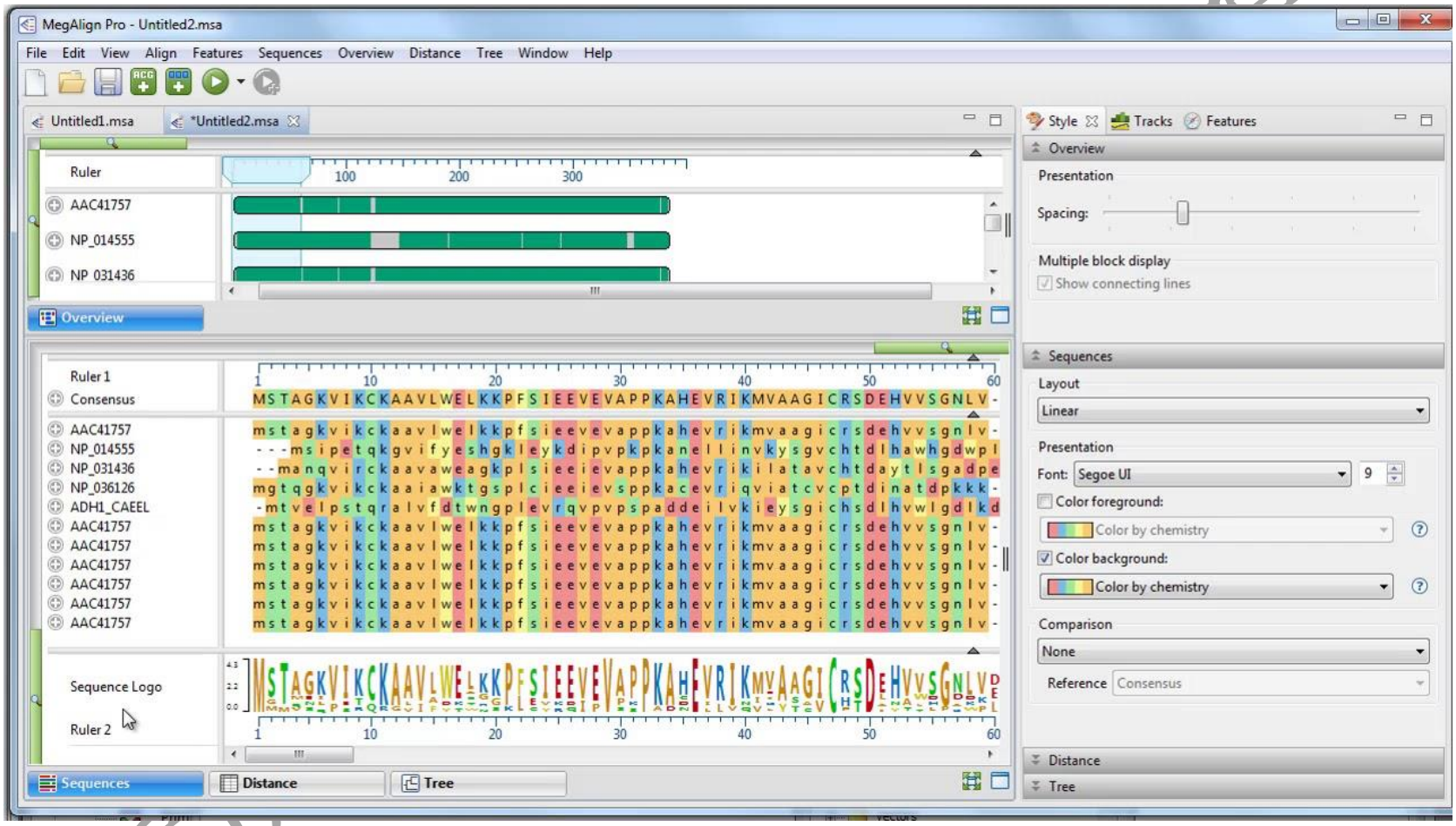
DNAstar软件包有PC Windows和Macintosh两种版本，它的一个主要功能是有7种程序可以针对不同的应用，用户可根据自己的需要进行选择。

有关DNAstar软件包更详细的信息查询网站：

http://www.dnastar.com

❖ 4.2 DNAstar软件包

http://www.dnastar.com



DNAstar软件包可进行分子生物学中的**DNA**和蛋白小规模序列分析和多序列比对。**DNAstar**有**PC Windows**和**Macintosh**两种版本，它的一个主要功能是有**7**种程序可以针对不同的应用，用户可根据自己的需要进行选择。

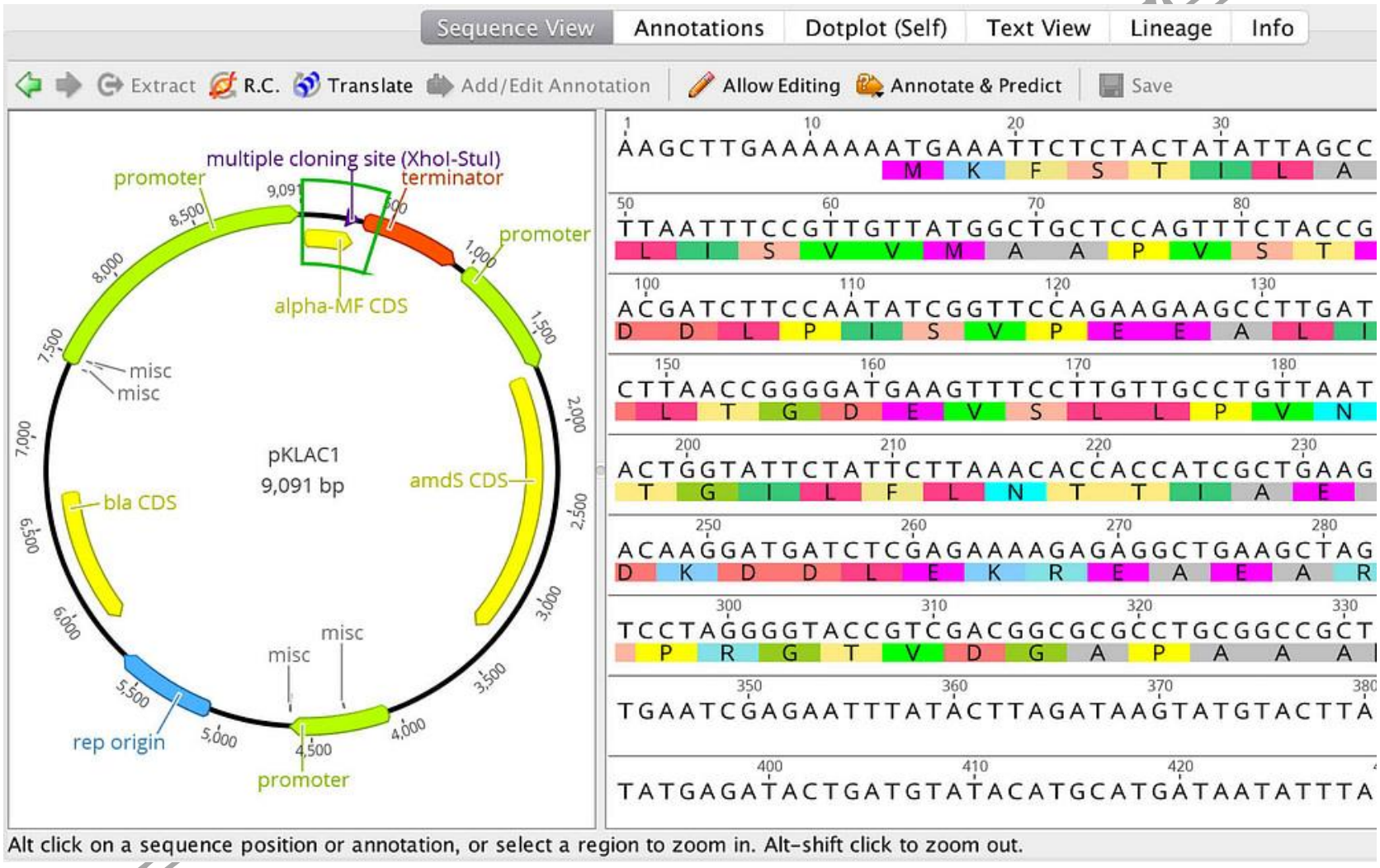
❖ 4.3 Clustal软件包

http://www.clustal.org/

Clustal Omiga是一款强大的蛋白质、核酸分析软件，可以实现对核酸序列和蛋白序列分析的大部分功能，同时它还兼有引物设计的功能。

❖ 4.4 Vector NTI软件包

http://www.informaxinc.com/



Vector NTI是Informax开发的一种高度集成、功能齐全的分子生物学应用软件，可以对DNA、蛋白质序列进行分析和操作。

Vector NTI主要功能

1. DNA序列的开放阅读框、序列模式、功能区搜索、限制酶图谱、蛋白质翻译。
2. PCR引物、测序引物、杂交探针的设计和评价。
3. DNA测序片断的拼接。
4. 同源比较和系统发育树构建。
5. 蛋白质结构预测：三维结构、化学键、翻译后修饰位点、结构域等。
6. 模拟电泳：琼脂糖电泳、PAGE。

重庆师范大学
CHONG QING NORMAL UNIVERSIYT

Thanks for your attention!

Acknowledgement

College of Life Sciences, Chongqing Normal University

2022, Chongqing of P. R. C