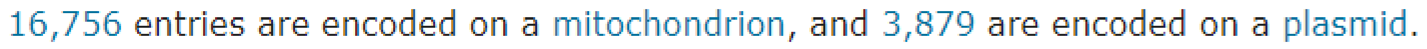
PRACTICAL 3
ANALYSIS OF SEQUENCE CHARACTERISTICS

2021.4

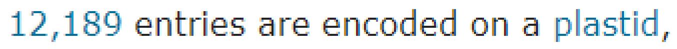
PREVIEW

- Review: References should be listed.

- https://www.uniprot.org/statistics/Swiss-Prot



16,756 entries are encoded on a mitochondrion, and 3,879 are encoded on a plasmid.



12,189 entries are encoded on a plastid,

When you have a sequence

- Is it likely to be a gene?
- What is the possible expression level?
- What is the possible protein product?
- Can we get the protein product?
- Can we figure out the key residue in the protein product?
- ……

基因预测方法分类

- 序列比对：
  - 和已知物种基因集进行同源序列比对，筛选出同源比对区域（利用已知的信息去预测未知）
- 从头预测：基于序列特征
  - 利用软件对物种的基因组直接进行预测。
  - 基因的编码区CDS与开放阅读框ORF
  - 核糖体RNA的保守域
  - 转运RNA的倒三叶草结构
  - 。。。。。。

基于同源性的基因预测

- Pros
  - 基于已有的生物学数据, 结果更有生物学意义

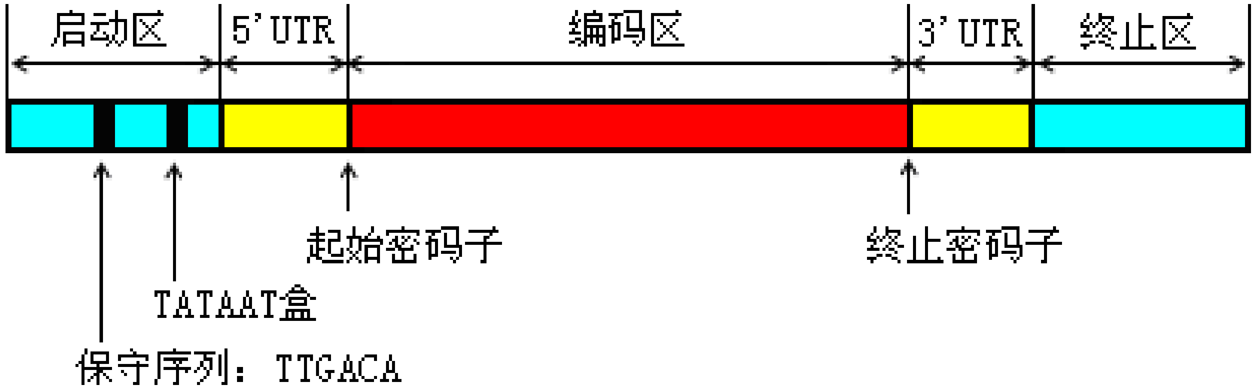
- Cons
- 受限于已有的生物学数据
  - 数据库可能存在的误差
  - 对于相似程度应如何定义

DNA序列特征分析

- 分析DNA序列，除了进行序列比对之外，更重要的工作是从序列中找到基因及其表达调控信息。
  - 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
  - 预测基因的编码区域，或预测外显子所在的区域

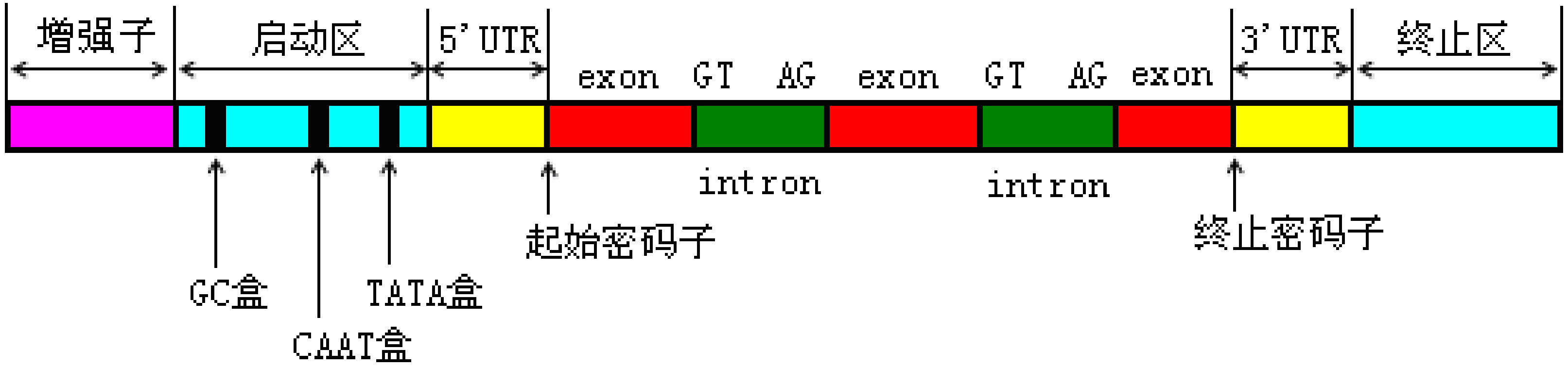
- 绝大部分基因表达调控信息隐藏在基因序列的上游区域，在组成上具有一定的特征

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，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(open reading frame, ORF)。

真核生物基因结构



- 基因由蛋白质编码序列（外显子 exon）和非编码序列（内含子 intron）组成
- 各个外显子被长度不同的内含子所隔离
- GT-AG法则：内含子5‘端是GT，3’端是AG，这两段高度保守序列与剪切机制有关，是RNA剪切的识别信号

TERMS

- 启动子（promoter）：与RNA聚合酶结合并能起始mRNA合成的序列。一般选择上游2 kb，下游 500 nt，也有选上下游各1 kb的
- 转录起始点（TSS）：转录时，mRNA链第一个核苷酸相对应DNA链上的碱基，通常为一个嘌呤。
- UTR（Untranslated Regions)：即非翻译区，mRNA分子编码区(CDS)两端的非编码片段。
- 5’-UTR从mRNA起点的甲基化鸟嘌呤核苷酸帽延伸至AUG起始密码子，3’-UTR从编码区末端的终止密码子延伸至Poly-A的末端。

“从头开始”基因预测

- Pros: 使用基因组序列本身信息预测
  - polyA信号(AATAAA)
  - 起始和终止: AUG; UAA, UAG, UGA
  - 序列中编码与非编码区域中密码子的不同使用情况
  - 上游调控信号(TATA boxes) 以及序列具体特征(CpG islands)
  - 剪切识别信号(如GT-AG)

- Cons
  - 对于预测可变剪切、嵌套或有重叠的基因作用不大

“从头开始”基因预测程序

- 刚开始只能预测单个exons, 如GRAIL、MZEF
- 后来可以预测整个基因，如Genscan、Fgenesh
- 对exons的预测，是基于密码子的使用、各种信号(起始, 终止, 剪切位点)。然后把预测到的可能exons 拼接成基因

- 单独使用这些方法，不能完全准确地预测出基因组中所有基因

Example1

IL17A interleukin 17A[Homo sapiens]

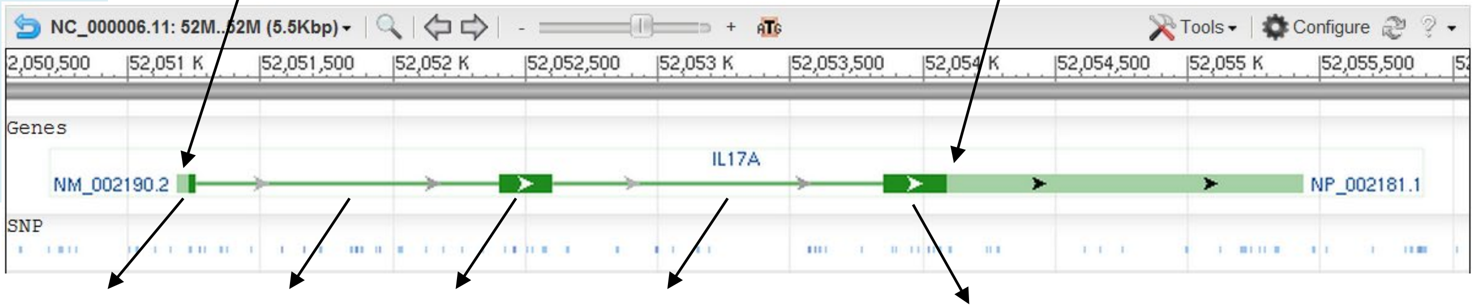
Location : 6p12

Sequence : Chromosome: 6; NC_000006.11 (52051185..52055436)

gene	1..4252
mRNA	join(1..72,1217..1419,2669..4252)
CDS	join(46..72,1217..1419,2669..2906)

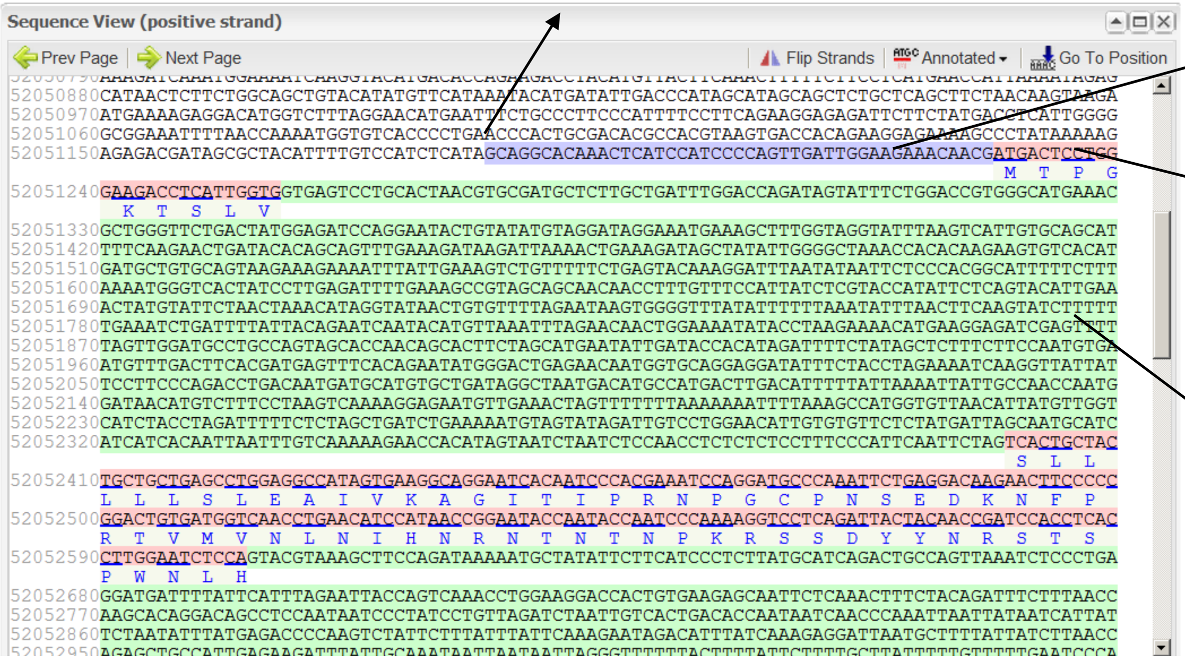
5'UTR

3'UTR



Exon1 Intron1 Exon2 Intron2 Exon3

TSS: transcription start site



```
52050790AAAGATCAAATGGAAAATCAAGGTACATGACACCAGAAGACCTACATGTTACTTCAAACTTTTTCTTCCTCATGAACCATTAAAATAGAG
52050880CATAACTCTTCTGGCAGCTGTACATATGTTCATAAATACATGATATTGACCCATAGCATAGCAGCTCTGCTCAGCTTCTAACAAGTAAGA
52050970ATGAAAAGAGGACATGGTCTTTAGGAACATGAATTTCTGCCCTTCCCATTTTCCTTCAGAAGGAGAGATTCTTCTATGACCTCATTGGGG
52051060GCGGAAATTTTAACCAAAATGGTGTCACCCCTGAACCCACTGCGACACGCCACGTAAGTGACCACAGAAGGAGAAAAGCCCTATAAAAAG
52051150AGAGACGATAGCGCTACATTTTGTCCATCTCATAGCAGGCACAAACTCATCCATCCCCAGTTGATTGGAAGAAACAACGATGACTCCTGG
                                                                                   M  T  P  G
52051240GAAGACCTCATTGGTGGTGAGTCCTGCACTAACGTGCGATGCTCTTGCTGATTTGGACCAGATAGTATTTCTGGACCGTGGGCATGAAAC
          K  T  S  L  V
52051330GCTGGGTTCTGACTATGGAGATCCAGGAATACTGTATATGTAGGATAGGAAATGAAAGCTTTGGTAGGTATTTAAGTCATTGTGCAGCAT
52051420TTTCAAGAACTGATACACAGCAGTTTGAAAGATAAGATTAAAACTGAAAGATAGCTATATTGGGGCTAAACCACACAAGAAGTGTCACAT
52051510GATGCTGTGCAGTAAGAAAGAAAATTTATTGAAAGTCTGTTTTTCTGAGTACAAAGGATTTAATATAATTCTCCCACGGCATTTTTCTTT
52051600AAAATGGGTCACTATCCTTGAGATTTTGAAAGCCGTAGCAGCAACAACCTTTGTTTCCATTATCTCGTACCATATTCTCAGTACATTGAA
52051690ACTATGTATTCTAACTAAACATAGGTATAACTGTGTTTTAGAATAAGTGGGGTTTATATTTTTAAATATTTAACTTCAAGTATCTTTTT
52051780TGAAATCTGATTTTATTACAGAATCAATACATGTTAAATTTAGAACAACTGGAAAATATACCTAAGAAAACATGAAGGAGATCGAGTTTT
52051870TAGTTGGATGCCTGCCAGTAGCACCAACAGCACTTCTAGCATGAATATTGATACCACATAGATTTCTATAGCTCTTTCTTCCAATGTGA
52051960ATGTTTGACTTCACGATGAGTTTCACAGAATATGGGACTGAGAACAATGGTGCAGGAGGATATTTCTACCTAGAAAATCAAGGTTATTAT
52052050TCCTTCCCAGACCTGACAATGATGCATGTGCTGATAGGCTAATGACATGCCATGACTTGACATTTTTATTAAAATTATTGCCAACCAATG
52052140GATAACATGTCTTTCCTAAGTCAAAAGGAGAATGTTGAAACTAGTTTTTTAAAAAAATTTAAAGCCATGGTGTTAACATTATGTTGGT
52052230CATCTACCTAGATTTTTCTCTAGCTGATCTGAAAAATGTAGTATAGATTGTCCTGGAACATTGTGTGTTCTCTATGATTAGCAATGCATC
52052320ATCATCACAATTAATTTGTCAAAAAGAACCACATAGTAATCTAATCTCCAACCTCTCTCTCCTTTCCCATTCAATTCTAGTCACTGCTAC
                                                                                    S  L  L
52052410TGCTGCTGAGCCTGGAGGCCATAGTGAAGGCAGGAATCACAATCCCACGAAATCCAGGATGCCCAAATTCTGAGGACAAGAACTTCCCCC
        L  L  L  S  L  E  A  I  V  K  A  G  I  T  I  P  R  N  P  G  C  P  N  S  E  D  K  N  F  P
52052500GGACTGTGATGGTCAACCTGAACATCCATAACCGGAATACCAATACCAATCCCAAAAGGTCCTCAGATTACTACAACCGATCCACCTCAC
        R  T  V  M  V  N  L  N  I  H  N  R  N  T  N  T  N  P  K  R  S  S  D  Y  Y  N  R  S  T  S
52052590CTTGGAATCTCCAGTACGTAAAGCTTCCAGATAAAAATGCTATATTCTTCATCCCTCTTATGCATCAGACTGCCAGTTAAATCTCCCTGA
        P  W  N  L  H
52052680GGATGATTTTATTCATTTAGAATTACCAGTCAAACCTGGAAGGACCACTGTGAAGAGCAATTCTCAAACTTTCTACAGATTTCTTTAACC
52052770AAGCACAGGACAGCCTCCAATAATCCCTATCCTGTTAGATCTAATTGTCACTGACACCAATAATCAACCCAAATTAATTATAATCATTAT
52052860TCTAATATTTATGAGACCCCAAGTCTATTCTTTATTTATTCAAAGAATAGACATTTATCAAAGAGGATTAATGCTTTTATTATCTTAACC
52052950AGAGCTGCCATTGAGAAGATTTATTGCAAATAATTAATAATTAGGGTTTTTTACTTTTATTCTTTTGCTTATTTTTGTTTTTGAATCCCA
```

Blue area: 5'UTR (1-45)

CDS (46-72)

Start codon: ATG

Green area: intron

Intron1 (73-1216)

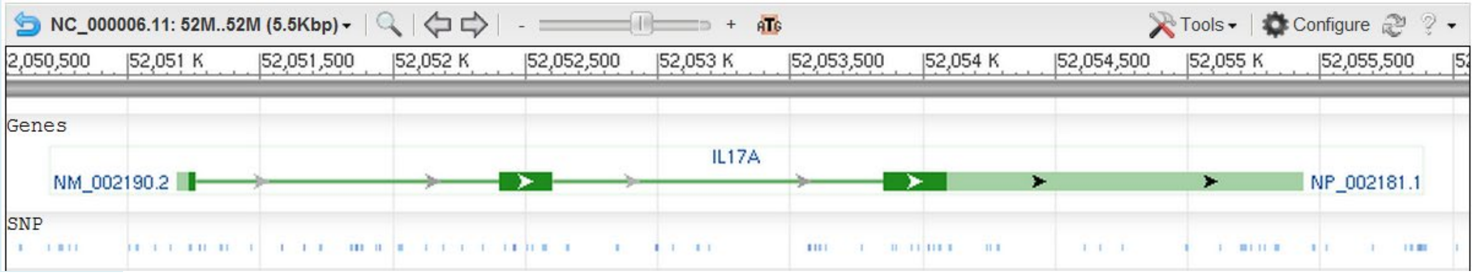
Example1

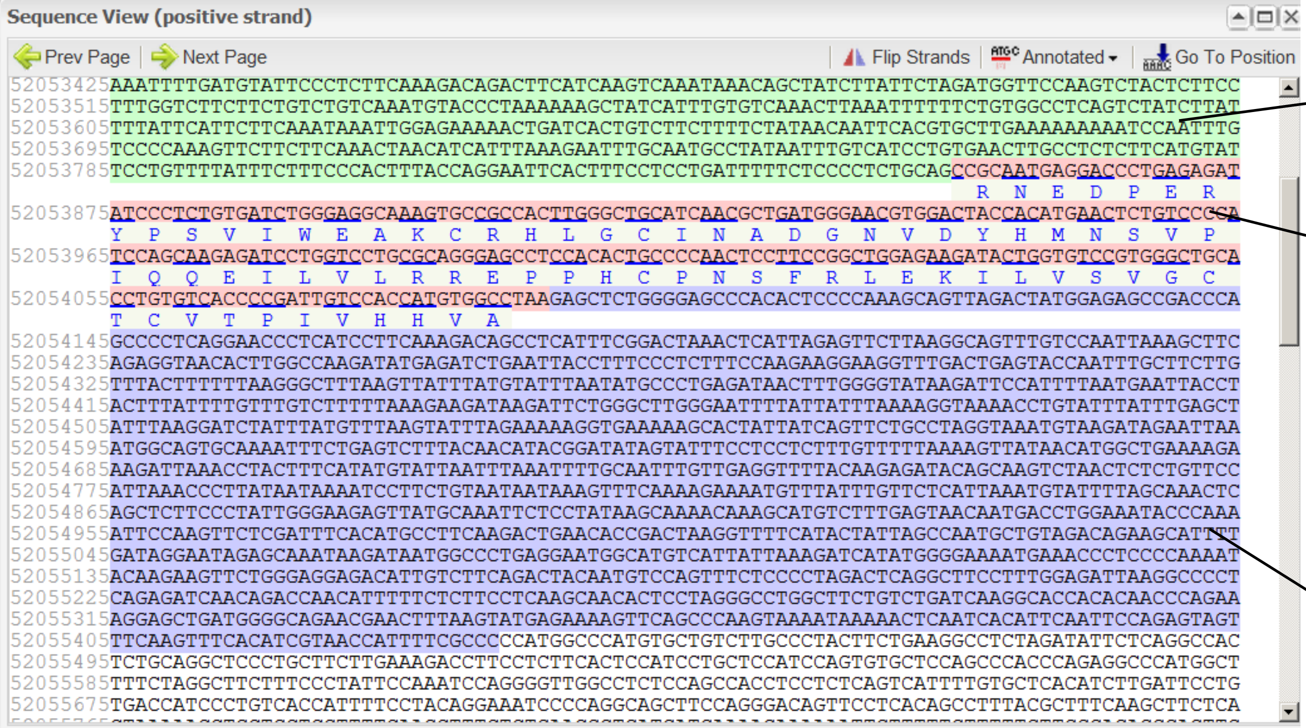
IL17A interleukin 17A[Homo sapiens]

Location :   6p12

Sequence : Chromosome: 6; NC_000006.11 (52051185..52055436)

gene	1..4252
mRNA	join(1..72,1217..1419,2669..4252)
CDS	join(46..72,1217..1419,2669..2906)





```
52053425AAATTTTGATGTATTCCCTCTTCAAAGACAGACTTCATCAAGTCAAATAAACAGCTATCTTATTCTAGATGGTTCCAAGTCTACTCTTCC
52053515TTTGGTCTTCTTCTGTCTGTCAAATGTACCCTAAAAAAGCTATCATTTGTGTCAAACTTAAATTTTTCTGTGGCCTCAGTCTATCTTAT
52053605TTTATTCATTCTTCAAATAAATTGGAGAAAAACTGATCACTGTCTTCTTTTCTATAACAATTCACGTGCTTGAAAAAAAAATCCAATTTG
52053695TCCCCAAAGTTCTTCTTCAAACTAACATCATTTAAAGAATTTGCAATGCCTATAATTTGTCATCCTGTGAACTTGCCTCTCTTCATGTAT
52053785TCCTGTTTTATTTCTTTCCCACTTTACCAGGAATTCACTTTCCTCCTGATTTTTCTCCCCTCTGCAGCCGCAATGAGGACCCTGAGAGAT
                                                                               R  N  E  D  P  E  R
52053875ATCCCTCTGTGATCTGGGAGGCAAAGTGCCGCCACTTGGGCTGCATCAACGCTGATGGGAACGTGGACTACCACATGAACTCTGTCCCCA
         Y  P  S  V  I  W  E  A  K  C  R  H  L  G  C  I  N  A  D  G  N  V  D  Y  H  M  N  S  V  P
52053965TCCAGCAAGAGATCCTGGTCCTGCGCAGGGAGCCTCCACACTGCCCCAACTCCTTCCGGCTGGAGAAGATACTGGTGTCCGTGGGCTGCA
         I  Q  Q  E  I  L  V  L  R  R  E  P  P  H  C  P  N  S  F  R  L  E  K  I  L  V  S  V  G  C
52054055CCTGTGTCACCCCGATTGTCCACCATGTGGCCTAAGAGCTCTGGGGAGCCCACACTCCCCAAAGCAGTTAGACTATGGAGAGCCGACCCA
         T  C  V  T  P  I  V  H  H  V  A
52054145GCCCCTCAGGAACCCTCATCCTTCAAAGACAGCCTCATTTCGGACTAAACTCATTAGAGTTCTTAAGGCAGTTTGTCCAATTAAAGCTTC
52054235AGAGGTAACACTTGGCCAAGATATGAGATCTGAATTACCTTTCCCTCTTTCCAAGAAGGAAGGTTTGACTGAGTACCAATTTGCTTCTTG
52054325TTTACTTTTTTAAGGGCTTTAAGTTATTTATGTATTTAATATGCCCTGAGATAACTTTGGGGTATAAGATTCCATTTTAATGAATTACCT
52054415ACTTTATTTTGTTTGTCTTTTTAAAGAAGATAAGATTCTGGGCTTGGGAATTTTATTATTTAAAAGGTAAAACCTGTATTTATTTGAGCT
52054505ATTTAAGGATCTATTTATGTTTAAGTATTTAGAAAAAGGTGAAAAAGCACTATTATCAGTTCTGCCTAGGTAAATGTAAGATAGAATTAA
52054595ATGGCAGTGCAAAATTTCTGAGTCTTTACAACATACGGATATAGTATTTCCTCCTCTTTGTTTTTAAAAGTTATAACATGGCTGAAAAGA
52054685AAGATTAAACCTACTTTCATATGTATTAATTTAAATTTTGCAATTTGTTGAGGTTTTACAAGAGATACAGCAAGTCTAACTCTCTGTTCC
52054775ATTAAACCCTTATAATAAAATCCTTCTGTAATAATAAAGTTTCAAAAGAAAATGTTTATTTGTTCTCATTAAATGTATTTTAGCAAACTC
52054865AGCTCTTCCCTATTGGGAAGAGTTATGCAAATTCTCCTATAAGCAAAACAAAGCATGTCTTTGAGTAACAATGACCTGGAAATACCCAAA
52054955ATTCCAAGTTCTCGATTTCACATGCCTTCAAGACTGAACACCGACTAAGGTTTTCATACTATTAGCCAATGCTGTAGACAGAAGCATTTT
52055045GATAGGAATAGAGCAAATAAGATAATGGCCCTGAGGAATGGCATGTCATTATTAAAGATCATATGGGGAAAATGAAACCCTCCCCAAAAT
52055135ACAAGAAGTTCTGGGAGGAGACATTGTCTTCAGACTACAATGTCCAGTTTCTCCCCTAGACTCAGGCTTCCTTTGGAGATTAAGGCCCCT
52055225CAGAGATCAACAGACCAACATTTTTCTCTTCCTCAAGCAACACTCCTAGGGCCTGGCTTCTGTCTGATCAAGGCACCACACAACCCAGAA
52055315AGGAGCTGATGGGGCAGAACGAACTTTAAGTATGAGAAAAGTTCAGCCCAAGTAAAATAAAAACTCAATCACATTCAATTCCAGAGTAGT
52055405TTCAAGTTTCACATCGTAACCATTTTCGCCCCCATGGCCCATGTGCTGTCTTGCCCTACTTCTGAAGGCCTCTAGATATTCTCAGGCCAC
52055495TCTGCAGGCTCCCTGCTTCTTGAAAGACCTTCCTCTTCACTCCATCCTGCTCCATCCAGTGTGCTCCAGCCCACCCAGAGGCCCATGGCT
52055585TTTCTAGGCTTCTTTCCCTATTCCAAATCCAGGGGTTGGCCTCTCCAGCCACCTCCTCTCAGTCATTTTGTGCTCACATCTTGATTCCTG
52055675TGACCATCCCTGTCACCATTTTCCTACAGGAAATCCCCAGGCAGCTTCCAGGGACAGTTCCTCACAGCCTTTACGCTTTCAAGCTTCTCA
```

Green area: intron

Intron2 (1420-2668)

CDS (2669-4252)

End codon: TAA

Blue area:

3'UTR (2907-4252)

启动子 PROMOTERS

- 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。

开放阅读框ORF

- 开放阅读框(open reading frame, ORF)指的是从5‘端翻译起始密码子（AUG）到终止密码子（UUA、UAG、UGA）的蛋白质编码碱基序列

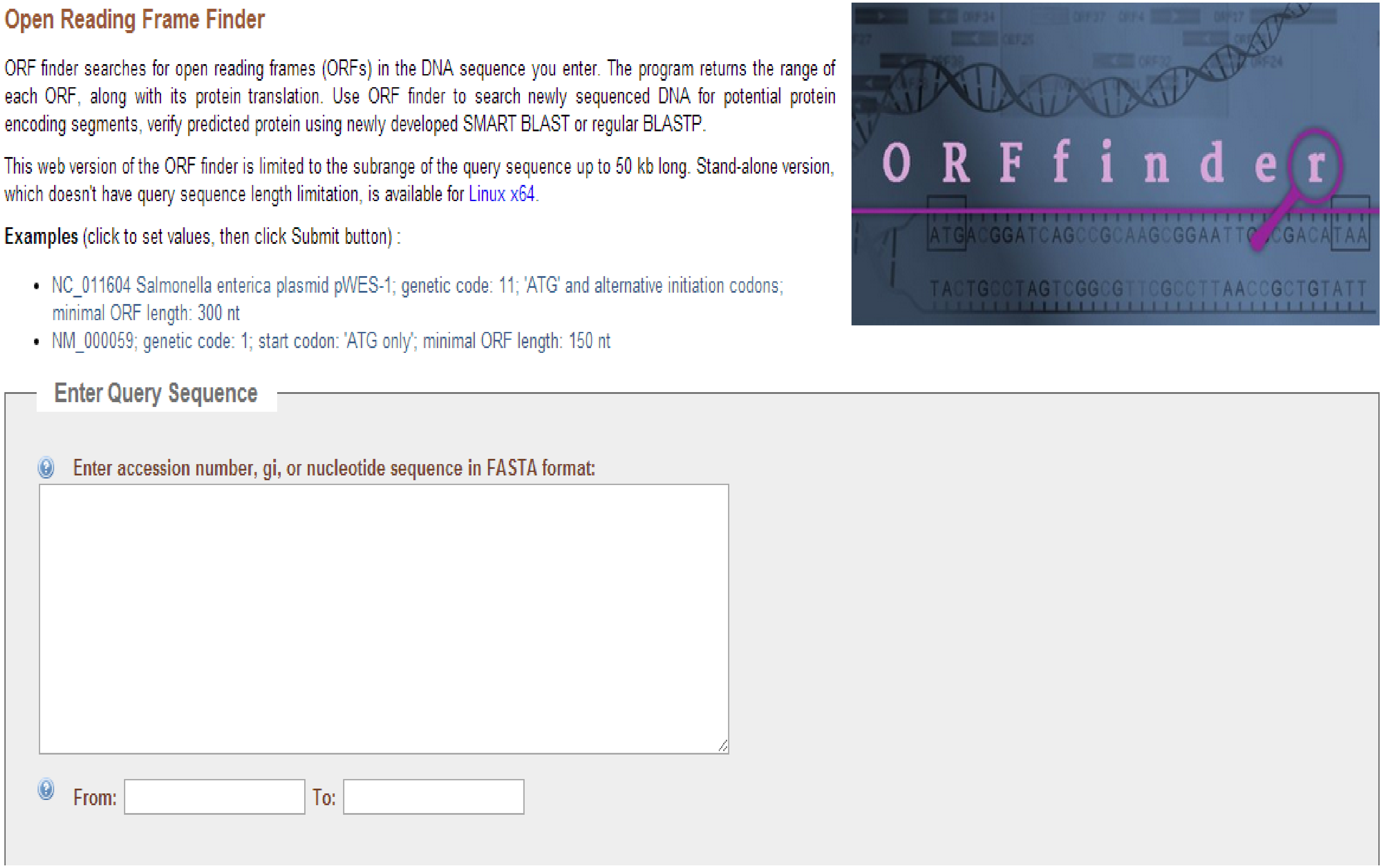
- DNA双链正反向共6种可能的阅读方式，分析的目的是从中找出一个正确的ORF

- 真核生物的内含子GT-AG法则有助于开放阅读框的识别

ORF & CDS

- ORF：理论上的氨基酸编码区。
  - 程序在DNA序列中寻找启动因子（AUG），然后按每3个核酸一组，一直延伸寻找下去，直到碰到终止因子（UGA,UAA或UAG）。这个区域为ORF区，理论上可以编码一组氨基酸。
- CDS：编码一段蛋白产物的序列。
  - CDS必定是一个ORF，也可能包括很多ORF。

ORF Finder



Open Reading Frame Finder

ORF finder searches for open reading frames (ORFs) in the DNA sequence you enter. The program returns the range of each ORF, along with its protein translation. Use ORF finder to search newly sequenced DNA for potential protein encoding segments, verify predicted protein using newly developed SMART BLAST or regular BLASTP.

This web version of the ORF finder is limited to the subrange of the query sequence up to 50 kb long. Stand-alone version, which doesn't have query sequence length limitation, is available for Linux x64.

Examples (click to set values, then click Submit button) :

- NC_011604 Salmonella enterica plasmid pWES-1; genetic code: 11; 'ATG' and alternative initiation codons; minimal ORF length: 300 nt
- NM_000059; genetic code: 1; start codon: 'ATG only'; minimal ORF length: 150 nt

ORFfinder
ATGACGGATCAGCCGCAAGCGGAATTCGCGACATAA
TACTGCCTAGTCGGCGTTCGCCTTAACCGCTGTATT

Enter Query Sequence

Enter accession number, gi, or nucleotide sequence in FASTA format:

From: To:

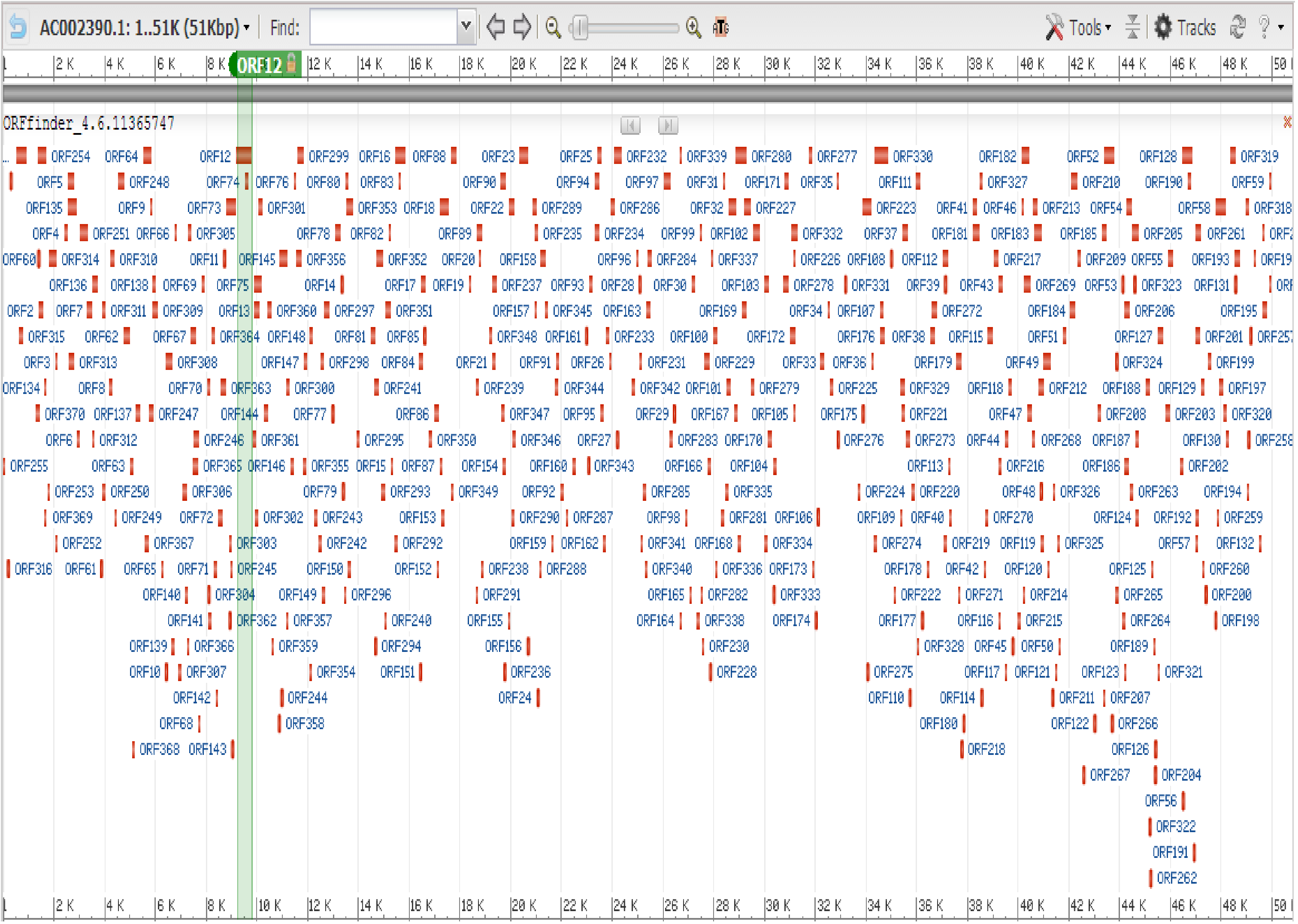
```
CDS             join(1325..1552,3620..3738,5216..5340,6405..6595)
                /gene="RAB3A"
                /note="Derived by automated computational analysis using
                gene prediction method: BestRefseq."
                /codon_start=1
```

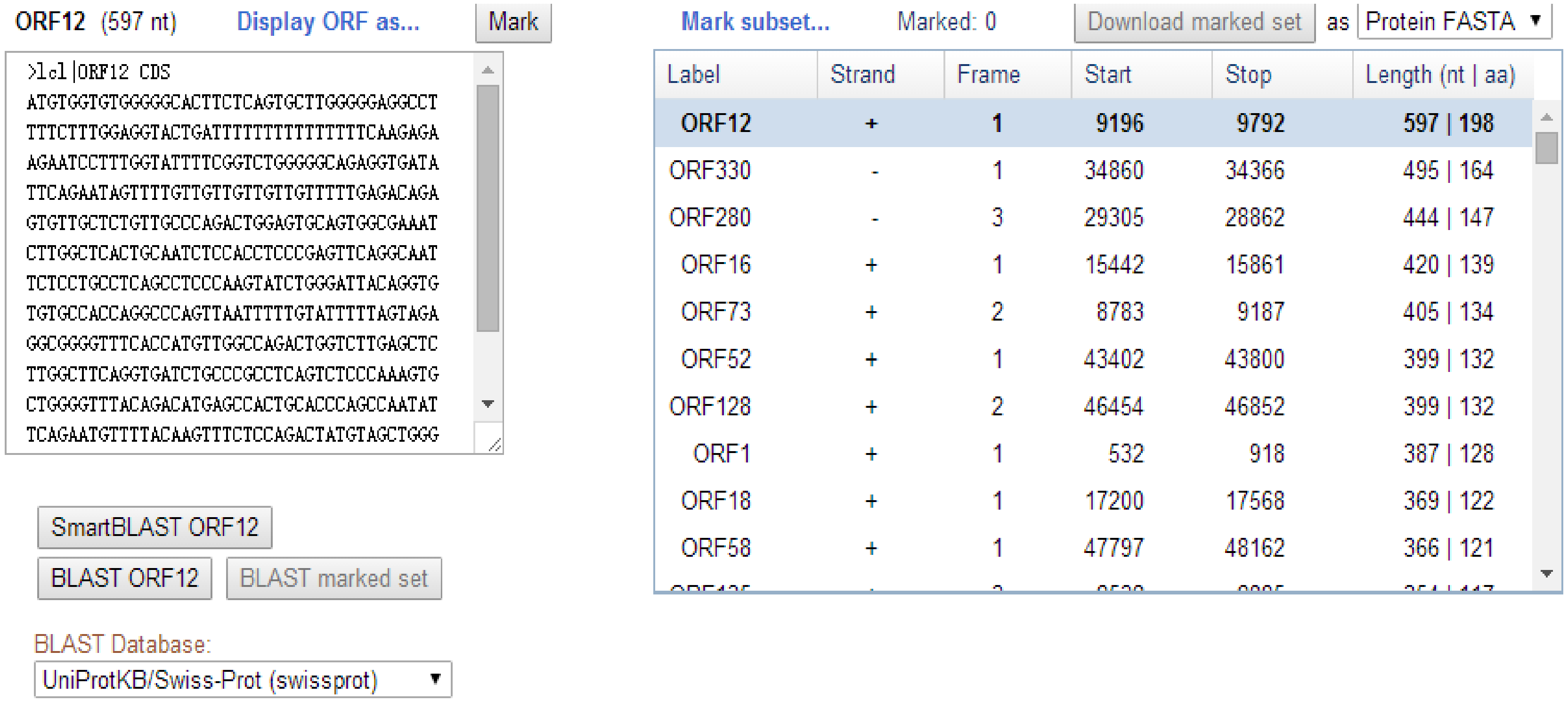
AC002390.1 Human DNA from overlapping chromosome 19-specific cosmids R30072 and R28588, genomic sequence, complete sequence

ORFs found: 370   Genetic code: 1   Start codon: 'ATG' only

ORFs were calculated on the interval from 1 to 50000 nt

ORIG


AC002390.1: 1..51K (51Kbp)
Find:
Tools
Tracks
1 2 K 4 K 6 K 8 K 10 K 12 K 14 K 16 K 18 K 20 K 22 K 24 K 26 K 28 K 30 K 32 K 34 K 36 K 38 K 40 K 42 K 44 K 46 K 48 K 50 K
ORF12
ORFfinder_4.6.11365747
ORF254 ORF64 ORF12 ORF299 ORF16 ORF88 ORF23 ORF25 ORF232 ORF339 ORF280 ORF277 ORF330 ORF182 ORF52 ORF128 ORF319
ORF5 ORF248 ORF74 ORF76 ORF80 ORF83 ORF90 ORF94 ORF97 ORF31 ORF171 ORF35 ORF111 ORF327 ORF210 ORF190 ORF59
ORF135 ORF9 ORF73 ORF301 ORF353 ORF18 ORF22 ORF289 ORF286 ORF32 ORF227 ORF223 ORF41 ORF46 ORF213 ORF54 ORF58 ORF318
ORF4 ORF251 ORF66 ORF305 ORF78 ORF82 ORF89 ORF235 ORF234 ORF99 ORF102 ORF332 ORF37 ORF181 ORF183 ORF185 ORF205 ORF261
ORF60 ORF314 ORF310 ORF11 ORF145 ORF356 ORF352 ORF20 ORF158 ORF96 ORF284 ORF337 ORF226 ORF108 ORF112 ORF217 ORF209 ORF55 ORF193
ORF136 ORF138 ORF69 ORF75 ORF14 ORF17 ORF19 ORF237 ORF93 ORF28 ORF30 ORF103 ORF278 ORF331 ORF39 ORF43 ORF269 ORF53 ORF323 ORF131
ORF2 ORF7 ORF311 ORF309 ORF13 ORF360 ORF297 ORF351 ORF157 ORF345 ORF163 ORF169 ORF34 ORF107 ORF272 ORF184 ORF206 ORF195
ORF315 ORF62 ORF67 ORF364 ORF148 ORF81 ORF85 ORF348 ORF161 ORF233 ORF100 ORF172 ORF176 ORF38 ORF115 ORF51 ORF127 ORF201 ORF257
ORF3 ORF313 ORF308 ORF147 ORF298 ORF84 ORF21 ORF91 ORF26 ORF231 ORF229 ORF33 ORF36 ORF179 ORF49 ORF324 ORF199
ORF134 ORF8 ORF70 ORF363 ORF300 ORF241 ORF239 ORF344 ORF342 ORF101 ORF279 ORF225 ORF329 ORF118 ORF212 ORF188 ORF129 ORF197
ORF370 ORF137 ORF247 ORF144 ORF77 ORF86 ORF347 ORF95 ORF29 ORF167 ORF105 ORF175 ORF221 ORF47 ORF208 ORF203 ORF320
ORF6 ORF312 ORF246 ORF361 ORF295 ORF350 ORF346 ORF27 ORF283 ORF170 ORF276 ORF273 ORF44 ORF268 ORF187 ORF130 ORF258
ORF255 ORF63 ORF365 ORF146 ORF355 ORF15 ORF87 ORF154 ORF160 ORF343 ORF166 ORF104 ORF113 ORF216 ORF186 ORF202
ORF253 ORF250 ORF306 ORF79 ORF293 ORF349 ORF92 ORF285 ORF335 ORF224 ORF220 ORF48 ORF326 ORF263 ORF194
ORF369 ORF249 ORF72 ORF302 ORF243 ORF153 ORF290 ORF287 ORF98 ORF281 ORF106 ORF109 ORF40 ORF270 ORF124 ORF192 ORF259
ORF252 ORF367 ORF303 ORF242 ORF292 ORF159 ORF162 ORF341 ORF168 ORF334 ORF274 ORF219 ORF119 ORF325 ORF57 ORF132
ORF316 ORF61 ORF65 ORF71 ORF245 ORF150 ORF152 ORF238 ORF288 ORF340 ORF336 ORF173 ORF178 ORF42 ORF120 ORF125 ORF260
ORF140 ORF304 ORF149 ORF296 ORF291 ORF165 ORF282 ORF333 ORF222 ORF271 ORF214 ORF265 ORF200
ORF141 ORF362 ORF357 ORF240 ORF155 ORF164 ORF338 ORF174 ORF177 ORF116 ORF215 ORF264 ORF198
ORF139 ORF366 ORF359 ORF294 ORF156 ORF230 ORF328 ORF45 ORF50 ORF189
ORF10 ORF307 ORF354 ORF151 ORF236 ORF228 ORF275 ORF117 ORF121 ORF123 ORF321
ORF142 ORF244 ORF24 ORF110 ORF114 ORF211 ORF207
ORF68 ORF358 ORF180 ORF122 ORF266
ORF368 ORF143 ORF218 ORF126
ORF267 ORF204
ORF56
ORF322
ORF191
ORF262

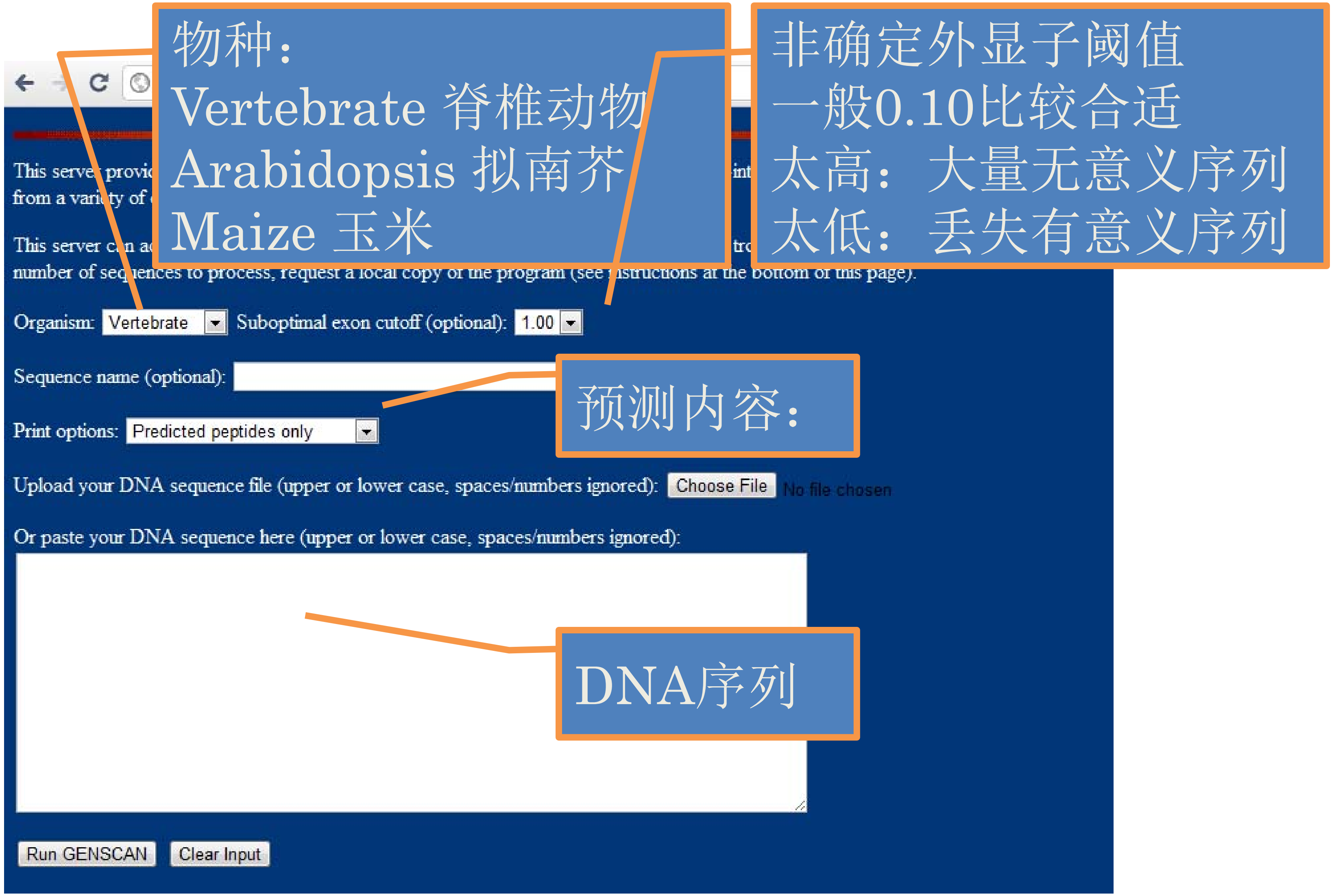


```
>lcl|ORF12 CDS
ATGTGGTGTGGGGGCACTTCTCAGTGCTTGGGGGAGGCCT
TTTCTTTGGAGGTACTGATTTTTTTTTTTTTTTCAAGAGA
AGAATCCTTTGGTATTTTCGGTCTGGGGGCAGAGGTGATA
TTCAGAATAGTTTTGTTGTTGTTGTTGTTTTTGAGACAGA
GTGTTGCTCTGTTGCCCAGACTGGAGTGCAGTGGCGAAAT
CTTGGCTCACTGCAATCTCCACCTCCCGAGTTCAGGCAAT
TCTCCTGCCTCAGCCTCCCAAGTATCTGGGATTACAGGTG
TGTGCCACCAGGCCCAGTTAATTTTTGTATTTTTAGTAGA
GGCGGGGTTTCACCATGTTGGCCAGACTGGTCTTGAGCTC
TTGGCTTCAGGTGATCTGCCCGCCTCAGTCTCCCAAAGTG
CTGGGGTTTACAGACATGAGCCACTGCACCCAGCCAATAT
TCAGAATGTTTTACAAGTTTCTCCAGACTATGTAGCTGGG
```

Label	Strand	Frame	Start	Stop	Length (nt | aa)
ORF12	**+**	**1**	**9196**	**9792**	**597 | 198**
ORF330	-	1	34860	34366	495 | 164
ORF280	-	3	29305	28862	444 | 147
ORF16	+	1	15442	15861	420 | 139
ORF73	+	2	8783	9187	405 | 134
ORF52	+	1	43402	43800	399 | 132
ORF128	+	2	46454	46852	399 | 132
ORF1	+	1	532	918	387 | 128
ORF18	+	1	17200	17568	369 | 122
ORF58	+	1	47797	48162	366 | 121

GENSCAN识别ORF

[HTTP://GENES.MIT.EDU/GENSCAN.HTML](http://genes.mit.edu/GENSCAN.html)



This serve provi … int
from a vari ty of …

This server c n a … tro
number of seq ences 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): Choose File No file chosen

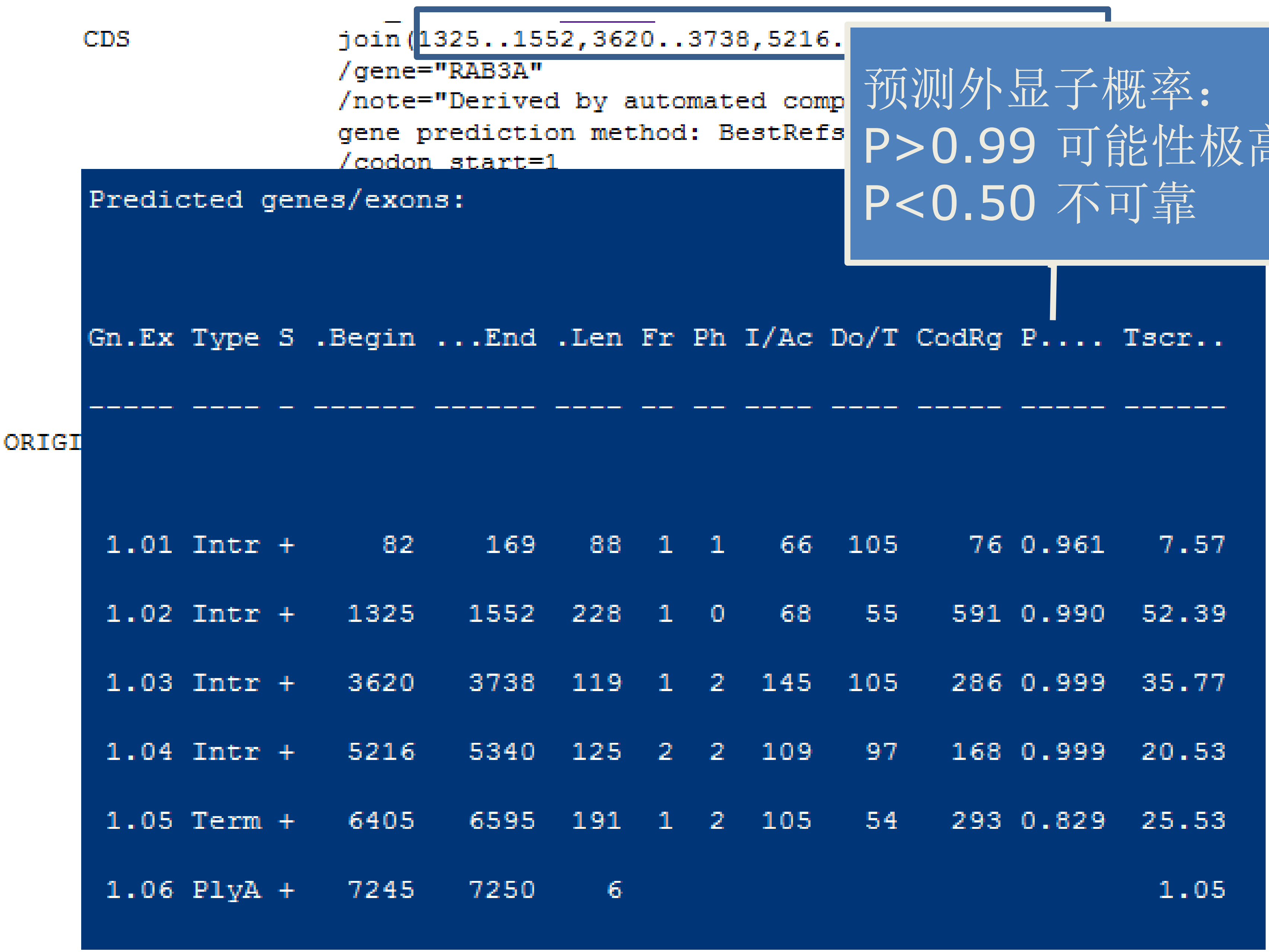
Or paste your DNA sequence here (upper or lower case, spaces/numbers ignored):

物种：
Vertebrate 脊椎动物
Arabidopsis 拟南芥
Maize 玉米

非确定外显子阈值
一般0.10比较合适
太高：大量无意义序列
太低：丢失有意义序列

预测内容：

DNA序列



```
     CDS             join(1325..1552,3620..3738,5216.
                     /gene="RAB3A"
                     /note="Derived by automated comp
                     gene prediction method: BestRefs
                     /codon_start=1
ORIGI
```

```
Predicted genes/exons:

Gn.Ex Type S .Begin ...End .Len Fr Ph I/Ac Do/T CodRg P.... Tscr..
----- ---- - ------ ------ ---- -- -- ---- ---- ----- ----- ------

 1.01 Intr +     82    169   88  1  1   66  105    76 0.961   7.57
 1.02 Intr +   1325   1552  228  1  0   68   55   591 0.990  52.39
 1.03 Intr +   3620   3738  119  1  2  145  105   286 0.999  35.77
 1.04 Intr +   5216   5340  125  2  2  109   97   168 0.999  20.53
 1.05 Term +   6405   6595  191  1  2  105   54   293 0.829  25.53
 1.06 PlyA +   7245   7250    6                                1.05
```

预测外显子概率：
P>0.99 可能性极高
P<0.50 不可靠

PARAMETER

- Gn.Ex gene number, exon number (for reference)
- Type： Init = Initial exon （ATG to 5’ splice site）

  Intr = Internal exon

  Term = Terminal exon

  Sngl = Single-exon gene

  Prom = Promoter

  PlyA = poly-A signal
- S DNA strand (+ = input strand; - = opposite strand)

FR "ABSOLUTE READING FRAME"

- relative to start of sequence.
  - if nucleotides 1,2,3 of the sequence are read as a codon, that's called reading frame 0.
  - If 2,3,4 are read as a codon, that's reading frame 1.
  - If 3,4,5 are read as a codon, that's reading frame 2, and so on.

PH "NET PHASE" OF EXON (EXON LENGTH MODULO 3)

- an exon of length 15 bp has net phase 0 since 15 is divisible by 3,
- an exon of length 16 bp has net phase 1 because 16 divided by 3 leaves a remainder of 1,
- an exon of length 17 bp has net phase 2, and an exon of length 18 bp has net phase 0 again.
- The point of this is that exons whose net phase is 0 can be omitted from the gene without disrupting the reading frame: such exons are candidates for being either 1) incorrect, or 2) alternatively spliced.

Parameters continue

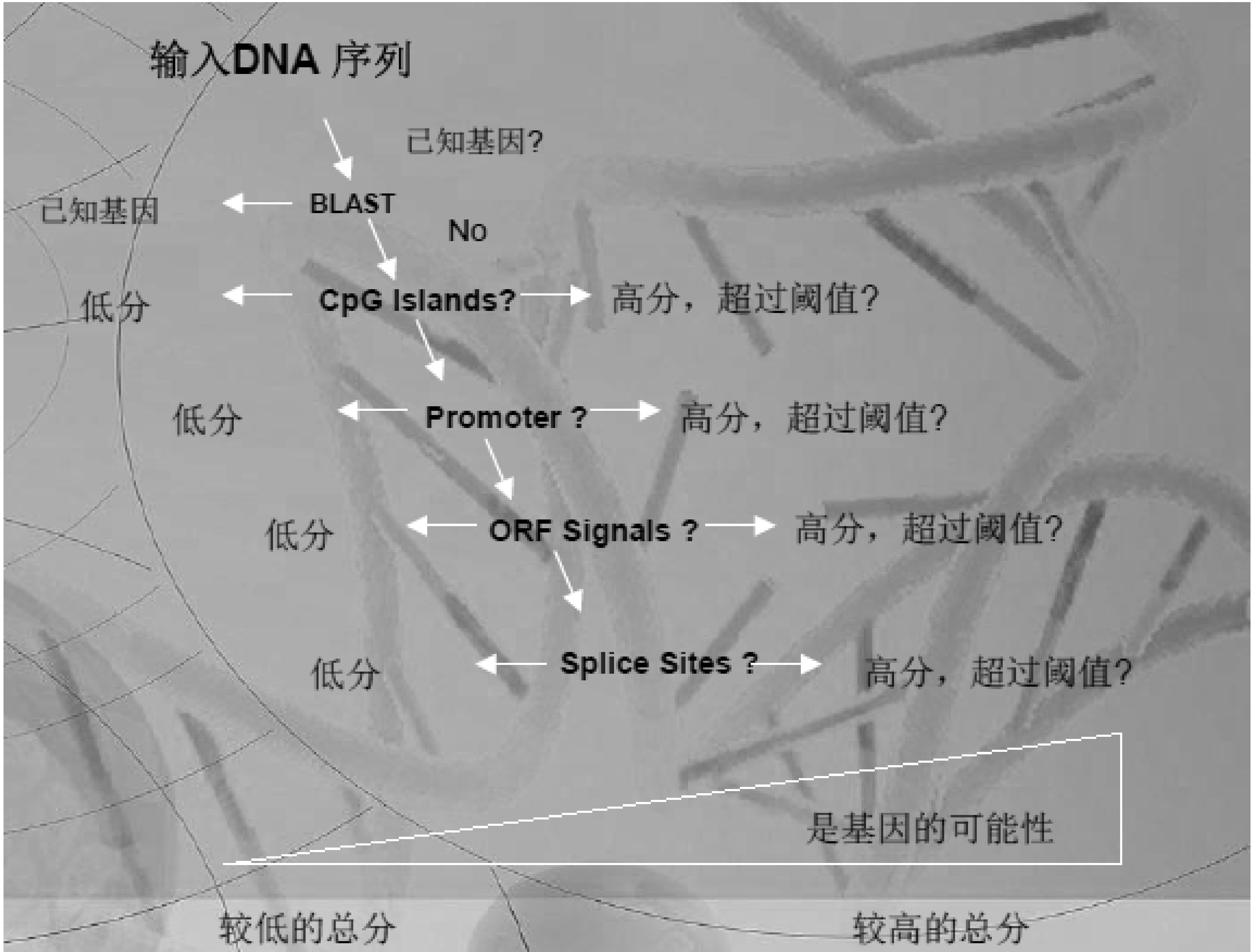
- I/Ac initiation signal or acceptor splice site score (x 10)
- Do/T donor splice site or termination signal score (x 10)
- CodRg coding region score (x 10)
  - Low coding region scores may indicate potentially incorrect predictions or genes with unusual amino acid and/or codon usage patterns.

PARAMETERS CONTINUE

- P probability of exon (sum over all parses containing exon)
  - This quantity is close to the actual probability that the predicted exon is correct.
- Tscr exon score (depends on length, I/Ac, Do/T and CodRg scores)
  - An overall measure of exon quality based on local sequence properties

GENSCAN的局限

- 重叠的转录单元
- 可变剪切
- 物种
- 准确率:
  - 中间exons > 初始或终止exons
  - exons > polyA 或启动子信号.



输入DNA 序列

已知基因?

已知基因 ← BLAST

No

低分 ← CpG Islands? → 高分，超过阈值?

低分 ← Promoter ? → 高分，超过阈值?

低分 ← ORF Signals ? → 高分，超过阈值?

低分 ← Splice Sites ? → 高分，超过阈值?

是基因的可能性

较低的总分

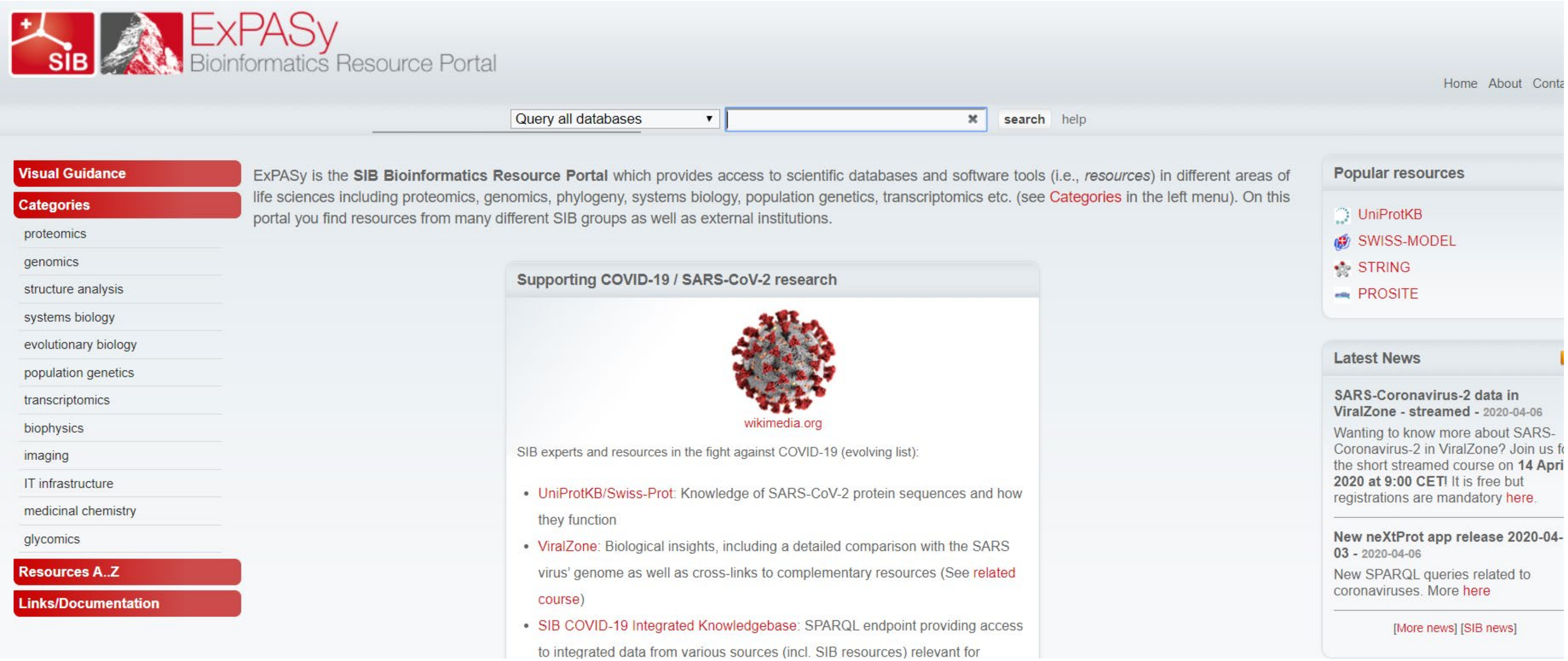
较高的总分

综合型基因识别方法

- 综合相似性比较结果及“从头开始”技术的方法
- 结合不同物种间同线性（ synteny） 的方法
- 整合几种预测基因不同部分的方法
- 整合几种不同的基因预测程序的结果

ExPASy（Expert Protein Analysis System）

- 瑞士生物信息学中心维护
- 提供系列蛋白质分析工具



ExPASy is the **SIB Bioinformatics Resource Portal** which provides access to scientific databases and software tools (i.e., *resources*) in different areas of life sciences including proteomics, genomics, phylogeny, systems biology, population genetics, transcriptomics etc. (see Categories in the left menu). On this portal you find resources from many different SIB groups as well as external institutions.

Supporting COVID-19 / SARS-CoV-2 research

SIB experts and resources in the fight against COVID-19 (evolving list):

- UniProtKB/Swiss-Prot: Knowledge of SARS-CoV-2 protein sequences and how they function
- ViralZone: Biological insights, including a detailed comparison with the SARS virus' genome as well as cross-links to complementary resources (See related course)
- SIB COVID-19 Integrated Knowledgebase: SPARQL endpoint providing access to integrated data from various sources (incl. SIB resources) relevant for

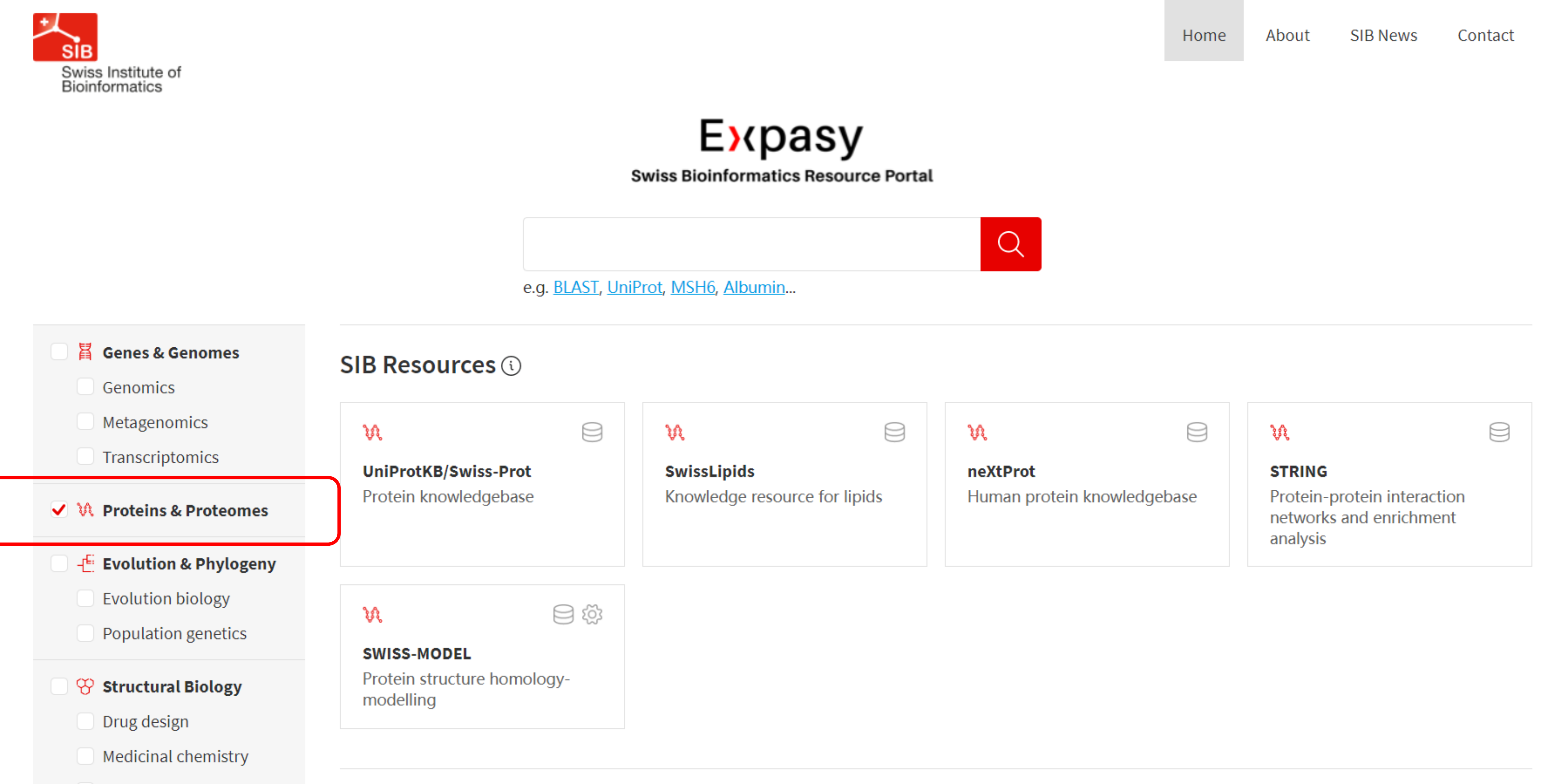
Latest News

SARS-Coronavirus-2 data in ViralZone - streamed - 2020-04-06

Wanting to know more about SARS-Coronavirus-2 in ViralZone? Join us f the short streamed course on **14 Apri 2020 at 9:00 CET!** It is free but registrations are mandatory here.

New neXtProt app release 2020-04-03 - 2020-04-06

New SPARQL queries related to coronaviruses. More here



SIB Swiss Institute of Bioinformatics

Home | About | SIB News | Contact

Expasy

Swiss Bioinformatics Resource Portal

e.g. BLAST, UniProt, MSH6, Albumin...

- □ **Genes & Genomes**
  - □ Genomics
  - □ Metagenomics
  - □ Transcriptomics
- ☑ **Proteins & Proteomes**
- □ **Evolution & Phylogeny**
  - □ Evolution biology
  - □ Population genetics
- □ **Structural Biology**
  - □ Drug design
  - □ Medicinal chemistry

SIB Resources ⓘ

UniProtKB/Swiss-Prot
Protein knowledgebase

SwissLipids
Knowledge resource for lipids

neXtProt
Human protein knowledgebase

STRING
Protein-protein interaction networks and enrichment analysis

SWISS-MODEL
Protein structure homology-modelling

蛋白质的理化性质

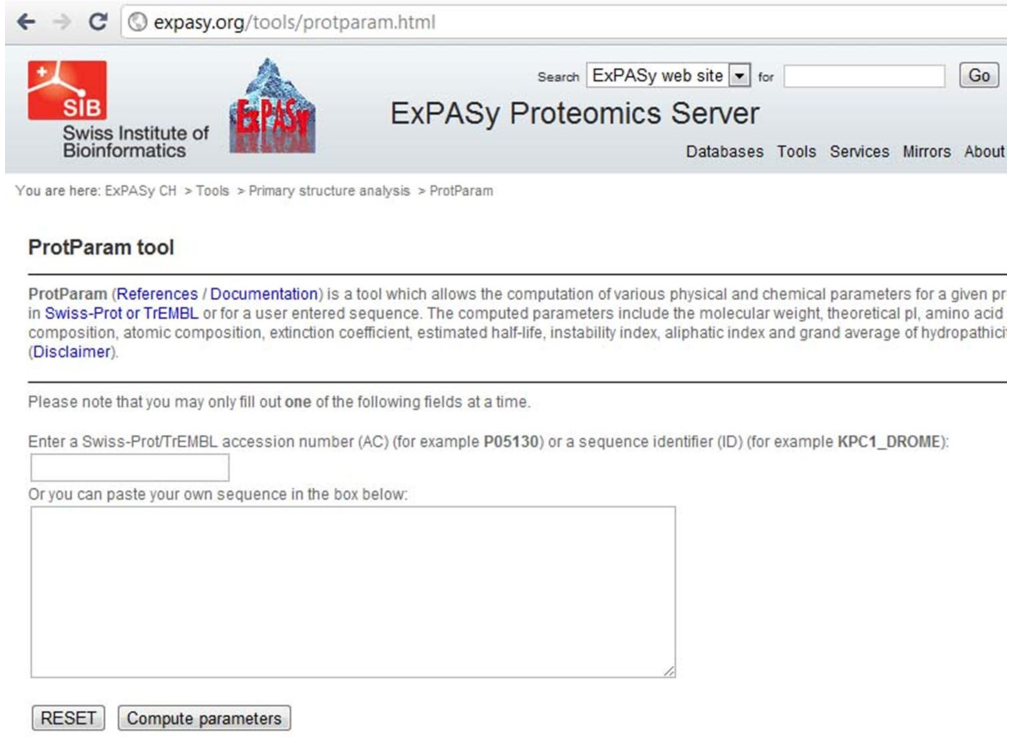
- 蛋白质是由氨基酸组成的大分子化合物，对组成蛋白质的氨基酸进行理化性质的统计分析是对一个未知蛋白质进行分析的基础。

- 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。

PROTPARAM分析蛋白质理化性质

- physico-chemical parameters of a protein sequence
  - https://web.expasy.org/protparam/

未考虑蛋白质翻译后修饰、蛋白质多聚体



expasy.org/tools/protparam.html
SIB
Swiss Institute of Bioinformatics
ExPASy
Search ExPASy web site for Go
ExPASy Proteomics Server
Databases Tools Services Mirrors About

You are here: ExPASy CH > Tools > Primary structure analysis > ProtParam

ProtParam tool

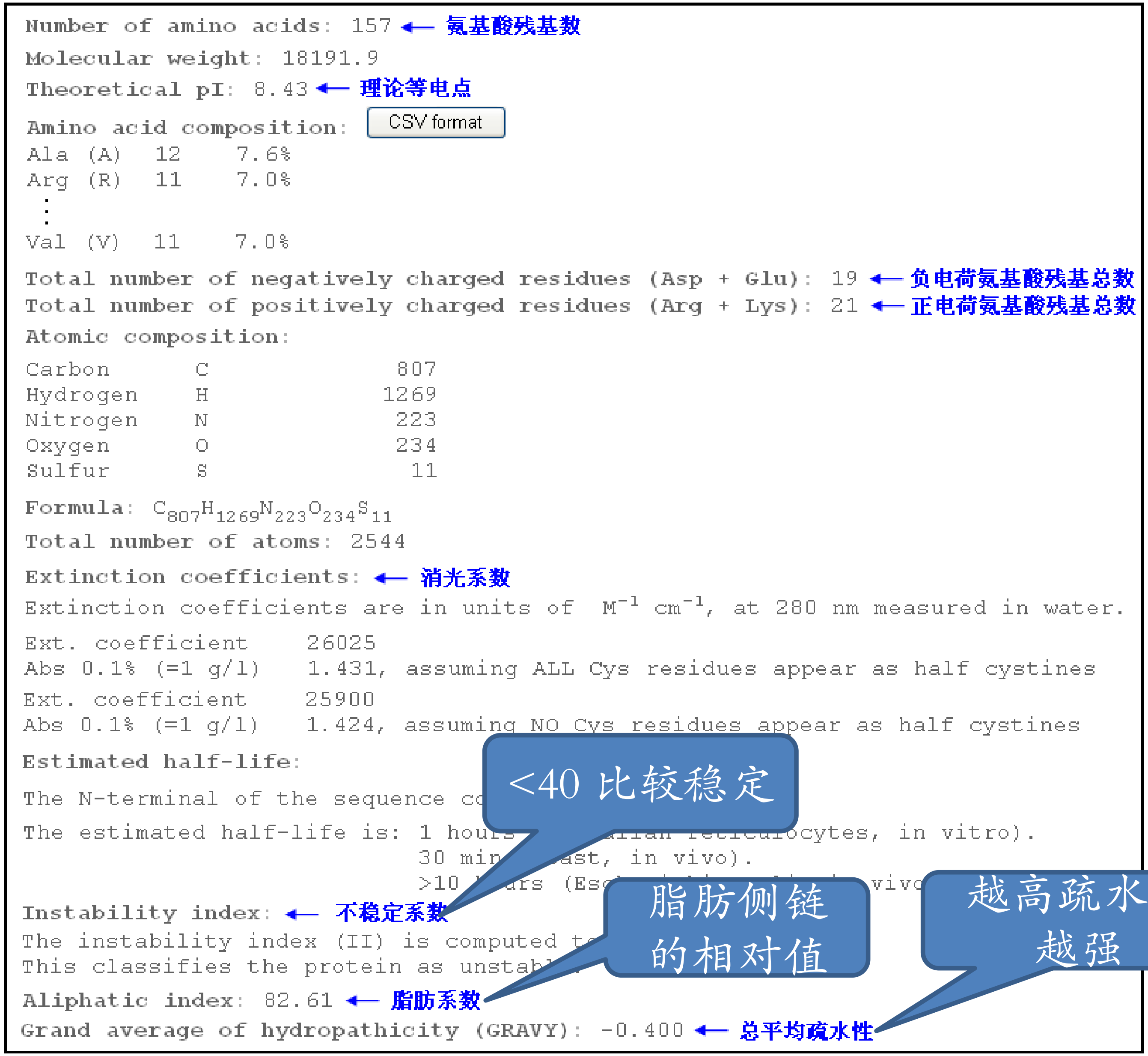
ProtParam (References / Documentation) is a tool which allows the computation of various physical and chemical parameters for a given pr in Swiss-Prot or TrEMBL or for a user entered sequence. The computed parameters include the molecular weight, theoretical pI, amino acid composition, atomic composition, extinction coefficient, estimated half-life, instability index, aliphatic index and grand average of hydropathici (Disclaimer).

Please note that you may only fill out **one** of the following fields at a time.

Enter a Swiss-Prot/TrEMBL accession number (AC) (for example **P05130**) or a sequence identifier (ID) (for example **KPC1_DROME**):

Or you can paste your own sequence in the box below:

用ProtParam分析Q28332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 co

The estimated half-life is: 1 hou                 reticulocytes, in vitro).
                            30 min       ast, in vivo).
                            >10         rs (Es                    vivo

Instability index: ← 不稳定系数
The instability index (II) is computed t
This classifies the protein as unstab

Aliphatic index: 82.61 ← 脂肪系数
Grand average of hydropathicity (GRAVY): -0.400 ← 总平均疏水性
```

<40 比较稳定

脂肪侧链的相对值

越高疏水性越强

蛋白质的亲水性或疏水性

- 非极性氨基酸（疏水氨基酸）：
  - 丙氨酸（Ala）缬氨酸（Val）亮氨酸（Leu）异亮氨酸（Ile）苯丙氨酸（Phe）色氨酸（Trp）甲硫氨酸(Met) 脯氨酸（Pro）
- 极性氨基酸（亲水氨基酸）：
  - 1）极性不带电荷/极性中性氨基酸

甘氨酸（Gly） 苏氨酸（Thr）丝氨酸（Ser）半胱氨酸（Cys）天冬酰胺（Asn）谷氨酰胺（Gln）酪氨酸（Tyr）

- 2）带正电氨基酸（碱性氨基酸）

赖氨酸（Lys）精氨酸（Arg）组氨酸（His）

- 3）带负电氨基酸（酸性氨基酸）

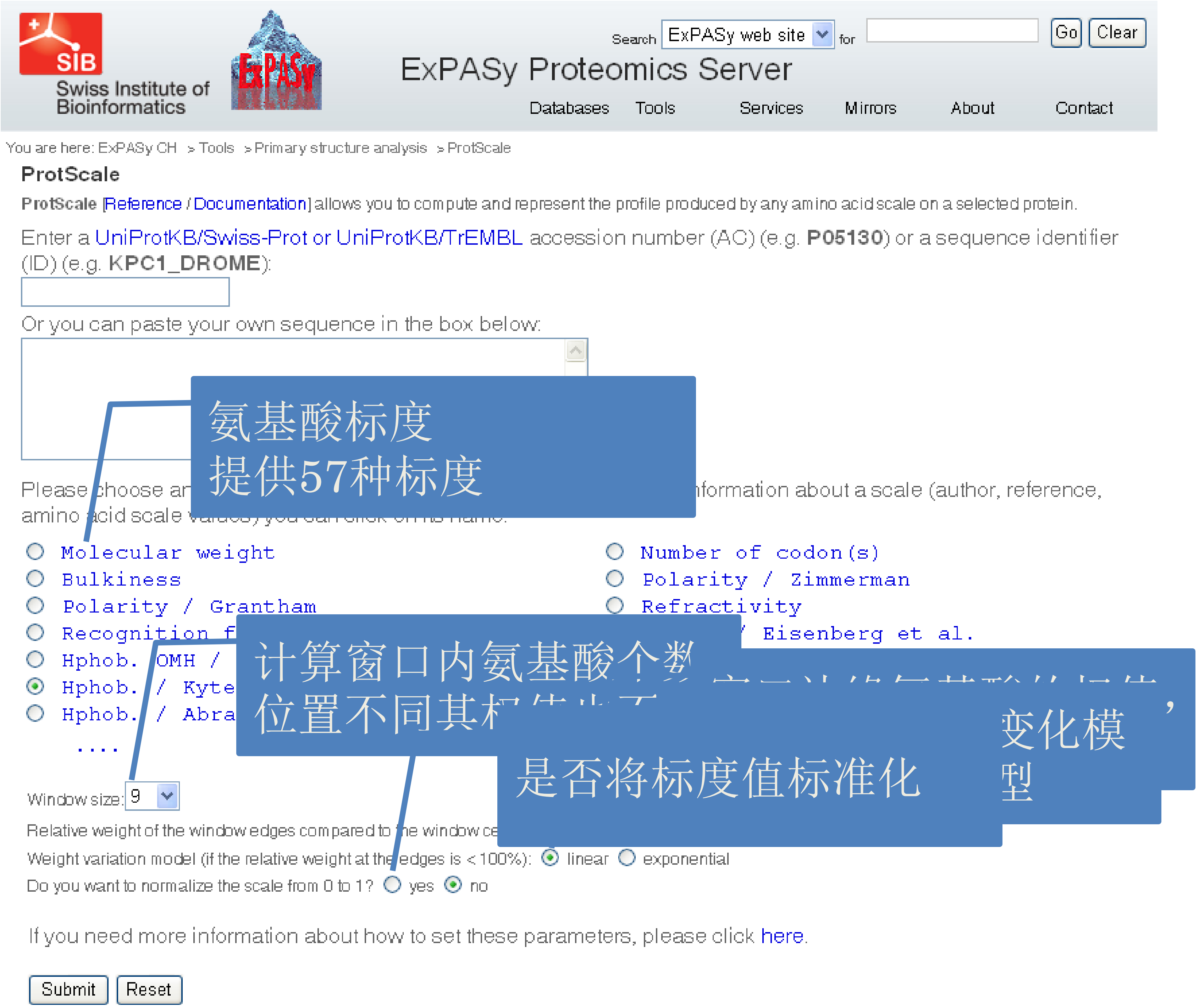
天冬氨酸（Asp）谷氨酸（Glu）

蛋白质的亲水性或疏水性

- 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内部疏水和外部亲水，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位置。

- ExPASy的ProtScale程序
  - https://web.expasy.org/protscale/



You are here: ExPASy CH > Tools > Primary structure analysis > ProtScale

ProtScale

ProtScale [Reference / Documentation] allows you to compute and represent the profile produced by any amino acid scale on a selected protein.

Enter a UniProtKB/Swiss-Prot or UniProtKB/TrEMBL accession number (AC) (e.g. **P05130**) or a sequence identifier (ID) (e.g. **KPC1_DROME**):

Or you can paste your own sequence in the box below:

Please choose an ... information about a scale (author, reference, amino acid scale values) you can click on its name:

○ Molecular weight
○ Bulkiness
○ Polarity / Grantham
○ Recognition f...
○ Hphob. OMH / ...
● Hphob. / Kyte ...
○ Hphob. / Abra...
....

○ Number of codon(s)
○ Polarity / Zimmerman
○ Refractivity
... / Eisenberg et al.

Window size: 9

Relative weight of the window edges compared to the window ce...

Weight variation model (if the relative weight at the edges is < 100%): ● linear ○ exponential

Do you want to normalize the scale from 0 to 1? ○ yes ● no

If you need more information about how to set these parameters, please click here.

Submit Reset

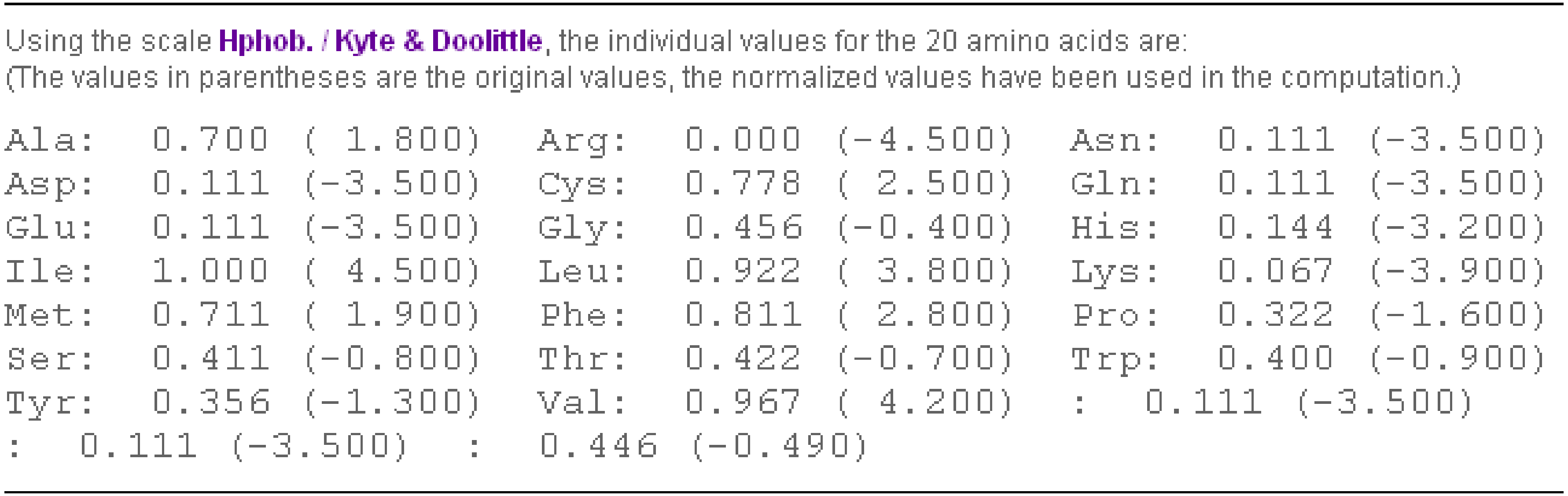
氨基酸标度
提供57种标度

计算窗口内氨基酸个数
位置不同其权

，
变化模型

是否将标度值标准化

Hohob./Kyte & Doolittle标度



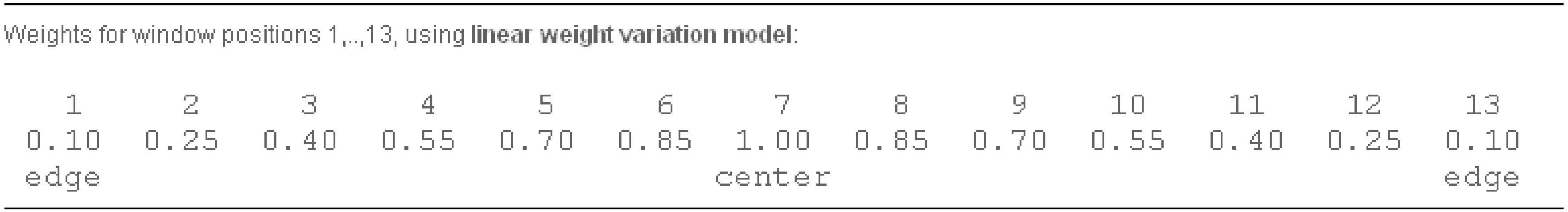
Using the scale **Hphob. / Kyte & Doolittle**, the individual values for the 20 amino acids are:
(The values in parentheses are the original values, the normalized values have been used in the computation.)

```
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)
```

计算窗口内每个位置上氨基酸的标度权值

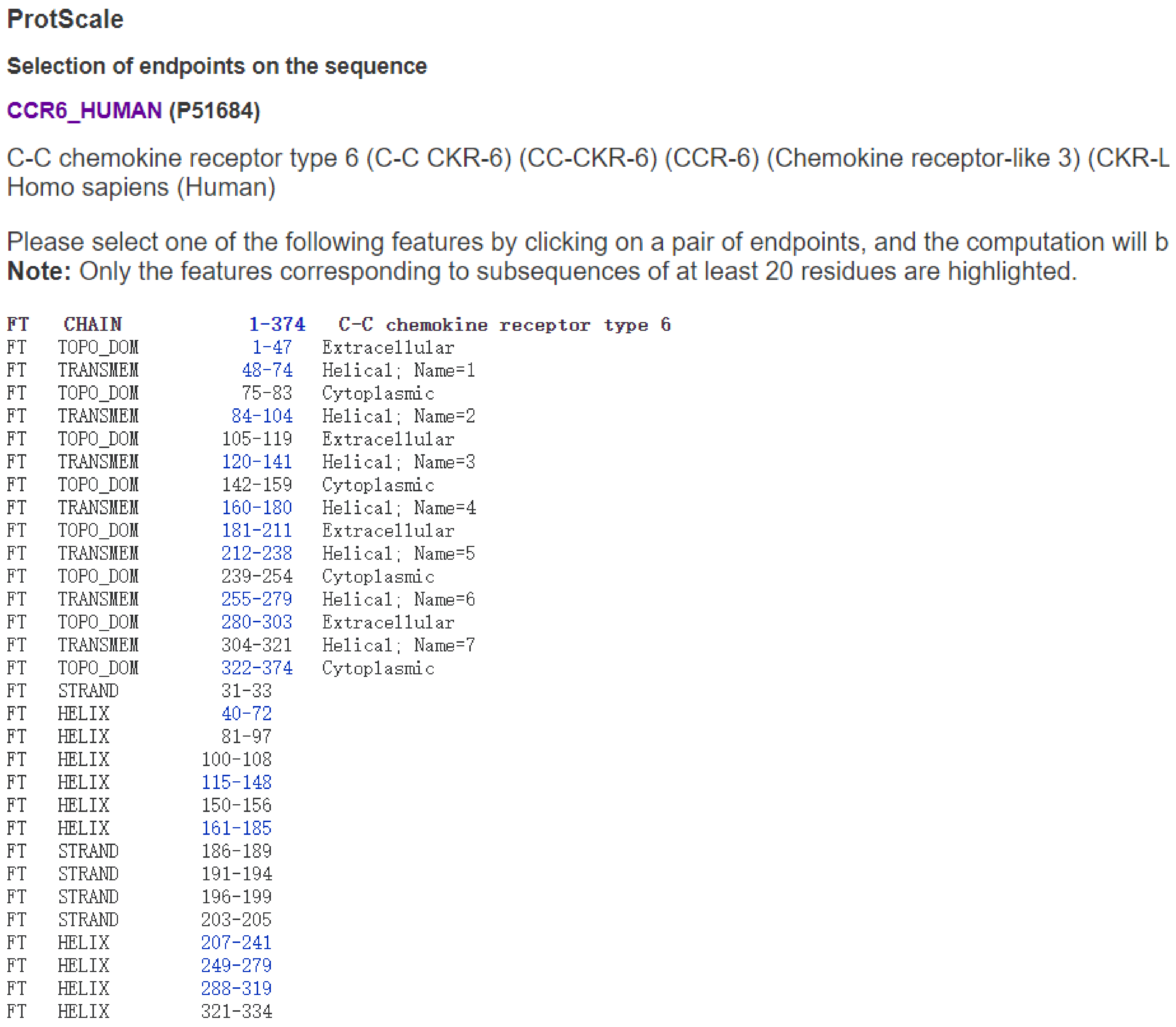
Window size=13， window edges=10%

Weight variation model=linear



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
```



ProtScale

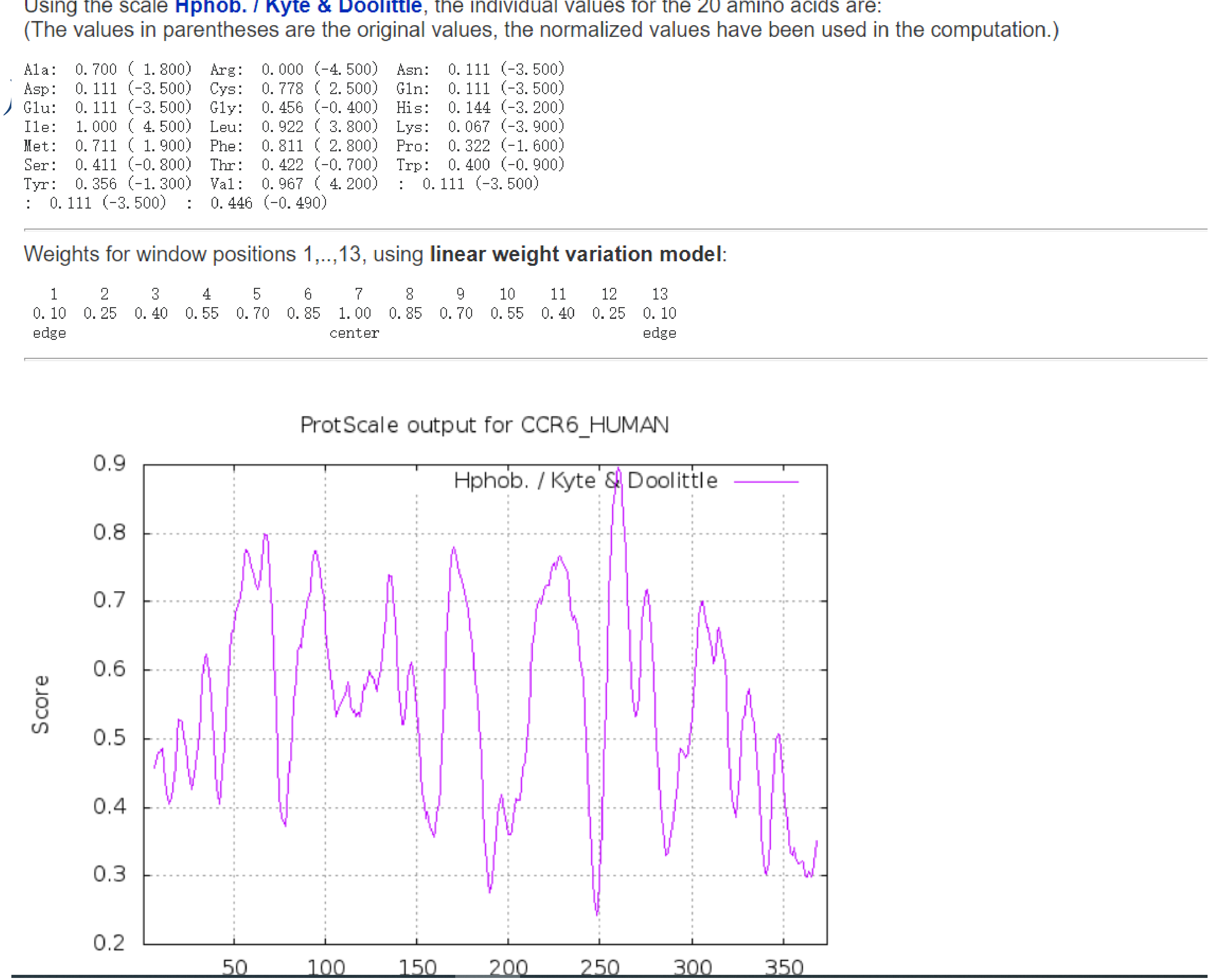
Selection of endpoints on the sequence

CCR6_HUMAN (P51684)

C-C chemokine receptor type 6 (C-C CKR-6) (CC-CKR-6) (CCR-6) (Chemokine receptor-like 3) (CKR-L
Homo sapiens (Human)

Please select one of the following features by clicking on a pair of endpoints, and the computation will b
Note: Only the features corresponding to subsequences of at least 20 residues are highlighted.

```
FT   CHAIN            1-374   C-C chemokine receptor type 6
FT   TOPO_DOM         1-47    Extracellular
FT   TRANSMEM        48-74    Helical; Name=1
FT   TOPO_DOM        75-83    Cytoplasmic
FT   TRANSMEM        84-104   Helical; Name=2
FT   TOPO_DOM       105-119   Extracellular
FT   TRANSMEM       120-141   Helical; Name=3
FT   TOPO_DOM       142-159   Cytoplasmic
FT   TRANSMEM       160-180   Helical; Name=4
FT   TOPO_DOM       181-211   Extracellular
FT   TRANSMEM       212-238   Helical; Name=5
FT   TOPO_DOM       239-254   Cytoplasmic
FT   TRANSMEM       255-279   Helical; Name=6
FT   TOPO_DOM       280-303   Extracellular
FT   TRANSMEM       304-321   Helical; Name=7
FT   TOPO_DOM       322-374   Cytoplasmic
FT   STRAND         31-33
FT   HELIX          40-72
FT   HELIX          81-97
FT   HELIX         100-108
FT   HELIX         115-148
FT   HELIX         150-156
FT   HELIX         161-185
FT   STRAND        186-189
FT   STRAND        191-194
FT   STRAND        196-199
FT   STRAND        203-205
FT   HELIX         207-241
FT   HELIX         249-279
FT   HELIX         288-319
FT   HELIX         321-334
```



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)
```

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
```

蛋白质的跨膜区

- 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。

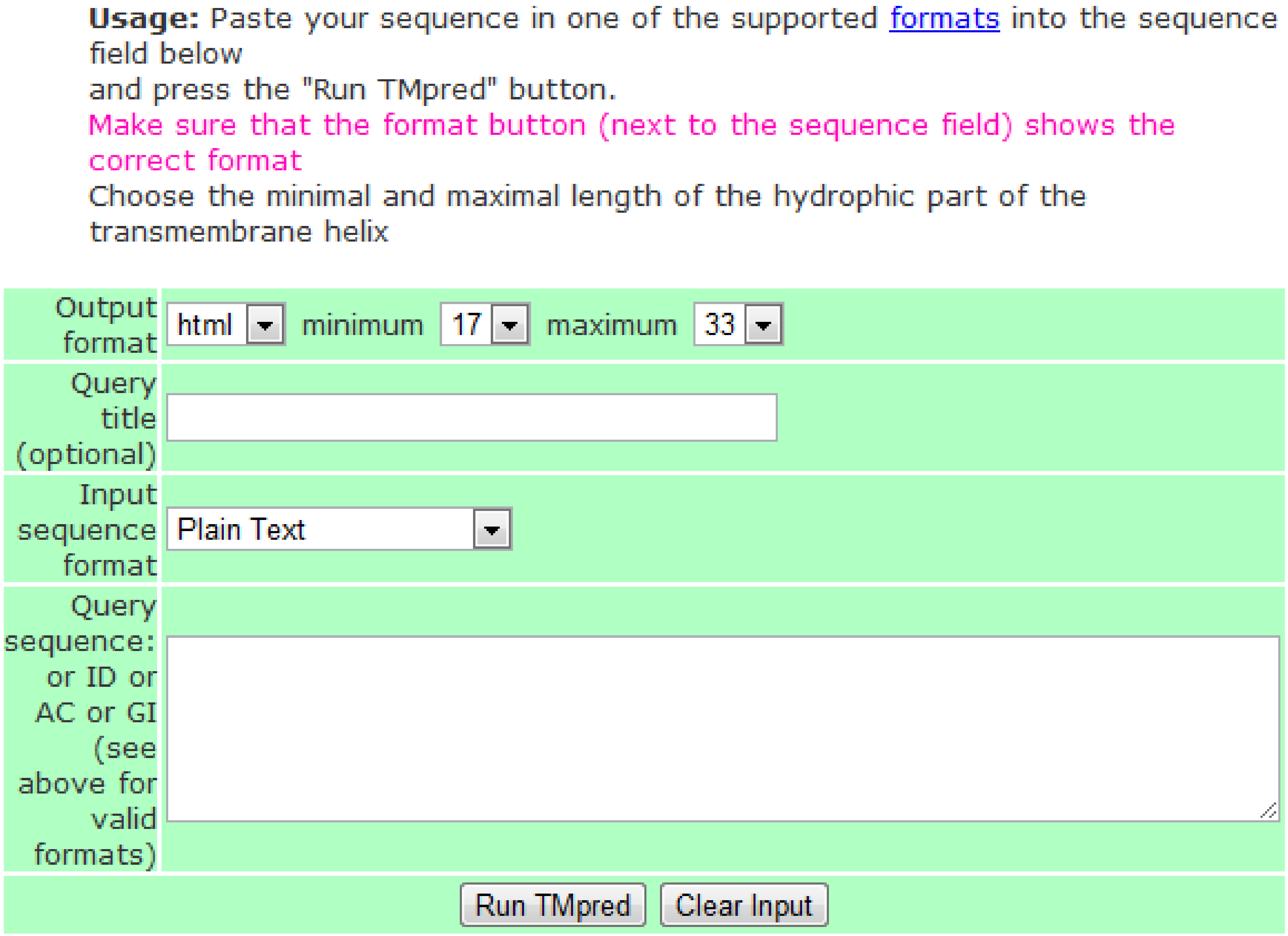
- 目前仅有少数膜蛋白的结构可被实验测得。

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。

- TMpred是EMBnet开发的一个分析蛋白质跨膜区的在线工具

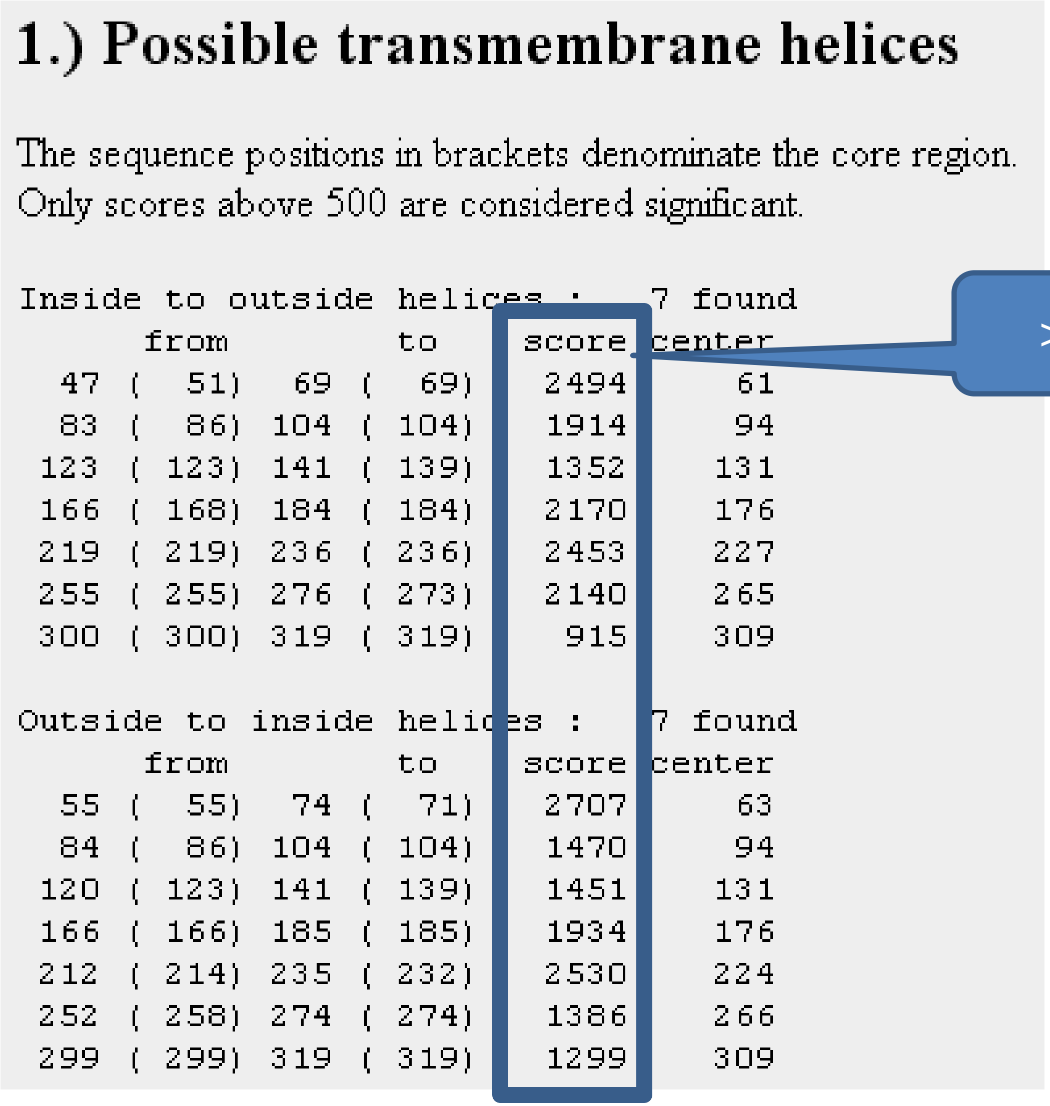
  https://embnet.vital-it.ch/software/TMPRED_form.html



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)	

用TMPRED分析P51684序列所得到的可能的7个跨膜螺旋区



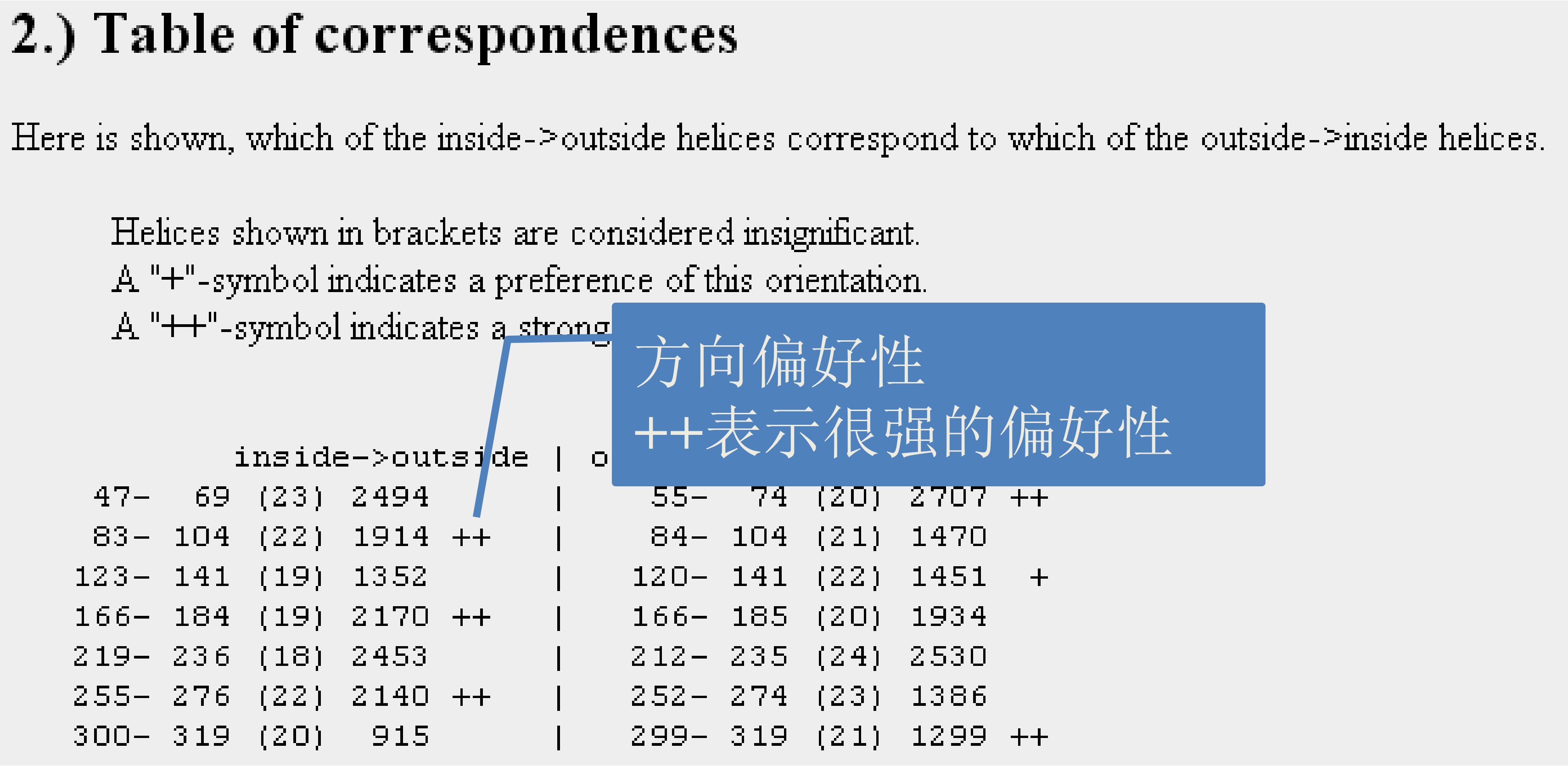
1.) Possible transmembrane helices

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
```

>500

可能的跨膜螺旋区的列表



2.) Table of correspondences

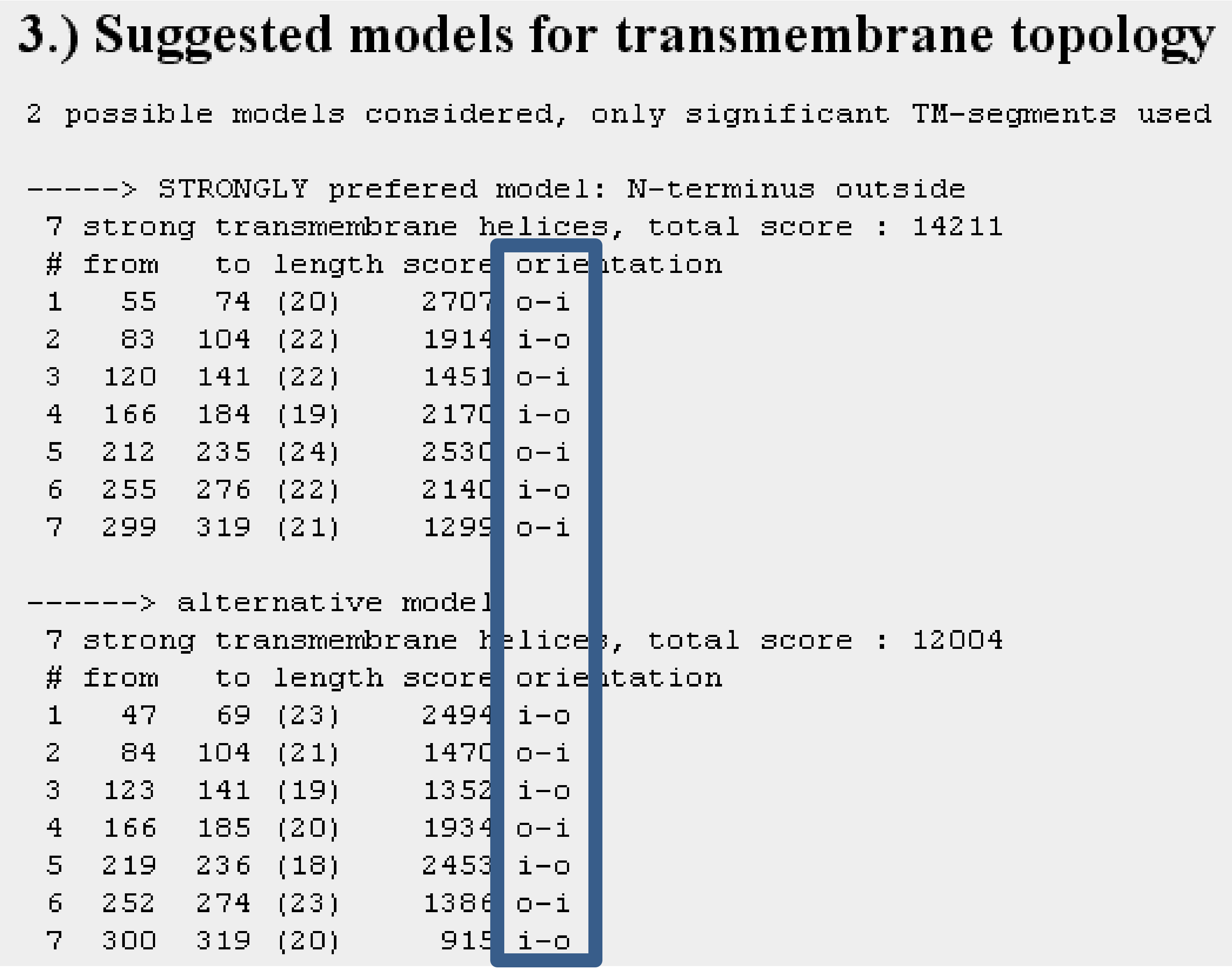
Here is shown, which of the inside->outside helices correspond to which of the outside->inside helices.

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

```
           inside->outside | o
 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 ++
```

方向偏好性
++表示很强的偏好性

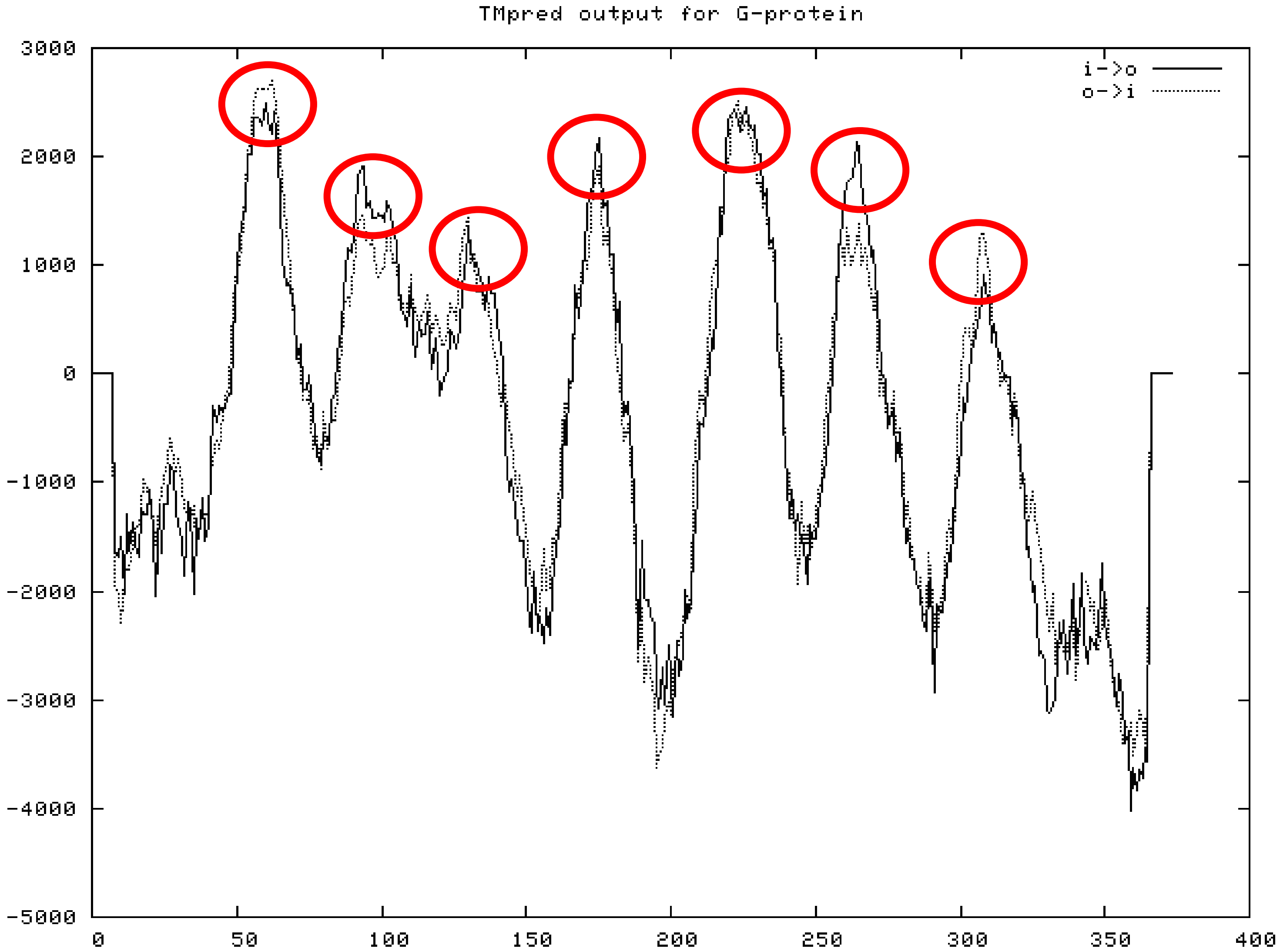
建议的跨膜拓扑模型



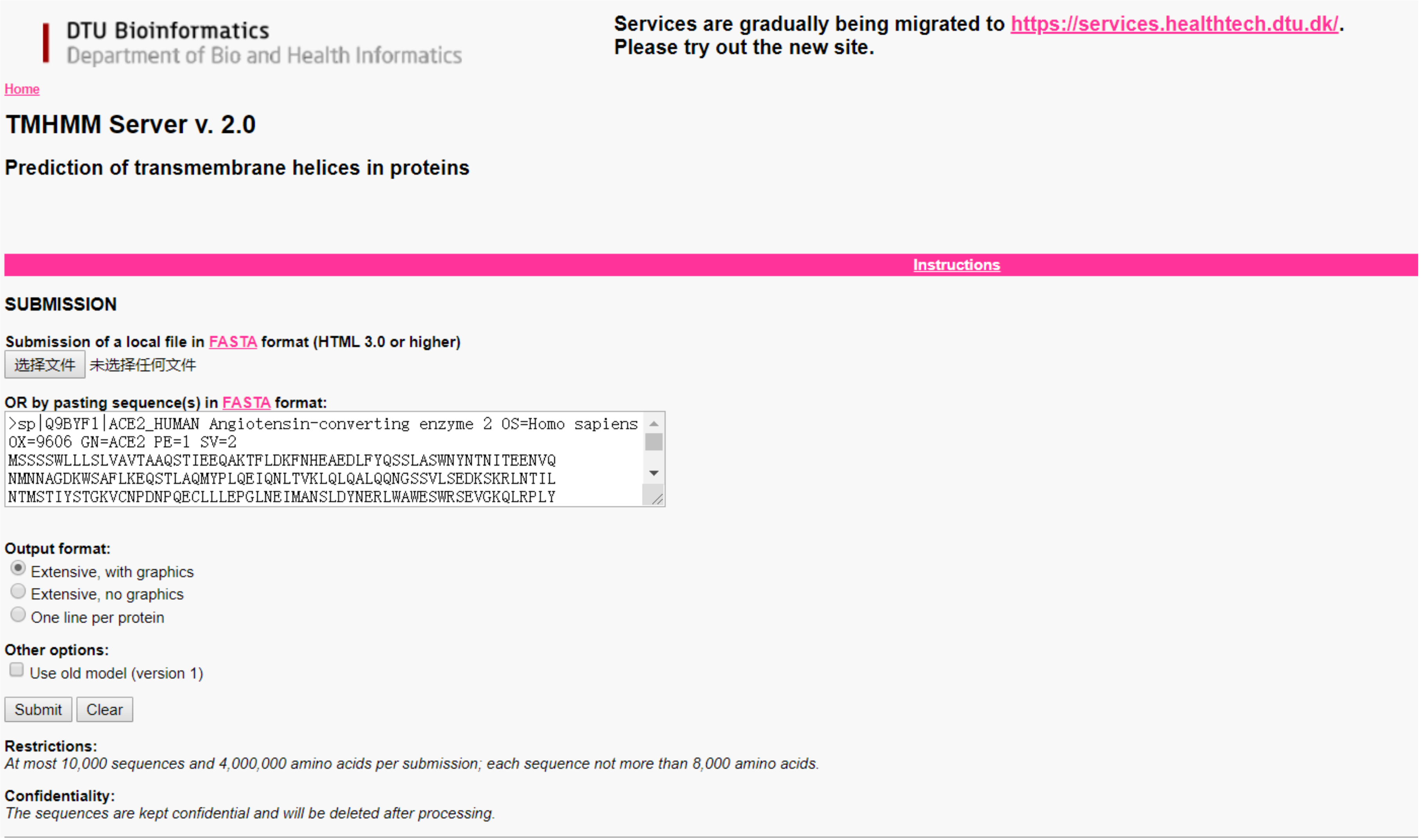
3.) Suggested models for transmembrane topology

```
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



TMHMM [HTTP://WWW.CBS.DTU.DK/SERVICES/TMHMM/](http://www.cbs.dtu.dk/services/TMHMM/)



DTU Bioinformatics
Department of Bio and Health Informatics

**Services are gradually being migrated to https://services.healthtech.dtu.dk/.
Please try out the new site.**

Home

TMHMM Server v. 2.0

Prediction of transmembrane helices in proteins

SUBMISSION

Submission of a local file in FASTA format (HTML 3.0 or higher)

OR by pasting sequence(s) in FASTA format:

```
>sp|Q9BYF1|ACE2_HUMAN Angiotensin-converting enzyme 2 OS=Homo sapiens
OX=9606 GN=ACE2 PE=1 SV=2
MSSSSWLLLSLVAVTAAQSTIEEQAKTFLDKFNHEAEDLFYQSSLASWNYNTNITEENVQ
NMNNAGDKWSAFLKEQSTLAQMYPLQEIQNLTVKLQLQALQQNGSSVLSEDKSKRLNTIL
NTMSTIYSTGKVCNPDNPQECLLLEPGLNEIMANSLDYNERLWAWESWRSEVGKQLRPLY
```

Output format:

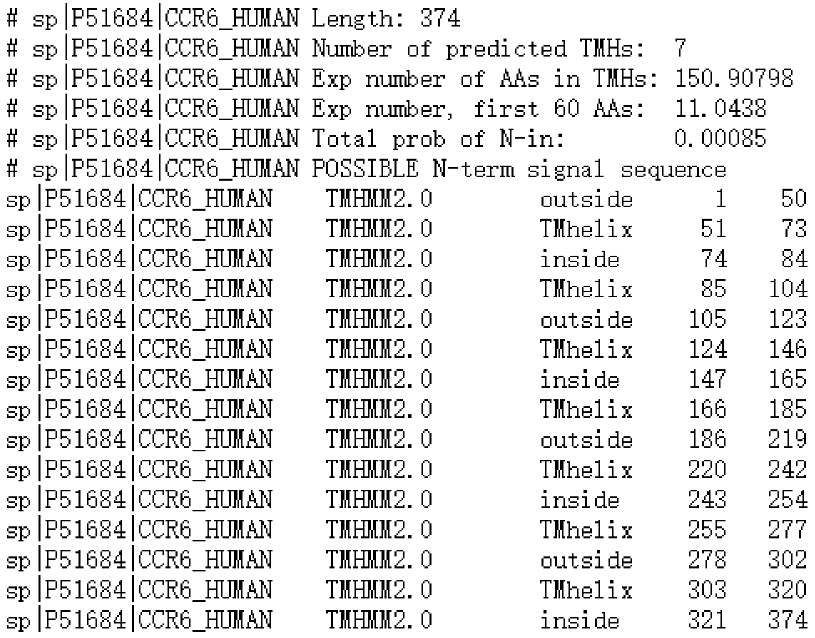
● Extensive, with graphics
○ Extensive, no graphics
○ One line per protein

Other options:

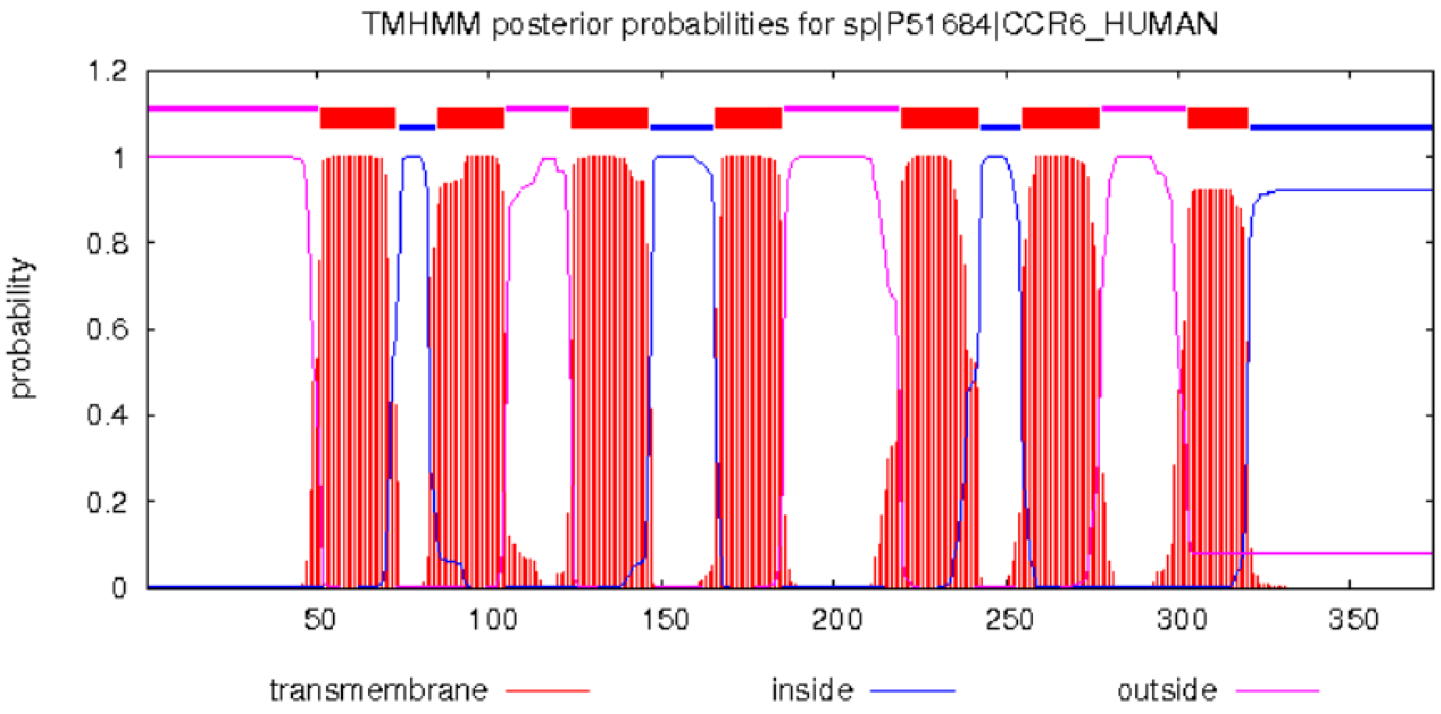
□ Use old model (version 1)

Restrictions:
At most 10,000 sequences and 4,000,000 amino acids per submission; each sequence not more than 8,000 amino acids.

Confidentiality:
The sequences are kept confidential and will be deleted after processing.



```
# sp|P51684|CCR6_HUMAN Length: 374
# sp|P51684|CCR6_HUMAN Number of predicted TMHs:  7
# sp|P51684|CCR6_HUMAN Exp number of AAs in TMHs: 150.90798
# sp|P51684|CCR6_HUMAN Exp number, first 60 AAs:  11.0438
# sp|P51684|CCR6_HUMAN Total prob of N-in:        0.00085
# sp|P51684|CCR6_HUMAN POSSIBLE N-term signal sequence
sp|P51684|CCR6_HUMAN    TMHMM2.0        outside      1    50
sp|P51684|CCR6_HUMAN    TMHMM2.0        TMhelix     51    73
sp|P51684|CCR6_HUMAN    TMHMM2.0        inside      74    84
sp|P51684|CCR6_HUMAN    TMHMM2.0        TMhelix     85   104
sp|P51684|CCR6_HUMAN    TMHMM2.0        outside    105   123
sp|P51684|CCR6_HUMAN    TMHMM2.0        TMhelix    124   146
sp|P51684|CCR6_HUMAN    TMHMM2.0        inside     147   165
sp|P51684|CCR6_HUMAN    TMHMM2.0        TMhelix    166   185
sp|P51684|CCR6_HUMAN    TMHMM2.0        outside    186   219
sp|P51684|CCR6_HUMAN    TMHMM2.0        TMhelix    220   242
sp|P51684|CCR6_HUMAN    TMHMM2.0        inside     243   254
sp|P51684|CCR6_HUMAN    TMHMM2.0        TMhelix    255   277
sp|P51684|CCR6_HUMAN    TMHMM2.0        outside    278   302
sp|P51684|CCR6_HUMAN    TMHMM2.0        TMhelix    303   320
sp|P51684|CCR6_HUMAN    TMHMM2.0        inside     321   374
```

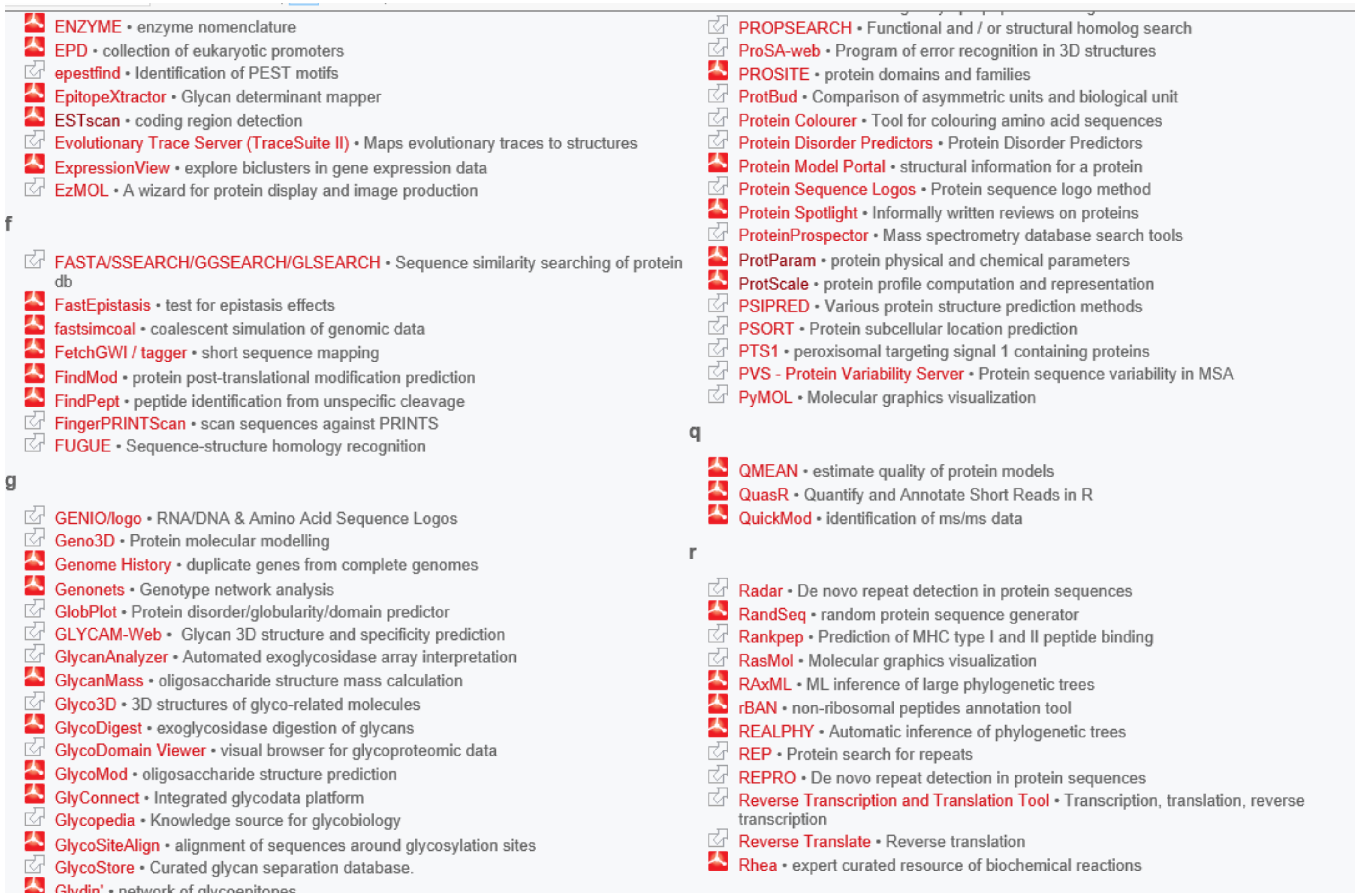


TMHMM

- Length: the length of the protein sequence.
- Number of predicted TMHs: The number of predicted transmembrane helices.
- Exp number of AAs in TMHs: The expected number of amino acids in transmembrane helices. If this number is larger than 18 it is very likely to be a transmembrane protein (OR have a signal peptide).
- Exp number, first 60 AAs: The expected number of amino acids in transmembrane helices in the first 60 amino acids of the protein. If this number more than a few, you should be warned that a predicted transmembrane helix in the N-term could be a signal peptide.
- Total prob of N-in: The total probability that the N-term is on the cytoplasmic side of the membrane.
- POSSIBLE N-term signal sequence: a warning that is produced when "Exp number, first 60 AAs" is larger than 10.

OTHER TOOLS

- https://www.expasy.org/resources



- ENZYME • enzyme nomenclature
- EPD • collection of eukaryotic promoters
- epestfind • Identification of PEST motifs
- EpitopeXtractor • Glycan determinant mapper
- ESTscan • coding region detection
- Evolutionary Trace Server (TraceSuite II) • Maps evolutionary traces to structures
- ExpressionView • explore biclusters in gene expression data
- EzMOL • A wizard for protein display and image production

f

- FASTA/SSEARCH/GGSEARCH/GLSEARCH • Sequence similarity searching of protein db
- FastEpistasis • test for epistasis effects
- fastsimcoal • coalescent simulation of genomic data
- FetchGWI / tagger • short sequence mapping
- FindMod • protein post-translational modification prediction
- FindPept • peptide identification from unspecific cleavage
- FingerPRINTScan • scan sequences against PRINTS
- FUGUE • Sequence-structure homology recognition

g

- GENIO/logo • RNA/DNA & Amino Acid Sequence Logos
- Geno3D • Protein molecular modelling
- Genome History • duplicate genes from complete genomes
- Genonets • Genotype network analysis
- GlobPlot • Protein disorder/globularity/domain predictor
- GLYCAM-Web • Glycan 3D structure and specificity prediction
- GlycanAnalyzer • Automated exoglycosidase array interpretation
- GlycanMass • oligosaccharide structure mass calculation
- Glyco3D • 3D structures of glyco-related molecules
- GlycoDigest • exoglycosidase digestion of glycans
- GlycoDomain Viewer • visual browser for glycoproteomic data
- GlycoMod • oligosaccharide structure prediction
- GlyConnect • Integrated glycodata platform
- Glycopedia • Knowledge source for glycobiology
- GlycoSiteAlign • alignment of sequences around glycosylation sites
- GlycoStore • Curated glycan separation database.
- Glydin' • network of glycoepitopes

- PROPSEARCH • Functional and / or structural homolog search
- ProSA-web • Program of error recognition in 3D structures
- PROSITE • protein domains and families
- ProtBud • Comparison of asymmetric units and biological unit
- Protein Colourer • Tool for colouring amino acid sequences
- Protein Disorder Predictors • Protein Disorder Predictors
- Protein Model Portal • structural information for a protein
- Protein Sequence Logos • Protein sequence logo method
- Protein Spotlight • Informally written reviews on proteins
- ProteinProspector • Mass spectrometry database search tools
- ProtParam • protein physical and chemical parameters
- ProtScale • protein profile computation and representation
- PSIPRED • Various protein structure prediction methods
- PSORT • Protein subcellular location prediction
- PTS1 • peroxisomal targeting signal 1 containing proteins
- PVS - Protein Variability Server • Protein sequence variability in MSA
- PyMOL • Molecular graphics visualization

q

- QMEAN • estimate quality of protein models
- QuasR • Quantify and Annotate Short Reads in R
- QuickMod • identification of ms/ms data

r

- Radar • De novo repeat detection in protein sequences
- RandSeq • random protein sequence generator
- Rankpep • Prediction of MHC type I and II peptide binding
- RasMol • Molecular graphics visualization
- RAxML • ML inference of large phylogenetic trees
- rBAN • non-ribosomal peptides annotation tool
- REALPHY • Automatic inference of phylogenetic trees
- REP • Protein search for repeats
- REPRO • De novo repeat detection in protein sequences
- Reverse Transcription and Translation Tool • Transcription, translation, reverse transcription
- Reverse Translate • Reverse translation
- Rhea • expert curated resource of biochemical reactions

Summary

- Analysis of DNA Sequence Characteristics

- Analysis of protein Sequence Characteristics

- Some tools