豌豆开花后特异表达基因 PPF-1 核酸和蛋白质序列分析

姓名 陈耿佳 学号 1301214752 组号 G01C

1. 研究背景和文献阅读

1) 利用各种不同方法，检索 PubMed 数据库中收录的开花后特异表达基因（Pea Post-floral-specific gene，PPF-1）相关研究论文。

2) 认真阅读上述论文，简述 PPF-1 序列特征、表达特异性，以及可能的生物学功能。

2. 数据库注释

1) 检索 UniProtKB 序列数据库中豌豆内膜蛋白 PPF-1 的蛋白质序列，归纳总结该序列条目的一般注释信息、序列注释信息、数据库交叉链接。

2) 利用注释信息中序列相似性和蛋白质家族注释信息，找出拟南芥中 PPF-1 同源蛋白 ALB3_ARATH。

3) 浏览该同源蛋白的注释信息、文献报道和数据库交叉链接，说明其功能、亚细胞定位、组织特异性、互作蛋白、结构域特征、剪接变体、序列特征、基因结构、基因组定位、表达特异性等。

4) 通过上述 ALB3_ARATH 序列中的数据库交叉链接 AT2G28800，浏览该基因在拟南芥信息资源系统（TAIR）中的注释信息，归纳总结其基因结构、可变剪接方式、突变体、文献资源等，并通过交叉链接 e-FP Browser 查看该基因的在不同组织中的表达，通过交叉链接 Phytozome Plant Gene Families 查看该基因在其它植物中的同源基因。

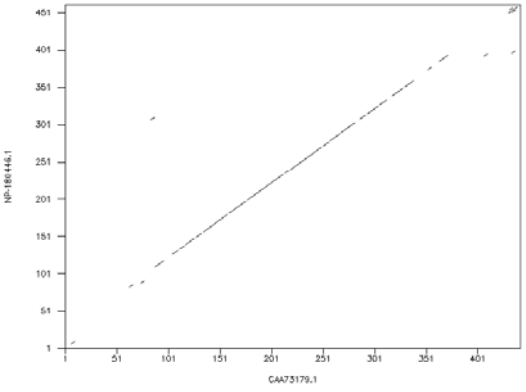
3. 序列相似性分析

1) 利用 WebLab 中的点阵图程序 Dottup、DotMatcher、DotPath 对 PPF1_PEA 和 ALB3_ARATH 蛋白质序列及其编码基因的编码区序列进行比对，分析比较比对结果，说明上述程序的适用范围。

Dottup:

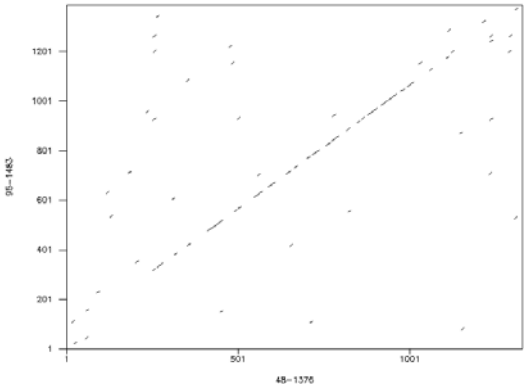
(A) Protein

(Word size = 4)



(B) CDS

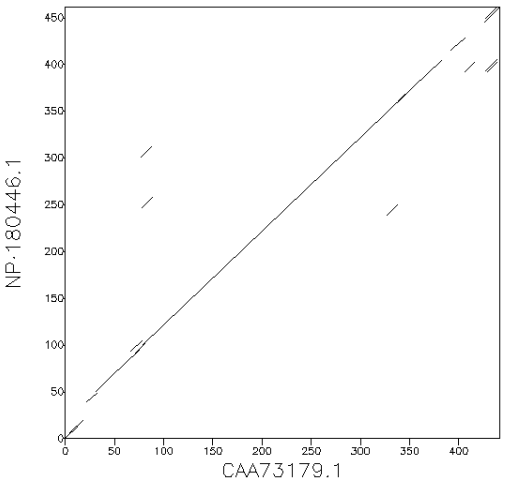
(Word size = 8)



DotMatcher

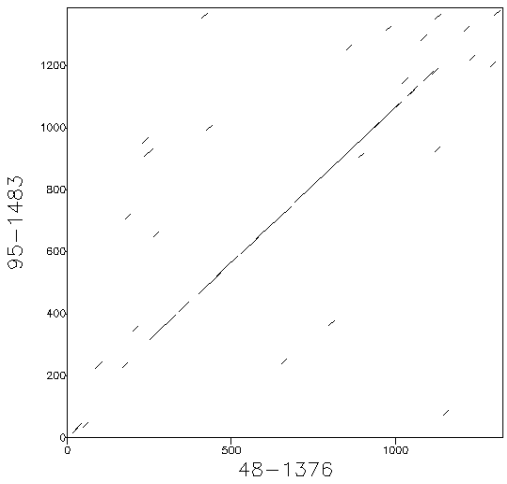
(A) Protein

(EBLOSUM62, Windowsize = 10, Threshold = 23)



(B) CDS

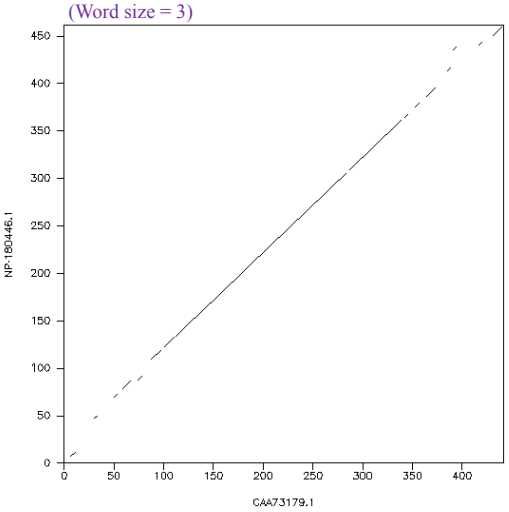
(EDNAFULL, Windowsize = 15, Threshold = 45)



DotPath

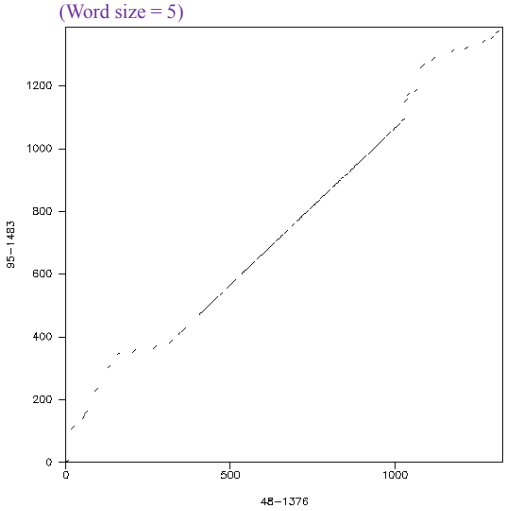
(A) Protein

(Word size = 3)



(B) CDS

(Word size = 5)



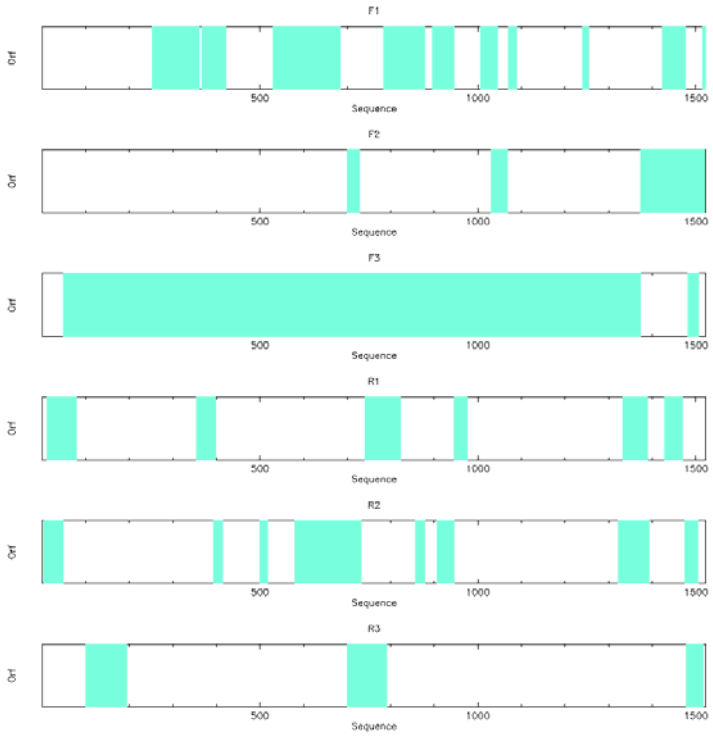
三种程序都可用于两个序列的比对。Dottup 是精确匹配，两个序列比对中，word size 内精确匹配时以图上的点表示，从点阵图则可直观看到两个序列的相似程度及匹配范围，当两个序列匹配程度较高时适用，如果两个序列差异较大则不适用；会重复匹配，存在符合 word size 的重复序列时会予以显示，当定义的 word size 较小时往往会在对角线外出现较多短线。DotMatcher 是近似匹配，用给定的计分矩阵对 windowsize 内的序列进行打分，高于 threshold 的在图上以点显示，可用于匹配程度不高的两个序列；也存在重复匹配。DotPath 与 Dottup 相似，是 word size 内的精确匹配，但不进行重复匹配（non-overlapping wordmatch）。三种程序均有相应的参数设置，需根据具体情况及实验目的进行设置，用 RNA 序列进行匹配时，相对于蛋白序列而言 word size 或 windowsize 应适当提高。PPF1_PEA 和 ALB3_ARATH 的序列相似程度并不高，相同位点为 327/466 (70.2%)，用 Dottup 和 DotPath 绘制点阵图时匹配程度不高，尤其是序列两段几乎没有点，用 DotMatcher 则可较好地匹配；用蛋白序列进行点阵图绘制，相对于核酸序列而言更加准确，主要是因为蛋白序列具有较高的特异性和保守性。

2) 利用 WebLab 中的全局比对程序 Needle 和局部比对程序 Water，对 PPF1_PEA 和 ALB3_ARATH 蛋白质序列及其编码基因的编码区序列进行比对，分析比较比对结果，说明上述程序的实用范围。

3. 读码框分析

1) 提取豌豆内膜蛋白编码基因 PPF-1 全长 mRNA 序列，用 WebLab 中 PlotORF 程序分析其可能的读码框。

结果如下图所示：



由上图可见，F3 中起止密码子间有较长的连续序列，可能为正确的读码框；其它读码框均只能得到较短的序列，可能性较低。

2) 用 WebLab 中 ShowORF 程序分析 PPF-1 全长 mRNA 序列读码框特征。

WebLab 中 ShowORF 程序不可用，改用 Mobyle 平台的，结果如下（仅给出 1-150）

```
SHOWORF of  from 1 to 1523

          ---------|---------|---------|---------|---------|
         1 CTCAAGCCTTCAAGCCTGAAGCGTCTCGTACACAAACCTTCTCATCCATG 50
F1       1 L  K  P  S  S  L  K  R  L  V  H  K  P  S  H  P  W  17
F2       1  S  S  L  Q  A  *  S  V  S  Y  T  N  L  L  I  H  G 11
F3       1   Q  A  F  K  P  E  A  S  R  T  Q  T  F  S  S  M   16
R1       3   E  L  R  *  A  Q  L  T  E  Y  V  F  R  R  M  W   25
R2      33    *  A  K  L  G  S  A  D  R  V  C  V  K  E  D  M  19
R3      13     L  G  E  L  R  F  R  R  T  C  L  G  E  *  G  H 9

          ---------|---------|---------|---------|---------|
        51 GCGAAGACACTGATTTCTTCTCCATCATTCCTCGGTACTCCACTTCCTTC 100
F1      18  R  R  H  *  F  L  L  H  H  S  S  V  L  H  F  L  H 13
F2      12   E  D  T  D  F  F  S  I  I  P  R  Y  S  T  S  F   27
F3      17 A  K  T  L  I  S  S  P  S  F  L  G  T  P  L  P  S  33
R1      24 P  S  S  V  S  K  K  E  M  M  G  R  Y  E  V  E  K  8
R2      18  A  F  V  S  I  E  E  G  D  N  R  P  V  G  S  G  E 2
R3       8   R  L  C  Q  N  R  R  W  *  E  E  T  S  W  K  R   1

          ---------|---------|---------|---------|---------|
       101 ACTTCACCGTACTTTCTCCCCTAATCGCACCAGGCTTTTCACCAAAGTTC 150
F1      14   F  T  V  L  S  P  L  I  A  P  G  F  S  P  K  F   29
F2      28 T  S  P  Y  F  L  P  *  S  H  Q  A  F  H  Q  S  S  9
F3      34  L  H  R  T  F  S  P  N  R  T  R  L  F  T  K  V  Q 50
R1       7  V  E  G  Y  K  R  G  *  D  C  W  A  K  *  W  L  E 2
R2       1   S  *  R  V  K  E  G  L  R  V  L  S  K  V  L  T   1
R3      34 *  K  V  T  S  E  G  R  I  A  G  P  K  E  G  F  N  19
```

F3 可翻译为较完整蛋白序列，而其他读码框均为多条小片段，可见 F3 可能为正确的读码框。

3) 用 WebLab 中 SixPack 程序分析 PPF-1 全长 mRNA 序列读码框特征。

用 Mobyle 中 SixPack 进行分析，翻译结果如下图所示（仅显示 1-120）：

```
   L  K  P  S  S  L  K  R  L  V  H  K  P  S  H  P  W  R  R  H     F1
    S  S  L  Q  A  *  S  V  S  Y  T  N  L  L  I  H  G  E  D  T    F2
     Q  A  F  K  P  E  A  S  R  T  Q  T  F  S  S  M  A  K  T  L   F3
1  CTCAAGCCTTCAAGCCTGAAGCGTCTCGTACACAAACCTTCTCATCCATGGCGAAGACAC 60
   ----:----|----:----|----:----|----:----|----:----|----:----|
1  GAGTTCGGAAGTTCGGACTTCGCAGAGCATGTGTTTGGAAGAGTAGGTACCGCTTCTGTG 60
    X  L  G  E  L  R  F  R  R  T  C  L  G  E  *  G  H  R  L  C    F6
   X  *  A  K  L  G  S  A  D  R  V  C  V  K  E  D  M  A  F  V     F5
     E  L  R  *  A  Q  L  T  E  Y  V  F  R  R  M  W  P  S  S  V   F4


   *  F  L  L  H  H  S  S  V  L  H  F  L  H  F  T  V  L  S  P     F1
    D  F  F  S  I  I  P  R  Y  S  T  S  F  T  S  P  Y  F  L  P    F2
     I  S  S  P  S  F  L  G  T  P  L  P  S  L  H  R  T  F  S  P   F3
61  TGATTTCTTCTCCATCATTCCTCGGTACTCCACTTCCTTCACTTCACCGTACTTTCTCCC 120
   ----:----|----:----|----:----|----:----|----:----|----:----|
61  ACTAAAGAAGAGGTAGTAAGGAGCCATGAGGTGAAGGAAGTGAAGTGGCATGAAAGAGGG 120
    Q  N  R  R  W  *  E  E  T  S  W  K  R  *  K  V  T  S  E  G    F6
   S  I  E  E  G  D  N  R  P  V  G  S  G  E  S  *  R  V  K  E     F5
     S  K  K  E  M  M  G  R  Y  E  V  E  K  V  E  G  Y  K  R  G   F4
```

在给出的翻译序列列表中，最长的序列如下：

```
>_3_ORF1  Translation of  in frame 3, ORF 1, threshold 1, 457aa
QAFKPEASRTQTFSSMAKTLISSPSFLGTPLPSLHRTFSPNRTRLFTKVQFSFHQLPPIQ
SVSHSVDLSGIFARAEGLLYTLADATVAADAAASTDVAAQKNGGWFGFISDGMEFVLKVL
KDGLSSVHVPYSYGFAIILLTVIVKAATLPLTKQQVESTLAMQNLQPKIKAIQERYAGNQ
ERIQLETSRLYTQAGVNPLAGCLPTLATIPVWIGLYQALSNVANEGLLTEGFLWIPSLGG
PTSIAARQSGSGISWLFPFVDGHPLLGWYDTAAYLVLPVLLIVSQYVSMEIMKPPQTNDP
NQKNTLLIFKFLPLMIGYFSLSVPSGLTIYWFTNNVLSTAQQVWLRKLGGAKPAVNENAG
GIITAGQAKRSASKPEKGGERFRQLKEEEKKKKLIKALPVEEVQPLASASASNDGSDVEN
NKEQEVTEESNTSKVSQEVQSFSRERRSKRSKRKPVA
```

其余序列均为短序列（最长为 91aa），可见上述 457aa 序列最可能为真实翻译结果，F3 可能为正确读码框。

4) 用 WebLab 中 GetORF 程序提取 PPF-1 全长 mRNA 序列中编码区核苷酸序列和所编码的氨基酸序列。

运行结果与 SixPack 给出的翻译序列相似，其中最长的序列如下：

```
>Y12618.1_32 [3 - 1373] Pisum sativum mRNA for PPF-1 protein
QAFKPEASRTQTFSSMAKTLISSPSFLGTPLPSLHRTFSPNRTRLFTKVQFSFHQLPPIQ
SVSHSVDLSGIFARAEGLLYTLADATVAADAAASTDVAAQKNGGWFGFISDGMEFVLKVL
KDGLSSVHVPYSYGFAIILLTVIVKAATLPLTKQQVESTLAMQNLQPKIKAIQERYAGNQ
ERIQLETSRLYTQAGVNPLAGCLPTLATIPVWIGLYQALSNVANEGLLTEGFLWIPSLGG
PTSIAARQSGSGISWLFPFVDGHPLLGWYDTAAYLVLPVLLIVSQYVSMEIMKPPQTNDP
NQKNTLLIFKFLPLMIGYFSLSVPSGLTIYWFTNNVLSTAQQVWLRKLGGAKPAVNENAG
GIITAGQAKRSASKPEKGGERFRQLKEEEKKKKLIKALPVEEVQPLASASASNDGSDVEN
NKEQEVTEESNTSKVSQEVQSFSRERRSKRSKRKPVA
```

其余序列均为短序列（最长为 91aa），可见上述氨基酸序列最可能为真实翻译结果，相应的编码区核苷酸序列为全长 mRNA 序列的 3-1373 位。

5) 比较上述读码框分析软件，说明其用途和特点。

PlotORF 是根据给定的起止密码子来划分 ORF，并以图形形式展示所有 6 中读码框下的 ORF，非常直观但没有显示精确位置，可以根据实际需求定义起止密码子，适用于原核或者 mRNA 真核序列，但当所给真核序列不包含起始密码子时会丢失部分外显子。

ShowORF 将所给核酸序列按照所选读码框（或全部 6 种读码框）及密码子翻译为氨基酸序列，将所给核酸序列及所选/全部读码框下翻译的氨基酸序列并排显示，起止密码子以星号显示，根据所翻译的氨基酸序列可以推断读码框是否合理，可以直接看到所选/全部读码框下的翻译情况，但不够直观。

SixPack 与 ShowORF 相似，展示所给核酸序列及六种读码框下翻译的氨基酸序列，还统计了每种读码框下 ORF 的个数，并将每种读码框下得到的符合限定要求的氨基酸序列以 fasta 格式给出，可以较直观得到所有可能的氨基酸序列。

GetORF 根据给定的起止密码子来划分 ORF，并将符合限定要求的 ORF 翻译为氨基酸序列以 fasta 格式给出，可得到符合限定要求的所有翻译序列及其对应的读码框和起止位置。

4. 核苷酸序列分析

1) 利用 WebLab 中密码子统计程序，分析豌豆 PPF-1 和拟南芥中同源基因。

用 Cusp 程序，分别以 PPF1_PEA 和 ALB3_ARATH 编码基因的编码区序列作为输入序列进行密码子统计，两者的密码子使用偏好如下图所示：

	PPF1_PEA	ALB3_ARATH
Coding GC	43.04%	44.49%
1st letter GC	51.92%	52.92%
2nd letter GC	43.34%	44.49%
3rd letter GC	33.86%	36.07%

可见，豌豆 PPF-1 和拟南芥中同源基因 ALB3_ARATH 密码子使用偏好很接近，以 AT 结尾的密码子的使用频率都高于以 GC 结尾的密码子，且两者三个位置的 GC 使用偏好也都较为接近。

2) 利用 WebLab 中内切酶分析程序，分析豌豆 PPF-1 基因的酶切位点。

用 Remap 程序对 PPF1_PEA mRNA 序列进行分析，限制只输出单一酶切位点的酶，得 103 个酶及其酶切位点，其中第 1-60 位序列的酶切位点如下所示：

```
                                        StyI
                                        NcoI
                                        ErhI
                                        BtgI
                                        BstDSI
                                        BssT1I
                                AcuI    Bsp19I
                        Esp3I   Eco57I  EcoT14I
                        BsmBI   Eco57MI Eco130I
                        \       \       \
CTCAAGCCTTCAAGCCTGAAGCGTCTCGTACACAAACCTTCTCATCCATGGCGAAGACAC
         10        20        30        40        50        60
----:----|----:----|----:----|----:----|----:----|----:----|
GAGTTCGGAAGTTCGGACTTCGCAGAGCATGTGTTTGGAAGAGTAGGTACCGCTTCTGTG
                              /  /              /
                              |  Eco57MI        Eco130I
                              |  Eco57I         EcoT14I
                              |  AcuI           Bsp19I
                              BsmBI             BssT1I
                              Esp3I             BstDSI
                                                BtgI
                                                ErhI
                                                NcoI
                                                StyI
```

报告还给出了有多个酶切位点的酶以及对该序列没有酶切位点的酶的列表，根据需求可以更改参数设置得到所需的酶及酶切位点信息，可以根据实验目的及这些酶切位点信息进行实验方案设计。

3) 利用 WebLab 中引物设计程序，设计 PPF-1 基因 mRNA 序列的引物。

用 Eprimer32 对 PPF1_PEA mRNA 序列进行分析，设置产物长度范围为 1380-1500，从输出结果列表中选择如下引物：

	Start	Len	Tm	GC%	Sequence
Forward primer	9	25	59.29	52.00	TTCAAGCCTGAAGCGTCTCGTACAC
Reverse primer	1365	27	57.66	48.15	GTATGTGGTCCACTATCATGCAACAGG
Product size	1383				

利用这对引物可以获得 PPF1_PEA mRNA 序列的 9-1392 位的 PCR 产物，其中包含了全部编码区信息（48-1376）。

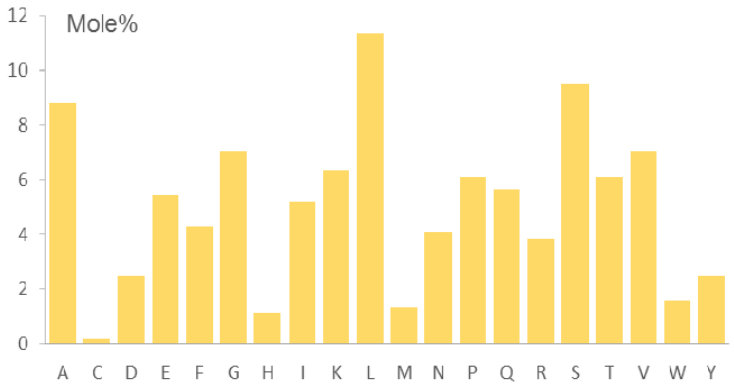
5. 蛋白质序列分析

1) 利用 WebLab 和 ExPASy 网站提供的氨基酸组成分析程序，统计 PPF-1 蛋白质 20 种不同氨基酸的组成。

用 WebLab 上的 Pepstats 对 PPF1_PEA 氨基酸序列进行分析，其氨基酸组成如下：

Residue	Number	Mole%	DayhoffStat
A = Ala	39	8.824	1.026
C = Cys	1	0.226	0.078
D = Asp	11	2.489	0.452
E = Glu	24	5.430	0.905
F = Phe	19	4.299	1.194
G = Gly	31	7.014	0.835
H = His	5	1.131	0.566
I = Ile	23	5.204	1.156
K = Lys	28	6.335	0.960
L = Leu	50	11.312	1.529
M = Met	6	1.357	0.799
N = Asn	18	4.072	0.947
P = Pro	27	6.109	1.175
Q = Gln	25	5.656	1.450
R = Arg	17	3.846	0.785
S = Ser	42	9.502	1.357
T = Thr	27	6.109	1.001
V = Val	31	7.014	1.063
W = Trp	7	1.584	1.218
Y = Tyr	11	2.489	0.732

作柱状图如下所示：



报告还给出了氨基酸属性的统计信息，如下所示：

Property	Residues	Number	Mole%
Tiny	(A+C+G+S+T)	140	31.674
Small	(A+C+D+G+N+P+S+T+V)	227	51.357
Aliphatic	(A+I+L+V)	143	32.353
Aromatic	(F+H+W+Y)	42	9.502
Non-polar	(A+C+F+G+I+L+M+P+V+W+Y)	245	55.430
Polar	(D+E+H+K+N+Q+R+S+T)	197	44.570
Charged	(D+E+H+K+R)	85	19.231
Basic	(H+K+R)	50	11.312
Acidic	(D+E)	35	7.919

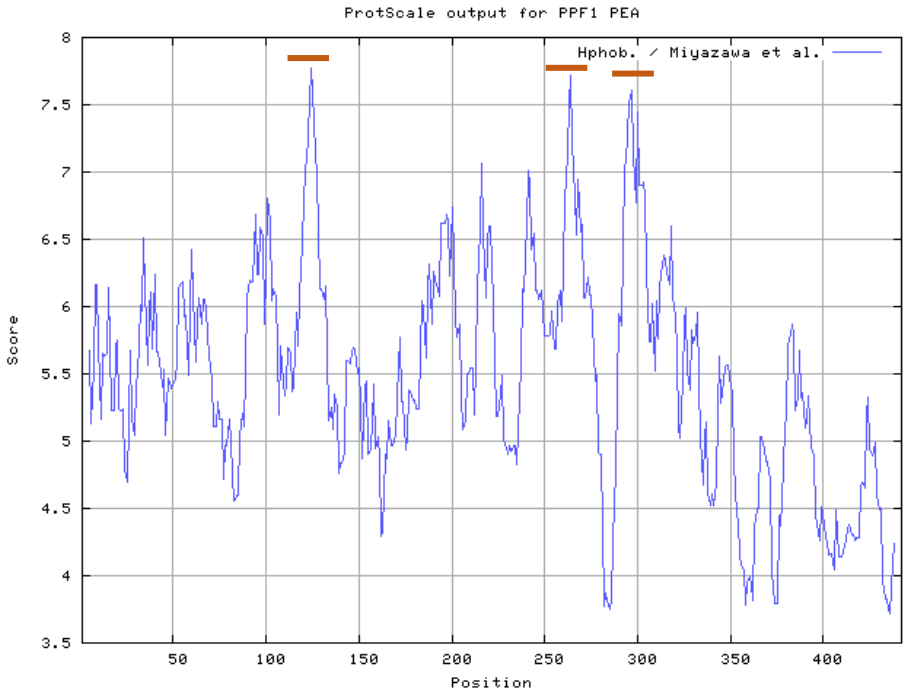
用 ExPASy 上的 ProtParam 工具对 PPF1_PEA 氨基酸序列进行分析，统计结果如下：

Residue	Number	Mole%
A	39	8.82
C	1	0.23
D	11	2.49
E	24	5.43
F	19	4.30
G	31	7.01
H	5	1.13
I	23	5.20
K	28	6.33
L	50	11.31
M	6	1.36
N	18	4.07
P	27	6.11
Q	25	5.66
R	17	3.85
S	42	9.50
T	27	6.11
V	31	7.01
W	7	1.58
Y	11	2.49

2) 利用 WebLab 和 ExPASy 网站提供的一级结构特征分析程序，分析 PPF-1 蛋白质不同区域疏水和亲水、柔性和刚性、溶剂可及性、空间位阻、二级结构等序列特征。

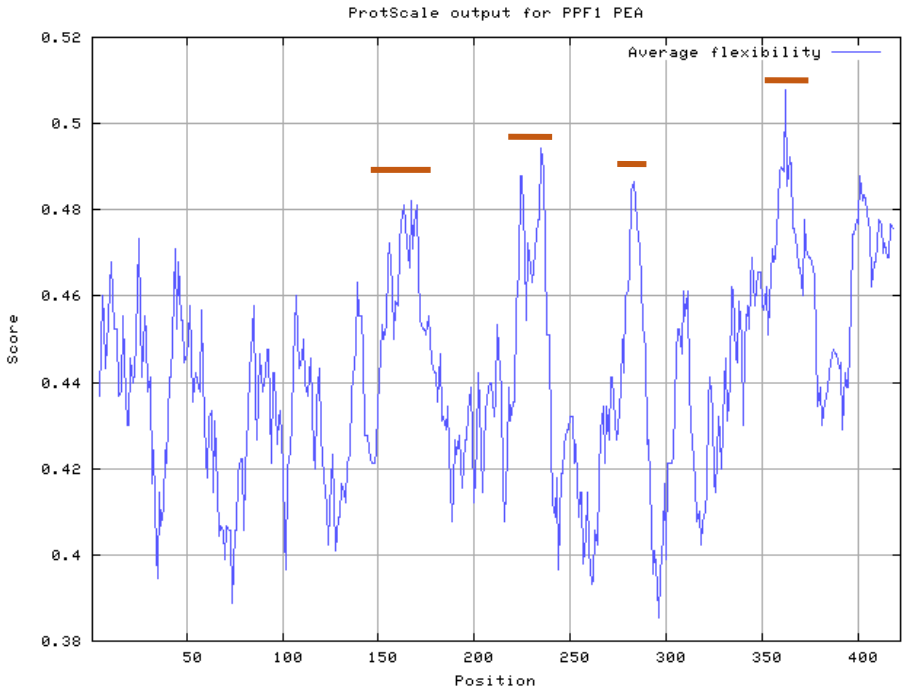
用 ExPASy 上的 ProtScale 工具对 PPF1_PEA 氨基酸序列进行分析：

疏水和亲水：（Hphob. / Miyazawa et al）



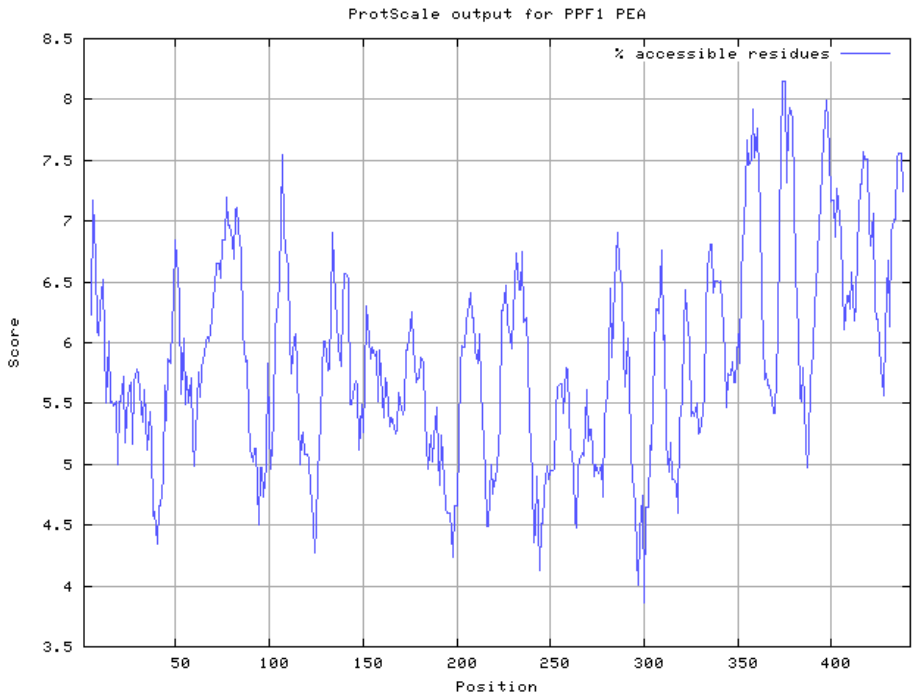
如上图所示为 PPF1_PEA 蛋白质的疏水性图谱，数值越高表明该区域多肽链的疏水性越强，PPF1 存在数段较为显著的疏水区域，特别是红线所标注的区域

柔性和刚性：（Average flexibility）



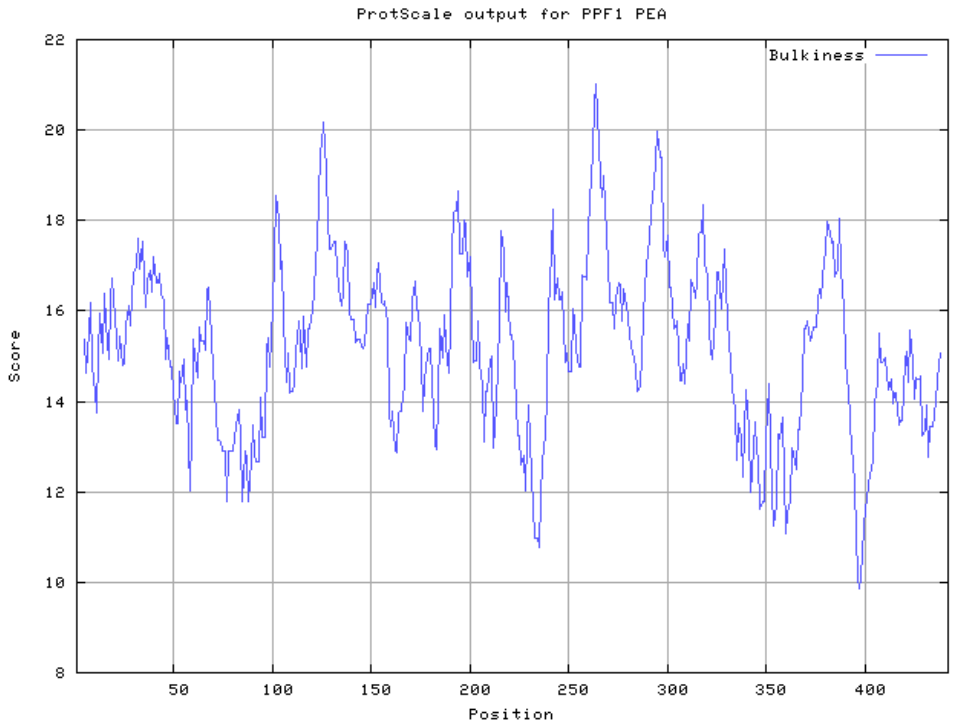
如上图所示为 PPF1_PEA 蛋白质序列的柔性图谱，数值越高，表明该区域多肽链的柔性越强，也即可变性越高，对比疏水性图谱可发现，两者大致呈现相反趋势，即疏水性较强的区域柔性一般比较低。

溶剂可及性：（% accessible residues）



如上图所示为 PPF1_PEA 氨基酸残基溶剂可及性图谱，得分越高表明该区域氨基酸残基的溶剂可及性越高。

空间位阻：（Bulkiness）

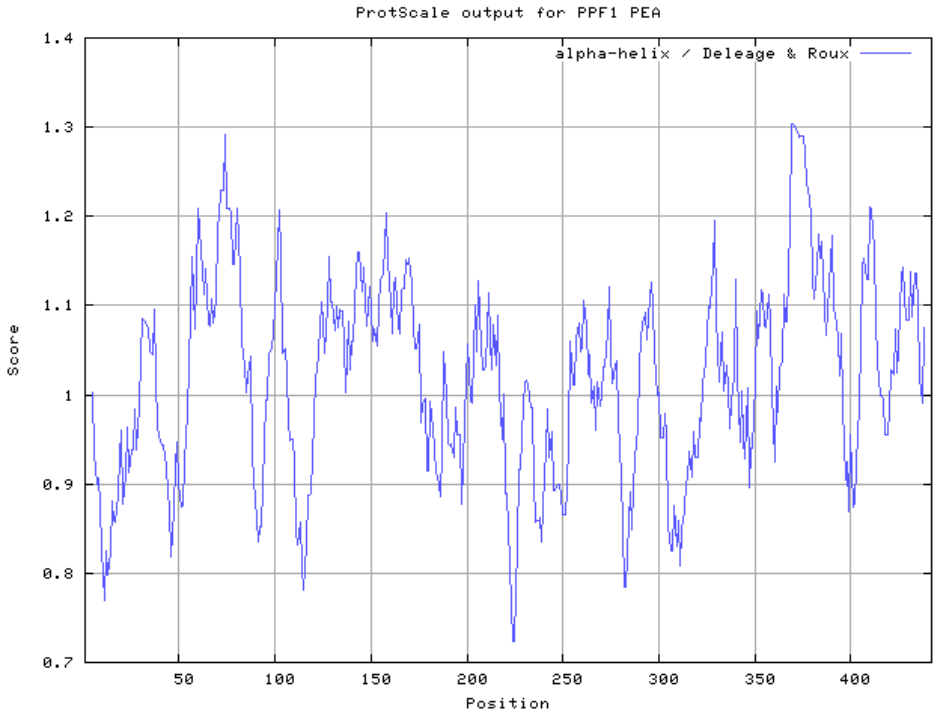


如上图所示为 PPF1_PEA 氨基酸残基空间位阻图谱，得分越高表明该区域氨基酸残基的空间位阻越大。

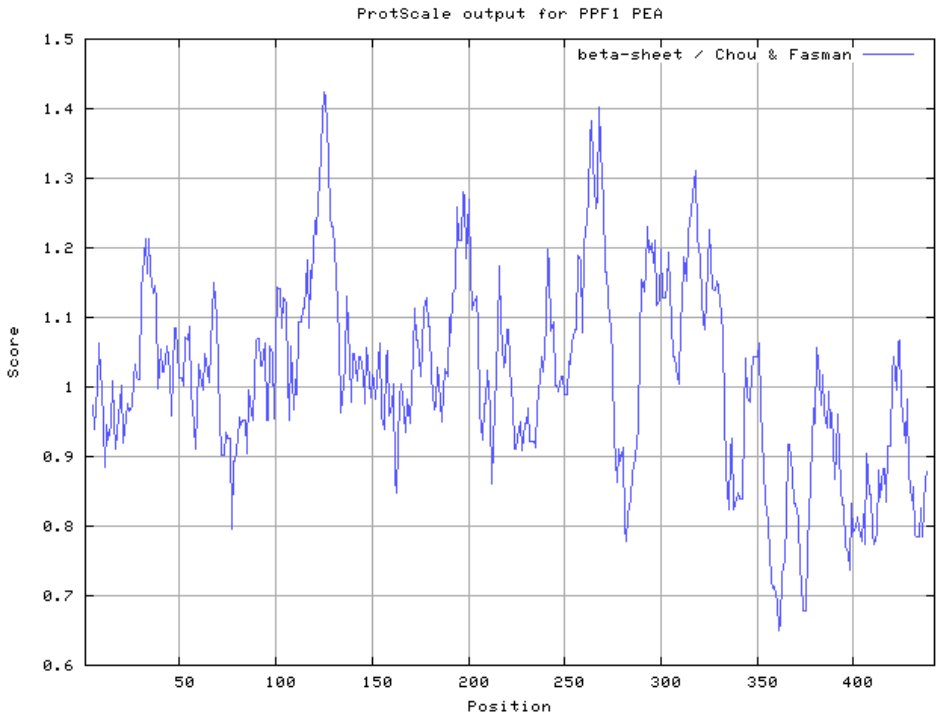
二级结构：

下面采用各种二级结构的分析程序分析 PPF1_PEA 蛋白序列的二级结构，得分越高表明该区域为该二级结构的可能性越高。

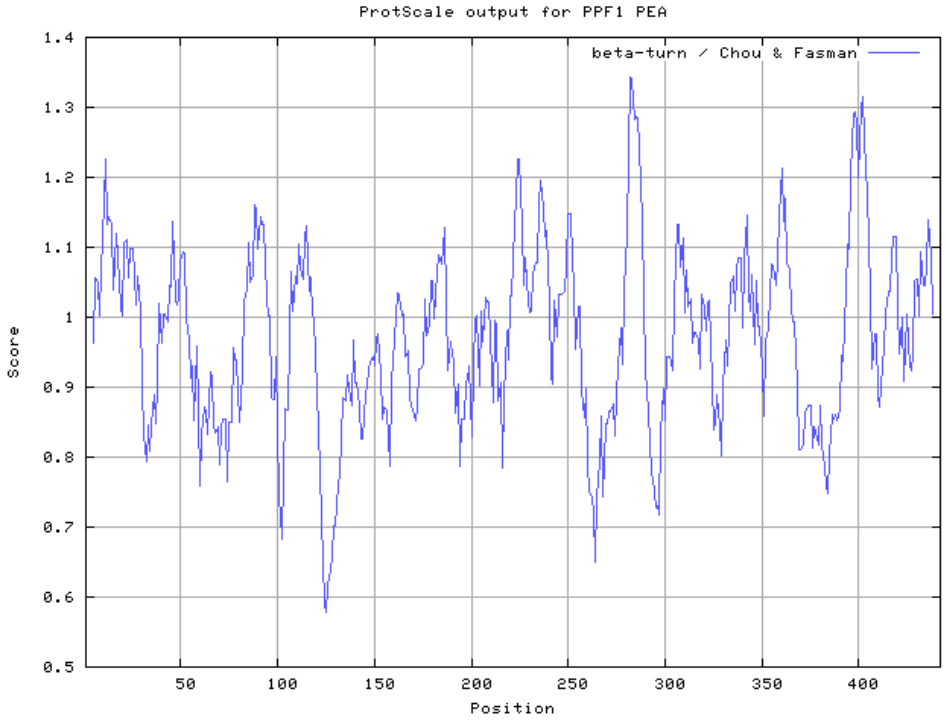
α-Helix（alpha-helix / Deleage & Roux）



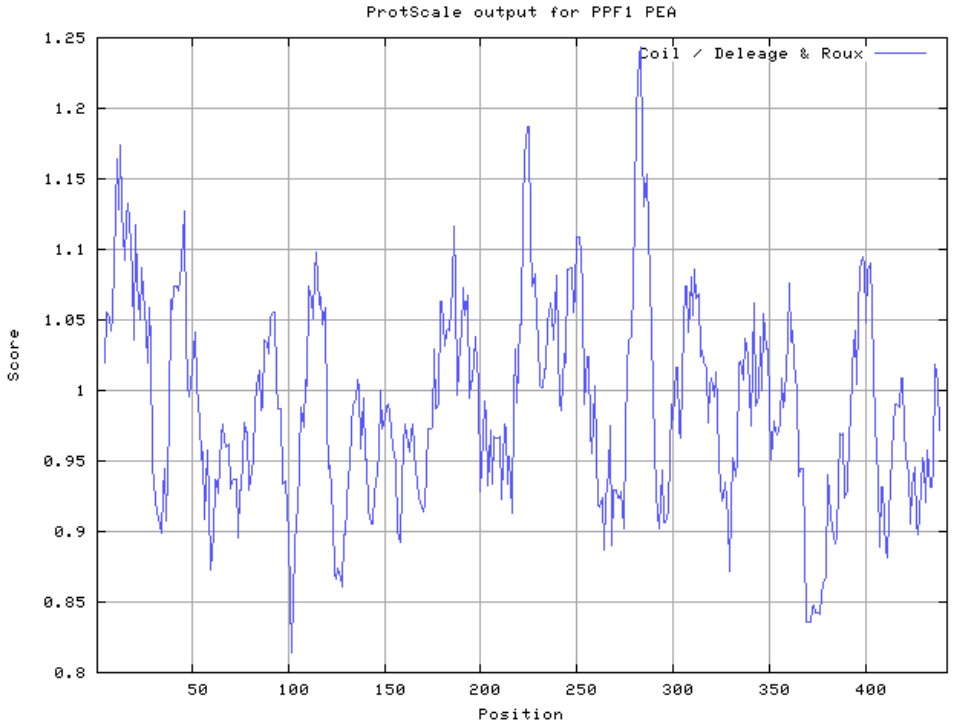
β-sheet（beta-sheet / Chou & Fasman）：



β-turn（beta-turn / Chou & Fasman）：

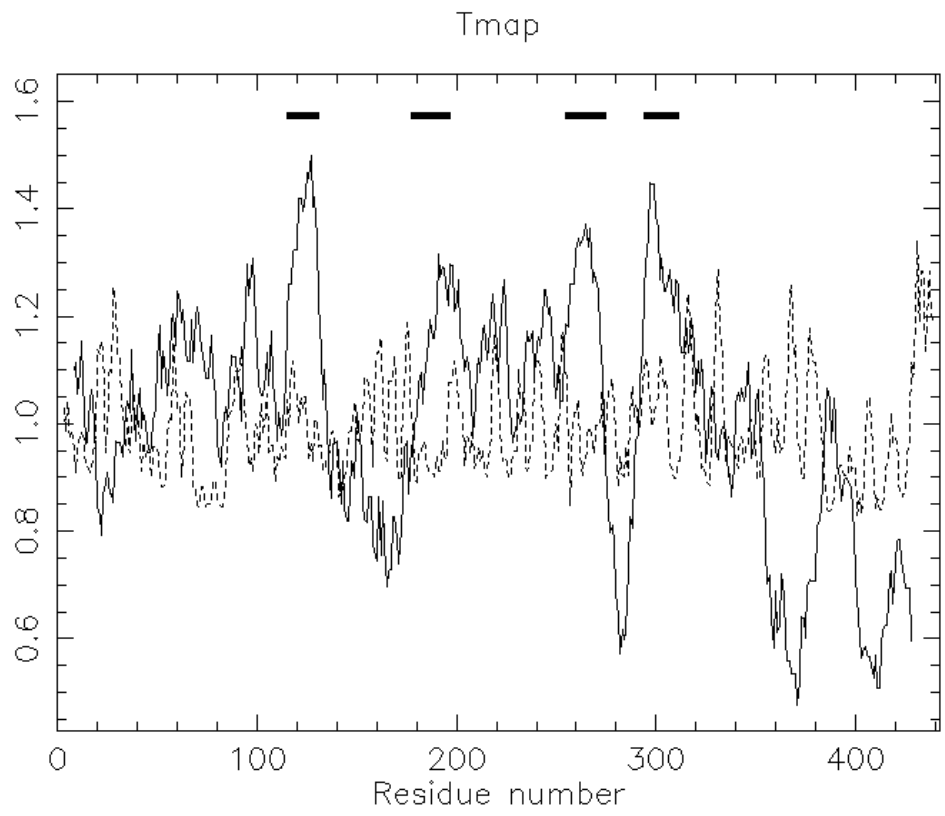


Coil（Coil / Deleage & Roux）：



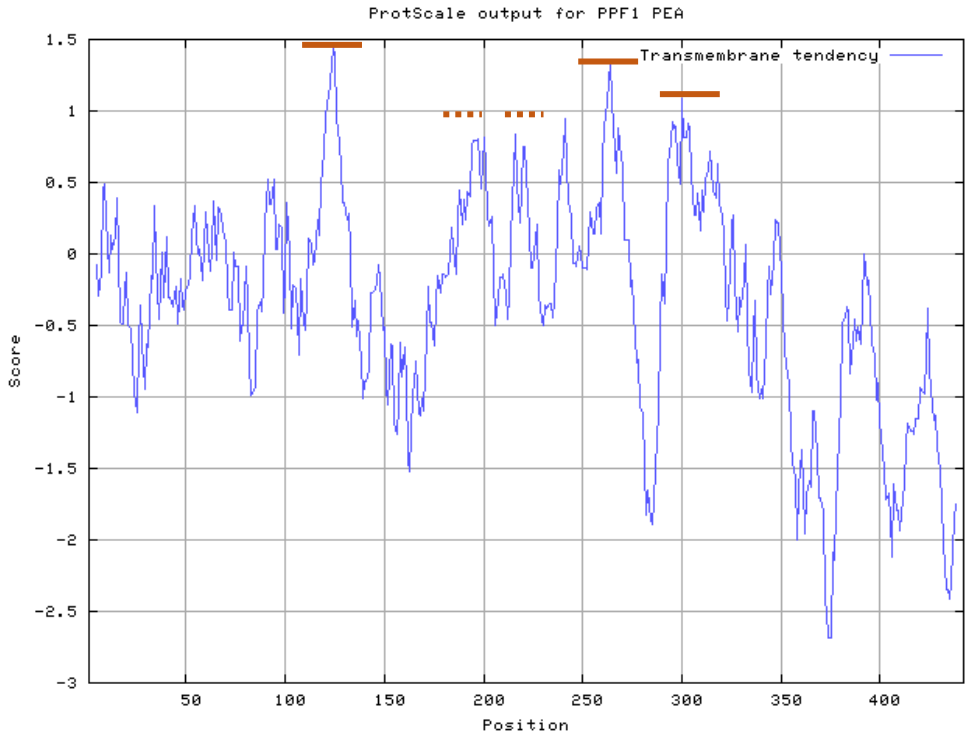
3) 利用 WebLab 中和 ExPASy 网站提供的跨膜螺旋预测程序，预测 PPF-1 蛋白质序列中可能的跨膜螺旋，并比较预测结果的异同。

用 WebLab 中 Tmap 程序对 PPF1_PEA 蛋白序列进行分析，结果如下图所示：



Tmap 预测了四段跨膜螺旋区，如上图所标示。

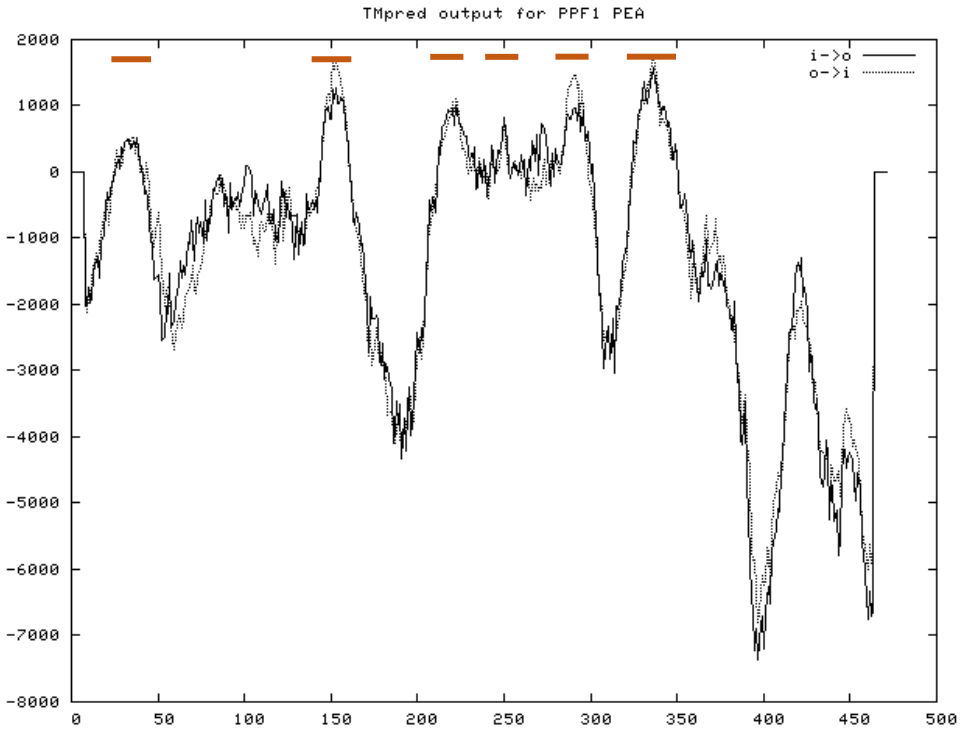
用 ExPASy 中 ProtScale 工具的 Transmembrane tendency 程序对 PPF1_PEA 进行分析，结果如下图所示：



该跨膜趋势图谱与 Tmap 大体一致，但没有明确标示出预测的跨膜螺旋区域。红实线所

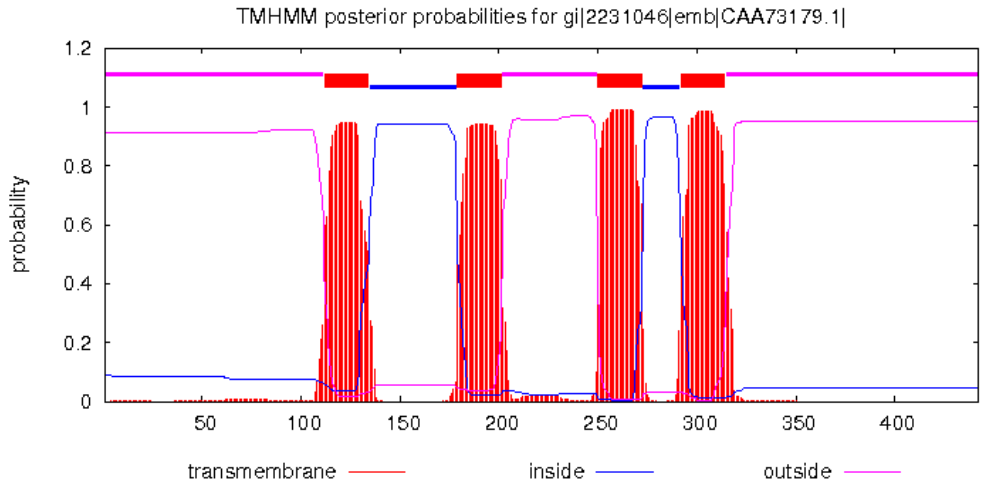
标示（手动）的为较突出的峰，与 Tmap 上相应位置的跨膜螺旋区域基本一致，而其他区域并不突出。

用 ExPASy 上的 TMpred 程序对 PPF1_PEA 进行分析，其预测的跨膜螺旋如下图所示：



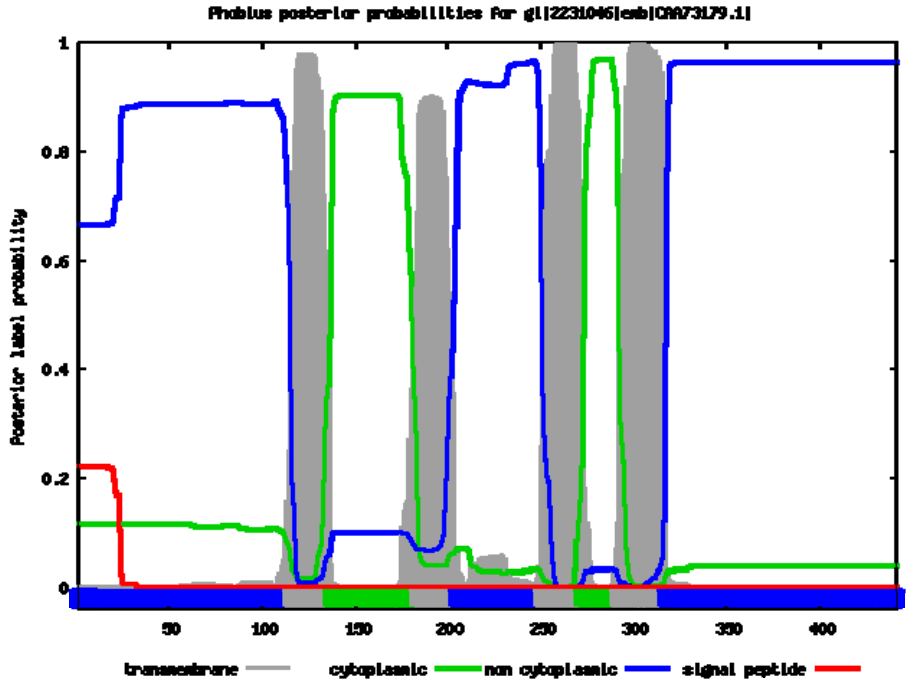
TMpred 预测出 6 段跨膜螺旋，图中红线是根据程序给出的统计结果手动标示的，其中第 2,3,5,6 段跨膜螺旋与 Tmap 预测结果基本一致。

用 TMHMM 对 PPF1_PEA 进行分析，其预测的跨膜螺旋如下图所示：



TMHMM 预测 4 段跨膜螺旋，各段螺旋区域与 Tmap 预测结果基本一致，Tmap 的图谱中第 2,3 段螺旋区域的得分显著低于第 1,4 段，而 TMHMM 预测的 4 段跨膜螺旋的得分相差不大且分值都较高。

用 Phobius 程序对 PPF1_PEA 进行分析，其预测的跨膜螺旋如下图所示：



与 TMHMM 预测结果基本一致。

用 DAS-TMfilter 对 PF1_PEA 进行分析，预测得 4 段跨膜螺旋，如下表所示：

Core	Value	Start	End	E Value
124	4.741	116	132	3.438e-05
194	2.978	189	199	1.733e-02
265	4.548	256	271	6.805e-05
300	3.804	293	312	9.398e-04

用 HMMTOP 对 PF1_PEA 进行分析，预测得 5 段跨膜螺旋：143-162，205-224，237-256，275-294，325-344。如下所示：

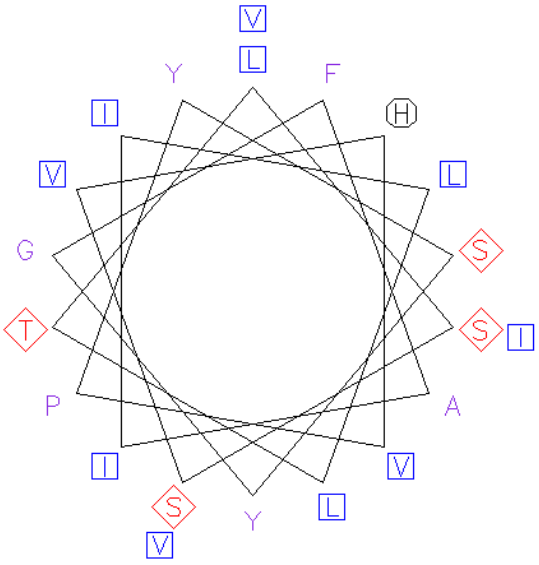
```
seq   GIEMCAAPPF PRTEINPISM SATIVMMAKT LISSPSFLGT PLPSLHRTFS    50
pred  0000000000 0000000000 0000000000 0000000000 0000000000
seq   PNRTRLFTKV QFSFHQLPPI QSVSHSVDLS GIFARAEGLL YTLADATVAA   100
pred  0000000000 0000000000 0000000000 0000000000 0000000000
seq   DAAASTDVAA QKNGGWFGFI SDGMEFVLKV LKDGLSSVHV PYSYGFAIIL   150
pred  0000000000 0000000000 0000000ooo oooooooooo ooHHHHHHHH
seq   LTVIVKAATL PLTKQQVEST LAMQNLQPKI KAIQERYAGN QERIQLETSR   200
pred  HHHHHHHHHH HHiiiiiiii iiiiiiiIII IIIIIIIIIi iiiiiiiiii
seq   LYTQAGVNPL AGCLPTLATI PVWIGLYQAL SNVANEGLLT EGFLWIPSLG   250
pred  iiiiHHHHHH HHHHHHHHHH HHHHoooooo oooooHHHH HHHHHHHHHH
seq   GPTSIAARQS GSGISWLFPF VDGHPLLGWY DTAAYLVLPV LLIVSQYVSM   300
pred  HHHHHHiiii iiiiiiiiii iiiiHHHHHH HHHHHHHHHH HHHHoooooo
seq   EIMKPPQTND PNQKNTLLIF KFLPLMIGYF SLSVPSGLTI YWFTNNVLST   350
pred  oooooooooo oooooooooo ooooHHHHHH HHHHHHHHHH HHHHiiiiii
seq   AQQVWLRKLG GAKPAVNENA GGIITAGQAK RSASKPEKGG ERFRQLKEEE   400
pred  iiiiiiiiiI IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII
seq   KKKKLIKALP VEEVQPLASA SASNDGSDVE NNKEQEVTEE SNTSKVSQEV   450
pred  IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII
seq   QSFSRERRSK RSKRKPVA  468
pred  IIIIIIIIII IIIIIIII
```

总结：各跨膜螺旋预测程序所得结果大同小异，其中 4 段跨膜螺旋在各程序中都可预测得到，可能性较大。UniProt 中 PPF1_PEA 的序列特征注释中显示 109-129,184-204,297-317 为跨膜螺旋区，与 Tmap，TMHMM 和 Phobius 预测得到第 1,2,4 段跨膜螺旋区域大体一致。除了 HMMTOP 程序预测结果偏差较大外，各程序的预测结果基本都涉及这三段区域，但都预测了其他区域。在各程序中 TMHMM 和 Phobius 的显示方式较为直观，DAS-TMfilter 给出了螺旋的具体位置和预测得分，比较清晰明了。

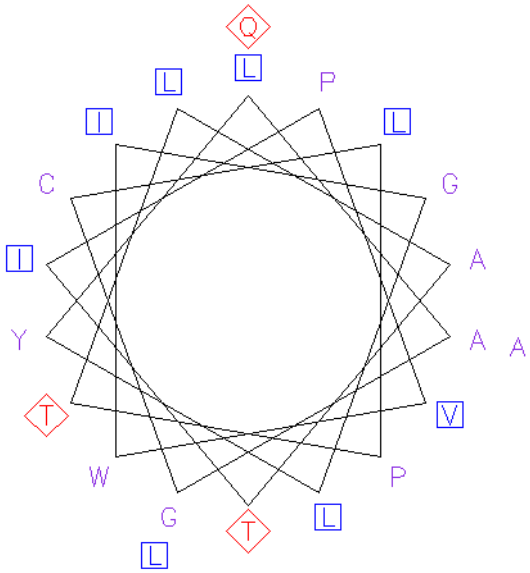
4) 利用 WebLab 中 alpha-螺旋轮显示程序，绘制 PPF-1 蛋白质序列中预测到的跨膜螺旋的螺旋论，说明跨膜区的序列特征 。

UniProt 中 PPF1_PEA (Q9FY06)的序列特征注释中显示 109-129,184-204,297-317 为跨膜螺旋区，此处用 WebLab 中 Pepwheel 程序对此三段跨膜螺旋区进行分析：

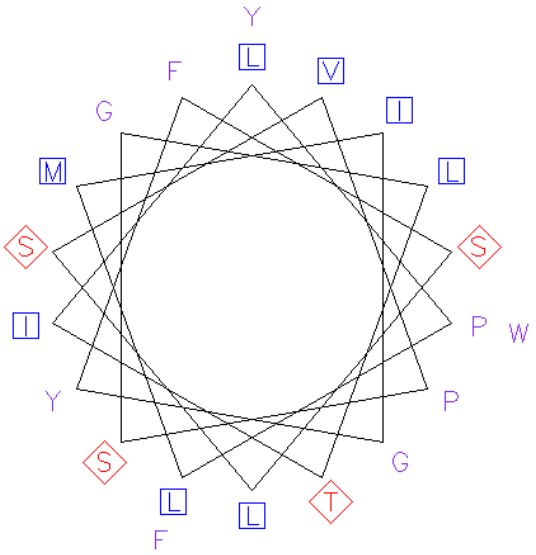
>sp|Q9FY06|109-129
LSSVHVPYSYGFAIILLTVIV



>sp|Q9FY06|184-204
LAGCLPTLATIPVWIGLYQAL



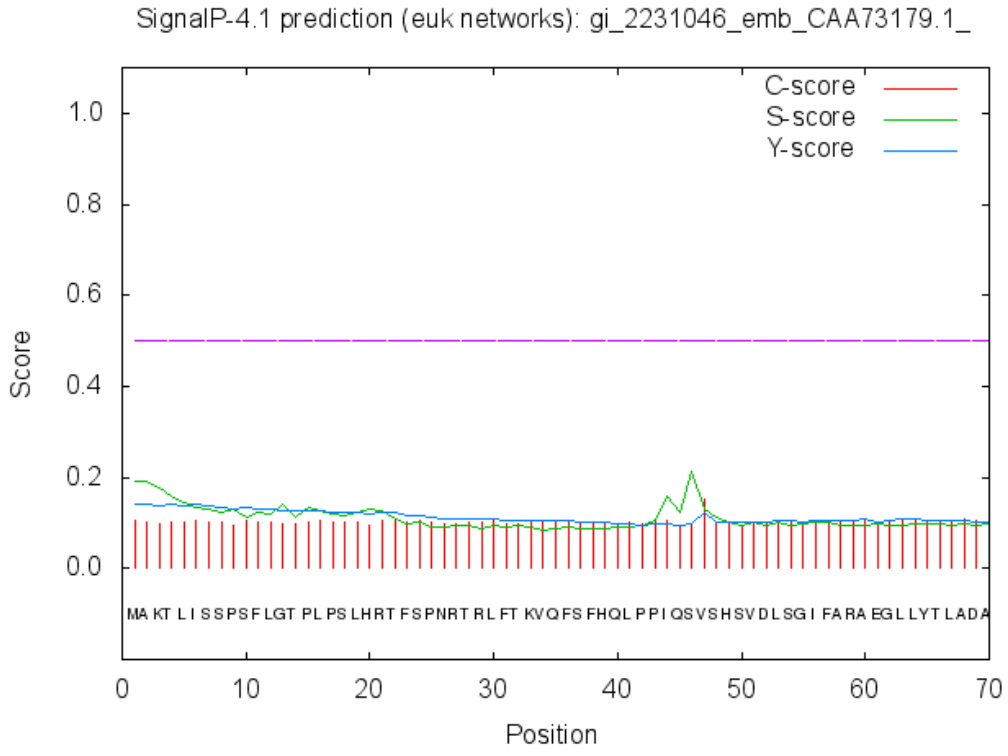
>sp|Q9FY06|297-317
LPLMIGYFSLSVPSGLTIYWF



这三个跨膜螺旋都以疏水性氨基酸为主，尤其是强疏水性的脂肪族氨基酸 Leu, Ile, Val 等（螺旋轮中以蓝色方框标示），以及其他疏水性氨基酸如 Ala, Phe, Trp, Tyr 等；同时也存在少数不带电荷的极性氨基酸如 Ser, Thr, Gln 等（螺旋轮中以红色菱形标示）；带电荷的氨基酸如 His 等很少（螺旋轮中以黑色八边形标示）。

5) 利用 CBS 网站提供的程序，预测 PPF-1 蛋白质的信号肽、亚细胞定位。

用 SignalP 程序分析 PPF1_PEA 蛋白序列，预测的信号肽图谱如下图所示：



统计结果如下所示：

```
# Measure  Position  Value    Cutoff   signal peptide?
  max. C    47       0.152
  max. Y    11       0.131
  max. S    46       0.214
  mean S     1-10    0.149
       D     1-10    0.141   0.450   NO
Name=gi_2231046_emb_CAA73179.1_ SP='NO' D=0.141 D-cutoff=0.450 Networks=SignalP-noTM
```

预测结果表明 PPF-1 存在信号肽段的可能性很小。

用 TargetP 程序分析 PPF1_PEA 蛋白序列的亚细胞定位情况，Organism Group 选择植物，预测结果如下所示：

```
### targetp v1.1 prediction results ###################################
Number of query sequences:  1
Cleavage site predictions included.
Using PLANT networks.

Name                    Len      cTP    mTP     SP  other  Loc  RC  TPlen
-----------------------------------------------------------------------
gi_2231046_emb_CAA73  442    0.844  0.178  0.020  0.039    C    2     29
-----------------------------------------------------------------------
cutoff                         0.000  0.000  0.000  0.000
```

预测结果显示，PPF-1 定位在叶绿体的可能性较高，cTP（chloroplast transit peptide）得分为 0.844，远高于线粒体 mTP（mitochondrial targeting peptide）得分 0.178，RC（Reliability class）值为 2（分 1-5，1 的可靠性最高），表明预测的可靠性较高。

6. 课题相关蛋白质和核酸序列分析

1) 检索 UinProtKB 数据库中和你研究课题相关的蛋白质序列，总结其注释信息、相关文献和数据库交叉链接提供的信息。

研究的蛋白信息如下：

UniProt Entry: Q6ZMU5 (TRI72_HUMAN)
名称：Tripartite motif-containing protein 72（TRIM72）或 Mitsugumin-53（MG53）
蛋白长度 477 AA　　Organism：Homo sapiens (Human)
RefSeq 序列登录号：NP_001008275.2. NM_001008274.3.

涉及对比时，将其与非洲爪蟾的同源基因作对比：

UniProt Entry: Q6PGR9 (TRI72_XENLA)
蛋白长度 477 AA　　Organism：Xenopus laevis (African clawed frog)
RefSeq 序列登录号：NP_001079922.1. NM_001086453.1.

2) 将以上豌豆内膜蛋白 PPF-1 序列分析思路和方法用于和你研究课题相关的蛋白质及其编码基因 mRNA 序列的分析，说明分析结果。

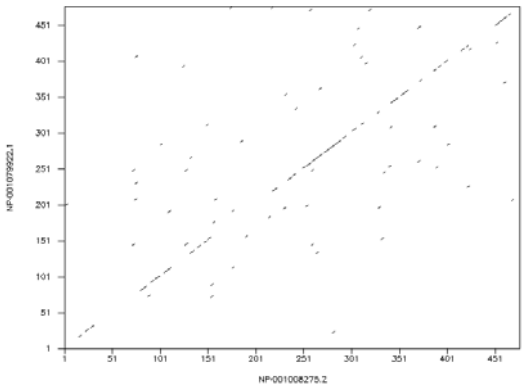
A. 序列相似性分析

1) 利用 WebLab 中的点阵图程序 Dottup、DotMatcher、DotPath 对 TRI72_HUMAN 和 TRI72_XENLA 蛋白质序列及其编码基因的编码区序列进行比对:

Dottup:

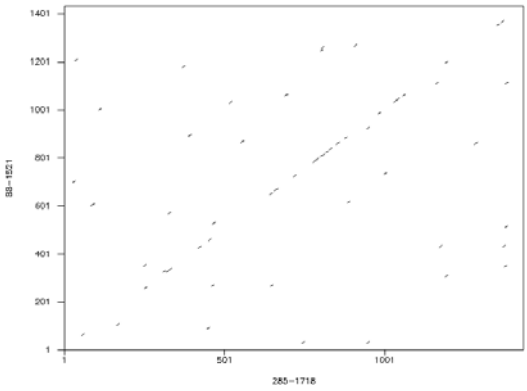
(A) Protein

(Word size = 3)



(B) CDS

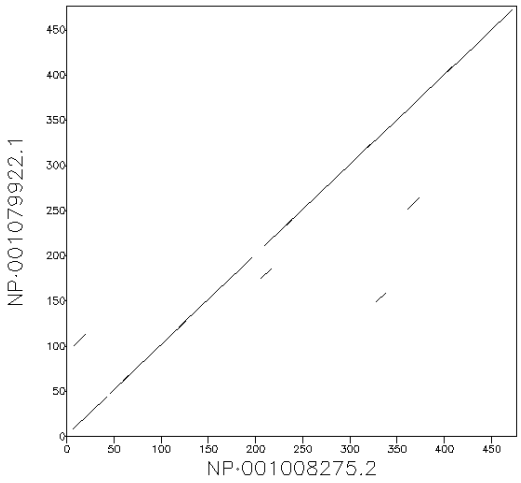
(Word size = 8)



DotMatcher

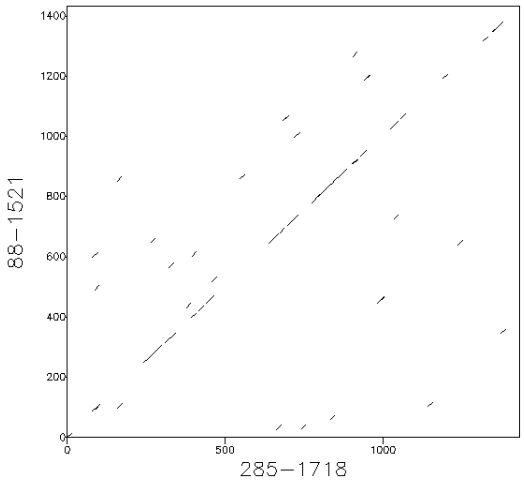
(A) Protein

(EBLOSUM62, Windowsize = 10, Threshold = 23)



(B) CDS

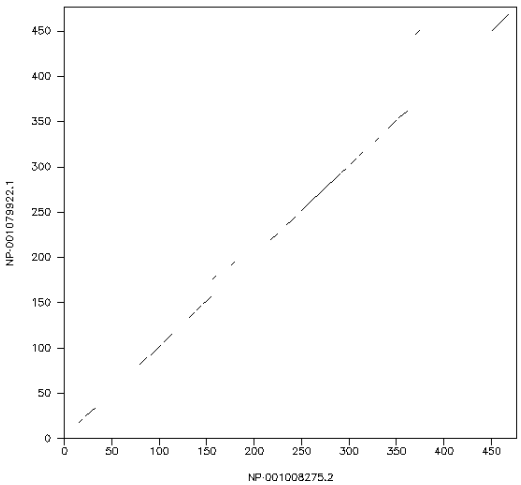
(EDNAFULL, Windowsize = 13, Threshold = 40)



DotPath

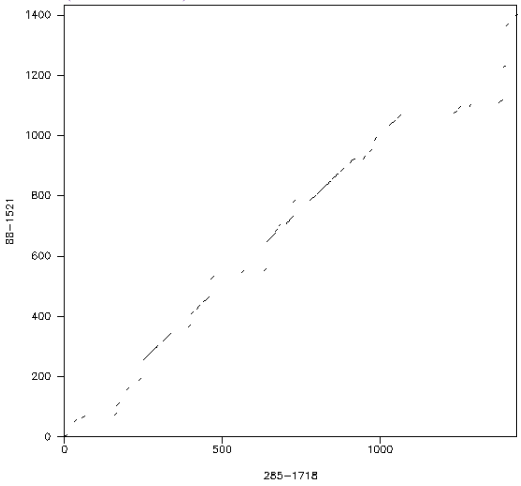
(A) Protein

(Word size = 3)



(B) CDS

(Word size = 5)



用 Dottup 和 DotPath 绘制点阵图时匹配程度不高，有多处中断，用 DotMatcher 则可较好地匹配；用蛋白序列进行点阵图绘制，相对于核酸序列而言匹配程度更高，主要是因为蛋白序列具有较高的特异性和保守性。总体而言 TRI72_HUMAN 和 TRI72_XENLA 的亲缘性不高，序列差异较大。

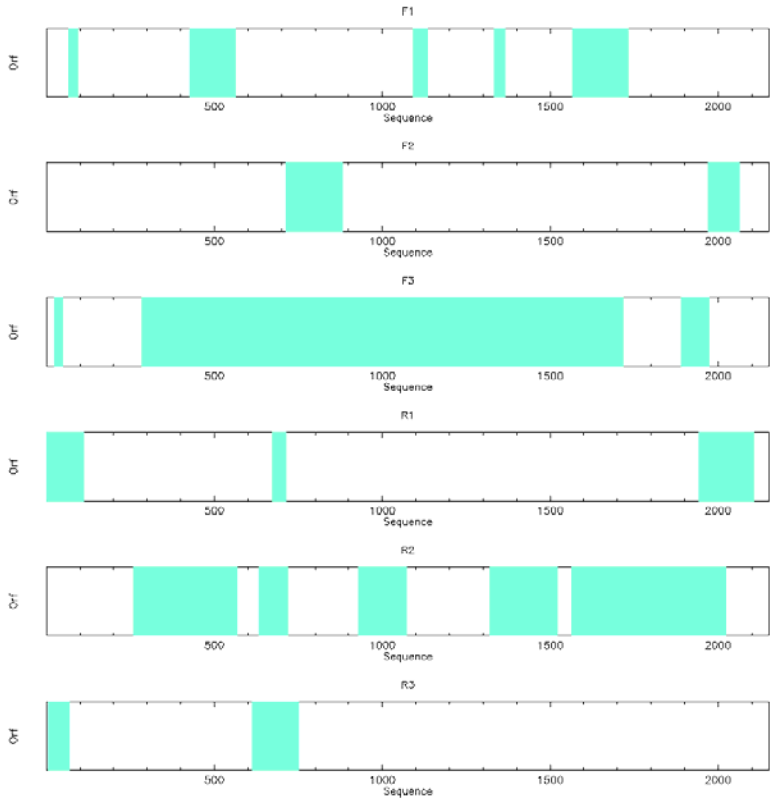
2) 利用 WebLab 中的全局比对程序 Needle 和局部比对程序 Water，对 TRI72_HUMAN 和 TRI72_XENLA 蛋白质序列及其编码基因的编码区序列进行比对，比对结果如下表所示：

Program	Item	Length	Score	Identity	Similarity	Gaps
needle	Protein	479	1415.0	271/479 (56.6%)	351/479 (73.3%)	4/479 (0.8%)
	CDS	1440	4816.5	878/1440 (61.0%)	878/1440 (61.0%)	12/1440 (0.8%)
water	Protein	465	1468.0	271/465 (58.3%)	350/465 (75.3%)	2/465 (0.4%)
	CDS	1440	4786.5	878/1440 (61.0%)	878/1440 (61.0%)	12/1440 (0.8%)

Needle 为全局比对，将全序列进行比对，在序列差异较大时难免出现 gap 或者较多差异位点；Water 为局部比对，选择两序列匹配较好地区域进行比对，相同位点比例增多。当两序列差异较大特别是长度相差较大时用局部比对更加准确。比对结果表明 TRI72_HUMAN 和 TRI72_XENLA 的相似性不是很高，相似氨基酸越 75%；蛋白序列相似性比 CDS 高。

(B) 读码框分析

1） 提取 TRI72_HUMAN 全长 mRNA 序列，用 WebLab 中 PlotORF 程序分析其可能的读码框，结果如下图所示：



由上图可见，F3 中起止密码子间有较长的连续序列，可能为正确的读码框；其它读码框均只能得到较短的序列，可能性较低。

2) 用 WebLab 中 GetORF 程序提取 TRI72_HUMAN 全长 mRNA 序列中编码区核苷酸序列和所编码的氨基酸序列。得到所有可能的翻译序列，其中最长的序列如下：

>NM_001008274.3_18 [261 - 1715] Homo sapiens tripartite motif containing 72, E3 ubiquitin protein ligase (TRIM72), mRNA
DLQPDLPAMSAAPGLLHQELSCPLCLQLFDAPVTAECGHSFCRACLGRVAGEPAADGTVL
CPCCQAPTRPQALSTNLQLARLVEGLAQVPQGHCEEHLDPLSIYCEQDRALVCGVCASLG
SHRGHRLLPAAEAHARLKTQLPQQKLQLQEACMRKEKSVAVLEHQLVEVEETVRQFRGAV
GEQLGKMRVFLAALEGSLDREAERVRGEAGVALRRELGSLNSYLEQLRQMEKVLEEVADK
PQTEFLMKYCLVTSRLQKILAESPPPARLDIQLPIISDDFKFQVWRKMFRALMPALEELT
FDPSSAHPSLVVSSSGRRVECSEQKAPPAGEDPRQFDKAVAVVAHQQLSEGEHYWEVDVG
DKPRWALGVIAAEAPRRGRLHAVPSQGLWLLGLREGKILEAHVEAKEPRALRSPERRPTR
IGLYLSFGDGVLSFYDASDADALVPLFAFHERLPRPVYPFFDVCWHDKGKNAQPLLLVGP
EGAEA

其余序列较短，上述氨基酸序列最可能为真实翻译结果，相应的编码区核苷酸序列为全长 mRNA 序列的 261 - 1715 位。

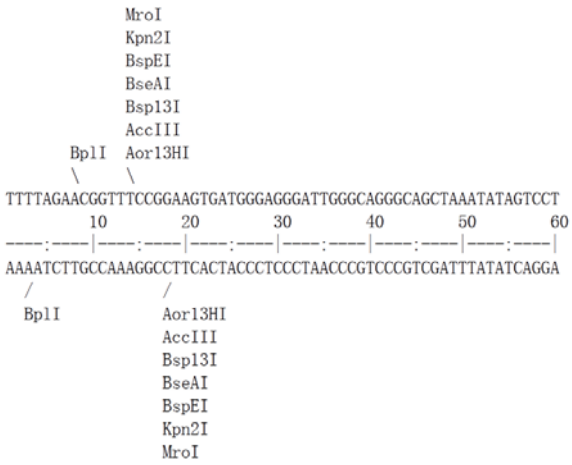
(C) 核苷酸序列分析

1) 利用 WebLab 中密码子统计程序 Cusp，分析 TRI72_HUMAN 和 TRI72_XENLA，两者的密码子使用偏好如下所示：

	TRI72_HUMAN	TRI72_XENLA
Coding GC	68.76%	51.39%
1st letter GC	74.06%	58.16%
2nd letter GC	44.77%	36.82%
3rd letter GC	87.45%	59.21%

可见，TRI72_HUMAN 和 TRI72_XENLA 密码子使用偏好相差较大，在 TRI72_HUMAN 中以 CG 结尾的密码子的使用频率显著高于以 AT 结尾的密码子，而 TRI72_XENLA 中这一现象并不显著，表明在进化过程中两个直系同源基因的密码子使用偏好发生了较显著的改变。

2） 利用 WebLab 中内切酶分析程序 Remap，对 TRI72_HUMAN mRNA 序列进行分析，限制只输出单一酶切位点的酶，得 124 个酶及其酶切位点，其中第 1-60 位序列的酶切位点如下所示：



报告还给出了有多个酶切位点的酶以及对该序列没有酶切位点的酶的列表，根据需求可以更改参数设置得到所需的酶及酶切位点信息，可以根据实验目的及这些酶切位点信息进行实验方案设计。

3) 利用 WebLab 中引物设计程序 Eprimer32，设计 TRI72_HUMAN mRNA 序列的引物，设置产物长度范围为 1500-1800，从输出结果列表中选择如下引物：

	Start	Len	Tm	GC%	Sequence
Forward primer	258	25	59.77	56.00	TAAGATCTCCAACCAGACCTGCCCG
Reverse primer	1750	25	59.94	56.00	AGGAGACCTAAGCTTCAACCCAGGC
Product size	1517				

利用这对引物可以获得 TRI72_HUMAN mRNA 序列的 258-1775 位的 PCR 产物，其中包含了全部编码区信息（285-1718）。

(D) 蛋白质序列分析

1) 利用 WebLab 网站提供的氨基酸组成分析程序 Pepstats，统计 TRI72_HUMAN 蛋白质 20 种不同氨基酸的组成。结果如下：

Residue	Number	Mole%	DayhoffStat
A = Ala	52	10.901	1.268
C = Cys	16	3.354	1.157
D = Asp	20	4.193	0.762
E = Glu	43	9.015	1.502
F = Phe	16	3.354	0.932
G = Gly	32	6.709	0.799
H = His	15	3.145	1.572
I = Ile	8	1.677	0.373
K = Lys	18	3.774	0.572
L = Leu	63	13.208	1.785
M = Met	7	1.468	0.863
N = Asn	3	0.629	0.146
P = Pro	34	7.128	1.371
Q = Gln	28	5.87	1.505
R = Arg	37	7.757	1.583
S = Ser	27	5.66	0.809
T = Thr	10	2.096	0.344
V = Val	36	7.547	1.144
W = Trp	5	1.048	0.806
Y = Tyr	7	1.468	0.432

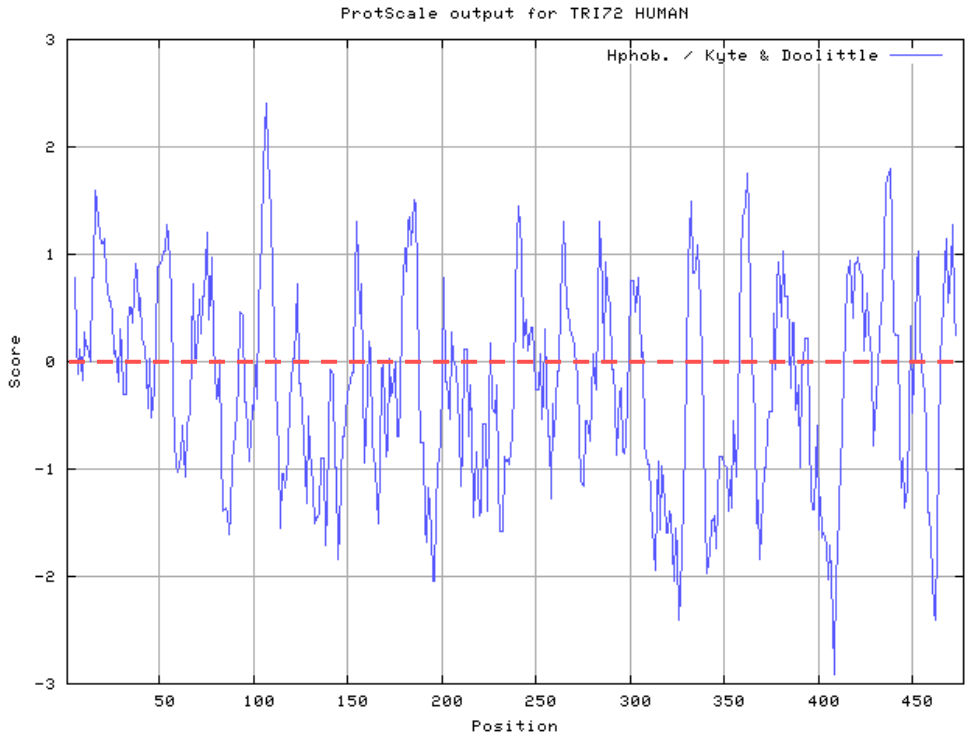
作柱状图如下所示：

报告还给出了氨基酸属性的统计信息，如下所示：

Property	Residues	Number	Mole%
Tiny	(A+C+G+S+T)	137	28.721
Small	(A+C+D+G+N+P+S+T+V)	230	48.218
Aliphatic	(A+I+L+V)	159	33.333
Aromatic	(F+H+W+Y)	43	9.015
Non-polar	(A+C+F+G+I+L+M+P+V+W+Y)	276	57.862
Polar	(D+E+H+K+N+Q+R+S+T)	201	42.138
Charged	(D+E+H+K+R)	133	27.883
Basic	(H+K+R)	70	14.675
Acidic	(D+E)	63	13.208

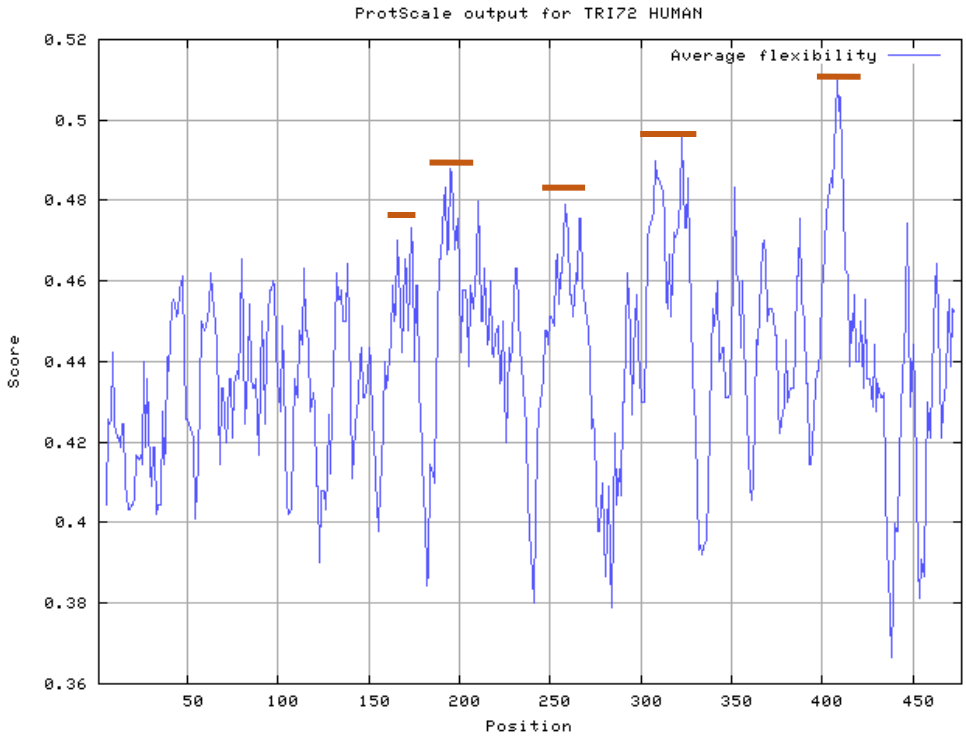
2) 利用 ExPASy 网站提供的一级结构特征分析程序 ProtScale，分析 TRI72_HUMAN 蛋白质不同区域疏水和亲水、柔性和刚性、溶剂可及性、空间位阻、二级结构等序列特征。

疏水和亲水：（Hphob. /Kyte & Doolittle）



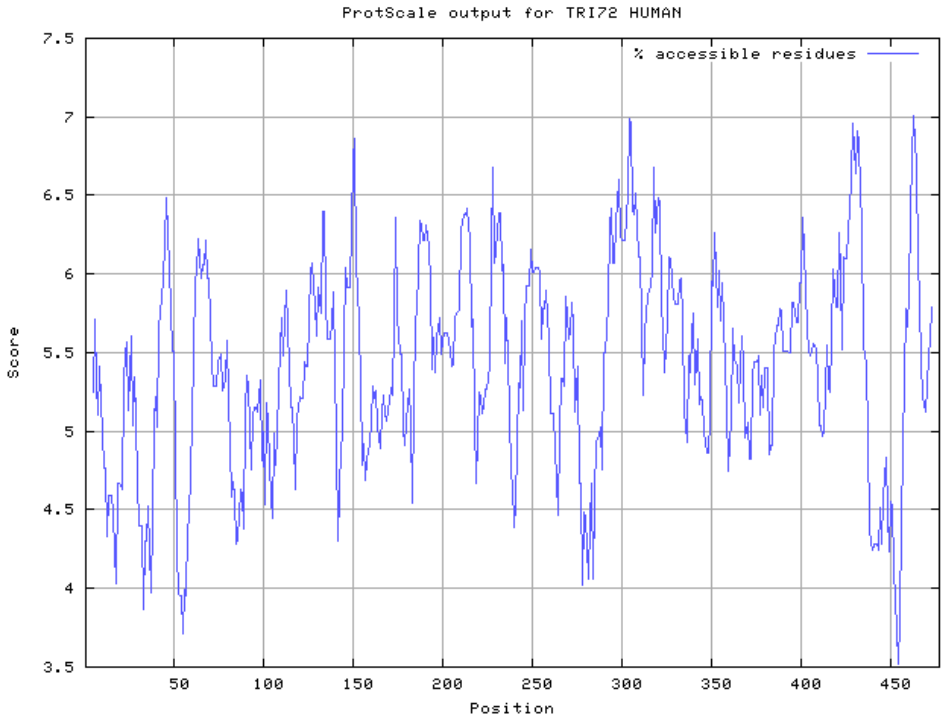
如上图所示为 TRI72_HUMAN 蛋白质的疏水性图谱，数值越高表明该区域多肽链的疏水性越强，TRIM72 蛋白存在数段疏水性较强的区域，但总体而言亲水性较强，表明 TRIM72 可能分散于介质中。

柔性和刚性：（Average flexibility）



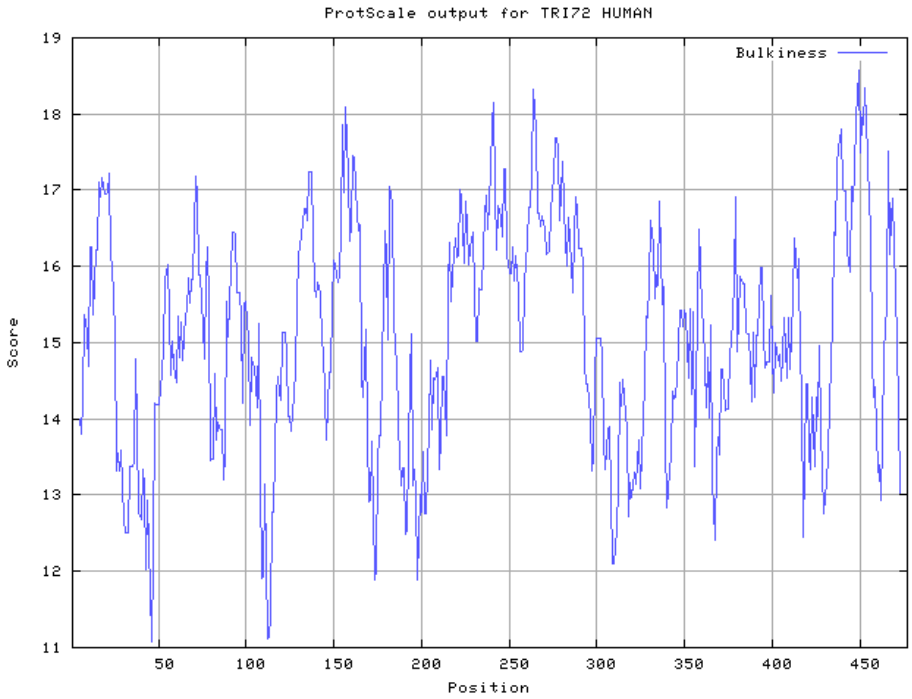
如上图所示为 TRI72_HUMAN 蛋白质序列的柔性图谱，数值越高，表明该区域多肽链的柔性越强，也即可变性越高。TRIM72 蛋白存在多段柔性较强区域，整体刚性不强，可能具有较灵活的三维结构。

溶剂可及性：（% accessible residues）



如上图所示为 TRI72_HUMAN 氨基酸残基溶剂可及性图谱，得分越高表明该区域氨基酸残基的溶剂可及性越高。TRIM72 溶剂可及性氨基酸残基分布总体而言较为均匀，表明 TRIM72 可能稳定分散于水介质中。

空间位阻：（Bulkiness）

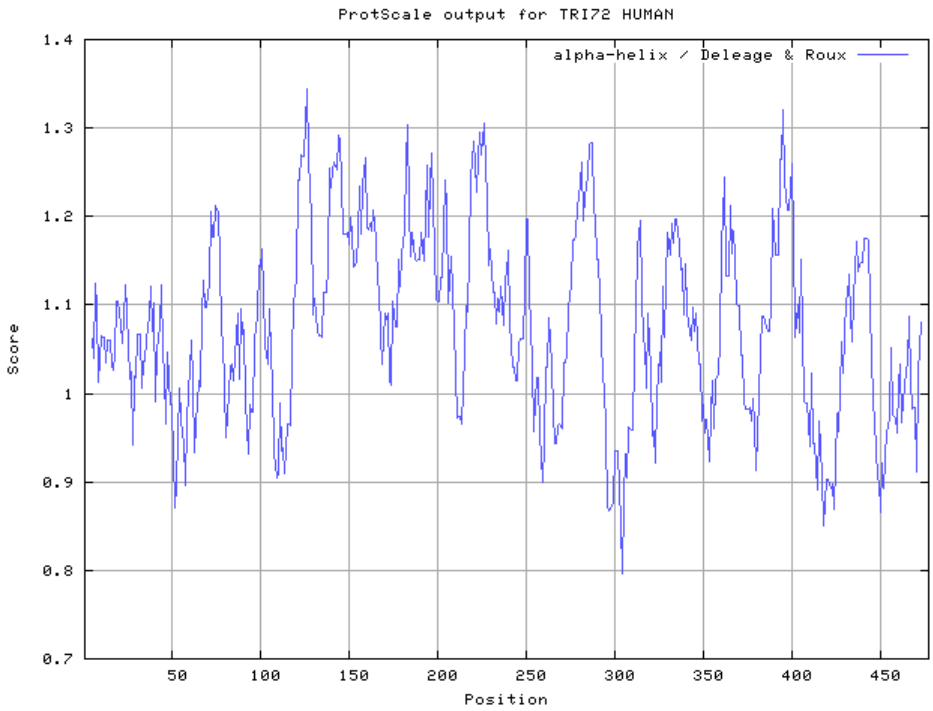


如上图所示为 TRI72_HUMAN 氨基酸残基空间位阻图谱，得分越高表明该区域氨基酸残基的空间位阻越大，空间位阻较大的区域在折叠时不易形成 α 螺旋，而倾向于形成 β 折叠。

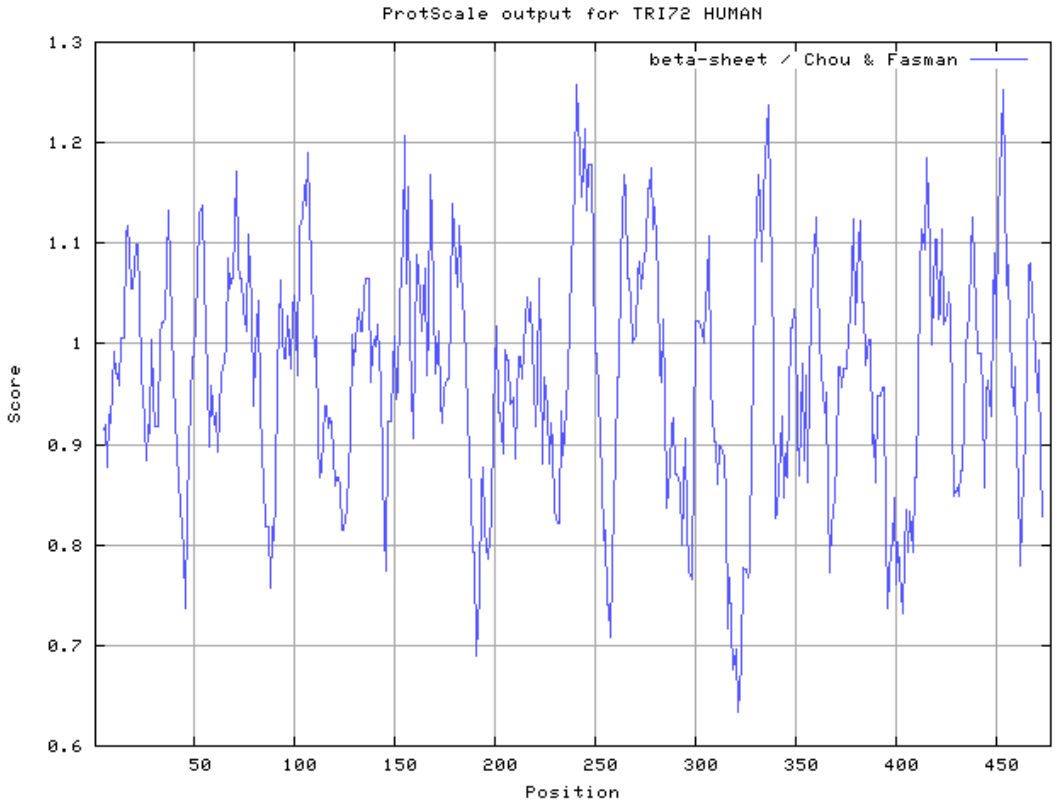
二级结构：

下面采用各种二级结构的分析程序分析 TRI72_HUMAN 蛋白序列的二级结构，得分越高表明该区域为该二级结构的可能性越高。

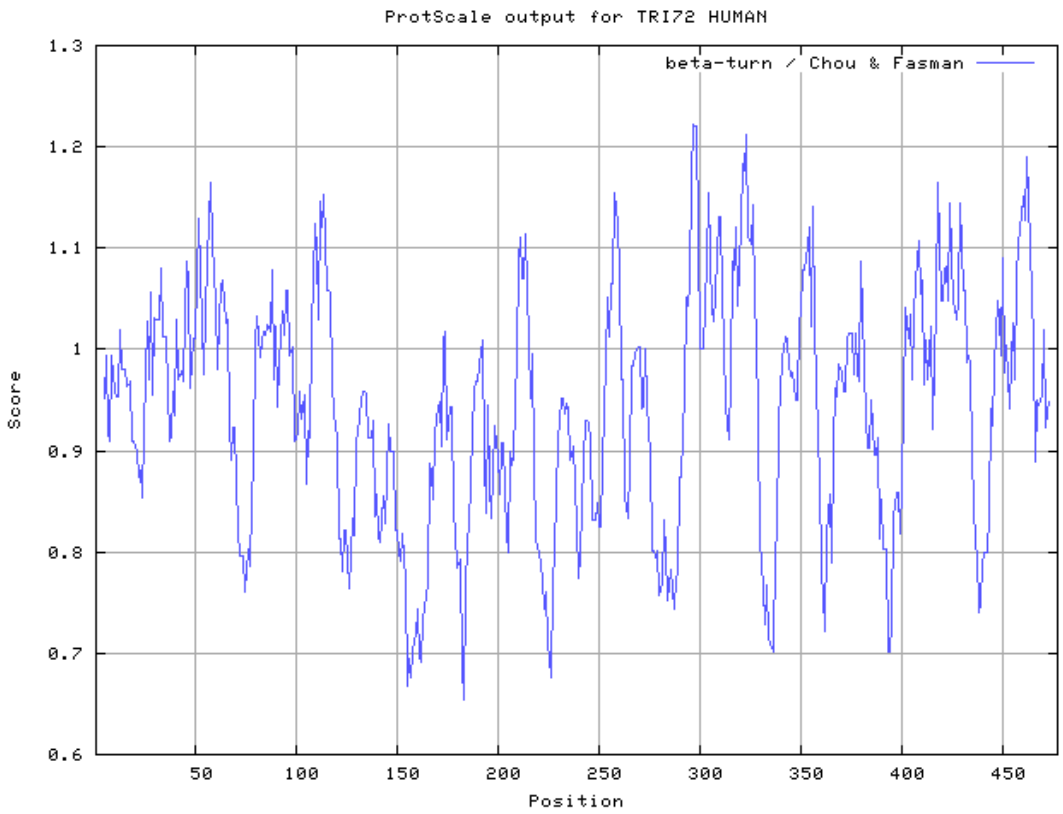
α-Helix（alpha-helix / Deleage & Roux）



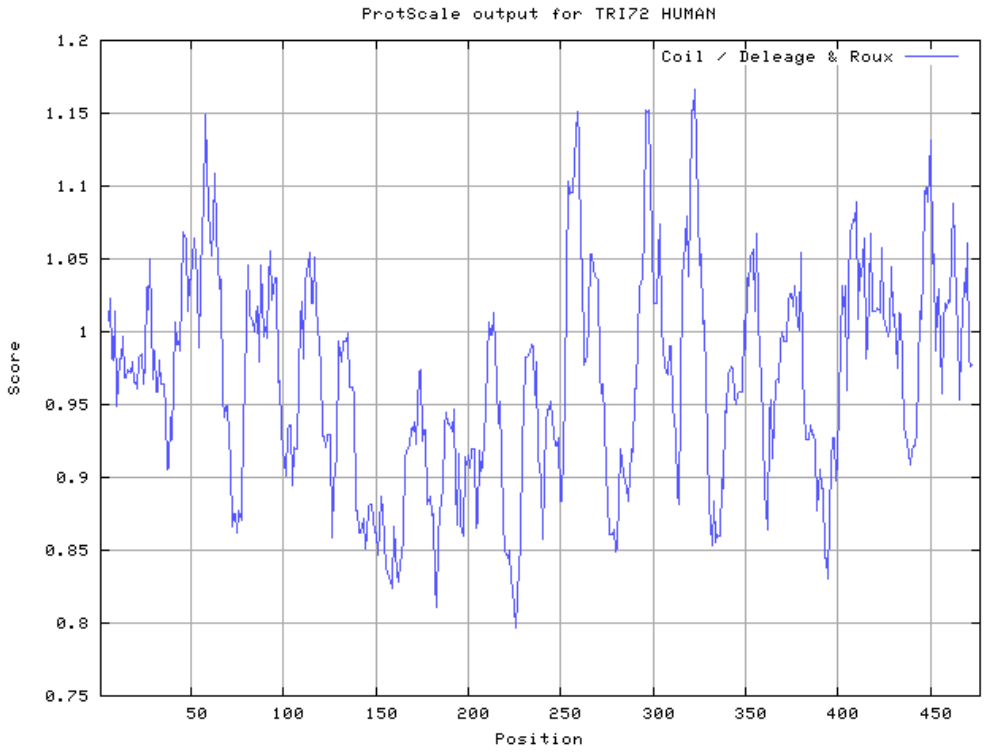
β-sheet（beta-sheet / Chou & Fasman）：



β-turn（beta-turn / Chou & Fasman）：



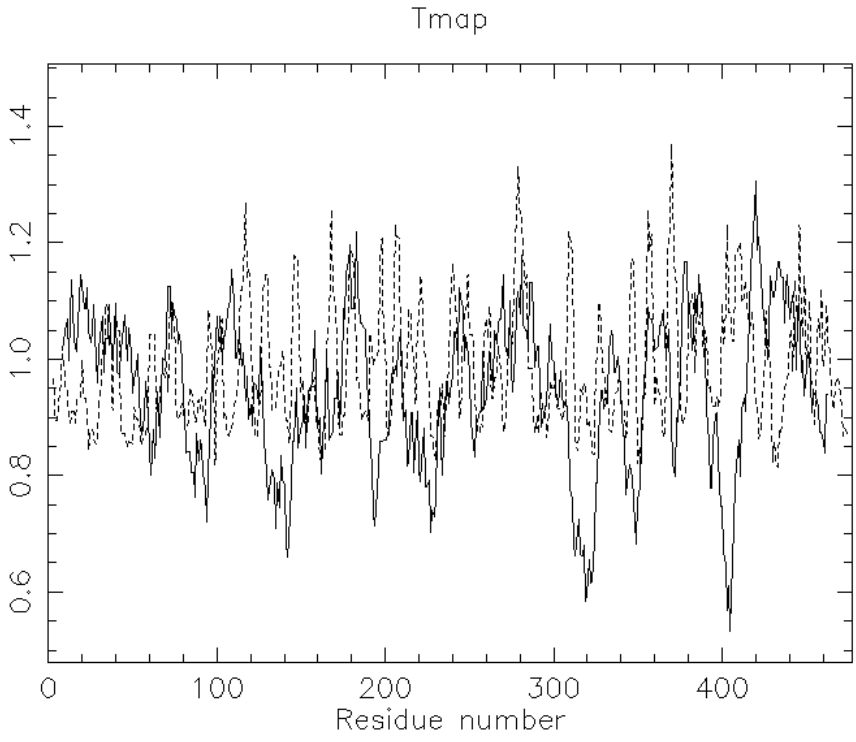
Coil（Coil / Deleage & Roux）：



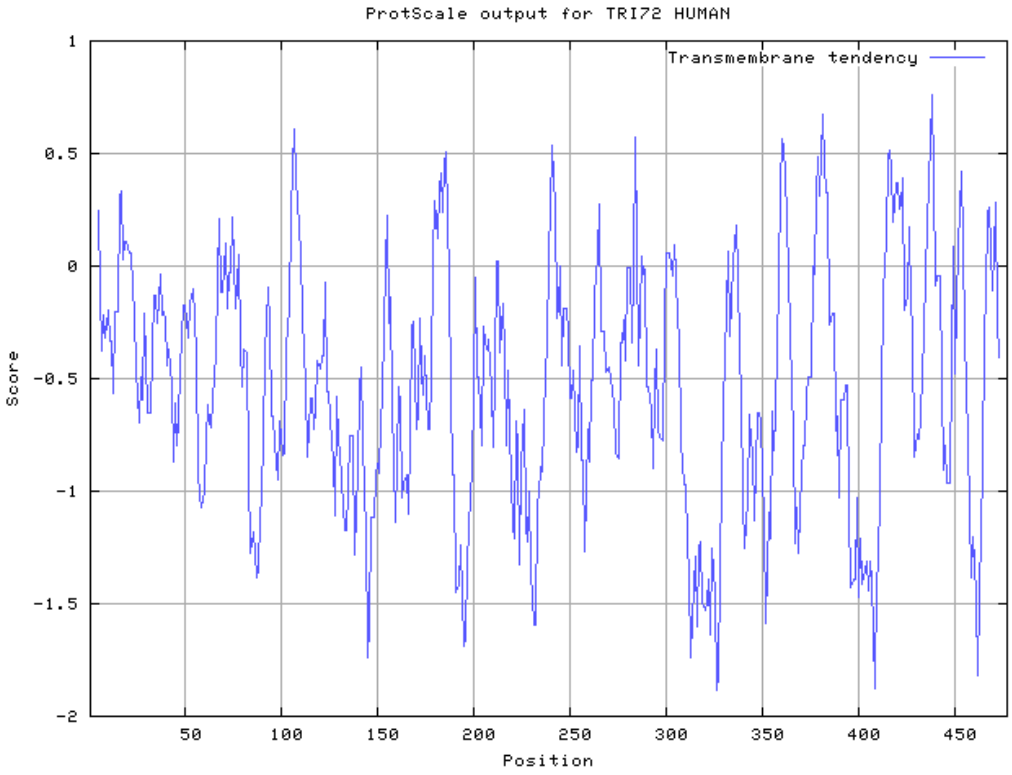
通过上面各种二级结构分析，可大致推断 TRIM72 的结构为多段 α 螺旋和 β 折叠交替。

3) 利用 WebLab 中和 ExPASy 网站提供的跨膜螺旋预测程序，预测 TRI72_HUMAN 蛋白质序列中可能的跨膜螺旋。

用 WebLab 中 Tmap 程序对 TRI72_HUMAN 蛋白序列进行分析，预测结果表明 TRIM72 没有跨膜螺旋区域（没有跨膜螺旋标示），如下图所示：

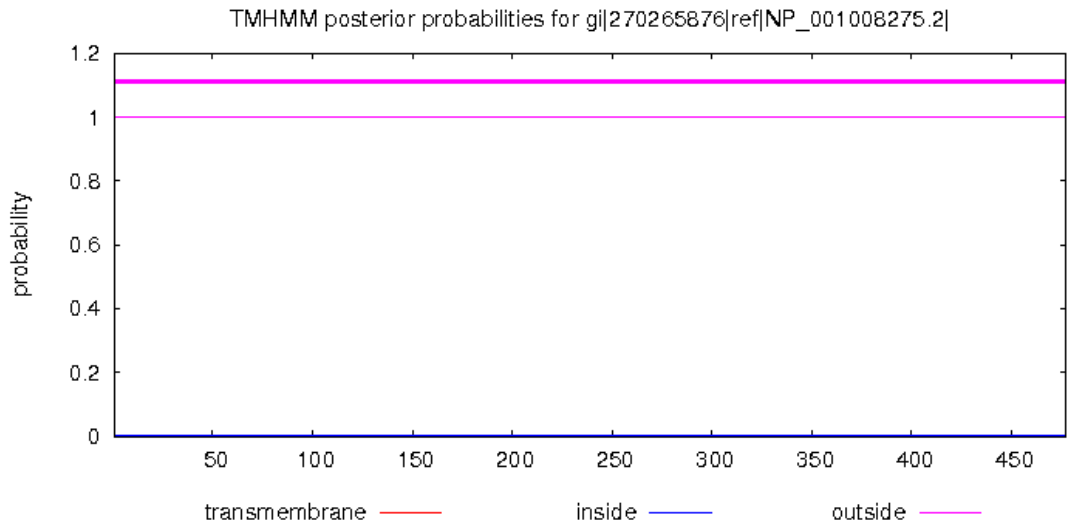


用 ExPASy 中 ProtScale 工具的 Transmembrane tendency 程序对 TRI72_HUMAN 进行分析，结果如下图所示：



该跨膜趋势图谱与 Tmap 走向大体一致，得分总体较低，基本都在 0.5 以下，表明 TRIM72 存在跨膜螺旋区域的可能性较低。

用 TMHMM 对 TRI72_HUMAN 进行分析，其预测结果表明 TRIM72 不存在跨膜螺旋区域，如下图所示：



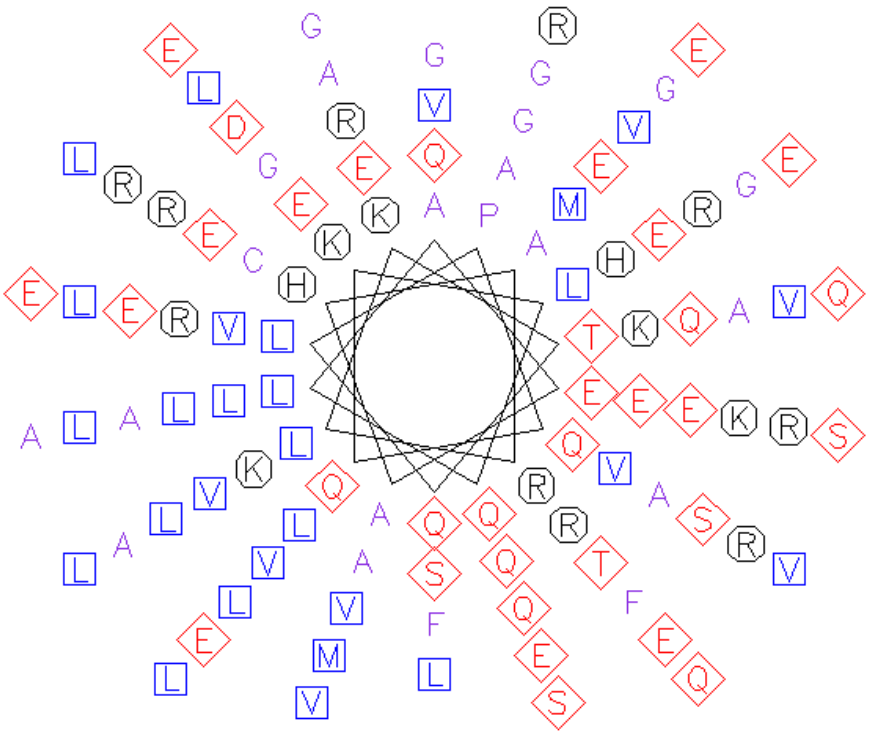
各跨膜螺旋预测程序分析结果都表明 TRIM72 存在跨膜螺旋区域的可能性较小。

4) 利用 WebLab 中 alpha-螺旋轮显示程序，绘制 TRI72_HUMAN 蛋白质序列中预测到的螺旋的螺旋轮。

用 Phyre2 预测 TRI72_HUMAN 蛋白质的结构，二级结构预测结果表明 TRI72_HUMAN

蛋白质存在多段 α 螺旋区，其中[122-230]为一段长螺旋，此处用 WebLab 中 Pepwheel 程序对此段 α 螺旋进行分析：

>sp|Q6ZMU5|122-230
AEAHARLK TQLPQQKLQL QEACMRKEKS VAVLEHQLVE VEETVRQFRG AVGEQLGKMR VFLAALEGSL DREAERVRGE AGVALRRELG SLNSYLEQLR QMEKVLEEVA



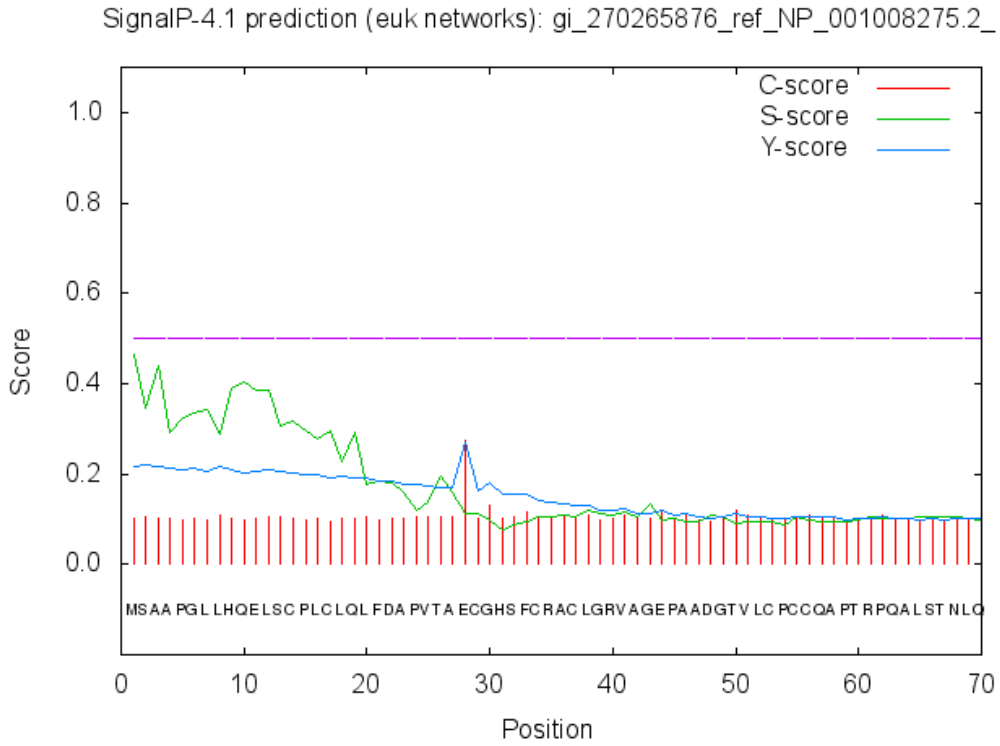
如上图螺旋轮所示，左侧以疏水性氨基酸为主，尤其是强疏水性的脂肪族氨基酸 Leu, Ile, Val 等（以蓝色方框标示），以及其他疏水性氨基酸如 Ala, Gly 等；同时也存在部分极性氨基酸如 Glu, Asp, Gln 等（以红色菱形标示）和带正电荷的氨基酸如 Arg, Lys, His 等（以黑色八边形标示），总体而言左侧以疏水性为主。右侧则明显以亲水性氨基酸为主，大部分为极性氨基酸如 Glu, Gln, Ser, Thr 等（以红色菱形标示）也有较多的带正电荷氨基酸如 Arg, Lys, His 等（以黑色八边形标示），总体而言螺旋轮右侧呈现出显著的亲水性。可见该段螺旋具有一定的两亲性，其亲水侧有利于其在水性介质中保持稳定的结构，疏水侧可能位于蛋白三维结构内侧，也有可能便利于与其他蛋白质结合等。

5) 利用 CBS 网站提供的程序，预测 TRI72_HUMAN 白质的信号肽、亚细胞定位。

用 SignalP 程序分析 TRI72_HUMAN 蛋白序列，预测结果如下：

```
# Measure  Position  Value    Cutoff   signal peptide?
  max. C    28       0.274
  max. Y    28       0.270
  max. S     1       0.464
  mean S     1-27    0.286
       D     1-27    0.279   0.450   NO
Name=gi_270265876_ref_NP_001008275.2_   SP='NO' D=0.279 D-cutoff=0.450 Networks=SignalP-noTM
```

预测结果表明 TRI72_HUMAN 蛋白存在信号肽的可能性较低。相应的信号肽预测图谱如下图所示：



用 TargetP 程序分析 TRI72_HUMAN 蛋白序列的亚细胞定位情况，Organism Group 选择动物，预测结果如下所示：

```
### targetp v1.1 prediction results ###################################
Number of query sequences:  1
Cleavage site predictions included.
Using NON-PLANT networks.

Name                    Len             mTP     SP   other   Loc   RC   TPlen
--------------------------------------------------------------------------------
gi_270265876_ref_NP_   477           0.033  0.654  0.477    S     5      27
--------------------------------------------------------------------------------
cutoff                               0.000  0.000  0.000
```

预测结果显示，TRIM72 为分泌性蛋白的可能性较高，SP（Secretory pathway）得分为 0.654，定位在线粒体 mTP（mitochondrial targeting peptide）的可能性不大，得分 0.033，也有可能定位在其他亚细胞结构中，other 项的得分为 0.477。预测的结论是 TRIM72 为分泌性蛋白，可能存在于胞质等水性介质中，RC（Reliability class）值为 5（分 1-5，1 的可靠性最高），表明预测的可靠性不高。