中国农业科学院研究生院实用生物信息学汇报

# 绵羊MHC基因的生物信息学分析

小　组：G14
报告人：朱才业
组　员：李西良
郭　义
刘志英

2013年12月2日

# 报告提纲

Contents

# 研究背景

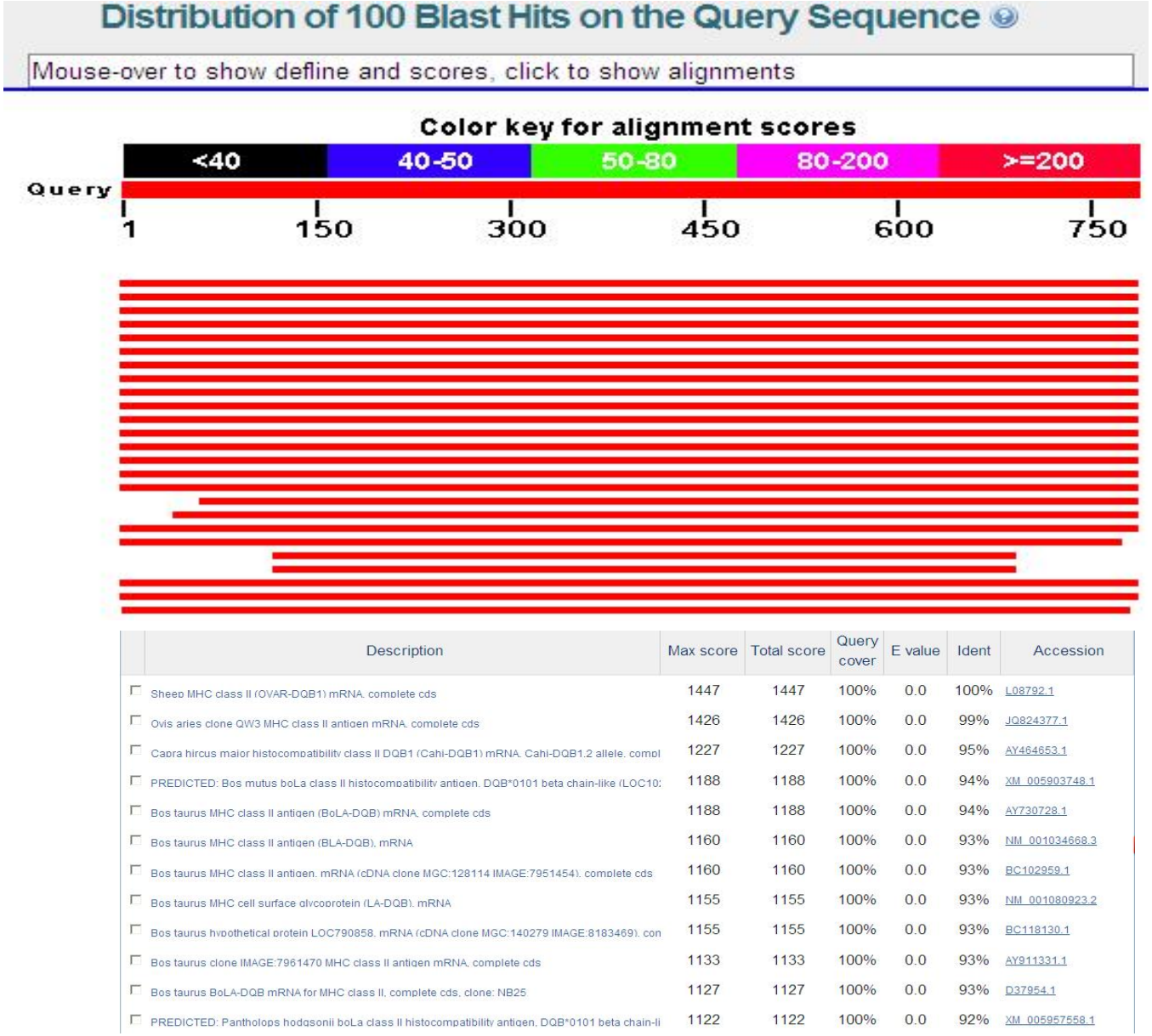
- 概念：主要组织相容性复合体(major histocompatibility complex，MHC)是由一群紧密连锁的基因群组成，定位于动物或人某对染色体的特定区域，呈高度多态性。

- 其编码的分子表达于不同细胞表面，参与抗原递呈，制约细胞间相互识别及诱导免疫应答。

# 研究意义

- 对绵羊的MHC进行生物信息学分析，找出其核酸序列和蛋白质序列，对其进行进化树分析、疏水性、三级结构等进行生物学特征进行分析，为下一步进化分析、育种做基础。

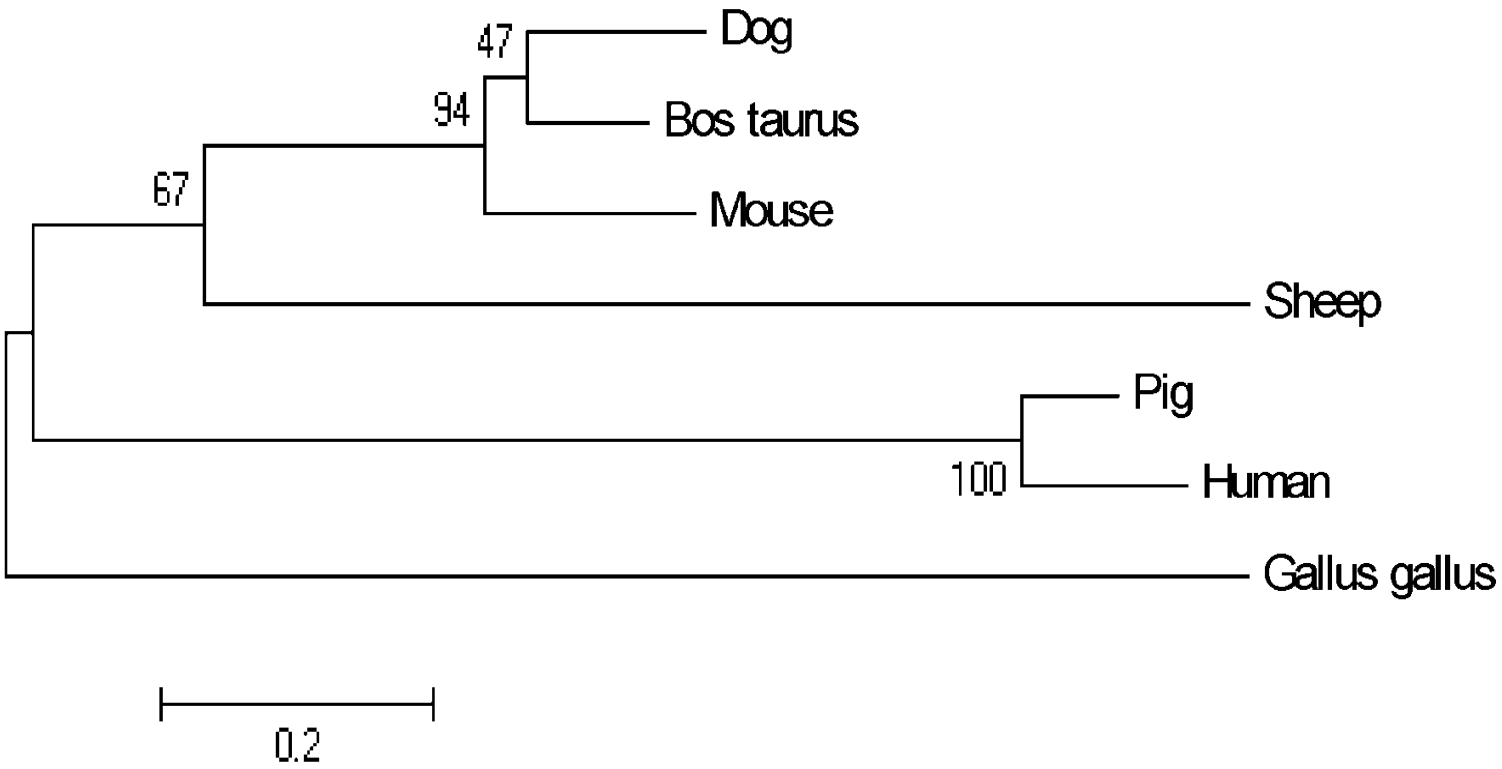
# 序列比对与系统发育分析



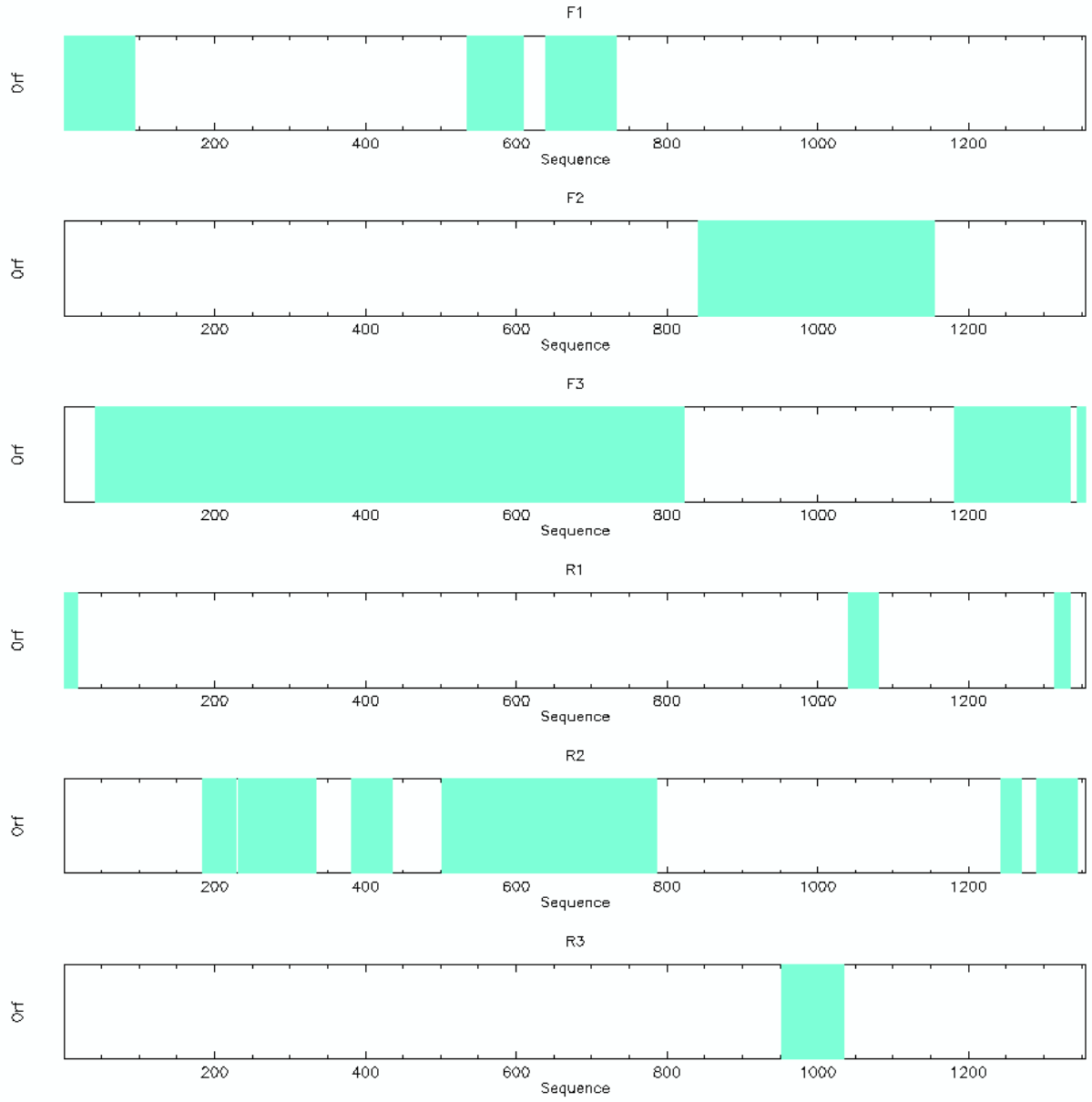

| Description | Max score | Total score | Query cover | E value | Ident | Accession |
|---|---|---|---|---|---|---|
| Sheep MHC class II (OVAR-DQB1) mRNA, complete cds | 1447 | 1447 | 100% | 0.0 | 100% | L08792.1 |
| Ovis aries clone QW3 MHC class II antigen mRNA, complete cds | 1426 | 1426 | 100% | 0.0 | 99% | JQ824377.1 |
| Capra hircus major histocompatibility class II DQB1 (Cahi-DQB1) mRNA, Cahi-DQB1.2 allele, compl | 1227 | 1227 | 100% | 0.0 | 95% | AY464653.1 |
| PREDICTED: Bos mutus bola class II histocompatibility antigen, DQB*0101 beta chain-like (LOC10 | 1188 | 1188 | 100% | 0.0 | 94% | XM_005903748.1 |
| Bos taurus MHC class II antigen (BoLA-DQB) mRNA, complete cds | 1188 | 1188 | 100% | 0.0 | 94% | AY730728.1 |
| Bos taurus MHC class II antigen (BLA-DQB), mRNA | 1160 | 1160 | 100% | 0.0 | 93% | NM_001034668.3 |
| Bos taurus MHC class II antigen, mRNA (cDNA clone MGC:128114 IMAGE:7951454), complete cds | 1160 | 1160 | 100% | 0.0 | 93% | BC102959.1 |
| Bos taurus MHC cell surface glycoprotein (LA-DQB), mRNA | 1155 | 1155 | 100% | 0.0 | 93% | NM_001080923.2 |
| Bos taurus hypothetical protein LOC790858, mRNA (cDNA clone MGC:140279 IMAGE:8183469), con | 1155 | 1155 | 100% | 0.0 | 93% | BC118130.1 |
| Bos taurus clone IMAGE:7961470 MHC class II antigen mRNA, complete cds | 1133 | 1133 | 100% | 0.0 | 93% | AY911331.1 |
| Bos taurus BoLA-DQB mRNA for MHC class II, complete cds, clone: NB25 | 1127 | 1127 | 100% | 0.0 | 93% | D37954.1 |
| PREDICTED: Pantholops hodgsonii bola class II histocompatibility antigen, DQB*0101 beta chain-li | 1122 | 1122 | 100% | 0.0 | 92% | XM_005957558.1 |

- 比对结果相似性较高的有多个，该基因是在大多数生物中都普遍存在的。

# 序列比对与系统发育分析



# 基因mRNA全长序列分析




- **PlotORF从6个frame着手寻找mRNA的开放阅读框。F3是最有可能的读码框，可看出正向从第一个碱基（F3）开始读起才能在42-824左右读到最完整的可连续编码蛋白质序列的读码框。**

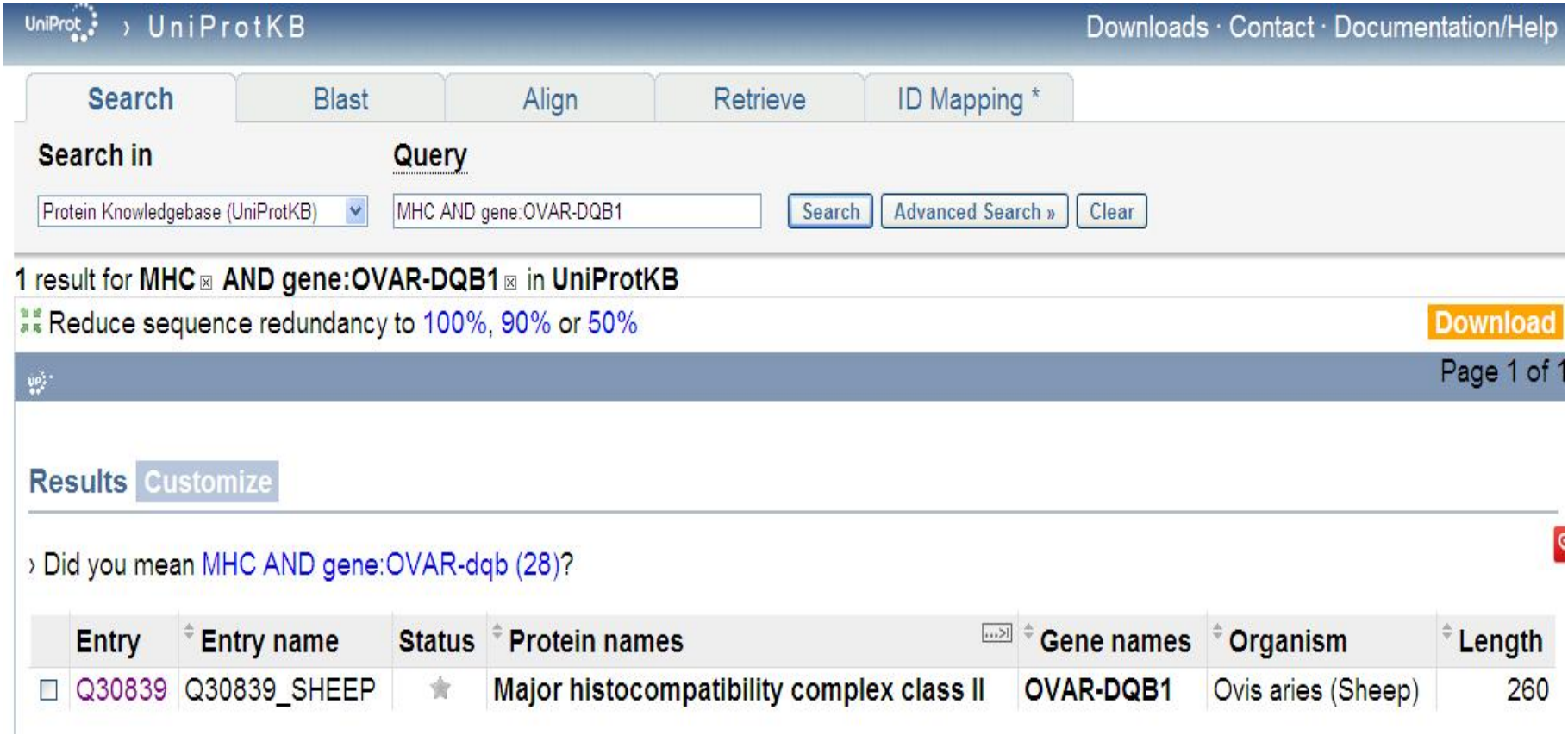
# 基因mRNA全长序列分析

```
---------|---------|---------|---------|---------|
1 atgggtgttgactaccattacttcttgctttgttctctattatgtctggg 50

---------|---------|---------|---------|---------|
51 atggtggctctgcggatccccagaggcctttggacggcagctgtgatggt 100

---------|---------|---------|---------|---------|

---------|---------|---------|---------|---------|
801 agccagaaggggcttgtgcgctgactcccgaggatactttgatggagatt 850

---------|---------|---------|---------|---------|
851 ggacttcgctcttctgtaatagctgcgtgtcttgccagaattcccagctg 900
```

ShowORF用特殊的格式陈列全长mRNA核酸序列和翻译的蛋白质序列，从42bp到824bp之间的开放阅读框最长，最可信。

# 蛋白质序列分析

在UniProt中搜索结果如下：



# 氨基酸组成分析

PEPSTATS of Q30839_SHEEP from 1 to 260

Molecular weight = 29768.95 Residues = 260
Average Residue Weight = 114.496 Charge = 8.0
Isoelectric Point = 9.1336
A280 Molar Extinction Coefficient = 45660
A280 Extinction Coefficient 1mg/ml = 1.53
Improbability of expression in inclusion bodies = 0.753

# 氨基酸组成分析

| Residue | Number | Mole% | DayhoffStat |
|---|---|---|---|
| A = Ala | 14 | 5.385 | 0.626 |
| B = Asx | 0 | 0.000 | 0.000 |
| C = Cys | 4 | 1.538 | 0.531 |
| D = Asp | 10 | 3.846 | 0.699 |
| E = Glu | 17 | 6.538 | 1.090 |
| F = Phe | 9 | 3.462 | 0.962 |
| G = Gly | 19 | 7.308 | 0.870 |
| H = His | 6 | 2.308 | 1.154 |
| I = Ile | 12 | 4.615 | 1.026 |
| J = --- | 0 | 0.000 | 0.000 |
| K = Lys | 5 | 1.923 | 0.291 |
| L = Leu | 23 | 8.846 | 1.195 |
| M = Met | 7 | 2.692 | 1.584 |

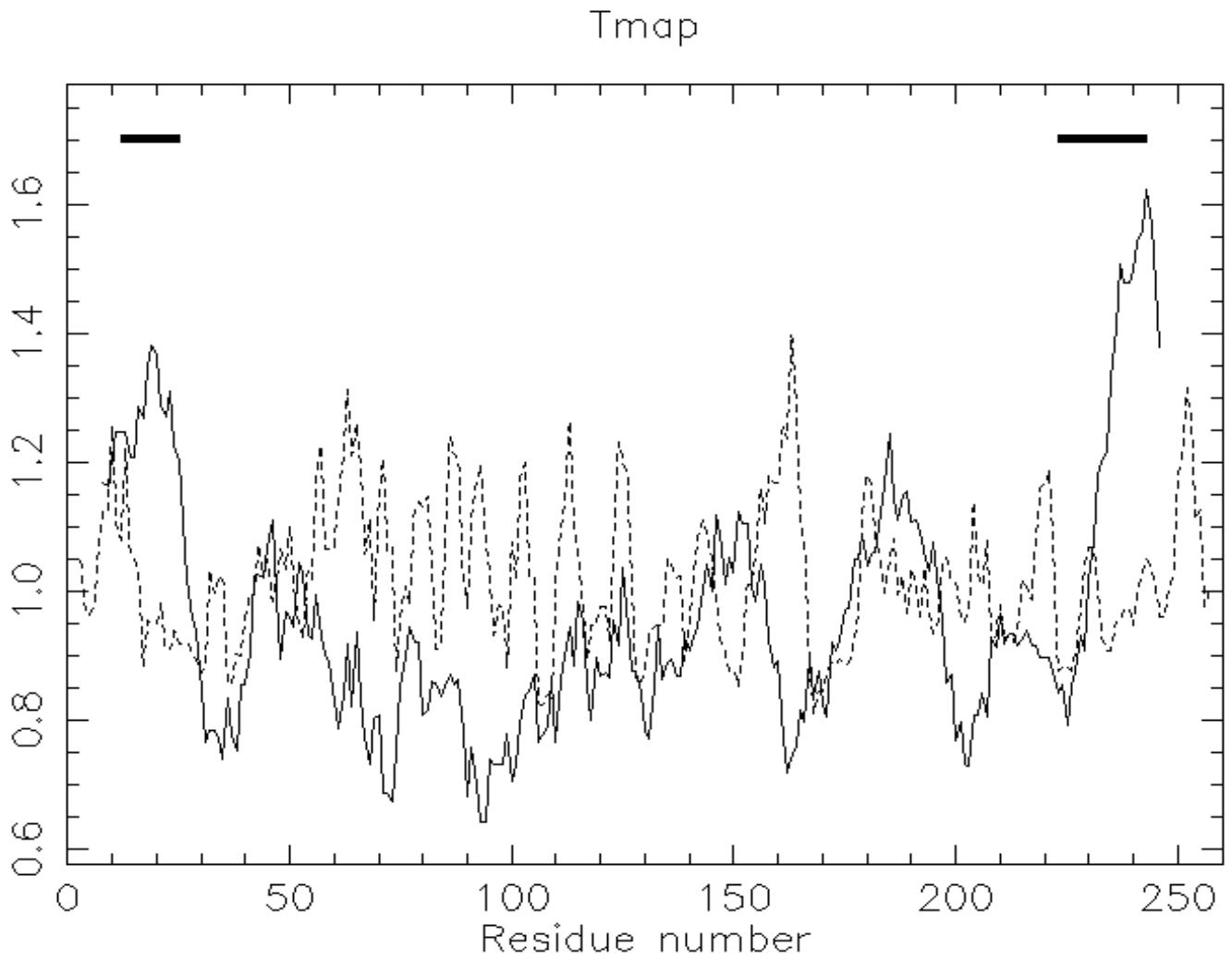
| Residue | Number | Mole% | DayhoffStat |
|---|---|---|---|
| N = Asn | 8 | 3.077 | 0.716 |
| O = --- | 0 | 0.000 | 0.000 |
| P = Pro | 10 | 3.846 | 0.740 |
| Q = Gln | 13 | 5.000 | 1.282 |
| R = Arg | 27 | 10.385 | 2.119 |
| S = Ser | 16 | 6.154 | 0.879 |
| T = Thr | 20 | 7.692 | 1.261 |
| U = --- | 0 | 0.000 | 0.000 |
| V = Val | 25 | 9.615 | 1.457 |
| W = Trp | 6 | 2.308 | 1.775 |
| X = Xaa | 0 | 0.000 | 0.000 |
| Y = Tyr | 9 | 3.462 | 1.018 |
| Z = Glx | 0 | 0.000 | 0.000 |

# 氨基酸组成分析

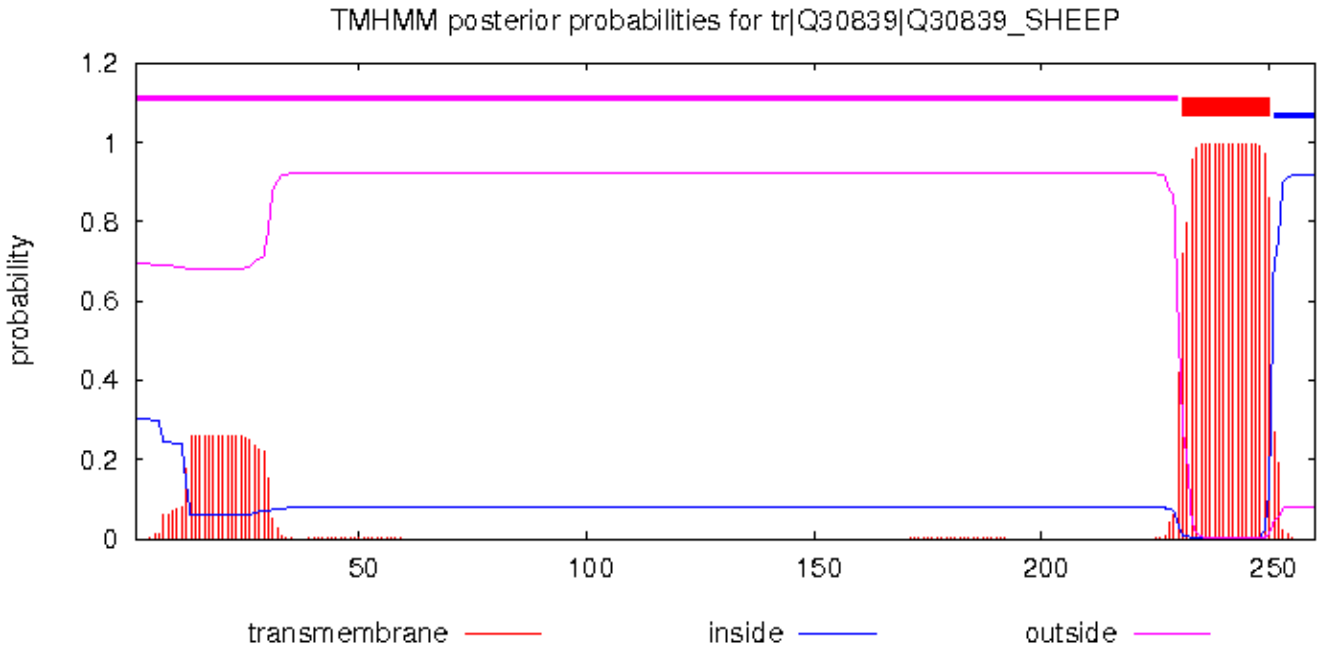
| Property | Residues | Number | Mole% |
|---|---|---|---|
| Tiny (A+C+G+S+T) | | 73 | 28.077 |
| Small (A+B+C+D+G+N+P+S+T+V) | | 126 | 48.462 |
| Aliphatic (A+I+L+V) | | 74 | 28.462 |
| Aromatic (F+H+W+Y) | | 30 | 11.538 |
| Non-polar (A+C+F+G+I+L+M+P+V+W+Y) | | 138 | 53.077 |
| Polar (D+E+H+K+N+Q+R+S+T+Z) | | 122 | 46.923 |
| Charged (B+D+E+H+K+R+Z) | | 65 | 25.000 |
| Basic (H+K+R) | | 38 | 14.615 |
| Acidic (B+D+E+Z) | | 27 | 10.385 |

该酶的等电点为9. 1336，含有的极性氨基酸占46. 923%，非极性占53. 077%，带电的氨基酸有65个，占25. 00%, 总带的电荷量为8. 0。

# 跨膜螺旋预测



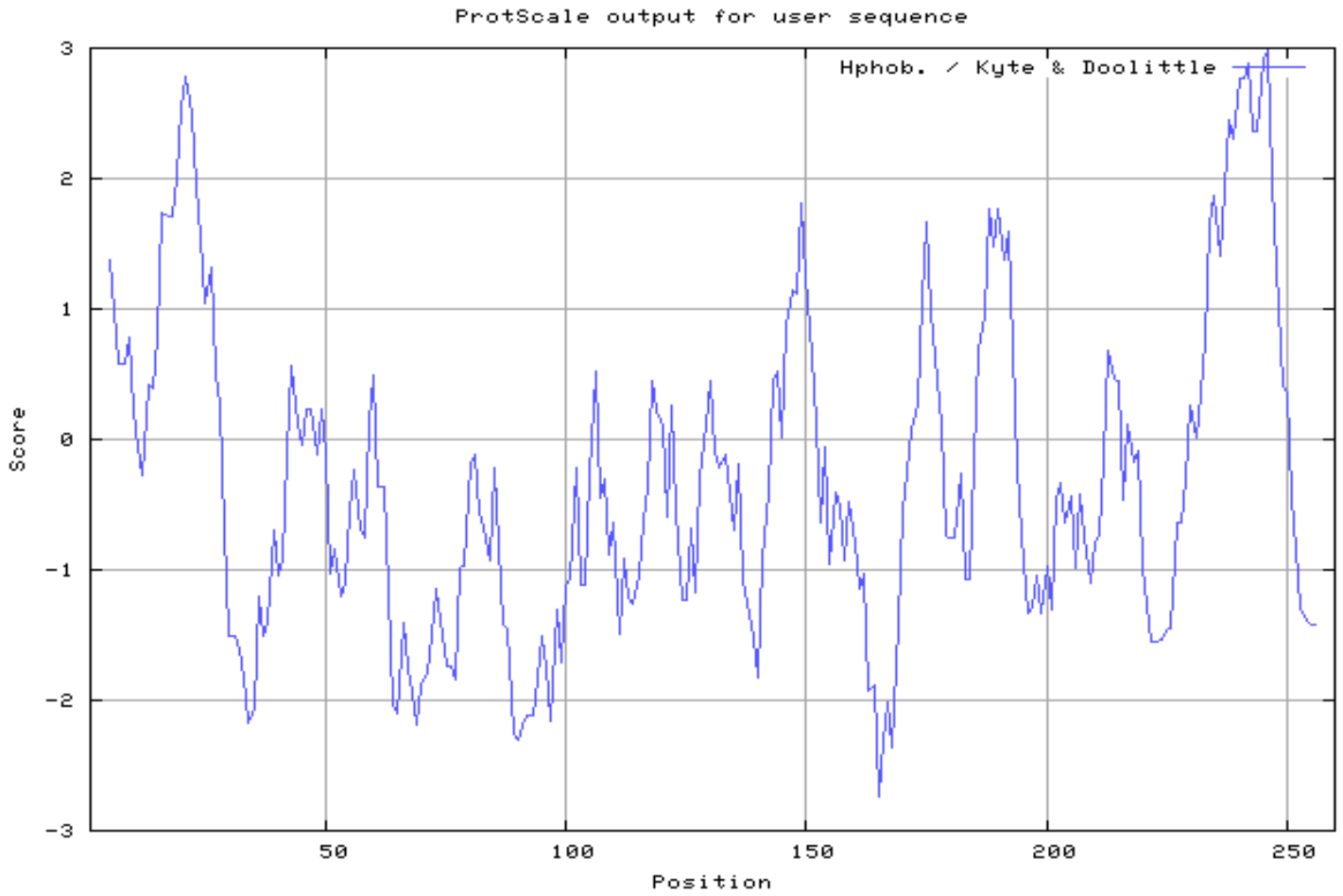
# 跨膜螺旋分析




```
# tr|Q30839|Q30839_SHEEP Length: 260
# tr|Q30839|Q30839_SHEEP Number of predicted TMHs:  1
# tr|Q30839|Q30839_SHEEP Exp number of AAs in TMHs: 25.4142
# tr|Q30839|Q30839_SHEEP Exp number, first 60 AAs:  5.13484
# tr|Q30839|Q30839_SHEEP Total prob of N-in:        0.30254
tr|Q30839|Q30839_SHEEP  TMHMM2.0        outside      1   230
tr|Q30839|Q30839_SHEEP  TMHMM2.0        TMhelix    231   250
tr|Q30839|Q30839_SHEEP  TMHMM2.0        inside     251   260
```

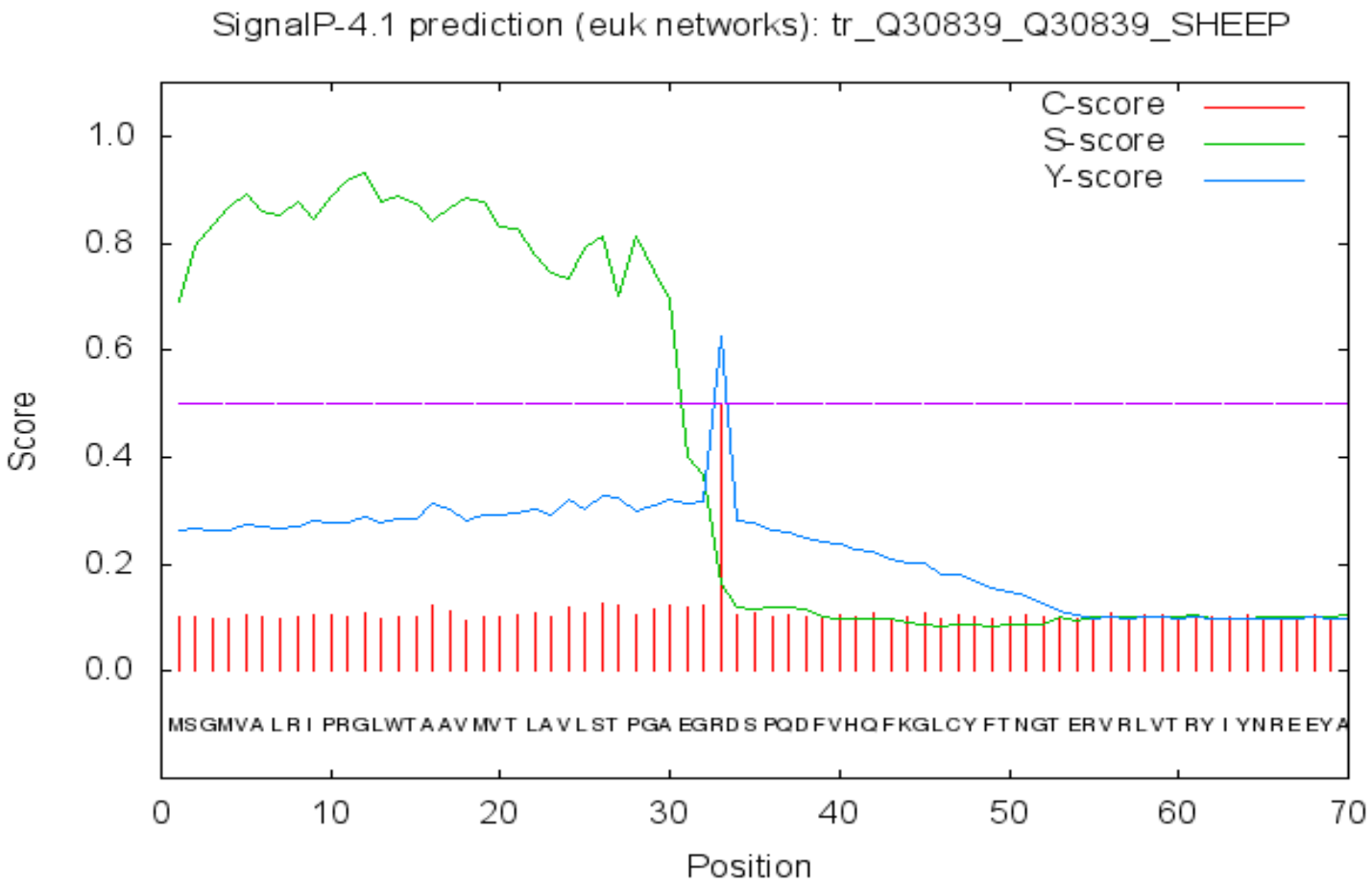
该蛋白质有1个跨膜区。曲线的纵坐标是概率，横坐标是序列，一共260个氨基酸，红色表示跨膜区，蓝色inside即在膜内部。

# 疏水性和亲水性分析




图中以0为界，正值表示疏水，负值表示亲水，该蛋白为亲水蛋白。

# 信号肽分析

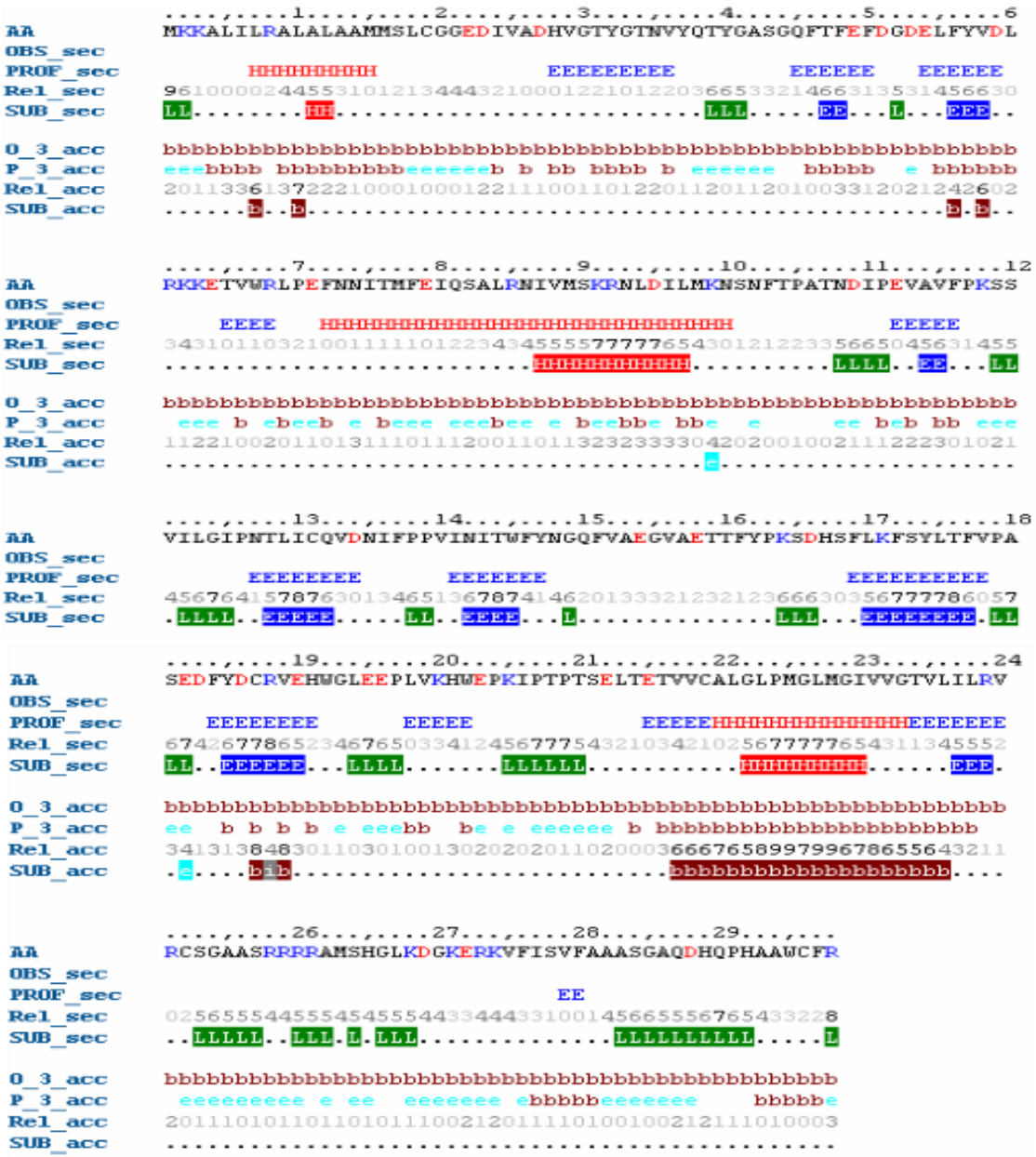



| # Measure | | Position | Value | Cutoff | signal peptide? |
|---|---|---|---|---|---|
| max. | C | 33 | 0.502 | | |
| max. | Y | 33 | 0.626 | | |
| max. | S | 12 | 0.931 | | |
| mean | S | 1-32 | 0.800 | | |
| | D | 1-32 | 0.720 | 0.450 | YES |

# 亚细胞定位

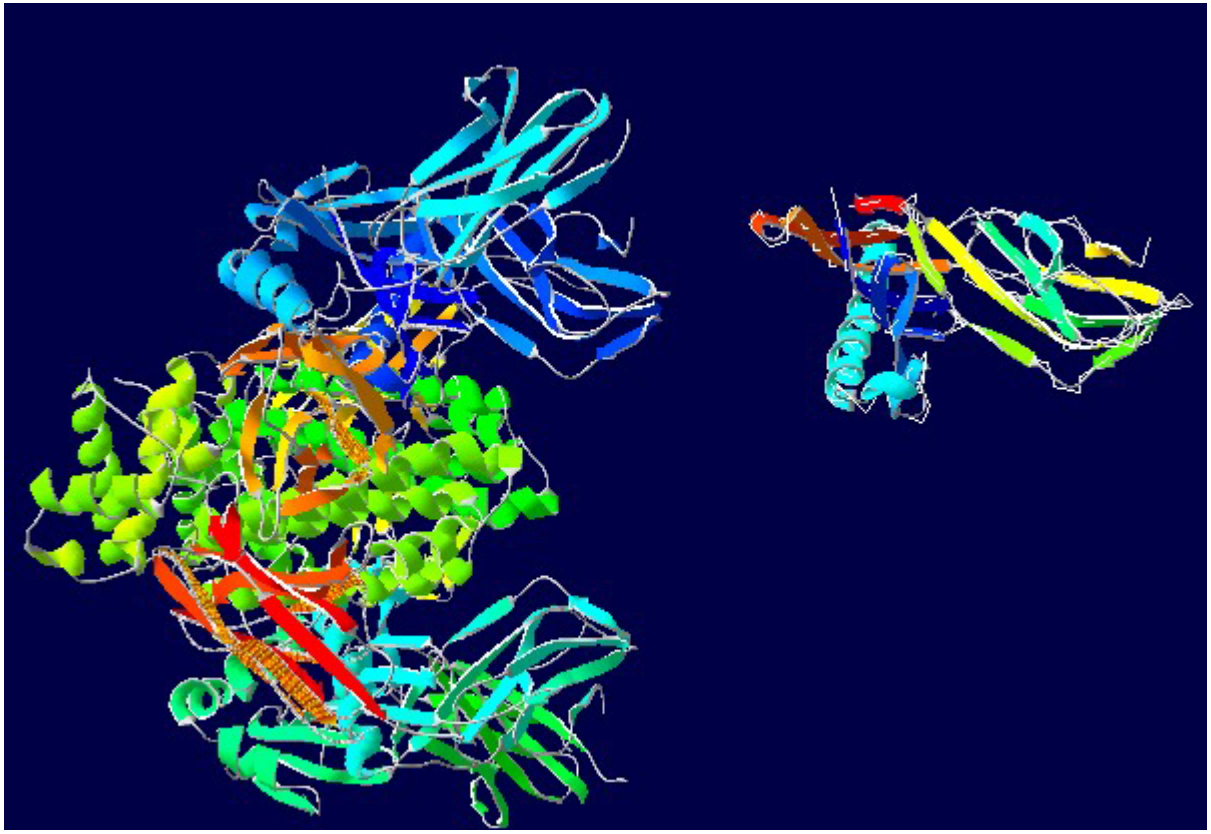
```
### targetp v1.1 prediction results ###################################
Number of query sequences:  1
Cleavage site predictions not included.
Using NON-PLANT networks.

Name                   Len            mTP     SP  other  Loc  RC
----------------------------------------------------------------------
tr_Q30839_Q30839_SHE  260          0.160  0.879  0.025   S    2
----------------------------------------------------------------------
cutoff                             0.000  0.000  0.000
```

# 蛋白二级结构预测

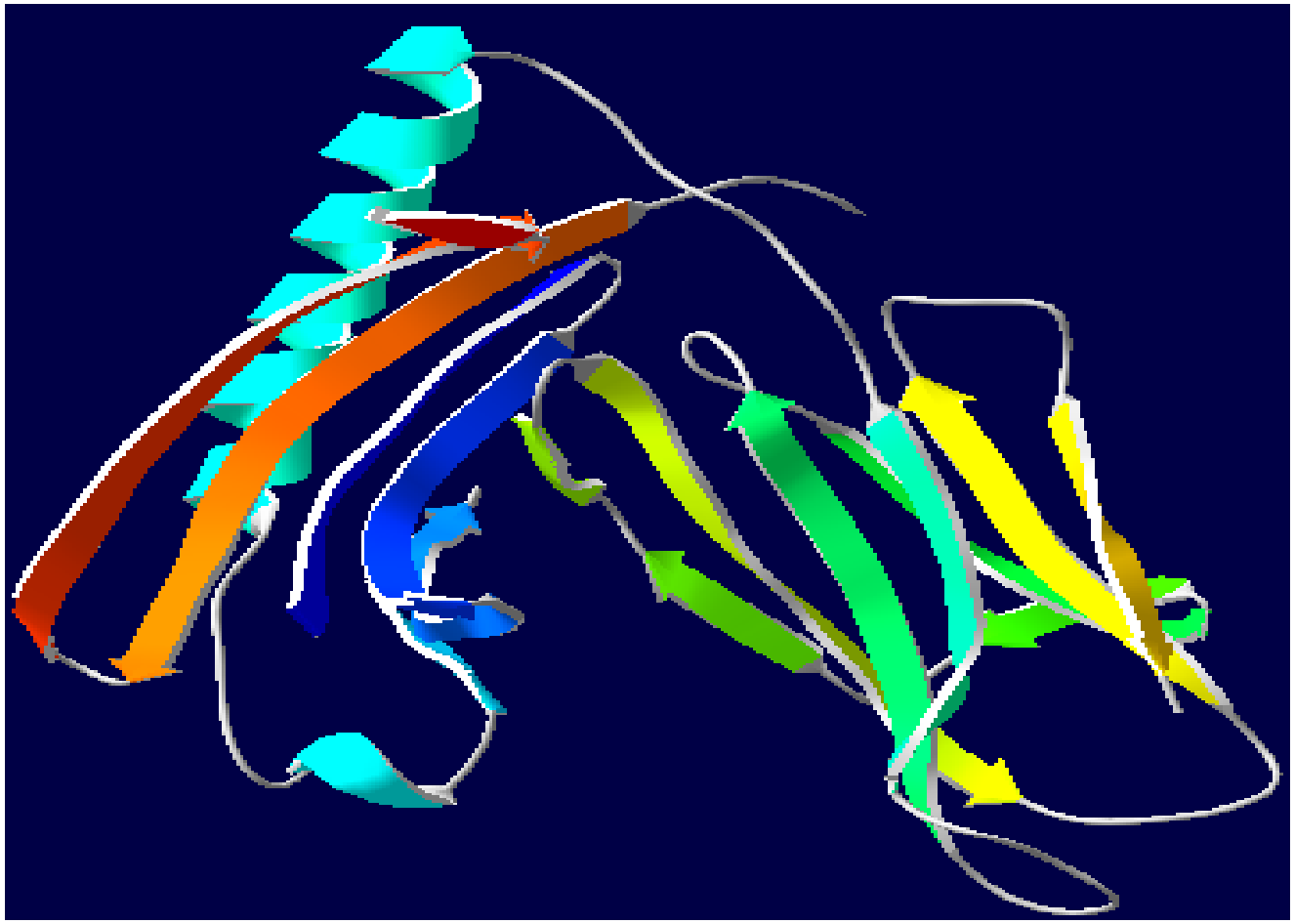


螺旋结构占总序列的18.06%，折叠结构占28.47%，卷曲结构占53.47%，该蛋白质卷曲程度较高。

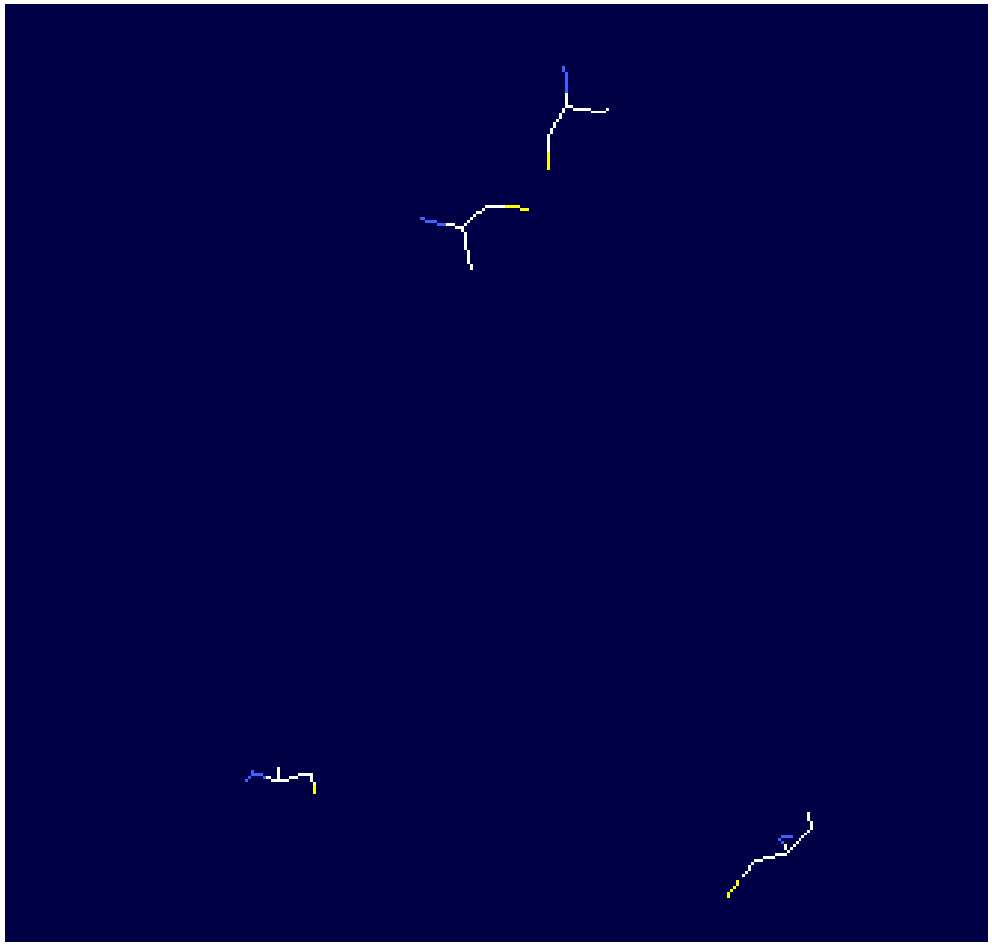
# 蛋白质三级结构预测



该蛋白的三级结构是基于TCR的三级结构为模板，两者序列一致性达80.22%。



该蛋白有2个结构域。



该蛋白有4个半胱氨酸，无二硫键。

# 致谢

衷心感谢罗老师对我们的悉心指导和谆谆教诲，让我们发现和体会到了生物信息学的乐趣和价值！

非常感谢组员及同学们的支持和帮助，使我们的学习更轻松！