胡杨*Nac*基因的生物信息学分析及其功能预测

- 14组汇报人：连通
- 组员：王华青　师双锋　陈茂然
- 生物所

2014.01.13

汇报内容

- 1.研究背景
- 2.基因的生物信息学分析
- 3.基因的功能预测
- 4.总结

研究背景

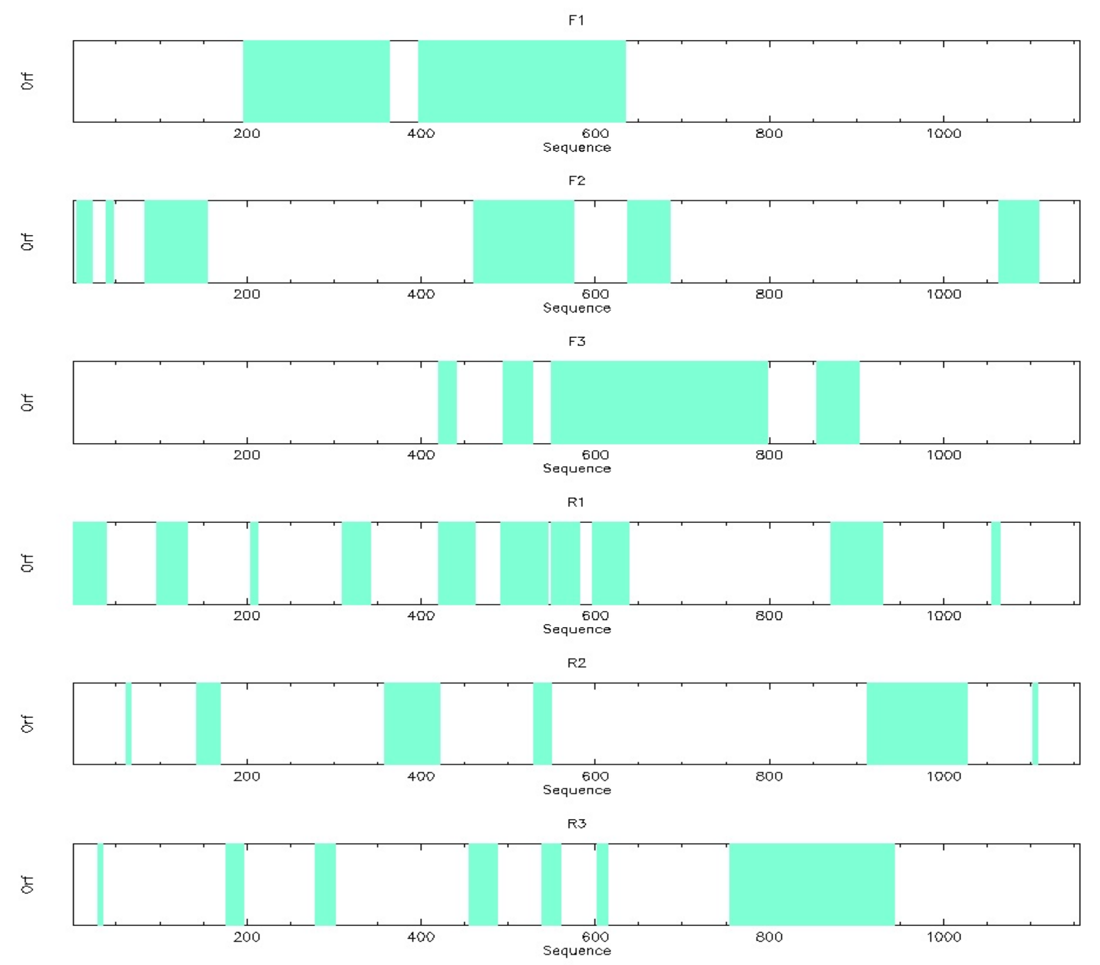
- *NAC*家族是近年来发现的植物特有的且具有多种生物功能的转录因子。研究表明该家族基因广泛参与枝顶端分生组织发育、花器官形成、侧根发育、叶序、逆境诱导开花、胚发育、细胞次生壁合成以及从叶到粒的营养代谢控制等。*NAC*家族基因也参与生物与非生物逆境反应，如枯萎病、干旱、高盐等。因此研究*NAC*基因家族对于植物改良具有重要意义，已成为植物学者们关注的热点。

研究背景

- 目前对*NAC*基因的主要研究主要集中在模式生物拟南芥和水稻，对于*NAC*蛋白涉及的调控途径及其组成因子知之甚少，因此*NAC*基因的功能还有待深入研究
- 本课题所用的胡杨NAC基因由导师提供，该课题目前处于保密研究阶段，因此信息较为有限。

生物信息学分析

综合运用plotorf、getorf、sixpack以及transeq等工具获得蛋白质序列信息

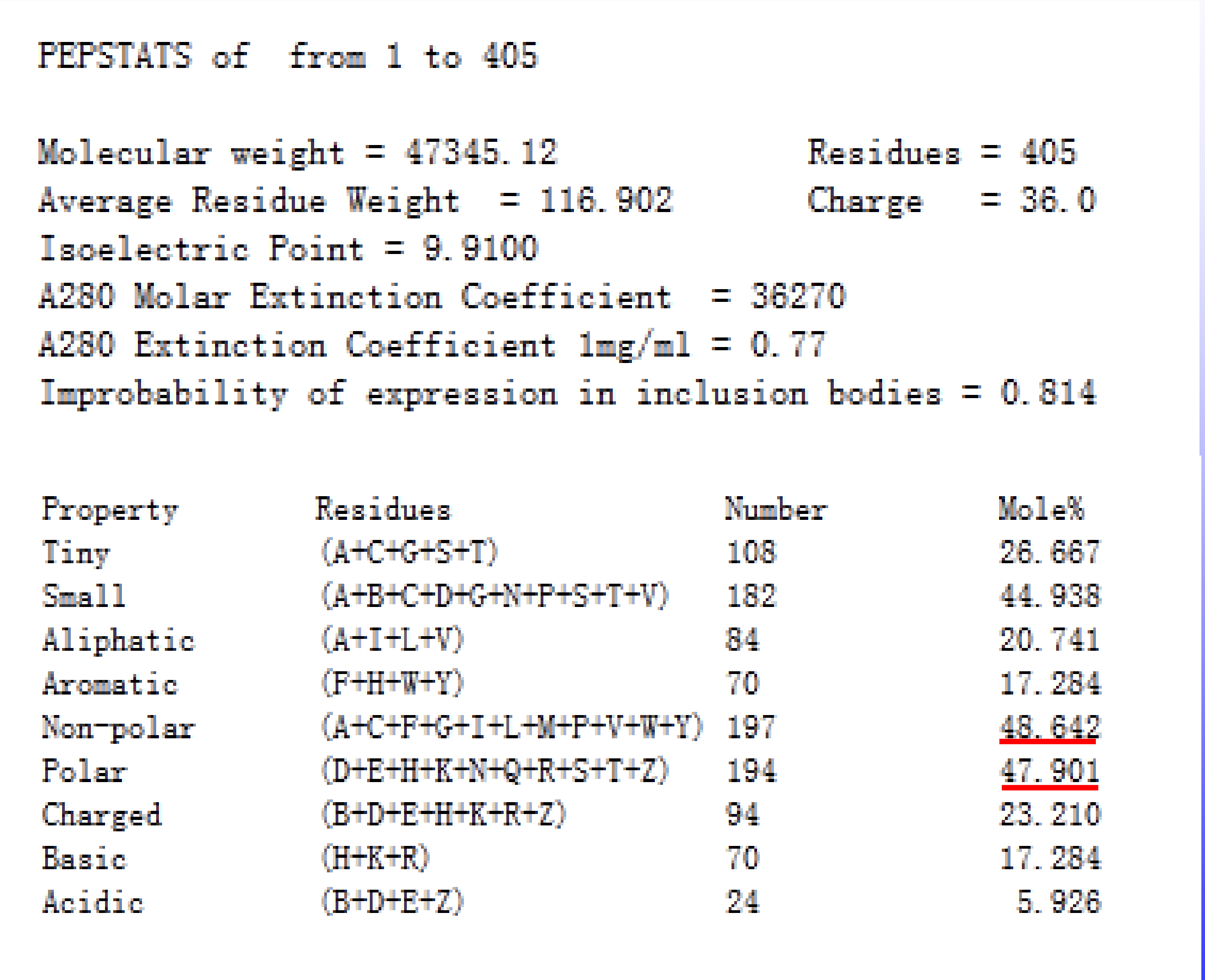


获得蛋白质序列

>Populus euphratica

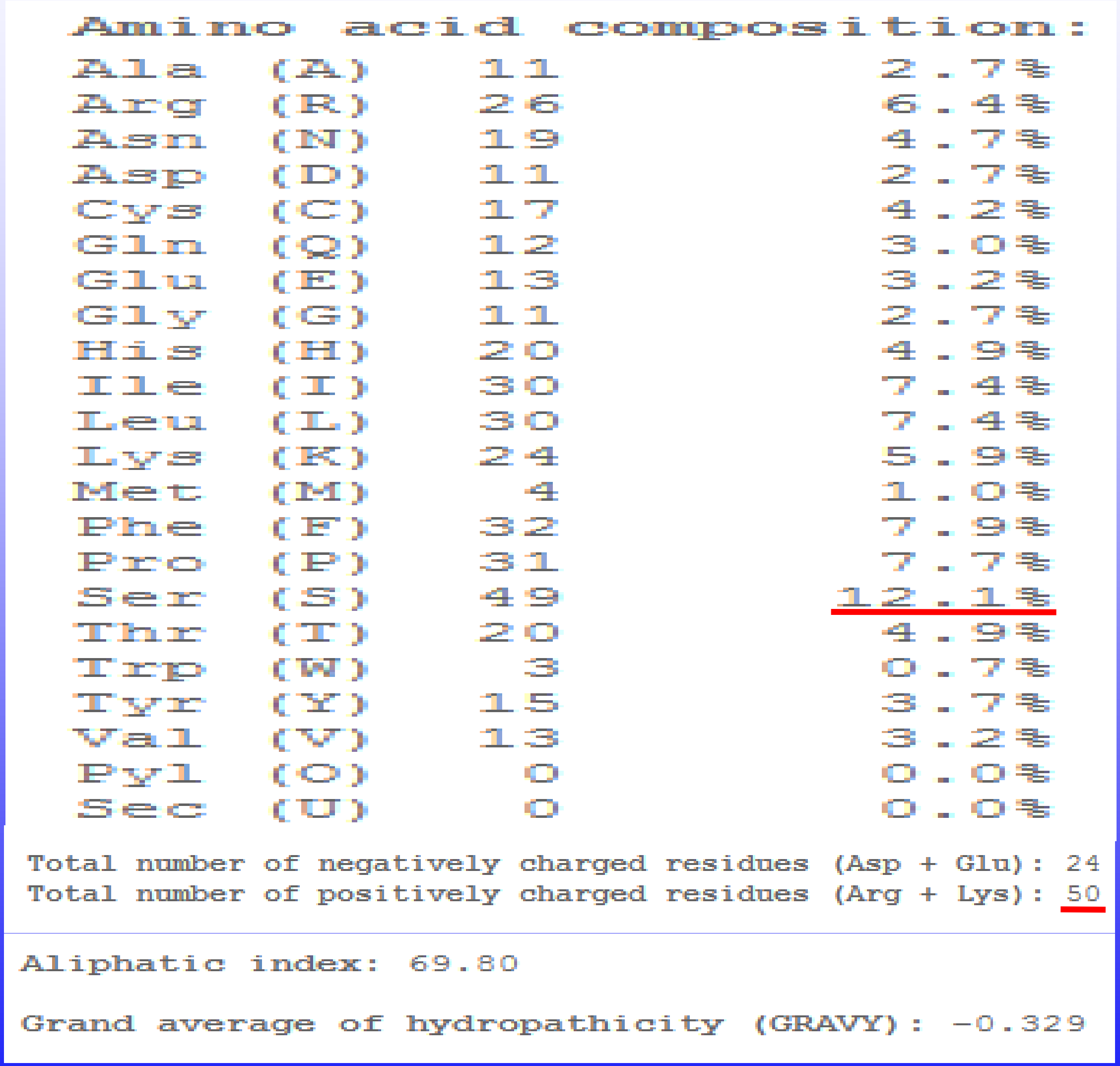
......NLKFNYIYIYLNGKCLVQQGKLFHTCSVNKIDYPIEXLIXWFKFIARYXCY
RFSSSFSFYFFPTREKHIVLYNKNKKHDLKDPXTEQSSDSPPKSSGSCSRSSFSS
PCPVPEHEIRWAVFPEIENTYLSFAYDLAKNAINFTRSNESPTIIVMCRKDNNX
IDXSIPXISPLFFYHIRMHQI......YTESFIPFLSSLFSHCQNSPVLACSRSHVGLPL
SAHLPFNWRLLHSLPTYRLTGGSLSKRALRSKRFIKTRPIPIRXIPGTKIFFFFFS
ARSPGCMDSYTNTTVQNKCRKQXPNHPHCEPLTRTHPHFVSCHVSRQSGKE
RKNXSLTCILEIIPATCPSISLINSCFSLEIXSFKRRHCIRP......

（1）运用Weblab中的Pepstats工具得到氨基酸组成的基本信息



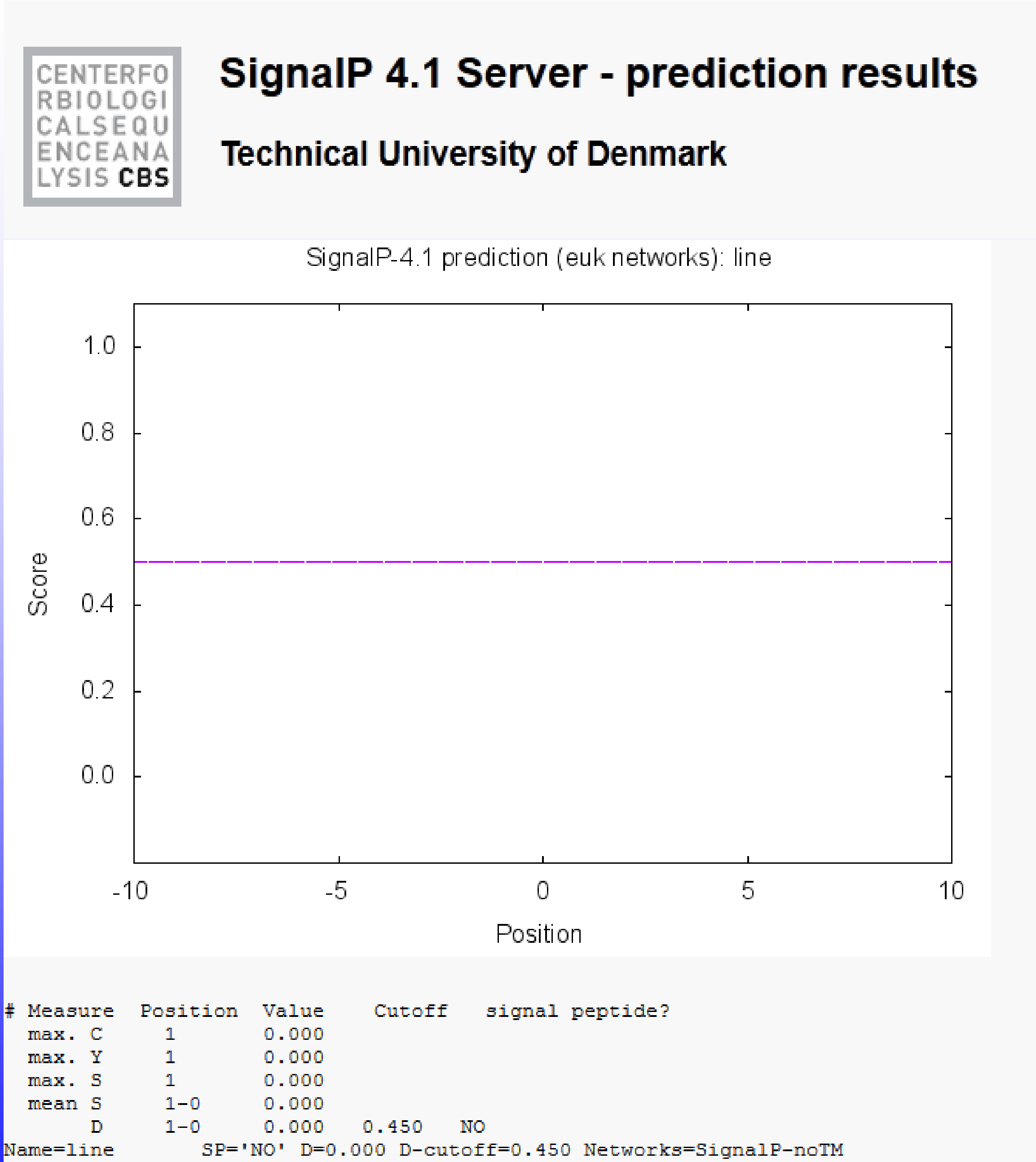
```
PEPSTATS of  from 1 to 405

Molecular weight = 47345.12              Residues = 405
Average Residue Weight  = 116.902        Charge   = 36.0
Isoelectric Point = 9.9100
A280 Molar Extinction Coefficient  = 36270
A280 Extinction Coefficient 1mg/ml = 0.77
Improbability of expression in inclusion bodies = 0.814

Property      Residues                 Number        Mole%
Tiny          (A+C+G+S+T)              108           26.667
Small         (A+B+C+D+G+N+P+S+T+V)    182           44.938
Aliphatic     (A+I+L+V)                84            20.741
Aromatic      (F+H+W+Y)                70            17.284
Non-polar     (A+C+F+G+I+L+M+P+V+W+Y)  197           48.642
Polar         (D+E+H+K+N+Q+R+S+T+Z)    194           47.901
Charged       (B+D+E+H+K+R+Z)          94            23.210
Basic         (H+K+R)                  70            17.284
Acidic        (B+D+E+Z)                24             5.926
```

Protparam分析氨基酸组成信息



```
Amino acid composition:
Ala (A)  11          2.7%
Arg (R)  26          6.4%
Asn (N)  19          4.7%
Asp (D)  11          2.7%
Cys (C)  17          4.2%
Gln (Q)  12          3.0%
Glu (E)  13          3.2%
Gly (G)  11          2.7%
His (H)  20          4.9%
Ile (I)  30          7.4%
Leu (L)  30          7.4%
Lys (K)  24          5.9%
Met (M)   4          1.0%
Phe (F)  32          7.9%
Pro (P)  31          7.7%
Ser (S)  49         12.1%
Thr (T)  20          4.9%
Trp (W)   3          0.7%
Tyr (Y)  15          3.7%
Val (V)  13          3.2%
Pyl (O)   0          0.0%
Sec (U)   0          0.0%

Total number of negatively charged residues (Asp + Glu): 24
Total number of positively charged residues (Arg + Lys): 50

Aliphatic index: 69.80

Grand average of hydropathicity (GRAVY): -0.329
```

（2）运用SignalP 4.1进行信号肽预测



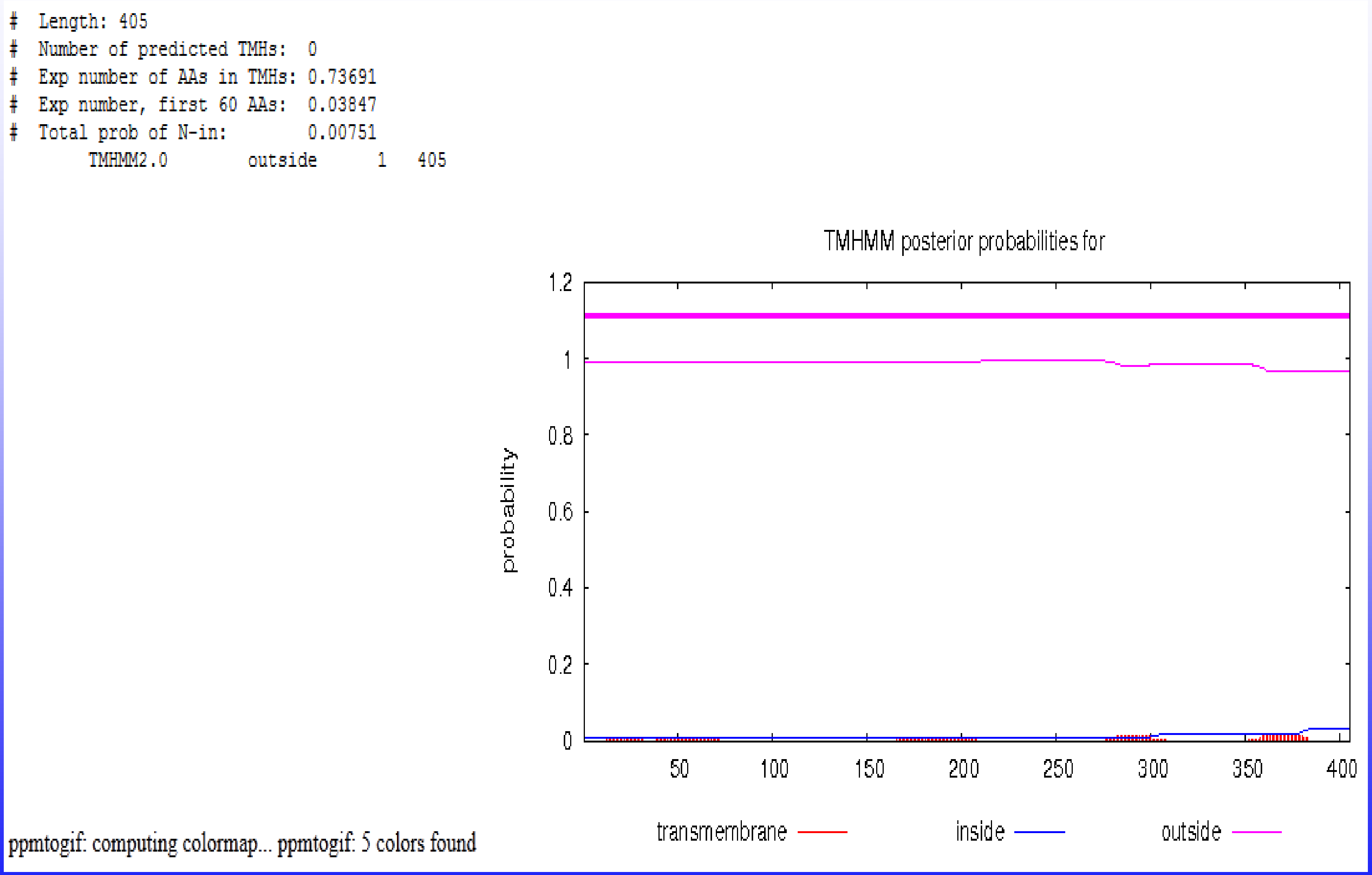
SignalP 4.1 Server - prediction results

Technical University of Denmark

```
# Measure  Position  Value    Cutoff   signal peptide?
  max. C     1       0.000
  max. Y     1       0.000
  max. S     1       0.000
  mean S     1-0     0.000
       D     1-0     0.000   0.450   NO
Name=line       SP='NO' D=0.000 D-cutoff=0.450 Networks=SignalP-noTM
```

结论：预测无信号肽

（3）TMHMM的跨膜螺旋预测

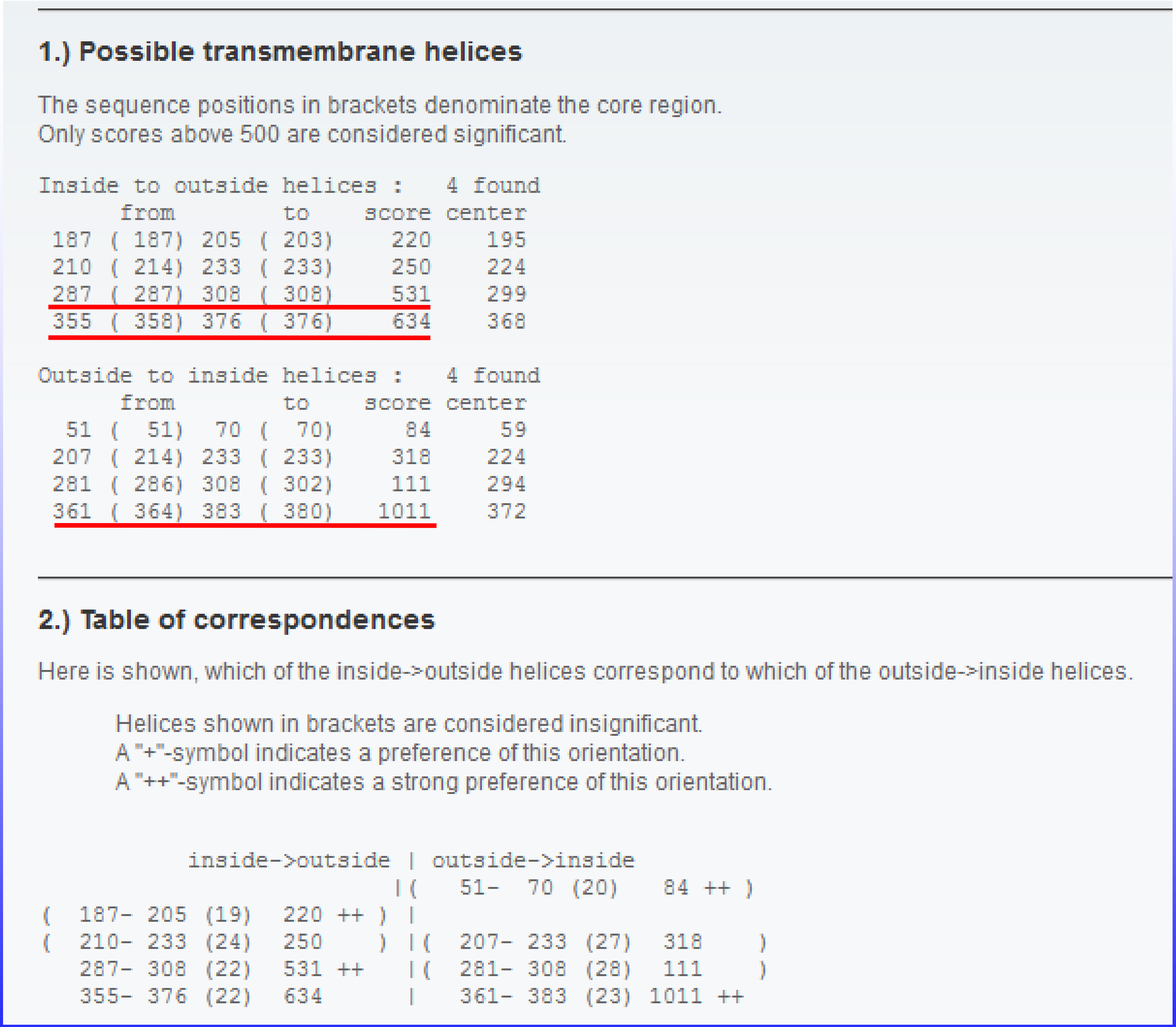


```
# Length: 405
# Number of predicted TMHs:  0
# Exp number of AAs in TMHs: 0.73691
# Exp number, first 60 AAs:  0.03847
# Total prob of N-in:        0.00751
     TMHMM2.0        outside      1   405
```

ppmtogif: computing colormap... ppmtogif: 5 colors found

Tmpred

通过分析疏水区域预测蛋白质的跨膜区域



1.) Possible transmembrane helices

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

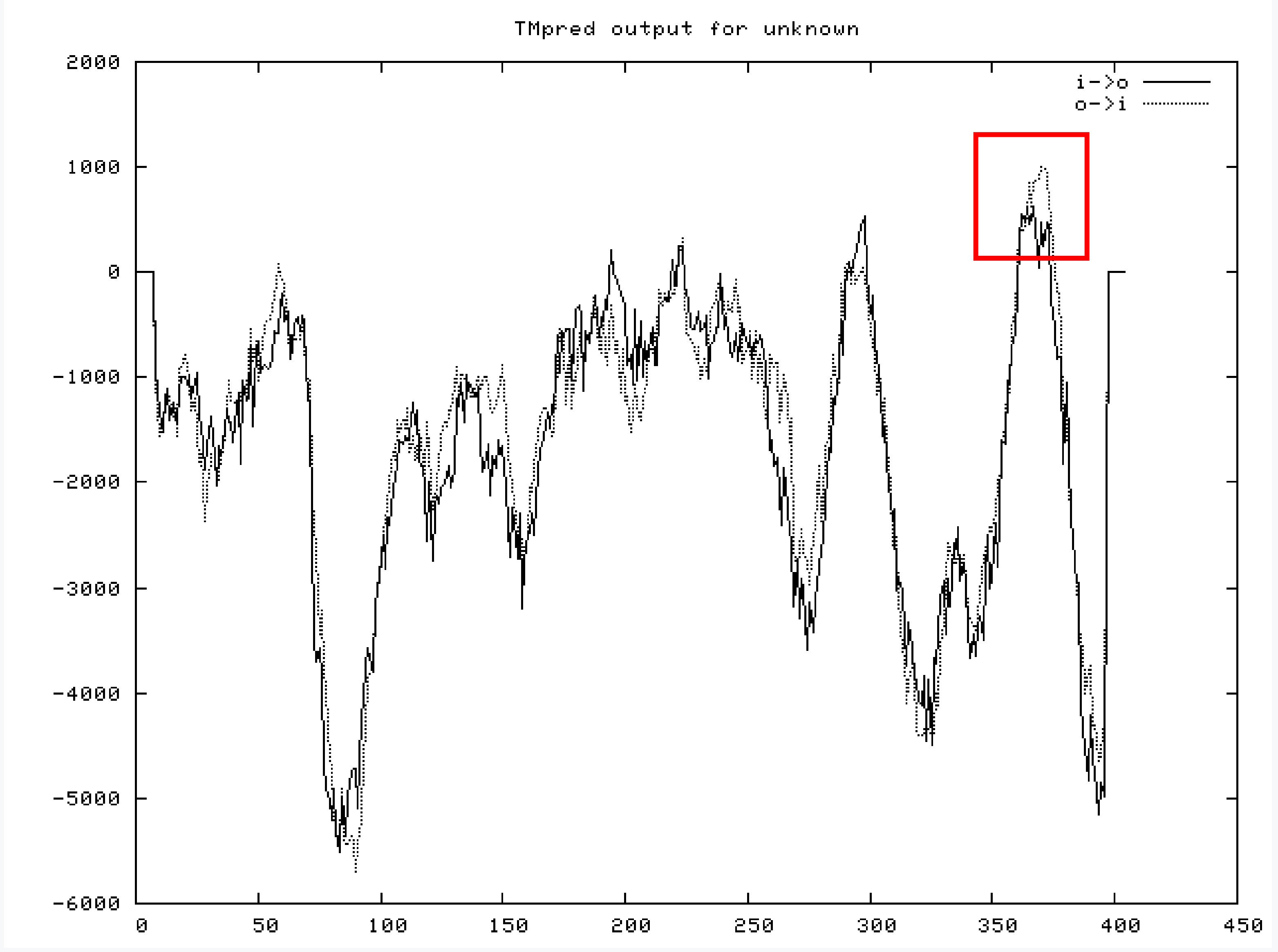
```
Inside to outside helices :   4 found
      from        to    score center
 187 ( 187) 205 ( 203)    220    195
 210 ( 214) 233 ( 233)    250    224
 287 ( 287) 308 ( 308)    531    299
 355 ( 358) 376 ( 376)    634    368

Outside to inside helices :   4 found
      from        to    score center
  51 (  51)  70 (  70)     84     59
 207 ( 214) 233 ( 233)    318    224
 281 ( 286) 308 ( 302)    111    294
 361 ( 364) 383 ( 380)   1011    372
```

2.) Table of correspondences

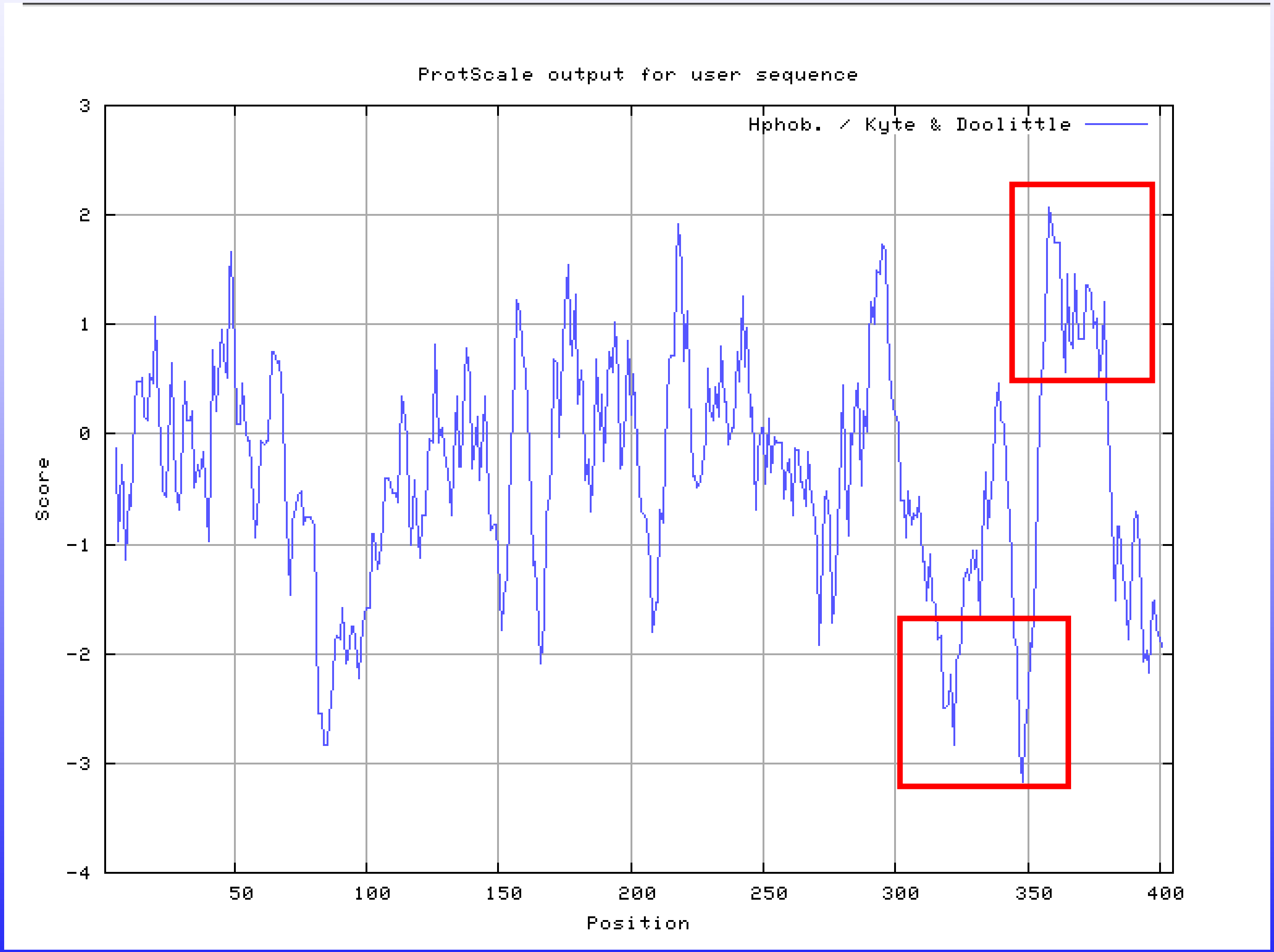
Here is shown, which of the inside->outside helices correspond to which of the outside->inside helices.

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
                           |(  51-  70 (20)   84 ++ )
(  187- 205 (19)  220 ++ ) |
(  210- 233 (24)  250    ) |(  207- 233 (27)  318    )
   287- 308 (22)  531 ++   |(  281- 308 (28)  111    )
   355- 376 (22)  634      |   361- 383 (23) 1011 ++
```


TMpred output for unknown
i->o
o->i
2000
1000
0
-1000
-2000
-3000
-4000
-5000
-6000
0
50
100
150
200
250
300
350
400
450

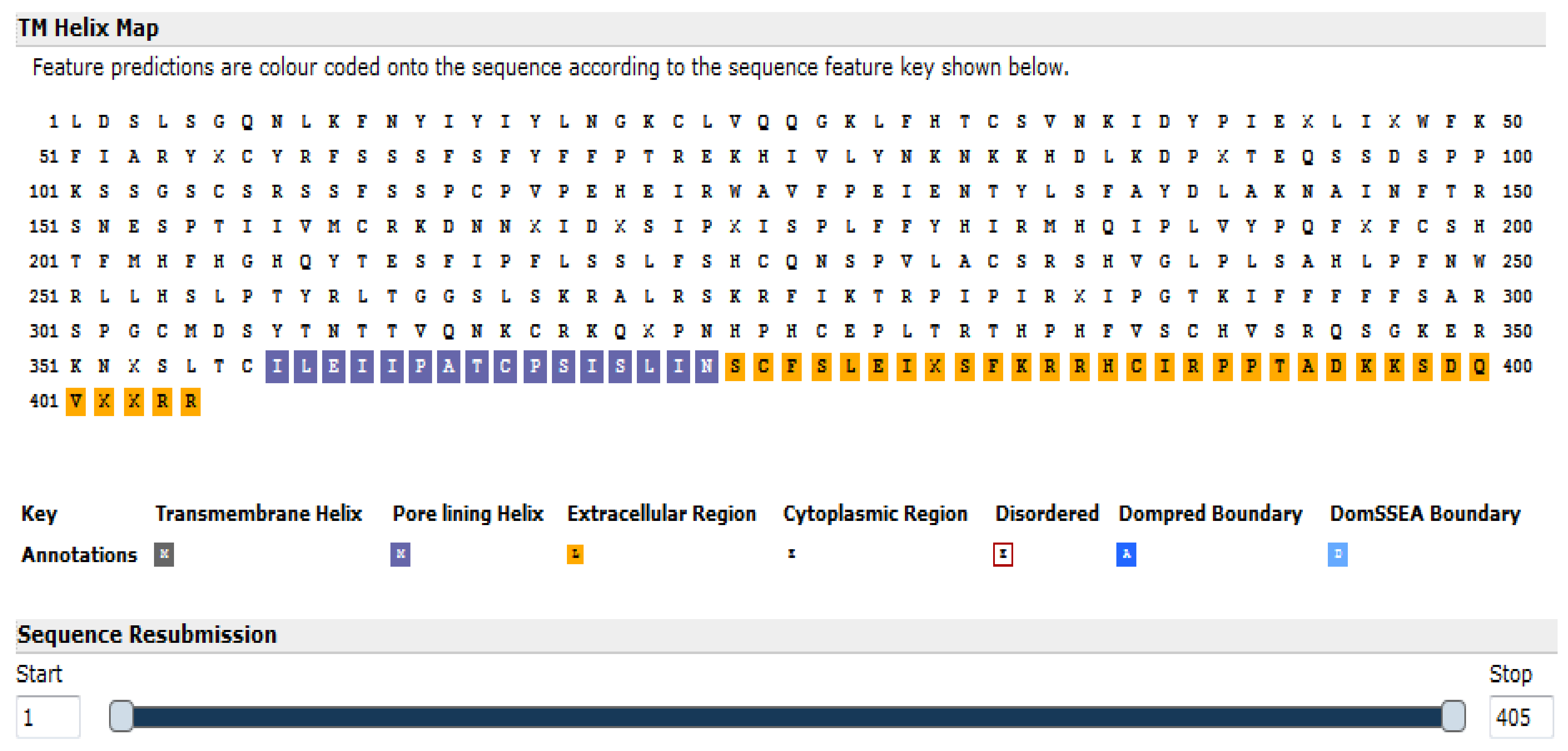
protscale的疏水性分析



Psipred跨膜预测

UCL Department Of Computer Science

Bioinformatics Group



TM Helix Map

Feature predictions are colour coded onto the sequence according to the sequence feature key shown below.

```
  1 L D S L S G Q N L K F N Y I Y I Y L N G K C L V Q Q G K L F H T C S V N K I D Y P I E X L I X W F K  50
 51 F I A R Y X C Y R F S S S F S F Y F F P T R E K H I V L Y N K N K K H D L K D P X T E Q S S D S P P 100
101 K S S G S C S R S S F S S P C P V P E H E I R W A V F P E I E N T Y L S F A Y D L A K N A I N F T R 150
151 S N E S P T I I V M C R K D N N X I D X S I P X I S P L F F Y H I R M H Q I P L V Y P Q F X F C S H 200
201 T F M H F H G H Q Y T E S F I P F L S S L F S H C Q N S P V L A C S R S H V G L P L S A H L P F N W 250
251 R L L H S L P T Y R L T G G S L S K R A L R S K R F I K T R P I P I R X I P G T K I F F F F F S A R 300
301 S P G C M D S Y T N T T V Q N K C R K Q X P N H P H C E P L T R T H P H F V S C H V S R Q S G K E R 350
351 K N X S L T C I L E I I P A T C P S I S L I N S C F S L E I X S F K R R H C I R P P T A D K K S D Q 400
401 V X X R R
```

Key	Transmembrane Helix	Pore lining Helix	Extracellular Region	Cytoplasmic Region	Disordered	Dompred Boundary	DomSSEA Boundary
Annotations	M	M	L	I	I	A	D

Sequence Resubmission

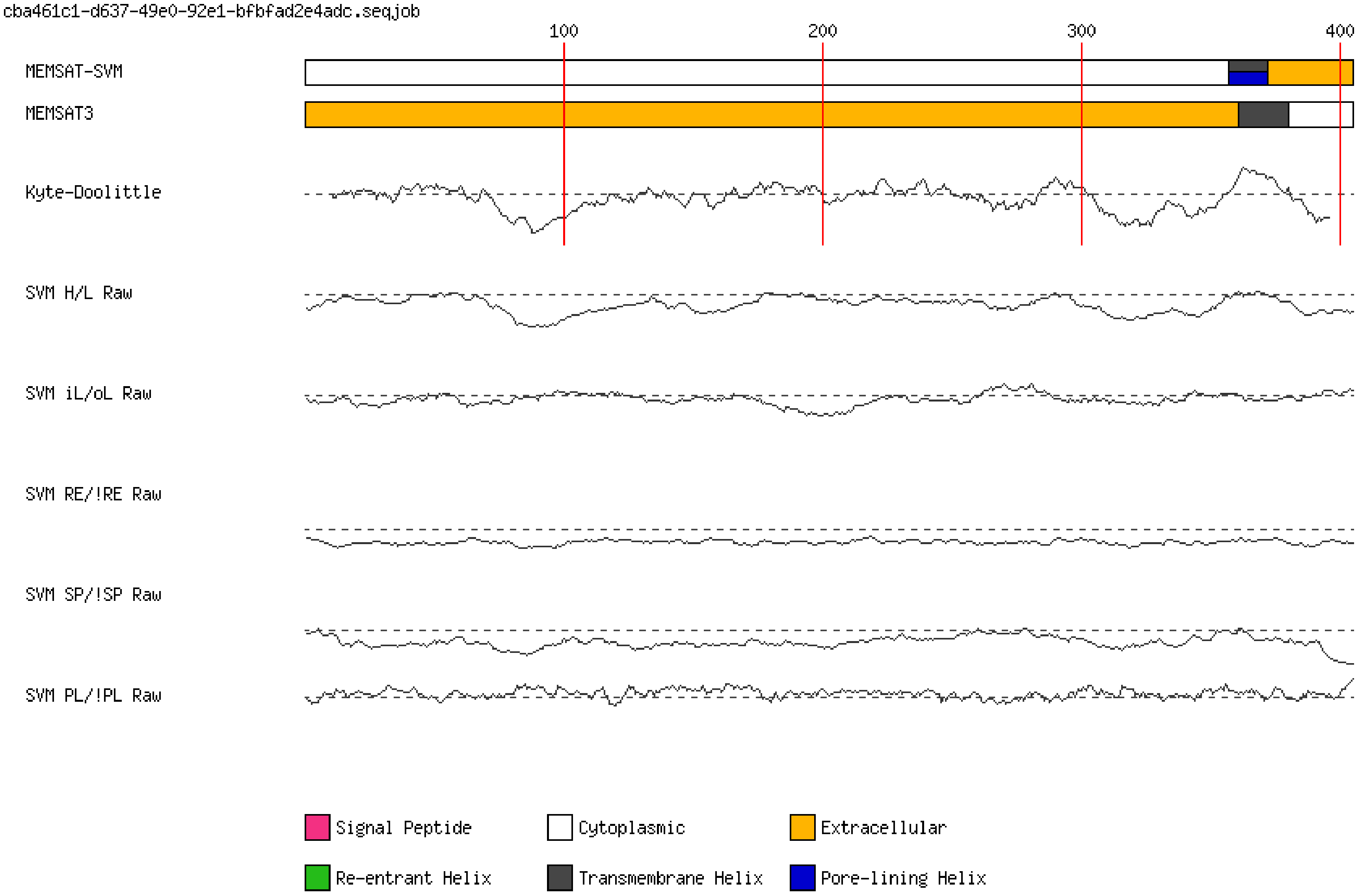
Start

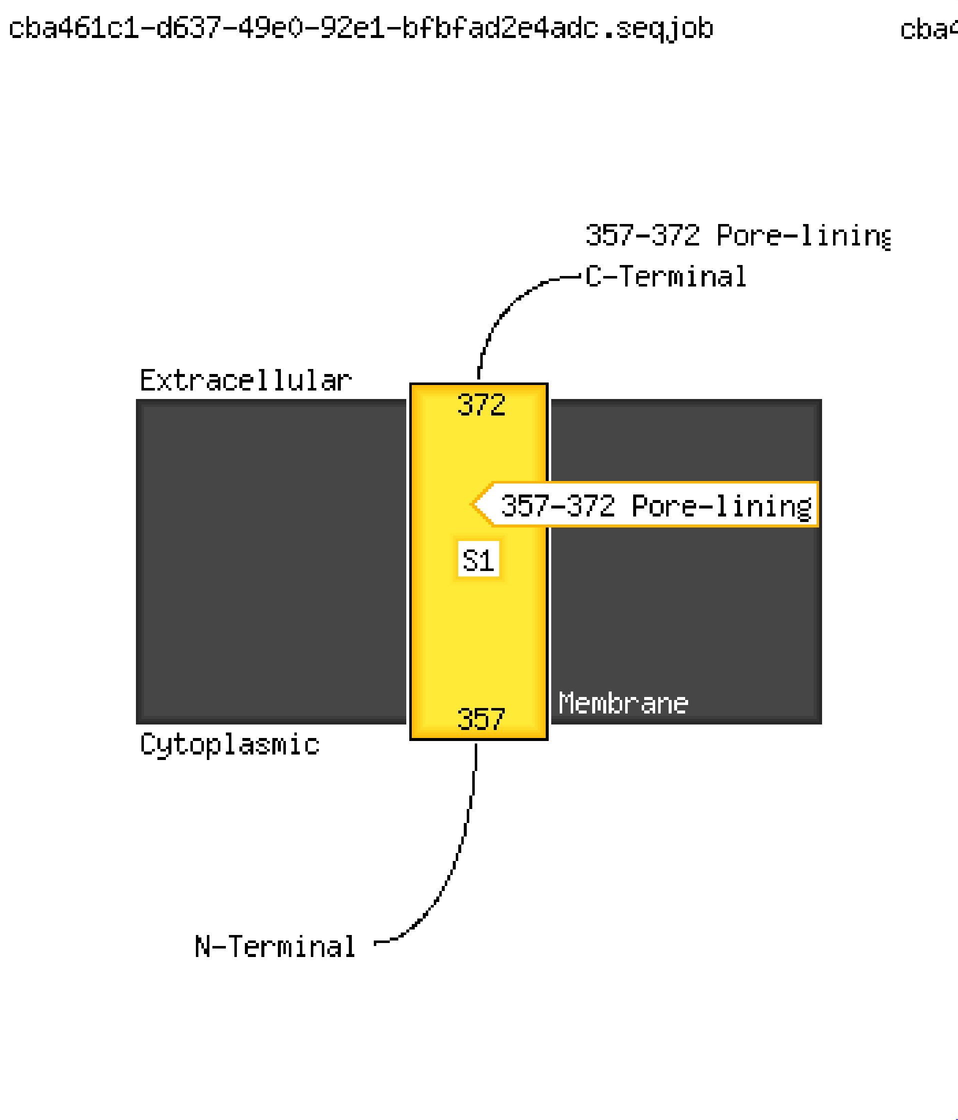
1

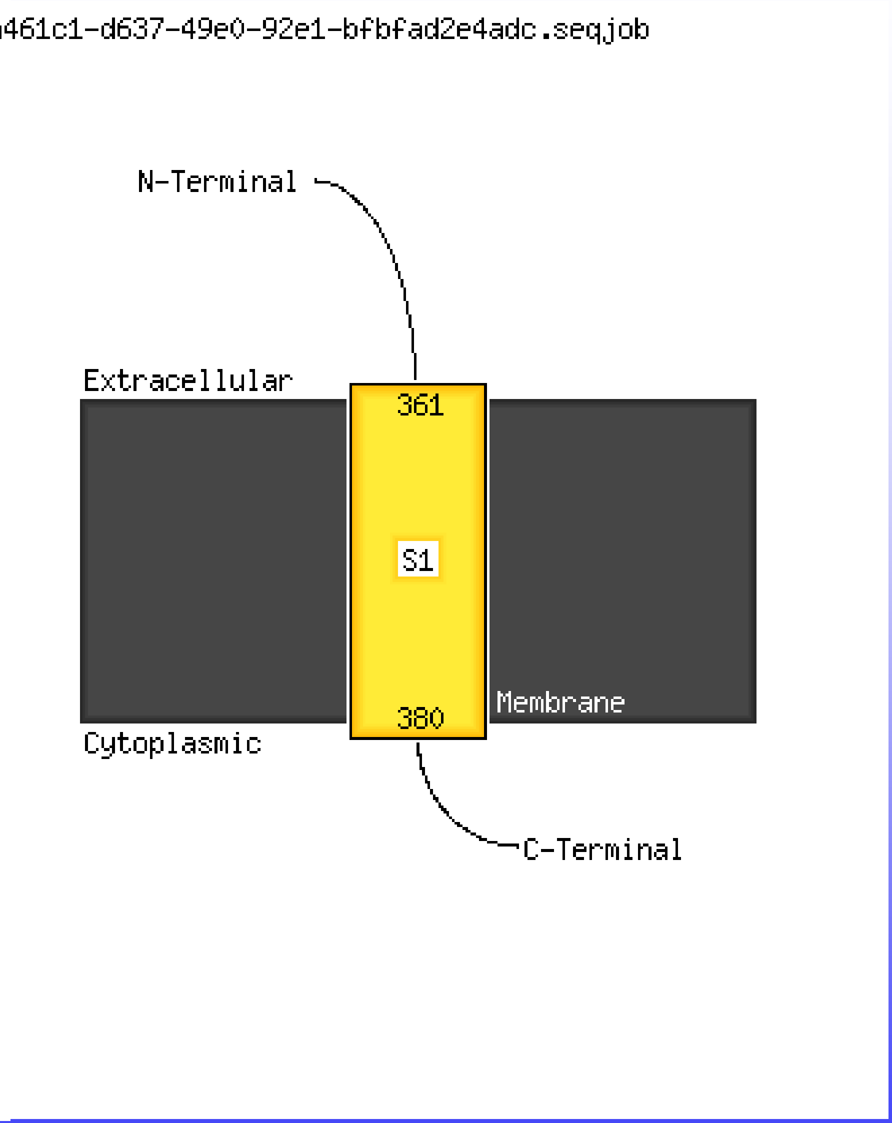
Stop

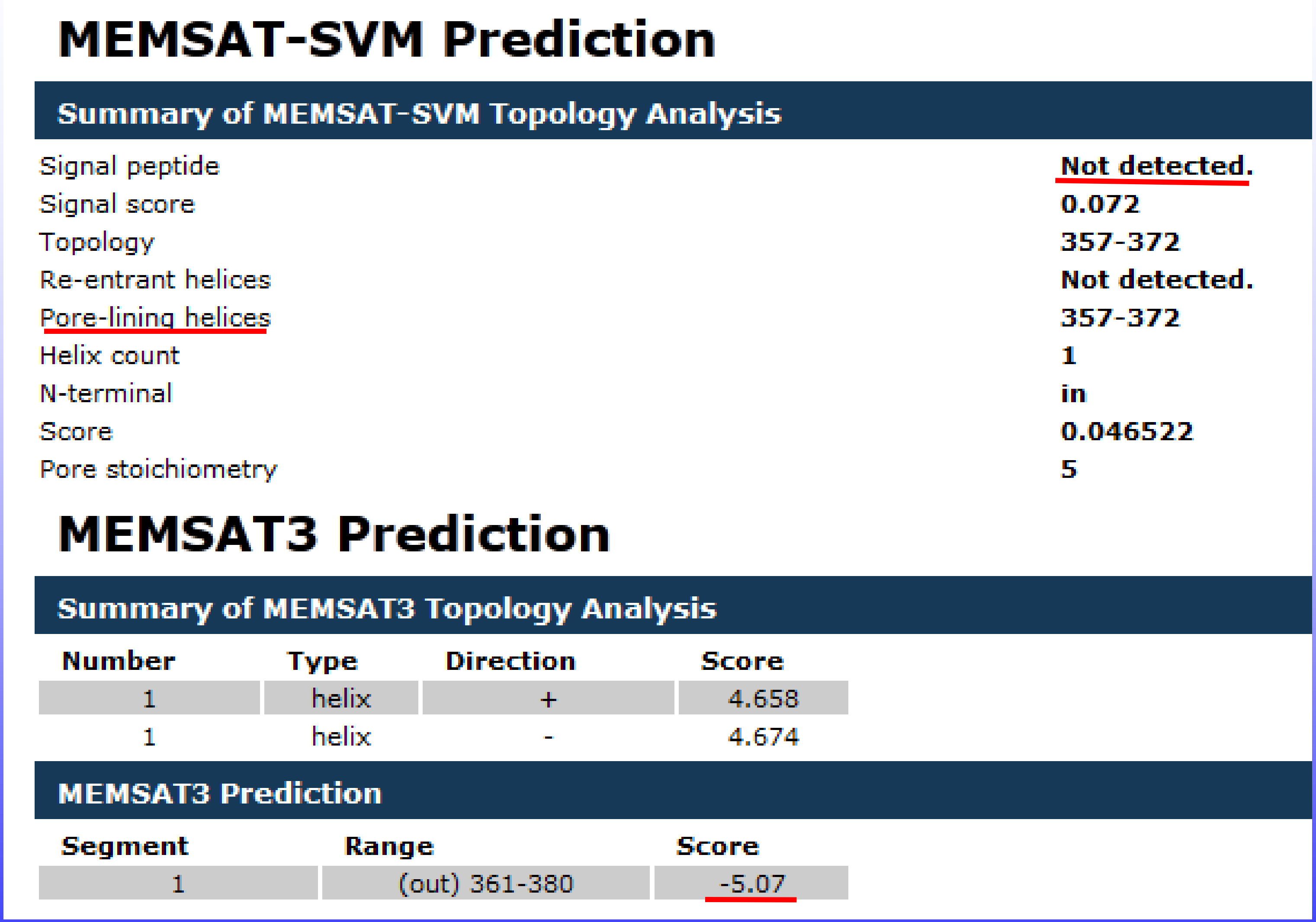
405

MEMSAT-SVM Schematic




cba461c1-d637-49e0-92e1-bfbfad2e4adc.seqjob
357-372 Pore-lining
C-Terminal
Extracellular
372
357-372 Pore-lining
S1
357
Membrane
Cytoplasmic
N-Terminal


cba461c1-d637-49e0-92e1-bfbfad2e4adc.seqjob
N-Terminal
Extracellular
361
S1
380
Membrane
Cytoplasmic
C-Terminal



MEMSAT-SVM Prediction

Summary of MEMSAT-SVM Topology Analysis

Signal peptide	**Not detected.**
Signal score	**0.072**
Topology	**357-372**
Re-entrant helices	**Not detected.**
Pore-lining helices	**357-372**
Helix count	**1**
N-terminal	**in**
Score	**0.046522**
Pore stoichiometry	**5**

MEMSAT3 Prediction

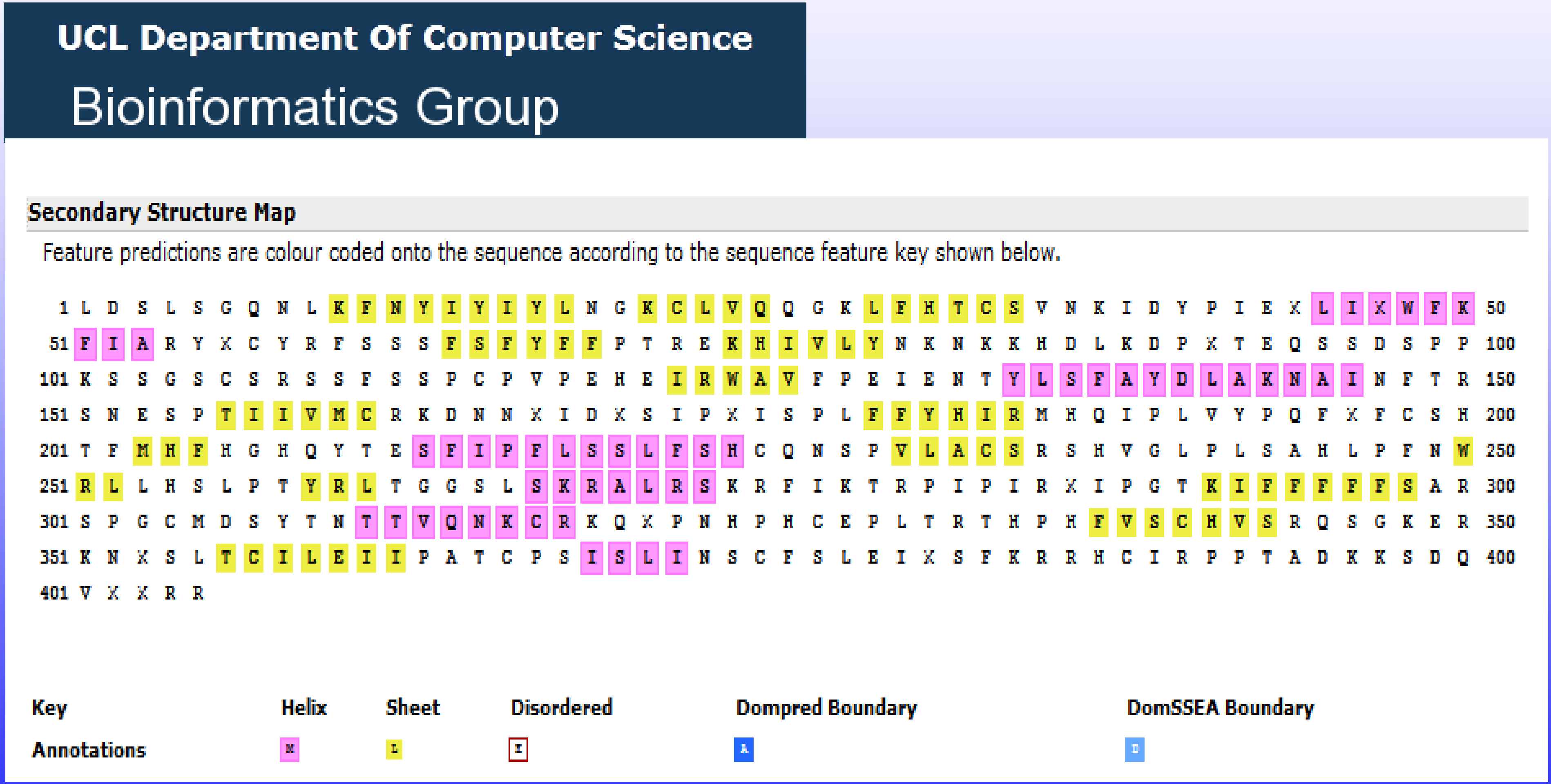
Summary of MEMSAT3 Topology Analysis

Number	Type	Direction	Score
1	helix	+	4.658
1	helix	-	4.674

MEMSAT3 Prediction

Segment	Range	Score
1	(out) 361-380	-5.07

（4）利用Expasy服务器下的SWISS-MODEL中的Psipred工具预测基因编码蛋白质的二级结构



UCL Department Of Computer Science

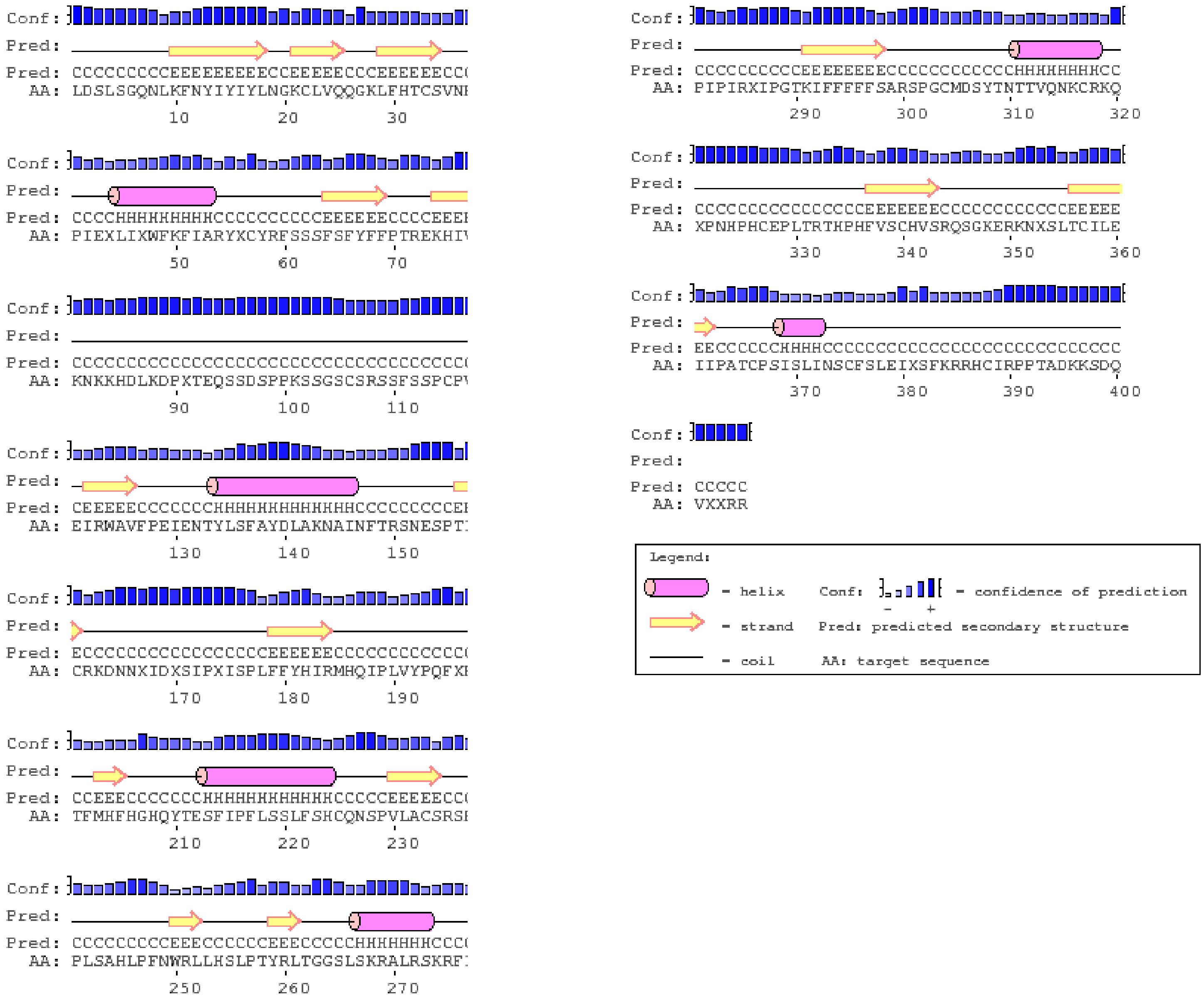
Bioinformatics Group

Secondary Structure Map

Feature predictions are colour coded onto the sequence according to the sequence feature key shown below.

```
  1 L D S L S G Q N L K F N Y I Y I Y L N G K C L V Q Q G K L F H T C S V N K I D Y P I E X L I X W F K 50
 51 F I A R Y X C Y R F S S S F S F Y F F P T R E K H I V L Y N K N K K H D L K D P X T E Q S S D S P P 100
101 K S S G S C S R S S F S S P C P V P E H E I R W A V F P E I E N T Y L S F A Y D L A K N A I N F T R 150
151 S N E S P T I I V M C R K D N N X I D X S I P X I S P L F F Y H I R M H Q I P L V Y P Q F X F C S H 200
201 T F M H F H G H Q Y T E S F I P F L S S L F S H C Q N S P V L A C S R S H V G L P L S A H L P F N W 250
251 R L L H S L P T Y R L T G G S L S K R A L R S K R F I K T R P I P I R X I P G T K I F F F F F S A R 300
301 S P G C M D S Y T N T T V Q N K C R K Q X P N H P H C E P L T R T H P H F V S C H V S R Q S G K E R 350
351 K N X S L T C I L E I I P A T C P S I S L I N S C F S L E I X S F K R R H C I R P P T A D K K S D Q 400
401 V X X R R
```

Key	Helix	Sheet	Disordered	Dompred Boundary	DomSSEA Boundary
Annotations	M	L	I	A	D



```
Conf:

Pred:

Pred: CCCCCCCCCEEEEEEEEECCEEEEECCCEEEEEECC
  AA: LDSLSGQNLKFNYIYIYLNGKCLVQQGKLFHTCSVN
               10        20        30

Conf:

Pred:

Pred: CCCCHHHHHHHHHCCCCCCCCCCEEEEEECCCCEEE
  AA: PIEXLIXWFKFIARYXCYRFSSSFSFYFFPTREKHI
               50        60        70

Conf:

Pred:

Pred: CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
  AA: KNKKHDLKDPXTEQSSDSPPKSSGSCSRSSFSSPCP
               90        100       110

Conf:

Pred:

Pred: CEEEEECCCCCCCHHHHHHHHHHHHHCCCCCCCCCE
  AA: EIRWAVFPEIENTYLSFAYDLAKNAINFTRSNESPT
               130       140       150

Conf:

Pred:

Pred: ECCCCCCCCCCCCCCCCCEEEEEECCCCCCCCCCCC
  AA: CRKDNNXIDXSIPXISPLFFYHIRMHQIPLVYPQFX
               170       180       190

Conf:

Pred:

Pred: CCEEECCCCCCCHHHHHHHHHHHHCCCCCEEEEECC
  AA: TFMHFHGHQYTESFIPFLSSLFSHCQNSPVLACSRS
               210       220       230

Conf:

Pred:

Pred: CCCCCCCCCEEECCCCCCEEECCCCCHHHHHHHCCC
  AA: PLSAHLPFNWRLLHSLPTYRLTGGSLSKRALRSKRF
               250       260       270

Conf:

Pred:

Pred: CCCCCCCCCCEEEEEEEECCCCCCCCCCCCHHHHHHHHCC
  AA: PIPIRXIPGTKIFFFFFSARSPGCMDSYTNTTVQNKCRKQ
               290       300       310       320

Conf:

Pred:

Pred: CCCCCCCCCCCCCCCCEEEEEEECCCCCCCCCCCCEEEEE
  AA: XPNHPHCEPLTRTHPHFVSCHVSRQSGKERKNXSLTCILE
               330       340       350       360

Conf:

Pred:

Pred: EECCCCCCHHHHCCCCCCCCCCCCCCCCCCCCCCCCCCCC
  AA: IIPATCPSISLINSCFSLEIXSFKRRHCIRPPTADKKSDQ
               370       380       390       400

Conf:

Pred:

Pred: CCCCC
  AA: VXXRR
```

Legend:

- helix

- strand

- coil

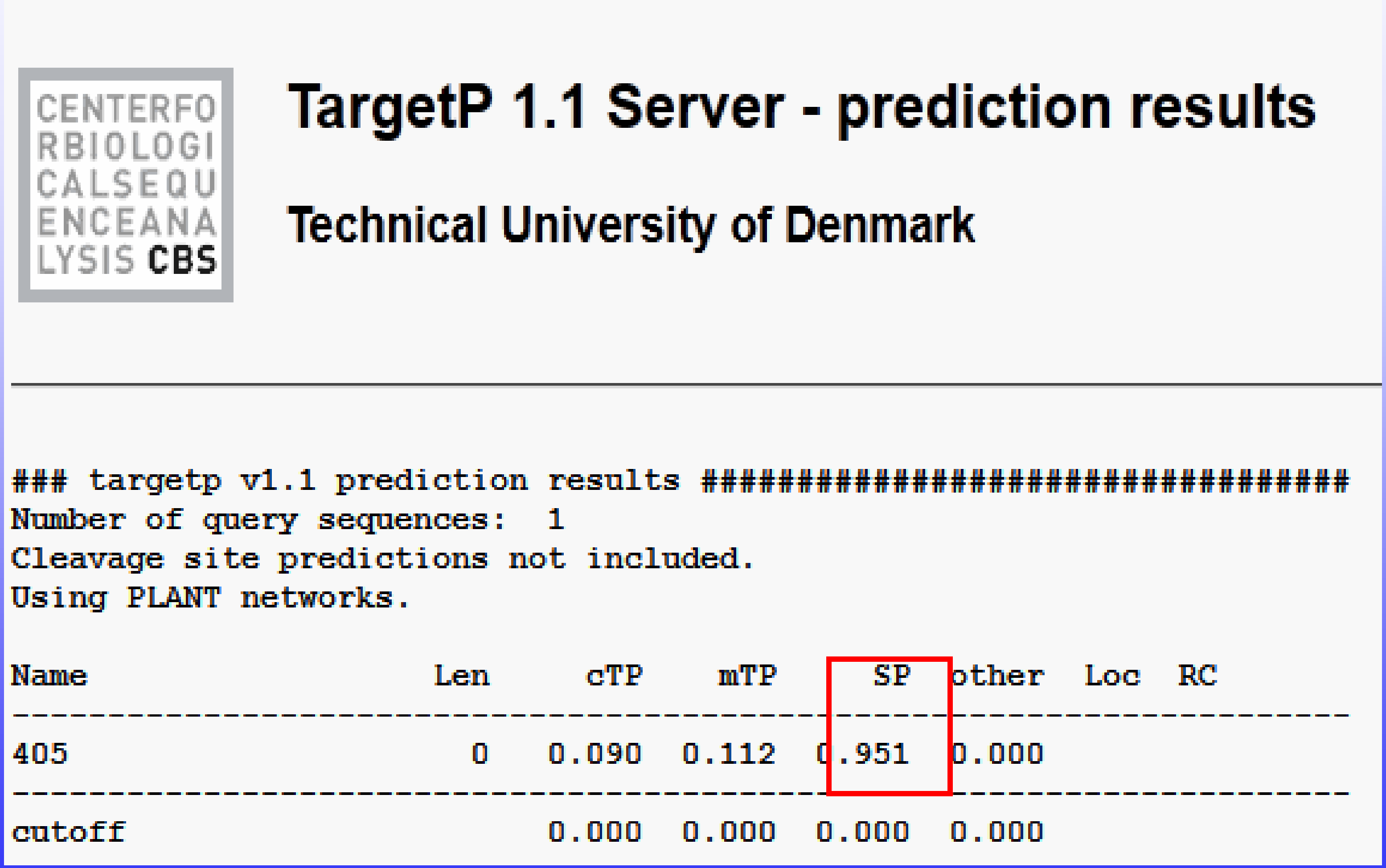
Conf: - confidence of prediction

\- +

Pred: predicted secondary structure

AA: target sequence

(5)亚细胞定位

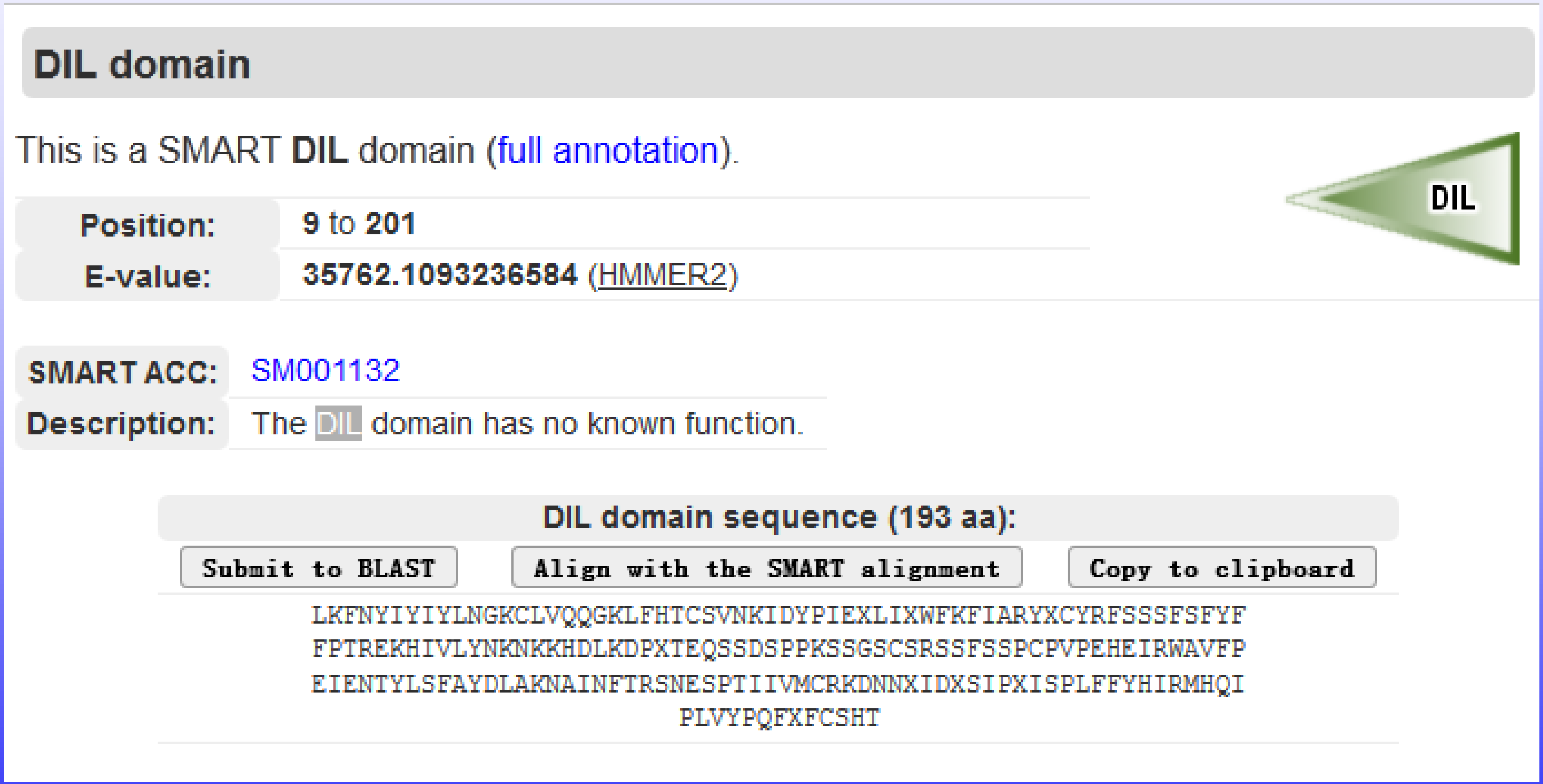


TargetP 1.1 Server - prediction results

Technical University of Denmark

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

Name                  Len     cTP    mTP     SP  other  Loc  RC
----------------------------------------------------------------------
405                     0   0.090  0.112  0.951  0.000
----------------------------------------------------------------------
cutoff                      0.000  0.000  0.000  0.000
```

(6)SMART预测结构域



DIL domain

This is a SMART **DIL** domain (full annotation).

Position: 9 to 201

E-value: 35762.1093236584 (HMMER2)

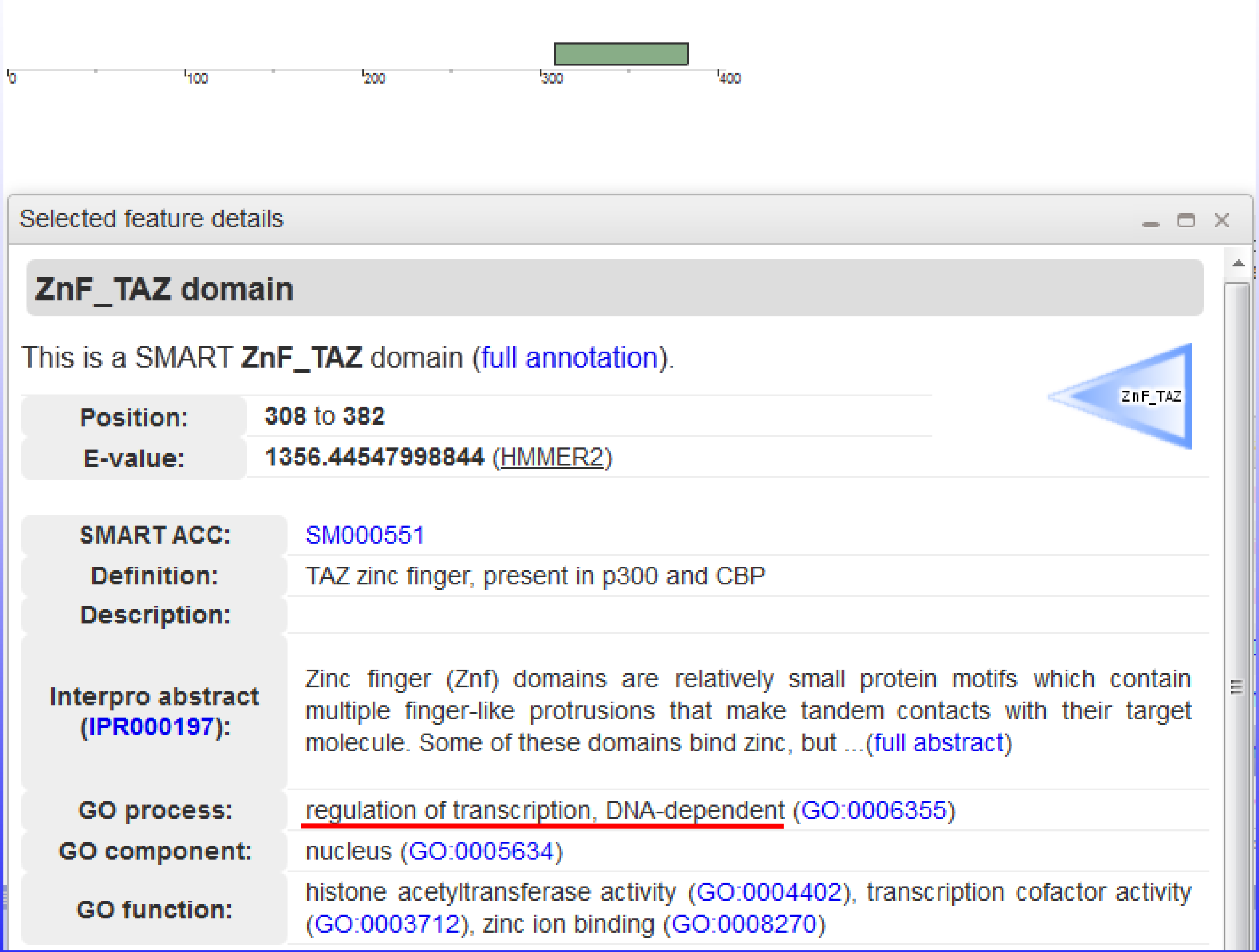
SMART ACC: SM001132

Description: The DIL domain has no known function.

DIL domain sequence (193 aa):

```
LKFNYIYIYLNGKCLVQQGKLFHTCSVNKIDYPIEXLIXWFKFIARYXCYRFSSSFSFYF
FPTREKHIVLYNKNKKHDLKDPXTEQSSDSPPKSSGSCSRSSFSSPCPVPEHEIRWAVFP
EIENTYLSFAYDLAKNAINFTRSNESPTIIVMCRKDNNXIDXSIPXISPLFFYHIRMHQI
PLVYPQFXFCSHT
```

SMART预测结构域



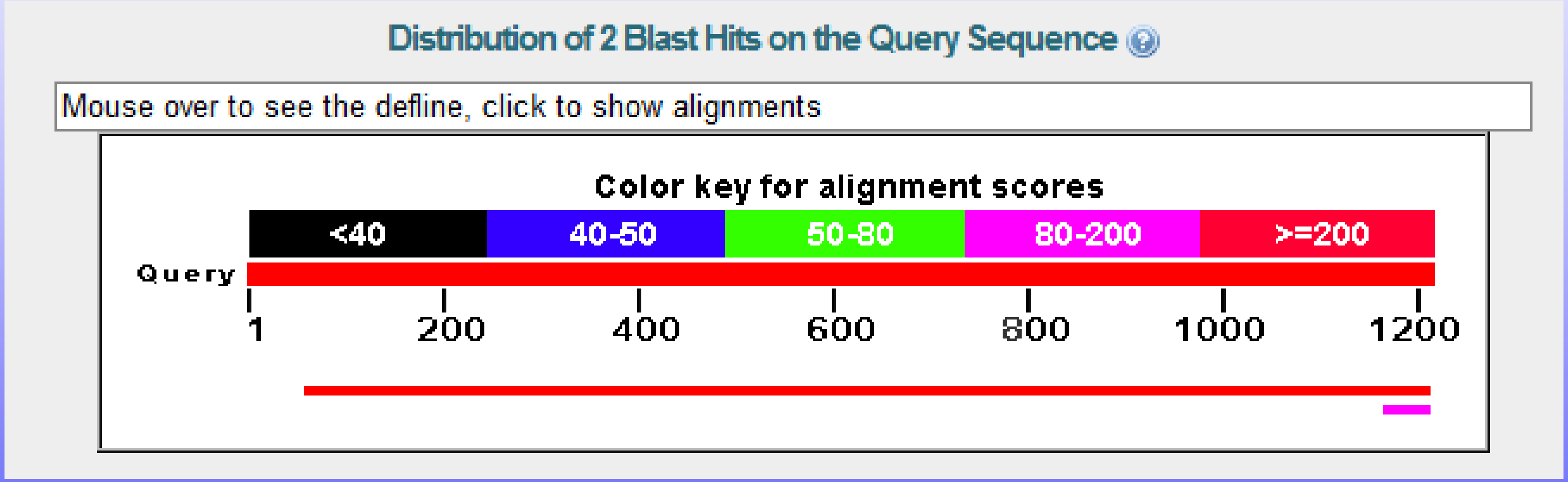
ZnF_TAZ domain

This is a SMART **ZnF_TAZ** domain (full annotation).

Position:	**308** to **382**
E-value:	**1356.44547998844** (HMMER2)
SMART ACC:	SM000551
Definition:	TAZ zinc finger, present in p300 and CBP
Description:	
Interpro abstract (IPR000197):	Zinc finger (Znf) domains are relatively small protein motifs which contain multiple finger-like protrusions that make tandem contacts with their target molecule. Some of these domains bind zinc, but ...(full abstract)
GO process:	regulation of transcription, DNA-dependent (GO:0006355)
GO component:	nucleus (GO:0005634)
GO function:	histone acetyltransferase activity (GO:0004402), transcription cofactor activity (GO:0003712), zinc ion binding (GO:0008270)

功能预测

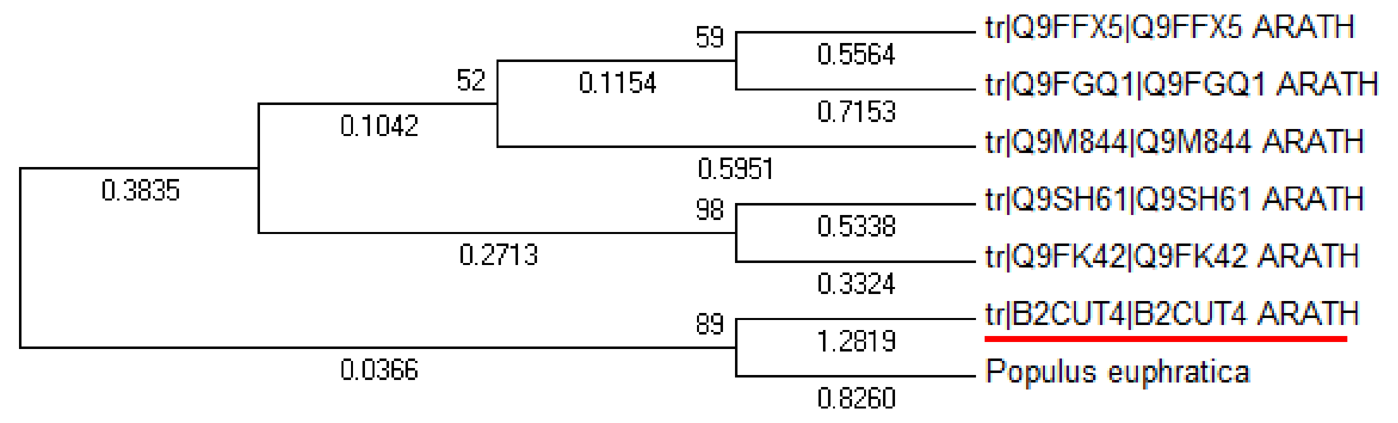
(1)在NCBI上进行蛋白质序列的blast分析

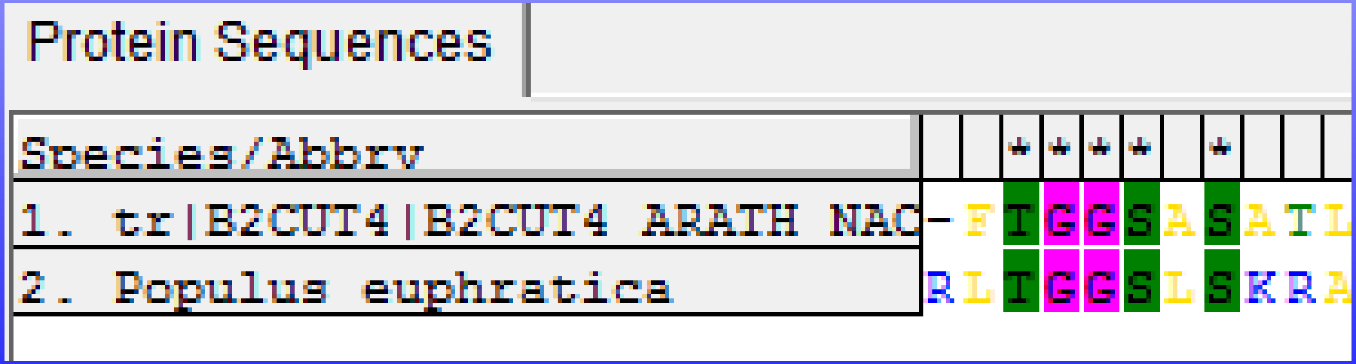


Distribution of 2 Blast Hits on the Query Sequence

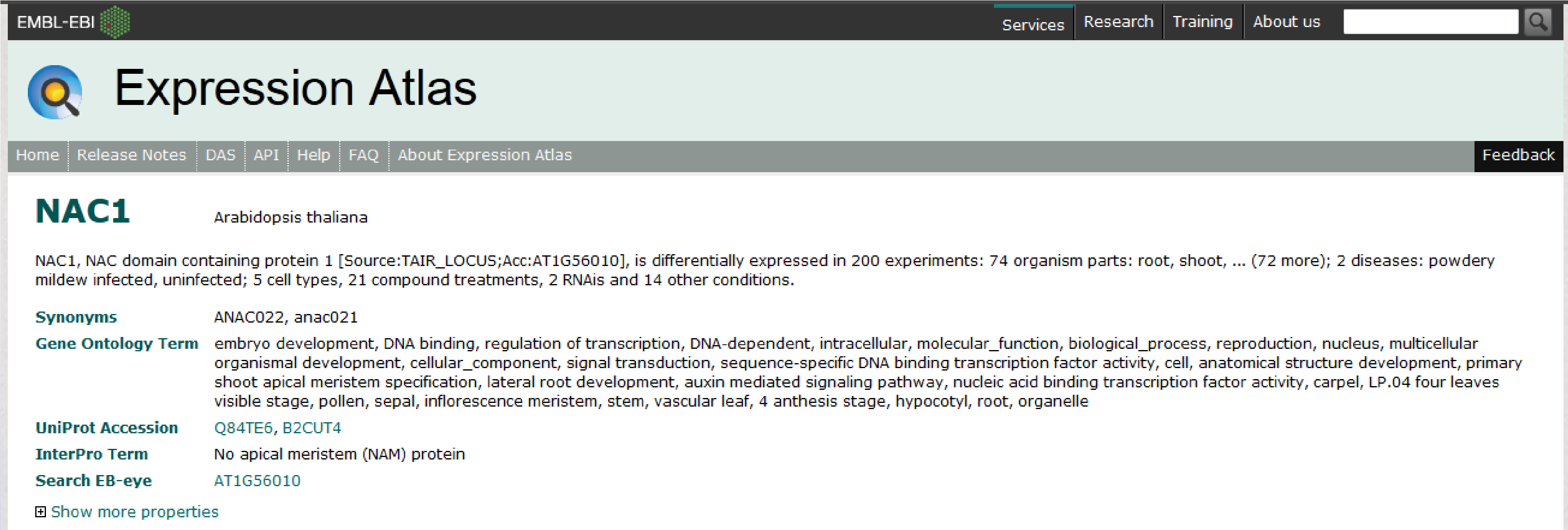
Mouse over to see the defline, click to show alignments

(2)比对105条拟南芥nac基因编码的蛋白质序列及45条大叶杨nac基因编码的蛋白质序列，构建进化树得到亲缘关系较为接近的一条序列——拟南芥nac1蛋白




Protein Sequences

Species/Abbrv	
	* * * * *
1. tr|B2CUT4|B2CUT4 ARATH NAC	-FIGGSASATL
2. Populus euphratica	RLIGGSLSKRA

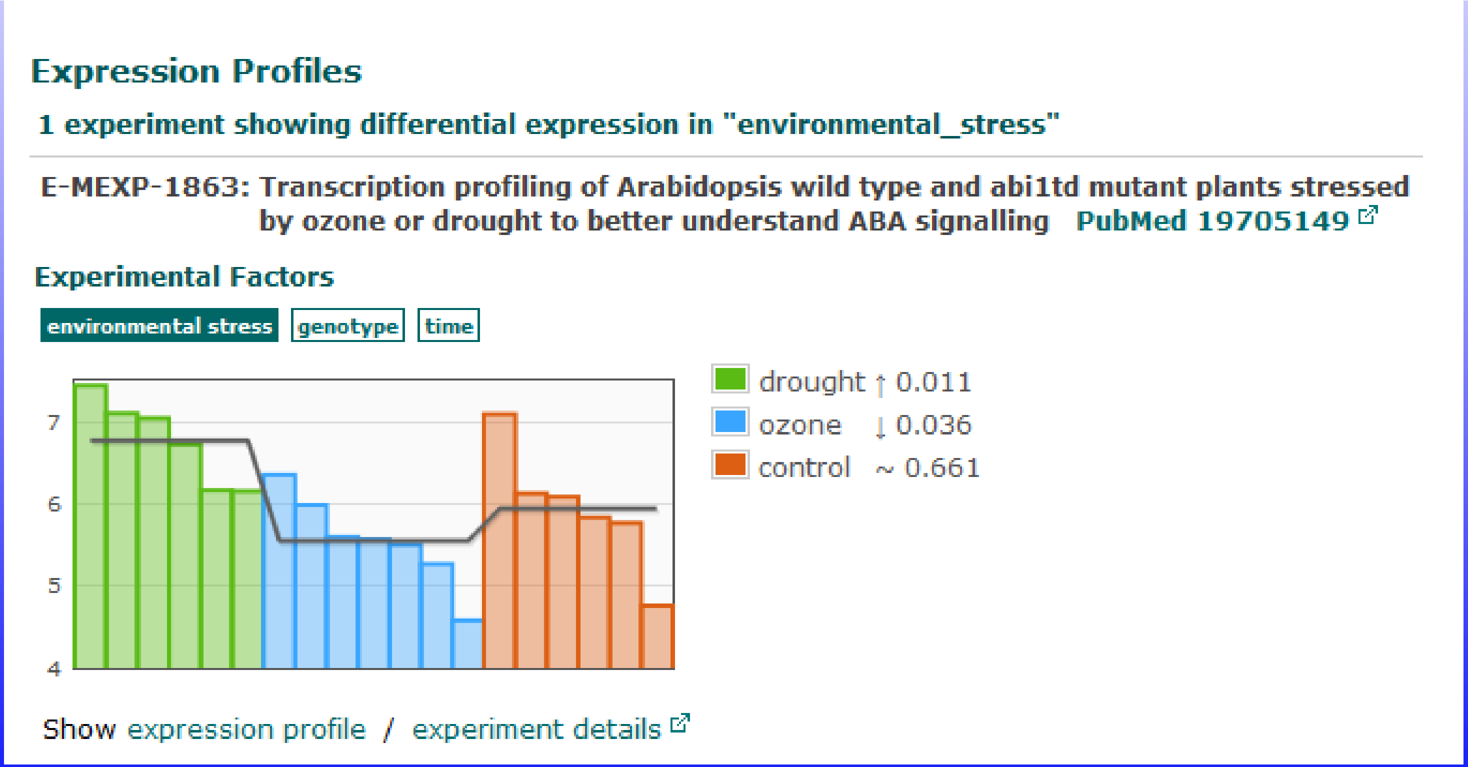


NAC1 Arabidopsis thaliana

NAC1, NAC domain containing protein 1 [Source:TAIR_LOCUS;Acc:AT1G56010], is differentially expressed in 200 experiments: 74 organism parts: root, shoot, ... (72 more); 2 diseases: powdery mildew infected, uninfected; 5 cell types, 21 compound treatments, 2 RNAis and 14 other conditions.

Synonyms	ANAC022, anac021
Gene Ontology Term	embryo development, DNA binding, regulation of transcription, DNA-dependent, intracellular, molecular_function, biological_process, reproduction, nucleus, multicellular organismal development, cellular_component, signal transduction, sequence-specific DNA binding transcription factor activity, cell, anatomical structure development, primary shoot apical meristem specification, lateral root development, auxin mediated signaling pathway, nucleic acid binding transcription factor activity, carpel, LP.04 four leaves visible stage, pollen, sepal, inflorescence meristem, stem, vascular leaf, 4 anthesis stage, hypocotyl, root, organelle
UniProt Accession	Q84TE6, B2CUT4
InterPro Term	No apical meristem (NAM) protein
Search EB-eye	AT1G56010

⊞ Show more properties



Expression Profiles

1 experiment showing differential expression in "environmental_stress"

E-MEXP-1863: Transcription profiling of Arabidopsis wild type and abi1td mutant plants stressed by ozone or drought to better understand ABA signalling PubMed 19705149

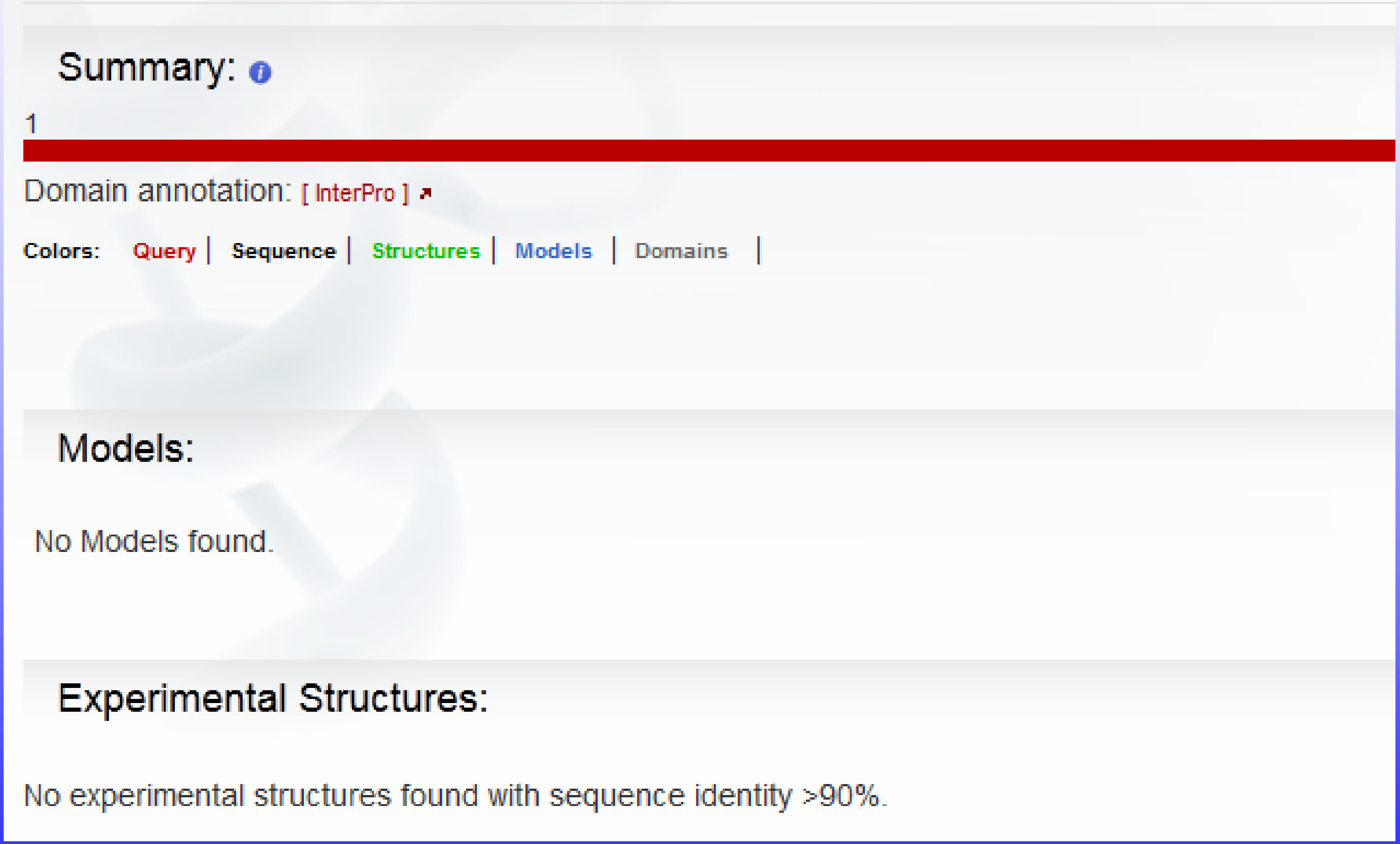
Experimental Factors

Show expression profile / experiment details

(3)拟南芥Nac1蛋白的基本信息

- NAC1（Unreviewed）
- Arabidopsis thaliana
- NAC1, NAC domain containing protein 1 [Source:TAIR_LOCUS;Acc:AT1G56010], is differentially expressed in 200 experiments: 74 organism parts: root, shoot, ... (72 more); 2 diseases: powdery mildew infected, uninfected; 5 cell types, 21 compound treatments, 2 RNAis and 14 other conditions.
- Synonyms ANAC022, anac021
- Gene Ontology Term embryo development, DNA binding, regulation of transcription, DNA-dependent, intracellular, molecular_function, biological_process, reproduction, nucleus, multicellular organismal development, cellular_component, signal transduction, sequence-specific DNA binding transcription factor activity, cell, anatomical structure development, primary shoot apical meristem specification, lateral root development, auxin mediated signaling pathway, nucleic acid binding transcription factor activity, carpel, LP.04 four leaves visible stage, pollen, sepal, inflorescence meristem, stem, vascular leaf, 4 anthesis stage, hypocotyl, root, organelle
- UniProt Accession Q84TE6, B2CUT4
- InterPro Term No apical meristem (NAM) protein
- Search EB-eye AT1G56010

(4)未能找到与之类似的空间结构



Summary:

1

Domain annotation: [InterPro]

Colors: Query | Sequence | Structures | Models | Domains |

Models:

No Models found.

Experimental Structures:

No experimental structures found with sequence identity >90%.

总结

- 预测胡杨*NAC*基因编码的蛋白无信号肽，但为跨膜蛋白。
- 其跨膜方式可能为孔隙螺旋而不是通常的跨膜螺旋。
- 预测该蛋白为外分泌蛋白，符合*NAC*转录因子的特点。
- 预测蛋白的C端具有锌脂结构域，可作用于DNA并进行转录调控。
- 预测该蛋白的功能和拟南芥*NAC1*蛋白相似，主要作用部位在植物中的根、叶中，行使与抗旱方面有关功能。

问题与不足

- 预测为分泌蛋白但没有预测到信号肽。
- 与胡杨nac基因相关的信息极少。
- 尚未进行同源模建，未能得到直观的蛋白质三级结构空间模型。

致谢

- 感谢罗老师的悉心指导！
- 感谢小组成员的共同努力 ！
- 请老师和同学们批评指正！