茶树查尔酮异构酶基因mRNA及蛋白质序列的生物信息学分析

小组：G10

报告人:王丹

组员:周艳华 胡娟
李丽田

2013/6/21

一　研究背景

二　序列比对和系统发育分析

三　基因mRNA全长序列分析

四　蛋白质序列分析

研究背景

花卉颜色物质以类黄酮色素中主要色素——花色素苷对花色作用最重要

花色素合成生物途径中查尔酮合酶含量直接影响着花色素在花瓣中的含量高低，改变了花的颜色。

科学家们通过基因工程技术改变了查尔酮的含量，得到了颜色各异，绚烂多姿的花朵。

矮牵牛因查尔酮合酶基因改变而颜色各异

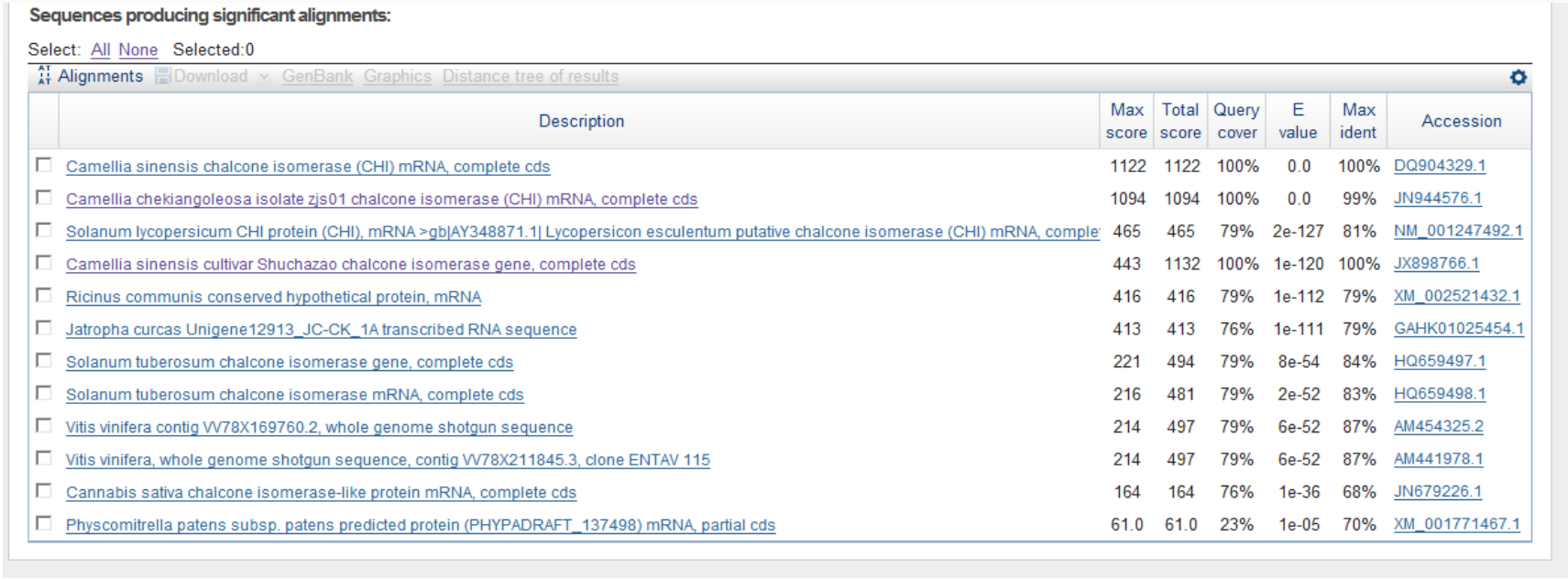
- 茶叶次生代谢产物儿茶素因较强的抗氧化作用越来越受到人们的重视。

- 儿茶素合成途径中的第一个中间产物是查尔酮，由查尔酮合成酶 (CHS)催化。

- 查尔酮异构酶 (CHI)是黄酮类代谢途径中的早期酶，也是增加黄酮醇产物的关键酶之一，它催化查尔酮异构化为黄烷酮柑橘素。

- 黄烷酮柑橘素是类黄酮合成途径中第一个稳定中间产物，再经**F3H**、**DFR**等的催化作用可生成一系列常见的次生代谢产物，从而形成其他黄酮类物质。
- 从**Genbank**搜索**DQ904329**登录号，然后对茶树查尔酮异构酶基因（**CHI**）**mRNA**全长序列进行生物信息学分析

序列比对和系统发育分析



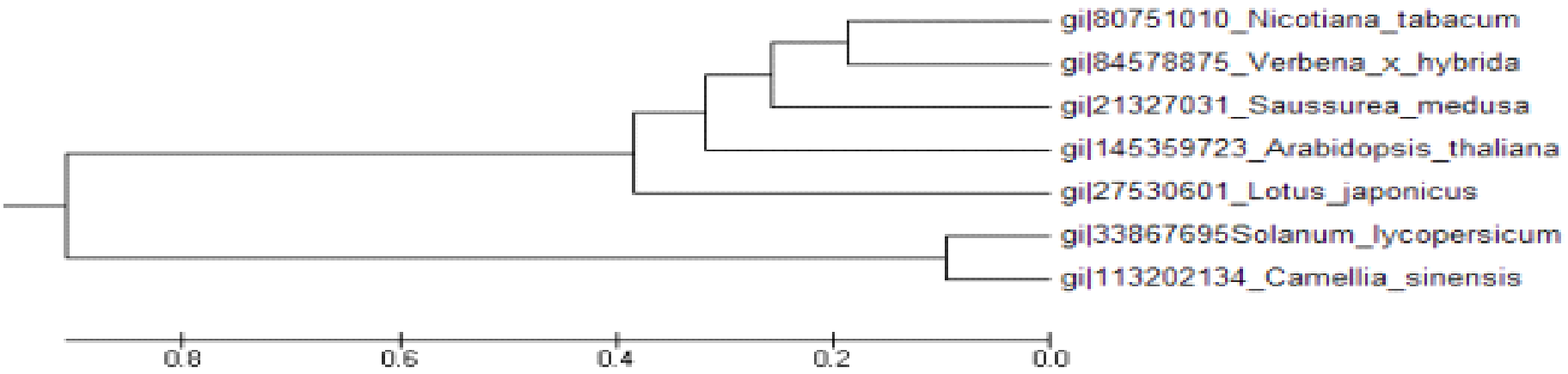
Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenBank Graphics Distance tree of results

Description	Max score	Total score	Query cover	E value	Max ident	Accession
Camellia sinensis chalcone isomerase (CHI) mRNA, complete cds	1122	1122	100%	0.0	100%	DQ904329.1
Camellia chekiangoleosa isolate zjs01 chalcone isomerase (CHI) mRNA, complete cds	1094	1094	100%	0.0	99%	JN944576.1
Solanum lycopersicum CHI protein (CHI), mRNA >gb|AY348871.1| Lycopersicon esculentum putative chalcone isomerase (CHI) mRNA, comple	465	465	79%	2e-127	81%	NM_001247492.1
Camellia sinensis cultivar Shuchazao chalcone isomerase gene, complete cds	443	1132	100%	1e-120	100%	JX898766.1
Ricinus communis conserved hypothetical protein, mRNA	416	416	79%	1e-112	79%	XM_002521432.1
Jatropha curcas Unigene12913_JC-CK_1A transcribed RNA sequence	413	413	76%	1e-111	79%	GAHK01025454.1
Solanum tuberosum chalcone isomerase gene, complete cds	221	494	79%	8e-54	84%	HQ659497.1
Solanum tuberosum chalcone isomerase mRNA, complete cds	216	481	79%	2e-52	83%	HQ659498.1
Vitis vinifera contig VV78X169760.2, whole genome shotgun sequence	214	497	79%	6e-52	87%	AM454325.2
Vitis vinifera, whole genome shotgun sequence, contig VV78X211845.3, clone ENTAV 115	214	497	79%	6e-52	87%	AM441978.1
Cannabis sativa chalcone isomerase-like protein mRNA, complete cds	164	164	76%	1e-36	68%	JN679226.1
Physcomitrella patens subsp. patens predicted protein (PHYPADRAFT_137498) mRNA, partial cds	61.0	61.0	23%	1e-05	70%	XM_001771467.1

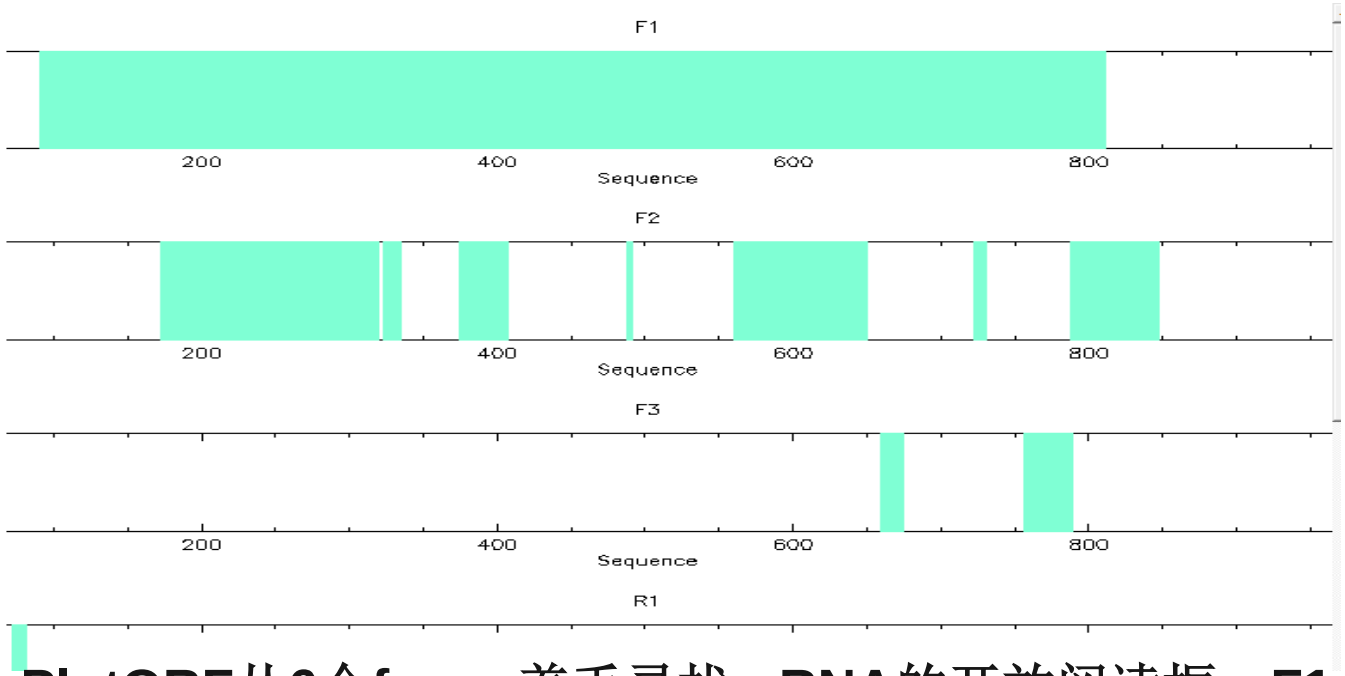
用blastn进行同源比对，参数经多次调试为期望值：1e-1比较理想。空位罚分0.4，茶树的类黄酮异构酶基因与同属的红花油茶（Camellia chekiangoleosa）隔离种zjs01中的查尔酮异构酶基因相似性达到了99%。与番茄中的 CHI protein相似性达到了81%。与马铃薯的查尔酮异构酶也都具有较高的相似性。



应用MAGA软件将茶树CHI编码的氨基酸序列及从GenBank中获取的其他植物CHI氨基酸序列进行系统树分析，发现茶树与番茄最先聚类合并，而与其它物种，如烟草、拟南芥等在进化上的亲缘关系则较远。

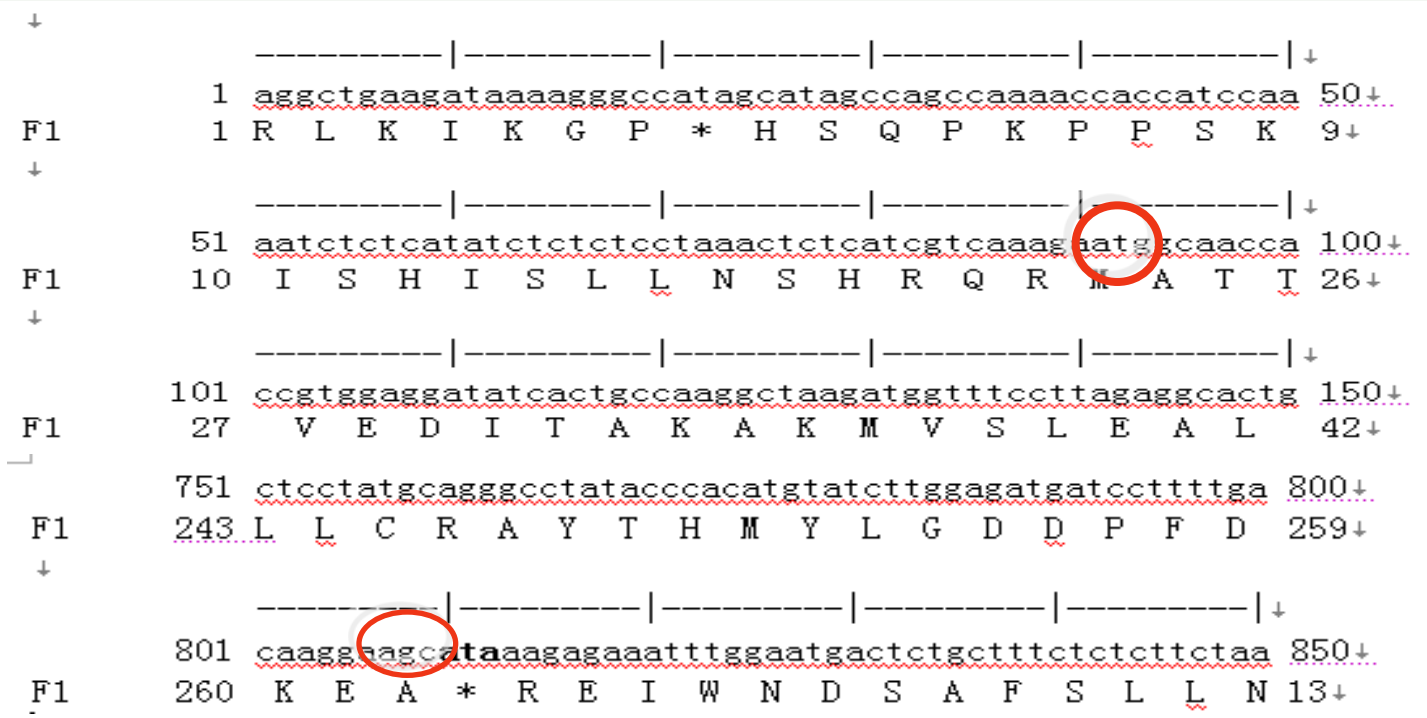
基因mRNA全长序列分析

PLOTORF



PlotORF从6个frame着手寻找mRNA的开放阅读框。F1是最有可能的读码框，可看出正向从第一个碱基（F1）开始读起才能在91-810左右读到最完整的可连续编码蛋白质序列的读码框。

SHOWORF of DQ904329 from 1 to 1002



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90bp到810bp之间的开放阅读框最长，最可信。1-90虽然也能连续编码氨基酸，但不是有起始密码子开始的，810之前的三个密码子是终止密码子，因此此后的序列不属于编码区。

sixpack

>DQ904329_1_ORF2 Translation of DQ904329 in frame 1, ORF 2, threshold 100, 262aa
HSQPKPPSKISHISLLNSHRQRMATTVEDITAKAKMVSLEALTPK
EEKVNGPESNKIADGEMGKADEEPQMGKKDDVPVETEPKTGV
SFPIKLDDGKQLNAVGLRKKSVLGIGIKIYGFGIYADNETLKDLLR
TKIGKAPTKPTKEMYQLVIDSDVGMLVRLVMVFSNLTMSMVRK
NFDEVLGASIKKLTGGKNDELTKKIMGEASDDIKLTCGSIIEISRL
PGYILQTKVMDEVVSKVESELLCRAYTHMYLGDDPFDKEA

SixPack程序：用于寻找开放阅读框，以多肽序列呈现。需要将最后一条参数（ORF start with an M）设置为Yes，运行后有多个结果，其中一条结果最可信（如图）有262个氨基酸残基。

getorf

>DQ904329_5 [91 - 810] Camellia sinensis chalcone isomerase (CHI) mRNA, complete cds.

atggcaaccaccgtggaggatatcactgccaaggctaagatggtttccttagaggcactgacacctaaa
gaggagaaagtgaatggccctgaatcaaataagattgccgatggtgagatggggaaagctgatgaag
agccacaaatgggcaagaaagatgacgtgccggttgagactgaacccaagaccggggtctcctttcc
gattaagttggatgatgggaagcagttgaatgcggttgggttgaggaaaaaaagcgtgcttggcatcgg
catcaaaatctatggcttcggaatatatgcagataatgagacactgaaagatcttctgaggacaaaaatt
gggaaagcaccaacaaaacctaccaaggaaatgtaccaactggtaattgacagtgatgtaggaatgct
ggtgcgattggtaatggtgttttccaacctcacaatgagcatggtaagaaagaactttgatgaagttcttg
gagcatctatcaaaaagctcactggtggaaagaatgacgagctcacaaagaagattatgggtgaagct
tcagatgacataaagctgacatgtggttcaataattgagatttctcggcttccaggatacattctccaaaca
aaagtgatggatgaagttgtgagcaaggttgaaagtgaactcctatgcagggcctatacccacatgtatc
ttggagatgatccttttgacaaggaagca

getorf(v6.0.1)：该软件用于寻找和提取开放阅读框。将输出类型（**Type of output**）设置为“在起始密码子和终止密码子之间翻译”（**Nucleic sequences between START and STOP codon**），其它参数不变，输出的是可能的编码区结果。编码区长度不同，选取最长的那条（如图），该结果有**720**个碱基。

getorf

- >DQ904329_5 [91 - 810] Camellia sinensis chalcone isomerase (CHI) mRNA, complete cds.

  MATTVEDITAKAKVSLEALTPKEEKVNGPESNKIADGEMGKADEEPQM
  GKKDDVPVETEPKTGVSFPIKLDDGKQLNAVGLRKKSVLGIGIKIYGFGI
  YADNETLKDLLRTKIGKAPTKPTKEMYQLVIDSDVGMLVRLVMVFSNLT
  MSMVRKNFDEVLGASIKKLTGGKNDELTKKIMGEASDDIKLTCGSIIEIS
  RLPGYILQTKVMDEVVSKVESELLCRAYTHMYLGDDPFDKEA

将输出类型（Type of output）设置为“在起始密码子和终止密码子之间翻译”（Translation of regions between START and STOP codon），其它参数不变。有24个结果，翻译的蛋白质长度不同，选取最长的那条（如图），共240氨基酸残基。

chips和cusp

密码子偏好性分析

 # CHIPS codon usage statistics

Nc = 58.300

- **CUSP**

Coding GC 44.03%

#1st letter GC 52.08%

#2nd letter GC 33.33%

#3rd letter GC 46.67%

Nc数在20到61之间，数字越高代表很低的密码子偏好程度，越低代表越强的密码子偏好程度。Nc=58.3表明CHI基因密码子使用的偏好度低。 GC含量大概是44%左右。

蛋白质序列分析

1、用weblab 中pepstats分析查尔酮异构酶氨基酸组成情况：

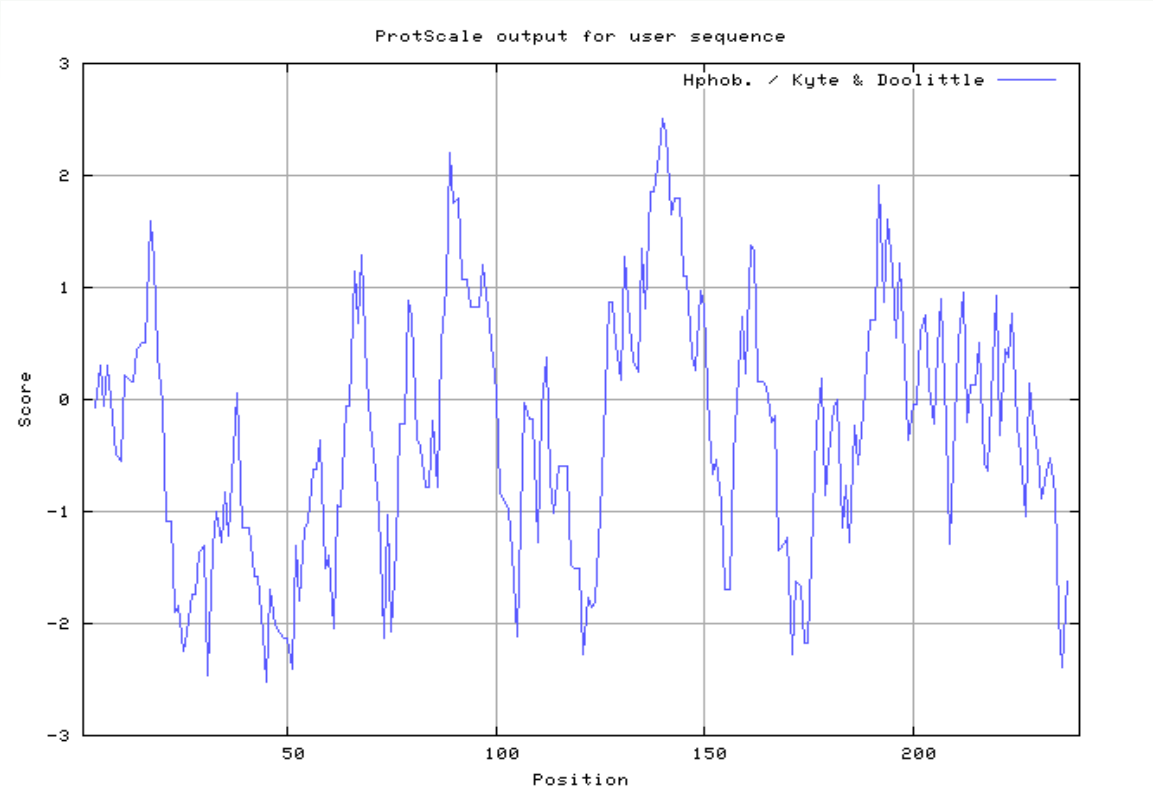
PEPSTATS of Q0G877_CAMSI from 1 to 240
Molecular weight = 26354.53 Residues = 240
Average Residue Weight = 109.811 Charge = -3.5
Isoelectric Point = 5.0902
A280 Molar Extinction Coefficient = 7680
A280 Extinction Coefficient 1mg/ml = 0.29
Improbability of expression in inclusion bodies = 0.527

Residue	Number	Mole%	DayhoffStat
A = Ala	13	5.417	0.630
B = Asx	0	0.000	0.000
C = Cys	2	0.833	0.287
D = Asp	19	7.917	1.439
E = Glu	20	8.333	1.389
F = Phe	5	2.083	0.579
G = Gly	20	8.333	0.992
H = His	1	0.417	0.208
I = Ile	16	6.667	1.481
J = ---	0	0.000	0.000
K = Lys	29	12.083	1.831
L = Leu	22	9.167	1.239
M = Met	12	5.000	2.941
N = Asn	7	2.917	0.678
O = ---	0	0.000	0.000
P = Pro	10	4.167	0.801
Q = Gln	4	1.667	0.427
R = Arg	6	2.500	0.510
S = Ser	13	5.417	0.774
T = Thr	16	6.667	1.093
U = ---	0	0.000	0.000
V = Val	19	7.917	1.199
W = Trp	0	0.000	0.000
X = Xaa	0	0.000	0.000
Y = Tyr	6	2.500	0.735
Z = Glx	0	0.000	0.000

Property	Residues	Number	Mole%
Tiny	(A+C+G+S+T)	64	26.667
Small	(A+B+C+D+G+N+P+S+T+V)	119	49.583
Aliphatic	(A+I+L+V)	70	29.167
Aromatic	(F+H+W+Y)	12	5.000
Non-polar	(A+C+F+G+I+L+M+P+V+W+Y)	125	52.083
Polar	(D+E+H+K+N+Q+R+S+T+Z)	115	47.917
Charged	(B+D+E+H+K+R+Z)	75	31.250
Basic	(H+K+R)	36	15.000
Acidic	(B+D+E+Z)	39	16.250

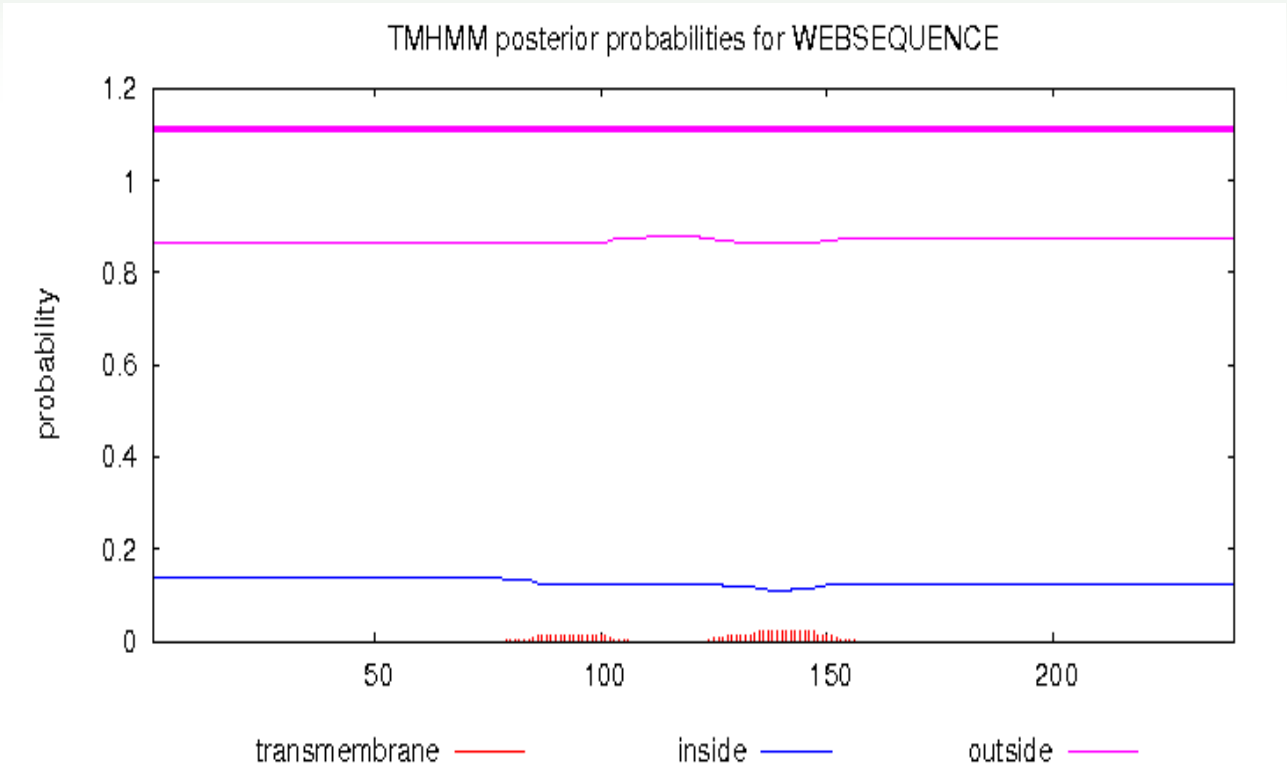
该酶的等电点为5.0902，含有的极性氨基酸占47.917%，非极性占52.083%，带电的氨基酸有75个，占31.250%,总带的电荷量为-3.5。

2、用ExPASy中protscale对该酶进行疏水性和亲水性分析



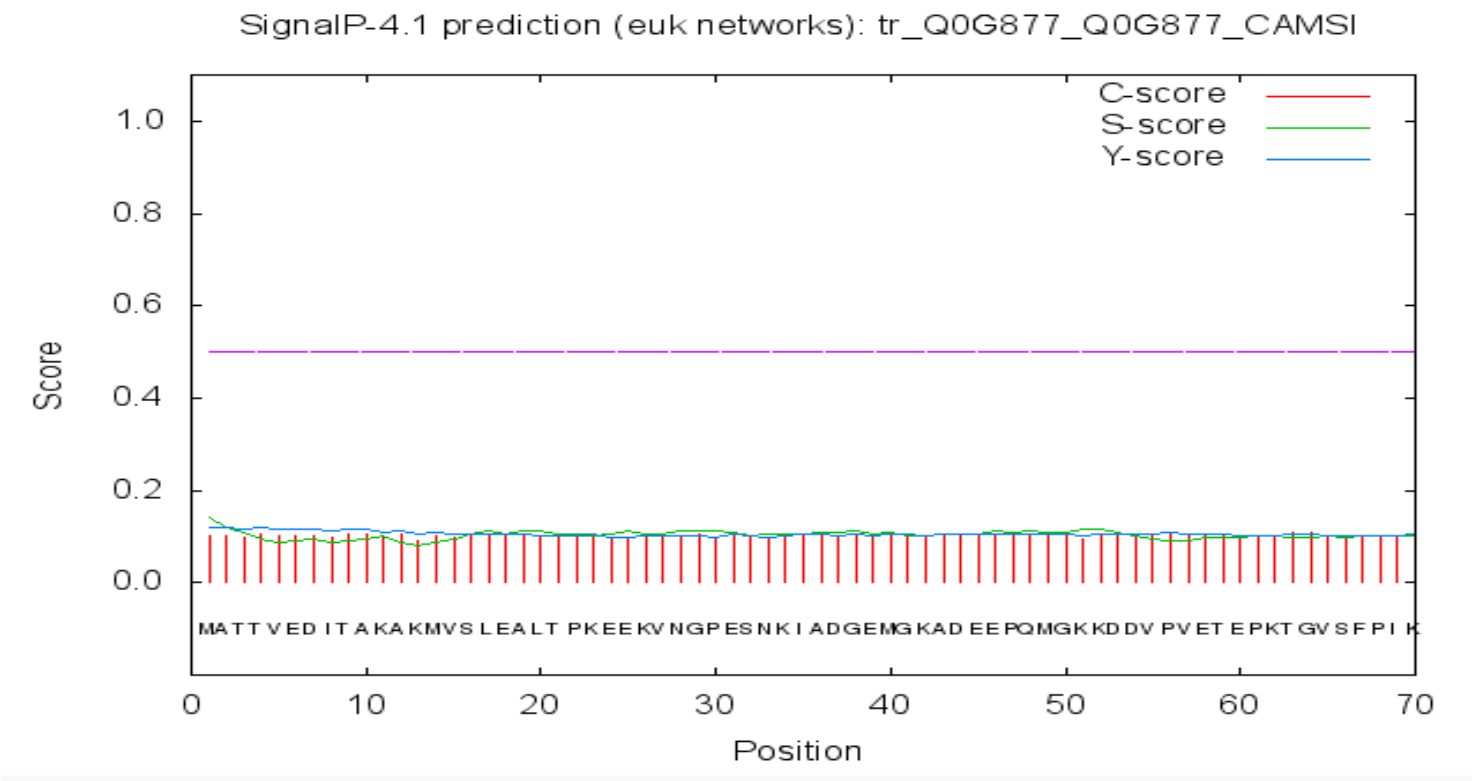
图中以0为界，正值表示疏水，负值表示亲水。

3、用ExPASy中TMHMM分析跨膜螺旋



由上图可以看出，该蛋白质无跨膜区，蛋白全部在膜外。曲线的纵坐标是概率，横坐标是序列，一共240个氨基酸，红色表示跨膜区，几乎都在5%概率下，蓝色inside即在膜内部，概率极低；相反紫色细线表示在膜外的概率有85%。因此1到240位氨基酸全部是outside。

4、用CBS中signal对该酶进行信号肽分析



图中可见，C 、S、Y三项得分均较低，因此该蛋白不含信号肽

5、用CBS 中targetP对该酶进行亚细胞定位

```
### 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
-----------------------------------------------------------------
tr_Q0G877_Q0G877_CAM  240   0.112  0.087  0.274  0.845   _    3
-----------------------------------------------------------------
cutoff                      0.730  0.860  0.430  0.840
```

loc中显示的是 “-”表示该蛋白在任何位置都有存在。

6、通过ExPASy中SOPMA对该酶二级结构进行预测

```
         10         20         30         40         50         60         70
          |          |          |          |          |          |          |
MATTVEDITAKAKMVSLEALTPKEEKVNGPESNKIADGEMGKADEEPQMGKKDDVPVETEPKTGVSFPIK
hccHHHHHHHHHeeeeettcHHHHHcccccccccchhhhcccccccccccccccceeeecttccccceee
LDDGKQLNAVGLRKKSVLGIGIKIYGFGIYADNETLKDLLRTKIGKAPTKPTKEMYQLVIDSDVGMLVRL
ecttheeeeetcccHHeeeeeeeeeeeeeeccHHHHHHHHHHHHcccccccchhhhhhhhhhhhhheeee
VMVFSNLTMSMVRKNFDEVLGASIKKLTGGKNDELTKKIMGEASDDIKLTCGSIIEISRLPGYILQTKVM
eeeettccHHHHHHHHHHHHHHHHHHHcccccHHHHHHHHHHHcttccccttceeeeeccttceeeeeet
DEVVSKVESELLCRAYTHMYLGDDPFDKEA
thhhhhhhhhhhhhhhhhhheecccccccccc

Sequence length :    240

SOPMA :
    Alpha helix       (Hh) :     97 is   40.42%
    3(10) helix       (Gg) :      0 is    0.00%
    Pi helix          (Ii) :      0 is    0.00%
    Beta bridge       (Bb) :      0 is    0.00%
    Extended strand   (Ee) :     52 is   21.67%
    Beta turn         (Tt) :     17 is    7.08%
    Bend region       (Ss) :      0 is    0.00%
    Random coil       (Cc) :     74 is   30.83%
    Ambigous states   (?)  :      0 is    0.00%
    Other states           :      0 is    0.00%
```

由图中可以看出该酶中的二级结构主要有α螺旋（40.42%），无规则卷曲（30.83%）和β折叠（7.08%）。

Thank You!