关于基因*Seita.5G258300*的生物信息学分析
Bioinformatics Analysis of the Gene *Seita.5G258300*

汇报人：田　宝

时　间：2018.6.23

成　员：3G01、3G02、3G03、3G11

主要内容

一、研究背景

二、基因结构分析

三、序列比对及分子进化树

四、基因功能预测分析

五、蛋白结构及功能的预测分析

六、CRISPR构建载体

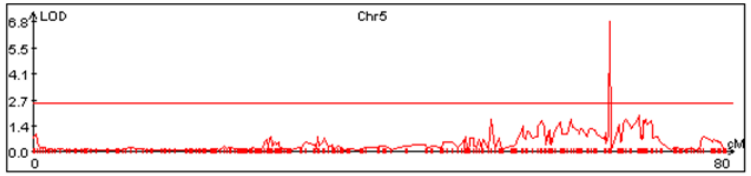
七、使用到的工具

一、研究背景

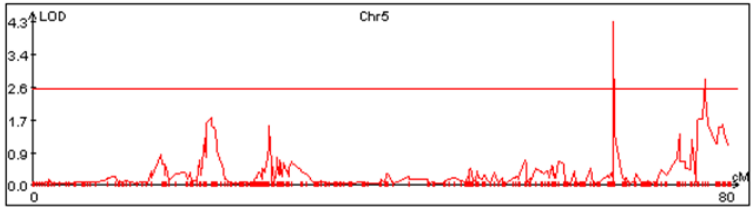
- 1、与课题相关（3G01A_田宝）：
- 谷子叶夹角QTL位点QLA_5A候选基因的鉴定及功能研究
- 以豫谷1×青24重组自交系（RI群体）作为材料，使用QTL定位的方法，定位到一段与控制叶夹角有关QTL区间，本研究对该位点的候选基因进鉴定及功能研究。

- 谷子：属于C_4光合途径的禾本科作物
- 二倍体自花授粉，基因组小（450Mb）
- 正式成为功能基因组研究的新模式作物

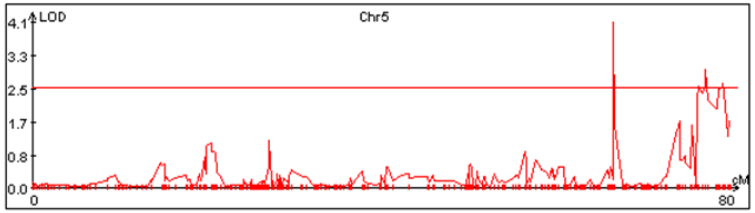
2、实验室前期准备：



2015 河南安阳



2016
黑龙江齐齐哈尔



2017
黑龙江齐齐哈尔

豫谷1×青24重组自交系

QTL定位到了110kb左右区间

Chr.5 32071638-32182208

候选基因：

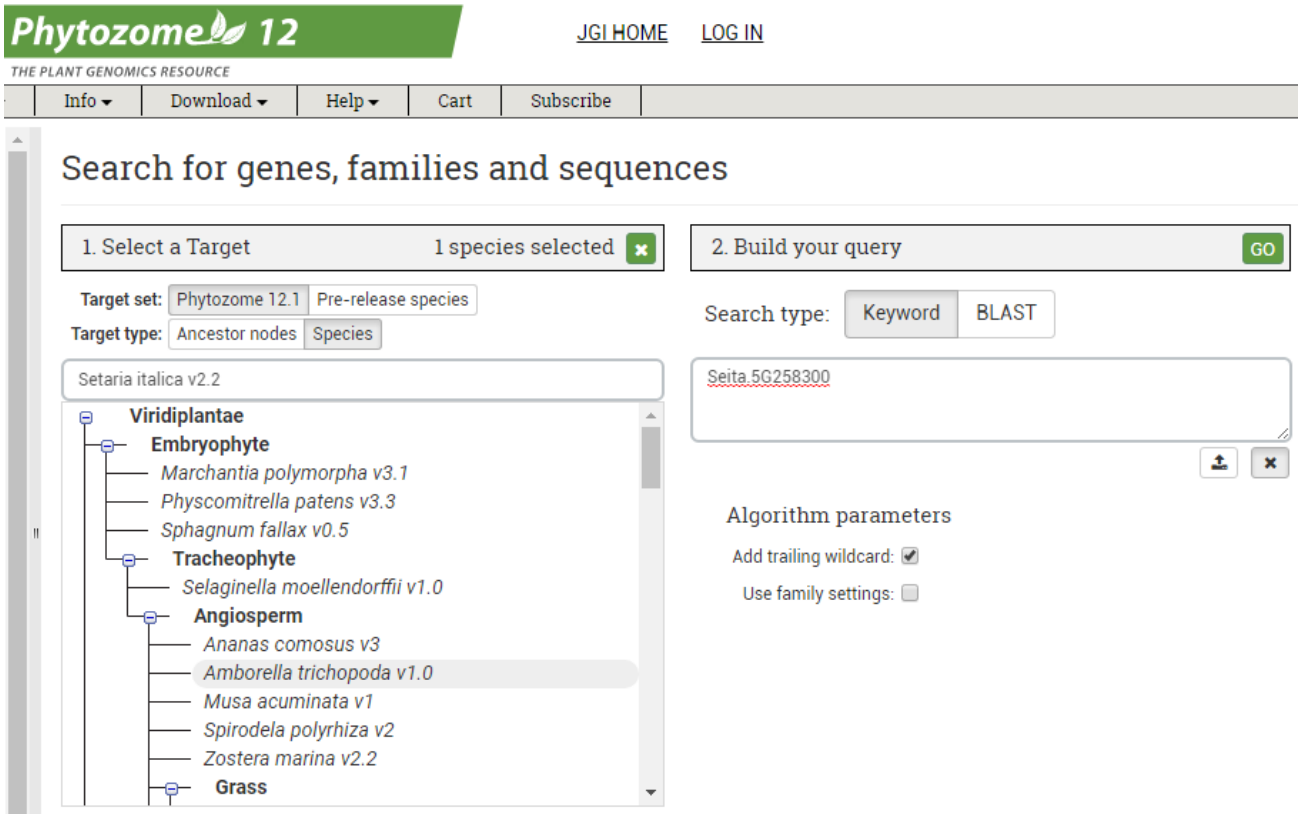
Seita.5G258000
Seita.5G258100
Seita.5G258200
Seita.5G258300
Seita.5G258400
Seita.5G258400
Seita.5G258500
Seita.5G258600
Seita.5G258700
Seita.5G258800
Seita.5G258900

scaffold_5： 32098582（G → A）

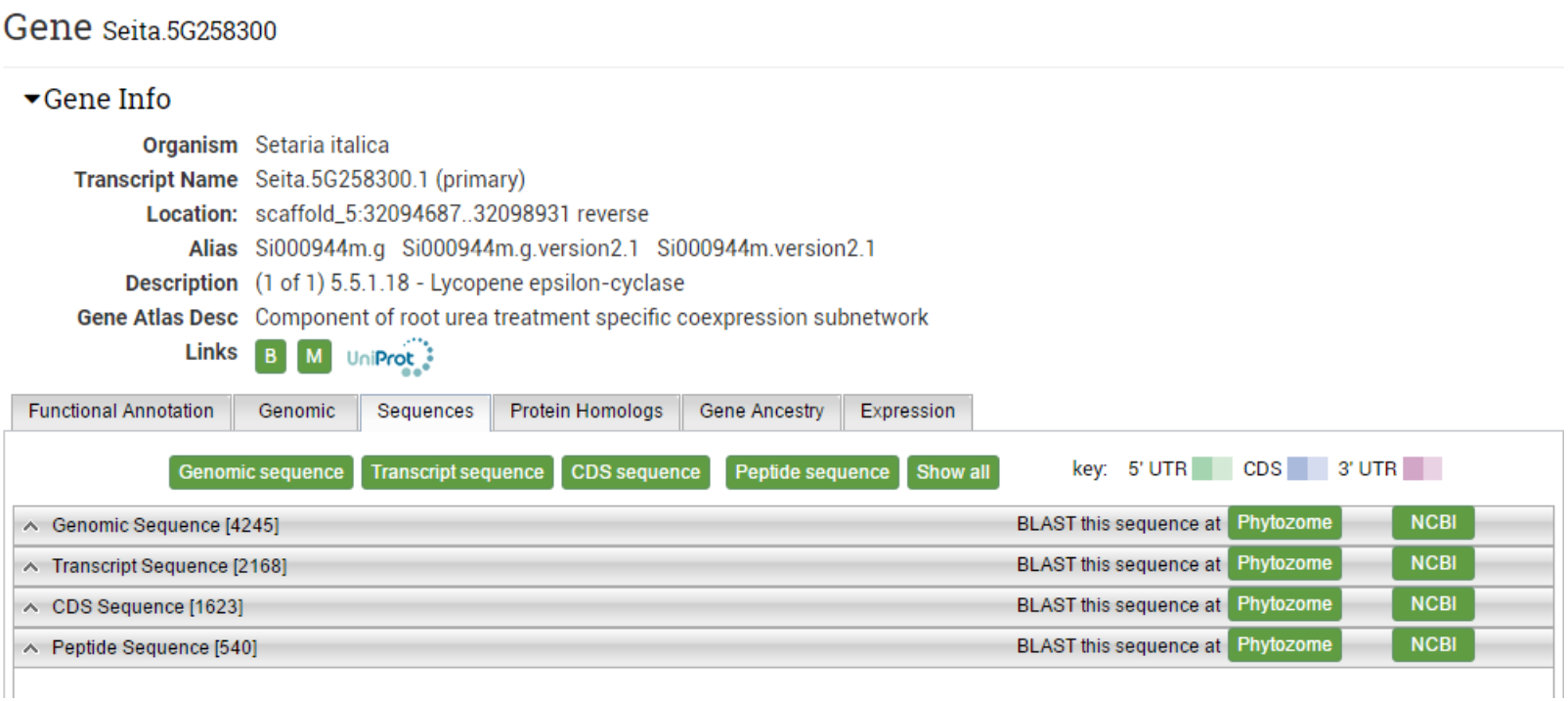
NON_SYNONYMOUS_CODING (Seita.5G258300)

二、基因结构分析

- 1、在Phytozome中查找*Seita.5G258300*基因序列

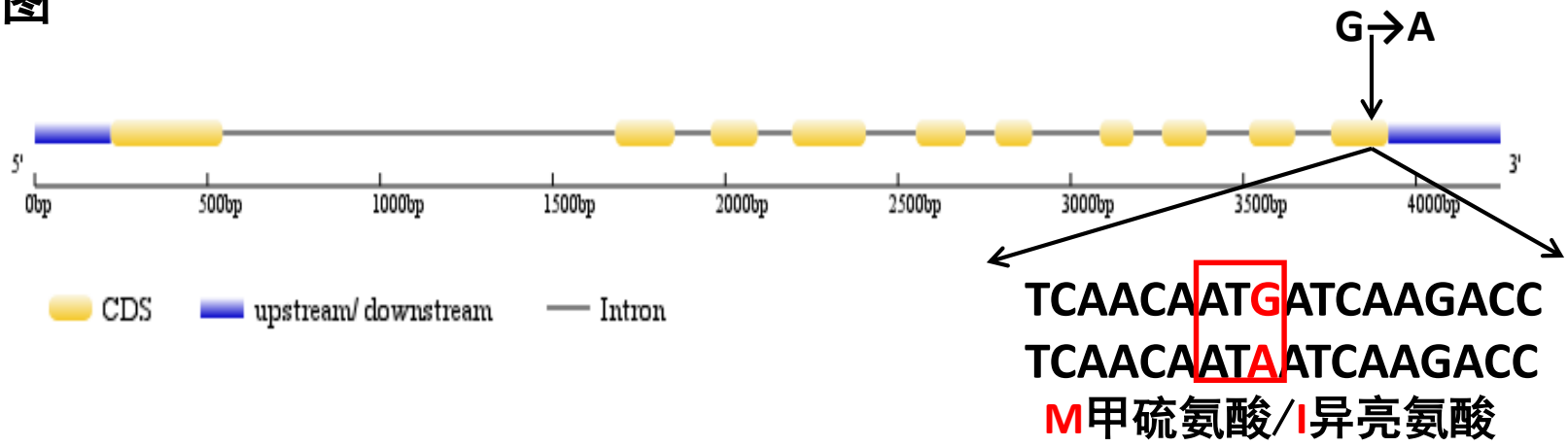


注释信息和序列如下：

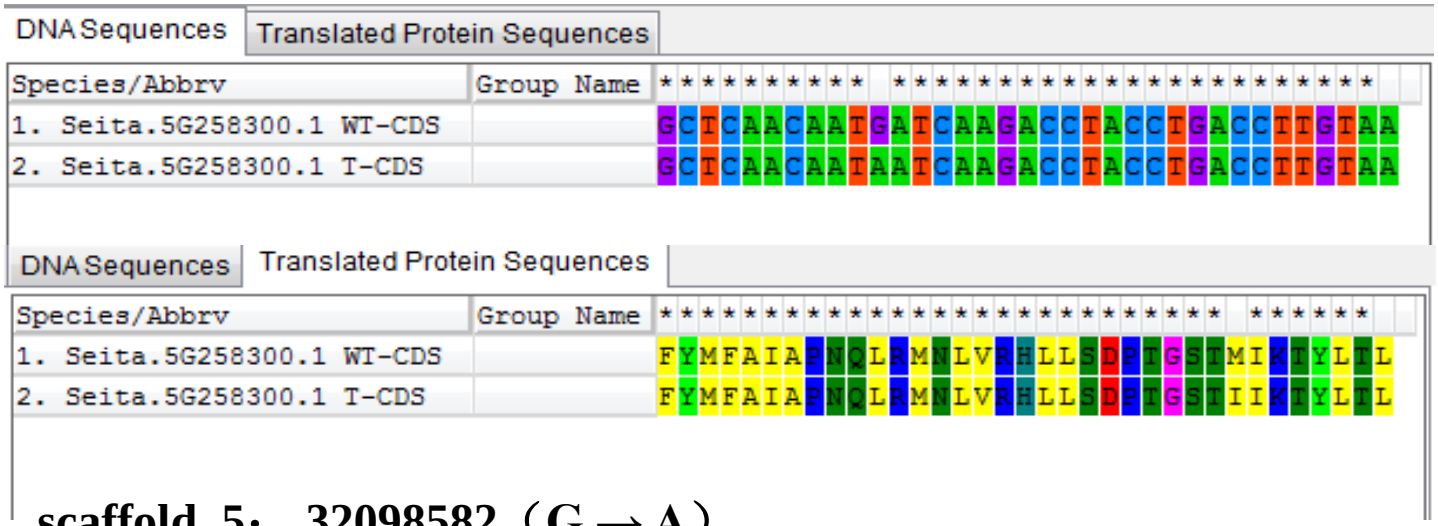


该基因编码番茄红素 ε -环化酶，位于5号染色体上，物理位置为32094687-32098931 ，基因组序列全长4245bp，转录组序列2168bp，编码序列1623bp，有10个外显子，9个内含子，编码540个氨基酸。

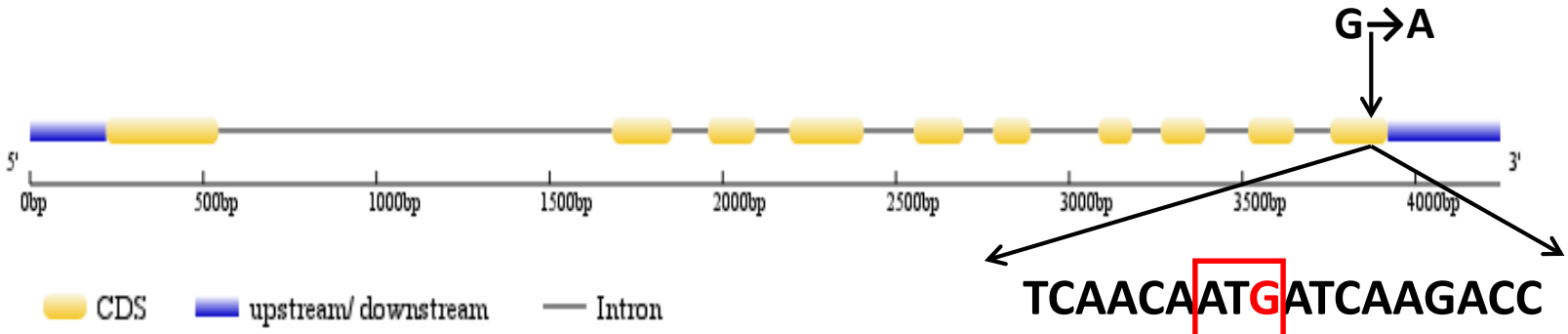
2、利用GSDS（Gene Structure Display Server）2.0在线软件作基因结构图



3、找到突变位点的物理位置，用MEGA做序列对比



scaffold_5： 32098582（G → A）



第10外显子中有一个碱基变化基因组序列3896bp处：

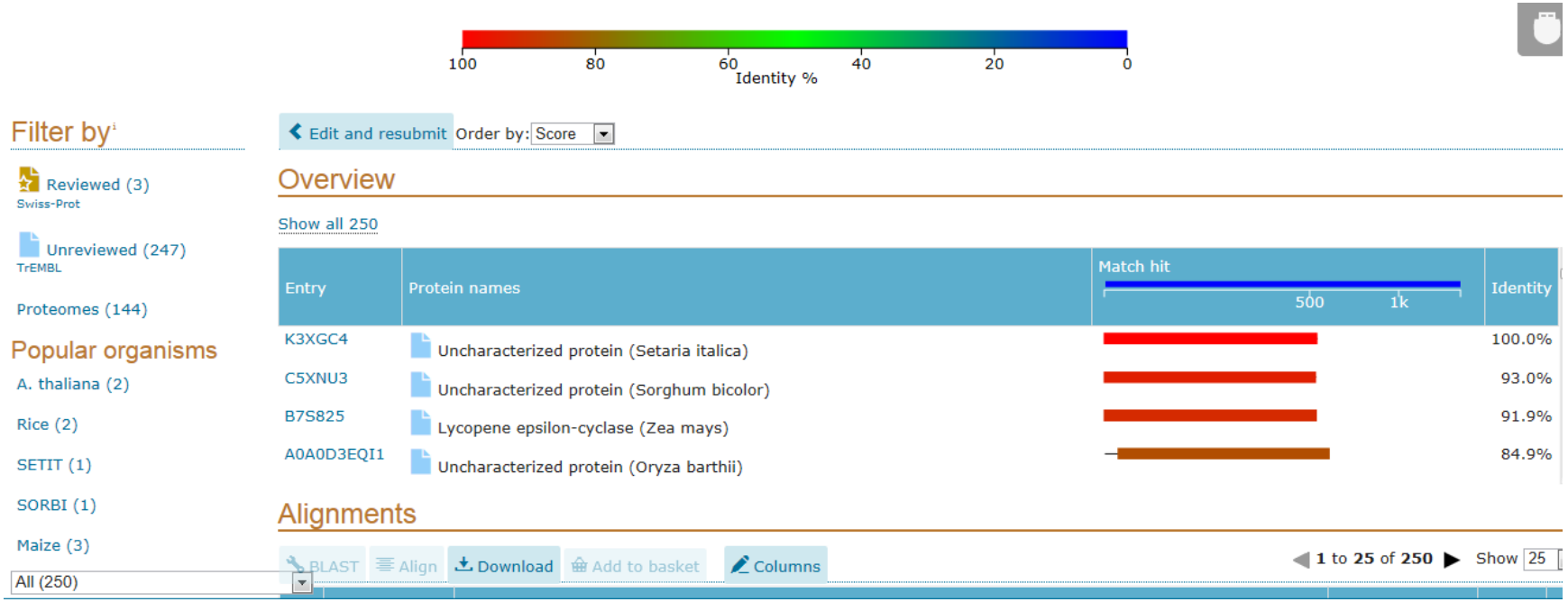
DNA： ATG→ATA

氨基酸：M甲硫氨酸→I异亮氨酸

甲硫氨酸（M）	化学式：$C_5H_{11}O_2NS$ 分子量：149.21 密度：1.340g/cm^3 等电点:5.74 沸点：186.8℃ 熔点：280~281℃ 溶于水
异亮氨酸（L）	化学式：$C_6H_{13}NO_2$ 分子量：131.17 密度为 1.038g/cm^3 熔点：332℃ 沸点为 217.7°C 略溶于水

三、序列比对及分子进化树

1、在Uniport中blast搜索



将氨基酸列进行blast搜索，发现有250个同源蛋白，其中有三个已经被审阅，相似度最高的是高粱和玉米，得分都在90%以上。

2、MEGA7做蛋白分子进化树

用氨基酸序列，在Uniprot中blast找到和该蛋白同源度较高的，用MEGA7做蛋白质分子进化树。与高粱中的蛋白同源度最高。

- 玉米：GRMZM2G012966，相似度高达90.9%，是玉米合成类胡萝卜素的主要基因。
- 小麦：radi2g41890，相似度达89.4%，与类胡萝卜素（维生素A的前体）的合成，小麦籽粒的颜色，胚乳中黄色素的含量有关。
- 水稻：

位点名称： LOC_Os01g39960

基因产物： lycopene epsilon cyclase, chloroplast precursor, putative, expressed

- LOC_Os01g39960基因，相似度达88.5%，在水稻的1号染色体上22534999-22538743位置，有一个可变剪接。它的基因产物是产物是番茄红素ε-环化酶。它与类胡萝卜素生物合成有关，且是促进番茄红素转化合成β-胡萝卜素的关键激素，通过使线性的番茄红素环化，通过加入环来产生叶黄素。同时该蛋白定位在叶绿体中且跨膜，所以猜测该蛋白除行使催化功能之外还可能参与物质运输或信号传递。

四、基因功能预测

1、在Plant-mPLoc网页和Uniport中预测亚细胞定位预测结果

Plant-mPLoc:Predicting subcellular localization of plant proteins including those with multiple sites

| Read Me | Data | Citation |

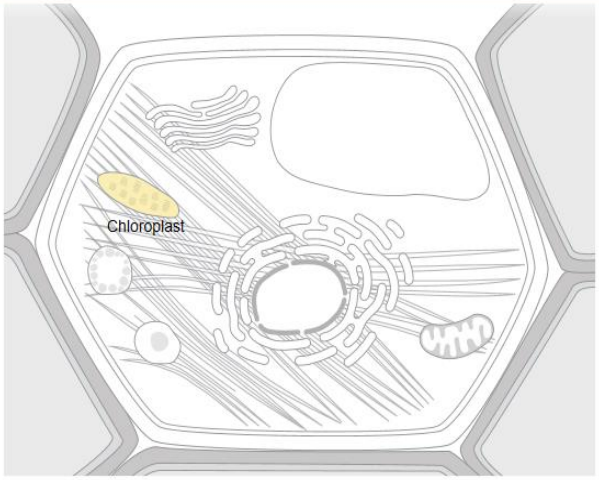
Your input sequence (540aa) is:

>Seita.5G258300.1
MGLSGAAISAPLGCRGLPHGAVGGGSKVRRAEVERWRRREGAGRRVAGPKVRCVATEKHD
ETAAVGAAAAGVEFADEEDYRKGGGGELLYVQMQATKPMESQSKIASKLLPISNENSVLD
LVIIGCGPAGLSLASESAKKGLTVGLIGPDLPFTNNYGVWEDEFKDLGLESCIEHVWKDT
IVYLDNNEPILIGRPYGRVHRDLLHEELLRRCYEAGVTYLNSKVDKIIESPDGHRVVCCE
RGREILCRLAIVASGAAASGRLLEYEVGGPRVCVQTAYGVEVEVENNPYDPSLMVFMDYRD
CFKEKFSHSEQENPTFLYAMPMSSTRVFFEETCLASKDAMPFDVLKKRLMYRLDAMGVRI
LKVHEEEWSYIPVGGSLPNTDQKNLAFGAAASMVHPATGYSVVRSLSEAPRYASVISDIL
RNRVPAQYLPGSSQNYSPSMLAWRTLWPQERKRQRSFFLFGLALIIQLNNEGIQTFFEAF
FRVPKWMWRGFLGSTLSSVDLILFSFYMFAIAPNQLRMNLVRHLLSDPTGSTMIKTYLTL

---------- Plant-mPLoc Computation Result ----------

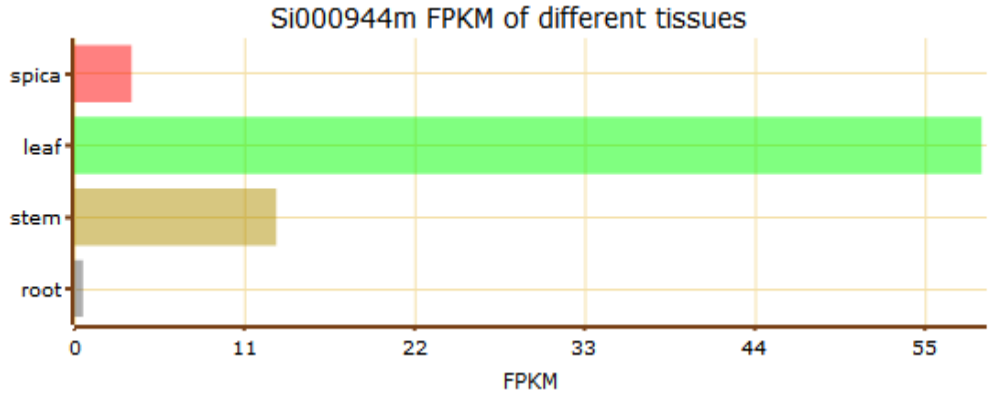
Query protein	Predicted location(s)
Seita.5G258300.1	Chloroplast.

Subcellular location[i]



这两个结果都表明该基因表达的部位是叶绿体

2、利用Setaria italica Functional Genomics Database数据库分析织组表达特异性：



在谷子的功能基因组数据库中找到，织特组异性表达主要在叶片部位高表达，其次是茎，再其次是穗部，再根上表达量很低。

3、在KEGG查找该基因参与的通路

Functional enrichments in your network

Note: some enrichments may be expected here (why?)

KEGG Pathways			
pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
00906	Carotenoid biosynthesis	4	5.96e-11
01110	Biosynthesis of secondary metabolites	4	4.58e-05
01100	Metabolic pathways	4	0.000283

PFAM Protein Domains			
pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
PF05834	Lycopene cyclase protein	2	0.000102
PF01593	Flavin containing amine oxidoreductase	2	0.0043

INTERPRO Protein Domains and Features			
pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
IPR010108	Lycopene cyclase, beta/epsilon	2	6.13e-05
IPR008671	Lycopene cyclase-like	2	9.2e-05
IPR002937	Amine oxidase	2	0.00517

该基因主要参与两条通路，类胡萝卜素生物合成，次生代谢产物的生物合成

五、蛋白结构及功能的预测分析

1、利用ExPASy中蛋白质组学内的ProtParam tool对*Seita.5G258300*氨基酸序列的理化性质进行分析:

氨基酸数量为540个，分子量为60065.98，理论等电点为7.12，各种氨基酸个数及所占比例如上所示。

Number of amino acids: 540

Molecular weight: 60065.98

Theoretical pI: 7.12

Amino acid composition:

Ala (A)	43	8.0%
Arg (R)	39	7.2%
Asn (N)	18	3.3%
Asp (D)	22	4.1%
Cys (C)	11	2.0%
Gln (Q)	13	2.4%
Glu (E)	40	7.4%
Gly (G)	46	8.5%
His (H)	10	1.9%
Ile (I)	24	4.4%
Leu (L)	55	10.2%
Lys (K)	23	4.3%
Met (M)	16	3.0%
Phe (F)	22	4.1%
Pro (P)	28	5.2%
Ser (S)	41	7.6%
Thr (T)	20	3.7%
Trp (W)	8	1.5%
Tyr (Y)	20	3.7%
Val (V)	41	7.6%
Pyl (O)	0	0.0%
Sec (U)	0	0.0%
(B)	0	0.0%
(Z)	0	0.0%
(X)	0	0.0%

Formula: $C_{2679}H_{4207}N_{739}O_{777}S_{27}$
Total number of atoms: 8429

Extinction coefficients:

Extinction coefficients are in units of M^{-1} cm^{-1}, at 280 nm measured in water.

Ext. coefficient 74425
Abs 0.1% (=1 g/l) 1.239, assuming all pairs of Cys residues form cystines

Ext. coefficient 73800
Abs 0.1% (=1 g/l) 1.229, assuming all Cys residues are reduced

Estimated half-life:

The N-terminal of the sequence considered is M (Met).

The estimated half-life is: 30 hours (mammalian reticulocytes, in vitro).
>20 hours (yeast, in vivo).
>10 hours (Escherichia coli, in vivo).

Instability index:

The instability index (II) is computed to be 42.13
This classifies the protein as unstable.

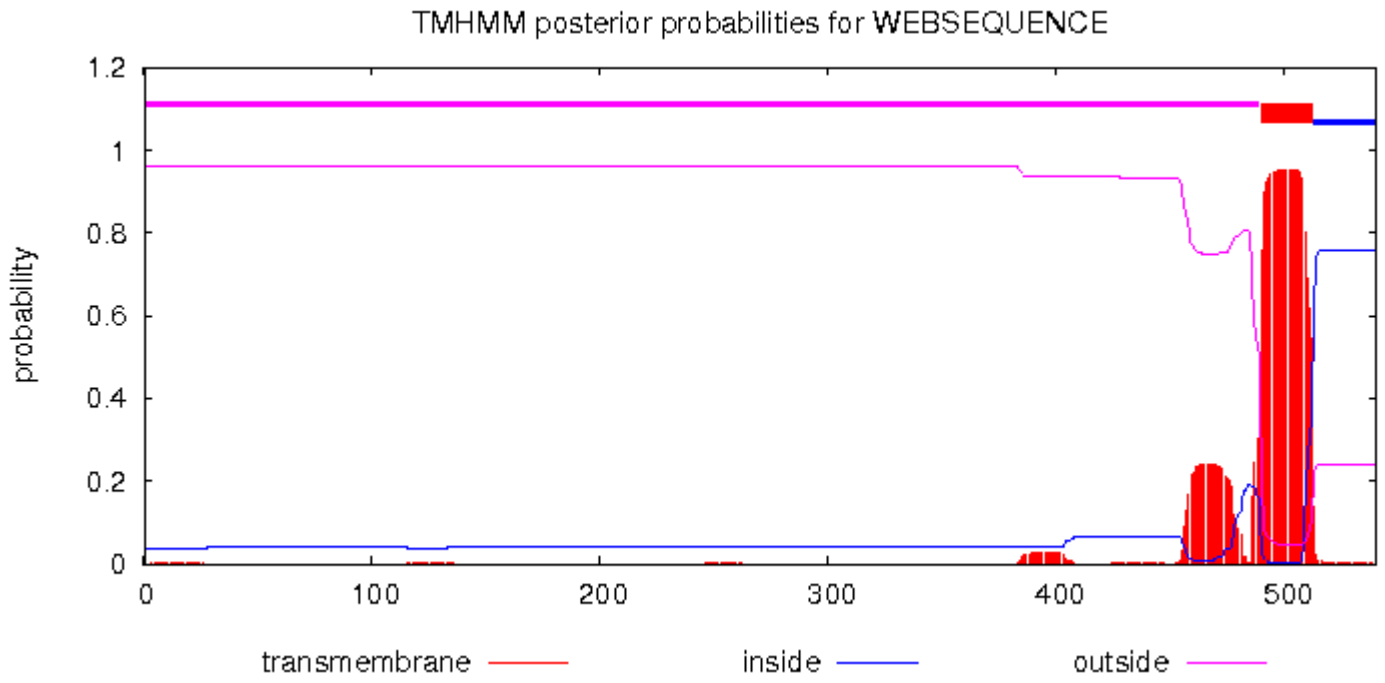
Aliphatic index: 87.04

Grand average of hydropathicity (GRAVY): -0.148

不稳定系数（Instability index）为42.13，说明稳定性不好；脂肪族氨基酸系数（Aliphatic index）为87.04%，即脂肪族氨基酸含量高；平均亲水系数（Grand average of hydropathicity）为负，说明总体为疏水性较强。

2、预测氨基酸的保守型Conservation

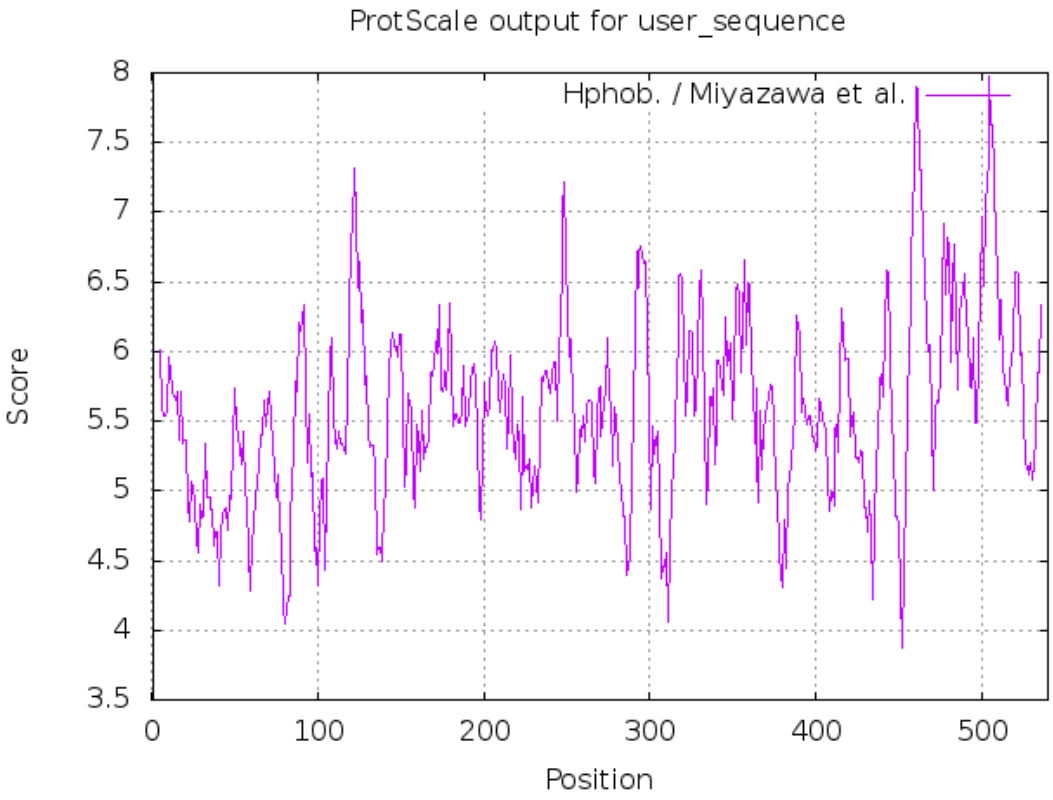
3.用ExPASy蛋白质组学中的TMHMM工具预测的跨膜区域



该工具预测出一个在第500个氨基酸位置附近的跨膜区域，大多数氨基酸都分布在膜外，只有跨膜之后的区域被预测在膜内，推测这与其功能密切相关。

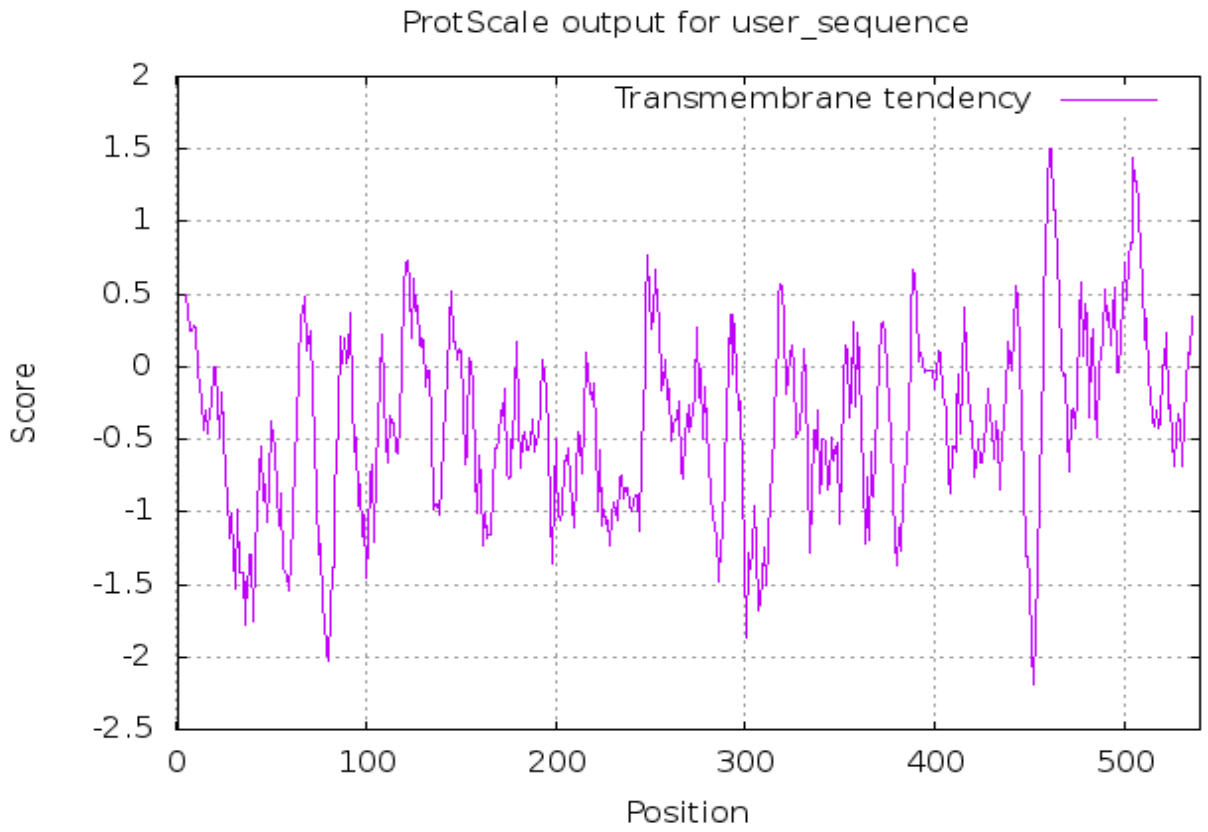
4.使用 ExPASy 内proteomics中的 ProtScale 工具对其氨基酸序列进行分析：

1）疏水性分析结果：



由结果可知在第450-500个氨基酸位置有两个疏水性较强的区域，这段疏水性较强的区域与预测出的其跨膜区域大致相同。

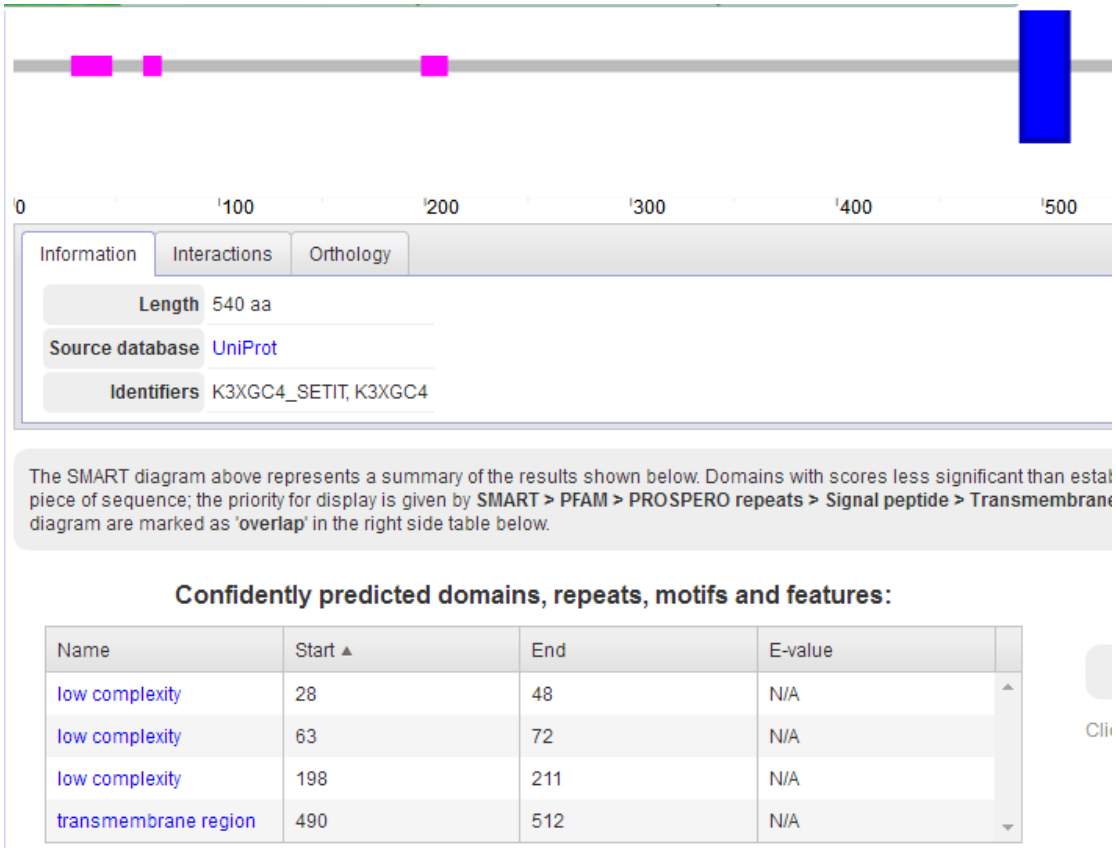
2）利用ProtScale对该蛋白的跨膜趋势进行分析：



该结果预测的跨膜区域与TMHMM预测的区域相同。

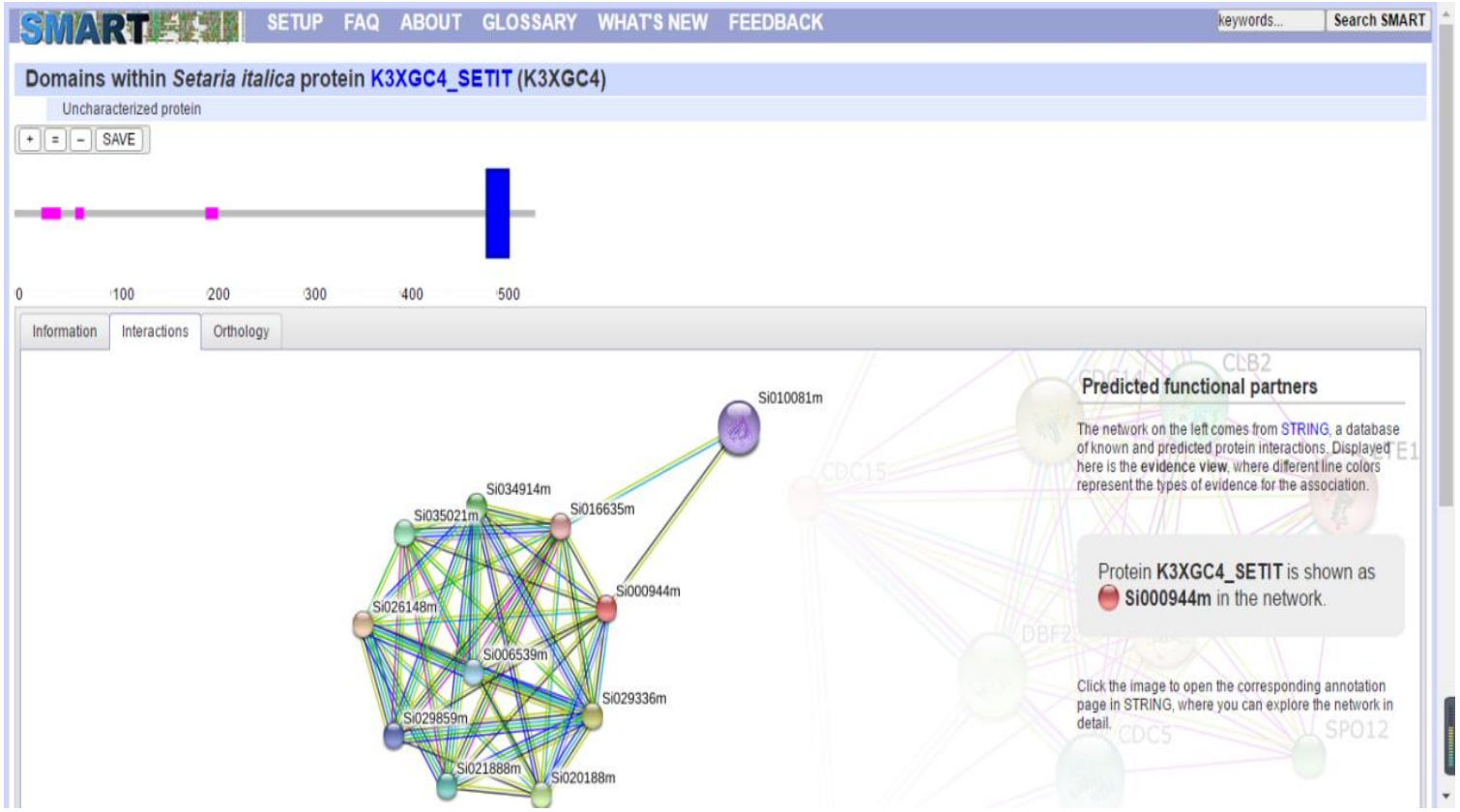
5.用SMART对蛋白序列进行分析结果如下：

1）Information窗口：



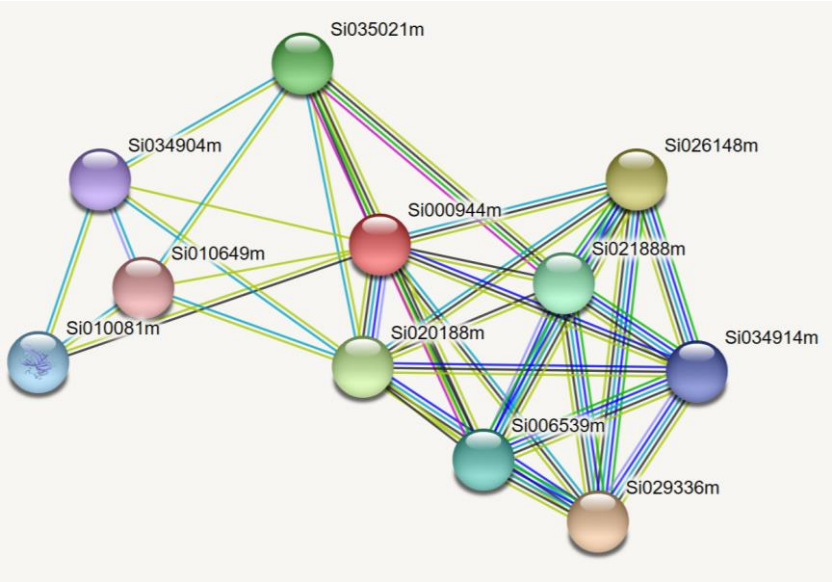
Information窗口结果显示其含有三个低复杂度区域和一个跨膜区域，紫红色部分为低复杂度区域，有可能存在结构域，蓝色部分为预测的跨膜区域，与TMHMM预测区域位置相同，推测其功能有可能参与跨膜运输。

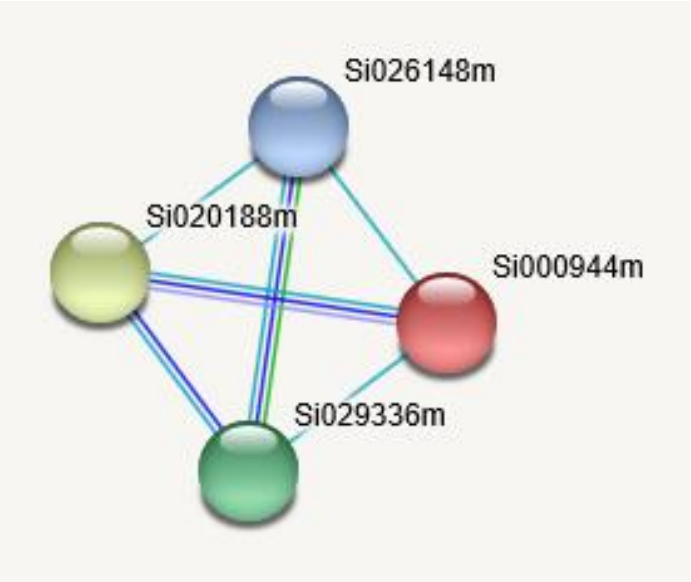
2）Interaction窗口：



Interaction窗口信息中包含了与该蛋白相互作用的蛋白的信息，这些蛋白质的功能之间有密切的联系，对该功能预测提供重要的参考作用。

STRING



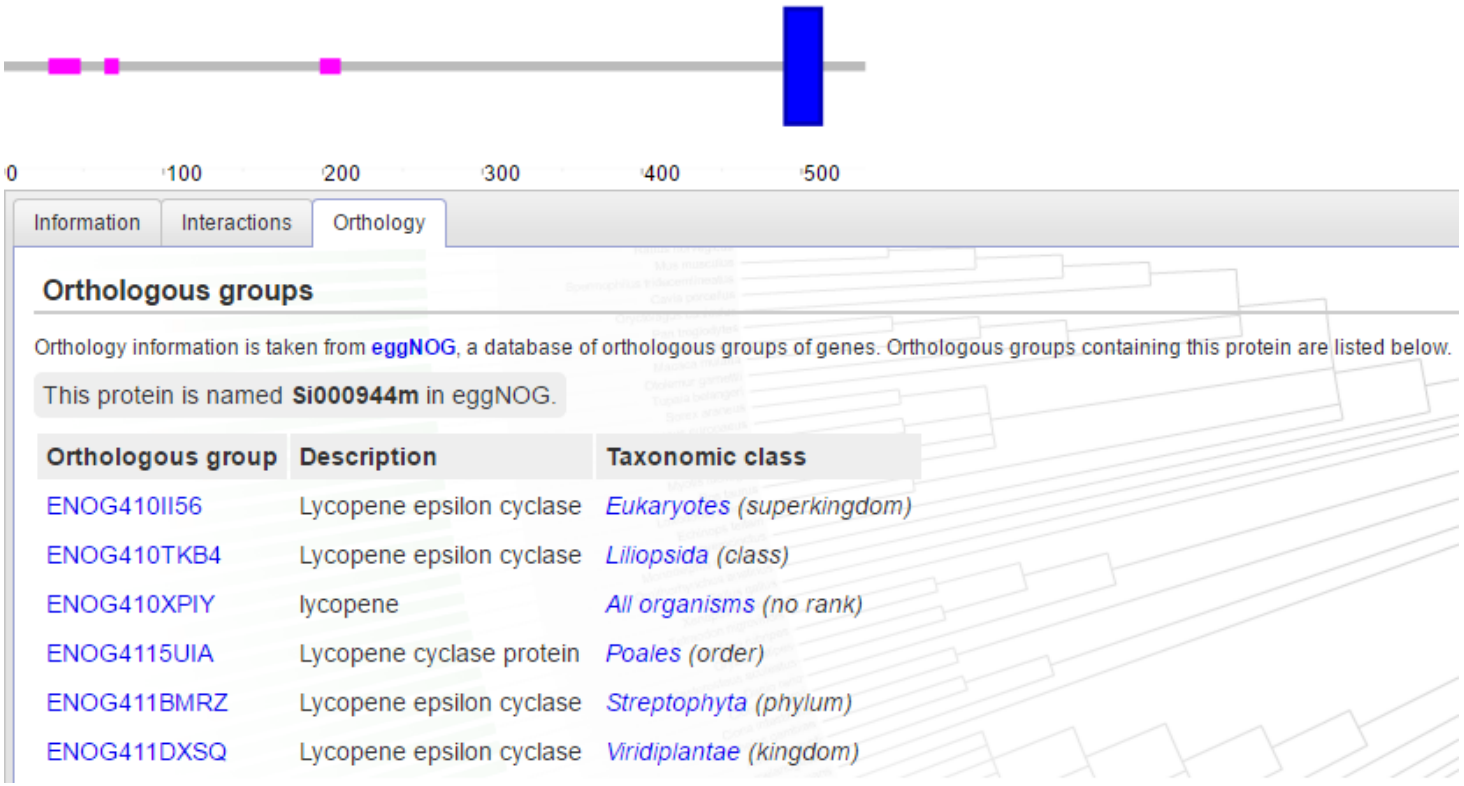


Known Interactions
- *from curated databases*
- *experimentally determined*

Predicted Interactions
- *gene neighborhood*
- *gene fusions*
- *gene co-occurrence*

Others
- *textmining*
- *co-expression*
- *protein homology*

3）Orthology窗口：

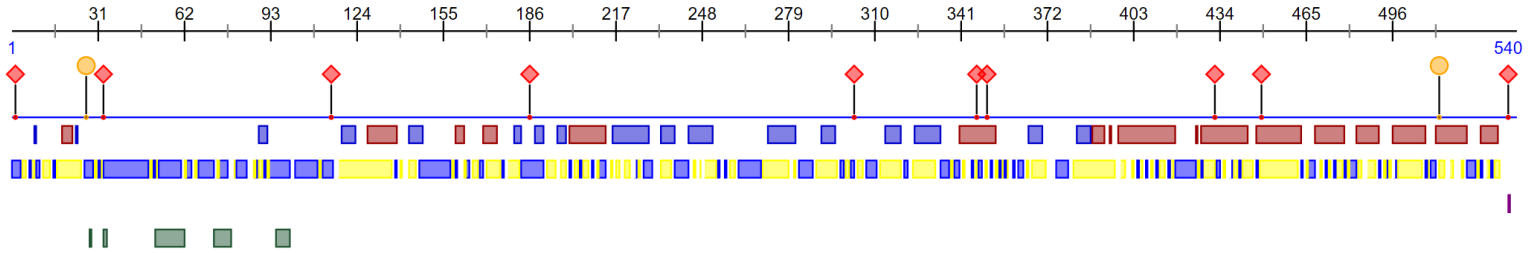


Orthology窗口找到了该基因的直系同源基因：番茄红素环化酶、番茄红素、番茄红素环化酶蛋白，推测该基因可能与其直系同源基因在细胞内行使相同的功能。

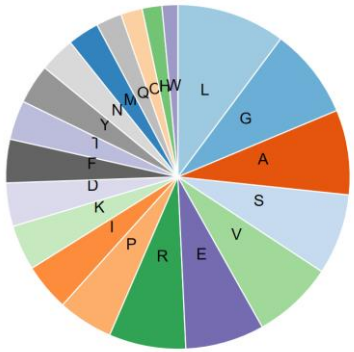
Protein prediction

Summary

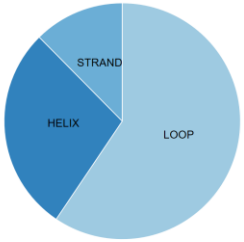
Sequence Length	540
Number of Aligned Proteins	49



Amino Acid composition



Secondary Structure Composition



Solvent Accessibility

Proteins can be classified as **mixed** given the following classes:

- 'all-alpha': %H > 45% AND %E < 5%
- 'all-beta': %H < 5% AND %E > 45%
- 'alpha-beta': %H > 30% AND %E > 20%
- 'mixed': All others

Pr protein binding region

polynucelotide binding region

helix

5、利用其蛋白序列，用Phyre2预测出的蛋白质三维结构如下：

Top model

Image coloured by rainbow N → C terminus

Model dimensions (Å): **X**:58.555 **Y**:64.621 **Z**:47.311

Model (left) based on template c2qa1A

Top template information

PDB header:oxidoreductase
Chain: A: **PDB Molecule:**polyketide oxygenase pgae;
PDBTitle: crystal structure of pgae, an aromatic hydroxylase involved in2 angucycline biosynthesis

Confidence and coverage

Confidence: **100.0%** Coverage: **64%**

346 residues (64% of your sequence) have been modelled with 100.0% confidence by the single highest scoring template.

Additional confident templates have been detected (see Domain analysis) which cover other regions of your sequence.

442 residues (82%) could be modelled at >90% confidence using multiple-templates.

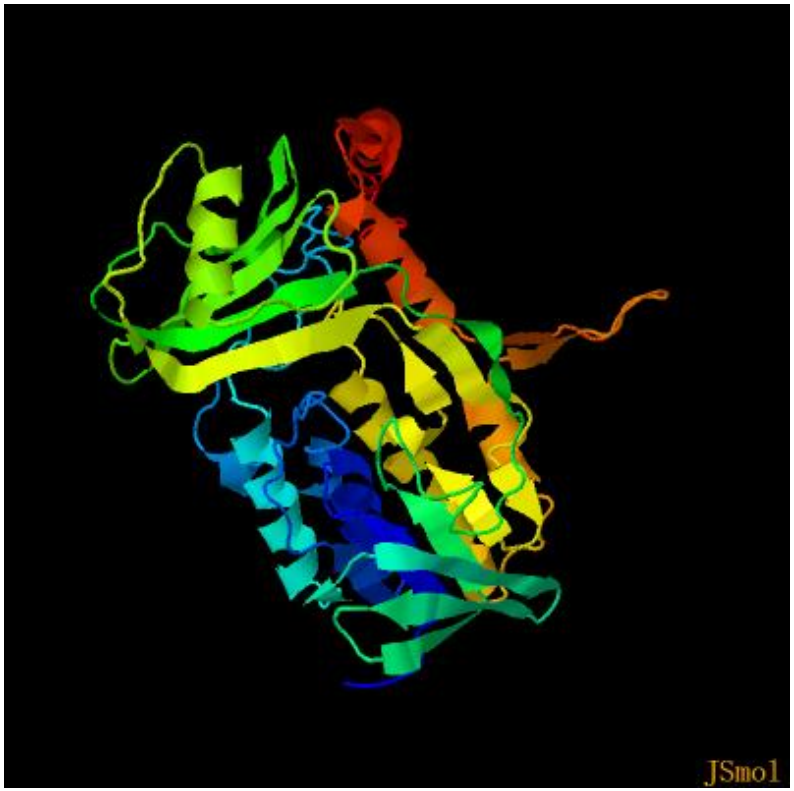
You may wish to try resubmitting your sequence in "intensive" mode to model more of your sequence.

3D viewing

Interactive 3D view in JSmol

For other options to view your downloaded structure offline see the FAQ

正常蛋白和突变蛋白的结构对比：



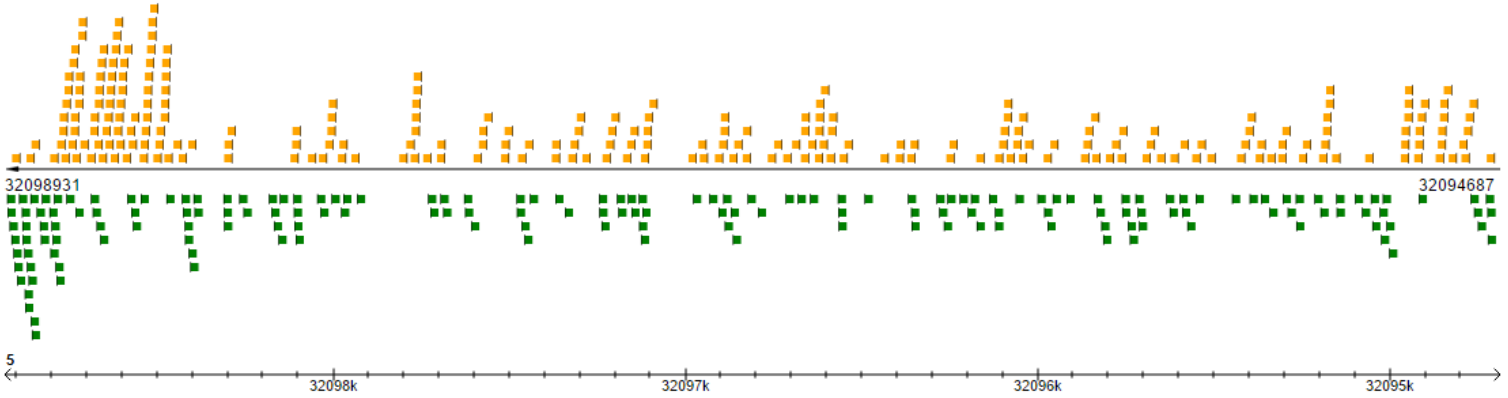
突变位点是在螺旋结构上

六、CRISPR构建载体

1、用CRISPR-P 2.0在线设计靶点

2. Mapping the reads...

ORG: Setaria italica (JGIv2.0), **Position**: 5:32098931..32094687, **Length**: 4245



Start with 'A' 'G' The current sgRNAs are G(N)20GG or A(N)20GG depending on if U6 or U3 promoters are used for transcribing the RNA molecules.

Sort by 'score' 'gc' 'position'

	On-score ⓘ	Sequence	Region	%GC	
guide1	0.9722	TCTTGCTTCGAATTCCGCGGCGG	CDS	55%	☑
guide2	0.8604	GCAGTCTGCACACAGACACGGGG	CDS	60%	☑
guide3	0.8297	CATCTACTCTCTGATCCGACCGG	CDS	50%	☑
guide4	0.8247	CGGAATTCGAAGCAAGACAGCGG	CDS	50%	☑
guide5	0.8050	CTGTCTTGCTTCGAATTCCGCGG	CDS	50%	☑
guide6	0.7763	ATACCGGATATGGCACGTGGCGG	Intergenic	55%	☑
guide7	0.7579	AAGATCATCGAATCTCCAGATGG	CDS	40%	☑
guide8	0.7570	GATTGGCCGTCCATATGGCAGGG	CDS	55%	☑

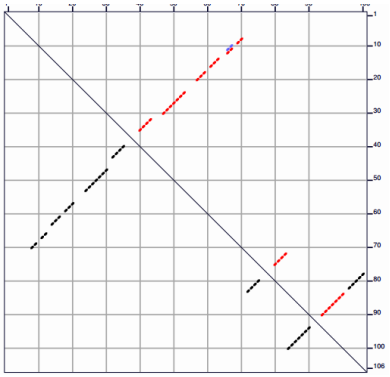
2、靶点的位置及二级结构预测

Guide-1：GAAGCTAGAGAAAGACCGGC

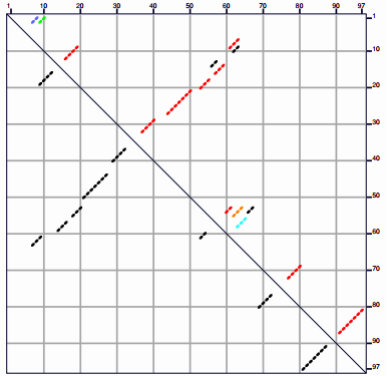
GAAACATGTGCTGTTAAAAATATGTGCACATTAACAATGTTTTAGTTCTTGTTTTATCGTAAACCTATTTTCTTAGGGAATGAAGTATTTCCTTACATGA
TGAAGTAGCTAAAAATCGTACTGGCATAGGTTCGCTGTATGGTTCAGGAAAAATGAATATGTGGTCTTTTCTTTTTCGCAGCTGCTGCCCATATCCAATG
AAAATTCAGTACTTGATCTGGTTATCATTGGCTGTGGTCCTGCCGGTCTTTCTCTAGCTTCAGAGTCAGCCAAGAAAGGCCTCACTGTTGGTCTTATTGG
CCCTGACCTTCCGTTCACGAATAACTATGGTGTGTGGGAGGATGAATTCAAAGGTATTATATTATTTGCATTGCTACGATGAAGAGTTTTTGCATAATAT
CTTTATCAACATAATTTACTTTGACGATACTTATTCTTTTCTCTCTTTTTCTGTCCAGATCTTGGTCTAGAGAGCTGTATTGAGCATGTCTGGAAGGATA
CCATTGTCTATCTAGACAATAATGAACCAATACTGATTGGCCGTCCATATGGCAGGGTGCACCGTGACCTGCTGCATGAGGAGTTGCTGAGAAGGTAAAT

Guide-2：GTTCCTGCGCAATATTTGCC

GTATGGAACTTCCATTACATCTGTGATTCTACAACCAGTACTTCCATGCTTCAAAGTTTCCGATCAAATTCTTTATCACAGGAAACATGCTTAGCCTCTA
AAGATGCGATGCCCTTTGATGTACTTAAGAAAAGGTTGATGTATCGGTTGGATGCAATGGGAGTTCGGATCCTGAAAGTTCATGAGGAGGTAAGAAGTTA
AGGGTCACTAGCATGTTCGGCTATGATTCTTGCCGCTGCGCTGCAGCTGCGCCTGCAGCACCAACAACTCTGGTTCCTGTAGTTTTAGAAGGCTAGCAGC
TGCAGCAATCTCAGCCAAACAGCTTGTTAATTTCTCATTCATATGTATTGTGAACATAAACTGAAATGCTGAAGCTTGGTTTCAGGAATGGTCCTACATT
CCTGTTGGAGGTTCCTTACCAAATACAGATCAGAAAAATCTTGCATTTGGTGCTGCAGCAAGTATGGTGCACCCTGCAACTGGTATGGCCAAATCCTTAA
TTTTTACACACCATGTCCTTCCTTGCAATCTAGCTGATATTCACAACGTGTTGGAAAATTCATAGGCTACTCAGTGGTCAGATCTTTGTCTGAGGCTCCA
AGATATGCCTCTGTAATATCAGATATCTTAAGGAACCGAGTTCCTGCGCAATATTTGCCAGGAAGTTCTCAAAATTACAGTCCATCAATGCTTGGTAAGT
ATTCTGCTGGTTTTTACTCTCGTAGACATCACTTCTGGACAAGTACAGCTTCAGTCTTTTTAAATTTTAGACAAGTGAGCAAAACCTTCTGCTTCTAATT



Guide-1



Guide-2

七、使用到的工具

- 查找基因序列——**phytozome**
- 画基因结构图——**GSDS（Gene Structure Display Server）2.0**
- 序列比对、分子进化树——**MEGA 7**
- 亚细胞结构预测——**Plant-mPLoc**
- 组织表达分析——**Setaria italica Functional Genomics Database**
- **ProtParam tool**——氨基酸的性质
- **TMHMM**——预测跨膜结构域
- **ProtScale**——氨基酸的跨膜趋势和亲疏水性质
- **SMART**——预测结构域、互作蛋白、直系同源
- **Protein prediction** ——蛋白质预测
- **STRING**——蛋白的交互作用
- 蛋白质结构预测——**Phyre 2 / PDB**软件
- 设计**CRISPR**靶点——**CRISPR-P 2.0**
- 检查引物的二级结构——**The mfold Web Server**

谢谢