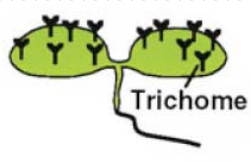
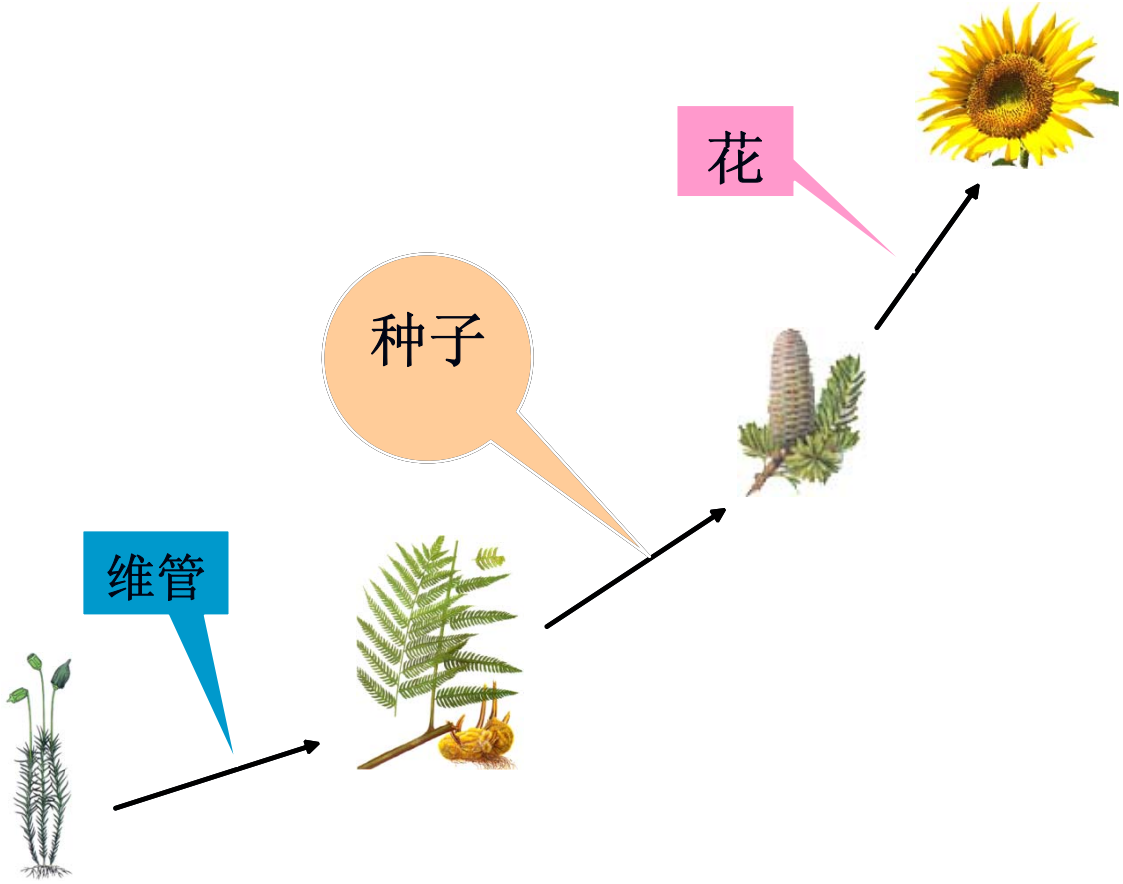
拟南芥CCAAT-HAP3基因家族分析

徐礼鸣，边洋，李晶，付立文

2008-6-16

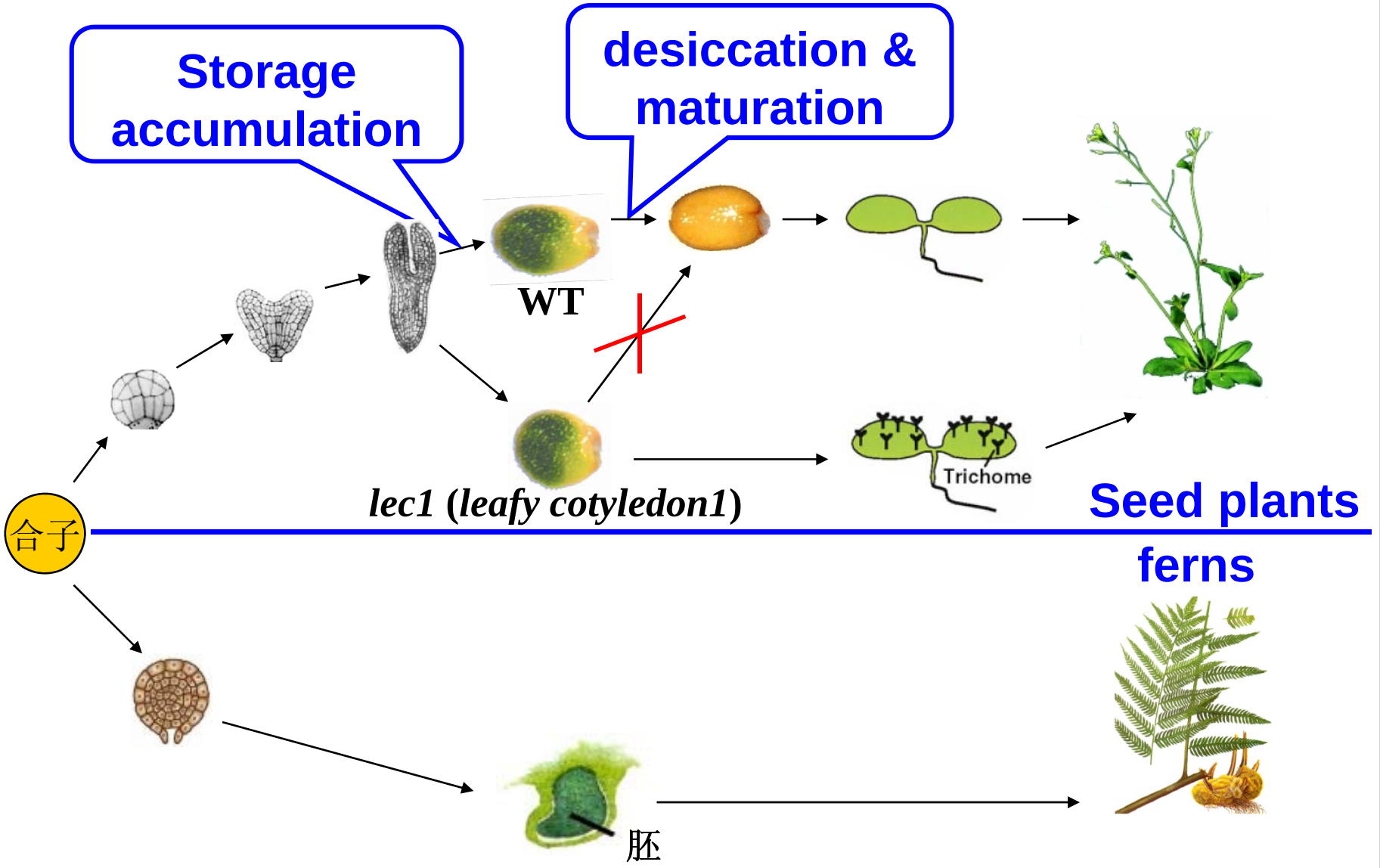


研究背景



Three key innovations of land plants

Life histories of seed plant and fern



研究背景

- 拟南芥基因组中CCAAT-HAP3 基因家族共有**10个**成员， LEC1 和L1L 属于**LEC1-type**，另外8 个属于**non-LEC1-type**，不具有LEC1 基因的功能，(Kwong et al., 2003)。
- 从蛋白序列上看，与所有的HAP3 一样，LEC1 中部有一个**保守的B 区(B domain)**，N 端和C 端分别有一个**不保守的A 区(A domain)和C 区(C domain)**
- 但LEC1 和L1L 在**B 区有16 个氨基酸残基**与该家族其它的成员不同，这些残基**与LEC1 的功能**有重要关系。

研究目标

- 研究LEC1基因在拟南芥中的进化，分析其进化地位
- 分析LEC1型基因和非LEC1型基因编码蛋白的主要差异，寻找关键的变异位点

研究方法

- LEC1基因在拟南芥中的进化
  - 收集拟南芥中所有10个HAP3家族的基因和所有可变剪切体
  - 分析这些可变剪切体的特性，进行蛋白序列的取舍
  - 分析序列特征，获取基本序列特征信息
  - 寻找合适的外类群建树，分析LEC1基因在拟南芥中的进化

研究方法

- LEC1型基因和非LEC1型基因编码蛋白的主要差异
  - 多序列比对，MEME分析，寻找保守motif，建立motif的系统发育树
  - 对于保守Motif进行Blast分析，寻找保守结构域
  - 对Motif进行多序列比对，寻找可能的引起功能变化的位点
  - 实验验证
  - 三维结构验证

数据收集

No	Name	Gene Symbol	TAIR ID	RefSeq ID	N	GenPet ID	L	Gene ID
1	atlec1a	LEC1	AT1G21970.1	NM_102046.4	2	NP_173616.2	238	838800
2	atlec2a	L1L	AT5G47670.1	NM_124141.3	2	NP_199578.2	234	834818
3	atlec2b	L1L	AT5G47670.2	NM_001085258.1	1	NP_001078727.1	205	834818
4	atlec3a	AT1G09030	AT1G09030.1	NM_100774.1	1	NP_172377.1	139	837424
5	atlec4a	AT2G13570	AT2G13570.1	NM_126937.1	1	NP_178981.1	215	815843
6	atlec5a	AT2G37060	AT2G37060.1	NM_001036423.1	6	NP_001031500.1	173	818282
7	atlec5b	AT2G37060	AT2G37060.2	NM_179946.3	6	NP_850277.2	173	818282
8	atlec5c	AT2G37060	AT2G37060.3	NM_201888.2	6	NP_973617.1	173	818282
9	atlec6a	ATHAP3	AT2G38880.1	NM_129445.2	5	NP_030436.1	141	818472
10	atlec6b	ATHAP3	AT2G38880.2	NM_001036434.1	5	NP_001031511.1	141	818472
11	atlec6c	ATHAP3	AT2G38880.3	NM_001036435.1	3	NP_001031512.1	112	818472
12	atlec6d	ATHAP3	AT2G38880.4	NM_179974.3	4	NP_850305.2	140	818472
13	atlec6e	ATHAP3	AT2G38880.5	NM_001036433.1	3	NP_001031510.1	112	818472
14	atlec6f	ATHAP3	AT2G38880.6	NM_179973.3	5	NP_850304.2	141	818472
15	atlec7a	At2g47810	At2g47810.1	NM_130348.2	1	NP_182302.1	160	819393
16	atlec8a	AT3G53340	AT3G53340.1	NM_115194.2	6	NP_190902.2	176	824502
17	atlec9a	AT4G14540	AT4G14540.1	NM_117534.2	1	NP_193190.1	161	827101
18	atlec10a	At5g47640	At5g47640.1	NM_124138.2	1	NP_199575.1	190	834815

拟南芥中所有10个HAP3家族的基因及其所有可变剪切体

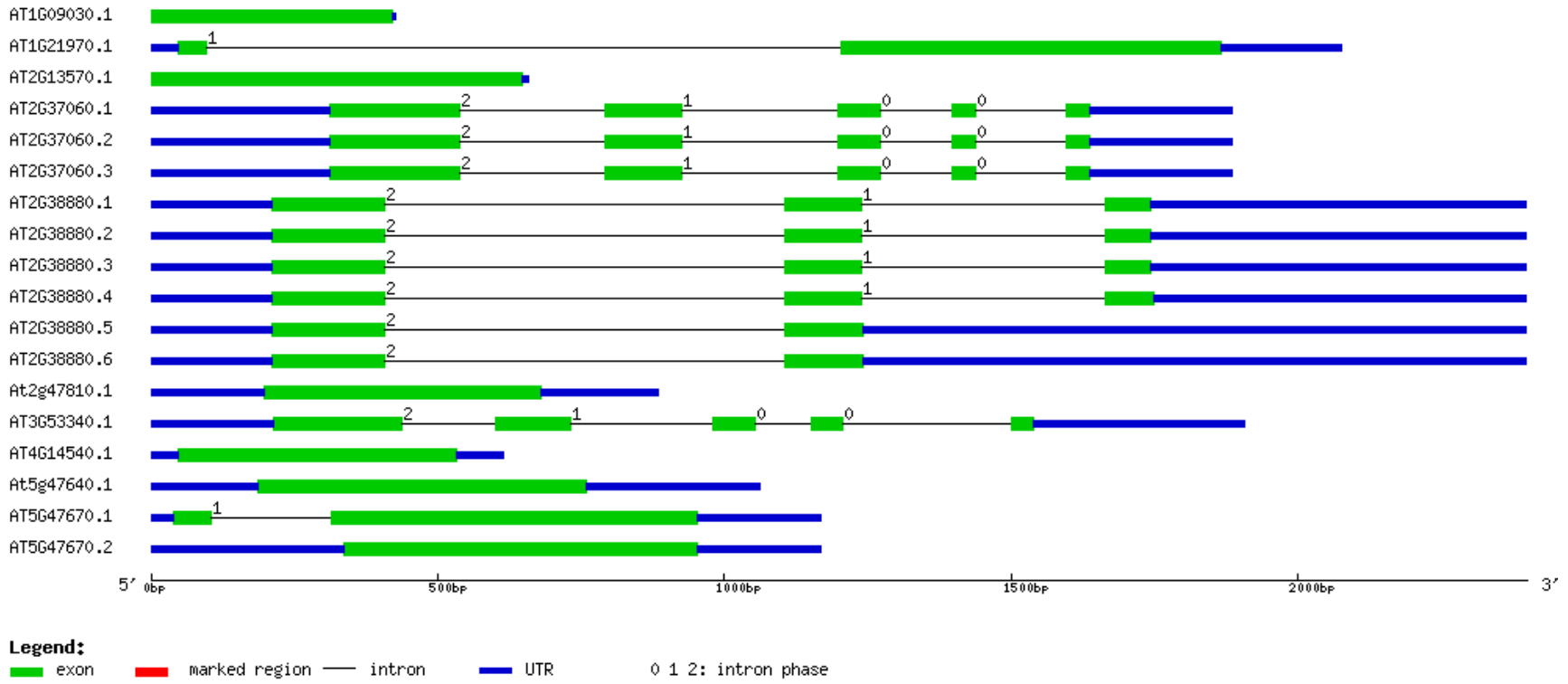
N为外显子数目
L为蛋白序列长度

数据收集

Name	TAIR ID	GenPet ID	L	uniprot AC	Uniprot ID	Length
atlec1a	AT1G21970.1	NP_173616.2	238	Q9SFD8	NFYB9_ARATH	208
atlec2a	AT5G47670.1	NP_199578.2	234	Q84W66	NFYB6_ARATH	234
atlec2b	AT5G47670.2	NP_001078727.1	205	Q84W66	NFYB6_ARATH	205
atlec3a	AT1G09030.1	NP_172377.1	139	O04027	NFYB4_ARATH	139
atlec4a	AT2G13570.1	NP_178981.1	215	Q9SIT9	NFYB7_ARATH	215
atlec5a	AT2G37060.1	NP_001031500.1	173	Q8VYK4	NFYB8_ARATH	173
atlec5b	AT2G37060.2	NP_850277.2	173	Q8VYK4	NFYB8_ARATH	173
atlec5c	AT2G37060.3	NP_973617.1	173	Q8VYK4	NFYB8_ARATH	173
atlec6a	AT2G38880.1	NP_030436.1	141	Q9SLG0	NFYB1_ARATH	141
atlec6b	AT2G38880.2	NP_001031511.1	141	Q9SLG0	NFYB1_ARATH	141
atlec6c	AT2G38880.3	NP_001031512.1	112	Q3EBK1	Q3EBK1_ARATH	112
atlec6d	AT2G38880.4	NP_850305.2	140			
atlec6e	AT2G38880.5	NP_001031510.1	112	Q3EBK1	Q3EBK1_ARATH	112
atlec6f	AT2G38880.6	NP_850304.2	141	Q9SLG0	NFYB1_ARATH	141
atlec7a	At2g47810.1	NP_182302.1	160	O82248	NFYB5_ARATH	160
atlec8a	AT3G53340.1	NP_190902.2	176	Q67XJ2	NFYBA_ARATH	176
atlec9a	AT4G14540.1	NP_193190.1	161	O23310	NFYB3_ARATH	161
atlec10a	At5g47640.1	NP_199575.1	190	Q9FGJ3	NFYB2_ARATH	190

在Uniprot数据库中寻找这拟南芥中所有10个HAP3家族的基因的蛋白序列

GSDS分析基因结构


AT1G09030.1
AT1G21970.1
AT2G13570.1
AT2G37060.1
AT2G37060.2
AT2G37060.3
AT2G38880.1
AT2G38880.2
AT2G38880.3
AT2G38880.4
AT2G38880.5
AT2G38880.6
At2g47810.1
AT3G53340.1
AT4G14540.1
At5g47640.1
AT5G47670.1
AT5G47670.2
5′
0bp
500bp
1000bp
1500bp
2000bp
3′
Legend:
exon
marked region
intron
UTR
0 1 2: intron phase

分析引起蛋白序列改变引起的可变剪切

- 拟南芥CCAAT-HAP3基因家族共有10个基因，18个剪接体，
- 大部分只产生一个序列
- ATLEC6(AT2G38880)剪接以后都分别产生3个不同的蛋白剪接体
- ATLEC2(AT5G47670)剪接以后都分别产生2个不同的蛋白剪接体

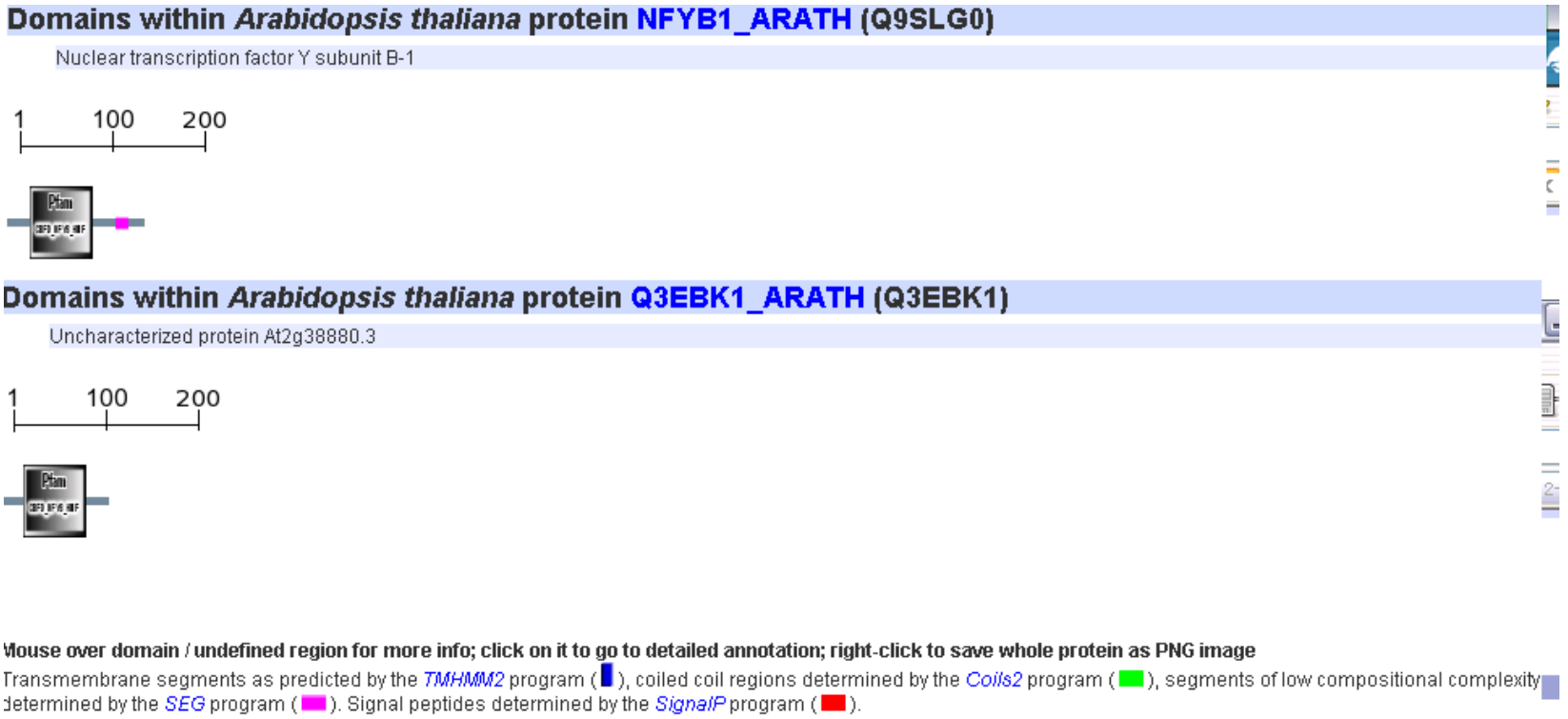
ATLEC6产生的3个剪接体分析

```
# Aligned_sequences: 2
# 1: Q3EBK1_ARATH
# 2: NFYB1_ARATH
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 141
# Identity:     110/141 (78.0%)
# Similarity:   111/141 (78.7%)
# Gaps:          29/141 (20.6%)
# Score: 569.0
#
#
#=======================================

Q3EBK1_ARATH       1 MADTPSSPAGDGGESGGSVREQDRYLPIANISRIMKKALPPNGKIGKDAK     50
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
NFYB1_ARATH        1 MADTPSSPAGDGGESGGSVREQDRYLPIANISRIMKKALPPNGKIGKDAK     50

Q3EBK1_ARATH      51 DTVQECVSEFISFITSEASDKCQKEKRKTVNGDDLLWAMATLGFEDYLEP    100
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
NFYB1_ARATH       51 DTVQECVSEFISFITSEASDKCQKEKRKTVNGDDLLWAMATLGFEDYLEP    100

Q3EBK1_ARATH     101 LKIYLARYREVW                                    112
                     ||||||||||:.
NFYB1_ARATH      101 LKIYLARYRELEGDNKGSGKSGDGSNRDAGGGVSGEEMPSW    141
```

Needle比对ATLEC6产生的序列长度为141和112的剪接体

SMART分析结果

（http://smart.embl-heidelberg.de/）



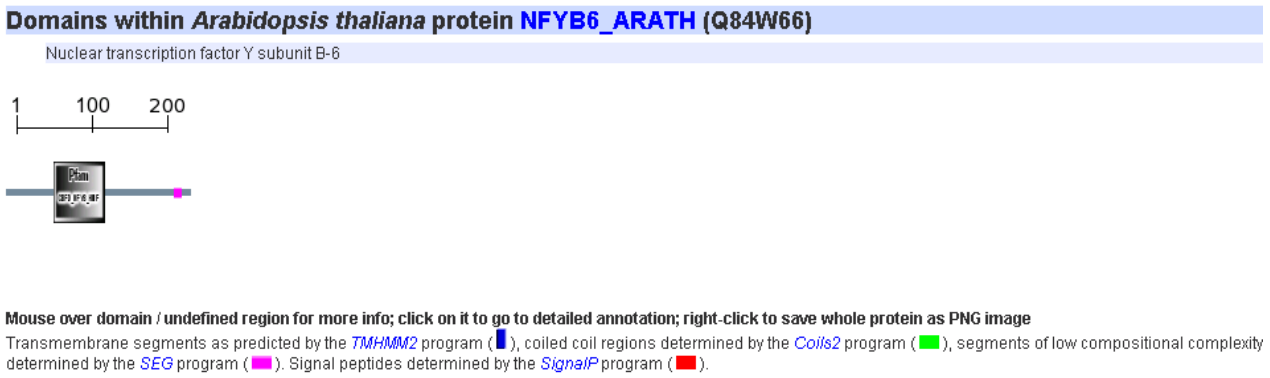
上图为141个氨基酸的序列的分析结果，下图为112个的分析结果

总结

- 含有140个氨基酸的序列与141个氨基酸的序列只有C端最后一两个氨基酸的差异，在这里不予考虑，可能是测序错误，很难影响功能
- 含有141个氨基酸的序列的蛋白SMART分析发现，存在一个含112个氨基酸的剪接体所没有的，富含Gly的低重复区域，序列为“GDNKGSGKSGDGSN”，因此推测较短的剪接体的功能可能被影响了。
- 这次的分析采用较长的含141个氨基酸的蛋白序列。

ATLEC2产生的2个剪接体分析

- 在Swiss-prot的注释信息中标明，较短的剪切体缺少了N端前1-29个氨基酸
- SMART分析显示，这个蛋白含有与上述atlec6一样的2个结构域，分别位于61-126和213-222，没有影响到缺少的那一部分碱基
- 这段区域的缺失对于整个蛋白的结构是否有影响需要实验证实。这次选用较长剪接体。



蛋白序列选择

- 蛋白序列基本采用Swiss-Prot注释信息中的较长的序列，LEC1则采用GenPet中的序列

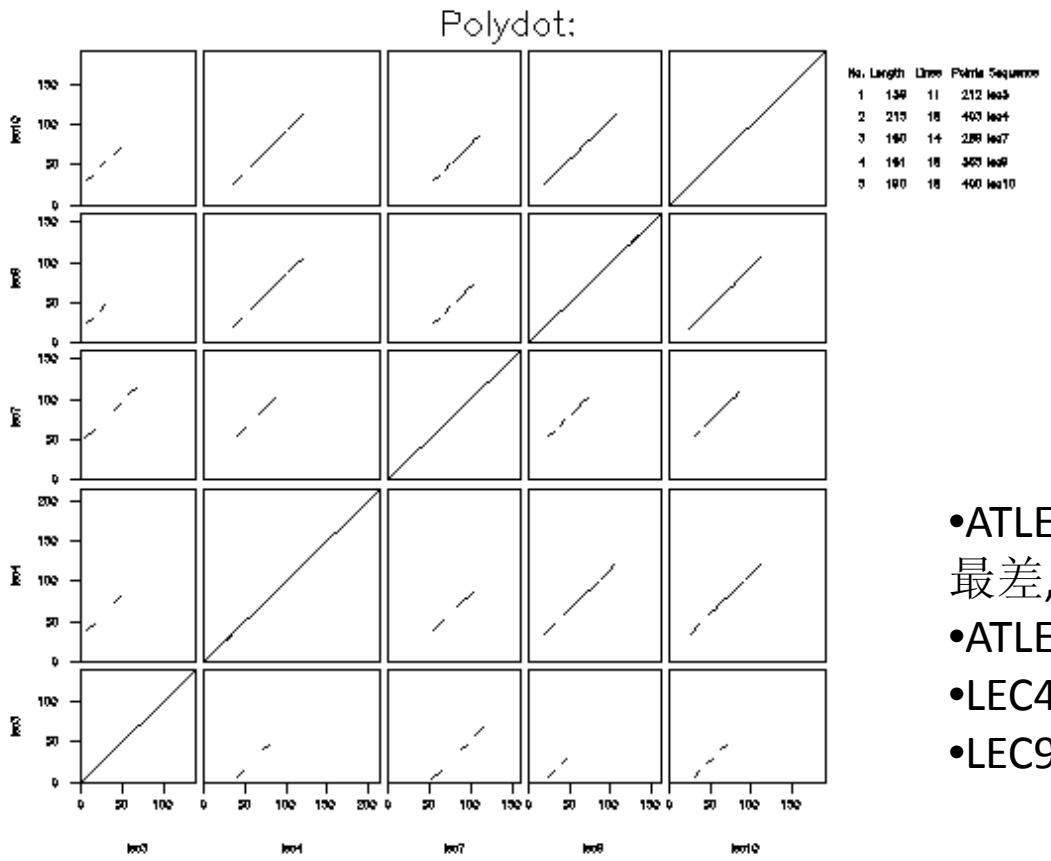
- 在所有的HAP3家族的基因中，仅有ATLEC1(NFYB9_ARATH)、ATLEC2(NFYB6_ARATH)和ATLEC3(NFYB4_ARATH)是经过证实在蛋白水平上存在的，其他只是在转录水平上证明是存在的。

序列特征分析

Name	Brief	TAIR ID	RefSeq ID	N	GenPet ID	L	Group
atlec3a	LEC3	AT1G09030.1	NM_100774.1	1	NP_172377.1	139	G1
atlec4a	LEC4	AT2G13570.1	NM_126937.1	1	NP_178981.1	215	G1
atlec7a	LEC7	At2g47810.1	NM_130348.2	1	NP_182302.1	160	G1
atlec9a	LEC9	AT4G14540.1	NM_117534.2	1	NP_193190.1	161	G1
atlec10a	LEC10	At5g47640.1	NM_124138.2	1	NP_199575.1	190	G1
atlec1a	LEC1	AT1G21970.1	NM_102046.4	2	NP_173616.2	238	G2
atlec2a	LEC2	AT5G47670.1	NM_124141.3	2	NP_199578.2	234	G2
atlec5a	LEC5	AT2G37060.1	NM_001036423.1	6	NP_001031500.1	173	G3
atlec6a	LEC6	AT2G38880.1	NM_129445.2	5	NP_030436.1	141	G3
atlec8a	LEC8	AT3G53340.1	NM_115194.2	6	NP_190902.2	176	G3

根据外显子数目对蛋白分组产生的分组表

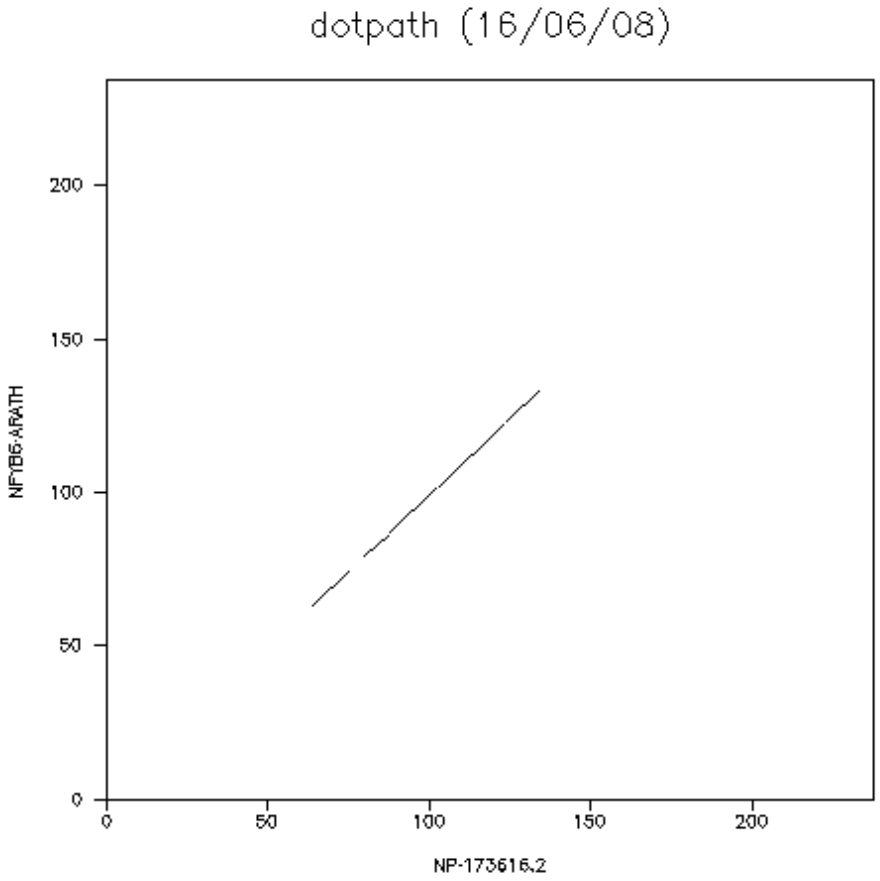
G1组序列分析



- ATLEC3序列与其他蛋白相似性最差,其中和LEC7最像
- ATLEC7稍好
- LEC4和LEC9,LEC1O相似性较高
- LEC9和LEC10相似性最高

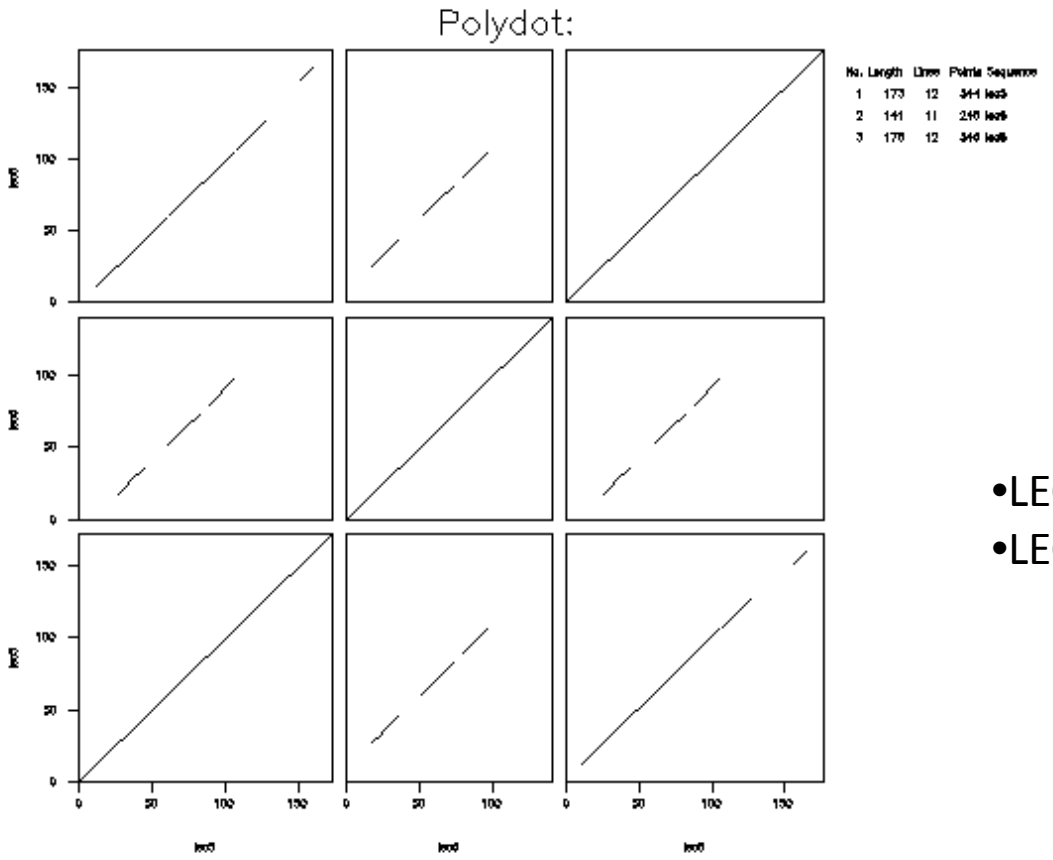
Polydot对于G1组中的5个序列进行分析（wordsize=6）

G2组序列分析



Dotpath对于G2组中的2个序列进行分析（wordsize=6）

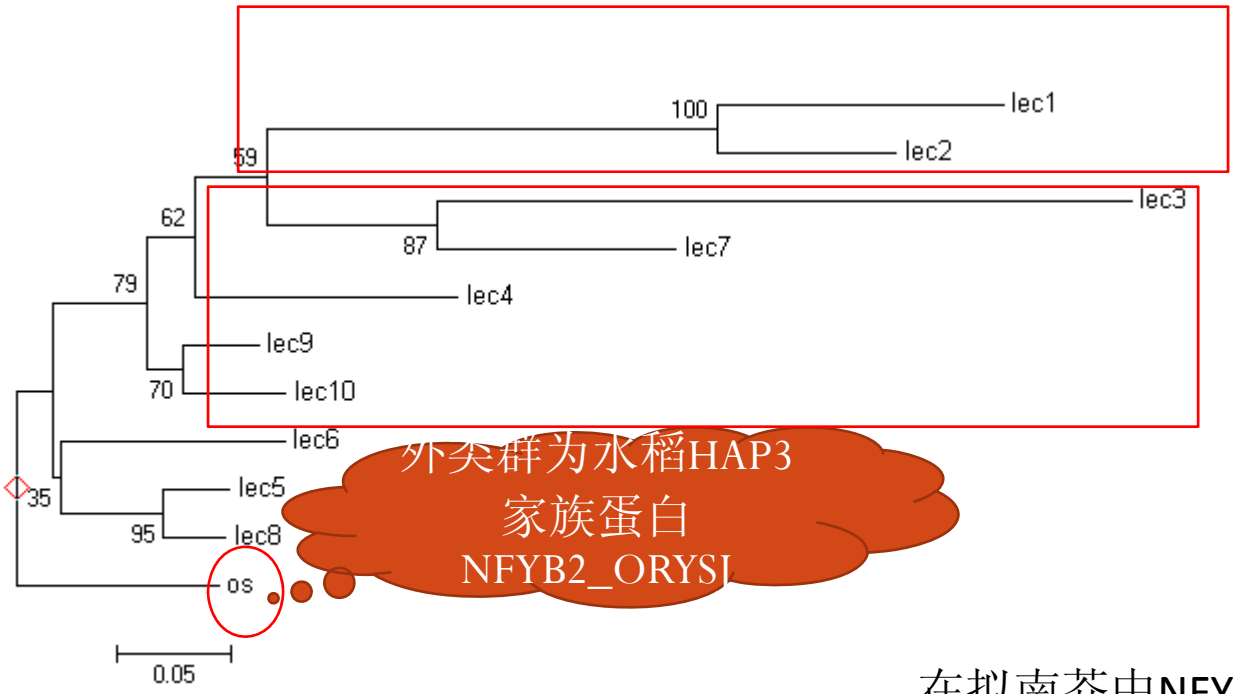
G3组序列分析



- LEC5和LEC8相似性最高
- LEC6与这两者相似性较低

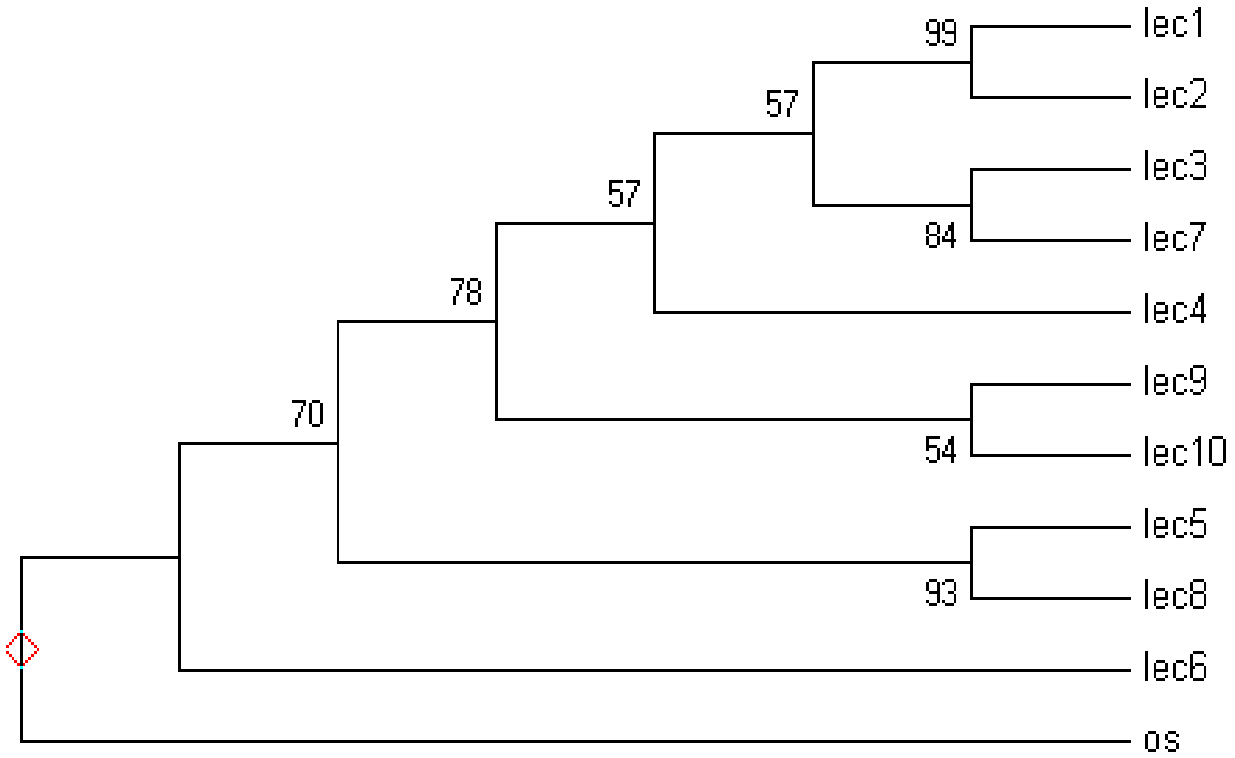
Polydot对于G3组中的2个序列进行分析（wordsize=6）

构建拟南芥HAP3家族的蛋白系统发育树



在拟南芥中NFYB2是LEC10

使用MEGA4.0对这10个蛋白建的NJ树



使用MEGA4.0对这10个蛋白建的MP树

结论1

- 拟南芥HAP3家族的10个蛋白构建的NJ树与系列分析结果很好地吻合了
- 在拟南芥中，HAP3家族的一部分基因有5-6个外显子，而另一部分则只有很少的1-2个外显子，这两部分蛋白似乎正向两个不同的方向变化，产生不同的作用
- 在只有1个外显子的蛋白中，又开始产生较多的外显子（LEC2还有一个可能有功能差异的一个外显子的剪接体），并且正在产生新的功能。

多序列比对寻找保守区域

```
CLUSTAL W (1.81) multiple sequence alignment


lec9            -------------MADSDNDS---GGHKDGGN----------------------ASTREQD
lec10           -------------MGDSDRDS---GGGQNGNNQNGQS----------------SLSPREQD
lec5            -------------MAESQAKSPGGCGSHESGGDQSPR----------------SLHVREQD
lec8            -------------MAESQTGG-GGGGSHESGGDQSPR----------------SLNVREQD
lec6            -------------MADTPSSPAGDGG--ESGG-----------------------SVREQD
lec4            ------------MTEESPEEDHGSPGVAETNPGSPSSKTNNN-----------NNNNKEQD
lec7            --MAGNYHSFQNPIPRYQNYNFGSSSSNHQHEHDGLVVVVEDQQQ-----EESMMVKEQD
lec3            ------------------------------------------------------MTDED
lec1            -------------MTSSVVVAGAGDKNNGIVVQ--QQPP----------------CVAREQD
lec2            MERGGFHGYRKLSVNNTTPSPPGLAANFLMAEGSMRPPEFNQPNKTSNGGEEECTVREQD
                                        .                               ::*

lec9            RFLPIANVSRIMKKALPANAKISKDAKETVQECVSEFISFITGEASDKCQREKRKTINGD
lec10           RFLPIANVSRIMKKALPANAKISKDAKETMQECVSEFISFVTGEASDKCQKEKRKTINGD
lec5            RFLPIANISRIMKRGLPANGKIAKDAKEIVQECVSEFISFVTSEASDKCQREKRKTINGD
lec8            RFLPIANISRIMKRGLPLNGKIAKDAKETMQECVSEFISFVTSEASDKCQREKRKTINGD
lec6            RYLPIANISRIMKKALPPNGKIGKDAKDTVQECVSEFISFITSEASDKCQKEKRKTVNGD
lec4            RFLPIANVGRIMKKVLPGNGKISKDAKETVQECVSEFISFVTGEASDKCQREKRKTINGD
lec7            RLLPIANVGRIMKNILPANAKVSKEAKETMQECVSEFISFVTGEASDKCHKEKRKTVNGD
lec3            RLLPIANVGRLMKQILPSNAKISKEAKQTVQECATEFISFVTCEASEKCHRENRKTVNGD
lec1            QYMPIANVIRIMRKTLPSHAKISDDAKETIQECVSEYISFVTGEANERCQREQRKTITAE
lec2            RFMPIANVIRIMRRILPAHAKISDDSKETIQECVSEYISFITGEANERCQREQRKTITAE
                : :****: *:*:. ** :.*:..::*: :***.:*:***:* **.::*::*:***:..:

lec9            DLLWAMTTLGFEDYVEPLKVYLQKYREVEGEKTTTAGRQGDKEG---GGGGGGAGSGSGG
lec10           DLLWAMTTLGFEDYVEPLKVYLQRFREIEGERTGLGRPQTGGEV---GEHQRDAVGDGGG
lec5            DLLWAMATLGFEDYMEPLKVYLMRYREMEGDTKGSAKGGDPNAK---KDGQSSQNGQFSQ
lec8            DLLWAMATLGFEDYIDPLKVYLMRYREMEGDTKGSGKGGESSAK---RDGQPSQVSQFSQ
lec6            DLLWAMATLGFEDYLEPLKIYLARYRELEGDNKGSGKSGDGSNR---DAG-----------
lec4            DIIWAITTLGFEDYVAPLKVYLCKYRDTEGEKVNSPKQQQQRQQ---QQQIQQQNHHNYQ
lec7            DICWAMANLGFDDYAAQLKKYLHRYRVLEGEKPNHHGKGGPKSS---PDN-----------
lec3            DIWWALSTLGLDNYADAVGRHLHKYREAERERTEHNKGSNDSGN---EKETNTRSDVQNQ
lec1            DILWAMSKLGFDNYVDPLTVFINRYREIETDRGSALRGEPPSLRQTYGGNGIGFHGPSHG
lec2            DVLWAMSKLGFDDYIEPLTLYLHRYRELEGERGVSCSAGSVSMT-----NGLVVKRPNGT
                *: **::.**:::*   :  .: ::*  * :           .:

lec9            APMYG-------GGMVTTMGHQFSHHFS--------------------------------
lec10           FYGGG-------GGMQYHQHHQFLHQQNHMYGATGGGSDSGGGAASGRTRT----------
lec5            LAHQG-----PYGNSQAQQHMMVPMPGTD--------------------------------
lec8            VPQQGSFSQGPYGNSQGS-NMMVQMPGTE--------------------------------
lec6            -------------GGVSGE-----EMPSW--------------------------------
lec4            FQEQDQ----NNNNMSCTSYISHHHPSPFLPVDHQPFPNIAFSPKSLQKQFPQQHDNNID
lec7            ------------------------------------------------------------
lec3            -------------STKFIRVVEKGSSSSAR-------------------------------
lec1            LPPPGPYGYGMLDQSMVMGGGRY-YQNG--SSGQDESSVGGGSSSSING--MPAFDHYGQ
lec2            MTEYGAYGP---VPGIHMAQYHYRHQNGFVFSGNEPNSKMSGSSSGASGARVEVFPTQ-Q
                   . . .                         . . .. .           .
```

CLUSTALW比对结果

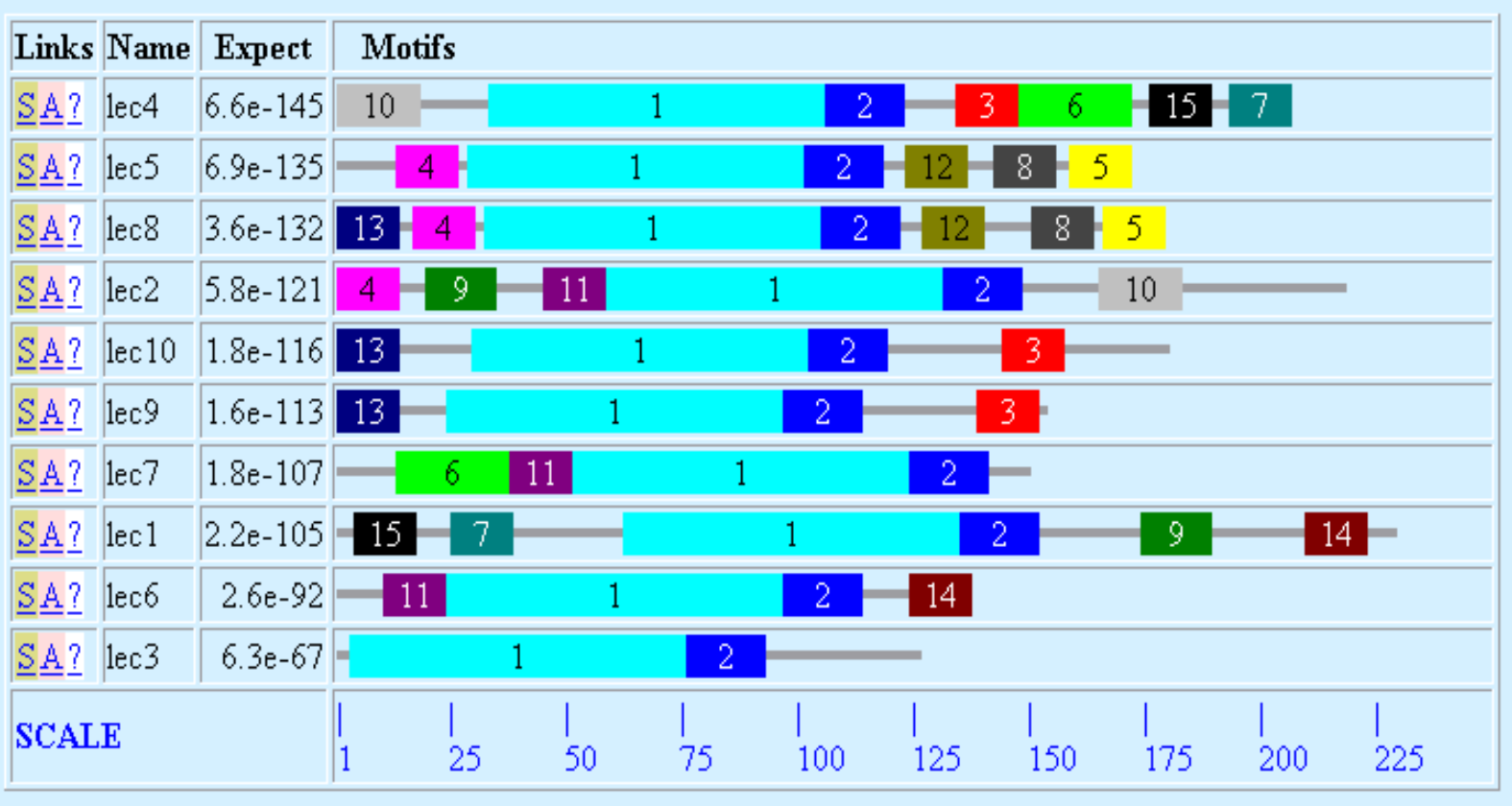
```
lec1            -EQDQYMPIANVIRIMRKTLPSHAKISDDAKETIQECVSEYISFVTGEAN
lec2            -EQDRFMPIANVIRIMRRILPAHAKISDDSKETIQECVSEYISFITGEAN
lec3            TDEDRLLPIANVGRLMKQILPSNAKISKEAKQTVQECATEFISFVTCEAS
lec4            -EQDRFLPIANVGRIMKKVLPGNGKISKDAKETVQECVSEFISFVTGEAS
lec5            -EQDRFLPIANISRIMKRGLPANGKIAKDAKEIVQECVSEFISFVTSEAS
lec6            -EQDRYLPIANISRIMKKALPPNGKIGKDAKDTVQECVSEFISFITSEAS
lec7            -EQDRLLPIANVGRIMKNILPANAKVSKEAKETMQECVSEFISFVTGEAS
lec8            -EQDRFLPIANISRIMKRGLPLNGKIAKDAKETMQECVSEFISFVTSEAS
lec9            -EQDRFLPIANVSRIMKKALPANAKISKDAKETVQECVSEFISFITGEAS
lec10           -EQDRFLPIANVSRIMKKALPANAKISKDAKETMQECVSEFISFVTGEAS


lec1            ERCQREQRKTITAEDILWAMSKLGFDNYVDPLTVFINRYREIETD-R---
lec2            ERCQREQRKTITAEDVLWAMSKLGFDDYIEPLTLYLHRYRELEGE-R---
lec3            EKCHRENRKTVNGDDIWWALSTLGLDNYADAVGRHLHKYREAERE-R---
lec4            DKCQREKRKTINGDDIIWAITTLGFEDYVAPLKVYLCKYRDTEGE-K---
lec5            DKCQREKRKTINGDDLLWAMATLGFEDYMEPLKVYLMRYREMEGDTK---
lec6            DKCQKEKRKTVNGDDLLWAMATLGFEDYLEPLKIYLARYRELEGDNK---
lec7            DKCHKEKRKTVNGDDICWAMANLGFDDYAAQLKKYLHRYRVLEGE-KPN-
lec8            DKCQREKRKTINGDDLLWAMATLGFEDYIDPLKVYLMRYREMEGDTK---
lec9            DKCQREKRKTINGDDLLWAMTTLGFEDYVEPLKVYLQKYREVEGE-KTT-
lec10           DKCQKEKRKTINGDDLLWAMTTLGFEDYVEPLKVYLQRFREIEGE-R--T
```

POA比对结果

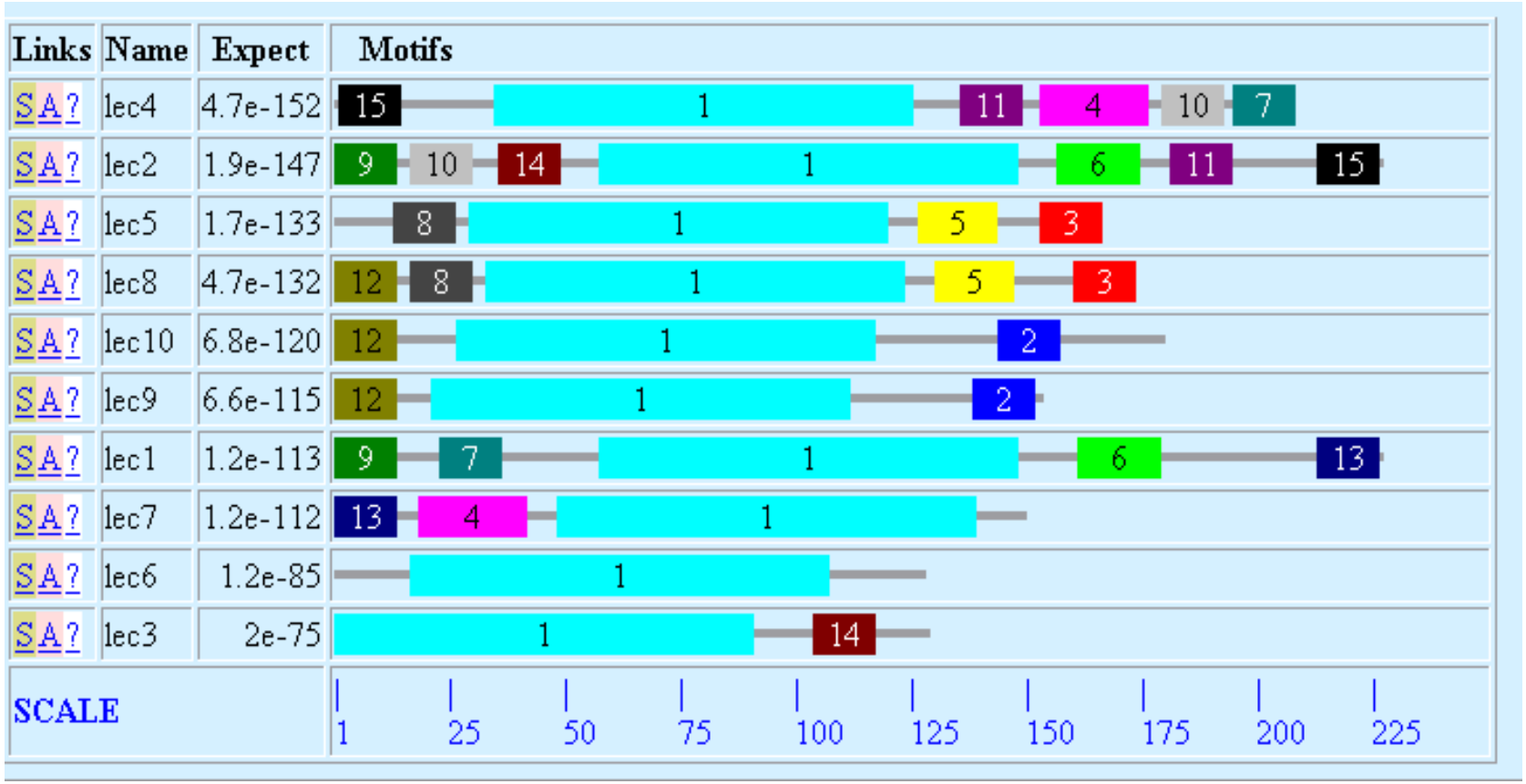
结果：在序列长度接近的时候，CLUSTALW的比对结果更为理想

MEME分析寻找保守模体

http://meme.sdsc.edu/



参数设置：Number of different motifs: 15， Maximum number of sites: 10，
Minimum motif width: 10， Maximum motif width: 80



参数设置：Number of different motifs: 15， Maximum number of sites: 10，
Minimum motif width: 10， Maximum motif width: 100

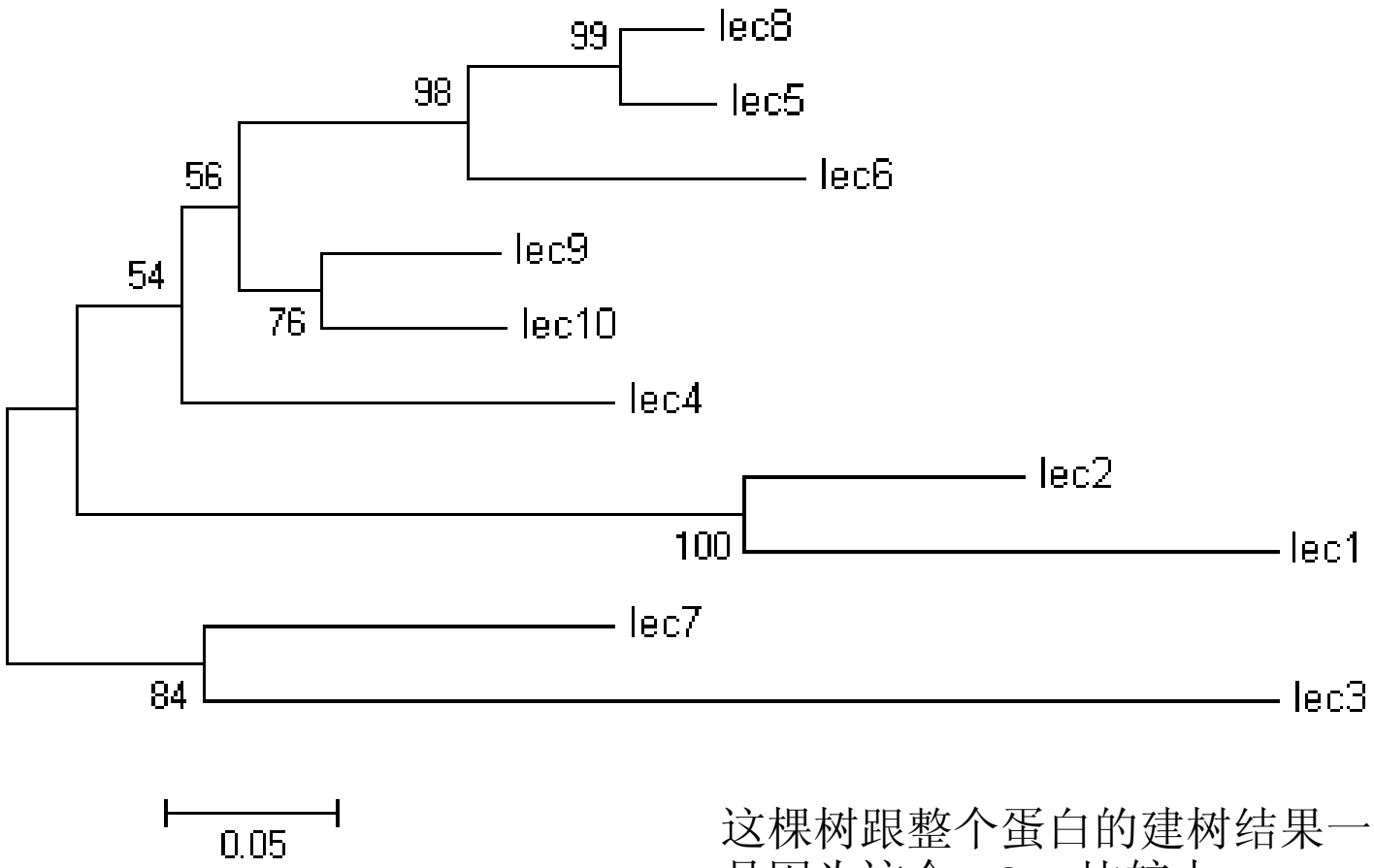
MOTIF 1的序列

>lec9 (start= 19)
TREQDRFLPIANVSRIMKKALPANAKISKDAKETVQECVSEFISFITGEASDKCQREKRKTINGDDLLWAMTTLGFEDYVEPLKVYLQKYREVEGEKTTT
>lec10 (start= 25)
PREQDRFLPIANVSRIMKKALPANAKISKDAKETMQECVSEFISFVTGEASDKCQKEKRKTINGDDLLWAMTTLGFEDYVEPLKVYLQRFREIEGERTGL
>lec8 (start= 27)
VREQDRFLPIANISRIMKRGLPLNGKIAKDAKETMQECVSEFISFVTSEASDKCQREKRKTINGDDLLWAMATLGFEDYIDPLKVYLMRYREMEGDTKGS
>lec5 (start= 28)
VREQDRFLPIANISRIMKRGLPANGKIAKDAKEIVQECVSEFISFVTSEASDKCQREKRKTINGDDLLWAMATLGFEDYMEPLKVYLMRYREMEGDTKGS
>lec6 (start= 19)
VREQDRYLPIANISRIMKKALPPNGKIGKDAKDTVQECVSEFISFITSEASDKCQKEKRKTVNGDDLLWAMATLGFEDYLEPLKIYLARYRELEGDNKGS
>lec4 (start= 34)
NKEQDRFLPIANVGRIMKKVLPGNGKISKDAKETVQECVSEFISFVTGEASDKCQREKRKTINGDDIIWAITTLGFEDYVAPLKVYLCKYRDTEGEKVNS
>lec7 (start= 49)
VKEQDRLLPIANVGRIMKNILPANAKVSKEAKETMQECVSEFISFVTGEASDKCHKEKRKTVNGDDICWAMANLGFDDYAAQLKKYLHRYRVLEGEKPNH
>lec2 (start= 56)
VREQDRFMPIANVIRIMRRILPAHAKISDDSKETIQECVSEYISFITGEANERCQREQRKTITAEDVLWAMSKLGFDDYIEPLTLYLHRYRELEGERGVS
>lec1 (start= 57)
AREQDQYMPIANVIRIMRKTLPSHAKISDDAKETIQECVSEYISFVTGEANERCQREQRKTITAEDILWAMSKLGFDNYVDPLTVFINRYREIETDRGSA
>lec3 (start= 1)
MTDEDRLLPIANVGRLMKQILPSNAKISKEAKQTVQECATEFISFVTCEASEKCHRENRKTVNGDDIWWALSTLGLDNYADAVGRHLHKYREAERERTEH

对MOTIF1 进行多序列比对

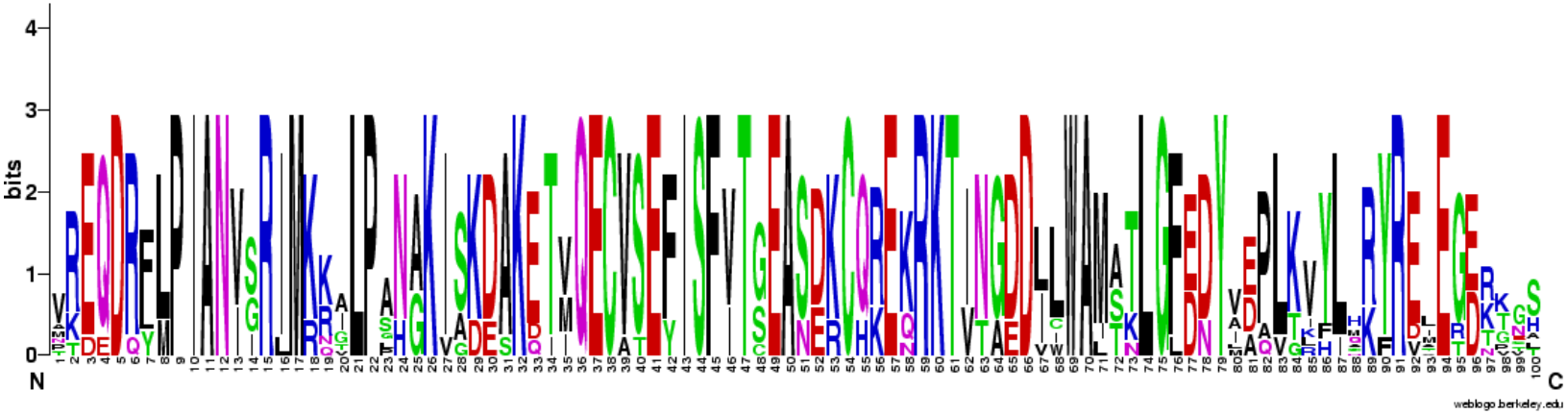
```
!Domain=Data;
#lec9_(_start  TREQDRFLPI ANVSRIMKKA LPANAKISKD AKETVQECVS EFISFITGEA
#lec10_(_start P......... .......... .......... ....M..... .....V....
#lec8_(_start  V......... ..I.....RG ..L.G..A.. ....M..... .....V.S..
#lec5_(_start  V......... ..I.....RG ....G..A.. ...I...... .....V.S..
#lec6_(_start  V.....Y... ..I.....  ..P.G..G.. ..D.....   .....S..
#lec4_(_start  NK.....   ...G.....V ..G.G..... .......... .....V....
#lec7_(_start  VK....L... ...G....NI ......V..E ....M..... .....V....
#lec2_(_start  V......M.. ...I...RRI ...H....D. S...I..... .Y.....
#lec1_(_start  A....QYM.. ...I...R.T ..SH....D. ....I..... .Y...V....
#lec3_(_start  MTDE..L... ...G.L..QI ..S......E ..Q.....AT .....V.C..

#lec9_(_start  SDKCQREKRK TINGDDLLWA MTTLGFEDYV EPLKVYLQKY REVEGEKTTT
#lec10_(_start .....K.... .......... .......... ........RF ..I...R.GL
#lec8_(_start  .......... .......... .A.......I D......MR. ..M..DTKGS
#lec5_(_start  .......... .......... .A.......M .......MR. ..M..DTKGS
#lec6_(_start  .....K.... .V.....   .A.......L ....I..AR. ..L..DNKGS
#lec4_(_start  .......... ......II.. I.....    A......C.. .DT....VNS
#lec7_(_start  ....HK.... .V....IC.. .AN...D..A AQ..K..HR. .VL....PNH
#lec2_(_start  NER....Q.. ..TAE.V... .SK...D..I ...TL..HR. ..L...RGVS
#lec1_(_start  NER....Q.. ..TAE.I... .SK...DN.. D..T.FINR. ..I.TDRGSA
#lec3_(_start  .E..H..N.. .V....IW.. LS...LDN.A DAVGRH.H.. ..A.R.R.EH
```

使用MEGA4.0对这个Motif建NJ树

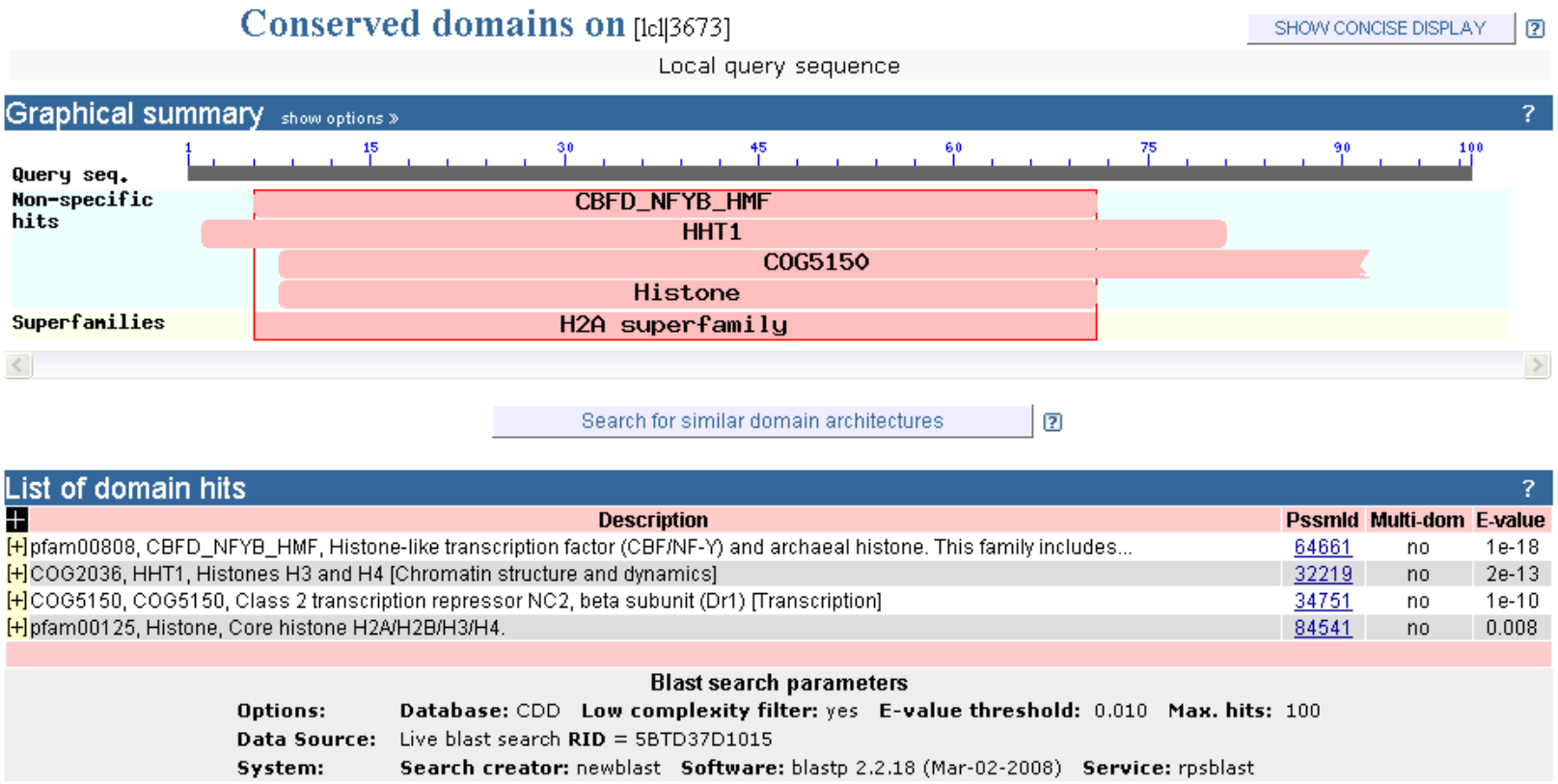


这棵树跟整个蛋白的建树结果一致，可能是因为这个MOTIF比较大

WebLogo显示该motif的保守性



在NCBI中通过BLAST寻找这个结构域的同源

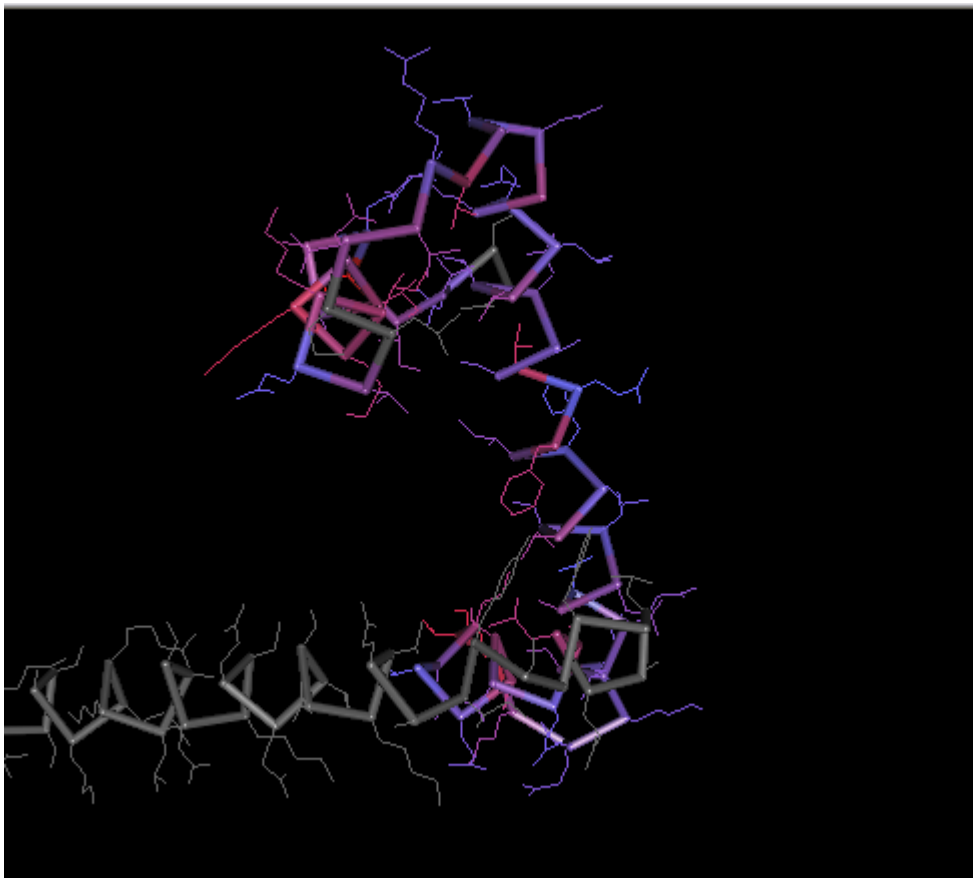


	Description	PssmId	Multi-dom	E-value
[+]	pfam00808, CBFD_NFYB_HMF, Histone-like transcription factor (CBF/NF-Y) and archaeal histone. This family includes...	64661	no	1e-18
[+]	COG2036, HHT1, Histones H3 and H4 [Chromatin structure and dynamics]	32219	no	2e-13
[+]	COG5150, COG5150, Class 2 transcription repressor NC2, beta subunit (Dr1) [Transcription]	34751	no	1e-10
[+]	pfam00125, Histone, Core histone H2A/H2B/H3/H4.	84541	no	0.008

Blast search parameters

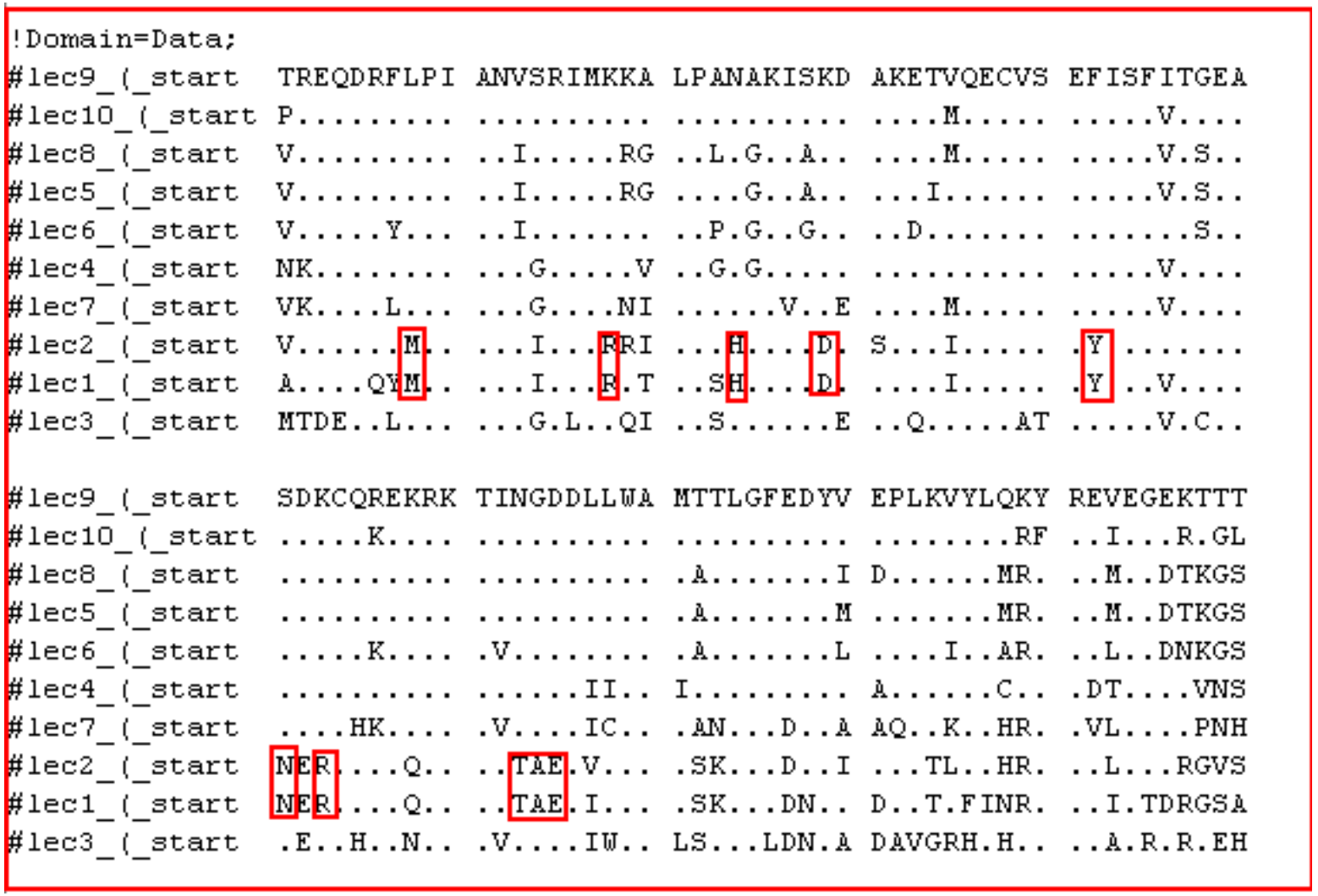
Options: **Database:** CDD **Low complexity filter:** yes **E-value threshold:** 0.010 **Max. hits:** 100
Data Source: Live blast search **RID** = 5BTD37D1015
System: **Search creator:** newblast **Software:** blastp 2.2.18 (Mar-02-2008) **Service:** rpsblast

Pfam中寻找这个结构域的结构和功能



- CCAAT结合因子 (CBF)是一个哺乳动物转录因子，结合在很多种基因的CCAAT模体上，
- 这个转录因子是一个由A.B两个不同的蛋白亚基共同作用形成的一个蛋白。
- 没有DNA存在时，这两个亚基也是结合在一起的。
- A、B亚基中的保守区域对蛋白结构起到非常重要的作用。
- 推测这10个HAP3家族的基因编码的为B亚基

寻找仅仅LEC1突变体含有的氨基酸变异



```
!Domain=Data;
#lec9_(_start  TREQDRFLPI ANVSRIMKKA LPANAKISKD AKETVQECVS EFISFITGEA
#lec10_(_start P......... .......... .......... ....M..... .....V....
#lec8_(_start  V......... ..I.....RG ..L.G..A.. ....M..... .....V.S..
#lec5_(_start  V......... ..I.....RG ....G..A.. ...I...... .....V.S..
#lec6_(_start  V.....Y... ..I..... ..P.G..G.. ..D..... .......S..
#lec4_(_start  NK........ ...G.....V ..G.G..... .......... .....V....
#lec7_(_start  VK....L... ...G....NI ......V..E ....M..... .....V....
#lec2_(_start  V......M.. ...I...RRI ...H....D. S...I..... .Y.......
#lec1_(_start  A....QYM.. ...I...R.T ..SH....D. ....I..... .Y..V....
#lec3_(_start  MTDE..L... ...G.L..QI ..S......E ..Q.....AT .....V.C..

#lec9_(_start  SDKCQREKRK TINGDDLLWA MTTLGFEDYV EPLKVYLQKY REVEGEKTTT
#lec10_(_start .....K.... .......... .......... ........RF ..I...R.GL
#lec8_(_start  .......... .......... .A.......I D......MR. ..M..DTKGS
#lec5_(_start  .......... .......... .A.......M .......MR. ..M..DTKGS
#lec6_(_start  .....K.... .V........ .A.......L ....I..AR. ..L..DNKGS
#lec4_(_start  .......... ......II.. I......... A......C.. .DT....VNS
#lec7_(_start  ....HK.... .V....IC.. .AN...D..A AQ..K..HR. .VL....PNH
#lec2_(_start  NER....Q.. ..TAE.V... .SK...D..I ...TL..HR. ..L...RGVS
#lec1_(_start  NER....Q.. ..TAE.I... .SK...DN.. D..T.FINR. ..I.TDRGSA
#lec3_(_start  .E..H..N.. .V....IW.. LS...LDN.A DAVGRH.H.. ..A.R.R.EH
```

该Motif对蛋白功能的作用

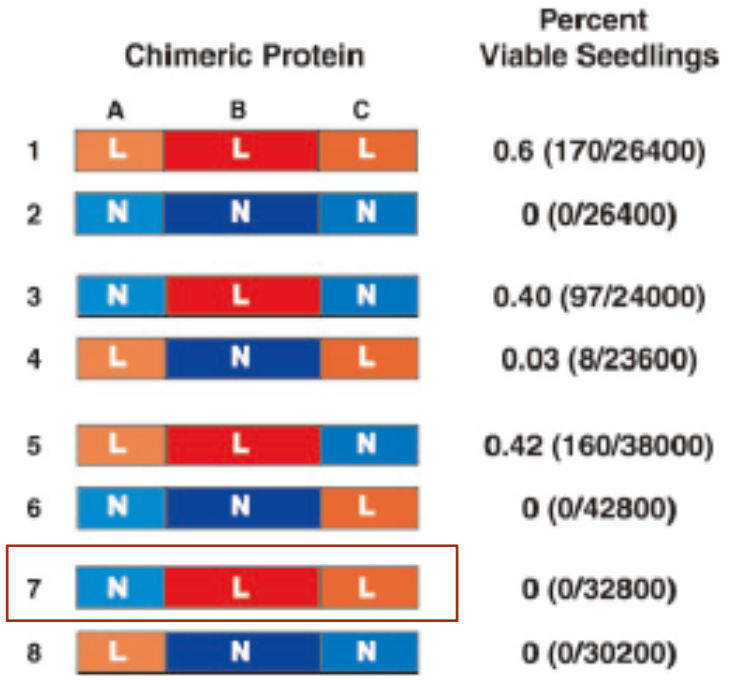


Fig. 2. LEC1 B domain is necessary and sufficient for its activity in embryogenesis. The diagrammatic representation shows LEC1 (construct 1), the non-LEC1-type AHAP3 subunit, At4g14540 (construct 2), and chimeric proteins containing A, B, and C domains from LEC1 (L) or At4g14540 (N). Constructs encoding these chimeric proteins were transformed into *lec1-1* null mutants and, after drying of the T_1 seeds, the number of viable seedlings generated from seeds tested (shown in parentheses) was determined. Transformation experiments were repeated multiple times with similar results, and the total values for all experiments are reported.

B结构域也就是我们找到的motif对于LEC1基因的功能是必须的。

充分？

Lee et al，2003

实验结果

Table 1. Identification of a LEC1 B domain residue required for LEC1 activity

Amino acid substitution in LEC1*	Percent viable seedlings†
M34L	0.8 (110 of 14,400)
R44K	0.6 (82 of 12,800)
H50N	0.7 (176 of 24,000)
D55K	0.01 (5 of 55,600)
N77S	0.4 (48 of 12,000)
Q84K	0.4 (75 of 20,800)
TAE89-91NGD	0.4 (73 of 16,800)
K99T	0.4 (95 of 22,080)

*Mutant forms of LEC1 are designated with the wild-type amino acid and its position within LEC1, followed by the amino acid that was inserted. cDNA clones encoding LEC1 with the indicated single or triple amino acid substitutions were fused with *LEC1* 5′ and 3′ flanking sequences and transferred into *lec1-1* mutants.

†T1 *lec1-1* seeds were collected and dried for 2 weeks at 28°C before germination tests were performed. Values obtained reflect the ability of the construct to suppress the *lec1* mutation and the transformation efficiency. For the wild-type *LEC1* gene, this value was 0.6 (see Fig. 2). The total number of viable seedlings obtained and seeds tested in all experiments are given in parentheses. Independent replicates of transformation experiments gave values similar to the percentages reported here.

Table 2. A single amino acid substitution confers LEC1 activity to a non-LEC1-type B domain

Construct*	Percent viable seedlings†
L^A-N^B-L^C	0.03 (8 of 23,600)‡
L^A-N^B-L^C (K55D)	0.23 (70 of 30,400)

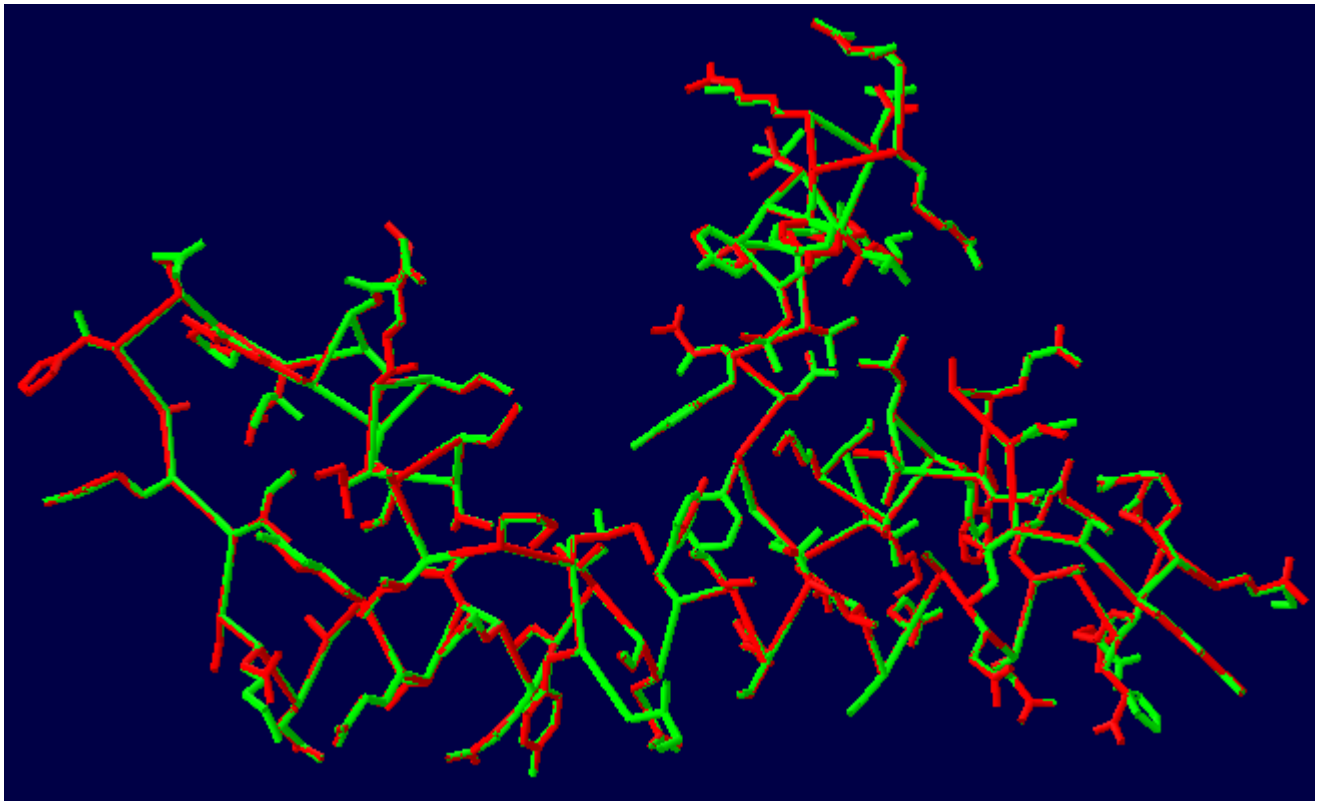
*L^A-N^B-L^C consists of A and C domains from LEC1 and the B domain from the non-LEC1-type AHAP3, At4g14540 (see Fig. 2).

†Percent viable seedlings was determined as described. Consistent values were obtained from independently repeated experiments. The numbers of seedlings and seeds represent the sum of replicate experiments.

‡Data from Fig. 2.

Lee et al，2003

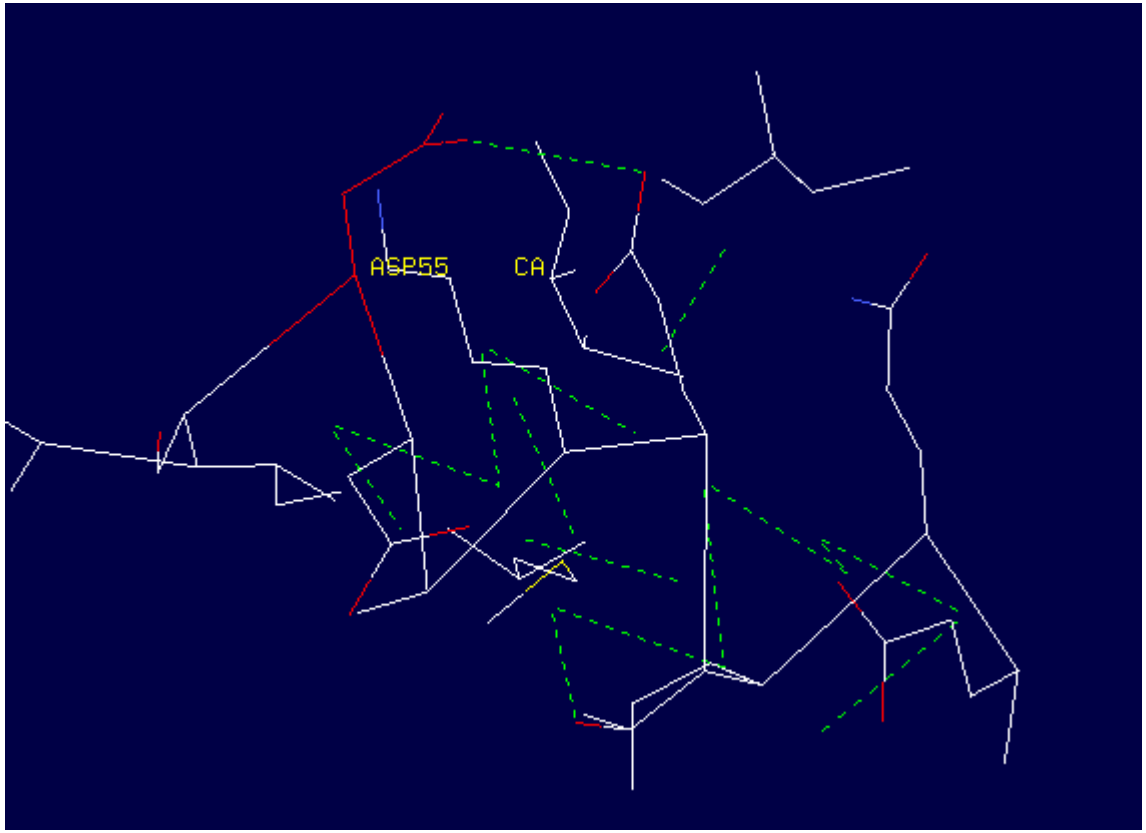
SwissModel中预测的LEC1蛋白部分结构



以NFYB_HUMAN为模板建立的LEC1部分结构模型

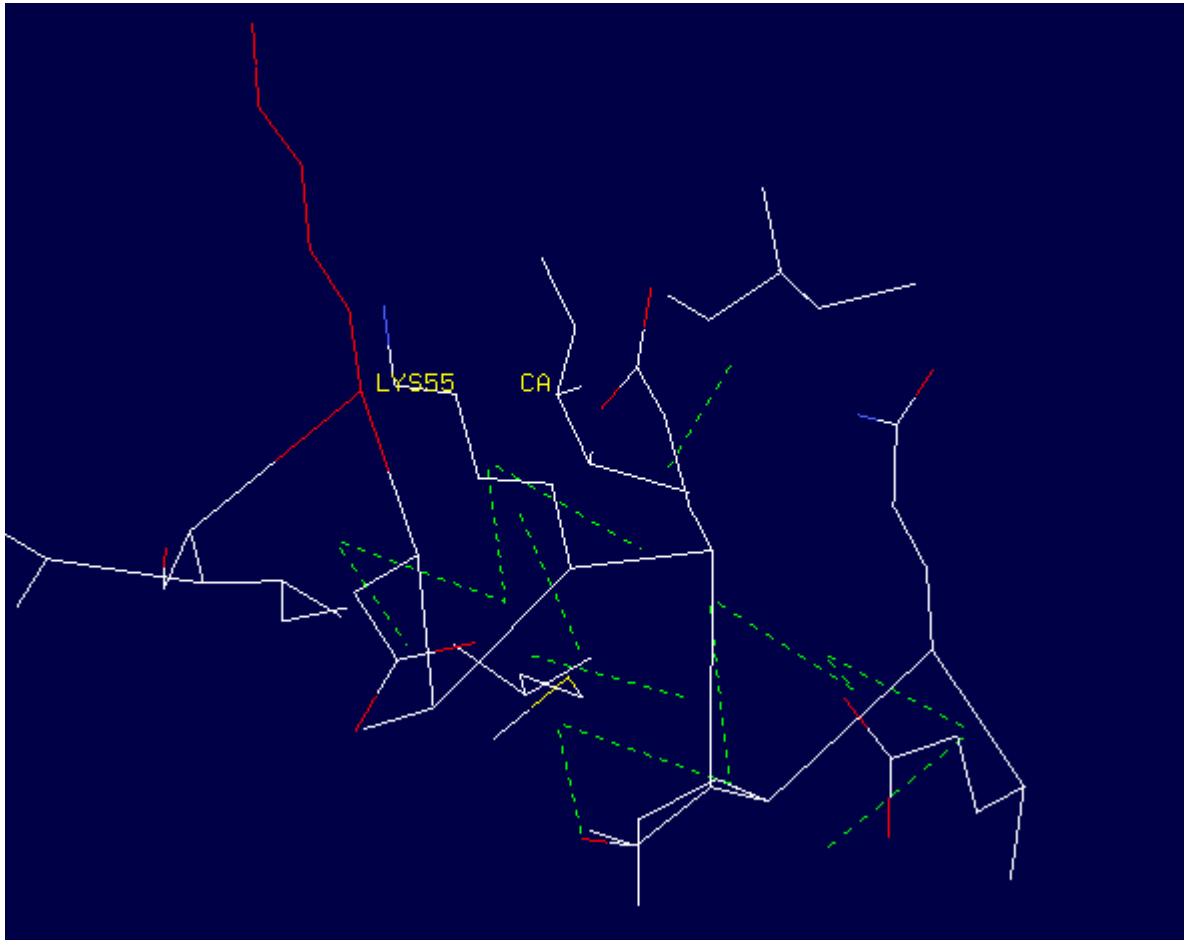
RMS:0.07

点突变分析蛋白结构



未突变前，Asp55和Glu59形成氢键

点突变分析蛋白结构



Asp55Lys突变后，Lys55和Glu59不形成氢键

结论2

- HAP3家族的这10个蛋白的保守模体在整个蛋白的功能中起到了关键性的作用
- 关键基因的关键性结构域的变化，导致蛋白结构的变化
- 蛋白结构的变化导致了新的功能的产生
- 功能发生改变的方式需要蛋白结构信息来进行解释

讨论

- 建立系统发育树时外类群的选择还需要有更多的考虑
- LEC1基因似乎是个刚刚进化出来的基因，是研究基因进化的很好地材料
- D55K似乎对蛋白产生新的功能起到了关键性的作用，对于这个蛋白结构的分析是很有意义的一件工作

致谢

感谢罗老师的教导和全组同学的共同努力
谢谢！