基因家族分析培训

学习文档

目录

一、基因家族数据下载

基因家族成员鉴定可以采用 blastp 和 hmmer3 鉴定，而这两个工具需要提供目前已知的基因家族的蛋白序列和研究物种的全基因组序列信息。那么如何获取这些信息呢?

比较直接的方法是在一些网站上下载他们已经整理好的某个家族序列即可，如 TAIR、PlantTFDB、NCBI 和 Uniprot 等，然后利用 blastp 进行鉴定；同时也可以在 Pfam 上下载该家族特有的结构域信息，使用 HMMER3 进行进一步筛选和鉴定。

基因组信息，一般通过基因组数据库下载，本处主要介绍三个常用的基因组数据下载网站，用于下载文献中报道的特定物种的基因家族序列和自己研究物种的全基因组序列。

1.1 基础知识

1、FASTA 格式。

用于表示**核酸序列或多肽序列**的格式。其中核酸或氨基酸均以单个字母来表示，且允许在序列前添加序列名及注释。该格式已成为生物信息学领域的一项标准。

FASTA 格式首先以大于号“>”开头，接着是序列的标识符（ID），然后是序列的描述信息（非必须），最后换行后是序列信息，序列中允许换行，空行，直到下一个大于号，表示该序列的结束。如：

```
>LOC_Os01g01019.1 gene=LOC_Os01g01019
MEEAGERDADETHAWSGTASPAALWKTVASSAAMLKLALAMISAAFRTTPFSMSMQLCPNATMSLHSPSI
FDVVSSITPIMSCIINNRLVAEKAGATMQRWRAHSSPSAMTRPLPNMGMRLSSYDIVCQLAHLHFSHVCC
LV.
>LOC_Os01g01030.1 gene=LOC_Os01g01030
MDSIRRRSAGGILGILFLVLLRWAGAGDPYAYYEWEVSYVWGAPLGGVKKQEAIGINGQLPGPALNVTTN
WNLVVNVRNGLDEPLLLTWHGVQQRKSPWQDGVGGTNCGIPPGWNWTYQFQVKDQVGSFFYAPSTALHRA
AGGYGAITINNRDVIPLPFPLPDGGDITLFLADWYARDHRALRRALDAGDPLGPPDGVLINALGPYRYND
```

2、GFF3 格式

为 general feature format 缩写，目前采用的是 version 3，即我们常说的 gff3 文件。该文件常用来对基因组进行注释，表示基因，外显子，CDS，UTR 等在基因组上的位置。

seqname - 序列的 ID，可以是染色体的 ID 也可以是 Scaffold 或者 Contig 的 ID。

source - 产生此文件的软件，如 Stringtie 产生的则为 Stringtie，Cufflinks 产生的则为 Cufflinks，不知道的使用点 "." 表示。

feature - 后面 start 和 end 之间区域代表的特征，如果此区域是基因，则此处为 gene，如果是外显子，则为 exon，如果是转录本，则为 transcript，如果是非编码 RNA 则为 lncRNA，如果是重复序列，则为 TE，等等，主要表明这一块区域的特征。

start -上述 feature 的在序列上的起始位置。

end - 上述 feature 的在序列上的终止位置。

score - 一个浮点数值，也可以为点 "." 。有值的时候代表上述 feature 的可靠性。因为无论是 gene 还是 mRNA，都是基于预测差生的，因而必然会有一个值来衡量预测准确性。

strand - + (forward)或者 - (reverse)，代表上述 feature 是位于正链还是负链上。

frame - 内含子相位，只能为'0', '1' or '2'，或者为点 "." 。 '0' 代表 feature 起始碱基为三联体密码子的第一个碱基, '1' 代表三联体密码子的第 1 个碱基, 2 代表第 2 个碱基。

attribute -备注列。主要备注该 feature 的一些信息，常见的是 gene 或者 transcript 等的 ID 信息以及 FPKM 值等，多个备注信息之间通常用分号分隔。

1.2 TAIR 和 TWGFD 数据库下载家族成员

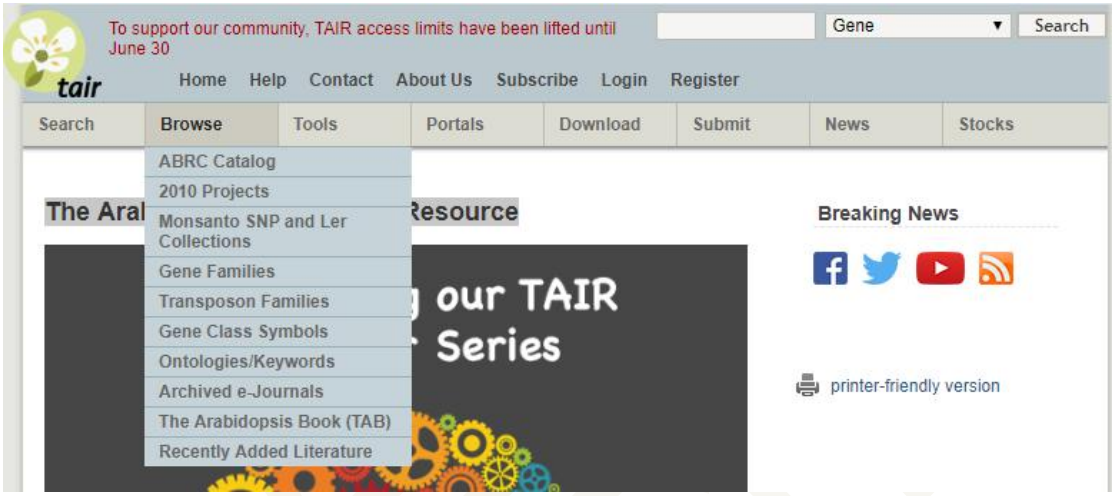
A、TAIR 数据库

拟南芥信息资源网站（TAIR，The Arabidopsis Information Resource）是存储高等植物拟南芥遗传和分子生物学数据的数据库。TAIR 提供包括完整的基因组序列以及基因结构，基因产物信息，基因表达，DNA 种子库，基因组图谱，遗传和物理标记，发表文献以及有关

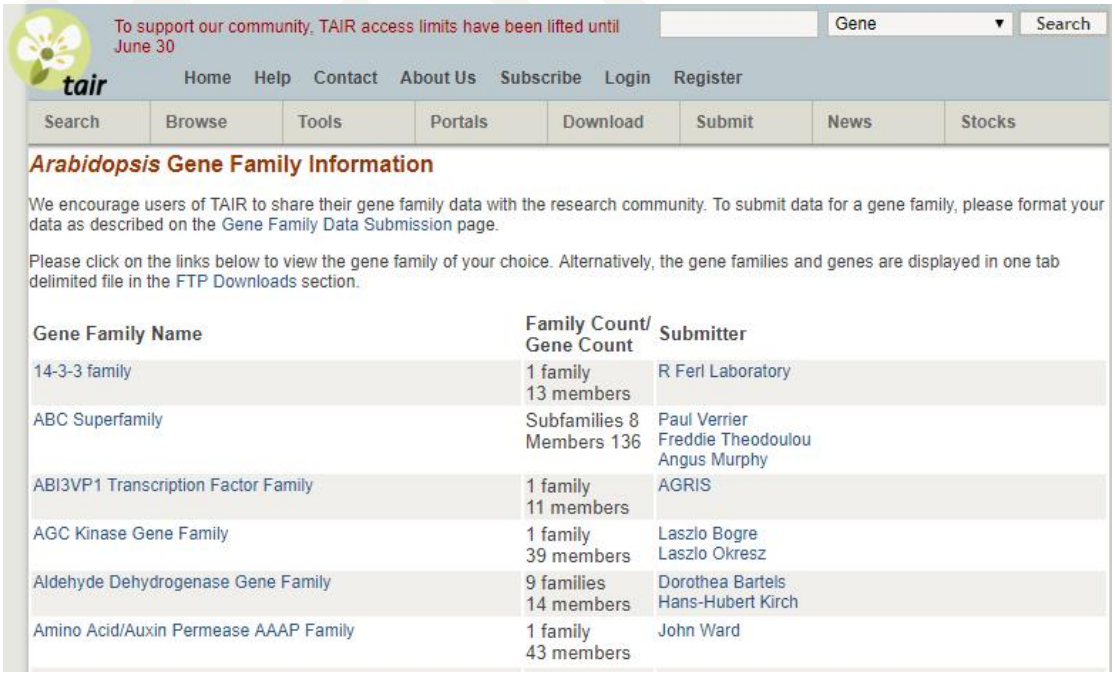
拟南芥研究社区的信息。每周都会根据最新出版的研究文献和社区数据来更新基因产物的功能数据。网址：https://www.arabidopsis.org/。

涉及到的基因家族信息的具体使用方法如下：

1、打开基因家族页面。从 Browse->Gene Families



2、出现如下界面。这个界面包含了目前（2017 年 12 月 12 日前）在拟南芥中阐明的基因家族名称，成员数目和递交者。其中有些超家族(Super family)可以进一步划分成多个亚家族（Sub family）。



3、以 14-3-3 family 为例，我们点击打开对应的链接如下：

***Arabidopsis thaliana* 14-3-3 protein gene family**

Source:
- Website: 14-3-3 Proteins
- **Sehnke, P.C., Chung, H.J., Wu, K., Ferl, R. J.** (2001) Regulation of starch accumulation by granule-associated plant 14-3-3 proteins. *PNAS* **98: 765-770**
- **DeLille, J. M., Sehnke, P. C., Ferl, R. J.** (2001) The arabidopsis 14-3-3 family of signaling regulators. *PLANT PHYSIOLOGY* **126: 35-38**
- **Rosenquist, M., Sehnke, P., Ferl, R. J., Sommarin, M., Larsson, C.** (2000) Evolution of the 14-3-3 Protein Family: Does the Large Number of Isoforms in Multicellular Organisms Reflect Functional Specificity?. *JOURNAL OF MOLECULAR EVOLUTION* **51: 446-458**
- **Wu, K., Rooney, M. F., Ferl, R. J.** (1997) The Arabidopsis 14-3-3 multigene family. *PLANT PHYSIOLOGY* **114: 1421-1431**

Gene Family Criteria:

Contact: R Ferl Laboratory

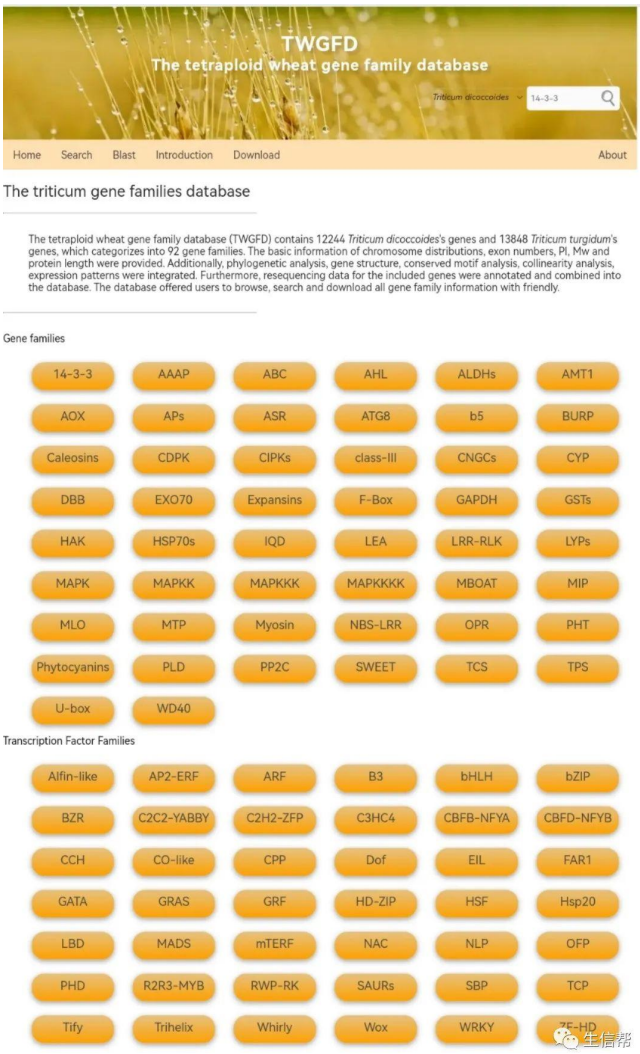
Gene Family Name:	Gene Name:	Protein Name:	Genomic Locus:	Accession:	TIGR Annotation:
14-3-3 proteins	GRF1	Chi	F23J3.30 At4g09000	L09112	14-3-3 protein GF14chi (grf1)
	GRF2	Omega	F3F9.16 At1g78300	M96855	14-3-3 protein GF14omega (grf2)
	GRF3	Psi	MXI10.21 At5g38480	L09110	14-3-3 protein GF14psi (grf3/RCI1)
	GRF4	Phi	T32G9.30 At1g35160	L09111	14-3-3 protein GF14phi (grf4)
	GRF5	Upsilon	F1N13.190 At5g16050	L09109	14-3-3 protein GF14upsilon (grf5)
	GRF6	Lambda	F12B17.200 At5g10450	U68545	14-3-3 protein GF14lambda (grf6/AFT1)
	GRF7	Nu	F16B3.15 At3g02520	U60445	14-3-3 protein GF14nu (grf7)
	GRF8	Kappa	MNA5.16 At5g65430	U36447	14-3-3 protein GF14kappa (grf8)
	GRF9	Mu	F14N22.14 At2g42590	U60444	14-3-3 protein GF14mu (grf9)
	GRF10	Epsilon	T16E15.8 At1g22300	U36446	14-3-3 protein GF14epsilon (grf10)
	GRF11	Omicron	F21H2.3 At1g34760	AF323920	14-3-3 protein GF14omicron (grf11)
	GRF12	Iota	T1K7.15 At1g26480	AF335544	14-3-3 protein GF14iota (grf12)
	GRF13	Pi	T11I11.16 At1g78220		14-3-3 protein (grf13), putative

4、提取蛋白序列。准备好以 AT 开始的基因编号，然后可以使用（03.scripts/getseq_fromID.pl）直接提取。

B 、TWGFD 单子叶基因家族数据库

四倍体小麦基因家族数据库（TWGFD，http://triticumgfdb.com/index.html）包含了 12244 个 Triticum dicoccoides 基因和 13848 个 Triticum turgidum 基因，这些基因被分类进了 92 个基因家族。数据库提供了染色体分布、外显子数目、等电点（PI）、分子量（Mw）和蛋白长度的基本信息。此外，数据库还整合了系统发育分析、基因结构、保守模体分析、同源性分析和表达模式等内容。此外，还将所包含基因的重测序数据进行了注释，并加入

到了数据库中。该数据库以用户友好的方式提供了浏览、搜索和下载所有基因家族信息的功能。



1.3 PlantTFDB 数据库下载家族成员

该数据从 165 种植物中鉴定了 320，370 个转录因子，并将其分为 58 个家族，包含了绿色植物的主要谱系。主要用于转录因子家族的分析，尤其是水稻和拟南芥的转录因子家族成员，可以直接作为已知的，用于未知物种的该家族成员的鉴定。

该网站为了给 TF 家族提供全面的信息，对每个家族都做了简要的介绍和主要的参考资料。对每个已识别的 TF 进行了全面的注释，包括功能域、基因本体（GO）、功能描述、结合基序、参考文献等。网站是 http://planttfdb.gao-lab.org/index.php。

具体用法如下：

1、进入主页后，呈现下述界面。主要分为按物种查找本物种的所有基因家族和按照基因家族名字查找所有物种的分布。其中含有的物种列表如下：

绿藻（16 种）

轮藻（1 种）

地钱（1 种）

苔藓植物（2 种）

蕨类植物（1 种）

裸子植物（5 种）

被子植物基部群（木兰类）（1 种）

单子叶植物（38 种）

双子叶植物（100 种）

Browse by Species

open all | close all

Taxonomic Group (165 species) (G)-species with genome sequence

- Chlorophytae (16 species)
- Charophyta (1 species)
- Marchantiophyta (1 species)
- Bryophyta (2 species)
- Lycopodiophyta (1 species)
- Coniferophyta (5 species)
- Basal Magnoliophyta (1 species)
- Monocots (38 species)
- Eudicots (100 species)

Quick links: Arabidopsis thaliana, Glycine max, Oryza sativa, Populus trichocarpa, Solanum lycopersicum, Zea mays

Couldn't find the species interested? Please try the extended TF repertoires or identify TFs by the prediction server.

Browse by Family

AP2 (4461)	ARF (4578)	ARR-B (2354)	B3 (10609)	BBR-BPC (1256)
BES1 (1549)	C2H2 (17740)	C3H (9693)	CAMTA (1343)	CO-like (2125)
CPP (1612)	DBB (1651)	Dof (5655)	E2F/DP (1781)	EIL (1234)
ERF (21129)	FAR1 (7527)	G2-like (9874)	GATA (5335)	GRAS (9304)
GRF (1876)	GeBP (1564)	HB-PHD (477)	HB-other (2277)	HD-ZIP (8602)
HRT-like (249)	HSF (4574)	LBD (7216)	LFY (253)	LSD (957)
M-type_MADS (7541)	MIKC_MADS (6918)	MYB (22032)	MYB_related (15369)	NAC (19997)
NF-X1 (403)	NF-YA (2461)	NF-YB (3099)	NF-YC (2446)	NZZ/SPL (109)
Nin-like (2766)	RAV (690)	S1Fa-like (359)	SAP (164)	SBP (4168)

2、我们以物种为例点开 Basal Magnoliophyta (1 species)这个。可以拿到本物种所有转录因子家族的成员分布列表。

Amborella trichopoda Transcription Factors

The gene annotaion from TAGP is used to identify transcription factors (TFs) of *Amborella trichopoda* (See datasource). Acorrding to the family assignment rules, 900 TFs (900 loci) are identified and classified into 58 families. You can download the TF list (here) and protein sequences of TFs (here), or download them on the download page.

Browse by Family

AP2 (11)	ARF (13)	ARR-B (7)	B3 (19)	BBR-BPC (4)
BES1 (5)	C2H2 (67)	C3H (29)	CAMTA (3)	CO-like (5)
CPP (3)	DBB (3)	Dof (18)	E2F/DP (5)	EIL (2)
ERF (61)	FAR1 (9)	G2-like (27)	GATA (20)	GRAS (43)
GRF (6)	GeBP (5)	HB-PHD (2)	HB-other (5)	HD-ZIP (21)
HRT-like (1)	HSF (12)	LBD (23)	LFY (1)	LSD (2)
M-type_MADS (19)	MIKC_MADS (15)	MYB (60)	MYB_related (41)	NAC (46)
NF-X1 (2)	NF-YA (5)	NF-YB (9)	NF-YC (8)	NZZ/SPL (1)
Nin-like (8)	RAV (1)	S1Fa-like (1)	SAP (1)	SBP (12)
SRS (5)	STAT (1)	TALE (10)	TCP (15)	Trihelix (29)
VOZ (1)	WOX (9)	WRKY (32)	Whirly (2)	YABBY (5)
ZF-HD (9)	bHLH (84)	bZIP (37)		

3、我们点击一下 ARF 家族，他总共有 13 个成员。点击 Download Sequences 即可下载全部序列（**注意部分物种中存在可变剪接，一般前缀一样后面以.1 .2 .3 等等结尾均为可变剪接，保留一个最长的**）。

Amborella trichopoda ARF Family

- ARF Family Introduction
- Download Sequences
- Multiple Sequences Alignment
- Phylogenetic Tree

Species	TF ID	Description
Amborella trichopoda (13)	evm_27.model.AmTr_v1.0_scaffold00007.382	ARF family protein
	evm_27.model.AmTr_v1.0_scaffold00016.128	ARF family protein
	evm_27.model.AmTr_v1.0_scaffold00017.97	ARF family protein
	evm_27.model.AmTr_v1.0_scaffold00021.168	ARF family protein
	evm_27.model.AmTr_v1.0_scaffold00021.210	ARF family protein
	evm_27.model.AmTr_v1.0_scaffold00025.251	ARF family protein
	evm_27.model.AmTr_v1.0_scaffold00029.187	ARF family protein
	evm_27.model.AmTr_v1.0_scaffold00057.126	ARF family protein
	evm_27.model.AmTr_v1.0_scaffold00092.36	ARF family protein
	evm_27.model.AmTr_v1.0_scaffold00148.24	ARF family protein
	evm_27.model.AmTr_v1.0_scaffold00155.39	ARF family protein
	evm_27.model.AmTr_v1.0_scaffold00155.56	ARF family protein
	evm_27.model.AmTr_v1.0_scaffold00211.4	ARF family protein

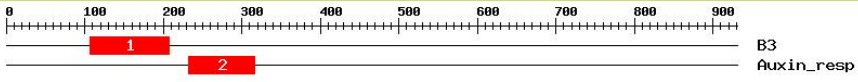
ARF Family Introduction

Auxin response factors (ARF) are transcription factors that regulate the expression of auxin response genes. ARFs bind

4、点击其中的一个 TF ID 进入如下界面，可以看到关于该基因的一些详细信息，包括基因的长度，结构域位置，序列信息，与拟南芥对应基因信息，GO 注释信息，发表文献等等。

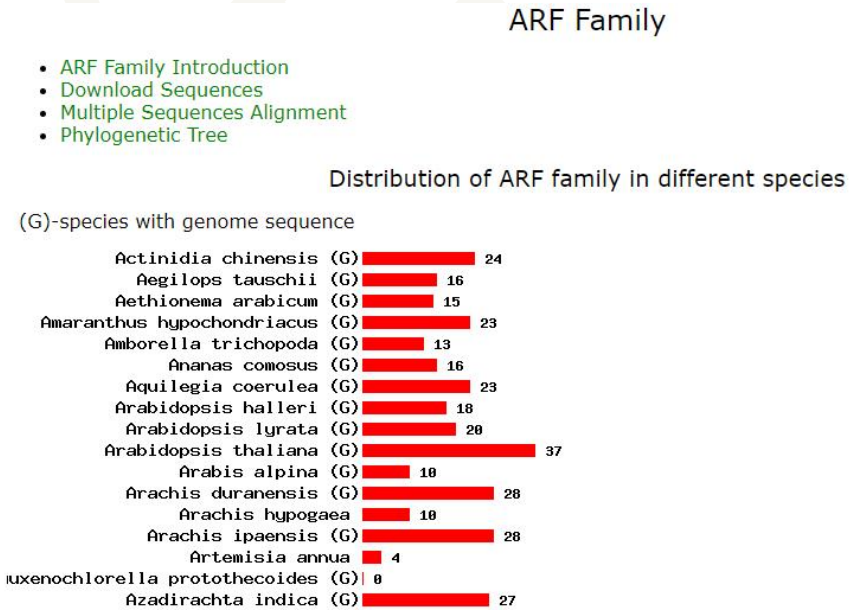
Basic Information ? help				Back to Top
TF ID	evm_27.model.AmTr_v1.0_scaffold00007.382			
Common Name	AMTR_s00007p00267580			
Organism	*Amborella trichopoda*			
Taxonomic ID	13333			
Taxonomic Lineage	cellular organisms; Eukaryota; Viridiplantae; Streptophyta; Streptophytina; Embryophyta; Tracheophyta; Euphyllophyta; Spermatophyta; Magnoliophyta; basal Magnoliophyta; Amborellales; Amborellaceae; Amborella			
Family	ARF			
Protein Properties	Length: 931aa MW: 104541 Da PI: 7.0449			
Description	ARF family protein			
Gene Model	**Gene Model ID**	**Type**	**Source**	**Coding Sequence**
	evm_27.model.AmTr_v1.0_scaffold00007.382	genome	TAGP	View CDS

Signature Domain ? help — Back to Top



No.	Domain	Score	E-value	Start	End	HMM Start	HMM End
1	B3	74.9	9e-24	107	208	1	99

5、我们按基因家族去看每个家族在物种中的分布情况。回到首页，点击 ARF,即可。在这里我们可以拿到所有物种的 ARF 家族序列（点击 Download Sequences）以及查看各物种的分布。

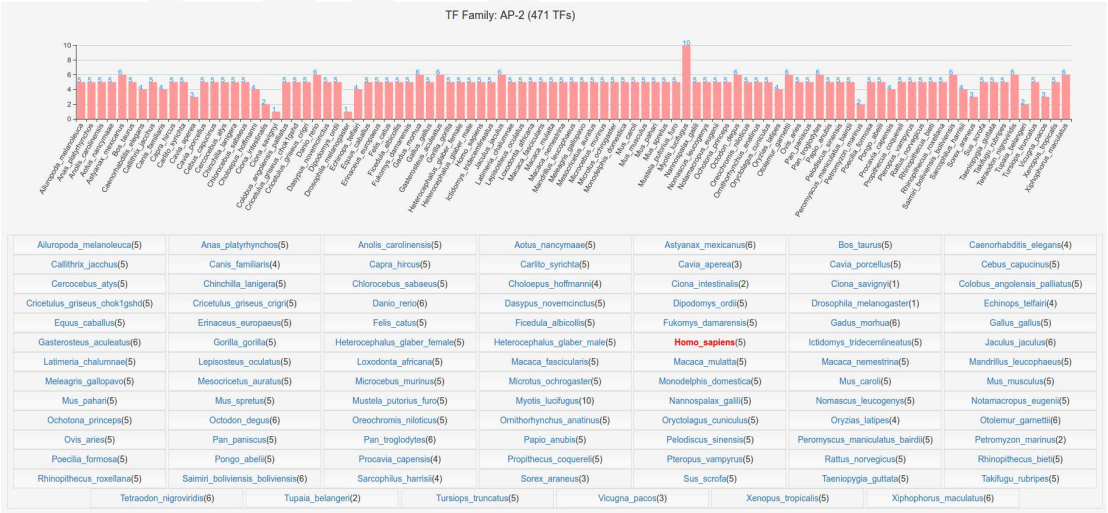


1.4 动物转录因子和辅助因子基因家族数据库下载

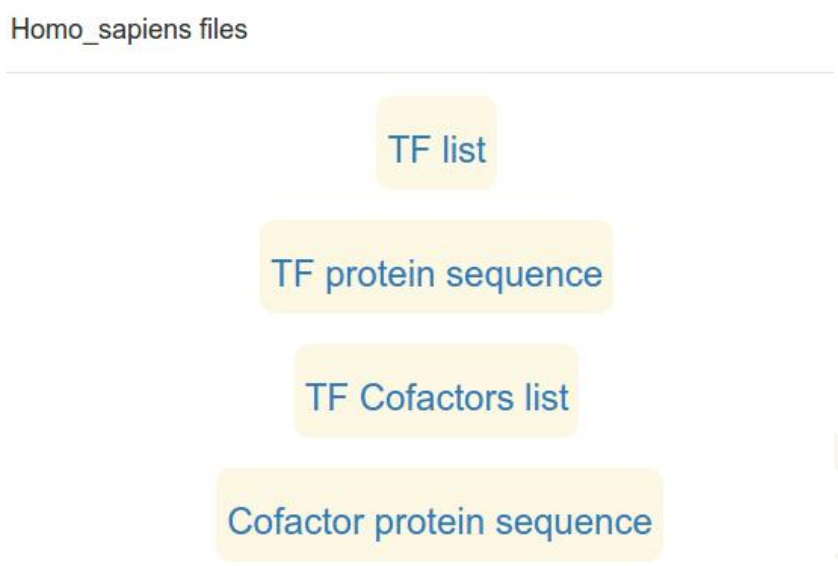
AnimalTFDB 是一个全面的数据库，其中包括 97 个动物基因组中全基因组范围内的转录因子（TFs）和转录辅因子的分类和注释。 TF 基于其 DNA 结合结构域（DBD）进一步划分为 73 个家族，辅助因子分为 83 个家族。具体用法如下：

1、进入网站：http://bioinfo.life.hust.edu.cn/AnimalTFDB/#!/

2、点 Family，出现如下界面，分为两类，一类是 TF，一类是 TF Cofactor 家族。我们以 TF AP-2 家族为例进行展示。下图展示了每个物种的分布情况。



3、数据下载。点击主页面中 Download，选择相应的物种，下载相应的 list 和 sequence 文件即可。



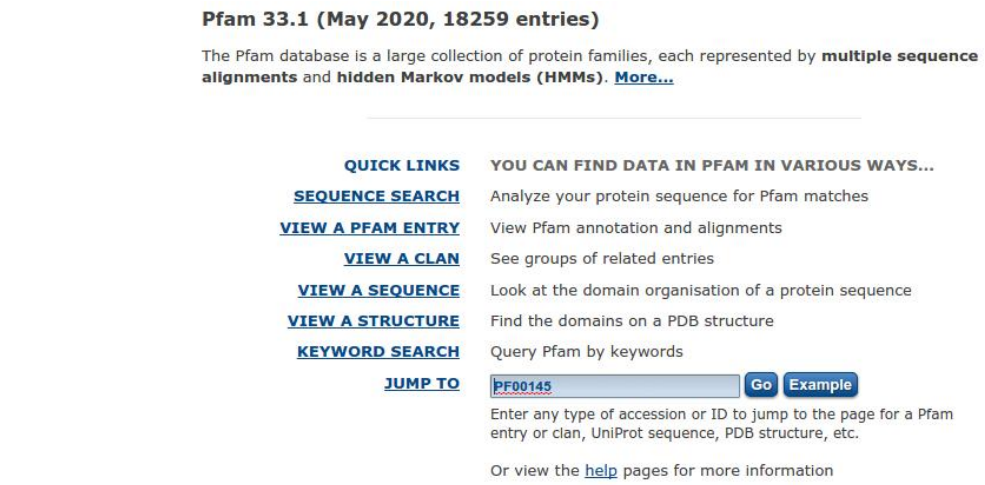
提取序列(03.scripts/getseq_fromID.pl)

1.5 Pfam 数据库下载家族成员

Pfam 数据库是大量蛋白质家族的集合，每个蛋白质家族均由多个序列比对和隐马尔可夫模型（HMM）表示。 蛋白质通常由一个或多个功能区（通常称为结构域）组成。 不同结构域的组合产生了自然界中各种各样的蛋白质。 因此，鉴定蛋白质中存在的结构域可以实现对基因家族的鉴定。网址为：http://pfam.xfam.org/。

具体我们以下载 DNA 甲基化酶基因家族必须要有的 C-5 cytosine-specific DNA methylase(PF00145)的 HMM 文件为例进行说明如下：

1、进入主页，输入 PF00145,点击 GO。

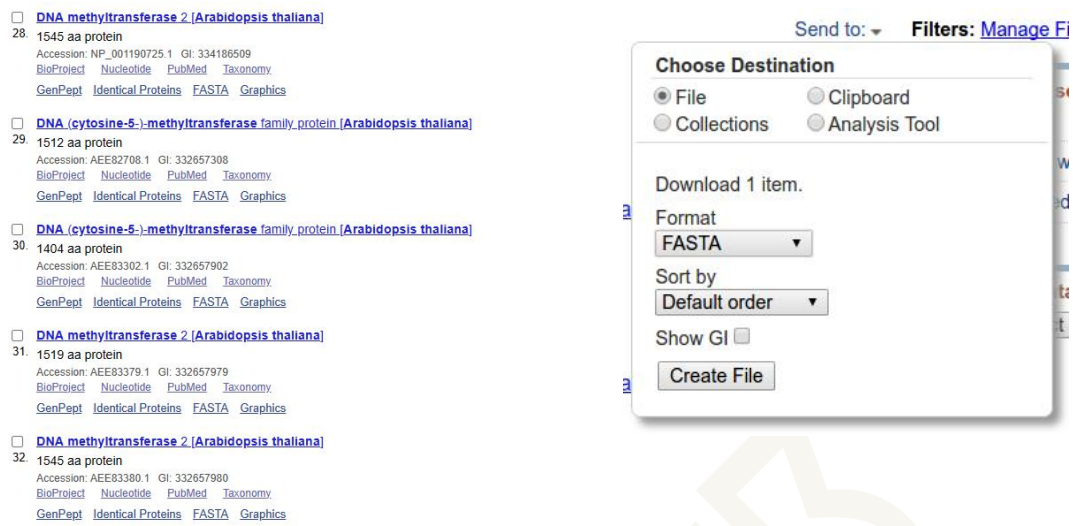


2、在跳转后的页面，点击左侧 Curation &model ，然后找到最下方一个 Download 的一个链接，即可下载到 DNA_methylase.hmm 这个文件，然后上传服务器，用于后续该家族的鉴定。

1.6 NCBI 数据库下载家族成员

可以直接检索，但要确保关键词合适，防止遗漏和增加。最好结合前人文献报道，进行定向检索，一般选择常见的模式物种如水稻、拟南芥、果蝇、家蚕、斑马鱼、人等。具体用法如下：

进入网站：https://www.ncbi.nlm.nih.gov/，选择 Protein 数据库，并输入 DNA (cytosine-5)-methyltransferase（大部分输入家族名字全称即可）,呈现如下输出结果，右侧 Results by taxon 是按照物种分类过滤选项，可以选择一类物种，也可以选择具体的物种，来进行过滤。本处我们选择拟南芥（点击[list]后选择就可以了）。出现下述页面。勾选，然后点击 Send to 选择 FASTA 输出序列即可。

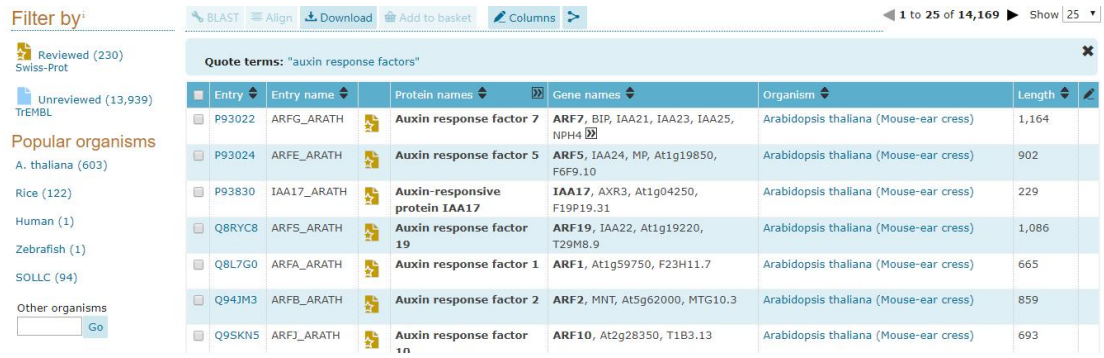


1.7 Uniprot 数据库下载家族成员

UniProt 是 Universal Protein 的英文缩写，是信息最丰富、资源最广的蛋白质数据库。它由整合 Swiss-Prot、 TrEMBL 和 PIR-PSD 三大数据库的数据而成。他的数据主要来自于基因组测序项目完成后，后续获得的蛋白质序列。它同时包含了大量来自文献的蛋白质的生物功能的信息。网址：https://www.uniprot.org/

1、打开网站，输入查找的基因家族名字,然后点击 Search。

2、结果呈现如下界面。可以在左边按照 Reviewed (人工注释的，可靠性高)和 Unreviewed (基于比对自动化注释的)物种进行过滤。并可以选择符合条件成员后点击 Download 进行序列下载。



1.8 NCBI 数据库下载基因组

主要是位于 Assembly 数据库中，包含了目前发表的大多数动植物基因组、微生物基因组的相关信息。具体我们以下载猕猴桃基因组为例进行操作。

1.9 Ensembl 数据库下载基因组

网站：http://asia.ensembl.org/index.html，此网站为该数据库的主界面，包含从灵长类（Primates）、啮齿目动物（Rodents）、劳亚兽目（Laurasiatheria），鸟类、爬行类及鱼类等脊椎动物的基因组，此外在网站最下方有 Ensembl Bacteria (细菌基因组数据库)；Ensembl Fungi（真菌基因组数据库）；**Ensembl Plants（植物基因组数据库）**；Ensembl Protists（原生生物基因组数据库）；Ensembl Metazoa（后生动物中的无脊椎动物数据库）几个子数据库。

1.10 JGI（Phytozome）下载基因组

为植物基因组数据库（https://phytozome.jgi.doe.gov/pz/portal.html），目前存储了下面这些基因组：

1.11 提取蛋白序列

做一下链接: ln -s path ~/Gene_Family

进入分析目录：cd ~/Gene_Family/02.data

利用下载的基因组和 gff 文件提取蛋白序列，cds 序列和外显子序列：

添加 gffread 环境变量（注意一定 v0.99 版本）：

执行拟南芥基因组信息提取操作：

perl ~/Gene_Family/03.scripts/prepare_gff_cds_pep.pl -gff Arabidopsis_thaliana.TAIR10.gff3 -genome Arabidopsis_thaliana.TAIR10.genome.fa -index AT -od AT_Longest/

提取前面下载的成员：

perl ~/Gene_Family/03.scripts/getseq_fromID.pl Ac_Longest/Longest/Ac.longest.protein.fa AT.id AT_methylases.fa

练习 1：提取猕猴桃基因组的蛋白序列及最长转录本对应的蛋白序列(基因组信息来源于

EnsemblPlants:https://plants.ensembl.org/Actinidia_chinensis/Info/Index)

perl ~/Gene_Family/03.scripts/prepare_gff_cds_pep.pl -gff Actinidia_chinensis.Red5_PS1_1.69.0.48.gff3 -genome Actinidia_chinensis.Red5_PS1_1.69.0.dna.toplevel.fa -index Ac -od Ac_Longest/

练习 2：提取番茄基因组的蛋白序列及最长转录本对应的蛋白序列

perl ~/Gene_Family/03.scripts/prepare_gff_cds_pep.pl -gff Solanum_lycopersicum.Ensembl_SL3.0.gff3 -genome Solanum_lycopersicum.Ensembl_SL3.0.genome.fa -index Sl -od SL_Longest

二、家族成员鉴定

基因家族候选成员鉴定的方法（两种方法可以联合使用）：

（1）利用已知的（拟南芥或者水稻等相关数据库中下载的序列）XX 条序列采用 blastp 比对（E 值<1e−5，或者 1e-3 都可）鉴定候选的成员；

（2）利用 Pfam 下载的家族特有的 HMM 模型，采用 HMMER3 进行比对（E 值 <1e−5，或者 1e-3 都可）。

基因家族成员过滤方法：

（1） BLASTP 去除重复基因（非必须）；

（2） 利用 Pfam 和 CDD 等鉴定候选基因是否存在本家族所具有的特定结构域，即可获得候选基因。

基因家族基本信息统计：

（1）通过 ExPASy 服务器（http://www.expasy.org/）中的 Compute pI / Mw 工具计算理论等电点（PI）和分子量（MW）。

（2）亚细胞定位预测采用 WoLF PSORT (https://wolfpsort.hgc.jp/）。

（3）基因结构信息统计：氨基酸长度，外显子数目，内含子数目，染色体位置。

本处我们练习采用的文献是：

Zhang, Y., He, X., Zhao, H., Xu, W., Deng, H., Wang, H., Wang, S., Su, D., Zheng, Z., Yang, B., Grierson, D., Wu, J., & Liu, M. (2020). Genome-Wide Identification of DNA Methylases and Demethylases in Kiwifruit (Actinidia chinensis). Frontiers in Plant Science, 11, 514993. https://doi.org/10.3389/fpls.2020.514993

拟南芥，番茄和水稻中已知的家族成员信息来源于 NCBI，可以采用前面已经介绍的方法下载。

2.1 BLAST 方法

利用已下载其他物种已知基因家族的序列借助于序列相似性进行本物种候选基因家族成员的鉴定。本处我们采用拟南芥中的已知的 DNA Methylases 成员鉴定猕猴桃中 DNA Methylases 基因家族。

1、BLAST 的安装。

从此网站下载即可使用：ftp://ftp.ncbi.nlm.nih.gov/blast/executables/LATEST，下载适合自己系统的 blast 版本，解压即可使用。

2、BLAST 用法。

运行下面这些 BLAST 命令之前，首先添加环境变量：

```
echo 'PATH=PATH/ncbi-blast-2.9.0+/bin/:$PATH' >>~/.bashrc
source ~/.bashrc
```

然后分两步，第一步构造数据库，第二步比对。

(1)利用猕猴桃最长转录本对应的全部蛋白序列构建 BLAST 数据库。

如果要比对的对象是蛋白序列，我们数据库类型则是蛋白序列库，具体的命令是

makeblastdb -in Ac_Longest/Longest/Ac.longest.protein.fa -dbtype prot -out refpep.db

makeblastdb 参数介绍：

-dbtype <String, `nucl', `prot'>： 数据库类型，核酸或者蛋白，选择其一。

-in <File_In>： 输入蛋白文件，即要鉴定物种的全部蛋白序列。

-out <String>： 创建的数据库名字

注意：若数据库为核酸序列 -dbtype 选择 nucl。:

(2)BLAST 比对。

blastp -query AT_methylases.fa -db refpep.db -out all.blast -evalue 1e-3 -num_threads 1 -outfmt 6

注：

-query： 已知的某家族的蛋白质序列文件

-out：比对结果的输出文件路径及文件名

-db：数据库路径及数据库名

-evalue：设置输出结果的 e-value 值

-num_threads：线程数，根据自己机器设置，越大越快。

-outfmt：输出文件格式，总共有 15 种格式，一般设置为 6。6 是 tabular 格式对应 BLAST 的 m8 格式 。此外还能自定义输出格式主要针对上述的 6, 7, and 10 三种格式，如下针对 6 格式，进行自定义输出：

-outfmt '6 qseqid qlen sseqid slen engths pident mismatch qcovs qstart qend sstart send evalue '

这样最终输出结果的的每一列信息会按照上述信息输出，具体每个单词的意思如下：

qseqid means Query Seq-id

qlen means Query sequence length

sseqid means Subject Seq-id

slen means Subject sequence length

qstart means Start of alignment in query

qend means End of alignment in query

sstart means Start of alignment in subject

send means End of alignment in subject

evalue means Expect value

length means Alignment length

pident means Percentage of identical matches

mismatch means Number of mismatches

qcovs means Query Coverage Per Subject

(3)示例结果如下：

AT4G08990	gene:CEY00_Acc21305	58.240	1523	589	16	7	1502	51	1553	0.0	1729	
AT4G08990	gene:CEY00_Acc06669	60.552	905	338	6	607	1502	1	895	0.0	1075	
AT4G08990	gene:CEY00_Acc29050	38.567	363	196	9	1161	1502	535	891	1.67e-60		225
AT4G08990	gene:CEY00_Acc29050	43.396	53	25	2	1071	1118	362	414	2.30e-04		45.8
AT4G08990	gene:CEY00_Acc22670	35.389	373	204	9	1161	1502	659	1025	1.34e-55		211
AT4G08990	gene:CEY00_Acc22670	38.571	70	38	2	1054	1118	470	539	7.82e-05		47.4
AT4G08990	gene:CEY00_Acc10156	35.580	371	205	7	1161	1501	438	804	3.23e-55		208
AT4G08990	gene:CEY00_Acc30443	32.540	378	206	11	1157	1501	427	788	1.08e-43		172

图中每一列的含义：

[1] Query id：拟南芥序列 ID

[2] Subject id:比对到数据库中的猕猴桃 ID

[3] % identity :相似度

[4] alignment length：比对长度

[5] mismatches ：错配数目

[6] gap openings：gap 的数目

[7] q. Start：我们的序列比对起始位置

[8] q. End：我们的序列比对终止位置

[9] s. Start：数据库中参考序列比对起始位置

[10] s. End；数据库中参考序列比对终止位置

[11] E value；比对的 E 值

[12] score；比对的得分

(4)筛选候选的家族基因的 id，并提取序列。

```
cut -f 2 all.blast |sort|uniq >id
perl ~/Gene_Family/03.scripts/getseq_fromID.pl Ac_Longest/Longest/Ac.longest.protein.fa id Ac.fa
```

2.2 HMMER 鉴定

HMM（Hidden Markov Model,隐马尔科夫模型），是一种序列特征谱（Profile），其搜索是基于蛋白质多序列比对结果中的保守序列区域进行检索，并考虑了不同保守度的氨基酸在相应位置的权重，可以更为敏感的检测到进化距离较远的蛋白质相关性，得到更

为灵敏的比对结果。HMMER 即是基于此方法开发的软件。HMMER 可以作为命令行工具，进行本地下载和安装，现在也可以通过 EBI 的服务器在线访问。

其中本地下载地址：http://eddylab.org/software/hmmer/hmmer-3.3.tar.gz

安装：

tar -zxvf hmmer-3.3.tar.gz

cd hmmer-3.3

./configure --prefix=$PWD

make

make install

安装完成后添加环境变量：

export PATH=PATH/hmmer-3.3/bin/:$PATH

具体用法如下：首先从 Pfam 下载某基因家族特有的 HMM，然后使用 hmmsearch 扫描你 blastp 鉴定到的候选的基因家族序列，即可获得获得该物种进一步筛选后的基因家族成员。当然也可以自己利用已知的家族成员序列，进行多序列比对后，使用 hmmbuild 构建 HMM 数据集，然后再 hmmsearch 进行比对鉴定。

1、自己构建 HMM 数据集（**可选**，本处我们直接使用前面下载的 PF00145 的 hmm 文件即可）。

命令：

hmmbuild [-options] <hmmfile_out> <msafile>

示例命令：

hmmbuild DNA_methylase.hmm　methylases.pep.align.fa

输入文件 methylases.pep.align.fa 为已知的序列经过 mafft 等工具多序列比对后的结果。

输出文件一般命名为.hmm 后缀。上述步骤完成后即可用 hmmsearch，寻找候选基因中相似的序列。

2、比对。

hmmsearch [options] <hmmfile> <seqdb>

示例命令：

hmmsearch --domtblout hmm.txt --incdomE 0.001 DNA_methylase.hmm Ac.fa

3、输出结果介绍。

一般输出参数控制为 --domtblout 。具体为输出结果中分为两类一类是针对序列的（full sequence），另一类是针对 domain 的（主要基于一条序列存在多个 domain）。这两种格式涉及到的每一列信息解释如下

(1) target name: 目标序列名字。

(2) accession: 一般为-。

(3) query name: 检索序列的名字（Query)。

(4) accession: 一般为-。

(5) hmmfrom: hmm 文件比对起始。

(6) hmm to: hmm 文件比对终止。

(7) alifrom: 目标序列比对的起始位置.

(8) ali to: 目标序列比对的终止位置.

(9) envfrom: 结构域的起始位置。

(10) env to: 结构域的终止位置。

(11) q len: 序列长度.

(12) strand: 比对的正负链.

(13) E-value: 针对整条序列而言的显著性评估指标

(14) score (full sequence): 整条序列打分

(15) Bias (full sequence): bias 是 score 的偏差

(17) c-Evalue: 针对结构域而言的显著性评估指标

(18) i-Evalue:针对鉴定的结构域而言的显著性评估指标。比 c-Evalue 更加严格。

4、提取最终结果，得到猕猴桃中 DNA Methylases 基因家族序列 Ac.fa。

```
cut -f 1 -d " " hmm.txt|grep -v "\#"|sort|uniq >final.id
perl ~/Gene_Family/03.scripts/getseq_fromID.pl Ac_Longest/Longest/Ac.longest.protein.fa final.id Ac.fa
```

2.3 结构域注释

我们通过各种方法鉴定到了某个基因家族成员，但需要进一步去鉴定结构域的位置，一般采用 Pfam 的结果，当 Pfam 未鉴定到相关的结构域，可以考虑联合采用 SMART 和 CDD。

1、 Pfam 预测结构域。

1、进入下述网址：http://pfam.xfam.org/search#tabview=tab1，在左边点击 Batch search 进行批量的同时对多条序列（鉴定的多个候选的家族成员）进行结构域注释。

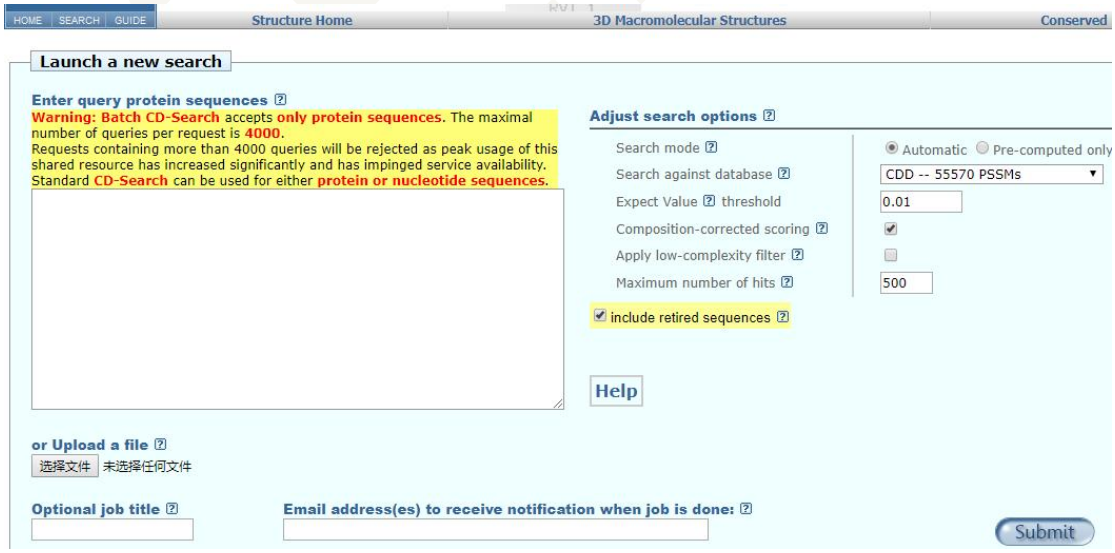
2、点击选择文件，上传相应的 Fasta 格式文件，同时写上邮箱，运行完成后就会把结果发到邮箱里。

3、在集群中使用 vi 命令创建一个叫 **pfam.txt** 的文件，把结果粘贴进去，并上传到集群： ~/Gene_Family/02.data/

2、NCBI CDD 使用。

CDD 全称为 Conserved Domain Database，可以接受不超过 4000 条蛋白序列的批量递交。网址：https://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi

打开网址，页面如下。



上传或者粘贴序列。

写上邮箱，点击 Submit 查收结果。

结果介绍。

Query	Hit type	PSSM-ID	From	To	E-Value	Bitscore	Accessior	Short name	Incomplet	Superfamily
Q#1 - >gene:CEY00_Acc02430	superfami	392266	465	579	7.54E-09	56.8598	cl30459	Cyt_C5_DNA_methylase_su	C	-
Q#1 - >gene:CEY00_Acc02430	superfami	389751	146	179	0.00118496	36.6528	cl21463	UBA_like_SF superfamily	-	-
Q#1 - >gene:CEY00_Acc02430	superfami	392266	271	452	0.00264961	39.911	cl30459	Cyt_C5_DNA_methylase_su	N	-
Q#2 - >gene:CEY00_Acc05665	specific	223348	434	551	1.12E-11	65.9158	COG0270	Dcm	C	cl30459
Q#2 - >gene:CEY00_Acc05665	superfami	389751	127	159	0.00119881	36.6528	cl21463	UBA_like_SF superfamily	-	-
Q#3 - >gene:CEY00_Acc06669	specific	240059	277	477	5.24E-81	259.701	cd04708	BAH_plantDCM_II	-	cl02608
Q#3 - >gene:CEY00_Acc06669	specific	223348	467	898	1.57E-48	175.312	COG0270	Dcm	-	cl30459
Q#3 - >gene:CEY00_Acc06669	superfami	351822	120	241	1.28E-25	102.483	cl02608	BAH superfamily	-	-

其中

Hit type：代表各种置信度级别（特定匹配，非特定匹配）和域模型范围（超家族，多结构域）。

From 和 To 是结构域的起始和终止位置；

E-Value 是结构域预测的显著性指标；

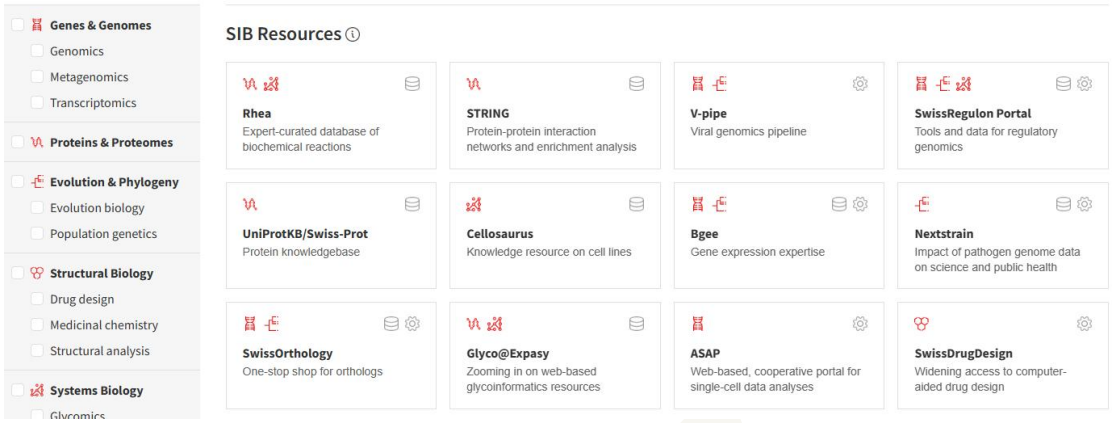
Short name：结构域名称。

Incomplete：当为-时表示结构域完整，其还可以为 N（氮端不完整）， C（碳端不完整）， NC（两末端均不完整）。

2.4 各成员基本信息统计

1 蛋白质相关特征信息。

ExPASy 是 SIB（Swiss Institute of Bioinformatics）生物信息学资源门户，可访问生命科学各个领域的科学数据库和软件工具（即资源），包括蛋白质组学，基因组学，系统发育，系统生物学，种群遗传学，转录组学等。详情见网站：https://www.expasy.org/。该网站提供了众多工具都有链接，大家可以根据网站提供的分类去找自己想要采用的工具。

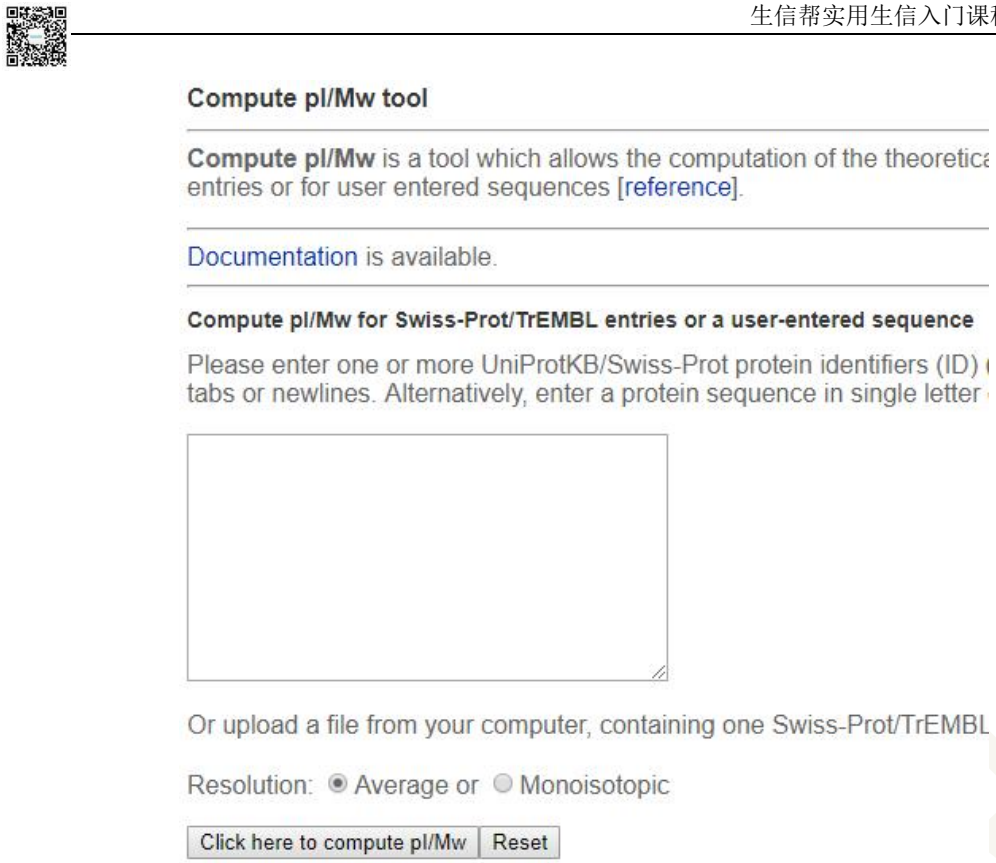


此处我们主要使用 Compute pI / Mw 工具计算蛋白质序列的理论等电点（PI）和分子量（MW），网站如下：https://web.expasy.org/compute_pi/。

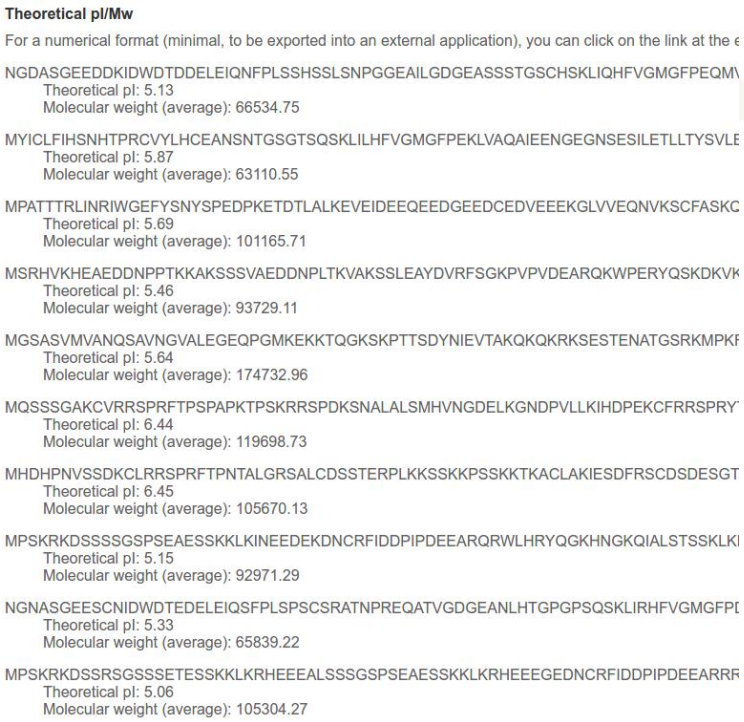
理论等电点（PI）：在某一 pH 的溶液中，氨基酸解离成阳离子和阴离子的趋势及程度相等，所带净电荷为零，呈电中性，此时溶液的 pH 称为该氨基酸的等电点，结果显示所有 HSP 的 PI 的平均值为 6.36。分子量（Molecular weight）：指每个蛋白的分子大小，单位为 Da.

我们将鉴定得到的家族序列提交这个网站即可，注意此网站需要的格式是一个基因序列必须位于一行，不保留基因的 id,可以通过下面命令得到：

```
grep -v ">" Ac.fa >Ac.txt
```



输出结果见下图：

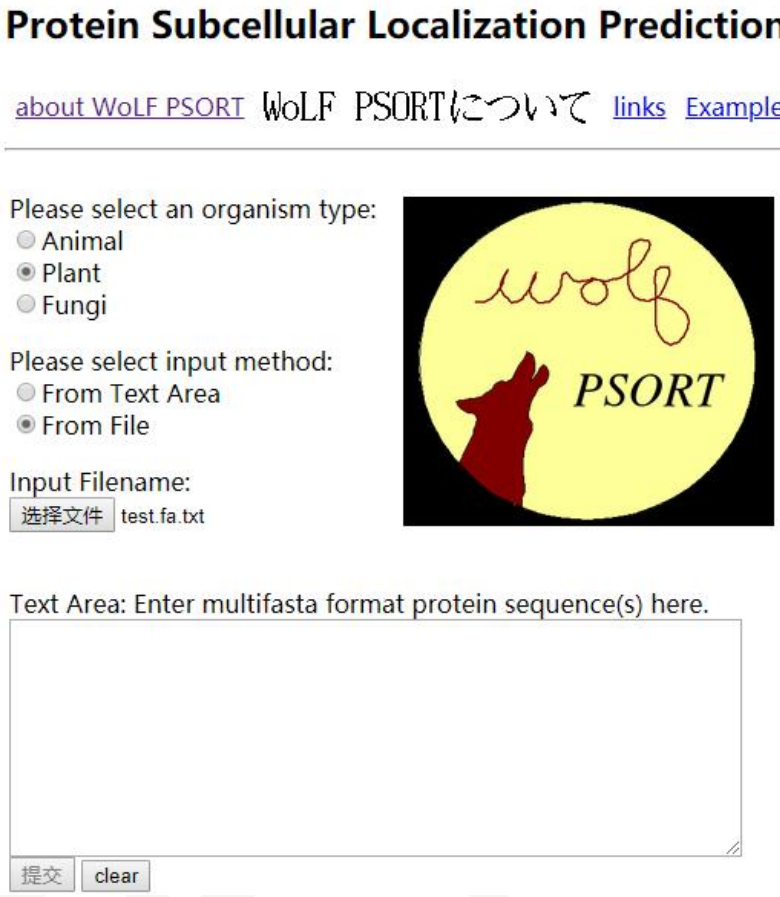


2 亚细胞定位。

预测每个蛋白在细胞中的位置，其中 Nuclear 代表定位在细胞核中，Cytoplasmic 代表定位在细胞质中，同理 Chloroplast、Mitochondrial、PlasmaMembrane、Extracellular

分别代表定位在叶绿体、线粒体、质膜、细胞外。

亚细胞定位预测采用 WoLF PSORT (https://wolfpsort.hgc.jp/)。进入网页，见下图，我们选择相应的物种分类，选择 From File，然后上传文件，提交即可。



结果解读：

#Warning sequence CEY00_Acc02430 does not start with M #Warning sequence CEY00_Acc30617 c
chlo: 1
CEY00_Acc05665 details cyto: 7, chlo: 3, nucl: 2, mito: 1, cysk: 1
CEY00_Acc06669 details nucl: 4.5, cyto_nucl: 4, cyto: 2.5, chlo: 2, plas: 2, mito: 1, vacu: 1, golg: 1
CEY00_Acc10156 details nucl: 14
CEY00_Acc21305 details nucl: 13, cyto: 1
CEY00_Acc22670 details nucl: 13, chlo: 1
CEY00_Acc29050 details nucl: 10, cyto: 3, cysk: 1
CEY00_Acc30443 details nucl: 13, cyto: 1
CEY00_Acc30617 details nucl: 12.5, cyto_nucl: 7, chlo: 1
CEY00_Acc32510 details nucl: 14

也可以采用 ProtComp (http://linux1.softberry.com/)

3 基因结构信息统计。

包括最终确定的基因号（gene ID）、染色体位置，蛋白质序列长度、CDS 长度、外显

子数目、内含子数目。同时为了方便分析，我们根据每个基因在染色体的位置将基因重新命名，称为基因名字（Gene name）；

(1)提取该基因家族成员对应的 GFF3 信息。

perl ~/Gene_Family/03.scripts/accordingIDgetGFFV1.pl -id final.id -combine Ac_Longest/Longest/Ac.longest.gff3 -o final.gff3

(2)统计信息。

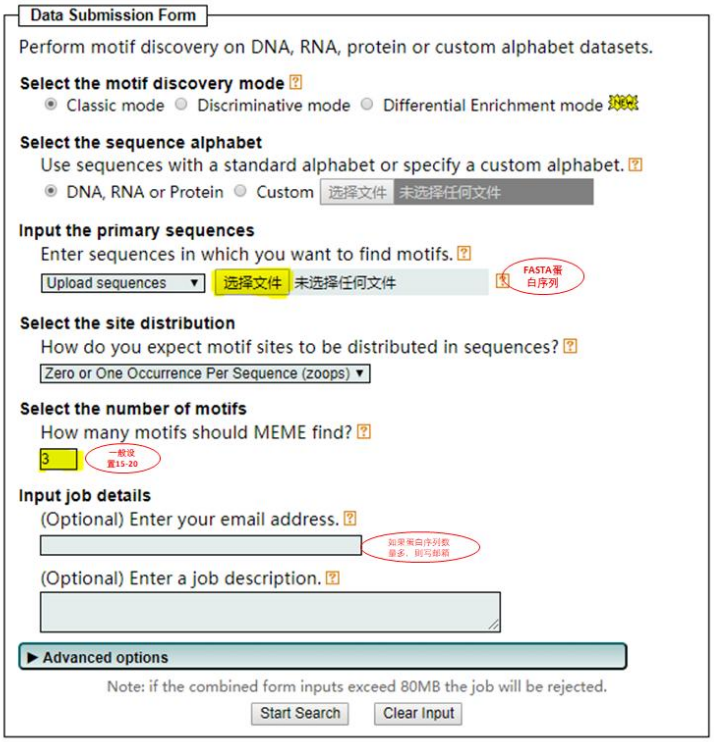
perl ~/Gene_Family/03.scripts/GFF3_statistics.pl -i final.gff3 -o final.gff3.stat

2.5 MEME Motif 鉴定

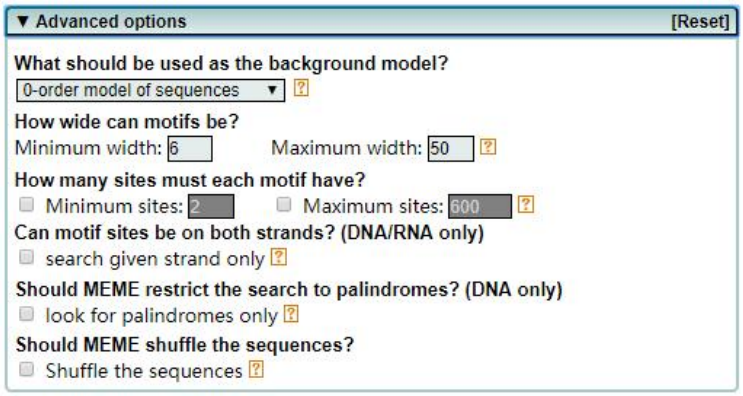
motif 是出现在一组相关基因中含有某种特征的短序列，其上面往往含有一些转录因子识别的位点。motif 分析则是通过将得到的蛋白序列提交到 MEME（http://meme-suite.org/index.html）进行预测，参数设置为最大 motif 数设为 15，motif 长度设为 6 到 200aa。基因结构和 motif 展示利用在线工具 evolview 进行可视化和美化。

目前 motif 预测主要通过 MEME 在线网站进行分析 (http://meme-suite.org/tools/meme)。

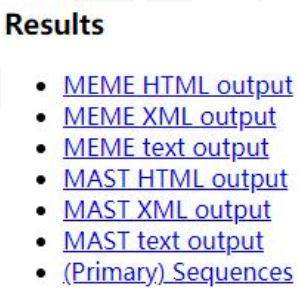
1、进入网站：http://meme-suite.org/tools/meme



其中 motif 个数一般设置 15 或者 20（具体看别人发表文献）即可，高级参数这里只需要选择 Maximum width 设成 50（或参考文献）即可。设置完成后点击 Start Search 提交运行即可。



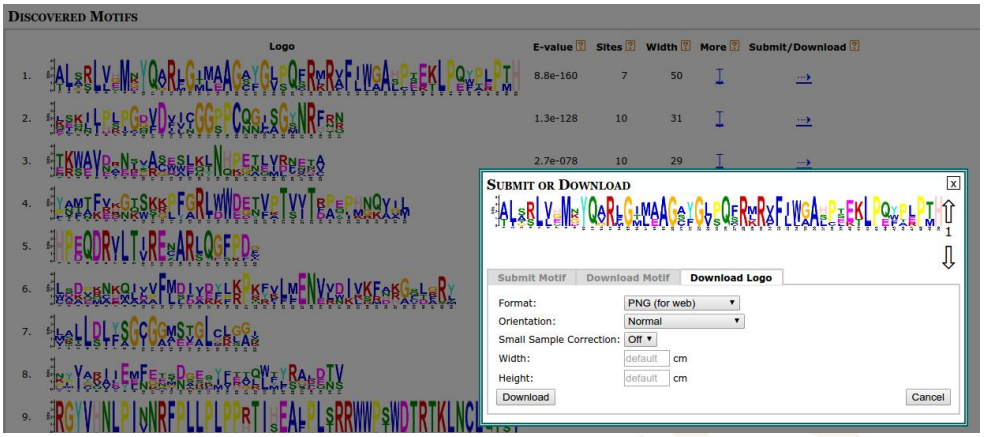
2、运行结果完成后出现如下界面，主要看 HTML 格式的，同时讲 MEME text output 鼠标右键另存为 meme.txt，并上传集群 ~/Gene_Family/02.data/。



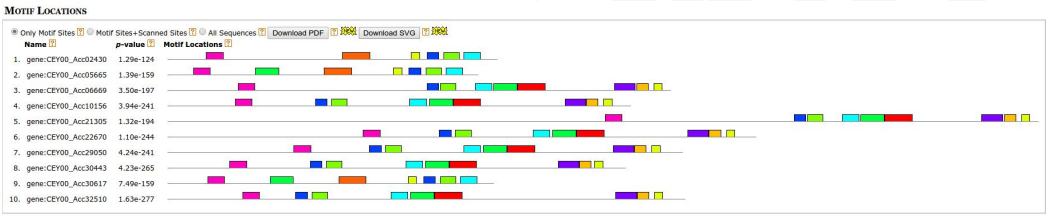
3、结果下载。

我们打开 MEME HTML output 输出结果，在这里面我们可以下载如下图片，可以进行组合获得最终用于发表的图片。

每个 motif 的 Logo 图：



Motif 在基因组上的分布图：



4、位置提取。主要采用上述 txt 格式利用下面脚本进行解析后获取。然后 **meme.evolview** 采用 Evolview 进行可视化。

perl ~/Gene_Family/03.scripts/make_meme_motif.pl **meme.txt** final.gff3.stat **meme.evolview**

2.6 基因在染色体上定位展示

采用在线网站进行绘制 http://mg2c.iask.in/mg2c_v2.1/

染色体长度获取，用于输入 2：

perl ~/Gene_Family/03.scripts/get_Seq_length.pl Actinidia_chinensis.Red5_PS1_1.69.0.dna.toplevel.fa chr.len

输入 1：

CEY00_Acc32510 15169385 15181698 LG28

CEY00_Acc02430 13147187 13161342 LG2

CEY00_Acc29050 16658886 16677120 LG25

CEY00_Acc30443 19829011 19840660 LG26

CEY00_Acc21305 12151974 12162752 LG19

CEY00_Acc22670 4250119 4266395 LG20

CEY00_Acc05665 11554590 11564191 LG5

CEY00_Acc30617 1537771 1548409 LG27

CEY00_Acc10156 4688362 4699999 LG9

CEY00_Acc06669 11943186 11957240 LG6

输入 2:

LG2 14622108

LG5 18594566

LG6 17379730

LG9 16576380

LG19 15405244

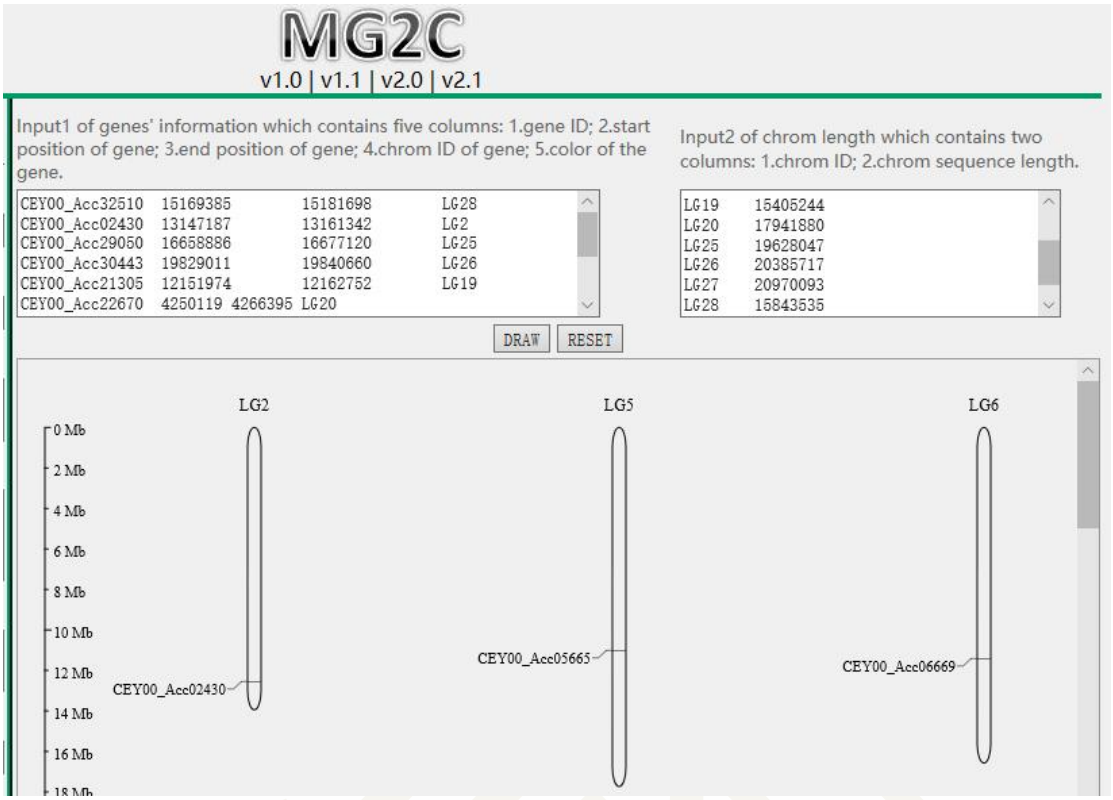
LG20 17941880

LG25 19628047

LG26 20385717

LG27 20970093

LG28 15843535



三、进化树构建

通过进化树可以实现对基因家族的进一步分类，如基因家族的进化分析，可以将其划分成 3 个亚组，每个亚组可能发挥不同的功能；其次进化树中位于同一分枝的成员之间往往具有更近的亲缘关系，因而推测其功能也更加相近；最后我们通过进化树分析能比较直观的发现基因家族各成员的扩增情况。

主要分析流程为：研究基因家族各成员蛋白序列准备 - >多序列比对 - >ML/NJ 法构建进化树 - >进化树美化 - >结果保存。这些操作均可以通过 MEGA 软件单独实现。

在进化树构建过程中，常用的多序列比对软件有 Muscle ， MAFFT ， Clustal W，T-Coffee 等。常用构建系统进化树的方法是 NJ 和 ML 法。如果序列相似性较高，选择什么样的方法都能得到不错的结果。建树采用的软件包括 MEGA（NJ 法首选），RaxML，PhyML 和 IQ-TREE 等。另外建完树后可以对进化树进行美化，常用的软件比如：MEGA、iTOL、Figtree、Evolview、ggtree 等。

3.1 进化树构建的基本原理

3.1.1 常用的进化树构建方法

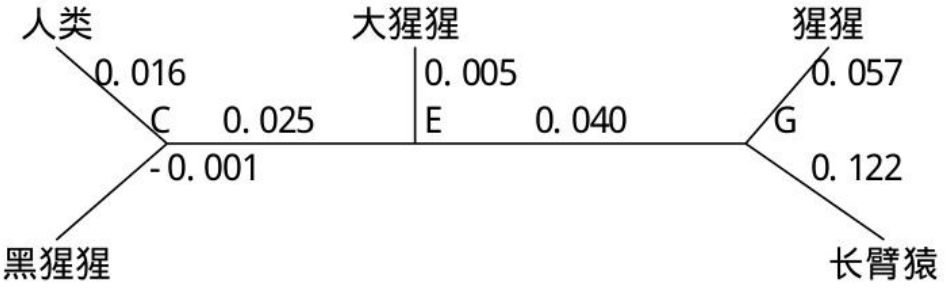
本处主要介绍 NJ 邻接法（Neighbour Joining）和 ML 最大似然法(Maximal likelihood)，另外还有贝叶斯法，运用较少，本文不做介绍。

1　NJ 邻接法（Neighbour Joining）

1、原理。

从一个距离矩阵开始，采用一定的标准，递归的合并矩阵中距离最短的两个节点，并重构新的距离矩阵，如此反复，直到剩下最后一个分类单元。常用分析软件为 MEGA。

举个例子如下图：人与黑猩猩是相邻的,人与大猩猩则不是;如果人与黑猩猩组成一个新类,则该新类与大猩猩又成为相邻。总之,通过循序地将相邻点合并成新的点,就可以建立一个相应的拓扑树。



2、优点。

假设少，算法简单易懂，运算速度快，树的构建相对准确，可以分析较多的序列；

缺点。

序列上的所有位点等同对待，只能应用于进化距离小，序列相似度较高的序列。

.2 ML 最大似然法(Maximal likelihood)

1、原理。

与 NJ 法不同的是，该方法不计算序列间的距离，而是将序列中有差异的位点作为单独的特征来构建进化树。具体为以一个特定的替代模型对每个位点所有可能出现的残基替换的概率进行累加，产生特定位点的似然值，然后对对所有可能的系统发育树都计算似然函数值，最后选择似然函数值最大的树作为最终结果。常用软件为 RAxML，phyML 和 iqtree。

2、优点。

在进化模型确定的情况下，ML 法是与进化事实最吻合的建树算法。

3、缺点。

所有可能的系统发育树都计算似然函数，计算强度非常大，极为耗时；依赖于合适的替代模型。

3.1.2 方法选择

对于单个基因家族的分析，由于序列相似度较高，可以采用 NJ 法进行构树，常用的软件为 MEGA 软件。ML 方法可以用 IQ-TREE。

3.1.3 进化树可靠性评估

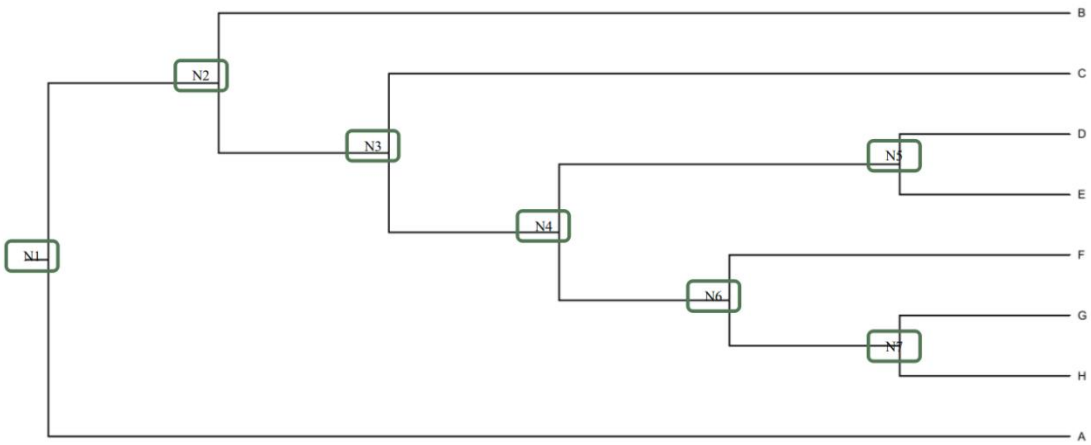
一般采用 Bootstrap 值，基因家族分析一般采用 1000 次。

Bootstrap 检测的具体方法为，经过多序列比对后，序列长度为 L ，我们从这 L 个位点进行 L 次有返回抽样，构成一组新的多序列比对结果，然后对其进行似然估计，构建进化树，如此重复N次，如果有 M 次与我们之前构建的进化树的拓扑结构一致，则我们就可以得到反映进化树可信度的 bootstrap 值，即 M/N*100。一般我们认为 bootstrap >70 即为比较可靠。

另外在 MEGA 中同时构建了 Boostrap consensus tree，即通过 “多数规则”(majority rule) 建立一致性树 (consensus tree)。

3.1.4 Newick 格式介绍

Newick 是最常见的进化树文件格式，其为一个文本文件。在了解这种格式之前，先看一下图形化的树的结构。



在上述示例中，一共有 A-H 共 8 个叶子节点，N2-N7 共 6 个内部节点以及 N1 这一个根节点。

所有节点之间的层级关系是 N2 和 A 这 2 个节点直接和根节点N1 相连，是树形结构中的第一层；B 和 N3 相连，是树形结构中的第二层；C 和 N4 相连，是树形结构中的第三层；N5 和 N6 相连，是树形结构中的第四层；D、E、F 和 N7 相连，是树形结构中的第五层；G 和 H 相连，是树形结构中的第六层。位于同一层级的节点，互称为同胞节点(sliding nodes)；层级关系中位于上一层的节点，称之为父节点(parent node)，比如 N5 就是 D 和 E 的父节点；类似的，G 和 H 称为 N7 的子节点(child node)。

当我们表示一个树状结构时，本质上是表示节点和分支的信息。对于 Newick 格式，采用圆括号将同胞节点括起来，多个节点之间用逗号相连，比如 G 和 H 表示为(G,H)，对于父节点 N7，直接写在子节点圆括号的外面为(G,H)N7，通过圆括号的嵌套区分不同层级，然后就可以表示出一棵完整的树，上述的 tree 我们按照一个层级一个层级去写如下：

(A,N2)N1->(A,(B,N3)N2)N1->(A,(B,(C,N4)N3)N2)N1
->(A,(B,(C,(N5,N6)N4)N3)N2)N1-> (A,(B,(C,((D,E)N5,(F,N7)N6)N4)N3)N2)N1->
(A,(B,(C,((D,E)N5,(F,(G,H)N7)N6)N4)N3)N2)N1

最终得到这棵树的 Newick 格式为：

(A,(B,(C,((D,E)N5,(F,(G,H)N7)N6)N4)N3)N2)N1；

(A,(B,(C,((D,E),(F,(G,H))))));

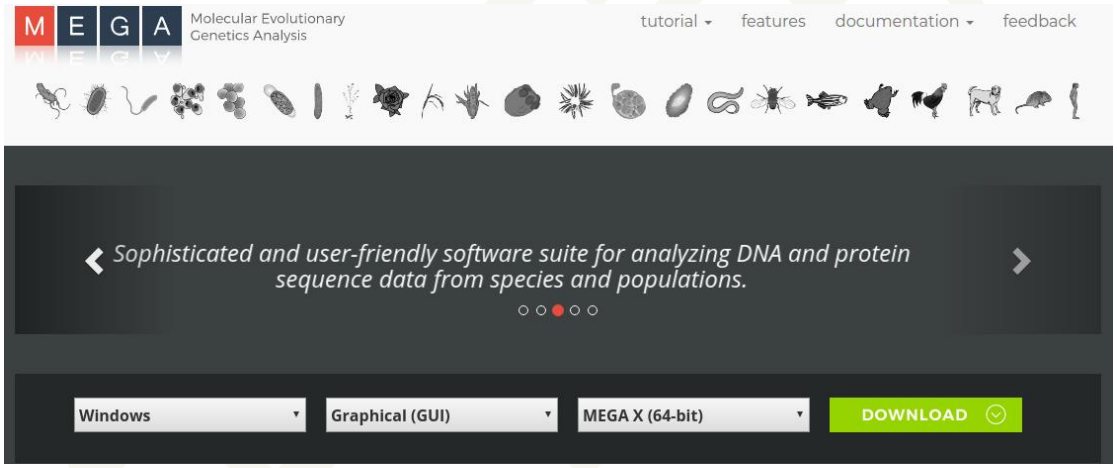
但上述的表示方式缺少了分支长度信息，如分化时间，遗传距离等。对于分支的信息，我们用节点的属性来表示，与节点的名称之间用冒号:分割，比如 G:0.1。这种表示方式涵盖了 tree 文件中所有的信息，但在实际使用中，我们通常更关注的是叶子节点，内部结点只是用来呈现树的层级结构，其名称并不是很重要，所以没必要在 Newick 格式中显示。同时根节点也可以忽略不显示，所以上述 tree 比较完整的表示方式为
(A:0.1,(B:0.1,(C:0.1,((D:0.1,E:0.1),(F:0.1,(G:0.1,H:0.1):0.1):0.2):0.1):0.3):0.1)；

3.2 MEGA 构建进化树

3.2.1 MEGA 安装

MEGA (Molecular Evolutionary Genetics Analysis)由在亚利桑那州立大学生命科学院的 Sudhir Kumar 教授主持开发和维护，是一款在 Windows, Linux 和 Mac OS X 平台下均能够运行的分子进化和遗传学分析的软件,目前最新版本为 MEGA X。因为采用了图形化界面和详尽的帮助系统,使得 MEGA 成为分子进化和系统发育分析领域内的拥有最为友好界面和最便利操作的软件，可用于序列比对、进化树推断与美化、估计分子进化速度、验证进化假说等。

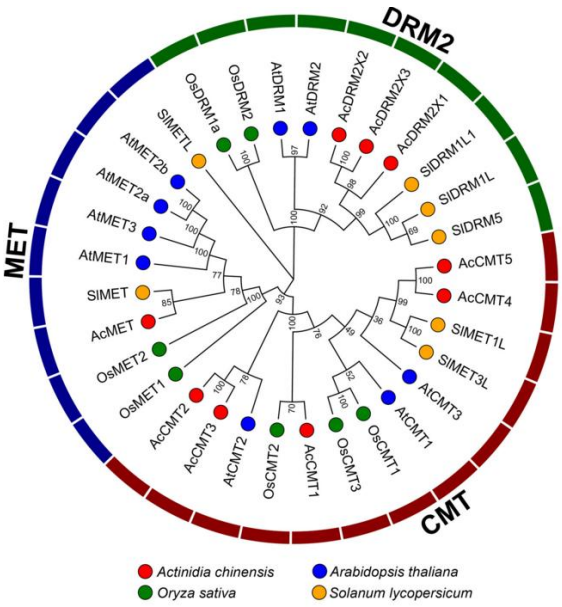
MEGA 下载地址为 https://www.megasoftware.net。进入此网站后，在下图界面中，选择相应系统和版本的 MEGA 即可，本处我们以 Window 的 64bit 系统的图形化界面(GUI)的软件为例。



点击 DOWNLOAD，下载完成后，双击 EGA_X_10.0.5_win64_setup.exe 进行安装即可。

3.2.2 MEGA 使用

我们主要以中 Figure 1（见下图）C5-MTases 基因家族为例进行进化树构建。

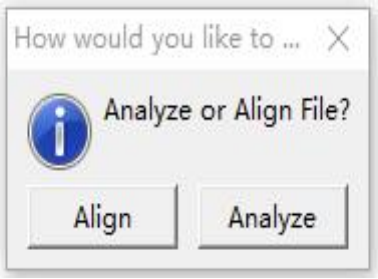


3.2.2.1 蛋白序列数据准备

我们用猕猴桃中鉴定到的 C5-MTases 基因家族，外加拟南芥、番茄和水稻的该家族成员，进行进化树构建，并将序列名字命名为 C5_MTase.fa，要求所有序列名字只保留 ID，不要保留序列的描述部分！

3.2.2.2 载入文件

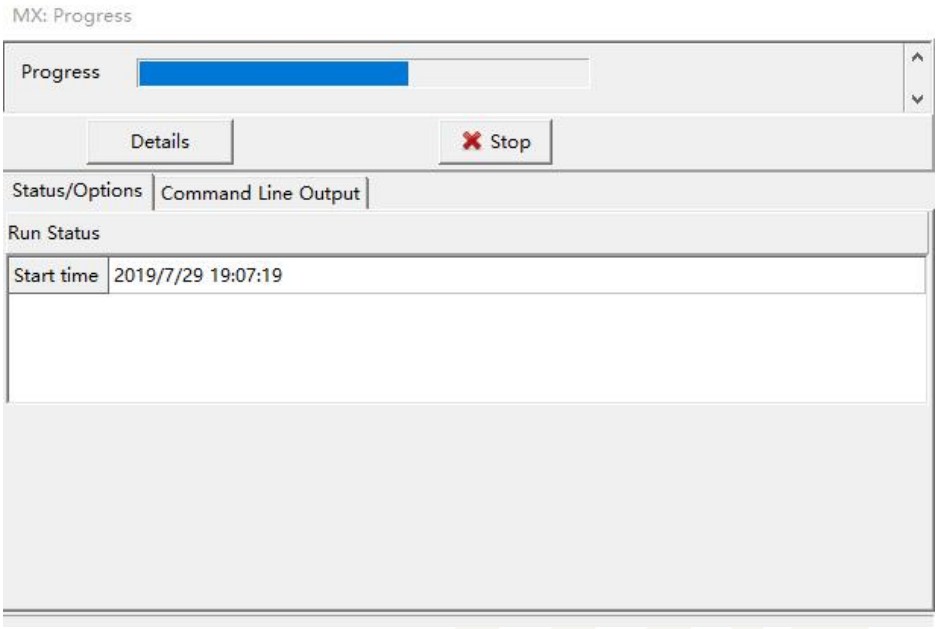
双击打开 MEGA 软件后，执行 File->Open A File/Session..->找到并载入 C5_MTase.fa 文件后弹出下面虽对话框：



我们选择 Align 进行下一步的多序列比对即可。

3.2.2.3 多序列比对

点击 Edit->select All 选中所有序列，然后点击 Alignment > Align by ClustalW（产生参数页面，一般默认即可）->OK，开始进行比对，并呈现以下状态：



比对完成后，我们将比对结果保存为 meg 格式（FASTA 格式如何输出？），具体为点击 Data > Export Alignment > MEGA Format，选择保存位置和文件名字，然后弹出对话框，输入一个标题即可，如 gene family。最后关闭这个窗口，即完成了比对结果的输出。

3.2.2.4 构建进化树

点击 MEGA 操作主界面 PHYLOGENY 中的 Constrcuct/Test Neighbor-Joining Tree .. 进行 NJ 树的构建，(也可以选择 Constrcuct/Test Maximum likelihood Tree ..进行 ML 树的构建，两种方法差别不大）。选择后会弹出载入数据的窗口，选择刚才保存的 mega 格式文件，载入后显示参数对话框：

左侧为可选的参数，右侧为相应参数的赋值，下面是比较重要的几个参数

1、Phylogeny Test

一般要选择 Bootstrap test，以及对应的次数。一般设置 1000 次。

2、Substitution Model。

可以选择替换模型，一般默认即可。

3、Data Subset to Use

主要用于处理多序列比对产生的 Gap 和缺失数据的 ，可以选择移除所有的含 gap 的比对结果（Complete deletion option)， 也可以在计算两序列之间的遗传距离时，进行两序列间的 Gap 删除(Pairwise deletion option)，或者根据 gap 的覆盖度进行删除(Partial deletion option)，不满足我们设定的百分比(Site Coverage Cutoff)，就将包含 gap 的比对结果删除。一般选择 Partial deletion option。

上述参数设置完毕后，点击 OK 按钮，即可开始建树。

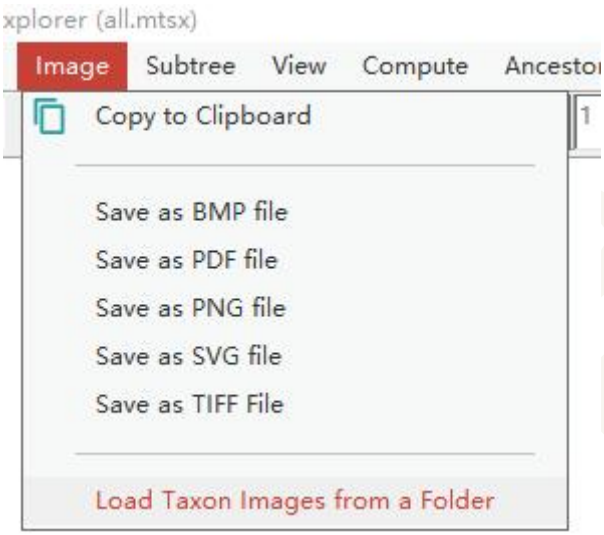
并最终得到两张图，即 Original tree / Bootstrap consensus tree。一般采用 Original tree 进行进一步的美化。

3.2.4 结果保存

主要将美化后进化树图片文件保存和进化树的文本文件保存。

3.2.4.1 进化树图片文件保存。

在 Tree Explorer 窗口中位于 Image 菜单下面（见下图）。我们一般另存成 SVG (PDF) 和 TIFF 两种格式，其中前者是矢量图。可以通过点击 Image->Save as SVG（PDF） file，保存至 filename.pdf。选择 Image->Save as TIFF file，保存至 filename.tiff。



3.2.4.2 保存进化树的文本文件。

选择 File->Export Current Tree (Newick)则以 Newick 格式保存树文件，不含有原先美化的各种效果，为文本文件，后续可以通过 Evolview 美化。

4.2.4.3 保存进化树的编辑后文件。

在 Tree Explorer 窗口中。点击 File->Save Current Session，会以 mtsx 格式保存树文件，此文件可以用 MEGA 重新打开，在原基础上继续编辑美化。

3.3 IQ-TREE

此处我们主要练习只用**猕猴桃的基因家族成员**建树的步骤，所有序列建树也是一样的操作。

3.3.1 多序列比对

将两条以上序列进行对齐，则需要依靠多序列比对的软件，如常见的 MUSCLE,mafft 等。多序列比对结果经常用于寻找保守结构域或 motif 以及进化树构建等。MUSCLE 嵌合于 MEGA 中，有 window 版本可以使用，我们此处重点介绍 Linux 下的 mafft 用法。

1、mafft 软件安装。

下载;

https://mafft.cbrc.jp/alignment/software/mafft-7.429-without-extensions-src.tgz

解压：tar -zxvf mafft-7.429-without-extensions-src.tgz

cd mafft-7.429-without-extensions/core

更改安装目录：

```
perl -p -i -e 's#PREFIX =.*#PREFIX =PATH/mafft-7.429-without-extensions/#' Makefile
perl  -p -i -e 's#BINDIR=.*#BINDIR=PATH/mafft-7.429-without-extensions/bin/#' Makefile
```

编译：

```
make clean
make
make install
```

添加环境变量：

```
echo 'PATH=PATH/mafft-7.429-without-extensions/bin:$PATH:' >> ~/.bashrc
```

环境变量生效：source ~/.bashrc

2、用法。

```
mafft --maxiterate 1000 --localpair  input [> output]
```

示例命令：

```
mafft  --maxiterate 1000 --localpair  --thread 1 Ac.fa > Ac.align.fa
```

其他参数：

--thread：可以写线程数，越大速度越快。

3.3.2 IQ-TREE

IQ-TREE 在很大程度上解决了最大似然法建树软件在速度上的缺陷。主要表现在 IQ-TREE 在 Bootstrap 检测上比 MEGA 等快 10 - 40 倍；在模型选择上是 jModelTest and ProtTest 的 10 – 100 倍；同时可以处理大批量的数据，如上千条序列或者是 Mb 级别的多序列比对位点。

1、命令：

iqtree -s Ac.align.fa -bb 1000 -m TEST -nt 1 -pre IQ_TREE

2、参数：

-s：输入的多序列比对文件，可以接受以下多种格式 PHYLIP/FASTA/NEXUS/CLUSTAL/MSF。

-st：数据类型，如 DNA, AA, NT2AA, CODON 等。根据自己序列进行选择。

-m：最佳模型选择的范围。

-bb：进行 ultrafast bootstrap 值检验 1000 次。

-nt：根据自己的机器设置相应的 CPU，资源充足情况下，要设置大一些，能够加快运行速度。

3、主要的结果：

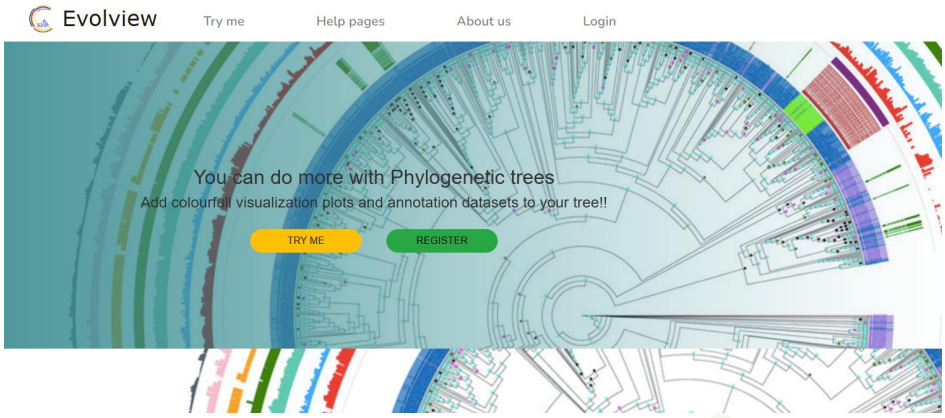
IQ-TREE 构建的进化树: IQ_TREE.treefile，为 newick 格式，跟 MEGA 输出是一致的。

IQ-TREE 构建的进化树的运行信息：IQ_TREE.iqtree。其中在 ModelFinder 这一部分有采用的最佳模型。

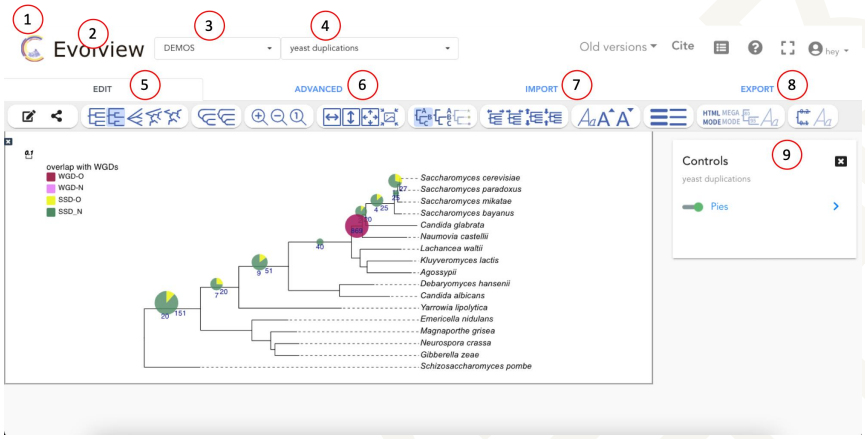
3.4 Evolview 进化树美化

Evolview 是一个功能强大的在线工具，可以帮助用户可视化和编辑系统树，并且支持各种注释和注释类型。用户可以轻松地导入、导出和共享他们的树和数据集，并且还可以使用详细的帮助和支持资源来解决任何问题。开始使用 Evolview，只需登录或注册一个帐户，上传您的树文件并开始进行可视化和编辑。Evolview 网址：

https://www.evolgenius.info/evolview/#login，免费注册后可以登录使用。

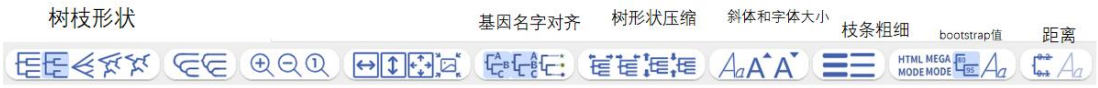


用户界面：

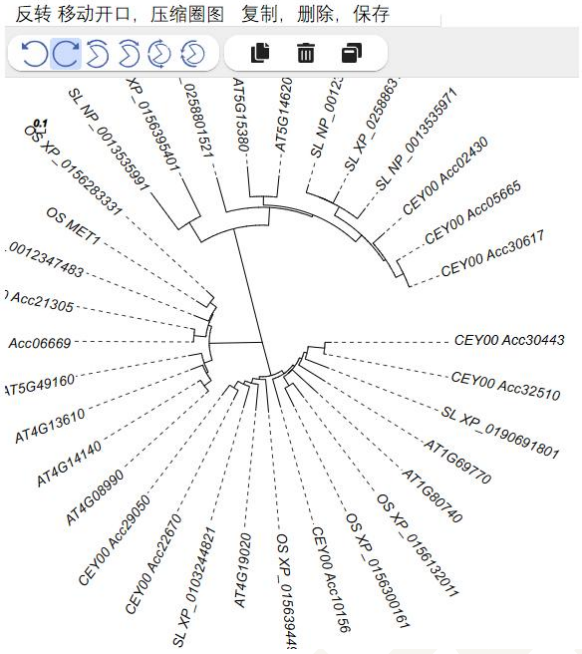


Edit 界面意思

编辑工具栏：提供快速树编辑选项的工具栏。编辑工具栏将按钮根据功能分组；高级工具栏：提供一些高级树编辑操作，如交换树旋转角度、显示/隐藏控制面板、删除树；导入工具栏：提供所有文件导入相关操作；导出工具栏：提供图像和数据导出选项；控制面板：数据集控制面板。



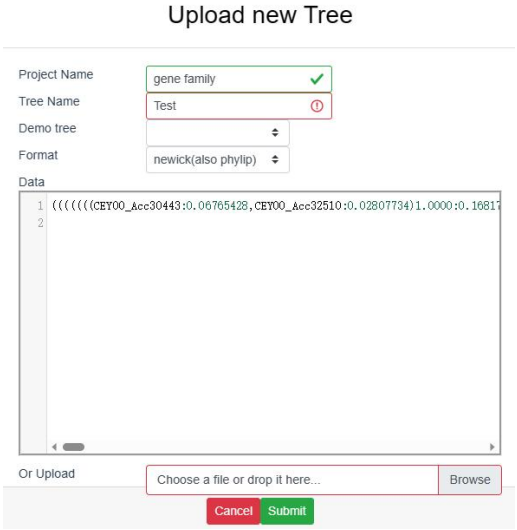
Advance 界面意思



Import 界面意思

树上传和注释。当前版本的 Evolview 支持完全二十种注释类型，包括新增加的一种。其中最受欢迎的包括叶标签背景颜色、群组标签、颜色条和颜色形状、叶子颜色和热图等。一般来说，这些注释类型可以根据其预期功能分为以下几个子类别：叶子样式和注释、分支样式和注释以及其他注释的显示。

数据读入。Evolview 支持两种类型的树上传操作：拖放树文件和使用上传按钮上传树。这两种上传选项都需要提供一些详细信息，如项目名称和树名称，以便进行方便管理支持的输入格式 newick/phylip/nhx/nexus/phyloXML。如下填写项目信息后，点击 submit。



分支样式和注释。该类别包括八种注释类型，如分支颜色、饼图、Bootstrap 值样式、时间线绘图和折叠内部节点等。分支颜色注释类型有助于修改分支颜色，饼图数据集类型有

助于在分支和内部节点上插入饼图绘图。处理极大树的可视化呈现往往伴随着一些困难，折叠内部节点可用于对节点进行分组，通过节点分组提供更好的视角。

叶子样式和注释。总共有四种注释类型属于此类别，包括叶子颜色、叶子背景颜色、叶子标签修饰、群组标签和叶子图像。这些注释类型侧重于帮助用户进行基于叶子标签的修改，例如更改叶子标签颜色、更改背景或在叶子标签和分支之间添加其他装饰。其中许多注释类型便于将多个数据集上传到同一棵树上，为用户提供不同可视化的支持。

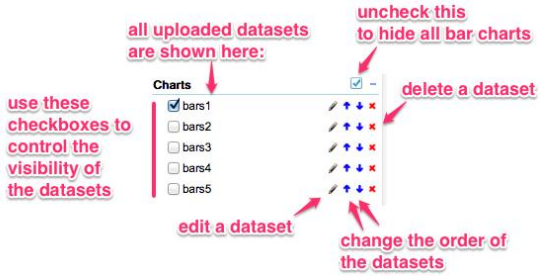
其他注释。Evolview 一直在支持将各种树与特定类型的图或图表结合和可视化。目前的版本 3 完全支持七个数据集，包括条形图、热图、蛋白质结构域图、柱状图、点图、颜色条和形状。这些注释类型为用户提供了额外的可视化选项，可以与系统树集成，为树相关信息提供更大的可视化平台。

自动生成注释。从 Evolview 4 版本开始，我们为注释数据集提供了模板生成选项，这将帮助用户尝试任何可视化选项。注释模板是为用户选择的树生成的，供用户选择注释类型。用户可以编辑、修改注释定义以满足其自己的可视化样式选择。

Export 界面

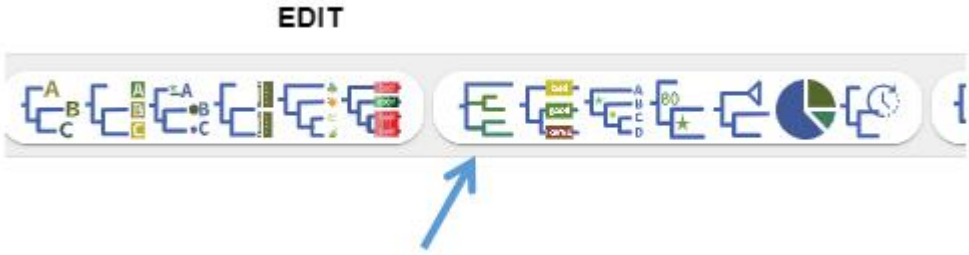
导出选项包括图像（PNG、SVG）、数据集以及可嵌入网页的交互式树。通过导入和导出功能，用户可以方便地共享和保存他们的树和数据集。

绘制完成后，各种类型的数据都可以选择保留或者改变顺序以进一步美化。下面将讲解每种类型的配置，需要注意关键字与数据之间用制表符分隔！



1、枝条标注颜色。

可以按照前缀（prefix）标注所有枝条的颜色，也可以分组标记一组枝条的颜色（ad）。




```
INT2   red     ad
INT14  blue    ad
INT22  green   ad
```

2、叶子标注背景颜色。

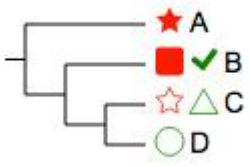
注意给基因名字或者物种名字添加背景颜色。Datasets 示例如下：

```
INT2   red     ad
INT14  blue    ad
INT22  green   ad
```

2）叶子标记装饰符号。

叶子装饰是要在叶子节点标签和分支之间显示的颜色对象或者形状，一个叶子允许多个装饰类型，支持三角形（triangle）、圆形（circle）、矩形（rect）、星形（star）对号（check）

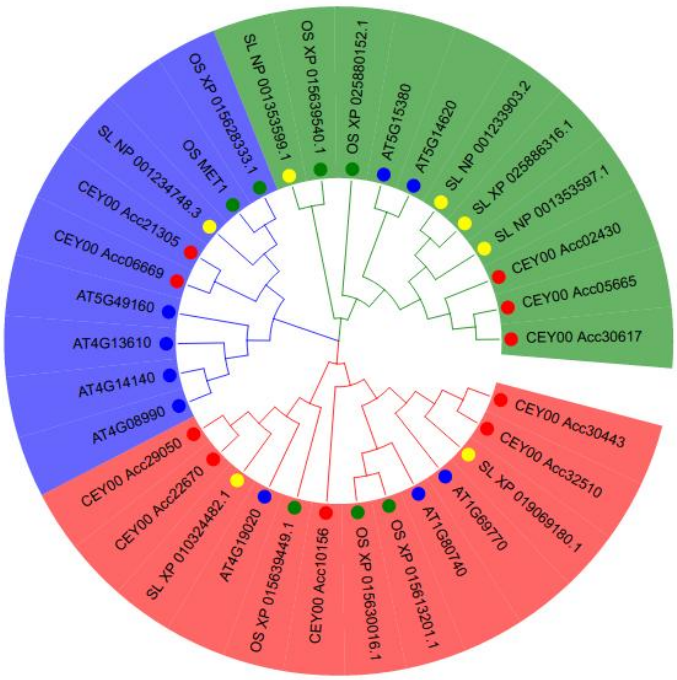
```
!defaultstrokewidth  0.7
A   star,red
B   rect,red    check,green##形状和颜色是一组数据
C   star,white:red     triangle,white:green##white 不填充颜色
D   circle,white:green
```



本处我们采用下述命令后在 excel 中制备：

grep ">" C5_MTase.fa |sed 's:>::g'

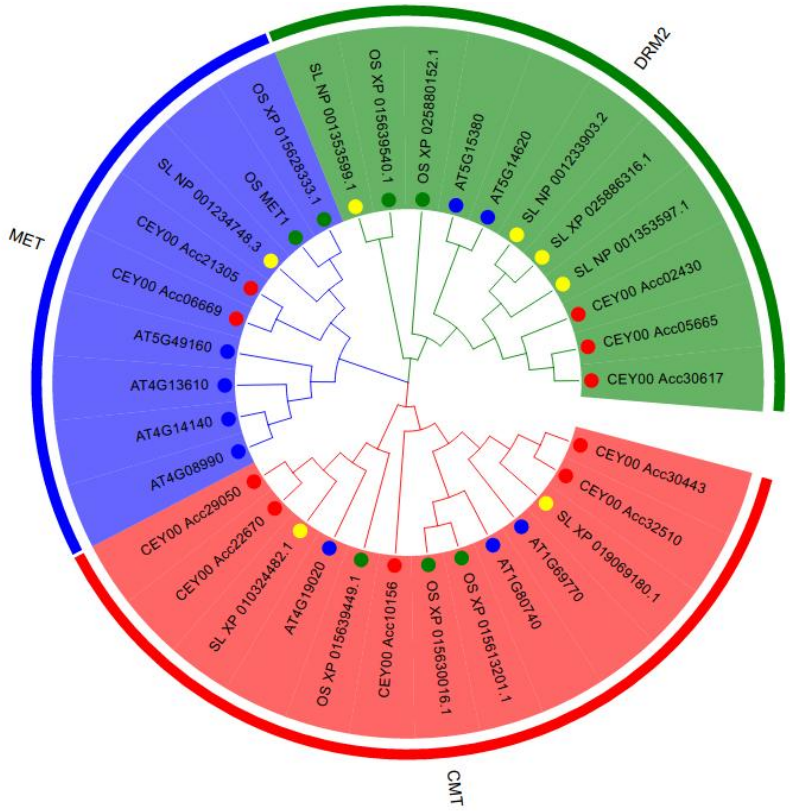
上述三步效果图：



3、群组标记。

Textalign:文字对齐方式；textorientation：文字横排还是竖排展示；linestyle：支持虚线展示；通过起始和终止控制画圈分组的范围。

```
CEY00_Acc30443,CEY00_Acc29050
   text=CMT,linecolor=red,linewidth=10,fontsize=16
AT4G08990,OS_XP_015628333.1
   text=MET,linecolor=blue,linewidth=10,fontsize=16
SL_NP_001353599.1,CEY00_Acc30617
   text=DRM2,linecolor=green,linewidth=10,fontsize=16
```



4、点图。

主要针对本物种的进化树外加其他数据进行联合绘图操作。可以在物种树中展示各个物种某个基因家族成员的数量，也可以展示基因表达量，或者启动子调控元件的数量等。

本处练习数据来源于 promoter.evolview.txt.class(在启动子分析中生成的文件)

##dot plots

!groups dot 1,dot 2,dot 3##三组数据

!colors #028482,#7ABA7A,yellow##颜色

!title Example of dot plot##图例标题

!legendstyle circle##图例是圆形

!showLegends 1##是否显示图例

!byarea 1

!grid0##是否显示背景网格线

!gridlabel 1##显示每组名字

!dotplots shape=circle,margin=2,colwidth=30

#colwidth ：每组宽度；margin：组之间间距

!showdataValue show=1,fontsize=12,fontcolor=white, valuesToHide=5:9##

valuesToHide 隐藏值得范围

5、柱形图。

本处练习数据来源于 promoter.evolview.txt.class(在启动子分析中生成的文件)

!groups group 1,group 2,group 3#分类，画堆积柱形图

!colors #028482,#7ABA7A,#B76EB8##可以设置颜色，几组写几个

!title barplot with data shown

!Fanplot##圈图必须加

!align##不花堆积柱形图，画三组单独柱形图

!itemHeightPX 10##柱形图宽度

!itemHeightPCT 80##柱形图总高度

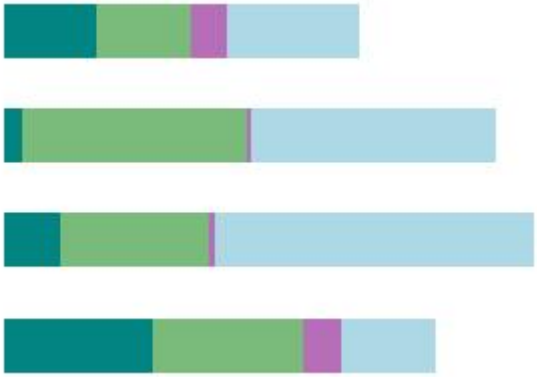
!plotWidth 150##打印的宽度

!showData##展示数据到柱形图中

!showDataFontSize 10##字体大小

!showDataFontColor white##字体颜色

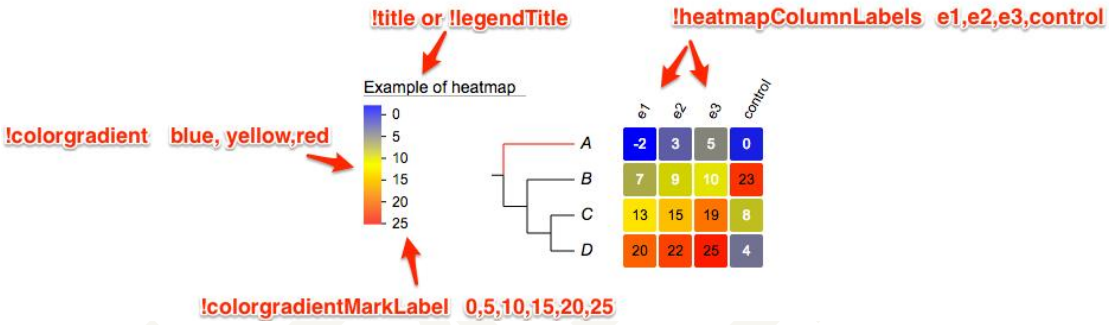
!showDataTextAlign middle##数字的位置



6、热图。

展示基因表达量等均可。

perl ~/Gene_Family/03.scripts/make_RNAseq.pl RNASeq.FPKM.txt RNASeq.FPKM.evolview



#heatmap

!title Example of heatmap##图例标题

!showLegends 1##展示图例

!colorgradient red,white,blue##使用的颜色

!colorgradientMarkLabel 0,5,10,15,20,25##图例标注尺度

!showHeatMapColumnLabel 1##开启展示热图每列名字

!heatmapColumnLabels e1,e2,e3,control##具体每列名字

!columLabelStyle

show=1,fontsize=20,fontitalic=0,fontbold=1,textangle=60,fontcolor=red##列名设置

!heatmap margin=2,colwidth=30,roundedcorner=2

margin 列之间间距；colwidth 单个热图元素的宽度。

!showdataValue show=1,fontsize=12,fontitalic=0,textalign=middle

具体示例：

!title RNAseq mRNA expression heatmap

!showLegends 1

!colorgradient red,white,blue

!colorgradientMarkLabel 0,2,4,6,22

!showHeatMapColumnLabel 1

!heatmapColumnLabels Root,stem,leaf,flower

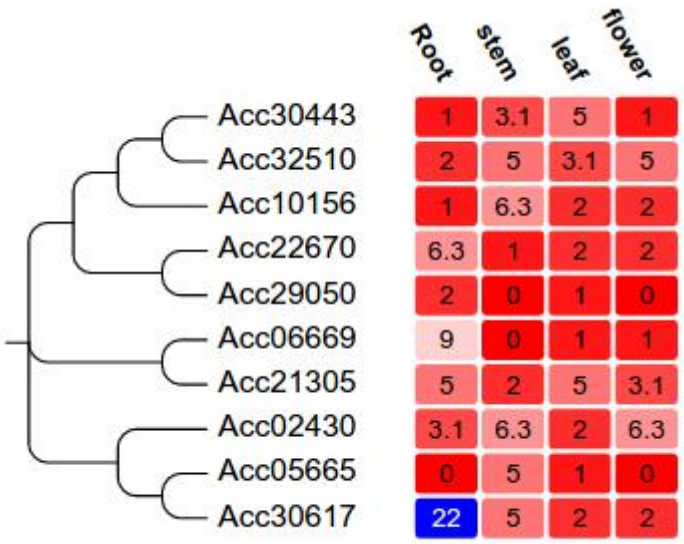
!columLabelStyle

show=1,fontsize=12,fontitalic=0,fontbold=1,textangle=60,fontcolor=black

!heatmap margin=2,colwidth=30,roundedcorner=2

!showdataValue show=1,fontsize=12,fontitalic=0,textalign=middle

示例图:



7、结构域图

基因名字+蛋白长度+结构域 1(起始，终止，名字)+ 结构域 2(起始，终止，名字)+…

可以采用此脚本生成配置文件(pfam 注释结果保存为 pfam.txt，同时此脚本生成 pfam.evolview.txt.gsds 用于 GSDS2 绘图):

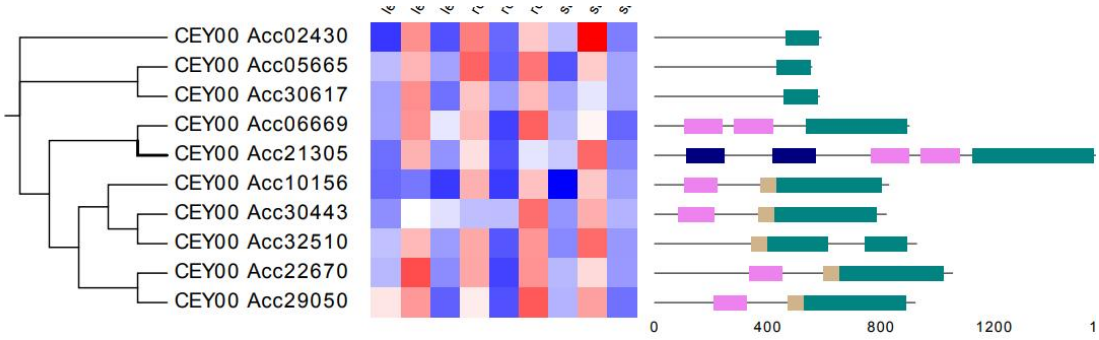
perl ~/Gene_Family/03.scripts/make_domain_format.pl final.gff3.stat pfam.txt pfam.evolview.txt

##the following modifiers are generated by EvolView!!

!groups DNA_methylase,Chromo,BAH,DNMT1-RFD

!colors #028482,Tan,Violet,Navy

##<<<< automatically generated modifiers end here!!



8、基因结构图。

基因名字+基因长度+UTR1(起始，终止，5 UTR)+ CDS(起始，终止，CDS)+ UTR2(起始，终止，3 UTR)

perl ~/Gene_Family/03.scripts/make_gene_struture_format.pl final.gff3 gene_struture.txt

示例：

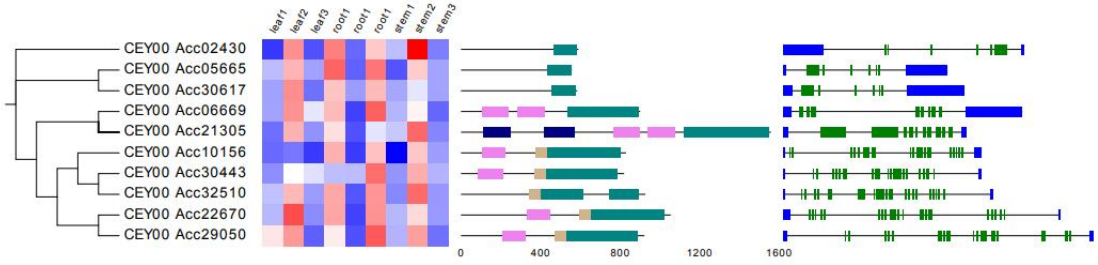
##the following modifiers are generated by EvolView!!

!groups UTR,CDS

!colors blue,green

!showLegends 1

##<<<< automatically generated modifiers end here!!



9、MEME 结构图。

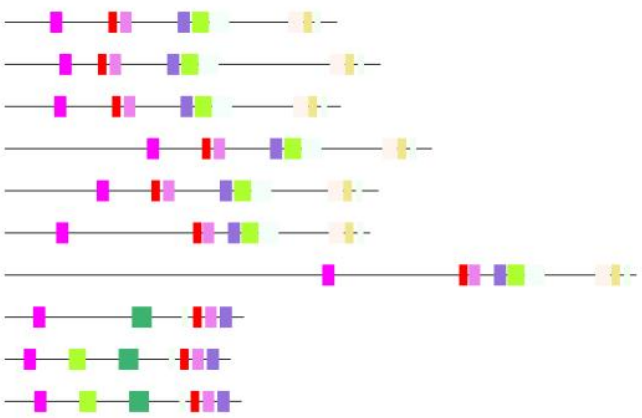
需要制作下表：基因名字+蛋白长度+Motif1（起始，终止，名字）+ Motif2（起始，终止，名字）

perl ~/Gene_Family/03.scripts/make_meme_motif.pl **meme.txt** final.gff3.stat **meme.evolview**

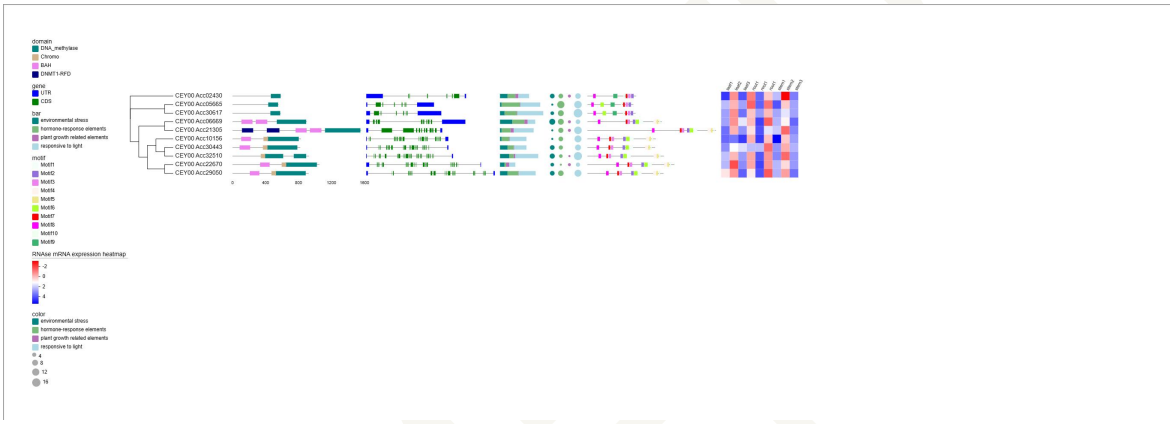
```
##the following modifiers are generated by EvolView!!
!showLegends  1
!groups
    Motif1,Motif2,Motif3,Motif4,Motif5,Motif6,Motif7,Motif8,Motif10,Motif9
```

!colors
MintCream,MediumPurple,Violet,SeaShell,Khaki,GreenYellow,Red,Magenta,HoneyDew,MediumSeaGreen
##<<<< automatically generated modifiers end here!!

CEY00_Acc32510 928 479,528,Motif1 401,431,Motif2
259,287,Motif3 803,840,Motif4 841,862,Motif5
436,478,Motif6 230,251,Motif7 135,165,Motif8
872,887,Motif10
CEY00_Acc06669 902 627,676,Motif1 551,581,Motif2
489,517,Motif3 800,837,Motif4 840,861,Motif5
584,626,Motif6 465,486,Motif7 127,157,Motif8
871,886,Motif10
CEY00_Acc02430 591 531,561,Motif2 495,523,Motif3
465,486,Motif7 70,100,Motif8 314,363,Motif9 437,452,Motif10
CEY00_Acc30443 821 505,554,Motif1 427,457,Motif2
285,313,Motif3 700,737,Motif4 738,759,Motif5
462,504,Motif6 256,277,Motif7 112,142,Motif8
765,780,Motif10
CEY00_Acc21305 1560 1284,1333,Motif1 1208,1238,Motif2
1146,1174,Motif3 1458,1495,Motif4 1498,1519,Motif5
1241,1283,Motif6 1122,1143,Motif7 784,814,Motif8
1529,1544,Motif10
CEY00_Acc22670 1055 733,782,Motif1 655,685,Motif2
516,544,Motif3 932,969,Motif4 970,991,Motif5
690,732,Motif6 487,508,Motif7 351,381,Motif8
1001,1016,Motif10
CEY00_Acc30617 585 525,555,Motif2 489,517,Motif3
184,226,Motif6 459,480,Motif7 73,103,Motif8 307,356,Motif9
431,446,Motif10
CEY00_Acc10156 830 512,561,Motif1 434,464,Motif2
294,322,Motif3 712,749,Motif4 750,771,Motif5
469,511,Motif6 265,286,Motif7 122,152,Motif8
781,796,Motif10
CEY00_Acc29050 923 609,658,Motif1 531,561,Motif2
391,419,Motif3 798,835,Motif4 836,857,Motif5
566,608,Motif6 362,383,Motif7 227,257,Motif8
867,882,Motif10
CEY00_Acc05665 558 499,529,Motif2 463,491,Motif3
158,200,Motif6 433,454,Motif7 47,77,Motif8 281,330,Motif9
405,420,Motif10



注意：所有绘图信息参见共享文件 evoview.all.txt



四、基因结构特征分析

得到的基因家族各个成员的 CDS 和 DNA 序列提交到 GSDS 2.0（Gene Structure Display Server）进行基因结构展示。

基因的顺式调控元件，则是提取每个基因上游 2kb 序列。将所有序列提交至 PlantCARE 数据库（http：// bioinformatics.psb.ugeniant.be/webtools/plantcare/html/），即可获得顺式作用调控元件。

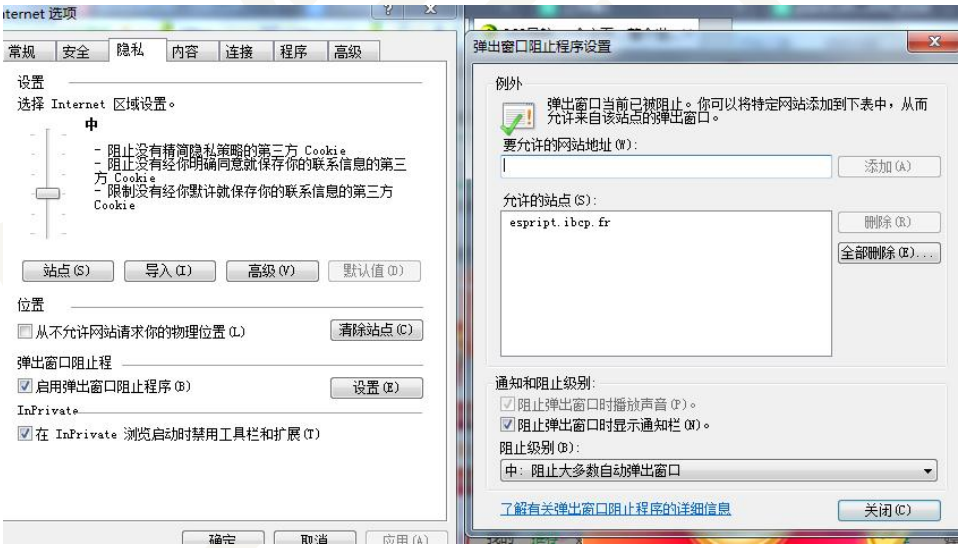
4.1 保守结构域序列展示

1、多序列比对结果的可视化**（使用 IE 浏览器，至少 11 版本）**。

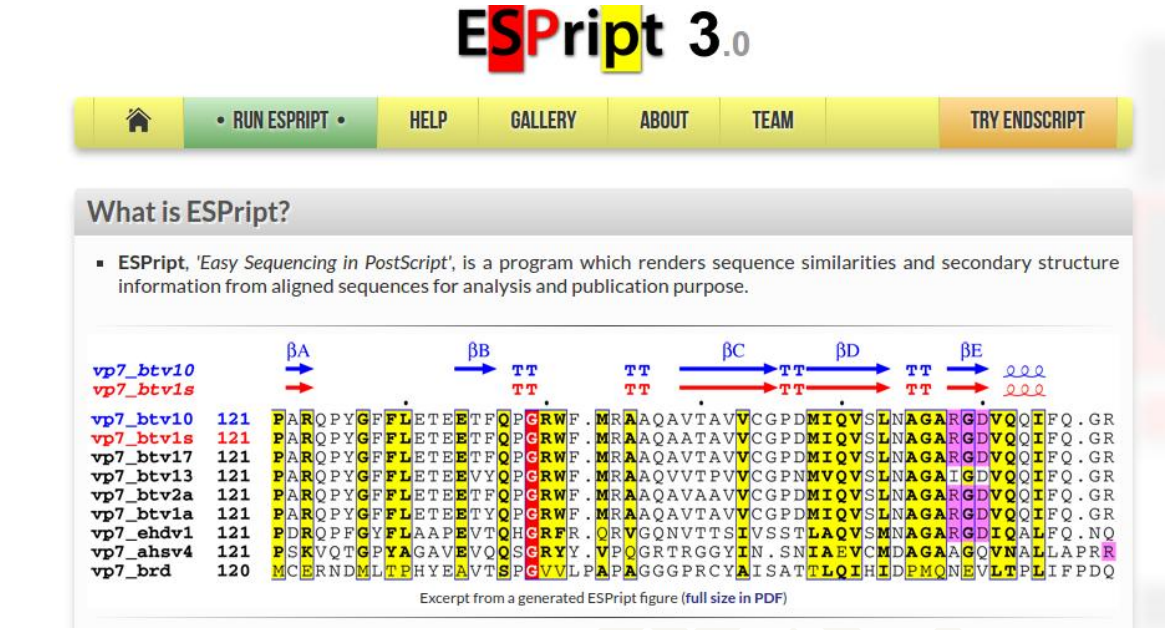
利用 MEGA 等工具比对完成后，保存成 fasta 或者 aln 格式的多序列比对文件，可以利用 ESPript 3.0 进行可视化输出。具体用法如下：

1）如果采用 IE 浏览器(**Microsoft Edge 应该不用**)必须设置允许弹窗 (一般在隐私设置里面）：

Internet 选项->隐私，然后按照下面截图设置：



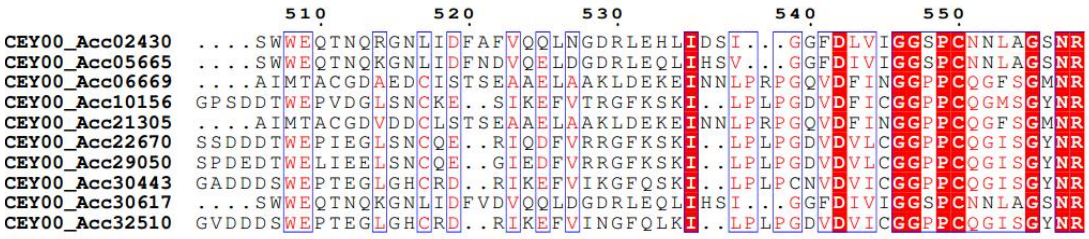
2）登陆网站 http://espript.ibcp.fr/ESPript/ESPript/，点击 RUN ESPRIPT。



3）上传多序列比对文件，必须是 FASTA、aln 等格式的。

4）点击 SUBMIT 提交即可。

5）结果。下载 pdf 格式，呈现结果如下图。



2、weblogo 展示多序列比对结果。

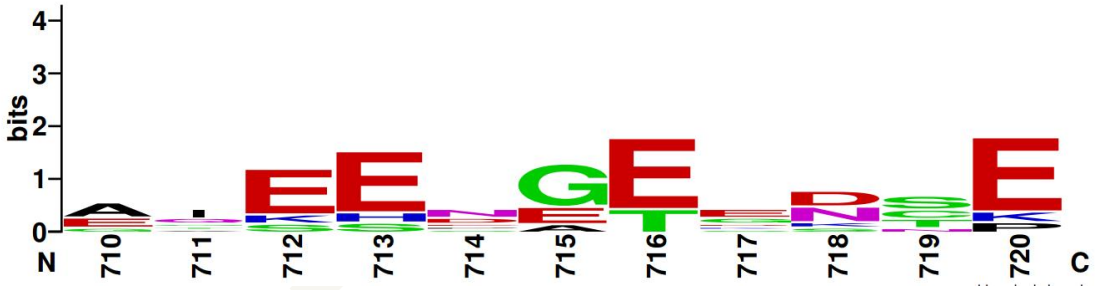
将多序列比对结果（主要是咱的结构域所在的位置的多序列比对结果），输入到 weblogo 中进行做图。具体为：

1）登陆网站：http://weblogo.berkeley.edu/logo.cgi

2）上传多序列比对文件或者 MEGA 粘贴出来的保守位点序列，并选择输出 pdf.

3)高级选项中选择输出的序列范围：

4)点击 Create Logo,生成如下图片：



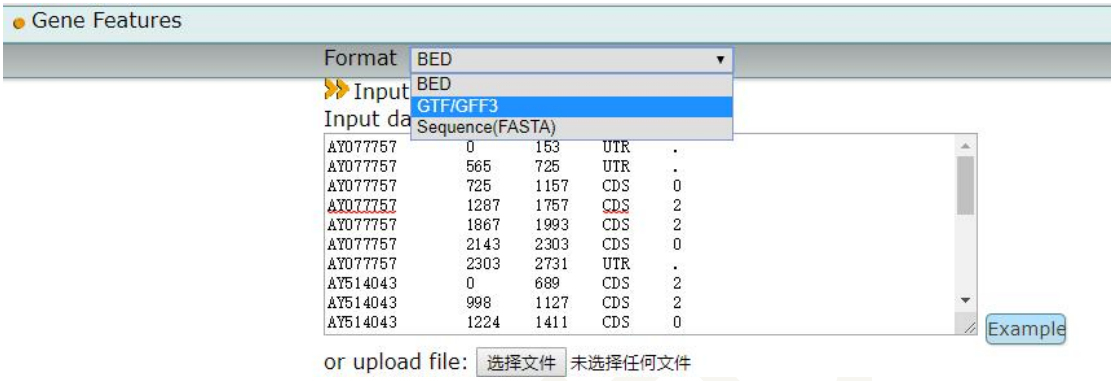
4.2 GSDS 基因结构展示

1、网址：

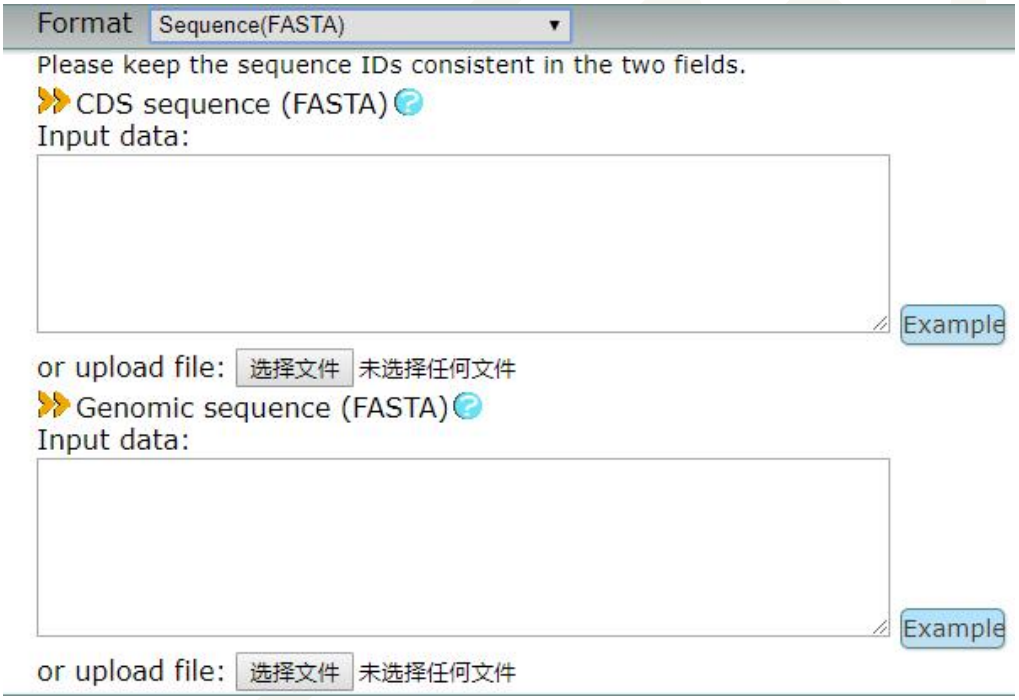
http://gsds.gao-lab.org/

2、用法。

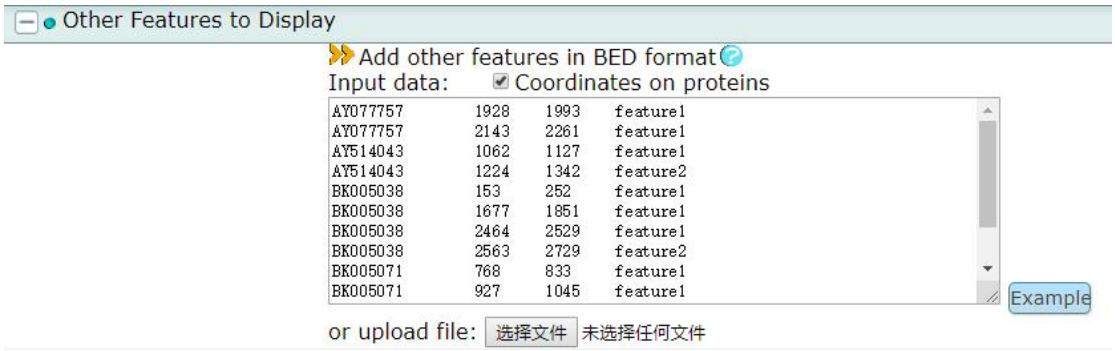
1）输入文件。输入文件支持 bed,gtf/gff3 和 FASTA 格式。其中 bed 可由 gff3 格式转换而来，gtf/gff3 格式一般通过数据库下载获得。下面以 bed 文件为例进行展示(前面提取 gff3 文件时已经获取，文件名字为 **final.gff3.bed**，直接上传即可)



如果没有上述格式文件，其也支持 FASTA 格式，但需要同时提供 cDNA 和基因组文件。



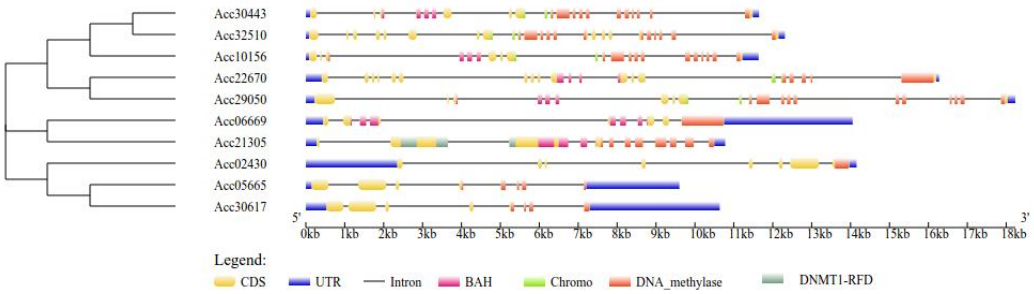
2）其他特征展示。可以在基因结构图上展示其他特征，如结构域位置。其中结构域位置来源于前面获取的 pfam.evolview.txt.gsds（make_domain_format.pl），注意此处结构域位置是基于蛋白序列的，所以要勾选 Coordinates on proteins。



3）进化树文件/输出顺序。进化树需要 Newick 格式的，用 IQ-TREE 或者 MEGA 建的研究物种的基因家族树文件，本处为 IQ_TREE.treefile。顺序参数则是与第一步输入文件 ID 相同的一列数据。

4）输出格式。支持 PNG 和 SVG。设置完成后点击 Submit。

5）结果。点击相应的输出格式即可。



4.3 启动子鉴定

此部分主要针对植物基因家族研究。主要采用 PlantCARE。

1、网址：http://bioinformatics.psb.ugent.be/webtools/plantcare/html/

2、用法。

此网址一次最多提交 100kb 序列，因此不要超量。如果你要预测数量较多，应分次提交。

1）获取启动子序列。

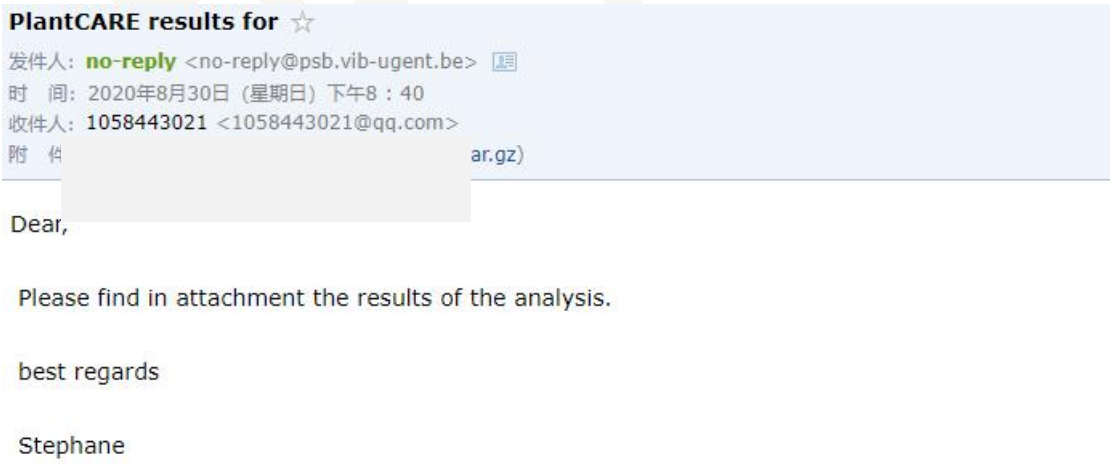
获取基因起始密码子上游 2 Kb 序列，并当大于 100Kb 时分割成多个文件。

perl ~/Gene_Family/03.scripts/get_promoter.pl -p final.gff3.stat -s Actinidia_chinensis.Red5_PS1_1.69.0.dna.toplevel.fa -o Ac.promoter.fa -len 2000

2)提交。填写邮箱，并上传序列即可。

3）结果。

运行完成后，会发送邮件到上述指定的邮箱中



解压后会产生一个 TAB 格式文本文件（plantCARE_output_PlantCARE_2670.tab），可以用 excel 打开，进行相关的统计。此外还有网页文件，可以用 IE 浏览器打开。

整理数据准备的原则：

1、保留有功能注释的（或比较关注的）元件，无实际功能的不做保留（核心启动子：CAAT-box TATA-box 也丢掉）。

2、一般分为至少四类：激素类，光反应、环境胁迫、生长发育

perl ~/Gene_Family/03.scripts/make_promoter_evolview.pl promoter.txt 2000 promoter.evolview.txt

注：promoter.txt 为整理好的 promoter 文件

2000：为提取上游启动子序列长度

结果生成 promoter.evolview.txt.class 和 promoter.evolview.txt

五、共线性及复制基因分析

主要阐明基因家族扩张的机制与进化。

5.1 MCScanX 共线性分析

5.1.1 MCScanX 软件安装

下载：http://chibba.pgml.uga.edu/mcscan2/MCScanX.zip

解压：unzip MCScanX.zip

安装：cd MCScanX

vi msa.cc，在顶部加上#include <unistd.h>

vi dissect_multiple_alignment.cc，在顶部加上#include <getopt.h>

vi　detect_collinear_tandem_arrays.cc，在顶部加上#include <getopt.h>

make

cd downstream_analyses/

make

5.1.2 输入文件

１、BLAST 文件。

BLAST 为你要分析物种基因的蛋白质文件或者 CDS 文件与参考物种（或自身）的蛋白质或者 CDS 文件的 BLAST 比对结果文件，格式一般为采用 BLAST 的 m8 或者 BLAST+的 -outfmt 6 格式。此处以 BLAST+为例，阐述此文件的制备。

进入分析目录：cd　~/Gene_Family/02.data/

1）BLAST 建库。

如果是利用参考蛋白建库，比较两个物种的共线性：

makeblastdb -in Ac_Longest/Longest/Ac.longest.protein.fa -dbtype prot -out refpep.db

makeblastdb 参数介绍：

-dbtype <String, `nucl', `prot'>：　数据库类型，核酸或者蛋白，选择其一。

-in <File_In>：　输入蛋白文件

-out <String>：　创建的数据库名字

注意：此处以蛋白质为例阐述，若数据库为核酸序列 -dbtype 选择 nucl

2）BLAST 比对。

注意：因为 blastp 运行时长较长，所以大家用我已经跑好的 blastp 结果来练习，实际项目，不需要拷贝，用自己产生 blastp 结果！

cp PATH/02.data/Ac_Vs_Ac.blast ~/Gene_Family/02.data/

cp PATH/02.data/Sl_Vs_Ac.blast ~/Gene_Family/02.data/

注意：大家上课时候就不要运行下面这两个 blastp 了，因为运行时长较长，所以大家用我已经跑好的 blastp（课后练习可以用自己跑的），通过上述命令拷贝的结果来练习即可！

两个物种的 BLAST，鉴定物种之间的共线性:

nohup *blastp -query SL_Longest/Longest/Sl.longest.protein.fa -db refpep.db -out Sl_Vs_Ac.blast1 -evalue 1e-5 -num_threads 1 -outfmt 6* ***&***

awk '{if($3>=50)print }' Sl_Vs_Ac.blast1 >Sl_Vs_Ac.blast

自身做 BLAST，鉴定内部共线性区块，按照如下命令书写：

nohup blastp -query Ac_Longest/Longest/Ac.longest.protein.fa -db refpep.db -out Ac_Vs_Ac.blast -evalue 1e-5 -num_threads 1 -outfmt 6 &

注：

blastp:将待查询的蛋白质序列一起对蛋白质序列数据库进行查询；

blastn:将待查询的核酸序列及其互补序列一起对核酸序列数据库进行查询

3）示例结果如下：

Solyc04g007470.3	CEY00_Acc32928	46.838	506	234	16	3	475	1	504	2.38e-130	387
Solyc04g007470.3	CEY00_Acc14054	46.278	497	223	14	3	456	1	496	1.74e-121	364
Solyc04g007470.3	CEY00_Acc13564	46.154	117	59	2	3	116	1	116	9.04e-28	114
Solyc04g007470.3	CEY00_Acc19175	49.074	108	52	1	3	107	1	108	2.43e-27	113
Solyc04g007470.3	CEY00_Acc26509	48.148	108	53	1	3	107	1	108	2.02e-26	110
Solyc04g007470.3	CEY00_Acc08292	24.587	484	299	14	7	456	14	465	3.73e-25	107
Solyc04g007470.3	CEY00_Acc02517	45.299	117	61	1	6	119	14	130	3.90e-25	107
Solyc04g007470.3	CEY00_Acc26483	47.321	112	55	2	3	111	1	111	1.42e-24	105

2、 GFF 文件。

此处我们要从两个物种的 GFF3 文件中提取上述做 BLAST 的基因在染色体上的位置信息文件。前面提取最长转录本时已经获得。该文件具体格式为：

基因组序列 ID	基因 ID	基因起始位置	基因终止位置

我们进一步仔细看基因 ID 在 BLAST 和 GFF 文件是否匹配。

匹配然后将你要分析物种的和参考物种的基因位置信息最终应合并到一个文件中。我们可以将其命名为 Sl_Vs_Ac.gff。

```
cat SL_Longest/Longest/Sl.gff_for_MCSCANX.txt Ac_Longest/Longest/Ac.gff_for_MCSCANX.txt >Sl_Vs_Ac.gff
```

```
ln -fns Ac_Longest/Longest/Ac.gff_for_MCSCANX.txt Ac_Vs_Ac.gff
```

3、注意事项。

1、blast 和 gff 文件命名规则为*.gff 和*.blast，即后缀名不变，前缀名字必须一致！

2、blast 和 gff 文件里面的基因 ID 必须名字必须一致！

5.1.3 运行命令

1、参数介绍。

-s：要求一个共线性区块应具有的最少基因数目

-m：一个 block 中允许的最大空位数目，降低默认值，可以保证更好的共线性结果。

2、运行命令

```
MCScanX Sl_Vs_Ac -m 15
```

```
MCScanX Ac_Vs_Ac -m 15
```

若产生结果不理想，可以过滤一下 BLAST 结果（根据 identity 和比对长度），或者使用-m 降低 gap 数目（默认 25，可以降低到 5 或者更小）。

5.1.4 结果介绍

生成一个文本文件（all.collinearity）和包含网页的文件夹（all.html）。

1、文本文件。

给出了每一个 Block 由那些基因组成。下面是一个例子：

```
############### Parameters ###############
# MATCH_SCORE: 50
# MATCH_SIZE: 5
# GAP_PENALTY: -1
# OVERLAP_WINDOW: 5
# E_VALUE: 1e-05
# MAX GAPS: 15
############### Statistics ###############
# Number of collinear genes: 22577, Percentage: 68.32
# Number of all genes: 33044
#########################################
## Alignment 0: score=329.0 e_value=2.3e-10 N=7 LG1&LG10 plus
  0-  0:        CEY00_Acc00952  CEY00_Acc11808    3e-10
  0-  1:        CEY00_Acc00957  CEY00_Acc11810    1e-41
  0-  2:        CEY00_Acc00961  CEY00_Acc11815    4e-87
  0-  3:        CEY00_Acc00966  CEY00_Acc11819    1e-51
  0-  4:        CEY00_Acc00971  CEY00_Acc11822   2e-127
  0-  5:        CEY00_Acc00972  CEY00_Acc11824   7e-127
  0-  6:        CEY00_Acc00977  CEY00_Acc11827    9e-25
```

最上面######Parameters 部分是我们运行的参数

往下######Statistics 部分是我们分析的结果统计，其中：

Number of collinear genes: , Percentage:###共线性基因数目及占比。

再往下是具体的共线性信息了：

Alignment 0: score=329.0 e_value=2.3e-10 N=7 LG1&LG10 plus

0 为 Block 的 ID，N 表示在这个 Block 中有多少对基因；LG1&LG10 两个基因组序列的染色体 ID；

plus 共线性的方向。

0- 0: CEY00_Acc00952 CEY00_Acc11808 3e-10

0：共线性 Block 的 ID

2：这个 Block 的第几个基因

0：BLAST 的 E 值。

5.2 复制基因鉴定和 Ka/Ks 计算

我们主要关注 segment duplication（全基因组复制或大片段重复产生的复制基因对）和 tandem duplication（串联重复）这两种类型复制基因。

1、提取复制基因对。

输入文件 gene_family_file 为一行信息，需自行制备此文件：

DNA_methyltransferases CEY00_Acc02430 CEY00_Acc05665 CEY00_Acc06669 CEY00_Acc10156 CEY00_Acc21305 CEY00_Acc22670 CEY00_Acc29050 CEY00_Acc30443 CEY00_Acc30617 CEY00_Acc32510

注意：使用 vi 创建文件后，粘贴进入，观察每个基因是否有制表符或空格分割

执行程序：

perl ~/Gene_Family/03.scripts/detect_collinearity_within_gene_families.pl -i gene_family_file -d Ac_Vs_Ac.collinearity -o Ac.duplicated_pairs

2、对复制基因对计算 Ka/Ks。

对于一个没有受到自然选择压力的基因来说， Ka/Ks 一般会等于 1。但实际情况下，这个比值都是远小于 1 的，因为从蛋白质进化的角度而言，为了维持蛋白质功能的稳定性，其非同义替换的数量往往较少，因此非同义突变 Ka 要远小于 Ks。因此我们根据 Ka/Ks 大小，将基因划分为三类：

Ka>Ks->Ka/Ks > 1，基因受正选择(positive selection)

Ka = Ks->Ka/Ks=1，基因中性进化(neutral evolution)

Ka<Ks->Ka/Ks < 1，基因受纯化选择(purify selection)

时间计算一般采用 此公式：T = Ks/2λ，其中λ需查阅文献，不同物种不太一样。

具体计算步骤如下：

(1)mafft 比对。

perl ~/Gene_Family/03.scripts/mafft_align.pl -list Ac.duplicated_pairs --pep Ac_Longest/Longest/Ac.longest.protein.fa --mafft mafft --outdir Duplicate_Mafft

(2)反转为密码子比对。（为避免大家路径问题，程序调用改为绝对路径）

ls Duplicate_Mafft/*.mfa|awk '{print " perl ~/Gene_Family/03.scripts/pepMfa_to_cdsMfa.pl "$1" 02.data/Ac_Longest/Longest/Ac.longest.cds.fa 1>"$1".condon 2>log.txt" }' |sh

(3)转化为 axt 格式。（为避免大家路径问题，程序调用改为绝对路径）

ls Duplicate_Mafft/*.mfa.condon |awk '{print "perl ~/Gene_Family/03.scripts/mfa_to_axt.pl "$1" " }' |sh

(4)KaKs_Calculator 计算 ka ks。

合并所有的 axt 文件：cat Duplicate_Mafft/*axt >all.duplicated.pairs.axt

执行 ka ks 计算，注意一个模型选择参数(NG,YN 等)即可。

~/Gene_Family/01.Soft/KaKs_Calculator2.0/bin/Linux/KaKs_Calculator -i all.duplicated.pairs.axt -o ka_ks.xls -m NG

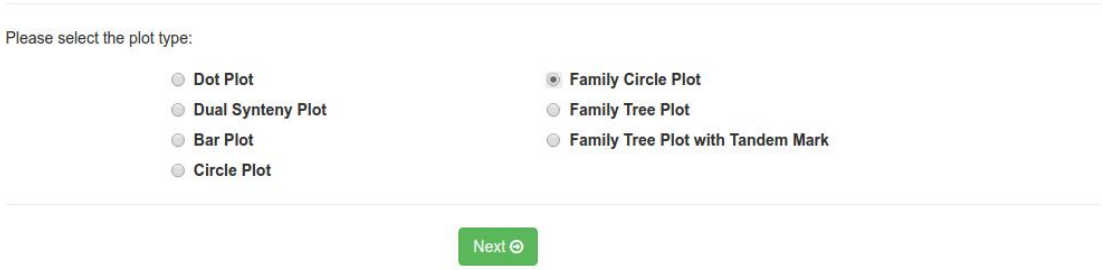
5.3 基因组内共线性圈图展示

采用 VGSC 进行在线绘图。

网站：https://dvb.ac.cn/vgsc2/service/home.php

用法：

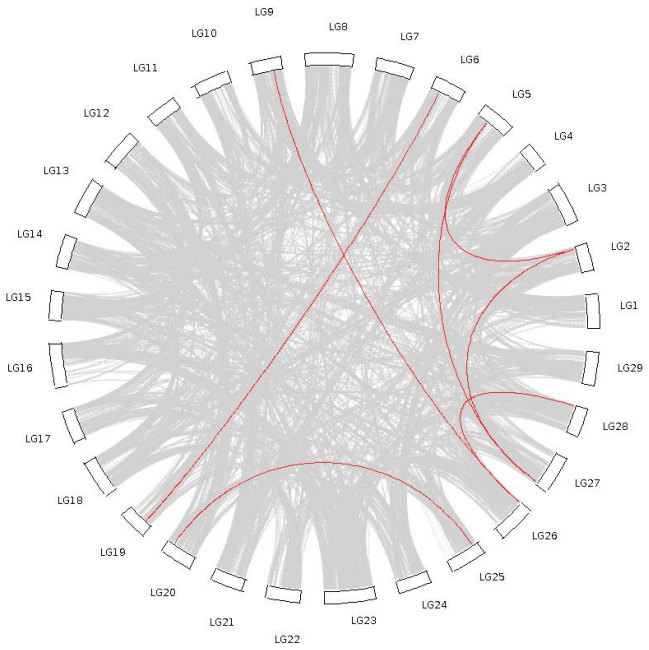
1、选择 Family Circle Plot



2、上传数据。

3、结果输出。

4、结果。红色部分为共线性的基因家族成员，背景是全基因组共线性。



5.4 基因组间共线性展示（JCVI）

1、安装。

JCVI 下载地址：https://github.com/tanghaibao/jcvi，安装最好让管理员安装：

pip install jcvi

2、使用。

1）准备 seqid 文件。采用 vi 命令创建。

seqids 文件是用来选择对哪些染色体进行画图，文件格式如下所示：

1,2,3,4,5,6,7,8,9,10,11,12,

LG1,LG2,LG3,LG4,LG5,LG6,LG7,LG8,LG9,LG10,LG11,LG12,LG13,LG14,LG15,LG16,LG17,LG18,LG19,LG20,LG21,LG22,LG23,LG24,LG25,LG26,LG27,LG28,LG29

其中第一行为物种 1 所包含的 12 个染色体，第二行为物种 2 包含的 29 个染色体。

2)运行。

需要多个输入文件，见下面例子：

perl ~/Gene_Family/03.scripts/JCVI.pl SL_Longest/Longest/Sl.longest.gff3 SL_Longest/Longest/Sl.longest.protein.fa SL Ac_Longest/Longest/Ac.longest.gff3 Ac_Longest/Longest/Ac.longest.protein.fa Ac Sl_Vs_Ac.collinearity Sl_Vs_Ac.blast seqid

其中在生成的文件中：

前两行展示了两个物种的基础绘图信息，如果画三个物种，那么可以写三行，以此类推。其中前三列分别指定了物种 1 和物种 2 染色体在图上 Y 轴的位置，X 轴的起始和终止位置，即我们共线性图在画布上的位置；第四列颜色可以默认，不填写；第五列是旋转角度，默认零度则是画两条平行的；第六列是标记物种名；第七列是染色体 ID 标记位置；第八列是 bed 文件。最后一行是画共线性的数据文件，其中 0 和 1 代表前述第一个物种和第二个物种，进行共线性的绘制。后面跟上的是相关的数据。如果有 3 个物种，可以继续写 1 和 2。如下例：

#y, xstart, xend, rotation, color, label, va, bed

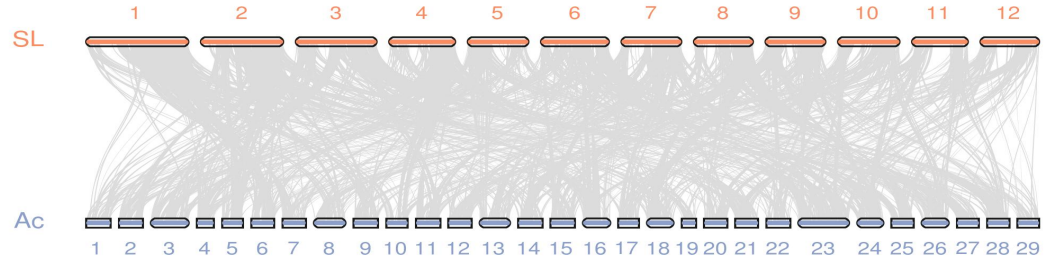
.6, .1, .95, 0, ,SL,top,SL.bed

4, .1, .95, 0, ,Ac,bottom,Ac.bed

edges

e, 0, 1, SL_Vs_Ac.collinearity.anchors.simple

3)结果图。



4)高亮基因家族成员。

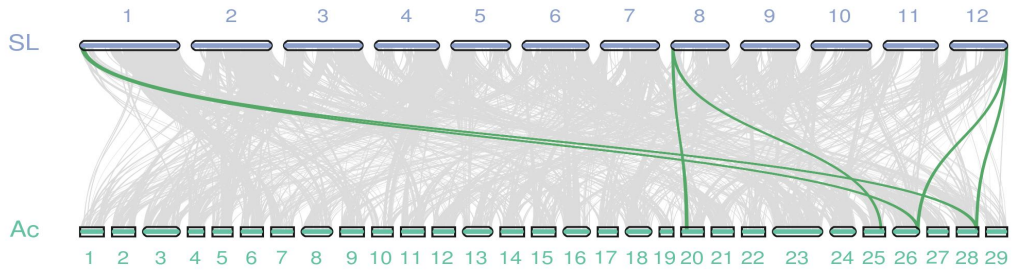
修改 SL_Vs_Ac.collinearity.anchors.simple 文件，注意颜色值为 16 进制，*隔开； 没有添加颜色值的行默认灰色。我新提供一个脚本可以实现自动处理，形成新的 SL_Vs_Ac.collinearity.anchors.simple 文件：

```
perl ~/Gene_Family/highlight_coline_gene_block.pl SL_Vs_Ac.collinearity.anchors gene_family_file SL_Vs_Ac.collinearity.anchors.simple
```

```
g*Solyc01g005510.3      Solyc01g006660.2      CEY00_Acc30423  CEY00_Acc30489  87      -
Solyc01g028930.3        Solyc01g044550.2      CEY00_Acc30458  CEY00_Acc30478  25      -
Solyc01g008490.3        Solyc01g009020.3      CEY00_Acc30349  CEY00_Acc30388  44      -
Solyc01g081570.3        Solyc01g073940.3      CEY00_Acc30430  CEY00_Acc30469  39      -
Solyc01g005210.3        Solyc01g005460.3      CEY00_Acc31138  CEY00_Acc31164  24      +
Solyc01g107300.3        Solyc01g107990.3      CEY00_Acc30968  CEY00_Acc30999  49      +
Solyc01g010135.1        Solyc01g010410.3      CEY00_Acc31419  CEY00_Acc31474  33      +
Solyc01g079210.3        Solyc01g079650.3      CEY00_Acc31527  CEY00_Acc31560  40      +
Solyc01g009070.3        Solyc01g009560.1      CEY00_Acc32044  CEY00_Acc32116  60      +
Solyc01g006900.3        Solyc01g007840.3      CEY00_Acc32049  CEY00_Acc32081  41      +
Solyc01g008730.3        Solyc01g009070.3      CEY00_Acc32391  CEY00_Acc32451  45      +
g*Solyc01g005510.3      Solyc01g006650.2      CEY00_Acc32482  CEY00_Acc32563  95      -
```

然后重新运行：

```
python3 ~/Gene_Family/03.scripts/karyotype.py seqid SL_Vs_Ac.layout --format pdf --notex
python3 ~/Gene_Family/03.scripts/karyotype.py seqid SL_Vs_Ac.layout --format png --notex
```



六、转录组分析（数据下载）

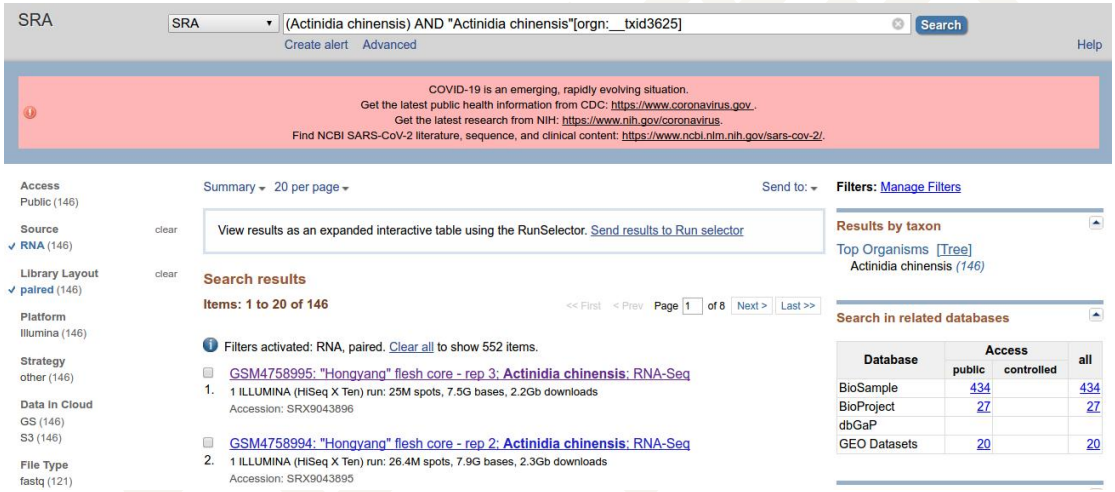
为了深入了解基因家族各成员在不同组织，发育阶段下的表达的特征，我们从 NCBI SRA 数据库获得了转录组测序数据。

6.1 SRA 数据下载

SRA（Sequence Read Archive）是 NCBI SRA 数据库用于存储二代测序原始数据的一种压缩格式，我们从 NCBI 下载的公共测序数据，一般是此格式。

NCBI 网址：https://www.ncbi.nlm.nih.gov/sra

1、检索和筛选。我们研究对象是猕猴桃(**Actinidia chinensis)**的转录组数据。

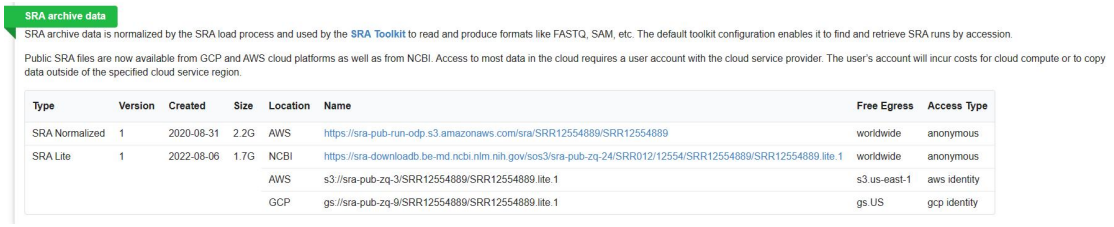


2、下载。点击第一个，出现下面界面：

Runs: 1 run, 25M spots, 7.5G bases, 2.2Gb

Run	# of Spots	# of Bases	Size	Published
SRR12554889	24,957,118	7.5G	2.2Gb	2020-09-01

点击 SRR12554889



SRA archive data

SRA archive data is normalized by the SRA load process and used by the SRA Toolkit to read and produce formats like FASTQ, SAM, etc. The default toolkit configuration enables it to find and retrieve SRA runs by accession.

Public SRA files are now available from GCP and AWS cloud platforms as well as from NCBI. Access to most data in the cloud requires a user account with the cloud service provider. The user's account will incur costs for cloud compute or to copy data outside of the specified cloud service region.

Type	Version	Created	Size	Location	Name	Free Egress	Access Type
SRA Normalized	1	2020-08-31	2.2G	AWS	https://sra-pub-run-odp.s3.amazonaws.com/sra/SRR12554889/SRR12554889	worldwide	anonymous
SRA Lite	1	2022-08-06	1.7G	NCBI	https://sra-downloadb.be-md.ncbi.nlm.nih.gov/sos3/sra-pub-zq-24/SRR012/12554/SRR12554889/SRR12554889.lite.1	worldwide	anonymous
				AWS	s3://sra-pub-zq-3/SRR12554889/SRR12554889.lite.1	s3.us-east-1	aws identity
				GCP	gs://sra-pub-zq-9/SRR12554889/SRR12554889.lite.1	gs.US	gcp identity

点击下载连接即可：

https://sra-pub-run-odp.s3.amazonaws.com/sra/SRR12554889/SRR12554889

3、转换为 fastq。主要采用 fastq-dump。

fastq-dump --split-files SRR12554889

注：下载文件在 RNAseq/SRR12554889

４、最终得到：SRR12554889_1.fastq 和 SRR12554889_1.fastq