家族基因的鉴定分析

家族基因鉴定主要有 blast 和 hmmer 两种方法，根据不同的家族基因情况，单独使用其中的一种或者两种同时使用取并集，候选基因后续验证。在进行家族基因鉴定分析前，准备工作主要是查找家族相关文章，下载已发表家族基因序列和（或）pfam 数据库中对应 hmm 文件。

1 BLAST 的使用

BLAST+与 BLAST 相比，有很多改进和提高，NCBI 强烈推荐放弃 BLAST，使用 BLAST+。BLAST+主要包括四个常用程序（makeblastdb，blastn，blastp，blastx）和其他一些屏蔽重复序列等程序，makeblastdb 使用 fasta 文件生成本地搜索库，blastn、blastp、blastx 三支程序包含有相似的命令参数。

BLAST 本地化安装

首先，在 ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST 目录下载对应系统版本的程序包。win 环境下，推荐安装在非系统盘，安装成功后在安装路径（例如 D:\blast-2.2.29+）下生成 bin、doc 两个子目录，其中 bin 是程序目录，doc 是文档目录，另外在安装路径下新建一个文件夹“data”保存数据库文件。

添加环境标量，右键点击“我的电脑”－“属性”，然后选择“高级系统设置”标签“环境变量”，在用户变量下方“Path”添加其变量值“D:\blast-2.2.29+\bin”，Path 原有变量值不可更改，不同变量间使用“;”分割。此时点击“新建”－变量名“BLASTDB”，变量值为“D:\blast-2.2.29+\data”(即数据库路径)

点击 Windows 的“开始”菜单，输入“cmd”调出 MS-DOS 命令行，转到 Blast 安装目录，输入命令“blastn -version”即可查看版本，如图说明本地 blast 已经安装成功。

构建本地库

makeblastdb.exe [-h] [-help] [-in input_file] [-input_type type]
-dbtype molecule_type [-title database_title] [-parse_seqids]
[-hash_index] [-mask_data mask_data_files] [-mask_id mask_algo_ids]
[-mask_desc mask_algo_descriptions] [-gi_mask]
[-gi_mask_name gi_based_mask_names] [-out database_name]
[-max_file_sz number_of_bytes] [-taxid TaxID] [--taxid_map TaxIDMapFile]
[-logfile File_Name] [-version]

示例：

makeblastdb -in reference.fa -dbtype nucl -parse_seqids -hash_index -out db_name

必须参数说明

-in：待格式化的序列文件

-dbtype：数据库类型，prot 或 nucl 必选其一

-out：数据库名，后续 blast 使用

复制将要用来构建blast数据库的fasta序列文件到data文件夹（例如D:\blast-2.2.29+\data）下，运行上面命令，生成 blast 数据库

本地 blast 搜索

根据已报道的家族基因氨基酸或核苷酸序列同源搜索目标数据库，具体有 blastn 和 blastp 两种方法，原理、实现上高度相似，通过设置 e-value 控制假阳性，家族基因鉴定这一步设置通常较为宽松（一般为 evalue=1e-5，不同家族可响应调整设置），尽可能发现全部目标物种全部家族基因，后续验证步骤排除。

示例：

blastn -query seq.fa -db db_name -outfmt 7 -out blast_results.txt -evalue 1e-5

必须参数说明:

-query： 输入文件路径及文件名

-out：输出文件路径及文件名

-db：格式化的数据库路径（数据库加入环境标量后，不需路径）及数据库名

-outfmt：输出文件格式，默认 xml，须更改为 7（输出为文本）

-evalue：设置输出结果的 e-value 值，默认 10

运行 blastp 时命令行参数设置为和 blastn 一致即可。

结果解释

```
# BLASTP 2.2.18 [Mar-02-2008]
# Query: At1g06400.1
# Database: D:\SeqHunter\db\G.raimondii_pep
# Fields: Query id, Subject id, % identity, alignment length, mismatches, gap openings, q. start, q. end, s. start, s. end, e-value, bit score
At1g06400.1 Gorai.009G160700.1  77.63   219 45  2   1   216 1   218 1e-098   355
At1g06400.1 Gorai.002G065100.1  78.24   216 45  2   1   215 1   215 3e-097   351
At1g06400.1 Gorai.010G161100.1  75.69   218 51  1   1   216 1   218 5e-097   350
At1g06400.1 Gorai.011G037400.1  76.26   219 48  2   1   216 1   218 1e-096   348
```

第二列对应即是 blast 所有匹配上的序列 ID，输出文件使用 excel 打开，选中第二列，复制粘贴到新的 excel 文件，去重复即得到 blast 发现的所有候选家族基因。

SeqHunter 工具（注意：SeqHunter 文件夹需放在 D 盘根目录下）集成了 BLAST 2.2.18，家族基因鉴定中也可用来进行家族基因同源搜索。

2 HMMER 的使用

HMMER 本地配置

首先在 http://hmmer.org 下载系统对应安装包，win 系统解压到适当路径即可，添加完整路径到系统环境变量（添加系统环境变量详细方法如上）。hmmer 包含以下几个程序：

phmmer：与 blastp 类似，使用一个蛋白质序列搜索蛋白质序列库；

jackhmmer：与 psiBlast 类似，蛋白质序列迭代搜索蛋白质序列库；

hmmbuild：用多重比对序列构建 HMM 模型；

hmmsearch：使用 HMM 模型搜索序列库（家族基因鉴定）；

hmmscan：使用序列搜索 HMM 库；

hmmalign：使用 HMM 为线索，构建多重比对序列；

hmmconvert：转换 HMM 格式

hmmemit：从 HMM 模型中，得到一个模式序列；

hmmfetch：通过名字或者接受号从 HMM 库中取回一个 HMM 模型；
hmmpress：格式化 HMM 数据库，以便于 hmmscan 搜索使用；
hmmstat：显示 HMM 数据库的统计信息；

使用 HMM 模型搜索序列数据库

hmm 文件可直接从 pfam 网站（http://pfam.xfam.org/）下载或者自己根据 Stockholm、FASTA 格式的多重比对序列文件使用 hmmbuild 构建。使用多重比对序列文件（globins4.sto）构建 HMM 文件命令如下：

hmmbuild globins4.hmm globins4.sto

hmmsearch 使用 globins4.hmm 文件搜索蛋白质序列数据库，蛋白质序列数据库须为 FASTA 格式，cmd 运行命令如下：

hmmsearch globins4.hmm protein.fasta > globins4.out

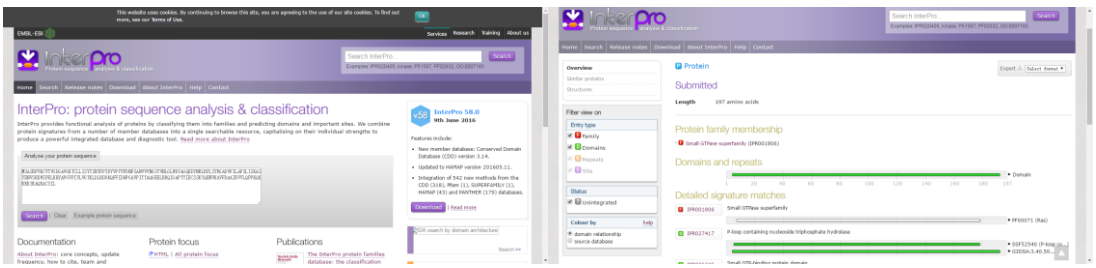
hmmsearch 输出解释

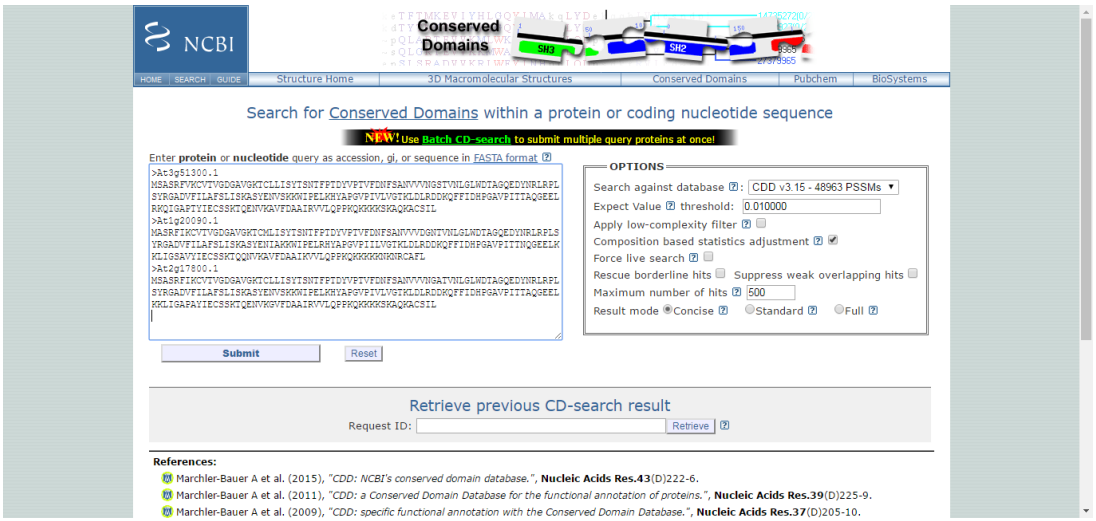
```
# hmmscan :: search sequence(s) against a profile database
# HMMER 3.0 (March 2010); http://hmmer.org/
# Copyright (C) 2010 Howard Hughes Medical Institute.
# Freely distributed under the GNU General Public License (GPLv3).
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
# query sequence file:             7LESS_DROME
# target HMM database:             minifam
# per-seq hits tabular output:     7LESS.tbl
# per-dom hits tabular output:     7LESS.domtbl
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

Query:       7LESS_DROME  [L=2554]
Accession:   P13368
Description: RecName: Full=Protein sevenless;          EC=2.7.10.1;
Scores for complete sequence (score includes all domains):
   --- full sequence ---   --- best 1 domain ---    -#dom-
    E-value  score  bias    E-value  score  bias    exp  N  Model    Description
    ------- ------ -----    ------- ------ -----   ---- --  --------  -----------
    5.6e-57  178.0   0.4    3.5e-16   47.2   0.7    9.4  9  fn3       Fibronectin type III domain
    1.1e-43  137.2   0.0    1.7e-43  136.5   0.0    1.3  1  Pkinase   Protein kinase domain
```

Model 对应列即为符合 HMM 文件设定条件的氨基酸序列 ID，输出文件可使用 excel 操作得到候选基因 ID。

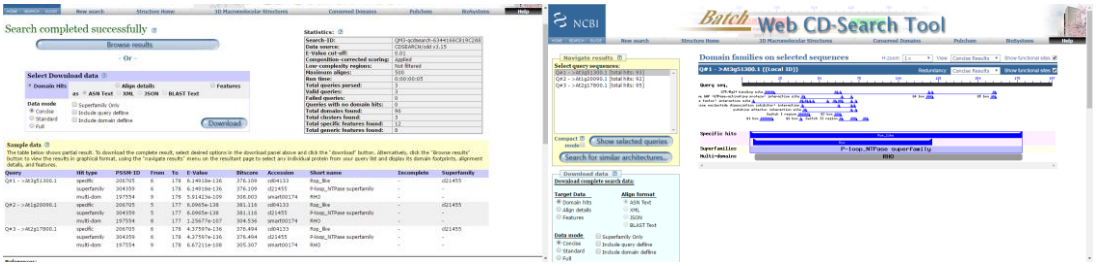
3 在线数据库分析验证

经过前两步搜索我们已经得到了一大批候选家族基因，其中部分基因可能是相似性较高的其他家族，或者基因错误注释为两个基因，在或两个基因嵌合转录本，因此需对候选基因基因进行验证，在线网站 InterPro（http://www.ebi.ac.uk/interpro/）和 CDD（http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi）常用来进行验证。其中 InterPro 网站预测结果结果较准确，但是不能批量进行，只能对氨基酸序列进行注释；而 CDD 数据库氨基酸和核苷酸序列都能够进行预测，可批量运行。





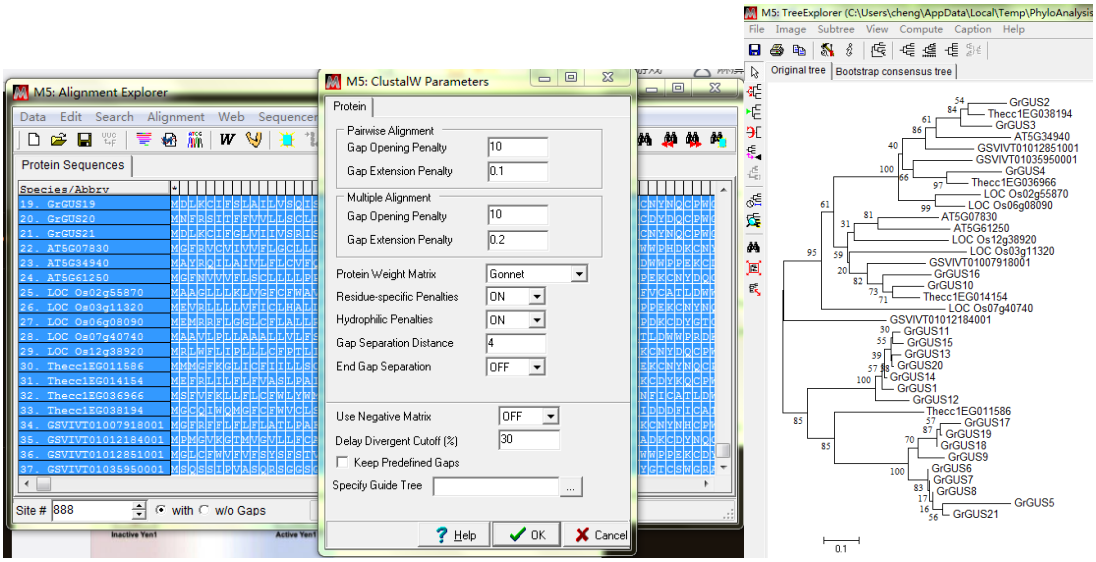
CDD 结果如下图，详细预测结果可选择下载或在线浏览，下载文本格式如网页下部分所示，borwse results 结果：



4 进化树重构

所有验证成功确定的家族基因序列用来重构分子进化树，win 环境下常用 MEGA 来构建进化树，具体步骤包括 fasta 序列导入、多序列对比、设置建树参数、树结果进行操作。

打开 MEGA 软件，选择 Align 菜单 —> Edit/Build alignment —> Create new alignment —> protein/DNA —> 导入序列（可直接复制粘贴 fasta 格式序列）



多序列对比完成之后，将多序列对比结果应用到建树中，具体操作是 Data —> phylogenetic analysis，然后返回到 MEGA 软件主界面，phylogeny —> Construct/Test Maximum LiKelihood Tree（速度较慢，NJ tree 运行快） —> Yes —> Compute(Test of Phylogeny 设置为 Bootstrap method， NO. of bootstrap replications 设置为 1000，其他参数默认即可)。建树步骤运行结束，产生下面结果，根据目的对其进行添加背景等操作，左边纵向排列图标表示了不同的操作手段。

5 多序列对比

基因多序列一般对比采用 ClustalX2 完成，打开软件导入 fasta 格式文本，Alignment —> Output Format Options —> 勾选 GCG/MSF format —> OK，Alignment —> Do Complete Alignment —> OK。运行结束，采用 GENEDOC 软件查看多序列对比结果（alignment.msf）。

6 基因结构图示

基因结果显示通过在线服务 Gene Structure Display Server 2.0（http://gsds.cbi.pku.edu.cn/）完成，主要有 GFF 注释文件、CDS 和基因组序列、BED 注释文件、NCBI 登录号四种方式，其中通过 GFF 注释文件、CDS 和基因组序列最常用。

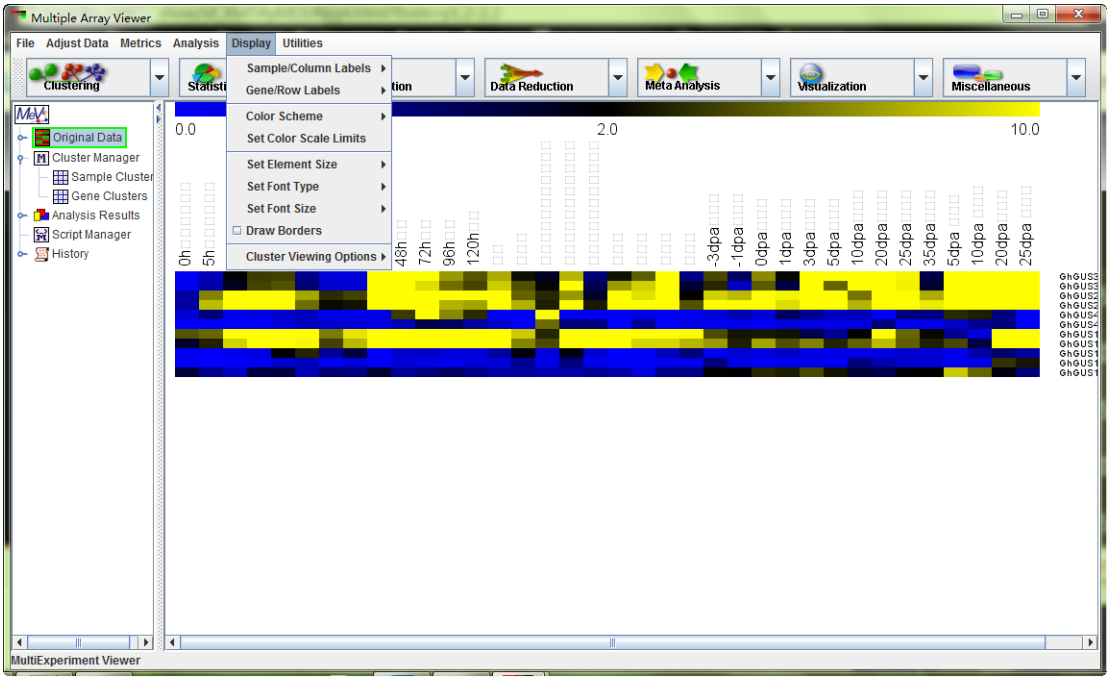
由 CDS 和基因组序列对比产生基因结构图存在错误对比的情况，推荐使用 GFF3 注释文件生成基因结构图。在基因组 GFF3 注释文件中筛选需要的家族基因注释信息，选择文件格式为 GTF/GFF3，复制粘贴到网页对应位置，提交即可，注意事项：网站会根据基因 ID 对基因结构出现先后顺序进行排序。

7 表达热图

基因表达热图有 java 工具 MeV 完成，基因表达数据整理到文本文件中，格式如下：

	0h种子萌	5h种子萌	10h种子萌	24h子叶	48h子叶	72h子叶	96h子叶	120h叶	24h根	48h根	72h根
GhGUS3-At	0.091268	0.526867	1.96869	3.93317	4.05845	0.915336	0.282448	0.301953	12.0584	11.0641	10.82385
GhGUS3-Dt	0.11291	0.680022	1.754875	4.15413	4.58789	1.11977	0.393489	0.325027	10.0571	10.60055	9.00763
GhGUS2-At	0.743472	6.20297	11.36585	16.4873	24.1216	7.25188	3.40997	4.14927	55.5147	39.45585	20.11915
GhGUS2-Dt	0.743472	7.99517	11.1534	16.7384	15.4984	5.51073	2.34082	2.63104	33.044	19.3084	12.36485
GhGUS4-At	0.363275	0	0.109691	0.209414	0.387047	0.486106	0.110431	0.121658	0.178033	3.885	11.29385
GhGUS4-Dt	0.0155838	0.0363296	0.069005	0.0564729	0.292193	0.458727	0.382216	0.31577	0.219765	0.5499595	1.36757
GhGUS16-At	4.64634	5.2906	12.2492	16.2697	18.8005	11.3405	5.89942	9.22946	23.0235	26.2869	21.41255
GhGUS16-Dt	1.66654	3.09737	8.631575	11.4781	14.0793	9.82034	4.11847	7.84303	21.8818	24.84125	24.1199
GhGUS10-Dt	0.0885237	0.0515941	0.1468872	0.21423	2.00542	3.15316	0.807608	1.48106	0.0489654	0.3804175	0.38627
GhGUS10-At	0.0555124	0.0161776	0.0614043	0.100929	0.390246	0.550528	0.10975	0.275909	0.0460679	0.3199115	0.177464
GhGUS1-At	1.53608	0.90705	0.49515	1.4726	1.48949	0.547213	1.10913	1.02212	1.59218	0.7951445	1.0317045

MeV 工具解压即可使用，点击“TMEV.bat”文件运行程序。File —> load data —> browse —> 选中目标文件 —> load



Display 菜单下设置图片显示，颜色(color scheme)，标尺(set color scale limits)，大小(set element size)，字体，字号，边框有无(draw borders)等。绘制热图其他菜单项不使用。

聚类分析，聚类分析菜单位于左上方，聚类算法常用层次聚类和 k 均值聚类两种，聚类参数根据聚类结果适当调整，一般默认参数。