中国农业大学学报 2011,16(4):134-140
Journal of China Agricultural University

苹果 EST 数据分析平台的构建及初步应用

岳 鹏[1,2] 钟 翡[1] 雷恒久[1] 柳爱玲[1] 刘兰英[3] 李天红[1*]

(1.中国农业大学 农学与生物技术学院/北京市果树逆境生理与分子生物学重点开放实验室,北京 100193;
2.张家口广播电视大学,河北 张家口 075000;3.北京市海淀区植物组织培养技术实验室,北京 100091)

摘 要 为更有效地利用苹果 EST 公共数据资源,方便信息交流与共享,本研究构建了苹果 EST 数据分析平台 *Malus domestica* Micro-Workstation,并整合了专为苹果定制的序列延伸系统。该数据分析平台具有本地信息查询、序列简单注释和短序列延伸等功能。以所建平台为辅助工具,从已公布的 324 308 条苹果 EST 序列中预测干旱胁迫相关基因,共有 1 979 条 EST 延伸序列可能与干旱胁迫相关,其中预计下调的数量为 959,上调的为 1 020,共计 109 条干旱胁迫相关序列可能具有全长编码区;从细胞组分、分子功能和生物过程等 3 方面将这些延伸序列归类,为进一步克隆苹果干旱胁迫相关基因提供线索。

关键词 苹果;EST;数据分析平台;序列延伸;干旱胁迫

中图分类号 S 661.1 **文章编号** 1007-4333(2011)04-0134-07 **文献标志码** A

Construction of EST-data analysis platform for *Malus domestica* and its preliminary application

YUE Peng[1,2], ZHONG Fei[1], LEI Heng-jiu[1], LIU Ai-ling[1], LIU Lan-ying[3], LI Tian-hong[1*]

(1. College of Agronomy and Biotechnology/Key Laboratory of Stress Physiology and Molecular Biology for Fruit Trees of Beijing, China Agricultural University, Beijing 100193, China; 2. Zhangjiakou Radio and TV University, Zhangjiakou 075000, China; 3. Laboratory of Plant Tissue Culture Technology, Haidian District, Beijing 100091, China)

Abstract In order to use ESTs data for *Malus domestica* more efficiently and conveniently, an EST-data analysis platform as being named "*Malus domestica* Micro-Workstation" was constructed and a sequence extending system was built on this basis. Upon test with the platform, the drought stress-related genes from 324 308 ESTs of *Malus domestica* were examed for prediction. The results showed that 1 979 sequences of the drought stress-related genes were found, in which 109 sequences were identified with potentially full-length coding regions. A total of 959 drought stress-related sequences may be down-regulated, while the other 1 020 sequences may be up-regulated. Then these drought stress-related sequences were classified into various functional types, according to cellular component, molecular function or biological process. The data analysis platform could be used for apple biotechnology research such as local inquiry, sequence annotation, sequence extension and so on.

Key words *Malus domestica*; EST; data analysis platform; sequence extend; drought stress

苹果(*Malus domestica* Borkh.)属蔷薇科苹果亚科[1],是世界范围内栽培的重要果树之一,分布广泛,品种繁多,具有很高的经济价值[2],已成为研究果树重要商业特性的模式植物[3]。截至 2009 年 8 月,在 GenBank 中已有 324 308 条苹果表达序列标签(Expressed Sequence Tag,EST),其中的大部分来源于新西兰和美国的大规模测序工作[4]。EST 可被用来发现新基因、开发分子标记、辅助注释基因组并识别基因结构[5]、指导单核苷酸多态性的研究及辅助分析蛋白质组[6],已被广泛应用于苹果分子

收稿日期:2010-11-05
基金项目:国家自然科学基金资助项目(30871696);公益性行业(农业)科研专项经费项目(201003021)
第一作者:岳鹏,硕士,E-mail:cengsan@126.com
通讯作者:李天红,教授,博士,主要从事果树生理与分子生物学研究,E-mail:lith@cau.edu.cn

生物学研究之中。

已公布的苹果 EST 来源于上百个 cDNA 文库；这些文库代表多种不同的转录模式，包括不同的器官、果实的不同部分、不同的发育阶段以及不同的生物胁迫和非生物胁迫[7]。所以，有必要将这些 EST 相关资源适当整合，方便信息的查询与挖掘。但 EST 仅是 cDNA 序列的片段[8]，这给从 EST 序列集中识别编码基因带来困难；若能把现有 EST 进行序列延伸并加以功能注释，将有助于提高通过实验克隆目的基因全长 cDNA 序列的效率。因此，构建苹果 EST 数据分析平台并实现前述功能具有重要意义。

当前，国外专门针对苹果 EST 数据资源的分析平台较少，如 Genome Database for Rosaceae[6] 和 Tree Fruit Technology[7] 是 2 个包含多种常见园艺作物信息的数据平台；除此之外，大部分是类似 NCBI(http://www.ncbi.nlm.nih.gov)的大型综合数据库应用系统。上述网站都提供了许多生物信息分析工具，能将大量序列拼接以获得单一序列(Unique Sequence)，但都不能直接延伸单条苹果 EST。而国内运用生物信息学研究苹果的实验室较少，迄今未见有关苹果生物数据分析平台的报道。

为更有效利用苹果 EST 公共数据资源，实现直接对单条 EST 序列的延伸和功能预测，本研究构建了苹果 EST 数据分析平台 *Malus domestica* Micro-Workstation(MdMW)，整合了专为苹果序列定制的延伸系统 *Malus domestica* Extending System (MdES)；以该平台为辅助工具，将已报道的苹果 EST 进行序列延伸；从所得结果序列中预测了可能存在的干旱胁迫相关基因，以期为苹果重要功能基因的深入研究提供科学线索。

1　材料与方法

1.1　运行环境

苹果 EST 数据分析平台采用 Fedora10 操作系统(kernel：Linux 2.6.27.5-117.fc10.i686)，所需软件包括 Apache 服务器(httpd-2.2.10-2.i386)、MySQL 数据库(mysql-5.0.67-2.fc10.i386)、PHP 解释器(php-5.2.6-5.i386)及 Perl 解释器(perl-5.10.0-49.fc10.i386)等。与生物数据分析相关的软件包括序列比对软件 BLAST(blast-2.2.19-ia32-linux)[9]、序列预处理软件 SeqClean[10] 和 RepeatMasker[11]、序列拼接软件 CAP3[12] 和 TGICL[13] 及 BioPerl 模块(BioPerl-1.6.0)[14] 等。

1.2　数据准备

从公共网站 NCBI 下载苹果 EST 序列，并运行 Perl 脚本 est_arrange.pl 加以整理，生成 est_infor_*和 lib_infor_*两个文件。其中，*代表源序列文件名，est_infor_*是 EST 信息表的内容，lib_infor_*是文库信息表的内容。EST 源序列文件还需要用软件 SeqClean 修剪并去掉载体序列，再用软件 RepeatMasker 屏蔽简单重复序列，所得结果文件被重命名为 Md_est。

对序列注释需要从 TAIR(The Arabidopsis Information Resource，http://arabidopsis.org)下载拟南芥生物序列、拟南芥 GOslim 文件 ATH_GO_GOSLIM、GOslim 与 GO 的映射文件 GO_GOSLIM_Map 及 GOslim 归类文件 TAIR_GO_slim_categories 等[15]；还要从 the Gene Ontology (GO)网站下载 GO 的 MySQL 数据库文件 go_200904-termdb-tables.tar.gz[16]。

1.3　数据库设计

本研究选择 MySQL 作为平台的数据库管理系统(DataBase Management System，DBMS)。在设计关系型数据库时，为防止数据过于冗余并使存储利用达到最佳，较好的方法是确保各关系模式至少遵守第三范式(Third Normal Form，3NF)；还要满足完整性约束条件以保证数据的正确性和一致性[17]。此外，必须在 DBMS 允许的范围内定义数据库对象和数据，执行合法操作。

1.4　构建平台

苹果 EST 数据分析平台是一个数据库应用系统(DataBase Application System，DBAS)。本研究以软件开发瀑布模型为基础，引入快速原型模型和增量模型的开发思路，渐进、迭代地开发 DBAS。在一次迭代开发过程中，通过项目规划、需求分析、系统设计和原型构建等基本活动，开发出一个满足部分需求的原型模型；然后从该原型出发，在下一次迭代开发中构造一个功能更为完善的 DBAS 原型；通过多次迭代，逐步扩展各原型系统的功能，形成最终的 DBAS 产品。

1.5　初步应用

应用平台将 324 308 条苹果 EST 序列逐条进行延伸，参数均取默认值；其中，如果 1 条 EST 已经包含在由其他 EST 延伸得到的序列中，则这条 EST 将不作为延伸种子序列。再通过序列注释系统以 Blastx 方式对获得的 EST 延伸序列加以功能注释；

在 *E*-value(Expect value)≤1e-10 时,延伸序列和与之对应的拟南芥基因会收录到文件 match_list 中。

Huang 等运用统计学方法 SAM(Significance Analysis of Microarray)分析拟南芥基因芯片,在 FDR(False Discovery Rate)<0.05 时,找到拟南芥干旱胁迫相关基因;这些基因的名称保存在文件 At_drought 中[16]。脚本 drought_search.pl 从 At_drought 中提取拟南芥基因名称,再到 match_list 中寻找能与这些基因匹配的苹果延伸序列;cds_find.pl 等一系列脚本被用来寻找可能具有全长编码区的苹果干旱胁迫相关序列;go_category.pl 则利用 TAIR 提供的 GO 分类法,分别从细胞组分、分子功能和生物过程等 3 方面将苹果干旱胁迫相关序列归类。

2 结果与分析

2.1 数据库结构

本研究利用 IDEF1X(Integrated computer aided manufacturing DEF inition method 1X)法进行数据建模(图 1),描述了各关系表的名称、属性及它们之间的联系。数据库 my_apple 是平台的中心数据库,共包含 8 个关系表。表 est_infor 存储 EST 信息;lib_infor 存储 EST 所属文库的信息;uniseq_infor 存储延伸序列信息; assemble_infor 存储延伸

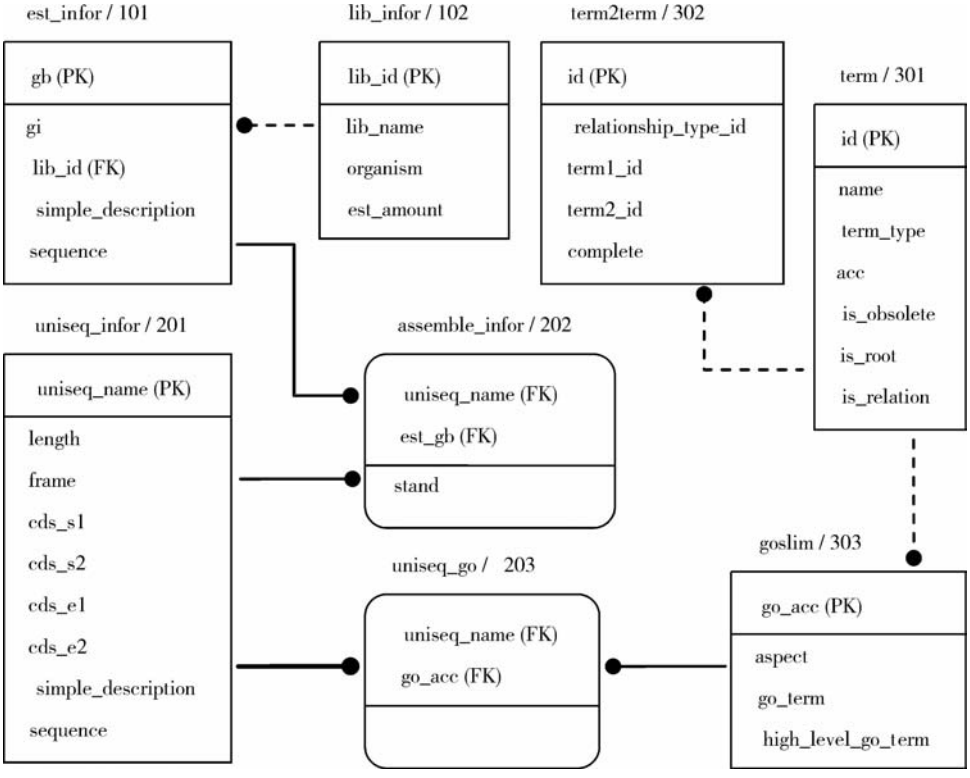

图 1 my_apple 数据库 IDEF1X 建模图

Fig. 1 IDEF1X diagram of database my_apple

序列与 EST 之间的对应关系;uniseq_go 存储延伸序列与 GO 检索号之间的对应关系;term 存储 GO 术语的详细信息;term2term 存储 GO 术语间的相互关系;goslim 存储注释延伸序列的 GO 术语,它来源于拟南芥 GOslim 文件,是对 term 表的简化。另外,在表 est_infor 和 lib_infor 上建立视图 est_list 以方便对 EST 信息的查询;还有建立在各个表上的索引、触发器等,用以提高数据库性能。

2.2 平台功能

图 2 是 DBAS 功能建模的第二层数据流图(Data Flow Diagram,DFD),是对平台结构和功能的概括。MdMW 是关于苹果 EST 的 DBAS,具有本地信息查询(图 2,P1.5)、序列简单注释(图 2,P1.6)和短序列延伸(图 2,P1.7)等功能。

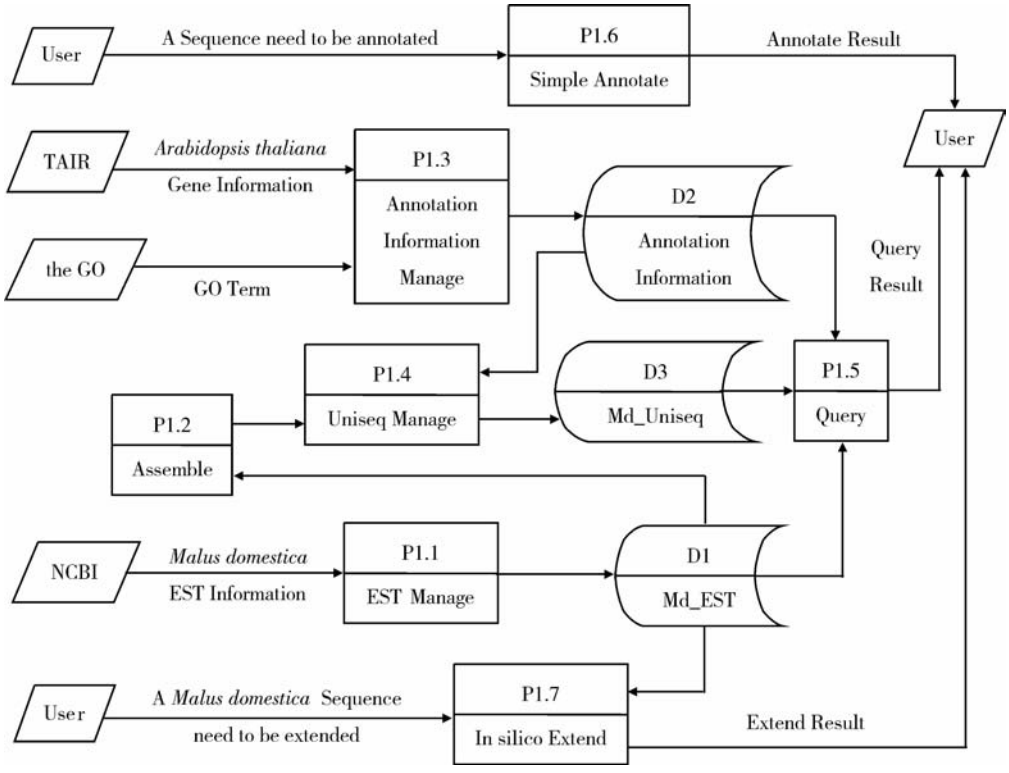

图 2　MdMW 功能建模的第二层数据流图

Fig. 2　Second layer of the data flow diagram of MdMW

2.2.1　本地信息查询

本地信息内容丰富，检索手段多样（图 3）。查询 EST 信息，可使用 GB、GI 或关键词检索。查询文库信息，可使用本地 ID 或关键词检索；使用 ID 查询，不仅可以得到文库的基本信息，还有文库所包含 EST 的列表。查询延伸序列信息，可使用名字或关键词检索；使用名字查询，还能获得延伸序列所含 EST 的列表。还可以使用 GO 检索号或 GO 关键词检索延伸序列信息；使用 GO 检索号查询，能得到全部相关延伸序列的列表。所有关键词检索都支持在结果中再进行与或非的查找，并显示查询次数和命中记录的数量。此外，还有全部 EST、文库和延伸序列的分页列表，使用者可以逐页浏览，选取感兴趣的记录。所有查询结果，都有与其他查询相关的超链接，将所有检索方式连成一个整体。

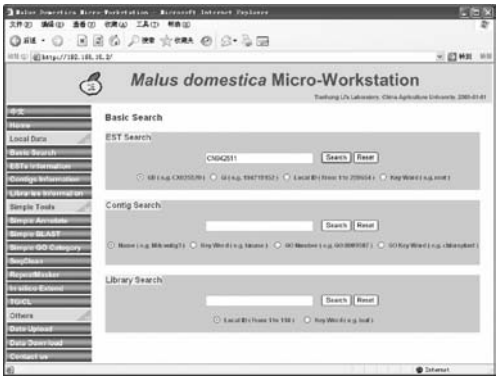

图 3　信息查询界面

Fig. 3　Snapshot of basic search

2.2.2　序列简单注释

对一条核酸序列进行简单注释，必须先以 FASTA 格式输入或上传，使用者可以自行设定 BLAST 参数。平台只提供 Blastn 和 Blastx 两种比对方式，默认为 Blastx；使用 Blastn 对拟南芥 cDNA 序列集检索，使用 Blastx 则对拟南芥蛋白质序列集检索。注释过程具有一定的可控性，系统会把最匹配拟南芥序列的信息赋给查询序列；如果匹配情况不理想，系统会给出警告，但不会中止运行。注释结果包括匹配概要信息和原拟南芥序列的描述信息及 GO 信息。

2.2.3　苹果序列延伸系统

MdES 是一个专为苹果定制的可通过 web 使

用的序列延伸系统,嵌入在平台 MdMW 中,其主要作用是将较短的 EST 序列逐步延伸成较长的 cDNA 序列。使用时,必须先以 FASTA 格式输入或上传一条待延伸序列,或输入某 EST 的 GB,系统会据此从 Md_est 库中提取一条待延伸序列;还可以设定 BLAST 期望值(*E*-value,默认为 1e-10)和最小延伸长度(默认为 10,设为 min_length)。在每次延伸后都显示效果图,增强了延伸结果的直观性,让使用者有机会自主选择新的待延伸序列以进一步延伸(图 4)。

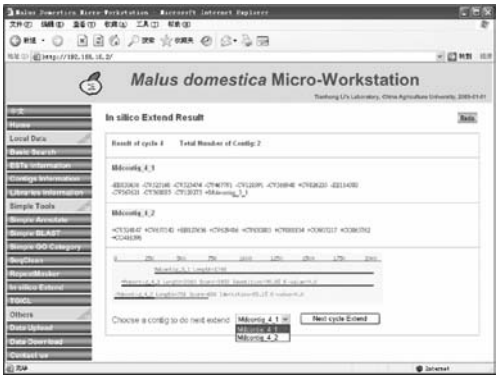

图 4 MdES 界面

Fig. 4 Snapshot of MdES

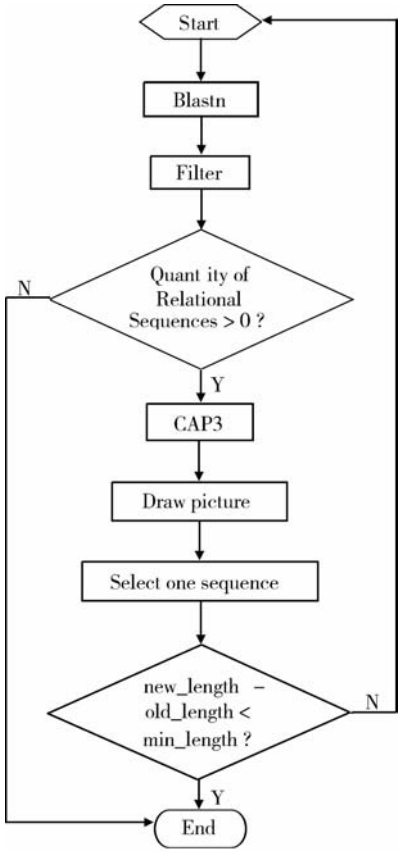

图 5 MdES 工作流程图

Fig. 5 Workflow diagram of MdES

MdES 的工作流程如下(图 5):

1)使用 Blastn 将待延伸序列对 Md_est 库检索。

2)根据比对结果,从 Md_est 库中提取出最多 20 条相关序列,同时滤掉相似性低的和左右两边都不长于待延伸序列的 EST。若相关序列的数量为 0,则延伸结束。

3)使用软件 CAP3 拼接所有相关序列,生成延伸后序列。

4)图形显示待延伸序列和拼接序列的相对位置,选择合适的拼接序列作为新的待延伸序列。

5)判断延伸是否结束。设原待延伸序列的长度为 old_length,新待延伸序列的长度为 new_length,若 new_length - old_length < min_length,则延伸结束;否则,重复上述各步骤。

2.3 初步应用结果

应用本研究平台将 324 308 条苹果 EST 进行序列延伸,获得了 24 392 条延伸序列,其中有 18 383条延伸序列与拟南芥基因相匹配(*E*-value≤1e-10),同时将匹配序列进行了功能注释。文件 match_list 存放延伸序列和与之对应的拟南芥基因;文件 At_drought 存放已公布的拟南芥干旱胁迫相关基因[18]。Perl 脚本 drought_search. pl 从 At_drought 中提取拟南芥基因名称,再到 match_list 中寻找能与这些基因匹配的苹果延伸序列;共找到 1 979 条延伸序列,被认为是苹果中可能存在的干旱胁迫相关基因的 cDNA 序列,其中预计下调的数量为 959,上调的为 1 020。运行 cds_find. pl 等一系列脚本,找到可能包含全长编码区的苹果干旱胁迫相关序列 109 条;其中,为保证数据的可靠性,包含不确定碱基的序列被排除,导致这一数量可能被低估。脚本 go_category. pl 根据 GO 分别从细胞组分、分子功能和生物过程等 3 方面将找到的干旱胁迫相关序列归类(表 1),显示出苹果干旱胁迫相关基因存在于各类别中;进一步的研究将根据可能的全长编码序列设计引物,以克隆感兴趣的全长 cDNA 序列。

表 1　苹果干旱胁迫相关序列分类表*

Table 1　Classification of drought stress-related sequences

细胞组分		分子功能		生物过程	
类别	数量	类别	数量	类别	数量
内质网	68(2)	DNA 或 RNA 连接	85(4)	DNA 或 RNA 代谢	19
高尔基体	25	水解酶	210(12)	细胞组织和生物发生	172(23)
细胞壁	167(13)	激酶	70(2)	发育过程	143(10)
叶绿体	616(25)	核酸连接	38(3)	电子传递或能量转换途径	38(7)
细胞溶胶	243(36)	核苷酸连接	88(6)	其他生物过程	246(11)
细胞外组分	134(8)	其他连接	251(9)	其他细胞过程	984(74)
线粒体	150(12)	其他酶	370(14)	其他代谢过程	976(74)
细胞核	211(18)	其他分子功能	65(1)	蛋白质代谢	456(47)
其他细胞组分	134	蛋白质连接	171(6)	生物或非生物刺激响应	388(17)
其他细胞质组分	803(62)	受体结合	15	胁迫响应	374(16)
其他细胞内组分	831(67)	结构分子	244(35)	信号转导	69(2)
其他膜	522(39)	转录因子	79(2)	转录	63(2)
原生质膜	426(38)	转移酶	174(7)	转运	202(12)
质体	369(16)	转运蛋白	131(7)	未知生物过程	308(6)
核糖体	232(34)	未知分子功能	276(6)		
未知细胞组分	213(17)				
总量	1 817(106)		1 813(101)		1 779(103)

注：* 括号中的数字表示可能具有全长编码区的苹果干旱胁迫相关序列的数量。

3　讨　论

苹果是温带果园的主要落叶果树，在国际鲜果贸易中占很大比重[19]，因其营养丰富且口感好而得到消费者的青睐[20]，国内外对苹果的生物学研究已日渐增多。随着生物技术的高速发展，苹果 EST 数据激增；虽然完全基因组测序发展迅速，但 EST 测序及分析仍然是发现真核生物新基因的主要工具[21-22]，是许多苹果分子生物学研究的切入点。本研究构建的苹果 EST 数据分析平台 MdMW 是在相关研究方向的初步尝试，已集成的和将要收集的都是苹果 EST 相关信息，具有序列延伸和序列注释等功能，为实验室提供了简单易用的辅助分析工具。

作为一种常用分析方法，EST 序列延伸在苹果新基因的发现和基因功能的探索中已有大量应用[3,7,23-24]。这些研究中的序列延伸过程类似，都是先把大量 EST 序列根据同源性聚类，再将聚集到同一类中的 EST 拼接以实现短序列的延伸。本研究中的序列延伸工作流程与前述过程不同，主要表现为：

1)延伸起始只需要 1 条种子序列。这样设计的好处是，当只获得 1 条苹果 cDNA 片段序列时，可直接将它作为种子序列，利用已公布的序列资源通过逐步延伸获得全长序列。

2)没有聚类过程，只提取最多 20 条与种子序列相似性高的 EST，同时滤掉对延伸没有价值的短序列。这样既保证了延伸结果的有效性，又减少了对系统资源的占用，避免长时间得不到系统响应。

3)图形显示待延伸序列和拼接序列的相对位置，允许使用者选择新的待延伸序列。此设计将延伸效果可视化处理，突出了 web 界面的优势，提高了延伸过程的可控性。

在对苹果干旱胁迫相关基因的预测中，上述特点都得到了很好的体现。但是，实践表明，将大量序

列逐条延伸时，MdES需要相当多的人工操作，应给予适当改进。另外，全长编码区的识别和GO功能分类在实践中发挥了很大的作用，应当把Perl脚本cds_find.pl和go_category.pl等也整合到MdMW平台中，这将在下一次DBAS迭代开发中重点完成。未来，平台需要在算法、结构、功能、安全性和稳定性等方面继续完善和优化，使之更好地服务于苹果分子生物学研究。

参 考 文 献

[1] 周云龙. 植物生物学[M]. 北京：高等教育出版社，1999：432-434

[2] 束怀瑞. 苹果学[M]. 北京：中国农业出版社，1999：1-28

[3] Newcomb R D，Crowhurst R N，Gleave A P，et al. Analyses of expressed sequence tags from apple[J]. Plant Physiol，2006，141：147-166

[4] Riccardo V，Andrey Z，Jason A，et al. The genome of the domesticated apple (*Malus* × *domestica* Borkh.)[J]. Nat Genet，2010，42：833-839

[5] Gasic K，Gonzalez D O，Thimmapuram J，et al. Comparative analysis and functional annotation of a large expressed sequence tag collection of apple[J]. Plant Genome，2009，2(1)：23-38

[6] Shulaev V，Korban S S，Sosinski B，et al. Multiple models for Rosaceae genomics[J]. Plant Physiol，2008，147：985-1003

[7] Park S，Sugimoto N，Larson M D，et al. Identification of genes with potential roles in apple fruit development and biochemistry through large-scale statistical analysis of expressed sequence tags[J]. Plant Physiol，2006，141：811-824

[8] Adams M D，Kelley J M，Gocayne J D，et al. Complementary DNA sequencing：expressed sequence tags and human genome project[J]. Science，1991，252(5013)：1651-1656

[9] Altschul S F，Madden T L，Schäffer A A，et al. Gapped BLAST and PSI-BLAST：a new generation of protein database search programs[J]. Nucl Acids Res，1997，25(17)：3389-3402

[10] Computational Biology and Functional Genomics Laboratory. DFCI Gene Indices Software Tools [EB/OL]. http：//compbio.dfci.harvard.edu/tgi/software/

[11] Smit A F A，Hubley R，Green P. Repeat Masker[EB/OL]. http：//repeatmasker.org

[12] Huang X，Madan A. CAP3：A DNA sequence assembly program[J]. Genome Res，1999，9：868-877

[13] Pertea G，Huang X，Liang F，et al. TIGR gene indices clustering tools (TGICL)：a software system for fast clustering of large EST datasets[J]. Bioinformatics，2001，19(5)：651-652

[14] Stajich J E，Block D，Boulez K，et al. The bioperl toolkit：Perl modules for the life sciences[J]. Genome Res，2002，12：1611-1618

[15] Berardini T Z，Mundodi S，Reiser L，et al. Functional annotation of the *Arabidopsis* genome using controlled vocabularies[J]. Plant Physiol，2004，135：745-755

[16] Ashburner M，Ball C A，Blake J A，et al. Gene Ontology：tool for the unification of biology[J]. Nat Genet，2000，25：25-29

[17] 王珊，萨师煊. 数据库系统概论[M]. 4版. 北京：高等教育出版社，2006：151-164，198-233

[18] Huang D，Wu W，Abrams S R，et al. The relationship of drought-related gene expression in *Arabidopsis thaliana* to hormonal and environmental factors[J]. J Exp Bot，2008，59(11)：2991-3007

[19] Zohary D，Hopf M. Domestication of Plants in the Old World[M]. 3 ed. New York：Oxford University Press，2000

[20] Harker F R，Gunson F A，Jaeger S R. The case for fruit quality：an interpretive review of consumer attitudes，and preference for apples[J]. Postharvest Biol Tec，2003，28：333-347

[21] Lee Y，Tsai J，Sunkara S，et al. The TIGR Gene Indices：clustering and assembling EST and known genes and integration with eukaryotic genomes[J]. Nucl Acids Res，2005，33：D71-D74

[22] Sanzol J. Dating and functional characterization of duplicated genes in the apple (*Malus domestica* Borkh.) by analyzing EST data[J]. BMC Plant Biology，2010，10：87

[23] Wisniewski M，Bassett C，Norelli J，et al. Expressed sequence tag analysis of the response of apple (*Malus* × *domestica* 'Royal Gala') to low temperature and water deficit [J]. Physiol Plant，2008，133：298-317

[24] Seo Y S，Kim W T. A Genomics approach using expressed sequence tags and microarrays in ripening apple fruit (*Malus domestica* Borkh.)[J]. J Plant Biol，2009，52：35-40

(责任编辑：袁文业)