生物技术通报
·特约综述· BIOTECHNOLOGY BULLETIN 2015, 31(11):68-77

植物转录因子分类、预测和数据库构建

靳进朴[1] 郭安源[1,2] 何坤[1,3] 张禾[1,4] 朱其慧[1,5] 陈新[1] 高歌[1] 罗静初[1]
(1. 北京大学生命科学学院 北京大学蛋白质与植物基因研究重点实验室 北京大学生物信息中心，北京 100871；2. 华中科技大学生命科学与技术学院，武汉 430074；3. 孟山都公司，美国；4. 密歇根大学，美国；5. 杰克森基因组医学实验室，美国)

摘 要： 转录因子在植物生长发育和应对胁迫等过程中具有重要调控作用，基因组水平上系统预测植物转录因子是研究其功能和演化的基础。通过深入全面的文献调研，总结了一套完整的植物转录因子家族分类规则，开发了转录因子预测流程，构建了转录因子预测平台。基于该流程，从83种绿色植物中预测到129 288个转录因子，分属58个家族。对预测到的转录因子，从家族和个体两个层次进行了详尽注释，构建了植物转录因子数据库PlantTFDB（http://planttfdb.cbi.pku.edu.cn）。PlantTFDB提供了覆盖绿色植物主要谱系的转录因子，其中大部分为具有重要经济价值的单子叶和双子叶植物，已成为植物转录因子功能和演化研究的重要信息资源和分析平台。

关键词： 转录因子；转录因子家族分类；转录因子预测；植物转录因子数据库
DOI：10.13560/j.cnki.biotech.bull.1985.2015.11.004

Classification, Prediction and Database Construction of Plant Transcription Factors

Jin Jinpu[1] Guo Anyuan[1,2] He Kun[1,3] Zhang He[1,4] Zhu Qihui[1,5] Chen Xin[1] Gao Ge[1] Luo Jingchu[1]
(1. *College of Life Sciences, the State Key Laboratory of Protein and Plant Gene Research, Center for Bioinformatics, Peking University, Beijing* 100871; 2. *College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan* 430074; 3. *Monsanto Company, USA*; 4. *University of Michigan, USA*; 5. *The Jackson Laboratory for Genomic Medicine, USA*)

Abstract: Transcription factors (TFs) play key regulatory roles in both plant development and stress response. Genome-wide prediction of TFs is essential for functional and evolutionary studies of plant TFs. We made extensive literature review with thousands papers related to plant TFs and summarized a set of classification rules for plant TFs and developed a TF prediction pipeline. Using this pipeline, we predicted 129 288 TFs, classified into 58 families, from 83 plant species covering the main lineages of green plants including many economically important monocot and dicot crops. We made high-quality annotations for these TFs and built a plant TF database PlantTFDB (http://planttfdb.cbi.pku.edu.cn). A TF prediction server was also developed for users to predict TFs from their own sequences. PlantTFDB has been served as an important portal for the functional and evolutionary studies of plant TF research community.

Key words: transcription factor; classification of transcription factor; prediction of transcription factor; plant transcription factor database

基因表达调控在动植物生长发育过程中具有重要作用，是植物适应外界环境的分子基础，转录调控是基因表达调控的关键步骤。转录调控通过转录因子（Transcription factor）蛋白质序列中的DNA结合结构域和靶基因上游启动子区域特异DNA序列模体结合而实现。除DNA结合结构域（DNA binding domain，DBD）外，转录因子通常还包含转录调控结构域（Transcription regulation domain），主要用于

收稿日期：2015-10-31
基金项目：国家自然科学基金项目（31071160，31171242，31470330），博士后科学基金项目（2014M560017，2015T80015）
作者简介：靳进朴，男，博士后，研究方向：生物信息学；E-mail：jinjp@mail.cbi.pku.edu.cn
通讯作者：罗静初，男，教授，研究方向：生物信息学；E-mail：luojc@pku.edu.cn
高歌，男，教授，研究方向：生物信息学；E-mail：gaog@mail.cbi.pku.edu.cn

调控靶基因转录活性，既可激活转录，也可抑制转录。转录因子中的核定位信号（Nuclear localization signal，NLS）可引导转录因子在胞浆内合成后通过核膜进入细胞核。此外，有些转录因子含寡聚化结构域可形成二聚体或多聚体复合物，具有更为复杂的调控机制。

转录因子种类繁多、功能复杂，它们通过与靶基因启动子结合，激活或抑制其转录活性，调控靶基因在不同组织、不同细胞、不同环境条件下特异表达，并通过转录因子级联调控网络，对许多生命过程进行调控。例如，果蝇体节发育由一类称为同源异型框（Homeobox）的基因调控，它们所编码的蛋白质为转录因子，含长度为 60 个氨基酸的 DNA 结合结构域。植物特异转录因子家族 SQUAMOSA promotor binding protein（SBP）成员具有调控玉米果实发育和水稻分蘖等多种功能。

20 世纪 90 年代开始的人类基因组计划，开创了生命科学研究的新时代。人类基因组计划指定的模式生物酿酒酵母、秀丽线虫和果蝇的基因组测序于 2000 年前先后完成。拟南芥基因组测序于 2000 年底完成。2000 年 12 月 15 日，就在 *Nature* 杂志发表拟南芥基因组序列分析论文[1]的第 2 天，*Science* 杂志发表了题为《拟南芥转录因子：从基因组水平上比较真核生物转录因子》的论文[2]，首次系统预测了拟南芥基因组中的 1 533 个转录因子，将它们分为 28 个家族，并与酵母、线虫和果蝇等其它 3 个真核生物进行了系统比较，发现拟南芥中转录因子在整个基因组中所占比例远高于上述 3 个物种。

2004 年，北京大学生命科学学院朱玉贤、邓兴旺主持的国家自然科学基金国际合作项目，对拟南芥中预测到的转录因子按家族逐个克隆，并对结果进行了初步分析[3]。为配合该课题的顺利进行，我们构建了拟南芥转录因子数据库[4]（Database of *Arabidopsis* transcription factors，DATF）。DATF 中预测到的转录因子数共 1 922 个，分为 64 个家族。此后不久，水稻和杨树基因组序列发布，我们又先后构建了水稻转录因子数据库[5]（Database of rice transcription factors，DRTF）和杨树转录因子数据库[6]（Database of poplar transcription factors，DPTF）。与此同时，苔藓类植物小立碗藓（*Physcomitrella patens*）和绿藻类植物莱茵衣藻（*Chlamydomonas reinhardtii*）基因组测序也先后完成，我们又构建了植物主要谱系中这两个代表性物种的转录因子数据库。

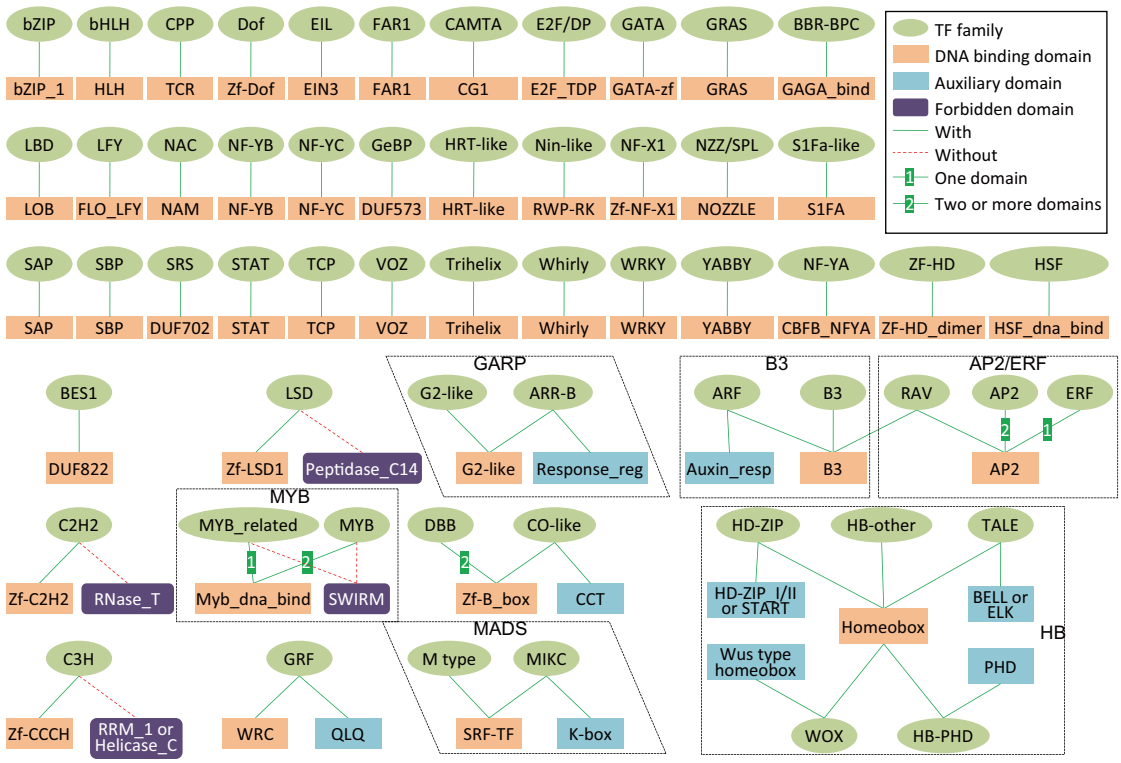
截止 2007 年，玉米、高粱、棉花、大豆、葡萄等重要经济作物的基因组测序尚未完成，但美国爱荷华州立大学植物基因组数据库 PlantGDB 收录了大量植物代表性转录本（Plant unique transcripts，PUT）序列数据[7]。这些 PUT 序列是由表达序列标签（Expressed sequence tag，EST）拼接而成，有些是全长 mRNA 序列，有些则是 mRNA 序列片段。我们从 17 个物种 PUT 序列中预测了转录因子，并和上述 DATF 等 5 个已完成基因组测序物种的转录因子数据库整合在一起，构建了植物转录因子数据库[8]（Plant transcription factor database，PlantTFDB），为植物基因组学、遗传学和植物分子生物学研究提供宝贵的数据资源。2010 年，玉米、高粱、大豆、葡萄等 18 个被子植物，代表性蕨类植物江南卷柏（*Selaginella moellendorffii*），以及 9 个绿藻基因组测序相继完成。此外，PlantGDB 数据库也进行了更新，并增加了不少新物种。与此同时，许多转录因子家族、特别是植物特异转录因子家族的起源、演化、功能等研究成果相继发表，转录因子家族分类也得以更新。为此，我们对 PlantTFDB 进行了大规模更新，更新后的第 2 版包括从 49 个物种中预测到的 53 315 个转录因子，分为 58 个家族[9]。随着基因组测序技术不断改进，测序速度不断加快。2013 年，已有 67 种植物的基因组测序完成，我们对 PlantTFDB 再次进行更新。更新后的第 3 版共包括 129 288 个转录因子，来自 83 个物种，其中 67 个已完成基因组测序，覆盖绿色植物各大门类[10]。

本文介绍植物转录因子分类规则和预测方法，以及植物转录因子数据库 PlantTFDB 的概况和注释信息。

1　植物转录因子家族分类

转录因子蛋白质序列中的 DNA 结合结构域 DBD 在很大程度上决定其与基因上游启动子区域 DNA 顺式元件结合的序列特异性[11]。DBD 在演化上比较保守，通常用作区分不同转录因子家族的主要依据。2000 年，Riechmann 等[2]归纳整理了拟

南芥中转录因子家族及其特征，将其分为 28 个家族。10 多年来，我们先后检索和阅读了大量植物转录因子相关文献，文章总数累计达 7 000 余篇。在 Riechmann 等工作基础上，根据已有文献报道，总结了植物转录因子家族及其结构域序列特征，改进了植物转录因子家族分类规则，并不断加以修改和完善，用于植物转录因子家族划分和植物基因组中未知转录因子的预测（图 1）。



绿色椭圆表示转录因子家族，红色矩形表示 DNA 结合结构域，蓝色矩形表示辅助结构域，紫色矩形表示禁止结构域。绿色实线连接表示转录因子家族和 DNA 结合结构域、辅助结构域之间的对应关系，红色虚线连接转录因子家族和禁止结构域。数字“1”和“2”表示连接的 DNA 结合结构域数

图 1 植物转录因子家族分类规则

1.1 单一DNA结合结构域

一般说来，根据转录因子蛋白质序列中所含 DNA 结合结构域种类，即可确定其属于某个特定家族。第 3 版 PlantTFDB 数据库 58 个转录因子家族中，36 个家族（~62%）符合这种家族与 DBD 一一对应的简单规则，如调控植物生长发育的乙烯不敏感（Ethylene insensitive-like，EIL）转录因子家族均含 EIN 结构域，调控植物花、果实发育的 SQUAMOSA 基因启动子结合蛋白（SQUAMOSA-promoter binding protein，SBP）均含 SBP 结构域。

1.2 禁止结构域

除上述具有简单对应关系的转录因子外，某些蛋白质家族情况比较复杂。例如，由两个半胱氨酸（Cys，C）和两个组氨酸（His，H）组成的 C2H2 锌指结构，是重要的蛋白质序列模体。这类蛋白质分子中，有些能与 DNA 结合，具有转录活性；有些则与 RNA 结合，具有核酸酶活性，除了能与 RNA 结合的 C2H2 锌指结构外，它们同时包含核酸酶相关 RNase_T 结构域。因此，我们将 RNase_T 结构域称为“禁止结构域”（Forbidden domain），用来降低转录因子预测中含 C2H2 锌指结构的蛋白质预测的假阳性率。又如，半胱氨酸型肽段内切酶 MCP1B 和 AtMC2 均具有 DNA 结合结构域 Zf-LSD，但目前尚无证据表明它们具备转录调控功能。我们用禁止结

构域“Peptidase_C14”用来滤除包含 Zf-LSD 结构域蛋白质中的非转录因子。除上述两个家族外，C3H 和 MYB 家族也含禁止结构域。

1.3　辅助结构域

有些转录因子中除了 DBD 外，还有其它一些特征结构域，称为“辅助结构域”（Auxiliary domain）。辅助结构域也可用作转录因子家族分类的依据。例如，生长调控因子（Growth regulation factor，GRF）转录因子家族中均含 WRC 结构域，该结构域中的特征序列为色氨酸（Trp，W）- 精氨酸（Arg，R）- 半胱氨酸（Cys，C）序列模体 WRC。但并非所有含 WRC 序列模体的蛋白质都具有转录活性，只有既有 WRC 序列模体又有 QLQ 序列模体［谷氨酰胺（Gln，Q）- 亮氨酸（Leu，L）- 谷氨酰胺（Gln，Q）］的蛋白质才是转录因子。

1.4　DBD结构域数

有些转录因子中含两个或两个以上 DBD，因此，DBD 数目也常常用来区分不同转录因子家族。典型实例为 AP2 和 ERF 家族。这两个家族转录因子中均含 AP2 结构域，同属于 AP2/ERF 超家族，其中仅含一个 AP2 结构域的为 ERF 家族，含两个或两个以上的则为 AP2 家族。又如，MYB 转录因子超家族均含 Myb_dna_bind 结构域，仅含一个的为 MYB_related 家族，而含两个或两个以上的为 MYB 家族。

1.5　超家族

除上述基于 DNA 结合结构域、利用禁止结构域和辅助结构域对不同转录因子家族进行分类外，有些转录因子家族之间的关系比较复杂。例如，具有 DNA 结合结构域 G2-like 的转录因子均属于 GARP 超家族，其中同时还含 Response_reg 结构域，而有的则仅有 G2-like 结构域。我们将仅含 G2-like 结构域的转录因子归为 G2-like 家族，而把兼有 G2-like 和 Response_reg 结构域的转录因子归为 ARR-B 家族。

更为复杂的是，AP2/ERF 超家族中的另外一个家族 RAV 同时含有两个 DNA 结合结构域，一个为 AP2，另一个为 B3。而 B3 结构域又是另外一个超家族 B3 中两个家族的 DNA 结合结构域。该超家族中仅含 B3 结构域的为 B3 家族，同时含 B3 结构域和 Auxin_resp 辅助结构域的为 ARF 家族。

具有同源异型结构域（Homeodomain）的转录因子是一个具有多个家族的超家族，根据是否具有辅助结构域及辅助结构域类别，可细分为 HD-ZIP、TALE、WOX 等家族。

2　植物转录因子预测

2.1　预测方法

利用上述家族分类规则，可以将文献中已经报道的植物转录因子分为若干家族，并以此为依据预测已经完成基因组测序的绿色植物基因组中未知转录因子。早期的预测主要采用 BLAST 序列相似性搜索，即以不同家族的已知转录因子 DBD 序列为检测序列，设置恰当的参数，用安装到本地的 BLAST 软件包，逐个搜索不同物种基因组中蛋白质编码序列，并对搜索结果进行计算机和人工筛选，剔除假阳性结果。

基于隐马氏模型（Hidden markov model，HMM）的序列分析软件包 HMMER 在蛋白结构域识别方面具有灵敏度高、特异性好的优势，多用于预测同一家族的远缘同源序列[12]。其主要原理为适当选取若干已知种子序列并进行多序列比对，基于隐马氏模型对序列比对结果进行分析并构建隐马氏模型，给出模型参数。因此，我们采用 HMMER 软件包为主要转录因子预测工具。欧洲生物信息学研究所（European bioinformatics institute，EBI）Bateman 领导的研究组，利用 HMMER 软件包构建了蛋白质结构域数据库 Pfam[13]。该数据库还无偿提供他们构建的用于预测蛋白质结构域的隐马氏模型。上述转录因子分类规则中共用到 63 个隐马氏模型，其中 52 个取自 Pfam 数据库，另外 11 个当时发布的第 27 版（Pfam V27.0）尚未公布。为此，基于文献和收集到的转录因子序列，利用 HMMER 软件包，我们构建了这 11 个结构域的隐马氏模型，用于预测植物基因组中的转录因子（表 1）。为提高预测的准确性，我们基于 GO 注释[14]、拟南芥信息资源数据库[15]（The Arabidopsis information resource，TAIR）和国际蛋白质序列和功能知识库 UniProtKB[16]等相关信息，人工检查序列比对结果，并参考 Pfam 确定阈值的方法，为每个结构域模型确定了一个阈值。

基于上述方法和隐马氏模型，我们构建了植物

表 1 用于转录因子预测的隐马氏模型

结构域类型	Pfam	自建	总数
DNA 结合结构域	39	8	47
辅助结构域	8	3	11
禁止结构域	5	0	5
总数	52	11	63

转录因子预测流程，用于预测植物基因组中未知转录因子[17]。

2.2 预测平台

上述用于转录因子预测的隐马氏模型可免费提供国内外用户，便于用户自行构建本地转录因子预测系统，从基因组水平系统预测新测定的基因组中未知转录因子。为方便不具备自行构建本地转录因子预测系统的广大用户，我们在 PlantTFDB 数据库网站中构建了在线转录因子预测平台，用户可以上载序列，预测未知蛋白序列中的转录因子。目前，模式植物拟南芥的转录因子调控机制研究最为清楚，在 PlantTFDB 中注释信息也最为详尽。用户若在提交页面勾选“Best hit in *Arabidopsis thaliana*”，预测结果中则包括相似拟南芥转录因子的超链接，供用户参考。

3 植物转录因子数据库构建

3.1 数据库概况

2013 年更新的第 3 版植物转录因子数据库 PlantTFDB 收录了从 83 个物种预测到的 129 288 个转录因子，分属 58 个家族（表 2）。这 83 个物种覆盖了绿色植物各大谱系，包括 10 个绿藻、1 个苔藓植物、1 个蕨类植物、4 个裸子植物、1 个被子植物基部类群、17 个单子叶植物和 49 个双子叶植物。裸子植物中欧洲云杉（*Picea abies*）的基因组测序已经完成，填补了旧版 PlantTFDB 中没有裸子植物全基因组预测所得转录因子的空白。显然，这 83 个物种中，被子植物占绝大多数（~81%），包括单子叶植物水稻、玉米、高粱、小麦、大麦等主要粮食作物，双子叶植物中棉花、烟草、大豆、番茄、马铃薯、黄瓜、西瓜等重要经济作物，以及葡萄、苹果、梨、橙、橘等水果，为作物分子育种研究提供了宝贵资源。而与模式植物拟南芥同一属的琴叶拟南芥（*Arabidopsis lyrata*）、同为十字花科的小盐芥（*Thellungiella halophila*）和条叶蓝芥（*Thellungiella parvula*）的转录因子数据，则为转录因子家族的起源、演化和功能研究提供了基础。

植物从水生到陆生的演变是生命演化史上的重要事件。横跨绿色植物各大分支的转录因子全谱的发布，使我们可以从转录调控水平研究这一重要历史进程。与绿藻相比，陆生植物无论在转录因子家族数目、转录因子数目及转录因子在基因组中所占比例等方面都明显高于绿藻，与陆生植物更加复杂的多细胞形态发育相关[18]。

3.2 数据库注释

高质量的注释信息是植物转录因子数据库 PlantTFDB 的重要特色。通过查看注释信息，从事植物转录调控研究的生物学工作者可获取该转录因子序列、功能、表达、调控等相关信息，并通过文献信息了解其研究现状。PlantTFDB 中的注释信息可以分为两个层次，第一个层次为单个转录因子的注释，第二个层次为家族水平的注释。

单个转录因子的注释，除名称、序列、结构域等基本信息外，也包括与其它重要数据库的链接。此外，我们从 TAIR、UniProtKB 和 AthMap[19] 等公共数据库中全面收集专家校验的功能描述、结合位点 / 矩阵、microRNA 调控、激素调控、相互作用、突变和表型等信息。同时，还通过整合 Entrez Gene[20]、GeneRIF[20] 以及通过文本挖掘和人工校验获得的文献信息[18]，为收录的转录因子提供了相关的参考文献列表。此外，我们还收录了分别基于 9 个十字花科物种的基因组比对和 20 个被子植物基因组比对所得到的转录因子结合位点保守元件序列[21, 22]（表 3）。

家族水平的注释除了该家族简介和相关文献信息外，还包括该家族成员的演化信息，包括所有物种每个家族成员和每个物种内每个家族成员两类比对信息，以序列图标（Sequence logo）（图 2-A）和系统发生树方式（图 2-B）展示。

4 结论与展望

自 2005 年首次发表拟南芥转录因子数据库 DATF[4] 至今已有 10 年，10 年来，我们不断扩充和

表 2　植物转录因子数据库 PlantTFDB 中 83 个物种转录因子及其家族统计

谱系	拉丁文学名	中文名	英文名	基因	转录因子	百分比 /%	家族
绿藻	*Bathycoccus prasinos*	超微浮游藻		7919	139	1.76	26
	Chlamydomonas reinhardtii	莱茵衣藻		19526	230	1.18	29
	Chlorella sp. NC64A	小球藻		9791	163	1.66	28
	Coccomyxa sp. C-169	胶球藻		9629	138	1.43	27
	Micromonas pusilla CCMP1545	细小微胞藻		10658	150	1.41	32
	Micromonas sp. RCC299	细小微胞藻		9891	153	1.55	31
	Ostreococcus lucimarinus CCE9901	微小海洋绿藻		7645	111	1.45	30
	Ostreococcus sp. RCC809	微小海洋绿藻		7492	102	1.36	29
	Ostreococcus tauri	微小海洋绿藻		7664	99	1.29	26
	Volvox carteri	团藻		15285	125	0.82	27
苔藓	*Physcomitrella patens*	小立碗藓	Moss	32273	1079	3.34	53
蕨类	*Selaginella moellendorffii*	江南卷柏	Spikemoss	22271	665	2.99	54
裸子植物	*Picea abies*	欧洲云杉	Norway spruce	71158	1851	2.60	55
	*Picea glauca**	白云杉	White spruce	16496	559	3.39	49
	*Picea sitchensis**	北美云杉	Sitka spruce	11351	362	3.19	48
	*Pinus taeda**	火炬松	Loblolly pine	13188	442	3.35	47
被子植物基部类群	*Amborella trichopoda*	无油樟		26846	900	3.35	58
单子叶植物	*Aegilops tauschii*	节节麦	Tausch's goatgrass	33849	1439	4.25	55
	Brachypodium distachyon	二穗短柄草	Purple false brome	26552	1557	5.86	56
	Hordeum vulgare	大麦	Barley	24211	1198	4.95	56
	Musa acuminate	小果野蕉	Dwarf banana	36519	2896	7.93	57
	Oryza barthii	短舌野生稻	African wild rice	31675	1507	4.76	56
	Oryza brachyantha	短花药野生稻	Malo sina	32037	1444	4.51	56
	Oryza glaberrima	非洲栽培稻	African rice	33164	1579	4.76	56
	Oryza punctate	斑点野生稻		32139	1718	5.35	56
	Oryza sativa subsp. *indica*	籼稻	Indian rice	40745	1891	4.64	56
	Oryza sativa subsp. *japonica*	粳稻	Japanese rice	55803	1859	3.33	56
	Phoenix dactylifera	海枣	Date palm	28882	1426	4.94	56
	Phyllostachys heterocycla	毛竹	Moso bamboo	31987	1768	5.53	54
	*Saccharum officinarum**	甘蔗	Sugarcane	21082	672	3.19	48
	Setaria italic	小米	Foxtail millet	40599	1994	4.91	56
	Sorghum bicolor	高粱	Sorghum	33032	1826	5.53	56
	*Triticum aestivum**	普通小麦	Wheat	56068	1940	3.46	56
	Triticum urartu	乌拉尔图小麦		24169	888	3.67	50
	Zea mays	玉米	Maize	38914	2231	5.73	55
真双子叶植物	*Aquilegia coerulea*	蓝花耧斗菜	Columbine	24823	1158	4.67	58
	Arabidopsis lyrata	琴叶拟南芥	Lyrate rockcress	32670	1759	5.38	58
	Arabidopsis thaliana	拟南芥	Thale cress	27416	1716	6.26	58
	*Arachis hypogaea**	花生	Peanut	18677	799	4.28	52
	*Artemisia annua**	黄花蒿	Sweet wormwood	15732	625	3.97	49
	Azadirachta indica	印楝	Neem	40482	1900	4.69	58
	*Brassica napus**	欧洲油菜	Rape	30365	1343	4.42	53
	*Brassica oleracea**	甘蓝	Wild cabbage	12061	477	3.95	51
	Brassica rapa	芜青	Field mustard	41019	3026	7.38	57
	Cajanus cajan	木豆	Pigeon pea	40071	1886	4.71	56
	Cannabis sativa	大麻	Hemp	22670	1061	4.68	56

续表

谱系	拉丁文学名	中文名	英文名	基因	转录因子	百分比/%	家族
	Capsella rubella	荠菜	Shepherd's purse	28447	1900	6.68	58
	*Capsicum annuum**	辣椒	Chilli pepper	19674	922	4.69	53
	Carica papaya	木瓜	Papaya	27765	1379	4.97	58
	Cicer arietinum	鹰嘴豆	Chickpea	27809	1897	6.82	56
	Citrullus lanatus	西瓜	Watermelon	23440	1355	5.78	58
	Citrus clementina	克莱门橙	Clementine	33929	1905	5.61	58
	Citrus sinensis	甜橙	Sweet orange	46147	2256	4.89	58
	Cucumis melo	甜瓜	Muskmelon	27427	1322	4.82	58
	Cucumis sativus	黄瓜	Cucumber	21603	1412	6.54	57
	Eucalyptus grandis	巨桉	Rose gum	36376	1729	4.75	56
	Fragaria vesca	野草莓	Wild strawberry	32831	1485	4.52	58
	Glycine max	大豆	Soybean	54175	3714	6.86	57
	*Gossypium hirsutum**	陆地棉	Upland cotton	21087	1151	5.46	50
	Gossypium raimondii	雷蒙德氏棉	Cotton	37505	2634	7.02	58
	*Helianthus annuus**	向日葵	Sunflower	8716	288	3.30	46
	Jatropha curcas	麻疯树	Jatropha curcas	52782	1467	2.78	57
	Linum usitatissimum	亚麻	Flax	43484	2481	5.71	57
	Lotus japonicus	百脉根	Crowtoe	26119	1311	5.02	56
	*Lactuca sativa**	莴苣	Garden lettuce	19676	1036	5.27	55
	Malus domestica	苹果	Apple	63516	3119	4.91	58
	Manihot esculenta	木薯	Cassava	34151	2247	6.58	58
	Medicago truncatula	苜蓿	Barrel medic	50952	1577	3.10	56
	Mimulus guttatus	猴面花	Spotted monkey flower	28282	1733	6.13	57
	Nelumbo nucifera	莲	Sacred lotus	26473	1476	5.58	57
	*Nicotiana tabacum**	烟草	Tobacco	19090	820	4.30	52
	Populus trichocarpa	毛果杨	Western balsam poplar	41335	2455	5.94	58
	Prunus persica	桃	Peach	28701	1529	5.33	58
	Pyrus bretschneideri	白梨	Chinese white pear	42812	2353	5.50	57
	*Raphanus sativus**	萝卜	Radish	17565	803	4.57	49
	Ricinus communis	蓖麻	Castor bean	31221	1299	4.16	57
	Solanum lycopersicum	番茄	Tomato	34727	1845	5.31	58
	Solanum tuberosum	马铃薯	Potato	51472	2406	4.67	56
	Thellungiella halophile	小盐芥	Salt cress	29284	1892	6.46	58
	Thellungiella parvula	条叶蓝芥		27132	1672	6.16	58
	Theobroma cacao	可可	Cocoa	29452	1449	4.92	58
	Utricularia gibba	丝叶狸藻	Humped bladderwort	27465	1651	6.01	55
	*Vigna unguiculata**	豇豆	Cowpea	12202	488	4.00	48
	Vitis vinifera	葡萄	Wine grape	26346	1276	4.84	58

注：带“*”标记的物种构建数据库时其基因组注释尚未发布

多次更新植物转录因子数据库PlantTFDB。在此期间，德国波茨坦大学、丹麦奥胡斯大学、美国俄亥俄州立大学、日本理化学研究所等单位也构建了相应的植物转录因子数据库（表4）。与这些数据库相比，PlantTFDB包括的物种最多、注释信息最丰富、更新最及时。目前，该数据库年访问量逾千万次，已成为植物转录因子功能和演化研究的权威数据库和重要数据资源，我们构建的植物转录因子家族分类规则也被国内外同行用于新测序物种转录因子预测。

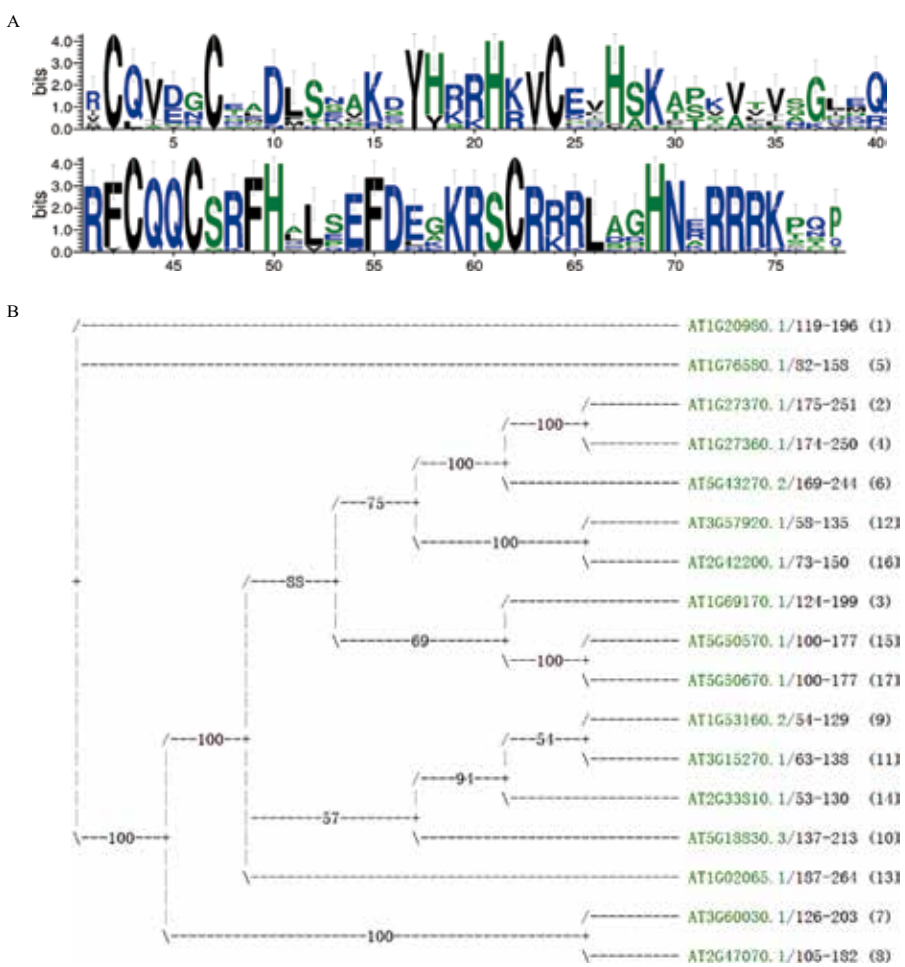
利用上述数据库资源，我们与其他课题组合作，

表 3　转录因子个体水平注释

注释类型	物种数	转录因子数	条目数
专家校验的功能描述	22	2128	6649
表达信息			
UniGene	44	44862	45239
芯片数据	14	15424	31975
Plant Ontology	5	6850	174162
调控信息			
结合位点或矩阵	24	541	729
ChIP-chip/ChIP-seq	1	54	75
microRNA	1	28	43
激素	1	417	803
相互作用信息	10	992	3101
保守元件	2	3709	63859
表型信息	2	4704	147684
参考文献	59	5004	20255

对 AP2/EREBP、MYB、SBP 等植物转录因子家族进行了演化和功能分析[32-34]。同时，对拟南芥转录调控网络进行了深入分析，揭示了植物转录调控网络在结构和演化上的新特征[18]。

不言而喻，随着测序技术的飞速发展，更多植物基因组测序将完成，大量基因组、转录组数据不断发布。随着转录调控研究不断深入，转录因子分类规则有待改进。此外，SELEX 等高通量 DNA 结合特异性测定技术的发展，为深入研究植物转录调控提供了新的契机。结合表达数据、启动子区域和保守元件等信息，预测转录因子下游靶基因，进而构建高质量转录调控网络，探索转录调控的分子机制，必将成为新的研究热点。开发转录调控分析平台，



A：拟南芥 SBP 转录因子 DNA 结合结构域序列图标；B：拟南芥 SBP 转录因子 DNA 结合结构域系统发生树

图 2　转录因子家族水平注释

表 4 国际上主要植物转录因子数据库

数据库名称	物种（物种数）	网址	创建单位（创建 / 更新日期）
PlantTFDB	绿色植物（83）	http://planttfdb.cbi.pku.edu.cn/	北京大学（2007/2013）[8, 9, 10]
PlnTFDB	绿色植物（19）	http://plntfdb.bio.uni-potsdam.de/	德国波茨坦大学（2007/2009）[23]
DATFAP	被子植物（12） 绿藻（1）	http://cgi-www.daimi.au.dk/cgi-chili/datfap/frontdoor.py	丹麦奥胡斯大学（2007/NA）[24]
TreeTFDB	树（6）	http://treetfdb.bmep.riken.jp/	日本理化学研究所（2013/2013）[25]
GRASSIUS	禾本科（5）	http://grassius.org/grasstfdb.html	美国俄亥俄州立大学（2009/2014）[26]
LegumeTFDB	豆科植物（3）	http://legumetfdb.psc.riken.jp/	日本理化学研究所（2009/NA）[27]
DATF	拟南芥	http://datf.cbi.pku.edu.cn/	北京大学（2005/2006）[4]
RARTF	拟南芥	http://rarge.gsc.riken.jp/rartf/	日本理化学研究所（2005/2006）[28]
AGRIS	拟南芥	http://arabidopsis.med.ohio-state.edu/	美国俄亥俄州立大学（2003/2010）[29]
TOBFAC	烟草	http://compsysbio.achs.virginia.edu/tobfac/	美国维吉尼亚大学（2008/NA）[30]
DPTF	杨树	http://dptf.cbi.pku.edu.cn/	北京大学（2007/NA）[5]
DRTF	水稻	http://drtf.cbi.pku.edu.cn/	北京大学（2006/NA）[6]
wDBTF	小麦	http://wwwappli.nantes.inra.fr：8180/wDBFT/	法国克莱蒙费朗第二大学（2009/NA）[31]

将植物转录因子数据库与数据分析整合起来，则是下一步研究目标。

参 考 文 献

[1] Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana* [J]. Nature, 2000, 408：796-815.

[2] Riechmann JL, Heard J, Martin G, et al. *Arabidopsis* transcription factors：genome-wide comparative analysis among eukaryotes [J]. Science, 2000, 290：2105.

[3] Gong W, Shen YP, Ma LG, et al. Genome-wide ORFeome cloning and analysis of *Arabidopsis* transcription factor genes [J]. Plant Physiol, 2004, 135：773-782.

[4] Guo A, He K, Liu D, et al. DATF：a database of *Arabidopsis* transcription factors [J]. Bioinformatics, 2005, 21：2568.

[5] Gao G, Zhong Y, Guo A, et al. DRTF：a database of rice transcription factors [J]. Bioinformatics, 2006, 22：1286.

[6] Zhu QH, Guo AY, Gao G, et al. DPTF：a database of poplar transcription factors [J]. Bioinformatics, 2007, 23：1307.

[7] Duvick J, Fu A, Muppirala U, et al. PlantGDB：a resource for comparative plant genomics [J]. Nucleic Acids Res, 2008, 36：D959-965.

[8] Guo AY, Chen X, Gao G, et al. PlantTFDB：a comprehensive plant transcription factor database [J]. Nucleic Acids Res, 2008, 36：D966-969.

[9] Zhang H, Jin J, Tang L, et al. PlantTFDB 2.0：update and improvement of the comprehensive plant transcription factor database [J]. Nucleic Acids Res, 2011, 39：D1114-1117.

[10] Jin J, Zhang H, Kong L, et al. PlantTFDB 3.0：a portal for the functional and evolutionary study of plant transcription factors [J]. Nucleic Acids Research, 2014, 42：D1182-D1187.

[11] Weirauch MT, Yang A, Albu M, et al. Determination and inference of eukaryotic transcription factor sequence specificity [J]. Cell, 2014, 158：1431-1443.

[12] Eddy S. HMMER User's Guide：Biological sequence analysis using profile hidden Markov models [W]. 2010, http：//hmmer.janelia.org/.

[13] Punta M, Coggill PC, Eberhardt RY, et al. The Pfam protein families database [J]. Nucleic Acids Research, 2012, 40：D290-D301.

[14] Ashburner M, Ball CA, Blake JA, et al. Gene ontology：tool for the unification of biology. The Gene Ontology Consortium [J]. Nat Genet, 2000, 25（1）：25-29.

[15] Lamesch P, Berardini TZ, Li D, et al. The *Arabidopsis* Information Resource（TAIR）：improved gene annotation and new tools [J]. Nucleic Acids Res, 2012, 40：D1202-210.

[16] UniProt Consortium. Activities at the universal protein resource（UniProt）[J]. Nucleic Acids Research, 2014, 42：D191-D198.

[17] He K, Guo AY, Gao G, et al. Computational identification of plant transcription factors and the construction of the PlantTFDB

database [M] //Computational Biology of Transcription Factor Binding. Humana Press, 2010：351-368.

[18] Jin J, He K, Tang X, et al. An *Arabidopsis* transcriptional regulatory map reveals distinct functional and evolutionary features of novel transcription factors [J]. Molecular Biology and Evolution, 2015, 32：1767-1773.

[19] Bulow L, Engelmann S, Schindler M, et al. AthaMap, integrating transcriptional and post-transcriptional data [J]. Nucleic Acids Res, 2009, 37：D983-D986.

[20] Maglott D, Ostell J, Pruitt KD, et al. Entrez Gene：gene-centered information at NCBI [J]. Nucleic Acids Research, 2011, 39：D52-D57.

[21] Haudry A, Platts AE, Vello E, et al. An atlas of over 90, 000 conserved noncoding sequences provides insight into crucifer regulatory regions [J]. Nature Genetics, 2013, 45：891-898.

[22] Baxter L, Jironkin A, Hickman R, et al. Conserved noncoding sequences highlight shared components of regulatory networks in dicotyledonous plants [J]. The Plant Cell Online, 2012, 24：3949-3965.

[23] Pérez-Rodríguez P, Riaño-Pachón DM, Corrêa LGG, et al. PlnTFDB：updated content and new features of the plant transcription factor database [J]. Nucleic Acids Research, 2010, 38：D822-827.

[24] Fredslund J. DATFAP：a database of primers and homology alignments for transcription factors from 13 plant species [J]. BMC Genomics, 2008, 9：140.

[25] Mochida K, Yoshida T, Sakurai T, et al. TreeTFDB：An integrative database of the transcription factors from six economically important tree crops for functional predictions and comparative and functional genomics [J]. DNA Research, 2013, 20：151-162.

[26] Yilmaz A, Nishiyama Jr MY, Fuentes BG, et al. GRASSIUS：a platform for comparative regulatory genomics across the grasses [J]. Plant Physiology, 2009, 149：171.

[27] Mochida K, Yoshida T, Sakurai T, et al. LegumeTFDB：an integrative database of *Glycine max*, *Lotus japonicus* and *Medicago truncatula* transcription factors [J]. Bioinformatics, 2010, 26：290-291.

[28] Iida K, Seki M, Sakurai T, et al. RARTF：database and tools for complete sets of *Arabidopsis* transcription factors [J]. DNA Res, 2005, 12：247-256.

[29] Yilmaz A, Mejia-Guerra MK, Kurz K, et al. AGRIS：the *Arabidopsis* gene regulatory information server, an update [J]. Nucleic Acids Res, 2011, 39：D1118-1122.

[30] Rushton PJ, Bokowiec MT, Laudeman TW, et al. TOBFAC：the database of tobacco transcription factors [J]. BMC Bioinformatics, 2008, 9：53.

[31] Romeuf I, Tessier D, Dardevet M, et al. wDBTF：an integrated database resource for studying wheat transcription factor families [J]. BMC Genomics, 2010, 11：185.

[32] Feng JX, Liu D, Pan Y, et al. An annotation update via cDNA sequence analysis and comprehensive profiling of developmental, hormonal or environmental responsiveness of the *Arabidopsis* AP2/EREBP transcription factor gene family [J]. Plant Mol Biol, 2005, 59：853-68.

[33] Chen YH, Yang XY, He K, et al. The MYB transcription factor superfamily of *Arabidopsis*：expression analysis and phylogenetic comparison with the rice MYB family [J]. Plant Mol Biol, 2006, 60：107-124.

[34] Guo AY, Zhu QH, Gu X, et al. Genome-wide identification and evolutionary analysis of the plant specific SBP-box transcription factor family [J]. Gene, 2008, 418：1-8.

（责任编辑　李楠）