北京大学

博士研究生学位论文

题目：植物转录因子的系统识别和注释及拟南芥转录调控网络分析

姓　　名：靳　进　朴

学　　号：1101110416

院　　系：生命科学学院

专　　业：生物学（生物信息学）

研究方向：生物信息学

导师姓名：罗静初 教授　高歌 研究员

2013 年 11 月

版权声明

摘　　要

经过数亿年的演化，植物形成了一套独特的系统来精确调控生长发育和快速响应多变的环境。作为调控基因转录的开关，转录因子及转录调控网络在植物生长发育及应对各种生物和非生物胁迫中具有关键作用。转录因子及转录调控网络的演化也在很大程度上造就了形态迥异、适应不同生境的植物物种。植物从水生到陆生的演化过程中，产生了很多新的转录因子家族。这些新类型的转录因子和古老类型的转录因子共同组建了一个新的转录调控系统，以适应截然不同的陆地生境和调控复杂的多细胞发育过程。从基因组水平上系统识别植物转录因子和转录调控关系、分析植物发育过程和应激过程中转录调控网络的架构、探索新老类型转录因子如何参与转录调控网络构建等将有助于我们深入理解植物转录调控系统的设计原则和演化特征。

为系统识别植物转录因子，构建了一套数据整合流程和转录因子预测流程。鉴于目前植物基因组注释的不完整性，通过整合基因组注释、RefSeq、PlantGDB 和 UniGene 等数据源，为每个物种构建了一套完整的蛋白组。在本实验室先前工作基础上并通过系统浏览 7000 余篇相关文献，构建了一套完整的植物转录因子分类规则。通过参考 GO 注释、UniProt、TAIR 和 Pfam 中的信息，为每个特征结构域模型确定了合理的阈值，以提高预测准确性。基于上述转录因子预测流程，预测了 83 个物种的转录因子，共计 129288 个转录因子，分为 58 个家族。其中 67 个物种具有基因组序列，覆盖了绿藻、苔藓、蕨类、裸子植物和被子植物等绿色植物各大门类。与绿藻相比，陆生植物的转录因子数、家族数和基因组中转录因子所占比例都有明显提高，推测与陆生植物具有更加复杂的形态相关。基于上述预测结果，对所构建的植物转录因子数据库 PlantTFDB 进行了两次更新。为收录的每个转录因子都做了详尽的注释，包括基本信息、结构域特征、GO 注释、表达信息、相关文献及到各大知名数据库的跨库链接；在 PlantTFDB 3.0 中还增加了专家描述、调控、相互作用、表型等重要信息。通过构建直系同源群和系统发生树，展现了各转录因子之间的演化关系。此外，还搭建了应用程序接口（Web service）和转录因子预测平台，供用户自动获取数据和预测转录因子。该数据库年访问量达千万次，为植物转录因子功能和演化研究提供了宝贵资源。

通过系统的文献发掘和人工校对，我们收集了 1431 个功能确定的转录调控关系并构建了一个高质量的拟南芥转录调控网络（ATRM）。该网络涉及 47 个家族、388 个转录因子，从基因组水平上展示了拟南芥特定生物过程及各过程间相互调控的概况。通过系统识别拟南芥转录调控网络中三节点的结构元件，共找到 5 种结构元件，其中三种是单细胞的大肠杆菌和酿酒酵母中没有的，它们富集于多细胞发育过程中。动力学模拟表明，这些新结构元件能完成状态的维持与转换等多细胞发育所需要的功能。通过比较发育子网络和应激子网络，发现它们在结构元件组成、全局拓扑结构和参与的转录因子性质等方面都存在明显差异。通过比较新老类型转录因子在转录调控系统中的角色，我们发现新类型的转录因子具有更高的调控特异性并倾向于参与发育过程中新类型结构元件和复杂调控网络等的构建。上述新类型转录因子高特异性和它们参与网络构建的倾向性，为转录因子和转录调控的演化以及它们在多细胞演化中的作用提供了新见解。

关键词：植物转录因子、数据库、转录调控网络、网络构建

Systematic identification and annotation of plant transcription factors

and

Analyses of *Arabidopsis* transcriptional regulatory networks

Jinpu Jin (Bioinformatics)

Directed by Professor Jingchu Luo and Professor Ge Gao

By regulating the transcription of their targets temporarily and spatially, transcription factors (TFs) play key roles in regulating development precisely and responding to stress rapidly. Evolution of TFs and their transcriptional regulatory networks contribute to the modification of morphologies and the adaption to local habitats to a large extent. During the evolutionary process of plant landing, many novel TF families emerged and built a new transcriptional regulatory system together with ancient TFs to adapt the dramatic changes in habits and complex multicellular development. Systematic identification of plant TFs and transcriptional regulatory networks would help us to understand their design principles and evolutionary features.

By integrating data from annotated genes of genome projects and other databases such as RefSeq, PlantGDB and UniGene, we built a comprehensive proteome for each species. According to an extensive literature review, we improved the TF family assignment rules used to identify TFs. Based on information from GO annotations, UniProt, TAIR, and Pfam, we defined thresholds for HMM models of signature domains to improve the prediction accuracy. Using the above prediction pipeline, we systematically identified 129,288 TFs from 83 plant species and classified them into 58 families. Compared with green alga, land plants had a large increase in the number of TF families, TFs and the percentage of TFs in their genomes, which might correlate with morphological complexity of land plants. We constructed a plant TF databases PlantTFDB for the identified TFs and made two updates. Furthermore, we made detailed annotations for these TFs, including basic information, domain features, GO annotation, express information, related references and cross-links to well-known databases. In an attempt to construct a knowledgebase for plant TFs, we further collected useful information such as expert-curated description, regulation, interaction,

and phenotype data for identified TFs in the latest version of PlantTFDB 3.0. The orthologous groups and phylogenetic trees for each family showed the evolutionary relationship of identified TFs. We also implemented a Web service for accessing these TF data automatically, and developed a TF prediction server for users to identify TFs from their own sequences.

Through a comprehensive literature mining and manual curation, we collected 1,431 functional confirmed regulatory pairs and constructed an *Arabidopsis* transcriptional regulatory network (ATRM). This map involved in 388 TFs from 47 families and offered a landscape of transcriptional regulation in *Arabidopsis*. By surveying three-node regulatory patterns in ATRM, we identified five types of network motifs in *Arabidopsis*, three of which were novel motifs absent from unicellular organisms *E. coli* and *S. cerevisiae*. These motifs were enriched in multicellular developmental processes, and kinetic simulations suggested their functions in the transition and maintenance of states which were required for multicellular development. Systematic analysis of this map revealed the significant differences between developmental sub-network and stress response sub-network in the composition of network motifs, global topological structure, and the property of TFs constructing them. Moreover, evolutionary novel TFs posed higher binding specificity and preferred being wired into more complex developmental sub-network than those of ancient ones. These results will provide insights into the evolution of TFs and transcriptional regulation, and their roles in evolution of multicellular organisms.

Key words: plant transcription factor, database, transcriptional regulatory network, network construction

目录

第1章　绪论

1.1　植物基因组时代

1.1.1　植物的多样性

植物有广义和狭义两种定义。广义的植物包括红藻和绿色植物等；狭义的植物即绿色植物，也就是通常所说的植物。绿色植物通过光合作用将 CO_2 同化为糖类并释放氧气，为动物和其它异养生物提供了生存的物质基础，在整个生物界发挥着举足轻重的作用。

植物的分布极为广泛，从茫茫大海到巍巍高山，从热带沃壤到极地冻土都有植物的身影。除生境不同外，植物形态亦是多种多样。现存植物大约有 35 万种，可分为绿藻、苔藓植物、蕨类植物和种子植物等四大类（图 1-1）。

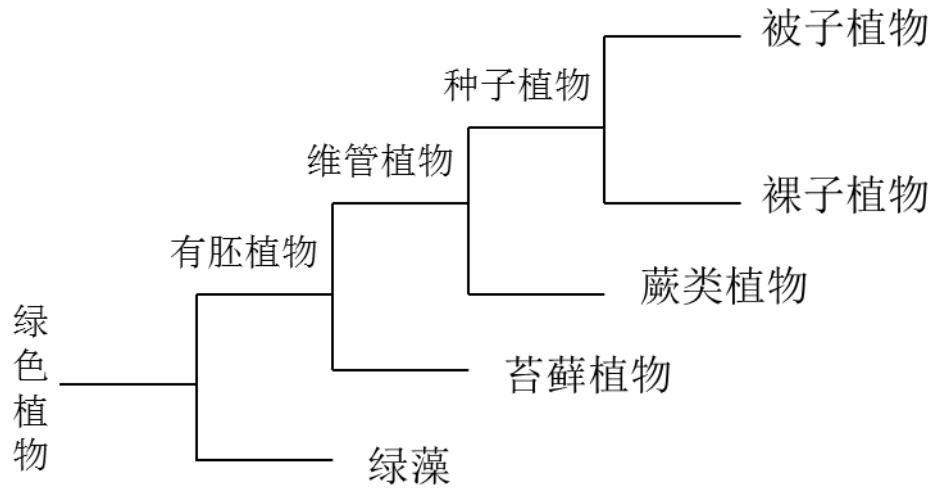

图 1-1　绿色植物主要门类的种系发生

1.1.2　植物登陆——植物演化历程中的重大事件

距今 6 亿年前，植物还限于生活在水中的绿藻。此后，植物开始了从水生到陆生的漫长演化历程。距今约 5.2 年前，陆地出现了两栖的类苔藓植物。经过大约一亿年的演化，出现了真正的陆生植物——维管束植物，使得植物可以树立在陆地上。此后又经过漫长的演化，陆生植物占领了地球各个角落。陆生植物的出现改变了大气组成和环境，为其它陆生生命提供了生存的物质基础和栖息场所，从而造就了丰富多彩的陆生世界。因此，植物登陆是植物乃至生命演化历程中的重大

事件。

由于陆生环境与水生环境存在极大差异，要占领陆地，陆生植物需要适应与水中截然不同的环境，包括温度的巨大波动、干旱以及强烈的辐射等。除了生境上的剧烈改变，陆生植物在形态发育上也发生了巨大的改变，如更加复杂的组织和器官的分化等。

1.1.3 植物基因组数据

染色体是遗传物质最主要的载体，DNA 测序技术的迅猛发展使我们获取植物基因组序列变得相对容易一些。基因组序列所承载的遗传信息则是植物适应不同生境和形态多样化的物质基础。

1999 年，第一个植物基因组——拟南芥基因组正式发布，测出的基因组序列覆盖了全部基因组 1.25 亿个碱基对中的 1.15 亿个。科学家从中识别出 25498 个编码基因，分属 11000 个不同家族。通过与酿酒酵母、黑腹果蝇和秀丽线虫等三个真核生物的比较，拟南芥基因组提供了理解植物与其它生物差异的物质基础(*Arabidopsis* Genome Initiative, 2000)。随后小立碗藓基因组的发布则提供了一个重构植物登陆过程的平台。通过与绿藻基因组的比较，小立碗藓基因组很好地解释了植物登陆过程，包括：1)基因家族复杂性的增加；2)水生环境相关基因的丢失；3)适应陆生环境各种非生物胁迫（如温度的剧烈波动、干旱等）相关基因的出现；4) 生长素和脱落酸信号通路的完善来调控多细胞发育和干旱胁迫(Rensing, *et al.*, 2008)。

从上面两个例子可以看出，基因组序列为理解植物形态和适应性提供了重要线索。随着测序技术快速发展，近年来大量植物基因组得以测定和注释。截止 2013 年 6 月，已经有 67 个植物基因组测序已完成，覆盖了绿色植物的各大门类（表 1-1），为我们系统研究植物形态和适应性的演化提供了重要资源。

表 1-1 基因组测序已完成的植物物种的门类分布

类别	物种数
绿藻 (Green Alga)	10
苔藓 (Moss)	1
蕨类 (Fern)	1
种子植物 (Seed plants)	
裸子植物 (Gymnosperms)	1
被子植物 (Angiosperms)	54
总数	67

1.2 转录因子及其演化

1.2.1 转录因子与转录调控

存贮在基因组中的遗传信息需要传递到 RNA 或蛋白质才能发挥作用。Crick 提出的分子生物学中心法则很好地阐释了遗传信息如何在 DNA、RNA 和蛋白质之间传递(Crick, 1970)，其中转录过程就是遗传信息从 DNA 传递到 RNA 的过程。而转录调控是基因表达调控的核心环节，涉及到染色质的修饰、转录调控蛋白与染色质的结合、转录复合体的组装、转录起始与延伸等众多环节。转录调控在调节形态发育、细胞分化和命运决定、防御和响应外部刺激等方面起着非常重要的作用(Latchman, 2008)。

真核生物中，DNA 以折叠组装的形式存在，阻碍了基本转录复合体对核心启动子的识别。通过染色质修饰使启动子暴露出来便成了转录的前提条件，而转录因子在其中的作用尤为重要。在转录过程中需要的蛋白主要分为以下四类：通用转录因子、转录辅助因子、序列特异性结合 DNA 的转录因子和染色质相关蛋白。通用转录因子包括 RNA 聚合酶(Pol I、 Pol II、 Pol III)、辅助 RNA 聚合酶识别启动子和准确起始转录的蛋白如 TFIIA、TFIIB、TFIID、TFIIE、TFIIF、TFIIH、TATA-box 结合蛋白（TBP），以及一些 TBP 相关因子(TAFs)。转录辅助因子本身并不结合 DNA，它通过与序列特异性结合 DNA 的转录因子、核心转录复合体或者转录桥接复合体相互作用来调控转录。序列特异性结合 DNA 的转录因子即我们通常所说的转录因子，它通过序列特异性的结合 DNA 来激活或抑制相关基因的转录。染色质相关蛋白包括共价修饰组蛋白和改变染色质结构的复合体(Latchman, 2008, Riechmann, 2002)。

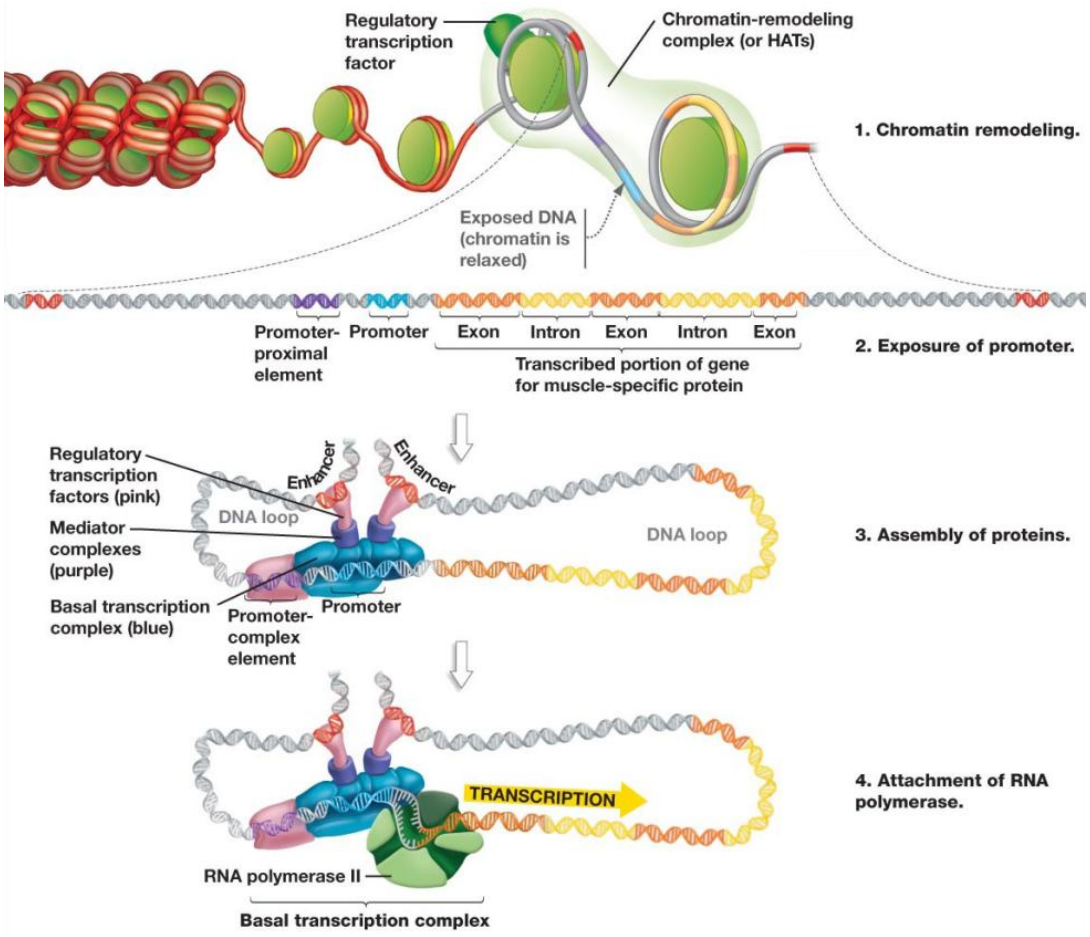

图 1-2　真核生物的转录调控过程

（http://www.uic.edu/classes/bios/bios100/lectures/genetic_control.htm）

图 1-2 是真核生物转录调控过程的基本模型：通过修饰染色质改变其结构以暴露启动子；通用转录因子通过识别 TATA-box 和转录起始位点结合到核心启动子区域。此时通用转录因子在启动子区域的结合并不稳定，只能提供很低的转录活性。接着序列特异性的转录因子结合到增强子、沉默子或近启动子区域的顺式元件上促进转录复合体的装配和稳定，增加启动子的活性。增强子可通过招募组蛋白修饰酶（如组蛋白乙酰转移酶）创造有利于转录的染色质环境或招募激酶磷酸化 RNA 聚合酶 II 羧基端结构域来促进转录起始和延伸(Farnham, 2009)。

本文研究的转录因子（transcription factor, TF）特指序列特异性结合 DNA 的转录因子，即上述四类转录调控蛋白中的第三类。它由 DNA 结合结构域和激活结构域组成，其中 DNA 结合结构域是很保守的。根据 DNA 结合结构域的不同，转录因子通常划分为不同的家族(Riechmann, *et al.*, 2000, Luscombe, *et al.*, 2000)。除了保守的 DNA 结合结构域外，转录因子通常还包含激活结构域。激活结构域包括酸性

结构域、富含谷氨酸的结构域和富含脯氨酸的结构域等几种类型(Carroll, 2000)。虽然大部分转录因子起激活转录的作用，但也有一部分转录因子能抑制基因的转录。转录因子抑制转录有直接和间接两种方式：直接方式如降低基本转录复合体的活性、改变染色质的结构或在转录延伸阶段抑制转录；间接方式如通过与激活作用的转录因子相互作用而抑制其功能(Carroll, 2000)。

1.2.2 转录因子的系统识别

1.2.2.1 拟南芥转录因子的系统识别

转录因子使用 DNA 结合结构域序列特异性的结合 DNA 来激活或抑制下游靶基因的转录。保守的 DNA 结合结构域常用来识别转录因子和将其划分到不同的家族。参考前人的工作，Riechmann 等总结出拟南芥转录因子家族及其演化关系（图 1-3）。该工作是对植物转录因子家族及其规则的第一次系统总结，它一方面反映了各家族的特征结构域，另一方面展现了不同家族之间的演化关系。特征结构域主要是 DNA 结合结构域，一部分家族如 MADS MIKC-type 还包括 K-box 等其它类型的结构域，可据此识别转录因子并划分到不同家族。

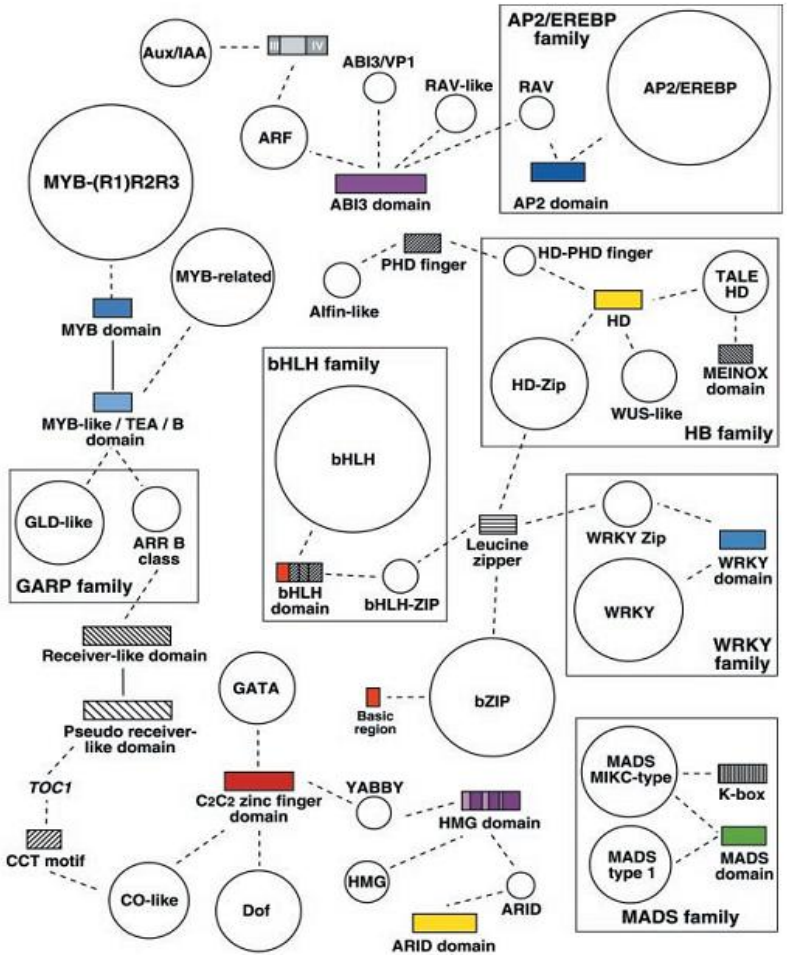

图 1-3　拟南芥转录因子家族及其演化关系(Riechmann, *et al.*, 2000)

第一个植物基因组——拟南芥基因组发布不久，Riechmann 等就使用 BLAST 和结构域模体(Motif)搜索的方式系统识别出 1533 个转录因子，约占当时已注释编码基因的 5.9%。与酿酒酵母、线虫和果蝇等其它真核生物相比，拟南芥具有更多的转录因子而且转录因子占基因组基因的比例也更高（表 1-2），体现了植物转录调控系统的高度复杂性。

表 1-2　真核生物基因组中的转录因子统计

物种	基因数	转录因子数	（%）
拟南芥　（*A. thaliana*）	~26000	1533	5.9
酿酒酵母（*S. cerevisiae*）	~6000	209	3.5
秀丽线虫（*C. elegans*）	~19000	669	3.5
黑腹果蝇（*D. melanogaster*）	~14000	635	4.5

注：表中数据来自(Riechmann, *et al.*, 2000)

1.2.2.2 植物转录因子数据库

Riechmann 等在拟南芥基因组发布不久便系统识别出其中的转录因子。在随后的几年内，致力于植物转录因子研究的学者通过总结文献不断更新和优化转录因子家族分类规则并构建了各种植物相关的转录因子数据库（表 1-3 和表 1-4）。其中既有包含单个物种的（如 DATF、AGRIS 等，表 1-3），也有包含多个物种的综合数据库（如 PlantTFDB、PlnTFDB 等， 表 1-4）；既有基于基因组注释的基因识别的（如 DATF、PlnTFDB 等），也有基于表达数据识别的（如 TOBFAC），还有综合基因组数据和表达数据的（如 PlantTFDB）。这些数据库为植物转录因子的功能和演化研究提供了宝贵的资源。

表 1-3 单物种的植物转录因子数据库

数据库	物种	网址链接	初建时间
AGRIS	拟南芥	http://arabidopsis.med.ohio-state.edu/	2003
DATF	拟南芥	http://datf.cbi.pku.edu.cn/	2005
RARTF	拟南芥	http://rarge.gsc.riken.jp/rartf/	2005
DRTF	水稻	http://drtf.cbi.pku.edu.cn/	2006
DPTF	杨树	http://dptf.cbi.pku.edu.cn/	2007
TOBFAC	烟草	http://compsysbio.achs.virginia.edu/tobfac/	2008
SoyDB	大豆	http://casp.rnet.missouri.edu/soydb/	2010
wDBTF	小麦	http://wwwappli.nantes.inra.fr:8180/wDBFT/	2010

表 1-4 多物种的植物转录因子数据库

数据库	物种数	网址链接	初建时间
PlantTFDB	83	http://planttfdb.cbi.pku.edu.cn/	2007
PlnTFDB	19	http://plntfdb.bio.uni-potsdam.de/	2007
PlanTAPDB	6	http://www.cosmoss.org/bm/plantapdb	2007
DATFAP	13	http://cgi-www.daimi.au.dk/cgi-chili/datfap/frontdoor.py	2008
GRASSIUS	5	http://grassius.org/	2009
LegumeTFDB	3	http://legumetfdb.psc.riken.jp/	2010
TreeTFDB	6	http://treetfdb.bmep.riken.jp/	2013

1.2.3 转录因子的演化

1.2.3.1 转录因子家族的起源与演化

在转录因子的演化历程中，通过产生新的特征结构域、特征结构域与原有结构域发生实质性改变（仅在序列上已无法判断它们之间的同源关系）或者通过原有结构域的重组等方式不断产生新的家族(Riechmann, *et al.*, 2000)。通过与酵母、线虫、果蝇等其它真核生物相比，Riechmann 等发现拟南芥具有很多植物特异的转录因子家族。这些植物特异的转录因子约占拟南芥转录因子总数的 45%，在酿酒酵母、线虫、果蝇中支系特异的转录因子所占比例也分别达到 32%、47%和 14% (Riechmann, *et al.*, 2000)。Charoensawan 等通过研究古细菌、真细菌和真核生物等三界 500 多个物种发现在所有 131 个家族中，只有 2 个是三界共有的(Charoensawan, *et al.*, 2010)。这些结果体现了转录因子家族的高度支系特异性。

不同类别生物包含的转录因子家族有很大不同，每个家族所占的比例亦有所不同。如在拟南芥中，锌指结构蛋白占转录因子的比例小于 22%，而在果蝇、线虫和酵母中的比例则分别高达 51%、64%和 56%。在拟南芥中，没有哪个家族像线虫的核荷尔蒙受体、果蝇的 C2H2、酵母的 C6 和 C2H2 扩张到如此高的比例，它们分别占各自基因组转录因子总数的~38%、~46%、~25%和~25% (Riechmann, *et al.*, 2000)。D. Lang 等通过系统识别 19 个植物中的转录相关蛋白，同样发现转录因子家族存在很多枝系特异性的扩张和收缩，而转录因子家族的大规模出现和扩增与植物登陆和被子植物的出现在时间上吻合，这或许暗示着转录因子在植物形态复杂性中的作用(Lang, *et al.*, 2010)。

1.2.3.2 转录因子在功能上的分化

基因复制为演化提供了原材料，是演化新颖性的源泉(Ohno, 1970)。基因在不断的发生复制，特别是在植物中存在多次全基因组复制，那么复制后的基因在功能上是如何演化的？在拟南芥中的研究表明复制基因在表达、蛋白蛋白相互作用、功能等方面都存在快速的分化(Blanc, *et al.*, 2004, *Arabidopsis* Interactome Mapping Consortium, 2011)。

保留下来的复制基因在功能上发生分化是多倍化后长期演化的主要特征(Blanc,

et al., 2004)，那么转录因子又是怎样？通过对酵母转录调控网络的研究，Conant 发现基因组复制后转录因子很快就会在调控网络中完成功能的分歧(Conant, 2010)。在真核生物不同模式物种中的研究表明转录因子的功能存在很大程度的演变：在不同物种中，相似的功能可能由不同源的转录因子来完成，同家族的转录因子之间则完成不同的功能(Riechmann, *et al.*, 2000)。对植物各转录因子家族的分析也表明各家族的成员之间在功能上存在明显的分化(Heim, *et al.*, 2003, Jin, *et al.*, 1999, Smaczniak, *et al.*, 2012)。

1.2.3.3 转录因子演化的作用

通过在特定时间和地点调控相应靶基因的转录，转录因子在植物形态发育和应对胁迫中发挥关键作用。转录因子和转录调控的演化在植物形态改变和环境适应中发挥着极为重要的作用(de Bruijn, *et al.*, 2012)。极为明显的例子就是转录因子和转录调控的改变在作物驯化中的作用。

玉米在大约 9000 年前由墨西哥蜀黍驯化而来，但是它们在分枝和果实形态上存在很大差异（图 1-4）。与玉米相比，蜀黍具有更多的分枝（图 1-4A）。研究发现在驯化过程中选择作用于调控形态的转录因子 tb1 的调控区域，其表达水平的变化引起分枝数目的改变(Doebley, *et al.*, 1997, Wang, *et al.*, 1999)。玉米果实外面没有外壳包被，而蜀黍的果实则有一层厚厚的外壳（图 1-4B）。该性状与 SBP 家族蛋白 TGA1 有关，如果将玉米的 TGA1 转到蜀黍中，蜀黍的果壳就会开裂(Wang, *et al.*, 2005)。这些例子充分说明转录因子和转录调控的演化对植物形态改变的重要性。

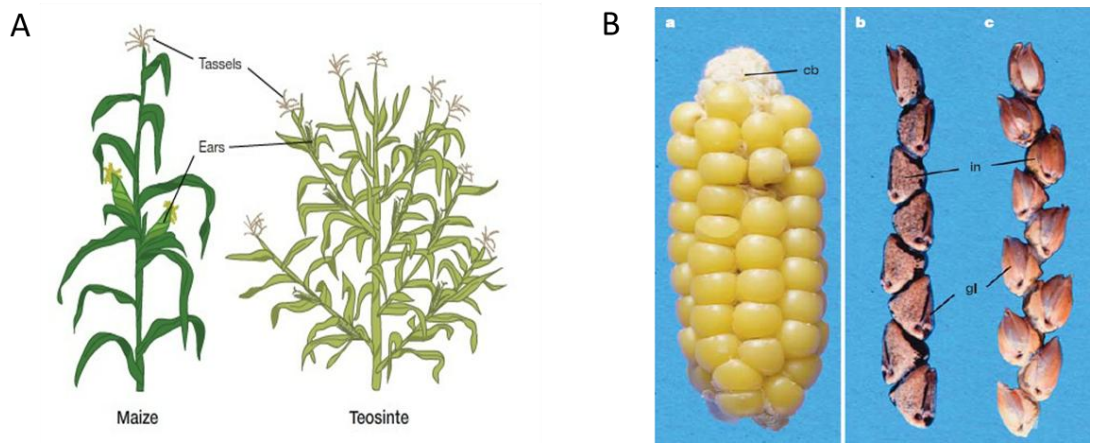

图 1-4　玉米与蜀黍在分枝（A）与果实形态（B）上的差异

图 A: http://learn.genetics.utah.edu/content/variation/corn/; 图 B: Wang, *et al.*, 2005

1.3转录调控网络

1.3.1 转录调控网络数据

大规模的转录调控网络数据是研究转录调控系统网络结构和演化规律的基础。作为研究常用的模式生物，大肠杆菌、酿酒酵母等物种内部很多生物过程的转录调控机理已经研究清楚并分散在大量的科学文献中。通过从文献中收集这些调控关系而构建的数据库如 RegulonDB、YEASTRACT 等为相关领域的研究提供了很多便利。近些年，随着 ChIP-chip、ChIP-seq 等高通量技术的发展使基因组范围内识别转录因子的结合位点成为可能。一些大型项目如 ENCODE、modENCODE 识别了人（ENCODE）、线虫和果蝇（modENCODE）中很多转录因子的结合位点。

表 1-5 收录转录调控数据的数据库

数据库名	物种	网址
RegulonDB	大肠杆菌	http://regulondb.ccg.unam.mx/
EcoCyc	大肠杆菌	http://ecocyc.org
CoryneRegNet	棒状杆菌	http://www.coryneregnet.de/
DBTBS	枯草芽孢杆菌	http://dbtbs.hgc.jp/
RhizoRegNet	中华根瘤菌	http://rhizoregnet.cebitec.uni-bielefeld.de/
RegTransBase	原核生物	http://regtransbase.lbl.gov
YEASTRACT	酵母	http://www.yeastract.com/
REDfly	果蝇	http://redfly.ccr.buffalo.edu
modENCODE	果蝇、线虫	http://www.modencode.org
ENCODE	人	http://genome.ucsc.edu/ENCODE/
AGRIS	拟南芥	http://arabidopsis.med.ohio-state.edu/
ORegAnno	真核生物	http://www.oreganno.org
TRANSFAC	真核生物	http://www.gene-regulation.com/cgi-bin/pub/databases/transfac

参考相关文献和网站，转录调控数据的数据库整理成表（表 1-5）。根据文献中确定调控关系的方法（小规模实验与高通量实验），这些数据库又可分为不同类型：基于小规模实验的（如 RegulonDB）、高通量实验的（如 AGRIS）和二者皆有的（如

YEASTRACT）。

1.3.2 转录因子与其靶基因之间的表达相关性

由于转录因子调控下游靶基因的转录，人们预期转录因子与其靶基因在表达上存在一定的相关性。这种表达相关性也广泛应用于通过表达数据推测转录调控网络中(Basso, *et al.*, 2005, Bansal, *et al.*, 2007)。然而转录因子与其靶基因之间的表达相关性到底如何，人们并不清楚。随着大肠杆菌和酵母中积累了足够多的转录调控数据，人们有机会去系统研究这两个模式生物中转录调控在表达相关性上的总体模式。通过分析大肠杆菌和酵母中转录因子与其靶基因之间的表达相关性，Herrgard 等发现只有小部分转录调控具有明显的表达相关性(Herrgard, *et al.*, 2003)。Wu 等在酵母中的研究发现了类似的模式。意外的是他们发现转录因子与靶基因间的总体表达相关性竟然低于随机基因对之间的表达相关性 (Wu, *et al.*, 2012)。这些结果体现了转录调控的复杂性，也体现了通过表达数据预测表达相关性存在很多困难。

1.3.3 转录调控网络的架构

1.3.3.1 生物网络的一般特征

生物网络和万维网等复杂网络中节点的连接数（即度）均服从幂律（Power law）分布，这种分布与网络的尺度无关（scale-free）(Barabasi, *et al.*, 1999)。通过模拟分析，Barabasi 等发现这种现象是两种机制共同作用的结果：1）网络通过不断加入新节点来扩张；2）新节点加入网络时倾向于连接到连接比较好的节点上；以上两条件缺一不可(Barabasi, *et al.*, 1999)。与其它复杂网络不同，生物网络在不断的发生复制和演变，那么复制是否会导致生物网络出现上述特性呢？Teichmann 等通过研究靶基因复制对大肠杆菌和酵母转录调控网络节点度分布的影响，发现靶基因的复制不是导致节点度呈幂律分布的原因(Teichmann, *et al.*, 2004)。

1.3.3.2 转录调控网络的结构

转录调控网络在结构上是否存在一些设计模式呢？转录调控数据的不断积累使科学家有机会去研究这一问题。通过对单细胞生物大肠杆菌和酿酒酵母的转录

调控网络进行系统分析，Alon 等发现它们的转录调控网络中存在一些重复出现的调控模式，这些调控模式出现的频率显著高于相对应的随机网络(Shen-Orr, *et al.*, 2002, Milo, *et al.*, 2002)。他们将这些出现次数显著多于相应随机网络的调控模式定义为结构元件（network motifs）(Shen-Orr, *et al.*, 2002, Milo, *et al.*, 2002, Alon, 2007)。

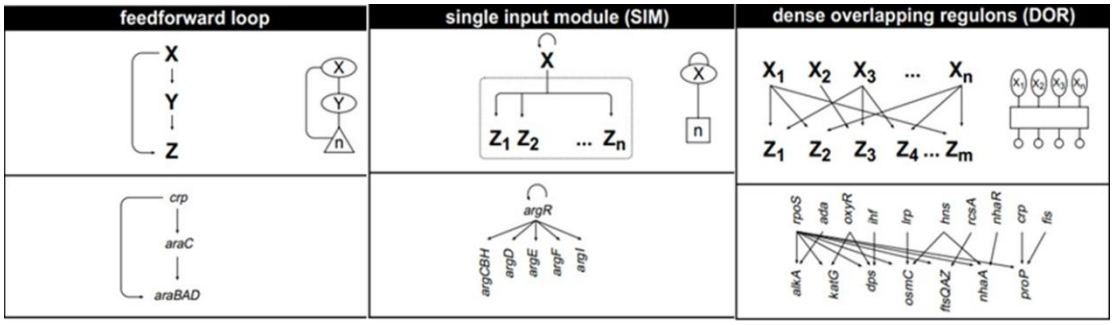

图 1-5　大肠杆菌转录调控网络中的结构元件(Shen-Orr, *et al.*, 2002)

如图 1-5 所示，大肠杆菌转录调控网络中有三类网络结构元件，包括前馈环(Feedforward Loop, FFL)、简单的输入调控模块(Simple Input Module, SIM)和高度重叠的调控模块（Dense Overlapping Regulons, DOR）等调控模式(Shen-Orr, *et al.*, 2002)，下面的一层为每种结构元件在大肠杆菌中的实例。

为什么单细胞生物的转录调控网络中会有这些结构元件呢？它们有什么功能吗？动力学模拟和实验研究表明这些结构元件能够完成某些特定的生物学功能(Rosenfeld, *et al.*, 2002, Becskei, *et al.*, 2000, Kalir, *et al.*, 2004, Mangan, *et al.*, 2003, Alon, 2007)。比如 FFL 能过滤噪音信号或者在信号来临（或结束）时延迟对信号的响应(Mangan, *et al.*, 2003)。如图 1-6 所示，基因 X 代表某种输入信号，基因 Z 则代表系统的响应。如果只有 X 和 Y 同时存在才能启动 Z 的转录，则当 X 已表达而 Y 的表达量比较低时，Z 的转录不能被启动，从而过滤信号中的噪音。由于 Y 的转录也是由 X 启动的，只有 Y 的表达量到达一定水平才能和 X 一起启动 Z 的转录，从而在信号来临时推迟对信号的响应。如果 X 或 Y 单独就能启动 Z 的转录，则不会在信号来临时延迟 Z 的转录，而会在信号结束后的一段时间内继续激活 Z 的转录。由于单细胞生物要根据环境变化迅速作出调整，这些结构元件对于它们完成这些功能起着重要的作用(Alon, 2007)。

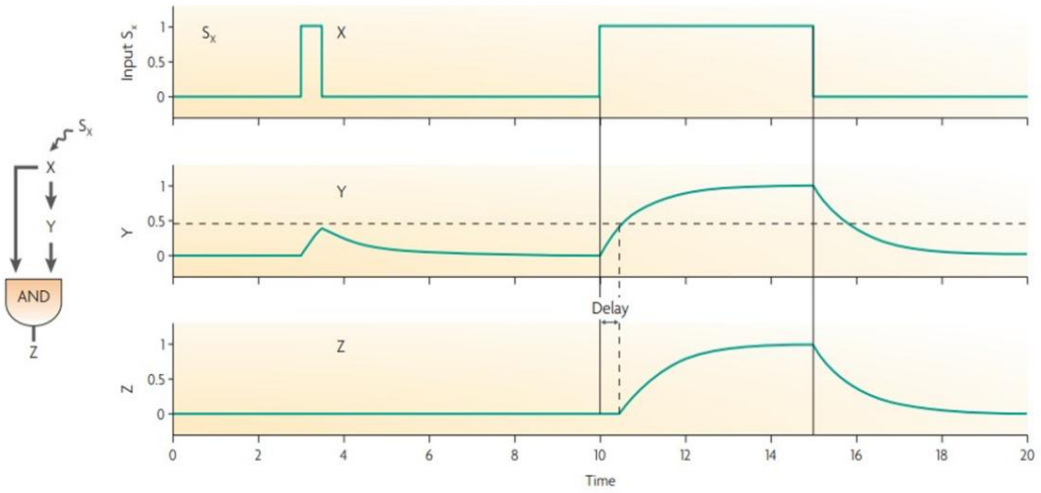

图 1-6　动力学模拟前馈环(FFL)的功能(Alon, 2007)

在高等生物的发育网络中，人们发现其中有一些反馈环（Feedback Loop，FBL），动力学模拟表明这些反馈环能在信号结束后保持某种状态（图 1-7a）或者从一种状态转变到另一种状态（图 1-7b）。这些功能是细胞分化等多细胞发育所需要的(Alon, 2007)。

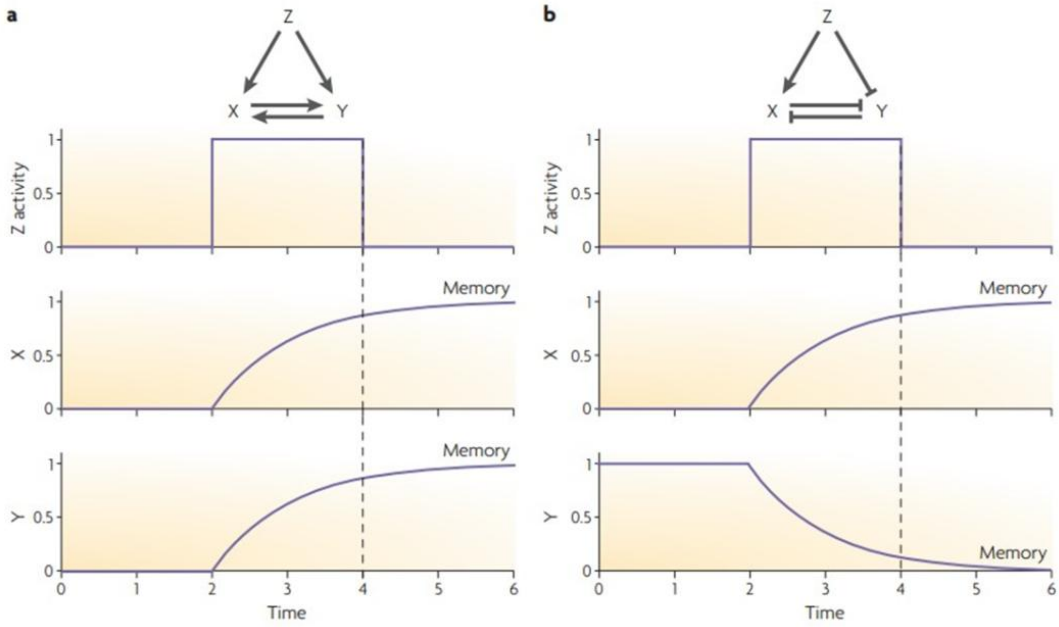

图 1-7　发育网络中的结构元件及其功能(Alon, 2007)

上述所述，对单细胞生物转录调控网络的系统分析以及高等生物某些特定调控通路的分析发现转录调控网络是由某些结构元件组成的，这些结构元件同时也是

能完成某些特定生物学功能的功能模块。

1.3.3.3 内源系统与外源系统在网络架构上的差异

Luscombe 等通过整合遗传、生化、ChIP-chip 等方面的数据,构建了一个酿酒酵母转录调控网络，并将其划分为细胞周期(Cell cycle)、孢子形成(Sporulation)、双峰转换(Diauxic shift)、DNA 损伤(DNA damage)和应对胁迫(Stress response)等 5 种不同状态下的调控网络，最后将以上 5 个子网络分为内源性和外源性两种类型。内源性的（细胞周期、孢子形成）包含多个不同状态并且是内部发育机制的渐序发展，而外源性的（双峰转换、DNA 损伤和应对胁迫）则是通过迅速改变大量基因的转录状态以应对外部的各种刺激(Luscombe, *et al.*, 2004)。

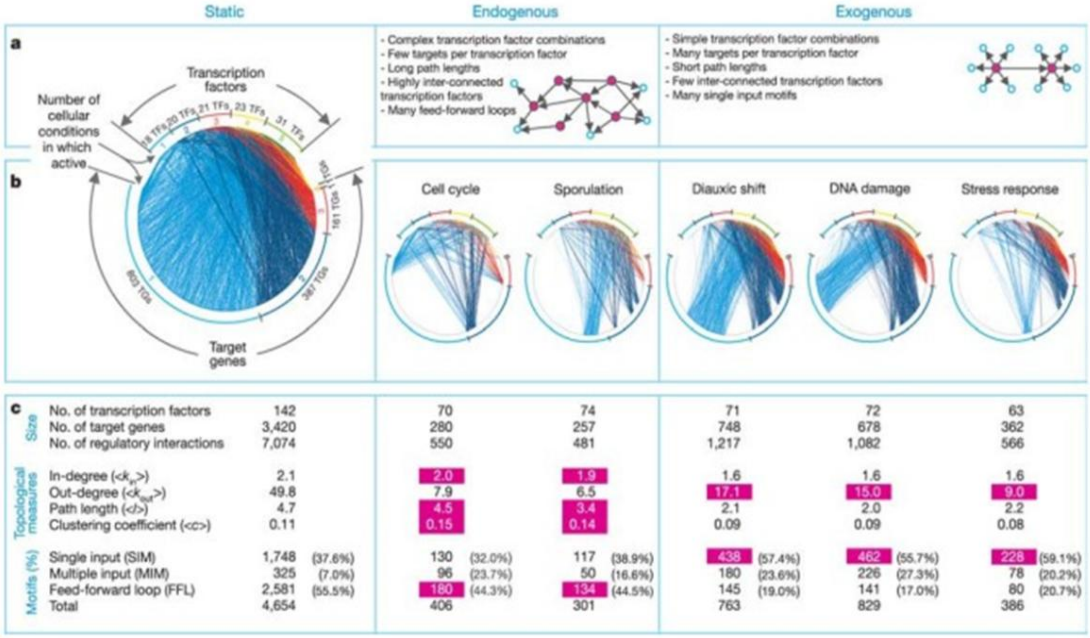

图 1-8　酿酒酵母内源性系统和外源性系统在网络架构上的差异(Luscombe, *et al.*, 2004)

对以上两类转录调控网络的系统分析则发现它们无论在结构元件的组成上还是全局拓扑结构上都有着显著的不同。如图 1-8 所示，在结构元件的组成上，内源性网络更多的使用前馈环(FFL)而外源性系统则更多的使用简单的调控模块 SIM。在全局拓扑结构上，内源性系统中的基因处于更多转录因子的调控下，而其中的转录因子则具有相对较少的靶基因，其调控网络具有更长的路径长度和更大的聚类系数。从生物学角度来讲，这些结果表明内源性系统使用更加复杂的调控精确

的推进各生命过程，而外源性系统则使用更具影响力的转录因子通过简单的调控快速响应外部的各种刺激(Luscombe, *et al.*, 2004)。

1.3.3.4 转录因子的性质与网络构建

研究表明原核生物和真核生物的转录调控网络呈层次性结构(Yu, *et al.*, 2006, Ma, *et al.*, 2004, Farkas, *et al.*, 2006)。通过将酵母转录调控网络中的转录因子划分为顶层、中心层和底层三类，Jothi 等系统研究了不同层次转录因子的静态和动态性质，以探索节点的性质与它在网络中所起作用的关系。结果发现同一层次的转录因子在性质上更为类似，不同层次的转录因子则具有不同的性质。与中心层和底层转录因子在蛋白水平上相比，顶层转录因子具有更高的表达丰度、更长的半衰期和更高的表达噪音（图 1-9）(Jothi, *et al.*, 2009)。

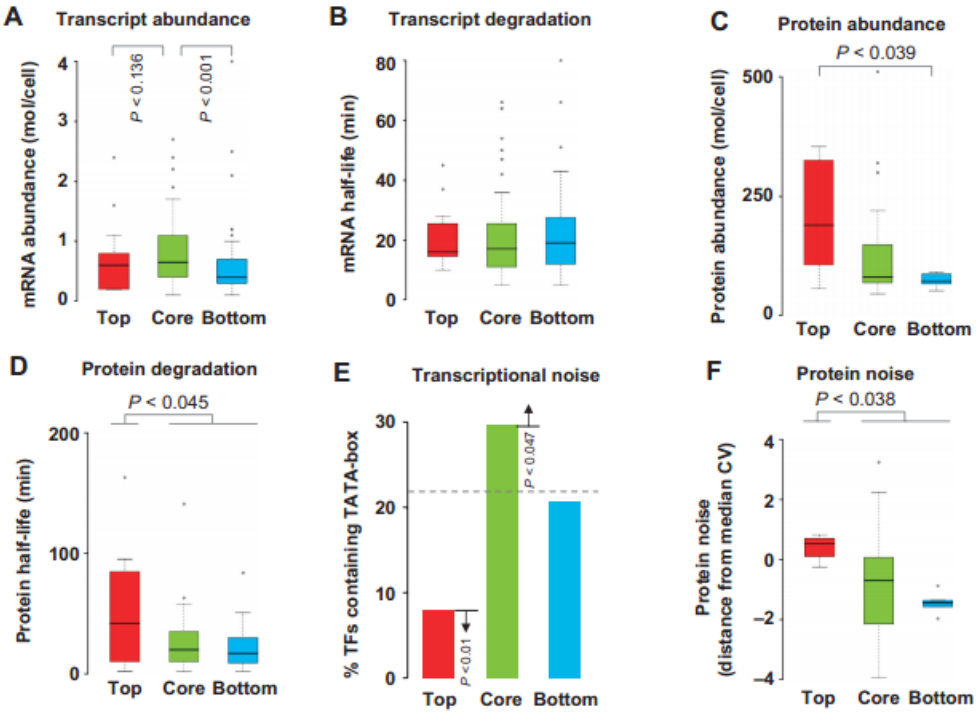

图 1-9　不同网络层次的转录因子在动态性质上的比较(Jothi, *et al.*, 2009)

此外，Luscombe 等人的分析揭示了内源系统和外源系统在网络结构上的显著差异及构建它们的转录因子在靶基因数目上的不同(Luscombe, *et al.*, 2004)。Wang 等在大肠杆菌中的研究发现转录本具有较短半衰期的转录因子显著富集于转录调控的结构元件和集散节点（hub）上(Wang, *et al.*, 2005)。这些结果暗示转录因子的

性质或许与它构建的网络之间存在某种内在的联系。

1.3.4 转录调控的演化

1.3.4.1 转录调控在快速的演化

转录调控的演化是形态多样性的重要源泉。早在基因调控发现不久，人们就意识到调控的演化在形态差异上的重要性。Britten 等通过研究重复序列发现其可能通过重塑基因调控带来演化上的新颖性(Britten, *et al.*, 1971)。鉴于人和黑猩猩在蛋白序列上的差异无法解释他们在形态上的巨大差异，King 等认为调控上的变化可能导致了他们在表型上的差异(King, *et al.*, 1975)。近来的一些研究如人中一些群体的乳糖不耐症(Tishkoff, *et al.*, 2006)、昆虫翅膀形态和色斑的改变(Gompel, *et al.*, 2005, Werner, *et al.*, 2010)等则充分说明了转录调控的演化对形态和生理改变的重要性。

近年来，ChIP-chip/ChIP-seq 技术的发展使基因组水平上研究转录因子结合位点的改变成为可能。在酵母种间和种内的研究都发现直系同源基因的调控位点发生了很大程度的改变(Borneman, *et al.*, 2007, Zheng, *et al.*, 2010)。在果蝇近缘种、人和小鼠的比较中也发现了类似的模式(Bradley, *et al.*, 2010, Schmidt, *et al.*, 2010)。这些研究结果表明转录调控在快速的演化，而顺式元件的改变在其中起主要作用(Borneman, *et al.*, 2007)。

从上面的结果我们知道转录调控改变之快。那么与其它类型的生物网络相比，转录调控网络的演化速度处于怎样的水平呢？Shou 等通过比较蛋白蛋白相互作用网络、遗传相互作用网络、代谢网络、转录调控网络、激酶磷酸化网络和 miRNA 调控网络等生物网络的演化，发现转录调控网络在这些类型的网络中具有最高的演化速率（图 1-10）(Shou, *et al.*, 2011)。

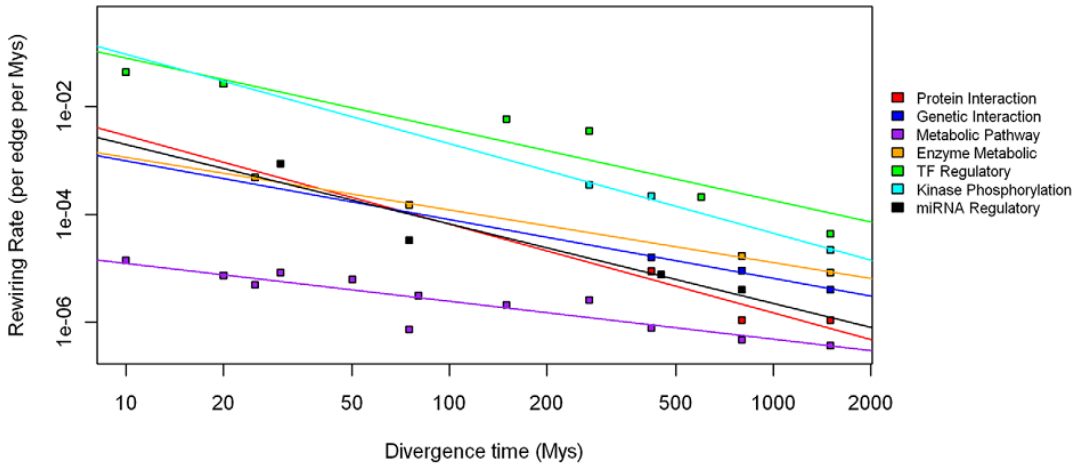

图 1-10 各类型生物网络的演化速率(Shou, *et al.*, 2011)

1.3.4.2 转录调控的演化模式

基因组水平的研究表明转录调控网络在快速的演化，即便亲缘关系很近的物种其调控通路都发生了显著的改变。Tuch 等通过总结现有研究提出真核生物转录调控演化的几个特点：1)在较短的演化时间内，转录调控就会发生很大程度的分歧；2)在不同物种内，一套相同的共表达基因可能由不同的调控机制所调控；3)生物体内存在很多转录调控蛋白之间的相互作用，这种共同调控的方式也有利于调控通路的改变(Tuch, *et al.*, 2008)。

在真核生物中，酵母转录调控的演化研究的相对比较清楚。Li 等通过总结酵母转录调控通路的演化提出转录调控演化的 3 个基本模式：1)转录因子及相互作用的蛋白在结合特异性上是保守的，调控的靶基因发生了改变（图 1-11a）；2)调控的靶基因是保守的，不过由一个转录因子的调控变为另外一个转录因子调控(图 1-11b)；3)调控蛋白之间产生新的组合方式（图 1-11c）。以上三种情况是转录调控网络演化的简单模式，实际情况可能是三种情况的组合或者更复杂的情况(Li, *et al.*, 2010)。

与真核生物相比，基因的水平转移在细菌转录调控的演化中发挥着重要的作用(Perez, *et al.*, 2009)。在大肠杆菌中的研究发现其中绝大多数的邻接性转录因子（转录因子与调控的靶基因在相邻位置上）是通过基因水平转移获得的，大多数全局转录因子则是演化上更加保守通过垂直演化从祖先继承而来的(Price, *et al.*, 2008)。与其它基因相比，水平转移进来的基因通常处于更加复杂的控制之下，防止它对宿主产生有害影响(Price, *et al.*, 2008, Perez, *et al.*, 2009)。

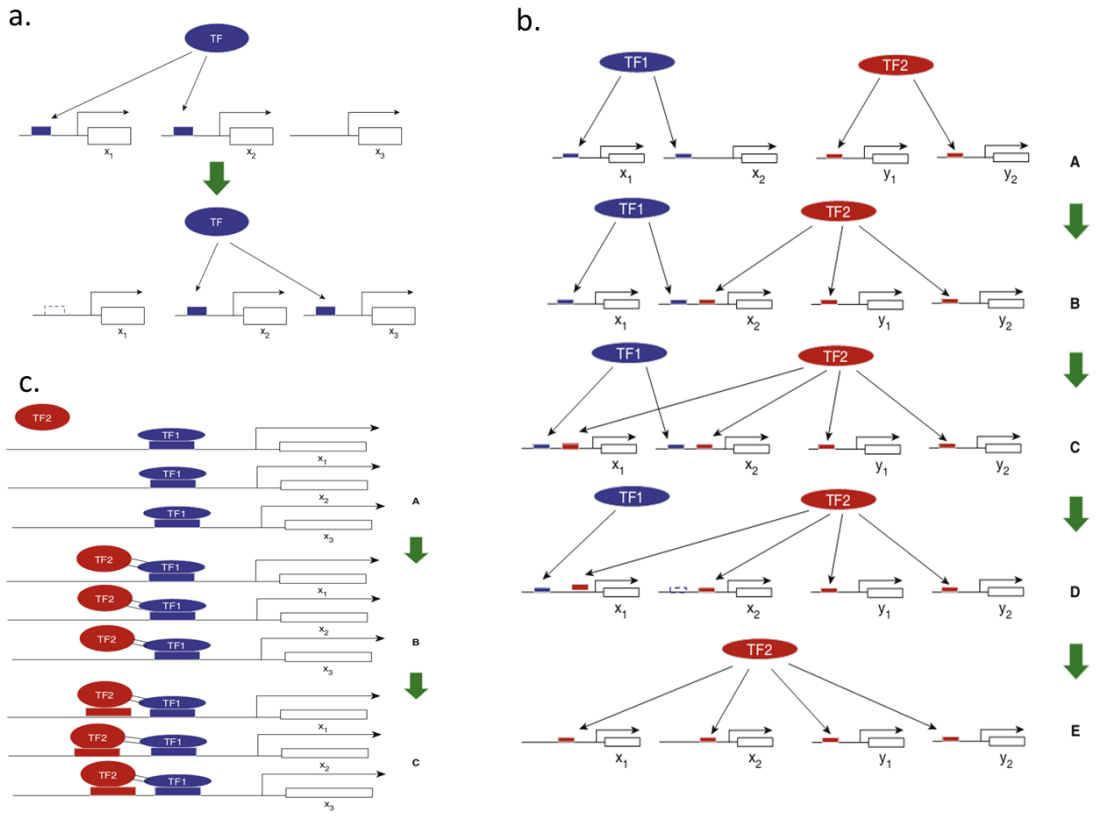

图 1-11 酵母中转录调控演化的基本模式(Li, *et al.*, 2010)

1.4 问题的提出与本文章节安排

1.4.1 问题的提出

目前已测序植物基因组已覆盖了绿色植物的各大分枝，各植物门类基因组中有多少转录因子、转录因子家族以及转录因子占基因组基因的比例等问题都尚不清楚。虽然目前已测定了 67 个植物的基因组，然而像小麦、甘蔗、烟草等重要经济作物的基因组尚未发布（或尚未注释）。对于此类物种，我们怎样才能识别尽量全的转录因子供相关领域的科学家使用。一个完整的转录因子家族分类规则是系统识别转录因子的前提，然而目前最新的转录因子分类规则也只是基于数年前文献总结的。随着科学研究的不断推进，近些年是否发现了新的转录因子家族，又是否有旧的家族被证明不是转录因子，科学界对某个家族规则的普遍认识是否发生了改变？当我们识别出转录因子后，提供什么样的注释才能更好的服务于本领域的科研工作者，又如何让用户能方便的获取我们的数据和使用我们的转录因子预

测流程？

识别出转录因子只是理解植物转录调控系统的第一步。作为研究最为清楚的植物模式物种，拟南芥中很多生物过程的转录调控机制已经研究清楚并记录在文献中。我们能否从文献中将这些转录调控关系收集起来构建一个高质量、基因组范围的拟南芥转录调控网络。如果在植物领域能有此转录调控网络，我们就可以研究它的转录调控在表达相关性上的总体模式。在大肠杆菌和酵母等单细胞生物中的研究表明它们的转录调控网络由一些结构元件组成，那么植物转录调控网络又由哪些结构元件组成呢？与单细胞的生物相比，它是否存在一些新的结构元件？作为植物转录调控系统的两个主要部分，发育系统和应激系统在网络架构上是否有所不同？

在植物登陆期间产生了很多新的转录因子家族，这些新类型转录因子和古老类型转录因子是如何参与转录调控系统构建的？与古老类型相比，新类型转录因子在性质上有何特点？它们在参与发育系统和应激系统的构建上是否存在某种偏好性？如果存在的话，它的性质与这种偏好性之间是否存在某种内在关系？

1.4.2 本文章节安排

本文是对上述问题的一些探索和研究。文章安排如下：

第 1 章简要介绍与本文研究相关的背景知识和本文要研究的生物学问题。其中背景知识主要包括植物的特点、转录因子和转录调控的重要性、转录调控网络的架构与演化等。

第 2 章介绍转录因子的系统识别与两个版本转录因子数据库的构建。本章前四节介绍 PlantTFDB 2.0 的构建，重点介绍数据整合、转录因子家族分类规则、转录因子注释及供用户获取数据的 Web service 搭建。第 5 节介绍 PlantTFDB 3.0 的构建，重点介绍覆盖绿色植物各大分枝的转录因子全谱、知识型数据的收集、演化相关的分析及转录因子预测平台的搭建。

第 3-5 章分别介绍拟南芥转录调控网络的收集与构建、网络的架构和演化分析。其中第 3 章介绍拟南芥转录调控网络的收集、网络质量评估、网络涉及的生物过程及拟南芥转录调控的一些总体模式等。第 4 章系统研究一个植物的转录调控系统是如何架构的。重点介绍了拟南芥转录调控网络的结构元件、发育系统和应激系统在结构元件、网络全局拓扑结构和构建它们的转录因子性质上的差异等。第 5

章介绍转录因子在参与转录调控系统构建时的倾向性。重点讨论了新类型和古老类型转录因子在性质和转录调控系统构建上的差异、转录因子的性质与网络构建的关系等问题。

本文最后一章（第 6 章）总结了本文的工作并展望了未来可能的工作方向。

第 2 章 植物转录因子的系统识别与数据库构建

2.1概述

2.1.1 植物转录因子相关数据库

转录因子是一类序列特异性结合 DNA 并能激活和/或抑制相应基因转录的蛋白，它在植物的生长发育和应对生物和非生物胁迫中起着关键的作用(Riechmann, *et al.*, 2000, Riechmann, 2006)。系统识别植物体内的转录因子并提供相应注释有助于研究转录因子的功能和演化。

拟南芥基因组序列发布后不久，Riechmann 等基于前人研究总结出一套转录因子家族分类图并从拟南芥基因组中识别出 1533 个转录因子，占基因组注释基因的 5.9% (Riechmann, *et al.*, 2000)。在随后几年里，随着植物基因组序列陆续发布，多个植物相关的转录因子数据库相继建立。它们中既有针对单个物种的转录因子数据库如 AGRIS (Yilmaz, *et al.*, 2011)、RAPTF (Iida, *et al.*, 2005)和 TOBFAC (Rushton, *et al.*, 2008)，也有针对某类植物或者综合的转录因子数据库如 PlnTFDB (Pérez-Rodríguez, *et al.*, 2010)、PlantTAPDB (Richardt, *et al.*, 2007)、GRASSIUS (Yilmaz, *et al.*, 2009)和 LegumeTFDB (Mochida, *et al.*, 2010)等。与此同时，随着拟南芥基因组(*Arabidopsis* Genome Initiative, 2000)、水稻基因组(Goff, *et al.*, 2002, Yu, *et al.*, 2002)以及杨树基因组(Tuskan, *et al.*, 2006)的发布，我们实验室通过收集其中的转录因子先后构建了拟南芥转录因子数据库 DATF (Guo, *et al.*, 2005)、水稻转录因子数据库 DRTF (Gao, *et al.*, 2006)和杨树转录因子数据库 DPTF (Zhu, *et al.*, 2007)。随着对相关文献的综述和三个数据库的相继构建，我们也发展了一套自己的分类规则。2007 年，我们实验室整合了这 3 个数据库并加上其它 19 个物种构建了一个综合的植物转录因子数据库 PlantTFDB (Guo, *et al.*, 2008)，成为当时最为全面的植物转录因子数据库。

2.1.2 PlantTFDB 存在的问题

作为一个综合的植物转录因子数据库，PlantTFDB 自建立以来已收到上千万次

访问，成为该领域的权威数据库之一。但它在以下几方面仍存在问题:

1）**数据不足**

PlantTFDB 中只有 5 个物种使用基因组注释的基因(Guo, *et al.*, 2008)。随着测序技术快速发展，越来越多植物基因组测序完成。截止到 2010 年 7 月，已有 29 个植物基因组测序完成。因此，有必要系统识别这些物种的转录因子以辅助相关领域的研究。此外，目前基因组注释大多是自动注释的，存在很多问题，比如注释不完整和注释错误等(Ouyang, *et al.*, 2009)。

2）**转录因子分类规则有待更新和优化**

- PlantTFDB 中包含 64 个转录因子家族，但是这些家族并非全都是转录因子家族，里面混着一些起染色质修饰、甲基化调控等功能的蛋白。
- 近几年随着研究的深入，发现和定义了一些新的转录因子家族。与此同时，也发现一些原来认为的转录因子家族并非转录因子。
- 并非所有包含 DNA 结合结构域（DNA-binding domain，DBD）的蛋白都是转录因子，有必要设计合理的规则过滤包含 DNA 结合结构域而没有转录因子活性的蛋白，降低转录因子预测的假阳性。
- 合理的阈值是提高预测准确性的关键，原先阈值统一设为 E 值（E value）小于 0.01 不能同时适用于所有的结构域模型。

3）注释不足

一个好的注释能帮助用户了解该转录因子的功能并为后续分析提供重要线索。但原来数据库只有一些简单的序列、结构域方面的注释，到知名数据库的跨库链接也不是很多。

4）数据库结构和用户界面设计不统一

由于 PlantTFDB 是在整合了 DATF、DRTF 和 DPTF 的基础上构建的，这就造成了各物种间相互独立、数据库结构和用户界面不统一等情况，不便于用户浏览和使用。

针对上述问题，我们从数据、分类规则、注释、数据库结构和用户界面设计等四方面入手，对原有数据库做了大的改动，构建了 PlantTFDB 2.0 (Zhang, *et al.*, 2011)，并于今年 6 月更新到 PlantTFDB 3.0 (Jin, *et al.*, 2013)。下面第 2-4 节将依次介绍我们如何对这四部分进行改进和构建新的植物转录因子数据库 PlantTFDB 2.0，在第 5 节将介绍 PlantTFDB 3.0 的构建及其新特征。

2.2数据整合

2.2.1 数据源

为尽可能全的识别植物转录因子，我们需要一个足够完整的蛋白组。虽然基因组测序项目注释的基因是蛋白组数据最主要的来源，但是目前大多数基因组注释严重依赖自动注释流程，容易导致基因注释错误和注释不完整(Ouyang, *et al.*, 2009)。另外并不是所有植物都具有基因组序列，一些重要的经济作物如小麦、大麦、白菜等的基因组在当时都还未测定。除基因组项目注释的基因外，一些综合的数据库如 RefSeq (Pruitt, *et al.*, 2009)、UniGene (Sayers, *et al.*, 2010)和 PlantGDB (Duvick, *et al.*, 2008)也都包含大量的基因注释和表达数据。

下面先简要介绍一下基因组注释外的其它三个数据源：

1） RefSeq

全称 NCBI 参考序列数据库，旨在整合一套完整非冗余的转录本和蛋白的数据集(Pruitt, *et al.*, 2009)。RefSeq 当时包含 10 个植物物种，其中部分物种已经过人工校对，是一个高质量的数据源。

2） UniGene

NCBI 收录表达序列标签（EST）的数据库，它依靠计算的方式将可能来源于同一基因位点的EST聚到一起，从中选取最长的EST来代表这一簇EST，即“Unique UniGene”(Sayers, *et al.*, 2010)。非特别说明，本文中所说的 UniGene 均指“Unique UniGene”。

3） PlantGDB

对 GenBank 中转录本条目大于 10000 的植物物种，PlantGDB 使用特定的流程为每一物种都拼装了一套转录本，称为 PUT （PlantGDB-assembled unique transcripts）(Duvick, *et al.*, 2008)。

为制定合理的方法来整合不同数据源，我们首先要了解不同数据源的数据特点和它们之间的关系。因为基因组注释的基因和 RefSeq 注释的基因使用同一参考基因组，所以只有当序列完全相同时我们才认为它们为同一序列。由于 EST 测序会引入较高的测序错误，UniGene 和 PUT 在比对到相应基因组时设定的阈值为比对上的长度占总体的比例（Coverage）不小于 90%、序列一致性（Identity）不小于

95%。为避免引入过多错误，这类数据与其它数据源比较时，只要比对上的长度占总体的比例不小于 90%、序列一致性不小于 95%，则认为它们是同一序列。

下面以拟南芥和水稻（*Oryza sativa subsp. japonica*）为例，比较这四个数据源之间相互包含的情况（表 2-1）。从表 2-1 中可以看出，四个数据源之间不存在一个数据源能完全包含其它数据源，每个数据源都能提供自己特有的基因注释。即便在注释最完整的拟南芥中，RefSeq、PlantGDB 和 UniGene 也分别能为现有基因组注释提供 228、3521 和 681 条转录本。在注释相对较差的其它物种如水稻，这些数据源则提供更多的新转录本。通过将这些数据源特有的序列比对到基因组上，我们发现这些序列大多为新的可变剪接体（图 2-1）。

表 2-1　四个数据源之间的相互包含情况。其中(A、B)、（C、D）分别为拟南芥和水稻中的情况。（A、C）为数据源间重叠的部分占列中数据源的比例，（B、D）为列中的数据源相对行中数据源所特有的转录本。

A.

	RefSeq	**PUT**	**UniGene**	**基因组注释**
RefSeq	-	0.66	0.95	0.98
PUT	0.59	-	0.72	0.59
UniGene	0.72	0.77	-	0.68
基因组注释	0.99	0.85	0.97	-

B.

	Refseq	**PUT**	**UniGene**	**基因组注释**
Refseq	-	8083	1120	612
PUT	13638	-	5778	13669
UniGene	9236	5497	-	10555
基因组注释	228	3521	681	-

C.

	Refseq	**PUT**	**UniGene**	**基因组注释**
Refseq	-	0.53	0.81	0.36
PUT	0.54	-	0.72	0.33
UniGene	0.59	0.61	-	0.33
基因组注释	0.75	0.77	0.89	-

D.

	Refseq	**PUT**	**UniGene**	**基因组注释**
Refseq	-	12771	3747	43297
PUT	12434	-	5524	45465
UniGene	10913	10643	-	45282
基因组注释	6784	6391	2137	-

注：本表中 PUT 和 UniGene 均为可比对到基因组上的 PUT 和 UniGene。

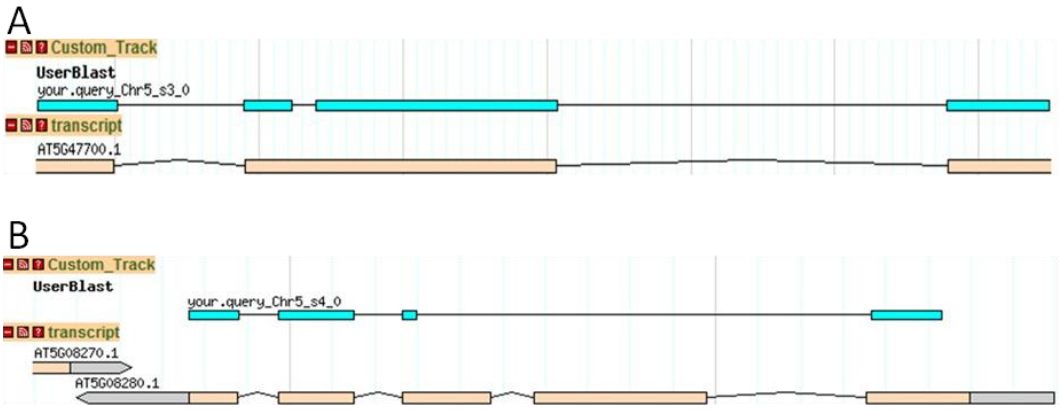

图 2-1 PUT 和 UniGene 可提供新的可变剪接体。（A）中的可变剪接模式为内含子保留，（B）中的可变剪接模式为可变的 5’和嵌套外显子。

除基因组测序已完成的植物外，我们还选取了重要的经济作物（如小麦、大麦、小米等）和代表性植物（如裸子植物）共 49 个物种用于转录因子的系统识别。所用数据源及注释版本见表 2-2。

表 2-2　整合蛋白组所用数据源及注释版本

物种	基因组注释（参考文献）	RefSeq	UniGene	PlantGDB
Arabidopsis lyrata	JGI, v1.0	-	-	-
Arabidopsis thaliana	TAIR, v9	2010-3-14	Build #70	169a
Brachypodium distachyon	JGI, v1.0	-	-	157a
Chlorella sp. NC64A	JGI, v1.0	-	-	-
Carica papaya	ASGPD	-	-	167a
Chlamydomonas reinhardtii	JGI, v4.0	2010-3-14	Build #26	163a
Cucumis sativus	JGI, v1.0	-	-	157a
Coccomyxa sp. C-169	JGI, v1.0	-	-	-
Glycine max	JGI, v1.01	-	Build #38	169a
Lotus japonicas	Kazusa	-	Build #6	171a
Manihot esculenta	JGI, v1.1	-	-	165a
Mimulus guttatus	JGI, v1.0	-	-	173a
Micromonas pusilla CCMP1545	JGI, v2.0	-	-	-
Micromonas sp. RCC299	JGI, v3.0	2010-3-14	-	-
Medicago truncatula	MGSC, v3.0	-	Build #35	169a
Ostreococcus lucimarinus CCE9901	JGI, v2.0	2010-3-14	-	161a
Ostreococcus sp. RCC809	JGI, v2.0 [$]	-	-	-
Oryza sativa indica	RIS, glean	-	Build #80	171a
Oryza sativa japonica	MSU v6.1	2010-3-14	Build #80	163a
Ostreococcus tauri	JGI, v2.0	-	-	-
Physcomitrella patens subsp. patens	JGI, v1.1	2010-3-14	Build #15	169a
Prunus persica	IPGI, v1.0	-	Build #6	161a
Populus trichocarpa	JGI, v2.0	-	Build #9	157a
Ricinus communis	JCVI, v0.1	2010-3-14	-	163a
Sorghum bicolor	JGI, v1.4	2010-3-14	Build #28	157a
Selaginella moellendorffii	JGI,v1.0	-	-	-
Volvox carteri	JGI, v2.0	-	-	-
Vitis vinifera	Genoscope	2010-3-14	Build #10	169a
Zea mays	maizesequence, 4a.53	2010-3-14	Build #77	171a
Picea glauca	-	-	Build #13	175a
Picea sitchensis	-	-	Build #14	175a
Pinus taeda	-	-	Build #11	175a
Arachis hypogaea	-	-	Build #1	171a
Artemisia annua	-	-	Build #3	177a
Brassica napus	-	-	Build #18	173a
Brassica rapa	-	-	Build #5	171a
Citrus sinensis	-	-	Build #11	167a
Gossypium hirsutum	-	-	Build #10	165a
Helianthus annuus	-	-	Build #9	169a
Malus x domestica	-	-	Build #8	173a
Nicotiana tabacum	-	-	Build #12	173a
Raphanus sativus	-	-	Build #3	165a
Solanum lycopersicum	-	-	Build #36	171a
Solanum tuberosum	-	-	Build #34	157a
Theobroma cacao	-	-	Build #1	169a
Vigna unguiculata	-	-	Build #2	167a
Hordeum vulgare	-	-	Build #56	169a
Panicum virgatum	-	-	Build #1	169a
Saccharum officinarum	-	-	Build #14	157a
Triticum aestivum	-	-	Build #56	163c

2.2.2　数据整合流程

结合数据源特点和比较的结果，我们设计了两套数据整合流程分别用于具有基因组序列和没有基因组序列的物种。

对于基因组测序已完成的物种，基因组注释的基因是蛋白组主要来源，RefSeq 的基因注释以及 PlantGDB 和 UniGene 的表达数据则作为基因组注释的有效补充

（图 2-2）。整合蛋白组的具体流程如下：

1. 在基因组注释中存在一些蛋白内部包含终止密码子，这些基因可能是假基因或者注释错误，所以首先将这部分蛋白过滤掉。然后将过滤后的基因组注释蛋白与 Refseq 注释蛋白（如果存在）合并，使用 MD5 算法(Rivest, 1992)合并序列完全一致的蛋白，所得到的蛋白数据集称为 RGset。
2. 使用 ESTScan (Iseli, *et al.*, 1999)预测 PlantGDB 拼装的相应物种的 PUT 和 UniGene 收录的 EST 序列的编码区（Coding Sequence，CDS）及相应的蛋白序列。只保留 CDS 长度大于 150、ESTScan 得分大于 200 的序列用于接下来的分析。
3. 使用 BLAT (Kent, 2002)将 CDS 比对到基因组上，要求和基因组的一致性大于 95%，比对上的 CDS 长度占全长比例大于 90%，而且 CDS 中不能有序列的删除。去掉在此标准下无法比对到基因组上的序列。
4. 使用 RGset 除去步骤 3 所得 CDS 编码蛋白中与之冗余的蛋白，并过滤低质量的蛋白。使用 BlastClust 合并冗余的蛋白（序列一致性大于 0.95、重叠的比例大于 0.9），得到的蛋白集称为 PUset。
5. 合并 RGset 和 PUset 得到一个完整的蛋白组。

对于基因组测序尚未完成的物种，使用 PlantGDB 拼装的 PUT 和 UniGene 收录的 EST 数据构建统一的蛋白组用于转录因子的识别（图 2-3）。具体流程如下：

1. 使用 ESTScan 识别 PlantGDB PUT 和 UniGene EST 中的 CDS，保留 CDS 长度大于 150，ESTScan 得分大于 200 的序列。
2. 过滤低质量的蛋白。
3. 使用 BlastClust 合并冗余的蛋白（序列一致性大于 0.95、重叠的比例大于 0.9），得到的蛋白集称为 PUset。PUset 即为尚未完成基因组测序物种的蛋白组。

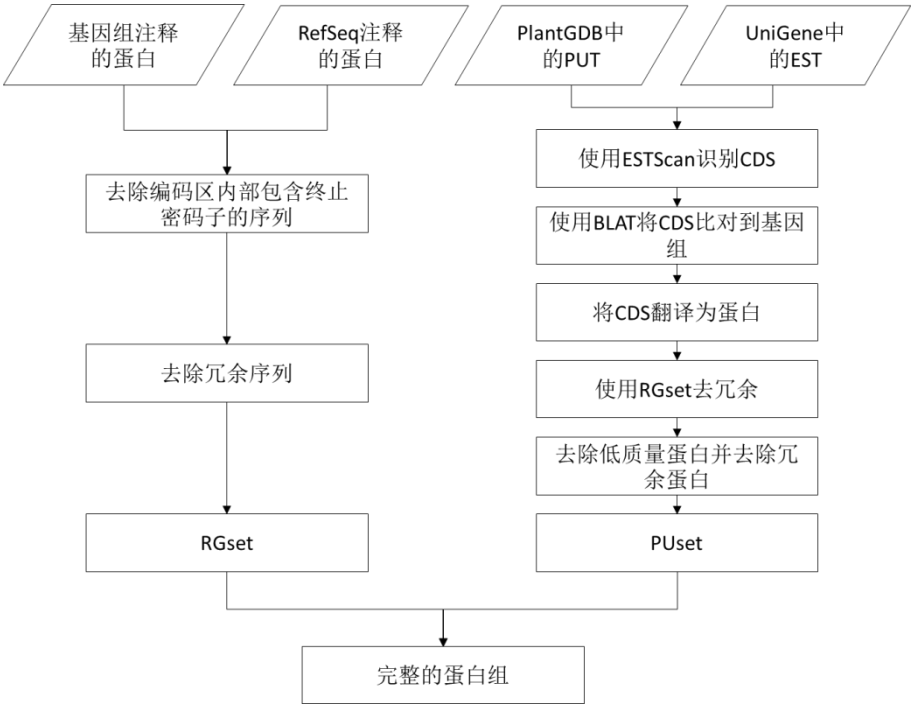

图 2-2　已完成基因组测序物种的蛋白组整合流程

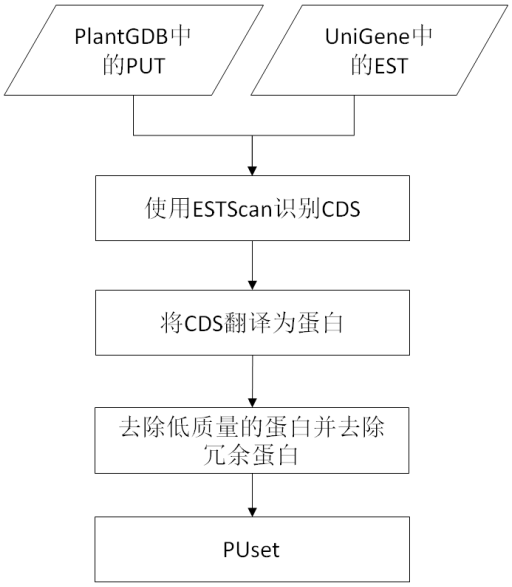

图 2-3　基因组测序尚未完成物种的蛋白组整合流程

2.2.3　数据整合结果

使用上述数据整合流程，我们构建了 49 个植物物种的蛋白组（表 2-3）用于植物转录因子的系统识别。在这 49 个物种中，28 个具有基因组序列，10 个物种具有 Refseq 注释，40 个物种使用了 PUT 数据，33 个物种使用了 UniGene 数据。各物种蛋白组包含的蛋白数及数据来源见表 2-3。在有基因组注释的物种中，

PlantGDB 和 UniGene 平均贡献了 3.7%的蛋白，可见整合这些数据源是必要的。

表 2-3 各物种蛋白组包含的序列数及数据来源

物种	总数	RGset	基因组注释	RefSeq	PUset	PlantGDB 中的 PUT	UniGene 中的 EST
Arabidopsis lyrata	32233	32233	32670	0	0	0	0
Arabidopsis thaliana	32125	31221	33410	33200	904	722	399
Brachypodium distachyon	30726	30159	32253	0	567	572	0
Chlorella sp. NC64A	9762	9762	9791	0	0	0	0
Carica papaya	27829	26954	27082	0	875	927	0
Chlamydomonas reinhardtii	23042	22201	16709	14412	841	847	164
Cucumis sativus	27725	27652	32527	0	73	74	0
Coccomyxa sp. C-169	9900	9900	9994	0	0	0	0
Glycine max	48707	45993	46244	0	2714	2479	983
Lotus japonicas	27974	26381	26700	0	1593	1501	344
Manihot esculenta	46478	45504	47163	0	974	989	0
Mimulus guttatus	27989	26760	27504	0	1229	1287	0
Micromonas pusilla CCMP1545	10518	10518	10537	0	0	0	0
Micromonas sp. RCC299	10074	10074	9891	10044	0	0	0
Medicago truncatula	52086	51172	53412	0	914	844	327
Ostreococcus lucimarinus CCE9901	7960	7676	7645	7603	284	391	0
Ostreococcus sp. RCC809	7484	7484	7492	0	0	0	0
Oryza sativa subsp. Indica	43027	40550	40745	0	2477	2539	446
Oryza sativa subsp. japonica	58760	57680	50939	26777	1080	1042	195
Ostreococcus tauri	7654	7654	7664	0	0	0	0
Physcomitrella patens subsp. patens	40604	35809	35938	35809	4795	4921	1279
Prunus persica	28299	27937	28689	0	362	350	94
Populus trichocarpa	45183	44353	45778	0	830	811	508
Ricinus communis	31953	31583	31221	31221	370	376	0
Sorghum bicolor	35810	32881	33038	32889	2929	2887	919
Selaginella moellendorffii	32969	32969	34677	0	0	0	0
Volvox carteri	15416	15416	15544	0	0	0	0
Vitis vinifera	47097	45883	30434	23335	1214	1228	285
Zea mays	62184	59636	53764	17706	2548	2474	1072
Arachis hypogaea	7243	0	0	0	7243	5952	5146
Artemisia annua	13062	0	0	0	13062	7336	12667
Brassica napus	30482	0	0	0	30482	35656	15087
Brassica rapa	14313	0	0	0	14313	13368	9702
Citrus sinensis	13522	0	0	0	13522	11930	6559
Gossypium hirsutum	20862	0	0	0	20862	20409	13115
Helianthus annuus	8634	0	0	0	8634	7673	6087
Hordeum vulgare	24020	0	0	0	24020	22906	13669
Malus x domestica	15173	0	0	0	15173	15329	7313
Nicotiana tabacum	18898	0	0	0	18898	18846	11462
Panicum virgatum	30078	0	0	0	30078	33194	15067
Picea glauca	15376	0	0	0	15376	14977	11183
Picea sitchensis	10989	0	0	0	10989	7943	10374
Pinus taeda	13275	0	0	0	13275	12982	8138
Raphanus sativus	14799	0	0	0	14799	8426	12110
Saccharum officinarum	21172	0	0	0	21172	20925	9387
Solanum lycopersicum	15722	0	0	0	15722	14611	12919
Solanum tuberosum	17445	0	0	0	17445	16440	10418
Theobroma cacao	7493	0	0	0	7493	5900	5993
Triticum aestivum	20494	0	0	0	20494	5501	20053
Vigna unguiculata	12205	0	0	0	12205	10932	8935

可变剪接能增加转录本的可塑性和蛋白组的复杂性(Filichkin, *et al.*, 2010)。目前已知的可变剪接类型主要包含外显子嵌套（Cassette exons 或 Exon skipping）、外显子互斥（Mutually exclusive exons）、5’可变（Competing/alternative 5’end）、3’可变（Competing/alternative 3’end）、内含子保留（Retained intron）、多启动子（multiple promoters 或 alternative initiation）、多 PolyA(Mutiple/alternative polyadenylation)等 7 种类型（图 2-4）。基因组范围的研究表明，拟南芥中至少 42%的多外显子基因都具有可变剪接体(Filichkin, *et al.*, 2010)。

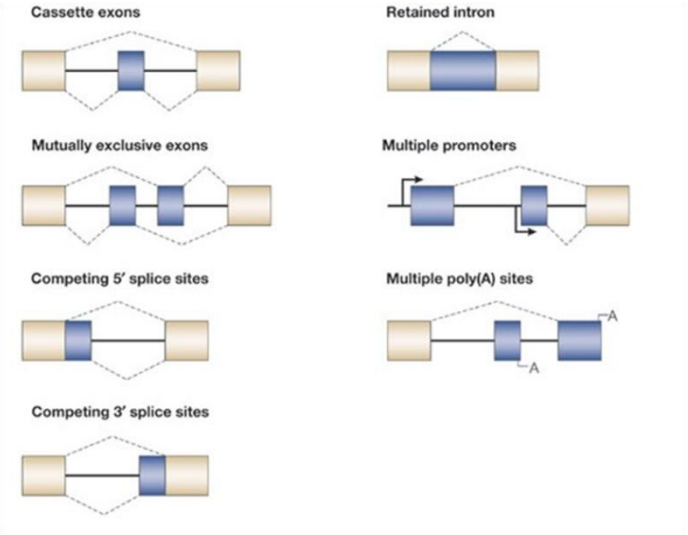

图 2-4　现有的可变剪切模式(Matlin, *et al.*, 2005)

前面提到，PlantGDB 和 UniGene 能提供一些新的基因和可变剪接体注释。现在以拟南芥的 PUset 为例研究 PUset 能补充些什么注释。在拟南芥中，PUset 包含 904 个蛋白，其中 70 个为尚未注释的基因，834 是已注释基因的可变剪接体。通过使用 GMAP (Wu, *et al.*, 2005)将它们的转录本比对到基因组上，然后使用 ASpipe (Wang, *et al.*, 2008)分析这些转录本涉及的可变剪接事件（只涉及基因内部的可变剪切，不包括多启动子和多 PloyA 这两种类型），结果如表 2-4 所示。

表 2-4 拟南芥 PUset 涉及的可变剪接事件

类型	AltA	AltD	AltP	ExonS	IntronR	总数
数目	121	80	97	192	427	917
比例（%）	13.2	8.7	10.6	20.9	46.6	-

注：AltA 为 3’不同，5’相同；AltD 为 5’不同，3’相同；AltP 为 5’, 3’均不同；ExonS 为外显子嵌套；IntronR 为内含子保留。

2.3转录因子家族分类规则的优化

2.3.1 收录转录因子家族的调整

在本研究中采纳了一个较为严格的转录因子定义，即转录因子是“能序列特异性的结合 DNA 并激活和/或抑制基因转录的蛋白”(Riechmann, 2006)。通过浏览 7000 余篇植物转录因子相关的文献，我们对转录因子家族做了较大的调整（表 2-5）。一方面，去掉了 17 个不符合转录因子定义的家族，比如转录辅助因子和染色质相关蛋白。转录辅助因子虽能调控转录，但是它并没有结合 DNA 的结构域，自己本身不能结合 DNA，故不属于上述严格定义下的转录因子。诸如染色质重构蛋白、组蛋白去甲基化酶、DNA 甲基化转移酶、组蛋白乙酰化转移酶等染色质相关蛋白虽能结合 DNA，但是这种结合并不是序列特异性的，因此也不属于转录因子的范畴。根据最新实验证据，以前认为的转录因子家族 TUBBY-like (Carroll, *et al.*, 2004) 和 Alfin-like (Lee, *et al.*, 2009)其实不是或者没有足够的证据证明是转录因子，所以也将这些家族去掉了。另一方面，添加了 5 个近几年新发现或定义的转录因子家族（DBB、FAR1、LSD、NF-X1 和 STAT）。由于结构域组成的不同，AP2/ERF 和 HB 内部的成员在 DNA 结合能力和功能上有很大差别，参考相关文献按结构域组成的不同对其进行了细分。研究报道 MADS 家族的一些 M-type 基因可能是假基因或者一种新类型的转座子元件(Riechmann, 2006)，所以我们对 MADS 家族中的 M-type 和 MIKC 做了区分。此外，我们根据文献中的习惯用法更改了 5 个家族的名称。最后，共得到 58 个植物转录因子家族。

表 2-5　收录转录因子家族的调整

调整类型		家族
去掉	转录辅助因子	AUX/IAA、GIF、TAZ、LUG、MBF1
	染色质相关	ARID、HMG、JUMONJI、PcG、PLATZ、ULT、PHD
	其它	FHA、LIM、ZIM、TUB、Alfin-like
新加		DBB、FAR1、LSD、NF-X1、STAT
细分		AP2/ERF→ RAV、AP2、ERF
		HB → HD-ZIP、TALE、WOX、HB-PHD、HB-other
		MADS → M type、MIKC
重命名		ABI3/VP1 → B3
		CCAAT-HAP2 → NF-YA
		CCAAT-HAP3、CCAAT-Dr1 → NF-YB
		CCAAT-HAP5 → NF-YC

2.3.2　转录因子分类规则

转录因子通过 DNA 结合结构域来结合 DNA 元件并调控转录，其 DNA 结合结构域在演化上很保守(Riechmann, *et al.*, 2000)。DNA 结合结构域的功能特点和其序列上的保守性成为我们识别转录因子并将其划分到不同家族的依据。参阅文献，我们总结了这 58 个家族的结构域特征和分类规则（图 2-5）。

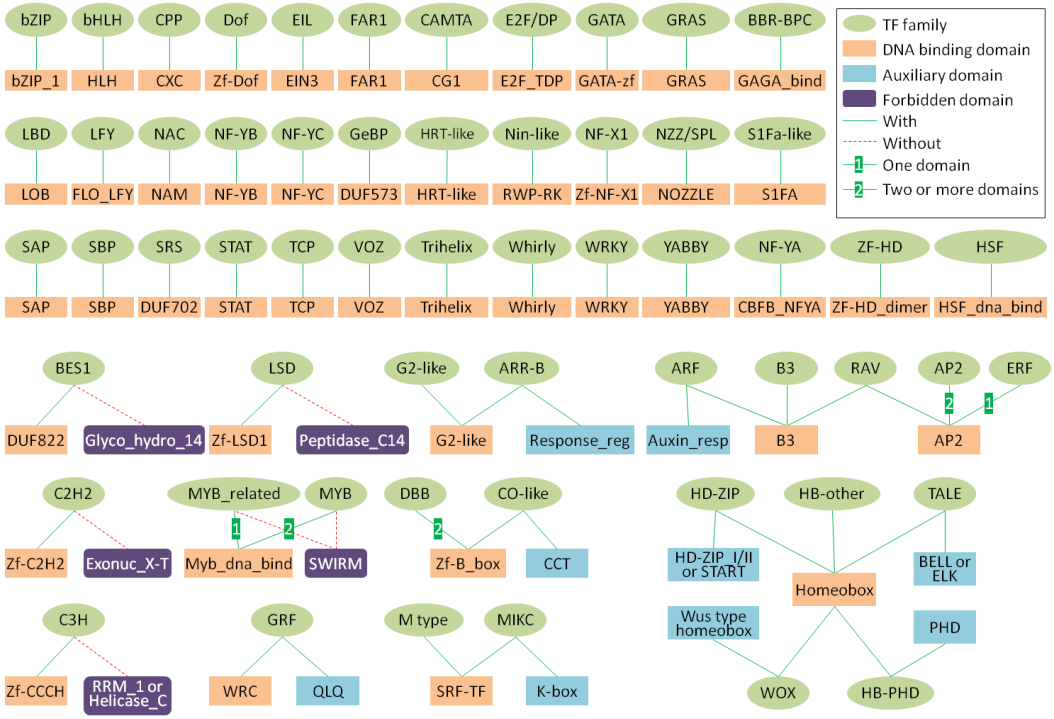

图 2-5　转录因子家族分类规则。图中绿色椭圆表示转录因子家族，红色矩形表示 DNA 结合结构域，蓝色矩形表示辅助分类的结构域，紫色矩形表示不能含有的结构域。绿线表示某家族应包含该结构域，红色虚线表示某家族不应包含该结构域。绿色线上面的数字表示包含该结构域的数目。

如图 2-5 所示，我们在识别转录因子并将其划分到不同的家族中使用了三类结构域，即 DNA 结合结构域（DNA-binding domain， DBD）、辅助结构域（Auxiliary domain）和禁止出现的结构域(Forbidden domain)。在通常情况下，仅依靠 DNA 结合结构域就可以将转录因子划分到特定家族。在某些情况下则需要借助辅助结构域将转录因子划分到特定家族。例如，如果一个蛋白包含 DNA 结合结构域“B3”，我们只能将它指定到 B3 超家族。B3 超家族包含两个家族 ARF 和 B3，此时需要借助辅助结构域“Auxin-response”来区分它们。如果这个蛋白包含“Auxin-response”则属于 ARF 家族，否则属于 B3 家族。由于并非所有包含 DNA 结合结构域的蛋白都是转录因子，因此我们定义了一种新的结构域类型“禁止出现的结构域”来去掉包含 DNA 结合结构域但不具有转录因子活性的蛋白，从而降低预测的假阳性。例如包含 DNA 结合结构域“Zf-LSD”的蛋白 AtMC2 和 MCP1B 是半胱氨酸型的肽段内切酶，没有任何证据表明它们属于转录因子，所以使用禁止出现的结构域“Peptidase_C14”过滤掉混入 LSD 家族中的此类蛋白。

2.3.3 结构域模型的构建及阈值确定

本研究中，使用 HMMER 3.0 识别蛋白中的结构域。为了提高预测的准确性，需要构建合适的隐马尔科夫（HMM）模型和选择恰当的阈值。在上述转录因子家族分类规则中，总 64 个 HMM 模型，其中 53 个取自 Pfam 24.0 (Finn, *et al.*, 2010)，另外 11 个是通过自己收集的序列构建的（图 2-6）。

表 2-6 用于识别转录因子并划分家族的模型统计

结构域类型	总数	Pfam (v24)	自建
DNA 结合结构域	47	39	8
辅助结构域	11	8	3
禁止出现的结构域	6	6	0
总数	64	53	11

E 值(E value)本身依赖于数据库的大小，一个统一的 E 值未必适合所有的结构域模型，因此我们和 Pfam 一样使用得分代替 E 值作为结构域的阈值(Punta, *et al.*, 2012)。参考 Pfam，判断一条序列是否包含某个结构域需要达到两个阈值，即序列阈值和结构域阈值(Finn, *et al.*, 2010)。辅助结构域的阈值（不包括自建的）和禁止结构域的阈值直接取自 Pfam 24.0。 DNA 结合结构域和自建辅助结构域的阈值是按照下面的方法确定的：

1. 使用 GO 注释（不包括电子注释，即证据类型为 IEA 的注释）初步确定结构域模型的阈值，包含 TC（Trusted Cutoff）和 NC（Noise Cutoff）。其中，TC 为包含该结构域并具有转录因子活性的蛋白的最低得分，NC 为不含有该结构域或者没有转录因子活性的蛋白的最高得分。
2. 对于 GO 注释未能提供足够信息的结构域，参考 Pfam 的阈值（TC 和 NC）来调整相应结构域的 TC 和 NC。
3. 使用 TAIR 注释和 Uniprot 注释进一步优化 TC 和 NC。
4. 通过人工检查序列与 HMM 模型的比对进一步优化 TC 和 NC，并在 TC 和 NC 之间选择一个合理的分数作为阈值。

取自 Pfam 的 HMM 模型和自建 HMM 模型中 DNA 结合结构域的阈值见表 2-7 和 2-8。

表 2-7 取自 Pfam 的 HMM 模型的阈值

DNA 结合结构域	结构域阈值	序列阈值
AP2	21.5	21.5
B3	30.2	30.2
CBFB_NFYA	22	22
CG-1	25	25
CXC	20.9	20.9
DUF260	22	22
DUF573	29.5	29.5
DUF702	21.3	21.3
DUF822	25	25
E2F_TDP	23	23
EIN3	22	22
FAR1	21.3	21.3
FLO_LFY	20.7	20.7
GAGA_bind	23	23
GATA	22	22
GRAS	20.5	20.5
HLH	11	11
HSF_DNA-bind	22.5	22.5
Homeobox	15	14
Myb_DNA-binding	21.5	21.5
NAM	21.5	21.5
NOZZLE	20.2	20.2
RWP-RK	20.8	20.8
S1FA	22	22
SBP	23	23
SRF-TF	21	21

TCP	21	21
WRC	21.3	21.3
WRKY	22	22
Whirly	20.2	20.2
YABBY	21	21
ZF-HD_dimer	20.5	20.5
bZIP_1	19.7	19.7
zf-B_box	16	16
zf-C2H2	12.2	11
zf-CCCH	17	17
zf-Dof	26	24
zf-LSD1	20.2	20.2
zf-NF-X1	16	16

表 2-8 自建 HMM 模型的阈值

DNA 结合结构域	**结构域阈值**	**序列阈值**
G2-like	20	20
HRT-like	24	24
NF-YB	27	27
NF-YC	24	22.9
SAP	50	50
STAT	150	150
VOZ	100	100
Trihelix	16	16

2.4植物转录因子数据库 PlantTFDB 2.0

2.4.1 PlantTFDB 2.0 的构建

基于整合的蛋白组和总结的转录因子分类规则，我们从 49 个物种系统识别出 53574 个转录因子，分属 58 个家族。对于识别的每个转录因子和每个物种的每个家族，我们都做了详尽的注释，并构建了转录因子数据库 PlantTFDB 2.0（图 2-6）。为方便用户使用，我们为所有物种设计了统一的数据库结构和用户界面。如图 2-6 所示，用户可以按物种和家族来浏览数据库。此外，还搭建了 Web Service 供用户通过程序方便的查询和批量下载数据库中的数据。

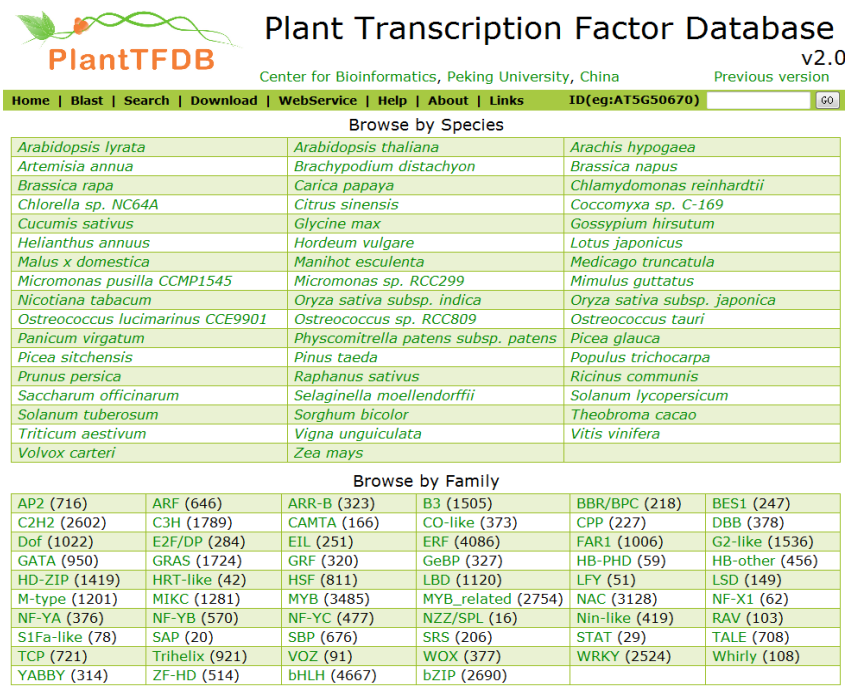

图 2-6　植物转录因子数据库 PlantTFDB 2.0 首页

基因组测序已完成的 28 个物种和基因组测序尚未完成的 21 个物种的转录因子信息统计分别见表 2-9 和 2-10。由于 PlantTFDB 2.0 收录的 49 个物种包含 9 个绿藻、1 个苔藓、1 个蕨类、3 个裸子植物和 35 个被子植物，覆盖了植物的各大分支，为研究植物转录因子的演化提供了很好的素材。从表 2-9 中可以看出，与绿藻相比，陆生植物无论在转录因子家族数、转录因子数还是转录因子占基因组基因的比例上都有明显的提高，预示着植物在登陆过程中重构了一个更加复杂的转录调控系统来调控复杂的多细胞生长发育过程和适应截然不同的陆生环境。

表 2-9　基因组测序已完成物种的转录因子统计

类群	物种	中文名	蛋白	转录因子	%	家族	OG*	TFOG*
单子叶植物	*Brachypodium distachyon*	二穗短柄草	30,726	1,687	5.49	56	1,016	1,271
	Oryza sativa subsp. indica	籼稻	43,027	1,936	4.50	56	1,427	1,692
	Oryza sativa subsp. japonica	粳稻	58,760	2,424	4.13	56	1,422	1,636
	Sorghum bicolor	高粱	35,810	1,819	5.08	54	1,252	1,583
	Zea mays	玉米	62,184	3,355	5.40	56	1,208	1,762
双子叶植物	*Arabidopsis lyrata*	琴叶鼠耳芥	32,233	1,729	5.36	58	1,298	1,604
	Arabidopsis thaliana	拟南芥	32,125	2,016	6.28	58	1,297	1,609
	Carica papaya	番木瓜	27,829	1,387	4.98	58	881	1,203
	Cucumis sativus	黄瓜	27,725	1,769	6.38	57	894	1,153
	Glycine max	大豆	48,707	3,546	7.28	57	1,148	3,057
	Lotus japonicus	百脉根	27,974	1,275	4.56	56	752	986
	Manihot esculenta	木薯	46,478	2,201	4.74	58	1,084	1,922

	Medicago truncatula	苜蓿	52,086	1,605	3.08	56	823	1,272
	Mimulus guttatus	猴面花	27,989	1,681	6.01	57	863	1,345
	Populus trichocarpa	毛果杨	45,183	2,585	5.72	58	1,086	2,195
	Prunus persica	桃	28,299	1,513	5.35	58	1,006	1,380
	Ricinus communis	蓖麻	31,953	1,291	4.04	57	994	1,170
	Vitis vinifera	葡萄	47,097	2,436	5.17	58	921	1,207
蕨类	*Selaginella moellendorffii*	江南卷柏	32,969	971	2.95	55	411	856
苔藓	*Physcomitrella patens subsp. patens*	小立碗藓	40,604	1,188	2.93	53	322	863
绿藻	*Chlamydomonas reinhardtii*	莱茵衣藻	23,042	224	0.97	30	123	136
	Chlorella sp. NC64A	小球藻	9,762	163	1.67	28	94	120
	Coccomyxa sp. C-169	胶球藻	9,900	123	1.24	29	82	90
	Micromonas pusilla CCMP1545	细小微胞藻	10,518	141	1.34	32	119	124
	Micromonas sp. RCC299	-	10,074	153	1.52	32	124	134
	Ostreococcus lucimarinus CCE9901	-	7,960	118	1.48	30	100	103
	Ostreococcus sp. RCC809	-	7,484	100	1.34	29	95	97
	Ostreococcus tauri	-	7,654	97	1.27	26	89	91
	Volvox carteri	团藻	15,416	168	1.09	28	125	137

* OG：包含至少两个转录因子的直系同源群的数目；TFOG：直系同源群中包含的转录因子数。

表 2-10 基因组测序尚未完成物种的转录因子统计

类群	物种名	中文名	蛋白	转录因子	%	家族
单子叶植物	*Hordeum vulgare*	大麦	24,020	778	3.24	54
	Panicum virgatum	柳枝稷	30,078	1,140	3.79	52
	Saccharum officinarum	甘蔗	21,172	671	3.17	48
	Triticum aestivum	小麦	20,494	746	3.64	53
双子叶植物	*Arachis hypogaea*	花生	7,243	219	3.02	39
	Artemisia annua	黄花蒿	13,062	514	3.94	48
	Brassica napus	油菜	30,482	1,334	4.38	53
	Brassica rapa	白菜	14,313	718	5.02	49
	Citrus sinensis	甜橙	13,522	534	3.95	46
	Gossypium hirsutum	陆地棉	20,862	1,111	5.33	50
	Helianthus annuus	向日葵	8,634	279	3.23	44
	Malus x domestica	苹果	15,173	658	4.34	51
	Nicotiana tabacum	烟草	18,898	793	4.20	52
	Raphanus sativus	萝卜	14,799	573	3.87	45
	Solanum lycopersicum	番茄	15,722	799	5.08	54
	Solanum tuberosum	马铃薯	17,445	776	4.45	52
	Theobroma cacao	可可	7,493	239	3.19	44
	Vigna unguiculata	豇豆	12,205	475	3.89	48
裸子植物	*Picea glauca*	白云杉	15,376	508	3.30	48
	Picea sitchensis	北美云杉	10,989	319	2.90	47
	Pinus taeda	火炬松	13,275	434	3.27	47

2.4.2 详尽的注释

2.4.2.1 个体水平注释

对于识别的转录因子，提供了基本信息、特征结构域、蛋白结构域、GO 注释、表达信息、直系同源群、核定位信号、相关文献及到知名数据库的跨库链接等注释（图 2-7）。下面简要介绍一下各类型注释：

- **基本信息：** 包括该转录因子所属的物种和家族、蛋白基本特性（包括长度、分子量和等电点等）、描述及该转录因子对应的基因模型。
- **特征结构域：** 特征结构域用来识别转录因子并将其划分到特定家族。本注释提供了特征结构域的位置信息和与 HMM 模型的比对。
- **蛋白结构域:** 蛋白结构域组成与其功能之间存在密切关系，因此蛋白的结构域信息可为推测其功能提供重要线索。InterPro (Hunter, *et al.*, 2012)是通过整合 Pfam、PROSITE、PRINTS 等 11 个结构域相关数据库构建而成的最全面的蛋白结构域特征数据库。它提供的工具 InterProScan (Quevillon, *et al.*, 2005)用来识别蛋白序列中的各结构域。

图 2-7 PlantTFDB 2.0 中转录因子个体水平的注释

- **GO 注释：** GO（Gene Ontology）使用特定的词汇来描述基因的功能(Ashburner, *et al.*, 2000)。GO 注释分为细胞定位（Cellular Component）、分子功能（Molecular Function）和生物过程（Biological Process）三类，其中细胞定位描述该分子所在的亚细胞结构或者大分子复合体，分子功能描述该基因在分子水平发挥的功能，生物过程描述该基因参与的生物过程。在 PlantTFDB 2.0 中，InterProScan 用来预测转录因子的 GO 注释。
- **表达信息：** 包含该转录因子 EST 的表达信息以及到 GEO (Barrett, *et al.*, 2013)、ArrayExpress (Rustici, *et al.*, 2013)和 GeneVestigator (Zimmermann, *et al.*, 2004)的链接，供用户根据表达位置、差异表达的时期或胁迫条件来推测其功能。
- **直系同源群：** 直系同源群（Orthologous group，OG）指各物种中从同一共同祖先演化而来的一群基因。直系同源基因之间通常具有类似的功能，因此可用于可推测研究尚不清楚基因的功能。本数据库中的直系同源群是使用 OrthoMCL (Li, *et al.*, 2003)基于 29 个具有基因组序列物种的蛋白比对信息构建的。
- **核定位信号：**核定位信号（Nuclear Localization Signals，NLS）是标记一个蛋白是否转运到核内的短序列。PredictNLS 用来识别转录因子中的核定位信号。
- **相关文献：**通过文本挖掘和人工校对，提供了与该转录因子相关的文献列表。通过这些文献，用户可以了解该转录因子的功能和研究进展。
- **跨库链接：**为方便用户到各大数据库查询相关信息，提供了到 Refseq、Swissprot、TrEMBL、TransFac、STRING、PDB 等知名数据库的跨库链接。

2.4.2.2 家族水平注释

为了展现各转录因子间的演化关系，我们为所有物种和单个物种的每个家族都构建了多序列比对、序列 logo 图和系统发生树（图 2-8）。其中基于序列全长的多序列比对是用 T-Coffee (Notredame, *et al.*, 2000)构建的，基于 DNA 结合结构域的多序列比对是用 HMMER 3.0 构建的。为便于用户了解各位点的保守情况，我们使用 Weblogo 3.0 (Crooks, *et al.*, 2004)绘制了多序列比对的序列 logo 图（图 2-8B）。所有物种每个家族的系统发生树是用 FastTree 2.1.3 (Price, *et al.*, 2010)构建的，单个物种每个家族的系统发生树是用 MrBayes 3.2 (Ronquist, *et al.*, 2003)构建的。

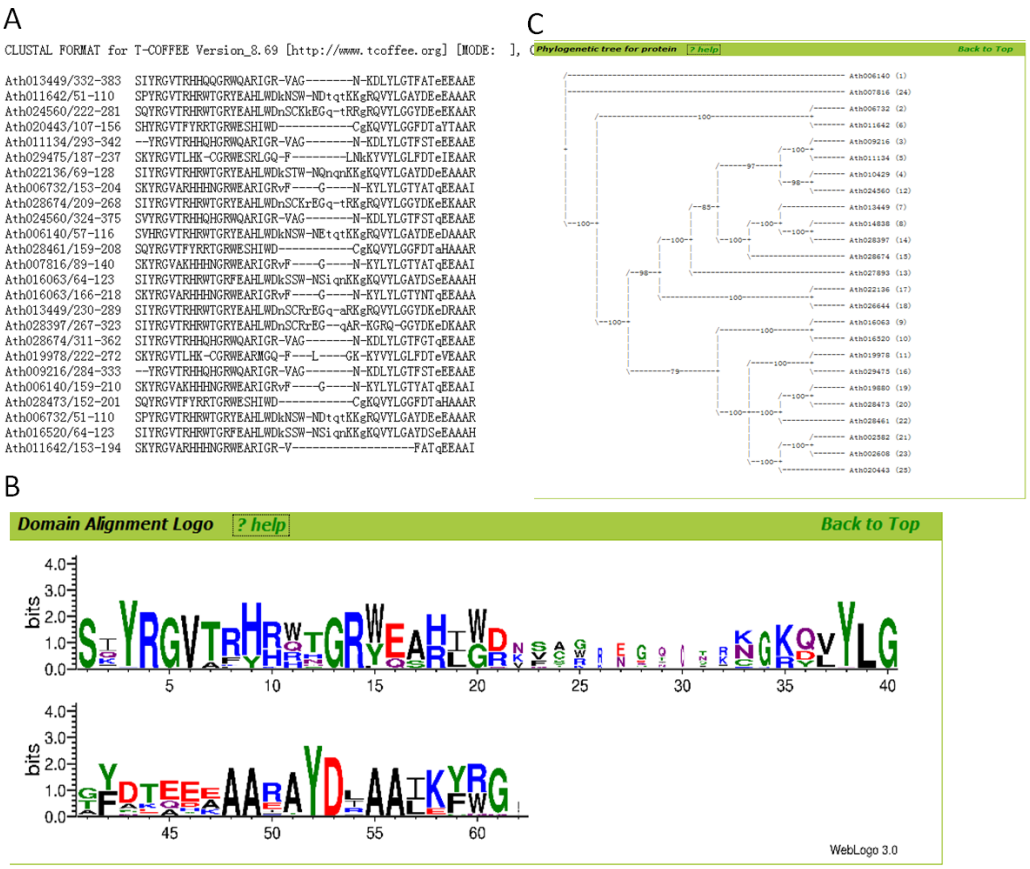

图 2-8　PlantTFDB 2.0 中家族水平的注释。A 和 B 分别为某家族转录因子 DNA 结合结构域的多序列比对及其序列 logo 图，C 为某家族的系统发生树。

2.4.3　用户界面优化

在 PlantTFDB 2.0 中，我们重新设计了数据库的用户界面。所有物种都拥有一个统一的用户界面（图 2-6），用户可以方便地浏览转录因子和家族的注释信息。除提供快速搜索外，进一步完善了高级搜索功能（图 2-9）。使用高级搜索，用户可选择搜索的范围，包括物种、家族和特定注释等。此外，还可定制显示搜索结果，并将搜索结果保存成不同格式用于后续分析。

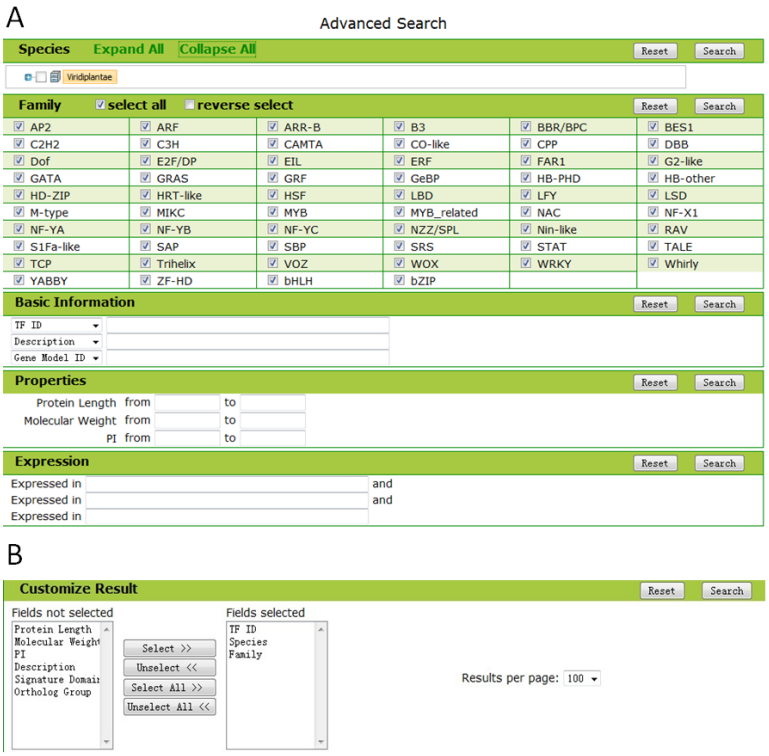

图 2-9　PlantTFDB 2.0 中的高级搜索页面

除通过界面浏览获取 PlantTFDB 2.0 中的信息外，用户还可以通过提供的 Web Service（http://planttfdb_v2.cbi.edu.cn/webservice/server.php）接口批量查询和下载数据库中的数据。为便于用户使用 PlantTFDB 2.0 提供的 Web Service，我们编写了一个 PHP 版本的示例（图 2-10）供用户下载、学习和使用。

图 2-10 PlantTFDB 2.0 提供的 Web Service 示例

2.5 PlantTFDB 3.0——植物转录因子功能和演化分析的资源平台

2.5.1 概述

在前几节中，我们详述了为什么要构建 PlantTFDB 2.0 以及如何构建的。PlantTFDB 2.0 收录了来自 49 个物种的 53319 个转录因子，其中 28 个物种具有基因组序列。自 2010 年 7 月上线以来，PlantTFDB 2.0 的访问量已达数千万次，其数据和分类规则已广泛应用于植物转录因子的功能和演化分析以及新测序物种的转录因子注释(Shulaev, *et al.*, 2010, Jia, *et al.*, 2013)。

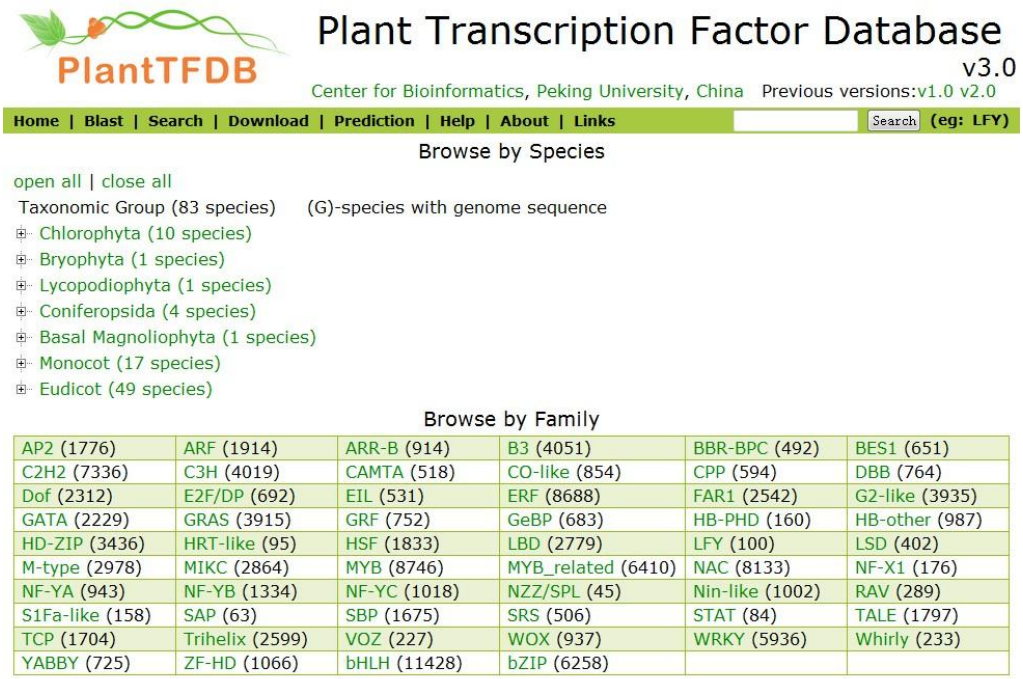

图 2-11 植物转录因子数据库 PlantTFDB 3.0 的首页

随着测序技术迅猛发展，近 3 年来又有 40 余个新的植物基因组被测定和注释（附录 1）。具有基因组的物种越来越多，特别是一个裸子植物基因组的发布(Nystedt, *et al.*, 2013)使我们有机会首次揭示覆盖绿色植物各大分支的转录因子全谱。虽然 PlantTFDB 2.0 中已有的注释类型也可以为下一步的分析提供某些线索，但是诸如功能描述、结合位点/矩阵、相互作用等信息可以提供更为直接和宝贵的证据。因此，我们有必要从各大数据库中收集此类信息进而构建一个植物转录因子的知识库。由于目前植物转录因子的研究还主要局限于模式物种，推断转录因

子的系统发生关系不但有利于研究它们的演化，也可以辅助推测功能尚未明确蛋白的功能。此外，随着我们的分类规则和流程被大家广泛认可和使用，也有必要构建一个植物转录因子预测平台供用户从自己提供的序列中识别转录因子。基于以上目的，我们将 PlantTFDB 更新到 3.0 (http://planttfdb.cbi.pku.edu.cn) (图 2-11)。

在本次更新中，共收集了 83 个物种，其中 67 个具有基因组注释（附录 1）。通过参考文献和 TAIR、UniProt 中的注释信息，对转录因子分类规则和预测流程进行了调整和优化。使用优化后的转录因子预测流程系统地从上述物种中识别转录因子，并对识别的转录因子做了详尽的注释，包括功能和演化方面的注释。此外，还搭建了一个转录因子预测平台供用户从自己提供的序列中识别转录因子（图 2-16）。

图 2-12 展示了 PlantTFDB 3.0 的主要构建流程，包括数据整合、转录因子预测流程的优化、转录因子注释以及直系同源群的构建。对于 PlantTFDB 3.0 中新添加的注释类型以下划线标注。

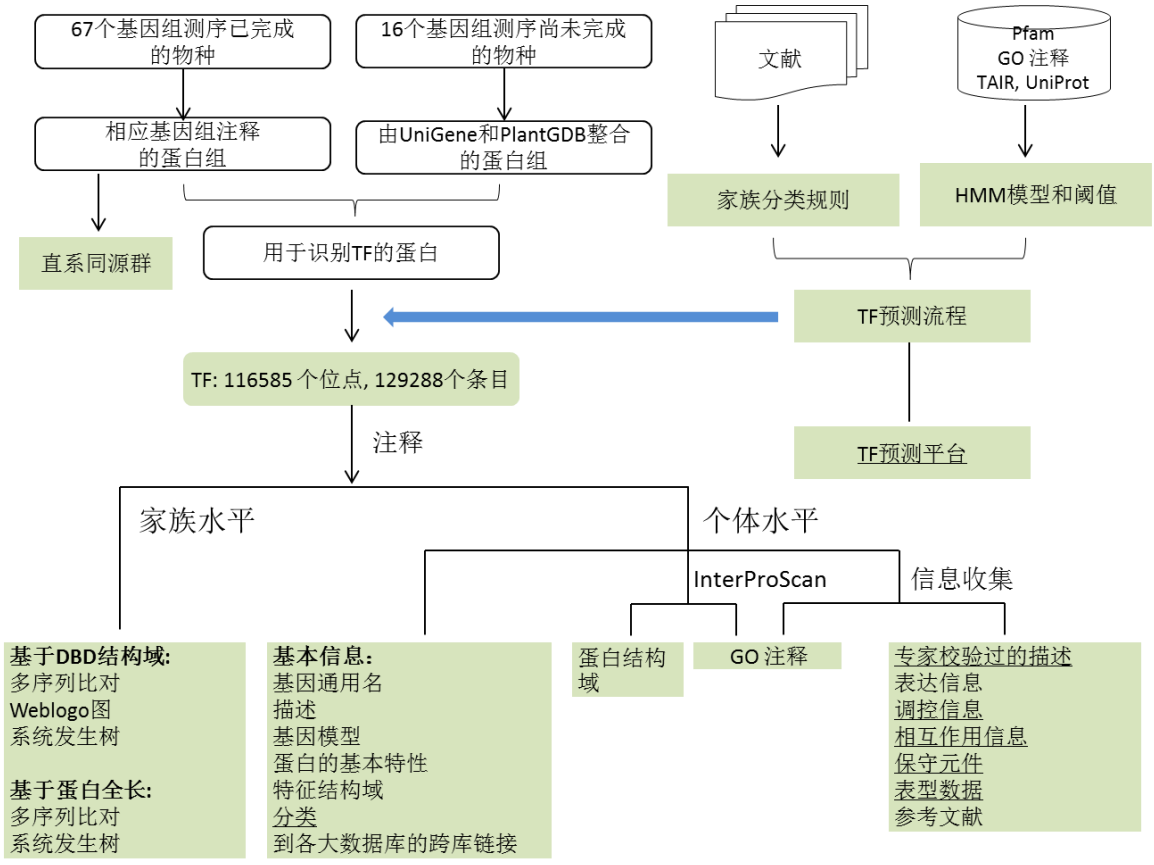

图 2-12 PlantTFDB 3.0 的构建流程图

与前两版本相比，PlantTFDB 3.0 包含更多的物种和转录因子，具有更多类型

的注释，构建了直系同源群水平的系统发生树，而且包含一个转录因子预测平台供用户从自己提供的蛋白序列中识别转录因子（表 2-11）。

表 2-11 3 个不同版本 PlantTFDB 的比较

PlantTFDB	Version 1.0	Version 2.0	Version 3.0
物种	22	49	83
基因组测序已完成的物种	5	28	67
基因组测序尚未完成的物种	17	21	16
TF 家族数	64	58	58
TF 数	26 402	53 574	129 288
注释类型			
专家校验的注释	No	No	Yes
表达信息	Yes	Yes	Yes
调控信息	No	No	Yes
相互作用信息	No	No	Yes
表型数据	No	No	Yes
参考文献	Yes	Yes	Yes
直系同源群	Yes	Yes	Yes
系统发生树			
家族水平	No	Yes	Yes
直系同源群	No	No	Yes
Web service	No	Yes	No
TF 预测平台	No	No	Yes

下面将分别从数据源、转录因子分类流程的优化、识别的转录因子及其注释、转录因子预测平台等几个方面详细介绍 PlantTFDB 3.0 的构建。

2.5.2 数据源

随着测序技术迅猛发展，在最近几年越来越多的植物基因组得以测序和注释。通过从发表的文献和各大基因组测序平台（如 JGI）收集基因组测序已完成的物种，我们共收集到 67 个物种（68 个基因组，附录 1）。与上一版本相比，具有基因组序列的物种在数目上增长了 139%。由于以 RNA-seq 为代表的表达数据已广泛应用到基因的注释中，从这版开始我们不再为已具有基因组注释的物种整合蛋白组。对这些物种而言，相应基因组测序项目注释的基因在过滤掉假基因后直接用于转录因子识别。

对于 16 个基因组序列尚未发布但在 UniGene (Sayers, *et al.*, 2011)和 PlantGDB (Duvick, *et al.*, 2008)中已有大量 EST 数据并且经济上重要的物种（表 2-12），我们

使用先前发布的流程(Zhang, *et al.*, 2011)基于以上两个数据源为其中的每一个物种都构建了一个完整的蛋白组用于转录因子的识别。

表 2-12 16 个基因组序列尚未发布的物种中用于构建蛋白组的数据来源及其版本

类群	拉丁名	常用名	UniGene	PlantGDB
裸子植物	*Picea glauca*	白云衫	Build #15	175a
	Picea sitchensis	北美云杉	Build #16	183a
	Pinus taeda	火炬松	Build #13	157a
单子叶植物	*Saccharum officinarum*	甘蔗	Build #15	157a
	Triticum aestivum	小麦	Build #63	163b
双子叶植物	*Arachis hypogaea*	花生	Build #4	171a
	Artemisia annua	青蒿	Build #6	183a
	Brassica napus	油菜	Build #19	173a
	Brassica oleracea	甘蓝	Build #5	163a
	Capsicum annuum	辣椒	Build #4	171a
	Gossypium hirsutum	陆地棉	Build #14	165a
	Helianthus annuus	向日葵	Build #12	169a
	Lactuca sativa	莴苣	Build #17	187a
	Nicotiana tabacum	烟草	Build #17	173a
	Raphanus sativus	萝卜	Build #6	187a
	Vigna unguiculata	豇豆	Build #4	167a

最后，结合这两部分数据，PlantTFDB 3.0 共收录了 83 个物种，这些物种覆盖了绿色植物的各大分支，并且每一分支都至少有一个具有基因组序列的物种（表 2-13 和附录 2）。

表 2-13 PlantTFDB 3.0 中收录的 83 个物种的门类分布

门类	基因组测序已完成	基因组测序尚未完成	总数
绿藻	10	0	10
苔藓	1	0	1
蕨类	1	0	1
裸子植物	1	3	4
被子植物			
被子植物基部物种	1	0	1
单子叶植物	15	2	17
双子叶植物	38	11	49
总数	67	16	83

2.5.3 转录因子预测流程的优化

根据特征结构域的不同（主要是 DNA 结合结构域），人们可以识别转录因子并将其划分到不同的家族(Zhang, *et al.*, 2011)。通过系统浏览转录因子相关文献，我们更新了转录因子家族分类规则（图 2-13）。最新研究表明具有“Glyco_hydro_14”结构域的 BES1 家族蛋白具有转录因子活性(Reinhold, *et al.*, 2011)，据此我们将先前 BES1 家族的这条过滤规则（forbidden domain）去掉了。为了便于显示家族间的关系，属于同一超家族的家族使用虚线圈在一起（图 2-13）。

除更新转录因子家族分类规则外，我们还从 HMM 模型和它们的阈值两个方面对转录因子预测流程进行优化。我们将从 Pfam 下载的用于识别转录因子的 HMM 模型更新到最新版本（v27.0）(Punta, *et al.*, 2012)，并根据先前发布的方法确定了这些模型的阈值(Zhang, *et al.*, 2011)。基于优化后的预测流程，我们使用 HMMER 3.0 (Eddy, 2010)系统地从 83 个物种中识别转录因子并根据上述转录因子分类规则将它们划分到不同的家族。

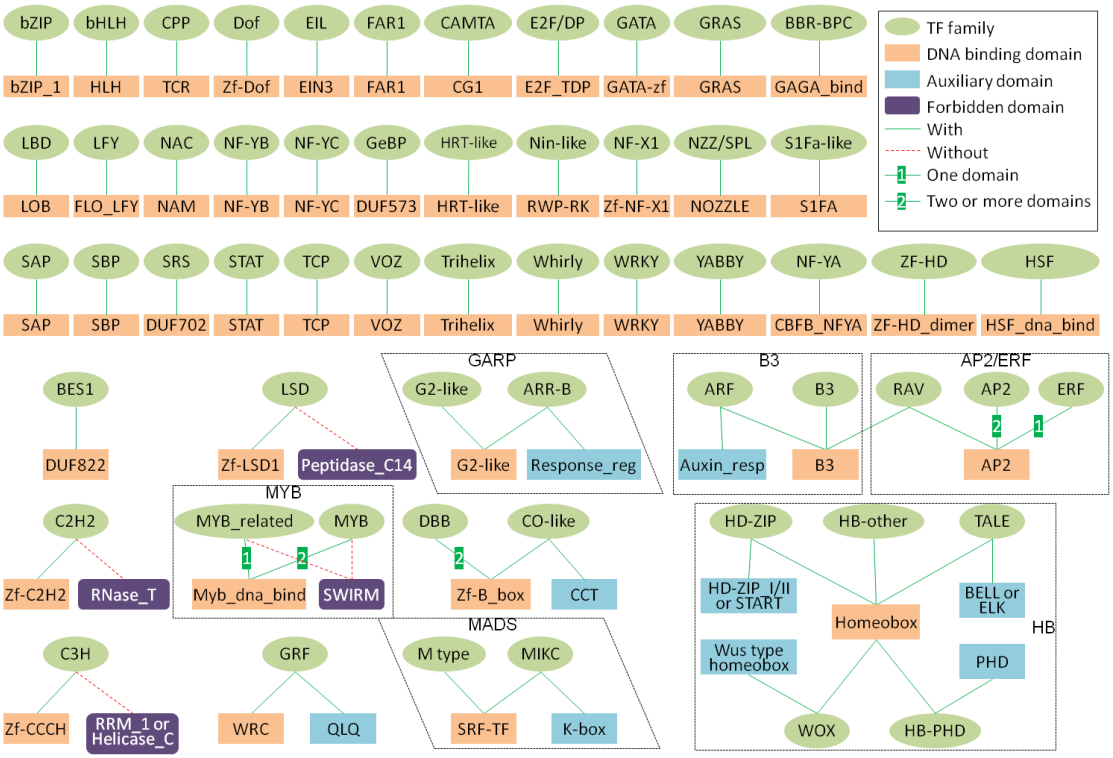

图 2-13 更新后的转录因子家族分类规则

2.5.4 横跨绿色植物各大分支的转录因子全谱

使用优化后的转录因子预测流程，我们从上述 83 个物种（共 2691496 条蛋白，

分属于 2437666 个基因）中系统识别出 129288 个转录因子（分属于 116585 个基因）(表 2-14、表 2-15、附录 2)。由于当前收录的 67 个具有基因组序列的物种已覆盖了绿色植物的各大分支（表 2-14），特别是随着一个裸子植物基因组的发布(Nystedt, *et al.*, 2013)使得我们有机会首次展示一个横跨绿色植物各大分支的转录因子全谱(表 2-14 和附录 2)。与绿藻相比，陆生植物在转录因子家族数、转录因子数和转录因子占基因组基因的比例等方面有明显的提高 (表 2-14)，可能与陆生植物具有更加复杂的转录调控系统有关(Lang, *et al.*, 2010)。

表 2-14 基于 67 个具有基因组序列的物种统计的绿色植物不同分支转录因子的平均数目

类群	物种数	基因数	TF (%)	家族数
绿藻	10	10 550	141 (1.34)	35
苔藓 [1]	1	32 273	1 079 (3.34)	53
蕨类 [2]	1	22 271	665 (2.99)	54
裸子植物 [3]	1	71 158	1 851 (2.60)	55
被子植物基部植物 [4]	1	26 846	900 (3.35)	58
单子叶植物	15	34 017	1 701 (5.00)	58
双子叶植物	38	34 798	1 861 (5.35)	58

[1] *Physcomitrella patens*, [2] *Selaginella moellendorffii*, [3] *Picea abies*, [4] *Amborella trichopoda*

表 2-15 从 16 个基因组测序尚未完成的物种中识别的转录因子统计

类群	拉丁名	蛋白数	TF	(%)	家族数
裸子植物	*Picea glauca*	16 496	559	3.39	49
	Picea sitchensis	11 351	362	3.19	48
	Pinus taeda	13 188	442	3.35	47
单子叶植物	*Saccharum officinarum*	21 082	672	3.19	48
	Triticum aestivum	56 068	1 940	3.46	56
双子叶植物	*Arachis hypogaea*	18 677	799	4.28	52
	Artemisia annua	15 732	625	3.97	49
	Brassica napus	30 365	1 343	4.42	53
	Brassica oleracea	12 061	477	3.95	51
	Capsicum annuum	19 674	922	4.69	53
	Gossypium hirsutum	21 087	1 151	5.46	50
	Helianthus annuus	8 716	288	3.30	46
	Lactuca sativa	19 676	1 036	5.27	55
	Nicotiana tabacum	19 090	820	4.30	52
	Raphanus sativus	17 565	803	4.57	49
	Vigna unguiculata	12 202	488	4.00	48

2.5.5 注释

2.5.5.1 转录因子相关知识的收集

一个好的注释应该让用户通过这些注释了解该转录因子的研究现状并为其下一步研究提供线索。虽然上一版本中已有的注释类型也可以为下一步的分析提供某些线索，但是诸如功能描述、结合位点/矩阵、相互作用等信息则可以提供更为直接的证据。致力于建成植物转录因子的知识库，我们从 TAIR (Lamesch, *et al.*, 2012)、UniProt (Consortium, 2013)等公共数据库中全面收集转录因子的相关信息，包括专家描述、表达、调控、相互作用、突变和表型等重要信息（表 2-16）。通过整合 Entrez Gene (Maglott, *et al.*, 2011)、UniProtKB (Consortium, 2013)、GeneRIF (Maglott, *et al.*, 2011)中的参考文献以及自己通过文本挖掘收集的文献，为收录的转录因子提供了相关文献列表。这些文献不但能准确反映转录因子的研究进展，也便于用户获取进一步的知识。

研究表明很多保守元件可作为转录调控元件发挥作用(Baxter, *et al.*, 2012, Haudry, *et al.*, 2013)。因此，基于基因组比对识别的保守元件也能为研究提供一些线索。目前 PlantTFDB 收集了两个不同来源的保守元件，一个是基于 9 个十字花科物种的基因组比对识别的保守元件(Haudry, *et al.*, 2013)，另一个是基于 20 个被子植物的基因组比对识别的保守元件(Hupalo, *et al.*, 2013)。拟南芥作为目前研究最为深入的植物模式物种，也是 PlantTFDB 中注释最全的物种。因此，我们使用 BLAST (Altschul, *et al.*, 1997)为其它物种的转录因子创建了到拟南芥序列相似性最高的转录因子的链接，为功能尚不明确的转录因子提供参考。

PlantTFDB 3.0 中收录的转录因子个体水平的注释如表 2-16 所示，其中新类型的注释以下划线标注。

表 3-16 PlantTFDB 3.0 中收录的转录因子个体水平注释

注释类型	物种数	TF	条目数
专家描述	22	2 128	6 649
表达信息			
UniGene	44	44 862	45 239
芯片数据	14	15 424	31 975
Plant Ontology	5	6 850	174 162
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	3 101
保守元件	2	3 709	63 859
表型信息	2	4 704	147 684
参考文献	59	5 004	20 255

下面简要介绍一下新类型的注释及其收集流程：

- **专家描述**：从 UniProt、TAIR 和 GeneRIF 中收集，主要包括功能描述、表达描述、调控描述和突变表型描述四部分。
- **Plant Ontology**：Plant Ontology(PO)是描述基因表达的时期和组织的特定词汇。在 PlantTFDB 3.0 中，拟南芥的 PO 注释是从 TAIR 中下载的，其它物种的 PO 注释来自 Plant Ontology Consortium.
- **调控信息**：包括转录因子结合位点/矩阵、ChIP-chip/ChIP-seq 实验、microRNA 调控信息和激素调控信息等。其中转录因子结合位点和矩阵是从 AthMap、AtProbe、TRANSFAC 和 JASPAR 上收集的；ChIP-chip 和 ChIP-seq 实验是从 GEO 和 SRA 上收集的；microRNA 调控信息和激素调控信息则分别是从 miRTarBase 和 AHD 中下载的。
- **相互作用信息**：包括蛋白-启动子相互作用信息和蛋白-蛋白相互作用信息，是从 BioGRID、IntAct 和 BIND 收集的。
- **保守元件**：基于多基因组比对识别的 DNA 保守元件。如上所述，PlantTFDB 收集了两个不同来源的保守元件，一个是基于 9 个十字花科物种的基因组比对识别的保守元件(Haudry, *et al.*, 2013)，另一个是基于 20 个被子植物的基因组比对识别的保守元件(Hupalo, *et al.*, 2013)。
- **表型信息**：包括突变信息和 QTL 信息。突变信息来自 UniProt、T-DNA express 和 riceGE，QTL 数据来自 Gramene。

2.5.5.2 演化层面注释

PlantTFDB 收录了 83 个物种的转录因子谱，这些数据将为分析植物转录因子的演化提供宝贵的数据资源。为了展现转录因子之间的演化关系，我们分别使用保守的DNA结合结构域和蛋白全长为所有物种的每个家族和每个物种的每个家族构建了多序列比对和系统发生树。前者是使用 FastTree (v2.1.3) (Price, *et al.*, 2010)基于 100 次抽样构建的;后者是使用 MrBayes (v3.2.1) (Ronquist, *et al.*, 2003)基于 Dayhoff 模型运行 50000 代构建的。这些系统发生树一方面展现了植物转录因子的演化概况，另一方面也可用于推测研究尚不清楚的转录因子的功能。

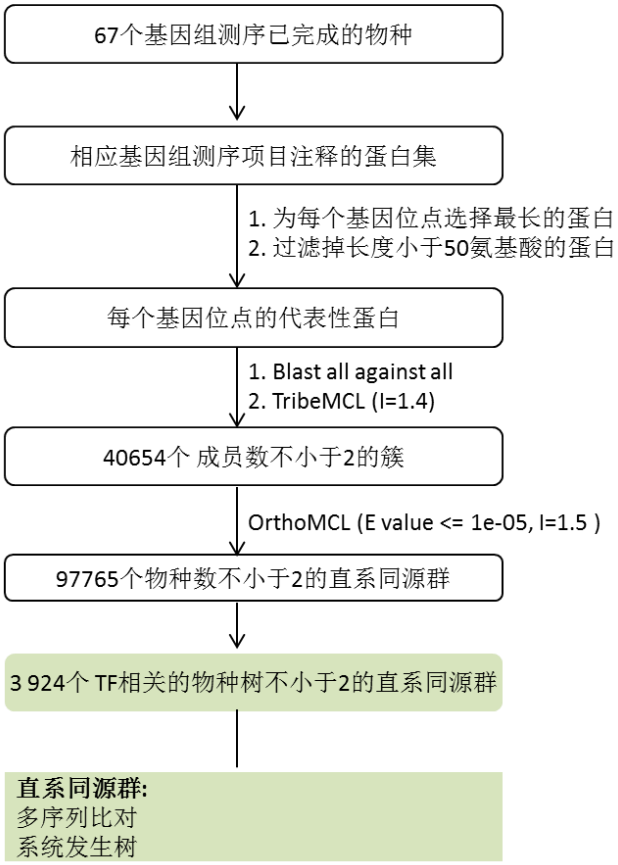

图 2-14　　直系同源群的构建流程

共同祖先基因由于物种分化事件而分离到不同物种中的基因互为直系同源。直系同源基因之间通常具有类似的功能并广泛应用于推测研究尚不明确基因的功能。直系同源群则指在多个物种中由一个共同祖先分化出来的一群基因。借鉴 Plaza (Van Bel, *et al.*, 2012)中构建直系同源群的方法，我们为 67 个具有基因组序列的物种构建了直系同源群（图 2-14）。首先每个基因位点选取最长的转录本用于接下来

的分析，并过滤掉长度小于 50AA 的蛋白；接着使用 TribeMCL (Enright, *et al.*, 2002) 将这些基因划分成不同的簇；然后使用 OrthoMCL (Li, *et al.*, 2003)在这些簇内部推测直系同源群。最后，共得到 97765 个至少包含 2 个物种的直系同源群，其中 3914 个为转录因子相关的直系同源群，包含了 69450 个转录因子（图 2-14）。

为了更好的展示直系同源群内部转录因子间的演化关系，我们使用 MrBayes 基于与上面相同的参数为每个直系同源群构建了系统发生树（图 2-15）。在这一版本中，我们在系统发生树中添加了到转录因子页面的超链接便于用户查看相关转录因子的注释信息。PlantTFDB 中所有蛋白序列、多序列比对和系统发生树都可以免费下载以供后续分析。

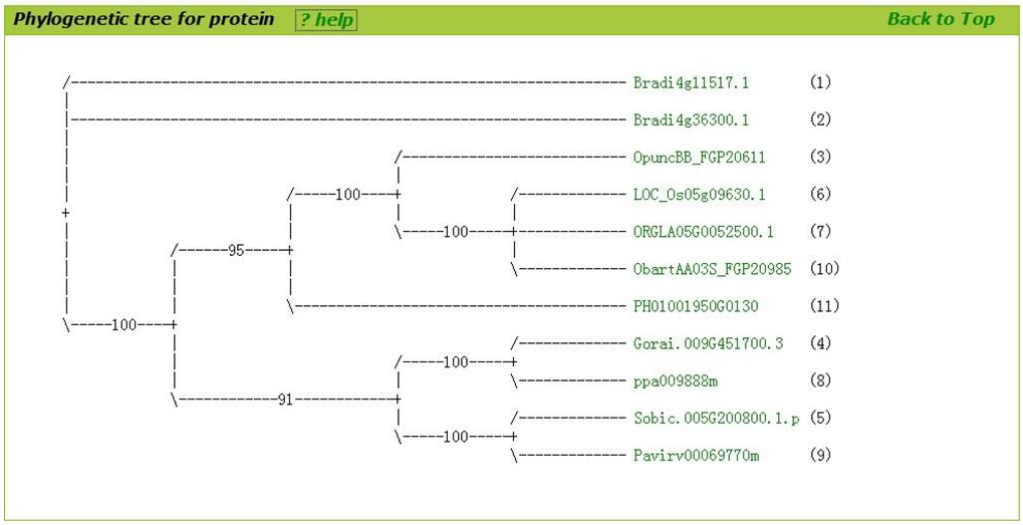

图 2-15 PlantTFDB 3.0 中的系统发生树

2.5.6 转录因子预测平台的构建

近些年，我们的转录因子预测流程已广泛应用到新测序物种的转录因子预测中 (Shulaev, *et al.*, 2010, Jia, *et al.*, 2013)。为了方便用户识别尚未收录的物种或新注释序列中的转录因子，我们基于已建立的预测流程搭建了一个植物转录因子预测平台(http://planttfdb.cbi.pku.edu.cn/prediction.php) (图 2-16)。它可以较为迅速地从多条蛋白序列中识别出其中的转录因子。由于拟南芥是 PlantTFDB 3.0 中注释信息最全的物种，用户若勾上“the best hit in *Arabidopsis thaliana*”，后台程序将把识别出来的转录因子与拟南芥的转录因子做 BLAST，并在结果中给出到序列相似性最高的拟南芥转录因子的链接和描述（图 2-17）。

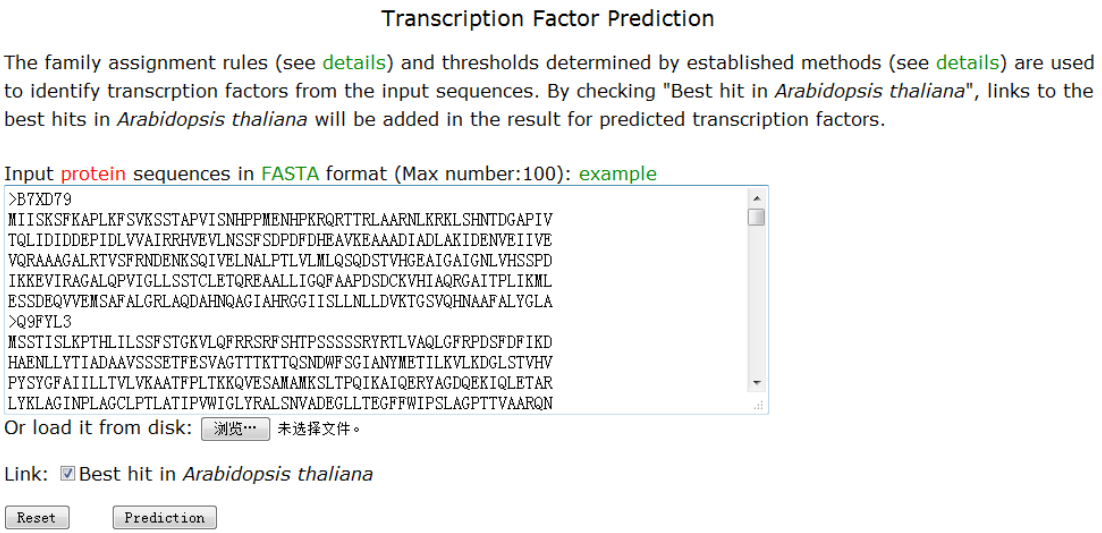

图 2-16 植物转录因子预测平台

Result

Number of input sequences: 10 (Download list)
Number of transcription factors identified from them: 2

Result of identified transcription factors: (Download list)

TF ID	Family	Best hit in *A. thaliana*	Blast e-value	Description for the best hit
C0SUU6	GeBP	AT1G11510.1	0.0	DNA-binding storekeeper protein-related transcriptional regulator
Q9S840	SBP	AT5G43270.3	0.0	squamosa promoter binding protein-like 2

图 2-17 转录因子预测结果

2.6 总结

通过整合基因组注释、Refseq、PlantGDB 和 UniGene 等四个数据源，我们为每个物种构建了一个完整的蛋白组。通过系统浏览 7000 余篇植物转录因子相关的文献，构建一套新的转录因子分类规则。通过参考 GO 注释、TAIR 和 SwissProt 等注释并经人工检查为每个 HMM 模型确定了一个合理的阈值，以提高预测的准确性。使用自己整合蛋白组和总结的转录因子分类规则，我们从 49 个物种中系统识别出 53574 个转录因子，并将它们划分到 58 个家族。为便于用户使用，为每个家族和每个转录因子都做了详尽的注释并重新设计了数据库底层结构和用户界面，构建了 PlantTFDB 2.0。该数据库覆盖了绿藻、苔藓、蕨类、裸子植物和被子植物等主要的植物分支，为相关领域工作人员研究转录因子功能和家族演化提供了宝贵的资源。此外，我们还完善了高级搜索功能并提供了 Web Service 便于用户查询和批量下载数据库中的数据。

为了向研究植物转录因子功能和演化的科学家提供更新更全的数据，我们将 PlantTFDB 更新到 3.0。使用优化后的转录因子预测流程，我们从 83 个物种中系统

识别出 129288 个转录因子。由于收录的具有基因组的 67 个物种已覆盖绿色植物的各大分支，PlantTFDB 首次提供了一个横跨绿色植物各大分支的转录因子全谱。为便于用户更好地了解转录因子的功能并为进一步研究提供线索，我们为收录的转录因子做了详尽的注释。除了上一版本已有注释类型外，我们还从公共数据库中收集了专家描述、表达信息、调控信息、相互作用信息、保守元件、表型信息以及参考文献等重要信息，为用户了解该转录因子的研究现状和制定下一步研究计划提供更直接的证据。为了展现转录因子间的演化关系，为所有物种的每个家族和每个物种的每个家族都构建了系统发生树。除此之外，使用 67 个具有基因组序列物种的蛋白组构建了直系同源群，并为每个直系同源群构建了系统发生树以便在更精细尺度展示转录因子的演化。这些系统发生树不仅能为转录因子的演化分析提供数据资源，也有助于推测研究尚不明确的转录因子的功能。此外，基于已有的预测流程搭建了一个植物转录因子预测平台供用户从自己提交的序列中识别转录因子。

目前，PlantTFDB 年访问量达千万次，已广泛应用于植物转录因子的功能和演化研究以及新测序物种的转录因子注释中。PlantTFDB 3.0 中的所有转录因子已被国际权威的数据索引数据库 Thomson Reuters Data Citation Index（http://wokinfo.com/products_tools/multidisciplinary/dci/）收录。

第 3 章 拟南芥转录调控网络

3.1概述

3.1.1 转录调控网络的重要性

转录调控网络是生物体转录调控过程的集中体现，是理解特定生物过程和生物过程之间调控机理的基础。通过转录因子之间形成的级联调控网络，植物精确地调控复杂的生长发育过程和迅速响应各种生物/非生物胁迫。例如植物通过转录因子之间的正、负反馈依次开启 A、B、C 类基因的表达从而决定花分生组织依次分化为花萼、花瓣、雄蕊和心皮（图 3-1A）(Irish, 2010)；当遇到低温和干旱等胁迫时，植物通过转录因子间形成的级联调控网络最终开启抗逆基因的表达以应对外界的不利环境（图 3-1B）(Zhang, *et al.*, 2004)。

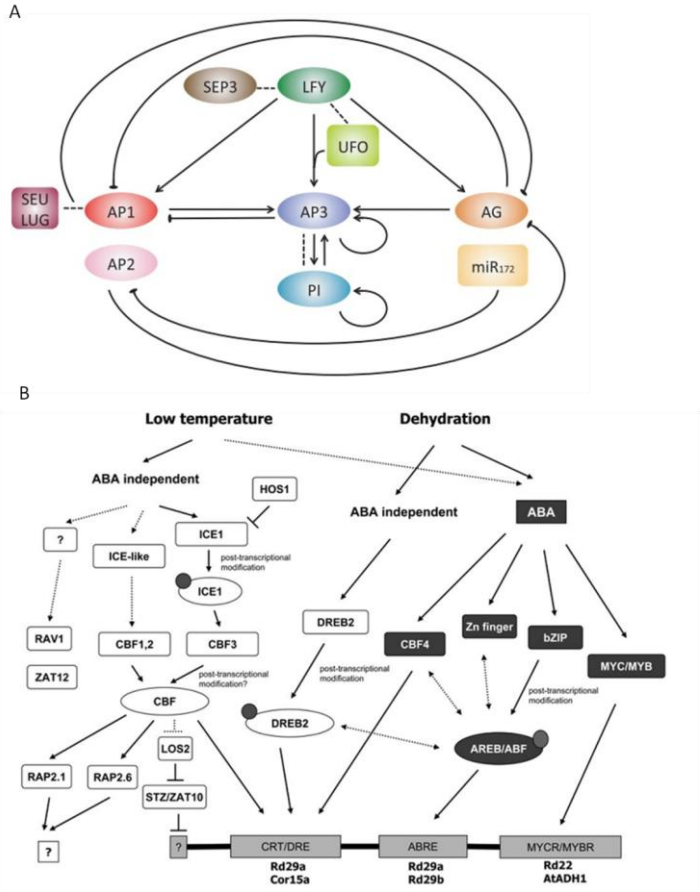

图 3-1 植物通过转录调控网络调控生长发育和应对各种胁迫。（A）为调控花分生组织分化的转录调控通路(Irish, 2010)，（B）为低温和干旱引起的级联转录调控(Zhang, *et al.*, 2004)。

一个高质量、基因组范围的转录调控网络是理解转录调控系统如何设计以及为何如此的基础。在过去的十数年，正是由于有了从文献收集的大肠杆菌(Gama-Castro, *et al.*, 2011)和酵母(Costanzo, *et al.*, 2001)的转录调控网络，人们才得以系统分析并深入理解转录调控网络在结构和演化等方面的特征，比如网络的拓扑结构(Luscombe, *et al.*, 2004)、构成元件(Shen-Orr, *et al.*, 2002, Lee, *et al.*, 2002, Alon, 2007)以及它们的演化规律(Teichmann, *et al.*, 2004, Tuch, *et al.*, 2008)等。与上述物种不同，植物演化出一个独特的转录调控系统来精确调控生长发育和快速响应外部胁迫。然而目前植物转录调控研究还主要局限于特定的生物过程，比如花器官(Lau, *et al.*, 2012, Wellmer, *et al.*, 2010, Irish, 2010)、叶(Byrne, 2005, Townsley, *et al.*, 2012)和根(Brady, *et al.*, 2011, Montiel, *et al.*, 2004, Ishida, *et al.*, 2008)的发育等。缺少一个基因组范围的植物转录调控网络阻碍我们理解植物是如何设计自己的转录调控网络以完成生长发育和应激所需。

3.1.2 收集拟南芥转录调控网络的可行性

科技文献是研究工作的总结，对科技文献进行挖掘分析也能带来新发现和新想法(Rebholz-Schuhmann, *et al.*, 2012)。拟南芥是植物研究中最重要的模式物种，数十年研究积累了大量特定生物过程内部的调控信息。这些调控信息分散在海量的科技文献中，其中大多数调控信息是由多方面证据支持的，在生物过程中发挥着某种特定的功能。如能将这些调控信息收集起来，有望构建一个高质量的拟南芥转录调控网络，进而促进我们理解植物转录调控网络的设计原则和演化特征。

在 PubMed 中，拟南芥转录调控相关的文献多达 5000 余篇。一个人如果想通过阅读每一篇文章从中提取转录调控的信息将花费太长的时间，因此需要借助某种技术手段辅助我们阅读这些文献，从而缩小阅读范围或者直接提取所需的调控信息。文本挖掘作为一种自然语言处理的技术，目前已逐渐应用于信息提取和数据库构建中(Hunter, *et al.*, 2008, Rzhetsky, *et al.*, 2008)。例如收集文献中报道的药物副作用构建的数据库 SIDER (Kuhn, *et al.*, 2010)和收集遗传变异对药物反应的影响构建的 PharmGKB (Thorn, *et al.*, 2010)等都用到了文本挖掘的技术。转录调控关系在文本中表现为两个分子实体间的相互作用，挖掘这种二元关系在技术上是相对容易实现的。目前已有多个工具如 MedScan (Novichkova, *et al.*, 2003)、iHOP (Hoffmann, *et al.*, 2005)和 Textpresso (Muller, *et al.*, 2004)等可用于挖掘此类型的关系。文本挖掘技术的进步使得我们可以在合理的时间内尽可能多的收集散布在海

量文献中的拟南芥转录调控信息。

3.2拟南芥转录调控网络的收集

3.2.1 拟南芥转录调控信息的收集流程

在收集转录调控信息之前首先要确定收集所用的数据源。ResNet plant 3.0 (Nikitin, *et al.*, 2003)从PubMed和11种植物相关的期刊全文中收集蛋白相关的关系构建而成的一个以蛋白为中心的知识库，里面包含了很多转录因子之间和转录因子与其它蛋白的作用，包括转录调控、磷酸化调控、蛋白蛋白相互作用等。PubMed则是世界上收集生物医学文献摘要最大和最全的数据库。因此，我们确定使用ResNet 3.0和PubMed作为收集转录调控信息的数据源。

在确定了数据源后，按照如下方法收集拟南芥转录调控信息（图3-2）：

1. 以PlantTFDB 2.0 (Zhang, *et al.*, 2011)中拟南芥所有转录因子作为输入，从ResNet plant 3.0中检索所有与输入转录因子相关的两个基因间的关系/作用，得到4150个转录因子相关的两基因间的关系及它们出现的文献和语句。
2. 以PlantTFDB 2.0中拟南芥所有转录因子作为输入，使用文本发掘工具MedScan (Novichkova, *et al.*, 2003)提取PubMed摘要（截止到2011年5月）中的此类关系，得到3200个转录因子相关的两基因间的关系及它们出现的文献和语句。
3. 使用Pathway Studio (Nikitin, *et al.*, 2003)合并步骤1和2中得到的关系，共得到4663个转录因子相关的两个基因间的关系。
4. 参考原文人工检查每一条关系是否为所要收集的转录调控关系。在此过程中，去掉不属于转录调控的关系，如蛋白蛋白相互作用、泛素化修饰、遗传相互作用等；去除文本发掘工具错误识别的关系和添加文本发掘工具遗漏的转录调控关系；确定调控的类型（激活和/或抑制）。
5. 去掉调控类型不明确的转录调控关系，最后得到1431个调控类型明确的转录调控关系。

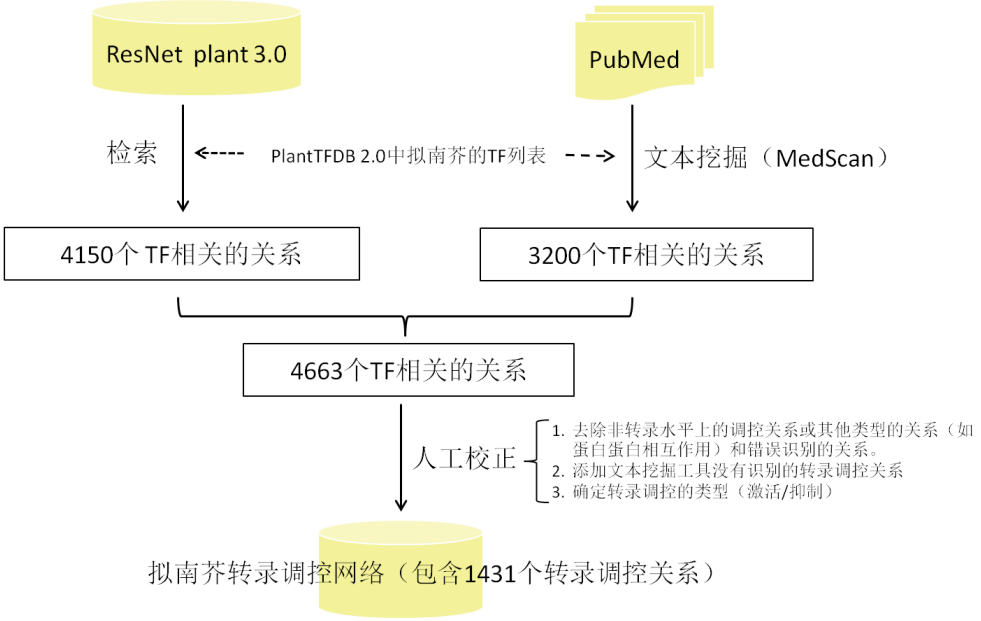

图 3-2　拟南芥转录调控网络的收集流程

3.2.2　拟南芥转录调控网络的内容和特征

使用收集到的 1431 条转录调控信息，我们构建了一个拟南芥转录调控网络（*Arabidopsis* transcriptional regulatory map, ATRM）（图 3-3）。在图 3-3 中，根据 TAIR 中具有实验证据的 GO 注释（注释版本 2012-6-5，实验证据代码为 EXP、IDA、IPI、IMP、IGI 或 IEP。若非特殊说明，本文研究中所用的 TAIR GO 注释均为此版本且有实验证据支持的 GO 注释）(Lamesch, *et al.*, 2012),我们将 ATRM 包含的基因分为 4 类：只参与发育过程的基因（红色的圆）、只参与应激（应对胁迫）的基因（绿色的圆）、同时参与发育和应激的基因（黄色的圆）及不属于以上三种类型或未注释的其它基因（黑色的圆）。红线表示转录因子 A 激活基因 B，绿线表示 A 抑制 B，蓝线则表示在某些情况下 A 激活 B 而在某些情况下 A 抑制 B。

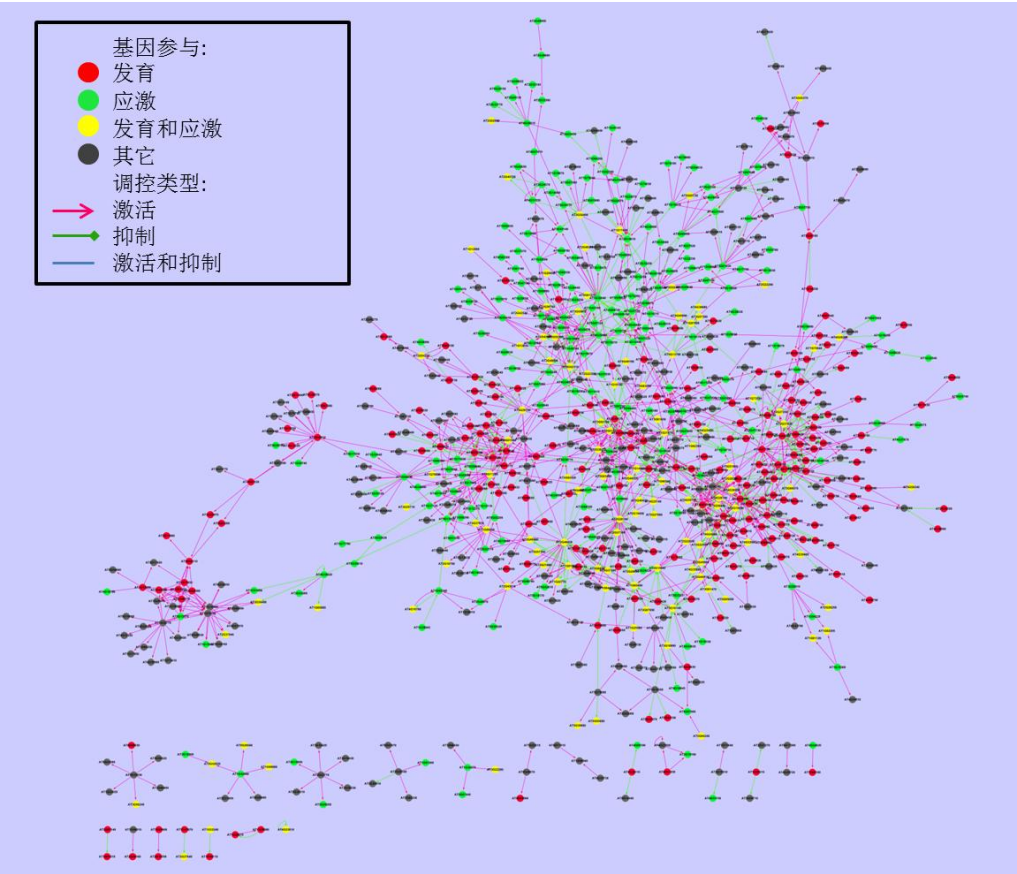

图 3-3　拟南芥转录调控网络（ATRM）

表 3-1　拟南芥转录调控网络中基因和调控的统计

类型		**数目**
基因(生物过程)	发育	222
	应激	221
	发育和应激	116
	其它	231
基因	转录因子	388
(转录因子/非转录因子)	非转录因子	402
调控类型	激活	998
	抑制	430
	激活和抑制	3

ATRM 中的调控涉及 974 篇文献，其中包含基因和调控的统计见表 3-1。ATRM 包含 47 个家族 388 个转录因子，分别覆盖了拟南芥中转录因子家族数的 81.0%和所有转录因子的 22.8%。为研究 ATRM 包含调控所在的具体生物过程，我们使用

map2slim 工具将 ATRM 中的基因映射到植物的 GOslim 上（表 3-2）。从表 3-2 可以看出，这些基因主要集中在应对生物/非生物胁迫等应激过程和多细胞器官发育、胚的发育等发育过程。

表 3-2　ATRM 中基因参与生物过程的分布情况

Biological process（生物过程）	**基因数目**
Response to stress（应激过程）	337
GO:0006950: response to stress	257
GO:0009607: response to biotic stimulus	91
GO:0009628: response to abiotic stimulus	243
Developmental process （发育过程）	338
GO:0030154: cell differentiation	76
GO:0007275: multicellular organismal development	182
GO:0009791: post-embryonic development	132
GO:0009790: embryo development	40
GO:0009653: anatomical structure morphogenesis	111
GO:0009908: flower development	116
Other（其它）	231
With biological process annotation	117
Without biological process annotation	114

通过研究复杂的社会网络和生物网络，Barabasi 等发现复杂网络中度的分布服从幂律分布(Barabasi, *et al.*, 1999)。因此，我们也查看了 ATRM 中基因入度、出度和度的分布，并使用 R 拟合一幂律函数来检验其是否遵从该分布。和前人发现一致，ATRM 在这些方面亦服从幂律分布（图 3-4）。

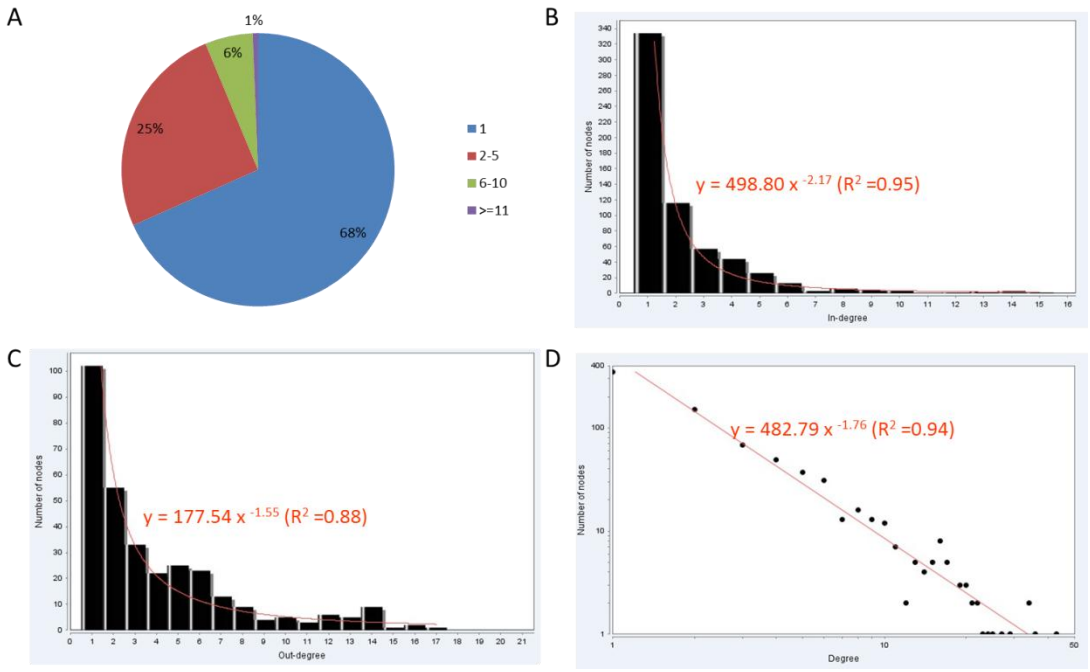

图 3-4　拟南芥转录调控网络的统计。（A）为每个调控所涉及的参考文献数目的分布情况。（B）为 ATRM 中基因入度（TFs/靶基因）的分布。（C）为 ATRM 中基因出度（靶基因/TF）的分布。（D）为 ATRM 中基因度的分布情况。

3.3ATRM 的质量评估

3.3.1　评估方法和数据

转录因子调控其靶基因的转录，转录因子和它们的靶基因应该能同时出现在某个生物过程中（共过程）而且在表达上具有较高的相关性。高质量的转录调控网络通常具有更高比例的调控能同时出现在某个生物过程中和具有高的表达相关性。因此，功能相似性和表达相关性广泛应用于网络的总体质量评估，包括蛋白蛋白相互作用网络和转录调控网络等(Wang, *et al.*, 2012, Marbach, *et al.*, 2012)。

拟南芥基因调控数据库 AGRIS 的 AtRegNet 从 76 篇文献中收集了 11227 条调控信息，其中的 769 条具有两个或两个以上的证据，这部分数据被标记为“AtRegNet(confirmed)”(Yilmaz, *et al.*, 2011)。拥有此数据，一方面可通过比较 AtRegNet 和 AtRegNet（confirmed）判断共表达和表达相关性能否用于评估转录调控网络的质量，另一方面也可作为背景衡量 ATRM 的总体质量。

3.3.2　转录调控对的共过程

在使用转录因子和靶基因的共过程评估转录调控网络的质量以前，我们还需要

构建一个相应的转录调控背景（background）做对比。转录调控的背景是按照如下方法构建的：

1. 合并 AtRegNet 和 ATRM 的 TF 列（调控关系在列表中的格式为“TF target”，第一列为 TF 列），得到一个 TF 列表 A。
2. 将 TF 列表 A 中 TF 的“生物过程”（Biological process，BP）GO 注释使用 map2slim 工具映射到植物的 GOslim 上，选择至少有 10 个基因的 GOslim 项（term），去掉其中对 TF 而言太泛的项比如“调控 RNA 合成”等得到一个 GOslim 列表（表 3-3）用于下面的分析。
3. TF 列表 A 和拟南芥基因中可以映射到（2）中选取的 GOslim 项上的转录因子和基因分别标记为“mapped TFs”和“mapped 基因”。
4. “mapped TFs”和“mapped 基因”间所有可能的调控（不包括自我调控）为下面分析中用到的转录调控背景。选取 AtRegNet 和 ATRM 中被转录调控背景包含的调控关系用于下面的质量评估。

表 3-3 选取的映射到该项的转录因子（TF）不小于 10 个的 GOslim 列表

GO term	**描述**	**TF 的数目**
GO:0007275	multicellular organismal development	185
GO:0009791	post-embryonic development	128
GO:0000003	Reproduction	118
GO:0009719	response to endogenous stimulus	108
GO:0006950	response to stress	101
GO:0009628	response to abiotic stimulus	100
GO:0009908	flower development	77
GO:0007154	cell communication	65
GO:0009653	anatomical structure morphogenesis	63
GO:0007165	signal transduction	57
GO:0030154	cell differentiation	50
GO:0009607	response to biotic stimulus	36
GO:0019748	secondary metabolic process	29
GO:0016043	cellular component organization	23
GO:0040007	growth	20
GO:0009790	embryo development	19
GO:0009605	response to external stimulus	11
GO:0007049	cell cycle	10

转录调控对只要能同时出现在以上所选取的任一生物过程，就称这对转录调控是共过程的。首先，我们比较了 AtRegNet 和 AtRegNet（confirmed）中共过程的转录调控对占总体的比例，结果显示数据质量更高的 AtRegNet（confirmed）中共

过程的转录调控对所占比例显著高于质量相对较低的 AtRegNet（单侧二项式检验 *P* <2.2e-16，图 3-5）。通过比较 ATRM 与背景、AtRegNet、AtRegNet（confirmed）中共过程的转录调控对的比例并使用二项式检验进行统计检验，结果表明 ATRM 的高质量（图 3-5）。

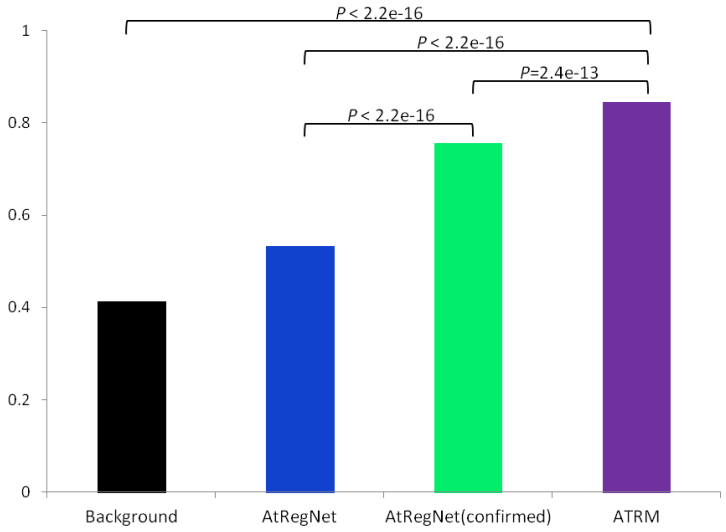

图 3-5　背景（Background）、AtRegNet、AtRegNet（confirmed）和 ATRM 中共过程的转录调控对所占的比例。线上的 *P* 值是单侧二项式检验的结果。

3.3.3　转录调控对的表达相关性

Pearson 相关系数（Pearson correlation coefficient, PCC）已广泛应用于衡量共表达基因的表达相关性。ATTED-II 提供了拟南芥各基因在表达上的 Pearson 相关系数(Obayashi, *et al.*, 2009)。ATTED-II 中基因间的 Pearson 相关系数是基于来自正常发育过程、生物和非生物胁迫、激素和光处理等 58 个实验（详见：http://atted.jp/help/experiment_GeneExp_v3.shtml ） 1388 张芯片（详见：http://atted.jp/help/slide_GeneExp_v3.shtml）的表达数据计算的。下面分析使用的 Pearson 相关系数均是从 ATTED-II 中下载的。所有 TF 列表 A 中 TF 与拟南芥基因的可能调控作为下面分析比较的背景。

首先，比较了 AtRegNet 和 AtRegNet（confirmed）中具有高表达相关性（PCC>=0.5）的调控对所占的比例，质量较高的 AtRegNet（confirmed）中高表达相关性的调控对所占的比例要明显高于质量较低的 AtRegNet（单侧二项式检验 *P* = 3.1e-05，图 3-6）。通过研究 ATRM、背景、AtRegNet 和 AtRegNet（confirmed）中转录调控对 PCC 的分布情况及高 PCC 的转录调控对所占的比例（图 3-6），也显示

了 ATRM 的高质量。

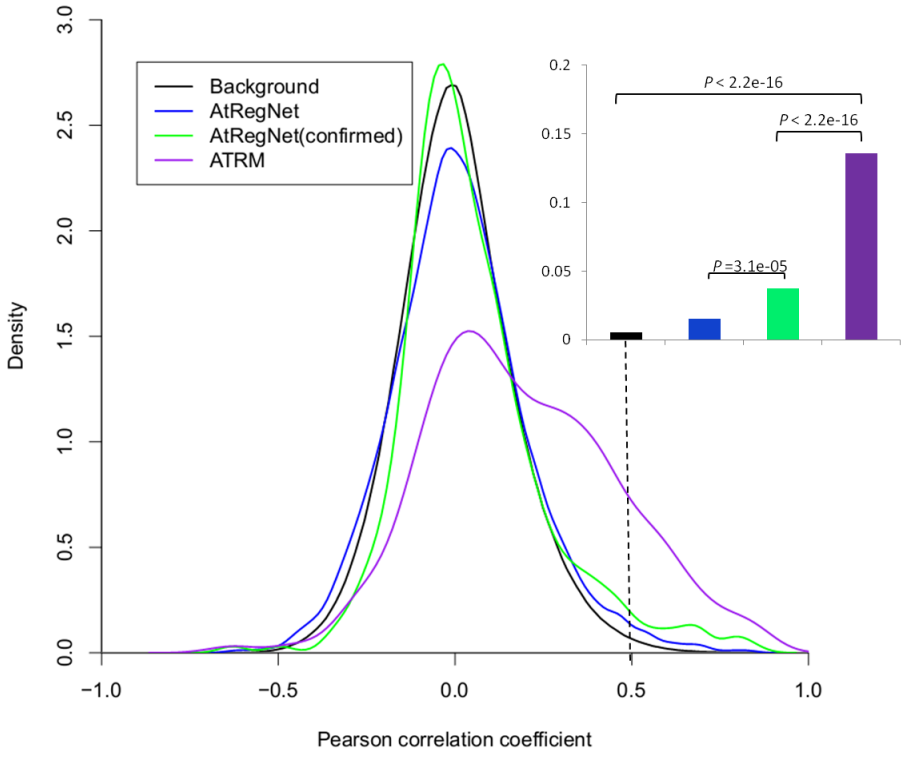

图 3-6 背景（Background）、AtRegNet、AtRegNet（confirmed）和 ATRM 中调控对之间的 Pearson 相关系数（PCC）的密度分布曲线。右上角的图为 PCC 不低于 0.5 的调控对所占的比例，线上标注的 *P* 值是单侧二项式检验的结果。

以上通过转录调控对共过程的比例和高表达相关性的比例两个方面评估了 ATRM 的质量，二者都说明 ATRM 中收录的转录调控对的质量在总体上是相对较高的。

3.4 ATRM——拟南芥转录调控过程的集中展现

转录调控网络是由生物过程内部的调控和生物过程之间调控组成的。因此，ATRM 是特定生物过程内部调控和生物过程之间相互调控的集中展现。本节我们以花分生组织的确立与分化和脱落酸介导的信息通路为例说明 ATRM 在研究和了解拟南芥转录调控过程中重要性。

3.4.1 特定生物过程转录调控通路的重现与扩展

转录因子间的正、负反馈调控在花发育过程中起着关键的作用。早在上世纪

90 年代，科学家就通过遗传分析提出了花器官分化和一致性确立的“ABC”模型(Bowman, *et al.*, 1991, Coen, *et al.*, 1991, Weigel, *et al.*, 1994)。其背后的分子机制也逐渐被揭示，然而目前仍有很多地方不太清楚。通过与文献综述(Irish, 2010)中总结的花分生组织确立和分化的转录调控通路比较，ATRM 能够重现 89%（24/27）的调控并能添加 27 个新调控（图 3-7）。这些新增加的调控有助于我们理解花发育背后的分子调控机理。比如转录因子 AP1 和 AP2 都是 A 功能所需要的，从文献综述总结的通路中可以清晰看出 AP1 作为 A 功能基因是如何发挥作用的，然而 AP2 是如何工作的在这个调控通路上并不明晰（图 3-1A）。ATRM 补充的 4 个 AP2 相关调控则能告诉我们 AP2 通过抑制 TFL1 调控花序分生组织到花分生组织的转变（Bradley, *et al.*, 1997），能激活 B 功能基因 AP3 和 PI 的表达和与 C 功能基因 AG 相互遏制（图 3-8）。这些 ATRM 补充的调控说明 AP2 以类似于另一个 A 类基因 AP1 的方式调控花分生组织的分化。

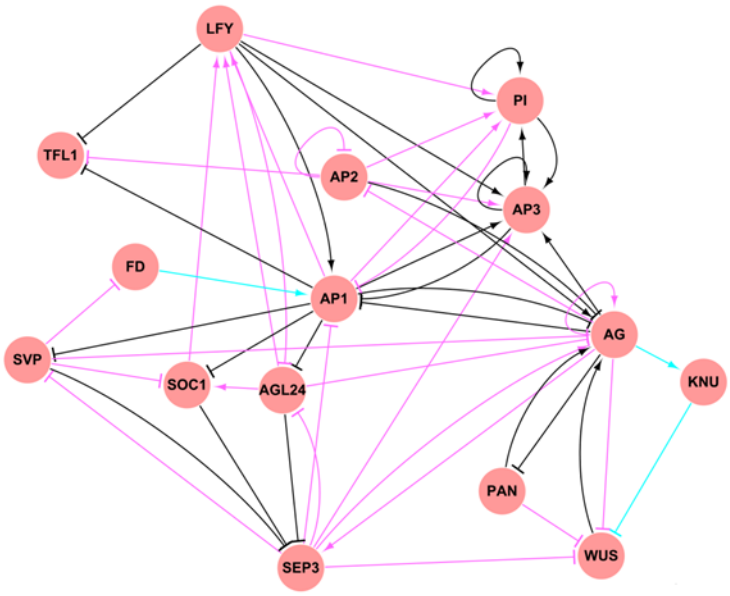

图 3-7　ATRM 与文献综述中总结的拟南芥花分生组织确立和分化的转录调控通路(Irish, 2010)的比较。图中黑线表示调控同时存在于总结的通路和 ATRM 中，青线表示调控不在 ATRM 中（比较后已加入 ATRM），粉红线表示相应调控不在总结的通路中。

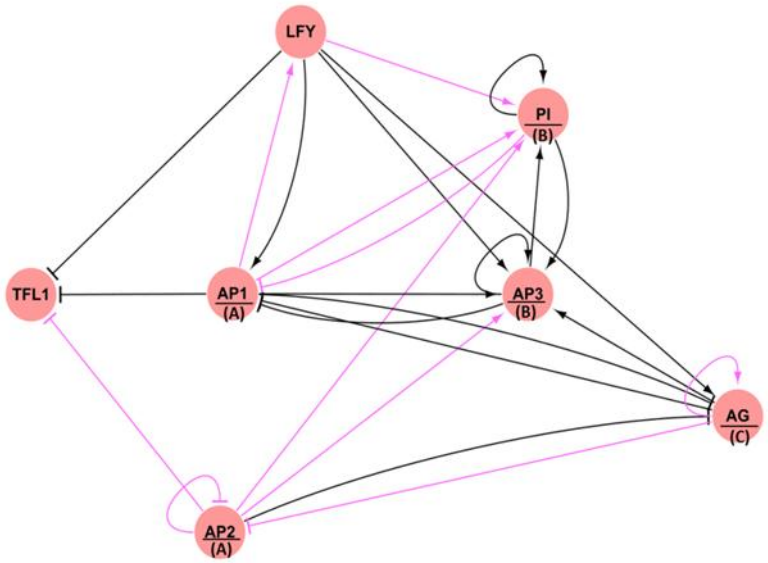

图 3-8　拟南芥花分生组织分化的转录调控通路。图中黑线表示相应的调控同时存在于文献综述总结的通路(Irish, 2010)和 ATRM 中，粉红线表示相应的调控不在总结的通路中。

3.4.2　生物过程内部及过程间的转录调控

复杂的生物网络中存在一些内部调控多于外部调控的模块（Community）,它们在生物体内发挥着某种相对独立的功能(Fortunato, 2010)。使用 CytoMCL(Guzzi, *et al.*, 2012)，我们将 ATRM 划分为 156 个模块，其中 62 个模块拥有 5 个或以上成员。通过使用 topGO (Alexa, *et al.*, 2010)对这 62 个模块进行 GO 富集分析，发现绝大多数模块（58/62）在特定的生物过程中发挥调控作用。根据前 5 个富集的 GO 项，我们对这些模块进行了命名（附录 3）。从附录 3 中可以看出，这些模块主要是一些信号转导通路以及发育和应激过程的调控通路。通过查看模块内部和模块间的调控，ATRM 提供了一个理解拟南芥特定生物过程和过程间转录调控的平台。

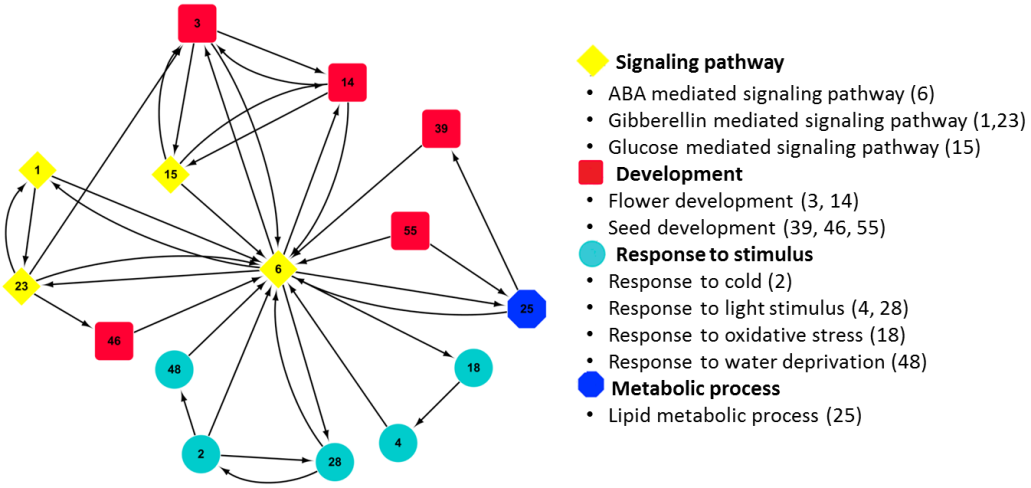

图 3-9　ATRM 中与 ABA 介导的信号通路相互调控的生物过程

上节花分生组织确立和分化的例子说明 ATRM 能辅助我们理解某特定生物过程内部的调控。现在以脱落酸（Abscisic acid，ABA）介导的信号通路为例，说明 ATRM 如何展现生物过程间的相互调控。研究表明 ABA 通过与其它激素、糖信号介导的通路相互调控在很多发育和应激过程中发挥着重要的调控作用(Cutler, *et al.*, 2010)。通过识别 ATRM 中 “ABA 介导的信号通路”（模块 6）及与其有相互调控的模块，可以看到 ABA 介导的信号通路与赤霉素（Gibberllin）、糖介导的信号通路、花和种子发育相关的通路以及各种胁迫相关通路之间的相互调控（图 3-9）。如果进一步细看 ATRM 包含的相应模块之间是如何调控的，则能帮助我们理解 ABA 介导的信号通路是如何在转录调控水平上完成上述过程的。

3.5拟南芥转录调控在表达相关性上的总体模式

3.5.1　拟南芥转录调控的总体表达相关性

转录因子与其靶基因之间的表达相关性是依靠芯片、RNA-seq 等表达数据构建转录调控网络的基础。有研究报道，在酵母中转录因子与其靶基因的总体表达相关性比背景还低(Wu, *et al.*, 2012)。在拟南芥中，转录因子与其靶基因在总体表达相关性上又是怎样的？基于 ATTED-II 计算的 Pearson 相关系数，我们研究了拟南芥转录调控在表达相关性上的关系。与酵母中的情况不同(Wu, *et al.*, 2012)，我们的结果显示拟南芥转录调控对在总体表达相关性上显著高于背景（单侧 Wilcoxon 秩和检验 $P < 2.2e\text{-}16$，图 3-10）。不过，我们也应该意识到多数转录调控对的总体

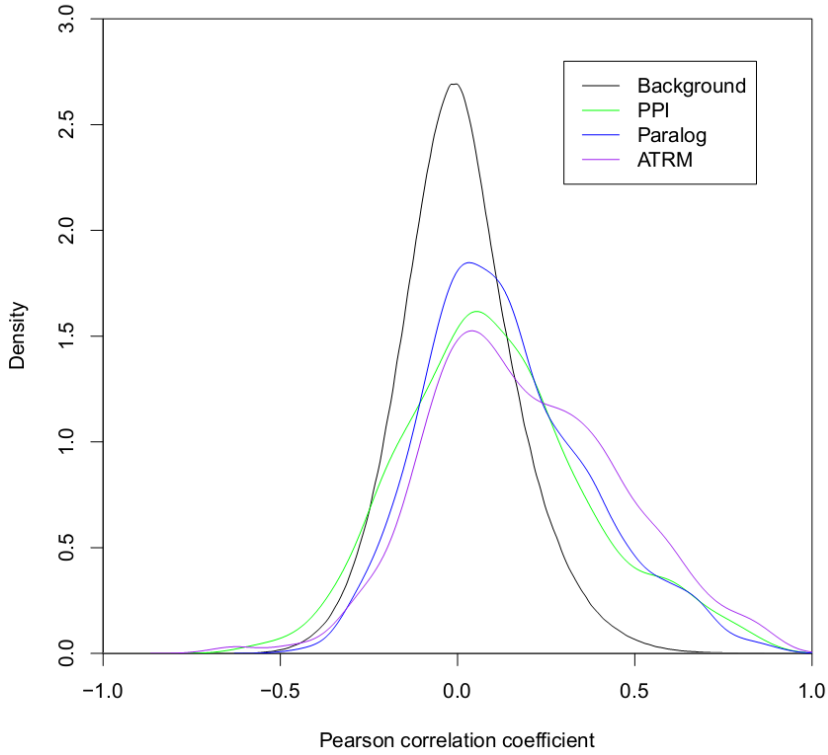

图 3-10 转录因子相关的蛋白蛋白相互作用（PPI）、旁系同源基因（Paralog）以及 ATRM 中转录调控在表达上的 Pearson 相关系数（PCC）的密度分布曲线。

表达相关性依然较低，只有接近 14%转录调控对在表达上的 Pearson 相关系数大于 0.5。与此同时，一些非转录调控因素，如相互作用的蛋白和旁系同源基因之间也具有某种程度的表达相关性（图 3-10）。这些因素加大了使用表达数据预测转录调控网络的难度。

3.5.2 不同类型的转录调控在总体表达相关性上的比较

转录调控可分为不同的类型，如转录因子之间的调控与转录因子和非转录因子之间的调控，激活调控与抑制调控。这些不同类型的调控在总体表达相关性上有何模式尚不清楚。基于 ATRM 中的转录调控关系和 ATTED-II 计算的 Pearson 相关系数，我们比较了不同类型的转录调控在总体表达相关性上的特征。与转录因子和非转录因子（TF-nonTF）间的调控相比，转录因子间（TF-TF）的调控具有更多抑制类型的调控（表 3-4）。在表达相关性上，TF-TF 调控要明显高于 TF-nonTF 调控（单侧 Wilcoxon 秩和检验 P = 8.6e-06,图 3-11A）。和预期一样，激活类型的转录调控在总体表达相关性上要显著高于抑制类型的调控（单侧 Wilcoxon 秩和检验 P = 6.7e-11,图 3-11B）。出乎意料，抑制类型转录调控在总体表达相关性也是正的

表 3-4　转录因子间（TF-TF）调控与转录因子和非转录因子（TF-nonTF）调控的调控类型比较

	激活	抑制
TF-TF	416	221
TF-nonTF	582	209

单侧 Fisher 精确检验 $P = 0.0004$

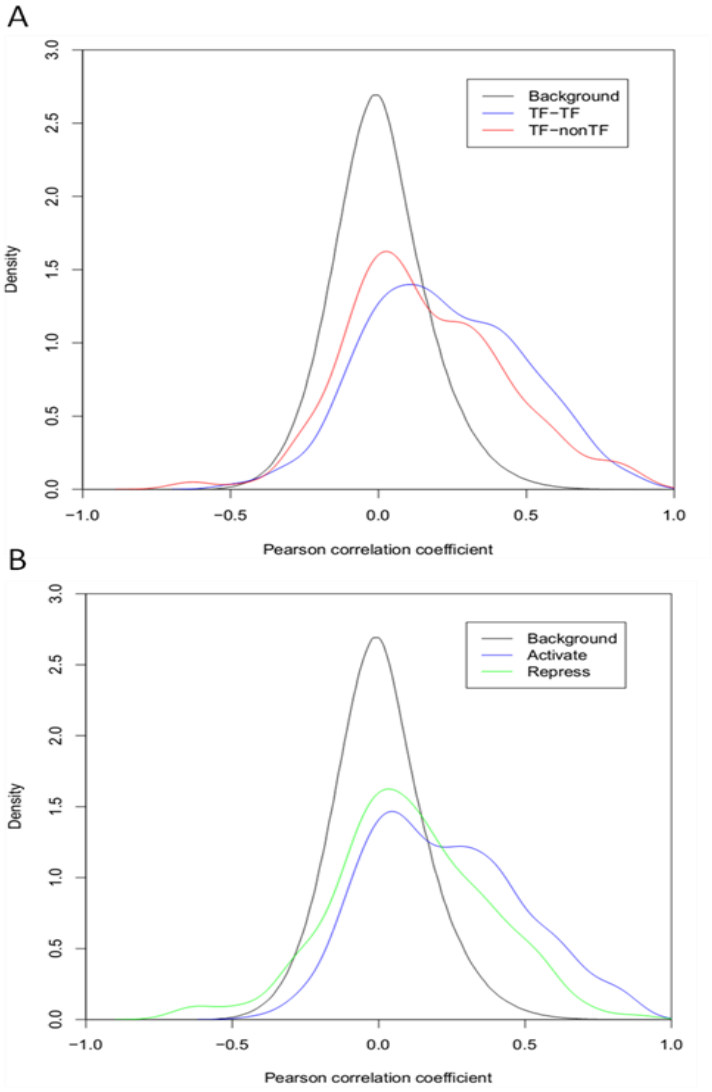

图 3-11 不同类型的转录调控在总体表达相关性上的比较。其中（A）为转录因子间（TF-TF）调控与转录因子和非转录因子（TF-nonTF）调控在表达上的 Pearson 相关系数的密度分布曲线，（B）为激活和抑制类型转录调控 Pearson 相关系数的密度分布曲线。

（中值 0.09，均值 0.11），显著高于背景值（中值 0，均值 0.01；单侧 Wilcoxon 秩和检验 $P < 2.2e-16$；图 3-11B）。这可能是由于转录因子和调控基因间存在某种反馈造成的。

3.6本章小结

通过系统的文献挖掘和人工校正，我们收集并构建了一个拟南芥转录调控网络（ATRM）。它包含 1431 个调控，涉及 47 个家族、388 个转录因子，分别覆盖了拟南芥家族和成员数的 81.0%和 22.8%。这些调控主要集中在生长发育和应对生物/非生物胁迫等生物过程。共过程和共表达两个方面的评估结果都表明 ATRM 的高质量。通过与文献综述中总结的花分生组织确定和分化的通路相比，ATRM 除能很好地重现其中的转录调控外，其补充的调控将辅助理解生物过程背后的分子调控机理。通过划分生物模块，ATRM 在基因组水平上展示了拟南芥生物过程内部和各过程间相互调控的概况。转录因子与靶基因在表达上的相关性是依靠表达数据预测转录调控网络的基础。拥有一个拟南芥转录调控网络，我们揭示了拟南芥转录调控在表达相关性上的总体特征、不同类型的转录调控在表达相关性上的差异等，这些发现有利于转录调控网络预测研究的改进。

第 4 章 拟南芥转录调控网络的架构

4.1概述

在大肠杆菌和酿酒酵母等单细胞生物中的研究发现它们的转录调控网络是由一些结构元件组成的(Shen-Orr, *et al.*, 2002, Lee, *et al.*, 2002, Milo, *et al.*, 2002)，这些结构元件同时也是一些能够完成特定生物功能的功能元件(Rosenfeld, *et al.*, 2002, Becskei, *et al.*, 2000, Kalir, *et al.*, 2004, Mangan, *et al.*, 2003, Alon, 2007)。对酵母内源和外源系统的研究则发现不同生物系统可能采取不同的架构来完成相应的功能(Luscombe, *et al.*, 2004)。

与低等的单细胞生物不同，高等植物既要精确的调控复杂的多细胞发育过程如组织和器官的发育，又要对外界的各种生物和非生物胁迫做出迅速的响应，那么植物到底演化出一个怎样的转录调控系统来满足以上需求的？

拥有一个基因组范围的高质量的拟南芥转录调控网络 ATRM，我们有机会去研究一个植物的转录调控网络是如何构建的？与单细胞生物相比，拟南芥的转录调控网络中是否存在一些新的结构元件？它的发育系统和应激系统在结构元件组成、全局拓扑结构和参与构建它们的转录因子的性质上又有何不同？本章将针对这些问题进行研究和探讨。

4.2拟南芥转录调控网络中的结构元件

4.2.1 转录调控网络数据

拥有一个高质量的基因组范围的拟南芥转录调控网络 ATRM，我们有机会去系统研究一个植物转录调控网络的结构元件及其组成。为将结果与单细胞生物比较，我们从 RegulonDB 8.0 (Salgado, *et al.*, 2013) 下载了大肠杆菌的转录调控网络，从 YEASTRACT (Abdulrehman, *et al.*, 2011)中提取了酿酒酵母中具有直接调控证据并且功能已确定的转录调控关系。三个模式物种转录调控网络数据的统计见表 4-1。

表 4-1 大肠杆菌、酿酒酵母和拟南芥转录调控网络数据的统计

物种	节点数	调控数
大肠杆菌 (*E. coli*)	1707	3787
酿酒酵母 (*S. cerevisiae*)	1532	2479
拟南芥 (*A. thaliana*)	789	1406

注：未包含自调控

4.2.2 结构元件的识别方法

结构元件是指与其相对应的随机网络相比，在实际网路中显著富集的调控模式(Shen-Orr, *et al.*, 2002, Milo, *et al.*, 2002, Alon, 2007)。为研究拟南芥调控网络中的结构元件组成，我们使用 Mfinder 1.2 (Milo, *et al.*, 2002)系统识别出所有 3 个基因间的调控模式，并通过统计分析找出其中的结构元件。在不考虑调控类型（激活或抑制）的情况下，3 个基因间的调控模式共有 13 种（图 4-1）。

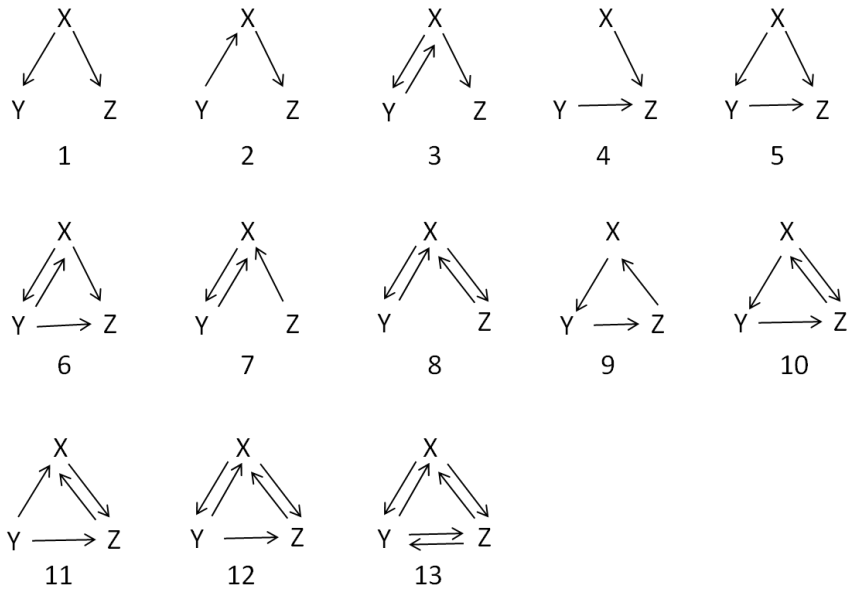

图 4-1 3 个基因间所有可能的 13 种调控模式

在保持节点出度、入度和相互调控的边数不变的情况下，我们使用 Mfinder 生出 1000 个与 ATRM 相对应的随机网络。然后通过与随机网络中各调控模式的数目比较，找出 ATRM 中出现次数显著大于随机网络的调控模式。在判断一个调控模式是否显著出现时，使用了 Mfinder 默认的阈值，包括 $P < 0.01$, Mfactor > 1.10 和 Uniqueness >=4 等三个方面。其中，P 值是基于生成的 1000 个随机网络计算的，Mfactor (N_{real}/N_{rand})是某种模式在真实网络中的数目与随机网络中数目的比值，

Uniqueness 指实际网络中构建某种类型结构元件的不同基因集合的数目(Milo, *et al.*, 2002)。

4.2.3 拟南芥转录调控网络中的结构元件

4.2.3.1 结构元件的系统识别

使用 Mfinder 1.2，我们从大肠杆菌、酿酒酵母和拟南芥转录调控网络中系统识别出所有 3 节点的调控模式，并找出其中显著富集的调控模式，即结构元件（表 4-2、表 4-3 和表 4-4，表中下划线标注的为结构元件）。其中在大肠杆菌中有两种结构元件（Motif 5 和 Motif 6）（表 4-2），在酿酒酵母中有一种结构元件（Motif 5）(表 4-3)。前人的动力学模拟和实验研究表明，这些元件在它们应对各种环境胁迫中发挥着非常重要的作用(Rosenfeld, *et al.*, 2002, Becskei, *et al.*, 2000, Kalir, *et al.*, 2004, Mangan, *et al.*, 2003, Alon, 2007)。

表 4-2 大肠杆菌转录调控网络中的调控模式，下划线标注的为结构元件

Motif id	N_{real}	$N_{Random} \pm SD$	*P*	Uniqueness
1	281705	282482.2 ±56.3	1.00	144
2	2561	3329.8 ±56.0	1.00	28
3	1105	1467.5 ±61.3	1.00	6
4	3384	4333.1 ±63.4	1.00	36
5	**1145**	**376.4 ±56.0**	**0**	**23**
6	**230**	**49.6 ±30.8**	**0**	**6**
7	36	54.6 ±3.7	1.00	4
8	0	3.6 ±0.7	1.00	0
9	0	0.3 ±0.5	1.00	0
10	2	1.0 ±1.0	0.27	1
11	12	3.5 ±1.8	0	2
12	1	0.4 ±0.6	0.35	1
13	1	0.0 ±0.1	0.01	1

表 4-3 酿酒酵母转录调控网络中的调控模式，下划线标注的为结构元件

Motif id	N_{real}	$N_{Random} \pm SD$	*P*	Uniqueness
1	89808	89897.8 ±10.4	1.00	86
2	1202	1291.8 ±10.3	1.00	21
3	271	271.9 ±4.1	0.69	2
4	1581	1671.6 ±10.5	1.00	39
5	**147**	**57.0 ±10.3**	**0**	**12**
6	4	3.5 ±2.0	0.44	2
7	11	10.3 ±1.0	0.64	2
8	1	0.9 ±0.3	0.93	1
9	0	0.0 ±0.2	1.00	0
10	0	0.1 ±0.3	1.00	0
11	0	0.2 ±0.5	1.00	0
12	0	0.1 ±0.3	1.00	0
13	0	0.0 ±0.0	1.00	0

表 4-4 拟南芥转录调控网络中的调控模式，下划线标注的为结构元件

Motif id	N_{real}	$N_{Random} \pm SD$	*P*	Uniqueness
1	4944	5230.2 ±7.9	1.00	109
2	2058	2355.1 ±9.0	1.00	52
3	439	577.3 ±4.9	1.00	17
4	1763	2068.3 ±7.7	1.00	77
5	**303**	**46.2 ±7.4**	**0**	**37**
6	**53**	**4.5 ±2.1**	**0**	**10**
7	286	386.0 ±5.3	1.00	15
8	44	70.9 ±2.5	1.00	6
9	9	1.9 ±1.4	0.001	3
10	**22**	**3.2 ±1.9**	**0**	**6**
11	**34**	**4.6 ±2.3**	**0**	**8**
12	**25**	**2.6 ±1.6**	**0**	**8**
13	2	0.5 ±0.7	0.08	1

在拟南芥转录调控网络 ATRM 中有 5 种结构元件（图 4-4）。与单细胞的大肠杆菌和酿酒酵母相比，拟南芥有 3 个新结构元件（Motif 10、Motif 11 和 Motif 12）（图 4-2）。其中的 2 个新结构元件（Motif 11 和 Motif 12）在人的转录调控网络中也是富集的(Neph, *et al.*, 2012, Gerstein, *et al.*, 2012)。如图 4-2 所示，“Motif 5（303）”中“Motif 5”为结构元件的编号，括号中的数字“303”为 ATRM 中该结构元件的数目。与单细胞生物相比，拟南芥中新出现的结构元件用红色表示。

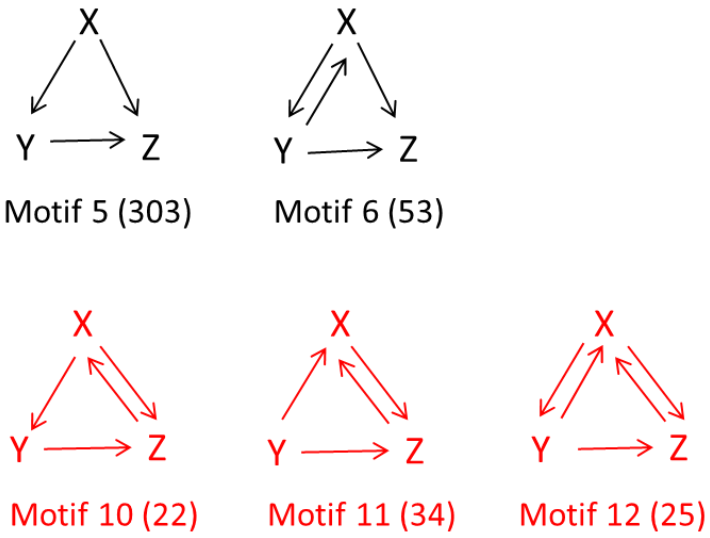

图 4-2　拟南芥转录调控网络 ATRM 中的结构元件。与单细胞的大肠杆菌和酿酒酵母相比，ATRM 中新出现的结构元件用红色标记

4.2.3.2 拟南芥转录调控网络中新结构元件的功能

与单细胞生物相比，拟南芥转录调控网络中存在 3 个新结构元件。为什么拟南芥转录调控网络中会出现这些新结构元件？它们有什么生物学功能？本节将针对这些问题进行探讨。

通过对构建结构元件的转录因子作 GO 生物过程（Biological process, BP）的富集分析（表 4-5 和表 4-6），发现：与参与单细胞生物中已存在结构元件构建的转录因子相比（表 4-5），参与新结构元件构建的转录因子更倾向于参与多细胞发育过程、生殖过程和器官发育等生物过程(表 4-6)。在富集分析中，拟南芥中具有 GO 生物过程注释的转录因子作为背景，Benjamin 和 Hochberg 的方法用于调整 *P* 值。

表 4-5 参与单细胞的大肠杆菌和酿酒酵母中已存在结构元件（Motif 5 和 Motif 6）构建的转录因子所富集的生物过程（前 20）

GO id	**GO term**	***P***	**Adjusted *P***
GO:0051093	negative regulation of developmental process	4.0e-07	0.0005
GO:0048518	positive regulation of biological process	2.1e-06	0.0013
GO:0031323	regulation of cellular metabolic process	4.8e-06	0.0017
GO:0019222	regulation of metabolic process	6.0e-06	0.0017
GO:0050794	regulation of cellular process	9.8e-06	0.0017
GO:0009893	positive regulation of metabolic process	1.0e-05	0.0017
GO:0031325	positive regulation of cellular metabolic process	1.0e-05	0.0017
GO:0048522	positive regulation of cellular process	1.6e-05	0.0021
GO:0044237	cellular metabolic process	1.8e-05	0.0021
GO:0065007	biological regulation	2.0e-05	0.0021
GO:0001708	cell fate specification	2.1e-05	0.0021
GO:0032501	multicellular organismal process	2.1e-05	0.0021
GO:0009987	cellular process	2.4e-05	0.0021
GO:0048856	anatomical structure development	2.5e-05	0.0021
GO:0050789	regulation of biological process	2.6e-05	0.0021
GO:0080090	regulation of primary metabolic process	3.9e-05	0.0026
GO:0010077	maintenance of inflorescence meristem identity	4.0e-05	0.0026
GO:0010187	negative regulation of seed germination	4.2e-05	0.0026
GO:0008152	metabolic process	4.2e-05	0.0026
GO:0048646	anatomical structure formation involved in morphogenesis	4.6e-05	0.0027

表 4-6　参与新结构元件（Motif 10、Motif 11 和 Motif12）构建的转录因子富集的 GO 生物过程（前 20）

GO id	GO term	*P*	Adjusted *P*
GO:0048731	system development	3.2e-16	3.8e-13
GO:0048856	anatomical structure development	7.9e-16	4.7e-13
GO:0007275	multicellular organismal development	3.6e-15	1.4e-12
GO:0032501	multicellular organismal process	7.9e-15	2.4e-12
GO:0032502	developmental process	4.2e-14	1.0e-11
GO:0048608	reproductive structure development	1.8e-13	3.6e-11
GO:0003006	developmental process involved in reproduction	9.6e-13	1.6e-10
GO:0000003	reproduction	2.1e-12	2.8e-10
GO:0022414	reproductive process	2.1e-12	2.8e-10
GO:0009888	tissue development	7.5e-12	8.9e-10
GO:0010073	meristem maintenance	3.4e-11	3.7e-09
GO:0007389	pattern specification process	9.6e-11	9.5e-09
GO:0003002	regionalization	1.1e-10	1.0e-08
GO:0048513	organ development	1.3e-10	1.1e-08
GO:0048507	meristem development	7.6e-10	6.0e-08
GO:0030154	cell differentiation	3.9e-09	2.9e-07
GO:0001708	cell fate specification	5.0e-09	3.4e-07
GO:0045165	cell fate commitment	5.1e-09	3.4e-07
GO:0009908	flower development	6.5e-09	4.1e-07
GO:0045596	negative regulation of cell differentiation	9.1e-09	5.3e-07

为进一步研究结构元件是如何参与调控系统构建的，按照如下标准将识别的结构元件划分为参与发育系统构建的结构元件和参与应激系统构建的结构元件。对于参与发育系统构建的结构元件，要求其 3 个基因中至少有 2 个基因注释为参与发育过程，并且没有一个基因仅注释为参与应激过程，反之亦然。表 4-7 统计了 ATRM 中各结构元件在发育过程和应激过程中的分布情况。从表中可以看出新结构元件全部参与到发育系统构建中，发育过程和应激过程在结构元件的类型和组成上有着明显的不同。

表 4-7 ATRM 中的结构元件在发育系统和应激系统中的分布情况，新结构元件以下划线标记

编号	发育	应激
Motif 5	107	54
Motif 6	25	4
Motif 10	**17**	**0**
Motif 11	**32**	**0**
Motif 12	**23**	**0**

以上结果显示新结构元件倾向于构建发育系统的网络。前人研究表明 Motif 11 能在信号出现后转换到某个状态并维持下去，这些功能是多细胞的发育过程所需要的(Alon, 2007)。拟南芥中的新结构元件是否也在这方面起作用呢？参考 Mangan 等的工作(Mangan, *et al.*, 2003)，我们使用动力学模拟的方法（公式 1-3）模拟了这些结构元件可能的功能。

$$dX/dt = B_x + \alpha_x f(Y, K_{yx}, Z, K_{zx}) - \beta_x X \quad (1)$$

$$dY/dt = B_y + \alpha_y f(X, K_{xy}, Z, K_{zy}) - \beta_y Y \quad (2)$$

$$dZ/dt = B_z + \alpha_z f(X, K_{xz}, Y, K_{yz}) - \beta_z Z \quad (3)$$

在公式 1-3 中，B_i 为基因 i 的基础（本底）转录率，K_{ij} 为基因 i 对基因 j 的转录激活或者抑制系数。对于激活来说 $f(i, K_{ij}) = K_{ij}C_i/(1+ K_{ij}C_i)$，抑制 $f(i, K_{ij}) = (1-K_{ij}C_i)/(1+ KC_{max})$ （C_i 是基因 i 的转录水平，在初始状态时假定 X 开始处于最高转录速率 1 然后被基因 Y 或者 Z 竞争性的抑制）。β_i为基因 i 的降解速率，如果一个基因被两个转录因子同时调控，比如 Z 被 X 和 Y 激活，则 $f(X, K_{xz}, Y, K_{yz}) = (K_{xz}C_x+K_{yz}C_y)/(1+ K_{xz}C_x+K_{yz}C_y)$。在模拟中假定基因 i 的最低转录水平为 0，最高为 1。为了满足这一约束，我们取参数 $B_i = 0$, $\alpha = 1$, $\beta = 0.5$，$K_{ij} = 1$。如果$\sum KC > 1$，则取上限 1。在动力学模拟中，初始状态为 X 处于最高转录水平 1，Y 和 Z 处于转录水平 0。当 Y 和 Z 的转录水平不低于 0.5 时，它们处于活跃状态并能激活/抑制相关基因的转录。

基于上述公式，使用 MATLAB 中的 ODE45 模块对 Motif 11、Motif 10 和 Motif 12 的可能功能进行了模拟。结果如图 4-3 所示，其中的 A、B、C 分别是对 Motif 11、

Motif 10 和 Motif 12 中的一个例子进行的模拟，图中不同背景颜色代表不同状态（使用 X 和 Z 的表达来代表不同的状态）。动力学模拟表明它们可以完成状态的转换和维持，这些功能是多细胞发育尤其是细胞的命运决定中所需要的。

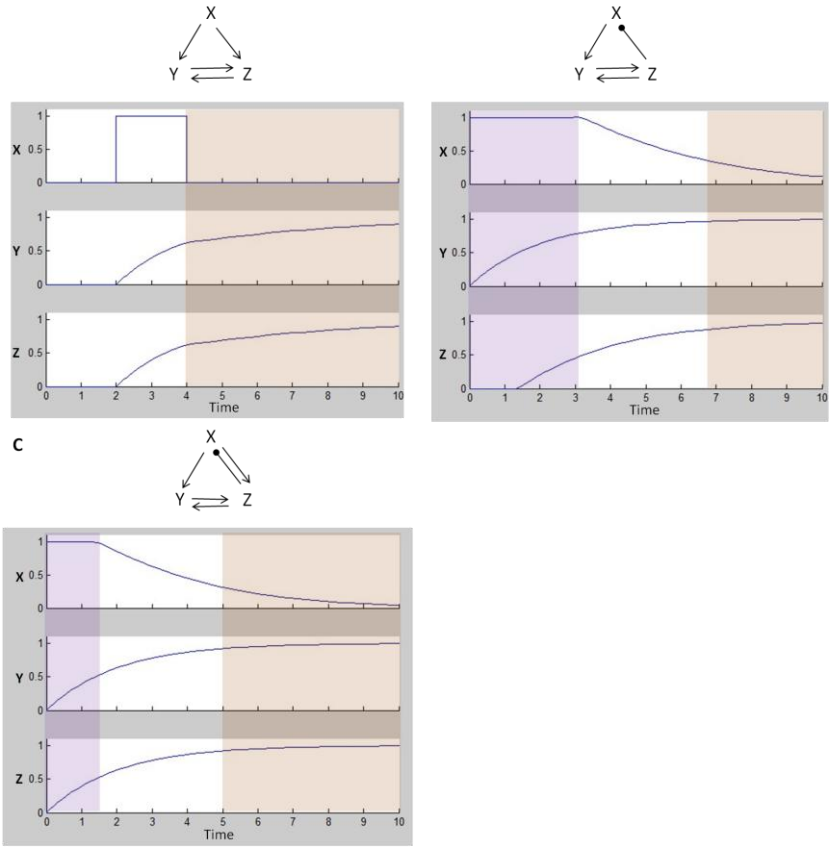

图 4-3　动力学模拟 ATRM 中新结构元件的功能

4.3 发育系统和应激系统在网络全局拓扑结构上的差异

4.3.1　发育子网络和应激子网络的划分

发育系统和应激系统是植物转录调控系统的两个重要组成部分。为了研究这两部分在网络全局拓扑结构上是否有所差异，按照如下方法分别识别出发育子网络和应激子网络：

(1) 根据具有实验证据的 GO 注释识别出参与发育过程的基因（GO:0032502）和参与应激过程的基因(GO:0006950、GO:0009607 和 GO:0009628),只有只被注释为发育过程或应激过程的基因用于第 2 步的分析。

(2) 识别出只参与发育过程的基因之间的调控（发育子网络）和只参与应激过程的基因之间的调控(应激子网络)，结果如表 4-8 所示。

表 4-8 发育子网络和应激子网络的数据统计

	发育子网络	应激子网络
转录因子数	109	67
靶基因数	117	118
调控数	315	224

4.3.2 发育系统和应激系统在网络全局拓扑结构上的差异

4.3.2.1 发育系统和应激系统的网络全局拓扑结构

结构元件组成是网络局部结构的反映，下面的四个参数则是对网络全局拓扑结构的衡量：包括平均的出度(<Out-degree>)、入度(<In-degree>)、路径长度(<Path length>)和聚类系数(<Clustering coefficient>)，其中<>表示在网络中该参数的平均值。在转录调控网络中，上面四个参数分别对应<Targets per TF>、<TFs per target>、<Path length>和<Clustering coefficient>。其中，<Targets per TF>衡量转录因子可以瞬时启动多少靶基因的转录；<TFs per target>衡量基因被转录因子调控的情况，即调控的复杂程度；<Path length>指调控的路径长度，即信号从转录因子传递到末端基因所需要的路径长度；<Cluster coefficient>衡量调控的复杂程度，尤其是转录因子和转录因子之间的调控(Luscombe, *et al.*, 2004)。

为降低比较散在的基因影响参数计算的稳定性，并最大限度的降低可能潜在的研究偏差带来的影响，我们选取发育子网络和应激子网络中最大的连通组分（去除了比较分散的少数基因和调控）并使用 igraph 0.6 (Csardi, *et al.*, 2006)来估计它们的全局拓扑参数。结果如表 4-9 所示，与应激子网络相比，参与发育子网络构建的转录因子具有较少的靶基因，相关的基因被更多的转录因子调控，具有更长的调控路径和更高的调控复杂性。

表 4-9　发育子网络和应激子网络的全局拓扑结构

	发育	应激
<Targets per TF>	2.96	3.77
<TFs per target>	2.75	1.97
< Path length >	3.77	1.74
< Clustering coefficient >	0.21	0.08

4.3.2.2 发育系统和应激系统在网络全局拓扑结构上的差异是显著和稳定的

发育系统和应激系统在网络全局拓扑结构上存在差异，但是基于文献收集的转录调控网络能否正确反映两个子系统之间的差异？它们之间的差异是否显著？我们将从下面几个方面说明它们之间差异是显著和稳定的（Robust）。

二者在收集深度上没有显著差异

如果一个子系统的研究比另一个子系统的研究更加透彻可能会对评估它们之间的差异带来偏差。为了检查它们在这方面是否存在系统性偏差，我们比较了发育子网络和应激子网络中转录因子在连接数（代表收集的深度）上是否存在显著的差异。结果显示，无论在两个子网络中还是最终用于计算拓扑参数的最大连通组分中，二者在收集的深度上都没有显著的差异（图 4-4）。

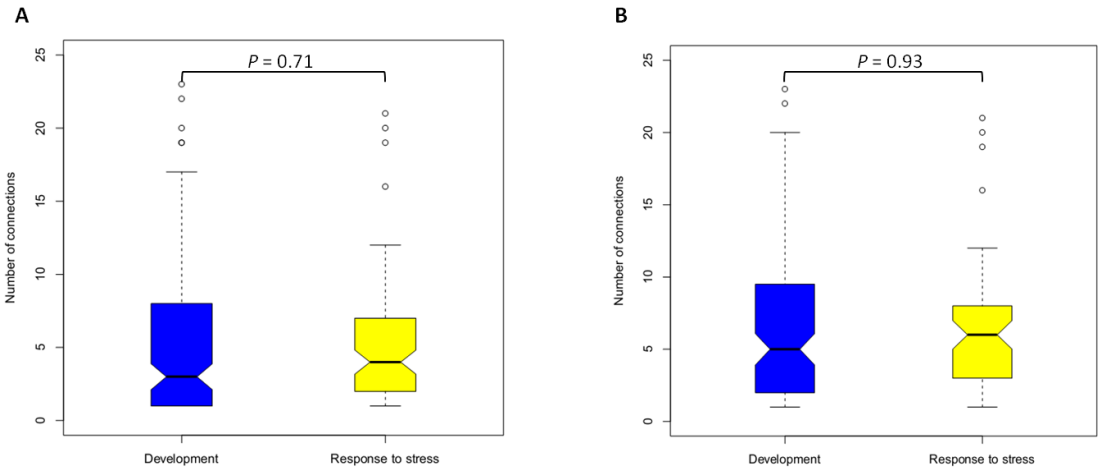

图 4-4　ATRM 中参与发育过程和应激过程的转录因子在收集深度上的比较。(A) 为 ATRM 中各转录因子的连接数，（B）为用于拓扑参数计算的最大连通组分中各转录因子的连接数。图上标注的是双侧 Wilcoxon 秩和检验的 *P* 值。

基于文献收集的转录调控网络能正确地反映两个系统之间的差异

基于 ChIP-chip 数据构建的转录调控网络，Luscombe 等发现酿酒酵母的细胞周期和双峰转换过程在全局拓扑结构上存在显著差异(Luscombe, *et al.*, 2004)。因此，我们使用人工收集的调控数据检查此类数据能否正确反映二者之间的差异。

酿酒酵母的转录调控数据是从 YEASTACT (Abdulrehman, *et al.*, 2011)检索提取的文献中报道的具有直接调控证据和功能确定的调控关系。Luscombe 等确定的参与细胞周期和双峰转换的基因列表(Luscombe, *et al.*, 2004)用于识别细胞周期子网络和双峰转换子网络（相应过程基因之间的相互调控组成的网络，表 4-10）。

表 4-10 酿酒酵母细胞周期子网络和双峰转换子网络的数据统计

	细胞周期子网络	双峰转换子网络
转录因子数	46	61
靶基因数	161	626
调控数	259	1046

由于细胞周期子网络中的调控数远小于双峰转换子网络的调控数（表 4-9），我们通过从双峰转换子网络中抽样的方式来比较两个子网络间的差异。通过从双峰转换网络中抽取与细胞周期网络相同边数的子网络，并重复抽样 10000 次，然后分别选取其中最大的连通组分计算全局拓扑结构参数。最后的结果显示基于文献收集的调控网络揭示的模式与 Luscombe 等的发现完全一致（图 4-5）。

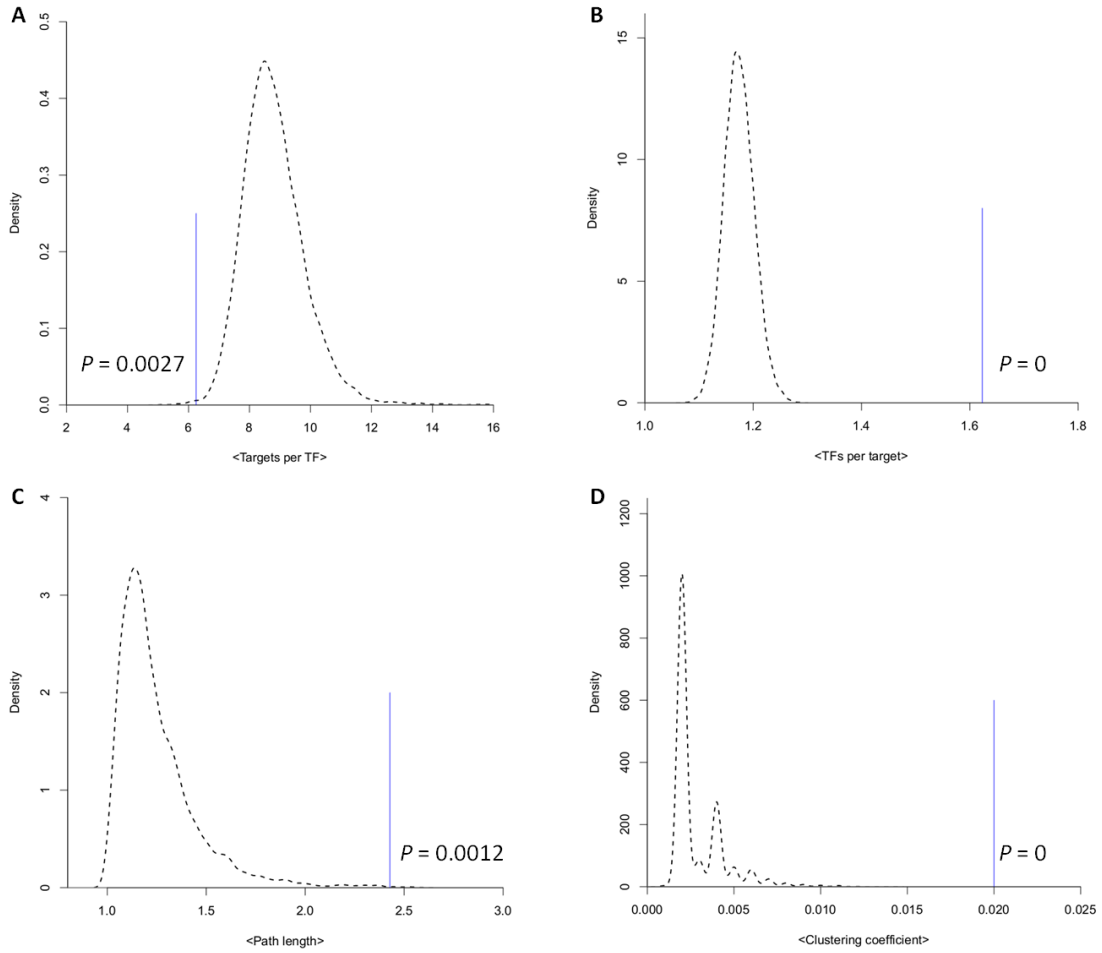

图 4-5　双峰转换子网络与细胞周期子网络在全局拓扑结构上的比较。图中的虚线是从双峰转换子网络中抽样 10000 次计算的全局拓扑结构参数的分布情况，蓝色竖线则是细胞周期子网络中的相应参数的值。图中标注的 *P* 值是基于 10000 次重复抽样计算的。

二者之间的差异是显著和稳定的

发育子网络和应激子网络在全局拓扑结构上存在差异，但是这种差异是否显著和稳定？通过抽样检验、使用不同的 GO 注释版本和将同时参与发育和应激过程的基因考虑进来的方式评估了该差异的显著性和稳定性。

通过从发育子网络中抽取与应激子网络相同的调控数，并重复抽样 10000 次，我们比较了发育子网络和应激子网络在全局拓扑结构上的差异。结果显示它们之间的差异是显著的(图 4-6)。

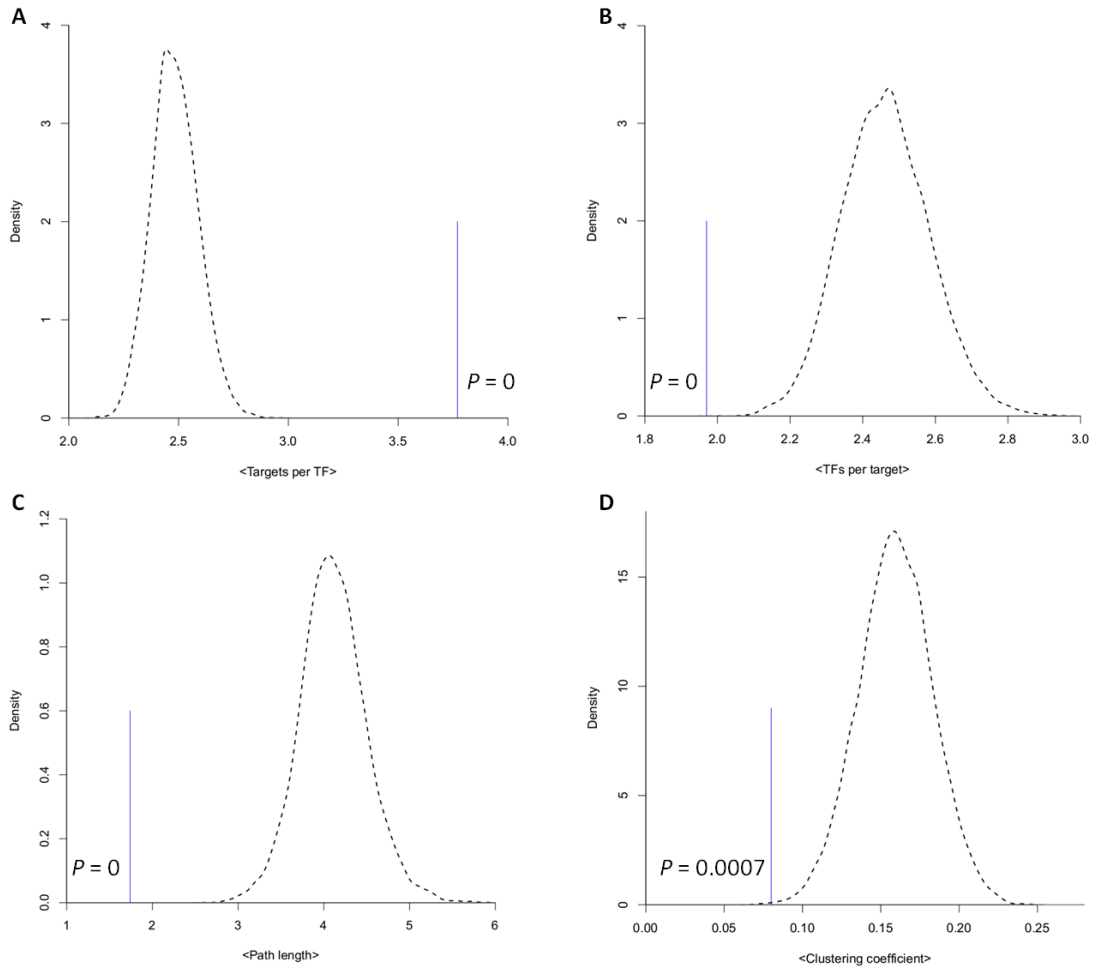

图 4-6　发育子网络和应激子网络在全局拓扑结构上的差异。图中的虚线是从发育子网络中抽样 10000 次计算的全局拓扑结构参数的分布情况，蓝色竖线则是应激子网络中相应参数的值。图中标注的 *P* 值是基于 10000 次抽样计算的。

为了检查网络大小和研究深度对估计全局拓扑结构参数的影响，分别随机从发育子网络和应激子网络中抽取 50%、60%、70%、80%和 90%的边各 1000 次，然后选取其中最大的连通组分来计算相应的参数。结果如图 4-7 所示，发育系统和应激系统转录调控网络在全局拓扑结构上的差异是显著和稳定的。

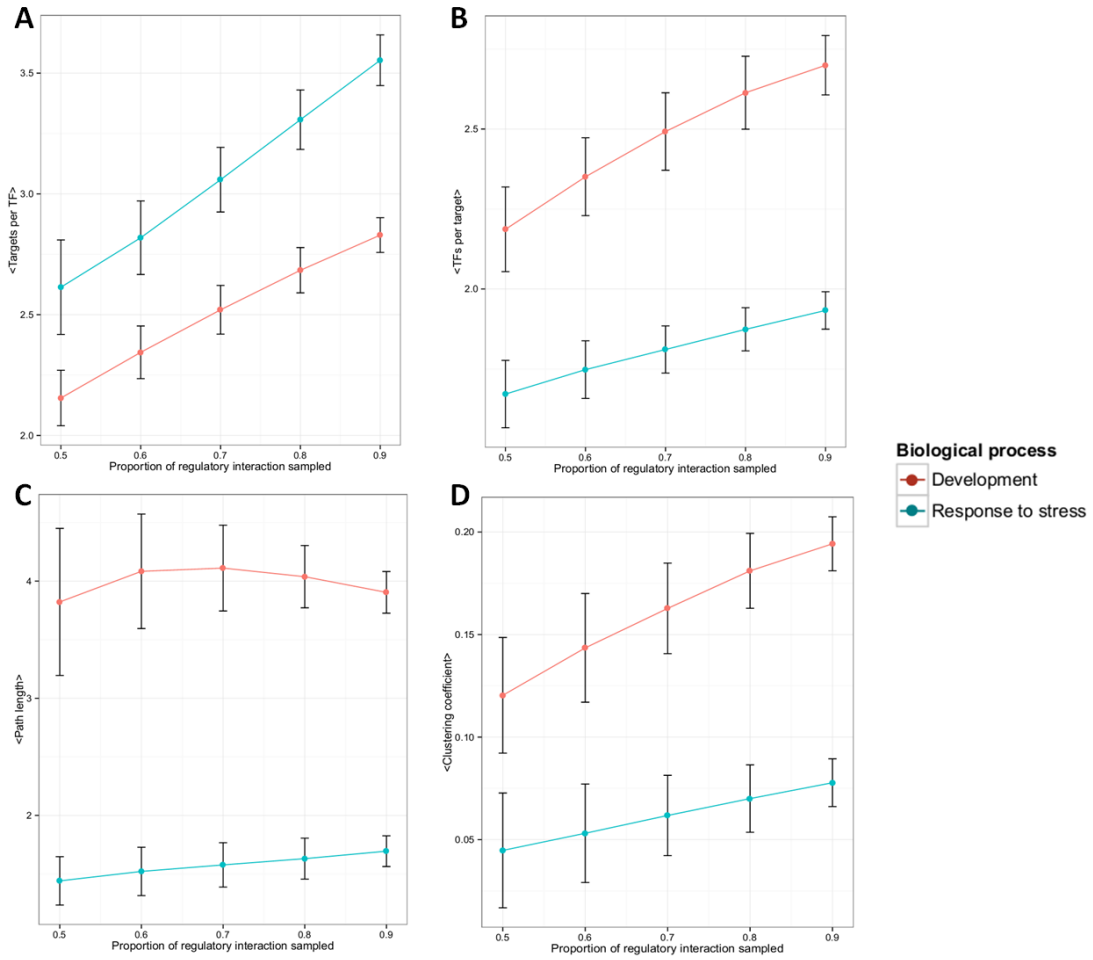

图 4-7　抽样检测发育子网络和应激子网络在全局拓扑结构差异上的显著性和稳定性。通过从发育子网络和应激子网络中分别抽取 50%、60%、70%、80%和 90%的边各 1000 次，检查了其对全局拓扑结构参数计算的影响。图中的竖线为各抽样结果的标准差。

在识别发育子网络和应激子网络时，我们使用了 TAIR 10 中的 GO 注释。如果使用其它注释版本，是否依旧有一致的模式？通过整合 TAIR 10 和 EBI 的 GO 注释重新划分发育子网络和应激子网络，重复以上抽样检验过程并得到了一致的结果（图 4-8）。

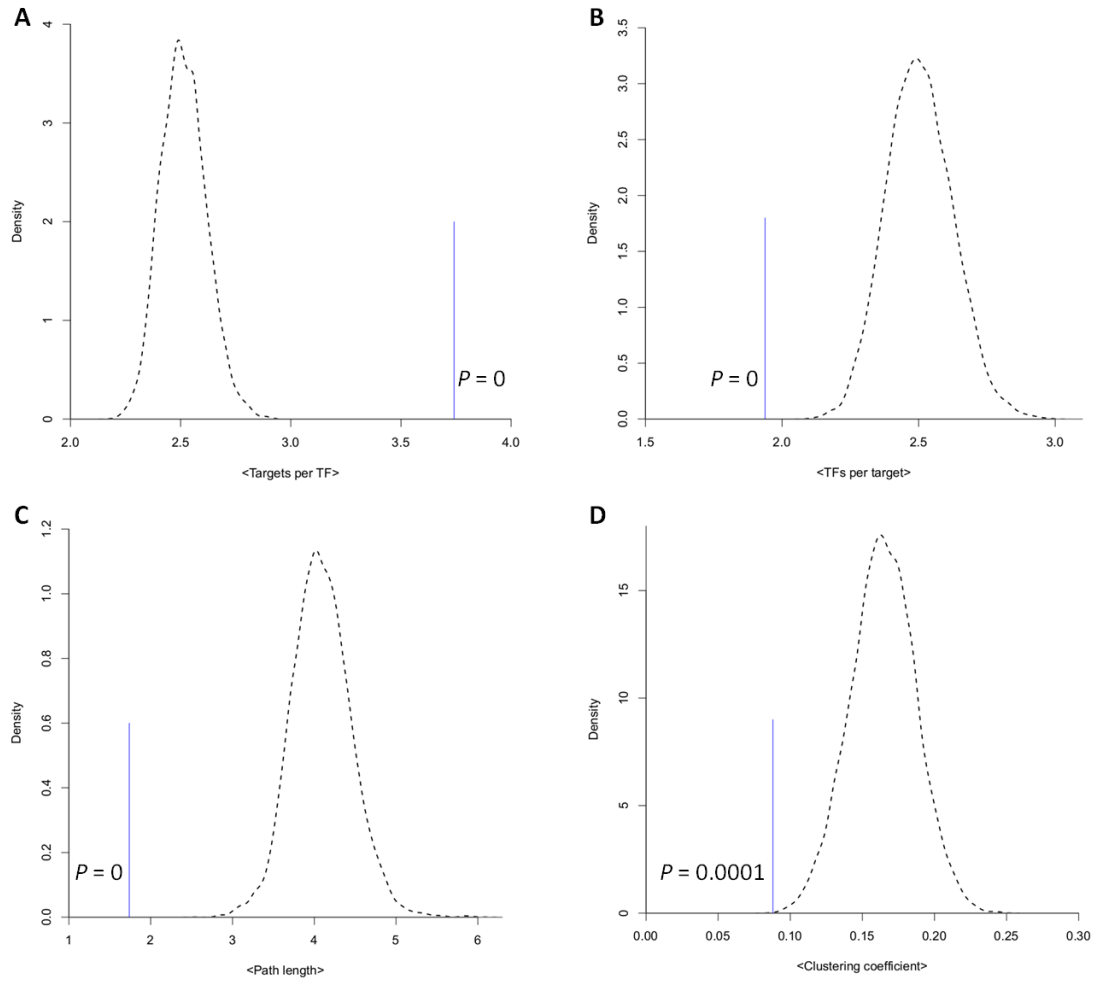

图 4-8　抽样检测发育子网络和应激子网络在全局拓扑结构上的差异（GO 注释整合自 TAIR10 和 EBI,版本 4/9/2013）。图中的虚线是从发育子网络中抽样 10000 次计算的全局拓扑结构参数的分布情况，蓝色竖线则是应激子网络中相应参数的值。图中标注的 *P* 值是基于 10000 次抽样计算的。

在上面识别发育子网络和应激子网络时，没有考虑同时参与发育过程和应激过程的基因。如果将这些基因考虑在内，结果会怎样？在将这些基因考虑在内后确定发育子网络和应激子网络并重复上述抽样检查，得到的结果与以前结果一致（图 4-9）。

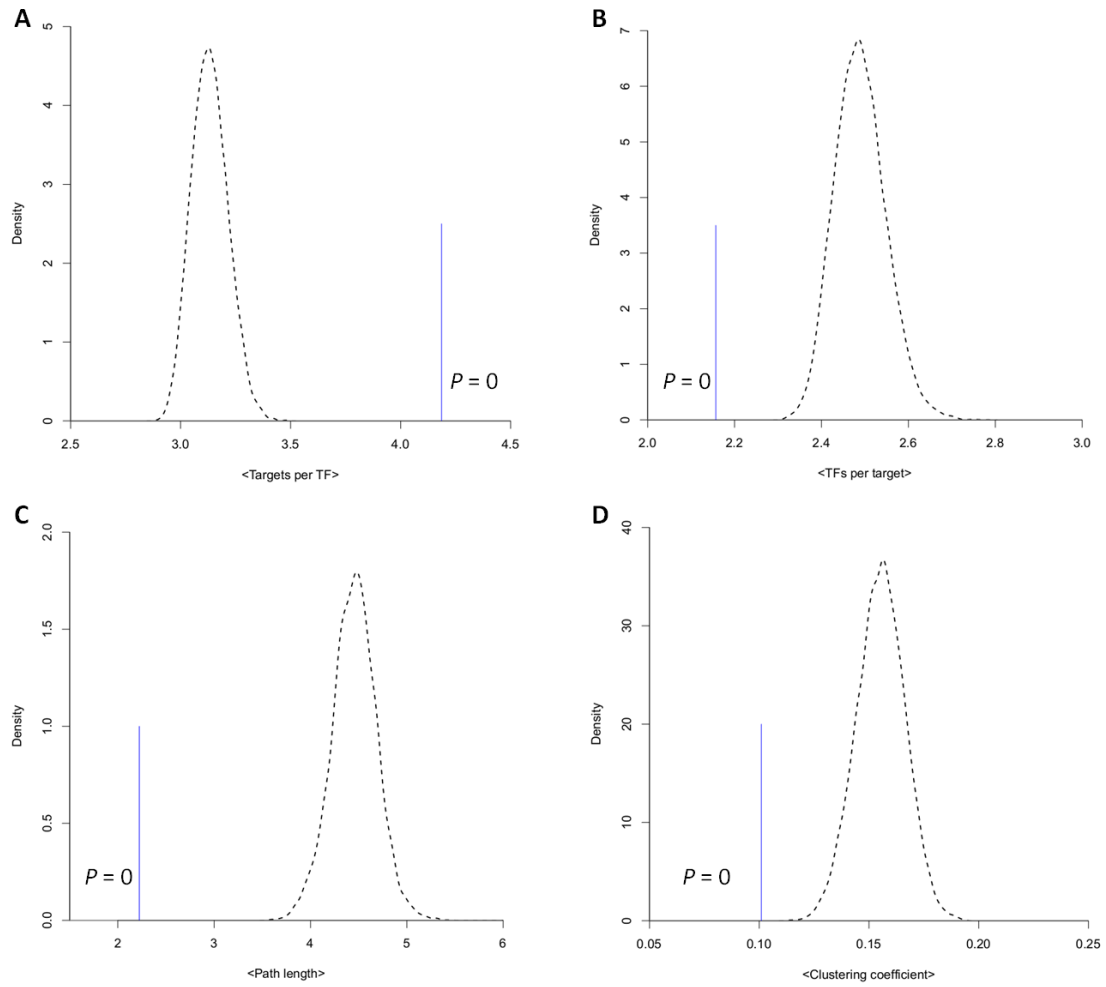

图 4-9　抽样检测发育子网络和应激子网络在全局拓扑结构上的差异（将既参与发育过程又参与应激过程的基因考虑在内）。图中的虚线是从发育子网络中抽样 10000 次计算的相应拓扑结构参数的分布情况，蓝色竖线则是应激子网络中相应参数的值。图中标注的 *P* 值是基于 10000 次抽样计算的。

在预测的拟南芥转录调控网络中的情况

使用下面的方法，我们预测了一个拟南芥的转录调控网络：（1）使用 Match 程序和 TRANSFAC 中的转录因子结合矩阵，在拟南芥基因转录起始位点（TSS）上游 1000bp 扫描相应的结合位点，对于上游拥有 2 个或者 2 个以上结合位点的基因作为其候选靶基因用于接下来的分析。（2）通过表达相关性过滤表达相关性太低的调控关系。所用的标准分别为转录因子与候选靶基因间的表达相关性不小于 0.30、0.35 和 0.40（表 4-11）。由于已确定参与发育过程和应激过程并在 TRANSFAC 中具有结合矩阵的转录因子比较少，很难通过上面的 4 个参数来衡量它们的全局拓扑结构。前面分析显示发育过程的转录因子具有更少的靶基因和更加复杂的转录因子和转录因子之间的调控，所以我们主要查看了它们在这些方面的差异。结果

显示发育子网络的转录因子具有较少的靶基因和更高比例的转录因子和转录因子之间的调控，这与前面揭示的模式在说明的问题上是一致的（表 4-9）。

表 4-11 预测的拟南芥转录调控网络中的情况

PCC[a]	发育子网络				应激子网络			
Threshold	TF	Edge	<Targets per TF>	TF-TF(%)[b]	TF	Edge	<Targets per TF>	TF-TF(%)
0.30	10	235	23.5	28.9	12	449	37.4	21.6
0.35	9	122	13.6	36.1	11	292	26.5	23.0
0.40	8	54	6.8	40.7	10	181	18.1	26.5

[a] PCC: Pearson 相关系数是从 ATTED-II 上下载的

[b] TF-TF(%):转录因子之间的调控所占的比例

以上几方面的分析结果说明发育系统和应激系统在全局拓扑结构上的差异是显著和稳定的。在酿酒酵母中的研究发现其内源系统和外源系统在全局拓扑结构上具有显著的差异(Luscombe, *et al.*, 2004)，我们的结果与其较为相似。结合上一节揭示的它们在网络结构元件组成上的差异，我们的结果充分说明发育系统通过更加复杂的调控（涉及更多转录因子和转录因子间的调控和更长的调控路径）来精确的调控发育状态的维持和转换，而应激系统则倾向于使用更具影响力的转录因子通过简单的调控在短时间内改变更多基因的转录状态来快速响应外部的各种生物和非生物胁迫。

4.4参与发育系统和应激系统构建的转录因子在性质上的差异

4.4.1 转录因子的调控特异性

无论在结构元件的组成上还是网络的全局拓扑结构上，发育系统和应激系统都存在明显的差异。那么参与发育系统和应激系统构建的转录因子在性质上又是否有所不同呢？

转录因子通过结合序列上相似的特定顺式元件来调控下游靶基因的转录。通过其多个结合位点可构建出这个转录因子的结合矩阵来代表它的结合谱。结合矩阵的信息量（Information content, IC）则可以衡量它们的结合谱与随机序列之间的区分度(Schneider, *et al.*, 1986, Hertz, *et al.*, 1999)。在 TRANSFAC (2011 专业版)中收录了 142 个植物转录因子相关的结合矩阵。参考文献 (Hertz, *et al.*, 1999) 中的方法，使用公式 4-6 来计算转录因子结合矩阵的信息量。

由于在构建结合矩阵的序列数上的不同可能会给信息量的计算带来某些偏差，通过根据序列数不同引入不同伪数（pseudocount）的方式来尽量避免这一情况。通过使用一个转录因子具有两个或者两个以上结合矩阵的情况进行评估，公式 4 中的 k 值可确保使用不同序列数构建的矩阵在信息量的计算上没有系统性的偏差。

$$k = \begin{cases} 0.1s & if\ s \leq 10 \\ 1 + 0.02(s - 10) & if\ 10 < s\ \leq 20 \\ 1.2 + 0.005(s - 20) & if\ s\ > 20 \end{cases} \quad (4)$$

$$f_{i,j} = \frac{n_{i,j} + p_i k}{\sum_{i=1}^{4} n_{i,j} + k} \quad (5)$$

$$I = \sum_{j=1}^{w} \sum_{i=1}^{4} f_{i,j} \ln\left(\frac{f_{i,j}}{p_i}\right) \quad (6)$$

在公式 4-6 中，k 是根据构建矩阵所用的序列数 s 确定的伪数(pseudocount)；$f_{i,j}$ 是校正后的核苷酸 i 在位置 j 上的频率；$n_{i,j}$ 是在结合矩阵中核苷酸i 在位置 j 上频数；p_i 为核苷酸 i 的先验概率；w 是结合矩阵的长度；I 是最终计算的结合矩阵的信息量。

我们只选取 TRANSFAC 中高质量的转录因子结合矩阵进行下面的研究，并使用下面的方法将 TRANSFAC 中的 ID 与 TAIR 的 ID 对应起来以尽可能的使用现有数据: 1)对于 TRANSFAC 中拥有通用名的拟南芥转录因子，则使用通用名直接对应；2)对于其它物种的转录因子或者使用通用名无法与拟南芥中的转录因子直接对应的，我们使用 BLAST 构建其与拟南芥转录因子的一对一关系；3)对于有两个或者两个以上结合矩阵的转录因子，只保留拥有最多序列数的结合矩阵。由于 RAV 家族的 AT1G13260 具有两种不同类型的 DNA 结合结构域，我们将相应的两个结合矩阵结合起来计算其信息量。各转录因子、结合矩阵及其信息量见附录 4。

为评估上述方法计算的信息量能否正确代表其调控的特异性，我们使用 Match 工具和转录因子结合矩阵在拟南芥基因转录起始位点上游 1000bp 扫描结合位点，进而统计预测的下游靶基因的数目。为了将预测的假阳性和假阴性综合起来降到最低，我们使用了 TRANSFAC 中预先为每个结合矩阵设定的阈值。转录因子结合矩阵的信息量和预测的靶基因数目见附录 5。通过使用 Spearman 秩相关检验衡量结合矩阵的信息量与预测的靶基因数之间的相关性，我们发现二者之间具有高度

的负相关性（ρ = -0.76 和 *P* = 7.72e-15）。这也说明使用上述方法计算的信息量能很好的代表转录因子调控的特异性。

4.4.2 参与发育系统和应激系统构建的转录因子在调控特异性上的差异

通过比较参与发育过程和应激过程的转录因子结合矩阵的信息量，我们发现参与发育过程的转录因子具有更高的信息量（单侧 Wilconxon 秩和检验 *P*=0.04; 图 4-10），即更高的调控特异性。这个结果与上面全局拓扑结构参数<Targets per TF>揭示的结果是一致的，反映了构建发育系统和应激系统的转录因子在调控特异性上的不同。考虑到上述两个系统在网络结构方面的显著差异，这个结果暗示转录因子的调控特异性可能与它们参与构建的网络之间存在着某种关联。

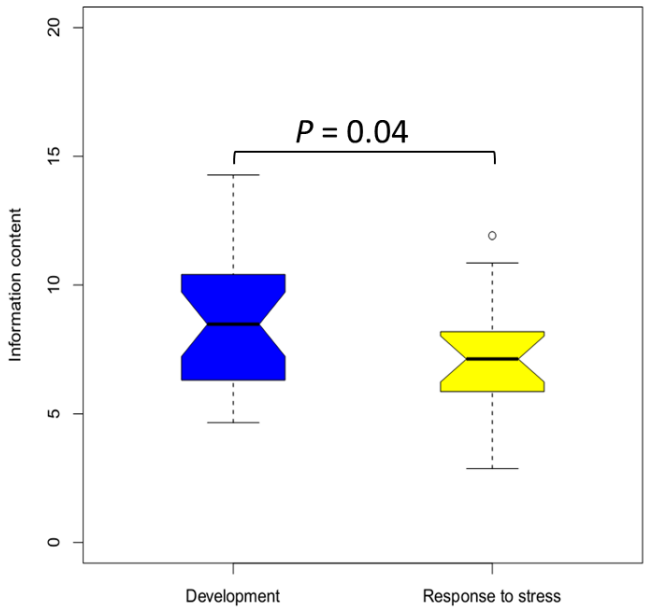

图 4-10 参与发育系统和应激系统构建的转录因子在结合矩阵信息量（IC）上的比较

4.5本章小结

使用拟南芥转录调控网络 ATRM，我们系统研究了一个植物的转录调控网络是如何构建的。在 ATRM 中，有 5 个 3 节点的结构元件。与单细胞的大肠杆菌和酿酒酵母相比，拟南芥有 3 个新结构元件(Motif 10、Motif 11 和 Motif 12)，其中的 2 个结构元件也在人的转录调控网络中富集。通过对参与结构元件构建的转录因子做 GO 富集分析和分析这些元件参与构建的网络，我们发现这些新结构元件主要

参与构建复杂的发育过程网络。随后的动力学模拟也说明这些元件能完成状态的维持与转换，这些功能是多细胞发育过程尤其是细胞的命运决定中所需要的。

通过将 ATRM 划分为发育子网络和应激子网络，我们发现发育系统和应激系统在结构元件组成、全局网络拓扑结构以及参与构建它们的转录因子的调控特异性上都有明显的不同。这些结果显示发育系统通过更加复杂的调控（涉及更多的转录因子间的调控和更长的调控路径）来精确调控发育状态的维持和转换，而应激系统则更倾向于使用更具影响力的转录因子通过更加简单的调控在短时间内启动大量基因的转录来快速响应外部的各种生物和非生物胁迫。这也体现了转录调控网络构建中的设计原则，通过特定的结构形式的网络来完成相应的功能。

第 5 章 转录因子在参与转录调控系统构建中的倾向性

5.1概述

5.1.1 植物登陆——植物演化历程中的大事件

植物为动物提供了食物来源和栖息场所，在生物界起着举足轻重的作用。从水生藻类到陆生植物的演化是植物演化历程中的重大事件。为了占领陆地，陆生植物需要适应与水中截然不同的陆地环境，包括温度的巨大波动、干旱以及暴露在强烈的辐射下。除了生境上的剧烈改变，陆生植物有了更加复杂的组织和器官的分化。

转录因子通过在恰当的时间和地点开启或者关闭相应基因的转录，在植物发育和应对各种生物和非生物胁迫中发挥着关键的作用。目前研究表明转录因子及转录调控网络的演化在植物形态的改变和对特定生境的适应中发挥着至关重要的作用(de Bruijn, *et al.*, 2012)。在植物登陆期间产生了很多新类型的转录因子(Lang, *et al.*, 2010, Zhang, *et al.*, 2011)，它们与古老类型的转录因子共同构建了一个新的转录调控系统以调控更加复杂的发育过程和应对剧烈改变的环境。

5.1.2 本章问题的提出

陆生植物既要精确调控复杂的组织和器官的发育又要快速响应外界的各种胁迫。上一章的结果显示植物的这两大部分在结构元件组成和全局拓扑结构方面都有明显的差异。在植物从水生到陆生的演化历程中，产生了很多新类型的转录因子(Lang, *et al.*, 2010, Zhang, *et al.*, 2011)，但这些新类型转录因子和古老类型的转录因子是如何参与转录调控系统构建的以及什么因素在其中起作用目前都尚不清楚。

本章将针对以下问题进行探讨，包括在植物登陆期间有哪些新类型的转录因子产生、新类型和古老类型的转录因子是如何参与转录调控系统构建的、新类型的转录因子具有什么性质以及为什么会是这样等。

5.2在植物登陆期间产生了 19 个新的转录因子家族

根据特征结构域（绝大多数是 DNA 结合结构域）的不同，转录因子可以划分为不同的家族(Riechmann, *et al.*, 2000, Zhang, *et al.*, 2011)。在植物演化历程中，通过产生新的特征结构域、特征结构域与古老类型发生实质性的分化（从序列上已无法分辨它们之间的同源关系）、或者通过结构域重组产生新的组合等方式产生了很多新的转录因子家族(Riechmann, *et al.*, 2000)。

PlantTFDB 2.0 中收录了从 28 个具有基因组序列的物种（其中 9 个为绿藻，19 个为陆生植物）中系统识别的转录因子全谱。根据特征结构域的不同，PlantTFDB 2.0 将它们划分为 58 个家族(Zhang, *et al.*, 2011)。根据各家族出现时间的不同，我们将陆生植物最近共同祖先（MRCA）中存在的 54 个家族划分为两类：古老类型、新类型。其中古老类型的家族为绿藻中已存在的 35 个家族（至少存在于 9 种绿藻中的一种）；新类型的家族是指存在于 19 个陆生植物的最近共同祖先中而在 9 种绿藻中没有的家族。

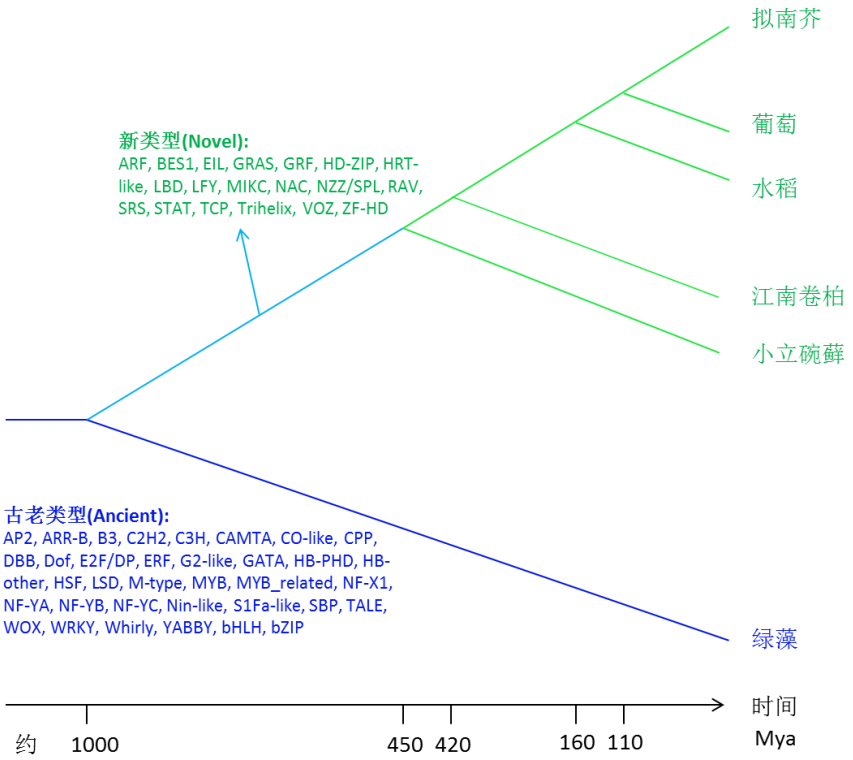

图 5-1　在植物登陆期间产生了 19 个新的转录因子家族

如图 5-1 所示，在植物登陆期间（特指图 5-1 中用青色标记的时间段，下同）共产生了 19 个新的转录因子家族。由于通过产生新的特征结构域和特征结构域已

与古老类型发生实质性分化这两种方式在序列上难于区分，我们将这两种方式合并在一起并标记为“新的特征结构域”。通过结构域重组产生的新类型标记为“新的组合方式”。这 19 个新家族的产生方式见表 5-1。

表 5-1　19 个新家族的产生方式

方式	家族
新的特征结构域	BES1, EIL, GRAS, GRF, HRT-like, LBD, LFY, NAC, NZZ/SPL, SRS, STAT, TCP, Trihelix, VOZ, ZF-HD
新的组合方式	ARF, HD-ZIP, MIKC, RAV

5.3转录因子在参与转录调控系统构建中的倾向性

5.3.1　新类型和古老类型转录因子在参与生物过程中的倾向性

为了探究新类型和古老类型的转录因子是如何参与植物转录调控系统两个主要部分（发育系统和应激系统）的构建，我们分别在家族水平和个体成员水平研究了它们参与以上两种生物过程构建的情况。

为便于在家族水平研究新类型和古老类型转录因子参与发育过程和应激过程的情况，我们定义了“倾向性指数”来代表一个家族参与生物过程的倾向性。首先，选取具有实验证据支持的（参考 GO 注释）参与发育过程或应激过程的转录因子，使用 BLASTClust (Altschul, *et al.*, 1997)将高度冗余的基因聚为一簇。BLASTClust 覆盖度（coverage）和一致度(identify)的阈值均取 0.9。聚在一起的每簇基因都作为一个参考基因(Refgene)用于下面的分析（下面所说的个体水平分析使用的也是此参考基因）。然后取明确参与发育过程或应激过程的参考基因和具有至少 5 个此类参考基因的家族计算“倾向性指数”。最终参与发育过程的“倾向性指数”指该家族中参与发育过程的参考基因所占的比例，对于应激过程的“倾向性指数”也是一样的。各家族参与发育过程和应激过程的倾向性如图 5-2 所示。图中的每个点都代表一个家族，为了避免有些点重合在一起，使用 R 中的 jitter 功能对点的位置进行了微调。与古老类型的家族（图中红色的点）相比，新类型的家族（图中青色的点）更倾向于参加发育过程（单侧 Wilcoxon 秩和检验 $P = 0.04$）。

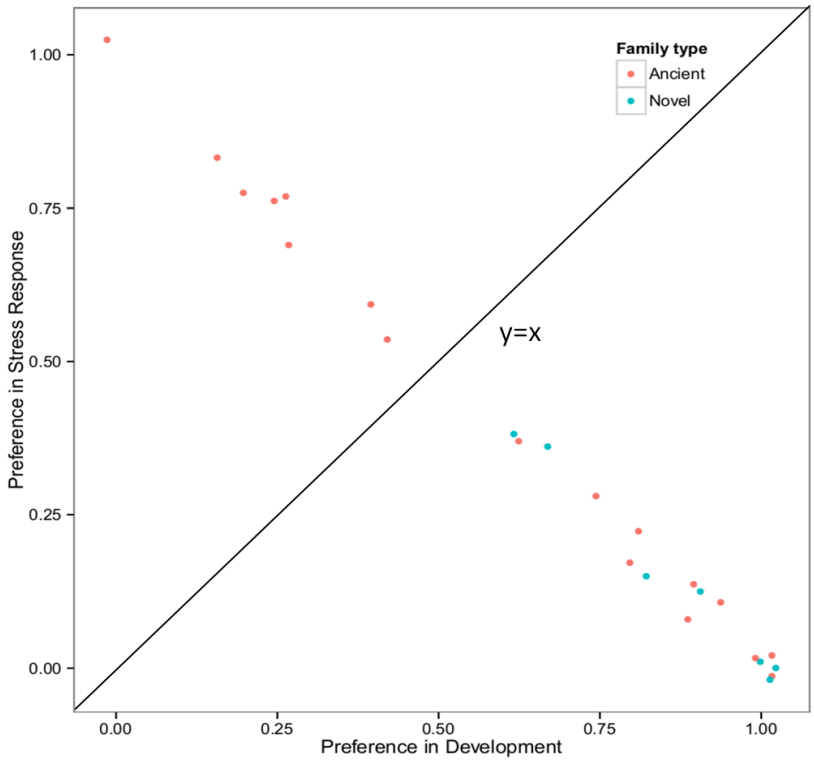

图 5-2　新类型和古老类型家族参与发育过程和应激过程的倾向性

此外，我们还在个体水平检查了它们参与发育过程和应激过程的倾向性。与基于家族水平得到的模式一致，新类型转录因子在个体水平上也倾向于参加发育过程（单侧 Fisher's 精确检验 P = 1.36e-8；表 5-2）。

表 5-2　新类型和古老类型的转录因子参与生物过程的倾向性

类型	发育	应激
新类型	85	19
古老类型	157	150

单侧 Fisher's 精确检验: P = 1.36e-08, odds ratio= 4.26

上面结果显示新类型转录因子倾向于参与发育过程。为进一步查看其参与的具体生物过程，使用具有 GO 生物过程注释（实验证据支持的）的转录因子作为背景，对新类型转录因子作生物过程方面的富集分析。结果显示新类型转录因子富集在系统发育、器官发育等多细胞发育过程中（表 5-3）。

表 5-3 新类型转录因子富集的生物过程（前 20）

GO ID	GO term	*P* 值	调整后的 *P* 值
GO:0048731	system development	3.9e-14	3.6e-11
GO:0048856	anatomical structure development	5.8e-14	3.6e-11
GO:0048513	organ development	1.3e-12	5.4e-10
GO:0032502	developmental process	1.8e-11	4.5e-09
GO:0007275	multicellular organismal development	1.8e-11	4.5e-09
GO:0032501	multicellular organismal process	2.4e-11	5.0e-09
GO:0022621	shoot system development	1.4e-10	2.2e-08
GO:0048367	shoot development	1.4e-10	2.2e-08
GO:0048366	leaf development	5.5e-09	7.6e-07
GO:0048827	phyllome development	1.3e-08	1.6e-06
GO:0010016	shoot morphogenesis	1.9e-08	2.1e-06
GO:0050793	regulation of developmental process	2.2e-07	2.3e-05
GO:0009965	leaf morphogenesis	1.3e-06	0.0001
GO:0051093	negative regulation of developmental process	2.8e-06	0.0003
GO:0009908	flower development	3.0e-06	0.0003
GO:0009791	post-embryonic development	5.9e-06	0.0005
GO:0009653	anatomical structure morphogenesis	8.8e-06	0.0006
GO:0048608	reproductive structure development	9.0e-06	0.0006
GO:0061458	reproductive system development	9.0e-06	0.0006
GO:0045962	positive regulation of development, heterochronic	1.8e-05	0.0011

通过在家族水平和个体水平分析新类型和古老类型转录因子参与发育过程和应激过程的情况，发现新类型转录因子更倾向于参与发育过程的构建。进一步 GO 富集分析的结果则显示新类型转录因子在系统发育、器官发育等多细胞发育过程中发挥作用。与之相反，应激过程则主要由古老类型的转录因子构建，从基因数目的层面来讲占到了 89% (表 5-2)。

5.3.2 新类型和古老类型的转录因子在参与网络构建中的倾向性

上一节的分析结果显示新类型和古老类型的转录因子在参与发育过程和应激过程时具有明显的倾向性。由于发育过程和应激过程在转录调控网络的构造上存在明显差异，那么新类型和古老类型的转录因子在参与网络的构建上是否亦有所不同呢？

通过上一章的分析，我们得知拟南芥中除了具有单细胞生物存在的 2 种网络结构元件（Motif 5 和 Motif 6），还有 3 种新的网络结构元件（Motif 10、Motif 11 和 Motif 12）；既有简单的调控模式也有很多转录因子之间的复杂调控。为了弄清楚新类型和古老类型的转录因子是如何参与转录调控网络构建的，我们对其如何参

与上述两类结构元件的构建以及调控什么等方面展开了研究。为了分析它们调控什么，我们按照靶基因的功能将其划分为转录因子和非转录因子两类来统计靶基因中转录因子所占的比例。

在分析之前，首先去掉 ATRM 中的自调控关系，然后选取至少具有 4 个调控关系的转录因子（4 为 ATRM 中转录因子具有连接数的中位数）用于本节的研究。在统计转录因子靶基因中转录因子所占的比例时则选取至少具有 4 个靶基因的转录因子用于此项分析。最后，在五个方面比较了新类型和古老类型转录因子在 ATRM 中的具体情况。其中<Targets per TF>和<TFs per target>两项用于检查这两类转录因子在连接数上是否有显著的偏差，参与构建 Motif(5,6)和 Motifs(10，11，12)的数目、靶基因中转录因子的比例则用于分析转录因子在网络中的位置。

在上述五个方面，首先计算出相关项的平均值，然后分别统计出新类型和古老类型中大于和小于此数值的转录因子数目，最后使用单侧 Fisher’s 精确检验检测新类型与古老类型的转录因子在网络构建中是否有明显的差别，结果见表 5-4。

表 5-4　新类型和古老类型的转录因子参与网络构建的情况

	Targets per TF		TFs per target		Motifs (5, 6)		Motifs (10, 11, 12)		TF rate of targets	
Mean	7.13		2.76		5.13		1.46		0.42	
</> mean	<	>	<	>	<	>	<	>	<	>
Novel TF	33	17	29	21	41	9	33	17	12	30
Anicent TF	79	33	68	44	75	37	94	18	57	33
Odds ratio	0.81		0.89		2.24		0.37		0.23	
P[a]	0.34		0.44		0.04		0.01		1.80e-04	

[a] One-tailed Fisher's exact test

从表 5-4 可以看出，新类型和古老类型的转录因子在前两项上没有显著差异，说明在 ATRM 中这两类转录因子在调控关系的数目上没有明显的不同。在参与网络构建方面，与古老类型的转录因子相比，新类型的转录因子更倾向于调控转录因子和参与较复杂的新结构元件的构建（Motif 10、11 和 12）。

5.4转录因子的性质与参与网络构建的倾向性

5.4.1　新类型和古老类型转录因子在调控特异性上的差异

参与发育系统和应激系统构建的转录因子在调控特异性方面存在显著的差异，而新类型和古老类型的转录因子在参与发育系统和应激系统的构建上又有明显的倾向性。那么新类型和古老类型的转录因子在调控特异性上是否亦有所不同呢？

使用先前计算的转录因子结合矩阵的信息量，我们比较了新类型和古老类型转

录因子的调控特异性。结果发现新类型转录因子具有更高的信息量（图 5-3），即具有更高的调控特异性。

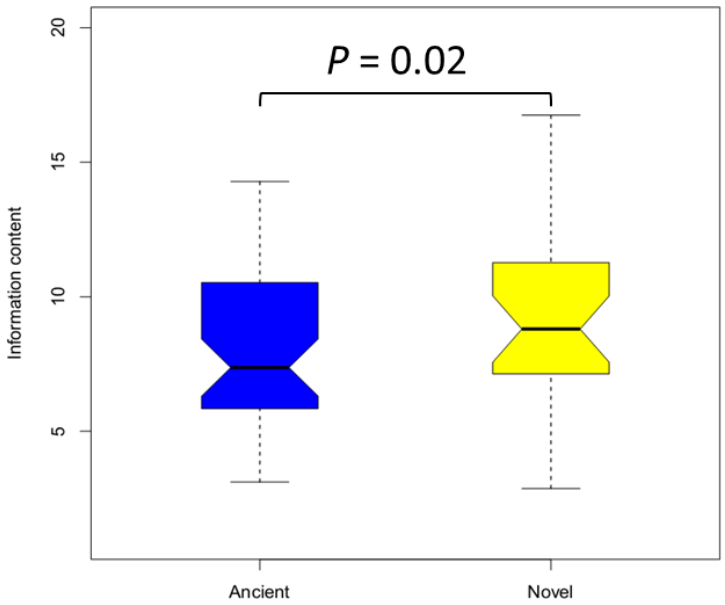

图 5-3 新类型和古老类型转录因子结合矩阵的信息量(IC)，图中标注的 *P* 值是单侧 Wilcoxon 秩和检验的结果。

使用至少有 3 个转录因子具有结合矩阵的家族，我们在家族水平也比较了新类型和古老类型转录因子的调控特异性，结果如表 5-5 所示。按照结合矩阵信息量的高低，将各家族分为高信息量（大于中位数 7.85）和低信息量（小于中位数 7.85）两类，它们在信息量高低中的分布情况见表 5-5B。这些结果表明在家族水平上新类型转录因子也具有比古老类型更高的调控特异性。

表 5-5：拟南芥转录因子家族结合矩阵的信息量。（A）为各家族的类型及信息量的中位数，（B）为新类型和古老类型家族在信息量高低上的分布。

A

家族	类型	信息量的中位数
Dof	古老类型	5.33
ERF	古老类型	6.27
HD-ZIP	新类型	12.67
MIKC	新类型	8.92
MYB	古老类型	8.17
NAC	新类型	5.86
TCP	新类型	7.96
WRKY	古老类型	7.74
bHLH	古老类型	8.89
bZIP	古老类型	7.34

B

	低信息量	高信息量
新类型	1	3
古老类型	4	2

抽样估计: Odds ratio = 0.20

5.4.2 转录因子的调控特异性与参与网络构建的倾向性

新类型的转录因子具有更高的调控特异性，而前面的分析暗示转录因子的调控特异性与其参与构建的网络之间可能存在某种内在的联系（详见 4.4.2）。那么新类型转录因子的高调控特异性是否与其参与网络构建的倾向性有关呢？

一个转录因子去调控什么与它在网络中的位置相关。为研究转录因子的调控特异性与它调控什么是否具有某种内在联系，我们将靶基因按照调控功能分为两类：转录因子和非转录因子。然后使用 Spearman’s 秩相关检验检测转录因子结合矩阵的信息量与相应靶基因中转录因子所占的比例之间是否存在相关性。结果显示二者之间存在显著的正相关性（$\rho = 0.46$ 和 $P = 0.02$；表 5-6），说明转录因子的调控特异性越高，越倾向于调控转录因子。

表 5-6 拟南芥中转录因子的调控特异性与靶基因中转录因子所占比例的关系

转录因子	信息量	靶基因中转录因子所占比例 (%)
AT1G09530	10.85	16.67
AT1G09770	7.86	33.33
AT1G19850	6.75	37.5
AT1G24260	9.84	100
AT1G45249	10.86	0
AT1G75080	6.12	20
AT2G16910	5.33	0
AT2G20180	6.94	30.23
AT2G36010	7.67	25
AT2G38470	7.13	55.56
AT2G40220	14.15	42.86
AT2G46830	12.04	28.57
AT2G47460	6.12	0
AT3G20770	5.86	25
AT3G27920	10.98	100
AT3G56400	8.52	12.5
AT4G18960	9.06	78.57
AT4G25490	3.11	21.43
AT4G31550	6.38	60
AT4G37750	11.03	90.91
AT4G38620	4.43	0
AT5G11260	7.2	18.18
AT5G13790	8.92	33.33
AT5G20730	6.75	46.15
AT5G41315	10.98	70
AT5G62000	6.75	16.67

为了深入研究转录因子的调控特异性与它们参与网络构建的情况，我们统计了转录因子结合矩阵的信息量与其参与结构元件构建的关系，结果如表 5-7 所示。为便于比较分析，根据转录因子结合矩阵信息量的高低将转录因子划分为高信息量（高于中位数 7.67）和低信息量（低于中位数 7.67）两类。表中的度（Degree）指这个转录因子在 ATRM 的连接数，Motifs（5，6）和 Motifs（10，11，12）则指该转录因子参与构建的这两类结构元件的数目。通过使用单侧 Wilcoxon 秩和检验检测转录因子调控信息量的高低与其参与结构元件构建的关系，我们发现在它们的连接数没有显著区别的情况下（Degree: $P = 0.456$），信息量高的转录因子倾向于参与到更多新类型结构元件（Motif 10、11 和 12）构建中（$P = 0.005$）。

表 5-7　转录因子结合矩阵的信息量与参与网络构建的情况

TF id	Degree	Motifs (5, 6)	Motifs (10, 11, 12)	IC	Type
AT1G09530	12	2	0	10.85	high
AT1G09770	6	4	0	7.86	high
AT1G19850	8	0	0	6.75	low
AT1G24260	13	2	11	9.84	high
AT1G45249	6	1	0	10.86	high
AT1G75080	5	0	0	6.12	low
AT2G16910	16	1	0	5.33	low
AT2G20180	43	9	0	6.94	low
AT2G36010	5	3	0	7.67	-
AT2G38470	9	3	0	7.13	low
AT2G40220	9	9	0	14.15	high
AT2G46830	25	19	1	12.04	high
AT2G47460	5	2	0	6.12	low
AT3G20770	12	1	0	5.86	low
AT3G23250	4	5	0	11.92	high
AT3G27920	11	7	1	10.98	high
AT3G56400	10	2	0	8.52	high
AT3G58780	7	3	0	8.69	high
AT4G08150	9	3	0	5.13	low
AT4G18960	34	9	25	9.06	high
AT4G25490	19	16	0	3.11	low
AT4G31550	6	4	0	6.38	low
AT4G34000	4	1	0	10.86	high
AT4G37750	16	7	5	11.03	high
AT4G38620	5	0	0	4.43	low
AT5G11260	22	2	0	7.2	low
AT5G13790	10	0	0	8.92	high
AT5G20730	13	5	0	6.75	low
AT5G22220	5	3	0	7.67	-
AT5G41315	13	9	1	10.98	high
AT5G62000	6	0	0	6.75	low

以上结果说明转录因子的调控特异性与其在网络中的位置有关，调控特异性高的转录因子更倾向于调控转录因子和参与单细胞中没有的新结构元件的构建。

5.4.3　其它生物界中转录因子调控特异性与参与网络构建的关系

在拟南芥中，转录因子的调控特异性与其在网络中的位置有关。这一模式只局限于拟南芥中还是普遍存在于各个生物界？为了弄清这一问题，我们收集了大肠杆菌（*E. coli*，细菌）、酿酒酵母（*S. cerevisiae*，真菌）和人（*H. sapiens*，后生动物）中的转录因子结合矩阵和转录调控网络，并分析了在这些生物中的情况。

大肠杆菌、酿酒酵母和人的转录调控网络数据分别是从 RegulonDB 8.0 (Salgado, *et al.*, 2013)、YEASTACT (Abdulrehman, *et al.*, 2011)和 ENCODE 项目 (Gerstein, *et al.*, 2012)下载的。大肠杆菌的转录因子结合矩阵是从 RegulonDB 下载的。人的转录因子结合矩阵是从 TRANSFAC (2011 专业版)下载的，选取使用 SELEX 方法确定的矩阵用于下面的分析。对于大肠杆菌和人，转录因子结合矩阵的信息量用于代表转录因子的调控特异性，结合矩阵的信息量是使用先前介绍的方法计算的。对于酿酒酵母，当前数据库中收录的结合矩阵在信息量上区分度很低，而基于基因组范围识别的转录调控数据则非常丰富。因此，直接使用 ChIP-chip 实验确定的靶基因的数目代表酿酒酵母中转录因子的调控特异性，而功能确定的转录调控数据则用于计算靶基因中转录因子的比例。Ecocyc & Ecoli Hub（3/15/2013）、SDG（4/43/2013）和 EBI （4/15/2013）中的 GO 注释分别用于识别大肠杆菌、酿酒酵母和人中的转录因子。

大肠杆菌、酿酒酵母和人中转录因子调控特异性和靶基因中转录因子的比例分别见附录 6、附录 7 和附录 8。通过使用 Spearman's 秩相关检验，发现在这三个物种中调控特异性和靶基因中转录因子的比例都有显著的相关性（大肠杆菌：$\rho = 0.31$ 和 $P = 0.03$；酿酒酵母：$\rho = -0.36$ 和 $P = 0.0003$；人：$\rho = 0.47$ 和 $P = 0.009$）。由于酿酒酵母中使用 ChIP-chip 实验确定的靶基因数代表调控特异性，因此在大肠杆菌、酿酒酵母、人中发现的模式与拟南芥的模式是一致的：转录因子的调控特异性越高，越倾向于调控转录因子。

5.4.4 讨论

通过在拟南芥（植物）、大肠杆菌（细菌）、酿酒酵母（真菌）和人（后生动物）中的分析，我们发现了一个在生物界中转录调控网络构建的普遍模式，揭示了转录因子的调控特异性与其在网络中的位置之间的内在联系。鉴于转录调控网络的高重搭建（rewring）速率(Borneman, *et al.*, 2007, Shou, *et al.*, 2011)和 DNA 结合结构域的高度保守性(Riechmann, *et al.*, 2000)，此模式说明转录因子在网络中的位置已经收到自然选择的作用。转录因子通过长期的演化可能已经连接到一个适合它本身特点的位置上。这个各生物界转录调控网络构建的通用模式也在一定程度上解释了新类型转录因子和具有高调控特异性的转录因子为什么更倾向于参与构建发育网络中新结构元件和转录因子与转录因子之间的复杂调控。

5.5新类型转录因子参与生物系统倾向性的其它可能模型

为什么新类型的转录因子倾向于参与发育过程而不是应激过程？除了上面提到的新类型转录因子的高调控特异性外，诸如转录因子成员的非对称复制、植物登陆期间对发育系统的定向选择、后产生的转录因子成员参与网络构建的倾向性等原因都可能导致新类型转录因子倾向于参与发育系统。下面将分别讨论这三个方面的可能性。

5.5.1 转录因子成员的非对称复制

如果参与到发育系统中的新类型转录因子具有更高的复制率会导致在统计时发现新类型转录因子更倾向于参与发育系统。为避免这一因素造成统计上的偏差，我们选取了拟南芥演化历程中的四个重要时间点，将在所选时刻属于同一共同祖先的转录因子合并为一个参考基因来讨论它们参与生物过程的倾向性。

如图 5-4 所示，这四个关键时间点分别是拟南芥与小立碗藓（*P. patens*）的最近共同祖先（MRCA）、拟南芥与水稻（*O. Sativa*）的 MRCA、拟南芥与葡萄（*V. vinifera*）的 MRCA、拟南芥与琴叶鼠耳芥(*A. lyrata*)的 MRCA。

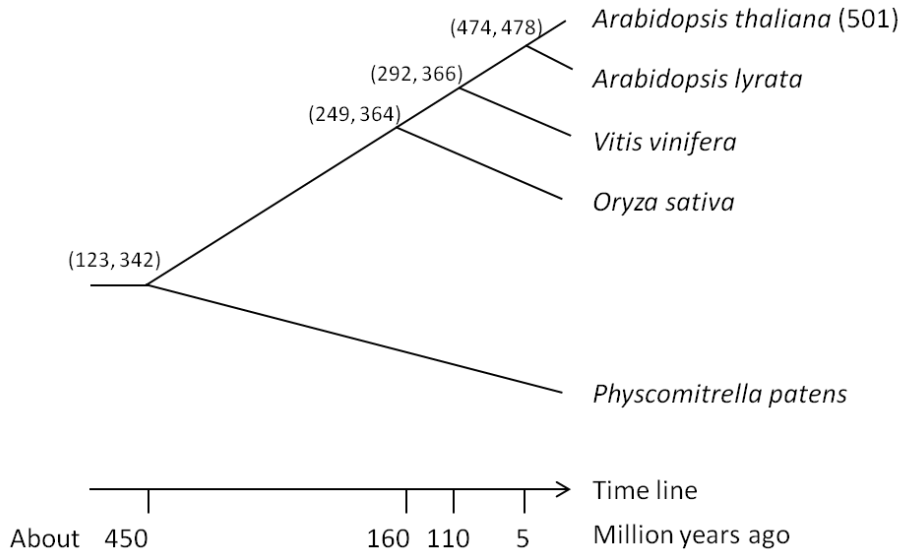

图 5-4 四个用于合并从同一共同祖先演化而来的拟南芥转录因子的关键时间点

表 5-8 新类型和古老类型转录因子参与发育过程和应激过程的情况。表中的数据分别为未聚类之前（A），使用琴叶鼠耳芥（B）、葡萄（C）、水稻（D）和小立碗藓（E）中的直系同源基因聚类后的数据。

A

类型	发育过程	应激过程
新类型	109	24
古老类型	187	181

单侧 Fisher's 精确检验: *P* = 8.69e-11, odds ratio= 4.38

B

类型	发育过程	应激过程
新类型	98	24
古老类型	179	173

单侧 Fisher's 精确检验: *P* = 4.03e-09, odds ratio= 3.94

C

类型	发育过程	应激过程
新类型	70	17
古老类型	106	100

单侧 Fisher's 精确检验: *P* = 1.79e-06, odds ratio= 3.87

D

类型	发育过程	应激过程
新类型	54	9
古老类型	95	91

单侧 Fisher's 精确检验: *P* = 4.27e-07, odds ratio= 5.71

E

类型	发育过程	应激过程
新类型	26	7
古老类型	49	43

单侧 Fisher's 精确检验: *P* = 0.008, odds ratio= 3.23

基于 Ensembl Compara 流程构建的系统发生树(Vilella, *et al.*, 2009)，Ensembl Plants 识别了多个植物物种之间的直系同源关系(Kersey, *et al.*, 2012)。使用从 Ensembl Plants (版本号：15)中的下载的拟南芥与琴叶鼠耳芥、葡萄、水稻和小立碗藓的直系同源关系，将拟南芥中从上述时间点的共同祖先演化出来的转录因子合并为一个参考基因（在此分析中使用的是具有实验证据支持的参与发育过程或者应激过程的转录因子）。图 5-4 中括号中的数字分别为那个时期涉及到的共同祖先数目及其包含的拟南芥转录因子数。

通过在拟南芥演化历程中的四个重要时间点将来源于同一祖先的基因聚在一起，我们使用单侧 Fisher’s 精确检验检查了新类型和古老类型转录因子参与发育过程和应激过程的情况。与没聚类之前一样，新类型转录因子仍然倾向于参与发育过程（表 5-8）。这些结果一方面说明转录因子成员的非对称复制不能解释新类型转录因子倾向于参与发育过程，另一方面说明无论是在现在还在拟南芥的演化历程中，新类型转录因子都是倾向于参与发育过程的。

5.5.2 植物登陆期间对发育系统的选择压力

由于植物登陆后具有更加复杂的发育系统，是否由于这个时间段（本文所指的植物登陆期间特指图 5-1 青色标记的时间段）对植物发育系统的选择导致新类型转录因子倾向于进入发育过程？在本节，我们通过查看其它时间段产生的转录因子家族的情况和这一时间段产生的转录因子成员的情况来评估其影响。

基于 PlantTFDB 2.0 的转录因子预测流程，使用一个放松的阈值（序列水平和结构域水平的阈值均为 0.01）在大肠杆菌、酿酒酵母和人中识别植物中存在的转录因子家族。根据古老类型的家族是否在以上物种（之一）出现，将古老类型的家族分为两类：Ancient1 和 Ancient2。如果在这些物种（之一）中出现，则属于 Anicent1，否则属于 Ancient2。然后基于前面计算的各家族的“倾向性指数”，发现其它时间段产生的家族（Ancient2）也比更加古老的古老类型转录因子（Ancient1）倾向于参与发育过程（图 5-5）。

在前面的分析中，我们已将拟南芥中的转录因子聚为拟南芥和小立碗藓 MRCA 中的 123 个参考基因。然后根据 EnsemblPlants (版本号：15)中的直系同源关系，我们找出在 *C. merolae*or 和 *C. reinhardtii* 中没有直系同源基因的参考基因作为在植物登陆期间诞生的基因。如果 *C. reinhardtii* 中有参考基因的直系同源基因，参考基因中与 *C. reinhardtii* 的直系同源基因具有最高序列相似性的基因作为在它们分歧前就有的基因，其它的参考基因作为在植物登陆期间产生的基因。然后我们统计了新类型转录因子和在植物登陆期间产生的古老类型转录因子参与发育过程和应激过程的情况，结果如表 5-9 所示。这些结果表明，与在这一期间产生的古老类型转录因子相比，新类型的转录因子仍然倾向于参与发育过程。

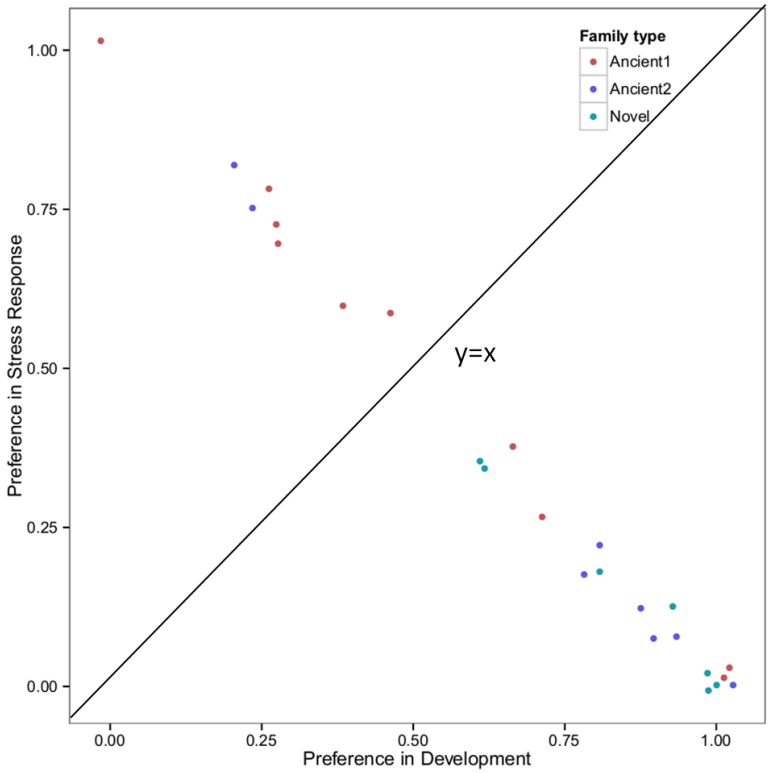

图 5-5 不同时间起源的转录因子参与生物过程的倾向性

表 5-9：新类型转录因子和植物登陆期间产生的古老类型中的新成员参与发育过程和应激过程的情况。(A)中的数据是聚类前的数据，（B）是使用小立碗藓中的直系同源基因聚成参考基因后的结果。

A

类型	发育过程	应激过程
新类型	77	19
古老类型	80	81
单侧 Fisher's 精确检验:	*P* = 6.64e-07, odds ratio= 4.08	

B

类型	发育过程	应激过程
新类型	25	6
古老类型	34	32
单侧 Fisher's 精确检验:	*P* = 0.007, odds ratio= 3.87	

以上结果说明这段时间对发育的可能选择压不能解释新类型转录因子倾向于参与发育过程。

5.5.3 后来产生的转录因子参与网络构建的倾向性

由于新类型转录因子的整体年龄比古老类型转录因子年轻，如果后面产生的转

录因子成员倾向参与发育过程，则可能是由于时间因素而不是由于转录因子的性质导致这种倾向性？使用上节提到的方法，我们将拟南芥和小立碗藓 MRCA 中的123个参考基因划分为在它们MRCA中就存在的老成员和分歧之后产生的新成员。新老成员参与发育过程和应激过程的情况见表 5-10。结果显示，与老成员相比，新成员在参与以上两个生物过程中并没有明显的倾向性。这也说明是由于新类型转录因子的性质而不是它的产生时间晚这一特点导致它们参与生物过程的倾向性。

表 5-10 新老转录因子成员参与发育过程和应激过程的情况

类型	发育过程	应激过程
新成员	123	95
老成员	76	46

单侧 Fisher's 精确检验: P = 0.87, odds ratio= 0.78

5.6其它生物界中新类型转录因子的性质及讨论

在细菌中的研究发现水平转移进来的转录因子通常处于更加复杂的调控之下，以尽量避免其对转入生物产生有害的影响(Price, *et al.*, 2008, Rajewsky, *et al.*, 2002, Perez, *et al.*, 2009)。新产生的 microRNA 也通常具有很低的表达水平(Lu, *et al.*, 2008)。这些研究都说明新的调控基因对原有系统的影响越小那么它可能就有更大的机会保留下来。这些结果提示新类型的转录因子具有更高的调控特异性可能是一个更通用的模式。

为了弄清这一问题，我们进一步研究了大肠杆菌（细菌）、酿酒酵母（真菌）和人（后生动物）等其它生物界中的情况。由于 DBD (Wilson, *et al.*, 2008)包含了横跨各生物界的转录因子全谱，因此使用其中包含的上述三个物种的转录因子列表（使用基于 Pfam HMM 识别的）用于下面的分析。根据 Charoensawan 等确定的各家族谱系分布局限 (Charoensawan, *et al.*, 2010)，在大肠杆菌、酿酒酵母和人中的新家族分别指分布谱系局限在变形菌（Proteobacteria）、真菌（Fungi）和后生动物（Metazoa）的转录因子家族。使用 DBD 中的转录因子列表，在以上三个物种识别的新类型和古老类型转录因子的数据统计见表 5-11。

表 5-11 大肠杆菌、酿酒酵母和人中新类型和古老类型转录因子的数据统计

物种		古老类型	新类型
大肠杆菌	家族数	30	8
	转录因子数	216	9
	具有调控特异性的转录因子数	64	1
酿酒酵母	家族数	23	4
	转录因子数	104	62
	具有调控特异性的转录因子数	56	21
人	家族数	32	24
	转录因子数	1234	211
	具有调控特异性的转录因子数	85	50

从表 5-11 可以看出，在单细胞的大肠杆菌和酿酒酵母中只有个别的新类型转录因子或新家族，而在多细胞的人中则有很多新类型的家族和转录因子。通过比较人中新类型和古老类型转录因子结合矩阵的信息量，发现新类型的转录因子具有更高的调控特异性（图 5-6）。接着我们在家族水平上也比较了新类型和古老类型转录因子的调控特异性，结果同样表明新类型比古老类型具有更高的调控特异性（单侧 Wilcoxon 秩和检验 $P = 0.036$；表 5-12）。

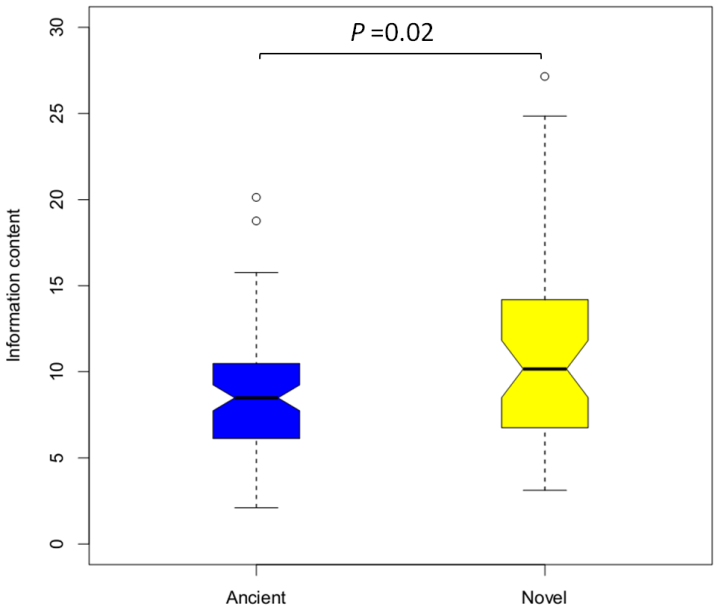

图 5-6 人中古老类型和新类型转录因子结合矩阵的信息量，图中标注的是单侧 Wilcoxon 秩和检验的 *P* 值。

表 5-12 人中各转录因子家族的调控特异性。至少有 3 个转录因子具有结合矩阵的家族用于本表的统计，表中的信息量是该家族各转录因子结合矩阵信息量的中位数。

家族	类型	信息量中位数
DM	新类型	10.16
Ets	新类型	8.64
Fork_head	古老类型	8.88
GATA	古老类型	6.18
HLH	古老类型	9.73
Homeobox	古老类型	8.34
PAX	新类型	8.48
Pou	新类型	12.48
STAT_bind	新类型	10.50
T-box	新类型	21.27
bZIP_1	古老类型	8.62
bZIP_2	古老类型	8.49
bZIP_Maf	新类型	6.44
zf-C2H2	古老类型	8.34
zf-C4	新类型	10.83

当新的转录因子类型整合入原有系统时，如果它具有较低的调控特异性，则有更大的几率导致有害的影响而被选择所淘汰。这可能是新类型转录因子具有更高调控特异性的一个原因。从某种角度来说，很多具有较高调控特异性的新类型转录因子加入到生物系统，这些转录因子又倾向于参与构建更加复杂的发育网络或许暗示着它们在生命从单细胞到多细胞演化中的重要作用。

5.7本章小结

通过识别在植物登陆期间产生的 19 个新的转录因子家族，我们系统研究了新类型和古老类型的转录因子是如何参与植物转录调控系统构建的。与古老类型的转录因子相比，新类型的转录因子具有更高的调控特异性，而且倾向于参与构建发育网络中新的结构元件和转录因子之间的复杂调控。

新类型转录因子的高调控特异性是影响其参与网络构建倾向性的因素吗？为弄清转录因子调控特异性与其参与构建什么网络之间的关系，根据靶基因的功能将其划分为转录因子和非转录因子两类。在拟南芥中的分析表明转录因子的调控特异性与调控的靶基因中转录因子所占的比例呈明显的正相关性，调控特异性越

高越倾向于调控转录因子。随后在大肠杆菌、酿酒酵母和人中的分析展现了与拟南芥中同样的模式，提示这一模式是转录调控网络构建的通用模式。这一模式也在一定程度上解释了为什么在拟南芥中新类型转录因子和具有高调控特异性的转录因子倾向于参与构建发育网络中新结构元件和转录因子之间的复杂调控。此外，我们还探讨了转录因子成员的非对称复制、植物登陆期间对发育的选择压力、后来产生的转录因子成员参与生物过程的倾向性等是否可能导致我们所观察的现象，结果表明它们不能解释或者不能完全解释这一现象。这些结果更加突出了转录因子本身的性质在新类型转录因子参与生物过程的倾向性中可能发挥的作用。

接下来探讨了为什么后来产生的新类型转录因子具有更高的调控特异性。当新类型转录因子整合入原有系统时，如果它具有较低的调控特异性，则有更大的几率导致有害的影响而被选择所淘汰。这可能是新类型转录因子具有更高调控特异性的一个原因。

综合上面发现的几个模式，从某种角度来说，很多具有较高调控特异性的新类型转录因子加入到生物系统，这些转录因子又倾向于参与更加复杂的发育网络的构建或许暗示着它们在生命从单细胞到多细胞演化中的重要作用。

第 6 章 总结和展望

6.1本文工作总结

通过在特定时间和地点调控相应靶基因的转录，转录因子在植物发育和应对胁迫中发挥着关键的作用。转录因子及转录调控网络的演化是植物形态改变和适应生境的重要因素。植物登陆期间产生了很多新的转录因子家族，这些新类型的转录因子与古老类型的转录因子共同构建了一个新的转录调控系统来调控更加复杂的多细胞发育过程和应对剧烈改变的生境。系统识别植物转录因子及其调控关系将有助于研究转录因子的功能、转录调控网络的构建原理和演化特征，进而揭示它们在形态改变和适应环境中的重要作用。本文针对植物转录因子的系统识别、植物转录调控网络的架构和演化特征两个方面展开探索研究并得到一些初步结果，分别总结如下：

6.1.1 植物转录因子的系统识别和注释

既然转录因子在植物发育和应对胁迫中起着这么重要的功能，植物体内到底有多少转录因子呢？它们分属多少家族？不同门类植物的转录因子谱又有何不同？

一个完整的蛋白组是系统识别植物转录因子的前提。通过反复比较分析，我们构建了一个数据整合流程为每个收录的物种（特别是没有基因组的物种）构建一个尽可能完整的蛋白组用于识别转录因子。

除了完整的蛋白组，系统识别植物转录因子还需要一套完整的转录因子分类规则。通过浏览七千余篇植物转录因子相关的文献，我们总结出一套完整的转录因子分类规则。该规则包含三种类型的结构域：DNA 结合结构域、辅助结构域和禁止出现的结构域。使用前两种结构域，我们从序列中识别出转录因子并将其划分到 58 个不同的家族。“禁止出现的结构域”则用于过滤包含 DNA 结合结构域而不具有转录因子活性的蛋白，以降低转录因子预测的假阳性。为了进一步提高预测的准确率，我们参考 GO 注释、Pfam、TAIR 和 UniProt 中的信息分别确定了每个特征结构域模型的阈值。

转录因子分类规则、特征结构域模型及其阈值构成了我们的转录因子预测流程。使用此预测流程，我们从 83 个物种中系统识别出 129288 个转录因子，其中 67 个物种具有基因组序列。由于具有基因组序列的物种覆盖了绿藻、苔藓、蕨类、裸

子植物和被子植物等绿色植物的各大分支，我们有幸提供了第一个覆盖绿色植物各大分支的转录因子全谱。与绿藻相比，陆生的高等植物在转录因子家族数、转录因子数和转录因子占基因组中所有基因的比例等方面都有了很大的提升。这或许与陆生高等植物具有更加复杂的形态发育相关。

为了方便用户使用识别出来的转录因子信息，我们先后构建了植物转录因子数据库 PlantTFDB 2.0 和 PlantTFDB 3.0。深知一个好的资源平台不是仅仅为用户提供一些转录因子列表，而且要提供注释帮助用户了解该转录因子的功能和研究现状并为下一步分析提供重要的线索。秉承这一原则，我们为识别的每个转录因子都做了详尽的注释。特别是在第三版中，更是从各大公共数据库系统收集了专家描述、表达信息、调控信息、相互作用信息、突变和表型信息、保守元件等知识型数据。这些注释将为用户研究转录因子的功能提供更加直接的线索。

除了个体水平的注释，我们还为每个物种的每个家族和所有物种的每个家族构建了系统发生树。为了在更细的精度上展示转录因子之间的演化关系，使用 67 个具有基因组序列的物种中的基因构建了直系同源群，还为每个直系同源群构建了系统发生树。这些系统发生树不但有助于用户研究转录因子之间的演化关系，也能为推测研究尚不清楚的转录因子的功能提供帮助。

此外，还搭建了应用程序接口 Web Service 和转录因子预测平台分别供用户批量获取 PlantTFDB 中的数据和从自己提供的序列中识别转录因子。

综上所述，PlantTFDB 包含了一套完整的转录因子预测流程、一个覆盖绿色植物各大分支的转录因子全谱、详尽的个体水平注释和演化注释以及一个植物转录因子预测的平台。它将为用户研究植物转录因子的功能和演化提供重要的资源。

6.1.2 拟南芥转录调控网络的架构和演化特征

在过去十数年，通过对大肠杆菌、酿酒酵母和人的转录调控网络进行系统研究，人们揭示了很多转录调控网络的构建原理和演化特征。不同于上述模式物种，植物既需要精确调控复杂的发育又需要快速响应外界的各种胁迫。缺少一个基因组范围的高质量的转录调控网络阻碍我们去理解植物转录调控系统的设计原理和演化特征。

通过系统识别和人工检查文献中报道的转录调控关系，我们构建了一个基因组范围的高质量的拟南芥转录调控网络 ATRM。使用该调控网络，我们揭示了一个植物物种内部转录因子和其靶基因在表达相关性上的总体模式。通过划分模块，

我们从 ATRM 中识别出 58 个特定生物过程的调控模块。这些模块内部和模块之间的调控为我们提供了生物过程内部和不同生物过程之间调控的概况。

通过研究大肠杆菌和酿酒酵母等单细胞生物的转录调控网络，人们发现转录调控网络是由一些结构元件组成的。这些结构元件同时也是能完成特定生物学功能的功能元件。通过系统识别 ATRM 中三节点的调控模式，我们在拟南芥的转录调控网络中识别出 5 种结构元件。与上述两种单细胞生物相比，拟南芥具有 3 种新的结构元件，其中的两个也在人的转录调控网络中富集。通过 GO 生物过程富集分析和研究这些结构元件参与构建的网络，发现这些结构元件倾向出现在发育过程的网络中。随后的动力学模拟说明这些元件能在信号出现时完成状态的转换与维持，这些功能是多细胞发育特别是组织和器官的分化所需要的。

通过比较发育子网络和应激子网络，我们发现无论是在结构元件的组成、网络的全局拓扑结构还是参与构建的转录因子的性质上都存在明显的不同。与应激系统相比，发育系统具有更加复杂的调控模式，参与发育系统构建的转录因子具有更高的调控特异性。这些特点与发育和应激系统的功能特点相符合。发育系统通过复杂的调控（体现在使用更加复杂的新结构元件和更多转录因子之间的调控）精确调控发育状态的维持和转换，而应激系统则倾向于使用具有影响力的转录因子通过简单的调控模式来改变大量基因的表达状态以迅速响应各种胁迫。

在植物登陆期间产生很多新的转录因子家族，这些新类型的转录因子是如何参与转录调控系统构建的？使用 PlantTFDB 2.0 中 28 个具有基因组序列的物种的转录因子全谱，将转录因子家族按起源时间的早晚划分为新类型和古老类型两类。通过研究转录因子是如何参与生物过程和转录调控网络构建的，发现新类型的转录因子更倾向于参与构建发育过程中新类型结构元件和更加复杂的转录因子之间的调控。

通过比较新类型和古老类型转录因子的性质，我们发现新类型的转录因子具有更高的调控特异性。新类型转录因子的高调控特异性与其参与生物过程的倾向性之间有什么联系呢？为了研究调控特异性和调控什么之间的关系，我们将靶基因划分为转录因子和非转录因子两类。进而发现转录因子的调控特异性与靶基因中转录因子的比例呈明显的正相关性。转录因子的调控特异性越高越倾向于调控转录因子。随后在真细菌（大肠杆菌）、真菌（酿酒酵母）和后生动物（人）中发现了一致的模式，说明这个模式是转录调控网络构建的通用模式。鉴于转录调控网络的高重搭率和 DNA 结合结构域的高保守性，说明转录因子在网络中的位置已受

到自然选择的作用。经过长期的演化，转录因子已经连接到一个适合它性质的位置上。这个转录调控网络构建的通用模式也在一定程度上解释了具有高调控特异性的新类型转录因子为什么倾向于参与构建发育过程中新类型结构元件和复杂的转录因子之间的调控。

通过对水平转移的转录因子和新产生的 microRNA 的研究表明当一个新转录因子整合入一个新系统时，它对原有系统的影响越小则有更大的几率保留下来。如果是这样的话新类型转录因子具有高调控特异性或许是一个更加通用的模式。为此，我们还研究了大肠杆菌、酿酒酵母和人中新类型转录因子的情况。结果发现单细胞的大肠杆菌和酿酒酵母只有少数的新类型转录因子或者家族，而多细胞的人中则有很多新类型的转录因子。与古老类型相比，人中新类型的转录因子同样具有更高的调控特异性。当新类型的转录因子整合到原有系统时，如果调控特异性比较低，则有更大的可能带来有害的影响从而被选择所淘汰，这或许是新类型转录因子具有更高调控特异性的一个原因。很多具有高调控特异性的新类型转录因子产生，它们又倾向于参与构建新结构元件和复杂的转录因子之间的调控等发育类型的网络，或许暗示着它们在生命从单细胞到多细胞演化中的重要作用。

总之，通过系统识别和核对文献中报道的转录调控关系，我们构建了一个高质量的拟南芥转录调控网络 ATRM。这个网络展现了一个植物的转录调控概况。通过分析发育和应激这两个植物转录调控系统的主要组成部分，我们发现它们在结构元件的组成、网络全局拓扑结构和构建它们的转录因子的性质上都有着明显的不同。通过研究转录因子是如何参与以上两部分构建的，发现新类型的转录因子具有更高的倾向性而且倾向于构建发育网络中新结构元件和转录因子之间更加复杂的调控。这些结果为研究转录因子的演化命运和它们在多细胞生物演化中的作用提供了新的见解。

6.2展望

PlantTFDB 包含 83 个植物的转录因子谱，覆盖了绿色植物的各大分支。自发布以来，年访问量达千万次，已广泛应用于植物转录因子的功能和演化研究中。虽然目前的转录因子家族分类规则已是最新，但是随着研究的不断深入，家族数目和规则或许会有所变化。因此，还需要通过不断浏览文献来更新转录因子分类规则和优化转录因子预测的流程。

随着芯片和测序技术的广泛使用，诸如 GEO、ArrayExpress、AtGeneExpress 等公共数据库已存储了大量的表达数据。如果通过整合这些数据计算出转录因子的表达谱、共表达的基因及其差异表达的时期或胁迫处理，将为用户研究该转录因子的功能提供更加直接的线索。

目前研究表明很多保守元件在转录调控中作为顺式元件起作用，系统识别植物基因组中的保守元件将有助于研究转录因子的调控。但目前 PlantTFDB 只收录了个别物种如拟南芥的保守元件。将来需要通过构建植物的全基因组比对系统识别其中的保守元件。如能做到，一方面将展现植物基因组中保守元件的总体概况，为研究其演化模式提供重要的数据资源，另一方面也有助于研究相应转录因子的调控。

目前 PlantTFDB 只提供序列搜索和转录因子的预测服务。待表达信息整合进来后，用户输入某个基因将能搜出与其表达相关性很高的转录因子，然后可以进一步通过 GO 富集分析等研究其功能。随着收录的转录因子结合位点和矩阵越来越多，还可以为用户提供转录调控的预测服务。通过提交基因上游启动子序列，用户可以得到序列中包含已知转录因子的结合位点或元件。

通过系统识别和构建一个高质量的拟南芥转录调控网络，我们研究了植物的转录调控是如何构建的。结合系统识别出来的植物转录因子，我们探讨了新类型转录因子的性质和它们是如何参与转录调控系统构建的。该研究主要探讨了新类型和古老类型转录因子在调控特异性上的差异及调控特异性与网络构建的关系。除了调控特异性外，在转录因子的其它性质比如长度、表达量或 GC 含量上或许也存在某种模式。此外，由于目前拟南芥转录调控数据、转录因子结合矩阵等信息还不是太多，所以目前研究还比较粗浅，里面的很多发现和观点还有待进一步的验证。随着以后相关数据越来越多，一方面可以验证我们的结果和观点，另一方面还可以研究转录调控网络演化的基本模式和速率，不同类型的转录因子在这些方面的异同等问题。

参考文献

Abdulrehman D, Monteiro PT, Teixeira MC, Mira NP, Lourenco AB, dos Santos SC, Cabrito TR, Francisco AP, Madeira SC, Aires RS, *et al.* YEASTRACT: providing a programmatic access to curated transcriptional regulatory associations in Saccharomyces cerevisiae through a web services interface. *Nucleic Acids Res*. 2011. **39**: D136-40.

Alexa A and Rahnenfuhrer J. topGO: topGO: Enrichment analysis for Gene Ontology. *R package version 2.10.0*. 2010.

Alon U. Network motifs: theory and experimental approaches. *Nat Rev Genet*. 2007. **8**: 450-61.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 1997. **25**: 3389-402.

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

Arabidopsis Interactome Mapping Consortium. Evidence for network evolution in an Arabidopsis interactome map. *Science*. 2011. **333**: 601-7.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*. 2000. **25**: 25-9.

Bansal M, Belcastro V, Ambesi-Impiombato A and di Bernardo D. How to infer gene networks from expression profiles. *Mol Syst Biol*. 2007. **3**: 78.

Barabasi AL and Albert R. Emergence of scaling in random networks. *Science*. 1999. **286**: 509-12.

Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, *et al.* NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res*. 2013. **41**: D991-5.

Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R and Califano A. Reverse engineering of regulatory networks in human B cells. *Nat Genet*. 2005. **37**: 382-90.

Baxter L, Jironkin A, Hickman R, Moore J, Barrington C, Krusche P, Dyer NP, Buchanan-Wollaston V, Tiskin A and Beynon J. Conserved noncoding sequences highlight shared components of regulatory networks in dicotyledonous plants. *The Plant Cell Online*. 2012. **24**: 3949-3965.

Becskei A and Serrano L. Engineering stability in gene networks by autoregulation. *Nature*. 2000. **405**: 590-3.

Blanc G and Wolfe KH. Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. *Plant Cell*. 2004. **16**: 1679-91.

Borneman AR, Gianoulis TA, Zhang ZD, Yu H, Rozowsky J, Seringhaus MR, Wang LY, Gerstein M

and Snyder M. Divergence of transcription factor binding sites across related yeast species. *Science*. 2007. **317**: 815-9.

Bowman JL, Smyth DR and Meyerowitz EM. Genetic interactions among floral homeotic genes of Arabidopsis. *Development*. 1991. **112**: 1-20.

Bradley D, Ratcliffe O, Vincent C, Carpenter R and Coen E. Inflorescence commitment and architecture in Arabidopsis. *Science*. 1997. 275: 80-83.

Bradley RK, Li X-Y, Trapnell C, Davidson S, Pachter L, Chu HC, Tonkin LA, Biggin MD and Eisen MB. Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related Drosophila species. *PLoS biology*. 2010. **8**: e1000343.

Brady SM, Zhang L, Megraw M, Martinez NJ, Jiang E, Yi CS, Liu W, Zeng A, Taylor-Teeples M, Kim D, *et al.* A stele-enriched gene regulatory network in the Arabidopsis root. *Mol Syst Biol*. 2011. **7**: 459.

Britten RJ and Davidson EH. Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. *The Quarterly Review of Biology*. 1971. **46**: 111-138.

Byrne ME. Networks in leaf development. *Curr Opin Plant Biol*. 2005. **8**: 59-66.

Carroll K, Gomez C and Shapiro L. Tubby proteins: the plot thickens. *Nat Rev Mol Cell Biol*. 2004. **5**: 55-63.

Carroll SB. Endless forms: the evolution of gene regulation and morphological diversity. *Cell*. 2000. **101**: 577-80.

Charoensawan V, Wilson D and Teichmann SA. Lineage-specific expansion of DNA-binding transcription factor families. *Trends Genet*. 2010. **26**: 388-93.

Coen ES and Meyerowitz EM. The war of the whorls: genetic interactions controlling flower development. *Nature*. 1991. **353**: 31-7.

Conant GC. Rapid reorganization of the transcriptional regulatory network after genome duplication in yeast. *Proc Biol Sci*. 2010. **277**: 869-76.

Consortium U. Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Research*. 2013. **41**: D43-47.

Costanzo MC, Crawford ME, Hirschman JE, Kranz JE, Olsen P, Robertson LS, Skrzypek MS, Braun BR, Hopkins KL, Kondu P, *et al.* YPD, PombePD and WormPD: model organism volumes of the BioKnowledge library, an integrated resource for protein information. *Nucleic Acids Res*. 2001. **29**: 75-9.

Crick F. Central dogma of molecular biology. *Nature*. 1970. **227**: 561-3.

Crooks GE, Hon G, Chandonia JM and Brenner SE. WebLogo: a sequence logo generator. *Genome Res*. 2004. **14**: 1188-90.

Csardi G and Nepusz T. The igraph software package for complex network research. *InterJournal, Complex Systems*. 2006. **1695**.

Cutler SR, Rodriguez PL, Finkelstein RR and Abrams SR. Abscisic acid: emergence of a core signaling network. *Annu Rev Plant Biol*. 2010. **61**: 651-79.

de Bruijn S, Angenent GC and Kaufmann K. Plant 'evo-devo' goes genomic: from candidate genes to regulatory networks. *Trends Plant Sci*. 2012. **17**: 441-7.

Doebley J, Stec A and Hubbard L. The evolution of apical dominance in maize. *Nature*. 1997. **386**: 485-8.

Duvick J, Fu A, Muppirala U, Sabharwal M, Wilkerson MD, Lawrence CJ, Lushbough C and Brendel V. PlantGDB: a resource for comparative plant genomics. *Nucleic Acids Res*. 2008. **36**: D959-65.

Eddy S. HMMER User's Guide: Biological sequence analysis using profile hidden Markov models. 2010.

Enright AJ, Van Dongen S and Ouzounis CA. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res*. 2002. **30**: 1575-84.

Farkas IJ, Wu C, Chennubhotla C, Bahar I and Oltvai ZN. Topological basis of signal integration in the transcriptional-regulatory network of the yeast, Saccharomyces cerevisiae. *BMC Bioinformatics*. 2006. **7**: 478.

Farnham PJ. Insights from genomic profiling of transcription factors. *Nat Rev Genet*. 2009. **10**: 605-16.

Filichkin SA, Priest HD, Givan SA, Shen R, Bryant DW, Fox SE, Wong WK and Mockler TC. Genome-wide mapping of alternative splicing in Arabidopsis thaliana. *Genome Res*. 2010. **20**: 45-58.

Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, *et al.* The Pfam protein families database. *Nucleic Acids Res*. 2010. **38**: D211-22.

Fortunato S. Community detection in graphs. *Physics Reports*. 2010. **486**: 75-174.

Gama-Castro S, Salgado H, Peralta-Gil M, Santos-Zavaleta A, Muniz-Rascado L, Solano-Lira H, Jimenez-Jacinto V, Weiss V, Garcia-Sotelo JS, Lopez-Fuentes A, *et al.* RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic Acids Res*. 2011. **39**: D98-105.

Gao G, Zhong Y, Guo A, Zhu Q, Tang W, Zheng W, Gu X, Wei L and Luo J. DRTF: a database of rice transcription factors. *Bioinformatics*. 2006. **22**: 1286.

Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, Mu XJ, Khurana E, Rozowsky J, Alexander R, *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature*. 2012. **489**: 91-100.

Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, *et al.* A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). *Science*.

2002. **296**: 92-100.

Gompel N, Prud'homme B, Wittkopp PJ, Kassner VA and Carroll SB. Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in Drosophila. *Nature*. 2005. **433**: 481-487.

Guo AY, He K, Liu D, Bai S, Gu X, Wei L and Luo J. DATF: a database of Arabidopsis transcription factors. *Bioinformatics*. 2005. **21**: 2568-9.

Guo AY, Chen X, Gao G, Zhang H, Zhu QH, Liu XC, Zhong YF, Gu X, He K and Luo J. PlantTFDB: a comprehensive plant transcription factor database. *Nucleic Acids Research*. 2008. **36**: D966.

Guzzi PH, Cannataro M. CytoMCL: A Cytoscape plugin for fast clustering of protein interaction networks. *CBMS*. 2012:1-5

Haudry A, Platts AE, Vello E, Hoen DR, Leclercq M, Williamson RJ, Forczek E, Joly-Lopez Z, Steffen JG and Hazzouri KM. An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nature Genetics*. 2013, **45**: 891–898

Heim MA, Jakoby M, Werber M, Martin C, Weisshaar B and Bailey PC. The basic helix-loop-helix transcription factor family in plants: a genome-wide study of protein structure and functional diversity. *Mol Biol Evol*. 2003. **20**: 735-47.

Herrgard MJ, Covert MW and Palsson BO. Reconciling gene expression data with known genome-scale regulatory network structures. *Genome Res*. 2003. **13**: 2423-34.

Hertz GZ and Stormo GD. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*. 1999. **15**: 563-77.

Hoffmann R and Valencia A. Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics*. 2005. **21 Suppl 2**: ii252-8.

Hunter L, Lu Z, Firby J, Baumgartner WA, Johnson HL, Ogren PV and Cohen KB. OpenDMAP: An open source, ontology-driven concept analysis engine, with applications to capturing knowledge regarding protein transport, protein interactions and cell-type-specific gene expression. *BMC Bioinformatics*. 2008. **9**: 78.

Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A, Bernard T, Binns D, Bork P, Burge S, *et al.* InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res*. 2012. **40**: D306-12.

Hupalo D and Kern AD. Conservation and functional element discovery in 20 angiosperm plant genomes. *Mol Biol Evol*. 2013. **30**: 1729-44.

Iida K, Seki M, Sakurai T, Satou M, Akiyama K, Toyoda T, Konagaya A and Shinozaki K. RARTF: database and tools for complete sets of Arabidopsis transcription factors. *DNA Res*. 2005. **12**: 247-56.

Irish VF. The flowering of Arabidopsis flower development. *Plant J*. 2010. **61**: 1014-28.

Iseli C, Jongeneel CV and Bucher P. ESTScan: a program for detecting, evaluating, and

reconstructing potential coding regions in EST sequences. *Proc Int Conf Intell Syst Mol Biol*. 1999. 138-48.

Ishida T, Kurata T, Okada K and Wada T. A genetic regulatory network in the development of trichomes and root hairs. *Annu Rev Plant Biol*. 2008. **59**: 365-86.

Jia J, Zhao S, Kong X, Li Y, Zhao G, He W, Appels R, Pfeifer M, Tao Y and Zhang X. Aegilops tauschii draft genome sequence reveals a gene repertoire for wheat adaptation. *Nature*. 2013. **496**: 91-5

Jin H and Martin C. Multifunctionality and diversity within the plant MYB-gene family. *Plant Mol Biol*. 1999. **41**: 577-85.

Jin J, Zhang H, Kong L, Gao G and Luo J. PlantTFDB 3.0: a portal for the functional and evolutionary study of plant transcription factors. *Nucleic Acids Research*. 2013. gkt1016.

Jothi R, Balaji S, Wuster A, Grochow JA, Gsponer J, Przytycka TM, Aravind L and Babu MM. Genomic analysis reveals a tight link between transcription factor dynamics and regulatory network architecture. *Mol Syst Biol*. 2009. **5**: 294.

Kalir S and Alon U. Using a quantitative blueprint to reprogram the dynamics of the flagella gene network. *Cell*. 2004. **117**: 713-20.

Kent WJ. BLAT--the BLAST-like alignment tool. *Genome Res*. 2002. **12**: 656-64.

Kersey PJ, Staines DM, Lawson D, Kulesha E, Derwent P, Humphrey JC, Hughes DS, Keenan S, Kerhornou A, Koscielny G, *et al.* Ensembl Genomes: an integrative resource for genome-scale data from non-vertebrate species. *Nucleic Acids Res*. 2012. **40**: D91-7.

King MC and Wilson A. Evolution at Two Levels Humans and Chimpanze. *Pan*. 1975. **11**: I111-1.

Kuhn M, Campillos M, Letunic I, Jensen LJ and Bork P. A side effect resource to capture phenotypic effects of drugs. *Mol Syst Biol*. 2010. **6**: 343.

Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, *et al.* The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res*. 2012. **40**: D1202-10.

Lang D, Weiche B, Timmerhaus G, Richardt S, Riaño-Pachón DM, Corrêa LGG, Reski R, Mueller-Roeber B and Rensing SA. Genome-Wide Phylogenetic Comparative Analysis of Plant Transcriptional Regulation: A Timeline of Loss, Gain, Expansion, and Correlation with Complexity. *Genome Biology and Evolution*. 2010. **2**: 488.

Lau S, Slane D, Herud O, Kong J and Jurgens G. Early embryogenesis in flowering plants: setting up the basic body pattern. *Annu Rev Plant Biol*. 2012. **63**: 483-506.

Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, *et al.* Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*. 2002. **298**: 799-804.

Lee WY, Lee D, Chung WI and Kwon CS. Arabidopsis ING and Alfin1-like protein families localize to the nucleus and bind to H3K4me3/2 via plant homeodomain fingers. *Plant J*. 2009. **58**: 511-24.

Li H and Johnson AD. Evolution of transcription networks--lessons from yeasts. *Curr Biol*. 2010. **20**: R746-53.

Li L, Stoeckert CJ and Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome research*. 2003. **13**: 2178.

Lu J, Shen Y, Wu Q, Kumar S, He B, Shi S, Carthew RW, Wang SM and Wu CI. The birth and death of microRNA genes in Drosophila. *Nat Genet*. 2008. **40**: 351-5.

Luscombe NM, Austin SE, Berman HM and Thornton JM. An overview of the structures of protein-DNA complexes. *Genome Biol*. 2000. **1**: REVIEWS001.

Luscombe NM, Babu MM, Yu H, Snyder M, Teichmann SA and Gerstein M. Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*. 2004. **431**: 308-12.

Ma HW, Buer J and Zeng AP. Hierarchical structure and modules in the Escherichia coli transcriptional regulatory network revealed by a new top-down approach. *BMC Bioinformatics*. 2004. **5**: 199.

Maglott D, Ostell J, Pruitt KD and Tatusova T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research*. 2011. **39**: D52-57.

Mangan S and Alon U. Structure and function of the feed-forward loop network motif. *Proc Natl Acad Sci U S A*. 2003. **100**: 11980-5.

Marbach D, Roy S, Ay F, Meyer PE, Candeias R, Kahveci T, Bristow CA and Kellis M. Predictive regulatory models in Drosophila melanogaster by integrative inference of transcriptional networks. *Genome Res*. 2012. **22**: 1334-49.

Matlin AJ, Clark F and Smith CW. Understanding alternative splicing: towards a cellular code. *Nat Rev Mol Cell Biol*. 2005. **6**: 386-98.

Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D and Alon U. Network motifs: simple building blocks of complex networks. *Science*. 2002. **298**: 824-7.

Mochida K, Yoshida T, Sakurai T, Yamaguchi-Shinozaki K, Shinozaki K and Tran LS. LegumeTFDB: an integrative database of Glycine max, Lotus japonicus and Medicago truncatula transcription factors. *Bioinformatics*. 2010. **26**: 290-1.

Montiel G, Gantet P, Jay-Allemand C and Breton C. Transcription factor networks. Pathways to the knowledge of root development. *Plant Physiol*. 2004. **136**: 3478-85.

Muller HM, Kenny EE and Sternberg PW. Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol*. 2004. **2**: e309.

Neph S, Stergachis AB, Reynolds A, Sandstrom R, Borenstein E and Stamatoyannopoulos JA. Circuitry and dynamics of human transcription factor regulatory networks. *Cell*. 2012. **150**:

1274-86.

Nikitin A, Egorov S, Daraselia N and Mazo I. Pathway studio--the analysis and navigation of molecular networks. *Bioinformatics*. 2003. **19**: 2155-7.

Notredame C, Higgins DG and Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*. 2000. **302**: 205-17.

Novichkova S, Egorov S and Daraselia N. MedScan, a natural language processing engine for MEDLINE abstracts. *Bioinformatics*. 2003. **19**: 1699-706.

Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin Y-C, Scofield DG, Vezzi F, Delhomme N, Giacomello S and Alexeyenko A. The Norway spruce genome sequence and conifer genome evolution. *Nature*. 2013. **497:** 579-84

Obayashi T, Hayashi S, Saeki M, Ohta H and Kinoshita K. ATTED-II provides coexpressed gene networks for Arabidopsis. *Nucleic Acids Res*. 2009. **37**: D987-91.

Ouyang S, Thibaud-Nissen F, Childs KL, Zhu W and Buell CR. Plant genome annotation methods. *Methods Mol Biol*. 2009. **513**: 263-82.

Pérez-Rodríguez P, Riaño-Pachón DM, Corrêa LGG, Rensing SA, Kersten B and Mueller-Roeber B. PlnTFDB: updated content and new features of the plant transcription factor database. *Nucleic Acids Research*. 2010. **38**: D822-827.

Perez JC and Groisman EA. Evolution of transcriptional regulatory circuits in bacteria. *Cell*. 2009. **138**: 233-44.

Price MN, Dehal PS and Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One*. 2010. **5**: e9490.

Price MN, Dehal PS and Arkin AP. Horizontal gene transfer and the evolution of transcriptional regulation in Escherichia coli. *Genome Biol*. 2008. **9**: R4.

Pruitt KD, Tatusova T, Klimke W and Maglott DR. NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res*. 2009. **37**: D32-6.

Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G and Clements J. The Pfam protein families database. *Nucleic Acids Research*. 2012. **40**: D290-301.

Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R and Lopez R. InterProScan: protein domains identifier. *Nucleic Acids Res*. 2005. **33**: W116-20.

Rajewsky N, Socci ND, Zapotocky M and Siggia ED. The evolution of DNA regulatory regions for proteo-gamma bacteria by interspecies comparisons. *Genome Res*. 2002. **12**: 298-308.

Rebholz-Schuhmann D, Oellrich A and Hoehndorf R. Text-mining solutions for biomedical research: enabling integrative biology. *Nat Rev Genet*. 2012. **13**: 829-39.

Reinhold H, Soyk S, Šimková K, Hostettler C, Marafino J, Mainiero S, Vaughan CK, Monroe JD and Zeeman SC. β-Amylase–Like Proteins Function as Transcription Factors in Arabidopsis,

Controlling Shoot Growth and Development. *The Plant Cell Online*. 2011. **23**: 1391-1403.

Rensing SA, Lang D, Zimmer AD, Terry A, Salamov A, Shapiro H, Nishiyama T, Perroud PF, Lindquist EA, Kamisugi Y, *et al.* The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. *Science*. 2008. **319**: 64-9.

Richardt S, Lang D, Reski R, Frank W and Rensing SA. PlanTAPDB, a phylogeny-based resource of plant transcription-associated proteins. *Plant Physiol*. 2007. **143**: 1452-66.

Riechmann J. 2 Transcription factors of Arabidopsis and rice: a genomic perspective. *Regulation of transcription in plants*. 2006. 28.

Riechmann J, Heard J, Martin G, Reuber L, Keddie J, Adam L, Pineda O, Ratcliffe O, Samaha R and Creelman R. Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. *Science*. 2000. **290**: 2105-2110.

Riechmann JL. Transcriptional regulation: a genomic overview. *The Arabidopsis Book*. 2002. **1**: 1-46.

Rivest R. The MD5 message-digest algorithm. 1992.

Ronquist F and Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*. 2003. **19**: 1572-4.

Rosenfeld N, Elowitz MB and Alon U. Negative autoregulation speeds the response times of transcription networks. *J Mol Biol*. 2002. **323**: 785-93.

Rushton PJ, Bokowiec MT, Laudeman TW, Brannock JF, Chen X and Timko MP. TOBFAC: the database of tobacco transcription factors. *BMC Bioinformatics*. 2008. **9**: 53.

Rustici G, Kolesnikov N, Brandizi M, Burdett T, Dylag M, Emam I, Farne A, Hastings E, Ison J, Keays M, *et al.* ArrayExpress update--trends in database growth and links to data analysis tools. *Nucleic Acids Res*. 2013. **41**: D987-90.

Rzhetsky A, Seringhaus M and Gerstein M. Seeking a new biology through text mining. *Cell*. 2008. **134**: 9-13.

Salgado H, Peralta-Gil M, Gama-Castro S, Santos-Zavaleta A, Muniz-Rascado L, Garcia-Sotelo JS, Weiss V, Solano-Lira H, Martinez-Flores I, Medina-Rivera A, *et al.* RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Res*. 2013. **41**: D203-13.

Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M and Federhen S. Database resources of the national center for biotechnology information. *Nucleic Acids Research*. 2011. **39**: D38-51.

Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Federhen S, *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. 2010. **38**: D5-16.

Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S,

Martinez-Jimenez CP, Mackay S, *et al.* Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science*. 2010. **328**: 1036-40.

Schneider TD, Stormo GD, Gold L and Ehrenfeucht A. Information content of binding sites on nucleotide sequences. *J Mol Biol*. 1986. **188**: 415-31.

Shen-Orr SS, Milo R, Mangan S and Alon U. Network motifs in the transcriptional regulation network of Escherichia coli. *Nat Genet*. 2002. **31**: 64-8.

Shou C, Bhardwaj N, Lam HY, Yan K-K, Kim PM, Snyder M and Gerstein MB. Measuring the evolutionary rewiring of biological networks. *PLoS computational biology*. 2011. **7**: e1001050.

Shulaev V, Sargent DJ, Crowhurst RN, Mockler TC, Folkerts O, Delcher AL, Jaiswal P, Mockaitis K, Liston A and Mane SP. The genome of woodland strawberry (Fragaria vesca). *Nature Genetics*. 2011. **43**: 109-16.

Smaczniak C, Immink RG, Angenent GC and Kaufmann K. Developmental and evolutionary diversity of plant MADS-domain factors: insights from recent studies. *Development*. 2012. **139**: 3081-98.

Teichmann SA and Babu MM. Gene regulatory network growth by duplication. *Nat Genet*. 2004. **36**: 492-6.

Thorn CF, Klein TE and Altman RB. Pharmacogenomics and bioinformatics: PharmGKB. *Pharmacogenomics*. 2010. **11**: 501-5.

Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Powell K, Mortensen HM, Hirbo JB and Osman M. Convergent adaptation of human lactase persistence in Africa and Europe. *Nature Genetics*. 2006. **39**: 31-40.

Townsley BT and Sinha NR. A new development: evolving concepts in leaf ontogeny. *Annu Rev Plant Biol*. 2012. **63**: 535-62.

Tuch BB, Li H and Johnson AD. Evolution of eukaryotic transcription circuits. *Science*. 2008. **319**: 1797-9.

Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, *et al.* The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science*. 2006. **313**: 1596-604.

Van Bel M, Proost S, Wischnitzki E, Movahedi S, Scheerlinck C, Van de Peer Y and Vandepoele K. Dissecting plant genomes with the PLAZA comparative genomics platform. *Plant Physiol*. 2012. **158**: 590-600.

Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R and Birney E. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res*. 2009. **19**: 327-35.

Wang BB, O'Toole M, Brendel V and Young ND. Cross-species EST alignments reveal novel and conserved alternative splicing events in legumes. *BMC Plant Biol*. 2008. **8**: 17.

Wang E and Purisima E. Network motifs are enriched with transcription factors whose transcripts have short half-lives. *Trends in Genetics*. 2005. **21**: 492-494.

Wang H, Nussbaum-Wagler T, Li B, Zhao Q, Vigouroux Y, Faller M, Bomblies K, Lukens L and Doebley JF. The origin of the naked grains of maize. *Nature*. 2005. **436**: 714-9.

Wang RL, Stec A, Hey J, Lukens L and Doebley J. The limits of selection during maize domestication. *Nature*. 1999. **398**: 236-9.

Wang X, Wei X, Thijssen B, Das J, Lipkin SM and Yu H. Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nat Biotechnol*. 2012. **30**: 159-64.

Weigel D and Meyerowitz EM. The ABCs of floral homeotic genes. *Cell*. 1994. **78**: 203-9.

Wellmer F and Riechmann JL. Gene networks controlling the initiation of flower development. *Trends Genet*. 2010. **26**: 519-27.

Werner T, Koshikawa S, Williams TM and Carroll SB. Generation of a novel wing colour pattern by the Wingless morphogen. *Nature*. 2010. **464**: 1143-1148.

Wilson D, Charoensawan V, Kummerfeld SK and Teichmann SA. DBD--taxonomically broad transcription factor predictions: new content and functionality. *Nucleic Acids Res*. 2008. **36**: D88-92.

Wu M and Chan C. Learning transcriptional regulation on a genome scale: a theoretical analysis based on gene expression data. *Brief Bioinform*. 2012. **13**: 150-61.

Wu TD and Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*. 2005. **21**: 1859-1875.

Yilmaz A, Mejia-Guerra MK, Kurz K, Liang X, Welch L and Grotewold E. AGRIS: the Arabidopsis Gene Regulatory Information Server, an update. *Nucleic Acids Res*. 2011. **39**: D1118-22.

Yilmaz A, Nishiyama MY, Jr., Fuentes BG, Souza GM, Janies D, Gray J and Grotewold E. GRASSIUS: a platform for comparative regulatory genomics across the grasses. *Plant Physiol*. 2009. **149**: 171-80.

Yu H and Gerstein M. Genomic analysis of the hierarchical structure of regulatory networks. *Proc Natl Acad Sci U S A*. 2006. **103**: 14724-31.

Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, *et al.* A draft sequence of the rice genome (Oryza sativa L. ssp. indica). *Science*. 2002. **296**: 79-92.

Zhang H, Jin J, Tang L, Zhao Y, Gu X, Gao G and Luo J. PlantTFDB 2.0: update and improvement of the comprehensive plant transcription factor database. *Nucleic Acids Res*. 2011. **39**: D1114-7.

Zhang JZ, Creelman RA and Zhu JK. From laboratory to field. Using information from Arabidopsis to engineer salt, cold, and drought tolerance in crops. *Plant Physiol*. 2004. **135**: 615-21.

Zheng W, Zhao H, Mancera E, Steinmetz LM and Snyder M. Genetic analysis of variation in transcription factor binding in yeast. *Nature*. 2010. **464**: 1187-1191.

Zhu QH, Guo AY, Gao G, Zhong YF, Xu M, Huang M and Luo J. DPTF: a database of poplar

transcription factors. *Bioinformatics*. 2007. **23**: 1307.

Zimmermann P, Hirsch-Hoffmann M, Hennig L and Gruissem W. GENEVESTIGATOR. Arabidopsis microarray database and analysis toolbox. *Plant Physiol*. 2004. **136**: 2621-32.

附录

附录 1　67 个基因组测序已完成的物种中用于转录因子识别的基因注释版本

类群	拉丁名	常用名	注释版本
绿藻	*Bathycoccus prasinos*	-	ORCAE
	Chlamydomonas reinhardtii	-	JGI, v4.0
	Chlorella sp. NC64A	-	JGI, v1.0
	Coccomyxa sp. C-169	-	JGI, v2
	Micromonas pusilla CCMP1545	-	JGI, v3.0
	Micromonas sp. RCC299	-	JGI, v3.0
	Ostreococcus lucimarinus CCE9901	-	JGI, v2.0
	Ostreococcus sp. RCC809	-	JGI, v2.0 [$]
	Ostreococcus tauri	-	JGI, v2.0
	Volvox carteri	-	JGI, v2.0
苔藓	*Physcomitrella patens*	-	JGI, v1.6
蕨类	*Selaginella moellendorffii*	-	JGI, v1.0
裸子植物	*Picea abies*	挪威云杉	ConGenIE, v1.0
被子植物基部植物	*Amborella trichopoda*	互叶梅	AGD, EVM27
单子叶植物	*Aegilops tauschii*	节节麦	BGI
	Brachypodium distachyon	二穗短柄草	JGI, v1.2
	Hordeum vulgare	大麦	IBSC, v2.18
	Musa acuminata	粉蕉	Cirad
	Oryza barthii	非洲野生稻	AGI
	Oryza brachyantha	-	OGE, v1.4b
	Oryza glaberrima	非洲稻	AGI, v1.1
	Oryza punctata	-	AGI
	Oryza sativa subsp. indica	籼稻	RIS, glean
	Oryza sativa subsp. japonica	粳稻	MSU, v7.0
	Phoenix dactylifera	海枣	PDK, v3
	Phyllostachys heterocycla	竹子	ICBR, v1.0
	Setaria italica	粟	JGI, v2.1
	Sorghum bicolor	高粱	JGI, v2.1
	Triticum urartu	乌拉尔图小麦	BGI
	Zea mays	玉米	MaizeSequence, 5b[#]
双子叶植物	*Aquilegia coerulea*	耧斗菜	JGI, v1.1
	Arabidopsis lyrata	琴叶拟南芥	JGI, v1.0
	Arabidopsis thaliana	拟南芥	TAIR10
	Azadirachta indica	印楝	NGD
	Brassica rapa	芥菜	BRAD, v1.2

Cajanus cajan	木豆	IIPG
Cannabis sativa	大麻	CCBR
Capsella rubella	-	JGI, v1.0
Carica papaya	木瓜	ASGPB
Cicer arietinum	鹰嘴豆	BGI
Citrullus lanatus	西瓜	ICuGI, v1
Citrus clementina	克莱门柚	ICGC, v1.0
Citrus sinensis	甜橙	JGI, v1.1
Cucumis melo	甜瓜	MELONOMICS, v3.5
Cucumis sativus	黄瓜	JGI, v1.0
Eucalyptus grandis	巨桉	JGI, v1.1
Fragaria vesca	野生草莓	GDR, v1.1
Glycine max	大豆	JGI, v1.1
Gossypium raimondii	棉花	JGI, v2.1
Jatropha curcas	-	Kazusa, v4.5
Linum usitatissimum	亚麻	BGI, v1.0
Lotus japonicus	-	Kazusa, v2.5
Malus domestica	苹果树	GDR, v1.0
Manihot esculenta	木薯	JGI, v4.1
Medicago truncatula	蒺藜苜蓿	Mt, v3.0
Mimulus guttatus	猴面花	JGI, v1.1
Nelumbo nucifera	莲花	CAS, v2.0
Populus trichocarpa	毛果杨	JGI, v3.0
Prunus persica	桃树	JGI, v1.0
Pyrus bretschneideri	梨树	CPETR, v1.0
Ricinus communis	蓖麻	JCVI, v0.1
Solanum lycopersicum	番茄	ITAG, v2.3
Solanum tuberosum	土豆	PGSC, v3.4
Thellungiella halophila	小盐芥	JGI
Thellungiella parvula	盐芥	Thellungiella, v2.0
Theobroma cacao	可可	CGD, v1.1
Utricularia gibba	狸藻	UGSP
Vitis vinifera	葡萄	Genoscope, v1.0

[#]过滤后的数据集

附录 2　67 个基因组测序已完成的物种中识别的 TF 统计

类群	拉丁名	基因数	TF 数	(%)	家族数
绿藻	*Bathycoccus prasinos*	7 919	139	1.76	26
	Chlamydomonas reinhardtii	19 526	230	1.18	29
	Chlorella sp. NC64A	9 791	163	1.66	28
	Coccomyxa sp. C-169	9 629	138	1.43	27
	Micromonas pusilla CCMP1545	10 658	150	1.41	32
	Micromonas sp. RCC299	9 891	153	1.55	31
	Ostreococcus lucimarinus CCE9901	7 645	111	1.45	30
	Ostreococcus sp. RCC809	7 492	102	1.36	29
	Ostreococcus tauri	7 664	99	1.29	26
	Volvox carteri	15 285	125	0.82	27
苔藓	*Physcomitrella patens*	32 273	1 079	3.34	53
蕨类	*Selaginella moellendorffii*	22 271	665	2.99	54
裸子植物	*Picea abies*	71 158	1 851	2.60	55
被子植物基部植物	*Amborella trichopoda*	26 846	900	3.35	58
单子叶植物	*Aegilops tauschii*	33 849	1 439	4.25	55
	Brachypodium distachyon	26 552	1 557	5.86	56
	Hordeum vulgare	24 211	1 198	4.95	56
	Musa acuminate	36 519	2 896	7.93	57
	Oryza barthii	31 675	1 507	4.76	56
	Oryza brachyantha	32 037	1 444	4.51	56
	Oryza glaberrima	33 164	1 579	4.76	56
	Oryza punctate	32 139	1 718	5.35	56
	Oryza sativa subsp. indica	40 745	1 891	4.64	56
	Oryza sativa subsp. japonica	55 803	1 859	3.33	56
	Phoenix dactylifera	28 882	1 426	4.94	56
	Phyllostachys heterocycla	31 987	1 768	5.53	54
	Setaria italic	40 599	1 994	4.91	56
	Sorghum bicolor	33 032	1 826	5.53	56
	Triticum urartu	24 169	888	3.67	50
	Zea mays	38 914	2 231	5.73	55
双子叶植物	*Aquilegia coerulea*	24 823	1 158	4.67	58
	Arabidopsis lyrata	32 670	1 759	5.38	58
	Arabidopsis thaliana	27 416	1 716	6.26	58
	Azadirachta indica	40 482	1 900	4.69	58
	Brassica rapa	41 019	3 026	7.38	57
	Cajanus cajan	40 071	1 886	4.71	56
	Cannabis sativa	22 670	1 061	4.68	56
	Capsella rubella	28 447	1 900	6.68	58
	Carica papaya	27 765	1 379	4.97	58
	Cicer arietinum	27 809	1 897	6.82	56
	Citrullus lanatus	23 440	1 355	5.78	58
	Citrus clementine	33 929	1 905	5.61	58

Citrus sinensis	46 147	2 256	4.89	58
Cucumis melo	27 427	1 322	4.82	58
Cucumis sativus	21 603	1 412	6.54	57
Eucalyptus grandis	36 376	1 729	4.75	56
Fragaria vesca	32 831	1 485	4.52	58
Glycine max	54 175	3 714	6.86	57
Gossypium raimondii	37 505	2 634	7.02	58
Jatropha curcas	52 782	1 467	2.78	57
Linum usitatissimum	43 484	2 481	5.71	57
Lotus japonicas	26 119	1 311	5.02	56
Malus domestica	63 516	3 119	4.91	58
Manihot esculenta	34 151	2 247	6.58	58
Medicago truncatula	50 952	1 577	3.10	56
Mimulus guttatus	28 282	1 733	6.13	57
Nelumbo nucifera	26 473	1 476	5.58	57
Populus trichocarpa	41 335	2 455	5.94	58
Prunus persica	28 701	1 529	5.33	58
Pyrus bretschneideri	42 812	2 353	5.50	57
Ricinus communis	31 221	1 299	4.16	57
Solanum lycopersicum	34 727	1 845	5.31	58
Solanum tuberosum	51 472	2 406	4.67	56
Thellungiella halophile	29 284	1 892	6.46	58
Thellungiella parvula	27 132	1 672	6.16	58
Theobroma cacao	29 452	1 449	4.92	58
Utricularia gibba	27 465	1 651	6.01	55
Vitis vinifera	26 346	1 276	4.84	58

附录 3　ATRM 中 62 个具有 5 个或以上成员的生物模块名称

模块编号	名称
1	gibberellin mediated signaling pathway
2	response to cold
3	floral meristem determinacy
4	response to far red light
5	circadian rhythm
6	response to abscisic acid stimulus
7	pollen development
8	response to jasmonic acid stimulus
9	secondary cell wall biogenesis
10	lignin metabolic process
11	proanthocyanidin biosynthetic process
12	meristem development
13	meristem development/cytokinin mediated signaling pathway
14	floral organ development
15	glucose mediated signaling pathway
16	lateral root development/response to auxin stimulus
17	flavonoid biosynthetic process
18	response to oxidative stress
19	root epidermal cell differentiation
20	glycosinolate biosynthetic process
21	response to ethylene stimulus
22	regulation of short-day photoperiodism, flowering
23	response to gibberellin stimulus/regulation of seed dormancy process
24	response to auxin stimulus
25	lipid metabolic process
26	response to xenobiotic stimulus/salicylic acid mediated signaling pathway
27	anther development
28	response to high light intensity
29	sulfur compound biosynthetic process
30	cellular response to phosphate starvation
31	pigment metabolic process
32	ER body organization
33	NA
34	transition metal ion transport
35	asymmetric cell division
36	response to heat
37	NA
38	maintenance of inflorescence meristem identity
39	seed maturation
40	regulation of abscisic acid mediated signaling pathway
41	fatty acid metabolic process

42	positive regulation of meiotic cell cycle
43	organ morphogenesis
44	response to auxin stimulus
45	response to UV
46	regulation of seed maturation
47	formation of organ boundary
48	response to water deprivation
49	auxin polar transport
50	systemic acquired resistance
51	specification of axis polarity
52	Mo-molybdopterin cofactor biosynthetic process
53	vernalization response
54	defense response, incompatible interaction
55	seed oilbody biogenesis
56	jasmonic acid biosynthetic process
57	NA
58	response to light stimulus
59	NA
60	regulation of cell cycle
61	trichome morphogenesis
62	iron ion transport

注：根据富集的前 5 个 GO 项来命名每个生物模块。其中 “NA”指由于富集的 GO 项功能不统一，因此未能给出一个特定的名称。

附录 4　拟南芥中转录因子结合矩阵的信息量

TRANSFAC 中矩阵 ID	TRANSFAC 中 TF ID	TAIR ID	信息量
M00089	T01474	AT3G01470	13.39
M00151	T01007	AT4G18960	9.06
M00218	T01568	AT2G32460	14.28
M00226	T01590	AT2G47460	6.12
M00343 + M00344	T02637	AT1G13260	14.25
M00352	T01059	AT1G21340	5.64
M00353	T15264	AT1G51700	5.33
M00354	T02691	AT3G55370	5.31
M00358	T02790	AT3G62420	12.84
M00361	T02855	AT1G09770	7.86
M00371	T09449	AT5G28770	5.98
M00375	T00830	AT2G40950	6.11
M00376	T09478	AT5G65210	5.5
M00392	T03025	AT2G03710	8.72
M00417	T04001	AT1G30490	29.69
M00435	T04492	AT1G09530	10.85
M00438	T04507, T04514, T04516, T04528, T04529,T04530, T04531	AT1G19850, AT1G59750, AT5G62000, AT1G30330, AT5G20730, AT5G37020, AT5G60450	6.75
M00439	T01592, T02946	AT4G38620, AT4G34990	4.43
M00441	T01079, T01080, T13945	AT4G01120, AT2G46270, AT2G35530	8.41
M00442	T03820, T03823, T03824, T03825	AT1G49720, AT1G45249, AT4G34000, AT3G19290	10.86
M00501	T02639	AT4G37750	11.03
M00502	T04804	AT3G20770	5.86
M00503	T04066	AT5G65310	22.68
M00635	T04454	AT1G13450	4.74
M00660	T02786	AT5G24800	5.57
M00681	T09147, T09148	AT3G56400, AT4G24240	8.52
M00697	T00938	AT3G12250	8.59
M00702	T03975	AT2G38470	7.13
M00735	T03722	AT2G04880	8.36
M00798	T05553	AT3G11440	4.66
M00819	T02063	AT4G08150	5.13
M00820	T08621	AT2G46680	8.19
M00948	T10960	AT2G45680	7.07
M00952	T05738	AT3G15030	7.59
M00958	T05743	AT2G40220	14.15
M00968	T02590	AT3G50060	8.48
M00969	T02534	AT3G23250	11.92

M00970	T02597	AT3G49690	12.03
M01006	T06043	AT1G08010	7.37
M01021	T03994	AT4G02670	7.66
M01050	T05654	AT4G31920	5.87
M01052	T06533	AT2G01060	8.1
M01055	T06532	AT3G04070	13.4
M01057	T02654	AT4G17500	6.27
M01059	T03022	AT3G58780	8.69
M01061	T03024	AT5G15800	9.09
M01114	T05582, T06393	AT2G36010, AT5G22220	7.67
M01126	T09002	AT2G01930	4.8
M01128	T08411	AT3G61850	5.62
M01135	T09610	AT5G06100	5.84
M01156	T09158	AT1G75080	6.12
M01179	T08415	AT4G16150	10.53
M01180	T09195	AT1G20980	7.48
M01188	T14491	AT4G35580	5.86
M01191	T14550	AT5G52170	16.75
M01192	T14551	AT5G17320	10.59
M01193	T14552	AT4G21750	11.95
M01194	T14553	AT4G04890	7.2
M01581	T03018	AT5G13790	8.92
M01582	T03032	AT1G24260	9.84
M01583	T01588, T06338	AT3G27920, AT5G41315	10.98
M01584	T02812	AT5G11260	7.2
M01585	T06460	AT2G20180	6.94
M01586	T02795	AT5G06950	12.64
M01700	T02636	AT4G25490	3.11
M01701	T17708	AT1G01720	2.87
M01740	T07646	AT4G31550	6.38
M01799	T14421	AT2G16910	5.33
M01806	T15670	AT4G36730	7.21
M01809	T14801	AT3G02310	8.9
M01815	T02796	AT1G22070	7.47
M01819	T02860	AT2G46830	12.04
M01829	T08410, T14568	AT4G24060, AT5G60200	4.93
M01839	T06407, T07257, T07424	AT5G14960, AT3G01330, AT3G48160	6.34
M01848	T21746	AT1G69690	8.12
M01849	T13992	AT3G27010	7.96
M01850	T21745	AT2G37000	8.88

附录 5　转录因子结合矩阵的信息量和预测的靶基因数

结合矩阵 ID	信息量	预测的靶基因数
M00089	13.39	1031
M00151	9.06	17212
M00218	14.28	1737
M00226	6.12	4608
M00343	7.96	27858
M00352	5.64	14151
M00353	5.33	26754
M00354	5.31	27735
M00358	12.84	43
M00361	7.86	1728
M00371	5.98	4092
M00375	6.11	2870
M00376	5.5	2327
M00392	8.72	9132
M00417	29.69	27
M00435	10.85	200
M00439	4.43	26353
M00441	8.41	1186
M00442	10.86	540
M00501	11.03	124
M00502	5.86	30760
M00503	22.68	2911
M00635	4.74	33388
M00660	5.57	9452
M00681	8.52	24920
M00697	8.59	969
M00702	7.13	12011
M00735	8.36	1942
M00798	4.66	22709
M00819	5.13	19340
M00820	8.19	3256
M00948	7.07	1081
M00952	7.59	920
M00958	14.15	300
M00968	8.48	1993
M00969	11.92	6
M00970	12.03	33
M01006	7.37	3740
M01021	7.66	14305
M01050	5.87	27794
M01052	8.1	655

M01055	13.4	8
M01057	6.27	6353
M01059	8.69	5493
M01061	9.09	5525
M01114	7.67	1258
M01126	4.8	22904
M01128	5.62	17541
M01135	5.84	33331
M01156	6.12	1706
M01179	10.53	84
M01180	7.48	1716
M01188	5.86	26078
M01191	16.75	73
M01192	10.59	409
M01193	11.95	57
M01194	7.2	9324
M01581	8.92	9132
M01582	9.84	3471
M01583	10.98	654
M01584	7.2	9703
M01585	6.94	5837
M01586	12.64	61
M01700	3.11	13042
M01701	2.87	33597
M01740	6.38	6326
M01799	5.33	4443
M01806	7.21	2482
M01815	7.47	3310
M01819	12.04	388
M01829	4.93	32827
M01839	6.34	1064

附录 6　大肠杆菌中转录因子的调控特异性与靶基因中转录因子所占比例

TF ID	信息量	靶基因中 TF 的比例 (%)
AGAR	12.21	9.09
ARAC	7.38	9.09
ARCA	6.16	2.98
ARGP	6.35	7.14
ARGR	9.35	5.41
CPXR	6.09	6.9
CRA	9.63	7.79
CRP	7.59	8.37
CSGD	4.57	4.35
CYSB	12.68	8.33
CYTR	7.11	16.67
DGSA	15.65	10
DNAA	8.22	8.33
EVGA	13.58	20
FHLA	7.59	6.67
FIS	4.61	4.04
FLHDC	7.34	1.25
FNR	7.3	4.41
FUR	10.16	6.87
GADE	11.34	17.14
GADW	8.31	28.57
GADX	6.72	17.86
GALR	8.95	20
GALS	9.27	20
GNTR	9.92	8.33
IHF	5.56	2.67
ISCR	8.85	3.23
LEUO	12.83	15
LEXA	11.78	3.57
LRP	4.16	1.96
MARA	7.75	16.22
METJ	4.45	6.67
MODE	12.36	2.17
NAC	6.53	13.33
NAGC	11.78	5.88
NARL	4.18	0.83
NSRR	7.33	7.23
NTRC	9.34	6.82
OMPR	7.99	5.88
OXYR	6.87	6.25

PDHR	9.69	2.38
PHOB	7.54	5.26
PHOP	7.59	9.09
PURR	12.13	3.23
RCSB	8.2	12.82
ROB	7.62	8.33
RSTA	12.3	10
RUTR	10.79	25
SOXS	8.01	11.11
TORR	5.95	8.33
TRPR	11.76	8.33

附录 7　酿酒酵母中转录因子的调控特异性与靶基因中转录因子所占比例

TF ID	靶基因数	靶基因中 TF 的比例 (%)
ABF1	669	3.57
ACE2	170	7.89
ADR1	443	3.17
AFT1	1114	3.11
AFT2	193	4.44
ARO80	97	33.33
ARR1	743	2.91
AZF1	127	4
CAD1	478	2.5
CRZ1	299	3.1
CST6	193	4.19
DAL80	78	11.54
ECM22	270	3.79
FHL1	884	1.81
FKH1	241	6.45
FKH2	313	6.45
GAL4	158	12
GAT1	150	17.65
GAT3	193	2.04
GCN4	1260	4.29
GCR1	281	2.36
GCR2	596	6.21
GLN3	667	2.51
GZF3	149	4.35
HAA1	89	5.81
HAC1	208	2.54
HAP1	189	1.45
HAP2	197	3.57
HAP3	185	3.7
HAP4	425	3.57
HAP5	195	4
HCM1	249	33.33
HMS1	221	1.24
HSF1	571	2.96
IFH1	359	2.1
INO2	165	3.23
INO4	637	2.08
LEU3	495	6.23
MAL13	6	33.33
MBP1	498	4.94

MCM1	402	5.88
MET31	126	2.56
MET32	101	7.69
MET4	1260	3.26
MGA1	291	4.5
MIG1	239	6.57
MIG2	44	4.35
MIG3	15	16.67
MOT3	135	2.35
MSN2	1187	2.06
MSN4	739	2.52
NDT80	35	6.25
NRG1	399	2.03
OAF1	263	1.47
PDC2	19	8.33
PDR1	653	2.66
PDR3	547	0.67
PHO2	173	2.27
PHO4	379	2.61
PIP2	150	2.68
PUT3	157	6.25
RAP1	1502	3.1
RFX1	492	2.63
RGM1	119	2.27
RIM101	213	5.97
RLM1	205	1.1
RME1	237	5.1
ROX1	406	4.44
RPN4	1032	2.57
RSF2	52	17.65
RTG1	129	13.16
RTG3	221	11.11
SFL1	47	8.33
SFP1	2183	2.87
SIP4	98	15.38
SKN7	640	6.8
SKO1	627	8.22
SOK2	1034	4.81
STB5	339	2.7
STE12	2142	3.14
STP1	232	8.6
STP2	342	4.68
SUM1	159	1.79
SWI4	614	5.21
SWI5	231	7.14

TEC1	570	3.85
TOS8	293	1.69
UME6	238	4.76
UPC2	208	4.65
XBP1	501	4.85
YAP1	1824	2.78
YHP1	307	6.43
YOX1	462	8.21
YRM1	35	5.88
YRR1	104	7.32
ZAP1	185	0.84

附录 8　人中转录因子的调控特异性与靶基因中转录因子的比例

TF ID	信息量	靶基因中 TF 的比例 (%)
BRCA1	4.37	4.3
EGR1	11.47	8.2
ELK4	8.64	11.1
ESRRA	7.15	5
ETS1	6.29	1.8
GATA1	6.12	6.8
GATA2	6.81	7
GATA3	6.23	5.3
HNF4A	10.03	5.9
JUN	8.47	6.5
MAX	9.22	5.8
MEF2A	11.16	9.3
MYC	10.3	4.9
NR3C1	9.89	4.7
RXRA	42.43	10
SP1	4.86	5.7
SP2	8.34	7.1
SPI1	14.27	5.3
SREBF1	8.86	7.5
SRF	14.97	8.5
STAT1	12.98	6.9
STAT3	15.2	8.6
TAL1	10.45	5.5
TBP	6.25	2.1
TCF4	10.45	8.6
TFAP2A	5.22	6.2
TFAP2C	5.66	6.4
USF1	7.12	6.3
ZBTB33	8.37	2.6
ZEB1	6.58	5.6

附录 9　常用缩略词汇表

简写	英文全称	中文名称
ATRM	*Arabidopsis* transcriptional regulatory map	拟南芥转录调控网络
BP	Biological process	生物过程
CDS	Coding sequence	编码序列
DBD	DNA-binding domain	DNA 结合结构域
DNA	Deoxyribonucleic acid	脱氧核糖核酸
EST	Expressed sequence tag	表达序列标签
FFL	Feed-forward loop	前馈环
GO	Gene Ontology	基因本体
HMM	Hidden Markov model	隐马尔科夫模型
IC	Information content	信息量
MRCA	Most recent common ancestor	最近共同祖先
NLS	Nuclear localization signal	核定位信号
OG	Orthologous group	直系同源群
PCC	Pearson correlation coefficient	Pearson 相关系数
PUT	PlantGDB-assembled unique transcripts	PlantGDB 拼接的唯一转录本
QTL	Quantitative trait loci	数量性状位点
RNA	Ribonucleic acid	核糖核酸
SIM	Simple Input Module	简单的输入模块
TF	Transcription factor	转录因子

附录 10 在学期间的研究成果

*并列一作，# 通讯作者

已发表论文：

1. **Jin JP**, Zhang H, Kong L, Gao G# and Luo JC#. PlantTFDB 3.0: A portal for the functional and evolutionary study of plant transcription factors. *Nucleic Acids Res*, 2013, doi: 10.1093/nar/gkt1016.
2. Zhang H*, **Jin JP***, Tang L, Zhao Y, Gu XC, Gao G# and Luo JC#. PlantTFDB 2.0: update and improvement of the comprehensive plant transcription factor database. *Nucleic Acids Res*, 2013, **39**(Database issue): D1114-1117.
3. Liu XC*, Li ZH*, Jiang ZQ, Zhao Y, Peng JY, **Jin JP**, Guo HW# and Luo JC#. LSD: a leaf senescence database. *Nucliec Acids Res*, 2013, 39(Database issue): D1103-1107.

待发表论文：

1. **Jin JP**, Tang X, Lv L, Li Z, Zhao Y, Luo JC, He K# and Gao G#. An *Arabidopsis* transcriptional regulatory map provides insights into the wiring preference of transcription factors. (In submitting)
2. Hu Bo*, **Jin JP***, Guo AY, Zhang H, Luo JC, Gao G#. GFDS: An upgraded gene feature display system. (In preparation)
3. Zhao Y, Tang L, Li Z, **Jin JP**, Wei T, Luo JC, Gao G#. Comprehensive profiling of unitary gene loss in Poaceae. (In preparation)
4. Ji LX*, Wang J*, **Jin JP**, Liao WH, Ye MX, Chen Z, An XM#. Dynamic transcriptional regulatory network under ABA stimulation in *Populus hopeiensis*. (In submitting)
5. Wang SX*, Zhao XY*, Li ZY, Shen HY, Lai WJ, Zhang XF, Ma SP, **Jin JP**, Fang QJ, Yin YX, Wang QS#, Ji JG#. A global survey of protein phosphorylation reveals its extensive regulatory network in rat fetal neural stem cells. *Stem Cells*. (Submitted)

致 谢

随着毕业论文接近尾声在燕园的研究生学习生涯也即将结束。五年多时光匆匆而过，从当初对科学一无所知到如今有些浅薄的认识。在此，通过此文感谢这五年多在学习、科研和生活等方面所有帮助过我的人。

首先，我要感谢我的导师罗静初教授和高歌研究员。在这五年间，罗老师在学习、科研和生活等方面给我提供了很多细心的指导和帮助，罗老师对科研的态度和对细节的把握是我学习的榜样。高歌老师学术渊博，对很多方面都有深刻的见解。感谢高歌老师在课题选择、科研和论文写作上给我的细心指导和帮助。正是由于您们的细心指导，我才慢慢对科研有了自己的认识。

感谢龙漫远教授在拟南芥转录调控网络演化中的指导。您对科研的理解和态度是我学习的榜样。

感谢顾孝诚老师在 PlantTFDB 2.0 论文写作和平时给予的帮助。您是一位谦虚的智者，您对工作的热爱和处世态度是我学习的榜样。虽然您已经不在了，但是您的笑容和声音永远留在我们心中。

感谢魏丽萍教授和陶乐天研究员，您在如何做科研、做什么样的科研等方面的看法让我受益匪浅。

感谢白书农、顾红雅、郭红卫、李程、陆剑等老师在论文修改上的建议。

感谢曲红老师、唐汶老师在平时的帮助。

感谢何坤师兄在拟南芥转录调控网络上的指导和帮助。这个课题是在孟山都开始的，也正是走出学校的这段时间使我开始独立的思考科研。同时感谢孟山都的吴昕师兄、吕乐博士、孙建东博士和 Zhang Ray 在平时的讨论和帮助。

感谢我的师兄张禾，他在技术和科研上给我很大的帮助，也感谢他在合作课题 PlantTFDB 2.0 中的出色工作。

感谢我的同学唐星、赵汗青、谢忱在平时的帮助，你我共同奋斗，一起度过这五年时光。

感谢李哲师兄、唐亮师兄、赵义在课题上的讨论和帮助。感谢孔雷师兄、赵树起师兄、刘小桥师兄在技术上的支持和帮助。感谢王珺师姐、陈文博师姐、李小沫师姐和刘欢在平时生活和学习上的帮助。感谢赵敏师兄、钟应福师兄、刘小川师兄、边洋师兄、徐礼鸣师姐、黄岳、王聪等在平时的帮助。感谢颜林林在服务器管理上所做的工作。感谢呼波在 GSDS 升级工作中的优异表现。感谢各位师弟

师妹，你们的到来使 CBI 充满了新的活力。

同时感谢所有已毕业的师兄和师姐，你们优秀表现为我们的工作打下良好的基础。

最后，感谢我的父母和家人，您们的理解和支持是我学习和科研的动力。

北京大学学位论文原创性声明和使用授权说明

原创性声明

本人郑重声明： 所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

论文作者签名： 日期： 年 月 日

学位论文使用授权说明

（必须装订在提交学校图书馆的印刷本）

本人完全了解北京大学关于收集、保存、使用学位论文的规定，即：

- 按照学校要求提交学位论文的印刷本和电子版本；
- 学校有权保存学位论文的印刷本和电子版，并提供目录检索与阅览服务，在校园网上提供服务；
- 学校可以采用影印、缩印、数字化或其它复制手段保存论文；
- 因某种特殊原因需要延迟发布学位论文电子版，授权学校□一年/□两年/□三年以后，在校园网上全文发布。

（保密论文在解密后遵守此规定）

论文作者签名： 导师签名：

日期： 年 月 日