科学通报 2009 年 第 54 卷 第 23 期: 3710 ~ 3715
论 文 www.scichina.com csb.scichina.com
《中国科学》杂志社
SCIENCE IN CHINA PRESS

拟南芥代谢通路下基因调控网络的构建

焦清局[①], 杨仲南[②], 黄继风[①*]

① 上海师范大学信息与机电工程学院, 上海 200234;
② 上海师范大学生命与环境科学学院, 上海 200234
* 联系人, E-mail: jfhuang@shnu.edu.cn

2009-06-18 收稿, 2009-10-09 接受
上海市教育委员会教育科学研究资助项目(批准号: 07ZZ60)

摘要 基因调控网络在研究基因之间的调控关系及揭示复杂的生命现象方面有着重要的意义. 拟南芥整个生长过程是由基因网络所调控. 本文利用拟南芥代谢通路下基因共表达这一属性, 结合启动子序列分析的生物信息学方法来预测拟南芥代谢通路下基因的调控关系. 基于这种方法, 一共预测到 2268 对具有调控关系的基因对, 其中 91 对为高可信度的调控关系. 在我们预测到的调控关系中, 有 4 对调控关系已被实验验证, 实验表明本文预测的结果有一定的可靠性. 我们预测的拟南芥代谢通路下基因的调控网络, 为深入研究代谢通路在植物生长过程中所起的作用提供了方便, 有助于进一步研究拟南芥未知基因的功能.

关键词
拟南芥
代谢通路
基因调控网络
生物信息学
模体

随着生物分子数据的不断积累, 人们越来越多地利用计算机处理和分析生物数据, 代谢通路研究则成为其中的一个研究热点. 人们建立了各种有关动物和植物的通路数据库, 像 Reactome 人类生物学过程的通路数据库[1]、KEGG 通路数据库[2]、MetaCyc[3]、TRANSPATH 关于信号转换事件的数据库[4], 以及 TAIR 上有关拟南芥通路的数据库[5]. 通路作为一条维持生命的分子反应链, 不同的生命过程有不同的通路, 如细胞对外界环境变化相应的通路、光合作用、细胞代谢和信号通路. 一条通路功能的中断可能引起动物严重的疾病或者阻碍植物的生长. 根据代谢通路下基因的特征可以预测基因的调控网络. 因此, 代谢通路在研究动植物生长和基因的调控方面起着重要的作用.

一个基因的表达受其他基因的调控或影响, 而这个基因又调控或影响其他基因的表达, 这种相互调控或影响的关系构成了复杂的基因表达调控网络. 基因调控网络中基因的相互作用关系能够帮助研究者更加深入地认识真实的调控过程[6], 对调控过程认识的加深将会对药物研制和生物医学产生深远的影响. 基因调控网络是功能基因组学研究的一个热点, 它从基因之间互相调控或影响的角度揭示复杂的生命现象.

目前, 人们不仅构建了多种动植物的基因调控网络, 而且将其研究广泛地应用到各种模式物种当中[7~10]. 如 Divya Mathur 利用调控网络分析了老鼠的胚胎干细胞, 并绘制了转录因子 OCT4 和 NANOG 从数据集到染色体坐标的序列结合位点, 为进一步研究老鼠的胚胎干细胞提供了珍贵的资源[11]. Tsai 等人[12]利用基因表达数据和染色质免疫沉淀反应数据(ChIP-chip)结合统计模型对酵母细胞分裂周期调控网络进行了预测, 分析了细胞分裂过程中调控因子之间相互作用的过程. 然而, 由于缺少较为可靠的拟南芥基因表达数据以及一些模体(motif)预测工具准确度不够高, 鲜有大规模的拟南芥转录调控网络被构建出来.

本文基于同一代谢通路下的基因共表达这一原理[13,14], 并利用一种整合基因表达数据与启动子区

引用格式: 焦清局, 杨仲南, 黄继风. 拟南芥代谢通路下基因调控网络的构建. 科学通报, 2009, 54: 3710~3715
Jiao Q J, Yang Z N, Huang J F. Construction of a gene regulatory network for *Arabidopsis* based on metabolic pathway data. Chinese Sci Bull, doi: 10.1007/s11434-009-0728-8

域转录因子结合位点预测的生物信息学方法来构建拟南芥代谢通路下的基因调控网络. 通过启动子区域的转录因子结合位点(TBFS)的预测, 得出了一些高可信的基因调控关系, 利用这些调控关系构建了一个拟南芥代谢通路下基因的调控网络.

1 材料和方法

(ⅰ) 拟南芥代谢通路数据. (1) 拟南芥数据资源(TAIR)提供了拟南芥各方面的资源[5]. 从互联网上(ftp://ftp.arabidopsis.org/home/tair/pathways/)下载了所有拟南芥代谢通路数据. (2) ATTED-Ⅱ (http://atted.jp/)数据库提供了拟南芥基因的共表达关系, 用来预测未知基因的功能[15].

(ⅱ) 共表达基因组的筛选. 基于同一代谢通路下的基因共表达这一原理. 我们对下载得到的代谢通路数据进行整理. 首先, 把每一条代谢通路下功能不同却是同一个基因的合并为一个基因, 即去掉每一条代谢通路下重复的基因. 这样共得到 309 条通路(相当于 309 组共表达基因组). 然后, 用 MEME[16]软件预测这 309 组共表达基因组中的模体. MEME 软件对输入的一组共表达基因有数量上的要求, 如果输入的共表达基因数量较少, 则会影响预测结果的正确性. 为了得到可靠的模体预测结果, 设定输入 MEME 软件的每条通路下基因的数量大于等于 5. 最后, 通过整理共得到 174 条代谢通路数据, 即 174 组共表达基因组.

(ⅲ) 启动子区域的模体预测. 一般认为, 转录因子结合位点主要在转录起始位点(transcription start sites, TSSs)附近出现, 有些研究发现: 真核生物转录因子识别的模体一般分布在基因上游–1~500 的区域[17]. 但是有些转录因子结合在基因上游很远的区域, 被称为远程作用. 我们要根据研究问题的不同, 适当地选取转录起始位点碱基的长度. 如果序列太短会丢失部分结合位点. 序列太长会引入大量的背景噪声, 使真正的转录因子结合位点在噪声中无法区分.

为了可靠地预测共表达基因中的模体, 我们把得到的 174 条代谢通路下的基因及其对应的序列输入到 MEME 软件中, 并在 MEME 软件中设置了不同的参数, 做了不同的实验:

(1) 启动子序列为基因上游长 500 bp 的序列. 在每组共表达基因组中设置了 3 个不同的模体个数: ① 模体的个数为: 基因的个数除以 2, 然后对结果取整; ② 模体的个数为 3 个; ③ 模体的个数为 5 个. 使用 MEME 软件的 ZOOPS 模式; 其余参数采用 MEME 软件默认值.

(2) 启动子序列为基因上游长 1000 bp 的序列. 在每组共表达基因组中设置了 3 个不同的模体个数: ① 模体的个数为: 基因的个数除以 2, 然后对结果取整; ② 模体的个数为 3 个; ③ 模体的个数为 5 个. 使用 MEME 软件的 ZOOPS 模式; 其余参数采用 MEME 软件默认值.

(ⅳ) 调控关系的预测. 通过 MEME 软件预测出启动子上模体信息后, 利用 STAMP[18]工具将模体信息转化成转录因子信息. STAMP 将模体序列与数据库中已知转录因子结合位点序列比对, 从而确定一个基因及其启动子上结合的转录因子. 具体步骤如下:

(1) 使用 Perl 程序批量向 STAMP 服务器提交预测出来的模体信息, STAMP 服务器根据其转录因子识别序列数据库将模体信息转化成转录因子信息. STAMP 服务器收集了 TRANSFAC[19], AthaMap[20]和 AGRIS[21]这 3 个数据库存储的转录因子识别序列信息, 其中 TRANSFAC 数据库包含了各个物种的识别数据, AthaMap 包含了几种植物中的转录因子识别序列数据, 而 AGRIS 数据库则只包含拟南芥已知的转录因子识别序列数据.

(2) 从 STAMP 工具获得的转录因子信息作为启动子上可能结合的转录因子. 如果预测的转录因子为其他物种, 则根据同源性分析转换为拟南芥的转录因子. 一个基因与其启动子上可能结合的转录因子构成具有调控关系的基因对.

(ⅴ) 预测结果的筛选. 通过预测共得到 2268 对转录因子与其靶基因之间的转录调控关系. 为了提高结果的可靠性, 利用 ATTED-Ⅱ数据库对初步结果进行筛选, 从网站(http://atted.jp/download.shtml)(c4.1)下载了拟南芥全基因组的共表达数据. 这些数据包含了 20906 个文件, 每一文件对应一个拟南芥基因(目标基因), 并以该基因的 AGI 号命名. 每一个文件都有 22572 行, 每一行代表一个基因, 在每一个基因后面都有一个范围从–1.0~1.0 的相关系数来表示该基因与目标基因之间的共表达关系, 系数越大, 表示这两个基因的共表达性越高. 利用拟南芥全基因组的共表达数据, 编写 perl 程序对得到的 2268 对转

录调控关系进行筛选，如果转录调控关系对中的两个基因之间的共表达系数大于等于 0.6，则认为这一调控关系对为高可信的[22].

2 结果

2.1 实验结果的筛选

为了建立较为精确的基因调控网络，我们对同一数据集做了不同的实验，并对不同的实验结果用实验验证的 348 对调控关系(从 AGRIS 数据库得到)进行筛选，不同的实验结果如表 1 所示.

从表 1 可以看出：第一，当我们设置的启动子序列长 1000 bp 时，无论模体的个数怎样变化，预测到的调控关系对中包含已被实验验证的关系对始终为 0，说明 1000 bp 启动子序列中包含大量的背景噪声，使 MEME 工具无法正确地识别真正的转录因子结合位点；第二，当我们设置启动子序列为 500 bp，模体的个数为 3 时，在预测到的转录调控关系对中包含 2 个已被实验验证的调控关系对 (E2Fa 调控 At2g21790; E2Fa 调控 At3g46940)；模体的个数为基因个数除以 2 或等于 5 时，我们均预测到 4 对已被实验验证的基因对(E2Fa 调控 At2g21790; E2Fa 调控 At3g46940; HY5 调控 At5g38410; HY5 调控 At1g67090). 然而在一些代谢通路下基因个数较少(少于或等于 5 个)而又设置模体的个数为 5 时，预测结果中就会包含一些假阳性的转录因子结合位点. 所以在最后的实验中，我们设置启动子序列的长度为 500 bp, motif 的个数为基因个数除以 2，使用 MEME 软件的 ZOOPS 模式，其余参数为工具默认. 从实验结果我们得到与真核生物转录因子识别的模体一般分布在基因上游–1~500 的区域[17]相吻合的观点.

2.2 代谢通路下基因调控网络的构建

我们从 TAIR 网站上下载了所有的拟南芥 309 条生物代谢通路，通过整理共有 1620 个基因(不同代谢通路下的基因有重叠). 为提高 MEME 软件预测模体的可靠性，去掉代谢通路下基因个数小于 5 的通路，这样，经过整理共得到 174 条生物代谢通路(网络版附录 1)，即 174 组共表达基因. 然后，使用模体预测软件 MEME 对每一组基因启动子序列进行预测，找到这些基因启动子区域上的模体. 这些模体可能是转录因子的识别结合位点. 为了确定识别这些模体的转录因子，使用 STAMP 工具将模体信息转换成转录因子信息，从而确定一个基因启动子上可能结合的转录因子(具体流程见图 1). 经过分析，共预测到 2268 对模式植物拟南芥代谢通路表达转录因子与其靶基因之间的调控关系.

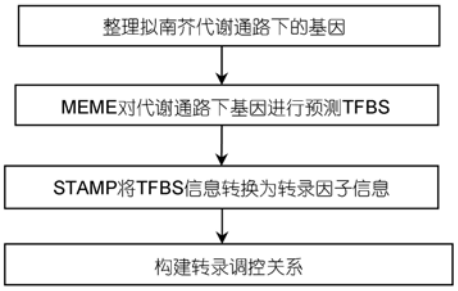

图 1 构建调控网络流程图

2.3 高可信度的基因调控关系

为了进一步提高结果的准确性，我们使用 ATTED-Ⅱ数据库中的基因共表达数据对以上的结果进行筛选，根据材料与方法(ⅴ)中的筛选条件，得到了 91 对高可信度的转录调控关系(网络版附录 2). 对预测出来的调控关系一共包含了 96 个基因，通过检索转录因子数据库 AGRIS 后，发现有 21 个编码转录因子的基因(网络版附录 3). 同时，利用 pajek 网络分析软件(http://vlado.fmf.uni-lj.si/pub/networks/pajek/)将这 91 对转录调控关系绘制成了可视化的图形(图 2).

表 1 实验结果

启动子序列长度	motif 的个数	预测到总的转录调控关系的个数	包含已被实验验证的调控关系对个数
500 bp	基因个数除以 2	2268	4
500 bp	3	2153	2
500 bp	5	2259	4
1000 bp	基因个数除以 2	2067	0
1000 bp	3	2026	0
1000 bp	5	2045	0

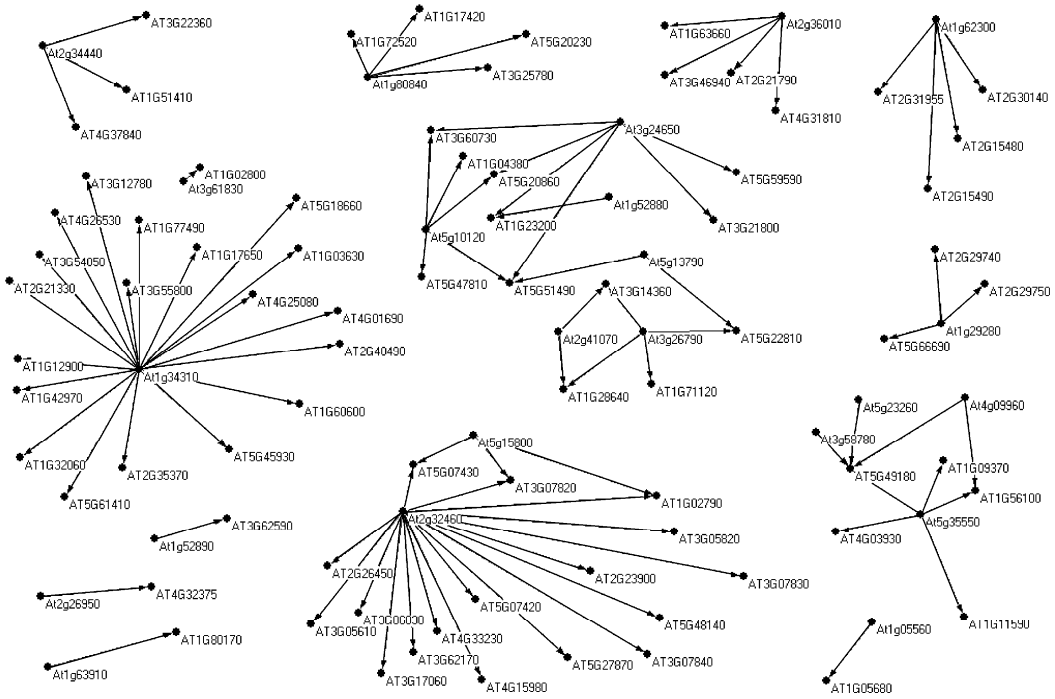

图 2　代谢通路下基因的调控网络

基因命名中 t 和 g 小写的表示转录因子，大写的表示靶基因，箭头表示转录因子和靶基因之间的调控关系

2.4　被实验验证的转录调控关系

我们从 AGRIS 数据库收集到了 348 对实验验证的转录调控关系，并将这些转录调控关系与我们预测的结果进行比较. 结果表明，在我们预测到的调控关系中，有 4 对调控关系被之前的实验所验证. 这 4 对调控关系是: E2Fa 调控 At2G21790，E2Fa 调控 At3G46940，HY5 调控 At5G38410，HY5 调控 At1G67090. 有研究发现, E2Fa 和 HY5 在拟南芥生长过程中起着非常重要的作用，其中异二聚体 E2F-DP 转录因子通过基因的调控转录来控制细胞周期，而这些基因是 DNA 复制和细胞周期不可缺少的[23]. 对于哺乳类动物, E2Fs 的 8 个转录因子都已经被克隆并注释. 在拟南芥基因组中，人们也识别出了 3 个 E2F (E2Fa, E2Fb, E2Fc)和 2 个 DP (DPa, DPb)[23]，它们是很重要的转录因子. 而 HY5 是拟南芥转录因子中注释最详细的转录因子之一，它是一个基本的亮氨酸拉链类型的转录因子，也是一个光合作用的正向调节元件[24]. 这些结果表明，我们所预测的基因调控关系有一定的准确性. 因此，本研究所预测的拟南芥代谢通路下基因调控网络将为研究代谢通路的作用提供有意义的数据.

3　讨论

3.1　代谢通路下的基因

本文只下载了 TAIR 网站上有关拟南芥的生物代谢通路数据，这一数据并不能包含拟南芥所有的代谢通路信息. 因此，无法精确地预测拟南芥代谢通路下基因的调控关系. 但是，由于这些数据只是我们输入 MEME 工具的原始数据，从统计学意义上看，一组共表达基因(一条代谢通路)中个别基因的缺失，对模体预测的可靠性影响不大. 把所得到的 174 条代谢通路基因(174 组共表达基因)作为整理后的最终结果，以此为基础预测到 2268 对基因调控关系，其中 91 对为高可靠的调控基因对. 在预测到的基因调控关系中有 4 对已被实验验证，说明预测结果有一定的可靠性，为拟南芥生长发育的分子机理研究提供了有用的参考信息. 在所预测出的 2268 对基因调控关系中，有部分为功能未知基因. 根据调控关系中的已知基因功能，可以预测未知基因的功能. 因此我们所构建的基因调控网络也为拟南芥发育中未知功能基因的研究提供了参考信息.

3.2 预测结果的生物学意义

MapMan 根据代谢通路和生物学功能对拟南芥蛋白质进行分类. 在第一层次, MapMan 将拟南芥基因产物分为 35 类, 其中第 27 类包含绝大部分的转录因子. 根据转录因子与被调控基因的关系发现: 在高可信度的 91 个基因对中, 大部分转录因子(除 At1g05560, At3g24650 外)都处在 27.3 层, 而在 MapMan 的分层说明里, 27.3 这一层的蛋白质基本是转录因子.

转录因子既可以调控靶基因, 也可以调控转录因子. 在 91 对高可信的基因对中, 转录因子 At1g34310 调控的基因为 19 个, 转录因子 At2g32460 调控的基因为 17 个, 说明它们在拟南芥生长过程中起着非常重要的作用. 在转录因子 At2g36010 调控的 4 个靶基因中, At2g21790 和 At3g46940 已被实验验证, 于是, 我们可以预测另外两个基因(At1g63660 和 At4g31810)与实验验证的靶基因有相似的基因功能.

在我们预测到的已被实验验证的 4 对调控关系对中, 其中 HY5 调控的 At5g38410 和 At1g67090 这两个基因均在光合作用这一代谢通路下, 与光合作用中的蓝光、红光、远红光的反应都有密切的联系. 而光为光反应提供了能量来源. E2Fa 调控的 At2g21790 和 At3g46940 基因在花瓣分化和生长等多个发育阶段表达, 其中, At2g21790 基因参与 DNA 复制和修复, 而 At3g46940 基因参与 dUTP 的代谢过程. 我们以附录的形式给出这 4 个靶基因的功能以及在拟南芥生长过程中发挥的作用(网络版附录 4).

3.3 文中研究方法的优点和缺点

相对实验方法, 文章所采用的生物信息学方法有着快速、高通量的优势, 可以建立一个大规模的基因调控网络. 与生物信息学中的反向工程的方法相比, 避免了收集大量的基因表达数据的困难. 但是我们所使用的方法亦存在着许多不足之处: 首先, 由于拟南芥代谢通路数据不足, 目前启动子上模体的预测算法的准确率还不够高, 假阳性结果较多; 其次, 由于缺乏足够的拟南芥转录调控实验数据, 难以对生物信息学的预测结果进行全面的评估. 尽管如此, 本研究方法仍然能预测已知调控关系中的 4 对, 表示该方法有一定的准确性, 该方法可以为拟南芥其他生物学过程分子机理研究和构建基因调控网络提供一种新的研究思路.

参考文献

1 Matthews L, Gopinath G, Gillespie M, et al. Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Res, 2009, 37(Database issue): D619—D622

2 Kanehisa M, Araki M, Goto S, et al. KEGG for linking genomes to life and the environment. Nucleic Acids Res, 2008, 36(Database issue): D480—D484[doi]

3 Caspi R, Foerster H, Fulcher C A, et al. The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res, 2008, 36(Database issue): D623—D631

4 Krull M, Pistor S, Voss N, et al. TRANSPATH: An information resource for storing and visualizing signaling pathways and their pathological aberrations. Nucleic Acids Res, 2006, 34(Database issue): D546—D551[doi]

5 Swarbreck D, Wilks C, Lamesch P, et al. The Arabidopsis Information Resource (TAIR): Gene structure and function annotation. Nucleic Acids Res, 2007: 1—6

6 Jong H D. Modeling and simulation of genetic regulatory systems: a literature review. J Comput Biol, 2002, 9: 67—103[doi]

7 Espinosa-Soto C, Padilla-Longoria P, Alvarez-Buylla E R. A gene regulatory network model for cell-fate determination during *Arabidopsis thaliana* flower development that is robust and recovers experimental gene expression profiles. Plant Cell, 2004, 16: 2923—2939[doi]

8 Keurentjes J J B, Fu J, Terpstra I R, et al. Regulatory network construction in *Arabidopsis* by using genome-wide gene expression quantitative trait loci. Proc Natl Acad Sci USA, 2007, 104: 1708—1713[doi]

9 Lackner D H, Beilharz T H, Marguerat S, et al. A network of multiple regulatory layers shapes gene expression in fission yeast. Mol Cell, 2007, 26: 145—155[doi]

10 McAdams H H, Shapiro L. A bacterial cell-cycle regulatory network operating in time and space. Science, 2003, 301: 1874—1877[doi]

11 Mathur D, Danford T W, Boyer L A, et al. Analysis of the mouse embryonic stem cell regulatory networks obtained by ChIP-chip and ChIP-PET. Genome Biology, 2008, 9: R126[doi]

12 Tsai H K, Lu H H, Li W H. Statistical methods for identifying yeast cell cycle transcription factors. Proc Natl Acad Sci USA, 2005, 102: 13532—13537[doi]

13 Wei H, Persson S, Mehta T, et al. Transcriptional coordination of the metabolic network in *Arabidopsis*. Plant Physiol, 2006, 142: 762—774[doi]
14 Srinivasasainagendra V, Page G P, Mehta T, et al. CressExpress: A tool for large-scale mining of expression data from *Arabidopsis*. Plant Physiol, 2008, 147: 1004—1016[doi]
15 Obayashi T, Hayashi S, Saeki M, et al. ATTED-II provides coexpressed gene networks for *Arabidopsis*. Nucleic Acids Res, 2009, 37(Database issue): D987—D991[doi]
16 Bailey T L, Williams N, Misleh C, et al. MEME: Discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res, 2006, 34(Web Server issue): W369—W373[doi]
17 Caselle M, Di Cunto F, Provero P. Correlating overrepresented upstream motifs to gene expression: A computational approach to regulatory element discovery in eukaryotes. BMC Bioinformatics 2002, 3: 7[doi]
18 Mahony S, Benos P V. STAMP: A web tool for exploring DNA-binding motif imilarities. Nucleic Acids Res, 2007, 35(Web Sever issue): W253—W258[doi]
19 Matys V, Fricke E, Geffers R, et al. TRANSFAC: Transcriptional regulation, from patterns to profiles. Nucleic Acids Res, 2003, 31: 374—378[doi]
20 Steffens N O, Galuschka C, Schindler M, et al. AthaMap: an online resource for in sillico transcription factor binding sites in the Arabidopsis thaliana genome. Nucleic Acids Res, 2004, 32(Database issue): D368—D372[doi]
21 Palaniswamy S K, James S, Sun H. AGRIS and AtRegNet. A platform to link cis-regulatory elements and transcription factors into regulatory networks. Plant Physiol, 2006, 140: 818—829[doi]
22 黄继凤, 杨晶晶, 王冠, 等. 拟南芥花药表达基因调控关系的预测. 科学通报, 2008, 53: 2054—2059
23 Vandepoele K, Vlieghe K, Florquin K, et al. Genome-wide identification of potential plant E2F target genes. Plant Physiol, 2005, 139: 316—328[doi]
24 Gao Y, Li J, Strickland E, et al. An *Arabidopsis* promoter microarray and its initial usage in the identification of HY5 binding targets *in vitro*. Plant Mol Biol, 2004, 54: 683—699[doi]