学校编码：10384　　分类号＿＿密级＿＿

学号：200326059　　UDC＿＿

厦门大学

硕士学位论文

水稻和拟南芥中碱性/螺旋－环－螺旋转录因子家族基因组水平分析

Genome-Wide Analysis of Basic/Helix-Loop-Helix Transcription Factor Family in Rice and Arabidopsis

作者姓名：李晓星

指导教师：陈　亮　　教授

张大兵　　教授

专业名称：细 胞 生 物 学

论文提交日期：2006 年 4 月 25 日

论文答辩时间：2006 年 6 月　5 日

学位授予日期：

答辩委员会主席：陶　涛　教授

评　　阅　　人：＿＿＿＿＿＿

＿＿＿＿＿＿

2006 年　6 月

厦门大学学位论文原创性声明

兹呈交的学位论文，是本人在导师指导下独立完成的研究成果。本人在论文写作中参考的其他个人或集体的研究成果，均在文中以明确方式标明。本人依法享有和承担由此论文产生的权利和责任。

声明人（签名）：

2006 年　　月　　日

厦门大学学位论文著作权使用声明

本人完全了解厦门大学有关保留、使用学位论文的规定。厦门大学有权保留并向国家主管部门或其指定机构送交论文的纸质版和电子版，有权将学位论文用于非赢利目的的少量复制并允许论文进入学校图书馆被查阅，有权将学位论文的内容编入有关数据库进行检索，有权将学位论文的标题和摘要汇编出版。保密的学位论文在解密后适用本规定。

本学位论文属于

1、保密（　），在　　年解密后适用本授权书。

2、不保密（√）

（请在以上相应括号内打“√”）

作者签名：　　日期：　　年　月　日

导师签名：　　日期：　　年　月　日

目 录

厦门大学学位论文原创性声明……I
厦门大学学位论文著作权使用声明……II
目 录……III
摘 要……1
ABSTRACT……2
第一章 绪 论……3
1.1 研究背景……3
1.1.1 生物信息学概述……3
1.1.2 生物信息学数据库介绍……4
1.1.3 生物信息学工具介绍……7
1.1.4 植物中的生物信息学研究……10
1.2 bHLH基因家族研究现状……10
1.2.1 bHLH结构域的组成……10
1.2.2 动物中的bHLH蛋白研究进展……11
1.2.3 植物中的bHLH蛋白研究现状……11
第二章 bHLH家族生物信息学分析……14
2.1 研究目的……14
2.2 研究方法……14
2.2.1 在数据库中搜索水稻bHLH序列……14
2.2.2 多序列联配 (Multiple sequence alignments)……15
2.2.3 构建系统发生树……15
2.2.4 水稻bHLH基因在水稻染色体组上的定位……16
2.2.5 预测miRNA在水稻和拟南芥bHLH基因中的靶基因……16
2.3 实验结果……16
2.3.1 确定水稻中bHLH家族含有至少 165 个成员……16
2.3.2 水稻中bHLH蛋白的序列联配、残基保守性分析及DNA结合活性预测……22
2.3.3 水稻bHLH基因家族系统发生分析……27
2.3.4 水稻和拟南芥bHLH结构域中内含子/外显子结构分析……30
2.3.5 水稻bHLH基因家族成员在水稻各染色体上的分布情况……32
2.3.6 水稻与拟南芥中bHLH基因家族的比较……34
2.3.7 在水稻与拟南芥bHLH基因家族成员中进行miRNA靶基因的预测……38
2.4 分析讨论……39
2.4.1 水稻与拟南芥bHLH基因家族成员的基本信息……39
2.4.2 水稻与拟南芥bHLH基因家族进化分析……40
2.4.3 水稻与拟南芥bHLH基因家族序列保守性与功能分析……40
第三章 bHLH家族成员表达活性分析……41
3.1 研究目的……41
3.2 材料方法……41
3.2.1 植物材料来源……41
3.2.2 试剂与仪器……42
3.2.3 植物组织提取RNA……42
3.2.4 检测表达活性所用引物的设计……42

3.2.5 反转录和PCR……47
3.2.6 水稻和拟南芥bHLH基因其他表达数据的获取……47
3.3 实验结果……48
3.3.1 水稻和拟南芥bHLH基因的表达模式……48
3.3.2 已研究的水稻bHLH基因的功能……54
3.4 分析讨论……54
3.4.1 水稻bHLH基因表达谱……54
3.4.2 水稻bHLH基因与拟南芥bHLH基因表达谱的比较……54
3.4.3 序列的信息学分析和表达谱研究的意义……55
第四章 结论与展望……56
参 考 文 献……58
致 谢……67

摘　要

碱性/螺旋－环－螺旋 (bHLH) 转录因子家族是一个在动植物中广泛存在的基因家族。这个家族成员众多并且在动植物的生长发育过程中也起到十分重要的调控作用。但目前，这个家族成员的功能在植物中的研究还比较少。近年来，植物模式生物拟南芥和水稻的全基因组测序工作，为从基因组水平研究和比较开花植物的 bHLH 基因家族提供了坚实的基础。通过数据库搜索和分析，在水稻基因组中找到了 165 个可能的 bHLH 基因家族成员。系统发生分析表明，这些 bHLH 基因可以依据支持度划分成多个亚家族。同时，bHLH 转录因子结合 DNA 活性预测、bHLH 结构域外的保守基序 (motif) 分析、内含子/外显子保守结构分析等多项信息学分析的结果初步揭示了这个家族的成员在进化和功能上的相关性。bHLH 家族成员在水稻和拟南芥基因组上的分布和复制 (duplication) 分析也为染色体区域复制的理论提供了有力证据。此外，对拟南芥和水稻中 bHLH 成员所做的系统发生分析表明，单、双子叶植物 bHLH 基因含有 66 个以上的共同的祖先。这一分析也支持了基因家族进化的“发生和消亡”理论。生物信息学分析也归纳了 bHLH 蛋白与其他蛋白相互作用来调控各种基因转录的可能规律。对拟南芥和水稻 bHLH 家族基因序列和表达活性分析表明，序列相似的 bHLH 基因在水稻和拟南芥中可能有相似的功能。

关键词：碱性/螺旋－环－螺旋 (bHLH) 基因家族；系统进化分析；表达谱分析

ABSTRACT

The basic/helix-loop-helix (bHLH) transcription factors and their homologs form a large family in plant and animal genomes and play important roles in the specification of tissue type in animals. However, few plant bHLH proteins have been studied functionally. Recent whole genome sequences of model plants *Arabidopsis thaliana* and *Orysa sativa* allow genome-wide analysis and comparison of the bHLH family in flowering plants. We have identified 165 bHLH genes in the rice genome, and their phylogenetic analysis indicates that they form well-supported clades, which are defined as subfamilies. In addition, sequence analysis of potential DNA binding activity, the finding of motifs outside the bHLH domain, the conservation of intron/exon structural patterns further support the evolutionary and potential functional relationships among these proteins. The genome distribution of rice bHLH genes strongly supported the hypothesis that genome duplication(s) contributed to the expansion of the bHLH gene family. Furthermore, phylogenetic studies of both rice and Arabidopsis bHLH genes estimate that 66 bHLH genes were present in the most recent common ancestor of monocots and eudicots, represented by rice and Arabidopsis, respectively. Also, this analysis also provides strong support for the "birth-and-death" theory of gene family evolution. Bioinformatic analysis suggests that rice bHLH proteins can potentially participate in a variety of combinatorial interactions, endowing them with the capacity to regulate a multitude of transcriptional programs. In addition, similar expression patterns support functional conservation between some rice bHLH genes and their close Arabidopsis homologs.

Key Words： Basic/Helix-Loop-Helix (bHLH) Gene Family; Phylogenetic Analysis; Expression Pattern

第一章　　绪　论

1.1　研究背景

1.1.1 生物信息学概述

随着生物学的迅速发展，特别是各物种基因组测序计划的顺利推进，研究者们已获得大量的生物分子数据，并且其积累速度还在不断地增加。这些数据具有非常丰富的内涵，揭示这些数据的内涵，得到对人类有用的信息，这将是生物学家和数学家们所面临的一个严峻的挑战。生物信息学是近年来为迎接这种挑战而发展起来的一个新型交叉学科[1, 2]。目前，生物信息学的研究对象主要是DNA序列和蛋白质序列，其主要任务是分析研究序列数据中所含的各种信息，特别是DNA序列中的遗传及调控信息，研究蛋白质序列与结构及功能的关系。

生物信息学的涵盖范围很广，在揭示生物分子数据内涵方面，主要包括以下一些手段：

1) 生物分子数据的收集与管理：随着DNA序列测定技术的不断进步，已测定的DNA序列信息呈指数增长。对这些海量的数据进行有组织的搜集和管理进行后续各项研究工作的前提。目前，国际上有许多公用大型数据库，专门负责整理这些序列数据供研究者共享。尤其是是欧洲分子生物学实验室的EMBL (http://www.expasy.ch/sprot/)、美国生物技术信息中心 (NCBI) 的GeneBank (http://www.ncbi.nlm.nih.gov/) 和日本遗传研究所的DDBJ (http://www.ddbj.nig.ac.jp/)。这三个组织相互合作，各数据库中的数据保持一致，对于特定的查询，三个数据库的响应结果一样。

2) 数据库搜索及序列比较：对于许多测序得到的基因序列，研究者们并不知道其对应的生物学功能。生物学研究人员希望能够通过搜索序列数据库找到与新序列同源的已知序列，并根据同源的已知序列来推测新序列的结构或功能，从而发现新的生物分子数据的内涵。搜索同源序列在一定程度上就是通过相似比较寻找相似序列。目前比较常用的方法是BLAST (Basic Local Alignment Search Tool)[3, 4]。

3) 序列分析：核酸序列是遗传信息的基础，而识别编码区域或寻找基因正

是序列分析中最关键的。由于存在大量的DNA序列数据，因此，发展识别编码区域和基因的算法是最大限度利用生物分子数据的重要环节。另外，从实验和计算的关系来看,在有些情况下，由实验测定的编码区域并不一定完整，需结合计算找到并证实所有的外显子 (extron)。从编码区域就可以推导出其对应的蛋白质序列。还有就是生物信息学可以通过相关的算法，找到一些用于识别、翻译和转录特征以及功能位点[5, 6]，比如，启动子、起始编码、剪切位点、内含子、外显子等。

4) 蛋白质结构预测：蛋白质一级序列根据基因的核酸序列就可以得到，但是蛋白质的功能很大程度上与蛋白质的空间结构有关。要预测蛋白质的功能，就必须预测蛋白质的结构。蛋白质结构预测分为二级结构预测和空间结构预测。目前，基于神经网络或HMM (Hidden Markov Model，隐马可夫模型) 预测二级结构的方法已经较为成熟，预测的准确率也相当高。而对蛋白质的空间结构进行预测，难度就比二级结构预测大得多，需要也更加迫切。在蛋白质结构数据库PDB (Protein Data Bank，http://www.pdb.org/pdb/) [7]中，到2006年2月16日，仅有35144个蛋白质的空间结构数据，这一数据与海量的蛋白质序列数据相比是不相称的。在空间结构预测方面，运用同源模型方法可以完成蛋白质10 %～30 %的空间结构预测工作。预测蛋白质结构以后就可以进一步预测、分析和研究蛋白质的生物学功能。

正因为生物信息学在研究基因功能方面，比单纯采用实验手段，具有速度快、成本低廉等优势，因此，将生物信息学和实验相结合来展开生物学研究成为一种更有效、更经济的研究方法。

1.1.2 生物信息学数据库介绍

近年来大量生物学实验的数据积累，形成了当前数以百计的生物信息数据库。它们各自按一定的目标收集和整理生物学实验数据，并提供相关的数据查询、数据处理的服务。随着因特网的普及，这些数据库大多可以通过网络来访问，或者通过网络下载。

一般而言，这些生物信息数据库可以分为一级数据库和二级数据库。一级数据库的数据都直接来源于实验获得的原始数据，只经过简单的归类整理和注释；

二级数据库是在一级数据库、实验数据和理论分析的基础上针对特定目标衍生而来，是对生物学知识和信息的进一步整理。国际上著名的一级核酸数据库有Genbank数据库、EMBL核酸库和DDBJ库等；蛋白质序列数据库有SWISS-PROT、PIR等；蛋白质结构库有PDB等。国际上二级生物学数据库非常多，它们因针对不同的研究内容和需要而各具特色。

1) 基因和基因组数据库

Genbank[8]包含了所有已知的核酸序列和蛋白质序列，以及与它们相关的文献著作和生物学注释。它是由NCBI建立和维护的。它的数据直接来源于测序工我们提交的序列、由测序中心提交的大量EST序列和其它测序数据、以及与其它数据机构协作交换数据而来。Genbank每天都会与EMBL的数据库，和日本的DNA数据库 (DDBJ) 交换数据，使这三个数据库的数据同步。Genbank的数据可以从NCBI的FTP服务器上免费下载完整的库，或下载积累的新数据。NCBI还提供广泛的数据查询、序列相似性搜索以及其它分析服务。NCBI的数据库检索查询系统是Entrez[9]。Entrez是基于Web界面的综合生物信息数据库检索系统。利用Entrez系统，用户不仅可以方便地检索Genbank的核酸数据，还可以检索来自Genbank和其它数据库的蛋白质序列数据、基因组图谱数据、来自分子模型数据库的蛋白质三维结构数据库MMDB、种群序列数据集、以及由PubMed获得Medline的文献数据。

EMBL[10]核酸序列数据库由欧洲生物信息学研究所 (EBI) 维护的核酸序列数据构成，由于与Genbank和DDBJ的数据合作交换，它也是一个全面的核酸序列数据库。该数据库由Oracal数据库系统管理维护，查询检索可以通过因特网上的序列提取系统 (SRS) 服务完成。

DDBJ[11] (日本DNA数据库) 也是一个全面的核酸序列数据库，与Genbank和EMBL核酸库合作交换数据。可以使用其主页上提供的SRS工具进行数据检索和序列分析。

TIGR[12] (The Institute for Genomic Research，基因组研究所数据库) 是国际上重要的测序中心之一，其重点是假设的基因组学和比较基因组的研究。它有大量的基因组数据和标记表达序列数据，其资源以微生物为主，真核生物和人类基因组为辅。TIGR数据库包括了微生物、植物及人类的DNA及蛋白质序列，基因

表达，细胞的作用，蛋白质家族及分类数据，是一套大型综合数据库。在它收录的多种多样的数据库中，微生物基因组数据库是世界上著名的基因组数据库。此外，其还拥有世界上最大的cDNA数据库。

2) 蛋白质数据库

SWISS-PROT[13]和PIR[14]是国际上二个主要的蛋白质序列数据库，目前这二个数据库在EMBL和GenBank数据库上均建立了镜像站点。SWISS-PROT数据库包括了从EMBL翻译而来的蛋白质序列，这些序列经过检验和注释。该数据库主要由日内瓦大学医学生物化学系和欧洲生物信息学研究所 (EBI) 合作维护。SWISS-PROT的数据存在一个滞后问题，即把EMBL的DNA序列准确地翻译成蛋白质序列并进行注释需要时间。一大批含有开放阅读框 (ORF) 的DNA序列尚未列入SWISS-PROT。为了解决这一问题，TREMBL [15] (Translated EMBL) 被建立了起来。TREMBL也是一个蛋白质数据库，它包括了所有EMBL库中的蛋白质编码区序列，提供了一个非常全面的蛋白质序列数据源，但是其注释质量不如SWISS-PROT。PIR数据库的数据由美国国家生物技术信息中心 (NCBI) 翻译自GenBank的DNA序列。

3) 水稻和拟南芥专项数据库

Gramene[16] (谷类比较图谱资源)是一个协助性的、以网络为基础的公开性数据资源，致力于稻科植物类的比较基因组分析。他们的目标是使用公用工程信息促进交叉物种的同源关系研究，这些公用工程包括基因组、EST序列、蛋白质结构和功能分析、遗传学和物理图谱、生物化学通路的阐述、表型特征和突变的QTL定位及描述。作为一个信息源，Gramene的目的是在公共资源中为资料提供更多的价值，便于研究者用水稻基因组序列来鉴定和阐述稻科作物的相应基因、通路和表型。

TAIR[17] (拟南芥信息资源网) 是拟南芥研究工作者的首选网站。该网站提供了一系列对拟南芥基因组进行诸如BLAST、FASTA 等分析的工具。此外,研究者也可以以FTP的形式下载部分数据库并进行本地分析。该网站也包括与其他拟南芥网站的链接。

TIGR Arabidopsis thaliana Annotation Database (TIGR拟南芥注释数据库) 和TIGR Rice Annotation Database[18] (TIGR水稻注释数据库) 由TIGR (基因组研究

所) 维护，包括拟南芥和水稻测序计划的所有序列，这些序列已经以统一的形式被注释。

1.1.3 生物信息学工具介绍

1) 综合工具包

EMBOSS[19] (欧洲分子生物学开放软件系统) 是一个开放源代码的序列分析软件包，是为分子生物学研究的特别需要而发展起来的。该软件能够自动识别处理以不同格式存储的数据，甚至可以通过互联网提取数据，并且，因为该软件包同时提供了一个扩展库，它也是允许其他科学家依据自由软件精神编制、发布软件的一个平台。EMBOSS同时将现在可以得到的一系列序列分析工具整合成一个无缝的整体。在EMBOSS中有一百多个应用程序，它们涵盖了序列分析和联配，数据库检索，蛋白质结构域识别，EST分析等方面。

BioPerl[20]是一组Perl模块，它主要目的在于利用Perl解决生物学研究中的一些问题，如获取分子生物学数据，分析序列文件，序列间比对，大批量BLAST，数据挖掘等。它主要的用途不在于直接提供可以使用的程序，而是提供大量可扩展的模块，使得生物学家可以利用它们很方便的写出满足自己需要的 Perl 脚本。

2) 两两序列比对工具

FASTA[21]是第一个被广泛应用的序列比对和搜索工具包，包含若干个独立的程序。FASTA为了提供序列搜索的速度，会先建立序列片段的“字典”，查询序列先会在字典里搜索可能的匹配序列，字典中的序列长度由ktup参数控制，缺省的ktup=2。FASTA的结果报告中会给出每个搜索到的序列与查询序列的最佳比对结果，以及这个比对的统计学显著性评估E值。

BLAST[22]是现在应用最广泛的序列相似性搜索工具，相比FASTA有更多改进，速度更快，并建立在严格的统计学基础之上。BLAST包含五个程序和若干个相应的数据库，分别针对不同的查询序列和要搜索的数据库类型。其中翻译的核酸库在搜索比对时会把核酸数据按密码子按所有可能的阅读框架转换成蛋白质序列。

3) 多序列联配工具

ClustalW[23]是一个最广泛使用的多序列比对程序，在任何主要的计算机平台上都可以免费使用。这个程序基于渐进比对的思想，得到一系列序列的输入，对于每两个序列进行双重比对并且计算结果。基于这些比较，计算得到一个距离矩阵，反映了每对序列的关系，这个矩阵被用来计算出一个系统发生辅助树。这个辅助树，加权后可以证实极相近的序列，然后以双重比对极相近的序列开始，为组建比对提供基础，然后重新比对下一个加入的比对，依次类推。如果加入的序列较多，那么毫无疑问，必须加入空位以适应序列的差异，但是加入空位必须接受空位开放罚分和空位扩展罚分。在绝大多数情况下，使用者不会在比对时加入结构信息，但是空位开放补偿利用了可以出现在α-螺旋或β-折叠末端的特殊残基以及空位罚分所偏好的残基。

MultAlin[24]也是基于用一系列双重比对开始的思想，然后基于双重比对的打分值进行一个分层次的聚类的软件。当序列都分成类后，开始进行多序列比对，计算出多序列比对中的两个序列比对的新值，基于这些新值，重新构建一棵树。这个过程不断进行，直到分值不再上升，完成序列比对。

4) 系统发育分析方法和工具

三种主要的建树方法分别是距离法、最大简约法 (maximum parsimony, MP) 和最大似然法 (maximum likelihood，ML)。最大似然方法考察数据组中序列的多重比对结果，优化出拥有一定拓扑结构和树枝长度的进化树，这个进化树能够以最大的概率导致考察的多重比对结果。距离树考察数据组中所有序列的两两比对结果，通过序列两两之间的差异决定进化树的拓扑结构和树枝长度。最大节约方法考察数据组中序列的多重比对结果，优化出的进化树能够利用最少的离散步骤去解释多重比对中的碱基差异。

距离方阵方法简单的计算两个序列的差异数量。这个数量被看作进化距离，而其准确大小依赖于进化模型的选择。然后运行一个聚类算法，从最相似 (也就是说，两者之间的距离最短) 的序列开始，通过距离值方阵计算出实际的进化树，或者通过将总的树枝长度最小化而优化出进化树。用最大节约方法搜索进化树的原理是要求用最小的改变来解释所要研究的分类群之间的观察到的差异。最大似然方法评估所选定的进化模型能够产生实际观察到的数据的可能性。进化模型可能只是简单地假定所有核苷酸 (或者氨基酸) 之间相互转变的概率一样。程序会

把所有可能的核苷酸轮流置于进化树的内部节点上，并且计算每一个这样的序列产生实际数据的可能性 (如果两个姐妹分类群都有核苷酸“A”，那么，如果假定原先的核苷酸是“C”，得到现在的“A”的可能性比起假定原先就是“A”的可能性要小得多)。所有可能的再现 (不仅仅是比较可能的再现) 的几率被加总，产生一个特定位点的似然值，然后这个数据集的所有比对位点的似然值的加和就是整个进化树的似然值[25]。

PHYLIP[26]是一个包含了大约30个程序的软件包，这些程序基本上囊括了系统发育的所有方面。PHYLIP是免费软件，并且可以在很多平台上运行 (Mac, DOS，Unix，VAX/VMS及其它)。PHYLIP目前已经是最广泛使用的系统发育程序。

MEGA [27]软件用于检验和分析DNA、蛋白质序列的演化。MEGA强调了序列获得和演化分析的整合；该软件允许多种格式数据输入，用户可以在多个窗口检视结果，进行序列数据的操作和编辑、系列比对和系统发育关系树推断，并进行演化距离估计。结果输出窗口 (results explorers) 允许使用者进行浏览、编辑、总结和输出结果。MEGA还包括距离矩阵、系统发育关系展示窗口 (explorers)，以及一些用于直观呈现输入数据和输出结果的高级图形模块。

开发PAUP[28]的目的是为系统发育分析提供一个简单的，带有菜单界面的，与平台无关的，拥有多种功能 (包括进化树图) 的程序。PAUP可以建立应用MP和ML和距离方法建立进化树，是一个商业软件。

5) 线上结构域预测工具

Pfam[29]是一收录大量序列比较和基于隐马可夫链算法的蛋白质家族比较的数据库及服务器。Pfam 19.0 (2005年12月) 版本包含了8183个蛋白家族连配序列数据和模型，序列数据来自Swissprot 48.1 和 SP-TrEMBL 31.1蛋白数据库。

SMART[30](简单模块架构搜索工具)也是基于隐马可夫链算法的蛋白质家族比较的工具。最初用来研究涉及真核生物信号转导的蛋白质结构域。现已扩展到细胞外蛋白质的活性结构域、细菌调控系统以及与DNA、RNA、染色体和细胞骨架功能有关的结构域。

InterPro[31](蛋白质集成的文档资源) 是一个关于蛋白质家族、结构域和功能位点知识的集成文档资源，将保存在蛋白质结构数据库如PROSITE、PRINTS、

Pfam和ProDom中的信息统一起来，因此只要访问这个站点就可以获得所有这些数据库相关的信息。来自这些数据库的合并的注释构成了InterPro的核心，每个条目包括蛋白质的功能描述和文献参考，以及与相关数据库相连的链接。

1.1.4 植物中的生物信息学研究

水稻和拟南芥作为植物分子生物学研究中两种常见的模式生物，分别是单子叶植物和双子叶植物的代表。它们的基因组相对较小，便于测序和分析。而水稻作为世界最重要的农作物之一，其基因功能和生物信息分析更是具有重要的意义。拟南芥全基因组约为 125 Mbp，其测序工作已于 2000 年完成[32]。而水稻全基因组约 390 Mbp中 95％以上序列的测序和拼接也由IRGSP (International Rice Genome Sequencing Project，国际水稻基因组测序计划) 完成 (http://rgp.dna.affrc.go. jp/cgi-bin/statusdb/irgsp-status.cgi，data of 11 Jul，2005)[33]。除全基因组测序的数据外，还有一些水稻或拟南芥的专有公共数据库可供使用，比如拟南芥的数据库TAIR (http://www.arabidopsis.org/)。

植物中各基因家族分子进化的研究，也在植物生物信息学研究中占很大比重。比如水稻和拟南芥中ERF基因家族的研究[34]，与植物花形态建成相关的MADS-盒基因家族的信息学研究[35, 36]，转录因子家族MYB基因家族的研究[37, 38]等等。

大量的公共数据，以及各种基因家族的信息学分析，给用生物信息学方法研究植物基因的功能打下了坚实的基础。此外在揭示植物各个物种的进化过程和相互间亲缘关系，以及物种内各基因家族成员进化关系方面，生物信息学也正在占据越来越重要的地位[39, 40]。

1.2 bHLH 基因家族研究现状

1.2.1 bHLH 结构域的组成

从Murre 等 1989 年发现了bHLH (basic/helix-loop-helix，碱性/螺旋-环-螺旋)结构[41]以来，研究者们已经在真核生物中发现了许多bHLH转录因子超家族成员。bHLH结构域是一个转录因子特有的结构域。这个结构域的基序包含了约 60

个氨基酸，由一个碱性氨基酸区 (basic region)和一个螺旋-环-螺旋 (HLH region)区组成[41]。碱性区约含 15 个氨基酸，其中含有较多的碱性氨基酸，这个区域的活性主要与转录因子与DNA特定序列的结合有关。螺旋-环-螺旋区的主要作用是与其他蛋白结合形成二聚体，协同行使功能。

1.2.2 动物中的 bHLH 蛋白研究进展

动物中的bHLH蛋白已经有较为深入的研究，bHLH 蛋白的主要功能为调节各种干细胞向终末细胞的分化。它的高表达可促进干细胞的分化，相反低表达则抑制干细胞的分化，使干细胞处于持续的增殖状态。它们在骨骼肌、果蝇的神经干细胞以及脊椎动物的脊髓、端脑皮层的发育过程中发挥着重要作用[42-44]。

已知的动物bHLH蛋白可以根据其结合DNA的活性划分成 6 类[45-47]：1) A类蛋白能结合序列为 5’-CAGCTG-3’ (E-盒序列) 的DNA序列，包括的蛋白有Lyl、Twist、Hen、Atonal、Delilah、dHand、AC-S、MyoD、E12 和Da[46]。2) B类成员如 Myc 、 Max 、 USF 、 SREBP 、 MITF 等可以结合G-盒序列，也就是5’-CACGTG-3’[48-50]。3) C类成员除了bHLH结构域之外，还含有另一个蛋白相互作用的结构域——PAS结构域[51]，这类bHLH蛋白可以结合非E-盒序列5’-NACGTG-3’或 5’-NGCGTG-3’。4) D类由不含碱性氨基酸区的HLH蛋白构成。由于不含有碱性氨基酸区，因此这类蛋白不能结合DNA，但其可以通过竞争性结合其他的bHLH蛋白，起到调节下游基因转录的作用，Id蛋白就属于这一类[52]。5) E类bHLH蛋白包含了Gridlock、E (spl)、Hey和Hairy[47]。这类蛋白在碱性区中含有脯氨酸和甘氨酸，并且可以结合 5’-CACGNG-3’[53, 54]。6) F类成员由含有COE-bHLH结构的蛋白组成，这个结构与DNA结合和二聚体都相关[47, 53, 55]。

动物 bHLH 蛋白功能繁多，各个成员在动物的神经系统和骨骼肌发育等生理过程中发挥非常重要的作用。因此，了解 bHLH 蛋白的功能，对于解释相关的生理和病理现象，促进医学和生物学相关领域的研究将有很大大的帮助。

1.2.3 植物中的 bHLH 蛋白研究现状

在植物中，从控制花青素合成的R基因及其产物Lc蛋白的发现到现在已经过去十多年了。尽管如此，仅仅是很少的一部分bHLH蛋白被发现和研究，这个家