第30卷 第4期
2009年10月

华南农业大学学报
Journal of South China Agricultural University

Vol. 30, No. 4
Oct. 2009

粳稻 MYB 蛋白家族的生物信息学分析

吴家胜[1]，赵彦宏[2]，汪旭升[3]

（1 浙江林学院 林业与生物技术学院，浙江 临安 311300；2 鲁东大学 生命科学学院，
山东 烟台 264025；3 浙江大学 生物信息研究所，浙江 杭州 310029）

摘要：运用隐马尔柯夫模型（Hidden Markov model，HMM），对粳稻 *Oryza sativa* L. ssp. *japonica* 的蛋白质数据库进行搜索，共找到192个MYB蛋白同源序列，其中有126个R2R3-MYB蛋白，6个R1R2R3-MYB蛋白以及60个MYB相关蛋白（MYB-related protein）. 进一步分析所有粳稻的MYB蛋白基因在染色体上的分布，结果发现，MYB基因基本上是成簇分布的，在染色体1～8号特别明显. 运用基于EM（Expectation maximization）算法的MEME程序，对MYB蛋白的功能域进行多序列比对分析，确定了R1、R2与R3结构域在每个位置上最大可能氨基酸的频率. 此外，还对水稻中的MYB蛋白家族作了初步的进化分析.

关键词：粳稻；MYB蛋白；基因家族；功能域；进化分析

中图分类号：Q816　　**文献标识码**：A　　**文章编号**：1001-411X(2009)04-0043-05

Bioinformatic Analysis of MYB Protein Family in *Oryza sativa* L. ssp. *japonica*

WU Jia-sheng[1], ZHAO Yan-hong[2], WANG Xu-sheng[3]

(1 College of Forestry & Biotechnology, Zhejiang Forestry University, Lin'an 311300, China; 2 College of Life Sciencse, Ludong University, Yantai 264025, China; 3 Institute of Bioinformatics, Zhejiang University, Hangzhou 310029, China)

Abstract: 192 MYB proteins were identified in rice genome by searching the *Oryza sativa* L. ssp. *japonica* protein database using Hidden Markov model (HMM) scan strategy. Among them, 126 were R2R3-MYB, 6 were R1R2R3-MYB, and 60 were MYB-related. Further studies showed that the MYB genes were not randomly distributed in chromosomes, but clustered in chromosomes, especially in chromsome 1, 2, 3, 4, 5, 6, 7 and 8. Multiple alignment was carried out with MYB proteins using MEME method and the most favored amino acid at each site of R1, R2 and R3 domain were determined. In addition, a preliminary analysis for evolution of the rice MYB protein family also was performed in this paper.

Key words: *Oryza sativa* L. ssp. *japonica*; MYB protein; gene family; functional domain; evolution analysis

MYB蛋白是植物体内最大的一类转录因子，调控着植物体内众多基因的转录，因此在植物代谢和发育以及抗逆等方面起着重要的作用. MYB蛋白最早是在1982年由Klempnauer等[1]首先从鸟类病毒AMV中发现的. 此后，人们不断地从人、小鼠、酵母、玉米、拟南芥、金鱼草、水稻和棉花等生物中鉴定出与MYB蛋白同源的蛋白[2-3]. 对MYB转录因子的深入研究，将有助于理解基因转录的调控机制.

目前，对MYB转录因子已有了一定的认识，但了解还非常有限，尤其是植物中存在的大量MYB还尚未被鉴定出来. 从植物中发现更多的MYB转录因子则是人们当前的追求目标. Riechmann等[4]和Stracke等[5]曾利用生物信息学的方法在拟南芥中发现了125个R2R3-MYB家族基因成员. Dias等[6]对R2R3-MYB家族在植物中的进化关系进行了研究. Zimmermann等[7]对MYB转录因子在拟南芥整个基

收稿日期：2008-11-27

作者简介：吴家胜（1969—），男，教授，博士；通讯作者：汪旭升（1978—），男，副教授，博士，E-mail：xwang39@utmem.edu

基金项目：国家自然科学基金（30671704，30800682）

因组中的分布进行了研究. Jiang 等[8]对拟南芥和水稻的 MYB 的基因结构进行了研究,认为不同亚类间的内含子和外显子的结构不一致,但在亚类内有相同的基因结构. Chen 等[9]对拟南芥 MYB 家族进行了进化和表达的分析,发现 MYB 相关蛋白在进化上更古老,而且变化迅速,同时发现绝大部分 MYB 基因响应 1 个或多个激素和胁迫处理.

本文基于已测序的水稻全基因组序列和 MYB 蛋白氨基酸序列保守的特性,利用 Pfam 中对应的保守特征序列搜索水稻全基因组,寻找水稻 MYB 相似的候选基因,并对其进行基因保守功能域结构分析和基因定位. 此外,还对所有的水稻 MYB 的基因家族在基因结构、进化上的关系等进行了初步分析.

1 材料与方法

1.1 数据库搜索

粳稻日本晴的基因组数据下载自基因组测序中心 TIGR(http://www.tigr.org)水稻数据库中第5号版本. 籼稻 9311 的基因组数据下载自华大基因组数据库(http://www.genomics.org.cn). 用于隐马尔柯夫模型(Hidden Markov model,HMM)搜索的 MYB 特征文件来自 Pfam(http://www.sanger.ac.uk). 使用软件 hmmer2.2 的 hmmsearch 功能模块,搜索粳稻 IRGSP 蛋白数据库,采用默认参数,将 E(Expectation value)≤ 10^{-20}的序列认为候选蛋白,由相应的程序和索引文件提取所有含 MYB 的蛋白序列,并构建列表.

1.2 序列联配和基因结构分析

利用 SIM4 程序[10],对 MYB 的 CDS 序列与基因组序列进行比较,从而了解基因的内含子和外显子结构. 在此基础上,运用 CLUSTALX (v1.81)程序[11]对水稻 MYB 蛋白家族的氨基酸序列进行多序列联配,程序运行均采用默认参数设置.

1.3 保守功能域的分析

运用基于 EM(Expectation maximization)算法的 MEME 程序[12],对 MYB 蛋白序列进行分析,找出 MYB 蛋白的保守功能域,运行参数采用默认设置.

1.4 系统进化树分析

根据上述多序列联配中的保守部分,使用 BioEdit[13]对联配结果和格式进行编辑. 然后运用 Mega 2.0[14]软件构建 MYB 蛋白家族的系统发育树,采用的算法为 Neighbor-joining (NJ)法. 获取籼稻 9311 各基因的 CDS 序列,使用 PAML 中的 yn00[15]程序计算水稻 2 亚种的 R2R3-MYB 基因的非同义替换率(K_a)和同义替换率(K_s),并计算 K_a/K_s 值,进而分析进化选择.

2 结果与分析

2.1 水稻中的 MYB 相似蛋白

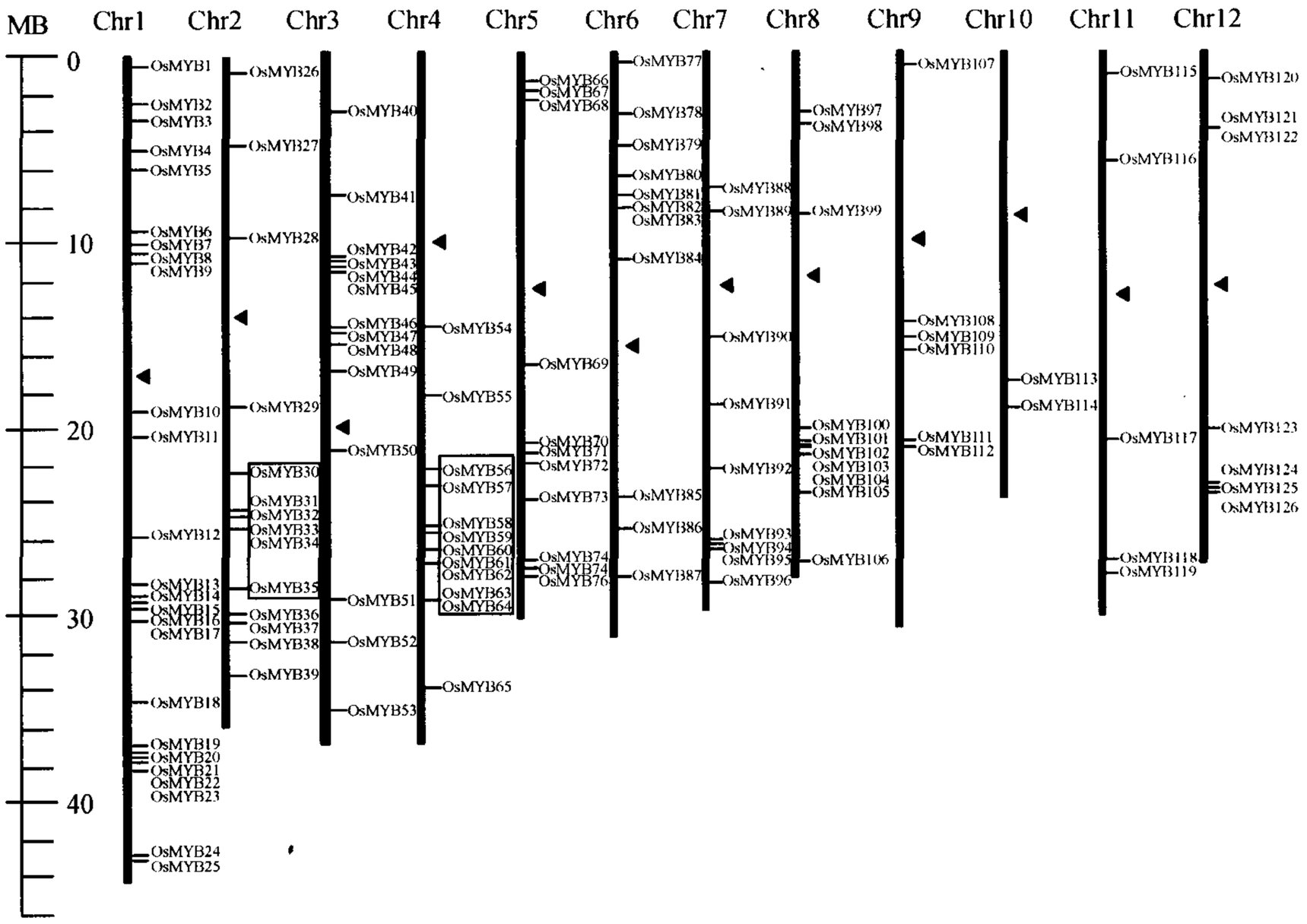
通过 HMM 对 MYB 功能域的搜索,在 IRGSP 数据库水稻蛋白中共鉴定出 192 个属于 MYB 超家族的蛋白同源序列,其中包含 126 个 R2R3-MYB 蛋白,6 个 R1R2R3-MYB 蛋白以及 60 个 MYB 相关蛋白,本文搜索到的 192 个 MYB 基因的详细信息参见 http://ibi.zju.edu.cn/bcl/publish/data/gene_list.xls. 根据这些蛋白基因的类型、在染色体上的位置及顺序来分别对其进行命名. 本文对以前曾在 Genbank 中登录过的 23 个 MYB 蛋白,也分别重新进行了统一命名,具体的命名结果参见 http://ibi.zju.edu.cn/bcl/publish/data/gene_list.xls.

根据 TIGR 提供的基因组序列,本文分析了这 192 个 MYB 超家族蛋白基因在染色体上的分布情况与具体位置,详细结果参见 http://ibi.zju.edu.cn/bcl/publish/data/gene_list.xls. 将其中 126 个 R2R3-MYB 基因在染色体上的分布以图谱的形式给出(图 1). 从染色体分布可知,第 1 号染色体上的 R2R3-MYB 基因最多,占到所有 R2R3-MYB 总数的 19.05%(24/126);第 10 号染色体上的最少,只有 2 个 R2R3-MYB 基因. 从基因组上的分布来看,R2R3-MYB 基因基本上是成簇分布的,尤其在染色体 1 ~8 号上表现得特别明显. 目前 TIGR 中水稻基因组数据(Vension 5.0)已注释的基因达到56 279个,R2R3-MYB 基因约占水稻基因组基因总数的 0.22%(126/56 279).

对所有 126 个 R2R3-MYB 基因在基因组中的复制情况进行了分析,发现总共有 31 对基因发生了复制,大部分是由于水稻在进化过程中发生的染色体复制造成了染色体片段间的重组,从而导致 MYB 基因之间的复制. 2 号染色体与 4 号染色体上的复制尤为明显(图 1). 所有的复制基因都列在 http://ibi.zju.edu.cn/bcl/publish/data/gene_list.xls.

2.2 MYB 蛋白的功能域多序列联配分析

运用 MEME 方法,将包含多个 MYB 重复功能域的 MYB 基因(指 R2R3-MYB 与 R1R2R3-MYB)的功能域序列进行多序列比较,进而推断一致性序列,并且分别确定 R1、R2 与 R3 结构域在每个位置上最大可能氨基酸的频率(图 2). 通过氨基酸频率的计算,确定出了这 3 个结构域最可能的氨基酸序列,同时也推断出这 3 个结构域内各氨基酸的保守程度. 比如,R2 结构域的第 32 ~38 个氨基酸处,保守性很强,这暗示了这些位点对维持 R2 结构域的功能具有重要的作用(图 2b).



图中方框代表 MYB 基因在染色体间的复制

图 1　R2R3-MYB 蛋白基因在水稻基因组上的分布

Fig. 1　The distribution of R2R3-MYB protein on rice genome

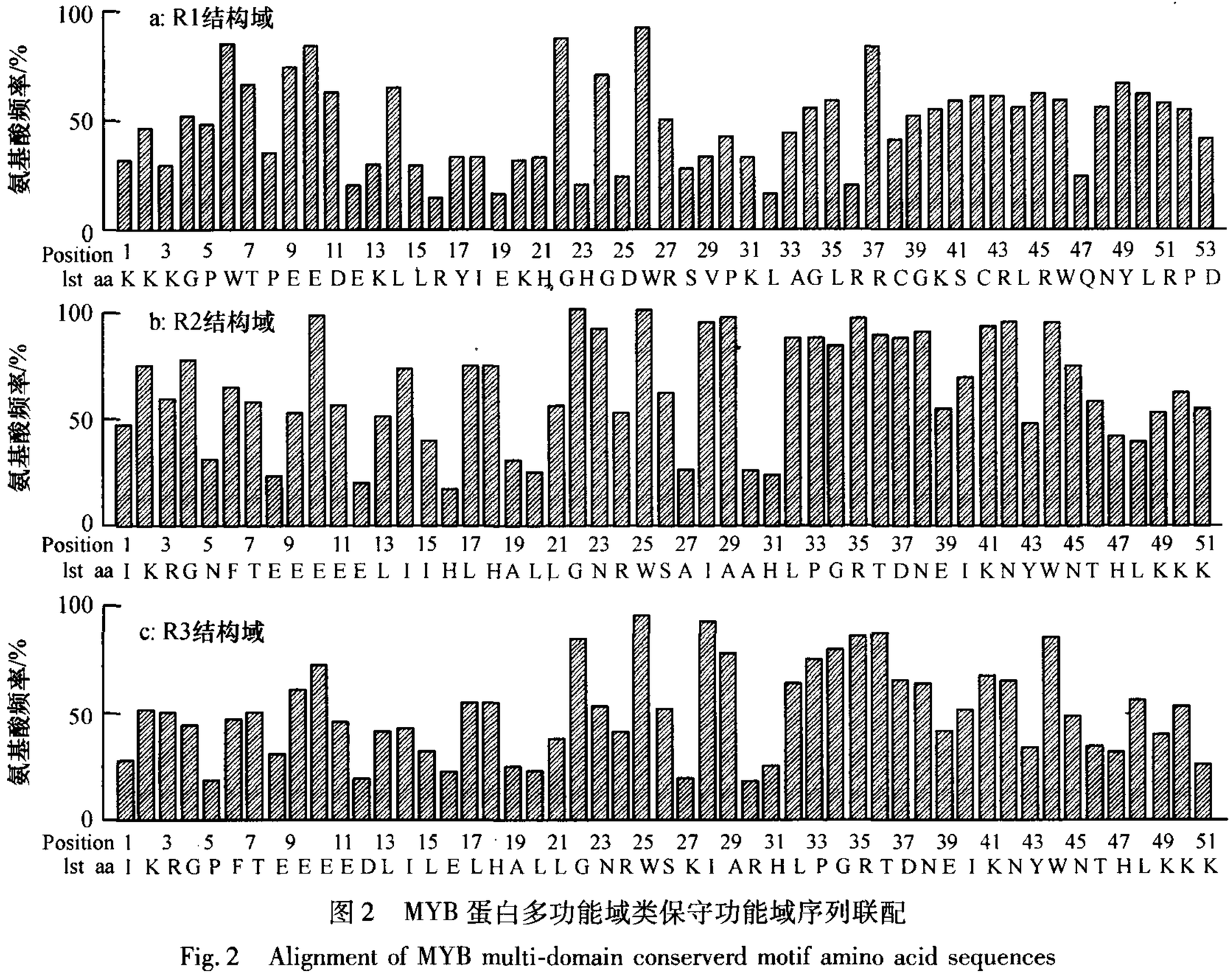


图 2　MYB 蛋白多功能域类保守功能域序列联配

Fig. 2　Alignment of MYB multi-domain conserverd motif amino acid sequences

2.3　系统进化树分析

采用 Neighbor-joining (NJ)方法,对 MYB 蛋白进行了系统进化分析,并构建了系统发育树(图 3). 根据构建的系统发育树,通过对染色体组复制数据的比对发现,水稻 MYB 基因家族有相当一部分基因(52个基因)经历了全染色体的大片段复制. 在图1

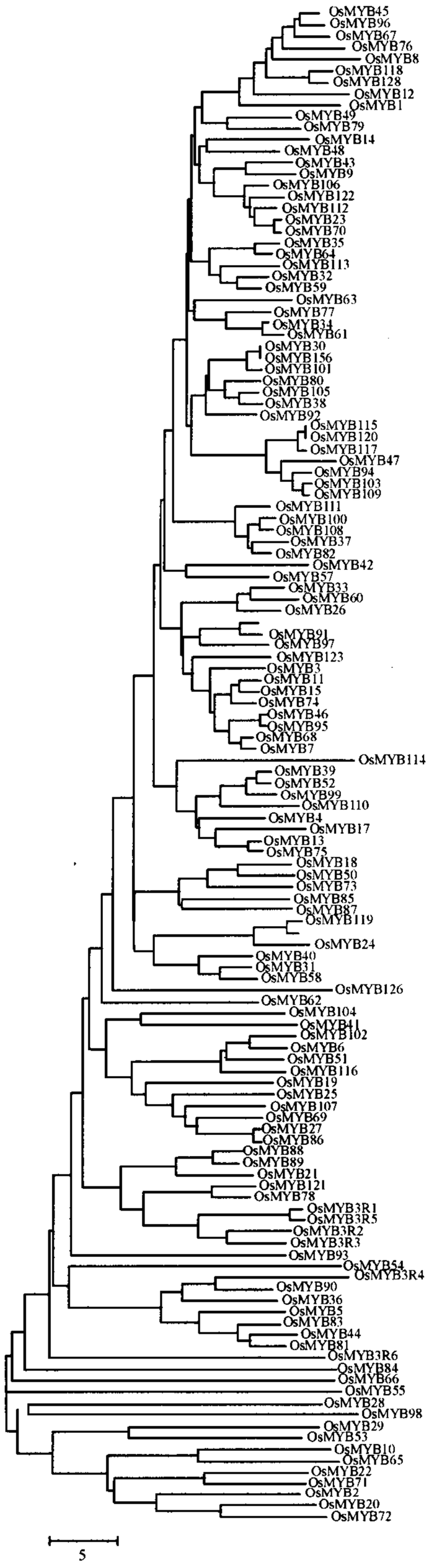


图3 水稻 MYB 蛋白的系统进化树

Fig. 3 The phylogenic tree of MYB proteins in rice

中标出了2段主要的复制区段,该区段包含了6个R2R3-MYB蛋白,这些基因之间的进化关系可以从图3的系统发育树中看到. 水稻 MYB 基因家族在系统发育树上并没有区分出明显的大类,这与拟南芥中的报道类似[5,16]. 值得注意的是,大多数成簇分布的 MYB 基因在系统发育树上分别属于不同的分支,这可能和 MYB 基因簇的不同基因分别控制不同的调控有关. 为了了解水稻 MYB 基因的进化,本文比对了水稻2个亚种的 MYB 序列,找出了51对匹配完好的 MYB 基因,并分别计算出了其 CDS 序列的 K_a 和 K_s 值,发现其中14对基因的 K_a/K_s 值大于1,18对基因的编码序列没有发生任何变化,而剩余19对基因的 K_a/K_s 值小于1,这表明27.45%的水稻 MYB 基因可能受到了正向选择. 具体的 K_a/K_s 数据结果详见 http://ibi.zju.edu.cn/bcl/publish/data/kaksdata.xls.

3 讨论与结论

本文利用 HMM 模型对水稻蛋白质数据库进行搜索,选用 $E \leqslant 10^{-20}$ 作为筛选门槛,发现了192个 MYB 蛋白,并在水稻基因组数据库中找出了其相应的基因,而且将这些基因定位于染色体上,确定了这些基因之间的物理距离. Chen 等[9]也采用了生物信息学的方法在粳稻中找出了183个 MYB 蛋白基因,其中有109个 R2R3-MYB 蛋白基因,4个 R1R2R3-MYB 蛋白基因和70个 MYB 相关蛋白基因,但其在搜索时将满足 $E \leqslant 0.1$ 的蛋白序列看作是 MYB 家族的候选成员. 相比之下,本文在搜索时采用的 E 值更低,将 $E \leqslant 10^{-20}$ 的蛋白看作是候选蛋白,鉴定出的 MYB 蛋白更可靠,假阳性更低. 本文在粳稻日本晴数据库中搜索到 MYB 蛋白基因后,同时又在籼稻9311的基因组数据库中进行了验证. 本文中鉴定出的192个水稻 MYB 蛋白与 Chen 等鉴定出的183个 MYB 蛋白存在着一定的差异.

通过对水稻 MYB 蛋白和拟南芥 MYB 蛋白相比较,发现两者在数量上非常接近,说明被子植物中单、双子叶植物并没有明显差异,可能的原因是单、双子叶植物在分化前 MYB 蛋白的亚类分化已经完成. 但是将水稻 MYB 蛋白系统进化树和结构域结合起来分析后发现水稻中 MYB 基因的相似程度并不是按照上述模型中提出的进化顺序聚类的,不同功能域的蛋白在进化树中交互出现,这表明现有植物中的 R1、R2 和 R3 这几个主要的结构域产生较早,在后期的进化过程中这些结构域的基因之间又发生了大量的重组. 这种大量的基因重组也带来了 MYB 基因家族功能的复杂性. MYB 蛋白进化的保守程度应该与其功能有一定关系. 水稻的 MYB 基因家族由众多的成员构成,属于调控基因,一般来说控制发育和基本代谢的 MYB 基因相对要比控制器官形态的

基因保守得多.

在搜索MYB蛋白的过程中,我们发现了一个编码c-MYB的蛋白,但是没有包括在本文的192个MYB蛋白中.许多单一MYB结构域蛋白的主要功能为转录因子,其中一些与基因表达的昼夜节律变化有关,有一些与根毛的形成有关.

参考文献:

[1] KLEMPNAUER K H, GONDA T J, BISHOP J M. Nucleotide sequence of the retroviral leukemia gene v-myb and its cellular progenitor c-myb: the architecture of a transduced oncogene[J]. Cell, 1982, 31: 453-463.

[2] RABINOWICZ P D, BRAUN E L, WOLFE A D, et al. Maize R2R3 Myb genes: Sequence analysis reveals amplification in the higher plants[J]. Genetics, 1999, 153: 427-444.

[3] GAO Ge, ZHONG Ying-fu, GUO An-yuan, et al. DRTF: A database of rice transcription factors[J]. Bioinformatics, 2006, 22: 1286-1287.

[4] RIECHMANN J L, RATCLIFFE O J. A genomic perspective on plant transcription factors[J]. Curr Opin Plant Biol, 2000, 3: 423-434.

[5] STRACKE R, WERBER M, WEISSHAAR B. The R2R3-MYB gene family in *Arabidopsis thaliana*[J]. Curr Opin Plant Biol, 2001, 4: 447-456.

[6] DIAS A P, BRAUN E L, MCMULLEN M D, et al. Recently duplicated maize R2R3 Myb genes provide evidence for distinct mechanisms of evolutionary divergence after duplication[J]. Plant Physiol, 2003, 131: 610-620.

[7] ZIMMERMANN I M, HEIM M A, WEISSHAAR B, et al. Comprehensive identification of *Arabidopsis thaliana* MYB transcription factors interacting with R/B-like BHLH proteins[J]. Plant J, 2004, 40: 22-34

[8] JIANG Ci-zhong, GU Xun, PETERSON T. Identification of conserved gene structures and carboxy-terminal motifs in the Myb gene family of *Arabidopsis* and *Oryza sativa* L. ssp. *indica*[J]. Genome Biol, 2004, 5: R46.

[9] CHEN Yan-hui, YANG Xiao-yuan, HE Kun, et al. The MYB transcription factor superfamily of *Arabidopsis*: Expression analysis and phylogenetic comparison with the rice MYB family[J]. Plant Mol Biol, 2006, 60: 107-124.

[10] FLOREA L, HARTZELL G, ZHANG Z, et al. A computer program for aligning a cDNA sequence with a genomic DNA sequence[J]. Genome Res, 1998, 8: 967-974.

[11] JEANMOUGIN F, THOMPSON J D, GOUY M, et al. Multiple sequence alignment with Clustal X[J]. Trends Biochem Sci, 1998, 23: 403-405.

[12] BAILEY T L, ELKAN C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers[C]// ALTMAN R C, BRUTLAG D, KARP P D, et al. Proceedings of the 2nd International Conferece on Intellegent Systems for Molecular Biology. Menlo Park: AAAI Press, 1994: 28-36.

[13] HALL T A. BioEdit: A user friendly biological sequence alignment editor and analysis program for windows 95/98/NT[J]. Nucleic Acids Res Symp Ser, 1999: 95-98.

[14] KUMAR S, TAMURA K, JAKOBSEN I B, et al. MEGA 2.0: Molecular evolutionary genetics analysis software[J]. Bioinformatics, 2001, 17: 1244-1245.

[15] YANG Zi-heng, NIELSEN R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models[J]. Molecular Biology and Evolution, 2000, 17: 32-43.

[16] JIN Hai-ling, MARTIN C. Multifunctionality and diversity within the plant MYB-gene family[J]. Plant Mol Biol, 1999, 41: 577-585.

【责任编辑　李晓卉】