生 物 数 学 学 报 2006 ,21(1) : 81-88

Journal of Biomathematics

利用基因组数据分析水稻基因顺式作用元件

张 扬[1] 徐国华[2] 徐 飞[2] 樊龙江[1,2]

(1 浙江大学 IBM 生物计算实验室, 浙江 杭州 310029 ; 2 浙江大学 作物科学研究所, 浙江 杭州 310029)

摘 要: 为了研究籼粳亚种基因调控序列的总体特性, 我们利用籼粳稻以及拟南芥基因组和全长 mRNA 序列获取了大量高可信度的调控序列, 通过这些序列, 分析了水稻基因调控序列顺式作用元件 (信号) 的数量、分布以及与 GC 含量的关系等. 研究结果表明: 一些信号在水稻基因调控序列中发生显著的数量变化, 同时一些信号数量在水稻与拟南芥基因间存在明显差异, 这说明这两种单双子叶植物间信号的使用上存在偏好, 同时水稻不同类型基因以及特有与非特有基因间在信号的使用上也存在差异. 这些差异信号的分布直接导致了调控序列 GC 含量的波动. 本研究没有发现水稻籼粳两个亚种间在调控序列方面 (顺式调节因子和 GC 含量等) 存在明显差异.

关键词: 水稻; 基因组; 顺式作用元件; 籼粳亚种; 拟南芥; GC 含量

中图分类号: Q78; S565 **MR 分类号:** 92D20; 92C40 **文献标识码:** A

文章编号: 1001-9626(2006)01-0081-08

0 引 言

水稻、拟南芥等植物基因组测序的相继完成为许多植物相关研究, 特别是功能基因组研究提供了一个前所未有的机遇和条件. 启动子是基因表达至关重要的部分, 已有大量研究成果和相应的序列数据. EPD(Eukaryotic Promoter Database ; http://epd.isb-sib.ch) 是目前主要的一个由实验获得的启动子序列数据库, 目前 (Release 74) 共有植物启动子序列 198 条, 其中水稻 7 条, 拟南芥 14 条 [1]. 顺式作用元件 (*cis*-acting element) 是基因启动子的重要组成部分, 是真核生物基因启动时转录因子的结合位点 (binding site) , 一般长度为 8-11bp , 所以往往又将这些位点称为信号 (signal) , 目前也有其相应的数据库, 如 PLACE 、 TRANSFAC 等 [2,3]. PLACE (Plant *cis*-acting regulatory DNA elements; www.dna.affrc.go.jp/htdocs/PLACE/) 数据库收录了大量来自实验验证的信号序列, 目前共收录 380 条记录 (2003-6-30 , Release 13.0). 一直以来, 由于这些启动子序列数量非常有限, 根本无法在大尺度上或基因组水平上分析特定物种基因启动子区域的一些信号特征等. 大量基因组序列的出现使这种分析成为可能. 长期以来我们已经对植物基因进行了大量研究, 其相当数量的序列 (大多数为 mRNA 序列) 已被确定并

收稿日期: 2004-03-05

基金项目: 国家自然科学基金项目资助 (30170181 ; 90208022)

作者简介: 张扬 (1979-) , 男, 浙江杭州人, 硕士研究生.

储存在公共数据库 (GenBanK/EMBL/DDBJ) 中，如水稻目前在 GenBank 数据库中约有 1000 条左右全长 mRNA 序列. 利用那些具有全长的基因序列，通过序列联配 (alignment) 的方法就可以将这些基因序列定位到其基因组上，由此可以确定外显子位置并截取该基因的上游启动子序列. 这些序列便可以用来分析特定物种基因启动子区域信号特征等. 目前利用这一技术路线已研究了人类基因的启动信号特征 [4−6]，但在植物上还未见报道.

本文利用了水稻籼粳两个亚种和拟南芥基因组数据等分析了水稻籼粳亚种以及拟南芥基因间顺式作用元件 (信号) 的使用偏好、分布频率以及 GC 含量的变化等，获得了一些水稻基因信号变化规律、特异性和亚种特异性信号特征等，为水稻功能基因组研究提供了重要信息.

1 材料与方法

1.1 数据来源

基因组序列数据分别来自 Syngenta 公司 [7](粳稻草图) 、中科院北京基因组研究所 [8](籼稻草图) 和 TAIR[9] (The *Arabidopsis* Information Resource)(拟南芥)； mRNA 全长序列来自 GenBank 和 TAIR；其他一些数据取自 SWISS-PROT 、 EPD 、 PLACE 等数据库.

1.2 方法

(1) 全长 mRNA 序列及其调控序列的获取. 开展本研究的一个关键步骤是获得一定数量可信度高的水稻基因调控序列，即基因从转录起始位点 (TSS) 上游 2000bp 至下游 100bp 一段序列 (记为 −2000bp∼ +100bp，其中基因转录起始位点为 0 位).

水稻基因的全长 mRNA 序列筛选自 GenBank 核酸数据库. 先从 SWISS-PROT 蛋白质数据库中获取所有水稻基因的 DNA 序列记录号 (ACCESSION NUMBER)，然后根据这些记录号可以从 GenBank 中提取相应的 mRNA 序列. 因为 SWISS-PROT 是目前世界公认可信度最高的数据库，所以成为我们取得全长 mRNA 的主要途径. 但是由于该数据库注释 (annotation) 速度远远落后于 GenBank 数据库 DNA 序列的增长，所以我们通过如下途径从 GenBank 中筛选了另一部分符合我们要求的序列数据：在 NCBI 中打入关键字 oryza sativa 、complete mRNA，然后再在查询得到的序列中删除那些尚未在刊物上发表的、非全长的、 CDS 起始位点从 1 开始 (因为 UTR 不可能为 0) 的序列. 这样我们共获得了 579 条序列，它们组成了我们的水稻全长 mRNA 数据集. 拟南芥基因的全长 mRNA 序列取自 TAIR 数据库，共 6022 条.

水稻基因调控序列的获取：首先利用 BLAST 搜索工具将以上获得的水稻全长 mRNA 序列数据集序列在籼稻和粳稻基因组中定位，得到了每一条 mRNA 具体在哪条或哪几条水稻基因组片段 (contig) 上的信息. 然后根据此信息，使用 sim4[10] 程序将 mRNA 与相应的基因组片段 (contig 序列) 做联配，确定基因转录起始位点，并以此获取转录起始位点附近 (−1999bp∼ +100bp) 区域的序列数据. 调控序列获取的标准：要求 sim4 结果中 mRNA 序列 5' 端必须与基因组序列完全匹配，同时匹配区段的匹配相同率不得低于 90%，剪接信号 (GT/AG) 明显；部分调控序列长度由于 contig 的长度限制不足 2100bp，但不得短于 135bp 即 -35∼ +100bp. 同时去除数据集的冗余性：我们把任意两条相似性大于 95 % 的序列进行去冗余处理 (通过限定 BLAST 搜索时的期望值实现). 最终共获得两个数据集：基于籼稻基因组的籼稻基因调控序列 354 条和

基于粳稻基因组的粳稻基因调控序列 416 条. 两个数据集中均出现的基因共 301 个，将这 301 个基因的调控序列再分别组成两个子集，用于籼粳亚种间的比较分析. 拟南芥基因的调控序列同样取自 TAIR 数据库.

(2) 水稻基因功能分类和特有基因的确定. 对以上获得的水稻基因数据集进行 GO 分类[11]. GO(Gene Ontolgoy, www.geneontology.org) 是目前在人类和水稻等基因组分析中被广泛应用的一种基因功能分类标准，它将基因按照分子功能、生化过程和细胞组分三大类进行具体分类. 使用 GO Editor(Version 1.113)(function ontology, Version 2.594; component ontology, Version 2.303; process ontology, Version 2.670) 对我们的水稻基因数据集 579 条基因进行分类. 对于一些同义异名的条目，如酶的名称等则参考《英汉生物化学及分子生物学词典》(科学出版社，2001) 及一些相应的文献. 对于无法归入具体类别的 53 条序列 (占全部序列的 9.15%)，归入 "molecular function unknown" 条目. 最终在如下 2 级条目名称下分入如下相应的序列数：Enzyme, 206 条，占 35.58%; Cell Growth and/or Maintenance, 116 条，占 20.03%; Binding, 96 条，占 16.58%; Transcription Regulator, 60 条，占 10.36%; Transporter, 56 条，占 9.67%; Molecular Function Unknown, 53 条，占 9.15%; Cell, 43 条，占 7.43%；其他条目 (序列数均小于 30) 略. 根据这一分类可以进一步获取籼稻和粳稻基因相对应的调控序列数据并建立子数据集 (共 $2*7=14$ 个子数据集)(表 1).

表 1　不同 GO 分类条目下籼稻和粳稻的基因调控序列数量

Table 1　Regulatory Sequence Numbers of *Indica* and *Japonica* Rice Genes under GO Classification Items

GO 编号	基因分类	籼稻	粳稻
0010086	transporter	28	35
0012175	enzyme	123	142
0001131	transcription regulator	42	46
0001320	cell growth and/or maintenance	76	83
0008412	cell	24	27
0008800	molecular function unknown	27	34
0009074	binding	62	74

最近水稻基因组的分析结果[8,12,13] 表明，水稻基因组中约有 50% 的基因在拟南芥基因组中通过同源序列搜索可以找到，但另一半基因却找不到，即所谓水稻特有基因 (rice-specific gene). 为了了解水稻基因组中这两类基因在调控序列上差异，我们对 579 条水稻基因在拟南芥基因组上进行 TBLASTN 同源序列搜索，结果返回 494 条 (E 值小于 10^{-5})，另 85 条找不到同源序列. 这两组序列分别组成水稻特有和非特有基因调控序列数据集.

(3) 基因顺式作用元件 (信号) 的搜索和分析. 通过 PLACE 数据库 (http://www.dna.affrc.go.jp/htdocs/PLACE/signalup.html) 的信号搜索 (Signal Scan) 程序进行上述各数据集基因调控序列顺式作用元件的搜索 (2003-01-12). PLACE 返回的搜索结果中，包括了 361 种顺式作用元件特征序列出现的位置、数量等信息. 通过编写相应程序提取并进一步统计出序列中出现的信号位点及其数量的信息. 由此可以计算出所有信号的频率 (fre_n)，即每 1000bp (1kb) 碱基长度出现多少个信号：

$$fre_n = (S_n)/\sum len(AS) * 10^3, \qquad n \in set(PLACE\_SIGNAL)$$

其中 S_n 表示第 n 个信号在该数据集里共出现的次数；AS(All Sequence)：所有序列；$\sum$len(AS) 表示该子数据集中所有序列的长度总和. PLACE_SIGNAL 即为 PLACE 中所有的信号. 由于拟南芥调控序列条数较多 (2984 条)，为了和水稻调控总序列数保持一致性，我们通过均匀分布的抽样办法在拟南芥调控序列中随机抽取了相应数量 (如 301 个记录) 用于与水稻的比较分析.

序列 GC 含量 (X_n) 按如下公式计算：

$$X_n = [(\sum G_n + \sum C_n)/(\sum A_n + \sum G_n + \sum C_n + \sum T_n)]^m, \quad n\text{的取值范围：}(-1999, +100)$$

其中 X_n 表示 m 条序列按照转录起始位点排齐，以 n 个碱基长度为窗口以 1 个碱基长度为步长滑动时，序列块中 $G+C$ 所占的百分比. 在计算时，每条基因序列从 5′ 端到 3′ 端以转录启始位点对齐进行计算. 几个主要信号的组合 GC 含量计算方法同上，只是在计算前先获得每个信号的屏蔽序列，即将每条序列中特定信号未覆盖的部分先屏蔽掉，然后再将这些序列对齐并按一个窗口和步长进行 GC 含量计算.

信号和 GC 含量的背景值均是在相应的基因组中随机 (3 次重复) 截取相应长度或条数的基因组序列并统计得出.

以上计算与分析程序均用 PERL 编写.

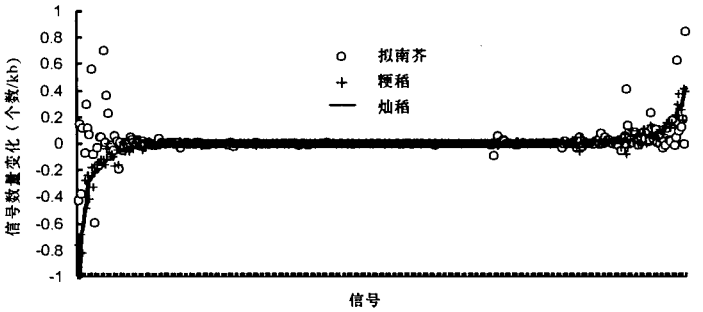


图 1 籼粳稻以及拟南芥基因顺式作用元件数量变化趋势. 横坐标为各个元件 (信号)，并按照籼稻基因信号数量变化排列

2 结果与分析

2.1 水稻顺式作用元件 (信号) 的数量变化

我们对水稻和拟南芥 600bp 长度的启动子序列 ($-499 \sim +100$bp) 以及随机截取的背景基因组序列进行了顺式作用元件 (信号) 的搜索. 图 1 给出了它们的所有信号增加和减少情况 (与它们相应的背景序列相比较). 横坐标的信号排列顺序是按照籼稻基因 (线) 信号数量从最小减少数到最大增加数顺序排列的，粳稻和拟南芥基因对应的信号数量变化情况分别用细线和虚线表示. 同时它们之间信号变化最大的 10 个信号也被列在表 2 中. 图 1 表明，水稻与拟南芥作为单子叶和双子叶植物的典型代表，它们在信号的使用上既有相同的趋势也有着明显的不同. 首先总体上说，它们使用的信号还是相对一致的，即使用数量发生明显变化主要集中在一些信号上，而有些信号大家都不使用或非常稀少. 同时两者明显的不同点是它们在一些信号的使用上有明显的偏好. 在启动子区域，有一些信号，水稻基因的使用数量明显减少，但拟南芥基因明显增加；同时有些信号，水稻使用数量有所增加，而拟南芥则增加更明显 (表 2). 另外还有一个明显的趋势，对于水稻而言，其基因在启动子序列中使用数量或出现次数明显减少的信号，

在拟南芥中有增有减；相反，水稻基因中，使用数量增加的信号，在拟南芥中则只增不减. 这些信号两者都不减少，它们是否是一些重要的和不可或缺的信号？比较这些信号的序列，至少它们在序列碱基的组成上有所不同：水稻基因信号数量增加的或偏好使用是 GC 含量相对较高的信号，而拟南芥基因信号数量增加的信号则 GC 含量相对较低，这样导致的结果是使它们启动子区域的 GC 含量变化趋势有所不同 (详见 2.2 部分). 同时有些信号本身出现的数量就很少，是一些特异表达或特定基因所拥有的信号，这些信号变化规律还有待进一步研究.

除了个别信号有较小的差异外，籼粳稻基因间的信号使用数量上几乎相同 (图 1 和表 2).

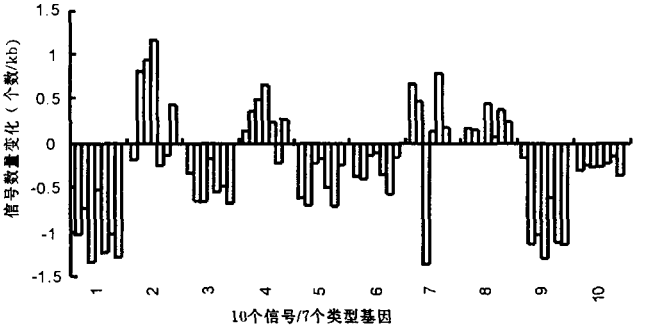


图 2　水稻几个主要顺式作用元件 (信号) 在不同类型基因上的数量变化. 横坐标为 10 个信号序号 (见表 2)，每个信号包括 7 个主要类型基因 (从左到右分别为转录调控基因、细胞生长和维持基因、细胞组分基因、分子功能基因、结合蛋白基因、转运蛋白基因和酶类基因)

水稻不同类型基因的信号使用上也表现出特异性 (图 2). 图 2 仅列出了 7 种类型基因在 10 个主要信号上的数量差异情况. 从这 10 个信号数量变化看，大多信号在各类基因上的数量均有一个趋同趋势，即要么都增加，要么都减少；同时各类基因在信号使用上也有其特异性或偏好，如 S000265 信号，在细胞组分基因中明显减少，但在其他类基因中均有所增加.

表 2 籼粳稻以及拟南芥基因间使用数量变化差异最大的 10 个顺式作用元件 (信号)

Table 2　Ten Signals with the Most Different Usages among *Indica*, *Japonica* and *Arabidopsis*

序号	PLACE 信号编号	信号序列	每千碱基 (每基因) 信号数变化		
			籼稻	粳稻	拟南芥
1	S000039	GATA	−1.02(7.28)	−0.77(7.99)	−0.43(10.36)
2	S000028	CAAT	−0.97(9.01)	−1.04(9.97)	0.15(13.85)
3	S000378	GTGA	−0.67(7.01)	−0.69(7.64)	−0.38(8.14)
4	S000098	ATATT	−0.65(5.17)	−0.82(5.75)	0.12(7.73)
5	S000176	CNGTTR	−0.49(3.07)	−0.27(3.43)	−0.07(3.42)
6	S000203	TTATTT	−0.46(2.46)	−0.48(2.72)	0.29(3.89)
7	S000265	AAAG	0.43(11.42)	0.40(12.40)	0.85(17.77)
8	S000179	CCWACC	0.39(1.46)	0.41(1.62)	0.01(0.91)
9	S000205	CGACG	0.33(2.21)	0.18(2.30)	0.19(1.22)
10	S000337	TAAAG	−0.29(2.62)	−0.24(2.80)	0.12(4.19)

水稻特有和非特有基因间信号数量的变化同样较大 (图 3). 虽然两者间的差异没有水稻与拟南芥基因间大，但可以看出，还是有约有近 10 个信号信号数量明显增加或减少了. 这些信号与前面水稻与拟南芥基因间差异的信号 (表 2) 有所不同，非特有基因信号使用数较特有基因增加的主要有 S000265 和 S000215，而减少的信号主要包括 S000368、S999154、S000122、S000186、S000226 和 S000385 等. 籼粳稻间在这一趋势上基本一致，但在个别信号 (如 S000226) 和程度

变异程度上有所差异.

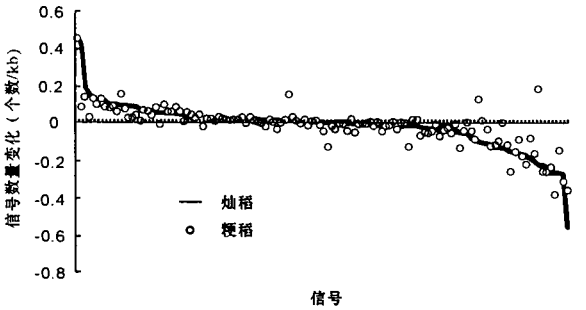


图 3 水稻特有与非特有基因间信号数量的比较. 横坐标的各个信号按照籼稻数据进行排列

2.2 几个主要顺式调节因子的分布以及与 GC 含量变化的关系

如果将信号沿调控序列上出现次数或分布频率画出，即横坐标为信号出现的位点，纵坐标为以 50bp 长度为窗口滑动时信号出现的次数，我们会发现从 -500bp 左右位置开始，信号的频率发生比较大的波动 (数据略). 图 4 给出了水稻和拟南芥调控序列信号数量发生变化最大的 5 个信号分布频率. 可以看出，这 5 个信号在 $-499 \sim +100$ 区域的分布很不均匀，出现频率最高或最低的位点附近一般是这些信号主要的功能位点. 我们已知的一些重要信号，如 "TATA" 盒、"CAAT" 盒等，它们的分布高频位点分别在 -10 和 -70 等位点左右，这与它们的已知功能位点相一致. 根据 600bp 启动子序列信号数量平均，一些主要信号在水稻在单条基因上出现的次数分别为："TATA":2.46，"CAAT": 9.01 等；而根据实际信号搜索结果，启动子序列中不含有如下主要信号的比率分别为："TATA":25%, "CAAT":4% 等.

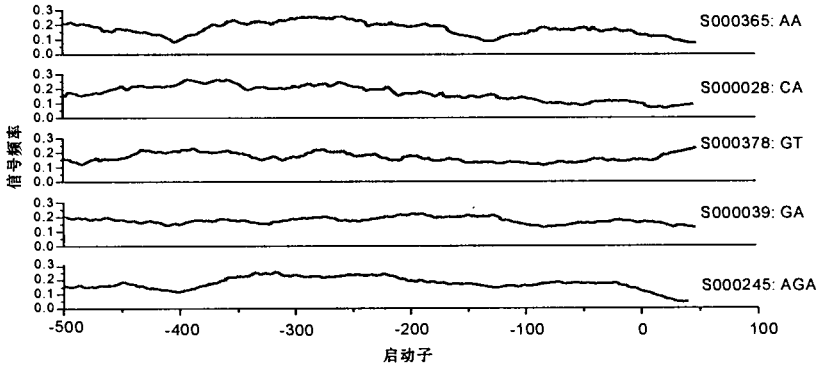
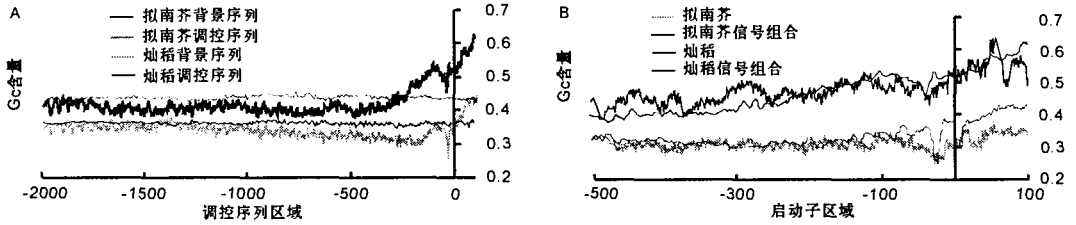


图 4 水稻基因中主要几个信号在启动子区域的分布频率

根据以上一些主要信号分布频率及其序列组成，使我们自然联想到它们与启动子序列 GC 含量变化的关系. 图 5 给出了籼稻和拟南芥调控序列以及它们背景基因组序列的 GC 含量图 (A) 和由水稻基因启动子序列与背景序列信号数量差异最大的 10 个信号的组合 GC 含量曲线 (B). 在基因组中随机获取的背景序列可能含有一定比例 ($< 5\%$) 的编码区序列 (其 GC 含量较高)，其实际 GC 含量应比现含量略低. 如果考虑到该因素，从图 5A 可以看到，水稻和拟南芥启动子序列 GC 含量在 -1200bp 左右的位点上开始与背景序列分离，并在 -500bp 左右位点开始迅速升高，除了在转录启始位点附近 ("TATA" 盒位置) 有一个明显的下降外， GC 含量

一直呈上升趋势. 这一 GC 含量增高趋势将在编码区启始位点处达到最高 [14]. 同时，水稻基因 GC 含量从 −1200bp 位点处就直接开始升高，并一直呈上升的趋势，但拟南芥基因先有一个缓慢下降过程，然后在 −500bp 处才开始迅速增加. 水稻籼粳两个亚种间调控序列 GC 含量和背景序列 GC 含量的差异非常秸叩 C 含量曲线几乎完全重合 (图中未画出). 由 10 个主要差异信号计算的组合 GC 含量曲线与相应物种 (水稻和拟南芥) 基因调控序列平均 GC 含量曲线走势非常吻合 (图 5B)，这说明它们调控序列的 GC 含量变化主要是由相应的信号造成的. 水稻和拟南芥这两个单双子叶植物的这几个信号有相同的，也有不同的，部分信号的使用上有偏好，即水稻使用的是 GC 含量偏高的信号 (如 CCWACC 、 CGACG)，而拟南芥 GC 含量偏低的信号 (如 CAAT 、 ATATT 、 TTATTT 、 TAAAG). 当然这些信号的 GC 含量均平均高于它们的背景序列，只有这样它们的总体 GC 含量才会升高. 基因编码区序列 GC 含量则更高.



A：籼稻 (左) 和拟南芥 (右) 启动子序列以及她们背景基因组序列的 GC 含量

B：利用水稻 (左) 和拟南芥 (右) 启动子序列中 10 个主要信号拟合的组合 GC 含量及其它们的实际 GC 含量

图 5　籼稻和拟南芥启动字序列的 GC 含量变化.

3　讨　论

我们用 mRNA 和基因组联配来获取大规模的调控序列的方法是可行的，而且获取的调控序列的 GC 含量变化和 Yu 等水稻基因组草图分析的结果基本一致 [8]，从而更进一步证明了我们数据的可信性. 通过本研究，获得了水稻基因顺式调节因子的使用偏好、分布等特性，为进一步相关研究，特别是水稻功能基因组研究提供了重要依据；与拟南芥基因的比较研究结果，为探明单双子叶植物基因组上的差异提供了一些线索.

自基因组测序开始以来，籼粳亚种基因组间的差异性一直来是值得探索的一个问题. 通过籼粳亚种同源区段的比较，发现了这两个亚种间在基因组序列上存在一定程度的插入、删除和重排，导致了该区段内个别基因在籼粳亚种的差异 [15]. 这是否是导致籼粳两个亚种间差异的主要来源呢？本研究没有发现水稻籼粳两个亚种间在调控序列方面 (顺式调节因子和 GC 含量等) 存在明显差异. 是否籼粳两个亚种间在信号组合和调控位点方面存在差异还有待进一步研究.

GC 含量是基因组研究中一个重要的单参数指示指标. 本研究证实调控序列的 GC 含量变化主要由基因信号数量和位点的变化造成的. 这些信号序列对植物，特别是水稻基因组基因注释等分析中具有应用价值.

参　考　文　献

[1] Praz V, Périer R C, Bonnard C, Bucher P. The Eukaryotic Promoter Database, EPD: new entry types and links to gene expression data[J]. *Nucleic Acids Res*, 2002, **30**(1): 322-324.

[2] Matys V, Fricke E, Geffers R, et al. 2003 TRANSFAC: transcriptional regulation, from patterns to profiles[J]. *Nucleic Acids. Res*, 2003, **31**(1): 374–378.

[3] Higo K, Ugawa Y, Iwamoto M, Korenaga T. PLACE: Plant *cis*-acting regulatory DNA elements database[J]. *Nucleic Acids Res*, 1999, **27**(1): 297–300.

[4] Suzuki Y, Yamashita R, Nakai K, Sugano S. DBTSS: database of human transcriptional start sites and full-length cDNAs[J]. *Nucleic Acids Res*, 2002, **30**(1):328–331.

[5] Majewski J, Ott J. Distribution and Characterization of Regulatory Elements in the Human Genome[J]. *Genome Res*, 2002, **12**(12): 1827–1836.

[6] Suzuki Y, Tsunoda T, Sese J, et al. Identification and characterization of the potential promoter regions of 1031 kinds of human genes[J]. *Genome Res*, 2001, **11**(5):677–84.

[7] Goff S A, Darrell R, Lan T, et al. A draft sequence of the rice genome (Oryza sativa L. ssp. Japonica)[J]. *Science*, 2002, (296):92–100.

[8] Yu J, Hu S, Wang J, et al. A draft sequence of the rice genome (Oryza sativa L. ssp. Indica)[J]. *Science*, 2002, (296): 79–92.

[9] Rhee S Y, Beavis W, Berardini T Z, et al. The Arabidopsis information resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community[J]. *Nucleic Acids Res*, 2003, **31**(1):224.

[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] The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology[J]. *Nat Genet*, 2000, 25:25–29.

[12] Feng Q, Zhang Y, Wang S, et al. Sequence and analysis of rice chromosome 4[J]. *Nature*, 2002, (420): 316–320.

[13] The Rice Chromosome 10 Sequencing Consortium. In-Depth View of Structure, Activity, and Evolution of Rice Chromosome 10[J]. *Science*, 2003, (300):1566–1569.

[14] Wong G K, Wang J, Tao L, et al. Compositional Gradients in Gramineae Genes[J]. *Genome research*, 2002, (12):851–856.

[15] Han B, Xue Y. Genome-wide intraspecific DNA-sequence variations in rice[J]. *Current Opinion in Plant Biology*, 2003, (6):134–138.

Analysis of Rice Cis-Acting Elements Based on Genomic Sequences

ZHANG Yang[1] XU Guo-hua[2] XU Fei[2] FANG Long-jiang[1,2]

(1 *IBM Biocomputation Lab, Zhejiang University, Hangzhou Zhejiang* 310029 *China*)

(2 *Institute of Crop Science, Zhejiang University, Hangzhou Zhejiang* 310029 *China*)

Abstract: In order to characterize regulatory region of *oryza sativa* ssp. *indica* and *japonica* in genome-scale, we mapped and collected near 600 regulatory sequences which were further used for analysis of rice *cis*-acting element(signal)'s amount, distribution and GC content, based on rice and *arabidopsis* genome and their full-length mRNA sequences under quality control. The results indicate that there are significant quantitative changes of *cis*-acting elements in rice regulatory region. A diverse signal usage between rice and *arabidopsis* are observed. Meanwhile the difference between different types of rice genes or rice specific and un-specific genes in signal usage are also recognized. Those different usages of signals result in fluctuation of GC content in regulatory region of genes. No significant difference in usage of *cis*-acting elements and GC content between two rice subspecies (*indica* and *japonica*) is found in this study.

Key words: *Oryza sativa* ; Genome ; *Cis*-Acting Element ; *Indica* and *Japonica* Subspecies.; *Arabidopsis thaliana*; GC content