浙江大学学报(农业与生命科学版) 32(2): 119~122, 2006
Journal of Zhejiang University (Agric. & Life Sci.)

文章编号:1008-9209(2006)02-0119-04

植物基因转录起始频率分析

张 扬[1], 徐国华[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: 利用 UniGene 数据库转录本序列数据,通过生物信息学方法,对已有全长 mRNA 序列数据的基因进行其转录本 5'端序列的比对,获得该基因编码区前所有转录本的启始位点信息. 通过 7 种植物 17437 个基因的分析表明,植物基因平均在 171 bp(mRNA 水平上)或 174 bp(基因组水平上)的区间内转录启始,转录频率分布基本呈正态分布. 为此我们研发了基因转录启始频率分析程序包 PIFMaker,并基于以上分析获得的数据,建立了植物基因转录频率数据库(PIFdb, http://ibi.zju.edu.cn/bioinplant/). 本研究分析基于该数据库第 2 版(Release 2.0)的数据.

关 键 词: 转录位点; 转录起始频率; UniGene; PIFdb; 水稻; 拟南芥
中图分类号: Q78 **文献标识码**: A

ZHANG Yang[1], XU Guo-hua[2], FAN Long-jiang[1,2] (*1. Institute of IBM Biocomputation Laboratory, Zhejiang University, Hangzhou 310029, China; 2. Institute of Crop Science, Zhejiang University, Hangzhou 310029, China*)

***In silico* analysis of transcriptional initiation frequency of plant genes.** Journal of Zhejiang University (Agric. & Life Sci.), 2006, 32(2): 119-122

Abstract: Transcriptional initiation frequency of 17,437 genes from 7 plants was estimated based on data of UniGene database using bioinformatics method. The results indicated that average distance of transcriptional initiation of plant genes was 171 bp (at mRNA level) or 174 bp (at genomic DNA level) and distribution of their transcriptional initiation frequencies presented a normal distribution in general. A pipeline program, PIFMaker and a database, PIFdb (Potential transcriptional initiation frequency database), were developed in this study. The data of Release 2.0 of PIFdb database was used in this analysis.

Key words: transcriptional initiation site; transcriptional initiation frequency; UniGene; PIFdb; *Oryza sativa*; *Arabidopsis*

真核生物基因往往可以在不同的位点上开始转录,也就是说其转录起始位点是分布在一定的区域内,而非某一固定的位置. Suzuki 等人[1]使用寡核苷酸加帽方法精确获得了不同条件下 mRNA 转录起始位点的信息,并用此方法测序了人的 276 个基因的 5880 条 mRNA 序列,表明人的基因转录起始位点在平均长度为 61.7 bp 区间内起始. 通过该方法进行基因转录起始位点多态性分析,需要大规模测序,资金和时间投入巨大,目前只有在人类基因方面进行了该研究.

收稿日期: 2004-10-16
基金项目: 国家自然科学基金资助项目(30170181;90208022).
作者简介: 张 扬(1979—),男,浙江杭州人,硕士研究生,从事植物基因组学与生物信息学方面的研究.
通讯作者: 樊龙江,男,教授,从事作物遗传育种与植物基因组学方面的研究. Tel: 0571-86971730; E-mail: fanlj@zju.edu.cn.

通过几十年来分子生物学家的不懈努力，特别是近年来开展的大规模测序，目前国际公共核酸数据库(NCBI/EMBL/DDBJ)中已有大量植物基因mRNA序列，其中相当部分是全长序列，特别是水稻和拟南芥大规模全长cDNA测序项目的完成，这两种植物基因的全长mRNA序列分别达到1.5万和3万条左右[2-3]。同时植物EST测序项目也产生了大量EST序列。美国国家生物技术信息中心(NCBI)利用上述序列，通过序列比对等生物信息学方法对这些序列进行了聚类，将来自特定物种的某一特定基因的所有转录本(序列)搜集在一起，建立了一个所谓"UniGene"数据库[4](ftp://ftp.ncbi.nlm.nih.gov/repository/unigene)。通过分子生物学手段进行基因调控序列的分析由来已久，并在实验数据的基础上建立了一些基因启动子序列(如EPD数据库)[5]和人类基因转录起始位点数据库DBTSS[6]等等。

本研究在UniGene数据库的基础上，通过比对某一特定基因的所有转录本5'端序列，并以全长mRNA记录序列为标准，获得了该基因的编码区前所有转录本的启始位点信息。通过7种植物17437个基因的分析表明，植物基因的转录平均在222 bp的区间内启始，转录频率分布呈正态分布。基于上述分析获得的数据，建立了植物基因转录频率数据库(PIFdb, http://ibi.zju.edu.cn/bioinplant/)。本研究分析结果基于该数据库第2版(Release 2.0)的数据。

1 材料与方法

1.1 材 料

本研究所用序列来自：①NCBI的UniGene数据库(ftp://ftp.ncbi.nlm.nih.gov/repository/unigene)。该数据库目前版本包括48000个记录。每一个Unigene记录都有一个参照序列(Reference)，该序列通过一定标准确定，是该基因记录中具有最高测序质量的最长序列，其中相当部分Reference来自全长cDNA测序项目[2,3]；②GenBank/EMBL核酸序列数据库；③NCBI的RefSeq数据库。RefSeq数据库是NCBI搜集和注释的高质量全长mRNA序列数据库；④水稻和拟南芥基因组序列：水稻基因组数据来自美国基因组研究所(TIGR)(osa1, version1.0, http://www.tigr.org/tdb/e2k1/osa1/pseudomolecules/info.shtml)，拟南芥基因组5条染色体序列数据来自GenBank。

1.2 方 法

序列分析程序cap3[7]用于本研究中启始位点信息的确定。以每个UniGene记录中的"Reference"序列为模版，与其他序列进行比对。

分析流程：①原始数据下载：从UniGene数据库下载水稻、拟南芥、大麦、小麦、大豆、玉米、马铃薯共7个植物所有记录；从EMBL中下载相关基因记录；从RefSeq数据库下载所有植物基因记录(ACCESSION号以NM_开头)；②全长mRNA序列的筛选：为了保证获得的每个基因转录区间的准确性，仅对已有全长mRNA序列的植物基因进行分析。采用多个途径来保证和筛选全长mRNA序列：UniGene记录中必须含有来自植物大规模全长cDNA测序项目的序列(如水稻和拟南芥)或RefSeq数据库序列，对没有上述序列的物种，进行了人工注释和筛选。筛选标准为：必须包含开放阅读框(ORF)和5'端序列不以ATG开头，即5'端的非编码区域(5'UTR)长度不为零；③序列比对：利用cap3进行转录本序列的比对，并提取相应位点信息；④基因组定位：利用sim4将步骤3确定的位点信息定位到基因组水平上。由于数据问题，目前只对水稻和拟南芥基因进行了定位；⑤PIFdb数据库记录的产生。至少含有3个或3个以上位点数据的基因才产生记录。

以上分析流程均用PIF Maker程序包完成。该程序用PERL编写，适合于LINUX和WINDOWS系统下运行，可以用于来自任何物种的基因转录启始频率分析。

2 结果与分析

2.1 植物基因转录启始区间

根据制定的筛选标准和现有的序列数据，共对来自水稻、拟南芥、大麦、小麦、大豆、玉米、马铃薯共7个物种的17437个基因进行了分析，获得了相应的PIFdb数据库记录(版本

2.0)．该版本记录中各物种数量见表 1．

表 1　植物基因转录启始区间大小

Table 1　Distance of transcriptional initiation region

物种	基因数量/个*	转录启始区间/bp**	
		平均值	中值
水稻	3127(2767)	121(129)	89(87)
大麦	151	(174)	(118)
小麦	324	(231)	(187)
玉米	113	(167)	(113)
拟南芥	10722(10702)	118(126)	93(90)
大豆	129	(253)	(227)
马铃薯	168	(139)	(84)
合计(平均)	17437(14354)	171(174)	130(129)

注：数据来自本研究产生的 PlFdb 数据库记录(2.0 版)．*括号内数据为有基因组水平转录启始位点数据的基因数量；**括号内数据基于 mRNA 水平上的位点数据统计．

分别统计各个物种基因转录起始区域的平均长度．转录起始区域定义为基因最上游起始位点到最下游位点的区间．表 1 结果表明，植物基因的转录起始区域在 mRNA 水平上平均在 171 bp 之间，7 个物种间大小有所不同，其中水稻和拟南芥分别为 121 和 118 bp．在基因组水平上，植物基因的转录起始区域在 174 bp 之间，水稻基因较拟南芥基因略小，平均值分别为 129 bp 和 126 bp．它们的长度分布呈偏态分布，起始区域长度大部分集中在 100 bp 以内，其中水稻中转录起始区域长度为 60 bp 最多，占 13.7%，而在拟南芥中则是 80 bp，占 13.0%(图 1)．另外 5 个物种转录起始区域的平均长度基本在 100～300 bp．

2.2　植物基因转录启始频率

进一步分析水稻和拟南芥基因组水平上基因转录起始的频率(图 2)．结果显示，它们的转录起始位点基本上在参考序列 5'端(即 0 位点)附近，形成明显的一个单峰，同时在 15～23 bp 的位置上水稻和拟南芥基因绝大多数集中在－50～＋100 bp 的区域内(图 2 右列)；与编码起始位点(ATG)对齐得到的结果表明(图 2 左列)，水稻和拟南芥基因大多在距 ATG －100 bp的范围内转录起始，水稻基因在－57 bp位点上起始频率最高，拟南芥基因的最高频率位点则在略后一点的位置上(－39 bp)．

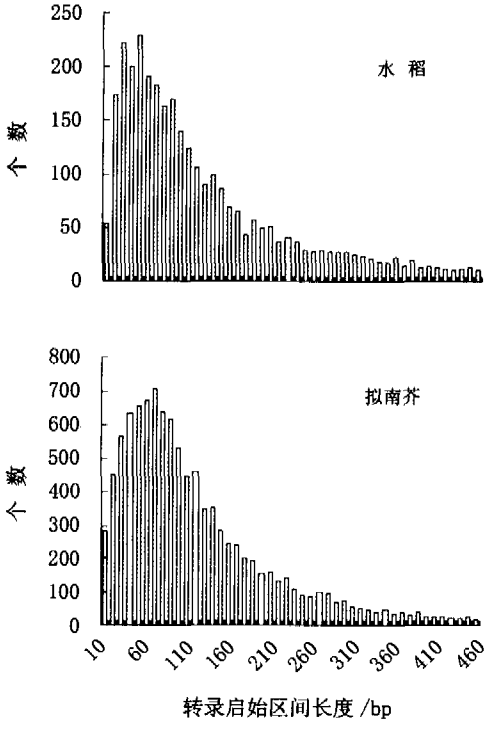


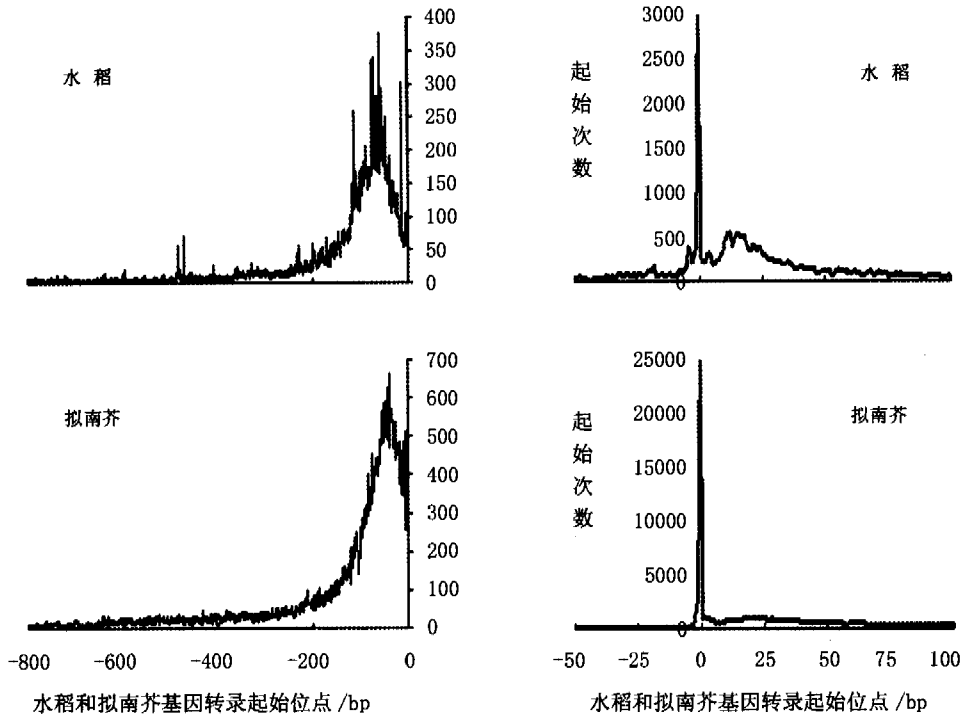
图 1　水稻和拟南芥基因转录启始区间长度分布

Fig. 1　Distribution of transcription initiation distances of rice and *Arabidopsis* genes

3　讨　论

本研究的一个关键问题是分析获得的基因转录位点信息是否真正反映了基因的实际转录状况，也就是说如何保证数据的准确性和进行质量控制．其基本思路是：世界各地的分子生物学家对某一基因表达序列进行研究，他们在不同的环境处理条件和植物生长发育阶段等条件下分别获得了该基因的转录本，其中部分转录本序列是从 5'端测序，这些序列或长(也许是全长序列)或短(如 EST 序列)，其 5'端起始位点也有所不同，这些差异实际是由于基因在不同条件下差异表达的结果，它们所有起始的位点代表了该基因所有可能起始的位点，不同位点的不同起始次数则代表了该基因转录起始的特征．所以，本研究所获得的信息基本反映了基因的实际转录状况，给出了基因转录的一个大致结构．本研究数据可能的误差主要来自两个方面，一是 EST 的测序误差，二是 5'端测序的完整性．为此，考虑了如下质量控制途径：①选用高质量的基础数

据:本研究选用 UniGene 数据库序列做为基础数据,保证了序列的可靠性. UniGene 数据库数据的产生流程中有一个完善和严格的序列质量控制标准,排除了载体等可能造成的测序误差;②为了对转录区间的严格界定:制定了严格的基因全长 mRNA 序列筛选标准,只有真正的全长 mRNA 序列才能进入分析流程中. 以全长 mRNA 序列为模版进行分析,就可保证转录区间的可靠性;③至少含有 3 个或 3 个以上位点数据的基因才产生 PIFdb 记录并用于本研究.



左列:基因序列按编码起始位点 ATG 对齐;右列:基因序列按 PIFdb 数据库记录的"0"起始位点对齐.

图 2 水稻和拟南芥基因转录起始位点频率

Fig. 2 Transcription initiation frequency of rice and *Arabidopsis* genes

本研究得到的植物基因转录起始区间大约在 120 bp 左右(如水稻 121 bp 和拟南芥 118 bp),大约是人类基因平均长度 61.7 bp 的两倍[1].这说明植物基因较人类基因可能在更宽泛的区间内转录起始.同时,它们起始区间长度的分布特征基本一致.

References:

[1] 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):677-684.

[2] Motoaki Seki, Mari Narusaka, Asako Kamiya, *et al*. Functional annotation of a full-length arabidopsis cDNA collection [J]. **Science**, 2002,296: 141-145.

[3] Kikuchi S, Satoh K, Nagata T, *et al*. Collection, mapping, and annotation of over 28000 cDNA clones from japonica rice [J]. **Science**, 2003, 301: 376-379.

[4] Pontius J U, Wagner L, Schuler G D. UniGene: a unified view of the transcriptome [M]//Bethesda. **The NCBI Handbook**. National Center for Biotechnology Information, 2003.

[5] Christoph D, Schmid, Viviane Praz, *et al*. The Eukaryotic promoter database EPD: the impact of in silico primer extension [J]. **Nucleic Acids Research**, 2004, 32: 82-85.

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

[7] Huang X, Madan A. CAP3: A DNA sequence assembly program [J]. **Genome Res**, 1999,9:868-877.