# 全基因组预测目标基因的新方法及其应用

张菁晶[1], 冯　晶[2], 朱英国[1], 李阳生[1]

(1. 武汉大学生命科学院植物发育生物学教育部重点实验室, 武汉 430072; 2. 武汉大学高科技研究与发展中心, 武汉 430072)

**摘　要:** 运用隐马尔可夫模型, 利用Perl编程, 以几种模式生物的蛋白质数据库为基础, 构建了目标基因的全基因组预测的新方法。该方法具有高通量, 准确度高且操作简易等优点, 特别在多结构域蛋白家族预测上更显优势。应用该方法对几种模式生物的全基因组 PPR 和 TPR 蛋白家族进行了预测, 其中粳稻日本晴中含有 536 个 PPR 蛋白、199 个 TPR 蛋白; 籼稻 9311 中含有 519 个 PPR 蛋白、177 个 TPR 蛋白; 拟南芥中含有 735 个 PPR 蛋白、292 个 TPR 蛋白; 红藻中 6 个 PPR 蛋白、32 个 TPR 蛋白; 蓝细菌以及古细菌中没有 PPR 蛋白, 但蓝细菌含有 10 个 TPR 蛋白, 古细菌有 4 个 TPR 蛋白, 并对所得结果进行了进一步生物信息学分析。

**关键词**: 基因预测; Perl; HMM; 全基因组; PPR; TPR

**中图分类号:** Q75　　**文献标识码:** A　　**文章编号:** 0253-9772(2006)10-1299-07

# A Novel Method of the Genome-Wide Prediction for the Target Genes and Its Application

ZHANG Jing-Jing[1], FENG Jing[2], ZHU Ying-Guo[1], LI Yang-Sheng[1]

(1. *Key Laboratory of Ministry of Education for Plant Developmental Biology*, *College of Life Sciences*, *Wuhan University*, *Wuhan* 430072, *China*; 2. *Advanced Research Center for Science and Technology*, *Wuhan University*, *Wuhan* 430072, *China*)

**Abstract:** Based on the protein databases of several model species, this study developed a new method of the Genome-wide prediction for the target genes, using Hidden Markov model by Perl programming. The advantages of this method are high throughput, high quality and easy prediction, especially in the case of multi-domains proteins families. By this method, we predicted the PPR and TPR proteins families in whole genome of several model species. There were 536 PPR proteins and 199 TPR proteins in *Oryza sativa* ssp. *japonica*, 519 PPR proteins and 177 TPR proteins in *Oryza sativa* L. ssp. *indica*, 735 PPR proteins and 292 TPR proteins in *Arabidopsis thaliana*, 6 PPR proteins and 32 TPR proteins in *Cyanidioschyzon merolae*. *Synechococcus* and *Thermophilic archaebacterium* did not have PPR proteins. By contrast, 10 TPR proteins were found in *Synechococcus* and 4 TPR proteins were found in *Thermophilic archaebacterium*. Moreover, of these results, some further bioinformatics analyses were conducted.

**Key words:** gene prediction; Perl; HMM; genome; PPR; TPR

随着人类基因组测序计划的实施, 拟南芥、水稻等模式生物的全基因组测序的相继完成, 许多物种的基因组学研究工作正在全面展开。根据美国国家生物信息中心(NCBI)的资料(http://www.ncbi.nlm.nih.gov/

**收稿日期:** 2006-04-25; **修回日期:** 2006-06-19

**基金项目:** 国家重点基础研究发展规划(973 计划)项目(编号: 2001CB108805)、创新研究群体科学基金(编号: 30521004)、长江学者和创新团队发展计划资助(PCSIRT)[Supported by Key Project of Chinese National Programs for Fundamental Research and Development (973 Program)(No. 2001CB108805), the National Natural Science Foundation of China for Innovative Research Team (No. 30521004) and the Program for Changjiang Scholars and Innovative Research Team]

**作者简介:** 张菁晶(1981－), 男, 湖北人, 硕士, 专业方向: 发育遗传学。E-mail: zjj33@yahoo.com.cn

**通讯作者:** 李阳生(1964－), 男, 湖南人, 教授, 博士生导师, 研究方向: 发育遗传学。E-mail: yangshengl@yahoo.com.cn

genomes/static/gpstat.html), 目前有 1 113 种生物正在进行基因组的测序工作, 其中原核生物 890 种, 真核生物 223 种; 已完成全基因组测序的生物有 330 种, 其中原核生物 310 种, 真核生物 20 种。基因组学技术的进步导致了生物学数据量的大爆炸, 对如此巨大和宝贵的生物学数据如何充分的加以利用, 是当今生物学面临的新的挑战和难题。尽管人类现在对基因的了解达到了前所未有的程度, 但是对几种模式生物基因组测序结果的分析却表明还有大量功能未知的基因存在。因此, 阐明这些基因的未知功能是后基因组时代的一个重要研究课题。如果能对目的基因进行有效的全基因组预测, 再将得到的序列加以实验验证, 对于新基因的发现和功能分析, 其针对性、正确性和工作效率将会得到明显的改善。

目前对具有某一类功能的目的基因的全基因组预测的方法主要有基于序列相似性的Blast搜索法 [1], 进化分析法 [2,3], 基于隐马尔可夫模型的HMMer搜索法 [4], 神经网络预测法 [5]和机器学习预测法 [6,7]等。其中使用最广泛的预测方法主要是依靠Blast的相似性分析, 根据和已知同源基因序列的相似度来打分, 高于一定域值的序列则被判断为候选基因, 如文献[8～10]的文章都用到了该方法。但是该方法存在一个难以避免的缺点, 即两个核酸的全序列的相似度较高, 并不一定意味着它们就具有相同或类似的核心功能模块, 也许它们的核酸序列中起着表达调控的那部分关键序列的相似度并不高; 而两个相似度低于设定域值的序列, 却有可能拥有相同的功能模块, 只是因为关键序列相对于它们的全长来说并不长, 其他部分序列相似度不高, 在做全序列Blast分析的时候, 这条序列就有可能因为低于域值标准而被筛选掉了。这样就会导致一部分候选序列并不是需要的结果, 而另一部分具有相同或类似功能的基因却因为全序列相似度不高而被漏掉。基于以上的考虑, 我们建立了以序列中功能模块/结构域(domain)为基础的基因全基因组预测系统。结构域是指蛋白质序列中局部的保守区域, 拥有相同结构域的蛋白质大部分都有着相同或类似的功能。在本文中, 我们将运用隐马尔可夫模型(HMM)[11,12], 利用Perl编程, 以已经完成全基因组测序的模式生物的蛋白质数据库为基础, 构建具有特定保守结构域的目的蛋白质全基因组预测系统。该预测方法所用程序源代码可于ftp://ricelab.vicp.net免费下载, 也可与作者联系索要。

为了详细阐明该预测方法的应用与实施过程, 我们将挑选两大蛋白家族进行了全基因组预测。其中之一是PPR(pentatricopeptide repeat)蛋白家族。PPR蛋白是由 35 个氨基酸组成的序列单元经串联重复排列而组成的一个基因家族, 这些蛋白中的一大部分都被预测为线粒体或叶绿体的靶标 [13]。近年来的研究发现, 植物细胞质雄性不育的育性恢复基因都含有PPR结构域 [14~17]。与PPR蛋白类似 [18], 另一大蛋白家族——TPR(tetratricopeptide repeat)蛋白是由 34 个氨基酸组成的序列单元经串联重复排列组成, 该蛋白家族在原核和真核细胞中均广泛存在, 参与许多重要的生命活动 [19]。TPR结构域可能介导蛋白之间的相互作用, 对于某些蛋白复合物的形成非常重要, TPR蛋白中串行排列的多个TPR序列可能分别介导与不同蛋白的作用, 从而使TPR蛋白在不同的情况下发挥不同的功能 [20]。TPR基序还可能成为蛋白复合物构成中的重要的桥梁, 在钙调素与蛋白结合的过程中发挥着重要作用 [21]。如果能找出模式生物全基因组所有的PPR和TPR蛋白, 将对研究植物细胞核与细胞质相互作用、细胞质雄性不育的育性恢复的机理及蛋白质互作起到很大的促进作用。

# 1 材料和方法

## 1.1 全基因组分析系统开发环境

考虑到该方法的通用性和易用性, 我们没有采用生物软件常用的 Linux 系统, 而是在 PC/Windows 上进行程序开发。PC 机配置为 CPU Intel Pentium 4 3.0CG/内存 2G/硬盘 160G SATA×2(RAID0)。编程语言采用 Perl, Web 开发软件为 dreamweaver。

## 1.2 全基因组预测的相关程序设计

为了达到高通量的预测要求, 避免遗漏掉可能的目标序列, 我们采用了提取 NCBI (http://www.ncbi.nlm.nih.gov/)的 Protein-protein BLAST (blastp)后台蛋白数据库的方法, 并加入了 TIGR(ftp://ftp.tigr.org/pub/data/)以及 RiceGAAS (ftp://ftp.dna.affrc.go.jp/pub/RiceGAAS)的蛋白数据库作为对比和补充。同时为了方便而有效的进行预测, 我们利用 Perl 和 CGI(公用网关接口)技术开发一个可通过 web 服务器预测的方法, 主要编译了两个程序, 一个是 ExtractCDS_Fasta.pl, 它负责将下载的 Fasta 格式的蛋白序列格式化为我们预测所需的格式; 另一个是

SearchResult.pl, 它是蛋白结构域预测的主程序, 负责将格式化后的序列逐个自动提交给网上的 HMM 服务器进行计算, 并将含有符合我们设定的关键词的结构域的蛋白序列保存下来。我们选用了位于 Washington University in St.Louis 的 Pfam 数据库(http://pfam.wustl.edu/hmmsearch.shtml)。该数据库是囊括了多重序列比对和隐马尔可夫模型方法组建的蛋白质家族数据库, 并基于 Swissprot 48.1 和 SP-TrEMBL 31.1 蛋白质数据库, 目前的版本是 19.0 (December 2005), 含有 8183 个蛋白家族的比对和模型数据。

### 1.3　几种模式生物 PPR 和 TPR 蛋白家族的全基因组搜索

我们从 NCBI 蛋白质数据库中调出水稻粳稻日本晴、拟南芥、红藻、蓝细菌、古细菌的全基因组蛋白序列, 下载到本地。从 Beijing Genomics Institute (BGI)(http://rise.genomics.org.cn/rice2/link/download.jsp)网站上下载水稻籼稻 9311 的全基因组蛋白序列, 利用 Perl 语言编写的 ExtractCDS_Fasta.pl 程序, 格式化全部蛋白序列; 利用 SearchResult.pl 程序添加接口, 设定关键词为 PPR 和 TPR, 使其自动提交单个蛋白序列, 在 Pfam 服务器(http://pfam.wustl.edu/hmmsearch.shtml)上完成全基因组蛋白结构域预测, 并将符合要求含有 PPR 和 TPR 结构域的蛋白序列及其登录号自动保存。最后, 每一个物种将分别得到一个所有含有 PPR 和 TPR 结构域的蛋白序列。预测程序中的关键词具体设置见表 1。

### 1.4　序列分析和系统发生树构建

多重序列比对分析使用 clustalx1.81 软件, 并参照 HMMER2.3.2 软件包(http://hmmer.wustl.edu/)中 hmmalign.exe 程序使用 TPR 或 PPR HMM Profile 的比对结果。序列联配结果经过人工校正后利用邻接法(NJ)自展后构建系统发生树。

### 1.5　结果分析所需软件

为了验证该方法的准确性和全面性, 我们编写了比对程序 compare.pl, 将本研究得到的拟南芥 PPR 蛋白全基因组分析结果与文献[13]的结果进行分析比较。该程序能找出两组序列中完全相同和不同的序列, 并分别将序列的名字和内容保存下来。为方便研究, 我们构建了本地化的 Blast (ftp://ftp.ncbi.nlm.nih.gov/blast/)。用 blastall.exe 程序进行比对, 设定 $E$<1e-200, 即 $10^{-200}$。

## 2　结果与分析

### 2.1　全基因组预测目标基因自动化搜索系统的建立

我们建立了全基因组预测目标基因自动化搜索系统。该系统由 SearchResult.pl 等主要程序和若干辅助程序组成, 利用其进行全基因组预测的流程如下: 第一步, 从 NCBI 或者其他数据库中下载 Fasta 格式的蛋白数据, 该数据可以是全基因数据, 或者是某条染色体上的数据, 也可以是具体实验中得到的蛋白数据; 第二步, 将得到的大量数据用 filter.pl 程序过滤掉冗余序列, 即序列完全相同, 命名却不同的序列。因为我们的预测是基于蛋白序列, 相同的序列得出的结构域结果是一样的, 所以该步骤能大大缩短我们的预测时间。序列相同名字不同的冗余序列被合并为一条序列, 其新的命名包含上述所有冗余序列的名字; 第三步, 将去掉重复序列的数据库用 ExtractCDS_Fasta.pl 程序格式化成下一步搜索所需的格式; 第四步, 设定好序列应该含有的结构域名称, 可以是含有一个或者多个结构域, 以及该蛋白家族名称, 然后运行 SearchResult.pl 程序, 进行全自动搜索, 程序自动将得到的结果按先前设定好的蛋白家族名称分别保存; 第五步, 将所得结果转化为 Fasta 格式。

以上就是该预测系统的基本流程, 通过改变关键词, 我们便能很方便的预测出具有某种或某些结构域的蛋白序列。

### 2.2　水稻、拟南芥、红藻、蓝细菌以及古细菌的全基因组 PPR 和 TPR 家族预测结果

我们从 NCBI 蛋白质数据库中调出水稻粳稻日

**表 1　预测程序的关键词设置**

**Table 1　Key words settings in prediction program**

| PPR | TPR |
|---|---|
| if(($feedback =~ /name=PPR/)){<br>print PPR "$temp\n";<br>print "$num--$CDS[0]--$CDS[1]\tMATCHED\n";<br>} | if(($feedback =~ /name=TPR/)){<br>print TPR "$temp\n";<br>print "$num--$CDS[0]--$CDS[1]\tMATCHED\n";<br>} |

本晴、拟南芥、红藻、蓝细菌、古细菌的全基因组蛋白序列, 从 BGI 网站上下载水稻籼稻 9311 的全基因组蛋白序列, 从 TIGR 和 RiceGAAS 网站上下载水稻粳稻日本晴的全基因组蛋白序列作为补充, 从 Cyanidioschyzon merolae Genome Project 网站上(http://merolae.biol.s.u-tokyo.ac.jp/)下载红藻的全基因组蛋白序列作为补充, 格式化后得到所有等待计算结构域的全基因组蛋白。几种模式生物的待计算的蛋白数分别为: 水稻粳稻(*Oryza sativa* ssp. *Japonica*) 105 303 个, 水稻籼稻(*Oryza sativa* L. ssp. *indica*) 48 832 个, 拟南芥(*Arabidopsis thaliana*) 123 227 个, 红藻(*Cyanidioschyzon merolae*)5 632 个, 蓝细菌(*Synechococcus* sp. WH 8102)5 341 个, 古细菌(嗜热性古细菌 *Thermophilic archaebacterium*)4 598 个。

运用该方法, 经过逐一计算每个蛋白的结构域, 设定搜索关键词为 PPR 或 TPR, 我们将符合条件的蛋白保存下来, 最后得到的全基因组 PPR 和 TPR 蛋白预测结果见表 2。

### 2.3 PPR 和 TPR 蛋白家族分类

为指导和方便日后的研究, 我们用得到的 TPR 蛋白序列按物种分别进行多重序列比对, 通过构建系统发生树, 将预测的 TPR 蛋白按物种进行了家族分类(图 1)。如图 1 所示, 各物种的 TPR 蛋白在进化上拥有很高的一致性。按照系统发生的关系, 我们将这些 TPR 蛋白分为了四大亚家族, 在图中用弧线框出。每个亚家族中, 又做了进一步细分, 在图中用不同背景色表明。从古细菌到蓝细菌, 从蓝细菌到红藻, 再到高等植物, 我们可以直观的看到 TPR 蛋白家族进化的方向和趋势。各大 TPR 亚家族的成员数目, 从古细菌的单个基因逐渐增多, 通过基因复制进化出越来越多功能的旁系同源基因。其中, Ⅰ型和Ⅱ型 TPR 蛋白在进化上相对保守, Ⅲ型和Ⅱ型 TPR 蛋白随着物种的从低到高, 进化出了更多的成员, 表明Ⅲ型和Ⅳ型 TPR 蛋白行使着更加复杂的作用。这为将来进一步蛋白功能研究的设计提供了有利的启示。

用同样的方法, 我们对预测到的各物种 PPR 蛋白也进行了亚家族分类, 与 TPR 蛋白的亚家族分类图一样, 各物种的 PPR 亚家族在进化上都有较高的一致性。由于进化树的构建是以序列为依据, 所以这从侧面上也有力的证明了我们预测结果的准确性。相同亚家族在图中用相同背景色表明, 详细结果见图 2。

### 2.4 全基因组的目标基因预测方法的比较分析

利用上文所述的比对程序 compare.pl, 我们将拟南芥全基因组 PPR 蛋白的分析结果与文献[13]的结果进行分析比较, 得到下列的结果(图 3)。

运用我们设定的方法预测到拟南芥全基因组的 PPR蛋白数为 735 个, Lurin C等[3]预测到的为 441 个。经过比较, 两种方法共同预测到的完全一样的序列的蛋白个数为 364 个, 占他们预测总数的 82.5%, 占我们预测总数的 49.5%。我们预测到的而他们没有预测到的PPR蛋白数为 371 个, 他们预测到的而我们没有预测到的PPR蛋白数为 77 个。

从这 77 个 PPR 蛋白中随机抽出几个与我们预测到的 735 个蛋白序列进行比对, 发现这几个序列都和我们预测的 735 个蛋白相对应的序列 100%比对, 也就是说 77 个 PPR 蛋白中的这几个序列皆为我们预测的序列的一部分片段序列。这可能是他们在预测时所做的 EST 电子拼接延伸不完全所致。为了验证我们的猜想, 我们建立了本地化的 Blast, 将他们预测的这 77 个序列和我们预测的 735 个序列放在一起用 blastall.exe 软件进行比对, 设定 $E$<1e-200, 详细结果见附录。

由比对结果我们可以看到, 他们预测的这 77 个序列, 有 72 个的比对结果$E$值为 0, 也就是说这些序列的93.5%均为我们预测的序列的一部分片段, 加上我们共同预测到的 364 个序列, Lurin C等[3]预测到的序列有 98.9%都包含在我们预测到的序列中, 这说明我们的预测结果更准确和全面。

**表 2 全基因组 PPR 和 TPR 蛋白预测结果**

**Table 2 Genome-wide prediction results of PPR and TPR proteins**

| | 粳稻 *Oryza sativa* ssp. *japonica* | 籼稻 *Oryza sativa* L. ssp. *indica* | 拟南芥 *Arabidopsis thaliana* | 红藻 *Cyanidioschyzon merolae* | 蓝细菌 *Synechococcus* sp. WH 8102 | 古细菌 *Thermophilic archaebacterium* |
|---|---|---|---|---|---|---|
| PPR | 536 | 519 | 735 | 6 | 0 | 0 |
| TPR | 199 | 177 | 292 | 32 | 10 | 4 |

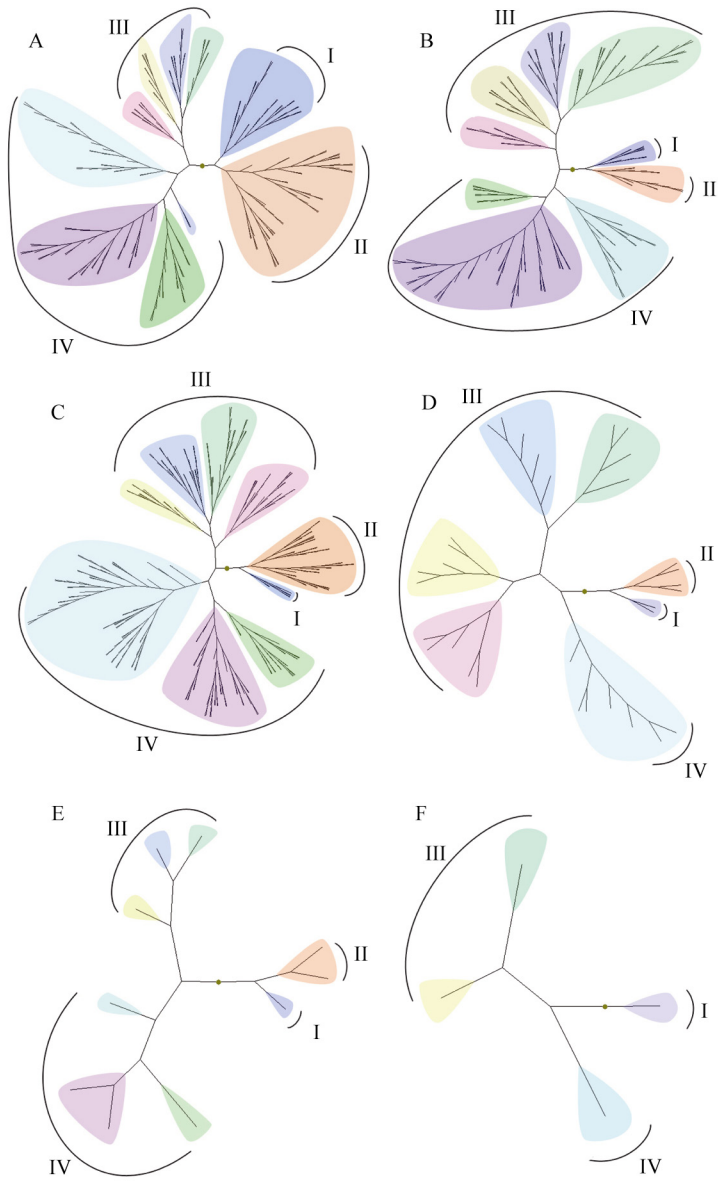


**图 1　TPR 蛋白亚家族分类**

A: 粳稻日本晴; B: 籼稻 9311; C: 拟南芥; D: 红藻; E: 蓝细菌。

**Fig. 1　The classification of TPR proteins' subfamilies**

A: *Oryza sativa* ssp. *japonica* ; B: *Oryza sativa* L. ssp. *indica*; C: *Arabidopsis thaliana*; D: *Cyanidioschyzon merolae*; E: *Synechococcus* sp. WH 8102; F: *Thermophilic archaebacterium*.

## 3　讨　论

为了便于跨平台运行，我们使用 Perl 语言进行了程序设计，该预测系统支持用户使用各种操作系统对数据库进行访问，目前已在 Windows/DOS、Linux 等平台中调试通过，且都能很好地显示结果。通过 Perl 编程，充分利用了 Pfam 数据库资源。Pfam 数据库可用于识别未知蛋白序列所包含的结构域，它不同于标准的序列比对，即使相似性较低也能识别。对于蛋白家族的预测和识别，其模型的准确度和复杂度也比Blast要高的多 [12]。本方法充分的利用了互联网资源，避免了繁琐的数据库设置和大量的数据计算过程，大大缩短了预测所需的时间，具有高通量、准确度高等特点。该预测系统可用于全基因组基因

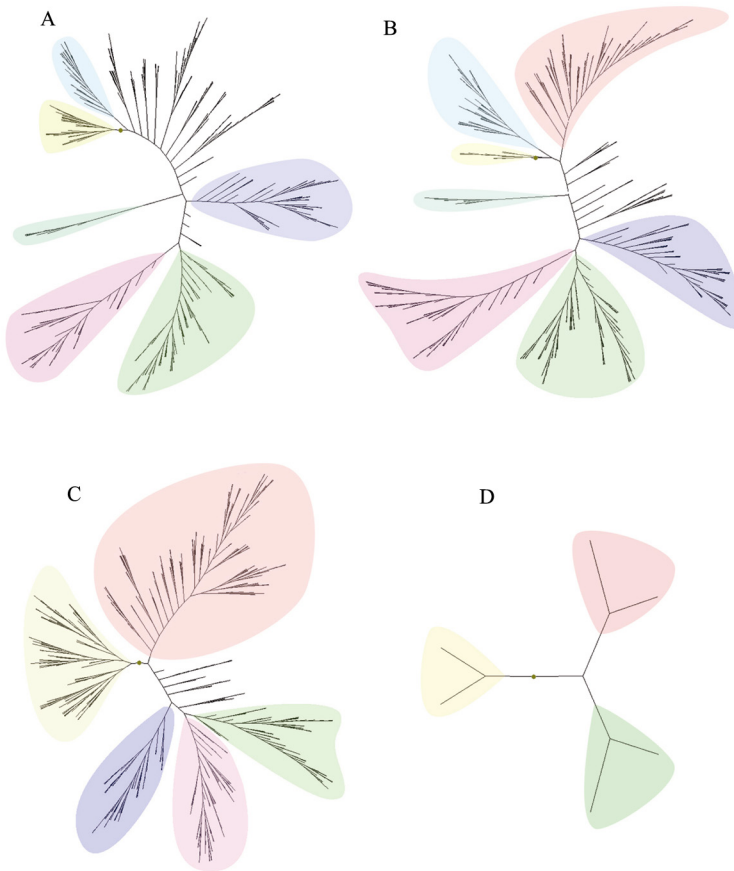


**图 2 PPR 蛋白亚家族分类**

A: 粳稻日本晴; B: 籼稻9311; C: 拟南芥; D: 红藻。

**Fig. 2 The classification of PPR proteins' subfamilies**

A: *Oryza sativa* ssp. *japonica*; B: *Oryza sativa* L. ssp. *indica*; C: *Arabidopsis thaliana*; D: *Cyanidioschyzon merolae*.


**图 3 PPR 蛋白预测结果比较**

**Fig. 3 The comparison of some PPR Proteins' prediction results**

预测及蛋白家族分类, 也可用于实验数据分析等特定的大量蛋白序列的功能搜寻及预测。与 HMMER 软件包中的结构域搜索程序 Hmmsearch 相比 (http://hmmer.wustl.edu/), 本方法能够更有效的进行蛋白多结构域预测。Hmmsearch 需要用到的 HMM Profile 对于那些未过多涉及生物信息学的研究者来说获取不易, 特别是当蛋白家族含有多个特定结构域, 甚至这些结构域间还穿插有其他非特定结构域时, 则得到唯一而准确的HMM profile更加困难 [22]。

虽然本系统是基于 HMM 的结构域预测, 但通过 Perl 编程, 我们有效而巧妙的解决了上述问题, 该方法适用于任意多个结构域的各种排列组合, 即使设定的特定结构域间穿插有其他非特定结构域, 也不会影响预测的准确度, 研究人员只需普通的文本操作, 双击设定好关键词的程序, 便能自动得到结果, 为全基因组新基因的预测和发现提供了方便而有效的新思路。由于国外各大数据库的基因预测程序已相对完善, 我们没有采取自己进行基因预测或者 EST 电子拼接的方法, 而是直接下载各大数据库预测和实验的蛋白数据结果, 进行结构域预测和家族分类。与Lurin C等 [3]预测的结果比较中可以看到, 该方法在一定程度上也大大提高了预测的准确度。另外, 我们预测方法的基础是利用各大数据库的蛋白数据库, 我们预测方法的准确性在很大程度上与那些数据库基因预测和发现软件的准确度及数据库的序列质量

紧密相关。随着生物信息学的不断发展，这些数据库的数据资源和整合数据库资源的技术平台将越来越完善，大规模预测基因也必将更加准确。

利用我们的全基因组预测结果，可以进行更深入细致的研究。蛋白家族分类是其中最基础的一步。我们还可以对每个亚家族进行更细致的分类和研究，并将预测到的全部序列进行染色体定位，研究基因在染色体上的分布情况，以及亚细胞定位预测等。在本文中，通过研究几种模式生物全基因组全部的 TPR 和 PPR 蛋白，我们可以清楚的看到它们间的进化方向，并从整体上掌握各蛋白间的功能联系，为将来的实验研究奠定了基础。近年来的研究表明植物细胞质雄性不育的育性恢复基因都含有 PPR 结构域，那么这些恢复基因是否就存在于我们预测的 PPR 序列中呢？这些都将会为我们的实验研究提供更多的有价值信息。通过对预测结果的观察和分析，我们可以产生很多有趣的推论，比如在低等生物中没有 PPR 蛋白，而 TPR 蛋白却已经出现。TPR 蛋白序列和 PPR 又非常相似，这能否说明 PPR 蛋白是由 TPR 蛋白进化而来？随着物种的不断进化，到了高等植物，全基因组 PPR 蛋白总数却反而是 TPR 蛋白总数的两倍以上，这进化上的变化又是什么原因造成的呢？是否说明 TPR 在进化上更保守，行使更重要的功能？而 PPR 却更多的与环境有关呢？这些问题都值得进一步分析探讨。

随着大量的模式生物基因组完成全基因组测序工作，生物科学已从以往的纯实验科学转向实验分析与生物信息分析相结合的综合科学。随着计算机技术和网络技术的高速发展，各种生物信息学数据的大爆炸，使我们迫切的需要更多更有效的数据挖掘方法。通过对各种生物信息学数据库的整理、总结和深入研究，使我们以全新的视角来利用其他生物学家研究所取得的成就。站在全基因组的高度来分析问题，将有助于更全面、更系统理解复杂的生命现象、生命过程和生命本质。

## 参 考 文 献(References):


[1] Altschul S F, Gish W, Miller W, Myers E W, Lipmann D J. Basic local alignment search tool. *J Mol Biol*, 1990, 215: 403~410.[DOI]

[2] Eisen J A. Phylogenomics: Improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res*, 1998, 8: 163~167.

[3] Pellegrini M, Marcotte E M, Thompson M J, Eisenberg D, Yeates T O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci USA*, 1999, 96: 4285~4288. [DOI]

[4] Eddy S. Profile hidden Markov models. *Bioinformatics*, 1998, 14: 755~763. [DOI]

[5] Jensen L J, Gupta R, Staerfeldt H H, Brunak S. Prediction of human protein function according to gene ontology categories. *Bioinformatics*, 2003, 19: 635~642. [DOI]

[6] King R D, Karwath A, Clare A, Dehaspe L. Accurate prediction of protein functional class from sequence in the Mycobacterium tuberculosis and *Escherichia coli* genomes using data mining. *Yeast*, 2000, 17: 283~293. [DOI]

[7] King R D, Karwath A, Clare A, Dehaspe L. The utility of different representations of protein sequence for predicting functional class. *Bioinformatics*, 2001, 17: 445~454. [DOI]

[8] Wang G, Kong H, Sun Y, Zhang X, Zhang W, Altman N, DePamphilis C W, Ma H. Genome-wide analysis of the cyclin family in Arabidopsis and comparative phylogenetic analysis of plant cyclin-like proteins. *Plant Physiology,* 2004, 135: 1084~1099. [DOI]

[9] Tian C, Wan P, Sun S, Li J, Chen M. Genome-wide analysis of the GRAS gene family in rice and *Arabidopsis*. *Plant Mol Biol*, 2004, 54: 519~532. [DOI]

[10] La H, Li J, Ji Z, Cheng Y, Li X, Jiang S, Venkatesh P N, Ramachandran S. Genome-wide analysis of cyclin family in rice (*Oryza Sativa* L.). *Mol Genet Genomics*, 2006, 275: 374~386. [DOI]

[11] Sonnhammer E L, Eddy S R, Birney E, Bateman A, Durbin R. Pfam: Multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res*, 1998, 26: 320~322. [DOI]

[12] Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy S R, Griffiths-Jones S, Howe K L, Marshall M, Sonnhammer E L. The pfam protein families database. *Nucleic Acids Research,* 2002, 30: 276~280[DOI]

[13] Lurin C, Andres C, Aubourg S, Bellaoui M, Bitton F, Bruyere C, Caboche M, Debast C, Gualberto J, Hoffmann B, Lecharny A, Le Ret M, Martin-Magniette M L, Mireau H, Peeters N, Renou J P, Szurek B, Taconnat L, Small I. Genome-wide analysis of Arabidopsis pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. *The Plant Cell*, 2004, 16: 2089~2103. [DOI]

[14] Bentolila S, Alfonso A A, Hanson M R. A pentatricopeptide repeat-containing gene restores fertility to cytoplasmic male-sterile plants. *Proc Natl Acd Sci USA*, 2002, 99: 10887~10892. [DOI]

[15] Koizuka N, Imai R, Fujimoto H, Hayakawa T, Kimura Y, Kohno-Murase J, Sakai T, Kawasaki S, Imamura J. Genetic characterization of a pentatricopeptide repeat protein gene, orf687, that restores fertility in the cytoplasmic male-sterile Kosena radish. *Plant J*, 2003, 34: 407~415. [DOI]

[16] Kazama T, Toriyama K. A pentatricopeptide repeat-containing gene that promotes the processing of aberrant atp6 RNA of cytoplasmic male-sterile rice. *FEBS Letters*, 2003, 544: 99~102. [DOI]

[17] Akagi H, Nakamura A, Yokozeki-Misono Y, Inagaki A, Takahashi H, Mori K, Fujimura T. Positional cloning of the rice Rf-1 gene, a restorer of BT-type cytoplasmic male sterility that encodes a mitochondria-targeting PPR protein. *Theor Appl Genet*, 2004, 108: 1449~1457. [DOI]

[18] Small I D, Peeters N. The PPR motif——a TPR-related motif prevalent in plant organellar proteins. *Trends Biochem Sci*, 2000, 25: 46~47.

[19] Sikorski R S, Boguski M S, Goebl M, Hieter P. A repeating amino acid motif in CDC23 defines a family of proteins and a new relationship among genes required for mitosis and RNA synthesis. *Cell*, 1990, 60: 307~317. [DOI]

[20] D'Andrea L D, Regan L. TPR proteins: the versatile helix. *Trends Biochem Sci*, 2003, 28: 655~662. [DOI]

[21] Buchner J. Hsp90 & Co.——a holding for folding. *Trends Biochem Sci*, 1999, 24: 136~141. [DOI]

[22] Zhang Z, Wood W I. A profile hidden Markov model for signal peptides generated by HMMER. *Bioinformatics*, 2003, 19: 307~308. [DOI]