# 真核生物启动子预测相关数据库资源概述

刘玉瑛, 张江丽　(1 . 首都师范大学生命科学学院, 北京 100037 ;2 . 廊坊师范学院生命科学学院, 河北廊坊 065000)

**摘要**　启动子是基因表达调控的重要元件, 深入研究启动子的结构和功能, 是理解基因转录调控机制和表达模式的关键。随着生物技术和计算机技术的高速发展, 应用生物信息学技术对启动子进行预测和分析的方法得到了很大发展。对目前常用的真核生物启动子预测相关数据库和软件资源作了简单介绍。

**关键词**　真核生物; 启动子; 数据库; 预测
**中图分类号**　Q24　　**文献标识码**　A　　**文章编号**　0517 - 6611( 2007) 24 - 07418 - 02

**The Databases of Eukaryotic Promoters and Related Software Resources**

**LIU Yu-ying et al**　( College of Life Science , Capital Normal University , Beijing 100037)

**Abstract**　Eukaryotic promoters are important elements in regulation of the expression . To study the structure and function of a promoter deeply , it is the key to know how the gene regulates its transcription and starts its expression . With the fast development of biological and computer technology , significant achievements have been made in computational prediction on Eukaryotic promoters . In this paper mainly introduces the progress made in the databases of predicting Eukaryotic promoters as well as the related software resources was introduced .

**Key words**　Wikipedia ; Promoter ; Database ; Prediction

作为基因表达所必需的重要序列信号和基因转录水平上一种重要的调控元件, 真核生物的启动子一直是现代分子生物学的研究热点。用实验的方法分析和鉴定启动子是多年以来进行启动子研究的主要途径。近年来, 随着人类基因组测序的完成和根据实验获得的对启动子的序列特征与结构功能的认识, 利用生物信息学的方法, 通过计算机模拟和计算来预测基因启动子的相关信息获得越来越多的应用。笔者对目前常用的几个启动子预测数据库和相关软件资源作一简单介绍。

## 1　真核生物启动子的基本结构

真核生物的启动子有3 种类型, 分别由 RNA 聚合酶Ⅰ、Ⅱ和Ⅲ进行转录。典型的真核生物启动子由核心启动子、上游元件和应答元件构成。

核心启动子包括起始子和基本启动子。其中起始子是 DNA 解链并起始转录的位点。基本启动子序列为中心在- 25 ～- 30 的 7bp 保守区, 其碱基频率为: T85 A97 T93 A85 A- 63 A83 A50 , 通常被称为 TATA 框或 Goldberg- Hogness 框, 具有选择正确的起始位点, 保证精确起始的功能。同时, TATA 框还能影响转录速率。如兔的珠蛋白基因中 TATA 框的保守序列 ATAAAA 人工突变为 ATGTAA 时, 转录效率会下降 80 %。

上游元件主要包括 CAAT 框和 GC 框两种, 均具有增强转录活性的功能。其中, CAAT 框的保守序列是 GGCT- CAATCT , 一般位于上游- 75 紧靠- 80 , 与其相互作用的因子有 CTF 家族的成员 CP1 、CP2 和核因子 NF-1 等; GC 框的保守序列是 GTGGGCGGGGCAAT , 常以多拷贝形式存在- 90 处, 识别该序列的转录激活因子为 Sp1 。两种上游元件同时存在或者只存在其中之一, 但并非所有真核基因的启动子都存在上游启动子元件, 有些植物细胞中几乎不存在 CAAT 框。

应答元件通常位于基因上游, 能被转录因子识别和结合, 从而调控基因的专一性表达。如热激应答元件、激素应答元件、cAMP 应答元件、金属应答元件、糖皮质激素应答元件和血清应答元件等。应答元件含有短重复序列, 不同基因中应答元件的拷贝数相近。

## 2　真核生物启动子预测相关数据库资源

**2 .1　EPD( Eukaryotic promoter database)** [1]　EPD 数据库 ( http :/ / www . epd . isb- sib . ch/ 或者 ftp :/ / ftp . epd . isb- sib . ch/ pub/ databases/ epd ) 是一个针对真核 RNA 聚合酶Ⅱ型启动子的非冗余数据库。现有启动子序列数据 1 500 多个, 按层次组织。关于启动子的描述信息直接摘自科学文献。该数据库中所有的启动子均经过一系列实验证实, 如: 是否为真核 RNA 聚合酶Ⅱ型启动子、是否在高等真核生物中有生物学活性、是否与数据库中的其他启动子有同源性等。同时, EPD 与其他的相关数据库如 EMBL 、SWISS- PROT 、TRANSFAC 等, 实现了数据的交叉链接。在其最新版本( 第 76 版) 中, EPD 将收集的启动子分为 6 大类: 植物启动子、线虫启动子、拟南芥启动子、软体动物启动子、棘皮类动物启动子和脊椎动物启动子, 共 2 997 个条目, 其中人类启动子有 1 871 个, 约占总数的 62 %。EPD 数据库是目前唯一一个源自实验数据的真核生物启动子数据库, 是常用的预测软件测评的手段之一。

**2 .2　PLACE( Plant cis-acting regulatory DNA elements)** [2]　PLACE 数据库 ( http :/ / www . dna . affrc . go . jp/ htdocs/ PLACE/ , FTP 服务器为 ftp :/ / ftp . dna . affrc . go . jp/ ) 是从已发表文献中搜集植物顺式作用元件资料而建立的模体数据库( motif database) , 始于 1991 年。目前服务器位于日本农林渔业部。PLACE 数据库中只囊括维管植物的信息, 其他与植物顺式作用元件同源的非植物模体也同时被收录。并且所收录信息根据实验最新进展随时得到更新。同时, PLACE 数据库中还包括了对每个模体的描述和在 PubMed 中的相关文献编号, 以及在 DDBJ/ EMBL/ GenBank 的核酸序列数据库的登录号, 点击后可阅读相关文献摘要等信息。登陆 PLACE 数据库界面, 用户可通过关键词、SRS 关键词或者同源序列查询顺式作用元件的信息。关键词可以是模体名称、涉及的诱导子或者植物激素、胁迫类型、该基因表达的组织或者器官、原始文献的作者、模体序列、植物种属等。查询结果显示位点( 模体) 名称、位置、序列和 PLACE 登录号, 同时, 也可以用 FASTA 格式批量上传序列信息。

作者简介　刘玉瑛( 1982 - ) , 女, 北京人, 硕士研究生, 研究方向: 生物化学与分子生物学。

收稿日期　2007- 04- 23

**2.3　TRRD(Transcription regulatory regions database)**[3]　TRRD 数据库(http://www.bionet.nsc.ru/trrd/),即转录调控区数据库。其数据来源于已发表的科学论文,包含特定基因各种结构与功能特性,包括转录因子结合位点、启动子、增强子、沉默子的位置以及基因表达调控模式等。2001 年的6.0 版本综合了3 898 篇科学文献中的1 167 个基因,5 537 个转录因子结合位点,1 714 个调控区域,14 个座位控制区和5 335 个表达模式。在TRRD 数据库中,所有信息被分列于5 个相关的数据表中:TRRDGENES(包含所有TRRD 库基因的基本信息和调控单元信息);TRRDSITES(包括调控因子结合位点的具体信息);TRRDFACTORS(包括TRRD 中与各个位点结合的调控因子的具体信息);TRRDEXP(包括对基因表达模式的具体描述);TRRDBIB(包括所有注释涉及的参考文献)。TRRD 的主页提供了对这几个数据表的检索服务。除此之外,数据库还提供了另外2 个工具:①序列获得系统(SRS),用于搜索TRRD 和与外部信息和软件资源进行整合;②TRRD Viewer,以基因图谱的形式提供相关信息的描述。

**2.4　TRANSFAC(Transcriptional regulation,from patterns to profiles)**[4]　TRANSFAC 数据库(http://www.gene-regulation.com/ 或者http://transfac.gbf.de/TRANSFAC/)是一个真核基因顺式调控元件和反式作用因子数据库,数据搜集的对象从酵母到人类。TRANSFAC 数据库中的数据资源被分为6 大类别:SITE 类数据是关于真核基因的不同调控位点信息,GENE 类数据描述具有多个调控位点的基因信息,FACTOR 类数据描述结合于这些位点的蛋白质因子信息,CELL 类数据则说明蛋白质因子的细胞来源,CLASS 类数据包含转录因子分类的基本信息,MATRIX 数据以矩阵的形式定量描述结合位点核苷酸的统计分布。此外,还有几个与TRANSFAC 密切相关的扩展库:PATHODB 库收集了转录区域中可能导致病态的突变数据;S/MART DB 收集了蛋白质结合位点的特征信息及作用于这些位点的蛋白质信息;TRANSPATH 库用于描述与转录因子调控相关的信号传递的网络;CYTOMER 库表现了人类转录因子在各个器官、细胞类型、生理系统和发育时期的表达状况。

## 3　前景与展望

对真核生物启动子进行计算机预测和鉴定是一项具有挑战性的研究工作。到目前为止,尽管相关数据库和软件资源得到了很大的丰富和发展,但仍存在着明显不足,如:①大多数数据库对于数据的创新、精确性和准确性没有权威评价,数据过多、重复,分类较粗等;②人类公共数据库中,只有极少数被实验证实的顺式作用元件,绝大多数基因的启动子仍然未知;③采用人类基因组信息来预测植物、真菌等远缘物种的基因结构时,数据准确性不高,但目前针对植物、真菌等的生物信息学数据库远没有人类的全面和完善;④数据库中cDNA 和EST 簇经常是不完整序列,特别是5′端,故无法确定转录起始位点的确切位置,从而影响启动子的预测;⑤真核生物的顺式作用元件比原核生物复杂,需要考虑多种因素[5]。因此高效的实验方法和设计良好的预测软件仍是生物学家面临的严峻课题。

随着分子生物学、遗传学和生物信息学的高速发展,更多的真核生物启动子序列将得到分析,各顺式作用元件的功能也会逐渐明确,启动子的计算机预测研究工作也将有更广阔的发展空间。

## 参考文献

[1] CHRISTOPH D,VIVIANE P.The Eukaryotic promoter database EPD:the impact of in silico primer extension[J].Nucleic Acids Research,2004,32:82-85.

[2] KENICHI H,YOSHIHIRO U.Plant cis-acting regulatory DNA elements (PLACE) database:1999[J].Nucleic Acids Research,1999,27(1):297-300.

[3] KOLCHANOV N A,LGNATIEVA E V.Transcription regulatory regions database (TRRD):its status in 2002[J].Nucleic Acids Research,2002,30(1):312-317.

[4] MATYS V,FRICKE E.TRANSFAC:transcriptional regulation,from patterns to profiles[J].Nucleic Acids Research,2003,31(1):374-378.

[5] TOMPA M,LI N,BAILEY T L,et al.Assessing computational tools for the discovery of transcription factor binding sites[J].Nature Biotech,2006,23:137-144.