遗传 HEREDITAS (Beijing) 2010 年 10 月, 32(10): 1009－1017
ISSN 0253-9772 www.chinagene.cn

综 述

DOI: 10.3724/SP.J.1005.2010.01009

转录因子相关数据库

陈鸿飞, 王进科

东南大学生物电子学国家重点实验室, 生物科学与医学工程学院生物技术与材料实验中心, 南京 210096

摘要: 转录水平的调控是基因调控的重要环节, 其中转录因子(Transcription Factor, TF)和转录因子结合位点(Transcription Factor Binding Site, TFBS)是转录调控的重要组成部分。为了解析基因转录调控过程中 TF 与其 TFBS 相互作用的分子机理, 鉴定 TFBS 及构建基因转录调控网络, 需要对已发现的 TF 及其 TFBS 信息进行系统的收集、整理和分析。目前, 国际上已经出现不少关于 TF 及其 TFBS 的专业数据库, 这些数据库对基因转录调控及 TF 相关的分子生物学、系统生物学及生物信息学的研究非常重要, 对这些领域的研究起到了显著的推进作用。文章对 7 个目前比较著名的 TF 及其 TFBS 相关数据库, 包括 TRANSFAC、JASPAR、TFDB、TRRD、TRED、PAZAR、MAPPER 的特点、数据种类和数量及使用方法进行了详细综述, 并简要介绍了其他相关数据库。

关键词: 转录因子; DNA结合位点; 数据库; 生物信息学

The databases of transcription factors

CHEN Hong-Fei, WANG Jin-Ke

The State Key Laboratory of Bioelectronics, *The Experimental Center of Biotechnology and Biomaterials*, *School of Biological Sciences and Medical Engineering*, *Southeast University*, *Nanjing* 210096, *China*

Abstract: The control of gene transcription is a critical level of gene expression regulation. The interactions between transcription factors (TF) and their DNA binding sites (TFBS) play a key role at this level. In order to decipher the molecular mechanism of the interactions of TFs with TFBSs and construct transcription regulatory network, it is necessary to systematically collect, save, and analyze the information of discovered TFs and their TFBSs. In recent years, multiple TF and TFBS-related databases have been established. These databeses significantly promoted the TF-related studies in the fields of molecular biology, bioinformatics, and system biology. This paper summarized the contents, characteristics, access, and advances of main TFs and TFBSs-related databases, including TRANSFAC, JASPAR, TFDB, TRRD, TRED, PAZAR, MAPPER and others.

Keywords: transcription factor; DNA binding site; database; bioinformatics

真核生物基因的表达受多个层次的调控, 其中基因的转录调控就是一个非常重要的环节。该环节中, 转录因子与其 DNA 结合位点的相互作用发挥关键作用。转录因子包括基础转录因子(Basic TF)和调控性转录因子(Regulatory TF)两类, 其中基础转录因子与 RNA 聚合酶一起构成转录机器(transcription apparatus or machine), 通过与转录起点(Transcription start site, TSS)临近的 DNA 上的启动子区结合实现基因的转录; 而调控性转录因子一般与位置多样的增强子序列结合, 再与转录机器发生作用, 调控基

收稿日期: 2009–12–25; 修回日期: 2010–03–11
基金项目:国家自然科学基金项目(编号：60871014)资助
作者简介:陈鸿飞(1987–), 男, 硕士研究生, 专业方向：生物医学工程。E-mail: chenhf0001@gmail.com
通讯作者:王进科(1969–), 男, 博士, 教授, 博士生导师。研究方向：生物医学工程。Tel: 025-83793620; E-mail: wangjinke@seu.edu.cn

因转录的水平及组织、细胞特异性。增强子序列没有方向性，可位于基因的上游，也可以位于基因的下游，甚至基因内部。目前，已经证实位于基因上游10 kb远的增强子仍然对基因有调控作用。

目前，在人的基因中已经鉴定出 2 000 多个转录因子[1, 2]，NCBI (National Center for Biotechnology Information) 数据库中收录人转录因子基因1 962个，占人全基因总数(24 652)的 8%[3]。已经鉴定的转录因子中约有 700 多个是DNA结合转录因子(DNA-binding TF)。除了转录因子的鉴定外，目前已经发现一个DNA结合转录因子在基因组中存在成千上万的DNA结合位点(DNA-binding sites)。例如，通过染色质免疫沉淀(Chromatin immunoprecipitation, ChIP)结合DNA微阵列芯片(ChIP-chip) 或高通量DNA测序技术(ChIP-seq)，发现转录因子SP1 在基因组中有 12 000 个结合位点[4]，c-myc有 25 000 个结合位点[4]、p53 有 1 600~65 000 个结合位点[5]、CREB 有 19 000~40 000 个结合位点[6, 7]。通过这些位点转录因子控制着众多基因的表达，构成了复杂的基因转录调控网络(Gene transcription regulatory network)。

转录因子及其DNA结合位点的鉴定，以及它们构成的基因转录调控网络的构建已经成为目前系统生物学研究的重点领域，也是生命科学研究的热点之一。这一领域的研究，不仅具有重要的基础研究价值，而且在生物技术及生物医学领域具有重要的应用价值。很多转录因子(如 NF-κB、AP1、p53、PPAR、CREB、STAT、E2F 等)与重要疾病(如炎症、肿瘤等)的发生、发展具有密切的关系，因而成为疾病诊断的依据和药物开发的靶点。

为了系统收集该领域研究产生的大量数据信息并进行相关的生物信息学研究，最近数年国际上涌现出不少转录因子相关数据库，如 TRANSFAC (TRANScription FACtor)、JASPAR、TFdb (The Mouse Transcription Factor Database)、TRRD(Transcription Regulatory Regions Database)、TRED(Transcriptional Regulation Element Database)、PAZAR、MAPPER 等。这些数据库各有特色，提供了转录因子研究不同侧面的数据信息，促进了转录因子的研究，特别是对转录因子相关的生物信息学研究发挥了显著的推动作用。然而目前国内还没有专业的转录因子数据库建立，为了推动这一领域的研究，本文对目前国际上主要的转录因子数据库的相关内容、特点及使用方法予以综述。

1 主要转录因子数据库

1.1 TRANSFAC数据库

TRANSFAC数据库是基于真核生物转录调控所建立的数据库，其中收集了大量与基因转录水平有关的数据，如转录因子及其DNA结合位点和相应的靶基因等信息[8, 9]。TRANSFAC数据库由BIOBASE公司负责日常更新和维护工作，网址是http://www.gene-regulation.com。该数据库分为公开版本和专业版本两个部分，用户只需登陆该网站，按照要求完成相应的注册，利用所获得的账号可以免费查询公开版本中所有的信息，而专业版本则需要用户付费使用，对于国内用户需要付款约 800 欧元进行网上查询，如需下载则需要额外的 800 欧元。目前，公开版版本的版本号为TRANSFAC7.0；专业版本版本号为TRANSFAC2009.3。两个版本的最后更新日期及贮存的数据种类及数据量见表 1。相对于公开版本，专业版本还增加了小RNA(miRNA)及其靶序列、ChIP-chip实验序列片段，以及所有收录数据的相关参考文献、启动子序列等信息。

TRANSFAC数据库的公开版本中主要包括 6 个工作表文件[10]：(1) 位点工作表(Site table)：主要包括每个(推定的)调控蛋白各自的结合位点信息。其中既包括真核生物基因调控中转录因子的结合位点，也包括经诱变实验、体内随机选择所得到的人工序列信息。收录的所有序列经证实都与蛋白结合并且有着特定的功能，每一条序列条目都有相应的唯一序号。(2) 因子工作表(Factor table)：储存相关的转录因子数据信息。在位点工作表中所涉及的转录因子在此表中都有储存。同时还包括一些不与DNA直接结合或者需要与其他转录因子形成复合物才能与DNA结合的转录因子。此外TRANSFAC还对所收集的转录因子根据其DNA结合结构域类型进行分类，方便用户根据需要进行查找。(3) 基因工作表(Gene table)：包括与转录调控相关的基因信息。该工作表最初建立的目的是与其他数据库如TRRD、TRANSCompel的数据相连接；现在已经成为与其

表 **1**　转录因子相关数据库的主要属性

数据库	物种	数据含量	更新时间
TRANSFAC	真核生物	公开版: 6 133 个转录因子 7 915 个转录因子结合位点 2 397 个调控基因 专业版: 12 795 个转录因子 26 589 个转录因子结合位点 51 325 个调节基因	2005 年 2009 年
JASPAR	真核生物	457 个转录因子结合位点模体	2009 年
TFDB	小鼠	1 585 个转录因子	2004 年
TRRD	真核生物	2 344 个调控基因 3 490 个调控区域(启动子, 增强子, 沉默子) 10 135 个转录因子结合位点	2005 年
TRED	人类、 小鼠、 大鼠	1 765 个转录因子结合位点模体 (其中人 1 249 个, 小鼠 366 个, 大鼠 150 个) 4 996 个调控基因 (其中人 3 409 个, 小鼠 1 126 个, 大鼠 461 个) 13 306 个启动子 (其中人 9 085 个, 小鼠 3 089 个, 大鼠 1 132 个)	2007 年
PAZAR	真核生物	859 个转录因子 14 964 个调控序列 5 756 个调控基因	2010 年
MAPPER	真核生物	681 个转录因子 1 134 个转录因子结合位点模体	2010 年
ABS	脊椎动物	650 个实验验证的转录因子结合位点 211 个启动子 100 个调控基因	2006 年
ORegAnno	真核生物	465 个转录因子 3 853 个调控基因	2008 年
DBTSS	真核生物	658 342 502 个转录起始位点	2009 年
AGRIS	拟南芥	1 170 个转录因子 25 516 个启动子 10 653 个转录因子与启动子作用信息	2010 年
DBD	真核生物	700 多个转录因子	2008 年
PlantPromDB	植物	305 个启动子	2002 年
Redfly	果蝇	737 个顺式作用元件 1 342 个转录因子结合位点	2009 年
RegulonDB	大肠杆菌	3 356 个转录因子 1 771 个启动子 1 584 个转录因子结合位点	2010 年

注: 表中数据库除 TRANSFAC 专业版外, 都为免费使用数据库。本表所列数据为截止 2010 年 6 月各数据库贮存的数据量。

他主流数据库如欧洲分子生物学实验室 (The European Molecular Biology Laboratory, EMBL)、美国国立生物信息中心(NCBI)联系的重要组成部分。

(4) 细胞工作表(Cell table): 主要包括了与结合位点相互作用的蛋白的细胞相关信息。利用这些信息可以来确定所涉及的细胞、组织、器官甚至生物体。

(5) 分类工作表(Class table): 主要存放了以不同的DNA结合结构域类型分类的转录因子的家族信息。

(6)矩阵工作表(Matrix table): 利用在Site工作表和Factor工作表中储存的转录因子位点信息, 以及

EMBL 数据库和NCBI提供的参考序列数据库(Reference sequence database, RefSeq)中的基因组序列信息，对转录因子建立了相应的位点特异性权重矩阵，储存在此表中[11]。

登陆TRANSFAC网站，用户可以根据自己的需求(如转录因子名称、结合位点序列)对 6 个主要工作表中的条目进行搜索、查询。同时, BIOBASE公司还提供了与TRANSFAC主数据库相关联的其他数据库，如TRANSPATH、TRANSCompel。TRANSPATH数据库提供了有关转录因子参与信号转导的信息以及它们参与的反应信息，并提供了包含许多信号组件的复杂的信号调控网络信息[12]。TRANSCompel主要是关于真核生物中影响转录的复合调控元件的数据库。复合调控元件由两个不同的转录因子紧密契合的DNA结合位点构成，从而提供了不同信号交叉偶联的机制[13]。

TRANSFAC 作为著名的关于转录因子的数据库，其数据规模十分庞大，收集的信息比较全面。但是其仍存在一些问题，比如数据存在冗余现象，对于不同研究组发现的同一转录因子可能存在不同的条目，且不同物种的同一转录因子也被分开存放，用户在使用时需要注意。同时对于信息更为丰富的TRANSFAC 专业版需要付费，限制了普通用户的使用，特别是普通科研用户。

1.2 JASPAR数据库

JASPAR是收集有关转录因子与DNA结合位点模体(motif)的最全面的公开的数据库，该数据库是由哥本哈根大学(University of Copenhagen)负责日常数据更新维护工作，其网址为http://jaspar.genereg.net/cgi-bin/jaspar_db.pl。JASPAR数据库中所包含的数据，都经过严格筛选，有确切的实验依据，通过计算机辅助软件进行整合识别匹配并用生物学手段进行注释[14, 15]。

JASPAR的最新版本号是JASPAR4.0，相对之前的版本，增加了许多新内容[16]。(1) JASPAR核心数据库(JASPAR_CORE)增加了ChIP-chip 和ChIP-seq相关的信息。ChIP-chip与以前常用的配体DNA系统进化指数富集技术(DNA Systematic Evolution of Ligands by Exponential Enrichment, DNA SELEX)相比，提供的信息量更大，更加准确，已经越来越多地应用于基因研究之中。(2) JASPAR_CORE中增加了 177 个酵母转录因子相关信息，并且增加了关于果蝇和线虫的条目数量，从而使JASPAR基本覆盖了所有的真核顶端生物群(Eukaryote crown group)。JASPAR_CORE中所包含的非冗余的条目从123 个大幅度增加到 457 个。(3) 增加了关于蛋白结合微列阵技术(Protein binding microarray, PBM)相关的 3 个子数据库。PBM子数据库包含 104 个小鼠转录因子信息。PBM_HOMEO子数据库包含 176 个小鼠同源域(Mouse homeodomains)信息。PBM_HLH子数据库包含线虫bHLH转录因子二聚体信息。通过更新，JASPAR目前在核心数据库之外已经拥有 840 个转录因子结合谱[16]。

除核心数据库之外，JASPAR还包含其他几个子数据库[17]：(1) JASPAR_FARM数据库：由具有相似结合特性的转录因子的模型所构成。目前含有 11 个转录因子家族图谱信息。由于多个转录因子有着相似的结合位点图谱，这样存放可以降低结果的复杂性。同时，这些模型也可以为新的数据提供分类依据。(2) JASPAR_phyloFACTs数据库：包含 174 个图谱。这些图谱提取于系统发生上高度保守的基因上游元件。JASPAR_phyloFACTs主要作为JASPAR_CORE的补充，可以和JASPAR_CORE中的数据共同使用。(3) JASPAR_POL Ⅱ：包含 13 个已知的与RNA聚合酶 Ⅱ 核心启动子相关的DNA序列。这些序列与JASPAR_CORE中数据的区别是这些序列不一定有与之作用的特异蛋白。(4) JASPAR_CNE：该数据库由 233 个后生动物基因组中高度保守的非编码DNA元件所组成。这些序列被发现行使长距离增强子作用，参与调控基因表达，调控生物发育和分化。(5) JASPAR_SPLICE：目前仅包括了人类的 6 个拼接位点，以后会增加其他真核生物的拼接位点信息，以及新的外显子拼接的增强子和衰减子信息。

JASPAR 中的数据是完全公开的，用户可以通过主页对数据库进行直接访问。网站在最新一次更新中对 JASPAR_CORE 根据物种分成 5 类，即脊椎动物门(Vertebrata)、线虫纲(Nematoda)、昆虫纲(Insecta)、植物界(Plantae)和真菌界(Fungi)，以及根据结构归类，用户可以清晰地在主页上找到相应链接。网站还提供了根据序列号(ID)、物种等特性进行的搜索，还可以直接浏览数据库的内容。同时，用户

通过主页可以下载 JASPAR 中的数据到自己的电脑上。与同领域相似数据库相比, JASPAR 是一个非冗余的数据库, 数据来源经过严格筛选, 并且对所有数据提供免费下载, 并有相应软件配套使用。但是相对于 TRANSFAC 等其他数据库, JASPAR 所包含的数据量比较小, 用户可以根据需要选择相应的数据库。

1.3　TFdb数据库

TFdb 是一个专业的关于小鼠转录因子的非冗余的数据库。该数据库由 RIKEN 基因组科学中心(RIKEN Genomic Sciences Center, GSC)的基因组探测研究组实验室(Laboratory for Genome Exploration Research Group)进行日常维护, 其网址是 http://genome.gsc.riken.jp/TFdb/。TFdb 包含了小鼠转录因子基因和与之相关联的基因数据。TFdb 的建库宗旨是提供小鼠全基因组的非冗余的转录因子信息。现在的版本共有 1 585 个小鼠转录因子信息。

TFdb是从全基因组的视角来收集数据, 其数据与基因组的联系十分紧密[18]。该数据库的优点是: (1) 数据的收集工作十分严谨, 是非冗余的数据库, 集合了许多有用的信息, 是目前关于小鼠转录因子的最专业最全面的数据库。(2) 该数据库中的每个转录因子都与NCBI的位点链接(LocusLink) (http://www.ncbi.nlm.gove/Locuslink/)相关联。位点链接(LocusLink)是NCBI提供的一个单一的查询界面, 可用来找到某一个遗传位点的序列和描述性信息; 它展现了官方命名、别名、序列登陆、表型、EC号码、MIM号码、UniGene聚类、同源图谱位点和相关网站信息[19], 并提供这个链接可以准确定位的转录因子基因。(3) 该数据库中的每个转录因子都提供了Gene Ontology(GO)的ID序号。Gene Ontology包含基因或蛋白的细胞组分、生物过程及分子功能信息, 并将上述信息根据概念粗细不同组织成有向无环图(DAG)结构。利用GO的条目, 可以明确确定每个基因是否具有转录因子功能。(4) 该数据库还利用NCBI提供的同源基因数据库(HomoloGene)对人类和大、小鼠的基因进行同源比较, 选出功能上具有转录因子效果的基因, 添加进数据库, 确保转录因子信息的完整性。(5) 转录因子DNA结合结构域种类则参考了蛋白质组数据库(InterPro)的资料, 并给出了相关链接。

TFDB 无需用户注册, 使用者可以直接点击页面上方按钮对数据库中所含数据进行浏览, 并可以根据转录因子基因的名称、GO 或 InterPro 数据库的 ID 或名称进行搜索。点击每个条目之前的细节按钮, 用户可以进入单个基因的条目, 网站给出了各个数据库的相关链接, 方便用户查询。

1.4　TRRD数据库

TRRD是转录调控区数据库, 收集基因转录调控区域注释信息资源。该数据库由俄罗斯科学院西伯利亚部的细胞学与遗传学研究所(Institute of Cytology and Genetics, Siberian Department of the Russian Academy of Sciences, Novosibirsk, Russia)提供技术支持及日常维护[20, 21]。其网址为http://wwwmgs.bionet.nsc.ru/mgs/gnw/trrd/。目前最新版本号为TRRD7.0。TRRD数据库包含的数据种类及数据量见表 1。TRRD数据库还包含 7 609 篇相关科学文献。TRRD还有与内分泌调节、脂质代谢以及细胞凋亡相关的转录因子信息。

最新版TRRD数据库主要由 8 个子数据库所构成, 分别是TRRDGENES(TRRD库基因的基本信息和调控单元信息); TRRDLCR(调控区定位信息); TRRDUNITS(调控区的启动子、增强子、沉默子等具体信息); TRRDSTARTS(转录起始位点相关信息); TRRDSITES(转录调控位点信息); TRRDFACTORS(转录因子信息); TRRDEXP(基因表达模式的信息)和TRRDBIB(数据库涉及的实验出版物信息)[22]。

TRRD 网站提供了几个子数据库的链接及搜索按钮。用户不需注册可以直接在网站上浏览其数据库信息。用户可以根据自己需求, 选择浏览或者对特定的条目进行搜索。

1.5　TRED数据库

TRED 为转录调控元件数据库, 是基于研究基因调控网络的需要而建立的数据库, 收集有实验证据的哺乳动物顺式作用元件和反式作用因子。TRED为公开的数据库, 由冷泉港实验室(Cold Spring Harbor Laboratory)承担数据整理及维护工作; 其网址为http://rulai.cshl.edu/cgi-bin/TRED/tred.cgi?process= home。该数据库目前版本收集的转录因子相关数据种类及数量见表 1。该数据库还提供人类、小鼠、大鼠的全基因组序列, 并提供相应的序列分析查询工具[23]。

TRED数据库所提供的数据都是经过实验验证,

并且经过人工一一筛选，保证了数据的有效性。该数据库不但提供转录因子结合位点序列信息，还提供转录因子结合位点的基因组定位信息，为查询者提供了方便。另外，该数据库还引入了基因调控网络的概念，并给出了与癌症相关的 36 个转录因子之间的调控网络，供使用者查询[24]。

TRED 的网站架构简洁明了，数据库不需要注册，数据信息完全公开，在其网站的左边部分，提供了对于数据库各项信息(如启动子序列，转录因子靶基因，转录因子结合位点模体信息等)的链接，非常适合使用者进行查询并使用。

1.6 PAZAR数据库

PAZAR数据库是一个公开的有关转录因子和其调控序列的数据库。PAZAR数据库提供相应的网页界面和工具来方便用户查询；其网址是http://www.pazar.info。该数据库包含的数据种类及数据量见表 1。PAZAR数据库把数据库的结构做成小商店的形式，为用户提供了方便简洁的查询方式[25]。

PAZAR数据库的一大特点是允许每个用户登陆报告自己、本实验室或者是某个数据库的信息。这使得PAZAR数据库更像是一个为转录因子及其结合位点的研究人员所建立的数据交流平台。目前，PAZAR主要包含的数据库有：(1) ABS(a database of Annotated regulatory Binding Sites)：是一个公开数据库，收集从文献中人工筛选获得的直系同源脊椎动物基因启动子中鉴定的已知结合位点[26]；其网址为http://genome.imim.es/datasets/abs2005/index.html。(2) JASPAR_CORE：上文提到的JASPAR的核心数据库。(3) ORegAnno (The Open REGulatory ANNOtation database)：是一个资源及获取都开放的数据库和文献管理系统(literature curation system)，以便基于集体注释实验鉴定的DNA调控区域、转录因子结合位点和调控变体[27]；其网址是http://www.oreganno. org/。PAZAR数据库中还有个人以及相关实验室发布的转录因子和调控基因的信息。

PAZAR数据库还为用户提供了各种计算机辅助软件，使不同学者之间交流信息变得更加容易[28]。数据库允许匿名用户进行数据查询，其网站也提供了免费的注册系统，经过注册后，用户可以向网站提供自己的研究成果，并定制相关服务。网站提供了相关数据的搜索工具，且提供了所有公开数据的免费下载。

1.7 MAPPER数据库

MAPPER数据库收集整理两个著名的转录因子数据库TRANSFAC和JASPAR中有关转录因子的资料，经过加工形成 1 134 个转录因子结合位点的隐马尔可夫模型，并通过这种模型去搜索人、小鼠及果蝇的基因组转录起始位点上游 10 kb序列。数据库还提供用户自定义序列的搜索功能。MAPPER数据库为利用计算机分析手段进行转录因子结合位点的研究提供了平台[29, 30]。MAPPER数据库的网址是http://bio.chip.org/mapper。

MAPPER 数据库需要用户利用自己的邮箱进行免费注册。注册登陆之后，用户即可使用数据库的各种功能。在 Models 板块，用户可以对数据库所包含的结合位点模型进行逐一查询，也可以根据模型的名称和序列来查询自己所需要的结合位点模型。在 Tools 模块中，用户可以对数据库已知的基因序列，或是自己提供的序列，进行结合位点模型的搜索，搜索结果包括了转录因子的名称、结合位点所在的确切位置、相似程度和预测结果。此外，MAPPER 数据库还为每个用户提供了一块私人区域来储存搜索的历史记录和用户自设的模型。

2 其他数据库

除上述主要数据库外，还有转录因子及其结合位点的其他数据库，如 DBTSS (Database of Transcriptional Start Sites)、AGRIS(Arabidopsis Gene Regulatory Information Server)、DBD(DNA-binding domain)、Redfly (Regulatory Element Database for Drosophila)、RegulonDB(Regulon Database)、PlantProm DB(Plant Promoter Sequences Database)等。

DBTSS即转录起始位点数据库[31]，由东京大学人类基因组中心维护，是一个关于转录起始位点的数据库，只收集准确实验确定的 5 '末端的完整cDNA序列，并提供了已知转录因子结合位点的定位；其网址是http://dbtss.hgc.jp。AGRIS即拟南芥基因调控信息数据库，是专业的拟南芥启动子序列、转录因子及其靶基因数据库[32]；网址为http://arabidopsis.med.ohio-state.edu。DBD即DNA结合结构域数据库，

是一个预测与DNA特定序列结合的转录因子的数据库[33]; 其网址是http://www.transcriptionfactor.org。Redfly是一个收集果蝇转录顺式作用元件和转录因子靶序列的数据库[34]; 其网址是http://redfly.ccr.buffalo.edu/。RegulonDB是一个专业的有关大肠杆菌和其他生物转录起始和转录调控的数据库[35], 其网址是http://regulondb.ccg.unam.mx。PlantProm DB是一个有注解非冗余的RNA聚合酶II识别的植物启动子序列数据库[36], 其网址是http://mendel.cs.rhul.ac.uk/。以上数据库包含的数据种类及数量见表1。

3 结 语

将生物信息学技术与分子及细胞生物学技术交叉联合, 即干实验(Dry experiment, *in silico* experiment)与湿实验 (Wet experiment, *in vivo/vitro* experiment)的相辅相成, 是今后生命科学发展的趋势。在转录因子研究领域, 自从转录因子DNA结合靶点高通量信息获取技术, 包括ChIP-chip[37]、ChIP-seq[38~40]、Protein-binding microarray[41~46]等产生后, 转录因子相关研究产生了巨大的数据信息, 对这些数据信息的收集、整理及分析成为转录因子研究领域无法避免的生物信息学课题。因此, 今后转录因子的相关研究必然是干湿实验结合的研究[47, 48]。转录因子数据库的建立无疑会极大地促进转录因子有关的分子生物学、系统生物学及生物信息学研究。

参考文献(References):

[1] Brivanlou AH, Darnell JE Jr. Signal transduction and the control of gene expression. *Science*, 2002, 295(5556): 813–818.[DOI]

[2] Messina DN, Glasscock J, Gish W, Lovett M. An ORFeome-based analysis of human transcription factor genes and the construction of a microarray to interrogate their expression. *Genome Res*, 2004, 14(10B): 2041–2047.[DOI]

[3] Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigó R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X. The sequence of the human genome. *Science*, 2001, 291(5507): 1304–1351.[DOI]

[4] Cawley S, Bekiranov S, Ng HH, Kapranov P, Sekinger EA, Kampa D, Piccolboni A, Sementchenko V, Cheng J, Williams AJ, Wheeler R, Wong B, Drenkow J, Yamanaka M, Patel S, Brubaker S, Tammana H, Helt G, Struhl K, Gingeras TR. Unbiased mapping of transcription factor binding sites along hu-

man chromosome 21 and 22 points to widespread regulation of noncoding RNAs. *Cell*, 2004, 116(4): 499–509.[DOI]

[5] Wei CL, Wu Q, Vega VB, Chiu KP, Ng P, Zhang T, Shahab A, Yong HC, Fu Y, Weng Z, Liu J, Zhao XD, Chew JL, Lee YL, Kuznetsov VA, Sung WK, Miller LD, Lim B, Liu ET, Yu Q, Ng HH, Ruan Y. A global map of p53 transcription-factor binding sites in the human genome. *Cell*, 2006, 124(1): 207–219.[DOI]

[6] Hagiwara M, Brindle P, Harootunian A, Armstrong R, Rivier J, Vale W, Tsien R, Montminy MR. Coupling of hormonal stimulation and transcription via the cyclic AMP-responsive factor CREB is rate limited by nuclear entry of protein kinase A. *Mol Cell Biol*, 1993, 13(8): 4852–4859.

[7] Euskirchen G, Royce TE, Bertone P, Martone R, Rinn JL, Nelson FK, Sayward F, Luscombe NM, Miller P, Gerstein M, Weissman S, Snyder M. CREB binds to multiple loci on human chromosome 22. *Mol Cell Biol*, 2004, 24(9): 3804–3814.[DOI]

[8] Matys V, Fricke E. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res*, 2003, 31(1): 374–378.[DOI]

[9] Matys V, Fricke E, Geffers R, Gössling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Münch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, Wingender E. TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res*, 2000, 28(1): 316–319.[DOI]

[10] Wingender E. The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Brief Bioinform*, 2008, 9(4): 326–332.[DOI]

[11] Fu Y, Weng Z. Improvement of TRANSFAC matrices using multiple docal alignment of transcription factor binding site sequences. *Conf Proc IEEE Eng Med Biol Soc*, 2004, 4: 2856–2859.

[12] Kuang Z, Liu A, Beck TL. TRANSPATH: A computational method for locating ion transit pathways through membrane proteins. *Proteins*, 2008, 71(3): 1349–1359.[DOI]

[13] Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE, Wingender E. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res*, 2006, 34(Database issue): D108–110.[DOI]

[14] Vlieghe D, Sandelin A, De Bleser PJ, Vleminckx K, Wasserman WW, van Roy F, Lenhard B. A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res*, 2006, 34(Database issue): D95–97.[DOI]

[15] Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res*, 2004, 32(Database issue): D91–94.[DOI]

[16] Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, Valen E, Yusuf D, Lenhard B, Wasserman WW, Sandelin A. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res*, 2010, 38(Database issue): D105–110.[DOI]

[17] Bryne JC, Valen E, Tang MH, Marstrand T, Winther O, da Piedade I, Krogh A, Lenhard B, Sandelin A. JASPAR, the open access database of transcription factor-binding profiles-new content and tools in the 2008 update. *Nucleic Acids Res*, 2008, 36(Database issue): D102–106.

[18] Kanamori M, Konno H, Osato N, Kawai J, Hayashizaki Y, Suzuki H. A genome-wide and nonredundant mouse transcription factor database. *Biochem Biophys Res Comm*, 2004, 322(3): 787–793.[DOI]

[19] Pruitt KD, Katz KS, Sicotte H, Maglott DR. Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. *Trends Genet*, 2000, 16(1): 44–47.[DOI]

[20] Heinemeyer T, Wingender E, Reuter I, Hermjakob H, Kel AE, Kel OV, Ignatieva EV, Ananko EA, Podkolodnaya OA, Kolpakov FA, Podkolodny NL, Kolchanov NA. Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL. *Nucleic Acids Res*, 1998, 26(1): 362–367.[DOI]

[21] Stepanenko I, Kolchanov N. Apoptosis gene network: description in the GeneNet and TRRD databases. *Ann N Y Acad Sci*, 2003, 1010 (Apoptosis from Signaling Pathways to Therapeutic Tools): 16–18.[DOI]

[22] Kolchanov NA, Ignatieva EV, Ananko EA, Podkolodnaya OA, Stepanenko IL, Merkulova TI, Pozdnyakov MA, Podkolodny NL, Naumochkin AN, Romashchenko AG. Transcription Regulation Regions Database (TRRD): its status in 2002. *Nucleic Acids Res*, 2002, 30(1): 312–317.[DOI]

[23] Zhao F, Xuan Z, Liu L, Zhang MQ. TRED: a Transcriptional Regulatory Element Database and a platform for *in silico* gene regulation studies. *Nucleic Acids Res*, 2005, 33(Database issue): D103–107.[DOI]

[24] Jiang C, Xuan Z, Zhao F, Zhang MQ. TRED: a transcriptional regulatory element database, new entries and other development. *Nucleic Acids Res*, 2007, 35(Database issue): D137–140.[DOI]

[25] Portales-Casamar E, Kirov S, Lim J, Lithwick S, Swanson MI, Ticoll A, Snoddy J, Wasserman WW. PAZAR: a framework for collection and dissemination of cis-regulation sequence annotation. *Genome Biol*, 2007, 8(10): R207.[DOI]

[26] Blanco E, Farré D, Albà MM, Messeguer X, Guigó R. ABS: a database of Annotated regulatory Binding Sites

from orthologous promoters. *Nucleic Acids Res*, 2006, 34(Database issue): D63–67.[DOI]

[27] Griffith OL, Montgomery SB, Bernier B, Chu B, Kasaian K, Aerts S, Mahony S, Sleumer MC, Bilenky M, Haeussler M, Griffith M, Gallo SM, Giardine B, Hooghe B, Van Loo P, Blanco E, Ticoll A, Lithwick S, Portales-Casamar E, Donaldson IJ, Robertson G, Wadelius C, De Bleser P, Vlieghe D, Halfon MS, Wasserman W, Hardison R, Bergman CM, Jones SJ, Open Regulatory Annotation Consortium. ORegAnno: an open-access community-driven resource for regulatory annotation. *Nucleic Acids Res*, 2008, 36(Database issue): D107–113.[DOI]

[28] Portales-Casamar E, Arenillas D, Lim J, Swanson MI, Jiang S, McCallum A, Kirov S, Wasserman WW. The PAZAR database of gene regulatory information coupled to the ORCA toolkit for the study of regulatory sequences. *Nucleic Acids Res*, 2009, 37(Database issue): D54–60.[DOI]

[29] Marinescu VD, Kohane IS, Riva A. The MAPPER database: a multi-genome catalog of putative transcription factor binding sites. *Nucleic Acids Res*, 2005, 33(Database issue): D91–97.[DOI]

[30] Marinescu VD, Kohane IS, Riva A. MAPPER: a search engine for the computational identification of putative transcription factor binding sites in multiple genomes. *BMC Bioinformatics*, 2005, 6: 79.[DOI]

[31] Wakaguri H, Yamashita R, Suzuki Y, Sugano S, Nakai K. DBTSS: dataBase of human transcription start sites, progress report 2008. *Nucleic Acids Res*, 2008, 36(Database issue): D97–101. [DOI]

[32] Lichtenberg J, Yilmaz A, Welch JD, Kurz K, Liang X, Drews F, Ecker K, Lee SS, Geisler M, Grotewold E, Welch LR. The word landscape of the non-coding segments of the Arabidopsis thaliana genome. *BMC Genomics*, 2009, 10: 463.[DOI]

[33] Wilson D, Charoensawan V, Kummerfeld SK, Teichmann SA. DBD-taxonomically broad transcription factor predictions: new content and functionality. *Nucleic Acids Res*, 2008, 36(Database issue): D88–92.[DOI]

[34] Halfon MS, Gallo SM, Bergman CM. REDfly 2.0: an integrated database of cis-regulatory modules and transcription factor binding sites in Drosophila. *Nucleic Acids Res*, 2008, 36(Database issue): D594–598.[DOI]

[35] Gama-Castro S, Jiménez-Jacinto V, Peralta-Gil M, Santos-Zavaleta A, Peñaloza-Spinola MI, Contreras-Moreira B, Segura-Salazar J, Muñiz-Rascado L, Martínez-Flores I, Salgado H, Bonavides-Martínez C, Abreu-Goodger C, Rodríguez-Penagos C, Miranda-Ríos J, Morett E, Merino E, Huerta AM, Treviño-Quintanilla L, Collado-Vides J. RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res*, 2008, 36(Database issue): D120–124.[DOI]

[36] Shahmuradov IA, Gammerman AJ, Hancock JM, Bramley PM, Solovyev VV. PlantProm: a database of plant promoter sequences. *Nucleic Acids Res*, 2003, 31(1): 114–117.[DOI]

[37] Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL, Wilson CJ, Bell SP, Young RA. Genome-wide location and function of DNA binding proteins. *Science*, 2000, 290(5500): 2306–2309.[DOI]

[38] Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, Thiessen N, Griffith OL, He A, Marra M, Snyder M, Jones S. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods*, 2007, 4(8): 651–657.[DOI]

[39] Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of *in vivo* protein-DNA interactions. *Science*, 2007, 316(5830): 1497–1502.[DOI]

[40] Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. High-resolution profiling of histone methylations in the human genome. *Cell*, 2007, 129(4): 823–837.[DOI]

[41] Bulyk ML, Gentalen E, Lockhart DJ, Church GM. Quantifying DNA-protein interactions by double-stranded DNA arrays. *Nat Biotechnol*, 1999, 17(6): 573–577.[DOI]

[42] Bulyk ML, Huang X, Choo Y, Church GM. Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. *Proc Nalt Acad Sci*, 2001, 98(13): 7158–7163.[DOI]

[43] Wang J, Bai Y, Li T, Lu Z. DNA microarrays with unimolecular hairpin double-stranded DNA probes: Fabrication and exploration of sequence-specific DNA-protein interactions. *J Biochem Biophys Methods*, 2003, 55(3): 215–232. [DOI]

[44] Wang JK, Li TX, Bai F, Lu ZH. Evaluating the binding affinities of NF-kappaB p50 homodimer to the wild-type and single-nucleotide mutant Ig-kappaB sites by the unimolecular dsDNA microarray. *Anal Biochem*, 2003, 316(2): 192–201.[DOI]

[45] Kolchanov NA, Merkulova TI, Ignatieva EV, Ananko EA, Oshchepkov DY, Levitsky VG, Vasiliev GV, Klimova NV, Merkulov VM, Charles Hodgman T. Combined experimental and computational approaches to study the regulatory elements in eukaryotic genes. *Brief Bioinform*, 2007, 8(4): 266–274.[DOI]

[46] Elnitski L, Jin VX, Farnham J, Jones SJM. Locating mammalian transcription factor binding sites: A survey of computational and experimental techniques. *Genome Res*, 2006, 16(12): 1455–1464.[DOI]