

家猪 TBP 蛋白结构与理化性质的 生物信息学分析

胡慧艳¹, 贾青^{1,2*}, 陶隽¹, 魏星灿¹, 陶文欢¹, 墨锋涛¹, 邢增喜¹

(1. 河北农业大学 动物科技学院, 河北 保定 071001; 2. 国家北方山区农业工程技术研究中心, 河北 保定 071001)

摘要: TBP 是真核生物的通用转录因子, 在转录过程中扮演着重要的角色。采用生物信息学方法鉴定了家猪 TBP 蛋白家族成员, 对家猪 TBP 蛋白进行染色体定位、理化性质和系统进化分析, 并对其二级和三级结构进行预测。结果显示, 家猪 TBP 蛋白家族包含 3 个成员, 染色体定位发现其中 2 个 TBP 基因分布在第 1 染色体上, 另 1 个位于骨架上; 3 个 TBP 蛋白家族成员的氨基酸数目分别为 188、273、376, 其中 1 个为疏水性蛋白、2 个为亲水性蛋白; TBP 蛋白二级结构以 α -螺旋和无规则卷曲为主要组成部分, 其家族成员的三级结构基本相似。

关键词: 家猪; TBP; 蛋白家族; 生物信息学分析

中图分类号: S828.9 文献标志码: A 文章编号: 1004-3268(2014)03-0128-05

Bioinformatics Analysis of TBP Protein Structure and Physicochemical Property in *Sus scrofa*

HU Hui-yan¹, JIA Qing^{1,2*}, TAO Jun¹, WEI Xing-can¹, TAO Wen-huan¹,
MO Feng-tao¹, XING Zeng-xi¹

(1. College of Animal Science and Technology, Agricultural University of Hebei, Baoding 071001, China;

2. National Engineering Research Center for Agriculture in Northern Mountainous Areas, Baoding 071001, China)

Abstract: TBP (TATA binding protein) is a general kind of important transcriptional factors in eukaryotes. This article analyzed the phylogenesis of TBP protein family sequences and the localization of TBP genome by using bioinformatics method and then predicted and analyzed their amino acid composition, physicochemical property, as well as secondary and tertiary structures. The predicted results of the TBP protein family of *Sus scrofa* showed that it consisted of three members. Further genetic mapping of TBP genome localization found that 2 TBP genes distributed on chromosome 1 and another located in a scaffold. The number of amino acid and hydrophobic quality of amino acid sequences in TBP protein families presented some differences. However, the predictive results of the secondary structure found that the main composition of 3 amino acid sequences was random coil, and the tertiary structure of the 3 amino acid sequences was similar.

Key words: *Sus scrofa*; TBP; protein family; bioinformatics analysis

真核生物基因表达是一个十分复杂有序的过程。基因的表达在各个层次上都受到精密的调控, 其中转录水平的调控发生在基因表达的初期阶段, 是许多基因表达调控的主要方式之一。TBP (TA-

TA binding protein) 是广泛存在于真核生物体中的一类重要转录因子^[1], 也是古生菌和真核生物中必不可少的转录起始因子。该因子是 RNA pol I 依赖性(rRNA)、RNA pol III 依赖性(mRNA 和 snRNA)

收稿日期: 2013-10-22

作者简介: 胡慧艳(1989-), 女, 河北灵寿人, 在读硕士研究生, 研究方向: 动物遗传育种与繁殖。

E-mail: huhuiyan315@163.com

* 通讯作者: 贾青(1964-), 男, 河北河间人, 教授, 博士, 主要从事动物遗传育种研究。E-mail: jiaqing@hebau.edu.cn

以及 RNA pol III 依赖性(tRNA 和 5S rRNA)基因转录所需要的蛋白^[2]。其能够以高亲和力与许多 RNA pol II 启动子中的 TATA 盒结合,从而有利于转录。由于 TBP 蛋白是所有启动子起始 RNA 合成所需要的蛋白,而且也是 RNA 聚合酶启动子转录所需要的起始复合物的成分,因而称之为通用转录因子^[3-4]。研究^[5-7]表明,TBP 蛋白可以确保生物活体正常的转录水平。

目前,在家猪上关于 TBP 蛋白家族的研究较少。本研究采用生物信息学的方法对家猪 *TBP* 基因家族进行发掘,并对 TBP 蛋白家族成员的理化性质及其结构进行分析和预测,为深入研究家猪 TBP 转录因子的功能提供参考。

1 材料和方法

1.1 *TBP* 基因家族成员搜索

从 Pfam 数据库(<http://pfam.sanger.ac.uk/family>)^[8]中获得 *TBP* 基因家族的隐马尔科夫模型序列谱(HMM 文件),从家猪基因组序列数据库(http://asia.ensembl.org/Sus_scrofa/Info/Index)分别获取全基因组的蛋白序列和 CDS 序列^[9],利用基于隐马尔科夫模型的 HMMER3.0 软件^[10]对 *TBP* 基因的 HMM 文件以及获取的家猪全基因组的蛋白序列进行搜索,去除冗余后得到的序列确定候选蛋白。对鉴定的 TBP 蛋白,按照下列标准进行分类:如果 *TBP* 基因位于不同基因座位时,该基因

被认为是 1 个成员,同一基因座位的多个剪接体被认为是 1 个成员,选择最长剪接体为代表进行后续研究^[11]。

1.2 家猪 TBP 蛋白的生物信息学分析

将得到的家猪体内的 TBP 家族的所有蛋白序列汇总成一个总蛋白序列文件。在 MEME 网站预测 TBP 蛋白保守结构域^[12],利用 ProtParam^[13](<http://expasy.org/tools/protparam.html>)在线计算其氨基酸数目、分子量、理论等电点、正负电位氨基酸数、脂肪族氨基酸指数、蛋白质疏水性等理化性质,应用 SOMPA 软件预测蛋白质的二级结构,并应用 Phyre 软件预测蛋白质的三级结构。

1.3 家猪 TBP 蛋白系统进化分析

下载小鼠 TBP 家族蛋白序列(来源于 http://www.ensembl.org/Mus_musculus/Info/Index)、人 TBP 家族蛋白序列(http://www.ensembl.org/Homo_sapiens/Info/Index)。利用 MEGA5.04 对家猪 TBP 家族蛋白序列与小鼠、人 TBP 家族蛋白序列进行多序列比对,并用邻接算法(Neighbor-Joining)绘制系统进化树^[14]。

2 结果与分析

2.1 TBP 蛋白挖掘结果

根据隐马尔科夫模型分析结果,共得到 3 个家猪 TBP 氨基酸序列(表 1),蛋白质大小分别为 188、273、376 个氨基酸,外显子个数分别为 7、10、8。

表 1 家猪 TBP 蛋白家族成员数据库信息

转录因子	转录因子 ID 号	蛋白 ID 号	外显子数/个	蛋白质大小/aa	基因 ID 号
TBP-201	ENSSSCT00000027701	ENSSSCP00000022721	10	273	ENSSSCG00000022683
TBP-202	ENSSSCT00000029370	ENSSSCP00000020153	10	273	ENSSSCG00000022683
TBPL2-201	ENSSSCT00000005576	ENSSSCP00000005438	8	376	ENSSSCG00000005059
TBPL1-201	ENSSSCT00000030224	ENSSSCP00000026747	7	188	ENSSSCG00000022387

2.2 TBP 蛋白的染色体定位和理化性质分析

3 个家猪 TBP 蛋白的染色体定位、氨基酸组成成分及理化性质分析结果见表 2。由表 2 可以看出,TBPL2-201、TBPL1-201 基因分布在第 1 染色体上,TBP-202 位于骨架上;TBP 蛋白的理论等电点分别为 6.45、9.61、9.74;TBPL1-201 的不稳定系数为 38.44,低于 40,为稳定蛋白,其他 2 个的不稳定系数均高于 40,均为不稳定蛋白,容易解离。TBPL1-201 为疏水性蛋白,TBP-202、TBPL2-201 为亲水性蛋

白,其中 TBPL2-201 的亲水性更强。

2.3 TBP 蛋白保守结构域预测

MEME 系统预测 TBP 蛋白保守结构域结果显示,TBP 蛋白中有 3 个保守结构域。由图 1 可以看出,Motif1 和 Motif2 在这 3 个 *TBP* 基因中都有分布,而 Motif3 只在 2 个 *TBP* 基因中分布,推断 Motif1 和 Motif2 为 TBP 蛋白家族的 2 个保守域。其中 Motif1 包含了 50 个保守氨基酸,Motif2 包含了 41 个保守氨基酸。

表 2 家猪 TBP 蛋白的染色体定位和理化性质

TBP 蛋白	所在染色体/骨架	起始位置	终止位置	理论等电点	正电位点	负电位点	不稳定系数	总平均疏水性	分子量/Da	脂肪系数
TBP-202	894 396.1	4 755	24298	9.74	31	18	43.22	-0.110	30 378.8	90.73
TBPL2-201	1	205 296 096	205 322 382	6.45	39	41	62.26	-0.342	41 695.8	84.79
TBPL1-201	1	33 400 395	33 439 414	9.61	27	17	38.44	0.058	21 110.7	100.59

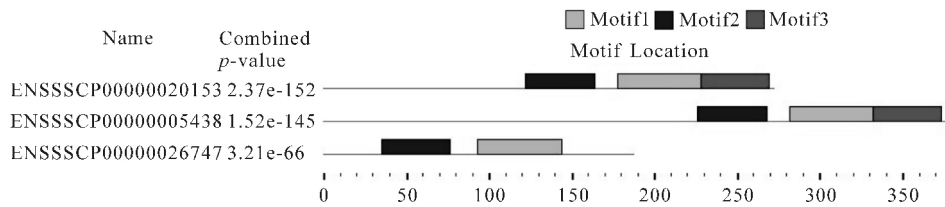


图 1 3 个保守结构域在各个基因上的位置分布

2.4 TBP 蛋白结构分析

TBP 蛋白的二级结构见表 3。由表 3 可见,组成 TBP 蛋白二级结构的有 4 种,即 α -螺旋、延伸链、

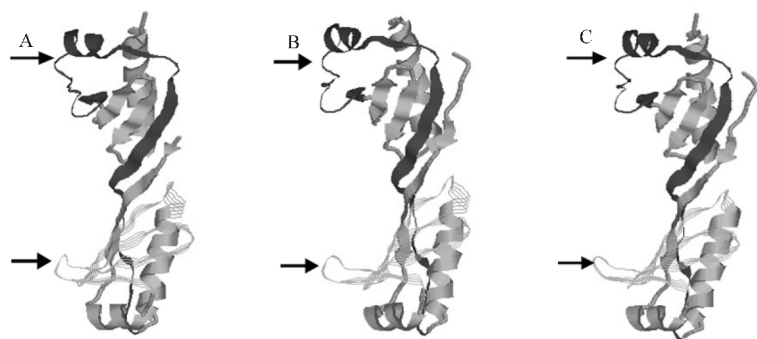
β -转角和无规则卷曲。其中无规则卷曲比例最高,TBP-202 中无规则卷曲占到了 54.58%,而 β -转角所占比例较小。

表 3 3 个 TBP 蛋白的二级结构

TBP 蛋白	(Hh) α -螺旋		(Ee)延伸链		(Tt) β -转角		无规则卷曲	
	残基个数	百分比	残基个数	百分比	残基个数	百分比	残基个数	百分比
TBP-202	61	22.34	45	16.48	18	6.59	149	54.58
TBPL2-201	118	31.38	57	15.16	19	5.05	182	48.40
TBPL1-201	67	35.64	43	22.87	15	7.98	63	33.51

TBP 蛋白的三级结构如图 2 所示。图中标出的丝带模式和网带模式分别是各自的 2 个结构域 Motif1 和 Motif2。在三级结构中 N 端 α -螺旋分布少,而

C 端分布较多,与预测的二级结构中 α -螺旋分布大致相同。而且发现 3 条序列都具有 4 个 α -螺旋,图 A、图 B 中有 18 个转角,图 C 中有 16 个转角。



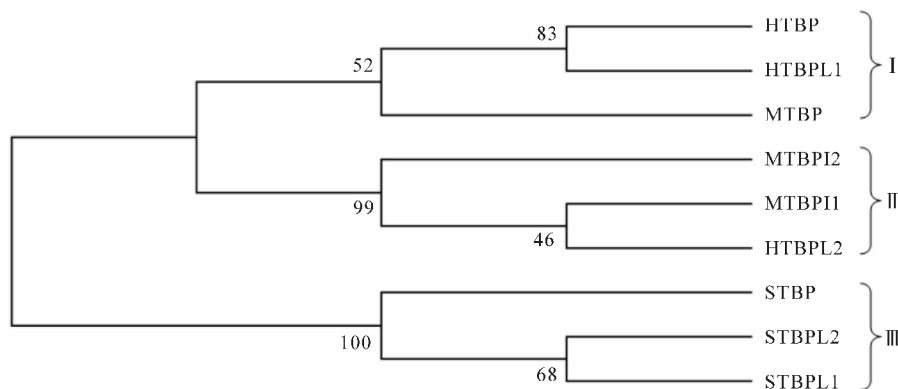
A. ENSSSCP00000020153; B. ENSSSCP00000005438; C. ENSSSCP00000026747

图 2 TBP 蛋白的三级结构

2.5 TBP 蛋白系统进化树

分别对家猪、人以及小鼠的 3 条 TBP 蛋白序列进化关联进行系统进化分析(图 3)。由图 3 可以看出,家猪、小鼠和人 TBP 蛋白序列分成了 3 个亚族,其中第 I 亚族含人的 2 条 TBP 蛋白序列和小鼠的 1 条 TBP

蛋白序列,第 II 亚族含小鼠的 2 条 TBP 蛋白序列和人的 1 条 TBP 蛋白序列,第 III 亚族为家猪 TBP 的 3 条蛋白序列。根据 3 个物种亚族分类情况推断人和小鼠的 TBP 家族基因在进化上具有一定的保守性,而家猪与人及小鼠 TBP 基因家族相似性较低。



STBP、STBPL1、STBPL2 为家猪 TBP 蛋白；HTBP、HTBPL1、HTBPL2 为人 TBP 蛋白；
MTBP、MTBPI1、MTBPI2 为小鼠 TBP 蛋白

图3 家猪、人及小鼠 TBP 蛋白家族的系统进化树

3 讨论

转录因子能够控制基因的表达,是调节各种生理活动的关键环节^[15]。氨基酸的亲疏水性是蛋白质各种物理化学性质中比较重要的性质。而蛋白质亲疏水性氨基酸的组成是蛋白质折叠的主要驱动力,疏水性分布图常用来反映蛋白质的折叠情况,蛋白质折叠会形成疏水内核和亲水表面,同时在潜在跨膜区出现高疏水值区域,据此可以测定跨膜螺旋等二级结构和蛋白质表面氨基酸分布^[16]。维持蛋白质的三级结构最重要的作用力是疏水基的相互作用。疏水作用对蛋白质的稳定性、构象和蛋白质功能具有重要意义。总平均疏水性可以体现蛋白质的亲疏水性质,数值越高代表疏水能力越强,反之则代表亲水性强,总平均疏水性为负值说明此蛋白为亲水性蛋白^[17]。本研究分析了3个TBP蛋白的理论等电点,除TBPL2-201在酸性范围内外,其余2个TBP蛋白均在碱性范围内。TBPL1-201的不稳定系数为38.44,低于40,为稳定蛋白,另外2个TBP蛋白的不稳定系数均高于40,为不稳定蛋白,容易解离。TBPL1-201为疏水性蛋白,另外2个TBP蛋白为亲水性蛋白,其中TBPL2-201的亲水性最强,这可能与TBP蛋白质结构的复杂性以及TBP蛋白质与其他物质间的相互作用的特性有关。

本研究采用基于SOMPA分析方法对TBP蛋白的二级结构进行预测,结果显示,3个TBP蛋白均含有2个结构域,与Bjorn等^[18]报道的一致,进而为三级结构的预测提供指导。蛋白质的功能与其三级结构密切相关。目前,常用的蛋白质三级结构预测的方法有同源建模、穿线法(折叠识别法)、从头预测法^[19]等。其中同源建模对于序列相似度大于

30%的序列模拟比较有效,也是最常用的方法;若序列相似度低于30%,从而会降低预测准确率^[20]。本研究利用同源建模的方法来建立结构模型,采用SWISS-MODEL工具进行预测,对于TBP2未找到Motif2保守域,这与二级结构预测结果不符,因而利用该方法来预测TBP蛋白的三级结构准确率不高。从头预测法是基于分子动力学,寻找能量最低的构象,计算量大,只能做小分子预测^[21]。而Phyre的3d-PSSM的升级版,增加了fold数据,并且采用了新的分析界面,并且在性能上提高10%~15%^[22]。因此,本研究又采用基于折叠模式识别的Phyre预测TBP蛋白质的三级结构,这样可增加预测的准确性。预测结果显示,家猪TBP基因家族的大部分成员间蛋白质三级结构十分相似。

系统进化分析发现,人和小鼠的TBP家族基因在进化上具有一定的保守性,而家猪与人及小鼠TBP基因家族相似性较低。这可能是由于家猪进化过程较快,从而在功能上物种间可能也会有所差别,但这有待进一步研究。本研究对蛋白质的分子进化树分析,可为从分子水平研究物种进化提供手段。其生物学功能及调控机制仍需要通过基因克隆、表达分析以及转基因等方法进行验证。

参考文献:

- [1] Marie K, Benoit C, Jack G, *et al.* Recombinant TBP, transcription factor II B, and RAP30 are sufficient for promoter recognition by mammalian RNA polymerase II[J]. *Biological Chemistry*, 1992, 267(14): 9463-9466.
- [2] Winfried H, Jorn W, Carina H, *et al.* Two transcription factors related with the eucaryal transcription factors TATA-binding protein and transcription factor II B direct promoter recognition by an archaeal RNA polymerase[J].

- Biological Chemistry, 1996, 271(47): 30144-30148.
- [3] Laszlo T. A unified nomenclature for TATA box binding protein (TBP)-associated factors (TAFs) involved in RNA polymerase II transcription [J]. *Genes & Development*, 2002, 16: 673-675.
- [4] Justyna Z, Alice T, Robert G R, *et al.* A novel TBP-TAF complex on RNA polymerase II-transcribed snRNA genes [J]. *Transcription*, 2012, 3(2): 92-104.
- [5] Hidefumi S, Ryo I, Kaori I, *et al.* TATA-binding protein (TBP)-like protein is required for p53-dependent transcriptional activation of upstream promoter of p21Waf1/Cip1 gene [J]. *Biological Chemistry*, 2012, 287(24): 19792-19803.
- [6] Mohamed-Amin C, Dominique K, Igor M, *et al.* Inter-conversion between active and inactive TATA-binding protein transcription complexes in the mouse genome [J]. *Nucleic Acids Research*, 2012, 40(4): 1446-1459.
- [7] Hobbs N K, Bondareva A A, Barnett S, *et al.* Removing the vertebrate-specific TBP Nterminus disrupts placental beta2m-dependent interactions with the maternal immune system [J]. *Cell*, 2002, 110: 43-54.
- [8] Sonnhammer E L, Eddy S R, Durbin R. Pfam: A comprehensive database of protein domain families based on seed alignments [J]. *Proteins*, 1997, 28(3): 405-420.
- [9] Paul F, Ikhlak A, Mridwan A, *et al.* Solution structure of CEH-37 homeodomain of the nematode *Caenorhabditis elegans* [J]. *Nucleic Acids Research*, 2013, 41(4): 48-55.
- [10] Robert D F, Jody C, Sean R E, *et al.* HMMER web server: interactive sequence similarity searching [J]. *Nucleic Acids Research*, 2011, 39(5): 29-37.
- [11] 朱红霞, 胡利宗, 邓小莉, 等. 大豆 *SBP* 基因家族的序列特征、表达及进化分析 [J]. *东北农业大学学报*, 2012, 43(7): 26-33.
- [12] Timothy L B, Mikael B, Fabian A B, *et al.* MEME SUITE: Tools for motif discovery and searching [J]. *Nucleic Acids Research*, 2009, 37: 202-208.
- [13] Elisabeth G, Christine H, Alexandre G, *et al.* The proteomics protocols handbook [M]. Totowa: Humana Press Inc, 2005: 571-608.
- [14] Koichiro T, Daniel P, Nicholas P, *et al.* MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods [J]. *Mol Biol Evol*, 2011, 27(10): 2731-2739.
- [15] 田李, 梁成伟, 杨雨, 等. 真核生物转录调控进化的研究进展 [J]. *生物学杂志*, 2008, 25(3): 1-4.
- [16] 薛庆中. DNA 和蛋白质序列数据分析工具 [M]. 北京: 科学出版社, 2009: 72-75.
- [17] 黄曼, 卞科. 蛋白质疏水性测定方法研究进展 [J]. *粮油食品科技*, 2004, 12(2): 31-32.
- [18] Bjorn B, Benoit H D, Corin Y, *et al.* Evolutionary history of the TBP-domain superfamily [J]. *Nucleic Acids Research*, 2013, 41(5): 2832-2845.
- [19] 张漫, 常延琦. 蛋白质三级结构预测方法简述 [J]. *中国动物检疫*, 2005, 22(5): 36-37.
- [20] Konstantin A, Lorenza B, Jurgen K, *et al.* The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling [J]. *Structural Bioinformatics*, 2006, 22(2): 195-201.
- [21] Gian G T, Andrea C, Michele V, *et al.* Prediction of local structural stabilities of proteins from their amino acid sequences [J]. *Structure*, 2007, 12(7): 139-143.
- [22] Chris S P, Liam J M, David T J, *et al.* Improving sequence-based fold recognition by using 3D model quality assessment [J]. *Structural Bioinformatics*, 2005, 21(17): 3509-3515.