研究报告

树鼩全基因组分泌蛋白的预测分析

罕园园,马开利

(中国医学科学院北京协和医学院医学生物学研究所药物安全性评价研究中心,云南 昆明 650118)

【摘要】 **目的** 综合了解树鼩的分泌蛋白的信息。**方法** 利用真核生物分泌蛋白预测流程 EuSecPred 2.0 对树鼩的分泌蛋白进行全基因组预测,并对获得的分泌蛋白及所携带的信号肽进行系统分析。**结果** 树鼩全基因组蛋白基因中鉴定获得 279 个分泌蛋白,占基因组全部编码基因的 7.2%。对其功能和结构域进行注释分析,发现了以水解酶类、行使蛋白结合与离子结合功能蛋白质及初级代谢过程相关蛋白为主的分泌蛋白。氨基酸组成分析表明,树鼩分泌蛋白及其信号肽主要由疏水性氨基酸组成,信号肽剪切位点有部分亲水性氨基酸的存在。基序分析发现,分泌蛋白信号肽中未存在保守基序,而分泌蛋白内部存在 GxHxCGG[FSV]L[IV][RAS][EP]D[WF]VLTAAHC、[KG]PPGV[YF]T[RK][VI][SC]x[YF][VL][DS]WIQx[TV][MI][RK]、[DT][SA][CF][QK]GDSGGPLVCNGV[LA]QG[IL]V、GY[HL][FL]CGG[SAT]L[ILV]S[EDP][CR]WV[LV][TS]AAHCF、N[IV][FI]FSP[LV]S[IV][SA][TA]ALAMLSLG[AT]xNDTLTQ[IL]L[EQ][GV]LGF[ND]LT[ES]T[SP]E 5 种基序。**结论** 在全基因组范围对树鼩分泌蛋白进行预测分析,有助于进一步研究树鼩机体功能调控机制及动物模型的开发。

【关键词】 树鼩; 基因组;分泌蛋白;信号肽;基序
【中图分类号】R33 【文献标识码】A 【文章编号】1671-7856(2014) 04-0033-05
doi: 10.3969.j.issn.1671.7856. 2014.004.008

A genome-wide prediction and analysis of secreted proteins of *Tupaia Chinensis*

HAN Yuan-yuan, MA Kai-li
(Center for Drug Safety Evaluation and Research, Institute of Medical Biology, Chinese Academy of Medical Sciences & Peking Union Medical College, Kunming 650118, China)

【**Abstract**】 **Objective** To better understand the information of secreted proteins of *Tupaia Chinensis*. **Methods** The internet-based software EuSecPred 2.0 and other mapping software were collectively used to predict and analyze the genome-wide secreted proteins. **Results** There were 279 secreted proteins, accounting for 7.2% of the total proteins encoded by genome sequences. The function and structure of the domain annotation analysis found that they were mainly hydrolases, protein/ion binding proteins and the primary metabolic process-associated proteins. Amino acid composition analysis showed that the signal peptide secreted proteins were mainly composed of hydrophobic amino acids, and there existed some hydrophilic amino acids in the signal peptide cleavage site. Motif analysis revealed no presence of conserved motifs in signal peptide, however, five kinds of motifs existed inside the secreted proteins. **Conclusion** Genome-wide predictive analysis of secreted proteins of tree shrews is helpful for further study of the regulation of body functions in tree

[**基金项目**]国家自然科学基金(81301073);高等学校博士学科点专项科研基金新教师类资助课题(20121106120056);云南省应用基础项目(2013FZ132);协和青年基金—中央高校基本科研业务费专项资金(3332013083)。

[**作者简介**]罕园园(1983 -),女,助理研究员。E-mail: hyy@imbcams.com.cn。
[**通讯作者**]马开利(1982 -),男,博士。Email: mklpumc@gmail.com。

shrews, and for developing mechanism studies and animal models researches.

【Key words】 Tree shrew; *Tupaia chinensis*; Genome; Genome sequences; Secreted proteins; Signal peptide

中缅树鼩(*Tupaia chinensis*)属于灵长目与食虫目之间独立的目—攀鼩目,由于与人类具有高度的同源性,广泛应用于人类病毒性疾病动物模型、细菌感染疾病、内分泌、神经系统疾病和肿瘤方面的研究[1-3],其基因组序列也在近期完成测序,并获得了高覆盖率(79X)的基因组序列,以及转录组鸟枪法组装序列数据库(Transcriptome Shotgun Assembly Sequence Database, TSA)[4],这些数据使得采用生物信息学方法分析树鼩分泌蛋白质组的研究成为可能。

分泌蛋白组指所有的分泌蛋白及蛋白质运输的途径,分泌蛋白在多细胞生物体中决定、控制和协调许多生物学过程,在生物体个体发育、生理功能的发挥及各种病理过程的演进中起着重要作用,分泌蛋白起到的核心作用使它们成为疾病诊断、治疗、药物干预中很好的标志物和靶标[5],研究和鉴定树鼩分泌蛋白组的结构和功能,有助于阐明树鼩免疫、内分泌调控、神经传导、细胞增殖、激素调节等生理活动的机制、阐释其生命现象和推动实验动物模型的开发。

利用基因组序列和可信度较高的生物信息学软件对生物分泌蛋白进行研究显示出强大的优越性,已对多种细菌[6,7]、酵母菌/真菌[8]、微孢子虫[9,10]、草鱼[11]等实验对象进行了分泌蛋白组的预测分析,建立了可信度较高的分泌蛋白预测系统[12-14],得到了许多有益的结果,但迄今树鼩分泌蛋白的分离鉴定主要集中于免疫因子类(如干扰素[15]、IL-2[16]等),尚未见到其它分泌蛋白的研究报道,本研究基于基因组测序数据,通过生物信息学方法对树鼩的分泌蛋白在全基因组范围进行预测,并对分泌蛋白的功能和序列特征进行分析,以期为树鼩免疫、内分泌调控、神经传导、细胞增殖、激素调节等生理活动机制的实验研究提供参考,使实验数据更具目的性和有效性。

1 材料和方法

1.1 材料及分析软件

用于分析的树鼩分泌蛋白的 3895 个来源为 *Tupaia chinensis* 的完整氨基酸序列来源于 uniprot 蛋白数据库(http://www.uniprot.org/)。真核生物分泌蛋白预测流程 EuSecPred 2.0(http://silkpathdb.swu.edu.cn/silkpathdb/eusecpred)。蛋白质 Gene Ontology(GO)注释及绘图程序分别为 InterProScan(http://www.ebi.ac.uk/Tools/pfa/iprscan/) 和 WEGO(http://wego.genomics.org.cn/cgi-bin/wego/index.pl)。氨基酸序列组成分析程序 WebLogo(http://weblogo.threeplusone.com/create.cgi),蛋白质序列基序分析软件 MEME(http://meme.nbcr.net/meme/cgi-bin/meme.cgi)。

1.2 分泌蛋白的预测

以上述树鼩的全部转录组蛋白质序列为基础数据,运行 EuSecPred 2.0 在线流程筛选分泌型蛋白质,该流程通过运行 TMHMM 程序过滤跨膜蛋白,利用 Kohgpi 程序剔除具有 GPI 锚定位点的蛋白质,通过 MitoProt 去除线粒体蛋白,运行 Nucpred、NLStradamus 和 PredictNLS 程序去除定位于细胞核的蛋白质,然后利用 SignalP 4.0 对过滤后的蛋白质序列进行信号肽预测,最后利用 TargetP、PSORT 和 WoLF PSORT 对具有信号肽的蛋白质进行亚细胞定位预测,筛选定位于细胞膜外的蛋白质。以上所有程序的运行及结果的处理通过 EuSecPred 2.0 在线流程完成[8]。筛选得到的分泌蛋白合集包括含有信号肽的经典型分泌(classical secreted protein, CSP)以及无信号肽的非经典型分泌蛋白(non-classical secreted protein, NCSP)两种[17]。

1.3 分泌蛋白的功能注释

针对以上预测结果,利用 InterProScan 以 Nr 和 Swiss-Prot 数据库中的蛋白质序列为参考序列,利用 BLASTP 程序进行比对,将最高得分同源序列的功能作为各分泌蛋白的参考功能,并利用 WEGO 将 InterProScan 获得的注释结果绘图[9]。

1.4 分泌蛋白的序列特征分析

根据信号肽的预测信息,利用 Perl 脚本截取所有分泌蛋白的信号肽序列,并统计信号肽序列的长度,然后通过 WebLogo 程序对所有分泌蛋白序列、信号肽序列及信号肽剪切位点前后 3 个氨基酸的组成进行统计分析,并利用 MEME 程序预测序列中的基序[8]。

2 结果

2.1 树鼩分泌蛋白信号肽的预测

对预测获得的 3 895 个树鼩蛋白质序列进行

EuSecPred 流程预测,发现其中 3 178 个蛋白质具有跨膜结构域,313 个蛋白质具有线粒体定位信号,570 个蛋白质具有细胞核定位信号,405 个蛋白质具有信号肽序列,最终筛选获得了 237 个具有信号肽的分泌蛋白和 42 个无信号肽的分泌蛋白,占已知树鼩蛋白序列的 7.2%。蛋白长度为 100 bp ~ 3896 bp,平均长度 466 bp,分布最多的区域集中在 100 bp ~500 bp,500 bp ~4 000 bp 稍有分布,现已知序列的树鼩分泌蛋白呈现偏态分布(图 1)。

2.2　树鼩分泌蛋白的功能分类

对预测获得的 237 个 CSP 和 42 个 NCSP 进行功能注释,279 个分泌蛋白均在公共蛋白质数据库中检索到明确功能信息的同源蛋白,在分子功能方面数目最多的是水解酶类,占 42.4%,具有蛋白结合功能的蛋白占 32.4%,具有离子结合功能的蛋白占 16.4%,在生物过程方面涉及初级代谢的蛋白最多,占 46.6%(图 2)。

图 1　树鼩信号肽分泌蛋白的 ORF 长度

Fig. 1　The length of secreted proteins with signal peptides of *Tupaia chinensis*

2.3　树鼩分泌蛋白的序列特征

引导分泌蛋白的信号肽长度介于 15 ~37 个氨基酸之间,长度主要集中在 18 aa ~20 aa,平均为 25 aa,中值为 19 aa(图 3)。如图 4 所示,根据 SignalP 3.0 的结果,从树鼩分泌型信号肽的 N 结构域、H 结构域和 C 结构域的变化来看,带正电荷的 N 结构域的长度变化为 2 ~ 21 aa,平均为 6 aa。H 结构域的长度变化为 6 ~ 14 aa,平均为 10 aa。分泌蛋白组成主要为疏水性氨基酸,占全部氨基酸组成的 41.5%,含量最高的是亮氨酸(L),占全部氨基酸组成的 10.2%,亲水性氨基酸占全部氨基酸组成的 34.6%,分泌蛋白氨基酸组成中色氨酸(W)的含量最低(图 4 纯色填充区);分泌蛋白信号肽氨基酸组成主要为疏水性氨基酸,占全部氨基酸组成的 65.7%,其中含量最高的也为亮氨酸(L),占全部氨基酸组成的 26.8%,亲水性氨基酸占 25.4%,碱性、酸性氨基酸的比率低于分泌蛋白中的相应比率,各占 6.8% 和 2.1%(图 4 渐变填充区)。

对树鼩分泌蛋白进行基序分析发现,在信号肽区域未发现有基序的存在,而在非信号肽区域发现有 5 种基序存在,分别为基序 1:GxHxCGG[FSV]L[IV][RAS][EP]D[WF]VLTAAHC、基序 2:[KG]PPGV[YF]T[RK][VI][SC]x[YF][VL][DS]WIQx[TV][MI][RK]、基序 3:[DT][SA][CF][QK]GDSGGPLVCNGV[LA]QG[IL]V、基序 4:GY[HL][FL]CGG[SAT]L[ILV]S[EDP][CR]WV[LV][TS]AAHCF、基序 5:N[IV][FI]FSP[LV]S[IV][SA][TA]ALAMLSLG[AT]xNDTLTQ[IL]L[EQ][GV]LGF[ND]LT[ES]T[SP]E(图 5)。

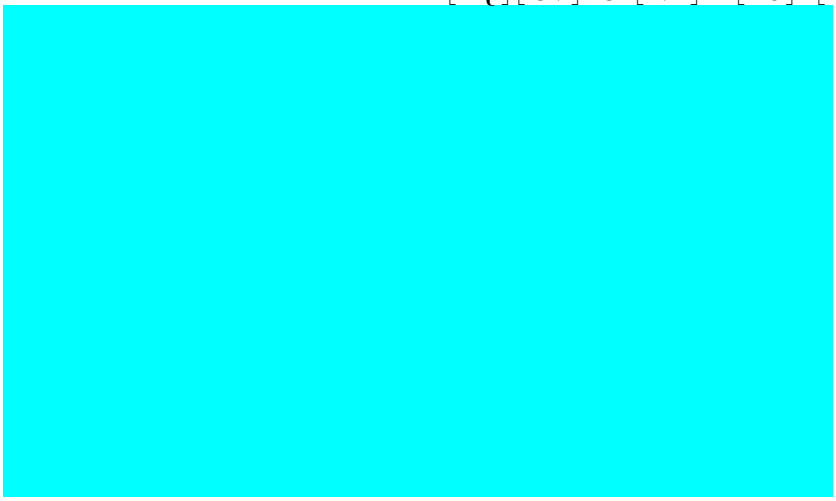

图 2　树鼩的分泌蛋白 GO 注释

Fig. 2　GO of the secretory proteins of *Tupaia chnensis*

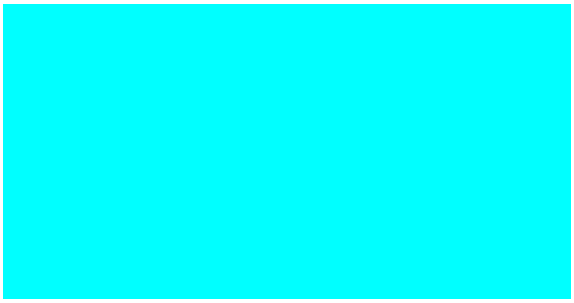

图 3 树鼩分泌蛋白信号肽长度分布

Fig. 3 Distribution of proteins with different lengths of signal peptides

图 4 树鼩分泌蛋白及信号肽序列的氨基酸组成

Fig. 4 Amino acid composition of *Tupaia chinensis* secreted proteins and their signal peptides

对树鼩分泌蛋白的信号肽剪切位点前后 3 个氨基酸进行统计分析发现,与整个信号肽的氨基酸组成稍有不同,剪切位点出现多种亲水性氨基酸及酸性、碱性氨基酸,其中甘氨酸(G)和丝氨酸(S)含量最为丰富;从各个位点来看,其基本序列组成为 Lxx[AV]x[AG],即剪切位点上游 - 3 位较保守,主要为亮氨酸(L);上游 - 2、-1 和 +2 位呈现随机分布状态,-2 位主要为甘氨酸(G)、丙氨酸(A)和脯氨酸(P),-1 位主要为亮氨酸(L)、甘氨酸(G)和丝氨酸(S),+2 位主要为亮氨酸(L)、丝氨酸(S)和谷氨酰胺(Q);+1 位主要为丙氨酸(A)和缬氨酸(V),+3 位主要为丙氨酸(A)和甘氨酸(G)(图 6)。

3 讨论

树鼩作为实验动物的开发还处于起步阶段,其分泌蛋白组的研究还在持续发展当中,除了少量细胞免疫因子类蛋白的分离外,其他分泌蛋白尚无报道。而近年来基于实验数据所建立的生物信息学算法的发展以及树鼩转录组数据的获得则为从基因组水平方面鉴定分泌蛋白提供了可能。本研究基于树鼩基因组数据,在全基因组范围内对分泌蛋白进行预测,获得了 279 个分泌蛋白,为树鼩分泌蛋白的后续实验研究提供了靶标和参考。另一方面,本研究采用的所有预测方法是一套非常严格的流程,保证了预测结果的可靠性,且 EuSecPred 2.0 加入了 SecretomeP 程序,使得非经典途径分泌的蛋白质也可以被预测。

本研究预测获得的 279 个树鼩分泌蛋白中,均能够在公共蛋白质数据库中检索到明确功能信息的同源蛋白,从而获得蛋白的 GO 注释,在分子功能方面数目最多的是分泌性的水解酶类,占 42.4%,具有蛋白结合功能的占 32.4%,具有离子结合功能的占 16.4%,在生物过程方面涉及初级代谢过程的蛋白最多,占 46.6%,这也验证了本研究所采用预测方法的可靠性,更重要的是为解析树鼩与人类的

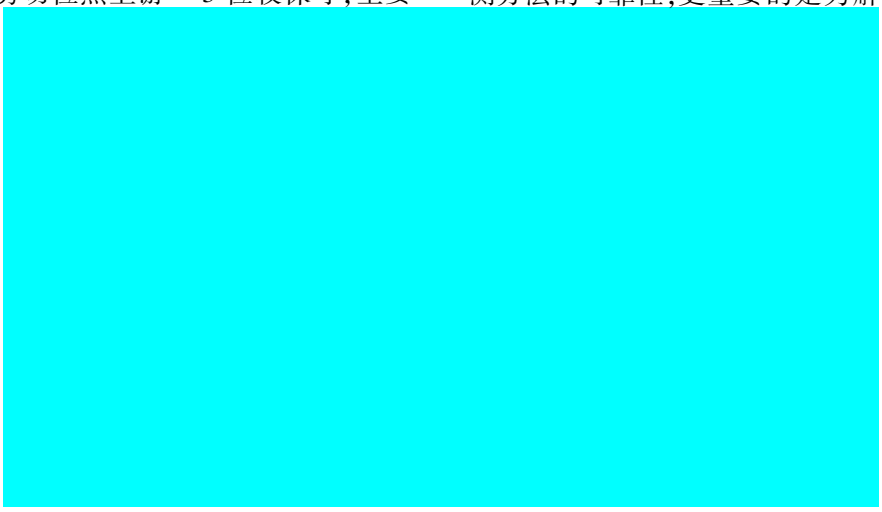

图 5 树鼩分泌蛋白基序

Fig. 5 Motifs in the *Tupaia chnensis* secreted proteins

剪切位点前后氨基酸

图 6　树鼩分泌蛋白信号肽

Fig. 6　The composition of three amino acids flanking splice site of *Tupaia chnensis* secreted proteins

同源性和细胞信息传递机制方面提供了更多的靶标和切入点。树鼩基因组中还有大量的通过预测(Predicted)得到的蛋白序列,由于其蛋白序列未被确定,因此在本研究中未进行预测分析,而这些预测得到的蛋白序列也有可能存在分泌蛋白,这类蛋白的预测分析还有待大量的工作对蛋白数据库进行进一步的完善。而从全基因组范围来看,树鼩的分泌蛋白也主要由疏水性氨基酸组成,是高度保守的,因此信号肽是高度进化的,在决定亚细胞定位上是非常精密的,这可能是分泌蛋白的特有特征。信号肽中亮氨酸的含量最高,表明亮氨酸可能是信号肽的一种关键氨基酸。树鼩分泌蛋白的信号肽区不存在基序,而位于序列内部的另外 5 个基序则可以对分泌蛋白的核定位功能研究提供重要参考。

生物信息学与生物学实验相结合的方法已经越来越广泛的应用于生物学的研究当中,通过多个软件结合对整个基因组的蛋白进行高通量和快速的分析,再用实验的方法加以验证,可以减少大量的实验工作和缩短科研耗时,本研究借助于现有的树鼩蛋白序列信息,对树鼩基因组蛋白进行了挖掘,并主要对经典的含信号肽的分泌蛋白进行了系统分析,对树鼩特异的生物信息通路、分泌蛋白表达谱研究有基础指导作用,树鼩分泌蛋白数据库的构建和完善,可为后续工作提供各种类的分泌蛋白进行针对性的研究,将大大加快分泌蛋白组的进程;同时结合液相色谱/质谱联用产生的大量数据,以及生物学技术 Western blot、免疫组化、Pull down、免疫共沉淀、酵母双杂交及蛋白质芯片技术等的联合,最终能达到由基因到功能的转换和互通[18]。

参考文献:

[1] 王晓娟, 杨春, 苏建家. 树鼩在医学实验研究中的新进展[J]. 中国比较医学杂志, 2010, 20(2):67 - 70.

[2] 徐林, 张云, 梁斌, 等. 实验动物树鼩和人类疾病的树鼩模型研究概述[J]. 动物学研究, 2013, 34 (2):59 - 69.

[3] 黄晓燕, 徐娟, 孙晓梅, 等. 树鼩在人类疾病动物模型中应用研究进展[J]. 实验动物科学, 2013, 30(2):59 - 64.

[4] Fan Y, Huang AY, Cao CC, et al. Genome of the Chinese tree shrew [J]. Nat Commun, 2013, 4:1426.

[5] Guerriero CJ, Brodsky JL. The delicate balance between secreted protein folding and endoplasmic reticulum-associated degradation in human physiology [J]. Physiol Rev, 2012, 92 (2): 537 - 576.

[6] Tjalsma H, Bolhuis A, Jonqbloed JD, et al. Signal peptide-dependent protein transport in Bacillus subtilis: a genome-based survey of the secretome [J]. Microbiol Mol Biol Rev, 2000, 64 (3):515 - 547.

[7] Vizca íno C, Restrepo-Montoya D, Rodríguez D, et al. Computational prediction and experimental assessment of secreted/surface proteins from Mycobacterium tuberculosis H37Rv [J]. Plos Comput Biol, 2010, 6(6):e1000824.

[8] Druzhinian IS, Shelest E, Kubicek CP, et al. Novel traits of Trichoderma predicted through the analysis of its secretome [J]. FEMS Microbiol Lett, 2012, 337(1):1 - 9.

[9] 李田, 刘显林, 韩冰, 等. 家蚕微孢子虫全基因组分泌蛋白的预测分析[J]. 蚕业科学,2013, 39 (2):295 - 301.

[10] 李田, 齐晓冉, 陶美林, 等. 4 种微孢子虫的分泌蛋白的比较基因组学分析[J]. 蚕业科学, 2013, 39 (3):527 - 536.

[11] 孙翰昌, 杨帆, 徐敬明, 等. 草鱼含信号肽分泌蛋白的预测分析[J]. 水产科学, 2011, 30 (3):164 - 167.

[12] Klee EW, Ellis LB. Evaluating eukaryotic secreted protein prediction [J]. BMC Bioimformatics, 2005, 14(6):256 - 263.

[13] Min XJ. Evaluation of computational methods for secreted protein prediction in different eukaryotes [J]. J Proteomics Bioinform, 2012, 3(5):143 - 147.

[14] Cui J, Liu Q, Puett D, et al. Computational prediction of human proteins that can be secreted into the bloodstream [J]. Bioinformatics, 2008, 25(20):2370 - 2375.

[15] 李明利, 田巍威, 高跃东, 等. 树鼩干扰素家族的基本构成及分子特征分析[J]. 动物学研究, 2012, 33 (1):67 - 74.

[16] 黄晓燕, 李明利, 徐娟, 等. 树鼩 IL - 2 全长编码序列的克隆及分子特征分析[J]. 动物学研究, 2013, 34 (2):121 - 126.

[17] Malhotra V. Unconventional protein secretion [J]. EMBO J, 2013, 32(12):1660 - 1664.

[18] 张楠楠, 刘欣, 孙晶, 等. 真核细胞非经典蛋白分泌途径[J]. 遗传, 2009, 31(1):29 - 35.

〔修回日期〕2014-02-18