

张永改,李祥龙,周荣艳,等. 不同物种 *CDK5* 基因编码区生物信息学分析[J]. 江苏农业科学,2014,42(7):32-34.

不同物种 *CDK5* 基因编码区生物信息学分析

张永改, 李祥龙, 周荣艳, 李兰会

(河北农业大学动物科技学院,河北保定 071001)

摘要:采用生物信息学方法比较分析了东非狒狒、黑猩猩、裸鼯鼠、牛、欧洲雪貂、人、灰鼠猴、大猩猩、长臂猿、猪 *CDK5* 基因编码区(CDS)的遗传多样性。结果表明,来自 10 个物种的 47 条基因序列中检测到个 532 多态位点,共生成 30 种单倍型,物种间及物种内 *CDK5* 基因序列编码区存在较丰富的遗传多样性;理论等电点均大于 6,N 端无信号肽,无跨膜结构域,肽链表现为亲水性;蛋白质二级结构主要结构元件是 α -螺旋、无规则卷曲。

关键词:*CDK5* 基因;物种;生物信息学

中图分类号: Q813 **文献标志码:** A **文章编号:** 1002-1302(2014)07-0032-03

CDK5 主要参与神经元正常功能的维持,但在非神经元的组织或细胞中也广泛表达并发挥作用^[1]。研究表明,*CDK5* 可能通过调节黑色素细胞核中 *MITF*mRNA、*TYR* 的表达,从而参与调控羊驼毛色的形成^[2]。在羊驼皮肤组织中,*Agouti* 基因、*MC1R* 基因、酪氨酸酶基因家族、*KIT* 基因、*CDK5* 基因都参与了羊驼毛色的形成与调控^[3]。本研究利用比较基因组学、生物信息学方法研究了该基因编码区种间、种内变异,以期探明该基因在不同种间及种内的遗传分化,进而为相关

黑色素基因的遗传学分析奠定基础。

1 材料与方法

1.1 序列来源

从 NCBI 网站 <http://www.ncbi.nlm.nih.gov/> 的 GenBank 中下载东非狒狒、黑猩猩、裸鼯鼠、牛、欧洲雪貂、人、灰鼠猴、大猩猩、长臂猿、猪等 10 个物种的 47 条 *CDK5* 基因的 CDS 序列(表 1)。

表 1 不同物种的 *CDK5* 基因序列来源

物种	序列数(条)	序列号
东非狒狒	2	XM-003896873.1, XR-161539.1
黑猩猩	2	XR-169918.1, XM-519523.2
裸鼯鼠	2	XM-004860387.21, XM-004877372.1
牛	3	NM-174017.2, L04798.1, BC120083.1
欧洲雪貂	11	XM-004765943.1, XM-004765944.1, XM-004810654.1, XM-004765942.1, XM-004810655.1, XM-004810653.1, XM-004765941.1, XM-004810652.1, XM-004810651.1, XM-004765940.1, XM-004765939.1
人	16	CR-457041.1, NM-004935.3, BT006680.1, BC005115.1, AY049778.1, AY889545.1, X66364.1, AB463550.1, BT020139.1, BT008036.1, AY892008.1, AY892007.1, AY891155.1, AY588641.1, AY893831.1, NG-022432.1
灰鼠猴	2	XM-003929869.1, XR-166425.1
大猩猩	2	XR-175359.1, XM-004046495.1
长臂猿	2	XM-003270919.2, XR-179558.1
猪	5	NM-001044621.2, AK391758.1, AK389504.1, AK349617.1, AY610190.1

收稿日期:2013-10-17

作者简介:张永改(1987—),女,河北定州人,硕士研究生,从事动物遗传学研究。E-mail:zhangyonggai1022@126.com。

通信作者:李祥龙,博士,教授,从事动物遗传学研究。E-mail:lixianglongen@aliyun.com。

1.2 方法

用 BioEdit 软件对 47 个不同序列的 *CDK5* 基因编码区进行比对分析,编辑并选取共有的编码区序列(长度为 879 bp)进行比较,利用 DanSP 4.20 软件对其进行遗传多态性分析,生成单倍型,并计算种间核苷酸歧异度(D_{xy})、遗传分化指数

[18] Bendtsen J D, Nielsen H, von Heijne G, et al. Improved prediction of signal peptides: SignalP 3.0 [J]. Journal of Molecular Biology, 2004, 340(4): 783-795.

[19] Marchler - Bauer A, Anderson J B, Chitsaz F, et al. CDD: specific functional annotation with the Conserved Domain Database [J]. Nucleic Acids Research, 2009, 37: D205-D210.

[20] Letunic I, Doerks T, Bork P. SMART 7: recent updates to the protein domain annotation resource [J]. Nucleic Acids Research, 2012, 40: D302-D305.

[21] Galagan J E, Calvo S E, Borkovich K A, et al. The genome sequence of the filamentous *Neurospora crassa* [J]. Nature, 2003, 422(6934): 859-868.

[22] Jores J L A, Appel B, Lewin A. Cloning and molecular Characterization of a Unique Hemolysin Gene of *Vibrio pommerensis* sp. nov. development of a DNA probe for the detection of the hemolysin gene and its use in identification of related *Vibrio* spp. from the Baltic Sea [J]. FEMS Micro Lett, 2003, 229(2): 223-229.

(G_{st})。再利用 MEGA 4.0 软件的 UPGMA 方法进行种间聚类分析,所利用的在线工具分别是:(1) ProtParam (<http://www.expasy.ch/tools/protparam.html>);(2) SignalP 3.0 Server (<http://www.cbs.dtu.dk/services/SignalP/>);(3) TMHMM 2.0 Server (<http://www.cbs.dtu.dk/services/TMHMM/>);(4) TargetP1.1 Server (<http://www.cbs.dtu.dk/services/TargetP/>);(5) ProtScaleSOPMA (<http://www.expasy.ch/tools/protscale.html>)。

2 结果与分析

2.1 不同物种 *CDK5* 基因核苷酸多样性

2.1.1 多态位点、单倍型及其多样性 在 47 条序列中发现了 532 个多态位点,百分率为 60.52%,其中单一多态位点 114 个,百分率为 12.95%。共发现 30 种单倍型,各物种的多态位点数及单倍型多样性不一致表明 *CDK5* 基因在种群间存在遗传变异(表 2)。平均核苷酸差异数为 83.501,核苷酸多样性为 0.095,其中裸鼯鼠、灰鼠猴的 *CDK5* 基因核苷酸变异较大,说明裸鼯鼠、灰鼠猴的 *CDK5* 基因多态性比较丰富。大猩猩、长臂猿的 *CDK5* 基因核苷酸变异较小。

2.1.2 物种间 *CDK5* 基因的遗传分化 由表 3 可知,不同种群间核苷酸歧异度(D_{st})为 0.037 54~0.162 37,遗传分化在

表 2 不同物种 *CDK5* 基因序列多态信息、单倍型及其多样性

物种	序列(条)	多态位点数(个)	单倍型数(种)	单倍型多样性	核苷酸多样性
东非狒狒	2	77	2	1.000	0.088
黑猩猩	2	63	2	1.000	0.072
裸鼯鼠	2	0	1	0.000	0.000
牛	3	3	3	1.000	0.002
欧洲雪貂	11	82	4	0.763	0.051
人	16	71	7	0.792	0.011
灰鼠猴	2	107	2	1.000	0.122
大猩猩	2	59	2	1.000	0.067
长臂猿	2	61	2	1.000	0.069
猪	5	9	5	1.000	0.004

0.000~0.333,不同物种间核苷酸歧异度、遗传分化变化范围都很大,说明不同物种间遗传分化很明显。其中人和黑猩猩以及人和大猩猩间的核苷酸歧异度最小,表明人与黑猩猩以及大猩猩的亲缘关系较近,其中欧洲雪貂与其他物种间的核苷酸歧异度最大,可见欧洲雪貂与本研究中其他物种间亲缘关系较远。依据不同物种之间的核苷酸歧异度构建的物种间分子聚类图也可看出人、黑猩猩、大猩猩的亲缘关系较近,欧洲雪貂与其他物种的亲缘关系最远(图 1)。

表 3 不同物种核苷酸歧异度、遗传分化

物种	东非狒狒	黑猩猩	裸鼯鼠	牛	欧洲雪貂	人	灰鼠猴	大猩猩	长臂猿	猪
东非狒狒	—	0.000	0.333	0.008	0.098	0.099	0.000	0.000	0.000	0.031
黑猩猩	0.064	—	0.333	0.008	0.098	0.099	0.000	0.000	0.000	0.031
裸鼯鼠	0.115	0.106	—	0.289	0.242	0.203	0.333	0.333	0.333	0.241
牛	0.103	0.098	0.102	—	0.077	0.074	0.008	0.008	0.008	0.008
欧洲雪貂	0.153	0.147	0.140	0.119	—	0.122	0.098	0.098	0.098	0.066
人	0.054	0.037	0.097	0.081	0.129	—	0.099	0.099	0.099	0.059
灰鼠猴	0.097	0.089	0.133	0.120	0.162	0.078	—	0.000	0.000	0.031
大猩猩	0.062	0.040	0.104	0.095	0.146	0.038	0.088	—	0.000	0.031
长臂猿	0.062	0.050	0.103	0.096	0.148	0.043	0.090	0.047	—	0.031
猪	0.109	0.104	0.095	0.067	0.121	0.089	0.125	0.101	0.102	—

注:上三角为遗传分化指数,下三角为核苷酸歧异度。

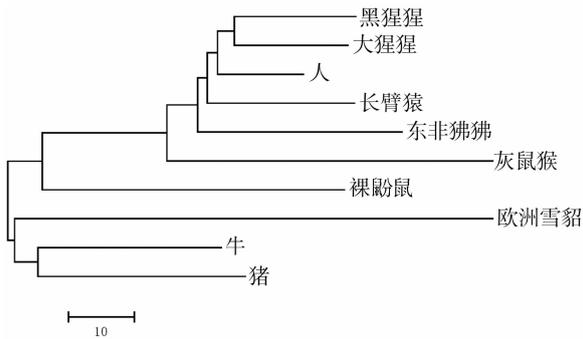


图 1 根据物种间的核苷酸歧异度构建的聚类图

2.2 不同物种 *CDK5* 氨基酸多样性

2.2.1 密码子偏爱性 所截选的不同物种 *CDK5* 基因序列编码区中密码子有效值(ENC)为 48.392 (< 61),偏爱指标(CBI)为 0.424 (> 0),经 χ^2 检验,计算得到未校正的 χ^2 值为 0.454, *CDK5* 基因对密码子有较强的偏爱性^[4]。

2.2.2 同义替换、非同义替换 由表 4 可知,不同物种 47 条

CDK5 基因序列编码区中不同物种同义替换位点数为 207.00~213.44,同义替换核苷酸多样性均值为 0.313 11。非同义替换位点数为 665.56~672.00,非同义替换核苷酸多样性均值为 0.024 73。灰鼠猴的非同义替换位点数较其他物种多,其次是黑猩猩。同义替换不受自然选择的控制,非同义的替代速率远远低于同义替代,且其发生的速率与基因密切相关,这被认为是净化选择的结果^[5]。在有些基因中,非同义替代速

表 4 不同物种 *CDK5* 基因同义替换和非同义替换

物种	同义替换位点数	非同义替换位点数
东非狒狒	209.17	669.83
黑猩猩	208.67	670.33
裸鼯鼠	210.67	668.33
牛	213.44	665.56
欧洲雪貂	212.74	666.26
人	213.08	665.92
灰鼠猴	207.00	672.00
大猩猩	209.17	669.83
长臂猿	211.25	667.75
猪	211.73	667.27

率远远高于同义替代,这是由于正向选择造成的。本研究发现,所选物种的 *CDK5* 基因的非同义替换位点数均明显高于同义替换位点数,说明种群进化过程中可能受到正向选择的影响。

2.3 不同物种氨基酸序列分析和预测

2.3.1 不同物种 *CDK5* 氨基酸序列理化性质分析 用 ProtParam 在线工具分析 10 个物种的 *CDK5* 基因编码的氨基酸序列,结果显示,等电点大于 6,Leu 是所测物种中共有的最主要的氨基酸,含量为 12.3% ~ 15.3%;其次是 Lys,含量为 6.1% ~ 8.1%;其他氨基酸含量波动不大。通过 ProtParam 分析这些物种的 *CDK5* 蛋白,不稳定系数为 28 ~ 32,说明这种蛋白质比较稳定。

2.3.2 导肽的预测和可靠性分析 导肽是一段新合成肽链进入细胞器的识别序列。一般带正电荷的碱性氨基酸含量较为丰富,这些氨基酸对于蛋白质定位非常重要^[6]。利用在线工具 TargetP 1.1 Server 对 10 个物种的 *CDK5* 氨基酸序列进行预测,这些序列没有氨基酸残基分裂位点,所以认为这些 *CDK5* 序列可能不存在导肽酶切位点,不具导肽^[7]。

2.3.3 信号肽的预测、分析 信号肽位于蛋白质的 N 端,用于指导分泌性蛋白到内质网膜上合成。利用在线工具 SignalP 3.0 Server 分析人 *CDK5* 氨基酸序列的信号肽的存在位置及序列,结果显示,人的 *CDK5* 氨基酸序列无信号肽^[7]。对其余几个物种 *CDK5* 氨基酸序列的信号肽的存在位置及其序列进行相应的分析,也得到了统一的结果。

2.3.4 跨膜结构域的预测和分析 跨膜结构域通常是由跨膜蛋白的效应区域所显示。利用在线工具 TMHMM 2.0 Server 预测人 *CDK5* 氨基酸序列的跨膜结构域,结果显示,其不具跨膜结构域,在细胞膜外与细胞膜内都有 *CDK5* 肽链^[8]。对其他物种 *CDK5* 氨基酸序列的跨膜结构域进行分析,得到的预测结果与人的相似。结合上述导肽的预测,可以推测 *CDK5* 在细胞质基质中合成后,很可能也要转运到膜外,该蛋白可以在细胞质与细胞膜外行使其功能。

2.3.5 疏水性/亲水性的预测和分析 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。蛋白质在折叠时会形成疏水内核、亲水表面,同时在潜在跨膜区会出现高疏水值区域,据此可以测定跨膜螺旋等二级结构及蛋白质表面有氨基酸分布^[9]。利用在线工具 ProtScale 预测人 *CDK5* 氨基酸序列的疏水性/亲水性,结果表明,精氨酸(Arg)具有最低分值 -4.500,亲水性最强,异亮氨酸(Ile)具有最高分值 4.500,疏水性最强^[10]。总体来看,亲水区域明显大于疏水区域。因此,整个多肽链表现为亲水性。对其他物种 *CDK5* 氨基酸序列的疏水性/亲水性进行预测,其预测结果与人的相似,因此认为 *CDK5* 蛋白是亲水性蛋白。

2.3.6 二级结构的预测和分析 蛋白质的二级结构是指氨基酸残基形成的 α -螺旋、 β -转角、 β -折叠片延伸链、无规则卷曲。用 SOPMA 对 10 个物种的 *CDK5* 氨基酸序列的二级结构进行预测,结果表明,*CDK5* 蛋白质二级结构的主要结构元件是 α -螺旋(39.73% ~ 42.81%)、无规则卷曲(34.93% ~

38.70%),其次是 β 折叠延伸链(13.46% ~ 14.73%)、 β -转角(6.75% ~ 8.85%)^[11-13]。

3 结论

本研究表明,不同物种间 *CDK5* 基因的核苷酸歧异度、净遗传距离、单倍型间的遗传距离差异都较大,种内及种间遗传分化明显。*CDK5* 基因对密码子有较强的偏爱性,非同义替换位点数均明显高于同义替换位点数,*CDK5* 基因在进化过程中可能受到正向选择的影响。*CDK5* 基因物种间的亲缘关系与动物学分类相符。*CDK5* 蛋白 N 端无信号肽、无导肽、无跨膜结构域,整个多肽链表现为亲水性,蛋白质二级结构的主要元件为无规则卷曲、 α -螺旋,除此之外还有少量 β 折叠、 β 转角。

参考文献:

- [1] 刘 佳,董常生,范瑞文,等. 周期素依赖性蛋白激酶-5 在不同毛色羊驼皮肤中的表达[J]. 畜牧兽医学报,2010,41(4):478-483.
- [2] 张瑞娜,范瑞文,程志学,等. *CDK5* 对羊驼皮肤黑色素细胞 *TYR* 和 *MITF* mRNA 表达的调节[J]. 畜牧兽医学报,2011,42(12):1712-1717.
- [3] 邸文达,杨 川. 羊驼毛色相关基因研究进展的综述[J]. 畜牧与饲料科学,2012,33(8):84-86.
- [4] Smith S D, Kelley P M, Kenyon J B, et al. Tietz syndrome (hypopigmentation/deafness) caused by mutation of *MITF* [J]. Journal of Medical Genetics, 2000, 37(6):446-448.
- [5] 李 易. 基因进化的同义与非同义替代计算及统计检验的比较分析[J]. 曲靖师范学院学报,2006,25(6):1-8.
- [6] 强 毅. 植物蔗糖磷酸合成酶的生物信息学分析[J]. 现代生物医学进展,2007,7(4):557-560,570.
- [7] Emanuelsson O, Nielsen H, Brunak S, et al. Predicting subcellular localization of proteins based on their N-terminus amino acid sequence [J]. Molecular Biology, 2000, 30(4):1005-1016.
- [8] Ikeda M, Arai M, Lao D M, et al. Transmembrane topology prediction methods: a re-assessment and improvement by a consensus method using a dataset of experimentally characterized transmembrane topologies [J]. In Silico Biology, 2002, 2(1):19-33.
- [9] 薛庆中. DNA 和蛋白质序列数据分析工具 [M]. 3 版. 北京:科学出版社,2010.
- [10] Kyte J, Doolittle R F. A simple method for displaying the hydropathic character of a protein [J]. Journal of Molecular Biology, 1982, 157(1):105-132.
- [11] Geourjon C, Deléage G. SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments [J]. Comput Appl Biosci, 1995, 11(6):681-684.
- [12] 薛永常, 裴会忠, 刘长斌. 木质素合成酶 *C3H* 基因的生物信息学分析 [J]. 生物信息学, 2009, 7(1):13-17.
- [13] 范瑞文, 刘 佳, 张俊珍, 等. 黄色羊驼皮肤细胞周期素依赖蛋白激酶 5 (*CDK5*) 完整 CDS 区结构域分析 [J]. 激光生物学报, 2011, 20(1):83-86.