

不同物种 *Hsf2* 基因编码区生物信息分析

马建青, 李祥龙*, 周荣艳, 李兰会, 任玉红

(河北农业大学 动物科技学院, 河北 保定 071001)

摘要: 采用比较基因组学和生物信息学方法, 分析了人、黑猩猩、长臂猿、猕猴、狨、欧洲兔、大熊猫、野猪、牛、非洲象、褐家鼠、小家鼠、荷兰猪和灰仓鼠共 14 个物种的热休克转录因子 2(或称热激因子 2, heat shock transcription factor 2, HSF2) 基因编码区(CDS)的遗传多样性, 并对该基因的氨基酸序列、组成成分等进行预测和推断。结果表明, 在来自 14 个物种的 40 条基因序列中共检测到多态位点数 203 个, 其中有单一多态位点 68 个, 百分率约为 9.81%, 检测到单体型 17 种, *Hsf2* 基因 CDS 在种内表现较为保守, 种间则表现有较丰富的遗传多样性。*Hsf2* 的氨基酸序列表现为亲水性, 理论等电点大多小于 7, 表现为酸性; 肽链的不稳定系数在 51.90~58.13, 表明多肽不稳定。

关键词: 物种; 热休克转录因子 *Hsf2* 基因; 生物信息学

中图分类号: Q786 **文献标志码:** A **文章编号:** 1004-3268(2012)06-0148-04

Bioinformatics Analysis of Coding Regions of *Hsf2*
Gene among Species

MA Jian-qing, LI Xiang-long*, ZHOU Rong-yan, LI Lan-hui, REN Yu-hong

(College of Animal Science and Technology, Agricultural University of Hebei, Baoding 071001 China)

Abstract: The genetic diversity of the CDS of *Hsf2* gene from 14 species, including *Homo sapiens*, *Pan troglodytes*, *Nomascus leucogeny*, *Macaca mulatta*, *Callithrix jacchus*, *Oryctolagus cuniculus*, *Ailuropoda melanoleuca*, *Sus scrofa*, *Bos Taurus*, *Loxodonta Africana*, *Rattus norvegicus*, *Mus musculus*, *Cavia porcellus* and *Cricetulus griseus* were analyzed using the method of comparative genomics and bioinformatics. The characteristics of composition of nucleic acid sequences and amino acid sequences were also analyzed. The results showed that a total of 203 polymorphic sites, including 68 single polymorphic sites, the percentage was about 9.81%, that could be sorted into 17 haplotypes were detected from 40 sequences of 14 species. Comparing with its CDS and the structure of protein conservative within species, it comes to relative diversity in *Hsf2* for nucleotide among species. The amino acid sequences of *Hsf2* presented hydrophilic, theoretical PI was less than 7 mostly. The nature of polypeptide was acid and the instability index of polypeptide was between 51.90 and 58.13, indicating that the polypeptide was not stable.

Key words: species; *Hsf2* gene; bioinformatics

生物体在应对高温和热胁迫等逆境时, 体内会产生一系列的应急反应, 一些相关基因的活性会急剧增加, 其中热激蛋白(heat shock proteins, HSPs)的大量积累尤为显著, 它们以分子伴侣的形式帮助相关蛋白组

装、折叠、胞内运输和降解^[1-2], 使生物体在胁迫环境中的伤害减小到最低。而热激因子(heat shock transcription factor, HSF)是调节这些相关基因包括热激蛋白基因(HSP)活性的信号传导链中的最终成员^[3-5]。目前研

收稿日期: 2012-01-15

基金项目: 国家自然科学基金项目(31172196)

作者简介: 马建青(1988-), 女, 河北省邢台人, 在读硕士研究生, 研究方向: 动物遗传。E-mail: majianqing2006@126.com

* 通讯作者: 李祥龙(1963-), 男, 河北丰南人, 教授, 博士, 博士生导师, 主要从事动物遗传育种研究。

E-mail: lixianglongcn@yahoo.com

究证实,动物细胞中热休克转录因子家族主要包括 4 种不同的类型:HSF1、HSF2、HSF3、HSF4^[6]。其中 HSF2 对热刺激信号耐受,对生物体生长、发育、分化的信号更为敏感,更多是在调控胚胎发育方面发挥作用^[7-8]。

目前国内外对 HSF2 的研究较少,鉴于此,利用生物信息学和比较基因组学的方法研究了不同物种间和物种内 HSF2 的基因编码区的相关特征,旨在探明该基因在所研究物种种内和种间的遗传分化,进而为生物体的生长、发育与分化及机体的热应激

反应机制和动物遗传育种研究提供基础资料。

1 材料和方法

1.1 序列来源

所需基因序列从 NCBI 网站 <http://www.ncbi.nlm.nih.gov/> 的 GenBank 中下载,在本研究中,分别下载了人、黑猩猩、长臂猿、猕猴、狨、欧洲兔、大熊猫、野猪、牛、非洲象、褐家鼠、小家鼠、荷兰猪和灰仓鼠共 14 个不同物种的 40 条 *Hsf2* 基因编码区序列(表 1)。

表 1 不同物种的 *Hsf2* 基因序列来源

| 物种 | 序列数 | 序列号 |
|--------------------------------------|-----|--|
| 人(<i>Homo sapiens</i>) | 16 | AK294624.1, NM_001243094.1, BC064622.1, BC005329.1, AK316377.1, NM_001135564.1, NM_004506.3, EU446897.1, BC128420.1, BC121051.1, BC121050.1, BC112323.1, M65217.1, AB463722.1, DQ492684.1, NG_029607.1 |
| 黑猩猩(<i>Pan troglodytes</i>) | 2 | XM_002817309.1, XM_002817308.1 |
| 长臂猿(<i>Nomascus leucogeny</i>) | 2 | XM_003255640.1, XM_003255639.1 |
| 猕猴(<i>Macaca mulatta</i>) | 1 | XM_001108944.2 |
| 狨(<i>Callithrix jacchus</i>) | 2 | XM_002746919.1, XM_002746918.1 |
| 欧洲兔(<i>Oryctolagus cuniculus</i>) | 2 | XM_002714793.1, XM_002714792.1 |
| 大熊猫(<i>Ailuropoda melanoleuca</i>) | 2 | XM_002919684.1, XM_002919683.1 |
| 野猪(<i>Sus scrofa</i>) | 1 | XM_003121229.3 |
| 牛(<i>Bos Taurus</i>) | 1 | NM_001083405.1 |
| 非洲象(<i>Loxodonta africana</i>) | 2 | XM_003404245.1, XM_003404244.1 |
| 褐家鼠(<i>Rattus norvegicus</i>) | 2 | NM_031694.2, AF172640.1 |
| 小家鼠(<i>Mus musculus</i>) | 4 | NM_008297.3, BC018414.1, X61754.1, AK052270.1 |
| 荷兰猪(<i>Cavia porcellus</i>) | 2 | XM_003479409.1, XM_003479408.1 |
| 灰仓鼠(<i>Cricetulus griseus</i>) | 1 | XM_003504214.1 |

1.2 分析方法

用生物学软件 BioEdit 对已下载的 40 条不同物种 *Hsf2* 的 CDS 序列进行比对分析,选取、编辑共有的编码区序列(长度为 693 bp)进行比较,并分析其氨基酸序列特征和 G+C 含量。之后再利用 DnaSP 5.10 软件对其进行遗传多态性分析,并生成单倍型,在此基础上计算种间核苷酸歧异度(Dxy)与遗传分化系数(Gst)。最后再利用软件 MEGA5.0 的 UPGMA 方法进行种间聚类分析,构建出所研究物种的聚类图。

2 结果与分析

2.1 基因核苷酸序列特征

2.1.1 不同物种 *Hsf2* 多态位点、单倍型及其多样性 在研究的不同物种共 40 条序列中,总共发现有多态位点数 203 个,其百分率约为 29.3%,其中包含 68 个单一多态位点,其百分率约为 9.81%;共发现单倍型 17 种(表 2)。其中人的 *Hsf2* 基因多态位点数与核苷酸多样性的值较其他物种高。

表 2 不同物种 *Hsf2* 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 | 单倍型多样性值 | 核苷酸多样性值 |
|-----|------|---------|--------|---------|---------|
| 人 | 16 | 8 | 3 | 0.425 0 | 0.003 5 |
| 黑猩猩 | 2 | 0 | 1 | 0.000 0 | 0.000 0 |
| 长臂猿 | 2 | 0 | 1 | 0.000 0 | 0.000 0 |
| 荷兰猪 | 2 | 0 | 1 | 0.000 0 | 0.000 0 |
| 小家鼠 | 4 | 1 | 2 | 0.500 0 | 0.000 7 |
| 褐家鼠 | 2 | 0 | 1 | 0.000 0 | 0.000 0 |
| 非洲象 | 2 | 0 | 1 | 0.000 0 | 0.000 0 |
| 大熊猫 | 2 | 0 | 1 | 0.000 0 | 0.000 0 |
| 欧洲兔 | 2 | 0 | 1 | 0.000 0 | 0.000 0 |
| 狨 | 2 | 0 | 1 | 0.000 0 | 0.000 0 |

2.1.2 不同物种 *Hsf2* 基因的遗传分化 所研究的各物种种群间核苷酸歧异度在 0.006 3~0.095 2,遗传分化系数在 0.295 3~1.000 0(表 3),由数据可看出,不同物种间核苷酸歧异度和遗传分化系数的变化范围均较大。依据物种间的核苷酸歧异度构建的分子聚类图可看出,人与黑猩猩、长臂猿的亲缘关系较近,大熊猫和欧洲兔的亲缘关系亦较近,而荷兰猪与所研究物种的亲缘关系最远(图 1)。

表 3 不同物种核苷酸歧异度和遗传分化

| 物种 | 人 | 黑猩猩 | 长臂猿 | 荷兰猪 | 小家鼠 | 褐家鼠 | 非洲象 | 大熊猫 | 欧洲兔 | 獭 |
|-----|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| 人 | — | 0.331 9 | 0.331 9 | 0.331 9 | 0.295 3 | 0.331 9 | 0.331 9 | 0.331 9 | 0.331 9 | 0.331 9 |
| 黑猩猩 | 0.006 3 | — | 1.000 0 | 1.000 0 | 0.505 1 | 1.000 0 | 1.000 0 | 1.000 0 | 1.000 0 | 1.000 0 |
| 长臂猿 | 0.010 6 | 0.010 1 | — | 1.000 0 | 0.505 1 | 1.000 0 | 1.000 0 | 1.000 0 | 1.000 0 | 1.000 0 |
| 荷兰猪 | 0.081 9 | 0.077 9 | 0.083 7 | — | 0.505 1 | 1.000 0 | 1.000 0 | 1.000 0 | 1.000 0 | 1.000 0 |
| 小家鼠 | 0.070 2 | 0.068 2 | 0.071 1 | 0.094 2 | — | 0.505 2 | 0.505 2 | 0.505 2 | 0.505 2 | 0.505 2 |
| 褐家鼠 | 0.066 9 | 0.064 9 | 0.067 8 | 0.092 4 | 0.032 1 | — | 1.000 0 | 1.000 0 | 1.000 0 | 1.000 0 |
| 非洲象 | 0.065 5 | 0.060 6 | 0.063 5 | 0.086 6 | 0.092 7 | 0.095 2 | — | 1.000 0 | 1.000 0 | 1.000 0 |
| 大熊猫 | 0.042 4 | 0.040 4 | 0.046 2 | 0.077 9 | 0.076 8 | 0.076 5 | 0.053 4 | — | 1.000 0 | 1.000 0 |
| 欧洲兔 | 0.041 0 | 0.041 6 | 0.041 9 | 0.079 4 | 0.068 2 | 0.066 4 | 0.064 9 | 0.040 4 | — | 1.000 0 |
| 獭 | 0.023 6 | 0.020 2 | 0.024 5 | 0.088 0 | 0.072 5 | 0.066 4 | 0.059 2 | 0.041 9 | 0.046 2 | — |

注:左三角为核苷酸歧异度;右三角为遗传分化系数。

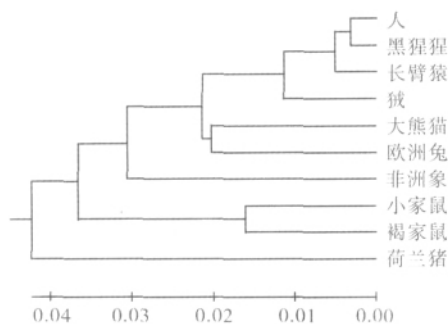


图 1 根据物种间的核苷酸歧异度进行的聚类分析结果

2.1.3 不同物种 *Hsf2* 基因的 G+C 含量 生物分类学上用 G+C 含量占全部碱基的比例来表示各类生物的 DNA 碱基组成特征。亲缘关系相近的生物,它们的核苷酸的碱基组成中应该具有相似的 G+C 含量,若生物之间 G+C 含量差别大,则表明它们的亲缘关系远。本研究表明,14 个物种 *Hsf2*

的碱基 G+C 含量在 37.81%~42.86%(表 4),再次说明 *Hsf2* 在不同物种间发生一定程度的遗传变异,同时也说明了所研究物种的亲缘关系远近。其中,荷兰猪 *Hsf2* 中 G+C 含量为 42.86%,与其他物种 *Hsf2* 基因中 G+C 含量相差最大,说明荷兰猪与所研究的其他物种亲缘关系最远。

2.2 氨基酸序列特征

2.2.1 不同物种 HSF2 氨基酸序列理化性质分析

利用在线软件工具 ProtParam (<http://www.expasy.ch/tools/protparam.html>)分析所研究物种的氨基酸序列理化性质(表 4)。结果表明,来自 14 个物种 *Hsf2* 基因的氨基酸序列亲水性值在 -0.706~-0.482,分子量在 27 026.9~77 203.8 bp,理论等电点在 4.68~9.67,不稳定系数在 51.90~58.13。

表 4 不同物种 *Hsf2* 基因 G+C 含量和氨基酸序列理化性质参数

| 物种 | 亲水性大小 | 氨基酸数量/个 | 分子量/bp | 理论等电点 | 不稳定系数 | G+C 含量/% |
|-----|--------|---------|----------|-------|-------|----------|
| 人 | -0.706 | 230 | 27 026.9 | 9.67 | 51.90 | 38.24 |
| 黑猩猩 | -0.576 | 519 | 58 323.2 | 4.70 | 54.03 | 39.25 |
| 长臂猿 | -0.570 | 536 | 60 222.3 | 4.68 | 52.44 | 38.82 |
| 猕猴 | -0.575 | 536 | 60 269.6 | 4.70 | 51.94 | 39.39 |
| 獭 | -0.581 | 518 | 58 284.1 | 4.70 | 55.13 | 38.96 |
| 欧洲兔 | -0.585 | 518 | 58 344.2 | 4.70 | 54.28 | 39.25 |
| 熊猫 | -0.566 | 517 | 58 115.0 | 4.69 | 54.34 | 39.11 |
| 野猪 | -0.570 | 535 | 60 179.3 | 4.70 | 53.59 | 38.53 |
| 牛 | -0.583 | 534 | 60 127.2 | 4.70 | 54.33 | 37.81 |
| 非洲象 | -0.587 | 517 | 58 191.1 | 4.70 | 54.80 | 38.53 |
| 褐家鼠 | -0.552 | 513 | 57 743.7 | 4.77 | 55.61 | 39.68 |
| 小家鼠 | -0.544 | 517 | 58 155.1 | 4.78 | 56.17 | 39.39 |
| 荷兰猪 | -0.576 | 536 | 60 256.4 | 4.70 | 53.67 | 42.86 |
| 灰仓鼠 | -0.482 | 701 | 77 203.8 | 5.26 | 58.13 | 41.13 |

2.2.2 不同物种 *Hsf2* 的密码子偏爱性 本研究中所编辑选取的各物种 *Hsf2* 基因序列编码区中密码子有效值(ENC)为 59.297(<61),偏爱指标(CBI)为 0.248(>0),经过 χ^2 检验并计算得到未校正的 χ^2 值为 0.187,说明 *Hsf2* 基因对密码子有较强的偏爱性^[9]。

2.2.3 不同物种 *Hsf2* 的同义替换和非同义替换 本

研究中的 14 个物种的 40 条 *Hsf2* 基因 CDS 区中有 140.97 个同义替换的平均位点数,552.03 个非同义替换平均位点数。不同物种同义替换位点数范围为 139.00~145.17,同义替换核苷酸多样性均值为 0.204 0。非同义替换位点数范围为 547.83~554.00,非同义替换核苷酸多样性均值为 0.008 9。分析发现,所研

究物种 *Hsf2* 基因的非同义替换位点数均明显高于同义替换位点数。

3 结论与讨论

3.1 不同物种 *Hsf2* 核苷酸特征分析

从所研究物种 *Hsf2* 基因序列多态信息、单倍型及其多样性数据得出各物种的遗传相关参数(多态位点数、单倍型多样性等)不一致,表明 *Hsf2* 基因在种群间存在遗传变异,*Hsf2* 基因序列编码区在种内表现较为保守,种间则表现有较丰富的遗传多样性,数据显示出 *Hsf2* 基因在一些物种间如黑猩猩与长臂猿、褐家鼠与非洲象、大熊猫、欧洲兔和狨之间比较保守。人与其他物种相比,*Hsf2* 基因多态位点数与核苷酸多样性的值最高,因此,说明人的 *Hsf2* 基因存在较丰富的遗传多样性。

表 3 数据显示,不同物种间核苷酸歧异度和遗传分化系数的变化范围均较大,说明不同物种间遗传分化较为明显。其中,人和黑猩猩以及人和长臂猿间的核苷酸歧异度最小,表明其亲缘关系较近;大熊猫和欧洲兔之间的核苷酸歧异度很接近,表明它们二者之间有较近的亲缘关系;同时数据和聚类图显示出小家鼠和褐家鼠之间亲缘关系也很近。分析发现,荷兰猪与其他物种的核苷酸歧异度最大,说明与其他物种间亲缘关系较远,上述结果与动物分类学相一致。另一方面,各物种基因中的 G+C 含量同样得出了上述结论。

3.2 不同物种 *Hsf2* 氨基酸特征分析

本研究结果表明,热激因子 *Hsf2* 多肽链表现为亲水性,理论等电点数值大多小于 7(人为 9.67),说明该多肽链为酸性。其中人的 *Hsf2* 理论等电点为 9.67,与其他物种相比差异较明显,可能是所编辑选取的氨基酸序列较短、样本含量限制有关,具体原因有待于进一步研究。不稳定系数在 51.90~58.13,软件分析得出的结论是 *Hsf2* 多肽不稳定。

蛋白质在翻译过程中,物种间或物种内的不同基因在密码子的使用上一般都具有明显的偏爱性^[10]。本研究也得出 *Hsf2* 基因对密码子具有较强的偏爱性。

蛋白质在翻译过程中某些碱基会发生一定程度的替换,包括同义替换和非同义替换。同义替换现象的发生大多不受自然选择的控制,同义替换速率远远高于非同义替换速率,且这种现象发生的速率与基因密切相关,这被认为是净化选择的结果^[11]。而在某些基因中,非同义替换速率则远远高于同义替换速率,原因在于达尔文的正向选择^[12]。本研究结果表明,所选物种 *Hsf2* 基因的非同义替换位点数均明显高于同义替换位点数,由此说明,所研究的物种在进化过程中很大

程度上可能受到达尔文的正向选择的影响。利用软件分析数据时发现,野猪 *Hsf2* 基因中非同义替换位点数为 554.00,与其他物种相比,其非同义替换位点数较多,说明野猪 *Hsf2* 基因编码区的非同义替换较其他物种更为明显,其原因有待进一步研究。

目前,国内外对热激因子的研究已经深入到该转录因子家族的蛋白结构层次,对于 HSF 与生物体热应激反应与应对环境胁迫的关系有待进一步研究。且国内外对 HSF2 的研究较少,因此,进一步研究 HSF2 对生物体生长、发育及分化和控制基因特异性表达方面的作用是以后研究的重要任务,这对动物的遗传育种与繁殖及动物品种的抗病抗逆性研究也具有重要意义。

参考文献:

- [1] Hartl F U, Hayer-Hartl M. Protein folding: molecular chaperones in the cytosol: from nascent chain to folded protein[J]. Science, 2002, 295: 1852-1858.
- [2] Clos J, Westwood J T, Becker P B, et al. Molecular cloning and expression of a hexameric *Drosophila* heat stress factor subject to negative regulation[J]. Cell, 1990, 63: 1085-1097.
- [3] Morimoto R I. Regulation of the heat shock transcriptional response: cross talk between a family of heat shock factors, molecular chaperones and negative regulators [J]. Genes Dev, 1998, 12: 3788-3796.
- [4] Schoffl F. Regulation of the heat-shock response[J]. Plant Physiol, 1998, 117: 1135-1141.
- [5] Baniwal S K. Heat stress response in plants: a complex game with chaperones and more than twenty heat stress transcription factors[J]. J Biosci, 2004, 29: 471-487.
- [6] Naki A A, Tanabe M, Kawazo E Y. HSF, a new member of the human heat shock factor family which lacks properties of a transcriptional activator[J]. Mol Cell Biol, 1997, 17: 469-481.
- [7] Liu X D, Liu P C, Santoro N, et al. Conservation of a stress response: Human heat shock transcription factors functional substitute for yeast HSF[J]. EMBO J, 1997, 16: 6466-6477.
- [8] Tanabe M, Kanazo E Y, Takea S, et al. Disruption of the HSF3 gene: Results in the severe reduction of heat shock gene expression and loss of the motolerance[J]. EMBO J, 1998, 17: 1750-1758.
- [9] 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: 446-448.
- [10] Ghosh T. Studies on codon usage in *Entamoeba histolytica* [J]. International Journal for Parasitology, 2000, 30: 715-722.
- [11] 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.
- [12] Guo Zhong-ping. Introduction to population genetics[M]. Beijing: Agricultural Press, 1993: 298-332.