

小鼠中血清因子结合位点位置分布及保守性的生物信息学分析 *

沈 霞 谭从娥 陈 莹 冯居君 张 鑫

(陕西中医学院药学院 陕西 西安 712046)

摘要 目的 :血清反应因子在与心血管相关的疾病基因调控方面的作用越来越重要。血清反应因子识别的结合位点 CArG 元件因其重要的基因调控作用近年来随之备受关。本研究目的是揭示血清因子结合位点的位置分布与功能的关系及 CArG 元件内部各个位点的保守性。方法 本研究应用生物信息学方法结合遗传学方法对小鼠中 CArG 元件的位置分布、位点替换率及 GO 分类进行深入研究。结果 结果表明 ,71%的功能 CArG 元件分布在转录起始位点上游 ,且距离转录起始位点越近 ,CArG 元件的数量越多。保守性分析发掘出元件内部的替换冷点、热点及替换规律。GO 分类结果显示 ,CArG 依赖性基因多为信号转导和细胞骨架蛋白。结论 :上述研究结果将为准确预测候选 CArG 元件提供重要理论基础 ,同时也将为更为深入阐述 SRF 的调控模式奠定基础。
关键词 血清因子结合位点 ,CArG 元件 ,生物信息学 ,替换率 ,位置分布
中图分类号 :Q95-3 Q-33 文献标识码 :A 文章编号 :1673-6273(2011)11-2041-04

Bioinformatics Analysis of Position Distribution and Conservation of Serum Response Factor Binding Site in Mice*

SHEN Xia, TAN Cong-e, CHEN Ying, FENG Ju-jun, ZHANG Xin
(Shanxi university of Chinese medicine, Xhanxi, Xi'an, 712046, China)

ABSTRACT Objective: The serum response factor is one of the valuable transcription factors because of its important role in gene expression. In turn, cis-elements CArG bound by serum response factor is getting more and more attention. This study is going to reveal the position distribution and conservation of CArG elements. **Methods:** We study the position distribution, substitution of CArG elements and GO of CArG dependent genes in mouse by the bioinformatics and genetic methods. **Results:** Our results show that 71% functional CArG elements exist upstream to the annotated TSS, with copy number increasing as one move closer to the TSS. The conservation analysis revealed the substitution pattern. What' more, GO analysis results show that the CArG-dependent genes mostly are signaling and cytoskeleton protein. **Conclusion:** The results will be a great help in predicting the candidate CArG elements and attempting to distinguish the functional CArG elements through alignments. And it will be fundamental in study the regulatory system of SRF.
Key words: Serum response factor; CArG elements; Bioinformatics; Substitution; Position distribution
Chinese Library Classification(CLC): Q-33 **Document code:** A
Article ID:1673-6273(2011)11-2041-04

前言

SRF(serum response factor)是有机体内广泛存在的转录因子 ,自 27 年前被发现 ,因其调控功能多样性和重要性随即成为生物学中研究的热点之一^[1]。近些年研究表明 ,SRF 是中胚叶形成和发育所必需的重要转录因子之一。其对心脏的发育与形成 ,神经元的迁移和聚集都有很重要的作用^[2]。SRF 对学习能力的发育和形成也起到关键作用^[3]。更为重要的是 SRF 的非正常调节作用对其下游基因的影响与人类疾病例如癌症和心脏病有明显相关性^[4]。总而言之 ,SRF 对与生物体发育的广泛影响和调节作用使其成为医学研究的热点转录因子之一。SRF 的结合位点是位于基因上游的 CArG 元件 ,CArG 元件的与 SRF 的准确识别和结果对于 SRF 发挥生物学功能有着不容忽视的作用^[5]。

CArG 元件由 10 个碱基组成 ,5' 端为 CC,3' 端为 GG,因其

内部富含 A 和 T 而被命名为 CC(A/T-rich)GG,缩写为 CArG。越来越多的研究发现 ,CArG 广泛存在于基因的启动子区 ,例如骨骼细胞中的 α 肌动蛋白 ,α 肌浆球蛋白的重链 ,心脏和骨骼的肌浆球蛋白的轻链以及心脏的肌钙蛋白 T^[6]。几乎所有含有 CArG 元件的基因都可以被归为两类 ,肌细胞限制型或者与生长相关的基因^[7]。人类和模式生物的基因组完成测序以后 ,越来越多的研究表明 ,有机体的正常生长发育不仅仅是由基因序列所控制的 ,基因的时空精确表达也是重要因素之一。随着研究的不断深入 ,顺式作用元件的重要性也日益体现出来。顺式作用元件的位置分布和替换规律与调节基因表达直接相关 ,因而受到研究人员的关注^[8]。

然而 ,迄今为止有关 CArG 元件在转录起始位点附件分布、位点保守性分析的研究尚未见报道。本研究应用 NCBI、DBTSS、GO 三大数据库 ,通过对 CArG 元件位置分布信息及保守性信息的深度挖掘解析功能 CArG 元件在转录起始位点附

* 基金项目 国家自然科学基金(NO:81072731)
作者简介 沈霞(1979-) ,女 ,博士 ,讲师 ,主要研究方向 :生物信息学。
电话 :029-38185165 E-mail :jxrain@163.com
(收稿日期 :2011-01-05 接受日期 :2011-01-28)

近分布、替换率、GO 分类与功能的关系,更深入阐述以上因素与 CArG 元件功能之间的关系。本研究结果将为深入研究哺乳动物 SRF 调控体系奠定科学依据,进而为阐明人中 SRF 的调控网络与人类疾病的关系提供技术方法和奠定理论基础。

1 材料与方法

1.1 实验数据下载

从 NCBI 数据库中下载得到小鼠全基因组序列(V3.61),其中包括基因所处的位置及基因的长度大小(<http://www.ncbi.nlm.nih.gov/>)。接着,由于研究所需,根据基因的序列号从 DBTSS(Transcription Start Sites, version6.0)数据库下载得到所需基因的转录起始位点上下游 4kb 的基因组序列。DBTSS 数据库涵盖了小鼠及人类转录起始位点的精确信息,启动子区的数据以及转录起始位点上游的基因组序列(<http://dbtss.hgc.jp/>)。Gene Ontology 可分为分子功能(Molecular Function)、生物过程(biological process)和细胞组成(cellular component)三个部分(<http://www.geneontology.org/>)。先利用从 NCBI 上查到的基因序列找到与之对应的 GO 分类号,进而通过 GO 分类号可对应查到基因的功能类别。

1.2 小鼠中功能 CArG 元件的搜集与整理

由于本研究旨在发掘功能 CArG 元件在转录起始位点附近的分布情况,所以本研究选取有功能的元件进行分析。首先搜索文献,从文献中提取出小鼠中的受功能 CArG 元件调控的基因的信息。从 NCBI 上下载 CArG 元件调控的基因上下游 4000bp 的序列并获取基因的基因序列号,然后用 DBTSS 数据库确定每条基因转录起始位点的位置。最后用计算机程序扫描确定功能 CArG 元件的位置,最终搜集整理得到小鼠功能 CArG 元件共 152 个(附件)。并通过 DBTSS 数据库确定 CArG 元件在启动子、内含子、外显子及非编码区的分布情况。

1.3 小鼠中功能 CArG 元件替换率分析

CArG 元件的序列组成比较特殊,5' 端的 CC,3' 端的 GG,中心则是 AT 的不同排列组合,所以统计其不同位点的替换数要采用不同的方法。5' 端的 CC 如果发生替换,将会是到 G、A 或者 T 的替换。所以会有三种替换形式的出现,3' 端的 GG 也是如此。中心序列与两边的序列则有所不同,中心序列由 A 和 T 不同排列组成,一共是 6 个碱基。所以发生 A 到 T,或是 T 到 A 的替换都不能从序列上体现出来,也不能记为一次替换。所以本研究定义中心序列只有两种替换形式。中心的六个碱基只有发生到 C 或者 G 的替换算作是一次替换。所以 CArG 元件的不同位点的替换数需要采取不同的计数方法。两端的替换形式为三种,而中心序列只有两种。基于以上理论编写程序统计了 CArG 元件各个位点的替换数,并计算出替换率。

2 结果

2.1 功能 CArG 元件在转录起始位点的分布

普遍来说,顺式作用元件在转录起始位点的分布与其功能是具有一定的相关性,功能 CArG 元件作为顺式作用元件也不例外。据此,功能 CArG 元件在转录起始位点附近的分布必定有其一定的规律性。本研究统计了功能 CArG 元件在基因转录起始位点的分布,结果如图 1 所示。图中,TSS 是指基因的转录

起始位点,其中横坐标为与基因转录起始位点的距离,纵坐标为每一个区间内 CArG 元件为数量。从图中可以看出功能 CArG 元件在转录起始位点附近的分布具有明显的偏好。图中显示,功能 CArG 元件在转录起始位点附近的分布明显具有向转录起始位点聚集的趋势,越接近基因转录起始位点,功能 CArG 元件数目越多。此外,绝大部分功能 CArG 元件分布在转录起始位点上游(108/152, 71%),且统计检验显示转录起始位点上游的功能 CArG 元件数要远远多于下游的功能 CArG 元件数(t-test, P= 3.27E-8)。

此外,本研究还就功能 CArG 元件在转录起始位点上游的分布做了回归分析,做了回归方程。发现上游的功能 CArG 元件是服从负幂指数分布的。方程式如下:

$$f(x)=0.0064(0.342e^{-0.00947x}+0.07492e^{-0.00063x})$$

f(x) 是 x 的密度分布,x 代表的是 CArG 元件距离基因转录起始位点的距离(Chi-square is 15.7326, P=0.071)。这一结果对于利用比较基因组学推测 CArG 元件的功能有着重要的理论意义。同时也为开发 CArG 元件预测软件奠定基础

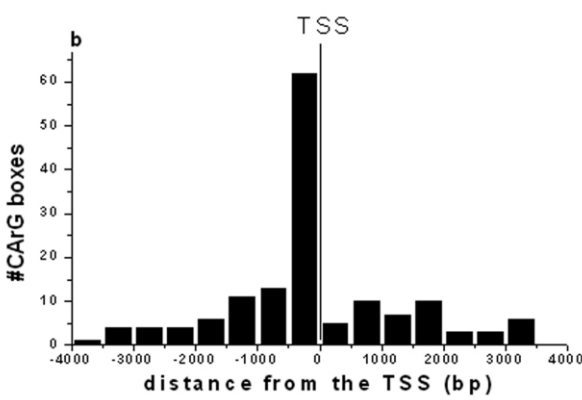


图 1 功能 CArG 元件在转录起始位点的分布
Figure1 Position distribution of CArG elements to the TSS

2.2 小鼠中功能 CArG 元件的替换率规律

功能 CArG 元件内部允许一个碱基替换。为了深入挖掘功能 CArG 元件的序列各个位点的替换规律,本研究对其每个位点的替换率分别进行统计,以期发现功能 CArG 元件的各个位点的替换率,揭示替换发生的热点和冷点。图 2 中横坐标为功能 CArG 元件的十个碱基,从 5' 端到 3' 端依次为 1 到 10,纵坐标为每个位点的替换率。图 2 表明,小鼠功能 CArG 元件的每一个位点的替换数都不尽相同,说明该元件中的不同位点因其功能的重要性不同而承受不同的选择压力。小鼠功能 CArG 元件 4 位点是保守性最高的位点,这一位点的替换数明显少于元件中的其它位点,为元件中发生替换的冷点。在小鼠也存在替换的热点,和元件中其他位点相比,小鼠中 5 和 9 位点的替换数最多,明显高于其它位点的替换率数值。据此推测,这一变化可能反映出在与 SRF 互动时,CArG 元件的各个位点的对形成正确结合构象的贡献不同。

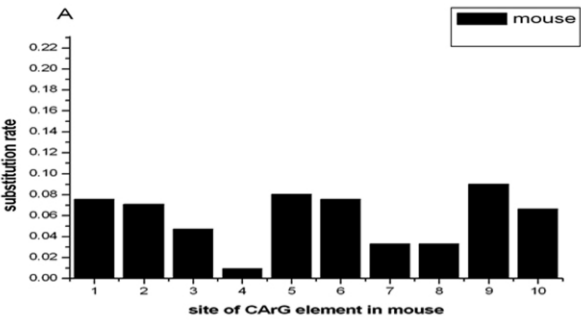


图2 CArG 元件各个位点替换率统计
Figure 2 Substitution pre site of CArG elements

2.3 CArG 元件在启动子、内含子、外显子等的分布
本研究统计分析 152 个 CArG 元件在基因外显子、内含

表 1 CArG 元件在启动子 ,内含子 ,外显子 ,非编码区等的分布
Table 1 CArG elements distribution in promoter,excon,intron and intergenic

	启动子区 Promoter	内含子 Intron	外显子 Excon	非编码区 Intergenic
数量 /Number	107	36	2	7
百分比 /Percent	70%	24%	1%	4%

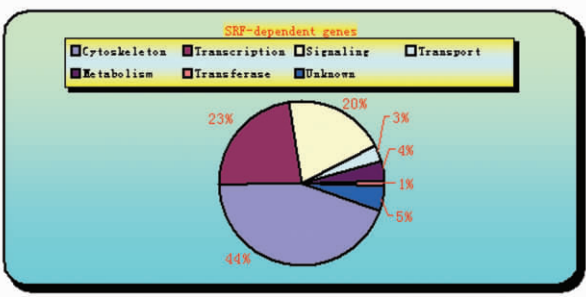


图3 GO 分类示意图
Figure 3 The results of GO analysis

研究发现从单细胞真核生物到小鼠 ,SRF 和功能性 CArG 元件在生物生长和发育过程中有很重要的作用。同时遗传学研究证明 SRF 和功能性 CArG 元件对细胞内各物质的迁移、维持正常肌动蛋白和收缩性蛋白的生物学过程起到不可替代的作用。新近研究表明超过半数的 SRF 目的基因编码与细胞骨架动态平衡相关的蛋白^[10]。GO 分类分析进一步证明 SRF 调控的基因多为细胞骨架蛋白相关基因。

3 讨论

CArG 元件在转录起始位点附近的分布与 SRF 调控机理密不可分 ,阐明 CArG 元件的分布规律有利于用比较基因组的方法预测新的功能 CArG 元件。在 c-fos 基因中 ,位于上游 -313 的 CArG 元件被 SRF 结合 ,SRF 又与 RNA 聚合酶 作用促使前转录起始复合物的形成^[11] ,离转录起始位点较近 ,利于全酶中的 TF F 因子与 SRF 直接作用^[12] ;在 SMα-actin 启动子中 ,研究表明 ,改变 CArG 元件与转录起始位点的距离 ,则使该基

子、启动子中的分布(表 1)。结果表明其中有 107 个分布于启动子区 ,占 74% ,36 个分布于内含子中 ,占 23% ,Dnajb1 和 Impact 分布在外显子中 ,这是在研究顺式作用元件 CArG 元件以来 ,最早发现的、也是目前为止仅有的两个功能性 CArG 元件位于编码区的基因^[9]。本研究结果表明 ,绝大多数 CArG 元件分布在启动子区。然而 CArG 元件分布的范围也很广泛 ,在基因上下游 4000bp 的范围内都有分布。

2.4 CArG 元件的 GO 分类

已发现的 122 个 SRF 目的基因 ,将它们按照 Gene Ontology(GO) annotation 分类 ,可以划分为细胞骨架、收缩性或支持蛋白 ,核内转录 ,信号转导 ,代谢基因等(图 3)。调控细胞骨架蛋白占 44% ,转录方面的占 23% ,信号转导方面占 20% ,相对来说 ,运输、代谢等方面的很少。

因的表达大幅度降低^[13]。由此可见 CArG 元件离转录起始位点的距离直接影响了其下游基因的表达量。本研究运用生物信息学方法结合统计学方法分析了小鼠中功能 CArG 元件的在基因转录起始位点附件的分布情况。研究发现 ,CArG 元件在基因上游的位置受到很强的功能限制 ,距离转录起始位点越近其有功能的几率愈大 ,并利用已有数据推测出其位置分布的方程。这些信息将为更加准确的设计、开发 CArG 元件预测的软件奠定坚实的理论基础。

人类基因组计划完成以后 ,越来越多的研究发现顺式作用元件的突变更容易改变生物体的表形 ,并且进化选择对于顺式作用元件的选择作用也比编码区更加有效^[8]。更有研究表明人类很多疾病和顺式作用元件的变化紧密连锁^[14]。血清反应因子的调控变化与人类心血管疾病有显著的相关性^[3] ,因此发掘 CArG 元件的替换规律对其功能的研究有着重要的价值。本研究统计分析功能 CArG 元件的替换速率 ,发现元件每个位点的替换规律 ,发现了替换热点和冷点。这一研究结果为有效、合理设计点突变实验研究 CArG 元件与 SRF 作用模式及对下游基因调控的影响提供数据支持。

Lilly 等报道 CSRPI 的内含子中有一个功能性 CArG ,在三类肌细胞(平滑肌细胞、骨骼肌细胞、心肌细胞)形成过程中 ,对半乳糖苷酶报告基因的表达起到增强子的作用 [15] ,在平滑肌细胞中 ,CSRPI 的 CArG 元件的增强子作用仅限于动脉平滑肌细胞^[16] ,在α-SMA 启动子区域内有多个 CArG 元件 ,James J 等通过实验指出 ,CArG 元件在物种中高度保守 ,如果位于内含子中的 CArG 发生突变 ,则影响平滑肌细胞的形成^[17]。本研究结果显示 70%的 CArG 元件分布在启动子区 ,24%的 CArG 元件分布在内含子中。结合前人实验研究 ,可以推测分布在内

含子中的 CArG 元件属于平滑肌细胞的特异性增强子元件。根据以上研究结果,可以根据未知功能的 CArG 元件的位置分布对其功能进行预测。这将有效的提高实验效率,减少实验消耗。综上所述,本研究首次用生物信息结合遗传学的方法分析小鼠中功能 CArG 元件的位置分布特征,并探究其位点的保守性和位点的替换规律。此项研究揭示了功能 CArG 元件的位置分布方程和特有的替换模式,为准确预测候选 CArG 元件,应用点突变实验验证 CArG 元件与 SRF 的作用机制提供了重要的理论基础。在此基础之上,可以更深入探讨 CArG 元件的变化对 SRF 调控网络模型的作用模式及对下游基因调控的影响,阐明与 SRF 调控相关的心血管疾病的致病机理。此外,本研究为揭示其它顺式作用元件序列特征建立了新的研究平台。对其它顺式作用元件的序列特征的深入发掘,将为准确预测顺式作用元件、分析基因表达以及研究基因表达调控网络模型奠定坚实理论基础。

致谢:感谢西北农林科技大学生物信息中心陶士珩教授为本研究提供的可行性建议和无私帮助。同时感谢生物信息中心各位同学的鼎力协助。

参考文献(References)

- [1] Greenberg ME, Ziff EB. Stimulation of 3T3 cells induces transcription of the c-fos proto-oncogene [J]. *Nature*, 1984, 311:433-438
- [2] Knoll, B., Kretz, O., Fiedler, C., et al. Serum response factor controls neuronal circuit assembly in the hippocampus [J]. *Nat. Neurosci*, 2006, 9: 195-204
- [3] Etkin, A., Alarcon, J.M., Weisberg, S.P, et al. A role in learning for SRF: Deletion in the adult forebrain disrupts LTD and the formation of an immediate memory of a novel context [J]. *Neuron*, 2006, 50: 127-143
- [4] Miano J M. Role of serum response factor in the pathogenesis of disease [J]. *Laboratory Investigation*, 2010, 90 :1274-1284
- [5] Miano J M. Deck of CArGs [J]. *Circ. Res*, 2008; 103; 13-15
- [6] Zhang SX, Gras EG, Wycuff DR, et al. Identification of direct serum response factor gene targets during DMSO induced P19 cardiac cell differentiation [J]. *J Biol Chem*, 2005, 280:19115-19126
- [7] Miano J M., Long X, and Fujiwara K. Serum response factor: master regulator of the actin cytoskeleton and contractile apparatus [J]. *Am J Physiol Cell Physiol*, 2007, 292: C70-C81
- [8] Sethupathy P, Giang H, Plotkin J B., and Hannenhalli S. Genome-Wide Analysis of Natural Selection on Human Cis-Elements [J]. *PLoS ONE*, 2008, 3(9): e3137
- [9] Sun Q, Chen G, Streb JW, et al. Defining the mammalian CArGome [J]. *Genome Res*, 2006, 16:197-207
- [10] Halene S, Gao Y, Hahn K, et al. Serum response factor is an essential transcription factor in megakaryocytic maturation [J]. *Blood*, 2010, 116(11): 1942-1950
- [11] McDonald O G., Wamhoff B R., Owens G K., et al. Control of SRF binding to CArG box chromatin regulates smooth muscle gene expression in vivo [J]. *J. Clin. Invest*, 2006, 116:36-48
- [12] Joliot V, Demma M, Prywes R. Interaction with RAP74 subunit of TFIIF is required for transcriptional activation by serum response factor [J]. *Nature*, 1995, 373:632-635
- [13] Mack CP, Thompson MM, Owens GK, et al. Smooth muscle α -actin CArG elements coordinate formation of a smooth muscle cell-selective, serum response factor-containing activation complex [J]. *Circ Res*, 2000, 86:221-232
- [14] Kleinjan, D.A., Heyningen, van V. Long-range control of gene expression: Emerging mechanisms and disruption in disease [J]. *Am J Hum Genet*, 2005, 76: 8-32
- [15] Lilly B, Olson EN, Beckerle MC. Identification of a CArG box-dependent enhancer within the cysteine-rich protein 1 gene that directs expression in arterial but not venous or visceral smooth muscle cells[J]. *Dev Biol*, 2001,240:531-547
- [16] Kumar MS, Owens GK: Combinatorial control of smooth muscle-specific gene expression [J]. *Arterioscler Thromb Vasc Biol*, 2003, 23: 737-747
- [17] James J. Tomasek, Joel McRae, Gary K. Owens, and Carol J. Haaksma Regulation of α -Smooth Muscle Actin Expression in Granulation Tissue Myofibroblasts Is Dependent on the Intronic CArG Element and the Transforming Growth Factor- β 1 Control Element [J]. *Am J Pathol*, 2005, 166:1343-1351