Serum Response Factor Binding Sites in Human and Mouse Genome: Conservation，Diversity and Evolution*

SHEN Xia[1△], WU Wen-wu[2] , TAN Cong-e [1] , FENG Ju-jun [1]

(1 College of Chinese traditional herb, Shaanxi university of Chinese medicine Shaanxi, China;

2 Bioinformatics Center, College of Life Science,Northwest A&F University, Yangling, Shaanxi, China)

ABSTRACT Objective: CArG cis-elements, short DNA consensus sequences that binding by serum response factors, are presently being intensively studied, but little is known about the sequence type and the evolutionary pattern of functional CArG elements. Methods: The present study was performed a genome-scale diversity and evolutionary analysis of the CArG element in the mammalian genome. Results :Diversity analysis showed that the sequences type of CArG element were significantly diverse in both human and mouse genomes. The main sequence types of CArG element were not entirely similar in the two genomes. Orthologous analysis indicated that functionally important CArG elements probably evolved from two different origins. Conclusion: The results presented here will fundamentally improve future CArG elements prediction, regulatory determinant pattern detection and analysis of SRF-dependent gene expression.

Key words: CArG elements; Conservation; Diversity; Evolution; Bioinformatics

Chinese Library Classification(CLC): Q95-3 Q-33 Document code: A

Article ID:1673-6273 (2011) 10-1821-06

Introduction

It was indicated that subtle alterations in the timing, location, and level of protein synthesis also had considerable impact on gene regulation at both the molecular and organismal levels [1]. Changes in the cis-regulatory elements that control gene expression could influence a variety of functionally significant traits, including morphology, behavior, and physiology[2]. Thus, additional mechanisms likely exist that can explain variations in gene expression. Consistent with potential role of cis-elment in determining phenotypic variability, there is now evidence for natural selection acting on specific gene regulatory elements[3]. A surprisingly large number of cis-regulatory elements have been subject to coordinated genome-wide modifications during vertebrate evolution, such changes in cis-regulatory elements can have a substantial impact upon the relationship between the transcription factor and its binding sites throughout the genome[4] .

Serum response factor (SRF), which was first defined by Richard Treisman[5], is a member of the MADS-box familyof transcription factors and is one of the best understood DNA-binding proteins in the human [6]. Studies on the signaling path of SRF revealed its essential role on the normal development and maintenance of smooth muscle cells and neurons.And a growing number of experimental and human diseases are associated with changes in SRF expression, suggesting that SRF has a role in the pathogenesis of disease [7]. Indeed, DNA-binding domains of transcription factors are often highly conserved across species [8]. For this reason , studies regarding cis-regulatory element of the SRF binding site will provide key information for understanding the expression mechanism of SRF-dependent genes.

The CArG elements, the binding site of SRF, have been well studied by a large number of experiments. This short DNA motif usually contains ten nucleotides with the consensus sequence CC $(A/T)_6$GG. The functional studies of CArG element, the position distribution of CArG element in the promoter region and CArG flanking sequences[9] have been studied widely, much less information is available on the properties of its sequence type and evolution. A more in-depth understanding of interaction of CArG elements and the SRF is needed in order to document the sequence type and the diversity of this consensus.

At present, there are 185 CArG elements experimentally validated in the human genomes[10] and 210 CArG elements experimentally validated in the mouse genome[11,12]. This study had performed sequence type, diversity and evolutionary analysis of the CArG element by using bioinformatics methods.

1 Materials and Methods

1.1 Sequence data

Information on CArG-containing target genes and CArG elements in the human and mouse genomes was collected from the literature [10-12]. The genome with the elementsof which had been experimentally validated were considered in the present study. Mouse and Human Genome data were downloaded from the NCBI (human v 36.3, mouse v37.1). The genomic sequences from +4 to -4kb around the annotated transcription start site (TSS) in the pro - moter region of all genes were downloaded from the database of

*Fund：the National Science Foundation of China (NO:81072731)

Author: Shen Xia (1979-) , female, doctor degree, study on the bioinformatics, Tel:029-38185165, Email :jxrain@gmail.com

△Corresponding author: Xia Shen, Email address:jxrain@gmail.com

(Receive: 2011-02-03 Accepted:2011-02-28)

Transcription Start Sites (DBTSS, version6.0). The DBTSS contained detailed information as to the genomic positions of the annotated TSS (based on UCSC hg18, mm8) and the adjacent promoters for human and mouse RefSeq transcripts (http://dbtss.hgc.jp/). This database also contained comparative promoter information of orthologous genes for mice and humans.

1.2 Selection of the background DNA

Previous studies had shown that almost every mouse and human gene contained more than one CArG-like sequence in the promoter region [13]. For this, CArG-like sequences can be viewed as representing the best control sets for the analysis of functional sequences. The term CArG-like sequence which was used as background DNA in this study refers to non-functional CArG-like motifs. There are four parameters used to define CArG-like Sequence: 1 non-functional; 2 lie in the promoter region; 3 The sequence arrangement of CArG-like Sequence is CC $(A/T)_6GG$, similar to the CArG elements; 4 The sequence has ten base pairs.

Since all of the known CArG elements reside within 8kb of an annotated TSS in the promoter region, the genomic sequences from +4kb~-4kb around the annotated TSS in the promoter region of the control genes of mouse and human genome were downloaded from DBTSS. Then, a computer program was built that could search for CArG-like sequences in these genomic sequences. Given that sequences with one nucleotide substitution from the classical motif are often encountered and can still be functional, setting a standard that allowed one site mismatch for the scanning was considered reasonable. However, these searches did not take into account that some substitutions in CArG-elements were probably more favorable than others. At last, 500 CArG-like sequences were separately selected from the computer-searched CArG-like sequences in mice and humans that lay within 8kb of the annotated TSS in the promoter region of control genes. These 500 CArG-like sequences were used as the background DNA for study of functional elements.

1.3 Sequence diversity analysis of functional CArG elements

The sequence characters of CArG elements suggest that there are 1216 permutations of these elements. It could imagine that sequence type of CArG elements should under a strong functional strain. It will not be possible that 1216 permutations of CArG elements have the same affinity when interact with SRF. There will be the CArG element of certain sequence pattern that is more important than others. In order To distinguish the major type of CArG sequence, a computer program was run to scan all of the known functional CArG elements in mouse and human genomes to find all the type of CArG sequence. Each type was separately enumerated in each of the two species.

1.4 Evolutionary analysis of functional CArG elements

Previous study had shown that most CArG elements appear to be evolutionarily conserved in humans and mice. Therefore, the functional CArG elements in the mice and humans were compared by orthologous analysis to ascertain the evolutionary origin of the CArG elements.

This study used only CArG elements from the literature that had been experimentally validated previously. First, the orthologous pairs of the CArG-containing target genes of mice and humans were detected according to the NCBI release of homologous data. Second, the orthologous promoters of these orthologous genes were identified on the DBTSS. Third, the pairs of CArG elements were manually selected on the premise that (i) the CArG elements should be similar in both sequence and relative position to the TSS in their respective genes; (ii) the CArG elements had been previously experimentally validated. Owing to limited data on functional CArG elements, only 25 pairs of candidate CArG elements were found by this analysis.

2 Results

2.1 Sequence pattern of functional CArG elements

Based on 20 years of DNA-protein and promoter analyses, as well as comparative genomics, it was thought that SRF may potentially bind to 1216 permutations of a CArG element, with CC $(A/T)_6GG$ emerging as a consensus sequence. However, after the sequence logo analysis we found that A\T in the core region of CArG element was not arranged randomly. Fig1 indicated an obviously TATA "motif" in the core region within the functional CArG elements in both human and mouse genomes.

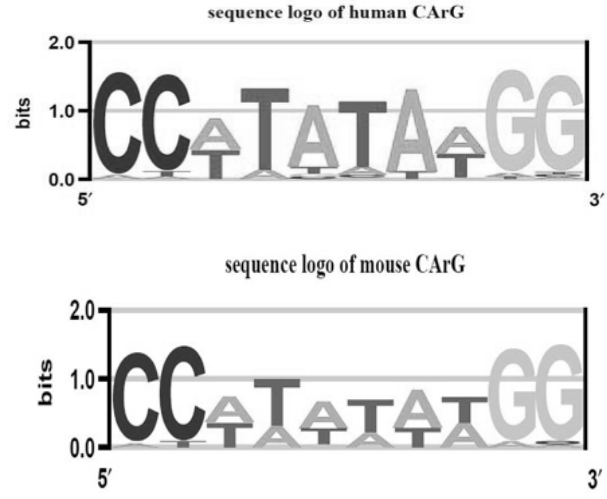

Fig. 1 Sequence logo of CArG elements generated from 185 functional CArG element and 210 functional CArG element in mouse in human using enoLOGOS[14]

The substitution rate (i.e., substitution rate inferred nucleotide changes) of each site in both human and mouse genome within the CArG elements were computed in order to find the best conserved site in the elements. Table 1 showed that in the mouse genome, site 4 was the best conserved site within the functional CArG elements. While in the human genome, site 3 showed the least amount of variation . On the contrary, most substitution occurred at position 2 , 5 in human genome and position 9 in the mouse genome.

Table 1 The substitution rate of functional CArG elements compared to the background

Position	1	2	3	4	5	6	7	8	9	10
Mouse/ substitution rate	0.0758	0.071	0.047	0.009	0.081	0.076	0.0332	0.0332	0.09	0.0664
Human/ substitution rate	0.049	0.076	0.011	0.016	0.076	0.054	0.027	0.038	0.054	0.0595

2.2 The diversity of sequence type of the CArG element in human and mouse genomes

Based on the sequence characteristics of the CArG element, an estimated 1216 permutations of this element would occur. In order to find the diversity of sequence arrangement and determine the main sequence type of the CArG element in mammals, a computer program that could document every sequence type in both the human and mouse genomes was developed. The program revealed 65 types among the total 185 CArG elements in humans and 122 types among the total 210 CArG elements in mice. In this study, a main type which appears more than four times among all of the type involved was defined. Because of the different sample sizes for both genomes (185 CArG elements for humans, 210 CArG elements for mice), the frequencies of each main type were calculated for both organisms. In this study, the frequency was the proportion of a main type in the total CArG elements of each genome.

Table 2 showed that over half of the main type in the mouse genome and nearly half of the main type in the human genome that were found in functional CArG elements were not found in the background DNA. Furthermore, the main type in the humans and mouse genomes also differed. A main type in the human genome, such as CCATATAAGG with a frequency of 0.049, was not main type in the mouse genome, with a frequency of 0.001. Similarly, the consensus CCTTATTTGG, which was main type in the mouse genome, was not main type in the human genome (frequency 0.032). In addition, one main type found in the human genome was not found in mice (CCATATAGGG).

Table 2 Comparison of the frequency of the main type of the CArG elements between the mouse genome and the human genome

	Human			Mouse		
Haplotypes	Functional	Background	Odd's ratio	Functional	Background	Odd's ratio
CCATAAAAGG	0.049	0.002	24.5	0.019	0	INF
CCATATAAGG	0.049	0	INF	0.001	0	INF
CCTTATAAGG	0.043	0	INF	0.019	0	INF
CCTTTTAAGG	0.043	0	INF	0.019	0	INF
CCAAATATGG	0.038	0	INF	0.024	0.002	12
CCTTATTTGG	0.032	0.002	16	0.043	0.002	21.5
CCTTTTATGG	0.032	0.002	16	0.033	0	INF
CCTTATATGG	0.032	0.002	16	0.029	0	INF
CCTAATATGG	0.027	0	INF	0.005	0	INF
CCTTAAAAGG	0.027	0.004	8	0.014	0.002	7
CCATATAGGG	0.027	0	INF	0	0	

This table contains the main types in both genomes. The frequency of each main types was calculated. Although the main type in the two genomes differed, certain similarities still remained. All of the main types in both genomes were fully perfect CArG consensuses without any substitution. The core regions of these perfect CArG consensuses presented an obvious TATA bias. Therefore, the functional CArG element still had some meaningful and common sequence characters.

2.3 Most CArG elements are conserved in the human and the mouse genomes

Only 25 pairs of orthologous candidates were found in the present study because of limited relevant experimental data (Table 3). Among these 25 pairs, most CArG element pairs (88%, 22) were perfectly conserved between the human and mouse genomes. Their substitution sites and directions, as well as their position relative to the TSS of the CArG- containing gene, were also entirely similar, indicating that these were orthologous pairs. In this table, the sequence and the relative position to the TSS are showed of the orthologous CArG element of human and mouse respectively. The grey filling is the different between the two genomes. Position is the relative position to the TSS obtained from the literatures.

Table 3 The orthologous of the CArG elements of the mouse genome and the human genome

Gene Symbol	RefSeq	Mouse CArG Sequence	Position	Human CArG Sequence	Position	RefSeq	Gene Symbol
Acta2	NM_007392	CCCTATATGG	-140	CCCTATATGG	-134	NM_001613	ACTA2
Acta2	NM_009610	GCTTATAAGG	-278	GCTTATAAGG	-268	NM_001615	ACTG2
		CCTTATATGG	-110	CCTTATATGG	-108		
Myh11	NM_013607	CCTTTTATGG	-1227	CCTTTTATGG	-1226	NM_022844	MYH11
Dmd	NM_007868	CCTTATATGG	-129	CCTTATATGT	-120	NM_004006	DMD
Srf	NM_020493	CCATATAAGG	-60	CCATATAAGG	-55	NM_003131	SRF
		CCATAAAAGG	-40	CCATAAAAGG	-40		
Vcl	NM_009502	CCTTATAAGG	-253	CCTTATAAGG	-253	NM_014000	VCL
Fos	NM_010234	CCATATTAGG	-313	CCATATTAGG	-314	NM_005252	FOS
FosB	NM_008036	CCTTATATGG	-273	CCTTATATGG	-273	NM_006732	FOSB
Egr1	NM_007913	CCATATATGG	-95	CCATATATGG	-95	NM_001964	EGR1
		CCATATTAGG	-120	CCATATTAGG	-199		
		CCTTATTTGG	-376	CCTTATTTGG	-378		
		CCTTATTTGG	-393	CCTTATTTGG	-394		
		CCATATAAGG	-430	CCATATAAGG	-432		
Egr2	NM_010118	CCATATATGG	-70	CCATATATGG	-70	BC035625	EGR2
Ier2	NM_010499	CCAAATTTAG	-1200	CCTAATATGG	-1225	BC003625	IER2
Il2RA	NM_008367	CCTTTTATGG	-266	CCTTTTATGG	-268	ENST00000281359	IL2RA
Mcl1	NM_008562	CCTTTTACGG	-55	CCTTTTATGG	-39	NM_182763	MCL1
Tuft1	NM_011656	CCTTTTAAGG	-611	CCTTTTAAGG	-616	NM_020127	TUFT1
Dusp2	NM_010090	CCTTGTATGG	-63	CCTTGTATGG	-64	NM_004418	DUSP2
Ctgf	NM_010217	CCATATACGG	-3658	CCATATACGG	-3652	BC087839	CTGF
Dusp5	AK153769	CCATATTTGG	-90	CCATATTTGG	-88	NM_004419	DUSP5
Cfl1	NM_007687	CCTTATTAGG	-1400	CCTTATTAGG	-1390	NM_005507	CFL1

However, there were 3 pairs of CArG elements that differed between the two genomes (Table 3, grey filled areas). One of these was Dmd in which two consensuses had only one different base at site 10. At this site, the base was G in mice and T in humans. The other pairs were Ier2 which had three different sites, and Mcl1which had two different sites. Since it was impossible for the orthologous pairs to differ in more than one site in such a short sequence as a CArG element (only ten base pairs) after speciation, the CArG elements pairs in Ier2 and Mcl1 were definitely not orthologous.

3 Discussion

Fig1 showed that the CArG element in mouse genome and human genome had the same reversible motif: CCWTA/TAWGG. The A\T in the core region of the CArG element was not randomly arranged, but showed an obvious TATA bias. The three-dimensional structure of SRF binding to the CArG provides that the central A\T base pairs produced a narrow minor groove and could be specifically recognized by the SRF [15]. However , the same sequence pattern in human and mouse did not mean that the CArG element was under the same selective pressure during evolution in different species. The best conserved site (site 4 in mouse and site 3 in human) is different between the two genomes. This result provided useful information for the design of site-directed mutagenesis experiments that could be used to compare relative the SRFbinding efficiencies of CArG promoters in functional assays.

Recently, an increasing number of human diseases have been linked to the changes in gene expression attributable to polymorphisms within regulatory elements[16]. The majority of known human polymorphisms occurring in non-coding regions are likely to underlie gene expression variation between humans[17]. The present study revealed the diversity of sequence types in the functional CArG element in mouse and human genome. The diversity of sequence type probably indicates the involvement of the changes

within functional CArG sequencese in regulating gene expression. Because of the reduced affinity of degenerate CArG elements bound to SRF, the main types in both genomes were all perfect consensuses. The sequence characteristics of the functional CArG elements found in this study proved to be a great help in the prediction of candidate CArG elements.

CArG elements often appear to be evolutionarily conserved between relatively distant species, such as humans, mice, chickens, rats and frogs. Interestingly, the presence of a CArG element in both human and mouse promoters is a strong indication of SRF binding [10]. The orthologous analysis in this study was consistent with previous results that showed most CArG elements found in the orthologous promoters of human beings and mouse to be evolutionarily conserved. Two exceptions was not found in the orthologous promoters of Ier2 and Mcl1 (Table 3). The significant differences in the sequences of CArG elements between these two pairs of genes strongly suggest that they may originate from different ancestries. A previous study on the distribution of functional CArG elements around the TSS of CArG-containing genes in the mouse showed that the promoter region of the CArG-containing genes contained more CArG-like sequences than did the background genes[13]. These CArG-like sequences would have more chance to be selected for bind with SRF under proper conditions and that these would then become fixed. Based on this assumption, we speculated that functional CArG elements probably evolve in two different ways. The main way would be to originate from a common ancestry this would explain the appearance of functional CArG elements that are conserved in distant taxa. The other way was most likely through evolution toward functionality after speciation; However, this is only a speculation. For a more complete understanding of the origination of CArG elements, further studies on the phylogenetics of these elements in mammalian populations will be necessary.

This study revealed the sequence patterns and the diversification of the functional CArG elements in both the human and mouse genomes. The results presented here provide a platform for study of the cis-element through sequence analysis and evolutionary methods. A better understanding of the sequence characteristics, main sequence type and evolution of CArG elements will fundamentally improve future cis-element prediction, regulatory determinant pattern detection and analysis of gene expression. Lastly, it would provide insight into how CArG elements impact the SRF regulation of gene expression in processes such as embryonic development, heart disease, and cancer.

References

[1] Averof M, Patel NH. Crustacean appendage evolution associated with changes in Hox gene expression [J]. Nature, 1997, 388(6643): 682-686

[2] Li H, Johnson AD. Evolution of transcription networks—lessons from yeasts [J].Curr Biol, 2010, 20: 746-753

[3] Hahn MW. Detecting natural selection on cis-regulatory DNA [J]. Genetica, 2007,129(1): 7-18

[4] Yokoyama KD,Thorne JL, Wray GA. Coordinated genome-wide modifications within proximal promoter cis-regulatory elements during vertebrate evolution [J]. Genome Biol Evol, 2011,3: 66-74

[5] Treisman R. Identification of a protein-binding site that mediates transcriptional response of the c-fos gene to serum factors [J]. Cell, 1986, 46:567-574

[6] Miano JM. Role of serum response factor in the pathogenesis of disease[J]. Laboratory Investigation, 2010, 90:1274-1284

[7] Kwon CY,Kim KR, Choi HN,et al. The role of serum response factor in hepatocellular carcinoma: implications for disease progression [J]. Int J Oncol. 2010,37(4):837-44

[8] Rorick MM, Wagner GP. The origin of conserved protein domains and amino acid repeats via adaptive competition for control over amino acid residues [J].J Mol Evol, 2010,70:29-43

[9] Wu WW, Shen X, Tao SH. Characteristics of the CArG-SRF binding context in mammalian genomes [J]. Mamm. Genome, 2010, 21(1-2): 104-113

[10] Cooper SJ, Trinklein ND, Nguyen L,et al. Serum response factor binding sites differ in three human cell types [J]. Genome Res, 2007, 17: 136-144

[11] Sun Q, Chen G, Streb JW, et al. 2006. Defining the mammalian CArGome [J]. Genome Res, 2006, 16: 197-207

[12] Balza RJ, Misra RP. Role of the serum response factor in regulating contractile apparatus gene expression and sarcomeric integrity in cardiomyocytes[J]. J Biol Chem, 2006,281:6498-6510

[13] Shen X, Walsh B, Li JJ, et al. The correlations of the function and positional distribution of the cis-elements CArG around the TSS in the genes of Mus musculus[J]. Genome, 2009, 52(3): 217-221

[14] Workman CT, Yin Y, Corcoran DL, et al. enoLOGOS: a versatile web tool for energy normalized sequence logos [J]. Nucleic Acids Res,2005, 33:389-392

[15] Pellegrini L, Tan S, Richmond TJ. Structure of serum response factor core bound to DNA [J]. Nature, 1995, 376:490 - 498

[16] Kleinjan DA, Heyningen van V. Long-range control of gene expression: Emerging mechanisms and disruption in disease [J]. Am J Hum Genet, 2005, 76: 8-32

[17] Spielman RS, Bastone LA, Burdick JT, et al. Common genetic variants account for differences in gene expression among ethnic groups [J]. Nat Genet, 2007, 39:226-231

血清反应因子结合位点在小鼠及人基因组中保守性、多样性及进化的研究*

沈 霞[1] 吴文武[2] 谭从娥[1] 冯居君[1]

(1 陕西中医学院药学院 陕西 西安 712046;2 西北农林科技大学 生命科学学院 生物信息中心 陕西 杨凌 712000)

摘要 **目的**:CArG 元件因其为血清反应因子识别的结合位点近年来备受关注。然而迄今为止尚未见到有关 CArG 元件的序列特征及进化模式的研究。**方法**: 本研究应用生物信息学方法结合遗传学方法对小鼠及人基因组中 CArG 元件的位置分布序列类型、多样性及保守性进行深入研究。**结果**:多样性研究结果显示,CArG 元件的序列在小鼠及人类基因组存在大量的不同类型。但是,小鼠和人基因组中 CArG 元件的主要类型又存在明显差异。同源性分析结果表明人类和小鼠中的 CArG 元件存在两种进化历程,一部分 CArG 元件拥有共同的祖先,一部分是在物种分化以后突变产生的。**结论**: 上述研究结果将为更为深入阐述 SRF 的调控模式奠定理论基础,同时为更清楚的阐释 CArG 元件序列变化对下游基因的表达影响提供理论支持。

关键词:血清因子结合位点;CArG 元件;生物信息学;序列特征;进化

中图分类号:Q95-3 Q-33 **文献标识码**:A **文章编号**:1673-6273(2011)10-1821-06

* **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**:2011-02-03 **接受日期**:2011-02-28)