J Mol Evol (1991) 33:237–240

Journal of Molecular Evolution

Nonrandom CpG Mutations Affect the Synonymous Codon Usage of Moderately GC-Rich Single Copy Actin Genes

Guy Drouin

Department of Biology, University of Ottawa, 30 George Glinsky, Ottawa, Ontario K1N 6N5, Canada

Summary. In species where actin genes exist as single copies, analysis of their synonymous codon usage and of the substitutions occurring between the genes of closely related species shows that there is a positive selection for codons that do not have highly mutable CpG dinucleotides in codon positions 2 and 3 when the GC content of these genes is less than 57%.

Key words: Codon usage — Actin genes — CpG dinucleotides — Nonrandom mutations — 5-Methylcytosine

Introduction

CpG dinucleotides occur at only about 20% of their expected frequency in vertebrate genomes (Russell et al. 1976). This CpG avoidance has been attributed to the propensity of 5-methylcytosine (5-mC) to be deaminated to thymine, and to the fact that 70–80% of the 5-mC present in vertebrate genomes is found in CpG dinucleotides (Bird 1986). The occurrence of such nonrandom mutations has been reported in several instances (e.g., Savatier et al. 1985; Green et al. 1990), but previous studies found that they did not have a significant influence on the usage of synonymous codons of several genes from the genomes of *Escherichia coli, Salmonella typhimurium,* bacteriophage T7, yeast, and human (Li et al. 1985; Hanai and Wada 1988).

Here I report that the high mutation rate of 5-mC present in CpG dinucleotides does affect the synonymous codon usage of single copy actin genes when the overall GC content of these genes is lower than 57%.

Results

The occurrence of particular dinucleotides within a coding region can most easily be observed when it is scored relative to its codon position. A striking feature of the dinucleotide distributions shown in Table 1 is the significant ($P < 0.01$) CpG deficiency when the cytosine occurs in position 2 of codons in the eight genes that have a GC content of less than 57%. There is also a significant CpG deficiency when cytosine occurs in position 1 of codons in *Tetrahymena pyriformis* and *Tetrahymena thermophila.* On the other hand, CpG frequency is not significantly different from random expectation in any of these 10 genes when cytosine occurs in position 3 of codons (with guanine following in position 1 of the next codon).

A high mutability of CpG, when compared with other nucleotides, can be observed when the actin sequences of closely related species are compared. Comparing the genes of closely related species ensures that the pattern of substitutions observed will not be overly affected by multiple substitution events at the same sites. Table 2 shows the analysis of the distribution of nucleotide substitutions among the 16 classes of dinucleotides between the actin genes of two *Tetrahymena* species, and the actin genes of three closely related Ascomycetes fungi (Drouin and Gilbert, unpublished). The observed distribution of single nucleotide substitutions among the 16 classes of dinucleotides is compared to that expected from a random distribution of nucleotide variation (Savatier et al. 1985). Note that in order to assess the effect of a neighboring nucleotide, only those substitutions of the sites where the 3′ neighboring nucleotide was conserved were scored. All three comparisons show the same substitution patterns: (1)

Table 1. Number of dinucleotides relative to codon position in single copy actin genes

Dinucleotide	f(1,2)										f(2,3)			
	Ca	Sc	Kl	Sp	Tt	Tp	Gl	An	Pm	Tl	Ca	Sc	Kl	Sp
AA	27	27	28	30	36	36	35	28	33	28	56+	46+	42+	31+
AC	19	21	21	19	20	22	20	21	26	23	31	39	43	28
AG	20	15	16	5−	19	18	17	6	1−	5−	5	13	17	31
AT	46	46	45	44+	42	42	31	44+	47+	42+	21−	14−	10−	29−
CA	26+	24	28	23	14	14	20	23	16	23	19	16	18	3
CC	18	19	18	19	23+	21	17	19	19	18	17	31	27	30
CG	2	5	3	16	0−	0−	13	18	21+	19	0−	0−	0−	2−
CT	0−	4−	3−	15	16	18	28	21	15	21	57	48	51+	55
GA	48	47	47	49	51	51	51	47	50	49	18	13	15	7
GC	26	26	28	31	26	24	30	25	26	29	0−	0−	1−	8
GG	31	29	28	30	26	26	26	30	28	30	4	4	4	4
GT	23	26	27	22	24	23	30	29	28	26	39	40+	35+	41
TA	12−	14	14	17	19	18	14	14	14	14	6−	4−	6−	2
TC	30	29	29	21	24	26	15	25	19	22	19	40	48	26
TG	8	8	8	9	12	12	12	8	10	8	35+	35+	37+	28
TT	40+	35	37	25	24	24	16	17	22	18	49	32	21	50
% GC	39	44	44	45	45	49	53	55	57	59	39	44	44	45

The number of dinucleotides relative to codon position for each gene. The genes are arranged by order of increasing GC content, which is indicated at the bottom of the table. The notation f(1,2) represents the first and second bases of codons, f(2,3) the second and third bases of codons, and f(3,1) the last base of a codon and the first base of the following codon. The significance of dinucleotide distribution bias was assessed by a χ^2, defined as (observed frequency − random expectation)^2/random expectation. The random expectation was calculated by pX·pY·n, where pX and pY are the proportions, respectively, of base X and Y at a given codon position, and n is the number of codons in the coding region of the gene being analyzed. Values of 6.6 are significant at $P < 0.01$ and are indicated to the right of the dinucleotide numbers; "−" indicates a significant deficiency; "+" indicates a significant overrepresentation. Species abbreviations and references are: An, *Aspergillus nidulans* (Fidel et al. 1988); Ca, *Candida albicans* (Losberger and Ernst 1989); Gl, *Giardia lamblia* (Drouin, unpublished); Kl, *Kluveromyces lactis* (Deshler et al. 1989); Pm, *Phytophthora megasperma* (Dudler 1990); Sc, *Saccharomyces cerevisiae* (Gallwitz and Sures 1980); Sp, *Schizosaccharomyces pombe* (Mertins and Gallwitz 1987); Tl, *Thermomyces lanuginosus* (Wildeman 1988); Tp, *Tetrahymena pyriformis* (Hirono et al. 1987); Tt, *Tetrahymena thermophila* (Cupples and Pearlman 1986).

CpG substitutions occur significantly more often than expected by chance; (2) most, if not all, CpG substitutions from one species are found as TpG in the other species; (3) all substituted CpG dinucleotides are in position 3,1 of codons (not shown).

Discussion

The fact that most, if not all, substitutions occurring in the highly mutable CpG dinucleotides found in position 3,1 of one species are found as TpG dinucleotides in a closely related species suggests that this mutational susceptibility is due to the deamination of 5-mC to thymine, and that 5-mC preferentially occurs in CpG dinucleotides.

The observation that the actin genes of species, which avoid CpG when cytosine is in position 1 and/or 2 of codons, also show high levels of substitutions from CpG to TpG in position 3,1 of codons, suggests that the absence of CpG in position 2,3 of codons (and, to a lesser extent in position 1,2) is the result of the positive selection for codons that do not have highly mutable CpG dinucleotides in these positions. Interestingly, the relatively GC-rich actin genes of *Phytophthora megasperma* (57.4% GC) and *Thermomyces lanuginosus* (59.5% GC) do not show CpG avoidance in any codon position. This observation suggests that GC-rich actin genes might be undermethylated compared to actin genes with a lower GC content and that this relaxes the selection for codons that do not have CpG dinucleotides in positions 1,2 and 2,3. This suggestion is consistent with the observation that CpG discrimination is attenuated at high GC content in the human genome (Hanai and Wada 1988).

The fact that previous studies did not find that nonrandom CpG mutations had a significant influence on synonymous codon usage could be a consequence of the fact that they analyzed a collection of sequences, and that these sequences did not all represent highly expressed genes. Furthermore, these sequences might not all have been under a nonrandom mutation pressure as strong as the one observed here. Selection is likely to be stronger on the codons of highly expressed genes as they are used more often, especially under strong nonrandom mutation pressure.

Table 1. Continued

f(2,3)						f(3,1)									
Tt	Tp	Gl	An	Pm	Tl	Ca	Sc	Kl	Sp	Tt	Tp	Gl	An	Pm	Tl
40+	36+	9	9	13	4	32	30	32	11	21	18	16	4	5	2
35	44	36	41	51	37−	14	12	9	9	9	9	11	4	9	1
24	24	54+	50+	40	55+	34	23	27	16	19	17	25	1	19	4
21−	15−	21	12−	9	18	19	14	13	7	14	13	7	5	1	0
0−	2−	24+	4	8	1	27	27	28	24	41	54	36	47	49	42
43	62+	30	52	31	45	8	15	13	17	20	26	29	35	20	35
1−	0−	9−	5−	37	25	20	40	45	25	38	54	34	56	47	54
49+	29	19	29	14	21	11	27	32	25	36	40	23	35	32	33
17	18+	9	2	6	2	14	20	19	22	19	20	31	28	43	38
9	8−	27	14−	35	20	3	3	6	13	5	7	25	16	35	24
3	3	15	4	12	5−	13	18	18	18	15	14	35	25	38	31
28	27+	17	42+	7	35+	14	11	15	12	8	8	16	13	28	22
5−	1−	17	1	7	0	38	31	30	40	35	25	19	19	9	15
49	61	29	67	32	63	21	22	19	34	19	12	13	26	7	21
19	22	30	23	55	30	61	47	40	73	56	38	43	47	28+	45+
33	23	29	20	18	14	46	34	28	28	21	20	11	11	4	7
45	49	53	55	57	59	39	44	44	45	45	49	53	55	57	59

Table 2. Observed and expected distributions of dinucleotide differences between actin genes

Dinucleotide	*T. pyriformis* (48.9%) vs *T. thermophila* (45.5%)				*C. albicans* (38.9%) vs *K. lactis* (44.1%)				*S. cerevisiae* (43.9%) vs *K. lactis* (44.1%)			
	N_{obs}	S_{obs}	S_{exp}	χ^2	N_{obs}	S_{obs}	S_{exp}	χ^2	N_{obs}	S_{obs}	S_{exp}	χ^2
AA	90	1	5.6	3.7	115	6	12.9	3.7	103	1	8.8	6.9**
AC	75	2	4.7	1.6	64	6	7.2	0.2	72	3	6.1	1.6
AG	59	1	3.7	2.0	59	7	6.6	0.0	51	0	4.4	4.4*
AT	70	2	4.4	1.3	86	12	9.6	0.6	71	5	6.1	0.2
CA	70	13	4.4	16.8**	72	4	8.1	1.9	67	9	5.7	1.9
CC	109	6	6.8	0.1	43	2	4.8	1.6	65	3	5.5	1.1
CG	54	21 (21)	3.4	91.1**	22	4 (3)	2.5	0.9	45	14 (11)	3.8	27.4**
CT	87	7	5.4	0.5	68	1	7.6	5.7*	79	9	6.7	0.8
GA	89	1	5.5	3.7	80	2	9.0	5.4*	80	2	6.8	3.4
GC	39	2	2.4	0.1	29	1	3.2	1.5	29	1	2.5	0.9
GG	43	0	2.7	2.7	48	2	5.4	2.1	51	2	4.3	1.2
GT	58	1	3.6	1.9	76	4	8.5	2.4	77	2	6.6	3.2
TA	44	3	2.7	0.0	56	11	6.3	3.5	49	9	4.2	5.5*
TC	99	2	6.2	2.8	70	11	7.8	1.3	91	5	7.7	0.9
TG	72	3 (3)	4.5	0.5	104	29 (28)	11.6	26.1**	90	18 (16)	7.7	13.8**
TT	67	5	4.2	0.2	135	24	15.1	5.2*	101	13	8.6	2.3
Total	1125	70	70		1127	126	126		1125	96	96	

N_{obs}, the number of observed dinucleotides in the first species; S_{obs}, the number of dinucleotides substituted in the second species; S_{exp}, the expected distribution of substitutions in the second species assuming a random chance of substitution. Expected values were calculated by (total number of substitutions in the second species/total number of dinucleotides in the first species) × observed number of a given dinucleotide in the first species. Observed and expected values were compared using a χ^2 test. A value of 3.8 is significant at $P < 0.05$ (indicated by *) and a value of 6.6 is significant at $P < 0.01$ (indicated by **). The percentages in parentheses are the GC content of each gene, and the numbers in parentheses indicate the number of times CpG was substituted for TpG and vice versa

Acknowledgments. I thank Allan Wilson and anonymous referees for invaluable suggestions on a first draft of this paper. The comments of Robert Dorit and Dipankar Sen were much appreciated. This work was supported in part by a Postdoctoral Scholarship from the National Science and Engineering Research Council of Canada.

References

Bird AP (1986) CpG-rich islands and the function of DNA methylation. Nature 321:209–213

Cupples CG, Pearlman RE (1986) Isolation and characteriza-

tion of the actin gene from *Tetrahymena thermophila*. Proc Natl Acad Sci USA 83:5160–5164

Deshler JO, Larson GP, Rossi JJ (1989) *Kluveromyces lactis* maintains *Saccharomyces cerevisiae* intron-encoded splicing signals. Mol Cell Biol 9:2208–2213

Dudler R (1990) The single-copy actin gene of *Phytophthora megasperma* encodes a protein considerably diverged from any other known actin. Plant Mol Biol 14:415–422

Fidel S, Doonan JH, Morris NR (1988) *Aspergillus nidulans* contains a single actin gene which has unique intron locations and encodes a γ-actin. Gene 70:282–293

Gallwitz D, Sures I (1980) Structure of a split yeast gene: complete nucleotide sequence of the actin gene in *Saccharomyces cerevisiae*. Proc Natl Acad Sci USA 77:2546–2550

Green PM, Montandon AJ, Bentley DR, Ljung R, Marie Nilsson I, Giannelli F (1990) The incidence and distribution of CpG → TpG transitions in the coagulation factor IX gene. A fresh look at mutational hotspots. Nucleic Acids Res 18:3227–3231

Hanai R, Wada A (1988) The effects of guanine and cytosine variation on dinucleotide frequency and amino acid composition in the human genome. J Mol Evol 27:321–325

Hirono M, Endoh H, Okada N, Numata O, Watanabe Y (1987) *Tetrahymena* actin. Cloning and sequencing of the *Tetrahymena* actin gene and identification of its gene product. J Mol Biol 194:181–192

Li W-H, Luo C-C, Wu C-I (1985) Evolution of DNA sequences. In: MacIntyre RJ (ed) Molecular evolutionary genetics. Plenum, New York, pp 1–94

Losberger C, Ernst JF (1989) Sequence of the *Candida albicans* gene encoding actin. Nucleic Acids Res 17:9488

Mertins P, Gallwitz D (1987) A single intronless actin gene in the fission yeast *Schizosaccharomyces pombe*: nucleotide sequence and transcripts found in homologous and heterologous yeast. Nucleic Acids Res 15:7369–7379

Russell GJ, Walker PMB, Elton RA, Subak-Sharpe JH (1976) Doublet frequency analysis of fractionated vertebrate nuclear DNA. J Mol Biol 108:1–23

Savatier P, Trabuchet G, Faure C, Chebloune Y, Gouy M, Verdier G, Nigon VM (1985) Evolution of the primate β-globin gene region. High rate of variation in CpG dinucleotides and in short repeated sequences between man and chimpanzee. J Mol Biol 182:21–29

Wildeman AG (1988) A putative ancestral actin gene present in a thermophilic eukaryote: novel combination of intron positions. Nucleic Acids Res 16:2553–2564

Received December 7, 1990/Accepted March 12, 1991