安徽农业大学学报, 2004, 31(1): 1~5
Journal of Anhui Agricultural University

Synonymous Codon Usage of Both Alternatively and Commonly Spliced Genes in Human Chromosome 1 ①

I: Synonymous Codon Usage Bias Analysis

CHEN Xue-ping[1], WU Yao-ting[2], GUO Jia-min[1], ZHANG Cheng[3], MA Fei[4 *]
(1. College of Economics and Technology, University of Science and Technology of China, Hefei 230052; 2. National Key Biotechnology Laboratory for Tropical Crops, Chinese Academy of Tropical Agricultural Sciences, Haikou 571101; 3. Station of Plant Protection of Hefei City, Hefei 230031; 4. College of Life Science, Xiamen University, Xiamen 361005)

Abstract: It is already clear that alternative splicing has an extremely important role in expanding the protein diversity. Comparative study of the codon usage patterns of alternatively and commonly spliced genes may thereby be necessary. In this paper, the patterns of codon usage bias of two kinds of human genes, alternatively spliced genes and commonly spliced genes, were formulated through analyzing 344 non-redundant protein coding sequences from alternatively spliced genes (188183 codons) and 386 from commonly spliced genes (223116 codons) in human chromosome 1. Overall codon usage data analysis indicated that the alternatively spliced genes showed a stronger codon usage bias than commonly spliced genes. Very extensive heterogeneity of G+C content in silent third codon position (GC3s) was evident among these genes, and GC3s content of alternatively spliced genes was higher than that of commonly spliced genes. G- or C-ending codons were more abundant in alternatively spliced genes than commonly spliced genes in human chromosome 1. The causation of differences created could be explained by pre-mRNA structural characteristics of alternatively spliced genes influencing their codon usage bias.

Key words: alternative splicing; common splicing; codon usage; human
CLC number: Q987 **Document code**: A **Article ID**: 1672-352X(2004)01-0001-05

Alternative splicing has been found in nearly all metazoan organisms as a means for producing functionally diverse polypeptides from a single gene, where variation in mRNA structure may take many different forms[1,2]. The positions of either 5′ or 3′ splice sites can shift to result in longer or shorter exons. Introns that are normally excised can be retained in the mRNA. In addition to above changes in splicing, alterations in transcriptional start site or polyadenylation site may also allow production of multiple mRNAs from a single gene. Exonic splicing enhancers are often found at the 5′ and 3′ ends, and sometimes in the middle of an exon and may regulate the accessibility of different exons to the splicing machinery through the formation of secondary structures[3~5]. Furthermore, secondary structure of RNA can also alter the splicing pattern of human mRNA transcripts[6]. Recent estimates, based on analyses of ESTs, have suggested that the transcripts from 35% ~59% of human genes are alternatively spliced[7~9]. Findings above have strongly suggested an important role for alternative splicing in the formation of biological complexity of human.

The previous studies on codons usage did not distinguish alternative splicing genes from common splicing ones. Whether synonymous codon usage has the difference between them or not, which is yet unclear. Thereby, the main aim of this study is to investigate differences of codon biased usage between alternative and common splicing genes in human chromosome 1. In the present work, the 344 protein-coding sequences (188183

① **Received date**: 2003-05-07
Foundation item: China postdoctoral programs foundation(200211).
Biography: CHEN Xue-ping(1956—), Man, Doctor, Associate professor. * Corresponding author

codons) of alternatively spliced genes and 386 protein-coding sequences (223116 codons) of commonly spliced genes in human chromosome 1 were used to explore codon usage patterns of the human genes, and to identify the differences of codon usage patterns between alternatively spliced genes and commonly spliced genes.

1 Materials and methods

1.1 Complete protein Coding sequences of human chromosome 1

All nuclear protein-coding sequences annotated as deriving from human chromosome 1 were extracted from the GenBank/EMBL/DDBJ DNA sequence database. Duplicate sequences, partial sequences, short sequences (less than 100 codons), sequences predicted, and sequences containing ambiguous codons or multiple stop codons were excluded, and a total dataset of 730 complete Coding Sequences (CDS) was yielded for data analysis. They could be divided into two subsets based on annotation. One subset included 344 CDS of alternatively spliced genes representing 188183 codons. Another subset contained 386 commonly spliced genes with 223116 codons.

1.2 Codon usage bias indices

Nc is the effective number of codons used by a gene, generally used to measure the bias of synonymous codons and independent of amino acid compositions and codon numbers[10]. The values of Nc range from 20 (when one codon is used per amino acid) to 61 (when all the codons are used with equal probability). Nc appears to be a good measure of general codon usage bias. The lower the Nc, the higher the codon bias is.

Relative synonymous codon usage (RSCU) was used to study the overall codon usage variation among the genes. RSCU is defined as the ratio of the observed frequency of codons to the expected frequency if all the synonymous codons for those amino acids are used equally[11]. RSCU values greater than 1.0 indicate that the corresponding codon is more frequently used than expected, whereas the reverse is true for RSCU values less than 1.0.

GC3s is the frequency of use of G+C in synonymously variable third positions of sense codons (i.e., Met, Trp, and 3 termination codons are excluded).

The parameters mentioned above were estimated by using the program CodonW 1.3 (written by John Peden, obtained from ftp://molbiol.ox.ac.uk/Win95.codonW.zip).

2 Results

2.1 Level of codon usage among alternatively and commonly spliced genes

Several measures of the degree of codon bias for a given gene have been developed. Here we used the effective number of codons (Nc) to estimate the level of codon usage of gene[10]. Nc appears to be a good measure of general codon usage bias. The lower the Nc, the higher codon bias is. The 344 protein-coding sequences (188183 codons) of alternatively spliced genes and 386 protein-coding sequences (223116 codons) of commonly spliced genes in human chromosome 1 were employed to analyze codon usage level. The level of codon usage skewness across these genes was represented in Fig. 1. The mean Nc of alternatively spliced genes was somewhat less than that for commonly spliced genes. It indicated that alternatively spliced genes had a greater proportion of very highly preferred genes than commonly spliced genes did. When we considered extreme bias such as an Nc of 40 or less, 12.8% of alternatively spliced genes and 10.9% of commonly spliced genes were in this category. Moreover, commonly spliced genes exhibited a greater variance (SD) in codon usage than alternatively spliced genes. Those conclusions revealed that the level of codon usage in alternatively spliced genes was higher than that in commonly spliced genes.

2.2 Relationship between Nc and GC3s

The heterogeneity of codon usage bias among genes was examined in human chromosome 1. Effective number of codons (Nc) used by a gene and (G + C) percentage at the third synonymous were used to explore

the codon usage variation. For the genes from alternatively spliced genes subset, GC3s value varied from 27.5% to 92.8% with a mean of 64.66% and standard deviation of 15.17%; while in commonly spliced genes subset, GC3s values varied from 24.7% to 94.4% with a mean of 59.97% and standard deviation of 15.67%. The findings indicated that alternatively spliced genes had a rather higher content of G+C in silent the third codon position than commonly spliced genes. Further analyzing the heterogeneity of GC3s showed that there were the mean Nc of 41.45, 51.06, 51.56 of the high G + C range (0.70—0.93; 44.93% of the genes), midterm G + C range (0.45—0.70; 42.03% of the genes), and low G + C range (0.27—0.45; 13.33% of the genes) in alternatively spliced genes subsets respectively; whereas, in commonly spliced genes subsets, there were the mean Nc of 41.11, 51.97, 52.21, from the high G + C range (0.70 ~0.95; 32.12% of the genes), midterm G + C range (0.45—0.70; 48.96% of the genes), and low G + C range (0.24—0.45; 19.17% of the genes) respectively. These investigations proved that the G + C contents of the third codon position of human genes were scattered in the G + C range of 0.24 ~0.95 in the third codon position, and the higher the G + C content of the third codon position (GC3s), the higher level of codon bias.

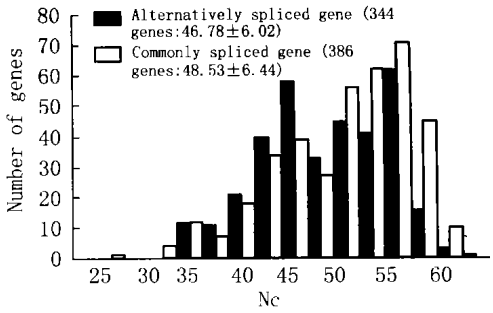

Figure 1　Nc distribution of alternatively spliced genes(■) and commonly spliced genes(□) in chromosome 1. The number of genes for alternative splicing and common splicing and the mean Nc±SD are shown in brackets

Plotting Nc values against GC3s was used to explore the codon usage heterogeneity among the genes from human chromosome 1 (Fig. 2). The Nc plot revealed that there were very similar patterns of codon usage bias between alternatively spliced and commonly spliced genes. It was very clear from Fig. 2 that the most number of points lay in the GC-rich regions, only smaller points in the expected curve. For these genes from both alternatively spliced and commonly spliced ones, however, a majority of sequences had rather lower Nc values than expected, which indicated that these genes had additional codon usage skewness, other than genomic GC composition. This position in the lower part of the plot might imply a pattern of codon usage bias generated by selection for translationally optimal codons. These results suggest that apart from compositional constraints, other trends might influence the overall codon usage variation among these genes of human chromosome 1.

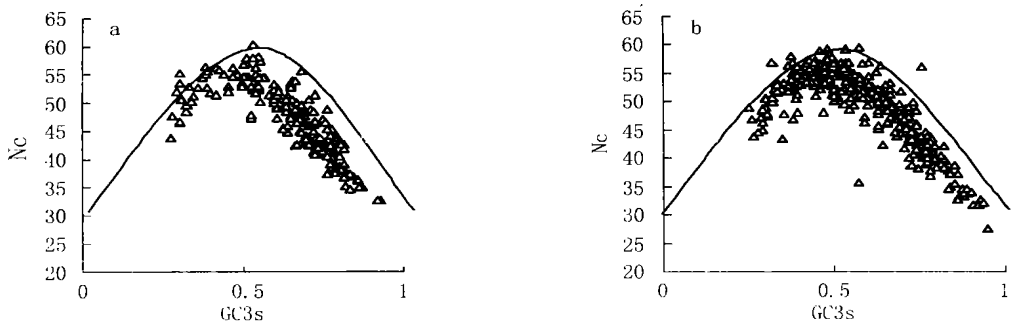

Figure 2　Nc plot of alternatively spliced genes (a) and commonly spliced genes (b) in human chromosome 1. The continuous curve represents the expected curve between GC3s and Nc under random codon usage

2.3　Overall codon usage analysis among alternatively and commonly spliced genes

The codon usage across these genes was outlined in Table 1. It was clear from Table 1 that the most markedly used codons were those ending in either C or G. This result intensively supports that the human genes do prefer those codons ending in either C or G. To further analyze, for each of the six amino acids with two-fold degenerate sets of synonyms ending in U or C, the C-ending codon usage (average RSCU of 1.182)

in alternatively spliced genes was more frequent than that (average RSCU of 1.105) in commonly spliced genes. Similarly, among amino acids with two-fold degenerate sets of synonyms ending in A or G, the G-ending codon (average RSCU of 1.287) in alternatively spliced genes was somewhat higher than those (average RSCU of 1.257) in commonly spliced genes, and in alternatively spliced genes, among amino acids with six-fold degeneracy both C and G ending codons had average RSCU values of 1.328, 1.204, respectively; while in commonly spliced genes, they was 1.268, 1.186, respectively. Between amino acids with four-fold degeneracy, C ending codons in alternatively spliced genes had higher average RSCU values of 1.402, while in commonly spliced genes, it was 1.366; in both alternatively spliced genes and commonly spliced genes, however, the G ending codons had lower average RSCU values than 1 (Table 1). In addition, when we subject RSCU values of ending in G and C to t- tests, the C- ending codons usage has a significant difference($t = 3.5494$, $P = 0.0032 < 0.005$), and the G-ending codons usage also has a significant difference($t = 2.4141$, $P = 0.0327 < 0.05$) between alternatively spliced genes and commonly spliced genes. Those results implied that GC-ending codons are more abundant in alternatively spliced genes than commonly spliced genes in human chromosome 1.

Table 1 Overall codon usage in human chromosome 1 genes

AA	Codon	AS genes		CS genes		AA	Codon	AS genes		CS genes	
		N	RSCU	N	RSCU			N	RSCU	N	RSCU
Phe	UUU	2 602	0.74	3 744	0.91	Ser	UCU	2 683	1.05	3 436	1.12
	UUC	4 450	1.26	4 479	1.09		UCC	3 334	1.30	4 068	1.33
Leu	UUA	1 145	0.37	1 562	0.42		UCA	2 258	0.88	2 741	0.90
	UUG	2 211	0.72	2 845	0.76		UCG	892	0.35	905	0.30
	CUU	2 317	0.75	3 011	0.80	Pro	CCU	3 170	1.14	3 862	1.13
	CUC	3 866	1.26	4 408	1.17		CCC	3 702	1.33	4 592	1.35
	CUA	1 110	0.36	1 688	0.45		CCA	2 815	1.01	3 730	1.09
	CUG	7 767	2.53	9 021	2.40		CCG	1 420	0.51	1 471	0.43
Ile	AUU	2 791	1.01	3 558	1.07	Thr	ACU	2 159	0.82	2 960	0.98
	AUC	4 321	1.56	4 885	1.48		ACC	4 292	1.63	4 376	1.46
	AUA	1 186	0.43	1 489	0.45		ACA	2 819	1.07	3 420	1.14
Met	AUG	4 238	1.00	4 900	1.00		ACG	1 289	0.49	1 273	0.42
Val	GUU	1 922	0.68	2 353	0.70	Ala	GCU	3 421	1.06	4 245	1.08
	GUC	2 850	1.01	3 226	0.96		GCC	5 233	1.62	6 479	1.65
	GUA	1 080	0.38	1 577	0.47		GCA	2 877	0.89	3 537	0.90
	GUG	5 423	1.92	6 355	1.88		GCG	1 373	0.43	1 412	0.36
Tyr	UAU	2 071	0.79	2 712	0.87	Cys	UGU	2 052	0.85	2 242	0.88
	UAC	3 167	1.21	3 545	1.13		UGC	2 802	1.15	2 867	1.12
Ter	UAA	101	0.88	90	0.70	Ter	UGA	174	1.52	204	1.59
	UAG	69	0.60	92	0.72	Trp	UGG	2 873	1.00	2 769	1.00
His	CAU	1 746	0.78	2 213	0.82	Arg	CGU	1 013	0.57	1 029	0.51
	CAC	2 728	1.22	3 177	1.18		CGC	2 143	1.20	2 274	1.12
Gln	CAA	2 028	0.52	2 931	0.53		CGA	1 051	0.59	1 509	0.74
	CAG	5 711	1.48	8 170	1.47		CGG	2 390	1.34	2 601	1.28
Asn	AAU	2 893	0.84	3 807	0.95	Ser	AGU	2 252	0.88	2 755	0.90
	AAC	3 964	1.16	4 233	1.05		AGC	3 965	1.55	4 426	1.45
Lys	AAA	4 027	0.79	5 079	0.84	Arg	AGA	2 154	1.21	2 346	1.16
	AAG	6 187	1.21	6 999	1.16		AGG	1 920	1.08	2 424	1.19
Asp	GAU	3 917	0.91	4 864	0.94	Gly	GGU	1 884	0.58	2 279	0.62
	GAC	4 656	1.09	5 497	1.06		GGC	4 616	1.42	5 167	1.41
Glu	GAA	5 872	0.83	6 861	0.86		GGA	2 946	0.91	3 539	0.96
	GAG	8 226	1.17	9 085	1.14		GGG	3 569	1.10	3 722	1.01

Note: AA, amino acids; N, number of codons; RSCU, cumulative relative synonymous codon usage; AS, alternatively spliced; CS, commonly spliced.

3 Discussion

This is the first attempt to systematically compare the differences of synonymous codon usage bias be-

tween alternatively and commonly spliced genes by analyzing the large data set from human chromosome 1. The results show that the level of codon usage in alternatively spliced genes is relatively higher than that in commonly spliced genes, and alternatively spliced genes have a higher content of G+C in silent third codon position than commonly spliced genes do. In addition, commonly spliced genes have a more heterogeneity than alternatively spliced genes, and the high G + C content of the third codon position demonstrates a higher codon usage preference. All these imply that mutation bias has an influence on synonymous codon usage in human, although other factors such as the sharp heterogeneity of GC3s may also play an important role in shaping codon usage patterns. Overall codon usage analyses indicate that G- and C-ending codons are predominant in alternatively and commonly spliced genes, revealing that the human genes do prefer those codons ending in either C or G, and GC-ending codons are more frequent in alternatively spliced genes than commonly spliced genes in human chromosome 1. We suggest that the pre-mRNA structural characteristics of alternative splicing may strongly influence their codon usage bias, resulting in the differences of alternatively genes from commonly spliced genes in codon usage bias.

References:

[1] Lopez A J. Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation[J]. Ann Rev Genet, 1998, 32: 279~305

[2] Smith C W and Valcarcel J. Alternative pre-mRNA splicing: the logic of combinatorial control[J]. Trends Biochem Sci, 2000, 25: 381~388

[3] Wang Y-C, Selvakumar M, Helfman D M. Alternative Pre-mRNA Splicing[A]. In: Krainer A R (Ed.). Eukaryotic mRNA Processing[C]. Oxford University Press, Oxford, 1997. 242~279

[4] König H, Ponta H, Herrlich P. Coupling of signal transduction to alternative pre-mRNA splicing by a composite splice regulator[J]. EMBO J, 1998, 17: 2904~2913

[5] Muro A F, Iaconcig A, Baralle F E. Regulation of the fibronectin EDA exon alternative splicing. Cooperative role of the exonic enhancer element and the 50 splicing site[J]. FEBS Lett, 1998, 437: 137~141

[6] Mikheeva S, Hakim-Zargar M, Carson D, Jarrell K A. Use of an engineered ribozyme to produce a circular human exon[J]. Nucleic Acids Res, 1997, 25: 5085~5094

[7] Modrek B, Resch A, Grasso C, Lee C. Genome-wide analysis of alternative splicing using human expressed sequence data [J]. Nucleic acids Res, 2001, 29: 2850~2859

[8] Moderk B and Lee C. A genomic view of alternative splicing[J]. Nat Genet, 2002, 30: 13~19

[9] Brett D, Pospisil H, Valcrel J, Reich J, Bork P. Alternative splicing and genome complexity[J]. Nat Genet, 2002, 30: 29~30

[10] Wright F. The 'effective number of codons' used in a gene[J]. Gene, 1990, 87: 23~29

[11] Sharp P M and Li W-H. An evolutionary perspective on synonymous codon usage in unicellular organisms[J]. J Mol Evol, 1986, 24: 28~38

人类 1 号染色体可变剪接与普通剪接基因同义密码子的使用分析 I. 同义密码子偏爱使用分析

陈学平[1], 武耀廷[2], 郭家明[1], 张　成[3], 马　飞[4*]

(1. 中国科学技术大学经济技术学院, 合肥 230052; 2. 中国热带农业科学院热带作物生物技术国家重点实验室, 海口 571101; 3. 合肥市植保站, 合肥 230031; 4. 厦门大学生命科学学院, 厦门 361005)

摘　要: 人类 1 号染色体可变剪接(选择性剪接)基因 344 非冗余蛋白质编码序列(188183 密码子)和普通剪接(非可变剪接)基因的 386 蛋白质编码序列(223116 密码子)被用于研究人类密码子使用偏爱模式。全部密码子使用数据分析表明, 人类可变剪接基因密码子的偏爱水平显著高于普通剪接基因。在人类 1 号染色体基因中, 密码子第三位置的 G+C 含量有很大的异质性(0.24~0.95), 并且可变剪接基因密码子第三位置平均 G+C 含量(64.66%)大于普通剪接基因(59.97%)。Nc 值对 GC3s 图显示密码子偏爱使用除了受核苷酸组成制约外, 其它的因子可能也影响密码子的使用变化。此外, 可变剪接基因中以 G 或 C 结尾的密码子比普通剪接基因出现的频率高。密码子使用的差异可能是由可变剪接基因 pre-mRNA 特有的结构特征和多种剪接模式决定的。

关键词: 可变剪接; 普通剪接; 密码子使用; 人类