Available online at www.sciencedirect.com
SCIENCE DIRECT
Gene 376 (2006) 199–206
GENE
SECTION
EVOLUTIONARY GENOMICS
www.elsevier.com/locate/gene

Nucleotide substitution pattern in rice paralogues: Implication for negative correlation between the synonymous substitution rate and codon usage bias

Xiaoli Shi [a,b,1], Xiyin Wang [a,c,1], Zhe Li [a], Qihui Zhu [a,b], Wen Tang [a], Song Ge [b,*], Jingchu Luo [a,*]

[a] College of Life Sciences, National Laboratory of Plant Genetic Engineering and Protein Engineering, Center of Bioinformatics, Peking University, Beijing 100871, China
[b] Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China
[c] College of Sciences, Hebei Polytechnic University, Tangshan, Hebei 063009, China

Received 9 November 2005; received in revised form 2 March 2006; accepted 10 March 2006
Available online 18 March 2006

Abstract

Understanding the correlation between synonymous substitution rate and GC content is essential to decipher the gene evolution. However, it has been controversial on their relationship. We analyzed the GC content and synonymous substitution rate in 1092 paralogues produced by two large-scale duplication events in the rice genome. According to the GC content at the third codon sites (GC_3), the paralogues were classified into GC_3-rich and GC_3-poor genes. By referring to their outgroup sequences, we inferred the last common ancestor of sister paralogues and, consequently, calculated the average synonymous substitution rate for two gene classes. The results suggest that average synonymous substitution rate is lower in GC_3-rich genes than that in GC_3-poor genes, indicating that the synonymous substitution rate is negatively correlated with GC content in the rice genome. Through characterizing the synonymous nucleotide substitution pattern, we found a strong synonymous nucleotide substitution frequency bias from AT to GC in GC_3-rich genes. This indicates possible limitations of commonly used methods developed to estimate the synonymous substitution rate. Their estimates might produce misleading results on correlation between the synonymous substitution rate and GC content.

Keywords: GC content; Two gene classes; Nucleotide substitution bias; Methodology

1. Introduction

Analysis of synonymous substitution rate in homologues from the same or different evolutionary lineages provides a powerful tool to detect the driving forces and mechanisms of molecular evolution (Yang and Nielsen, 1998) and has been used for the estimation of divergence time of organisms at different taxonomic levels (Bowers et al., 2003). Based on the study on 242 duplicated gene pairs on chromosomes 2 and 4 of *Arabidopsis thaliana*, Zhang et al. (2002) found that synonymous substitution rate among genes varied much wider than that previously reported and suggested an effect of physical location on synonymous substitution rates.

GC content is one of the most important factors to affect synonymous substitution rate and the correlation between GC content and synonymous substitution rate has been extensively examined in many species, such as Drosophila (Sharp and Li, 1989; Moriyama and Hartl, 1993), mammals (Wolfe et al., 1989; Bulmer et al., 1991), enteric bacteria (Smith and Eyre-Walker, 2001) and plants including *Arabidobsis* (Zhang et al.,

Abbreviations: GC_3, GC content at the third codon sites; LCA, last common ancestor; ASR, average synonymous substitution number per third codon site; PNS frequency, partially normalized synonymous nucleotide substitution frequency; WNS frequency, wholly normalized nucleotide substitution frequency.

* Corresponding authors. Luo is to be contacted at Center of Bioinformatics, Peking University, Beijing 100871, P.R. China. Tel.: +86 10 6275 7281; fax: +86 10 6275 9001. Ge, Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, P.R. China. Tel.: +86 10 6283 6097; fax: +86 10 6259 0843.
E-mail addresses: gesong@ibcas.ac.cn (S. Ge), luojc@pku.edu.cn (J. Luo).
[1] These authors contributed equally to this work.

0378-1119/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2006.03.003

2002), Gramineae (Alvarez-Valin et al., 1999) and conifers (Kusumi et al., 2002). However, whether they are positively or negatively correlated or independent of each other, has been in a hot debate. Some researchers proposed that synonymous substitution rate was negatively correlated with GC content in Drosophila (Sharp and Li, 1989; Moriyama and Hartl, 1993) and mammals (Wolfe et al., 1989; Bulmer et al., 1991). Conversely, others reported that the correlation was positive (Smith and Hurst, 1999; Betancourt and Presgraves, 2002) or synonymous substitution rate was independent of GC content (Dunn et al., 2001; Zhang et al., 2002).

As pointed out by Bielawski et al. (2000) that the correlation between GC content and synonymous substitution rate might be methodologically sensitive. To date, different methods have been designed to estimate synonymous substitution rate (Li et al., 1985; Nei and Gojobori, 1986; Goldman and Yang, 1994; Muse and Gaut, 1994; Ina, 1995; Yang and Nielsen, 2000) and commonly used methods could be classified into two categories (Nei and Kumar, 2000), the approximate (evolutionary pathway) method and the maximum-likelihood method. A commonly used approximate method was proposed by Nei and Gojobori (1986), referred to NG, involving three steps: (i) counting the synonymous sites (*S*), (ii) counting the synonymous differences between two sequences and (iii) performing multiple-hit correction. An improved approximate method proposed by Yang and Nielsen (2000), referred to YN, takes into account two important evolutionary features of DNA sequence: transition/transversion rate bias and codon usage bias. Based on an explicit model of codon substitution, Goldman and Yang (1994) developed a maximum-likelihood method (ML), which is flexible since knowledge such as transition/transversion rate and codon usage bias can be easily integrated into the model. After evaluating the relationship between synonymous substitution rate and GC content of 83 orthologs of Drosophila, Dunn et al. (2001) found that synonymous substitution rate was positively correlated with the GC content when the NG method was used but the rate was independent of the GC content when the ML method was used.

Negative correlation between synonymous substitution rate and GC content might be explained by natural selection (Bielawski et al., 2000; Smith and Eyre-Walker, 2001; Urrutia and Hurst, 2001). Although synonymous substitution was regarded as neutral by most researchers previously (Miller et al., 2004), evidences indicated that selection acts on synonymous substitution in numerous species, such as bacteria, Drosophila (Urrutia and Hurst, 2001) and mammals (Chamary and Hurst, 2005). Synonymous substitution relates to selection for translational accuracy, which is irrelevant of the change of amino acids at the protein level. If selection acts, it would increase GC content and the sequence under greater selective pressure would evolve more slowly (Eyre-Walker and Hurst, 2001). However, positive correlation between synonymous substitution rate and GC content has been supported by Francino and Ochman (1999) who found that interspecific GC variation in two pseudogenes was led by the effect of differential GC mutation pressure. Therefore, it has been a subject of controversy regarding the correlation between synonymous substitution rate and GC content.

In our previous study in rice, we found 10 large duplicated regions in the rice genome produced by two duplications, and identified GC_3-rich genes and GC_3-poor genes in these 10 regions (Wang et al., 2005). This provided us with simultaneously duplicated paralogues having unambiguous phylogenetic patterns. Here, we estimated the synonymous substitution rate of GC_3-rich and GC_3-poor genes based on the synonymous substitution number between the sister paralogues and their last common ancestor (LCA). We analyzed the relationship between synonymous substitution rate and GC content, as well as synonymous nucleotide substitution bias of these two gene classes. Then, we characterized the synonymous nucleotide substitution pattern and explored whether the synonymous substitution was biased and whether the codon/nucleotide frequencies were constant. Our analysis indicated possible limitations of the approximate and maximum-likelihood methods when estimating the synonymous substitution rate.

2. Materials and methods

2.1. Materials

The rice genome sequences were obtained from RISE (http://rise.genomics.org.cn) and the rice genes were predicted by BGF (http://btn.genomics.org.cn/rice). There are 10 duplicated blocks in rice genome with one produced by a segmental duplication event ~5 mya, after the divergence of rice from other cereals, and the other nine by a whole-genome duplication ~70 mya, before the divergence of cereals but after their divergence from Arabidopsis (Wang et al., 2005). We retrieved 1738 pairs of paralogues in colinearity and searched their orthologs in maize and Arabidopsis, respectively. The maize and Arabidopsis sequences were retrieved from GenBank. The paralogues which have cDNA (Kikuchi et al., 2003) (http://cdna01.dna.affrc.go.jp/cDNA/) and outgroup sequences were further analyzed. This resulted in a dataset containing 1092 triplets of paralogueous genes and corresponding outgroups.

2.2. Calculation of average synonymous substitution rate

We estimated the synonymous substitution rate using two approximate methods (Nei and Gojobori, 1986; Yang and Nielsen, 2000) and one maximum-likelihood method (Goldman and Yang, 1994) encapsulated in PAML (Yang, 1997).

To explore the relationship between the synonymous substitution rate and GC content, we computed the average synonymous substitution number per third codon site (ASR) for the GC_3-rich and GC_3-poor genes. We aligned two paralogues and their outgroups by codons using ClustalW (Thompson et al., 1994). To restrict our analysis to codon triplets without involving nonsynonymous differences, we considered nucleotide substitutions at the third site of fourfold degenerate LCA codons only, which were inferred by

adopting two different approaches, the maximum-parsimony and maximum-likelihood. In the parsimonious analysis, the LCA state was revealed by the outgroup codon site if it was identical to the corresponding nucleotide in at least one of the sister paralogues. In the likelihood analysis, the possible LCA states were inferred using codeml in PAML (Yang, 1997). We assumed that the observed number of nucleotide differences between the LCA and the sister paralogues was exactly the nucleotide substitution number occurred after the duplication event (Bazykin et al., 2004). The synonymous nucleotide substitutions at the third codon sites between LCA and the sister paralogues were counted. We calculated ASR by dividing the total number of the observed synonymous substitutions with the number of triplets with four-fold degenerate LCA codons. To assess the confidence of the ASRs in two gene classes, we performed bootstrap tests by calculating the ASRs for artificial sequences constructed by picking codon triplets from the paralogues at random with replacement for the GC_3-rich and GC_3-poor genes, respectively.

2.3. Calculation of synonymous nucleotide substitution frequency

We characterized the synonymous nucleotide substitution frequencies between different nucleotides at the third codon sites using the maximum-parsimony and maximum-likelihood methods. With the 108,924 matched synonymous codons in 1092 gene triplets, we counted the number of synonymous substitutions between the LCA and the paralogues. Then, we computed the frequency of each type of synonymous substitution ($b_i \rightarrow b_j$), where b_i stands for A, T, C or G, by dividing the observed number of this type of substitution (on ($b_i \rightarrow b_j$)) with the number of the LCA codons (pn ($b_i \rightarrow b_j$)), at the third sites of which the synonymous substitution of this type could occur. For example, when computing the frequency of the synonymous substitution A→T, we counted the number of synonymous substitution A→T and the number of the codons ending with A in LCA, which could be synonymously substituted by T. In this case, 11 types of codons (ATA, GTA, CCA, ACA, GCA, GGA, CTA, TCA, CGA, TTA and AGA) were involved. We estimated the frequency of synonymous nucleotide substitution as follows:

$$f(b_i \rightarrow b_j) = \frac{\mathrm{on}(b_i \rightarrow b_j)}{\mathrm{pn}(b_i \rightarrow b_j)} \qquad (i,j = 1 \sim 4,\ i \neq j)$$

The substitution frequency was normalized by the sum of the frequencies of the nucleotide substitutions starting from the same nucleotide, referred to partially normalized synonymous nucleotide substitution frequency (PNS frequency):

$$f^*(b_i \rightarrow b_j) = \frac{f(b_i \rightarrow b_j)}{\sum_{k=1(k \neq i)}^{4} f(b_i \rightarrow b_k)}, \qquad (i,j,k = 1 \sim 4,\ i \neq j)$$

The PNS frequencies of two classes of genes were computed individually. To display the substitution bias among all types of substitutions, we further normalized the substitution frequencies by the sum of the total substitution frequencies, referred to wholly normalized nucleotide substitution frequency (WNS frequency):

$$f^{**}(b_i \rightarrow b_j) = \frac{f(b_i \rightarrow b_j)}{\sum_{m=1}^{4} \sum_{n=1(n \neq m)}^{4} f(b_m \rightarrow b_n)}, \ (i,j,m,n = 1 \sim 4,\ i \neq j)$$

The WNS frequencies of genes in both recent and ancient blocks were computed individually.

3. Results

3.1. GC content

We characterized the GC content of 1092 paralogues and found a distinct bimodal distribution with peak positions at 0.46 and 0.69 (Fig. 1A). We also analyzed the GC content at each codon site separately and found a bimodal pattern for the GC content at the third codon site (Fig. 1D) and a unimodal pattern for the GC content at the first and second codon sites (Fig. 1B

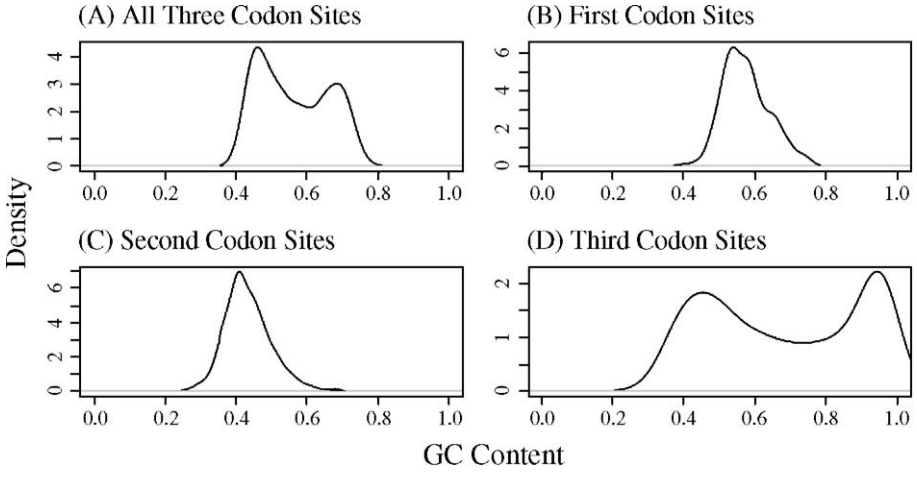

Fig. 1. GC content at three codon sites. The distributions of GC content at the first and second codon sites are unimodal and distributions at the third and all three codon sites are bimodal.

Table 1
Correlation between Ks computed by three different methods and GC_3

Blocks	GC_3	Number of genes	Rho[a], *P*-value		
			NG	ML	YN
Recent block (block 1)	0.27–1	62	0.0947, 0.4745	0.4110, 0.0014	0.4560, 0.0003
Ancient blocks (blocks 2–10)	0.27–1	1030	−0.6295, <2.2e−16	0.4457, <2.2e−16	0.4988, <2.2e−16

[a] Rho: Spearman correlation coefficient.

and C). Therefore, the bimodal pattern was mainly attributed to the GC content at the third codon site. According to GC_3, ranging from 0.27 to nearly 1, we divided the paralogues into two classes at $GC_3 = 0.75$ by performing hierachical average clustering analysis and obtained 469 GC_3-rich and 623 GC_3-poor genes.

3.2. Nucleotide substitution rate

We computed three different estimates of the synonymous substitution rate (Ks) using the NG, ML and YN methods. The correlations between Ks calculated by these three methods and GC_3 of the paralogues were inconsistent. We found that Ks was negatively correlated with GC_3 according to the NG method but positively correlated with GC_3 when either ML or YN method was applied (Table 1).

In order to study the correlation between synonymous substitution rate and GC_3 content, we calculated the Ks values for paralogues in the ancient blocks with the NG and ML methods. The NG estimates showed a clear bimodal distribution of Ks when all genes were used (Fig. 2A), but a unimodal distribution when either GC_3-rich or GC_3-poor genes were used (Fig. 2B and C). The peaks of two unimodal distributions corresponded perfectly to the two peaks of the bimodal distribution, with the peak position of the GC_3-rich genes (0.40) much lower than that of the GC_3-poor genes (0.85). This suggested that the Ks was negatively correlated with the GC_3 content while using the NG method. The ML estimates of Ks for all paralogues displayed a unimodal distribution with a long upper tail (Fig. 2D) and the peak position of unimodal distribution for the GC_3-rich genes (1.55) was higher than that for GC_3-poor genes (0.85) (Fig. 2E and 2F), indicating a positive correlation between Ks and GC_3.

We further calculated the correlation coefficient between Ks and GC_3 (Table 1). When the NG method was used, a negative correlation was found between Ks and GC_3 for the ancient paralogues, while Ks was independent of GC_3 for the recent paralogues. However, when the ML method was used, Ks was positively correlated with GC_3 for the ancient and recent paralogues. The YN method showed a similar result to the ML

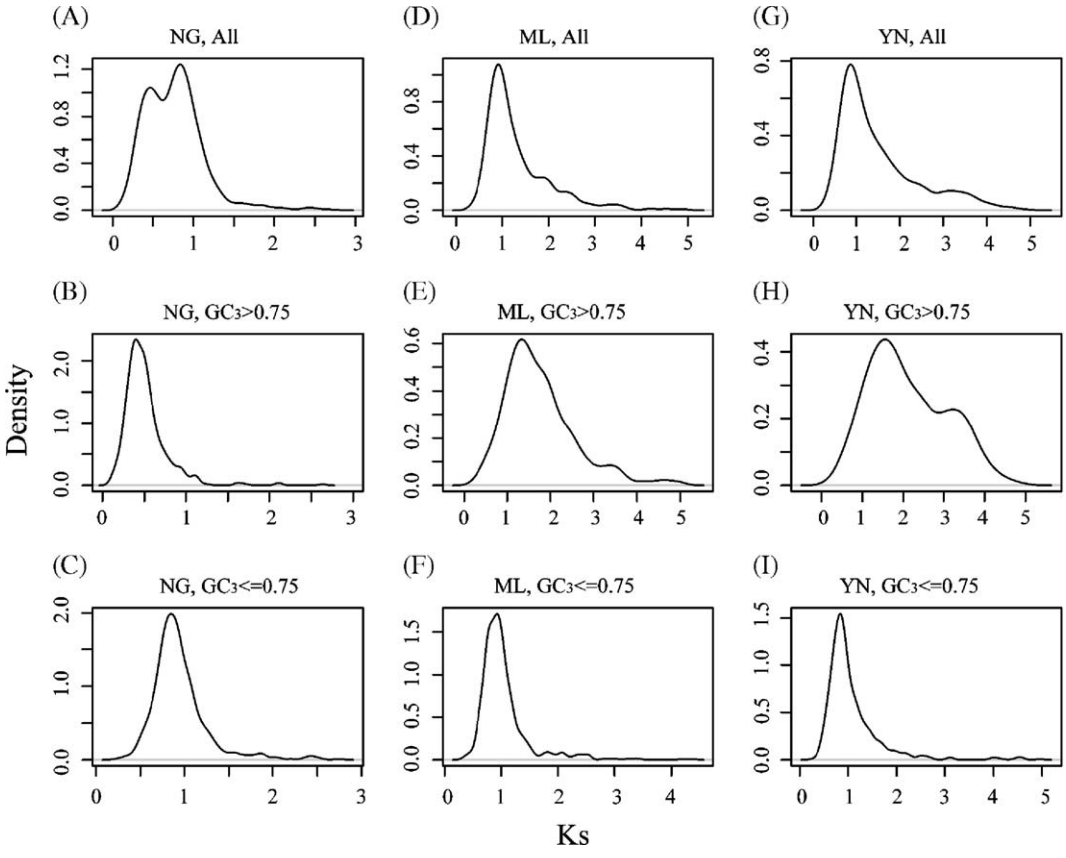

Fig. 2. Distributions of Ks estimated by three methods on homologues including all ancient paralogues, GC_3-rich and GC_3-poor ancient paralogues. (A–C) Distributions of Ks calculated using the NG method; (D–F) distributions of Ks calculated using the ML method; (G–I) distributions of Ks calculated using the YN method.

method (Fig. 2G–I, Table 1), which was supported by computer simulation (Yang and Nielsen, 2000).

3.3. Average synonymous substitution rate

By referring to the LCA states estimated by maximum-parsimony and maximum-likelihood methods, we computed the average synonymous substitution rate (ASR) for the two gene classes in ancient blocks. Based on the LCA states inferred by the maximum-parsimony method, the estimates of the ASR were 0.557 and 0.654 for GC_3-rich and GC_3-poor genes, respectively, implying a negative correlation between synonymous substitution rate and GC_3. The maximum-likelihood analysis came to the same conclusion though the corresponding ASR estimates, 0.605 and 0.711, were a little larger, which was caused by involving the multiple substitution sites in the maximum-likelihood analysis but not in the maximum-parsimony analysis. Using the bootstrap tests (1000 repetitions), we assessed the significance of the negative correlation between synonymous substitution rate and GC_3. For both methods, the estimations of ASR of artificial GC_3-rich sequences were constantly smaller than those of artificial GC_3-poor sequences, indicating a significant negative correlation between synonymous substitution rate and GC_3, consistent with the NG method. We considered that the positive correlation proposed by the ML and YN methods was a methodological artifact. We did not compute the ASRs of the recent duplicated genes for the number of nucleotide differences between the paralogues was too small to accurately estimate ASRs for two gene classes.

3.4. Synonymous nucleotide substitution bias

To explore the discrepancy of the correlation between Ks and GC_3, we computed the WNS frequencies for both recent and ancient paralogues and PNS frequencies for two gene classes. The synonymous nucleotide substitution patterns estimated by the parsimony and likelihood methods are shown in Table 2. The calculated WNS frequencies for the recent and ancient paralogues by the parsimony method were similar to those obtained by the likelihood method (*T*-test *P*-value > 0.827). We analyzed substitution patterns estimated by the parsimony method hereafter.

We found that there were biased nucleotide substitutions frequencies and the biases were consistent in the recently and anciently duplicated paralogues. For the recent paralogues, the WNS frequencies of AT↔GC were 0.465 and 0.245 in two directions, and the corresponding frequencies for ancient paralogues were 0.402 and 0.264. Frequency of AT→GC is the total substitution frequencies from A or T to C or G and frequency of GC→AT is vice versa.

The excess of AT→GC substitution frequency was observed in paralogues duplicated at different times and the biases were at the same level (t=0.958, *P*-value=0.409); therefore, we would discuss the substitution frequencies in two gene classes regardless of their duplication times. We computed WNS and PNS frequencies for GC_3-rich and GC_3-poor genes in 10 duplicated regions (Table 2). According to the WNS frequencies, the substitutions frequencies were balanced in both directions of AT↔GC (0.378 vs. 0.366) in GC_3-poor genes (t=0.2932, *P*-value=0.7885). By contrast, in GC_3-rich genes, there was a significant excess of AT→GC (0.640) over GC→AT (0.090) (t=−5.045, *P*-value=0.0150). Therefore, in GC_3-rich genes, synonymous substitution frequencies AT→GC were considerably preferred to GC→AT. The PNS frequencies were very different among the substitutions starting from the same nucleotide. In GC_3-poor genes, the transition frequencies were about two to three folds of the transversion frequencies. For example, the substitution frequency of A→G was 0.543, while the frequencies of A→C and A→T were 0.185 and 0.272, respectively. The higher frequencies of transitions than transversions has been widely noted (Brown et al., 1982; Gojobori et al., 1982; Curtis and Clegg, 1984; Wakeley, 1996; Yang and Yoder, 1999). However, we found a different pattern of synonymous nucleotide substitution frequency bias in the GC_3-rich genes. For substitutions at the third site of codons ending with G and C, the transversions C→G and G→C were extremely preferred, about three to five folds of the corresponding transition of C→T and G→A, e.g., the substitution frequency of transversion C→G (0.650) was almost three times greater than that of transition C→T (0.226).

Table 2
WNS frequency and PNS frequency of paralogues

	A	T	C	G	A	T	C	G
Maximum-parsimony								
PNS frequency, GC_3 > 0.75					*PNS frequency, $GC_3 \le 0.75$*			
A	–	0.109	0.334	0.557	–	0.272	0.185	0.543
T	0.071	–	0.646	0.283	0.280	–	0.537	0.183
C	0.124	0.226	–	0.650	0.194	0.585	–	0.221
G	0.156	0.106	0.738	–	0.545	0.210	0.245	–
WNS frequency, GC_3 > 0.75					*WNS frequency, $GC_3 \le 0.75$*			
A	–	0.038	0.117	0.193	–	0.075	0.051	0.150
T	0.025	–	0.228	0.100	0.069	–	0.132	0.045
C	0.017	0.031	–	0.089	0.049	0.148	–	0.056
G	0.025	0.017	0.118	–	0.122	0.047	0.055	–
WNS frequency, recent block					WNS *frequency, ancient blocks*			
A	–	0.044	0.058	0.131	–	0.073	0.061	0.160
T	0.068	–	0.177	0.099	0.066	–	0.147	0.052
C	0.033	0.090	–	0.071	0.036	0.100	–	0.085
G	0.086	0.036	0.107	–	0.091	0.037	0.092	–
Maximum-likelihood								
WNS frequency, recent block					WNS *frequency, ancient blocks*			
A	–	0.071	0.075	0.156	–	0.083	0.083	0.128
T	0.053	–	0.125	0.076	0.064	–	0.121	0.075
C	0.031	0.068	–	0.081	0.037	0.069	–	0.100
G	0.079	0.063	0.122	–	0.065	0.056	0.118	–

4. Discussion

4.1. Synonymous substitution rate is negatively correlated with codon usage bias

Estimating relationship between the synonymous substitution rate and codon usage bias is important for understanding

molecular evolution. The nature of this relationship has been studied in many species, including Drosophila, mammals and enteric bacteria using different methods. Similar to the rice genes with the GC_3 ranging from 27% to nearly 100%, Drosophila also has wide range in GC_3 from 28% to 93% (Dunn et al., 2001). In this study, we found a negative correlation between synonymous substitution rate and GC_3, in agreement with some previous studies on Drosophila (Sharp and Li, 1989; Moriyama and Hartl, 1993; Betancourt and Presgraves, 2002; Bierne and Eyre-Walker, 2003). However, positive correlation (Smith and Hurst, 1999; Bielawski et al., 2000; Betancourt and Presgraves, 2002) or non-correlation (Dunn et al., 2001) has also been reported in Drosophila. Based on a computer simulation, Bielawski et al. (2000) observed that the correlation might be sensitive to methodology, i.e., a positive correlation was often reached when the ML method was used, whereas a negative correlation was obtained by the NG method. They further indicated that the negative correlation resulted from the NG method was incorrect, due to the ignorance of transition/transversion bias and codon frequency bias (Dunn et al., 2001). However, Bierne and Eyre-Walker (2003) argued that the positive correlation or non-correlation, calculated by the ML method, was misleading because this method relied on a mutational-opportunity definition of a site. They obtained a negative correlation between the synonymous substitution rate and codon bias in Drosophila while using a physical definition of a site.

In our study, the ASRs in two gene classes of rice support a negative correlation between the synonymous substitution rate and codon usage bias measured by GC_3 (Bierne and Eyre-Walker, 2003). We retrieved 1030 ancient simultaneously duplicated paralogues from the rice genome. Using these paralogues and their outgroups, we inferred the LCA of sister paralogues and calculated the ASRs using the four-fold degenerate codons in two gene classes. This approach is free of restriction of many assumptions adopted in the NG and ML methods, such as the equal nucleotide frequencies, the equal nucleotide substitution frequencies and the equilibrium distribution of codons. This makes it possible to obtain a more accurate estimate for the evolutionary rate, especially for datasets violating above assumptions. As a result, we come to a conclusion that there is a negative correlation between the synonymous substitution rate and GC_3 of rice duplicated paralogues, suggesting a greater selective pressure on the silent sites of genes with high codon bias (Duret and Mouchiroud, 1999; Dunn et al., 2001).

4.2. Limitations of the NG method

The prerequisite assumptions of NG method, equal nucleotide frequencies and random nucleotide substitution (Nei and Gojobori, 1986), are unrealistic for some datasets. In our study, highly biased synonymous codon usage was observed in 1092 rice duplicates whose GC_3 ranges from 27% to ~100%. According to WNS frequencies, there is significant excess of AT→GC over GC→AT, indicating an extreme synonymous nucleotide substitution frequency bias in the GC_3-rich genes. It is evident that the assumption of the NG method is not reconciled with the characteristics of our dataset.

It was proposed that the major source of bias in Ks arose from the estimation of synonymous sites, referred to S (Dunn et al., 2001). Therefore, we quantitatively analyze the bias of the estimates of S. While counting synonymous sites for each codon, referred to s, the NG method assumes single and constant substitution rate, i.e., 1/3. We consider that the empirical PNS frequencies reflect the synonymous substitution pattern more objectively. We compute two S estimates adopting equal substitution frequencies and PNS frequencies, respectively, and then calculate their difference. Our computation is restricted to the third codon sites where most synonymous substitutions occurred (Nei and Gojobori, 1986). Different types of codons make different contributions to the estimation of S. For zero-fold degenerate codons, all substitutions are non-synonymous and thus they do not contribute to s. For two-fold degenerate codons, all synonymous substitutions at the third codon site are transitions. For four-fold degenerate codons, the estimate of s is 1, no matter what nucleotide substitution frequencies are used. Therefore, we exclude them from our computation. For six-fold degenerate codons, the six synonymous codons for each group could be divided into two subgroups, one consisting of two-fold degenerate codons and the other consisting of four-fold degenerate codons. For three-fold degenerate codons, two types of synonymous transversions, A↔T and A↔C, could occur at the third codon site. To simplify the calculation, we neglect three-fold degenerate codons, which account for only 4% of the total codons in the dataset. Therefore, only synonymous transitions that occurred at the third codon sites are considered while computing the difference of S:

$$\begin{aligned}\Delta S = {} & NP(\mathrm{G})[F_{\mathrm{PNS}}(\mathrm{G}\rightarrow\mathrm{A}) - F_{\mathrm{NG}}(\mathrm{G}\rightarrow\mathrm{A})] \\ & + NP(\mathrm{C})[F_{\mathrm{PNS}}(\mathrm{C}\rightarrow\mathrm{T}) - F_{\mathrm{NG}}(\mathrm{C}\rightarrow\mathrm{T})] \\ & + NP(\mathrm{A})[F_{\mathrm{PNS}}(\mathrm{A}\rightarrow\mathrm{G}) - F_{\mathrm{NG}}(\mathrm{A}\rightarrow\mathrm{G})] \\ & + NP(\mathrm{T})[F_{\mathrm{PNS}}(\mathrm{T}\rightarrow\mathrm{C}) - F_{\mathrm{NG}}(\mathrm{T}\rightarrow\mathrm{C})]\end{aligned}$$

where N denotes the number of codons in the observed sequences; $P(b)$ denotes proportion of nucleotide b at the third site of all codons except for zero-, three- and four-fold degenerate codons; $F_{\mathrm{PNS}}(b_i \rightarrow b_j)$ denotes PNS frequency from b_i to b_j; and $F_{\mathrm{NG}}(b_i \rightarrow b_j)$ denotes substitution rate from b_i to b_j used in NG method and equals to 1/3. For GC_3-rich genes and GC_3-poor genes, we obtain different ΔS:

$$\begin{aligned}\Delta S_{\mathrm{GC_3-poor}} = {} & N[0.211{*}P(G) + 0.251{*}P(C) \\ & + 0.210{*}P(A) + 0.204{*}P(T)]\end{aligned}$$

$$\begin{aligned}\Delta S_{\mathrm{GC_3-rich}} = {} & N[-0.177{*}P(G) - 0.107{*}P(C) \\ & + 0.224{*}P(A) + 0.313{*}P(T)]\end{aligned}$$

The ΔS of GC_3-poor genes, which account for 57% of all collinear paralogues, is always positive, no matter what proportions of nucleotides are. For GC_3-rich genes, the ΔS can be negative or positive. Using the average base frequencies of GC_3-rich paralogues, we compute the $\Delta S_{\mathrm{GC_3}}$-rich and obtain

a negative result. Therefore, for GC_3-poor paralogues, the NG method underestimates *S* and overestimates synonymous substitution rate, whereas for major GC_3-rich paralogues, the NG method overestimates *S* and underestimates synonymous substitution rate. Based on the Ks distribution of the NG method, the positions of the two peaks corresponding to GC_3-rich genes and GC_3-poor genes are 0.40 and 0.85, whereas the ASRs of the two gene classes are 0.56 and 0.65, using LCA estimated by the parsimony method, 0.61 and 0.71 using LCA estimated by the likelihood method. Although these measurements reveal negative correlation between the synonymous substitution rate and GC_3, the correlation obtained using ASR is considerably weaker than that obtained using the NG estimates.

4.3. Limitations of the ML method

On the basis of the above analysis, the actual peak positions of distribution of synonymous substitution rate of GC_3-rich and GC_3-poor genes should be between 0.40 and 0.85. However, using the ML method, the positions of the two peaks are 1.55 and 0.85, respectively. These suggest that the ML method overestimate the synonymous substitution rate, especially for the GC_3-rich genes. Unlike the NG method, the ML method has the transition/transversion rate bias and different codon frequencies under consideration (Yang and Nielsen, 2000). However, it failed to accommodate the complexity of synonymous nucleotide substitution pattern as indicated by PNS and WNS frequencies, especially in the GC_3-rich genes.

The biased estimation of the ML method can arise from the unrealistic definition of the substitution rate from any codon *i* to codon *j* (q_{ij}, $i \neq j$), which depends on three factors, transition/transversion rate ratio (κ), nonsynonymous/synonymous rate ratio and frequency of codon *j* (π). For example, given a four-fold degenerate codon TCG to which TCC, TCA and TCT might synonymously transit, three substitution rates are π_{TCG}, $\kappa\pi_{TCG}$ and π_{TCG} using the ML method. Based on PNS frequency of GC_3-rich genes, however, the corresponding codon substitution frequencies are 0.650, 0.557 and 0.283, respectively. The ML method assumes that substitution rates of TCC→TCG and TCT→TCG are equal, but the PNS frequencies show that the former could be more than two folds of the latter. In another word, the probability of codon *i* changing to codon *j* is affected by both the starting and ending states instead of the latter only. The PNS and WNS frequencies clearly indicate that neither nucleotide/codon frequencies nor transition/transversion rate bias could fully represent the complexity in codon substitution.

In summary, we found a negative correlation between synonymous substitution rate and GC content, although this correlation is weaker than that obtained using the NG method and is contrary to that obtained by the ML method. We also proposed that nucleotide substitution patterns adopted in NG and ML method could not reflect the actual nucleotide substitution bias in rice, which might be the major factor of improper predictions produced by these two methods. Carels and Bernardi (2000) tested homologous genes in maize, rice and barley, and found similar compositional distribution of genes in three Gramineae species. Therefore, many other cereals might hold similar synonymous substitution pattern as found in rice although additional evidence is needed from other grasses. Consequently, it should be cautious when Ks of homologues with strong codon bias is used to study various evolutionary issues, including the correlation between synonymous substitution rate and codon usage bias, the dating of evolutionary events, etc., especially for genes in Gramineae.

One important unsolved problem is how to explain the observed negative correlation between the synonymous substitution rate and GC content. Piganeau et al. (2002) analyzed the relationship between substitution rate and GC content under different models and concluded that a positive correlation between Ks and GC content is expected under mutation bias model. However, our study in the rice genome revealed a negative correlation between Ks and GC content and the relationship might occur under both mutation bias and selection models introduced by Piganeau et al. (2002). Therefore, the above models could not be used to determine whether mutation bias or selection produces the observation of negative correlation in the rice genome and further theoretical and empirical investigations are needed to determine the biological mechanisms of the negative correlation.

Acknowledgements

This study is supported by the National Key Basic Research Program of China (2003CB715900), the National Natural Science Foundation of China (90408015, 30121003) and the grants from the China High-Tech program.

References

Alvarez-Valin, F., Jabbari, K., Carels, N., Bernardi, G., 1999. Synonymous and nonsynonymous substitutions in genes from Gramineae: intragenic correlations. J. Mol. Evol. 49, 330–342.

Bazykin, G.A., Kondrashov, F.A., Ogurtsov, A.Y., Sunyaev, S., Kondrashov, A.S., 2004. Positive selection at sites of multiple amino acid replacements since rat–mouse divergence. Nature 429, 558–562.

Betancourt, A.J., Presgraves, D.C., 2002. Linkage limits the power of natural selection in Drosophila. Proc. Natl. Acad. Sci. U. S. A. 99, 13616–13620.

Bielawski, J.P., Dunn, K.A., Yang, Z., 2000. Rates of nucleotide substitution and mammalian nuclear gene evolution. Approximate and maximum-likelihood methods lead to different conclusions. Genetics 156, 1299–1308.

Bierne, N., Eyre-Walker, A., 2003. The problem of counting sites in the estimation of the synonymous and nonsynonymous substitution rates: implications for the correlation between the synonymous substitution rate and codon usage bias. Genetics 165, 1587–1597.

Bowers, J.E., Chapman, B.A., Rong, J., Paterson, A.H., 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature 422, 433–438.

Brown, W.M., Prager, E.M., Wang, A., Wilson, A.C., 1982. Mitochondrial DNA sequences of primates: tempo and mode of evolution. J. Mol. Evol. 18, 225–239.

Bulmer, M., Wolfe, K.H., Sharp, P.M., 1991. Synonymous nucleotide substitution rates in mammalian genes: implications for the molecular clock and the relationship of mammalian orders. Proc. Natl. Acad. Sci. U.S.A 88, 5974–5978.

Carels, N., Bernardi, G., 2000. Two classes of genes in plants. Genetics 154, 1819–1825.

Chamary, J., Hurst, L.D., 2005. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. Genome Biol. 6, R75.

Curtis, S.E., Clegg, M.T., 1984. Molecular evolution of chloroplast DNA sequences. Mol. Biol. Evol. 1, 291–301.

Dunn, K.A., Bielawski, J.P., Yang, Z., 2001. Substitution rates in Drosophila nuclear genes: implications for translational selection. Genetics 157, 295–305.

Duret, L., Mouchiroud, D., 1999. Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. Proc. Natl. Acad. Sci. U. S. A. 96, 4482–4487.

Eyre-Walker, A., Hurst, L.D., 2001. The evolution of isochores. Nat. Rev. Genet. 2, 549–555.

Francino, M.P., Ochman, H., 1999. Isochores result from mutation not selection. Nature 400, 30–31.

Gojobori, T., Ishii, K., Nei, M., 1982. Estimation of average number of nucleotide substitutions when the rate of substitution varies with nucleotide. J. Mol. Evol. 18, 414–423.

Goldman, N., Yang, Z., 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. 11, 725–736.

Ina, Y., 1995. New methods for estimating the numbers of synonymous and nonsynonymous substitutions. J. Mol. Evol. 40, 190–226.

Kikuchi, S., et al., 2003. Collection, mapping, and annotation of over 28,000 cDNA clones from japonica rice. Science 301, 376–379.

Kusumi, J., Tsumura, Y., Yoshimaru, H., Tachida, H., 2002. Molecular evolution of nuclear genes in Cupressacea, a group of conifer trees. Mol. Biol. Evol. 19, 736–747.

Li, W.H., Wu, C.I., Luo, C.C., 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol. Biol. Evol. 2, 150–174.

Miller, W., Makova, K.D., Nekrutenko, A., Hardison, R.C., 2004. Comparative genomics. Genomics Hum. Genet. 5, 15–56.

Moriyama, E.N., Hartl, D.L., 1993. Codon usage bias and base composition of nuclear genes in Drosophila. Genetics 134, 847–858.

Muse, S.V., Gaut, B.S., 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol. Biol. Evol. 11, 715–724.

Nei, M., Gojobori, T., 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 3, 418–426.

Nei, M., Kumar, S., 2000. Molecular Evolution and Phylogenetics. Oxford University Press, Inc.

Piganeau, G., Mouchiroud, D., Duret, L., Gautier, C., 2002. Expected relationship between the silent substitution rate and the GC content: implications for the evolution of isochores. J. Mol. Evol. 54, 129–133.

Sharp, P.M., Li, W.H., 1989. On the rate of DNA sequence evolution in Drosophila. J. Mol. Evol. 28, 398–402.

Smith, N.G., Eyre-Walker, A., 2001. Nucleotide substitution rate estimation in enterobacteria: approximate and maximum-likelihood methods lead to similar conclusions. Mol. Biol. Evol. 18, 2124–2126.

Smith, N.G., Hurst, L.D., 1999. The effect of tandem substitutions on the correlation between synonymous and nonsynonymous rates in rodents. Genetics 153, 1395–1402.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673–4680.

Urrutia, A.O., Hurst, L.D., 2001. Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. Genetics 159, 1191–1199.

Wakeley, J., 1996. The variance of pairwise nucleotide differences in two populations with migration. Theor. Popul. Biol. 49, 39–57.

Wang, X., Shi, X., Hao, B., Ge, S., Luo, J., 2005. Duplication and DNA segmental loss in the rice genome: implications for diploidization. New Phytol. 165, 937–946.

Wolfe, K.H., Sharp, P.M., Li, W.H., 1989. Mutation rates differ among regions of the mammalian genome. Nature 337, 283–285.

Yang, Z., 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13, 555–556.

Yang, Z., Nielsen, R., 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J. Mol. Evol. 46, 409–418.

Yang, Z., Nielsen, R., 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol. Biol. Evol. 17, 32–43.

Yang, Z., Yoder, A.D., 1999. Estimation of the transition/transversion rate bias and species sampling. J. Mol. Evol. 48, 274–283.

Zhang, L., Vision, T.J., Gaut, B.S., 2002. Patterns of nucleotide substitution among simultaneously duplicated gene pairs in *Arabidopsis thaliana*. Mol. Biol. Evol. 19, 1464–1473.