J Mol Evol (1999) 49:330–342

JOURNAL OF MOLECULAR EVOLUTION

Synonymous and Nonsynonymous Substitutions in Genes from *Gramineae*: Intragenic Correlations

Fernando Alvarez-Valin,[1,3] Kamel Jabbari,[1] Nicolas Carels,[1,2] Giorgio Bernardi[1,2]

[1] Laboratoire de Génétique Moléculaire, Institut Jacques Monod, 2 Place Jussieu, F-75005 Paris, France
[2] Stazione Zoologica Anton Dohrn, Laboratorio de Evoluzione Molecolare, Villa Comunale I, 80121 Napoli, Italy
[3] Sección Biomatemática, Facultad de Ciencias, Universidad de la República, Montevideo, Uruguay

Received: 2 July 1998 / Accepted: 18 April 1999

Abstract. In this work, we have investigated the relationships between synonymous and nonsynonymous rates and base composition in coding sequences from *Gramineae* to analyze the factors underlying the variation in substitutional rates. We have shown that in these genes the rates of nucleotide divergence, both synonymous and nonsynonymous, are, to some extent, dependent on each other and on the base composition. In the first place, the variation in nonsynonymous rate is related to the GC level at the second codon position (the higher the GC_2 level, the higher the amino acid replacement rate). The correlation is especially strong with T_2, the coefficients being significant in the three data sets analyzed. This correlation between nonsynonymous rate and base composition at the second codon position is also detectable at the intragenic level, which implies that the factors that tend to increase the intergenic variance in nonsynonymous rates also affect the intragenic variance. On the other hand, we have shown that the synonymous rate is strongly correlated with the GC_3 level. This correlation is observed both across genes and at the intragenic level. Similarly, the nonsynonymous rate is also affected at the intragenic level by GC_3 level, like the silent rate. In fact, synonymous and nonsynonymous rates exhibit a parallel behavior in relation to GC_3 level, indicating that the intragenic patterns of both silent and amino acid divergence rates are influenced in a similar way by the intragenic variation of GC_3. This result, taken together with the fact that the number of genes displaying intragenic correlation coefficients between synonymous and nonsynonymous rates is not very high, but higher than random expectation (in the three data sets analyzed), strongly suggests that the processes of silent and amino acid replacement divergence are, at least in part, driven by common evolutionary forces in genes from *Gramineae*.

Key words: Nucleotide substitutions — Nonsynonymous substitutions — Monocots — Base composition

Introduction

Rates of synonymous and nonsynonymous substitutions have been shown to vary among genes of both animals and bacteria (Li et al. 1985; Wolfe and Sharp 1993; Bernardi et al. 1993). The variability in nonsynonymous rates has been attributed to different intensities of negative selection acting towards maintaining amino acids. The larger the proportion of amino acids under functional constraint, the lower the substitution rate. On the other hand, several factors have been postulated to explain the variation in synonymous rates: variation in the rate and pattern of mutation among different regions of the genome (Wolfe et al. 1989), base composition (Ticher and Graur 1989), and, in the case of enterobacterial genes, selection for codon usage (Sharp and Li 1987).

Correspondence to: G. Bernardi at Stazione Zoologica Anton Dohrn; *e-mail:* bernardi@alpha.szn.it

Table 1a. Homologous genes between maize and rice[a]

								Intragenic correlation			
Gene product	Maize	Rice	Length	GC_3 (average)	GC_2 (average)	Syn. dist.	Non-syn. dist.	*r*(GC_2–NSd)	*r*(GC_3–NSd)	*r*(GC_3–Sd)	*r*(NSd–Sd)
1. Chlorophyll a/b binding protein	ZMCABM7	RICLHCP1	265	0.97	0.49	0.36	0.07	0.62*	0.45	−0.09	0.54
2. Cyclophilin (*CyP*)	MZECYP	RICCYP2G	172	0.97	0.45	0.38	0.12	−0.06	−0.27	−0.43	−0.04
3. Lactate dehydrogenase	ZMLACDEHG	RICLDH	352	0.96	0.47	0.46	0.10	0.47	−0.76**	−0.74**	0.62*
4. Phospholipid transfer protein	MZEPLTP	OSLPTPRA	117	0.96	0.64	0.43	0.18	—	—	—	—
5. Embryogenic abscisic acid-inducible gene	ZMEACI	OSEMP1G	91	0.96	0.52	0.31	0.14	—	—	—	—
6. Ferredoxin (Fd) isoprotein pFD5	MZEFD5	RICFERR	134	0.95	0.46	0.54	0.19	—	—	—	—
7. Pyruvate decarboxylase	ZMPDCMRNA	OSU07339	602	0.94	0.44	0.34	0.06	0.43	0.06	−0.8****	−0.16
8. Histone H3 (H3C4)	MZEH3C4	OSHIS311	136	0.94	0.49	0.37	0.02	—	—	—	—
9. Rubisco, small subunit	MZEPCSSU	RICRUBPC1	169	0.94	0.45	0.37	0.16	−0.12	0.16	−0.12	−0.43
10. Acidic class I chitinase	MZECHITC	RICCHT1	316	0.94	0.58	0.52	0.25	−0.02	−0.04	−0.75*	0.26
11. Heat shock protein 17.2	ZMHSP172	OSLMWHSP	150	0.92	0.38	0.42	0.05	0.72	0.33	−0.88*	−0.33
12. Dehydrin (dhn3)	HVDHN3	OSRAB21	155	0.91	0.53	0.43	0.20	0.62	0.35	−0.94***	−0.26
13. Amyloplast-specific transit protein	MZEWAXY	OSWAXY	603	0.87	0.44	0.50	0.10	0.59**	0.25	−0.53*	−0.06
14. α-Amylase	MZEALAM	OSRAMY3B	437	0.86	0.43	0.62	0.07	−0.28	−0.25	−0.86***	0.25
15. Metallothionein	ZMMETALL	OSU18404	74	0.86	0.64	0.53	0.16	—	—	—	—
16. Knotted-1 (*Kn-1*)	ZMKN1	RICOSH1	350	0.83	0.41	0.42	0.09	0.68*	0.26	−0.64*	−0.07
17. Viviparous-1, transcript. act.	MZEREGPRO	RICOSVP1	679	0.83	0.54	0.52	0.20	0.51*	−0.18	−0.36	0.52*
18. Glutamine synthetase	MZEGS1B	OSSIGS28	356	0.82	0.46	0.56	0.03	0.08	−0.18	−0.79*	0.27
19. Alcohol dehydrogenase 2 (*Adh2-N*)	ZMADH2NR	RICADH2A	369	0.79	0.41	0.53	0.07	0.33	−0.64*	−0.53	0.32
20. Fructose bisphos. aldolase	ZMALDOAR	RICCYTALD	355	0.79	0.42	0.55	0.05	0.12	0.34	−0.30	−0.30
21. Proliferating cell nuclear antigen	ZMPCNAR	OSPCNAGEN	263	0.76	0.35	0.60	0.03	0.16	−0.55	−0.73*	0.24
22. Cystatin I.	MZECYS	RICCPI	102	0.73	0.29	0.47	0.22	—	—	—	—
23. Sus1.	MZESUS1	ORRSS2	816	0.70	0.36	0.60	0.03	−0.08	0.00	0.18	0.21
24. ZAG1 (homeotic gene)	MZEZAG1A	RICOSMAB3A	232	0.70	0.35	0.78	0.20	−0.06	−0.76*	−0.23	0.27
25. Calmodulin	ZMCAM2	RICCALMODL	149	0.69	0.28	1.21	0.01	—	—	—	—
26. Manganese super-oxide dismutase (SOD-3)	MZESOD3A	RICRMSO	231	0.67	0.41	0.54	0.07	0.81*	0.43	−0.22	−0.10
27. β-6 tubulin	MZEBTUB6A	RICBTA	446	0.67	0.40	0.57	0.00	−0.63*	0.40	0.16	0.16
28. QM protein	ZMU06108	OSSG12G	218	0.65	0.46	0.50	0.07	−0.04	−0.38	−0.46	0.87**
29. Alcohol dehydrogenase 1 (*Adh1-1F*)	ZMADH1FA	OSACLDE1	376	0.64	0.41	0.54	0.03	0.24	−0.37	−0.47	−0.14
30. β-Amylase	ZMBAMYL	RICAMYBA	488	0.62	0.37	0.49	0.09	−0.22	−0.02	−0.40	−0.21
31. α-Tubulin	MZEBLMEXSW	OSTUBA1M	450	0.62	0.40	0.45	0.01	0.09	−0.03	−0.30	0.06
32. Regulatory protein GF14-12	MZEREGP	RICS94A	247	0.62	0.37	1.01	0.10	0.55	0.14	0.51	0.16
33. Enolase	ZMENOLA	OSU09450	446	0.56	0.39	1.18	0.08	−0.02	0.18	−0.20	−0.49

Table 1a. Continued

								Intragenic correlation			
Gene product	Maize	Rice	Length	GC_3 (average)	GC_2 (average)	Syn. dist.	Non-syn. dist.	$r(GC_2–NSd)$	$r(GC_3–NSd)$	$r(GC_3–Sd)$	$r(NSd–Sd)$
34. ATP synthase β subunit	ZMATP2MT	RICATPB	550	0.55	0.44	0.49	0.02	0.72**	0.53*	0.14	0.18
35. NADP-dependent malic enzyme (*Me1*)	MZENDMEX	RICME6	631	0.55	0.43	0.64	0.11	0.61**	0.81****	−0.04	0.04
36. Cdc2 kinase	MZEKINAA	OSRCDC21	294	0.54	0.35	0.59	0.03	0.14	0.12	−0.46	−0.18
37. Catalase-1 isoenzyme	ZMCAT1	RICPOSCATB	492	0.54	0.39	0.68	0.03	0.01	0.25	−0.05	0.70**
38. Adenine nucleotide translocator	ZMANT2MU	RICATADPT	381	0.53	0.44	0.57	0.05	0.29	0.14	−0..22	−0.07
39. Triosephosphate isomerase	MZETPI	RICRIC	253	0.48	0.42	0.62	0.06	0.15	−0.06	0.01	0.62*
40. Starch branching enzyme II	MZEGLUCTRN	RICBCE3	798	0.39	0.41	0.67	0.10	0.59*	0.74**	0.19	0.04
Probability								1.8×10^{-5}	9.1×10^{-4}	1.6×10^{-8}	0.0128

[a] The length of genes is given in codons. $r(GC_2–NSd)$, $r(GC_3–NSd)$, $r(GC_3–Sd)$, and $r(Sd–NSd)$ are the intragenic correlation coefficients between the variables included in the parentheses, where NSd is the profile of nonsynonymous distance, Sd is the profile of synonymous distance, and GC_2 and GC_3 are the profiles of GC content at the second and third codon positions. The superscript asterisks indicate the statistical significance of the correlation coefficients: * significant at the 5% level; ** significant at the 1% level; *** significant at the 0.1% level; **** significant at the 0.01% level. The probability values in the bottom row indicate the probability of obtaining this group of correlation coefficients by chance (see Materials and Methods). The dashes indicate that the genes were excluded in the window analysis because of their small size.

Moreover, consistent evidence has been presented indicating that in mammalian genes silent positions could be under selective constraints. In fact, Mouchiroud et al. (1995) found that the synonymous rates are gene specific, since independent processes of divergence (i.e., human–calf vs rat–mouse) produce strongly correlated distances. Further evidence supporting this point comes from the fact that synonymous and nonsynonymous evolutionary rates are correlated (Graur 1985; Li et al. 1985; Wolfe and Sharp 1993; Mouchiroud et al. 1995; Ohta and Ina 1995). While this correlation was attributed to doublet mutation (Wolfe and Sharp 1993), more recent investigations ruled out this possibility and suggested that similar constraints acting on synonymous and nonsynonymous mutations could be responsible for the correlations (Mouchiroud et al. 1995). The latter hypothesis has been gaining support from several lines of evidence. Cacciò et al. (1995) reported that the degree of synonymous divergence in duet codons is correlated with that of quartet codons. Moreover, Zoubak et al. (1995) demonstrated that conserved positions (especially in GC_3-rich genes) exhibit a synonymous base level that differs substantially from what would be expected in sequences subjected to a random substitution process. Finally, Alvarez-Valin et al. (1998) found that the intragenic spatial pattern of synonymous substitutions is correlated with that of nonsynonymous substitutions and with GC_3 level. In the present work we have investigated the relationship between synonymous and nonsynonymous rates and base composition of coding sequences from *Gramineae.*

Materials and Methods

The analysis was carried out on three sets of homologous nuclear coding sequences from *Gramineae.* The first set consisted of 40 sequences from maize (*Zea mays*) and rice (*Oriza sativa*) (Table 1a); the second set, of 47 sequences from maize and either wheat (*Triticum aestivum*) or barley (*Hordeum vulgare*) (Table 1b). The last set included 32 sequences from rice and either wheat or barley, which were considered as a single taxon (Table 1c). This is justified by the fact that, for any particular gene, the distance between wheat and barley is much smaller (by far) than that between wheat and rice (or maize) or that between barley and rice (or maize). In addition, phylogenetic studies indicate that wheat and barley are very close to each other, forming a monophyletic clade compared with maize or rice (Duval and Morton, 1996). It is worth noting that not all genes displaying a certain degree of homology between the species analyzed were included in the data sets. Specifically, we restricted the data sets to those homologous genes exhibiting at least 60% amino acid identity. The exclusion of loosely related homologous genes gives a certain level of confidence that the genes used in this work are not paralogous.

Nucleotide distances (synonymous and nonsynonymous) were estimated by Nei and Gojobori's method (1986). The GC level as well as the level of each individual base was determined in the three codon positions for each gene.

The present analysis was also extended to the intragenic level. For this purpose, the variation in substitution rates along the genes was determined by using a sliding window. Two profiles were obtained, representing synonymous and nonsynonymous divergence, where each

Table 1b. Homologous genes between maize and wheat–barley[a]

								Intragenic correlation			
Gene product	Maize	Wheat–barley	Length	GC_3 (average)	GC_2 (average)	Syn. dist.	Nonsyn. dist.	*r*(GC_2–NSd)	*r*(GC_3–NSd)	*r*(GC_3–Sd)	*r*(Sd–NSd)
1. Heat shock protein, 18 kDa	ZMHSP18K2	TAHSP173	157	0.99	0.41	0.29	0.11	0.58	−0.32	−0.85*	0.60
2. Hstone H3	MZEH3A	TAHI02	136	0.99	0.49	0.26	0.01	—	—	—	—
3. Flavanone 3-β-hydroxylase (*fht*)	ZMFHT	HVFL3DOX	369	0.97	0.43	0.27	0.13	−0.45	0.13	−0.20	−0.45
4. Ferredoxin I (Fd) isoprotein, pFD1′	MZEFD1P	TAPETFFE	143	0.96	0.48	0.48	0.24	—	—	—	—
5. Transmembrane protein	ZMTRAPRO	HVEMIP	287	0.96	0.47	0.28	0.07	0.65*	0.14	−0.64*	−0.27
6. Chalcone synthase	ZMC2CS	X58339	398	0.96	0.43	0.60	0.07	−0.06	−0.15	−0.11	−0.05
7. Histone H4	MZEH4C14	TAHI01	103	0.95	0.48	0.25	0.00	—	—	—	—
8. Lactate dehydrogenase	ZMLACDEHG	BLYLDHA13	353	0.95	0.48	0.54	0.13	0.52*	−0.77**	−0.72**	0.88***
9. Ubiquitin fusion protein (UBF9)	MZEUBFA	BLYMUB1	155	0.94	0.34	0.29	0.02	0.15	−0.16	0.11	0.05
10. UDP glucose flavonoid glycosyl-transferase	ZMBZMCC	HVBRNZ1H	452	0.94	0.56	0.44	0.16	−0.12	−0.23	0.16	0.00
11. Chlorophyll a/b binding protein (*CAB-m7* gene)	ZMCABM7	HVCAB2	261	0.94	0.47	0.72	0.10	0.78**	0.05	0.11	0.24
12. MFS18	ZMFS18	HVSTRPR	121	0.93	0.63	0.49	0.31	—	—	—	—
13. α-Amylase	MZEALAM	BLYAMY1A	425	0.92	0.42	0.51	0.22	−0.36	−0.43	−0.86***	0.72**
14. Amyloplast-specific transit protein	MZEWAXY	HVWAXYG	600	0.92	0.43	0.50	0.11	0.71****	0.01	−0.62**	0.38
15. Acidic class I chitinase	MZECHITC	TACHIG	313	0.91	0.58	0.63	0.22	0.05	−0.48	−0.88***	0.19
16. Rubisco small subunit	MZEPCSSU	WHTRUBIAB	169	0.90	0.46	0.59	0.17	0.13	0.39	0.00	−0.70
17. Heat shock protein 26 (HSP26)	MZEHSP26X	TAHSP266	231	0.90	0.47	0.51	0.12	0.67*	−0.12	−0.79**	−0.04
18. Histone H2A	ZMUO8225	WHTPH2AD	138	0.90	0.46	0.53	0.11	—	—	—	—
19. Embryogenic abscisic acid-inducible gene	ZMEACI	HVLEAB191	91	0.89	0.52	0.37	0.12	—	—	—	—
20. Metallothionein	ZMMETALL	WHTWALI1A	72	0.89	0.60	0.44	0.23	—	—	—	—
21. Chlorophyll a/b binding protein (*Cab-1* gene)	ZMCAB1	HVLHBC	261	0.88	0.46	0.64	0.24	0.81*	0.44	−0.82**	−0.37
22. Phospholipid transfer protein	MZEPLTP	HVU18127	115	0.88	0.62	0.65	0.35	—	—	—	—
23. H2B histone (gH2B4)	SMH2B4A	WHTPH2B12C	121	0.87	0.38	0.37	0.09	—	—	—	—
24. Heat shock protein 17.2	ZMHSP172	HVHSP17A	150	0.86	0.39	0.98	0.17	0.77*	−0.01	−0.70	−0.51
25. Pathogenesis-related protein	ZMPRMS	HVPATHRP1	164	0.86	0.49	0.71	0.24	−0.57	−0.66	−0.79*	0.90**
26. Dehydrin (*dhn3*)	HVDHN3	HVDHN17	146	0.86	0.51	0.58	0.15	—	—	—	—
27. Lipase (LIP)	MZELIPASE	HVPAF93	248	0.86	0.38	0.67	0.22	0.19	−0.04	−0.36	−0.50
28. Chlorophyll a/b binding protein (*Cab-m9* gene)	ZMLHCABB	WHTCAB	264	0.85	0.48	0.83	0.08	0.51	−0.36	−0.44	−0.33

Table 1b. Continued

								Intragenic correlation			
Gene product	Maize	Wheat–barley	Length	GC_3 (average)	GC_2 (average)	Syn. dist.	Non-syn. dist.	r(GC_2–NSd)	r(GC_3–NSd)	r(GC_3–Sd)	r(Sd–NSd)
29. Knotted-1 (*Kn-1*) gene	ZMKN1	HVKNOX3	349	0.84	0.42	0.53	0.10	0.69*	0.28	−0.58	0.08
30. Glutamine synthetase	MZEGS1A	HVCGSA	356	0.83	0.46	0.46	0.06	0.12	−0.10	−0.70*	0.28
31. Calmodulin	ZMCAM2	BLYCAMA	149	0.79	0.28	0.70	0.01	−0.81*	0.66	−0.40	−0.84*
32. Adh2-N alcohol dehydrogenase 2	ZMADH2NR	HVADH3	379	0.78	0.41	0.64	0.06	−0.51	0.16	−0.49	0.52
33. Cysteine proteinase, clone CCP2	MZECYPA	HVLEU	359	0.74	0.47	0.57	0.09	−0.18	0.34	−0.71*	−0.10
34. *Ssus1* gene	MZESUS1	HVRNASS	816	0.69	0.36	0.86	0.06	−0.04	0.02	0.05	0.33
35. Protein disulfide isomerase (pdi)	MZEPDI	BLYPDIA	508	0.68	0.36	0.62	0.11	0.52	0.37	−0.26	0.29
36. *GapC2* gene	ZMGAPC2	HVGADPH	337	0.66	0.40	0.49	0.03	−0.16	0.29	0.01	0.33
37. Regulatory protein GF14-12	MZEREGP	HV1433PH	248	0.66	0.37	1.02	0.11	0.57	0.27	0.26	0.12
38. Alcohol dehydrogenase (*Adh1-S*)	MZEADH1SA	HVADH1	379	0.65	0.42	0.66	0.03	−0.16	−0.06	−0.11	0.35
39. ß-Amylase	ZMBAMYL	BLYBAA	488	0.62	0.38	0.65	0.13	−0.45	0.02	−0.15	−0.08
40. Mit. ATP synthase ß subunit	ZMATP2MT	TAATP2	552	0.58	0.43	0.44	0.05	0.61*	0.73***	0.03	0.20
41. ADP–glucose pyrophosphorylase	ZMADPGLPP	HVBEPL	517	0.58	0.43	1.01	0.10	0.26	0.65**	0.18	0.08
42. Catalase-1 isoenzyme	ZMCAT1	HVU20777	492	0.57	0.39	0.73	0.05	0.49	0.53	0.09	−0.02
43. Porin	ZMPOR1	TAVDAC3	275	0.57	0.41	0.67	0.08	0.29	0.50	−0.32	−0.29
44. Cysteine synthase	ZMCSOATL	WHTCYS1	324	0.55	0.43	0.59	0.06	0.16	−0.54	−0.41	0.56
45. Acyl carrier protein	ZMACPAA	BLYACL3	121	0.54	0.41	0.73	0.18	—	—	—	—
46. Dihydrodipicolinate synthase	ZMDHPS	WHTDHDPD26	376	0.51	0.45	0.97	0.15	0.81**	0.56	0.46	0.16
47. TATA-binding protein	MZETBPA	TATFIIDMR	200	0.46	0.38	1.01	0.03	0.06	0.88**	−0.51	−0.55
Probability								7.3410^{-7}	3.6×10^{-4}	4.24×10^{-8}	4.4×10^{-3}

[a] See Table 1a, footnote a.

point corresponds to the pairwise distance (estimated, as for the whole gene, using Nei and Gojobori's method) for each window. Similarly, the variation in GC level (for each codon position) along the gene was obtained by using a sliding window. It should be noted that the window size cannot be excessively small (fewer than 20 codons), otherwise distance estimates are subject to large stochastic errors. Yet it cannot be too large either, otherwise the number of points to be compared becomes too small (especially in short genes). It is worth clarifying that the number of points is never very large due to the fact that only nonoverlapping windows were used. For this reason, those genes having fewer than 150 codons were not considered in this part of the analysis. Furthermore, the window sizes were somewhat proportional to the sizes of the genes. For genes between 150 and 200 codons long, the window size used was 20 codons. For genes having between 201 and 300 codons, the window size was 25 codons, and for genes larger than 300 codons the window size was 30 codons. This proportionality is for the sake of reducing as much as possible the two sources of sampling variance. In long genes it is possible to use larger window sizes, thus rendering more reliable within-window distance estimations, even if this reduces the total number of points to be compared. In shorter genes, in contrast, it is not possible to use larger window sizes. For instance, if in genes less than 200 codons long the profile was obtained using a window size of 30 codons, the number of independent sampling points would be between 5 and 6 (3 or 4 degrees of freedom).

Table 1c. Homologous genes between rice and wheat–barley[a]

								Intragenic correlation			
Gene product	Rice	Wheat–barley	Length	GC_3 (average)	GC_2 (average)	Syn. dist.	Non-syn. dist.	r(GC_2–NSd)	r(GC_3–NSd)	r(GC_3–Sd)	r(Sd–NSd)
1. γ-Tip	RICYK333	HVGTIPP	250	0.98	0.50	0.26	0.07	−0.07	−0.61	−0.25	−0.12
2. Histone H3	OSHIS311	TAHI02	136	0.98	0.49	0.28	0.02	—	—	—	—
3. Chloroplast transit peptide	OSGOS5G	HVPSAH	142	0.97	0.53	0.46	0.08	—	—	—	—
4. Heat shock protein, 16.9 kD	OSLMWHSP	TAHSPLW	150	0.96	0.37	0.32	0.10	0.52	0.55	0.34	0.72
5. *Emp1* gene	OSEMP1G	TAEMAAA	93	0.96	0.52	0.32	0.07	—	—	—	—
6. Ferredoxin	RICFERR	TAPETFFE	136	0.96	0.44	0.41	0.15	—	—	—	—
7. Lactate dehydrogenase	RICLDH	BLYLDHA13	352	0.95	0.47	0.53	0.08	0.02	0.19	0.28	−0.23
8. Type I light-harvesting chlorophyll a/b	RICLHCP1	HVCAB2	261	0.95	0.48	0.66	0.11	0.88***	−0.38	−0.53	0.48
9. Lipoxygenase L-2	OSLRNA	BLYLOXA	855	0.95	0.41	0.40	0.17	0.34	−0.31	−0.53**	0.38*
10. Endochitinase (*Cht-2* gene)	RICCHT2	BLYCHI33A	332	0.95	0.59	0.53	0.22	−0.25	−0.53	−0.41	0.84**
11. ß-*O*-Glucanase	OSGNS1	HVBDG	334	0.95	0.49	0.40	0.12	0.40	0.13	−0.42	0.02
12. Lectin	RICLECTIN	BLYLEC	212	0.94	0.63	0.38	0.27	−0.28	−0.45	−0.44	0.88***
13. Endochitinase (*Cht-1* gene)	RICCHT1	TACHIG	320	0.93	0.57	0.50	0.17	0.48	−0.30	−0.69*	0.11
14. Thaumatin-like protein	OSTHLP	TATHAU	169	0.90	0.59	0.52	0.29	0.26	−0.65	−0.71*	0.59
15. α-Amylase (amy2A)	RICAMY2A	BLYAMY2	435	0.89	0.44	0.54	0.12	−0.08	0.20	−0.30	0.30
16. Water stress-inducible protein	OSRAB21	HVDHN17	152	0.88	0.54	0.57	0.19	0.49	−0.21	−0.56	0.86***
17. Chloroplast carbonic anhydrase	OSU08404	BLYCA	262	0.88	0.47	0.40	0.13	0.78**	−0.07	−0.69*	0.68*
18. Homeobox protein (*OSH1* gene)	RICOSH1	HVKNOX3	355	0.88	0.41	0.37	0.08	0.59*	0.26	−0.86***	0.05
19. Type II light-harvesting chlorophyll a/b	RICLHCP2	HVLHBC	258	0.87	0.45	0.76	0.22	0.59	0.19	0.11	0.12
20. Glycogen (starch) synthetase	OSWAXY	HVWAXYG	602	0.86	0.42	0.62	0.10	0.65	0.06	−0.61**	0.34
21. Peroxidase	RICPERX	BLYPRX	312	0.83	0.50	0.61	0.17	0.09	−0.16	−0.79**	0.55
22. Metallothionein-like protein	OSU18404	HVIDS1	74	0.80	0.59	0.50	0.25	—	—	—	—
23. Cytosolic glutamine synthetase	OSSIGS28	HVCGSA	356	0.77	0.45	0.33	0.06	0.21	0.03	−0.21	0.47
24. Calmodulin	RICCALMODU	BLYCAMA	149	0.71	0.28	0.79	0.00	—	—	—	—
25. Sucrose–UDP glucosyltransferase	ORRSS2	HVRNASS	816	0.64	0.36	0.77	0.06	0.02	0.11	0.00	0.28
26. Alcohol dehydrogenase 1	OSACLDE1	HVADH1	376	0.64	0.41	0.52	0.04	−0.02	0.13	−0.11	−0.14
27. Kinase C inhibitor homologue	RICS94A	HV1433PH	260	0.60	0.38	0.55	0.02	−0.36	0.46	−0.09	0.22
28. Chloroplastic glutamine synthetase	OSSIGS31	HVGLN2R	428	0.60	0.48	0.63	0.09	0.61**	0.45	0.26	0.84***

Table 1c. Continued

								Intragenic correlation			
Gene product	Rice	Wheat–barley	Length	GC_3 (average)	GC_2 (average)	Syn. dist.	Non-syn. dist.	$r(GC_2$–NSd)	$r(GC_3$–NSd)	$r(GC_3$–Sd)	r(Sd–NSd)
29. Sucrose synthase	OSRSS1A	HVSSYNMR	807	0.59	0.36	0.57	0.04	0.09	0.35	−0.03	−0.04
30. Mitochondrial F1-ATPase	RICATPB	TAATP2	550	0.58	0.43	0.53	0.07	0.59**	0.75***	0.05	0.19
31. Aspartic protease	RICAPA	HVASPROT	506	0.57	0.44	0.46	0.08	−0.03	0.38	−0.13	−0.01
32. ADP–glucose pyrophosphorylase	RICADP	TAAGPSMR	472	0.40	0.38	0.55	0.05	0.30	0.40	0.58*	0.65**
Probability								1.03×10^{-4}	0.0257	3.29×10^{-6}	9.4×10^{-8}

[a] See Table 1a, footnote a.

The degree of similarity between the synonymous and the nonsynonymous profiles as well as between the profiles of divergence and those of base level was determined by Pearson's correlation coefficient. In this respect, one should consider that when one calculates several correlation coefficients at the same time, some of these coefficients are expected to be significant just by chance. For instance, for the maize–rice data set, where 33 genes were submitted to the window analysis, we would expect, by chance alone, to obtain 1.65 correlation coefficients (positive or negative) that are significant at the 5% level [$0.05 * 33$], of which we expect 1.32 to be significant only at the 5% level [$(0.05–0.01) * 33$], 0.297 significant only at the 1% level [$(0.01–0.001) * 33$], and 0.033 significant at the 0.1% level [$0.001 * 33$]. Therefore, we need to know if, in a set of observed correlations (i.e., in a set of correlation coefficients, each having its own p value), the number of genes that display significant correlations exceeds random expectation by a significant amount. To know this, we have to calculate the probability that by chance alone we could obtain results that are as far, or farther, from random expectation than our results. If we choose the same significance level α for all n genes in our set, the probability that k or more correlations, positive or negative, are significant at that level is

$$P_{\alpha}(k) = \Sigma_{r=k,..,n}(n!/(r!(n-r)!))\alpha^{r}(1-\alpha)^{n-r}$$

i.e., the area under the binomial distribution to the right of $k - 1$. We can, however, obtain a fairer idea of how far our results are from random expectation if we consider three significance levels, $\alpha_1 > \alpha_2 > \alpha_3$. The probability that k correlations, positive or negative, are significant at level α_1 but not at level α_2, l correlations are significant at level α_2 but not at level α_3, and m correlations are significant at level α_3, is again obtained by summing up the probability that by chance alone we could obtain results that are as far, or farther, from random expectation than our results. This probability is given by

$$P_{\alpha1\alpha2\alpha3}(k,l,m) = \Sigma_{rst}(n!/(r!s!t!(n-r-s-t)!))$$
$$(\alpha_1-\alpha_2)^{r}(\alpha_2-\alpha_3)^{s}\alpha_3{}^{t}(1-\alpha_1)^{n-r-s-t}$$

Here, the sum is over all terms for which the probability is equal to, or lower than, that of the term corresponding to the observed set of correlations. In other words, the cutoff for the sum is given by the "contour" of the multinomial distribution (see, e.g., Feller 1950) corresponding to the observed correlation set. The probabilities shown in Tables 1a–1c (at the bottom of columns) are for $\alpha_1 = 0.05$, $\alpha_2 = 0.01$, and $\alpha_3 = 0.001$.

Results and Discussion

Intergenic Correlations: Relationship Between Nucleotide Divergence and GC Level

The synonymous and nonsynonymous distances are given in Tables 1a, 1b, and 1c along with the gene length and GC_2 and GC_3 level for each gene. Remarkably, approximately the same behavior was found in the three data sets. In this respect, it should be mentioned that the average nonsynonymous divergence is always (in the three data sets) significant and positively correlated with the GC level at the second codon position (GC_2), the correlation coefficients being $r = 0.47$ ($p < 0.01$) for the data set including rice and maize, $r = 0.60$ ($p < 0.0001$) for the second data set (maize/wheat–barley), and $r = 0.74$ ($p < 0.0001$) for the third data set (rice/wheat–barley) (Figs. 1a, 2a, and 3a). Interestingly, when the frequency of individual bases is considered separately, T_2 always exhibits strong negative correlations, but C_2 and G_2 vary among data sets. For the maize/wheat–barley data set both C_2 and G_2 exhibit significant positive correlation coefficients with the nonsynonymous distance, but only C_2 displays significant correlation in the rice/wheat–barley data set, and only G_2 in the maize–rice data set (Table 2).

Two other common behaviors are present in the three data sets. On the one hand, the GC_3 level is always negatively correlated with the synonymous distance (Figs. 1b, 2b and 3b), the correlation coefficients being very significant in all data sets ($r = -0.52$, $p < 0.001$, in maize/rice; $r = -0.62$, $p < 0.0001$, in maize/wheat–barley; $r = -0.43$, $p < 0.05$, in rice/wheat–barley). On the other hand, GC_3 is positively correlated with nonsynonymous distances in the three data sets (Figs. 1c, 2c, and 3c). These correlation coefficients are significant in rice/maize and rice/wheat–barley data sets ($r = 0.40$, $p < 0.05$, and $r = 0.43$, $p < 0.01$, respectively), but in the

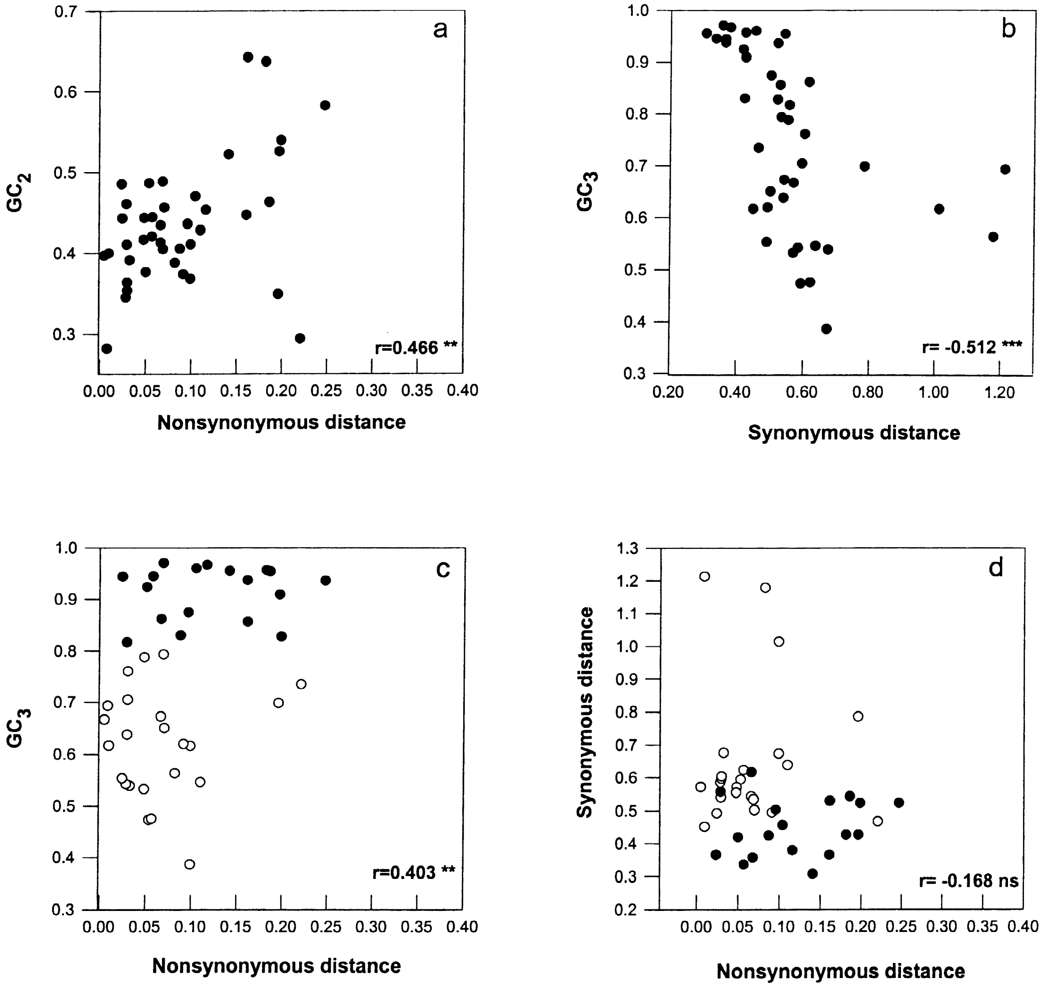

Fig. 1. **a, c** Nonsynonymous distances between homologous genes of maize/rice are plotted against GC$_2$ and GC$_3$. **b** Synonymous distances are plotted against GC$_3$. **d** Nonsynonymous distances are plotted against synonymous distances. *Filled circles* represent genes with GC$_3$ values ≥80%; *open circles* represent genes with GC$_3$ values <80%.

maize/wheat–barley data set the correlation coefficient between GC$_3$ and nonsynonymous distance happens to be at the border of significance ($r = 0.285$, $p = 0.052$). However, as we discuss later, these positive correlations between nonsynonymous distance and GC$_3$ rely on the existence of two subpopulations of genes.

As far as the correlation between synonymous and nonsynonymous distances is concerned, the results obtained in this work show that, contrary to what has been described for several mammalian gene data sets (Graur 1985; Li et al. 1985; Wolfe and Sharp 1993; Mouchiroud et al. 1995; Ohta and Ina 1995), here none of the data sets analyzed exhibited a significant correlation. However, this lack of correlation does not necessarily mean lack of relationship. In effect, the plot of synonymous vs nonsynonymous distances (see Figs. 1d, 2d, and 3d) shows the existence of two populations of genes: the GC$_3$-poorer population (GC$_3$ <80%) and the GC$_3$-richer population (GC$_3$ >80%). The former contains genes that exhibit low rates of nonsynonymous substitutions and high rates of synonymous divergence, while the latter is composed of genes having higher rates of nonsynonymous divergence and lower synonymous rates. The difference between these two populations of genes is quite evident in the three data sets analyzed in this work.

Two Subpopulations of Genes in Gramineae

To investigate further the implication of these two populations of genes, we analyzed each subpopulation separately (Table 3). Remarkably, the correlations between nonsynonymous distance and GC$_2$ level are positive and significant for the three data sets in the GC$_3$-richer subpopulation, while in the GC$_3$-poorer subpopulations the correlation coefficients vary from data set to data set. The correlation coefficients between GC$_3$ level and synonymous distances are negative in both subpopulations, but only the GC$_3$-richer subpopulations display high and significant correlation values in the three data sets. In contrast to what happens when the data sets are not split, the correlation between nonsynonymous distances and

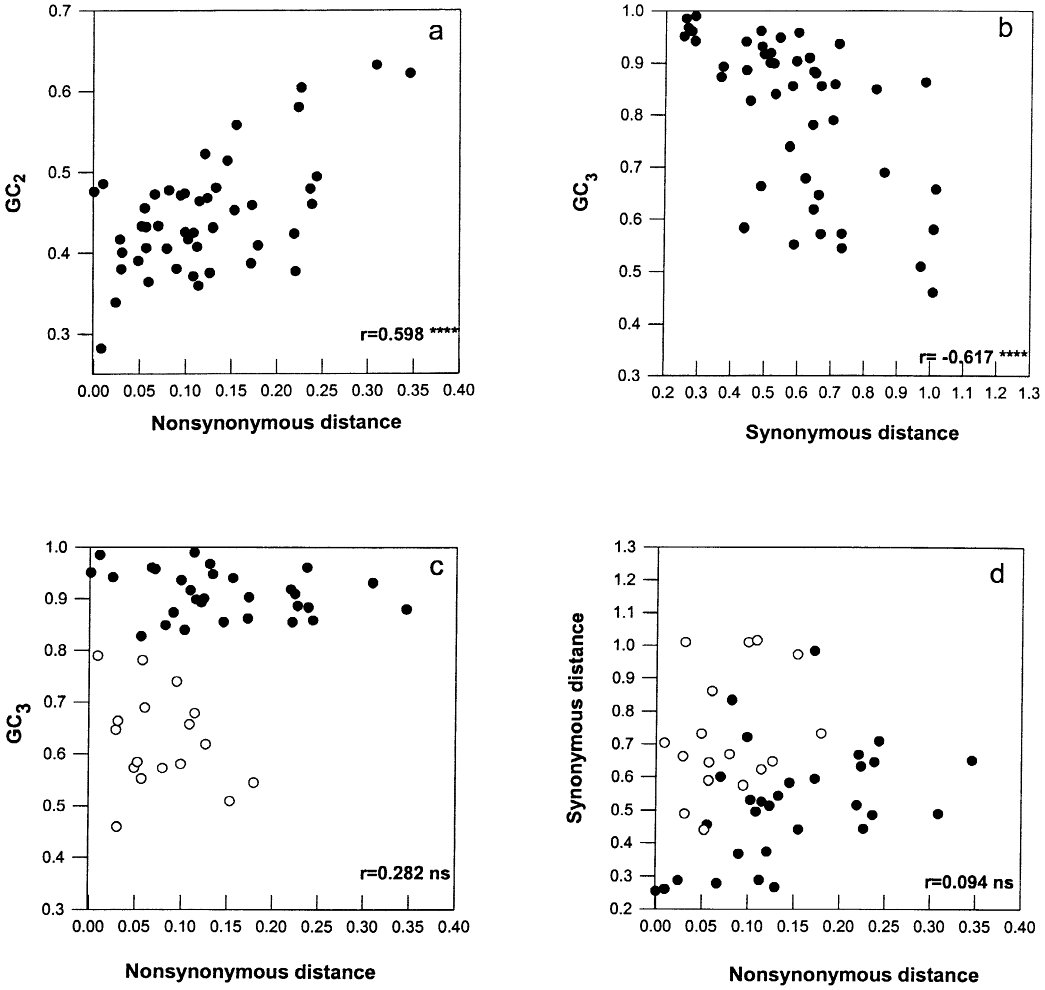

Fig. 2. **a, c** Nonsynonymous distances between homologous genes of maize/wheat–barley are plotted against GC_2 and GC_3. **b** Synonymous distances are plotted against GC_3. **d** Nonsynonymous distances are plotted against synonymous distances. *Filled circles* represent genes with GC_3 values ≥80%; *open circles* represent genes with GC_3 values <80%.

GC_3 level varies from relatively high negative values to slightly positive ones in the GC_3-richer subpopulations, while in the GC_3-poorer subpopulation figures vary from frankly positive to negative. This result indicates that the positive correlation already described between GC_3 and nonsynonymous distance is in fact the result of these two populations of genes that differ in their average nonsynonymous rate, rather than the result of a monotonous relationship. More interesting, though, is the fact that in the GC_3-richer group there is a positive correlation between the synonymous and the nonsynonymous rates, the correlation coefficients being significant in the maize/wheat–barley and rice/wheat–barley data sets. In contrast, no correlation was found between synonymous and nonsynonymous substitutional rates in the population of genes that have a lower GC_3 level.

These two groups of genes that were recognized here on the basis of their base composition and evolutionary rates were also found to differ in other respects. In effect, Carels and Bernardi (paper in preparation) reported that if the genes are classified according to their level of GC in the third codon position (i.e., the same criterion as used in this work), the two groups differ remarkably in the number and length of their introns. This result gives additional evidence in the direction that these two groups of genes very likely represent two sharply differentiated subpopulations that, besides being submitted to different kinds of constraints, also differ in their structure and perhaps in their functions.

Intragenic Correlations

The intragenic correlations between synonymous and nonsynonymous rates, as well as between GC level (in the different codon positions) and substitutions rates, were also investigated. The results for each data set are given in Tables 1a, 1b, and 1c. The first point that emerges from these tables is that in the three data sets analyzed, the intragenic correlations between GC_2 and nonsynonymous rate follow the pattern already described for the entire genes. That is, they are positive in the vast

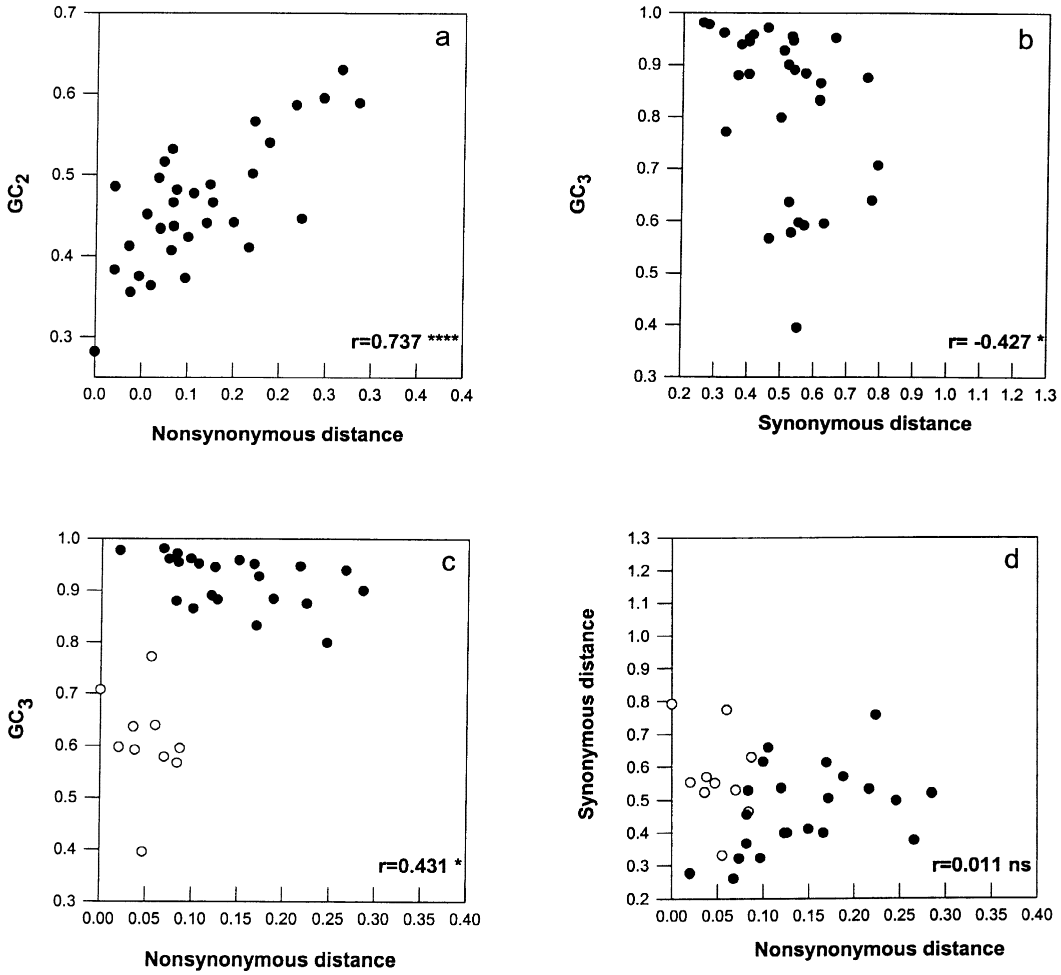

Fig. 3. **a, c** Nonsynonymous distances between homologous genes of rice/wheat–barley are plotted against GC$_2$ and GC$_3$. **b** Synonymous distances are plotted against GC$_3$. **d** Nonsynonymous distances are plotted against synonymous distances. *Filled circles* represent genes with GC$_3$ values ≥80%; *open circles* represent genes with GC$_3$ values <80%.

Table 2. Correlation coefficients between synonymous and nonsynonymous distances and base composition

	Correlation between[a]					
Data set	Sd–GC$_3$	NSd–GC$_2$	NSd–GC$_3$	NSd–C$_2$	NSd–G$_2$	NSd–T$_2$
Maize/rice	−0.512***	0.452**	0.405**	0.303	0.402**	−0.568***
Maize/wheat–barley	−0.617****	0.598****	0.282	0.541***	0.333*	−0.428**
Rice/wheat–barley	−0.427*	0.737****	0.431*	0.273	0.662****	−0.530**

[a] *Significant at the 5% level; **significant at the 1% level; ***significant at the 0.1% level; ****significant at the 0.01% level.

Table 3. Correlations in the two gene subpopulations

	Number of genes	Correlation between[a]			
		NSd–GC$_2$	NSd–GC$_3$	Sd–GC$_3$	Sd–NSy
GC$_3$-richer population (GC$_3$ ≥ 0.8)					
Maize/rice	18	0.64**	0.1	−0.61**	0.18
Maize/wheat–barley	30	0.53**	−0.25	−0.58***	0.44*
Rice/wheat–barley	22	0.59**	−0.47*	−0.53**	0.42*
GC$_3$-poorer population (GC$_3$ ≤ 0.8)					
Maize/rice	22	−0.26	0.08	−0.09	0.03
Maize/wheat–barley	17	0.32	−0.29	−0.35	0.26
Rice/wheat–barley	10	0.82**	0.43	−0.02	−0.33

[a] See Table 2, footnote a.

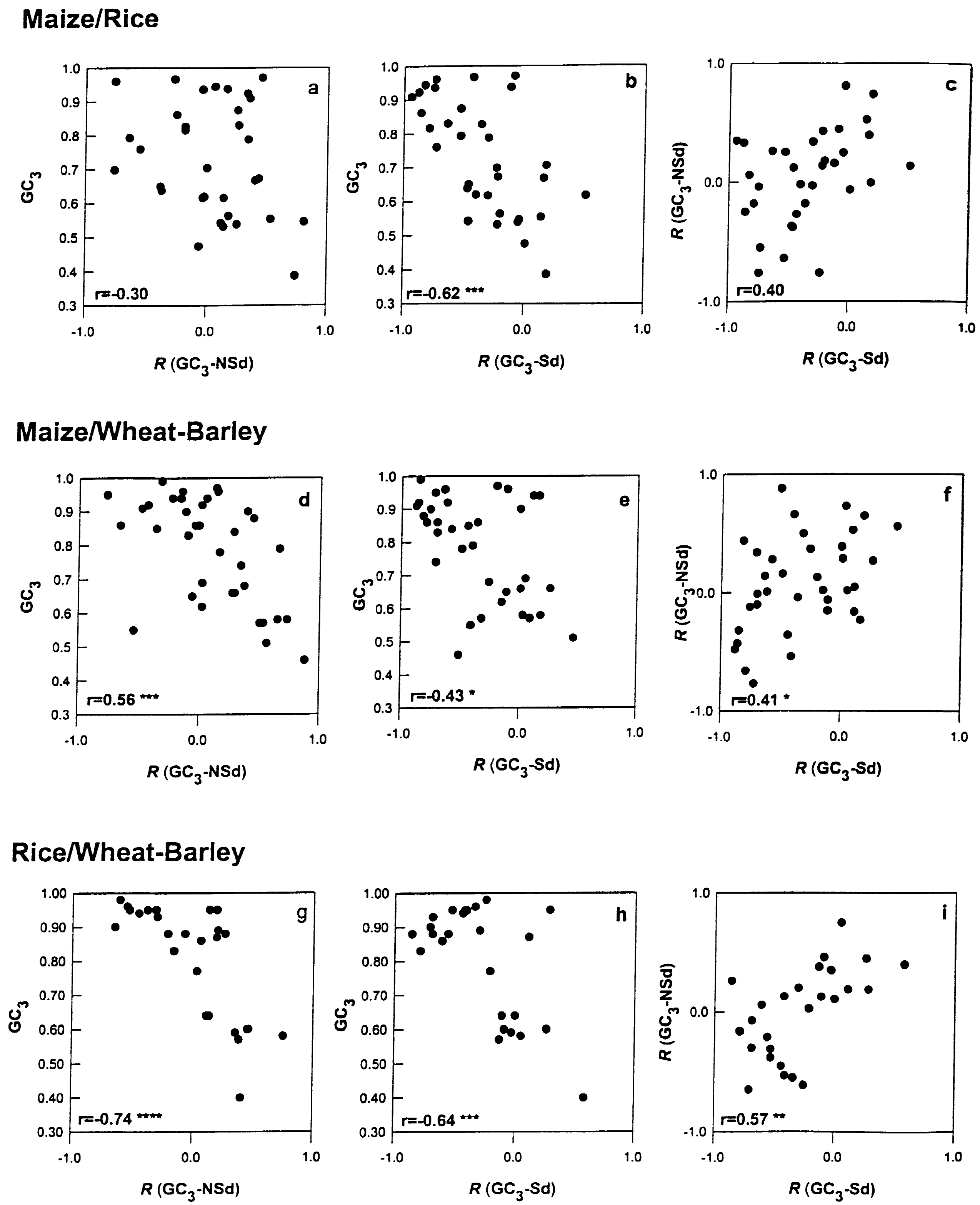

Fig. 4. **a, b, d, e, g, h** Relationships between the intragenic correlation coefficients and the GC$_3$ level (total) of genes (**c, f,** and **i**). Scatterplots between the intragenic correlation coefficients of synonymous distances and GC$_3$ vs the intragenic correlation coefficients of nonsynonymous distance and GC$_3$. *R* (GC$_3$–Sd) and *R* (GC$_3$–NSd), respectively, are the intragenic correlation coefficients between GC$_3$ vs synonymous and nonsynonymous distances.

majority of genes, being significant in about one-third of them. On the other hand, the intragenic correlation coefficients between GC_3 level and synonymous rates are negative in most genes, yet they vary from very negative in the very GC_3-rich genes to zero or slightly positive values in those genes with a lower GC_3 level. Similarly, the intragenic correlations between GC_3 level and nonsynonymous distance span from negative to positive figures. The proportion of genes displaying significant correlation coefficients is much higher than random expectation for these three comparisons (see bottom rows in Tables 1a, 1b, and 1c).

Interestingly, the correlation coefficients between the profiles of nonsynonymous divergence and GC_3 level are dependent on the total GC_3 level of the gene, in such a way that the very GC_3-rich genes exhibit negative intragenic correlation coefficients, but genes with a lower GC_3 level display positive correlation coefficients. In fact, the total GC_3 level of the genes is negatively correlated with the intragenic correlation coefficients between GC_3 and nonsynonymous distance (Figs. 4a, d, and g). As already mentioned, a similar relationship was observed between synonymous divergence and GC_3 level; in fact the intragenic correlation coefficients between GC_3 and synonymous rates are also negatively correlated with the total GC_3 of the genes. However, in this case the correlation coefficients span from very negative in the very GC_3-rich genes to zero or slightly positive values in genes with a lower GC_3 level (Figs. 4b, e, and h). These results mean that in GC_3-richer genes those segments with a higher GC_3 level tend to be more conserved, at both the synonymous and the amino acid level, while in genes with a lower GC_3 level, the GC_3-richer segments tend to evolve more rapidly. The important point that emerges from these results is that synonymous and nonsynonymous divergences exhibit a parallel behavior in relation to the GC_3 level, indicating that the intragenic patterns of both silent and amino acid divergence are related in a similar way to the intragenic variation of GC_3. This parallelism between synonymous and nonsynonymous divergence is clearly shown in Figs. 4c, f, and i, which indicate that in the three data sets there is a positive and significant correlation between the intragenic correlation coefficients of GC_3-synonymous distances and the intragenic correlation coefficients of GC_3-nonsynonymous divergence.

This behavior observed at the intragenic level is compatible with that described at the entire gene level if account is taken of the existence of two populations of genes. As noted above, at the entire gene level the correlation between GC_3 and synonymous distance was negative in the GC_3-richer population and very close to zero in the GC_3-poorer one. This is in line with what is observed at the intragenic level, where very GC_3-rich genes display a negative intragenic correlation between GC_3 and synonymous distance and genes with a lower GC_3 display no correlation or slightly positive correlation values.

As far as the intragenic correlations between synonymous and nonsynonymous distances are concerned, they range from positive to negative values, but high and significant correlation coefficients are positive in sign. Even if in the three data sets the number of genes that exhibit significant correlation coefficients between synonymous and nonsynonymous profiles is higher than random expectations, it is evident that they represent a modest proportion of the total sample. In fact, only in the rice/wheat–barley data set is the proportion of genes displaying significant coefficients comparable to that described in mammals. This scarcity of genes with significant correlations between synonymous and nonsynonymous profiles is fairly striking since, as mentioned previously, the processes of synonymous and nonsynonymous divergence are indeed related in genes from *Gramineae.* In this respect, it should be considered that at the entire gene level, the populations of genes with a higher GC_3 level display a significant correlation between synonymous and nonsynonymous distances. In addition, the intragenic patterns of synonymous and nonsynonymous divergence exhibit a parallel behavior in relation to GC_3 level. It is therefore likely that in several genes the lack of significance is due to the small number of codons, this claim being supported by the fact that there are several examples of small genes having correlations coefficients at the limit of significance.

Acknowledgments. We thank Oliver Clay for his critical reading of the manuscript as well as his help in improving it. One of us (F.A.-V.) was the recipient of a Unesco/Twas Fellowship.

References

Alvarez-Valin F, Jabbari K, Bernardi G (1998) Synonymous and nonsynonymous substitutions in mammalian genes: Intragenic correlations. J Mol Evol 46:37–44

Bernardi G, Mouchiroud D, Gautier C (1993) Silent substitutions in mammalian genomes and their evolutionary implications. J Mol Evol 37:583–589

Cacciò S, Zoubak S, D’Onofrio G, Bernardi G (1995) Nonrandom frequency patterns of synonymous substitutions in homologous mammalian genes. J Mol Evol 40:280–292

Carels N, Bernardi G (1999) Two classes of genes in plants (submitted for publication)

Duvall MR, Morton BR (1996) Molecular phylogenetics of Poaceae: An expanded analysis of rbcl sequence data. Mol Phyl Evol 5:352–358

Feller W (1950) An introduction to probability theory and its applications, 3rd ed, Vol 1. Wiley & Sons, New York, p 167

Graur D (1985) Amino acid composition and the evolutionary rates of protein-coding genes. J Mol Evol 22:53–62

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

Mouchiroud D, Gautier C, Bernardi G (1995) Frequencies of synonymous substitutions in mammals are gene-specific and correlated with frequencies of non-synonymous substitutions. J Mol Evol 40: 107–113

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

Ohta T, Ina Y (1995) Variation in synonymous substitutions rates among mammalian genes and correlations between synonymous and nonsynonymous divergences. J Mol Evol 41:717–720

Sharp PM, Li W-H (1987) The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol Biol Evol 4:222–230

Ticher A, Graur D (1989) Nucleic acid composition, codon usage, and the rate of synonymous substitution in protein-coding genes. J Mol Evol 28:286–298

Wolfe KH, Sharp PM (1993) Mammalian gene evolution: Nucleotide sequence divergence between mouse and rat. J Mol Evol 37:441–456

Wolfe KH, Sharp PM, Li W-H (1989) Mutation rates differ among regions of the mammalian genome. Nature 337:283–285

Zoubak S, D’Onofrio G, Cacciò S, Bernardi G, Bernardi G (1995) Specific compositional patterns of synonymous positions in homologous mammalian genes. J Mol Evol 40:293–307