J Mol Evol (1997) 45:227–231

JOURNAL OF MOLECULAR EVOLUTION

© Springer-Verlag New York Inc. 1997

The Influence of Specific Neighboring Bases on Substitution Bias in Noncoding Regions of the Plant Chloroplast Genome

Brian R. Morton,[1] Virginia M. Oberholzer,[2] Michael T. Clegg [2]

[1] Department of Biological Sciences, Barnard College, Columbia University, 3009 Broadway, New York, NY 10027, USA
[2] Department of Botany and Plant Sciences, University of California, Riverside, CA 92521, USA

Received: 18 October 1996 / Accepted: 12 April 1997

Abstract. Substitutions occurring in noncoding sequences of the plant chloroplast genome violate the independence of sites that is assumed by substitution models in molecular evolution. The probability that a substitution at a site is a transversion, as opposed to a transition, increases significantly with increasing A + T content of the two adjacent nucleotides. In the present study, this dependency of substitutions on local context is examined further in a number of noncoding regions from the chloroplast genome of members of the grass family (Poaceae). Two features were examined; the influence of specific neighboring bases, as opposed to the general A + T content, on transversion proportion and an influence on substitutions by nucleotides other than the two immediately adjacent to the site of substitution. In both cases, a significant effect was found. In the case of specific nucleotides, transversion proportion is significantly higher at sites with a pyrimidine immediately 5′ on either strand. Substitutions at sites of the type YNR, where N is the site of substitution, have the highest rate of transversion. This specific effect is secondary to the A + T content effect such that, in terms of proportion of substitutions that are transversions, the nucleotides are ranked T > A > C > G as to their effect when they are immediately 5′ to the site of substitution. In the case of nucleotides other than the immediate neighbors, a significant influence on substitution dynamics is observed in the case where the two neighboring bases are both A and/or T. Thus, substitutions are primarily, but not exclusively, influenced by the composition of the two nucleotides that are immediately adjacent. These results indicate that the pattern of molecular evolution of the plant chloroplast genome is extremely complex as a result of a variety of inter-site dependencies.

Key words: Nucleotide substitution — Neighboring base — Nucleotide composition — A + T content

Introduction

Biases in the process of nucleotide substitution have long been recognized as a central feature of molecular evolution and as an important factor for comparative sequence analyses (Kimura 1980; Li et al. 1984). The most prevalent bias that has been identified is a strong preference for transition substitutions over transversions in many sequences (Brown et al. 1982; Gojobori et al. 1982; Tamura 1992). This transition bias has been incorporated into a wide range of molecular analyses (Nei 1987; Swofford and Olsen 1990).

It is also becoming evident that substitution dynamics are more complex than a simple transition bias. A number of studies have shown that the substitution process can be context dependent, that is, neighboring base composition can influence the substitution bias at a particular site, such that substitution dynamics vary from site to site (Bulmer 1986; Mendelman et al. 1989; Blake et al. 1992) and, as a result, over time at a particular site.

Abbreviations: Tv = transversion; Ts = transition; ORF = Open Reading Frame; Y = pyrimidine; R = purine

Correspondence to: B.R. Morton

Context dependency of spontaneous mutations can result from influences of neighboring bases on the process of misincorporation or the process of mismatch repair, or both. Influences on misincorporation have been observed from replication by *Drosophila melanogaster* DNA polymerase α for which sites with a pyrimidine 5′ were found to have higher rates of misincorporation (Mendelman et al. 1989). Mismatch repair has also been shown to have context dependency in certain cases. Repair by *Escherichia coli* enzymes in vitro is strongly dependent on the composition of nucleotides surrounding the mismatch (Radman and Wagner 1986; Jones et al. 1987). The efficiency of mismatch repair by 3′–5′ exonuclease proofreading has also been found to depend on local composition. Unstable duplex regions are repaired more efficiently so that G+C rich regions are repaired with a low efficiency (Bessman and Reha-Krantz 1977; Petruska and Goodman 1985). In addition to influences through the process of misincorporation and repair, neighboring bases could have influences for very specific reasons. For example, methylation of CpG sites in vertebrate genomes appears to lead to a high rate of transition at these dinucleotides (Bulmer 1986).

The current work was undertaken to further address the issue of context dependency of substitutions in noncoding regions of the flowering plant chloroplast genome, a widely used genome in plant molecular evolutionary studies (Clegg 1993). It has been shown that there is a strong influence of neighboring base composition on the process of substitution in these sequences (Morton and Clegg 1995; Morton 1995) as well as in coding sequences (Morton 1997). Sites that are flanked both 5′ and 3′ by a G and/or C have a much higher proportion of transitions than sites flanked on both sides by A and/or T. This influence is observed throughout the entire genome with the result that substitution dynamics, in terms of transition: transversion ratio, can vary noncoding regions due to differences in composition (Morton 1995).

The current study addresses this issue in further detail. The influence of neighbors other than the immediate neighbors is tested as well as the influence of specific flanking nucleotides. Both tests give significant results. Sites with a 5′ pyrimidine are found to have higher rates of misincorporations which lead to a transversion than do sites with a 5′ purine. In addition, when both flanking bases are A and/or T, the composition of the two sites that are one nucleotide further removed are found to be correlated substitution dynamics. As a result, sites in highly unstable (A + T rich) regions have a much higher proportion of transversions.

Materials and Methods

Substitutions from noncoding regions of the grass chloroplast genome were analyzed. The regions studied are those described previously (Morton and Clegg 1995; Morton 1995) in addition to a multiple sequence alignment of the region flanked by the gene *psbM* and ORF29.

Table 1. Nucleotide composition of noncoding regions from the rice chloroplast genome

Region	% A + T	Length
1	67.7	548
2	71.3	1061
3	67.9	346
4	64.3	1307
5	69.6	1084
6	67.4	794
7	63.3	371
8	68.8	600
9	65.7	600
10	68.3	486
11	66.5	704
12	68.6	784
13	64.9	1001
14	66.6	1196
15	69.2	318
16	69.6	441
17	69.6	510
18	64.8	1530
19	73.5	446
psbM-ORF29	71.3	766

Sequence was generated for the region flanked by *psbM* and *trnC* employing PCR amplification with the primers ATGAAGTCAATATTCTCGCATTTAT and CCAGTTCTAAATCTGGGTGCCGCCT following the conditions from Morton and Clegg (1995). PCR products were cleaned following the GeneClean protocol from BIO101 and then both strands were sequenced directly using the fmol protocol from Promega. Species sequenced were *Avena sativa*, *Bambusa multiplex*, *Eragrostis japonica*, *Hordeum vulgare*, *Muhlenbergia setaroides*, *Pennisetum glaucum*, *Sorghum bicolor*, *Zea mays*, and *Zizania texana* (see Morton and Clegg 1995 and Duvall and Morton 1996 for DNA sources). This region from the *Oryza sativa* complete genome sequence (Hiratsuka et al. 1989) was also included. All of the noncoding regions used in this study are quite similar in terms of composition, being within the range of 63.3 to 73.5% A + T (Table 1) and are, therefore, directly comparable. The combined length of the regions analyzed is 14,894 nucleotides in the rice chloroplast genome.

Substitutions were scored as described in Morton and Clegg (1995) and Morton (1995). Every comparison involved the same strand of the chloroplast genome, the strand defined as forward in the rice genome (Hiratsuka et al. 1989). Therefore, neighboring base composition is directly comparable. Substitution type was scored as a transversion or transition and for each substitution neighboring base composition was recorded for two pairs of nucleotides separately (see Fig. 1). These are the N1 pair (the two nucleotides immediately 5′ and 3′ to the site of substitution) and the N2 pair (the two nucleotides immediately 5′ and 3′ to the N1 pair). In all cases, substitutions recorded were those for which neighboring base composition was unambiguous.

Results and Discussion

ORF29 Evolution Indicates a Coding Function

In the rice genome, an open reading frame, ORF29, lies within the region flanked by *psbM* and *trnC*. The homologous region in the other grass species sequenced

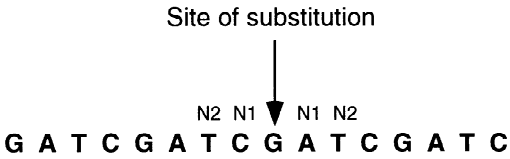

Fig. 1. Definition of the N1 and N2 neighboring base pairs referred to in the text.

here was found to be highly conserved suggesting that ORF29 is a coding sequence. Although chloroplast noncoding regions undergo a very high rate of insertion/deletion (indel) mutations (Morton and Clegg 1993) and the region flanked by *psbM* and *trnC* has a large number, no indels are observed within ORF29 (data not shown). Further, the region of ORF29 has a very low rate of substitution relative to flanking sequences, and all of the substitutions that are observed have occurred at putative third codon positions (data not shown). These two features indicate that ORF29 is most likely a functional protein-coding sequence and it was therefore excluded from the analyses of substitutions. A Blast search yielded no homologous protein to the putative ORF29 polypeptide.

Effect of Specific Nucleotides on Substitution Dynamics

The effect of specific combinations of N1 nucleotides on substitution bias is summarized in Table 2. Substitution bias is given for each of the 16 possible specific N1 compositions grouped by the A + T content of the N1 nucleotides and ranked by transversion proportion within each group. As observed in separate regions (Morton and Clegg 1995; Morton 1995), substitutions overall are strongly dependent upon neighboring base composition. Sites flanked by A and/or T both 5′ and 3′ have a high proportion of transversions, roughly twice the rate observed when both N1 nucleotides are G or C. The difference in transversion proportion between the different contexts is highly significant ($\chi^2 = 46.81$, $p < 0.001$). Since the regions are so similar in composition (Table 1), and since context dependency is observed in separate regions (Morton and Clegg 1995) as well as within specific coding regions (Morton 1997), it is unlikely that the results are an artifact of composition heterogeneity among regions.

It is apparent from Table 2 that, when specific neighboring base combinations are considered, there is considerable variation within each of the three N1 A + T composition classes, as well as some overlap between them. When N1 = 0 A + T, the GNC context, where N designates the site of substitution, has the highest proportion of transversions. Also noticeable is that when N1 = 2, A + T, the ANT context has a much lower proportion of transversions than do the other three specific nucleotide combinations. In both of these cases, it is also apparent that the transversion proportion is much lower

Table 2. Influence of specific base combinations on substitutions

N1 A + T content[a]	5′	3′	Ts	Tv	Tv / Tv + Ts
0	G	C	20	6	0.231
	C	C	17	7	0.292
	G	G	25	11	0.306
	C	G	19	11	0.367
			81	35	0.302
1	A	C	26	8	0.235
	C	T	40	16	0.286
	G	T	28	12	0.300
	A	G	39	26	0.400
	T	C	25	17	0.405
	G	A	32	23	0.418
	T	G	36	29	0.446
	C	A	38	31	0.449
			264	162	0.380
2	A	T	45	36	0.444
	A	A	33	52	0.612
	T	A	35	53	0.624
	T	T	30	54	0.643
			143	195	0.577

[a] A + T content of the two N1 nucleotides (see Fig. 1)

than the reverse complement context (CNG and TNA, respectively) which is unexpected since the reverse complement is the same context for substitutions on the complementary strand. In other reverse complement pairings, the transversion proportion is very similar in the complementary contexts, with the exception of ANG and CNT.

This difference in substitution bias between certain reverse complement pairings suggests a reason for the variation among specific nucleotide combinations of the same A + T content. The two contexts noted above, GNC and ANT, which have low proportions of transversions for their A + T content are of the type RNY. On the other hand, an examination of N1 combinations that are of the type YNR shows that they have higher transversion proportions. A comparison of substitutions to the 5′ or 3′ nucleotide separately shows that, for both A versus T and G versus C, a 5′ pyrimidine, or a 3′ purine (which is a 5′ pyrimidine on the opposite strand), is associated with a higher average rate of transversion. This ranking is secondary to A + T content so that, with respect to transversion proportion, 5′ nucleotides are ranked T > A > C > G (see data from Table 2).

This association of a 5′ pyrimidine with increased rates of transversion can be examined further. A comparison of substitutions to the purine/pyrimidine content of the two N1 nucleotides together shows that the presence of a 5′ pyrimidine is highly correlated with transversion proportion. If we consider both strands, then sites can be classified as having no 5′ pyrimidines (YNR), one 5′ pyrimidine (YNY/RNR), or two 5′ pyrimidines (RNY). As shown in Table 3, transversion proportion increases significantly as the number of 5′ pyrimidines

Table 3. Effect of a 5′ pyrimidine on substitutions

N1 A + T content	Number of 5′ pyrimidines	Ts	Tv	Tv / Tv + Ts	Probability
	0	20	6	0.231	
0	1	42	18	0.300	$\chi^2 = 1.23$
	2	19	11	0.367	(NS)
	0	54	20	0.270	
1	1	136	82	0.376	$\chi^2 = 13.3$
	2	54	60	0.526	($p < 0.01$)
	0	45	36	0.444	
2	1	63	106	0.627	$\chi^2 = 7.77$
	2	35	53	0.602	($p < 0.05$)
	0	119	62	0.343	
All	1	241	206	0.461	$\chi^2 = 15.2$
	2	108	124	0.534	($p < 0.01$}

increases. This increase is not significant when N1 = 0 A + T but the existence of the trend is interesting and may become significant with more data.

The increased proportion of transversions at sites with a 5′ pyrimidine raises an interesting possibility in relation to the factors that may be responsible for the observed increase in the proportion of transversions in certain contexts. Two main possibilities exist concerning the influence of neighboring bases on substitution, an influence by context on mismatch repair or on replication (including repair-associated replication). Since replication proceeds 5′ to 3′, if misincorporation by DNA polymerase is influenced by neighboring bases it is most likely to be affected by the 5′ nucleotide. The results presented here are consistent with a model in which misincorporation by the DNA polymerase in the flowering plant chloroplast has different dynamics when it occurs downstream of a pyrimidine as opposed to a purine. The increase in transversion proportion from zero through two 5′ pyrimidines indicates that, if replication is responsible, this effect occurs on both leading and lagging strand. It is suggestive that a very similar effect has been observed during replication by DNA polymerase α from *Drosophila* (Mendelman et al. 1989).

If replication is a factor, then there are three possibilities for how such an increase in transversions could occur. One is that an upstream pyrimidine could increase the absolute rate of misincorporations leading to transversions while not affecting rate of misincorporation leading to transitions. The second is the opposite, a decrease in the rate of misincorporation leading to transitions. The third possibility is that the absolute rates are affected in opposing manners. Using the data from the pairwise comparison of rice and maize (Morton 1995), the absolute numbers of transversions and transitions per total number of sites observed were calculated separately for sites of the type YNR (two flanking 5′ pyrimidines) and RNY (no 5′ pyrimidines). The results are shown in Table 4. The two contexts do not differ significantly in terms of absolute rate of transitions per site ($\chi^2 = 1.68$, N. S.) but significantly more YNR sites have undergone a transversion than have RNY sites ($\chi^2 = 11.1$, $p < 0.01$). Therefore, the high proportion of transversions at sites of the type YNR appears to be at least partially due to an increase in the absolute rate of transversions, with no change in the absolute rate of transitions. If replication, including repair-related DNA replication, is a major factor in the context dependency observed here, then the data from Tables 1 to 3 suggest that a 5′ pyrimidine leads to an increased rate of misincorporations leading to transversion substitutions.

Table 4. Influence of 5′ nucleotide on number of transitions and transversions per site

Context[a]	Sites	Ts	Tv	Ts/site	Tv/site
YNR	1381	83	55	0.060	0.040
RNY	1416	69	26	0.049	0.018

[a] N represents the site of substitution

Table 5. Influence of N2 pair composition on substitutions

N1 A + T content	N2 A + T content	Ts	Tv	Tv / Tv + Ts	Probability
	0	6	6	0.500	
0	1	38	15	0.283	$\chi^2 = 4.67$ (NS)
	2	31	7	0.184	
	0	29	12	0.293	
1	1	107	58	0.352	$\chi^2 = 3.05$ (NS)
	2	116	83	0.417	
	0	12	11	0.478	
2	1	52	56	0.519	$\chi^2 = 6.45$ ($p < 0.05$)
	2	59	112	0.655	

Effect of A + T Content of Nonadjacent Neighbors

When the composition of the N1 pair is held constant, the transversion proportion can be compared to the composition of the N2 pair as shown in Table 5. When the N1 pair composition is 0 or 1 A + T, the composition of the N2 pair is not observed to have a significant influence on substitution type. However, when the N1 nucleotides are both A or T, the composition of the N2 pair is correlated with transversion proportion. The relationship, increasing transversion proportion with increasing A + T content, mimics the effect of the N1 nucleotides and indicates that the composition of nonadjacent neighboring bases can influence substitutions. As a result, sites flanked by four A and/or T nucleotides have the highest proportion of transversion. This result, an influence of nonadjacent nucleotides only when the N1 composition is 2 A + T, is identical to what has been observed in chloroplast coding sequences (Morton 1997).

A possibility is that the overall A + T content of four flanking nucleotides (N1 and N2 pair combined) influences substitution bias regardless of the composition of the N1 pair, so that the influence of N1 that is seen in Table 2 might be a result of the different distribution of the four nucleotide A + T compositions among each of the three N1 composition classes. However, comparing the cases where the A + T content of the four nucleotides flanking a site of substitution is 50% regardless of which of the nucleotides are A or T (N1 = 0 and N2 = 2, N1 = 1 and N2 = 1, and N1 = 2 and N2 = 0) shows that there is a significant difference in transversion proportion between these cases ($\chi^2 = 6.27$, $p < 0.05$). Therefore, the composition of the N1 pair appears to have the dominant effect on substitution dynamics at a site with the influence of the N2 pair being a function of the N1 composition.

Conclusions

It is becoming clear that the pattern of substitution in noncoding regions of the plant chloroplast genome is complex and involves a great deal of inter-site dependency (Morton and Clegg 1995; Morton 1995, 1997). Substitutions occurring in a context of high A + T content have a much greater frequency of transversions than those occurring in a high G + C context. Further, it appears that substitutions are not just affected by the composition of nucleotides that immediately flank the site but can be influenced by nucleotides further upstream or downstream.

As well as a general effect of neighboring base A + T content on substitution dynamics, specific nucleotides influence substitutions in different manners. The influence of specific neighboring base composition, however, appears to be secondary to the effect of A + T content. Substitutions are increasingly biased as the A + T content increases, but within this general increase, transversion bias increases as the number of 5′ pyrimidines increases. The result is a layered effect, with substantial variation between sites in terms of substitution dynamics.

This complex context dependency of substitutions in the angiosperm chloroplast genome raises important questions for molecular evolutionary studies that use chloroplast DNA. One issue is that the assumption of site independence, which underlies all current models of molecular evolution, clearly does not hold. What effect this has on analyses that employ these models is not clear, but it is certainly possible that their use creates biases in certain instances. In particular, phylogenetic analyses that employ noncoding regions, whether spacer regions or introns, could be affected by the violation of site independence. Whether or not this violation has a significant effect should be determined if such methods are to be employed.

Acknowledgments. This work was supported in part by NIH grant GM 45144.

References

Bessman MJ, Reha-Krantz LJ (1977) Studies on the biochemical basis of spontaneous mutation. V. Effect of temperature on mutation frequency. J Mol Biol 116:115–123

Blake RD, Hess ST, Nicholson-Tuell J (1992) The influence of nearest neighbors on the rate and pattern of spontaneous point mutations. J Mol Evol 34:189–200

Brown WM, Prager EM, Wang A, Wilson AC (1982) Mitochondrial DNA sequences of primates: tempo and mode of evolution. J Mol Evol 18:225–239

Bulmer M (1986) Neighboring base effects on substitution rates in pseudogenes. Mol Biol Evol 3:322–329

Clegg MT (1993) Chloroplast gene sequences and the study of plant evolution. Proc Natl Acad Sci USA 90:363–367

Duvall MR, Morton BR (1996) Molecular phylogenetics of Poaceae: an expanded analysis of *rbc*L sequence data. Mol Phylogenetics and Evol 5:352–358

Gojobori T, Li WH, Graur D (1982) Patterns of nucleotide substitution in pseudogenes. J Mol Evol 18:360–369

Hiratsuka J, Shimada H, Whittier R, Ishibashi T, Sakamoto M, Mori M, Kondo C, Honji Y, Sun CR, Meng BY, Li YQ, Kanno A, Nishizawa Y, Hirai A, Shinozaki K, Sugiura M (1989) The complete sequence of the rice (*Oryza sativa*) chloroplast genome: intermolecular recombination between distinct tRNA genes accounts for a major plastid DNA inversion during the evolution of the cereals. Mol Gen Genet 217:185–194

Jones M, Wagner R, Radman M (1987) Repair of a mismatch is influenced by the base composition of the surrounding nucleotide sequence. Genetics 115:605–610

Kimura M (1980) A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111–120

Li WH, Wu CI, Luo CC (1984) Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. J Mol Evol 21:58–71

Mendelman LV, Boosalis MS, Petruska J, Goodman MF (1989) Nearest neighbor influences on DNA polymerase insertion fidelity. J Biol Chem 264:14415–14423

Morton BR, Clegg MT (1993) A chloroplast DNA mutational hotspot and gene conversion in a noncoding region near *rbc*L in the grass family (Poaceae). Curr Genet 24:357–365

Morton BR, Clegg MT (1995) Neighboring base composition is strongly correlated with base substitution bias in a region of the chloroplast genome. J Mol Evol 41:597–603

Morton BR (1995) Neighboring base composition and transversion/transition bias in a comparison of rice and maize chloroplast noncoding regions. Proc Natl Acad Sci USA 92:9717–9721

Morton BR (1997) The influence of neighboring base composition on substitutions in plant chloroplast coding sequences. Mol Biol Evol 14:189–194

Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York

Petruska J, Goodman MF (1985) Influence of neighboring bases on DNA polymerase insertion and proofreading. J Biol Chem 260: 7533–7539

Radman M, Wagner R (1986) Mismatch repair in *Escherichia coli.* Ann Rev Genet 20:523–538

Swofford DL, Olsen GJ (1990) Phylogeny reconstruction. In: DM Hillis and C Moritz (eds) Molecular systematics. Sinauer Associates, Sunderland, MA, pp 411–501

Tamura K (1992) The rate and pattern of nucleotide substitution in *Drosophila* mitochondrial DNA. Mol Biol Evol 9:814–825.