ISSN : 0974 - 7435 Volume 10 Issue 17

2014

BioTechnology

An Indian Journal

FULL PAPER

BTAIJ, 10(17), 2014 [9772-9780]

Neighboring sites effect and substitution trends in poaceae chloroplast genome

Zhuoran Huang[1,2], Mingchuan Fu[1,2], Jian Cheng[1,2], Shiheng Tao[1,2*]

[1]College of Life Sciences and State Key Laboratory of Crop Stress Biology in Arid Areas, Northwest A&F University, Yangling, Shaanxi, 712100, (CHINA)

[2]Bioinformatics Center, Northwest A&F University, Yangling, Shaanxi, 712100, (CHINA)

E-mail: shihengt@nwsuaf.edu.cn, Shiheng Tao shihengt@nwsuaf.edu.cn, Zhuoran Huang hzrds@hotmail.com, Mingchuan Fu fmcsky@nwsuaf.edu.cn, Jian Cheng chengjian2007@126.com

ABSTRACT

Recent research on chloroplast genome focused on the sequencing of new plastid genomes and comparing related genomes. Influence exerted by non-adjacent sites is rarely mentioned. In the current study, the substitution sites were counted based on the pairwise comparison of five Poaceae chloroplast genomes, using *Phalaenopsis aphrodite* chloroplast genome as the outgroup. The relationship between mutation patterns and the base composition of flanking sites was detected. A significant flanking sites effect was observed. Substitutions to and from each dinucleotide were calculated, and three strong "losers" (AA, AT, and TA) and four strong "gainers" (CC, CG, GC, and GG) were found. The number of AA is higher in the ancestral sequence which gradually decreased in the evolution process. The reduction in A and T with C and G accumulation are reported as well. The dinucleotide substitution trends are largely determined by the mononucleotide substitution trends. Results indicate that context significantly influences mutations, further enhancing our understanding of context dependency and mutation dynamics in chloroplast genomes.

KEYWORDS

Chloroplast genome; Poaceae; Neighboring sites effect; Context dependency; Nucleotide substitution; Mutation pattern.

INTRODUCTION

Chloroplasts, the semi-autonomous organelles in plant cells, have their own genome that encodes a number of photosynthesis proteins and several housekeeping proteins. Chloroplast genomes (cpDNA) are highly conserved in the organization where most plant plastid genomes are composed of a single circular double-stranded DNA molecule containing large and small single-copy regions separated by two copies of inverted repeats[1-3]. The complete cpDNA sequences of tobacco (*Nicotiana tabacum*)[4] and liverwort (*Marchantia polymorpha*)[5] were first established as a result of the development in DNA sequencing technology. Consequently, the cpDNA of many species were sequenced completely, including those of rice (*Oryza sativa*), maize (*Zea mays*), wheat (*Triticum aestivum*), barley (*Hordeum vulgare*), and sorghum (*Sorghum bicolor*).

The most recent studies on cpDNA concentrated on the sequencing of new plastid genomes and comparative analysis of related genomes. For instance, 12 *Gossypium* chloroplast genomes were sequenced in 2012, with the detection of sequence variations among species using sequence alignment[6]. Recently, the comparative analysis of Proteaceae chloroplast genome showed that species-rich lineages tend to have significantly higher chloroplast substitution rates, which may be one of the influences on the plant populations divergence speed[7]. Similarly, we compared the 5 Poaceae chloroplast genomes with the use of sequence alignment, in order to explore the substitution patterns of these genomes.

The investigation of the neighbor-dependent mutation in chloroplasts has been progressing for a number of years. Morton[8] found that substitution bias is notably correlated with the base composition of the immediate neighboring sites both in the non-coding and coding sequences, except for the *rbcL* gene. In addition, substitutions have been observed to favor transversions over transitions in neighboring sites with higher A+T content[9]. Further research showed that the number of flanking pyrimidines on the same strand significantly influences the substitution properties as well[10]. In their research on maize nuclear genome, several relationships between the flanking base composition and the mutation pattern have been reported. The A+T content of the two sites immediately flanking the mutation site is correlated with the rate, transition bias, and GC→AT pressure[11]. However, the studies on neighbor-dependent mutation in chloroplasts cited above failed to report on complete cpDNA genome profiles and the effects of the composition of more distant nucleotide sites. Accordingly, we used the recent sequenced results to facilitate former research on neighbor effect in cpDNA.

In nuclear genes, the most apparent neighboring nucleotide effect that has been studied to date is the CpG effect. CpG deficiency in vertebrate genomes and human sequences is widely accepted to be the result of cytosine methylation and deamination of 5-methylcytosine leading to TpG and CpA dinucleotides. However, the observed context dependency in cpDNA is not consistent with CpG deamination, and CpG methylation has not been established to occur in cpDNA. Further understanding of context dependency and mutation dynamics in cpDNA is necessary. Thus, in our study, the existence of a dinucleotide bias in plastid genomes is investigated, and the relationship between mutation patterns and the base composition of flanking sites is determined by comparing five Poaceae cpDNA using classical chi-square tests. The mechanism underlying the selection effects on the substitution pattern is investigated more intensively by considering more flanking bases.

MATERIALS AND METHODS

CpDNA Sequences

The complete cpDNA sequences of *H. vulgare* (NC_008590)[12], *T. aestivum* (NC_002762)[13], O. sativa (NC_001320)[14], Z. mays (NC_001666)[15], S. bicolor (EF115542)[12], and P. aphrodite (NC_007499)[16] were downloaded from GenBank (http://www.ncbi.nlm.nih.gov/genbank). We counted 16 dinucleotide compositions in each Poaceae cpDNA, and calculated the dinucleotide frequencies.

Substitution inference and Sequence alignment

The substitution sites were calculated based on the pairwise comparison of five Poaceae cpDNA, using *P. aphrodite* as the outgroup. The tri-species alignments were performed using CLUSTAL W ver. 2.0.12[17]. The method employed to infer a nucleotide substitution in a tri-species alignment has been described previously[18]. Given that at a certain nucleotide site, the wheat cpDNA has A, and both the rice and *P. aphrodite* cpDNA sequences have C, then nucleotide C is assumed to be substituted with A in the wheat cpDNA, i.e., C→A. Twelve substitution categories exist, i.e., each type of nucleotide can be substituted with any of the other three types. Deletions and insertions were excluded. The position of the substitution site was arbitrarily labeled as "zero." Subsequently, the positions at the 5′ flank were designated as negative numbers and the positions at the 3′ flank as positive numbers[19]. The base compositions from the -3 to +3 sites were calculated.

Statistical test

The 4×4 χ^2 test was used to assess the statistical differences between cases and controls in dinucleotide category and dinucleotide frequency. The influence from neighbor site to substitution site was estimated by the 12×4 test. Statistical differences with $p<0.01$ were considered significant. χ^2 tests were performed using SPSS software version 15.0.

RESULTS

Dinucleotide frequency

The dinucleotide frequencies of five Poaceae cpDNA are shown in Figure 1. The dinucleotide compositions are highly conserved in the Poaceae plant cpDNA[10, 12, 16]. AA, TT, AT and TA are the four most frequent dinucleotides. CG and GC

appear as the two least dinucleotides, which is consistent with CpG deficiency caused by CpG methylation[20, 21]. TABLE 1, which shows the dinucleotide numbers of wheat cpDNA, was taken as a 4×4 cross-tabulation (χ^2=3385.41, *P*-value<0.0001). Thus, a significant relationship between nearest-neighbor sites exists. The chi-square contributions of 16 dinucleotides (TABLE 1), including AA, AC, CC, GG, GT, TA, and TT, with the largest contributions to overall significance (χ^2>300) were calculated. Of these seven cells, AA, CC, GG, and TT showed exceedingly positive deviations, whereas AC, GT, and TA showed exceedingly negative deviations. Research on the other four Poaceae plant cpDNA yielded the same results.

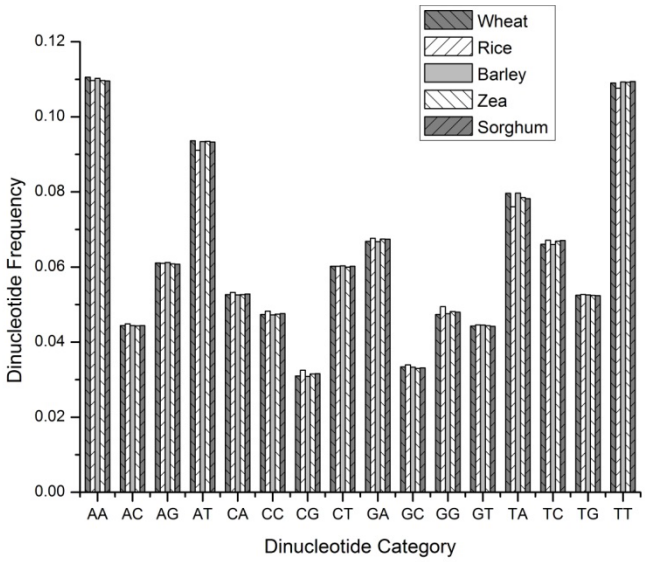

Figure 1 : The dinucleotide frequencies are almost consistent in these five Poaceae cpDNA (wheat, rice, barley, maize, and sorghum). The four most frequent dinucleotides are AA, TT, AT, and TA. The two least frequent dinucleotides are CG and GC.

TABLE 1 : Chi-square tests used for dinucleotide numbers of wheat cp DNA

	A	C	G	T	Total
A					
count	14876	5973	8220	12597	41666
expected count	12903.16	7967.47	7996.58	12798.80	41666.00
chi-square	301.64*	499.27*	6.24	3.18	863.11
C					
count	7084	6373	4170	8101	25728
expected count	7967.47	4919.77	4937.74	7903.03	25728.00
chi-square	97.96	429.27*	119.37	4.96	651.56
G					
count	8992	4496	6373	5961	25822
expected count	7996.58	4937.74	4955.78	7931.90	25822.00
chi-square	123.91	39.52	405.29*	489.72*	1058.44
T					
count	10714	8886	7059	14670	41329
expected count	12798.80	7903.03	7931.90	12695.28	41329.00
chi-square	339.59*	122.26	96.06	307.16*	865.08
Total count	41666	25728	25822	41329	134545
chi-square	863.11	1090.31	626.96	805.03	3385.41
DF			9		
P-Value			0.00		

***Large chi-square contribution to overall significance.**

NEIGHBORING-NUCLEOTIDE EFFECTS ON THE SUBSTITUTION SITES

TABLE 2 : Chi-square tests used for barley cpDNA -1 sites (Barley/Sorghum comparison)

	A	C	G	T	Total
A→C					
count	22	26	25	54	127
expected count	37.86	28.14	22.05	38.95	127.00
chi-square	6.64	0.16	0.39	5.82	13.02
A→G					
count	123	86	105	84	398
expected count	118.65	88.18	69.11	122.06	398.00
chi-square	0.16	0.05	18.63*	11.87	30.71
A→T					
count	42	30	13	60	145
expected count	43.23	32.13	25.18	44.47	145.00
chi-square	0.03	0.14	5.89	5.43	11.49
C→A					
count	18	7	12	47	84
expected count	25.04	18.61	14.59	25.76	84.00
chi-square	1.98	7.24	0.46	17.51*	27.19
C→G					
count	16	10	11	20	57
expected count	16.99	12.63	9.90	17.48	57.00
chi-square	0.06	0.55	0.12	0.36	1.09
C→T					
count	81	83	46	90	300
expected count	89.44	66.47	52.10	92.00	300.00
chi-square	0.80	4.11	0.71	0.04	5.66
G→A					
count	102	71	55	89	317
expected count	94.50	70.23	55.05	97.22	317.00
chi-square	0.59	0.01	0.00	0.69	1.30
G→C					
count	10	15	8	26	59
expected count	17.59	13.07	10.25	18.09	59.00
chi-square	3.27	0.28	0.49	3.45	7.51
G→T					
count	31	22	13	46	112
expected count	33.39	24.81	19.45	34.35	112.00
chi-square	0.17	0.32	2.14	3.95	6.58
T→A					
count	49	20	6	45	120
expected count	35.77	26.59	20.84	36.80	120.00
chi-square	4.89	1.63	10.57	1.83	18.91
T→C					
count	144	125	84	114	467
expected count	139.22	103.47	81.10	143.22	467.00
chi-square	0.16	4.48	0.10	5.96	10.71
T→G					
count	59	23	28	42	152
expected count	45.31	33.68	26.40	46.61	152.00
chi-square	4.13	3.38	0.10	0.46	8.07
Total					
count	697	518	406	717	2338
chi-square	22.90	22.37	39.61	57.37	142.25
DF			33		
P-Value			0.00		

***Large chi-square contribution to overall significance.**

As described in the Materials and Methods section, 2,338 substitutions were obtained after excluding adjacent multiple-base substitutions, deletions, and insertions based on the tri-alignment among barley (136,462 bp), sorghum (140,754 bp), and *P. Aphrodite* (148,964 bp). TABLE 2 shows the -1 base composition relative to each substitution pattern in barley cpDNA, e.g., 127 A→C substitutions at the 0 sites occurred, where 22 A, 26 C, 25G, and 54 T at the -1 sites were obtained. A chi-square test was performed on this 12×4 cross-tabulation (χ^2=142.25, *P*-value<0.0001). The context effect showed a tendency to be significant on the -1 site. The chi-square contributions of 48 cells are also shown in TABLE 2. The cell -1 base has G with an A→G substitution, and the cell -1 base has T with a C→A substitution. These two bases provided the two largest contributions. The conditional probability at the -1 and +1 sites in barley cpDNA were calculated (TABLE 3). The cell -1 base with G and an A→G substitution, as well as the cell -1 base with T and a C→A substitution, expressed the two largest deviations compared to the relative substitution rate. At the +1 site, the cell +1 base with A and a T→C substitution and the cell +1 base with G and a G→A substitution provided the two largest deviations. The chi-square test results at the -3, -2, +1, +2, and +3 sites in barley cpDNA (barley/sorghum comparison) are shown in TABLE 4. The chi-square value is considerably larger at the -1 and +1 sites, and decreased at the more distant sites. Even at the -3 and +3 sites, no significant relationship (*P*-value >0.01) between the base composition and mutation patterns was observed. Similar analyses were performed on the other four cpDNA in 8 comparison results. Barley/wheat and zea/sorghum comparisons were ignored because of the insufficient number of substitution sites.

TABLE 3 : Conditional probability at -1 and +1 sites in barley cpDNA (Barley/Sorghum comparison)

	E	-1	A	C	G	T	+1	A	C	G	T
A→C	0.054		0.032	0.050	0.062	0.075		0.048	0.064	0.053	0.057
A→G	0.170		0.176	0.166	0.259*	0.117		0.170	0.173	0.180	0.161
A→T	0.062		0.060	0.058	0.032	0.084		0.068	0.041	0.038	0.087
C→A	0.036		0.026	0.014	0.030	0.066*		0.037	0.036	0.019	0.048
C→G	0.024		0.023	0.019	0.027	0.028		0.028	0.031	0.013	0.026
C→T	0.128		0.116	0.160	0.113	0.126		0.115	0.128	0.133	0.138
G→A	0.136		0.146	0.137	0.135	0.124		0.141	0.115	0.193*	0.098
G→C	0.025		0.014	0.029	0.020	0.036		0.028	0.018	0.021	0.030
G→T	0.048		0.044	0.042	0.032	0.064		0.070	0.031	0.036	0.044
T→A	0.051		0.070	0.039	0.015	0.063		0.076	0.026	0.051	0.041
T→C	0.200		0.207	0.241	0.207	0.159		0.134*	0.268	0.193	0.233
T→G	0.065		0.085	0.044	0.069	0.059		0.085	0.071	0.071	0.037

***The two largest deviations compared to the expect substitution probabilities, which were shown in column E.**

TABLE 4 : Chi-square value at-3, -2, -1, +1, +2, and +3 sites in barley cpDNA (Barley/Sorghum comparison)

Sites	-3	-2	-1	1	2	3
Chi-Square	54.75	113.49*	142.25*	122.59*	73.72*	37.32

******P*-value<0.01.**

Mononucleotide and dinucleotide substitution trends

The substitutions to and from each dinucleotide were measured to determine the relationship between the base composition of adjacent sites and the chi-square values of different dinucleotides. We found that seven substitutions showed consistent substitution trends, i.e., the gain and loss[22-24] of dinucleotides in all comparison results (TABLE 5). Three strong substitution "losers" (AA, AT, and TA) are reduced, and four strong "gainers" (CC, CG, GC, and GG) accumulated in all categories.

TABLE 5 : The strong "gainer" and "loser" dinucleotides in 8 comparison results

		Substitutions to and from a dinucleotide*						
Comparison	Taxon	AA[a]	AT[a]	CC[b]	CG[b]	GC[b]	GG[b]	TA[a]
b/s	barley	349/390	286/466	303/176	261/196	225/133	285/209	361/407
		-0.055	-0.239	0.265	0.142	0.257	0.154	-0.060
	sorghum	329/402	222/426	305/173	314/145	194/142	319/183	254/380
		-0.100	-0.315	0.276	0.368	0.155	0.271	-0.199
b/r	barley	352/382	286/413	288/183	231/206	209/131	262/206	342/372
		-0.041	-0.182	0.223	0.057	0.229	0.120	-0.042
	rice	327/450	242/490	371/160	347/143	247/119	370/188	247/461
		-0.158	-0.339	0.397	0.416	0.350	0.326	-0.302
b/z	barley	329/387	289/447	298/183	251/195	225/132	317/212	355/401
		-0.081	-0.215	0.239	0.126	0.261	0.198	-0.061
	zea	322/433	246/406	304/175	302/149	187/148	318/168	284/395
		-0.147	-0.245	0.269	0.339	0.116	0.309	-0.163
r/s	rice	282/316	171/422	314/168	291/151	218/124	324/167	211/383
		-0.057	-0.423	0.303	0.317	0.275	0.320	-0.290
	sorghum	317/338	237/355	277/180	231/164	190/129	250/178	279/332
		-0.032	-0.199	0.212	0.170	0.191	0.168	-0.087
r/w	rice	294/399	205/416	326/150	295/124	228/111	328/155	206/396
		-0.152	-0.340	0.370	0.408	0.345	0.358	-0.316
	wheat	344/393	302/378	293/174	219/187	215/131	238/194	337/355
		-0.066	-0.112	0.255	0.079	0.243	0.102	-0.026
r/z	rice	266/393	191/438	329/168	281/148	219/115	310/170	218/390
		-0.193	-0.393	0.324	0.310	0.311	0.292	-0.283
	zea	316/356	265/382	303/198	249/181	193/146	251/206	305/368
		-0.060	-0.181	0.210	0.158	0.139	0.098	-0.094
w/s	wheat	333/373	250/416	323/142	256/173	229/128	302/187	306/386
		-0.057	-0.249	0.389	0.193	0.283	0.235	-0.116
	sorghum	286/359	201/391	305/163	305/127	191/124	294/156	245/363
		-0.113	-0.321	0.303	0.412	0.213	0.307	-0.194
z/w	zea	333/457	259/427	309/194	299/152	186/133	316/174	299/401
		-0.157	-0.245	0.229	0.326	0.166	0.290	-0.146
	wheat	360/428	258/426	330/162	272/182	243/130	316/200	348/409
		-0.086	-0.246	0.341	0.198	0.303	0.225	-0.081

***The number of substitutions creating (C) and removing (R) of seven dinucleotides, which showed consistent substitution trends in all comparison results, together with their normalized difference D= (C-R)/(C+R)[22], [a] Strong loser dinucleotides, D<0, [b] Strong gainer dinucleotides, D>0**

TABLE 6 : The strong gainer and loser mononucleotides in 8 comparison results

		Substitutions to and from a mononucleotide*			
Comparison	**Taxon**	**A[a]**	**C[b]**	**G[b]**	**T[a]**
b/s	barley	521/670	653/441	607/488	557/739
		-0.125	0.194	0.109	-0.140
	sorghum	483/700	620/368	642/400	431/708
		-0.183	0.255	0.232	-0.243
b/r	barley	523/647	576/459	559/474	571/649
		-0.106	0.113	0.082	-0.064
	rice	462/805	718/345	730/386	427/801
		-0.271	0.351	0.308	-0.305
b/z	barley	515/684	612/445	641/484	561/716
		-0.141	0.158	0.140	-0.121
	zea	487/724	613/388	647/396	463/702
		-0.196	0.225	0.241	-0.205
r/s	rice	403/644	600/320	615/357	377/674
		-0.230	0.304	0.265	-0.283
	sorghum	459/579	562/393	504/415	473/611
		-0.116	0.177	0.097	-0.127
r/w	rice	393/687	633/316	654/317	358/718
		-0.272	0.334	0.347	-0.335
	wheat	522/635	558/413	550/446	523/659
		-0.098	0.149	0.104	-0.115
r/z	rice	413/684	623/337	637/336	397/713
		-0.247	0.298	0.309	-0.285
	zea	492/614	585/433	542/432	533/673
		-0.110	0.149	0.113	-0.116
w/s	wheat	488/656	635/399	613/434	488/735
		-0.147	0.228	0.171	-0.202
	sorghum	423/661	597/338	586/353	413/667
		-0.220	0.277	0.248	-0.235
z/w	zea	493/734	632/409	655/410	515/742
		-0.196	0.214	0.230	-0.181
	wheat	539/709	654/421	664/446	511/792
		-0.136	0.217	0.196	-0.216

***The number of substitutions creating (C) and removing (R) of all mononucleotides, which showed consistent substitution trends in all comparison results, together with their normalized difference D= (C-R)/(C+R)[22], [a] Strong loser mononucleotides, D<0, [b] Strong gainer mononucleotides, D>0**

The substitutions to and from each mononucleotide were also measured. The reduction in A and T with the accumulation of C and G is shown in TABLE 6. The A+T content is considerably higher than the G+C content in Poaceae cpDNA.

DISCUSSION AND CONCLUSIONS

The major finding of our study is that the adjacent neighboring sites exerted a significant influence on the mutations. However, no significant CpG effect was observed. The composition of the three immediate neighbors of the mutation site is correlated with the mutation patterns. These effects are similar to those obtained by previous studies and are possibly due to the influence of the local composition on polymerase misincorporation or mismatch repair[10, 25]. Previous studies on mouse and human SNPs have indicated that nucleotides beyond the immediate neighbors can influence nucleotide mutation biases[26, 27]. The compositions of non-adjacent neighboring nucleotide sites do not exert as much influence as the two immediately flanking sites, which is similar to the results of the current study[28, 29].

According to TABLE 5, the number of AA substitutions should be reduced based on the substitutions in barley cpDNA using the barley/sorghum comparison. However, the chi-square test shows that the number of AA substitutions significantly exceeded the expected count (TABLE 1). It could be concluded that the number of AA substitutions was higher in the ancestral sequence and gradually decreased in the evolution process. Otherwise, the substitution trends of TA, CC, and GG are consistent with the deviation in TABLE 1. The consistence is probably due to the substitution bias. The number of AG and CT substitutions is almost stable in the evolution process. The data of mononucleotide substitution show that the A+T content is believed to decline with the increase in the G+C content. Considering the further analysis, the dinucleotide substitution trends are largely determined by the mononucleotide substitution trends.

A complete comparison of these cpDNA sequences is convenient because of the small scale of cpDNA. However, the smallness of the genome scale restricted our investigation and consequently, insufficient substitution data was achieved. On the other hand, these methods may be very useful in interpreting mammalian nuclear genomes because of their considerably larger genome size.

ACKNOWLEDGMENTS

We gratefully acknowledge The Ph.D. Programs Foundation of Ministry of Education of China (20100204110026) in SHT Lab. We thank Xu Zhao for their helpful discussion and comments on statistical analysis. We appreciate Jiang Xiaoqian and Wu Wenwu for instructive advices in manuscript writing.

REFERENCES

[1] J.D.Palmer; Plastid chromosomes: Structure and evolution, The molecular biology of plastids, 7, 5-53 **(1991)**.

[2] J.D.Palmer, J.M.Nugent, L.A.Herbon; Unusual structure of geranium chloroplast DNA: A triple-sized inverted repeat, extensive gene duplications, multiple inversions, and two repeat families. Proceedings of the National Academy of Sciences **84**, 769-773 **(1987)**.

[3] L.A.Raubeson, R.K.Jansen; Chloroplast genomes of plants, Plant diversity and evolution: genotypic and phenotypic variation in higher plants, **45**, **(2005)**.

[4] K.Shinozaki, M.Ohme, M.Tanaka, T.Wakasugi, N.Hayashida, T.Matsubayashi, N.Zaita, J.Chunwongse, J.Obokata, K.Yamaguchi-Shinozaki; The complete nucleotide sequence of the tobacco chloroplast genome: Its gene organization and expression, The EMBO journal, **5**, 2043 **(1986)**.

[5] K.Ohyama, H.Fukuzawa, T.Kohchi, H.Shirai, T.Sano, S.Sano, K.Umesono, Y.Shiki, M.Takeuchi, Z.Chang; Chloroplast gene organization deduced from complete sequence of liverwort Marchantia polymorpha chloroplast DNA, **(1986)**.

[6] Q.Xu, G.Xiong, P.Li, F.He, Y.Huang, K.Wang, Z.Li, J.Hua; Analysis of complete nucleotide sequences of 12 Gossypium chloroplast genomes: Origin and evolution of allotetraploids, PloS one, 7, e37128, **(2012)**.

[7] D.Duchene, L.Bromham; Rates of molecular evolution and diversification in plants: Chloroplast substitution rates correlate with species-richness in the Proteaceae, BMC evolutionary biology, **13**, 65 **(2013)**.

[8] B.R.Morton; The influence of neighboring base composition on substitutions in plant chloroplast coding sequences. Molecular Biology and Evolution, **14**, 189-194 **(1997)**.

[9] B.R.Morton, V.M.Oberholzer, M.T.Clegg; The influence of specific neighboring bases on substitution bias in noncoding regions of the plant chloroplast genome, Journal of molecular evolution, **45**, 227-231 **(1997)**.

[10] B.R.Morton; The role of context-dependent mutations in generating compositional and codon usage bias in grass chloroplast DNA, Journal of molecular evolution, **56**, 616-629 **(2003)**.

[11] B.R.Morton, I.V.Bi, M.D.McMullen, B.S.Gaut; Variation in mutation dynamics across the maize genome as a function of regional and flanking base composition, Genetics, **172**, 569-577 **(2006)**.

[12] C.Saski, S.B.Lee, S.Fjellheim, C.Guda, R.K.Jansen, H.Luo, J.Tomkins, O.A.Rognli, H.Daniell, J.L.Clarke; Complete chloroplast genome sequences of Hordeum vulgare, Sorghum bicolor and Agrostis stolonifera, and comparative analyses with other grass genomes, Theoretical and applied genetics, **115**, 571-590 **(2007)**.

[13] Y.Ogihara, K.Isono, T.Kojima, A.Endo, M.Hanaoka, T.Shiina, T.Terachi, S.Utsugi, M.Murata, N.Mori; Chinese spring wheat (Triticum aestivum L.) chloroplast genome: complete sequence and contig clones, Plant Molecular Biology Reporter **18**, 243-253 **(2000)**.

[14] J.Hiratsuka, H.Shimada, R.Whittier, T.Ishibashi, M.Sakamoto, M.Mori, C.Kondo, Y.Honji, C.R.Sun, B.Y.Meng; 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, Molecular and General Genetics MGG, **217**, 185-194 **(1989)**.

[15] R.M.Maier, K.Neckermann, G.L.Igloi, H.Kössel; Complete sequence of the maize chloroplast genome: Gene content, hotspots of divergence and fine tuning of genetic information by transcript editing, Journal of molecular biology, **251**, 614-628 **(1995)**.

[16] C.C.Chang, H.C.Lin, Lin, T.Y.Chow, H.H.Chen, W.H.Chen, C.H.Cheng, C.Y.Lin, S.M.Liu, C.C.Chang; The chloroplast genome of Phalaenopsis aphrodite (Orchidaceae): comparative analysis of evolutionary rate with that of grasses and its phylogenetic implications, Molecular Biology and Evolution, **23**, 279-291 **(2006)**.

[17] J.D.Thompson, D.G.Higgins, T.J.Gibson; CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, Nucleic acids research, **22**, 4673-4680 **(1994)**.

[18] T.Gojobori, W.H.Li, D.Graur; Patterns of nucleotide substitution in pseudogenes and functional genes, Journal of Molecular Evolution, **18**, 360-369 **(1982)**.

[19] L.Ma, T.Zhang, Z.Huang, X.Jiang, S.Tao; Patterns of nucleotides that flank substitutions in human orthologous genes, BMC genomics, **11**, 416 **(2010)**.

[20] B.K.Duncan, J.H.Miller; Mutagenic deamination of cytosine residues in DNA, Nature, **287**, 560-561 **(1980)**.

[21] A.P.Bird; DNA methylation and the frequency of CpG in animal DNA, Nucleic Acids Research, **8**, 1499-1504 **(1980)**.

[22] I.K.Jordan, F.A.Kondrashov, I.A.Adzhubei, Y.I.Wolf, E.V.Koonin, A.S.Kondrashov, S.Sunyaev; A universal trend of amino acid gain and loss in protein evolution, Nature, **433**, 633-638 **(2005)**.

[23] L.D.Hurst, E.J.Feil, E.P.Rocha; Protein evolution: Causes of trends in amino-acid gain and loss, Nature, **442**, E11-E12 **(2006)**.

[24] K.Misawa, N.Kamatani, R.Kikuno; The universal trend of amino acid gain–loss is caused by CpG hypermutability, Journal of Molecular Evolution, **67**, 334-342 **(2008)**.

[25] B.R.Morton; Neighboring base composition and transversion/transition bias in a comparison of rice and maize chloroplast noncoding regions, Proceedings of the National Academy of Sciences, **92**, 9717-9721 **(1995)**.

[26] Z.Zhao, E.Boerwinkle; Neighboring-nucleotide effects on single nucleotide polymorphisms: A study of 2.6 million polymorphisms across the human genome, Genome.Res., **12**, 1679-1686 **(2002)**.

[27] F.Zhang, Z.Zhao; The influence of neighboring-nucleotide composition on single nucleotide polymorphisms (SNPs) in the mouse genome and its comparison with human SNPs, Genomics, **84**, 785-795 **(2004)**.

[28] P.A.Nevarez, C.M.DeBoever, B.J.Freeland, M.A.Quitt, E.C.Bush; Context dependent substitution biases vary within the human genome, Bmc Bioinformatics, **11**, 462 **(2010)**.

[29] A.Panchin, S.Mitrofanov, A.Alexeevski, S.Spirin, Y.Panchin; New words in human mutagenesis, Bmc Bioinformatics, **12**, 268 **(2011)**.