

Research Article

ISSN 2320-4818
JSIR 2013; 2(6): 1017-1022
© 2013, All rights reserved
Received: 27-07-2013
Accepted: 31-12-2013

Binod C Sharma

PG Department of Botany,
Darjeeling Government College,
Darjeeling 734101, India

Dipen Ghimiray

PG Department of Botany,
Darjeeling Government College,
Darjeeling 734101, India

Abhijit Chhetri

Department of Microbiology, St.
Joseph's College, Singamari,
Darjeeling 734104, India

Correspondence:

Binod C Sharma

PG Department of Botany,
Darjeeling Government College, 18
Lebong Cart Road, Darjeeling
734101, India

Tel: +919434575465

Fax: +913542254078

E-mail: bcsgc@gmail.com

Study of the AT/GC composition in chloroplast genomes of Solanaceae shows strong conservation of base-composition in genes of plastomes

Binod C Sharma*, Dipen Ghimiray, Abhijit Chhetri

Abstract

AT/GC composition from different protein encoding genes of chloroplast genomes from the family Solanaceae taking four species of *Nicotiana* and three species of *Solanum* was studied. We selected genes of photosystem I and II, larger sub-unit of RUBISCO, NADH dehydrogenase, ATPase, 4 RNA polymerases, conserved genes designated as *ycf* (except *ycf1* and *ycf2*) and other three genes *matK*, *accD* and *clpP*; and also, ribosomal genes of 30S and 50S subunits. rRNA genes and ORFs were evaluated in *Nicotiana tabacum* only. Genes of tRNAs and AT/GC composition in introns were excluded in this study. In all genes AT%, GC% and AT/GC ratio was calculated first manually, and then by our own developed tool AACAL 1.2. This study shows that the highest AT% and lowest GC% is present in *ycf3* (70.14 and 29.86), highest GC% and lowest AT% is present in the gene of 16S (58.56 and 41.44), and, highest and lowest AT/GC ratio was found to be present in *ycf3* (3.05) and gene of 16S (0.70), respectively. Mean AT%, GC% and AT/GC ratio was found to be 60.803, 39.197 and 1.583 respectively for protein encoding genes; for rRNA encoding genes of *Nicotiana tabacum* it was respectively 62.665, 37.335 and 1.678; similarly in the ORFs of *Nicotiana tabacum* AT%, GC% and AT/GC ratio was found to be 63.466, 36.534 and 1.737 respectively. Therefore, we can conclude that the base compositions in the plastomes (especially in family Solanaceae) are conserved and more or less uniform in their GC%. These may be because of their occurrence in the same taxonomic group or due to less diversion in the plastomes vis-a-vis time.

Keywords: Chloroplast genome, Solanaceae, Protein encoding genes, GC%, AT%, AT/GC ratio.

Introduction

After the pioneering sequencing of chloroplast genome in tobacco¹, international databanks like NCBI, EMBL and DDBJ show the total number of sequenced chloroplast genome more than 300, and surely this number will go on increasing, as now there occurs several sophisticated methods of genome sequencing. These chloroplast genomes are more or less 120-150 kb long and contains nearly 120 genes¹⁻⁵ which can be grouped into different categories like genes of photosystem I (designated as *psa*), photosystem II (*psb*), cytochrome b6/F(*pet*), atp synthase (*atp*), larger sub-unit of RUBISCO (*rbcL*), different components of RNA polymerases (*rpo*), NADH oxidoreductase (*ndh*), 50S and 30S subunits of ribosomal proteins (*rpl* and *rps* respectively), miscellaneous proteins (*accD*, *clpP*, *matK*), certain conserved proteins (*ycf*), unassigned open reading frames (ORFs) and genes of tRNAs and rRNAs.

Comparative study of the sequenced chloroplast genome (otherwise known as Chloroplast DNA, cpDNA, Plastome) has been done by various workers. Shimada and Sugiura² were first to analyse the initial three sequenced chloroplast plastomes. The sequence of sugarcane chloroplast genome was studied and a comparative analysis on four monocots was performed.⁶ A comparative study of various genes rice chloroplast genomes is described.⁷ Loss of phylogenetic distribution of *rps12* and *clpP* intron among legumes are also reported.^{8, 9} The plastome of an orchid (*Phalaenopsis aphrodite*) was sequenced and the evolutionary rate as well as phylogenetic implications of this was compared with chloroplast genome of grasses.¹⁰ A comparative study of *Solanum bulbocastanum* and *S. lycopersicum* genomes is previously done.¹¹ Similarly, comparative study of chloroplast genomes of cotton and other eight dicots are reported.^{12, 13} Loss of inverted repeat copies and influence of heterotachy from the chloroplast genome of members of pinaceae and cupressophytes are described.¹⁴ Chloroplast genomes from cryptophyta are more or less intron free and contain many additional genes.⁵

Chloroplast genomes of 7 species of family solanaceae have been sequenced, of which, 3 members belonged to the genus Solanum: *S. tuberosum* {NCBI GenBank Reference Sequence NC_007500}¹⁵, *S. bulbocastanum* {NC_007943 and *S. lycopersicum* {NC_007898},¹¹ and 4 members belonged to the genus Nicotiana: *N. tabacum* {NC_001879}¹, *N. sylvestris* {NC_007500}¹⁵, *N. tomentosiformis* {NC_007602}¹⁵ and *N. undulate* {NC_016068}.¹⁶ In this communication, we studied genes of photosystem I and II, larger sub-unit of RUBISCO, NADH dehydrogenase, ATPase, 4 RNA polymerases, conserved genes designated as *ycf* (except *ycf1* and *ycf2*) and other three genes *matK*, *accD* and *clpP*; and also, ribosomal genes of 30S and 50S subunits. rRNA genes and

ORFs were evaluated in *Nicotiana tabacum* only. Genes of tRNAs were excluded in this study. AT%, GC% and AT/GC ratio from all genes under study are deduced.

Material and methods

Genomes of seven species of solanaceae as mentioned above were extracted from the genbank. Genes were separated according to their annotation following their grouping convention. For this we manually checked the length of different genes under study, separated them according to their start and stop sites, and calculated the total number of adenine and thymine from each gene. Total number of cytosine and guanine was derived by deducting the total number of adenine and thymine from the full gene length, as well as by individual guanine and cytosine counting and subtracting their total number from full gene length to cross check the numbers of adenine and thiamine. These data were again recalculated using calculation software AACAL 1.2.¹⁷

AT% was calculated by the simple formula – total number of adenine and thymine divided by the full gene length and multiplying the result by one hundred. GC% was also calculated in the same way. AT/GC ratio was determined by standard method of total adenine and thymine divided by total number of guanine and cytosine. In each case, first individual percentage and ratio in each gene was calculated, and then an average for the gene group was deduced. Finally, all data obtained against individual species were compared among all seven species of solanaceae.

Result

Average AT%, GC% and AT/GC ratio in the available sequenced chloroplast genomes of seven species of solanaceae is shown in Table.1

Table 1: AT/GC composition in chloroplast genome of seven species of Solanaceae

GENE	AT%	GC%	AT/GC		GENE	AT%	GC%	AT/GC
<i>psa</i>	59.344	40.656	1.460		<i>rpo</i>	62.665	37.335	1.678
<i>psb</i>	61.108	38.892	1.571		<i>ycf</i>	67.792	32.208	2.105
<i>pet</i>	61.065	38.935	1.568		<i>rRNA</i>	62.665	37.335	1.678
<i>atp</i>	49.401	50.599	0.976		<i>pseudo</i>	65.100	34.900	1.865
<i>ndh</i>	64.417	35.583	1.810		<i>ORF</i>	63.466	36.534	1.737
<i>rbcL</i>	56.140	43.860	1.280		Protein	60.803	39.197	1.583
<i>others</i>	62.820	37.180	1.690		rRNA	62.665	37.335	1.678
<i>rps</i>	62.040	37.960	1.634		Pseudo	65.100	34.900	1.865
<i>rpl</i>	62.044	37.956	1.635		ORF	63.466	36.534	1.737

Genes of photosystem I

Out of 5 genes of photosystem I – *psaA*, *psaB*, *psaC*, *psal* and *psaJ*, present in all seven species of the family, highest AT% was found in the *psal* with 65.76%, which corresponds to the lowest GC% of 34.24; followed by *psaB* with AT% of 56.76 and thus GC% of 40.90. Similarly, highest AT/GC ratio of the group occurred in the *psal* with 1.92, followed by *psaB* with 1.44. Lowest AT/GC ratio was found in the *psaA*, in which AT% corresponds to 56.76 and GC% was 43.24.

Genes of photosystem II

Fifteen genes of photosystem II were found in the seven species of solanaceae under study. *psbM* showed highest AT% of 69.52, lowest GC% 30.48 and highest AT/GC ratio of 2.28. This was followed by *psbI* and *psbL*, both with AT% 66.67, GC% 33.33, AT/GC ratio 2. Lowest AT% (55.28) and highest GC% (44.72) was found in *psbA*; followed by second lowest percentage of AT and second highest of GC as 56.18 and 43.82 respectively in *psbB*; with the second lowest AT/GC ratio of 1.28. Except for minor variation in *psbI* in *Nicotiana tabacum* and *psbC* in *Solanum bulbocastanum*, all genes of this group show the conserved base composition among the plastomes of the family.

Genes of cytochrome b6/F

Out of six genes of cytochrome b6/F, highest AT% (64.91), lowest GC% (35.09) and highest AT/GC ratio (1.85) was observed in the *petG*. This was followed by the second highest AT%, second lowest GC% and second highest AT/GC ratio in *petL* with the respective value of 64.58, 35.42 and 1.82. Lowest AT% highest GC% and lowest AT/GC of the group was found in *petN* with 57.78, 42.22 and 1.36 respectively.

Genes of ATP-synthase

Out of six genes in this group, all across the studied seven species of solanaceae; *atpI* shows the highest AT% (62.36), lowest GC% (37.64) and highest AT/GC ratio (1.65) which is also equal in case of *atpF*; though the latter shows slightly lower AT% of 62.34 and a bit higher 37.66 GC%. Lowest AT% and AT/GC ratio was found in the *atpH* with the values of 53.65 and 1.15 respectively, which corresponds to the highest GC% for the group with value of 46.35.

Genes of RUBISCO

Plastomes of all the seven species of solanaceae show only one copy of gene that encodes larger subunit of the RUBISCO enzyme. Designated as *rbcL*, with a constant length of 1434 basepairs among the plastomes in the family, its AT%, GC% and AT/GC ratio was respectively 56.14, 43.86 and 1.279. Total AT length was 805 bp and rest 629 bp were GC.

Genes of NADH-oxidoreductase

Eleven genes of NADH oxidoreductase system were found to be present in the solanaceae plastome, their length were also constant among the studied 7 species. In this group, highest AT% was observed in case of *ndhA* with the value of 70.14, lowest GC% with value of 29.86 and AT/GC ratio of high 2.34. Similarly *ndhE* shows AT% of 67.64, GC% of 32.36 and AT/GC ratio of 2.09. Lowest AT/GC ratio and AT% was found in *ndhJ* with values of 1.47 and 59.53, respectively. Second lowest AT/GC ratio was that of *ndhH* with 1.54, second lowest AT% was of *ndhK* with 62.23.

Genes of RNA polymerase

Constant numbers of 4 genes of the plastome among the member of the solanaceae family were observed. The longest gene *rpoC2* (2567 bp) shows AT% 61.60, GC% 38.40 and AT/GC ratio 1.60. Highest AT% was found in the *rpoA* (65.78%), highest GC% was observed in *rpoC1* (38.85) which is also the only intron containing gene of the group, whilst highest AT/GC ratio was observed in *rpoA* with value of 1.92.

Genes of ribosomal larger sub-unit

Nine genes for proteins of larger subunit of plastidal ribosome are present in the plastome of solanaceae. Their number is fairly constant among the two investigated genera. Highest AT% and AT/GC ratio was observed in *rpl32* with the respective value of 67.26 and 2.05, which corresponds to the lowest GC% of 32.74. Alternatively, lowest AT% and AT/GC ratio of 56.60 and 1.30 respectively was found to be present in the *rpl2*, which also has the lowest GC% value 43.40.

Genes of ribosomal smaller sub-unit

There are nine genes of smaller subunit of ribosome in the solanaceae plastid genome. Longest of the group is *rps2* (711 bp) and shortest is the *rps14* (303 bp). Highest AT% was found in *rps15* with 69.31%, which also, corresponds to the lowest GC% of 30.69; followed by *rps18* with AT% of 65.03 and thus GC% of 34.97. Similarly, highest

AT/GC ratio of the group was found in *rps15* with 2.25, followed by *rps8* and *rps18* with 1.85. Lowest AT/GC ratio was found in the *rps11* (1.20), in which AT% corresponds to 56.76 and GC% was 43.24.

Conserved genes

Out of six conserved genes, we studied four genes – *ycf3*, *ycf4*, *ycf5* (= *ccsA*) and *ycf10* (= *cemA*). Highest AT% and AT/GC ratio was found to be 75.34 and 3.05 respectively in *ycf3*, which corresponds to the lowest GC% of 24.66. Lowest AT% and AT/GC ratio among this group of genes was obtained in the *ycf4* with the corresponding value of 61.44 and 1.59, with the highest GC content of 38.56%.

Other genes

Three genes – *matK*, *accD* and *clpP* are considered as other genes (sometimes genes of miscellaneous function). They show AT% of 67.05, 63.54 and 57.84, respectively. Similarly, in the same sequence AT/GC ratio found to be 2.03, 1.74 and 1.37 respectively for these three genes.

Genes for rRNAs

There are four genes encoding rRNAs – 4.5S, 5S, 16S and 23S. AT% for these four genes are respectively 49.51, 47.93, 41.44 and 44.23. Also, there AT/GC ratio was found to be 0.98, 0.92, 0.70 and 0.79, respectively.

Pseudogene and ORF of *Nicotiana tabacum*

One pseudogene for translation initiation factor is present which is of 318 bp. It's AT%, GC% and AT/GC ratio was found to be 65.10, 34.90 and 1.86 respectively. Similarly 14 ORFs were found in the plastome of *N. tabacum*; of which, ORF103 shows the highest AT% (70.19), while ORF75 had the highest GC% (45.62), and highest AT/GC ratio was found in ORF103 (2.35) (Table 1).

Discussion

In our study, AT and GC percentage were calculated from different protein encoding genes of the chloroplast DNA from the seven species of solanaceae. We found that, AT/GC ratio, AT% and GC% were identical among all the seven species under study (Fig.1). The least deviation and highly conserve base pattern in this family's plastome suggests that there is very little evolutionary deviation among the member at family level. On a whole, AT% of *psa* was 59.34, *psb* was 61.10, *pet* has 61.06, *atp* has 49.40, *ndh* has 64.41, *rbcL* shows 56.14, *rps* and *rpl* shows 62.04, *rpo* has 62.66, *ycf* shows 67.79. Other three genes – *matK*, *clpP* and *accD* shows on the average, AT% of 62.82. Similarly, in the plastome of *Nicotiana tabacum*, AT% of rRNA genes, on the average, was found to be 62.66; open reading frames shows 63.46; and one pseudogene of translation factor has AT% of 65.10. On a whole, *ycf* shows highest AT% of 67.792 while lowest AT% was found to be present in *atp* with 49.401.

Similarly, average GC% was calculated and it was found to be highest in *atp* with 50.599% and the lowest was observed in *ycf* with 32.208%. In *psa* it was 40.65, *psb* 38.89, *pet* 38.93, *atp* 50.59, *ndh* 35.58, *rbcL* 43.86, *rps* and *rpl* 37.96, *ycf* 32.20%; while, remaining three genes – *matK*, *clpP* and *accD* shows average GC% of 37.18; whereas, rRNA genes, pseudo gene, ORFs in *N. tabacum* shows respectively 37.33, 34.90, 36.53 GC%. AT/GC ratio among different genes were found as follows – *psa* 1.46, *psb* 1.57, *pet* 1.56, *atp* 0.97, *ndh* 1.81, *rbcL* 1.28, *rps* and *rpl* 1.63, *rpo* 1.67, *ycf* 2.1, miscellaneous (others) genes 1.69; whereas, rRNA genes, pseudo gene and ORFs of *N. tabacum* shows 1.67, 1.86 and 1.73, respectively. On a whole, protein encoding genes shows AT% of 60.80, GC% 39.19 and AT/GC ratio of 1.58. Similarly, rRNA genes have AT%, GC% and AT/GC ratio of 62.66, 37.33 and 1.67, respectively. These figures exclude the pseudogene and ORFs of *N. tabacum*.

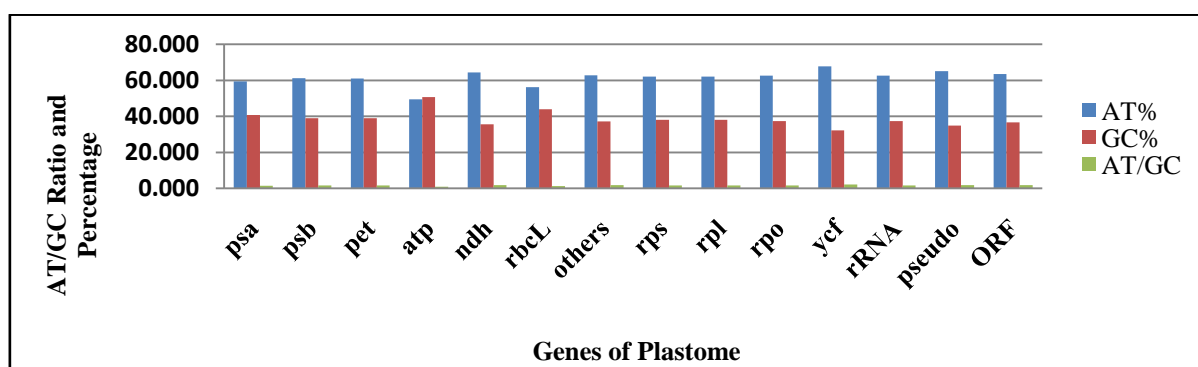


Figure 1: Average AT/GC composition in plastomes of seven species Solanaceae family

Conclusion

Slightly higher AT/GC composition in conserved genes like *ycf* may be due to exclusion of the introns in their study. In the study of the intronic genes, only exon portion was calculated, while introns were excluded in the evaluation of the AT/GC composition. In all the seven species under study, these ratios and percentages were found to be constant; suggesting the conserved pattern in their evolutionary course, which may be because of the study at the family level only, and it further requires analyzing at the higher taxonomic level for better understanding for the base composition in the plastome genomes.

Acknowledgements

The authors are grateful to Department of Science and Technology, Govt. of India for providing infrastructural facilities in the form of DST-FIST scheme to the Department.

References

1. Shinozaki K, Ohme M, Tanaka M, Wakasugi T, Hayashida N, Matshubayashi T, *et al.* The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. *EMBO J.* 1986; 5(9): 2043-2049
2. Shimada H & Sugiura M. Fine structural features of the chloroplast genome: comparison of the sequenced chloroplast genomes. *Nucleic Acids Res.* 1991; 19(5): 983-995
3. Sato S, Nakamura Y, Kaneko T, Asamizu E and Tabata S. Complete structure of the chloroplast genome of *Arabidopsis thaliana*. *DNA Res.* 1999; 6(5): 283-290
4. Robbens S, Derelle E, Ferraz C, Wuyts J, Moreau H & Deeper YV. The Complete Chloroplast and Mitochondrial DNA Sequence of *Ostreococcus tauri*: Organelle Genomes of the Smallest Eukaryote are Example of Compaction. *Mol Biol Evol.* 2007; 24(4): 956-968
5. Sharma BC and Ghimiray D. Analysis of the Cryptophyta chloroplast genome reveal presence of additional genes and absence of introns in their genome. *Biological Forum.* 2012; 4(2): 1-7
6. Asano T, Tsudzuki T, Takahashi S, Shimada H and Kadowaki K. Complete nucleotide sequence of the sugarcane (*Saccharum officinarum*) chloroplast genome: a

comparative analysis of four monocot chloroplast genomes. *DNA Res.* 2004; 11(2): 93-99

7. Tang J, Xia H, Cao M, Zhang X, Zeng W, Hu S, *et al.* A comparison of rice chloroplast genomes. *Plant Physiol.* 2004; 135(1): 412-420
8. Sasaki C, Lee SB, Daniell H, Wood TC, Tomkins J, Kim HG and Jansen RK. Complete chloroplast genome sequence of *Glycine max* and comparative analyses with other legume genomes. *Plant Mol. Biol.* 2005; 59(2): 309-322
9. Jansen RK, Wojciechowski MF, Sanniyasi E, Lee SB and Daniell H. Complete plastid genome sequence of the chickpea (*Cicer arietinum*) and the phylogenetic distribution of rps12 and clpP intron losses among legumes (Leguminosae). *Mol. Phylogenet. Evol.* 2008; 48(3): 1204-1217
10. Chang CC, Lin HC, Lin IP, Chow TY, Chen HH, Chen WH, *et al.* The chloroplast genome of *Phalaenopsis aphrodite* (Orchidaceae): comparative analysis of evolutionary rate with that of grasses and its phylogenetic implications. *Mol. Biol. Evol.* 2006; 23(2): 279-291
11. Daniell H, Lee SB, Grevich J, Sasaki C, Quesada-Vargas T, Guda C, *et al.* Complete chloroplast genome sequences of *Solanum bulbocastanum*, *Solanum lycopersicum* and comparative analyses with other Solanaceae genomes. *Theor. Appl. Genet.* 2006; 112(8):1503-1518
12. Ibrahim RI, Azuma J and Sakamoto M. Complete Nucleotide Sequence of the Cotton (*Gossypium barbadense* L.) Chloroplast Genome with a Comparative Analysis of Sequences among 9 Dicot Plants. *Genes Genet. Syst.* 2006; 81(5): 311-321
13. Lee SB, Kaittanis C, Jansen RK, Hostetler JB, Tallon LJ, Town CD and Daniell H. The complete chloroplast genome sequence of *Gossypium hirsutum*: organization and phylogenetic relationships to other angiosperms. *BMC Genomics.* 2006; 7:61
14. Wu CS, Wang YN, Hsu CY, Lin CP and Chaw SM. Loss of Different Inverted Repeat Copies from the Chloroplast Genomes of Pinaceae and Cupressophytes and Influence of Heterotachy on the Evaluation of Gymnosperm Phylogeny. *Genome Biol Evol.* 2011; 3:1284-1294

15. Yukawa M, Tsudzuki T and Sugiura M. The chloroplast genome of *Nicotiana sylvestris* and *Nicotiana tomentosiformis*: complete sequencing confirms that the *Nicotiana sylvestris* progenitor is the maternal genome donor of *Nicotiana tabacum*. *Mol. Genet. Genomics*. 2006; 275 (4): 367-373

16. Thyssen G, Svab Z and Maliga P. NCBI Genome Project. 2011. NIH, Bethesda, USA

17. Ghimiray D, Shah KB and Sharma BC. AACAL 1.2: A simple tool for the calculation and analysis of amino acids composition from polypeptides and proteins. Unpublished