STRUCTURAL ANALYSIS OF THE COMMON RAGWEED (*AMBROSIA ARTEMISIIFOLIA* L.) CHLOROPLAST GENOME

Erzsébet Nagy, János Taller*

Department of Plant Science and Biotechnology, University of Pannonia

Keszthely, Festetics Str. 7. H-8360 Hungary

**e-mail address:nagy.erzsebet@georgikon.hu*

Abstract

Chloroplast is an essential organelle in plants. Besides its key role in photosynthesis fatty acids and many amino acids are synthesized in chloroplasts. Due to the endosymbiotic origin chloroplasts have their own DNA, the cpDNA that called also plastome. The chloroplast genome has well defined structure and due to its conservative nature slow evolution is characteristic for it. In the present study we analyzed the plastome structure of common ragweed (*Ambrosia artemisiifolia*), the most widespread and highly allergenic weed in many parts in the world. Compared to other chloroplast genomes, the analysis indicates the differences in start codon of some genes, in gene content and structure, in intron content, as well as the repeated sequence motifs are identified, with special attention on the potential microsatellites. Results of this study can be utilized in future phylogenetic analyses and genotyping.

Keywords: common ragweed, chloroplast genome, plastome genes, repeats, microsatellites

Összefoglalás

A kloroplasztisz a növények létfontosságú sejtszervecskéje, mely a fotoszintézisben betöltött kulcsfontosságú szerepe mellett a zsírsavak és több aminosav bioszintézisének helye is egyben. Endoszimbionta származásának köszönhetően a kloroplasztisz saját DNS-el, a cpDNS-el rendelkezik, melyet plasztomnak is nevezünk. A kloroplasztisz genom meghatározott szerkezeti felépítése és konzervatív természete miatt lassú evolúció jellemzi. A jelen tanulmányban a világ számos régiója leggyakoribb gyomnövényének és legfőbb allergén növényének számító ürömlevelű parlagfű (*Ambrosia artemisiifolia*) plasztomjának szerkezetét elemeztük. Más kloroplasztisz genomokhoz hasonlítva a parlagfű plasztom egyes gének esetében eltérő start kódonnal rendelkezik, különbségek tapasztalhatók a géntartalomban és génszerkezetben, illetve az intron tartalomban. Továbbá meghatároztuk a parlagfű plasztom ismétlődő szekvencia motívumait, különös tekintettel a lehetséges mikroszatellitekre. A jelen tanulmány eredményei jól hasznosíthatók a továbbiakban filogenetikai és genotipizálási vizsgálatokban egyaránt.

Kulcsszavak: ürömlevelű parlagfű, kloroplasztisz genom, plasztom gének, ismétlődések, mikroszatellitek

Introduction

The chloroplast is an essential organelle of green plants that converts the energy of sunlight into energy storing organic molecules by using carbon dioxide and water. Besides capturing and utilizing the light-energy, in chloroplasts such essential molecules as the fatty acids and many of the amino acids are synthesized (Alberts et al., 2015). In today's agriculture chemicals

blocking either the photosynthetic or the biosynthetic processes which take place in chloroplast are widely used for weed control (Délye et al., 2013). However, chloroplasts have their own genome, processes in chloroplast are under control of the nuclear genome of the plant cell. Chloroplasts are able to divide in the plant cell, and generally they are inherited only maternally. A photosynthesizing plant cell contains about 100 chloroplasts, and each of them has 10-100 chloroplast DNA (cpDNA) molecules (Alberts et al., 2015). This high copy number on the one hand ensures high gene dosage and the possibility of variation that occasionally occurs in low copy number and is called heteroplasmy. On the other hand, since many genes of the chloroplast are essential for the life of the plant, the effect of mutations which are deleterious for the function of that gene will be alleviated or neutralized by the high number normal alleles (Birky, 2001).

The cpDNA or plastome, is generally visualized in circular form (Sakamoto and Takami, 2018). It has conservative structure with a determined order of genes. It is considered, that chloroplasts originate from a single endosymbiotic event, when a eukaryotic cell engulfed a photosynthetic cyanobacterium. This common ancestry, the highly conservative nature and the maternal inheritance are the reasons for the widespread use of cpDNA in phylogenetic studies (Alberts et al., 2015).

Recently, we published the chloroplast genome sequence of *Ambrosia artemisiifolia* (Nagy et al. 2017), the most widespread and highly allergenic weed in many parts in the world. In that article the main structure of the *A. artemisiifolia* plastome was determined and the gene content was characterized. In the present study we analyzed the gene structure and sequence repeats of the common ragweed cpDNA and identified the microsatellites of it. Our findings can be utilized in future evolutionary studies and chloroplast genome based genotyping.

Materials and methods

The assembled common ragweed chloroplast genome which we used in this study was uploaded and can be found at the NCBI (National Centre for Biotechnology Information, USA) MF362689.1 GenBank identification number.

The *A. artemisiifolia* plastome was compared to the *Helianthus annuus* cpDNA sequence.

Search for repeats was performed with the REPuter (Kurtz et al., 2001) and Tandem Repeat Finder (Benson, 1999) programs. The use of both programs was necessary, because the REPuter identifies just those repeats which are scattered on the sequence of analysis, while the Tandem Repeat Finder identifies repeats which are neighboring each other. For REPuter, the set parameters were as follows: Hamming distance: 3, number of hits 5000, minimum repeat length 30 bp, and all repeat types were marked. Results of REPuter were filtered with a self-developed program, called Clean Repeats, which we made freely accessible to the scientific community on the following link: http://cleanrepeats.georgikon.hu/. Tandem Repeat Finder was used with default parameters. For exploring the variability of the identified repeats plastome sequences of *Helianthus annuus* (NC_007977.1), *Arabidopsis thaliana* (NC_000932.1), *Artemisia argyi (*KM386991.1), *Ambrosia trifida* (MG029118.1) and of *Taraxacum officinale* (KU361241.1) were downloaded and sequence motives similar to *A. artemisiifolia* tandem repeats were identified.

The occurrence of repeats with length from 1 to 6 nucleotides, which are considered as potential microsatellites or SSR (simple sequence repeat) markers was investigated with the Msatcommander (Faircloth, 2008) and the WebSat (Martins et al., 2009) programs. The minimal number of repeats of mono-, di-, tri-, tetra-, penta- and hexanucleotides was adjusted to 10, 5, 4, 3, 3, 3 for both program. That means that a sequence motive should be constituted

from at least 10 mononucleotides, five dinucleotides, four trinucleotides or three tetra-, penta- or hexanucleotides to be recognized by either program as a microsatellite.

Results and discussion

Analysis of plastome genes of the common ragweed

In the common ragweed plastome from among the 80 protein coding genes three genes were identified which have different start codon from the ATG, which is the start codon of eukaryotic organisms. These start codons and genes are the followings: ACG in *ndhD*, ATC in *rpl16* and GTG in *rps19*. In other plants similar phenomenon was already reported for the *atpE, psaI, rps11, petB* (Wiegert et al., 2012), *rps19* (Kim et al., 2014), *ndhD, psbL, rps19* (Curci et al., 2015) chloroplast genes. These alternative start codons indicate to the prokaryotic origin of the chloroplast.

Compared to *A. thaliana*, the model plant in plant genetics, *A. artemisiifolia* has two more protein coding genes. These are the *infA*, which is a translation initiation factor, and the *ycf15*, which codes for a protein with unknown function.

In the IRA and IRB regions six protein coding genes (*rpl2, rpl23, ycf2, ndhB, rps17, ycf15*) as well as the 3'end of the *rps12* gene are repeated.

Compared to the *H. annuus* reference genome length differences in 11 common ragweed plastome genes were detected. Among these, the *trnT-GGU, trnA-UGC, psbH, rpoC2, rps16,* and the *ycf2* genes were larger, while the *rrn23, ccsA, matK, rpoA* and the *ycf1* genes were smaller in *A. artemisiifolia* than in sunflower. Length variations in these chloroplast genes were already reported in several studies. Timme et al. (2007) described sequence variations in the sunflower *ccsA, matK, psbH, ycf1* genes when compared to *Lactuca sativa*. In rice, *rpoC2*

length variations were detected (Shimada et al., 1990). Further, comparing 84 chloroplast protein and rRNA coding genes of 17 plant species significant sequence variations were registered for the *ycf2, rps16, rpoA* and *psbH* genes (Kim and Lee, 2004). Due to their length variation in different plant species it is suggested to analyze the usefulness of the above genes of *A. artemisiifolia* in phylogenetic and evolutionary studies. Whether the length variation of these genes is affecting their function needs investigation.

One intron was detected in the *atpF, petB, petD, ndhA, ndhB, rpl2, rps16, rpoC1* protein coding genes and in the *trnA-UGC, trnG-UCC, trnI-GAU, trnK-UUU, trnL-UAA, trnV-UAC* tRNA genes of the common ragweed. Two introns were detected in the *ycf3* and *clpP* genes. *A. thaliana* plastome has three more intron containing genes, the *rpl12, rpl16* and *rps12* genes, than the *A. artemisiifolia* (Sato et al., 1999), while *H. annuus* has one more intron containing gene, the *rps12* (Timme et al., 2007).

Further, the *trnE-UUC* gene of *A. artemisiifolia* is missing from the *H. annuus* plastome. The presence/absence of this gene is very variable in species of the *Asteraceae* family (Wang et al., 2018, Kim et al., 2017, Chen et al., 2018).

It is concluded, that except the above the *A. artemisiifolia* protein coding plastome genes do not show significant structural differences compared to related species or to *A. thaliana*.

Detection of repeats in the common ragweed plastome

The REPuter program with the adjusted parameters provided more than 2000 hits. Nevertheless, these hits were overwhelmingly overlapping repeats. To eliminate overlaps, we have created a program called Clean Repeats (http://cleanrepeats.georgikon.hu/) that will be published elsewhere. The filtered REPuter results were combined with results of the Tandem Repeat Finder program. In total 67 repeats were detected, from which one third is scattered throughout

of the plastome, and the two-third is constituted from neighboring repeats, which follow each other in the plastome.

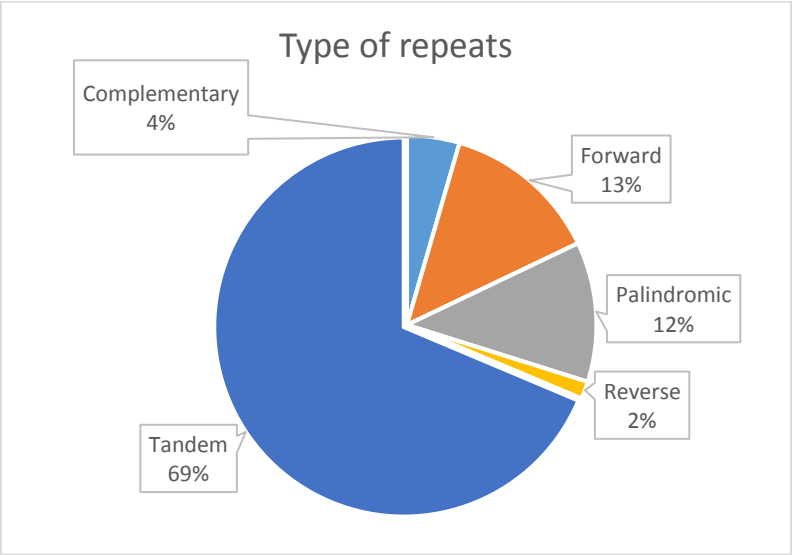

Figure 1: Distribution of the five repeat types in plastome of the common ragweed. Values are calculated from the combined results of the Tandem Repeat Finder and the filtered REPuter program.

Most of the repeats, 69%, belong to the tandem type, while the percentage of forward and palindromic repeats is just slightly above 10%. The reverse and complementary repeats have the lowest proportion with less than 5%. (Fig. 1.). Most repeats (52%) belong to the 15-30 bp long category (Fig. 2.). The number of smaller (<15 bp) and larger (>30 bp) repeats was almost similar. Besides the two inverted repeat regions two repeats above 100 bp were detected.

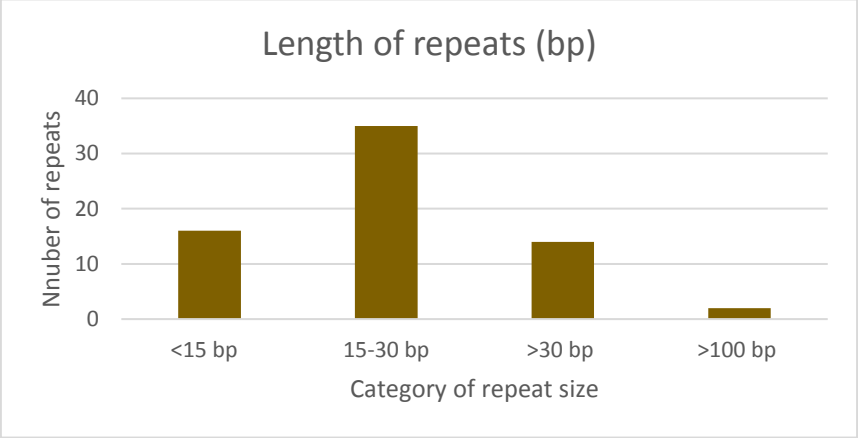

Figure 2: Repeat size categories in plastome of the common ragweed. Values are calculated from the combined results of the Tandem Repeat Finder and the filtered REPuter program.

However, the LSC region accounts for about 55% of the common ragweed plastome, 48 repeats (72%) occurred in the LSC region and three in LSC – inverted repeat region junctions (Fig. 3.).

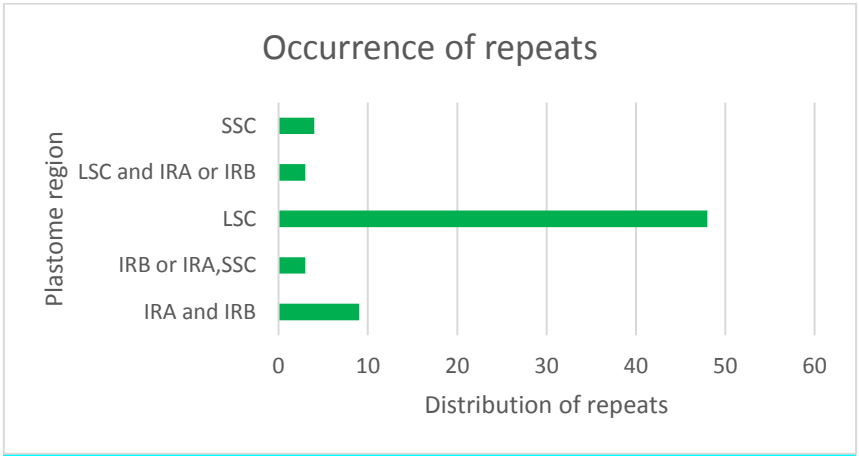

Figure 3: Number of repeats in different plastome regions of the common ragweed. Values are calculated from the combined results of the Tandem Repeat Finder and the filtered REPuter program. SSC (Small Single-Copy region, LSC (Large Single-Copy region), IRA and IRB (Inverse Repeat regions A and B)

Regarding the distribution, 47 repeats (70%) were found in intergenic spacer regions, and just six (9%) were identified in exons and three repeats (4%) in introns (Fig.4.). Interestingly, seven

repeats (10%) occurred in intergenic spacer region – intron junctions, and four (6%) in intergenic spacer region – exon junctions.

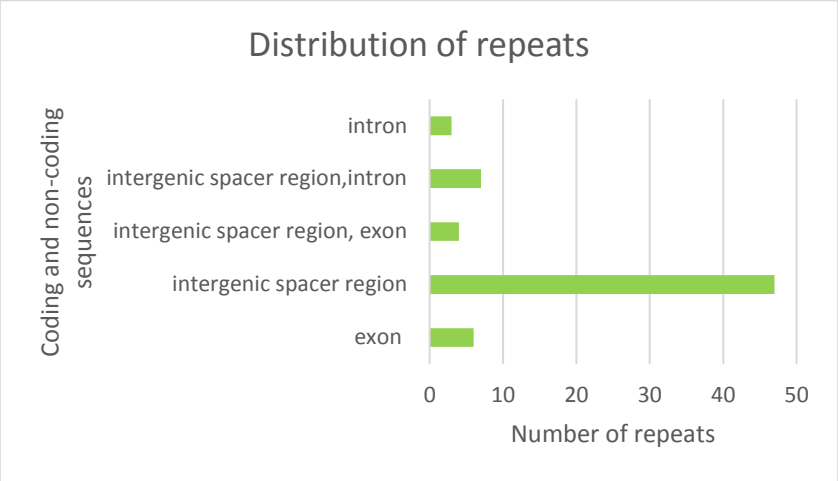

Figure 4: Number of repeats in different types of coding and non-coding sequences in plastome of the common ragweed. Values are calculated from the combined results of the Tandem Repeat Finder and the filtered REPuter program.

Comparing tandem repeats of *Ambrosia artemisiifolia*, *Helianthus annuus*, *Arabidopsis thaliana*, *Artemisia argyi*, *Ambrosia trifida* and *Taraxacum officinale* sequence similarity was detected in 45 cases. Only the *ycf2* gene occurred in all of these species with at least 70% similarity. Except *A. thaliana* repeats in the *atpI-atpH*, *rbcL-accD* and *rrn5-rrn4.5* intergenic non-coding regions and in *rps8* gene were common among the analyzed species. Except one of the remaining repeats of *A. artemisiifolia* all repeats could be found in *A. trifida*, four in *H. annuus*, four in *T. officinale*, and three in *A. argyi* while in *A. thaliana* no similar motif of these repeats could be detected. These results are in harmony with the evolutionary distances of *A. artemisiifolia* and the analyzed species, and reveal the variability of tandem repeats and their applicability in evolutionary studies.

Identification and characterization of microsatellites

With the adjusted parameters 43, as well as 53 microsatellites were detected with the Msatcommander and with the WebSat program, respectively (Table 1.).

Table 1. List of microsatellites in the common ragweed chloroplast genome

Start position (bp)	**Repeat motive**	**Repeat number**	**Repeat type**	**Position of the microsatellite**			**program**
41	A	10	mono	*rpl2-trnH-GUG*	IGS	LSC	W, M
2147	T	17	mono	*trnK-UUU-matK*	IGS	LSC	W, M
2490	T	11	mono	*matK*	exon	LSC	W, M
4696	A	30	mono	*trnK-UUU-rps16*	IGS	LSC	W, M
4897	A	77	mono	*trnK-UUU-rps16*	IGS	LSC	W, M
11220	T	22	mono	*psbM-trnD-GUC*	IGS	LSC	W, M
13129	T	16	mono	*trnE-UUC-rpoB*	IGS	LSC	W, M
13251	T	16	mono	*trnE-UUC-rpoB*	IGS	LSC	W, M
13940	A	10	mono	*rpoB*	exon	LSC	W, M
17170	A	11	mono	*rpoC1*	intron	LSC	W, M
17714	G	10	mono	*rpoC1*	intron	LSC	W, M
19025	A	10	mono	*rpoC1*	exon	LSC	W, M
21485	T	10	mono	*rpoC2*	exon	LSC	W, M
26699	T	10	mono	*atpI-atpH*	IGS	LSC	W, M
28640	T	18	mono	*atpF-atpA*	IGS	LSC	W, M
30438	A	11	mono	*trnR-UCU-trnG-UCC*	IGS	LSC	W, M
30792	A	12	mono	*trnG-UCC*	intron	LSC	W, M
35374	T	18	mono	*psbC-trnS-UGA*	IGS	LSC	W, M
44414	A	20	mono	*ycf3*	intron	LSC	W, M

47168	A	20	mono	*trnT-UGU-trnL-UAA*	IGS	LSC	W, M
54715	T	25	mono	*atpB-rbcL*	IGS	LSC	W, M
70612	T	23	mono	*clpP*	intron	LSC	W, M
71234	A	13	mono	*clpP*	intron	LSC	W, M
74255	A	10	mono	*psbT-psbN*	IGS	LSC	W, M
74837	A	21	mono	*psbH-petB*	IGS	LSC	W, M
75450	A	36	mono	*petB*	intron	LSC	W, M
78191	T	10	mono	*petD-rpoA*	IGS	LSC	W, M
80639	T	36	mono	*rps8-rpl14*	IGS	LSC	W, M
81171	T	15	mono	*rpl14-rpl16*	IGS	LSC	W, M
84349	T	10	mono	*rps19-rpl2*	IGS	LSC, IRA	W, M
109673	T	12	mono	*ycf1*	exon	SSC	W, M
110286	A	11	mono	*ycf1*	exon	SSC	W, M
111364	A	16	mono	*ycf1*	exon	SSC	W, M
112589	A	10	mono	*ycf1*	exon	SSC	W, M
120420	T	11	mono	*psaC-ndhD*	IGS	SSC	W, M
122012	A	13	mono	*ndhD-ccsA*	IGS	SSC	W, M
124427	A	12	mono	*rpl32-ndhF*	IGS	SSC	W, M
19249	AT	5	di	*rpoC1*	exon	LSC	W, M
20248	AT	5	di	*rpoC2*	exon	LSC	W, M
46687	TA	5	di	*rps4-trnT-UGU*	IGS	LSC	W
59205	TA	5	di	*accD-psaI*	IGS	LSC	W
117148	TA	5	di	*ndhA*	intron	SSC	W
35161	TTC	4	tri	*psbC*	exon	LSC	W
60959	ATT	4	tri	*ycf4-cemA*	IGS	LSC	W, M
109561	AGA	5	tri	*ycf1*	exon	SSC	W
109563	AAG	4	tri	*ycf1.*	exon	SSC	M
4882	ATAA	3	tetra	*trnK-UUU-rps16*	IGS	LSC	W

63329	TTCT	3	tetra	*petA-psbJ*	IGS	LSC	W
63331	CTTT	3	tetra	*petA-psbJ*	IGS	LSC	M
67995	TATT	3	tetra	*rpl33-rps18*	IGS	LSC	W
117209	TATC	3	tetra	*ndhA*	intron	SSC	W
121797	TATT	3	tetra	*ndhD*	exon	SSC	W
123754	ATTT	3	tetra	*trnL-UAG-rpl32*	IGS	SSC	W, M
68056	AACCA	3	penta	*rpl33-rps18*	IGS	LSC	W
56532	GGATAA	3	hexa	*rbcL*	exon	LSC	W

IGS= Intergenic spacer region; LSC= Large Single Copy; SSC= Small Single Copy; IRA, IRB=Inverted Repeat A, B; W=WebSat; M= Msatcommander

Mononucleotide microsatellites accounted for 86% and 70% of all SSR-s with the Msatcommander and WebSat program, respectively (Fig. 5. and 6.).

However, penta- and hexanucleotide type microsatellites were not detected by the Msatcommander (Fig. 5.), with the WebSat one of each of these type SSR-s could be detected (Fig. 6.). The difference between the numbers of detected microsatellites is due to the differing algorithm of the two programs.

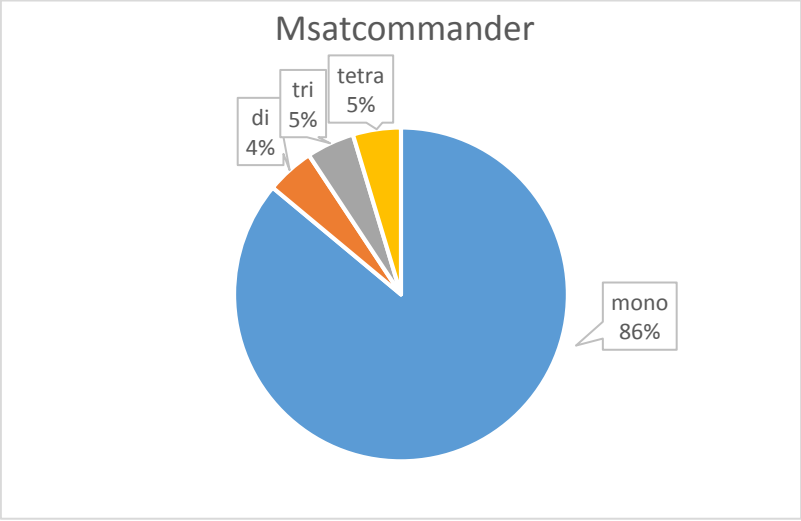

Figure 5: Distribution of microsatellites detected with the Msatcommander program in plastome of the common ragweed.

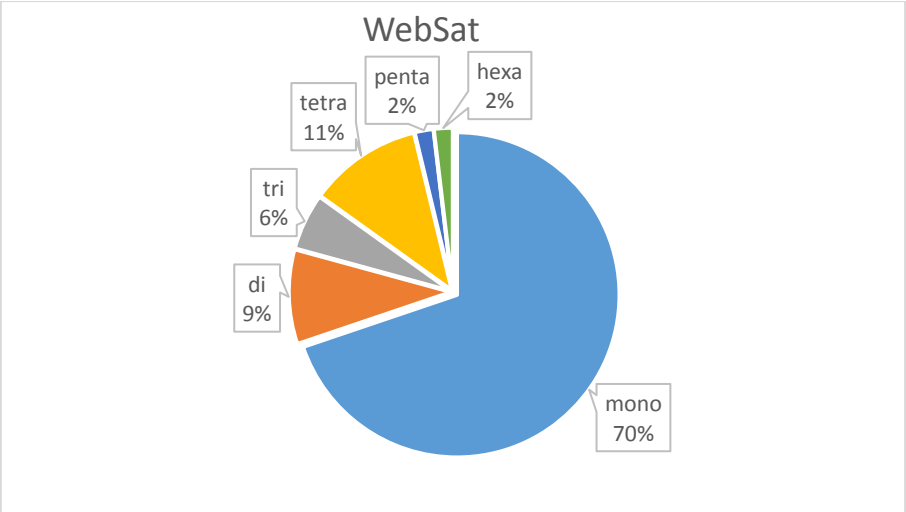

Figure 6: Distribution of microsatellites detected with the WebSat program in plastome of the common ragweed.

With one exception, only A (adenine) and T (thymine) mononucleotide microsatellites were found in the common ragweed plastome. Similar observation was already reported for other *Asteraceae* species, such as artichoke (Curci et al., 2015), *Jacobaea vulgaris* (Doorduin et al., 2011), and *Artemisia frigida* (Liu et al., 2013).

The microsatellites reported in the present publication could be efficient tools in future phylogenetic analyses and genotyping studies.

Acknowledgements

This research was supported by the Hungarian Government and the European Union, with the co-funding of the European Regional Development Fund in the frame of Széchenyi 2020 Programme GINOP-2.3.2-15-2016-00054 project.

References

Alberts, B., Johnson, A., Lewis, J. et al. 2015. Molecular Biology of the Cell. 6th edition. New York, Garland Science. Energy Conversion: Mitochondria and Chloroplasts. 753-812.

Benson, G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research*. **27**. (2). 573-580.

Birky, C. W. 2001. The Inheritance of Genes in Mitochondria and Chloroplasts: Laws, Mechanisms, and Models. *Annu. Rev. Genet*. **35**. 125–148.

Chen, X., Zhou, J., Ciu, Y., Wang, Y., Duan, B., Yao, H. 2018. Identification of *Ligularia* herbs using the complete chloroplast genome as a super-barcode. *Front. Pharmacol*. **9**. 695.

Curci, P.L., De Paola, D., Danzi, D., Vendramin, G.G., Sonnate, G. 2015. Complete chloroplast genome of the multifunctional crop globe artichoke and comparison with other *Asteraceae*. *Plos One*. **10.** (3). e0120589.

Délye, C., Jasieniuk, M., Le Corre, V. 2013. Deciphering the evolution of herbicide resistance in weeds. *Trends Genet.* **29**. 649–658.

Doorduin, L., Gravendeel, B., Lammers, Y., Ariyurek, Y., Chin-A-Woeng, T., Vrieling, K. 2011. The complete chloroplast genome of 17 individuals of pest species *Jacobaea vulgaris*: SNPs, microsatellites and barcoding markers for population and phylogenetic studies. *DNA Research*. **18**. 93-105.

Faircloth, B. 2008. Msatcommander: detection of microsatellite repeat arrays and automated, locus-specific primer design. *Molecular Ecology Resources*. **8**. 92-94.

Kim, H.T., Chung, M.G., Kim, K.-J. 2014. Chloroplast genome evolution in early diverged Leptosporangiate ferns. *Mol. Cells*. **37**. (5). 372-382.

Kim, K.-J., Lee, H.-L. 2004. Complete chloroplast genome sequence from Korean ginseng (*Panax schinseng* Nees) and comparative analysis of sequence evolution among 17 vascular plants. *DNA Research*. **11**. 247-261.

Kim, I., Park, J. Y., Lee, Y.S., Lee, H.O., Park, H.-S., Jayakodi, M., Waminal, N.E., Kang, J.H., Lee, T.J., Sung, H.S., Kim, K.Y., Yang, T.-J. 2017. Discrimination and authentication of *Eclipta prostrata* and *E. alba* based on the complete chloroplast genomes. *Plant Breed. Biotech*. **5**. (4). 334-343.

Kurtz, S., Choudhuri, J.V., Ohlebusch, E., Schleiermacher, C., Stoye, J., Giergerich, R. 2001. REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Research*. **29**. (22). 4633-4642.

Liu, Y., Huo, N., Wang, Y., Zhang, S., Young, H.A., Feng, X., Gu, Y.Q. 2013. Complete chloroplast genome sequences of mongolian medicine *Artemisia frigida* and phylogenetic relationships with other plants. *Plos One*. **8**. (2). e575333.

Martins, W.S., Lucas, D.C.S., de Sousa Neves, K.F., Bertioli, D.J. 2009. WebSat- a web software for microsatellite marker development. *Bioinformation*. **3**. (6). 282-283.

Nagy, E., Hegedűs, G., Taller, J., Kutasy, B., Virág, E. 2017. Illumina sequencing of the chloroplast genome of common ragweed (*Ambrosia artemisiifolia* L.). *Data in Brief*. **15**. 606-611.

Sakamoto, W., Takami, T. 2018. Chloroplast DNA Dynamics: Copy Number, Quality Control and Degradation. *Plant & Cell Physiology*. **59.** (6). 1120–1127.

Sato, S., Nakamura, Y., Kaneko, T., Asamizu, E., Tabata, S. 1999. Complete structure of the chloroplast genome of *Arabidopsis thaliana. DNA Reasearch*. **6**. 283-290.

Shimada, H., Fukuta, M., Ishikawa, M., Sugiura, M. 1990. Rice chloroplast RNA polymerase genes: the absence og an intron in *rpoC1* and the presence of an extra sequence in *rpoC2*. *Molecular and General Genetics MGG*. **221**. (3). 395-402.

Timme, R.E., Kuehl, J.L., Boore, J.L., Jansen, R.K. 2007. A comparative analysis of the *Lactuca* and *Heliathus* (*Asteracea*) plastid genomes: identification of divergent regions and categorization of divergent regions and categorization of shared repeats. *American Journal of Botany*. **94**. (3). 302-312.

Wang, X.-Y., Zhou, Z.-S., Liu, G., Quian, Z.-Q. 2018. Characteritation of the complete chloroplast genome of the invasive weed *Galinsoga quadriradiata* (Asterales: *Asteraceae*). *Consevation Genet. Resour*. **10**. 89-92.

Wiegert, K.E., Benett, M.S., Triemer, R.E. 2012. Evolution of the chloroplast genome in photosynthetic euglenoids: a comparison of *Eutreptia viridis* and *Euglena gracilis* (*Eugleniphyta*). *Protist*. **163**. 832-843.