https://doi.org/10.17221/562/2020-PSE

Identification of the optimal codons for acetolactate synthase from weeds: an *in-silico* study

Madhab Kumar Sen[1], Kateřina Hamouzová[1], Sunil Kanti Mondal[2], Josef Soukup[1]*

[1]Department of Agroecology and Crop Production, Faculty of Agrobiology, Food and Natural Resources, Czech University of Life Sciences Prague, Prague, Czech Republic
[2]Department of Biotechnology, The University of Burdwan, Burdwan, West Bengal, India
**Corresponding author: hamouzova@af.czu.cz*

Citation: Sen M.K., Hamouzová K., Mondal S.K., Soukup J. (2021): Identification of the optimal codons for acetolactate synthase from weeds: an *in-silico* study. Plant Soil Environ., 67: 331–336.

Abstract: Although various studies of codon usage bias have been reported in a broad spectrum of organisms, no studies to date have examined codon usage bias for herbicide target genes. In this study, we analysed codon usage patterns for the acetolactate synthase (ALS) gene in eight monocot weeds and one model monocot. The base composition at the third codon position follows C3 > G3 > T3 > A3. The values of the effective number of codons (ENC or Nc) indicate low bias, and ENC or Nc *vs*. GC3 plot suggests that this low bias is due to mutational pressure. Low codon adaptation index and codon bias index values further supported the phenomenon of low bias. Additionally, the optimal codons, along with over- and under-represented codons, were identified. Gene design using optimal codons rather than overall abundant codons produce improved protein expression results. Our results can be used for further studies, including eliciting the mechanisms of herbicide resistance (occurring due to elevation of gene expression levels) and the development of new compounds, their efficiency and risk assessment for herbicide resistance evolution.

Keywords: herbicide-resistant weeds; heterologous gene expression; primer designing; recombinant *ALS* protein

Weeds compete with the major crop, thus reducing their yield and productivity. Economically, they can be regarded as a more damaging agent than other crop pests (in several situations). Globally, herbicide resistance has been documented in a wide range of weed species. Acetolactate synthase (ALS) catalyses the first step in the synthesis of the branched-chain amino acids (Duggleby et al. 2008, Hamouzová et al. 2014) and is the target for a large number of herbicides. Continuous use of the same herbicide with the same mode of action has allowed for the selection of weed populations resistant to the overused herbicide or mechanism of herbicide action. The widespread evolution of multiple-herbicide resistance in weedy species makes their control more difficult. The most common mechanisms of evolving resistance to herbicides by plants include metabolic changes, mutations in the DNA of the target gene and overexpression of the target protein (Jugulam and Shyam 2019, Murphy and Tranel 2019). Unfortunately, for nearly 20 years, no new mode of action has been introduced into the market. Furthermore, with the release of glyphosate-resistant crops, the efforts for herbicide discovery reduced significantly (Powles and Yu 2010).

Due to their favourable effects on the efficiency and accuracy of the translation, certain codons are preferred over the others, leading to differential codon usage patterns (codon-usage bias) (Je et al. 2019). Optimal codons contribute to the accuracy as well as the speed of the translation elongation (Wright 1990). Thus, it is very useful to know the rules which govern the synonymous codon selection of the target gene. This knowledge can be extremely useful to design a heterologous gene, having the most efficient expressional efficiency. Apart from playing important roles in various physiological processes,

Supported by the National Agency for Agricultural Research, Project No. QK1820081.

https://doi.org/10.17221/562/2020-PSE

codon bias and codon optimisation, finds huge applications in industrial biotechnology, whose major goal is to produce recombinant proteins. Even though heterologous gene expression studies are among the most well-appreciated studies related to specific protein interactions, efficiency in the expression of heterologous genes remains the most challenging part. Codon usage patterns in non-model plants, especially weeds, are not well understood, mainly due to limitations in data availability. The development of new compounds with herbicidal properties along with an assessment of their efficiencies and risk of resistance may require designing synthetic genes based on their codon usage patterns. Synthetic genes find important applications in heterologous gene expression experiments. Heterologous gene expression studies can be very useful in basic biological research areas, including protein interactions studies and the development of new herbicidal compounds (Quax et al. 2015, Zhou et al. 2016).

In this study, we analysed codon usage patterns for ALS in eight monocot weeds (*Alopecurus myosuroides* Huds., *Apera spica-venti* L., *Beckmannia syzigachne* L., *Bromus tectorum* L., *Echinochloa crus-galli* (L.) P. Beauv., *Echinochloa oryzicola* (Vasinger) Ohwi, *Poa annua* L. and *Lolium rigidum* Gaud.) and one model monocot (*Zea mays* L.). Relative synonymous codon usage (RSCU), codon adaptation index (CAI), codon bias index (CBI), the effective number of codons (ENC or Nc), positional GC contents and CG dinucleotide suppression values, were analysed for ALS-coding sequences. Optimal codons, along with over- and under-represented codons, were identified. Our results would help in additional investigations with *acetolactate synthase* gene, including eliciting the mechanisms of herbicide resistance (occurring due to elevation of gene expression levels) and development of new herbicidal compounds with synthetic genes based on their codon usage pattern.

MATERIAL AND METHODS

Retrieval of sequences. Full-length *ALS* coding sequences were retrieved from the National Center for Biotechnology Information website (https://www.ncbi.nlm.nih.gov/). The list of organisms and their accession numbers are in Table 1.

Base composition parameters. Total GC content, the GC content at the different positions of a codon (GC1, GC2 and GC3) were calculated using MEGAX (Kumar et al. 2018). Because of the degenerate property of the codons, the 3rd position of a codon (also called as wobble position) has less discriminatory for an amino acid than the other two bases (Elhaik and Tatarinova 2012). Moreover, CG dinucleotides are the potential target sites for methylation (Elhaik et al. 2014). The XCG/XCC ratio based on RSCU values was used to calculate CG dinucleotide suppression values (Mazumdar et al. 2017).

Analysis of codon usage for acetolactate synthase in weeds. MEGAX and CodonW (http://codonw.sourceforge.net), with in-house PERL script and standard genetic codon table, were used for computing the RSCU values. RSCU value of 1 indicates no codon usage bias, while values above and below 1 indicate codons are utilised more and less frequently (respectively) than expected (Mondal et al. 2016). Based on the RSCU values, optimal codons and their frequencies (F_{op}) were calculated. CAI and CBI values were determined using CodonW. CAI values range from 0 (random codon usage) to 1 (extreme codon bias) (Mondal et al. 2016). CBI values measure the extent to which a gene uses the set of its optimal codons. CBI values range between 0 (random codon usage) and 1 (extreme codon bias) (Bennetzen and Hall 1982).

GC3 *vs.* the effective number of codons (ENC or Nc) plot. Mutational pressure and natural selection are the two well-known core factors responsible for codon biasness. Plot between GC3 and expected the effective number of codons can be a good measure to determine that among the mentioned core factors, which factor is the driving force (Sharp et al. 1993). ENc values

Table 1. List of organisms used for analysis along with their GenBank accession numbers

Organism	Accession number (CDS)
Zea mays ALS1	NCVQ01000006.1
Z. mays ALS2	NM_001148702.2
Apera spica-venti	JN646110.1
Bromus tectorum	MK492423.1
Alopecurus myosuroides	AJ437300.2
Beckmannia syzigachne	MG891930.1
Lolium rigidum	MK492446.1
Echinochloa crus-galli ALS1	KY071206.1
E. crus-galli ALS2	KY071207.1
E. crus-galli ALS3	KY071208.1
E. oryzicola ALS1	KY071209.1
E. oryzicola ALS2	KY071210.1
Poa annua ALSa	KT346395.1
P. annua ALSb	KT346396.1

https://doi.org/10.17221/562/2020-PSE

Table 2. Nucleotide composition for acetolactate synthase of the weeds of interest

Organism		T-3	C-3	A-3	G-3	GC1	GC2	GC3	GC	Nc	Codon adaptation index	Codon Bias index
		(%)										
Apera spica-venti		23.2	35.4	10.8	30.7	58.3	63.2	45.7	66	55.9	0.3	0.1
Bromus tectorum		20.5	37.2	11.3	31	59.1	63.6	45.4	68.2	54.6	0.3	0.1
Alopecurus myosuroides		23.6	34.3	12.8	29.3	57.4	62.7	45.9	63.7	57.3	0.2	0.1
Beckmannia syzigachne		23.3	35.8	11.3	29.7	57.9	62.3	45.8	65.5	56.3	0.3	0.1
Lolium rigidum		18.7	42	8.6	30.8	60.6	63	46	72.8	51.4	0.3	0.1
Echinochloa crus-galli	*ALS1*	18.8	40.6	9.3	31.3	60.6	63.3	46.7	71.9	52.1	0.3	0.1
	ALS2	18.5	40.4	9.3	31.8	60.5	63.2	46.1	72.2	51.9	0.3	0.1
	ALS3	19.3	39.8	8.7	32.2	60.7	63.6	46.5	72	52	0.3	0.2
Zea mays	*ALS1*	21.1	37.1	10	31.8	59.3	63.5	45.4	68.9	54.3	0.2	0.1
	ALS2	22.8	36.5	11.6	29.1	57.9	62.8	45.4	65.6	56.2	0.2	0.1
Echinochloa oryzicola	*ALS1*	18.8	40.4	9.3	31.4	60.4	63.1	46.2	71.9	52.2	0.3	0.1
	ALS2	18.5	40.4	9.2	32	60.6	63.2	46.3	72.4	51.7	0.3	0.1
Poa annua	*ALSa*	22.4	36.9	8.9	31.8	59	62.8	45.5	68.7	54.3	0.3	0.1
	ALSb	19.7	40.1	9.0	31.2	60.0	62.6	46.1	71.3	52.5	0.3	0.1

Nc – number of codons

were calculated from GC3s under the null hypothesis (i.e., no selection) according to the given equation by Wright (1990). The values of ENc (or Nc) might vary from extreme (20) to least bias (61) (Mondal et al. 2016).

RESULTS AND DISCUSSIONS

Base compositional parameters and correlation analysis. For these species, the base composition at the third codon position follows C3 > G3 > T3 > A3 (Table 2). Analysis of XCG/XCC ratio showed values of 0.4 (*A. spica-venti* L., *Z. mays* L. *ALS2* and *P. annua* L. *ALSb*), 0.5 (*B. tectorum* L., *A. myosuroides* Huds., *B. syzigachne* L., *E. crus-galli* (L.) P. Beauv., *E. oryzicola* (Vasinger) Ohwi, *P. annua* L. *ALSa* and *Z. mays* L. *ALS1*) and 0.6 (*L. rigidum* Gaud.). Except for *A. spica-venti*, *Z. mays* L. *ALS2* and *P. annua* L. *ALSb*; our results indicate moderate CG dinucleotide suppression. The values of Nc varied from 51.4 to 57.3. This indicates weak bias. Moreover, the Enc (or Nc) *vs.* GC3 plot suggests that the low bias might be due to mutational pressure and not a natural selection (Figure 1). The CAI values ranged from 0.2–0.3 (i.e., random codon usage). Low CBI values further supported the fact of random codon usage (Table 2).

Analysis of relative synonymous codon usage and determination of optimal codons. To analyse the codon usage patterns in the selected weed species, RSCU values were calculated. Based on the cluster analysis, primarily two clusters were formed: one with maize ALS and the other with the rest species. Within the cluster containing weedy species ALS, two further sub-clusters were formed. ALS from *B. tectorum* L. and *L. rigidum* Gaud. showed similar patterns with the ALS from the *Echinochloa* sp. (Figure 2). In all the cases, the number of codons having an RSCU value less than 1 is found to be greater than the codons having an RSCU value higher than optimum. Nine codons (GAA, GGA, AAA, UUA, CUA, CAA, AGA, CGA and GUA) were under-represented in all cases, whereas CAG, CGC and UCC were over-represented in all cases. Interestingly, all the nine

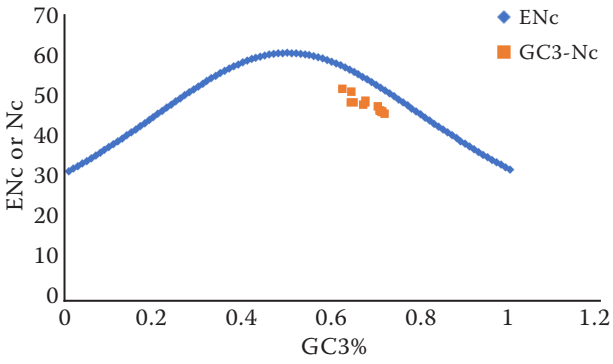

Figure 1. GC3 vs. number of codons (ENC or Nc) plot. The acetolactate synthase (ALS) genes which are positioned on or close to the curve line are considered to be under mutational pressure

https://doi.org/10.17221/562/2020-PSE

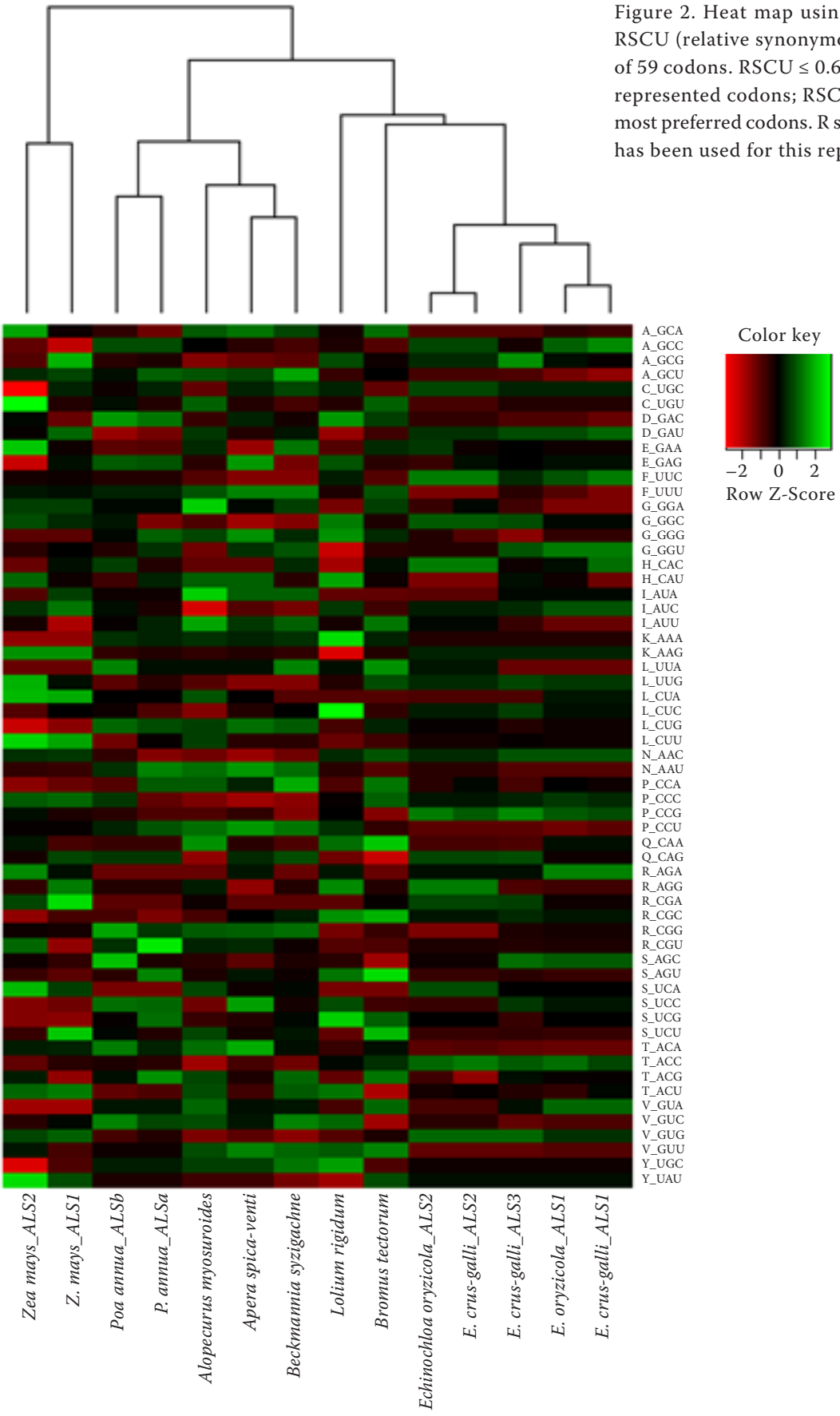

Figure 2. Heat map using raw Z-score of RSCU (relative synonymous codon usage) of 59 codons. RSCU ≤ 0.6 indicates under-represented codons; RSCU ≥ 1.6 indicates most preferred codons. R statistical software has been used for this representation

https://doi.org/10.17221/562/2020-PSE

Table 3. Optimal codons and their frequencies

Organism	A	F_{op}	C	F_{op}	D	F_{op}	E	F_{op}	F	F_{op}	G	F_{op}
Apera spica-venti	GCC	0.41	TGC	0.80	GAC	0.53	GAG	0.91	TTT	0.57	GGC	0.28
Bromus tectorum	GCC	0.39	TGC	0.67	GAC	0.55	GAG	0.82	TTT	0.52	GGC	0.35
Alopecurus myosuroides	GCC	0.43	TGC	0.67	GAT	0.52	GAG	0.82	TTT	0.52	GGC	0.33
Beckmannia syzigachne	GCC	0.40	TGC	0.83	GAC/GAT	0.50	GAG	0.79	TTT	0.57	GGC	0.30
Lolium rigidum	GCC	0.42	TGC	0.80	GAC	0.60	GAG	0.88	TTC	0.59	GGC	0.43
Echinochloa crus-galli ALS1	GCC	0.49	TGC	0.80	GAT	0.55	GAG	0.85	TTC	0.70	GGC	0.37
E. crus-galli ALS2	GCC	0.46	TGC	0.83	GAT	0.52	GAG	0.85	TTC	0.70	GGC	0.42
E. crus-galli ALS3	GCC	0.42	TGC	0.80	GAT	0.53	GAG	0.84	TTC	0.61	GGC	0.41
Zea mays ALS1	GCC	0.34	TGC	0.80	GAT	0.55	GAG	0.85	TTC	0.55	GGC	0.39
Z. mays ALS2	GCC	0.39	TGC/TGT	0.50	GAC	0.52	GAG	0.75	TTC	0.54	GGC	0.41
Echinochloa oryzicola ALS1	GCC	0.48	TGC	0.80	GAT	0.53	GAG	0.85	TTC	0.65	GGC	0.37
E. oryzicola ALS2	GCC	0.46	TGC	0.83	GAT	0.52	GAG	0.82	TTC	0.70	GGC	0.42
Poa annua ALSa	GCC	0.46	TGC	0.80	GAC	0.58	GAG	0.88	TTC	0.52	GGC/GGG/GGT	0.30
P. annua ALSb	GCC	0.47	TGC	0.75	GAC	0.60	GAG	0.89	TTC	0.52	GGC	0.38
Organism	H	F_{op}	I	F_{op}	K	F_{op}	L	F_{op}	N	F_{op}	P	F_{op}
Apera spica-venti	CAC	0.69	ATC	0.52	AAG	0.80	CTG	0.40	AAC/AAT	0.50	CCA	0.33
Bromus tectorum	CAC	0.77	ATC	0.54	AAG	0.80	CTG	0.35	AAC	0.82	CCC	0.38
Alopecurus myosuroides	CAC	0.69	ATC/ATT	0.44	AAG	0.79	CTG	0.37	AAC	0.56	CCA	0.35
Beckmannia syzigachne	CAC	0.79	ATC	0.50	AAG	0.79	CTG	0.39	AAC	0.56	CCA	0.39
Lolium rigidum	CAC	0.64	ATC	0.61	AAG	0.71	CTC	0.44	AAC	0.76	CCC	0.31
Echinochloa crus-galli ALS1	CAC	0.83	ATC	0.63	AAG	0.83	CTC/CTG	0.31	AAC	0.82	CCC	0.35
E. crus-galli ALS2	CAC	0.85	ATC	0.59	AAG	0.83	CTC/CTG	0.31	AAC	0.76	CCC	0.33
E. crus-galli ALS3	CAC	0.75	ATC	0.60	AAG	0.83	CTC	0.33	AAC	0.82	CCC	0.34
Zea mays ALS1	CAC	0.77	ATC	0.65	AAG	0.88	CTC	0.30	AAC	0.79	CCC	0.39
Z. mays ALS2	CAC	0.69	ATC	0.61	AAG	0.88	CTC	0.25	AAC	0.78	CCC	0.38
Echinochloa oryzicola ALS1	CAC	0.77	ATC	0.63	AAG	0.83	CTC/CTG	0.31	AAC	0.82	CCC	0.36
E. oryzicola ALS2	CAC	0.85	ATC	0.59	AAG	0.83	CTC/CTG	0.31	AAC	0.76	CCC	0.33
Poa annua ALSa	CAC	0.73	ATC	0.56	AAG	0.80	CTG	0.38	AAC	0.53	CCA	0.35
P. annua ALSb	CAC	0.80	ATC	0.58	AAG	0.79	CTG	0.40	AAC	0.65	CCC	0.35
Organism	Q	F_{op}	R	F_{op}	S	F_{op}	T	F_{op}	V	F_{op}	Y	F_{op}
Apera spica-venti	CAG	0.90	CGC	0.54	TCC	0.44	ACC	0.33	GTC	0.42	TAC	0.72
Bromus tectorum	CAG	0.79	CGC	0.64	TCC	0.35	ACC	0.42	GTC	0.36	TAC	0.59
Alopecurus myosuroides	CAG	0.82	CGC	0.50	TCC	0.33	ACC/ACT	0.27	GTC	0.44	TAC	0.72
Beckmannia syzigachne	CAG	0.92	CGC	0.56	TCC	0.37	ACC	0.31	GTC	0.46	TAC	0.78
Lolium rigidum	CAG	0.83	CGC	0.62	TCC	0.41	ACC	0.39	GTC	0.45	TAC	0.82
Echinochloa crus-galli ALS1	CAG	0.88	CGC	0.55	TCC	0.39	ACC	0.43	GTC/GTG	0.39	TAC	0.65
E. crus-galli ALS2	CAG	0.92	CGC	0.55	TCC	0.35	ACC	0.47	GTG	0.42	TAC	0.65
E. crus-galli ALS3	CAG	0.92	CGC	0.56	TCC	0.40	ACC	0.44	GTG	0.42	TAC	0.65
Zea mays ALS1	CAG	0.92	CGC	0.50	TCC	0.33	ACC	0.35	GTC/GTG	0.42	TAC	0.59
Z. mays ALS2	CAG	0.88	CGC	0.46	TCC	0.32	ACC	0.31	GTC/GTG	0.40	TAT	0.56
Echinochloa oryzicola ALS1	CAG	0.88	CGC	0.55	TCC	0.39	ACC	0.46	GTC/GTG	0.39	TAC	0.65
E. oryzicola ALS2	CAG	0.92	CGC	0.55	TCC	0.35	ACC	0.46	GTG	0.42	TAC	0.65
Poa annua ALSa	CAG	0.91	CGC	0.47	TCC	0.42	ACC	0.35	GTC	0.44	TAC	0.69
P. annua ALSb	CAG	0.91	CGC	0.50	TCC	0.43	ACC	0.36	GTC	0.46	TAC	0.69

https://doi.org/10.17221/562/2020-PSE

under-represented are A-ending codons, whereas the three over-represented codons are G/C-ending. Furthermore, based on the RSCU values, the F_{op} values were calculated (Table 3). Overall, the set of optimal codons for *ALS* gene in weeds is as follow: Ala (GCC), Cys (TGC/TGT), Asp (GAC/GAT), Glu (GAG), Phe (TTT/TTC), Gly (GGC/GGT/GGG), His (CAC), Ile (ATC/ATT), Lys (AAG), Leu (CTG/CTC), Asn (AAC/AAT), Pro (CCA/CCC), Gln (CAG), Arg (CGC), Ser (TCC), Thr (ACC/ACT), Val (GTC/GTG), Tyr (TAC/TAT).

Despite of its ubiquitous nature, the mechanism of codon bias is not fully understood. Studies showed that synonymous codon usage may alter the expression of the gene of interest, and this effect can reach up to 1 000-fold or even more (Stoletzki and Eyre-Walker 2007, Quax et al. 2015). Although several field-based studies on weeds and their herbicide-resistant properties were conducted but work related to their molecular properties are still at their infancies. Heterologous gene expression studies can be very useful to study specific protein interactions. Studies involving codon optimisation will allow researchers to develop synthetic heterologous genes involved in herbicide resistance with the most efficient expressional efficiencies. These heterologous expression studies with optimised codons will have the potential to prove their efforts in the development of new herbicidal compounds. Hence, the present study was conducted to gain insight into the codon usage pattern of the acetolactate synthase gene in weedy species. The results obtained from the current study will enhance our understanding of the major factors and the pattern of codon usage in the ALS gene of weeds. Additionally, these results will help further investigations with the ALS gene and the development of new herbicidal compounds, which may require synthetic gene design based on codon usage patterns.

Acknowledgment. The authors would also like to acknowledge Dr. Pavel Hamouz, Department of Agroecology and Crop Production, Faculty of Agrobiology, Food and Natural Resources, Czech University of Life Sciences Prague, for his help regarding the images. The authors would also like to thank Dr. Amit Roy, Excellent Team for Mitigation, Faculty of Forestry and Wood Sciences, Czech University of Life Sciences Prague and Dr. Maor Matzrafi, Department of Plant Pathology and Weed Research, Agricultural Research Organisation-Volcani Center, Israel for their helpful suggestions regarding the manuscript.

REFERENCES

Bennetzen J.L., Hall B.D. (1982): Codon selection in yeast. Journal of Biological Chemistry, 257: 3026–3031.

Duggleby R.G., McCourt J.A., Guddat L.W. (2008): Structure and mechanism of inhibition of plant acetohydroxyacid synthase. Plant Physiology and Biochemistry, 46: 309–324.

Elhaik E., Pellegrini M., Tatarinova T.V. (2014): Gene expression and nucleotide composition are associated with genic methylation level in *Oryza sativa*. BMC Bioinformatics, 15: 23.

Elhaik E., Tatarinova T. (2012): GC3 biology in eukaryotes and prokaryotes. In: Tatarinova T. (ed.): DNA Methylation — From Genomics To Technology. London, IntechOpen. ISBN 978-953-51-0320-2.

Hamouzová K., Košnarová P., Salava J., Soukup J., Hamouz P. (2014): Mechanisms of resistance to acetolactate synthase-inhibiting herbicides in populations of *Apera spica-venti* from the Czech Republic. Pest Management Science, 70: 541–548.

Je M.Y., Kim H.Y., Son H.S. (2019): Analysis of the codon usage pattern of the RdRP gene of mycovirus infecting *Aspergillus* spp. Virology Journal, 16: 10.

Jugulam M., Shyam C. (2019): Non-target-site resistance to herbicides: recent developments. Plants, 8: 417.

Kumar S., Stecher G., Li M., Knyaz C., Tamura K. (2018): MEGA X: molecular evolutionary genetics analysis across computing platforms. Molecular Biology and Evolution, 35: 1547–1549.

Mazumdar P., Othman R.Y.B., Mebus K., Ramakrishnan N., Harikrishna J.A. (2017): Codon usage and codon pair patterns in non-grass monocot genomes. Annals of Botany, 120: 893–909.

Mondal S.K., Kundu S., Das R., Roy S. (2016): Analysis of phylogeny and codon usage bias and relationship of GC content, amino acid composition with expression of the structural nif genes. Journal of Biomolecular Structure and Dynamics, 34: 1649–1666.

Murphy B.P., Tranel P.J. (2019): Target-site mutations conferring herbicide resistance. Plants (Basel), 8: 382.

Powles S.B., Yu Q. (2010): Evolution in action: plants resistant to herbicides. Annual Review of Plant Biology, 61: 317–347.

Quax T.E.F., Claassens N.J., Söll D., van der Oost J. (2015): Codon bias as a means to fine-tune gene expression. Molecular Cell, 59: 149–161.

Sharp P.M., Stenico M., Peden J.F., Lloyd A.T. (1993): Codon usage: mutational bias, translational selection, or both? Biochemistry Society Transaktions, 21: 835–841.

Stoletzki N., Eyre-Walker A. (2007): Synonymous codon usage in *Escherichia coli*: selection for translational accuracy. Molecular Biology and Evolution, 24: 374–381.

Wright F. (1990): The "effective number of codons" used in a gene. Gene, 87: 23–29.

Zhou Z., Dang Y., Zhou M., Li L., Yu C.H., Fu J., Chen S., Liu Y. (2016): Codon usage is an important determinant of gene expression levels largely through its effects on transcription. Proceedings of the National Academy of Sciences, 113: E6117–25.

Received: October 27, 2020
Accepted: February 15, 2021
Published online: May 10, 2021