IN SILICO IDENTIFICATION OF PUTATIVE BARLEY PHYTOCHROME INTERACTING FACTORS (PIFs)

*Krisztián Gierczik [1, 2, *], Attila Vágújfalvi [2], Gábor Galiba [1, 2], Balázs Kalapos [1, 2]*

[1] Festetics Doctoral School, Georgikon Faculty, University of Pannonia, Keszthely, H-8360 Hungary

[2] Agricultural Institute, Centre for Agricultural Research, Hungarian Academy of Sciences, Martonvásár, H-2462 Hungary

[*] E-mail address: gierczik.krisztian@agrar.mta.hu

Abstract

Phytochrome Interacting Factors (PIFs) are plant transcription factors; members of the basic helix-loop-helix (bHLH) protein family. PIFs interact with the light sensitive phytochrome photoreceptors, thus they play pivotal role in light signaling, influencing a vast number of physiological processes. In spite of their importance, only a few studies focus on the identification of PIFs in different plant species until now. In this *in silico* study, we identified barley (*Hordeum vulgare* L.) bHLH proteins, and divided them into subfamilies by phylogenetic analysis. A total of 9 barley bHLH sequences were classified as VII (a+b) subfamily members. Since this group contains also those sequences, which were applied as reference PIF proteins (isolated from other plant species), we consider these 9 proteins as putative barley PIFs. These results provide a useful dataset for the forthcoming verification, analysis and functional analysis of barley PIF proteins.

Keywords: barley, *in silico* analysis, phylogenetic tree, bHLH protein family, PIF

Összefoglalás

A PIF (Phytochrome Interacting Factor) transzkripciós faktorok a bHLH (bázikus hélix-hurok-hélix) domént tartalmazó fehérjecsaládba tartoznak. Fontos szerepet játszanak a fény által aktivált jelátviteli útvonalakban, mivel képesek kapcsolatot kialakítani a fitokróm fotoreceptorokkal, amelyek a spektrum vörös és távoli-vörös tartományában rendelkeznek elnyelési maximummal. Mindezidáig csak kevés publikáció ismert, amelyek különböző növények PIF génjeinek azonosításáról szól. Ebben a tanulmányban *in silico* módszerekkel azonosítottuk az árpa (*Hordeum vulgare* L.) bHLH fehérjéit, majd filogenetikai módszerekkel alcsoportokba osztottuk őket. Összesen 9 egyedi árpa fehérjét azonosítottunk a VII (a+b) alcsoportból, amely más növény fajokban a PIF szekvenciákat is tartalmazta, tehát feltételezésünk szerint sikerült árpa PIF fehérjéket azonosítanunk. Eredményeink alapul szolgálhatnak a jövőben az árpa PIF fehérjék összehasonlító analíziséhez, majd azok funkcionális vizsgálataihoz.

Kulcsszavak: árpa, *in silico* analízis, filogenetikus fa, bHLH fehérje család, PIF

Introduction

For higher plants, light is one of the most important environmental factors that affect almost every process during the whole life cycle. The well characterized photoreceptors are responsible for detecting and absorbing light, each in a specific wavelength-range. Phytochromes (Phys) are red (R; λ_{max} ~660 nm) and far-red (FR; λ_{max} ~730 nm) light absorbents, while cryptochromes, phototropins, the Zeitlupe family and UVR8 photoreceptors are blue light and/or UV radiation sensitives (reviewed by Demarsy et al., 2018). In *Arabidopsis thaliana* (*At*) the Phy family is one of the most well characterized photosensors, its genome encodes 5 members of them, designated PhyA to PhyE (Clack et al., 1994). The

monocot species, such as *Oryza sativa* (*Os*) or *Hordeum vulgare* (*Hv*) have 3 members of the Phy family, namely PhyA, PhyB and PhyC (Mathews and Sharrock, 1996). They exist in two interconvertible forms, the red light (λ_{max} ~660 nm) absorbing Pr form (biologically inactive) and the far-red light (λ_{max} ~730 nm) absorbing Pfr form (biologically active) in response to light signals. The Pr form is photoconverted to Pfr form under R light, which is reversible by exposure to FR light or independently from light (dark revision), but it is a slower conversion (Rockwell et al., 2006). The bioactive Pfr form is involved in two signaling pathways, one involves Constitutive Photomorphogenic 1 (COP1) and Elongated Hypocotil 5 (HY5) proteins, the other involves the Phytochrome Interacting Factors, hereafter PIFs (Ni et al., 1998, 1999; Lu et al., 2015).

PIFs are basic helix-loop-helix (bHLH) transcription factors, they play crucial roles in regulation of the expression level of numerous target genes, controlling hormonal signaling, abiotic (high temperature, light, circadian) and biotic (defense responses) pathways (reviewed by Paik et al., 2017). The bHLH structural motif is specific for a superfamily of dimerizing transcription factors that is characterized by two helical sequences connected with a loop (Toledo-Ortiz et al., 2003). Most of these proteins have an extra basic region in the N-terminal part of the HLH (reviewed by Jones, 2004). PIFs have been characterized mostly in the dicot model organism *Arabidopsis*, and almost nothing has been studied in the economically and agronomically important monocot species. The *Arabidopsis* genome encodes at least 7 PIF proteins (reviewed by Pham et al., 2018) and the rice genome encodes almost the same number (six) of PIF-like sequences (Nakamura et al., 2007), but no studies have been published about PIFs in barley or in bread wheat yet.

The aim of this study was to identify the bHLH proteins in barley and classify them into phylogenetic subfamilies, then to select those subgroup(s) that contains putative PIFs, based on reference sequence comparison.

Materials and Methods

bHLH domain sequences

The latest version (release-40) of the barley proteome (ftp://ftp.ensemblgenomes.org/pub/release-40/plants/fasta/hordeum_vulgare/pep/) was obtained from the Ensembl Genomes (Kersey et al., 2018) site for *in silico* analysis. Hidden Markov Model (HMM) based search with the Pfam (Finn et al., 2016) HLH domain profile (PF00010) was performed to identify putative bHLH protein sequences using the HMMER v3.0 software package (Eddy, 2009). The predicted bHLH protein sequences were manually curated for increasing redundancy. The sequence logos as graphical representation of protein alignments were generated by WebLogo v2.8.2 tool (Crooks et al., 2004).

Sequence alignment, phylogenetic analysis and motif search

For the identification of barley bHLH protein subfamilies and especially the barley PIFs, representative bHLH protein sequences from each subfamily from *At* and *Os* were collected and analyzed. AtPIF, AtPIL, OsPIF and *Glycine max* (Gm) PIF sequences belonging to the VII (a+b) subfamily (Pires and Dolan, 2010) were also added to the alignment (Table 1). The Clustal Omega (EMBL-EBI) web tool (Sievers et al., 2011) and the MEGA X software package (Kumar et al., 2018) were used for the multiple alignment of the bHLH protein sequences and Simple Phylogeny (EMBL-EBI) web tool (Larkin et al., 2007) was applied to generate phylogenetic tree data by Neighbor-Joining method as well. The phylogenetic tree

was edited and visualized using FigTree v1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/). The phylogenetic tree was represented in a radial tree layout and cladogram transformed branched.

Table 1 Representative members of bHLH proteins from each subfamily from At and Os; AtPIF, AtPIL, OsPIF and GmPIF sequences, used in multiple sequence alignment along with the barley bHLH protein sequences. At sequences were obtained from the literature (Pires and Dolan, 2010), *Os and Gm sequences were downloaded from the NCBI database (OsPILlike13: XP_015618074.1, OsPILlike15: XP_015619034.1, GmPIF3: NP_001340167.1).*

Subfamily	bHLH protein sequences
Ia	AtbHLH094 AtbHLH099 OsbHLH044
Ib (1)	AtbHLH095 OsbHLH146
Ib (2)	AtbHLH118 AtbHLH162 OsbHLH168
II	AtbHLH089 AtbHLH091 OsbHLH142
III (a+c)	AtbHLH027 AtbHLH029 OsbHLH156
IIIb	AtbHLH061 AtbHLH116 OsbHLH004
III (d+e)	AtbHLH004 AtbHLH005 OsbHLH009
IIIf	AtbHLH002 AtbHLH012 OsbHLH016
IVa	AtbHLH018 AtbHLH025 OsbHLH018
IVb	AtbHLH011 AtbHLH047 OsbHLH061
IVc	AtbHLH105 AtbHLH115 OsbHLH059
IVd	AtbHLH041 AtbHLH092 OsbHLH026
Va	AtbHLH046 AtbHLH141 OsbHLH031
Vb	AtbHLH030 AtbHLH032 OsbHLH042
VII (a+b)	OsbHLH152 AtPIL1 AtPIL2 AtPIF3 AtPIF4 AtPIL5 AtPIL6 AtPIF7 OsPILlike13 OsPILlike15 GmPIF3
VIIIa	AtbHLH052 AtbHLH053 OsbHLH178
VIIIb	AtbHLH040 AtbHLH140 OsbHLH123
VIIIc (1)	AtbHLH083 AtbHLH086 OsbHLH127
VIIIc (2)	AtbHLH054 AtbHLH085 OsbHLH128
IX	AtbHLH122 AtbHLH128 OsbHLH111
X	AtbHLH110 AtbHLH112 OsbHLH073
XI	AtbHLH069 AtbHLH082 OsbHLH114
XII	AtbHLH060 AtbHLH062 OsbHLH095
XIII	AtbHLH156 AtbHLH157 OsbHLH149
XIV	AtbHLH143 AtbHLH145 OsbHLH138
XV	AtbHLH134 AtbHLH136 OsbHLH154

Results

The barley proteome contains at least 163 bHLH domains

To identify the putative HvPIF sequences, the latest version of the whole barley proteome was used for a Hidden Markov Model based search (*hmmsearch*) applying the HLH profile. 1036 bHLH domains were obtained, then this protein records were manually curated for increasing the redundancy. This reduction revealed that the originally obtained 1036 bHLH domains mean 163 unique bHLH sequences in barley. Even so 183 records were used in the phylogenetic analysis, because in some cases we could not select only one gene variant.

Key amino acids in the barley bHLH domain

To get the most correct (without repetitions that distort the result) barley bHLH consensus sequence database, multiple sequence alignment was performed with the 163 unique bHLH domains, and a graphical representation as sequence logos of this alignment were generated (Figure 1). We found that the consensus amino acid residues of barley bHLH domain (with the minimal 50% identity) in the basic region are: E_4, R_5, R_8, R_9; while in the Helix 1 region: E_{10}, I_{12}, N_{13}, L_{19}, L_{22}, V_{23}, P_{24}; in the Loop region: D_{54}; and finally in the Helix 2 region: A_{56}, L_{59}, A_{62}, I_{63}, Y_{65} and L_{70}.

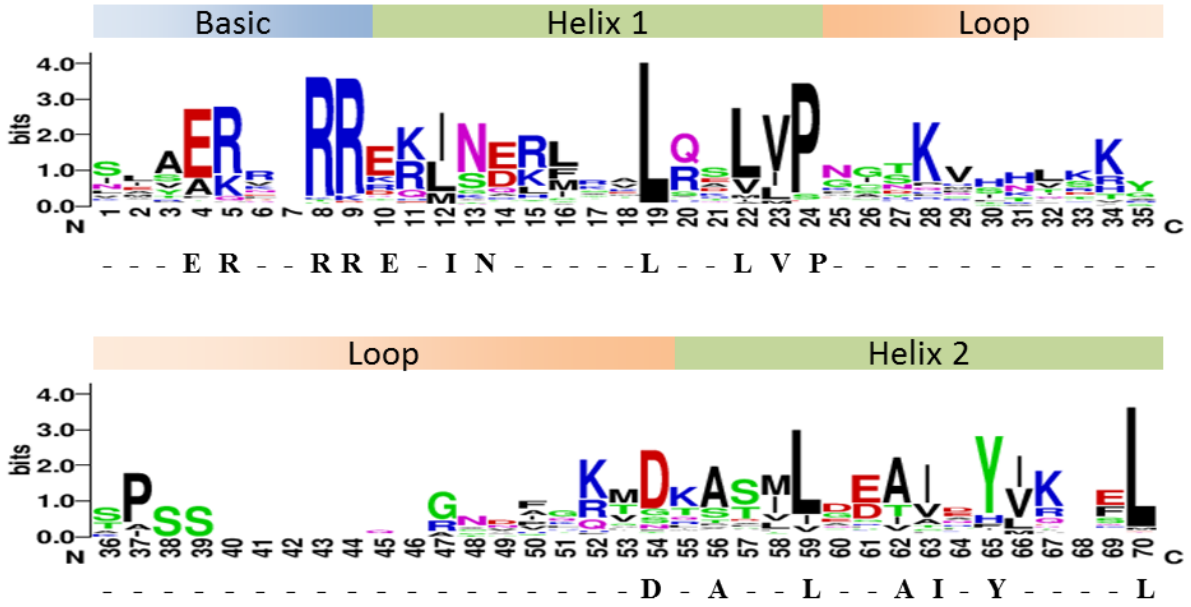

Figure 1 Graphical representation of barley bHLH sequence alignments. The labels at top represent the basic, helix 1, loop and helix 2 regions of the barley bHLH domain. The total height of the letters means the sequence conservation in that position in bits, while the proportions of amino acids in one position are shown as relative heights of individual letters. Coloring scheme was applied according to chemical properties. The bold letters under logos represent the consensus sequences of barley bHLH sequence with 50% identity. The numbering starts from a hypothetical site at the end of basic region.

Phylogenetic subfamilies of the barley bHLH proteins

To categorize the barley bHLH proteins into subfamilies, phylogenetic analyses were performed with the 183 amino acid sequences. To label the identified subfamilies and find the putative HvPIF sequences, we included several known sequences from each subfamily and also several known PIF amino acid sequences from other plant species (listed in Table 1). The phylogenetic analyses revealed that the barley bHLH domains can be separated into 25 subfamilies (Figure 2). The labels are numbered according to the system used for At and Os by Pires and Dolan (2010).

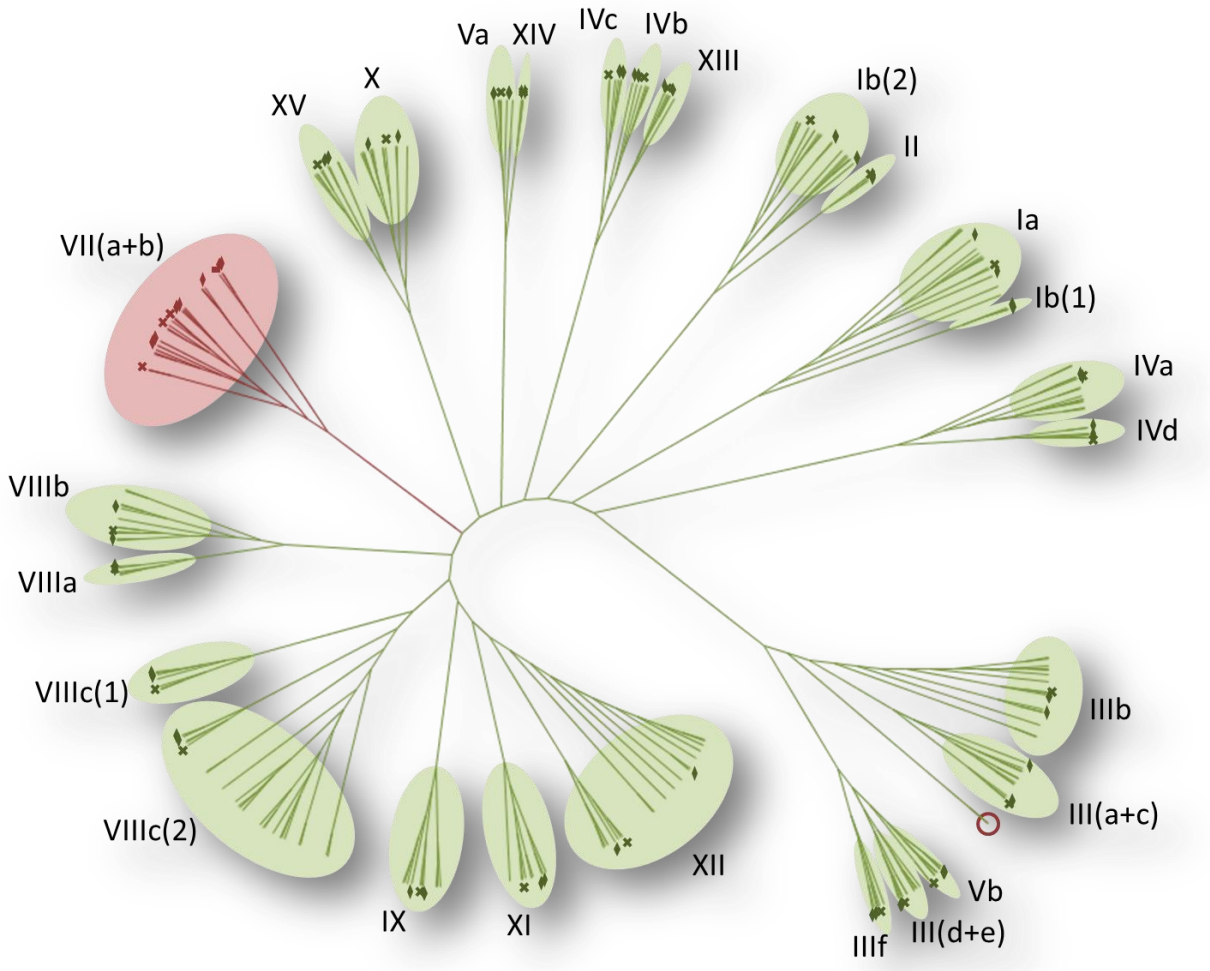

Figure 2 Phylogenetic tree of the barley bHLH proteins, shown as an unrooted cladogram. Representative members from At, Os and Gm were added to the analyses and marked at the tip of branches with diamond, cross and rectangle, respectively. The colored subgroups represent the barley bHLH subfamilies. The red bubble serves the VII (a+b) subfamily that probably contains the putative HvPIFs. Red circle marks the orphan members.

It turned out that two barley bHLH domain sequences cannot be grouped into any subfamilies. They showed a high level of divergence from other bHLH proteins, therefore these sequences (namely HORVU1Hr1G020370.2 and HORVU4Hr1G049550.3) were classified as ‘orphans’ (in the Figure 2 marked with red circle). On the other hand, we found that the XIV subfamily does not contain any barley bHLH sequences, suggesting that this subfamily is not highly conserved between plant species.

VII (a+b) subfamily contains at least 9 barley sequences

The phylogenetic analyses show that the VII (a+b) subfamily contains at least 11 sequences out of the 183 bHLH hits. Three (namely HORVU1Hr1G054260.3, HORVU1Hr1G054260.4 and HORVU1Hr1G054260.9) of the 11 sequences were not considered as unique hits, because these hits are from those ones that we could not classify as unique bHLH proteins. Thorough examination of these 3 sequences revealed that only one gene variants (namely HORVU1Hr1G054260.3) contains bHLH domain in the N-terminal region, so the other two variants were not further used in this study. Overall, we found that a total of 9 unique members of the VII (a+b) subfamily contain putative PIF sequences among the 163 identified barley bHLH proteins.

Discussion

In order to find the putative bHLH sequences in barley, we performed a Hidden Markov Model based search (*hmmsearch*) using the whole barley proteome and the Pfam (Finn et al., 2016) HLH profile. This search identified 163 unique bHLH sequences being encoded in the barley genome. Using *hmmsearch*, another study revealed that the number of these sequences are just the same in the *Arabidopsis* proteome (158) or in *Oryza sativa*, where 173 bHLH sequences were reported (Pires and Dolan, 2010). Multiple sequence alignment was performed to find the consensus sequence of these 163 bHLH proteins. The consensus amino acid (AA) residues of barley bHLH domain showed high similarities with the bHLH domains from *Arabidopsis* (Toledo-Ortiz et al., 2003) and *Zea mays* (Kumar et al., 2016) as well. In the *Arabidopsis* bHLH (*At* bHLH in Figure 3) sequence 16 AAs were found with at least 50% identity, whereas 26 AAs were reported from *Zea mays* (*Zm* bHLH in Figure 3) with the same

similarity. In this current study, 18 residues were found (with at least 50% identity) in the barley bHLH sequences (Figure 3).

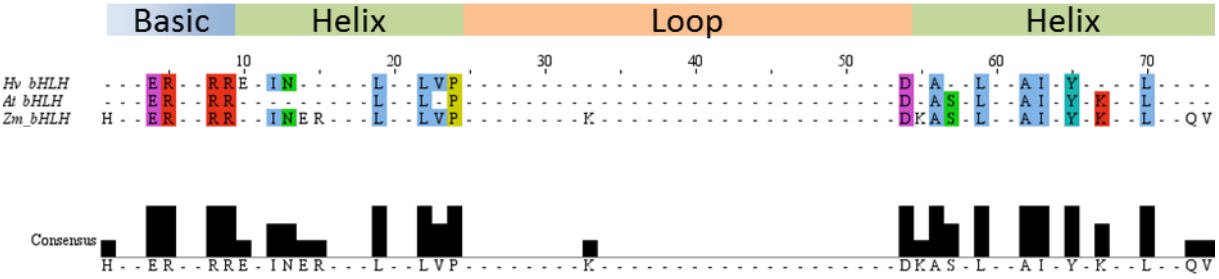

Figure 3 Multiple sequence alignment of the barley bHLH, the Arabidopsis bHLH (Toledo-Ortiz et al., 2003) *and the Zea mays bHLH* (Kumar et al., 2016) *domains with 50% identity . The labels at top represent the basic, helix 1, loop and helix 2 regions of bHLH domain. The total height of the black columns represents the sequence conservation in that position. Coloring scheme was applied for visualizing the residues. The numbering starts from a hypothetical site at the end of basic region.*

Comparing the three species, we assume that the common, i.e. consensus AAs represent pivotal residues with essential functions, therefore they are unchanged, conserved elements of the bHLH domain in Angiosperms. Thorough examination of those AAs, which are the same in *Arabidopsis* and *Zea mays* (S_{57} and K_{67}), but differ in barley reveals that they are just the same in barley, just the level of their identity is slightly less than the applied 50% threshold (S_{57}: 45.6% and K_{67}: 49.4%), therefore, these two AAs were also identified as elements of the consensus sequence. Based on the similarities of these elements, we assume that the functions of the bHLH proteins are conserved in *Arabidopsis*, maize and barley (and probably in Angiosperms). There was only one position (namely E_{10}) that was conserved neither in *Arabidopsis* nor in *Zea mays*, but indeed, it was in barley. Whether this AA has any special effect on the barley bHLH protein functions or not is still not known.

To find those barley bHLH sequences, which can be considered as putative PIFs, first we studied the 183 bHLH amino acid sequences compared to the 'reference' sequences (listed in Table 1) to compute a phylogenetic tree, and then, based on the topology of the tree, we

defined 25 subfamilies of barley bHLH proteins. These subfamily separations were highly consistent with the previously published phylogenetic analyses carried out on *Arabidopsis* and *Oryza sativa* (Pires and Dolan, 2010), as well as on *Zea mays* sequences (Kumar et al., 2016). Of the 268 proteins analyzed, only two (HORVU1Hr1G020370.2 and HORVU4Hr1G049550.3) were not clearly classified into any of the 25 subfamilies, thus they were marked as 'orphans'. These two protein sequences show high degree of sequence divergence from the other bHLH domains, therefore they may be considered as pseudogenes or they may represent a special subfamily, which is not present in the *Arabidopsis* bHLH tree. It is also possible that these two hits are members of the III (a+c) subfamily, which is closely located to them.

Subfamily XIV does not contain any of the studied barley sequences. Only one study, Imai et al. (2006) has been published in which a member of this minor subfamily was functionally characterized.

To identify the most relevant bHLH hits as putative PIF sequences from the 183 proteins, known PIFs, as reference sequences, were collected and added to the analysis from other plant species (*At, Os, Gm*). The phylogenetic analyses revealed that all of the studied PIFs were classified into the VII (a+b) subfamily, as it could be predicted from the literature (Pires and Dolan, 2010), thus confirming the accuracy of this novel phylogenetic tree. Overall, 9 unique barley members of the VII (a+b) subfamily were identified which probably contain the barley PIF sequences (Table 2). To find out which one of these are actually members of the HvPIF transcription factor family, *in vivo* function analyses are needed as further studies.

Table 2 Characteristics of the barley VII (a+b) subfamily from the bHLH proteins.

Transcript ID	Chr	Genomic DNA size (bp)	Number of Amino Acids
HORVU1Hr1G017900.1	1H	3463	493
HORVU1Hr1G054260.3	1H	1736	464
HORVU2Hr1G060680.1	2H	2239	362
HORVU2Hr1G104040.6	2H	567	188
HORVU5Hr1G011780.1	5H	4871	341
HORVU5Hr1G093310.11	5H	2104	547
HORVU5Hr1G102240.5	5H	3033	396
HORVU6Hr1G088020.4	6H	1685	296
HORVU7Hr1G026560.2	7H	1876	338

Conclusion

In this study, *in silico* analyses have provided information about the barley bHLH protein family and the putative PIF transcription factors in barley. Phylogenetic analysis identified the subfamilies of the barley bHLH domain sequences and revealed 9 unique sequences that show high level of similarity to PIFs from other plant species. This result will serve as valuable resources for future research in plant molecular biology, especially in light signaling studies.

Acknowledgement

The publication is supported by the EFOP-3.6.3-VEKOP-16-2017-00008 project. The project is co-financed by the European Union and the European Social Fund.

References

Clack T., Mathews S., and Sharrock R.A., 1994. The phytochrome apoprotein family in *Arabidopsis* is encoded by five genes: the sequences and expression of *PHYD* and *PHYE*. *Plant Molecular Biology*. **25**, 413–427.

Crooks G.E., Hon G., Chandonia J.-M., and Brenner S.E., 2004. WebLogo: A Sequence Logo Generator. *Genome Research*. **14**, 1188–1190.

Demarsy E., Goldschmidt-Clermont M., and Ulm R., 2018. Coping with 'Dark Sides of the Sun' through Photoreceptor Signaling. *Trends in Plant Science*. **23**, 260–271.

Eddy S.R., 2009. A New Generation of Homology Search Tools Based on Probabilistic Inference. *Genome Informatics*., 205–211.

Finn R.D., Coggill P., Eberhardt R.Y. *et al.*, 2016. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research*. **44**, D279–D285.

Imai A., Hanzawa Y., Komura M., Yamamoto K.T., Komeda Y., and Takahashi T., 2006. The dwarf phenotype of the *Arabidopsis acl5* mutant is suppressed by a mutation in an upstream ORF of a bHLH gene. *Development*. **133**, 3575–3585.

Jones S., 2004. An overview of the basic helix-loop-helix proteins. *Genome Biology*. **5**, 226.

Kersey P.J., Allen J.E., Allot A. *et al.*, 2018. Ensembl Genomes 2018: an integrated omics infrastructure for non-vertebrate species. *Nucleic Acids Research*. **46**, D802–D808.

Kumar S., Stecher G., Li M., Knyaz C., and Tamura K., 2018. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Molecular Biology and Evolution*. **35**, 1547–1549.

Kumar I., Swaminathan K., Hudson K., and Hudson M.E., 2016. Evolutionary divergence of phytochrome protein function in *Zea mays* PIF3 signaling. *Journal of Experimental Botany*. **67**, 4231–4240.

Larkin M.A., Blackshields G., Brown N.P. *et al.*, 2007. Clustal W and Clustal X version 2.0. *Bioinformatics*. **23**, 2947–2948.

Lu X.D., Zhou C.M., Xu P.B., Luo Q., Lian H.L., and Yang H.Q., 2015. Red-Light-Dependent Interaction of phyB with SPA1 Promotes COP1-SPA1 Dissociation and Photomorphogenic Development in *Arabidopsis*. *Molecular Plant*. **8**, 467–478.

Mathews S. and Sharrock R.A., 1996. The Phytochrome Gene Family in Grasses (Poaceae): A Phylogeny and Evidence that Grasses Have a Subset of the Loci Found in Dicot Angiosperms. *Molecular Biology and Evolution*. **13**, 1141–1150.

Nakamura Y., Kato T., Yamashino T., Murakami M., and Mizuno T., 2007. Characterization of a Set of Phytochrome-Interacting Factor-Like bHLH Proteins in *Oryza sativa*. *Bioscience, Biotechnology, and Biochemistry*. **71**, 1183–1191.

Ni M., Tepperman J.M., and Quail P.H., 1998. PIF3, a Phytochrome-Interacting Factor Necessary for Normal Photoinduced Signal Transduction, Is a Novel Basic Helix-Loop-Helix Protein. *Cell*. **95**, 657–667.

Ni M., Tepperman J.M., and Quail P.H., 1999. Binding of phytochrome B to its nuclear signalling partner PIF3 is reversibly induced by light. *Nature*. **400**, 781–784.

Paik I., Kathare P.K., Kim J. Il, and Huq E., 2017. Expanding Roles of PIFs in Signal Integration from Multiple Processes. *Molecular Plant*. **10**, 1035–1046.

Pham V.N., Kathare P.K., and Huq E., 2018. Phytochromes and Phytochrome Interacting Factors. *Plant Physiology*. **176**, 1025–1038.

Pires N. and Dolan L., 2010. Origin and Diversification of Basic-Helix-Loop-Helix Proteins in Plants. *Molecular Biology and Evolution*. **27**, 862–874.

Rockwell N.C., Su Y.-S., and Lagarias J.C., 2006. Phytochrome Structure and Signaling Mechanisms. *Annual Review of Plant Biology*. **57**, 837–858.

Sievers F., Wilm A., Dineen D. *et al.*, 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*. **7**.

Toledo-Ortiz G., Huq E., and Quail P.H., 2003. The Arabidopsis Basic/Helix-Loop-Helix Transcription Factor Family. *The Plant Cell*. **15**, 1749–1770.