DOI: 10.2298/ABS151120023S

DUF1070 AS A SIGNATURE DOMAIN OF A SUBCLASS OF ARABINOGALACTAN PEPTIDES

Ana D. Simonović*, Milan B. Dragićević, Milica D. Bogdanović, Milana M. Trifunović-Momčilov, Angelina R. Subotić and Slađana I. Todorović

Institute for Biological Research "Siniša Stanković", Department for Plant Physiology, University of Belgrade,
Bulevar despota Stefana 142, 11060 Belgrade, Serbia

***Corresponding author:** ana.simonovic@ibiss.bg.ac.rs

Received: November 20, 2015; **Revised**: January 31, 2015; **Accepted**: February 1, 2016; **Published online**: March 17, 2016

Abstract: Over 20% of all protein domains are currently annotated as "domains of unknown function" or DUFs. In a recently identified *Centaurium erythraea* arabinogalactan peptide, *CeAGP3* (AGN92423), a conserved DUF1070 domain was found. Since identifying functions for DUFs is important in systems biology, we have analyzed the distribution and structure of DUF1070 domain (pfam06376) using a set of bioinformatics tools. There are 271 publically available DUF1070 members from 25 diverse families of vascular plants, and most are short sequences (50-100 aa). The N-terminal signal peptide (Nsp) was found in almost all complete sequences. In 233 sequences, at least two noncontiguous prolines were found as clustered dipeptides predicted to be hydroxylated and glycosylated with type II arabino-3,6-galactans, thus representing AG-II glycomodules. In addition, 35 sequences contained a region rich in basic residues (basic linker, BL). The N-terminal part of the DUF1070 domain is comprised of (part of) AG-II and/or BL, while the highly conserved C-terminus is a region of 26 aa, termed SH26. In 212 sequences, SH26 was a typical glycosylphosphatidylinositol lipid anchor signal peptide (GPIsp), but in 83 cases GPIsp was not predicted due to software constraints. In sequences where both Nsp and GPIsp were predicted, the length of mature peptides could be calculated, and it was 10-16 aa. Our analysis suggests that DUF1070 members are arabinogalactan (AG) peptides, of which the majority are GPI-anchored. DUF1070 is the only conserved domain found in classical arabinogalactan proteins and AG peptides. The SH26 region can be used for mining and annotation of AG peptides.

Key words: AG peptides; arabinogalactan proteins; DUF1070; GPI anchor; pfam06376

Abbreviations: aa – amino acids; AGPs – arabinogalactan proteins; AG – arabinogalactan; AG-II – type II arabino-3,6-galactans; GPI – glycosylphosphatidylinositol lipid anchor; GPIsp – GPI lipid anchor signal peptide; DUF – domain of unknown function; Nsp – N-terminal signal peptide; PAST – Pro/Hyp, Ala, Ser and Thr; SH26 – highly conserved domain of 26 amino acids; BL – basic linker; Hyp – hydroxyproline

INTRODUCTION

Arabinogalactan proteins (AGPs) are ubiquitous plant cell surface glycoproteins, located in the apoplast, and anchored to the plasma membrane or secreted. Along with extensins and proline-rich proteins, AGPs comprise a superfamily of hydroxyproline-rich glycoproteins [1]. AGPs are implicated in cell division, plant growth and in diverse developmental processes [2,3], including organogenesis and somatic embryogenesis *in vitro* [4].

The diversity of AGP biological functions is attributed to their structural heterogeneity, based partly on different protein backbones encoded by large multigene families [1,5-7] and partly on the fact that one backbone may undergo different glycosylation patterns yielding different glycoforms [5]. Nevertheless, all AGPs share some common features, such as decoration with glycans composed predominantly of arabinose and galactose that constitutes over 90% of their molecular weight, protein backbones rich in Pro (commonly modified to Hyp), Ala, Ser, Thr and Gly, and N-terminal signal peptide (Nsp) guiding their translation via secretory pathway [3,8]. The AGP glycan moieties are typically branched type II arabino-3,6-galactans (AG-II) and oligoarabinosides, which are *O*-linked to the Hyp residues [3, 9]. The glycosylation sites (glycomodules) in the protein backbones can be identified following the so-called Hyp contiguity hypothesis, whereby AG-IIs are *O*-linked to noncontiguous Hyp residues organized as clustered dipeptides (AG-II glycomodules), while *O*-glycosylation leading to arabinosylation occurs on extensin glycomodules (Ser-(Hyp)$_{3-5}$ contigs) [1,3,5,8]. Literature sources [1,3,10] suggest that dipeptides found in AG-II glycomodules include Ala-Hyp, Hyp-Ala, Ser-Hyp, Hyp-Ser, Thr-Hyp, Hyp-Thr, and less often Val-Hyp, Hyp-Val and Gly-Hyp. While sequence conservation of the AG-II glycomodules hardly exists, the abundance of Pro/Hyp, Ala, Ser and Thr (P, A, S, T) causes a biased amino acid composition of AGPs, so that a high percentage of these residues (% PAST) is used as a criterion in mining for AGP sequences [1].

Many AGPs are attached to plasma membrane by a glycosylphosphatidylinositol (GPI) lipid anchor [3,5,11]. GPI anchors are not only an alternative to transmembrane domains, but are thought to improve lateral mobility of the anchored proteins, allow polarized cell surface targeting and facilitate inclusion in lipid rafts [9,11,12]. Cotranslational addition of the GPI moiety to the ω-site of a nascent substrate protein is catalyzed by the transamidase complex at the luminal side of the ER membrane, following proteolytic cleavage of the C-terminal propeptide between ω and ω+1 residues [3,13]. This modification is directed by the C-terminal GPI lipid anchor signal peptide (GPIsp), a sequence that is not conserved, but has the characteristic pattern of amino acid residues around and following the ω-site, with position-specific physicochemical requirements and size restrictions [5,11-13].

Classical AGPs and their short counterparts, AG peptides, are AGP classes with simple structure, comprised, in immature proteins, of Nsp, Hyp-rich glycosylation scaffold (AG-II glycomodules) and often C-terminal GPIsp [5,6,10,13]. While other AGP classes may contain conserved functional domains (such as fasciclin) or specific motifs (e.g. Lys-rich motifs) and other distinguishable features [7,9], neither classical AGPs nor AG peptides contain any known conserved regions [5,6]. We have recently identified several AGPs expressed in common centaury (*Centaurium erythraea* Rafn), including an AG peptide, *CeAGP3* (AGN92423), with a conserved DUF1070 domain [4]. DUFs (domains of unknown function) represent a large set of families within the Pfam database that do not include any protein of known function [14]. Over 20% of all protein domains are currently annotated as DUFs, and it is estimated that there are over 1500 DUF domains (or Pfam families) in eukaryotes and even more in bacteria [14,15]. From the perspective of systems biology, identifying functions for DUFs is important for characterizing lists of biological "parts" [14]. The aim of the present work was to analyze all publically available sequences containing the DUF1070 domain (pfam06376) in order to determine whether this domain is present only in AG peptides or also in other proteins, and to elucidate its function using a bioinformatics approach.

MATERIALS AND METHODS

The most diverse members of the DUF1070 family were identified using the Conserved Domain Database (CDD, [16], www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml), while other sequences with similar domain architecture were found using the Conserved Domain Architecture Retrieval Tool (CDART, [17]). A total of 271 DUF1070-containing non-redundant protein sequences were downloaded from the NCBI Entrez GenPept database as a FASTA file on 31 March, 2015.

For sequence analyses and manipulations, Jalview version 2 was used [18] (www.jalview.org). The presence of Nsp was determined using SignalP 4.0. [19] (www.cbs.dtu.dk/services/SignalP/) with a discrimination score cutoff of D=0.450, using a method adopted for sequences that do not include transmembrane regions. Possible targeting to other cellular compartments was analyzed using TargetP 1.1 [20] (www.cbs.dtu.dk/services/TargetP/). The identification of GPIsps was performed with Big-Pi Plant Predictor [13] (mendel.imp.ac.at/gpi/plant_server.html). Possible *N*-glycosylation sites (Asn-Xaa-Ser/Thr consensus, where Xaa≠Pro) were predicted with the NetNGlyc 1.0 Server (www.cbs.dtu.dk/services/NetNGlyc/). AG-II glycomodules were found manually, searching for the dipeptides Ala-Hyp, Hyp-Ala, Ser-Hyp, Hyp-Ser, Thr-Hyp, Hyp-Thr, Val-Hyp, Hyp-Val and Gly-Hyp. Possible arabinosylation sites (Ser-(Hyp)3-5) were also found manually.

The DUF1070-containing protein sequences without the predicted Nsp were aligned using Kalign multiple sequence alignment [21] (www.ebi.ac.uk/Tools/msa/kalign/), with manual adjustments around a highly conserved region of 26 amino acids (termed SH26) present in all the examined sequences. The

SH26 average distance tree was constructed using the Blosum62 substitution matrix [22]. Calculations on sequence text strings such as % PAST were done using the stringr package [23] in R [24]. For graphical presentations R packages ggplot2 [25], gridExtra [26] and Venn diagram [27] were used.

RESULTS AND DISCUSSION

Number and origin of DUF1070 family members

The search for DUF1070-containing sequences was initiated with 15 of the most diverse DUF1070 (pfam06376) family members that were identified in CDD (Supplement A). For each of the 15 initially identified sequences, the same set of 269 non-redundant protein sequences with similar domain architecture was retrieved using CDART. Since all but two out of the 15 initial sequences were already represented in the 269 sequence set, a total of 271 DUF1070-containing protein sequences (Supplement A) were further analyzed. At the time of the download (31 March, 2015), there was a total of 479 protein and 597 DNA or RNA entries in the GenPept and GenBank collections, respectively, but all of them were already represented in the 271-sequence set, and as redundant were not further considered. The majority, i.e. 144 out of 271 sequences, were annotated as unknown or hypothetical proteins, 108 as AG peptides and 18 as AGPs, while XP_009377255 was probably wrongly annotated.

DUF1070-containing proteins were found in 61 species of vascular plants belonging to 25 diverse families, including *Lycopodiophyta*, the oldest extant division of the vascular plants. Since AGPs are found in all land plants and their charophyte ancestors [28], it remains to be clarified whether the members of the DUF1070 family are limited to vascular plants or are taxonomically as widespread as AGPs.

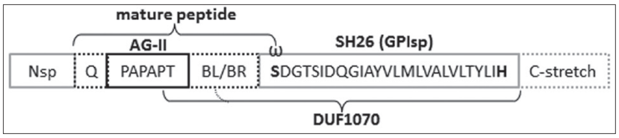

Fig. 1. General structure of DUF1070 members. Parts of the immature peptides that are present in almost all analyzed sequences are indicated as full lines, while those present in some are indicated as dotted lines. Black lines represent the regions of a mature peptide, while gray lines are parts that are predicted to be cleaved. **Nsp** – N-terminal signal peptide; **Q** – N-terminal Gln that can be modified to pyroglutamate; **AG-II** – glycomodule with PAPAPT consensus; **BL/BR** – basic linker or region; **SH26** – conserved region of 26 aa with indicated consensus, that is often recognized as glycosylphosphatidylinositol lipid anchor signal peptide (**GPIsp**); **ω** – site of GPI attachment. The N-terminal side of the DUF1070 is not well conserved, and consists of (part of) AG-II or (part of) BL/BR.

Since the DUF1070-containing peptide that caught our attention, CeAGP3 (AGN92423), is an AG peptide predicted to be GPI-anchored, the sequence analysis of the pfam06376 family was focused on features characteristic for AG peptides: short length (10-15 aa) and the presence of Nsp, AG-II glycomodule and GPIsp [10]. The general structure of DUF1070 members is presented in Fig. 1, and discussed in the following sections.

DUF1070-containing sequences are short and can be considered as peptides

The length of the 271 analyzed sequences ranges from 35 to 305 aa. However, the majority of the sequences are 50-100 aa long, with only 6 shorter and 11 longer sequences (Fig. 2A). According to Showalter et al. [29], the length of unprocessed AG peptides should

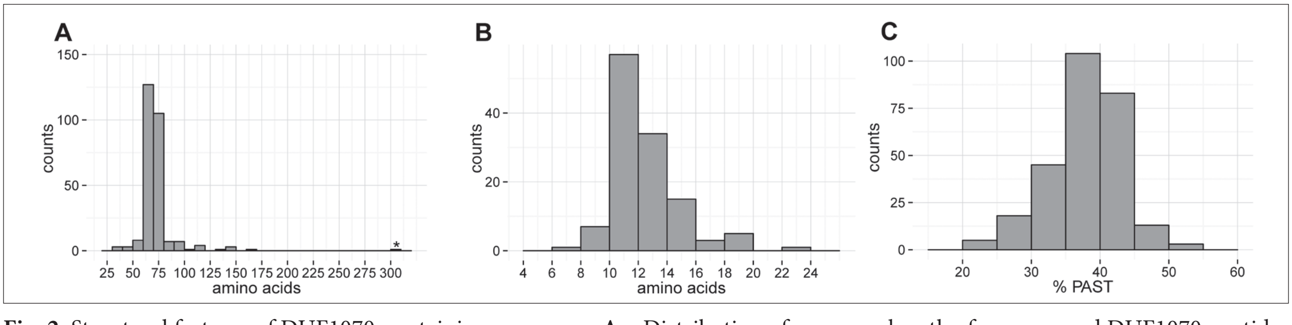

Fig. 2. Structural features of DUF1070-containing sequences. **A** – Distribution of sequence length of unprocessed DUF1070 peptides (indicated with asterisk is sequence XP_003620524 with 2 DUF1070 domains). **B** – Distribution of sequence length of mature peptides in a subset of sequences where Nsp and GPIsp are predicted. **C** – Content of Pro, Ala, Ser and Thr (% PAST) in all analyzed sequences.

be between 50 and 90 aa, so the DUF1070 members fulfill this qualification. The nine shortest sequences that are ≤54 aa appear to be incomplete (Supplement A). Of these, 7 lack the predicted Nsp, while ACG30862 and XP_008775127 are just 7 aa longer than their predicted Nsp, which overlaps with the DUF1070 region, suggesting that these sequences are also truncated. For the same reason, we suspect that XP_008775377 is also truncated. The longest sequence, XP_003620524 from *Medicago truncatula* (305 aa), contains two DUF1070 domains. For a subset of sequences where both Nsp and GPIsp were unambiguously recognized, the length distribution of the majority of mature proteins ranges from 10 to16 aa (Fig. 2B), which is in accordance with the predicted length of mature AG peptides [10].

DUF1070 members are predicted to be synthetized via the secretory pathway and are probably stabilized by N-terminal modification

The N-terminal SP cleavage sites (Nsp) were predicted for all but 19 sequences. In addition, Nsp was probably wrongly predicted in three apparently truncated sequences (the previously discussed ACG30862, XP_008775127 and XP_008775377), amounting to 22 sequences without Nsp (Table 1). Among them, five are considered as "ambiguous" regarding the presence of Nsp, including EYU40039, KHG25041 and XP_003529073, for which the discrimination score is somewhat below the default SignalP threshold, as well as XP_006397819 and XP_007152940, which likely possess N-terminal artificial adducts pre-

Table 1. Structural elements of the DUF1070 family members

group	Nr	Families	Nsp				Q		Hyp/Pro in AG-II				Linker				GPIsp				C
			Y	Am	Inc	N	Y	1st	0	1	2	≥3	BL	BR	nBL	N	Y	PP	Sh	N	Y
1	221	25 families	210	5	4	2	182	164	4	2	8	207	2	4	6	209	123$^{\#}$	82	3	13	142
2	2	Rosaceae	1		1				1			1	1	1			1		1		2
3	3	*Pinaceae*	3				3	2				3				3	1			2	
4	2	*Poaceae*	2				2	2		2					2		2				
5	1	*Rosaceae*	1				1	1				1				1	1				
6	2	*Musaceae, Poaceae*	2				1	1		2			2							2	
7	1	*Poaceae*	1						1				1							1	
8	1	*Myrtaceae*				1	1					1			1		1				
9	3	*Poaceae*			3				2	1			1	2					3		
10	1	Poaceae				1						1	1							1	
11	1	*Poaceae*				1			1					1						1	1
12	3	*Poaceae*			2	1	1		2			1	1	2					2	1	
13	18	*Poaceae*	18				17	7	3	12	2	1	18							18	1
14	1	*Fabaceae*	1									1				1	1				
15	2	*Fabaceae*	2				2	2				2				2				2	2
16*		Fabaceae																			
17	5	*Poaceae*	4			1	1			3		2	5							5	
18	1	*Poaceae*	1							1			1							1	1
19	1	*Poaceae*	1							1			1							1	1
20	1	*Vitaceae*	1				1	1				1				1				1	1
21	1	*Vitaceae*	1				1	1				1				1				1	
total	271	25	249	5	10	7	213	181	14	24	10	223	34	10	9	218	130$^{\#}$	82$^{\#}$	9	50	151
%			91.9	1.8	3.7	2.6	78.6	66.8	5.2	8.9	3.7	82.3	12.5	3.7	3.3	80.4	48.0	30.3	3.3	18.5	55.7

The DUF1070 sequences were clustered into 21 groups based on homology of their SH26 regions. Group 1 contains species from 25 families – *Lycopodiophyta: Selaginellaceae; Pinopsida: Pinaceae; Eudicots: Amaranthaceae, Brassicaceae, Cleomaceae, Cucurbitaceae, Euphorbiaceae, Fabaceae, Gentianaceae, Malvaceae, Moraceae, Myrtaceae, Nelumbonaceae, Pedaliaceae, Phrymaceae, Rosaceae, Rubiaceae, Rutaceae, Salicaceae, Solanaceae, Theaceae* and *Vitaceae; Monocots: Arecaceae, Musaceae* and *Poaceae.* The numbers in the table represent the number of sequences with specific structural element. **Nsp** – N-terminal signal peptide (**Y** – present, **Am** – ambiguous, **Inc** – incomplete sequence, **N** – not present); **Q** – Gln at the beginning of mature peptide (**Y** – number of sequences that contain Gln at the first position in mature peptide or within several amino acids preceding the AG-II glycomodule, **1st** – number of sequences where Gln is the N-terminal amino acid in mature peptide); **Hyp/Pro in AG-II** – the number of (hydoxy)prolines in AG-II glycomodule; **Linker** – stretch of amino acids preceding the SH26 region that is not an AG-II glycomodule and can be rich in Arg, Lys or His (termed basic linker – **BL**, when linking AG-II and SH26, or basic region – **BR** in sequences lacking Nsp and AG-II), without basic residues (non-basic linker, **nBL**) or absent (**N**); **GPIsp** – GPI lipid anchor signal peptide (**Y** – predicted, **PP** – not predicted due to propeptide length, **Sh** – not predicted in sequences ≤ 54 aa, **N** – not predicted for other reasons); **C** – C-terminal stretch found after the SH26 region. * – group 16 contains only the second DUF1070 domain of the sequence XP_003620524, while its first DUF1070 belongs to group 15, were all structural elements of this sequence are presented. # – Big-Pi predicted GPIsp in 129 sequences, but NP_566070, where GPIsp was experimentally confirmed is included in this group instead of "PP" group.

ceding the Nsp. Nsp was not found in 7 very short and apparently incomplete sequences (ACG26444, ACG27819, ACG29429, EAZ22191, XP_004951483, XP_009336272 and XP_010251935), leaving only 7 sequences without explanation for the Nsp absence (Table 1 and Supplement A). Of these, three are relatively long (143-148 aa) sequences from *Eucalyptus grandis*, while the remaining four are all from monocots (EEE62548 from rice, EMS58237 from wheat and XP_008649433 and XP_008657847 from maize), with atypical or absent AG-II glycomodules. Most of the 249 predicted Nsps are 20-35 aa. In few probably incomplete monocot sequences, a potential chloroplast transit peptide or mitochondrial targeting peptide was found, but with low reliability (Supplement A). Identification of Nsp in 92% of the sequences suggests that DUF1070-containing sequences, with few exceptions of questionable quality, are synthetized as peptides for export via the secretory pathway.

An interesting feature in the analyzed sequences is the presence of a Gln (Q) preceding the AG-II glycomodule (Fig. 1). Namely, in 78.6% of cases, Gln is within several aa before the AG-II region, while in 66.8% of sequences Gln is predicted to be the N-terminal residue in mature peptides (Table 1 and Supplement B). The N-terminal Gln is modified to pyroglutamate in many proteins. This is a common posttranslational modification that may occur either spontaneously or by the action of glutaminyl cyclase (E.C. 2.3.2.5) [30,31]. Modification of N-terminal Gln to pyroglutamate may contribute to protein stability, since it protects the proteins from aminopeptidases. In animal systems, this modification was shown to increase protein hydrophobicity, propensity to aggregate and to affect protein function [30]. Plant glutaminyl cyclases have low substrate specificity and are synthetized by the secretory pathway, along with their substrates, so that both proteolytic N-terminal processing and the consequent glutaminyl cyclization occur in the ER [31]. DUF1070 members AtAGP16, AtAGP20 and AtAGP41 (NP_566070, NP_191723 and NP_974828, respectively) were shown to have N-terminal pyroglutamate [5,10]. The occurrence of N-terminal Gln in 66.8% of sequences suggests their proteolytic stability and slower turnover. AGPs are generally stable, and when they accumulate to a required level, their expression may decline, while the protein level persists [1]. In 32 sequences, however, Gln is not the N-terminal aa, but precedes the AG-II glycomodule (Table 1 and Supplements 1 and 2). However, in most of these cases, SignalP predicted several potential Nsp cleavage sites, including those preceding Gln, so the number of DUF1070 sequences starting with Gln could be even higher.

Most of the DUF1070 peptides contain typical AG-II glycomodule

In most of the analyzed sequences, the AGII glycomodule can be readily identified by the presence of Pro alternating with Ala, Ser or Thr. According to the Hyp contiguity hypothesis, these prolines are hydroxylated to serve as glycosylation scaffolds [1]. The number of Pro/Hyp in the AG-II glycomodules varies, but in 80.3% of the DUF1070 members it is 3 or more (Table 1 and Supplement B). Pro/Hyp that occurs singly, as in 24 of the sequences, is rarely glycosylated [7]. Even though the AG-II glycomodules are not conserved [1,5,6], the "PAPAPT" motif is frequent among the analyzed set of sequences (Fig. 1)

None of the seven short and apparently incomplete sequences that lack Nsp possess AG-II glycomodule, except XP_004951483, which has a single "PA" dipeptide (Supplement B). Likewise, AG-II glycomodule is absent in sequences ACG30862, XP_008775127 and XP_008775377 that were previously presumed to be truncated. Sequence EMS58237 from *Triticum urartu* also lacks both Nsp and AG-II glycomodule. Only three sequences from sorghum (XP_002458076, XP_002458079 and XP_002458080) and one from rice (ABB46792) contain a predicted Nsp but lack AG-II glycomodule. It can be concluded that the vast majority of complete and well-assembled DUF1070-containing sequences are predicted to be *O*-glycosylated with type II arabino-3,6-galactans, while only 14 sequences lack the AG-II glycomodule (Table 1). Each of the eight longest sequences (118-305 aa, Supplements A and B) have several additional Pro/Hyp residues (dipeptides) either in their long N-terminal region (XP_008657847, KCW79699, KCW79700, KCW82460, KJB09691 and XP_00352907), following the DUF1070 domain (BAD52702), or between two DUF1070 domains (XP_003620524), but only in the latter case is this actually a second AG-II domain (duplicated along with DUF1070), while in the other cases, these dipeptides are scattered. In addition, a single sequence (XP_004503418) was found to contain an SPPP motif, a possible extensin module that might be arabinosylated [1]. Possible *N*-glycosylation sites were predicted in 11 sequences, but in 10 cases this site was within Nsp, while in one it was within the C-stretch. Since all found *N*-glycosylation sites (indicated in Supplement A) are predicted to be cleaved, it can be concluded that DUF1070 family members are not *N*-glycosylated, which is a modification common in some chimeric AGPs, but not in AG peptides [3,8].

The % PAST for AG peptides should be above 35% [1,6]. In the analyzed sequences, the % PAST is between 20.4% and 51%, but in 248 sequences it is over 30% (Fig. 2C). The slightly lower % PAST in some of the analyzed sequences than is commonly found in AG peptides is due to the presence of DUF1070 domains, which are not rich in PAST.

Structure of the DUF1070 domain

The common feature of all analyzed sequences is, of course, the DUF1070 domain. The C-terminal part of DUF1070 is comprised of highly conserved 26 residues termed the SH26 region, after Ser and His that often encompass it, while the N-terminal part is less conserved and in the majority of members is actually a part of the AG-II glycomodule (Fig. 1). However, in 53 sequences the SH26 region is preceded by a non-conserved stretch of amino acids of variable length. This region is often rich in basic amino acids, and if it connects AG-II and SH26, it is called a basic linker (BL, found in 34 sequences), or basic region (BR) in 10 sequences lacking Nsp and AG-II. In 9 sequences this region does not contain basic residues (non-basic linker, nBL, Table 1). In these 53 sequences, the N-terminal part of the DUF1070 domain is (a part of) BL, BR or nBL. This region considerably contributes to the diversity of the DUF1070 domains. In addition, in sequences containing BL, the SH26 region actually does not start with Ser, but usually with Ile or Val (Supplement B). BL is usually up to 8 amino acids long, and beside basic residues it mostly contains aliphatic amino acids. The rice sequence ABB46792 has an exceptionally long BL containing six Arg and one His. BL and BR are found only in monocots and in few *Rosaceae* sequences (Table 1).

Analysis of *Centaurium CeAGP3* (AGN92423) and its orthologs revealed that SH26 in these sequences represents typical GPIsp [4]. However, since even close sequence homology does not guarantee GPI lipid anchoring throughout a protein family [13], the DUF1070-containing sequences were further analyzed for the presence of GPIsp.

Is a DUF1070 domain a conserved GPI attachment signal sequence?

The prediction of a GPI-anchoring signal peptide in proteins and the position of its ω-site, the residue where GPI is supposed to be attached, is based on the specific requirements that the C-terminus needs to meet in order to be a substrate for the transamidase complex [13,32]. The necessary elements of GPIsp include (i) a polar and flexible linker (ω-11 to ω-1); (ii) small amino acids at positions ω-1 to ω+2, of which ω, ω+1 and ω+2 are usually Ala, Asn, Asp, Cys, Gly or Ser; (iii) a relatively polar spacer of 5-10 residues (usually ω+3 to ω+9), followed by (iv) a hydrophobic tail (ω+10 to the end) [11,13,32]. The consensus sequence of the aligned SH26 regions, “SDGTSIDQGIAYVLMLVALVLTYLIH” (Fig. 1 and Supplement B) perfectly fits the description for elements ii-iv, as it starts with the small amino acids “SDG”, followed by relatively polar “TSIDQG” linker, and ending with a hydrophobic “IAYVLMLVALVLTYLI” tail. The significance of conserved His at the end of the SH26 region is unclear, but few polar or even charged residues may occur at the C-terminal end of the hydrophobic tail [13]. Nevertheless, the GPIsp prediction software, Big-Pi [13], which takes into account the abovementioned position-specific structural and physicochemical requirements and restrictions, found that only 129 out of the 271 analyzed sequences actually have predicted GPIsp (Table 1 and Supplements A and B). Of these 129 sequences, the majority (98) has Ser, the first amino acid of the SH26 motif, as the ω-site, which is in accordance with the finding that Ser is the most preferred amino acid at the ω-site in GPI-anchored proteins [13].

The analysis as to why in 142 sequences the GPIsp was not predicted even though most of them have the described necessary elements (ii-iv) was possible thanks to the detailed Big-Pi output that lists as many as 20 score "Terms" or penalties. It was found that in 83 out of 142 cases where GPIsp was not predicted this was due to a too long propeptide (Term 7 – propeptide length, Table 1). Namely, as many as 151 sequences have an additional C-terminal stretch of variable length, termed the C-stretch (Fig. 1 and Supplement B), which is not part of the DUF1070 domain, even though it is quite conserved. The C-terminal hydrophobic tail (starting with ω+9) has to be between 9 and 24 aa [13], while in 83 sequences with the C-stretch this tail is somewhat longer, resulting in the rejection of these sequences by Big-Pi. However, this criterion is discussible, because in sequence NP_566070 (AtAGP16 from *A. thaliana*) the C-terminal hydrophobic tail is 28 aa long (IAYLLMVVALVLTYLIHPLDASSSYSFF, with C-stretch underlined), and yet this sequence was experimentally proven to be GPI-anchored [10]. NP_566070 is a representative example of this group, since a number of other sequences discarded by Big-Pi have a very similar, if not identical, C-stretch (Supplement B), implying that they also might be anchored. Schultz et al. [10] noted that the length constraint for permissive C termini is just a built-in Big-Pi property, that parameters other than propeptide length are more important for accurate GPIsp prediction, and that the program correctly identifies the ω-sites even in such cases (indicated as black/bold residues both in the Big-Pi output and in Supplement A). Accordingly, the authors suggest that AtAGP20 (NP_191723) is also GPI-anchored, even with its 29 aa-long C-terminal tail (C-stretch: PLDASSSSYTFF). Considering the conservation of the SH26 domain and the fact that in each of 83 cases the ω-site was found when the C-stretch was truncated *in silico*, it can be speculated that all of the discarded 83 sequences are indeed GPI-anchored. As suggested for AtAGP16, the C-stretch could also be preprocessed prior to presentation to the transamidase complex [10], but there is no experimental evidence to support this.

Along with 83 sequences that are considered ambiguous with regard to the GPIsp presence, there are 9 incomplete sequences ≤54 aa, which are discarded by Big-Pi by default. This leaves 50 sequences where GPIsp was not predicted for some other reason (Table 1). Some generalizations can be drawn from the Venn diagram (Fig. 3), were 59 sequences lacking GPIsp (for reasons other than propeptide length) are presented along with subsets of sequences lacking Nsp or AG-II glycomodule, as well as with sequences containing basic linker or region (BL and BR). It is obvious that GPIsp is not predicted in any of the 14 sequences lacking AG-II glycomodule, or in 43 of 44 sequences with BL or BR. This finding emphasizes the importance of the sequence preceding the ω-site, which in the case of DUF1070 members actually has to be AG-II glycomodule and not BL/BR if GPIsp is to be predicted. It is unknown whether in 42 sequences that have both Nsp and AG-II but are not predicted to be GPI-anchored (Fig. 3), the SH26 region serves as a transmembrane domain, and is either proteolytically processed by enzymes other than the transamidase complex, or these proteins are secreted as they are.

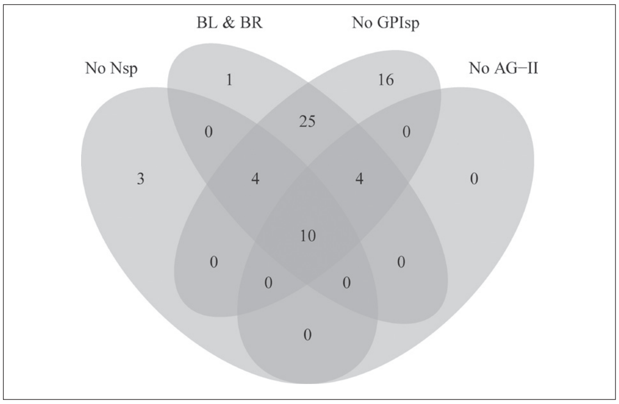

Fig. 3. DUF1070-containing sequences lacking regions typical for AG peptides or containing atypical regions. Presented are subsets of sequences lacking N-terminal signal peptide (**No Nsp**), GPI lipid anchor signal peptide (**No GPIsp**) or AG-II glycomodule (**No AG-II**). The "No GPIsp" group does not include sequences where GPIsp was not predicted only because of the length of the propeptide. Presented are also sequences that possess basic linker or region (**BL & BR**).

Homology and phylogenetic relations among the DUF1070 family members

In order to present structural features and to analyze homology relations among members of the DUF1070 family, the sequences were aligned and an average distance tree was constructed based exclusively on the SH26 region (Supplement B). The average distance tree was clustered at a third of the total distance between the most divergent sequences, resulting in the generation of 21 groups (Table 1 and Supplements A and B). In this alignment, the Nsp (where present) was omitted for the sake of simplicity, while the remaining parts of the sequences were manually split into

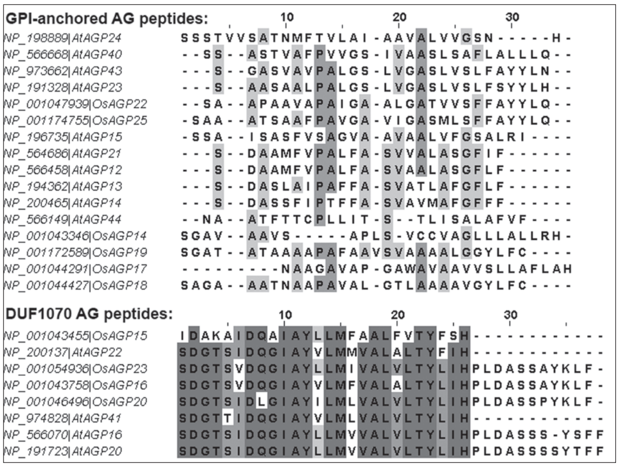
GPI-anchored AG peptides:
10
20
30
NP_198889|AtAGP24 SSSTVVSATNMFTVLAI-AAVALVVGSN-----H-
NP_566668|AtAGP40 --SS--ASTVAFPVVGS-IVAASLSAFLALLLQ--
NP_973662|AtAGP43 ---S--GASVAVPALGS-LVGASLVSLFAYYLN--
NP_191328|AtAGP23 ---S--AASAALPALGS-LVGASLVSLFSYYLH--
NP_001047939|OsAGP22 --SA--APAAVAPAIGA-ALGATVVSFFAYYLQ--
NP_001174755|OsAGP25 -SAA--ATSAAFPAVGA-VIGASMLSFFAYYLQ--
NP_196735|AtAGP15 -SSA--ISASFVSAGVA-AVAALVFGSALRI----
NP_564686|AtAGP21 ---S--DAAMFVPALFA-SVVALASGFIF------
NP_566458|AtAGP12 ---S--DAAMFVPALFA-SVAALASGFLF------
NP_194362|AtAGP13 ---S--DASLAIPAFFA-SVATLAFGFLF------
NP_200465|AtAGP14 ---S--DASSFIPTFFA-SVAVMAFGFFF------
NP_566149|AtAGP44 --NA--ATFTTCPLLIT-S--TLISALAFVF----
NP_001043346|OsAGP14 SGAV--AAVS-----APLS-VCCVAGLLLALLRH-
NP_001172589|OsAGP19 SGAT--ATAAAAPAFAAVSVAAAALGGYLFC----
NP_001044291|OsAGP17 ----------NAAGAVAP-GAWAVAAVVSLLAFLAH
NP_001044427|OsAGP18 SAGA--AATNAAPAVAL-GTLAAAAVGYLFC----
DUF1070 AG peptides:
10
20
30
NP_001043455|OsAGP15 IDAKAIDQAIAYLLMFAALFVTYFSH------------
NP_200137|AtAGP22 SDGTSIDQGIAYVLMMVALALTYFIH------------
NP_001054936|OsAGP23 SDGTSVDQGIAYLLMIVALVLTYLIHPLDASSAYKLF-
NP_001043758|OsAGP16 SDGTSVDQGIAYVLMFVALALTYLIHPLDASSAYKLF-
NP_001046496|OsAGP20 SDGTSIDLGIAYILMLVALVLTYLIHPLDASSPYKLF-
NP_974828|AtAGP41 SDGTTIDQGIAYVLMLVALVLTYLIH-----------
NP_566070|AtAGP16 SDGTSIDQGIAYLLMVVALVLTYLIHPLDASSS-YSFF
NP_191723|AtAGP20 SDGTSIDQGIAYLLMVVALVLTYLIHPLDASSSSYTFF

Fig. 4. Alignment GPI signal sequences and the SH26 and C-stretch domains from *Arabidopsis thaliana* and *Oryza sativa* AG peptides. Darker shades of gray indicate higher conservation.

four motifs: N-terminal part with the conserved Gln, AG-II glycomodule, BL/BR/nBL (if present) and SH26 with C-stretch (where present). The first and largest of the generated groups containing 221 sequences is characterized by Pro-rich AG-II glycomodule and the absence of BL (with few exceptions, Table 1). Members of this group have a highly conserved SH26 region, almost always starting with Ser. In this group, the GPIsp was predicted in 122 cases and experimentally proven in one, and where GPIsp was not predicted, this was due to a longer C-stretch. No taxonomical regularity was observed within this group: on the contrary, representatives of *Lycopodiophyta*, *Gymnospermae*, dicots and monocots, as well as sequences from sequenced plant genomes, were represented, suggesting that peptides of this type are ubiquitous in vascular plants. For example, of 16 *Arabidopsis* AG peptides [1], 14 are predicted to be GPI-anchored [1,10], of which four are DUF1070 members: AtAGP16 (NP_566070), AtAGP20 (NP_191723), AtAGP22 (NP_200137) and AtAGP41 (NP_974828), all classified in group 1 (Fig. 4 and Supplement B).

The remaining 50 sequences were more heterogeneous and distributed in 20 groups, of which group 13 was the largest, with 18 members (Table 1 and Supplement B). The majority of these 50 sequences have BL and no predicted GPIsp. Most of these 50 sequences belong to *Poaceae*, suggesting diversification of the DUF1070 members among the monocots. One example of this diversification is rice, where of 15 AG peptides identified in the genome, 11 are predicted to be GPI-anchored according to Ma and Zhao [7], and four are DUF1070 members: OsAGP15, OsAGP16, OsAGP20 and OsAGP23 (accessions NP_001043455, NP_001043758, NP_001046496 and NP_001054936, respectively). However, our prediction of GPIsp presence in rice AG peptides completely differs from that of Ma and Zhao [7], even though we used the same Big-Pi software: we found that only six rice AG peptides have GPIsp, none of which is a DUF1070 member. Sequences NP_001043758, NP_001046496 and NP_001054936 are among the previously discussed 83 sequences with longer propeptides, and belong to conservative group 1, while diversification is seen in NP_001043455, which has an AG-II domain, but also has BL and is not predicted to be GPI-anchored, and it fell into group 13. Nevertheless, it should be pointed out that Ma and Zhao [7] noticed the conserved motif that they termed the "SDGT region" (the N-terminal part of SH26) and clustered all four *Arabidopsis* and three rice DUF1070 peptides (OsAGP16, OsAGP20 and OsAGP23) together. They missed OsAGP15, but erroneously included OsAGP26 (NP_001057717), which neither has "SDGT" nor is related to DUF1070. The alignment of *Arabidopsis* and rice AG peptides with GPIsp and those with DUF1070 (Fig. 4) illustrates the difference in conservation of DUF1070 regions and other GPIsp sequences.

Among peptides from other groups, sequence XP_003620524 from *Medicago truncatula* is noteworthy since it has two DUF1070 domains, of which the first belong in group 15 and the second in group 16. The unique presence of two DUF1070 domains as well as two AG-II glycomodules might be either a result of domain duplication or a misassembled contig. In addition, *M. truncatula* has another sequence in group 15 (KEH27090), one in group 14 (XP_003620523) and as many as seven in group 1.

Functions of DUF1070 peptides

Since a vast majority of the DUF1070-containing sequences are short, containing Nsp and classical AG-II glycomodule, it can be concluded that they represent AG peptides, most of which are predicted to be GPI-anchored. There is limited information on the expression of DUF1070 AG peptides in *Arabidop-*

sis and rice. AtAGP20 and AtAGP41 have very low expression in different tissues [7,10], so nothing is known about their function. AtAGP16 is expressed thoughout the plant, while AtAGP22 is primarily expressed in roots and pollen [7]. Ma and Zhao [7] have investigated the expression of rice AG peptides and found that OsAGP15 is expressed in leaves, panicles, mature pollen and seeds, and that it is upregulated by salt stress, drought and ABA treatment. OsAGP16 is specifically expressed in pollen and might be involved in pollen development. OsAGP20 is expressed in vegetative meristems, in roots and inflorescence, and is upregulated by cold stress and gibberellins and downregulated by salt stress, drought and ABA. OsAGP23 is expressed in vegetative tissues and is downregulated by drought and salt stress. Gene CAN82800 from *Vitis vinifera* was found to be upregulated in grapevine overexpressing the C-repeat binding factor gene VvCBF4, and thus might be involved in freezing tolerance [33]. Gene CeAGP3 from *C. erythraea* (AGN92423) has very low expression in leaves and roots, but is strongly induced during morphogenesis *in vitro*, particularly during somatic embryogenesis, and, to a lesser extent, during direct organogenesis and rhizogenesis [4]. It would be interesting to explore the functions of these and other GPI-anchored DUF1070 members, because GPI anchoring extends the signaling possibilities of AGPs: GPI-anchored AGPs can be later processed by GPI-specific phospholipases and glycosidases, thus releasing diffusible AGPs and/or their parts as extracellular signals and biologically active lipids as intracellular signals [1,2,5,11,12]. The functions of the more distant and diversified members of the monocot species, particularly those lacking AG-II glycomodule, are unknown.

Finally, what is really intriguing about this family is that the most conserved part of the analyzed sequences, the SH26 region, in the majority of sequences is predicted to be cleaved and discarded during the GPI attachment process. Knowing that only the basic structure, but not the actual sequence of GPIsp is conserved [12] (Fig. 4), raises the question of why a sequence predicted to be discarded is so conserved if this is not required for transamidase complex substrate specificity? The endoplasmic reticulum translocon can distinguish GPIsp from other hydrophobic sequences, thus allowing GPIsp sequences be fully translocated into the ER lumen [34]. This means that the SH26 region, in cases when it is recognized by the ER translocon and the transamidase complex, remains in the ER lumen and is then probably secreted by vesicular transport. Considering that sequence conservation usually implies some function, it remains to be elucidated whether cleaved SH26 regions have any function of their own.

CONCLUSIONS

The DUF1070 domain is a signature motif of a subset of AG peptides, which in the majority of sequences represents typical GPIsp. The DUF1070 peptides appear to be conserved and ubiquitous in vascular plants, but in monocots some of the members show significant diversification. Most of the family members are probably stable peptides, resistant to aminopeptidases. The significance of conservation of the SH26 region, particularly in sequences where it is predicted to be cleaved as GPIsp, is unknown, but implies that SH26 has some additional function. We suggest that the conserved SH26 region, rather than entire DUF1070 domain, should be used in the mining and annotation of AG peptides.

Acknowledgments: This work was financially supported by the Serbian Ministry of Education, Science and Technological Development, Grant TR31019.

Authors' contributions: ADS conceived and designed the research and wrote the manuscript. MBD performed the calculations and sequence manipulations in R, designed the figures and drafted the manuscript. MDB and MMT helped in sequence analyses. ARS and SIT helped in data interpretation. All authors read and approved the manuscript.

Conflict of interest disclosure: The authors declare that they have no conflict of interest.

REFERENCES

1. Showalter AM. (2001) Arabinogalactan-proteins: structure, expression and function. Cell Mol Life Sci. 2001;58(10):1399-417.
2. Seifert GJ, Roberts K. The biology of arabinogalactan proteins. Annu Rev Plant Biol. 2007;58:137-61.
3. Ellis M, Egelund J, Schultz CJ, Bacic A. Arabinogalactan-proteins: key regulators at the cell surface? Plant Phys. 2010;153(2):403-19.
4. Simonović AD, Filipović BK, Trifunović MM, Malkov SN, Milinković VP, Jevremović SB, Subotić AR. Plant regenera-

tion in leaf culture of *Centaurium erythraea* Rafn. Part 2: The role of arabinogalactan proteins. Plant Cell Tiss Org. 2015;121(3):721-39.

5. Schultz CJ, Johnson KL, Currie G, Bacic A. The classical arabinogalactan protein gene family of Arabidopsis. Plant Cell. 2000;12(9):1751-67.
6. Schultz CJ, Rumsewicz MP, Johnson KL, Jones BJ, Gaspar YM, Bacic A. Using genomic resources to guide research directions. The arabinogalactan protein gene family as a test case. Plant Physiol. 2002;129(4):1448-63.
7. Ma H, Zhao J. Genome-wide identification, classification, and expression analysis of the arabinogalactan protein gene family in rice (Oryza sativa L.). J Exp Bot. 2010;61(10):2647-68.
8. Tan L, Showalter AM, Egelund J, Hernandez-Sanchez A, Doblin MS, Bacic A. Arabinogalactan-proteins and the research challenges for these enigmatic plant cell surface proteoglycans. Front Plant Sci. 2012;3:140.
9. Gaspar Y, Johnson KL, McKenna JA, Bacic A, Schultz CJ. The complex structures of arabinogalactan-proteins and the journey towards understanding function. Plant Mol Biol. 2001;7:161-76.
10. Schultz CJ, Ferguson KL, Lahnstein J, Bacic A. Post-translational Modifications of Arabinogalactan-peptides of *Arabidopsis thaliana* endoplasmic reticulum and glycosylphosphatidylinositol-anchor signal cleavage sites and hydroxylation of proline. J Biol Chem. 2004;279(44):45503-11.
11. Schultz C, Gilson P, Oxley D, Youl J, Bacic A. GPI-anchors on arabinogalactan-proteins: implications for signalling in plants. Trends Plant Sci. 1998;3(11):426-31.
12. Borner GH, Sherrier DJ, Stevens TJ, Arkin IT, Dupree P. Prediction of glycosylphosphatidylinositol-anchored proteins in Arabidopsis. A genomic analysis. Plant Phys. 2002;129(2):486-99.
13. Eisenhaber B, Wildpaner M, Schultz CJ, Borner GHH, Dupree P, Eisenhaber F. Glycosylphosphatidylinositol lipid anchoring of plant proteins. Sensitive prediction from sequence- and genome-wide studies for Arabidopsis and rice. Plant Physiol. 2003;133(4):1691-701.
14. Bateman A, Coggill P, Finn RD. DUFs: families in search of function. Acta Crystallogr Sect F Struct Biol Cryst Commun. 2010;66(10):1148-52.
15. Goodacre NF, Gerloff DL, Uetz P. Protein domains of unknown function are essential in bacteria. mBio. 2013;5(1):e00744-13.
16. Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, Geer RC, He J, Gwadz M, Hurwitz DI, Lanczycki CJ, Lu F, Marchler GH, Song JS, Thanki N, Wang Z, Yamashita RA, Zhang D, Zheng C, Bryant SH. CDD: NCBI's conserved domain database. Nucleic Acids Res. 2015;43:D222-6.
17. Geer LY, Domrachev M, Lipman DJ, Bryant SH. CDART: protein homology by domain architecture. Genome Res.2002;12(10):1619-23.
18. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview Version 2 – a multiple sequence alignment editor and analysis workbench. Bioinformatics. 2009;25(9):1189-91.
19. Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods. 2011;8:785-6.
20. Emanuelsson O, Nielsen H, Brunak S, von Heijne G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol. 2000;300(4):1005-16.
21. Lassmann T, Sonnhammer E. Kalign - an accurate and fast multiple sequence alignment algorithm. BMC Bioinformatics. 2005;6:298.
22. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. P Natl Acad Sci USA. 1992;89:10915-9.
23. Wickham H. stringr: Make it easier to work with strings; R package version 0.6.2. 2012.
24. R Core Team. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2014.
25. Wickham H. ggplot2: elegant graphics for data analysis. New York: Springer Science & Business Media; 2009.
26. Auguie B. gridExtra: functions in Grid graphics. R package version 0.9.1. 2012.
27. Chen H. VennDiagram: Generate high-resolution Venn and Euler plots. R package, version 1.6.9. 2014.
28. Popper Z. Evolution and diversity of green plant cell walls. Curr. Opin. Plant Biol. 2008;11(3)286-92.
29. Showalter AM, Keppler BD, Lichtenberg J, Gu D, Welch LR. A bioinformatics approach to the identification, classification, and analysis of hydroxyproline-rich glycoproteins. Plant Phys. 2010;153:485-513.
30. Kumar A, Bachhawat AK. Pyroglutamic acid: throwing light on a lightly studied metabolite. Curr Sci. 2012;102(2):288-97.
31. Schilling S, Stenzel I, von Bohlen A, Wermann M, Schulz K, Demuth HU, Wasternack C. Isolation and characterization of the glutaminyl cyclases from *Solanum tuberosum* and *Arabidopsis thaliana*: implications for physiological functions. Biol Chem. 2007;388(2):145-53.
32. Eisenhaber B, Bork P, Eisenhaber F. Sequence properties of GPI-anchored proteins near the omega-site: constraints for the polypeptide binding site of the putative transamidase. Protein Eng. 1998;1(12):1155-61.
33. Tillett RL, Wheatley MD, Tattersall EA, Schlauch KA, Cramer GR, Cushman JC. The Vitis vinifera C-repeat binding protein 4 (VvCBF4) transcriptional factor enhances freezing tolerance in wine grape. Plant Biotechnol J. 2012;10(1):105-24.
34. Dalley JA, Bulleid NJ. The endoplasmic reticulum (ER) translocon can differentiate between hydrophobic sequences allowing signals for glycosylphosphatidylinositol anchor addition to be fully translocated into the ER lumen. J Biol Chem. 2003;278(51):51749-57.

Supplementary Material

Supplementary File 1. Tabular presentation of characteristics of all analyzed sequences. The data can be viewed via the following link:
http://www.serbiosoc.org.rs/sup/SupplementaryFile1.xls

Supplementary File 2. Alignment and distance tree of all analyzed sequences. The data can be viewed via the following link:
http://www.serbiosoc.org.rs/sup/SupplementaryFile2.png