Family-Specific Gains and Losses of Protein Domains in the Legume and Grass Plant Families

Evolutionary Bioinformatics
Volume 16: 1–15

DOI: 10.1177/1176934320939943

Akshay Yadav[1], David Fernández-Baca[2] and Steven B Cannon[3]

[1]Bioinformatics and Computational Biology Graduate Program, Iowa State University, Ames, IA, USA. [2]Department of Computer Science, Iowa State University, Ames, IA, USA. [3]Corn Insects and Crop Genetics Research Unit, USDA-Agricultural Research Service, Ames, IA, USA.

ABSTRACT: Protein domains can be regarded as sections of protein sequences capable of folding independently and performing specific functions. In addition to amino-acid level changes, protein sequences can also evolve through domain shuffling events such as domain insertion, deletion, or duplication. The evolution of protein domains can be studied by tracking domain changes in a selected set of species with known phylogenetic relationships. Here, we conduct such an analysis by defining domains as "features" or "descriptors," and considering the species (target + outgroup) as instances or data-points in a data matrix. We then look for features (domains) that are significantly different between the target species and the outgroup species. We study the domain changes in 2 large, distinct groups of plant species: legumes (Fabaceae) and grasses (Poaceae), with respect to selected outgroup species. We evaluate 4 types of domain feature matrices: domain content, domain duplication, domain abundance, and domain versatility. The 4 types of domain feature matrices attempt to capture different aspects of domain changes through which the protein sequences may evolve—that is, via gain or loss of domains, increase or decrease in the copy number of domains along the sequences, expansion or contraction of domains, or through changes in the number of adjacent domain partners. All the feature matrices were analyzed using feature selection techniques and statistical tests to select protein domains that have significant different feature values in legumes and grasses. We report the biological functions of the top selected domains from the analysis of all the feature matrices. In addition, we also perform domain-centric gene ontology (dcGO) enrichment analysis on all selected domains from all 4 feature matrices to study the gene ontology terms associated with the significantly evolving domains in legumes and grasses. Domain content analysis revealed a striking loss of protein domains from the Fanconi anemia (FA) pathway, the pathway responsible for the repair of interstrand DNA crosslinks. The abundance analysis of domains found in legumes revealed an increase in glutathione synthase enzyme, an antioxidant required from nitrogen fixation, and a decrease in xanthine oxidizing enzymes, a phenomenon confirmed by previous studies. In grasses, the abundance analysis showed increases in domains related to gene silencing which could be due to polyploidy or due to enhanced response to viral infection. We provide a docker container that can be used to perform this analysis workflow on any user-defined sets of species, available at https://cloud.docker.com/u/akshayayadav/repository/docker/akshayayadav/protein-domain-evolution-project.

KEYWORDS: Protein domain evolution, legumes, grasses, mutual-information, Fisher's exact test, Wilcoxon Rank-Sum test

RECEIVED: June 9, 2020. **ACCEPTED:** June 15, 2020.

TYPE: Original Research

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported in part by the NSF project "Federated Plant Database Initiative for the Legumes," award #1444806, and by the US Department of Agriculture, Agricultural Research Service, project 5030-21000-069-00D. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the US Department of Agriculture. USDA is an equal opportunity provider and employer.

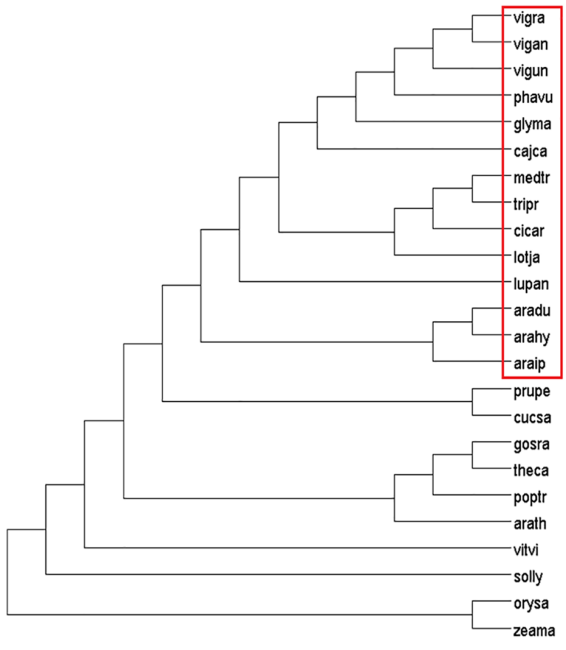
DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Akshay Yadav, Bioinformatics and Computational Biology Graduate Program, Iowa State University, Ames, IA 50011, USA.
Email: aayadav@iastate.edu

Introduction

Protein domains are independent evolutionary units of proteins that enable proteins to evolve in a modular fashion through domain insertion, deletion, duplication, or substitution, in addition to evolution through point mutations.[1,2] In this ability of protein domains to fold and function independently of other domains, they can be considered as "lego bricks" that can be recombined in various ways to build new proteins.[3,4] Small proteins are usually made up of just one domain, whereas large proteins are formed by combinations of multiple domains.[5] Roughly two-thirds of the prokaryotic proteins and four-fifths of the eukaryotic proteins are multidomain proteins that are formed through recombination of 2 or more domains.[6,7] The "combinability" of domains makes them prime candidates for studying evolution—both of proteins and species. For example, protein domains have been used to study evolution on genome-wide and species-wide scales by examining the protein-domain content of the species.[8-10] Protein-domain content is defined by the presence or absence of protein domains in complete genomes of the species. The importance of protein domains in studying evolution can be verified from the ability of protein-domain content in reconstructing the phylogeny of life, in comparison to trees obtained from standard phylogenetic and phylogenomic approaches that utilize information from molecular markers, gene content, and gene order.[10]

In this study, we examine the domain combinations present in 2 groups of plant species—the legumes (Fabaceae) and grasses (Poaceae), treating the protein domains as species "features" that may be present or absent in the focal species. Accordingly, a data matrix was defined with rows representing species, columns representing the protein domains, and the cells containing domain feature values for the respective species. We used standard feature selection and statistical testing techniques to identify protein domains that differ between the target set of species and their respective outgroups.



Content	Duplication	Abundance	Versatility
FANCI_S2	*DUF812*	*Tmemb_14*	*zf-UDP*
absent	1	12.0718	0.3333
absent	1	12.6208	0.3333
absent	1	12.4968	0.5
absent	1	12.3965	0.5
absent	1	12.3422	0.5
absent	1	12.3898	0.25
absent	0	12.8436	0.5
absent	1	12.3924	0.3333
absent	1	12.249	0.5
absent	1	12.5598	0.5
absent	1	12.3356	0.5
absent	1	12.1926	0.5
absent	1	12.2013	0.5
absent	1	12.3197	0.5
present	2	11.9889	1
absent	2	12.0285	1
present	2	12.0135	1
present	2	11.9136	1
present	2	11.6849	1
present	2	11.5728	1
present	2	11.8553	1
present	2	11.9151	1
present	0	11.6984	1
present	0	11.2111	0.25

Figure 1. Phylogeny of legumes with legume outgroups (left) and table (right) showing the 4 types of domain changes analyzed in this study using example domains mentioned in the second row of the table.

Gain or loss of particular domains in a group of species can provide a means of understanding trait evolution in those species.[11,12] Protein domains can duplicate locally, giving significantly different counts of certain domains. This may provide some useful information about functions associated with those domains.[13,14] Counts of protein domains can also increase or decrease along with the proteins that they comprise.[15] Finally, "versatile" domains can partner with multiple different domains; and versatility values can be used to study the evolution of associated functions.[3,16,17] We evaluated domain evolution using these types of domain feature matrices: domain content, duplication, abundance, and versatility.

We used 2 types of statistical methods: mutual-information (MI) and nonparametric statistical tests. MI measures mutual dependence between 2 random variables by quantifying the amount of information communicated about one random variable from another random variable.[18] MI has been routinely used for selecting meaningful features, in classification and pattern recognition problems.[19-21] Here, we used MI to quantify the mutual dependence between domain feature values and the classification between target and outgroup species. We also employed tests for significance of differences in domain feature values between the target and outgroup species. We applied Fisher's exact tests[22] for feature matrices containing discrete values, and Wilcoxon rank-sum tests[23] for feature matrices containing continuous values.

Material and Methods

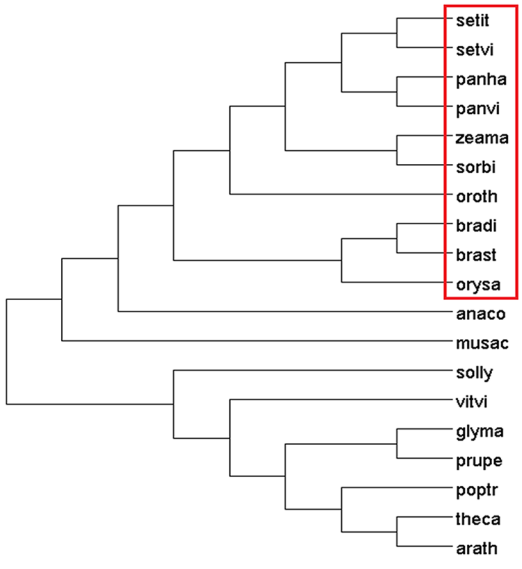
We used 2 sets of plant species to study the species-level changes in protein-domain characteristics for a given set of target species (Figures 1 and 2). The first set consisted of 14 legumes (from the Papilionoideae subfamily within the legume/Fabaceae family), and 10 outgroup species defined concerning the legumes (Table 1).[24-45] The second set consisted of 10 grass species (Poaceae) and 9 outgroup species defined concerning the grasses (Table 2).[27,36,37,39,40,42-54]

All target proteomes from legumes and grasses, together with their respective outgroup proteomes, were searched against domain Hidden Markov Models (HMMs) from the Pfam database (release 32)[55] to assign domains to the protein sequences. The *pfam_scan.pl* script[56] was used to assign domains to proteomes, which internally uses the *hmmscan* program from the HMMER package.[57] Subsequently, the domain assignments from target proteomes and their respective outgroup proteomes were used to calculate the 4 types of domain feature matrices.

Calculation of domain feature matrices

The domain content matrix was calculated to represent the presence or absence of domains in target and outgroup species. Columns of the content matrix represent individual Pfam domains and rows represent species. Each cell was assigned a value of "1" if the corresponding domain was detected in the species, else the cell was a value of "0." Columns with domains that were present in all the target and outgroup species were uninformative and therefore removed.

The domain duplication matrix contains the most frequent copy number of each Pfam domain in species, which was calculated as the modal value of list all possible copy counts of that domain in the corresponding species. The modal value of the list was added to each domain column and corresponding species row. Columns with constant duplication values across all



Content	Duplication	Abundance	Versatility
P_C	*Sec39*	*E1_FCCH*	*Mur_ligase_M*
present	1	12.2434	1
present	1	12.229	1
present	1	12.4331	1
present	1	12.5435	1
present	1	12.4742	1
present	1	12.4644	1
present	1	12.4178	1
present	1	12.4316	1
present	1	12.3605	1
present	1	12.5057	1
absent	1	14.0995	0.5
absent	2	13.0542	0.5
absent	2	13.5001	0.5
absent	2	13.1773	0.5
absent	2	13.3422	0.5
absent	2	13.3108	0.5
absent	2	13.2699	0.5
absent	2	13.2356	0.5
absent	2	13.3801	0.5

Figure 2. Phylogeny of grasses with grass outgroups (left) and table (right) showing the 4 types of domain changes analyzed in this study using example domains mentioned in the second row of the table.

Table 1. Legumes and legume outgroups used to study protein-domain evolution in the legumes.

SPECIES	ABBREV.	CLASS	GENOTYPE	ASSEMBLY	ANNOT.	PUBLICATION	SOURCE
Arachis duranensis	aradu	Legume	V14167	1	1	Bertioli *et al*[24]	PeanutBase
Arachis ipaensis	araip	Legume	K30076	1	1	Bertioli *et al*[24]	PeanutBase
Arachis hypogaea	arahy	Legume				Bertioli *et al*[24]	PeanutBase
Cajanus cajan	cajca	Legume	ICPL87119	1	1	Varshney *et al*[25]	LegumeInfo
Cicer arietinum	cicar	Legume	Frontier	1	1	Varshney *et al*[26]	LegumeInfo
Glycine max	glyma	Legume	Williams 82	2	1	Schmutz *et al*[27]	Phytozome
Lotus japonicus	lotja	Legume	MG20	3	1	Sato *et al*[28]	Phytozome
Lupinus angustifolius	lupan	Legume				Hane *et al*[29]	LegumeInfo
Medicago truncatula	medtr	Legume	A17_HM341	4	2	Tang *et al*[30]	Phytozome
Phaseolus vulgaris	phavu	Legume	G19833	2	1	Schmutz *et al*[31]	Phytozome
Trifolium pretense	tripr	Legume				De Vega *et al*[32]	LegumeInfo
Vigna angularis	vigan	Legume	Va3.0	1	3	Kang *et al*[33]	LegumeInfo
Vigna radiate	vigra	Legume	VC1973A	6	1	Kang *et al*[34]	LegumeInfo
Vigna unguiculata	vigun	Legume	IT97K	1	1	Phytozome[35]	Phytozome
Prunus persica	prupe	Outgroup	Lovell	2	2.1	IPGI[36]	Phytozome
Vitis vinifera	vitvi	Outgroup	PN40024	12X	12X	Jaillon *et al*[37]	Phytozome
Cucumis sativus	cucsa	Outgroup		1	1	Phytozome[38]	Phytozome
Arabidopsis thaliana	arath	Outgroup	Col-0	TAIR10	TAIR10	Berardini *et al*[39]	Phytozome
Solanum lycopersicum	solly	Outgroup	LA1589	ITAG2.4	ITAG2.4	Tomato Genome Consortium[40]	Phytozome

(Continued)

Table 1. (Continued)

SPECIES	ABBREV.	CLASS	GENOTYPE	ASSEMBLY	ANNOT.	PUBLICATION	SOURCE
Gossypium raimondii	gosra	Outgroup		2	2.1	Paterson *et al*[41]	Phytozome
Oryza sativa	orysa	Outgroup		7	7.0	Ouyang *et al*[42]	Rice Genome Annotation Project
Populus trichocarpa	poptr	Outgroup		3	3.1	Tuskan *et al*[43]	Phytozome
Theobroma cacao	theca	Outgroup		2	2.1	Motamayor *et al*[44]	Cacao Genome Project
Zea mays	zeama	Outgroup		6	6a	Schnable *et al*[45]	

Table 2. Grasses and grass outgroups used to study protein-domain evolution in the grasses.

SPECIES	ABBREV.	CLASS	GENOTYPE	ASSEMBLY	ANNOT.	PUBLICATION	SOURCE
Setaria italica	Setit	Grass	Yugu1	2	2.2	Bennetzen *et al*[46]	Phytozome
Setaria viridis	Setvi	Grass		2	2.1	Phytozome[47]	Phytozome
Panicum hallii	Panha	Grass	filipes	3	3.1	Phytozome, 2017	Phytozome
Panicum virgatum	Panvi	Grass		5	5.1	Phytozome[48]	Phytozome
Zea mays	Zeama	Grass		6	6a	Schnable *et al*[45]	
Sorghum bicolor	Sorbi	Grass		3.1	3.1.1	McCormick *et al*[49]	Phytozome
Oropetium thomaeum	Oroth	Grass		1	1.0	VanBuren *et al*[50]	Phytozome
Brachypodium distachyon	Bradi	Grass		3	3.1	International Brachypodium Initiative[51]	Phytozome
Brachypodium stacei	Brast	Grass		1	1.1	Phytozome[52]	Phytozome
Oryza sativa	Orysa	Grass		7	7.0	Ouyang *et al*[42]	Rice Genome Annotation Project
Arabidopsis thaliana	Arath	Outgroup	Col-0	TAIR10	TAIR10	Berardini *et al*[39]	Phytozome
Theobroma cacao	Theca	Outgroup		2	2.1	Motamayor *et al*[44]	Cacao Genome Project
Populus trichocarpa	Poptr	Outgroup		3	3.1	Tuskan *et al*[43]	Phytozome
Prunus persica	Prupe	Outgroup	Lovell	2	2.1	IPGI[36]	Phytozome
Glycine max	Glyma	Outgroup	Williams 82	2	1	Schmutz *et al*[27]	Phytozome
Vitis vinifera	Vitvi	Outgroup	PN40024	12X	12X	Jaillon *et al*[37]	Phytozome
Solanum lycopersicum	Solly	Outgroup	LA1589	ITAG2.4	ITAG2.4	Tomato Genome Consortium[40]	Phytozome
Ananas comosus	Anaco	Outgroup		3	3	Ming *et al*[53]	Phytozome
Musa acuminata	music	Outgroup		1	1	Droc *et al*[54]	Banana Genome Hub

the species (target + outgroup) were removed from the matrix. Also, columns with domain duplication values $\leqslant 1$ across all the rows were removed.

The domain abundance matrix was built to represent the abundance value of protein domains in target and outgroup species. Here, we define the abundance value of each domain in each species as the proportion of protein sequences from the entire proteome that contains the domain. The abundance value of each domain in each species is calculated using the inverse domain frequency (IDF) function (equation 1) which is inspired

by the inverse document frequency function used in text mining and natural language processing (NLP) applications

$$\mathrm{IDF}(S,d) = \log_2 \frac{\mathrm{N}(S)}{\mathrm{N}(S,d)} \quad (1)$$

where N(S) is the total number of proteins in species "S" and N(S, d) is the number of proteins containing domain "d" in species "S"

The domain versatility matrix was calculated to represent the changes in the versatility values of the domains across the species. Versatility value (equation 2) for a given domain and species combination was calculated as the reciprocal of the number of domains immediately adjacent to the given domain in protein sequences in the corresponding species. Here too, the columns with constant versatility values across all species (target + outgroup) were removed from the matrix

$$V(S,d) = \frac{1}{F(S,d)} \quad (2)$$

where F(S, d) is the number of different domains adjacent to domain "d" in species "S"

Finally, an additional "species label" column containing value "1" for target species and "0" for outgroup species was attached to all 4 domain feature matrices to represent the classification between target and outgroup species.

Statistical analysis of domain feature matrices

We applied 2 types of statistical analyses to the domain feature matrices. The MI function (equation 3) was used to calculate the MI score for each domain feature by comparing it against the species label column. The MI quantity measures how much information, on average, is communicated in the domain feature column about the classification between target and outgroup species (species label column). Feature columns of the duplication and abundance matrices were subjected to "L^2" normalization before application of MI scoring. The L^2 normalization technique modifies the column values such that in each column, the sum of the squares will always have a maximum value of 1

$$I(X;Y) = \sum_{x\in X}\sum_{y\in Y} P(x,y)\, log\, \frac{P(x,y)}{P(x)P(y)} \quad (3)$$

We also tested feature columns for significance, calculating P values to measure the difference in domain feature values between target and outgroup species. We used Fisher's exact test to evaluate feature columns from the duplication and versatility matrices. Fisher's exact test was applied to contingency tables built using the discrete values from each domain column and the species labels. The dimensions of the contingency tables, in case of the content matrix, were always 2×2 because each domain column can have only 2 possible values for each species row—whereas, in case of duplication and versatility matrices, the dimensions were $r \times 2$, where "r" is the number of discrete values observed in the corresponding domain column. The Wilcoxon rank-sum test was applied for significance testing of the domain abundance matrix due to the continuous values of the domain features. The P values obtained for domains were corrected for multiple testing using the false discovery rate (FDR) method.[58] The FDR-adjusted P values were reported for the domains.

Table 3. Domains gained or lost in legumes concerning legume outgroups (top 11 by MI score).

DOMAIN NAME	MI SCORE	FDR-ADJUSTED *P* VALUES	GAIN-LOSS (+/–) STATUS
FANCI_S2	0.6082	.0017	–
FANCI_S1	0.5804	.0017	–
FANCI_HD1	0.5804	.0017	–
SHNi-TPR	0.5781	.0017	+
WD-3	0.5666	.0017	–
TPMT	0.5527	.0017	–
FANCI_HD2	0.5527	.0017	–
FANCI_S4	0.5527	.0017	–
FA_FANCE	0.516	.0099	–
FANCL_C	0.4604	.0099	–
FANCF	0.4395	.0099	–

Abbreviations: MI, mutual information; FDR, false discovery rate.

Results

All 4 types of domain feature matrices were calculated for 2 sets of plants—the first containing 14 legume and 10 outgroup species, and the second containing 10 grass and 9 outgroup species. For all feature matrices, we applied MI scoring and significance testing.

Domain content analysis

In legumes and grasses, 13 and 55 domains, respectively, showed significant presence/absence differences relative to their respective outgroups. The results show a loss of 12 domains and gain of the *SHNi-TPR* domain in legumes and loss of 33 domains and gain of 22 domains in grasses. The Pfam domains showing the most significant gain or loss in legumes and grasses are listed in Tables 3 and 4. The gained *SHNi-TPR* domain in the legumes contains an interrupted form of the TPR repeat. The *SHNi-TPR* family includes proteins such as Sim3 (yeast), NASP(Human) and N1/N2(Xenopus), which are responsible

Table 4. Domains gained or lost in grasses concerning grass outgroups (top 10 by MI score).

DOMAIN NAME	MI SCORE	FDR-ADJUSTED *P* VALUES	GAIN-LOSS (+/–) STATUS
P_C	0.7188	.0011	+
Mur_ligase	0.7188	.0011	–
Glutenin_hmw	0.7188	.0011	+
DUF1618	0.7188	.0011	+
MFS18	0.7188	.0011	+
SEO_C	0.7188	.0011	–
ACCA	0.7188	.0011	–
SEO_N	0.7188	.0011	–
DUF1719	0.7188	.0011	+
DUF1110	0.7188	.0011	+

Abbreviations: MI, mutual information; FDR, false discovery rate.

for delivering histone proteins such as H3 to centromeric chromatin.[59] Most of the missing domains in legumes are parts of multidomain proteins found in the Fanconi anemia (FA) pathway. The FA pathway is responsible for maintaining chromosomal stability through the repair of interstrand DNA crosslinks in a replication-dependent manner.[60] Most of the proteins in the FA pathway form a core complex known as the FA core complex which is responsible for the ubiquitination of FANCD2 and FANCI proteins.[61] Both the proteins are then localized to the site of DNA repair along with few other proteins. The FANCI is a multidomain protein made up of 5 domains: *FANCI_S2, FANCI_S1, FANCI_HD1, FANCI_HD2*, and *FANCI_S4*. All the 5 FANCI domains are missing in legumes, which means that the entire FANCI protein is lost in legumes. In addition, FANCD2-binding *FA_FANCE* domain[62] and the C-terminal domain of FANCL protein (*FANCL_C* domain) are also missing in legumes. The missing *WD-3* domain belongs to the family of WD-repeats region, which is approximately 100 residues long and is contained within the FANCL protein, the putative E3 ubiquitin ligase subunit of the FA core complex (p. 40).[63] The only protein involved in the FA pathway that is present in legumes is the single domain FANCD2 nuclease containing the *FancD2* domain.[64] In addition to the domains from the FA pathway, the thiopurine-S-methyltransferase (*TPMT*) domain was also detected as lost from the legumes. This is a cytosolic enzyme involved the catalysis of S-methylation of aromatic and heterocyclic sulfhydryl compounds, such as anticancer and immunosuppressive thiopurines.[65]

Among the top 10 protein domains in grasses, 6 were detected as gained and 4 were detected as lost concerning the grass outgroups. There were 3 domains with unknown functions—*DUF1618, DUF1719, DUF1110*, and 3 domains with known functions—*P_C, Glutenin_hmw, MFS18*, that were detected as present in grasses. The *P_C* domain is present at the C terminus of plant P proteins. The P proteins in maize act as transcriptional regulators of enzymes involved in a red phlobaphene pigment-producing arm of the flavonoid biosynthesis pathway.[66,67] The domain *Glutenin_hmw* is the high molecular subunit of glutenin protein responsible for the elastic properties of gluten. The elastomeric glutenin proteins form a network that can withstand significant deformations without breaking, and return to the original conformation when the stress is removed—the property important for making dough.[68] The male flower specific protein 18 (*MFS18*) domain found in the MFS18 protein in maize is rich in glycine, proline, and serine. The MFS18 mRNA is found to accumulate in a vascular bundle in the glumes, anther walls, paleas, and lemmas of mature florets.[69]

The 4 domains *Mur_ligase, SEO_N, SEO_C*, and *ACCA*, were among the top 10 domains detected as lost in most grasses concerning the selected outgroups. The *Mur_ligase* domain is the catalytic domain found in the Mur ligase family of enzymes that catalyze the successive steps in the synthesis of peptidoglycan.[70] The *SEO_N* and *SEO_C* in domains are respectively found at the *N* and *C* terminus of sieve element occlusion (SEO) proteins also known as phloem proteins or forisomes. These phloem proteins remain associated with cisternae of the endoplasmic reticulum of the sieve elements after differentiation and provide rapid protection against wounding of sieve tubes by forming a gel-like mass.[71] The *ACCA* domain is the alpha isoform of the carboxyltransferase subunit of Acetyl Co-A carboxylase enzyme. The *ACCA* domain is known to play an important role in the production of Malonyl-CoA in fatty acid synthesis.[72]

Domain duplication analysis

Application of MI-scoring and Fisher's exact tests on domain features of duplication matrices revealed a single domain (of unknown function) in legumes and 8 types of domains in grasses that were significantly different (FDR ≤ 0.05) in their copy numbers as compared to the copy numbers observed in their respective outgroup sets. The domain *DUF812* is present in 1 copy in all legume sequences except *Medicago*, and in 2 copies in all outgroups except rice and maize (MI score = 0.519444; FDR = 0.000993). Among the 8 significantly different domains in grasses (Table 5), 4 of the domains have increased in copy numbers and 4 have decreased in copy numbers. The domains *DUF775, SPX, zf-PARP*, and *FANCF* are present in 2 copies in most grass sequences and 1 copy in most outgroup sequences. The *SPX* domain is a 180 residue-long protein domain found at the *N*-terminus of a family of proteins involved in *G*-protein-associated signal transduction.[73-75] The *zf-PARP* domain resides at the amino-terminal region of Poly (ADP-ribose) polymerase protein, which is an important regulatory component in the cellular response to DNA damage.

Table 5. Domains with significant differences in copy numbers between grasses and grass outgroups (top 10 by MI score).

DOMAIN NAME	MI SCORE	FDR-ADJUSTED *P* VALUES	GAIN-LOSS (+/–) STATUS
Sec39	0.6168	.0178	–
Prenyltrans	0.5845	.0339	–
DUF775	0.5642	.0178	+
Nop16	0.5193	.0356	–
SPX	0.4798	.0356	+
zf-PARP	0.4715	.0445	+
mTERF	0.4447	.0356	–
FANCF	0.4329	.0359	+

Abbreviations: MI, mutual information; FDR, false discovery rate.

This domain is known to act as a DNA nick sensor.[76] The *FANCF* domain is present in the FA group F protein involved in FA DNA repair pathway. Inactivation of the FANCF protein induced by methylation may play an important role in the occurrence of ovarian cancers.[77]

The domains *Sec39, Prenyltrans, Nop16*, and *mTERF* show a decrease in copy numbers, with 2, 2, 3 to 5, and 2 copies in most of the outgroup species and 1, 1, 1 to 3 and 1 copies, respectively, in most grasses. The *Sec39* domain is a part of "secretory pathway protein 39," which is involved in ER-Golgi transport.[78,79] The *Prenyltrans* domain-containing enzymes are responsible for the transfer of allylic prenyl groups to acceptor molecules.[80,81] The *Nop16* domain is part of a protein involved in ribosome biogenesis.[82] The *mTERF* protein domain is a part of the "mitochondrial transcription termination factor" (mTERF) protein, containing 3 leucine zipper motifs, and known to bind to the DNA.[83]

Domain abundance analysis

The analysis of domain abundance matrices revealed 111 domains in legumes and 497 domains in grasses that have expanded or contracted significantly (FDR ≤ 0.05), as compared to their respective outgroup sets. In the legumes relative to outgroups, 51 domains have expanded significantly in abundance and 60 domains have contracted. In the grasses, 196 domains have expanded significantly in abundance and 301 domains have contracted. The top 10 significantly expanded or contracted domains in legumes and grasses are listed in Tables 6 and 7.

Among the top 10 domains showing expansions or contractions in abundance in the legumes, the *ThylakoidFormat, GST_C_6, DUF726, FERM_M, DAO_C, Aa_trans*, and *SURNod19* domains have expanded, and the *Tmemb_14, DUF724*, and *DUF563* domains have contracted. The thylakoid formation protein (*ThylakoidFormat*) domain is present in the outer plastid membrane and the stroma. This protein is known to have roles in sugar signaling, chloroplast and leaf development, and vesicle-mediated thylakoid membrane biogenesis.[84] The C-terminal domain of Glutathione-S-transferase (*GST_C_6*) is known to conjugate reduced glutathione to auxin-regulated proteins in plants.[85] The *FERM_M* domain is the middle domain of FERM protein and is involved in localizing proteins from cytosol to plasma membrane.[86] The *DAO_C* domain is present at the C-terminal

Table 6. Domains with significant differences in abundance values between legumes and legume outgroups (top 10 by MI score).

DOMAIN NAME	MI SCORE	FDR-ADJUSTED *P* VALUES	GAIN-LOSS (+/–) STATUS
Tmemb_14	0.6272	.0199	–
ThylakoidFormat	0.5976	.0199	+
GST_C_6	0.5804	.0462	+
DUF724	0.5698	.0199	–
DUF726	0.5644	.0199	+
FERM_M	0.5399	.0213	+
DUF563	0.5393	.0233	–
DAO_C	0.5325	.0372	+
Aa_trans	0.5284	.0372	+
SURNod19	0.5148	.0233	+

Abbreviations: MI, mutual information; FDR, false discovery rate.

Table 7. Domains with significant difference in abundance values between grasses and grass outgroups (top 11 by MI score).

DOMAIN NAME	MI SCORE	FDR-ADJUSTED *P* VALUES	GAIN-LOSS (+/–) STATUS
E1_FCCH	0.7188	.0159	+
TruD	0.7188	.0159	+
Kelch_6	0.7188	.0159	–
NT-C2	0.7188	.0159	–
Peptidase_C12	0.7188	.0159	+
HD-ZIP_N	0.7188	.0159	–
DUF1442	0.7188	.0159	–
TK	0.7188	.0159	–
SNARE	0.7188	.0159	–
Pec_lyase_C	0.7188	.0159	–
Pectinesterase	0.7188	.0159	–

Abbreviations: MI, mutual information; FDR, false discovery rate.

region of alpha-glycerophosphate oxidase enzyme. The transmembrane region of amino-acid transporter protein (*Aa_trans*) is found in many amino-acid transporters like the amino-butyric acid (GABA) transporter.[87]

The *Tmemb_14* domain is the only one among the 10 domains in Table 6 to have contracted in legumes. This domain belongs to a family of uncharacterized short transmembrane proteins.

Among the top 11 domains in grasses to have expanded or contracted in abundance relative to outgroups, only 3 have expanded—specifically, sequences containing the *E1_FCCH, TruD*, and *Peptidase_C12* domains have increased in abundance the grasses. The *E1_FCCH* domain is found in the E1 family of ubiquitin-activating enzymes,[88] which is involved in protein degradation cascades. The tRNA-pseudouridine synthase D (*TruD*) protein is involved in the synthesis of pseudouridine from uracil-13 in transfer RNAs. The *Peptidase_C12* domain, also known as a Ubiquitin C-terminal hydrolase, is a deubiquitination enzyme involved in hydrolysis of adducts from the C-terminus of ubiquitin.[89]

Sequences containing the *Kelch_6, NT-C2, HD-ZIP_N, DUF1442, TK, SNARE, Pec_lyase_C*, and *Pectinesterase* domains have decreased in proportion in grasses. The Kelch (*Kelch_6*) motif contains about 50 amino acids and is found in a variety of proteins with diverse functions including functions related to actin dynamics and cell adhesion.[90] The *N*-terminal C2 (*NT-C2*) domain is found in plant proteins involved in the regulation of rhizobium-directed polar growth and intracellular movement of chloroplasts in response to blue light.[91] The *HD-ZIP_N* domain is present at the *N*-terminal of plant homeobox-leucine zipper protein which is known to regulate interfascicular fiber differentiation in *Arabidopsis*.[92] The thymidine kinase (*TK*) domain is a phosphotransferase enzyme (EC 2.7.1.21) that catalyzes the transfer of a single phosphate group from adenosine triphosphate (ATP) to thymidine and is required for DNA synthesis in cell division. The *SNARE* domain acts as a module for protein-protein interaction in the assembly of SNARE machinery, which in turn mediates membrane fusion events in eukaryotic cells.[93] The *Pec_lyase_C* domain is a part of the Pectate Lyase enzyme (EC 4.2.2.2), which is known to be involved in maceration and soft rotting of plant tissue and pectin degradation during pollen tube growth.[94,95] The *Pectinesterase* domain is a cell-wall-associated enzyme (EC 3.1.1.11) involved in cell-wall modification and breakdown.[96]

Domain versatility analysis

The analysis of domain versatility matrices revealed a single domain in legumes and 12 domains (Table 8) in grasses with significantly increased or decreased versatility values with respect to their outgroup sets. In legumes, the *zf-UDP* domain co-occurs with 2 to 4 different domains but partners with only one other domain in all outgroup species except maize (FDR-adjusted *P* value = .0019). The *zf-UDP* domain is a RING/U-box type zinc-binding domain frequently found in the catalytic subunit of cellulose synthase enzyme (EC: 2.4.1.12). This enzyme catalyzes the addition of glucose to the growing cellulose from UDP-glucose.

Table 8. Domains with significant differences in versatility values between grasses and grass outgroups.

DOMAIN NAME	MI SCORE	FDR-ADJUSTED *P* VALUES	GAIN-LOSS (+/–) STATUS
Mur_ligase_M	0.7188	.0043	–
CG-1	0.7188	.0043	+
zf-met	0.6344	.0129	–
DOMON	0.6344	.0043	–
WRC	0.6212	.0203	–
RPN13_C	0.6037	.0155	–
HATPase_c	0.5861	.0203	–
CBS	0.5467	.0302	–
Jacalin	0.5291	.035	+
Biotin_lipoyl	0.4715	.0302	–
GST_N	0.4583	.0442	–
zf-CCHC	0.4359	.0412	+

Abbreviations: MI, mutual information; FDR, false discovery rate.

The *CG-1, Jacalin*, and *zf-CCHC* domains have all gained additional domain partners in grasses as compared to their outgroups. The most prominent of the 3, the *CG-1* domain, co-occurs with 2 domains in outgroups but partners with 3 to 4 domains in grasses. Similarly, *Jacalin* and *zf-CCHC* domains also have gained 2 to 5 additional domain partners in grasses. The *CG-1* domains are highly conserved, 130 amino acid long DNA-binding protein domains associated with light signal transduction[97] and calmodulin-binding transcriptional activators containing ankyrin motifs.[98] The *Jacalin* domain is a mannose-binding lectin domain with a beta-prism fold.[99] The zinc knuckle (*zf-CCHC)* domain is a zinc-binding motif composed of the CX2CX4HX4C motif (where X can be any amino acid).

Among the protein domains that have lost domain partners in grasses as compared to the outgroups, the *Mur_ligase_M* domain has the highest MI score value. This is the middle domain found adjacent to the *N*-terminal *Mur_ligase* domain in grass outgroups but has lost the *N*-terminal partner in grasses (as found in the domain content analysis). The *zf-met, DOMON, WRC*, and *RPN13_C* domains also have lost, respectively, 2 to 3, 1 to 3, 1 to 2, and 2 to 3 adjacent domain partners in grasses. The *zf-met* domain is another zinc-finger domain, containing the CxxCx(12)Hx(6)H motif, and is associated with RNA binding. The *DOMON* domain is 110 to 125 residues long and is found in heme- and sugar-binding proteins.[100] The *WRC* domain is known for containing the conserved

Table 9. Enriched GO terms from protein domains that were detected as gained in grasses as compared to grass outgroups.

BIOLOGICAL PROCESS (BP)				
GO ID	GO TERM	CATEGORY	*Z* SCORE	FDR
GO:0009058	Biosynthetic process	Highly general	2.89	2.47e−02
GO:0006725	Cellular aromatic compound metabolic process	Highly general	2.79	2.47e−02
GO:1901360	Organic cyclic compound metabolic process	Highly general	2.7	2.47e−02

Abbreviation: FDR, false discovery rate.

Trp-Arg-Cys motif, along with a putative nuclear localization signal and a zinc-finger motif with involvement in DNA binding. The *RPN13_C* domain is an all-helical C-terminal domain that forms a binding surface for ubiquitin-receptor proteins for deubiquitination.[101,102]

Domain-centric gene ontology enrichment analysis

To check if the significantly evolving domains (FDR ⩽ 0.05), selected from an analysis of feature matrices, map to any particular gene ontology (GO) terms, we used "dcGO," the domain-centric ontology database that provides associations between GO terms and protein domains from Pfam.[103] The GO enrichment analysis was performed on domain lists obtained from the content, duplication, abundance, and versatility matrices from both the species sets, to check for significantly enriched GO terms from the 3 GO subontologies: biological process (BP), cellular component (CC), and molecular function (MF).

GO enrichment analysis was performed for the 13 domains from legumes and 55 domains from grasses, that were identified from the analysis of content matrices. Separate enrichment analyses were performed for domains that were detected as gained in target species and domains that were detected as lost in the target species. No GO term enrichment was found for the single *SHNi-TPR* domain that was gained in legumes concerning the legume outgroups. However, for the 12 domains that seem to have been lost in legumes, weak enrichment (*Z* score = 2.86, FDR = 1.93e−02) was observed for the highly general CC term "nuclear lumen" (GO:0031981). In grasses, weak enrichment for 3 highly general BP terms was found (Table 9) for the 22 domains that seem to be gained concerning the grass outgroup. Again, no GO term enrichments were found for the 33 domains that were detected as lost in grasses concerning their outgroups.

As a single domain of unknown function (*DUF812*) was detected as significantly different in terms of copy number in legumes versus legume outgroups, from the analysis of domain duplication matrices, no enrichment of GO terms was observed in legumes. Similarly, in grasses, the 4 protein domains that show an increase in copy numbers and 4 domains that show a decrease in copy numbers did not contain any enriched GO categories.

In domain-centric GO analyses of domains showing significant increase in abundance values, in legumes, enrichment of 3 BP terms and 5 CC terms (Table 10) was found. There is enrichment in biological metabolic processes involving glycosyl compounds (GO:1901659, FDR = 4.80e−03), ribonucleosides (GO:0009119, FDR = 1.39e−02), and isoprenoids (GO:0008299, FDR = 1.39e−02), with involvement in organelle membranes (GO:0098805, FDR = 1.27e−03).

GO analyses of domains that showed significant decrease in abundance values between legumes and legume outgroups found enrichment of 10 BP terms and 11 MF terms (Table 11). Among the BP terms, strongest enrichment was found for purine nucleobase metabolic process (GO:0006144, FDR = 9.85e−07) and hydrogen peroxide metabolic process (GO:0042743, FDR = 1.25e−03). Among the MF terms, very strong enrichment was observed for specific MF terms such as xanthine dehydrogenase activity (GO:0004854, FDR = 8.10e−10), oxidoreductase activity, acting on CH or CH2 groups, oxygen as acceptor (GO:0016727, FDR = 8.10e−10), oxidoreductase activity, acting on the aldehyde or oxo group of donors, oxygen as acceptor (GO:0016623, FDR = 8.10e−10), molybdopterin cofactor binding (GO:0043546, FDR = 8.10e−10) and 2 iron, 2 sulfur cluster binding (GO:0051537, 8.25e−08).

In grasses, GO enrichments of 16 BP, 5 CC, and 4 MF terms were found for domains that showed significant increase in abundance values in comparison to the abundance values in grass outgroups (Table 12). Strongest enrichment was observed for specific BP term chromatin silencing (GO:0006342, FDR = 2.02e−05) with relatively moderate enrichments for BPs including protein unfolding (GO:0043335, FDR = 4.26e−03), negative regulation of translational initiation (GO:0045947, FDR = 4.12e−03), positive regulation of nuclear-transcribed mRNA poly(A) tail shortening (GO:0060213, FDR = 4.26e−03), miRNA-mediated inhibition of translation (GO:0035278, FDR = 5.63e−03), small RNA loading onto RISC (GO:0070922, FDR = 5.87e−03), production of siRNA involved in RNA interference (GO:0030422, 7.51e−03), mRNA cleavage (GO:0006379, FDR = 7.51e−03) and pre-miRNA processing (GO:0031054, FDR = 7.51e−03). Enrichments in the CC terms correlated with the BP terms, with general and specific CCs like polysome (GO:0005844, FDR = 8.05e−03), RNAi effector complex (GO:0031332, FDR = 2.93e−03), microribonucleoprotein complex (GO:0035068, FDR = 2.93e−03),

Table 10. Enriched GO terms from protein domains that show significant increase in abundance values in legumes as compared to legume outgroups.

GO ID	GO TERM	CATEGORY	*Z* SCORE	FDR
BIOLOGICAL PROCESS (BP)				
GO:1901659	Glycosyl compound biosynthetic process	Specific	10.39	4.80e−03
GO:0009119	Ribonucleoside metabolic process	Specific	7.16	1.39e−02
GO:0008299	Isoprenoid biosynthetic process	Specific	7.16	1.39e−02
CELLULAR COMPONENT (CC)				
GO:0098805	Whole membrane	Highly general	5.28	1.27e−03
GO:0031090	Organelle membrane	Highly general	3.50	1.34e−02
GO:0031300	Intrinsic component of organelle membrane	General	5.46	1.34e−02
GO:0019867	Outer membrane	General	4.88	1.34e−02
GO:0044437	Vacuolar part	General	3.55	4.23e−02

Abbreviation: FDR, false discovery rate.

Table 11. Enriched GO terms from protein domains that show significant decrease in abundance values in legumes as compared to legume outgroups.

GO ID	GO TERM	CATEGORY	*Z* SCORE	FDR
BIOLOGICAL PROCESS (BP)				
GO:0009056	Catabolic process	Highly general	3.68	6.12e−03
GO:0017144	Drug metabolic process	General	5.79	1.42e−04
GO:1901361	Organic cyclic compound catabolic process	General	5.14	1.62e−03
GO:0044270	Cellular nitrogen compound catabolic process	General	4.75	3.56e−03
GO:0046700	Heterocycle catabolic process	General	4.75	3.56e−03
GO:0019439	Aromatic compound catabolic process	General	4.57	4.38e−03
GO:0046113	Nucleobase catabolic process	Specific	15.1	9.85e−07
GO:0006144	Purine nucleobase metabolic process	Specific	15.1	9.85e−07
GO:0072523	Purine-containing compound catabolic process	Specific	11.86	9.22e−06
GO:0042743	Hydrogen peroxide metabolic process	Specific	9.35	1.25e−03
MOLECULAR FUNCTION (MF)				
GO:0016491	Oxidoreductase activity	Highly general	4.87	1.68e−04
GO:0005506	Iron ion binding	General	10.94	6.03e−07
GO:0051536	Iron-sulfur cluster binding	General	9.88	6.66e−06
GO:0016903	Oxidoreductase activity, acting on the aldehyde or oxo group of donors	General	9.18	1.17e−05
GO:0050662	Coenzyme binding	General	4.99	1.37e−03
GO:0042803	Protein homodimerization activity	General	4.52	2.42e−03
GO:0004854	Xanthine dehydrogenase activity	Specific	22.11	8.10e−10
GO:0016727	Oxidoreductase activity, acting on CH or CH2 groups, oxygen as acceptor	Specific	22.11	8.10e−10
GO:0016623	Oxidoreductase activity, acting on the aldehyde or oxo group of donors, oxygen as acceptor	Specific	22.11	8.10e−10
GO:0043546	Molybdopterin cofactor binding	Specific	22.11	8.10e−10
GO:0051537	2 iron, 2 sulfur cluster binding	Specific	15.49	8.25e−08

Abbreviation: FDR, false discovery rate.

Table 12. Enriched GO terms from protein domains that show significant increase in abundance values in grasses as compared to grasses outgroups.

GO ID	GO TERM	CATEGORY	*Z* SCORE	FDR
BIOLOGICAL PROCESS (BP)				
GO:0006950	Response to stress	Highly general	3.65	7.51e−03
GO:0009056	Catabolic process	Highly general	3.76	8.06e−03
GO:0016458	Gene silencing	General	7.49	4.42e−05
GO:0040029	Regulation of gene expression, epigenetic	General	6.10	9.47e−04
GO:0009615	Response to virus	General	5.12	5.87e−03
GO:0098542	Defense response to other organisms	General	4.58	5.87e−03
GO:0016567	Protein ubiquitination	General	4.56	6.40e−03
GO:0006342	Chromatin silencing	Specific	9.17	2.02e−05
GO:0045947	Negative regulation of translational initiation	Specific	7.01	4.12e−03
GO:0043335	Protein unfolding	Specific	9.16	4.26e−03
GO:0060213	Positive regulation of nuclear-transcribed miRNA poly(A) tail shortening	Specific	7.49	4.26e−03
GO:0035278	miRNA mediated inhibition of translation	Specific	7.10	5.63e−03
GO:0070922	Small RNA loading onto RISC	Specific	6.75	5.87e−03
GO:0030422	Production of siRNA involved in RNA interference	Specific	6.16	7.51e−03
GO:0006379	mRNA cleavage	Specific	6.16	7.51e−03
GO:0031054	Pre-miRNA processing	Specific	6.16	7.51e−03
CELLULAR COMPONENT (CC)				
GO:0005844	Polysome	General	5.08	8.05e−03
GO:0031332	RNAi effector complex	Specific	7.27	2.93e−03
GO:0035068	Microribonucleoprotein complex	Specific	6.89	2.93e−03
GO:0070578	RISC-loading complex	Specific	6.89	2.93e−03
GO:0005845	mRNA cap binding complex	Specific	6.55	3.23e−03
MOLECULAR FUNCTION (MF)				
GO:0004839	Ubiquitin activating enzyme activity	Specific	11.95	2.87e−05
GO:0070551	Endoribonuclease activity, cleaving siRNA-paired mRNA	Specific	9.62	2.06e−04
GO:0016778	Diphosphotransferase activity	Specific	8.85	3.14e−04
GO:0000340	RNA 7-methylguanosine cap binding	Specific	7.69	8.12e−04

Abbreviations: FDR, false discovery rate; RISC, RNA-induced silencing complex.

RISC-loading complex (GO:0070578, FDR = 2.93e−03) and mRNA cap-binding complex (GO:0005845, FDR = 3.23e−03) showing moderate enrichments. In addition to BP and CC terms, enrichment for specific MF terms such as endoribonuclease activity, cleaving siRNA-paired mRNA (GO:0070551, FDR = 2.06e−04), diphosphotransferase activity (GO:0016778, 3.14e−04) and RNA 7-methylguanosine cap binding (GO:0000340, FDR = 8.12e−04) was found, with strongest enrichment observed for MF involving ubiquitin-activating enzyme activity (GO:0004839, FDR = 2.87e−05).

For domains that showed significant decrease in abundance value in grasses, GO enrichment for 6 BP and 2 MF terms were observed (Table 13). Among the BP terms, there was moderate enrichments for the specific process, acetyl-CoA metabolic process (GO:0006084, FDR = 2.61e−03) and 2 highly specific processes, namely cellular response to azide (GO:0097185, FDR = 5.64e−03) and cellular response to copper ion starvation (GO:0035874, FDR = 5.64e−03).

Finally, the domain-centric GO-enrichment analyses of domains that have significant different versatility values in

Table 13. Enriched GO terms from protein domains that show significant decrease in abundance values in grasses as compared to grasses outgroups.

GO ID	GO TERM	CATEGORY	*Z* SCORE	FDR
BIOLOGICAL PROCESS (BP)				
GO:0009056	Catabolic process	Highly general	5.19	5.80e−04
GO:0006810	Transport	Highly general	4.11	5.64e−03
GO:0006790	Sulfur compound metabolic process	General	5.31	1.92e−03
GO:0006084	Acetyl-CoA metabolic process	Specific	6.50	2.61e−03
GO:0097185	Cellular response to azide	Highly specific	8.09	5.64e−03
GO:0035874	Cellular response to copper ion starvation	Highly specific	8.09	5.64e−03
MOLECULAR FUNCTION (MF)				
GO:0043167	Ion binding	Highly general	4.90	3.19e−04
GO:0004478	Methionine adenosyltransferase activity	Specific	7.83	4.25e−03

Abbreviation: FDR, false discovery rate.

legumes and grasses concerning their outgroup species did not show enrichment of GO terms from any of the 3 subontologies.

Discussion

In this study, we describe evolutionary patterns in species from 2 large plant families: legumes and grasses, by tracking changes in their species-level protein-domain characteristics relative to selected outgroup species. We analyzed 4 types of domain characteristics to study gain and loss of domains, changes in duplication counts of domains along the sequences, expansion and contraction of domains, and changes in the partnering tendency of domains.

The work presents a generic framework for studying evolution of a chosen set of target species using protein domains as a unit of evolution instead of entire protein sequences. The feature-selection techniques used in data science and machine learning like the MI and statistical tests like Fisher's exact test and Wilcoxon rank-sum test can be used to select or filter-out significantly evolving domains in the target set of species relative to an outgroup set of species, which can be mapped to gain/loss or increase/decrease of particular biological functions in the target species. We have also containerized this entire analysis workflow inside a docker container which can be downloaded from the following URL: https://cloud.docker.com/u/akshayayadav/repository/docker/akshayayadav/protein-domain-evolution-project. The container is designed to accept user-defined set of target and outgroup proteomes along with the Pfam domain database and output domain sets for all 4 feature categories that have significantly different domain feature values (FDR $\leqslant$ 0.05) in target species as compared to the outgroup species.

It should be noted that the FDR-adjusted *P* values assigned to the domains by the statistical tests could be underestimated due to the statistical dependence between species in the target and outgroup set. In other words, even though the species are evolving independently, they are not statistically independent units, which could result in higher Type I error while testing the significance of the difference in values for domains, between the target species and outgroup species. Therefore, we recommend using the MI score, instead of the FDR-adjusted *P* values, as the primary indicator for detecting differential evolution of domains between the target and outgroup set of species.

Domain content analysis in legumes shows a striking loss of protein domains from FA pathway, the pathway which is responsible for the repair of interstrand DNA crosslinks. The FA pathway consists of a core complex that ubiquitinates the FANCD2-FANCI complex, which then localizes to the site of DNA repair. It seems that all the proteins from FA core complex and the FANCI protein, except the FANCD2 nuclease, are lost in the legumes. Although one of the repair proteins (FANCD2) is present in legumes, the core complex protein (FANCL) that monoubiquitinates the FANCD2, is absent. As ubiquitination of the FANCD2[60] is an indispensable part of the DNA repair process, this could mean that legumes might have lost the ability to repair interstrand DNA crosslinks or that the FA-mediated repair of interstrand DNA crosslinks is carried out without the ubiquitination of FANCD2. In grasses, domains showing gains include those involved in flavonoid biosynthesis (well-studied in maize), as well as structural proteins found in gluten and male florets. The domains that were detected as lost in grasses are involved in functions such as peptidoglycan biosynthesis, wound repair in sieve tubes, and fatty acid synthesis. Fatty acid synthesis may be reduced in the sampled monocots, due to relatively greater production of carbohydrates in grass seeds, and the differences in sieve tube structure in monocots as compared to dicots.[104]

Analyses of duplication feature matrices revealed a single domain of unknown function to have significantly decreased in copy number in legumes sequences. In grasses, an increase in

copy number of domains such *zf-PARP* and *FANCF* shows the evolution of enhanced DNA repair mechanisms because both the domains are involved in the detection of DNA nicks and interstrand DNA crosslinks, respectively. On the contrary, domains with functions related to ER-Golgi transport, enzymatic transfer of prenyl groups, and termination of mitochondrial transcription were found to be decreased in copy numbers. A study on the role of plastidic protein BELAYA SMERT (BSM) of the mitochondrial transcription termination family in embryogenesis and postembryonic development in plant cells shows that proteins from this family are not essential for cell viability in monocotyledonous grasses[105] thus explaining the decreased copy number of the *mTERF* domain in grasses.

Domains with significantly increased abundance values in legumes were found to be associated with functions involving Thylakoid formation, Glutathione metabolism, and enriched with GO terms related to biosynthetic/metabolic processes involving glycosyl compounds, ribonucleosides, and isoprenoids. For domains that showed significant decrease in abundance values in legumes, GO terms related to specific BPs and MFs involving oxidation of purine nucleobase xanthine were found to be significantly enriched. A study on xanthine oxidizing enzymes isolated from leaves of legumes confirms that these oxidoreductases do not react with molecular oxygen and are essentially dehydrogenases.[106] The decrease in abundance of domains involved in purine catabolism may also be attributed to the availability of fixed nitrogen and remobilization of nitrogen from breaking down purine rings is no longer required.[107] In grasses, domains showing significant increase in abundance values revealed domains involved in functions related to gene silencing with GO terms such as chromatin silencing, regulation of translational initiation, protein unfolding, micro/si-RNA-mediated gene regulation, showing significant enrichment. The microRNA-related enrichments could be attributed to the regulation of floral organ genes in grasses such as rice and maize influencing various features of flower structure.[108] An increase in gene-silencing-related domains could also be attributed to polyploidy in grasses[109] or enhanced response to viral infection.[110] On the contrary, domains with significant decrease in abundance values, in grasses, showed involvement in functions such as cell adhesion, intracellular chloroplast movement, interfascicular fiber differentiation, DNA synthesis, and pectin metabolism with enrichment of GO terms such as acetyl-CoA metabolism and response to azide.

Finally, the increase in the versatility of the zinc-binding domain in legumes could be related to root nodule symbiosis and compound leaf morphology. Nitrogen fixation through root nodule symbiosis is one of the salient features in legumes and studies have shown the involvement of nodule-specific zinc-binding domain-containing proteins in symbiosis establishment and nodule function.[111] The zinc-finger domain-containing transcription factor has also been shown to be involved in trifoliate compound leaf morphology in *Medicago truncatula*.[112] In grasses, increased versatility of DNA-binding domain involved in ultraviolet (UV)-light-related signal transduction, and calmodulin-binding could be due to an increase in the number of proteins involved in abiotic stress tolerance.[113] The increase in the versatility of the *Jacalin* domain also suggests increased adaptation of grasses to stressful environments.[114]

This method can be effectively used to study characteristic biological functions/processes for a selected group of species by filtering out protein domains that seem to have differently evolved in the group, with respect to an outgroup set of species. By closely studying the most significantly evolved protein domains and GO terms associated with significantly evolved protein domains, we might be able to explain the molecular mechanisms responsible for characteristic biological features observed in our target group of species.

Author Contributions

AY concieved and designed the research and analysis, and drafted the manuscript. SC and DFB supervised the analysis. All authors read, edited, and approved the manuscript.

ORCID iD

Akshay Yadav https://orcid.org/0000-0002-7313-1234

REFERENCES

1. Liu J, Rost B. CHOP: parsing proteins into structural domains. *Nucleic Acids Res*. 2004;32:W569-W571. doi:10.1093/nar/gkh481.
2. Bornberg-Bauer E, Beaussart F, Kummerfeld SK, Teichmann SA, Weiner J. The evolution of domain arrangements in proteins and interaction networks. *CMLS*. 2005;62:435-445. doi:10.1007/s00018-004-4416-1.
3. Vogel C, Teichmann SA, Pereira-Leal J. The relationship between domain duplication and recombination. *J Molec Biol*. 2005;346:355-365. doi:10.1016/j.jmb.2004.11.050.
4. Das S, Smith TF. Identifying nature's protein Lego set. *Adv Protein Chem*. 2000;54:159-184.
5. Chothia C, Gough J, Vogel C, Teichmann SA. Evolution of the protein repertoire. *Science*. 2003;300:1701-1703. doi:10.1126/science.1085371.
6. Teichmann SA, Park J, Chothia C. Structural assignments to the mycoplasma genitalium proteins show extensive gene duplications and domain rearrangements. *PNAS*. 1998;95:14658-14663.
7. Koonin EV, Aravind L, Kondrashov AS. The impact of comparative genomics on our understanding of evolution. *Cell*. 2000;101:573-576. doi:10.1016/S0092-8674(00)80867-3.
8. Lin J, Gerstein M. Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels. *Genome Res*. 2000;10:808-818. doi:10.1101/gr.10.6.808.
9. Caetano-Anollés G, Caetano-Anollés D. An evolutionarily structured universe of protein architecture. *Genome Res*. 2003;13:1563-1571. doi:10.1101/gr.1161903.
10. Yang S, Doolittle RF, Bourne PE. Phylogeny determined by protein domain content. *PNAS*. 2005;102:373-378. doi:10.1073/pnas.0408810102.
11. Nasir A, Kim KM, Caetano-Anollés G. Global patterns of protein domain gain and loss in superkingdoms. *PLOS Comput Biol*. 2014;10:e1003452. doi:10.1371/journal.pcbi.1003452.
12. Buljan M, Frankish A, Bateman A. Quantifying the mechanisms of domain gain in animal proteins. *Genome Biol*. 2010;11:R74. doi:10.1186/gb-2010-11-7-r74.
13. Björklund ÅK, Ekman D, Elofsson A. Expansion of protein domain repeats. *PLOS Comput Biol*. 2006;2:e114. doi:10.1371/journal.pcbi.0020114.
14. Yasutake Y, Watanabe S, Yao M, Takada Y, Fukunaga N, Tanaka I. Structure of the monomeric isocitrate dehydrogenase: evidence of a protein monomerization by a domain duplication. *Structure*. 2002;10:1637-1648. doi:10.1016/S0969-2126(02)00904-8.
15. Vogel C, Chothia C. Protein family expansions and biological complexity. *PLOS Comput Biol*. 2006;2:e48. doi:10.1371/journal.pcbi.0020048.
16. Basu MK, Poliakov E, Rogozin IB. Domain mobility in proteins: functional and evolutionary implications. *Brief Bioinform*. 2009;10:205-216. doi:10.1093/bib/bbn057.

17. Forslund K, Sonnhammer ELL. Evolution of protein domain architectures. In: Anisimova M, ed. Evolutionary Genomics: Statistical and Computational Methods, Volume 2. Methods in Molecular Biology. Totowa, NJ: Humana Press; 2012:187-216. doi:10.1007/978-1-61779-585-5_8.
18. Kraskov A, Stögbauer H, Grassberger P. Estimating mutual information. *Phys Rev E*. 2004;69:066138. doi:10.1103/PhysRevE.69.066138.
19. Amiri F, Rezaei Yousefi M, Lucas C, Shakery A, Yazdani N. Mutual information-based feature selection for intrusion detection systems. *J Netw Comput Appl*. 2011;34:1184-1199. doi:10.1016/j.jnca.2011.01.002.
20. Kraskov A, Stögbauer H, Andrzejak RG, Grassberger P. Hierarchical clustering based on mutual information. http://arxiv.org/abs/q-bio/0311039 Updated 2003. Accessed October 2, 2019.
21. Beraha M, Metelli AM, Papini M, Tirinzoni A, Restelli M. Feature selection via mutual information: new theoretical insights. http://arxiv.org/abs/1907.07384 Updated 2019. Accessed October 2, 2019.
22. Fisher RA. On the interpretation of χ 2 from contingency tables, and the calculation of P. *J R Stat Soc*. 1922;85:87-94. doi:10.2307/2340521.
23. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. *Ann Math Statist*. 1947;18:50-60. doi:10.1214/aoms/1177730491.
24. Bertioli DJ, Cannon SB, Froenicke L, *et al*. The genome sequences of Arachis duranensis and Arachis ipaensis, the diploid ancestors of cultivated peanut. *Nature Genet*. 2015;47:438.
25. Varshney RK, Chen W, Li Y, *et al*. Draft genome sequence of pigeonpea (Cajanus cajan), an orphan legume crop of resource-poor farmers. *Nature Biotechnol*. 2012;30:83.
26. Varshney RK, Song C, Saxena RK, *et al*. Draft genome sequence of chickpea (Cicer arietinum) provides a resource for trait improvement. *Nature Biotechnol*. 2013;31:240.
27. Schmutz J, Cannon SB, Schlueter J, *et al*. Genome sequence of the palaeopolyploid soybean. *Nature*. 2010;463:178.
28. Sato S, Nakamura Y, Kaneko T, *et al*. Genome structure of the legume, Lotus japonicus. *DNA Res*. 2008;15:227-239.
29. Hane JK, Ming Y, Kamphuis LG, *et al*. A comprehensive draft genome sequence for lupin (Lupinus angustifolius), an emerging health food: insights into plant–microbe interactions and legume evolution. *Plant Biotechnol J*. 2017;15:318-330. doi:10.1111/pbi.12615.
30. Tang H, Krishnakumar V, Bidwell S, *et al*. An improved genome release (version Mt4.0) for the model legume Medicago truncatula. *BMC Genomics*. 2014;15:312. doi:10.1186/1471-2164-15-312.
31. Schmutz J, McClean PE, Mamidi S, *et al*. A reference genome for common bean and genome-wide analysis of dual domestications. *Nature Genet*. 2014;46:707-713. doi:10.1038/ng.3008.
32. De Vega JJ, Ayling S, Hegarty M, *et al*. Red clover (Trifolium pratense L.) draft genome provides a platform for trait improvement. *Sci Rep*. 2015;5:17394. doi:10.1038/srep17394.
33. Kang YJ, Satyawan D, Shim S, *et al*. Draft genome sequence of adzuki bean, Vigna angularis. *Sci Rep*. 2015;5:8069. doi:10.1038/srep08069.
34. Kang YJ, Kim SK, Kim MY, *et al*. Genome sequence of mungbean and insights into evolution within Vigna species. *Nature Commun*. 2014;5:5443. doi:10.1038/ncomms6443.
35. Vigna unguiculata v1.1. (Cowpea). https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Vunguiculata_er. Accessed February 11, 2019.
36. Prunus persica v2.1. (Peach). https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Ppersica. Accessed February 11, 2019.
37. Jaillon O, Aury J-M, Noel B, *et al*. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*. 2007;449:463-467. doi:10.1038/nature06148.
38. Phytozome 12, Cucumis sativus v1.0. (Cucumber). https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Csativus. Accessed February 11, 2019.
39. Berardini TZ, Reiser L, Li D, *et al*. The Arabidopsis information resource: making and mining the "gold standard" annotated reference plant genome. *Genesis*. 2015;53:474-485.
40. Solanum lycopersicum iTAG2.4 (Tomato). https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Slycopersicum. Accessed February 11, 2019.
41. Paterson AH, Wendel JF, Gundlach H, *et al*. Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres. *Nature*. 2012;492:423-427. doi:10.1038/nature11798.
42. Ouyang S, Zhu W, Hamilton J, *et al*. The TIGR rice genome annotation resource: improvements and new features. *Nucleic Acids Res*. 2007;35:d883-d887. doi:10.1093/nar/gkl976.
43. Tuskan GA, Difazio S, Jansson S, *et al*. The genome of black cottonwood, Populus trichocarpa (Torr & Gray). *Science*. 2006;313:1596-1604. doi:10.1126/science.1128691.
44. Motamayor JC, Mockaitis K, Schmutz J, *et al*. The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color. *Genome Biol*. 2013;14:r53. doi:10.1186/gb-2013-14-6-r53.
45. Schnable PS, Ware D, Fulton RS, *et al*. The B73 maize genome: complexity, diversity, and dynamics. *Science*. 2009;326:1112-1115. doi:10.1126/science.1178534.
46. Bennetzen JL, Schmutz J, Wang H, *et al*. Reference genome sequence of the model plant Setaria. *Nat Biotechnol*. 2012;30:555-561. doi:10.1038/nbt.2196.
47. Setaria viridis v2.1—Phytozome v12.1: info. https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Sviridis_er. Accessed October 8, 2019.
48. Panicum virgatum v5.1—Phytozome v12.1: info. https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Pvirgatum_er. Accessed October 8, 2019.
49. McCormick RF, Truong SK, Sreedasyam A, *et al*. The Sorghum bicolor reference genome: improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. *Plant J*. 2018;93:338-354. doi:10.1111/tpj.13781.
50. VanBuren R, Bryant D, Edger PP, *et al*. Single-molecule sequencing of the desiccation-tolerant grass Oropetium thomaeum. *Nature*. 2015;527:508-511. doi:10.1038/nature15714.
51. International Brachypodium Initiative. Genome sequencing and analysis of the model grass Brachypodium distachyon. *Nature*. 2010;463:763-768. doi:10.1038/nature08747.
52. Brachypodium stacei v1.1—Phytozome v12.1: info. https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Bstacei. Accessed October 8, 2019.
53. Ming R, VanBuren R, Wai CM, *et al*. The pineapple genome and the evolution of CAM photosynthesis. *Nat Genet*. 2015;47:1435-1442. doi:10.1038/ng.3435.
54. Droc G, Larivière D, Guignon V, *et al*. The banana genome hub. *Database (Oxford)*. 2013;2013:bat035. doi:10.1093/database/bat035.
55. El-Gebali S, Mistry J, Bateman A, *et al*. The Pfam protein families database in 2019. *Nucleic Acids Res*. 2019;47:D427-D432. doi:10.1093/nar/gky995.
56. Mistry J, Bateman A, Finn RD. Predicting active site residue annotations in the Pfam database. *BMC Bioinformatics*. 2007;8:298. doi:10.1186/1471-2105-8-298.
57. Eddy SR. A new generation of homology search tools based on probabilistic inference. In: Genome Informatics 2009. London, England: Imperial College Press and distributed by World Scientific; 2009:205-211. doi:10.1142/9781848165632_0019.
58. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Royal Stat Soc B (Methodological)*. 1995;57:289-300. doi:10.1111/j.2517-6161.1995.tb02031.x.
59. Dunleavy EM, Pidoux AL, Monet M, *et al*. A NASP (N1/N2)-related protein, Sim3, binds CENP-A and is required for its deposition at fission yeast centromeres. *Mol Cell*. 2007;28:1029-1044. doi:10.1016/j.molcel.2007.10.010.
60. Moldovan G-L, D'Andrea AD. How the Fanconi Anemia pathway guards the genome. *Annu Rev Genet*. 2009;43:223-249. doi:10.1146/annurev-genet-102108-134222.
61. Joo W, Xu G, Persky NS, *et al*. Structure of the FANCI-FANCD2 complex: insights into the Fanconi anemia DNA repair pathway. *Science*. 2011;333:312-316. doi:10.1126/science.1205805.
62. Nookala RK, Hussain S, Pellegrini L. Insights into Fanconi Anaemia from the structure of human FANCE. *Nucleic Acids Res*. 2007;35:1638-1648. doi:10.1093/nar/gkm033.
63. Gurtan AM, Stuckert P, D'Andrea AD. The WD40 repeats of FANCL are required for Fanconi anemia core complex assembly. *J Biol Chem*. 2006;281:10896-10905. doi:10.1074/jbc.M511411200.
64. MacKay C, Déclais A-C, Lundin C, *et al*. Identification of KIAA1018/FAN1, a DNA repair nuclease recruited to DNA damage by monoubiquitinated FANCD2. *Cell*. 2010;142:65-76. doi:10.1016/j.cell.2010.06.021.
65. Fessing MY, Krynetski EY, Zambetti GP, Evans WE. Functional characterization of the human thiopurine S-methyltransferase (TPMT) gene promoter. *Eur J Biochem*. 1998;256:510-517. doi:10.1046/j.1432-1327.1998.2560510.x.
66. Chopra S, Athma P, Peterson T. Alleles of the maize P gene with distinct tissue specificities encode Myb-homologous proteins with C-terminal replacements. *Plant Cell*. 1996;8:1149-1158. doi:10.1105/tpc.8.7.1149.
67. Grotewold E, Drummond BJ, Bowen B, Peterson T. The myb-homologous P gene controls phlobaphene pigmentation in maize floral organs by directly activating a flavonoid biosynthetic gene subset. *Cell*. 1994;76:543-553. doi:10.1016/0092-8674(94)90117-1.
68. Tatham AS, Shewry PR. Elastomeric proteins: biological roles, structures and mechanisms. *Trends Biochem Sci*. 2000;25:567-571. doi:10.1016/s0968-0004(00)01670-4.
69. Wright SY, Suner MM, Bell PJ, Vaudin M, Greenland AJ. Isolation and characterization of male flower cDNAs from maize. *Plant J*. 1993;3:41-49. doi:10.1046/j.1365-313x.1993.t01-2-00999.x.
70. Bertrand JA, Auger G, Fanchon E, *et al*. Crystal structure of UDP-N-acetylmuramoyl-L-alanine: D-glutamate ligase from Escherichia coli. *EMBO J*. 1997;16:3416-3425. doi:10.1093/emboj/16.12.3416.
71. Rüping B, Ernst AM, Jekat SB, *et al*. Molecular and phylogenetic characterization of the sieve element occlusion gene family in Fabaceae and non-Fabaceae plants. *BMC Plant Biology*. 2010;10:219. doi:10.1186/1471-2229-10-219.

72. Munday MR, Hemingway CJ. The regulation of acetyl-CoA carboxylase—a potential target for the action of hypolipidemic agents. *Adv Enzyme Regul*. 1999;39:205-234.
73. Battini JL, Rasko JE, Miller AD. A human cell-surface receptor for xenotropic and polytropic murine leukemia viruses: possible role in G protein-coupled signal transduction. *Proc Natl Acad Sci USA*. 1999;96:1385-1390. doi:10.1073/pnas.96.4.1385.
74. Spain BH, Koo D, Ramakrishnan M, Dzudzor B, Colicelli J. Truncated forms of a novel yeast protein suppress the lethality of a G protein alpha subunit deficiency by interacting with the beta subunit. *J Biol Chem*. 1995;270:25435-25444. doi:10.1074/jbc.270.43.25435.
75. Lenburg ME, O'Shea EK. Signaling phosphate starvation. *Trends Biochem Sci*. 1996;21:383-387.
76. de Murcia G, Ménissier de Murcia J. Poly(ADP-ribose) polymerase: a molecular nick-sensor. *Trends Biochem Sci*. 1994;19:172-176. doi:10.1016/0968-0004(94)90280-1.
77. Wang Z, Li M, Lu S, Zhang Y, Wang H. Promoter hypermethylation of FANCF plays an important role in the occurrence of ovarian cancer through disrupting Fanconi anemia-BRCA pathway. *Cancer Biol Ther*. 2006;5:256-260. doi:10.4161/cbt.5.3.2380.
78. Mnaimneh S, Davierwala AP, Haynes J, *et al*. Exploration of essential gene functions via titratable promoter alleles. *Cell*. 2004;118:31-44. doi:10.1016/j.cell.2004.06.013.
79. Kraynack BA, Chan A, Rosenthal E, *et al*. Dsl1p, Tip20p, and the novel Dsl3(Sec39) protein are required for the stability of the Q/t-SNARE complex at the endoplasmic reticulum in yeast. *Mol Biol Cell*. 2005;16:3963-3977. doi:10.1091/mbc.e05-01-0056.
80. Poralla K, Hewelt A, Prestwich GD, Abe I, Reipen I, Sprenger G. A specific amino acid repeat in squalene and oxidosqualene cyclases. *Trends Biochem Sci*. 1994;19:157-158. doi:10.1016/0968-0004(94)90276-3.
81. Wendt KU, Poralla K, Schulz GE. Structure and function of a squalene cyclase. *Science*. 1997;277:1811-1815. doi:10.1126/science.277.5333.1811.
82. Harnpicharnchai P, Jakovljevic J, Horsey E, *et al*. Composition and functional characterization of yeast 66S ribosome assembly intermediates. *Mol Cell*. 2001;8:505-515. doi:10.1016/s1097-2765(01)00344-6.
83. Fernandez-Silva P, Martinez-Azorin F, Micol V, Attardi G. The human mitochondrial transcription termination factor (mTERF) is a multizipper protein but binds to DNA as a monomer, with evidence pointing to intramolecular leucine zipper interactions. *EMBO J*. 1997;16:1066-1079. doi:10.1093/emboj/16.5.1066.
84. Huang J, Taylor JP, Chen J-G, *et al*. The plastid protein THYLAKOID FORMATION1 and the plasma membrane G-protein GPA1 interact in a novel sugar-signaling mechanism in Arabidopsis. *Plant Cell*. 2006;18:1226-1238. doi:10.1105/tpc.105.037259.
85. Armstrong RN. Structure, catalytic mechanism, and evolution of the glutathione transferases. *Chem Res Toxicol*. 1997;10:2-18. doi:10.1021/tx960072x.
86. Chishti AH, Kim AC, Marfatia SM, *et al*. The FERM domain: a unique module involved in the linkage of cytoplasmic proteins to the membrane. *Trends Biochem Sci*. 1998;23:281-282. doi:10.1016/s0968-0004(98)01237-7.
87. McIntire SL, Reimer RJ, Schuske K, Edwards RH, Jorgensen EM. Identification and characterization of the vesicular GABA transporter. *Nature*. 1997;389:870-876. doi:10.1038/39908.
88. Lee I, Schindelin H. Structural insights into E1-catalyzed ubiquitin activation and transfer to conjugating enzymes. *Cell*. 2008;134:268-278. doi:10.1016/j.cell.2008.05.046.
89. Johnston SC, Larsen CN, Cook WJ, Wilkinson KD, Hill CP. Crystal structure of a deubiquitinating enzyme (human UCH-L3) at 1.8 a resolution. *EMBO J*. 1997;16:3787-3796. doi:10.1093/emboj/16.13.3787.
90. Adams J, Kelso R, Cooley L. The kelch repeat superfamily of proteins: propellers of cell function. *Trends Cell Biol*. 2000;10:17-24.
91. Zhang D, Aravind L. Identification of novel families and classification of the C2 domain superfamily elucidate the origin and evolution of membrane targeting activities in eukaryotes. *Gene*. 2010;469:18-30. doi:10.1016/j.gene.2010.08.006.
92. Zhong R, Ye Z-H. IFL1, a gene regulating interfascicular fiber differentiation in arabidopsis, encodes a homeodomain-leucine zipper protein. *Plant Cell*. 1999;11:2139-2152. doi:10.1105/tpc.11.11.2139.
93. Weimbs T, Low SH, Chapin SJ, Mostov KE, Bucher P, Hofmann K. A conserved domain is present in different families of vesicular fusion proteins: a new superfamily. *Proc Natl Acad Sci USA*. 1997;94:3046-3051. doi:10.1073/pnas.94.7.3046.
94. Yoder MD, Keen NT, Jurnak F. New domain motif: the structure of pectate lyase C, a secreted plant virulence factor. *Science*. 1993;260:1503-1507. doi:10.1126/science.8502994.
95. Wing RA, Yamaguchi J, Larabell SK, Ursin VM, McCormick S. Molecular and genetic characterization of two pollen-expressed genes that have sequence similarity to pectate lyases of the plant pathogen Erwinia. *Plant Mol Biol*. 1990;14:17-28. doi:10.1007/bf00015651.
96. Fries M, Ihrig J, Brocklehurst K, Shevchik VE, Pickersgill RW. Molecular basis of the activity of the phytopathogen pectin methylesterase. *EMBO J*. 2007;26:3879-3887. doi:10.1038/sj.emboj.7601816.
97. da Costa e Silva O. CG-1, a parsley light-induced DNA-binding protein. *Plant Mol Biol*. 1994;25:921-924. doi:10.1007/BF00028887.
98. Bouché N, Scharlat A, Snedden W, Bouchez D, Fromm H. A novel family of calmodulin-binding transcription activators in multicellular organisms. *J Biol Chem*. 2002;277:21851-21861. doi:10.1074/jbc.M200268200.
99. Sankaranarayanan R, Sekar K, Banerjee R, Sharma V, Surolia A, Vijayan M. A novel mode of carbohydrate recognition in jacalin, a Moraceae plant lectin with a beta-prism fold. *Nat Struct Biol*. 1996;3:596-603.
100. Iyer LM, Anantharaman V, Aravind L. The DOMON domains are involved in heme and sugar recognition. *Bioinformatics*. 2007;23:2660-2664. doi:10.1093/bioinformatics/btm411.
101. Yao T, Song L, Xu W, *et al*. Proteasome recruitment and activation of the Uch37 deubiquitinating enzyme by Adrm1. *Nat Cell Biol*. 2006;8:994-1002. doi:10.1038/ncb1460.
102. Chen X, Lee B-H, Finley D, Walters KJ. Structure of proteasome ubiquitin receptor hRpn13 and its activation by the scaffolding protein hRpn2. *Mol Cell*. 2010;38:404-415. doi:10.1016/j.molcel.2010.04.019.
103. Fang H, Gough J. dcGO: database of domain-centric ontologies on functions, phenotypes, diseases and more. *Nucleic Acids Res*. 2013;41:D536-D544. doi:10.1093/nar/gks1080.
104. Botha T. A tale of two neglected systems—structure and function of the thin- and thick-walled sieve tubes in monocotyledonous leaves. *Front Plant Sci*. 2013;4:297. doi:10.3389/fpls.2013.00297.
105. Babiychuk E, Vandepoele K, Wissing J, *et al*. Plastid gene expression and plant development require a plastidic protein of the mitochondrial transcription termination factor family. *PNAS*. 2011;108:6674-6679. doi:10.1073/pnas.1103442108.
106. Montalbini P. Xanthine dehydrogenase from leaves of leguminous plants: purification, characterization and properties of the enzyme. *J Plant Physiol*. 2000;156:3-16. doi:10.1016/S0176-1617(00)80266-7.
107. Werner AK, Witte C-P. The biochemistry of nitrogen mobilization: purine ring catabolism. *Trends Plant Sci*. 2011;16:381-387. doi:10.1016/j.tplants.2011.03.012.
108. Smoczynska A, Szweykowska-Kulinska Z. MicroRNA-mediated regulation of flower development in grasses. *Acta Biochimica Polonica*. 2016;63:687-692. doi:10.18388/abp.2016_1358.
109. Levy AA, Feldman M. The impact of polyploidy on grass genome evolution. *Plant Physiol*. 2002;130:1587-1593. doi:10.1104/pp.015727.
110. Ratcliff F, Harrison BD, Baulcombe DC. A similarity between viral defense and gene silencing in plants. *Science*. 1997;276:1558-1560. doi:10.1126/science.276.5318.1558.
111. Yuan S, Li X, Li R, *et al*. Genome-wide identification and classification of soybean C2H2 zinc finger proteins and their expression analysis in legume-rhizobium symbiosis. *Front Microbiol*. 2018;9:126. doi:10.3389/fmicb.2018.00126.
112. Chen J, Yu J, Ge L, *et al*. Control of dissected leaf morphology by a Cys(2) His(2) zinc finger transcription factor in the model legume Medicago truncatula. *PNAS*. 2010;107:10754-10759. doi:10.1073/pnas.1003954107.
113. Mizoi J, Shinozaki K, Yamaguchi-Shinozaki K. AP2/ERF family transcription factors in plant abiotic stress responses. *Biochim Biophys Acta—Gene Regulat Mech*. 2012;1819:86-96. doi:10.1016/j.bbagrm.2011.08.004.
114. Song M, Xu W, Xiang Y, Jia H, Zhang L, Ma Z. Association of jacalin-related lectins with wheat responses to stresses revealed by transcriptional profiling. *Plant Mol Biol*. 2014;84:95-110. doi:10.1007/s11103-013-0121-5.