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

Akshay Yadav1, David Fernández-Baca2 and Steven B Cannon3

1Bioinformatics and Computational Biology Graduate Program, Iowa State University, Ames, IA, USA. 2Department of Computer Science, Iowa State University, Ames, IA, USA. 3Corn 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/akshayyadav/repository/docker/akshayyadav/protein-domain-evolution-project.

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

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

Material and Methods
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. MI has been routinely used for selecting meaningful features, in classification and pattern recognition problems. 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 for feature matrices containing discrete values, and Wilcoxon rank-sum tests for feature matrices containing continuous values.

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.
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    |          |          |        |                    |            |
| Cajanus cajan           | cajca   | Legume    | ICPL87119| 1        | 1      | Varshney et al[25] | LegumelInf |
| Cicer arietinum         | cicar   | Legume    | Frontier | 1        | 1      | Varshney et al[26] | LegumelInf |
| Glycine max             | glyma   | Legume    | Williams 82| 2        | 1      | Schmutz et al[27]  | Phytomez   |
| Lotus japonicus         | lotja   | Legume    | MG20     | 3        | 1      | Sato et al[28]     | Phytozome  |
| Lupinus angustifolius   | lupan   | Legume    |          |          |        |                    |            |
| Medicago truncatula     | medtr   | Legume    | A17_HM341| 4        | 2      | Tang et al[29]     | Phytomez   |
| Phaseolus vulgaris      | phavu   | Legume    | G19833   | 2        | 1      | Schmutz et al[30]  | Phytomez   |
| Trifolium pretense      | tripr   | Legume    |          |          |        |                    |            |
| Vigna angularis         | vigan   | Legume    | Va3.0    | 1        | 3      | Kang et al[31]     | LegumelInf |
| Vigna radiate           | vigra   | Legume    | VC1973A  | 6        | 1      | Kang et al[32]     | LegumelInf |
| Vigna unguiculata       | vigin   | Legume    | IT97K    | 1        | 1      | Phytozome[33]      | Phytomez   |
| Prunus persica          | prupe   | Outgroup  | Lovell   | 2        | 2.1    | IPG[34]            | Phytomez   |
| Vitis vinifera          | vitvi   | Outgroup  | PN40024  | 12X      | 12X    | Jaillon et al[35]  | Phytomez   |
| Cucumis sativus         | cucsa   | Outgroup  |          | 1        | 1      | Phytozome[36]      | Phytomez   |
| Arabidopsis thaliana    | arath   | Outgroup  | Col-0    |          |        | TAIR10             | Berardinid |
| Solanum lycopersicum    | solly   | Outgroup  | LA1589   | ITAG2.4  | ITAG2.4 | Tomato Genome     |            |

(Continued)
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    | Benetzen 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\(^{46}\)  |                    |
| 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 | Puptri | Outgroup | Lovell   | 2        | 3.1    | Tuskan et al\(^{43}\)    | Phytozome          |
| Prunus persica   | Prupe   | Outgroup |          | 2        | 2.1    | IPGII\(^{36}\)           | Phytozome          |
| Glycine max      | Glyma   | Outgroup | Williams 82 | 2        | 1      | Schmutz et al\(^{27}\)  | Phytozome          |
| Vitis vinifera   | Viti    | 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      | Droct et al\(^{54}\)    | Banana Genome Hub  |

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

$$\text{IDF}(S, d) = \log_2 \frac{N(S)}{N(S, d)}$$  \hspace{1cm} (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

$$V(S, d) = \frac{1}{F(S, d)}$$  \hspace{1cm} (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. The duplication and versatility 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

$$I(X; Y) = \sum_{x, y} P(x, y) \log \frac{P(x, y)}{P(x)P(y)}$$  \hspace{1cm} (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 $2 \times 2$ 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. The FDR-adjusted $P$ values were reported for the domains.

**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
for delivering histone proteins such as H3 to centromeric chromatin. 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. 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. 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 and the C-terminal domain of FANCL protein (FA_NC 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). The only protein involved in the FA pathway that is present in legumes is the single domain FANCD2 nuclease containing the FancD2 domain. 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 in the catalysis of S-methylation of aromatic and heterocyclic sulfhydryl compounds, such as anticancer and immunosuppressive thiopurines.

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 photophenene pigment-producing arm of the flavonoid biosynthesis pathway. 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. 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.

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. The SEO_N and SEO_C domains are respectively found at the N and C terminus of sievel 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. The ACCA domain is the alpha isoform of the carbamyltransferase subunit of Acetyl CoA carboxylase enzyme. The ACCA domain is known to play an important role in the production of Malonyl-CoA in fatty acid synthesis.

**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. 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 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                 | –                     |
| SEO_N       | 0.7188   | .0011                 | –                     |
| DUF1719     | 0.7188   | .0011                 | +                     |
| DUF1110     | 0.7188   | .0011                 | +                     |

Abbreviations: MI, mutual information; FDR, false discovery rate.
This domain is known to act as a DNA nick sensor. 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.

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. The Prenyltrans domain-containing enzymes are responsible for the transfer of allylic prenyl groups to acceptor molecules. The Nop16 domain is part of a protein involved in ribosome biogenesis. 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.

**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. The C-terminal domain of Glutathione-S-transferase (GST_C_6) is known to conjugate reduced glutathione to auxin-regulated proteins in plants. The FERM_M domain is the middle domain of FERM protein and is involved in localizing proteins from cytosol to plasma membrane. The DAO_C domain is present at the C-terminal

### 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                 |                | +      |
| zl-PARP     | 0.4715    | .0445                 |                | +      |
| mTERF       | 0.4447    | .0356                 |                | -      |
| FANCF       | 0.4329    | .0359                 |                | +      |

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

### 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.

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, 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.

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. 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. 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. 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. The Pec_lyase_C domain is a part of the Pectolyase 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. The Pectinesterase domain is a cell-wall-associated enzyme (EC 3.1.1.11) involved in cell-wall modification and breakdown.

**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.

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 and calmodulin-binding transcriptional activators containing ankyrin motifs. The Jacalin domain is a mannos-binding lectin domain with a beta-prism fold. The zinc knuckle (zf-CCHC) domain is a zinc-binding motif composed of the CX2C4H4X4C motif (where X can be any amino acid).

Among the protein domains that have lost domain partners in grasses as compared to their 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 CxxCxs(12)Hxs(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.

The WRC domain is known for containing the conserved

| 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.
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.\textsuperscript{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.\textsuperscript{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 SHN1-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 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).

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 11) 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).

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:006379, 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).

| 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.*

Table 9. Enriched GO terms from protein domains that were detected as gained in grasses as compared to grass outgroups.
### 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:0019887 | 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:006144  | 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.
Yadav et al.

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, FDR = 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 enrichment 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

| GO ID         | GO TERM                                          | CATEGORY      | Z SCORE | FDR       |
|---------------|--------------------------------------------------|---------------|---------|-----------|
| 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                              | General       | 4.56    | 6.40e−03 |
| GO:0045947    | Negative regulation of translational initiation  | Specific      | 7.01    | 4.12e−03 |
| GO:0043335    | Protein unfolding                                | Specific      | 9.17    | 2.02e−05 |
| 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 |
| 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 |
| 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.
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 $\leq 0.05$) 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 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

| 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:0043187  | 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.
copy number of domains such as 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 crossovers, 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 grasses103 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 transcriptional initiation, protein unfolding, micro/si-RNA-mediated gene regulation, showing significant enrichment. The micro-RNA-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 grasses109 or enhanced response to viral infection.110 In grasses, domains showing 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.117 The zinc-finger domain-containing transcription factor has also been shown to be involved in triloiate 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 conceived and designed the research and analysis, and drafted the manuscript. SC and DBF supervised the analysis. All authors read, edited, and approved the manuscript.

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

REFERENCES
1. Lin J, Rose B. CHOP: parsing proteins into structural domains. Nucleic Acids Res. 2004;32:W569-W571. doi:10.1093/nar/gkh481.
2. Bornberg-Bauer E, Beauasset F, Kummerfeld SK, Teichmann SA, Weiner J. The evolution of domain arrangements in proteins and interaction networks. CME. 2005;6:435-445. doi:10.1007/s00018-004-4416-1.
3. Vogel C, Teichmann SA, Perest-Leaf J. The relationship between domain duplication and recombination. J Proteol. 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;95:14683-14663.
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)08673-1.
8. Lin J, Gerstein M. Whole-genome trees based on the occurrence of folds and orthologous 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.116903.
10. Yang S, Doolittle RF, Bourne PE. Phylogeny determined by protein domain content. PNAS. 2007;104:373-378. 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 and loss in animal proteins. Genome Biol. 2010;11:R74. doi:10.1186/gb-2010-11-7-74.
13. Björklund ÅK, Ekman D, Elofsson A. Expansion of protein domain repeats. Adv Protein Chem. 2005;62:435-445. doi:10.1016/j.apsb.2004.11.050.
14. Yasutake Y, Watanabe S, Yao M, Takada Y, Fukunaga N, Tanaka I. Structure of the monomeric isocitrate dehydrogenase: evidence of a protein monomerizability 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. Forlund K, Sonnhammar E.L. 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. Kraikov A, Stoibauer H, Grasserberger P. Estimating mutual information. Phys. Rev. E. 2009;60:066138. doi:10.1103/PhysRevE.69.066138.

19. Amiri F, Rezaei Yousef 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. Kraikov A, Stoibauer H, Andreszak RG, Grasserberger P. Hierarchical clustering based on mutual information. http://arxiv.org/abs/1003.0130v3 Updated 2003. Accessed October 2, 2019.

21. Beraha M, Metelli AM, Papini M, Tiriinzoni A, Restelli M. Feature selection via mutual information: new theoretical insights. http://arxiv.org/abs/1907.03843 Updated 2019. Accessed October 2, 2019.

22. Fisher RA. On the interpretation of g 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 into evolution within Vigna species. Nat Genet. 2011;43:67-70. doi:10.1038/nature09874.

24. Bhattacharyya D, Chakraborty A, Das B, et al. A NASP (N1/N2)-related protein, A3 from contingency tables, and the calculation of P. J R Stat Soc. 1922;85:87-94. doi:10.2307/2340521.

25. Mann HB, Whitney DR. On a test of whether one of two random variables is into evolution within Vigna species. Nat Genet. 2011;43:67-70. doi:10.1038/nature09874.
Chishti AH, Kim AC, Marfatia SM, McIntire SL, Reimer RJ, Schuske K, Edwards RH, Jorgensen EM. Identification and characterization of the vesicular GABA transporter. *Science*. 1998;23:281-282. doi:10.1016/s0968-0004(98)01237-7.

Trends Biochem Sci. 1996;21:383-387.

doi:10.1016/s0968-0004(96)80004-9.

Moaimneh S, Davierwala AP, Haynes J, Wang Z, Li M, Lu S, Zhang Y, Wang H. Promoter hypermethylation of one transferases. *Biochim Biophys Acta*. 1997;10:2-18. doi:10.1021/tx960072x.

Mnaimneh S, Davierwala AP, Haynes J, Wang Z, Li M, Lu S, Zhang Y, Wang H. Promoter hypermethylation of one transferases. *Biochim Biophys Acta*. 1997;10:2-18. doi:10.1021/tx960072x.

Fries M, Ihirj J, Brocklehurst K, Shevchik VE, Pickersgill RW. Molecular basis of the activity of the phytotypathogen pectin methyltransferase. *EMBO J*. 2007;26:3879-3887. doi:10.1093/emboj/dem160.

do:10.1093/mbio/0015651.

da Costa e Silva O. CG-1, a paraldehyde light-induced DNA-binding protein. *Plant Mol Biol*. 1994;25:291-294. doi:10.1007/BF00288887.

Bouché N, Scharlat A, Snedden W, Bouché D, Fromm H. A novel family of calmodulin-binding transcription activators in multicellular organisms. *J Biol Chem*. 2002;277:21831-21836. doi:10.1074/jbc.M200268200.

Sankaranarayanan R, Sekar K, Banerjee R, Sharma V, Sarbha A, Vijayan M. A novel model of carbohydrate recognition in jacinula, a Moraceae plant lectin with a beta-p3 fold. *Nat Struct Biol*. 1996;3:596-603.

Iyer LM, Anantharaman V, Aravind L. The DOMON domains are involved in heme and sugar recognition. *Bioinformatics*. 2005;21:2660-2664. doi:10.1093/bioinformatics/bti360.

Utigur S, Zhou Y, Xie W, et al. Proteasome recruitment and activation of the Uch7 deubiquitinating enzyme by Adml. *Nat Cell Biol*. 2006;8:994-1002. doi:10.1038/ncl364.

Yao T, Song L, Xu W, et al. Proteasome recruitment and activation of the Uch7 deubiquitinating enzyme by Adml. *Nat Cell Biol*. 2006;8:994-1002. doi:10.1038/ncl364.

Bouché N, Scharlat A, Snedden W, Bouché D, Fromm H. A novel family of calmodulin-binding transcription activators in multicellular organisms. *J Biol Chem*. 2002;277:21831-21836. doi:10.1074/jbc.M200268200.

Sankaranarayanan R, Sekar K, Banerjee R, Sharma V, Sarbha A, Vijayan M. A novel model of carbohydrate recognition in jacinula, a Moraceae plant lectin with a beta-p3 fold. *Nat Struct Biol*. 1996;3:596-603.

Iyer LM, Anantharaman V, Aravind L. The DOMON domains are involved in heme and sugar recognition. *Bioinformatics*. 2005;21:2660-2664. doi:10.1093/bioinformatics/bti360.

Yao T, Song L, Xu W, et al. Proteasome recruitment and activation of the Uch7 deubiquitinating enzyme by Adml. *Nat Cell Biol*. 2006;8:994-1002. doi:10.1038/ncl364.

Yao T, Song L, Xu W, et al. Proteasome recruitment and activation of the Uch7 deubiquitinating enzyme by Adml. *Nat Cell Biol*. 2006;8:994-1002. doi:10.1038/ncl364.

Bouché N, Scharlat A, Snedden W, Bouché D, Fromm H. A novel family of calmodulin-binding transcription activators in multicellular organisms. *J Biol Chem*. 2002;277:21831-21836. doi:10.1074/jbc.M200268200.

Sankaranarayanan R, Sekar K, Banerjee R, Sharma V, Sarbha A, Vijayan M. A novel model of carbohydrate recognition in jacinula, a Moraceae plant lectin with a beta-p3 fold. *Nat Struct Biol*. 1996;3:596-603.

Iyer LM, Anantharaman V, Aravind L. The DOMON domains are involved in heme and sugar recognition. *Bioinformatics*. 2005;21:2660-2664. doi:10.1093/bioinformatics/bti360.

Yao T, Song L, Xu W, et al. Proteasome recruitment and activation of the Uch7 deubiquitinating enzyme by Adml. *Nat Cell Biol*. 2006;8:994-1002. doi:10.1038/ncl364.

Bouché N, Scharlat A, Snedden W, Bouché D, Fromm H. A novel family of calmodulin-binding transcription activators in multicellular organisms. *J Biol Chem*. 2002;277:21831-21836. doi:10.1074/jbc.M200268200.

Sankaranarayanan R, Sekar K, Banerjee R, Sharma V, Sarbha A, Vijayan M. A novel model of carbohydrate recognition in jacinula, a Moraceae plant lectin with a beta-p3 fold. *Nat Struct Biol*. 1996;3:596-603.

Iyer LM, Anantharaman V, Aravind L. The DOMON domains are involved in heme and sugar recognition. *Bioinformatics*. 2005;21:2660-2664. doi:10.1093/bioinformatics/bti360.

Yao T, Song L, Xu W, et al. Proteasome recruitment and activation of the Uch7 deubiquitinating enzyme by Adml. *Nat Cell Biol*. 2006;8:994-1002. doi:10.1038/ncl364.

Bouché N, Scharlat A, Snedden W, Bouché D, Fromm H. A novel family of calmodulin-binding transcription activators in multicellular organisms. *J Biol Chem*. 2002;277:21831-21836. doi:10.1074/jbc.M200268200.

Sankaranarayanan R, Sekar K, Banerjee R, Sharma V, Sarbha A, Vijayan M. A novel model of carbohydrate recognition in jacinula, a Moraceae plant lectin with a beta-p3 fold. *Nat Struct Biol*. 1996;3:596-603.

Iyer LM, Anantharaman V, Aravind L. The DOMON domains are involved in heme and sugar recognition. *Bioinformatics*. 2005;21:2660-2664. doi:10.1093/bioinformatics/bti360.

Yao T, Song L, Xu W, et al. Proteasome recruitment and activation of the Uch7 deubiquitinating enzyme by Adml. *Nat Cell Biol*. 2006;8:994-1002. doi:10.1038/ncl364.

Bouché N, Scharlat A, Snedden W, Bouché D, Fromm H. A novel family of calmodulin-binding transcription activators in multicellular organisms. *J Biol Chem*. 2002;277:21831-21836. doi:10.1074/jbc.M200268200.

Sankaranarayanan R, Sekar K, Banerjee R, Sharma V, Sarbha A, Vijayan M. A novel model of carbohydrate recognition in jacinula, a Moraceae plant lectin with a beta-p3 fold. *Nat Struct Biol*. 1996;3:596-603.

Iyer LM, Anantharaman V, Aravind L. The DOMON domains are involved in heme and sugar recognition. *Bioinformatics*. 2005;21:2660-2664. doi:10.1093/bioinformatics/bti360.