PCP
PLANT & CELL PHYSIOLOGY

Special Online Collection – Database Paper

Plant-PrAS: A Database of Physicochemical and Structural Properties and Novel Functional Regions in Plant Proteomes

Atsushi Kurotani[1,2], Yutaka Yamada[1], Kazuo Shinozaki[1], Yutaka Kuroda[2] and Tetsuya Sakurai[1,*]

[1]RIKEN Center for Sustainable Resource Science, Yokohama, Kanagawa, 230-0045 Japan
[2]Department of Biotechnology and Life Sciences, Faculty of Technology, Tokyo University of Agriculture and Technology, Koganei, Tokyo, 184-8588 Japan
*Corresponding author: E-mail, tetsuya.sakurai@riken.jp; Fax, +81-45-503-9665.
(Received August 28, 2014; Accepted October 31, 2014)

Arabidopsis thaliana **is an important model species for studies of plant gene functions. Research on Arabidopsis has resulted in the generation of high-quality genome sequences, annotations and related post-genomic studies. The amount of annotation, such as gene-coding regions and structures, is steadily growing in the field of plant research. In contrast to the genomics resource of animals and microorganisms, there are still some difficulties with characterization of some gene functions in plant genomics studies. The acquisition of information on protein structure can help elucidate the corresponding gene function because proteins encoded in the genome possess highly specific structures and functions. In this study, we calculated multiple physicochemical and secondary structural parameters of protein sequences, including length, hydrophobicity, the amount of secondary structure, the number of intrinsically disordered regions (IDRs) and the predicted presence of transmembrane helices and signal peptides, using a total of 208,333 protein sequences from the genomes of six representative plant species,** ***Arabidopsis thaliana***, ***Glycine max*** **(soybean),** ***Populus trichocarpa*** **(poplar),** ***Oryza sativa*** **(rice),** ***Physcomitrella patens*** **(moss) and** ***Cyanidioschyzon merolae*** **(alga). Using the PASS tool and the Rosetta Stone method, we annotated the presence of novel functional regions in 1,732 protein sequences that included unannotated sequences from the Arabidopsis and rice proteomes. These results were organized into the Plant Protein Annotation Suite database (Plant-PrAS), which can be freely accessed online at http://plant-pras.riken.jp/.**

Keywords: Database • Gene function • Physicochemical property • Plant protein • Protein property.

Abbreviations: IDR, intrinsically disordered region; GRAVY, grand average of hydrophobicity; MSU Rice, Michigan State University Rice Genome Annotation Project; Plant-PrAS, Plant Protein Annotation Suite database; RAP-DB, the Rice Annotation Project database; TAIR, The Arabidopsis Information Resource.

Introduction

The flowering plant Arabidopsis has a small genome and a short life cycle. Therefore, it is considered an important model plant. After the whole-genome sequence of Arabidopsis was published in 2000 (Arabidopsis Genome Initiative 2000), the information related to Arabidopsis research was organized into The Arabidopsis Information Resource (TAIR; http://arabidopsis.org/), comprising various types of data such as DNA and seed stocks, literature citations, gene functions and protein structures (Lamesch et al. 2012). Nevertheless, one-third of all the proteins of Arabidopsis still lack functional annotations in terms of biological roles (Kourmpetis et al. 2011, Li et al. 2012) in spite of the extensive experimental and computational studies undertaken by many researchers. Similarly, the whole-genome sequencing of rice, one of the most important model crop plants, was recently completed (International Rice Genome Sequencing Project 2005, Yu et al. 2002). Subsequently, all the functional annotations for proteins and non-coding RNAs (ncRNAs) were manually curated (Rice Annotation Project 2007). The genome and the functional gene annotations of rice have been updated in the Michigan State University Rice Genome Annotation Project database (MSU Rice; http://rice.plantbiology.msu.edu/) (Kawahara et al. 2013) and in the Rice Annotation Project database (RAP-DB; http://rapdb.dna.affrc.go.jp/) (Sakai et al. 2013). Those annotations, however, also include information on genes with insufficient experimental evidence. Thus, *Arabidopsis thaliana* and rice, two well-studied plant species, still harbor unannotated genes.

In order to improve functional annotation of genes in plants, various initiatives have been undertaken, such as inclusion of an experimental method that uses cross-species expressed sequence tag (EST) information (Chen et al. 2007), integration of plant genomic information (Asamizu et al. 2014), integration of Arabidopsis transcriptomic information (Obayashi et al. 2014), utilization of transcriptomic and metabolic profiles among plant tissues (Sakurai et al. 2013), integrative analysis of plant hormone accumulation and gene expression among rice tissues (Kudo et al. 2013), inclusion of the phenotypic information on mutant Arabidopsis lines (Sakurai et al. 2011, Myouga et al. 2013, Akiyama et al. 2014), inclusion of experimental and computational methods using gene expression data and experimentally derived (or predicted) protein–protein interactions (Kourmpetis et al. 2011), and inclusion of similarity clustering among protein sequences in the SALAD database (http://salad.dna.affrc.go.jp/salad/) (Mihara et al. 2010).

Plant Cell Physiol. 56(1): e11(1–9) (2015) doi:10.1093/pcp/pcu176, Advance Access publication on 29 November 2014, available online at www.pcp.oxfordjournals.org

To improve the annotations further, we attempted to utilize the proteome information. In this study, we adopted a new method, which we use to study predicted secondary structures and functions of proteins to make plant gene annotations easier to understand. Because proteins possess specific structures and functions, obtaining this information helps us to elucidate the corresponding gene functions. Here, we report analyses of multiple physicochemical and secondary structural parameters of whole-protein sequences obtained from representative data sets of six plant species, *A. thaliana*, *Glycine max* (soybean), *Populus trichocarpa* (poplar), *Oryza sativa* (rice), *Physcomitrella patens* (moss) and *Cyanidioschyzon merolae* (alga). The genome sequences of these six species have been completely determined previously. We propose new annotations for the predicted functional regions corresponding to the unannotated genes of Arabidopsis and rice. We also developed the Plant-PrAS (Plant Protein Annotation Suite) database, which includes the annotations generated in this study.

Results and Discussion

Protein sequence sets

We prepared non-redundant sequence sets from the whole-protein sequences, using the procedure that was described in our previous study (Kurotani et al. 2014). Protein sequences with length ranging from 50 to 2,000 amino acid residues were extracted from the databases for analysis in this study. Redundant data in these protein sequences were removed using the OrthoMCL software (Chen et al. 2006). The final filtered proteomes contained 26,326, 34,972, 35,791, 40,087, 35,908, 30,654 and 4,595 non-redundant protein sequences corresponding to Arabidopsis, soybean, poplar, MSU Rice, RAP-DB (rice), moss and algae, respectively. In addition, 20,572 and 6,216 non-redundant protein sequences of the mouse and yeast, respectively, were also prepared as a reference (mammals and fungi).

Secondary structural properties of proteins

Transmembrane helices, domain linkers and signal peptides. We calculated the number of transmembrane helices, domain linkers and signal peptides in the protein sequences using the TMHMM (Krogh et al. 2001), DROP (Ebina et al. 2011) and SignalP (Petersen et al. 2011) software packages. For example, transmembrane helices in proteins play an important role in the transport of various substances across biological membranes, and signal peptides are present either in secreted proteins or in transmembrane proteins. Predicting the numbers of transmembrane helices, domain linkers and signal peptides in a protein sequence does not lead to the prediction of protein function directly but does elucidate the corresponding intramolecular interactions. The results obtained using the above-mentioned analytical tools suggested that *P. patens* and *C. merolae* possess a smaller number of transmembrane helices, domain linkers and signal peptides than do the other plant species examined (**Supplementary Table S1**). We can speculate that the physiology of higher order plants (vascular plants) involves a variety of functions that require the presence of a greater number of transmembrane helices, domain linkers and signal peptides compared with lower order plants.

Intrinsically disordered regions (IDRs) and post-translational modifications. Recently, it was reported that the number of IDRs in proteins is higher among the monocots compared with other types of plants (Kurotani et al. 2014). In our processed data sets, the IDR content of the monocot rice calculated using the RONN software (Yang et al. 2005) was higher than that of the other five plant species, in agreement with our recent study (Kurotani et al. 2014). Moreover, in angiosperms, the proteins showing high IDR content generally show higher reactivity in these regions (e.g. post-translational modifications such as phosphorylation and *O*-glycosylation) (Iakoucheva et al. 2004, Gao and Xu 2012, Yao et al. 2012). The IDRs are considered vulnerable to an attack by a reactive molecule owing to their high flexibility and easy accessibility. The frequencies of *N*-glycosylation sites in Arabidopsis, soybean and poplar (all dicots) were higher than those in the monocot rice (**Supplementary Table S1**). On the other hand, the frequency of *O*-glycosylation in the monocot rice was higher than that in the dicot species (**Supplementary Table S1**). The reason is that *O*-glycosylation occurs preferably in IDRs as a non-conservation property involved in functional diversity and structural stability (Nishikawa et al. 2010), whereas *N*-glycosylation does not strongly correlate with IDR content; this is because *N*-glycosylation is known to occur co-translationally before a protein is fully folded (Petrescu et al. 2004, Kurotani et al. 2014). Moreover, a higher IDR content results in unstable protein structures and problems with crystallization (Oldfield et al. 2013). Accordingly, we observed that rice proteins, as a whole, tend to show higher susceptibility to phosphorylation and *O*-glycosylation but fail to crystallize during three-dimensional structural analysis owing to the presence of a greater number of IDRs compared with Arabidopsis, soybean and poplar.

Functional regions. In order to obtain useful information on the functional regions in the protein sequence data sets, Plant-PrAS prepares the results by means of the PASS tool, which identifies highly conserved sequence regions using existing protein sequence sets (Kuroda et al. 2000), and by means of the Rosetta Stone method, which identifies the regions likely to be involved in protein–protein interactions, using a comparative genomic approach (Enright et al. 1999, Marcotte 1999). 'Rosetta Stone composites' are paired regions in a protein sequence, and 'Rosetta Stone components' are the elements of the Rosetta Stone composites (Enright et al. 1999). Plant-PrAS provides the results on both the Rosetta Stone composites and components to help find functional regions. As a result of the calculations on the six plants species, we obtained 32,158 protein sequence hits with the PASS tool, 19,627 with the Rosetta stone composites and 13,428 with the Rosetta Stone components (**Supplementary Table S2**). In addition, Plant-PrAS can combine and provide the results of the PASS and Rosetta Stone methods to improve the reliability of the functional region annotations. Finally, we identified functional regions in 52,049

non-overlapping protein sequence hits by means of PASS and Rosetta Stone composites/components from the six plant species.

Detection of novel functional regions in the unannotated protein sequences of Arabidopsis and rice. We extracted the unannotated protein sequences of Arabidopsis and rice from the annotation information file, which contained 5,180 sequences for Arabidopsis, 15,322 for MSU Rice and 14,716 for RAP-DB (see the Materials and Methods and **Supplementary Table S3**). Subsequently, we identified candidate protein sequences, including the novel functional regions in the unannotated sequences in Arabidopsis and rice, by using PASS and Rosetta Stone composite/component methods and the Pfam database (Finn et al. 2014). As a result, we assigned 2,470 proteins to Pfam. For those proteins not assigned to Pfam, we found novel functional regions in 523 proteins (PASS), in 1,008 proteins (Rosetta Stone composites) and in 700 proteins (Rosetta Stone components; **Table 1**). Finally, we annotated 1,732 non-overlapping proteins from the unannotated sequences in Arabidopsis and rice using the methods for detection of functional regions. With regard to the above analyses using the PASS tool and the Rosetta Stone methods, we applied this tool and the methods to UniProt-plant (UniProt Consortium 2014), which is a collection of plant protein sequences that includes abundant as well as unknown functional protein sequences. The above results have the possibility that novel functional regions are identified on the information on unannotated proteins from this study.

The search interface of Plant-PrAS

We developed a publicly accessible web-based database, Plant-PrAS (http://plant-pras.riken.jp/), which currently stores 208,333 protein sequence records derived from genome-wide analysis of six major plant species (*A. thaliana*, soybean, poplar, rice, *P. patens* and *C. merolae*) and 26,788 protein sequence records derived from the two reference species (the mouse and yeast). Each protein sequence is annotated with information on the calculated protein properties and classified physicochemical properties [length, percentage of charged amino acids, percentage of non-polar amino acids, percentage of acidic amino acids, percentage of basic amino acids, percentage low complexity, the grand average of hydrophobicity (GRAVY) and the pI], and protein secondary structural properties [percentage solvent accessibility, percentage of β-sheets, percentage of IDRs, and the presence of a signal peptide(s), transmembrane helices, S–S bonds and domain linkers]; functional annotations against the eukaryotic orthologous groups (KOG) of Clusters of Orthologous Groups of proteins (COGs) (Tatusov et al. 2000, Tatusov et al. 2003), Protein Data Bank (PDB) (Berman et al. 2000, Berman et al. 2013), UniProt-SwissProt and UniProt-plant; functional regions detected using the PASS tool and Rosetta Stone methods; and other properties such as protein solubility, subcellular localization, the number of *N*/*O*-glycosylation sites and the number of ubiquitination sites. Plant-PrAS offers powerful search features and statistical information on various calculations. An entire data set can be downloaded as a file. The database has three types of search functions: 'Property Search', 'Keyword Search' and 'ID Search'.

Property Search. Plant-PrAS allows users easily to combine search results using a Property Search, designed for obtaining abundant proteomic information all at once. The Property Search can extract data from multiple species in our data set and from multiple protein sequence properties such as length, percentage of charged amino acids and GRAVY; from protein structural properties such as S–S bonds, transmembrane helices and percentage IDRs; from protein annotation data such as the information on Pfam, UniProt and the Enzyme Commission (EC) number; from protein modification/localization data such as *O*/*N*-glycosylation and subcellular localization; and from functionally conserved regions and interaction regions (**Fig. 1A**). On this page, for instance, a user can select an annotated or unannotated sequence from Arabidopsis and rice. Combined selection of the unannotated sequences and the calculation tools is available, helping to find a novel annotation corresponding to the unannotated sequences. For example, when a user performs a search by checking the options 'unannotated sequences', 'Rosetta Stone composite hit UniProt-plant' and 'Pfam not-hit/unknown' for Arabidopsis, MSU Rice and RAP-DB, the results show the presence of 421, 280 and 307 candidate protein sequences, respectively, including those corresponding to a novel functional region (**Table 1**). On the Results page of the 'Property Search', users can browse through

Table 1 Detection of novel functional regions in the unannotated protein sequences of Arabidopsis and rice by means of Plant-PrAS (Plant Protein Annotation Suite database)

Plant species	Unannotated sequences	Pfam(+)[a]	Pfam(−)		
			PASS(+)[b]	Rosetta Stone	
				Composite(+)[c]	Component(+)[d]
Arabidopsis	5,180	312	111	421	63
MSU Rice	15,322	640	111	280	225
RAP-DB (rice)	14,716	1,518	301	307	412
Total	35,218	2,470	523	1,008	700

[a] The number of protein hits in the Pfam database.
[b] The number of proteins whose functional regions were detected by PASS but not by Pfam [Pfam(−)].
[c] The number of proteins whose functional regions were detected as Rosetta Stone composites with Pfam(−).
[d] The number of proteins whose functional regions were detected as Rosetta Stone components with Pfam(−).

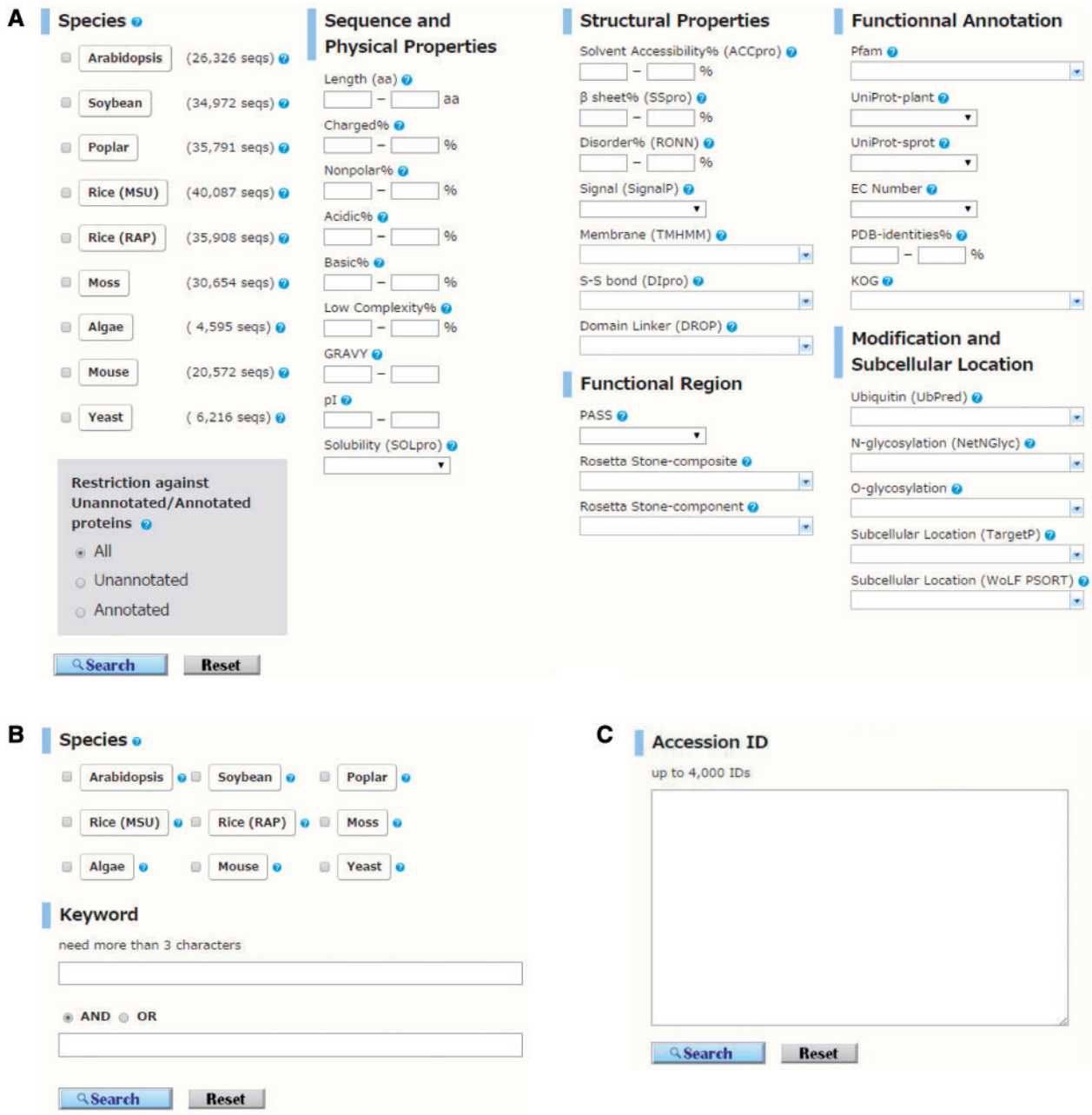
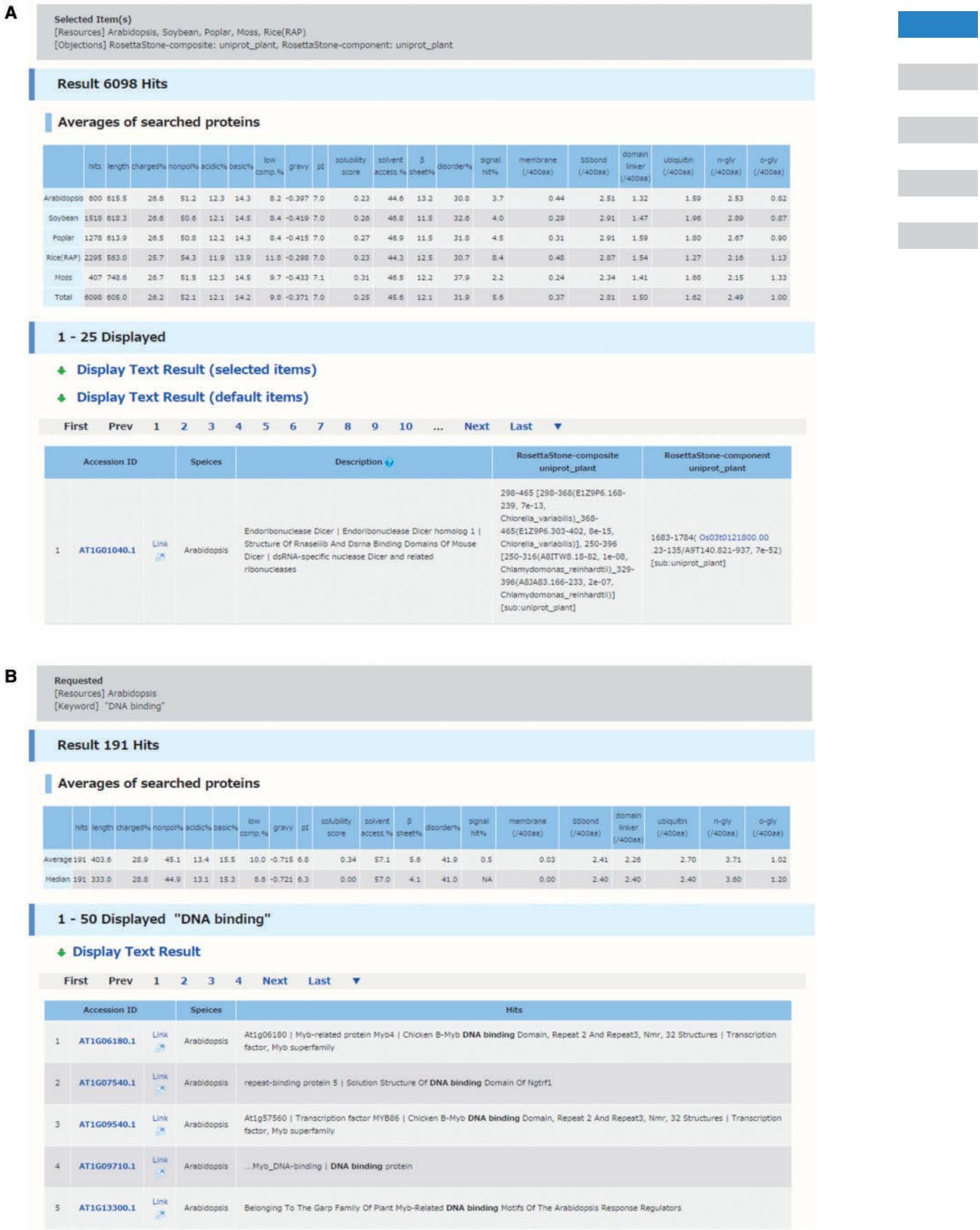


Fig. 1 Search interfaces of Plant-PrAS. A user can search for multiple protein sequence properties on the 'Property Search' page (A). The user can also search for objective records using the 'Keyword Search' function (B). 'ID Search' makes it possible to search for objective records by IDs from public databases (C).

the protein features by the averages of the extracted sequence properties in the statistics table (**Fig. 2A**). The search results can also be downloaded as a text file. Plant-PrAS houses information on charged amino acids, IDRs and solvent accessibility. Thus, the Property Search feature can be utilized for plant proteomic analyses.

Keyword Search. This option can be used to find protein sequences in our data sets, by using any keywords containing three characters corresponding to the protein descriptions from Pfam, PDB, KOG and UniProt (**Fig. 1B**). This feature allows the user to select the AND/OR function during a multiple keyword search. The extracted records are listed on the results page with short descriptions (**Fig. 2B**). The user can click on an 'ID' to obtain detailed information on a protein.

ID Search. Plant-PrAS allows a user to extract general IDs supported by the public databases pertaining to our data sets, by using the ID Search function (**Fig. 1C**). The extracted records are listed on the Results page with short descriptions (**Fig. 2B**). The user can click on an 'ID' to obtain detailed information on a protein.



A

Selected Item(s)
[Resources] Arabidopsis, Soybean, Poplar, Moss, Rice(RAP)
[Objections] RosettaStone-composite: uniprot_plant, RosettaStone-component: uniprot_plant

Result 6098 Hits

Averages of searched proteins

	hits	length	charged%	nonpol%	acidic%	basic%	low comp.%	gravy	pI	solubility score	solvent access.%	β sheet%	disorder%	signal hit%	membrane (/400aa)	SSbond (/400aa)	domain linker (/400aa)	ubiquitin (/400aa)	n-gly (/400aa)	o-gly (/400aa)
Arabidopsis	600	615.5	26.6	51.2	12.3	14.3	8.2	-0.397	7.0	0.23	44.6	13.2	30.8	3.7	0.44	2.51	1.32	1.59	2.53	0.82
Soybean	1518	618.3	26.6	50.6	12.1	14.5	8.4	-0.419	7.0	0.26	46.8	11.5	32.6	4.0	0.29	2.91	1.47	1.96	2.89	0.87
Poplar	1278	613.9	26.5	50.8	12.2	14.3	8.4	-0.415	7.0	0.27	46.9	11.5	31.8	4.5	0.31	2.91	1.59	1.80	2.67	0.90
Rice(RAP)	2295	563.0	25.7	54.3	11.9	13.9	11.8	-0.298	7.0	0.23	44.3	12.5	30.7	8.4	0.48	2.87	1.54	1.27	2.16	1.13
Moss	407	748.6	26.7	51.5	12.3	14.5	9.7	-0.433	7.1	0.31	46.5	12.2	37.9	2.2	0.24	2.34	1.41	1.68	2.15	1.33
Total	6098	605.0	26.2	52.1	12.1	14.2	9.8	-0.371	7.0	0.25	45.6	12.1	31.9	5.6	0.37	2.81	1.50	1.62	2.49	1.00

1 - 25 Displayed

Display Text Result (selected items)
Display Text Result (default items)

First Prev 1 2 3 4 5 6 7 8 9 10 ... Next Last ▼

	Accession ID		Speices	Description	RosettaStone-composite uniprot_plant	RosettaStone-component uniprot_plant
1	AT1G01040.1	Link	Arabidopsis	Endoribonuclease Dicer | Endoribonuclease Dicer homolog 1 | Structure Of Rnaseiiib And Dsrna Binding Domains Of Mouse Dicer | dsRNA-specific nuclease Dicer and related ribonucleases	298-465 [298-368(E1Z9P6.168-239, 7e-13, Chlorella_variabilis)_368-465(E1Z9P6.303-402, 8e-15, Chlorella_variabilis)], 250-396 [250-316(A8ITW8.18-82, 1e-08, Chlamydomonas_reinhardtii)_329-396(A8JA83.166-233, 2e-07, Chlamydomonas_reinhardtii)] [sub:uniprot_plant]	1683-1784(Os03t0121800.00 .23-135/A9T140.821-937, 7e-52) [sub:uniprot_plant]

B

Requested
[Resources] Arabidopsis
[Keyword] "DNA binding"

Result 191 Hits

Averages of searched proteins

	hits	length	charged%	nonpol%	acidic%	basic%	low comp.%	gravy	pI	solubility score	solvent access.%	β sheet%	disorder%	signal hit%	membrane (/400aa)	SSbond (/400aa)	domain linker (/400aa)	ubiquitin (/400aa)	n-gly (/400aa)	o-gly (/400aa)
Average	191	403.6	28.9	45.1	13.4	15.5	10.0	-0.715	6.8	0.34	57.1	5.6	41.9	0.5	0.03	2.41	2.26	2.70	3.71	1.02
Median	191	333.0	28.8	44.9	13.1	15.3	8.6	-0.721	6.3	0.00	57.0	4.1	41.0	NA	0.00	2.40	2.40	2.40	3.60	1.20

1 - 50 Displayed "DNA binding"

Display Text Result

First Prev 1 2 3 4 Next Last ▼

	Accession ID		Speices	Hits
1	AT1G06180.1	Link	Arabidopsis	At1g06180 | Myb-related protein Myb4 | Chicken B-Myb **DNA binding** Domain, Repeat 2 And Repeat3, Nmr, 32 Structures | Transcription factor, Myb superfamily
2	AT1G07540.1	Link	Arabidopsis	repeat-binding protein 5 | Solution Structure Of **DNA binding** Domain Of Ngtrf1
3	AT1G09540.1	Link	Arabidopsis	At1g57560 | Transcription factor MYB86 | Chicken B-Myb **DNA binding** Domain, Repeat 2 And Repeat3, Nmr, 32 Structures | Transcription factor, Myb superfamily
4	AT1G09710.1	Link	Arabidopsis	...Myb_DNA-binding | **DNA binding** protein
5	AT1G13300.1	Link	Arabidopsis	Belonging To The Garp Family Of Plant Myb-Related **DNA binding** Motifs Of The Arabidopsis Response Regulators

Fig. 2 Examples of search results in Plant-PrAS. (A) The results of Property Search. (B) The results of Keyword or ID Search.

Annotation details of proteins in Plant-PrAS. The Annotation Details page of Plant-PrAS displays basic information on each protein, such as protein sequence and similar proteins in the same species and among other species (**Fig. 3A**). Similarly, the page contains information on physical and sequence properties (**Fig. 3B**), structural properties (**Fig. 3C**), detected functional regions (**Fig. 3D**), functional annotation (**Fig. 3E**) and modifications and subcellular localization (**Fig. 3F**). To facilitate evaluation of various protein properties, the page shows the summary with average, median and percentile values in relation to proteins from the same species as a background distribution (**Fig. 3G**).

Exploration of the properties of unannotated proteins. We wanted to determine whether a data set obtained using Plant-PrAS provides new insights into the functions of unannotated proteins. Here, we present an example of deduction of such a function.

Generally, the propensity for solubility or cell-free synthesis of a protein in *Escherichia coli* can be predicted by analyzing various properties of the protein sequence (Luan et al. 2004, Tartaglia et al. 2009, Kurotani et al. 2010, Agostini et al. 2012). The results produced by the protein solubility tool showed that the percentage of soluble proteins was higher among the unannotated proteins than among the annotated proteins ($P < 0.05$ *in the t*-test of differences between the annotated and unannotated proteins; **Table 2**). This result shows that unannotated proteins may contribute to the success of protein solubilization experiments. Moreover, the functional regions extracted using the Rosetta Stone method have the potential to interact with each partner region. Therefore, functional region candidates of this property identified by Plant-PrAS may aid in the discovery

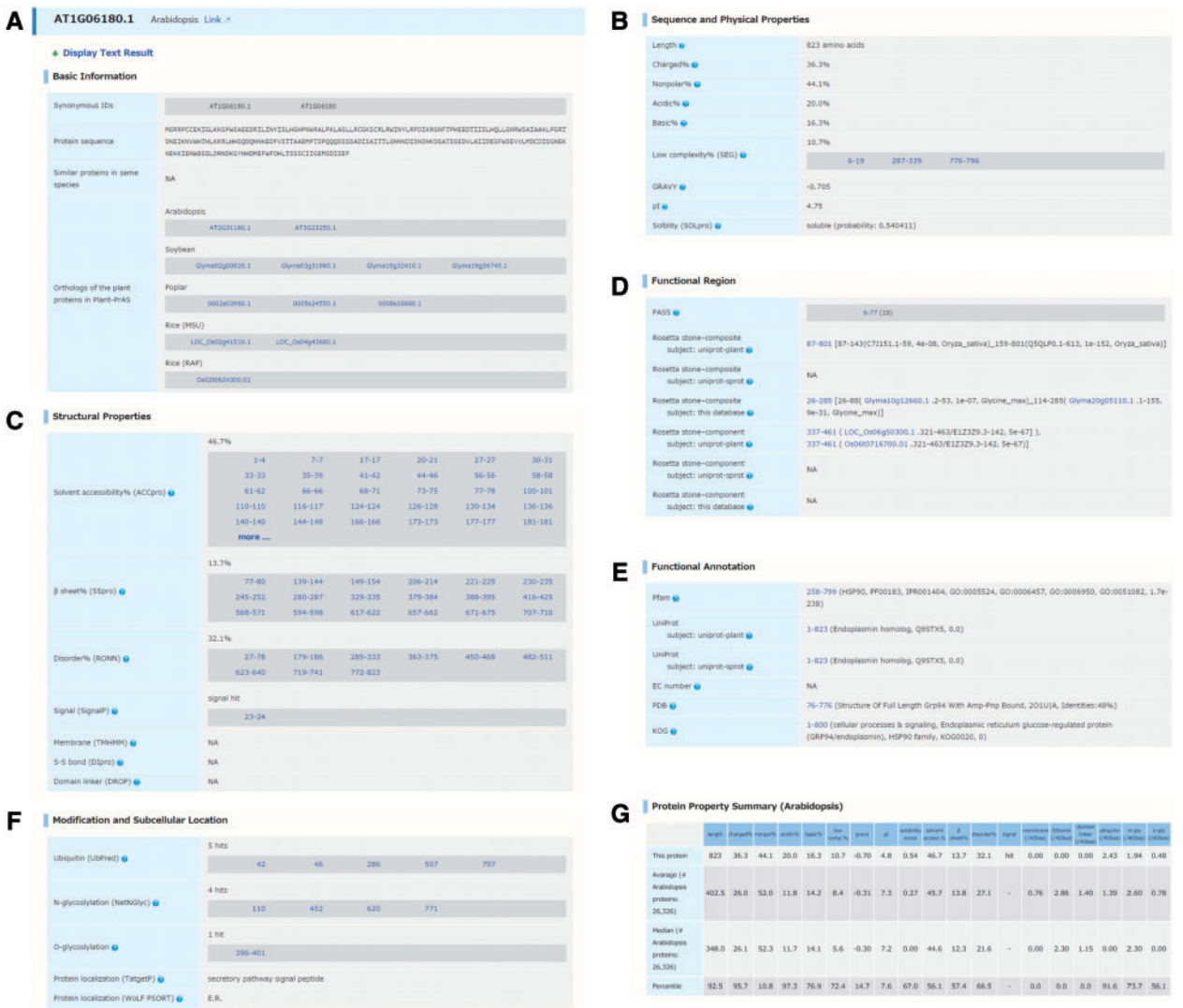


Fig. 3 Typical examples of the annotation details of proteins in Plant-PrAS. (A) Basic information on a protein in Plant-PrAS. (B) Physical and sequence properties. (C) Structural properties. (D) The detected functional regions. (E) Functional annotation. (F) Modifications and subcellular localization. (G) Summary with average, median and percentile values in relation to proteins from the same species (as a background distribution).

Table 2 The percentage of soluble proteins (among all proteins) in Arabidopsis and rice

Species	Category	No. of sequences (soluble/total)[a]	Percentage of soluble proteins
Arabidopsis	Annotated	7,545/21,146	35.7%
	Unannotated	2,389/5,180	46.1%
MSU Rice	Annotated	8,432/24,765	34.0%
	Unannotated	8,177/15,322	53.4%
RAP-DB (rice)	Annotated	7,579/21,192	35.8%
	Unannotated	7,746/14,716	52.7%

In Arabidopsis and rice, there is a greater number of soluble proteins among unannotated proteins than among annotated proteins ($P < 0.05$ in the t-test of the differences between annotated and unannotated proteins).

[a] Proteins that have a solubility score >0.5 according to the SOLpro software were regarded as soluble proteins.

of novel annotated proteins that contain the novel functional regions.

Materials and Methods

Protein sequence resources

We analyzed the whole-protein sequences derived from the genome sequences of six major model plant species, namely Brassicaceae (Arabidopsis) (Arabidopsis Genome Initiative 2000), Fabaceae (soybean) (Schmutz et al. 2010), Salicaceae (poplar) (Tuskan et al. 2006), Poaceae (rice) (Yu et al. 2002, International Rice Genome Sequencing Project 2005), Funariaceae (*P. patens*) (Rensing et al. 2008) and (*C. merolae*) (Matsuzaki et al. 2004). The Arabidopsis proteomic sequence set was retrieved from TAIR (Lamesch et al. 2012). The rice sequences were retrieved from RAP-DB (Sakai et al. 2013) and from the MSU Rice Genome Annotation Project website (Ouyang et al. 2007, Kawahara et al. 2013). *Cyanidioschyzon merolae* sequences were retrieved from the *C. merolae* Genome Project website. The other plant sequences were retrieved from Phytozome (Goodstein et al. 2012). In addition, mouse (Mouse Genome Sequencing Consortium 2002) and yeast (Mewes et al. 2002) sequences were retrieved from the National Center for Biotechnology Information (NCBI) (ftp://ftp.ncbi.nih.gov/genomes/M_musculus/protein/) and from Munich Information Center for Protein Sequences (MIPS) (ftp://ftpmips.gsf.de/fungi/yeast/), respectively. They were used as reference proteome sets. Subsequently, we prepared non-redundant proteome sequence sets of the target organisms using the OrthoMCL software (Chen et al. 2006) with the runtime options pi_cutoff = 90, pmatch_cutoff = 90 and pv_cutoff = 1e-30.

Analysis of a protein sequence

Physicochemical properties. The percentage of polar, charged, acidic and basic amino acids as well as the isoelectric point were calculated using the ProteoMix software (Chikayama et al. 2004). The GRAVY index was calculated using the GRAVY algorithm (Kyte and Doolittle 1982).

Secondary structural properties. For prediction of these properties, we used the following software tools: SignalP (Petersen et al. 2011) to detect the presence of signal peptides, TMHMM (Krogh et al. 2001) to identify transmembrane helix domains, DROP (Ebina et al. 2011) to find interdomain linkers, DIpro (Cheng et al. 2006) to find S–S bonds, SSpro (Cheng et al. 2005) to identify secondary structures, ACCpro (Cheng et al. 2005) to analyze solvent accessibility and RONN (Yang et al. 2005) to find IDRs.

Functional and structural annotations. We used all the protein sequences for searches in KOG (Tatusov et al. 2000, Tatusov et al. 2003) and in UniProt-SwissProt/UniProt-plant (UniProt Consortium 2014) using BLASTP with the runtime options 'cutoff E-value' 1e-10 or 1e-5, respectively. Similarly, we used all protein sequences for searches in the PDB (Berman et al. 2000, Berman et al. 2013) using BLASTP with >50% identity. The Pfam annotations (Finn et al. 2014) and the EC number (Bairoch 2000) were obtained using the InterProScan software (Hunter et al. 2012).

Other properties. To analyze other properties, we used the following software packages: SOLpro (Magnan et al. 2009) for protein solubility, TargetP (Emanuelsson et al. 2000) and WoLF PSORT (Horton et al. 2007) for subcellular localization, NetNglyc (R. Gupta et al. unpublished) for *N*-glycosylation, the Gomond's algorithm (Gomord et al. 2010) for *O*-glycosylation, and UbPred (Radivojac et al. 2010) for ubiquitination.

Detection of the functional regions. This procedure was performed on protein sequences by means of the proteome sequence set of the UniProt-plant database and the PASS tool (Kuroda et al. 2000), with the runtime options 'cutoff E-value'≤1e-7 and 'cutoff homolog'≥100, and the Rosetta Stone method (Enright et al. 1999, Marcotte 1999), with the cutoff E-value ≤1e-5, identities ≥35%, component length ≥50 amino acids and a component range from 10 to 30 amino acids, with runtime options similar to those described previously (Uversky 2002, Chia and Kolatkar 2004, Enault et al. 2005, Wallner and Elofsson 2005, Nayeem et al. 2006).

Extraction of the unannotated sequences in Arabidopsis and rice. The unannotated sequences of Arabidopsis and rice (MSU Rice and RAP-DB) were extracted from whole-protein sequences using the description terms shown in **Supplementary Table S3**.

Availability and implementation of the system

Plant-PrAS was implemented in a web application framework, MENTA, with MySQL as a database engine, and was tested in the following web browsers: Internet Explorer 11, Chrome 36 and Firefox 31.

Supplementary data

Supplementary data are available at PCP online.

Funding

This work was supported the Japan Society for the Promotion of Science [a Grant-in-Aid for Young Scientists (B) (18700106 to T.S.)].

Acknowledgments

We thank Alexander A. Tokmakov (Kobe University) and Takuhiro Yoshida and Kenji Akiyama (RIKEN) for their helpful comments and for management of the development environment.

Disclosures

The authors have no conflicts of interest to declare.

References

Agostini, F., Vendruscolo, M. and Tartaglia, G.G. (2012) Sequence-based prediction of protein solubility. *J. Mol. Biol.* 421: 237–241.

Akiyama, K., Kurotani, A., Iida, K., Kuromori, T., Shinozaki, K. and Sakurai, T. (2014) RARGE II: an integrated phenotype database of Arabidopsis mutant traits using a controlled vocabulary. *Plant Cell Physiol.* 55.

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* 408: 796–815.

Asamizu, E., Ichihara, H., Nakaya, A., Nakamura, Y., Hirakawa, H., Ishii, T. et al. (2014) Plant Genome DataBase Japan (PGDBj): a portal website for the integration of plant genome-related databases. *Plant Cell Physiol.* 55: e8.

Bairoch, A. (2000) The ENZYME database in 2000. *Nucleic Acids Res.* 28: 304–305.

Berman, H.M., Coimbatore Narayanan, B., Di Costanzo, L., Dutta, S., Ghosh, S., Hudson, B.P. et al. (2013) Trendspotting in the Protein Data Bank. *FEBS Lett.* 587: 1036–1045.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H. et al. (2000) The Protein Data Bank. *Nucleic Acids Res.* 28: 235–242.

Chen, F., Mackey, A.J., Stoeckert, C.J. Jr. and Roos, D.S. (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.* 34: D363–D368.

Chen, F.C., Wang, S.S., Chaw, S.M., Huang, Y.T. and Chuang, T.J. (2007) Plant Gene and Alternatively Spliced Variant Annotator. A plant genome annotation pipeline for rice gene and alternatively spliced variant identification with cross-species expressed sequence tag conservation from seven plant species. *Plant Physiol.* 143: 1086–1095.

Cheng, J., Randall, A.Z., Sweredoski, M.J. and Baldi, P. (2005) SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Res.* 33: W72–W76.

Cheng, J., Saigo, H. and Baldi, P. (2006) Large-scale prediction of disulphide bridges using kernel methods, two-dimensional recursive neural networks, and weighted graph matching. *Proteins* 62: 617–629.

Chia, J.M. and Kolatkar, P.R. (2004) Implications for domain fusion protein–protein interactions based on structural information. *BMC Bioinformatics* 5: 161.

Chikayama, E., Kurotani, A., Kuroda, Y. and Yokoyama, S. (2004) ProteoMix: an integrated and flexible system for interactively analyzing large numbers of protein sequences. *Bioinformatics* 20: 2836–2838.

Ebina, T., Toh, H. and Kuroda, Y. (2011) DROP: an SVM domain linker predictor trained with optimal features selected by random forest. *Bioinformatics* 27: 487–494.

Emanuelsson, O., Nielsen, H., Brunak, S. and von Heijne, G. (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* 300: 1005–1016.

Enault, F., Suhre, K. and Claverie, J.M. (2005) Phydbac 'Gene Function Predictor': a gene annotation tool based on genomic context analysis. *BMC Bioinformatics* 6: 247.

Enright, A.J., Iliopoulos, I., Kyrpides, N.C. and Ouzounis, C.A. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature* 402: 86–90.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R. et al. (2014) Pfam: the protein families database. *Nucleic Acids Res.* 42: D222–D230.

Gao, J. and Xu, D. (2012) Correlation between posttranslational modification and intrinsic disorder in protein. *Pac. Symp. Biocomput.* 94–103.

Gomord, V., Fitchette, A.C., Menu-Bouaouiche, L., Saint-Jore-Dupas, C., Plasson, C., Michaud, D. et al. (2010) Plant-specific glycosylation patterns in the context of therapeutic protein production. *Plant Biotechnol. J.* 8: 564–587.

Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J. et al. (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* 40: D1178–D1186.

Horton, P., Park, K.J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C.J. et al. (2007) WoLF PSORT: protein localization predictor. *Nucleic Acids Res.* 35: W585–W587.

Hunter, S., Jones, P., Mitchell, A., Apweiler, R., Attwood, T.K., Bateman, A. et al. (2012) InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.* 40: D306–D312.

Iakoucheva, L.M., Radivojac, P., Brown, C.J., O'Connor, T.R., Sikes, J.G., Obradovic, Z. et al. (2004) The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res.* 32: 1037–1049.

International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. *Nature* 436: 793–800.

Kawahara, Y., de la Bastide, M., Hamilton, J.P., Kanamori, H., McCombie, W.R., Ouyang, S. et al. (2013) Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. *Rice* 6: 4.

Kourmpetis, Y.A., van Dijk, A.D., van Ham, R.C. and ter Braak, C.J. (2011) Genome-wide computational function prediction of Arabidopsis proteins by integration of multiple data sources. *Plant Physiol.* 155: 271–281.

Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* 305: 567–580.

Kudo, T., Akiyama, K., Kojima, M., Makita, N., Sakurai, T. and Sakakibara, H. (2013) UniVIO: a multiple omics database with hormonome and transcriptome data from rice. *Plant Cell Physiol.* 54: E9.

Kuroda, Y., Tani, K., Matsuo, Y. and Yokoyama, S. (2000) Automated search of natively folded protein fragments for high-throughput structure determination in structural genomics. *Protein Sci.* 9: 2313–2321.

Kurotani, A., Takagi, T., Toyama, M., Shirouzu, M., Yokoyama, S., Fukami, Y. et al. (2010) Comprehensive bioinformatics analysis of cell-free protein synthesis: identification of multiple protein properties that correlate with successful expression. *FASEB J.* 24: 1095–1104.

Kurotani, A., Tokmakov, A.A., Kuroda, Y., Fukami, Y., Shinozaki, K. and Sakurai, T. (2014) Correlations between predicted protein disorder and post-translational modifications in plants. *Bioinformatics* 30: 1095–1103.

Kyte, J. and Doolittle, R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* 157: 105–132.

Lamesch, P., Berardini, T.Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R. et al. (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* 40: D1202–D1210.

Li, D., Berardini, T.Z., Muller, R.J. and Huala, E. (2012) Building an efficient curation workflow for the Arabidopsis literature corpus. *Database (Oxford)* 2012: bas047.

Luan, C.H., Qiu, S., Finley, J.B., Carson, M., Gray, R.J., Huang, W. et al. (2004) High-throughput expression of C. elegans proteins. *Genome Res.* 14: 2102–2110.

Magnan, C.N., Randall, A. and Baldi, P. (2009) SOLpro: accurate sequence-based prediction of protein solubility. *Bioinformatics* 25: 2200–2207.

Marcotte, E.M. (1999) Detecting protein function and protein–protein interactions from genome sequences. *Science* 285: 751–753.

Matsuzaki, M., Misumi, O., Shin, I.T., Maruyama, S., Takahara, M., Miyagishima, S.Y. et al. (2004) Genome sequence of the ultrasmall unicellular red alga Cyanidioschyzon merolae 10D. *Nature* 428: 653–657.

Mewes, H.W., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M. et al. (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* 30: 31–34.

Mihara, M., Itoh, T. and Izawa, T. (2010) SALAD database: a motif-based database of protein annotations for plant comparative genomics. *Nucleic Acids Res.* 38: D835–D842.

Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420: 520–562.

Myouga, F., Akiyama, K., Tomonaga, Y., Kato, A., Sato, Y., Kobayashi, M. et al. (2013) The Chloroplast Function Database II: a comprehensive collection of homozygous mutants and their phenotypic/genotypic traits for nuclear-encoded chloroplast proteins. *Plant Cell Physiol.* 54: E2.

Nayeem, A., Sitkoff, D. and Krystek, S. (2006) A comparative study of available software for high-accuracy homology modeling: from sequence alignments to structural models. *Protein Sci.* 15: 808–824.

Nishikawa, I., Nakajima, Y., Ito, M., Fukuchi, S., Homma, K. and Nishikawa, K. (2010) Computational prediction of O-linked glycosylation sites that preferentially map on intrinsically disordered regions of extracellular proteins. *Int. J. Mol. Sci.* 11: 4992–5009.

Obayashi, T., Okamura, Y., Ito, S., Tadaka, S., Aoki, Y., Shirota, M. et al. (2014) ATTED-II in 2014: evaluation of gene coexpression in agriculturally important plants. *Plant Cell Physiol.* 55: e6.

Oldfield, C.J., Xue, B., Van, Y.Y., Ulrich, E.L., Markley, J.L., Dunker, A.K. et al. (2013) Utilization of protein intrinsic disorder knowledge in structural proteomics. *Biochim. Biophys. Acta* 1834: 487–498.

Ouyang, S., Zhu, W., Hamilton, J., Lin, H., Campbell, M., Childs, K. et al. (2007) The TIGR Rice Genome Annotation Resource: improvements and new features. *Nucleic Acids Res.* 35: D883–D887.

Petersen, T.N., Brunak, S., von Heijne, G. and Nielsen, H. (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* 8: 785–786.

Petrescu, A.J., Milac, A.L., Petrescu, S.M., Dwek, R.A. and Wormald, M.R. (2004) Statistical analysis of the protein environment of N-glycosylation sites: implications for occupancy, structure, and folding. *Glycobiology* 14: 103–114.

Radivojac, P., Vacic, V., Haynes, C., Cocklin, R.R., Mohan, A., Heyen, J.W. et al. (2010) Identification, analysis, and prediction of protein ubiquitination sites. *Proteins* 78: 365–380.

Rensing, S.A., Lang, D., Zimmer, A.D., Terry, A., Salamov, A., Shapiro, H. et al. (2008) The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. *Science* 319: 64–69.

Rice Annotation Project (2007) Curated genome annotation of Oryza sativa ssp. japonica and comparative genome analysis with Arabidopsis thaliana. *Genome Res.* 17: 175–183.

Sakai, H., Lee, S.S., Tanaka, T., Numa, H., Kim, J., Kawahara, Y. et al. (2013) Rice Annotation Project Database (RAP-DB): an integrative and interactive database for rice genomics. *Plant Cell Physiol.* 54: e6.

Sakurai, T., Kondou, Y., Akiyama, K., Kurotani, A., Higuchi, M., Ichikawa, T. et al. (2011) RiceFOX: a database of Arabidopsis mutant lines overexpressing rice full-length cDNA that contains a wide range of trait information to facilitate analysis of gene function. *Plant Cell Physiol.* 52: 265–273.

Sakurai, T., Yamada, Y., Sawada, Y., Matsuda, F., Akiyama, K., Shinozaki, K. et al. (2013) PRIMe Update: innovative content for plant metabolomics and integration of gene expression and metabolite accumulation. *Plant Cell Physiol.* 54: E5.

Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J., Mitros, T., Nelson, W. et al. (2010) Genome sequence of the palaeopolyploid soybean. *Nature* 463: 178–183.

Tartaglia, G.G., Pechmann, S., Dobson, C.M. and Vendruscolo, M. (2009) A relationship between mRNA expression levels and protein solubility in *E. coli*. *J. Mol. Biol.* 388: 381–389.

Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V. et al. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4: 41.

Tatusov, R.L., Galperin, M.Y., Natale, D.A. and Koonin, E.V. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* 28: 33–36.

Tuskan, G.A., DiFazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U. et al. (2006) The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science* 313: 1596–1604.

UniProt Consortium (2014) Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* 42: D191–D198.

Uversky, V.N. (2002) Natively unfolded proteins: a point where biology waits for physics. *Protein Sci.* 11: 739–756.

Wallner, B. and Elofsson, A. (2005) All are not equal: a benchmark of different homology modeling programs. *Protein Sci.* 14: 1315–1327.

Yang, Z.R., Thomson, R., McNeil, P. and Esnouf, R.M. (2005) RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics* 21: 3369–3376.

Yao, Q., Gao, J., Bollinger, C., Thelen, J.J. and Xu, D. (2012) Predicting and analyzing protein phosphorylation sites in plants using musite. *Front. Plant Sci.* 3: 186.

Yu, J., Hu, S., Wang, J., Wong, G.K., Li, S., Liu, B. et al. (2002) A draft sequence of the rice genome (Oryza sativa L. ssp. indica). *Science* 296: 79–92.