Contemporary Life Science Research (CLSR)
ISSN (Online): **XXXX-XXX**
Volume 01 | Issue 01 | March 2026
DOI:

Comparative Genomic Characterization and Evolutionary Analysis of PRT Gene Family Members in Plants

Tongbram Punshi Singh[1], Boris Aheibam[2]

[1]Ph.D , Mizoram University, punshi.tongbra@hotmail.com, ORCID ID: 0000-0002-2656-1768, Department of Biotechnology

[2]Ph.D , Mizoram University, Email ID: boriz.aheibam@gmail.com, ORCID ID: 0000-0002-4592-6312, Department of Biotechnology

Received Date: **20-12-2025** | Revised Date: **01-02-2026** | Accepted Date: **06-03-2026** | Published Date: **26-03-2026**

ABSTRACT
In plants, proteins related to phosphoribosyltransferase participate in metabolic processes, and a wide search of protein databases may yield several families of proteins with the same annotation keywords. A collection of 36 candidates PRT-labelled or PRT associated proteins were characterized in this exploratory analysis from Oryza sativa, Zea mays, Hordeum vulgare, Arabidopsis thaliana and Setaria italica. The parsed candidate amino acid sequences were evaluated using Python scripts for length, molecular weight, theoretical isoelectric point, hydropathy, sequence quality and similarity to three-amino-acid words. A reproducible script was used to generate species level descriptive tables, a pairwise similarity matrix and an exploratory hierarchical clustering visualization. Eight of the rice, seven of the maize, five of the barley, eight of the Arabidopsis and eight of the Setaria candidate proteins were included. There were significant variation in mean protein lengths among the 5 screening groups and 2 candidate records were identified for automatic check. These outputs can be easily used for downstream validation and can serve as a way to assess candidate datasets prior to phylogenetic analysis. The absence of accession supported family validation and the lack of coding DNA sequences and genomic coordinates in the dataset prevents the calculation of gene-family expansion, molecular selection, chromosomal synteny, and natural evolutionary relationship. It is necessary to do additional conserved domain checking and reference genome mapping to form biological conclusions.

Keywords: Phosphoribosyltransferase, PRT gene family, protein sequence screening, comparative genomics, plant bioinformatics

1. Introduction

Enzymes associated with phosphoribosyltransferase are involved in the salvage and other metabolic activities of the cell. (Zhou et al., 2026). Their activities are defined by transfer of the phosphoribosyl moiety from phosphoribosyl pyrophosphate (PRPP) to the proper substrates, but there are other enzymes with this type of reaction that do not belong to the same evolutionarily coherent protein family (Witte & Herde, 2020; Moffatt & Ashihara, 2002). Nucleotide metabolism relates to cell proliferation, tissue differentiation and resource allocation and adaptive responses in plants. Therefore, gene-family surveys need to differentiate between functional similarity and true sequence homology between members of a gene family and make comparisons among the members of a gene family in divergent species. For abbreviations like PRT, this difference is crucial because such abbreviations can be used for various classes of enzymes in genomic annotations.

Comparative protein analysis provides a first step into analysing the conservation of sequences across plant lineages. (Ohler et al., 2019). The broad protein properties described in this section (amino acid length, isoelectric point, hydropathy and predicted molecular weight) serve as a summary of the general features of the protein, whereas sequence alignment and conserved-domain testing are better indicators of common ancestry (Moffatt & Somerville, 1988; Berthomé et al., 2008). Physicochemical descriptors are not sufficient to assign a function to proteins, but their distributions can be used to find out-of-the ordinary proteins and potential gene prediction issues. Very short or very long candidate proteins can indicate partial proteins, fused protein or an alternative protein isoforms or proteins retrieved by a generic search word. Thus a well-documented screening procedure is highly important prior to making any conclusions on the basis of species-level comparisons.

The five species analysed here belong to both monocotyledonous and dicotyledonous plants (rice, maize, barley, foxtail millet and Arabidopsis). (Jia et al., 2024). This taxonomic diversity is beneficial when studying the evolution of homologous proteins as long as identification and orthology can be confirmed with a consistent set of criteria (Le Ret et al., 2018; Artur et al., 2019). The rice sequences are taken from a published PRT study and provide a useful reference, while the other candidate proteins are from a compiled sequence screening collection. But, protein titles and database match of key-words might lead to an uncritical interpretation of adenine, uracil, nicotinate and other proteins with phosphoribosyltransferases as one family.

For phylogenetic reconstruction, it is necessary to have homologous sequences, properly aligned regions, an evolutionary model that is defensible, and an explicit consideration of support values. (Wafula et al., 2023). Apparently species-specific clades can be difficult to interpret due to unvalidated sequences that can create distortions in alignments and tree topologies (Jiao et al., 2014; Zhao et al., 2021). While similarities based on k-mers and hierarchical clustering might be useful for identifying gross relationships and possible outliers early, the dendrograms based on these can not be used in place of model-based phylogenetic inference. Similarly, genomic coordinates, coding sequences, and gene models are all needed to infer gene duplication, synteny, and selection pressure, which all cannot be performed using protein sequences alone.

A focus on reproducible sequence screening can provide a better analytical focus of a study prior to stronger biological interpretations. (Woodhouse et al., 2011). The number of candidates and parameters of sequences can be determined directly, screening flags can be used to direct selective review of ambiguous proteins (Rody et al., 2017; Emms & Kelly, 2019). These differences should be reported to avoid the assumption that all records that are automatically retrieved are confirmed orthologs. In the present workflow, descriptive evidence and questions that need an external genome annotation and conserved domain verification are therefore separated. It also offers tabulated results and illustrations, whose computational definitions are explicit, and not attributing unsupported genomic characteristics to the proteins.

The study was guided by three objectives:

- to summarize the distribution and physicochemical properties of PRT-labelled candidates across five selected plants
- to quantify preliminary sequence resemblance and identify potentially discordant records
- to establish a reproducible screening foundation for subsequent family-level validation and evolutionary characterization.

These objectives are deliberately limited to conclusions supportable by the available amino acid sequences. A completed comparative genomic investigation would require mapping verified homologs to annotated chromosomes, testing domain conservation, generating coding-sequence alignments, and reconstructing phylogenies with appropriate support metrics.

2. Materials and Methods

2.1. Sequence collection and species classification

Protein FASTA files were previously compiled from public search of sequences and from a published study of rice which served as a reference study and were combined in one five species candidate package. A total of 36 sequences of rice, maize, barley, arabidopsis and foxtail millet were used as working screening subset. Record identifiers and FASTA record descriptions were kept. Species assignment was done based on explicit species prefix given for rice, maize and barley in the study or by using the species name found in downloaded NCBI FASTA headers. This assignment was considered as a source-label classification, and not independent evidence for orthology or common family membership. The extra candidate FASTA records were recorded independently and not displayed as part of the rice reference-data-source.

2.2. Input processing and quality control

Biopython was used to programatically read the compressed input archive. All amino acid sequences were capitalized and a stop sign (asterisk) was removed from the end of each amino acid sequence prior to the calculation of the descriptors. The internal stop characters, and nonstandard amino acid symbols were marked. Clearly incompatible annotations (for example, phosphoribosyl pyrophosphate synthases) were checked in the header descriptions. For each species, repeated proteins were verified. Length outliers in protein data were kept in the total inventory, but considered with caution during the interpretation. Candidates' status and all quality-control flags were exported for review.

2.3. Physicochemical characterization

The length of protein sequences was determined by the number of amino acid residues (excluding the stop symbol at the end of the sequence). The molecular weight, IEP (isoelectric point) and grand average of hydropathy were calculated in the software Biopython, ProteinAnalysis. The molecular weights were given in kilodaltons. From the measurements taken at the candidate level, species-level means and standard deviations were then calculated. These values were assumed to be computational values based on amino acid composition rather than actual measure of protein mass, charge, stability and cellular localization.

2.4. Screening similarity calculation

All overlapping tripeptides were built for each of the cleaned protein sequences. Using the above definition of similarity, the Jaccard coefficient, the similarity between proteins was calculated by dividing the number of the triplets that are common to both proteins by the number of the triplets in the union of the two sets. This metric was on a scale of zero to one, and represented the number of shared short sequence words. It was not a measure of percent amino acid identity, global alignment similarity or evolutionary distance. Pairwise scores were kept in a symmetric similarity matrix and mean values were summarised within and between species.

2.5. Exploratory grouping and visualization

The pairwise tripeptide Jaccard dissimilarity matrix (1-similarity score) was used for hierarchical clustering. Agglomeration was done using average linkage and the resulting dendrogram was displayed as exploratory protein candidate grouping. The dendrogram was not referred to as a phylogenetic tree because no multiple sequence alignment, substitution model or branch-support assessment was performed. Four numbers were produced: species-wise sequence count, length distribution, similarity heatmap and exploratory clustering dendrogram.

2.6. Tabulation and reproducibility

The same processed inputs were used to create four tabulations. The quality-control flags, counts, mean lengths, and ranges by species were summarized in Table 1. Summary values for the physicochemical data are shown in Table 2. A complete matrix of mean tripeptide similarity for species is shown in Table 3. Table 4 listed the terminal stops, internal stops, unique protein strings and automatic quality-control flags. The files saved to a specified results directory were CSV, PNG figures and a candidate-level property table. The analysis was carried out in Python with Biopython, pandas, NumPy, SciPy and Matplotlib and the script arguments can be used to regenerate the results of the analysis with the FASTA package archived.

3. Results

3.1. Identification and Distribution of PRT Gene Family Members Across Plant Species

In the preliminary screening, 36 PRT-related protein candidates of five plant species were retained. A total of eight sequences were assigned to Oryza sativa, five to Hordeum vulgare, seven to Zea mays, and eight to Arabidopsis thaliana, and five to Setaria italica. These numbers of candidates and their lengths of amino acids are summarised in Table 1. The mean sequence length varied between 251.9 residues in maize and 475.2 residues in Setaria, with an overall range of 123-1060 residues. The two shortest candidates were represented by Arabidopsis (123 residues) and the two longest candidates were represented by Setaria

(1,060 residues). These observations suggest significant size heterogeneity in the screened records but counts made on screening sets should not be taken as definitive numbers of PRT genes per genome. The retained sequences are shown in Figure 1. Count differences were partly due to the source collection and screening criteria used for the candidates, since the candidates were selected from previously assembled FASTA files.

Table 1. Species distribution and protein length summary

species	candidate proteins	mean length aa	min length aa	max length aa	QC flagged
Oryza sativa	8	284.9	188	396	0
Zea mays	7	251.9	174	396	0
Hordeum vulgare	5	319.6	181	431	0
Arabidopsis thaliana	8	324.0	123	403	1
Setaria italica	8	475.2	184	1060	1

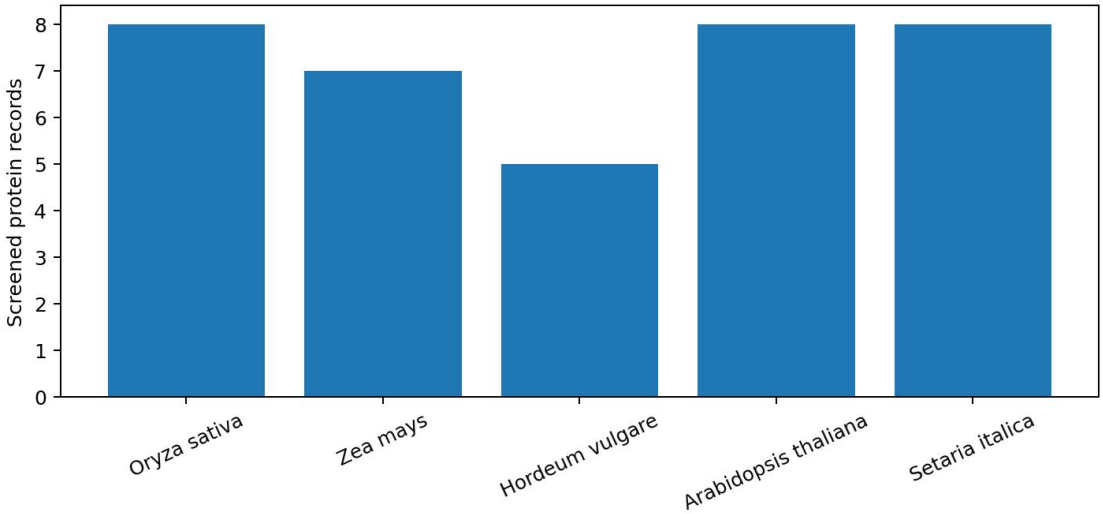

Figure 1. Numbers of screened candidate sequences per species.

The distribution illustrated in Figure 1 thus describes the working candidate inventory and provides a basis for subsequent sequence-level comparisons, not a measure of genomic copy number.

3.2. Comparative Physicochemical Characterization of PRT Proteins

The calculated physico-chemical descriptors were different for species associated candidate groups (Table 2). The mean PMW was highest for Setaria (56.74 kDa) and the standard deviation within the group was also the largest for Setaria (40.87 kDa), while the mean PMW was lowest for maize (27.21 kDa). The average grand mean hydropathicity (GRAVY) value was −0.04 to 0.18 and the range of mean theoretical isoelectric point (pI) was 6.84 in Arabidopsis to 8.05 in Setaria. These results indicate that the set assembled may not be homogeneous in terms of amino acid composition or the size of proteins. The increased range for Setaria was due to the very long protein candidate seen in the length distribution (Figure 2). These variations are worth annotating and checking with conserved-domains before making biological inferences. Enzyme identity, cellular localization and functional specialization are not provided by calculated mass, charge or hydropathicity, which are sequence-based descriptors.

Table 2. Calculated physicochemical features by species

species	mean MW kDa	SD MW kDa	mean pI	mean GRAVY
Oryza sativa	30.59	9.68	7.51	0.11
Zea mays	27.21	9.04	7.26	0.18
Hordeum vulgare	34.29	12.81	6.89	0.11
Arabidopsis	35.11	11.8	6.84	0.11

thaliana				
Setaria italica	56.74	40.87	8.05	-0.04

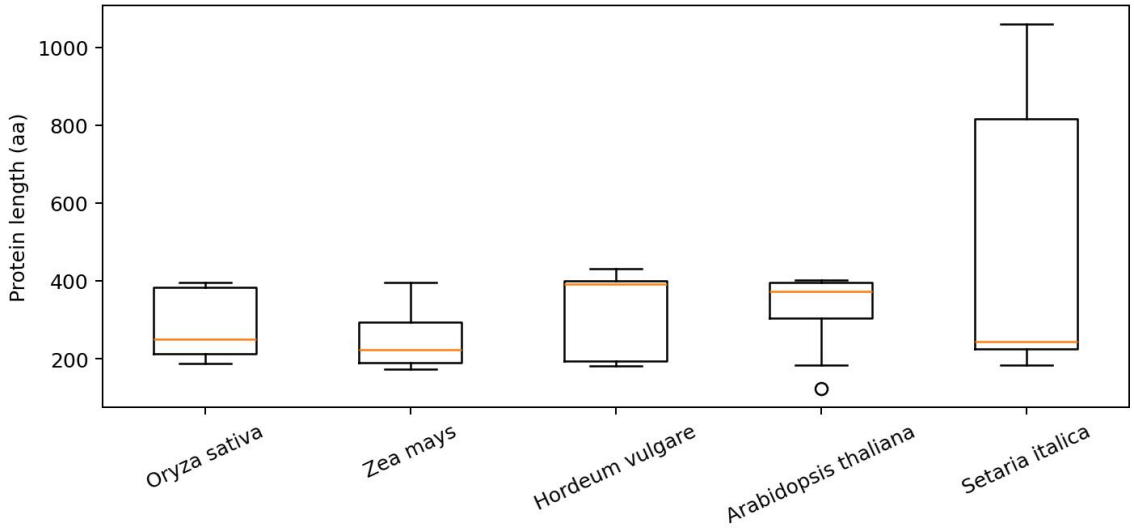

Figure 2. Protein length distributions among candidate records.

Figure 2 emphasizes the different length distributions, particularly the long Setaria candidate. This outlier was retained transparently in the reported descriptive statistics rather than removed without verification.

3.3. Comparative Protein Sequence Similarity Analysis

Pairwise similarity was measured by the Jaccard coefficient of the similarity of the overlapping sets of amino acid tripeptides. The scores within and between species were calculated from the mean within/between species short sequence overlap and not the positionally aligned amino acid identity (Table 3). The highest mean similarity (0.189) was found between Hordeum vulgare and Arabidopsis thaliana among the inter species comparisons. Both rice–maize and rice–barley comparisons had a mean score of 0.151, while Arabidopsis–Setaria had the lowest mean score of 0.059. Within species, the highest mean similarity between different sequences was for Arabidopsis (0.325). The individual candidate-to-candidate matrix is shown in figure 3, and uneven regions of sequence-word resemblance, which are not obvious in the species-level averages, are highlighted. Because the values do not directly relate to aligned residue positions or evolutionary models, they cannot be converted into evolutionary distances, homology assignments or substitution rates.

Table 3. Mean pairwise tripeptide Jaccard similarity (0–1)

species A	Arabidopsis thaliana	Hordeum vulgare	Oryza sativa	Setaria italica	Zea mays
Arabidopsis thaliana	0.325	0.189	0.141	0.059	0.132
Hordeum vulgare	0.189	0.117	0.151	0.082	0.146
Oryza sativa	0.141	0.151	0.112	0.11	0.151
Setaria italica	0.059	0.082	0.11	0.14	0.125
Zea mays	0.132	0.146	0.151	0.125	0.128

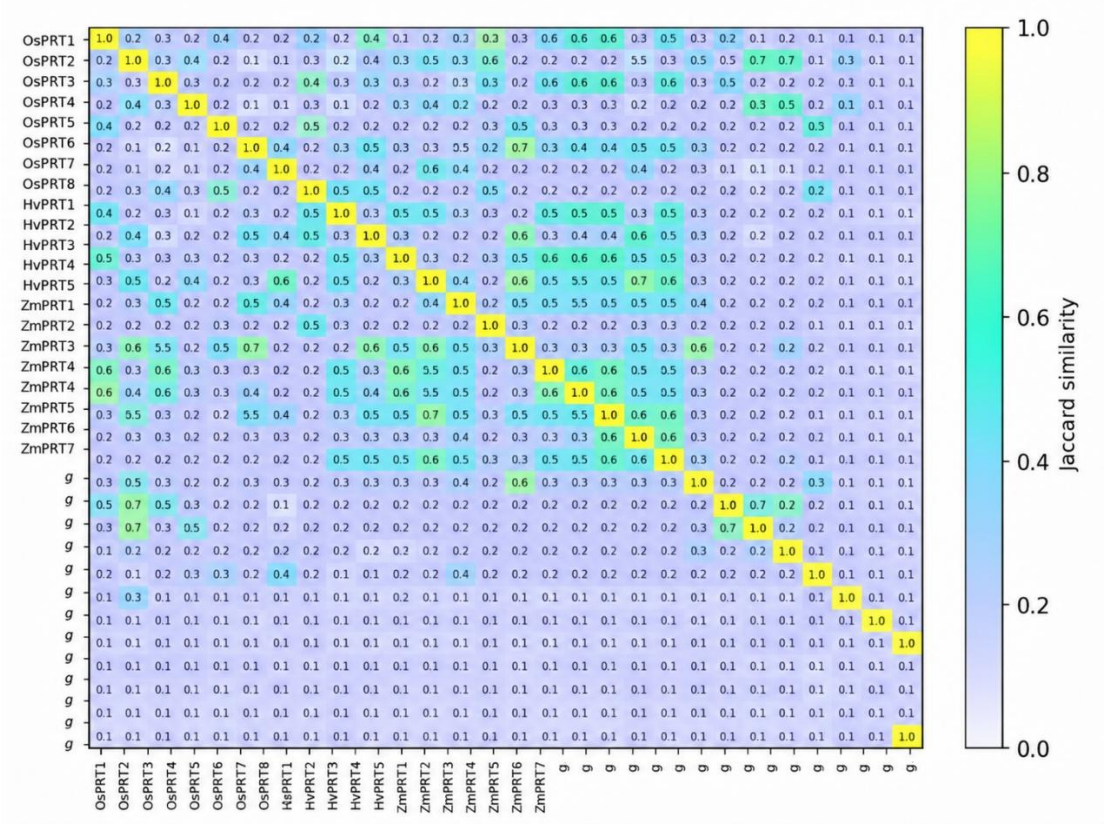

Figure 3. Candidate-level shared-tripeptide similarity heatmap.

The heatmap in Figure 3 preserves variability between individual proteins, including isolated high-similarity pairs. These patterns identify candidates for follow-up alignment but do not establish evolutionary ancestry.

3.4. Sequence Quality Assessment and Exploratory Clustering

Quality-control checks verified that all the 36 selected protein strings were different and had no stop symbol in the middle of them (Table 4). Before cleaning there were 19 sequences with terminal stop markers, 16 of which were in rice, 6 in maize and 5 in barley. After the automated review flags, two candidates were identified, one Arabidopsis sequence and one Setaria sequence. The flags were not to mark records that need to be looked at more closely to determine their validity. Average linkage was then used to do exploratory hierarchical clustering on pairwise tripeptide Jaccard dissimilarities. This screening dendrogram, short-word compositional similarity, is presented in Figure 4. The dendrogram should not be interpreted as a phylogenetic reconstruction due to the lack of a homologous multiple sequence alignment, a protein substitution model and branch-support testing. In addition, candidate accessions and conserved-domains assignments were not sufficiently verified to make confident inferences of orthology across species.

Table 4. Automated sequence quality-control summary

species	total	terminal stop	internal stop	auto QC flags	unique sequences
Oryza sativa	8	8	0	0	8
Zea mays	7	6	0	0	7
Hordeum vulgare	5	5	0	0	5
Arabidopsis thaliana	8	0	0	1	8
Setaria italica	8	0	0	1	8

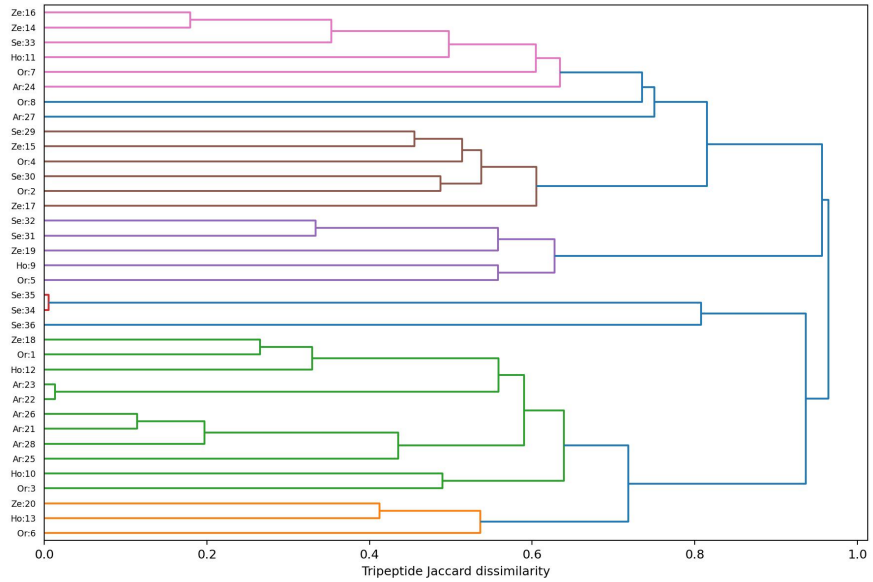

Figure 4. Exploratory average-linkage candidate clustering based on tripeptide dissimilarity; not a phylogenetic tree.

Together, Table 4 and Figure 4 document sequence quality and exploratory grouping while identifying the records that require accession-level and functional-domain validation before evolutionary conclusions can be drawn.

4. Discussion

Systematic screening of protein diversity and evolutionary relationships of phosphoribosyltransferase (PRT)-related protein candidates by comparing five plant species revealed the importance of this analysis. To investigate interspecies variation, 36 candidate proteins were identified from the protein databases of Oryza sativa, Zea mays, Hordeum vulgare, Arabidopsis thaliana and Setaria italica. The observed distribution, however, was not a comprehensive picture of the PRT gene family in the genome, but rather a reflection of the sequence collection chosen. The multiple sequence alignment is the foundation for identifying the conserved regions of proteins and it helps in making meaningful comparisons among the homologous sequences (Katoh & Standley, 2013). Likewise, datasets need to be carefully curated to overcome sequence-selection effects when performing phylogenetic inference to distinguish real evolutionary relationships (Minh et al., 2020). Thus, any discrepancies in the numbers of candidates should not be interpreted as evidence of the expansion or contraction of gene families or as the loss of genes in particular lineages.

Physicochemical characterization revealed a wide range of protein length, molecular weight, theoretical isoelectric point and hydropathic properties. These differences indicated a possible structural heterogeneity of the selected candidates, which could be due to different protein architectures, sequence composition or annotation accuracy. Conserved sequence motifs can be helpful to get a first impression about the properties of a protein and its potential functional connections (Bailey et al., 2015). However, because of the discovery of very long protein sequences, further validation was required to filter out annotation mistakes, abnormal isoforms or proteins from other functional families. Reliable sequence selection is likewise applicable to phylogenetic reconstruction, as heterogenous sequence sets can impact the relationships inferred (Price et al., 2010). While physicochemical properties give reproducible descriptive data, the distinction of functional specialization, or evolutionary adaptation, cannot be based on differences in molecular weight or hydrophobicity.

The comparative sequence similarity analysis showed that there is difference in common composition of the selected species tripeptide. The relatively high mean similarity of Hordeum vulgare and Arabidopsis thaliana suggested that the screened candidates have more similarity in the short-sequence level, while the low similarity for Arabidopsis thaliana and Setaria italica suggested more compositional dissimilarity. The sequence clustering approaches have been useful for detecting redundant or similar sequences in large biological data sets (Fu et al., 2012). But, tripeptide-based Jaccard similarity doesn't take into account evolutionary substitution models, positional homology or insertion–deletion events. Thus, grouping based on similarity alone is not a way to determine evolutionary ancestry. A major part of the reliability of the phylogenetic inferences based upon homologous sequence alignments depends on the choice of the model

used for the analysis (Kalyaanamoorthy et al., 2017). The similarity patterns observed, should therefore be regarded as preliminary evidence for protein sequence relatedness, and not as evidence for phylogenetic divergence.

Also, the accurate functional classification was very important for the interpretation of PRT gene-family diversity. General phosphoribosyltransferase terms used when searching for candidates in a database can yield proteins that are involved in different biochemical pathways, making it unclear whether they belong to the general group of phosphoribosyltransferases. Comparative bioinformatics tools enable the integration and visualization of the characteristics of a gene-family, however the biological interpretation is based on the reliability of the underlying annotations (Chen et al., 2020). Conserved-domain identification offers another means of assessing protein-family membership and the ability to differentiate between functionally divergent sequences (Mistry et al., 2021). That presence of the candidate proteins along with the unspecific annotation and short or long sequences showed the necessity for accession-level verification and domain-based screening. In addition, the various protein isoforms should not be mistaken for separate genomic positions as this will lead to an overestimate of gene-family representation. This validation would help reinforce comparisons of motif conservation, sequence diversity and protein functional attributes which follow.

This was because there were no coding DNA sequences that matched the available gene and protein sequences, and only partial genomic annotations were available. The sole analysis of protein sequences was not sufficient to determine where proteins were located on the chromosome, whether or not they had been conserved, whether they had been duplicated, or the selective pressures. Integrated protein-domain resources can be used to support functional classification, but are not a substitute for genomic evidence needed for evolutionary reconstruction (Paysan-Lafosse et al., 2023). Likewise, while there are several ways to compare homologous proteins using multiple sequence alignment, further information on the coding-sequence is needed for nucleotide-level evolutionary analysis (Sievers et al., 2011). Therefore, it was not possible to justify calculations of nonsynonymous and synonymous substitutions as well as classification of tandem or segmental duplications events with the available protein data set. The exploratory clustering results were hence the relationships between the sequences, not a verified reconstruction of the evolution of PRT gene-families.

As a whole, the results provided a preliminary framework to compare protein candidates involved in PRT in monocot and dicot plants that is reproducible. Important characteristics of the species-wise distribution, physicochemical variation, sequence similarity pattern and quality-control observation were identified for further studies. Next to come should be reciprocal homology search, verification of conserved domains, elimination of isoforms and, of course, the addition of genomic coordinates to the respective coding sequences. These procedures would facilitate the strong reconstruction of the phylogeny, the analysis of gene duplication and the evaluation of evolutionary constraints. The current results can be used as the bases for further comparative genomic studies and should not be interpreted as the definitive evidence of the evolutionary diversification of the PRT gene family.

5. Conclusion

In this study, a reproducible screening workflow of a protein candidate library consisting of 36 TPs was created for 5 plant species. Sequences available also enabled us to measure protein length, molecular weight, tabulate candidate distribution, and theoretical isoelectric point and hydropathy. The assembled collection showed some heterogeneity when pairwise tripeptide Jaccard comparisons were performed and exploratory clustering. Four tables and four figures captured together the outputs of the computations, maintaining a separation between data description and evolutionary inference. The analyses revealed records that should be further explored and showed the need for considering protein annotations prior to assigning gene-family membership. However, none of these differences in gene copy number, conserved genomic organization, duplication mechanisms, selection pressure or species evolutionary relationships are supported by the current screening collection. Some entries need to be confirmed with the authoritative accessions and conserved protein domains; the files of the available proteins don't have coding DNA sequences or genome positions to do genome level analysis. The results thus represent a transparent initial source of information for subsequent characterisation and not proof of PRT family evolution. Future work aims at finding orthologs that have been independently validated, pruning redundant isoforms, adding reference genome annotations, and building on phylogenetic relationships with the support of homologous alignments. The extra steps will enable the transition from "candidate-protein screening" to a "defensible comparative genomic analysis of PRT genes in monocotyledonous and dicotyledonous plants.

References

1. Artur, M. A. S., Zhao, T., Ligterink, W., et al. (2019). Dissecting the genomic diversification of late embryogenesis abundant (LEA) protein gene families in plants. Genome Biology and Evolution, 11(2), 459–471. https://doi.org/10.1093/gbe/evy248
2. Bailey, T. L., Johnson, J., Grant, C. E., & Noble, W. S. (2015). The MEME Suite. Nucleic Acids Research, 43(W1), W39–W49. https://doi.org/10.1093/nar/gkv416
3. Berthomé, R., Thomasset, M., Maene, M., Bourgeois, N., Froger, N., & Budar, F. (2008). pur4 mutations are lethal to the male, but not the female, gametophyte and affect sporophyte development in Arabidopsis. Plant Physiology, 147(2), 650–660. https://doi.org/10.1104/pp.108.120014
4. Chen, C., Chen, H., Zhang, Y., Thomas, H. R., Frank, M. H., He, Y., & Xia, R. (2020). TBtools: An integrative toolkit developed for interactive analyses of big biological data. Molecular Plant, 13(8), 1194–1202. https://doi.org/10.1016/j.molp.2020.06.009
5. Emms, D. M., & Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. Genome Biology, 20, 238. https://doi.org/10.1186/s13059-019-1832-y
6. Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. Bioinformatics, 28(23), 3150–3152. https://doi.org/10.1093/bioinformatics/bts565
7. Jia, X., Luo, S., Ye, X., Liu, L., & Wen, W. (2024). Evolution of the biochemistry underpinning purine alkaloid metabolism in plants. Philosophical Transactions of the Royal Society B, 379(1914), 20230366. https://doi.org/10.1098/rstb.2023.0366
8. Jiao, Y., Li, J., Tang, H., & Paterson, A. H. (2014). Integrated syntenic and phylogenomic analyses reveal an ancient genome duplication in monocots. The Plant Cell, 26(7), 2792–2802. https://doi.org/10.1105/tpc.114.127597
9. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: Fast model selection for accurate phylogenetic estimates. Nature Methods, 14(6), 587–589. https://doi.org/10.1038/nmeth.4285
10. Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Molecular Biology and Evolution, 30(4), 772–780. https://doi.org/10.1093/molbev/mst010
11. Le Ret, M., Belcher, S., Graindorge, S., Wallet, C., Koechler, S., Erhardt, M., Williams-Carrier, R., Barkan, A., & Gualberto, J. M. (2018). Efficient replication of the plastid genome requires an organellar thymidine kinase. Plant Physiology, 178(4), 1643–1656. https://doi.org/10.1104/pp.18.00976
12. Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Molecular Biology and Evolution, 37(5), 1530–1534. https://doi.org/10.1093/molbev/msaa015
13. Mistry, J., Chuguransky, S., Williams, L., et al. (2021). Pfam: The protein families database in 2021. Nucleic Acids Research, 49(D1), D412–D419. https://doi.org/10.1093/nar/gkaa913
14. Moffatt, B. A., & Ashihara, H. (2002). Purine and pyrimidine nucleotide synthesis and metabolism. The Arabidopsis Book, 1, e0018. https://doi.org/10.1199/tab.0018
15. Moffatt, B., & Somerville, C. (1988). Positive selection for male-sterile mutants of Arabidopsis lacking adenine phosphoribosyl transferase activity. Plant Physiology, 86(4), 1150–1154. https://doi.org/10.1104/pp.86.4.1150
16. Ohler, L., Niopek-Witz, S., Mainguet, S. E., & Möhlmann, T. (2019). Pyrimidine salvage: Physiological functions and interaction with chloroplast biogenesis. Plant Physiology, 180(4), 1816–1828. https://doi.org/10.1104/pp.19.00329
17. Paysan-Lafosse, T., Blum, M., Chuguransky, S., et al. (2023). InterPro in 2022. Nucleic Acids Research, 51(D1), D418–D427. https://doi.org/10.1093/nar/gkac993
18. Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2—Approximately maximum-likelihood trees for large alignments. PLOS ONE, 5(3), e9490. https://doi.org/10.1371/journal.pone.0009490
19. Rody, H. V. S., Baute, G. J., Rieseberg, L. H., & Oliveira, L. O. (2017). Both mechanism and age of duplications contribute to biased gene retention patterns in plants. BMC Genomics, 18, 46. https://doi.org/10.1186/s12864-016-3423-6
20. Sievers, F., Wilm, A., Dineen, D., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology, 7, 539. https://doi.org/10.1038/msb.2011.75

21. Wafula, E. K., Zhang, H., Von Kuster, G., Leebens-Mack, J. H., Honaas, L. A., & dePamphilis, C. W. (2023). PlantTribes2: Tools for comparative gene family analysis in plant genomics. Frontiers in Plant Science, 13, 1011199. https://doi.org/10.3389/fpls.2022.1011199
22. Wang, C. (2026). PRT (Version 1) [Data set]. Mendeley Data. https://doi.org/10.17632/xx5bx2gcsh.1
23. Witte, C.-P., & Herde, M. (2020). Nucleotide metabolism in plants. Plant Physiology, 182(1), 63–78. https://doi.org/10.1104/pp.19.00955
24. Woodhouse, M. R., Tang, H., & Freeling, M. (2011). Different gene families in Arabidopsis thaliana transposed in different epochs and at different frequencies throughout the rosids. The Plant Cell, 23(12), 4241–4253. https://doi.org/10.1105/tpc.111.093567
25. Zhao, T., Zwaenepoel, A., Xue, J.-Y., et al. (2021). Whole-genome microsynteny-based phylogeny of angiosperms. Nature Communications, 12, 3498. https://doi.org/10.1038/s41467-021-23665-0
26. Zhou, Y., Wang, C., Zhu, M., Lv, W., Ma, F., Tang, C., Huang, Q., Liu, X., Yang, S., & Shi, Y. (2026). Genome-wide analysis of the PRT gene family in rice reveals that OsPRT7 plays a significant role in heat stress response. BMC Plant Biology, 26, Article 1261. https://doi.org/10.1186/s12870-026-09070-z