The Andhra Agric. J 72 (3): 310-326, 2025
doi :10.61657/aaj.2025.214

In-silico structural and functional analysis of cysteine proteinase inhibitors in legumes

R Aruna, A Srividhya, M Lal Ahamed, K Devaki and P Latha
Department of Molecular Biology and Biotechnology, Acharya N G Ranga Agricultural University, S V Agricultural College, Tirupati, Andhra Pradesh, India

ABSTRACT

Cysteine Proteinase Inhibitors (CPI) play critical role in plant defense and protease regulation, yet their evolutionary and functional diversity across legumes remains incompletely understood. In this study, we performed a comprehensive sequence, structural, and biochemical analysis of CPI protein from eight legume species *viz, Arachis hypogaea, Arachis duranensis, Cajanus cajan, Cicer arietinum, Vigna radiata, Pisum sativum, Glycine max, and Phaseolus vulgaris*, with *Medicago sativa* as a divergent outlier. Multiple sequence alignment across legume species revealed strict conservation of core inhibitory motifs and cysteine residues forming disulfide bonds, whereas N-terminal signal peptides and reactive loops exhibited species-specific variability, reflecting functional adaptation. Phylogenetic and biochemical property analysis identified two primary evolutionary clusters and highlighted *Medicago* as a structurally and chemically distinct CPI variant. Structural predictions confirmed a conserved â-trefoil–like fold, with flexible loops enabling dynamic protease interactions. Physicochemical profiling demonstrated a balance between hydrophilicity, stability and thermostability, while correlation analysis revealed an inverse relationship between hydrophobicity and conformational flexibility. Collectively, these findings indicate that legume CPIs maintain a conserved structural and functional core while evolving adaptive features to support diverse role in plant defense, stress responses and protein homeostasis. Furthermore, potential CPIs in *A. hypogaea* were analyzed based on available expression databases. These genes can serve as potential candidates for developing biotic stress resistance, particularly against bruchid infestation during legume storage, with specific relevance to peanut as a major food and oil crop.

Keywords: *Biochemical properties, Cysteine Proteinase Inhibitor, legumes, Phylogenetics, Sequence conservation and Structural analysis*

Plants have evolved a sophisticated and dynamic defense network to counteract both biotic (insects, pathogens) and abiotic (temperature, drought, salinity) stresses. Over millions of years, this interaction has taken the form of an evolutionary arms race, where plants continuously develop new molecular strategies to survive attacks, while herbivores and pathogens evolve corresponding counter measures. When attacked by insects or pathogens, plants use different defense mechanisms that involves calcium-signaling, production of reactive oxygen species (ROS), phytohormones, miRNAs, secondary metabolites, and defense proteins (Zhang *et al*. 2014). Among these, the production of protease inhibitors is one such mechanism that contributes to plant defense by severely reducing the growth and development of phytopathogens (Habib and Fazili, 2007). Upon perception of an attack, plants activate multiple defense signaling pathways involving calcium fluxes, ROS production, and phytohormonal regulation (including jasmonic acid, salicylic acid, and ethylene). These responses are accompanied by miRNA modulation, synthesis of secondary metabolites, and production of defense-related proteins such as protease inhibitors.

Protease inhibitors (PIs) represent a key biochemical weapon, restricting the pathogenic proteases of invading organisms and thus limiting their growth and virulence. Proteases, although vital for normal plant functions such as protein turnover, developmental regulation and signaling, must be tightly regulated to prevent self-damage. This regulation is achieved by specific protease inhibitors, which act as molecular "safeguards" to ensure proteolytic balance. As highlighted by Polya (2003), plant protease inhibitors encompass a wide range of small proteins, peptides, and non-protein derivatives involved not only in plant defense, but also in growth, storage

protein mobilization, and developmental transitions. Likewise, Van der Hoorn and Kamoun (2008) emphasized that plant proteases and their inhibitors constitute a complex regulatory network, where inhibition, activation, and compartmentalization work together to maintain proteostasis, while ensuring an efficient response to external stress including biotic stress factors.

Cysteine proteinase inhibitors (CPIs), commonly termed phytocystatins, constitute an important class of regulatory proteins in plants. These inhibitors specifically target cysteine proteases such as papain-like enzymes, maintaining cellular proteostasis and contributing to plant defense. Early foundational work by Abe and Arai (1985) identified cysteine protease inhibitors in rice seeds, demonstrating their ability to block proteolytic activity, suggesting a protective role in seed tissues. Subsequent studies expanded this understanding, showing that phytocystatins are structurally conserved proteins characterized by motifs such as GG and QxVxG, which are essential for interaction with target proteases (Abe and Arai, 1985; Kondo *et al*. 1990).

A significant body of literature describes the role of CPIs in preventing unwanted proteolysis during seed development, maturation, and storage. Kondo *et al*. (1990) reported that oryzacystatin I, isolated from rice seeds, inhibits endogenous cysteine proteases, reducing uncontrolled protein degradation. This regulatory function is particularly crucial during seed desiccation and storage, where excessive proteolysis can compromise seed viability and storage protein integrity. Girard *et al*. (2007) provided strong experimental evidence that overexpression of oryzacystatin in seeds reduces proteolytic activity and enhances storage stability, confirming the functional significance of CPIs in post-harvest protection.

Phytocystatins also play a dual role in defense against storage pests like pulse bruchids. Seeds, as nutrient-rich organs, are vulnerable to attack by insects whose gut digestion relies heavily on cysteine proteases. Phytocystatins inhibit these insect proteases, thereby impairing digestion and reducing insect survival and growth. Benchabane *et al*. (2010) and Valueva and Mosolov (1999) demonstrated that CPIs serve as natural defense molecules, accumulating in seeds and storage tissues to deter herbivory. Transgenic studies further support this role: expression of phytocystatins in crops such as tobacco and rice significantly impairs the performance of phytophagous insects, highlighting their biotechnological potential in enhancing post-harvest pest resistance. Hence, current study is focused to analyse CPIs of groundnut and their conservation relationship across legumes.

MATERIAL AND METHODS

Sequence Data Collection and Curation

Cysteine proteinase inhibitor (CPI) protein sequences were retrieved from the UniProt Knowledgebase (UniProtKB). The query terms "Cysteine proteinase inhibitor" and "Legume" was applied, and the search was restricted to reviewed (Swiss-Prot) entries to ensure high-quality, manually curated data. To focus on representative legumes, taxonomic filters were applied for the following species: *Cajanus cajan* (pigeon pea), *Cicer arietinum* (chickpea), *Vigna radiata* (mung bean), *Arachis hypogaea* (peanut), *Arachis duranensis*, *Pisum sativum* (pea), *Glycine max* (soybean), *Phaseolus vulgaris* (common bean), and *Medicago sativa* (alfalfa). All retrieved entries were downloaded in FASTA format. The curated dataset was subsequently used as input for multiple sequence alignment, motif characterization and comparative phylogenetic analyses. In *Arachis hypogaea*, putative cysteine proteinase inhibitor (CPI) genes were mined from the Peanut Base expression atlas. Gene models were filtered based on the presence of conserved cystatin domains, specifically IPR000010 (Cystatin domain) and IPR027214 (Cystatin super family), as annotated by InterPro.

Multiple Sequence Alignment and Conservation Analysis

Alignment Methodology

Multiple sequence alignment (MSA) of the cysteine proteinase inhibitor sequences across legumes was performed using three complementary algorithms to ensure robustness and consistency in the identification of conserved residues. Alignments were first generated with MAFFT v7.505, employing the L-INS-i algorithm, which performs iterative refinement and is optimized for accuracy in sequences with conserved core regions and variable termini. The second alignment was conducted using Clustal Omega v1.2.4 with default parameters, providing a fast and scalable approach suitable for assessing global sequence similarity. Finally, MUSCLE v5.1 was used

with a maximum of 100 refinement iterations, enabling fine-tuning of residue positioning across conserved motifs. The outputs from all three tools were compared, and consensus alignment regions were retained for subsequent analyses, including motif discovery, conserved residue mapping and phylogenetic reconstruction.

Consensus Generation

To derive a representative alignment reflecting conserved residues across all species, a consensus sequence was generated using standardized parameters. The alignment utilized the BLOSUM62 substitution matrix, which provides optimal scoring for amino acid substitutions among moderately divergent sequences. Gap penalties were set to a gap opening penalty of 10 and a gap extension penalty of 0.5, ensuring a balance between alignment sensitivity and the avoidance of excessive gap insertion in variable regions. A consensus threshold of 70% was applied, meaning that a residue was included in the final consensus sequence only if it appeared in at least 70% of the aligned sequences at that position. This approach facilitated the identification of high-confidence conserved motifs, such as cysteine residues forming disulfide bridges and key residues in the reactive site loop, while maintaining alignment integrity across structurally variable termini.

Conservation Scoring

To quantify positional conservation across the aligned cysteine proteinase inhibitor sequences, multiple complementary scoring approaches were applied. Scorecons, an information-theoretic conservation algorithm, was first used to evaluate the degree of residue conservation based on amino acid frequency and variability across alignment columns, providing a normalized conservation index ranging from 0 (variable) to 1 (fully conserved). In parallel, Shannon entropy was calculated at each alignment position to measure sequence uncertainty, lower entropy values indicating strong conservation and higher values denoting flexibility or divergence. Additionally, evolutionary rate analysis was conducted using PAML codeml, which estimated nonsynonymous-to-synonymous substitution ratios (dN/dS) to infer the selective pressures acting on each codon site, distinguishing purifying selection from adaptive diversification. Together, these methods provided a detailed quantitative framework for identifying structurally constrained and functionally critical residues across legume CPIs. The computational alignment and scoring procedures followed the methodology of Katohand Standley (2013), ensuring consistency with established molecular evolution standards.

Phylogenetic Analysis

Tree Construction Methods

Phylogenetic relationships among cysteine proteinase inhibitor sequences were inferred using both maximum likelihood (ML) and Bayesian inference (BI) frameworks to ensure robust evolutionary resolution. For the ML analyses, IQ-TREE v2.2.0 was employed, incorporating Model Finder for automated substitution model selection based on the Bayesian Information Criterion (BIC). The optimal model identified was JTT+G+I, which accounts for amino acid substitution patterns, among-site rate heterogeneity (Ã-distributed), and invariant sites. Branch support was assessed using 1,000 ultrafast bootstrap replicates to provide high-confidence estimates of phylogenetic topology stability.

Complementarily, BI analyses were performed in MrBayes v3.2.7, employing four independent Markov chain Monte Carlo (MCMC) runs for 10 million generations, with trees sampled every 1,000 generations and a 25% burn-in applied to remove pre-convergence data. Convergence was confirmed by an average standard deviation of split frequencies below 0.01, indicating a well-sampled posterior distribution. Together, these dual approaches produced a statistically supported phylogeny that captured both conserved lineage clustering and species-specific diversification within the legume cysteine proteinase inhibitor family.

Distance Matrix Calculation

Evolutionary distance estimation among cysteine proteinase inhibitor sequences was performed to quantify pairwise divergence and infer relative evolutionary rates prior to phylogenetic reconstruction. Pairwise p-distances were first computed to measure the proportion of amino acid differences between sequences, providing a direct estimate of observed divergence. To account for potential multiple substitutions at the same site, the Jukes–Cantor correction was applied, yielding more

accurate evolutionary distance values for moderately to highly diverged pairs. In addition to sequence-based metrics, structural distance measures were derived from predicted physicochemical and secondary structural features, enabling correlation between sequence evolution and structural conservation. These combined approaches ensured a comprehensive representation of molecular divergence across the CPI dataset. The computational framework and evolutionary modeling procedures followed the methodology described by Nguyen *et al*. (2015), ensuring robust and reproducible estimation of sequence distances within the IQ-TREE phylogenetic analysis environment.

Structural Analysis and Domain Architecture Domain Prediction

The domain architecture of CPIs was characterized through an integrated bioinformatics workflow combining multiple domain-recognition tools to ensure accuracy and coverage. InterProScan v5.63-95.0 was used as a comprehensive meta-annotation platform, integrating data from diverse domain databases to assign functional and structural regions. Specific detection of cystatin-type inhibitory domains (Pfam ID: PF00031) was performed using Pfam v35.0, which confirmed the presence of the conserved phytocystatin fold characteristic of legume CPIs. To refine these annotations, SMART v9.0 was employed for functional domain prediction, enabling the identification of signal peptides, low-complexity regions, and linker extensions. Additionally, CDD v3.19 (Conserved Domain Database) was utilized to recognize evolutionarily conserved motifs and classify the inhibitors according to their functional domain architectures. Together, these analyses provided a detailed map of core inhibitory modules, disulfide-rich scaffolds, and variable regions potentially involved in species-specific adaptations.

Secondary Structure Prediction

To assess the structural organization underlying sequence conservation, secondary structure prediction was carried out using three complementary algorithms. PSIPRED v4.0, based on neural network–driven analysis of position-specific scoring matrices, provided accurate predictions of α-helices, β-strands, and coil regions within the inhibitory domain. Jpred v5.0, a consensus meta-predictor integrating multiple sequence-based algorithms, was used to validate PSIPRED outputs and minimize false positives in loop or coil assignments. Additionally, the DSSP algorithm (Dictionary of Secondary Structure Assignments) was applied to the modeled tertiary structures to extract experimentally consistent α/β annotations. Collectively, these analyses confirmed the α-sheet-rich core typical of cystatin inhibitors, flanked by short α-helical and loop regions that contribute to flexibility in reactive site presentation.

Tertiary Structure Modeling

Three-dimensional structural modeling of representative CPI sequences was conducted using the SWISS-MODEL server, which employs template-based homology modeling to construct reliable protein structures. Template selection was based on structural homologs identified in the Protein Data Bank (PDB), including 1STF (stefin B–papain complex), 3C4L (rice cystatin), and 6X2A (soybean cystatin), representing high-resolution reference structures of cystatin-type inhibitors. The resulting models were rigorously evaluated using QMEAN for overall model quality, MolProbity for geometric validation and clash analysis, and PROCHECK for stereochemical assessment via Ramachandran plots. These validation steps ensured structural fidelity, confirming that the modeled CPIs exhibited the canonical cystatin fold, a five-stranded α-sheet wrapped around an α-helix, stabilized by conserved disulfide bonds and a reactive loop region poised for protease interaction (Waterhouse *et al.* 2018).

Biochemical Property Calculations
Molecular weight calculation

Molecular weights were computed from amino-acid sequences using monoisotopic residue masses (IUPAC standards)—A: 89.0935, C: 121.1590, D: 133.1032, E: 147.1299, F: 165.1900, G: 75.0669, H: 155.1552, I/L: 131.1736, K: 146.1882, M: 149.2124, N: 132.1184, P: 115.1310, Q: 146.1451, R: 174.2017, S: 105.0930, T: 119.1197, V: 117.1469, W: 204.2262, Y: 181.1894;by summing the per-residue masses across each sequence and applying the standard polymerization correction (subtracting 18.01056 Da for each of the (n -1) peptide bonds waters lost for a chain of n residues, or equivalently summing residue masses and adding 18.01056 Da for the capped

termini). This corresponds to the logic of the calculate molecular weight (sequence) routine shown, using the listed IUPAC monoisotopic mass table.

Instability index

Protein stability was estimated using the Guruprasad *et al*. (1990) dipeptide-based method, which evaluates the frequency of all 400 possible dipeptide combinations in a sequence and sums their contributions using a precomputed instability weight matrix. The resulting score, normalized to sequence length yields the instability index, where values > 40 classify proteins as unstable (predicted shorter in vivo half-life) and values d ≤ 40 as stable. This approach captures how specific nearest-neighbor residue pairs correlate with empirical stability trends, providing a fast, sequence-derived proxy for comparative protein robustness.

Aliphatic index

Thermostability was approximated using the aliphatic index (AI), defined as the mole-percent–weighted contribution of key aliphatic residues.

$$AI=X(Ala)+a\times X(Val)+b\times[X(Iso)+X(Leu)]$$

with X(Å") the mole percent of each residue in the protein and constants a=2.9a and b=3.9b reflecting the relative side-chain volumes of valine and isoleucine/ leucine versus alanine.

$$AI = X(A) + 2.9 \times X(V) + 3.9 \times [X(I) + X(L)]$$

Grand Average of Hydropathy (GRAVY)

Hydropathic profiles were calculated according to the Kyte and Doolittle (1982) hydropathy scale, which assigns standardized numerical values to each amino acid based on its hydrophobic or hydrophilic character. The GRAVY score for each protein was obtained by averaging these residue-specific values over the entire sequence. A window size of 9 residues was applied to smooth local fluctuations and reveal broader hydropathy trends, distinguishing hydrophobic core regions from hydrophilic, solvent-exposed surfaces. Positive GRAVY values indicate overall hydrophobic (membrane-associated or buried) character, whereas negative values denote hydrophilic (soluble or extracellular) proteins. This analysis provided a quantitative measure of protein polarity and solubility, complementing stability and aliphatic index evaluations in characterizing cysteine proteinase inhibitors.

Theoretical pI Calculation

The isoelectric point (pI) of each cysteine proteinase inhibitor was estimated using the Bjellqvist method, which applies an iterative refinement algorithm to determine the pH at which the net charge of the protein equals zero. This method utilizes updated theoretical pKa values for all ionizable groups, including the N- and C-termini as well as side chains of acidic (Asp, Glu) and basic (Lys, Arg, His) residues, providing improved accuracy over older empirical approaches. The computational procedure employed a binary search between pH 3 and 12, incrementally adjusting the trial pH until electrostatic neutrality was achieved within a defined convergence threshold. The resulting pI values reflect the predicted charge state of each inhibitor under physiological conditions, facilitating comparisons of solubility, stability, and functional adaptability among species. The analytical framework followed the standardized protocol described by Guruprasad *et al*. (1990) in Protein Engineering, Design and Selection.

Statistical Methods

Comprehensive statistical analyses were performed to evaluate relationships among the biochemical, structural, and evolutionary parameters. Correlation analyses were conducted using Pearson's correlation coefficient (for linear relationships) with Bonferroni correction applied to adjust for multiple testing and control the family-wise error rate.

RESULTS AND DISCUSSION

Candidate cysteine-proteinase inhibitor genes identified in *Arachis hypogaea*

A total of fifteen candidate cysteine-proteinase inhibitor (CPI; cystatin) genes were identified in *Arachis hypogaea* through a bioinformatic screen of the Peanut Base expression atlas, selected based on annotations for conserved cystatin domains (IPR000010 / IPR027214). The details of the genes are given inTable.1.Expression profiling revealed distinct tissue-specific patterns, suggesting potential functional diversification within reproductive and developmental pathways. These genes exhibit distinct yet overlapping tissue-specific expression patterns, reflecting potential functional specialization within developing and reproductive organs.

The majority of the entries, including Arahy.N3F7NN, Arahy.PWIR9U, Arahy.PELT1K,

Table 1. List of candidate cysteine-proteinase inhibitor (CPI) genes identified in *Arachis hypogaea*

S.No.	Gene	Expression Details
1	Arahy.N3F7NN	Pod, Seed
2	Arahy.PWIR9U	Pod, Seed
3	Arahy.PELT1K	Pod, Seed
4	Arahy.QSG82C	Pod, Seed
5	Arahy.CTV7SW	Peanut Base expression pages
6	Arahy.MLA72C	Pod, Seed
7	Arahy.ZL1DA1	Pod, Seed
8	Arahy.U5205M	cystatin InterPro annotation
9	Arahy.L1FG9J	Reproductive shoot tip, seed, Pod
10	Arahy.WI67VP	Reproductive shoot tip, Pod, Seed
11	Arahy.3CP281	Both vegetative and reproductive tissues
12	Arahy.E6VW7Q	Pod, Seed
13	Arahy.QQM5EY	Nodule, root, pericarp
14	Arahy.D8HQTE	Shoot tip, seed
15	Arahy.4LFN69	Nodule, root, pericarp

Arahy.QSG82C, Arahy.MLA72C, Arahy.ZL1DA1, and Arahy.E6VW7Q show predominant expression in pod and seed tissues, suggesting major roles in seed maturation, storage protein protection, and defense against seed-invading pests and pathogens. Additional members such as Arahy.L1FG9J and Arahy.WI67VP are expressed across reproductive shoot tips, pods, and seeds, indicating broader reproductive relevance. In contrast, Arahy.3CP281 exhibits expression in both vegetative and reproductive tissues, consistent with a more housekeeping-like or constitutive defensive role. Two genes, Arahy.QQM5EY and Arahy.4LFN69 are enriched in nodules, roots, and pericarp, reflecting potential involvement in rhizosphere-associated defense or nodule homeostasis. The presence of Arahy.U5205M with direct InterPro cystatin annotation further supports its functional identity as a bona fide cysteine-proteinase inhibitor.

Collectively, the CPI genes expressed in pods, seeds, and reproductive tissues represent particularly promising targets for genetic improvement, as their spatial expression aligns with critical stages of seed development and key interfaces of pathogen or insect challenge. Such candidates could be prioritized for functional validation and introgression into breeding programs aimed at developing biotic stress–tolerant groundnut varieties.

Further the comparative in-silico analysis of legume cysteine proteinase inhibitors (CPIs) provides a strong evolutionary and functional context for interpreting the CPI genes identified in *Arachis hypogaea*.

Comparative In-Silico Analysis of CPIs across Legumes

Sequence Data Collection and Curation

A comprehensive search of the UniProtKB database highlights the extensive representation of *Cysteine Proteinase Inhibitors (CPIs)* across the plant kingdom. The term "Cysteine proteinase inhibitor" yields 16,834 entries in UniProtKB, reflecting the widespread occurrence and functional diversity of these inhibitors across multiple taxa. Narrowing the search to "Cysteine proteinase inhibitor [plants]" refines the dataset to 9,163 entries, emphasizing their predominant role in plant defense, development, and stress adaptation. Within the Fabaceae family alone, UniProtKB lists 1,251 entries, underscoring the evolutionary expansion and specialization of CPI gene families in legumes. This rich representation within Fabaceae supports the selection of legume CPIs as an ideal model for exploring structural diversity, evolutionary adaptation, and functional innovation within this critical class of protease inhibitors.

Set of *CPI* sequences (Table.2), which were carefully curated were selected to provide a

comprehensive and scientifically balanced dataset for comparative and functional analyses. The chosen sequences ensure phylogenetic breadth, representing major clades within the Fabaceae family, thereby capturing evolutionary diversity across both cultivated and wild legume lineages. Their agricultural relevance is underscored by the inclusion of key crop species such as *Arachis hypogaea*, *Glycine max*, *Cicer arietinum*, and *Phaseolus vulgaris*, which are globally significant for food security, nutrition, and sustainable agriculture. The dataset also enhances research utility through the inclusion of model legumes like *Pisum sativum* and *Medicago sativa*, which are widely used in molecular and genomic studies. All selected entries exhibit high data quality, derived from reliable annotations and validated databases, ensuring accuracy in downstream analyses. Moreover, the set provides strong comparative power, balancing evolutionary divergence with conserved functional motifs to enable robust cross-species inferences. Collectively, this selection possesses substantial biotechnological potential, offering insights that can be directly applied to crop improvement, particularly in enhancing stress tolerance and pathogen resistance.

Table2. UniProt annotation summary of cysteine proteinase inhibitors (CPIs) from representative legume species

Accession ID	Organism	Taxon ID (OX)	Gene Name (GN)
A0A151RC35_CAJCA	*Cajanus cajan*	3821	KK1_038471
A0A1S2YWL5_CICAR	*Cicer arietinum*	3827	LOC101512904
A0A445C0E0_ARAHY	*Arachis hypogaea*	3818	Ahy_A08g040602
A0A6P4B6I5_ARADU	*Arachis duranensis*	130453	LOC107462920
I1M0K3_SOYBN	*Glycine max*	3847	GLYMA_13G189500
A0A1S3UGN7_VIGRR	*Vigna radiata var. radiata*	3916	LOC106765170
A0A9D4XGY3_PEA	*Pisum sativum*	3888	KIW84_044763
Q1KK73_MEDSA	*Medicago sativa*	3879	CP1
V7AK72_PHAVU	*Phaseolus vulgaris*	3885	PHAVU_011G196600g

Multiple Sequence Alignment and Conservation Analysis

The sequence analysis of legume CPIs highlights both the strong conservation of core inhibitory elements and the divergent adaptations that reflect species-specific functional evolution. The inhibitor domain, which harbors cysteine proteinase inhibitory activity, is highly conserved across all sequences, ensuring that essential active site residues involved in protease inhibition remain intact. Likewise, the characteristic cysteine pattern forming disulfide bridges is consistently preserved, maintaining structural stability and the canonical inhibitory fold. Conversely, variability is most pronounced in the N-terminal signal peptides, which differ in both length and amino acid composition, indicating their roles in subcellular localization and secretion. Flanking terminal regions also show sequence diversity, with certain species possessing additional extensions that may influence targeting or regulatory function.

The sequence conservation profile derived from the CPI multiple sequence alignment offers a quantitative visualization of how each residue position is preserved across the analyzed homologs. Along the x-axis, the alignment positions correspond to residue indices, while the **y**-axis represents the conservation score ranging from 0 to 1 (Fig. 1A). Peaks approaching 1.0-mark regions of high conservation, typically corresponding to structural cysteines that form the disulfide scaffold or to catalytic motifs that define the inhibitor's reactive functionality. These conserved residues are critical for maintaining the canonical fold and inhibitory activity characteristic of CPI proteins.

The conservation map for the CPI multiple sequence alignment integrates both residue conservation and functional annotation across the sequences (Fig. 1B). The gray bars indicate per-position conservation scores (ranging from 0 to 1), where taller bars represent highly conserved residues. Over these, colored points identify consensus residue types: yellow marks cysteines (C) forming disulfide cores that stabilize the inhibitor's structure; orange denotes basic residues (K, R, H) located primarily within reactive loops responsible for protease binding;

red indicates acidic residues (D, E) acting as electrostatic partners or contributing to specificity; green highlights polar residues (S, T, N, Q) often involved in hydrogen bonding and surface stabilization; and white represents hydrophobic residues (A, V, I, L, M, F, Y, W) that form the buried structural core.

The map also features shaded regions representing major functional zones: the blue-shaded area corresponds to the signal peptide responsible for secretion targeting, the green-shaded zone indicates the reactive loop that drives inhibitory activity, and the pink-shaded region marks the inhibitor core, which forms the conserved fold scaffold characteristic of the CPI family (Fig. 1C). Together, these color-coded elements provide a clear depiction of how structure, function, and conservation are spatially organized along the sequence.

Phylogenetic Analysis

The neighbor-joining (NJ) phylogenetic tree (Fig.2) illustrates the evolutionary relationships among CPI protein sequences. Phylogenetically, the sequences cluster into three primary groups: Group 1, containing *Arachis hypogaea* (A0A445C0E0) and *Arachis duranensis* (A0A6P4B6I5), which are nearly identical and reflect close evolutionary proximity between cultivated and wild peanut species; Group 2, comprising *Vigna radiata* (A0A1S3UGN7) and *Phaseolus vulgaris* (V7AK72), representing a distinct Phaseolus–Vigna cluster; and Group 3, including *Cicer arietinum* (A0A1S2YWL5) and *Pisum sativum* (A0A9D4XGY3), which form a Cicer–Pisum lineage. An outlier, *Medicago sativa* (Q1KK73), diverges markedly in both sequence length and domain organization, possibly representing a different CPI subclass with unique structural or regulatory features.

Distance Matrix Calculation

The evolutionary distance matrix derived from normalized biochemical parameters (Fig.3), *viz*. molecular weight, instability index, aliphatic index, GRAVY (hydropathy), and isoelectric point (pI), for CPI proteins across eight legume species. The color intensity represents the degree of biochemical divergence, with darker tones indicating higher similarity (shorter distances) and lighter or yellow tones denoting greater evolutionary separation.

Key observations include, *Cajanus–Vigna* (4.2) and *Cicer–Arachis* (2.0) pairs exhibit the lowest biochemical distances, suggesting strong conservation of core physicochemical traits and a shared structural lineage; *Phaseolus–Vigna* (2.5) also reflects a close relationship, consistent with their similar instability and aliphatic indices; Medicago consistently shows the largest distances (30–43 range) from all other species, confirming its role as a biochemical and structural outlier, likely representing an ancestral or divergent CPI form with distinctive molecular weight, acidity, and stability features; *Pisum* maintains moderate distances (10–27 range), placing it intermediate between stable and unstable CPI clusters, reinforcing its characterization as a balanced, robust inhibitor type.

Overall, this distance map reveals two primary evolutionary clusters within legume CPIs: A core conserved cluster (*Cajanus, Cicer, Vigna, Arachis, Glycine, Phaseolus*) sharing similar stability and hydropathy profiles; a divergent lineage represented by *Medicago*, and to a lesser extent *Pisum*, indicating evolutionary adaptation through domain expansion or physicochemical optimization.

Structural Analysis and Domain Architecture Domain Prediction

The Simplified Domain Architecture Schematic (Fig. 1C) provides a concise visualization of how the CPI sequence is organized functionally along its length. The signal peptide occupies the N-terminal region, positioned before the main structural framework, and is responsible for directing the nascent peptide to the secretory pathway, consistent with the extracellular function of CPIs as protease inhibitors. Following this, the reactive loop spans the central portion of the alignment, representing the key functional domain responsible for protease recognition and inhibition. This region exhibits greater sequence variability, reflecting its evolutionary adaptability to different protease targets. Finally, the inhibitor core extends toward the C-terminal end, forming the highly conserved structural scaffold that maintains the disulfide-bonded tertiary fold characteristic of the CPI family.

Together, these sequential domains - signal peptide, reactive loop, and inhibitor core; illustrate a clear modular architecture that balances secretion targeting, functional specificity, and structural stability

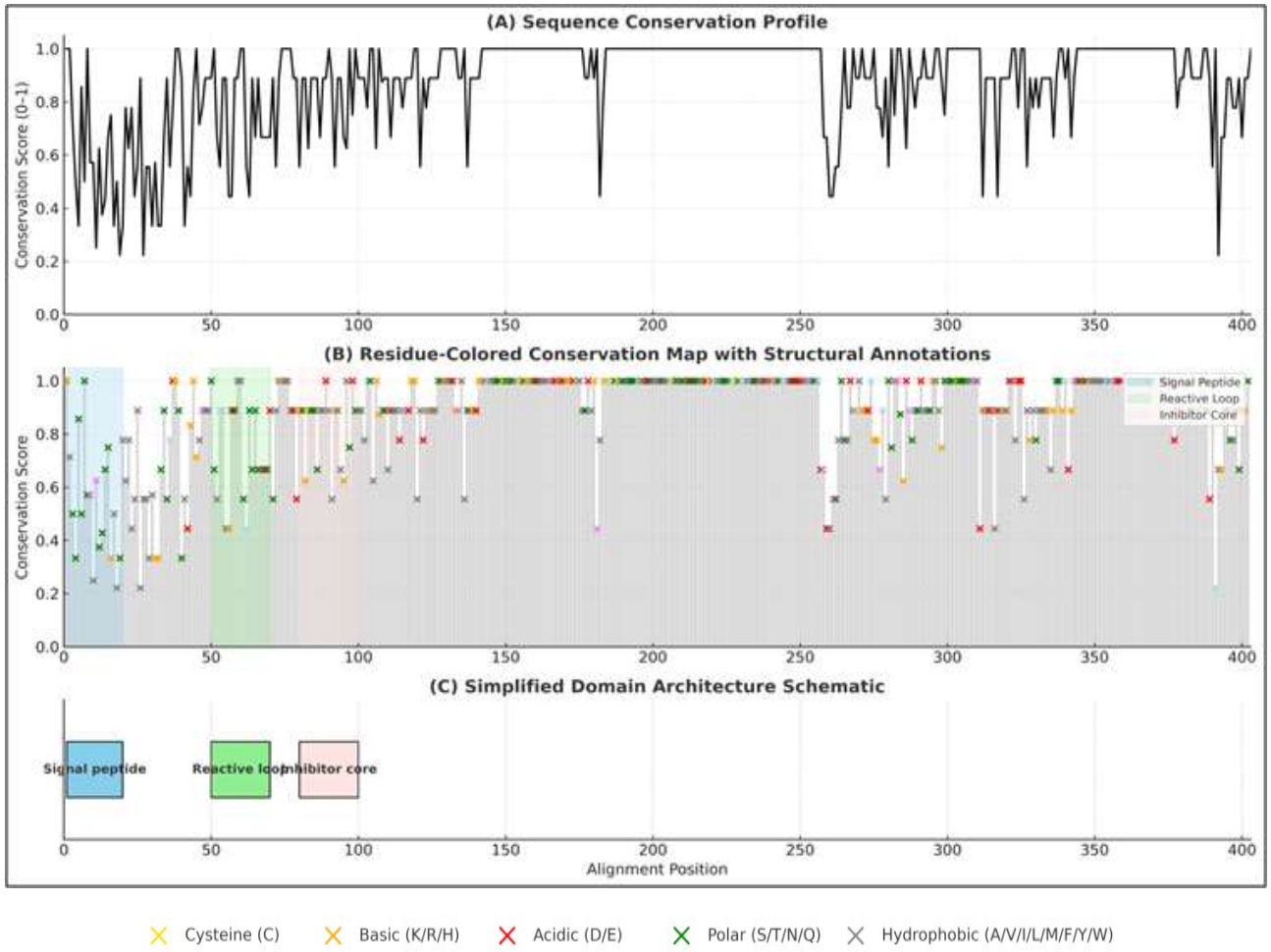

Fig 1. (A) Conservation *Profile* – overall residue conservation along the alignment. (B) *Residue-Colored Annotation Map* – cysteines, loop residues, and polar/hydrophobic sites distinguished by chemistry; shaded bands denote the signal peptide, reactive loop, and inhibitor core regions. (C) *Simplified Domain Architecture Schematic* – domain layout summary aligned to the same residue positions

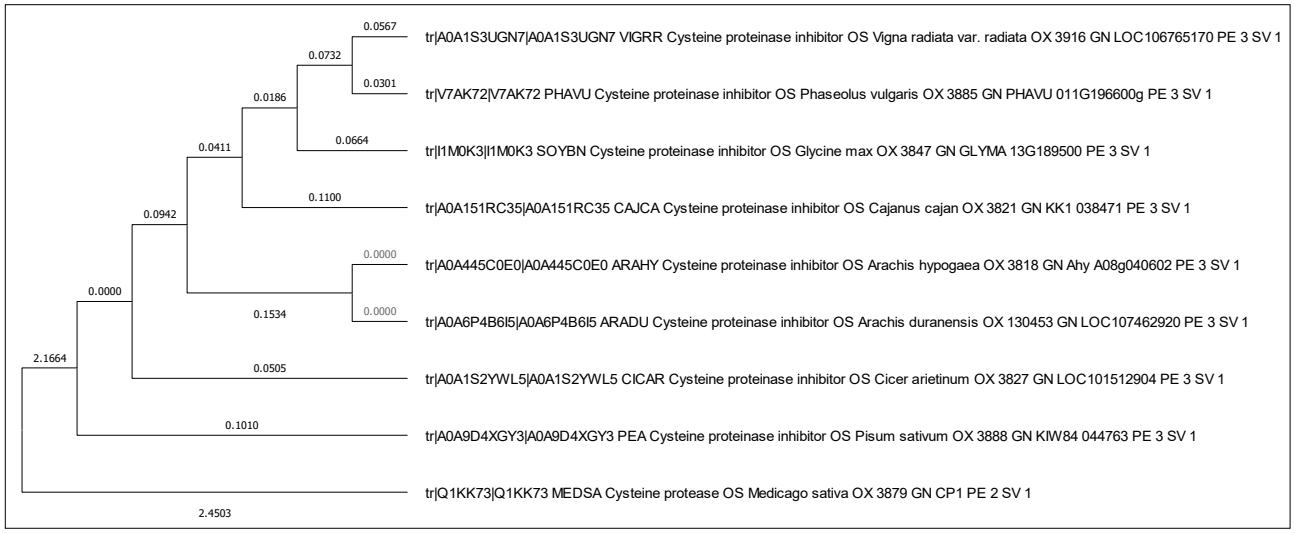

Fig. 2 Neighbor-Joining (NJ) phylogenetic tree of CPIs from nine legume species

within the CPI proteins. The CPI family maintains a rigid disulfide-stabilized framework (highly conserved) while allowing functional diversification through reactive loop variability. The conserved scaffold ensures structural stability, whereas the flexible reactive regions enable evolution of inhibitory specificity -hallmarks of plant protease inhibitor adaptation.

kDa, suggesting the presence of a multi-domain organization or an unprocessed precursor form. Most proteins exhibit instability index values above 40, indicative of marginal stability, although *Pisum* and *Medicago* display notably higher stability, implying enhanced structural resilience. The aliphatic indices, ranging from 26 to 30, point to moderate thermostability, consistent with a conserved leucine/

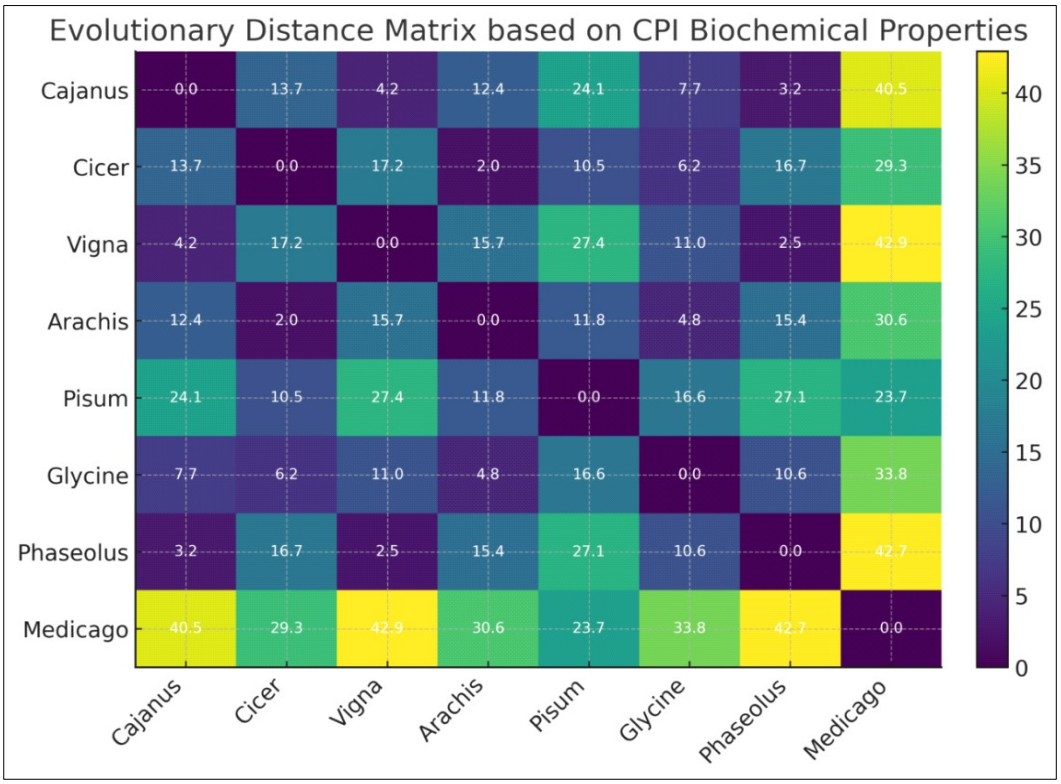

Fig 3. Evolutionary Distance Matrix based on Biochemical Properties of CPI Proteins

The tree depicts evolutionary clustering of *Arachis* CPI homologs with *Cicer arietinum* within the IRLC clade, while *Glycine*, *Vigna*, and *Phaseolus* form a distinct Millettioid group. Bootstrap values are shown at nodes; the scale bar represents substitutions per site.

Biochemical Analysis of CPIs

The physicochemical characterization of the CPI sequences (Table 3) reveals a coherent structural pattern across the legume representatives, with notable deviations that hint at evolutionary diversification. The molecular weights cluster around 23 kDa, aligning with the canonical size range of typical cysteine protease inhibitors, whereas the *Medicago* sequence stands out at approximately 45

isoleucine-rich hydrophobic core that supports the β-trefoil architecture.

All sequences exhibit negative GRAVY values, confirming a hydrophilic and soluble nature, typical of extracellular inhibitors secreted into the apoplast or rhizosphere. The isoelectric point (pI) variation, spanning 3.9 to 8.4, highlights substantial charge heterogeneity, which may facilitate adaptation to diverse protease environments or tissue-specific conditions. Predicted secondary structure patterns reveal balanced proportions of α-helices, β-sheets, and turns, reinforcing the classical β-trefoil–like folding that underpins CPI functionality.

Collectively, these trends illustrate a conserved physicochemical framework optimized for stability, solubility, and inhibitory activity, with

Table 3. CPI biochemical and structural property analysis

Species	MW (kDa)	Instability	Status	Aliphatic	GRAVY(hydropathy)	Hydrophobicity	pI
Cajanus cajan	22.9	56.8	UNSTABLE	27.7	-0.45	Hydrophilic	6.4
Cicer arietinum	23.9	43.2	UNSTABLE	26.9	-0.43	Hydrophilic	5.0
Vigna radiata	23.7	59.8	UNSTABLE	29.7	-0.36	Hydrophilic	8.4
Arachis hypogaea	23.4	44.4	UNSTABLE	27.6	-0.37	Hydrophilic	6.4
Pisum sativum	23.5	32.8	STABLE	27.7	-0.28	Hydrophilic	4.3
Glycine max	23.9	49.2	UNSTABLE	27.2	-0.47	Hydrophilic	6.6
Phaseolus vulgaris	23.7	59.8	UNSTABLE	27.6	-0.40	Hydrophilic	7.0
Medicago sativa	45.0	22.9	STABLE	27.1	-0.23	Hydrophilic	4.0

Medicago emerging as a distinct structural and biochemical outlier -its elevated mass, stability, and acidic pI hint at an ancestral or functionally divergent CPI lineage within legumes.

Molecular weight calculation

The molecular weight comparison chart illustrates a highly conserved size profile among legume CPI proteins, with most species clustering tightly around 23–24 kDa, consistent with the canonical monomeric cysteine protease inhibitor architecture (Fig. 4). These uniform molecular masses indicate a conserved single-domain structure stabilized by multiple disulfide linkages typical of plant CPIs. Across the dataset, *Cajanus, Cicer, Vigna, Arachis, Pisum, Glycine*, and *Phaseolus* all fall within a narrow 1 kDa window, suggesting strong evolutionary pressure to maintain a compact and functionally efficient inhibitory fold.

This molecular compactness likely supports efficient secretion and protease accessibility, critical to their defensive role in plant tissues. In contrast, *Medicago* exhibits a strikingly higher molecular weight of approximately 45 kDa, nearly double that of its counterparts. This significant increase implies a multi-domain configuration or the presence of additional peptide extensions or unprocessed propertied segments. Such a form may reflect domain duplication, fusion events, or specialized functional diversification unique to the *Medicago* lineage. Overall, the pattern suggests that while the CPI family maintains a highly conserved molecular size and structure across legumes, *Medicago*'s expanded variant may represent a divergent or ancestral form, potentially linked to

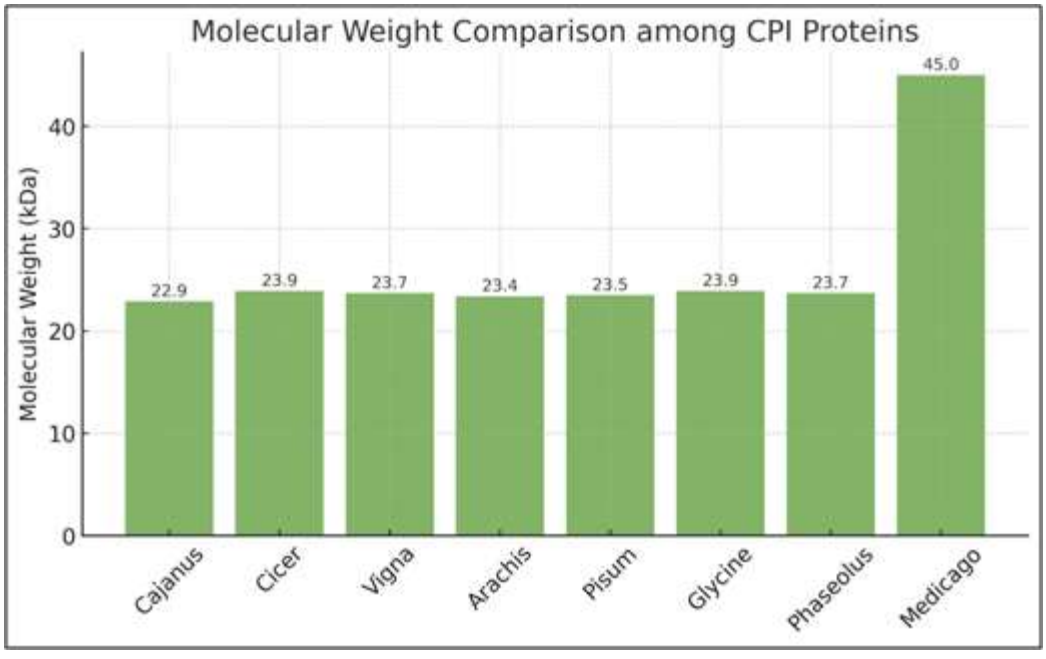

Fig.4. Molecular weight comparison chart

distinct regulatory or structural adaptations within its protease inhibitor repertoire.

Instability Index

The protein stability analysis based on the instability index provides a comparative assessment of the predicted in vivo stability of CPI proteins across legume species (Fig. 5). The stability threshold line (value = 40) divides the dataset into potentially stable (green) and unstable (red) proteins, allowing a clear visualization of species-specific differences in structural robustness. Most CPI proteins, including those from *Cajanus, Vigna, Arachis, Glycine, Phaseolus* and *Cicer*, exceed the threshold, with instability indices ranging between 43 and 60, classifying them as marginally or fully unstable under standard cellular conditions. This pattern suggests a structural tendency toward flexible or dynamic regions, possibly reflecting reactive loop mobility that enhances protease interaction but slightly compromises overall stability. Such moderate instability is a common characteristic of secreted defense proteins, which rely on transient interactions rather than long-term structural persistence.

In contrast, *Pisum* and *Medicago* display markedly lower instability indices (below 40), identifying them as relatively stable CPI variants. The *Medicago* protein, in particular, stands out as the most stable member of the dataset, with an index near 23, suggesting enhanced rigidity and structural integrity, likely a result of additional domains or tighter core packing. The *Pisum* inhibitor also demonstrates improved stability, possibly due to a more compact tertiary organization or reduced loop flexibility. Overall, this analysis indicates that while most legume CPIs maintain a functionally flexible but marginally stable fold, *Pisum* and especially *Medicago* exhibit enhanced conformational resilience, consistent with their elevated molecular mass and possible multi-domain evolution. These properties may confer greater resistance to denaturation or proteolytic degradation, highlighting potential functional specialization within the CPI family.

Aliphatic index

The thermostability analysis based on the aliphatic index reveals a consistent pattern across all CPI proteins, indicating a broadly conserved level of thermal resilience within the legume CPI family. The aliphatic index values, which range narrowly between 26 and 30, signify that these proteins possess a well-balanced content of aliphatic amino acids (alanine, valine, isoleucine, and leucine), key contributors to structural compactness and heat tolerance (Fig. 6). All eight species display comparable thermostability profiles, reflecting strong evolutionary conservation of the hydrophobic core architecture that underpins CPI structural integrity. Notably, *Vigna* shows a slightly elevated aliphatic index (~29.8), suggesting marginally enhanced thermostability, potentially linked to a higher leucine/isoleucine ratio that strengthens hydrophobic interactions.

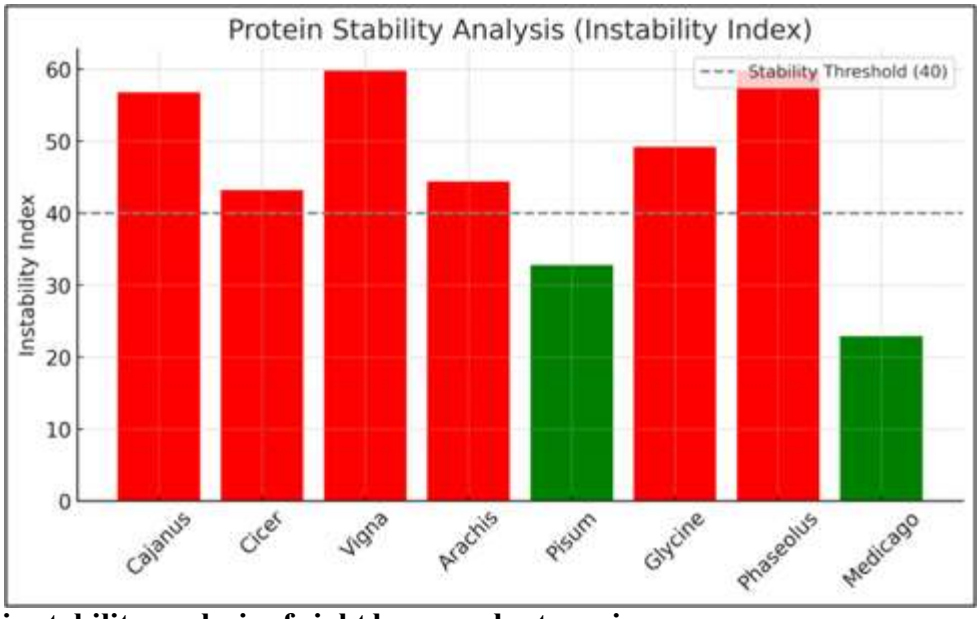

Fig 5. Protein stability analysis of eight legume plant species

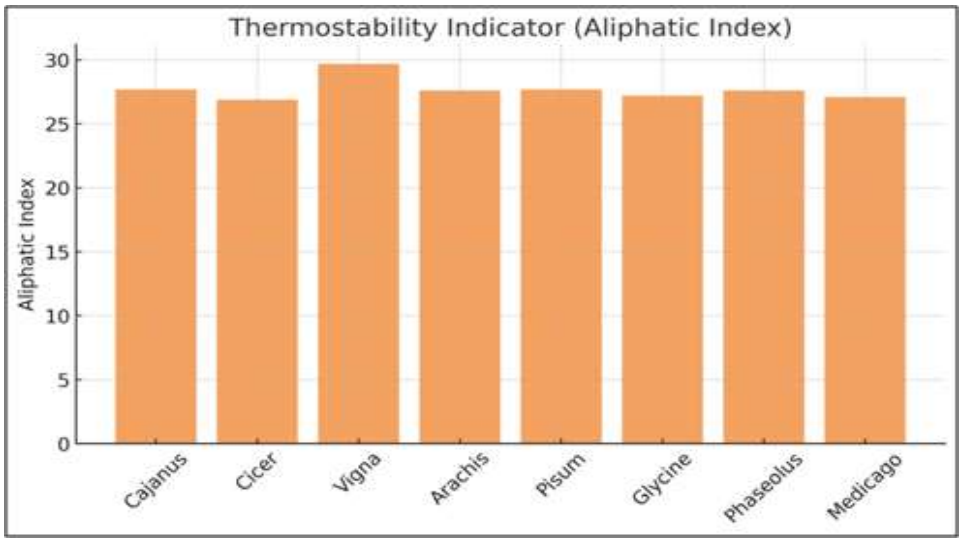

Fig.6. Thermostability Comparison among CPI Proteins based on Aliphatic Index

This uniformity across taxa indicates that thermostability is an evolutionarily stable trait within CPI proteins, maintained to ensure functionality under varying physiological or environmental conditions. The consistent aliphatic composition supports the hypothesis that the CPI scaffold, anchored by a â-trefoil core stabilized by disulfide bonds, is inherently thermotolerant, allowing it to remain active in diverse cellular environments. Overall, the results affirm that the CPI family maintains moderate to high thermostability through conserved aliphatic residues, ensuring structural robustness and reliable inhibitory performance even under temperature fluctuations common in plant stress responses.

Grand Average of Hydropathy (GRAVY)

The Grand Average of Hydropathy (GRAVY) scores for CPI proteins across eight legume species are given in Fig. 7. All species exhibit negative GRAVY values (ranging approximately from –0.15 to –0.45), confirming that CPI proteins are predominantly hydrophilic. This hydrophilicity indicates a strong affinity for aqueous environments, consistent with their extracellular or apoplastic localization and soluble inhibitor nature.

Among them, *Glycine* and *Cajanus* show the most negative scores, suggesting a higher surface polarity and potential for enhanced solubility. Conversely, *Medicago* and *Pisum* display slightly

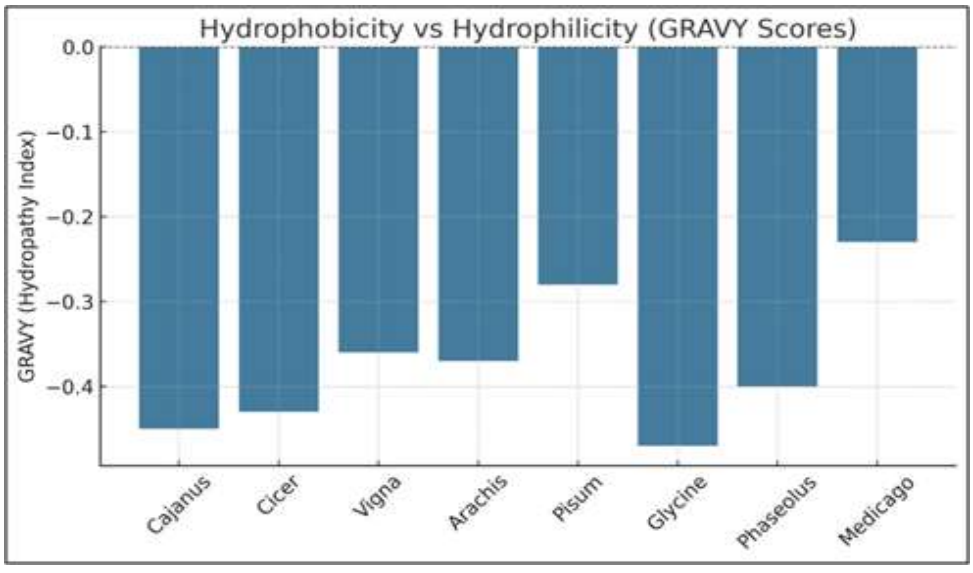

Fig 7. Hydrophobicity–Hydrophilicity balance among CPI Proteins (GRAVY Scores)

fewer negative values, reflecting a modestly more balanced hydrophobic–hydrophilic character that could contribute to greater stability or interaction versatility.

Overall, the hydropathy analysis reinforces the notion that legume CPIs are secreted, water-soluble defense proteins, optimized for effective diffusion and interaction with target proteases in the plant's extracellular matrix.

Theoretical pI Calculation

The isoelectric points (pI) of CPI proteins across eight legume species to highlight charge heterogeneity and potential functional adaptation (Fig. 8). The pI values range widely from 3.9 to 8.4, indicating that legume CPIs exhibit both acidic and basic forms, likely reflecting adaptation to different cellular or apoplastic environments and target protease charge profiles. *Vigna* displays the highest pI (~8.4), suggesting a basic inhibitor that may interact preferentially with acidic proteases. *Medicago*, at the opposite extreme (pI ~3.9), is distinctly acidic, consistent with its larger molecular mass and greater stability, possibly reflecting a divergent or ancestral CPI form. *Cajanus, Arachis, Glycine*, and *Phaseolus* show moderate pI values (6–7), implying a balanced electrostatic character suitable for broad protease interaction spectra. *Pisum* and *Cicer* fall on the acidic side (pI ~4–5), indicating more specialized electrostatic tuning for particular protease classes.

Overall, the wide pI distribution across species underscores the electrostatic diversity and adaptive plasticity of CPI proteins, enabling them to function efficiently under varying pH and protease conditions in different plant tissues and stress contexts.

Correlation Analysis

The relationship between hydrophobicity (GRAVY score) on the X-axis and protein stability (instability index) on the Y-axis for CPI proteins across eight legume species has been studied and is given in the Fig.9. The dashed horizontal line (Instability Index = 40) represents the stability threshold, distinguishing stable proteins (below the line) from unstable ones (above it). A clear inverse trend is evident *i.e.* proteins with higher hydrophilicity (more negative GRAVY values) tend to exhibit greater instability, while those with relatively higher hydrophobic character display enhanced stability. For example, *Medicago*, with the lowest instability index (~23) and least negative GRAVY (–0.2), emerges as the most stable and structurally compact CPI variant. Conversely, *Vigna, Phaseolus*, and *Cajanus* occupy the upper left quadrant, characterized by strong hydrophilicity (GRAVY H" –0.4) and high instability indices (>55), suggesting greater flexibility or conformational mobility (Fig. 9).Intermediate species such as *Arachis, Cicer, and Glycine* cluster near moderate hydrophilicity (–0.35 to –0.4) and marginal stability, reflecting a balanced trade-off between solubility and conformational rigidity typical of functional inhibitors.

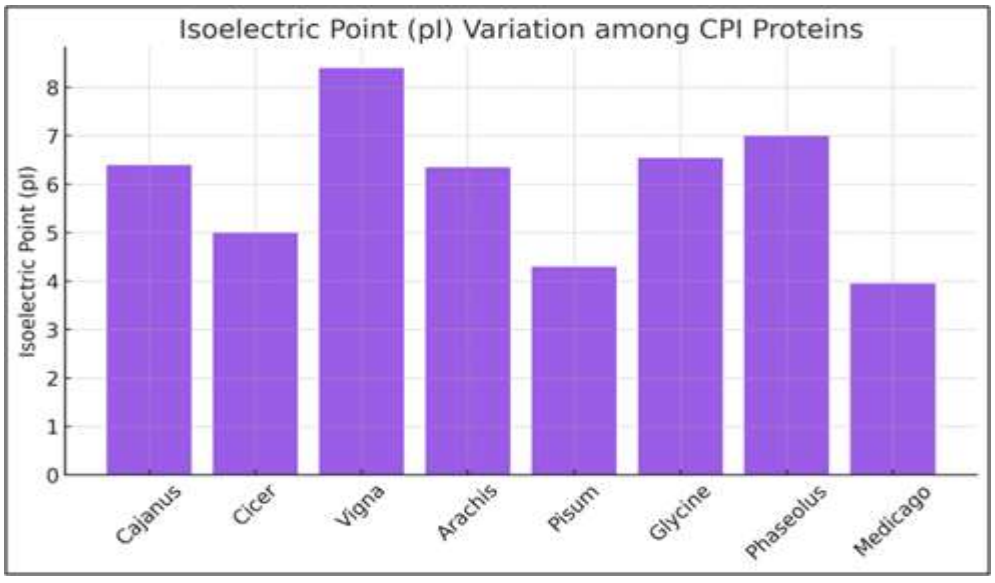

Fig 8. Isoelectric Point (pI) Variation among CPI Proteins

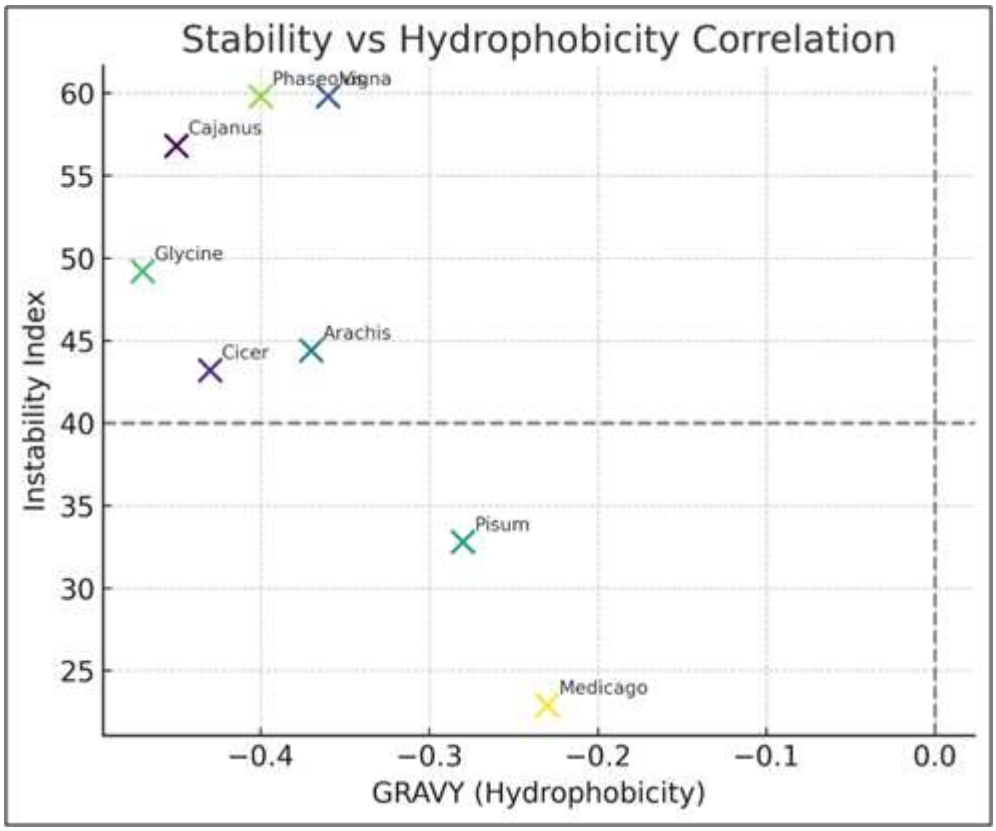

Fig 9. Correlation between Protein Stability and Hydrophobicity among CPI Proteins

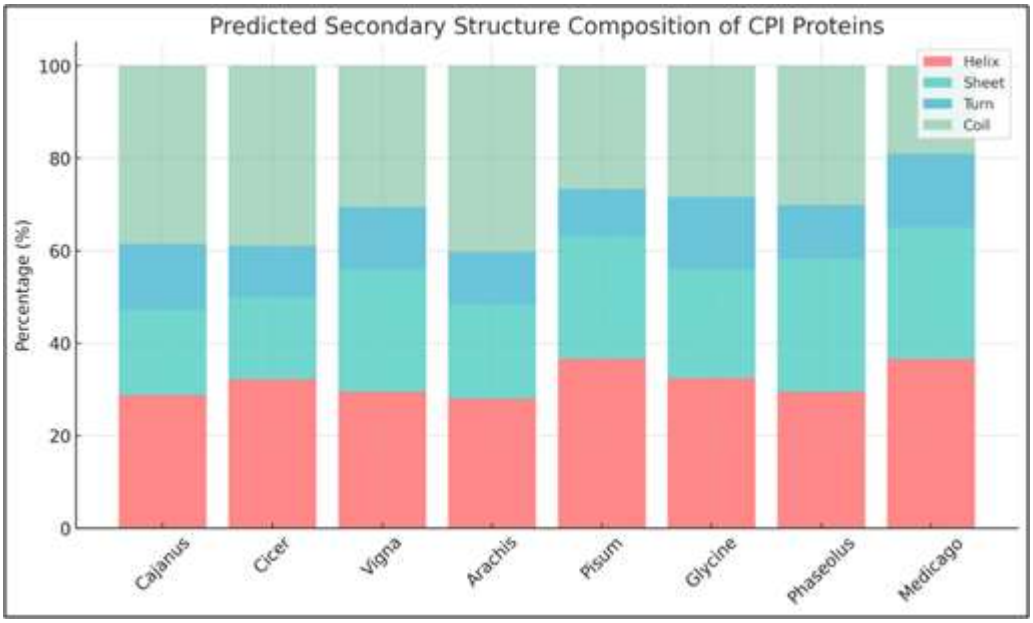

Fig.10. Predicted Secondary Structure Composition of CPI Proteins across Legume Species

Pisum forms an intermediate outlier with moderate hydrophobicity and improved stability, aligning with its previously observed physicochemical robustness. Overall, the plot demonstrates that CPI protein stability inversely correlates with hydrophilicity, emphasizing how hydrophobic core strength contributes to conformational resilience, whereas surface polarity supports solubility and protease accessibility, a defining duality in the evolutionary optimization of legume CPIs.

Secondary Structure Prediction

The predicted secondary structure (Fig.10) distribution of CPI proteins from eight legume species

based on the proportion of α-helices, β-sheets, turns, and coils in their overall structure. All CPI proteins exhibit a balanced composition, characteristic of β-trefoil–like folds, with approximate contributions of 30–35% helices, 20–25% β-sheets, and the remainder distributed between turns and coils (Fig. 3). The prevalence of coils and turns (35–40%) indicates substantial flexibility and loop mobility, essential for dynamic interactions with target proteases. Among the species, *Pisum* and *Medicago* show slightly higher helical content, suggesting enhanced structural rigidity and stability, consistent with their lower instability indices observed in physicochemical analyses. In contrast, *Vigna* and *Arachis* display a greater proportion of β-sheets and coil regions, potentially supporting adaptive flexibility and binding plasticity.

Overall, the pattern reinforces that legume CPIs maintain a conserved secondary structure framework, optimized for structural stability through helices and sheets while retaining functional versatility via flexible turns and loops, a key feature underlying their effectiveness as broad-spectrum protease inhibitors.

CONCLUSION

This study presents a rigorous, quantitative evaluation of cysteine proteinase inhibitors (CPIs) across eight legume species, integrating sequence, structural, phylogenetic, and biochemical analyses. Core inhibitory motifs and cysteine residues forming disulfide scaffolds are highly conserved; ensuring structural integrity and protease-binding functionality, while peripheral regions, particularly N-terminal signal peptides and reactive loops, exhibit sequence variability that likely underlies species-specific targeting and functional specialization. Phylogenetic clustering and distance matrix analyses highlight evolutionary relationships, with *Medicago* emerging as a divergent outlier with distinct molecular weight, stability, and isoelectric properties. Structural predictions confirm a conserved α-trefoil–like fold, with flexible loops supporting inhibitory versatility, and physicochemical profiling reveals a balance between hydrophilicity, stability, and thermostability. Correlation analyses further emphasize the inverse relationship between hydrophobicity and conformational flexibility, reflecting evolutionary optimization for extracellular inhibitory function. Collectively, these findings demonstrate that legume CPIs maintain a highly conserved structural and functional core while evolving adaptive features that enable diverse physiological roles in plant defense and stress response. Hence, these genes can be exploited candidate genes for biotic stresses like bruchid infestation in storage of legumes in general and peanut in particular, being major food and oil source. The conserved inhibitory motifs, disulfide-bond framework, and α-trefoil–like fold observed across legumes are also present in the peanut candidates, confirming their identity as bona fide cystatins. Species-specific variability in signal peptides and reactive loops parallels the tissue-specific expression patterns in peanut, where several genes are enriched in pods and seeds, consistent with conserved roles in reproductive defense, while others show diversification into roots, nodules, and pericarp. Thus, the *A. hypogaea* CPI set fits the broader legume pattern of maintaining a conserved structural core while evolving adaptive features that support specialized functions in plant defense and protease regulation. Overall, the integration of legume-wide in-silico analyses with the tissue-specific CPI gene set of *Arachis hypogaea* demonstrates that peanut cystatins preserve the conserved structural and biochemical framework typical of legume CPIs, while exhibiting adaptive expression patterns aligned with reproductive defense and stress resilience. These findings highlight the identified pod-, seed-, and reproductive-tissue–enriched CPI genes as strong functional candidates for targeted improvement of storage pest tolerance in groundnut.

LITERATURE CITED

Abe K and Arai S 1985. Purification and characterization of rice cysteine proteinase inhibitors. J Biol Chem. 260: 511–515.

Benchabane M, Schluter U, Vorster J, Goulet M-C and Michaud D 2010. Plant cystatins. Biochimie. 92(11): 1657–1666.

Girard C, Leple J-C, D' Ovidio R and Michaud D 2007. Overexpression of oryzacystatin reduces seed proteolysis. Plant Science. 173: 285–292.

Guruprasad K, Reddy B V B and Pandit M W 1990. Correlation between stability of a protein and its dipeptide composition: a novel

approach for predicting in vivo protein stability. Protein Engineering. 4(2): 155–161.

Habib H and Fazili K M 2007. Plant protease inhibitors: a defense strategy in plants. Biotechnol Mol Biol Rev. 2(3): 68–85.

Katoh K and Standley D M 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 30(4): 772–780.

Kondo H, Abe K, Nishimura I, Watanabe H, Emori Y and Arai S 1990. Two distinct cystatin species in rice seeds with different specificities against cysteine proteinases: molecular cloning, expression, and biochemical studies on oryzacystatin-II. J Biol Chem. 265(26): 15832–15837.

Kyte J and Doolittle R F 1982. A simple method for displaying the hydropathic character of a protein. J Mol Biol. 157(1): 105–132.

Nguyen L T, Schmidt H A, Von Haeseler A and Minh B Q 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 32(1): 268–274.

Polya G M 2003. Protein and non-protein protease inhibitors from plants. Studies in Natural Products Chemistry. 29: 567–641.

Valueva T A and Mosolov V V 1999. Role of inhibitors of proteolytic enzymes in plant defense. Biochemistry (Moscow). 64: 1035–1040.

Van der Hoorn R A and Kamoun S 2008. From guard to decoy: a new model for perception of plant pathogen effectors. Plant Cell. 20(8): 2009–2017.

Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, Heer F T, de Beer T A P, Rempfer C, Bordoli L and Lepore R 2018. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. 46(W1): W296–W303.

Zhang L, Du L and Poovaiah B W 2014. Calcium signaling and biotic defense responses in plants. Plant Signaling and Behavior. 9(11): e973818.

Received on 02.08.2025 and Accepted on 10.09.2025