RESEARCH ARTICLE

SABRAO Journal of Breeding and Genetics
55 (6) 1950-1962, 2023
http://doi.org/10.54910/sabrao2023.55.6.9
http://sabraojournal.org/
pISSN 1029-7073; eISSN 2224-8978

BIOINFORMATICS-BASED CHARACTERIZATION OF THE CHALCONE SYNTHASE (*CHS*) FAMILY GENES IN FLOWERING PLANTS

S. HUSSAIN[1*], A. HUSSAIN[2], I. AHMAD[1], F. WAHID[1], and M. SAJID[1]

[1]Department of Horticulture, The University of Agriculture, Peshawar, Pakistan
[2]Department of Agriculture, Abdul Wali Khan University Mardan, Pakistan
[*]Corresponding author's email: sajeelhussain01@gmail.com
Email addresses of co-authors: fazaliwahid@aup.edu.pk, msajid@aup.edu.pk, imran73pk@gmail.com, adilhussain@awkum.edu.pk

SUMMARY

Chalcone synthase (*CHS*) is an essential rate-limiting enzyme in the biosynthesis of anthocyanin pigments found in plant organs, such as, flowers and fruits. The *CHS* gene family appears in all flowering plants. Here, we searched and characterized the *CHS* genes from different flowering plants. Database search resulted in identifying Chalcone synthase genes from 29 diverse plant species. Phylogenetic analysis indicated significantly higher similarity between the various *CHS* genes, divided into at least six closely related rooted clades. Gene structure analysis identified the relative sizes and positions of UTRs, introns, and exons. Protein sequence alignment specified more than 95% similarity between the *CHS* genes, with eight highly conserved domains of different lengths. Likewise, in-depth analysis showed the presence of three highly conserved motifs in the protein sequence of all the 29 chalcone synthase genes. Physicochemical properties, such as, molecular weight, instability index, aliphatic index, hydropathicity (GRAVY), length, and isoelectric point (pI) of the *CHS* genes were significantly similar. Furthermore, the predicted 3D structures of *CHS* genes from different plant species highly remained and are homologous to each other, indicating that the *CHS* family genes have significantly conserved sequences and functionality across the plant kingdom.

Keywords: Phylogenetic tree, gene length, motif analysis, protein alignment, 3D protein structure, physicochemical properties

Key findings: The bioinformatics approach used in the study gainfully identified and analyzed Chalcone synthase genes of 29 flowering plants. The various physicochemical attributes, the DNA, and the 3D protein structures of *CHS* genes from different plant species appeared highly conserved across the plant kingdom.

Communicating Editor: Dr. Sajjad Hussain Qureshi

Manuscript received: July 21, 2023; Accepted: August 30, 2023.

Citation: Hussain S, Hussain A, Ahmad I, Wahid F, Sajid M (2023). Bioinformatics-based characterization of the Chalcone synthase (*CHS*) family genes in flowering plants. *SABRAO J. Breed. Genet.* 55(6): 1950-1962. http://doi.org/10.54910/sabrao2023.55.6.9

INTRODUCTION

Chalcone synthase is an influential and rate-limiting enzyme in the biosynthesis of anthocyanin pigments found in plant organs, such as, flowers and fruits. Still, the relationship between petal coloration level and *CHS* expression in different cultivars is not fully understood. The role of *CHS* genes has undergone thorough study for understanding the synthesis and accumulation of anthocyanin pigments. Chalcone synthases belong to a group of enzymes called type III polyketide synthases (PKSs), specific to plants. Genes encode *CHSs* by producing relatively small proteins (around 40–45 kDa) that form pairs. *CHSs* facilitate a chemical reaction involving the joining of malonyl-CoA and starter molecules linked to CoA, resulting in the production of specific compounds. Alongside *CHS*, there are other enzymes within the same group of type III PKSs found in plants. These include bibenzyl synthase, benzophenone synthase, curcumin synthase, acridone synthase, and stilbene synthase (STS) (Austin and Noel, 2003).

Earlier conclusions stated that vegetables like cucumber's ability to withstand waterlogging might connect to specific genes called *CsCHS*. When aphids feed on the cucumber, it increases the *CsCHS* genes' activity, which could also lead to activating other related genes in the cucumber. Additionally, researchers found that two similar genes, *CsERF1* and *CsERF3*, can interact with a particular *CsCHS* gene called *CsCHS2*. This interaction may influence how the cucumber's immune system responds to different types of stress. As a result, *CsCHS2* could play a crucial role in developing cucumber varieties that are resilient against various environmental challenges.

Several reports described the effects of altering the expression of the *CHS* gene in transgenic plants, e.g., in petunia, the manifestation of the antisense *CHS* gene results in pale or even white-colored flower formation due to the inhibition of anthocyanin production with a negative impact on plant fertility (Napoli *et al.*, 1990; Blokland *et al.*, 1994). The chalcone synthase enzyme is vital in synthesizing multiple secondary metabolites in plants, bacteria, and fungi. *CHS* is well-known as the gatekeeper of the anthocyanin pathway (Dao *et al.*, 2011). *CHS* are members of the plants-specific type III polyketide synthase (PKS) gene family (Austin and Noel, 2003; Wu *et al.*, 2020), which catalyze the main steps of the phenylpropanoid pathway leading to the synthesis of various flavonoids (Winkel-Shirley, 2001; Pang *et al.*, 2005). Flavonoids are a group of secondary metabolites that consist of several different classes of compounds, such as, flavones, chalcones, anthocyanin, and flavonol isoflavones. Flavonoids have many important biological functions, including flower pigmentation, protection against UV radiation, auxin transport, defense against phytopathogens, and pollen fertility (Winkel-Shirley, 2002; Falcone Ferreyra *et al.*, 2012).

After about two centuries of research, three distinct groups of plant-based pigments have been identified, i.e., carotenoids, flavonoids, and alkaloids, based on their chemical structure, cellular localization, and biochemical synthesis (Zhao and Tao, 2015). Of these, carotenoids are the most abundant pigments in nature, found in fruits, flowers, leaves, and roots of higher plants. These further subdivide into two groups—carotene and lutein. Flavonoids are a large group of secondary metabolites widely distributed across the plant kingdom. They are highly essential pigments producing the broadest spectrum of colors found in plant flowers, such as, chrysanthemums, dahlias (Thill *et al.*, 2012), groundcover roses (Schmitzer *et al.*, 2010), violets (Tatsuzawa *et al.*, 2012), and herbaceous peony (Zhao *et al.*, 2012). Alkaloids are cyclic organic substances containing negatively oxidized nitrogen atoms, including papaverine, betalain, and berberine. Among these, betalain is a water-soluble nitrogen found in red beets (also known as purple beetroots), some flowers, fruits, and leaves. Betacyanin and betaxanthin occur in these plants, where betacyanin is the main component, accounting for 75%–95% of the total betalain (Strack *et al.*, 2003). The study's objective was to search the genomes of different plant species, identify *CHS* genes, and characterize these using different bioinformatics tools, which will provide valuable information for breeders and future functional genomics research.

MATERIALS AND METHODS

Plant species and phylogenetic analysis

Selection totaled 29 species of flowering plants (Table 1) for chalcone synthase genes' characterization. All plant species' genomic, coding DNA sequence (CDS), and protein sequences of chalcone synthase genes downloaded *via* a search query in NCBI (https://www.ncbi.nlm.nih.gov/) used the amino acid sequence of the Arabidopsis *CHS* gene (Gene Bank: NP_16897.1). Downloading the respective protein sequences continued from the UniProt database (https://www.uniprot.org/) (UniProt, 2022). The protein sequences helped in phylogenetic analysis to determine their evolutionary relationships using the MEGA11 Software (Tamura *et al.*, 2021). It resulted in a phylogenetic tree drawn using the Neighbor-Joining method (Saitou and Nei, 1987). The constructed tree formed a scale with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances' computation used the Poisson correction (Zuckerkandl and Pauling, 1965) in the units of the number of amino acid substitutions per site. All ambiguous positions' removal ensued for each sequence pair (pair-wise deletion).

Gene structure analysis

Gene structure analyses show the numbers and locations of introns and exons in the genomic sequence of a gene. Constructing gene structures for all the chalcone synthase genes via the Gene Structure Display Server (http://gsds.gao-lab.org/) (Hu *et al.*, 2014) utilized the genomic DNA sequences [5'UTR::Exon(s)::Intron(s)::3'UTR].

Alignment and identification of conserved domains

All 29 protein sequence alignments used the ClustalW (Thompson *et al.*, 1994) in BioEdit (Hall, 1999). Visual observations and engaging the "find conserved regions" option in BioEdit identified conserved domains.

Motif analysis

Conducting motif analyses used MEME (Multiple EM for Motif Elicitation) https://meme-suite.org/meme/ (Bailey *et al.*, 2009) to find the number and location of highly conserved motifs in the *CHS* genes of the selected plant species. The recording of sites, patterns, and p-values followed. Downloading all the data as PDF or TIF images stemmed where required. The consensus motif logos saved as SVG included information regarding the E-values, site, and motif width.

Physicochemical properties

The study determined the physicochemical properties of all the chalcone synthase genes using the ProtParam platform of the Swiss Bioinformatics Resource PortalExPasy (https://web.expasy.org/protparam/) (Gasteiger *et al.*, 2005). Physicochemical properties, such as, molecular weight, instability index, aliphatic index, hydropathicity (GRAVY), length, and the isoelectric point (pI) values calculations commenced for all the chalcone synthase proteins.

The 3D protein structure

Representative 3D protein structures of chalcone synthase genes acquired downloading for six plant species among the selected 29 plant species from UniProt https://www.uniprot.org/(UniProt, 2022). These included original, as well as, predicted 3D structures where applicable. The downloaded structure source files as PDB files continued analysis in PyMol (DeLano, 2002) and saved them as JPEG files. The recorded structure confidence statistics for the predicted structures emerged in different colors in the protein cartoon structure as very high, confident, low, and very low confidence.

RESULTS

Identification of chalcone synthase genes, phylogenetic, and gene structure analysis

The database search resulted in the chalcone synthase genes' identification from 29 different plant species (Table 1). The traditional phylogenetic tree showed the evolutionary relationship between the chalcone synthase genes of the 29 varied plant species. Analysis indicated extremely high similarity between the different *CHS* genes, divided into at least six closely related rooted clades (Figure 1). Furthermore, gene structure analysis specified the number, position, and relative sizes of the introns and exons in the *CHS* genes (Figure 1). The *Arabidopsis thaliana CHS* gene appeared to have the largest 5' UTR of 2048bp and 542bp of 3' UTR and a single intron. Apart from the *CHS* gene of *Stenopetalum lineare,* which had seven introns, all the *CHS* genes had a single intron (Figure 1). Interestingly, the first exon of all the *CHS* genes appeared almost the same size, except the *CHS* gene from *Stenopetalum lineare* and *Yinshania acutangula*.

Alignment and identification of conserved domains

The amino acid sequences of all 29 *CHS* genes from different plants showed more than 95% similarity with each other (Figure 2). Furthermore, eight highly conserved regions (CRs) of different lengths occurred across the *CHS* genes. These regions have labels as A, B, C – F (red alphabets, Figure 2). The CR C was notably the extensive conserved region in the middle of the gene, spanning 78 total amino acids (R100 to R178). Interestingly, in 19 plant species, the CR C appeared to be 132aa long, traversing way up to the end of CR G (233aa). The CR G had the lowest number of conserved amino acids, starting from 342 up to 357 (16aa) (Figure 2). These conserved regions in the protein sequence alignment of the chalcone synthase gene show similarities of the chalcone synthase gene between these 29 selected plant species and also determine the evolutionary relationship of these plant species with each other.

Motif analysis

We found three highly conserved motifs (5' - red, purple, and green boxes - 3') in the protein sequence of all the 29 chalcone synthase genes (Figure 3). In total, 29 sites contributed to constructing all three motifs, with a width of 50 amino acids (Figure 3).

Physicochemical properties

Determining the physicochemical properties of all the chalcone synthase genes used the ProtParam database of the Swiss Bioinformatics Resource Portal ExPasy (https://web.expasy.org/protparam/) (Gasteiger *et al.*, 2005). These physicochemical properties are valuable as they determine the function of these proteins. These include molecular weight, instability index, aliphatic index, hydropathicity (GRAVY), length, and pI. The values of these properties of all the selected plant species were relatively close to each other (Table 2). *Halimolobos perplexus var. Perplexus* has the heaviest molecular weight of 43117.84da, whereas *Transberingia bursifolia CHS* had the lowest M.W. of 42283.82da. *Boechera fendleri and Boechera stricta CHS* genes had the lowest instability index of 30.45, whereas *Yinshania acutangula* had the highest instability index of 34.07. *Malcolmia maritime CHS* had the lowest aliphatic index of 89.62, while that of *Descurainia sophia* was 93.31, the highest aliphatic index compared with the other *CHS* genes. On the other hand, *Turritis glabra CHS* had the minimum hydropathicity value of -0.038, and *Descurainia preauxiana* had the maximum hydropathicity value of -0.103.

Table 1. Result of blast search for Chalcone synthase genes from different plant species.

No	Description	Scientific Name	Query Coverage	E: Value	Percent identity	Acc. Length	Accession
1	Chalcone synthase	*Arabidopsis thaliana*	100%	0	100	395	NP_196897.1
2	Chalcone synthase	*Camelina microcarpa*	99%	0	98.98	393	CDU44567.1
3	Chalcone synthase [*Halimolobos perplexus* var. perplexus]	*Halimolobos perplexus var. perplexus*	100%	0	98.73	395	AAF23569.1
4	Chalcone synthase [*Capsella rubella*]	*Capsella rubella*	100%	0	98.48	395	AAF23581.1
5	Chalcone synthase [*Murbeckiella boryi*]	*Murbeckiella boryi*	99%	0	98.98	393	CDU45783.1
6	Chalcone synthase [*Crucihimalaya himalaica*]	*Crucihimalaya himalaica*	100%	0	98.48	395	AAG43349.1
7	Chalcone synthase [*Alyssopsis mollis*]	*Alyssopsis mollis*	99%	0	98.73	393	CDU43751.1
8	Chalcone synthase [*Turritis glabra*]	*Turritis glabra*	100%	0	98.23	395	AAF23566.1
9	Chalcone synthase [*Boechera fendleri*]	*Boechera fendleri*	100%	0	97.97	395	AAF23565.1
10	Chalcone synthase [*Boechera stricta*]	*Boechera stricta*	100%	0	97.72	395	AAF23563.1
11	Chalcone synthase	*Descurainia paradisa*	99%	0	97.96	393	CDU45082.1
12	Chalcone synthase [*Ballantinia antipoda*]	*Ballantinia antipoda*	98%	0	99.23	388	ADE18841.1
13	Chalcone synthase [*Yinshania acutangula*]	*Yinshania acutangula*	99%	0	97.2	393	CDU46546.1
14	Chalcone synthase [*Malcolmia graeca*]	*Malcolmia graeca*	99%	0	97.46	393	CDU45861.1
15	Chalcone synthase [*Crucihimalaya wallichii*]	*Crucihimalaya wallichii*	98%	0	98.71	388	AQW79552.1
16	Chalcone synthase [*Stenopetalum lineare*]	*Stenopetalum lineare*	98%	0	98.71	388	AQW79573.1
17	Chalcone synthase [*Cuphonotus andraeanus*]	*Cuphonotus andraeanus*	98%	0	98.45	388	AQW79554.1
18	Chalcone synthase [*Crucihimalaya mollissima*]	*Crucihimalaya mollissima*	98%	0	98.45	388	ACN41919.1
19	Chalcone synthase [*Descurainia sophia*]	*Descurainia sophia*	99%	0	97.2	393	CDU45103.1
20	Chalcone synthase [*Transberingia bursifolia*]	*Transberingia bursifolia subsp. Bursifolia*	98%	0	98.45	388	ACN41920.1
21	Chalcone synthase [*Harmsiodoxa puberula*]	*Harmsiodoxa puberula*	98%	0	98.2	388	AQW79561.1
22	Chalcone synthase [*Malcolmia maritima*]	*Malcolmia maritima*	99%	0	96.95	393	CDU45877.1
23	Chalcone synthase [*Descurainia preauxiana*]	*Descurainia preauxiana*	98%	0	98.2	388	AQW79555.1
24	Chalcone synthase [*Menkea villosula*]	*Menkea villosula*	98%	0	98.2	388	AQW79564.1
25	Chalcone synthase [*Erysimum pseudorhaeticum*]	*Erysimum pseudorhaeticum*	99%	0	96.94	394	CDU45268.1
26	Chalcone synthase [*Smelowskia tibetica*]	*Smelowskia tibetica*	99%	0	96.69	393	CDU45443.1
27	Chalcone synthase [*Brachypus suffruticosus*]	*Brachypus suffruticosus*	98%	0	96.93	392	CDU45287.1
28	Chalcone synthase [*Ionopsidium abulense*]	*Ionopsidium abulense*	99%	0	97.19	394	CDU45460.1
29	Chalcone synthase [*Harmsiodoxa brevipes*]	*Harmsiodoxa brevipes*	98%	0	98.2	388	AQW79560.1

Notes: Blast result of the chalcone synthase gene of the selected 29 plant species shows that *CHS* gene of all selected 29 plant species has Query coverage between 98%–100%, E value 0, Percent identity 96.93–100, Acc. Length 388–395.

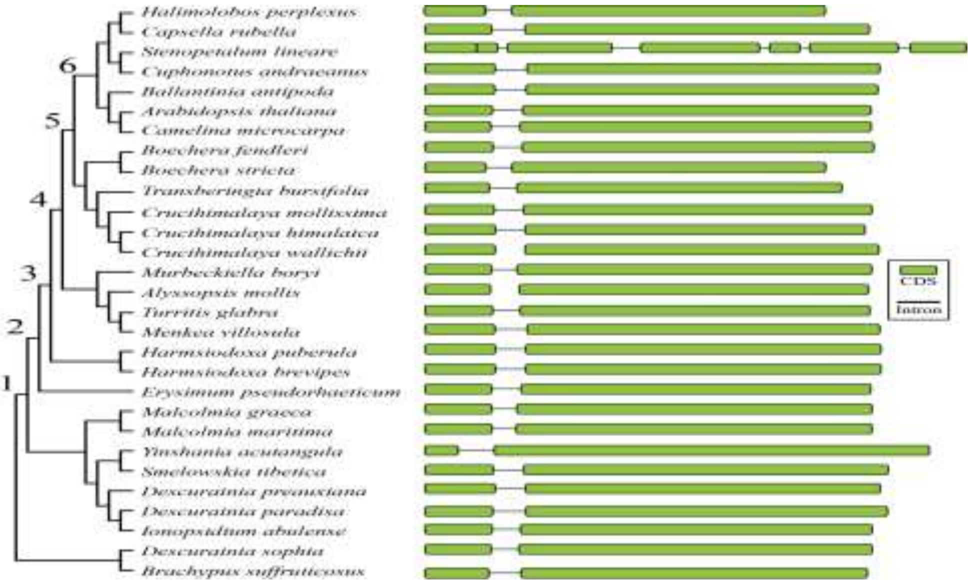


Figure 1. Phylogenetic and gene structure analysis.

The evolutionary history was inferred using the Neighbor-Joining method (Saitou and Nei, 1987). The optimal tree is shown. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Poisson correction method (Zuckerkandl and Pauling, 1965) and are in the units of the number of amino acid substitutions per site. The analysis involved 29 amino acid sequences. All ambiguous positions were removed for each sequence pair (pair-wise deletion). There were a total of 396 positions in the final dataset. Evolutionary analyses were conducted in MEGA11 (Tamura et al., 2021). Gene structure analysis showed the number, positions, and sizes of introns and exons.

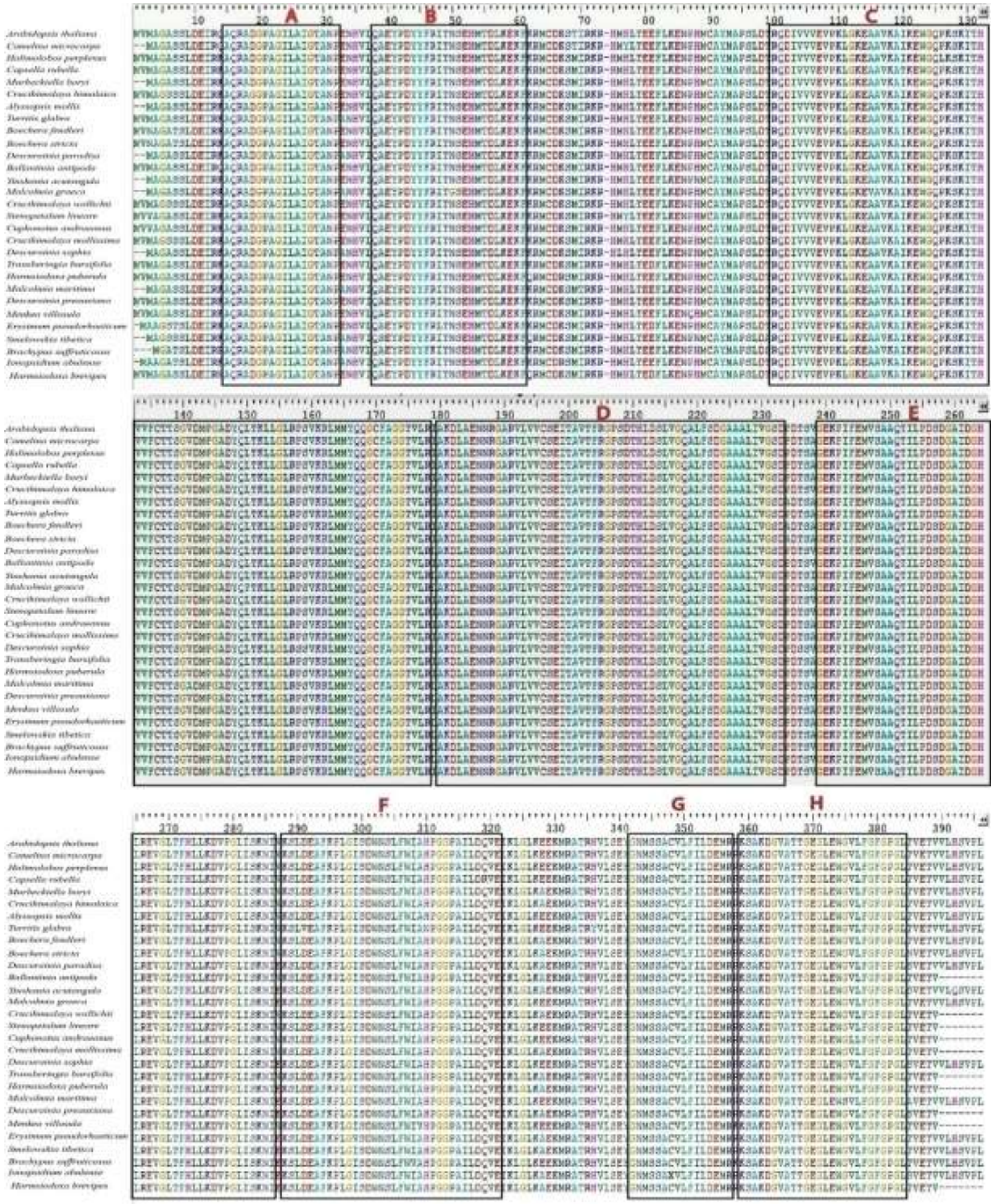
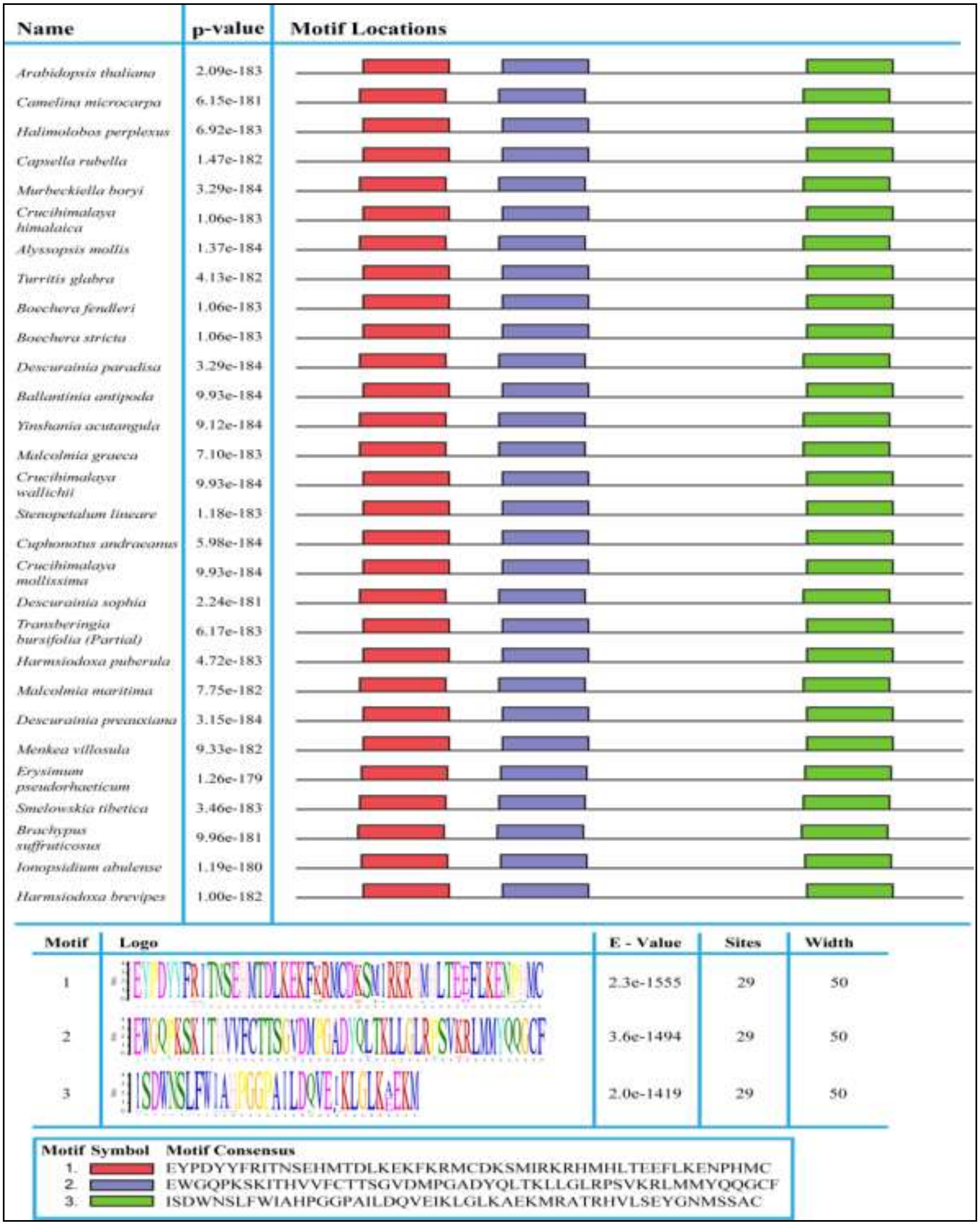


Figure 2. Alignment and identification of conserved domains in the protein sequence of chalcone synthase from 29 different plants.

The 29 CHS proteins from different plants appeared more than 95% similar, with eight highly conserved regions (CRs) of different lengths represented as A, B, C – F. The CR C showed the extensive conserved region (R100 to R178). The CR G had the lowest number of conserved amino acids, starting from 342 up to 357 (16aa).



Name	p-value
Arabidopsis thaliana	2.09e-183
Camelina microcarpa	6.15e-181
Halimolobos perplexus	6.92e-183
Capsella rubella	1.47e-182
Murbeckiella boryi	3.29e-184
Crucihimalaya himalaica	1.06e-183
Alyssopsis mollis	1.37e-184
Turritis glabra	4.13e-182
Boechera fendleri	1.06e-183
Boechera stricta	1.06e-183
Descurainia paradisa	3.29e-184
Ballantinia antipoda	9.93e-184
Yinshania acutangula	9.12e-184
Malcolmia graeca	7.10e-183
Crucihimalaya wallichii	9.93e-184
Stenopetalum lineare	1.18e-183
Cuphonotus andraeanus	5.98e-184
Crucihimalaya mollissima	9.93e-184
Descurainia sophia	2.24e-181
Transberingia bursifolia (Partial)	6.17e-183
Harmsiodoxa puberula	4.72e-183
Malcolmia maritima	7.75e-182
Descurainia preauxiana	3.15e-184
Menkea villosula	9.33e-182
Erysimum pseudorhaeticum	1.26e-179
Smelowskia tibetica	3.46e-183
Brachypus suffruticosus	9.96e-181
Ionopsidium abulense	1.19e-180
Harmsiodoxa brevipes	1.00e-182

Motif	Logo	E - Value	Sites	Width
1		2.3e-1555	29	50
2		3.6e-1494	29	50
3		2.0e-1419	29	50

Motif Symbol		Motif Consensus
1.	🟥	EYPDYYFRITNSEHMTDLKEKFKRMCDKSMIRKRHMHLTEEFLKENPHMC
2.	🟦	EWGQPKSKITHVVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCF
3.	🟩	ISDWNSLFWIAHPGGPAILDQVEIKLGLKAEKMRATRHVLSEYGNMSSAC

Figure 3. Identification of conserved motifs in the amino acids sequence of chalcone synthase from 29 different plants.

Three highly conserved motifs in the protein sequence of the chalcone synthase genes (5'-red, purple, and green boxes - 3') occurred. A total of 29 sites contributed to the construction of all three motifs with a width of 50 amino acids.

Table 2. Physicochemical attributes of the chalcone synthase protein sequences from 29 different plant species.

	Physicochemical properties					
Plant Species	M. Weight (da)	Instability Index	Aliphatic Index	Hydropathicity (GRAVY)	Length	pI
Arabidopsis thaliana	43115.72	32.16	92.08	-0.074	395	6.08
Camelina microcarpa	42867.38	31.36	91.32	-0.083	393	6.03
Halimolobos perplexus var. Perplexus	43117.84	31.40	92.58	-0.051	395	6.20
Capsella rubella	43117.80	32.31	90.61	-0.080	395	6.32
Murbeckiella boryi	42829.44	31.16	91.58	-0.077	393	6.32
Crucihimalaya himalaica	43029.78	32.29	92.08	-0.044	395	6.47
Alyssopsis mollis	42831.41	31.97	91.83	-0.076	393	6.20
Turritis glabra	43046.81	31.47	92.58	-0.038	395	6.38
Boechera fendleri	43014.67	30.45	92.10	-0.056	395	6.32
Boechera stricta	43027.66	30.45	92.10	-0.063	395	6.32
Descurainia paradise	42815.41	31.57	91.58	-0.078	393	6.32
Ballantinia antipoda	42325.86	31.44	90.49	-0.072	388	6.15
Yinshania acutangula	42825.40	34.07	91.83	-0.085	393	6.09
Malcolmia graeca	42864.44	32.59	90.33	-0.085	393	6.20
Crucihimalaya wallichii	42341.90	31.66	90.23	-0.079	388	6.28
Stenopetalum lineare	42377.91	30.60	90.98	-0.076	388	6.09
Cuphonotus andraeanus	42337.85	31.68	90.75	-0.088	388	6.15
Crucihimalaya mollissima	42313.84	31.88	89.74	-0.086	388	6.28
Descurainia sophia	42797.33	33.28	93.31	-0.085	393	6.20
Transberingia bursifolia (Partial)	42283.82	30.62	90.00	-0.079	388	6.28
Harmsiodoxa puberula	42399.85	32.01	89.74	-0.102	388	6.03
Malcolmia maritime	42847.33	33.46	89.62	-0.112	393	6.20
Descurainia preauxiana	42357.85	31.31	89.74	-0.103	388	6.15
Menkea villosula	42402.94	29.69	90.23	-0.102	388	6.15
Erysimum pseudorhaeticum	42913.43	31.10	90.36	-0.099	394	6.13
Smelowskia tibetica	42780.28	34.04	90.84	-0.090	393	6.15
Brachypus suffruticosus	42758.36	32.76	91.05	-0.088	392	6.20
Ionopsidium abulense	42936.72	31.20	91.60	-0.080	394	6.32
Harmsiodoxa brevipes	42343.74	30.63	89.74	-0.099	388	5.91

Notes: The physicochemical properties of all the chalcone synthase genes' determination used the ProtParam database of the Swiss Bioinformatics Resource Portal ExPasy. These physicochemical properties include molecular weight, instability index, aliphatic index, hydropathicity (GRAVY), length, and isoelectric point (pI). The values of each of these properties of all the selected plant species were relatively close.

Ballantinia antipoda, Crucihimalaya wallichii, Stenopetalum lineare, Crucihimalaya mollissima, Transberingia bursifolia (Partial), Harmsiodoxa puberula, Descurainia preauxiana, Menkea villosula, and *Harmsiodoxa brevipes CHS* genes were all 388 amino acids in length, which is the lowest as compared with other selected plant species. *Camelina microcarpa, Murbeckiella boryi, Alyssopsis mollis, Descurainia paradise, Yinshania acutangula, Malcolmia graeca, Descurainia sophia, Malcolmia maritima, Smelowskia tibetica CHS* genes were 393 amino acids in length. *Erysimum pseudorhaeticum* and *Ionopsidium abulense* had a range of 394 amino acids, and *Brachypus suffruticosus* had a reach of 392 amino acids. For *Arabidopsis thaliana, Halimolobos perplexus var. Perplexus, Capsella rubella, Crucihimalaya himalaica, Turritis glabra, Boechera fendleri,* and *Boechera stricta* all had a maximum length of 395 amino acids. *Harmsiodoxa brevipes CHS* had the lowest isoelectric point (pI) value of 5.91 versus the highest pI value of 6.47 for the CHS protein of *Crucihimalaya himalaica* (Table 2).

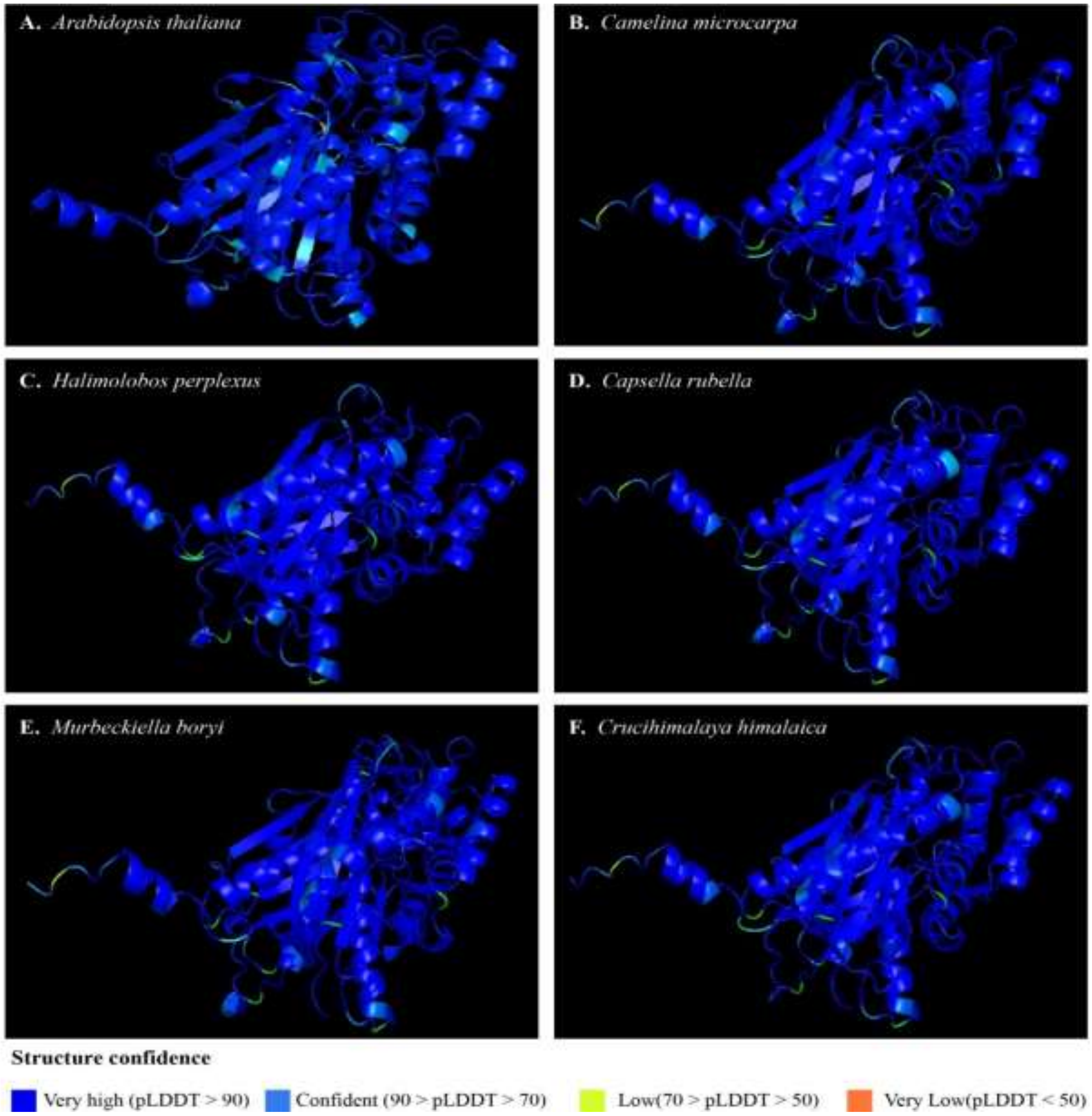


Figure 4. The 3D structures of chalcone synthase.

The predicted 3D structures of *Arabidopsis thaliana* (A)*, Camelina microcarpa* (B)*, C. Halimolobos perplexus* (C)*, Capesulla rubella* (D)*, Murbeckiellaboryi* (E), and *Crucihimalaya himalaica* (F). Different colors indicate confidence, ranging from very high (blue), confident (light blue), low (yellow), and very low (orange) confidence. The pLDDT corresponds to the model's prediction of its score on the Local Distance Difference Test (LDDT-Cα).

3D Protein structures

The downloaded 3D protein structures of chalcone synthase genes for 29 plant species came from UniProt https://www.uniprot.org/ (UniProt, 2022). These included original, as well as, predicted 3D structures where applicable. Structure source files came as PDB files. The structures continued further analysis in pymol (DeLano, 2002) and saved them as JPEG files. As expected, each other's 3D structures remained highly conserved and similar. Figure 4 shows the 3D rendering of six representative proteins. Structure confidence statistics' calculation for the predicted structures appears in Figure 4. The different colors in the protein cartoon structure indicate very high, confident, low, and least confidence. Blue shows very high confidence, light blue represents confidence (90 > pLDDT > 70), green represents low, and orange represents the minimum level of structural confidence. As distinct from the color scheme in Figure 4, all the 3D structures generated had a very high (pLDDT > 90) confidence rate.

DISCUSSION

Chalcone synthase (CHS) is a key enzyme in the flavonoid/isoflavonoid biosynthesis pathway. Besides regulating plant developmental programs, the *CHS* gene expression surfaces in plants under unfavorable biotic and abiotic stress conditions, such as, bacterial or fungal infections or UV light. The chalcone synthase gene causes the accumulation of flavonoid and isoflavonoid phytoalexins and also contributes to the salicylic acid defense pathway (Dao *et al.*, 2011; Aziz *et al*., 2016; Mohsin *et al*., 2023). This study analyzed 29 chalcone synthase genes of different plant species using various *in silico* techniques to determine their diverse characteristics and the evolutionary relationship between the highly contrasting 29 varied flowering plants. Our initial search using the different keywords and sequence-based blast results identified numerous *CHS* genes across the kingdom Plantae. However, the selected 29 species represent plants from almost every major taxonomic group of the flowering plants. Likewise, the study considered the available high-throughput DNA and protein sequences for the genes to allow robust analysis. The *CHS* genes of all the selected 29 plant species were participants in every analysis performed, with the *Arabidopsis thaliana CHS* gene chosen as a baseline for searching the *CHS* genes because it is a model organism for studying plant biology due to its relatively simple genetics, short life cycle, and well-characterized genome. Arabidopsis has a unique place in the plant kingdom, representing an eudicot model system. By comparing the *CHS* gene and its motifs in Arabidopsis with those in other plant taxa, researchers can gain insights into the evolution of the *CHS* gene family, the conservation of functional elements, and the adaptations that have occurred in different lineages. By leveraging the knowledge from Arabidopsis CHS, we can unravel the broader implications of *CHS* genes in plant biology and evolution, contributing to a comprehensive understanding of flavonoid biosynthesis and its diverse roles across the plant kingdom.

The constructed traditional phylogenetic tree showed a close association between all the different plant species, indicating their evolutionary origin from a common ancestor. It has further support from homology of more than 95% between the amino acid sequences of all the *CHS* genes. As such, protein homology carries higher importance than nucleic acid-based homology, specifically in the function of different genes involved in chief physiological processes. Protein sequences also have more diversity and complexity than DNA sequences and can provide more information about the structure and function of genes. Additionally, proteins are more conserved than DNA sequences because they are subject to robust selective pressures. Therefore, protein sequences are less likely to converge than DNA sequences, meaning that high similarity between any two proteins always indicates homology (common ancestry), whereas high similarity between any two DNA sequences may be due to chance or convergent evolution indicating independent origin (Panchen, 1999; Sommer, 2008; Scotland, 2010). The significantly higher homology of the *CHS* genes from entirely different plant families compelled us to investigate more deeply and identify highly conserved domains.

The value of conserved domains in protein sequences has gained scrutiny by other scientists for other gene sets. Conserved domains often serve as identifying factors for gene families and as regions of vital importance in the overall function of particular genes (Hao *et al.*, 2021). Deletion, silencing, or even small mutations in the conserved domains often render the proteins non-functional, resulting in significant physiological consequences (Wright *et al.*, 2022). The research found eight conserved regions, which we labeled A, B, C, to H. Among these, CR C appeared as the extensive conserved region in the middle of the gene, spanning 78 total amino acids. These conserved domains show the similarities of the chalcone synthase gene between these 29 selected plant species. An in-depth analysis further identified three highly conserved motifs in all the *CHS* genes from the 29 flowering plants with significantly higher confidence. Motif sites often serve as docking sites for connecting other essential functional regulators. Conserved motifs in the CHS sequence correspond to crucial regulatory elements that include binding sites for different transcription factors in the promoter regions. Identification of these motifs allows researchers to unravel the transcriptional control of *CHS* genes under basal, as well as, conditions induced by specific biotic and abiotic stresses.

On the other hand, conserving these motifs across different plant species also provides an insight into the evolutionary history of flavonoid biosynthesis, revealing common functional elements that remained conserved over time. It not only advances the understanding of plant molecular biology and evolution but also provides a base for the practical application of this information via functional genomic and biotechnology approaches for crop improvement. The physicochemical properties of the chalcone synthase gene in all 29 species were significantly similar following the analysis on the ExPASY server (Gasteiger *et al.*, 2003). The assessment indicated that the *CHS* gene family functions under various cellular environments with a significantly higher indication of functional redundancy, which means that multiple *CHS* genes work together to regulate physiological pathways. The 3D structural analyses were uniquely interesting. We determined the 3D protein structures of the chalcone synthase gene of six plant species among the selected 29 plant species. The 3D structures of each were significantly similar, with a higher confidence value indicating functional similarity and redundancies between the different members of the *CHS* gene family. Further in-depth analysis and practical genomics studies need elucidation of the role of this highly significant gene family in plant development and stress responses.

CONCLUSIONS

The study of the chalcone synthase gene across 29 different flowering plant species revealed that the *CHS* gene of selected plant species remained highly conserved. The phylogenetic tree of the *CHS* gene divides into at least six closely related rooted clades and protein sequence alignment, with eight conserved regions both showing a high degree of similarity, indicating that the gene has been conserved over a long time and across a wide variety of species. The three conserved motifs identified in the motif analysis further supported the idea of a highly conserved gene. The physicochemical properties of the protein sequences were also evident to be quite similar across the selected plant species, suggesting that the protein plays an influential role in the biology of these plants. The 3D protein structures constructed were significantly similar to each other with a higher confidence value, indicating functional similarity and redundancies between the different members of the *CHS* gene family. It can provide a basis for future studies investigating the practical significance of the conserved regions. Overall, this study provides valuable insights into the evolution and conservation of the chalcone synthase gene in flowering plants and has significant implications for understanding the biology of these species.

REFERENCES

Austin MB, Noel JP (2003). The chalcone synthase superfamily of type III polyketide synthases. *Nat. Prod. Rep.* 20(1): 79-110. https://doi.org/10.1039/b100917f.

Aziz SA, Azmi TKK, Sukma D, Qonitah FZ (2016). Morphological characters of triploids and tetraploids produced by colchicine on buds and flowers of *Phalaenopsis amabilis*. *SABRAO J. Breed. Genet*. 48(3): 352-358.

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS (2009). MEME SUITE: Tools for motif discovery and searching. *Nucleic acids research.* 37 (Web Server issue): 202-208. https://doi.org/ 10.1093/nar/gkp335.

Blokland RV, Geest NVD, Mol JN, Kooter JM (1994). Transgene-mediated suppression of chalcone synthase expression in *Petunia hybrida* results from an increase in RNA turnover. *Plant Journal.* 6861-877. https://doi.org/10.1046/j.1365-313X.1994.6060861.x.

Dao TTH, Linthorst HJM, Verpoorte R (2011). Chalcone synthase and its functions in plant resistance. *Phytochem. Rev.* 10(3): 397-412. https://doi.org/10.1007/s11101-011-9211-7.

DeLano WL (2002). Pymol: An open-source molecular graphics tool. *CCP4 Newsl. Protein Crystallogr.* 40(1): 82-92.

Falcone Ferreyra ML, Rius SP, Casati P (2012). Flavonoids: Biosynthesis, biological functions, and biotechnological applications. *Front Plant Sci.* 3222. https://doi.org/10.3389/fpls.2012.00222.

Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A (2003). ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic acids research.* 31(13): 3784-3788. https://doi.org/10.1093/nar/gkg563.

Gasteiger E, Hoogland C, Gattiker A, Duvaud SE, Wilkins MR, Appel RD, Bairoch A (2005). Protein identification and analysis tools on the ExPASy server. In: J.M. Walker (ed.). Springer Protocols Handbooks: Humana Press. https://doi.org/10.1385/1-59259-890-0:571.

Hall TA (1999). "BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT," in: *Nucleic acids symposium series.* 41. 95-98.

Hao Y, Zong X, Ren P, Qian Y, Fu A (2021). Basic Helix-Loop-Helix (bHLH) transcription factors regulate a wide range of functions in Arabidopsis. *Int J Mol Sci.* 22(13): https://doi.org/10.3390/ijms22137152.

Hu B, Jin J, Guo A-Y, Zhang H, Luo J, Gao G (2014). GSDS 2.0: An upgraded gene feature visualization server. *Bioinformatics.* 31(8): 1296-1297. https://doi.org/10.1093/ bioinformatics/btu817.

Mohsin RM, Abd Asal KN, Kamaluddin AA, Zaky AA (2023). Genotypes and storage duration effects on the quality of cut flower - gerbera (*Gerbera jamesonii* Hook). *SABRAO J. Breed. Genet*. 55(1): 260-267. http://doi.org/10.54910/sabrao2023.55.1.24.

Napoli C, Lemieux C, Jorgensen R (1990). Introduction of a chimeric chalcone synthase gene into Petunia results in reversible co-suppression of homologous genes in trans. *Plant Cell.* 2(4): 279-289. https://doi.org/ 10.1105/tpc.2.4.279.

Panchen AL (1999). "Homology — History of a Concept," in *Novartis Foundation Symposium 222 - Homology*. 5-23. https://doi.org/10.1002/9780470515655.ch2.

Pang Y, Shen G, Wu W, Liu X, Lin J, Tan F, Sun X-F, Tang K (2005). Characterization and expression of chalcone synthase gene from *Ginkgo biloba*. *Plant Sci.* 168(6): 1525-1531. https://doi.org/10.1016/j.plantsci. 2005.02.003.

Saitou N, Nei M (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution.* 4(4): 406-425. 10.1093/ oxfordjournals.molbev.a040454.

Schmitzer V, Veberic R, Osterc G, Stampar F (2010). Color and phenolic content changes during flower development in groundcover rose. *J. Amer. Soc. Hort. Sci.* 135(3): 195-202. https://doi.org/10.21273/jashs.135.3.195.

Scotland RW (2010). Deep homology: A view from systematics. *BioEssays.* 32(5): 438-449. https://doi.org/10.1002/bies.200900175.

Sommer RJ (2008). Homology and the hierarchy of biological systems. *BioEssays.* 30(7): 653-658. https://doi.org/10.1002/bies.20776.

Strack D, Vogt T, Schliemann W (2003). Recent advances in betalain research. *Phytochemistry.* 62(3): 247-269. https://doi.org/10.1016/S0031-9422(02) 00564-2.

Tamura K, Stecher G, Kumar S (2021). MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Mol Biol and Evo.* 38(7): 3022-3027. https://doi.org/10.1093/molbev/msab120.

Tatsuzawa F, Saito N, Toki K, Shinoda K, Honda T (2012). Flower colors and their Anthocyanins in *Matthiola incana* cultivars (Brassicaceae). *J. Jpn. Soc. Hortic. Sci.* 81(1): https://doi.org/10.2503/ jjshs1.81.91.

Thill J, Miosic S, Ahmed R, Schlangen K, Muster G, Stich K, Halbwirth H (2012). 'Le Rouge et le Noir': A decline in flavone formation correlates with the rare color of black dahlia (*Dahlia variabilis* hort.) flowers. *BMC Plant Biol.* 12(1): 225. https://doi.org/10.1186/ 1471-2229-12-225.

Thompson JD, Higgins DG, Gibson TJ (1994). CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research.* 22(22): 4673-4680. https://doi.org/10.1093/nar/22.22.4673.

Uniprot CT (2022). "UniProt: the Universal Protein Knowledgebase in 2023," in: *Nucleic Acids Res.*). https://doi.org/10.1093/nar/gkac1052.

Winkel-Shirley B (2001). It takes a garden. How work on diverse plant species has contributed to an understanding of flavonoid metabolism. *Plant Physiol.* 127(4): 1399-1404. https://doi.org/10.1104/pp.010675.

Winkel-Shirley B (2002). Biosynthesis of flavonoids and effects of stress. *Curr. Opin. Plant Biol.* 5(3): 218-223. https://doi.org/10.1016/s1369-5266(02)00256-x.

Wright CJ, Smith CWJ, Jiggins CD (2022). Alternative splicing as a source of phenotypic diversity. *Nat. Rev. Genet.* 23(11): 697-710. https://doi.org/10.1038/s41576-022-00514-4.

Wu X, Zhang S, Liu X, Shang J, Zhang A, Zhu Z, Zha D (2020). Chalcone synthase (CHS) family members analysis from eggplant (*Solanum melongena* L.) in the flavonoid biosynthetic pathway and expression patterns in response to heat stress. *PloS one.* 15(4): e0226537. https://doi.org/10.1371/journal.pone.0226537.

Zhao D and Tao J (2015). Recent advances on the development and regulation of flower color in ornamental plants. *Front Plant Sci.* 6https://doi.org/10.3389/fpls.2015.00261.

Zhao D, Tao J, Han C, Ge J (2012). Flower color diversity revealed by differential expression of flavonoid biosynthetic genes and flavonoid accumulation in herbaceous peony (*Paeonia lactiflora* Pall.). *Mol. Biol. Rep.* 39(12): 11263-11275. https://doi.org/10.1007/s11033-012-2036-7.

Zuckerkandl E, Pauling L (1965). "Evolutionary divergence and convergence in proteins," in *Evolving Genes and Proteins*. V. Bryson and H.J. Vogel (eds.). Academic Press, 97-166. https://doi.org/10.1016/B978-1-4832-2734-4.50017-6.