Araştırma Makalesi / Research Article

Bioinformatical Analyses of cinnamyl alcohol dehydrogenase (CAD) proteins from higher plant species

Ertuğrul FİLİZ[1], Fırat KURT[*2]

[1]*Duzce University, Cilimli Vocational School, Department of Crop and Animal Production, 81750, Duzce, Turkey*
[2]*Mus Alparslan University, Faculty of Applied Sciences, Department of Plant Production Technologies, Mus, Turkey*

Abstract
Cinnamyl alcohol dehydrogenase (CAD) (EC 1.1.1.195) is an enzyme functioning in the reduction of various phenylpropenyl aldehyde derivatives which are precursors in lignin and lignan production. Species-specific *CAD* genes have been extensively identified in recent years. In this study, we used bioinformatics tools to characterize and classify plant CADs. The amino acid and nucleotide sequences of 16 CADs from different plant species were used to compare their physiological properties, phylogeny, and conserved motifs. For this purpose, sequence, phylogenetical, structural analyses of proteins were conducted using various servers. All plant CADs had the characteristic alcohol dehydrogenase (PF08240) and zinc-binding dehydrogenase domains (PF00107). According to the physicochemical analysis, it was revealed that the most of plant CADs (81.25%) were in acidic character. Sequence length (aa) and molecular weight (kDa) of CAD proteins were found in range of 356 -367 and 38.6-40.5 respectively. The highest sequence similarities were found between *Sorghum bicolor* and *Zea mays* (95.3%), *Panicum virgatum* and *Sorghum bicolor* (90.9%), and *Oryza sativa* and *Zea mays* (87.1%) respectively. Plant *CAD*s showed divergent exon-intron structures in which exon numbers were ranged from two to six. Four monocot species (*S. bicolor, P. virgatum, Z. mays,* and *O. sativa*) have four exons, whereas *Brachypodium distachyon* contains only two exons. Phylogenetic analysis revealed that the CAD proteins mainly divided into two groups. The highest bootstrap values were found as follows: *Fragaria vesca-Prunus persica* clade (100%), *Glycine max-Medicago truncatula* (81%), and *S. bicolor-Z. mays* (72%). The 3D structures of plant CADs showed that *Oryza* and *Vitis* had the most divergent structures when compared to the other plant species. Eventually, the data represented here contribute to studies aiming at evaluating the plant CADs extensively and at identifying new *CAD* genes in other plants.

Keywords: 3D structure, Cinnamyl alcohol dehydrogenase (CADs), comparative phylogenetics, *in silico* analysis

Yüksek Bitki Türlerindeki Sinamil alkol dehidrogenaz (CAD) Proteinlerinin Biyoinformatiksel Analizi

Öz
Sinamil alkol dehidrogenaz (CAD) (EC 1.1.1.195) lignin ve lignin üretimindeki öncül çeşitli fenil propenil aldehit türevlerinin indirgenmesinde görev alan bir enzimdir. Türlere özgü olan *CAD* genleri, son yıllarda önemli derecede tanımlanmıştır. Bu çalışmada CAD genlerinin (enzim veya proteinlerinin) bioinformatik araçlar kullanılarak karakterize edilip, sınıflandırılması amaçlanmıştır. 16 farklı bitki türünden elde edilen CAD nükleotit ve amino asit dizileri fizyolojik özellikler, filogenetik ve korunmuş motif bölgelerinin karşılaştırılması için kullanılmıştır. Bu amaçla CAD proteinlerinin sekans, filojenik ve yapısal analizleri çeşitli sunucular yardımıyla yapılmıştır. Bütün incelenen CAD dizilerinin alkol dehidrogenaz (PF08240) ve çinko bağlayıcı dehidrogenaz domainlerine sahip oldukları gözlenmiştir (PF00107). Fizyokimyasal analiz sonuçlarına göre, CAD'lerin önemli bir kısmının (%81,25'i) asidik karakterde olduğu gözlenmiştir. Bu proteinlerin amino asit uzunlukları (aa) ve moleküler ağırlıklarının (kDa) 356 -367 ve 38,6-40,5 arasında sırasıyla değişmekte olduğu belirlenmiştir. Dizi benzerlikleri en yüksek *Sorghum bicolor* ile *Zea mays* (%95,3), *Panicum virgatum* ile *Sorghum bicolor* (%90,9) ve *Oryza sativa*

[*]Sorumlu yazar: f.kurt@alparslan.edu.tr
Geliş Tarihi: 22.09.2018, Kabul Tarihi: 05.02.2019

ile *Zea mays* (% 87,1) arasında bulunmuştur. İncelenen *CAD* genlerinin intron ve ekzon yapıları birbirlerinden farklılık göstermiş olduğu ve ekzon sayılarının iki ve altı arasında değiştiği belirlenmiştir. Çalışmadaki tek çenekli türler olan *S. bicolor, P. virgatum, Z. mays,* ve *O. sativa'*nın dört ekzona sahip olduğu; *Brachypodium distachyon'*un ise sadece iki ekzona sahip olduğu gözlenmiştir. Filogenetik analiz neticesinde CAD proteinlerinin sadece iki ana gruba ayrıldığı saptanmış; en yüksek bootstrap değerleri sırasıyla şu şekilde bulunmuştur: *Fragaria vesca-Prunus persica* grubu (%100), *Glycine max-Medicago truncatula* (%81), and *S. bicolor-Z. mays* (%72). İncelenen CAD'lerin 3 boyutlu analizlerine göre, *Oryza* ve *Vitis* CAD'leri, araştırmadaki diğer bitki CAD'lerinden en fazla ayrılma göstermiştir. Son olarak bu çalışmadaki veriler, farklı bitkilerdeki CAD genleri veya proteinlerinin tanımlanması ve değerlendirmesini amaçlayan yeni çalışmalara katkı sağlayacaktır.

Keywords: 3 boyutlu yapı, Sinamil Alkol Dehidrogenaz (CAD'ler), karşılaştırmalı filogenetik, bilgisayar simülasyonlu analiz.

1. Introduction

Lignin, the polymer of subunit monolignols, including p-coumaryl, coniferyl, and sinapyl alcohols, is the structural component of cell wall in vascular plants, supporting mechanical resistance against hydrophobicity, plant growth, development, and responses to environmental stresses [1, 2]. The cinnamyl alcohol dehydrogenase is used in the reduction of cinnamaldehydes into cinnamyl alcohols in the last step of monolignol biosynthesis in the cell wall before oxidative polymerization [3, 4].

CAD exhibits different features between gymnosperms and angiosperms. Gymnosperm *CAD* is encoded by single gene with highly characteristic for coniferyl aldehyde, whereas angiosperm *CAD* is encoded by multiple genes having crucial affinity for coniferyl and sinapyl aldehydes [5]. The *CAD* genes display nearly 80% and 70% nucleotide sequence identity in all published angiosperm and angiosperms & gymnosperms sequences [6].

CAD and *CAD-like* genes have been reported in many plant genomes, including *Populus trichocarpa*, [7] *Oryza sativa*, [8] *Eucalyptus globules*, [9] *Arabidopsis thaliana* [10], wheat [5], sorghum, [11] maize, [12] *Picea abies*,[13] and *Lolium perenne* [14]. Nonetheless, in *Arabidopsis*, 9 *CAD* genes were identified in *CAD* multigene families. Among the *AtCAD* genes, only *AtCAD1*, *AtCAD4* and *AtCAD5* were found to be related with lignin biosynthesis [3, 15, 39]. The anatomical parts of AtCAD4 and AtCAD5 differs; as AtCAD4 is primarily expressed in leaves and flowers, AtCAD5 expression particularly higher in roots [38]. In the rice genome, 12 distinct genes showed higher similarity to *CAD* genes [8]. It was reported that AtCAD4 and 5 were phylogenetically grouped in the same clad with *bona fide* ZmCAD2 (maize), OsCAD2 (rice), SbCAD2 (sorghum) and BdCAD5 (*Brachypodium*). BdCAD5 and BdCAD3 had similar tertiary structures with AtCAD5; however, in terms of kinetic parameters, BdCAD5 was more involved in lignin biosynthesis [40]. In wheat, a total of 11 wheat CAD sequences were identified within 6 groups based on the phylogenetic analysis. *TaCAD1* is very similar to the other *bona fide CADs* in lignin synthesis owing to resemblance of amino acid sequence and three-dimensional structure [5]. As for TaCAD12, it was suggested to involve in plant defense system against *Rhizoctonia cerealis* [39]. Presence of 15 *PoptrCAD* genes were reported in poplar [7].

Based on gene structure analysis, three patterns were detected and 14 of the 15 *CAD* genes distributed on duplicated regions. Also, *CAD* gene expressions exhibited different patterns. In sorghum, 14 *CAD-like* genes at seven different loci were identified. Phylogenetic analysis showed that *SbCAD* genes clustered into four groups. *SbCAD2* groups were similar to *bona fide CADs* from other species [11]. Jun et al. [41] reported that SbCAD2 and SbCAD4 have high structural and functional resemblance with AtCAD5. Consequently, in this study we performed *in silico* analysis of cinnamyl alcohol dehydrogenase nucleotide and protein sequences from higher plant species to characterize and classify *CAD* genes. For this purpose, we also include comparative motif and gene structure, physiochemical, and phylogenetic analyses.

2. Materials and methods

2.1. Sequence database searches

CAD protein sequences were collected from NCBI protein database (http://www.ncbi.nlm.nih.gov/protein) by using BLASTP program. The CAD sequences of nine *Arabidopsis* [15] and 12 *Oryza* [8] were retrieved from TAIR (http://www.tair.org) and TIGR (http://www.tigr.org), respectively and they were used as queries. The sequences were selected as predicted proteins if their E-value satisfied smaller than e^{-10}. Also, all candidate sequences were analyzed in the Pfam database [16] to detect alcohol dehydrogenase (PF08240) and zinc-binding dehydrogenase (PF00107) domains. Thus, 16 higher plant species (*Arabidopsis thaliana, Brachypodium distachyon, Sorghum bicolor, Panicum virgatum, Zea mays, Oryza sativa* indica, *Solanum lycopersicum, Glycine max, Citrus sinensis, Vitis vinifera, Fragaria vesca, Prunus persica, Cucumis sativus, Ricinus communis, Medicago truncatula,* and *Populus trichocarpa*) were used to analyze CAD protein sequences. The nucleotide sequences of plant CAD proteins were obtained from the Joint Genome Institute (JGI) (http://www.phytozome.net).

2.2. Prediction of conserved motifs and gene structures of CADs

Physiochemical data were generated from the Expasy'sProtParam server [17] including sequence length, molecular weight, and theoretical isoelectric point (*pI*) values. Conserved motifs of CAD proteins were identified by using MEME suite (http://meme.sdsc.edu/meme4_4_0/intro.html) [18]. The following parameters were adopted: the optimum motif width was set to ≥6 and ≤50; the maximum number was set to identify 15 motifs. The sub-cellular distribution and potential N-glycosylation sites of the CAD proteins were predicted by using TargetP 1.1 (http://www.cbs.dtu.dk/services/Target/) [19] and NetNGlyc 1.0 Server (http://www.cbs.dtu.dk/services/NetNGlyc/). A structural figure of *CAD* genes, including exon and intron numbers, was determined using the Gene Structure Display Server (GSDS) (http://gsds.cbi.pku.edu.cn/) [20]. All CAD protein sequences were aligned with the ClustalW multiple sequence alignment tool. Full protein sequences were taken to display the consensus sequence analysis. Weblogo 3 program was used to compare the conserved motifs of the species [21, 22] .Interacting partners of *Arabidopsis* CAD and its co-expressed genes were predicted using String 9.1 software (http://string-db.org/) [23].

2.3. Secondary and tertiary structures analysis

Secondary and tertiary structures of CAD proteins were predicted by using a web-based tool, PSIPRED v2.5 (http://bioinf.cs.ucl.ac.uk/psipred/) and BioSerf (http://bioinf.cs.ucl.ac.uk/psipred/?bioserf=1) [24, 25]. The tertiary structures of the most divergent CADs were compared to analyze the structural and possible functional differences. Swiss-PdbViewer (DeepView v4.1) program was used to design the CAD protein models (https://swissmodel.expasy.org)[26].The stereochemical qualities of the modeled proteins wereevaluated by RAMPAGE server [27].

2.4. Phylogenetic analysis

Amino acid sequences of the CAD proteins were aligned using Clustal W [28]. Phylogenetic analysis were performed by MEGA 5.1 program [29] using a neighbour-joining (NJ) tree method, based on the multiple sequence alignment with following parameters: Poisson correction, pair-wise deletion, and bootstrap analysis with 1000 replicates.

3.Results and discussion

3.1. Physicochemical analysis

We used totally 16 CAD protein sequences from 16 different plant species for *in silico* comparative analysis (Table 1). Physicochemical analysis showed that many CAD proteins (81.25%) were in acidic

character (*pI*≤7), while only three proteins belonging to *S. lycopersicum*, *G. max*, and *P. persica* had in basic character (*pI*≥7). The average molecular weights and sequence lengths of CADs were calculated as 39.10 kDa and 361 amino acids, respectively. Previous studies are in agreement with our findings, including *Cameliasinensis* [30], strawberry [31], *Eucalyptus globules* [9] and *Pyrus bretschneideri* [38]. Sub-cellular localizations were predicted and only three of 16 CAD proteins were found to be resided in mitochondria (*S. bicolor, and Z.mays*), and chloroplast (*M. truncatula*).

N-linked glycosylation cause a basic post-translational modification over proteins with formation of a covalent bounding on asparagine residues owing to oligosaccharide attachment in the polypeptide chains. The N-X-S/T consensus sequence is known as a general recognition element [32]. In our study, nine plants contained putative N- glycosylation sites (Table 1) suggesting that these sites may regulate protein structures of CADs with relation to various metabolic or physiological conditions. Pfam analyses revealed that all plant CADs had alcohol dehydrogenase GroES-like (PF08240) and zinc-binding dehydrogenase domains (PF00107) (Figure 1, Figure 2).

Table 1. Characteristics of CADs in higher plant species, including ORF length, exon and intron number, Pfam family, protein sequence length, molecular weight, predicted subcellular localizations, and N-glycosylation sites.

Species	Accession Num.	ORF length (bp)	Exon Num.	Pfam family	Seq. length (aa)	M. wt. (Da)	pI	SL	N-glycosylation sites
B. distachyon	XP_003581549	1098	5	Alcohol dehydrogenase	361	39.14	6.15	-	6 NHTQ, 82 NVST
S.bicolor	BAJ09366	1086	2	Alcohol dehydrogenase	365	38.65	5.84	M	-
P. virgatum	ACX50973	1098	4	Alcohol dehydrogenase	364	38.69	5.84	-	-
Z. mays	CAA74070	1104	4	Alcohol dehydrogenase	367	38.74	5.95	M	-
O.s. indica	ABB04029	1104	4	Alcohol dehydrogenase	363	38.64	5.94	-	26 NYTL
S. lycopersicum	XP_004233212	1092	4	GroES-like domain	360	39.56	8.2	-	83 NVSH, 258 NHSL
G. max	AEI54337	1083	5	GroES-like domain	360	39.02	7.09	-	320 NITA
C. sinensis	ABM67695	1083	5	GroES-like domain	357	38.93	5.88	-	-
V. vinifera	XP_002273147	1074	5	GroES-like domain	357	38.97	6.01	-	108 NQSL
F.vesca	XP_004291336	1074	6	GroES-like domain	370	40.54	6.16	-	86 NFSV, 271 NGTM, 320 NITC
P. persica	EMJ23293	1122	6	GroES-like domain	361	39.30	7.24	-	7 NHTQ, 271 NGTM, 320 NITC
C. sativus	XP_004166963	1086	6	GroES-like domain	356	39.28	5.91	-	179 NKSG
R. communis	EEF43600	1071	5	GroES-like domain	357	38.84	5.75	-	-
A. thaliana	NP_188576	1074	5	GroES-like domain	365	39.10	5.33	-	-
M. truncatula	AET03358	1083	5	GroES-like domain	360	38.65	6.16	C	110 NESL
P. trichocarpa	EEE87830	1074	5	GroES-like domain	357	38.96	5.76	-	-

M: Mitochondria, C: Chloroplast

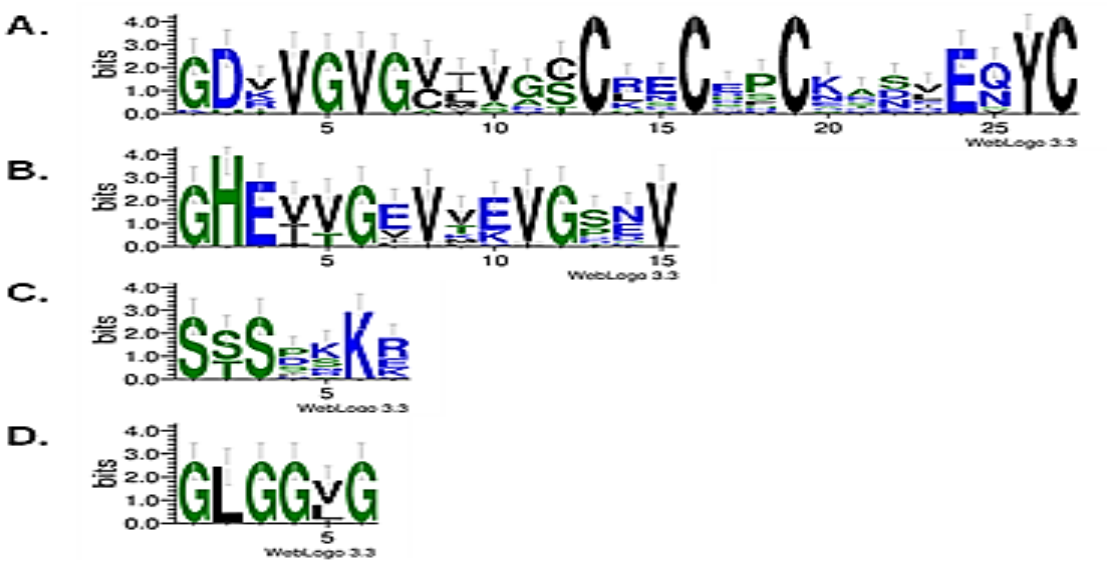


Figure 1. Comparison of critical domains of CAD proteins among 16 plant species. Logo analysis represents Zinc binding domain-Zn2 (A), Zinc binding domain-Zn1 (B), coenzyme specific domain (C), and NADPH binding domain (D).

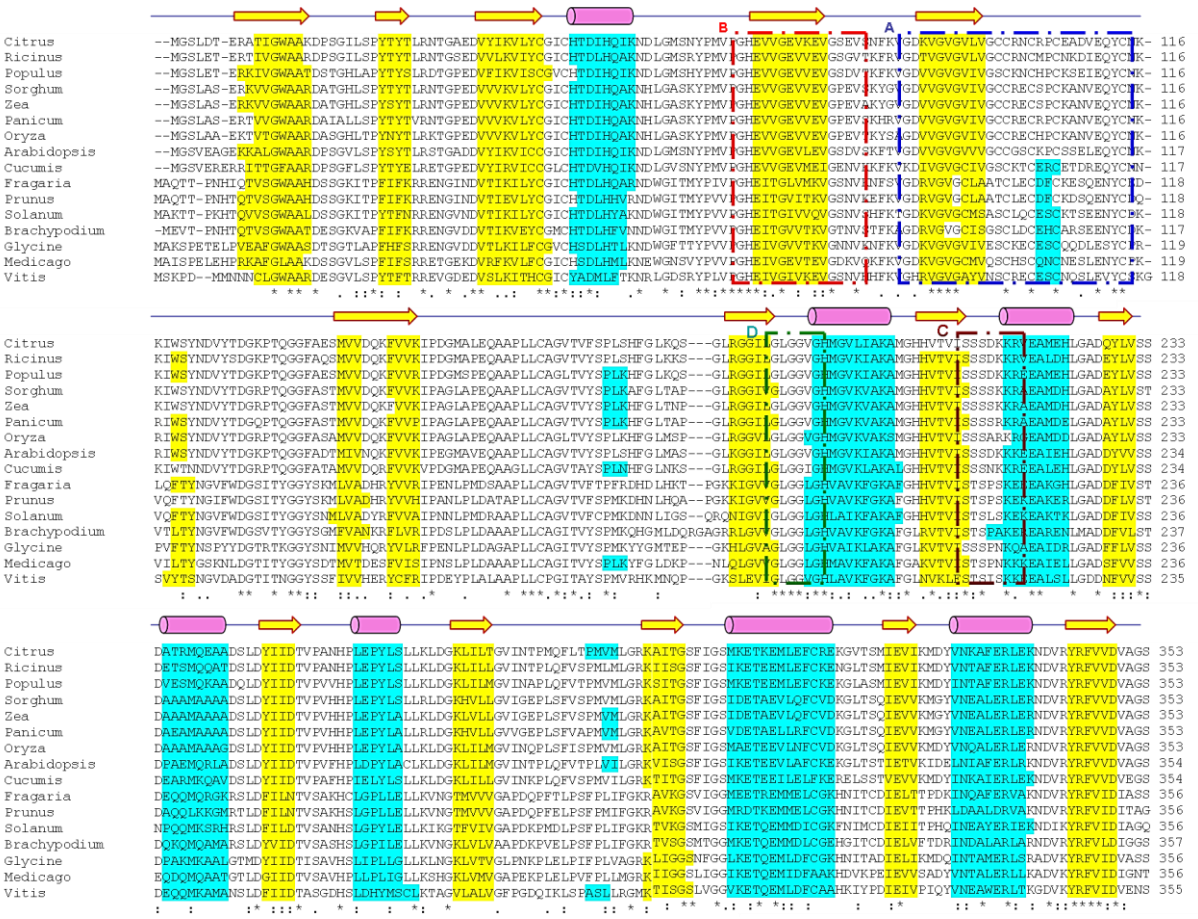


Figure 2. Alignment of 16 CAD proteins belonging to different plant species. Identical residues were labeled with asterisks (∗); similar alternate residues with (:); and dissimilar alternate residues with (.).

α-helix and β-sheet structures were predicted according to PSIPRED program and sequences represent those motifs were labeled with magenta and yellow, respectively. Lines indicated in the upside position shows the predicted loop regions. It gives evident that some CAD proteins' secondary structure differ among others which can affect their three dimensional structures and hence their conformations. Critical domains on CADs were shown into dashed frames: Zinc binding domain-Zn2 (A), Zinc binding domain-Zn1 (B), coenzyme specific domain (C), and NADPH binding domain (D) (Also, for the logo analysis, see in Figure 1).

3.2. The conserved motifs and sequence divergence of plant CADs

Conserved motif analysis of 16 plant CADs revealed that a total of 15 motifs were detected (Table 2 and Figure 3). The motif 1, 2, 3, 4, 5, 6, 7, and 9 were observed in all CADs. The motif 10 was only absent in *P. persica*, while motif 8 was not present in both *B. distachyon* and *S. lycopersicum*. The motif 12 (3 members), motif 13 (9 members), and motif 14 (2 members) were located in the first position. Interestingly, motif 15 was only observed in *Brachypodium* and *Solanum*, whereas motif 14 was only found in *Glycine* and *Medicago*. These unique motifs (motif 12, 13, 14, and 15) in CADs may be related to domain binding structures and be specific to these plants.

Based on the sequence identity matrix data (data not shown), the highest identity values were found among the monocot (grass) species; *Sorghum, Panicum, Zea*, and *Oryza*. The highest identity was observed between *Sorghum-Zea* (95.3%) followed by *Panicum-Sorghum* (90.9%), and *Oryza-Zea* (87.1%), respectively. It can be thought that *CAD* genes are well conserved in monocots. Surprisingly, *Brachypodium* had no higher similarity with the other monocots and this species had the highest identity with *Citrus* (27.7%) followed by *Arabidopsis* (26.8%), and *Vitis* (21.8%), respectively. In the *Brachypodium* genome, pseudo *CAD* genes may cause this identity. *ArabidopsisCAD* (*AtCAD4*), *bona fide* CAD, is related to lignin biosynthesis [10] and showed the highest identity with *CucumisCAD* (70.1%) suggesting that *CucumisCAD* gene may take part in lignifications. In dicots, the highest identity value was found between *Citrus* and *Ricinus* (86.2%), followed by *Populus-Ricinus* (84.8%), and *Prunus-Fragaria* (81.8%) respectively.

3.3. Predicted secondary and tertiary structure analysis of plant CADs

To compare the secondary and tertiary structures of plant CADs, the amino acid sequences were aligned to each other and possible critical domains were shown. Recent reports indicate that CAD proteins are composed of four critical domains: structural zinc binding domain-2 (Zn2), catalytic zinc binding domain-1 (Zn1), coenzyme specific domain, and NADPH binding domain. In this study, fourcritical domains were shown in the secondary and tertiary structure analysis (Figure 1 and Figure 5). NADPH binding domain, comprised of GLGGLG residues in TaCAD12, was reported to bond with pyrophosphate group of NADP [39].

Also, some CAD proteins' secondary structures differ affecting their three dimensional structures and hence their conformations. According to the Ramachandran plot obtained with the RAMPAGE server, 89.1% and 88.7% residues in favoured region, 9.2% and 8.7% in allowed region, and 1.7% and 2.6% in outlier region in *Oryza* and *Vitis*, respectively, indicating that the 3D models were reliable and good quality.

It is important to note that the secondary structure of zinc binding domain-Zn2 (Figure 4) was variable among plant CADs. For instance, half of the plant CADs was composed of only single β-sheet structure while the remaining plant CADs included β-sheet with an additional α-helix structure (Figure 1). Since this domain is very critical for the enzyme activity, the conformational alteration of CAD proteins may affect the enzyme activity positively, or *vice versa*. Also, in the phylogenetic analysis of CAD proteins, except the *C. sativus*, plants with an additional α-helix structure were grouped into the same clade (Figure 8). The similar domain organizations can make the plant CADs functionally identical.

As it was stated in Figure 2, the most divergent CAD genes were found between *O. sativa* and *V. vinifera*. To compare the tertiary structures of those CADs, we modeled their three dimensional configurations (Figure 4A and 4B). It was observed that *V. vinifera* CAD (*Vv*CAD) has a complete α-helix (Helix-A) structure between the two main dimeric structures where may serve as a catalytic site.

Instead, *O. sativa* (*Os*CAD) has a fragmented α-helix (Helix-A) structure and this critical structural difference may cause functional differentiation between the plant CADs. Also, the models exhibited that an additional β-sheet structure was resided in the Zn-binding motif of *Os*CAD protein. However, *Vv*CAD did not (Figure 5). This may affect the enzyme-substrate interaction and consequently the enzymes' activity. Moreover, six residues (Gly68, His69, Glu70, Gly73, Gly79, and Val82) were found to be interacting with Zn ion (Figure 5). Although the residues were exactly similar between the *Vv*CAD and *Os*CAD, their locations were found to be slightly different that may be the consequence of one-base insertion or deletion in the grapevine CAD protein, or *vice versa*.

Table 2. The most conserved protein motifs in CAD protein sequences of different plant species

Motif	Width	Protein sequences
1	50	PYTYTRRNTGPEDVTIKVLYCGICHTDIHQAKNDWGMSMYPMVPGHEIV
2	50	NDVYWDGRPTQGGFASMMVVDQRFVVRIPDNMPPEQAAPLLCAGVTVY
3	50	LGGVGHMAVKFAKAMGHHVTVISSSPKKREEAMEHLGADDYLVSSDQQ
4	50	GSFIGSMKETQEMLEFCKEHNITCQIEVIKMDYINEAWERLERNDVRYRF
5	50	QAAADSLDYIIDTVPAHHPLEPYLSLLKLDGKLILMGVINQPLQFPSPML
6	36	NVSKFKVGDRVGVGCIVGCCRECEPCKQNQEQYCNK
7	15	RTVFGWAARDPSGHL
8	15	HFGLTQPGLRGGILG
9	8	MLGRKAIT
10	8	VIDVAGSN
11	6	EVVEVG
12	8	MAQTTPNH
13	6	MGSLES
14	8	MAKSPETE
15	8	QNGMGDQR

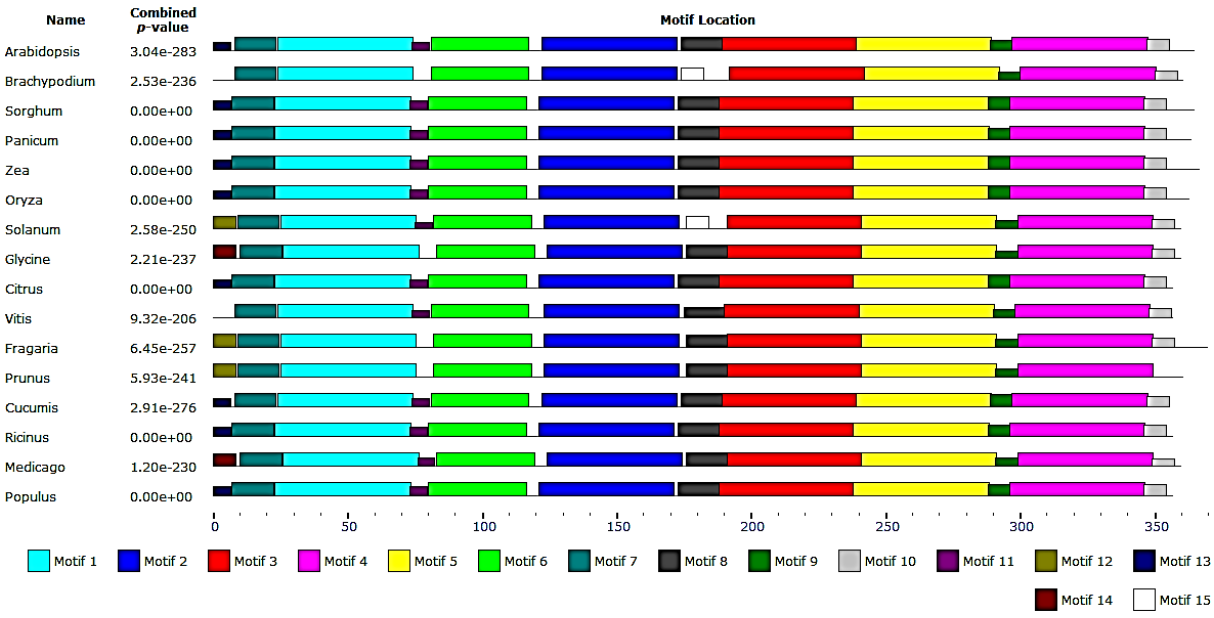


Figure 3. Schematic representation of conserved motifs of plant CADs by MEME server

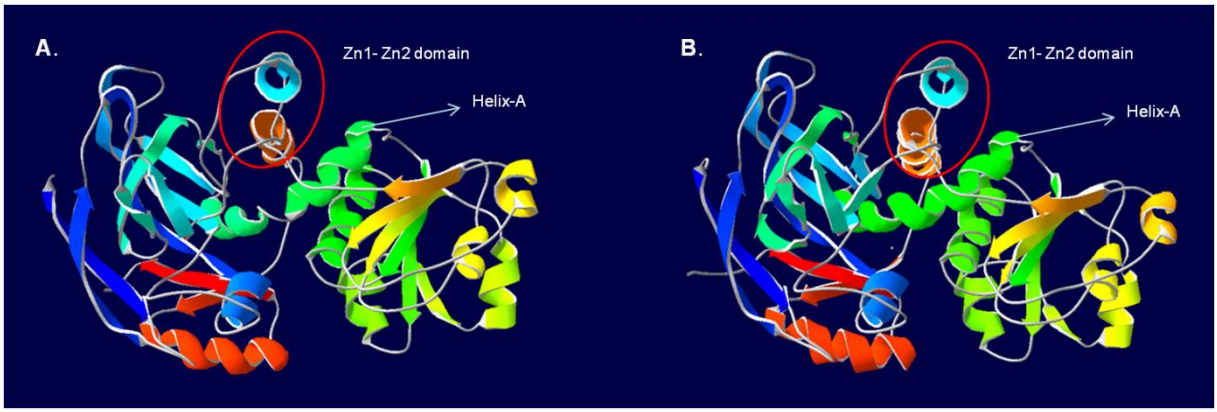


Figure 4. Tertiary structures of CAD proteins in *O. sativa* (A) and *V. vinifera* (B) having the most divergent secondary structures among the 16 CAD proteins. The predicted CAD models were obtained using the BioSerf automated homology and *de-novo* modeling server. N-terminal (blue) and C-terminal (red) residues were labeled according to rainbow color mode. Also, a partial α-helix structure can be clearly observable on *Os*CAD protein (A) and a complete α-helix (Helix-A) structure in VvCAD

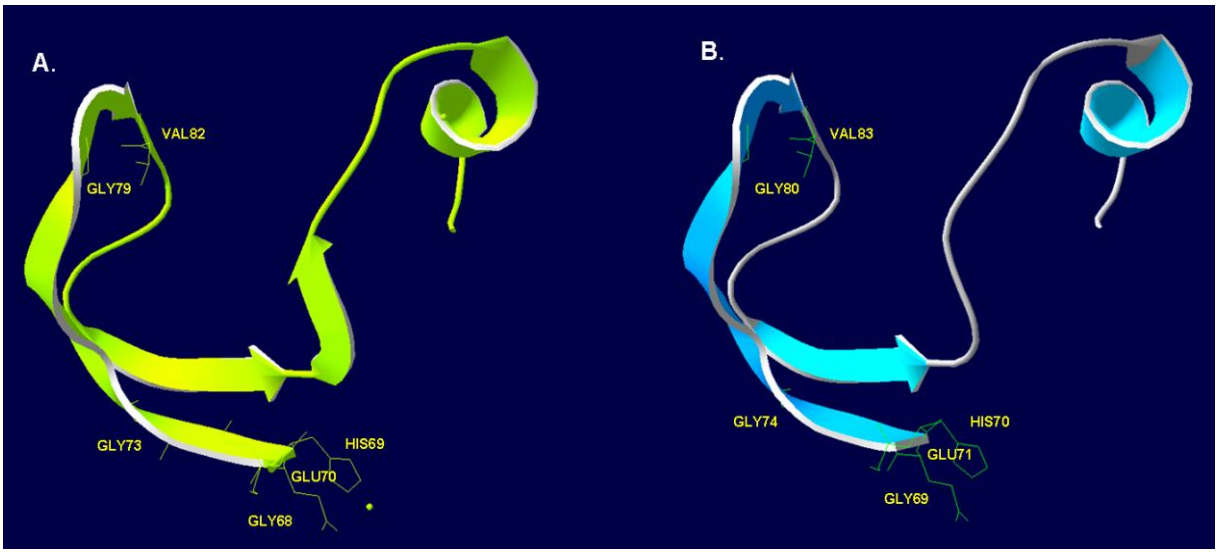


Figure 5. Representation of rice (*O. sativa*) (A) and grapevine (*V. vinifera*) (B) Zn-binding motif structures. Prosite patterns of the CAD proteins were searched for the zinc-binding alcohol dehydrogenase enzyme (PS00059) and the corresponding 6 residues interacting with Zn ions were determined (Gly68, His69, Glu70, Gly73, Gly79, and Val82 for *Os*CAD). Also, it is clearly observable that an additional β-sheet structure was available in *Os*CAD motif (A) which may affect the enzyme-substrate interaction and the enzymes' activity

3.4. The Gene structure and phylogenetic analysis of plant CADs

The ORF length of plant *CAD*s were ranged between 1071 bp (*C. sativus*) and 1122 bp (*F. vesca*) (Table 1). There were no intronless*CAD* genes. Exon numbers were varied in range oftwo and six (Figure 6). The exon numbers was found two (one member), four (four members), five (eight members), and six (three members). All monocots had four exons, except *Brachypodium* (2 exons). Exon-intron structure analysis of *Os*CAD genes revealed that *OsCADs* had two, four, five, and six exons [8]. *BrachypodiumCADs* were similar to *OsCAD8A*, *B*, *C*, and *D* with two exons. In this study, plant CADs exhibited various gene structures with diverse exon-intron numbers. Divergences in exon–intron structure have been observed in duplication events. Structural divergences may generate new protein domain with new biochemical functions [33]. Collectively, CADs in plants may be subjected to gene duplication that causes divergence in their numbers and structures.

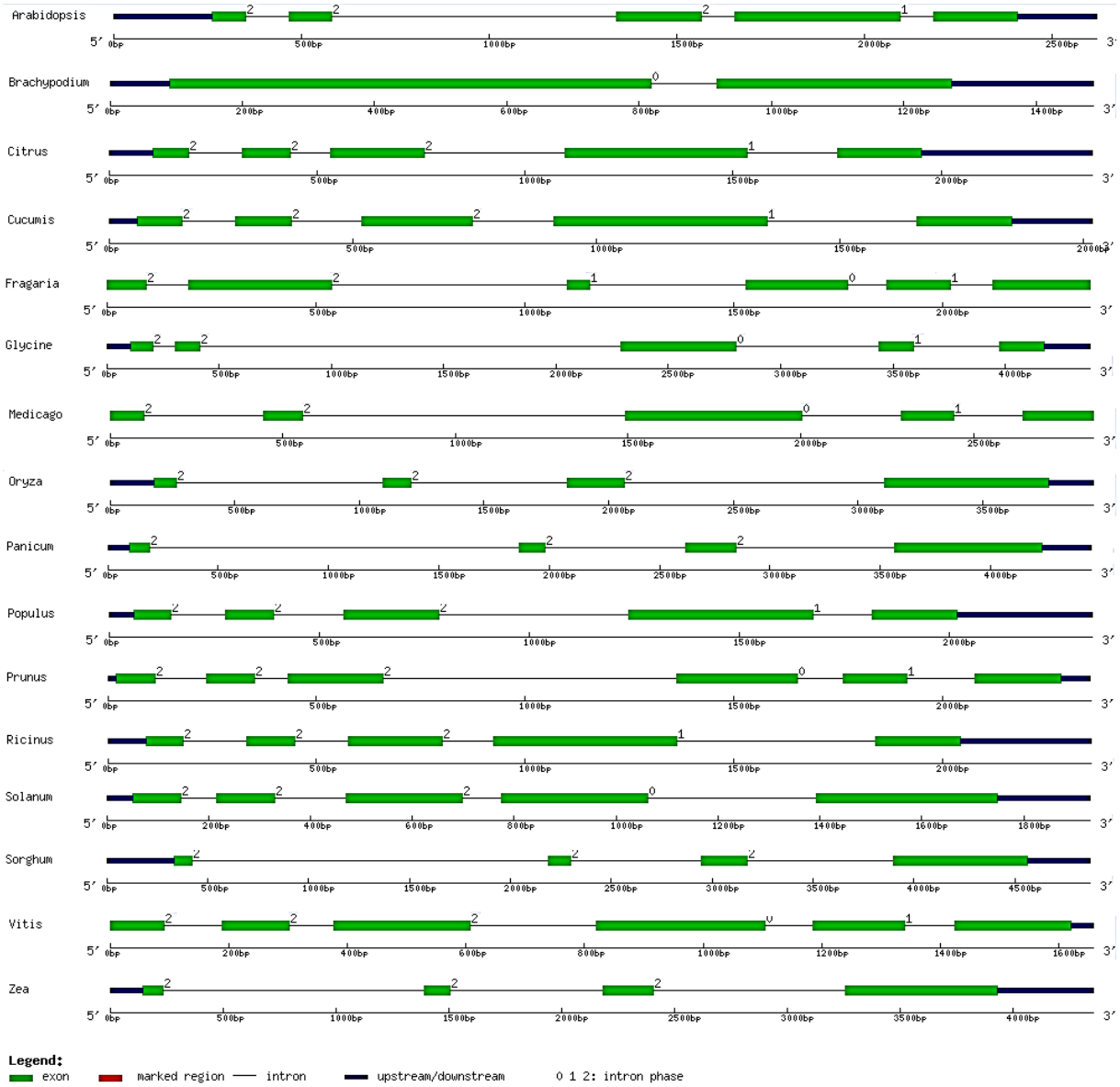
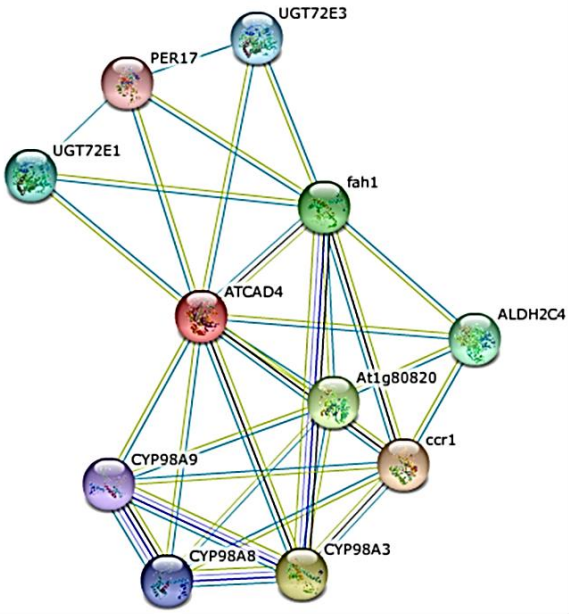
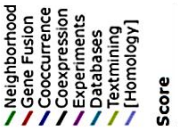


Figure 6. Gene structure of plant *CAD* genes. Exons and introns are depicted by green filled boxes and single lines, respectively. Intron phases 0, 1 and 2 are indicated by numbers 0, 1 and 2 in the figure. UTRs are displayed by thick blue lines at the two ends

The interacting partners of *Arabidopsis* CAD (*AtCAD4*) was predicted using String server and several annotated proteins were found, including cinnamoyl-coA reductase 1 (CCR1), cytochrome P450, cinnamoyl-coA reductase 2 (CCR2), ferulic acid 5-hydroxylase 1 (FAH1), 3-chloroallyl aldehyde dehydrogenase, UDP-glycosyltransferase, peroxidase 17 (PER17) (Figure 7). Cinnamoyl-CoA reductase is a key enzyme in lignin biosynthesis and the cinnamolyl-CoA esters are converted into monolignols by two enzymes with cinnamoyl-CoA reductase (CCR) and cinnamyl alcohol dehydrogenase (CAD) [34]. Ferulate 5-hydroxylase belongs to cytochrome P450-dependent monooxygenase in phenylpropanoid metabolisms and plays important roles in sinapic acid and syringyl lignin biosynthesis [15]. Peroxidases (class III plant peroxidases, EC 1.11.1.7) are the major enzymes involved in the process of monolignol assembly in lignin biosynthesis [35].





Predicted Functional Partners:

ccr1	CCR1 (CINNAMOYL COA REDUCTASE 1); cinnamoyl-CoA reductase; Encodes a cinnamoyl CoA reductase. I [...] (344 aa)				•		• •	0.980
CYP98A3	CYP98A3 (cytochrome P450, family 98, subfamily A, polypeptide 3); monooxygenase/ p-coumarate 3- [...] (508 aa)				•		• •	0.966
At1g80820	CCR2 (CINNAMOYL COA REDUCTASE); cinnamoyl-CoA reductase; Encodes an cinnamoyl CoA reductase iso [...] (332 aa)						• •	0.954
fah1	FAH1 (FERULIC ACID 5-HYDROXYLASE 1); ferulate 5-hydroxylase/ monooxygenase; encodes ferulate 5- [...] (520 aa)				•		• •	0.948
ALDH2C4	ALDH2C4; 3-chloroallyl aldehyde dehydrogenase/ aldehyde dehydrogenase (NAD)/ coniferyl-aldehyde [...] (501 aa)						•	0.906
UGT72E1	UGT72E1 (UDP-glucosyl transferase 72E1); UDP-glycosyltransferase/ coniferyl-alcohol glucosyltra [...] (487 aa)						•	0.904
UGT72E3	UGT72E3; UDP-glycosyltransferase/ coniferyl-alcohol glucosyltransferase/ transferase, transferr [...] (481 aa)						•	0.902
CYP98A8	CYP98A8; electron carrier/ heme binding / iron ion binding / monooxygenase/ oxygen binding; mem [...] (497 aa)						•	0.902
CYP98A9	CYP98A9; electron carrier/ heme binding / iron ion binding / monooxygenase/ oxygen binding; mem [...] (487 aa)						•	0.902
PER17	peroxidase 17 (PER17) (P17); peroxidase 17 (PER17) (P17); FUNCTIONS IN- electron carrier activi [...] (329 aa)						•	0.901

Figure 7. *In-silico* prediction of interacting partners for *CAD*gene of Arabidopsis by using STRING 9.1. The box shows list with putative interacting partners of barley *MoCo sulfurase* gene. STRING automatically highlighted the corresponding nodes in the network and the interactions contain direct (physical) and indirect (functional) associations [23]

Phylogenetic analysis of plant CADs revealed that plant CADs were divided into two main groups (Figure 8). The group I had nine species, while group II were composed of seven species. The most of monocots (4 of 5 species, except *Brachypodium*) was belonged to group I. The highest bootstrap value (100%) was found between *Fragaria* and *Prunus* (in *Rosaceae* family) followed by *Glycine-Medicago* (in *Fabaceae* family) (81%) and *Sorghum-Zea* (in *Poaceae* family) (72%). These clades with the highest bootstrap values share the same gene pool due to their belongingness to the same family. Thus, they have similar genetic background and those clades may be related to functional conservation of *CAD* genes.

It is noteworthy that the monocot *Brachypodium* clustered in group II with *Solanum*, *Fragaria*, and *Prunus* with the highest bootstrap value (100%). This close relationship may be related to similar physiological roles. *Oryza* clustered to *Panicum*, *Zea*, and *Sorghum* clade with the 100% bootstrap value accordingly. The previous studies showed that the monocot CADs were grouped together [7, 8, 11]. This data is consistent with our findings and it may prove that CADs were conserved well in monocots.

In this study, some sequences from various species grouped close to each other in phylogenetic tree. Gene duplication may play critical role for diversification [36]. The copy number variation and changes in gene family size affects genetic variations among closely related species and individuals [37]. Gene duplication events in CADs may result new catalytic functions that can create similar protein sequences.

In conclusion, CAD genes play significant role in various metabolic pathways such as lignification, pathogen defense, growth, and abiotic and biotic stress mechanism [39]. Consequently, the results presented here support to understand *CAD* genes functions in various biological processes in plants and to contribute to *in silico* and experimental studies about CADs involvement to these processes.

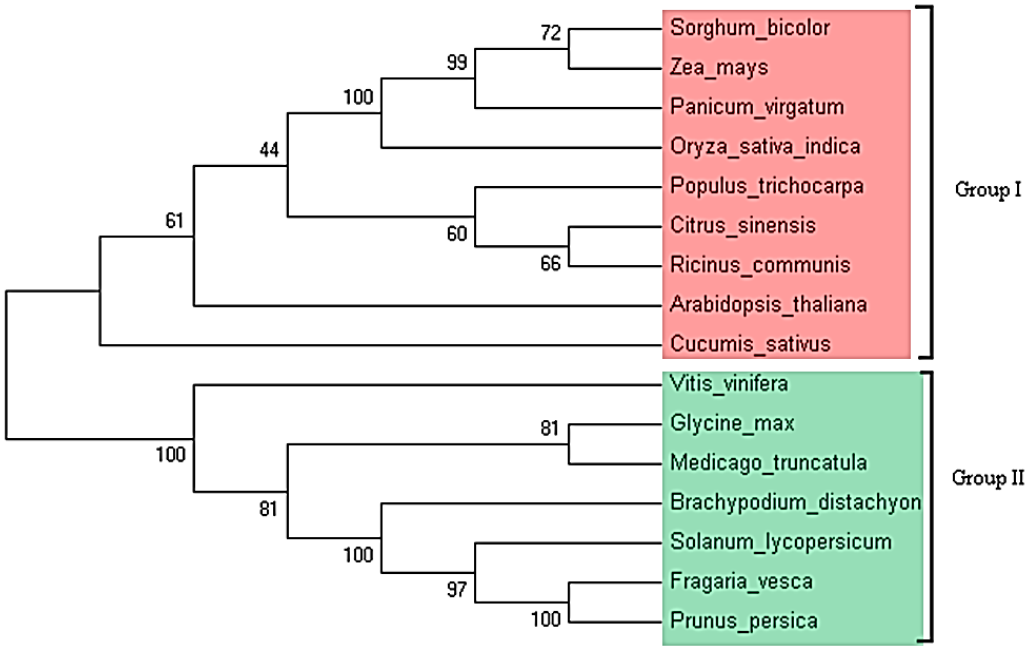


Figure 8. Phylogenetic tree of CAD protein sequences in higher plant species. Sequence alignment was performed using ClustalX and phylogenetic tree was drawn by Neighbour Joining method with MEGA 5.1

References

[1] Lange B.M., Lapierre C., Sandermann, J.H. 1995 Elicitor-Induced Spruce Stress Lignin. Structural Similarity to Early Developmental Lignins, Plant Physiol., 108: 1277-1287.

[2] Tronchet M., Balagué C, Kroj T., Jouanin L., Roby D. 2010. Cinnamyl Alcohol Dehydrogenases-C and D, Key Enzymes In Lignin Biosynthesis, Play An Essential Role In Disease Resistance in *Arabidopsis*, Molecular Plant Pathology, 11: 83-92.

[3] Sibout R., Eudes A., Mouille G., Pollet B., Lapierre C., Jouanin L., Se´guin A. 2005. Cinnamyl Alcohol Dehydrogenase-C and -D are the Primary Genes Involved in Lignin Biosynthesis in the Floral Stem of *Arabidopsis,* The Plant Cell, 17: 2059-2076.

[4] Santos W.D., Ferrarese M.L.L., Ferrarese-Filho O. 2006. High Performance Liquid Chromatography Method for the Determination of Cinnamyl Alcohol Dehydrogenase Activity in Soybean Roots, Plant Physiology and Biochemistry, 44: 511-515.

[5] Ma Q.H. 2010. Functional analysis of a Cinnamyl Alcohol Dehydrogenase Involved In Lignin Biosynthesis in Wheat, Journal of Experimental Botany, 61: 2735-2744.

[6] Li X., Ma D., Chen J., Pu G., Ji Y., Lei C., Du Z., Liu B., Ye H., Wang H. 2012. Biochemical Characterization and Identification of a Cinnamyl Alcohol Dehydrogenase from *Artemisia annua*, Plant Science, 193-194: 85-95.

[7] Barakat A., Bagniewska-Zadworna A., Choi A., Plakkat U., DiLoreto D.S., Yellanki P., Carlson J.E. 2009. The Cinnamyl Alcohol Dehydrogenase Gene Family In *Populus*: Phylogeny, Organization, and Expression, BMC Plant Biology, 9: 26.

[8] Tobias C.M., Chow E.K. 2005. Structure of the cinnamyl-alcohol dehydrogenase gene family in rice and promoter activity of a member associated with lignification. Planta, 220: 678-688.

[9] De Melis L.E., Whiteman P.H., Stevenson T.W. 1999. Isolation and characterisation of a cDNA clone encoding cinnamyl alcohol dehydrogenase in *Eucalyptus globulus* Labill. Plant Science, 143: 173-182.

[10] Kim S.J., Kim M.R., Bedgar D.L., Moinuddin S.G.A., Cardenas C.L., Davin L.B., Kang C., Lewis N.G. 2004. Functional reclassification of the putative cinnamyl alcohol dehydrogenase multigene family in *Arabidopsis*. PNAS, 101: 1455-1460.

[11] Saballos A., Ejeta G., Sanchez E., Kang C., Vermerris W. 2009. A Genome-Wide Analysis of the Cinnamyl Alcohol Dehydrogenase Family in Sorghum [*Sorghum bicolor* (L.) Moench] identifies SbCAD2 as the Brown midrib6 Gene. Genetics, 181:783-795.

[12] Halpin C.K., Holt J., Chojecki D., Oliver B., Chabbert B., Monties B., Edwards K., Barakate A., Foxon G.A. 1998. Brown-midribmaize (bm1): A Mutation Affecting the Cinnamyl Alcohol Dehydrogenase Gene, Plant J., 14: 545-553.

[13] Schubert R., Sperisen C., Müller-Starck G. La Scala S., Ernst D., Sandermann Jr. H., Hager K.P. 1998. The Cinnamyl Alcohol Dehydrogenase Gene Structure in *Picea abies* (L.) Karst.: Genomic Sequences, Southern Hybridization, Genetic Analysis and Phylogenetic Relationships, Trees, 12: 453-463.

[14] Lynch D., Lidgett A., McInnes R., Huxley H., Jones E., Mahoney N., Spangenberg G. 2002. Isolation and Characterisation of Three Cinnamyl Alcohol Dehydrogenase Homologue cDNAs from Perennial Ryegrass (*Lolium perenne* L.). J. Plant Physiol, 159: 653-660.

[15] Kim Y.J., Kim D.G., Lee S.H., Lee I. 2006. Wound-Induced Expression of the *ferulate 5-hydroxylase* gene in *Camptotheca acuminate*. Biochimica et Biophysica Acta, 1760: 182-190.

[16] Sonnhammer E.L., Eddy S.R., Durbin R. 1997. Pfam: A Comprehensive Database of Protein Domain Families Based on Seed Alignments. Proteins, 28: 405-420.

[17] Gasteiger E. 2005. Protein Identification and Analysis Tools on the ExPASy Server. In: John M. Walker ed, The Proteomics Protocols Handbook, Humana Press, 571-607.

[18] Timothy L., Mikael Bodén B.., Buske FA., Frith M., Grant C.E., Clementi L., Ren J., Li W.W., Noble W.S. 2009. MEME SUITE: Tools for Motif Discovery and Searching, Nucleic Acids Research, 37: 202-208.

[19] Emanuelsson O., Nielsen H., Brunak S., Heijne G. 2000. Predicting Subcellular Localization of Proteins Based on Their N-Terminal Amino Acid Sequence. J. Mol. Biol., 300: 1005-1016.

[20] Guo A.Y., Zhu Q.H., Chen X., Luo J.C. 2007. GSDS: A Gene Structure Display Server. Yi Chuan 29 (8):1023–1026.

[21] Schneider T.D., Stephens R.M. 1990. Sequence Logos: A New Way to Display Consensus Sequences. Nucleic Acids Res.,18: 6097-6100.

[22] Crooks G.E., Hon G., Chandonia J.M., Brenner S.E. 2004. WebLogo: A sequence logo Generator. Genome Research, 14: 1188-1190.

[23] Franceschini A., Szklarczyk D., Frankild S., Kuhn M., Simonovic M., Roth A., Lin J., Minguez P., Bork P., von Mering C., Jensen L.J., 2013. STRING V9.1: Protein-Protein Interaction Networks, with Increased Coverage and Integration, Nuc Acid Res 41: 1. doi-10.1093/nar/gks1094.

[24] McGuffin L.J., Bryson K., Jones D.T. 2000. The PSIPRED Protein Structure Prediction Server. Bioinformatics,16 (4): 404-405.

[25] Buchan D.W., Ward S.M., Lobley A.E., Nugent T.C., Bryson K., Jones D.T. 2010. Protein Annotation and Modelling Servers at University College London, Nucl. Acids Res., 38 Suppl, W563-W568.

[26] Guex N., Peitsch M.C., Schwede T. 2009. Automated Comparative Protein Structure Modeling with SWISS-MODEL and Swiss-PdbViewer: A historical perspective, Electrophoresis, 30: S162-S173.

[27] Lovell S.C., Davis I.W., Arendallp W.B. 3rd, de Bakker P.I., Word J.M., Prisant M.G., Richardson J.S., Richardson D.C. 2003. Structure Validation by Cα Geometry: φ, ψand Cβ Deviation, Proteins, 50: 437-450.

[28] Thompson J.D., Higgins D.G., Gibson T.J. 1994. CLUSTAL W: Improving the Sensitivity of Progressive Multiple Sequence Alignment Through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice. Nucleic Acids Res, 22:4673-4680.

[29] Tamura K., Peterson D., Peterson N., Stecher G., Nei M., Kumar S. 2011. MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Molecular Biology and Evolution, 28: 2731-2739.

[30] Deng W.W., Zhang M., Wu J.Q., Jiang Z.Z., Tang L., Li Y.Y., Wei C.L., Jiang C.J., Wan X.C. 2013. Molecular Cloning, Functional Analysis of Three Cinnamyl Alcohol Dehydrogenase (CAD) Genes in The Leaves of Tea Plant, *Camellia sinensis*. Journal of Plant Physiology, 170: 272-282.

[31] Blanco-Portales R., Medina-Escobar N., Lopez-Raez J.A., Gonzalez-Reyes J.A., Villalba J.M., Moyano E, Caballero J.L., Munoz-Blanco J. 2002. Clone, Expression and Immunolocalization

Pattern of a Cinnamyl Dehydrogenase Gene from Strawberry (*Fragaria ananassa* cv. Chandler), J Exp Bot., 375:1723-34.
[32] Schwarz F., Aebi M. 2011. Mechanisms and Principles of N-Linked Protein Glycosylation. Current Opinion in Structural Biology, 21: 576-582.
[33] Xu G., Guo C., Shan H., Kong H. 2011. Divergence of Duplicate Genes in Exon–Intron Structure. PNAS, doi/10.1073/pnas.1109047109.
[34] Boerjan W., Ralph J., Baucher M. 2003. Lignin Biosynthesis, Annu Rev Plant Biol., 54: 519-46.
[35] Herrero J., Esteban-Carrasco A., Zapata J.M. 2013. Looking for *Arabidopsis thaliana* Peroxidases Involved in Lignin Biosynthesis. Plant Physiol. Bioch., 67: 77-86.
[36] Lawton-Rauh A. 2003. Evolutionary Dynamics of Duplicated Genes in Plants. Molecular Phylogenetics and Evolution, 29: 396-409.
[37] Lynch M. 2007. The Origins of Genome Architecture. Sinnauer, Sunderland.
[38] Cheng X., Li M., Li D., Zhang J., Jin Q., Sheng L., Cai Y., Lin Y. 2017. Characterization and Analysis of CCR and CAD Gene Families at the Whole-Genome Level for Lignin Synthesis of Stone Cells in Pear (*Pyrus bretschneideri*) Fruit, Biology open, 6 (11): 1602-1613.
[39] Rong W., Luo M., Shan T., Wei X., Du L., Xu H., Zhang Z. 2016. A Wheat Cinnamyl Alcohol Dehydrogenase TaCAD12 Contributes to Host Resistance to the Sharp Eyespot Disease. Frontiers in Plant Science, 7: 1723.
[40] Bukh C., Nord-Larsen P.H., Rasmussen S.K. 2012. Phylogeny and Structure of the Cinnamyl Alcohol Dehydrogenase Gene family in *Brachypodium distachyon*. J Exp Bot., 63 (17): 6223-36.
[41] Jun S.Y., Walker A.M., Kim H., Ralph J., Vererris W., Sattler S.E., Kang C. 2017. The Enzyme Activity and Substrate Specificity of Two Major Cinnamyl Alcohol Dehydrogenases in Sorghum (*Sorghum bicolor*), SbCAD2 and SbCAD4. Plant Physiol., 174 (4): 2128-2145.