Evaluation of the diversity and phylogenetic implications of NAC transcription factor members of four reference species from the different embryophytic plant groups

Rakhi Chakraborty¹ · Swarnendu Roy²

Received: 6 March 2018 / Revised: 17 June 2018 / Accepted: 17 July 2018 / Published online: 25 July 2018
© Prof. H.S. Srivastava Foundation for Science and Society 2018

Abstract NAC transcription factors (TFs) are one of the largest and important TF family that are involved in the regulation of plant growth and development. They are characterized by a highly conserved N-terminal domain and a variable C-terminal domain. In the present study, the amino acid sequences of NAC TFs from four embryophytic plant species viz. *Arabidopsis thaliana* (Angiosperm), *Picea abies* (Gymnosperm), *Selaginella moellendorffii* (Pteridophyte) and *Physcomitrella patens* (Bryophyte) as reference of the different plant groups were collected from the Plant Transcription Factor Database (PTFD) and the phylogenetic relationships were evaluated. The phylogenetic tree revealed that the majority of the NAC members were interspersed in the major subgroups that indicated the expansion of the NAC members predates the speciation events. Thirty one (31), five (05), one (1) and ten (10) paralog pairs were determined respectively for *Arabidopsis*, *Picea*, *Selaginella* and *Physcomitrella*. The structure–function relationship of paralog pairs were inferred from the phylogenetic tree of combined set of paralogous gene pairs by studying the prevalence of flanking regions and motif analysis of the NAC proteins. The motif analysis revealed the presence of an N-terminal conserved domain, a characteristic of the majority of NAC family proteins. Conserved motifs in the C-terminal region were absent in the majority of the

protein sequences except few members in *Arabidopsis* and *Physcomitrella*. Also the time of gene duplication of the paralog pairs were calculated that revealed the duplication events occurred between 4.48 and 45.94 MYA *Arabidopsis*, 167.57–532.86 MYA in *Picea*, and 29.12–53.53 MYA in *Physcomitrella*.

Keywords NAC transcription factors · Phylogenetic tree · Gene duplication · Motif analysis · Conserved domain

Introduction

NAC transcription factors (TFs) are one of the major plant-specific TFs that are involved in regulation of plant growth and development (Nuruzzaman et al. 2013; Shao et al. 2015). These TFs derive its name from three different identified genes, viz. NAM (no apical meristems), ATAF 1/2 (*Arabidopsis* transcription activation factor) and CUC 2 (cup-shaped cotyledon) (Souer et al. 1996; Aida et al. 1997; Olsen et al. 2005). Members of NAC superfamily share a conserved N-terminal domain and a variable C-terminal domain (Xie et al. 2000). The DNA-binding N-terminal domain consists of approximately 150–160 amino acid residues which has been again divided into 5 sub-domains, but the C-terminal domain is highly variable both in length and amino acid residues (Ooka et al. 2003).

Distribution of NAC TFs in a wide range of plant species has lead to extensive investigation in the identification and characterization of these genes. Complete set of NAC TFs in different species of angiosperms have been reported, viz. 151 in rice, 117 in *Arabidopsis*, 152 in soyabean, 180 in apple, 204 in cabbage and so on (Nuruzzaman et al. 2010, 2013; Le et al. 2011; Su et al. 2013). 37 NAC TFs have been identified from *Pinus pinaster* (Pascual et al.

✉ Swarnendu Roy
swarnendu.kc@gmail.com

¹ Department of Botany, A.P.C. Roy Govt. College, Matigara, Siliguri, WB 734010, India

² Molecular and Analytical Biochemistry Laboratory, Department of Botany, University of Gour Banga, Mokdumpur, Malda, WB 732103, India

2015). 30 NAC TFs have been identified from *Selaginella moellendorffii* (Zhu et al. 2012). Though land plants have evolved from aquatic algae, but no NAC homologs have been identified in algae (Shen et al. 2009). Altogether more than 80 plant species with or without complete genome sequences have been characterised for NAC TFs (Jin et al. 2014). However, no NAC TFs from bacteria, algae and fungi have been reported till date (Kikuchi 2014).

NAC TFs are involved in a large number of functions including root and shoot development, floral morphogenesis, leaf senescence, seed and embryo development and cell cycle regulation in different plant species (Uauy et al. 2006). Apart from these, NAC TFs also plays modulatory roles in plant responses to biotic and abiotic stresses (Nakashima et al. 2012; Puranik et al. 2012). Upregulation of drought tolerance genes at transcriptional level was conferred by 3 NAC genes in *Arabidopsis thaliana*, viz. *ANAC019*, *ANAC055* and *ANAC072* (Tran et al. 2004). According to Jensen et al. 2010, ABA hypersensitivity is conferred through positive regulation of ABA signalling, which may be due to ectopic expression of *ANAC019*. Similarly, mitochondrial retrogressive regulation is mediated by *ANAC013* in response to oxidative stress (De Clercq et al. 2013). *SNAC1* gene expression in rice significantly increases crop production when subjected to severe drought and salinity stress (Hu et al. 2006). Recent studies in cassava isolated 96 NAC genes (*MeNAC*) which confer a wide degree of tolerance to several stress factors like salinity, temperature, ABA, and H_2O_2 (Hu et al. 2016). Also in soyabean, six NAC genes (*GmNAC1*, *GmNAC2*, *GmNAC3*, *GmNAC4*, *GmNAC5* and *GmNAC6*) are actively involved upon exposure to various stress conditions (Pinheiro et al. 2009).

The objective of this study was to evaluate the phylogenetic relationship of the members of NAC protein family across the 4 species of major plant groups, viz. *Arabidopsis thaliana*, *Picea abies*, *Selaginella moellendorffii* and *Physcomitrella patens*. Also, the study of the expansion of the NAC protein family and the possible implications of this expansion was undertaken by evaluating the paralogous gene pairs as inferred from the phylogenetic tree; for which an in silico approach has been undertaken with the aid of several bioinformatics tools.

Materials and methods

Retrieval of NAC TF sequences and construction of phylogenetic tree

Amino acid sequences encoding NAC TFs from 4 plants, viz. *Arabidopsis thaliana*, *Picea abies*, *Selaginella moellendorffii* and *Physcomitrella patens* were retrieved from

the Plant Transcription Factor Database v3.0 (PlantTFDB) (<http://planttfdb.cbi.pku.edu.cn/>) (Jin et al. 2014). The sequences were saved in FASTA format and renumbered for phylogenetic analyses. The combined unrooted phylogenetic tree of all the NAC TFs were constructed with MEGA 7.0 using the Neighbor-Joining (NJ) method with the bootstrap test carried out with 1000 iterations (Kumar et al. 2016). The resulting tree was critically analysed, remodelled and displayed using Fig Tree v1.4.2 (<http://tree.bio.ed.ac.uk/software/figtree/>).

Retrieval of genomic sequences and phylogenetic divergence of corresponding paralogs

The paralogs for NAC TFs in *Arabidopsis*, *Picea*, *Selaginella* and *Physcomitrella* were inferred from the phylogenetic tree. For identification of probable paralogs, a method similar to Yang et al. (2008) was followed with some modification. The transcription factor pairs that appeared at the terminal nodes of the phylogenetic tree were initially selected as paralogs and then the corresponding genomic sequences for the polypeptides were retrieved from different databases: The Arabidopsis Information Resource (TAIR, <https://www.arabidopsis.org/>) for *Arabidopsis*, Congenie (<https://congenie.org/>) for *Picea*, JGI Phytozome 11 (https://phytozome.jgi.doe.gov/pz/#!info?alias=Org_Smoellendorffii) for *Selaginella* and EnsemblPlants (http://plants.ensembl.org/Physcomitrella_patens/Info/Index) for *Physcomitrella*. A phylogenetic tree was again constructed using the nucleotide sequences of the probable paralogs with MEGA 7.0 and the pairs that appeared again at the terminal nodes were confirmed to have arisen due to duplication events. The paralog protein pairs were further reaffirmed by pairwise sequence alignment (https://www.ebi.ac.uk/Tools/psa/emboss_needle/) following the method of Singh and Hannenhalli (2008); protein pairs showing FASTA-alignable region of $\geq 70\%$ and identity (I) $\geq 5\%$ were paralogous. The corresponding intron–exon junctions of the nucleotide sequences were also displayed with the aid of Gene Structure Display Software (GSDS 2.0, <http://gsds.cbi.pku.edu.cn/>) (Hu et al. 2015).

Estimation of synonymous and non-synonymous substitution rates

Multiple sequence alignments of the amino acid sequences of the inferred paralog gene pairs of *Arabidopsis*, *Picea* and *Selaginella* was performed with MUSCLE 3.8 (<http://www.ebi.ac.uk/Tools/msa/muscle/>). The aligned amino acid sequences and their corresponding cDNA sequences were introduced into the server of PAL2NAL (<http://www.bork.embl.de/pal2nal/>) and the estimation of the

synonymous (Ks) and non-synonymous (Ka) substitution rates were performed using the CODEML program in PAML interface (Suyama et al. 2006). Time (million years ago, MYA) of duplication and divergence of each paralog gene pairs were calculated using the formula, $T = Ks/2\lambda$, where Ks refers to the rate of synonymous substitutions and λ refers to the clock like rates of synonymous substitutions that varies from species to species. For *Arabidopsis*, the value of λ was taken to be 1.5×10^{-8} (Koch et al. 2000); for *Picea*, the value of λ was taken to be 0.68×10^{-9} (Buschiazio et al. 2012) and for *Physcomitrella*, the value of λ was taken to be 0.94×10^{-8} (Rensing et al. 2007, 2016).

Prediction of stress responsive *cis*-acting regulatory DNA elements (CREs)

The promoter region of all the individual paralog genes of *Arabidopsis*, *Picea* and *Physcomitrella* were predicted on TSSPlant interface (<http://www.softberry.com/berry.phtml?topic=tssp&group=programs&subgroup=promoter>) (Solovyev et al. 2010). Then approximately one Kb upstream sequence from the transcription start sites were extracted using BioEdit Sequence Alignment Editor (Li et al. 2014). Analysis of the known stress responsive CREs was performed using New PLACE (<https://sogo.dna.affrc.go.jp/cgi-bin/sogo.cgi?lang=en&pj=640&action=page&page=newplace>) (Higo et al. 1999).

Identification of conserved motifs and domains

The program MEME version 4.11.2 was used for the elucidation of motifs in the NAC protein sequences of the paralogous genes (Bailey et al., 2009). MEME was run with the following parameters: number of repetitions—zero or one, maximum number of motifs—20, and the optimum motif widths were constrained to between 6 and 50 residues. The NCBI Conserved Domain Database search (<https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi>) was also performed with the same set of sequences to analyze the presence of conserved domains and the relative position of the conserved domains in the sequence (Marchler-Bauer et al. 2015).

Results

Collection of NAC TF sequences

The protein sequences of NAC TFs of 4 plant genomes, *Arabidopsis*, *Picea*, *Selaginella* and *Physcomitrella* were downloaded from PlantTFDB v3.0 and compiled into FASTA format for phylogenetic analysis. The number of

sequences that were retrieved and used for phylogenetic analysis was 136, 90, 22 and 35 respectively for *Arabidopsis*, *Picea*, *Selaginella* and *Physcomitrella*. NAC TF sequences in *Arabidopsis* ranged from 87 to 806 amino acid residues; in *Picea* ranged from 101 to 940 amino acid residues; in *Selaginella* ranged from 128 to 474 amino acid residues; and in *Physcomitrella* ranged from 243 to 710 amino acid residues.

Phylogenetic relationships of NAC TF sequences in *Arabidopsis*, *Picea*, *Selaginella* and *Physcomitrella*

The phylogenetic relationship among the NAC TF proteins in *Arabidopsis*, *Picea*, *Selaginella* and *Physcomitrella* was done by constructing an unrooted tree from alignments of the full-length NAC protein sequences (Fig. 1). The phylogenetic tree was constructed using MEGA 7.0 by Neighbor-Joining (NJ) method. A close introspection of the phylogenetic tree led to the distribution of the NAC TFs into 12 distinct subgroups that are depicted in different colours. The subgroups were designated as **A1**, **A2**, **B**, **C**, **D**, **E**, **F**, **G**, **H**, **I**, **J** and **K** for simplicity of analysis. Among the subgroups, **B** and **D** were the largest followed by **A1** and **A2**. Among these subgroups, the NAC TFs from all the 4 species were found to be interspersed that indicated that the expansion events of NAC proteins predates the divergence of the studied species. This revelation also holds correct for the subgroups **G** and **J**. In contrast, the subgroup **H** and **I** comprised of 15 and 5 AtNAC TFs respectively only, which suggests expansion of the member of this subgroup occurred most recently after the diversification of the angiosperm stock. Moreover, the subgroup **F** revealed the occurrence of 3 AtNACs (AtNAC13, AtNAC94, AtNAC129) alongside the *Picea* NAC TFs. This indicated the expansion of this subgroup following the divergence of gymnosperm—angiosperm stock from the lower group of plants and the subsequent divergence of the 3 AtNAC TFs from the gymnosperm stock.

Determination of paralogs and evaluation of phylogenetic relationship of the paralog genes

The paralog pairs were inferred from the phylogenetic tree of the full length protein sequences of the NAC TFs (Fig. 1). 31, 5, 1 and 10 paralog pairs were determined respectively for *Arabidopsis*, *Picea*, *Selaginella* and *Physcomitrella*. Both tandem and segmental duplication events contributed to the expansion of NAC TFs in all the plant groups; but in *Arabidopsis* and *Physcomitrella*, the contribution of these duplication events in the expansion of the gene family was more pronounced. A phylogenetic tree was reconstructed taking the genomic sequences of the paralog pairs using NJ method and the subsequent intron–

Fig. 1 Phylogenetic tree of NAC TFs from *Arabidopsis*, *Picea*, *Selaginella* and *Physcomitrella*. The deduced full-length amino acid sequences of 136, 90, 22 and 35 members in *Arabidopsis*, *Picea*, *Selaginella* and *Physcomitrella* NAC proteins respectively were

aligned by MUSCLE 3.8 and the phylogenetic tree was constructed using MEGA 7.0 by the Neighbor-Joining (NJ) method with 1000 bootstrap replicates. Each NAC subfamily has been separated and is depicted using different colours

exon junction was displayed using GSDS 2.0 (Fig. 2). The phylogenetic tree revealed that the paralog genes in *Selaginella* share a common ancestry with the paralog pairs in *Arabidopsis*. Interestingly, the paralog pairs in *Picea* with large introns shared a common ancestry with the paralog genes in *Arabidopsis*. The intron distribution also provided important evidence to support phylogenetic relationships of the paralog pairs of the species studied. In *Arabidopsis*, number of introns ranged from 2 (viz. *AtNAC109*, *AtNAC135*) to 6 (viz. *AtNAC17*, *AtNAC18*). The sizes of the introns were comparatively smaller as compared to that in other species, largest introns up to 1.5 Kb observed in *AtNAC98* and *AtNAC99*. In *Picea*, number of introns

ranged from 2 (viz. *PaNAC37*, *PaNAC48*) to 8 (viz. *PaNAC14*). The largest introns were observed in *Picea* which ranged from 8 to 10 Kb as seen in *PaNAC3*, *PaNAC7* and *PaNAC14*. Only one pair of paralog gene was determined in *Selaginella* that showed 1–2 small introns with a relatively large upstream non-coding region in *SmNAC10*. In contrast, in *Physcomitrella*, number of introns ranged from 1 (viz. *PpNAC14*, *PpNAC15*) to 3 (viz. *PaNAC33*, *PpNAC35*); with relatively large upstream regions (viz. *PpNAC14*, *PpNAC15*) and downstream regions (viz. *PpNAC6*).

Fig. 2 Phylogenetic tree of the combined paralog gene pairs from *Arabidopsis*, *Picea*, *Selaginella* and *Physcomitrella*, constructed using MEGA 7.0 by the Neighbor-Joining (NJ) method with 1000 bootstrap

replicates. The corresponding intron–exon junctions were displayed with GSDS 2.0

Evolutionary pattern of paralog genes

The presence of highly identical and conserved flanking regions for the pairs of paralogous genes, suggest that the expansion of paralogous NAC genes initiated from segmental duplication events, though tandem duplication events were not rare. We used Ks as the proxy for time to estimate the approximate dates of the duplication events. The Ks values and the estimated dates for all duplication events of the paralog pairs of NAC genes of *Arabidopsis*, *Picea* and *Physcomitrella* are listed in Tables 1, 2 and 3 respectively. Ks values $< 1e-05$ for the pair of paralogs were not used for the calculation of the time of duplication events. Also, the time of duplication event for the only paralog pair in *Selaginella* was not calculated, as no reliable records documenting the value of λ (clock like rates of synonymous substitution) for *Selaginella moellendorffii* could be found. In *Arabidopsis*, the values of Ks ranged from 0.13 to 1.37 and the time of duplication event was

calculated to have spanned from 4.48 to 45.94 MYA (Table 1). Similarly in *Picea*, the values of Ks ranged from 0.22 to 0.72 and the time of duplication event was calculated to have spanned from 167.57 to 532.86 MYA (Table 2); and in *Physcomitrella*, the values of Ks ranged from 0.54 to 1.01 and the time of duplication event was calculated to have spanned from 29.12 to 53.53 MYA (Table 3). Ka/Ks is an indicator of positive or negative selection applicable regarding the divergence of paralog pairs. In all the paralogous pairs of the 3 species, Ka/Ks value was < 1 in all instances.

Analysis of stress responsive CREs

The presence of 23 stress responsive CREs (S000007, S000008, S000022, S000024, S000054, S000119, S000133, S000134, S000135, S000140, S000141, S000159, S000174, S000175, S000176, S000177, S000152, S000200, S000232, S000406, S000415,

Table 1 Inference of duplication time of NAC paralogous pairs in *Arabidopsis* (MW and PI of individual paralog pairs are also recorded)

Sl. No.	Paralog pairs	MW	PI	Ks	Ka	Ka/Ks	Duplication time (MYA)
1	<i>AtNAC49-AtNAC104</i>	42.48, 43.51	5.40, 6.19	0.94	0.18	0.19	31.31
2	<i>AtNAC8-AtNAC35</i>	45.84, 45.58	5.83, 6.17	1.09	0.17	0.15	36.51
3	<i>AtNAC56-AtNAC88</i>	40.78, 37.86	6.27, 5.74	0.98	0.25	0.25	32.72
4	<i>AtNAC19-AtNAC93</i>	35.25, 38.95	8.07, 6.76	0.94	0.15	0.16	31.26
5	<i>AtNAC30-AtNAC31</i>	37.85, 42.93	5.46, 5.95	0.13	0.12	0.88	4.48
6	<i>AtNAC85-AtNAC128</i>	20.38, 24.16	5.15, 4.52	0.77	0.27	0.35	25.77
7	<i>AtNAC57-AtNAC116</i>	40.82, 38.73	5.53, 6.33	0.81	0.22	0.28	26.68
8	<i>AtNAC71-AtNAC106</i>	62.35, 63.51	4.63, 4.64	1.03	0.17	0.16	34.24
9	<i>AtNAC83-AtNAC124</i>	52.49, 50.89	5.76, 5.8	1.16	0.26	0.23	38.61
10	<i>AtNAC80-AtNAC122</i>	38.06, 38.04	5.46, 5.4	1.08	0.21	0.19	35.98
11	<i>AtNAC94-AtNAC129</i>	30.12, 33.58	8.51, 5.17	1.38	0.16	0.12	45.94
12	<i>AtNAC58-AtNAC117</i>	55.58, 54.83	5.37, 5.31	1.01	0.17	0.17	33.35
13	<i>AtNAC24-AtNAC75</i>	35.81, 35.41	6.02, 8.67	0.79	0.16	0.19	26.34
14	<i>AtNAC109-AtNAC135</i>	32.22, 35.99	5.55, 8.63	1.09	0.08	0.07	36.51
15	<i>AtNAC79-AtNAC125</i>	35.82, 32.98	8.23, 5.74	0.95	0.14	0.15	31.54
Average		–	–	0.94	0.18	0.24	31.42

Table 2 Inference of duplication time of NAC paralogous pairs in *Picea* (MW and PI of individual paralog pairs are also recorded)

Sl. No.	Paralog pairs	MW	PI	Ks	Ka	Ka/Ks	Duplication time (MYA)
1	<i>PaNAC3-PaNAC42</i>	56.99, 43.15	5.25, 5.45	0.72	0.26	0.19	532.86
2	<i>PaNAC48-PaNAC64</i>	35.62, 42.52	9.04, 5.01	0.23	0.09	0.26	167.57
3	<i>PaNAC7-PaNAC60</i>	35.69, 30.58	9.4, 6.01	0.25	0.16	0.16	186.54
4	<i>PaNAC9-PaNAC14</i>	29.29, 106.8	5.01, 5.53	0.41	0.15	0.49	304.77
Average		–	–	0.41	0.16	0.27	297.93

Table 3 Inference of duplication time of NAC paralogous pairs in *Physcomitrella* (MW and PI of individual paralog pairs are also recorded)

Sl. No.	Paralog pairs	MW	PI	Ks	Ka	Ka/Ks	Duplication time (MYA)
1	<i>PpNAC3-PpNAC20</i>	45.29, 45.84	7.3, 9.01	0.90	0.17	0.19	47.93
2	<i>PpNAC6-PpNAC9</i>	46.59, 48.23	6.68, 7.75	0.75	0.16	0.21	39.68
3	<i>PpNAC4-PpNAC30</i>	47.78, 48.14	9.09, 8.86	0.66	0.07	0.10	35.26
4	<i>PpNAC11-PpNAC24</i>	47.67, 46.99	5.53, 6.36	1.01	0.22	0.22	53.53
5	<i>PpNAC18-PpNAC21</i>	46.84, 46.48	5.93, 6.11	0.92	0.20	0.22	49.25
6	<i>PpNAC16-PpNAC29</i>	47.39, 47.28	5.18, 4.99	0.54	0.08	0.16	29.12
Average		–	–	0.79	0.15	0.19	42.46

S000412) obtained from the New PLACE dataset were analysed for the individual NAC paralog gene promoters. Out of the CREs scanned, only 6 CREs viz. S000024, S000174, S000175, S000176, S000177, S000415 were found to be spanned among the tested upstream promoter regions of *Arabidopsis*, *Picea* and *Physcomitrella* (Table 4).

Analysis of conserved motifs and domains

The protein sequences of all the NAC paralogs that ranged between 158 and 940 amino acid residues were introduced to MEME for motif analysis with a setting of 20 motifs to be discovered. The results of the motif analysis along with the conserved sequence of the motifs are shown in Fig. 3.

Table 4 Stress-responsive CREs determined in the upstream sequences of individual paralog gene promoters

Accession of CRE from New PLACE dataset	CRE ID and sequence recognised	Description	<i>Arabidopsis</i>	<i>Picea</i>	<i>Selaginella</i>	<i>Physcomitrella</i>
S000024	ASF1MOTIFCAMV (TGACG)	Stimulate under biotic and abiotic stress	<i>AtNAC49</i>	<i>PaNAC3</i>	–	<i>PpNAC3</i>
			<i>AtNAC104</i>	<i>PaNAC42</i>		<i>PpNAC20</i>
			<i>AtNAC8</i>	<i>PaNAC64</i>		<i>PpNAC6</i>
			<i>AtNAC56</i>	<i>PaNAC60</i>		<i>PpNAC9</i>
			<i>AtNAC88</i>			<i>PpNAC4</i>
			<i>AtNAC30</i>			<i>PpNAC30</i>
			<i>AtNAC31</i>			<i>PpNAC11</i>
			<i>AtNAC106</i>			<i>PpNAC24</i>
			<i>AtNAC83</i>			<i>PpNAC18</i>
			<i>AtNAC124</i>			<i>PpNAC21</i>
			<i>AtNAC80</i>			<i>PpNAC16</i>
						<i>PpNAC29</i>
S000174	MYCATRD22 (CACATG)	Binding site for MYC in dehydration-responsive gene	<i>AtNAC49</i>	<i>PaNAC48</i>	–	<i>PpNAC3</i>
			<i>AtNAC104</i>			<i>PpNAC6</i>
			<i>AtNAC8</i>			<i>PpNAC9</i>
			<i>AtNAC35</i>			<i>PpNAC24</i>
			<i>AtNAC56</i>			<i>PpNAC18</i>
			<i>AtNAC30</i>			<i>PpNAC29</i>
			<i>AtNAC31</i>			
			<i>AtNAC116</i>			
			<i>AtNAC124</i>			
S000175	MYBATRD22 (CTAACCA)	Binding site for MYB (ATMYB2) in dehydration-responsive gene, rd22	<i>AtNAC49</i>	<i>PaNAC48</i>	–	<i>PpNAC9</i>
			<i>AtNAC104</i>			
			<i>AtNAC8</i>			
			<i>AtNAC106</i>			
S000176	MYBCORE (CNGTTR)	Involved in regulation of genes that are responsive to water stress	<i>AtNAC49</i>	<i>PaNAC3</i>	<i>SmNAC10</i>	<i>PpNAC3</i>
			<i>AtNAC104</i>	<i>PaNAC42</i>		<i>PpNAC20</i>
			<i>AtNAC8</i>	<i>PaNAC48</i>		<i>PpNAC6</i>
			<i>AtNAC56</i>	<i>PaNAC64</i>		<i>PpNAC9</i>
			<i>AtNAC88</i>	<i>PaNAC7</i>		<i>PpNAC4</i>
			<i>AtNAC19</i>	<i>PaNAC60</i>		<i>PpNAC30</i>
			<i>AtNAC93</i>	<i>PaNAC9</i>		<i>PpNAC11</i>
			<i>AtNAC30</i>			<i>PpNAC24</i>
			<i>AtNAC31</i>			<i>PpNAC18</i>
			<i>AtNAC57</i>			<i>PpNAC21</i>
			<i>AtNAC116</i>			<i>PpNAC16</i>
			<i>AtNAC71</i>			<i>PpNAC29</i>
			<i>AtNAC106</i>			
			<i>AtNAC83</i>			
			<i>AtNAC124</i>			
			<i>AtNAC80</i>			
			<i>AtNAC122</i>			
			<i>AtNAC94</i>			

Table 4 continued

Accession of CRE from New PLACE dataset	CRE ID and sequence recognised	Description	<i>Arabidopsis</i>	<i>Picea</i>	<i>Selaginella</i>	<i>Physcomitrella</i>
S000177	MYB2AT (TAACTG)	Involved in regulation of genes that are responsive to water stress	<i>AtNAC49</i>	<i>PaNAC42</i>	–	<i>PpNAC18</i>
			<i>AtNAC104</i>	<i>PaNAC7</i>		
			<i>AtNAC8</i>	<i>PaNAC60</i>		
			<i>AtNAC88</i>	<i>PaNAC9</i>		
			<i>AtNAC116</i>			
S000415	ACGTATERD1 (ACGT)	Required for etiolation-induced expression of <i>erd1</i> under dehydration stress	<i>AtNAC49</i>	<i>PaNAC3</i>	–	<i>PpNAC3</i>
			<i>AtNAC104</i>	<i>PaNAC42</i>		<i>PpNAC20</i>
			<i>AtNAC8</i>	<i>PaNAC64</i>		<i>PpNAC6</i>
			<i>AtNAC56</i>	<i>PaNAC60</i>		<i>PpNAC4</i>
			<i>AtNAC88</i>			<i>PpNAC30</i>
			<i>AtNAC19</i>			<i>PpNAC11</i>
			<i>AtNAC93</i>			<i>PpNAC24</i>
			<i>AtNAC30</i>			<i>PpNAC18</i>
			<i>AtNAC31</i>			<i>PpNAC21</i>
			<i>AtNAC57</i>			<i>PpNAC16</i>
			<i>AtNAC106</i>			<i>PpNAC29</i>
			<i>AtNAC83</i>			
			<i>AtNAC124</i>			
			<i>AtNAC80</i>			

Among the discovered motifs, motifs 14 and 8 were the smallest with a width of 6 and 8 amino acids respectively; whereas the motifs 13, 15 and 19 were the largest with a width of 50 amino acids (Fig. 3). Also, the motifs 1–8 and 14 registered more than 80 hits among the total sequence of 94, which indicated the conserved nature of the sequences across the paralogous sequences and species (Fig. 3). The motif analysis revealed the presence of an N-terminal conserved domain, a characteristic of the majority of NAC family proteins. These findings were also validated by the CDD search which also revealed the presence of an NAM conserved domain in the N-terminal of the majority of proteins (Fig. 4). Interestingly, in few proteins, the NAC conserved domain could not be located in CDD search viz. PpNAC23, PpNAC25, AtNAC26, AtNAC57, PaNAC3, AtNAC30 and AtNAC31. Conserved motifs in the C-terminal region were absent in the majority of the protein sequences owing to their enormous variability (Fig. 4). Though some conserved motifs could be located viz. AtNAC17, AtNAC18, AtNAC102, AtNAC103, AtNAC83, AtNAC124, 20, AtNAC21, AtNAC15, AtNAC16, AtNAC98, AtNAC99 in *Arabidopsis*; PpNAC33, PpNAC35, PpNAC27, PpNAC28, PpNAC4, PpNAC30, PpNAC9, PpNAC3, PpNAC20 in *Physcomitrella* (Fig. 4). No conserved motifs in the C-terminal could be identified in the members of *Selaginella* and *Picea*.

Discussion

The origin and evolution of plants has been regarded as a complex biological phenomenon that leads to variation in the different groups and sub-groups. Fossil evidences from the past years and decades revealed that land habit in plant kingdom arose from the freshwater cyanobacterial members and other multicellular photosynthetic organisms as early as 1 billion years ago, which existed on land in the late Precambrian, around 850 MYA. The origin of land plants leads to the classification into two major categories: bryophytes (with amphibious habit and gametophytic life cycle) and tracheophytes (spore bearing vascular plants) (Kenrick 2000). *Physcomitrella patens* was one of the first land plants, that probably first appeared 500 MYA as described by the fossil records (Schaefer and Zryd 2001). *S. moellendorffii*, one of the first vascular land plants appeared in the fossil record some 400 MYA (Banks 2009). They later formed a dominant part of the world's flora during the Carboniferous period. Though there is no conclusive record of the origin of *Picea abies*, but its evolution is related to other conifers like pine that evolved around 150 MYA (Keeley 2012). The origin of *Arabidopsis thaliana* probably occurred during middle Miocene (approximately 13 MYA) that has been estimated through molecular phylogenetic analysis (Beilstein et al. 2010).

Fig. 3 Sequence logos for the discovered motifs in the protein sequences of the paralogous genes in MEME 4.11.2 along with the width and number of occurrence

The number of NAC protein members in angiosperms and gymnosperms outnumber the members in mosses and ferns (Soltis and Soltis 2013). This is due to the more frequent relative expansion of the gene family in higher plants by the duplication events, either tandem or segmental. The main difference in NAC protein family numbers within the angiosperms and gymnosperms could be due to common selective pressures including environmental stresses, which may have influence the regulation of plant growth and development (Hughes and Friedman 2003). Duplication events facilitate the TFs to accrue different functions from their ancestors and to be naturally selected for their novel functions (Force et al. 1999). Interestingly, the number of paralogous pairs in *Picea* was

comparatively lesser considering the number of NAC members present in this species.

NAC protein family is one of the largest protein families and the members of this family are both structurally and functionally diverse. Therefore, it has been a difficult task to assign or designate structure–function relationship to the individual NAC genes (Puranik et al. 2012). However, several stress related NACs have been reported to play regulatory function in biotic and abiotic stresses (Tran et al. 2010; Jeong et al. 2010; Jensen et al. 2007). 5 conserved subdomains have been identified in the N-terminal region of the NAC TFs (designated as A–E); out of which the subdomains B and E are somewhat divergent that may be related with the diverse function of the NAC members (Ooka et al. 2003). Also, the transcription regulatory region

Fig. 4 Sequence-wise distribution of the individual motifs and the corresponding results of the conserved domain database search for domain analysis

(TRR) lying at the C-terminal region is highly diverse and is associated with the activation or repression of transcription (Puranik et al. 2012). Although diverse, the TRR may or may not possess some specific motifs that are

sometimes conserved across the protein sub-families, if present these motifs impart variable dimensions to the functionality of the individual members, viz. the TRR of rice NAC proteins were found to contain ten C-terminal

motifs (Fang et al. 2008; Shen et al. 2009). In our findings, the presence of variable C-terminal motifs in *Arabidopsis* and *Physcomitrella* in few protein members indicates the possibility of these members to play regulatory function in stressed environments (Fig. 4). This inference is in accordance with the findings of Tran et al. (2004) where the conserved motifs in C-terminal half of related NACs in *Arabidopsis* viz. ANAC019, ANAC055 and ANAC072 were attributed to the regulation of the transcription of other stress related genes. Also, considering the time of divergence of paralogs in *Arabidopsis* and *Physcomitrella*, a clear indication towards the accrual of stress induced regulatory function during the expansion of NAC family can be obtained, as the expansion of paralogs in these two species took place very recently as compared to that in *Picea*.

Gene duplication events are important for the evolution of gene family, because it is associated with the structural divergence of new genes and facilitate the generation of novel functions (Kong et al. 2007). From our results, we could assume that the duplication events were more in *Arabidopsis* followed by *Physcomitrella*. Gene duplication and divergence in function are associated with positive Darwinian selection (Zhang 2003). To explore whether positive selection drove the divergence of the paralog pairs, we estimated the Ka/Ks ratio. The Ka/Ks ratio provides a sensitive measure of selective pressure on the protein and it is accounted as one of the major forces contributing to the variation of structural patterns in a functional protein that ultimately leads to the emergence of new motifs/functions in protein after gene duplication (Yang et al. 2006). Ka/Ks values = 1 indicates neutral evolution or no selection; whereas Ka/Ks values < 1 indicates purifying selection. Rarely, the Ka/Ks values of > 1 are observed, in which case, positive Darwinian selection is involved (Li and Gojbori 1983). The results obtained indicated a purifying selection among all the paralogous gene pairs in all the species studied.

The *cis*-regulatory elements (CREs) play an important role in gene regulation under conditions of abiotic and biotic stresses (Hernandez-Garcia and Finer 2014). Duplication events can either contribute to the evolution of novel functions or enhance the robustness of the existing functions of genes (Arsovski et al. 2015). Therefore, we have explored the presence of the stress-responsive CREs among the paralog genes of all the species under study. Out of the CREs tested, 6 of them have been found to be present in many of the paralog members of the NAC TF promoters of *Arabidopsis*, *Picea* and *Physcomitrella* (Table 4). The existence of CREs in most of the members could be correlated with the positive selection of these members for duplication, which is also supported by the presence of multiple copies of these CREs (Bilas et al. 2016).

In the present study, a comprehensive analysis of NAC proteins in 4 species of the major plant groups in terms of their phylogeny, gene structure, conserved domains and motifs, divergence time of paralogous gene pairs was performed. The phylogenetic tree of all the NAC proteins revealed the presence of 12 distinct subgroups and revealed that the expansion of the majority of NAC TFs in all the species occurred most recently in comparison to the speciation events. However, the functional attribution to the subgroups could not be performed owing to the large number of members in the NAC protein family. The paralogous pairs were inferred from the phylogenetic tree and the divergence time of the duplication events were calculated. The time of duplication event revealed that the expansion of the NAC TFs in *Picea* occurred much prior to that in *Arabidopsis* and *Physcomitrella*. This recent expansion of the NAC members in the 2 species could be related to the accrual of novel functions with the changing environmental conditions on the basis of motif analysis. Also, the expansion of the NAC protein family is driven by purifying selection as evident from the Ka/Ks ratio of the paralogous pairs.

Acknowledgements The corresponding author is grateful to Dr. Vinay Singh, Information Officer, Centre for Bioinformatics, Banaras Hindu University, Varanasi, India for providing valuable inputs for the present study. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

- Aida M, Ishida T, Fukaki H, Fujisawa H, Tasaka M (1997) Genes involved in organ separation in *Arabidopsis*: an analysis of the cup-shaped cotyledon mutant. *Plant Cell* 9(6):841–857
- Arsovski AA, Pradinuk J, Guo XQ, Wang S, Adams KL (2015) Evolution of *cis*-regulatory elements and regulatory networks in duplicated genes of *Arabidopsis*. *Plant Physiol* 169(4):2982–2991
- 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 Res* 37:W202–W208
- Banks JA (2009) Selaginella and 400 million years of separation. *Annu Rev Plant Biol* 60:223–238
- Beilstein MA, Nagalingum NS, Clements MD, Manchester SR, Mathews S (2010) Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana*. *PNAS* 107(43):18724–18728
- Bilas R, Szafran K, Hnatuszko-Konka K, Kononowicz AK (2016) *Cis*-regulatory elements used to control gene expression in plants. *Plant Cell Tiss Org* 127(2):269–287
- Buschiazzo E, Ritland C, Bohlmann J, Ritland K (2012) Slow but not low: genomic comparisons reveal slower evolutionary rate and higher dN/dS in conifers compared to angiosperms. *BMC Evol Biol* 12:8
- De Clercq I, Vermeirssen V, Van Aken O, Vandepoele K, Murcha MW, Law SR, Inze A, Ng S, Ivanova A, Rombaut D, Van de Cotte B, Jaspers P, Van de Peer Y, Kangasjarvi J, Whelan J, Van

- Breusegem F (2013) The membrane bound NAC transcription factor ANAC013 is a regulator of mitochondrial retrograde regulation of the oxidative stress response in *Arabidopsis*. *Plant Cell* 25(9):3472–3490
- Fang Y, You J, Xie K, Xie W, Xiong L (2008) Systematic sequence analysis and identification of tissue-specific or stress-responsive genes of NAC transcription factor family in rice. *Mol Genet Genom* 280:535–546
- Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151(4):1531–1545
- Hernandez-Garcia CM, Finer JJ (2014) Identification and validation of promoters and cis-acting regulatory elements. *Plant Sci* 217–218:109–119
- Higo K, Ugawa Y, Iwamoto M, Korenaga T (1999) Plant *cis*-acting regulatory DNA elements (PLACE) database. *Nucleic Acids Res* 27(1):297–300
- Hu H, Dai M, Yao J, Xiao B, Li X, Zhang Q, Xiong L (2006) Overexpressing a NAM, ATAF, and CUC (NAC) transcription factor enhances drought resistance and salt tolerance in rice. *Proc Natl Acad Sci* 103:12987–12992
- Hu B, Jin J, Guo A-Y, Zhang H, Luo J, Gao G (2015) GSDB 2.0: an upgraded gene feature visualization server. *Bioinformatics* 31(8):1296–1297
- Hu W, Yang H, Yan Y, Wei Y, Tie W, Ding Z, Zuo J, Peng M, Kaimian L (2016) Genome-wide characterization and analysis of bZIP transcription factor gene family related to abiotic stress in cassava. *Sci Rep* 6:22783
- Hughes AL, Friedman R (2003) Parallel evolution by gene duplication in the genomes of two unicellular fungi. *Genome Res* 13(5):794–799
- Jensen MK, Rung JH, Gregersen PL, Gjetting T, Fuglsang AT, Hansen M, Joehnk N, Lyngkjær MF, Collinge DB (2007) The HvNAC6 transcription factor: a positive regulator of penetration resistance in barley and *Arabidopsis*. *Plant Mol Biol* 65(1–2):137–150
- Jensen MK, Kjaersgaard T, Petersen K, Skriver K (2010) NAC genes: time-specific regulators of hormonal signaling in *Arabidopsis*. *Plant Signal Behav* 5(7):907–910
- Jeong JS, Kim YS, Baek KH, Jung H, Ha S-H, Choi YD, Kim M, Reuzeau C, Kim J-K (2010) Root-specific expression of OsNAC10 improves drought tolerance and grain yield in rice under field drought conditions. *Plant Physiol* 153(1):185–197
- Jin JP, Zhang H, Kong L, Gao G, Luo JC (2014) PlantTFDB 3.0: a portal for the functional and evolutionary study of plant transcription factors. *Nucleic Acids Res* 42(D1):182–187
- Keeley JE (2012) Ecology and evolution of pine life histories. *Ann For Sci* 69(4):445–453
- Kenrick P (2000) The relationships of vascular plants. *Phil Trans R Soc Lond B* 355:847–855
- Kikuchi S (2014) Genome-wide view of the expression profiles of NAC-domain genes in response to infection by rice viruses. In: Benkeblia N (ed) *Omics technologies and crop improvement*. CRC Press, Boca Raton, pp 127–152
- Koch MA, Haubold B, Mitchell-Olds T (2000) Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*, *Arabis*, and related genera (Brassicaceae). *Mol Biol Evol* 17:1483–1498
- Kong H, Landherr LL, Frohlich MW, Leebens-Mack J, Ma H, de Pamphilis CW (2007) Patterns of gene duplication in the plant SKP1 gene family in angiosperms: evidence for multiple mechanisms of rapid gene birth. *Plant J* 50:873–885
- Kumar S, Stecher G, Tamura K (2016) MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol Biol Evol* 33:1870–1874
- Le DT, Nishiyama R, Watanabe Y, Mochida K, Yamaguchi-Shinozaki K, Shinozaki K, Tran LS (2011) Genome-wide survey and expression analysis of the plant-specific NAC transcription factor family in soybean during development and dehydration stress. *DNA Res* 18:263–276
- Li WH, Gojobori T (1983) Rapid evolution of goat and sheep globin genes following gene duplication. *Mol Biol Evol* 1(1):94–108
- Li J, Yuan J, Li M (2014) Characterization of putative *cis*-regulatory elements in genes preferentially expressed in *Arabidopsis* male meiocytes. *BioMed Res Int* 2014: Article ID 708364
- Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, Geer RC, He J, Gwadz M, Hurwitz DI, Lanczycki CJ, Lu F, Marchler GH, Song JS, Thanki N, Wang Z, Yamashita RA, Zhang D, Zheng C, Bryant SH (2015) CDD: NCBI's conserved domain database. *Nucleic Acids Res* 43:D222–D226
- Nakashima K, Takasaki H, Mizoi J, Shinozaki K, Yamaguchi-Shinozaki K (2012) NAC transcription factors in plant abiotic stress responses. *Biochim Biophys Acta* 1819:97–103
- Nuruzzaman M, Manimekalai R, Sharoni AM, Satoh K, Kondoh H, Ooka H, Kikuchi S (2010) Genome-wide analysis of NAC transcription factor family in rice. *Gene* 465(1–2):30–44
- Nuruzzaman M, Sharoni AM, Kikuchi S (2013) Roles of NAC transcription factors in the regulation of biotic and abiotic stress responses in plants. *Front Microbiol* 4:248
- Olsen AN, Ernst HA, Leggio LL, Skriver K (2005) NAC transcription factors: structurally distinct, functionally diverse. *Trends Plant Sci* 10(2):79–87
- Ooka H, Satoh K, Doi K, Nagata T, Otomo Y, Murakami K, Matsubara K, Osato N, Kawai J, Carninci P, Hayashizaki Y, Suzuki K, Kojima K, Takahara Y, Yamamoto K, Kikuchi S (2003) Comprehensive analysis of NAC family genes in *Oryza sativa* and *Arabidopsis thaliana*. *DNA Res* 10:239–247
- Pascual MB, Canovas FM, Avila C (2015) The NAC transcription factor family in maritime pine (*Pinus Pinaster*): molecular regulation of two genes involved in stress responses. *BMC Plant Biol* 15:254
- Pinheiro GL, Marques CS, Costa MDBL, Reis PAB, Alves MS, Carvalho CM, Fietto LG, Fontes EPB (2009) Complete inventory of soybean NAC transcription factors: sequence conservation and expression analysis uncover their distinct roles in stress response. *Gene* 444(1–2):10–23
- Puranik S, Sahu PP, Srivastava PS, Prasad M (2012) NAC proteins: regulation and role in stress tolerance. *Trends Plant Sci* 17(6):369–381
- Rensing SA, Ick J, Fawcett JA, Lang D, Zimmer A, Van de Peer Y, Reski R (2007) An ancient genome duplication contributed to the abundance of metabolic genes in the moss *Physcomitrella patens*. *BMC Evol Biol* 7:130
- Rensing SA, Ick J, Fawcett JA, Lang D, Zimmer A, Van de Peer Y, Reski R (2016) Erratum to: an ancient genome duplication contributed to the abundance of metabolic genes in the moss *Physcomitrella patens*. *BMC Evol Biol* 16:184
- Schaefer DG, Zryd J-P (2001) The moss *Physcomitrella patens*, now and then. *Plant Physiol* 127(4):1430–1438
- Shao H, Wang H, Tang X (2015) NAC transcription factors in plant multiple abiotic stress responses: progress and prospects. *Front Plant Sci* 6:902
- Shen H, Yin Y, Chen F, Xu Y, Dixon RA (2009) A bioinformatic analysis of NAC genes for plant cell wall development in relation to lignocellulosic bioenergy production. *Bioenerg Res* 2:217–232
- Singh LN, Hannehalli S (2008) Functional diversification of paralogous transcription factors via divergence in DNA binding site motif and in expression. *PLoS ONE* 3(6):e2345

- Solovyev VV, Shahmuradov IA, Salamov AA (2010) Identification of promoter regions and regulatory sites. *Method Mol Biol* 674:57–83
- Soltis PS, Soltis DE (2013) A conifer genome spruces up plant phylogenomics. *Genome Biol* 14(6):122
- Souer E, Van Houwelingen A, Kloos D, Mol JNM, Koes R (1996) The no apical meristem gene of *Petunia* is required for pattern formation in embryos and flowers and is expressed at meristem and primordia boundaries. *Cell* 85:159–170
- Su H, Zhang S, Yuan X, Chen C, Wang XF, Hao YJ (2013) Genome-wide analysis and identification of stress-responsive genes of the NAM-ATAF1, 2-CUC2 transcription factor family in apple. *Plant Physiol Biochem* 71:11–21
- Suyama M, Torrents D, Bork P (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* 34:W609–W612
- Tran LS, Nakashima K, Sakuma Y, Simpson SD, Fujita Y, Maruyama K, Fujita M, Seki M, Shinozaki K, Yamaguchi-Shinozaki K (2004) Isolation and functional analysis of *Arabidopsis* stress-inducible NAC transcription factors that bind to a drought-responsive cis-element in the early responsive to dehydration stress 1 promoter. *Plant Cell* 16:2481–2498
- Tran LS, Nishiyama R, Yamaguchi-Shinozaki K, Shinozaki K (2010) Potential utilization of NAC transcription factors to enhance abiotic stress tolerance in plants by biotechnological approach. *GM Crops* 1:32–39
- Uauy C, Distelfeld A, Fahima T, Blechl A, Dubcovsky J (2006) A NAC gene regulating senescence improves grain protein, zinc, and iron content in wheat. *Science* 314(5803):1298–1301
- Xie Q, Frugis G, Colgan D, Chua N (2000) *Arabidopsis* NAC1 transduces auxin signal downstream of TIR1 to promote lateral root development. *Genes Dev* 14:3024–3036
- Yang X, Tuskan GA, Cheng MZ (2006) Divergence of the Dof gene families in poplar, *Arabidopsis*, and rice suggests multiple modes of gene evolution after duplication. *Plant Physiol* 142:820–830
- Yang Z, Zhou Y, Wang X, Gu S, Yu J, Liang G, Yan C, Xu C (2008) Genomewide comparative phylogenetic and molecular evolutionary analysis of tubby-like protein family in *Arabidopsis*, rice, and poplar. *Genomics* 92:246–253
- Zhang J (2003) Evolution by gene duplication: an update. *Trends Ecol Evol* 18(6):292–298
- Zhu T, Nevo E, Sun D, Peng J (2012) Phylogenetic analyses unravel the evolutionary history of NAC proteins in plants. *Evolution* 66(6):1833–1848