Indian Journal of Agricultural Sciences **83** (1): 21–5, January 2013/Article

Phylogenetic analysis and secondary structure prediction for drought tolerant Cap binding proteins of plant species

JYOTIKA BHATI[1], PAVAN K CHADUVULA[2], SANJEEV KUMAR[3] and ANIL RAI[4]

Indian Agricultural Statistics Research Institute, New Delhi 110 012

Received: 18 November 2011; Revised accepted: 17 December 2012

ABSTRACT

Cap binding proteins play an important role in improvement of drought tolerant varieties. Fifteen drought tolerant Cap binding proteins from different plant species were phylogenetically analysed and their secondary structure were predicted. On the rectangular cladogram *(Mirabilis jalapa)* was nearest to the origin and is placed separately with *Ricinus communis* forming separate cluster. The Cap binding proteins of *Nicotiana tabacum* were closely related with that of *Solanum tuberosum*, showing their orthologous behaviour. Random coils constituted the maximum coverage along the entire length is a common trend in 14 drought tolerant proteins, irrespective of size and species family.

Key words: Cap binding proteins, Drought tolerance, Phylogenetic analysis, Secondary structure prediction

Agricultural drought is one of the most important abiotic stresses and major constraint on crop growth with marked effect on productivity. Reduction in crop yield due to water stress probably exceeds reduction of yield jointly due to remaining factors (Kramer 1980). Water stress tolerance mechanism in crops is a complex phenomenon, comprising a number of physical and biochemical processes at both cellular and whole organism levels. Drought tolerant crops adopt several strategies to adjust water stress at different stages of growth, such as reduction in water loss by increasing stomatal resistance, increases of water uptake by developing large and deep root systems, and accumulation of osmolytes. The osmolytes accumulation includes amino acids such as proline, glutamate, glycine-betaine and sugars (mannitol, sorbitol and trehalose). These compounds play a key role in preventing membrane disintegration and enzyme inactivation in the low water activity environment (Patrizia *et al.* 2006). The stress signal is first perceived at the membrane level by the receptors and then transduced in the cell to switch on the stress responsive genes for mediating stress tolerance (Mahajan and Tuteja 2005).

There are many proteins which are responsible for drought tolerance in plants, viz. GRAS protein SCL7 (Ma *et al.* 2010), HD-START protein (Yu *et al.* 2008), SNF1-related protein kinase 2 (SRK2C) (Umezawa 2004), RAB18 (Mantyla 1995) and Cap binding protein (CBP) (Papp 2004). GRAS protein SCL7 is up-regulated under stress conditions and plays diverse roles in plant development and stress responses. Plant overexpressing SCL7 showed increased tolerance to environmental stress including high salinity, osmotic and drought stresses at seedling stage, without causing growth retardation (Ma *et al.* 2010). HD START proteins regulates a complex network of genes that increases the accumulation of abscisic acid (ABA), enhances ABA and calcium signalling, enhances oxidative stress tolerance, improves root architecture and reduces stomatal density. It can prove to be a key regulator to improve drought tolerance in plants as it has a conserved homeobox domain which is reported to play important role during abiotic stress tolerance (Yu *et al.* 2008). SRK2C is an osmotic-stress-activated protein kinase that impacts drought tolerance of plants. These proteins mediate stress signalling in plants by acting as a sensor of water and nutrient status in soil. It is a positive regulator of drought tolerance in *Arabidopsis* roots (Umezawa 2004). RAB18 polypeptides play a significant role in tolerance to water stress, play a protective function on thylakoid membranes and enzyme activities during freeze induced cellular dehydration (Mantyla 1995). The Cap Binding Complex (CBC) consists of a handful of proteins, including the cap binding proteins CBP80 and CBP20. Both proteins are essential for nCBC function. CBC plays role in splicing and recruits mRNA to ribosome. It plays role in the stability of polyA site cleavage complexes formed in the nuclear extract and helps in the protection of mRNA from RNAases. Mutations in either of these proteins

[1]Research Associate (e mail: jyotikabhati@iasri.res.in), [2]Research Associate (e mail: pavanck@iasri.res.in), [3]Scientist (Senior Scale) (e mail: sanjeevk@iasri.res.in), [4]Head & Principal Scientist (e mail: anilrai@iasri.res.in), Centre for Agricultural Bioinformatics

are ABA hypersensitive, which affect vegetative growth rate and tolerate water deficiency much better. Reduction of the cbp20 function also confers hypersensitivity to abscisic acid during germination, significant reduction of stomatal conductance, serrated leaf phenotype and greatly enhanced tolerance to drought. CBP20 provides a new target for breeding efforts that aim at the improvement of drought tolerance in plants (Papp 2004).

Proteins are classified according to their secondary structure content, considering α-helices, β-strands and other structural conformations such as loops, turns and coils. Computational approaches to predict protein structures prove to be very useful in creating insights into structural protein folding. Aligning the sequences proves to be useful for studying molecular evolution and analysing sequence-structure relationships. Phylogenetic study helps in elucidating the evolutionary relationships among various organisms. Sequences collected from different species for same protein family shows the path of their evolution with greater accuracy, as single amino acid change, leads to conformational changes in the proteins. The present study aims to predict secondary structures for 15 CBP, because of its significant role in drought tolerance, across different plant species and studying their evolutionary relationships through phylogenetic analysis.

MATERIALS AND METHODS

Collection of sequence data

Amino acid sequences of CBP were downloaded from NCBI GenBank across agriculturally important plant species, viz. *Arabidopsis thaliana*, *Oryza sativa*, *Solanum tuberosum*, *Zea mays*, *Sorghum bicolor*, *Ricinus communis*, *Triticum aestivum*, *Nicotiana tabacum*, *Arabidopsis lyrata*, *Mirabilis jalapa*, *Hordeum vulgare*, *Medicago truncatula*, *Vitis vinifera*, *Populus trichocarpa* and *Helianthus annuus*. Protein sequences were saved in FASTA format (Table 1).

Table 1 Sequences of cap-binding proteins retrieved from NCBI

Species	Common name	Gi Number
Arabidopsis thaliana	Thale cress	gi|20260618|gb|AAM13207.1
Oryza sativa	Rice	gi|75327885|sp|Q84L14.1
Solanum tuberosum	Potato	gi|224460063|gb|ACN43582.1
Zea mays	Maize	gi|6016335|sp|O81482.1
Sorghum bicolor	Jowar	gi|242065748|ref|XP_002454163.1
Ricinus communis	Castor	gi|255553647|ref|XP_002517864.1
Triticum aestivum	Wheat	gi|30923212|sp|P29557.3
Nicotiana tabacum	Tobacco	gi|261866539|gb|ACY02034.1
Arabidopsis lyrata	Thale cress	gi|297845870|ref|XP_002890816.1
Mirabilis jalapa	4 o' clock flower	gi|46949206|gb|AAT07459.1
Hordeum vulgare	Barley	gi|220936494|gb|ACL83596.1
Medicago truncatula	Barrel	gi|217073778|gb|ACJ85249.1
Vitis vinifera	Grape vine	gi|147770671|emb|CAN62481.1
Populus trichocarpa	Black cottonwood	gi|222869561|gb|EEF06692.1
Helianthus annuus	Sunflower	gi|309256635|gb|ADO62445.1

Multiple sequence alignment

Multiple sequence alignment is of fundamental importance in all aspects of DNA and protein sequence analysis. Multiple alignments constitute an extremely powerful means of revealing the constraints imposed by structure and function on the evolution of a protein family. CLUSTAL provides single environment, where user can perform multiple alignments, view results and, if needed, refine and improve the alignment, allowing highlighting low-scoring regions in the alignment. Low-scoring regions could be corrected automatically by realigning a misaligned sequence or a selected region of an alignment. CLUSTAL aligns all pairs of sequences separately in order to calculate a distance matrix giving the divergence of each pair of sequences. A quality score is calculated for each column in the alignment, which depends on the amino acid variability in the column. A high score indicates a highly conserved column; a low score indicates a less well-conserved position (Thomson *et al.* 1994). CLUSTAL-X was used for sequence alignment.

Phylogenetic analysis

The evolutionary connections between organisms are represented graphically through phylogenetic trees. Phylogenetics is the study of evolutionary relatedness among groups of organisms (eg species, populations), which is revealed through molecular sequencing data and morphological data matrices. The distance of a path in a phylogenetic tree must be as close as the evolutionary distance between two species. A phylogenetic tree must not have edge crossings, as such crossings prevent from recognising the phylogenetic information. PhyloDraw provides various editing functions to provide a user-friendly interface by selecting different tree type (rectangular cladogram, slanted cladogram, phylogram, unrooted tree, or radial tree), resize the tree, and change the branching patterns of the phylogenetic tree. PhyloDraw constructs the tree from the pairwise distance matrix, using either of the two clustering methods: Neighbor

Joining and Fitch-Margoliash. Further, PhyloDraw is a unified viewing tool that supports various multi-alignment formats (Dialign2, ClustalX, Phylip, NEXUS, MEGA, etc.) and visualizes various kinds of tree diagrams (Choi *et al.* 2000). The file saved in CLUSTAL-X with '.ph' extension was imported in PhyloDraw using rectangular cladogram. The phylogenetic clusters can be visualized along with root distance and pair distance. The final tree layout was exported as BMP (bitmap image format).

Structural prediction of proteins

Secondary structure of protein is defined by the patterns of hydrogen bonds between backbone amide and carboxyl groups. Amino acids vary in their ability to form the various secondary structure elements like alpha-helices, beta-strands, loops, and random coils. Secondary structure prediction is a set of techniques that aim to predict the local secondary structures of proteins sequences based only on knowledge of their primary structure. Secondary structure is the general three-dimensional form of local segments of proteins. In proteins, SOPMA (Self-optimized prediction method) is secondary structure prediction software which works on neural network based methods. The tools first searches the SWISSPORT database and extracts the most homologous sequences using FASTA program; secondly, aligning the sequences with the set of homologous proteins using CLUSTAL program and finally applying SOPM method to each aligned sequences. For each amino acid position in the alignment, the conformational score of each state is averaged over all the sequences at the given position (Georjon and Deleage 1995). The amino acid sequence in FASTA format was imported in SOPMA window and submitted to SOPMA server. The results appeared with secondary structure and with percentage of each secondary structure of the proteins. Disulfide bonds are the part of secondary structure and were predicted using DiANNA 1.1 web server (Ferre and Clote 2006).

RESULTS AND DISCUSSION

The present study focuses on the phylogenetic analysis among the cap binding proteins from 15 species of plant kingdom. The amino acid sequences of drought tolerant protein (Table 1) were retrieved in the FASTA format. Among these proteins, largest protein sequence was *Vitis vinifera* with 266 amino acids, while the smallest sequence was *Populus trichocarpa* with 140 amino acids.

The evolutionary relationships between the plants were evaluated by phylogenetic analysis of the aligned amino acids sequences of their cap binding proteins. Phylogenetic analysis of retrieved drought tolerant proteins was performed by PhyloDraw. The resultant cladogram was divided into two distinct clusters (Fig 1). *Mirabilis jalapa* was nearest to the origin and is placed separately with *Ricinus communis* forming separate cluster with root distance 0.019 and pair distance 0.571 with *Ricinus communis*. There are few ribosome-inactivating proteins (RIP) which showed similar action among species like *Mirabilis jalapa*, *Ricinus communis*, and *Phytolacca americana* (Kataoka *et al.* 1991). The present analysis supports the above fact, *Ricinus communis* and *Mirabilis jalapa* formed a cluster which is completely out grouped from the rest 13 species.

The second cluster comprised of remaining 13 species, and was divided into two sub-clusters formed. Sub-cluster I consisted of one sub-sub cluster of *Solanum tuberosum* and *Nicotiana tabacum*. It was found that the Cap binding proteins of *Nicotiana tabacum* was closely related with that of *Solanum tuberosum*, showing their orthologous behaviour as they belong to the same family Solanaceae. *Triticum aestivum* and *Arabidopsis lyrata* formed a cluster, having a largest root distance 0.882 and *Solanum tuberosum* and *Nicotiana tabacum* forming a cluster with lowest root distance and pair distance of 0.090 and 0.051 respectively. *Zea mays* differed and have a pair distance of 0.886. The sub-cluster 2 consisted of two sub-sub cluster with first cluster of *Medicago truncatula* and *Helianthus annuus* having largest pair distance of 0.934 (Table 2). It has been predicted that SERK gene in both *Medicago truncatula* and *Helianthus annuus* showed similar function towards environmental stresses (Nolan *et al.* 2003, Thomas *et al.* 2004). This phylogenetic study also showed maximum nearness between *Medicago truncatula* and *Helianthus annuus* and hence they belong to same cluster. The analysis of length, composition and divergence time of intraspecific comparisons of rice genome, indicated common and lineage-specific patterns of conservation between different grasses' genomes such as barley (*Hordeum vulgare*) and sorghum (*Sorghum bicolor*) (Salse *et al.* 2008). This provides support to the orthology between *Oryza sativa*, *Sorghum bicolor* and *Hordeun vulgare*. Somatic embryogenesis receptor kinase (SERK) genes play an important role in plant response to biotic and abiotic stimuli (Santos and Aragao 2009).

The secondary structure prediction of drought tolerant proteins among 15 species was obtained by using SOPMA online secondary structure prediction server. The content of α-helix in *Helianthus annus* was highest (60.33%) and was lowest (24.28%) in *Arabidopsis lyrata* (Table 3). The percentage of β-turn was the highest in *Mirabilis jalapa* (11.56%) and lowest in *Arabidopsis lyrata* (4.12%).

Table 2 Cluster form on cladogram with their root and pair distance

Cluster among species	Root distance	Pair distance
Mirabilis jalapa (nearest to origin)	0.019	0.571
Triticum aestivum/Arabidopsis lyrata	0.882	0.427
Zea mays (differed)	0.890	0.886
Oryza sativa/Sorghum bicolor	0.130	0.111
Solanum tuberosum/Nicotiana tabacum	0.090	0.051
Medicago truncatula/Helianthus annus	0.879	0.934

Table 3 Percentage of amino acids sequence forming secondary structure in SOPMA prediction and number of disulfide bonds in proteins

Species	No. of amino acids	No. of disulfide bonds	α-helix (Hh)(%)	β-turns (Tt)(%)	Extended strands (Ee)(%)	Random coils (%)
Arabidopsis thaliana	221	6	27.51	05.35	15.60	51.54
Oryza sativa	243	1	25.51	07.41	09.05	58.02
Solanum tuberosum	255	1	32.16	08.63	08.24	50.98
Zea mays	216	3	36.11	05.09	13.43	45.37
Sorghum bicolor	242	1	33.06	06.61	08.26	52.07
Ricinus communis	255	1	30.20	05.88	12.16	51.76
Triticum aestivum	215	6	38.60	05.88	14.88	40.93
Nicotiana tabacum	254	3	29.13	09.45	07.87	53.54
Arabidopsis lyrata	243	3	24.28	04.12	16.46	55.14
Mirabilis jalapa	147	1	32.65	11.56	16.33	39.46
Hordeum vulgare	241	1	29.05	07.05	10.37	53.53
Medicago truncatula	238	10	35.71	06.72	24.79	32.77
Vitis vinifera	266	28	34.21	05.26	15.79	44.74
Populus trichocarpa	140	28	27.86	08.57	24.29	39.29
Helianthus annuus	184	21	60.33	05.98	13.59	20.11

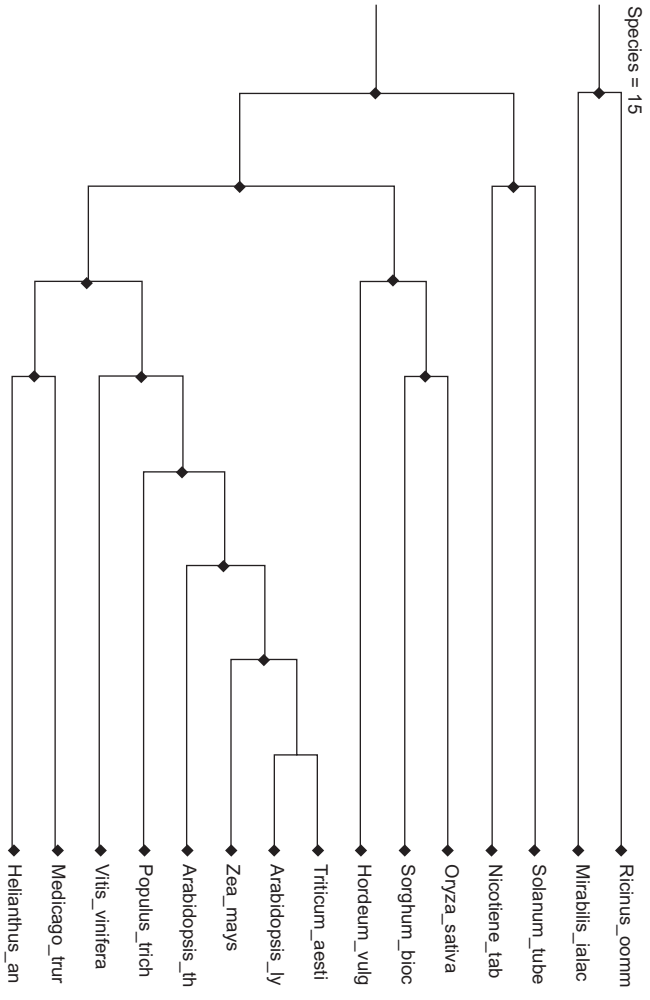

Fig 1 Cladogram of 15 drought tolerant cap binding proteins

Percentage of extended strands was lowest in *Nicotiana tabacum* (7.87%) and highest percentage of extended strands was 24.79% found in *Medicago truncatula*. The percentage of random coil was highest in *Oryza sativa* (58.02%) and lowest percentage of random coils was 20.11%, which was found in *Helianthus annuus* (Table 3).

By observing the colours, structural differences could be identified with respective amino acids changes in an individual species. The β-turns of *Ricinus communis* were homologous to *Triticum aestivum*. There were no β-bridges (Bb) found in all the 15 species of drought tolerant cap binding proteins. There is a common trend among all the 14 drought tolerant proteins, i e random coils constituted the maximum coverage along the entire length of the proteins, irrespective of the size of the protein or the family to which the species belong. This ranges between 40% to 58%. This was followed by α-helix, except in *Helianthus annuus*. The frequency of occurrence of extranded strands in the protein sequence was at the third position with least β-turns. The β-turns of *Ricinus communis* were homologous to *Triticum aestivum*. It has been observed that, occurrence of α-helix is more than that of β-extranded strands in a protein sequence irrespective of the species (Birve *et al.* 1996).

The presence of disulphide bonds are the characteristics of stress related proteins. The disulphide bonds were found in all 15 cap-binding proteins, which confer its stress tolerant characteristics (Table 3). Disulfide bond plays an important role in the folding and stability of the proteins. It is reported in the literature that during stress conditions, the proteins misfold or unfold to unregulate the gene expression which can trigger the endoplasmic reticulum stress responses which lead to apoptosis or programmed cell death (Liu and Howell 2010, Malhotra and Kaufman 2007).

Phylogenetic analysis reveals the presence of various orthologous plant species in the evolution of drought tolerant cap binding proteins. In this analysis it was proved that stress tolerant proteins have the disulphide bonds which are responsible for the protein folding to regulate the gene expression during stress responses. The result reveals that the secondary structure of proteins have the highest occurrence of α-helix as compared to β-extranded strands. Further studies can be done to elucidate the tertiary structure of the cap binding proteins.

ACKNOWLEDGEMENTS

We acknowledge support from National Agricultural Bioinformatics Grid (NABG), National Agricultural Innovation Project (NAIP), Indian Council for Agricultural Research (ICAR).

REFERENCES

Birve S J, Selstam E and Johansson L B A. 1996. Secondary structure of NADPH: protochlorophyllide oxidoreductase examined by circular dichroism and prediction methods. *Biochemistry Journal* **317**: 549–55.

Choi J, Yung H, Kim H and Cho H. 2000. PhyloDraw: a phylogenetic tree drawing system. *Bioinformatics* **16**(11): 1 056–8.

Ferre F and Clote P. 2006. DiANNA 1.1: an extension of the DiANNA web server for ternary cysteine classification. *Nucleic Acid Research* **34**(2): 182–5.

Georjon C and Deleage G. 1995. SOPMA: significant improvement in protein secondary structure prediction by consensus prediction from multiple alignments. *CABIOS* **11**(6): 681–4.

Kataoka J, Habuka N, Furuno M, Miyano M, Takanami Y and Koiwai A. 1991. DNA sequence of *Mirabilis* antiviral protein (MAP), a ribosome-inactivating protein with an antiviral property, from *Mirabilis jalapa* L. and its expression in *Escherichia coli*. *Journal of Biological Chemistry* **266**(13): 8 426–30.

Kataoka N, Ohno M, Kangawa K, Tokoro Y and Shimura Y. 1994. Cloning of complementary DNA encoding an 80 kilodalton nuclear cap binding protein. *Nucleic Acids Research* **22**: 3 861–5.

Kramer P J. 1980. Drought, stress, and the origin of adaptation. (*In*) *Adaptation of Plants to Water and High Temperature Stress*, pp 7–20. Turner N C and Kramer P J (Eds). John Wiley and Sons, New York, USA.

Liu J X and Howell S H. 2010. Endoplasmic reticulum protein quality control and its relationship to environmental stress responses in plants. *The Plant Cell* **22**: 2 930–42.

Ma H, Liang D, Shuai P, Xia X and Yin W. 2010. The salt- and drought-inducible poplar GRAS protein SCL7 confers salt and drought tolerance in *Arabidopsis thaliana*. *Journal of Experimental Botany* **61**(14): 4 011-9.

Mahajan S and Tuteja N. 2005. Cold, salinity and drought stresses: An overview. *Archives of Biochemistry and Biophysics* **444**: 139–58.

Malhotra J D and Kaufman R J. 2007. The endoplasmic reticulum and the unfolded protein response. *Semin Cell Develomental Biology* **18**: 716–31.

Mantyla E, Lang V and Palva E T. 1995. Role of abscisic acid in drought-induced freezing tolerance, cold acclimation, and accumulation of LT178 and RAB18 proteins in *Arabidopsis thaliana*. *Plant Physiology* **107**: 141–8.

Nolan K E, Irwanto R R and Rose R J. 2003. Auxin upregulates *MtSERK1* expression in both *Medicago truncatula* root-forming and embryogenic cultures. *Plant Physiology* **133**: 218–30.

Papp I, Mur L A, Dalmadi A, Dulai S, and Koncz C. 2004. A mutation in the Cap Binding Protein 20 gene confers drought tolerance to *Arabidopsis*. *Plant Molecular Biology* **55**: 679–86.

Patrizia Rampino et al.2006. Drought stress response in wheat: physiological and molecular analysis of resistant and sensitive genotypes. *Plant, Cell and Environment* **29**: 2 143–52.

Salse J, Bolot S, Throude M, Jouffe V, Piegu B, Quraishi U M, Calcagno T, Cooke R, Delseny M and Feuillet C. 2008. Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. *The Plant Cell* **20**: 11–24.

Santos M O and Aragao F J L. 2009. Role of SERK genes in plant environmental response. *Plant Signaling and Behavior* **4**(12): 1 111–3.

Thomas C, Meyer D, Himber C and Steinmetz A. 2004. Spatial expression of a sunflower *SERK* gene during induction of somatic embryogenesis and shoot organogenesis. *Plant Physiology Biochemistry* **42**: 35–42.

Thompson J D, Higgins D G and 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 Acid Research* **22**(22): 4 673–80.

Umezawa T, Yoshia R, Maruyama K, Yamaguchi-Shinozaki K and Shinozaki K. 2004. SRK2C, a SNK1 related protein kinase 2, improves drought tolerance by controlling stress-responsive gene expression in *Arabisopsis thaliana*. *Proceedings of National Academy of Science* **101**: 17 306–11.

Yu H, Chen X, Hong Y, Wang Y, Xu P, Ke S, Liu H, Zhu J, Oliver D J and Xiang C. 2008. Activated expression of an *arabidopsis* HD-START protein confers drought tolerance with improved root System and reduced stomatal density. *The Plant Cell* **20**: 1 134–51.