Chapter V

DISCUSSION
Chapter 5

Discussion

*C. borivilianum* is a wonderful herb, cultivated on a large scale and has vast demand all over the world. It is among the twenty odd species of natural medicinal plants that enjoy intensive therapeutic application & expansive global market. Besides being an aphrodisiac agent, it is helpful in curing impotency and is thought of as an effective substitute for Viagra. The amalgamation of *C. borivilianum* leaves with other herbs such as *Withania sominifera*, *Embilca officinalis* etc. makes body resistant against the attack of sex related diseases and it also delays the menopause (Kaushik, 2005). It is used as a curative for physical weakness, illnesses, natal and post natal problems. It is viewed as an immunity improving drug, a remedy for diabetics and arthritis. As a medicinal plant it is poised to open up new exciting avenues in healthcare (Kritikar and Basu 1975). As a commercial crop, it is positioned to provide unbeatable returns. Most biological effects of *C. borivilianum* are attributed to saponins.

Saponin biosynthesis involves several enzymatic reactions. A comprehensive understanding of the biochemical pathway and enzymes involved in saponin biosynthesis will facilitate the development of plants with altered saponin content or its production *in vitro*. In some cases, enhanced levels of saponins or synthesis of novel saponins may be desirable (for example: for drug production or improved disease resistance).

This study therefore, aims to understand saponins and their biosynthesis in *C. borivilianum*. A few expressed sequence tags (ESTs) have been generated by our lab (Kumar et al., 2012), with the aim to identify the differentially expressed genes in the leaf and root tissues of *C. borivilianum*, thus, the functional genomics studies in *C. borivilianum* is still in its juvenility. Also, efforts are being made to clone the genes of interest via the candidate gene approach for this plant.

The limited transcriptomic data hampers the study of saponin biosynthetic mechanisms in several plants including, *C. borivilianum*. Next generation sequencing of transcriptome has evolved to be a very useful technique for providing large expression data in much shorter time period (as compared to the conventional EST-based approach), depth and coverage to expedite understanding of metabolic pathway as well as contribute to comparative transcriptomics, evolutionary genomics and gene discovery (Shendure and Ji, 2008; Morozova et al., 2009). With this background, the work was initiated with an aim of identifying and characterizing early genes involved in saponin biosynthesis in *C. borivilianum*. 

- 112 -
5.1 Cloning of Early Genes Associated with Saponin Biosynthesis

To understand molecular basis of saponin biosynthesis, cloning of genes of the pathway occupies central role. None of the genes associated with steroid or triterpenoid saponin biosynthesis have been cloned in *C. borivilianum* so far. In order to have a comprehensive picture of saponin biosynthetic pathway, it was imperative to clone the genes involved and the present thesis reported cloning of three full-length cDNAs, namely *CbSqS*, *CbSE* and *CbCAS*, their analysis and characterization along with partial cloning of *CbBAS*.

5.1.1 Degenerate Oligonucleotide Primers Successfully Amplified Partial *CbSqS* and *CbSE*

Procedures for PCR-based cloning need standardization of various parameters like annealing temperature, MgCl₂ concentration and number of PCR cycles. Use of degenerate primers can result in non-specific amplification apart from the desired band and thus require cloning and screening of all those bands individually. The chances of getting the desired gene/gene fragment are probability based. A single primer pair may or may not work in all the cases. Therefore, use of more than one primer pair may be required to amplify the target gene under different amplification conditions. Thus the procedure costs on time, labor and money.

Degenerate oligonucleotide primer pair was adopted from previously available sources (Kribii et al., 1997; Uchida et al., 2007) which successfully amplified partial gene fragments of *CbSqS* (380 bp) and *CbSE* (529 bp). The fragments thus obtained showed significant similarity to *SqSs* and *SEs* from other plant species. A 380 bp partial fragment of *CbSqS* shared 77% and 78% similarity at nucleotide level and 83% and 84 % similarity at amino acid level with the *SqS* of *Z. mays* and *O. sativa* respectively. On the other hand, a 529 bp partial fragment of *CbSE* shared 77% and 79% similarity at nucleotide level and 85% and 86% similarity at amino acid level with *SE* of *Z. mays* and *O. sativa* respectively. This confirmed the cloning of desired genes. Thus full length genes of *CbSqS* and *CbSE* were then cloned through RACE PCR.

5.1.2 Full length cloning of *CbSqS* and *CbSE* through RACE

*CbSqS*

The sequence information from partial *CbSqS* gene was used to design primers for RACE PCR to amplify full-length cDNA sequence. Amplicons were generated, and sequence analysis of ten different clones obtained from both the 5'-
and 3′-RACE PCR revealed the full-length cDNA clone for CbSqS gene (1760 bp) (Kalra et al., 2013). Deduced amino acid sequence of CbSqS shared 81%, 77%, 81%, 78% and 73% similarity with Panax notoginseng (ABA29019.1), Centella asiatica (AAV58897.1), Euphorbia tirucalli (BAH23428.1), Glycyrrhiza uralensis (ADG36717.1) and Arabidopsis thaliana (NP195190.1) respectively. SqS is generally accepted to be a monomeric enzyme with molecular weight in a range from 40-50 kDa. However, some exceptions are indeed found only in prokaryotes. The molecular weight of Staphylococcus aureus SqS is 56 kDa (Holden et al., 2004) while SqS from Botryococcus braunii is 7.3 kDa (Kaneda et al., 2001).

Multiple sequence alignment of CbSqS with SqS gene from other plant, yeast, humans and prokaryotic species revealed six highly conserved signature domains (I-VI) of CbSqS of 14-23 amino acids. These domains are helpful in the catalytic activity of the enzyme Out of all these domains, domains III, IV and V showed a highly conserved consensus sequence (DDYDEYCHGYVAGLGLGS, GLFLKQTNIRDYLED and FCAIPQXMAIGTXLAXCYN respectively) with other SqS enzyme sequences while domains I and II were less conserved. Domain II was highly conserved with the aspartate-rich-motif with other SqS amino acid sequences (Devarenne et al., 2002). These aspartate-rich-motifs are proposed binding site for the phosphate groups of the prenyl acceptor (allylic substrate) via a magnesium salt bridge. To further corroborate the importance of aspartate rich motif, a study was performed by Gu et al.(1998) where site-directed mutagenesis of rat hepatic squalene synthase (RSS, EC 2.5.1.21) showed that the first Tyr residue (Tyr171, corresponding to Tyr173 of CbSqS) in the region III is essential for the activity of RSS and is likely involved in the first reaction, and mutation of Tyr171 to the mutants Y171F, Y171S, and Y171 completely abolished formation of PSPP or squalene from FPP (Pandit et al., 2000). This result also indicated that the phenyl ring is specifically required for activity and cannot be substituted with just either an aromatic or hydroxyl group (Pandit et al., 2000). The region IV also contains an aspartate-rich motif, 228-DYLED-232 in CbSqS, which overlaps with the 219-DYLED-223 sequence in human SQS. In the RSS, the two Asp residues Asp219 and Asp223 (equivalent to Asp224 and Asp234 of CbSqS, respectively) are essential; neutralization or reversal of charge (D219 N, D223 N, and D223 K) cause inactivation. The results were consistent with the participation of Asp219 and Asp224 in the binding of the diphosphates of two substrate FPP molecules via bridging Mg$^{2+}$ ions (Lee and Poulter, 2008). Region V is believed to be responsible for the catalysis of the second reaction F/VAIPQXMAIA/GTL (X=V/G/A) sequence is likely the NADPH binding motif
Site-directed mutagenesis showed that the RSS mutant F288L (corresponding to Phe298 in the motif FVAIP of CbSqS) caused almost a complete loss of the second activity, but led to accumulation of significant amounts of PSPP even in the presence of NADPH (Pandit et al., 2000). The aspartate-rich motif (81DTVED85 in CbSqS, relevant to 80DTLED84 in human SQS, data not shown) in region II was also predicted to participate in the binding of the substrates (Pandit et al., 2000). Most membrane proteins have a hydrophobic region that span the hydrophobic core of the membrane bilayer. Amino acids within the C-terminal region formed domain VI. These were the most variable and hydrophobic compared with the other regions of the sequence. These sequences have a role in anchoring the enzyme to endoplasmic reticulum membrane (Ashby and Edwards, 1990).

The predicted secondary structure by SOPMA program indicated the similarity of CbSqS to other plant SqS secondary structure. These percentages are again in range with SqS from other plant species encoding for saponin biosynthesis (Bhat et al., 2012). The structural similarity indicated it to be a functional protein.

Phylogenetically, a high degree of conservation among sequences suggests a low rate of evolution. Bacillus subtilis showed 20% identity to CbSqS amino acid sequence. Most of the plant SqS amino acid sequence showed 72.4%-78.2% identity with CbSqS amino acid sequence. SqS amino acid sequences from H. sapiens, S. cerevisiae and B. subtilis grouped together on the phylogenetic tree with 100% confidence value and as expected, these showed very less similarity of 42.8%, 36.9% and 20% respectively with CbSqS amino acid sequence. We observed lower similarity of yeast SqS to plant SqS, compared to human SqS, supporting that during the course of evolution yeast and plant SqS diverged earlier than yeast and human SqS. CbSqS amino acid sequence as expected was phylogenetically closer to other plant SqSs.

CbSqS core structure was similar to that of the several class I isoprenoid biosynthetic enzymes whose crystal structures are known and the conserved feature in all the structures was an α-helical core surrounding a central active site cavity regardless of amino acid sequence identity of which one end was predominantly hydrophobic, and the other end was more hydrophilic and contains a signature “aspartate-rich sequence” (Lee and Poulter, 2008).

The substrate binding site was present at the core region of the enzyme structure. The conserved “aspartate - rich region” in the active site of the compound is consistent with other class I isoprenoid enzymes (Ashby and Edwards, 1990).
aspartate side chains are involved in binding multiple Mg$^{2+}$ ions that stabilize binding of diphosphate groups in the substrate. CbSqS 3D structure matches closely with SqS crystal structures from human. This can further help in studying the crystal structure of CbSqS. Evolutionary conservation of amino acids at the surface of CbSqS sequence was performed using ConSurf program. Conservation analysis of positions among members from the same family can often reveal the importance of each position for the protein (or nucleic acid)'s structure or function. The degree to which an amino (or nucleic) acid position is evolutionarily conserved is strongly dependent on its structural and functional importance; rapidly evolving positions are variable while slowly evolving positions are conserved.

**CbSE**

The sequence information from partial CbSE gene was used to design primers for RACE PCR to amplify full-length cDNA sequence. Amplicons were generated, and sequence analysis of ten different clones obtained from both the 5'- and 3'-RACE PCR revealed the full-length cDNA clone for CbSE gene (1893 bp). Deduced amino acid sequence of CbSE shared 80%, 80%, 79%, 76% and 71% similarity with *V. vinifera* (XP_002271528.1), *O. sativa* (NP_001049463.1), *Z. mays* (NP_001151921.1), *G. max* (XP_003529275.1) and *A. thaliana* (NP_564734.1) respectively. Conserved domain database (CDD) search at NCBI indicated the presence of conserved squalene epoxidase domain (214-487) which is found in taxonomically diverse group of eukaryotes and also in bacteria (Pearson et al., 2003). This domain contains a putative FAD binding domain (171-404). Another major motif and functional domain that was observed in CDD was an NAD (P) binding site whose dinucleotide binding fold is known as Rossmann fold. The conserved domains in the CbSE amino acid sequence were also determined using SMART tool. It was predicted using SMART tool that CbSE sequence contains many overlapping domains which included SE domain (214-487) (as predicted by CDD), FAD binding domain_3 (63-408), FAD binding domain_2 (65-97), DAO (65-99). DAO (D-amino acid oxidase [EC. 1.4.3.3]) has an oxido-reductase activity. It is a FAD flavoenzyme that is found to oxidize neutral and basic D-amino acids into their corresponding keto acids. FAD binding domains are involved in the binding of enzymes using FAD as cofactor. This domain is responsible for many catalytic properties in living systems (Mushegian and Koonin 1995).

The predicted secondary structure by SOPMA program indicated the similarity of CbSE to other plant SE secondary structure. These percentages are again in range with SE from other plant species encoding for saponin biosynthesis (Razdan et al., 2012).
Phylogenetically, a high degree of conservation among sequences suggests a low rate of evolution. *Methylobacterium alcaphilum* showed 37% identity to CbSE amino acid sequence. Most of the plant SE amino acid sequence showed 71%-80% identity with CbSE amino acid sequence. SE amino acid sequences from *H. sapiens*, *S. cerevisiae* and *M. alcaphilum* grouped together on the phylogenetic tree with 100% confidence value and as expected, these showed very less similarity of 37%, 37% and 36% respectively with CbSE amino acid sequence. CbSE amino acid sequence as expected was phylogenetically closer to other plant SEs such as *V. vinifera*, *O. sativa*, *Z. mays* etc.

Many amino acids involved in FAD binding in pHBH are conserved in SE, suggesting similar overall structure. Both enzymes require a source of reducing equivalents for monooxygenation; pHBH binds NADPH directly, whereas SE uses cytochrome P450 reductase as a source of electrons. Interestingly, pHBH lacks the typical NADPH binding domain found in most enzymes that utilize this cofactor, and is thought instead to bind NADPH at the interface that forms between paired identical molecules (the "dimer interface"; Eppink et al. 1998). pHBH shared 19% identity with the CbSE sequence. The predicted structure of CbSE consists of αβ domains. The FAD2 binding domain consists of the typical βαβ sandwich which consists of consensus sequence ("DVIVGAVAGSALAYTLGKDGRVHIER") that is highly conserved for the binding of FAD cofactor in the flavoprotein aromatic hydroxylase class of protein family. This FAD-binding motif located in the N-terminal region is a part of the well-known Rossmann fold, which is broadly shared by flavoprotein aromatic hydroxylase class of enzymes. Another motif FAD3, in this class of enzymes is rather not conserved and is involved in the binding of both FAD and NADPH. In CbSE, consensus sequence "GDAFMMRHPLTGGM" represents the FAD3 domain. The substrate binding site was present at the core region of the enzyme structure. These binding sites are conserved and are a part of FAD binding domain and SE binding domain.

5.1.3 High Throughput Sequencing of Leaf Tissues of *C. borivilianum*

As the sequence information for CAS and BAS genes was scanty and degenerate PCR primer based approach was not successful; so to clone CAS and BAS from *C. borivilianum*, we opted for high throughput transcriptome sequencing based approach.
The Illumina high throughput sequencing technology yields huge amount of parallel sequence reads, though short reads but provide high coverage. Global analysis of *C. borivilianum* transcriptome from the leaf tissues was done. The study reported a strategy for *de novo* assembly of transcriptome using short-read sequence data generated by Illumina RNA-Seq method in lieu to discover the candidate genes that encode enzymes in the steroidal saponin biosynthetic pathway. The coverage of the transcriptome, at 22.42 megabase pairs, was comprehensive enough to discover all known genes of several major metabolic pathways.

In the annotated *C. borivilianum* transcriptome dataset, multiple transcripts encoding almost all known enzymes involved in the MVA pathway, MEP pathway and saponin biosynthesis pathway were identified. In almost all the cases, more than one unique sequence was annotated as the same enzyme. These unique sequences may either represent different fragments of a single transcript or different members of a gene or both. Steroidal saponins, borivilianosides, are the major type of saponins present in *C. borivilianum* (Acharya et al., 2009). There are yet no reports of presence of triterpenoid type of saponins in *C. borivilianum*. Surprisingly, two *C. borivilianum* singleton sequences (Transcript ID: 664097 and 798016) matched to β-amyrin synthase and Dammarenediol II synthase of *V. vinifera* and *R. communis* respectively. The copy number of these two genes was very low as compared the genes that are involved in steroidal saponin biosynthesis. Therefore, it can be presumed that in *C. borivilianum*, the triterpenoid saponins were present at levels too low to be detected phytochemically. For this reason it was not possible to clone β-amyrin synthase gene via candidate gene approach from this plant.

Surprisingly, some genes associated with flavonoid biosynthesis pathway were also detected in the plant. Genes of the phenylpropanoid pathway have already been reported in several plant species such as *Camellia sinensis* (Takeuchi et al., 1994, Lin et al., 2007), *Petroselinum hortense* (Parsley) (Kreuzaler et al., 1983), *Zea mays* (Goff et al., 1990), *Arabidopsis thaliana* (Shirley et al., 1992), *Vitis vinifera* (Boss et al., 1996a), citrus (*Citrus unshiu* Marc.) (Moriguchi et al., 2002) and strawberry plants (*Fragaria spp.*) (Manning, 1998). Flavonoids are a class of low-molecular weight phenolic compounds that are widely distributed in plant kingdom. Due to immense diversity in their structure, flavonoids possess antioxidant, anti-cancerous and several medicinal values. Genes obtained showed the vertical pathway responsible for the formation and conversion of the sub-categories of flavonoids. At present, there are no reports of existence of flavonoid pathway in *C. borivilianum*.
Alkaloids are heterocyclic nitrogen compounds biosynthesized from amino acids. With currently more than 12,000 known structures, alkaloids represent one of the biggest groups of natural products (Caporale, 1995). The potent biological activity of some alkaloids has led to their exploitation as pharmaceuticals, stimulants, narcotics and poisons. Plant-derived alkaloids currently in clinical use include, analgesics morphine and codeine, the anticancer agents vinblastine and taxol, the gout suppressant colchicine, the muscle relaxant (C)-tubocurarine, the antiarrhythmic ajmaline, the antibiotic sanguinarine, and the sedative scopolamine. Other important alkaloids of plant origin include caffeine, nicotine, cocaine, and the synthetic O, O-acetylated morphine derivative heroin. Tyrosine decarboxylase (TYDC), the only enzyme involved in the early steps of alkaloid biosynthesis has been purified (Marques and Brodelius, 1988), and for which the corresponding cDNA has been cloned (Facchini and Luca, 1994, Maldonado-Mendoza et al., 1996). TYDC cDNAs have also been reported from parsley (Kawalleck et al., 1993) and Arabidopsis thaliana (Trezzini et al., 1993), which do not accumulate tyrosine-derived alkaloids. TYDC mRNAs were shown to be rapidly induced in response to elicitor treatment (Kawalleck et al., 1993; Trezzini et al., 1993) and pathogen challenge (Schmelzer et al., 1989) in various plants. Induction of TYDC mRNAs in parsley and Arabidopsis suggests that tyramine serves as the precursor to a ubiquitous class of defense response metabolites, in addition to BIA. The SAM-dependent 6-O- and 4′-O-methyltransferases (6-O-MT and 4′-O-MT, respectively) have been purified from cultured Coptis japonica cells (Sato et al., 1994), and the corresponding cDNAs isolated and characterized (Morishige et al., 2000). However the enzymes that are involved beyond BBE in this pathway could not be discovered. The probability of the absence of such enzymes may be due to transient expression of those genes or their expression is too low to be discovered phytochemically.

Another revelation came out when we discovered some genes associated with isoflavonoid metabolism also in our transcriptome. C. borivilianum is not known to produce isoflavones. Similar kinds of findings have been reported in other non-iso flavone accumulating plants (Yu and McGonigle, 2005), suggesting homologs of isoflavone reductase and isoflavone O-methyltransferase may have general metabolic functions in many plant species.

We could also discover several Cytochrome P450 (CYP 450) genes that are involved in saponin and flavonoid biosynthesis pathways. CYP 450s are versatile biocatalysts. These enzymes form the largest family of plant proteins (http://drnelson.utmem.edu/CytochromeP450.html). CYP450s are involved in the
NADPH-dependent regio- and stereo-specific oxygenations during the biosynthesis of terpenoids, sterols, lignins, hormones, fatty acids, pigments, and phytoalexins in plants (Meijer et al., 1993). In the biosynthetic pathway of steroidal-type of saponins, CYP450 catalyzes the conversion of obtusifoliol to \( \Delta^{6,14} \) - sterol. Previous studies have characterized \textit{obtusifoliol 14-alpha-demethylase} (EC 1.14.13.70, 7 unigenes) of CYP51 family from \textit{Sorghum bicolor} (Bak et al., 1997) and \textit{Triticum aestivum} (Cabello-Hurtado et al., 1997) and CYP 710 (Morikawa et al., 2006) from \textit{Arabidopsis}, both of which are involved in steroid saponin biosynthesis. Therefore, the CYP450s belonging to CYP51 and CYP710 family might be involved in borivilianoside biosynthesis in \textit{Chlorophytum} genus.

Besides these, several CYP 450s involved in flavonoid biosynthesis were also discovered. Cinnamate 4-hydroxylase (C4H, EC 1.14.13.11), a member of CYP450 monooxygenase superfamily was discovered. The enzyme requires NADPH, oxygen and acts as a typical CYP450 system requiring a second enzyme system NADPH: CYP450 reductase. Another group of CYP450s is flavonoid hydroxylases (F3H, F3’H & F3’5’H). Flavanone 3-hydroxylase (F3H) catalyzes stereospecific 3β-hydroxylation of (2S)-flavanones to the (2R, 3R)-3-hydroxyflavanones (dihydroflavanols). Hydroxylation at the 3-position of both naringenin and eriodictyol \textit{in vitro} was observed by Forkmann and coworkers (1980) using the preparations of \textit{Mathiola incana} flowers. They showed that the enzyme was soluble and requires 2-oxoglutarate, \( \text{Fe}^{2+} \), ascorbate as cofactor and molecular oxygen for its activity.

Dihydrokaempferol (DHK), the product of F3H catalyzed hydroxylation of naringenin, can be further hydroxylated, either at the 3’ position or at both 3’ and 5’ positions of the B-ring. The former reaction leads to the formation of dihydroquercetin (DHQ) and ultimately to the production of cyanidin based pigments. This is carried out by the P450 hydroxylase flavonoid 3’-hydroxylase (F3’H). The latter hydroxylation steps are catalysed by the P450 enzyme flavonoid 3’, 5’-hydroxylase (F3’5’H), responsible for the conversion of DHK into dihidromyricetin (DHM), which is required for the production of delphinidin-based anthocyanins (Forkmann, 1991; Winkel-Shirley, 2001; Toda et al., 2002). Both enzymes require oxygen and NADPH as cofactor.

F3’H enzyme activity has been detected in floral extracts from a number of ornamental plants like \textit{Matthiola incana} (Forkmann et al., 1980), \textit{Antirrhinum majus} (Forkmann and Stotz, 1981), \textit{Dianthus caryophyllus} (Spribille and Forkmann, 1982) and \textit{Petunia hybrida} (Stotz et al., 1985). Brugliera et al., (1999) were first to report the F3’H cDNA corresponds to the \( Ht1 \) locus of petunia which controls the
hydroxylation of dihydrokaempferol to dihydroquercetin and of naringenin to eriodictyol. F3’H transcript was found to be most abundant in petals from flowers at early stage of development and level declined as flower matures. Transcripts were also detected in the ovaries, sepals, peduncles, stem and anthers. Later F3’H cDNA was isolated from Arabidopsis (Schoenbohm et al., 2000) and Torenia (Ueyama et al., 2002). Their activity towards naringenin has been detected using a yeast expression system.

F3’5’H belongs to the P450 monoxygenase superfamily of membrane bound homoproteins which interact with NADPH-P450 oxidereducatase. F3’5’H cDNAs have been isolated from Petunia hybrida (Holton et al., 1993), Solanum melongena (Toguri et al., 1993) and Gentiana triflora (Tanaka et al., 1996). Nielsen and Podivinsky (1997) isolated F3’5’H from Eustoma grandifolium petal cDNA library using petunia cDNA as probe. Northern analysis showed that the F3’5’H mRNA was transcribed late in petal development and was concomitant with appearance of colored anthocyanins in the petal tissue that is transcript was more abundant in outer petal tissue compared to the inner, deeply pigmented throat. Vetten et al. (1999) identified and additional petunia gene that is required for 3’, 5’ substitution of anthocyanins and purple flower colors. It encoded a cytochrome b5 and was expressed exclusively in the flowers. Inactivation of the gene by targeted transposon mutagenesis reduced F3’5’H enzyme activity and the accumulation of 5’-substitued anthocyanins, resulting in an altered flower color.

Among the reactions for the modification of secondary metabolites, glycosylation plays an important role in plants, contributing to the biosynthesis and storage of secondary metabolites (Bowles et al., 2006). UDP-glycosyltransferases (UGTs) catalyze the transfer of sugar residues from uridine diphosphate sugars to an acceptor. UGTs comprise a superfamily of enzymes ubiquitous in living organisms (http://afmb.cnrs-mrs.fr/CAZY/fam/acc_GT.html), the number of which in the plant kingdom is likely comparable to that of the cytochrome P450 enzymes. Soyasaponin III rhamnosyltransferase is involved in the biosynthesis of soyasaponin I in Glycine max (Shibuya et al., 2010). The enzyme has strong sugar donor specificity for UDP-rhamnose. It does not show any activity with UDP-glucose, UDP-galactose or UDP-glucuronic acid. Sterol 3- beta- glucosyltransferase on the other hand is involved in the glycosylation activity of different type of sterols (e.g. sitosterol, stigmasterol etc.) present in the plant. This enzyme is thought to be involved in steroid metabolism and steroid biosynthesis pathways. Presence of many glycosylation enzymes and
cytochrome P450 genes might contribute to the extensive modifications of various secondary compounds found in *C. borivilianum* leaf tissues.

Based on EC classification, a large amount of assembled transcripts belonged to non-specific protein tyrosine kinases enzyme class alone (14.7%). In *A. thaliana*, this enzyme controls the shoot and floral meristem size (Clark et al., 1997) and also contributes to signal transduction (Hirayama and Oka, 1992). Presence of this enzyme in such abundance may suggest that the plant is in active metabolic phase.

### 5.1.3.1 GC content and SSR analysis

Deep sequencing of *C. borivilianum* transcriptome rendered us with an opportunity to calculate the ratio of guanine cytosine (GC content) of all the transcripts that were obtained. Knowledge of GC content helps in understanding the ecology and evolution of particular taxa. It also plays a vital role in gene and genome regulation and also helps in determining the physical properties of the genome. It is an important indicator of stability of DNA (Carels et al., 1998). Average GC content of *C. borivilianum* transcripts is in range with the GC levels of coding sequences of *O. sativa* (43.6%) which also belong to monocot family

Assembly of *C. borivilianum* was further assessed for the molecular markers. Morphological as well as biochemical markers are used in the authentication of herbal drugs. SSRs (Simple Sequence Repeats) or microsatellite markers are polymorphic stretches of 1 to 6 nucleotide units repeated in tandem and randomly spread in eukaryotic genomes. SSRs are generally associated with functional and phenotypic variations. Moreover, SSRs allow identification of many alleles at a single locus, are easy to develop, distributed evenly all over the genome and are codominant (Wei et al., 2011). SSRs are particularly useful when genome information of the crop is lacking. Since *C. borivilianum* is a cross pollinated species (Geetha and Maiti, 2001), and hence the seed raised population will have variability. This variability accounts for the difference in saponin contents between two species of *C. borivilianum* collected from different environment samples. The identification of SSRs in *C. borivilianum* will help in distinguishing closely related individuals and will help in preservation of germplasm of the plant. Most prevalent of mononucleotide was poly-A. The edge of poly A mono-nucleotide SSRs over others may be due to the presence of poly A tails in the RNA sequences. The di-nucleotide SSRs were the most abundant in the transcripts, which is similar to results obtained from other plants (Senthilvel et al., 2008). The CG repeats were very infrequent in the plant (0.6%).
which is consistent with previous observations (Varshney et al., 2002; Senthivel et al., 2008). These findings indicated that unique sequences containing SSR markers were indeed abundant in *C. borivilianum*. These unique sequence-derived markers generated in this study represent a valuable genetic resource for future studies of this species as well as related Chlorophytum species.

### 5.1.3.2 Transcription Factors

Transcription factors are sequence specific DNA binding proteins that interact with the promoter regions of a gene and modulate its expression. Associative modulation of several plant processes suggests involvement of transcription factors (TFs) for coordinated regulation of gene expression. Transcription factor encoding transcripts were identified from *C. borivilianum* transcriptome by sequence comparison to known transcription factor gene families. MYB family proteins are characterized by DNA-binding domains and it is the largest transcription factor family in Arabidopsis as well, where it comprises of 163 genes (Yanhui et al., 2006). These TFs have been associated with varied processes. It has been observed that many protein members of the MYB, bZIP and WRKY transcription factor families insinuates the regulation of stress responses (Singh et al., 2002). On the other hand, members of C3H family are involved in embryogenesis (Li and Thomas, 1998) whereas members of PHD family TFs are involved in vernalization processes (Sung et al., 2006). Members of bHLH are involved in controlling cell proliferation and in the development of specific cell lineages (Heim et al., 2003) and bZIP, yet another family of TFs, regulate processes including pathogen defence light and stress signalling seed maturation and flower development. The complete picture about evolution of various transcription factor families will emerge once the complete genome sequence of *C. borivilianum* will be available.

**CbCAS**

Full length *CbCAS* gene (2499 bp) was discovered from the high throughput sequencing data. Deduced amino acid sequence of *CbCAS* shared 89%, 76%, 73%, 75% and 76% similarity with *Zea mays* (NP_001168640.1), *Avena strigosa* (AAT38891.1), *Ricinus communis* (Q2XPU6.1), *Panax ginseng* (O82139.1) and *Arabidopsis thaliana* (XP002863756.1) respectively, suggesting that *CbCAS* codes for cycloartenol synthase.

Currently there is no experimentally determined three-dimensional structural information available for OSCs, although studies with a related enzyme, squalene-hopene cyclase (SC; EC 5.4.99.7) have proved informative. Squalene cyclases are
involved in the direct cyclisation of squalene to pentacyclic triterpenoids known as hopanoids, which play an integral role in membrane structure in prokaryotes (Ourisson et al., 1987). A number of SC genes have been cloned from bacteria (Ochs et al., 1992; Reipen et al., 1995). The SC and OSC enzymes have related predicted amino acid sequences, and so should have similar spatial structures (Wendt et al., 1997). The crystal structure of recombinant SC from the Gram-positive bacterium *Alicyclobacillus acidocaldarius* has established that the enzyme is dimeric (Wendt et al., 1997). Each subunit consists of two α-α barrel domains that assemble to form a central hydrophobic cavity (Wendt et al., 1997, Wendt et al., 1999). The activity of 2, 3-oxidosqualene cyclases is associated with microsomes, indicating their membrane-bound nature. However, the predicted amino acid sequences of these enzymes generally lack signal sequences and obvious transmembrane domains. Addition of hydrophobic membrane-localizing regions to OSCs during evolution may have removed selection pressures that maintained alternate mechanisms for membrane localization (Buntel and griffin, 1994). Consistent with this, there is a non-polar plateau on the surface of the *A. acidocaldarius* SC enzyme which is believed to be immersed in the centre of the membrane. The squalene substrate for SC is likely to diffuse from the membrane interior into the central cavity of the enzyme via this contact region (Wendt et al., 1997, Wendt et al., 1999). Mechanism-based irreversible inhibitors and mutational analysis with OSCs have shown that the highly conserved amino acid motif DCTAE is required for substrate binding (Abe and Prestwich, 1994, Abe and Prestwich, 1995, Feil et al., 1996), and the conserved aspartate residue within this motif (D456) has been implicated as the likely electrophilic activator in the generation of the protosteryl cation for LAS (Abe and Prestwich, 1994, Abe and Prestwich, 1995, Corey et al., 1997). Similar experiments indicate that two aspartate residues at the homologous position of the *Alicyclobacillus acidocaldarius* SC amino acid sequence (amino acids 376 and 377 of the DDTAV motif) are also essential for enzyme activity (Feil et al., 1996). These residues are located in the large central cavity of the dimeric enzyme (Wendt et al., 1997). Interestingly, targeted mutations that convert the DDTAV motif of SC to DCTAE (the corresponding OSC motif) result in a change in substrate specificity from squalene to 2, 3-oxidosqualene (Dang and Prestwich, 2000).

In addition to the DCTAE/DDTAV motifs, a highly conserved repetitive β strand turn motif rich in aromatic amino acids (the QW motif) occurs in all OSCs and SCs and is repeated four to eight times. These repeats are likely to be important for protein structure and stability and also for catalytic activity (Wendt et al., 1997,
Poralla et al., 1994, Sato et al., 1998, Poralla, 1994). The aromatic amino acids of the QW motif have been proposed to constitute sites of negative point charge that may interact with the intermediate cations during the cyclization process (Poralla et al., 1994).

**CbBAS**

Partial 

Partial CbBAS gene (1449 bp) was discovered from the high throughput sequencing data. Analysis using BLASTX showed that the deduced sequence is of BAS gene and the cloned gene had ~700bp missing at the 5’ end of the gene. The CbBAS sequence obtained showed considerable homology to Isoprenoid Class_2 like superfamily in BLASTx analysis. We could obtain only a single copy of BAS in next generation sequencing data as compared to the another OSC, CAS for which 21 copies of the gene have been obtained. This may be because; the pathway of saponin biosynthesis in C. borivilianum is inclined to the production of steroidal saponins which proceeds from CAS as compared to triterpenoid production which involves BAS. So, no further analysis was carried out on CbBAS.

### 5.2 Expression patterns of CbSqS, CbSE and CbCAS in the leaf and root tissues of the plant

*C. borivilianum* is a traditional medicinal plant was originally grown in thick forest in natural form (Nayar and Sastry, 1988). It is a monocot and perenates in the soil by fleshy roots attached to a condensed stem disc which remains dormant in the soil for about seven months. The fleshy roots are economically important, which contain saponins and are used for the preparation of many Ayurvedic tonics. The stem is condensed to a disc from which is produced a whorl of leaves that are long, sessile and somewhat thick. Expression of CbSqS was, therefore, analysed in the leaves and root tissues of the *C. borivilianum* by semi quantitative gene expression studies. Based on the IDV calculation, it was seen that the expression of CbSqS was 14.3% higher and the expression of CbSE was 17.6% higher in leaves as compared to the root tissues. The result supports the expression of SqS from *Withania somnifera* which predicted that the level of expression of SqS was higher in leaves as compared to the stem and root tissues (Bhat et al., 2012). Kim et al., (2001) conducted the *in vitro* hybridization studies on various isoforms of SqS in *Panax ginseng* and reported the expression and accumulation of SqS near the conducting tissue of petioles. The RT-PCR analysis of different isoforms of SE in *Arabidopsis thaliana* revealed that SE1 expresses more in the leaf tissues than SE2 isoform of the gene (Rasberry et al., 2006). The expression pattern of CbSE
revealed that the cloned gene is thus SE1 type isoform. Jinting and Zhenghai, (2008) reported that saponins in Achyranthus bidentata were first synthesized in the leaves and then transported to the roots via phloem of veins and stem vascular bundles for storage. The expression pattern of CbSqS and CbSE supports the fact that initial reactions of isoprenoid biosynthetic pathway occur in leaves, while later step modifications and storage occurs in roots (Kumar et al., 2012). However, the expression of CbCAS was 21.4% higher in roots as compared to leaves. This probably suggests that OSC genes of the saponin biosynthesis pathway are more active in roots than in leaves, probably due to accumulation of saponins in the roots of C. borivilianum. In Avena strigosa, a novel type of OSC, bAS1, expresses highly in the root tissues with little or no detectable transcript in other plant tissues (Haralampidis et al., 2001) consistent with the organ-specific accumulation of the saponins (Turner, 1953) and with the biochemical information indicating that the root tips are the site of synthesis (Trojanowska et al., 2000; Trojanowska et al., 2001).

5.3 Heterologous Expression of CbSqS and CbSE in E. coli

Measurable levels of SqS polypeptide were observed in the extracts of E. coli that expressed the truncated SqS (pQE30-TruncCbSqS). Presence of microsomal membrane anchoring structure at the carboxy-terminal of the protein, might be the reason of low expression of the polypeptides in the E. coli cells transformed with pQE30-CbSqS. In accordance with this suggestion, Lee et al. 2002 reported that the carboxyl-terminal deletion of residues 389-411 in the hot pepper SQS resulted in the accumulation of a functionally soluble SQS protein. Similar kind of result was observed in Withania somnifera (Bhat et al., 2012). In case of CbSE, high levels of the CbSE protein were observed 2 h after the addition of 1mM IPTG and the expression of CbSE further increased after 4 h and 6 h. On the other hand, no expression was seen in the cells that were transformed with the vector sequence only. In W. somnifera the entire protein coding cDNA of WsSQE was cloned into expression vector pGEX-4T-2 and the expression of the protein was observed 2 h from induction with 1mMIPTG. (Razdan et al. 2012).

5.4 Future Prospects

Due to immense significance of C. borivilianum as a multipurpose medicinal plant, it is instigating to decipher the biosynthetic pathway regulating the saponin biosynthesis in this plant, and also to exploit the full biotechnological potential of this crop. Understanding of various enzymatic steps involved in the pathway is a prerequisite for metabolic engineering. These genes could be used for genetic
engineering for beneficial purposes such as increasing the saponins contents. Using degenerate primers and RACE PCR strategy, we have cloned full length cDNA encoding CbSqS and CbSE. Using Next Generation sequencing data, we could obtain full length sequence of CbCAS and partial sequence of CbBAS genes. The genes were analyzed in detail, expression analysis was performed in leaf and root tissues and recombinant protein was expressed in microbial expression system. Next Generation sequencing data gave further insights into several biosynthetic pathways operating in C. borivilianum. It was also helpful to calculate the GC content, SSRs in the transcriptome of C. borivilianum leaf tissues with this data. The cloning and characterization of these genes from C. borivilianum not only constitutes an important step unravelling the key committed step of the saponin biosynthetic pathway, but also throws a fresh prospect to understand the regulation of the isoprenoid pathway genes in this medicinal plant which represents a rich repository of bioactive metabolites. The present work could be followed in future as:

- Functional characterization of the promoters of genes involved in this study.
- Mining of isoforms of different genes involved in saponin biosynthesis.
- Strategies could be devised to transplant the saponin biosynthetic pathway in heterologous systems.
- Transcriptome data opens a new endeavor to study various metabolic networks in this plant, such as flavonoid and alkaloid pathway as well as transcription factors controlling secondary metabolism.