Plant Omics Journal POJ

POJ 4(7):354-363 (2011) ISSN:1836-3644

In silico approaches in comparative genomics, structure prediction and functional characterization of secondary metabolite proteins of Mentha sp.

Sudeep Roy[1]*, Rashi Chauhan[1], Nidhi Maheshwari[1], Sanchita Gupta[1], Dwijendra K Gupta[2] and Ashok Sharma[1]

[1]Biotechnology Division, Central Institute of Medicinal and Aromatic Plants, Council of Scientific and Industrial Research, Lucknow - 226015, India
[2]IIDS Center of Bioinformatics, Nehru Science Center, University of Allahabad, Allahabad - 211001, India

***Corresponding author: roysudeep28@gmail.com**

Abstract

Mentha species are the main source of a number of natural aroma chemicals. In this paper, an attempt has been made to study some important *Mentha* species, viz., *M. piperita, M. canadensis* and *M. arvensis. In-silico* approaches were used for EST assembly, comparative genomics, homology modeling, molecular threading, domain and fold recognition, secondary structure analysis, physicochemical and functional characterization. Totally, 164 different proteins were reported from contigs which were subjected to modeling. Detailed functional and structural characterizations were carried out for 53 different secondary metabolite proteins of *Mentha.* Physicochemical characterization of the proteins revealed their aliphatic index instability index and extinction coefficient values. These findings will help in further annotation of *Mentha* from EST level to protein level.

Keywords: EST assembly, Comparative genomics, Molecular modeling, Threading, Secondary metabolite proteins (SMPs)
Abbreviation: SMPs:_Secondary metabolite proteins, Expressed Sequence Tags (EST), Nucleotide Binding Protein (NDB3), Tata Binding Protein (TBP), Methionine adenosyltransferase 3 (MAT3), Arabidopsis thaliana binding protein (ATB2), Protein model database (PMDB), Root Mean Square Deviation (RMSD)

Introduction

Mentha species (Family Lamiaceae) are widely used for their flavoring and medicinal properties throughout the world. They are the main source of natural aroma chemicals namely menthol, carvone, linalool and linalyl acetate, commonly used in pharmaceutical preparations. Three species namely *Mentha piperita*, *M. arvensis* and *M. canadensis* are the most widely found species across the globe. *Mentha piperita* is currently one of the most economically important aromatic and medicinal crop. *Mentha* oil known as peppermint oil is used in toothache, rheumatism, muscular pains and to relieve menstrual cramps. *Mentha piperita* is currently used to treat irritable bowel syndrome, Crohn's disease, ulcerative colitis, gallbladder and bilary tract disorders and liver complaints. *Mentha arvensis*, a native of Japan, is cultivated extensively throughout India for its use as a food seasoner, household remedy and industrial purposes. The plant has been reported to possess diverse medicinal properties. *Mentha canadensis* (syn. *M. arvensis* var. *canadensis*) leaves have a distinct peppermint smell when pinched or crushed as (Husain et al., 1988, Husain, 1994, Husain et al., 1992). ESTs have now become a tool to refine the predicted transcripts for those genes, which leads to the prediction of their protein products and ultimately their function (Gupta et al., 2010). ESTs of three *Mentha* species were taken from NCBI. ESTs were assembled using EGassembler to generate contigs. Comparative genomic approach was used to retrieve the genes and chromosome location which is similar between *Mentha* species and the model plant *Arabidopsis thaliana*. Blastx was used for taking out homologous proteins that are present in three species of *Mentha.* In the present work, we have selected secondary metabolite proteins (SMPs) that are present in 3 species of *Mentha* for *In-silico* analysis. Plants produce numerous secondary metabolites that have versatile physiological and protective roles (Rohman et al., 2009; Pavli et al., 2011; Lee et al., 2011). Once the homologous proteins were taken out, SMP sequences were subjected to physio-chemical characterization. Disulfide bonds which play an important role in stability of protein were retrieved. The 3-D structure of protein is an important source of information to understand the function of a protein and its interactions with other compounds (ligands, proteins or DNA) (Tramontano, 1998). In case of *Mentha* only 2 protein structures are reported in Protein Data Bank. Hence in order to deduce 3D structures of *Mentha* proteins, comparative modeling has been performed. These putative 3D protein structures of *Mentha* have been validated with validation server and RMSD test was also performed to check the accuracy of these protein models. In case, protein sequences did not give any close homologues as possible templates for modeling, molecular threading has been performed. After building protein model structures, Profunc server was used to reveal the functional domains and folds that are present in SMPs of *Mentha*. Promotif analysis was done to deduce the secondary structures present in SMP, which ultimately fold into 3D structures of protein. In this way, starting from ESTs of *Mentha* available in NCBI, proceeding to contigs of assembled EST datasets of *Mentha*, we have functionally and structurally annotated the putative proteins of *Mentha* (Fig 1). The objective of these approaches is to generate a path

leading from ESTs to proteins for an organism with the usage of *In-silico* tools.

Results and discussions

EST assembly

Out of total 1316, 224 and 81 ESTs reported from *M. piperita*, *M. canadensis* and *M. arvensis*, 1211, 145, 60 respectively were successfully assembled into contigs. Remaining 105 (*M. piperita*), 79 (*M. canadensis*), 21 (*M. arvensis*) showed no overlap with any ESTs, thus called as 'singletons'. ESTs assembly resulted in the formation of 145 contigs in *M. piperita*. The data redundancy in terms of bp were recorded as 591215 bp. Out of 145 contig in *M. piperita*, 49 contigs reported expression at protein level in form of secondary metabolites .Similarly in case of *M. canadensis* EST's assembly resulted in the formation of 12 contigs that includes 96062bp data redundancy. Similarly for *M. arvensis* assembly output were 7 contigs, resulted in decrement of 22951bp data redundancy. For *M. arvensis* 1 contig out of 7 reported expressions at protein level in form of secondary metabolites. There were no contig expressed in form of secondary metabolite protein for *M. canadensis* (Table 1).

Comparative genomics

After EST assembly the resulted contigs obtained from 3 *Mentha* species were subjected to Blastn program. Location of these contigs was mapped against *Arabidopsis thaliana*. 48 out of 145 contigs in *M. piperita* gave us the regions in *Arabidopsis* genome having a similarity and these regions were found to occur maximum times on chromosome 5th of *A. thaliana* followed by 3rd chromosome. In *M. arvensis* homologous regions were found maximum in chromosome 3. In *M. canadensis* 1st, 2nd, and 5th chromosome showed homologous region (supplementary 1). Blastx was carried out between *A. thaliana* and 3 species of *Mentha*. Gene coding for chalcone-flavanone isomerase and chalcone synthase showed homology common between *Mentha* and *A. thaliana*. Some other genes having homology include MAT3, ATB2 etc. (Table 2, supplementary 2).

Blastx

Blastx was performed on each contig to search homologous proteins having significant match to translated contig sequence. Blastx was performed against non-redundant protein database. For this study, a significant match was defined as a sequence with E value <=3. For each contig, we took top 4 homologous proteins according to Bit's score. It is to mention that for contig numbers 25, 80,101,142 no suitable proteins were found in *Mentha piperita*. Similar types of observation were made for contig numbers 10, 11 in *M. canadensis* and for contig number 1 in *M. arvensis*. Thus Blastx result yielded 362 homologous proteins for *M. piperita*, 25 for *M. canadensis* and 20 for *M. arvensis*. Total 53 proteins were found as secondary metabolite proteins in *M. piperita* and 2 proteins in *M. arvensis* (Table 3).
Amongst all secondary metabolite proteins, 7 proteins were reported to be expressed by more than one contig for *M. piperita.* For *M. arvensis* no SMP protein is found to expressed by more than one contig and in *M. canadensis* no SMP were reported. Limonene hydroxylase is expressed by maximum 7 different contigs for *M. piperita*. It indicates that limonene hydroxylase is expressed in larger amount in *M. piperita* as compared to other proteins in *M. piperita*. A graphical representation for proteins expressed by individual contigs and their chromosome location obtained from Blastn run against *A. thaliana* is provided (Fig 2.). Proteins other than secondary metabolites in *M. piperita* with respect to their contig participation and its chromosome location have been provided (Fig 3.). In the figure 2, 1 to 5 chromosome are the chromosomes of *A. thaliana*, 6 to 28 are the numbers represented in following way. If contig location is obtained on both chromosome 1 and 2 of *A. thaliana* we indicated it as chromosome 6 for convenience to study in the form of graph. Similarly contig location on chromosome 1 and chromosome 3 of *A. thaliana* has been shown as chromosome 7. Applying same coding, chromosome 1 and chromosome 4 has been named as chromosome 8,chromosome 1 and 5 as chromosome 9, chromosome 2 and chromosome 3 as chromosome 10, chromosome 2 and chromosome 4 chromosome 11, chromosome 2 and chromosome 5 as chromosome 12, chromosome 3 and chromosome 4 as chromosome 13, chromosome 3 and chromosome 5 as chromosome 14, chromosome 4 and chromosome 5 as chromosome 15, chromosome 1 and chromosome 2 and chromosome 3 as chromosome 16, chromosome 2 and chromosome 3 and chromosome 4 as chromosome 19, chromosome 3 and chromosome 4 and chromosome 5 as chromosome 22, chromosome 1 and chromosome 2 and chromosome 4 as chromosome 23. chromosome 1 and chromosome 3 and chromosome 4 as chromosome 24, chromosome 1 and chromosome 3 and chromosome 5 as chromosome 25, chromosome 1 and chromosome 2 and chromosome 5 as chromosome 26, chromosome 2 and chromosome 4 and chromosome 5 as chromosome 27, chromosome 1 and chromosome 2 and chromosome 3 and chromosome 4 as chromosome 17, chromosome 2 and chromosome 3 and chromosome 4 and chromosome 5 as chromosome 20, chromosome 1 and chromosome 2 and chromosome 3 and chromosome 5 as chromosome 21, chromosome 1 and chromosome 3 and chromosome 4 and chromosome 5 as chromosome 28, chromosome 1 and chromosome 2 and chromosome 3 and chromosome 4 and chromosome 5 as chromosome 18. If no location was found then it was indicated as chromosome 29. Same method is followed for figure 3.

Comparative modeling

The 3D modeled structure for putative *Mentha* proteins have been generated by Modeller9v8. Blastp was performed to select templates for generating the putative 3D modeled structure. For each template modeler generates five models. The 3-D structures were visualized by Swiss-PDB Viewer. The modeled structure was further subjected to energy minimization by GROMOS96 program, implemented in Swiss model software. The best model was selected after analyzing the result generated for each models by four different validation programs present in SAVES server. The results of the Procheck analysis in the form of Ramachandran plot for all *Mentha* proteins indicate that a relatively low percentage of residues have phi/psi angles in the disallowed ranges. The percentage of residues in the core region was found to be within the range of 70% to 96%. Residues in the disallowed region were within the range of 0% to 1.5%. Bad contacts were removed and brought down within acceptable range of 0 to 2. Overall quality factor for the model was calculated by Errat program (Ramachandran et al., 1963) whose value ranges from 51 to 95. Verify_3D result showed that 55-90% of the residues of all proteins in *Mentha* had an

Table 1. Statistics of EST assembly in *Mentha species*

Species Name	Number of EST	Total number of assembled sequences	No of contigs	No of singletons	Longest contig (bp)	Smallest contig (bp)	Average contig length (bp)	Total size of examined sequences before assembly (bp)	Total size of examined sequences after assembly (bp)	No. of ESTs forming contig sequences	Reduction in redundancy (bp)
M. piperita	1316	250	145	105	1692	377	650	694554	103339	1211	591215
M. canadensis	224	91	12	79	866	260	432	101246	5184	145	96062
M. arvensis	81	28	7	21	1259	279	598	27138	4187	60	22951

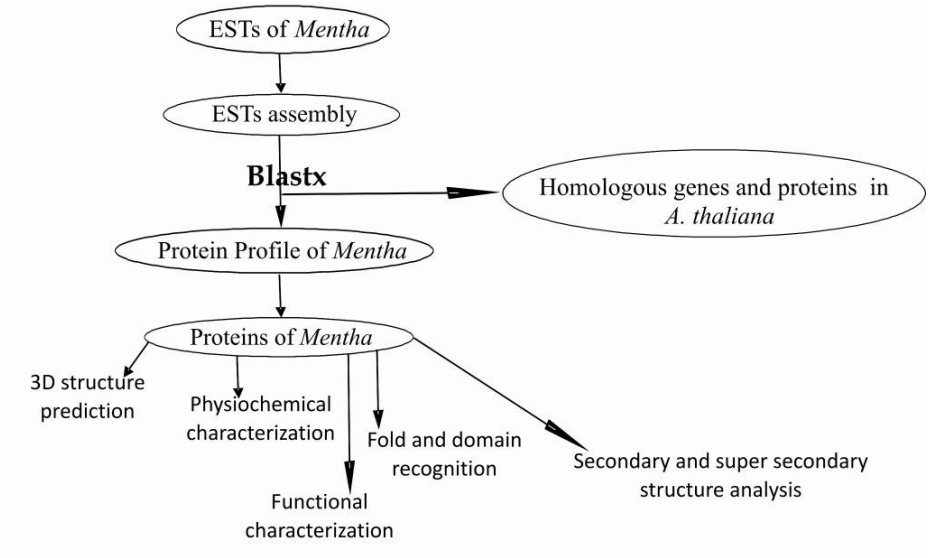

Fig 1. Methodology used in present study

average 3D-1D score > 0.2. Figure 4 represents modeled structure of Isopiperitone dehydrogenase.

Result of validation server

Procheck analysis of Isopiperitone dehydrogenase revealed that 86.1% of amino acid residues are in core region, 11.7% in allowed region, 1.7 % in generously allowed region and only 0.4% of amino acid are in disallowed region (Fig 5.). Bad contacts were calculated out to be 0. Verify-3d results show 69.17% of the residues had an average 3D-1D score >0.2. Errat showed an overall quality factor 85.547.

Submission of protein models in PMDB Database

The homology model of hitherto unreported 362 proteins of *M. piperata*, 25 of *M. canadensis* and 20 of *M. arvensis* was submitted to PMDB (http://mi.caspur.it/PMDB/). Submitted modeled structure has been assigned a specific accession number. This accession number can be used to retrieve the submitted protein structure (Table 4, supplementary 3).

Authentication of the predicted model

The weighted root mean square deviation of C trace between the template and final model for *Mentha* species modeled proteins were found to be between 0.22 to 3.23 as retrieved from TM align. In only 4 cases, the RMSD value came above 2Å. Remaining all case, RMSD value came below 2 which suggest further that the models are reliable. TM align also calculates TM align score. TM-score < 0.2 indicates that there is no similarity between two structures; where are a TM-score > 0.5 means the structures share the same fold. In all predicted models of *Mentha* TM Align score was computed greater than 0.5 which implies that template and modeled target structures of *Mentha* share the same fold. In only three cases TM aligns score came below than 0.5. TM align score finally prove that structure model of *Mentha* proteins are reliable. As judged by deep view-Swiss PDB viewer none of the residues were found to make clashes in their existing position which means that the residues occupied locations that were not impossible due to steric hindrances (Table 5, supplementary 4).

Threading

The blastp search for proteins for *Mentha piperita*, viz., dihydrin 13**,** vascular specific protein 4. Cytochrome P450 like_TBP in *M. arvensis* and plant senescence-associated protein in *M. canadensis* did not result in suitable templates. Therefore, protein fold recognition methods were applied .Molecular threading was performed using Biosuite software v3 .Threading module of Biosuite predicted five folds and their corresponding scores for each set of sequences. Biosuite calculates the folds using the fold database contained in it. Detailed result of the folds and their scores are given in the Table 4. Details of one fold, which is having maximum score amongst 5 folds as predicted by Biosuite for each *Mentha* protein sequences, are provided in Table 6, supplementary 5.

Physiochemical and functional characterization

Computation of various physical and chemical parameters for a given SMPs of *Mentha piperita* and *M. arvensis* was performed using ProtParam. Various physical and chemical calculations such as molecular weight, isoelectric point, and extinction coefficient have been calculated. Molecular weight was observed between the ranges of 10678.1-83708.5 for all secondary metabolite proteins in *Mentha*. Eleven proteins were found to have pI greater than 7 which indicates that they are basic in nature. Remaining SMPs computed to have Pi value below than 7 contribute to their acidic nature. Isoelectric point (overall charge) of SMPs can help in separation of proteins on a polyacrylamide gel using a technique called isoelectric focusing, which uses a pH gradient to separate proteins. The extinction coefficients for the proteins were also calculated which tells about wavelength-dependent molar absorptivity coefficient of the protein within the units of M^{-1} cm^{-1}. The extinction coefficient provides an indication of the amount of light that a given protein will absorb at a certain wavelength (usually 280 nm). Instability index analysis reveals the presence of certain dipeptides occurring at significantly different frequencies between stable and unstable proteins. Proteins with an instability index less than 40 are predicted to be stable, whereas those with a value greater than 40 are predicted to be unstable. Instability index for SMP has been calibrated.

Instability index analysis showed 24 SMP are unstable proteins that includes geranylgeranyl diphosphate synthase, 1-deoxy-D-xylulose 5-phosphate synthase, linalool synthase, geranyl diphosphate synthase small subunit. The aliphatic index refers to the relative volume of a protein that is occupied by aliphatic side chains (alanine, isoleucine, leucine and valine) and contributes to the increased thermo stability observed for globular proteins. Aliphatic index analysis reveals high value of aliphatic index for all SMP of *Mentha*. Higher aliphatic index of *Mentha* proteins indicates that their structures are more stable over a wide range of temperature. The GRAVY value for a peptide or protein is calculated as the sum of hydropathy values of all the amino acids, divided by the number of residues in the sequence. The SMPs which have large negative values indicates that these proteins have relatively more hydropathicity as compared to proteins which have less negative values. Physiochemical characterization for SMPs of *M. piperita* and *M. arvensis* is provided in supplementary 6 and 7, respectively.

Disulphide bonds play an important role in stability and folding of proteins. Presence of disulphide bond was also seen in some of SMPs including terpinolene synthase, limonene synthase, monoterpene synthase etc. Disulphide bridges were find using Cys_REC tool from softberry (Table 7).

Domain and fold recognition

The SMPs were studied by various computational tools to find out their 3-dimensional conserved region i.e. domains. These conserved regions (domains) are the 3-dimensional packing of amino acids, which play an important role in the activity of proteins. A domain is a compact arrangement of secondary structures connected by linker polypeptides (Campbell and Downing 1994). Several motifs pack together to form compact, local, semi-independent units called domains. Interproscan analysis resulted in the retrieval of 45 different kind of functional domains present in secondary metabolite proteins of *Mentha.* Among 45 domains, 14 were found to common in some of secondary metabolite proteins viz; terpenoid synthases, NAD (P)-binding domain etc. Some of the motif which showed high frequency of occurrence is shown in Figure 6. The Profunc server identified folds present in 51 proteins associated with secondary metabolites. Some of folds showing high frequency of occurrence include crystal structure of teas w273s/c440w and crystal structure

Table 2. Contigs of *M. piperita* showing 70% or above identity in Blastx search against *A. thaliana*

Contig number	Homologus protein in *A. thaliana*	Percentage identity	Coding gene with locus	Protein function
Contig2	S-adenosyl-L-homocystein hydrolase	89%	AAP92453	Catalyzes the hydrolysis of S-adenosyl-L-homocysteine(AdoHyc) to form adenosine (Ado) and homocysteine (Hcy)
Contig3	MAT3 (methionine adenosyltransferase 3)copper ion binding / methionine adenosyltransferase	88%	MAT3 at locus AT2G36880	Copper ion binding, methionine adenosyltransferase activity, one-carbon compound metabolic process, S-adenosylmethionine biosynthetic process
Contig6	Quinone oxidoreductase-like protein	75%	AAM63201	Zinc-binding dehydrogenase
Contig7	Geranylgeranyl pyrophosphate synthase	77%	AAA32797	Isoprenyl diphosphate synthasese which synthesis various chain length

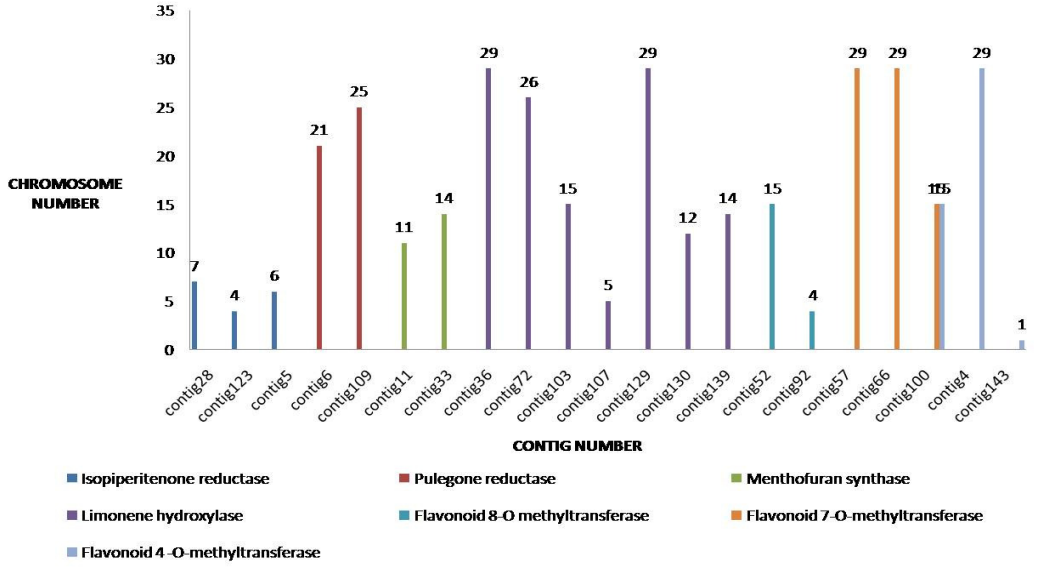

Fig 2. Contigs for SMPs with their corresponding chromosome location

analysis of chalcone o-methyltransferases. These folds are assumed to play a specific function in the secondary metabolite proteins of *Mentha* proteins (Fig 7.).

Secondary and super secondary structure analysis

Promotif analyses showed that the predicted structure of 53 secondary metabolite proteins of *Mentha piperita* have secondary structures which includes Beta sheets, Beta bulges, Beta turns, Beta hairpins, Gamma turns, helixes, Psi loops, 3-10 helix and super secondary structures includes Super secondary structures Beta hairpins (consists of two adjacent antiparallel β-strands joined by a small loop) and Beta alpha beta motif (used to connect two parallel β-strands). The Promotif results revealed that beta turns occur in high frequency among all SMPs in *Mentha piperita* followed by alpha helix and gamma turns. Beta sheets were found in lowest frequency than alpha helix in all SMPs. Beta hair pins, Beta alpha beta motif, showed less frequency of occurrence in SMPs. (supplementary 8).

Materials and methods

Retrieval of EST sequences

All the EST sequences of *Mentha* genus which include *Mentha piperita*, *M. arvensis* and *M. canadensis* were retrieved from dbEST database of NCBI (http://www.ncbi.nlm.nih.gov/dbEST). A total of 1632 ESTs comprising 1316, 224 and 81 ESTs each for *Mentha piperita, M. canadensis* and *M. arvensis,* respectively were retrieved.

EST sequence assembly

EST sequences were assembled using the contig assembly program EGassembler (http://egassembler.hgc.jp) (Masoudi-Nejad et al., 2006). EGassembler is a commonly used program which identifies overlapping sequences and generates contig with consensus sequences. EGassembler was run with default parameters used for assembly of EST sequences. The results were obtained in three different output

Table 3. Secondary metabolite proteins of *Mentha* species retrieved through Blastx and taken for study.

S.no		S.no	
	M. canadensis	24	Germacrene C synthase
1	No SMPs reported	25	Flavonoid 8-O-methyltransferase
	M. arvensis	26	Orcinol O-methyltransferase
1	Stress-associated protein 1	27	Resveratrol O-methyltransferase
2	Stress-associated protein 10	28	Flavonoid 7-O-methyltransferase
	M. piperita	29	Terpene synthase-like; Terpenoid synthase
1	(E)-beta-farnesene synthase	30	Isopiperitenol dehydrogenase
2	Sophorol reductase	31	Reticuline-7-O-methyltransferase
3	Isopiperitenone reductase	32	(+)-pulegone reductase
4	Menthol dehydrogenase	33	Chalcone isomerase
5	Pulegone reductase	34	Cytochrome P450 hydroxylase
6	Geranyl diphosphate synthase	35	Perakine reductase
7	Geranylgeranyl diphosphate synthase	36	Chalcone synthase
8	Geranylgeranyl pyrophosphate synthase	37	Cinenol synthase
9	Menthofuran synthase	38	Sabinene synthase
10	Gamma-cadinene synthase	39	Fenchol synthase
11	Flavonoid 4'-O-methyltransferase	40	1-Ddeoxyxylulose-5-phosphate synthase
12	d-Limonene synthase	41	(S)-N-methylcoclaurine 3'-hydroxylase
13	Cineole synthase	42	Geranyl diphosphate synthase small subunit
14	Terpinolene synthase	43	GGR (geranylgeranyl reductase); farnesyltranstransferase
15	Dihydroflavanol reductase 3	44	O-methyl transferase
16	3-Carene synthase	45	(-)P450 limonene-3-hydroxylase
17	Monoterpene synthase	46	1,8 cineole synthase 2
18	Vestitone reductase	47	Flavonoid 3'-O-methyltransferase
19	Germacrene A oxidase	48	Terpene synthase
20	Limonene hydroxylase	49	Linalool synthase
21	Shikimate dehydrogenase	50	4S Limonene synthase
22	Germacrene D synthase	51	Selinene synthase
23	Caryophyllene/alpha-humulene synthase	52	Salutaridine reducatse
24	Germacrene C synthase		

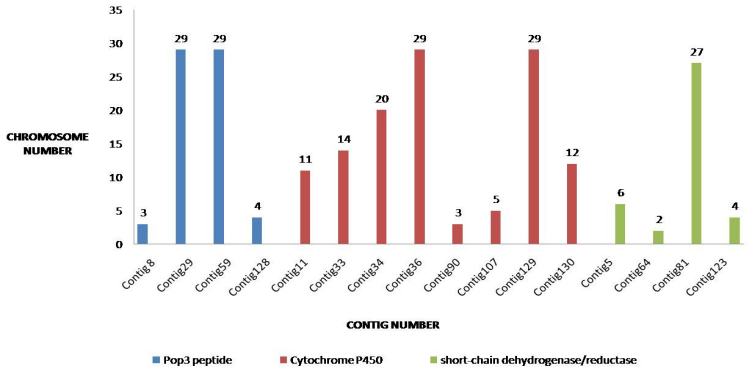

Fig 3. Contigs for Non SMPs and their corresponding chromosome location
In both the figures 3,4 graph is plotted between chromosomes of *A. thaliana* (five chromosomes) and contig number (expressed in the form of SMPs of *M. piperita* (145). Figure shows those SMPs which are expressed by more than one contig. Y axis shows chromosomes of *A. thaliana*.

files having contigs and singletons. A single output file also defined number of ESTs and their description that are responsible for formation of individual contigs. EG assembler run resulted in the formation of 145 contigs and 105 singletons, 12 contigs and 79 singletons, 7 contigs and 21 singletons for *Mentha piperita*, *M. canadensis* and *M. arvensis,* respectively.

Comparative genomics

In order to find the regions in *Arabidopsis* genome having similarity with 164 contigs in all three *Mentha* species, Blastn (http://blast.ncbi.nlm.nih.gov/Blast.cgi) was carried against *Arabidopsis thaliana*. Positions of contigs for all three *Mentha* species were found on five different chromosomes in *A. thaliana*. The generation of ESTs has proven to be a rapid and economical approach to identify and characterize expressed genes (Ton et al., 2000). Blastx (http://blast.ncbi.nlm.nih.gov/Blast.cgi) was carried out with each contig of all three *Mentha* species against *Arabidopsis thaliana* to find out the homologous genes, proteins and function of proteins. A threshold value of 70% identity is considered for doing this study.

Table 4. Listing of SMPs models submitted in PMDB

S.No	Protein	Organism	PMID
1	Limonene hydroxylase	*Mentha arvensis*	PM0076704
2	Isopiperitenone reductase	*Mentha piperita*	PM0076853
3	Menthol dehydrogenase	*Mentha piperita*	PM0076854
4	(+)-Pulegone reductase	*Mentha haplocalyx* var. *piperascens*	PM0077020
5.	(E)-Beta-farnesene_synthase	*Mentha piperita*	PM0076875
6	(-)P450 limonene-3-hydroxylase	*Mentha haplocalyx* var. *piperascens*	PM0076702

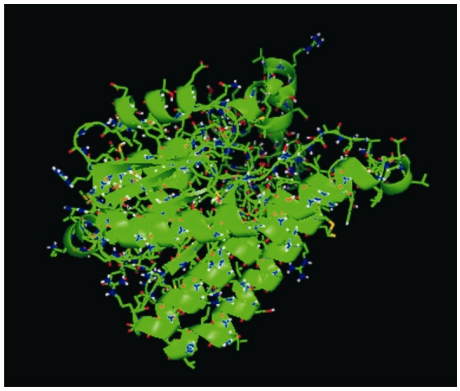

Fig 4. Modeled structure of isopiperitone as viewed by Pymol

Blastx

Blastx program available at NCBI was used to find out proteins which can be expressed by individual contig in all three *Mentha* species. 145 contigs of *Mentha piperita*, 12 contigs of *M. canadensis* and 7 contigs of *M. arvensis* were given as input to Blastx. Outputs of Blastx were proteins that can be expressed by all 164 contigs in all three *Mentha* species. Proteins were selected on the basis of high bit's score and low E-values. Proteins from both prokaryotes and eukaryotes were taken into consideration.

Selection of templates

The identification of templates for 145 proteins of *M. piperita, M. arvensis and M. canadensis* was carried out using Geno3d (Combet et al., 2002). For those proteins where we did not get any template, Blastp program was run against PDB database. The target-template showed sequence identity taken in range from 55% to 100% with E-value range from 0 to 4. Three templates were taken for each protein sequence or target sequence for homology modeling.

Model building

Knowledge gained from its 3-D structure of proteins with functionally important domains and structural features is essential for better understanding of regulatory mechanism at molecular level as well to target the protein metabolic engineering to enhance biosynthesis of menthol (monoterpenoids). Hence 3D modelingof 8 SMPs directly involved in menthol biosynthetic pathway was performed. Other secondary metabolite precursor proteins from *Mentha piperita* (44), *M. arvensis* (2) and remaining proteins were modeled by using the program Modeller9v8 (Sali and Blundell, 1993).

Energy minimization via GROMOS96

After obtaining the putative 3D modeled structure for 164 proteins which can be expressed by 145 contigs, the structures were subjected for energy minimization to obtain more optimized model. Energy minimization has been done by GROMOS96 force field implemented in Swiss model software v4.0.1 (Guex and Peitsch, 1997). The GROMOS96 helps in minimization of bond stretch energy of the modeled protein. It incorporates both bonded and non bonded form of energy occupied in the protein molecule.

Validation of 3-D models or model quality evaluation

The stereo chemical quality of the modeled protein was performed by Ramachandran plot analysis using the program Procheck (Laskowski et al., 1993). Further evaluation of modeled structure was done by VERIFY3D (Eisenberg et al., 1997). All validation test were performed using SAVES server (http://nihserver.mbi.ucla.edu/SAVES/).

Authentication of the predicted model

The quality of model was also assessed by comparing predicted structure to experimentally solved structure via superimposition and atoms RMSD assessment. (Rayan, 2009). The structural superimposition of template and query structures are estimated by using TM Align (http://zhanglab.ccmb.med.umich.edu/TM-align) (Zhang et al., 2005). An optimal alignment between two proteins, as well as the TM-score, is reported for each comparison. In general, a comparison of TM-score < 0.2 indicates that there is no similarity between two structures; a TM-score > 0.5 means that the structures share the same fold. Quality of the homology model was judged by selecting residue which makes clashes in their existing position.

Table 5. RMSD value and TM align scores

S.No	Protein	Organism	Amino acid making clashes	RMSD	TM Score
1	(E)-beta-farnesene synthase	*Menthax piperita*	NO	0.86	0.79650
2	Flavonoid 4'-O-methyltransferase	*Menthax piperita*	NO	0.84	0.97751
3	Isopiperitenone reductase	*Mentha piperita*	NO	0.71	0.88314
4	Menthol dehydrogenase	*Mentha piperita*	NO	0.66	0.97745
5	Pulegone reductase	*Mentha piperita*	NO	1.70	0.89650
6	Menthofuran synthase	*Mentha arvensis*	NO	0.69	0.97751

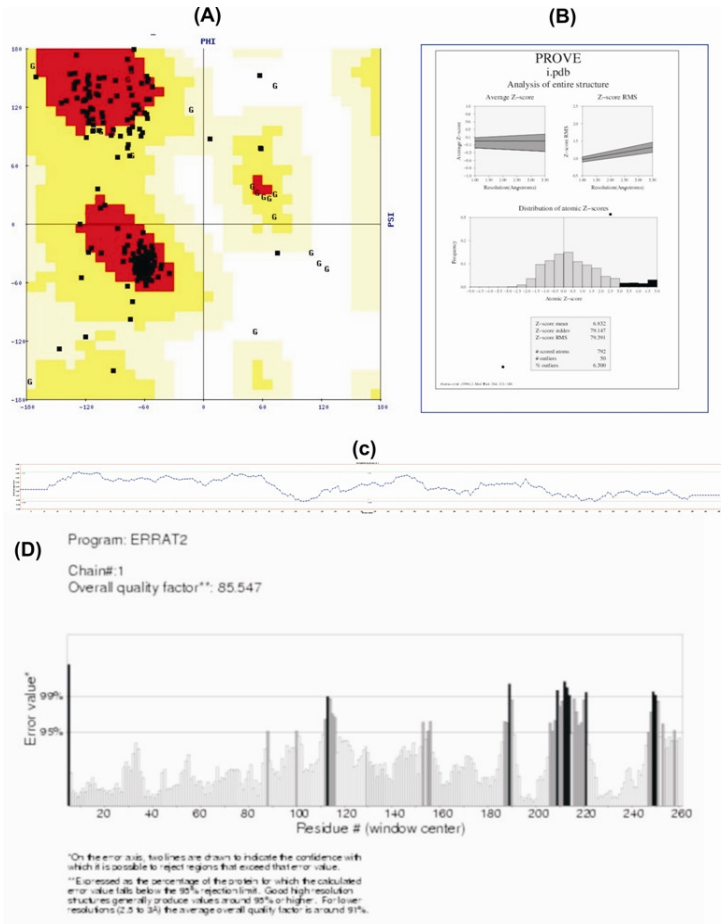

Fig 5. **(A)** Result of Ramachandran plot via procheck analysis. (**B)** Result of Prove analysis showing Z Score = 6.692 which is the statistical z-score deviation for the model from the highly resolved and refined pdb structures. (**C)** Result of Verify _3d. Showing that 69.17% of the residues had an average of 3D-1D score >0.2. (**D)** Result of Errat analysis showing an overall quality factor of 85.547.

Threading

Threading has been shown to make accurate predictions even in a "twilight zone" of <25% sequence identity, where sequence-based approaches normally fail. When we did not get sequences having more than 25% sequence similarity to 53 protein sequences in *Mentha piperita* and 10 in *M. arvensis*, these sequences were subjected to molecular threading. Biosuite v3 was used for threading. Proteins that were subjected to threading includes NDB3; NADH dehydrogenase, external rotenone-insensitive NADPH dehydrogenase for *Mentha piperita*, cytochrome P450 like_TBP for *M. arvensis*.

Calculation of physicochemical parameters and functional characterization

Various physicochemical characters, viz. molecular weight, theoretical pI, number of negatively and positively charged residues, extinction coefficients (Gill and Hippel 1989), instability index (Guruprasad et al., 1990), aliphatic index (Ikai, 1980) and grand average of hydropathicity (GRAVY) (Kyte and Doolittle, 1982) were computed for *Mentha* secondary metabolite proteins using Expasy's ProtParam proteomics server (Gasteiger et al., 2005). Functional characterization of disulfide bonds present in secondary metabolite proteins of *Mentha* species were done by Cys_REC tool from Softberry.

Table 6. Threading results of the proteins of *Mentha* sp.

Protein	Fold ID	Score	Fold	Organism
Dehydrin 13	1.1IN5A	1.6578	5. A. DNA/RNA-binding 3-helical bundle	*Thermotoga maritima*
	2.1ACC-	2.6334.5		
	3. 1I50A	3.6298.5		
	4. 1EGUA	4. 6292		
	5. 1EPUA	5.6279.5		
Cold and drought regulatory protein	1.1I50A	1.6324.5	Beta and beta-prime subunits of DNA dependent RNA- polymerase	*Saccharomyces cerevisae*
	2.1EGOA	2.6294		
	3.1EUHA	3.6267		
	4.1CROA	4.6264		
	5.1EW2A	5.6259		

Table 7. Position of disulphide bridges.

S.No	Protein name	Position of disulphide bond in SMP
1	Terpinolene synthase	43
2	Limonene synthase	49
3	Monoterpene synthase	30
4	Fenchol synthase	33
5	GGR (geranylgeranyl reductase); farnesyltranstransferase	227
6	(-)P450 limonene-3-hydroxylase	375
7	Raffionse synthase 3	48

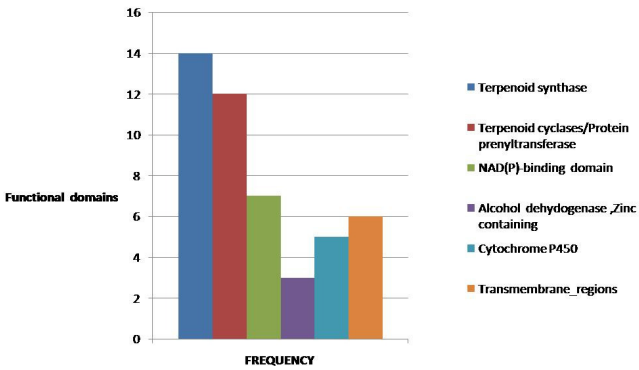

Fig 6. Motif showing high frequency of occurrence in SMPs

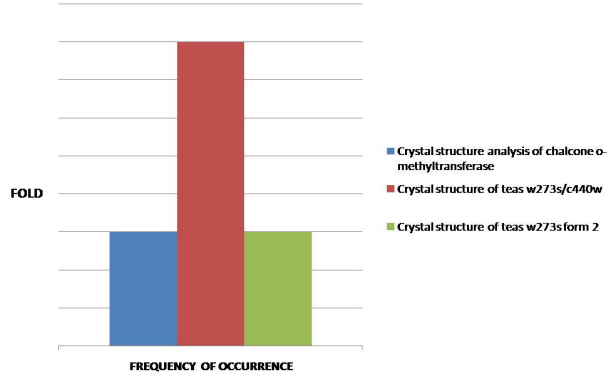

Fig 7. Folds showing high frequency of occurrence in SMP

Comparative analysis of functional domains and folds

Functional domain occurrences in secondary metabolite proteins were investigated through Interproscan run using Profunc server (http://www.ebi.ac.uk/thornton-srv/databases/ProFun) (Laskowski et al., 2005). The study is done to get insight into the short amino acid sequence pattern which plays an important functional role in secondary metabolite proteins. The Interproscan script is used to identify functional domains from the eight sources which include Prosites, Prints, and ProDom etc. To get insight in to the topology present in secondary metabolite proteins, fold detection was performed through Profunc server. The Profunc server uses the MSDfold (previously known as SSM) and DALI services to identify structural similarity.

Comparative analysis of secondary structures

Promotif (Chan et al., 1993) analysis was performed to get insight into the structural motifs. It may give insight into the relationships between proteins and their possible evolutionary origins. It also deepens our understanding of the relationship between the amino acid sequence and the tertiary structure of *Mentha* proteins. This in turn can be used to aid modeling by homology, ab-initio prediction of structure from sequence, and design of novel proteins. Super secondary structures are also good candidates for nucleation sites in protein folding. Many motifs, such as β turns and β bulges, are functionally important, as they have been found to be involved in active sites and ligand binding surfaces (Chan et al., 1993). Detailed analyses of these motifs have been carried out with the help of Promotif for 52 SMPs of *Mentha piperita* and 2 of *M. arvensis* e.g. β hairpins (Sibanda and Thornton, 1985), β-bulges (Richardson, 1981), and classification schemes have been devised to describe the conformations in which they occur.

Acknowledgements

Financial support of Department of Biotechnology (DBT) Govt. of India, New Delhi under BTISnet program is gratefully acknowledged. Sudeep Roy is thankful to Council

of Scientific and Industrial Research for Senior Research Fellowship.

Reference

Campbell ID, Downing AK (1994) Building protein structure and function from modular units. Trends Biotechnol. 12: 168-172.

Chan AWE, Hutchinson EO, Harris D, Thornton JM (1993) Identification, Classification and analysis of & bulges in proteins. Protein Sci. 2: 1574-1590.

Combet C, Jambon M, Deléage G, Geourjon C (2002) Geno3D: Automatic comparative molecular modelling of protein. Bioinformatics 18: 213-214.

Eisenberg D, Luthy R, Bowie JU (1997) VERIFY:3D Assessment of protein models with three-dimensional profiles. Method Enzymol. 277: 396-404.

Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins M R, Appel RD, Bairoch A (2005) Protein Identification and Analysis Tools on the ExPASy Server; in John M. Walker (ed): The Proteomics Protocols Handbook, (Humana Press) pp. 571-607.

Gill SC, Hippel PH Von (1989) Calculation of protein extinction coefficients from amino acid sequence data. Anal. Biochem. 182: 319-326.

Guex N, Peitsch MC (1997) SWISS-MODEL and the Swiss –PDBViewer: An environment for comparative protein modeling. Elerctrophoeresis 18: 2714-2723.

Gupta Sanchita, Shukla Rishi, Roy Sudeep, Sen Naresh, Sharma Ashok (2010) *In silico* SSRs and FDM analysis through EST sequences in *Ocimum basilicum*. Plant Omics J 3: 121-128.

Guruprasad K, Reddy BV, Pandit MW (1990) Correlation between stability of a protein and its dipeptide composition: A novel approach for predicting in vivo stability of a protein and its primary sequence. Protein Eng. 4: 155-161.

Husain A (1994) Essential oil plants and their constitution, Central Institute of Medicinal and Aromatic Plants, (CIMAP), Lucknow, India, pp 167-181.

Husain A, Virmani OP, Popli SP, Mishra LN, Gupta MM, Srivastva GN, Abraham Z, Singh AK (1992) Dictionary of Indian Medicinal Plants, Central Institute of Medicinal and Aromatic Plants, (CIMAP), Lucknow, India, pp 294-296.

Husain A, Virmani OP, Sharma A, Kumar A, Misra LN (1988) Major essential oil bearing plants of India, Central Institute of Medicinal and Aromatic Plants, (CIMAP), Lucknow, India, pp 167-181.

Ikai A (1980) Thermostability and aliphatic index of globular proteins. J Biochem. 88: 1895-1898.

Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol. 157: 105-132.

Laskowski RA, Macarthur MW, Moss DS, Thorntom JM (1993) PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Crystallogr. 26: 283-291.

Laskowski RA, Watson James D, Thornton Janet M (2005) ProFunc: a server for predicting protein function from 3D structure. Nucleic Acids Res. 33 (Web Server issue).

Lee KW, Choi GJ, Kim KY, Ji HC, Zaman R, Lee SH (2011) Identification of drought induced differentially expressed genes in barley leaves using the annealing control-primer-based GeneFishing technique. Aust J Crop Sci. 5(11): 1364-1369.

Masoudi-Nejad A, Tonomura K, Kawashima S, Moriya Y, Suzuki M, Itoh M, Kanehisa M, Endo T, Goto S (2006) EGassembler: online bioinformatics service for large-scale processing, clustering and assembling ESTs and genomic DNA fragments. Nucleic Acids Res. 34: W459-462.

Pavli OI, Ghikas DV, Katsiotis A, Skaracis GN (2011) Differential expression of heat shock protein genes in sorghum (Sorghum bicolor L.) genotypes under heat stress. Aust J Crop Sci. 5(5):511-515.

Ramachandran GN, Ramakrishnan C, Sasisekharan V (1963) Stereochemistry of polypeptide chain configuration. Journal of Molecular Biology 7: 95-99.

Rayan A (2009) New tips for structure prediction by comparative modeling. Bioinformation 3: 263-267.

Richardson JS (1981) The anatomy and taxonomy of protein structure. Adv. Protein Chem. 34: 167-339.

Rohman MM, Suzuki T, Fujita M (2009) Identification of a glutathione S-transferase inhibitor in onion bulb (*Allium cepa* L.). Aust J Crop Sci. 3: 28-36.

Sali A, Blundell TL (1993) Protein modelling by satisfaction of spatial restraints. J. Molecular Biology 234: 779-815.

Sibanda BL, Thornton JM (1985) Hairpin families in globular proteins. Nature 316: 170-175.

Ton Christopher, Hwang David M, Dempsey Adam A, Tang Hong-Chang, Yoon Jennifer, Lim Mindy, Mably John D, Fishman Mark C, Liew Choong-Chin (2000) Identification, Characterization, and Mapping of Expressed Sequence Tags from an Embryonic Zebrafish Heart cDNA Library. Genome Res. 10: 1915-1927.

Tramontano A (1998) Homology modeling with low sequence identity. Methods in Enzymology 14: 293-300.

Zhang Y, Skolnick J (2005) TM-align: A protein structure alignment algorithm based on TM-score. Nucleic Acids Res. 33: 2302-2309.