Evolutionary Bioinformatics



ORIGINAL RESEARCH

# Evolutionary Implications and Physicochemical Analyses of Selected Proteins of Type III Polyketide Synthase Family

V. Mallika[1], K.C. Sivakumar[2] and E.V. Soniya[1]

[1]Plant Molecular Biology, [2]Bioinformatics facility, Rajiv Gandhi Centre for Biotechnology, Thycaud P O, Poojappura, Thiruvananthapuram-695 014, Kerala, India. Corresponding author email: evsoniya@rgcb.res.in

**Abstract:** Type III polyketide synthases have a substantial role in the biosynthesis of various polyketides in plants and microorganisms. Comparative proteomic analysis of type III polyketide synthases showed evolutionarily and structurally related positions in a compilation of amino acid sequences from different families.Bacterial and fungal type III polyketide synthase proteins showed <50% similarity but in higher plants, it exhibited >80% among chalcone synthases and >70% in the case of non-chalcone synthases. In a consensus phylogenetic tree based on 1000 replicates; bacterial, fungal and plant proteins were clustered in separate groups. Proteins from bryophytes and pteridophytes grouped immediately near to the fungal cluster, demonstrated how evolutionary lineage has occurred among type III polyketide synthase proteins. Upon physicochemical analysis, it was observed that the proteins localized in the cytoplasm and were hydrophobic in nature. Molecular structural analysis revealed comparatively stable structure comprising of alpha helices and random coils as major structural components. It was found that there was a decline in the structural stability with active site mutation as prophesied by the *in silico* mutation studies.

**Keywords:** type III polyketide synthase, CHS, PKS, non-CHS, molecular phylogeny, chalcone synthase

*Evolutionary Bioinformatics* 2011:7 41–53

doi: 10.4137/EBO.S6854

This article is available from http://www.la-press.com.

## Introduction

Polyketide synthases (PKSs) are a family of enzyme complexes that produce the large class of secondary metabolites, known as polyketides. They exhibit many pharmacologically important properties including antibiotic, antifungal, antitumor and immunosuppressive activities. Three types of PKSs are known to date; type I, II and III which are quite different from each other in their structures and functions and can be easily distinguished by their physical composition.[1]

Type III PKSs are responsible for the synthesis of polyketides such as chalcones and stilbenes. Among the proteins mentioned in Table 1, the well studied and widely distributed family of chalcone synthases (EC 2.3.1.74) play a significant role in the biosynthesis of various polyketides.[2,3] They exist as multigene families in many of the plants[4,5] and are found to be involved flower pigmentation, defense response and plant fertility. Type III PKS proteins identified from *Zingiber officinale* (ZoPKS) show significant variations from the previously identified type III PKSs.[6] Extensive gene duplication followed by the functional divergence is believed to have played an important role in generating the biochemical diversity of PKS superfamily.[3] This diversity provides a rich source of pharmaceutically valuable antibiotics, anti-cancer drugs, biologically active pigments and allelochemicals. For example, resveratrol, a stilbene synthase derivative from grapes showed cancer chemopreventive activity in murine models.[7,8] Stilbene from red wine is also reported to have protective role against heart diseases.[2]

Type III PKSs are structurally simple, small homodimeric proteins of approximately 40–45 kDa, with a catalytic triad of $Cys^{164}$-$His^{303}$-$Asn^{336}$ at the active site. Structural insight into the reaction mechanism elucidated for *M. sativa* CHS indicates the presence of three inter connected cavities; the CoA binding tunnel, a coumaroyl-binding pocket and a cyclization pocket at the active site.[9] Type III PKS monomer utilizes the triad within an internal active site cavity, which is connected to the surrounding aqueous phase by a narrow CoA- binding tunnel.[10] Enzymes in this family have broad substrate specificities and they generate a wide diversity of polyketide derivatives by changing the starter substrates,[11] the number of acetyl additions and cyclization reactions. The reaction (Fig. 1) involves the formation of chalcone or stilbene by the stepwise decarboxylative condensation of coumaroyl CoA with three malonyl CoA followed by the intramolecular cyclization known as claisen type condensation[12,13] of the tetraketide product. Fatty acid biosynthesis also utilizes the same mechanism of carbon-carbon bond formation and both the synthesis pathways show many common features.[12]

The aim of the present study is to investigate the evolutionary relationship of Type III PKSs among different plant, bacterial and fungal species. The study also aims to find out the sequence conservation, structural features and physicochemical properties across the different type III PKS sequences.

**Table I.** Functionally divergent members of type III PKS superfamily.

| Major type III polyketide synthases | References |
|---|---|
| Chalcone synthase | Austin and Noel;[3] McKhann and Hirsch[43] |
| 2-pyrone synthase | Helariutta et al;[44] Koskela et al[45] |
| Phloroisovalerophenone synthase | Paniego et al[46] |
| Trihydroxybenzophenone synthase | Liu et al[47] |
| Acridone synthase | Wanibuchi et al[48] |
| Biphenyl synthase | Liu et al[49] |
| Stilbene synthase | Schroeder and Schroeder;[35] Schoppner and Kindl[50] |
| Pentaketide chromone synthase | Abe et al[51] |
| Octaketide synthase | Abe et al[52] |
| Benzalacetone synthase | Abe et al[53] |
| Aloesone synthase | Abe et al[54] |
| Bibenzyl synthase | Preisig-Muller et al[55] |
| Coumaroyl triacetic acid synthase | Akiyama et al[56] |
| Stilbenecarboxylate synthase | Eckermann et al;[57] Schroder[13] |

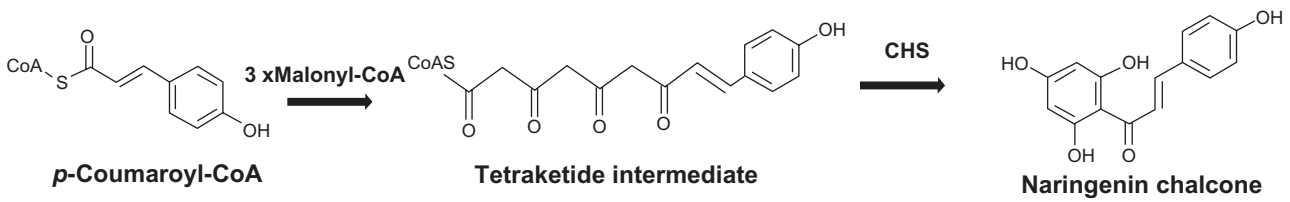


**Figure 1.** CHS reaction mechanism. Chalcone synthase condense three acetyl anions derived from the malonyl-CoA decarboxylation with the coumaroyl-CoA to form the tetraketide intermediate, which further aromatize and cyclise to form the chalcone.

## Materials and Methods

### Sequence data retrieval

Type III PKS amino acid sequences were retrieved in FASTA format from GenBank (http://www.ncbi.nlm.nih.gov/Genbank/) and Swiss-Port (http://www.expasy.ch/sprot/). Subsequently, a search was done using the BLASTp program[14] at NCBI to identify and retrieve potentially related sequences. A homology threshold (E-value) of 0.01 was set to get significant hits to be included in the alignment comprising subfamilies. *M. sativa* CHS (Swiss-Prot Acc. No- P30074) was used as query against the non-redundant protein sequence database. From the retrieved PKS sequences, about 56 Type III PKSs were selected to represent different families of plants, fungi and bacteria. Partial sequence entries were excluded. The species names and database accession numbers of the selected type III PKSs are given in Supplementary material 1.

### Alignment process

Type III PKS amino acid sequences, representing 56 families were aligned using the multiple sequence alignment program ClustalW2.[15,16] Default gap opening and extension penalty were used. A careful manual examination was done to avoid misalignment. Gaps in the alignment were treated as missing data. Proteins were named using an abbreviation followed by the family name in the alignment.

### Phylogenetic analysis

Phylogenetic analysis was carried out by using MEGA4.[17] Neighbor-joining, Maximum Parsimony and Minimum Evolution methods were used to infer the phylogeny across the data.[18–20] Bootstrap analysis was performed to check the reliability of the evolutionary tree.[21] Branches corresponding to partitions reproduced in less than 50% bootstrap replicates were collapsed.

### Cavity volume analysis

Substrate binding pockets and respective cavity volumes were predicted using CASTp, a fast and precise algorithm.[22] based on computational geometry methods, including alpha shape and discrete flow theory. Default probe radius was used (1.4 Angstroms). In the present work, we have compared cavity volume of eight different type III PKS proteins viz. Chalcone synthase (*M. sativa*), Stilbenecarboxylate synthase 2 (*M. polymorpha*), Type III PKS (*M. tuberculosis*), Tetrahydroxynaphthalene synthase (*S. coelicolor*), Stilbene synthase (*A. hypogaea*), 2-pyrone synthase (*G. hybrida*), Type III PKS (*N. crassa*) and Pentaketide chromone synthase (*A. arborescens*). The atomic coordinates of these type III PKS proteins were obtained from the RCSB PDB database.

### Physicochemical and structural analysis

Amino acid composition of selected type III PKS proteins were analyzed by ProtParam, an online tool available on ExPASy proteomics server.[23] Physicochemical properties and subcellular localization were predicted by protein calculator v3.3 (http://www.scripps.edu/~cdputnam/protcalc.html) and SubLoc v1.0 server[24] respectively. For subcellular localization prediction of bacterial type III PKSs, TargetP 1.1,[25] and PSORTb v.3.0[26] were used. The disordered regions in the above mentioned eight different type III PKS proteins were studied by PreDisorder[27] (http://casp.rnet.missouri.edu/predisorder.html). Single site mutation studies were performed on MUpro server[28] which is available at www.ics.uci.edu/~baldig/mutation.html. Cellular function and hydrophobicity of the proteins were analyzed by ProtFun 2.2[29,30] and pepstats[31] servers respectively. Secondary structures were predicted by GARNIER[32] on EMBOSS server.

## Results and Discussion

### Sequence alignment

Multiple alignments (Supplementary data 2) showed high sequence similarity across various type III PKS proteins. Around 80% similarity was observed among chalcone synthase superfamily of plant type III PKSs. This suggested a high conservation among CHSs from plant species. Fungal and bacterial type III PKSs shared less than 50% similarity, while the non-CHS proteins presented more than 70% similarity. Highly conserved amino acid residues were depicted in Figure 2.

### Residues in the catalytic site

Earlier studies on CHS from *M. sativa* indicated that four amino acid residues ($Cys^{164}$, $Phe^{215}$, $His^{303}$ and $Asn^{336}$) that belong to the catalytic region show conservation in all the members of CHS family.[9] Among these residues 'Cys-His-Asn' is known as catalytic triad in CHSs.[3]

In our studies, all type III PKSs including CHSs and non-CHSs showed absolute conservation in the case of catalytic triad. Proteins from bacteria and fungi also hold the conserved catalytic residues (Cys, His and Asn). '$Phe^{215}$' residue replaced with 'Leu' in BAS which is a non-CHS protein from *R. palmatum*

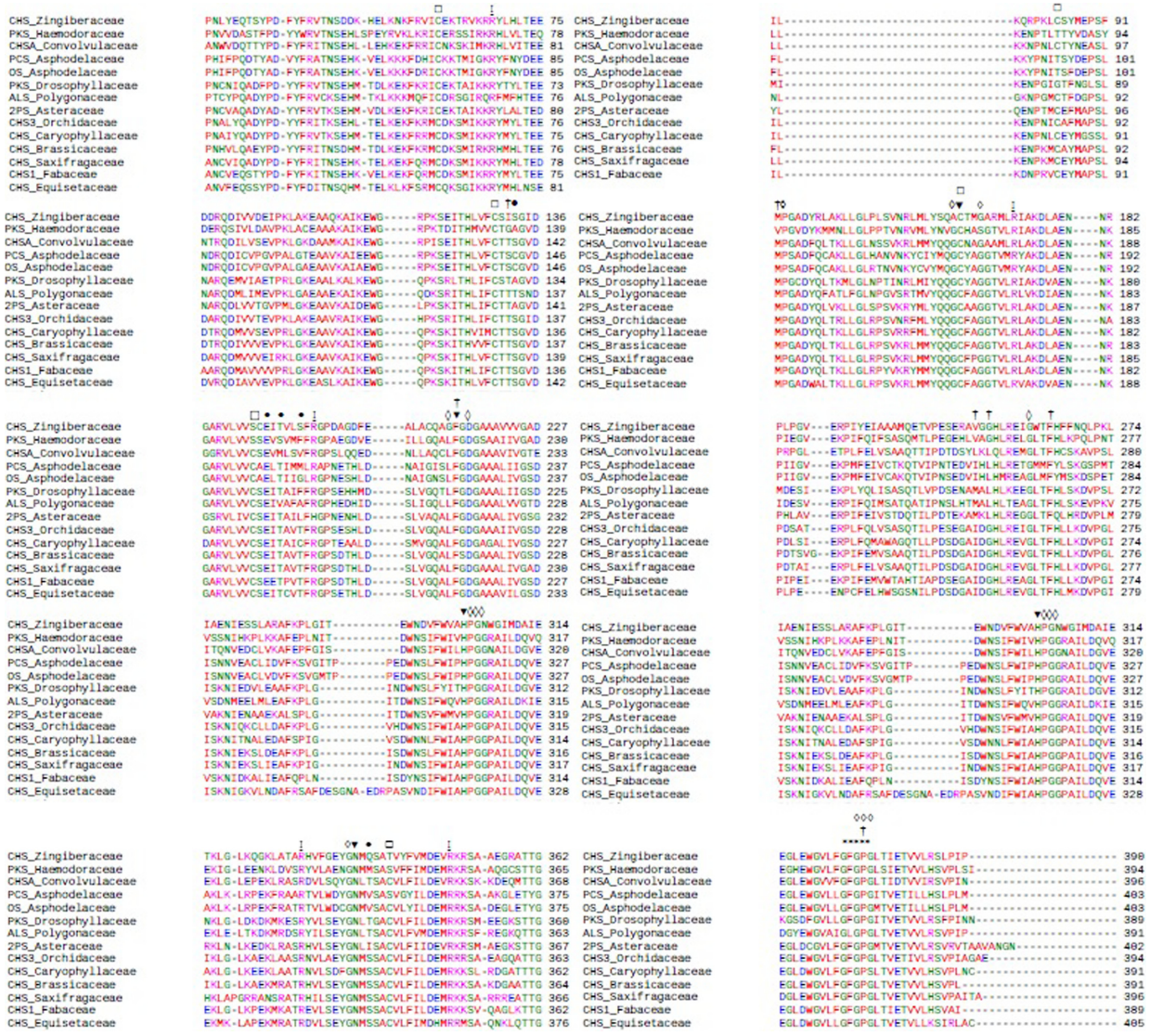

**Figure 2.** Alignment of type III PKS proteins. The amino acid residues that forms the catalytic triad (▼), substrate binding pocket (•), Cyclisation pocket (†) are shown in the alignment. The symbol "↕" represents the non-active arginine residue that is conserved. The asterisks (*) symbol represents the 'GFGPG loop' which serve as a part of active site scaffold. The residues, shaping the geometry of the active site are indicated by the symbol "◊". Symbol "□" represents conserved cysteine residues.

(Polygonaceae family). This indicates that this substitution is favoured. The conserved asparagine residue (Asn$^{336}$) was found to be substituted with 'Ala' in *P. americana CHS* (Lauraceae- Q9ZU06), eventhough it showed 75.3% similarity to *M. sativa* CHS.

Previous studies have reported that the mentioned active sites residues (Cys, His and Asn) are located at the lower side of hydrophobic pockets or crevices and that they vary according to the substrate.[12] It is found that the residue 'Cys$^{164}$' is located at the amino terminal end of the a-helix and 'Phe$^{215}$' at the bottom of the binding pocket. It is reported that 'Phe$^{215}$' is involved in the decarboxylation of malonyl-CoA and may also help in the orientation of substrates and reaction intermediates at the active site.[33]

## Residues in the substrate binding pocket

The amino acid residues (Ser$^{133}$, Glu$^{192}$, Thr$^{194}$, Thr$^{197}$ and Ser$^{338}$) located at the substrate binding pocket presented high degree of conservation with a number of exceptions (Supplementary data 3, Table 1). Sequence alignment clearly showed that all fungal and bacterial type III PKSs substituted 'Ser$^{133}$' with 'Thr'. Here both 'Ser' and 'Thr' belong to the same neutral polar amino acid group. 'Ser$^{133}$' also showed substitution in most of the non-CHS members. Absolute conservation of residue 'Glu$^{192}$' was found across the alignment while Thr$^{194}$, Thr$^{197}$ and Ser$^{338}$ showed variations. In most of the fungal and bacterial type III PKSs, 'Thr$^{194}$' was substituted with 'Cys'. 'Ser$^{338}$' showed lesser conservation among the non-CHSs of angiosperms and bacterial type III PKSs. In case of *M. sativa* (PDB:1BI5), structural analysis showed that the residue 'Ser$^{133}$' was located at the end of the fourth strand and 'Glu$^{192}$' at the end of 6th strand. Threonine residues of the substrate binding pocket were located at the 10th alpha helix, where 'Thr$^{194}$' was at the beginning and 'Thr$^{197}$' towards the end. 'Ser$^{338}$' was found to be positioned at the starting of the 15th alpha helix. Sequence conservation pattern also indicated that a single amino acid change can alter a CHS protein to a non-CHS protein. So substrate preference can alter and this may help in the drug development studies.

## Residues in the cyclization pocket

Residues that formed the cyclization pocket presented high degree of conservation (Thr$^{132}$, Met$^{137}$, Phe$^{215}$, Ile$^{254}$, Gly$^{256}$, Phe$^{265}$, and Pro$^{375}$) with a number of exceptions (Supplementary data 3, Table 2). Thr$^{132}$ was substituted with 'Cys' in most of the fungal and bacterial type III PKSs. CHSs from all plants exhibited 'Thr$^{132}$' conservation, with an exception of *Z. officinale,* where isoleucine (I) was found instead of Threonine. Type III PKS from the pteridophyte, *P. nudum* (wisk fern) displayed variations in the case of cyclization pocket residues. ie, VPS of *P. nudum* showed three amino acid substitutions (Ser$^{139}$, Val$^{261}$ and Leu$^{272}$ instead of Thr$^{132}$, Ile$^{254}$ and Phe$^{265}$ respectively) while STS displayed a single substitution ('Met$^{137}$' with 'Leu'). The residues corresponding to 'Pro$^{375}$' exhibited absolute conservation across the various type III PKSs. In the case of angiospermic CHSs, the residues Met$^{137}$, Ile$^{254}$ and Phe$^{265}$ were highly conserved. But in the case of non-CHSs, the fungal and bacterial type III PKSs exhibited variations within these residues. It has been reported that among these residues Phe$^{215}$ and Phe$^{265}$ were situated at the active site entrance (Jez et al, 2002). So in this regard we can infer that these residues might possess major catalytic roles.

## Cysteine conservation

Investigation on RS3 from *A. hypogaea* and CHS from *S. alba* showed cysteine conservation at six positions (Cys$^{65}$- Cys$^{89}$- Cys$^{135}$- Cys$^{169}$- Cys$^{195}$- Cys$^{347}$). These residues were also found to be conserved in all other chalcone synthases known so far.[34,35] In our studies, 'Cys$^{65}$' was accommodated with other residues in bacterial and fungal PKSs. Fungal type III PKSs substituted 'Ala' instead of 'Cys$^{65}$', while bacterial PKSs presented 'Ala', 'Phe', 'Ile' or 'Tyr' in the referred position. 'Cys$^{89}$' is substituted with 'Thr', 'Gly', 'Val', 'Ala', 'Arg', 'Ser' or 'Glu' residues in some type III PKSs. In most of the bacterial and fungal PKSs, 'V' was found in place of Cys$^{135}$. In THBS of *H. androsaemum* (Clusiaceae), CHS of *P. patens* (Funariaceae) and *M. Polymorpha* STCS2 (Marchantiaceae), 'Cys$^{135}$' was substituted with 'Ala'. Cys$^{195}$ was conserved in most cases leaving out Zingiberaceae, Haemodoraceae and Funariaceae. In fungal PKSs, non-polar neutral 'Ala' was found in place of 'Cys$^{195}$' with the exception of Pleosporaceae where polar residue 'Thr' was found. Fungal, bacterial, bryophytic and pteridophytic type III PKSs showed little conservation for the

'Cys[347]' residue. In addition to that, some angiospermic CHSs and non-CHSs also showed variations in the case of cysteine conservation (ie, *P. americana* (Lauraceae) and *Z. officinale* (Zingiberaceae).

### GFGPG loop

A highly conserved 'GFGPG' loop was detectable in all plant CHSs and non-CHSs. In this, the amino acid 'P' (proline) constituted the central part of cyclization scaffold.[36] Phenylalanine (F) in the 'GFGPG' loop was found to be replaced with leucine (L) in 'THBS' of *H. androsaemum* (Clusiaceae) and 'STS' of *P. nidum* (Psilotaceae). *In silico* 'single site amino acid mutation' predictions indicated that the stability of the protein structure increases when 'leucine' replaces 'phenylalanine' in the loop. As predicted by Mupro, there is a decline in the structural stability of THBS when leucine was mutated to phenylalanine. In fungal proteins, the 'GPG' residues of the above mentioned loop exhibited absolute conservation, while the very first 'Glycine' residue was found to be substituted with alanine (A) or serine (S). In *P*. nodorum (Phaeosphaeriaceae), phenylalanine (2nd residue of the loop) was assigned with 'valine'. In bacteria, this loop was conserved in most cases with very few exceptions. 'FGPG' residues of the loop were highly conserved but the first 'glycine' residue of the loop was substituted with 'serine' and 'alanine' in *B. subtilis* (Bacillaceae) and *N. farcinica* (Nocardiaceae) respectively.

### Other residues

Amino acid residue 'Gly[238]' is found only in CHS of Brassicaceae and rarely found in other plant PKSs.[34] Our analysis showed that THBS of Clusiaceae also possessed this 'glycine' residue. In addition to this, it was identified that the amino acid residues Pro[138], Gly[163], Gly[167], Leu[214], Asp[217], Gly[262], Pro[304], Gly[305], Gly[306], Gly[335], Gly[374], Pro[375] and Gly[376] presented high degree of sequence conservation, which may play a crucial role in shaping the correct geometry of the active site. The residues ('Gly[163]' and 'Phe[165]') surrounding the conserved active site residue 'Cys[164]' also appeared to have high degree of conservation. But in case of CHSs from *Z. officinale*, 'Ala' substituted the place of 'Gly', but it did not contribute much variation to the geometry of the active site, as 'Ala' belongs to the same category of non-polar amino acids. In most of the non-CHSs, the position of 'Phe[165]' was occupied by aromatic amino acids like 'Tyr' or 'His'. In *Z. officinale*, 'Phe[165]' was replaced with polar amino acid 'Thr'. PKS from Heamodoraceae presented basic polar amino acid 'His' at the position corresponding to 'Phe[165]'. '2PS' from Asteraceae and type III PKSs from fungi and bacteria also substituted 'Phe[165]' with 'Ala'. We can conclude that these changes might have occurred during gene duplication events in the course of evolution.

## Phylogenetic distribution of type III PKS proteins

To determine the evolutionary history of type III PKSs, it is necessary to examine how these proteins are distributed across the species. Comparative amino acid sequence analysis was performed to investigate the phylogenetic relationship among plants (monocots, dicots, bryophytes, pteridophytes and gymnosperm), fungi and bacteria. Separate clustering of the plant CHSs and non-CHSs has already been reported[37] and they suggest a repeated gene birth, death and reinvention of non-CHS functions throughout the evolution of angiosperms.

In our investigation, a total of 56 Type III PKSs were selected to represent various families in an attempt to study the phylogenetic relationship that exists between them. The evolutionary history was inferred using the Neighbor-Joining method.[18] The bootstrap consensus tree developed from 1000 replicates (Fig. 3) represents the evolutionary history of the taxa analyzed.[25] Branches corresponding to partitions reproduced in less than 50% bootstrap replicates were collapsed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test is shown next to the branches.[21] The evolutionary distances were computed using the Poisson correction method[38] and are in the units of the number of amino acid substitutions per site. All positions containing gaps and missing data were eliminated from the dataset (Complete deletion option). There were a total of 276 positions in the final dataset. The number on each branch indicated the bootstrap probability (%) by resamplings. The length of each branch is proportional to the estimated evolutionary distance.

Phylogenetic tree based on Maximum Parsimony (MP) was obtained using the Close-Neighbor-Interchange algorithm with search level 3[21,39] in which the initial trees obtained were with the

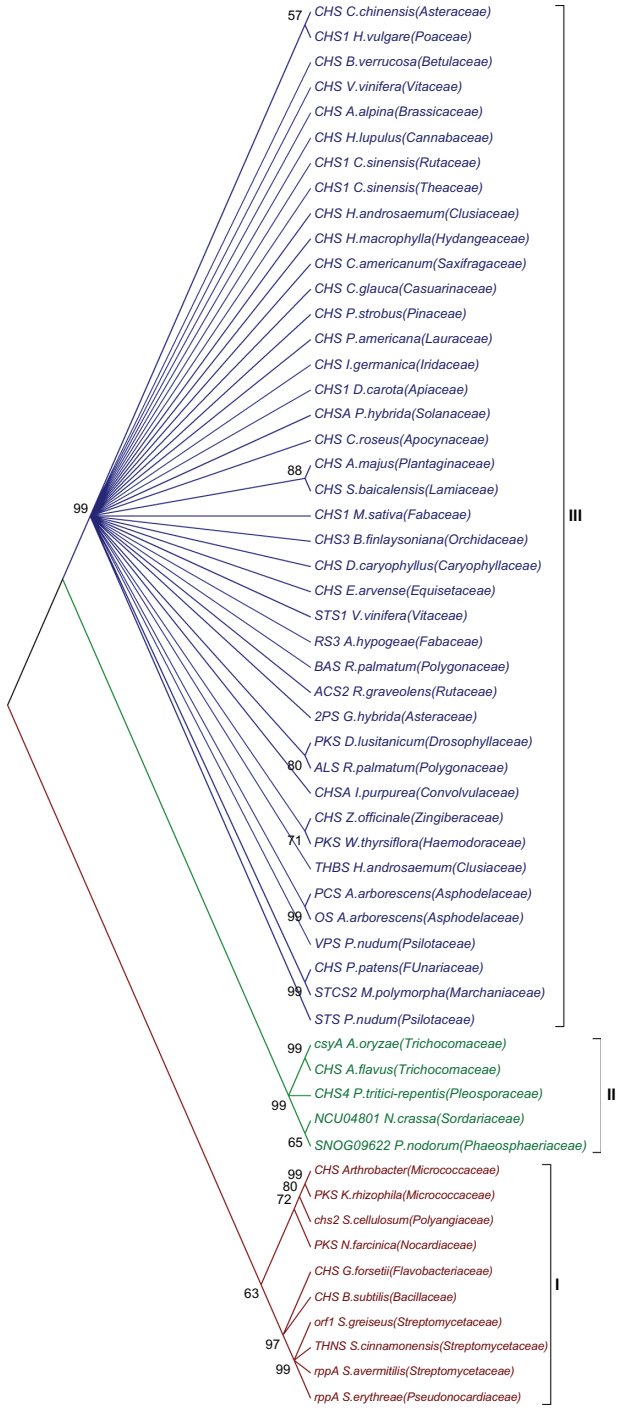


**Figure 3.** Neighbor Joining (NJ) tree showing the separate clusterring of plant, fungal and bacterial proteins. Numbering in the branch indicates bootstrap probability obtained (values lesser than 50% were not shown). 'I' and 'II' groups in the figure represent bacterial and fungal type III PKSs respectively and 'III' group indicates plant PKSs.

random addition of sequences (10 replicates). There were a total of 276 positions in the final dataset, out of which 253 were parsimony informative. The MP based phylogenetic tree for PKS superfamily diverged into different clusters, while the type III PKS from Nocardiaceae formed an out-group (Supplementary data 4).

The phylogenetic tree of the type III PKS superfamily diverged into three distinct groups (plant, fungal and bacterial proteins). Clusterring of plant and fungal proteins occurred with 99% bootstrap support (Fig. 3) while bacterial proteins with 63%. Non-CHS proteins showed sequential clustering followed by bryophytic and pteridophytic proteins. From this we can infer that plant type III PKS proteins might evolved successively from simple bryophytes to angiosperms (Fig. 4). When considering the plant type III PKSs alone, they showed a clear distinction as CHSs and non-CHSs with few exceptions (Fig. 5). CHSs from *Z. officinale* (Zingiberaceae), *I. purpurea* (Convolvulaceae) and *P. americana* (Lauraceae) were grouped along with non-CHSs (75%, 78% and 75.3% similarity to the *M. Sativa* CHS respectively). CHSs from *E. arvense* (Equisetaceae) and *P. patens* (Funariaceae) were also found to be clusterred with the non-CHSs.

The bacterial PKSs clustering occurred in two distinct clades with a bootstrap support of 63% and fungal PKSs were grouped (99% bootstrap support) very next to the bacterial clade. Type III PKS proteins from pteridophytes (VPS and STS from *P. nudum*) and bryophytes (CHS from *P. patens* and STCS2 from *M. polymorpha*) clustered well within the tree and are nested by the non-CHSs from monocot flowering plants (*A. arborescens*). The CHS from the gymnosperm *P. strobus* (Pinaceae) grouped along with angiosperms (83% identity to the typical *M. sativa* CHS). Putative type III PKS (Swiss-Prot ID: Q5YPW0) of *N. farcinica* (Nocardiaceae) showed 22.4% identity and 41.6% similarity to the typical *M. sativa* CHS (Swiss-Prot ID: P30073). Further, the clustering of bryophytic CHSs in between the fungal and plant proteins suggested the evolution of CHS proteins from lower to higher plants and hence it can be regarded as the ancestor of the plant type III PKS superfamily.

## Cavity volume and structural analysis on type III PKSs

Cavity volume plays an important role in the strength of the substrate binding interaction and hence any change in it alters the product formation profile[40] of type III PKSs. The potential substrate binding pockets and their corresponding cavity volumes

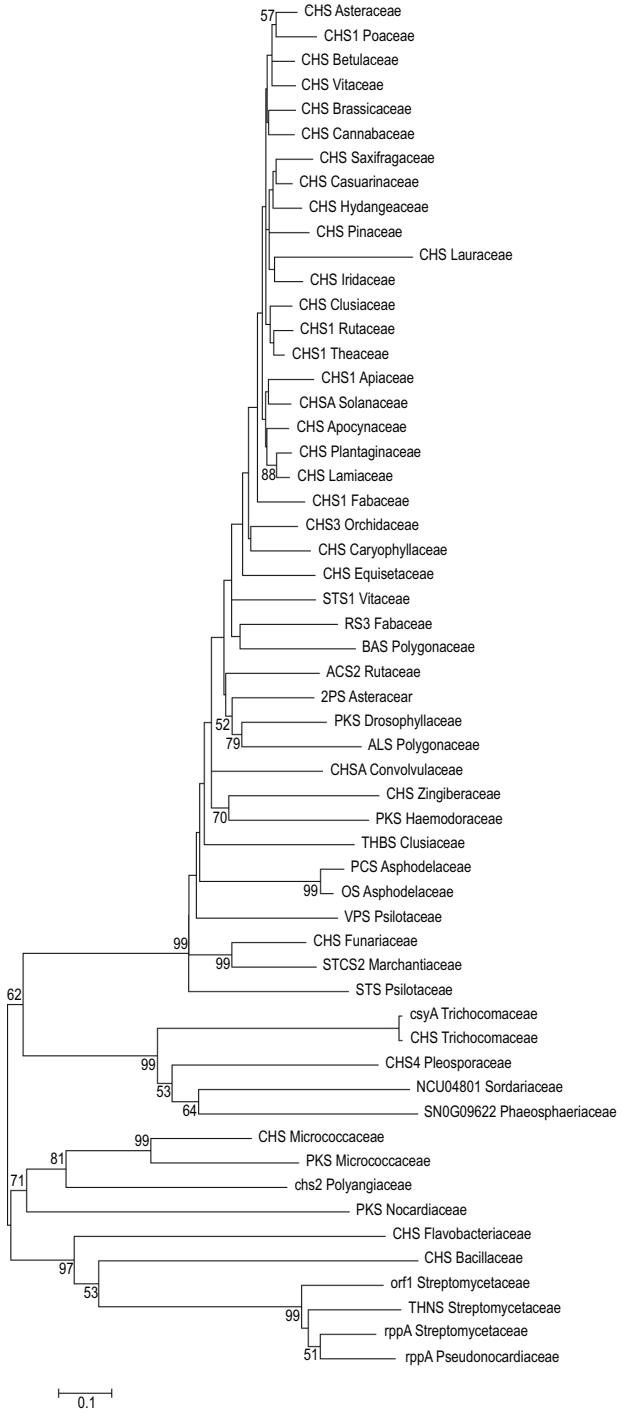


**Figure 4.** A Consensus tree derived from 1000 bootstrap replicates by Neighbor Joining (NJ) method. Numbering in the branch indicates bootstrap probability obtained.

can be predicted from its three dimensional structure using the program CASTp. The program analytically measures the area and the volume of each pocket and cavity, both in solvent accessible surface (SA, Richards' surface) and molecular surface (MS, Connolly's surface). Studies on protein structure, stability, design, and substrate binding interaction rely on the accurate prediction of cavities in it.

CASTp analyses of eight different type III PKS structures (Figs. 6 and 7) were carried out after removing the water molecules. The number of cavities per protein ranged from 40 to 65. The surface area and the volume of all the pockets in the proteins were measured and the best cavity was selected (as shown in Fig. 6). Usually the large pocket or cavity is the active site of the protein. The pockets or clefts, which are important for molecular recognition and protein function[41] possess a well developed mouth opening that provides a direct connection to the exterior of the protein. The mouth leads to a cavity which is tunnel shaped. In our analysis, flexible active site pocket volume varied over a range of 737Å-1683Å and tunnel length in between 310Å-685Å. The cavity volume of 2-pyrone synthase at the active site was found to be significantly smaller than that of chalcone synthase. The cavity volume (Fig. 7) of the eight proteins in the increasing order is as follows; *N. crassa* Type III PKS < *A. arborescens* PCS < *S. coelicolo*r THNS < *A. hypogaea* STS < *G. hybrida* 2PS < M. sativa CHS < *M. polymorpha* STCS2 < *M. tuberculosis* Type III PKS. The details of the values obtained by CASTp analysis and the changes in residues of the substrate binding pockets are given in tables (Supplementary data 3, Tables 3 and 4).

To find out the stability of the protein structure, we analyzed the disordered regions of the eight proteins using PreDisorder. The probability of disorder in the protein sequence is depicted in Figure. 8. From the evaluation of the disordered regions, it was found that the type III PKS proteins from plants adopted comparatively stable structure. In case of *N. crassa,* an increase in disorder was observed towards C- terminal end. According to the predictions from GARNIER, *M. sativa* CHS has a structure composed of coils, turns, alpha helices and beta strands in a proportion of 18%, 14.2%, 46.4% and 25.7% respectively (Supplementary data 3, Table 6). Coils and turns together constitutes about 32%. It was also found that the alpha helix formed the major structural component in type III PKS proteins.

Single site amino acid mutation predictions on the active site residues (Cys-His-Asn) performed by 'Mupro'. The results showed that both 'His' and 'Asn' mutations decreased the protein structure stability (Supplementary data 3, Table 5). A general decline in the

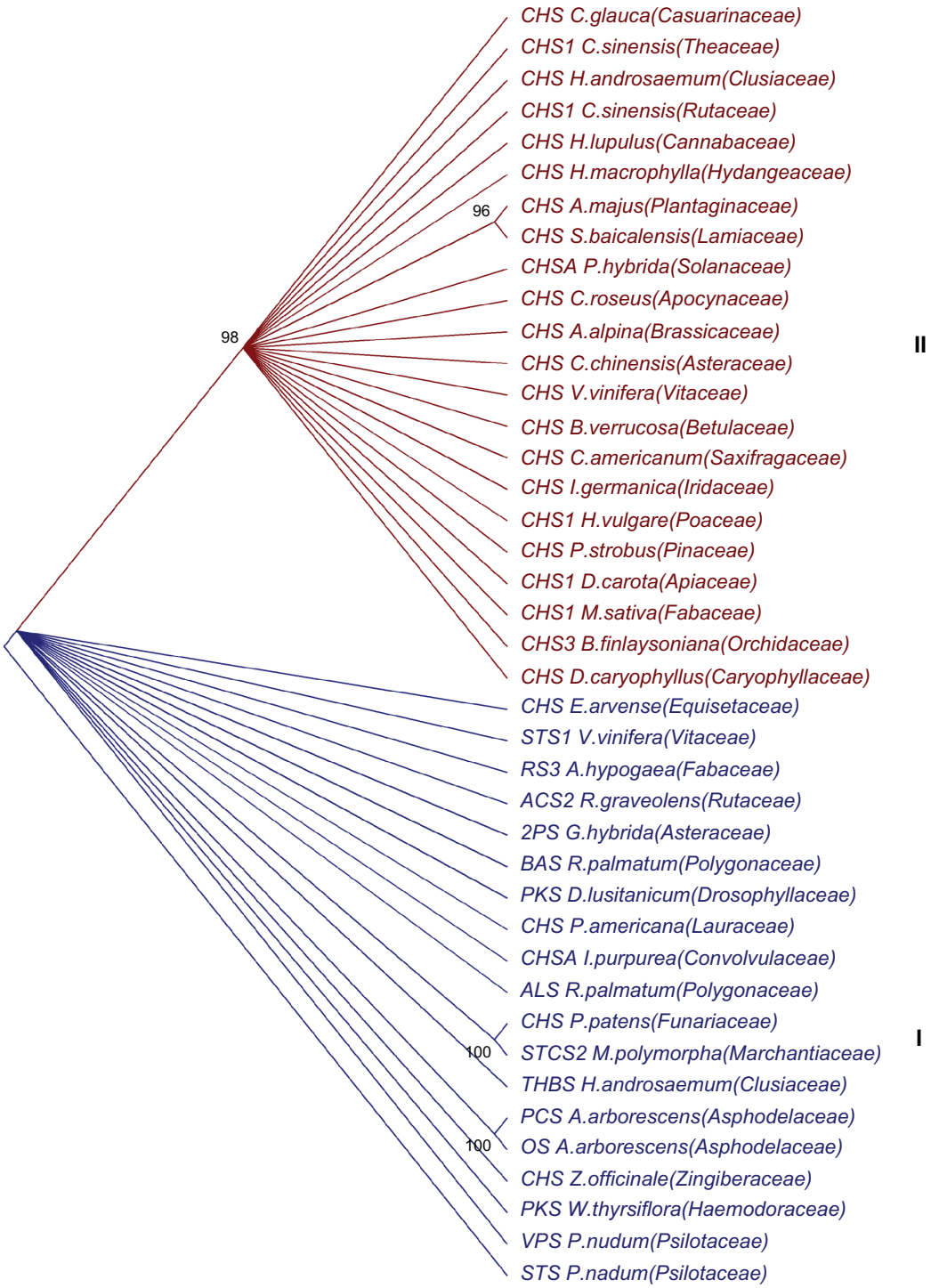


**Figure 5.** Phylogenetic tree derived from plant type III PKS alone, showing separate clustering of CHSs and non-CHSs. Ist group indicates non-CHS proteins and the IInd group indicates CHS.

stability of protein was also noticed upon mutation of the 'Cys' residue. Eventhough mutations with Leu, Ala, Ile, Met, Arg, Glu and Asp were resulted in enhanced stability, they might alter the substrate specificity and hence not preferred. This indicates that these residues possess major roles in the catalytic functions and specificities of type III PKSs. These predictions will help the researchers to carry out mutation studies in vitro.

## Physicochemical characteristics

Physicochemical analysis suggested that type III PKSs are hydrophobic in nature and are localized in

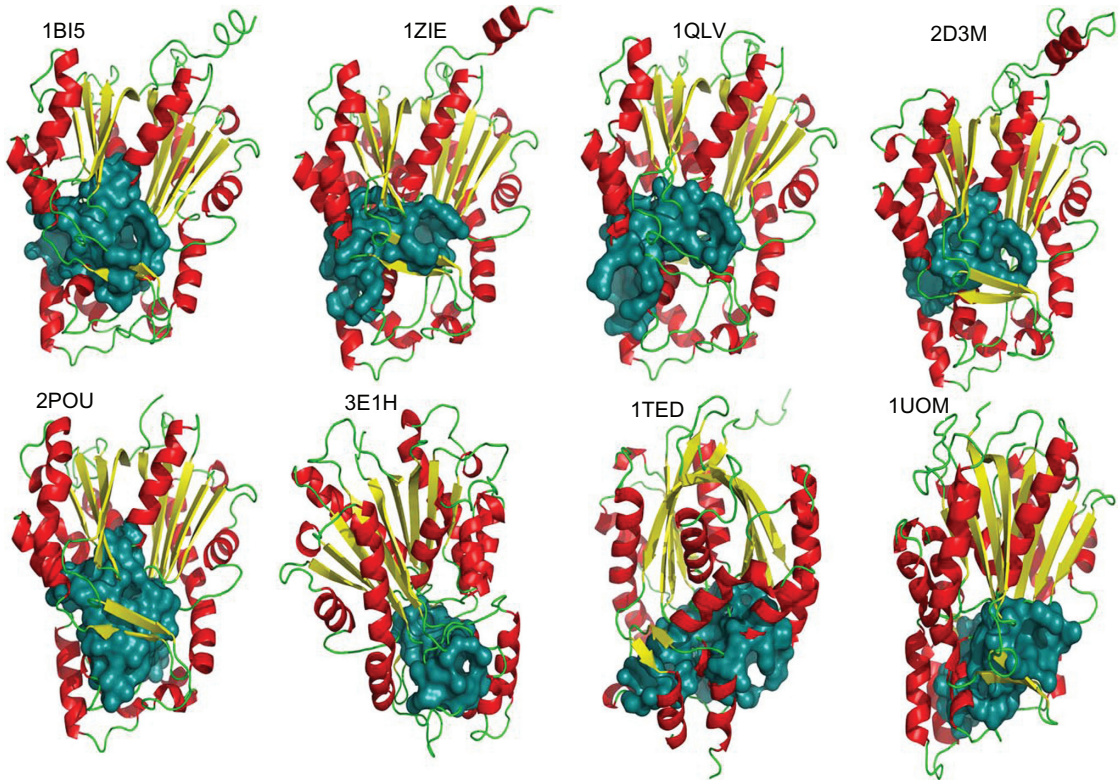


**Figure 6.** Molecular structures of type III PKS showing substrate binding pocket.

the cytoplasmic matrix. It is obvious that the disparity in the physicochemical properties and molecular structures is modest in case of type III PKSs from angiospermic plants. But the bacterial and fungal type III PKSs exhibited numerous variations (Supplementary data 3, Table 7). Analysis on amino acid composition (Supplementary data 3, Table 8) revealed that the residues 'Ala', 'Glu', 'Gly', 'Leu', 'Lys', 'Val' and 'Ile' were found abundant in proteins (totalling above 50% of all the amino acids in examined cases), while 'Cys', 'His' and 'Trp' in very low percentile (<3%).

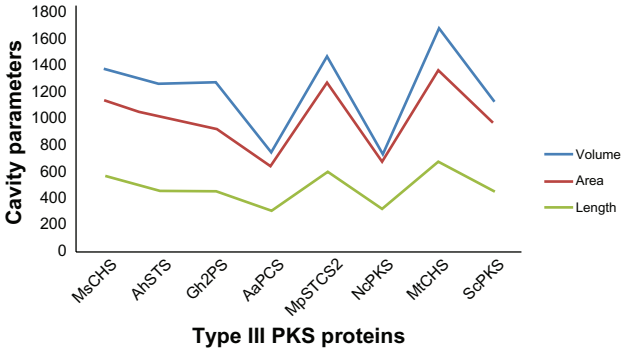


**Figure 7.** Variation of cavity parameters (Å) in different type III PKS. Here MtCHS shows highest values area, length and volume. While NcPKS shows the least volume and AaPCS shows the lowest area. (MsCHS- *M. sativa* CHS, AhSTS- *A. hypogaea* STS, Gh2PS- *G. hybrida* 2PS, AaPCS- *A. arborescens* PCS, MpSTCS2- *M. polymorpha* STCS2, NcPKS- *N. crassa* Type III PKS, MtCHS- *M. tuberculosis* Type III PKS, ScPKS- *S. coelicolor* THNS)

The sub cellular localization prediction by SubLoc v1.0 and PSORTb v.3.0 suggested that all the examined type III PKSs from plants and fungi were localized in the cytoplasmic matrix exclusive of any transmembrane peptides. This indicated that all plant and fungal type III PKSs are functional in the cytoplasmic matrix but in the case of bacteria (*M. tuberculosis*), CHS is located in the cytoplasmic membrane. PSORTb and TargetP prediction identified that the sequence contained mTP, a mitochondrial targeting peptide,[42] 80 amino acids in length. CHS was found to be localized in the bacterial cytoplasmic matrix in *S. coelicolor*.

## Summary

*In silico* sequence and structural analysis of type III PKSs showed evolutionarily and structurally related regions in a compilation of amino acid sequences from different families. CHSs from higher plants displayed more similarity when compared to type III PKSs from bacteria and fungi. Besides the conserved cysteine residues; the amino acid residues in the active site,

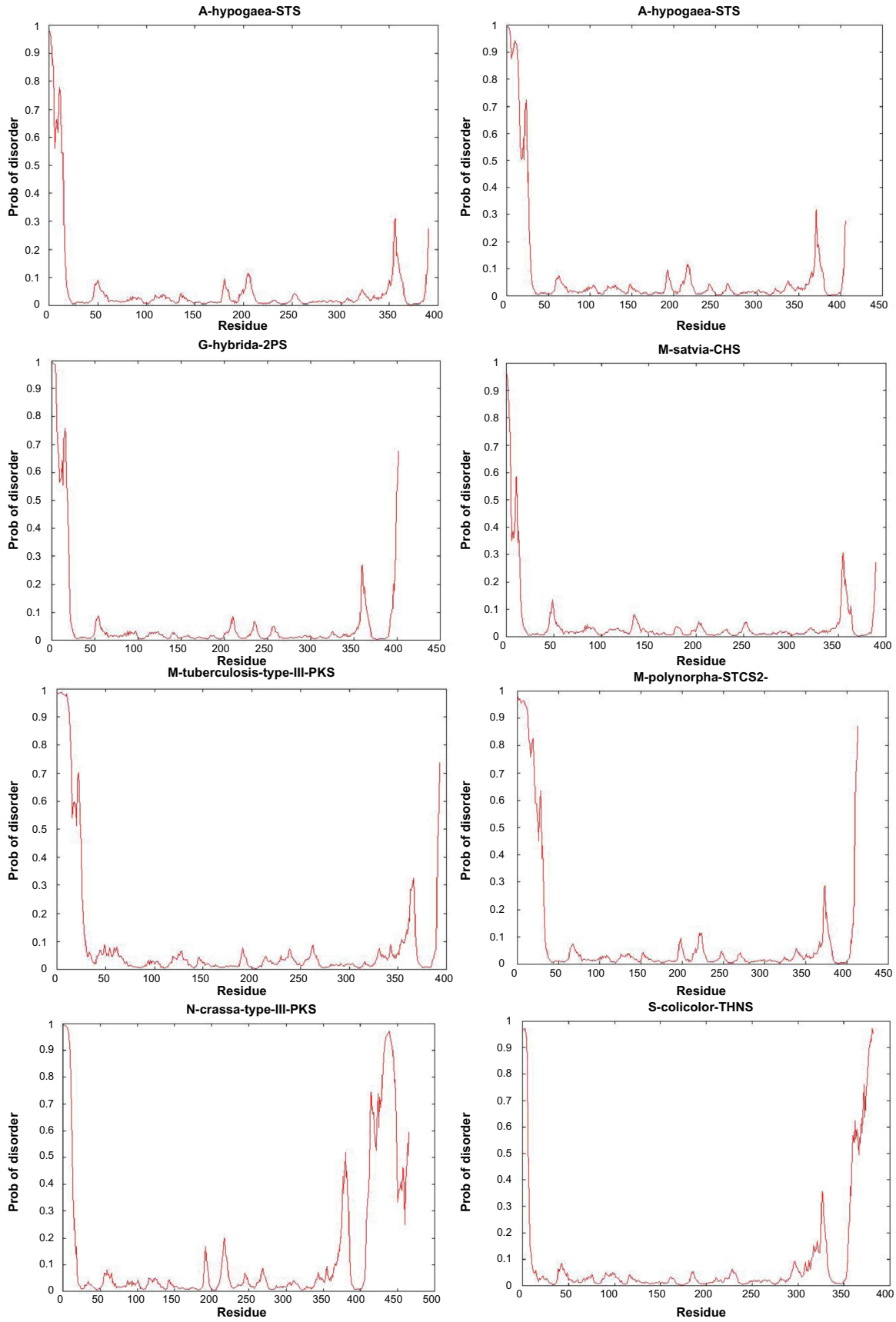


**Figure 8.** Probability of disorder in type III PKS proteins. The red curve shows the predicted probability of disorder for each residue in the protein sequence.

the substrate binding pocket, the cyclization pocket and the 'GFGPG' loop that forms the cyclization scaffold also exhibited a high degree of conservation across the various type III PKSs. Molecular phylogenetic analysis showed that the type III PKS proteins from plants, bacteria and fungi were clustered separately. Proteins from the primitive bryophytes and pteridophytes grouped immediately near the fungi, as a proof of the evolutionary divergence that occurred in the type III PKS superfamily. Structural analysis showed that the protein's secondary structures were mainly composed of alpha helices and random coils. *In silico* sequence analysis revealed that type III PKS proteins are highly hydrophobic in nature and mostly localized in the cytoplasmic matrix.

## Abbreviations

PKS, polyketide synthase; 2PS, 2-pyrone synthase; VPS, Phloroisovalerophenone synthase; THBS, Trihydroxybenzophenone synthase, ACS2, Acridone synthase; STS1, Stilbene synthase 1; PCS, Pentaketide chromone synthase; OS, Octaketide synthase; RS3, Resveratrol synthase 3; BAS, Benzalacetone synthase; ALS, Aloesone synthase; STCS2, Stilbenecarboxylate synthase 2; THNS, Putative 1, 3, 6, 8-tetrahydroxynaphthalene synthase; CHS, Chalcone synthase.

## Acknowledgements

This study was supported by the Department of Information Technology, Government of India. The authors acknowledge BTISNet, Department of Biotechnology, Government of India for the Bioinformatics facility.

## Disclosures

This manuscript has been read and approved by all authors. This paper is unique and not under consideration by any other publication and has not been published elsewhere. The authors and peer reviewers report no conflicts of interest. The authors confirm that they have permission to reproduce any copyrighted material.

## References

1. Watanabe K, Praseuth AP, Wang CCC. A comprehensive and engaging overview of the type III family of polyketide synthases. *Curr Opin Chem Biol*. 2007;11(3):279–86.
2. Winkel-Shirley B. Flavonoid biosynthesis. A colorful model for genetics, biochemistry, cell biology, and biotechnology. *Plant Physiol*. 2001;126:485–93.
3. Austin MB, Noel JP. The chalcone synthase superfamily of type III polyketide synthases. *Nat Prod Rep*. 2003;20:79–110.
4. Koes RE, Spelt CE, Mol JNM, Geratas AGM. The chalcone synthase multigene family of *Petunia hybrida* (V30) Sequence homology, chromosomal location and evolutionary aspects. *Plant Mol Biol*. 1987;10:375–85.
5. Ryder TB, Hedrick SA, Bell JN, Liang X, Clouse SD, Lamb CJ. Organisation and differential activation of a gene family encoding the plant defense enzyme CHS in *Phaseolus vulgaris*. *Mol Gen Genet*. 1987;210:219–33.
6. Radhakrishnan EK, Sivakumar KC, Soniya EV. Molecular Characterization of Novel form of Type III Polyketide Synthase from *Zingiber Officinale* Rosc. and its Analysis using Bioinformatics Method. *J Proteomics Bioinform*. 2009;2:310–5.
7. Jang M, Cai L, Udeani GO, et al. Cancer chemopreventive activity of resveratrol, a natural product derived from grapes. *Science*. 1997;275(5297):218–20.
8. Clement MV, Hirpara JL, Chawdhury SH, Pervaiz S. Chemopreventive Agent Resveratrol, a Natural Product Derived From Grapes, Triggers CD95 Signaling-Dependent Apoptosis in Human Tumor Cells. *Blood*. 1998;92(3):996–1002.
9. Ferrer JL, Jez JM, Bowman ME, Dixon RA, Noel JP. Structure of chalcone synthase and the molecular basis of plant polyketide biosynthesis. *Nat Struct Biol*. 1999;6:775–84.
10. Austin MB, Izumikawa M, Bowman ME, et al. Crystal Structure of a Bacterial Type III Polyketide Synthase and Enzymatic Control of Reactive Polyketide Intermediates. *J Biol Chem*. 2004;279:45162–74.
11. Jez JM, Bowman ME, Noel JP. Expanding the biosynthetic repertoire of plant type III polyketide synthases by altering starter molecule specificity. *Proc Natl Acad Sci U S A*. 2002;99:5319–24.
12. Heath RJ, Rock CO. The Claisen condensation in biology. *Nat Prod Rep*. 2002;19(5):581–96.
13. Schroder J. A family of plant-specific polyketide synthases: facts and predictions. *Trends in Plant Science*. 1997;2:373–8.
14. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215:403–10.
15. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. The CLUSTAL_X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res*. 1997;25:4876–82.
16. Larkin MA, Blackshields G, Brown NP, et al. Clustal W and Clustal X version 2.0. *Bioinformatics*. 2007;23:2947–8.
17. Tamura K, Dudley J, Nei M, Kumar S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol*. 2007;24:1596–9.
18. Saitou N, Nei M. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol Biol Evol*. 1987;4:406–25.
19. Eck RV, Dayhoff MO. Atlas of Protein Sequence and Structure. National Biomedical Research Foundation, Silver Springs, Maryland; 1966.
20. Rzhetsky A, Nei M. A simple method for estimating and testing minimum evolution trees. *Mol Biol Evol*. 1992;9:945–67.
21. Felsenstein J. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution*. 1985;39:783–91.
22. Dundas J, Ouyang Z, Tseng J, Binkowski A, Turpaz Y, Liang J. CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic Acids Res*. 2006;34:W116–8.
23. Gasteiger E, Hoogland C, Gattiker A, et al. Protein Identification and Analysis Tools on the ExPASy Server, In: John M. Walker, editor: The Proteomics Protocols Handbook, Humana Press; 2005:571–607.
24. Hua S, Sun Z. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*. 2001;17(8):721–8.
25. Emanuelsson O, Nielsen H, Brunak S, Heijne GV. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J Mol Biol*. 2000;300:1005–16.
26. Gardy JL, Laird MR, Chen F, et al. PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics*. 2005;21(5);617–23.
27. Deng X, Eickholt J, Cheng J. PreDisorder: ab initio sequence-based prediction of protein disordered regions. *BMC Bioinformatics*. 2009;10:436.

28. Cheng J, Randall A, Baldi P. Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins*. 2006;62(4):1125–32.
29. Jensen JL, Gupta R, Blom N, et al. Ab initio prediction of human orphan protein function from post-translational modifications and localization features. *J Mol Biol*. 2002;319:1257–65.
30. Jensen JL, Sterfeldt HH, Brunak S. Prediction of human protein function according to gene ontology categories. *Bioinformatics*. 2003;19:635–42.
31. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 2000;16(6):276-7.
32. Garnier J, Osguthorpe DJ, Robson B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J Mol Biol*. 1978;120(1):97–120.
33. Jez JM, Noel JP. Mechanism of chalcone synthase. pKa of the catalytic cystein and the role of the conserved histidine in a plant polyketide synthase. *J Biol Chem*. 2000;275:39640–6.
34. Lanz T, Tropf S, Marner FJ, Schroder J, Schroder G. The role of cysteines in polyketide synthases: site-directed mutagenesis of resveratrol and chalcone synthases, two key enzymes in different plant-specific pathways. *J Biol Chem*. 1991;266, 9971–6.
35. Schroeder J, Schroeder G. Stilbene and chalcone synthases: related enzymes with key functions in plant-specific pathways. Z. *Naturforsch C*. 1990;45(1–2):1–8.
36. Suh DY, Fukuma K, Kagami J, et al. Identification of amino acid residues important in the cyclization reactions of chalcone and stilbene synthases. *Biochem J*. 2000;350:229–35.
37. Jiang C, Kim SY, Suh DY. Divergent evolution of the thiolase superfamily and chalcone synthase family. *Mol Phylogenet Evol*. 2008;49(3): 691–701.
38. Zuckerkandl E, Pauling L. Evolutionary divergence and convergence in proteins. Evolving Genes and Proteins, Academic Press, New York; 1965:97–166.
39. Nei M, Kumar S. Molecular Evolution and Phylogenetics. Oxford University Press, New York; 2000.
40. Jez JM, Bowman ME, Noel JP. Structure-guided programming of polyketide chain-length determination in chalcone synthase. *Biochemistry*. 2001;40 (49):14829–38.
41. Laskowski RA, Luscombe NM, Swindells MB, Thornton JM. Protein clefts in molecular recognition and function. *Protein Sci*. 1996;5:2438–52.
42. Nielsen H, Engelbrecht J, Brunak S, Heijne GV. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Engineering*. 1997;10:1–6.
43. McKhann HI, Hirsch AM. Isolation of chalcone synthase and chalcone isomerase cDNAs from alfalfa (*Medicago sativa* L.): highest transcript levels occur in young roots and root tips. *Plant Mol Biol*. 1994;24(5):767–77.
44. Helariutta Y, Elomaa P, Kotilainen M, Griesbach RJ, Schroder J, Teeri TH. Chalcone synthase-like genes active during corolla development are differentially expressed and encode enzymes with different catalytic properties in *Gerbera hybrida* (Asteraceae). *Plant Mol Biol*. 1995;28(1):47–60.
45. Koskela S, Elomaa P, Helariutta Y, Soderholm P, Vuorela P. Two bioactive compounds and a novel chalcone synthase like enzyme identified in *Gerbera hybrida*. Proc. IV IS on In Vitro Cult. & Hort Breeding. *ISHS Acta Hort*. 2001;560:271–4.
46. Paniego NB, Zuurbier KW, Fung SY, et al. Phlorisovalerophenone synthase, a novel polyketide synthase from hop (*Humulus lupulus* L.) cones. *Eur J Biochem*. 1999;262(2):612–6.
47. Liu B, Falkenstein-Paul H, Schmidt W, Beerhues L. Benzophenone synthase and chalcone synthase from *Hypericum androsaemum* cell cultures: cDNA cloning, functional expression, and site-directed mutagenesis of two polyketide synthases. *Plant J*. 2003;34:847–55.
48. Wanibuchi K, Zhang P, Abe T, et al. An acridone-producing novel multifunctional type III polyketide synthase from *Huperzia serrata*. *FEBS J*. 2007;274(4):1073–82.
49. Liu B, Raeth T, Beuerle T, Beerhues L. Biphenyl synthase, a novel type III polyketide synthase. *Planta*. 2007;225(6):1495–503.
50. Schoppner A, Kindl H. Purification and properties of a stilbene synthase from induced cell suspension cultures of peanut. *J Biol Chem*. 1984; 259(11):6806–11.
51. Abe I, Utsumi Y, Oguro S, Morita H, Sano Y, Noguchi H. A Plant Type III Polyketide Synthase that Produces Pentaketide Chromone. *J Am Chem Soc*. 2005;127:1362–3.
52. Abe I, Oguro S, Utsumi Y, Sano Y, Noguchi H. Engineered biosynthesis of plant polyketides: chain length control in an octaketide-producing plant type III polyketide synthase. *J Am Chem Soc*. 2005;127(36):12709–16.
53. Abe I, Takahashi Y, Morita H, Noguchi H. Benzalacetone synthase. A novel polyketide synthase that plays a crucial role in the biosynthesis of phenylbutanones in *Rheum palmatum*. *Eur J Biochem*. 2001;268(11):3354–9.
54. Abe I, Utsumi Y, Oguro S, Noguchi H. The first plant type III polyketide synthase that catalyzes formation of aromatic heptaketide. *FEBS Lett*. 2004; 562(1–3):171–6.
55. Preisig-Muller R, Gnau P, Kindl H. The inducible 9, 10-dihydrophenanthrene pathway: characterization and expression of bibenzyl synthase and S-adenosylhomocysteine hydrolase. *Arch Biochem Biophys*. 1995;317(1):201–7.
56. Akiyama T, Shibuya M, Liu HM, Ebizuka Y. p-Coumaroyltriacetic acid synthase, a new homologue of chalcone synthase, from *Hydrangea macrophylla* var. thunbergii. *Eur J Biochem*. 1999;263(3):834–9.
57. Eckermann C, Schroder G, Eckermann S, et al. Stilbenecarboxylate biosynthesis: a new function in the family of chalcone synthase-related proteins. *Phytochemistry*. 2003;62(3):271–86.