Available online **www.ijpras.com**

International Journal of Pharmaceutical Research & Allied Sciences, 2017, 6(1):22-46

Research Article

ISSN : 2277-3657
CODEN(USA) : IJPRPM

Functional characterization of cytochrome P450 variant (CYP71) isolated from Artemisia annua L. Plants

Pravej Alam[1], Malik Zainul Abdin[2], Naveed Ahmad[3], Mirza Sarwar Baig[4], Mahmoud A. Sharaf-Eldin[1,5] and Shereen F. Elkholy[1,6]

[1]Sara Alghonaim Research Chair (SRC), Biology Department, College of Science and Humanities, Prince Sattam bin Abdulaziz University (PSAU), 11942 Alkharj, Kingdom of Saudi Arabia (KSA)
[2]Centre for Transgenic Plant Development, Department of Biotechnology, Faculty of Science, Jamia Hamdard, New Delhi, India -110062
[3]Department of Computer Sciences, Jamia Hamdard, New Delhi-110062, India
[4]Department of Biosciences, Faculty of Natural Sciences, Jamia Millia Islamia, New Delhi, India, 110025
[5]Department of Medicinal and Aromatic Plants Research, National Research Centre (NRC), 12622-Cairo, Egypt
[6]Plant Transformation and Biopharmaceuticals Lab, Agricultural Genetic Engineering Research Institute (AGERI), Agricultural Research Centre (ARC), Giza, Egypt

Running Title: Functional Characterization of CYP71 in Artemisinin

Corresponding author
Dr. P. Alam
Sara Alghonaim Research Chair (SRC),
Biology Department, College of Science and Humanities,
Prince Sattam bin Abdulaziz University (PSAU),
11942 Alkharj, Kingdom of Saudi Arabia (KSA)
E mail:alamprez@gmail.com

ABSTRACT

Cytochrome P450 monooxygenase isolated from anti-malarial Artemisia annua L. plants belongs to the family Asteraceae is involved in artemisinin biosynthesis considered as a vigorous anti-malarial drug. To understand the structural and functional features of amorpha-4,11-dienemonooxygenase, a variant of cytochrome P450 monooxygenase (CYP71) may help in the up-regulation of artemisinin. In this context, biocomputational approaches were used to study the amorpha-4,11-diene monooxygenase catalytic features related to other biosynthetic enzymes targeted from artemisinin biosynthesis. The full-length gene was cloned from a high yielding artemisinin strain of the A. annua L. plant with size of (2054 bp) and encodes for the CYP71 protein (495aa) protein residues with different conserved domains/motifs. The molecular weight ~55.73kDa and isoelectric point 9.13 of CYP were observed through in silico analysis. The phylogenetic analysis revealed that cyp71 gene of A. annua was evolutionary conserved with other plant species of cytochrome family. The Modelled 3D structure of CYP71 protein has shown the catalytic relationship between the three different ligand molecules such as farnesyl pyrophosphate (FPP), omega 4,11-diene (AD) and artemisinic alcohol (AA). The in silico information showed a key role of binding of these molecules for up-regulation of artemisinin biosynthesis and its accumulation.

Keywords: *in silico characterization, Artemisinin, cyp71 gene, Phylogenetic tree, Homology modelling, Docking.*

INTRODUCTION

Plant cytochrome P450 monooxygenases (CYPs) form a large super-family is composed of multifunctional, heme-iron complex. It is catalysing the NADPH and oxygen dependent hydroxylation reactions in plants system and also involved in the biosynthesis of a variety of bioactive compounds such as artemisinin, vinblastine, fatty acids, hormones, pigments etc. [1-3]. The biosynthetic route for artemisinin biosynthesis proposed by different investigators starting from farnesyl pyrophosphate (FPP) to artemisinin via different cascades [4]. The first step in artemisinin biosynthesis is catalysed by amorpha-4,11-diene synthase (ADS) which is a sesquiterpene cyclase, it is followed by the oxidation of amorpha-diene into artemisinic alcohol and artemisinic aldehyde/dihydroartemisinic aldehyde. Further step is catalysed by an enzyme, cytochrome P450 monooxygenase (amorpha-4,11-diene monooxygenase; CYP71) which is encoded by *cyp*71 gene. Thereafter, a cytochrome P450 monooxygenase oxidizes amorpha-4,11-diene at the C12 position to artemisinic acid in three sequential steps [5-6]. However, recent data proposed that the primary route of artemisinin biosynthesis is through the artemisinic aldehyde intermediate resulted from the second oxidation step of cytochrome P450 monooxygenase i.e. amorpha-4,11-diene monooxygenase [7-9]. Artemisinin is biosynthesized from *A. annua* L. plant native of china. The proportional low concentration of artemisinin (0.01-1.1% dry weight) production is the serious concern worldwide [10-11]. Biotechnological approaches have been considered as a feasible approach for its production [11-12]. Recent years, several studies have been conducted in different laboratories to understand artemisinin biosynthesis and its regulation in *A. annua* L. plants to improve the content of artemisinin [5,11-12]. Keeping the facts observed from literatures regarding the plant cytochrome P450 (CYPs) and its complexity in metabolic network between the enzymes, we have analysed *in silico* based study for the catalytic relationship of CYP71 with other enzymes linked to the artemisinin biosynthesis.

Materials and Methods

The *Artemisia annua* seeds (0.7-0.9% artemisinin DW, dry weight basis) were obtained from the Centre for Transgenic Plant Development, Department of Biotechnology, Hamdard University, New Delhi, India. Plants were grown in the experimental field of Jamia Hamdard, New Delhi, India.

Isolation and PCR amplification of amorpha-4,11-diene monooxygenase (CYP71)

The genomic DNA was isolated from leaves of high yielding morphotypes of *A. annua* L. plant, using the C-TAB Method [13]. The primers were used in this study, designed on according to the previously available sequence from NCBI data base (Accession No. DQ826743). The primers were used, Forward primer; 5' GAAGAGTATACTAAAAGCAAT3' and Reverse primer; 5'AAACTTGGAACGAGTAACAACTCA3' for PCR amplification of *cyp* gene. The PCR amplification was done using the method of [14]

Cloning and sequencing of amorpha-4,11-diene monooxygenase gene

The resulting PCR product was purified and ligated into the pGEM-T-easy cloning vector (Promega Inc.) following the manufacturer's instruction. The ligation mixture was then used to transform *E. coli* DH5α (100 µl) competent cells. The cells were mixed with 900 µl of Luria broth (LB) medium, incubated at 37°C for an hour and *E. coli* cell culture (100 µl) was plated onto X-gal/Amp/IPTG/LB plate. Following overnight incubation at 37°C, the recombinant colonies were picked and used for plasmid extraction. Plasmid DNA was purified using GeneJETTMPlasmid Mini prep Kit (Fermentas Life Science) following the manufacturer's instructions. The clones were confirmed through PCR amplification. The automated sequencing was revealed for *cyp* gene in both the forward and reverse directions using SP6 and T7 promoter universal primers using the ABI Prism automated DNA sequencer (Chromus Biotech, India).

Sequence and phylogenetic analysis

The nucleotide and protein sequence were analysed on the basis of homology searches using the basis local alignment search tools (BLAST) algorithm. Sequence analysis was performed by using online tools of the ExPASy Server and Splign (http://www.ncbi.nlm.nih.gov/sutils/splign/splign.cgi) for nucleotide and amino acid sequences

and generate comparative of CYP71 [15-16]. Primary structure analysis was done using the ProtParam (http://www.expasy.ch/tools/protparam.html) [17]. The ClustalW multiple sequence alignment program was performed to align the *cyp*71 sequences (http://www.ebi.ac.uk/Tools/msa/clustalw2/) [18]. Nucleotide sequences of *cyp*71 (FJ716705) from *A. annua* L. and other cytochrome P450 monooxygenase genes from *Artemisia compestris*, *Saussurea costus*, *Helianthus annuus*, *Lactuca sativa*, *Cichorium intybus*, *Barnadesia spinosa*, *Ammi majus*, *Panax notoginseng*, *Nicotiana tabacum*, *Solanumchacoense*, *Hyoscyamusmuticus*, *Solanumtuberosum*, *Medicagotruncatula, Glycine max* and *Paper somniferum* were obtained from NCBI GenBank. The evolutionary link between nucleotide sequences was generated by MEGA 5.13 program to create neighbour-Joining tree [19]. Preface probability was conducted to convince the statistical importance of the clusters in the phylogenetic tree [19].

Domains/motifs search

The highly conserved domains and motifs of amorpha-4,11-diene monooxygenase protein (CYP71) were computationally identified through online tools SMART (http://smart.embl.de/smart) [20], and TMHMM v2.0 [www.cbs.dtu.dk/services/TMHMM, 21] online programs.

Homology modelling and template identification of CYP71

The homology modelling of amorpha-4,11-diene monooxygenase (CYP71) of *A. annua* L. was accomplished in the following sequential steps. The protein data bank (PDB) through NCBI BLAST was used to distinguish the suitable template for protein modelling of CYP71 from *A. annua* L. In a trial to search the suitable template against PDB database for modelling the CYP71 protein, Phyre2 (http://www.sbg.bio.ic.ac.uk/phyre) server, for searching similar sequences and structure-wise symmetry, was used [22]. For 3D modelling, single template of the mammalian cytochrome [1nr6] was chosen on the basis of fold recognition.

Sequence alignment and validation of CYP71 structure

The amino acid sequence alignment of target and the template protein was derived using the Swiss-PDB viewer package [http://www.expasy.ch/spdbv/, 23]. Default parameters were utilized and the aligned sequences were observed and well attuned manually to decrease gaps and insertions. The three dimensional structure of CYP71 was predicted using phyre2 MODELLER. The scripts were also used to execute an alignment between both target and the template sequence. The predicted three dimensional model of CYP71 was thereafter selected from the script model default based on the previously generated alignment. The predicted constructed model of CYP71 was solvated and subjected to energy minimization (100 kJ/mol/$Å^2$) for protein atoms, using the advanced decline and conjugate gradient to eliminate the bad contacts between both protein and water molecules. *In vacuo* with the GROMOS96 43B1 computation programs were carried out. The stereo chemical quality of the predicted structure was evaluated by VADAR [24] and PROCHECK [25]. Ramachandran plot of CYP71 analysis was performed with favoured region for checking the accuracy of the predicted model. The stability during the MD simulation time was assessed by calculating the RMSD (root mean square deviation) in between the starting and structures generated from the treatments. The generated model was visualized, inspected and analysed through CHIMERA and PyMOL [26-27].

Docking studies of CYP71 with three different ligand molecules (FPP, AD and AA)

After obtaining the final model, the centre co-ordinates of expected binding sites of amorpha-4,11-diene monooxygenase were extracted using Q-site finder (http://bmbpcu36.leads.ac.uk) [28]. Farnesyl pyrophosphate (FPP), amorpha-4,11-diene (AD) and artemisinic alcohol (AA) were downloaded from Pubchem **(**http://www.ncbi.nlm.nih.gov/pccompound) and used as ligand molecules for docking studies (Table 1). The FPP, AD and AA were docked with CYP71 using the Lamarckian Genetic algorithm (LGA) provided by the Auto Dock program, version 4.2 (http://autodock.scripps.edu/, [29].The docking of FPP, AD and AA were performed with respect to all the 10 binding sites of the enzyme. The residues lining the cavities within 10Å region around the ligand molecule were extracted using an Autodock 4.2 for plot protein-ligand interactions.

Results and Discussion

In Plants, Cytochrome P450s (CYPs) play a leading role in the biosynthesis of varieties of compounds such as hormones, fatty acid, steroid, and terpenes. CYPs have a large binding site so that it can bind to exogenous and endogenous compound easily in plants. Cytochrome P450 monooxygenases are quite related in their physiochemical properties. The cloned *cyp*71 gene sequence (FJ716705) from *A. annua* showed 99% similarity with nucleotide derived from earlier published *A. annua* L. *cyp*71 gene by other investigators (Accession Numbers DQ826743 and DQ667170 obtained from NCBI database; Table 2). The deduced weight of the encoded protein is 55.7 kDa and the pI is calculated to be 9.13. On amino acid scale, CYP71 (495 aa) showed more than 98% identity with cytochrome P450 monooxygenase proteins of *A. annua* L. and other species.

The analysis of Phylogenetic tree showed relatively higher analogy of *cyp*71 gene from *Artemisia annua* L. with other cytochrome P450 monooxygenase of other plants on amino acid scale, especially for angiospermic species were evolutionary conserved Fig. 1 [30]. The evolutionary distances were calculated by using the method of Maximum Likelihood and history by using the Neighbour-Joining method in MEGA 5.10. Phylogenetic trees were constructed according to the original data set by the neighbour-joining method, as well as 1,000 bootstrap data sets generated by CLUSTAL X [31]. The tree was generated using as the cytochrome P450 of *Papaver somniferum* out group sequence to allow the tree to be rooted. The bootstrap analysis revealed that the phylogenetic tree generated is reliable because of high bootstrap value. This finding suggested that cytochrome P450 monooxygenase is most closely related to cytochrome P450 and amorpha-4,11-diene C-12 oxidase (NCBI genbank Accession numbers DQ826743 and DQ667170) of *Artemisia annua* than any other plant (Fig. 1). Different domains/motifs that were found in CYP71 protein are listed in Table 3 and 4.

Two highly conserved regions have been identified in the CYP protein among which the biggest one is P450 domain, having 454 residues of amino acid and another, is a short trans-membrane domain of only 23 residues were observed in Table 3. CYP is highly diverged enzyme family but when it is studied at the sequence level, some conserved motifs are identified in all CYP enzymes based on their tertiary structure and the function which it performs. One of the signature motif which is regarded as the heme binding site is conserved. In this motif only cysteine residue is highly conserved in all P450 enzymes.

Hence, it proved that the amorpha-4,11-diene monooxygenase belongs to the cytochrome P450 monooxygenase family. Another 37 less conserved motifs have also been identified in CYP71 protein of *A. annua* plant. SRP54_N (signal recognition particle 54-type N terminal region), IL4_13 (Interleukins 4 and 13 types), HintC (Hedgehog/Intein domain C-terminal region) and TFS2N (N-terminus of transcription elongation factor S-II type) are the most significant putative motifs present in the CYP71 protein having E-values 1.3, 16.3, 242 and 337 respectively (Table 4). The signal recognition particle (SRP) 54-type region is a helical bundle motif which is repeatedly present in the CYP71 amino acid sequence at positions 178-212, 229-277 and 320-349 with significant E-values 1.3, 530 and 710 Table 4. On the basis of *in silico* analyses of the observed conserved motifs present in CYP71 protein might be play an important role to upregulation of terpene cyclases in this plant [32-33].

The modelled 3D structure of CYP71 with the template protein (1NR6) with RMSD 0.322 Å (Fig. 2) contains 53.33% of α-helices, β-turns (4.6%) and 31.52% of random coils (Fig. 3). The structure of CYP71, α-helices and random coils are considered the most abundant elements of the protein CYP71 while β-turns are distributed in as symmetrical manner. The total values of energy 90 % (after energy minimization) of CYP71 were observed in Ramachandran plot. The CYP71 predicted model, was further analysed by VADAR to check the Ramachandran plot quality (Table 5). Further analysis of 3D structure will help in the identification of the binding sites for studying enzyme substrate mechanisms.

After the refinement of 3D structure, it was used for docking studies with three different ligand molecules such as FPP (Pub Chem. ID, 445713), AD (Pub Chem. ID, 11052747) and AA (Pub Chem. ID, 15983960) related to the artemisinin biosynthetic pathway. Recent studies have revealed that the core sequences of many proteins were nearly optimized for stability by natural evolution. Surface residues, by contrast, were not so optimized, presumably because protein function is mediated through surface interactions with other molecules.

Here, we thought to determine the extent to which the sequences of protein ligand-binding and enzyme active sites could be predicted by optimization of scoring functions based on protein ligand-binding affinity rather than structural stability [14]. In an attempt to find the possible binding sites of FPP, AD and AA on CYP71, the Q site finder was performed. The output of Q site finder centre coordinates for 10 binding sites. Docking of FPP, AD and AA were performed with respect to all the 10 binding sites of the enzyme on the basis of free energy. It was observed that the artemisinic alcohol (AA) showed better binding to the CYP71 (Fig. 4a, b) followed by farnesyl pyrophosphate (FPP) and amorpha-4,11-Diene (AD), respectively (Fig. 5a,b and 6a,b). It is due to the catalytic activity resembled between the enzyme and ligand molecules present in each other. The aim of molecular docking is to achieve an optimized conformation for protein, ligand and relative orientation between each other to minimize the free energy of the overall system to obtain the better binding stretches.

Lower energy corresponds to better binding therefore; initial 5 of 10, these binding sites were studied for the interaction with FPP, AD and AA (Fig.4, 5 and 6). Based on the docking studies, the following stretches are RIAWRGATNLRC, RAFWRDGATLLCV, RLRNMRMCVEGRM, WKRRLMRK, WKRRLMRK: farnesyl pyrophosphate; TTTQLFL, TTTQLFA, TTTQLFAL, TTQLPF, TTQLPFA: amorpha-4, 11-diene; TTQLPFA, TTQPFA, TTQLPFC, TTQLPFA, ILGCGLE: artemisinic alcohol constituted to binding with CYP71.

These results may have implications for understanding the role of CYP in the cyclization of FPP, AD as well as AA. The energy minimization and hydrogen bonding were more effective with the AA than the other (Fig. 4, 5 and 6). These findings may also lead to upregulate the enzyme fluxes by the over-expression of its genes and could enhance the production of metabolites concentration through the cyclization process.

The artemisinin pathway, is first elucidated by amorpha-4,11-diene synthase (ADS) an important enzyme, which converts farnesyl diphosphate into amorpha-4,11-diene [7-34]. Afterward, a cytochrome P450 monooxygenase (CYP71) oxidized amorpha-4,11-diene to artemisinic acid in three successive steps [6-7]. After completion of these steps, artemisinic aldehyde is then converted into artemisinin by other enzymatic and non-enzymatic reactions [35]. It was based on primary sequences on the alignment of the CYP71 family with P450 [9]. The docking of artemisinin alcohol in to CYP71 model proved that there is a loop located in the amino acid residue that covering the C12 position of alcohol and subsequently affect the binding pockets [9].

To understand the structural element and features of enzymes and ligands is the bottle neck for metabolic engineering. The expression of the CYP71 in *A. annua*, is a key enzyme to regulate the biosynthesis of artemisinin. Therefore, this *in silico* information may lead to the understanding of enzyme structure, function and regulation. The structure can give new functional access for the expression and cyclization of this key enzyme towards the artemisinin biosynthesis. The predicted motif or domain of CYP71 may also attribute the functional insight for artemisinin biosynthesis in *A. annua* L. plant. The high homology between different monooxygenases makes modelling of the three dimensional structure of the various cytochrome P450 monooxygenase may also possible. Phylogenetic analysis suggested that *cyp*71 genes among different species are evolutionary conserved.

The structure and function of CYP71 protein is now clearly showed that it is an important class of enzymes in the artemisinin biosynthetic pathway. This study can also be served as a platform for other protein to understand their role in biosynthesis of important metabolites in plants.

Acknowledgement

Author P.A., is thankful to UGC for financial support in the form of Dr. D. S. Kothari Post-Doctoral Fellowship. Authors are also thankful to Sara Alghonaim Research Chair (SRC), Biology Department, College of Science and Humanities, Prince Sattam bin Abdulaziz University (PSAU) for providing scientific facilities.

Conflict of Interest: Authors declare no conflict.

References

1. Fischer M, Knoll M, Sirim D, Wagner F, Funke S, Pleiss J. The Cytochrome P450 Engineering Database: a navigation and prediction tool for the cytochrome P450 protein family. Bioinformatics 2007; 23: 2015–7.

2. Nelson DR. Progress in tracing the evolutionary paths of cytochrome P450. Biochim Biophys Acta2011; 1814: 14–8.

3. Misra A, Chanotiya CS, Gupta MM, Dwivedi UN, Shasany AK. Characterization of cytochrome P450 monooxygenases isolated from trichome enriched fraction of *Artemisia annua* L. leaf. Gene 2012; 510: 193–201.

4. Abdin MZ, Israr M, Rehman RU, Jain SK. Artemisinin, a novel antimalarial drug: biochemical and molecular approaches for enhanced production. Planta Med 2003; 69: 289–99.

5. Aquil S, Husaini AM, Abdin MZ, Rather GM. Overexpression of the HMG-CoA reductase gene leads to enhanced artemisinin biosynthesis in transgenic *Artemisia annua* plants. Planta Med.2009; 75: 1453–8.

6. Bertea CM, Freije JR, van der Woude H, Verstappen FW, Perk L, Marquez V, De Kraker J-W, Posthumus MA, Jansen BJ, de Groot A, et al. Identification of intermediates and enzymes involved in the early steps of artemisinin biosynthesis in Artemisia annua. Planta Med 2005; 71: 40–7.

7. Wallaart T, van Uden W, Lubberink H, Woerdenbag H, Pras N, Quax W. Isolation and identification of dihydroartemisinic acid from *Artemisia annua* and its possible role in the biosynthesis of artemisinin. J Nat Prod 1999; 62: 430–3.

8. Teoh KH, Polichuk DR, Reed DW, Nowak G, Covello PS. *Artemisia annua* L. (Asteraceae) trichome-specific cDNAs reveal CYP71AV1, a cytochrome P450 with a key role in the biosynthesis of the antimalarial sesquiterpene lactone artemisinin. FEBS Lett2006; p 580: 1411–6.

9. Komori A, Suzuki M, Seki H, Nishizawa T, Meyer JJM, Shimizu H, Yokoyama S, Muranaka T. Comparative functional analysis of CYP71AV1 natural variants reveals an important residue for the successive oxidation of amorpha-4,11-diene. FEBS Lett 2013; 587: 278–84.

10. Dhingra V, Vishweshwar Rao K, Lakshmi Narasu M. Current status of artemisinin and its derivatives as antimalarial drugs. Life Sci 2000; 66:279–300.

11. Alam P, Abdin MZ. Over-expression of HMG-CoA reductase and amorpha-4,11-diene synthase genes in *Artemisia annua* L. and its influence on artemisinin content. Plant Cell Rep 2011; 30:1919–28.

12. Chen JL, Fang HM, Ji YP, Pu GB, Guo YW, Huang LL, Du ZG, Liu BY, Ye HC, Li GF, Wang H. Artemisinin biosynthesis enhancement in transgenic *Artemisia annua* plants by downregulation of the β-caryophyllene synthase gene. Planta Med.2011; 77: 1759–65

13. Doyle JL, Doyle J. A rapid total DNA preparation procedure for fresh plant tissue. Focus (Madison).1990; 12: 13–15.

14. Alam P, Kiran U, Ahmad MM, Kamaluddin, Khan MA, Jhanwar S, Abdin MZ. Isolation, characterization and structural studies of amorpha - 4, 11-diene synthase (ADS(3963)) from *Artemisia annua* L. Bioinformation2010; 4: 421–9.

15. Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A. ExPASy: The proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res 2003; 31: 3784–8.

16. Wilkins MR, Gasteiger E, Bairoch A, Sanchez JC, Williams KL, Appel RD, Hochstrasser DF. Protein identification and analysis tools in the ExPASy server. Methods Mol Biol.1999; 112: 531–52.

17. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, et al., 2007. Clustal W and Clustal X version 2.0. Bioinformatics. 23: 2947–8.

18. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol Biol Evol 2013; 30: 2725–9.

19. Letunic I, Doerks T, Bork P. SMART 7: recent updates to the protein domain annotation resource. Nucleic Acids Res 2012; 40: 302–5.

20. Kelley LA, Sternberg MJE. Protein structure prediction on the Web: a case study using the Phyre server. Nat Protoc 2009; 4: 363–71.

21. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol.2001; 305: 567–80.

22. Guex N, Peitsch MC. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis.1997; 18: 2714–23.

23. Willard L, Ranjan A, Zhang H, Monzavi H, Boyko RF, Sykes BD, Wishart DS. VADAR: a web server for quantitative evaluation of protein structure quality. Nucleic Acids Res.2003; 31: 3316–9.

24. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Crystallogr1993; 26: 283–291.

25. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem2004; 25: 1605–12.

26. Delano WL. PyMOL:An Open-Source Molecular Graphics Tool.2013;http://www.ccp4.ac.uk/newsletters/newsletter40/11_pymol

27. Laurie ATR, Jackson RM. Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. Bioinformatics2005; 21: 1908–16.

28. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, Olson AJ, Al MET. Automated Docking Using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function. J Comput Chem1998; 19: 1639–1662.

29. Chaban C, Waller F, Furuya M, Nick P. Auxin responsiveness of a novel cytochrome p450 in rice coleoptiles. Plant Physiol.2003; 133: 2000–9.

30. Saitou N, Nei M. The number of nucleotides required to determine the branching order of three species, with special reference to the human-chimpanzee-gorilla divergence. J Mol Evol.1986; 24: 189–204.

31. 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.

32. Nelson DR. Cytochrome P450 and the individuality of species. Arch Biochem Biophys. 1999; 369: 1–10.

33. Ayebi S, Akashi T. Cytochrome P450s in flavonoid metabolism. Phytochem Rev.2006; 5: 271–282.

34. Chang YJ, Song SH, Park SH, Kim SU. Amorpha-4,11-diene synthase of Artemisia annua: cDNA isolation and bacterial expression of a terpene synthase involved in artemisinin biosynthesis. Arch Biochem Biophys 2000; 383: 178–84.

35. Weathers PJ, Arsenault PR, Covello PS, McMickle A, Teoh KH, Reed DW. Artemisinin production in Artemisia annua: studies in planta and results of a novel delivery method for treating malaria and other neglected diseases. Phytochem Rev2011; 10: 173–183.

Legends and captions

Fig. 1: Phylogenetic tree constructed by parsimony analyses of cytochrome P450 monooxygenase amino acid sequences selected from the BLAST profile. CLUSTAL-W from the Molecular Evolutionary Genetics Analysis program (MEGA5.10) was used for multiple alignments. Numbers on the branches are bootstrap values obtained for 1000 replicates are shown. GenBank accession numbers are shown first and cytochrome P450 amino acid sequences with their respective species is shown on the right side. The cytochrome P450 of *Papaver somniferum* is selected as an out-group to draw the rooted tree.

Fig. 2: Diagrammatic representation of 3D structure of model protein (CYP71) and Yellow colour is template protein (1NR6) with RMSD 0.322 Å

Fig. 3: Prediction of three dimensional structure of Amorpha-4,11-diene monooxygenase (CYP71) showing N and C terminal extracted from *Artemisia annua* L. plant.

Fig. 4a: Interaction of artemisinic alcohol (AA) to the CYP71 protein with different binding sites having different binding energy.

Fig. 4b: Receptor molecule (CYP71) with ligand artemisinic alcohol (AA) docked into top (five) ranking binding sites. All the five figures in different colour correspond to five binding sites containing the ligand molecule AA docked into it.

Fig. 5a: Interaction of FPP to the CYP71 protein with different binding sites having different binding energy.

Fig. 5b: Receptor molecule (CYP71) with the ligand FPP docked in five top ranking binding sites. All the five figures in different colour correspond to five binding sites with the ligand molecule FPP.

Fig. 6a: Interaction of amorpha-4,11-diene to the CYP71 protein with different binding sites having different binding energy.

Fig. 6b: Receptor molecule (CYP71) with the ligand Amorpha-4,11-diene docked in five top ranking binding sites. All the five figures in different colour correspond to five binding sites containing the ligand molecule AD docked into it.

Table 1: Ligands information chart of artemisinic alcohol, amorpha,4-11,diene and farnesyl pyrophosphate.

Table 2: Comparative Nucleotide sequence analysis of amorpha-4,11-diene monooxygenase with other monooxygenase (EF197889). The deduced amino acid sequence is given in the one-letter code below above the corresponding nucleotide sequence. *cyp*71 nucleotide sequence has been deposited in GenBank under accession number FJ716705.

Table 3: Two putative conserved domains have been detected in amino acid sequence of amorpha-4,11-dienemonooxygenase (CYP71) of *A. annua.*

Table 4: Functional motif present in the amino acid sequence of amorpha-4,11-diene monooxygenase (CYP71) of *A. annua* L.

Table 5: Statistical analyses of amorpha-4,11-dienemonooxygenase CYP71 by using the VADAR (Volume, Area, Dihedral Angle Reporter).

Table 1: Ligands information chart of artemisinic alcohol, amorpha,4-11,diene and farnesyl pyrophosphate.

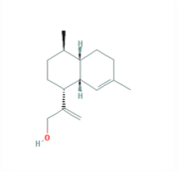
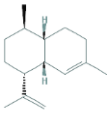
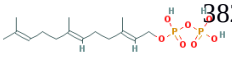
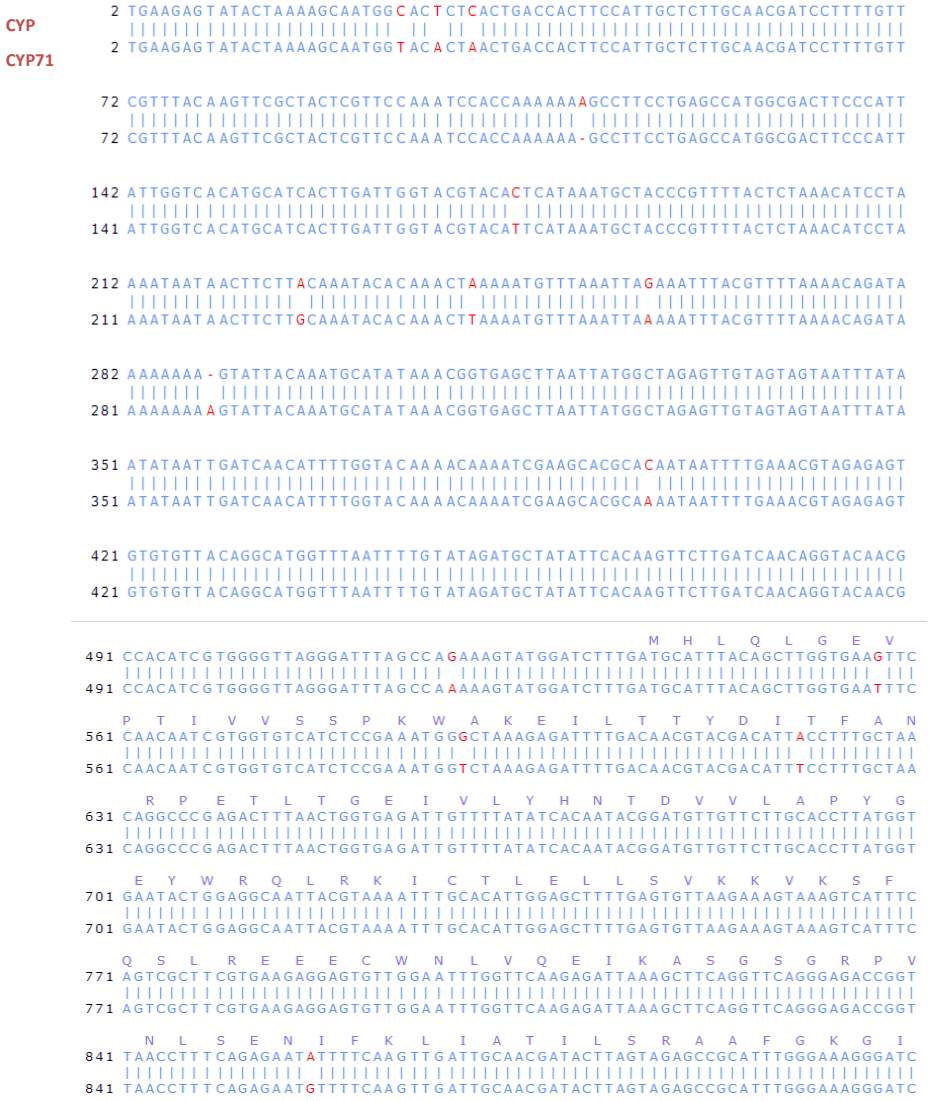
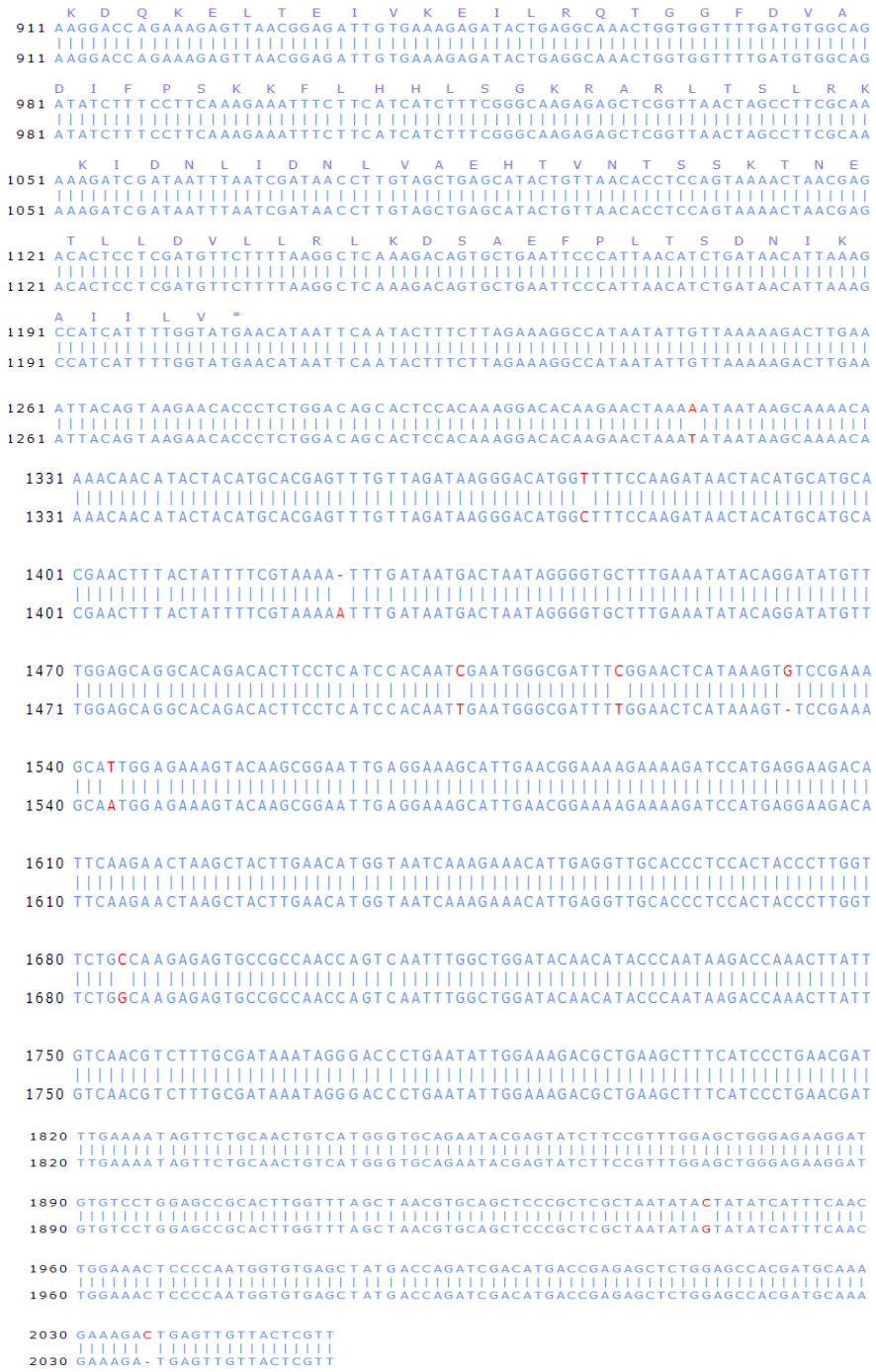
Ligands	Pubchem Compound ID	Image	Mol Weight [g/mol]	X-logP	H-Bond Donor	H-Bond Acceptor	Rotatable bond
Artemisinic alcohol ($C_{15}H_{24}O$)	15983960		220.35046	3.8	1	1	2
amorpha-4,11-diene ($C_{15}H_{24}$)	11052747		204.35106	5.1	0	0	1
farnesyl pyrophosphate ($C_{15}H_{28}O_7P_2$)	445713		382.32614	2.6	3	7	11

Table 2: Comparative Nucleotide sequence analysis of amorpha-4,11-diene monooxigenase with other monooxigenase (EF197889). The deduced amino acid sequence is given in the one-letter code below the corresponding nucleotide sequence. cyp71 nucleotide sequence has been deposited in GenBank under accession number FJ716705.





Coverage	99.37%	CDS	0.00%
Overall	98.26%	In-frame	0.00%
Exon	98.88%	Primary transcript	2050 bp

Mismatches and indels	25
Exons (min/max/ave), bp	2050 / 2050 / 2050
Introns (min/max/ave), bp	

Table 3: Two putative conserved domains have been detected in amino acid sequence of amorpha-4, 11- diene monooxygenase (CYP71) of *A. annua* L.

Name of the conserved Domains	Start Position	End Position	e-value
Transmembrane region	7	29	Not significant
Pfam p450	39	492	2.8e-97

Table 4: Other motif present in the amino acid sequence of amorpha-4, 11- diene monooxygenase (CYP71) of A. annua L.

Name	Start	End	E-value	Description
AIP3	1	343	59800	Actin interacting protein 3
low complexity	9	23	59800	This is a region of 15 amino acid long low compositional complexity,
T5orf172	47	117	148000	represents a DNA-binding domain found in bacteriophage T5, ORF172
HhH2	67	99	1020	Helix-hairpin-helix class 2 (Pol1 family) motifs
Spc7	75	356	75000	This domain is found in cell division proteins which are required for kinetochore-spindle association.
HintC	77	125	242	Hint (Hedgehog/Intein) domain C-terminal region split to accommodate large insertions of endonucleases.
RGS	102	233	3190	Regulator of G protein signalling domain
SEP	106	179	1130	Domain present in Saccharomyces cerevisiae Shp1, Drosophila melanogaster eyes closed gene (eyc), and vertebrate p47.
DEXDc2	121	339	2400	DEAD-like helicases superfamily
Elp3	149	409	3330	Elongator protein 3, MiaB family, Radical SAM
THUMP	149	244	141000	The THUMP domain is named after thiouridine synthases, methylases and PSUSs

MA	151	356	1200	Methyl-accepting chemotaxis-like domains (chemotaxis sensory transducer).
OmpH	158	396	144000	Outer membrane protein (OmpH-like)
BPI2	169	394	1180	Bactericidal permeability-increasing protein (BPI) / Lipopolysaccharide-binding protein (LBP) / Cholesteryl ester transfer protein (CETP) C-terminal domain
ADSL_C	173	241	45400	Adenylosuccinate lyase C-terminus
BP28CT	175	283	111000	found in BAP28-like nucleolar proteins
AgrD	175	209	9100	Staphylococcal AgrD protein
ClpB_D2-small	177	277	87300	C-terminal, D2-small domain, of ClpB protein
IL4_13	177	283	16.3	Interleukins 4 and 13
Pfam:SRP54_N	178	212	1.3	signal recognition particle (SRP)54-type protein, helical bundle domain
VHS	185	320	817	Domain present in VPS-27, Hrs and STAM
Resolvase	186	311	67900	Resolvase, N terminal domain
eIF3_N	187	296	13600	eIF3 subunit 6 N terminal domain
RAP	188	246	54400	RNA-binding domain abundant in Apicomplexans
RPOL4c	201	323	1590	DNA-directed RNA-polymerase II subunit
ENDO3c	211	358	3440	endonuclease III
Pfam:SRP54_N	229	277	530	signal recognition particle (SRP)54-type protein, helical bundle domain
SMR	230	298	838	Small MutS-related domain
ALBUMIN	232	377	1140	serum albumin
HLH	233	283	1770	helix loop helix domain
DALR_1	238	367	66200	DALR anticodon binding domain

B12-binding_2	244	338	132000	Cobalamin-dependent methionine synthase (b12)binding domain
GatB_Yqey	249	334	50000	GatB domain, the function of which is uncertain, is associated with aspartyl/glutamyl amidotransferase subunit B and glutamyl amidotransferase subunit E
TFS2N	251	326	337	Domain in the N-terminus of transcription elongation factor S-II
Ubox	288	343	1750	Modified RING finger domain Probable involvement in E2-dependent ubiquitination.
Pfam:SRP54_N	320	349	710	signal recognition particle (SRP)54-type protein, helical bundle domain
TDU	323	338	1240	Short repeats in human TONDU, fly vestigial and other proteins.
low complexity	359	371	1240	13 amino acid long region of low compositional complexity

Table 5: Statistical values of amorpha-4, 11- diene monooxygenase CYP71 protein of *Artemisia annua* L. by using the VADAR (Volume, Area, Dihedral Angle Reporter).

Statistical	**Observed**	**Expected**
Expected values for highly refined X-ray and NMR protein structures		
Helix	233 (50 %)	-
Beta	64 (13%)	-
Coil	165 (35 %)	-
Turn	116(25%)	-
Hydrogen Bonds(H-bonds)		
Mean bond distance	2.2 sd=0.4	2.2 sd=0.4
Mean bond energy	-1.6 sd=1.0	-2.0 sd=0.8
Residues with H-bonds	374 (80%)	346 (75%)
Dihedral Angles		
Mean Helix Phi	-66.8 sd=9.4	65.3 sd=11.9
Mean Helix Psi	-37.1 sd=19.1	-39.4 sd=25.5
Residues with Gauche +Chi	210(54%)	212 (55%)
Residues with Gauche-Chi	43(11%)	77 (20%)

Mean Chi Gauche+	133(34%)	96 (25%)
Mean Chi Gauche-	-65.9	-66.7 sd=15.0
Mean Chi Trans	63.5 sd=10.9	64.1 sd=15.7
Standard deviation of chi pooled	170.2	168.6
Mean Omega (|ω|>90)	10.83	15.70
Residues with (|ω|<90)	179.8 sd=1.4	180.0 sd=5.8
3D profile Quality index		
Resolution		-
R-value		-
Residues in phi-psi core	407(88%)	416 (90%)
Residues in phi-psi allowed	46(6%)	32(7%)
Residues in phi-psi generous	5 (0%)	5 (1%)
Residues in phi-psi outside	4(0%)	0 (0%)
Residues in omega core	461 (96%)	444(96%)
Residues in omega allowed	0(3%)	14 (3%)
Residues in omega generous	0 (0%)	0 (0%)
Residues in omega outside	1 (0%)	5 (1%)
Packing defects	109	32
Free energy of folding	-458.87	-441.36
Residues 95% buried	139	188
Buried charges	10	0

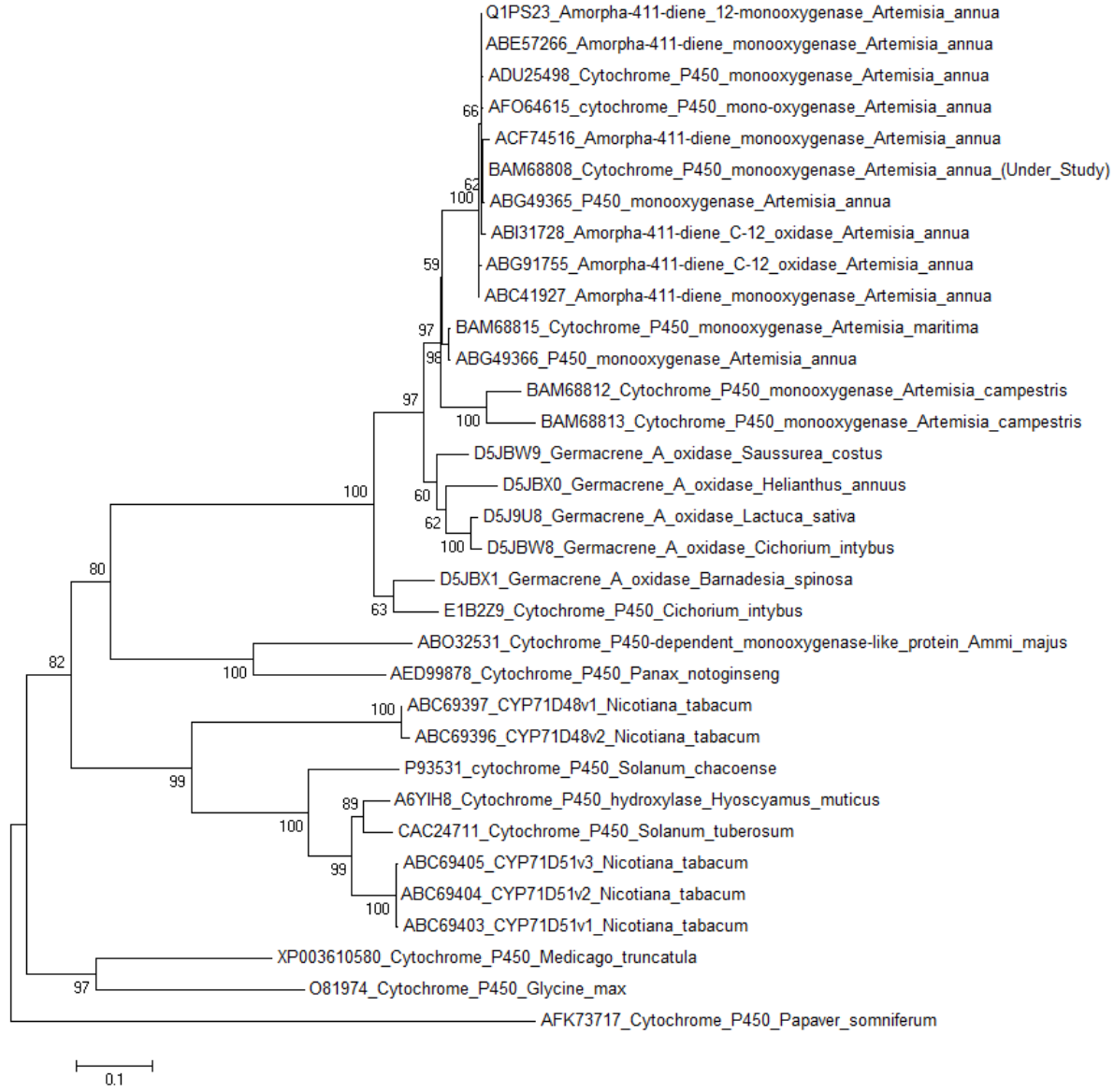


Fig. 1: Phylogenetic tree constructed by parsimony analyses of cytochrome P450 monooxygenase amino acid sequences selected from the BLAST profile. CLUSTAL-W from the Molecular Evolutionary Genetics Analysis program (MEGA5.10) was used for multiple alignments. Numbers on the branches are bootstrap values obtained for 1000 replicates are shown. GenBank accession numbers are shown first and cytochrome P450 amino acid sequences with their respective species is shown on the right side. The cytochrome P450 of *Papaver somniferum* is selected as an out-group to draw the rooted tree.

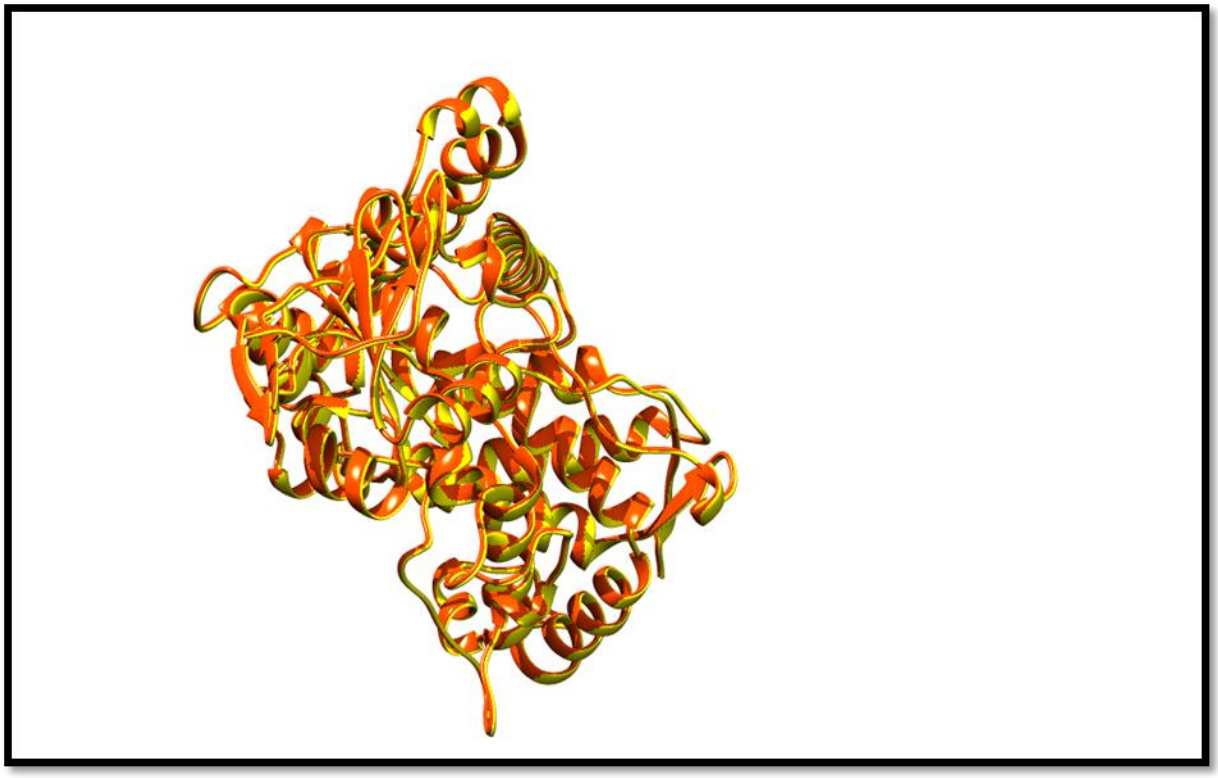


Fig. 2: Diagrammatic representation of 3D structure of model protein (CYP71) and Yellow colour represents template protein (1NR6) with RMSD 0.322 Å

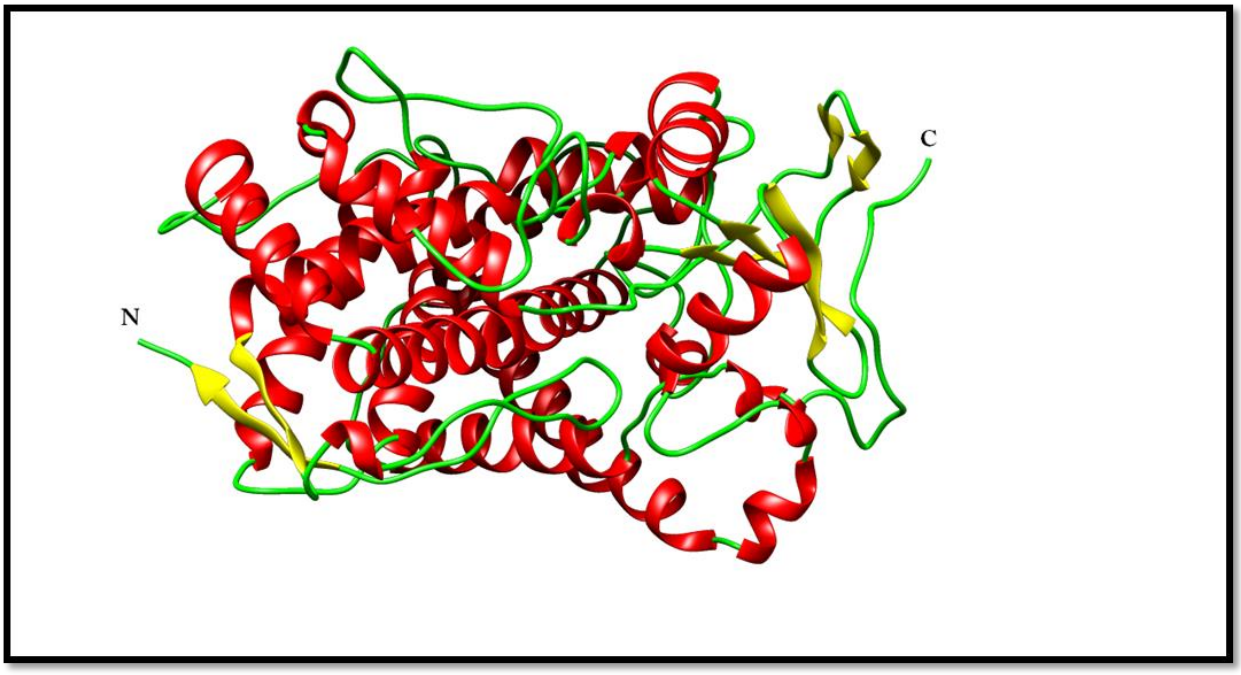


Fig. 3: Prediction of 3D model of Amorpha-4, 11-diene monooxigenase (CYP71)showing N and C terminal isolated from Artemisia annua L. plant.

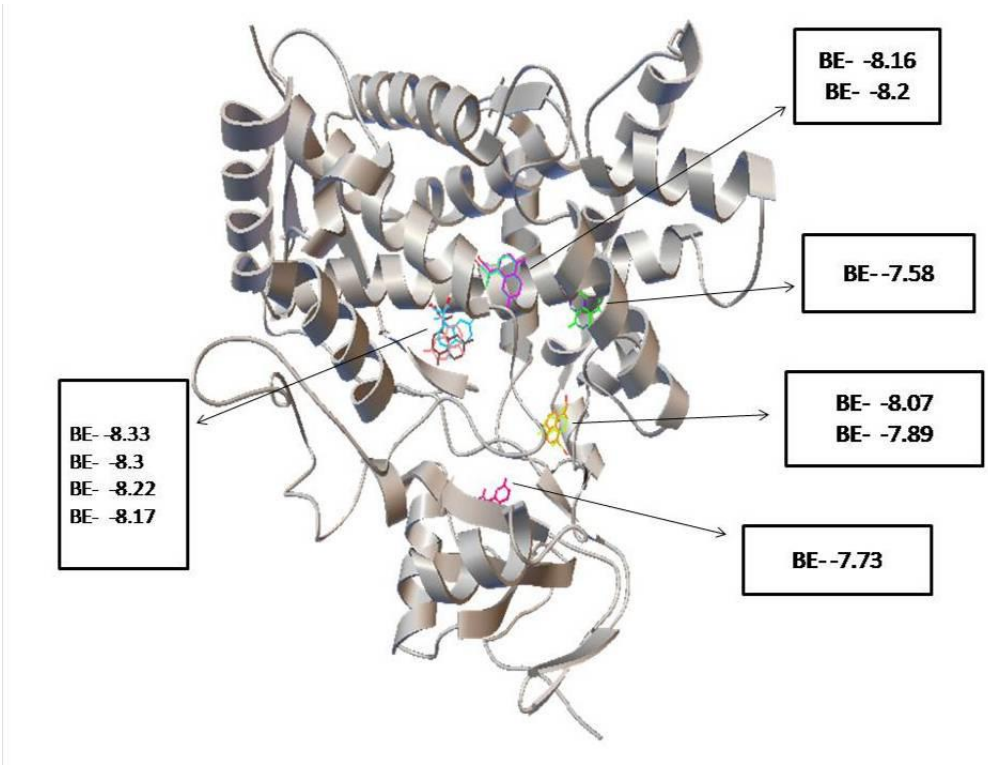
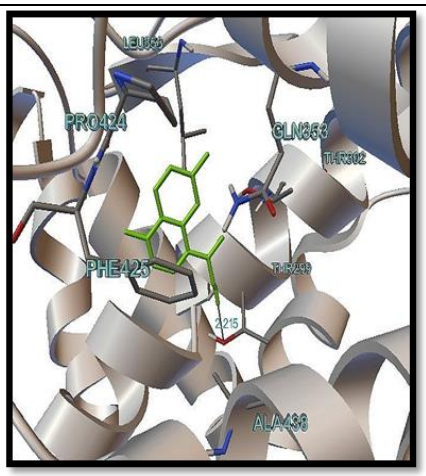
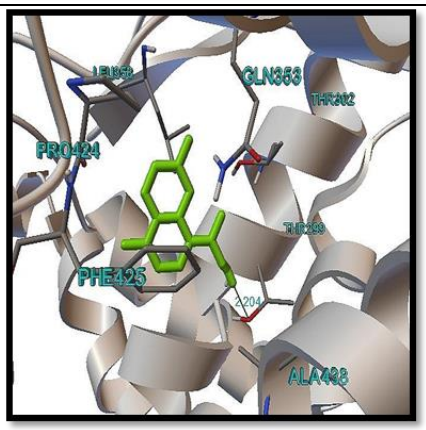
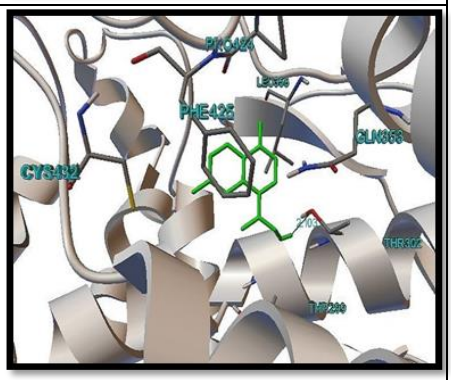


Fig. 4a: Interaction of artemisinic alcohol (AA) to the CYP71 protein with different binding sites having different binding energy.

Ligand	Confirmation	Active site residue	Binding energy	Hydrogen Bond (protein)	Docking
Artemisinic alcohol	1	TTQ LPFA	-8.33	THR299	
	2	TTQP FA	-8.3	THR299	
	3	TTQ LPFC	-8.22	THR302	

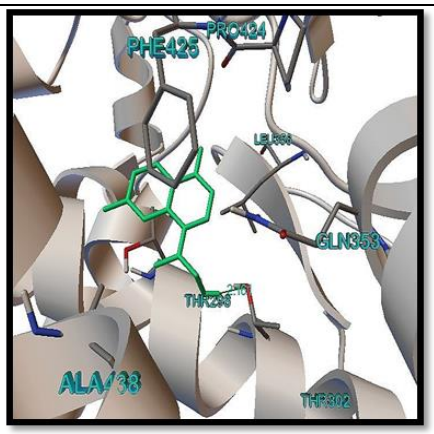
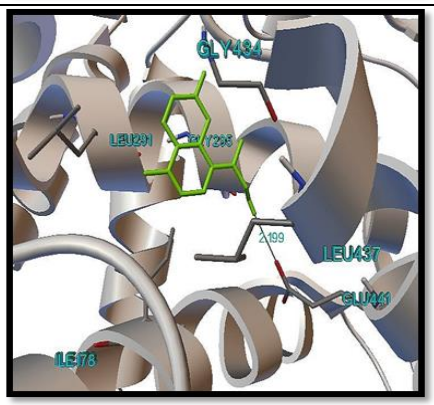
	4	**TTQ LPFA**	**-8.17**	**THR302**	
	5	**ILGC GLE**	**-8.2**	**GLU441**	

Fig. 4b: Receptor molecule (CYP71) with the ligand artemisinic alcohol (AA) docked in five top ranking binding sites. All the five figures in different colour correspond to five binding sites containing the ligand molecule AA docked into it.

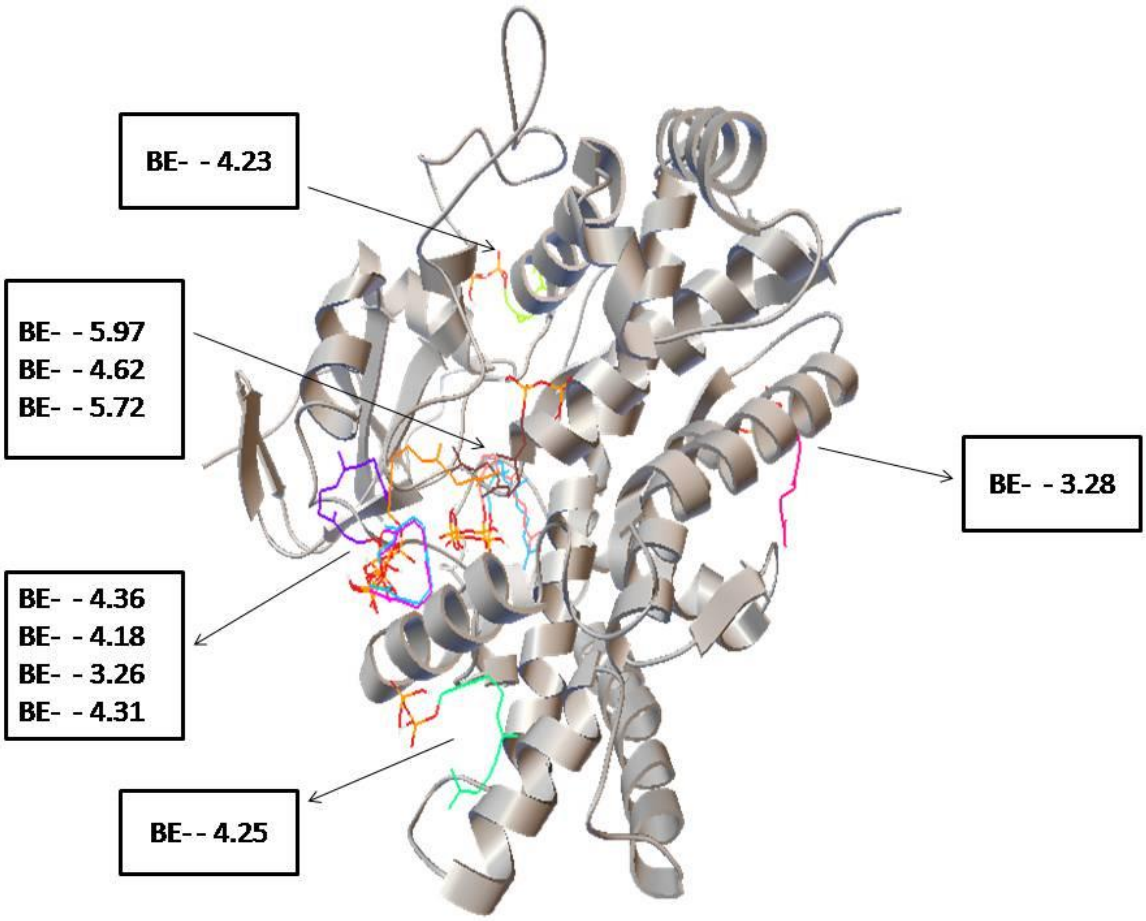
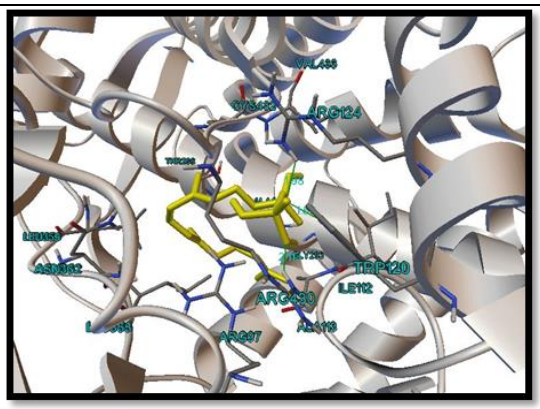
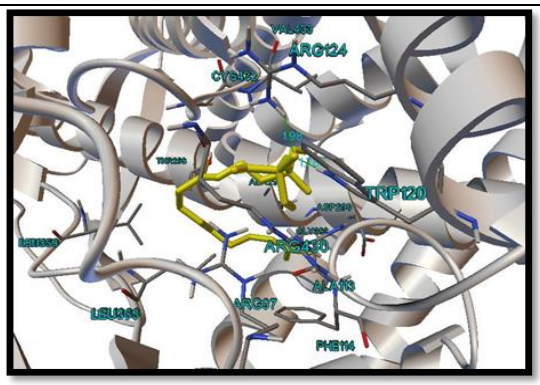
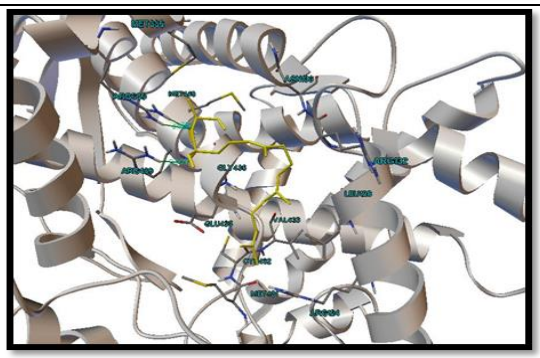
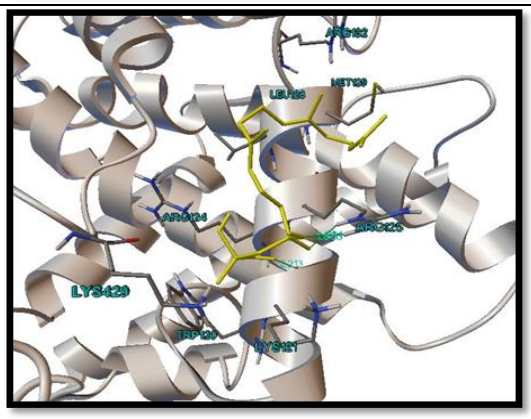


Fig. 5a: Interaction of FPP to the CYP71 protein with different binding sites having different binding energy.

Ligand	Confirmation	Active site Residue	Binding energy	Hydrogen Bond (protein)	Docking
FPP	1	RIAFWRDGATLNLRMCVG	-5.97	TRP120 ARG430 ARG124	
	2	RIAFWRDLGATLNLRMCV	-5.72	TRP120 ARG124	
	3	RLLRNMMCEGRM	-4.68	ARG339 ARG439 ARG339	
	4	WKRRLMRK	-4.36	LYS121 ARG125 ARG125	

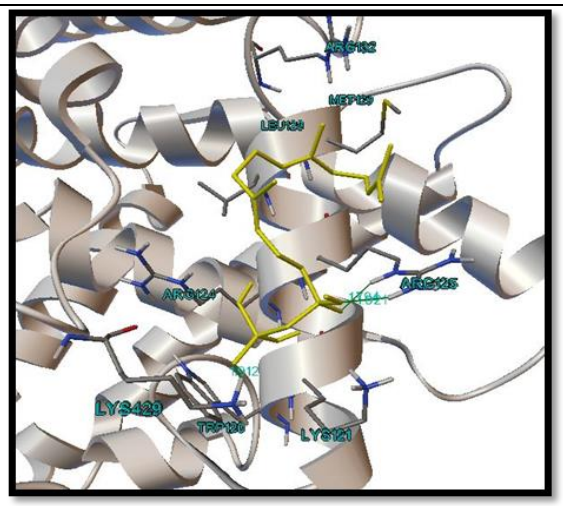
	5	**WKRR LMRK**	**-4.31**	**ARG125 LYS429 ARG125**	

Fig. 5b: Receptor molecule (CYP71) with the ligand FPP docked in five top ranking binding sites. All the five figures in different colour correspond to five binding sites containing the ligand molecule FPP docked into it.

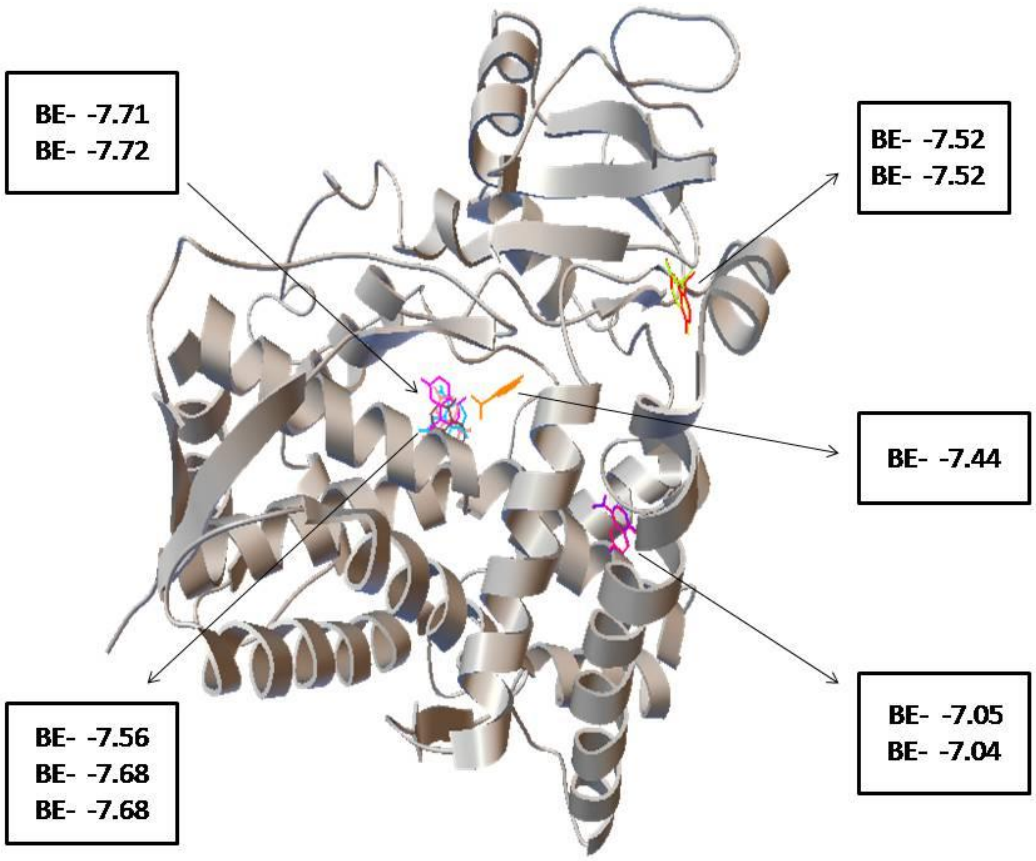
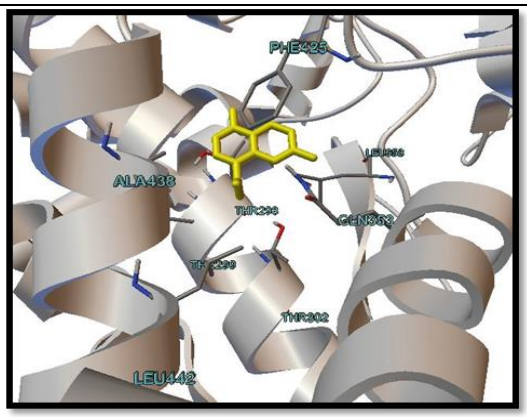
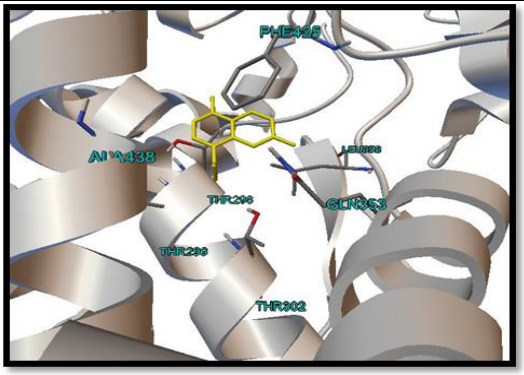
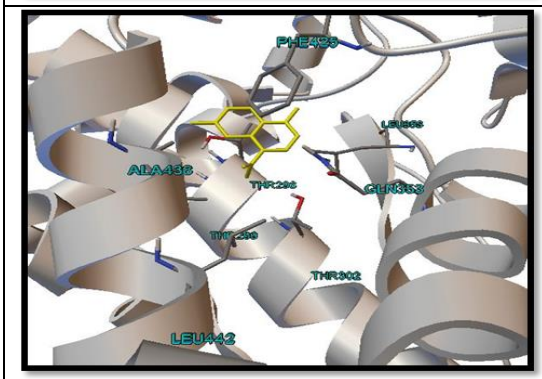
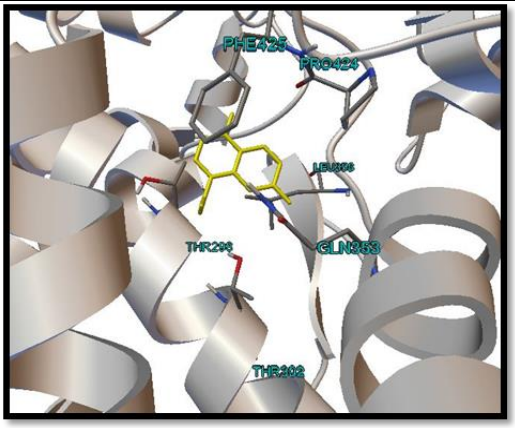


Fig. 6a: Interaction of amorpha- 4,11- diene to the CYP71 protein with different binding sites having different binding energy.

Ligand	Confirmation	Active site Residue	Binding Energy	Hydrogen Bond	Docking
Amorpha-4-11 Diene	1	TTTQLPFCAL	-7.72	No Hydrogen bond	
	2	TTTQLPFCAL	-7.71		
	3	TTTGLFAL	-7.68		
	4	TTTQLLPFA	-7.68		

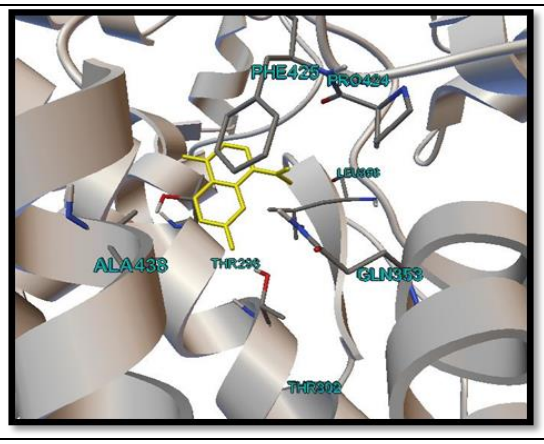
	5	**TTTQLPFAL**	**-7.56**		

Fig. 6b: Receptor molecule (CYP71) with the ligand Amorpha-4,11-diene. Docked in five top ranking binding sites. All the five figures in different colour correspond to five binding sites containing the ligand molecule AD docked into it.