In *silico* Analysis, Homology Modelling and Evolutinary Analysis of 5-enolpyruvyl shikimate 3-phosphate (EPSP) synthase enzyme from different plants

[1]Anupsingh Thakur, [2]Omkar Gogate, [3]Gaurav N. Chaudhari, [4]D. N Jagtap

[1,2,4] College of Agriculture Sangulwadi, Dr BSKKV Dapoli 416810
[3] Deptt. of Biotechnology, UAS, Dharwad-5

***Abstract*-** 5-enolpyruvylshikimate 3-phosphate (EPSP) synthase, an enzyme used in biosynthesis of aromatic amino acids in plants, many bacteria, and microbes, is a prime target for drugs and herbicides. The herbicide glyphosate (N-phosphonomethyl glycine) is a potent reversible inhibitor of the 5-enolpyruvylshikimate-3-phosphate (EPSP) synthase activity as it is competitive with respect to phosphoenolpyruvate and uncompetitive with respect to shikimate-3-phosphate. It is important to study this enzyme for elucidation of the active site of EPSP synthase and especially of the binding pattern of glyphosate provides a valuable roadmap for engineering new herbicides and herbicide-resistant crops, as well as new antibiotic and antiparasitic drugs. In this paper, a bioinformatics and molecular modeling approach was adopted to explore properties and structure of enzymes from different plants viz., *Nicotiana tabacum, Vitis vinifera, Amaranthus palmeri, Gossypium hirsutum, Brassica napus, Eleusine indica, Convolvulus arvensis* and *Capsicum annuum*. The properties of these proteins have been interpreted by Physico-chemical characterization including pI, EC, AI, GRAVY and instability index. Functional characterization was done by predicting motifs, patterns, disulfide bridges and secondary structure. Three dimensional structures for these proteins were not available as yet at PDB. Therefore, homology models for these enzymes were developed by using SWISS MODEL server. The model was analyzed for its Fold reliability by using server ProSA, ERRAT server was used for analyzes the statistics of non-bonded interactions between different atom types. The model was validated using protein structure checking tool WHAT IF. These structures will provide a good foundation for functional analysis of experimentally derived crystal structures. Phylogenic analysis was done by using MEGA 4.0. Two major sequence clusters were constructed by phylogenetic analysis showing *Nicotiana tabacum*, *Solanum lycopersicum* and *Capsicum annuum* showed in single cluster with other studied plants.

I. INTRODUCTION

The enzyme 5-enolpyruvylshikimate 3-phosphate (EPSP) synthase (EC 2.5.1.19) is an enzyme involved the shikimate pathway and is essential for the synthesis of aromatic amino acids. This pathway is found in algae, higher plants, bacteria, and fungi (1–3), as well as in apicomplexan parasites (4) but absent from mammals (2, 3). New drugs, herbicides and antimicrobial agents effective against Weeds, bacterial, parasitical and fungal pathogens had been designed against EPSP synthase. Glyphosate (N-phosphonomethyl glycine, 'Roundup') is a successful, broad-spectrum, postemergence herbicide, which has proven as potent and specific inhibitor of EPSP synthase (5). Glyphosate is successfully used as a herbicide, being the active ingredient of the widely used weed control agent Roundup, and was recently shown to inhibit the growth of the pathogenic parasites *Plasmodium falciparum* (malaria), *Toxoplasma gondii*, and *Cryptosporidium parvum* (4).

EPSP synthase catalyzes the transfer of the enolpyruvyl moiety from phosphoenol pyruvate (PEP) to shikimate-3-phosphate forming the products EPSP and inorganic phosphate (1, 6). The reaction is chemically unusual because it proceeds via C–O bond cleavage of PEP rather than via P–O bond cleavage (7) as in most PEP-utilizing enzymes. Glyphosate inhibits EPSP synthase in a slowly reversible reaction, which is competitive versus PEP and uncompetitive versus shikimate-3-phosphate (5, 8, 9).

Although EPSP synthase has been extensively studied over more than three decades (10), conclusions on the enzyme mechanism (8, 9, 10, 11) and especially on the mode of action of the herbicide glyphosate (9, 10, 11) remained controversial. *In silico* approach reveals the two diamentional and three diamentioal structural properties by Homology modeling of EPSP synthase from different plants viz., *Nicotiana tabacum, Vitis vinifera, Amaranthus palmeri, Gossypium hirsutum, Brassica napus, Eleusine indica, Convolvulus arvensis, Capsicum annuum* and further studied for elucidation of the active site of EPSP synthase and especially of the binding pattern of glyphosate by Docking study provides a valuable roadmap for engineering new herbicides and herbicide-resistant crops, as well as new antibiotic and antiparasitic drugs.

II. MATERIALS AND METHODS

Sequences of antioxidant proteins of spinach were retrieved from the NCBI (http://www.ncbi.nlm.nih.gov/), a public domain protein database **(12)** as shown in table 1. The EPSP synthase enzyme sequences were retrieved in FASTA format and used for further analysis.

For physico-chemical characterization, theoretical isoelectric point (pI), molecular weight, total number of positive and negative residues, extinction coefficient **(13)**, instability index **(14)**, aliphatic index **(15)** and grand average hydropathy (GRAVY) **(16)** were computed using the Expasy's ProtParam

server **(17)** (http://us.expasy.org/tools/protparam.html). The results were shown in Table 1.

SOPMA **(18)** was employed for calculating the secondary structural features of the EPSP synthase enzyme sequences considered for this study. The results were presented in Table 2.

The modeling of the three dimensional structure of the protein was performed by homology modeling server, Swissmodel (http://swissmodel.expasy.org/) **(19)**. The overall stereochemical property of the protein was assessed by Ramchandran plot analysis **(20)**. The validation for structure models obtained from the three software tools was performed by using WHAT IF **(21)**, ProSA **(22)** and ERRAT. The results of WHAT IF, ProSA (https://prosa.services.came.sbg.ac.at/prosa.php) and ERRAT (http://nihserver.mbi.ucla.edu/ERRATv2/) analysis were shown in figure 1. Structural analysis was performed and figures representations were generated with Swiss PDB Viewer **(23)**.

Phylogeneic analysis for EPSP synthase from different plants was done by MEGA 4 tool. Results were shown in Figure 2. There is single cluster to all proteins revealed that the functional motif are conserved in evolution.

III. RESULTS AND DISCUSSION

Table 1 shows accession numbers of EPSP synthase enzyme from different plants considered in this study and retrived in FASTA format for further analysis. The parameter pI, EC, AI, GRAVY and instability index, which was computed using Expasy's ProtParam tool had been shown in table 1. Isoelectric point (pI) value was calculated because at that value protein is stable and compact. pI value is the pH at which protein has no net charge. The computed pI value of EPSP synthase enzyme from plants *Brassica napus* and *Eleusine indica* (CAA35839.1, CAD01096.1, AAR87845.1) were less than 7 (pI<7) indicates that these antioxidant proteins were considered as acidic whereas for *Convolvulus arvensis* it is 7.05 showing the neutral. The EPSP synthase enzyme plants *Nicotiana tabacum, Vitis vinifera, Amaranthus palmeri* and *Gossypium hirsutum* showed pI value more than 7.0. The computed isoelctric point (pI) will be useful for developing buffer system for purification by isoelectric focusing method. Extinction coefficient computed for the wavelength 276, 278, 279, 280 and 282 nm of which 280 nm is favored because proteins (Cys, Trp and Tyr amino acids) absorb light strongly. The high EC of enzyme from plants *Nicotiana tabacum* and *Brassica napus* indicates presence of high concentration of Cys, Trp and Tyr. The computed extinction coefficients help in the quantitative study of protein–protein and protein–ligand interactions in solution. The instability index provides an estimate of the stability of protein in a test tube. A protein whose instability index is smaller than 40 is predicted as stable, a value above 40 predicts that the protein may be unstable **(14)**. The instability index value for the spinach antioxidant proteins were found to be ranging from 29.2 to 36.59. The results classified as enzyme of all plants were stable protein (Table 1).
The aliphatic index (AI) is defined as the relative volume of a protein occupied by aliphatic side chains. It is a positive factor for the increase of thermal stability of globular proteins. Aliphatic index for the antioxidant protein sequences ranged from 88.87 to 95.53. The very high aliphatic index of all enzymes sequences indicates that these enzymes may be stable for a wide temperature range whereas the lower thermal stability of enzyme from plant *Gossypium hirsutum* was indicative of a more flexible structure when compared to other antioxidant protein.

The secondary structure of EPSP synthase enzyme from different plants were predicted by SOPMA (Self Optimized Prediction Method with Alignment) with default parameter which correctly predicts 69.5% of amino acids for a state description of the secondary structure prediction **(18)**. Results are represented in table 3 with parameters Alpha helix Pi helix Beta bridge Extended strand Beta turn Bend region Random coil Ambigous states.

The modeling of the three dimensional structure of the protein was performed by three homology modeling server, Swiss Model. SWISS-MODEL is a fully automated protein structure homology-modeling server, accessible via the ExPASy web server, or from the program DeepView (Swiss Pdb-Viewer). The purpose of this server is to make Protein Modelling accessible to all biochemists and molecular biologists worldwide.

The model was analyzed for its Fold reliability by using server ProSA server. The recognition of errors in experimental and theoretical models of protein structures is a major problem in structural biology. ProSA calculates an overall quality score for a specific input structure. If this score is outside a range characteristic for native proteins the structure probably contains errors. A plot of local quality scores points to problematic parts of the model which are also highlighted in a 3D molecule viewer to facilitate their detection. The z-score indicates overall model quality (As shown in analysis). Its value is displayed in a plot that contains the z-scores of all experimentally determined protein chains in current PDB. Plot of residue scores groups of structures from different sources (X-ray, NMR) are distinguished by different color. It can be used to check whether the z-score of the input structure is within the range of scores typically found for native proteins of similar size. Results are shown in figure 1.
ERRAT server was used for analyzes the statistics of non-bonded interactions between different atom types. ERRAT is a protein structure verification algorithm that is especially well-suited for evaluating the progress of crystallographic model building and refinement. The program works by analyzing the statistics of non-bonded interactions between different atom types.

The modeled structures of EPSP synthase enzymes were also validated by other structure verification servers WHAT IF **(21)**. Standard bond angles of the four models are determined using WHAT IF. The results were shown in Table 3. The analysis revealed RMS Z-scores were almost equal to 1 suggesting high model quality. The predicted structures conformed well to the stereochemistry indicating reasonably good quality.

Phylogenic analysis was done by using MEGA 4.0. Two major sequence clusters were constructed by phylogenetic analysis showing *Nicotiana tabacum*, and *Capsicum annuum* showed in single cluster with other studied plants.

IV. CONCLUSION

The modeling of the three dimensional structure of the protein was performed by three homology modeling server, Swiss Model. Model of EPSP synthase enzyme from all plants

show reliable folding, validated by ERRAT and WHAT IF servers. These structures will provide a good foundation for functional analysis of experimentally derived crystal structures.

REFERENCES

[1] Bentley, R. (1990) Crit. Rev. Biochem. Mol. Biol. 25, 307–383.

[2] Haslam, E. (1993) Shicimic Acid: Metabolism and Metabolites (Wiley, Chichester, U.K.).

[3] Kishore, G. M. & Shah, D. M. (1988) Annu. Rev. Biochem. 57, 627–663.

[4] Roberts F., Roberts, C. W., Johnson, J. J., Kyle, D. E., Krell, T., Coggins, J. R., Coombs, G. H., Milhous, W. K., Tzipori, S., Ferguson, D. J. P., et al. (1998) Nature (London) 393, 801–805.

[5] Steinru¨cken, H. C. & Amrhein, N. (1980) Biochem. Biophys. Res. Commun. 94, 1207–1212.

[6] Levin, J. G. & Sprinson, D. B. (1964) J. Biol. Chem. 239, 1142–1150.

[7] Walsh, C. T., Benson, T. E., Kim, D. H. & Lees, W. J. (1996) Chem. Biol. 3, 83–91.

[8] Boocock, M. R. & Coggins, J. R. (1983) FEBS Lett. 154, 127–133.

[9] Steinru¨cken, H. C. & Amrhein, N. (1984) Eur. J. Biochem. 143, 351–357.

[10] American Chemical Society (1997). Glyphosate: A Unique Global Herbicide (American Chemical Society Monograph, Washington, D.C.).

[11] Sikorski, J. A. & Gruys, K. J. (1997) Acc. Chem. Res. 30, 2–8.

[12] Bairoch A.and Apweiler R. (2000) "The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000" Nucleic Acids Res 28: 45-48.

[13] Gill S. C and Von Hippel P. H. (1989) "Extinction coefficient" Anal Biochem 182: 319- 328.

[14] Guruprasad K., Reddy B. V. P. and Pandit M. W. (1990) "Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence" Prot Eng 4: 155-164.

[15] Ikai A. J. (1980) "Thermo stability and aliphatic index of globular proteins" J Biochem 88: 1895-1898

[16] Kyte J. and Doolottle R. F. (1982) "A simple method for displaying the hydropathic character of a protein" J Mol Biol 157: 105- 132

[17] Gasteiger E (2005) "Protein Identifi cation and Analysis Tools on the ExPASy Server" In: John M. Walker ed, The Proteomics Protocols Handbook, Humana Press 571-607

[18] Geourjon C. and Deléage G. (1995) "SOPMA: signifi cant improvements in protein secondary structure prediction by consensus prediction from multiple alignments" Comput Appl Biosci 11: 681-684.

[19] Arnold K., Bordoli L., Kopp J. and Schwede T. (2006) "The SWISS-MODEL workspace: a web-based environment for protein structure homology modeling" Bioinformatics 22: 195-201

[20] Ramachandran G. N., Ramakrishnan C. and Sasisekhran V. (1963) "Stereochemistry of polypeptide chain configuarations" J Mol Biol 7: 95-99.

[21] Vriend G. (1990) "WHAT IF: a molecular modelling and drug design program" J. Mol. Graph. 8, 52--56.

[22] Wiederstein M. and Sippl M. J. (2007) "ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins" Nucleic Acids Res 2007;35:W407-410.

[23] Guex N. and Manuel C. P. (1997) "Data Modeling, Analysis and Classifi cation SWISSMODEL and the Swiss-Pdb Viewer: An environment for comparative protein modeling" Electrophoresis 18: 2714-2723.

AUTHORS

First Author – Anupsingh Thakur, M.Sc Molecular Biology and biotechnology, Assistant professor, College of Agriculture Sangulwadi, Dr BSKKV Dapoli 416810 (anupthakur34@yahoo.com).

Second Author – Omkar Gogate, M.Sc Bio-Analytical Techniques, Assistant professor, College of Agriculture Sangulwadi, Dr BSKKV Dapoli 416810 (omkar.gogate@yahoo.in).

Third Author – Gaurav N. Chaudhari, M.Sc Agriculture Biotechnology, Deptt. of Biotechnology, UAS, Dharwad-5 (chaudharigaurav11@gmail.com).

Correspondence Author – Dr. D. N. Jagtap, Principal, of Agriculture Sangulwadi, Dr BSKKV Dapoli 416810 (mauli296@gmail.com).

	Accession No.	Sequence length	Molecular wt	pI	-R	+R	EC	II	AI	GRAVY
Nicotiana tabacum	AAA34071.1	518	55711.2	8.45	54	58	40380	31.62	92.37	-0.042
Vitisv inifera	ACY29662.2	521	55486.7	7.58	50	51	31900	34.61	89.98	-0.017
Amaranthus palmeri	ACV53022.1	518	55175.5	7.53	55	56	36370	36.59	91.08	-0.027
Gossypium hirsutum	ACF16410.1	521	55597.1	8.56	52	57	38890	27.7	88.87	-0.041
Brassica napus	CAA35839.1	516	55030.2	6.63	55	54	39015	33.33	91.63	-0.009
Eleusine indica	CAD01096.1	445	47397.7	5.52	54	48	33265	29.2	94.65	0.08
	AAR87845.1	445	47414.7	5.41	54	47	34775	29.21	95.53	0.09
Convolvulus arvensis	ACD80082.1	520	55498.9	7.05	56	56	34755	36.34	92.25	-0.037
Capsicum annuum	AEK22121.1	516	55377.7	7.54	56	57	33390	32.06	92.56	-0.047

Table 1 Protein sequences considered for the study and Parameters computed using Expasy'sProtParam tool And Transmembrane regions identified by SOSUI

	Nicotiana tabacum	*Vitis vinifera*	*Amaranthus palmeri*	*Gossypium hirsutum*	*Brassica napus*	*Eleusine indica*		*Convolvulus arvensis*	*Capsicum annuum*
	AAA34071.1	ACY29662.2	ACV53022.1	ACF16410.1	CAA35839.1	CAD01096.1	AAR87845.1	ACD80082.1	AEK22121.1
Alpha helix	31.47	28.79	33.2	30.13	31.78	33.48	32.36	30.77	34.88
310 helix	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Pi helix	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Beta bridge	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Extended strand	17.76	18.62	16.02	18.04	18.02	16.63	17.98	15.96	16.67
Beta turn	6.56	5.57	3.86	6.53	4.46	6.07	6.52	5.38	6.2
Bend region	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Random coil	44.21	47.02	46.91	45.3	45.75	43.82	43.15	47.88	42.25
Ambigous states	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

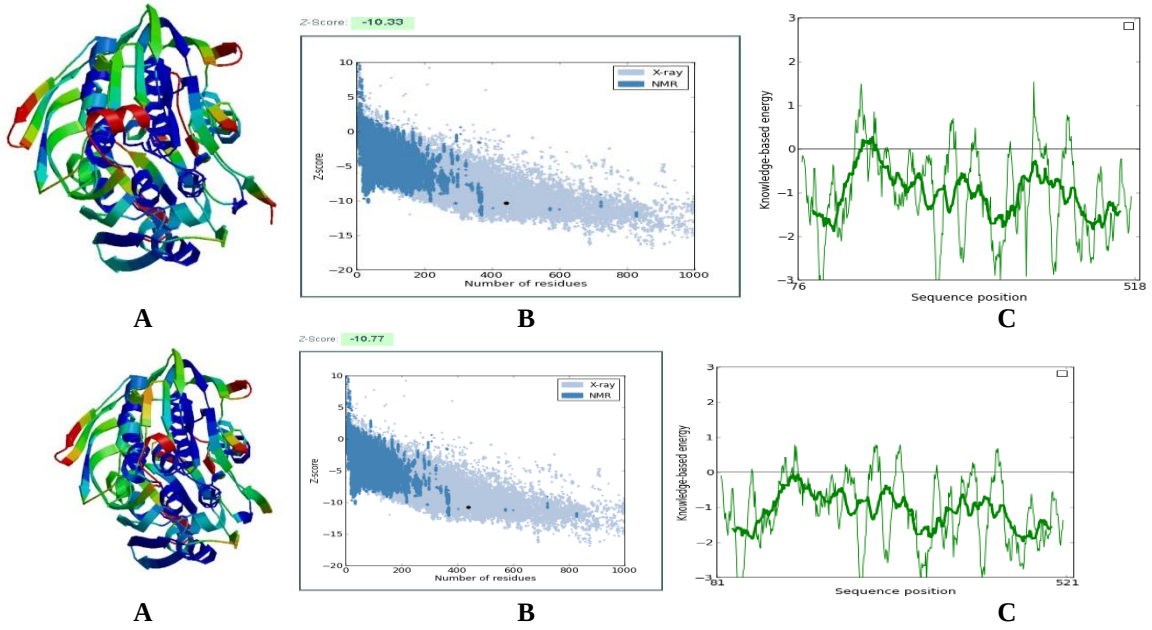
Table 2.Calculated secondary structure elements by SOPMA

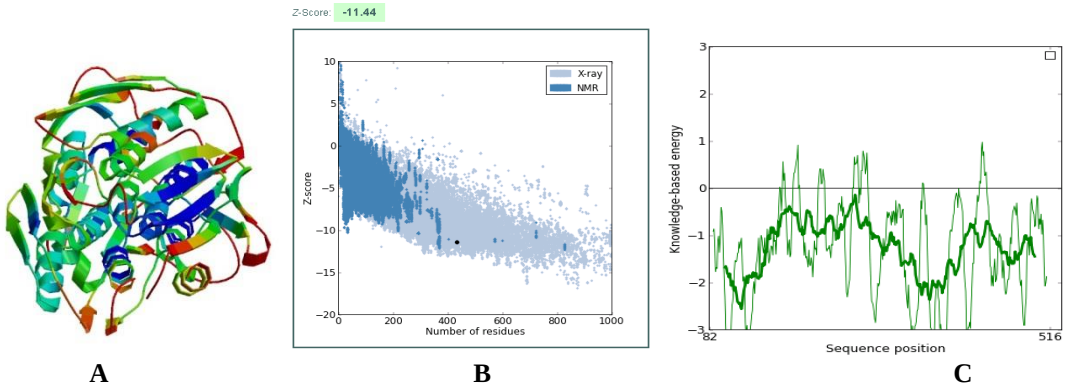
	Nicotiana tabacum	*Vitis vinifera*	*Amaranthus palmeri*	*Gossypium hirsutum*	*Brassica napus*	*Eleusine indica*		*Convolvulus arvensis*	*Capsicum annuum*
	AAA34071.1	ACY29662.2	ACV53022.1	ACF16410.1	CAA35839.1	CAD01096.1	AAR87845.1	ACD80082.1	AEK22121.1
Structure Z-scores, positive is better than average									
1st generation packing quality	-0.929	-0.955	-1.387	-0.835	-0.624	-0.746	-0.734	-0.874	-0.797
2nd generation packing quality	-2.093	-1.892	-2.87	-2.332	-1.953	-1.771	-1.664	-2.01	-2.064

Ramachandran plot appearance	-1.089	-0.895	-1.078	-0.886	-0.79	-0.931	-0.906	-1.027	-0.952
chi-1/chi-2 rotamer normality	0.961	1.321	0.073	1.104	1.163	1.291	1.227	1.376	0.852
Backbone conformation	-1.676	-2.126	-2.239	-1.76	-1.661	-2.188	-2.056	-1.937	-2.024
RMS Z-scores, should be close to 1.0:									
Bond lengths	0.652T	0.67	0.612	0.612T	0.616T	0.661T	0.641T	0.678	0.638T
Bond angles	1.086	1.14	1.355	1.118	1.127	1.172	1.166	1.101	1.088
Omega angle restraints	1.041	0.964	1.072	1.088	0.881	1.007	1.031	0.875	0.97
Side chain planarity	1.823	2.344L	2.191L	1.857	1.819	1.906	1.935	1.579	2.009L
Improper dihedral distribution	1.476	1.548L	1.763L	1.582L	1.516L	1.493	1.493	1.513L	1.516L
Inside/Outside distribution	0.976	0.972	1.027	0.976	0.975	0.993	0.993	0.958	0.975

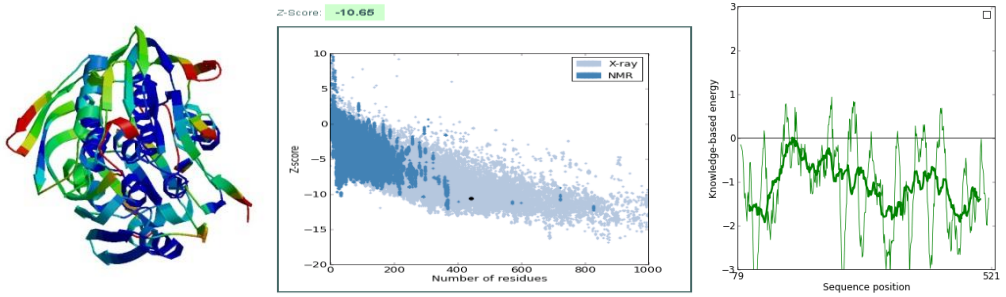
Table 3 What if analysis

Figure 1.Modeled Structure (SWISS MODE) andProSA analysis of Antioxidant proteins

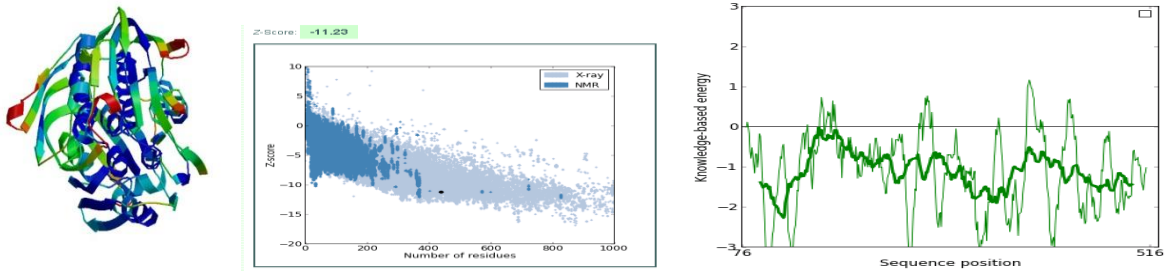




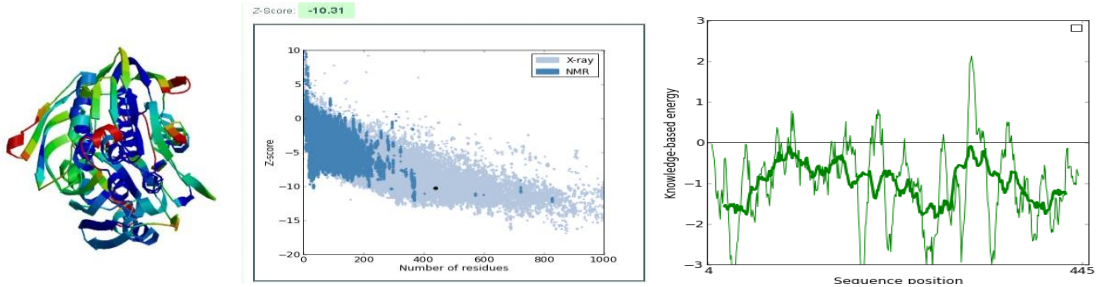
A **B** **C**



A **B** **C**



A **B** **C**



A **B** **C**

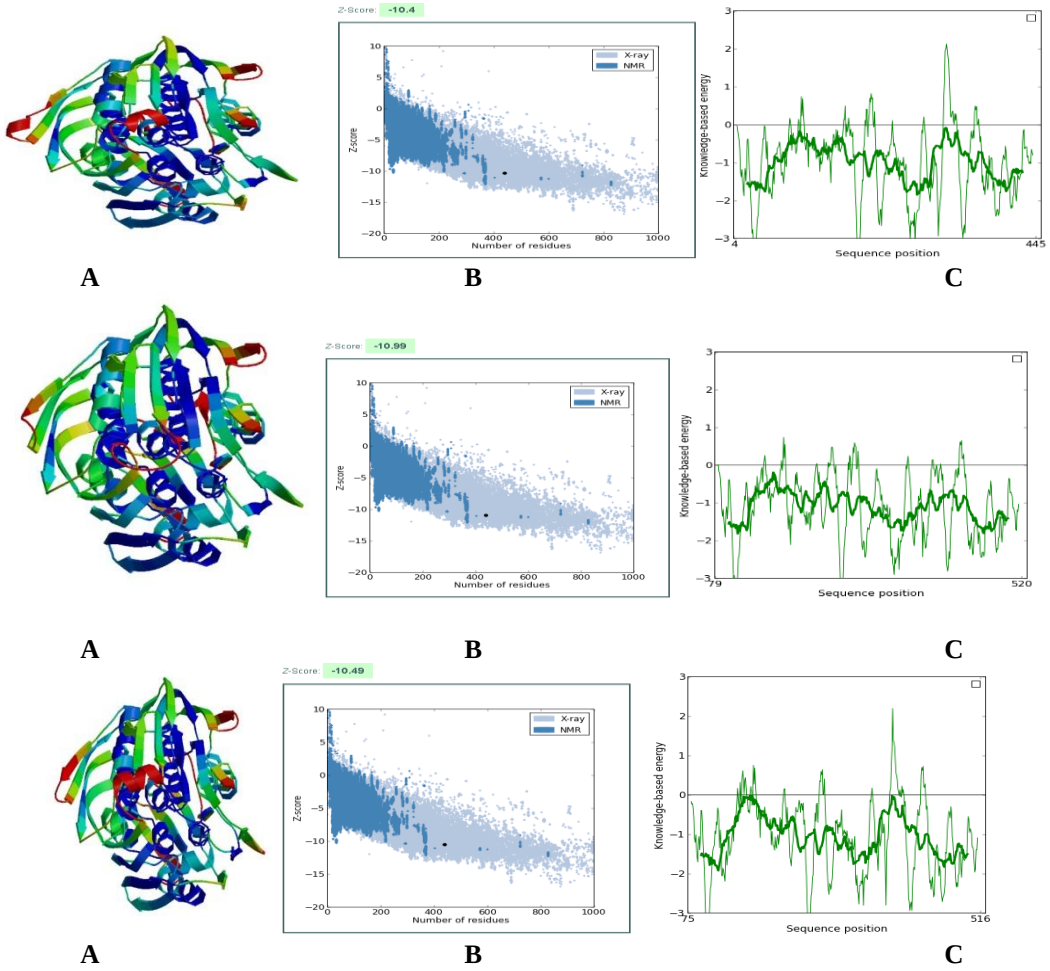

Z-Score: -10.4
X-ray
NMR
Z-score
Number of residues
Knowledge-based energy
Sequence position
4
445
A
B
C
Z-Score: -10.99
X-ray
NMR
Z-score
Number of residues
Knowledge-based energy
Sequence position
79
520
A
B
C
Z-Score: -10.49
X-ray
NMR
Z-score
Number of residues
Knowledge-based energy
Sequence position
75
516
A
B
C

Figure 2 Phylogenic analysis