In-silico Characterization, Structural Modelling, Docking Studies and Phylogenetic Analysis of 5-Enolpyruvylshikimate-3-Phosphate Synthase Gene of Oryza sativa L.

Ubaid Yaqoob1*, Tanushri Kaul2, Saurabh Pandey1 and Irshad Ahmad Nawchoo1

1Plant Reproductive Biology, Genetic Diversity and Phytochemistry Research Laboratory, Department of Botany, University of Kashmir, Srinagar, Jammu and Kashmir, India
2Plant Molecular Biology Lab, International Centre for Genetic Engineering and Biotechnology, New Delhi, India

Abstract

The 5-enolpyruvylshikimate-3-phosphate synthase (EPSPS), one of the key enzymes of the shikimate pathway which is involved in the biosynthesis of secondary metabolites and several amino acids. The multiple sequence alignment of these EPSPS protein sequences from different plants showed conserved regions at different stretches with maximum homology in amino acid residues. We revealed the homology model of Oryza sativa EPSPS (OsEPSPS) protein using the structure of E. coli EPSPS as template. The resulting model structure was refined by PROCHECK, RAMPAGE server, ProSA, Verify3D etc. that indicated the model structure is reliable. Ramachandran plot analysis showed that conformations for 94.3% of amino acid residues are within the most favoured regions. Through motif analysis, it was revealed that a conserved EPSPS domain is uniformly found in all EPSPS proteins irrespective of variable plant species suggesting its possible role in cellular and metabolic functions. The phylogenetic tree constructed revealed different clusters based on EPSPS in respect of bacteria, monocot and dicot plants. The interacting partners of the gene shows the importance of this gene family in regulating developmental and metabolic functions. The two conserved motifs LP(G/S)KSLN[R/L]LLAAL and LFLGNAGTAMRPL present in almost all EPSPS plant species may function as the catalytic domains of EPSPS enzymes and are supposed to contribute in the glyphosate binding site.

Keywords: EPSP synthase; Glyphosate; Herbicide; Oryza sativa; Shikimate pathway

Introduction

The 5-enolpyruvylshikimate-3-phosphate synthase (EPSPS), one of the key enzymes of the shikimate pathway is involved in the biosynthesis of several aromatic amino acids (Phenylalanine (Phe), Tyrosine (Tyr) and Tryptophan (Trp)) and other secondary products (auxin, salicylate, folic acid, phytoalexins, flavonoids, alkaloids etc.) essential for plant survival [1]. It is also verified as a specific target of broad spectrum herbicide glyphosate (N-phosphonomethyl glycine) [2]. EPSPS (aroA) plays a central role in catalysing the transfer of enolpyruvyl moiety from phosphoenol pyruvate (PEP) to shikimate-3-phosphate (S3P) forming EPSP and inorganic phosphate [3]. The reaction is chemically infrequent because it proceeds via C–O bond cleavage of phosphoenol pyruvate rather than via P–O bond cleavage [4]. Glyphosate (GPI) inhibits EPSPS in a slowly reversible reaction, which is competitive with respect to PEP and uncompetitive with respect to S3P [5,6]. In most of the crops and weeds, glyphosate can starve the plants of aromatic amino acids by competitively inhibiting the binding of EPSPS with PEP. Mutagenesis of EPSPS was done in various species so as to obtain glyphosate-tolerant EPSPS like proline-106 to serine in E. indica [7], proline-106 to leucine in N. tabacum [1], glycine-100 to alanine in agrobacterium sp. strain CP4 [8], proline-101 to serine in N. tabacum [9]. The occurrence of shikimate pathway in algae, bacteria, fungi and plants makes EPSPS a principal target for rising herbicide-resistant genetically modified crops [10]. Thus understanding its mechanism for regulating metabolic and developmental processes in diverse plant species would be a great revolution for engineering new herbicides, developing glyphosate resistant crops, new antibiotic and anti-parasitic drugs.

Methodology

Comparative modelling and structural analysis

The reference sequence of EPSPS from Oryza sativa was retrieved by using NCBI database (http://www.ncbi.nlm.nih.gov). By searching the PDB of known protein structures, the comparative modelling was performed with target sequence as the query [11]. The target sequence was searched for similar sequence using the BLAST (Basic Local Alignment Search Tool) [12] against Protein Database (PDB) (http://www.rcsb.org). The best template for query sequence was recognized based on the e-value, % sequence identity and % sequence coverage. The BLAST results yielded X-ray structure of EPSPS from E. coli with 53% similarity to our target protein (OsEPSPS). Using ClustalW [13], all the sequences of EPSPS were aligned to find out the similarity present among the sequences. 2D and 3D structure alignment was carried out using ClustalW [14] and MATRAS 1.2 [15], respectively. The sequences of the EPSPS were further analysed for the presence of specific EPSPS domains and motifs through motifscan (myhits.isb-sib.ch/cgi-bin/motif scan) and scan prosite (Prosite.expasy.nlm.nih.gov). Analysis of conserved motifs was done by MEME version 3.5.7.
The secondary structural features of the EPSPS sequences of template and target were calculated using SOPMA. The physico-chemical properties of EPSPS sequences like molecular weight, theoretical isoelectric point (pI), number of amino acids, total number of positive and negative residues, aliphatic index [17], grand average hydropathy (GRAVY) [18] extinction coefficient [19] and instability index [20] were evaluated by using Expasy’s ProtParam server (http://us.expasy.org/tools/protparam.html) [21]. The sub-cellular localizations were predicted by using CELLO v.2.5 [22]. Using NetNGlyc 1.0 server (http://www.cbs.dtu.dk/services/NetNGlyc/), the N-glycosylation sites of the EPSPS proteins were predicted. Using String software (http://string-db.org/) the interacting partners of EPSPS and its co-expressed genes were predicted [23].

Model validation of OsEPSPS

On the basis of geometrical and stereo-chemical constraints, the model was evaluated using RAMPAGE server (http://mordred.bioc.cam.ac.uk/~rapper/rampage.php), PROCHECK [24], Verify 3D [25] and ProSA-Web [26]. The model with the least number of residues in the disallowed region was selected for the further studies. The RMSD value between the template and target was calculated using MOE [27]. The best model structure was then compared with the template protein by superimposition using SuperPose Version 1.0 [28].

Active site prediction and molecular docking

Active sites of model and template proteins were identified using different binding site prediction servers like Q-site finder (http://bmmbpcu36.leeds.ac.uk/qsitefinder/), CASTp (http://sts-fw.bioengr.uiuc.edu/castp/) and PINUP server (http://sparks.informatics.iupui.edu/PINUP/) [29-31]. The refined protein model (OsEPSPS) was used to study its ligand binding mechanism. Docking analysis was performed by Sybyl 8.0 molecular modelling tool to identify active sites on protein structure where favourable protein-ligand interactions can occur [32]. The ligand molecules (S3P and GPJ) were docked inside the cavity of OsEPSPS protein.

Phylogenetic analysis

Using Molecular Evolutionary Genetic Analysis (MEGA) software Version 4.1 [33], phylogenetic analysis of the sequences was carried out by using UPGMA method. Each node was tested using the bootstrap approach by taking 5,000 replicates.

Results and Discussion

Comparative modelling and structural analysis

The *Oryza sativa* EPSPS (OsEPSPS) protein sequence comprises of 515 amino acid residues. Sequences that showed maximum identity with high score and low e-value were aligned. According to the result of BLAST search against PDB [34], three reference proteins (PDB ID: 3NV3, 1G6S, 3FIX) represented a high level of sequence identity - 54%, 53% and 53% respectively. The *E. coli* template (PDB ID: 1G6S) with an e-value of 2e-149 and a query cover of 84% was selected for homology modelling. Structurally conserved regions (SCRs) between model OsEPSPS (target) and homologous proteins (PDB: 1G6S, 3NV3, 3FIX) were determined by multiple sequence alignment (Figure 1).

Multiple sequence alignment of the EPSPS sequences highlighted the sequence conservation of amino acid residues among different species (Supplementary File 1). Structurally conserved regions (SCRs) between model OsEPSPS and template (PDB: 1G6S) were also determined (Figure 2). An extensive search of the motifs and their positions was done by MEME software which identified several conserved motifs in the protein sequences of EPSPS (Figure 3). Multilevel consensus sequences for the MEME defined motifs along with their functions are shown in Table 1. LP(G/S)KSLSNRLLLAAAL and LFLGNAGTAMRPL motifs were present in almost all selected species.

The initial model of OsEPSPS was built by homology modelling methods using Modeller 9.12 software [35]. The Modeller 9.12 software constructed five model structures for OsEPSPS and the model with the lowest Discrete Optimized Protein Energy (DOPE) score was visualized by Accelrys Discovery studio version 4.1. This model was used for the identification of active sites and for docking of the substrate with the EPSPS. The *E. coli* harbours both of the EPSPS domains which probably indicate toward similar mode of action as in microbes. In this study, predicted structure of OsEPSPS was generated and the N-terminal and C-terminal domains were identified (Figure 4). *E. coli* EPSPS consists of six aligned parallel alpha-helices in each of two similar EPSPS I domains [36]. Similar domain structures were detected by Gong et al. [37], Garg et al. [38] and Filiz and Kok [39]. Bacterial EPSPS are reported to fold in two globular domains and an inside-out α-β barrel domain with PEPS3P binding in the interdomain cleft region [40]. The secondary structural features of the EPSPS sequences of 1G6S and OsEPSPS were calculated using SOPMA [41] with default parameters (Table 2). The EPSPS protein is composed of 42.52% α-helices, 17.86% extended strands and 10.10% beta turn in rice. In case of *E. coli*, the EPSPS protein is composed of 38.88% α-helices, 20.61% extended strands and 11.48% beta turn. Thus the α-helices and the beta sheets cover comparatively larger portions of the rice and *E. coli* EPSPS enzymes. Similar results have been observed by Gong et al. [37], Garg et al. [38] and Filiz and Kok [39] in several plant species. ScanProsite server identified the two signature sequences LFLGNAGTAMRPLTA (166-180) and RVKETERVMAIRTETLTKLG (427-445) in both target and template. Several physico-chemical properties of EPSPS sequences were calculated by using Expasy’s ProtParam server [21]. The results are shown in Table 3. In developing buffer system for protein purification (isoelectric focusing method), the computed isolectric point (pI) will be useful. The very high aliphatic index of the EPSPS enzyme sequences indicate that these enzymes may be stable for a wide temperature range. The high extinction coefficient of enzyme in rice indicates the presence of more Cys, Trp and Tyr. The instability index value for the EPSPS proteins were found to be ranging from 28.78 to 33.83 indicating the stable nature of the proteins. Using NetNGlyc 1.0 server, the N-glycosylation sites (188 NATY and 464 NITA) of the OsEPSPS protein were predicted and may play role in posttranslational modifications for enzymatic function. N-glycosylation is an essential process for posttranslational modifications of proteins [42].

Using String software, the EPSPS interacting partners as well as its co-expression genes were predicted in both rice and *E. coli* (Figure 5). Some proteins such as 3-dehydroquininate synthase, 3-dehydroquininate dehydratase, shikimate kinase, chorismate synthase and shikimate-5-dehydrogenase are found to be common interacting partners of EPSPS in both rice and *E. coli*. In the second step of shikimate pathway, 3-dehydroquininate synthase converts the 3-deoxy-arabinoheptulosonate-7-phosphate to 3-dehydroquininate and is essential for basic cellular metabolism machinery. In the fifth step of shikimate pathway, Shikimate kinase, an ATP dependent enzyme
Figure 1: Comparative sequences structure alignment of OsEPSPS with other homologues.

Motif | Multilevel consensus sequences | Function
--- | --- | ---
1 | ITPPEKLNTETIDTDHRRMAMCSLAAADVNPVTDIPGCTKTFPDPYF | Protein kinase C phosphorylation site, Casein kinase II phosphorylation site and N-glycosylation site
2 | DVNMMNDPVAMTLAVALFADGPTAIRDVASVRKETERMIAICTELRK | EPSP synthase, Protein kinase C phosphorylation site
3 | EGDASSASYFLAGAATGTGTVVECCGTNSLQGVDVFAEVLKMGAKVTW | DLRB LDL-receptor class B (DLRBR), N-myristoylation site
4 | ISSQYTLALLMAAPLALGDEVIEDIKLISIPYVMETLMLERFGVESVEH | Protein Kinase C Phosphorylation Site
5 | VLOPIKEIGTALKPSKLSNRLLLAASEGTVVDNLLSDDHYYLM | Casein kinase II phosphorylation site, Protein kinase C phosphorylation site, Pumilio RNA-binding repeat profile
6 | AVTACGGNARYVLGVPIRMERPGDLVDGLKQLGADVDC | EPSP synthase
7 | SWDRFYIKGQKSYKSPGNYV | -

Table 1: Multilevel consensus sequences for the MEME defined motifs and their predicted functions.

<table>
<thead>
<tr>
<th>Secondary structure element</th>
<th>OsEPSPS</th>
<th>16S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpha helix</td>
<td>42.52%</td>
<td>38.88%</td>
</tr>
<tr>
<td>$\beta_1$ helix</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Pi helix</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Beta bridge</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Extended strand</td>
<td>17.86%</td>
<td>20.61%</td>
</tr>
<tr>
<td>Beta turn</td>
<td>10.10%</td>
<td>11.48%</td>
</tr>
<tr>
<td>Bend region</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Random coil</td>
<td>29.51%</td>
<td>29.04%</td>
</tr>
<tr>
<td>Ambiguous states</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Other states</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
</tbody>
</table>

Table 2: Details of the calculated secondary structure elements by SOPMA.
catalyzes the phosphorylation of shikimate to shikimate 3-phosphate. The seventh step of the shikimate pathway for the biosynthesis of aromatic amino acids is catalysed by chorismate synthase which is conserved in prokaryotes, fungi and plants [43].

Validation of OsEPSPS structure

RAMPAGE server and PROCHECK generated model revealed that 94.3% residues are falling in the most favoured region, 4.1% residues in allowed region, and 1.6% residues in outlier region of the Ramachandran plot (Figure 6). ProSA-Web analysis of the model revealed a Z-score value of target protein. The Z-score value of the target model OsEPSPS (-8.01) is located within the space of proteins determined by NMR and X-ray crystallography. The Z-score value is close to the value of template 1G6S (-11.83) which suggested that the obtained model was reliable and very close to experimentally determined structures (Figure 7a). Verify3D showed a score greater than 0.2 in 76% of the residues that corresponded to the quality of the OsEPSPS model that was acceptable and reliable. The value of RMSD indicates the degree to which the two three dimensional structures are similar. The Co RMSD and backbone RMSD deviation for the OsEPSPS model and the E. coli template (1G6S) crystal structure were 1.58 Å, and 1.56 Å, respectively and overall RMSD was 1.72 Å. Thus, the OsEPSPS model generated by Modeller 9.12 was confirmed to be reliable and accurate. The superimposition of the template and the model structure is shown in Figure 7b. It shows that the helix and the sheet regions of the template model structure superimposed in a better way and a large deviation can be observed mainly in loop regions. It is reported that the loop region is the main region where the accuracy of a model protein structure deviates from the templates [44]. The ribbon diagram shown in Figure 4. 14C shows the docking of glyphosate (white balls) and S3P (brown balls) into the structure of OsEPSPS (target).

Prediction of active sites and docking studies

After the final model was built, the possible binding sites of OsEPSPS were searched using various binding site prediction servers such as Q-site finder, CASTp and PINUP [29-31]. These studies showed that residues K, Q, D were highly conserved in active site of both model and the template protein and hence it could be predicted that their biological function would be identical. These conserved residues may function as the catalytic domains of EPSPS enzymes and could be in the glyphosate binding site as seen in bacterial EPSPS [7]. The mutation of a single amino acid (particularly lysine and arginine) can alter the binding site of glyphosate [37]. Molecular docking was performed by Sybyl 8.0 Surflex-Dock method (Tripos Inc., USA). We docked S3P and GPJ inside the cavity of OsEPSPS protein (Figure 7c). The Shikimate-3-phosphate (S3P) has ligand binding residues at 94, 99, 173, 251, 277, 280, 402, 429 and the binding residues are K, S, R, T, S, S, Q, S, Y, D, and K respectively. The glyphosate (GPJ) ligand has ligand binding residues at 94, 170, 172, 251, 402, 430, 474, 475, 500 and the binding residues are K, N, G, R, Q, D, E, R, H, and K respectively. Both GPJ and S3P have similar amino acids K, Q, D at positions 94, 251 and 402 respectively (Table 4). The glyphosate binding site is dominated by basic residues (Arg and Lys) [45] indicating their role in glyphosate-EPSPS binding.

Phylogenetic analysis

The phylogenetic analysis of EPSPS across the selected organisms showed a clear delineation of EPSPS into four clusters. Phylogenetic tree results outline the development of EPSPS in Arabidopsis thaliana, Amborella trichopoda, Brassica rapa, Brachypodium distachyon, Cucumis melo, Fragaria vesca, Glycine max, Malus domestica, Oryza

Figure 2: Comparative sequence alignment of OsEPSPS (target) and E. coli EPSPS (template) using superpose.
Figure 3: Block diagram of multilevel consensus sequences for the MEME defined motifs of EPSPS proteins: Seven motifs were obtained by MEME software. Different motifs are indicated by different filled boxes with numbers 1 to 7.

Figure 4: Cartoon structure of OsEPSPS showing its N- and C- termini in blue and red respectively.
Figure 5: EPSPS interacting partners as well as its coexpression genes predicted by STRING. (A) Rice (B) E. coli (C) The key to the putative interacting partners for OsEPSPS gene is listed. (D) The key to the putative interacting partners of E. coli EPSPS gene is listed.

Figure 6: The plot for OsEPSPS designed by Rampage.
In-silico Characterization, Structural Modelling, Docking Studies and Phylogenetic Analysis of 5-Enolpyruvylshikimate-3-Phosphate Synthase Gene of Oryza sativa L.

Figure 7: (A) Validation of OsEPSPS by ProSA tool. The Z-score value OsEPSPS (target) and E. coli EPSPS (template) protein were determined by NMR (represented in dark blue colour) and X-ray (represented in light blue colour). The two black dots represent Z-score value of target and the template. (B) Superposition of OsEPSPS (target) and E. coli EPSPS template (PDBID: 1G6S) shown in blue and green colour respectively. (C) Ribbon diagram showing docking of glyphosate (white balls) and S3P (brown balls).

Figure 8: Phylogenetic tree constructed by minimum evolution method of MEGA version 4.1 showing similarity of OsEPSPS with monocots, dicots and bacteria.

<table>
<thead>
<tr>
<th>Properties</th>
<th>OsEPSPS</th>
<th>1G6S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Molecular weight</td>
<td>54345.7</td>
<td>46055.7</td>
</tr>
<tr>
<td>Theoretical pI</td>
<td>8.04</td>
<td>5.37</td>
</tr>
<tr>
<td>Number of amino acids</td>
<td>515</td>
<td>427</td>
</tr>
<tr>
<td>-R</td>
<td>55</td>
<td>48</td>
</tr>
<tr>
<td>+R</td>
<td>57</td>
<td>38</td>
</tr>
<tr>
<td>Aliphatic index</td>
<td>93.42</td>
<td>94.66</td>
</tr>
<tr>
<td>Grand average of hydrophilicity (GRAVY)</td>
<td>0.101</td>
<td>-0.005</td>
</tr>
<tr>
<td>Extinction coefficients (M⁻¹ cm⁻¹)</td>
<td>34755</td>
<td>30745</td>
</tr>
<tr>
<td>Instability index</td>
<td>33.83</td>
<td>28.78</td>
</tr>
<tr>
<td>CELLO predicted location</td>
<td>Combined</td>
<td>Combined</td>
</tr>
<tr>
<td>Predicted N-glycosylation sites</td>
<td>188 NATY, 464 NITA</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 3: Physicochemical, structural and sequence properties, sub-cellular localizations and N-glycosylation sites of the EPSPS protein sequences.

<table>
<thead>
<tr>
<th>Ligand Name</th>
<th>Binding Residues</th>
</tr>
</thead>
<tbody>
<tr>
<td>S3P</td>
<td>94K 95S 99R 173T 249S 250S 251Q 277S 280Y 420D 429K</td>
</tr>
<tr>
<td>GPJ</td>
<td>94K 170N 172G 202R 251Q 402D 430E 433R 474H 475R 500K</td>
</tr>
<tr>
<td>PO₄</td>
<td>168L 169G 170N 171A 196V 199M</td>
</tr>
</tbody>
</table>

Table 4: Binding residues of different ligands of the OsEPSPS protein.

sativa, Populus trichocarpa, Phoenix dactylifera, Setaria italica, Sorghum bicolor, Solanum lycopersicum, Vitis vinifera, Zea mays, E. coli and V. chlorae. Many of these exhibited orthologous and paralogous relations with each other (Figure 8). However, B. distachyon showed highest sequence similarity to OSEPS. Amborella trichopoda is believed to be the most basal lineage in the clade of angiosperms. The results indicate that EPSPS protein gene family is strictly conserved and has evolved from bacteria.

Acknowledgements

The first author is grateful to Council of Scientific and Industrial Research (CSIR) for providing financial assistance.

Conflict of Interest

We declare that we have no conflict of interest.

References


