**In-Silico Characterization and Mutagenesis of β-Glucosidase from Anoxybacillus sp. SK3-4**

Kamal Usman\(^1\) *, Idris Shehu\(^2\), Abdulrahman Mohmoud Dogara\(^2\), Zainab Ladidi Idris\(^3\)

\(^1\)Department of Biological and Environmental Sciences, Qatar University, Doha, Qatar
\(^2\)Department of Microbiology, Kaduna State University, Kaduna, Nigeria
\(^3\)Department of Biochemistry, Bauchi State University, Gadau, Nigeria

**Email address:**
kusman@qu.edu.qa (K. Usman)
*Corresponding author

**To cite this article:**

**Abstract:** Structural analysis of β-glucosidase (Bgl; EC 3.2.1.21) isolated from *Anoxybacillus* sp SK3-4 were carried out with the aim of generating structural and functional information of the protein, as well as theoretically improve thermostability by *in silico* mutagenesis. Different bioinformatics databases, software’s and servers were used to generate both functional and structural information and results of the primary sequence analysis revealed that the protein has 455 amino acid residues with molecular weight of 52627.8 Da. Pattern and profile search using Interpro indicate the presence of β-glucosidase hydrolase catalytic domain. Biochemical function of the enzyme proved it catalyses the hydrolysis of 6-phospho-beta-D-glucosyl-(1, 4)-glucose, it also catalyze the hydrolysis of several phospho-beta-D-glucosides but not phosphorylated form. Tertiary structure modelling and subsequent validation of the models have identified CPH model as the best model compared to those built by Swiss and I-TASSER. Predicted active site residues were Thy (166) and Arg (355). Other Important residues forming the binding site were predicted to include Gln (23), His (120), Asn (165), Tyr (298), Trp (402) and Ser (410). Site directed mutagenesis was adopted to introduce the mutation at two sites: Arg (41) Lys and Lys (45) Glu, which improve the thermostability of the protein.

**Keywords:** Thermo-Stability, β-Glucosidase, Bioinformatics, Mutagenesis, *Anoxybacillus* sp SK3-4

**1. Introduction**

Exhaustive screening of microbial isolates over the last several years has resulted in the identification and commercialization of numerous biomolecules, many of which are the products of microbial secondary metabolism [1]. Thermophilic bacteria have attracted the interest of many scientists especially biotechnologist; not only out of scientific curiosity but also because of the biotechnological potential of these bacteria which includes their roles as source of thermo-stable enzymes. Thermophilic microorganisms have been isolated from different habitats such as hot springs, deep ocean-basin cores, shallow marine environments, petroleum reservoirs, deep-sea hydrothermal vents and the leachate of a waste pile from a canning factory [2]. Genus *Anoxybacillus* is separate from the genus Bacillus, comprising aero tolerant anaerobes or facultative anaerobes [1]. The genus *Anoxybacillus* belongs to the order *Bacillales* under the *Firmicutes* phylum in the domain Bacteria. Some identified species of this phylum include *A. contaminans*, *A. flavithermus*, *A. gonensis*, *A. ayderensis*, *A. kestanbolensis*, *A. amylyoliticus* and *Anoxybacillus salavatiensis* [3]. *Anoxybacillus* sp SK3-4 is a thermophilic, rod-shaped, Gram positive and endospore-forming bacterium firstly isolated from Sungai Klah hot spring in Malaysia [3]. Being relatively new genus as compared to the well documented *Geobacillus* or *Bacillus*, most of the reported data have revealed that *Anoxybacillus* spp produce interesting enzymes that are both thermostable and alkaline tolerant [2]. Thus, being one of the recently discovered bacteria, their potentials...
need to be explored for possible biotechnological applications.

β-glucosidases (Bgl; EC 3.2.1.21) constitute a major group among glycosyl hydrolase enzymes. Out of the 82 families classified under glycosyl hydrolase category, β-Glucosidases belong to family 1 and family 3 and they are known to catalyze the selective cleavage of glucosidic bonds; a function that is very pivotal in many essential biological pathways, such as degradation of structural and storage polysaccharides, cellular signaling, oncogenesis, host-pathogen interactions, as well as in a number of biotechnological applications [4, 5]. Recent reports has identified these enzymes as one that gained momentum owing to their biosynthetic abilities [6]. The enzymes exhibit utility in synthesises of diverse oligosaccharides, glycol-conjugates, alkyl- and amino-glucosides. Presently, attempts are being made to understand the structure-function relationship of these versatile biocatalysts. Earlier reviews have described the sources, properties, physiological functions, characteristics, and catalytic action of native β-glucosidases from various plant, animal, and microbial sources [4].

Enzymes involved in starch bioconversions such as β-glucosidase are of major industrial interest and considerable attention has been focused on obtaining new enzymes with improved properties or new applications [7, 8]. Site-Directed mutagenesis of DNA is a very important tool in proteins engineering. Changing the DNA sequence can facilitate the study of the structure–function relationships of DNA, RNA, or protein coded by the DNA sequence. A variety of methods have been applied for the introduction of specific changes at predetermined sites in DNA sequence [9].

Currently, bioinformatics tools have been very indispensable for the analysis of macromolecules. In this article, we document bio information obtained from primary sequence analysis, secondary structure prediction, 3D structure modeling and mutagenesis of targeted residues of β-glucosidase of *Anoxybacillus* sp. SK3-4; with a view to offer theoretical background for further studies.

### 2. Materials and Method

#### 2.1. Protein Identification

To identify the protein from the amino acid sequence, sequence similarity search of non-redundant proteins was performed using NCBI blast at (http://blast.ncbi.nlm.nih.gov/Blast.cgi). Gene ontology and enzyme classification were accessed through appropriate links on Uniprot database. From the BLAST output, a number of homologous proteins from different organisms were selected for Multiple Sequence Alignment (MSA) using ClustalW (http://www.ch.embnet.org/software/ClustalW.html). Protein tree was then generated from the aligned sequences for phylogenetic analysis.

#### 2.2. Primary Sequence Analysis

Physicochemical properties of the protein such as molecular weight, theoretical isoelectric point (pI), extinction coefficient, half-life, aliphatic and instability indices, number of positively and negatively charged residues were predicted using ProtParam tool available at (http://expasy.org/tools/protparam) [10]. Protein Pattern and profile search was also performed using InterProscan at (http://wwwdev.ebi.ac.uk/interpro/) in order to predict structural domain, active site and their related function.

### 2.3. Secondary Structure Prediction

Secondary structure of the protein was predicted using three different prediction servers namely: GOR4, SOPMA available at Expasy (http://www.expasy.org/tools) and Predict Protein (http://predictprotein.org), all predictions were done using default settings in each of the respective servers and in accordance with the prediction methods [11]. Results of these predictions were compiled and compared.

#### 2.4. Tertiary Structure (3-D) Prediction

The 3-D structure modeling of the protein was performed using three homology modeling tools namely: Swiss Model (http://swissmodel.expasy.org), CPHmodels (http://cbs.dtu.dk/services/CPHmodel) and I-TASSER (http://zhanglab.ccmb.med.umich.edu/I-TASSER). In each case, the query sequence was submitted to modeling programs under automatic mode settings, in which the servers automatically select the template from experimentally solved protein structures thereby generating the best possible model according to homology modeling method. The predicted 3-D models were then accessed in form of PDB files after which the files were separately uploaded and visualized in Pymol viewer. I-TASSER concurrently predicted the binding sites from the template used. Attempt was also made to label the important residues on the best model structure obtained.

#### 2.5. Models Validation

Validation of the predicted 3-D models was performed using various validation programs to evaluate their respective qualities in order to identify the best among three. This validation was performed using ERRATv2.0 (http://nihserver.mbi.ucla.edu/ERRATv2/), VERIFY_3D (http://nihserver.mbi.ucla.edu/Verify_3D/) and RAMPAGE Server (http://mordred.bioc.cam.ac.uk/~rapper/) [1]

### 2.6. Protein Mutagenesis

To mutate the protein, two mutation sites were initially chosen based on the previous mutation work on similar protein (β-glucosidase from soil meta-genome), which identified residues at positions 41 and 45 as critical for increased thermo-stability as reported by [1]. Multiple sequence alignment was then performed between this similar protein (eUsBgl) and the query protein (Bgl SK3-4) with a view to identify the appropriate residues to be mutated to obtain mutagenized form of the query protein (eBgl sk3.4).
Side-directed mutagenesis approach using Mega primer PCR was performed. The mutation resulted in the following amino acid substitutions Arg (41) Lys and Lys (45) Glu. Three primers (P1, P2 and P3) were carefully designed from the open reading frame [12] of the gene nucleotides sequence and used for the two PCR rounds. The first round involved the use of P1 (5’ATGTACACACGCGTAAATCG 3’) which served as external forward primer and P2 (5’CCCCTCTGGAATCGGCTTAGCATC 3’) which served as internal reverse mutagenic primer carrying the substitution codons for the two residues at the appropriate positions.

3. Results and Discussion

3.1. Primary Sequence Analysis

The output of BLAST search revealed a number of homologous proteins with different percentage similarity to the query sequence. According to this result, β-glucosidase from Anoxibacillus sp SK3-4 was found to have 100% identity with the query sequence. Hence, the protein sequence (query) was identified as β-glucosidase of Anoxibacillus sp SK3-4. It was found to have UniProt Entry name and Gene name of T0BDA2_BACI and C289_1346 respectively.

From the result of primary sequence analysis performed using Prot Param, the protein was predicted to have a molecular weight of 52627.8 Da., theoretical isoelectric point (pI) of 5.58, this implies that at pH 5.58 the protein is expected to have zero net charge. Extinction coefficient in 280 nm aqueous solutions was predicted to be 89395 M-1 cm-1 the half-life was 30 hrs, >20 hrs, >10 hrs respectively in mammalian reticulocytes (in vitro), yeast (in vivo) and Escherichia coli (in vivo). The instability index (II) was found to be 35.99. This classified the protein as stable for having a value of less 40 [13]. The protein was also predicted to have aliphatic index and Grand average of hydropathicity (GRAVY) of 82.77 and -0.411 respectively.

3.2. Structure Predictions

Table 1 shows the summary of results obtained from the secondary structure predictions by the three different servers. It appears that each of the three servers yield a different result, however the relative percentages of the three basic secondary structural elements (helix, sheet and coil) are within the same ratio. The slight difference among the prediction results produced by the respective servers could probably arise due to the use of different indices in making the prediction by different servers. Gor4 and SOPMA produce result in both numeric and graphical forms. Predict Protein on the other hand only gives the numerical value in percentage.

Table 1. Secondary Structure of β-glucosidase of Anoxibacillus sp SK3-4 as predicted by different servers.

<table>
<thead>
<tr>
<th>SERVERS</th>
<th>SECONDARY STRUCTURAL ELEMENTS</th>
<th>Helix</th>
<th>Extended Sheet</th>
<th>Coil</th>
</tr>
</thead>
<tbody>
<tr>
<td>GOR4</td>
<td>40.88%</td>
<td>19.34%</td>
<td>39.78%</td>
<td></td>
</tr>
<tr>
<td>SOPMA</td>
<td>39.34%</td>
<td>16.04%</td>
<td>35.82%</td>
<td></td>
</tr>
<tr>
<td>PREDICT PROTEIN</td>
<td>36.70%</td>
<td>15.60%</td>
<td>47.69%</td>
<td></td>
</tr>
</tbody>
</table>

The 3-D models predicted by the three modelling programs are shown in figures 4, 5 and 6. According to prediction outputs, the Swiss Modeller used 6-phospho-beta-glucosidase from Streptococcus pneumonia TIGR4 (4IPN) as template. CPH models used 6-phospho-beta-glucosidase BGLA from E. coli K-12 (2XHY). I- TASSER on the other hand predicted up to five different models using different templates. However the best model among the five (based on C-score) was build-using 6-phospho-beta-glucosidase from Streptococcus pneumoniae TIRG4 (4IPL).
In all the three model predicted, the structures are in agreement with the results of the secondary structure earlier predicted with loops dominating the structure. The model predicted by CPH modeler appeared to have good structural fold resembling jelly-roll as shown in Fig. 4, however, identifying the best model need to be based on structure validation results.

Active site residues were predicted to be at position 166 and 355 both of which correspond to Glutamine with Glu (355) as the nucleophile. The distance between these catalytic residues was found to be around 11.0 Å as shown in figure 5a below. Other important residues that form the protein binding sites were predicted to include Gln (23), His (120), Asn (165), Tyr (298), Trp (402) and Ser (410). This prediction of binding sites by I-TASSER has a confidence score (C-score) of 0.64. Theoretically, confidence score value of predicted binding site might range between zeros to one (0-1) with higher values indicating a more reliable ligand binding site prediction [11, 13, 14]. The binding site residues of a protein participate in binding with enzymes’ substrate and therefore important for catalytic activity of the protein.

3.3. Models Validation

While all models built by different modeling programs appeared to have satisfied the requirement for 3D-1D average profiling of greater 0.2 [15, 16], however validation by ERRAT2 revealed an overall quality factors of 72.955%, 81.406% and 48.993% for Swiss, CPH and I-Tasser respectively. Although CPH model emerged with the highest quality factor (81.406%), the value is only close to the least average value (90.00%) accepted which yields a resolution quality of 2.5-3.0Å [17]. Ramachandran plot analysis of the three models by RAMPAGE server produced residues percentage of 95.5, 90.2 and 88.7 in the most favored region, 2. 5, 7. 3 and 7. 9 in the allowed region and 2. 0, 2. 5 and 3. 4 in the outlier for Swiss, CPH and I-Tasser respectively.

Considering the results in the Table 2, Swiss model have recorded a better score with 95.5% of residues in the most favored regions of the Ramachandran plot. Table 3 shows comparison among the three templates used in terms of X-ray diffraction experimental data and the percentage identity between the query sequence and the templates for each of the three programs. This is important in order to establish a link between the quality of template used and that of the corresponding model predicted. It can be seen from the table that I-TASSER used a template with the percentage identity of 49.0% while SWISS and CPH used templates with 46.67% and 47.18% respectively.

Table 2. Comparison of Ramachandran Plot output for the three models.

<table>
<thead>
<tr>
<th>Modeler</th>
<th>Most Favoured</th>
<th>Allowed</th>
<th>Outlier</th>
</tr>
</thead>
<tbody>
<tr>
<td>SWISS</td>
<td>95.5%</td>
<td>2.5%</td>
<td>2.0%</td>
</tr>
<tr>
<td>CPH</td>
<td>90.2%</td>
<td>7.3%</td>
<td>2.5%</td>
</tr>
<tr>
<td>I-TASSER</td>
<td>88.7%</td>
<td>7.9%</td>
<td>3.4%</td>
</tr>
</tbody>
</table>

Table 3. X-ray Diffraction experimental data of the different templates used by the three Modelling programs.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Swiss model</th>
<th>CPH models</th>
<th>I-TASSER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein Model and PDB ID</td>
<td>6-Phospho-beta-glucosidase from</td>
<td>6-Phospho-beta-glucosidase</td>
<td>6-Phospho-beta-glucosidase</td>
</tr>
<tr>
<td></td>
<td>Streptococcus pneumonia TIGR4 (4IPN)</td>
<td>BGLA from E. coli K-12 (2XHY)</td>
<td>from Streptococcus pneumonia</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TIGR4 (4IPL)</td>
</tr>
<tr>
<td>Target and Template Identity (%)</td>
<td>46.67</td>
<td>47.18</td>
<td>49.0</td>
</tr>
<tr>
<td>Resolution of Template Structure (Å)</td>
<td>2.41</td>
<td>2.30</td>
<td>2.41</td>
</tr>
<tr>
<td>R-factor</td>
<td>0.178</td>
<td>0.174</td>
<td>0.178</td>
</tr>
</tbody>
</table>

Figure 5. (a) The two active site residues separated by a distance indicated by dotted line. (b) Predicted binding residues as viewed in Pymol with grey background (c) Protein surface with arrow indicating binding pocket.
The template used by CPHmodel appeared to have the lowest resolution value of 2.30Å which indicate higher resolution in solved structures and key to determining the structural features of proteins such as chain tracing, secondary structure, side chain conformations, orientation of peptide planes and by extension means a better model template [18]. Also R-factor for CPH template appeared to be the least, which is a statistical value that measure how well the coordinates reproduce the experimental data and theoretically the lower the R-factor the more likely the model structure will be accurately predicted.

Therefore, comparing the results of various validation and template experimental data, CPH model that has a relatively better scores from the three validations including the 90.2% score for residues in the most favoured regions which is slightly above the least accepted value for good protein model structure [19]. Hence, CPH Model is suggested to be the better predicted structure for the query protein when compared to the other two predicted models by Swiss and I-Tasser.
Figure 6. Ramachandran plots of predicted Swiss (i), CPH (ii) and I-TASSER (iii) models of the protein.
3.4. Protein Mutagenesis

As a fundamental technique, site-directed mutagenesis is a powerful tool for introducing changes at a desired position in a gene. A careful selection of mutation site and good primers design are critical to successfully accomplish the process, background information about the protein to be mutated is also very important. Literature search revealed that a number of proteins belonging to Glucosidase Hydrolase family 1 (GH1) have been previously engineered with a view to enhance their properties for industrial application. β-glucosidase from *Anoxybacillus* sp. SK3-4 is a novel enzyme with no record of previous engineering work. However, a similar protein (β-glucosidase) from a soil metagenome appeared to have been previously mutated at different positions to improve its thermo-stability and solubility [1].

Choice of the mutation sites for our protein (Bgl SK3-4) was based on the multiple sequence alignment with β-glucosidase from a soil metagenome (eUsBgl). Result indicate that the positions of residues that are critical for thermo-stability (41 and 45) are not conserved, and hence mutating the residues to form (eBgl SK3-4) that is expected to have improved thermo-stability.

4. Conclusion

In conclusion, β-glucosidase from *Anoxybacillus* sp. SK3-4 holds a tremendous potential for industrial application. It is possible to build a structure model of a protein if homologous protein (template) with solved structure exists. A careful selection of mutation site and good primers design are critical to successfully accomplish the process, background information about the protein to be mutated is also very important. Literature search revealed that a number of proteins belonging to Glucosidase Hydrolase family 1 (GH1) have been previously engineered with a view to enhance their properties for industrial application. β-glucosidase from *Anoxybacillus* sp. SK3-4 is a novel enzyme with no record of previous engineering work. However, a similar protein (β-glucosidase) from a soil metagenome appeared to have been previously mutated at different positions to improve its thermo-stability and solubility [1].

Choice of the mutation sites for our protein (Bgl SK3-4) was based on the multiple sequence alignment with β-glucosidase from a soil metagenome (eUsBgl). Result indicate that the positions of residues that are critical for thermo-stability (41 and 45) are not conserved, and hence mutating the residues to form (eBgl SK3-4) that is expected to have improved thermo-stability.

References


