Chapter 9

In-silico prediction of structural and functional aspects of salt responsive protein of SKP1-like protein 1A.

9.1 Introduction

In silico analysis of different gels provide a complete catalogue that exists between different sample protein patterns (Pirondini et al., 2010). Protein identification and analysis is usually based on in silico techniques that match the peptide sequences that paves the way in metabolomics, transcriptomics and other large pomics techniques into systems biology (Chen and Harmon, 2006). There are tremendous opportunities for molecular biologists to define the protein nature that the transducer genetic information and energy transfer that obtains from plants. These proteomics methods are very much important for the study of different components in the plant that states plant function (Roberts, 2002).

When carrying out analysis of protein sequences, the aim is to find out as much information as possible about potential relationships with other sequences as well as characterizing their physiochemical properties. The first step usually involves comparing the protein sequence against a non-redundant protein sequence database by using Blast (Altschul et al., 1997) or Fasta (Pearson and Lipman, 1988), which will reveal which sequence(s) are similar to the query sequence alone. To obtain further information about a proteins specific function, searches against secondary databases also known as pattern or signature databases are necessary. When such searches return significant matches or hits, these results help in the assignment of a particular function or functional domain to the query protein (Quevillon et al., 2005).

Proteins are important biological polymers formed from building blocks called amino acids. The three-dimensional structure and biological activity of proteins depend on the physicochemical properties of their constituent amino acids. The primary structure identifies a
protein unambiguously, determines its chemical and biological characteristics, and specifies the higher levels of protein structure (Cozzone, 2002).

Prediction of secondary structures is a fundamental basis for protein structure prediction. Protein structure determination and prediction has been a focal research subject in the field of bioinformatics due to the importance of protein structure in understanding the biological and chemical activities of organisms (Singh et al., 2008). The term "homology modeling", also called comparative modeling or template-based modeling (TBM), refers to modeling a protein 3D structure using a known experimentally determined structure of a homologous protein as a template. Protein modeling is the only way to obtain structural information if experimental techniques fail. Many proteins are simply too large for NMR analysis and cannot be crystallized for X-ray diffraction (Krieger et al., 2003).

In the present study, In silico characterization of amino acid sequence of SKP1-like protein 1A protein was carried out. Secondary structures of the protein based on their residues were predicted and classified. Type of residues and their physical properties have been estimated. By profiling the compositional analysis of the proteins the charge distribution, atomic composition, extinction coefficients, instability index and aliphatic index were predicted. The present investigation implies homology modeling to deduce a three dimensional structure of the uncharacterized protein under study, followed by validation and comparative modeling of the obtained structure for its conformational stability and further biological analysis.
9.2 Materials and Methods

9.2.1. Sequence analysis

The sequences that are sequenced in MALDI were selected for in silico analysis. The sequences were compared for detecting homologous sequences found in databases using Basic Local Alignment Search Tool (BLAST) (Altschul et al., 1997). BLAST from NCBI was used to compare the query sequence with the database sequence to find its homologues.

9.2.2. MOTIF search

The motifs were identified using the tool Motif Search (Thakallapally et al., 2000).

9.2.3. Conserved domain search

Conserved domain search was performed using CDD according to (Marchler-Bauer et al., 2011). Enter a protein or nucleotide query as an accession or GI number (e.g., AAC50285 or 463989) or as a sequence in FASTA format to identify the protein's conserved domains and therefore its putative function.

9.2.4. InterPro based protein signature recognition analysis

The InterPro project home page is at http://www.ebi.ac.uk/interpro. InterProScan can take either nucleotide or protein sequences in a recognized sequence format (such as raw, FASTA or EMBL). It will reformat and, if necessary, translates the sequences before beginning its search tasks. If raw format (free text) is used, it will be given the name "Sequence_n" by default, where n is the order in which it appeared in the input (Hunter et al., 2011).

9.2.5. ProtParam based physical and chemical computation in protein sequences

Using the primary sequence, the physicochemical properties of the protein were calculated with the aid of the tool ProtParam. The computed parameters include the molecular weight, theoretical pI, amino acid composition, atomic composition, extinction coefficient,
estimated half-life, instability index, aliphatic index and grand average of hydropathicity (Gasteiger et al., 2005).

9.2.6. Secondary structure prediction

The GOR IV server (Garnier et al., 1996) analyzes sequences to predict alpha helix, beta sheet, turn or random coil secondary structure at each position based on 17-amino-acid sequence windows.

9.2.7. Tertiary structure prediction

9.2.7.1. Homology modeling

For homology modeling, initially a suitable template was searched using Protein data bank (Altschul et al., 1997), SWISS model template library (Arnold et al., 2006) and PSI-BLAST (Berman et al., 2000). Amino acid sequence alignment of target and template proteins and rough 3D models (03 models) were constructed by using homology modeling servers Swiss models (Arnold et al., 2006).

9.2.7.2. Model structure validation

All the generated structures of the protein model were subjected to a series of tests for testing its internal consistency and reliability. Backbone conformations were evaluated by the inspection of the Psi/Phi Ramachandran plot obtained from PROCHECK analysis (Laskowski et al., 1993). The analyses can be generated by clicking on the PROCHECK button at the top of the PROCHECK summary page; and can get to this page from the PDB sum page of the structure in question either by clicking on PROCHECK in the index on the left hand side of the page, or else by clicking on the Ramachandran plot on the right.

9.2.7.3. Structure visualization

The modeled tertiary structure was visualized using Rasmol. It reads molecular structure files from the Protein Data Bank (PDB) in .pdb extension. The web server is available at http://www.openrasmol.org.
9.3 Result and discussion

9.3.1. Sequence analysis

*In silico* analysis of the obtained sequence was carried out. BLAST homology searches showed that this sequence shares 100% identity with PREDICTED: SKP1-like protein 1A (*Sesamum indicum*).

**PREDICTED: SKP1-like protein 1A (*Sesamum indicum*)**

MSSDGAAQKMIVLKSDDGTEFEVEEAAPESQTIKHMIEDNCAUTSIPLPVTSKILAKVIEYCKRHVDA AAASADATASDKVAEDDLKAFDAEFVKVDQGTLFDLILAANYLNIKSLDDLCQTVADMIGKTPEEIR KTFN1KNDFTPEEEEEVRRENAWF

9.3.2. MOTIF search

It was found that about 3 motifs were present in the test sequence such as SKP1, SKP1_POZ and Ribosomal_L7Ae. Positions and E-values are given in Fig 42.

![Result of MotifFinder](image)

**Fig 42: Predicted motif search of SKP1-like protein 1A**
9.3.3. Conserved domain search

Conserved domain search revealed that, sequence resembles S-phase kinase-associated protein 1 (SKP1) family. The domain identified in the protein sequence was positioned at 89 to 166 residues (Fig 43).

Fig 43: Conserved domain search (CDD) using BLAST for predicted SKP1-like protein 1A

9.3.4. Integrated protein signature

In the present study, the protein signature analysis for the SKP1-like protein1A sequences of the study species with highlighted domains corresponding to the matches in the InterPro database was obtained (Fig 44). This result again confirms that uncharacterized protein sequence belongs to SKP1 super family and containing three domains of SKP1 component POZ domain, POZ domain and SKP1 component dimerisation.
Fig 44: Integrated protein signature analysis for predicted SKP1-like protein 1A


9.3.5. Physicochemical properties of the protein

ProtParam results exhibit the physicochemical parameters of uncharacterized protein (Gasteiger et al., 2005). There are 166 amino acids in the sequence, its molecular weight was 18347.6 and theoretical pI was 4.53. The maximum number of amino acids present in the sequence were Alanine (12.7%) followed by Glutamine (10.8%) and Aspargene (9%) (Table 7).
The total numbers of negatively charged residues (Aspartic acid + Glutamic acid) were thirty three while the total numbers of positively charged residues (Arginine + Lysine) were nineteen. Atomic composition computed as C (802), H (1281), N (209), O (267) and S (7). Thus $C_{802}H_{1281}N_{209}O_{267}S_7$ has been arrived as the molecular formula for the SKP1-like protein A protein. The aliphatic index was calculated as 83.55. The instability index is computed to be 46.64 which classified the protein as unstable. The grand average of hydropathicity (GRAVY) was calculated to be -0.336.

**Number of amino acids: 166**

**Molecular weight: 18347.6**

**Theoretical pI: 4.53**

a. **Charge distribution**

- Total number of negatively charged residues (Asp + Glu): 33
- Total number of positively charged residues (Arg + Lys): 19

b. **Atomic composition:**

<table>
<thead>
<tr>
<th>Element</th>
<th>Symbol</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carbon</td>
<td>C</td>
<td>802</td>
</tr>
<tr>
<td>Hydrogen</td>
<td>H</td>
<td>1281</td>
</tr>
<tr>
<td>Nitrogen</td>
<td>N</td>
<td>209</td>
</tr>
<tr>
<td>Oxygen</td>
<td>O</td>
<td>267</td>
</tr>
<tr>
<td>Sulfur</td>
<td>S</td>
<td>7</td>
</tr>
</tbody>
</table>

c. **Formula:** $C_{802}H_{1281}N_{209}O_{267}S_7$

d. **Total number of atoms:** 2566
e. **Extinction coefficients:**

Extinction coefficients are in units of $M^{-1} \text{cm}^{-1}$, at 280 nm measured in water.

- Ext. coefficient $8605$
- Abs 0.1% ($=1 \text{ g/l}$) 0.469, assuming all pairs of Cys residues form cystines

- Ext. coefficient $8480$
- Abs 0.1% ($=1 \text{ g/l}$) 0.462, assuming all Cys residues are reduced
f. **Estimated half-life:**

   The N-terminal of the sequence considered is M (Met).

   The estimated half-life is: 30 hours (mammalian reticulocytes, in vitro).
   > 20 hours (yeast, in vivo).
   > 10 hours (Escherichia coli, in vivo).


g. **Instability index:**

   The instability index (II) is computed to be 46.64
   This classifies the protein as unstable.


h. **Aliphatic index: 83.55**

   Grand average of hydropathicity (GRAVY): -0.336


i. **Aminoacid composition**

   ![](image)

   **Table 7: Aminoacid composition of SKP1 like protein 1A**
9.3.6. Secondary structure prediction

The secondary structural analysis of the protein was done and alpha helix (Hh) was found to be most frequent (63.86%), followed by random coil (30.72%). Extended strand (Ee) was found to be least frequent (5.42%) (Table 8). The dominance of the coiled regions indicates the high level of conservation and stability of the protein structure (Neelamathi et al., 2000).

<table>
<thead>
<tr>
<th>Structural elements</th>
<th>Number of residues</th>
<th>Percentage of residues</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpha helix (Hh)</td>
<td>106</td>
<td>63.86 %</td>
</tr>
<tr>
<td>310 helix (Gg)</td>
<td>0</td>
<td>0.00 %</td>
</tr>
<tr>
<td>Pi helix (II)</td>
<td>0</td>
<td>0.00 %</td>
</tr>
<tr>
<td>Beta bridge (Bb)</td>
<td>0</td>
<td>0.00 %</td>
</tr>
<tr>
<td>Extended strand (Ee)</td>
<td>9</td>
<td>5.42 %</td>
</tr>
<tr>
<td>Beta turn (Tt)</td>
<td>0</td>
<td>0.00 %</td>
</tr>
<tr>
<td>Bend region (Ss)</td>
<td>0</td>
<td>0.00 %</td>
</tr>
<tr>
<td>Random coil (Cc)</td>
<td>51</td>
<td>30.72 %</td>
</tr>
<tr>
<td>Ambiguous states (?)</td>
<td>0</td>
<td>0.00 %</td>
</tr>
<tr>
<td>Other states</td>
<td>0</td>
<td>0.00 %</td>
</tr>
</tbody>
</table>

Table 8: Predicted secondary structure of SKP1 like protein 1A
Information about amino acid composition would also be useful for refining the analysis of truncated hydrophobic clusters that only cover a limited part of the associated regular secondary structure. Amino acids in the vicinity of these hydrophobic clusters, such as A, C and T that can substitute for strong hydrophobic residues, might indicate the overflowing of the hydrophobic cluster by the associated regular secondary structure.

Pilley, (2011) reported that modeled stress induced protein of *Capnocytophaga canimorsus Cc5* exhibited a maximum number of random coils (47.06%) with alpha helix (36.03%) and extended strands (16.91%) as secondary structural elements. In this study alpha helix (Hh) was found to be most frequent (63.86%), followed by random coil (30.72%) and extended strand (Ee) was found to be least frequent (5.42%). However, this study protein contains more structural elements.

9.3.7. Tertiary structure prediction

Template search with BLAST was performed against PDB and SWISS-model template library. A total of 32 templates were found. Suitable tree templates of PDB ID: 3ogm (*Arabidopsis thaliana*), 2ass (*Homo sapiens*), 3ogl (*Arabidopsis thaliana*), were selected on the basis of sequence similarity with a resolution of 78% with 3.34Å, 60.8% with 3Å and 79% with 3.18 Å respectively. For each identified templates, the template quality was predicted from the features of the target-template alignment and rough 3D models were built for three templates using SWISS modeler.

The models were validated in PROCHECK. The formation of helix-helix contacts is crucial for membrane protein folding. Residues involved in helix-helix contacts are therefore evolutionary conserved. Ramachandran plot shows the phi-psi torsion angles for all residues in the structures of the three models for target protein of SKP1-like protein A. The coloring/shading on the plot represents the different regions (Morris et al., 1992) the darkest areas shown in red correspond to the "core" regions representing the most favorable combinations of phi-psi values (Fig 45).
The percentage of residues in the “core” regions is one of the better guides to stereochemical quality (Table 9). The plot shows separate Ramachandran plots for each of the 20 different amino acid types. Shading on each plot indicates how favorable each region on the plot is; the darker the shade the more favorable the region. The numbers in brackets, following each residue name, show the total number of data points on that graph (Fig 46 a-c). The red numbers above the data points are the reside-numbers of the residues lying in unfavorable regions of the plot. This gives a visualization of which regions appear to have consistently poor or unusual geometry and which have more normal geometry (Berman et al., 2000).

Fig 45: Ramachandran analysis of the backbone dihedral angles Psi (j) and Phi (s) for the three template models PDB ID 3ogm, 2ass and 3ogl validated with ProCheck program. Red region represents the most favored region, yellow = allowed region, light yellow = generously allowed region, white = disallowed region.
<table>
<thead>
<tr>
<th>Percentage parameters</th>
<th>Model 1 (3ogm)</th>
<th>Model 2 (2ass)</th>
<th>Model 3 (3ogl)</th>
</tr>
</thead>
<tbody>
<tr>
<td>% of residue in most favored regions</td>
<td>79.9</td>
<td>80.1</td>
<td>82.8</td>
</tr>
<tr>
<td>% of residue in the additionally allowed zones</td>
<td>19.2</td>
<td>17.4</td>
<td>16.5</td>
</tr>
<tr>
<td>% of residue in the generously regions</td>
<td>0.2</td>
<td>2.5</td>
<td>0.7</td>
</tr>
<tr>
<td>% of residue in disallowed regions</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>% of non-glycine and non-proline residues</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 9: Percentage of residues falling in the core region of Ramachandran’s plot for the three template models of 1, 2 and 3 (PDB ID 3ogm, 2ass and 3ogl).

Fig 46 a: The plot shows separate Ramachandran plots for different amino acid types for model 1 SKP1-like protein1A protein sequence. The darker the shaded area on each plot, the more favorable the region. The red numbers above the data points are the residue-numbers lying in unfavorable regions of the plot.
Padaria et al., (2013) reported PROCHECK analysis of uncharacterized protein generated from heat responsive showed 65.2% residues in most favored region (A,B,L), 30.3% residues in additional allowed region and 4.5% residues in disallowed region. Kurkarni et al., (2014) reported ramachandran plot for salt stress responsive transcription factor SsMYB2R protein.
*Saccharum sportuneum* obtained through PROCHECK founded the residues of protein in the core region was 78.1%. In the present study, percentage of residues in the core region of protein sample was 79.9%, 80.1% and 82.8% for the three models. Ramachandran plot results highlighted that the SKP1-like protein1A protein sequences of model 3 showed high stereochemical quality when compared to that of model 1 and model 2 sequences. When comparing the results with previous reports of stress responsive proteins such as uncharacterized protein (Padaria *et al.*, (2013) and SsMYB2R protein (Kurkarni *et al.*, (2014)) the percentage of stereochemical quality of this SKP1-like protein1A was found to be comparatively high.

### 9.3.8. Tertiary structure of SKP1-like 1A

Based on the result of sequence similarity, resolution value and the validation report, model 3 (PDB ID: 3ogl *Arabidopsis thaliana*) was selected as a best template for the target sequence of SKP1-like protein1A and finally tertiary structure of the protein visualized using Rasmol (Fig 47).

**Fig 47: Tertiary structure view of SKP1-like protein 1A**

The sequence and structural analysis of S-phase kinase-associated protein1–like protein1A (SKP1-like protein1A) was carried out using bioinformatics tools. Sequence analysis of SKP1-like protein1A revealed the presence of highly conserved region of SKP1 protein family. The sequence analysis revealed that SKP1-like protein1A protein might be involved in ubiquitine dependent protein degradation in stress related signaling and response mechanism. Thus results aid to the experimental data and help to built up a complete view of SKP1-like protein1A protein role in plant stress response.