3. MATERIALS AND METHODS

The ability to assign function from protein structure is quite difficult but a very important task, especially for structural genomics projects, in which protein structures are experimentally determined but have very low sequence identity to any currently available protein in the Protein Data Bank (PDB) (Berman et al. 2002). Tools that allow for the prediction of protein function from structure are becoming increasingly important as these projects gather momentum. Comparing proteins, deriving structural patterns and ultimately utilizing such patterns are important steps in determining the structure of the protein and further predicting its function.

In the present study, we have integrated the available tools with the knowledge based approach to study the behaviour of osmotin protein as cell signaling modulator. Osmotin protein was analyzed with respect to its function as transcription factor and/or cell signal pathway modulator for proline biosynthesis and its accumulation, using softwares for protein data analysis and visualization, database searching, homology matching and molecular simulation. The bioinformatic tools that were used and the methodology that was followed in the present study are described in the following sections:

3.1. Osmotin protein: Tobacco (Nicotiana tabacum)

3.1.1 Retrieval of sequence and structural information on osmotin protein: Tobacco (Nicotiana tabacum) osmotin protein sequence and structural features (1PCV) were downloaded from UniProtKB/Swiss-Prot database and RCSB-Protein Data Bank (PDB), respectively, for analysis. Amino acid sequence of osmotin protein in FastA format was submitted to Compute pI/Mw at ExPASy proteomics server for calculating theoretical pI and molecular weight. The subcellular location of osmotin protein was predicted using TargetP 1.1 Server; http://www.cbs.dtu.dk/services/TargetP hosted at Technical University of Denmark, Denmark.

3.1.2 Functional characterization of osmotin as transcription factor: Prediction of DNA-binding motifs

To study the osmotin protein, its amino acid sequence and structure (1PCVA) were retrieved from NCBI and RCSB-Protein Data Bank, respectively. For any protein to act as transcription factor, it must have a DNA binding domain. The identification of DNA binding
site on protein is an integral part of a larger system that is required to make inferences about protein structure and function. This is done by studying the presence of binding clefts, identification of catalytic sites and binding sites of other molecule. This is generally achieved by integrating knowledge based approach with the use of available standalone as well as online softwares.

To analyze the sequence and structure of osmotin for DNA binding motif and domain, online DNA-binding motif prediction softwares were used. GT1, a known transcription factor and a member of trihelix family of transcription factors from *Arabidopsis thaliana* was taken as reference for this study. The three dimensional (3D) structure of GT1 (2jmwA) was downloaded from RCSB-PDB. Osmotin and GT1 amino acid sequences and structures (as required) were submitted to DNA-binding motif prediction softwares:

- 2-zip server,
- DBD transcription factor prediction database,
- GYM 2.0, and
- Predictdnahth.

3.1.2.1 2-zip server: The leucine zipper is a dimerization domain occurring mostly in regulatory and thus, present in many oncogenic proteins (Bornberg-Bauer et al. 1998). 2-zip is the program that identifies leucine zippers from sequence alone. It combines a standard coiled coil prediction algorithm with an approximate search for the characteristic leucine repeat. No further information from homologs is required for prediction. This approach scores significantly over existing methods, especially because the coiled coil prediction turns is highly informative and avoids large numbers of false positives.

3.1.2.2 DBD transcription factor prediction program: DNA-binding domain (DBD) is a database of predicted sequence-specific DNA-binding transcription factors (TFs) for all publicly available proteomes (Kummerfeld and Teichmann 2006). The access to TF predictions is provided through http://transcriptionfactor.org, where all data is available for viewing and immediate download. It allows browsing through the predictions for over 700 species (from *Arabidopsis thaliana* to *Zymomonas mobilis*) or DBD family (including helix–turn–helix, zinc-fingers, homeobox and many others), search predictions by sequence identifier or domain family, receive classifications for submitted protein sequences, and download domain assignments, as well as a manually curated list of DBDs. The prediction method in the DBD database uses hidden Markov models (HMMs) to identify domains in proteins from two databases: SUPERFAMILY and Pfam. The prediction of TFs begins.
with a domain annotation of all proteins from completely sequenced genomes with all HMMs from the SUPERFAMILY and Pfam databases. A protein is classified as a TF, if it has a significant match to a model annotated as being a DBD, with the significance thresholds for HMM matches taken from the Pfam and SUPERFAMILY databases. The TF predictions are limited to the families in the annotated collection, which means that the coverage is about two-thirds of known TFs. At the same time, up to an additional 50% of proteins are predicted as TFs that have annotations such as ‘hypothetical protein’, particularly in metazoan genomes. The prediction method is general and applicable to any proteome or sequence set.

3.1.2.3 GYM 2.0: GYM 2.0 program is used to detect Helix-Turn-Helix (HTH) motifs in protein sequences (Narasimhan et al. 2002). The HTH motif is one of the best studied motifs in proteins. Proteins with such motif are usually transcription factors. They bind to DNA and affect the function of RNA polymerase, thus regulating the gene expression. The HTH motif of these proteins is responsible for binding with DNA. Detecting motifs including HTH motif has become an important issue in biochemistry. The algorithm assumes that a motif is constituted by the presence of a "good" combination of residues in appropriate locations of the motif. The algorithm attempts to compile such good combinations into a "pattern dictionary" by processing an aligned training set of protein sequence. The dictionary is subsequently used to detect motifs in new protein sequence. Statistical significance of the detection results are ensured by statistically determining the various parameters of the algorithm. The Helix-Turn-Helix motif was used as a model system on which the program was tested. The program also detects homeodomain motifs.

3.1.2.4 Predictdnahth: Predictdnahth program predicts the DNA-binding helix-turn-helix (HTH) structural motif in a given three-dimensional protein structure (McLaughlin and Berman 2003). The program uses statistical models based on geometrical measurements of the motif; therefore, either a PDB file or a PDB id for the structure has to be provided. With a decision tree model, key structural features required for DNA binding are identified. These include a high average solvent-accessibility of residues within the recognition helix and a conserved hydrophobic interaction between the recognition helix and the second alpha helix preceding it. The Protein Data Bank is searched using a more accurate model of the motif created using the Adaboost algorithm to identify structures that have a high probability of containing the motif, including those that had not been reported previously.

Osmotin protein structure was also compared with the transcription factors listed in Database of Arabidopsis transcription factors (DATF) from Arabidopsis to find similar DNA binding domain on osmotin protein.
3.1.2.5 Database of *Arabidopsis* transcription factors (DATF) from *Arabidopsis*. DATF aims to provide comprehensive annotations for the transcription factors of *Arabidopsis*. It is based on the *Arabidopsis* sequence of TAIR version 6 (20051108). It uses not only locus (gene), but also gene model (transcript, protein). It has multiple alignment of the DNA-binding domain of each family, neighbour-joining phylogenetic tree of each family, the gene ontology (GO) annotation and homolog with the Database of Rice Transcription Factors (DRTF). It also keeps old information items such as the unique, cloned and sequenced information of about 1200 transcription factors, protein domains, 3-D structure information with BLAST hits against PDB, predicted Nuclear Location Signals, UniGene information, as well as links to literature references.

To find DNA binding domain on osmotin protein, comparative studies were carried out using structure of transcription factor families listed in DATF. Since, the crystal structure of these transcription factors were not available therefore, first their structures were build. Computationally, the structure can be predicted by inter relating various physical, chemical and mathematical rules. Mathematical models and algorithms are needed for incorporating these rules into a physical structure (Zhang and Skolnick 2005). The methods devised for predicting the structure are comparative or homology modelling and *ab initio* method with computer algorithm. Comparative or homology modelling uses experimentally determined protein structures to predict the conformation of other proteins with similar amino acid sequences (Figure 4).

Figure 4: Schematic representation of methodology followed for protein structure determination.
Therefore, to predict the structures, the amino acid sequences of transcription factors were retrieved. Randomly, amino acid sequences of three transcription factors were taken from each of 64 families, for the study. The amino acid sequences were then submitted to SWISS-MODEL workspace, online modelling software, for the computational structure predictions of these transcription factors. Amino acid sequences of those proteins for which SWISS-MODEL failed to provide a model, were submitted to BlastP individually, with default settings so as to get a close homolog on which modelling of these transcription factors can be performed. The best match was saved and used for homology modelling of the protein structure of transcription factor.

3.1.2.6 SWISS-MODEL workspace: It is a web-based integrated service dedicated to protein structure homology modelling. It assists and guides the user in building protein homology models at different levels of complexity. A personal working environment is provided for each user where several modelling projects can be carried out in parallel. Protein sequence and structure databases necessary for modelling are accessible from the workspace and are updated in regular intervals. Tools for template selection, model building and structure quality evaluation can be invoked from within the workspace. The SWISS-MODEL workspace can be accessed freely at http://swissmodel.expasy.org/workspace/. The SWISS-MODEL workspace integrates programs and databases required for homology modelling in an easy-to-use web-based modelling workbench. It allows the user to construct comparative protein models from a computer with web connection without the need of downloading and installing large program packages and databases.

Amino acid sequences of these transcription factors were again submitted to SWISS-MODEL workspace, alignment mode, for modelling of protein’s structure. The structures were visualized and analyzed using Chimera. SuperPose program was used to study the superimposition of the structures of transcription factors on osmotin as well as GT. SuperPose: Structure comparison provides tremendous insight into the origin, function, location, interaction and activity of a given protein. The fact that the structure is actually much more conserved than sequence, structure comparisons allow us to look back into biological prehistory (Maiti et al. 2004; Orengo et al. 2003). The most common method for three-dimensional (3D) structure comparison is called structure superposition where a molecule rotated oriented until it can be directly be overlaid on top of a similar molecule. The SuperPose web server rapidly and robustly calculates both pairwise and multiple protein structure superpositions using a modified quaternion eigenvalue approach. SuperPose generates sequence alignments, structure alignments, PDB (Protein Data Bank)
coordinates and RMSD statistics, as well as difference distance plots and images (both static and interactive) of the superimposed molecules. SuperPose employs a simple interface that requires only PDB files or accession numbers as input. All other superposition decisions are made by the program. SuperPose is uniquely able to superimpose structures that differ substantially in sequence, size or shape. It is also capable of handling a much larger range of superposition queries and situations than many standalone programs and yields results that are intuitively more in agreement with known biological or structural data. The SuperPose web server is freely accessible at http://wishart.biology.ualberta.ca/SuperPose/.

3.2 Osmotin/ thaumatin like protein in Plants

The gene and protein sequences of osmotin and osmotin/thaumatin- like proteins were retrieved from published completed genomes of *Arabidopsis lyrata*, *Arabidopsis thaliana*, *Carica papaya*, *Medicago truncatula*, *Glycine max*, *Cucumis sativus*, *Populus trichocarpa*, *Vitis vinifera*, *Oryza sativa*, *Zea mays* and *Sorghum bicolor*. Superfamily server was used for structural domain analysis for osmotin/thaumatin- like proteins family.

All the gene and protein sequences of osmotin and osmotin/thaumatin- like proteins were retrieved from published completed genome of *Arabidopsis thaliana* using Superfamily server. Amino acid sequences of all the proteins in Fasta format was submitted to Compute pI/Mw at ExPASy Proteomics Server for calculating theoretical pI and molecular weights. The subcellular locations of all the proteins were predicted using TargetP 1.1 Server; http://www.cbs.dtu.dk/services/TargetP hosted at Technical University of Denmark, Denmark. Clustal W was performed for the amino acid sequence alignment of members osmotin/thaumatin-like proteins family from *Arabidopsis* and *Nicotiana tabacum* to find the closest homolog of osmotin in *Arabidopsis* among osmotin/thaumatin-like proteins family.

3.3 Sequence analysis, homology modelling, model optimization, quality assessment and visualization of ATOSM34

The gene and protein sequences of ATOSM34 (At4G11650) were retrieved from published completed genomes of *Arabidopsis thaliana* (The *Arabidopsis* Genome Initiative (2000)). The 3D model of ATOSM34 was prepared using Discovery Studio 2.0 (Accelrys Inc., San Diego, CA, U.S.A).
3.3.1 Discovery Studio (http://www.accelrys.com)

Discovery Studio (DS) is a complete modelling and simulation environment for life scientists. DS assist scientists in various aspect of macromolecule-based research such as alignment of multiple sequences, determination of the three-dimensional structure and properties of a macromolecule, studying macromolecular interactions with their partners, conducting forcefield based simulations etc. The alignment of sequence of ATOSM34 to 1PCV (the template) was done using Align2D available in MODELER 8.2 (Sali et al. 1995) from Discovery studio. The Model was build using the co-ordinate assignment process. The obtained model was then energy minimized using the molecular mechanics force field, CHARMM (Brooks et al. 1983) implemented under Discovery Studio 2.0. The minimization was carried out with the initial 5000 steps of steepest descent followed by 1000 steps of conjugate gradient with convergence criterion of 0.05 Kcal/(mol× Angstrom). Verification of model in terms of the amount of misplaced phi and psi angles were determined by Ramachandran plot and by examination of the structure on a graphic display. The model was evaluated by using Verrify Protein (Modeler and Profile-3D) (Luthy et al. 1992).

The theoretical model generated was also subjected to validation using the program PROCHECK for assessing the stereochemistry of the model. The program PROCHECK (Morris et al. 1992) concentrates on the parameters such as bond length, bond angle, main chain and side chain properties, residue-by residue properties, RMS distance from planarity and distorted geometry plots. It assess how normal, or conversely how unusual, the geometry of the residues in a given protein structure is, as compared with stereo chemical parameters derived from well-refined, high-resolution structures (Laskowski et al. 1993).

3.4.1 Osmotin, a cell signaling molecule, for proline biosynthesis and accumulation under stress

A thorough literature search was done to identify the stress regulatory protein molecules in plants. Transcription factors, growth regulators and some inorganic molecules have been implicated in regulation of metabolic pathways. The type and the information about such molecules are however, limited, especially for proline biosynthesis. Literature suggested osmotin share some structural features with adiponectin, an adipocyte-derived hormone that regulates metabolism of lipids and glucose in humans.
The structural coordinates of human adiponectin (1C3H) was downloaded from RCSB protein data bank. Clustal W of adiponectin and osmotin (1PCV) proteins was done to find the similarity. Literature mining suggested that adiponectin mediate its effect through its receptor ADIPOR1 and ADIPOR2. The sequences of human ADIPOR1 (NP_057083) and human ADIPOR2 (NP_078827) were downloaded from Uniprot. The search for homologues of adiponectin receptors, ADIPOR1 and ADIPOR2 was done against Arabidopsis genome using BlastP. Five proteins were found to be significant homologues. Arabidopsis hepta helical family proteins (HHP); HHP1 (At5g20270, AY143975), HHP2 (At4g30850, AY267331), HHP3 (At2g24150, AF370179), HHP4 (At4g37680, AY056348) and HHP5 (At4g38320, AY070414) were retrieved from AITGR. The amino acid sequences of Arabidopsis HHP proteins and their homologues in human were aligned with the program Clustal W with Gonnet as protein weight matrix, multiple alignment gap opening penalty as 10 and multiple alignment gap extension penalty as 0.20. Topology predictions of HHP family proteins in terms of trans-membranous helices and domain analysis were conducted by using TMHMM 2.0 (http://www.cbs.dtu.dk/services/TMHMM).

### 3.4.2 Secondary structure prediction and modelling of ADIPOR1, receptor for adiponectin

Some relationships among proteins at the fold level are readily identified due to the sequence similarities among them. However, in many cases the sequence similarities can be very low and thus relationships not obvious. It is now well accepted that conservation at the structure level can be much higher and thus more detectable than at the sequence level (Lo et al. 2007). Therefore, in absence of a clear structural homologue due to very low sequence similarity, group based alignment of secondary structure was used to find the template for modelling. The modelling of ADIPOR1 consists of four sequential steps: secondary structure assignment and template selection, target–template alignment, model building, and model assessment.

The amino acid sequence of ADIPOR1 was submitted to various online secondary structure prediction softwares viz. Jpred(Jnet,Jpssm,Jhmm) nnPredict, GOR4, HNN, SOPMA, PSIPRED, Prepro and Sspro webservers to know the secondary structure composition.
3.4.2.1 Jpred(Jnet,Jpssm,Jhmm): Jpred is a web server that takes a protein sequence or multiple alignment of protein sequences, and from these predicts secondary structure using a neural network called Jnet. Jnet (Cuff and Barton 2000) is a program for protein secondary structure prediction which uses neural network prediction algorithm that works by applying multiple sequence alignments, alongside Position specific iterated - BLAST (PSIBLAST) and hidden markov model (HMM) profiles. The prediction is the definition of each amino acid residue into either alpha helix ('H'), beta sheet ('E') or random coil ('-') secondary structures. Jnet also makes predictions of relative solvent accessibility. Additionally, Jpred makes coiled-coil predictions using the 'COILS' algorithm developed by Lupas and co-workers (Lupas et al. 1991).

3.4.2.2 NN_PREDICT Protein Secondary Structure Prediction Server: Computational neural networks have recently been used to predict the mapping between protein sequence and secondary structure. NN_PREDICT is a program that predicts the secondary structure type for each residue in an amino acid sequence by using a two-layer, feed-forward neural network (Kneller et al. 1990). It takes as input a sequence consisting of one-letter amino acid codes or three-letter amino acid codes separated by spaces. The residues are classified either as: 'H', a helix element; 'E', a beta strand element, or '-', a turn element. The program uses the tertiary class of the protein (either none, all-alpha, all-beta, or alpha/beta) for prediction. For the best-case prediction, the accuracy rate of NN_PREDICT has been reported as being over 65%.

3.4.2.3 GOR IV: The Garnier-Osguthorpe-Robson (GOR) method uses both information theory and Bayesian statistics for predicting the secondary structure of proteins (Garnier et al. 1978). Over the years, the method has been improved by including larger databases and more detailed statistics, which account not only for amino acid composition, but also for amino acid pairs and triplets. GOR IV, uses all possible pair frequencies within a window of 17 amino acid residues. The program gives two outputs, the sequence and the predicted secondary structure in rows, H=helix, E=extended or beta strand and C=coil. The second output gives the probability values for each secondary structure at each amino acid position. The GOR method takes into account not only the probability of each amino acid having a particular secondary structure, but also the conditional probability of the amino acid assuming each structure, given that its neighbour assume the same structure. This method is both more sensitive and more accurate because the amino acid structural propensities are only strong for a small number of amino acids such as proline and glycine.

3.4.2.4 HNN SECONDARY STRUCTURE PREDICTION METHOD: Hierarchical neural network (HNN) is one of the most accurate secondary structure prediction methods for the past decade and is made up of two networks; a sequence-to-structure network and a structure-to-structure
network (Guermur et al. 1999). In neural network approach, computer programs are trained to be able to recognize amino acid patterns that are located in known secondary structure and to distinguish these patterns from the other patterns which are not located in these structures. In this method a sliding window of 13-17 amino acids is moved along the sequence. The sequence within each window is read and used as input to a neural network model previously trained to recognize the secondary structure most likely to be associated with that pattern. The network then predicts the secondary structures configuration of central amino acid as α-helix, β-sheet, and coil. Rules or other trained networks are then applied that make the prediction of a series of residues reasonable.

3.4.2.5 SOPMA: Self-Optimized Prediction Method with Alignment (SOPMA) is based on the homologue method (Geourjon and Deleage 1995). The improvement takes place in the fact that SOPMA takes into account information from an alignment of sequences belonging to the same family. This tool works on the basis of neural network method. The PHD algorithm first performs a database search for possible homologous proteins, then aligns and filters the sequences to decide on the most likely homologs, and finally feeds the sequences and alignment profile to a feed-forward neural network for secondary structure prediction. SOPMA accurately predicts 69.5% of amino acid for the three states describing the secondary structure (α-helix, β-sheet and coil). The program uses window width of 17 and predicts four conformational states, namely helix, sheet turn and coil.

3.4.2.6 PSIPRED: PSIPRED is a simple and accurate secondary structure prediction method, incorporating two feed-forward neural networks which perform an analysis on output obtained from PSI-BLAST (Jones 1999). The prediction method follows three sequential steps: generation of a sequence profile, prediction of initial secondary structure, and finally the filtering of the predicted structure. PSI-BLAST constructs a profile (Position Specific Scoring Matrix, PSSM) from a multiple alignment of the highest scoring hits in an initial BLAST search. Highly conserved positions receive high scores and weakly conserved positions receive scores near zero. The profile is used to perform a second blast search and the results of each iteration used to refine the profile. The iterative searching strategy results in increase sensitivity. The method is capable of an average $Q_3$ score of 80.7%. The results are returned in graphical representation of secondary structure prediction.

3.4.2.7 Prepro (http://www.predictprotein.org): PredictProtein (Prepro) is an Internet service for sequence analysis and the prediction of protein structure and function. When users submits protein sequences or alignments, then the program returns multiple sequence alignments, PROSITE sequence motifs, low-complexity regions (SEG), nuclear localization signals, regions lacking regular structure (NORS) and predictions of secondary structure, solvent accessibility, globular regions, transmembrane helices, coiled-coil regions, structural switch regions, disulfide-bonds, sub-cellular
localization and functional annotations. Upon request, fold recognition by prediction-based threading, CHOP domain assignments, predictions of transmembrane strands and inter-residue contacts are also available.

3.4.2.8 Sspro: The Sspro combines machine learning methods, evolutionary information in the form of profiles, fragment libraries extracted from the Protein Data Bank (PDB), and energy functions to predict protein structural features and tertiary structures. The prediction is based on an ensemble of one-dimensional recursive neural network (1D-RNN) architectures. These RNN architectures are based on the theory of probabilistic graphical models (Bayesian networks) meshed with a neural network parameterization to accelerate belief propagation and learning. These architectures systematically combine standard information contained in a local input window with more distant contextual information extracted by translation invariant recursive neural networks that are convolved along the entire length of the protein (1D) or of the contact maps (2D) from all possible directions.

For selection of suitable template for 3D modelling of ADIPOR1, manual secondary structure alignment for lengths of helices and loops between the helices of ADIPOR1 with all the proteins containing single subunit and 7 helix topology from OPM database (http://opm.phar.umich.edu/) was carried out. Initial N terminal (1-134 amino acids) contained cytoplasmic domain and, therefore, was excluded from modelling while only the region from 135–355 amino acids was taken. The consensus secondary structure of ADIPOR1 was submitted to SSEA (Secondary Structure Element Alignment) (http://protein.cribi.unipd.it/ssea/) to compute the compatibility in terms of secondary structure of ADIPOR1 against the secondary structure of all the proteins containing single subunit and 7 helix topology from OPM database in one vs. one alignment mode.

3.4.2.9 SSEA (Secondary Structure Element Alignment): SSEA is a web server which computes alignments of protein secondary structures. The server supports both performing pair-wise alignments and searching a secondary structure against a library of domain folds. It can calculate global and local secondary structure element alignments. A combination of local and global alignment steps can be used to search for domains inside the query sequence or help in the discrimination of novel folds. Both SCOP and the PDB fold libraries, clustered at 95% and 40% sequence identity, are available for alignment.

The manual alignment of the secondary structure and alignment scores for secondary structure alignment from SSEA program showed that the structure of ADIPOR1 and eubacteria (Anabena nostoc sp.) sensory rhodopsin (PDB code: 1XIO) have similar lengths.
Chapter 3  Materials and Methods

of helices and loops between the helices. Therefore, this protein was used as the template for predicting the three-dimensional structure of ADIPOR1 by comparative modelling strategy. The protein sequence of ADIPOR1 was submitted to Discovery studio 2.0 (DS) (Accelrys Inc., San Diego, CA, U.S.A). The alignment of sequence of ADIPOR1 to 1XIO (the template) was done using Align2D available within Modeller 8.2 (Sali et al. 1995) from Discovery studio 2.0. The Model was build using the co-ordinate assignment process. Energy minimization of the ADIPOR1 was carried out using the molecular mechanics force field CHARMM, an implicit distance-dependent dielectric constant and a non-bonded atom cut-off 14 angstrom, by the Adopted Basis Newton-Raphson energy minimization (Brooks et al. 1983), which is implemented under DS as well. The minimization was carried out with the initial 5000 steps of steepest descent followed by 1000 steps of conjugate gradient with convergence criterion of 0.05 Kcal/(mol× Angstrom). Verification of model in terms of the amount of misplaced phi and psi angles were determined by Ramachandran plot and by examination of the structure on a graphic display. The model was evaluated by using Verify Protein (Modeler and Profile-3D) (Luthy et al. 1992).

The theoretical model generated was also subjected to validation using the program PROCHECK for assessing the stereochemistry of the model. The program PROCHECK (Morris et al. 1992) concentrates on the parameters such as bond length, bond angle, main chain and side chain properties, residue-by-residue properties, RMS distance from planarity and distorted geometry plots. It assess how normal, or conversely how unusual, the geometry of the residues in a given protein structure is, as compared with stereo chemical parameters derived from well-refined, high-resolution structures (Laskowski et al. 1993). The model generated was then taken for the subsequent analysis and docking studies.

3.4.3 Secondary structure Prediction and modelling of HHP1, a homologue of ADIPOR1

The amino acid sequence of HHP1 was submitted to various online secondary structure prediction softwares similar to ADIPOR1: Jnet, GORIV, HNN, SOPMA, PSIPRED, Prepro webservers to know the secondary structure composition.

The selection of suitable template for modelling of HHP1 was done by similar method as used for ADIPOR1. Manual secondary structure alignment for lengths of helices and loops
between the helices of HHP1 with all the proteins containing single subunit and 7 helix topology from OPM database (http://opm.phar.umich.edu/) was carried out. Initial N terminal (1-96 amino acids) containing the cytoplasmic domain was excluded from modelling. The consensus secondary structure of HHP1 was submitted to SSEA (Secondary Structure Element Alignment) (http://protein.ccri.unipd.it/ssea/) to compute the compatibility in terms of secondary structure of HHP1 against the secondary structure of all the proteins containing single subunit and 7-helix topology from OPM database in one vs. one alignment mode.

As in case of ADIPOR1, the manual alignment of the secondary structure and alignment scores for secondary structure alignment from SSEA program showed that the structure of HHP1 and eubacteria (Anabena nostoc sp.) sensory rhodopsin (PDB code: 1XIO) have similar lengths of helices and loops between the helices. Therefore, this protein was used as the template for predicting the three-dimensional structure of HHP1 by comparative modelling strategy.

The protein sequence of HHP1 was submitted to Discovery studio 2.0. The alignment of sequence of HHP1 to 1XIO (the template) was done using Align2D available within Modeler 8.2 (Sali et al. 1995) from Discovery studio 2.0. The Model was build using the co-ordinate assignment process in Modeler. Energy minimization of the HHP1 was carried out using the molecular mechanics force field CHARMM (Brooks et al. 1983), which is implemented under DS as well. The energy minimization was carried out with the initial 5000 steps of steepest descent followed by 1000 steps of conjugate gradient with convergence criterion of 0.05 Kcal/(mol× Angstrom). Verification of model in terms of the amount of misplaced phi and psi angles were determined by Ramachandran plot and by examination of the structure on a graphic display. The model was evaluated by using Verrify Protein (Modeler and Profile-3D) (Luthy et al. 1992).

The theoretical model generated was also subjected to validation using the program PROCHECK for assessing the stereochemistry of the model. The model generated was then taken for the subsequent analysis and docking studies.

3.5 Molecular docking of osmotin-HHP1 and ATOSM34-HHP1 complex

It has been shown that the osmotin can activate AMP kinase in C2C12 myocytes and the suppression of ADIPORs expression by siRNA markedly reduced phosphorylation of AMP
kinase induced by osmotin (Narasimhan et al. 2005). These data suggested that the osmotin activates AMP kinase via ADIPORs in mammalian C2C12 myocytes. Further to explore the possibility of osmotin and ATOSM34 using HHP1 as a receptor to transduce the signals inside the cell and to study the nature of interaction among the two, we modelled the complex osmotin-HHP1 and ATOSM34-HHP1. The 3D co-ordinates of osmotin and HHP1 and ATOSM34 and HHP1 were uploaded in Discovery studio 2.0.

3.5.1 ZDOCKpro (Discovery studio): ZDOCKpro is a protein-protein docking method developed by Professor Zhiping Weng at Boston University and is included in Discovery studio 2.0 (DS) (Accelrys Inc., San Diego, CA, U.S.A). The program includes two main algorithms called ZDOCK (for fast, rigid-body, initial stage docking using pairwise shape complementarity) and RDOCK (for CHARMM-based refinement of the complex poses generated by ZDOCK). RDOCK also ranks the docked structures based on CHARMM electrostatic interaction energy and ACE desolvation energy.

3.6 Transduction of signal through the amino (N) terminal domain of HHP1, originated due to the interaction of ligand receptor

Amino acid sequence of N terminal cytoplasmic domain of HHP1 (1-96 aas) was submitted to Group-based phosphorylation scoring method (GPS) (Xue et al. 2005) in FastA format.

3.6.1 Group-based Phosphorylation Scoring Method (GPS): In the eukaryotic cells, protein phosphorylation is one of the most ubiquitous posttranslational modifications of proteins, orchestrating most of the cellular processes, including the cell cycle (Lou et al. 2004), transcriptional (Uddin et al. 2003) and translational regulations (Yoshizawa et al. 2002), metabolic pathways (Meijer and Dubbelhuis 2004), and signal transductions (Choudhary et al. 2004). In silico prediction of phosphorylation sites with their specific kinases may help and alleviate the labor-intensive in vivo or in vitro identification of phosphorylation sites greatly. GPS method is based on the observation that for two peptides with only one pair of different amino acids according to their positions, it may be assumed with confidence that they have similar 3D structures and biochemical characteristics, especially when the two different amino acids are a conserved pair, e.g. isoleucine (I) and valine (V), or serine (S) and threonine (T); thus, making information more meaningful to biologists. With data from the public database Phospho.ELM/PhosphoBase (Diella et al. 2004; Kreegipuu et al. 1999) and extensive literature curation, the method covers 71 Protein Kinase (PK) families/PK groups.
3.7 Cytosolic and nuclear targets for extracellular signal-regulated kinase (ERK) in proline biosynthesis.

3.7.1 Identification of nuclear targets (transcription factors) to be phosphorylated by extracellular signal-regulated kinase (ERK) in proline biosynthesis.

Based on the literature and genome information available on proline metabolism pathway in *Arabidopsis thaliana*, a set of seven genes was identified (Table 23). The promoter sequences of these seven genes were retrieved after mapping the Locus IDs to the ATIGRS gene identifiers.

3.7.1.1 Promoter sequence analysis of differentially expressed genes of proline biosynthesis

Promoters sequence analysis of differentially expressed genes of proline metabolism was done using ExPlain Plant Analysis system and BIOBASE Knowledge Library Plant Edition (BKL-Plant). The promoter window of -1,000 to +100bp was uploaded for all the genes into ExPlain PlantAnalysis System. The F-Match module was used to identify transcription factor binding sites overrepresented in differentially expressed proline biosynthesis gene sets against a background set of 200 ubiquitously expressed genes. The composite module analysis (CMA) was then used to determine which combination of binding sites or composite module was most commonly found within the sets of genes. Matrices with a Yes/No score >1.3, p-value <0.05, and Matched promoters p-value <0.1 from F-Match analyses were selected for the composite module analysis.

3.7.1.1.1 ExPlain™ Plant Analysis Platform:

ExPlain™ is a unique data analysis system that combines promoter and pathway analysis tools. Using the power of TRANSFAC®’s transcription factor binding site derived positional weight matrices, ExPlain™ identify transcription factors affecting gene expression in the test set, as well as predict how they, in combination, can induce changed gene expression patterns. It is an integrated computational tool interpretation of gene expression data. It analyzes microarray data and proposes complexes of transcription factors as well as “upstream” key signaling molecules that master the observed gene expression profile. The method utilizes data from (TRANSFAC® (Matys et al. 2006) and TRANSPATH® (Krull et al. 2006) databases. Composite Module Analyst (CMA) analyzes 5’-upstream regions of co-expressed genes and applies a genetic algorithm to reveal composite modules (CMs) consisting of co-occurring single TF binding sites and composite elements (Kel et al. 2006;
Waleev et al. 2006). The Match™ program of ExPlain™ predicts binding sites in the test sequence sets. The Match results were then subjected to F-Match analysis. The F-Match algorithm is designed to discover transcription factors whose binding sites are enriched in a sequence set of interest, the so-called “Yes” set, compared to a sequence set that represents genomic background, the so-called “No” set. The CMA tool of ExPlainTM searches for combinations of transcription factor binding sites that best discriminate a Yes sequence set against a No sequence set. The ExPlain™ Plant Analysis System offers a unique new tool for plant researchers that allows quick mapping and identification of transcription factor binding sites within large promoter sets, identifies combinations of binding sites overrepresented in a promoter set, Provides quick access to information on matrices and their binding factors through BKL-Plant, allowing assessment of potential new gene targets.

The composite module showed the presence of two TF binding sites matching matrices; Opaque2 (O2_Q2) and Ocs element binding factor (OCSBF-1) in the proline metabolism gene set. BlastP was performed against A. thaliana genome, after downloading the amino acid sequences of Opaque2 (O2_Q2) and Ocs element binding factor (OCSBF-1) from BKL-Plant.

3.7.2 Determination of the nuclear and cytosolic targets of extracellular signal-regulated kinase (ERK) in proline metabolism

The available programs such as GPS, Kinophos, etc. were able to find the sites for MAP kinases on amino acid sequence of proteins but not ERKs. Therefore, amino acid sequences of all the enzymes of proline biosynthesis (Table 23) and identified transcription factors AtbZIP10, AtbZIP25, AtbZIP53, AtbZIP2, MYBAS1 and GT-1 were manually searched for ERK phosphorylation sites, and docking motifs on the substrates. Finally a model was proposed, compiling the data generated from the study, explaining the behaviour of osmotin as cell signaling molecule, especially in biosynthesis and accumulation of proline in plants experiencing stress conditions.