1.1. Introduction

The whole genome consists of several thousand genes. Every gene needs to be expressed and regulated to do their functional roles. The organisms utilize various strategies in regulating the gene expression. Gene expression is a process that encodes DNA to its corresponding functional proteins. The following equation is vital in the mechanism of gene expression,

$$\text{DNA} \xrightarrow{\text{Transcription by RNA polymerase}} \text{RNA} \xrightarrow{\text{Translation by Ribosome}} \text{Protein}$$

Gene regulation is a mechanism that increases/decreases the expression level of a given gene. The process is controlled at multiple levels, mostly during transcription and translation processes. The transcription and translation are physically coupled in prokaryotes, but in eukaryotes these mechanisms are controlled in different compartments, namely nucleus and cytoplasm (Glisovic et al., 2008). Every operon of gene consists of many regions, the so called promoter, operator, structural gene etc. The operon model is shown in Fig. 1.1.

In bacteria, gene expression is regulated by regulatory proteins, which are of two types: positive regulators (activators) and negative regulators (repressors). Fig. 1.2 shows the mechanism of positive and negative regulating proteins.
**Fig. 1.1.** The *lac* operon is shown as a representative model of gene.

(The image was adapted from http://faculty.quinnipiac.edu/health/biology/buckley/bi_571/control_prok_genes/sld017.htm)

**Positive and negative regulation**

**Fig. 1.2.** The mechanism of regulatory proteins

(The picture was adapted from http://www.discoveryandinnovation.com/BIOL202/notes/lecture17.html).
The small RNA (sRNA) and RNA binding proteins are fundamental regulators in post-transcriptional gene regulatory mechanism. The RNA binding proteins bind with mRNA and control multiple aspects ranging from transcription, splicing, polyadenylation, mRNA stabilization, mRNA localization and translation.

1.2. RNA binding proteins

RNA binding proteins interact with RNA either by sequence specific or structure dependent. It can bind with single or double stranded RNA. RNA binding proteins have multiple structural domains, which are mediating the binding capability with various nucleic acid (DNA/RNA) components. The proteins contain the following domains: (i) RNA recognition motif (RRM), (ii) double-stranded RNA binding domain (dsRBD), (iii) K homology domain (KH), (iv) zing fingers domain (ZnF), (v) Pumilio homology domain (PUM-HD) and (vi) cold-shock domain (Glisovic et al., 2008).

RNA binding proteins play a central role in gene regulation. It is a crucial mediator of gene expression in both prokaryotes and eukaryotes. It selectively regulates the gene expression. The fundamental mechanism of gene regulation mediated by RNA binding protein is similar to gene regulation using transcription factors where the transcription factors bind with upstream DNA motif of gene. But in the case of RNA binding proteins, these regulatory proteins bind to the sequence/structural features of mRNA.

The regulatory proteins interact/bind with the nascent RNA transcript and modulating the activity of terminator, which is located in between the promoter and upstream genes of the operon (Manival et al., 1997).
1.3. Gene regulation by RNA binding proteins

Bacteria utilize multiple mechanisms to control the expression of genes with respect to environmental conditions. Expression of bacterial gene in operons is controlled at the transcriptional and translational levels. The site responsible for the transcription mechanism of operon is located within or at the end of the operons. RNA binding proteins regulate gene expression by two mechanisms; it acts as a regulatory molecule (activator/repressor) or by modulating the structure of the corresponding mRNA. The expression of bacterial genes is controlled by regulatory proteins through positive (antitermination) and negative regulations (attenuation) (Rutberg, 1997; Yanofsky, 2000).

Antitermination/attenuation mechanism of gene regulation for TRAP, *hutP*, *pgR*, *bglP*, *sacT*, *glpP*, *licT*, *SacY*, *amir* and *nasR* systems are reported earlier (Rutberg, 1997). In antitermination mechanism, the regulatory proteins/activators bind with specific sequence of mRNA at the region of terminator loop and allow the transcription to the level of downstream genes. It can happen in two ways, (i) modification of RNA polymerase and (ii) destabilizing the terminator structure. But in attenuation, the presence of regulatory proteins/repressors, the transcription is stopped/paused before the full length of transcription occurs.

HutP is a regulatory protein belongs to *hut* operon gene. When the hexameric form of HutP binds with its mRNA in the presence of histidine and divalent metal ions, the terminator loop is destabilized and regulates the expression of its structural genes (Fig.1.3A). The TRAP is the best example for attenuation process, which regulates *trp* operon. In the presence of tryptophan, 11 subunits of TRAP bind with *trp* leader mRNA and induce the terminator (Fig.1.3B).
Fig. 1.3. The antitermination/attenuation mechanism of gene regulation. A. Positive gene regulatory mechanism of HutP. B. The negative gene regulation by TRAP protein.

1.4. HutP-an RNA binding gene regulatory protein

Histidine is an important and essential amino acid, synthesized in cells involving different steps of biochemical mechanism. When nitrogen atoms and carbohydrates are lacking in their sources, the hut operon gets activated and utilizes the freely available histidine as nitrogen source; Oda and his coworkers initially characterized and cloned the hut genes from *B. subtilis* (Oda, Sugishita and Furukawa, 1988). **Histidine Utilizing Protein (HutP)** is an RNA binding protein, which controls the gene expression by antitermination mechanism. HutP is a positive gene regulatory protein. The hut operon has six genes wherein hutP is placed at the promoter region and the other five structural genes (hutH, hutU, hutL, hutG, & hutM) are located in the downstream region (Oda et al., 2000). The nucleotide sequence of hut mRNA has a termination/antitermination loop region located in between the hutP and its structural genes (Wray and Fisher, 1994; Oda et al., 2004). The formation of terminator hairpin structure and destabilization depend on protein-RNA interactions. The investigation reveals that HutP regulates the expression of downstream genes by antitermination mechanism when it binds with UAG sequence of mRNA with the presence of histidine (Oda et al., 2004).

HutP is capable of destabilizing the terminator structure when it binds with single stranded RNA. Kumarevel and his coworkers solved the first crystal structure of HutP from *Bacillus subtilis* (Kumarevel et al., 2004). In order to understand the structural features and mechanism behind the gene regulation of HutP protein, different complexes such as native HutP (1WPS), HutP-ligand (1VEA), HutP-ligand-metal (1WPV, 1WRN, 1WRO and 4H4L), and HutP-ligand-metal-RNA (1WPU, 1WMQ, 1WRQ and 3BOY) were prepared and solved. The quaternary structure of the HutP is shown in Fig. 1.4.
Fig. 1.4. Structure of activated HutP complex
HutP belongs to $\alpha/\beta$ family of proteins. The structures are arranged with four $\beta$ strands in the center and four $\alpha$ helices are in front ($\alpha_1$, $\alpha_2$) and back ($\alpha_3$, $\alpha_4$) of the strands. The strands and helices are connected through five loops. Each L-histidine and metal ions bind with individual monomeric unit. Hexameric form of HutP binds with specific sequence (UAG) of hut mRNA. HutP not only requires L-histidine for RNA binding but also the divalent metal ions like Mg$^{2+}$, Mn$^{2+}$, etc (Kumarevel, Mizuno and Kumar, 2005).

In the apo-form, HutP crystallized as a dimer and three dimers were related by 3-fold symmetry to form a hexamer. In binary complex, HutP recognizes the L-histidine at the dimer interface by forming an open hydrophobic pocket with its surrounding residues. This hydrophobic pocket is essential to clarify whether the incoming residue possess the imidazole group or not. HutP-L-histidine binding with mRNA confirms that the L-histidine analogs with imidazole ring possess higher affinity for binding. Once the ternary complex (HutP-L-histidine-Mg$^{2+}$) formation is achieved, the metal ion makes the structural rearrangement in L3 and L5 loop regions and L-histidine binding sites in the structure. Conformational changes are induced in the protein structure during ligand and metal binding which play a vital role in RNA binding of HutP (Kumarevel, Mizuno and Kumar, 2005).

The studies indicate that the termination occurs in hut mRNA structure between hutP and hutH genes within hut operon at +498 to +572 (Oda et al., 2000; Kumar, Kumarevel and Mizuno, 2006). Upon binding to HutP, RNA terminator structure is destabilized and expression of downstream genes of hutP is regulated using antitermination mechanism.
1.5. Steps followed for the experiments

The following steps are performed to express, purify, crystallize and solve the structure of HutP from *Geobacillus thermodenitrificans*.

- Gene synthesis
- Subcloning
- Transformation
- Plasmid preparation
- Expression
- Protein purification
- Crystallization
- Data collection
- Structure solution & Refinement
1.5.1. Gene synthesis

Gene synthesis is an alternative method to synthesize the gene of our interest by chemical method instead of cloning (expressing) from the genomic plasmid using primers.

1.5.2. Subcloning

Cloning is a widely used molecular biology technique, used to transfer the gene of our interest to the targeted vector. Subcloning is a method to re-clone a gene to the expression vector of our interest. In our study, the gene was chemically synthesized and cloned into vector and subcloning was done using the TA cloning method. The restriction enzymes are used to digest the gene of our interest from the cloned vector. The digested product was purified by gel extraction and purification methods using commercially available kits. The digested product then ligated to expression vector of our interest by ligation procedure.

1.5.3. Agarose gel electrophoresis

Agarose gels are used for separation and identification of nucleic acids. The gel was prepared by mixing agar powder with TAE buffer. When the electric field was applied, the negatively charged DNA moves towards the anode. The movements of DNA/RNA samples are controlled based on the molecular weight of molecule and pore size of the gel.

1.5.4. Plasmid preparation and expression

The ligated product of gene is used for transformation into DH5α cells. The transformed cells are grown in agar plates. The well-grown individual colonies are
collected and inoculated into the 5ml culture for further growth. The cells are harvested by centrifugation and the DNA is isolated by plasmid preparation protocols.

1.5.5. DNA sequencing

The isolated DNA is further confirmed through DNA sequencing. The capillary electrophoresis method is used for DNA sequencing. The experiment is performed using Applied Biosystem instrument. The results are analyzed through FinchTV software program.

1.5.6. Transformation

The cloned template is used for transformation into BL21 (DE3) expression vector. The transformed template is grown in agar plates. The colonies are used for culture preparation. The overnight culture is harvested by centrifugation. The pellet is sonicated in the presence of sonication buffer and the protein expression is analyzed through SDS-PAGE.

1.6. Protein purification

The impurities and co-expressed proteins are removed by various methods to achieve the highest purity of protein possible. Chromatographic technique is used to separate and purify the biological and synthetic compounds. Here in, the stationary and mobile phase are used to separate the materials. The stationary phase is made up of organic materials; the buffer and other organic solvents are used in mobile phase.

In protein biochemistry, we used the following techniques to separate the biological molecules.
1.6.1. Ion exchange chromatography

Ion exchange chromatography is used to purify the biological samples based on their charges. Ion exchange chromatography can be classified into two types based on their charge. The positive charged protein molecules bind with negative charged resin and was eluted by adding of suitable buffer samples, this method is named as “cation exchange”. However the exchange of negatively charged molecule with positive charged resin called as anion exchanger. The charge of amino acids, resin, buffer composition and pH of protein samples play a major role on ion exchange chromatography.

1.6.2. Gel filtration chromatography

This method is known as size exclusion chromatography. The separation of biological samples is done based on their molecular weight. The stationary phase materials are made up of pores. When the protein sample is applied to column, the protein molecules tend to penetrate the pores of the stationary phase with respect to their size. While adding the buffer, the lower molecular weight proteins are eluted initially followed by higher molecular weight proteins.

1.6.3. Affinity chromatography

Affinity chromatography is a specific method used to separate the biological samples based on their specific interactions. The metal ions/antigens/receptors and ligands are used as affinity materials. Ni-Affinity chromatography: The Ni-NTA (nitrilotriacetic acid) column is used to separate the tagged proteins, the stationary phase is made up of Ni$^{2+}$ ions. The His-tag fused protein strongly binds with Ni-NTA column. When the buffer is added to the column, unbounded proteins are easily eluted.
But the protein bound with Ni-NTA is eluted by adding of buffer containing imidazole. This affinity chromatography method is used to purify the fused proteins.

1.6.4. SDS-PAGE

The sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) is a popular technique used in molecular biology to separate and characterize the protein molecules based on their size. SDS non-covalently binds with proteins, denatures and mask the intrinsic charge of proteins. The tank buffer used to conduct the electricity during electrophoresis. Once the electric charge is applied, the proteins tend to move towards positive charge electrode. The movement of protein samples are visualized by Coomassie Brilliant Blue staining (CBB). The standard molecular markers are used to identify the molecular weight of the proteins.

1.6.5. Protein sequencing

The protein sequencing is a method used to identify the amino acids present in the unknown proteins. The sequence from N-terminal regions is identified from this method. The experiment is performed by Edman degradation method, using PROCISE protein sequencer for the purified protein sample.

1.6.6. Mass spectroscopy analysis

The matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) is a spectroscopy method to identify the molecular weight of unknown biological samples. The spectrum is recorded based on the mass of the samples. The Voyager-DE pro (Applied Bio System) machine is used for MALDI-TOF spectrum experiment.
1.7. Protein Crystallography

Protein crystallography is the field of science to study the structural and conformational details of the biological samples. Three dimensional structures of the biological molecules are determined by the following steps.

1.7.1. Crystallization

The achievement of meta stable supersaturated solution with stable lower energy state by reduction of solute concentration is called “crystallization”. The nucleation and crystal growth are the two major factors contribute in the formation of crystals. The protein is crystallized using any of the following variable methods, hanging drop vapour diffusion method, sitting drop vapour diffusion method, sandwich drop, and batch methods. The principle of crystallization and the various crystallization methods are shown in Figs.1.5 and 1.6.

![Fig.1.5. The steps involved in the formation of biological crystals.](http://www.mdpi.com/1422-0067/14/6/11643/htm)
Fig.1.6. Various crystallization methods are shown. A. Hanging drop vapour diffusion method, B. Sitting drop vapour diffusion method, C. Sandwich method and D. Batch method.

(The images adapted http://www-structmed.cimr.cam.ac.uk/Course/Crystals/Theory/methods.html)
1.7.2. Phase problem

The phase problem in X-ray crystallographic structure determination can be solved through techniques like Molecular Replacement (MR), Multiple Isomorphous Replacement (MIR), Multi-wavelength Anomalous Dispersion (MAD) or Single wavelength Anomalous Dispersion (SAD) etc. The molecular replacement technique is used to solve the structures presented in the thesis. Among the various methods MR is the simplest technique. A successful MR solution requires a protein model, with high structural similarity of the protein of interest. The programs AMORE (Navaza, 1994), MOLREP (Vagin and Teplyakov, 1997) and PHASER (McCoy et al., 2007) are used for solving the crystal structures in the present studies.

1.7.3. Structure refinement

The refinement of the structures is carried out using the program REFMAC (Murshudov, Vagin and Dodson, 1997). It can carry out rigid body, restrained or unrestrained refinement against X-ray data or idealization of a macromolecular structure. It minimizes the coordinate parameters, with the help of predefined libraries, to satisfy a maximum likelihood residual. It is a user friendly as one can generate the geometric restraints prior to running REFMAC. REFMAC also produces MTZ output file containing weighted coefficients for σA weighted mFo-nFc maps.

1.7.4. Electron density map

After each stage of refinement, the model is examined and manual rebuilding is done using the electron density maps with Fourier coefficients (2|Fo|- |Fc|) and (|Fo|- |Fc|). The two types of maps are contoured to the level of 1σ and 3.0σ, respectively where σ refers to the rms deviation in the mean density in electrons/Å³. Regions of poor electron densities are examined with the maps contoured at a lower σ
level. The molecular modeling package COOT (Emsley and Cowtan, 2004) is used to examine and interpret the model against the electron density maps.

### 1.7.5. Identification of solvent sites

Solvent sites are identified when the R-factor converged to a value of around 25 to 30%. The water molecules are manually picked using both (2|Fo|- |Fc|) and (|Fo|- |Fc|) maps contoured at 1σ and 3σ, respectively. Their positions are verified by subsequent refinements and map fitting. Several rounds of this procedure are carried out until most of the densities in the maps are accounted for along with the convergence values.

### 1.7.6. Structure validation

The stereochemical properties of the structure are analyzed by using PROCHECK program (Laskowski et al., 1993). It provides Ramachandran plot (Ramachandran, Ramakrishnan and Sasisekharan, 1963) along with a number of stereochemical parameters that can be used to validate the quality of the final model.

### 1.7.7. Structure visualization and molecular graphics

The programs widely used in the thesis are COOT (Emsley and Cowtan, 2004) for structure visualization and modeling, PyMOL (DeLano, 1998), CCP4mg (McNicholas et al., 2011) and Chimera (Pettersen et al., 2004) for graphical visualization and to produce high quality images.