Introduction
1.1 Introduction

Microbes make up about 60% of the earth's biomass, yet less than 1% of microbial species have been identified. Microbes have been found surviving and thriving in an amazing diversity of habitats, in extremes of heat, cold, radiation, pressure, salinity, and acidity, often where no other forms of life could exist. The diversity and range of their environmental adaptations indicate that microbes have long ago "solved" many problems for which scientists are still actively seeking solutions. Identifying and harnessing their unique capabilities, which have evolved over 3.8 billion years, will offer us new solutions to longstanding challenges in environmental and waste cleanup, energy production and use, medicine, industrial processes, agriculture, and other areas. Comparisons of the genomes of organisms from all three domains (plants, animals, bacteria) are helping researchers in understanding the evolution of life.

1.2 Functional Genomics

The 20th century has seen a remarkable number of inventions and technological advances in virtually all aspects of human life and health care. Many areas of biomedical research have made great strides in unraveling the cause of human disease and in developing new therapies to counter, or at least improve, outcome from disease. However, the cause of the vast majority of common diseases remains poorly defined. In the final year of the millennium, the release of the draft sequence of the human genome promises to bring in a new era for basic science research and, hopefully, unprecedented growth in our understanding of human
disease. For this to occur there is a critical need to annotate the genomic sequence with gene function and basic biology. The genome describes an organism's genetic content. One major goal of functional genomics is to understand the functions of every gene in the genome. Functional genomics combines high throughput experimental methodologies with statistical and computational analysis of the results to study genes or proteins in a systematic fashion (Hieter and Bogusky, 1997).

Functional genomics is a relatively new field of molecular biology that studies how genomic information defines the functions of proteins in living organisms. The complexity of the domain that functional genomics examines and the amount of data this field produces demand that we adopt powerful computational techniques for data analysis. High-throughput genome sequencing and assembly techniques, together with new information resources, such as structural proteomics, interactomics, transcriptome data from microarray analyses, or light microscopy images of living cells have led to a rapid increase in the amount of data available (Kitano, 2002; Tong, 2004).

Bioinformatics is responding to the challenge with new integrated management systems for data collection, validation and analysis. However, the predictive power of bioinformatic analysis is limited, and putting the role of genes into a biological context will require more definitive functional approaches. Multiple alignments of genomic and protein sequences provide an ideal environment for the integration of this mass of information. In the context of the
sequence family, structural and functional data can be evaluated and propagated from known to unknown sequences. Multiple sequence alignments play a central role in a wide range of applications, including in-depth database searching, functional residue identification, structure prediction techniques and of course, evolutionary studies.

Functional genomics is often described as one of the most important challenges in the post-genomics era. Now many genomes have been sequenced and the next step is to understand the functions of all the gene-products. The innovation in functional genomics is to extend the approach from studying single gene and protein to examining in a systemic way, the functional networks that genes and proteins form in a cell. As the number of bacterial genomes with known DNA sequences increases, the important problem of assigning functional roles to predicted features of these genomes is becoming more obvious. Some of this assignment, such as predicting open reading frames (ORFs) or finding similarities between previously characterized sequences and those in the target organism, is currently done via various bioinformatics tools.

Traditionally, computational methods used to assign protein function are based on the simple assumption that proteins with similar sequences (homologs) perform similar molecular function. In such homology-based methods, function of a protein is inferred by comparing its sequence against a database such as the nonredundant database or Swiss-Prot using powerful tools such as PSI-BLAST to pick up homologous protein (Aravind and Koonin, 1999). Unfortunately, many
ORFs have no similarity to proteins with known function and, even when a match can be made, obtaining bioinformatic predictions specific enough to fit ORFs into a biological context can be difficult, especially in the situation where many proteins are predicted to have similar and related biochemical functions. One step in understanding the specific functions of ORFs would be to clone them in a way that facilitates further analysis.

The ability to predict protein function from amino acid sequence is a central research goal of molecular biology. Such a capability would greatly aid the biological interpretation of the genomic data and accelerate its medical exploitation. Basic recombinant technology provided an early route to the cloning of genes of known function. Such genes were cloned by hybridization screening using oligonucleotide probes, designed on the basis of partial protein sequence. These gene products formed the first wave of candidate drugs derived from studies of the human genome. Epogen, human erythropoietin is a successful gene product drug and global portfolio of such drugs will continue to grow as human and model organism sequence data drive the expansion of fundamental biological knowledge. Functional genomics has simplified the cloning of known and/or unknown genes, just by designing two primers from the complete genome sequence available.

The availability of complete genome sequences is driving a major breakthrough in fundamental biology as researchers compare entire genomes to gain new insights into evolutionary, biochemical, genetic, metabolic, and physiological pathways. The first complete genome sequence of a free-living
organism (*Hemophilus influenzae*) was published by Fleischmann *et al.*, in 1995, who employed a strategy of random whole-genome ‘shotgun’ sequencing. Since then it became possible to sequence the entire genomes of prokaryotes with great rapidity and efficiency and now there are 411 complete genome sequences available in genome databases and 681 are in progress. This wealth of genome data has enabled the development of comparative genomics and functional genomics approaches to investigate the biology of these organisms.

The complete genome sequence (http://ecoli.aist-nara.ac.jp) of the model bacterium *Escherichia coli* (Blattner *et al.*, 1997) has revealed nearly 2000 open reading frames (ORF’s) that remain to be functionally characterized (Mori *et al.*, 2000). Functional analysis of most of the genes can be predicted from homology analysis. But prediction of function by homology analysis alone could not define the function of ~30% of the genes.

Recently 25 strictly orphan ORFs of *E. coli* were tested under two growth conditions, i.e., exponential and stationary phases. Transcripts were identified for a total of 19 orphan genes, with 2 genes found to be expressed in only one of the two growth conditions (Alimi *et al.*, 2000). These results suggested that a vast majority of *E. coli* ORFs presently annotated as “hypothetical” correspond to bonafide genes. By extension, this implies that randomly occurring “junk” ORFs have been actively counter selected during the evolution of the dense *E. coli* genome. Primary genome annotation summary of some bacteria are given in table 1.1.
Table 1.1: Primary genome annotation summary of some bacteria

<table>
<thead>
<tr>
<th>Primary annotation summary</th>
<th>B. anthracis Ames</th>
<th>B. subtilis 168</th>
<th>E. coli K12</th>
<th>M. tuberculosis H37Rv</th>
<th>S. aureus MRSA 252</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total genes</td>
<td>5636 100.00%</td>
<td>4100 100.00%</td>
<td>4289 100.00%</td>
<td>3918 100.00%</td>
<td>2744 100.00%</td>
</tr>
<tr>
<td>Protein coding genes</td>
<td>5636 100.00%</td>
<td>4100 100.00%</td>
<td>4289 100.00%</td>
<td>3918 100.00%</td>
<td>2744 100.00%</td>
</tr>
<tr>
<td>Genes assigned a role</td>
<td>3455 62.72%</td>
<td>2572 62.73%</td>
<td>2052 47.84%</td>
<td>2069 52.80%</td>
<td>1926 70.18%</td>
</tr>
<tr>
<td>hypothetical genes (role not assigned)</td>
<td>2181 37.28%</td>
<td>1528 37.27%</td>
<td>2237 52.16%</td>
<td>1849 47.20%</td>
<td>818 29.82%</td>
</tr>
</tbody>
</table>

1.3 *Bacillus subtilis* Functional Genomics

*Bacillus subtilis* is an endospore forming gram-positive bacterium which has served as a model for studies in biochemistry, genetics and molecular biology and is important for the environment, medicine and industry. The complete genome sequence of this bacterium, comprising 4,214,810 base pairs, has been determined (Kunst *et al.*, 1997) and it comprises of approximately 4100 putative protein-coding sequences. Of the 4,000 putative protein coding sequences (CDSs) covering 87% of the genome sequence, 78% of the genes start with ATG, 13% with TTG and 9% with GTG, which compares with 85%, 3% and 14%, respectively, in *E. coli*. CDSs of *B. subtilis* can be separated into three well-defined classes. Class 1 comprises the majority of the *B. subtilis* genes (3,375 CDSs), including most of the genes involved in sporulation. Class 2 (188CDSs) includes genes that are highly expressed under exponential growth conditions, such as genes encoding the transcription and translation machineries, core intermediary metabolism, stress proteins, and one-third of genes of unknown function. Class 3 (537 CDSs) contains
a very high proportion of genes of unidentified function (84%), and the members of this class have codons enriched in A +T residues.

Possible function was defined for about 2,300 genes, based on genetic and biochemical studies and sequence similarities with known proteins. However, no functional information was obtained from the sequence for the remaining 1,800 genes, including those conserved in a variety of organisms. The function as predicted by similarity with those found in other organisms may not be exerted in *B. subtilis*. Kunst *et al.*, (1997) assigned known function to 58% of the *B. subtilis* proteins. For up to 42% of the gene products, the function cannot be predicted by similarity to proteins of known function: 4% of the proteins are similar only to other unknown proteins of *B. subtilis*; 12% are similar to unknown proteins from some other organism; and 26% of the proteins are not significantly similar to any other proteins in databanks. Only, 1,200 gene functions (30%) have been experimentally identified in *B. subtilis*.

In order to identify a minimal gene set required to sustain bacterial life in nutritious conditions, Kobayashi *et al.*, (2003) carried out a systematic inactivation of *B. subtilis* genes. Among 4,100 genes of the organism, only 192 were shown to be indispensable. Another 79 genes were predicted to be essential. The vast majority of essential genes were categorized in relatively few domains of cell metabolism, with about half involved in information processing, one-fifth involved in the synthesis of cell envelope and the determination of cell shape and division, and one-tenth related to cell energetics. Only 4% of essential genes encode
unknown functions. Most essential genes are present throughout a wide range of bacteria, and almost 70% can also be found in archaea and eucarya.

1.3.1 Signal Transduction

Tyrosine phosphorylation plays a big role in signal transduction and bacterial physiology in *B. subtilis*. Mijakovic (2003) described the first complete system involving a protein tyrosine kinase (YwqD), a protein tyrosine phosphatase (YwqE), and their corresponding substrates (YwqF and TuaD), all involved in synthesis of acidic polysaccharides, such as teichuronic acid. *B. subtilis* homologues of the eukaryotic low-molecular-weight protein tyrosine phosphatases (LMPTPs) were also described (Mijakovic et al., 2005). The YfkJ and YwIE enzymes were shown to have phosphatase activity against pNPP (p-nitrophenyl phosphate) and were sensitive to phosphatase inhibitors in a manner similar to that of the eukaryotic LMPTPs. Madec et al. (2003) demonstrated that PrkC Ser/Thr kinase involved in development, biofilm formation, and swarming motility can autophosphorylate on eight distinct Thr residues, four of which, located on the activation loop are essential for kinase activity.

Two-component signal transduction systems are an important mechanism by which bacteria sense and respond to their environment. YvrGHb is a novel two-component system in *B. subtilis*. The yvrG and yvrHb genes are encoding sensor kinase and response regulator respectively. Serizawa et al. (2005) showed that the YvrGHb system positively regulates the 7 transcriptional units (*wprA, wapA-yyxG, dltABCDE, sunA, sunT-bdbA-yolJ-bdbB, yvrl-yvrHa, and sigX-rsiX*), and
negatively regulates the lytABC operon. wprA, wapA, lytB, and lytC encode the main cell surface proteins of *B. subtilis*. YvrGHb system controls the transcription of the lytC, wprA, and wapA genes, thus controlling the autolytic functions of *B. subtilis*.

Serizawa and Sekiguchi (2005) reported that the YdfHI two-component system regulates ydfJ transcription. The PhoPR two-component system participates in the cellular response to phosphate limitation, and the essential YycFG two-component system in *B. subtilis*. PhoR sensor kinase can activate the YycF response regulator during a phosphate limitation-induced stationary phase, and that this reaction occurs in the presence of the cognate YycG sensor kinase. Phosphorylation of YycF by PhoR also occur *in-vitro*, albeit at a reduced level. A second level of interaction between PhoPR and YycFG is indicated by the fact that cells depleted for YycFG have a severely deficient PhoPR-dependent phosphate limitation response and that YycF can bind directly to the promoter of the phoPR operon. YycFG-depleted cells neither activate expression of phoA and phoPR nor repress expression of the essential tagAB and tagDEF operons upon phosphate limitation. This effect is specific to the PhoPR-dependent phosphate limitation response because PhoPR-independent phosphate limitation responses can be initiated in YycFG-depleted cells (Howell *et. al.*, 2006).

1.3.2 Regulation

Meng and Switzer (2002) addressed transcription attenuation in the pyrimidine synthesis pathway by showing a novel molecular switch that responds
to CTP concentration and determines termination (high CTP) or antitermination (low CTP) of the pyrG gene, encoding CTP synthetase. Putzer et al. (2002) described transcription antitermination by in vitro reconstitution of thrS gene and tRNA<sup>Thr</sup>. Pellegrini et al. (2003) reported the functional characterization of the yajK gene, which encodes the homologue of the RNase Z enzyme, whose function is to endonucleolytically process tRNAs lacking the CCA motif. This was the first demonstration of endonucleolytic maturation of the 3 end of tRNAs in bacteria, a process generally assumed to be exonucleolytic from studies with E. coli, thus establishing a new bacterial paradigm for tRNA maturation. Schujman et al. (2003) discovered fapR, a transcription factor common to many gram-positive organisms, involved in the global regulation of fatty acids and phospholipid metabolism in response to the cellular pool of malonyl coenzyme A.

Spx is a global transcriptional regulator of the oxidative stress response in B. subtilis. Its target is RNA polymerase, where it contacts the α subunit C-terminal domain. Recent evidence indicates that Spx participates in sulfate-dependent control of organo-sulfur utilization operons, including the yml, yxel, ssu, and yrrT operons. The yrrT operon includes the genes that function in cysteine synthesis from S-adenosylmethionine through intermediates S-adenosylhomocysteine, ribosylhomocysteine, homocysteine, and cystathionine. These operons are also negatively controlled by CymR, the repressor of cysteine biosynthesis operons. All of the operons are repressed in media containing cysteine or sulfate but are derepressed in medium containing the alternative sulfur source, methionine. Spx was found to negatively control the expression of these operons in sulfate medium,
in part, by stimulating the expression of the \textit{cymR} gene. Spx directly activates \textit{yrrT} operon expression during growth in medium containing methionine as sole sulfur source (Choi et al., 2006)

1.3.3 Sporulation and Cell Division

The transcription factor Spo0A is a master regulator for entry into sporulation in \textit{B. subtilis}, but it has been uncertain whether activation of Spo0A is sufficient to trigger development. Spo0A, a member of the response regulator family of gene-control proteins, is activated by phosphorylation via a multicomponent phosphorelay in response to conditions of nutrient limitation. Fujita and Losick (2005) reported that sporulation can be triggered with high efficiency in cells in the exponential phase of growth in rich medium by artificial induction of the synthesis of any one of three histidine kinases that feed phosphoryl groups into the relay. They showed that the levels of Spo0A protein and activity increase gradually over the first 2 h of sporulation both under conditions of nutrient limitation and in response to induction of kinase synthesis. Spo0A regulates expression of the virulent \textit{B. subtilis} phage phi29. Spo0A binds to the origins of replication and prevents the initiation step of DNA replication of either genome (Castilla-Llorente et al., 2006).

\textit{Bacillus subtilis} aconitase, encoded by the \textit{citB} gene functions both an enzyme and an RNA binding protein. Mutated C-terminal region of \textit{citB} gene product had high catalytic activity but was defective in sporulation. The defect was at a late stage of sporulation, specifically affecting expression of $\sigma^E$-dependent
genes, which are important for spore coat assembly and require transcriptional activation by GerE. Aconitase binds to and stabilizes gerE mRNA in order to allow efficient GerE synthesis and proper timing of spore coat assembly (Serio et al., 2006).

*B. subtilis*, sfkABCDEFGH operon (ybcOPST and ybdABDE) is involved in the production and export of a sporulation killing factor while the sdpABC operon (yvaWXY) encodes an extracellular signaling protein. The concerted action of the killing factor and the signaling protein produced by cell that have entered the sporulation pathway result in inhibition of sporulation in the cells still in vegetative phase. These cells become more sensitive to the killing factor; thus, they lyse and provide nutrients for the sporulating cells to feed on and continue growing (Gonzalez-Pastor et al., 2003).

Smits et al., (2005) characterized eight *B. subtilis* genes encoding putative cytoplasmic thioredoxin-like proteins. Only the trxA mutant showed a phenotype, as this strain was deficient in competence development and sporulation. TrxA was shown to be involved in disulfide bond formation in the ComS peptide regulating ComK activity and to be involved in improving the secretion of SOS bond-containing proteins in *B. subtilis*.

Yamaguchi et al., (2004) reported that *B. subtilis* yvcE gene involved in peptidoglycan hydrolysis and belong to DL-endopeptidase family II. A new peptidoglycan hydrolase, CwlS, hydrolyzes the linkage of d-gamma-glutamyl-meso-diaminopimelic acid in the cell wall of *B. subtilis*, suggesting that CwlS is a
d, l-endopeptidase. The localization sites are similar to those of LytF and LytE, indicating that CwIS is involved in cell separation with LytF and LytE (Fukushima et al., 2006).

1.4 In-silico Analysis of yjbI

In-silico analysis of the B. subtilis genome reveals that >37% of genes have no known function. We have looked for hypothetical open reading frames, which have similar functions to known genes from other microorganisms. Using various template sequences of hemoglobins from bacterial species, BLAST was performed at the NCBI server. We found a putative truncated globin gene (yjbI) with a significant homology to many globin and globin-like proteins from different microbial genomes. The presence of globin fold in the sequence of yjbI indicates that it is a member of the protozoan/cyanobacterial globin family [www.sbg.bio.ic.ac.uk/~3dpsm/]. The yjbI gene length is 399 base pair and protein length is 132 amino acids. Interestingly, yjbI is truncated at the N-terminus, and its predicted molecular mass is 15.1 kDa. This hypothetical protein yjbI is named as trBHb (Bacillus hemoglobin). Alignment at the level of structure based amino acid was performed at biologic work bench [http://seagtool.sdsc.edu]. A sequence comparison of yjbI with known protein profiles in Pfam and InterPro databases revealed one Bac_globin domain.
1.5 Scope of the Present Study

The progress in genome sequencing has led to a rapid accumulation in GenBank submissions of uncharacterized ‘hypothetical’ genes (>30-40%). These genes, which have not been experimentally characterized and whose functions cannot be deduced from simple sequence comparisons alone, now comprise a significant fraction of the public databases. Most genes found in databases have only been predicted by in-silico methods but experimental validation of such genes is necessary and it poses a great challenge to scientists working in the area of functional genomics.

Truncated hemoglobins are found widely in micro-organisms and have role in oxygen transport. Many of the truncated globins have not been characterised functionally in many bacteria including pathogens like Bacillus anthracis, S. aureus. Since the ORF “yjbl” corresponding to truncated globin of B. subtilis is highly similar to truncated globins of deadly pathogens M. lepre (41% identities) and M. tuberculosis glbO (40% identities), this ORF has been studied by cloning, expressing in E. coli vector, and hypothetical protein characterized in order to understand the function of this protein in B. subtilis.
1.6 Aims and Objectives of the Present Study

The present study was initiated with the following objectives.

- Identification of hypothetical ORF yjbl as truncated hemoglobin using bioinformatics tools.
- PCR amplification, cloning and expression of corresponding gene in *E. coli* expression vector.
- Purification of recombinant protein
- Characterization of protein