



Jaypee University of Information Technology
Solan (H.P.)
LEARNING RESOURCE CENTER

Acc. Num. **SP03070** Call Num:

General Guidelines:

- ◆ Library books should be used with great care.
- ◆ Tearing, folding, cutting of library books or making any marks on them is not permitted and shall lead to disciplinary action.
- ◆ Any defect noticed at the time of borrowing books must be brought to the library staff immediately. Otherwise the borrower may be required to replace the book by a new copy.
- ◆ The loss of LRC book(s) must be immediately brought to the notice of the Librarian in writing.

Learning Resource Centre-JUIT



SP03070

Comparative Genomics of *Bacillus* species

By

PARUL AGARWAL (031531)

NEHA KESARWANI (031537)



MAY-2007

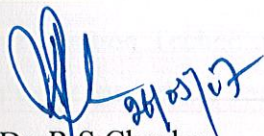
Submitted in partial fulfillment of the Degree of
Bachelor of Technology

DEPARTMENT OF BIOTECHNOLOGY AND
BIOINFORMATICS

JAYPEE UNIVERSITY OF INFORMATION
TECHNOLOGY-WAKNAGHAT

CERTIFICATE

This is to certify that the work entitled, “Comparative genomics of *Bacillus species*” submitted by Parul agarwal and Neha kesarawani in partial fulfillment for the award of degree of Bachelor of Technology in Bioinformatics of Jaypee University of Information Technology has been carried out under my supervision. This work has not been submitted partially or wholly to any other University or Institute for the award of this or any other degree or diploma.



Dr. R.S. Chauhan

H.O.D

Department of Biotechnology and Bioinformatics

Jaypee University of Information Technology

Waknaghat, solan-H.P

CANDIDATE'S DECLARATION

We hereby certify that the work, which is being presented in the thesis, entitled, "Comparative genomics of *Bacillus* species", is partial fulfillment of the requirements for the award of degree of Bachelor of Technology in Bioinformatics and submitted in bioinformatics and biotechnology Department of Jaypee University of Information Technology, Waknaghat (Distt. Solan), is an authentic record of our own work carried out under the supervision of Dr. R.S.Chauhan.

The matter presented in this thesis has not been submitted by us for the award of degree of any other degree of this or any other University.



(PARUL AGARWAL)



(NEHA KESARWANI)

This is to certify that the above statement made by the candidates is correct and true to the best of my knowledge.

Supervisor:



Dr. R.S.Chauhan

H.O.D

Department of Biotechnology and Bioinformatics

Jaypee University of Information Technology

Waknaghat, solan-H.P

ACKNOWLEDGEMENT


We wish to express our sincere gratitude to **Dr. R. S. Chauhan**, for providing us invaluable guidance and suggestions. He was always there to listen and to give advices.

He taught us how to ask questions and express our ideas. He showed us different ways to approach a research problem and the need to be persistent to accomplish any goal.

We would also like to thank all the staff members of Bioinformatics Department of **Jaypee University Of Information Technology**, Solan, for providing us all the facilities required for the completion of this thesis.

We wish to thank all our classmates and friends for their timely suggestions and cooperation during the period of our thesis.

Last, but not least, we thank our families for educating us with aspects from sciences, for unconditional support and encouragement to pursue our interests, even when the interests went beyond boundaries of language, field and geography, for listening to our complaints and frustrations, and for believing in us, for reminding us that our research should always be useful and serve good purposes for all humankind.



Parul Agarwal



Neha Kesarwani

TABLE OF CONTENTS

CERTIFICATE.....	2
CANDIDATE'S DECLARATION.....	3
ACKNOWLEDGEMENTS.....	4
CONTENTS.....	5-6
LIST OF FIGURES.....	7
LIST OF ABBREVIATIONS.....	8
ABSTRACT.....	9
CHAPTER 1	
INTRODUCTION.....	10
1.1 Identification of Drug targets in <i>Bacillus</i> species.....	10-11
1.2 Comparative Genomics.....	11
1.3 Recent use of Comparative Genomics in Microbial Pathogens.....	11-13
1.4 Rationale of the Project.....	13-14
1.5 <i>Bacillus</i> species.....	15
1.5.1 <i>Bacillus anthracis</i>	16-17
1.5.1(a) Pathogenesis.....	17
1.5.2 <i>Bacillus cereus</i>	18-19
1.5.3 <i>Bacillus halodurans</i>	19-20
1.5.3(a) Uses.....	20
1.5.4 <i>Bacillus licheniformis</i>	20
1.5.4(a) Uses.....	20
1.5.5 <i>Bacillus subtilis</i>	21-22
CHAPTER 2	
Materials and Methods.....	23
2.1 Download Of Genome sequence data.....	23
2.2 Identification of Unique regions in pathogenic <i>Bacillus</i> species.....	23

2.2.1 What are essential genes????.....	23-24
2.3 search for Unique regions/Proteins from pathogenic <i>Bacillus</i> species against human species and beneficial <i>Bacillus</i> species.....	25-26
2.3.1 what is S score and What is E value???	27
2.3.1(a)Defining E value.....	27
2.4 Prediction of function of the uniquely identified Proteins in Pathogenic <i>Bacillus</i> species.....	28
2.4(a) working of Profun 2.2 Server.....	29
CHAPTER 3	
RESULTS	
Identification Of Unique Proteins in Pathogenic <i>Bacillus</i> species	
3.1 Unique Proteins identified in <i>Bacillus anthracis</i> Along with their functions and accession ID's.....	32-34
3.2 Unique Proteins identified in <i>Bacillus cereus</i> Along with their functions and accession ID's.....	35
3.3 Identified Proteins common in <i>B.anthraxis</i> and <i>B.cereus</i>	35-36
CHAPTER 4	
Conclusion.....	37

LIST OF FIGURES

Fig1.6.1(a).....	<i>Bacillus anthracis</i>
Fig1.6.1(b).....	<i>Bacillus anthracis</i>
Fig1.6.1(c).....	<i>Bacillus anthracis</i>
Fig1.6.1(d).....	structure of <i>B.anthraxis</i>
Fig1.6.2.....	<i>B.cereus</i> on sheep blood agar plate
Fig1.6.3.....	Image of <i>B.halodurans</i>
Fig1.6.4(a).....	<i>B.cereus</i> image
Fig1.6.4(b).....	<i>B.cereus</i> image
Fig2.1.....	FGENSB tool for gene finding
Fig2.2.....	Output Format in FGENSB
Fig2.3	Flowchart for phase 3
Fig2.4	BLAST tool for comparison of unknown protein with human and bacillus sp
Fig 2.5.....	Protfun 2.2 server for the prediction of the function of unknown protein

LIST OF ABBREVIATIONS

B.....	<i>Bacillus</i>
Protfun.....	Protein Function
BA.....	<i>Bacillus anthracis</i>
BC	<i>Bacillus cereus</i>
BLAST.....	Basic Local Alignment Search Tool

ABSTRACT

Many projects have been completed till today for the sequencing of various bacterial species. The sequenced genomes of pathogenic bacteria provide useful information for understanding host-pathogen interactions. These data prove to be a new weapon in fighting against pathogenic bacteria by providing information about potential drug targets. But the limitation of computational tools for finding potential drug targets has hindered the process and further experimental analysis. There are many *in silico* approaches proposed for finding drug targets but only few have been automated .One such approach finds unique and essential genes in bacterial genomes with no human homologue and predicts these as potential drug target.

We have designed a tool to find genome regions which are unique to pathogenic *Bacillus* species. By using the program based on global alignment we found the unique regions in the various bacterial species predicted the genes in it and found the proteins coded by these genes. After getting the proteins we did the blast (BLASTp) of these proteins with *Bacillus* species and subsequently with humans. And considered only those proteins whose score value was greater than 100 and E-value less than 10^{-10} .To find the function of these proteins we used protfun2.2, an online tool. We got 35 unique proteins in *B.anthraxis* and 6 unique proteins in *B.cereus* along with 12 proteins common in *Bacillus anthracis* and *Bacillus cereus*.

CHAPTER 1

1. Introduction

1.1 Identification of drug targets in *Bacillus* species

Various *in silico* approaches are being used to find details for various diseases such as AIDS, cancer, anthrax etc. this project is also an effort to identify the potential drug targets in microorganisms and validate their function so that it can be used in drug formulation. Its not the case that drugs against these diseases are not available but the drugs are not effective due to many reasons:

- May be exact target to drug is unknown
- May be the bacteria has acquired resistance to the target drug.
- Various side effects occur because of prolonged use of drugs.

Traditionally, genes that are essential for bacterial survival or virulence were identified through random mutagenesis of bacterial genomes [Hood, 1999] and were used as potential drug targets. Genes having sequence similarities with the already identified essential genes can also be considered as essential for the bacterial survival. The effort of compiling currently available essential genes found by various experiments on one platform by Zhang *et al.*, 2004, has provided a method to find essential genes based on sequence similarities. The Database of Essential Genes (DEG) developed by them contains 3007 essential genes in its latest edition i. e. DEG 2.5. The DEG contains essential genes identified in the genomes of *Mycoplasma genitalium*, *Haemophilus influenzae*, *Vibrio cholerae*, *Staphylococcus aureus*, *Escherichia coli*, *Bacillus subtilis*, *Helicobacter pylori*, *Streptococcus pneumoniae* *Rx-1* and *Saccharomyces cerevisiae* All the essential genes found by the above method can be presented as target genes since the bacteria for survival would require all of them. A few among them may have sequence similarities with the host genes and drugs against these genes products may have toxic effects on the host. So, it is necessary to identify such similar genes and exclude them from the list of possible drug targets. The availability of all human proteins at NCBI has provided a platform to identify genes with human homologue.

We used comparative genomics approaches to identify common regions between the pathogenic and beneficial microbes belonging to the same bacterial species and then out from the complete data we extracted uncommon regions and proteins. These are going to directly guide to the key locations where we can target.

1.2 Comparative genomics

Comparative genomics is the analysis and comparison of genomes from different species. The purpose is to gain a better understanding of how species have evolved and to determine the function of genes and noncoding regions of the genome. Researchers have learned a great deal about the function of human genes by examining their counterparts in simpler model organisms such as the mouse. Genome researchers look at many different features when comparing genomes: sequence similarity, gene location, the length and number of coding regions (called exons) within genes, the amount of noncoding DNA in each genome, and highly conserved regions maintained in organisms as simple as bacteria and as complex as humans.

1.3 Recent Use of Comparative Genomics in Microbial Pathogens

Results from the completed prokaryotic genome sequences show that almost half of predicted coding regions identified are of unknown biological function. Comparative genomics exploits both similarities and differences in the proteins, RNA, and regulatory regions of different organisms to infer how selection has acted upon these elements. Those elements that are responsible for similarities between different species should be conserved through time (stabilizing selection), while those elements responsible for differences among species should be divergent (positive selection). Finally, those elements that are unimportant to the evolutionary success of the organism will be unconserved (selection is neutral).

The study of transport proteins in prokaryotic species elucidates the relationship between genome size and biological complexity. The ability to discriminate and transport appropriate compounds is an essential function of cell membranes and their resident proteins. The fidelity of these transport reactions is particularly critical at the

cytoplasmic membrane of prokaryotes since this is the primary barrier that separates the physiologic reactions of the cytosol from the external environment. Many bacterial pathogens face astounding chemical and biological challenges from their host environment (e.g., the extreme acidity of the gastrointestinal tract challenges *H. pylori*). In each host-pathogen relationship, the microbial membrane system contributes to the cell's strategy for energy production and carbon fixation while maintaining ionic homeostasis so that the enzymatic activities of the cytosol can proceed. In addition, all species encode proteins to expel toxic ions (particularly metals) and metabolites.

With complete genome sequences, evaluating the quantity and contribution of solute traffic across the membrane boundaries of pathogenic organisms is now possible. Comparisons between 11 sequenced bacterial pathogens indicate that approximately 6% of each genome encodes proteins (holoenzymes and subunits) involved in solute transport. This percentage is likely an underestimate since many of the gene products annotated as hypothetical proteins have hydropathy profiles reflective of known transporters. Genome size and the number of transport systems are directly related; the greatest number, 53, is annotated in the *M. tuberculosis* genome, and the smallest, 12, is found in the sequence of *M. genitalium*. Bacterial pathogens are heterotrophs; therefore, most of their import systems are used for the uptake of organic compounds (carbohydrates, organic alcohols, acids, amino acids, peptides, and amines). *M. tuberculosis* is the exception; it has 18 annotated transporters for organic substrates and 34 for the movement of ions. In this genome, there are nine copies of a P-type ATPase with a predicted substrate specificity for divalent cations. Whether this reflects a physiologic specialization allowing *M. tuberculosis* to be more resilient in its host environment is unknown.

Comparative genomics involves the use of computer programs that can line up multiple genomes and look for regions of similarity among them. Some of these sequence-similarity tools are accessible to the public over the Internet. One of the most widely used is BLAST, which is available from the National Center for Biotechnology Information. BLAST is a set of programs designed to perform

similarity searches on all available sequence data .Blast results are analysed on the basis of S score and E value.

1.4 Rationale of the Project

Project has been designed and worked out keeping in view that it should be able to give potential targets against the pathogens for which stable cure is not available. It should also keep in check that these proteins when targeted by various drugs should not harm the environment or other microbes. For this the proteins should be uniquely identified in each *Bacillus* species. We just know that antibiotics or various other drugs which we take for cure of the disease but have side effects

Side effects of antibiotics and other drugs on environment and other microbes

Various antibiotics are used against *B.anthraxis* and *Bacillus cereus*. Some of them are **penicillin, tetracycline, erythromycin, chloramphenicol, gentamycin and ciprofloxacin**. All these antibiotics when enter into the environment cause harm to environment as well as to other microbes. Various human and veterinary therapeutics are released to the environment by various routes. Residues released during the manufacturing process may ultimately enter surface waters. After administration, human medicines are absorbed, metabolized and then excreted to the sewer system. They usually go through a treatment works before they find their way into receiving waters or land by the application of sewage sludge. Antibacterial for the treatment of fish or shrimp in aquaculture are directly released to surface waters. Veterinary medicines used to treat pasture animals are excreted to soils or surface waters. In intensive livestock treatments, these medicines are likely to enter the environment indirectly through the application of slurry and manure as fertilizers. Other minor routes of entry include emissions to air and through the disposal of unused medicines and containers. Once released into the environment, pharmaceuticals will be transported and distributed to air, water, soil or sediment. A range of factors, such as the physico-chemical properties of the compound and the characteristics of the receiving environment, will affect their distribution. The degree to which a pharmaceutical is transported between the different environmental media primarily

depends on the sorption behavior of the substance in soils, sediment-water systems and treatment plants, which varies widely across pharmaceuticals. Reported sorption coefficients for several veterinary medicines in soils range from less than 1 litre per kilogram to more than 6,000 litres per kilogram.

Pharmaceutical compounds are designed either to be highly active and interact with receptors in humans and animals or to be toxic for many infectious organisms, including bacteria, fungi and parasites. But this does not mean that they affect only these living forms. Many lower animals have receptor systems similar to humans and animals used in agriculture. Furthermore, many groups of organisms that affect human and animal health, which are targeted by pharmaceuticals, have a crucial role in the functioning of ecosystems. It is therefore possible that pharmaceuticals may cause subtle effects on aquatic and terrestrial organisms that are not detected in standard studies. And as human medicines are almost continuously released to the environment, wildlife organisms are exposed for much longer durations than those used in standard tests. Researchers have therefore begun to look into some of the more subtle effects caused by long-term, low-level exposure to pharmaceuticals. A wide range of subtle impacts has been reported so far, including effects on oocytes and testicular maturation, impacts on insect physiology and behavior, effects on dung decomposition, inhibition or stimulation of growth in aquatic plant and algae species, and the development of antibacterial resistance in soil microbes. Steroids from contraceptives are strongly suspected to affect the fertility and development of fish, reptiles and aquatic invertebrates. Equally, antibiotics from human and veterinary use have an effect on soil microbes and algae. Macro cyclic lactones can affect invertebrate larvae in dung at fairly low concentrations; earthworms appear sensitive to the parasiticides used in veterinary medicine and plants may be sensitive to many antibiotic

Thus if we design such drugs which will only target the harmful species and the disease then the environmental hazards can be escaped.

1.5 *Bacillus* SPECIES

HIERARCHY

Scientific classification	
Kingdom:	Bacteria
Division:	Firmicutes
Class:	Bacilli
Order:	Bacillales
Family:	Bacillaceae
Genus:	<i>Bacillus</i>

Bacillus is a genus of rod-shaped, Gram-positive bacteria and a member of the division Firmicutes. *Bacillus* species are either obligate or facultative aerobes, and test positive for the enzyme catalase. Ubiquitous in nature, *Bacillus* includes both free-living and pathogenic species. Under stressful environmental conditions, the cells produce oval endospores that can stay dormant for extended periods. These characteristics originally defined the genus, but not all such species are closely related, and many have been moved to other genera.

Two *Bacillus* species are considered medically significant: *B. anthracis*, which causes anthrax, and *B. cereus*, which causes a food borne illness similar to that of *Staphylococcus*. A third species, *B. thuringiensis*, is an important insect pathogen, and is sometimes used to control insect pests. The type species is *B. subtilis*, an important model organism. It is also a notable food spoiler, as is *B. coagulans*.

An easy way to isolate *Bacillus* is by placing non-sterile soil in a test tube with water, shaking, placing in melted Mannitol Salt Agar, and incubating at room temperature for at least a day. Colonies are usually large, spreading and irregularly-shaped. Under the microscope, the *Bacillus* appear as rods, and a substantial portion usually contain an oval endospore at one end, making it bulge.

Work has been carried out on four species of bacteria, 2 of which are pathogenic and 2 species are non pathogenic.

Name of the bacteria	Pathogenic/non pathogenic
<i>Bacillus anthracis</i>	Pathogenic
<i>Bacillus cereus</i>	Pathogenic
<i>Bacillus halodurans</i>	Industrially important
<i>Bacillus licheniformis</i>	Industrially important

Brief description of various species of *Bacillus* species on which we have worked:-

1.5.1 *Bacillus anthracis*

Hierarchy

Kingdom: Bacteria

Phylum: Firmicutes

Class: Bacilli

Order: Bacillales

Family: Bacillaceae

Genus: *Bacillus*

Species: *B. anthracis*

Bacillus anthracis is a Gram-positive, facultative anaerobic, rod-shaped bacterium of the genus *Bacillus*. An endospore forming bacterium, *B. anthracis* is a natural soil-dwelling organism, as well as the causative agent of anthrax

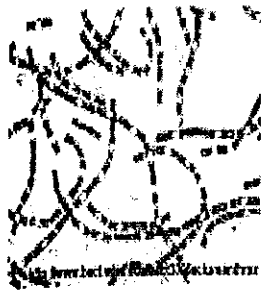


Fig 1.5.1a



Fig 1.5.1b

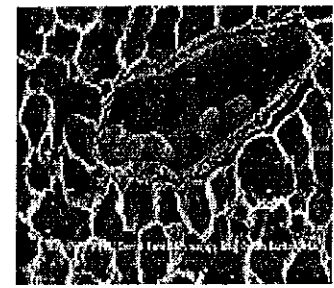


Fig 1.5.1c

B. anthracis was the first bacterium conclusively demonstrated to cause disease, by Robert Koch in 1877.

1.5.1(a) Pathogenesis

Under conditions of environmental stress, *B. anthracis* bacteria naturally produce endospores which rest in the soil and can survive for decades in this state. When ingested by a cattle, sheep, or other herbivores, the bacteria begin to reproduce inside the animal and eventually kill it, then continue to reproduce in its carcass. Once the nutrients are exhausted, new endospores are produced and the cycle repeats

The form associated with the 2001 anthrax attacks produced both toxin (consisting of three proteins: the protective antigen, the edema factor and the lethal factor) and a capsule (consisting of a polymer of glutamic acid). Infection with anthrax requires the presence of all three of these exotoxins

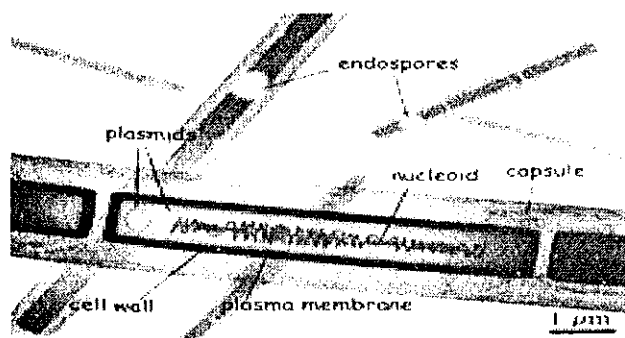


Fig 1.5.1(d) Structure of bacillus anthracis

1.5.2 *Bacillus cereus*

Hierarchy

Kingdom: Bacteria
Phylum: Firmicutes
Class: Bacilli
Order: Bacillales
Family: Bacillaceae
Genus: *Bacillus*
Species: *cereus*

Bacillus cereus is an endemic, soil-dwelling, Gram-positive, rod shaped, beta hemolytic bacteria that causes foodborne illness. *B. cereus* bacteria are facultative aerobes, and like other members of the genus *Bacillus* can produce protective endospores.

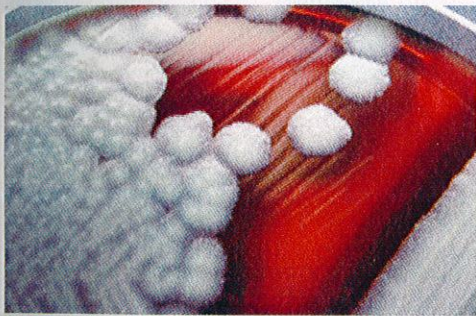


Fig 1.5.2 *Bacillus cereus* on sheep blood agar plate

Bacillus cereus is a normal inhabitant of the soil, but it can be regularly isolated from foods such as grains and spices. *B. cereus* causes **two types of food-borne intoxications** (as opposed to infections). One type is characterized by nausea and vomiting and abdominal cramps and has an incubation period of 1 to 6 hours. It resembles *Staphylococcus aureus* food poisoning in its symptoms and incubation period. This is the "short-incubation" or emetic form of the disease. The second type is manifested primarily by abdominal cramps and diarrhea with an incubation period of 8 to 16 hours. Diarrhea may be a small volume or profuse and watery. This type is referred to as the "long-incubation" or diarrheal form of the disease and it resembles

food poisoning caused by *Clostridium perfringens*. In either type, the illness usually lasts less than 24 hours after onset.

The short-incubation form is caused by a preformed heat-stable enterotoxin of molecular weight less than 5,000 Daltons. The mechanism and site of action of this toxin are unknown. The long-incubation form of illness is mediated by a heat-labile enterotoxin (molecular weight of approximately 50,000 Daltons) which activates intestinal adenylate cyclase and causes intestinal fluid secretion.

1.5.3 *Bacillus halodurans*

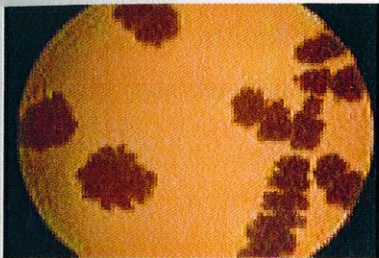


Fig 1.5.3

Bacillus halodurans is one of a group of rod-shaped, Gram-positive, aerobic or (under some conditions) anaerobic bacteria widely found in soil and water.

An alkaliphilic bacterium, strain C-125 (JCM9153), isolated in 1975, was identified as a member of the genus *Bacillus* and reported as a **b-galactosidase** and **xylanase** producer. It is the most thoroughly characterized strain, physiologically, biochemically, and genetically, among those in the collection of alkaliphilic *Bacillus* isolates. Recently, this strain was reidentified as *Bacillus halodurans* based on phylogenetic analysis using 16S rDNA sequence and DNA-DNA hybridisation analysis.

The *B. halodurans* genome contains 112 transposase genes, indicating that transposases have played an important evolutionary role in horizontal gene transfer and also in internal genetic rearrangement in the genome.

Out of 11 factors which belong to the extracytoplasmic function family, 10 are unique to *B.halodurans* , suggesting that they may have a role in the special mechanism for adaptation to an alkaline environment. The genome of *B.halodurans* is a single circular chromosome.

1.5.3(a) Uses

Bacillus halodurans produces many industrially useful alkaliphilic enzymes such as, **protease** (protein degrading enzyme), **cellulase** (cellulose degrading enzyme) and **amylase** (starch degrading enzyme). These enzymes are widely used as additives to laundry detergents. *Bacillus halodurans* also produces **keratin decomposing enzyme** which devolves keratinous proteins such as hair, nail and cock feathers which cause difficulty for their disposal. *Bacillus halodurans* also produces **xylanase** that bleaches pulp in the process of paper-making.

1.5.4 *Bacillus licheniformis*

Bacillus licheniformis is a bacterium commonly found in the soil. Recently, studies have also shown that it is found on bird feathers, especially chest and back plumage, and most often in ground dwelling birds (like sparrows) and aquatic species (like ducks). *B. licheniformis* has also been associated with septicaemia , peritonitis , ophthalmitis , and food poisoning in humans, as well as with bovine toxemia and abortions. *B. licheniformis* is a common contaminant of dairy products

1.5.4(a) Uses

Bacillus licheniformis is cultured in order to obtain **protease for use in biological washing powder**. The bacteria is well adapted to grow in alkaline conditions, and as such, the protease that it produces can withstand high pH levels, making it ideal for this use. The protease has a pH optimum of between 9 and 10 and is added to laundry detergents, in order to digest , and hence remove dirt comprised of proteins on

garments. This allows for much lower temperatures to be used, resulting in lower energy use and a lesser risk of shrinkage of garments or loss of colored dyes.

1.5.5 *Bacillus subtilis*

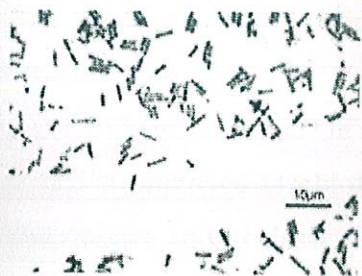


Fig 1.5.5 a



Fig 1.5.5 b

Hierarchy

Kingdom: Bacteria
Phylum: Firmicutes
Class: Bacilli
Order: Bacillales
Family: Bacillaceae
Genus: *Bacillus*
Species: *subtilis*



Bacillus subtilis is a Gram-positive, catalase-positive bacterium commonly found in soil. A member of the genus *Bacillus*, *B. subtilis* has the ability to form a tough, protective endospore, allowing the organism to tolerate extreme environmental conditions. Unlike several other well-known species, *B. subtilis* has historically been classified as an obligate aerobe, though recent research has demonstrated that this is not strictly correct.

B. subtilis is used as a soil inoculant in horticulture and agriculture. *B. subtilis* has been used for a biowarfare stimulant during Project SHAD. *B. subtilis* hazard status is under dispute.

Enzymes produced by *B. subtilis* and *B. licheniformis* are widely used as additives in laundry detergents.

Its other uses include the following:

- a model organism for laboratory studies

- a strain of *B. subtilis* formerly known as **Bacillus natto** is used in the commercial production of the Japanese delicacy **natto** as well as the similar Korean food **cheonggukjang**

- *B. subtilis* strain QST 713 (marketed as QST 713 or Serenade™) has a natural fungicidal activity, and is employed as a biological control agent

- **can convert nuclear waste and explosives into harmless compounds of nitrogen, carbon dioxide, and water**

- plays a role in safe radionuclide waste [e.g. Thorium (IV) and Plutonium (IV)] disposal with the proton binding properties of its surfaces

- recombinants *Bacillus subtilis* str. pBE2C1 and *Bacillus subtilis* str. pBE2C1AB were used in **production of polyhydroxyalkanoates (PHA)** and that they could use malt waste as carbon source for lower cost of PHA production.

CHAPTER 2

MATERIALS AND METHODS

Work on the project has been done in various parts or phases:

2.1 Download of Genome Sequence Data

Downloaded the raw genome sequence of *Bacillus anthracis*, *Bacillus cereus*, *Bacillus halodurans*, *Bacillus subtilis* and *Bacillus licheniformis* from National Centre of Biological Information (NCBI).

Made a program on global alignment to find the unique and common regions between the microorganisms with a window size of 300 and stringency value of 230. with the help of this program we did pair wise alignment of

Bacillus anthracis with *Bacillus halodurans*

Bacillus cereus with *Bacillus licheniformis*

2.2 Identification of Unique regions in Pathogenic *Bacillus* species

After getting the unique regions in each pathogenic *Bacillus* species we converted them in genes by an online tool FGENSEB. we pasted the unique fragments of genome of *B. anthracis*, obtained from pairwise alignment of *B. anthracis* with *B. halodurans*, one by one and converted them in genes. It gave the output telling us whether it was a positive strand or negative strand and also it depicted the proteins coded by those genes. Repeated the same process with *B. cereus* also.

We also checked whether the proteins coded by these genes are essential or not by doing BLASTp in DEG (database of essential genes).

2.2.1 What are essential genes????

Essential Genes are the genes that are indispensable to sustain cellular life. The functions encoded by essential genes are considered as a foundation of life and therefore are likely to be common for all cells. Users can Blast the query sequences against DEG. If the homologous genes are found, it is possible that the queried genes are also essential. Essential gene products comprise excellent targets for antibacterial drugs. The analysis of essential genes could help to answer the question what are the basic functions necessary to support cellular life

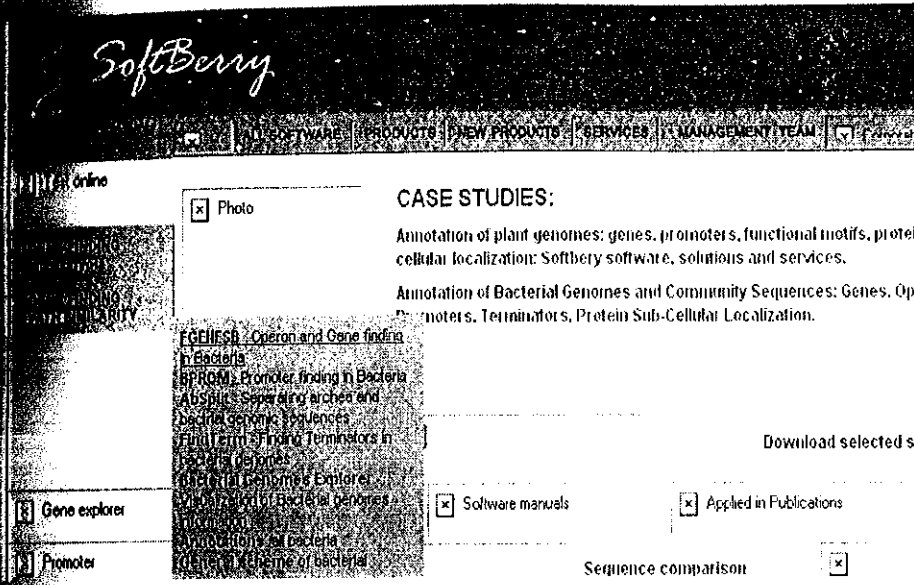


Fig 2.1 Homepage of softberry

Fig 2.2 Output format in FGENSEB

Prediction of potential genes in microbial genomes
 Tue Jan 1 00:00:00 2005
 Test sequence: test sequence
 Length of sequence - 86100 bp
 Number of predicted genes - 29
 Number of transcription units - 1, operons - 1

N	Tu/Op	Conserved S	Start	End	Score
1	1 Op 1	+	3	3014	2146
2	1 Op 2	+	3001	6045	2132
3	1 Op 3	+	5999	9025	1804
4	1 Op 4	+	9099	12005	1598
5	1 Op 5	+	12079	15036	1598
6	1 Op 6	+	15059	18016	1507
7	1 Op 7	+	17997	21011	1839
8	1 Op 8	+	20998	24006	1335
9	1 Op 9	+	23978	27034	1405
10	1 Op 10	+	27021	30014	2138
11	1 Op 11	+	30001	33045	2132
12	1 Op 12	+	32999	36025	1804
13	1 Op 13	+	36099	39005	1598
14	1 Op 14	+	39079	42036	1598
15	1 Op 15	+	42059	45016	1507
16	1 Op 16	+	44997	48011	1839
17	1 Op 17	+	47998	51006	1335
18	1 Op 18	+	50978	54034	1405
19	1 Op 19	+	54021	57014	2138
20	1 Op 20	+	57001	60045	2132
21	1 Op 21	+	59999	63025	1804
22	1 Op 22	+	63099	66005	1598
23	1 Op 23	+	66079	69036	1598

2.3 Search for Unique genome regions/proteins from pathogenic Bacillus species against human genome and beneficial *Bacillus* species.

We did the BLAST of the proteins coded by various genes in Bacillus anthracis and Bacillus cereus one by one with Bacillus database followed by human protein

database.

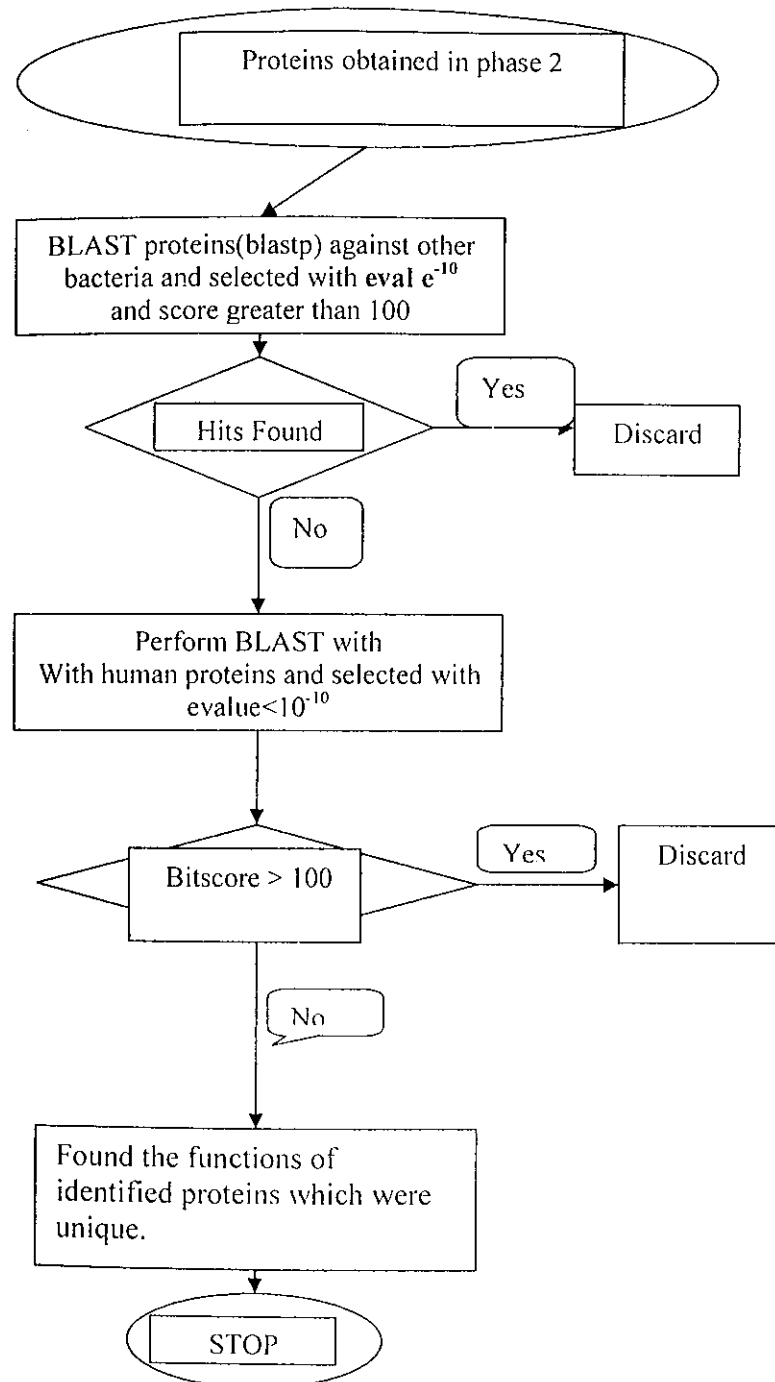


Fig 2.3 Flow chart of searching the unique genome regions/proteins from pathogenic sp of Bacillus against human genome and Benificial Bacillus species

Snapshot of BLAST tool for comparison of the unknown protein with human and bacillus species fig 2.4

Query-
Length=699

No significant similarity found. For reasons why, [click here](#).

Database: All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF exclud:
environmental samples from WGS projects

Posted date: May 12, 2007 5:53 PM

Number of letters in database: 68,962,236

Number of sequences in database: 192,717

Lambda K H
0.316 0.134 0.370

Gapped
Lambda K H
0.267 0.0410 0.140

Matrix: BLOSUM62

Gap Penalties: Existence: 11, Extension: 1

Number of Sequences: 192717

Number of Hits to DB: 548726

Number of extensions: 25881

Number of successful extensions: 122

Number of sequences better than 10: 0

Number of HSP's better than 10 without gapping: 0

Number of HSP's gapped: 122

BLAST results are analyzed on the basis of S score and E value.

2.3.1 What is S score and what is E value???

- The S score is a measure of the similarity of the query to the sequence shown.
- The E-value is a measure of the reliability of the S score.
- The definition of the E-value is: The probability due to chance, that there is another alignment with a similarity greater than the given S score.

2.3.1(a) Defining E-value

E-value Equation

The actual equation is $E = Kmn(e^{-\lambda S})$

- The parameters K and λ represent natural scales for the search space and the scoring system respectively.
- The rest of the equation represents the size of the query (m), the size of the database (n), and of course the S score.

The Size of the E-value

- The typical threshold for a good E-value from a BLAST search is $e^{-5} = (10^{-5})$ or lower.
- The reason for such low values is that an $E = 0.001$ in a million entry database would still leave 1000 entries due to chance. An $E = e^{-6}$ would only leave one entry due to chance.

Problems with the E-value

1. Tends to be conservative when the query sequence is short (simply cannot achieve high S scores).
2. Statistical theory breaks down with gaps in sequences, so gap scores are used.
3. Some sequences have areas of "low complexity," that will show artificial similarity with other sequences.

BLAST attempts to control for all of these problems, however, they are important to keep in mind.

E-value Summary

- Ideally one wants to run a query on BLAST with a long, unified sequence.
- An $E < e^{-5}$ of an alignment means that that alignment is highly unique, and not due to error.
- An $E \geq e^{-6}$ means that the alignment might be strong, but more research is needed to verify.

2.4 Prediction of function of the uniquely Identified Proteins in Pathogenic species of *Bacillus*

- Obtained the functions of proteins unique to *Bacillus anthracis* using profun 2.2 server.
- Obtained the functions of the proteins unique to *Bacillus cereus* using profun 2.2 server.

- Also got a set of 10 proteins common to *Bacillus anthracis* and *Bacillus cereus* and obtained their functional importance also by using protfun 2.2 server.

2.4(a) Working of Protfun 2.2 server

DESCRIPTION

For each input sequence the server predicts cellular role, enzyme class and Gene Ontology category.

The scores consist of two numbers. The first number is the estimated probability that the entry belongs to the class in question. It is influenced by the prior probability of that class. The second number represents the odds that the sequence belongs to that class/category. It is independent of the prior probability.

For each sequence the scores with the highest information content are marked with arrows (=>).

CELLULAR ROLE

The input protein is classified into 12 different functional categories based on the scheme developed by Monica Riley for *E. coli* in 1993.

- Amino acid biosynthesis
- Biosynthesis of cofactors
- Cell envelope
- Cellular processes
- Central intermediary metabolism
- Energy metabolism
- Fatty acid metabolism
- Purines and pyrimidines
- Regulatory functions
- Replication and transcription

- Translation
- Transport and binding

The categories may overlap, since a protein may belong to several categories.

ENZYME/NONENZYME

The protein is classified as enzyme or non-enzyme

- Enzyme
- Non-enzyme

ENZYME CLASS

The protein is classified into 6 different enzyme classes:

- Oxidoreductase (EC 1.-.-.)
- Transferase (EC 2.-.-.)
- Hydrolase (EC 3.-.-.)
- Isomerase (EC 4.-.-.)
- Ligase (EC 5.-.-.)
- Lyase (EC 6.-.-.)

The prediction of enzyme class is carried out even if the input protein is not predicted to be an enzyme. The purpose is to provide maximal information in borderline cases. However, no arrow marking is done for predicted non-enzymes.

GENE ONTOLOGY CATEGORY

The prediction scores are given for the following Gene Ontology categories:

- Signal transducer
- Receptor
- Hormone
- Structural protein
- Transporter
- Ion channel
- Voltage-gated ion channel

- Cation channel
- Transcription
- Transcription regulation
- Stress response
- Immune response
- Growth factor
- Metal ion transport

The arrow marking is not done if the score with the highest information content has odds lower than 1.

Snapshot of ProtFun server for the prediction of the function of the unknown proteins

CENTERED
 BIOLOGI
 CAL SEQU
 ENCEANA
 LYSIS CBS

ProtFun 2.2 Server - prediction results

Technical University of Denmark

ProtFun 2.2 predictions

>Sequence

# Functional category	Prob	Odds
Amino_acid_biosynthesis	0.012	0.554
Biosynthesis_of_cofactors	0.032	0.439
Cell_envelope	0.036	0.595
Cellular_processes	0.027	0.373
Central_intermediary_metabolism	0.043	0.679
Energy_metabolism	=> 0.215	2.387
Essential_metabolism	0.016	0.215

Fig 2.5

CHAPTER 3

RESULTS

Identification of Unique Proteins in Pathogenic *Bacillus* species

Bacillus anthracis

By doing global alignment of *Bacillus anthracis* and *Bacillus halodurans* with the help of the program for finding unique regions in pathogenic microorganisms with stringency value greater than 2.00 and window size 100 we obtained 129 unique regions in raw genome. These unique regions were converted into genes with the help of online tool FGENSEB.

We did the Blast of these proteins with *Bacillus* species database and humans. We considered only those proteins were bit score was greater than 100 n E-value less than 10^{-10} . After doing it we obtained 35 unique proteins.

Finally we found the function of these proteins with the help of an online tool Protfun2.2 which finds the function of the protein after taking the sequence.

3.1 Unique Proteins identified in *Bacillus anthracis* along with their putative functions and accession ID's.

Protein name	Function	GenBank ID's	Gene ontology category
1. lipoprotein	translation	gi 47525508	Not known
2. DNA binding domain, exonuclease family	translation	gi 30253892	transporter
3 Amino acid permease family protein	Energy metabolism	gi 47778142	transporter
4. Permeases of major facilitator super family)	Transport and binding	gi 65321851	transporter
5. Hypothetical protein BA0789	Energy metabolism	gi 30260931	Growth factor
6. Hypothetical protein BA0881	Energy	gi 30261016	Immune

	metabolism		response
7. Anti sigma b factor antagonist RsbV	Energy metabolism	gi 47526269	Not known
8. Transposase IS605 family	Translation	gi 47530709	Growth factor
9. RNA polymerase sigma C factor	Energy metabolism	gi 47777889	Structural protein
10. Competence transcription factor	Translation	gi 65318516	transcription
11. L lactate permease	Translation	gi 47530778	transporter
12. Hypothetical protein BA1276	Translation	gi 30261371	Structural protein
13. dltb protein	Transport and binding	gi 47526661	transporter
14. Acetolactate synthase(essential)	Translation	gi 30254616	Structural protein
15. Hypothetical protein BAS1360	Translation		Immune response
16. Hypothetical protein BA1523	Transport and binding	gi 30261597	Immune response
17. hypothetical protein BA1571	Energy metabolism	gi 30261644	Not known
18. 2-hydroxychromene-2-carboxylate isomerase family protein	Energy metabolism	gi 47527213	Transcription regulation
19. transcriptional regulator, MarR family	Amino acid biosynthesis	gi 67078221	Transcriptional regulator
20. transport ATP-binding protein CydC(essential)	Cell envelope	gi 47527240	Stress response
21. microcin immunity protein MccF()	Translation	gi 47527244	transcription

22.Hypothetical protein BA1985	Translation	gi 30262014	Growth factor
23.Exonucleas(essential)	Translation	gi 47528879	Growth factor
24.Carboxylesterase	Translation	gi 47530644	Not known
25.hypothetical protein BAS2538	Translation	gi 49179473	Voltage gated ion channel
26.BA3034	Transport and binding	gi 30262988	transporter
27.sugar-binding transcriptional regulator, LacI family(essential)	Amino acid biosynthesis	gi 47778294	Not known
28.serine dehydratase, iron-sulfur-dependent, beta subunit	Translation	gi 47529658	Immune response
29.prophage LambdaBa01, acyltransferase	Purines and pyrimidines	gi 47529089	receptor
30.hydrolase, alpha/beta fold family	Amino acid biosynthesis	gi 47778019	Growth factor
31.Serine protease, subtilase family	Transport and binding	gi 47529180	hormone
32.Dependent dipeptidase, microsomal dipeptidase homolog	Translation		Immune response
33.Renal dipeptidase family protein	Translation	gi 47529202	hormone
34.Prophage LambdaBa02, tape measure protein	Translation	gi 47529376	Not known

Bacillus cereus

By doing global alignment of *Bacillus cereus* and *Bacillus licheniformis* with the help of the program for finding unique regions in pathogenic microorganisms with stringency value greater than 230 and window size 100 we obtained 215 unique

regions in raw genome. These unique regions were converted into genes with the help of online tool FGENSEB.

We did the Blast of these proteins with *Bacillus* species database and humans. We considered only those proteins were bit score was greater than 100 n E-value less than 10^{-10} . After doing it we obtained 6 unique proteins.

Finally we found the function of these proteins with the help of an online tool Protfun 2.2 which finds the function of the protein after taking the sequence.

3.2 Unique Proteins identified in *Bacillus cereus* along with their functions and accession ID's

Protein name	Function	GenBank ID's	Gene ontology category
1.transcription antiterminator, LytR family(essential)	Cell envelope	gi 42784351	Not known
2.Hypothetical protein BCE_5514	Regulatory functions	gi 42784559	Growth factor
3.Cyclic nucleotide-binding domain protein (essential)	Energy metabolism	gi 42784583	Growth factor
4.superoxide dismutase	Amino acid biosynthesis	gi 42780667	transcription
5.Hypothetical protein BCE_5583	Transport and binding	gi 42784628	Not known
6.Transporter, putative ()	Transport and binding	gi 42784626	transporter

3.3 Identified Proteins Common in *B. anthracis* and *B. cereus* along with their accession ID's

Protein name	Function	GenBank ID's	Gene ontology category
1.Acetylornithine deacytlase	Translation	Yes	Growth factor

2.Spore germination protein GERLC	Amno acid biosynthesis	gi 50402123	Structural protein
3.Permease protein	Fatty acid metabolism	gi 30018856	receptor
4.Unknown name	Transport and binding	No	transporter
5.Flageller biosynthesis protein(essential)	Amino acid biosynthesis	gi 65319014	Structural protein
6.Peptidase E	Translation	gi 29895406	Structural protein
7.Unknown name	Central_intermediary_metabolism .		Not known
8.Gluconate permease	Transport and binding	gi 47530396	Immune response
9.Transcriptional regulator, ArsR family	energy metabolism	gi 47528545	growth factor
10.Stage III sporulation protein E(essential)	energy metabolism	gi 47529222	growth factor
11.ATP-dependent protease peptidase subunit	Translation	gi 47529258	Growth factor
12.PhoH family protein	Translation	gi 47529459	Growth factor

CHAPTER 4

Conclusion

The increase in drug resistance among the pathogenic microbes has increased the search for novel drug targets. The available genome data with the *in silico* tools has simplified the search for these targets. The availability of complete sequences of genomes of bacterial pathogens as well as of the human genome is of high importance for finding drug targets *in silico*. But very few of them have been automated and this has been a bottleneck in the *in silico* approach for finding drug targets. The process is further slowed down by the online methods since all methods have their own limitations.

Our project is an effort to provide some proteins against the diseases for which conventional methods are not so useful in producing drugs. Our project has identified 35 unique proteins in *B.anthraxis* out of which 4 proteins are essential, 6 proteins in *B.cereus* among them 2 are essential and 12 common proteins between *Bacillus anthracis* and *Bacillus cereus* among them also we have 12 proteins which are essential. The proteins discovered can be used to target against the diseases in which it is discovered and may be it can lead to some novel drugs against anthrax or food borne illness.

References

- D.A. Rodionov, I. Dubchak, A. Arkin, E. Alm, M.S. Gelfand. **Dissimilatory Metabolism of Nitrogen Oxides in Bacteria: Comparative Reconstruction of Transcriptional Networks.** *PLoS Comput Biol.* 2005 Oct;1(5):e55.
- http://www.entelos.com/pubArchive/2006/Michelson_Chapter_Final.pdf
- Nitesh Kumar Singh, S. Mahalaxmi Selva¹ and Paulsharma Chakravarthy T-**iDT: Tool for identification of drug target in bacteria and validation by Mycobacterium tuberculosis**
- Galperin, M. Y. and Koonin, E. V. (1999). **Searching for drug targets in microbial genomes.** *Curr. Opin. Biotechnol.* 10, 571-578.
- Dutta, A., Singh, S. K., Ghosh, P., Mukherjee, R., Mitter, S. and Bandyopadhyay, D. (2006). **In silico identification of potential therapeutic targets in the human pathogen *Helicobacter pylori*.** *In Silico Biol.* 6, 0005
- Hood, D. W. (1999). **The utility of complete genome sequences in the study of pathogenic bacteria.** *Parasitology* 118, S3-S9.
- Miesel, L., Greene, J. and Black, T. A. (2003). **Genetic strategies for antibacterial drug discovery.** *Nature Rev. Genet.* 4, 442-456.
- Sakharkar, K. R., Sakharkar, M. K. and Chow, V. T. K., (2004). **A novel genomics approach for the identification of drug targets in pathogens, with special reference to *Pseudomonas aeruginosa*.** *In Silico Biol.* 4, 0028.
- Zhang, R., Ou, H. Y. and Zhang, C. T (2004). **DEG: A database of essential genes.** *Nucleic Acids Res.* 32, D271-D272.
- Database of Essential Genes
- Reference paper of comparative genomics of *Bacillus cereus* with *Bacillus subtilis*
- www.ncbi.nlm.nih.gov/BLAST
- www.softberry.com/FGENSB
- www.2can Support Portal Genomes - All Genomes.htm
- www.wikipedia.org
- protfun 2.2 server