INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

The University of Oklahoma

Graduate College

The Working Draft Sequence of the Gram-Positive Bacteria *Bacillus stearothermophilus* Strain 10 Genome, its Analysis, and Annotation

A Dissertation

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

degree of

Doctor of Philosophy

By

Sharon A. Lewis
Norman, Oklahoma
2002

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

The Working Draft Sequence of the Gram-Positive Bacteria *Bacillus stearothermophilus* Strain 10 Genome, its Analysis, and Annotation

A Dissertation APPROVED FOR THE
DEPARTMENT OF CHEMISTRY AND BIOCHEMISTRY

BY

Dr Bruce A Roe

Dr Paul Cook

Dr John Downard

Dr George Richter-Addo

Dr Helen Zgurskaya

Acknowledgments

The author would like to thank her parents, Mr. & Mrs. Thomas A. Scott, for their support and never-ending talks of encouragement while pursuing her educational career. The author thanks her husband, Judge David B. Lewis for his daily example of the Christain walk, unconditional love, and prayers. The author thanks her children, David, Jr., and Danielle, for allowing her to win the many games of Putt-Putt the summer of 2002. The author would like to also thank members of her advisory committee, Drs. Bruce A. Roe, Paul F. Cook, John Downard, George Richter-Addo, and Helen Zgurskaya for their time and support.

The author would like to thank all of her co-workers, present and former, to include Dr. Fares Najar, Dr. Qun Ren, Dr. Doris Kupfer, Beth Perry, Rose Morales-Diaz, Kim Hyatt, Michael Sughrue, Julie Thompson, Michael Schade, Clayton S. Powell, Angie Prescott, Graham Wiley, Christina Barnes, Blake Fitzgerald, Brandon Drabek, Andrea McSpadden, Shaoping Lin, Ziyun Yao, Runying Tian, Steve Shaull, Mounir Elharam, Long Nyugen, Heather Wright, Phoebe Loh, Sulan Qi, Bart Ford, Dr. Stephane Descamps, Anh Phoung Do, Trang Do, Steve Kenton, Jim White, Hongshing Lai, Doug White, Dr. Axin Hua, Ping Hu, Yonas Tesfai, Randy Hine, Tammy Womack, Kaylynn Hale, Dixie Wishnuck, Mary Catherine Williams, Jing Tian, Yonathan Tilahun, George Martin, Wendy Martin, YingPing Wang, Dr. Feng Chen, Dr. Steve Toth, Dr. HuaQin Pan, Dr. Sandy Clifton, Dr. Akbar Khan, Dr. Dennis Burian, Min Zhan, Lisa Lane, Linda Cantu, Dr. Ryan Rahhal, Dr. Murli Rao, and Dr. Rusty Wayt for the team effort, encouragement, mutual respect, and friendly environment.

Financial support for her graduate study from the University of Oklahoma's Department of Chemistry and Biochemistry, the Graduate College, the Sponsor Research Initiative, the National Science Foundation's Experimental Program To Stimulate Competitive Research (EPSCoR), the Department of Energy GAANN Fellowship, and grants from the National Institute of Health, Human Genome Project, was truly appreciated.

Table of Contents

List of Tables

List of Figures

The Working Draft Sequence of the Gram-Positive Bacteria *Bacillus stearothermophilus*
Strain 10 Genome, its Analysis, and Annotation

Abstract

Bacillus stearothermophilus is a Gram-positive, obligate, thermophilic bacterium with an optimal growth temperature of 65 to 70 degrees Celsius. It is a non-photosynthetic, spore forming, rod-shaped bacterium. Strain 10 was originally isolated by Dr. Jochanan Stenesh from hot springs of Yellowstone National Park (1), although it initially was discovered in a spoiled can of corn.

A whole-genome random shotgun (WGS) sequencing method was used to obtain the working draft sequence. To date, over 99 percent of this approximately 3.4 Mbp *Bacillus stearothermophilus* genome has been sequenced from small-insert pUC and large insert P1 clones with average insert size between 1.5 and 6.0 kbp. A total of approximately 78,000 shotgun sequence reads with lengths averaging 505 bp were obtained from double stranded DNA sequencing templates. The sequence reads were assembled using the phred/phrap program after sequencing these subclones with various different chemistries, that included the ABI Big Dyes, Big Dye dGTP or dRhodamine mixes with nested fragment set resolution on the ABI 3700. Gaps in the sequences were closed using primer walking, uniplex and multiplex PCR (2).

The major goal of this research was to determine the working draft sequence of the *Bacillus stearothermophilus* genome, identify its approximately 2,731 open reading frames and describe the metabolic pathways while noting the differences. The lysine pathway in *Bacillus stearothermophilus* follows the succinylated pathway while *Bacillus subtilis* follows the acetylated pathway. In addition, *Bacillus stearothermophilus* encodes seventeen genes involved in the aerobic synthesis of cobalamin, which is a vitamin not synthesized by *Bacillus subtilis* strain 168.

Analysis of the genomic sequence of *Bacillus stearothermophilus* supports the hypothesis that several features in thermophilic organisms contribute to the ability of these organisms to thrive at elevated temperatures. In *Bacillus stearothermophilus*, these features include G-C and A-T clusters in its genome, the overabundance of saturated fatty acids in its lipids, and the overall amino acid composition of its proteins, which have elevated levels of arginine residues in comparison to mesophilic *Bacillus* species.

Chapter I: Introduction

A. Introduction to *Bacillus stearothermophilus*

1. General Information

Bacillus stearothermophilus is a Gram-positive, obligate, thermophilic bacterium with an optimal growth temperature of 65 to 70 degrees Celsius. It is a non-photosynthetic, spore forming, rod-shaped bacterium. *Bacillus* species are saprophytic producing no toxins and are not pathogenic to humans. The *Bacillus stearothermophilus* strain 10, originally isolated by Dr. Jochanan Stenesh from hot springs of Yellowstone National Park, Wyoming, was sequenced to working draft during the course of this dissertation research. Several members in the genus *Bacillus*, in addition to being able to sporulate, also are capable of complementation.

In 1917, P. J. Donk, who examined the spoiled remains of a can of "Standard Maine Style" corn, made the initial discovery of a thermophilic organism that appeared to be an unknown species and proposed the name *Bacillus stearothermophilus* (3), which is derived from *Bacillus*, which means rod shaped, *stearo*, which means fat, *thermo*, which means heat, and *philus*, which means loving.

Bacillus stearothermophilus has been found in all climatic zones, in soil, in deep cores of ocean sediments estimated to have an age of thousands of years, in compost heaps and in spoiled food products. *Bacillus stearothermophilus* strain 22 had less than twenty percent DNA homology with *Bacillus subtilis* and other mesophiles of the genus *Bacilli*. Researchers have determined that thermophilic *Bacilli* also are phenotypically and phylogenetically similar to microorganisms discovered in high temperature oilfields where the organic matter is liquid hydrocarbons. Thus based on fatty acid analysis, DNA-DNA hybridization studies, and the sequence analysis of the 16S rRNA in 2001, *Bacillus stearothermophilus* strain 22 was included in the new taxonomy of *Geobacillus*, with the new taxonomic name of *Geobacillus stearothermophilus* although it will be referred to as *Bacillus stearothermophilus* throughout this dissertation study (4). *Geobacillus* means

earth or soil small rod. During the rest of this dissertation, all reference to *Bacillus stearothermophilus* refers to strain 10 unless otherwise specified.

These vegetative cells occur singly or in short chains. If they are motile, it is with the use of peritrichous flagella. Vegetative cells produce one endospore per cell, which is ellipsoidal or cylindrical and located terminally or subterminally in swollen or non-swollen sporangia. *Bacillus stearothermophilus* strain 22 has growth occuring in a pH range of 6.5-7.5 (4).

Bacillus stearothermophilus is Gram-positive because it contains a multilayered (20-100 layers thick) peptidoglycan containing cell wall that is differentially stained with crystal violet (an aniline dye) primary stain in iodine and safranin (a red counter stain) giving a purple-black color (5)

Bacteria often are classified based on their cell growth conditions. Mesophiles, such as *Bacillus subtilis*, are bacteria that reproduce optimally at temperatures between 30 and 45 degree Celsius while, psychrophiles or psychrotrophes are bacteria that grow between 0 and 30 degree Celsius. There are four main groups of thermophiles. Extreme thermophiles, such as *Aquifex aeolicus* and *Thermotoga maritime*, are bacteria that can grow above 70 degree Celsius, but no *Bacilli* extreme thermophiles are listed in Bergey's Manual (6). Obligate or strict thermophilic bacteria, such as *Bacillus stearothermophilus*, have an optimal growth temperature of 65 and 70 degree Celsius, but do not grow below 40 degree Celsius. Facultative thermophiles have a maximal growth temperature between 50 and 65 degree Celsius but also are capable of growth at room temperature. Thermotolerant bacteria, such as *Bacillus pumilus*, have growth maxima between 45 and 50 degree Celsius and also grow at room temperature (7).

The kinetic theory, the lipid theory, and the macromolecular theory have been developed to explain themophily. In the kinetic theory, thermophilic bacteria are postulated to have a special metabolic state characterized by high rates of synthesis and degradation (turnover) of proteins and nucleic acids. In the lipid theory, it is proposed

that there are high melting point lipids covalently attached to macromolecules and subcellular structures of thermophilic bacteria. In contrast, the macromolecular theory proposes that physical-chemical structure and function of proteins, enzymes, nucleic acids and nucleoproteins for thermophiles differ when compared to mesophiles (8). Finally, in a combined hypothesis, Hendrix and Welker proposed that thermophily in most proteins may not be due to the individual structural properties of these proteins but rather due to the cumulative effect of many amino acid substitutions throughout the protein without a striking change in conformation (9).

In addition to *Bacillus stearothermophilus*, there are at least fifty-three other *Bacillus* species. Also, there are many different strains of each *Bacillus* species, for example, there are at least 20,000 strains of *B. thuringiensis* (7, 10, 6). To maintain this large variety of organisms, the *Bacillus* Genetic Stock Center was established in 1978 at Ohio State University as the centralized culture collection repository (11).

2. Six Applications of *Bacillus stearothermophilus*

Bacillus stearothermophilus and several of its enzymes have been used in numerous applications. For example, the large fragment of *Bacillus stearothermophilus* DNA polymerase (BST) has been used in DNA sequencing reactions (12), as it can accurately replicate nanogram amounts of template DNA, as well as, efficiently extend G-C rich templates. BST polymerase is well suited for the analysis of double-stranded templates, as it is highly processive, has a high polymerization rate and does not require an annealing step (13).

The baking industry has used *Bacillus stearothermophilus* enzymes, such as the amylases to prolong the shelf life of bakery products. These amylases act as anti-staling agents by hydrolyzing gelatinizing starch during baking (14). In 1986, The Food and Drug Administration accepted the first Generally Accepted As Safe (GRAS) petition for alpha-amylase from a genetically engineered organism from the Enzyme Bio-systems Company, making it the first such acceptance of any recombinant derived food ingredient

(15).

Bacillus stearothermophilus enzymes also are used in the dairy industry, where a *Bacillus stearothermophilus* Disc Assay (BSDA) is used to test for animal drug residues (beta lactam antibiotics, i.e., penicillin G, ampicillin, amoxicillin, ceftiofur, cephapirin, cloxacillin) in raw and bulk milk (16). Here, Bacto PM Indicator Agar is used in a microbiological test that measures the growth or inhibition of growth of *Bacillus stearothermophilus* to rapidly detect the presence of penicillin in milk. Milk containing as little as 0.005 IU penicillin/ml milk will produce a zone of inhibition, although the official definition of a positive test is a clear, well-defined zone of inhibition equivalent to the zone around a positive control disk containing 0.008 IU penicillin/ml milk (16).

Bacillus stearothermophilus also is an important bacterium for commercial enzyme manufacture (15, 17). For example, the *Bacillus stearothermophilus* arabinfuranosidase, xylanase, and xylosidase enzymes are used to hydrolyze xylan present in wood pulp. In the pulp and paper industry, these enzymes are used for treatment of pulp in order to reduce hemicellulose content for dissolving pulp production and for bleaching of chemical pulps. These hemicellulases and particularly, xylanases also are used as additives in fruit juices to facilitate their clarification (18).

Since *Bacillus stearothermophilus* spores are more resistant to heat than the mesophilic species in the genus, *Bacillus stearothermophilus* spores on paper strips and disks are used to evaluate the effectiveness of the sterilization processes. Here, the strips or disk are autoclaved, grown in media, and incubated. If thermophilic bacterium growth occurs, then the sterilization process was not effective (19, 11).

The hypothesis of this dissertation is that determining the working draft sequence of *Bacillus stearothermophilus* genome will provide direct evidence for the phenotypical properties of this organism, its metabolic pathways, and its thermophilic growth characteristics.

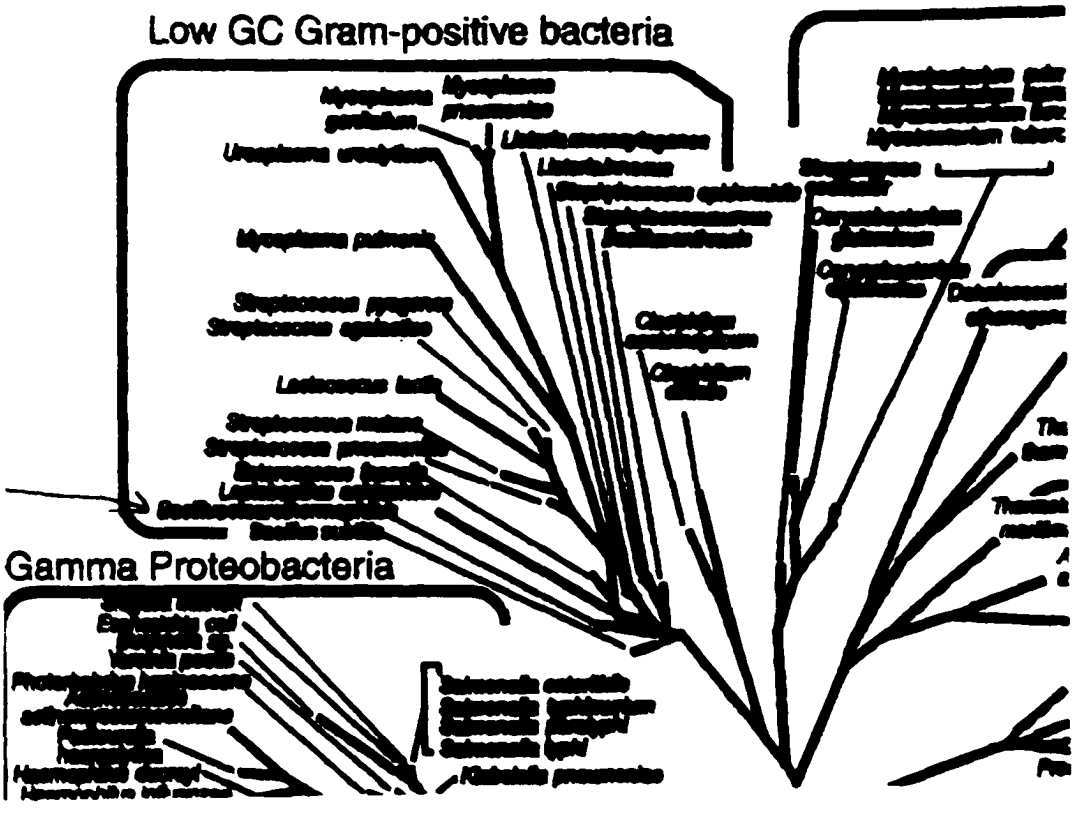
B. Introduction to *Bacillus* Species

1. Phylogenetic Analysis

Several *Bacillus* genomes are either completely sequenced or in the working draft, including the present working draft of *Bacillus stearothermophilus* strain 10, the working draft of *Bacillus anthracis* strain Ames (20), and *Bacillus cereus*, and the completely sequenced genomes of *Bacillus subtilis* strain 168 (21), and *Bacillus halodurans* strain C-125 (22).

Knowledge of the complete nucleotide sequence of the 4,214,810 bp *Bacillus subtilis* genome allowed for the identification of its 4,100 ORFs. *Bacillus subtilis* is a strict aerobic, endospore-forming, mesophilic bacterium found in soil and in association with plants. Phylogenetic analysis places *Bacillus stearothermophilus* just above *Bacillus subtilis* in the lower portion of the Low G-C Gram-positive bacteria group. Five organisms can be grouped as High G-C rich and include *Mycobacterium avium*, which TIGR is sequencing (percent G-C not available yet), *Mycobacterium leprae*, which is 57.79 percent G-C rich, and *Mycobacterium bovis*, which is 65.63 percent (23), *Bacillus flavothermus*, which is 61 percent and *Bacillus schlegli*, which is 64 percent (24). The unrooted phylogenetic tree shown in figure 1 below was based on 16SrRNA sequences for each of the prokaryotic organisms. The major prokaryotic phylogenetic groupings were derived from sequences from the Ribosomal Database Project using Phylip for tree construction where the lengths of the individual lines reflect the amount of sequence change (25).

Figure 1-Phylogenetic Tree of Low GC Gram-positive Bacteria (25)

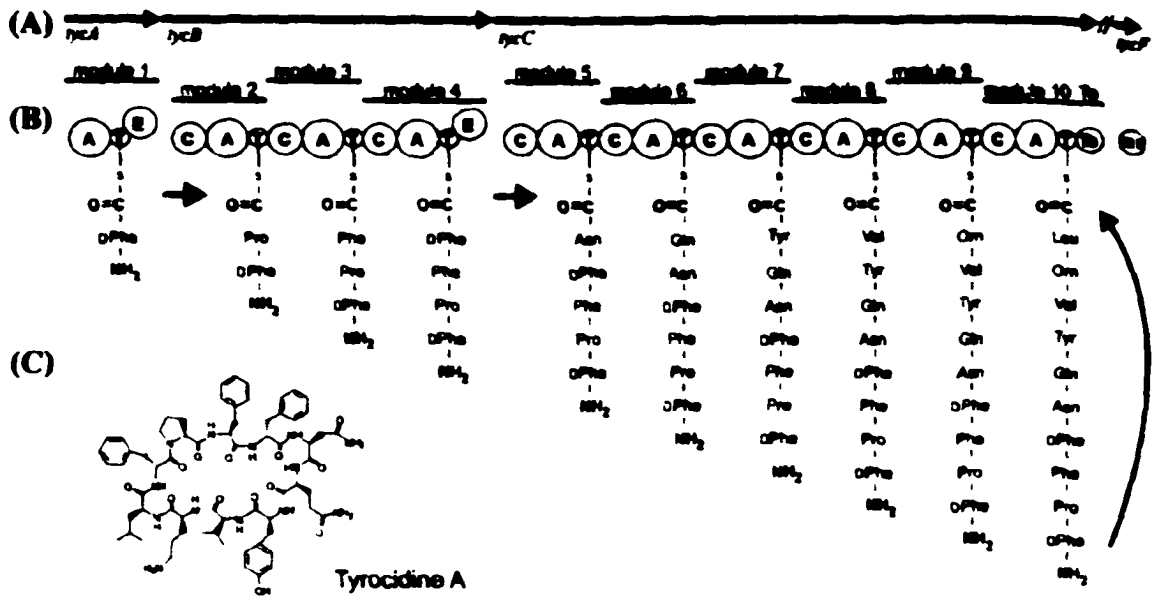


Bacillus species exhibit post-exponential growth responses of adaptation and differentiation that include production of antibiotics and hydrolytic enzymes, chemotaxis and associated motility, its general stress response, sporulation, and more. The first two responses allow them to survive during competition from other soil organisms with the increase in the range of carbon/nitrogen sources that can be utilized. *Bacillus subtilis*, a chemoorganotroph, i.e., bacteria that require an organic source of carbon and metabolic energy, differs from autotrophs, i.e., bacteria capable of using carbon dioxide or carbonates as a sole source of carbon and a simple inorganic nitrogen compound for metabolic synthesis, as it can grow on a salts medium containing glucose or other simple sugars as carbon and energy source and inorganic nitrogen if there is oxygen available (26).

2. Antibiotic Production

Approximately 45 antibiotics can be produced by a variety of *Bacilli*. For example, *Bacillus licheniformis* produces bacitracin, while *Bacillus polymyxa* produces polymyxin, and *Bacillus brevis* produces gramicidins, tyrocidines and edeines (27, 21). *Bacillus subtilis* alone produces ten antibiotics, including Subtilin, a 32 residue antibiotic formed by proteolytic cleavage of modified precursors, and subtilosin A, another 32 peptide produced at the end of vegetative growth just prior to spore formation, which is produced in sporulation medium by the mechanism of usual protein synthesis. Other *Bacillus subtilis* antibiotics such as alboleutin, bacillomycin, bacilysin, bacitracin, fengymycin, iturin, mycobacillin, mycosubtilin, rhizocticins, and surfactin are produced in a more complicated manner that does not involve mRNA and ribosomes (28). Nonribosomal peptide synthetases (NRPSs) are multimodular protein templates that are basically assembly lines of semiautonomous catalytic domains. The nature, number and order of these NRPSs determine the structure and size of many important antibiotics, siderophores, phytotoxins, and immunosuppressive agents. NRPSs are large, multifunctional enzymes with repetitive building blocks called modules, which catalyze one complete cycle of chain elongation and optional modification. NRPSs mode of action is called the multiple carrier thiotemplate mechanism shown in the figure below (24).

Figure 2-The Multiple Carrier Thiotemplate Mechanism (24)



Reprinted with permission from the American Society for Microbiology coypyright 2002 by ASM Press

As shown in part A of the figure 2, three NRPS, encoded by the genes *tycABC*, act together for the stepwise assembly of the cyclic decapeptide tyrocidine A as shown below (29). As shown in part B of figure 2, *tycA* encoding a synthetase is composed of one module, *tycB* is composed of three modules, and *tycC* is composed of six modules The modules are further subdivided into five functional domains. The adenylation (A) domain is where substrates are recognized and adenylated by the corresponding aminoacyl adenylate. The aminoacyl moieties are covalently tethered to the thiol group of cofactor 4' phosphopantetheinyl (Ppant) group. The thiolation (T) domain serves as an anchor for the peptidyl intermediates. The nascent peptidyl chain grows directionally in incremental steps of elongating acyl-A-Ppant intermediates. The condensation (C) domain catalyzes peptide bond formation and chain translocation occurs each time a peptidyl-S-Ppant intermediate is attached by a monomeric aminoacyl-S-Ppant nucleophile. The epimerization (E) domain converts L-Phe moieties into the D-isomer

The terminal (Te) domain is a cyclase and releases the final product. The end product, tyrocidine is in figure 2-C from the mechanism picture (24)

Although antibiotic synthesis has been associated with sporulation, antibiotic-deficient mutants have been isolated which sporulate normally. In addition to the antibiotics produced commercially in *Bacillus*, other products associated with *Bacillus* sporulation include insecticides, and hydrolases, proteolytic enzymes are used in the detergents, dairy, and leather industries, and carbohydrases that are used in the baking, brewing, distilling, starch, and textile industries (27).

3. Chemotaxis

Bacillus subtilis is found many places that include in soil as well as the ocean. This is different from *E. coli*, which exists only in the homeostatic environment of the gut. Chemotaxis is the process of bacteria traveling toward higher concentrations of attractants or lower concentrations of repellants. All twenty amino acids and many sugars are attractants for *Bacillus subtilis*. The repellants include toxins, local anesthetics, inhibitors of electron transport and uncouplers of oxidative phosphorylation. The bacterial sensory system of flagella-chemotaxis and motility is highly evolved. The flagellum, the motor that drives the bacterium, is connected to the bacterial cell surface through a hook and basal body (HBB). The flagellum of *Bacillus subtilis* is a left-handed helix. Chemotaxis is mediated through sets of cellular receptors that bind specific attractants. The binding starts a signal that results in specific methylation-demethylation and phosphorylation-dephosphorylation (24).

4. Proteases

The microbial endoproteases are classified, based on their mechanisms of action, into four categories, which are serine proteases, metalloproteases, thiol (sulfulhydryl) proteases, and acid (aspartic) proteases. To date, no thiol or acid proteases have been reported in *Bacillus*. Seven extracellular proteases are mapped to the *Bacillus subtilis* chromosome and shown to be nonessential for either growth or sporulation. The four

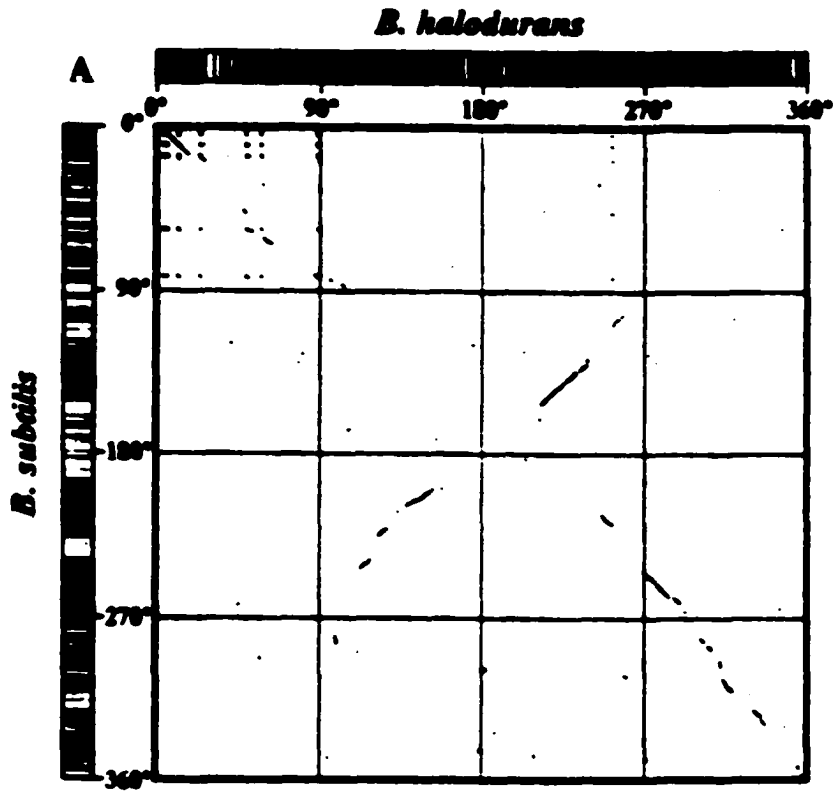
alkaline serine proteases and three neutral metalloprotease are synthesized as preproproteins. Subtilisin, an alkaline serine protease, occurs in greatest concentration in the media of *Bacillus subtilis*. Other minor serine proteases include bacillopeptidase F (*bpr*), *epr*, *vpr*. Serine proteases have a serine residue in the active site and are inhibited by hydroxyl-reactive organofluorides such as di-isopropylflurophosphate and phenylmethylsulfonyl fluoride. Metalloproteases require divalent metal cations and are inhibited by chelators of these cations, such as EDTA and 1, 10-phenanthroline. The metalloprotease genes include *mpr*, *nprB*, and *nprE* (24)

5. *Bacillus subtilis* and *Bacillus halodurans* Genetic Comparison

Bacillus subtilis is taxonomically related to the alkaliphilic *Bacillus halodurans* except for the alkaliphilic phenotype of being able to grow at pH 7-10.5 with optimal growth at pH above 9.5. The genome of *Bacillus halodurans* is 4,202,353 bp with 4,066 reported ORFs. Industrial applications of *Bacillus halodurans* have been investigated and several enzymes, including proteases, amylases, cellulases, and xylanases, have been commercialized (22)

A comparison of gene conservation between *Bacillus subtilis* and *Bacillus halodurans*, revealed 1,500 well conserved genes, that include several operons and the genes involved mobility, chemotaxis, protein secretion, cell division, the main glycolytic pathways, the TCA cycle, metabolism of nucleotides and nucleic acids, metabolism of coenzymes and prosthetic groups, DNA replication, RNA modification, ribosomal proteins, aminoacyl-tRNA synthetases, and protein folding. The dotplot between *Bacillus subtilis* and *Bacillus halodurans* shows that large segments of genomic similarity can be observed between these two bacteria (22)

Figure 3-Dotplot between *Bacillus subtilis* and *Bacillus halodurans* (22)



Reprinted with permission from Dr. Hideto Takami contained in *Nucleic Acids Research*, volume 28, coypyright 2000 by Oxford University Press.

6. Sporulation

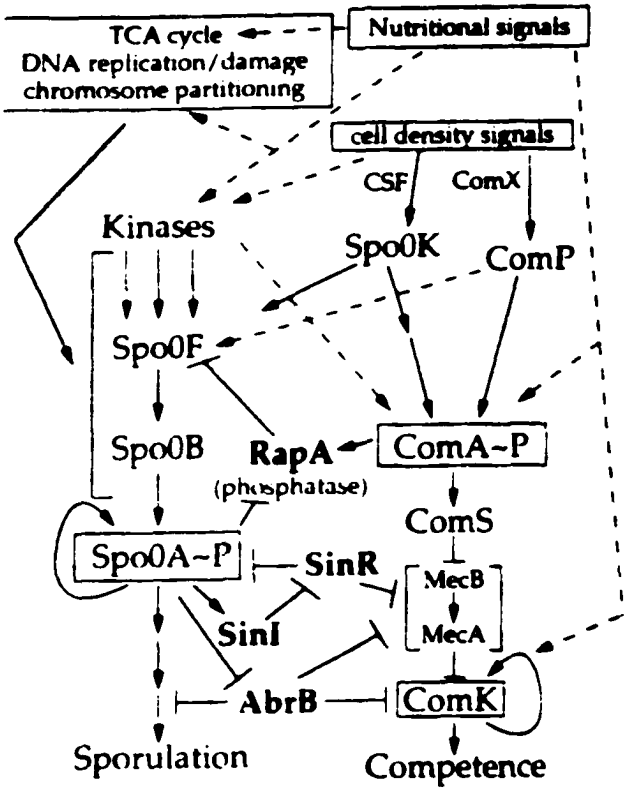
Sporulation, the ability of the vegatative cell to differentiate into a spore resistant to heat, radiation, drying, and chemicals, results in endospore formation through a process of asymmetric cell division that produces a mother cell and a smaller cell (forespore). Approximately 139 sporulation genes have been identified in *Bacillus subtilis*. In *Bacillus halodurans*, only 68 genes related to sporulation have been identified (22).

When in a nutrient deficient environment, *Bacillus subtilis* will initiate motility and associated chemotaxis, the induction of hydrolytic enzymes (proteases and carbohydrases), and competence for DNA uptake. If these responses do not restore *Bacillus subtilis* to acquire sufficient growth-sustaining nutrients, then it forms

endospores. The postexponential phase phenomena to include sporulation and competence have been studied in *Bacillus subtilis* for over forty years (30).

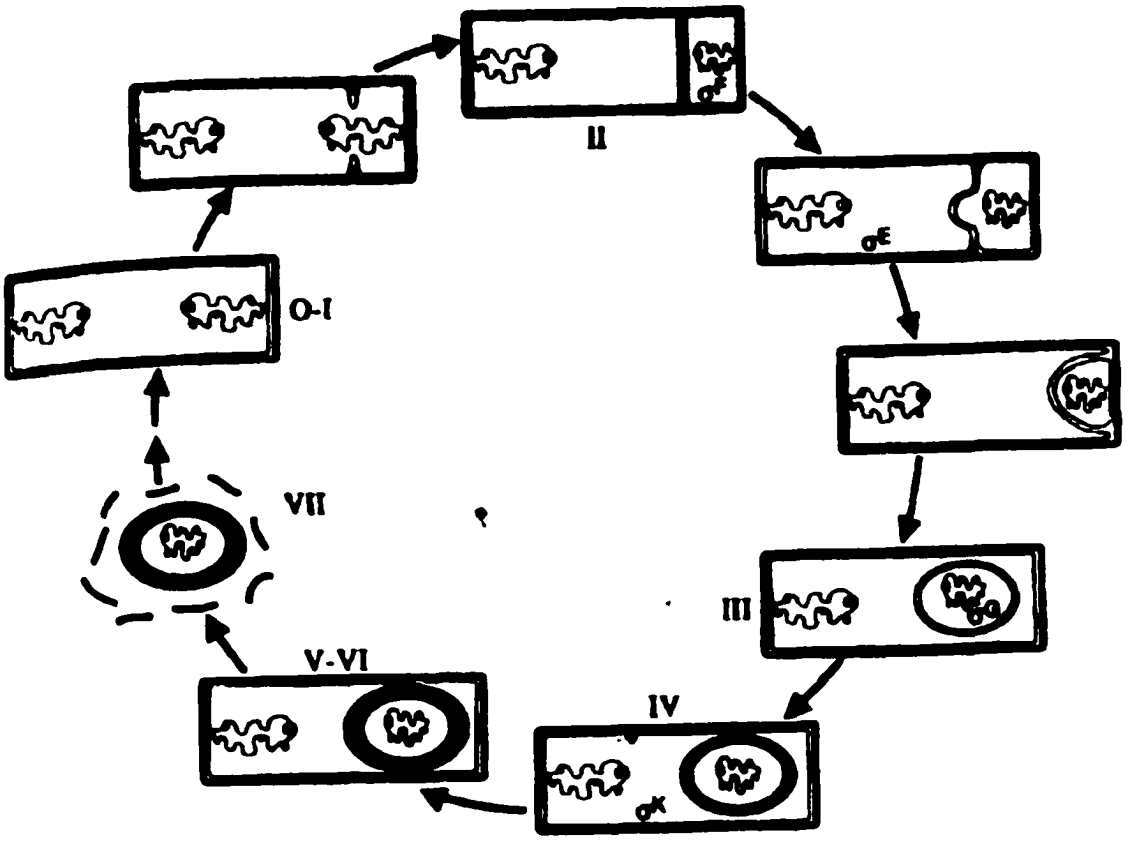
A phosphorelay component of the sporulation pathway integrates multiple signals, generated by conditions of nutrient depletion, high cell density, the TCA cycle, DNA replication, DNA damage, and the chromosome partitioning machinery, from histidine kinases, KinA, KinB, and KinC, which transfer phosphate to Spo0F, then to Spo0B, and finally to Spo0A (31). After histidine kinases indirectly phosphorylate Spo0A, phosphorylation of Spo0A is modulated by specific phosphatases, such as Spo0E, which dephosphorylates Spo0A~P and RapA which dephosphorylates Spo0F~P. Transcriptional regulation of *spo0E* and *rapA* involves a feedback loop and a connection between competence and sporulation (31).

Figure 4-Feedback Loop (32)



The process of sporulation begins immediately after nutrient deprivation and is completed in six to eight hours. The stages of sporulation include growth, axial filament formation, sporulation septum formation at the extreme polar position, prespore, cortex formation, formation of spore coats, mature spore formed, and spore liberation. The mother cell dies, and the prespore, now an immortal spore, gives rise to subsequent progeny (31). The spore can remain dormant for many years, but eventually, they germinate, lose their resistance properties and resume vegetative growth (32).

Figure 5-The Morphological Stages of Sporulation (31)



7. Competence

Competence for DNA uptake is the natural ability for DNA mediated transformation and homologous recombination (33). In *Bacillus subtilis*, competence develops from one percent to ten percent of the cells late in exponential growth or early in stationary phase. In *Bacillus subtilis*, cells that will develop competence receive signals from other cells in the population. Competence development is regulated by cell density signals and nutritional signals occurring in exponential phase or the stationary phase of the growth cycle (32).

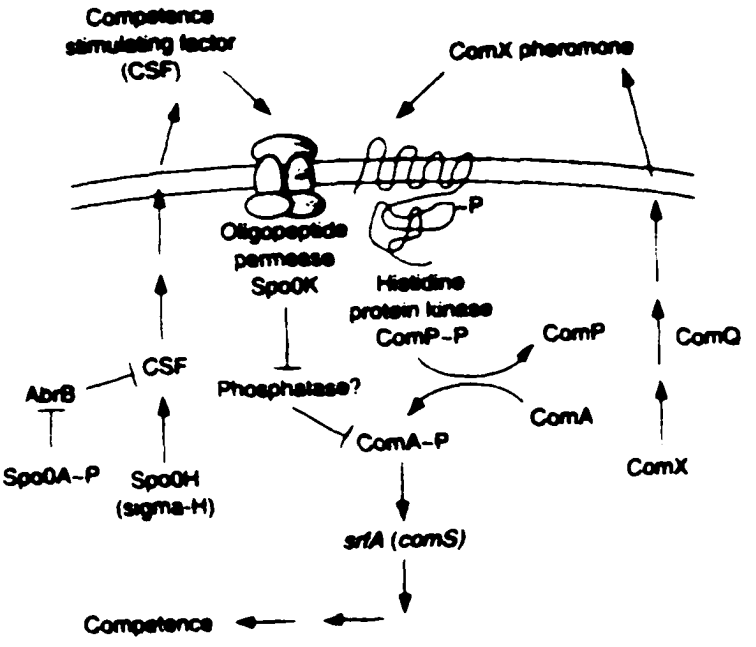
Bacillus subtilis has approximately forty genes involved in competence to include the *com* genes. Competence begins in *Bacillus subtilis* with DNA being bound, with no sequence specificity, to a number of sites on the cell surface. The DNA undergoes double-stranded cleavage and one strand is taken into the cell while the other strand is degraded outside the cell.

In *Bacillus subtilis*, competence stimulating factor (CSF) and ComX pheromone are the two extracellular peptide factors required for competence development. These two peptide factors accumulate in culture medium as cells grow to high density. Production of CSF is regulated by transcription factors that are required for competence and for the initation of sporulation, including SpoOA (the key sporulation transcription factor), AbrB (a transcription factor repressed by SpoOA), and SpoOH (encodes a sigma factor of RNA polymerase that is required for the initation of sporulation and the development of competence). The cellular response to CSF requires *comA* and *spoOk* (32).

In *Bacillus subtilis*, ComA is a transcription factor that is activated by phosphorylation and binds upsteam of the srfA promoter. ComP, a histidine protein kinase with an eight member-spanning domain, is required for the response to the ComX pheromone. The pheromone probably interacts directly with ComP to stimulate the autophosphorylation and transfer of phosphate to the ComA transcription factor. ComA is required for expression of *comS*, a gene necessary to activate the comK transcription

factor. The SpoOK oligopeptide permease is required for response to CSF and might act to transport CSF into the cell (32).

Figure 6-Competence (32)



Reprinted with permission from the *Annual Review of Genetics*, Volume 30 coypyright 1996 by Annual Reviews www.annualreviews.org

8. Metabolism

Genomic sequencing allows for direct comparison of the metabolism of different organisms. Plants, animals, and bacterial differ from each other in their biosynthetic and degradative pathways that dictate their nutritional demands. Bacteria have many nutritional requirements for growth to include carbon in the form of carbon dioxide, carbohydrates, and amino acids. Vitamins, purines, pyrimidines, and inorganic ions, such as nitrogen, phosphate, and magnesium are used by bacteria (5).

Bacteria use energy produced by respiration or fermentation. Aerobic respiration begins with glycolysis, which has pyruvate as an end product, continues with the

tricarboxylic acid cycle, which has carbon dioxide as an end product, and is facilitated by the electron transport system. This transport system has electron acceptors, such as, NADH, NADPH, and oxygen, embedded in the bacterial membrane. Electrons move through a series of electron carriers, such as, cytochromes, flavoproteins, and ubiquinones, with successively higher oxidation-reduction potenials. Energy is used to force a pair of electrons across an insulating membrane creating an electrochemical gradient. Oxidative phosphorylation results when those electrons return to the other side of the membrane (5).

Anaerobic respiration has nitrogen as a terminal electron acceptor. In fermentation, pyruvate, lactate, and oxaloacetate are products of the terminal electron acceptor. C4 and C5 compounds necessary for the synthesis of bacterial products are generated during the pentose phosphate shunt (5).

Of the twenty amino acids, ten are synthesized from two intermediates of the tricarboxylic acid cycle. Six amino acids come from oxaloacetate to include: aspartate, asparagine, methionine, lysine, threonine, and isoleucine. Four amino acids come from *alpha*-ketoglutarate to include glutamate, glutamine, arginine, and proline. Three glycolytic intermediates (pyruvate, 3-PGA, and PEP) are precursors to seven more amino acids. Alanine, valine, and leucine are synthesized from pyruvate. Serine, glycine, and cysteine are synthesized from 3-phosphoglycerate. The pentose phosphate pathway provides erythrose-4-phosphate, which binds with PEP to form the aromatic amino acids. Histidine metabolism starts with PRPP generated from pyrimidine biosynthesis (34).

Amino acids are used in the synthesis of proteins and other biomolecules. During starvation, amino acids often are degraded by a series of catabolic reactions that produces metabolic intermediates that can be converted into glucose or be oxidized by

the citric acid cycle. These carbon skeletons become pyruvate, acetyl CoA, acetoacetyl CoA, alpha-ketoglutarate, succinyl CoA, fumarate, and/or oxaloacetate (34)

9. *Bacillus subtilis* and *E. coli* Genetic Comparison

Differences exist between the Gram-positive and Gram-negative organisms. One difference between the two organisms is that Gram-positive organisms sporulate while Gram-negative organisms do not (33). Since Bacilli are rod shaped, they have a high surface to volume ratio allowing for higher rates of nutrient transfer. Therefore, the most abundant class of proteins found in *Bacillus subtilis* is the ABC transporters, which couple the hydrolysis of ATP to the translocation of solutes across a biological membrane. Seventy-seven transporters of which 71 are ATP-binding protein genes were found in the *Bacillus subtilis* genome (33). These transporters are extremely important in Gram-positive bacteria because these bacteria have an envelope consisting of a single membrane. There is no outer membrane or associated lipopolysaccharides (22)

Gram-positive and Gram-negative organisms also differ in their cell wall structure. The cell envelope in *Bacilli* consists of a multi-layered structure involving the protoplast membrane, cell wall, and possibly an S-layer. The protoplast membrane encloses the protoplasm and adds to cellular metabolism since it is the site of synthesis, energy generation, signal transduction, solute translocation, and DNA and phage binding. The cell wall acts as a sieving and ion exchange membrane. It is 20-50 nm thick which is about 20 times the thickness of the structural peptidoglycan layer of a typical Gram-negative bacterium, 1 to 2 layers thick. The S-layer is a proteinacious outer layer. The proteins and glycoprotein subunits are arranged in hexagonal or tetragonal arrays over the whole surface of the cell. During growth, S-layer proteins are shed into the culture supernatant (26).

The genome sequences of *Escherichia coli* and *Bacillus subtilis* provide a means of studying the evolutionary divergence of eubacteria into the Gram-positive and Gram-

negative groups which occurred approximately one billion years ago. Although conserved genes are found between *E. coli* and *Bacillus subtilis*, their relative chromosomal positions differ reflecting chromosomal rearrangements during a long evolutionary period (34). The complete genome sequence of the Gram-negative bacteria *E. coli* was found to contain 4,639,221 bp with 4,288 ORFs (35).

In *Bacillus subtilis*, thirty genes coding for ribosomal proteins are present in a large conserved cluster that show high degrees of sequence homology in *E. coli*. The gene order proceeds as follows in both *Bacillus subtilis and E. coli rplKAJL-rpoBC-rpsLG-fusA-tufA-rpsJ-rplCDWB-rspS-rplV-rpsC-rplP-rpmC-rpsQ rplNXE-rpsNH-rplFR-rpsE-rpmD-rplO-secY-rpmJ* (36). In *Bacillus subtilis*, there is an insertion of *adk*, *map*, and *infA* to the *spc* operon, which starts at *rplNXE* and ends with *rpmJ* above. In *E. coli* the gene order of the alpha operon is as follows *rpsMKD-rpoA-rplQ*. In *Bacillus subtilis*, the *rpsD* is translocated. Both organisms have the same gene order for the L13 operon (*rplM-rpsI*). In *Bacillus subtilis* and *E. coli* the RNA polymerase B-B' subunit appears as *rpoBC* (36).

In *Bacillus subtilis*, the replication origin has a gene order *gidBA*-orf50k-orf208-orf261-*rnpA-rpmH-dnaAN*-orf71-*recF-gyrB*. The identical gene order including 2 orfs (orf50k-orf261) appears in *E. coli* around its replication origin, however, *gidA* and *gidB* genes are located away from the cluster (36).

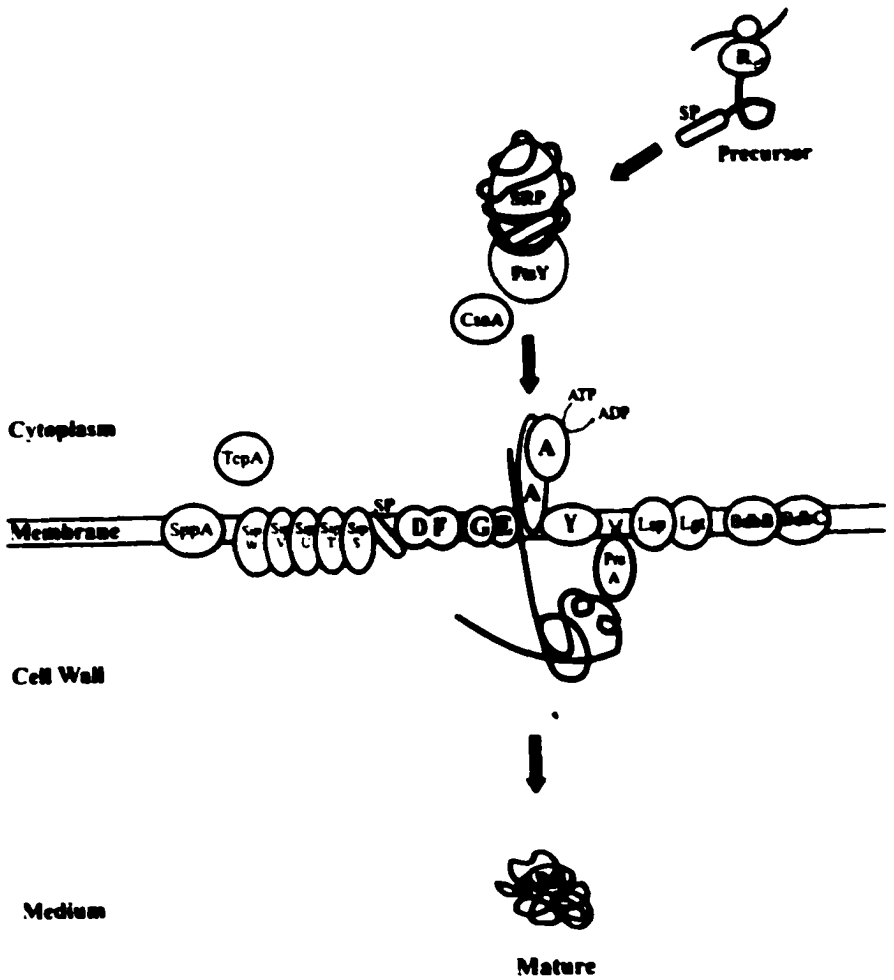
The ATP synthase operons conserved in *Bacillus subtilis* and *E. coli* appear as one operon. The F0F1 type ATP synthase genes are *atpBEFHAGDC* (37).

The *cta* operon in *Bacillus subtilis* and the *cyo* operon in *E. coli* are homologous with the following gene order *ctaB-ctaC-ctaD-ctaE-ctaF* in *Bacillus subtilis* and *cyoE-cyoA-cyoB-cyoC-cyoD-cyoE* in *E. coli*. The gene order in the two operons is identical except that the first gene *ctaB* corresponds to the last gene *cyoE* (36).

10. Secretion

Bacillus species have a high capacity to secrete proteins of the major secretory pathway directly into the growth medium (27). The absence of an outer membrane greatly simplifies the secretion process since exoproteins have to be translocated through a single membrane before reaching the culture medium. Secretion of the protein across the *Bacillus subtilis* cell envelop occurs in several steps. After the protein is inserted into the membrane, the protein is translocated across the membrane. Cleavage of the signal peptide can occur prior to or after completion of the translocation process. Folding of the mature protein occurs prior to its release from the membrane. Its passage through the cell wall results in secretion of the protein into the growth medium (24)

Figure 7-Components of Bacterial Protein-Secretion Pathways (24)



11. Introduction of Three Hyperthermophiles: *Aquifex aeolicus*, *Thermotoga maritime*, and *Methanococcus jannaschii*

Aquifex aeolicus is a hyperthermophilic bacterium with growth temperature maxima near 95 degree Celsius. The genera in this group are thermophilic, hydrogen-oxidizing, micro-aerophilic, obligate chemolithoautotrophs (an organism which uses an inorganic carbon source for biosynthesis and an inorganic chemical energy source). The 1,551,335 bp genome has been sequenced and shown to contain 1,512 ORFs. This organism has two 16S-23S-5S rRNA operons with a high G+C content of 65 percent. The G+C content of the stable RNAs is indicative of the high growth temperature of the organism (38).

Thermotoga maritime strain MSB8 is a hyperthermophilic, non-spore-forming, rod-shaped bacteria that grows at 80 degrees Celsius. This bacterium originally was isolated from geothermal heated marine sediment at Vulcano, Italy. Of the Eubacteria sequenced by 1999, this bacterium had the highest percentage (24 percent) of genes that were most similar to the Archaea in terms of gene content and overall genome organization. It has a 1,860,725 bp genome with 1,877 predicted ORFs where 21 percent had a highest similarity to *Bacillus subtilis*, and 15 percent had a highest similarity with *Aquifex aeolicus*. This bacterium metabolizes many simple and complex carbohydrates that include glucose, sucrose, starch, cellulose, and xylan. Several genes encoding proteins involved in the sequential degradation of xylan were present. Through conversion to fuels such as hydrogen, both cellulose and xylan have great potential as renewable carbon and energy sources (39).

The *Methanocaldococcus (Methanococcus) jannaschii*, strain DSM 2661, was the first archaeon genome completely sequenced and thus provides the first opportunity to compare the genetic complements and biochemical pathways among the three domains of

life (Archaea, Bacteria, Eukarya) from which all life forms evolved. This organism is a hyperthermophilic, methane producing, anaerobe with irregular cocci shaped cells. It was isolated from a depth of 2600 meters along the sea floor in the East Pacific Rise. This organism has adaptations for optimal growth at high temperature (85 degrees Celsius), and pressure (of up to more than 200 atmosphere) in moderately saline environments. This organism consists of three physically distinct elements: a large circular chromosome of 1,664,976 base pairs containing 1682 ORFs; a large extrachromosomal element of 58,407 base pairs containing 44 ORFs; and a small circular extrachromosomal element containing 16,550 base pair containing 12 ORFs. The archaeon genes involved in energy production, cell division, and metabolism are most similar to Eubacteria, however, the genes involved in replication, transcription, and translation, are most similar to Eukaryotes (40, 41)

C. Basic Knowledge

1. DNA Structure

DNA, the hereditary material, is a right-handed, double-helix composed of two single-stranded, anti-parallel polynucleotides bonded together by hydrogen-bonding following the Watson-Crick base pairing rules (42, 43, 44).

A nucleotide consists of the nitrogenous base, one or more phosphate groups, and a ribose sugar. Their alternating sugar and phosphate groups form a backbone for structural support. The bases are on the inside of the helix with the phosphates on the outside. The bases are paired with one of the pair being a purine and the other a pyrimidine for bonding to occur (45, 46, 47).

The four common ribonucleotide 5'-triphosphates are adenine (ATP), cytosine (CTP), guanine (GTP), and thymine (TTP). The ribose triphosphate portion and the linkages are identical in all four. The nitrogenous base is a derivative of pyrimidines

(adenine and guanine) and purines (cytosine and thymine). The adenine pairs with the thymine via two hydrogen bonds. The cytosine pairs with the guanine via three hydrogen bonds. The sugar in deoxyribonucleotide is deoxyribose meaning that the sugar lacks an oxygen atom. Of the two single-stranded polynucleotides, since the strands are anti-parallel, one chain is the complement of the other. Thus, if the actual order of the bases on one of the pair of chains is known, then the order of the bases on the complementary strand are known because of base pairing (45, 46, 47)

2. Replication

Chromosomal replication occurs once per cell division cycle as the growth rate decreases and the cell mass increases. DNA replication is catalyzed by DNA polymerase III, *polC*, an enzyme that synthesizes DNA in the 5' to 3' direction with 3' to 5' exonuclease proofreading ability. The DNA polymerase is linked to the DNA template by a sliding protein clamp known as the *beta*-clamp, DnaN. DNA replication begins at the origin (*oriC*) with DnaA, the recognition protein, binding to its DNA binding sites in DnaA boxes and forming a nucleoprotein structure that unwinds an adjacent AT-rich region, melts the DNA, and signals the replisome to come to the *oriC*. At *oriC*, the replisome and DnaC, the replicative helicase, are loaded onto the DNA at the AT-rich region. DNA helicase unwinds DNA strands with the help of ATPand DNA gyrase. This expands the single-stranded DNA bubble and introduces a negative supercoil. DnaB, DnaD, DnaI and primase help load helicase with DnaI intereacting strongly with the helicase (19).

DNA replication proceeds bidirectionally at a constant speed of ~500 to 1,000 nucleotides per second. The leading and lagging strand DNA polymerases are held together through interactions with a dimer of the scaffolding protein, DnaX. The leading

strand is synthesized continuously. Lagging strand, synthesized discontinuously, occurs in stretches of approximately 1000 nucleotides called Okazaki fragments. Primase synthesizes the primers needed for synthesis of the Okazaki fragments. The primers for each fragment are removed, and DNA polymerase I fills in the intervening gaps. DNA polymerase I is poorly processive with only 10 to 50 nucleotides per template binding event. DNA polymerase I has both a 3' to 5' proofreading exonuclease and a 5' to 3' exonuclease for primer removal. The DNA ligase joins the Okazaki fragments (24)

Since DNA duplex is a right-handed helix, positive supercoils are twisted in the same direction as the helix to overwind the helix. During DNA replication, DNA gyrase solves the problem of overwinding. It removes the positive supercoils that form in the unreplicated DNA ahead of the replication fork and converts them into negative supercoils. DNA gyrase makes a double-stranded break in the DNA, passing an unbroken portion through the gap and resealing the break. Single stranded binding proteins maintain and protect the single strands produced by helicase unwinding of the DNA duplex (24)

When the replisome encounters DNA that disrupts replication, the process of replication begins again. The following genes are used for the collapse of the replication folk: *priA*, *dnaC*, *dnaI*, *dnaB*, and *dnaG*. *priA* binds to the collapsed fork and has helicase activity. In addition to those genes, the genes used in homologous recombination and double-stranded break repair may also help with a collapsed fork (24).

3. Transcription

The central dogma of biochemistry is that DNA is transcribed into RNA (the production of rRNA, tRNA, mRNA, and snRNA from the DNA template) which is translated into protein (protein synthesis determined by the nucleotide sequence in DNA). Transcription is divided into several distinct phases to include promoter recognition, initiation, elongation and termination. Transcription is catalyzed by RNA polymerase

core enzyme, which has two *beta* and one *alpha* subunit, and has some accessory polypeptides such as, a *gamma* protein, and two *omega* polypeptides. This core enzyme has a high affinity for double-stranded DNA and can bind at any site to form a stable enzyme-DNA complex. A transient RNA-DNA hybrid duplex is formed in the region of the enzyme-DNA complex (24, 48).

The process of transcription begins when the RNA polymerase holoenzyme (the core enzyme) and sigma factors bind to a specific promoter region, conserved sequences near -35 and -10 (actually -70 to +20) relative to the start of transciption. Once bound to the promoter, the holoenzyme selectively unwinds the DNA and adds the first RNA nucleotide to the 5' end of the RNA transcript. Transcription starts with the polymerase synthesizing a strand of RNA complimentary to the DNA template, making it an exact copy of the nontemplate strand except for the substitution of uracil for thymine. After the polymerase clears the promoter region, sigma factor is released, and the polymerase becomes highly processive in RNA synthesis (24, 48)

In the elongation phase, RNA polymerase moves along the DNA helix, unwinding the strands to expose the single-stranded DNA template, supervising the correct base-pairing between incoming rNTPs and the template, and linking the nucleotides in the 5' to 3' direction with elimination of pyrophosphate. The DNA rewinds as the RNA peels away from the template. Elongation continues until the core enzyme recognizes a termination signal (24, 48).

In bacteria, there are two kinds of transcription terminators that vary in efficiency and mechanism of action. In a *rho*-independent terminator, the RNA transcript of the terminator region contains a GC rich Palindromic sequence which can form a stem and loop structure followed by a row of uridine residues. The stem and loop structures cause the RNA polymerase to pause at the uridine residue sequence and may cause the release of the RNA transcript. In a *rho*-dependent terminator, there may be a stem and loop structure (not particularly GC rich), but there is no row of uridine residues. Termination

of the transcript requires a *rho* protein that is active as a hexamer with RNA-dependent NTPase activity. This *rho* protein interacts with the transcript (24, 48).

4. Translation

After transcription of mRNA, the mature mRNA acts as a template for the assembly of the polypeptide chain. Amino acids cannot recognize their specific codons in the mRNA directly so each must be linked to a specific tRNA containing the correct anti-codon. Ribosomes organize the interaction between charged tRNA and the corresponding mRNA and their peptide bond formation. Each ribosome has the acceptor (A) site and the peptidyl (P) site for charged tRNA. A ribosome binds to the mRNA and moves along it so that each codon is exposed to the A site (48).

The amino acid is connected to its correct tRNA in a two step process. Activation occurs when the amino acid binds to a specific aminoacyl-tRNA synthetase to form an enzyme-bound aminoacyl AMP complex. Transfer occurs when the aminoacyl group is transferred to the 2' or 3' position of the terminal adenosine residue in the appropriate tRNA releasing AMP and the synthetases (48).

The initiator codon, AUG, is recognized by a distinct initiator tRNA that is charged with Met (methionine). The three translation initiation factors in *Bacillus subtilis* are IF1, IF2, and IF3 encoding *infA*, *infB*, and *infC* respectively. IF-1 stimulates IF-2 to form a pre-initiation complex with Met-tRNA in the presence of GTP. This complex binds to the 30S ribosomal subunit when the 30S ribosomal subunit binds to mRNA. IF-3 is necessary for the binding of 30S ribosomal subunit to mRNA molecule at the Shine-Dalgarno sequence, 3-9 purine nucleotides long. The 16SrRNA in the 30S ribosomal subunit has a complementary sequence to the Shine-Dalgarno sequence near the 3'end. This sequence aligns the ribosomal P site with the correct initiation codon. IF-1 is released, and the 50S ribosomal subunit then binds to the 30S initiation complex, releasing IF-3 to form the complete 70S ribosome. Finally, GTP is hydrolysed and IF-2 is released (24, 48).

Elongation of the polypeptide chain occurs in three steps that are repeated at each

codon. In the first step, decoding step, the translation elongation factor EF-Tu, *tufA*, brings aminoacylated tRNA (the second codon) into the A site of the elongating ribosome. EF-Tu-GTP complexes with any aminoacylated tRNA to form the ternary complex aminoacylated-tRNA-EF-Tu-GTP, which binds at the A site of a ribosome while the P site is occupied by an fMet-tRNA (24, 48).

In the second step of translation, transpeptidation, a peptide bond is formed between the alpha amino group of the aminoacylated tRNA in the A site and the carboxyl group of the fMet-tRNA in the P site leaving an uncharged tRNA in the P site and the dipeptidyl-tRNA in the A site (24, 48).

In the third step, translocation, the ribosome advances in the 5' to 3' direction one codon relative to the mRNA-dipeptidyl-tRNA. The uncharged tRNA in the P site is ejected. The dipeptidyl-tRNA in the A site advances to the P site. The vacated A site is positioned next to the next codon (24, 48). Many antibiotics interfer with the steps of the elongation process,

In *Bacillus subtilis*, the translation termination genes are *prfA*, a releasing factor that recognizes UAG and UAA nonsense codons, and *prfB*, a releasing factor that recognizes UAA and UGA. Binding of the releasing factor to a termination codon results in hydrolysis of the ester bond between the tRNA and the polypeptide chain. The *frr* gene is needed for dissociation of the ribosome from mRNA after termination (24, 48).

5. Recombination

Recombination is a highly conserved mechanism used in genetic transformation in competent cells and DNA-damage repair. The addAB protein is a helicase and nuclease binds to both strands of double stranded DNA and unwinds and degrades the duplex until the sequence 5'-AGCGG-3' is reached. The 5' strands continues to be degraded. Degradation of the 3' strand stops once that sequence is reached. With the help of single stranded binding protein, the recA protein catalyzes homologous base pairing by allowing the single stranded DNA to invade the double stranded DNA (forming a D-loop). The

single stranded DNA pairs with the complementary region of the double stranded DNA and displaces the homologous strand. The recA, ruvA, and ruvB proteins extend the heteroduplex DNA via branch migration to form a Holliday junction. The ruvA protein binds to the junction thereby changing the structure of the junction into a fourfold symmetrical structure. Finally, the strands of the junction are cleaved by a resolvase (24).

6. DNA Sequencing

The ability to generate the sequence of a large bacterial genome in a reasonable amount of time was made possible by three discoveries. Sanger sequencing, the polymerase chain reaction, and automated fluorescent DNA sequencers (49). Initially, radioactivity was incorporated using alpha-32 phosphorus dATP but later 5'-radiolabeled primers were included in the extension reaction that creates the nested fragment sets. However, since radioisotopes are hazardous, costly, unstable, and require manual data acquisition, fluorescent dyes and automated data collection instruments are used today. The two categories of fluorescent sequencing are dye primer sequencing and dye terminator sequencing. The advantage of dye primer sequencing is the use of any DNA polymerase as long as it accepts dideoxynucleotide triphosphates (ddNTPs) as substrates. The disadvantage is that in dye primer sequencing the high level of product fold-back compression results in less accurate DNA sequence data.

In dye terminator sequencing, the fluorescent dyes are attached to the ddNTPs and are incorporated into the DNA fragments in a single extension/termination reaction using an unlabeled primer. As in fluorescently-labeled dye primers, the reaction occurs in one tube, and four spectrally resolvable fluorescent dyes are required. The advantages include reduced product fold-back compression because of the large bulky fluorescent group on the dideoxyterminators and unlabeled primers can be used. The disadvantage is

that the DNA polymerase must be genetically altered to work with each set of dye labeled terminators.

The discovery of capillary sequencers allowed the sequencing capacity to increase tremendously. Two basic approaches are used to sequence genomes to include clone by clone sequencing and whole genome shotgun sequencing. In the clone by clone approach, the genome is divided and inserted into bacterial artificial chromosomes capable of holding approximately 150,000 base pairs (49). Since no phagemid, cosmid, bacterial artificial chromosome (BAC), yeast artificial chromosome (YAC), or bacteriophage artificial chromosome libraries existed for *Bacillus stearothermophilus* prior to the start of this sequencing project, a whole genome shotgun sequencing approach was used for this project. In the whole genome shotgun method, pUC clones are prepared directly from genomic DNA rather than from mapped BACs (49).

There are many steps involved in shotgun sequencing including the isolation of the DNA, fragment shearing of DNA, end-repair of overhangs, phosphorylation of ends, size selection of 2 kb to 6 kb DNA on an agarose gel, extraction and precipitation of the DNA, ligation of DNA into a pUC18 vector, electroporation of chimeric plasmids into *E.coli* cells, blue/white color assay of chimeric plasmids, growth of chimeric plasmids, subclone isolation, and subclone sequencing (2).

The first round of shotgun DNA sequencing used a primer that was complementary to a site on the pUC18 vector near the site of insertion of the cloned DNA. After the shotgun sequencing reactions were assembled using Phred and Phrap, a second primer was synthesized that was complementary to a segment of about 50 nucleotides from the end of the segment that was just sequenced. This process typically was repeated until the gap was closed and the entire insert was sequenced. Primers were generated with the aid

of a computer program, PrimOU. These customized primers were synthesized on a laboratory chemical dispensing robot, the LCDR/Mermade, which had the capacity to synthesize two standard 96 well plates of oligos in a single run using standard phosphoramidite chemistry (42).

The nested fragment sets produced by the dideoxysequencing reactions were electrophoresed on the Perkin Elmer ABI Prism 377 DNA Sequencer, 3700 DNA Analyzer, and the Amersham Pharmacia Biotech MegaBACE 1000 DNA Sequencing System. The gel reads were transferred to a UNIX based Sun Microsystems Sparc Station 20 for analysis and assembly. Production sequencing software included the Phred base caller, the Phrap assembler, and the Consed viewer and editor. Data analysis tools included BLAST, Crossmatch, Dotter, Repeatmasker, Exgap, and Finisher. The gene prediction tools included Genscan, tRNAscan, Artemis, KEGG, MIPS, SubtiList, BSORF, and Deepstar.

Sanger sequencing (also called the chain termination or dideoxy method) is an enzymatic procedure by which DNA chains of varying lengths were synthesized in the presence of the four dNTPs, DNA polymerases, and the four ddNTPs DNA replication terminators. The ddNTPs differ from dNTP in that they lack a hydroxyl residue at the 3' position of deoxyribose. Thus, when a trace amount of one ddNTP is included with the four conventional dNTPs in a reaction mixture for DNA synthesis, there is a competition between extension of the chain and infrequent, but specific, termination. The products of the reaction thus are a nested fragment set or a series of oligonucleotide chains whose lengths vary based on the distance between the terminus of the primer used to initiate DNA synthesis and the sites of premature termination (43, 44).

During this dissertation research, a modified Sanger method was used. Here, dye-

terminator sequencing incorporated fluorescent labels into DNA sequencing reaction products by using dye-labeled ddNTP terminators. Since each ddNTP is tagged with a different fluorescent dye (rhodamine110, rhodamine-6-G, tetramethyl rhodamine, rhodamine X) a single reaction mixture can be used thus simplifying the reaction set-up. Thus, the growing chain is simultaneously terminated and labeled with the dye that corresponded to that base. Advantages of using dye-labeled terminators are that an unlabeled primer is used and four sequencing reactions can be performed simultaneously in one tube. Dye Terminator Cycle Sequencing kits contained the AmpliTaq DNA polymerase enzyme, sequencing buffer, dNTP mix, dye-labeled ddNTP (terminators), control template, and control primer. AmpliTaq DNA polymerase is an ultra-pure thermostable polymerase obtained by expression of a modified form of the *Thermus aquaticus* DNA polymerase gene in *E. coli*. The enzyme provided rapid nucleotide incorporation, and has no 3' to 5' exonuclease activity. AmpliTaq retained substantial activity after repeated exposure to high temperatures of between 94 and 96 degree Celsius, thereby making it an ideal enzyme for cycle sequencing (2, 50).

7. Sequencers

The ABI Prism 377 automatically analyzed DNA molecules labeled with any of 4 different fluorescent dyes. These vertical slab gels allowed loading of 64 lanes per gel with a run time of 7 hours. When the sequencing reactions were electrophoresed on the ABI Prism 377, the beam of an argon laser mechanically scanned a horizontal segment at the bottom of the gel. As the fluorophore-labeled fragments undergoing electrophoresis passed through this region they were excited to fluorescence. The emitted light was collected and focused onto a detector after passing through four different filters on a rotating wheel. The identity of the terminating nucleotide analog of each passing

fragment was identified because of the unique fluorophore associated with it. The temporal order of passing fluorescent molecules was automatically translated into linear DNA sequence and stored by the computer (51).

These slab gels were tracked on the Power MacIntosh computers. In contrast, the ABI Prism 3700 and the MegaBACE employed capillary electrophoresis which had the greatest advantages of eliminating lane tracking errors and electrophoresing 96 simultaneous samples in a shorter time as the capillary electrophoresis used narrow-bore capillaries filled with a separation matrix. Both the ABI Prism 3700 and the MegaBACE provided automatic injection (electrokinetic injection) and electrophoresis of samples into 96 capillaries, each capillary corresponded to a well, with laser excitation of the fluorescently labeled samples (52).

With the ABI Prism 3700, an autoloader transfers up to 96 samples per run to the injection wells for electrokinetic injection, where the loading end of a polymer-filled capillary was dipped into the sample suspension. Then, as a voltage was applied across the sample, typically for less than a minute, a portion of the negatively charged DNA migrated into the polymer in the capillary. As the DNA molecules in the sample occupied very little volume, the volume of the sample remained about the same. After injection, the remaining sample was replaced with buffer and electrophoresis was initiated (53).

The MegaBACE used a linear polyacrylamide (LPA) long read matrix, with DYEnamic energy transfer (ET) terminators, and the Thermo Sequenase II DNA polymerase. LPA matrix enabled average read lengths in excess of 500 base pairs with PHRED accuracy scores of greater than 20 (54).

The capillary instruments required the use of energy transfer dye-labeled reaction product nested fragment sets. Therefore, each dideoxy terminator was labeled with two dyes. One of these dyes, fluorescein, has a large extinction coefficient at the wavelength (488 nm) of the argon ion laser in the instrument. The fluorescein donor dye absorbed

light energy from incident laser light and transfered the collected energy via radiationless energy transfer to an 'acceptor' dye. Each of the four chain terminators, ddG, ddA, ddT or ddC, has a different acceptor dye coupled with the fluorescein donor. The acceptor dyes emitted light at their characteristic wavelengths. The instrument detected fluorescence allowing identification of the terminator nucleotide. As a result, energy transfer-based terminator based DNA sequencing had signal strengths several fold greater than those obtained with conventional single-dye labeled molecules (54, 55).

D. DNA Base Calling and Assembly

1. Phred

The DNA sequence data collected on the ABI Prism 377, 3700, and MegaBACE DNA sequencers was transferred to a UNIX-based Sun Microsystems Sparc Station 20 for analysis and assembly using Phred/Phrap/Crossmatch Programs (56, 57, 58). Here, the individual sequence trace files were examined by the Phred program which called the bases, assigned quality values to each base, and wrote the base calls and quality values to output files which then were assembled by Phrap. The algorithm that Phred uses to call bases is a fast Fourier transformation method. This was a two-step approach. The first approach examined the four base traces in the region surrounding each point in the data set and predicted a series of evenly spaced locations. This prediction considered compressions, dropouts, or other factors shifting the peaks from their true location. The second approach examined each trace to determine the center of the peak and the areas of these peaks relative to their neighbors. A dynamic programming algorithm was used to match the observed peaks detected in the second step with the predicted peak locations found in the first step. Phred quantified the trace quality by assigning four or five quality values that were related to the base call error probability (56).

2. Phrap

Phrap (phragment assembly program), developed by Phil Green, is the program for assembling shotgun DNA sequence data by constructing contiguous overlapping

sequences as a mosaic of the highest quality regions of reads, rather than a consensus. A contig is an uninterrupted group of linked and ordered clones. The portion of each read, which is accurate enough to align against other reads, is used in the assembly. Phrap also uses a combination of user-supplied and internally computed data quality information to improve accuracy of assembly in the presence of repeats as well as input from two programs, SWAT and CrossMatch. SWAT searches one or more DNA or protein query sequences against a sequence database using the Smith-Waterman algorithms (57). Crossmatch screens out vector sequences, which produces vector masked versions of the reads, and performs sequence comparisons of repeat elements against the contigs within a database or against the sequence in another database, such as that of the entire *E. coli* host genome, to flag any repeats or any contaminating host genomic sequences (58).

3. Consed

Once the sequence data is assembled into a Phrap database, it is viewed using the Consed program. Consed is a graphical tool that allows viewing and editing of the Phrap sequence assembly and assessing the quality of the individual sequence reads by assigning a color from white to grey. Consed allows the picking of custom primers using various stringency values using the quality values assigned by Phred and Phrap and the consensus sequence created by Phrap (59, 60).

E. DNA Sequence Analysis

1. BLAST

The basic local alignment search tool (BLAST) program suite is a statistically driven search method that identified regions of similarity between the gapless alignment of the DNA or protein query and database sequences by calculating maximum segment pair (MSP) scores, which are the highest scoring pair of identical length segments between two sequences (61). BLAST uses a heuristic algorithm, which allows for identification of

short homologous sequences within databases. The BLASTN window indicates GenBank nucleotide sequences homologous to the sample nucleotide sequence. The BLASTX window indicates GenBank amino acid sequences homologous to the sample amino acid sequence, the accession numbers of the sequence producing significant alignments, the scores, and the error values (62). The BLASTP output has four parts: 1) an introduction giving where the search occurred, what database was searched and which query sequence was compared, 2) a list of the sequences in the database containing segment pairs whose scores were least likely to have occurred by chance; 3) a display of the alignments of the high-scoring segment pairs showing identical and similar residues, and 4) a complete list of the parameter settings used for the search (62).

2. PowerBLAST

PowerBLAST is a program that performs analysis by sequence similarity search. This program processes sequences of any length because it divides long query sequences into overlapping fragments and merges the results after searching. The program performs "masking" to reduce or eliminate misleading results. PowerBLAST was used for the four cosmids in the Human Genome Project but not used for analysis in the *Bacillus stearothermophilus* sequencing genome project as it was superseded by the more recent release of the BLAST program from the National Center for Biotechnology Information (NCBI) (64).

3. Genscan

Genscan, a gene identification program, determines the most likely gene structure using a probabilistic model of the gene structural and organism specific compositional properties of its genomic DNA. The program treates the most general case in which the sequence could contain no genes, one gene, or multiple genes on either or both DNA

strands. Partial genes, as well as, complete genes are considered. It also accounts for typical gene density, the typical number of exons per gene, the distribution of exon sizes for different types of exon, the reading frame-specific hexamer composition of coding regions versus the reading frame-independent hexamer composition of introns and intergenic regions, the position-specific composition of the translation initiation (Kozak) and termination signals, and of the TATA box, cap site and poly-adenylation signals. Genescan predicts protein coding genes versus tRNA or rRNA genes (45).

4. REPEATMASKER

REPEATMASKER screens for low complexity DNA sequences and interspersed repeats known to exist in mammalian genomes for example: (ALUs, MIRs, LINE1 and LINE2). This program determines the copy number of each element, the length of the repeat, and the percentage of sequence for each repeat element in addition to the total length of sequence in the directory and the G+C level (65).

5. EXGAP

Dr. Axin Hua in our laboratory designed a program called EXGAP that orders and orientates the contigs by pairing the forward and reverse reads from the same subclones (66). The subclones identified in the EXGAP program then are used in uniplex PCR and multiplex PCR (MPCR) followed by sequencing with either the extendc or extends primers picked by PrimOU or primers hand-picked using Consed's Primer Picking Preference to close many gaps. PrimOU is the University of Oklahoma's version of the University of Texas-Southwestern Medical Center's Primo Primer Picking Program.

6. Finisher

A program, Finisher, was constructed by Ziyun Yao, in our laboratory, and provides all possible gaps and a list of the subclones covering each gap (67).

7. Dotter

Dotter is a graphical dotplot program that provides a detailed comparison of two sequences by displaying the entire sequence for a comparison of every residue in two sequences. This program is used to visualize the similarity between the sequences and reveal repeated domains. When used in conjunction with Crossmatch, dotplot also is used to verify the ordering of contigs (68, 69).

8. tRNAscan

The tRNAscan v1.3 program uses multistep weight matrix algorithm for identification of tRNA genes. This program also identifies pseudogenes by a low score resulting from subtracting a tRNA similarity score to the primary structure-only model from the base-pairing cloverleaf secondary structure-only score (70).

9. Artemis

The Sanger Center developed the Artemis software program that was useful for annotation of the open reading frames in the *Bacillus stearothermophilus* sequencing project. Artemis is a DNA sequence viewer and annotation tool that allowed visualization of sequence features and the results of analyses within the context of the sequence and its six frame translation. The main Artemis edit window has several sections to include the menu, the name of the DNA segment analyzed, an overview of the sequence and the features from the active entries, in addition to a textual summary of the active features (71, 72).

F. Genome Metabolic Analysis

1. Monica Riley

In 1997, the complete genome sequence of *Escherichia coli*, strain K-12 was published

(35). However, before the paper was published, Dr. Monica Riley began annotating its genome. Dr. Riley compiled an extensive classification of *E. coli* gene products and their function in the cell from at least two sources. The gene names, also used by Kenneth E. Rudd and Gerard Bouffard (73, 74, 75), were listed and compiled into a bibliography of *E. coli* genes, phenotypes, and gene products by Barbara Bachmann (76). Dr. Riley's major categories include intermediary metabolism, biosynthesis of small molecules, macromolecule metabolism, cell structure, cellular processes, and other functions (77).

Table 1–Dr. Riley's Cellular Functional Categories

Classification of E. coli gene products	No. of genes
Category of function	
I. Intermediary metabolism	
A. Degradation	175
B. Central intermediary metabolism	54
C. Respiration (aerobic and anaerobic)	55
D. Fermentation	39
E. ATP-proton motive force interconversion	9
F. Broad regulatory functions	43
II. Biosynthesis of small molecules	
A. Amino acids	
1. Glutamate family/nitrogen assimilation	23
2. Aspartate family, pyruvate family	53
3. Glycine-serine family/sulfur metabolism	17
4. Aromatic amino acid family	24
5. Histidine	9
B. Nucleotides	
1. Purine ribonucleotides	20
2. Pyrimidine ribonucleotides	11
3. 2'-Deoxyribonucleotides	7
4. Salvage and interconversions	19
C. Sugars and sugar nucleotides	13
D. Cofactors, prosthetic groups, electron carriers	
1. Biotin	8
2. Folic acid	9
3. Lipoate	2
4. Molybdopterin	5
5. Pantothenate	4
6. Pyridoxine	4
7. Pyridine nucleotides	5

8. Thiamine	11
9. Riboflavine	3
10. Thioredoxin, glutaredoxin, and glutathione	5
11. Menaquinone and ubiquinones	15
12. Heme and porphyrins	11
E. Fatty acids and lipids	35
F. Polyamines	6
III. Macromolecule metabolism	
A. Synthesis and modification	
1. Ribosomal and "stable" RNAs	25
2. Ribosomal proteins and their modification	64
3. Ribosomes and their maturation and modification	8
4. tRNAs, aminoacyl-tRNA synthetases and their modification	133
5. RNA synthesis, modification, and DNA transcription	19
6. Basic proteins	4
7. DNA replication, restriction/modification, recombination, and repair	99
8. Proteins (translation and modification)	22
9. Polysaccharides (cytoplasmic)	5
B. Degradation of macromolecules	
1. RNA	12
2. DNA	8
3. Proteins	20
IV. Cell structure	
A. Membrane components	20
B. Murein sacculus	36
C. Surface polysaccharides and antigens	35
D. Surface structures	50
V. Cellular processes	
A. Transport/binding proteins	216
B. Cell division	33
C. Chemotaxis and mobility	15
D. Protein secretion	19
E. Osmotic adaptation	21
VI. Other functions	
A. Cryptic genes	31
B. Phage-related functions and prophages	31
C. Colicin-related functions	16
D. Plasmid-related functions	5
E. Drug/analog sensitivity	35
F. Radiation sensitivity	4
G. DNA sites	20
H. Adaptations to atypical conditions	20

2. KEGG

The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a knowledge base for the systemic analysis of gene functions, linking genomic information with higher order functional information that was developed at Kyoto University of Japan. The three KEGG databases are the GENES database, containing a collection of genes for all the genomes, the PATHWAY database, with a representation of higher order functions in terms of the network of interacting molecules, and the LIGAND database, which contained a collection of chemical compounds within a cell, enzyme molecules and enzymatic reactions. The functional assignments in KEGG resulted from a process of linking a set of genes in the genome with a network of interacting molecules in the cell, the pathway. Each reference pathway can be viewed as a network of enzymes or a network of E.C. numbers. The metabolic pathway is a network of enzyme-enzyme relations, and the regulatory pathway was a network of direct protein-protein interactions, such as binding and phosphorylation, and classes of indirect protein-protein interactions, which were relations of transcription factors and transcribed gene products via gene expressions (78, 79).

3. MIPS

For a comparative, independent analysis, the *Bacillus stearothermophilus* working draft was submitted to the Munich Information Centre for Protein Sequences (MIPS), a bioinformatics group of the GSF (National Research Center for Environment and Health) at the Max-Planck-Institute of Biochemie. The MIPS group used the protein extraction, description, and analysis tool (PEDANT) program that utilized bioinformatics methods to provide complete functional and structural characterization of protein sequence sets from individual sequences to complete genomes. PEDANT depends on the ORPHEUS

algorithm to predict genes in bacterial DNA sequences. Functional assignments of gene products were made based on PSI-BLAST similarity searches supplemented by detection of PROSITE patterns, motifs, and comparisons with conserved sequence blocks and protein domains. Sequences are characterized by PIR keywords and E.C. numbers. The MIPS homepage, accessible at URL http://pedant.gsf.de, contains a large selection of computational analysis of completed genomes and experimental and unfinished genomic sequences (80).

4. SubtiList

Ivan Moszer manages the SubtiList database, which contains the *Bacillus subtilis* genomic information and is maintained by the Institut Pasteur, France (81). SubtiList is implemented on a UNIX system using the Sybase relational data scheme in which genes are grouped together into tables and linked by logical relationships. The genomic features are located relative to their chromosomal sub-divisions, which then are mapped to their chromosomal location. This database also provides links to other websites such as SWISS-PROT and European Molecular Biology Laboratory (EMBL) (81).

5. BSORF

A *Bacillus subtilis* database, BSORF, is maintained by GenomeNet for the research community in Japan and allows for searching of ORFs by entering the gene name, accession number, description, category number, or KEGG pathway (82).

6. DEEP STAR

A *Bacillus halodurans* Extremo Base database, DEEP STAR, is maintained by Japan Marine Science and Technology Center (JAMSTEC), a general oceanographic research institution, for the purpose of conducting research of deep sea microorganisms. Basic research is being conducted to understand extreme microorganisms that live in the

environment of the dark sea bottom with low temperature, high pressure, where hot water ejects and cold water seeps out, and where no solar energy exists (83).

7. Get_Amino_Stats

In an effort to obtain the relative amino acid composition of individual proteins in the *Bacillus stearothermophilus* genome, the ORFs of interest were located using PEDANT and processed through Jim White's "get_amino_stats" Perl script to determine the relative usage for each amino acid in each predicted ORF

8. Codon usage

In bacteria, gene codon usage often can be correlated with the level of individual gene expression. Genes with different codon usage may indicate a recent horizontal gene transfer and thus intraspecies codon usage can provide information on genomic evolution. A Perl script written by Jim White was used to calculate codon usage

9. Di- and Trinucleotide Frequencies

It is important to correlate the di- and trinucletide frequencies with the environment of the organism. Since it has been postulated that thermophiles will have higher percentages of G-C and C-G frequencies, the di- and trinucleotide frequencies in *Bacillus stearothermophilus* were determined using a Perl script written by Dr. Fares Najar.

10. TMHMM2.0

The TMHMM2.0 program is used to predict transmembrane helices in proteins. The protein coding sequences for the genomes, in fasta format, are retrieved from the NCBI FTP server. The TMHMM program specifies the location and orientation of the ORFs, which contain more than one transmembrane (84). The algorithm used for this identification is called N-best. Once identified, the codon usage and predicted amino acid composition were determined using a Perl script written by Dr. Fares Najar.

Chapter II: Materials and Methods

A. Material Preparation

1. Small Scale DNA Isolation of Gram-positive Bacteria

Agar plates consisting of 5 gram Trypticase Soy Broth, 1 gram Yeast Extract, 10 grams Trypticase Soy Agar, with sterile double distilled water (sddwater) to 500 ml were streaked by Fares Najar with the *Bacillus stearothermophilus* glycerol stock culture. The plates were incubated at 65 degree Celsius overnight. Media consisting of 10 grams Trypticase Soy Broth, 2 grams Yeast Extract, and sterile double distilled water to 1 liter was inoculated with *Bacillus stearothermophilus* cells and incubated at 55 degree Celsius. The cells were harvested into 96 deep well blocks and stored in the -20 degree Celsius freezer (85).

Cell pellets were washed by resuspending in 1 ml 50/20 TE by vortexing and then centrifugation for 4 minutes in the cold room. After discarding the supernatant the pellets were resuspended in 1 ml of ice-cold acetone and vortexed. The suspension was incubated on ice for 5 minutes and centrifuged for 4 minutes in the cold room. The acetone supernatant was removed by aspiration in hood. Cells were resuspended in 500 ul 50/20 TE containing lysozyme. To partially degrade the cell wall, the mixtures were incubated at 37 degree Celsius for 1 or 2 hours until cell suspension became slightly viscous at which time 75 ul of 10 percent SDS and 292 ul of 3 M NaCl was added to the cell suspension (86).

After mixing by inversion, the cell suspension was placed in -70 degree Celsius freezer for 3 minutes. The samples were mixed again and transferred to a 65 degree Celsius water bath for 3 minutes. This hot/cold treatment was repeated three times, after which the samples were incubated on ice for 10 minutes and then centrifuged at room temperature for 6 minutes to pellet cell debris. Then, 10 ul (200 ug/ml) DNase-free

RNase A was added to the supernatant and this mixture was incubated at 37 degree Celsius for 15 minutes. Then 20 ul (50 ug/ml) Proteinase K was added, and the mixture was incubated at 37 degree Celsius for an additional 30 minutes followed by a phenol/chloroform extraction and ethanol precipitation. After the supernatant was aspirated, the DNA was dried and dissolved in 100 ul sterile double distilled water (86)

2. Gentra PureGene DNA Isolation Kit

Genomic DNA was isolated in 1.5 ul centrifuge tubes using the Gentra PureGene DNA Isolation Kit. Four sets of two cell pellets were resuspended in 300 ul Cell Suspension Solution (Tris, EDTA, and sorbitol) followed by the addition of 1.5 ul Lytic Enzyme Solution. This suspension was mixed and incubated at 37 degree Celsius for 30 minutes to digest the cell wall. After 1 minute of centrifugation, 300 ul Cell Lysis Solution (Tris, EDTA, and sodium dodecyl sulfate) was added to resuspend the pellet (87)

The suspension was heated in the 80 degree Celsius water bath for 5 minutes to complete cell lysis. Next, 1.5 ul RNase A was added to the cell lysate, mixed, incubated at 37 degree Celsius for 15 to 60 minutes. After the sample was cooled to room temperature, 100 ul Protein Precipitation Solution (ammonium acetate) was added. The sample was vortexed at high speed and placed on ice for 15 to 60 minutes. The sample was centrifuged for 3 minutes to pellet the genomic DNA. This DNA then was ethanol precipitated and dissolved in 100ul sterile double distilled water. To check the DNA concentration, 10ul of DNA was loaded onto a 0.7 percent agarose gel and electrophoresed in a Life Technologies Horizon 20-25 gel box. The DNA concentration was checked spectrophotometrically on the Molecular Dynamics Spectramax Plus Microplate Spectrophotometer with the wavelength set at 260 nanometers (87)

The DNA was diluted prior to the absorbance measurement. Here, 100 ul of DNA was added to 900 ul sterile double distilled water for a total volume of 1000 ul. The absorbance reading was multiplied by 10 ul (1000 ul/100 ul = 10 ul). Since 1 absorbance

unit at 260 unit of duplex DNA equaled 50 ug, the A260 reading was multiplied by 50 ug/ml. Typical yields of *Bacillus stearothermophilus* isolated were 77 mg/l of log phase cell growth.

3. Large Scale DNA Isolation of Four Cosmids

My preliminary work involved the sequencing of a 41 Kb cosmid in the Neurofibromatosis 1 region of Human Chromosome 22 and a three cosmid set (~90 Kb) in a Meningioma Deletion Region of Human Chromosome 22. The large scale DNA isolation started with a plate streaked of each cosmid. One liter of Luria Broth (LB) and 4 ml of kanamycin were combined and 3 ml of this antibiotic containing media was transferred to a large falcon tube. A colony of cosmid containing bacteria streaked on a petri dish was added to the 3 ml antibiotic containing media and grown in the 250 rpm shaking/incubator for 8 hours at 37 degree Celsius. The 3 ml culture then was transferred to a flask with 50 ml antibiotic containing media and grown for 8 hours in the 250 rpm incubator

The 50 ml culture was transferred to a one liter flask of media/antibiotic and incubated for 8 hours. The one liter culture was divided into two 500 ml centrifuge bottles and centrifuged in a DuPont Instruments Sorvall RC-5B Refrigerated Superspeed Centrifuge for 20 minutes at 7,000 rpm to give a cell pellet that was frozen at -80 degrees Celsius (2).

The large scale DNA manual cosmid isolation was performed using the double acetate precipitation procedure. Here, the cell pellet was resuspended in 35 ml GET buffer containing glucose to prevent the cells from premature lysis and EDTA to bind divalent metal cations. Cells lysis started by adding 70 mg lysozyme and incubating at room temperature for 10 minutes. Then 70 ml alkaline lysis solution was added by gentle swirling in the centrifuge bottles and incubating in an ice water bath for 5 minutes. To precipitate the protein and other unwanted cellular components, 52.5 ml 3 M NaOAc was added. This lysate was swirled and incubated in an ice water bath for 15 minutes

before being poured through a double layer of cheesecloth to collect the large debris, and the flow through was centrifuged in the Sorval RC-5B Refrigerated Superspeed Centrifuge in 250 ml bottles at 10,000 rpm for 30 minutes to clear the lysate (2)

The supernatant, containing the DNA, was transferred to a clean 500 ml centrifuge bottle and after adding an equal volume of isopropanol, mixing by inverting several times and incubating at room temperature for 5 minutes, was centrifuged at 9,000 rpm for 20 minutes. After discarding the supernatant, a visibly clear film along the side of the bottle contained the DNA, which was dissolved in 18 ml 10:1 TE to dissolve the DNA pellet, and divided into two orange screw-cap centrifuge tubes. Then 4.5 ml of 7.5 M KOAc was added to each tube and placed in the -70 degree Celsius freezer for 30 minutes. The tubes were centrifuged at 2000 rpm for 10 minutes (2)

The supernatant was transferred to clean tubes and 200 ul of DNase-free RNase A was added. The tubes were incubated in the 37 degree Celsius water bath for 1 hour and the DNA was ethanol precipitated. The pelleted DNA was dissolved in 750 ul sterile double distilled water prior to nebulization. To check the DNA concentration, 10ul of DNA isolation was loaded onto a 0.7 percent agarose gel and electrophoresed in a Life Technologies Horizon 20-25 gel box. The final DNA concentration also was checked spectrophotometrically on the Spectramax Plus Microplate Spectrophotometer from Molecular Dynamics at a wavelength of 260 nanometers (2) Typical yields of isolated cosmid DNA were 1,450 mg/l of log phase cell growth.

B. Random Shotgun DNA Subclone Library Construction

1. DNA Shearing

Nebulization was performed on the isolated DNA to fragment DNA by physical shearing. The nebulizer was a product of IPI Medical Products. Initially, an ethanol and sodium chloride ice bath was prepared to a temperature of -20 degree Celsius. The following were combined inside the nebulizer: 500 ul glycerol, 50 ug DNA and sterile double distilled water. This mixture was kept in the ice water bath until the nebulizer

containing the mixture was placed into ethanol and sodium chloride ice bath. The *Bacillus stearothermophilus* DNA was sheared at 12 psi of nitrogen and the cosmids were sheared at 8 psi of for 2.5 minutes. Under these conditions, the resulting DNA fragments were in the 1.5-3.0 kb range as measured by comparison with size markers on a low melting point agarose gel. The mixture was centrifuged briefly in the Beckman Allegra 6R Centrifuge to collect any condensate and divided into four micro-centrifuge tubes for ethanol precipitation (2)

2. End-repair and Phosphorylation

Since the nebulized fragments usually contained single-stranded ends, they were made blunt-ended and phosphorylated prior to ligation into the SmaI/pUC blunt-ended vector using a fill-in/kinase reaction. DNA polymerase I (Klenow) has 5'-3' polymerase and 3'-5' exonuclease activities but lacks the 5'-3' exonuclease domain, while T4 DNA polymerase has the 5'-3' polymerase activity. Thus, adding dNTPs to reaction containing two enzymes results in blunt ended double-stranded DNA. The cloned T4 Polynucleotide Kinase then phosphorylates the 5' end of fragmented DNA. The following reagents were added to dry DNA. 27 ul sddwater, 3 ul 1 M 10xTM buffer, 5 ul Kinase buffer, 5 ul 10 mM rATP, 7 ul 25 mM dNTP, 1 ul T4 Polynucleotide Kinase, T4 DNA polymerase, and 2 ul Klenow for a total volume of 50 ul. This mixture was incubated in 37 degree Celsius water bath for 1 hour (2)

3. DNA Size Selection

The reaction was terminated after 5 ul of blue loading dye (15 percent ficoll, 0.2 percent bromophenol blue, and 0.2 percent xylene cyanol FF in sterile double distilled water) was added. The end-repair and phosphorylation reaction product were electrophoresed in a low melting point 1.0 percent agarose gel for subclone size selection. Ethidium bromide was added to this 1.0 percent agarose gel to make the fragmented DNA visible under ultraviolet light. The ethidium bromide intercalated between the bases and fluoresces under the UV light. Large bands between 1.5 to 3.0 kb

and small bands between 800 to 1.5 kb were excised and eluted by phenol/chloroform extraction. The DNA was ethanol precipitated, dried overnight, and resuspended in 10 ul sterile double distilled water (2).

4. DNA Ligation

After end-repair and phosphorylation, the DNA fragments were ready for ligation into the pUC cloning vector that was linearized at the Sma I site and dephosphorylated. T4 DNA Ligase catalyzed formation of phosphodiester bonds between 5' phosphoryl and 3' hydroxyl groups in duplex DNA. The optimum ligation reaction was 1 ul of a 1:2 dilution of *Bacillus stearothermophilus* DNA added to 5 ul sterile double distilled water, 1 ul Ligase buffer, 1 ul pUC/Sma I vector, 1 ul Ligase (2).

The following controls were performed to check the vector, competent cells, and the enzymatic reactions. The vector only control (negative control) required 1 ul pUC/Sma I vector having no fragmented DNA added to the reaction mixture. The competent cell only control required 1 ul isolated pUC to the reaction, while the ligation only control (positive control) required 1 ul of both pUC/Sma I vector and Sma I cut cosmid DNA added to the reaction pot. Results from the controls were as follows: the vector only control yielded no colonies; the competent cell only control yielded a plate of blue colonies since no DNA was inserted; the ligation only control yielded a plate of white colonies. Each ligation reaction had a total volume of 9 ul. The ligation mixtures were allowed to incubate in the cold room from overnight to a few days (2).

5. DNA Transformation

Electro-competent cells for electroporation were produced by combining 3 ul of YENB, 6 ul tetracycline, and a stab of the colonies from a streaked plate containing the XL1BlueMRF' *E. coli* cells in a 50ml Falcon tube. After incubating in the 37 degree Celsius, 250 rpm shaker for 12 to 16 hours, the contents were poured into the 1 liter of sterile YENB, and allowed to grow for an additional 4 to 5 hours until the absorbance at 600 nM was between 0.5 and 0.6 (2).

Using a sterile graduated cylinder, 250 ml of cultured cells were dispensed into the four sterile 500 ml centrifuge bottles. The cells were set on ice in the cold room for 5 minutes and centrifuged at 5,000 rpm for 10 minutes using the DuPont Instruments Sorvall RC-5B Refrigerated Superspeed Centrifuge (2).

After discarding the supernatant and removing any residues from the mouth of the bottles, 100 ml of cold sterile double distilled water was added to each centrifuge bottle. The cells were resuspended and centrifuged at 5,000 rpm for 10 minutes and the supernatant was discarded. One hundred milliliters of cold sterile double distilled water was added to each bottle to resuspend the cells. The bottles were pooled for a total of two bottles each having 200 ml of suspension. The cells were centrifuged at 5,000 rpm for 10 minutes and the supernatant discarded. One bottle of cells was resuspended in 20 ml cold sterile 10 percent glycerol. This suspension was transferred to the second bottle and the cells were centrifuged at 5,000 rpm for 10 minutes. The supernatant was discarded and the cells were resuspended in a final volume of 2 or 3 ml of cold sterile 10 percent glycerol. Forty microliters aliquots of these competent cells were prepared and frozen on dry ice during the distribution of the cells prior to the storage of the cells at -80 degree Celsius until use (2).

For transformation into *E. coli* cells using the Bio-Rad *E. coli* Pulser Electroporator, 2 to 3 ul of the ligation mixture was added to 40 ul aliquot of electro-competent cells and placed on ice for 1 minute. The 40 ul of the mixture/cells was pipetted into the 2 mm gap of the electroporation cuvette. Afterwards, the cuvette in the holder was slid into the shocking chamber. The "pulse" option was pressed while the electroporation unit delivered a 5 msec time constant and 2.5 kV to the *E. coli* cells. Then, 1000 ul YENB was added to the cuvette and mixed. The contents of the cuvette were poured into a small falcon tube. The falcon tube was put into the 37 degree Celsius incubator shaking at 350 rpm for 20 minutes. The LB/Amp plates were pre-warmed. The cells were centrifuged at 3,000 rpm for 5 minutes and the supernatant decanted.

Two hundred microliters of YENB, 25ul IPTG, and 25ul X-gal were added to each falcon tube. The cells were streaked on LB/Amp plates and grown for 18 hours in the 37 degree Celsius incubator (88). Since the multiple cloning site is engineered in the lac Z' gene of the pUC cloning vector, the presence of an inserted DNA could be detected in the presence of x-gal and IPTG via blue/white color selection.

In this blue/white colony selection, the white colonies of chimeric plasmids were picked using toothpicks, from the petri dish into 96 deep-well blocks containing 1.5 ml of terrific broth (TB) media inoculated with ampicillin and terrific broth salt. Four milliliters of ampicillin and 100 ml of TB salts were added to one liter of TB media. The deep-well blocks were put in the 37 degree Celsius incubator shaking at 350 rpm between 20 and 22 hours. Cells were harvested by centrifuging for 7 minutes at 2500 rpm and the supernatant discarded. The cell pellets were frozen at -20 degree Celsius (2).

6. Semi-Automatic Isolation of Plasmid Subclones using the Biomek 2000 Automation Workstation and the Robbins Scientific Hydra

In the initial phases of this research, the Beckman Biomek 2000 laboratory automation workstation was used because four sets of 96 double-stranded DNA sequencing templates could be isolated in less than 4 hours. The automation workstation had the following positions: position A1-tool rack with MP200 tool in the right hand rack, position A2-P250 Biomek 2000 tips, position A3-P250 Biomek 2000 tips, left side of position A4-100 ml TE-RNase A + ribonuclease T1, right side of position A4-100 ml SDS/NaOH, position A5-96 well blocks containing cell pellets, position A6-96 well blocks containing cell pellets, position B1-empty, position B2-P250 Biomek 2000 tips, position B3-P250 Biomek 2000 tips, left side of position B4-adds 100 ml 3M NaOAc, right side of position B4-empty, position B5-96 well blocks containing cell pellets, position B6-96 well blocks containing cell pellets (89). To prepare 120 ml TE-RNase A+ Ribonuclease T1, the following chemicals were combined: 6 ml 1 M Tris-HCl, 2.4 ml .5 M EDTA, 25 0ul RNase A, 50 ul Ribonuclease T1, and 112 ml sterile double-

distilled water. To prepare 100 ml SDS, the following chemicals were combined: 10 ml 10 percent SDS, 0.8 gms NaOH pellets, and sterile double-distilled water to a final volume of 100 ml. To prepare 3M NaOAc, the following chemicals were combined: 408 grams NaOAc was dissolved in 1 liter of glacial acetic acid/sddwater. One hundred milliliters of 3 M NaOAc was needed to process 4 blocks of subclones on the Biomek 2000 (2).

The program dsisol.v1 added 200 ul TE-RNase A + t1 to each well of the 4 blocks and mixed 20 times for resuspension of the cell pellet to begin the DNA template isolation. Here, the RNase A destroyed the RNA in the cell pellet. This step took 1 hour and 6 minutes. All 4 boxes of P250 Biomek 2000 tips at positions A2, A3, B2, and B3 were changed. The next program continued automatically. At this point 200 ul SDS/NaOH was added to each well of the 4 blocks and mixed 10 times. SDS dissolved the phospholipids in components of the cell membrane and lysed the membrane to release the cellular content. This step took 35 minutes. The first row of tips at position A2 was refilled. The program was continued after selecting "replace all." Finally, 200 ul 3 M NaOAc was added to each well of the 4 blocks. The acetate precipitated the protein and other unwanted cellular components. This step took 10 minutes. The 4 blocks were removed from the workstation. The blocks were covered with a plate sealer, vortexed briefly, and the plate sealer was replaced with a lid. The blocks were put on a Labline Instruments titer plate shaker for 10 minutes, and put in 37 degree Celsius incubator shaking at 350 rpm for 10 minute (2, 89).

The blocks were stored in -20 degree Celsius freezer overnight. They were centrifuged for 45 minutes at 4,250 rpm on Jouan's KR 422'Refrigerated Centrifuge. Using the Robbins Scientific Hydra 96 Automated Plate Positioning System, the upper 400 ul of supernatant was transferred from each 96 well block to a clean 96 well block. Ethanol precipitation and drying preceded hydration of blocks with 100 ul sterile double distilled water. The subclone isolations were electrophoresed on a 0.7 percent agarose

gel to check purity and concentration of DNA (2).

7. Semi-Automatic Isolation of Plasmid Subclones using Flexys, HIGRO, and Hydra

At the later stages of this work, some subclones were picked using the Genomic Solution PBA Flexys Colony and Plaque Picker, a robotic instrument for automatic picking of biological samples from agar plates to microtiter plates. The flat-bottom microtiter plates contained 150 ul of TB media and TB salts inoculated with ampicillin. The benefits of using this instrument were that it only took eight minutes to inoculate one microtiter plate and that it was more accurate than manual picking (90).

After picking, the subclones were grown for 18 hours at 37 degree Celsius with shaking at 520 rpm in the GeneMachine HIGRO. The HIGRO was a high capacity, shaking incubator that was proven to enhance 96-well format growth for sequencing protocols. It featured oxygenation, high shaking speed, and a small shaking orbital of 8.0 mm, which helped optimize growth in shallow-well plates. This orbital was slightly smaller than the diameter of the wells in a 96-well plate, thus, enabling the use of very high speeds without splashing. This 8.0 mm orbital allowed the formation of a vortex in each well during shaking. Vortexing was important as it created vertical mixing in the well and exposed more cells to the air above the well. This complete aeration prevented cell death and allowed for optimal utilization of the available nutrients in the media. The shaking speed directly affected the rate of growth. The higher the shaking speed, the faster the cultures grew (91). Cells were centrifuged on the Beckman Allegra 6R Refrigerated Centrifuge for 10 minutes at 2500 rpm and frozen for at least 2 or 3 hours at -80 degree Celsius prior to isolation (92).

The subclone isolation consisted of adding 60 ul each of TE-RNase A + t1, SDS, and 3 M NaOAc. After the first two additions of the above reagents, the lysate was shaken on the titer shaker for 20 minutes. After the acetate addition, the lysate was placed in the HIGRO for shaking at 400 rpm for 30 minutes. The lysate was frozen

overnight at -80 degree Celsius. After thawing, the microtiter plates were centrifuged on Beckman's Allegre 6R Centrifuge at 3000 rpm for 45 minutes at 4 degree Celsius and 60 ul of the supernatant was transferred to a clean v-bottom microtiter plate followed by ethanol precipitation. Approximately 50 ul sterile, double distilled water was added to the dried DNA (92).

8. Characterization of P1 clones containing *Bacillus stearothermophilus* Genomic DNA

Bacteriophage P1 recombinant plasmids were isolated to provide large insert clones for the *Bacillus stearothermophilus* sequencing project. The P1 clones were provided by Dr. Jonathon Coren of Southwestern Oklahoma State University. The individual clones were amplified by inducing with 1 mM IPTG at an absorbance 600 of approximately 0.01 (64). Dr. Coren plated the P1 clones on LB agar inoculated with kanamycin. The cloned inserts averaged 6 kb. Approximately 30 percent had inserts between 5 to 10 kb and 12.5 percent had inserts greater than 10 kb (1 was 12 kb and 1 was 22 kb). The remainder was less than 5 kb. There were approximately 4,000 colonies, thus there were 500 clones with 10-20 kb inserts (93).

One liter of LB media inoculated with 5 ml kanamycin was used for P1 subclone growth. The P1 colonies were picked into 96 well, deep well blocks containing 1.5 ml of LB/Kan. These blocks grew in the 37 degree Celsius shaker/incubator for 22 hours. The cells were harvested and stored at -20 degree Celsius. The P1 isolation consisted of adding the following reagents to the deep well blocks: 200 ul each of TE-RNase A + T1, SDS, and 3 M NaOAc. After the first two additions of the above reagents, the lysate was placed on the titer shaker for 20 minutes. After the acetate addition, the lysate was placed on the titer shaker for 10 minutes and placed in the 37 degree Celsius shaker/incubator for 10 minutes. The lysate was frozen overnight at -20 degree Celsius.

After thawing, the deep well blocks were centrifuged in the Jouan KR 422 Refrigerated Centrifuge at 4250 rpm for 45 minutes at 4 degree Celsius. Next, 400 ul of the supernatant was transferred to clean deep well blocks.

Isopropanol was added to these blocks which were placed in the -20 degree Celsius freezer overnight. The blocks were centrifuged on the Beckman Allegra 6R Centrifuge for 30 minutes at 4 degree Celsius at 3000 rpm. The supernatant was decanted and 500 ul 70 percent ethanol was added to the DNA. The blocks were centrifuged for 15 minutes in the Beckman and the DNA was allowed to dry overnight. To the dried DNA, 30 ul sterile, double distilled water was added. The blocks were placed on the titer shaker for 2 hours for the DNA to totally resuspend in the water. To check the DNA concentration, the DNA was loaded onto a 0.7 percent agarose gel and electrophoresed in a Life Technologies Horizon 20-25 gel box (92). The resulting PAC library consisted of 16x96=1536 recombinant clones.

9. Dideoxynucleotide DNA Sequencing

The sequence of the universal forward primer used in pUC chimeric plasmids was 5'-TGT-AAA'ACG-ACG-GCC-AGT-3', and the sequence of the universal reverse primer was 5'-CAG-GAA'ACA-GCT-ATG-ACC-3'. The sequence of the forward primer used in the P1 clone sequencing reactions was T7 (5'-TAA- TA'-GAC-TCA-CTA-TAG-GG-3') and the reverse primer was SP6 (5'-ATT- TA'-GTG-ACA-CTA-TAG-AA-3'). Sequencing was performed in 384 well plates on the Perkins Elmer Applied Biosystems Gene Amp PCR System 9600 or 9700 (2).

Sixty rounds of cycle sequencing allowed the amplification of the labeled DNA fragments. Each cycle resulted in a linear amplification of the template DNA. Denaturation of double-stranded DNA occurred at 95 degree Celsius and this

temperature was maintained for 10 seconds. Annealing of primer to a predetermined segment of a strand of the pUC cloning vector near the insertion site of the cloned DNA occurred at 50 degree Celsius and incubation at this temperature was for 5 seconds. The primer provided a 3'-hydroxyl group for the initiation of synthesis of a complementary DNA strand. Since strand extension occurred at 60 degree Celsius, the incubation at this temperature was for 4 minutes. The cycle sequencing products then were resolved on either the Perkin Elmer ABI Prism 377 DNA Sequencer or the Perkin Elmer ABI Prism 3700 DNA Analyzer (94, 53). For reactions resolved on the Amersham Pharmacia Biotech MegaBACE 1000 DNA Sequencing System, the cycling times were 20 seconds at 95 degree Celsius, 15 seconds at 50 degree Celsius, and 1 minute at 60 degree Celsius using the Pharmacia sequencing enzyme kit rather than the ABI sequencing kit (54)

DNA chain elongation ceased when fluorescent 3' dye-labeled terminator ddNTP's were incorporated into DNA extension products because ddNTP's lacked a 3'OH group which was necessary to form the phosphodiester linkage that joined adjacent nucleotides (50).

When electrophoresing samples through gels on the ABI Prism 377, 2 ul of DNA template was used. A reaction mixture consisting of 100ul 5x dilution buffer, 400 ul primer and 300 ul Big Dye was made and 3 ul was dispensed into each well of a 384-well thermocycle plate using a Robbins Scientific Hydra 96 Automated Plate Positioning System. Thus, the total volume thermocycled per reaction for electrophoresis on the ABI Prism 377 was 5 ul. Only 1 ul of pUC chimeric DNA was used for cycling reactions electrophoresed on the ABI Prism 3700 and only 2 ul of reaction mixture was added to each well. Thus, the total volume thermocycled per reaction for electrophoresis on the 3700 was 3 ul. The thermocycle reactions for the P1 chimeric clones electrophoresed on

the ABI Prism 3700 contained 5 ul DNA, 2 ul primer, 1 ul 25 percent DMSO, and 2 ul 1:12 Big Dye dilution for a 10 ul total volume (92)

The MegaBACE required the following reagents for successful cycling reactions: 2 ul of pUC chimeric DNA, 1 ul primer, 4 ul DYEnamic energy transfer (ET) terminator reagent premix, and 3 ul sterile double distilled water for a total volume of 10 ul per reaction. Both forward and reverse primers were used on each subclone template (55)

Unincorporated dye-labeled terminators were removed prior to electrophoresis on the ABI Prism 377 by passage through Sephadex G50 Columns prepared in 96-well filtration plates (95). These plates had a 0.05 um pore size filter made of mixed cellulose esters Unincorporated terminators were removed from the thermocycle reactions prior to electrophoresis on the ABI Prism 3700 and MegaBACE by ethanol precipitation (2)

10. Primer Walking

After sequencing the P1 clones using Big Dye with the T7 forward primer and the SP6 reverse primer, and then primer walking off the P1 clones, I continued to primer walk the subclones with customized local and extendc primers. A total of 415 gap spanning clones were identified from approximately 60,000 subclones and arrayed into 96 well format. These spanning clones then were resequenced with universal forward and reverse primers using Big Dye and Big Dye dGTP. Of the 415 gap spanning clones, 135 were identified for sequencing with dRhodamine. Additional rounds of primer walking with customized primers then were used on the gap spanning clones, P1 clones, and subclones to close sequence gaps.

11. Polymerase Chain Reaction

The uniplex PCR reaction included the following reagents: 3 ul subclone DNA, 3 ul percent DMSO, 2 ul each of two primers, 5 ul 10X PCR buffer, 5 ul 2 mM dNTP, 30 ul

sddwater, 1 ul AmpliTaq DNA polymerase. The multiplex PCR reaction included the following reagents: 3 ul subclone DNA, 3 ul 100 percent DMSO, 2 ul each of eight primers, 10 ul 10X PCR buffer, 10 ul 2 mM dNTP and/or 5 ul 7-Deaza-2'-deoxy-GTP, 20 ul sddwater, 1 ul AmpliTaq DNA polymerase. 7-Deaza-2'-deoxy-GTP is a substrate for most DNA polymerases, including *Taq* DNA polymerase. Incorporation of 7-Deaza-2'-deoxy-GTP into DNA altered the fluorescent staining and electrophoretic mobility of the DNA. I also had some success with the XL PCR kit provided by PE Applied Biosystems. This kit included the rTth DNA polymerase, XL (Extra Long) enzyme designed to amplify target DNA sequences ranging from 5 to more than 40 kb in size. The successful XL PCR reactions included the following reagents: 3 ul subclone DNA, 3 ul 100 percent DMSO, 2 ul each of eight primers, 15 ul 3 3 XL Buffer II, 2 ul GeneAmp10 mM dNTP Blend, 3 ul 25 mM MgCl, 20 ul sddwater, and 1 ul rTth DNA polymerase, XL. After the PCR, I removed excess primers and excess dNTP using 5 ul Shrimp Alkaline Phosphatase and 5 ul Exonuclease I in each well of the PCR reaction. Finally, thermocycle reactions included one primer per well (96).

C. DNA Anaylsis

1. Phred/Phrap/Crossmatch

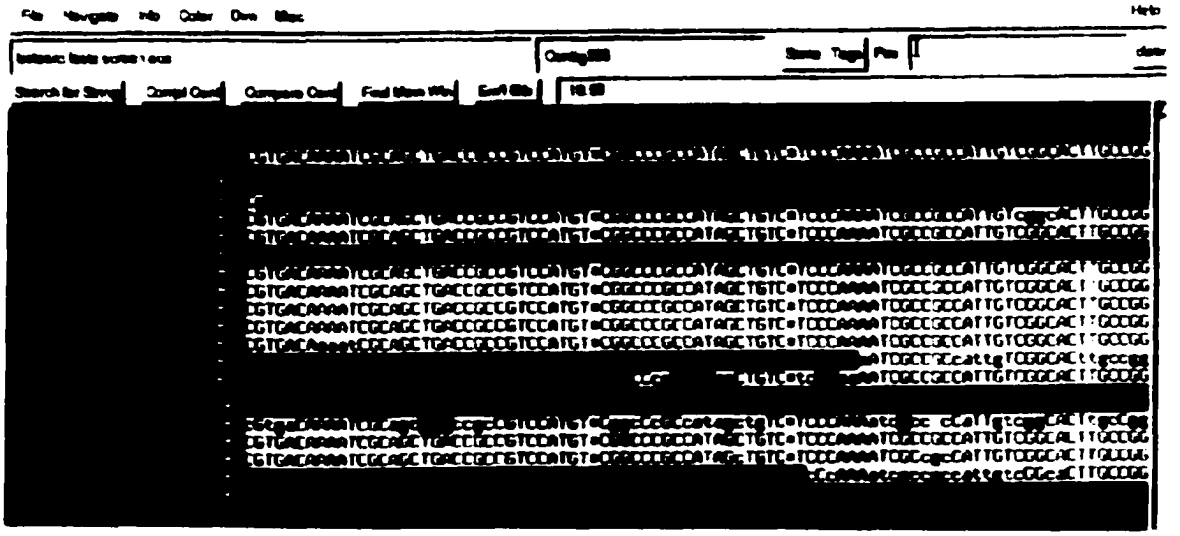
The DNA sequence data collected on the ABI Prism 377, 3700, and MegaBACE DNA sequencers was transferred to a UNIX-based Sun Microsystems Sparc Station 20 for analysis and assembly. After entering the sequence reads into the directory, the perlscript, "automake directory" controlled the Phred/Phrap/Crossmatch Programs (56, 57, 58), which resulted in basecalling of the recently entered DNA sequence reads, and assembling them into the database, and screening out *E. coli* vector contaminant.

2. Consed

Once the sequence data was assembled into a Phrap database, it was viewed using the

Consed program. The Consed aligned reads window had several lines. The top line displays the contig sequence, and below it the read sequences are displayed. The background gray scale indicates base quality with the highest quality being white and the lowest quality black. *E. coli* vector bases are identified with the letter x (59, 60).

Figure 8-Consed for *Bacillus stearothermophilus*



D. DNA Sequence Analysis

1. BLAST

Typically, a BLAST search yields a multi-column, tabular result containing the High Score, the likelihood of error (P), the contig number, the first base, the last base, the protein and its organism (97).

2. PowerBLAST

The PowerBLAST results were displayed as multiple alignments with annotated features derived from the GenBank records of matching sequences. The output file was viewed using MUSK Viewing Program (62).

3. Genscan

Genscan produces a graphic output file that displayes the location and DNA strand of

each predicted exon (64).

4. REPEATMASKER

A typical result of a repeatmaker screen gives multiple columns that include the following: the Smith-Waterman (SW) score of the match, the percent substitutions in matching region compared to the consensus, the percent of bases opposite a gap in the query sequence, the percent of the bases opposite a gap in the repeat consensus (inserted bp), the name of the query sequence, the starting and ending positions of match in the query sequence, the number of bases in query sequence past the ending position of match, tells whether or not the match is with the Complement of the consensus sequence, the name and class of the matching repeat, the number of bases in the repeat consensus sequence, the starting and ending position of the match in database sequence (65)

5. EXGAP

Exgap, a program written by Dr Axin Hua, graphically and tabularly displays the position of paired end reads from a sequence project. A Perl script, written by Steve Kenton, pairs all the forward reactions in one list and the reverse reactions in separate lists to reduce the chance of error in setup of the uniplex PCR plates (66)

6. Finisher

Finisher is a script that pairs subclones for use in PCR reactions to close gaps by identifying the 4 kb window spanning the gap and records the forward and reverse reads that are within that window (67).

7. Dotter

The dotplot output displays one sequence along the horizontal axis and the other along the vertical axis of a coordinate system with a dot printed where the two sequences matched at a user defined stringency and window. Thus a repeated region would run

diagonally across the dot matrix (68, 69)

8. Artemis

Once the *Bacillus stearothermophilus* genome reached 7.1 fold coverage at 3.1 Mb, the ORFs were predicted by Artemis and BLASTP homology to entries in the GenBank database. Prior to using the Artemis program, the *Bacillus stearothermophilus* genome was divided into ten pieces with each piece containing 300 kb to make the annotation more manageable. The open reading frames were identified that were at least 100 codons in length and /or greater than 3 kb using the built in Artemis function. The annotation of the *Bacillus stearothermophilus* genome was manually edited when each ORF was verified by BLASTP homology to a known gene scanning the probability scores and any insignificant ORFs were noted (71, 72)

E. Genome Metabolic Analysis

1. KEGG

The ORFs obtained from Artemis were entered into a modified listing similar to Dr Riley's classification scheme and pie charts were made representing this classification (35, 77). Hongshing Lai developed a Perl script that uses this modified annotation list that contains the E.C. numbers (numbers used in enzyme nomenclature to name reactions catalyzed) and the number of copies of each gene within the genome.

Table 2- Dr. Riley's Modified List of Categories of Cellular Functions as used in Dr Roe's lab

Categories of Cellular Functions
Physiological Function
I. Small-Molecules
Degradation
Energy Metabolism
Central Intermediary Metabolism
Amino Acid Biosynthesis
Polyamine Biosynthesis
Purines, Pyrimidines, Nucleosides, Nucleotides
Biosynthesis of Cofactors, Prosthetic Groups, and Carriers
Fatty Acid Biosynthesis
Broad Regulatory Functions
II. Macromolecules
Synthesis and Modification of Macromolecules
Nucleic Acid/Protein Degradation
Cell Envelope Metabolism
III. Processes
Transport Binding Protein
Chaperones
Cell Division
Chemotaxis & Mobility
Protein & Peptide Secretion
Osmotic Adaptation
Detoxification
IV. Other
Metabolism of Cofactors & Vitamins
Metabolism of Other Substances
V. Unknown

The Perl script output was converted to a table using Excel and after deleting the letters "E.C.", the actual numbers from the *Bacillus stearothermophilus* annotation were kept and the *Bacillus subtilis* genome in the KEGG program was opened (78, 98). This *Bacillus stearothermophilus* annotation then was entered into the *Bacillus subtilis* KEGG metabolic pathway.

3. MIPS

MIPS, a DNA Viewer and Protein Viewer, displays ORF analysis with selectable options. The *Bacillus stearothermophilus* MIPS directory was last updated in March 2002 with the results sent to a specific website (99)

4. SubtiList

SubtiList is a database of *Bacillus subtilis* genomic features. It can be searched using the gene, protein name, sequence analysis, or chromosome region. The results of a query give either a list of genes generated from the user query, or a graphical representation of the chromosome region selected. In addition, it also provides specific details of the selected gene, such as, mapping information, functional classification, and the opportunity to retrieve the sequence in UNIX, windows, or MacIntosh (81)

5. BSORF

Bacillus subtilis ORFs (BSORF) is a database that allows for entering the gene name, accession number, description, category number, or KEGG pathway, for retrieval of the SWISSPROT gene sequence used in a TBLASTN search at the *Bacillus stearothermophilus* genome BLAST server

6. Codon usage

Jim White wrote a Perl script to calculate total codon usage and a Perl script to calculate codon usage for each ORF that had at least one transmembrane helix.

7. Di- and Trinucleotide Frequencies

Jim White and Dr. Fares Najar wrote a series of Perl scripts, which calculate either the di- or trinucleotide frequencies of a DNA sequence.

Chapter III: *Bacillus stearothermophilus* Sequencing Results and Discussion

A. General Information and Genomic Comparisons

1. Statistics

A whole-genome random sequencing method was used to obtain the working draft sequence of the approximately 3.4 Mb *Bacillus stearothermophilus* genome. Through this approach, approximately 2,731 ORFs were identified in this genome, and its G+C content was determined to be 52.75 percent. The average length of sequencing reads is 506 bp and a total of over 78,000 sequence reads were assembled by phred/phrap. Approximately 600 microtiter plates, each containing 96 wells of DNA sequencing templates from individual subclones were isolated. There are 320 contigs greater than 2 kb totaling 3,434,800 bp. This project is at 9.6 fold coverage.

No plasmids were detected either by biochemical methods or through DNA sequencing although at least one plasmid was previously reported in other strains of *Bacillus stearothermophilus*. The genome sequence data and clone information for *Bacillus stearothermophilus* have been publically released on our website at URL http://genome.ou.edu/bstearo.html (97).

Table 3- Summary of the Recent Phrap High Stringency Statistics for *Bacillus stearothermophilus*

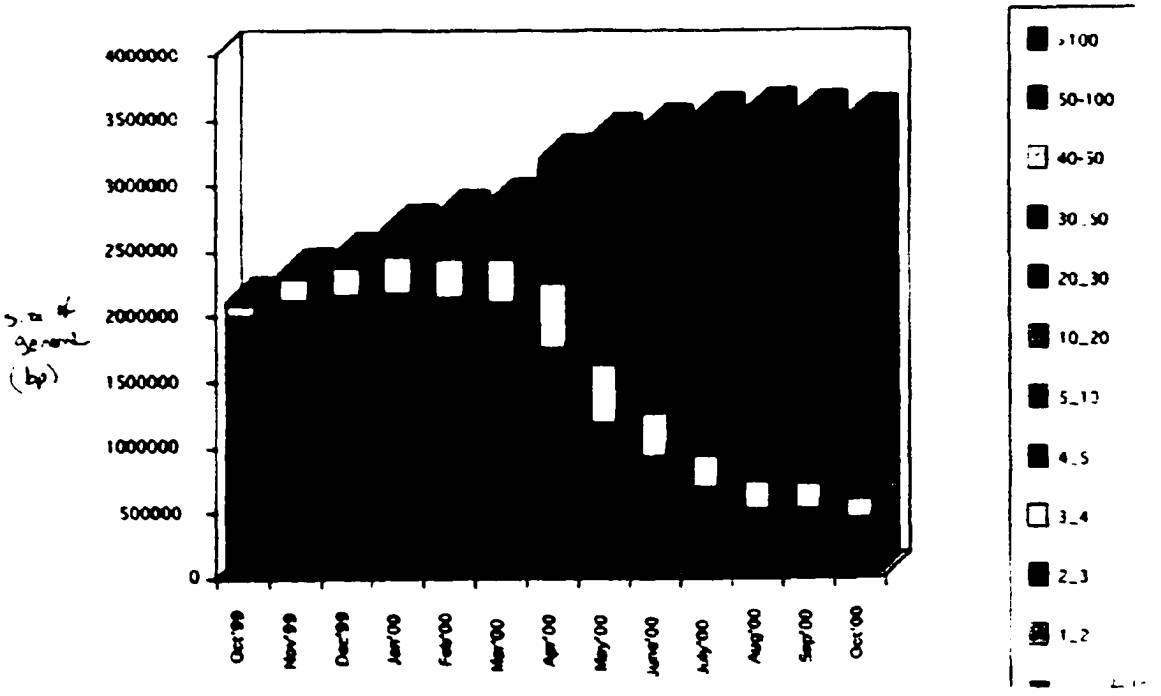
Summary for bstearo_high_stringency as of Mon Nov 19 18:21:01 CST 2001

Contig Size	Total Number	Total Length	% Of Cumulative
0 - 1 kb	59	39754	1.1%
1 - 2 kb	62	90439	2.5%
2 - 3 kb	47	117532	3.3%
3 - 4 kb	27	96380	2.7%
4 - 5 kb	22	97921	2.7%
5 - 10 kb	98	726424	20.4%
10 - 20 kb	81	1104969	31.0%
20 - 30 kb	26	657781	18.5%
30 - 40 kb	14	488332	13.7%
40 - 50 kb	2	88845	2.5%
50 - 100 kb	1	56056	1.6%
>100 kb	0	0	0.0%
Cumulative	439	3564433	Consed_Err/10KB = 132.1
Cumulative>1 kb	380	3524679	Consed_Err/10KB = 108.51
Cumulative>2 kb	318	3434240	Consed_Err/10KB = 92.7

Phrap Coverage: 9.4, Phrap Avg. Confirmed Length: 503.5, Confirmed Reads: 7055

2. Progress Report

Figure 9-Progress Report for *Bacillus stearothermophilus*



I began this project January, 1999. By March, 1999, the first sequencing reads went into the directory. The above chart shows that by the end of October, 1999, a cumulative 2.0 Mb of sequenced data was in the *Bacillus stearothermophilus* directory. A year later, by the end of October 2000, a cumulative of 3.4 Mb of sequenced data was in the directory. At this point, there were 404 contigs greater than 2 kb. By the end of October, 2001, a cumulative of 3.5 Mb of sequenced data was in the directory, with 316 contigs greater than 2 kb.

Presently, this genome is in the working draft. The percentage of the genome sequenced can be estimated following the Lander and Waterman application of the Poisson distribution. The Lander and Waterman application indicates that a sequencing project at 9 fold coverage should have 99.99 percent of the genome sequenced. Presently, the *Bacillus stearothermophilus* sequencing project is at 9.6 fold coverage. This Lander and Waterman application indicates that at 9 fold coverage at a genomic size of 4 Mb of which is 3.6 Mb totally sequenced, the total gap length should be approximately 400 bases. Presently, over 3.4 Mb of sequenced data is in the directory. This Lander and Waterman application indicates that at 9 fold coverage with a genome size of 4 Mb and with the average read length at 500 bases, the directory should contain 72,000 gel reads. Presently, this directory contains over 78,000 gel reads. This Lander and Waterman application indicates that this genome should have 7 gaps, with 57 bases per gap, instead of the 320 gaps greater than 2 kb which presently exist (100).

3. Genomic Comparison

Now that several bacterial genomes have been sequenced, a comparison can be made to elucidate the biochemical properties of the *Bacillus stearothermophilus* working draft genome. Although the exact size of the *Bacillus stearothermophilus* genome is not known, based on the genomic comparisons in Table 4, it is likely that *Bacillus stearothermophilus* may not be as large as *Bacillus subtilis*, and in the 2.7-3.4 Mb range.

The reason for this corrected projection is that the ratio between the genome size and the number of ORFs for *Bacillus subtilis* strain 168, *Bacillus halodurans* strain C-125, *Escherichia coli*, *Aquifex aeolicus* and *Thermotoga maritime* is approximately one ORF per thousand bases.

Table 4- Genomic Comparison

Organism	*B stearothermophilus*	*B subtilis*	*B halodurans*	*E coli*	*A aeolicus*	*T maritima*
Size (bp)	~3434240	4214810	4202353	46639221	1551335	1860725
ORFs	2731	4100	4066	4288	1512	1877
rRNA operons	8	10	8	7	2	1
tRNA operons	83	86	78	86	44	46
G+C ratio	52 75%	43 50%	43 70%	50 80%	43 40%	46%

Note: size of genome/#*ORFs*

B. subtilis: 4214814/4100=1028

B. halodurans: 4202353/4066=1034

E. coli: 4639221/4288=1081

A. aeolicus: 1551335/1512=1026

T. maritima: 1860725/1877=991

Bacillus stearothermophilus has approximately 2,731 ORFs on its chromosome with a G+C content of 52.75 percent as compared to *Bacillus subtilis* with 4,100 ORFs with a G+C content of 43.50 percent. Evidence presented in this research agrees with existing data that states that thermophiles have a higher G+C content as compared to mesophiles In his graduate research, Dr. Roe, working with the mesophile *Bacillus licheniformis* and the thermophile *Bacillus stearothermophilus*, studied the base composition of the DNA, using the nearest neighbor frequency analysis, concluded that the mesophilic G+C content was 45.2 percent, whereas the thermophilic G+C content was 53 4 percent at 65 degree Celsius. Dr. Roe determined that there was a preferential clustering of G and C in the thermophilic DNA and of A and T in the mesophilic DNA, allowing the DNA from

the thermophilic strain to be more heat stable than the DNA from the mesophilic strain (101).

At present, there are approximately 500 transposable elements identified in many different bacterial species. Although they are diverse in structural organization and primary sequence, sequence alignments are used to group the transposable elements into 17 families. Analysis of the *Bacillus stearothermophilus* genome reveals that it contains at least 24 transposons, 47 insertion sequences, and 106 transposase as compared to zero insertion sequences and 10 transposase and related genes reported in the published results of the *Bacillus subtilis* genome analysis and 112 transposase genes and related proteins found in *Bacillus halodurans*.

Presently, each completely sequenced genome has approximately 30-40 percent of its protein with unknown functions. Genes, which have known functional classifications from *Bacillus subtilis* and *Bacillus stearothermophilus* are listed in Appendices 2 and 3, respectively

B. Gene Conservation

1. Functional Classification

DNA replication involves DNA polymerase III subunits, DNA polymerase I, DNA gyrase subunits, DNA helicases, helicase loader, primosomal proteins, single strand DNA binding proteins, DNA ligase, and DNA primase. In the *Bacillus stearothermophilus* genome, when analyzed and viewed by Artemis reveals an operon in the replication origin at position 1,453,259 that had the following gene order: *gyrAB-recF*, and DNA polymerase III, B chain and *dnaA* at position 1,872,200. In *Bacillus subtilis*, the replication origin encompasses an operon with the gene order of *gidBA*-orf50k-orf208-orf261-*rnpA-rpmH-dnaAN*-orf71-*recF-gyrB*. The identical gene order including 2 orfs (orf50k-orf261) appears in *E. coli* around its replication origin, however, *gidA* and *gidB* genes are located away from the cluster (36). *dnaH* is homologous to *dnaA* in *E. coli*. *dnaG* is homologous to *dnaN* of *E. coli*. As shown in table 5, all the necessary proteins to carry out replication have been located in *Bacillus subtilis* (14), and *Bacillus stearothermophilus*.

Table 5-*Bacillus subtilis* Replication Genes

Bsu gene	*Bst* gene
dnaA	yes
dnaB	yes
dnaC	yes
dnaD	yes
dnaE	yes
dnaG	yes
dnaI	yes
dnaN	yes
dnaX	yes
gyrA	yes
gyrB	yes
holB	yes
lig	yes
pcrA	yes

polA	yes
polC	yes
priA	yes
ssb	yes

Transcription involves many accessory proteins, DNA-directed RNA polymerases *rpoA, E, D, B, C'*, and sigma factors. In the *Bacillus stearothermophilus* genome, Artemis reveals *rpoBC'* next to each other but as part of an operon that includes other ribosomal proteins. In *Bacillus subtilis* and *E. coli* the RNA polymerase B-B' subunit appears as *rpoBC'*. As shown in table 6, all of the genes that encode the proteins required for transcription (24) have been located in both *Bacillus subtilis* and *Bacillus stearothermophilus*.

Table 6-*Bacillus subtilis* RNA polymerase Core Subunits and Accessory Factors

Bsu gene	Function	*Bst* gene
rpoB	Core RNAP	yes
rpoC	Core RNAP	yes
rpoA	Core RNAP	yes
rpoE	Increases selectivity, stimulates recycling	yes
ykzG	Unknown	yes
yloH	Chaperone?	yes
greA	Elongation	yes
nusA	Elongation	yes
nusG	Elongation	yes
rho	Termination	yes
mfd	Transcription-repair coupling factor	yes

Bacillus subtilis has approximately 17 sigma factors, as shown in table 7, sixteen of which are members of the sigma70 family which recognize promoters near -35 and -10 with respect to the transcription start site (24).

Table 7-*Bacillus subtilis* Sigma Factors

Bsu gene	Function	*Bst* gene
sigA	RNA polymerase major factor (rpoD)	yes
sigB	general stress response	yes

sigD	flagella, chemotaxis, autolysins	yes
sigE	sporulation, early mother cell	yes
sigF	sporulation, late mother cell	yes
sigG	sporulation, late forespore	yes
sigH	sporulation, competence	yes
sigK	unknown	no
sigL	levanase, amino acid catabolism	yes
sigM	salt resistance	yes
sigV	extra cytoplasmic type	yes
sigW	antimicrobial resistance	yes
sigX	cell surface properties	yes
sigY	unknown	yes
sigZ	unknown	yes
ylaC	unknown	yes

Translation involves rRNA, ribosomal proteins, tRNA, aminoacyl-tRNA synthetases, translation factors, and modifying enzymes. Eight ribosomal RNA operons were identified in the *Bacillus stearothermophilus* genome. Of the fifty translation-related proteins found in *Bacillus stearothermophilus*, two proteins map to contig 89, two map to contig 119, seventeen map to contig 242, three map to contig 249, four map to contig 306, two map to contig 328, two map to contig 357, and three map to contig 359. The other fifteen proteins are located individually on different contigs. In *Bacillus subtilis* there are ten rRNA operons and *Bacillus halodurans* has eight rRNA operons, as shown in table 8, with tRNA-16S-23S-5S gene order. The tRNA genes are located between the 16S and 23S regions or downstream of the 5S region, such as, 16S-tRNA-23S-5S and 16S-23S-5S-tRNA. The rRNA genes are designated as rrn (24).

Table 8-*Bacillus Subtilis* Ribosomal RNA Operons (24, 83)

Bsu gene	*Bhal* gene
rrnA	yes
rrnB	yes
no	*rrnC*
rrnD	yes
rrnE	yes
no	*rrnF*
rrnG	yes
rrnH	yes
rrnI	no

rrnJ	no
rrnO	no
rrnW	no

The *Bacillus halodurans* rRNA operons were used to BLAST for rRNA genes encoded in the *Bacillus stearothermophilus* genome as shown in table 9 The contigs were recorded with the original intent of possibly tracking the origin of replication in this genome

Table 9-*Bacillus halodurans* rRNA (83)

Bhal rRNA	*Bst* high score	*Bst* probability	*Bst* contig
rrnA-16S	1233	9.50E-121	189
	1007	1.80E-100	293
rrnA-23S	1178	6.40E-146	221
	1632	7.80E-127	320
	874	1.30E-118	286
rrn-5S			160
			145
			162
			219
			270
			320
			391
			160

Table 10-*Bacillus subtilis* Ribosomal Proteins

Bsu gene	protein	*Bst* contig
rpsB	S2	388
rpsC	S3	242
rpsD	S4	229
rpsE	S5	242
rpsF	S6	359
rpsG	S7	119
rpsH	S8	242
rpsI	S9	249
rpsJ	S10	multiple
rpsK	S11	multiple
rpsL	S12	119
rpsM	S13	multiple
rpsN	S14	242
rpsO	S15	374

rpsP	S16	357
rpsQ	S17	242
rpsR	S18	359
rpsS	S19	242
rpsT	S20	120
rpsU	S21	208
rplA	L1	306
rplB	L2	242
rplC	L3	multiple
rplD	L4	multiple
rplE	L5	242
rplF	L6	242
rplI	L9	359
rplJ	L10	306
rplK	L11	306
rplL	L12	306
rplM	L13	249
rplN	L14	242
rplO	L15	242
rplP	L16	242
rplQ	L17	249
rplR	L18	242
rplS	L19	357
rplT	L20	328
rplU	L21	89
rplV	L22	242
rplW	L23	242
rplX	L24	242
rpmA	L27	89
rpmB	L28	error
rpmC	L29	error
rpmD	L30	242
rpmE	L31	267
rpmF	L32	error
rpmH	L34	error
rpmI	L35	328
rpmJ	L36	multiple

Of the eight ribosomal operons in *Bacillus stearothermophilus*, Artemis reveals three large ribosomal operons at different locations. The ribosomal proteins found in *Bacillus stearothermophilus* are listed in table 10. An operon appears at position 2,002,821 with the gene order of rpsH (RNA polymerase sigma H factor)-*rpl1-rpl-rpl0-rl7* (all four are 50 S ribosomal proteins)-*rpoBC*. An operon appears at position

939,400 with the following gene order *secY-rl15-rl5-rl18-rl6-rl5-rl24-rl14-rl16-rl3* (all are 50 S ribosomal proteins except rl5 and rl3, which are 30 S ribosomal proteins). Another operon appears at position 1,133,032 with the following gene order rpoa-rl17-rl13-rs9 (rl17 and rl13 are 50 S ribosomal proteins, and rs9 is a 30 S ribosomal protein). The gene order for ribosomal operons proceeds as follows in both *Bacillus subtilis and E. coli rplKAJL-rpoBC-rpsLG-fusA-tufA-rpsJ-rplCDWB-rpsS-rplV-rpsC-rplP-rpmC-rpsQ* ... *rplNXE-rpsNH-rplFR-rpsE-rpmD-rplO-secY-rpmJ* (35). In *Bacillus subtilis*, there is an insertion of *adk*, *map*, and *infA* to the *spc* operon that starts at *rplNXE* and ends with *rpmJ* above. In *E. coli* the gene order of the alpha operon is as follows *rpsMKD-rpoA-rplQ*. In *Bacillus subtilis*, the *rpsD* is translocated. Both organisms have the same gene order for the L13 operon, which is *rplM-rpsI*.

The tRNAscan 1.3 program was used to analyze the *Bacillus stearothermophilus* genome. It revealed eighty-three tRNA genes and 1 pseudogene. All amino acid tRNA synthetases were represented. In contrast, eighty-six tRNA genes were observed in *Bacillus subtilis*. These genes were clustered into eight single genes plus twelve operons, seven located between, adjacent to, or even withing rRNA operons (21). As mentioned in the Introduction, the *Bacillus subtilis* genome is missing the glutaminyl tRNA synthetase gene, but has the remaining nineteen aminoacyl tRNA synthetases. *Bacillus haloduranS* also lacks the glutaminyl-tRNA synthetase gene, one of two threonyl-tRNA synthetase genes, and two tyrosyl-tRNA synthetase genes (22).

Bacillus stearothermophilus has all of the protein secretion genes identified in *Bacillus subtilis* to include: *secA* (motor), *secY, secE, secG, secF, and secD* (protein conducting channel in the membrane), *yrbF, spoIIIF, comGA, comC* (translocase components), *ffh, ftsY, csaA* (secretion-dedicated chaperones), *groEL, groES, dnaK,*

dnaJ, *grpE* (general chaperones), and *prsA* (folding catalyst) (33). *Bacillus stearothermophilus* and *Bacillus subtilis* have five type I signal pepidase genes used to process secretory preproteins which have been identified to include: *sipS*, *sipT*, *sipU*, *sipV*, and *sipW* (33). The *dnaK* and *dnaJ genes* appear together in *Bacillus stearothermophilus*, while in *Bacillus subtilis*, the genes for heat shock proceed as follows *dnaK-dnaJ-rpoD-dnaG*. In *E. coli*, *dnaK* and *dnaJ* appear on a separate segment quite a distance from *rpo*D-*dna*G (36).

There were at least 123 sporulation genes identified in the *Bacillus stearothermophilus* genome compared to 155 in *Bacillus subtilis* and 38 in *Bacillus halodurans*. Some of the genes involved in sporulation initiation that are in the *Bacillus stearothermophilus* genome include *spoOA*, *spoOF*, *kinA*, *kinB*, *kinC*, and *spoIIE*. As in *Bacillus halodurans*, the *rapA-rapK* genes also were not found in *Bacillus stearothermophilus*, however, the *phrACEGIK* genes were found in *Bacillus stearothermophilus* although they were not found in *Bacillus halodurans*. The transcription initiation genes found in the *Bacillus stearothermophilus* genome include *sigA*, *sigG*, *sigK*, *sigE*, and *sigF*. Some of the genes for stage II cell division (septum formation) found in *Bacillus stearothermophilus* are *ftsA*, *ftsZ*, *divIB*, *spoOJ*, *spoIIE*, and *spoIIIE*, a translocase. There is a stage III gene for spore engulfment that Artemis reveals is *spoIIQ*. Genes involved in septal wall dissolution include *spoIIB*, *spoIID*, *spoIIM*, *spoIIP*, and *spoVG*. Stage IV involves cortex and coat formation. Artemis reveals some of the genes involved in cortex biosynthesis to include *spoVB*, *spoVD*, *spoVE*, *spoVK*, and *spoVR*. Artemis reveals genes involved in coat biosynthesis to include *bofA*, *cotD*, *cotE*, *cotF*, *cotJ*, *cotS*, *spoIIG*, *spoIVA*, *spoIVB*, *spoVD*, *spoVFA*, *spoVR*, and *spoVT*. The germination genes include *gerA*, *gerB*, *gerC*, *gerD*,

gerK, *gerM*, *gerN*, and *cwlC*. The following is a list of sporulation genes that are present in *Bacillus subtilis*. The asterick next to the genes, in table 11, represents the genes that are in the *Bacillus stearothermophilus* genome

Table 11-*Bacillus stearothermophilus* and *Bacillus subtilis* Sporulation Genes

abrA	*divIB**	*phrC*	*spoOC*	*spoIIIC**	*spoVN**
*abrB**	*divIC**	*phrE*	*spoOD*	*spoIIID**	*spoVR**
*bofA**	*ftsA**	*phrG*	*spoOE*	*spoIIIE**	*spoVS**
*bofC**	*ftsZ**	*phrI*	*spoOF**	*spoIIIF**	*spoVT**
				spoIIIG-	
cge	*gerA**	*phrK*	*spoOG*	*sigG**	*spoVIA*
*cotA**	*gerB**	*rapA**	*spoOH*	*spoIIIJ**	*spoVIB*
*cotB**	*gerC**	*rapB**	*spoOJ**	*spoIVA**	*spoVIC*
*cotC**	*gerD**	*rapC**	*spoOK*	*spoIVB**	*spoVID**
*cotD**	*gerE**	*rapD**	*spoOL*	*spoIVC**	*spsA**
*cotE**	*gerF**	*rapE**	*spoOM**	*spoIVD**	*spsB**
			spoIIA-		
*cotF**	*gerG*	*rapF**	*sigF**	*spoIVE**	*spsC**
*cotG**	*gerH*	*rapG**	*spoIIB**	*spoIVCA**	*spsD**
*cotH**	*gerI*	*rapH**	*spoIIC**	*spoIVCB**	*spsE**
*cotJA**	*gerJ*	*rapI**	*spoIID**	*spoIVF**	*spsF**
*cotJC**	*gerKA**	*rapJ**	*spoIIE**	*spoIVG*	*spsG**
*cotS**	*gerKB**	*rapK**	*spoIIF**	*spoVA**	*spsI**
			spoIIG-		
*cotT**	*gerKC**	*sigA**	*sigE**	*spoVB**	*spsJ**
*cotV**	*gerM**	*sigE**	*spoIIJ**	*spoVC**	*spsK**
*cotW**	*gpr**	*sigF**	*spoIIM**	*spoVD**	*spsL**
*cotX**	*gsp10*	*sigG**	*spoIIN**	*spoVE**	*sspA**
*cotY**	*hpr**	*sigK**	*spoIIP**	*spoVF**	*sspB*
*cotZ**	*kbaA**	*sinI*	*spoIIQ**	*spoVG**	*sspC**
*cwlC**	*kinA**	*spI**	*spoIIR**	*spoVH*	*sspD*
*cwlD**	*kinB**	*spoCM*	*spoIIS**	*spoVJ**	*sspE**
*dacB**	*kinC**	*spoOA**	*spoIIIA**	*spoVK**	*sspF*
*dacF**	*phrA*	*spoOB**	*spoIIIB**	*spoVM*	

The following competence genes are found in *Bacillus stearothermophilus* and *Bacillus subtilis*: *abrB*, *cinA*, *codY*, *comA*, *comB*, *comC*, *comD*, *comE*, *comF*, *comG*, *comI*, *comJ*, *comK*, *comL*, *comP*, *comQ*, *comEA*, *comEB*, *comEC*, *comER*, *comFA*, *comFC*, *comGA*, *comGB*, *comGC*, *comGD*, *degU*, *mecA*, *mecB (clpB)*, *mecC*, *sinR* (24). The transcription factor ComK is pivotal in competence regulatory pathway since

it activates expression of all of the identified competence genes that encode the DNA processing and uptake machinery.

ComK also increases the expression of recA to stimulate recombination between the incoming DNA and the chromosome. ComK encodes the competence transcription factor (CTF) which binds upstream of the late competence genes and is necessary for transcription from these promoters. Genes that control the activation and accumulation of ComK either directly or indirectly are *comQ*, *codY*, *abrB*, *sinR*, and *degU*. ComQ produces active ComX pheromone. MecA and MecB act together to inactivate CTF. CodY probably inhibits transcription of comK. AbrB, DegU, and SinR might stimulate transcription of *comK* (32).

I could not verify the following genes in either genome: *comM*, *comN*, *and com*. The following gene sequences were too short, resulting in an error message at the *Bacillus stearothermophilus* BLAST server: *comS*, *comX*, and *csf*. Only the following thirteen competence genes were identified in *Bacillus halodurans cinA*, *comC*, *comEA*, *comEB*, *comEC*, *comER*, *comFA*, *comFC*, *comGA*, *comGB*, *comGC*, *comGD*, and *mecA*. All thirteen were expressed in the late stage of competence. No genes were found in *Bacillus halodurans* that are expressed in early competence. Competence has not been experimentally demonstrated in *Bacillus halodurans* (22).

In *Bacillus stearothermophilus* and *Bacillus subtilis*, the following genes for genetic recombination are present (24).

Table 12-Recombination Genes

Bsu gene	*Bst* gene
addA	yes
addB	yes
recA	yes
recF	yes
recN	yes
recO	yes
recR	yes
recS	yes
recU	yes

ruvA	yes
ruvB	yes
sbcD	yes
ssb	yes
ywpH	yes
ylpB	yes
yocI	yes
yorK	yes
yqhH	yes
yrrC	yes
yrvE	yes
ywqA	yes

In *Bacillus stearothermophilus*, the eight proteins in the ATP synthase operon appear as follows: ATP synthase protein I-ATP synthase *ABDA*-gamma-*B*-epsilon. The ATP synthase operons conserved in *Bacillus subtilis* and *E. coli* appear as one operon. The F0F1 type ATP synthase genes are *atpBEFHAGDC'* (102).

In *Bacillus stearothermophilus*, Artemis reveals the *ctaA* and *ctaB* genes, and possibly the *cbdA* and *cbdB* genes were identified as part of the cta operon. The *cta* operon in *Bacillus subtilis* encoding cytochrome aa3 complex and the *cyo* operon in *E. coli* encoding cytochrome b03 ubiquinol oxidase are homologous with the following gene order *ctaB-ctaC'-ctaD-ctaE-ctaF* in *Bacillus subtilis* and *cyoE-cyoA-cyoB-cyoC'-cyoD-cyoE* in *E. coli*. The gene order in the two operons is identical except that the first gene *ctaB* corresponds to the last gene *cyoE* (36).

Artemis reveals the longest operon in the *Bacillus stearothermophilus* working draft genome encoding flagellar, chemotaxis, and motility genes appearing at location 2,812,750 as follows: *flgBC'-fliFGHIJK-flaA-flgG-fliLMY-cheY-fliZPR-flhBAF-cheBAWCD*. Other related genes found in the genome included: *fliS, flgM, flgK, flgL, fliD, and flit*, while a total of 39 genes were identified with the flagellar, chemotaxis, and motility functions. The gene order for flagellar, chemotaxis, and motility functions in

Bacillus subtilis cluster as follows: *flgBC-fliEFGHIJK-flgG-fliLMY-cheY-fliZPQR-flhBAF-cheB-cheAW* (36). All flagellin genes for *Bacillus subtilis* are found in table 13. In *E. coli*, the above genes are found in three separate locations along the chromosome. The *flgBC* genes are together. The *flgG* gene is alone. The remaining gene order is found: *flgNMABCDEFGHIJKL* (36).

Table 13-*Bacillus subtilis* Flagellin Proteins

Bsu gene	*Bsu* Function	*Bst* High Score	*Bst* P (N)	*Bst* Contig	*Bst* Identities
flhA	export apparatus/	280	9.0 e-209	388	93%
flhB	gate keeper	260	6.8 e-80	388	78%
sigD	sigma factor	275	2.8 e-47	388	85%
hag	flagellin	252	2.9 e-27	393	85%
fliD	HAP2	206	7.7 e-44	393	70%
fliE	rod protein	136	3.2 e-13	388	62%
fliF	MS ring	220	4.1 e-74	388	70%
fliG	C ring	245	1.1 e-105	388	78%
fliH	delivery complex	150	5.7 e-18	388	43%
fliI	delivery complex (ATPase)	265	1.8 e-113	388	85%
fliJ	delivery complex/chaperone	140	4.5 e-13	388	49%
fliK	hook length	127	7.9 e-10	388	53%
fliL	Unknown	140	4.4 e-13	388	46%
fliM	C ring	254	5.9 e-88	388	70%
fliP	C rod	229	1.5 e-91	388	70%
fliQ	C rod	213	3.2 e-25	388	66%
fliR	C rod	212	1.1 e-40	388	70%
fliS	Chaperone?	218	1.9 e-43	393	70%
fliT	Chaperone?	56	0.037	393	28%
fliZ	Unknown	153	3.3 e-14	388	56%
flgB	rod protein	139	5.9 e-22	388	63%
flgC	rod protein	214	8.1 e-36	388	66%
flgE	hook protein	215	3.8 e-39	388	74%
flhO	rod protein	192	1.7 e-34	165	63%
flgK	HAP1	178	9.6 e-64	393	57%
flgL	HAP3	177	2.1 e-26	393	60%
flgM	Anti-sigma	100	1.5 e-11	393	42%
motA	Protein conductor	258	1.3 e-90	332	81%
motB	Anchor to peptidoglycan	231	1.2 e-59	332	71%
flhF	GTP binding	214	2.2 e-54	388	70%
flxA	Rod cap, muramidase	249	9.1 e-	354	81%

yfxH	Cell division inhibitor	157	7.1 e-28	388	56%
yfxQ	Export apparatus	194	1.7 e-24	207	64%
yfxG	Hook cap	162	4.7 e-17	388	69%
yfxL	Chaperone	82	0.00012	388	48%
yfxY	Periplasmic chaperone	216	1.8 e-75	374	64%
flhP	rod protein	159	3.1 e-18	165	53%

The *Bacillus stearothermophilus* genome includes the following genes of the purine operon, *pur2, pur3, pur5, pur1, purL, purQ, pur7, pur8, purK,* and *pur6.* In *Bacillus subtilis,* the genes involved in purine biosynthesis have the gene order *purEKBCSLQFMNHD.* In *E. coli,* the purine biosynthesis genes appear in three clusters with much distance between them. The gene order is *purEK, purMN, purHD* (36, 102)

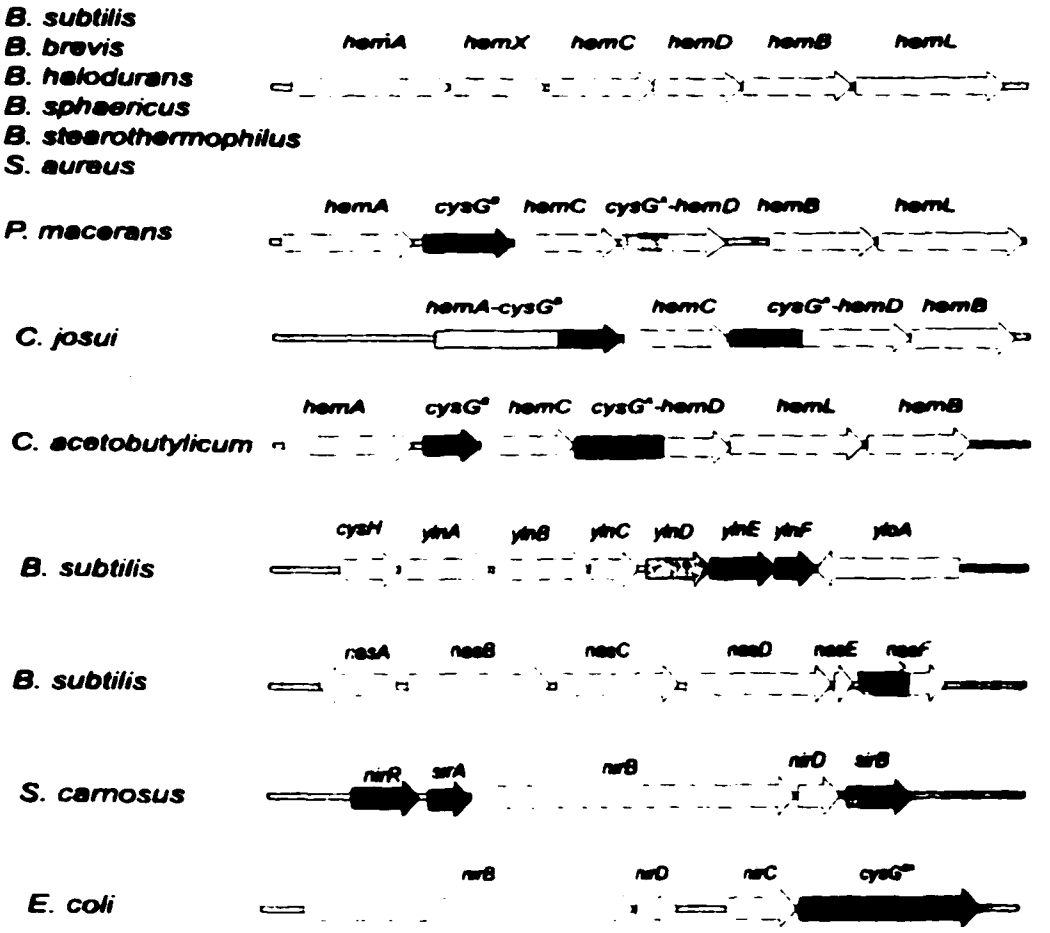
The pyrimidine operon in *Bacillus stearothermophilus* includes the following genes in order *pyrR-pyrP-pyrB-pyrC.* In *Bacillus subtilis,* the genes involved in pyrimidine biosynthesis are *purA, pyrB, pyrC, pyrD, pyrE, pyrF,* and *pyrG* (24)

All of the *trp* genes to include *trpE, trpC, trpD, trpF, trpB, and trpA* are identified in the *Bacillus stearothermophilus* genome. The *Bacillus subtilis* genes found in tryptophan biosynthesis operon are *trpEDCFBA.* The *trpG* gene is not located with the cluster but is found in the folate operon. In *E. coli,* the gene order is *trpEDCBA* (36)

Artemis reveals the extracellular neutral metalloproteases, *nprA, nprB, nprE,* a probable neutral zince metalloprotease, *len,* and a serine protease called bacillopeptidase F, *bprF. Bacillus stearothermophilus* antibiotic resistance genes revealed by Artemis include an acriflavin resistance protein, *yerP,* an antibiotic resistance protein, *ywoG,* and the subtilin biosynthetic regulatory protein, *spar.* A *BLAST* of the genome reveals mycosubtilin synthetase genes, *mycA* and *mycB,* the iturin synthetase A gene, *BAB6,* and bacitracin resistance protein and ABC transporter, *bcrA.*

Iron functions in sensing redox and oxygen levels and is required for growth in microorganisms. Siderophores have a high affinity for iron, which presents itself as iron-sulfur clusters or heme. The order of the genes in the *Bacillus subtilis* heme operon is *hemAX(')DBL* and the heme operon in various species is compared in figure 10 (24).

Figure 10-Heme Operon (24)



Reprinted with permission from the American Society for Microbiology coypyright 2002 by ASM Press.

Siderophores are important for the uptake of insoluble, oxidized iron by cells aerobically and anaerobically with nitrite versus those grown by fermentation. When siderophore production is coupled with a specific uptake system, they allow microorganisms to acquire iron in environments where there is a low concentration of

iron. Table 14 is a list of the genes required for synthesis and utilization of siderophore 2,3-dihydroxybenzoate.

Table 14-Siderophores

Bsu genes	*Bst* genes	Protein
dhbA	yes	2,3-dihydro-2,3-dihydroxybenzoate dehydrogenase
dhbC	yes	isochorismate synthase
dhbE	yes	2,3-dihydroxybenzoate-AMP ligase
dhbB	yes	isochorismatase
dhbF	yes	involved in 2,3-dihydroxybenzoate biosynthesis

Differentiation has many steps at which autolysins are active. During germination, enzymes degrade the cortex to allow spore outgrowth. *Bacillus subtilis* has approximately 35 peptidoglycan hydrolases called autolysins classified as muramidases, lytic transglycosylases, glucosaminidases, amidases, and endopeptidases (24). In the vegetative cell, peptidoglycan is continuously being synthesized, modified, and degraded to allow for cell growth, division, and other important cellular functions. After synthesis, the spore cortex is modified to allow for it to assume its final structure, which maintains spore dormancy. The autolysins found in *Bacillus stearothermophilus* and *Bacillus subtilis* are in the below table (27).

Table 15-*Bacillus subtilis* Autolysins

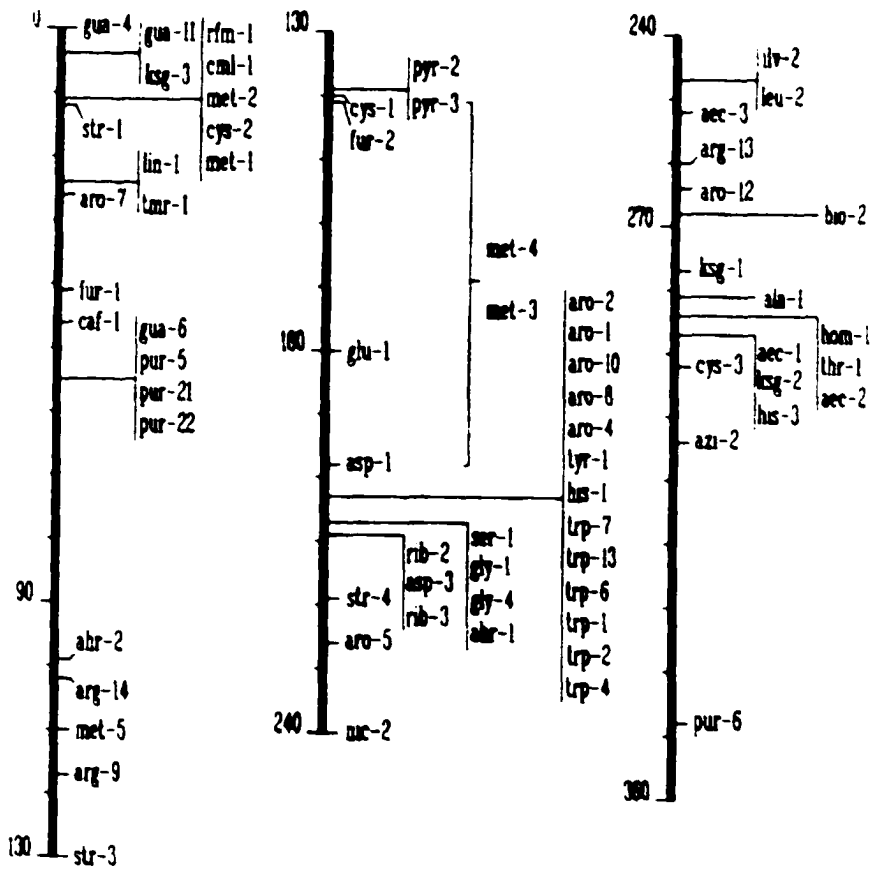
function of autolysin	*Bsu* gene	*Bst* gene
mother cell lysis	*lytC*	yes
	cwlC	yes
coat synthesis and cortex maturation	*cwlD*	yes
germination and outgrowth	*sleB*	yes
	cwlJ	yes
cell elongation	*lytC*	yes
peptidoglycan and cell wall turnover	*lytD*	no
cell separation	*lytC*	yes

		lytD	yes
		lytE	yes
		lytF	yes
		ywbG	yes
	motility	*lytC*	yes
		lytD	yes
		lytF	yes

2. *Bacillus stearothermophilus*, strain NUB36 Map Comparison with *Bacillus stearothermophilus*, strain 10

There are many similarities between the annotated working draft of *Bacillus stearothermophilus* strain 10 genome and a map of *Bacillus stearothermophilus* strain NUB36 Map published in 1993, which is represented below (29)

Figure 11-Linear Scale Drawings Representing the Circular Linkage Map of the *Bacillus stearothermophilus*, strain NUB36 Chromosome (29)



Printed by permission of Dr. Neil Welker and the American Society of Microbiology

Although there are similarities there are several differences in the naming convention between the two strains. There are no comparable genes in *Bacillus stearothermophilus* strain 10 for the following genes that exist in strain NUB36

Table 16-Genes Missing from strain 10 versus strain NUB36

strain NUB36 gene	strain NUB36 Phenotype
ahr-1	resistance to arginine hydroxamate
azi-2	resistance to sodium azide
glh-1	resistance to gamma-glutamylhydrazide
rfm-1	resistance to rifampin
tmr-1	resistance to tunicamycin
tsg-3	temperature sensitive growth at 70 degree Celsius

There are comparable genes in *Bacillus stearothermophilus* strain 10, however the names are different, for the following genes that exist in strain NUB36

Table 17-Differences in Gene Names between strain 10 and strain NUB36

strain NUB36 gene	strain NUB36 Phenotype	strain 10 gene	strain 10 Phenotype
caf-1	resistance to caffeine	no gene name given	synergistic inhibition of glycogen phosphorylaseA by a potential antidiabetic drug and caffeine
fur-1	resistance to 5-fluorouracil	no gene name given	dihydropyrimidine dehydrogenase (NadpH & 5-fluorouracil)
fut-1	fumarate utilization	*fumC*	fumarate hydratase
		frdA	fumarate reductase
lin-1	resistance to lincomycin	*lmbA*	

3. *Bacilli* Genetic Map Comparisons

The genetic maps for seven *Bacilli* are presented below in figure 12 and 13 with linearized chromosomes. The map comparisons of the six *Bacilli* to include *Bacillus*

megaterium, *Bacillus stearothermophilus*, *Bacillus thuringiensis*, *Bacillus amyloliquefaciens*, *Bacillus licheniformis*, and *Bacillus subtilis* are aligned when possible, but are not to scale (29). In figure 12, the 1993 map shows only the *Bacillus stearothermophilus* chromosome in one piece. There are similarities and differences between the gene order of *Bacillus stearothermophilus* and other *Bacilli* according to the map in figure 12. One of the differences in gene order between these organisms is the *Bacillus stearothermophilus* genes that are not listed on the other *Bacilli* maps and vice versa. The following observations are among similar genes and their location on the genetic maps between the two organisms being compared.

Bacillus stearothermophilus and *Bacillus megaterium* retain similar gene order for the following genes: *purA*, *rpsL*, *purB*, *arg*, *pyrB*, *hisH*, *trp*, *ilvB C*, *leu*, *his*, and *azi*. The difference involves the order of the subunits in the *trp* operon.

Bacillus stearothermophilus and *Bacillus thuringiensis* have similar order for *pur*, *rpoB*, *rpsL*, *arg*, *pyr*, *trp*, *his*, and *cys* even though, large amounts of the *Bacillus thuringiensis* map is missing. These two organisms also differ in their location of *metA* along the chromosome.

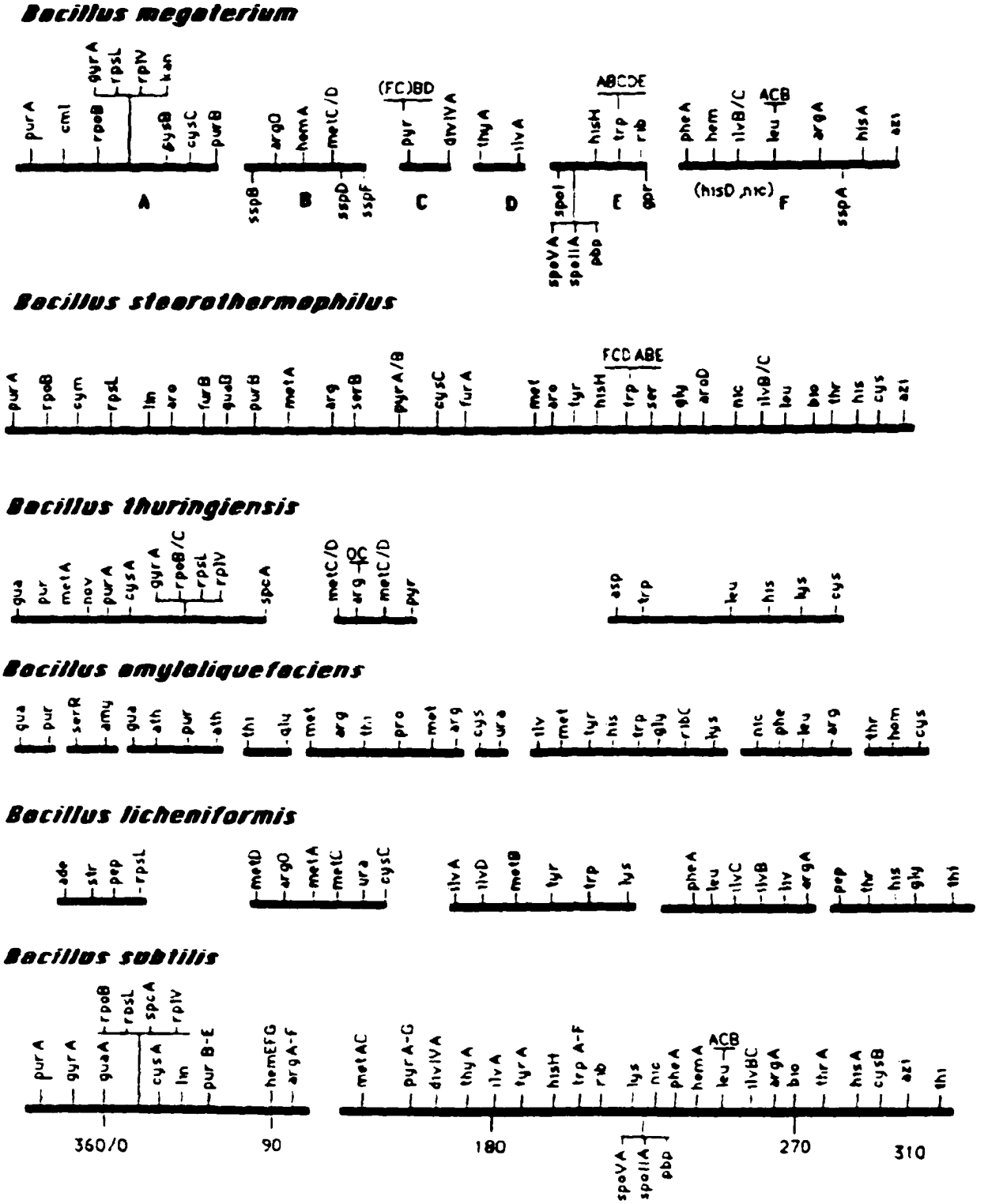
Bacillus stearothermophilus and *Bacillus amyloliquefaciens* have similar gene order for *pur*, *met*, *arg*, *cys*, *met*, *tyr*, *his*, *trp*, *gly*, *nic*, *leu*, *thr*,and *cys*. No differences in gene order of similar genes were noted.

Bacillus stearothermophilus and *Bacillus licheniformis* have similar gene order for *rpsL*, *argO*, *metA*, *cysC*, *met*, *tyr*, *trp*, *thr*, and *his*. The *leu* and *ilvC B* gene order is reversed.

Bacillus subtilis origin starts with *guaA* at zero degree. *Bacillus stearothermophilus* and *Bacillus subtilis* have similar gene order for *purA*, *rpoB*, *rpsL*, *lin*, *purB*, *arg*, *pyrAB*, *tyrA*, *hisH*, *trpA-F*, *nic*, *thr*, *his*, *cys*, and *azi*. These two organisms differ in the gene order of *leu* and *ilv*. Researchers working with *Bacillus stearothermophilus* DSM 22, who changed the taxonomy from *Bacillus stearothermophilus* to *Geobacillus*

stearothermophilus stated that genome organization is not conserved between *Bacillus stearothermophilus* and *Bacillus subtilis* citing the less than twenty percent similarity between the two genomes (4)

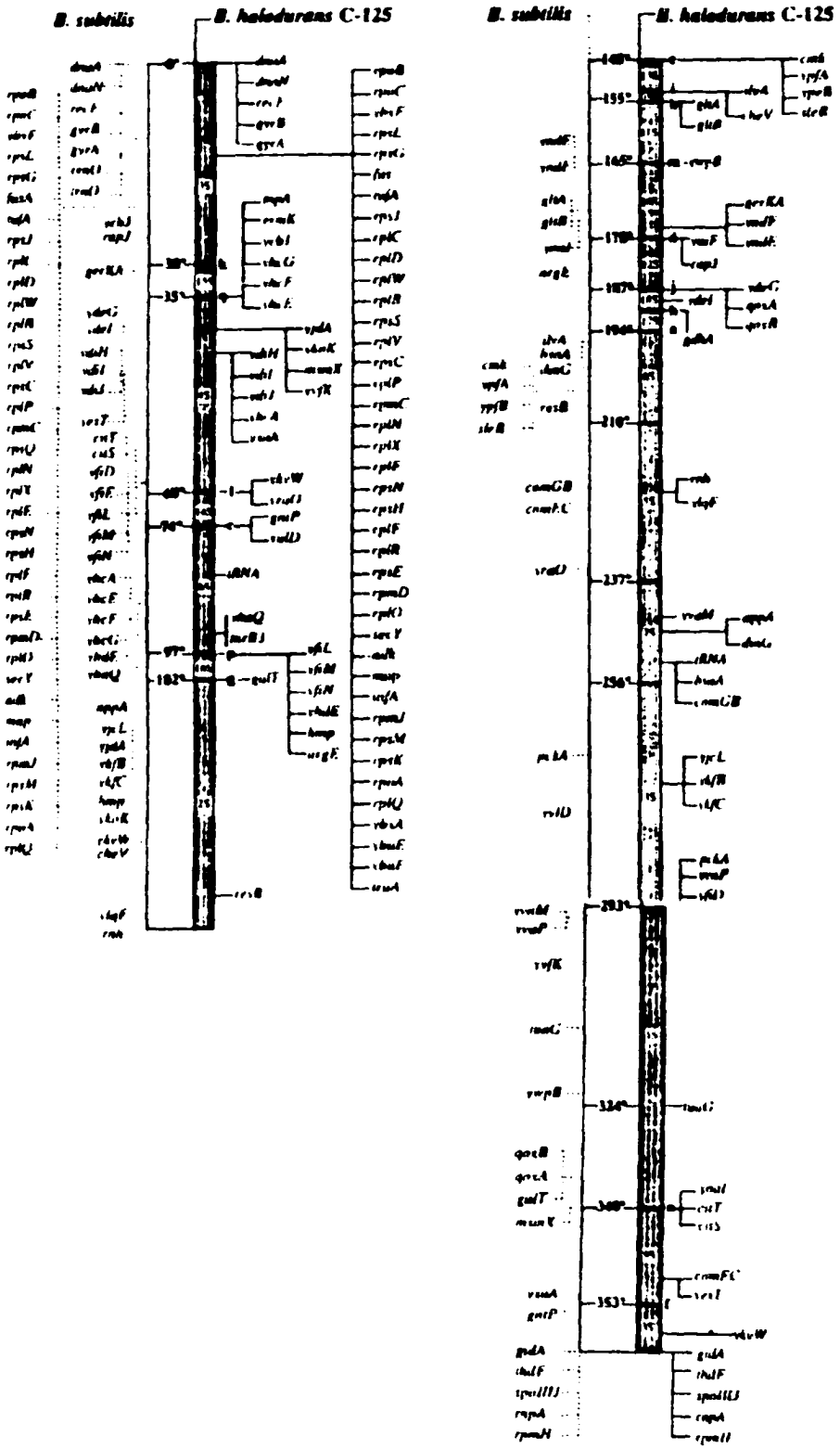
Figure 12-Comparison of the Genetic Map of Six *Bacilli* (29)



Reprinted with permission from the American Society for Microbiology coypyright 1993 by ASM Press.

The genetic map of *Bacillus halodurans*, represented with solid lines and *Bacillus subtilis*, which is represented with dashed lines shown in the below figure 13, is drawn to the scale of 360 degrees, beginning with zero at the *dnaA* locus (103). Throughout this dissertation, the genes in *Bacillus halodurans* are discussed.

Figure 13-Genetic Map of *Bacillus subtilis* and *Bacillus halodurans* (103)

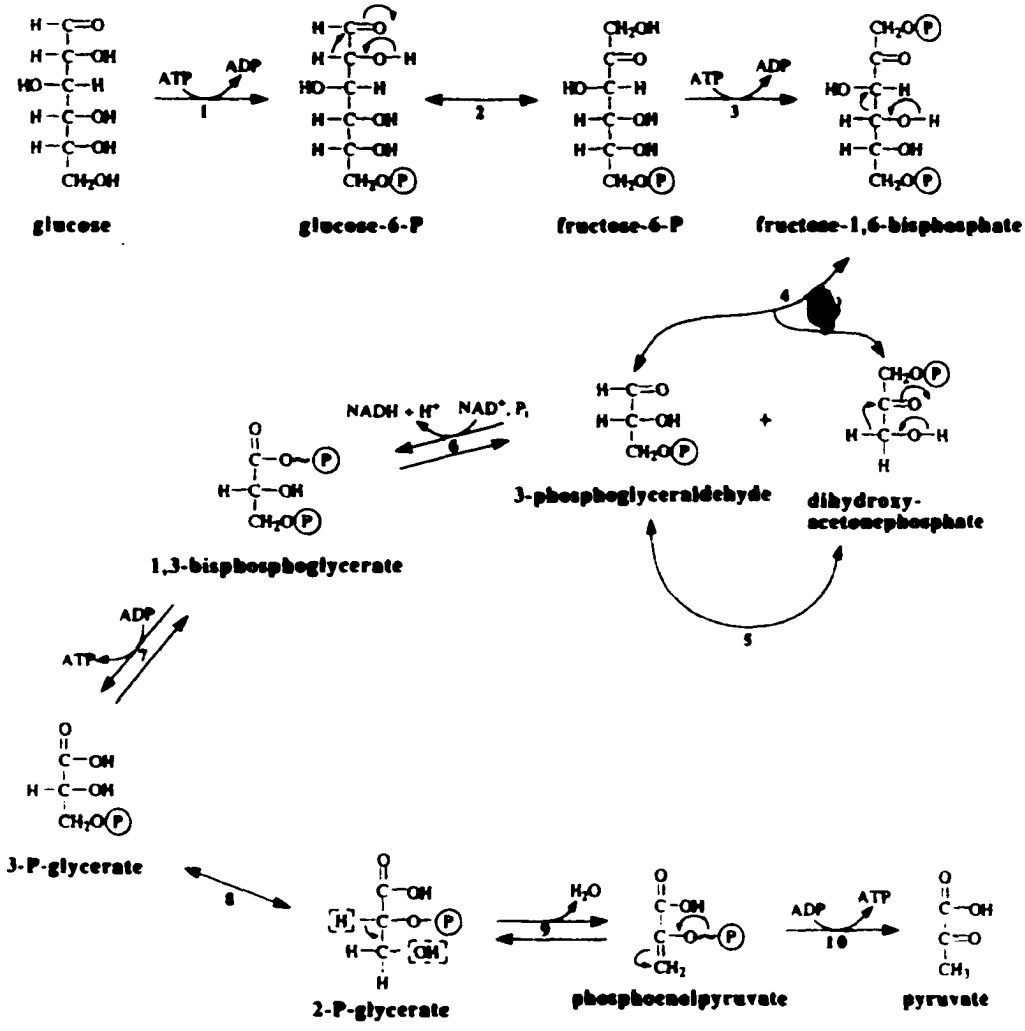


C. *Bacillus stearothermophilus* and *Bacillus subtilis* Metabolism

1. Carbohydrate Metabolism

The three pathways of glucose catabolism in microorganisms are the Embden-Meyerhof-Parnas pathway (EM), the Pentose Phosphate Pathway, and the Entner-Doudoroff pathway. Catabolism of galactose and maltose involves the formation of glucose 1-phosphate, which is then converted to glucose 6-phosphate by the enzyme phosphoglucomutase. The proteins from glycolysis that are present in both *Bacillus subtilis* and *Bacillus stearothermophilus* include glucokinase, glucose-6-phosphate isomerase, 6-phophofructokinase, fructose 1-6 bisphosphate aldolase, triosephosphate isomerase, glyceraldehyde 3-phosphate dehydrogenase, phosphoglycerate kinase, phosphoglycerate mutase, enolase, and pyruvate kinase. Thus, although, hexokinase is not present in either *Bacillus* species both *Bacillus subtilis* and *Bacillus stearothermophilus* encode the enzymes needed for the complete glycolytic pathway (105)

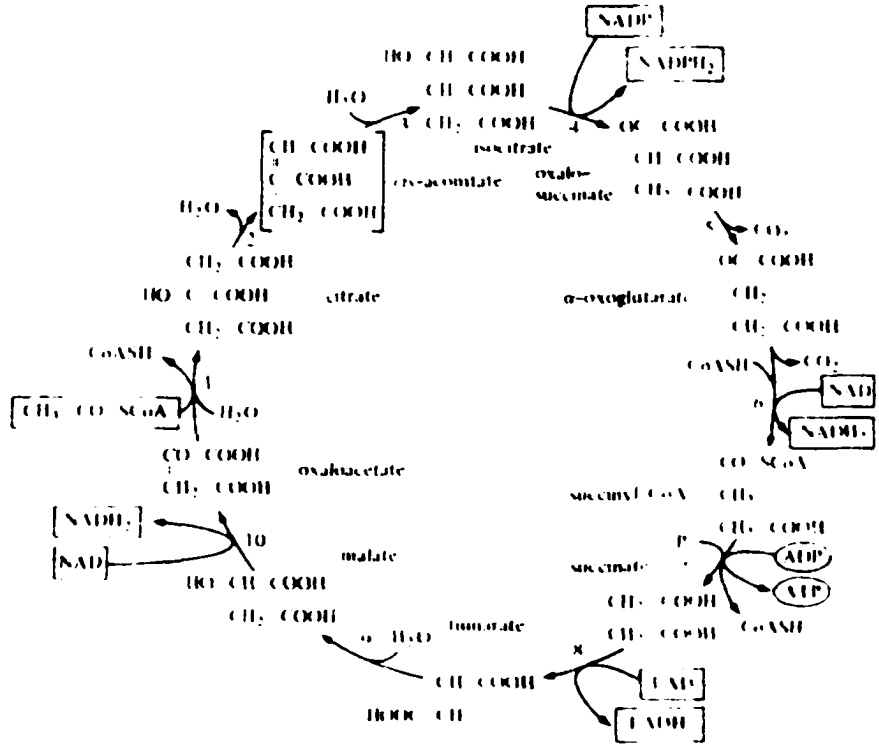
Figure 14-Glycolysis Pathway (105)



Tricarboxylic Acid Cycle

In the tricarboxylic acid cycle, the proteins that are present in both *Bacillus subtilis* and *Bacillus stearothermophilus* include citrate synthase I, aconitase hydratase, isocitrate dehydrogenase, 2-oxoglutarate dehydrogenase, 2-oxoglutarate dehydrogenase, pyruvate dehydrogenase, succinyl-CoA synthetase, succinate dehydrogenase, fumarase hydratase, malate dehydrogenase, and pyruvate carboxylase (105). This cycle is complete in both *Bacillus subtilis* and *Bacillus stearothermophilus*.

Figure 15-TCA Cycle (112)



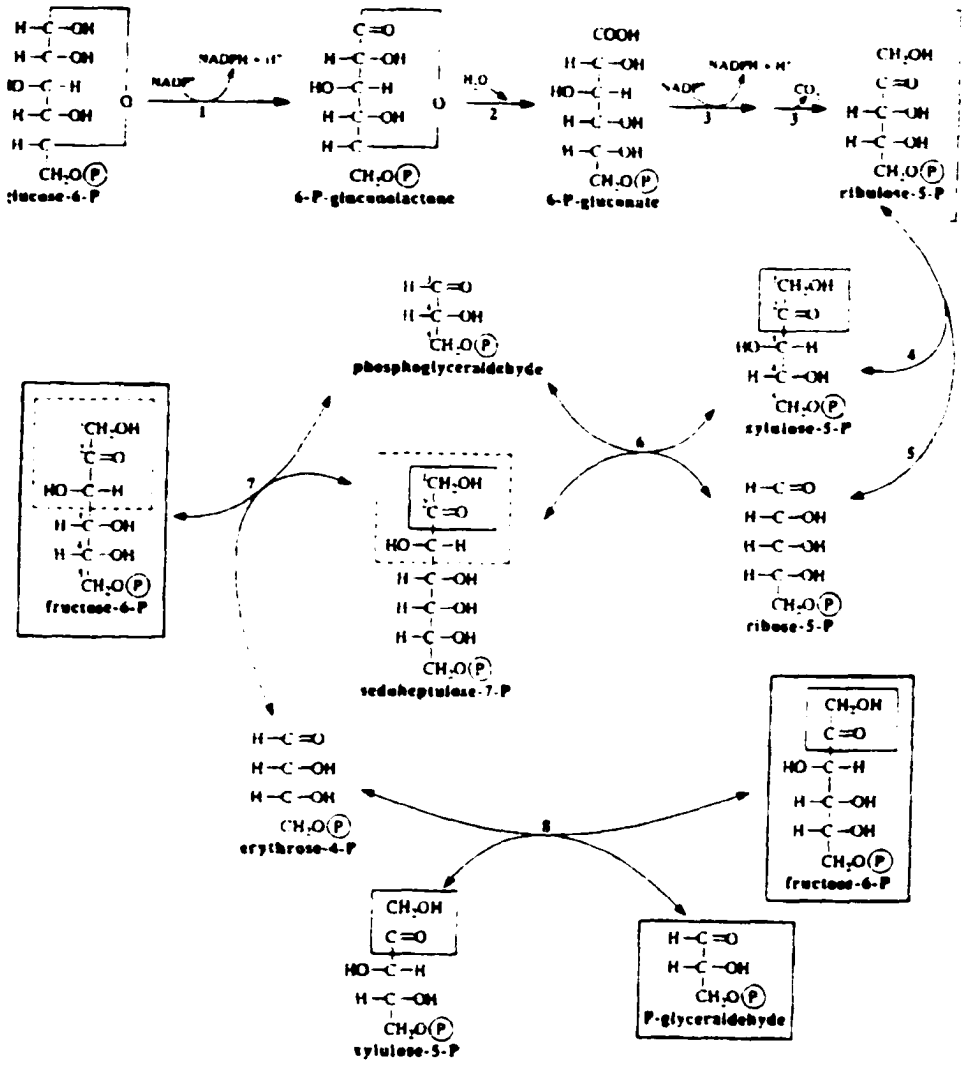
Reprinted by permission of Springer-Verlagо GmbH & Co.KG copyright by Springer-Verlag (1986).

Pentose Phosphate Cycle

This cycle is important because it produces the precursors (the pentose phosphates) to the ribose and deoxyribose, and this cycle provides precursors (erythrose phosphate) to the aromatic amino acids. This pathway results in NADPH, which is a major source of electrons for biosynthesis in most of the pathways. Organisms that lack stage 1 of the EMP pathway and the enzymes of the Entner-Doudoroff pathway oxidize glucose to phosphoglyceraldehyde using this pentose phosphate cycle

In the pentose phosphate pathway, the proteins that make a complete pathway for *Bacillus stearothermophilus* and *Bacillus subtilis* include, *ygiJ*, encoding glucose-6-phosphate dehydrogenase, *gntZ*, encoding gluconate-6-phosphate dehydrogenase, *ywlF*, encoding ribose 5-phosphate isomerase (pentose phosphate), *prs*, encoding ribose-P pyrophosphokinase, *yloR*, encoding ribulose-5-phosphate-3 epimerase, and *tkt*, encoding transketolase (24)

Figure 16-Pentose Phosphate Cycle (105)



2. Energy Metabolism

Electron Transport & Oxidative Phosphorylation

The electron transport system moves electrons through a series of reversible electron carriers with successively higher oxidation-reduction potentials. The spontaneous transfer in *Bacillus subtilis* is used to drive reactions that require energy, including solute transport, ATP synthesis through oxidative phosphorylation, and motility. Different membrane bound dehydrogenases (succinate, NADH, and glycerol-3-phosphate) oxidize water-soluble substrates to reduce menaquinone in the membrane to menaquinol (29).

Menaquinone consists of a naphthoquinone ring joined with many isoprenoid side chains with MK-7 (7 isoprenoid side chains) being the most abundant in *Bacillus subtilis*. Gram-positive bacteria have menaquinone only as their redox quinone versus both menaquinone and ubiquinone like Gram-negative bacteria. Succinate dehydrogenase, which is part of the TCA cycle by catalyzing the two electron oxidation of succinate to fumarate in the cytoplasm, couples this reaction to the reduction of menaquinone to menaquinol in the membrane. Succinate dehydrogenase is weakly oxidized in *Bacillus subtilis*. Each of the three NADH dehydrogenases, major MK reductases in *Bacillus subtilis*, consists of one polypeptide and one flavin adenine dinucleotide as a prosthetic group (29).

Under aerobic conditions, menaquinol is oxidized by one of several menaquinol oxidases or menaquinol: cytochrome *c* oxidoreductase (bc complex) (24). The bc complex consists of three proteins that include the Rieske protein, encoded by *qcrA*, containing one (2Fe-2S) cluster; the cytochrome b subunit, encoded by *qcrB*, containing two hemes, and the cytochrome c subunit, encoded by *qcrC*, containing one

heme C. One element of the bc complex, cytochrome *b*-559, is a transmembrane protein associated with succinate dehydrogenase possibly to anchor the enzyme to the cytoplasmic side of the bacterial membrane. Four c-type cytochromes are identified in *Bacillus subtilis* with *c*-550 being the most abundant. The cytochrome *c*-554 donates electrons to *c*-551 *ctaC* of the cytochrome *caa3* (106).

The final step in aerobic respiration, the four electron reduction of dioxygen to two water molecules, is catalyzed by membrane bound enzymes called terminal oxidases. The terminal oxidases are heme proteins or copper-heme proteins that oxidize lower potential reductants and use the electrons to reduce oxygen to water. Terminal oxidases either act as proton pumps or consume protons on the inner side of the cell membrane during the reduction of oxygen. There are three terminal oxidases in *Bacillus subtilis* to include cytochrome *o*, *caa3* and *aa3* cytochrome oxidase.

In *Bacillus stearothermophilus*, succinate dehydrogenase is associated with cytochrome *o* (B-558), *caa3* and *cao* cytochrome oxidase. Cytochrome *o* (B-558) has heme B prosthetic group. *Cao* oxidase has heme O as its oxygen-reactive center. Heme B and heme O have different structures at C-2, C-4, C-5, and C-8. The terminal oxidase *caa3* is encoded by the *cta* operon having the molecular genetics of *ctaCDEF* with *ctaA* and *ctaB* as regulatory genes located adjacent to the *cta* operon (29). *Bacillus stearothermophilus* also has the *Bacillus subtilis* terminal oxidase aa3 cytochrome which is encoded by *qoxABCD*.

Bacillus stearothermophilus is aerobic and facultatively anaerobic. Aerobically, the terminal electron acceptor is oxygen. Anaerobically, *Bacillus subtilis* can either undergo fermentation or respire with nitrate or nitrite as terminal electron acceptor. The control of *Bacillus subtilis* switching from aerobic to anaerobic metabolism is

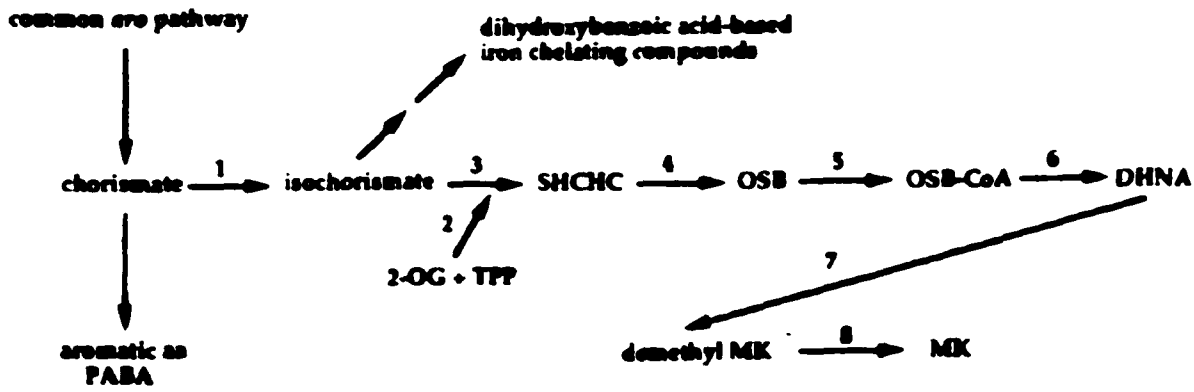
controlled at the transcriptional level

Menaquinone Biosynthesis

Menaquinone (MK) is a lipophilic, non protein redox component of the electron transport chain that shuttles electrons from dehydrogenases and cytochromes. MK-7, which is the major quinone in *Bacillus stearothermophilus*, is necessary for sporulation and proper regulation of cytochrome formation (4, 24). A precursor of menaquinone is vitamin K. MK is an anaerobic acceptor that is synthesized from chorismate. There are eight genes involved in menaquinone biosynthesis to include the following: 1) (*menF*) isochorismate synthase, 2) 2-oxoglutarate decarboxylase, 3) (*menD*) 2-succinyl-6-hydroxy-2,4-cyclohexadiene-1-carboxylate (SHCHC) synthase, 4) (*menC*) o-succinyl benzoate-CoA synthase, 5) (*menE*) o-succinylbenzoic acid-CoA ligase, 6) (*menB*) 1,4-dihydroxy-2-naphthoic acid synthase, 7) (*menA*) 1,4-dihydroxy-2-naphthoateoctaprenyltransferase, and 8) (*menH*) methyltransferase (107). *Bacillus stearothermophilus* and *Bacillus subtilis* have complete pathways.

In the menaquinone biosynthesis pathway below, the abbreviations are 2-OG, 2-oxoglutarate; TPP, thiamine pyrophosphate; OSB, o-succinyl benzoate synthase; OSB-CoA, o-succinyl benzoate-CoA synthase; DHNA, 1, 4-dihydroxy-2-naphthoic acid; and PABA, *p*-aminobenzoic acid.

Figure 17-Menaquinone Biosynthesis (107)



Reprinted from *Gene*, volume 167, Rowland, B., Hill, K., Miller, P., Driscoll, J., Taber, H., "Structural Organization of a *Bacillus subtilis* Operon Encoding Menaquinone Biosynthetic Enzymes," pages 105-109 (1995), with permission from Elsevier Science.

Fermentation

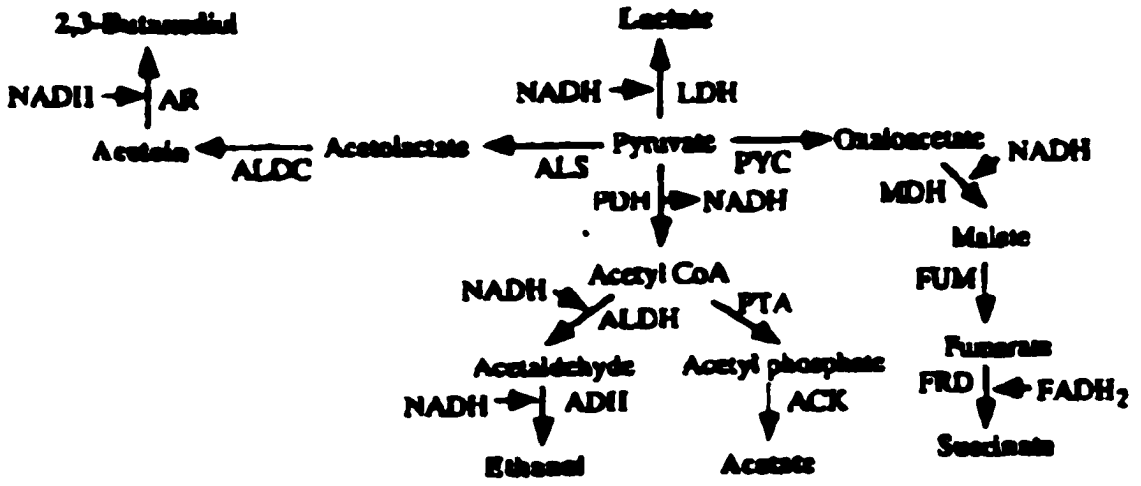
Anaerobically growing prokaryotes reoxidize NADH and other reduced electron carriers either by anaerobic respiration (using nitrate, sulfate, or fumarate as the electron acceptor) or they carry out fermentation. Fermentation is defined as a pathway in which a reduced electron acceptor NADH is reoxidized by metabolites produced by the pathway. The redox reactions occur in the cytosol. In fermentation, NADH generated by glycolysis cannot be reoxidized by electron transport systems. In the absence of external electron acceptors, such as, oxygen in aerobic respiration and nitrate or nitrite in anaerobic respiration, such organisms can grow anaerobically by fermenting sugars. Thus NAD+ is regenerated by the conversion of pyruvate to fermentative end products while ATP is generated by substrate-level phosphorylation (108).

Bacillus stearothermophilus and *Bacillus subtilis* grow anaerobically in the presence of nitrate and nitrite using the regulatory genes *fnr* and *resD-resE*, which are transcribed to activate the transcription of *narK-fnr* and *narGHJI* operons (108). Respiratory nitrate reductase encoded by the *narGHJI* operon is reponsible for respiratory nitrate reductase activity. Molybdenum is a cofactor for the *narG* subunit. Artemis reveals the molybdenum cofactor biosynthesis and transport permease genes as shown in table 18.

The genes for nitrite respiration, *nasDEF*, are found in both *Bacillus stearothermophilus* and *Bacillus subtilis* (108). The large subunit for nitrite reductase is *nasD*, the small subunit is *nasE*. *nasF* helps produce siroheme. The mechanism of nitrate and nitrite uptake is unknown.

Anaerobically, *Bacillus stearothermophilus* and *Bacillus subtilis* undergo mixed acid-butanediol fermentation involving these genes: *ack* encodes acetate kinase, *adh* encodes alcohol dehydrogenase, *aldc* encodes acetolactate decarboxylase, *aldh* encodes aldehyde dehydrogenase, *als* encodes acetolactate synthase, ar encodes acetoin reductase, *frd* encodes fumarate reductase, fum encodes fumarase, *ldh* encodes lactate dehydrogenase, mdh encodes malate dehydrogenase, *pdh* encodes pyruvate dehydrogenase, *pta* encodes phosphotransacetylase, *pyc* encodes pyruvate carboxylase (108).

Figure 18-Fermentation Pathway of *Bacillus subtilis* (108)



Reprinted with permission from Dr. Michiko Nakano and the American Society for Microbiology coypyright 1997 by ASM Press.

Table 18- Molybdenum Biosynthesis Genes

Bsu genes	Protein	*Bst* genes
mobA	molybdopterin-guanine dinucleotide biosynthesis	yes
moeB	molybdopterin biosynthesis protein	
moeA	molybdopterin biosynthesis protein	yes
mobB	molybdopterin-guanine dinucleotide biosynthesis	yes
moaE	molybdopterin converting factor, subunit 1	yes
moaD	molybdopterin converting factor, subunit 2	yes
no		*modB*

no	Molybdenum transport permease	yes
no	Molybdenum cofactor biosynthesis	yes

3. Lipid Metabolism

The three functions of fatty acids are that they function as building blocks of phospholipids and glycolipids, are energy molecules stored as triacylglycerols, and depend on their alkyl chain lengths and their degree of saturation. These chains may contain one or more double bonds, mostly in the cis configuration. Unsaturated fatty acids have a lower melting point than saturated fatty acids of the same length. Chain length also affects the melting point (34). *E. coli* and other Gram-negative bacteria synthesize even-numbered chains of saturated fatty acids and unsaturated fatty acids. In these organisms, unsaturated fatty acids are formed by anaerobic dehydration of a C10 hydroxy intermediate. In contract, *Bacillus* species produce odd and even carbon-numbered branched-chain fatty acids and glycolipids and possess an aerobic desaturation mechanism for unsaturated fatty acid biosynthesis (24)

Fatty acid biosynthesis differs from beta-oxidation (fatty acid metabolism) in the following ways: 1) the reductant is NAD not NADH; and 2) biosynthesis requires carbon dioxide and continues with the derivative malonyl-CoA with acyl carrier protein (ACP). Both the anabolism and catabolism take place in the cytosol. The fatty acid synthase system that is responsible for producing the multitude of fatty acid structures found in bacterial membranes is conserved in prokaryotes and eukaryotes and has an initiation stage and an elongation stage. Myristic and palmitic acids are the most common fatty acids in many organisms but are minor in the genus *Bacillus*. The three proteins involved in the initiation of fatty acid biosynthesis are acetyl-CoA carboxylase, malonyl transacylase, and *B*-ketoacyl-ACP synthase III (24). *Bacillus stearothermophilus* and *Bacillus subtilis* have these three proteins.

Figure 19-Initiation of Fatty Acid Biosynthesis (24)

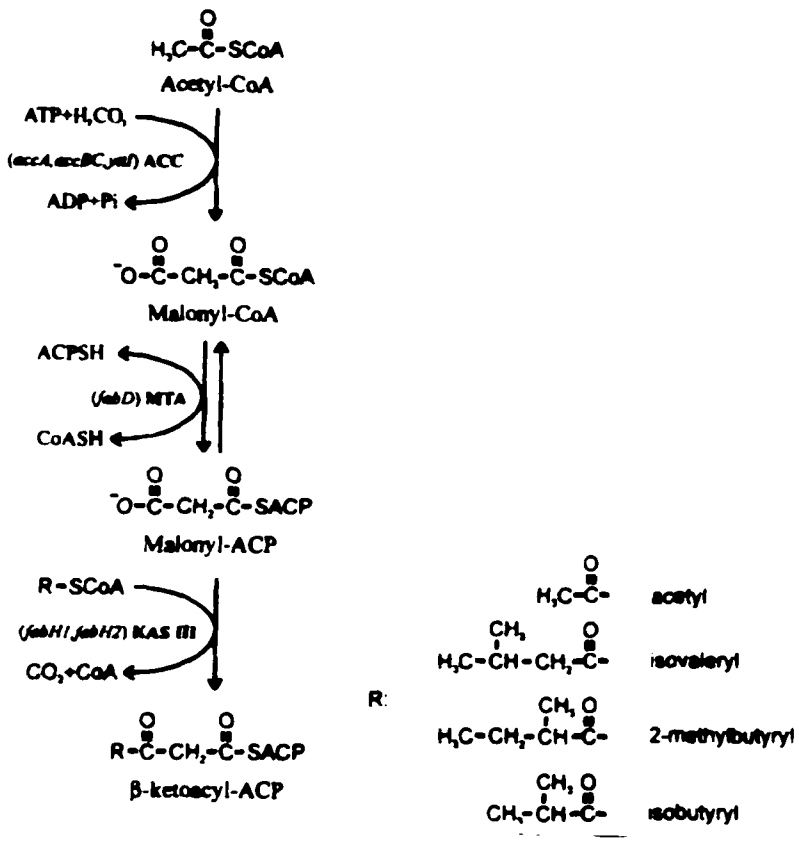


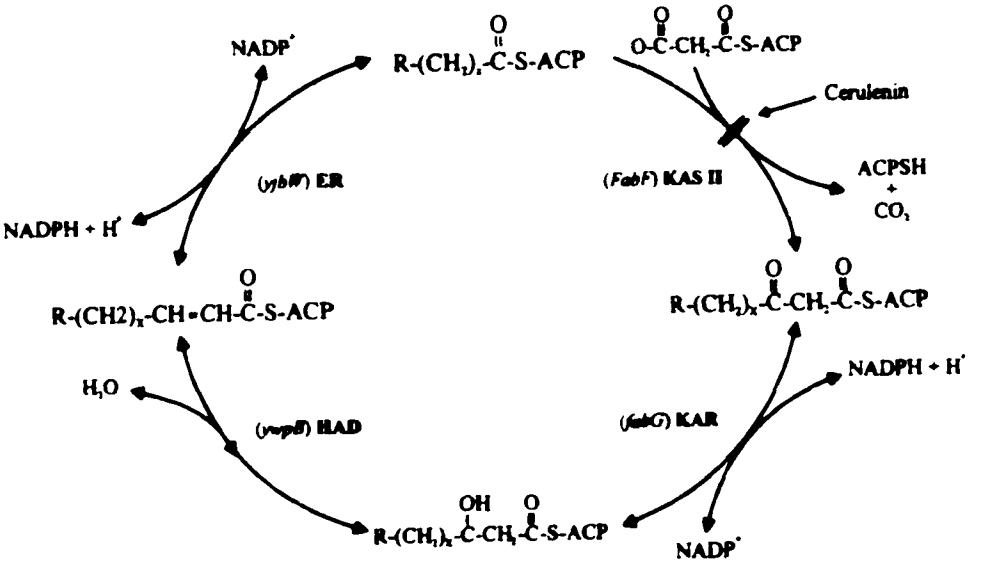
Table 19-Initiation of Fatty Acid Biosynthesis Genes

Bsu genes	Function	Bst
accA	Carboxyltransferase alpha subunit	yes
accB	Biotin carboxylase carrier subunit	yes
accD	Biotin carboxylase subunit	yes
acpA	Acyl carrier protein	yes
bcd	Branched-chain alpha keto acid dehydrogenase	yes
bkdA1	Branched-chain alpha keto acid dehydrogenase E1 alpha subunit	yes
bkdA2	Branched-chain alpha keto acid dehydrogenase E1 beta subunit	yes
bkdB	Branched-chain alpha keto acid dehydrogenase E2 subunit	yes

bkdR	Transcriptional activator control of BCKD	yes
fabH1	B-ketoacyl-ACP synthase III	yes
fabH2	B-ketoacyl-ACP synthase III	yes
ydcB	Similar to holo-acyl-carrier protein synthase	yes

The proteins involved in elongation of fatty acid biosynthesis are *B*-ketoacyl-ACP synthase II (*KAS II*), *B*-ketoacyl-ACP reductase (*KAR*), *B*-hydroxyacyl-ACP dehydrase (*HAD*), and enoyl reductase (*ER*) (24). *Bacillus stearothermophilus* and *Bacillus subtilis* have these four proteins.

Figure 20-Elongation of Fatty Acid Biosynthesis (24)



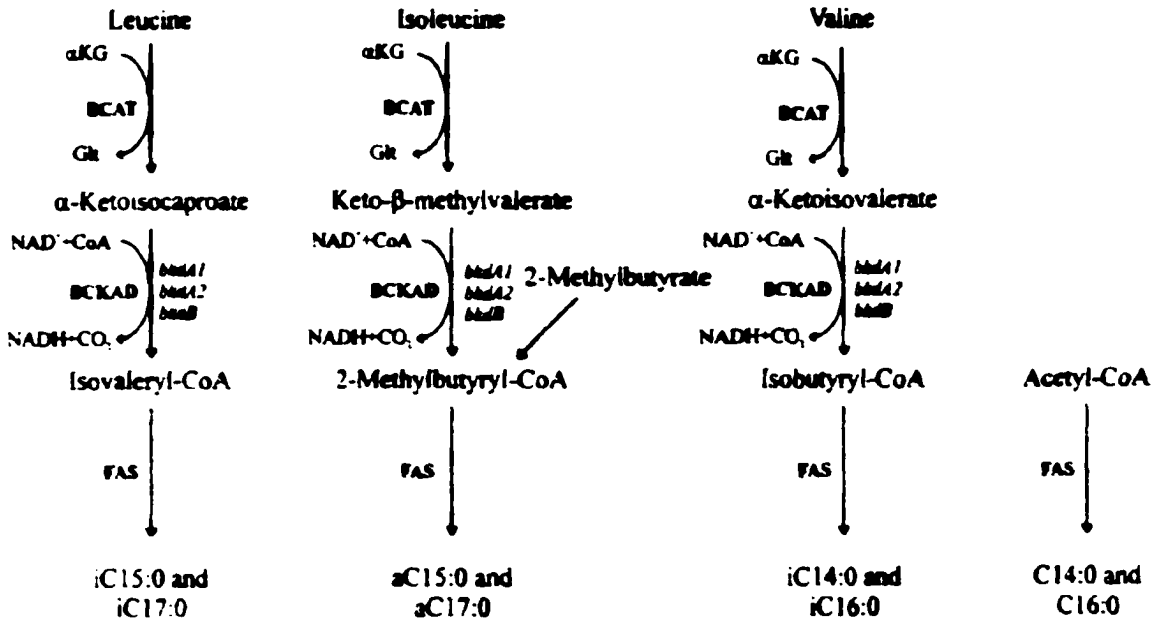
Reprinted with permission from the American Society for Microbiology coypyright 2002 by ASM Press.

Table 20- Elongation of Fatty Acid Biosynthesis Genes

Bsu genes	Function	Bst
fabD	Malonyl-CoA:ACP transacylase	yes
fabF	B-ketoacyl-ACP synthase II	yes
fabG	B-ketoacyl-ACP reductase	yes
yibW	Similar to enoyl-ACP reductase	yes
yidA	Similar to B-ketoacyl-ACP reductase	yes

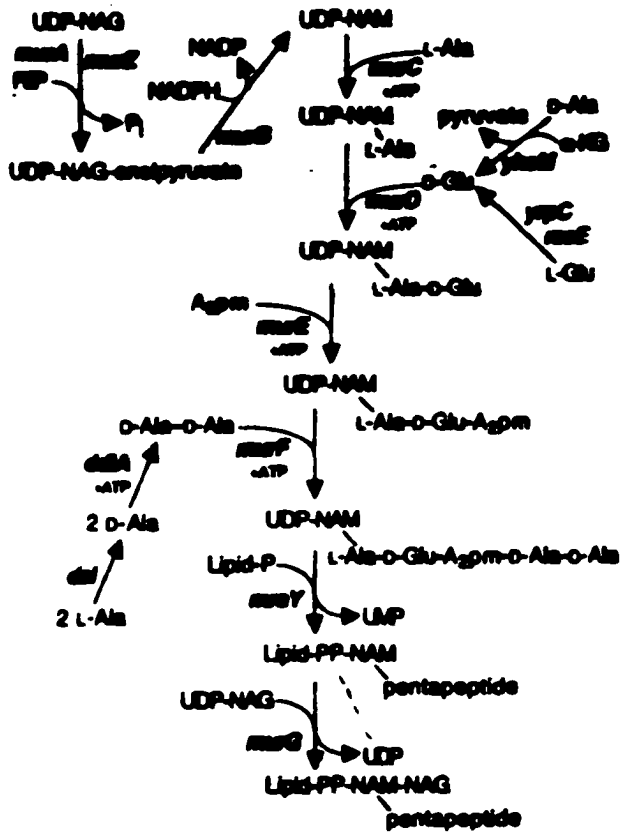
Branched-chain amino acids are converted to branched-chain 2-keto acid by a branched-chain amino acid transaminase (*BCAT*). Acyl-CoA intermediates are then synthesized by a branched-chain alpha-ketoacid dehydrogenase complex (*BCKAD*). With fatty acid synthase (*FAS*), the following fatty acids are produced: isovaleryl-CoA produces iso-branched C15:0 and C17:0; 2-methylbutyryl-CoA produces anteiso-branched C15:0 and C17:0; isobutyryl-CoA produces iso-branched C14:0 and C16:0, and acetyl-CoA produces C14:0 and C16:0 (19). The major components of *Bacillus subtilis* are anteiso-C15:0 and iso-C15:0 (19). The major components present in *Bacillus stearothermophilus* 22 were iso-15:0, iso-16:0, iso-17:0 since they accounted for 61.7-86.8 percent of the total fatty acids (4).

Figure 21-Proposed Pathway of Branched-Chain Keto Acid Incorporation into Fatty Acids (24)



Peptidoglycan biosynthesis genes in *Bacillus subtilis* were conserved in *Bacillus stearothermophilus*. Peptidoglycans consist of long glycan strands cross-linked via peptide side chains. The glycan strands consist of repeating disaccharide residues of N-acetyl glucosamine and N-acetylmuramic acid. The chain length in *Bacillus subtilis* 168 has been described as approximately 100 dischardes (24).

Figure 22-Peptidoglycan Precursor Synthesis (24)



Reprinted with permission from the American Society for Microbiology coypyright 2002 by ASM Press.

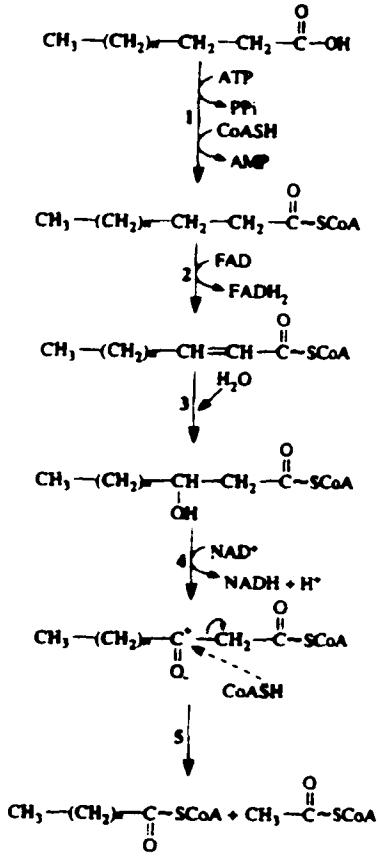
Note: If more than one gene is listed, they may have redundant functions or it is unknown which gene is required.
NAG-N-acetylglucosamine
NAM-N-acetylmuramic acid

Table 21-Peptidoglycan Precursor Synthesis Genes

Bsu gene	Function	Bst
dal (alr)	Alanine racemase	yes
ddlA	D-Alanyl-D-alanine ligase	yes
dat	D-Alanine aminotransaminase	yes
racE	Glutamate racemase	yes
yrpC	Glutamate racemase	yes
mraY	Phospho-*N*-acetylmuramoyl-pentapeptide translocase	yes
murB	UDP-*N*-acetylpyruvoylglucosamine reductase	yes
murC	UDP-*N*-acetylmuramate-alanine ligase	yes
murD	UDP-*N*-acetylmuramoylalanine-D-glutamate ligase	yes
murE	UDP-*N*-acetylmuramoylalanyl-D-glutamate-2,6-diaminopimelate ligase	yes
murF	UDP-*N*-acetylmuramoylalanyl-D-glutamyl-2,6-diaminopimelate-D-alanyl ligase	yes
murG	UDP-*N*-acetylglycosamine-N-acetylmuramyl-(pentapeptide) pyrophosphoryl-undecaprenol N-acetylglycosamine transferase	yes

The *B*-oxidation pathway (fatty acid catabolism) is complete in both *Bacillus stearothermophilus* and *Bacillus subtilis*. If the fatty acid has an even number of C atoms, then the entire chain is degraded to acetyl-CoA. If an odd-chain fatty acid, then the last fragment is propionyl-CoA. The five proteins involved in *B*-oxidation are acyl-CoA synthetase, fatty acyl-CoA dehydrogenase or butyryl-CoA dehydrogenase, enoyl-CoA hydratase, L-3-hydroxyacyl-CoA dehydrogenase, and *B*-ketothiolase, which is also referred to as acetyl-CoA C-acetyltransferase. *Bacillus stearothermophilus* calls the last enzyme *B*-ketothiolase. *Bacillus subtilis* calls the last enzyme acetyl-CoA C-acetyltransferase.

Figure 23-*Beta* Oxidation of Fatty Acids (105)



4. Nucleotide Metabolism

Purine Biosynthesis

Purines are synthesized stepwise from 5-phosphoribosylpyrophosphate (PRPP) which is synthesized during pyrimidine biosynthesis. The genes for all the enzymes for the purine biosynthetic pathway were found in the *Bacillus stearothermophilus* genome. This pathway as shown in figure 26 also is complete in *Bacillus subtilis*. The following proteins are involved in this pathway: glutamine PRPP amidotranferase (*purF*), phosphoribosylglysineamide synthetase (*purD*), phosphoribosylglycineamide formyltransferase (*purN*), GAR transformylase T (*purT*), phosphoribosylformylglycineamide synthetase (*purQ*, *purL*, *purS*), phosphoribosyl-aminoimidazole (AIR) synthetase (*purM*), phosphoribosylaminoimidazole (NCAIR) synthetase (*purK*), NCAIR mutase (*purE*), phosphoribosyl-aminoimidazole succinocarboxamide (SAICAR) synthetase (*purC*), adenylosuccinate lyase (*purB*), phosphoribosylaminoimidazolecarboxamide formyltransferase (*purH*), and IMP cyclohydrolase (*purH*) (24, 105).

Two intermediates in purine biosynthesis provide links to other biosynthetic pathways. The product of step five, 5'phosphoribosyl-5-aminoimidazole is used in thiamine synthesis. The product of step nine, 5'phosphoribosyl-4-carboxamide-5-formamidoimidazole is a product of histidine biosynthesis (24).

To form adenylate (6 aminopurine nucleotide) from IMP, the amino group is transferred from aspartate using GTP as an energy source and the following enzymes adenylosuccinate synthetase, adenylosuccinate lyase, and adenylate kinase. To form guanylate from IMP, then IMP is oxidized by DPN to xanthylate, which is then converted to GMP using ammonia in an amination reaction. The following enzymes are needed for this pathway: IMP dehydrogenase, GMP synthetase, guanylate kinase, and nucleoside diphosphokinase (24, 105).

Figure 24-Biosynthesis of Purine Nucleotide (105)

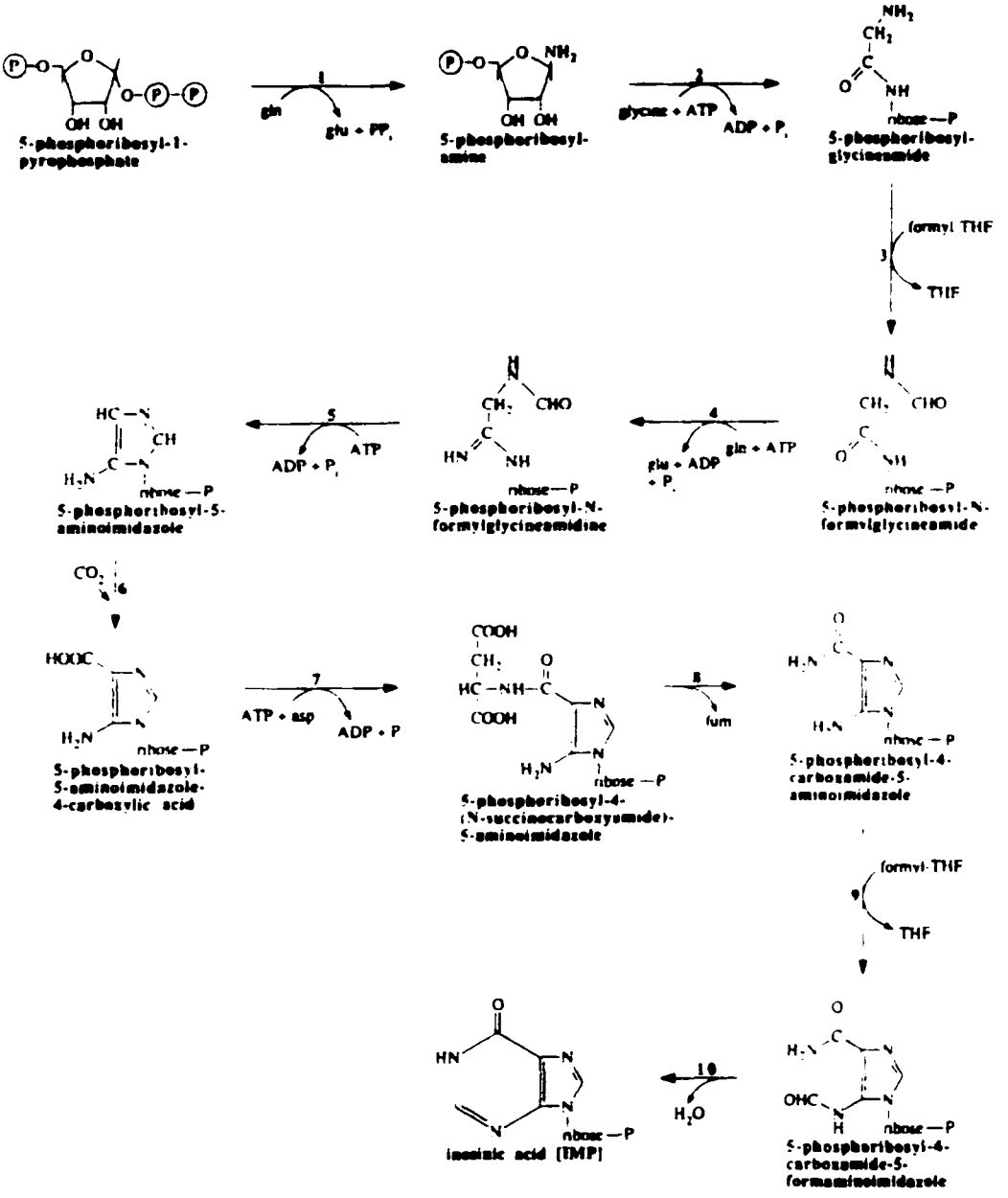
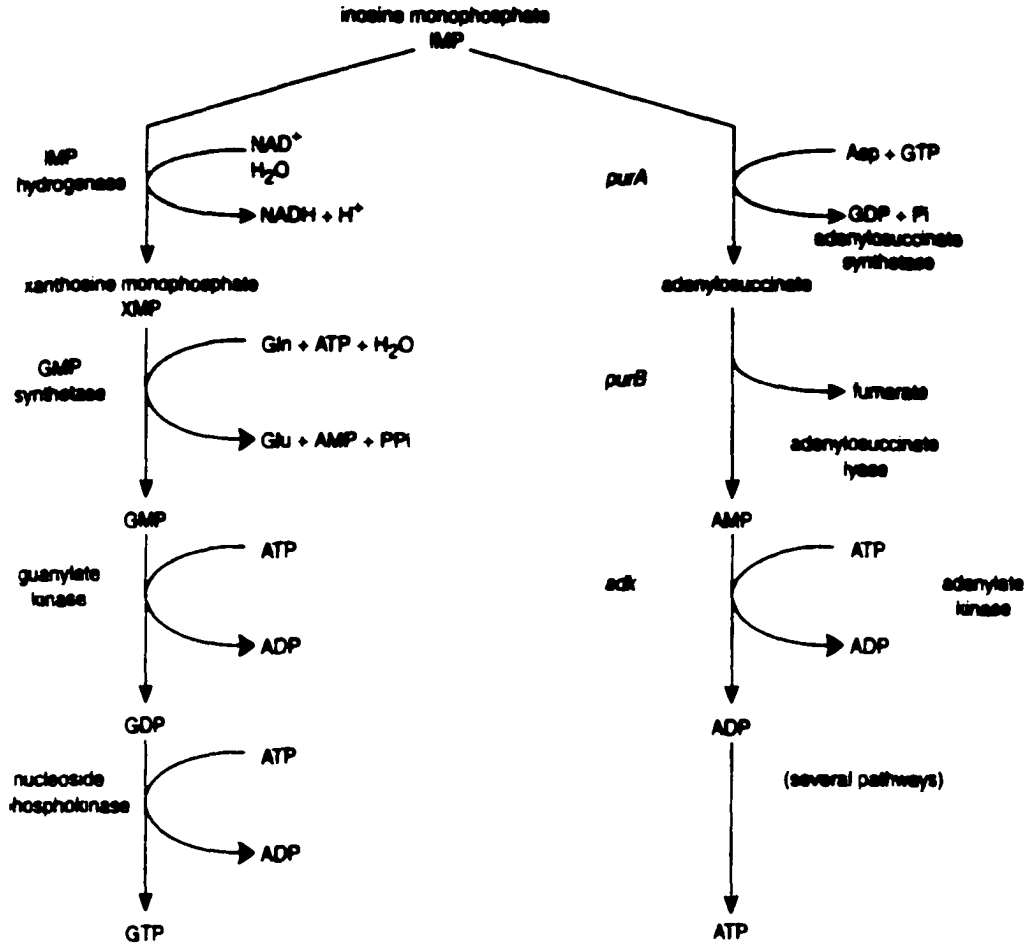


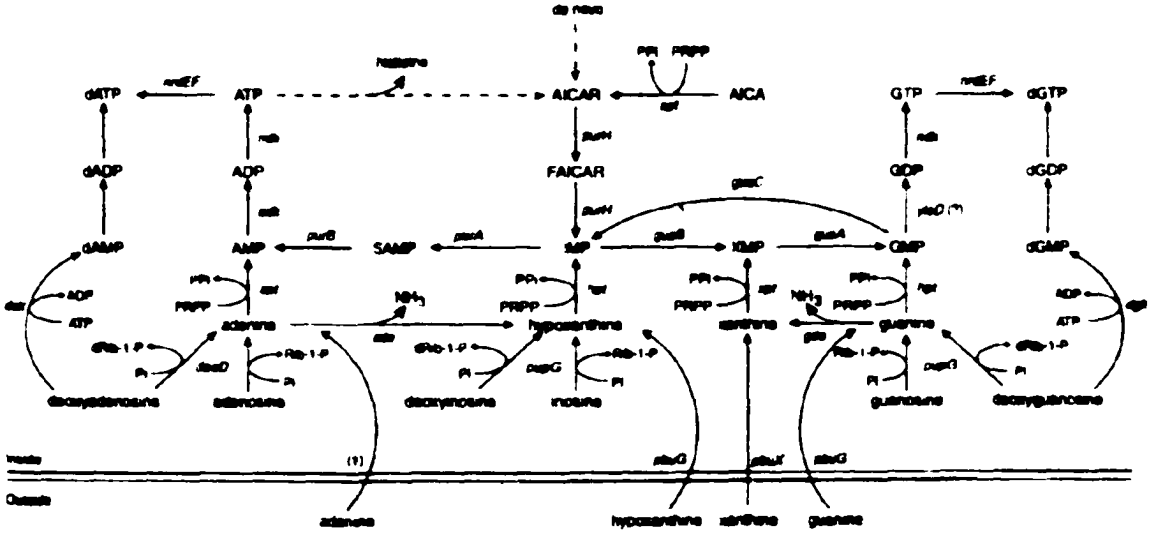
Figure 25-Synthesis of GMP and AMP (24)



Purine Salvage and Interconversion

Purine bases and nucleosides, resulting intracellularly from nucleic acid degradation or extracellularly from decaying cells or secretion from other organisms, are converted into nucleotides. The nitrogen and carbon sources for this interconversion result from the formation of pentose phosphates from the cleavage of nucleosides and formation of ammonia from the deamination of purine bases. After transport into the cell by uptake systems, purine bases are converted to nucleoside monophosphates by specific phosphoribosyltransferases (24) Except for lacking the *pupG* gene, the *Bacillus stearothermophilus* genome encodes the same purine salvage and interconversion genes as *Bacillus subtilis*

Figure 26-Purine Salvage and Interconversion (24)



Reprinted with permission from the American Society for Microbiology coypyright 2002 by ASM Press.

Table 22-Purine Salvage and Interconversion

Gene	Protein	Bsu	Bst
purB	adenylosuccinate lyase	yes	yes
purH	phosphoribosylaminoimidazolecarboxamide formyltransferase and IMP cyclohydrolas	yes	yes
purA	adenylosuccinate synthetase	yes	yes
guaB	IMP dehydrogenase	yes	yes
guaA	GMP synthetase	yes	yes

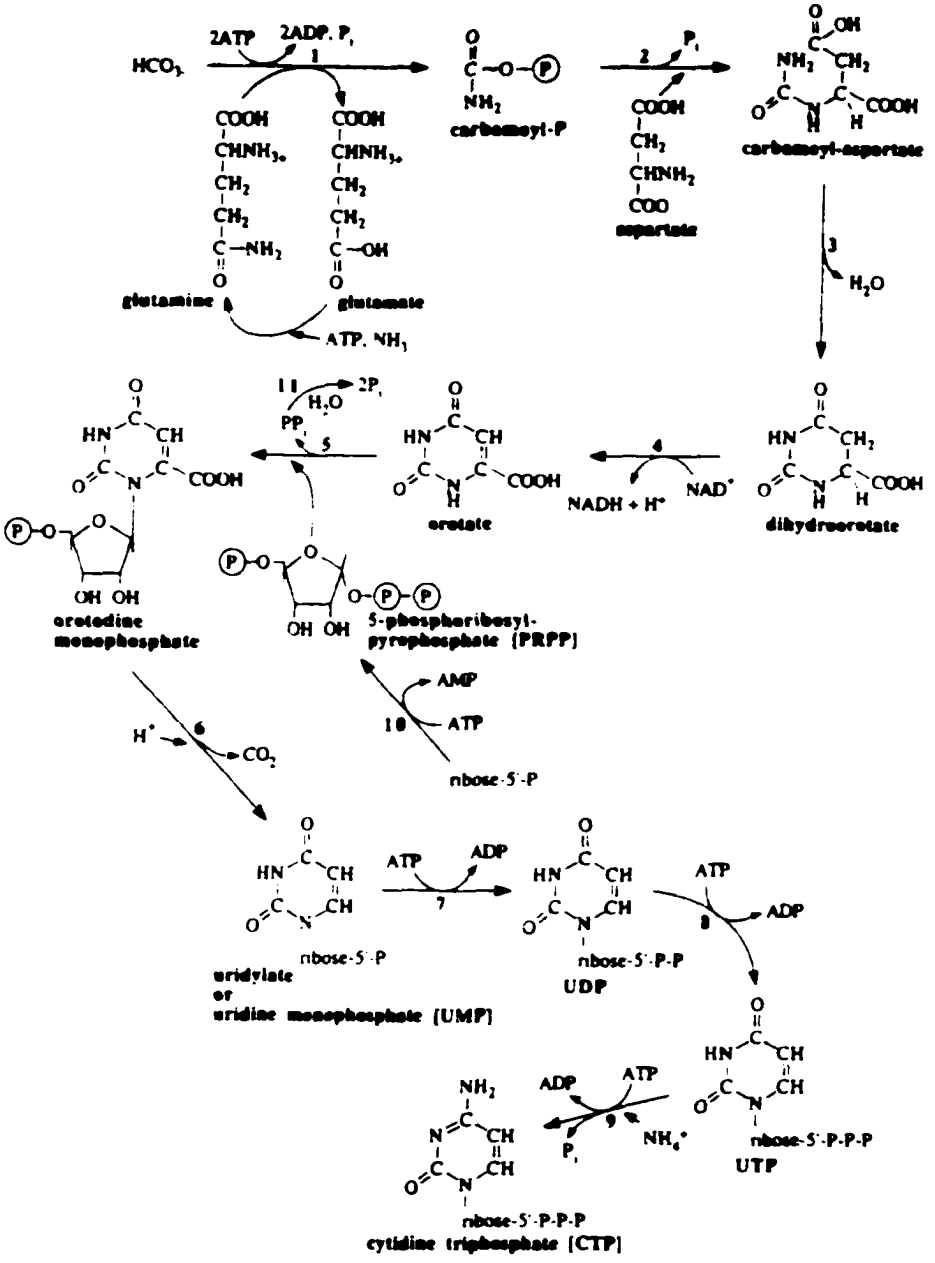
apt	adenine phosphoribosyltransferase	yes	yes
hpt	hypoxanthine-guanine phosphoribosyltransferase	yes	yes
xpt	xanthine phosphoribosyltransferase	yes	yes
guaC	GMP reductase	yes	yes
deoD	adenosine/deoxyadenosine phosphorylase	yes	yes
pupG	guanosine/deoxyguanosine-inosine/ deoxyinosine phosphorylase	yes	no
dck	deoxycytidine-deoxyadenosine kinase	yes	yes
dgk	deoxyguanosine kinase	yes	yes
ade	adenine deaminase	yes	yes
pbuG	hypoxanthine-guanine permease	yes	yes
adk	adenylate kinase	yes	yes
ndk	nucleoside diphosphokinase	yes	yes
nrdEF	ribonucleotide reductase	yes	yes
yloD	putative guanylate kinase	yes	yes
pbuX	xanthine permease	yes	yes

Pyrimidine Biosynthesis

Pyrimidine biosynthesis, which takes place in the cytosol, starts at glutamate metabolism with bicarbonate and glutamine as precursors and glutamate as an end product in the first reaction. Both *Bacillus stearothermophilus* and *Bacillus subtilis* have the complete pyrimidine biosynthetic pathway involving the following proteins carbamoyl phosphate synthetase (*pyrAA, pyrAB*), aspartate transcarbamoylase (*pyrB*), dihydroorotase (*pyrC*), dihydroorotate dehydrogenase (*pyrDI, pyrDII*), orotate phosphoribosyl transferase (*pyrE*), orotidine decarboxylase, cytidylate kinase (*smbA pyrH*), nucleoside diphosphate kinase (*ndk*), and CTP synthetase (*pyrG*) (24, 105).

To form dTTP from CTP, the pathway is CTP-CDP-dCDP-dCMP-dUMP-dTMP-dTTP (105).

Figure 27-Biosynthesis of Pyrimidine Nucleotides (105)



5. Vitamin Metabolism

Cobalamin Biosynthesis

The cobalamin (vitamin B12) is involved in rearrangement reactions. This operon in the Gram-positive *Bacillus megaterium* contains fourteen genes. This aerobic organism synthesizes cobalamin anaerobically adding cobalt at an early stage. Cobalamins, haem, chlorophyll, sirohaem, and coenzyme F430 all have the same precursor, uroporphyrinogen III. The two main coenzyme forms of vitamin B12 are methylcobalamin and adenosylcobalamin. The molecular genetics of cobalamin biosynthesis have been studied only in *Salmonella typhimurium* and *Pseudomonas denitrificans* prior to the work in *Bacillus megaterium* (109)

Bacillus stearothermophilus has seventeen genes involved in cobalamin biosynthesis to include the following proteins which are part of the aerobic pathway: precorrin methylase, precorrin-2 methylase, precorrin-3 methylase, precorrin-4 methylase, precorrin-6Y methylase, cobyrinic acid A,C-diamide synthase, *cobI*, *cobP*, *cobQ*, and *cobW*. cbiB, *cbiM*, *cbiN*, *cbiO* and *cbiQ* were also identified in the *Bacillus stearothermophilus* database. This pathway is not complete in *Bacillus stearothermphilus*.

When the *Bacillus subtilis* genomic database is queried for *cbi* or *cob* or cobalamin only one gene, uroporphyrin III C methyltransferase, appears. *Bacillus subtilis* 168 does not synthesize this vitamin, nor does it have homologs of any of the 30 genes involved in cobalamin biosynthesis.

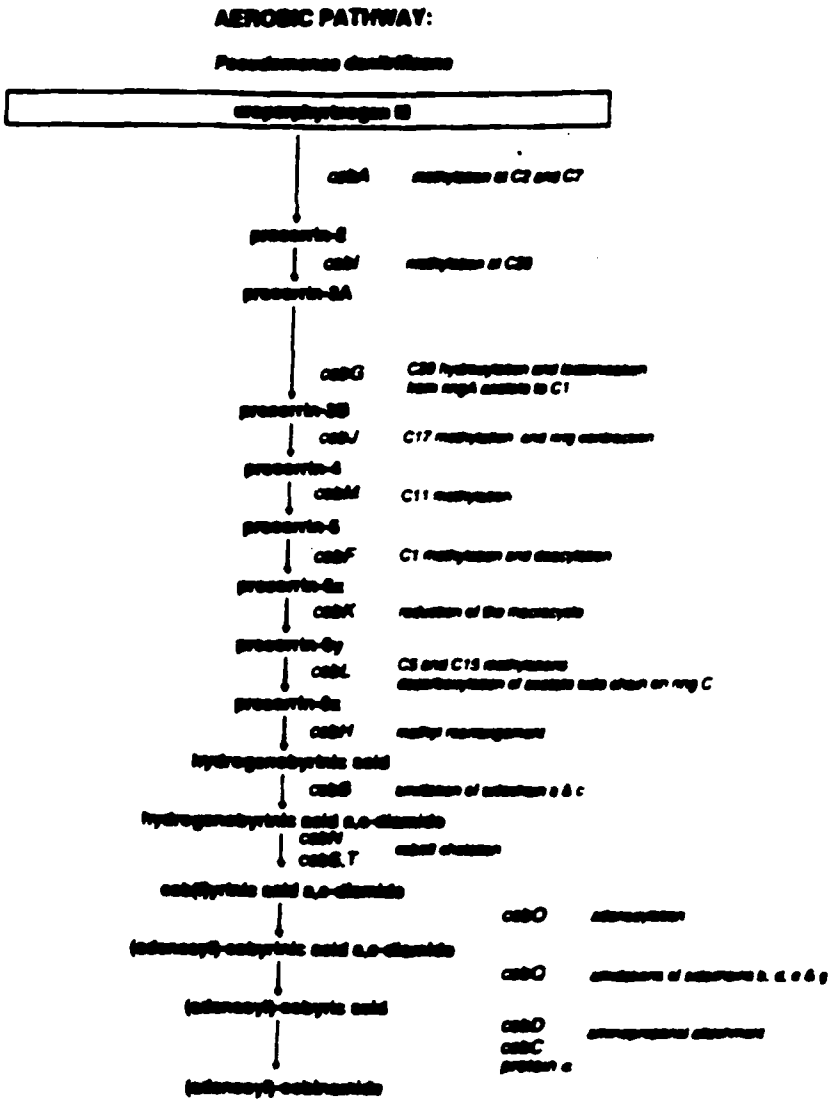
Bacillus megaterium is such a good producer of vitamin B12 that this organism has been used for its industrial production. The ORFs in *Bacillus megaterium* are assigned the prefix *cbi* named after *Salmonella typhimurium* since the genes were more similar to those than the genes in *Pseudomonas denitrificans* (*cob* prefix). Besides nomenclature, there are other differences between how these two Gram-negative organisms synthesize cobalamin. *P. denitrificans* does not contain equivalents of *cbiD*,

cbiG, and *cbiK*. *S. typhimurium* does not contain *cobE, F, G, N, S, T,* and *W*. This reflects the distinct synthesis pathways of the corrin component of cobalamin between these two organisms. *P. denitrificans* uses an aerobic pathway, whereas *S. typhimurium and B. megaterium* use an anaerobic pathway. The pathways also differ in the timing of the cobalt insertion. *P. denitrificans* inserts cobalt later in the pathway than *S. typhimurium* (109).

Of the fourteen genes in the *Bacillus megaterium* operon, the following five are transmethylases: *cysG, cbiE, L, F,* and *H*. Three genes from *Bacillus megaterium* were not found in either of the Gram-negative organisms. *cbiX* was deemed absolutely essential for cobyric acid synthesis in *Bacillus megaterium* (109).

The gene organization of the cobalamin operon in *Bacillus megaterium* is as follows: *cbiWYXJCDETLFGA-cysG-cbiY-btuR*-ORF1. The gene organization of the cobalamin operon in *S. typhimurium* is as follows: *cbiABCDETFGHJKLMNQOP-cobUST*. The genes for cobalamin biosynthesis in *P. denitrificans* appear in four separate locations along the chromosome. The following genes were not found in *P. denitrificans: cbiX, cbiD, cbiG, cbiD,* and *cbiK*. The following five genes are found in *P. denitrificans, S. typhimurium,* and *Bacillus megaterium: btuR, cbiP,* and *cobUST* (109).

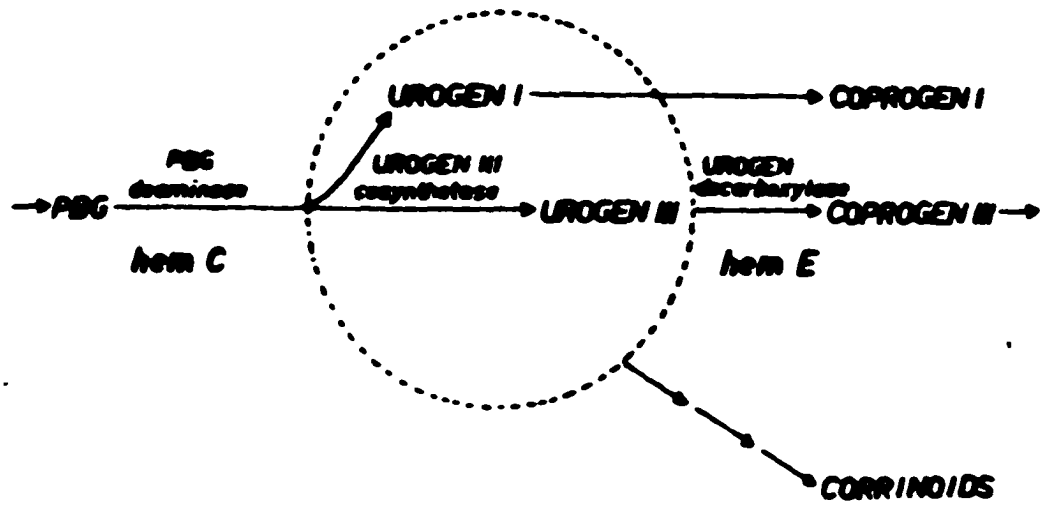
Figure 28-Cobalamin Biosynthesis (109)



Biosynthesis of Uroporphyrinogen III

Cobalamins, haem, chlorophyll, siroheme, and coenzyme F430 all have the same precursor, uroporphyrinogen III, which is formed from porphobilinogen. This formation in *Bacillus subtilis* involves porphobilinogen deaminase (*hemC*) and uroporphyrinogen III cosynthase (*hemD*) (110). *Bacillus stearothermophilus* has both genes, in addition to coproporphyrinogen III oxidase (*hemF*), uroporphyrinogin decarboxylase (*hemE*), and delta-aminolaevulinic acid dehydratase (*hemB*). This pathway is complete in both *Bacillus subtilis* and *Bacillus stearothermophilus*.

Figure 29-Uroporphyrinogen III Biosynthesis (110)

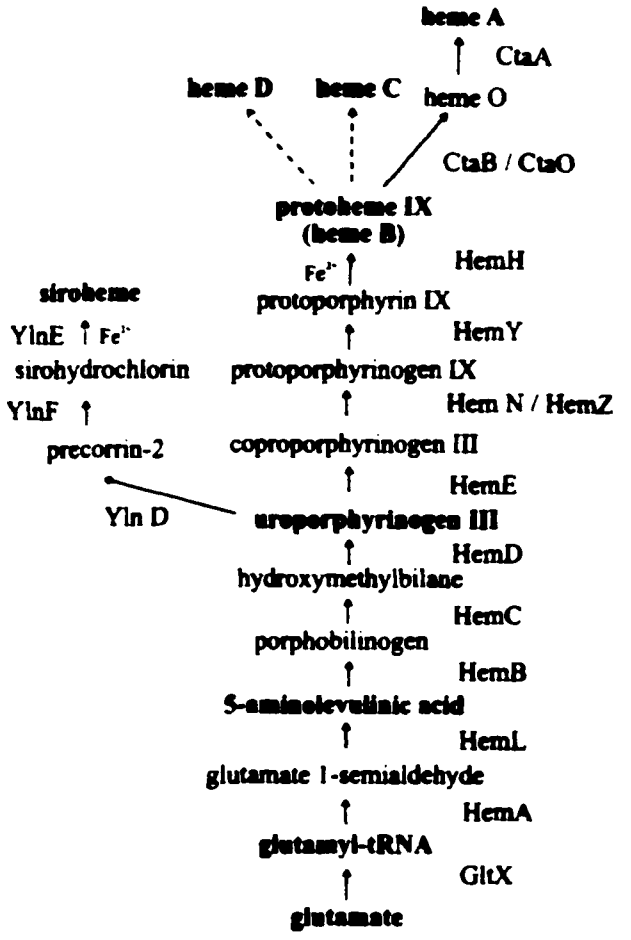


Reprinted with permission from American Society for Microbiology coypyright 1977 by ASM Press

Heme Biosynthesis

Heme is a vital prosthetic group of cytochromes and proteins that function in metabolism, respiration, detoxification and sensory. *Bacillus subtilis* has twenty-three different heme-containing proteins including membrane-bound cytochromes, water-soluble cytochrome P450 enzymes, and catalases (24). The cytochromes contain heme in their reaction centers. The basic structure of heme is an iron tetrapyrrole ring. The precursor in heme biosynthesis is glutamate. Eight molecules of 5-aminolevulinic acid are used to make one molecule of uroporphyrinogen III. Both *Bacillus stearothermophilus* and *Bacillus subtilis* encode the following genes for the heme pathway: *gltX* encodes glutamyl-tRNA synthetase, *hemA* encodes glutamyl-tRNA reductase and needs NADPH for this reaction, *hemL* encodes glutamate-1 semialdehyde 2,1-aminomutase, *hemB* encodes delta-aminolevulinic acid dehydratase, *hemC* encodes porphobilinogen deaminase, *hemD* encodes uroporphyrinogen III cosynthase, *hemE* encodes uroporphyrinogen decarboxylase, *hemN* encodes coproporphyrinogen III oxidase, *hemY* encodes protoporphyrinogen IX oxidase, *hemH* encodes ferrochelatase, *ctaB* encodes heme O synthetase, and *ctaA* encodes heme O oxygenase. Siroheme is a cofactor for nitrite and sulfite reductases. The latter is used in cysteine biosynthesis (24).

Figure 30-Heme Biosynthesis (24)

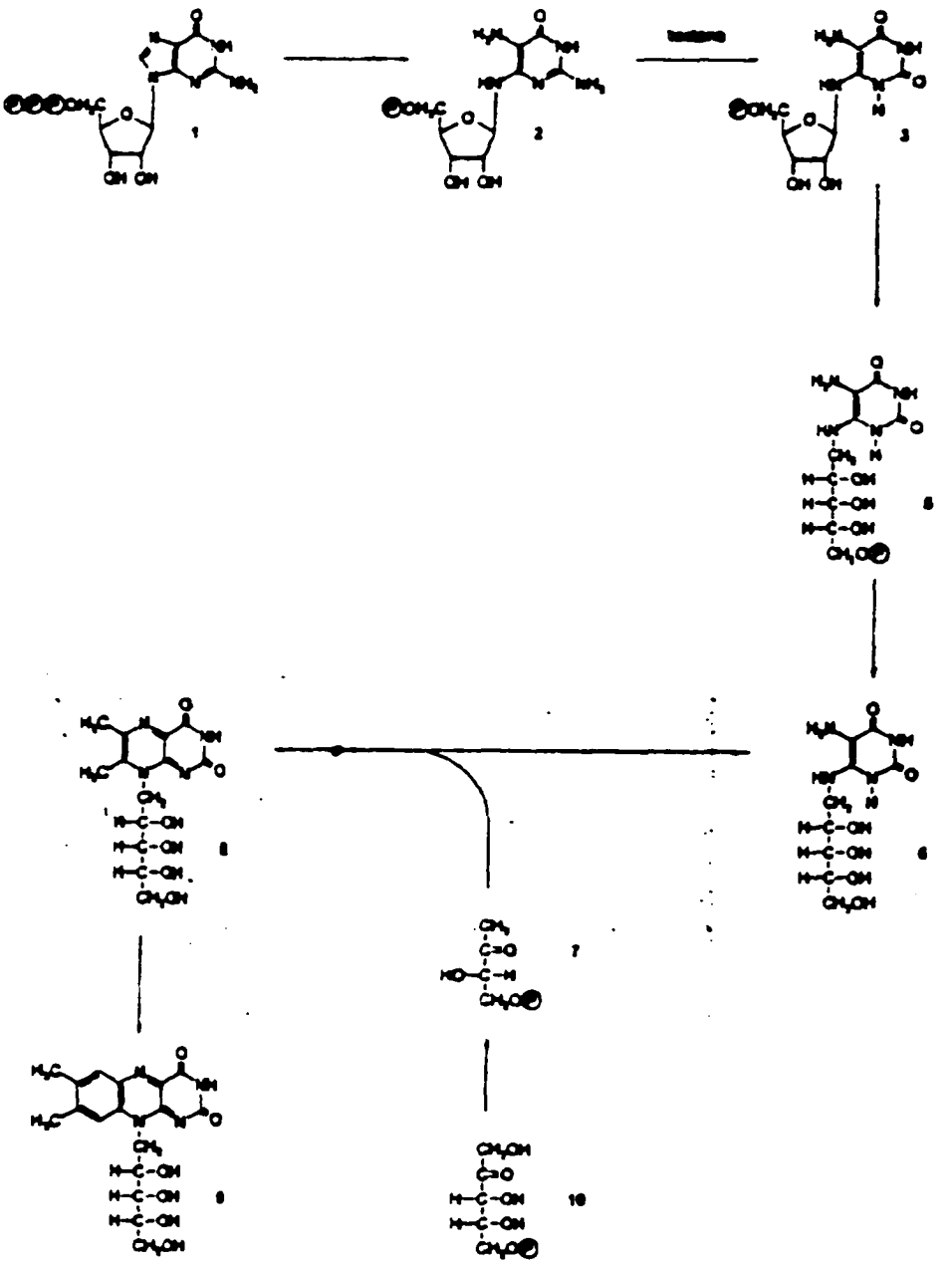


Riboflavin (Vitamin B-2) Biosynthesis

Riboflavin is a precursor of flavin mononucleotide (FMN) and flavin adenine dinucleotide (FAD), which are the prosthetic groups of flavoproteins. The riboflavin operon in *Bacillus subtilis* has the following gene order: *ribGBAH*. Guanosine triphosphate (GTP) is the precursor of riboflavin. In *Bacillus subtilis*, two of the genes in this operon encode bifunctional enzymes. The *ribA* gene encodes both 3,4-dihydroxy-2-butanone-4-phosphate synthase, located at the N terminus, and GTP cyclohydrolase II at the C terminus. The *ribG* gene encodes both pyrimidine deaminase and pyrimidine reductase. Both the deaminase and reductase require magnesium. The reductase requires NADPH or NADH as a cofactor (24).

Both *Bacillus stearothermophilus* and *Bacillus subtilis* have complete pathway for riboflavin biosynthesis. This pathway involves the following genes: from step 1 to step 2 *ribA* encodes GTP cyclohydrolase II, from step 2 to step 3 *ribG* encodes riboflavin-specific deaminase, from step 3 to step 5 *ribG* encodes riboflavin-specific reductase, from step 5 to step 6, no gene is identified, from step 10 to step 7 *ribA* encodes 3,4-dihydroxy-2-butanone-4-phosphate synthase, from step 6 to step 8 *ribH* encodes lumazine synthase, and from step 8 to step 9 *ribB* encodes riboflavin synthase (111)

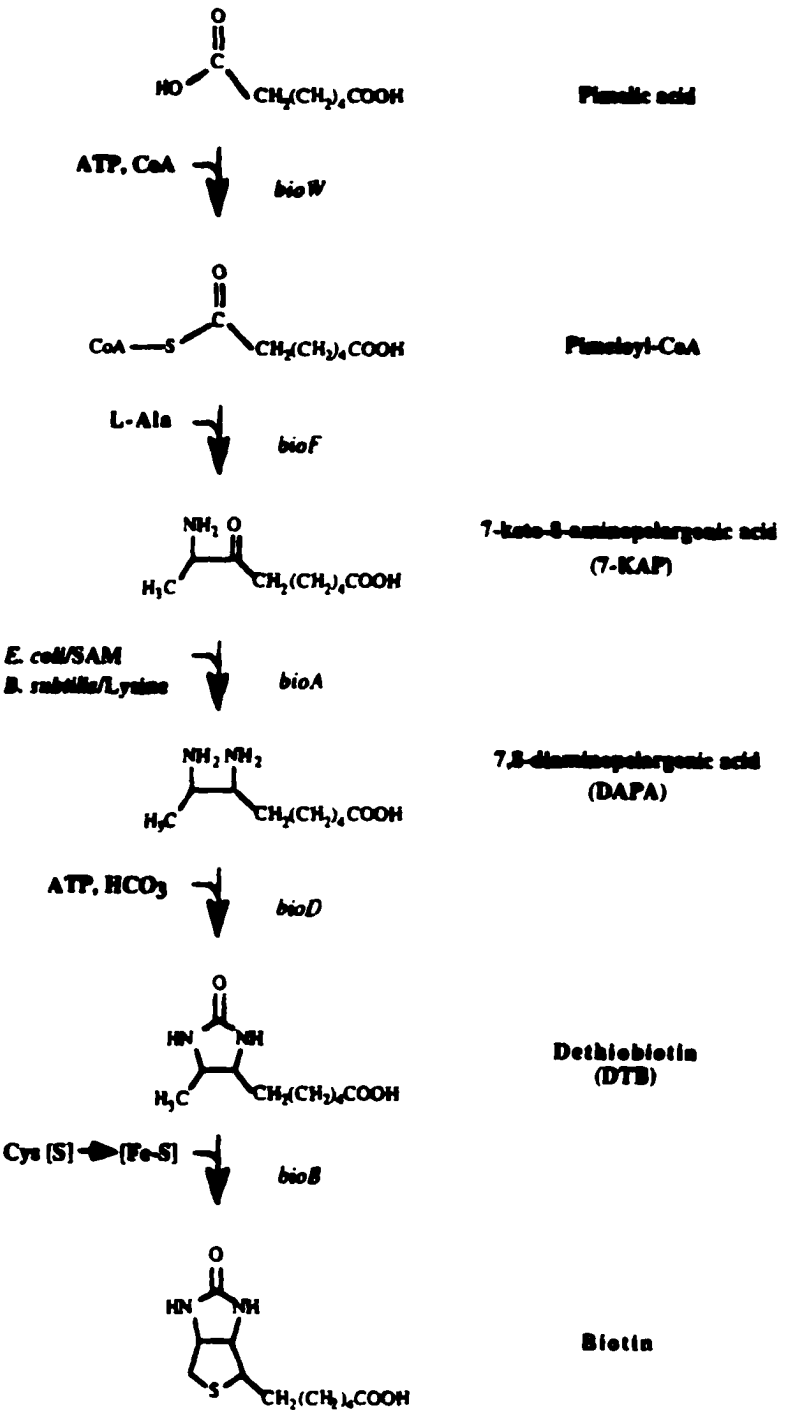
Figure 31-Riboflavin Biosynthesis (111)



Biotin (Vitamin H) Biosynthesis

Biotin is a required cofactor for carboxylation reactions such as for use in forming acetyl-CoA. The seven biotin synthetic genes of *Bacillus subtilis* are in an operon with the following gene order: *bioWAFDBI* and *ytbQ*, with unknown function. The genes for the complete pathway present in *Bacillus subtilis* also are observed in *Bacillus stearothermophilus*. These genes include the following: the *bioW* gene encodes pimeloyl-CoA synthase, the *bioF* gene encodes 7-KAP synthase, the *bioA* gene encodes DAPA Aminotransferase, the *bioD* gene encodes DTB synthetase, the *bioB* gene encodes biotin synthase and the *birA* gene encodes the biotin operon repressor (24).

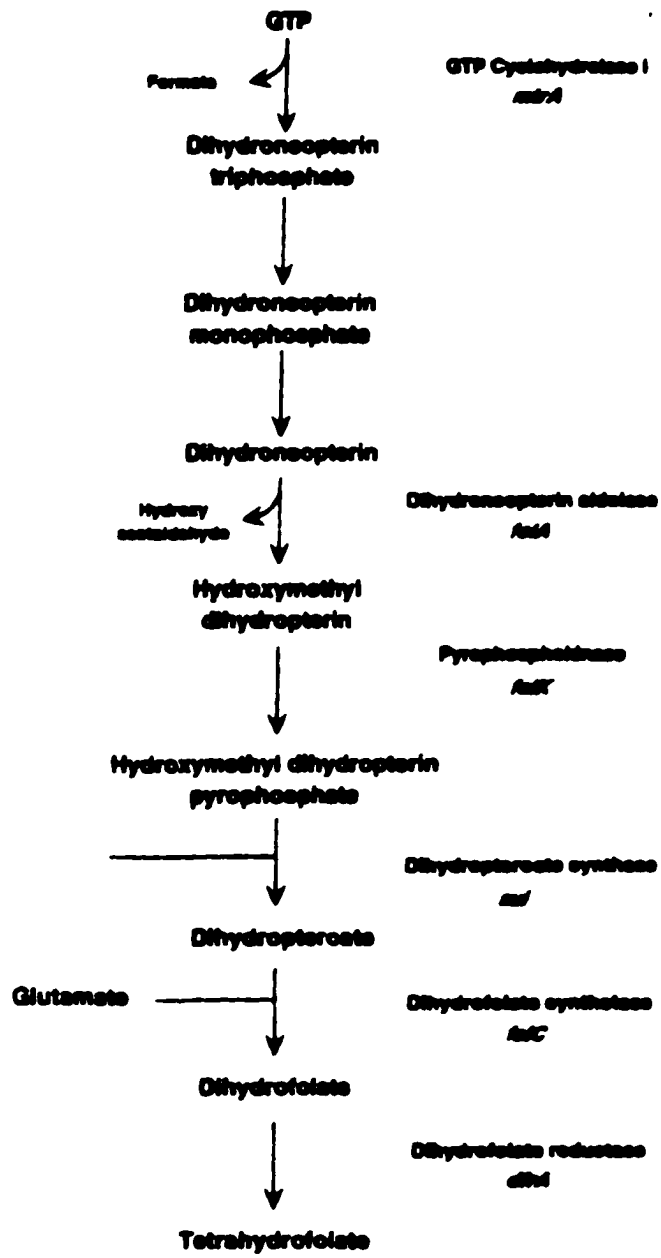
Figure 32-Biotin Biosynthesis (24)



Folic Acid Biosynthesis

The precursor of folic acid is GTP. Tetrahydrofolate, a folic acid derivative, functions as a coenzyme in numerous one carbon atom transfer reactions. P-aminobenzoic acid is a precursor of tetrahydrofolic acid. The genes for folic acid and tetrahydrofolate biosynthesis are encoded in the *Bacillus stearothermophilus* genome. The genes involved in folate biosynthesis are as follows: *mtrA* encodes GTP cyclohydrolase I, *folA* encodes dihydroneopterin aldolase, *folK* encodes pyrophosphokinase, *sul* encodes dihydropteroate synthase, *folC* encodes dihydrofolate synthetase, and *dfrA* encodes dihydrofolate reductase. *pabA* is also known as *trpG* since it functions in the biosynthesis of both folic acid and tryptophan. *pabA* combines with *pabB* to form p-aminobenzoate synthase (24).

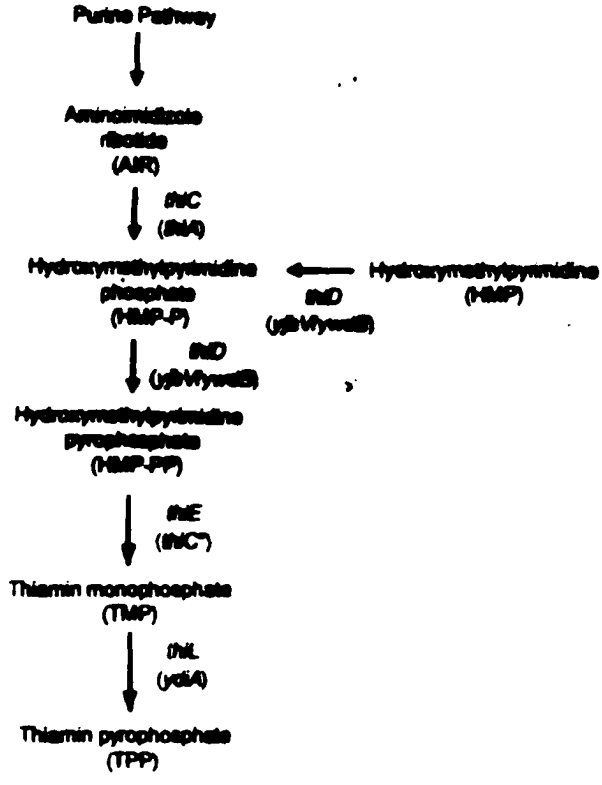
Figure 33-Folic Acid Biosynthesis (24)



Thiamin (Vitamin B1) Biosynthesis

The biologically active form of this vitamin, thiamin pyrophosphate (TPP), serves as a cofactor for pyruvate dehydrogenase, decarboxylase, pyruvate oxidase, and transketolase. All the genes involved in thiamine biosynthesis were found in the *Bacillus stearothermophilus* genome. The precursors for this pathway include derivatives from glycolysis, TPP, and the purine pathway. The proteins involved in this pathway include the following: 1-deoxy-D-xylulose synthase encoded by *yqiE*, the *thiSGF* operon, sarcosine oxidase encoded by *yjbR*, hydroxyethylthiazole kinase encoded by *thiK*, thiamin phosphate pyrophosphorylase encoded by *thiC*, possibly TMP kinase encoded by *ydiA*, a protein encoded by *thiA*, and a protein encoded by *thiI* (24).

Figure 34-Thiamin Biosynthesis (24)



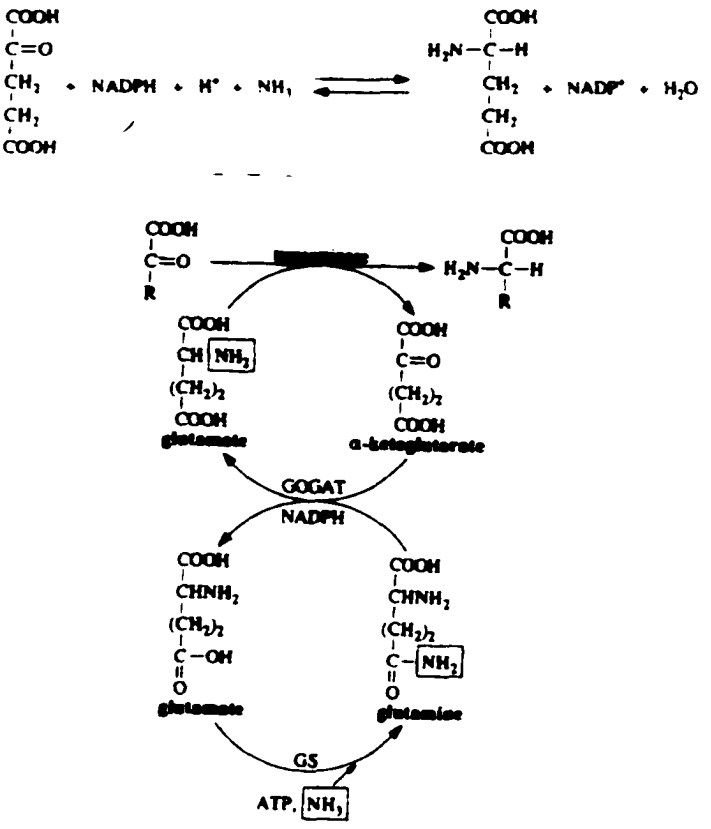
6. Amino Acid Metabolism

Glutamate and Glutamine Biosynthesis

Both the *Bacillus stearothermophilus* and *Bacillus subtilis* genomes encode the three proteins required for glutamate and glutamine biosynthesis. The biosynthesis of these amino acids begins when nitrogen-fixing microorganisms reduce nitrogen to ammonia, which assimilates into amino acids by way of glutamate and glutamine. The alpha amino group of most amino acids comes from the alpha amino group of glutamate by transamination. Glutamine contributes its side-chain nitrogen in the biosynthesis of purines, pyrimidines, amino sugars, histidine, tryptophan, asparagines, and NAD+ (34). Glutamate is synthesized by two routes. Glutamate is synthesized through the reductive transamination of *alpha*-ketoglutarate catalyzed by GOGAT, which is also called *gltA* and *gltB*. Glutamine is synthesized from glutamate with the help of ammonia using GS, which is also called *glnA* glutamine synthetase (105)

As mentioned earlier in the Methods section, the *Bacillus subtilis* genome is missing the glutaminyl tRNA synthetase gene, but has the remaining nineteen aminoacyl tRNA synthetases. *Bacillus halodurans* also lacks the glutaminyl-tRNA synthetase gene, one of two threonyl-tRNA synthetase genes, and two tyrosyl-tRNA synthetase genes (22)

Figure 35-Glutamate and Glutamine Biosynthesis (105)

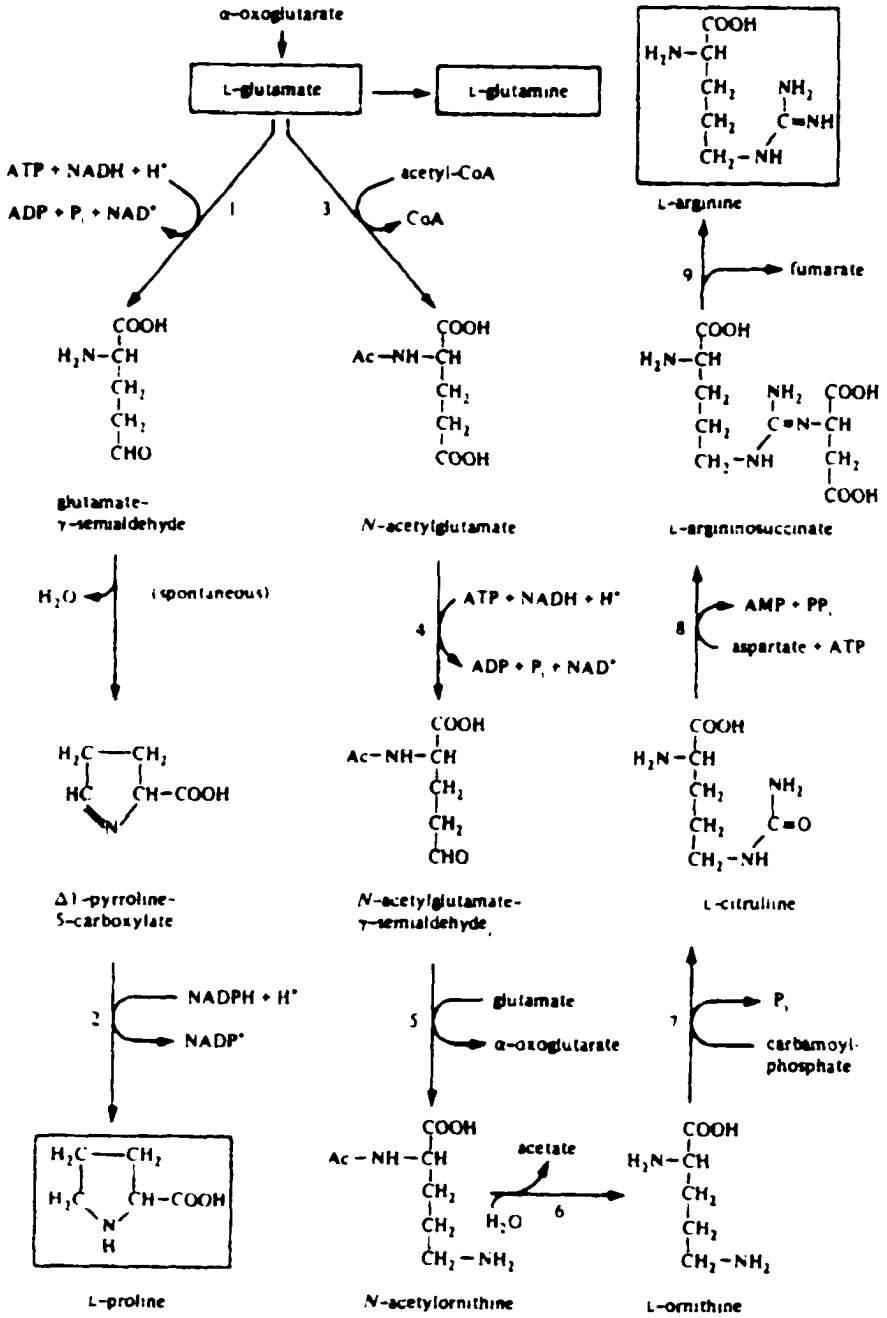


Proline and Arginine Biosynthesis

Both *Bacillus subtilis* and *Bacillus stearothermophilus* have complete pathways for proline biosynthesis. This pathway begins with L-glutamate and involves the following two proteins: 1-pyrroline-5-carboxylate dehydrogenase (*gamma*-glutamyl kinase and *gamma* glutamyl phosphate reductase) and pyrroline -5-carboxylate reductase. The biosynthesis of arginine starts from glutamate and involves the following proteins: ornithine N-acetyltransferase, N-acetylglutamate kinase, N-acetylglutamylphosphate reductase, N-acetylornithine *gamma*-aminotransferase, ornithine N-acetyltransferase, ornithine carbomoyltransferase, argininosuccinate synthetase, and argininosuccinase (112)

In both pyrimidine nucleotide biosynthesis and arginine biosynthesis, carbomoyl is used, yet the former takes place in the cytosol while the latter takes place in the mitochondria. Since bacteria cannot compartmentalize, they distinguish between using carbamoyltransferase in either pathway by regulating the components of the two pathways.

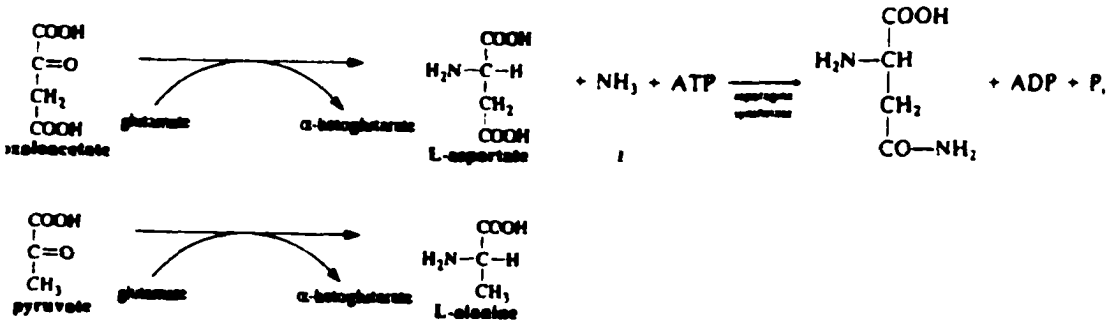
Figure 36-Proline and Arginine Biosynthesis (112)



Alanine, Aspartate, and Asparagine Biosynthesis

Transaminase is the protein that converts oxoloacetate into aspartate and pyruvate into alanine. Aspartate is converted into asparagine by way of asparagines synthetase (105, 112). Both transaminases are encoded in the *Bacillus stearothermophilus* and Bacillus subtilis genomes.

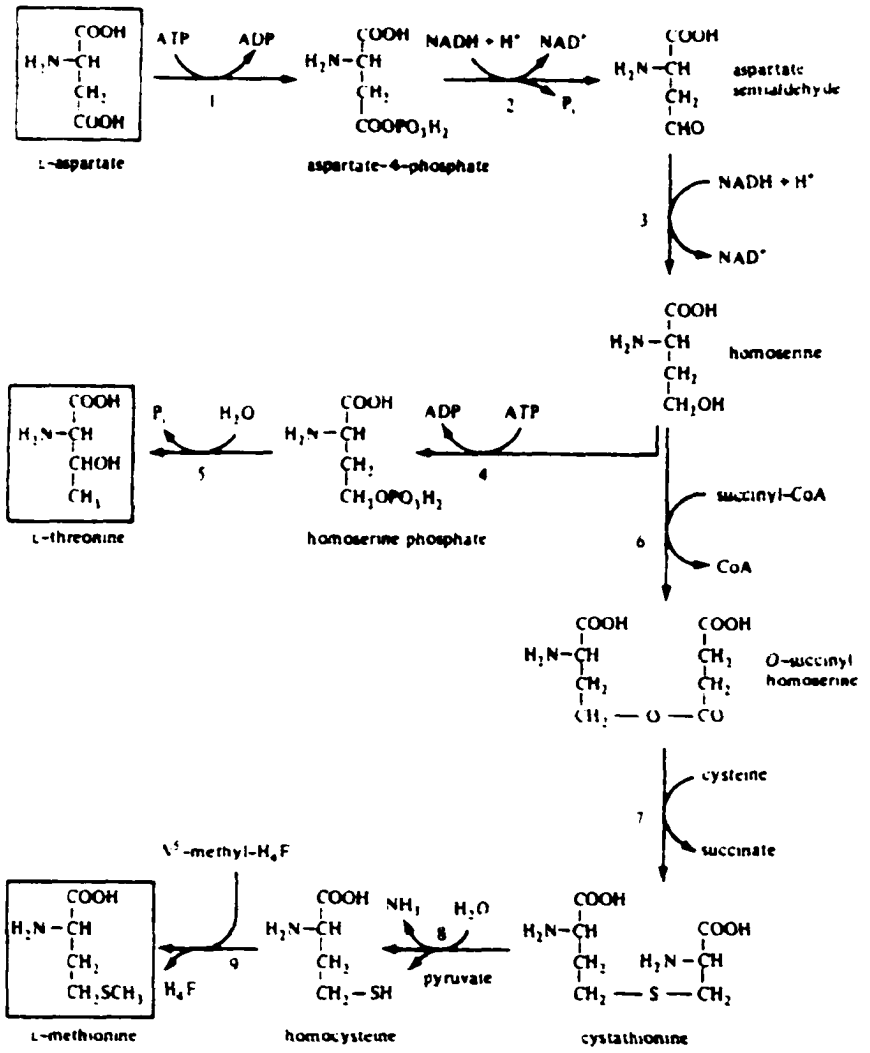
Figure 37-Alanine, Aspartate, and Asparagine Biosynthesis (105, 112)



Threonine and Methionine Biosynthesis

This biosynthesis starts from aspartate (precursor is oxaloacetate) and involves the following proteins: aspartate kinase, aspartate semialdehyde dehydrogenase, homoserine dehydrogenase, homoserine kinase, threonine synthase to produce threonine. To produce methionine, the first three enzymes from above are used in addition to homoserine O-succinyl transferase, cystathionine synthase, cystathionine lyase, and cobalamin-independent methionine synthase (112). This pathway is complete in both *Bacillus stearothermophilus* and *Bacillus subtilis*.

Figure 38-Threonine and Methionine Biosynthesis (112)



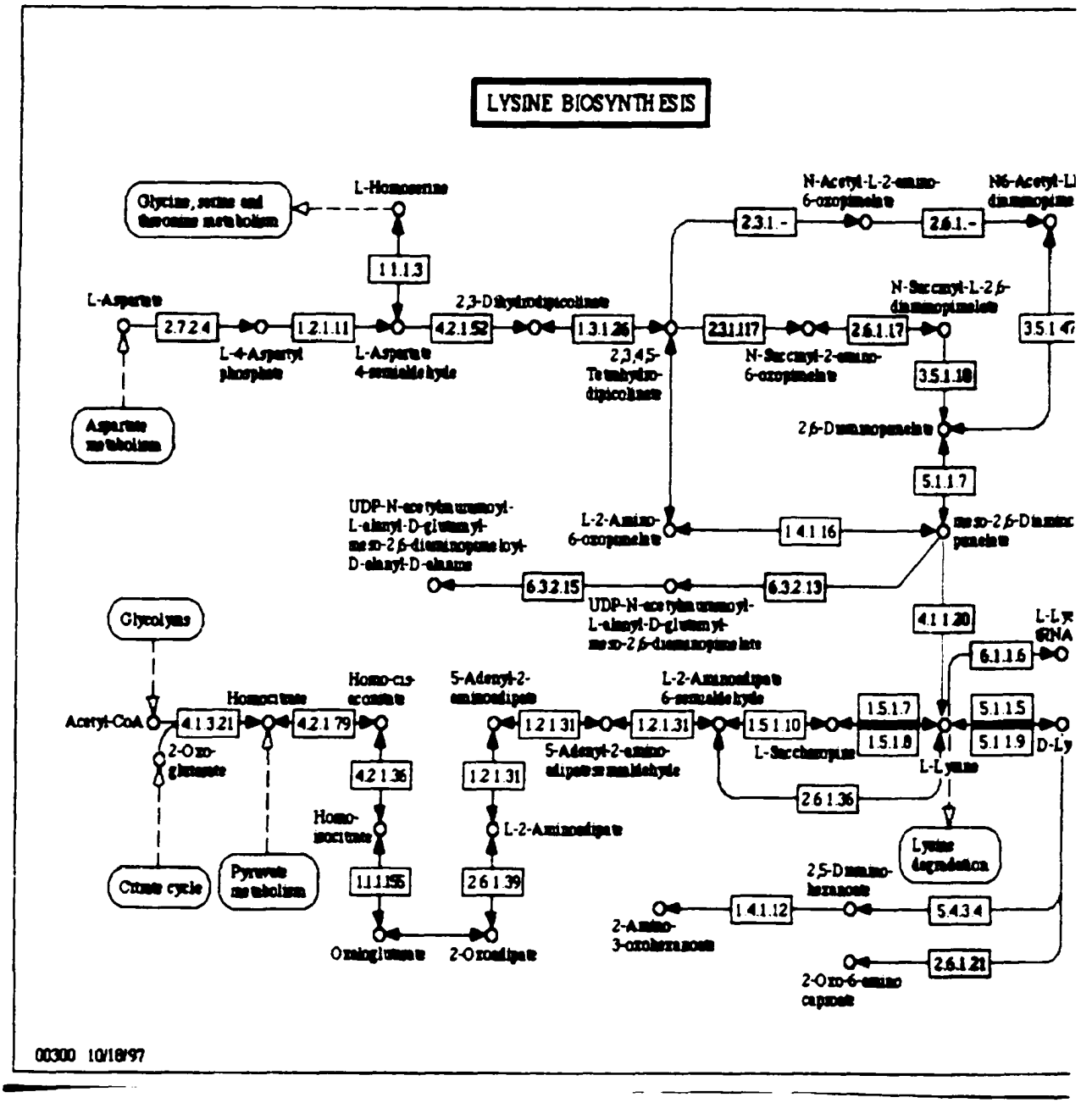
Lysine Biosynthesis

Three pathways of lysine biosynthesis are known in bacteria. Two pathways use succinylated (precursor is oxaloacetate) or acetylated derivatives for synthesis with the acetylation pathway being used by most *Bacilli* including *Bacillus subtilis*. The third pathway involves diaminopimelate (DAP) dehydrogenase (24). *Bacillus stearothemophilus* does not follow the acetylation pathway. However, *Bacillus stearothermophilus* does have the necessary proteins for the succinylated pathway. The precursor for this lysine biosynthesis begins with L-aspartate from aspartate metabolism. Lysine biosynthesis in *Bacillus stearothermophilus* proceeds with genes, to include, *dapG* encodes aspartate kinase (E.C. 2.7.2.4), *asd* encodes aspartate-semialdehyde dehydrogenase (E.C. 1.2.1.11), *dapA* encodes dihydrodipicolinate synthase (E.C. 4.2.1.52), *dapB* encodes dihydrodipicolinate reductase (E.C. 1.3.1.26), *dapD* encodes 2,3,4,5-tetrahydropyridine-2-carboxylate N-succinyltransferase (E.C. 2.3.1.117), *dapC* encodes a probable succinyldiaminopimelate aminotransferase (E.C. 2.6.1.17), *dapE* encodes succinyl-diaminopimelate desuccinylase (E.C. 3.5.1.18), *dapF* encodes diaminopimelate epimerase (E.C. 5.1.1.7), and *lysA* encodes meso-diaminopimelate decarboxylase (E.C. 4.1.1.20) (79).

The acetylation pathway, which is the pathway for *Bacillus subtilis*, has aspartate semialdehyde as the precursor with genes, to include, *dapA* encodes dihydrodipicolinate synthase, *dapB* encodes dihydrodipicolinate reductase, *dapD* encodes tetrahydrodipicolinate acetyltransferase (E.C. 2.3.1.-), *dapC* encodes N-acetyl-L,L-diaminopimelate aminotransferase (E.C. 2.6.1.-), *dapE* encodes N-acetyl-L,L-

diaminopimelate deacetylase (E.C. 3.5.1.47), *dapF* encodes diaminopimelate epimerase, and *lysA* encodes meso-diaminopimelate decarboxylase (79)

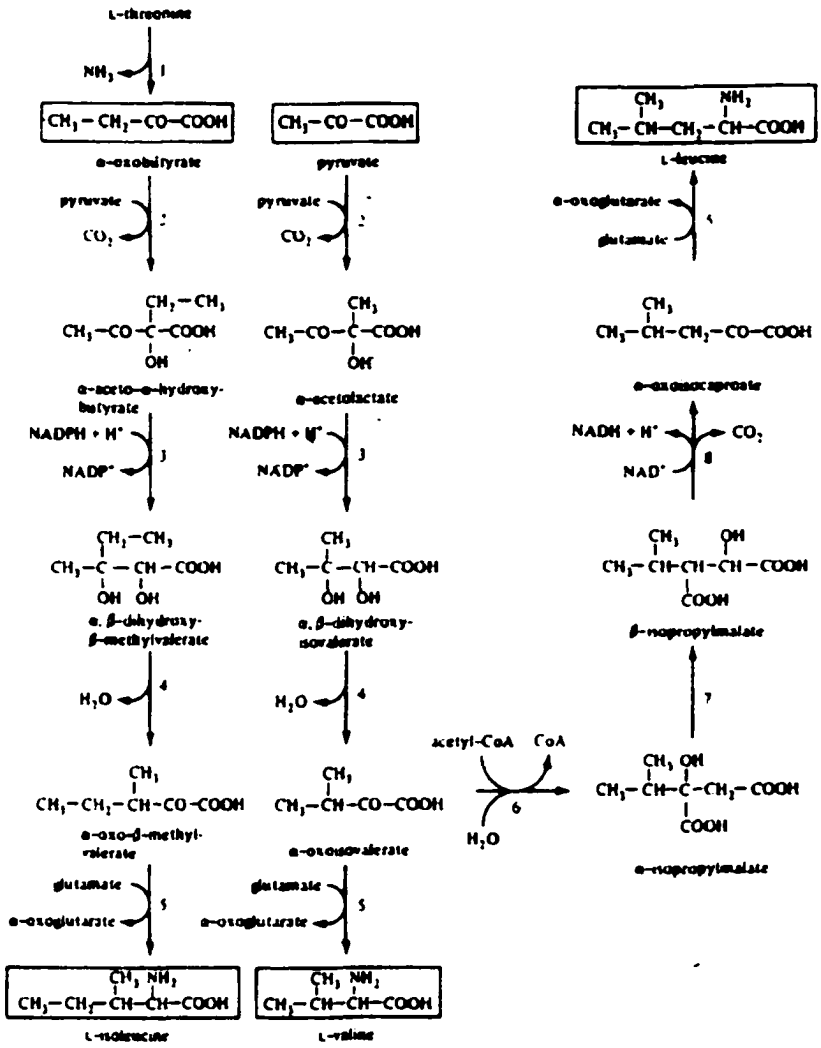
Figure 39-Lysine Biosynthesis (79)



Isoleucine, Valine, and Leucine Biosynthesis

The synthesis of isoleucine begins with threonine and involves the following proteins: threonine dehydratase, acetolactate synthase, ketol-acid reductoisomerase, dihydroxy acid dehydratase (ketol-acid reductoisomerase, and transaminase. The synthesis of valine and leucine starts with pyruvate and involves acetolactate synthase, ketol-acid reductoisomerase, dihydroxy acid dehydratase (ketol-acid reductoisomerase), transaminase, *alpha*-isopropylmalate synthase, isopropylmalate isomerase, and *beta*-isopropylmalate dehydrogenase (112). The pathway is complete in *Bacillus stearothemophilus* and *Bacillus subtilis*.

Figure 40-Isoleucine, Valine, and Leucine Biosynthesis (112)

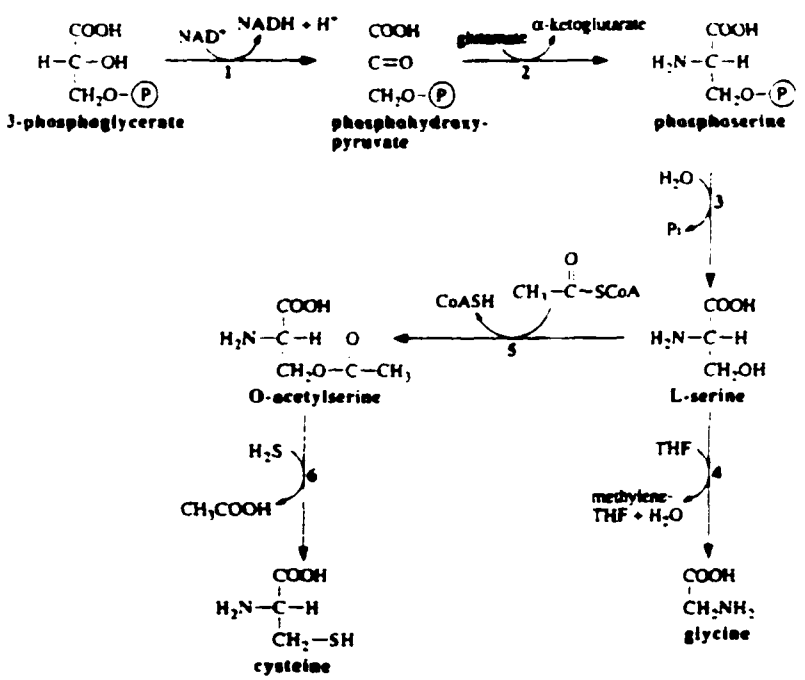


Reprinted by permission of Springer-Verlago GmbH & Co.KG copyright by Springer-Verlag (1986).

Serine, Glycine and Cysteine Biosynthesis

This biosynthesis begins with 3-PGA and involves the following proteins for production of serine: phosphoglycerate dehydrogenase, phosphoserine aminotransferase, homoserine kinase. All of those proteins are needed plus the following protein is needed at step #4 to produce glycine: serine hydroxymethyltransferase. The initial three proteins are needed to produce serine plus the following to produce cysteine: serine transacetylase, and cystathionine gamma-synthase (105). This pathway is complete in both *Bacillus stearothermophilus* and *Bacillus subtilis*

Figure 41-Serine, Glycine and Cysteine Biosynthesis (105)



The sulfur atom is derived from either sulfate or sulfonate with sulfate being preferred, however, sulfonate is the major form of organic sulfur found in soil, the natural habitat of *Bacillus subtilis*. *cysP* encoding sulfate permease is responsible for the internalization of sulfate in *Bacillus subtilis*. The internalized sulfate is activated and reduced to sulfite by the APS/PAPS (adenosine 5'phosphosulfate/3'-phosphoadenosine-5'phosphosulfate) sulfotransferase system. Sulfite reductase reduces sulfite further to sulfide. Sulfite reductase requires siroheme, which is produced in the synthesis of uroporphyrinogen III, as a cofactor. Only one of the four genes involved in the uptake of sulfonate (*ssuBAC'D*) was found in the *Bacillus stearothermophilus* genome (24).

Figure 42-Serine, Glycine, Cysteine, and Methionine Metabolism (24)

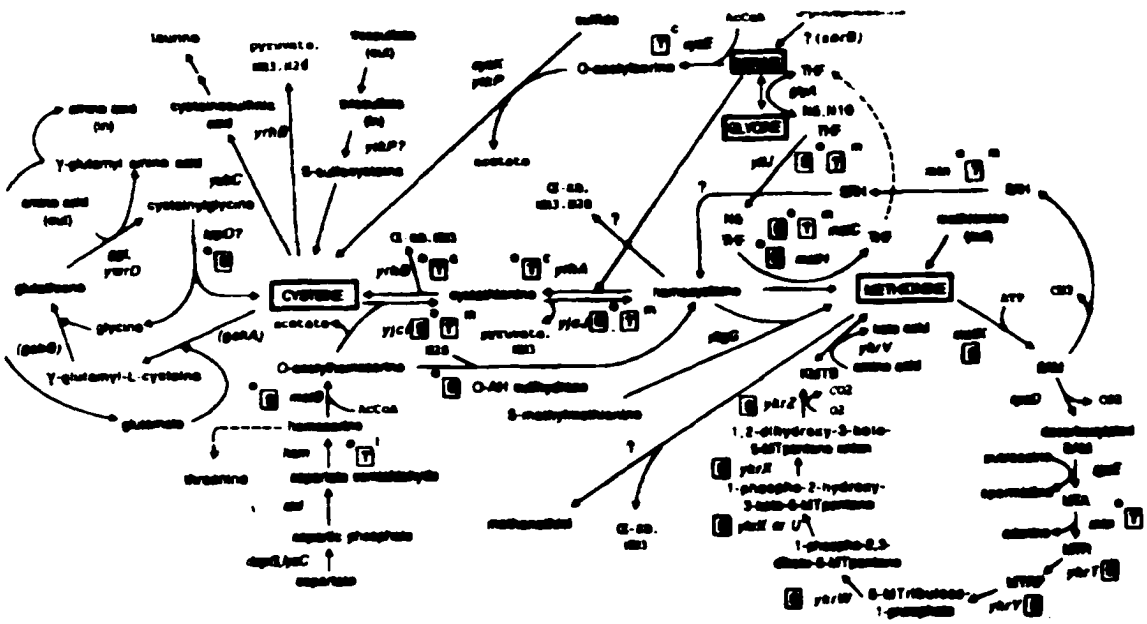


Table 23-Serine, Glycine, Cysteine, and Methionine Biosynthetic Genes in the *Bacillus subtilis* and *Bacillus stearothermophilus* Genomes

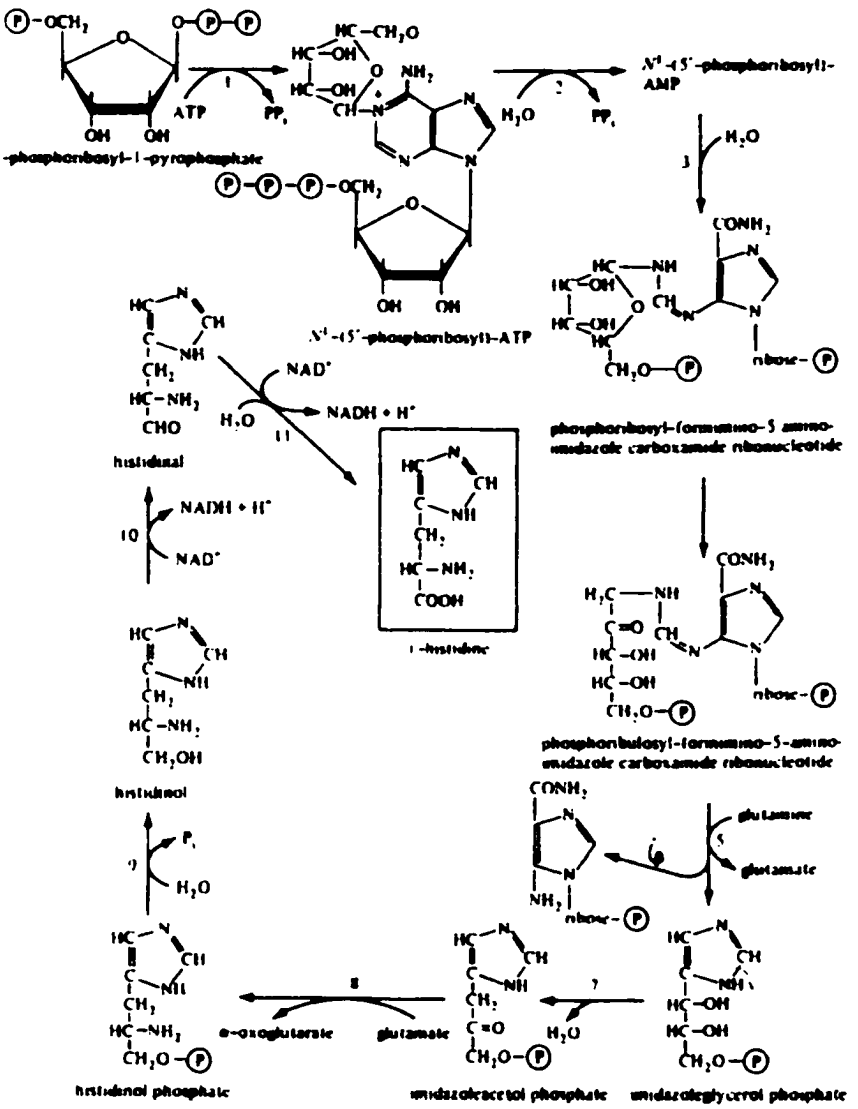
Bsu gene	*Bst*	Product
Serine-Glycine		
serA	yes	Phosphoglycerate dehydrogenase
serC	yes	Phosphoserine aminotransferase
glyA	yes	Serine hydroxymethyltransferase
Cysteine		
cysE	yes	Serine acetyltransferase
cysS	yes	Cysteinyl-tRNA synthetase
ytkP	yes	O-Acetylserine (thiol)-lyase
cysH	yes	PAPS sulfotransferase reductase
cysP	yes	Sulfate permease
ylnB	yes	Sulfate adenyltransferase
ylnC	yes	APS kinase
ylnD	yes	URO III methyltransferase
ylnE	yes	Unknown
ylnF	yes	Ferrochelatase
yitA	yes	Sulfate adenyltransferase
yisZ	yes	APS kinase?
yvgR	yes	Sulfite reductase flavoprotein
yvgQ	yes	Sulfite reductase hemoprotein
ssuB	yes	ABC transporter, ATP binding
ssuA	yes	Aliphatic sulfonate binding protein
ssuC	yes	ABC transporter, membrane

ssuD	no	FMNH2-dependent monooxygenase
Methionine		
hom	yes	Homoserine dehydrogenase
metB	yes	Homoserine O-acetyltransferase
yjcI	yes	Cystathionine *gamma*-synthase
yjcJ	yes	Cystathionine *beta*-lyase
CAC5	no	O-Acetylserine sulfhydrase
metH	yes	Methionine synthase, B12 dependent
yitJ	yes	Methylene tetrahydrafolate reductase
metS	yes	Methionyl-tRNA synthetase
metK	yes	SAM synthetase
speD	yes	SAM decarboxylase
speE	yes	Spermine/spermidine synthase
yrrT	yes	Methyltransferase
mtn	yes	MTA/SAH nucleosidase
yrhA	yes	Cystathionine *beta*-synthase
yrhB	yes	Cystathionine gamma-lyase
ykrT	yes	MTR kinase
ykrS	yes	MTA recycling
ykrW	yes	Rubisco homolog
ykrX	yes	MTA recycling
ykrY	yes	MTA recycling
ykrZ	yes	MTA recycling
ykrU	yes	MTA recycling
ykrV	yes	MTA recycling
ybgG	yes	S-Methylmethionine-homocysteine methyl transferase
yrvO	yes	Cysteine desulfhydrase
yubC	yes	Cysteine dioxygenase
yusC	yes	ABC transporter, ATP binding
yusB	yes	ABC transporter, binding protein
yusA	yes	ABC transporter, membrane
yoaD	yes	Similar to serA
yoaC	yes	Similar to sugar kinase
yoaB	yes	Similar to transporters
yxjG	yes	Similar to metC
yxjH	yes	Similar to metC
ytnJ	yes	Dibenzothiophene desulfurase
ribR	yes	Riboflavin kinase
hipO	no	Glutathione utilization

Histidine Biosynthesis

The biosynthesis of histidine begins with PRPP and involves phosphoribosyl ATP pyrophosphorylase, phosphoribosyl-ATP pyrophosphohydrolase, phosphoribosyl-AMP cyclohydrolase, phosphoribosyl formimino-5-aminoimidazole carboxamide ribonucleotide isomerase, phosphoribulosyl-formimino-5-aminoimidazole carboxamide ribonucleotide: glutamine aminotransferase, imidazoleglycerol phosphate dehydratase, histidinol-phosphate aminotransferase, histidinol phosphatase, and histidinol dehydrogenase (112). This pathway is complete in *Bacillus stearothermophilus* and *Bacillus subtilis*.

Figure 43-Histidine Biosynthesis (112)



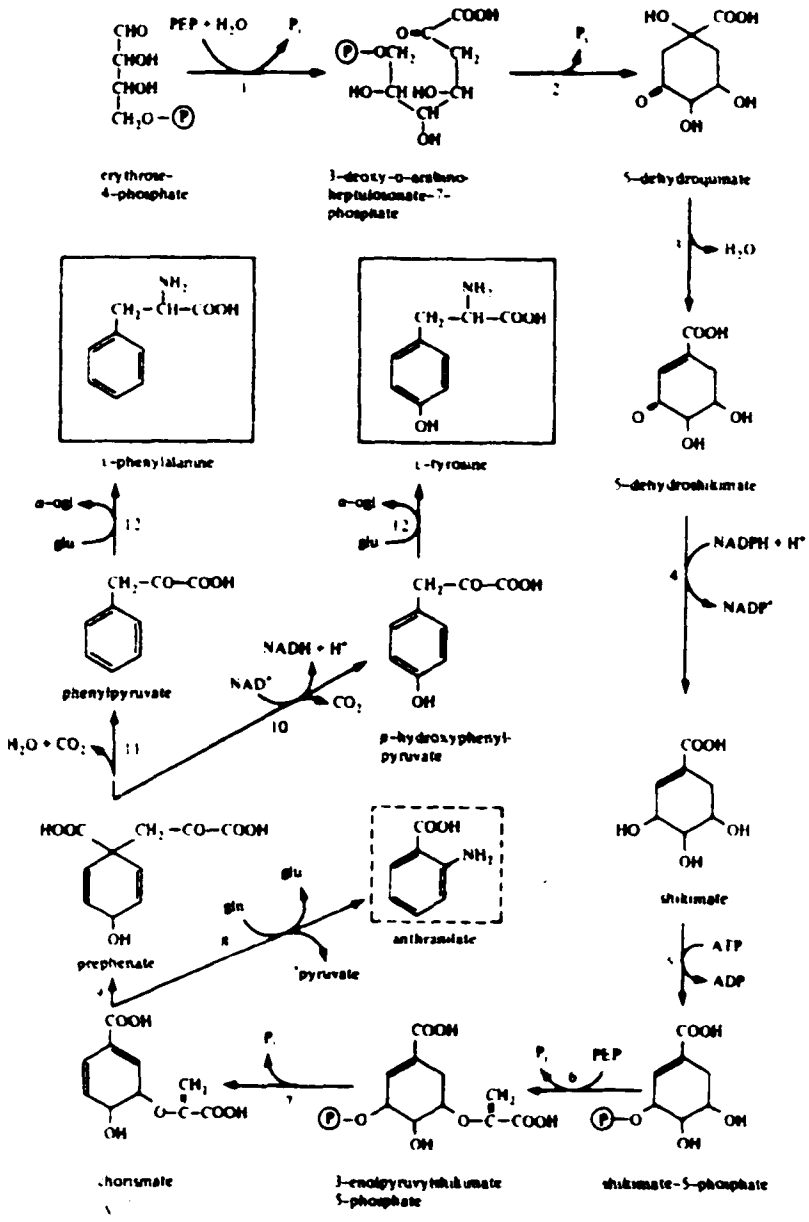
Reprinted by permission of Springer-Verlago GmbH & Co KG copyright by Springer-Verlag (1986).

Phenylalanine, Tyrosine, and Tryptophan Biosynthesis

Both the *Bacillus stearothermophilus* and *Bacillus subtilis* genomes encode the enzymes needed for aromatic amino acid biosynthesis. The two precursors for aromatic amino acid biosynthesis are phosphoenolpyruvate from glycolysis, and erythrose-4-phosphate from the pentose phosphate cycle. The proteins involved in the synthesis of all three amino acids include 3-deoxy-D-arabinoheptulosonate synthase, 5-dehydroquinate synthase, 5-dehydroquinate dehydratase, shikimate dehydrogenase, shikimate kinase, enolpyruvylshikimate-5-phosphate synthase, and chorismate synthetase. The remainder of the proteins involved in the phenylalanine biosynthetic pathway are chorismate mutase, prephenate dehydratase, and transaminase (tyrosine/phenylalamine amino-transferase), while the remainder of the proteins involved in the tyrosine biosynthetic pathway are chorismate mutase, prephenate dehydrogenase and transaminase (tyrosine/phenylalamine amino-transferase) (24, 112)

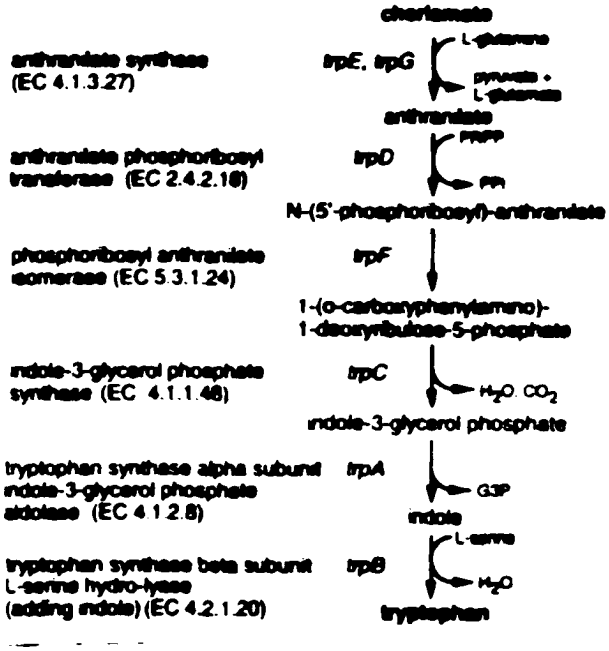
In addition to the two precursors mentioned above (phosphoenolpyruvate and erythrose-4-phosphate), PRPP is the third precursor for the biosynthesis of tryptophan. The rest of the proteins involved in the tryptophan biosynthetic pathway are anthranilate synthase, anthranilate phosphoribosyl transferase, phosphoribosyl-anthranilate isomerase, indole-3-glycerol phosphate synthase, and tryptophan synthase. Anthranilate synthase has two polypeptides encoding *trpE* and *trpG*. *trpG* catalyzes the conversion of glutamine to glutamate. Tryptophan synthase also has two polypeptides encoding *trpA* and *tryB* (24).

Figure 44-Phenylalanine and Tyrosine Biosynthesis (112)



Reprinted by permission of Springer-Verlago GmbH & Co KG copyright by Springer-Verlag (1986)

Figure 45-Tryptophan Biosynthesis (24)



Amino Acid Degradation

All twenty amino acids are degraded to the following seven intermediates: pyruvate, acetyl CoA, acetoacetyl-CoA, *alpha*-ketoglutarate, succinyl-CoA, fumarate, and oxaloacetate. More specifically, alanine, glycine, cysteine, serine, threonine, and tryptophan become pyruvate. Isoleucine, leucine, and tryptophan become acetyl-CoA. Leucine, lysine, phenylalanine, tyrosine, and tryptophan become acetoacetyl-CoA. Glutamine and glutamate become *alpha*-ketoglutarate. Isoleucine, methionine, and valine become succinyl-CoA. Tyrosine, aspartate, and phenylalanine become fumarate Aspartate and asparagines become oxaloacetate (105).

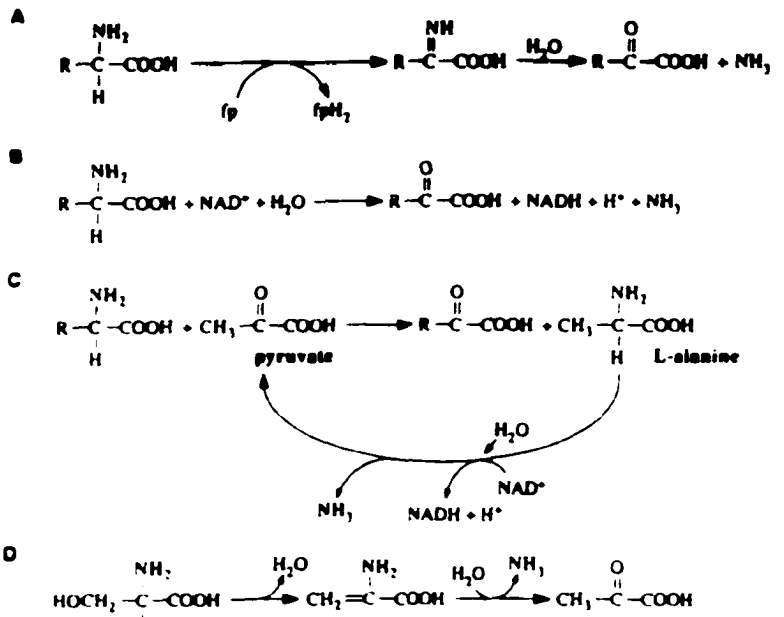
The first reaction of amino acid degradation involves the removal of the amino group to generate the *alpha* -keto acid that eventually enters the TCA cycle. There are four ways the amino groups are removed from amino acids. As viewed in pathway A below, amino acids can be oxidatively deaminated using a nonspecific flavoprotein oxidase to their corresponding keto acid, resulting in feeding the electrons directly into the electron transport chain. The *Bacillus stearothermophilus* genome encodes the FAD flavoprotein oxidase.

In pathway B, amino acid dehydrogenases that are NADP+ linked catalyze the reductive amination of the keto acid. Serine, glutamate, alanine, leucine, valine, aspartate and glycine dehydrogenases are identified in the *Bacillus stearothermophilus* genome. In pathway C, amino acids can be deaminated by a transaminase transferring an amino group to pyruvate forming L-alanine. The amino group then is released as ammonia while pyruvate is regenerated by alanine dehydrogenase. Glutamate and aspartate transaminases are identified in the *Bacillus stearothermophilus* genome.

Finally, in pathway D, the amino acids can be deaminated by dehydratases

(deaminases), some of which proceed by the elimination of water (105). Amino acids, such as, serine, threonine, leucine, and cysteine undergo deamination in *Bacillus stearothermophilus*. Too much cysteine is detrimental to bacteria causing inhibition of homoserine dehydrogenase, which results in a shortage of threonine and methionine. Cysteine can be deaminated to pyruvate, H2S, and NH3 (24). One of the characteristics of *Bacillus stearothermophilus* DSM 22 is that phenylalanine is not deaminated (4).

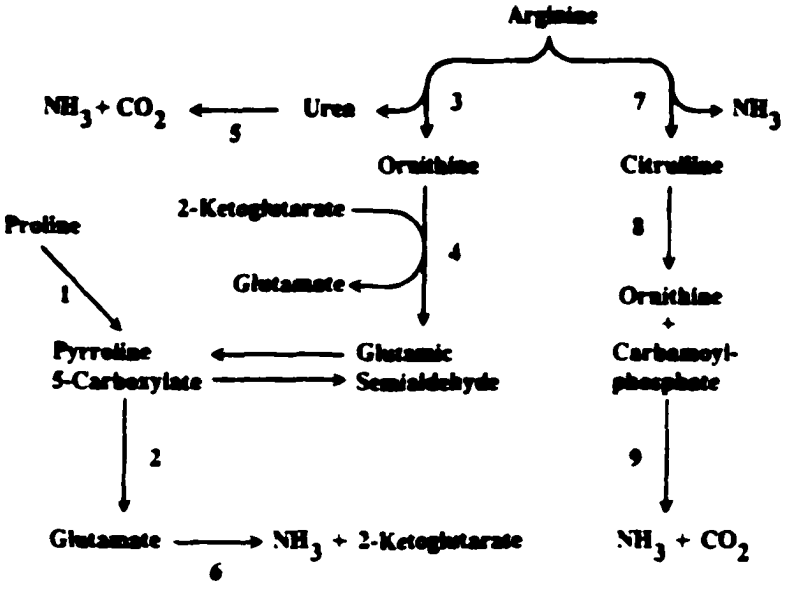
Figure 46-The Removal of Amino Groups from Amino Acids during Catabolism (105)



Methionine gamma lyase converts methionine to alpha-ketobutyrate, methanethiol, and ammonia (24). Another characteristic of *Bacillus stearothermophilus* DSM 22 is that tyrosine is not degraded (4).

The proline degradative enzymes, which are present both in *Bacillus stearothermophilus* and *Bacillus subtilis* are as follows: 1) proline oxidase, and 2) pyrroline-5-carboxylate dehydrogenase. The arginine degradative enzymes follow the arginase dependent pathway to include the following: 3) arginase, 4) ornithine transaminase, 5) urease. Both *Bacillus stearothermophilus* and *Bacillus subtilis* have these enzymes. Although *Bacillus stearothermophilus* has 7) arginine deiminase, and 8) ornithine carbamoyltransferase, it does not have 9) carbamate kinase. *Bacillus subtilis* also does not contain the arginine deiminase pathway.

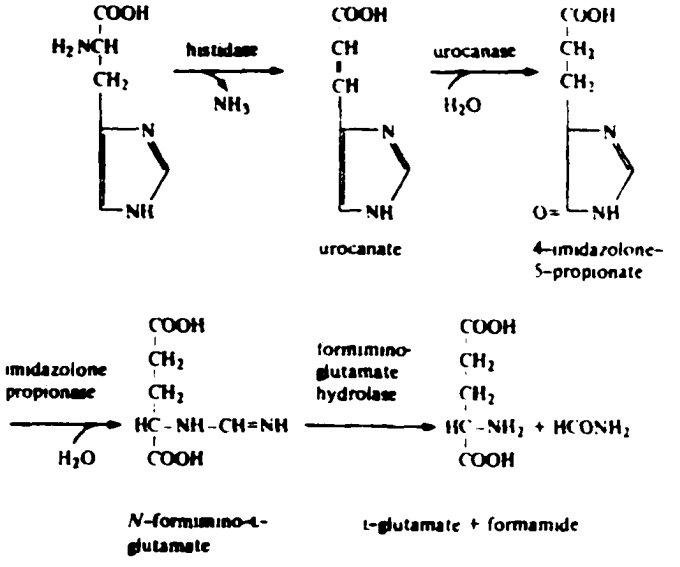
Figure 47-Arginine and Proline Degradation (24)



Histidine catabolism

The breakdown of histidine results in glutamate and formamide and both the *Bacillus stearothermophilus* and *Bacillus subtilis* genomes encode the following four genes required to include: *(hutH)* histidase, *(hutU)* urocanase, *(hutI)* imidazolone propionase, and *(hutG)* formiminoglutamate hydrolase (112).

Figure 48-Histidine Catabolism (112)

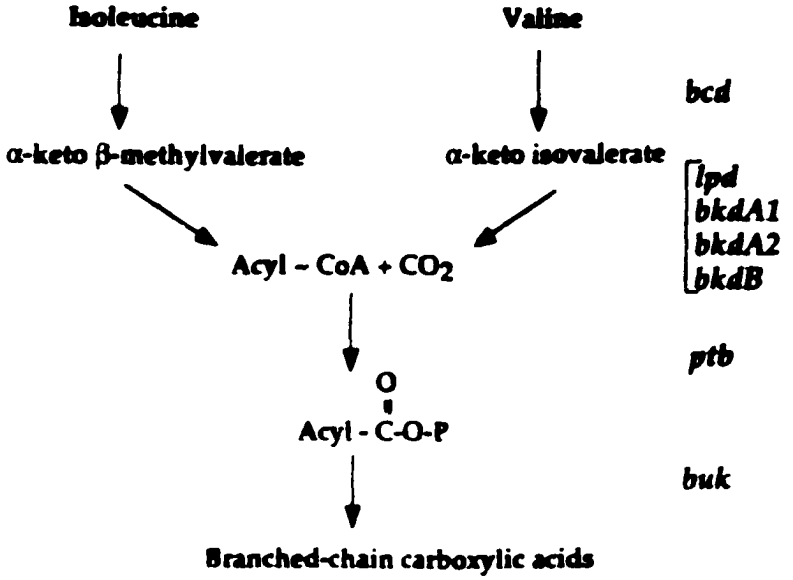


Reprinted by permission of Springer-Verlago GmbH & Co KG copyright by Springer-Verlag (1986).

Isoleucine and Valine Degradative Pathways

The isoleucine and valine degradative enzymes in *Bacillus stearothermophilus* and *Bacillus subtilis* include the following: *(bcd)* leucine dehydrogenase, (*bkdA1, bkdA2, bkdB, lpd*) branched-chain *alpha*-keto acid dehydrogenase, *(ptb)* phosphate butyryl-CoA transferase, and *(buk)* butyrate kinase (24). This pathway is complete in *Bacillus stearothermophilus* and *Bacillus subtilis*.

Figure 49-Isoleucine and Valine Degradative Pathways (24)

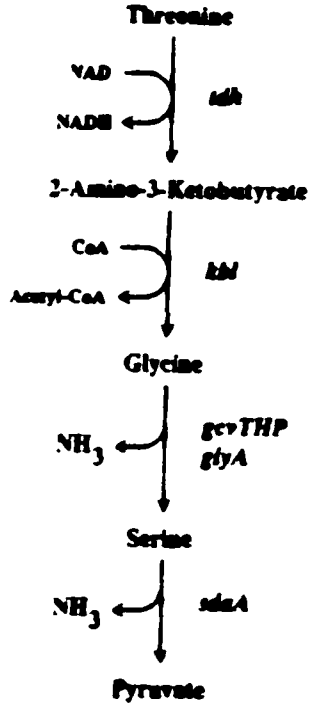


As shown in Figure 21, leucine, isoleucine, and valine also degrade into fatty acids

Threonine Degradative Pathway

Possibly, the enzymes involved in threonine degradation in *Bacillus stearothermophilus* and *Bacillus subtilis* are as follows: *(tdh)* threonine dehydrogenase, *(kbl)* 2-amino-3-ketobutyrate CoA ligase, (*gcvTHP*) glycine cleavage system, (*glyA*) serine hydroxymethyltransferase, and *(sdaA)* serine deaminase (24). This pathway is complete in *Bacillus stearothermophilus* and *Bacillus subtilis*.

Figure 50-Threonine Degradative Pathway (24)



Transport

Both *Bacillus stearothermophilus* and *Bacillus subtilis* are soil bacteria. One lipid bilayer acts as a barrier enabling uptake of sugars, amino acids, phosphates, sulfates, metallic cations, vitamins, organo-iron complexes, and other nutrients, and expulsion of drugs, animal toxins, and proteases. These organisms secrete polysaccharide degrading enzymes into the environment for the uptake and metabolism of these nutrients. The phosphotransferase system (PTS) is a carbohydrate transport and phosphorylation system using either of the following phosphorylation cascade enzymes: enzyme I, HPr, enzyme IIA, and enzyme IIB. Phosphoenolpyruvate (PEP) provides the phosphoryl group (24).

The ATP-binding cassette (ABC) transport system, known as ABC importers and extruders, is comprised of many membrane-bound protein subunits. Typically, the four subunits include two hydrophobic membrane-spanning domains (MDS) and two hydrophilic nucleotide-binding domains (NBD), which are localized at the cytoplasmic face of the cell membrane. The two MDS are the pathway through which the solute crosses the membrane. The two NBD act as pumps that couple ATP hydrolysis to transport. In addition, the uptake system has an extra cytoplasmic substrate-binding protein (SBP), which is a lipoprotein that interacts with the incoming substrate. The transport proteins shown in table 24 represent a small amount of the proteins involved in the phosphotransferase system (PTS) and the ATP-binding cassette (ABC) transport system (24).

Table 24-Transport Genes

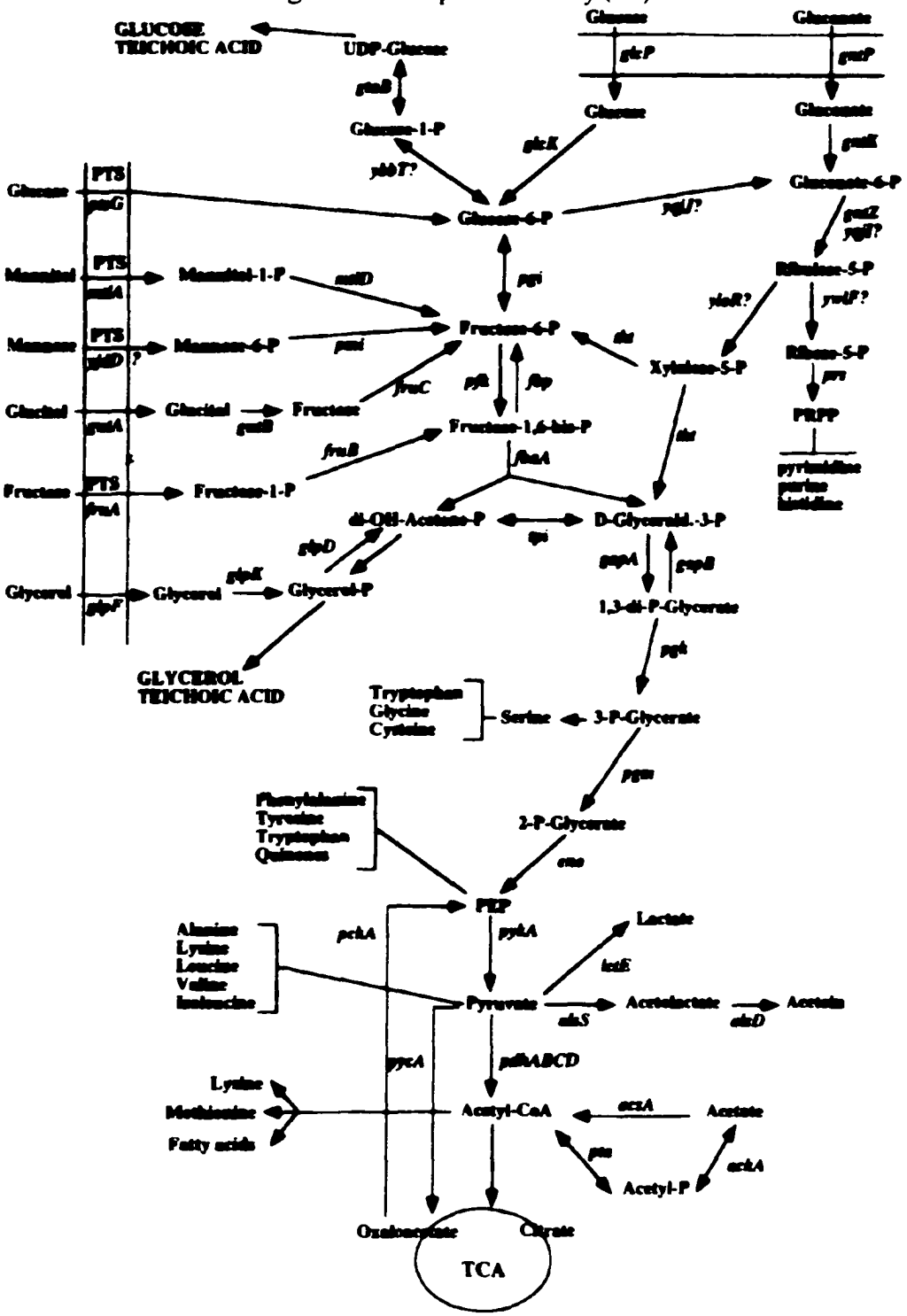
Bsu	*Bst*	PTS (phosphotransferase system)
malP	yes	PTS maltose specific enzyme IICB component
mtlA	yes	mannitol

fruA	yes	fructose
levF	yes	fructose
levG	yes	fructose
levD	yes	fructose
levE	yes	fructose
glpF	yes	glycerol
ptsG	yes	glucose
sacP	yes	sucrose
sacX	yes	sucrose
no	no	arbutin (glvC)
malP	yes	arbutin
no	no	(*ptfB*)
bglP	yes	beta-glucosides
no	Yes	cellobiose (*celB*)
no	Yes	*celC*
no	yes	*celA*
ydhO	yes	cellobiose
ydhN	yes	cellobiose
ydhM	yes	cellobiose
ywbA		?
treP	yes	trehalose
gamP	yes	glucosamine
		ATP-Binding Casette (ABC) Transport System
rbsB	yes	ribose
rbsC	yes	ribose
rbsD	yes	ribose
rbsA	yes	ribose
gntP	yes	gluconate
glcP	yes	glucose/mannose:H+ symporter
yurO	yes	maltose/lactose
yurN	yes	maltose/lactose
yurM	yes	maltose/lactose
yurJ	yes	maltose/lactose
natB	yes	sodium
natA	yes	sodium
opuBD	yes	choline
opuBC	yes	choline
opuBB	yes	choline
opuBA	yes	choline
opuCC	yes	glycine betaine/carnitine/choline
opuCB	yes	glycine betaine/carnitine/choline
opuCD	yes	glycine betaine/carnitine/choline
opuCA	yes	glycine betaine/carnitine/choline
glnH	yes	glutamine
glnM	yes	glutamine
glnP	yes	glutamine
glnQ	yes	glutamine
fhuC	yes	ferrous iron transport, protein B
fhuG	yes	ferrous iron transport, protein B
fhuB	yes	ferrous iron transport, protein B

fhuD	yes	ferrous iron transport, protein B
yvrA	yes	similar to iron transport system
yvrB	yes	similar to iron permease
yvrC	yes	similar to iron-binding protein
fhuC	yes	ferrichrome ABC transporter (ATP-binding protein)
fhuG	yes	ferrichrome ABC transporter (permease)
fhuB	yes	ferrichrome ABC transporter (permease)
fhuD	yes	ferrichrome ABC transporter (ferrichrome-binding protein)
oppA	yes	oligopeptide
oppB	yes	oligopeptide
oppC	yes	oligopeptide
oppD	yes	oligopeptide
oppF	yes	oligopeptide
appA	yes	oligopeptide
appB	yes	oligopeptide
appC	yes	oligopeptide
appD	yes	oligopeptide
appF	yes	oligopeptide
dppA	yes	dipeptide
dppE	yes	dipeptide
no	no	dppB
dppC	yes	dipeptide
dppD	yes	dipeptide
ykfD	yes	dipeptide
araP	yes	arabinose
araQ	yes	arabinose
araN	yes	arabinose
yvgM	yes	molybdenum
yvgL	yes	molybdenum
yufN	yes	lipoprotein
yufP	yes	lipoprotein
yufQ	yes	lipoprotein
yufO	yes	lipoprotein
ygbA	no	nitrate (*ygbA*)
ygaM	no	nitrate
ygaL	no	nitrate
ycdH	yes	manganese
yceA	yes	manganese
ycdI	yes	manganese
no	no	*ytgA*
no	no	*ytgC*
no	no	*ytgD*
no	no	*ytgB*
no	no	*phosphate (yqgG)*
no	no	*yqgH*
no	no	*yqgI*
no	no	*yqgJ*
no	no	*yqgK*
ybaR	yes	sulfate
yvdB	yes	sulfate

nrgA	yes	ammonium
ykrM	yes	potassium
yubG	yes	potassium
amyC	yes	maltose
amyD	yes	sugar
msmE	yes	sugar
		Polysaccharide Degrading Enzymes
amyE	yes	alpha amylase
amyX	yes	pullulanase
ydhT	yes	endo-*beta*-1,4-mannanase
sacC	yes	levanase
yvdF	yes	glucan-1,4-alpha-maltohydrolase
pel	yes	pectate lyases
pelB	yes	
bglS	yes	beta-1,3-1,4 endoglucanase
xynA	yes	endo-1,4-beta-xylanases
xynD	yes	endo-1,4-beta-xylanases

Figure 51-Complete Pathway (24)



D. Codon Usage

1. Comparison of Thermophilic and Mesophilic Amino Acid Residues

The authors of the *Aquifex aeolicus* genome sequencing project compared the relative amino acid compositions (in percentages) of the mesophiles *Haemophilus influenzae*, *Helicobacter pylori*, *Escherichia coli*, and *Synechosystis* sp., and the hyperthermophiles *Aquifex aeolicus* and *Methanococcus jannaschii* (38). The authors stated that the data obtained by comparing relative amino acid compositions might suggest trends that may correlate with the thermostability of the encoded proteins (38), i.e., that hyperthermophilic genomes encode more charged amino acid residues than mesophilic genomes, and that glutamine is rarely seen in the hyperthermophiles (38). Glutamine and asparagines are unstable amino acids at high temperatures, and it has been reported that thermophiles have low codon usage of glutamine and asparagines since these amino acids are easily deaminated (113). It also has been reported that no unusual structures were associated with thermophiles as compared to mesophiles since thermophilic proteins had similar helical, B-sheet, and were essentially homologous to their mesophilic counterparts. This group also showed that thermophiles have no additional hydrophobic residues as compared to mesophiles (7). Hocking and Harris reported that thermophiles should exhibit a decrease in cysteine content as compared to mesophiles, particularly a decrease of ancillary sulphydryl groups that would be exposed on the surface of the protein and subject to oxidation (8).

During this dissertation study, in addition to the above organisms from the *Aquifex aeolicus* paper, the codon usage of two mesophiles and one hyperthermophile, and three thermophiles are considered. The amino acid distribution for *Bacillus subtilis* was provided by Ivan Moszer of the Institut Pasteur, France (81). Codon usage of *Bacillus halodurans* was obtained from JAMSTEC (83). Codon usage of *Thermotoga maritime* was obtained from the TIGR website (41). The codon usage of *Bacillus stearothermophilus*, *Thermoplasma volcanium and Thermoplasma acidophilum* were

obtained after ORF analysis using a Perl script written by Jim White

Tables 25, 26, 27 and 28 show a comparison of relative amino acid compositions (in percentages) of the analyzed mesophile and thermophile genomes presented both tabularly and graphically. From this data, it can be seen that thermophilic charged residues account for 27.39 percent, but in mesophiles, the charged amino acid represent only 24.73 percent. This agrees with the observation that hyperthermophilic genomes encode more charged amino acid residues than mesophilic genomes. In thermophiles, the polar/uncharged residues account for 28.63 percent, while in mesophiles, their percentage is 30.91 percent. Thermophilic hydrophobic residues account for an average of 43.92 percent of the amino acid composition versus 44.24 percent in mesophiles, which agrees with a previous study that states that the hydrophobic residues in thermophiles averaged no additional hydrophobic residues as compared to mesophiles. The six thermophiles average 2.23 percent glutamine residues as compared to the mesophilic glutamine residues which average 4.33. This study agrees with the previous study that glutamine is rarely seen in the hyperthermophiles. The six thermophiles average 4.08 percent asparagines residues whereas the six mesophiles average 4.32 percent. The lower codon usages of glutamine and asparagines agree with previous statements that suggest possibly protection of the thermophiles from deamination at their higher growing temperatures with the use of lower percentages of these two amino acids. The six thermophiles average 0.84 percent cysteine residues in each genome whereas the six mesophiles average 0.96 percent. This study does agree with a previous study that states that thermophiles should exhibit a decrease in cysteine content as compared to mesophiles.

Table 25-Comparison of Relative Amino Acid Compositions (in percentages) of Thermophiles

	A. aeolicus	*T. maritima*	*M. jannaschii*	*B. stearothermophilus*	*T. acidophilum*	*T. volcanium*
Charged AA	**29.8**	**28.6**	**29.83**	**26.62**	**24.6**	**26**
aspartic acid	4.32	4.93	5.52	4.77	5.8	5.5
glutamic acid	9.63	8.86	8.67	7.06	6	6.4
histidine	1.54	1.58	1.43	2.39	1.6	1.5
lysine	9.4	7.58	10.36	5.46	5.7	6.9
arginine	4.91	5.55	3.85	6.94	5.5	4.7
Polar/uncharged AA	**26.31**	**26.99**	**27.21**	**28.68**	**31.3**	**31.4**
cysteine	0.79	0.71	1.27	1.09	0.6	0.6
glycine	6.75	6.87	6.41	7.41	7.3	7
asparagine	3.6	3.6	5.24	3.06	4.3	4.7
glutamine	2.04	2.02	1.44	3.55	2.2	2.1
serine	4.79	5.7	4.46	5.3	7.5	7.5
threonine	4.21	4.54	4.06	5.01	4.8	4.8
tyrosine	4.13	3.55	4.33	3.16	4.6	4.7
Hydrophobic AA	**43.77**	**44.38**	**42.84**	**44.8**	**44.2**	**43.5**
alanine	5.9	5.81	5.54	9.3	7	6.4
phenylalanine	5.13	5.17	4.2	4.19	4.7	4.7
isoleucine	7.32	7.13	10.45	6.06	9	9.2
leucine	10.57	10.02	9.38	9.56	8.4	8.7
methionine	1.92	2.29	2.33	2.46	3.1	2.7
proline	4.07	3.98	3.38	4.54	3.9	3.8
valine	7.93	8.62	6.85	7.38	7.2	7.2
tryptophan	0.93	1.36	0.71	1.31	0.9	0.8

Table 26-Comparison of Relative Amino Acid Compositions (in percentages) of Mesophiles

	H influenzae	*H pylori*	*E coli*	*Synechosystis*	*B subtilis*	*B halodurans*
Charged AA	**24.3**	**26.17**	**23.43**	**22.64**	**26.87**	**26.96**
aspartic acid	4.98	4.77	5.2	5.07	5.18	5.06
glutamic acid	6.48	6.88	5.91	6.2	7.24	7.82
histidine	2.05	2.12	2.26	1.93	2.27	2.41
lysine	6.32	8.94	4.48	4.26	7.05	5.85
arginine	4.47	3.46	5.58	5.18	4.13	4.82
Polar/uncharged AA	**31.37**	**31.24**	**30.68**	**31.67**	**30.69**	**29.93**
cysteine	1.03	1.09	1.11	1.01	0.8	0.73
glycine	6.65	5.76	7.42	7.77	6.91	6.99
asparagine	4.89	5.83	3.88	3.76	3.95	3.61
glutamine	4.64	3.7	4.42	5.26	3.84	4.1
serine	5.84	6.81	5.67	5.46	6.29	5.6
threonine	5.2	4.37	5.35	5.53	5.42	5.53
tyrosine	3.12	3.68	2.83	2.78	3.48	3.37
Hydrophobic AA	**44.23**	**42.47**	**46.79**	**46.67**	**43.43**	**43.84**
alanine	8.21	6.83	9.55	9.07	7.68	7.36
phenylalanine	4.46	5.41	3.87	3.75	4.5	4.44
isoleucine	7.1	7.2	5.95	6.31	7.36	6.91
leucine	10.5	11.18	10.56	10.93	9.65	9.98
methionine	2.44	2.28	2.86	2.12	2.78	2.68
proline	3.72	3.28	4.41	5.09	3.68	3.8
valine	6.68	5.59	7.11	7.1	6.75	7.46
tryptophan	1.12	0.7	1.48	1.3	1.03	1.21

Table 27

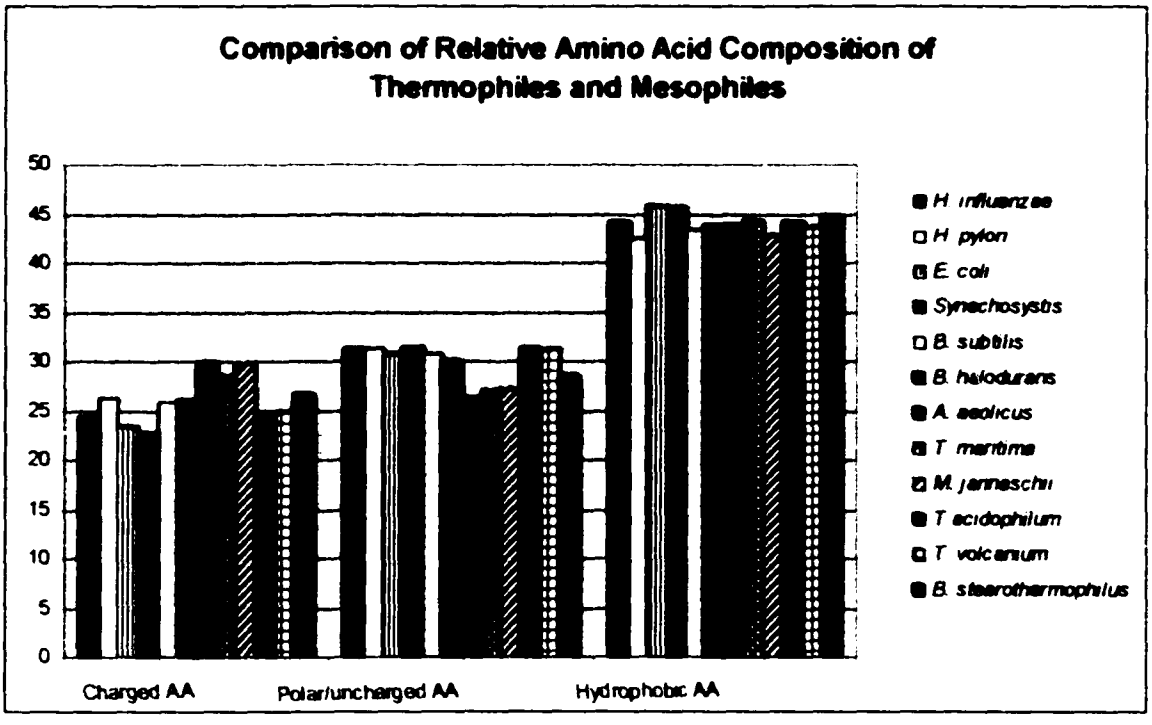
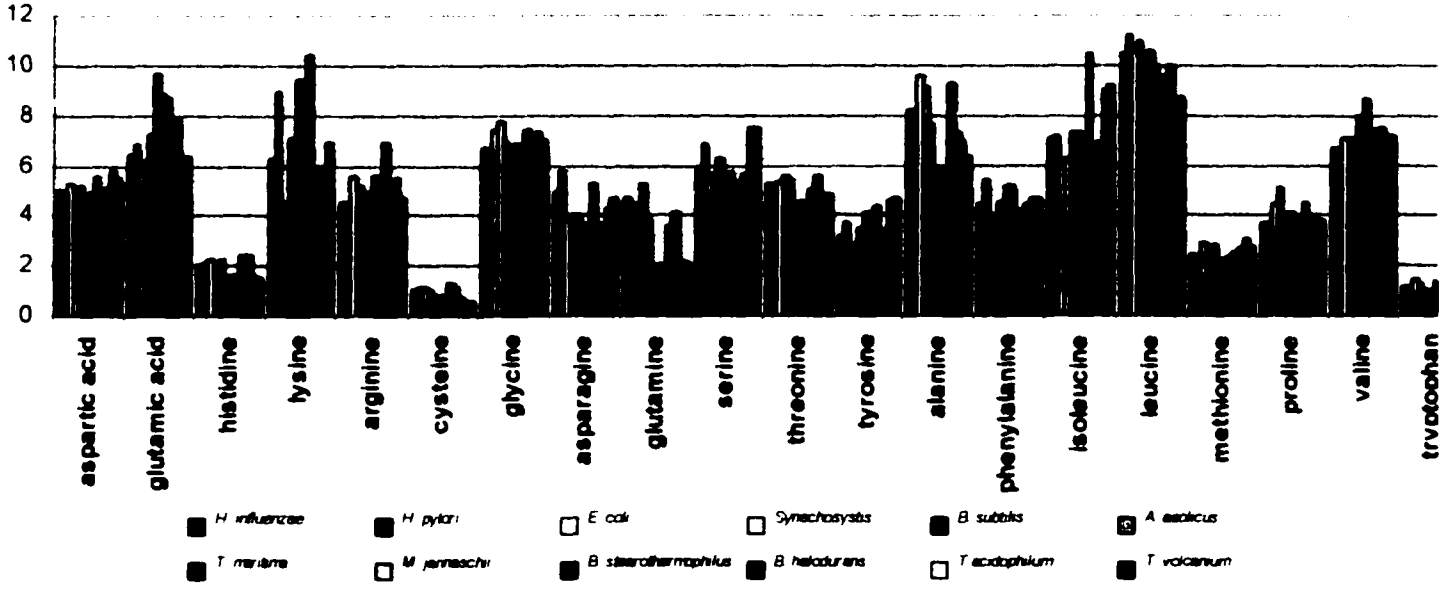


Table 28

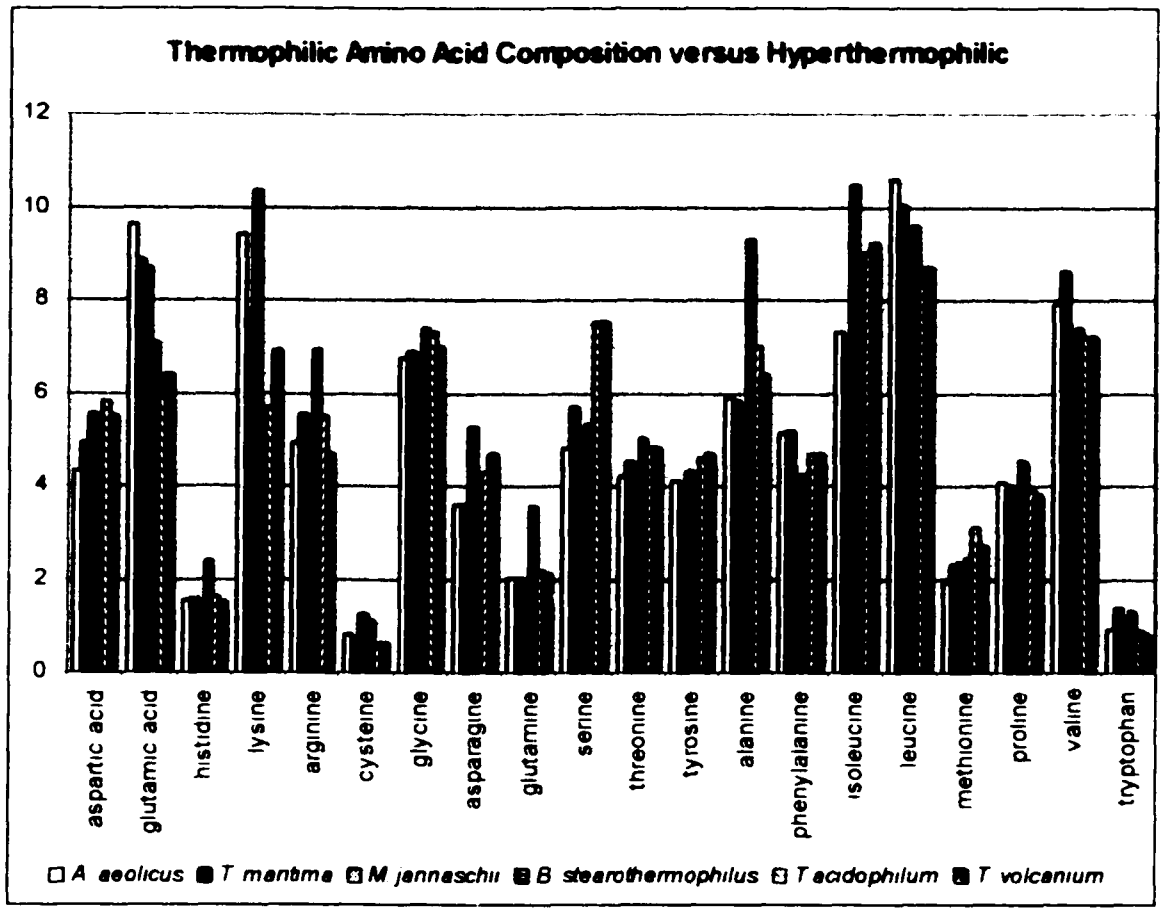
Relative Amino Acid Composition Among Thermophiles, Hyperthermophiles and Mesophiles



2. Comparison of Hyperthermophilic and Thermophilic Amino Acid Composition

It was reported that hyperthermophiles have a higher level of charged amino acids at the expense of polar/uncharged residues (38). As seen in table 25 through table 29, the relative amount of charged residues is 1.24 percent higher than polar/uncharged residues. More specifically, in the thermophiles *Bacillus stearothermophilus*, *Thermoplasma acidophilum*, and *Thermoplasma volcanium* the percentage of lysine is much lower at 5.46, 5.7 and 6.9 percent, respectively, than in the hyperthermophiles where lysine accounts for 9.4 percent in *Aquifex aeolicus*, 7.58 percent in *Thermotoga maritime*, and 10.36 in *Methanococcus jannaschii*. Arginine is present at 6.94 percent and histidine at 2.39 percent in *Bacillus stearothermophilu*, values that are highers than in the other two thermophiles and the three hyperthermophiles. It was reported that the amino acid glutamine is almost absent in hyperthermophiles (38), and indeed, *Bacillus stearothermophilus* has 3.55 percent glutamine, *Thermoplasma acidophilum* has 2.2 percent, and *Thermoplasma volucanium* has 2.1 percent, whereas the three hyperthermophiles *Aquifex aeolicus*, *Thermotoga maritime*, and *Methanococcus jannaschii* have only 2.04, 2.02, and 1.44 percent glutamine, respectively. *Bacillus stearothermophilus* proteins contain 9.3 percent alanine, while *Thermoplasma acidophilus* has 7.0 percent and *Thermoplasma volcanium* has 6.4 percent, values that are higher than in the hyperthermophiles, which have 5.9, 5.81, and 5.54 percent alanine in *Aquifex aeolicus*, *Thermotoga maritime*, and *Methanococcus jannaschii*, respectively. As also can be seen in tables 25 through 29, the three thermophiles also have more glycine and threonine residues than the hyperthermophiles.

Table 29



3. Comparison of *Bacillus stearothermophilus* and *Bacillus subtilis* Amino Acid Composition

In the amino acid comparison between *Bacillus stearothermophilus* and *Bacillus subtilis*, shown in table 30, the percentage of glutamine residues was similar with 3.55 percent in *Bacillus stearothermophilus* and 3.84 percent in *Bacillus subtilis*. The highest difference in relative amino acid compositions was observed in the relative arginine amounts in *Bacillus stearothermophilus* and *Bacillus subtilis*, where *Bacillus stearothermophilus* contained 6.94 percent arginine, and *Bacillus subtilis* contained 4.13 percent. The four hydrophobic residues that had higher percentages in the thermophile versus the mesophile were alanine, proline, valine, and tryptophan. Since thermophiles and the mesophiles have G+C content of 52.75 percent and 43.50 percent, respectively, it has been postulated that the differences in G+C content affects the amino acid

composition of species and particularly, the glycine, alanine, proline, and arginine residues, which are found at a much higher frequency in the predicted ORFs from G-C rich genomes (98).

Table 30-Comparison of the Relative Amino Acid Composition between *Bacillus stearothermophilus* and *Bacillus subtilis*

	B. stearothermophilus	*B. subtilis*	Difference
Charged Amino Acids			
aspartic acid	4.77	5.18	0.41
glutamic acid	7.06	7.24	0.18
histidine	2.39	2.27	0.12
lysine	5.46	7.05	1.59
arginine	6.94	4.13	2.81
Polar/uncharged AA			
cysteine	1.09	0.8	0.29
glycine	7.41	6.91	0.5
asparagine	3.06	3.95	0.89
glutamine	3.55	3.84	0.29
serine	5.3	6.29	0.99
threonine	5.01	5.42	0.41
tyrosine	3.16	3.48	0.32
Hydrophobic AA			
alanine	9.3	7.68	1.62
phenylalanine	4.19	4.5	0.31
isoleucine	6.06	7.36	1.3
leucine	9.56	9.65	0.09
methionine	2.46	2.78	0.32
proline	4.54	3.68	0.86
valine	7.38	6.75	0.63
tryptophan	1.31	1.03	0.28

4. Comparison of *Bacillus stearothermophilus*, *Bacillus subtilis*, and *Bacillus halodurans* Amino Acid Composition

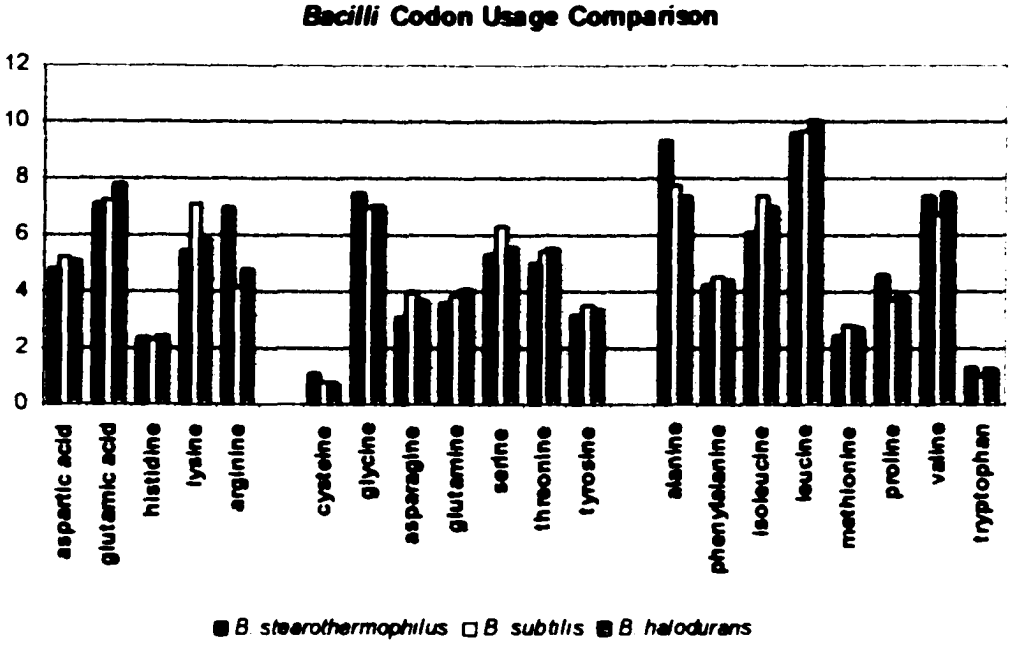
As shown in table 31 and 32, *Bacillus stearothermophilus* has the highest relative amino acid percentages in arginine, cysteine, glycine, alanine, and proline among the three *Bacilli* species. The most striking differences were observed for arginine with *Bacillus stearothermophilus*, *Bacillus subtilis*, and *Bacillus halodurans* having

percentages of 6.94, 4.13, and 4.82, respectively. Another observation is in the hydrophobic amino acid alanine with *Bacillus stearothermophilus*, *Bacillus subtilis*, and *Bacillus halodurans* having percentages of 9.3, 7.68, and 7.36, respectively

Table 31-Comparison of the Relative Amino Acid Composition between *Bacillus stearothermophilus*, *Bacillus subtilis*, *and Bacillus halodurans*

	B stearothermophilus	*B subtilis*	*B halodurans*
Charged AA	**26.62**	**26.87**	**26.96**
aspartic acid	4 77	5 18	5 06
glutamic acid	7 06	7 24	7 82
histidine	2.39	2 27	2 41
lysine	5 46	7 05	5 85
arginine	6 94	4 13	4 82
Polar/uncharged AA	**28.58**	**30.69**	**29.93**
cysteine	1 09	0 8	0 73
glycine	7 41	6 91	6 99
asparagine	3 06	3 95	3 61
glutamine	3 55	3 84	4 1
serine	5 3	6 29	5 6
threonine	5 01	5 42	5 53
tyrosine	3 16	3 48	3 37
Hydrophobic AA	**44.8**	**43.43**	**43.84**
alanine	9 3	7 68	7 36
phenylalanine	4 19	4 5	4 44
isoleucine	6 06	7 36	6 91
leucine	9 56	9 65	9 98
methionine	2 46	2 78	2 68
proline	4 54	3 68	3 8
valine	7 38	6 75	7 46
tryptophan	1 31	1 03	1 21

Table 32


Bacilli Codon Usage Comparison
12
10
8
6
4
2
0
aspartic acid
glutamic acid
histidine
lysine
arginine
cysteine
glycine
asparagine
glutamine
serine
threonine
tyrosine
alanine
phenylalanine
isoleucine
leucine
methionine
proline
valine
tryptophan
B. stearothermophilus
B. subtilis
B. halodurans

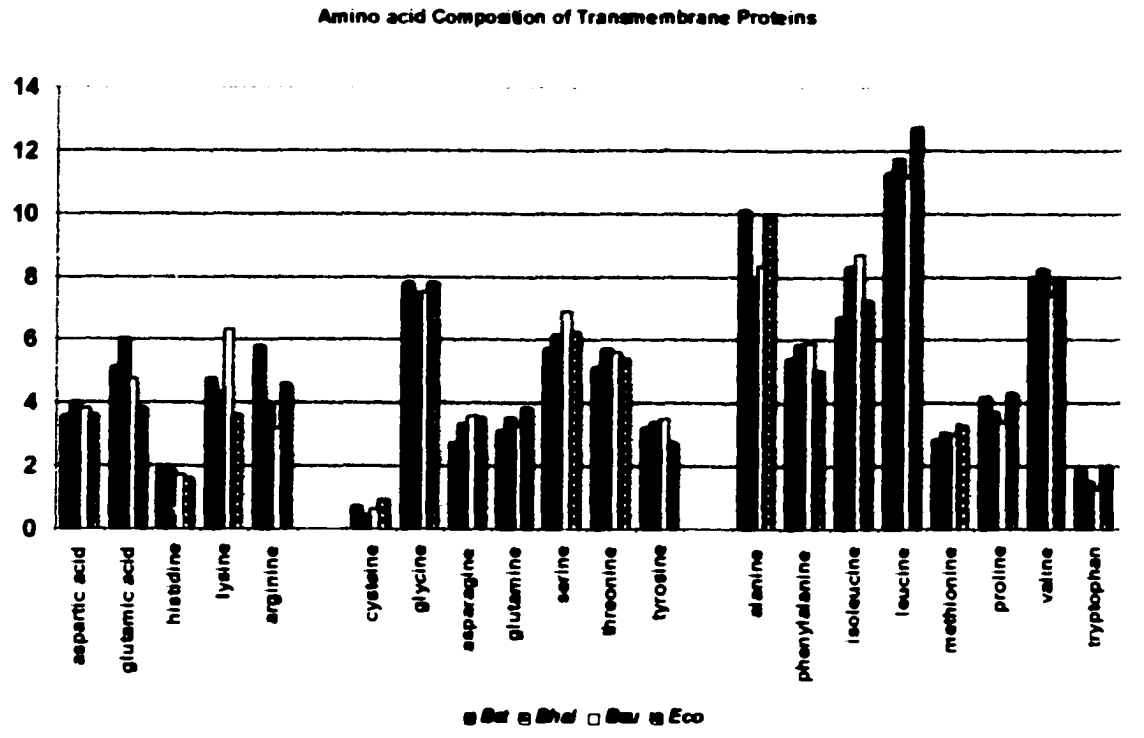
5. Transmembrane Helices Relative Amino Acid Comparison

The TMHMM2.0 program is used to predict transmembrane helices in proteins (84) Dr. Fares Najar wrote a Perl script to calculate codon usage for each ORF that has at least one transmembrane helix. In tables 33 and 34, the relative amino acid composition of *Bacillus stearothermophilus* is compared with three mesophiles *Bacillus stearothermophilus* has the highest percent composition of amino acids arginine, histidine, and alanine in the transmembrane helices. *Bacillus subtilis* has the highest percent composition of lysine, asparagine, serine, tyrosine, phenylalanine, and isoleucine. The acidophile, *Bacillus halodurans*, has the highest percent composition of aspartic acid, glutamic acid, threonine, and valine. *E. coli* has the highest percent composition of cysteine, glutamine, leucine, proline, and tryptophan.

Table 33-Amino Acid Composition Comparison of Transmembrane Proteins between *Bacillus stearothermophilus* and Three Mesophiles

	Bst	*Bhal*	*Bsu*	*Eco*
Charged Amino Acids	**21.2**	**20**	**19.8**	**17.2**
aspartic acid	3.6	4	3.8	3.6
glutamic acid	5.1	6	4.8	3.8
histidine	1.9	1.8	1.7	1.6
lysine	4.8	4.3	6.3	3.6
arginine	5.8	3.9	3.2	4.6
Polar/uncharged AA	**28.3**	**29.8**	**31**	**30.3**
cysteine	0.7	0.4	0.6	0.9
glycine	7.8	7.4	7.5	7.8
asparagine	2.7	3.3	3.6	3.5
glutamine	3.1	3.5	3.3	3.8
serine	5.7	6.1	6.9	6.2
threonine	5.1	5.7	5.6	5.4
tyrosine	3.2	3.4	3.5	2.7
Hydrophobic AA	**50.4**	**50**	**49.2**	**52.3**
alanine	10.1	7.8	8.3	9.9
phenylalanine	5.4	5.8	5.9	5
isoleucine	6.7	8.3	8.7	7.2
leucine	11.3	11.7	11.2	12.7
methionine	2.8	3	3	3.3
proline	4.2	3.7	3.4	4.3
valine	8	8.2	7.4	7.9
tryptophan	1.9	1.5	1.3	2

Table 34



In table 35, the relative amino acid composition of transmembrane helices from *Bacillus stearothemophilus*, three mesophiles, and three thermophiles is graphed As seen in table 34 and 35, *Bacillus stearothermophilus* has the highest percent of arginine, histidine, and alanine residues *Bacillus subtilis* has the highest percent of serine residues only in table 35 *Bacillus halodurans* has the highest percent of threonine residues only in both tables 34 and 35 As shown in tables 34 and 35, *E. coli* has the highest percent of glutamine, glutamic acid, methionine, tryptophan, and proline residues in its predicted transmembrane proteins, when compared to the various completed genomes. *Aquifex aeolicus* has the highest percent of leucine, phenylalanine and lysine residues. *Thermotoga maritima* has the highest percent of valine residues *Methanococcus jannaschii* has the highest percent of tyrosine, asparagines, and isoleucine.

Table 35

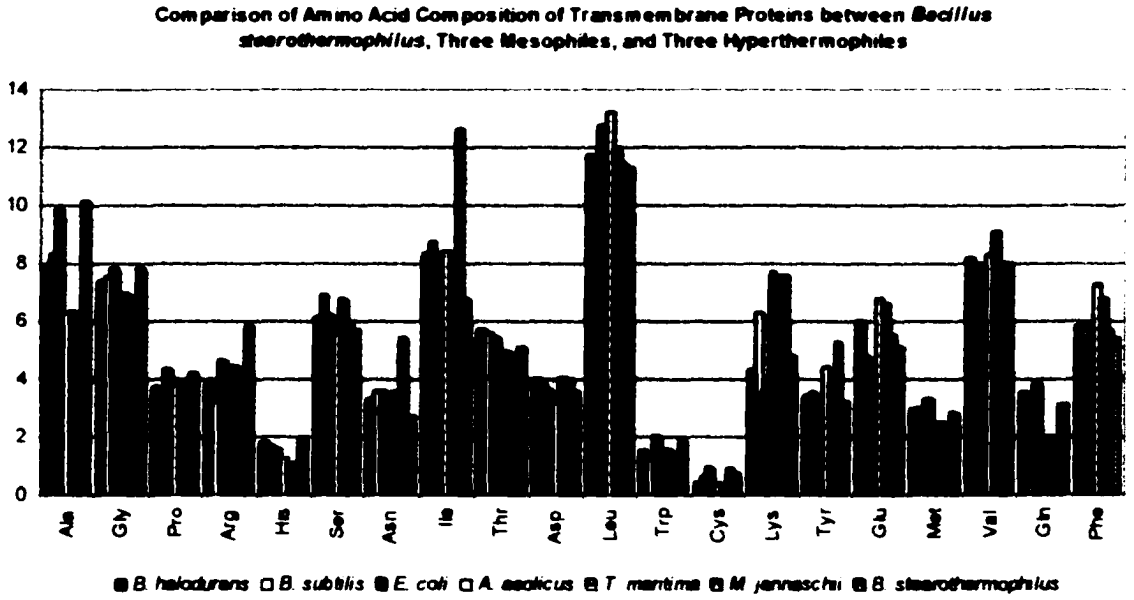


Table 36

	Bha	*Bsu*	*Eco*	*Aaeol*	*Tmari*	*Mjan*	*Bstearo*
Ala	7.8	8.3	9.9	6.3	6.3	5.7	10.1
Gly	7.4	7.5	7.8	6.7	6.9	6.8	7.8
Pro	3.7	3.4	4.3	3.8	3.8	3.2	4.2
Arg	3.9	3.2	4.6	4.1	4.4	2.5	5.8
His	1.8	1.7	1.6	1.3	1	1.1	1.9
Ser	6.1	6.9	6.2	5.6	6.7	5.9	5.7
Asn	3.3	3.6	3.5	3.2	3.5	5.4	2.7
Ile	8.3	8.7	7.2	8.4	7.8	12.6	6.7
Thr	5.7	5.6	5.4	4.5	4.9	4.8	5.1
Asp	4	3.8	3.6	3.2	4	4	3.6
Leu	11.7	11.2	12.7	13.2	11.9	11.4	11.3
Trp	1.5	1.3	2	1.1	1.5	1	1.9
Cys	0.4	0.6	0.9	0.4	0.3	0.9	0.7
Lys	4.3	6.3	3.6	7.7	6.1	7.6	4.8
Tyr	3.4	3.5	2.7	4.4	4	5.3	3.2
Glu	6	4.8	3.8	6.8	6.6	5.5	5.1
Met	3	3	3.3	2	2.5	2.4	2.8
Val	8.2	7.4	7.9	8.3	9.1	7	8
Gln	3.5	3.3	3.8	1.8	2	1.3	3.1
Phe	5.8	5.9	5	7.3	6.8	5.7	5.4

6. Relative Amino Acid Comparison of Heat Shock Proteins and Kinase Proteins

The amino acid comparison for three mesophiles, three thermophiles, and three hyperthermophiles was performed. *Bacillus stearothemophilus* had the highest percentage of histidine residues among heat shock proteins, as displayed in table 37 and 38, and kinase proteins, as displayed in table 39 and 40. *Thermoplasma volcanium* had the highest percentage of serine residues among the two types of proteins. That ends the consistency among the thermophiles. Among the hyperthermophiles, *Aquifex aeolicus* had the highest percentage of glutamine residues among all the organisms in both types of proteins. *Methanococcus jannaschii* had the highest percentage of asparagine residues in both types of proteins. That ends the consistency among the hyperthermophiles. Among the mesophiles, the only consistency observed was in *E. coli*, which had the highest percentage of alanine residues among the two heat shock and kinase proteins.

Table 37

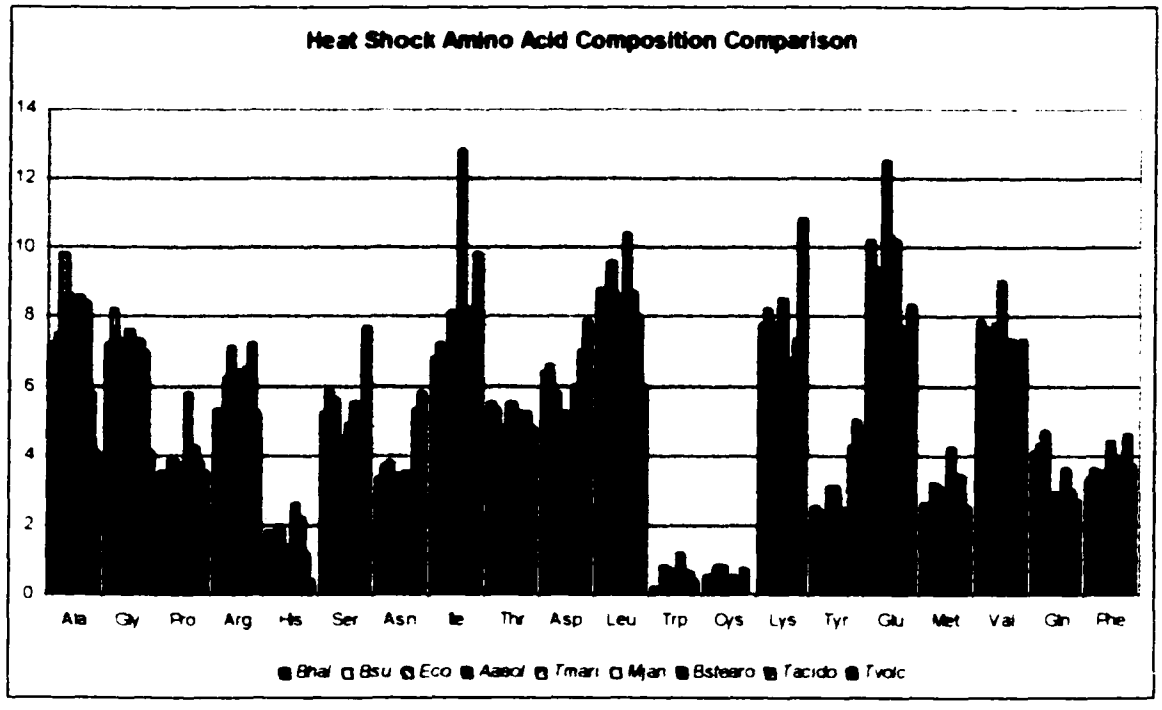


Table 38-Comparison of the Heat Shock Proteins for Three Mesophiles, Three Thermophiles, and Three Hyperthermophiles

	B. hal.	*B. subtilis*	*E. coli*	*A. aeolicus*	*T. maritima*	*M. jan.*	*B. stearo.*	*T. acido.*	*T. volcanium*
Ala	72	75	98	86	58	86	84	58	41
Gly	72	82	73	69	76	56	73	7	41
Pro	35	29	39	37	37	58	42	38	35
Arg	53	43	62	71	64	44	65	72	52
His	18	15	19	14	07	26	22	12	04
Ser	52	59	56	29	45	49	55	48	77
Asn	33	37	38	31	34	35	34	53	58
Ile	68	72	57	81	79	128	82	74	98
Thr	55	53	48	37	55	51	52	52	48
Asp	64	66	58	42	52	21	6	7	79
Leu	88	83	96	86	82	104	87	8	6
Trp	02	02	08	07	06	12	06	06	04
Cys	05	05	08	08	02	05	04	07	0
Lys	78	82	6	78	85	56	67	73	108
Tyr	25	24	21	31	31	23	24	42	5
Glu	102	94	75	125	103	102	77	68	83
Met	26	22	32	31	21	42	23	34	25
Val	79	77	69	78	9	49	73	68	73
Gln	41	43	47	26	29	16	36	3	27
Phe	33	36	35	35	44	39	34	46	37

Table 39

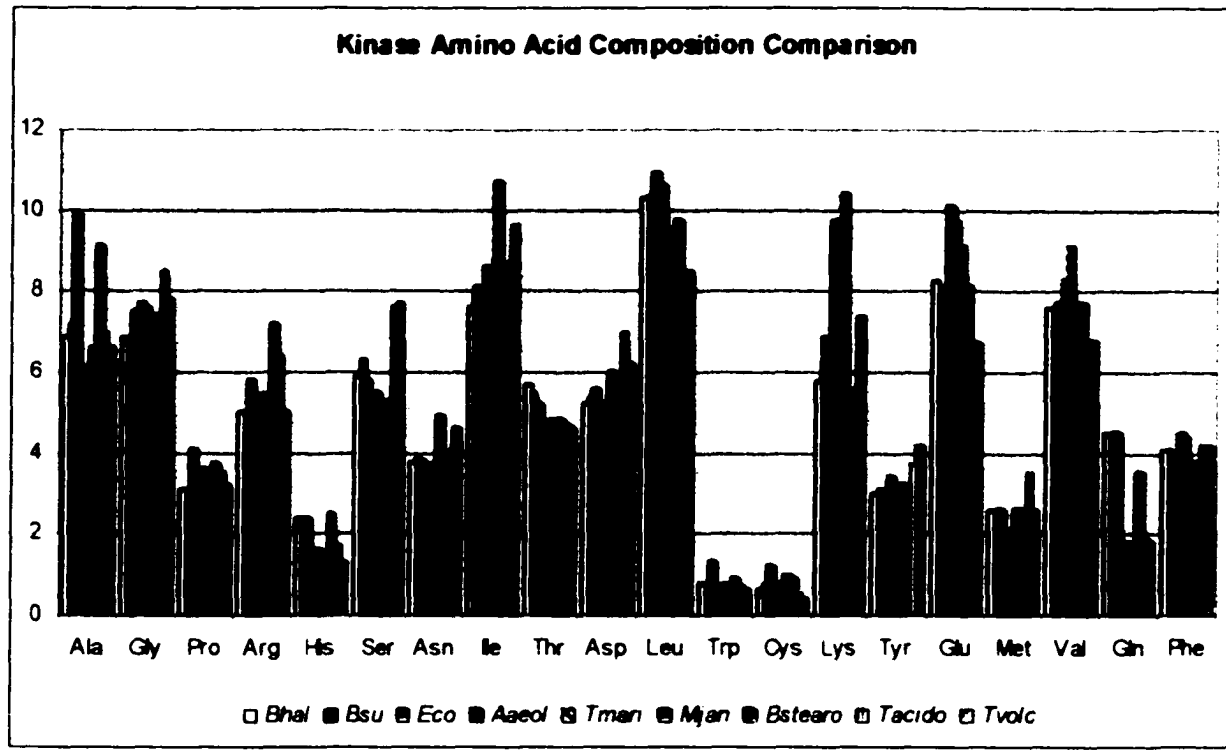


Table 40-Comparison of the Kinase Proteins for Three Mesophiles, Three Thermophiles, and Three Hyperthermophiles

	Bhal	B.subtilis	E.coli	Aaeol	Tmari	Mjan	Bstearo	Tacido	T. volcanium
Ala	69	72	99	59	61	66	91	7	66
Gly	69	68	75	77	76	73	74	85	78
Pro	31	3	41	34	36	28	37	35	32
Arg	5	43	58	49	55	36	72	64	5
His	24	24	24	14	16	12	25	17	13
Ser	59	63	58	49	55	48	53	76	77
Asn	38	39	38	35	37	49	29	41	46
Ile	76	81	59	86	75	107	67	87	96
Thr	57	55	52	39	48	48	48	47	46
Asp	52	54	56	45	52	6	51	7	62
Leu	103	98	109	106	95	87	97	74	85
Trp	08	08	13	06	08	04	09	07	06
Cys	06	08	12	08	07	1	09	05	04
Lys	58	69	4	97	8	104	54	56	74
Tyr	3	31	24	34	29	32	25	37	42
Glu	83	81	64	101	97	91	81	66	67
Met	26	26	26	15	21	26	26	35	26
Val	76	68	77	83	91	75	77	68	68
Gln	45	4	45	18	18	1	35	19	18
Phe	41	41	32	45	44	35	38	42	42

7. Di- and Trinucleotide Frequencies

The di- and trinucleotide frequencies of the thermophilic *Bacillus stearothermophilus* and three mesophilic genomes were studied. The two most common dinucleotide frequencies as shown in table 41 and 42 in *Bacillus stearothermophilus* are cg and gc which is expected since thermophiles grow at higher temperatures than mesophiles and thus need stronger bonds resulting in higher G-C percentages as compared to mesophiles. Other researchers suggested that the clustered G-C's act as clamps to improve double helical stability (113). The actual percentages for cg in *Bacillus stearothermophilus* is 9.40 whereas in *Bacillus halodurans*, *Bacillus subtilis*, and *E. coli* the percentages are 5.20, 4.90, and 7.50 respectively. The actual percentages for gc in *Bacillus stearothermophilus* is 8.60 as compared to *Bacillus halodurans*, *Bacillus subtilis*, and *E. coli* the percentages are 5.20, 6.00, and 8.30 respectively

Table 41

Dinucleotide Frequencies of *Bacillus stearothermophilus* and Three Mesophiles

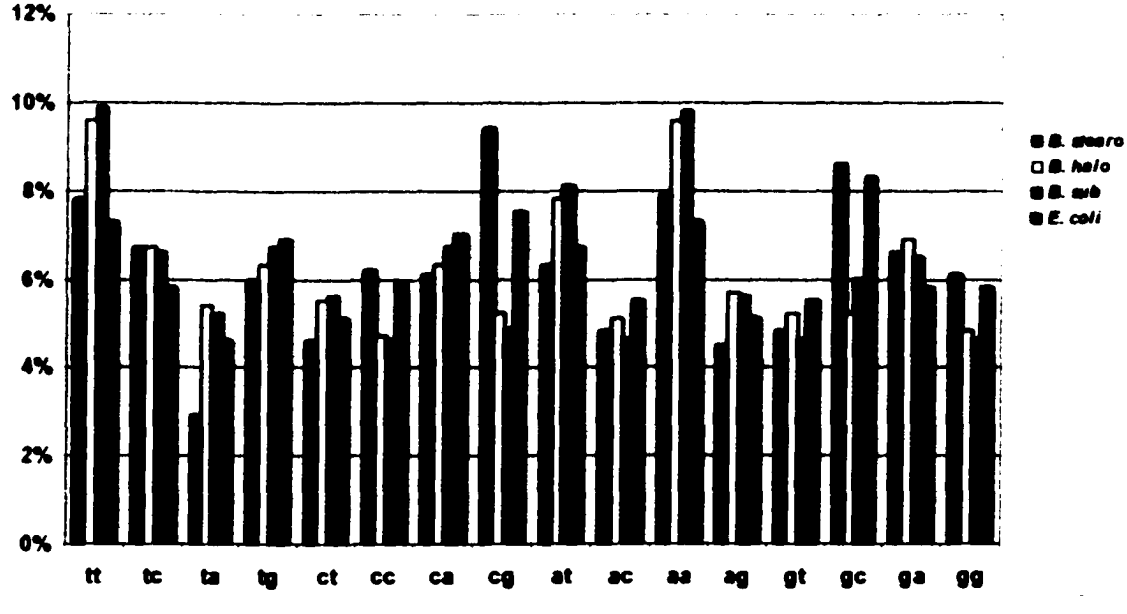


Table 42-Dinucleotide Frequencies for *Bacillus stearothermophilus* versus Three Mesophiles

	B. stearo	*B. halo*	*B. sub*	*E. coli*
tt	7.80%	9.60%	9.90%	7.30%
tc	6.70%	6.70%	6.60%	5.80%
ta	2.90%	5.40%	5.20%	4.60%
tg	6.00%	6.30%	6.70%	6.90%
ct	4.60%	5.50%	5.60%	5.10%
cc	6.20%	4.70%	4.60%	5.90%
ca	6.10%	6.30%	6.70%	7.00%
cg	9.40%	5.20%	4.90%	7.50%
at	6.30%	7.80%	8.10%	6.70%
ac	4.80%	5.10%	4.60%	5.50%
aa	7.90%	9.60%	9.80%	7.30%
ag	4.50%	5.70%	5.60%	5.10%
gt	4.80%	5.20%	4.60%	5.50%
gc	8.60%	5.20%	6.00%	8.30%
ga	6.60%	6.90%	6.50%	5.80%
gg	6.10%	4.80%	4.60%	5.80%

The most common trinucleotide frequencies as shown in tables 43 and 44 in *Bacillus stearothermophilus* are not as large a percentage as the dinucleotide frequencies. Surprisingly, the highest trinucleotides are ttt and aaa. However, as expected, where there are three g-c nucleotides, the percentage is high in the thermophile such as in cgc and gcg with 2.80 percent, and also ccg and cgg with 2.50 percent as compared to the mesophile.

Table 43

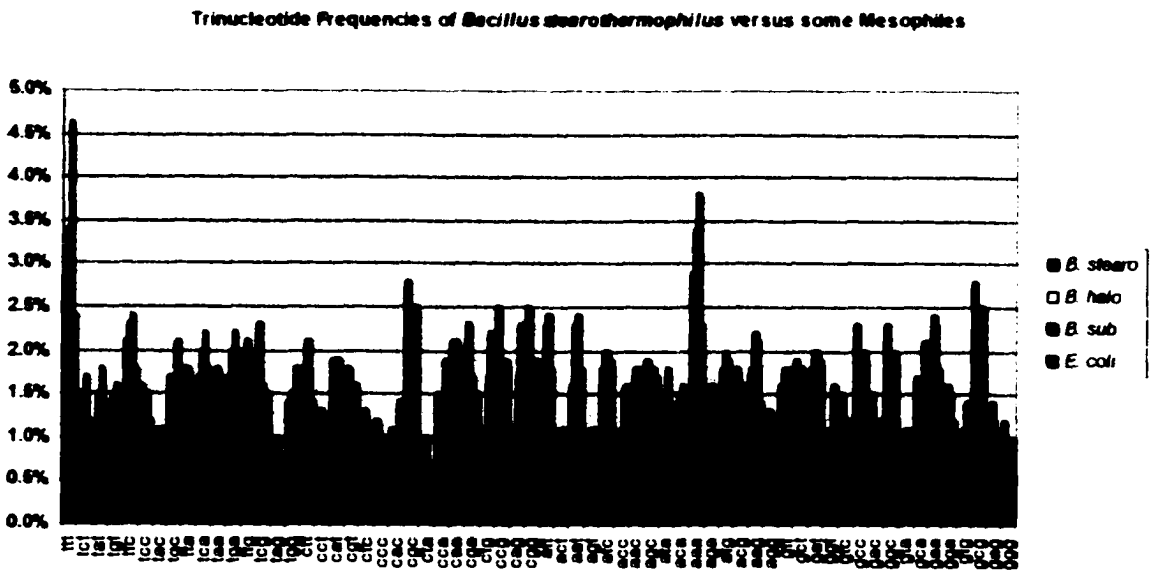


Table 44-Trinucleotide Frequencies for *Bacillus stearothermophilus* versus Three Mesophiles

	B. stearo	B. halo	B. sub	E. coli
ttt	2.90%	3.40%	4.62%	2.40%
tct	1.00%	1.50%	1.70%	1.20%
tat	1.00%	1.50%	1.80%	1.40%
tgt	1.30%	1.50%	1.60%	1.30%
ttc	2.10%	2.30%	2.40%	1.80%
tcc	1.60%	1.50%	1.50%	1.20%
tac	0.70%	1.10%	1.00%	1.10%
tgc	1.70%	1.40%	1.60%	2.10%
tta	0.80%	1.80%	1.70%	1.50%
tca	1.70%	2.00%	2.20%	1.80%
taa	0.80%	1.80%	1.70%	1.50%
tga	1.70%	2.00%	2.20%	1.80%
ttg	2.00%	2.10%	2.00%	1.70%
tcg	2.30%	1.60%	1.20%	1.50%
tag	0.40%	1.00%	0.70%	0.60%
tgg	1.40%	1.50%	1.20%	1.80%
ctt	1.80%	2.10%	2.10%	1.40%
cct	1.00%	1.30%	1.20%	1.10%
cat	1.90%	1.90%	1.90%	1.70%
cgt	1.80%	1.40%	1.10%	1.60%
ctc	1.20%	1.30%	1.10%	0.90%
ccc	1.20%	0.90%	0.80%	1.00%
cac	1.10%	1.10%	1.00%	1.40%
cgc	2.80%	1.00%	1.20%	2.50%
cta	0.40%	1.00%	0.70%	0.60%
cca	1.50%	1.40%	1.20%	1.90%

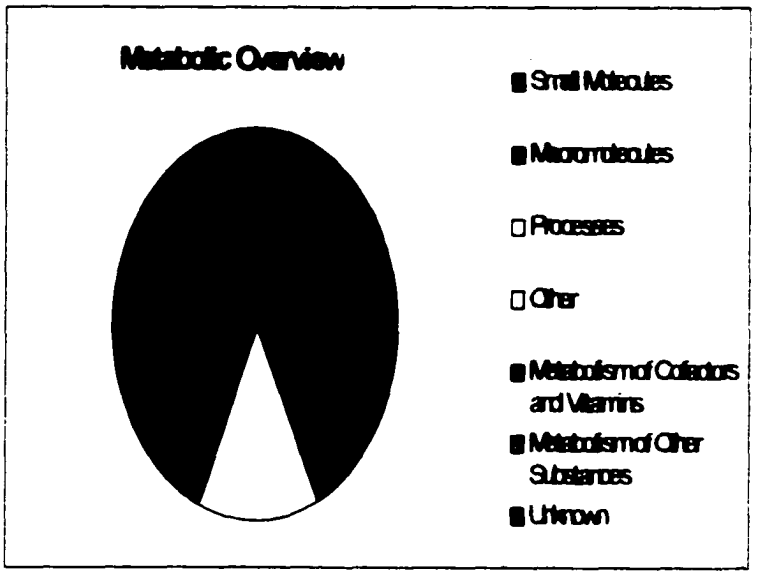
caa	2.10%	2.10%	2.00%	1.70%
cga	2.30%	1.70%	1.20%	1.50%
ctg	1.10%	1.10%	1.70%	2.20%
ccg	2.50%	1.10%	1.40%	1.90%
cag	1.10%	1.10%	1.70%	2.30%
cgg	2.50%	1.10%	1.40%	1.90%
att	1.60%	2.30%	2.40%	1.80%
act	0.70%	1.10%	1.00%	1.10%
aat	1.60%	2.30%	2.40%	1.80%
agt	0.70%	1.10%	0.90%	1.10%
atc	1.80%	2.00%	2.00%	1.90%
acc	1.00%	1.10%	0.90%	1.60%
aac	1.60%	1.80%	1.60%	1.80%
agc	1.90%	1.70%	1.80%	1.70%
ata	1.00%	1.50%	1.80%	1.40%
aca	1.30%	1.50%	1.60%	1.30%
aaa	2.90%	3.40%	3.80%	2.30%
aga	1.00%	1.60%	1.60%	1.20%
atg	1.80%	2.00%	1.90%	1.60%
acg	1.80%	1.40%	1.10%	1.60%
aag	1.80%	2.20%	2.10%	1.40%
agg	1.00%	1.30%	1.20%	1.10%
gtt	1.60%	1.80%	1.60%	1.80%
gct	1.90%	1.60%	1.80%	1.70%
gat	1.80%	2.00%	2.00%	1.90%
ggt	1.00%	1.10%	1.00%	1.60%
gtc	1.50%	1.10%	1.10%	1.20%
gcc	2.30%	1.10%	1.40%	2.00%
gac	1.50%	1.20%	1.10%	1.20%
ggc	2.30%	1.10%	1.30%	2.00%
gta	0.70%	1.10%	1.00%	1.10%
gca	1.70%	1.40%	1.60%	2.10%
gaa	2.10%	2.40%	2.30%	1.80%
gga	1.60%	1.60%	1.50%	1.20%
gtg	1.00%	1.10%	1.00%	1.40%
gcg	2.80%	1.10%	1.20%	2.50%
gag	1.20%	1.40%	1.10%	0.90%
ggg	1.20%	1.00%	0.80%	1.00%

Chapter IV: Conclusion

The major objective of this research was to determine the working draft sequence of the *Bacillus stearothermophilus* genome, and thereby obtain the sequence of its genomic features including, its ORFs, codon usage, operon structures, repeated elements, and nucleotide frequency distribution in an effort to characterize this organism. This goal was accomplished using a whole-genome shotgun approach, where over ninety-nine percent of the approximately 3.2 Mbp *Bacillus stearothermophilus* genome was sequenced to working draft stage, assembled, and annotated. The work presented in this dissertation confirms that the whole genome shotgun approach is a viable method for obtaining a thorough picture of the genetic information encoded in genomic DNA.

Of the approximately 2,731 genes discovered in the *Bacillus stearothermophilus* genome, slightly over 1500 are conserved in *Bacillus subtilis* and *Bacillus halodurans*. These represent genes involved in typical housekeeping functions, such as, glycolysis, the TCA cycle, metabolism of carbohydrates, nucleotides, nucleic acids, coenzymes and prosthetic groups, as well as, DNA replication, RNA modification, protein synthesis, cell mobility, chemotaxis, sigma factors, and many of the same proteins needed for competence and sporulation.

Figure 52



Two differences are noted between the *Bacillus stearothermophilus* and *Bacillus subtilis* metabolic pathways, the first being that *Bacillus stearothermophilus* lysine biosynthesis is by way of the succinylated pathway not the acetylated pathway used by *Bacillus subtilis*. Another difference is that B*acillus stearothermophilus* encodes for seventeen of the proteins involved in the aerobic synthesis of cobalamin, while *Bacillus subtilis* strain 168 does not synthesize this vitamin. Cobalamin (vitamin B12) is involved in rearrangement reactions. An interesting observation between the *Bacilli* genomes is that the *Bacillus stearothermophilus* genome contains 106 transposase genes as compared to 10 in *Bacillus subtilis* genome analysis and 112 in *Bacillus halodurans*. This observation may explain why the *Bacillus subtilis* genome is more stable than the genomes of both the thermophilic and halophilic *Bacilli* that grow either at high temperatures or under high salt conditions, respectively. Thus one mechanism possibly related to allowing these organisms to cope with stressed conditions, i.e., growth at high temperatures or growth under high salt conditions, is the ability of their genomes to reorganize in response to stressful ecosystem.

Another goal of this research was to determine if analysis of the working draft sequence of this thermophilic organism genome would help illuminate the biochemical

foundation for thermophily. The four major mechanisms that have been presented to explain themophily are the kinetic theory, the lipid theory, and the macromolecular theory (8) and a combined hypothesis (9). The kinetic theory states that thermophilic bacteria have a special metabolic state characterized by high rates of synthesis and degradation of proteins and nucleic acids. The lipid theory for explaining thermophily states that there are high melting point lipids on macromolecules and subcellular structures of thermophilic bacteria. The macromolecular theory states that there are physical-chemical differences (3 dimensional protein structure, G+C content, arginine percent composition) in the structure and function of proteins, enzymes, nucleic acids and nucleoproteins for thermophiles as compared to mesophiles. In a combined hypothesis, Hendrix and Welker proposed that thermophily in most proteins may not be due to the whole structural properties of these proteins but rather based on the cumulative effect of many amino acid substitutions throughout the protein without a striking change in conformation.

The work presented in this dissertation supports and extends the combined hypothesis since the evidence now indicates that this hypothesis should be extended to include all the major components in thermophilic organisms, including aspects of the carbohydrate associated kinetic theory, the lipid theory, the macromolecular theory, and the protein-based combined theory, as well as, the genomic sequence organization and the amino acid composition and sequence of the proteins.

Previous work in comparing the relative amino acid composition generated from codon usage of mesophiles and thermophiles established trends that could be correlated with the thermostability of the encoded proteins (38). In this dissertation research, when the relative amino acid composition of mesophiles (*H. influenzae, H. pylori, E. coli, Synechosystis, B. halodurans,* and *B. subtilis*), thermophiles (*B. stearothermophilus, T. acidophilum, T. volcanium*), and hyperthermophiles (*A. aeolicus, T. maritima, and M. jannaschii*) were compared, these trends were confirmed and extended to additional species. For example, the genomes of hyperthermophilic organisms do encode proteins

with slightly more charged amino acid residues than mesophilic genomes at the expense of polar/uncharged residues, while the genomes of thermophiles and mesophiles encode proteins with approximately the same relative amount of hydrophobic residues. In addition, glutamine is present at a much lower percentage in thermophiles and hyperthermophiles when compared to mesophiles, while thermophiles have less asparagine (an amino acid that is thermal labile) than mesophiles.

Further analysis of the amino acid composition of predicted proteins for thermophiles and hyperthermophiles reveals that *Bacillus stearothermophilus* has a lower percentage of lysine, and a higher percentage of arginine, histidine, glutamine, and alanine residues as compared to the three hyperthermophiles. When compared with *Bacillus subtilis*, thermophiles have higher percentages of arginine, alanine, proline, and valine residues This agrees with previous research stating that glycine, alanine, proline, and arginine residues are found at a much higher percentage in the ORFs from G-C rich genomes Analysis of the amino acid distribution of *Bacillus stearothermophilus*, *Bacillus subtilis*, and *Bacillus halodurans* indicates that *Bacillus stearothermophilus* has higher percentages of arginine, cysteine, glycine, alanine, and proline percentages.

The highest difference in relative amino acid composition between *Bacillus stearothermophilus* and *Bacillus subtilis* was observed for the arginine residues with a percentage of 6.94 in the thermophile and 4.13 in the mesophile. Although the intracellular concentration of glutamine and arginine in these organisms is not known, it has been reported that glutamine is the most efficient nitrogen source for *Bacillus subtilis* followed by arginine since these two amino acids preferentially support rapid cell growth (24). Since the amounts of glutamine are approximately equal in *Bacillus stearothermophilus* and *Bacillus subtilis*, it may be that the high level of arginine in *Bacillus stearothermophilus* could be a major component in support of the kinetic theory, which as discussed above, states that thermophilic bacteria have a special metabolic state characterized by high rates of synthesis and degradation of proteins and

nucleic acids since, as stated above, rapid cell growth must lead to rapid rates of synthesis and degradation. The high arginine percentage in Bacillus stearothermophilus supports the observation of more ionic interactions in thermophiles versus mesophiles

The major fatty acid components of *Bacillus subtilis* are anteiso-C15:0 and iso-C15:0 (24), while the major fatty acid components present in *Bacillus stearothermophilus* DSM 22 are iso-15:0, iso-16:0, iso-17:0 (4). This observation supports the lipid theory of thermophily, since a greater proportion of high melting point fatty acids exist in the lipids of thermophilic bacteria.

Finally, analysis of di- and trinucleotide frequencies revealed direct results that are consistent with the earlier nearest neighbor frequency analysis studies (101). Since *Bacillus stearothermophilus* has 52.75 percent G-C content, the highest frequencies are among the dinucleotides cg, and gc; and the highest trinucleotide frequencies are cgc, gcg, ccg, cgg, gcc, and ggc. Thus it is likely that the clustered G-C's act as clamps to improve double helical stability has been previously postulated

Through the analysis of the genomic sequence of *Bacillus stearothermophilus*, an expanded unified theory of thermophily emerges in which a combination of features, suggested as trends in the above discussion, provides the thermal advantage to the genome, its encoded proteins, its lipids, amino acids, and other cellular components. Thus, a thermophilic organism exists and functions optimally at elevated temperatures because its component compounds, its macromolecules, and its overall metabolism have evolved small changes in their properties that when accumulated, give the organism its selective advantage and flourish under those harsh conditions.

In this dissertation, I have presented the amino acid sequence, the metabolic pathways, and the gene organization for as much as is possible with the sequence still in the closure phase. To actually delineate why this organism grows at 65-70 degrees Celsius, researchers must analyze the three dimensional structure of the approximately 2,731 ORFs identified in this research. The science of proteomics developed as a result

from the science of genomics.

Chapter V: Literature Cited

1. Stenesh, J., Roe, B.A., Snyder T.L., "Studies of the Deoxyribonucleic Acid From Mesophilic and Thermophilic Bacteria," *Biochimica ET Biophysica ACTA*, vol. 161 p. 442-454 (1968).

2. Protocols For Recombinant DNA Isolation, Cloning, and Sequencing, Roe, B.A. (1997).

3. Donk, P.J., "A Highly Resistant Thermophilic Organism," *Journal of Bacteriology*, vol. V, p. 373-374 (1920).

4. Nazina, T.N., Tourova, T.P., Poltaraus, A.B., Novikova, E.V., Grigoryan, A.A., Ivanova, A.E., Lysenko, A.M., Petrunyaka, V.V., Osipov, G.A., Belyaev, S.S., Ivanov, M.V., "Taxonomic Study of Aerobic Thermophilic *Bacilli*: Descriptions of *Geobacillus subterraneus* gen. nov., sp. Nov. and *Geobacillus uzenensis* sp. nov. from Petroleum Reservoirs and Transfer of *Bacillus stearothermophilus*, *Bacillus thermocatenulatus*, *Bacillus thermoleovorans*, *Bacillus kaustophilus*, *Bacillus thermoglucosidasius* and *Bacillus thermodenitrificans* to *Geobacillus* as the new combinations *G. stearothermophilus*, *G. thermocatenulatus*, *G. thermoleovorans*, *G. kaustophilus*, *G. thermoglucosidasius* and *G. thermodenitrificans*," *International Journal of Systematic and Evolutionary Microbiology*, vol. 51, p. 433-446 (2001).

5. Microbiology and Infectious Diseases, 3rd edition, Virella, G., Williams & Wilkins, Chapter 1, p.3-19 (1997).

6. Bergey's Manual of Determinative Bacteriology, 8th edition, (ed. Buchanan, R.E., Gibbons, N.E.) Baltimore, Maryland, The Williams and Wilkins Company, p. 539-540 (1974).

7. Microbial Life in Extreme Environments, (ed. D.J. Kushner) Academic Press, Chapter 6, p. 217-278 (1978).

8. Amelunxen, R.E., Murdock, A.L., "Mechanisms of Thermophily," *CRC Critical Reviews of Microbiology*, vol.6, no.4, p.343-393 (1978).

9. Hendrix, J.D., Welker, N.E., "Isolation of a *Bacillus stearothermophilus* Mutant Exhibiting Increased Thermostability in Its Restriction Endonuclease," *Journal of Bacteriology*, vol. 162, no. 2, p. 682-692 (1985).

10. The Bacterial Spore, (ed. G.W. Gould and A. Hurst) Academic Press, New York, p. 129 (1968).

11. http://www.bacillus.bio-sci.ohio-state.edu

12. Mardis, E.R., Roe, B.A., "Automated Methods for Single-Stranded DNA Isolation and Dideoxynucleotide DNA Sequencing Reactions on a Robotic Workstation,"

Biotechniques, vol 7, no 8, p 840-850 (1988)

13 Mead, D A., McClary, J A., Luckey, J A., Kostichka, A J., Witney, F R., Smith, L M., "**Bst** DNA polymerase Permits Rapid Sequence Analysis from Nanogram Amounts of Template," *Biotechniques*, vol 11, no 1, p 76-88 (1991)

14 http www foodproductdesign com

15 http://www enzymebio com

16 http://www vm cfsan fda gov

17 Microbial Enzymes and Biotechnology, , (ed Fogarty, W M) Applied Science Publishers, New York, NY, Chapter 1, p 5 (1983)

18 Akiba, T, Horikoshi, K., "Xylanases of Alkalophilic Thermophilic *Bacillus*," *Methods in Enzymology*, vol 160, p 655-659 (1988)

19 http www namsa com

20 http sciencexpress org 9May2002/

21 Moszer, I , "The Complete Genome of *Bacillus subtilis* From Sequence Annotation to Data Management and Analysis," *FEBS Letters*, vol 430, p 28-36 (1998)

22 Takami, H , Nakasone, K , Takaki, Y , Maeno, G , Sasaki, R , Masui, N , Fuji, F , Hirama, C , Nakamura, Y , Ogasawara, N , Kuhara, S , Horikoshi, K , "Complete Genome Sequence of the Alkaliphilic Bacterium *Bacillus halodurans* and Genomic Sequence Comparison with *Bacillus subtilis*," *Nucleic Acids Research*, vol 28 no 21, p 4317-4331 (2000)

23 http www sanger ac uk

24 *Bacillus subtilis* and Its Closest Relatives, (ed Sonenshein, A L , Hoch, J A., Losick, R) American Society for Microbiology, Washington, D C (2002)

25 Nelson, K E., Paulson, I T , Heidelberg, J F., Fraser, C M., "Status of Genome Projects for Nonpathogenic Bacteria and Archaea," *Nature Biotechnology*, vol 18, p 1049-1054 (2000)

26 Molecular Biological Methods for *Bacillus*, (ed Harwood, C R., Cutting, S M) John Wiley & Sons New York (1990)

27 Harwood, C R., "*Bacillus subtilis* and its Relatives Molecular Biological and Industrial Workhorses," *Trends in Biotechnology*, vol 10, p 247-256 (1992)

28. Biochemistry of Peptide Antibiotics, (ed. Kleinkauf, v.D.) Walter de Gruyter, Berlin (1990).

29. *Bacillus subtilis* and Other Gram-Positive Bacteria, (ed. Sonenshein, A.L., Hoch, J.A., Losick, R.) American Society for Microbiology, Washington, D.C. (1993).

30. Microbiology, (2nd ed. Davis, B.D., Dulbecco, R., Eisen, H.N., Ginsberg, H.S., Wood, W.B., McCarty, M.) Harper and Row New York (1973).

31. Stragier, P., Losick, R., "Molecular Genetics of Sporulation in *Bacillus subtilis*," *Annual Review of Genetics*, vol. 30, p. 297-341 (1996).

32. Grossman, A.D., "Genetic Networks Controlling the Initiation of Sporulation and the Development of Genetic Competence in *Bacillus subtilis*," *Annual Review of Genetics*, vol. 29, p. 477-508 (1995).

33. Kunst, F., et. Al., "The Complete Genome Sequence of the Gram-positive Bacterium *Bacillus subtilis*," *Nature*, vol. 390, p. 249-256 (1997).

34. Biochemistry, 3rd edition, Stryer, Lubert, W. H. Freeman and Company (1975).

35. Blattner, F.R., Plinkett III, G., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B., Shao, Y., "The Complete Genome Sequence of *Escherichia coli* K-12," *Science*, vol.277, p.1453-1462 (1997).

36. Kunisawa, T., "Identification and Chromosomal Distribution of DNA Sequence Segments Conserved Since Divergence of *Escherichia coli* and *Bacillus subtilis*," *Journal of Molecular Evolution*, vol. 40, p. 585-593 (1995).

37. Terai, G., Takagi, T., Nakai, K., "Prediction of Co-regulated Genes in *Bacillus subtilis* on the Basis of Upsteam Elements Conserved Across Three Closely Related Species," *Genome Biology*, vol. 2, no. 11, p. 48.1-48.12 (2001).

38. Deckert, G., Warren, P.V., Gaasterland, T., Young, W.G., Lenox, A.L., Graham, D.E., Overbeek, R., Snead, M.A., Keller, M., Aujay, M., Huber, R., Feldman, R.A., Short, J.M., Olsen, G.J., Swanson, R.V., "The Complete Genome of the Hyperthermophilic Bacterium *Aquifex aeolicus*," *Nature*, vol. 392, p.353-358 (1998).

39. Nelson, K.E., Clayton, R.A., Gill, S.R., Gwinn, M.L., Dodson, R.J., Haft, D.H., Hickey, E.K., Peterson, J.D., Nelson, W.C., Ketchum, K.A., McDonald, L., Utterback, T.R., Malek, J.A., Linher, K.D., Garret, M.M., Stewart, A.M., Cotton, M.D., Pratt, M.S., Phillips, C.A., Richardson, D., Heidelberg, J., Sutton, G.G., Fleischmann, R.D., Eisen, J.A., White, O., Salzberg, S.L., Smith, H.O., Venter, J.C., Frazer, C.M., "Evidence for Lateral Gene Transfer between Archaea and Bacteria from Genome Sequence of *Thermotoga maritima*," *Nature*, vol. 399, p. 323-329 (1999).

40 Graham, D E., Kyrpides, N., Anderson, I J., Overbeek, R., Whitman, W B., "Genome of *Methanocaldococcus (Methanococcus) jannaschii*," *Methods in Enzymology*, vol. 330, p.40-123 (2001).

41 http://www.tigr.org

42 Bio-Automation, LCDR Mermade

43 Molecular Cloning: A Laboratory Manual, 2nd edition, (ed. Sambrook, J., Fritsch, E.F., Maniatis, T.) Cold Spring Harbor Press, p. 13.4-13.6 (1989)

44 Sanger, F., Nicklen, S., Coulson, A.R., "DNA Sequencing with Chain-Terminating Inhibitors, *Proceedings of the National Academy of Science*, vol.74, no.12, p 5463-5467 (1977).

45 Biochemistry, 4th edition, Stryler, L., W H. Freeman and Company, New York, NY, Chapter 4, p.75-76 (1995)

46 Watson, J D., Crick, F H.C., "Genetical Implications of the Structure of Deoxyribonucleic Acid," *Nature*, vol. 171, p. 964-967 (1953)

47 Watson, J D., Crick, F H.C., "Molecular Structure of Nucleic Acids," *Nature*, vol.171, p.737-738 (1953)

48 Meldrum, D R., "Sequencing Genomes and Beyond," *Science*, vol. 292, p. 515-517 (2001)

49 Dictionary of Microbiology and Molecular Biology, (2nd ed. Singleton, P., Sainsbury, D.) John Wiley and Sons, New York, 1995

50 Perkin-Elmer, ABI Prism DNA Sequencing Chemistry Guide, Version A, May 1995

51 Landegren, U., Kaiser, R., Caskey, C T., Hood, L., "DNA Diagnostics-Molecular Techniques and Automation," *Science*, vol. 242, p. 229-237 (1988)

52 Kambara, H., Takahashi, S., "Multiple-Sheathflow Capillary Array DNA Analyzer," *Nature*, vol. 361, p. 565-566 (1993)

53 http://www.Perkin-elmer.com/about our products/ABI Prism 3700 DNA Analyzer

54 http://www.mdyn.com/products/MegaBACE.

55 Amersham Pharmacia Biotech, Life Science News 2, p. 4-7 (1999).

56 http://www.genome.washington.edu/UWGC/analysistools/phred.htm.

57 http://www.genome.washington.edu/UWGC/analysistools/phrap.htm.

58 http://www.genome.washington.edu/UWGC/analysistools/crossmatch.htm.

59 http://www.genome.washington.edu/UWGC/analysistools/consed.htm.

60 Gordon, D., Abajian, C., Green, P., "Consed: A Graphical Tool for Sequence Finishing," *Genome Research*, vol.8, p.195-202 (1998).

61 Altschul, S. F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., "Basic Local Alignment Search Tool," *Journal of Molecular Biology*, vol.215, p. 403-410 (1990).

62 Zhang, J., Madden, T.L., "PowerBLAST: A New Network BLAST Application for Interactive or Automated Sequence Analysis and Annotation," *Genome Research*, vol.7, p.649-656 (1997).

63 http://www.ncbi.nlm.nih.gov.

64 Burge, C., Karlin, S., "Prediction of Complete Gene Structures in Human Genomic DNA," *Journal of Molecular Biology*, vol.268, p.78-94 (1997).

65 http://repeatmasker.genome.washington.edu/cgi-bin/RM2=req.pl

66 http://www.genome.ou.edu/pub/programs/exgap_src.tar

67 http://www.genome.ou.edu/informatics.html.

68 http://www.cgr.ki.se/cgr/groups/sonnhammer/Dotter.html

69 Sonnhammer, E.L.L., Durbin, R., "A Dot-Matrix Program with Dynamic Threshold Control Suited for Genomic DNA and Sequence Analysis," *Gene*, vol.167, GC1-10 (1995).

70 Fichant, G.A., Burks, C., "Identifying Potential tRNA Genes in Genomic DNA Sequences," *Journal of Molecular Biology*, vol. 220, p.659-671 (1991).

71 http://www.sanger.ac.uk/software/artemis/v3/manual

72 Rutherford, K.M., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M-A., Barnell, B., "Artemis: Sequence, Visualisation, and Annotation," *Bioinformatics*, in press.

73 Bouffard, G., Ostell, J., Rudd, K.E., "GeneScape: a Relational Database of *Escherichia coli* Genomic Map for MacIntosh Computers," *CABIOS* vol. 8, p. 563-567 (1992).

74 Rudd, K.E., "Alignment of *Escherichia coli* DNA Sequences to a Revised,

Integrated Genomic Restriction Map," p. 2.3-2.8, (edited by J. Miller), A Short Course in Bacterial Genetics: A Laboratory Manual, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York (1992).

75. Rudd, K.E., Bouffard, G., Miller, W., "Computer Analysis of *Escherichia coli* Restriction Maps," p. 1-38, (edited by K.E. Davies & S.M. Tilghmann), Genome Analysis, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York (1992).

76. Bachmann, B. J., "Linkage Map of *Escherichia coli*, K12," *Microbiological Review*, vol. 54, p. 130-197 (1990).

77. Riley, M.E., "Functions of the Gene Products of *Escherichia coli*," *Microbiological Reviews*, vol.57, no.4, p.862-952 (1993).

78. Kanehisa, M., Susumu, G., "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Research*, vol.28, no.1, p.27-30 (2000).

79. http://www.genome.ad.jp/kegg/kegg2.html

80. http://pedant.mips.biochem.mpg.de/cgi-bin

81. http://www.pasteur.fr/Bio/SubtiList.html

82. http://*bacillus*.genome.ad.jp

83. http://www.jamstec.go.jp/jamstec-e/bio_exbase.html

84. http://www.cbs.dtu.dk/services/TMHMM

85. Stenesh, J., Yang, C., "Characterization and Stability of Ribosomes from Mesophilic and Thermophilic Bacteria," *Journal of Bacteriology*, vol.93, No.3, p.930-936 (1967)

86. Kutchma, A.J., Roberts, M.A., Crawford, K., Crawford, D.L., "Small-Scale Isolation of Genomic DNA from *Streptomyces* Mycelia or Spores," *Biotechniques*, vol. 24, p.452-457 (1998).

87. http://www.gentra.com.

88. Bio-Rad, Complete Prokaryotic and Yeast Transformation: The E. coli Pulser Electroporator, Bulletin 1734.

89. Beckman, Biomek 2000 Laboratory Automation Workstation, User's Manual Beckman Instructions 015-609324-A (1994).

90. Genomic Solution, PBA Flexys Colony and Plaque Picker Product Manual.

91. GeneMachines, HIGRO User's Manual, Version: HGA.01F.11201999.

92. http://www.genome.ou.edu/protocols.

93. E-mail from Dr. Jonathon Coren.

94. http://www.Perkin-elmer.com/about our products/ABI Prism 377 DNA Sequencer.

95. http://www.millipore.com/multiscreen

96. Saiki, R.K., Scharf, S., Faloona, F., Mullis, K.B., Horn, G.T., Erlich, H.A., Arnheim, N.A., "Enzymatic Amplification of B-Globin Genomic Sequences and Restriction Site Analysis for Diagnosis of Sickle Cell Anemia," *Science*, vol 230, p.1350-1354 (1985).

97. http://www.genome.ou.edu/bstearo.html

98. Frazer, C.M., Eisen, J., Fleischmann, R.D., Ketchum, K.A., Petersen, S., "Comparative Genomics and Understanding of Microbial Biology," *Genomics*, vol. 6, no. 5, p.505-512 (2000).

99. http://pedant.gsf.de/cgi-bin/wwwfly.pl?Set=Bacillus_stearothermophilus_10&Page=index.

100. Lander, E.S., Waterman, M.S., "Genomic mapping by fingerprinting random clones: a mathematical analysis," *Genomics* 2 (3): 231-239 (1988).

101. Stenesh, J., Roe, B.A., "DNA Polymerase from Mesophilic and Thermophilic Bacteria,"*Biochimica Et Biophysica ACTA*, vol. 272, p. 167-178 (1972).

102. Honjo, M., Akaoka, A., Nakayama, A., Furutani, Y., "Secretion of Human Growth Hormone in *Bacillus subtilis* Using Prepropeptide Coding Region of *Bacillus amyloliquefaciens* Neutral Protease Gene," *Journal of Biotechnology*, vol. 4, p. 63-71 (1986).

103. Takami, H., Takaki, Y., Nakasone, K., Sakiyama, T., Maeno, G., Sasaki, R., Hirama, C., Fuji, F., Masui, N., "Genetic Analysis of the Chromosome of Alkaliphilic *Bacillus halodurans* C-125," *Extremophiles*, vol.3, p.227-233 (1999).

104. Okstad, O., Hegna, I., Lindback, T., Rishovd, A.L., Kolsto, A.B., "Genome Organization is not Conserved between *Bacillus cereus* and *Bacillus subtilis*, *Microbiology*, vol.145, p.621-631 (1999).

105. The Physiology and Biochemistry of Prokaryotes, 2nd edition, David White, Oxford University Press, Oxford (2000).

106. de Vrij, W., van den Burg, B., Konings, W.N., "Spectral and potentiometric analysis of Cytochromes from *Bacillus subtilis*," *European Journal of Biochemistry*,

vol. 166, p. 589-595 (1987).

107. Rowland, B., Hill, K., Miller, P., Driscoll, J., Taber, H., "Structural Organization of a *Bacillus subtilis* Operon Encoding Menaquinone Biosynthetic Enzymes," *Gene*, vol. 167, p. 105-109 (1995).

108. Nakano, M.M., Dailly, Y.P., Zuber, P., Clark, D.P., "Characterization of Anaerobic Fermentative Growth of *Bacillus subtilis*: Identification of Fermentation End Products and Genes Required for Growth," *Journal of Bacteriology*, vol. 179, no. 21, p. 6749-6755 (1997).

109. Raux, E., Lanois, A., Warren, M.J., Rambach, A., Thermes, C., "Cobalamin (vitamin B12) biosynthesis: identification and characterization of a *Bacillus megaterium* cobI operon," *Biochemistry Journal*, vol. 335, p. 159-166 (1998).

110. Miczak, A., "Porphyrin and Corrinoid Mutants of *Bacillus subtilis*," *Journal of Bacteriology*, vol. 131, no. 2, p. 379-381 (1977).

111. Bacher, A., Eberhardt, S., Fischer, M., Kis, K., Richter, G., "Biosynthesis of Vitamin B2 (Riboflavin)," *Annual Reviews in Nutrition*, vol. 20, p. 153-167 (2000).

112. Bacterial Metabolism, 2nd edition, Gerhard Gottschalk, Springer-Verlag, New York (1986).

113. Liang, S., and Wever, A., "Systemic Underutilization of Glutamine in Thermophile Proteins," The Joint Cold Spring Harbor Laboratory/Wellcome Trust Conference on Genome Informatics, September 4-8, 2002.

114. Williams, J.A., Henry, S.S., Bernards, A., Field, J., Sehgal, A., "A Circadian Output in *Drosophila* Mediated by Neurofibromatosis-1 and Ras/MAPK," *Science*, vol. 293, no. 5538, pg. 2251-2256 (2001).

115. Luijten, M., Wang, Y., Smith, B.T., Westerveld, A., Smink, L.J., Dunham, I., Roe, B.A., Hulsebos, T.J., "Mechanism of Spreading of the Highly Related Neurofibromatosis type 1 (NF1) pseudogenes on Chromsomes 2, 14, and 22," *European Journal of Human Genetics*, vol. 18, no. 3, pg. 209-214 (2000).

116. Dunham, I., Shimizu, N., Roe, B.A., Chissoe, S., "The DNA Sequence of Human Chromosome 22," *Nature*, vol. 402, p. 489-495 (1999).

117. Deprez, R.H.L., Riegman, P.H.J., Groen,, N.A., Warringa,, U.L., van Biezen, N.A., Molijn, A.C., Bootsma, D., de Jong, P.J., Menon,, A.G.,, Kley, N.A., Seizinger, B.R.,, Zwarthoff, E.C., "Cloning and Characterization of MN1, A Gene from Chromosome 22q11, Which is Disrupted by a Balanced Translocation in a Meningioma," *Oncogene*, vol. 10, p. 1521-1528 (1995).

Chapter VI: Appendix

Appendix 1-The Human Genome Project

During my research, in addition to sequencing the working draft of the *Bacillus stearothermophilus* genome, I sequenced a 41 Kb cosmid (cosi) on Human Chromosome 22 and a three cosmid set, cos50b11, cos58b12, and cos76A4,(~90Kb) in a Meningioma Deletion Region of Human Chromosome 22 as part of our laboratory's participation in the Human Genome Project. The Human Genome Project began at The Atomic Energy Commission and the Energy Research and Development Administration both of which are now presently known as the Department of Energy. The DOE originally set out to study the genetic and health effects of radiation and chemical by-products of energy production but later realized that the best way to learn about these effects was to study DNA directly. The Human Genome Project resulted as an expansion that included the NIH, which formed the National Human Genome Research Institute (NHGRI)

Analysis of the sequence revealed a gene for neurofibromatosis 1. In humans, neurofibromatosis 1 *(NF1)*, a tumor suppressor, is a frequent dominantly inheredited disorder occurring once in every 3,000-3,500 births with a gene product, neurofibromin, a protein of 2,818 amino acids. This protein inactivates the *Ras* oncogene through hydrolysis of guanosine triphosphate (GTP). Mutations in the *nf1* gene causing neurofibromatosis type 1 are phenotypically characterized variably to include the appearance of peripheral nerve sheath tumors called neurofibromas, a short statute, Lisch nodules of the iris, and a learning deficit. Tissues derived from the embryonic neural crest result in impaired intercellular communication within peripheral nerves (114).

Although neurofibromatosis 1 maps to chromosome band 17q11.2, *NF1*-related sequences (pseudogenes) are present on a number of chromosomes including 2, 12, 14, 15, 18, 21, and 22. Researchers have hypothesized that during evolution, these pseudogenes arose initially by duplication and transposition of the middle part of the functional *NF1* gene to chromosome 2, and subsequently by the same process arose

copies on chromosome 14 and 2 (115).

A. Analysis of the Cosi Sequencing Project

The results of the completed cosmid sequencing projects were included in the published results from Human Chromosome 22 (116). The cosmid, cosi, in the NF1 Region of Human Chromosome 22 was sequenced using a shotgun strategy in pUC18. Here, eighteen 96 well plates totally 1728 subclones were isolated and the sequence data was collected from forward and reverse universal primer reactions. Additional primer walking reactions using custom primers on subclones were performed to close gaps using 2 ul custom primer, 2 ul Big Dye, 1 ul 2 percent DMSO, and 2 ul DNA. Cosi had a total insert size of 41,958 bp with a G+C content of 41.63 percent. The average length of sequencing reads was 373 bp. A total of 1,456 gel reads was assembled by phred/phrap.

The Genscan output file predicted peptide sequences of 3 genes in cosi, the coding sequence of each gene and the start position of the first methionine. The BLAST for cosi revealed a homosapien gene similar to neurofibromin from 11,841 bp to 37,948 bp, an NF1 gene product, type 1 neurofibromatosis protein from 18,088 bp to 27,692 bp, a human inducible nitric oxide synthase mRNA at 34,592 bp; and a NEK2 gene, protein kinase, NIMA-like gene at 39,813 bp.

REPEATMASKER predicted that 33.84 percent of the 41,958 bp cosi region contained known repeat elements and had an overall G+C content of 42.47 percent. ALUs and L1 repeat elements were the main repeat elements present in the amount of 9.19 percent and 13.22 percent respectively.

Table 45-Repeatmasker for cosi

```
file name: cosi.fasta.screen.contig
sequences:          1
total length:    41958 bp
GC level:        42.47 %
bases masked     14199 bp ( 33.84 %)
**********************************************
                 number of      length   percentage
                 elements*    occupied  of sequence
----------------------------------------------------
SINEs:                19         4067 bp     9.69 %
      ALUs            17         3857 bp     9.19 %
      MIRs             2          210 bp     0.50 %

LINEs:                10         5872 bp    13.99 %
      LINE1            8         5548 bp    13.22 %
      LINE2            2          324 bp     0.77 %

LTR elements:         3          927 bp     2.21 %
      MaLRs           0            0 bp     0.00 %
      Retrov.         2          473 bp     1.13 %
      MER4_group      1          454 bp     1.08 %

DNA elements:         7         2413 bp     5.75 %
      MER1_type       3          380 bp     0.91 %
      MER2_type       3         1894 bp     4.51 %
      Mariners        0            0 bp     0.00 %

Unclassified:         0            0 bp     0.00 %

Total interspersed repeats:    13279 bp    31.65 %


Small RNA:            0            0 bp     0.00 %

Satellites:           1          336 bp     0.80 %
Simple repeats:       3          192 bp     0.46 %
Low complexity:       5          392 bp     0.93 %
```

B. Analysis of mn1_90kb Sequencing Project

Meningioma Deletion Regions are chromosomal segments that are lacking in the genomes. Present in primary tumors of the arachnoidal layer of the meninges, these meningioma tumors occur as benign and slow-growing tumors and represent between 13 and 19 percent of all primary intracranial and 25 percent of all intraspinal tumors. Both cytogenetic and molecular genetic studies have shown that complete or partial loss of Chromosome 22 is a very frequent event in the development of meningiomas. Specific chromosomal translocations are involved in the activation of proto-oncogenes and activation of tumor suppressor genes thus suggesting that tumor suppressor gene or genes located on the q-arm of Chromosome 22 play a crucial role in their pathogenesis (117).

Three cosmids, cos50b11, cos58b12, and cos76A4, (~90 Kb) in a Meningioma Deletion Region of Human Chromosome 22 were sequenced using a shotgun strategy in pUC18. Dr. Judy Crabtree began this project. To the data, I added an additional (33) 96 well plates totally 3,168 subclones to be completed. This sequence data from the forward and reverse primer reactions on these templates was done in collaboration with Dr. Qun Ren. As is typical of human genomic DNA this region contained numerous repeat sequences, which accounted for 33.58 percent of the total size of 90,223 bp. The average length of sequencing reads was 421 bp and the average G+C content was 51.22 percent. A total of 6,282 gel reads was assembled by phred/phrap. The only exon identified was a homosapien MN1 gene.

REPEATMASKER predicted that 33.58 percent of the 90,223 bp mn1_90kb region contained known repeat elements and has an overall G+C content of 51.22 percent. ALU repeat elements are the main repeat elements present.

Table 46-Repeatmasker for mn1_90kb

file name: mn1_90kb.fasta.screen.co
sequences: 1
total length: 90223 bp
GC level: 51.22 %
bases masked 30296 bp (33.58 %)

	number of elements*	length occupied	percentage of sequence
SINEs:	99	18874 bp	20.92 %
ALUs	41	11563 bp	12.82 %
MIRs	58	7311 bp	8.10 %
LINEs:	25	5002 bp	5.54 %
LINE1	4	1765 bp	1.96 %
LINE2	21	3237 bp	3.59 %
LTR elements:	6	2081 bp	2.31 %
MaLRs	4	1671 bp	1.85 %
Retrov.	2	410 bp	0.45 %
MER4_group	0	0 bp	0.00 %
DNA elements:	12	1844 bp	2.04 %
MER1_type	10	1644 bp	1.82 %
MER2_type	0	0 bp	0.00 %
Mariners	1	87 bp	0.10 %
Unclassified:	0	0 bp	0.00 %
Total interspersed repeats:		27801 bp	30.81 %
Small RNA:	1	44 bp	0.05 %
Satellites:	0	0 bp	0.00 %
Simple repeats:	29	1178 bp	1.31 %
Low complexity:	15	1292 bp	1.43 %

Appendix 2-Functional Classification of *Bacillus subtilis* Protein Coding Genes

		Bsu	%
1	Cellular envelope and cellular processes	867	21%
	1.1-Cell wall	93	
	1.2-Transport/binding proteins and lipoproteins	381	
	1.3-Sensors	38	
	1.4-Membrane bioenergetics	78	
	1.5-Mobility and chemotaxis	55	
	1.6-Protein secretion	18	
	1.7-Cell division	21	
	1.8-Sporulation	139	
	1.9-Germination	23	
	1.10-Transformation/competence	21	
2	Intermediary metabolism	742	18%
	2.1-Metabolism of carbohydrates and related molecules	261	
	2.1.1-Specific pathways	214	
	2.1.2-Main glycolytic pathways	28	
	2.1.3-TCA cycle	19	
	2.2-Metabolism of amino acids and related molecules	205	
	2.3-Metabolism of nucleotides and nucleic acids	83	
	2.4-Metabolism of lipids	77	
	2.5-Metabolism of coenzymes and prosthetic groups	99	
	2.6-Metabolism of phosphate	9	
	2.7-Metabolism of sulfur	8	
3	Information pathways	**482**	12%
	3.1-DNA replication	22	
	3.2-DNA restriction/modification & repair	39	
	3.3-DNA recombination	17	
	3.4-DNA packaging and segregation	10	
	3.5-RNA synthesis	244	
	3.5.1-Initiation	19	
	3.5.2-Regulation	213	
	3.5.3-Elongation	8	
	3.5.4-Termination	4	
	3.6-RNA Modification	19	
	3.7-Protein synthesis	96	

	3.7.1-Ribosomal proteins	56	
	3.7.2-Aminoacyl-tRNA synthetases	25	
	3.7.3-Initiation	6	
	3.7.4-Elongation	6	
	3.7.5-Termination	3	
	3.8-Protein modification	27	
	3.9-Protein folding	8	
		822	
4	Other functions	289	7%
	4.1-Adaptation to atypical conditions	72	
	4.2-Detoxification	68	
	4.3-Antibiotic production	30	
	4.4-Phage-related functions	83	
	4.5-Transposon and IS	10	
	4.6-Miscellaneous	26	
5	Similar to unknown proteins	667	16%
	5.1-From B. subtilis	177	
	5.2-From other organisms	490	
6	No similarity	1053	26%

Appendix 3-Functional Assignment from *Bacillus stearothermophilus* Working Draft Sequence

Categories	# of ORFs	%
Small Molecules	**836**	**30.61**
Degradation	**180**	**6.6**
Carbon Compounds	97	3.6
Amino Acids and Amines	32	1.2
Fatty Acids	45	1.6
Phosphorus Compounds	6	0.2
Energy Metabolism	**206**	**7.52**
Glycolysis	27	0.99
Pyruvate Dehydrogenase	22	0.81
TCA Cycle	12	0.44
Methane Metabolism	3	0.11
Oxidative Pentose Phosphate Pathway	2	0.07
Aerobic-Respiration	10	0.37
Anaerobic-Respiration	39	1.43
Electron Transport	31	1.14
Fermentation	37	1.35
ATP-Proton Motive Force Interconv	22	0.81
Central Intermediary Metabolism	**49**	**1.8**
General	32	1.17
Gluconeogenesis	1	0.04
Sugar-nucleotide Biosynthesis	6	0.22
Amino Sugars	4	0.15
Sulfur Metabolism	6	0.22
Amino Acid Biosynthesis	**160**	**5.86**
Glutamate Family	22	0.81
Aspartate Family	34	1.24
Serine Family	15	0.55
Cysteine Metabolism	1	0.04
Valine, Leucine&Isoleucine Metabolism	7	0.26
Lysine Metabolism	4	0.15
Arginine & Proline Metabolism	10	0.37
Aromatic Amino Acid Family	40	1.17
Urea Cycle&Metabolism-Amino Groups	2	0.07
Cyanoamino Acid Metabolism	1	0.04
D-Alanine	1	0.04
Histidine	19	0.7
Pyruvate Family	1	0.04
Branched-chain Family	11	0.4
Polyamine Biosynthesis	1	0.04
Purines,Pyrimidines,Nucleosides&Nucleotides	**77**	**2.81**
Purine Ribonucleotide Biosynthesis	37	1.35
Pyrimidine Ribonucleotide Biosynthe	14	0.51
2'Deoxyribonucleotide Metabolism	8	0.29
Salvage of Nucleosides & Nucleotides	13	0.48
Miscellaneous	5	0.18

Purines,Pyrimidines,Nucleosides&Nucleotides	**77**	**2.81**
Purine Ribonucleotide Biosynthesis	37	1.35
Pyrimidine Ribonucleotide Biosynthe	14	0.51
2'Deoxyribonucleotide Metabolism	8	0.29
Salvage of Nucleosides & Nucleotides	13	0.48
Miscellaneous	5	0.18
Biosynthesis of Cofactors, Prosthetic Groups and Carriers	**67**	**2.47**
Biotin	3	0.11
Folic Acid	4	0.15
Lipoate	1	0.04
Molybdopterin	6	0.22
Pantothenate	3	0.11
Pyridoxine	1	0.04
Pyridine nucleotide	3	0.11
Thiamine	10	0.37
Riboflavine	4	0.15
Thioredoxin, Glutaredoxin&Glutathione	1	0.04
Menaquinone & Ubiquinone	14	0.51
Heme & Porphyrin	14	0.51
Enterochelin	2	0.07
Cobalamine	1	0.04
Fatty Acid Biosynthesis	**27**	**0.99**
Broad Regulatory Functions	**62**	**2.27**
Macromolecules	**331**	**12.12**
Synthesis & Modification	**207**	**7.58**
Ribosome Maturation & Modification	30	1.09
Glycerolipid Metabolism	4	0.15
Inositol Phosphate Metabolism	2	0.07
Aminoacyl tRNA Synthesis&Modification	37	1.35
Aminosugars	12	0.44
Biosynthesis&Degradation of Glucoprotein	1	0.04
DNA Replication,Modification&Recombination	77	2.82
Protein Translation&Modification	10	0.37
RNA Synthesis,Transcription&Modification	28	1.03
Polysaccharides	4	0.15
Phospholipids	2	0.07
Degradation of Macromolecules	**47**	**1.73**
RNA&DNA	22	0.81
Proteins & Peptides	22	0.81
Polysaccharides	3	0.11
Cell Envelope	**77**	**2.82**

Surface Polysaccharides&Lipopsaccharides	13	0.48
Surface Structures	30	1.1
Murein Sacculus & Peptidoglycan	34	1.24
Processes	**371**	**13.58**
Transport Binding Proteins	**287**	10.51
Amino Acids & Amines	54	1.98
Cations	21	0.77
Anions	11	0.4
Carbohydrates,Organic Alcohols&Acids	94	3.44
Nucleosides,Purines&Pyrimidines	1	0.04
Other	106	3.88
Chaperones	**5**	0.18
Cell Division	**22**	0.81
Chemotaxis and Mobility	**4**	0.15
Protein and Peptide Secretion	**40**	1.46
Osmotic Adaptation	**9**	0.33
Detoxification	**4**	0.15
Other	**13**	**0.48**
Phage-related Functions & Prophages	1	0.04
Drug/Analog Sensitivity	9	0.33
Radiation Sensitivity	1	0.04
Adaptation & Atypical Conditions	2	0.07
Metabolism of Cofactors and Vitamins	**11**	**0.41**
Riboflavin Metabolism	1	0.04
Pantotenate&CoA Biosynthesis	4	0.15
Folate Biosynthesis	2	0.07
One Carbon Pool By Folate	1	0.04
Porphyrin&Chlorophyll Metabolism	3	0.11
Metabolism of Other Substances	**12**	**0.45**
Flavonoids, Stilbene&Lignin Biosynthesis	10	0.37
Alkaloid Biosynthesis I	1	0.04
Alkaloid Biosynthesis II	1	0.04
Unknown	**1157**	**42.37**
Total	**2731**	**100.02**