The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis* 

F. Kunst1, N. Ogawara2, I. Moszer3, A. M. Albertini4, G. Aloni5, V. Azevedo6, M. G. Bertero7,8, P. Bessières9, A. Bolotin10, S. Borchert11, R. Borris12, L. Boursier13, A. Brans14, M. Braun15, S. C. Brigell16, S. Bron17, S. Brouillet18, C. V. Bruschi19, B. Caldwell20, V. Capuano21, N. M. Carter22, S.-K. Choi23, J.-J. Codani24, I. F. Conerton25, N. J. Cummings26, R. A. Daniel27, F. Denizot19, K. M. Devine28, A. Düsterhöft29, S. D. Ehrlich30, P. T. Emmerson31, K. D. Entian32, J. Errington33, C. Fabret34, E. Ferrari35, D. Foulger36, C. Fritz37, M. Fujita38, Y. Fujita39, S. Fuma40, A. Galizi41, N. Galleron42, S.-Y. Ghim43, P. Glaser44, A. Goffeau45, E. J. Golightly46, G. Grandi47, G. Guiseppi19, B. J. Guy48, K. Haga49, J. Haich8, C. R. Harwood50, A. Hénaut51, H. Hilbert52, S. Holssen53, S. Hosono54, M.-F. Hulo55, M. Itaya56, J. Jones57, B. Joris58, D. Karamata59, Y. Kasahara60, A. Kawahara61, K. Kobayashi62, P. Koetter63, G. Konings64, S. Krogh65, M. Kumano66, K. Kurita67, A. Lapidus68, S. Lardinois69, J. Lauber70, V. Lazarevic71, S.-M. Lee72, A. Levine73, H. Liu74, S. Masuda75, C. Maéul76, N. Medina77, C. R. Harwood78, H. Hilbert79, S. Holsappel80, S. Hosono81, M.-F. Hulo82, M. Itaya83, J. Jones84, B. Joris85, D. Karamata86, Y. Kasahara87, A. Kawahara88, K. Kobayashi89, P. Koetter90, G. Konings91, S. Krogh92, M. Kumano93, K. Kurita94, A. Lapidus95, S. Lardinois96, J. Lauber97, V. Lazarevic98, S.-M. Lee99, A. Levine100, H. Liu101, S. Masuda102, C. Maéul103, N. Medina104, R. P. Mellado105, M. Mizuno106, D. Moestl107, S. Nakai108, M. Noback109, D. O'Reilly110, K. Ogawa111, A. Ogigawa112, B. Oudega113, S.-H. Park114, V. Parro115, T. M. Pohl116, D. Portetelle117, S. Porwollik118, A. M. Prescott119, E. Presecan120, P. Pujic121, B. Purnelle122, G. Rapoport123, M. Rey124, S. Reynolds125, M. Rieger126, C. Rivolta127, E. Rocha128, B. Roche129, M. Rose130, Y. Sadaie131, T. Sato132, E. Scanlan133, S. Schleich134, R. Schroeter135, F. Scocco136, J. Sekiguchi137, A. Sekowska138, S. J. Seror139, P. Serro140, B.-S. Shin141, B. Soldo142, A. Sorokin143, E. Taconni144, T. Takagi145, K. Takahashi146, K. Takekura147, M. Takeuchi148, A. Tamakoshi149, T. Tanaka150, P. Terpsstra151, A. Tognoni152, V. Tosato153, S. Uchiyama154, M. Vaknin155, A. Vassella156, A. Viari157, R. Wambutt158, E. Wedler159, H. Werder160, W. Weis161, Y. Yamamoto162, K. Yamane163, K. Yasumoto164, K. Yata165, S. Yoshida166, H.-F. Yoshikawa167, E. Zumstein168, H. Yoshikawa169, A. Danchin170, A. Danchin171, A. Danchin172, A. Danchin173, A. Danchin174, A. Danchin175, A. Danchin176, A. Danchin177, A. Danchin178, A. Danchin179.

---

*Bacillus subtilis* is the best-characterized member of the Gram-positive bacteria. Its genome of 4,214,810 base pairs comprises 4,100 protein-coding genes. Of these protein-coding genes, 53% are represented once, while a quarter of the genome corresponds to several gene families that have been greatly expanded by gene duplication, the largest family containing 77 putative ATP-binding transport proteins. In addition, a large proportion of the genetic capacity is devoted to the utilization of a variety of carbon sources, including many plant-derived molecules. The identification of five signal peptidase genes, as well as several genes for components of the secretion apparatus, is important given the capacity of *Bacillus* strains to secrete large amounts of industrially important enzymes. Many of the genes are involved in the synthesis of secondary metabolites, including antibiotics, that are more typically associated with *Streptomyces* species. The genome contains at least ten prophages or remnants of prophages, indicating that bacteriophage infection has played an important evolutionary role in horizontal gene transfer, in particular in the propagation of bacterial pathogenesis.
Techniques for large-scale DNA sequencing have brought about a revolution in our perception of genomes. Together with our understanding of intermediary metabolism, it is now realistic to envisage a time when it should be possible to provide an extensive chemical definition of many living organisms. During the past couple of years, the genome sequences of *Haemophilus influenzae*, *Mycoplasma genitalium*, *Synecocystis PCC6803*, *Methanococcus jannaschii*, *M. pneumoniae*, *Escherichia coli*, *Helicobacter pylori*, *Archaeoglobus fulgidus* and the yeast *Saccharomyces cerevisiae* have been published in their entirety\(^1\)-\(^8\), and at least 40 prokaryotic genomes are currently being sequenced. Regularly updated lists of genome sequencing projects are available at http://www.mcs.anl.gov/home/gaasterl/genomes.html (Argonne National Laboratory, Illinois, USA) and http://www.tigr.org (TIGR, Rockville, Maryland, USA).

The list of sequenced microorganisms does not currently include a paradigm for Gram-positive bacteria, which are known to be important for the environment, medicine and industry. *Bacillus subtilis* has been chosen to fill this gap\(^9\),\(^10\) as its biochemistry, physiology and genetics have been studied intensely for more than 40 years. *B. subtilis* is an aerobic, endospore-forming, rod-shaped bacterium commonly found in soil, water sources and in association with plants. *B. subtilis* and its close relatives are an important source of industrial enzymes (such as amylases and proteases), and much of the commercial interest in these bacteria arises from their capacity to secrete these enzymes at gram per litre concentrations. It has therefore been used for the study of protein secretion and for development as a host for the production of heterologous proteins\(^1\). *B. subtilis* (*natto*) is also used in the production of Natto, a traditional Japanese dish of fermented soya beans.

Under conditions of nutritional starvation, *B. subtilis* stops growing and initiates responses to restore growth by increasing metabolic diversity. These responses include the induction of motility and chemotaxis, and the production of macromolecular hydrolases (proteases and carbohydrate) and antibiotics. If these responses fail to re-establish growth, the cells are induced to form chemically, irradiation- and desiccation-resistant endospores. Sporulation involves a perturbation of the normal cell cycle and the differentiation of a binucleate cell into two cell types. The division of the cell into a smaller forespore and a larger mother cell, each with an entire copy of the chromosome, is the first morphological indication of sporulation. The former is engulled by the latter and differential expression of their respective genomes, coupled to a complex network of interconnected regulatory patterns and developmental checkpoints, culminates in the programmed death and lysis of the mother cell and release of the mature spore\(^2\). In an alternative developmental process, *B. subtilis* is also able to differentiate into a physiological state, the competent state, that allows it to undergo genetic transformation\(^13\).

### General features of the DNA sequence

**Analysis at the replicon level.** The *B. subtilis* chromosome has 4,214,810 base pairs (bp), with the origin of replication coinciding with the base numbering start point\(^14\), and the terminus at about 2,017 kilobases (kb)\(^15\). The average G + C ratio is 43.5%, but it varies considerably throughout the chromosome. This average is also different if one considers the nucleotide content of coding sequences, for which G and A (24% and 30%) are relatively more abundant than their counterparts C and T (20% and 26%). A significant inversion of the relative G – C/G + C ratio is visible at the origin of replication, indicating asymmetry of the nucleotide composition between the replication leading strand and the lagging strand\(^16\). Several A + T-rich islands are likely to reveal the signature of bacteriophage lysogens or other inserted elements (Fig. 1, see below).

We have analysed the abundance of oligonucleotides (‘words’) in the genome in various ways: absolute number of words in the genomic text, or comparison with the expected count derived from several models of the chromosome (for example, Markov models, or simulated sequences in which previously known features of the genome were conserved\(^17\)). Comparing the experimental data with various models allowed us to define under- and overrepresentation of words in the experimental data set by reference to the model chosen. In general, the dinucleotide bias follows closely what has been described for other prokaryotes\(^18\),\(^19\), in that the dinucleotides most overrepresented are AA, TT and GC, whereas those less represented are TA, AC and GT. Plots of the frequencies of AG, GA, CT and TC in sliding windows along the chromosome show dramatic decreases or increases around the origin and terminus of replication (data not shown). Trinucleotide frequency, directly related to the coding frame, will be discussed below. The distribution of words of four, five and six nucleotides shows significant correlations between the usage of some words and replication (several such oligonucleotides are very significantly overrepresented in one of the strands and underrepresented in the other one).

Setting a statistical cut-off for the significance of duplications at 10\(^{-3}\), we expected duplication by chance of words longer than 24 nucleotides to be rare\(^20\). In fact, the genome of *B. subtilis* contains a plethora of such duplications, some of them appearing more than

---

**Figure 1** Distribution of A + T-rich islands along the chromosome of *B. subtilis*, in sliding windows of 10,000 nucleotides, with a step of 5,000 nucleotides. Location of genes from class 3 according to codon usage analysis (see Fig. 4) is indicated by dots at the bottom of the graph. Known prophages (PBSX, SP\(\beta\) and skin) are indicated by their names, and prophage-like elements are numbered from 1 to 7.
twice. Among the duplications, we identified, as expected, the ribosomal RNA genes and their flanking regions, but also regions known to correspond to genes comprising long sequence repeats (such as pks and srf). We also found several regions that were not expected: a 182-bp repetition within the yyyL and yyyO genes; a 410-bp repetition between the yyyK and yyyL genes; an internal duplication of 174 bp inside yycf; and significant duplications in the regions involved in the transcriptional control of several genes (such as 118 bp repeated three times between yxyB and yxyC). Finally, we found several repetitions at the borders of regions that might be involved in bacteriophage integration.

The most prominent duplication was a 190-bp element that was repeated 10 times in the chromosome. Multiple alignment of the ten repeats showed that they could be classified into two subfamilies with six and three copies each, plus a copy of what appears to be a chimera. Similar sequences have also been described in the closely related species *Bacillus licheniformis*21,22. A striking feature of these repeats is that they are only found in half of the chromosome, at either side of the origin of replication, with five repeats on each side. Furthermore, with the exception of the most distal repeat at position 737,062, they lie in the same orientation with respect to the movement of the replication fork (Figs 2 and 3). Putative secondary structures conserved by compensatory mutations, as well as an insert in three of the copies, suggest that this element could indicate a structural RNA molecule.

**Analysis at the transcription and translation level.** Over 4,000 putative protein-coding sequences (CDSs) have been identified, with an average size of 890 bp, covering 87% of the genome sequence (Fig. 2). We found that 78% of the genes started with ATG, 13% with TTG and 9% with GTG, which compares with 85%, 3% and 14%, respectively, in *E. coli*.15 Fifteen genes (eight in the predicted CDSs in bacteriophage SPB) exhibiting unusual start codons (namely ATT and CTG) were also identified through their similarities to known genes in other organisms or because they had a good GeneMark prediction (see Methods). This has not yet been substantiated experimentally. However, in the case of the gene coding for translation initiation factor 3, the similarity with its *E. coli* counterpart strongly suggests that the initiation codon is ATT, as is the case in *E. coli*.

We have not annotated CDSs that largely or entirely overlap existing genes, although such genes (for example, conS inside srfA) certainly exist. It is also likely that some of the short CDSs present in the *B. subtilis* genome have been overlooked. For these reasons and possible sequencing errors, the estimated number of *B. subtilis* CDSs will fluctuate around the present figure of 4,100.

In several cases, in-frame termination codons or frameshifts were confirmed to be present on the chromosome (for example, an internal termination codon in yfjF, or the known programmed transcriptional frameshift in prfB), indicating that the genes are either non-functional (pseudogenes) or subject to regulatory processes. It will therefore be of interest to determine whether these gene features are conserved in related *Bacillus* species, especially as strain 168 is derived from the Marburg strain that was subjected to X-ray irradiation21.

A few regions do not have any identifiable feature indicating that they are transcribed: they could be ‘grey holes’ of the type described in *E. coli*.24 Preliminary studies involving all regions of more than 400 bp without annotated CDSs indicated that, of ~300 such regions, only 15% were likely to be really devoid of protein-coding sequences. One of the longest such regions, located between yfjO and yfjN, is 1,628 bp long. Grey holes seem generally to be clustered near the terminus of replication. However, a grey-hole cluster located at ~600 kb might be related to the temporary chromosome partition observed during the first stages of sporulation, when a segment of about one-third of the chromosome enters the prespore, and remains the sole part of the chromosome in the prespore for a significant transition period25.

The codon usage of *B. subtilis* CDSs was analysed using factorial correspondence analysis27. We found that the CDSs of *B. subtilis* could be separated into three well-defined classes (Fig. 4). Class 1 comprises the majority of the *B. subtilis* genes (3,375 CDSs), including most of the genes involved in sporulation. Class 2 (188 CDSs) includes genes that are highly expressed under exponential growth conditions, such as genes encoding the transcription and translation machineries, core intermediary metabolism, stress proteins, and one-third of genes of unknown function. Class 3 (537 CDSs) contains a very high proportion of genes of unidentified function (84%), and the members of this class have codons enriched in A + T residues. These genes are usually clustered into groups between 15 and 160 genes (for example, bacteriophage SPB) and correspond to the A + T-rich islands described above (Fig. 1). When they are of known function, or when their products display similarity to proteins of known function, they usually correspond to functions found in, or associated with, bacteriophages or transposons, as well as functions related to the cell envelope. This includes the region ydcy/yddy/yde (40 genes that are missing in some *B. subtilis* strains28), where gene products showing similarities to bacteriophage and transposon proteins are intertwined. Many of these genes are associated with virulence genes identified in pathogenic Gram-positive bacteria, suggesting that such virulence factors are transmitted horizontally among bacteria at a much higher frequency than previously thought. If we include these A + T-rich regions as possible cryptic phages, together with known bacteriophages or bacteriophage-like elements (SPβ, PBSX and the skin element), we find that the genome of *B. subtilis* 168 contains at least 10 such elements (Figs 2 and 3). Annotation of the corresponding regions often reveals the presence of genes that are similar to bacteriophage lytic enzymes, perhaps accounting for the observation that *B. subtilis* cultures are extremely prone to lysis.

The ribosomal RNA genes have been previously identified and

---

**Table 1** Functional classification of the *Bacillus subtilis* protein-coding genes

The genes of known function or encoding products similar to known proteins in *B. subtilis* or in other organisms have been classified into functional categories (2,379 genes). The total number of genes in each category is indicated after the category title. Genes are listed in alphabetical order within each category, and their positions (in kilobases) on the *B. subtilis* chromosome are indicated after the gene names. A brief description is given for each gene. In some cases, interacting proteins have been indicated between brackets (for example, histidine kinases and response regulators, phosphatases and their substrates). More detailed and constantly updated information is available in the SubList database (see Methods). A preliminary assessment of the significance of sequence similarities was obtained through an automated procedure involving a combination between the BLAST2P probability and the percentage of amino-acid identity. Matches considered significant were re-examined manually. It should be emphasized that functions assigned to γ genes are based only on sequence similarity information with the best counterparts in protein databanks. Genes whose products are only similar to other unknown proteins, or not significantly similar to any other proteins in databanks (categories V and VI), were omitted.

---

Figure 2 General view of the *B. subtilis* chromosome. Arrows indicate the orientation of transcription. Genes are coloured according to their classification into six broad functional categories (blue, category I; green, category II; red, category III; orange, category IV; purple, category V; pink, category VI; see Table 1). Class 2 CDSs according to codon usage analysis are indicated by oblique hatches, and class 3 CDSs are indicated by vertical hatches. Ribosomal RNA genes are coloured in yellow. Transfer RNA genes are marked by triangles. Other RNA genes are represented as white arrows. Known genes (non-γ genes) are printed in bold type. Putative transcription termination sites are represented as loops. Known prophages and prophage-like elements are indicated by brown hatches on the chromosome line. The 190-bp element repeated ten times is represented by hatched boxes.
shown to be organized into ten rRNA operons, mainly clustered around the origin of replication of the chromosome (Figs 2 and 3). In addition to the 84 previously identified tRNA genes, by using the Palingo27 and tRNAscan28 programs, we propose four putative new tRNA loci (at 1,262 kb, 1,945 kb, 2,003 kb and 2,899 kb), specific for lysine, proline and arginine (UUU, GGG, CCU and UCU anticodons, respectively). The 10S RNA involved in degradation of proteins made from truncated mRNA has been identified (ssrA), as well as the RNA component of RNase P (rrnP) and the 4.5S RNA involved in the secretion apparatus (scr).

There is a strong transcription orientation bias with respect to the movement of the replication fork: 75% of the predicted genes are transcribed in the direction of replication. Plotting the density of coding nucleotides in each strand along the chromosome readily identifies the replication origin and terminus (Fig. 3). To identify putative operons, we followed ref. 29 for describing Rho-independent transcription termination sites. This yielded ~1,630 putative terminators (340 of which were bidirectional). We retained only those that were located less than 100 bp downstream of a gene, or that were considered by the program to be 'very strong' (in order to account for possible erroneous CDSs). This yielded a total of ~1,250 terminators, with a mean operon size of three genes. A similar approach to the identification of promoters is problematical, especially because at least 14 sigma factors, recognizing different promoter sequences, have been identified in B. subtilis. Nevertheless, the consensus of the main vegetative sigma factor (σv) appears to be identical to its counterpart in E. coli (σ70): 5′-TTGACA-n17-TATAAT-3′. Relaxing the constraints of the similarity to sigma-specific consensus sequences led to an extremely high number of false-positive results, suggesting that the consensus-oriented approach to the identification of promoters should be replaced by another approach27.

**Classification of gene products**

Genes were classified according to ref. 14, based on the representation of cells as Turing machines in which one distinguishes between the machine and the program (Table 1). Using the BLAST2P software running against a composite protein databank compound of SWISS-PROT (release 34), TREMBL (release 3, update 1) and B. subtilis proteins, we assigned at least one significant counterpart with a known function to 58% of the B. subtilis proteins. Thus for up to 42% of the gene products, the function cannot be predicted by similarity to proteins of known function: 4% of the proteins are similar only to other unknown proteins of B. subtilis; 12% are similar to unknown proteins from some other organism; and 26% of the proteins are not significantly similar to any other proteins in databanks. This preliminary analysis should be interpreted with caution, because only ~1,200 gene functions (30%) have been experimentally identified in B. subtilis. We used the ‘y’ prefix in gene names to emphasize that the function has not been ascertained (2,853 ‘y’ genes, representing 70%).

**Regulatory systems.** Transcription regulatory proteins. Helix–turn–helix proteins form a large family of regulatory proteins found in both prokaryotes and eukaryotes. There are several classes, including repressors, activators and sigma factors. Using BLAST searches, we constructed consensus matrices for helix–turn–helix proteins to analyse the B. subtilis protein library. We identified 18 sigma or sigma-like factors, of which nine (including a new one) are of the SigA type. We also putatively identified 20 regulators (among which 18 were products of ‘y’ genes) of the GntR family, 19 regulators (15 ‘y’ genes) of the LysR family, and 12 regulators (5 ‘y’ genes) of the LacI family. Other transcription regulatory proteins were of the AraC family (11 members, 10 ‘y’), the Lrp family (7 members, 3 ‘y’), the DeoR family (6 members, 3 ‘y’), or additional families (such as the MarR, ArsR or TetR families). A puzzling observation is that several regulatory proteins display significant similarity to aminotransferases (seven such enzymes have been identified as showing similarity to repressors).

**Two-component signal-transduction pathways.** Two-component regulatory systems, consisting of a sensor protein kinase and a response regulator, are widespread among prokaryotes. We have identified 34 genes encoding response regulators in B. subtilis, most of which have adjacent genes encoding histidine kinases. Response regulators possess a well-conserved N-terminal phospho-acceptor domain30, whereas their C-terminal DNA-binding domains share similarities with previously identified response regulators in E. coli, Rhizobium meliloti, Klebsiella pneumoniae or Staphylococcus aureus. Representatives of the four subfamilies recently identified in E. coli31.
Protein secretion. It is known that *B. subtilis* and related *Bacillus* species, in particular *B. licheniformis* and *B. amyoliquefaciens*, have a high capacity to secrete proteins into the culture medium. Several genes encoding proteins of the major secretion pathway have been identified: *secA, secD, secE, secF, phr* and *ftsY*. Surprisingly, there is no gene for the SecB chaperone. It is thought that other chaperone(s) and targeting factor(s), such as Ffh and FtsY, may take over the SecB function. Further, although there is only one such gene in *E. coli*, five type I signal peptidase genes (*sipS, sipT, sipU, sipV* and *sipW*) have been found. The *lsp* gene, encoding a type II signal peptidase required for processing of lipo-modified precursors, was also identified. *PrsA*, located at the outer side of the membrane, is important for the refolding of several mature proteins after their translocation through the membrane.

Other families of proteins. ABC transporters were the most frequent class of proteins found in *B. subtilis*. They must be extremely important in Gram-positive bacteria, because they have an envelope comprising a single membrane. ABC transporters will therefore allow such bacteria to escape the toxic action of many compounds. We propose that 77 such transporters are encoded in the genome. In general they involve the interaction of at least three gene products, specified by genes organized into an operon. Other families comprised 47 transport proteins similar to facilitators (and perhaps sometimes part of the ABC transport systems), 18 amino-acid permeases (probably antiporters), and at least 16 sugar transporters belonging to the PEP-dependent phosphotransferase system.

General stress proteins are important for the survival of bacteria under a variety of environmental conditions. We identified 43 temperature-shock and general stress proteins displaying strong similarity to *E. coli* counterparts.

**Missing genes.** Histone-like proteins such as HU and H-NS have been identified in *E. coli*. We found that *B. subtilis* encodes two putative histone-like proteins that show similarity to *E. coli* HU, namely HBSu and YonN, but found no homologue to H-NS. It is known that the *hbs* gene encoding HBSu is essential, but we do not expect the *yonN* gene to be essential because it is present in the SPβ prophage. IHF is similar to HU, and it is not known whether HBSu plays a similar role to that of IHF in *E. coli*. Similarly, no protein similar to FIS could be found.

Genes encoding products that interact with methylated DNA, such as *seqA* in *E. coli*, involved in the regulation of replication initiation timing, or *mutH*, the endonuclease recognizing the newly synthesized strand during mismatch repair at hemi-methylated DNA, are also identified. *PrsA*, located at the outer side of the membrane, is important for the refolding of several mature proteins after their translocation through the membrane.

**Other families of proteins.** ABC transporters were the most frequent class of proteins found in *B. subtilis*. They must be extremely important in Gram-positive bacteria, because they have an envelope comprising a single membrane. ABC transporters will therefore allow such bacteria to escape the toxic action of many compounds. We propose that 77 such transporters are encoded in the genome. In general they involve the interaction of at least three gene products, specified by genes organized into an operon. Other families comprised 47 transport proteins similar to facilitators (and perhaps sometimes part of the ABC transport systems), 18 amino-acid permeases (probably antiporters), and at least 16 sugar transporters belonging to the PEP-dependent phosphotransferase system.

General stress proteins are important for the survival of bacteria under a variety of environmental conditions. We identified 43 temperature-shock and general stress proteins displaying strong similarity to *E. coli* counterparts.

**Missing genes.** Histone-like proteins such as HU and H-NS have been identified in *E. coli*. We found that *B. subtilis* encodes two putative histone-like proteins that show similarity to *E. coli* HU, namely HBSu and YonN, but found no homologue to H-NS. It is known that the *hbs* gene encoding HBSu is essential, but we do not expect the *yonN* gene to be essential because it is present in the SPβ prophage. IHF is similar to HU, and it is not known whether HBSu plays a similar role to that of IHF in *E. coli*. Similarly, no protein similar to FIS could be found.

Genes encoding products that interact with methylated DNA, such as *seqA* in *E. coli*, involved in the regulation of replication initiation timing, or *mutH*, the endonuclease recognizing the newly synthesized strand during mismatch repair at hemi-methylated DNA, are also identified. *PrsA*, located at the outer side of the membrane, is important for the refolding of several mature proteins after their translocation through the membrane.
GATC sites, are also missing. This is in line with the absence of
known methylation in *B. subtilis*, equivalent to Dam methylation in
*E. coli*. Similarly, *E. coli* 5′A, encoding an inhibitor of FisZ action in
the SOS response, has no counterpart in *B. subtilis*. In contrast, *B.
subtilis* replication initiation-specific genes, such as dnaB and dnaD,
are missing in *E. coli*. The exact counterpart of the *E. coli* mukB gene,
involved in chromosome partitioning, does not exist in *B. subtilis*,
but genes spo11 and smc (Smc is weakly similar to MukB), which are
suggested to be involved in partitioning of the *B. subtilis* chromo-
some, are missing in *E. coli*.

Turnover of mRNA is controlled in *E. coli* by a ‘degradosome’
comprising RNase E. It has a counterpart in *B. subtilis*, but we failed
to find a clear homologue of RNase E in this organism. Whether this
is related to the role of ribosomal protein S1 as an RNA helicase
involved in mRNA turnover in *E. coli* requires further investigation.
In particular, a homologue of rpsA (S1 structural gene), ypfD, might
be involved in a structure homologous to the degradosome.

Structurally unrelated genes of similar function. Several genes
encode products that have similar functions in *E. coli* and *B. subtilis*,
but have no evident common structure. This is the case for the
helicase loader genes, *E. coli* dnaC and *B. subtilis* dnaA; the genes
coding for the replication termination protein, *E. coli* tus and *B.
subtilis* rtp; and the division topology specifier genes, *E. coli* minE
and *B. subtilis* divIVA. The situation may even be more complex in
multisubunit enzymes: *B. subtilis* synthesizes two DNA polymerase
III α chains, one having 3′–5′ proofreading exonuclease activity (PolC)
and the other without the exonuclease activity (DnaE); in *E.
coli*, only the latter exists. *E. coli* DNA polymerase II is structurally
related to DNA polymerase α of eukaryotes, whereas *B. subtilis* YshC
is related to DNA polymerase β.

Metabolism of small molecules

The type and range of metabolism used for the interconversion of
low-molecular-weight compounds provide important clues to an
organism’s natural environment(s) and its biological activity. Here
we briefly outline the main metabolic pathways of *B. subtilis* before
the reconstruction of these pathways *in silico*, the correlation of
genes with specific steps in the pathway, and ultimately the predic-
tion of patterns of gene expression.

Intermediary metabolism. It has long been known that *B. subtilis*
can use a variety of carbohydrates. As expected, it encodes an
Embden–Meyerhof–Parnas glycolytic pathway, coupled to a func-
tional tricarboxylic acid cycle. Further, *B. subtilis* is also able to grow
anaerobically in the presence of nitrate as an electron acceptor. This
metabolism is, at least in part, regulated by the FNR protein,
binding to sites upstream of at least eight genes (four sites experi-
mentally confirmed and four putative sites). A noteworthy feature of
*B. subtilis* metabolism is an apparent requirement of branched
short-chain carboxylic acids for lipid biosynthesis. Branch-
chain 2-keto acid decarboxylase activity exists and may be linked
to a variety of genes, suggesting that *B. subtilis* can synthesize and
utilize linear branched short-chain carboxylic acids and alcohols.

Amino-acid and nucleotide metabolism. Pyrimidine metabolism of
*B. subtilis* seems to be regulated in a way fundamentally different from
that of *E. coli*, as it has two carbamylphosphate synthetases
(one specific for arginine synthesis, the other for pyrimidine).
Additionally, the aspartate transcarbamylase of *B. subtilis* does not
act as an allosteric regulator as it does in *E. coli*. As in other
microorganisms, pyrimidine deoxyribonucleotides are synthesized from
ribonucleoside diphosphates, not triphosphates. The cytidine
diphosphate required for DNA synthesis is derived from either the
salvage pathway of mRNA turnover or from the synthesis of
phospholipids and components of the cell wall. This means that
polynucleotide phosphorylase is of fundamental importance in
nucleic acid metabolism, and may account for its important role in
competence. Two ribonucleoside reductases, both of class I, 
NrdE type, are encoded by the *B. subtilis* chromosome, in one case
from within the SPβ genome. In this latter case, the gene corre-
sponding to the large subunit both contains an intron and codes for
an intein (V.L., unpublished data). The gene of the small subunit of
this enzyme also contains an intron, encoding an endonuclease, as
was found for the homologue in bacteriophage T4.

By similarity with genes from other organisms, there appears to be,
in addition to genes involved in amino-acid degradation (such as
the roc operon, which degrades arginine and related amino acids),
a large number of genes involved in the degradation of molecules
such as opines and related molecules, derived from plants. This is
also in line with the fact that *B. subtilis* degrades polygalacturonate,
and suggests that, in its biotope, it forms specific relations with
plants.

Secondary metabolism. In addition to many genes coding for
deradigative enzymes, almost 4% of the *B. subtilis* genome codes
for large multifunctional enzymes (for example, the srf, pps and pks
loci), similar to those involved in the synthesis of antibiotics in other
genera of Gram-positive bacteria such as *Streptomyces*. Natural
isolates of *B. subtilis* produce compounds with antibiotic activity,
such as surfactin, fengycin and difficidin, that can be related to the
above-mentioned loci. This bacterium therefore provides a simple
and genetically amenable model in which to study the synthesis of
antibiotics and its regulation. These pathways are often organized in
very long operons (for example, the pks region spans 78.5 kb, about
2% of the genome). The corresponding sequences are mostly
located near the terminus of replication, together with prophages
and prophage-like sequences.

Paralogues and orthologues

It is important to relate intermediary metabolism to genome
structure, function and evolution. We therefore compared the *B.
subtilis* proteins with themselves, as well as with proteins from
known complete genomes, using a consistent statistical method that
allows the evaluation of unbiased probabilities of similarities
between proteins. For Z-scores higher than 13, the number of
proteins similar to each given protein does not vary, indicating that
this cut-off value identifies sets of proteins that are significantly
similar.

Families of paralogues. Many of the paralogues constitute large
families of functionally related proteins, involved in the transport of
compounds into and out of the cell, or involved in transcription
regulation. Another part of the genome consists of gene doublets
(568 genes), triplets (273 genes), quadruplets (168 genes) and
quintuplets (100 genes). Finally, about half of the genome is made of
genes coding for proteins with no apparent paralogues (Fig. 5).
No large family comprises only proteins without any similarity to
proteins of known function.

The process by which paralogues are generated is not well
understood, but we might find clues by studying some of the
duplications in the genome. Several approximate DNA repetitions,
associated with very high levels of protein identity, were found,
mainly within regions putatively or previously identified as pro-
phages. This is in line with previous observations about PBSX and
the skin element, and suggests that these prophage-like elements
share a common ancestor and have diverged relatively recently. In
addition, several protein duplications are in genes that are located
very close to each other, such as *yukL* and *dhbF* (the corresponding
proteins are 65% identical in an overlap of 580 amino acids), *ygJ*
and *yugK* (proteins 73% identical), *yxH* and *yxI* (proteins 70%
identical), and the entire *opuB* operon, which is duplicated 3 kb
away (*opuC* operon, yielding ~80% of amino-acid identity in the
corresponding proteins).

The study of paralogues showed that, as in other genomes, a few
classes of genes have been highly expanded. This argues against the
idea of the genome evolving through a series of duplications of
ancestral genomes, but rather for the idea of genes as living
organisms, subject to evolutionary constraints, some being sub-

---

254 Nature © Macmillan Publishers Ltd 1997
mitted to expansion and natural selection, and others to local duplications of DNA regions.

Among paralogue doublets, some were unexpected, such as the three aminocetyl tRNA synthetases doublets (hisS (2,817 kb) and hisZ (3,588 kb); thrS (2,960 kb) and thrZ (3,855 kb); tyrS (3,036 kb) and tyrZ (3,945 kb)) or the two mutS paralogues (mutS and yshD). This latter situation is similar to that found in Synechocystis. In the case of _B. subtilis_, the presence of two MutS proteins could indicate that there are two different pathways for long-patch mismatch repair, possibly a consequence of the active genetic transformation mechanism of _B. subtilis_.

_Families of orthologues._ Because _Mycoplasma_ spp. are thought to be derived from Gram-positive bacteria similar to _B. subtilis_, we compared the _B. subtilis_ genome with that of _M. genitalium_. Among the 450 genes encoded by _M. genitalium_, the products of 300 are similar to proteins of _B. subtilis_. Among the 146 remaining gene products, a further 3 are similar to proteins of other _Bacillus_ species, and 9 to proteins of other Gram-positive bacteria; 25 are similar to proteins of Gram-negative bacteria; and 19 are similar to proteins of other _Mycoplasma_ spp. This leaves only 90 genes that would be specific to _M. genitalium_ and might be involved in the interaction of this organism with its host.

The _B. subtilis_ genome is similar in size to that of _E. coli_. Because these bacteria probably diverged more than one billion years ago, it is of evolutionary value to investigate their relative similarity. About 1,000 _B. subtilis_ genes have clear orthologous counterparts in _E. coli_ (one-quarter of the genome). These genes did not belong either to the prophage-like regions or to regions coding for secondary metabolism (~15% of the _B. subtilis_ genome). This indicates that a large fraction of these genomes shared similar functions. At first sight, however, it seems that little of the operon structure has been conserved. We nevertheless found that ~100 putative operons or parts of operons were conserved between _E. coli_ and _B. subtilis_. Among these, ~12 exhibited a reshuffled gene order (typically, the arabinose operon is araABD in _B. subtilis_ and arabBAD in _E. coli_). In addition to the core of the translation and transcription machinery, we identified other classes of operons that were well conserved between the two organisms, including major integrated functions such as ATP synthesis (atp operon) and electron transfer (cya and qox operons). As well as being well preserved, the murein biogenesis region was partly duplicated, allowing creation of part of the murein biosynthesis machinery. A major technical difficulty was the inability to construct in _E. coli_ gene banks representative of the entire _B. subtilis_ chromosome using vectors that have proved efficient for other sources of bacterial DNA (such as bacteriophage or cosmids vectors). This was due to the generally very high level of expression of _B. subtilis_ genes in _E. coli_, leading to toxic effects. This limitation was overcome by: cloning into a variety of vectors44,45; using an _E. coli_ strain maintaining low-copy number plasmids46; using an integrative plasmid/marker rescue genome-walking strategy47; and in vitro amplification using polymerase chain reaction (PCR) techniques48,49.

Although cloning vectors were used in the early stages as templates for sequencing reactions, they were largely superseded in the later stages by long-range and inverse PCR techniques. To reduce sequencing errors resulting from PCR amplification artefacts, at least eight amplification reactions were performed independently and subsequently pooled. The various sequencing groups were free to choose their own strategy, except that all DNA sequences had to be determined entirely on both strands.

**Sequence annotation and verification.** The sequences were annotated by the groups, and sent to a central depository at the Institut Pasteur. The Japanese sequences were also sent through the Japanese depository at the Nara Institute of Science and Technology. The same procedures were used to identify CDSs and to detect frameshifts. They were embedded within a cooperative computer environment dedicated to automatic sequence annotation and analysis50. In a first step, we identified in all six possible frames the open reading frames (ORFs) that were at least 100 codons in length. In a second step, three independent methods were used: the first method used the GeneMark coding-sequence prediction method51 together with the search for CDSs preceded by typical translation initiation signals (5′-AAGGAGGTG-3′), the second method used the results of a BLAST2X analysis performed on the entire _B. subtilis_ genome against the non-redundant protein database at the NCBI; and the third method was based on the distribution of non-overlapping trinucleotides or hexanucleotides in the three frames of an ORF52.

In general, frameshifts and nonsense mutations generating termination codons or eliminating start codons are relatively easy to detect. We shall devise a procedure for detecting another type of error, GC instead of CG or vice versa, which are much more difficult to identify. It should be noted that putative frameshift errors should not be corrected automatically. The sequences of the flanking regions of a 500-bp fragment centred around a putative error were sent to an independent verification group, which performed PCR amplifications using chromosomal DNA as template, and sequenced the corresponding DNA products.

**Organization and accessibility of data.** The _B. subtilis_ sequence data have been combined with data from other sources (biochemical, physiological and genetic) in a specialized database, Subtilist53, available as a Macintosh or Windows stand-alone application (4th Dimension runtime) by anonymous FTP at ftp://ftp.pasteur.fr/pub/GenomeDB/Subtilist.html, or through a World-Wide Web server at http://www.pasteur.fr/Bio/Subtilist.html.
where it has been implemented on a UNIX system using the Symbase relational database management system. A completely rewritten version of SubtiList is available on the World Wide Web site.

Received 16 July; 29 September 1997.

1. Fleischmann, R. D. et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 269, 496–512 (1995).

2. Fraser, C. M. et al. The minimal gene complement of Mycoplasma genitalium. Science 270, 397–403 (1995).

3. Kaneko, T. et al. Sequence analysis of the unc111(Cyano bacterium Sycystis sp. strain PCC6803). II. Sequence determination of the entire genome and assignment of potential protein-coding regions. DNA Res. 3, 109–136 (1996).

4. Bult, C. J. et al. Complete genome sequence of the methanogenic archaeon, Methanobrevibacter smithii. Science 273, 1058–1073 (1996).

5. Himmelreich, R. et al. Comprehensive analysis of the Bacillus subtilis 4.0 genome. Nucleic Acids Res. 24, 292–305 (1996).

6. Daniel, R. A., Drake, S., Buchanan, C. E., Scholle, R. & Errington, J. The Bacillus subtilis chromosome structure. J. Bacteriol. 175, 741–749 (1993).

7. Billoud, B., Konttis, M. & Viari, A. Palings: a declarative programming language to describe nucleic acids' secondary structures and to scan sequence database. Nucleic Acids Res. 24, 1395–1403 (1996).

8. Fichant, G. A. & Burke, C. Identifying potential tRNA genes in genomic DNA sequences. J. Mol. Biol. 255, 1126–1135 (1996).

9. d'Aubenton Carafa, Y., Brody, E. & Therrier, C. Prediction of rho-independent Escherichia coli transcription terminators. A statistical analysis of their RNA stem-loop structures. J. Mol. Biol. 216, 369–390 (1991).

10. Ogasawara, N. & Yoshikawa, H. The systematic sequencing of the Bacillus subtilis genome fragment using a co-operative computer system prototype. Gene 185, GC37–GC51 (1995).

11. Harwood, C. R. The Bacillus subtilis genome project: technical overview. Microbiology 141, 297–344 (1995).

12. Solomon, J. M. & Grossman, A. D. Who's competent and when: regulation of natural genetic competence in bacteria. Trends Genet. 12, 150–155 (1996).

13. Solomon, J. M. & Grossman, A. D. Who's competent and when: regulation of natural genetic competence in bacteria. Trends Genet. 12, 150–155 (1996).

14. Moszer, I., Kunst, F. & Daniel, A. The European Bacillus subtilis genome sequencing project: current status and accessibility of the data from a new World Wide Web site. Microbiology 141, 297–345 (1995).

15. Harwood, C. R. The Bacillus subtilis genome project: technical overview. Microbiology 141, 297–344 (1995).

16. Himmelreich, R. et al. Comprehensive analysis of the Bacillus subtilis 4.0 genome. Nucleic Acids Res. 24, 292–305 (1996).

17. Harwood, C. R. The Bacillus subtilis genome project: technical overview. Microbiology 141, 297–344 (1995).

18. Solomon, J. M. & Grossman, A. D. Who's competent and when: regulation of natural genetic competence in bacteria. Trends Genet. 12, 150–155 (1996).

19. Moszer, I., Kunst, F. & Daniel, A. The European Bacillus subtilis genome sequencing project: current status and accessibility of the data from a new World Wide Web site. Microbiology 141, 297–345 (1995).

20. Harwood, C. R. The Bacillus subtilis genome project: technical overview. Microbiology 141, 297–344 (1995).

21. Solomon, J. M. & Grossman, A. D. Who's competent and when: regulation of natural genetic competence in bacteria. Trends Genet. 12, 150–155 (1996).

22. Moszer, I., Kunst, F. & Daniel, A. The European Bacillus subtilis genome sequencing project: current status and accessibility of the data from a new World Wide Web site. Microbiology 141, 297–345 (1995).

23. Harwood, C. R. The Bacillus subtilis genome project: technical overview. Microbiology 141, 297–344 (1995).

24. Solomon, J. M. & Grossman, A. D. Who's competent and when: regulation of natural genetic competence in bacteria. Trends Genet. 12, 150–155 (1996).

25. Moszer, I., Kunst, F. & Daniel, A. The European Bacillus subtilis genome sequencing project: current status and accessibility of the data from a new World Wide Web site. Microbiology 141, 297–345 (1995).
## 1. Functional classification of the *Bacillus subtilis* protein-coding genes.

| Process | Genes | Description |
|---------|-------|-------------|
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | yua | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | yug | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | yul | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | yul | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | yud | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | yue | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | yuu | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | yuv | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | yuw | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | yux | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | yuy | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | yuz | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | yvA | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | yvB | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | yvC | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | yvD | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | yvE | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | yvF | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | yvG | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | yvH | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | yvI | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | yvJ | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | yvK | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | yvL | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | yvM | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | yvN | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | yvO | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | yvP | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | yvQ | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | yvR | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | yvS | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | yvT | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | yvU | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | yvV | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | yvW | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | yvX | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | yvY | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | yvZ | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | ywA | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | ywB | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | ywC | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | ywD | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | ywE | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | ywF | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | ywG | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | ywH | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | ywI | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | ywJ | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | ywK | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | ywL | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | ywM | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | ywN | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | ywO | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | ywP | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | ywQ | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | ywR | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | ywS | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | ywT | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | ywU | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | ywV | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | ywX | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | ywY | ABC transporter (membrane protein) |
| LI. TRANSPORT/REGULATORY PROTEINS AND LIPOPROTEINS | ywZ | ABC transporter (membrane protein) |

**Note:** The table represents a portion of the functional classification of *Bacillus subtilis* protein-coding genes, focusing on specific enzymes and components related to various processes and cellular functions. The complete classification would include a more extensive list of genes and their respective roles in cellular metabolism and biosynthesis.
| Gene | Function |
|-----|----------|
| ytdA | antiporter (membrane protein) |
| ytdB | antiporter (ATP-binding protein) |
| ytdC | antiporter (membrane protein) |
| yscP | antiporter (permease) |
| yqjV | multidrug resistance protein |
| yqiY | amino acid ABC transporter (permease) |
| yqgJ | phosphate ABC transporter (ATP-binding protein) |
| yocN | permease |
| ynaJ | H+-symporter |
| yloB | calcium-transporting ATPase |
| ykpA | ABC transporter (ATP-binding protein) |
| yknY | ABC transporter (ATP-binding protein) |
| yknV | ABC transporter (ATP-binding protein) |
| yknU | ABC transporter (ATP-binding protein) |
| ykbA | amino acid permease |
| yjmG | hexuronate transporter |
| yjdD | fructose phosphotransferase system enzyme II |
| yhfQ | iron(III) dicitrate-binding protein |
| yheL | Na+/H+ antiporter |
| yheI | ABC transporter (ATP-binding protein) |
| yhdG | amino acid transporter |
| yhcL | sodium-glutamate symporter |
| yfmM | ABC transporter (ATP-binding protein) |
| yfmF | ferrichrome ABC transporter (ATP-binding protein) |
| yfmE | ferrichrome ABC transporter (permease) |
| yflA | aminoacid carrier protein |
| yfkL | multidrug resistance protein |
| yfiU | multidrug-efflux transporter |
| yfiS | multidrug resistance protein |
| yesO | sugar-binding protein |
| yeaB | cation efflux system membrane protein |
| ydjD | H+-symporter |
| ydgK | bicyclomycin resistance protein |
| ydgH | transporter |
| ydfA | arsenical pump membrane protein |
| ydeR | antibiotic resistance protein |
| ycbM | two-component sensor histidine kinase (YcbL) |
| citS | two-component sensor histidine kinase (CitT) |
| yycB | ABC transporter (permease) |
| yxlH | multidrug-efflux transporter |
| yxlA | purine-cytosine permease |
| yxkJ | metabolite-sodium symport |
| yxjA | pyrimidine nucleoside transport |
| yxeR | ethanolamine transporter |
| yxdL | ABC transporter (ATP-binding protein) |
| ywrK | arsenical pump membrane protein |
| ywrB | chromate transport protein |
| ywpC | large conductance mechanosensitive channel |
| ywaA | bacteriocin transport permease |
| ywcA | Na+-dependent symport |
| yvsH | ABC transporter (amino acid permease) |
| yvqJ | macrolide-efflux protein |
| yvkA | multidrug-efflux transporter |
| yvgW | heavy metal-transporting ATPase |
| yvgM | molybdenum transport permease |
| yvgK | molybdenum-binding protein |
| yvfR | ABC transporter (ATP-binding protein) |
| yvdG | maltose/maltodextrin-binding protein |
| yvdB | transporter |
| yvaE | multidrug-efflux transporter |
| yutK | Na+/nucleoside cotransporter |
| yusV | iron(III) dicitrate transport permease |
| yurM | sugar permease |
| yugO | potassium channel protein |
| yubD | multidrug resistance protein |
| yttB | multidrug resistance protein |
| ytrE | ABC transporter (ATP-binding protein) |
| ytrB | ABC transporter (ATP-binding protein) |
| ytmL | amino acid ABC transporter (permease) |
| ytrP | two-component sensor histidine kinase |
| ykoH | two-component sensor histidine kinase |
| yhcY | two-component sensor histidine kinase |
| ydfH | two-component sensor histidine kinase |
| ydbF | two-component sensor histidine kinase |
| yclK | two-component sensor histidine kinase |
| cccB | cytochrome c551 |
| cccA | cytochrome c550 |
| atpG | ATP synthase (subunit) |
| atpF | ATP synthase (subunit) |
| atpE | ATP synthase (subunit) |
| atpD | ATP synthase (subunit) |
| atpC | ATP synthase (subunit) |
| atpB | ATP synthase (subunit) |
| atpA | ATP synthase (subunit) |
| atpH | ATP synthase (subunit) |
| atpG | ATP synthase (subunit) |
| atpF | ATP synthase (subunit) |
| atpE | ATP synthase (subunit) |
| atpD | ATP synthase (subunit) |
| atpC | ATP synthase (subunit) |
| atpB | ATP synthase (subunit) |
| atpA | ATP synthase (subunit) |
| atpH | ATP synthase (subunit) |
| atpG | ATP synthase (subunit) |

I.A. MEMBRANE BIOENERGETICS (ELECTRON TRANSPORT CHAIN AND ATP SYNTHESIS)
Table 1. (continuation) Functional characterization of the Bacillus subtilis protein-coding genes.

| Gene | Product | Functional Class | Subunit | Subunit | Subunit | Subunit | Subunit | Subunit | Subunit | Subunit | Subunit | Subunit | Subunit | Subunit | Subunit | Subunit | Subunit | Subunit |
|------|---------|-----------------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| yqA  | 2436    | 2,9-dihydroxy-9,10-anthracene-9-carboxylic acid dioxygenase | (pyrene biosynthesis) |
| yqB  | 2437    | 2,9-dihydroxy-9,10-anthracene-9-carboxylic acid dioxygenase | (pyrene biosynthesis) |
| yqC  | 2438    | 2,9-dihydroxy-9,10-anthracene-9-carboxylic acid dioxygenase | (pyrene biosynthesis) |
| yqD  | 2439    | 2,9-dihydroxy-9,10-anthracene-9-carboxylic acid dioxygenase | (pyrene biosynthesis) |
| yqE  | 2440    | 2,9-dihydroxy-9,10-anthracene-9-carboxylic acid dioxygenase | (pyrene biosynthesis) |
| yqF  | 2441    | 2,9-dihydroxy-9,10-anthracene-9-carboxylic acid dioxygenase | (pyrene biosynthesis) |
| yqG  | 2442    | 2,9-dihydroxy-9,10-anthracene-9-carboxylic acid dioxygenase | (pyrene biosynthesis) |
| yqH  | 2443    | 2,9-dihydroxy-9,10-anthracene-9-carboxylic acid dioxygenase | (pyrene biosynthesis) |
| yqI  | 2444    | 2,9-dihydroxy-9,10-anthracene-9-carboxylic acid dioxygenase | (pyrene biosynthesis) |
| yqJ  | 2445    | 2,9-dihydroxy-9,10-anthracene-9-carboxylic acid dioxygenase | (pyrene biosynthesis) |
| yqK  | 2446    | 2,9-dihydroxy-9,10-anthracene-9-carboxylic acid dioxygenase | (pyrene biosynthesis) |
| yqL  | 2447    | 2,9-dihydroxy-9,10-anthracene-9-carboxylic acid dioxygenase | (pyrene biosynthesis) |
| yqM  | 2448    | 2,9-dihydroxy-9,10-anthracene-9-carboxylic acid dioxygenase | (pyrene biosynthesis) |
| yqN  | 2449    | 2,9-dihydroxy-9,10-anthracene-9-carboxylic acid dioxygenase | (pyrene biosynthesis) |
| yqO  | 2450    | 2,9-dihydroxy-9,10-anthracene-9-carboxylic acid dioxygenase | (pyrene biosynthesis) |
| yqP  | 2451    | 2,9-dihydroxy-9,10-anthracene-9-carboxylic acid dioxygenase | (pyrene biosynthesis) |
| yqQ  | 2452    | 2,9-dihydroxy-9,10-anthracene-9-carboxylic acid dioxygenase | (pyrene biosynthesis) |
| yqR  | 2453    | 2,9-dihydroxy-9,10-anthracene-9-carboxylic acid dioxygenase | (pyrene biosynthesis) |
| yqS  | 2454    | 2,9-dihydroxy-9,10-anthracene-9-carboxylic acid dioxygenase | (pyrene biosynthesis) |
| yqT  | 2455    | 2,9-dihydroxy-9,10-anthracene-9-carboxylic acid dioxygenase | (pyrene biosynthesis) |

Nature © Macmillan Publishers Ltd 1997
IV.7. TERMINATION ..................................................3

ytfA 1784
ytfB 1785
ytfC 1788
ytfD 1789
ytfE 1790
ytfF 1791
ytfG 1792

IV.2. PROTEIN MODIFICATION .....................................2

amM 325
amyA 326
bgl 328
bglG 328
bglH 328

V.2. FROM OTHER ORGANISMS .................................491

yurV 3357
yuaG 3181
yrvO 2811
yrkA 2720
yplQ 2295
ycbU 287
ybaL 157
yacI 102
eag 1430
csfB 36
ydcR 535
ydcP 533
yddB 537
ydcL 530
ycdD 304
xtrA 1324
xtmA 1325
xkdY 1345
xkdS 1340
xkdR 1340
xkdM 1333
xkdK 1332
xkdI 1331
xkdC 1322

V.3 Antibiotic Resistance ..........................................59

yurV 332
yurS 333
yurT 334
yurU 335

V.4. PHAGE-RELATED FUNCTIONS .............................83

ytfG 1792
ytfH 1793
ytfI 1794
ytfJ 1795
ytfK 1796
ytfL 1797
ytfM 1798
ytfN 1799
ytfO 1800
ytfP 1801
ytfQ 1802
ytfR 1803
ytfS 1804
ytfT 1805
ytfU 1806
ytfV 1807
ytfW 1808
ytfX 1809
ytfY 1810
ytfZ 1811

V.5. TRANSDUCTION AND IS ..................10

yjF 3110
yjG 3111
yjH 3112
yjI 3113
yjJ 3114
yjK 3115
yjL 3116
yjM 3117
yjN 3118
yjO 3119
yjP 3120
yjQ 3121
yjR 3122
yjS 3123
yjT 3124
yjU 3125
yjV 3126
yjW 3127
yjX 3128
yjY 3129
yjZ 3130

V.6. MISCELLANEOUS ...............................................26

yurV 3357
yurS 333
yurT 334
yurU 335

VI. Similarity to unknown Proteins .........................46

ydiE 643
ydiC 642

Nature © Macmillan Publishers Ltd 1997