FEBS Letters 389 (1996) 84–87 FEBS 17135

Minireview

Sequencing and functional analysis of the genome of *Bacillus subtilis* strain 168

Colin R. Harwood*, Anil Wipat

Department of Microbiology, School of Microbiological, Immunological and Virological Sciences, The Medical School, University of Newcastle upon Tyne, Framlington Place, Newcastle upon Tyne NE2 4HH, UK

Received 10 May 1996

Abstract The international programme to sequence the 4.2 Mb genome of *Bacillus subtilis*, a model Gram-positive bacterium, is a joint project involving European, Japanese and US research groups. To date ca. 3.0 Mb of the genome has been sequenced, with the remaining 1.2 Mb expected to be completed in 1997. The amenability of *B. subtilis* to genetic manipulation, combined with the availability of extensive expertise on its biochemistry and physiology, makes this bacterium a valuable organism in which to investigate the properties of genes for which functions cannot be readily ascribed by standard methods.

Key words: Genome sequencing; Gene function; Gram-positive bacteria; Yeast artificial chromosome; Chromosome walking; Integrational vector

1. Introduction

Bacteria of the genus *Bacillus* (type strain *Bacillus subtilis* Marburg) are aerobic, endospore-forming Gram-positive rods. The genus is one of the most widely distributed and metabolically diverse, with representatives found in the soil and associated water sources [1]. *Bacillus* species have a long history of exploitation by mankind [2,3]; their metabolic diversity has been employed in a wide range of industrial processes, including the production of hydrolytic enzymes (e.g. alkaline proteases, α-amylases [4]), polypeptide antibiotics (e.g. bacitracin, gramicidins, polymyxins [5]), biochemicals (e.g. nucleosides for conversion to flavour enhancers [6]) and insecticides (e.g. δ-endotoxins [7]). *Bacillus subtilis* (var. *natto*) has been used throughout this century for the fermentation of soybeans into Natto, a traditional Japanese food. The consumption of some 10^8 kg of Natto annually, the low reported incidence of pathogenicity and the widespread use of its products, and those of its close relatives, in the food, beverage and detergents industries has resulted in the granting of GRAS (generally regarded as safe) status to *B. subtilis* by the U.S. Food and Drug Administration (FDA).

Our understanding of the biochemistry, physiology and genetics of *B. subtilis* is second only to that of *Escherichia coli* (see [8]), from which it diverged more than 2 billion years ago [9]. Three important characteristics of *B. subtilis* are responsible for the extent of this knowledge base: (i) the elaboration of an efficient natural genetic transformation system [10], the first discovered in non-pathogenic micro-organism [11]; (ii) the production of commercially important hydrolytic enzymes and bioactive compounds [2]; (iii) the ability to differentiate into heat-, desiccation- and chemical-resistant endospores [12]. Most molecular biological and genetic methods that are available for *E. coli* have been developed for *B. subtilis* [13,14]. However, one particular advantage this bacterium has over *E. coli* is that genes are very easily introduced into the chromosome using the active recombination pathways in *B. subtilis* [15]. This greatly facilitates the generation of transcriptional and translation fusions, insertional mutations and merodiploids, and the introduction of controllable promoters upstream of target genes. This property is being fully exploited in the gene functional analysis programme (see Section 4, below).

B. subtilis and closely related bacilli secrete industrially important enzymes directly into the growth medium at concentrations in excess of 10 g/l and this has provided the basis of much of the commercial interest in these bacteria [2,16]. However, attempts to use *B. subtilis* to secrete heterologous proteins at commercially significant concentrations have, with few exceptions, met with little success. The reasons for these failures have proved to be complex, but include the stability of cloning vectors, the production of extracellular proteases and incompatibilities between the secretion apparatus and target proteins. Many of these problems are currently being tackled in a EU-funded research programme, co-ordinated by Dr Sierd Bron (Groningen, The Netherlands).

2. Genome sequencing programme

The genome of *B. subtilis* strain 168 is approximately 4.2 Mb in size and has the benefit of extensive genetic [17] and physical [18] maps. The international programme to sequence the genome involves European, Japanese and US research groups. The genome has been divided into regions which, in most cases, are bordered by previously sequenced genes and assigned to participants in the consortium (Fig. 1). The European groups are responsible for sequencing 3.9 Mb. They are funded by the European Commission under its Biotechnology Programme and are co-ordinated by Dr Frank Kunst (Institut Pasteur, Paris [19]). The Japanese groups, responsible for 1.3 Mb, are supported by the Ministry of Education, Science and Culture of Japan and are co-ordinated by Professor Naotake Ogasawara (Nara Institute of Science and Technology [20]). To date more than 70% of the genome has been sequenced (3 Mb) and the entire sequence should be available within 12 months.

The European sequence data are collected and verified at the Institut Pasteur under the supervision of Dr Antoine Danchin [21], where it is complied into a database called *SubtiList*.

*Corresponding author. Fax: (44) (191) 222 7736.
E-mail: colin.harwood@ncl.ac.uk

PII S0014-5793(96)00524-8

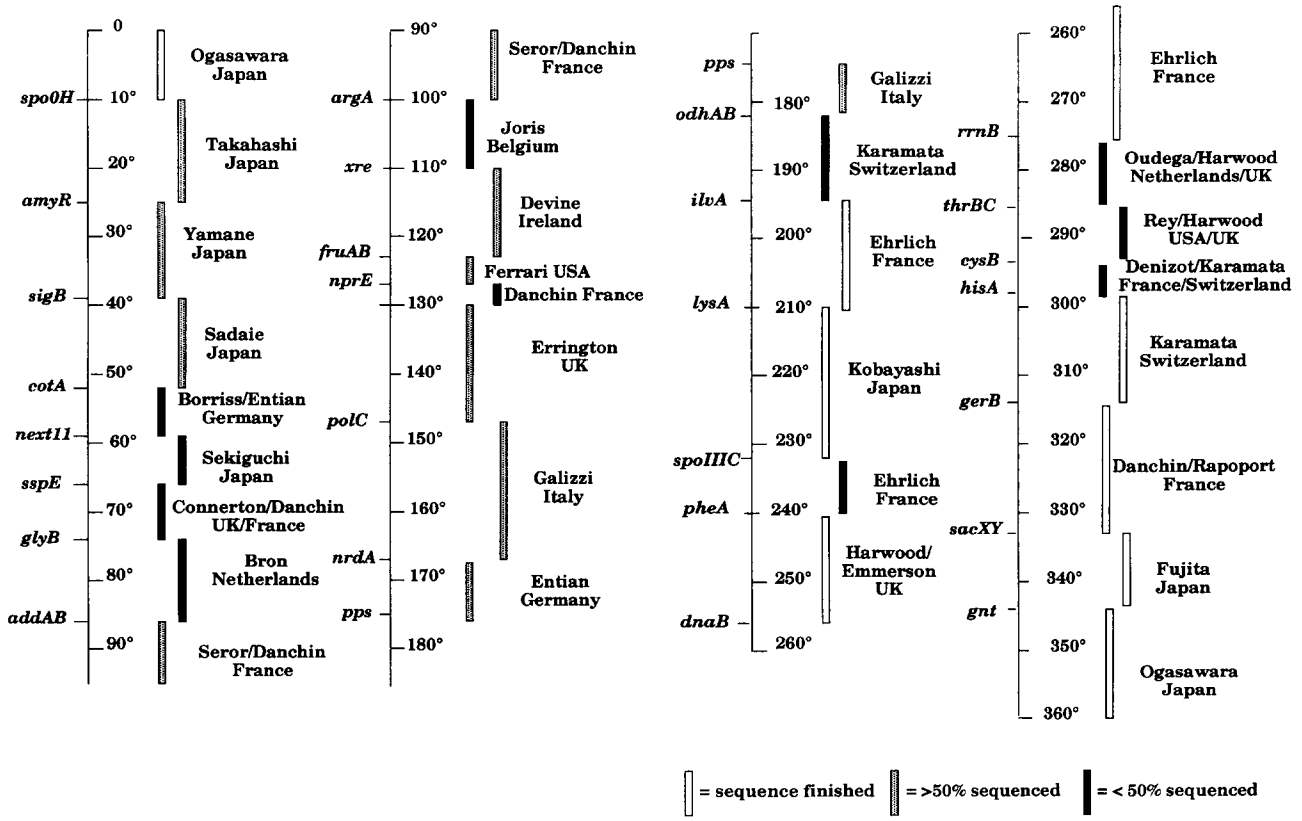

Fig. 1. Assignment of regions of the *B. subtilis* genome.

SubtiList is implemented on microcomputers with the Macintosh operating system via the 4th Dimension© relational database software (4D, ACI). Its user-friendly interface allows easy access to information on the genetic map, DNA and protein sequences, sequenced contigs, locations of control regions, gene descriptions and selected literature. The latest publicly available version of the *SubtiList* (currently ver. 10) is available on-line through an anonymous ftp from 'ftp.pasteur.fr' or IP number 157.99.64.12 in the directory '/pub/GenomeDB/SubtiList'.

3. Sequencing strategies

Systemic sequencing of microbial genomes has been achieved either by shotgun cloning the entire chromosome and assembling the resulting sequences in silico [22] or by creating an ordered library of large, overlapping chromosomal fragments which are subsequently sequenced individually [23]. The multi-national architecture of the *B. subtilis* sequencing consortium has dictated the latter approach, with individual groups being responsible for developing their own sequencing strategy. Whilst cosmid libraries have been used to create overlapping libraries of individual *Saccharomyces cerevisiae* [23] and *Caenorhabditis elegans* [24] chromosomes and for the genome of *Mycobacterium leprae* [25], and lambda libraries for the *E. coli* genome [26], it has proved impossible to construct cosmid or lambda libraries that completely cover the *B. subtilis* chromosome. This is because large fragments of *B. subtilis* DNA have been found to be unstable in *E. coli* due to their high AT content (43% GC) and the elevated levels of expression of many *B. subtilis* genes in this host [27,28].

The frequent instability and/or toxicity of *B. subtilis* DNA in *E. coli* has necessitated the development of alternative genome sequencing strategies. Azevedo and co-workers [29] avoided the use of *E. coli* as an intermediate host for generating a collection of large, overlapping DNA fragments by constructing an ordered library of the *B. subtilis* genome in yeast artificial chromosomes (pYACs). A collection of 59 pYAC clones was mapped by hybridisation, using previously cloned genes as probes, and the pYACs ordered in four contigs that covered >98% of the *B. subtilis* chromosome. PFGE-purified pYAC DNA was fragmented and cloned into M13-based vectors for sequencing [30]. The generation of sufficient quantities of pYAC DNA for subsequent fragmentation and cloning remains the main limitation of this approach.

Glaser and colleagues [27] addressed the problem of the instability of large fragments of *B. subtilis* DNA in *E. coli* by developing a method for directed chromosome walking. This method exploits the ease with which *B. subtilis* can take up exogenous DNA and integrate it into its chromosome by homologous recombination [15]. An *E. coli* plasmid (pDIA5304), unable to replicate in *B. subtilis* and carrying a fragment from the end of a cloned DNA region, is integrated into the chromosome by a Campbell-type recombination event [27]. The adjacent chromosomal DNA is then rescued by digestion with appropriate restriction endonucleases, self-ligated and then transformed back into *E. coli* using the origin of replication and antibiotic resistance marker of the original integration vector. Even using this methodology, instability/toxicity can still present problems. However, the use of an *E. coli* host which maintains normally high copy number plasmids at a low copy number, such as *E. coli* TP611 [31], has

facilitated the cloning and maintenance of many previously unstable fragments.

Despite the development of new methodologies, we and others have continued to find lambda clones a valuable starting material for sequencing, particular when used in combination with newer technologies for filling gaps between sequence islands. A particularly useful innovation has been to use pYAC clones from the Azevedo collection [29] as hybridisation probes for the selection of a set of λ-phage clones covering specific regions of the chromosome.

The advent of long accurate PCR permits the amplification of chromosomal segments of up to 35 kb [32]. In addition, the optimisation of the enzymology for dye terminator and dye primer technology for automated sequencing has simplified methods for the direct sequencing of PCR products and for the automation of template preparation. This has led us to develop an approach based on the use of a conventional shotgun sequencing procedure into a general purpose *E. coli* sequencing vector [33]. DNA from lambda clones or DNA generated by long PCR from lambda inserts or chromosomal DNA is randomly fragmented by treatment with DNase I in the presence of manganese ions and fragments of between 0.5 and 2.5 kb cloned into pUC18. After automated sequencing, DNA sequences are assembled using a contig building programme and any gaps filled by primer walking, by sequencing directly from PCR products or by the fragmentation of gap-filling PCR products and cloning as smaller fragments (250–500 bases).

We have used these techniques to complete the 114 kb region from *pheA* (240°) to *dnaB* (256°). Typically, we and others have found that $>90\%$ of the coding capacity encodes for ORFs, four out of five of which are orientated in the direction of movement of the replication fork. All three initiation codons are used, with a frequency ATG $\gg$ TTG $>$ GTG. In addition we have found the unusual ATT start codon in front of the gene coding for initiation factor 3 (IF3, unpublished). Putative ribosome binding sites are generally easy to identify because they are more conserved with respect to the 3′-end of 16S rRNA than are their counterparts in *E. coli*. This has been attributed to the absence in *B. subtilis* of a ribosomal protein equivalent to the S1 protein of *E. coli* [34]. In contrast, *B. subtilis* elaborates at least 10 distinct sigma factors [35], each of which direct RNA polymerase to a distinct set of promoter sequences. Consequently, searches for promoter sequences in silico are at best problematical and sometimes misleading, and we prefer to await confirmation by promoter mapping studies.

Another feature that is emerging from sequencing and concomitant functional analysis studies of *B. subtilis* is the apparent duplication of enzymic activities. For example, *B. subtilis* encodes at least three signal peptidases [36,37], up to four distinct alkaline phosphatases [38] and two Lon proteases ([39] and unpublished). Whilst it is not always obvious what benefits are conferred by this redundancy, it appears in some cases that distinct enzyme activities are required during vegetative growth and sporulation.

4. Systematic function analysis

About half of the genes identified during the systematic sequencing of bacterial genomes have no function ascribed to them – they either have no homology with sequences in the databases or the function of their homologues is also unknown. In order to optimise the value of the data from the *B. subtilis* genome sequencing project, and to facilitate increased industrial usage of *B. subtilis*, a consortium of 19 European laboratories, co-ordinated by Dr S. Dusko Ehrlich (INRA, Jouy en Josas, France), has recently initiated a systematic function analysis of *B. subtilis* genes. A parallel programme has been initiated in Japan on genes identified in their sequencing programme. The prime objective is to assign target genes to categories of cellular function including: (i) the metabolism of small molecules; (ii) macromolecular metabolism; (iii) cell structures; (iv) stress and stationary phase; and (v) cell processes. More detailed secondary and tertiary analysis will follow. The work programme is divided into resource and function-oriented consortia. The resource consortium is responsible for the production of a collection of standard mutant strains and their primary characterisation under defined growth conditions by transcriptional analysis and 3D protein profiles. Mutants are generated with integrational vectors, developed in Ehrlich's laboratory, that facilitate tandem duplication of the target gene. One copy of the gene is a non-functional deletion under the control of its native promoter and transcriptionally fused to a *lacZ* reporter. The second, wild-type copy of the gene is under the control of an IPTG-inducible promoter. The latter permits the selection and analysis of mutations even of genes that are essential for growth and/or viability.

The function-oriented consortium will assign target genes to the various function categories. This will be achieved systematically at three levels, initially as a tentative assignment by high throughput phenotypic tests, then by more elaborated tests and finally by specific detailed analysis. A relational database, MICADO (MICrobial Advanced Database Organisation), in under construction at INRA for the collation of genetic and physical maps, DNA sequences, microbial strains and the data derived from the functional analysis programme. The database can be accessed on the World Wide Web at 'http://locus.jouy.inra.fr/'.

The combination of highly co-ordinated genome sequencing and functional analysis programmes is unique among bacterial systems, the result of unprecedented co-operation between researchers in three continents. In addition to their value to the *Bacillus* scientific and industrial communities, the results will be of interest to those involved in the analysis of organisms for which genome sequences are known, but for which genetic manipulation and physiological studies are either impossible or, at best, extremely difficult to perform.

Acknowledgements: Our work on the sequencing and functional analysis of the *B. subtilis* genome is funded by the European Commission under the Biotechnology Programme, Contract BIO2-CT93-0272 and BIO4-CT96-0278. We also acknowledge information from Dr Frank Kunst and Dr Ivan Moszer for the compilation of Fig. 1.

References

[1] Priest, F.G. (1989) In: Biotechnology Handbook 2: *Bacillus* (C.R. Harwood, Ed.) pp. 27–56. Plenum, New York.
[2] Harwood, C.R. (1992) TIBTECH 10, 247–256.
[3] Priest, F.G. and Harwood, C.R. (1995) In: Food Biotechnology; Micro-organisms (Y.H. Hui and G.G. Khachatourians, Eds.) pp. 377–421. VCH Publishers, New York.
[4] Farrari, E., Jarnagin, A.S. and Schmidt, B.F. (1993) In: *Bacillus subtilis* and Other Gram-positive Bacteria (A.L. Sonenshein, J.A.

Hoch and R. Losick, Eds.) pp. 917–937. ASM Press, Washington, DC.
[5] Kleinkauf, H. and von Dohren, H. (1992) Eur. J. Biochem. 192, 1–15.
[6] Kuninaka, A. (1986) In: Biotechnology (H.J. Rehm and G. Reed, Eds.) pp. 71–114. VCH, Weinheim.
[7] Feitelson, J.S., Paynes, J. and Kim, L. (1992) Bio/Technology 10, 271–275.
[8] Sonnenshein, A.L., Hoch, J.A. and Losick, R. (1993) *Bacillus subtilis* and Other Gram-positive Bacteria. ASM Press, Washington, DC.
[9] Woese, C.R. (1987) Microbiol. Rev. 51, 221–271.
[10] Dubnau, D. (1991) Microbiol. Rev. 55, 395–424.
[11] Spizizen, J. (1958) Proc. Natl. Acad. Sci. USA 44, 407–408.
[12] Errington, J. (1993) Microbiol. Rev. 57, 1–33.
[13] Cutting, S.M. and Vander Horn, P.B. (1992) In: Molecular Biological Methods for *Bacillus* (C.R. Harwood and S.M. Cutting, Eds.) pp. 29–74. John Wiley and Sons, Chichester.
[14] Errington, J. (1990) In: Molecular Biological Methods for *Bacillus* (C.R. Harwood and S.M. Cutting, Eds.) pp. 175–220. John Wiley and Sons, Chichester.
[15] Youngman, P. (1990) In: Molecular Biological Methods for *Bacillus* (C.R. Harwood and S.M. Cutting, Eds.) pp. 221–265. John Wiley and Sons, Chichester.
[16] Simonen, M. and Palva, I. (1993) Microbiol. Rev. 57, 109–137.
[17] Anagostopoulos, C., Piggot, P.J. and Hoch, J.A. (1993) In: *Bacillus subtilis* and Other Gram-positive Bacteria (A.L. Sonenshein, J.A. Hoch and R. Losick, Eds.) pp. 425–461. ASM Press, Washington, DC.
[18] Itaya, M. (1993) In: *Bacillus subtilis* and Other Gram-positive Bacteria (A.L. Sonenshein, J.A. Hoch and R. Losick, Eds.) pp. 463–471. ASM Press, Washington, DC.
[19] Kunst, F., Vassarotti, A. and Danchin, A. (1995) Microbiology 141, 249–255.
[20] Ogasawara, N., Fujita, Y., Kobayashi, Y., Sadaie, Y., Tanaka, T., Takahashi, H., Yamane, K. and Yoshikawa, H. (1995) Microbiol. 141, 257–259.
[21] Moszer, I., Glaser, P. and Danchin, A. (1995) Microbiol. 141, 261–268.
[22] Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.-F., Dougherty, B.A., Merrick, J.M., Mckenney, K., Sutton, G., Fitzhugh, W., Fields, C.A., Gocayne, J.D., Scott, J.D., Shirley, R., Liu, L-I., Glodek, A., Kelley, J.M., Weidman, J.F., Phillips, C.A., Spriggs, T., Hedblom, E., Cotton, M.D., Utterback, T.R., Hanna, M.C., Nguyen, D.T., Saudek, D.M., Brandon, R.C., Fine, L.D., Fritchman, J.L., Fuhrmann, J.L., Geoghagen, N.S.M., Gnehm, C.L., Mcdonald, L.A., Small, K.V., Fraser, C.M., Smith, H.O. and Venter J.C. (1995) Science 269, 496–512.
[23] Oliver, S.G. et al. (1992) Nature 357, 38–46.
[24] Sulston J., Du, Z., Thomas, K., Wilson, R., Hillier, L., Staden, R., Halloran, N., Green, P., Thierry-Mieg, J., Qiu, L., Dera, S., Coulson, A., Craxton, M., Durbin, R., Berks, M., Metzstein, M., Hawkins, T., Ainscough, R. and Waterston, R. (1992) Nature 356, 37–41.
[25] Honoré, N., Bergh, S., Chanteau, S., Doucet-Populaire, F., Eiglmeier, K., Garnier, T., Georges, C., Launois, P., Limpalboon, T., Newton, S., Niang, K., del Portillo, P., Ramesh, G.R., Reddi, P., Ridel, P.R., Sittisombut, N., Wu-Hunter, S. and Cole, S.T. (1993) Mol. Microbiol. 7, 207–214.
[26] Kohara, Y., Akiyama, K. and Isono, K. (1987) Cell 50, 495–508.
[27] Glaser, P., Kunst, F., Arnaud, M., Coudart, M.-P., Gonzarles, W., Hullo, M.-F., Ionescu, M., Lubochinsky, S., Marcelino, L., Moszer, I., Presecan, E., Santana, M., Schneider, E., Schweizer, J., Vertes, A., Rapoport, G. and Danchin, A. (1993) Mol. Microbiol. 10, 371–384.
[28] Ogasawara, N., Nakai, S. and Yoshikawa, N. (1994) DNA Res. 1, 1–14.
[29] Azevedo, V., Alvarez, E., Zumstein, E., Damiani, G., Sgaramella, S., Ehrlich, S.D. and Serror, P. (1993) Proc. Natl. Acad. Sci. USA 90, 6047–6051.
[30] Sorokin, A., Zumstein, E., Azevedo, S., Ehrlich, S.D. and Serror, P. (1993) Mol. Microbiol. 10, 385–395.
[31] Hedegaars, L. and Danchin, A. (1985) Mol. Gen. Genet. 201, 38–42.
[32] Barnes, W.M. (1994) Proc. Natl. Acad. Sci. USA 90, 6047–6051.
[33] Viera, J. and Messing, J. (1987) Methods Enzymol. 153, 3–15.
[34] Farwell, M.A., Roberts, M.W. and Rabinowitz, J.C. (1992) Mol. Microbiol. 6, 3375–3383.
[35] Moran, C.P. (1993) In: *Bacillus subtilis* and Other Gram-positive Bacteria (A.L. Sonenshein, J.A. Hoch and R. Losick, Eds.) pp. 653–667. ASM Press, Washington, DC.
[36] Van Dijl, J.M., de Jong, A., Vehmaanperä, J., Venema, G. and Bron, S. (1992) EMBO J. 11, 2819–2828.
[37] Akagawa, E., Kurita, K., Sugawara, T., Nakamura, K., Kasahara, Y., Ogasawara, N. and Yamane, K. (1995) Microbiology 141, 3241–3245.
[38] Hulett, F.M. (1993) In, *Bacillus subtilis* and Other Gram-positive Bacteria (A.L. Sonenshein, J.A. Hoch and R. Losick, Eds.) pp. 229–235. ASM Press, Washington, DC.
[39] Riethdorf, S., Völker, U., Gerth, U., Winkler, A., Engelmann S. and Hecker M. (1994) J. Bacteriol. 176, 6518–6527.