Hypothesis

Duplicated sporulation genes in bacteria

Implications for simple developmental systems

J. Errington, P. Fort* and J. Mandelstam+

Microbiology Unit, Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, England and **Laboratoire de Biologie Moléculaire, Université de Sciences et Techniques du Languedoc, Place E. Bataillou, 34060 Montpellier Cedex, France*

Received 24 June 1985

Statistical analysis indicates that about 150 genes might be needed for spore formation in *Bacillus subtilis* [1]. We have sequenced 11 of these and examined the predicted amino acid sequences for the possibility that present-day sporulating bacteria may have evolved from ancestral forms in which sporulation was governed by a smaller number of genes. We have identified two pairs of duplicated genes. The first of them involves a sequence which shows major homologies with both the sigma factor (σ) of RNA polymerase in *Escherichia coli* and σ^{29} in *B. subtilis*. Another example of homology involves no less than 4 separate genes coding for the acid-soluble proteins of the spore [2]. We conclude that sporulation may have evolved from a much smaller number of genes that we first envisaged and it is likely, anyway, to be no more complex than replication of the larger coliphages.

Bacterial sporulation *Ribosome* *Phage replication* *Genetic regulation of development* *Sigma factor*

1. INTRODUCTION

Spore formation in *Bacilli* is an example of a very primitive developmental change. As such, it should be susceptible to description in terms of molecular biology and it may be useful to compare it with 2 other types of structure that are assembled in the cytoplasm of the bacterial cell. The first is ribosome formation which involves self-assembly and can take place in vitro. Next in apparent complexity, is phage replication which requires the metabolic equipment of the host cell and which involves both regulated gene expression and self-assembly. Sporulation at first sight appears to be even more complex, beginning with an asymmetric cell division, followed by the engulfment of the 'daughter cell' by the 'mother cell' and ending with the self-assembly of the proteins of the outer structure [3].

Sporulation is divided into 6 stages defined on the basis of the morphological changes observed by electron microscopy [4]. The process is blocked by mutations in a number of genetic loci which are designated according to the stage that is affected. Thus, the mutations that stop sporulation at, say, stage III are contained in the loci *spoIIIA, spoIIIB* etc. About 50 loci, which will be referred to as operons, are known [5] and the upper limit to their number has been estimated to be about 70 [1]. The morphological changes have been shown, mainly by studies in *Bacillus subtilis*, to be integrally associated with specific biochemical events such as the formation of alkaline phosphatase, glucose dehydrogenase, dipicolinic acid, etc. A

+ To whom correspondence should be addressed

Published by Elsevier Science Publishers B.V. (Biomedical Division)
00145793/85/$3.30

mutation that blocks sporulation at a particular stage prevents the occurrence of the biochemical events associated with later stages.

Most of the operons so far studied in detail appear to contain between 1 and 3 genes, *spoVA* with 5 genes (see below) being an exception. Assuming about 70 operons containing on average 2 genes each, the total number of genes governing spore formation might therefore be about 150.

We have already referred to the common observation that when sporulation is blocked at a particular stage the 'biochemical markers' associated with earlier stages occur normally while later ones do not. This leads to the simplest of models, the linear dependent sequence, which assumes that operons are expressed in a linear sequence in which the transcription of any sporulation operon leads to the synthesis of a protein that acts to 'switch on' the next [6]. Other models e.g., parallel or branched pathways have also been considered [5].

Losick and his colleages suggested in a number of papers [7–11] that the sporulation sequence might be controlled by successive changes in the sigma factors of RNA polymerase and it has, in fact, been shown that one of these factors, σ^{29}, plays an essential part in regulating spore formation [8]. However, although σ-factors are clearly involved at an early stage, the ordered expression of 50 or more operons over a period of several hours is likely to involve additional, and probably complex, control elements [11].

Recently, Stragier et al. [12] sequenced the locus *spoIIG*, which is concerned with an early stage of sporulation. Comparison of the predicted amino acid sequence for this gene with that of other proteins revealed a striking homology with the amino acid sequence of the main sigma factor of *Escherichia coli*, the product of the *rpoD* gene [13]. This in turn showed homology with another sigma-like factor of *E. coli*, *htpR*, which is not found under ordinary vegetative conditions but which is induced by subjecting the cells to heat shock [14]. It has since been shown that *spoIIG*, in fact, codes for σ^{29} [15].

Eleven genes specifically associated with sporulation have now been cloned and sequenced in this laboratory [5 genes in *spoVA* [16], 3 in *spoIIA* [17], and one each in *spoIID* (S. Clarke, personal communication), *spoIIIB* (M. Deadman, personal communication), and *spoVE* (U. Bugaichuk, personal communication)]. To these we added the sequence of the *spoIIG* gene [12] and then sought evidence for gene duplication among them by comparing the sequences using a computerised analysis.

2. RESULTS AND DISCUSSION

We found that the 12 sequences listed contained the 2 gene duplications shown in figs 1 and 2. The *spoIIA* operon, according to its nucleotide sequence, consists of 3 genes, A, B and C, and according to the size of its mRNA transcript, it appears that they generate a single polycistronic message [18]. From earlier work it is known that this operon has a regulatory function in sporulation. Thus, mutations which are quite close together in gene A produce a variety of phenotypic effects on sporulation and sporulation-associated biochemical marker events. The same assortment of pleiotropic effects is associated with very closely linked mutations in gene C of the same operon [19]. In addition, there is interaction (the nature of which is unknown) between the products of genes A and C because if a mutation in gene A, producing partial abolition of spore formation or enzyme production, is transferred into a strain with a mutation in gene C causing a similar phenotype, the double mutant exhibits total loss of both spore formation and the associated marker events. It is also known from earlier work that mutations in *spoIIA* could affect not only the degree of oligosporogeny but also the rate at which sporulation occurs [20].

We now find that the amino acid sequence of gene C exhibits a striking homology with σ^{29} (coded for by *spoIIG*) of *B. subtilis* and also with the sigma factor of *E. coli* (*rpoD*) and the heat-shock regulator protein (*htpR*) (fig.1). (Gene A, in spite of its similar function, bears no resemblance to either.)

It thus appears that stage II in sporulation involves the expression of at least 2 proteins, encoded by the *spoIIA* and *spoIIG* loci, with a significant degree of protein homology. Furthermore, examination by the Northern blot technique of the production of mRNA corresponding to *spoIIA* in different genetic backgrounds shows that the operon continues to be expressed in mutants damaged in any of the 6–7 operons, in-

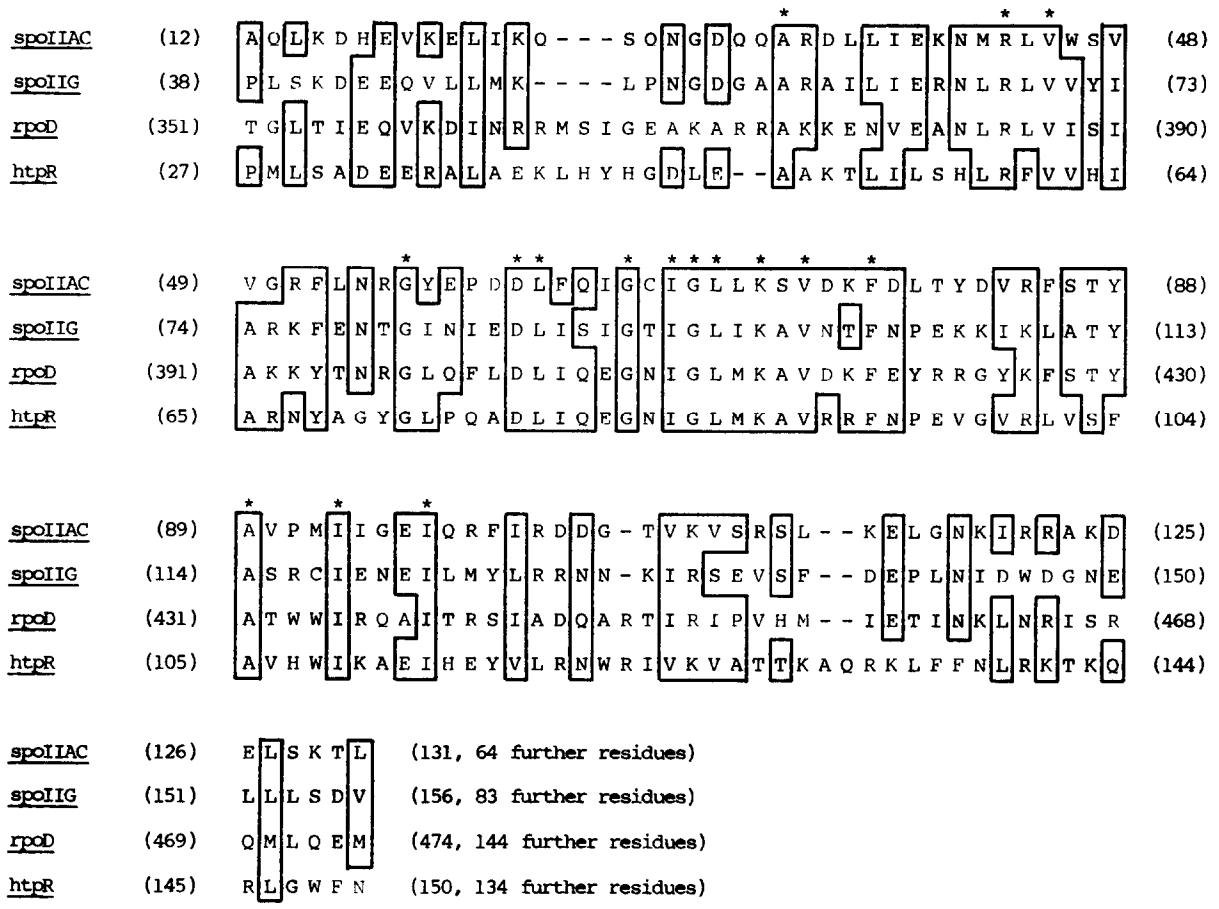

Fig.1. Alignment of the predicted amino acid sequences for *B. subtilis* sporulation genes *spoIIAA* and *spoIIG* and *E. coli* vegetative sigma factor (*rpoD*) and heat-shock regulatory protein (*htpR*). Positions in which all 4 proteins match exactly are indicated by asterisks. Positions in which at least 3 out of 4 residues are identical or show conservative changes are boxed. Conservative changes are defined as those taking place within one of the following groups of amino acids: (A, G, P, S, T); (D, E, N, Q); (F, W, Y); (H, K, R); (I, L, M, V) [26]. The numbers in parentheses at the beginning and end of each line refer to the numbers of the amino acid residues [12,13,14,17]. Gaps, shown as hyphens, were introduced as indicated by treatment of the data using the programme 'DIAGON' [27].

cluding *spoIIG*, known to control stage II [18]. It is apparent that transcription of *spoIIA* does not depend on that of *spoIIG*, and indeed, almost certainly precedes it.

SpoVA is the second sporulation operon that we have cloned and sequenced [16]. From the results it appears that it codes for 5 genes and a comparison of the predicted polypeptides (denoted A–E) showed that 2 of them, C and E, containing 150 and 323 amino acid residues, respectively, have 2 regions of homology, the first near the N-terminal extending over 17 residues and the second covering 78 residues of which 25 are identical (fig.2).

Eight strains with mutations in the *spoVA* locus have been examined [21] and the phenotypes of these, unlike those of the *spoIIA* mutants, appear to be identical. Other studies [16,18] show that the 5 genes of the *spoVA* locus generate a single polycistronic mRNA. We thus have the somewhat curious finding that 2 genes in the same operon are the result of a gene duplication.

In addition to the sporulation genes we have discussed, it has very recently been shown that at

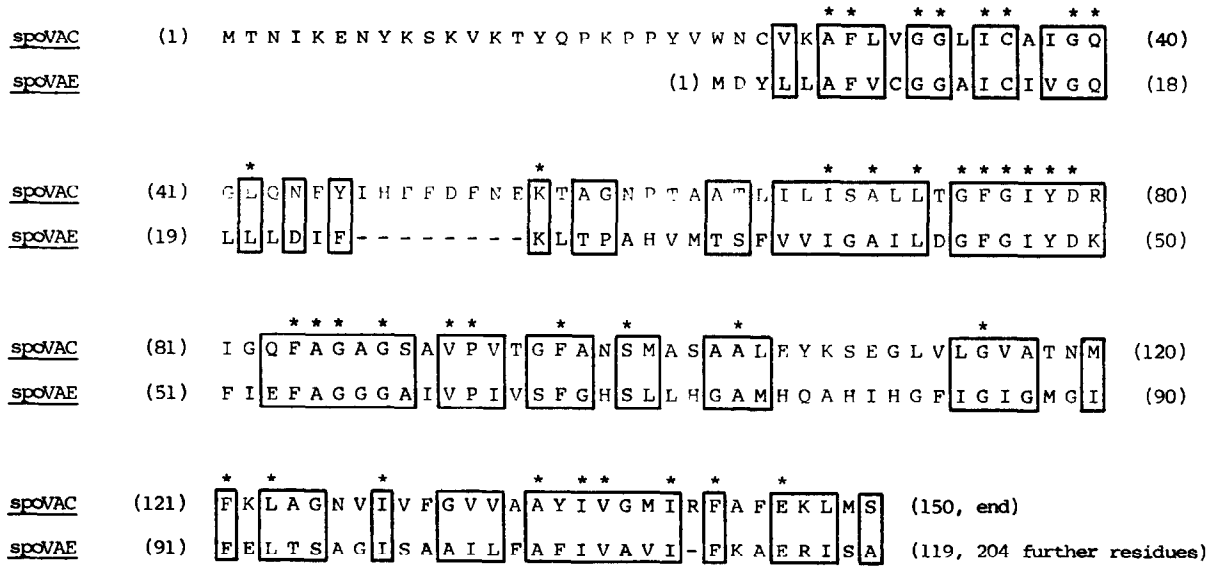

Fig.2. Alignment of the predicted amino acid sequences for *B. subtilis* sporulation genes *spoVAC* and *spoVAE* [16]. Positions of exact match or conservative changes (see fig.1) are boxed: exact matches are further indicated by an asterisk. For further details, see fig.1 legend.

least 4 acid soluble proteins of the spore in *B. subtilis* are closely related to one another [2] and the corresponding genes may be presumed to have arisen by gene duplication. These facts suggest that as more sporulation genes are sequenced and further comparisons are made it is very likely that more examples of duplication will be found. Although it would be premature to speculate on the number, it is nevertheless probable that sporulation will be shown to have arisen as a result of duplication among a substantially smaller number of genes than the 150 postulated earlier in this paper.

This number can first be considered in relation to the numbers of genes needed for ribosome assembly. Three genes are required for 3 species of RNA and about 52 more for their associated proteins [22]. (For purpose of this argument we can ignore the fact that the chromosome may have several copies of any particular gene.) To this should be added a few genes for enzymes that might be needed for processing, bringing the total number to about 60. It does not appear probable that bacterial sporulation could be carried out with a smaller number of genes than this.

For phages, the numbers show considerable variability. In smaller phages, for instance ϕX174, the genome contains information for only about 10 genes which overlap [23]. In lambda phage, 46 genes have been identified and these account for the majority of the DNA [24]. In the larger phage, T4, the number of genes is about 105 and the genome does not have the capacity for many more [25].

In the light of these numbers it is conceivable, and indeed likely, that bacterial sporulation will turn out to be no more complex a system than the assembly of the larger bacteriophages and the conceptual gap between understanding the molecular biology of both viruses and simple developmental systems may be smaller than was originally thought.

ACKNOWLEDGEMENTS

We thank Dr J. Gagnon for assistance with the computer analysis. This work was supported by the Science and Engineering Research Council and by ELF Aquitaine UK.

REFERENCES

[1] Hranueli, D., Piggot, P.J. and Mandelstam, J. (1974) J. Bacteriol. 119, 684–690.

[2] Connors, M.J. and Setlow, P. (1985) J. Bacteriol. 161, 333–339.
[3] Jenkinson, H.F., Sawyer, W.D. and Mandelstam, J. (1981) J. Gen. Microbiol. 123, 1–16.
[4] Ryter, A. (1965) Ann. Inst. Pasteur, Paris 108, 40–60.
[5] Piggot, P.J. and Coote, J.G. (1976) Bacteriol. Rev. 40, 908–962.
[6] Mandelstam, J. (1976) Proc. R. Soc. London, Ser. B 193, 89–106.
[7] Haldenwang, W.G. and Losick, R. (1979) Nature 282, 256–260.
[8] Haldenwang, W.G., Lang, N. and Losick, R. (1981) Cell 23, 615–624.
[9] Losick, R. and Pero, J. (1981) Cell 25, 582–584.
[10] Johnson, W.C., Moran, C.P. jr and Losick, R. (1983) Nature 302, 800–804.
[11] Losick, R. and Youngman, P. (1984) in: Microbial Development (Losick, R. and Shapiro, L. eds) pp.63–88, Cold Spring Harbor Laboratory, New York.
[12] Stragier, P., Bouvier, J., Bonamy, C. and Szulmajster, J. (1984) Nature 312, 376–378.
[13] Burton, Z., Burgess, R.R., Lin, J., Moore, D., Holder, S. and Gross, C.A. (1981) Nucleic Acids Res. 9, 2889–2903.
[14] Landick, R., Vaughan, V., Lau, E.T., Van Bogelen, R.A., Erickson, J.W. and Neidhardt, F.C. (1984) Cell 38, 175–182.
[15] Trempy, J.E., Bonamy, C., Szulmajster, J. and Haldenwang, W.G. (1985) Proc. Natl. Acad. Sci. USA, in press.
[16] Fort, P. and Errington, J. (1985) J. Gen. Microbiol. 131, 1091–1105.
[17] Fort, P. and Piggot, P.J. (1984) J. Gen. Microbiol. 130, 2147–2153.
[18] Savva, D. and Mandelstam, J. (1985) in: The Molecular Biology of Microbial Differentiation (Setlow, P. and Hoch, J. eds) in press, American Society for Microbiology, Washington, DC.
[19] Errington, J. and Mandelstam, J. (1983) J. Gen. Microbiol. 129, 2091–2101.
[20] Yudkin, M.D. and Turley, L. (1980) J. Gen. Microbiol. 121, 69–78.
[21] Errington, J. and Mandelstam, J. (1984) J. Gen. Microbiol. 130, 2115–2121.
[22] Nomura, M., Gourse, R. and Baughman, G. (1984) Annu. Rev. Biochem. 53, 75–117.
[23] Sanger, F., Coulson, A.R., Friedmann, T., Air, G.M., Barrell, B.G., Brown, N.L., Fiddes, J.C., Hutchison, C.A. III, Slocombe, P.M. and Smith, M. (1978) J. Mol. Biol. 125, 225–246.
[24] Sanger, F., Coulson, A.R., Hong, G.F., Hill, D.F. and Petersen, G.B. (1982) J. Mol. Biol. 162, 729–773.
[25] Freifelder, D. (1983) Molecular Biology, pp.613–614, Science Books International, Boston.
[26] Dayhoff, M.O., Schwartz, R.M. and Orcutt, B.C. (1978) in: Atlas of Protein Sequence and Structure (Dayhoff, M.O. ed.) vol.5, suppl.3, pp.345–352, National Biomedical Research Foundation, Washington, DC.
[27] Staden, R. (1982) Nucleic Acids Res. 10, 2951–2961.

NOTE ADDED IN PROOF, 8 August 1985.

We have found two errors in the nucleotide sequence of the *spoIIAC* gene published by Fort and Piggot (1984): an extra 'T' after residue 1722 and an extra 'A' after residue 1880. As kindly pointed out by dr P. Stragier, the insertion of two extra bases in the published sequence increases the predicted protein product from 195 to 255 amino acid residues with a corresponding extension of the homology to other sigma factors at the C-terminal end (see Stragier et al. (1985) FEBS Lett. 187, 11).