The Possible DNA-binding Nature of the Regulatory Proteins, Encoded by spolIID and gerE, Involved in the Sporulation of Bacillus subtilis

By SUSAN K. HOLLAND,1 SIMON CUTTING2 AND JOEL MANDELSTAM3∗

1 Laboratory of Molecular Biophysics, University of Oxford, South Parks Road, Oxford OX1 3QU, UK
2 Microbiology Unit, Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, UK

(Received 17 December 1986; revised 25 May 1987)

The predicted polypeptide products of two genes, spolIID and gerE, which appear to be concerned in the regulation of spore formation in Bacillus subtilis have been compared by modelling methods with known DNA-binding proteins. The results indicate that both polypeptides may have DNA-binding properties and the conclusion is drawn that this may account for their regulatory action.

INTRODUCTION

The formation of spores in Bacillus is the simplest example of cellular differentiation that has been extensively studied and the process has been divided into six stages on the basis of the appearance of cell sections in the electron microscope (Ryter et al., 1966). It is now known that the first two of these stages (0 and I) have, in fact, nothing to do with the developmental process and can be disregarded as far as this paper is concerned.

For the molecular biologist the questions posed are those that arise in any study of differentiation. (a) How many operons – gene clusters functioning as units – control the whole developmental change? (b) What is the order of expression of these operons? (c) What kinds of molecule are involved in switching on these operons?

The answer to the first question is partly known. About 50 genetic loci, presumed to act as operons, have been identified (see Piggot & Coote, 1976; Piggot & Hoch, 1985). Some of these have been shown by RNA–DNA hybridization to produce mRNA transcripts corresponding to several genes (Savva & Mandelstam, 1986) whereas others apparently code for only one gene each (see below). New sporulation loci are still being discovered, though at a diminishing rate, and the total number has been estimated on a statistical basis at about 60 (Hranueli et al., 1974). It thus appears that over two-thirds of the genetic loci have been identified.

The second question has now also been partly answered. It was evident from early work on sporulation that the expression of operons must be in some way sequential. This is because in a mutant blocked at any particular stage, the morphological events and the biochemical events associated with them occur normally until the block is encountered. The simplest model is the ‘linear dependent sequence’ (see Mandelstam, 1976) in which the turning on of each operon causes the formation of a molecule that regulates the expression of the next operon. However, some experiments designed to test the hypothesis suggested that two or more operons might be activated in parallel (Coote & Mandelstam, 1973). More recent work has shown that the original model was too simple and that, though there is a dependent sequence, it has at least three branches, one of which leads to the expression of some of the sporulation operons in the spore compartment while the other two lie in the mother cell compartment (Turner et al., 1986; Cutting & Mandelstam, 1987). It is already clear, although the data are far from complete, that the number of sequential steps is large (Mandelstam & Errington, 1987).
The final question, the nature of the molecules that control the sporulation operons, remains unsolved. It was proposed mainly by Losick and his colleagues (see e.g. Losick & Sonenshein, 1969; Losick & Pero, 1981) that the successive switching on of operons depended on the production of a succession of different types of sigma factor subunits for RNA polymerase. So far, however, only one sigma factor, $\sigma^{29}$ coded for by spoIIG, has been found to be associated with core polymerase and also necessary for sporulation. A second sigma factor, also with $M_r$ 29000, has been inferred from the sequence of one of the sporulation operons, spoIJA, whose nucleotide sequence (see Fort & Piggot, 1984) showed a high degree of homology to known sigma factors of B. subtilis and Escherichia coli (Errington et al., 1985). Both of these sigma factor genes are concerned with stage II, which is (see above) the first relevant stage of sporulation. However, two of the genes that have been sequenced in this laboratory and that have been shown to have a regulatory function in sporulation, gerE and spoIID, appear to code for typical DNA-binding proteins and this paper describes the observations that point to this conclusion.

METHODS

The predicted amino acid sequences of the Bacillus subtilis gerE and spoIID gene products were searched for similarities with other protein sequences held in the NBRF Database using the SEARCH program (Protein Identification Resource, National Biomedical Foundation, Georgetown University Medical Center, Washington DC, USA) on the Molecular Biophysics VAX computer (Cutting & Mandelstam, 1986; Lopez-Diaz et al., 1986; Dayhoff et al., 1983). The observed similarities were aligned using the ALIGN program (Needleman & Wunsch, 1970; Orcutt et al., 1984). The potential DNA-binding regions were modelled upon a DNA-binding protein, the phage $\lambda$ cro structure (Anderson et al., 1981), using the co-ordinates supplied by the Protein Databank and the Molecular Biophysics Evans and Sutherland PS300 graphics system. The main backbone of the $\lambda$ cro structure and the side chains of the spoIID and gerE sequences corresponding to those of the $\lambda$ cro side sequence were all held fixed. Where possible, all similar atoms were spatially matched and additional atoms were placed with regard to ideal stereochemistry as defined by the FRODO dictionary. Further manipulation was required to relieve unfavourable Van der Waals contacts.

RESULTS AND DISCUSSION

The amino acid sequences of the B. subtilis gerE and spoIID polypeptides predicted from the nucleotide sequences are shown in Fig. 1. The sequences were initially searched for homologies with other proteins by analysis of their sequences with the NBRF Database using the SEARCH program. The database search revealed that there were significant similarities with several regulatory proteins. Detailed inspection of the sequences revealed that both the gerE and spoIID polypeptides contained regions which showed significant homology with other DNA-binding proteins over the characteristic helix-turn-helix region, suggesting that these two proteins may also possess a DNA-binding substructure which interacts with a specific operator sequence. The potential DNA-binding regions of the gerE and spoIID sequences are shown underlined in Fig. 1. The same segments are aligned with several DNA-binding proteins in Fig. 2. We have included in Fig. 2 the sequence data corresponding to another sporulation gene, spoIIC (J. Errington, personal communication), which shows remarkable similarity to trpR.

Characterization of sequence-specific DNA-binding proteins

Regulatory proteins which interact with specific DNA sequences have been widely studied at the tertiary and primary level. The determination of the three-dimensional structures of the E. coli catabolite activator protein (CAP) (McKay et al., 1982), E. coli trp apo-repressor (Schevitz et al., 1985), and phage $\lambda$ cI (Pabo et al., 1982; Pabo & Lewis, 1982) and cro repressors (Anderson et al., 1981) and the phage 434 cI repressor (Anderson et al., 1985) has revealed certain characteristics of DNA-binding proteins; these have enabled specific models to be proposed for the recognition of operator DNA sequences by each of the proteins, and a comparison of primary sequences with other regulatory proteins has suggested similar modes of recognition of specific DNA sequences (Ptashne et al., 1980; Pabo & Sauer, 1984; Ohlendorf & Matthews, 1983; Sauer et al., 1982; Ohlendorf et al., 1983; Steitz et al., 1982).

Characterization of sequence-specific DNA-binding proteins
Role of spoIID and gerE in sporulation

Fig. 1. Predicted primary sequences of the polypeptide products of spoIID and gerE. (a) spoIID; broken underlining indicates a strongly hydrophobic region (see Discussion); the continuous line indicates the proposed DNA-binding region. (b) gerE; the consensus of secondary structure predictions using the methods of Chou & Fasman (1974), Lim (1974) and Gamier et al. (1978) is shown under the sequence. Alpha helices are shown as bold arrows and turns are labelled.

The proteins interact as dimers, with an operator sequence exhibiting two-fold symmetry. Although their overall tertiary structures are different, they all possess a common structure consisting of two consecutive helices separated by a tight turn. The second helix protrudes from the surface of the molecule and fits into the major groove of B-DNA (Warrant & Kim, 1978). The recognition of the operator is by multiple hydrogen bonds or hydrophobic interactions between amino acid side-chains of this helix and the exposed edges of the bases in the major groove (Seeman et al., 1976; Ebright et al., 1984; von Hippel et al., 1986). The phosphate groups of the DNA backbone are also involved, interacting mainly with the first helix of the structural unit, adding to the total binding energy. The DNA-binding sequences of several DNA-binding proteins are shown in Fig. 2 and the conserved characteristics of each position within the 20-residue unit are shown. The most strongly conserved residues flank the tight turn, and the residues at positions 5 and 15, which contribute to the structural integrity of the unit by hydrophobic interactions, maintain the angle between the helices.
The possible DNA-binding region of spoIID

The spoIID segment (residues 292-311) contains similar residues to those observed at the strongly conserved positions 8-9-10 (valine-glycine-methionine) about the tight turn, and also at positions 5 and 15 there are two residues (glycine and alanine) which maintain the angle between the two helices. The numbering of all DNA-binding segments is as given in Fig. 2. The interior facing triplets of the two helices (4-5-8 and 15-18-19) all contain non-polar residues, indicating a conserved hydrophobic environment in the interior of the unit. Comparison of the spoIID sequence with the known DNA-binding segments suggests that residues 11 (serine), 12 (glutamine), 13 (tyrosine), 16 (asparagine), 17 (phenylalanine) and 20 (lysine) are capable of interacting with the nucleotide bases of a DNA operator (Wharton & Ptashne, 1985;
Role of spoIID and gerE in sporulation

Hochschild & Ptashne, 1986). In particular, the residues serine and glutamine are often found at positions 11, 12 and 13. The models of the DNA–protein complexes suggest that residues at these positions form bidentate hydrogen bonds with bases at position +1 and +2 of the operator. By analogy, residues 13 (tyrosine) and 16 (asparagine) may contact positions +4 and +5 of an operator sequence. These are only tentative assignments for specific interactions because the recognition code is complex and the operator sequence for the spoIID protein is unknown. Other non-specific interactions may be formed between residues 2 (arginine), 6 (histidine), 11 (serine) and 16 (asparagine) and the phosphate groups of the backbone of the DNA.

The spoIID segment (residues 292–311) was modelled upon the backbone of the phage λ cro structure over the DNA-binding region, substituting the corresponding side chains of the λ cro for those of spoIID. The superimposed structure is shown in Fig. 3. The λ cro DNA-binding structure is shown in blue and the spoIID structure in orange. Overlap of similar atoms is shown in white. Several of the contacting residues are visible on the exterior of the unit. Examination of the internal packing between the two helices, based on Van der Waals radii, revealed a good fit between interface residues with no room for solvent molecules within the interior of the unit. Notably, the planar ring of residue 4 (phenylalanine) packed in between 15 (alanine) and the aliphatic part of 1 (threonine), provided a good hydrophobic contact.

The possible DNA-binding region of gerE

The gerE segment (47–66) shows a good correspondence with the residues of the known DNA-binding sequences, with similar characteristics in most positions. Of the most strongly conserved residues in the unit, the tight turn appears to be conserved in the gerE sequence at positions 8–9–10 (leucine, glycine, valine) and methionine at position 5. The last of these residues does not appear to be unacceptable, as the trp apo-repressor contains a lysine at this position. However, at position 15 only hydrophobic residues have previously been observed and here the gerE sequence contains a glutamine. To test the feasibility of the gerE polypeptide as a DNA-binding protein the segment was imposed upon the λ cro structure – as was done with the spoIID segment – again with the main backbone held fixed (Fig. 3). Apart from the glutamine side chain the remainder of the gerE sequence fitted well onto the bi-helical unit with no steric hindrance between residues. A study of 200 determined protein structures showed that 7% of all glutamates are found 95% or more buried within a protein (Chothia, 1976). It is therefore possible to accommodate this residue in the interface between the helices, provided that the polar head group can be hydrogen-bonded. Possible hydrogen bonds could be achieved between the carbonyl of the glutamate and the hydroxyl group of residue 14 (serine). Also, the amide of the glutamate could interact with the main chain carbonyl of residue 15 without induced stress in the conformation of the side chains. In the gerE structure this may indicate that the maintenance of the angle between helices is contributed to by these hydrogen bonds in addition to hydrophobic interactions.

From the comparison of the segment with the consensus of DNA-binding characteristics, there are several potential DNA-contacting residues. Those involved in specific base interactions may be 11 (lysine) which by analogy with the proposed complexes may contact position +1 of the unknown operator. Residue 13 (arginine) may contact position +4, with residues 14 (serine) and 17 (valine) in the vicinity of position +5. Residue 19 (glutamic acid) may also contact an exposed base. Non-specific interactions with the phosphate backbone could occur with residues 2 (serine), 3 (asparagine), 6 (glutamine), 7 (lysine), 11 (lysine) and 19 (glutamine).

Further study of the gerE primary sequence reveals that the amino-terminal arm is highly charged like that of the phage λ cI repressor, containing two glutamic acid residues and two lysines. In the λ cI repressor there are three lysines, a serine and a threonine which have been shown to contact the far side of the DNA, interacting with the bases and phosphates and aiding specificity (Pabo et al., 1982; Eliason et al., 1985). It is possible that the gerE protein may also similarly contact DNA, possessing a flexible arm to interact with the bases of the operator sequence.
Secondary structure predictions, based upon the methods of Chou & Fasman (1974), Lim (1974) and Garnier et al. (1978) suggest, with good agreement, that the gerE gene product is an all-helical protein, containing four helices of which the third and fourth (c and d) comprise the DNA-binding unit (see Fig. 1).

The NBRF protein database search also indicated several similarities between the gerE sequence and other regions of DNA-binding proteins. Significant similarities were found by running the ALIGN program for the complete gerE sequence and the whole of the aminoterminal DNA-binding domain of the E. coli lexA repressor (residues 1–58) (Pabo & Sauer, 1984; Brent & Ptashne, 1981; Hurstel et al., 1986) as shown in Fig. 4. This may indicate that the two structures adopt similar conformations.

Other similarities were found for residues 10–31 of the gerE sequence and a region in the carboxyl domain of the E. coli lexA repressor and the C helix of the E. coli trp repressor (Fig. 4) (Schevitz et al., 1985). The phage repressors and lexA repressor exhibit good homology over their carboxyl domains, which are known to contain strong oligomerization sites that create additional stabilizing dimer contacts (see Fig. 5) (Sauer et al., 1982). Similarly helix C of the trp repressor is known to form dimer contacts and the repressor itself contains only x-helices (Schevitz et al., 1985). These similarities between gerE and the DNA-binding proteins suggest not only that the gerE protein contains a DNA-binding unit, but that it also utilizes the rest of its relatively short length to form regions of inter-domain contacts, along one or more helices (a and/or b) (see Fig. 1b). As little of the chain extends beyond the potential DNA-binding region, it would appear likely that any dimer contacts would lie in the amino-terminal portion of the polypeptide. As the polypeptide is quite short it appears unlikely that the protein contains more than one domain. This homology strongly suggests that the gerE and lexA repressor contain similar helices to the helix C of the trp repressor and that the lexA repressor proteins and phage repressors use a similar helical region in the carboxyl domain to form dimer contacts.

The regulatory function of the B. subtilis potential DNA-binding proteins is indeterminable from sequence studies alone and this point will be considered in more detail below.

Negative gene regulation involves the repressor molecule binding at, or near to, the origin of transcription, and thus sterically preventing the access of the RNA polymerase to the initiation site (Maurer et al., 1980; Takeda et al., 1983). There is some evidence for a ‘direct contact’ mechanism in positive gene regulation by λ cII and P22 λ cI (Johnson et al., 1981). This suggests that negatively charged residues clustered on the surface of the activator are in close contact with a positive area on the RNA polymerase. Mutations in P22 cI at 42 (glutamic acid) prevent stimulation of transcription. Comparable residues occur in CAP at 192 (aspartic acid) and 191 (glutamic acid).

The spoIID and gerE sequences contain glutamic acid residues in similar locations to those in the P22 cI and CAP sequences. Glutamic acid residues occur at position 312 in spoIID and at positions 71 and 73 in gerE. Like the CAP protein, the DNA-binding units are towards the C-terminal of the chain.

As described in the Introduction, sporulation operons in general are expressed in a dependent sequence which has been partly elucidated. The stage 0 operons, designated as spo0A, spo0B etc., are now known to be expressed in the vegetative cell and can therefore be excluded in a consideration of the developmental process. They must be presumed to determine basic structures and functions of the cell which do not affect growth but which are a pre-requisite for sporulation although they are not specifically concerned with it. It is known that mutations in stage 0 operons have widely varying pleiotropic effects, for example, on competence of cells in genetic transformation, on susceptibility to infection of the cells by a variety of phages, on the secretion of exo-enzymes, etc.

---

Fig. 3 (on facing page). Structures based on modelling studies of protein products of spoIID and gerE.

(a) The modelled structure of the spoIID DNA-binding region based upon the λ cro structure. Overlap of similar atoms is shown in white; the λ cro structure is in blue, and the spoIID structure in orange.

(b) The modelled structure of the gerE DNA-binding region based upon the λ cro structure. Overlap of similar atoms is shown in white; the λ cro structure is in blue, and the gerE structure orange.
Coming now to the 'real' developmental operons, i.e. those that determine events from stage II onwards, we find that spoIIA, consisting of three genes (Fort & Piggot, 1984), has one which codes for a protein having strong homology with known sigma factors in both B. subtilis and E. coli (Errington et al., 1985). This seems to be one of the earliest operons in the sequence and the proper functioning of all the subsequent operons in the dependent sequence is blocked partly or completely by mutations in it. Another operon, spoIIG, codes for a product which, after proteolytic processing, is converted to the sigma factor $\sigma^{29}$. Whereas the polypeptide coded for by one of the genes in spoIIA has been presumed to code for a sigma factor on the basis of homology, the gene in spoIIG encodes a product that has actually been found in association with RNA polymerase (Losick & Pero, 1981). It is curious that the only two sigma factors that are apparently concerned with sporulation are both involved in the earliest stage, stage II, and are expressed independently of one another (Clarke & Mandelstam, 1987).

To explain the successive expression of a large number of operons we either have to assume that there are a large number of other sigma factors which are non-homologous and whose amino acid sequences have therefore as yet given no clue to their function (see Doi & Wang,
Role of spoIID and gerE in sporulation

1986), or that other types of molecule are involved in the transcription of sporulation operons. It is in this context that the findings reported in this paper acquire significance. Of some 15 sporulation genes sequenced in this laboratory three, spoIID, gerE and spoIIIC, appear to have the requisite properties of DNA-binding proteins. The simplest assumption is that they might function as activators rather than repressors. The first of these, the spoIID product, is of special interest because, at its N-terminal end, 20 residues of 26 are hydrophobic (see Fig. 1), which suggests that the polypeptide might enter or cross membrane structures. In addition, the N-terminal sequence has the characteristics of a signal peptide that is cleaved from proteins that are exported from the cell. Since we know that some sporulation operons are expressed in the spore compartment it follows that a molecule of some type must cross from the mother cell into the forespore where it sets in train a separate branch of the dependent sequence. From its structure it seems that the protein encoded by spoIID might have the requisite properties.

The operon that appears to code for the DNA-binding protein, gerE, is also monocistronic. It constitutes part of the dependent sequence of events expressed in the mother cell, and mutations in it affect some of the properties of the spore that develop at a relatively late stage (S. Cutting & J. Mandelstam, unpublished).

In summary, the predicted protein products of three genes (spoIID, spoIIIC and gerE) that are known to be regulatory share the helix-turn-helix structure at the C-terminal end that characterizes known activators and repressors binding to DNA. We have already emphasized that the structure of a protein cannot be accurately predicted from its sequence, and the probability of error in modelling of the kind we have used is generally estimated at 45% to 50% (say, 47.5%). In fact, when the sequences are treated by an alternative method (J. Wootton, personal communication) the spoIID product does not appear to have the helix-turn-helix motif; gerE, on the other hand, is a potential DNA-binding protein but the relevant domain on this model contains the residues 29 to 48 and is thus in a different part of the molecule to that indicated by us. However, other parts of the gerE product show homology to portions of the trpR repressor outside the DNA-binding motif. The regulatory protein encoded by spoIIIC has not been analysed by either method, but its close homology to trpR makes it very likely that it would score as a DNA-binding protein in both programs.

The probability of mis-classification in all three proteins is only about 0.1 (0.4753). On this basis, we propose as a hypothesis that the helix-turn-helix structure is responsible for the regulatory function of these molecules and that its activity is exerted by DNA-binding. It is in any case apparent that these proteins are distinct from the homologous set of sigma factors described by Doi & Wang (1986) and that they account, at least in part, for the sequential expression of sporulation operons.

We are indebted to Dr J. Errington for the sequence data for spoIIIC, to Dr Colin Blake for advice on the preparation of the manuscript, and to Dr R. Pickersgill for help with the graphics.

This work was done by S. K. H. on an MRC Research Studentship and by S. C. on an SERC-CASE award; it was supported by the SERC.

REFERENCES


Cutting, S. & Mandelstam, J. (1986). The nucleotide sequence and transcription during sporulation of the


Role of spoIID and gerE in sporulation


