Proc. Natl. Acad. Sci. USA
Vol. 80, pp. 6829–6833, November 1983
Biochemistry

Structural similarities between the development-specific protein S from a Gram-negative bacterium, *Myxococcus xanthus*, and calmodulin

(Ca^{2+}-binding protein/myxospore/protein secretion/codon usage)

SUMIKO INOUYE, THOMAS FRANCESCHINI, AND MASAYORI INOUYE

Department of Biochemistry, State University of New York at Stony Brook, Stony Brook, NY 11794

Communicated by Arthur B. Pardee, July 27, 1983

ABSTRACT **During differentiation of *Myxococcus xanthus*, a large amount of protein S is produced and assembled on the surface of the myxospore by a process that specifically requires Ca^{2+}. The gene for protein S has been cloned, and two tandemly repeated homologous genes have been found to be within a short distance of each other in the *M. xanthus* chromosome. We determined the DNA sequence of 3,692 bp encompassing both genes and deduced the amino acid sequences of the two gene products. The gene 1 (upstream) product and the gene 2 (downstream) product show extensive amino acid sequence homology (88%). However, from their structures, protein S was found to be produced from gene 2, indicating that gene 2 is specifically turned on during differentiation. The structure of protein S shows striking similarities with calmodulin: protein S is composed of four internally homologous domains. In particular, the first and the third domains, consisting of 38 residues each, show a high level of homology (79%), and the second and the fourth domains, consisting of 40 residues each, show homology of 65%. In the first and the third domains, there is a common sequence of nine residues, Glu (or Asp)-Asn-Asn-Thr-Ile-Ser-Ser-Val-Lys, which is highly homologous to one of the proposed Ca^{2+}-binding sequences in bovine brain calmodulin, Asp-Gly-Asn-Gly-Thr-Ile-Thr-Thr-Lys.**

A Gram-negative bacterium, *Myxococcus xanthus* moves by gliding on a solid surface and undergoes a unique developmental cycle, forming fruiting bodies upon starvation (for review, see ref. 1). During fruiting-body formation, production of a development-specific protein called protein S is induced, and it accumulates in a large amount (2, 3). Protein S assembles on the surface of myxospores in the presence of Ca^{2+} (3). Protein S has been purified and crystallized (4) and a preliminary x-ray crystallographic study also has been carried out (5). Recently, its gene was cloned, and it has been shown that, in the *M. xanthus* chromosome, the two genes for protein S are tandemly repeated in the same direction within a short distance (6). We determined the DNA sequence of the entire region covering both genes, enabling us to deduce the amino acid sequences of the products of these genes. The primary structure of protein S and its other properties show several common features with eukaryotic calmodulin.

MATERIALS AND METHODS

Plasmids. Plasmids pSI001, pSI002, and pSI003 (6) were used for DNA sequence determinations. The 3.0-kilobase (kb) *Bam*HI fragment harboring gene 1 (6) also was cloned into pBR322 in order to determine the DNA sequence of the region upstream of gene 1, and the new clone was designated pSI004.

DNA Sequence Determination. The DNA sequence was determined by the method of Maxam and Gilbert (7).

Other Materials and Methods. Restriction enzymes were obtained from either Bethesda Research Laboratories or New England BioLabs.

Restriction fragments were labeled at their 3′ end by the method of Sakano *et al.* (8) with [α-P^{32}]dNTPs (2,000–3,000 Ci/mmol, Amersham; 1 Ci = 37 GBq) and DNA polymerase Klenow fragment (New England Nuclear).

RESULTS AND DISCUSSION

DNA Sequence of Gene 1 and Gene 2. The genes for protein S have been identified and cloned with the use of mixed probes consisting of eight synthetic oligodeoxyribonucleotides (6). It was found that the two homologous genes for protein S (gene 1 and gene 2) are tandemly repeated in the same direction within a short distance on the *M. xanthus* chromosome. In order to determine the entire DNA sequence of the region encompassing both gene 1 and gene 2, the following four clones were used: pSI001, carrying the 9.7-kb *Hin*dIII fragment, which harbors both gene 1 and gene 2; pSI002, carrying the 1.0-kb *Hin*dIII–*Bam*HI fragment, which harbors only gene 1; pSI003, carrying the 2.2-kb *Bam*HI–*Bam*HI fragment, which harbors only gene 2 (6); and pSI004, carrying the 3.0-kb *Bam*HI fragment, which harbors gene 1. The sequencing strategy is summarized in Fig. 1. Thus, the DNA sequence of 3,692 bp encompassing both gene 1 and gene 2 was determined and is shown in Fig. 2. In this sequence, we found an extra *Bam*HI site (residues 1,368–1,373) that is only 40 bp from the first *Bam*HI site (residues 1,408–1,413; see Fig. 1). The small *Bam*HI fragment of 40 bp could not be detected in the agarose gel system used previously (6).

The DNA sequences (A-T-G-A-A-C-A-A-C-A-A-C-A-C) that hybridized with the synthetic oligonucleotide probes are from residue 970 to 983 for gene 1 and from residue 2,879 to 2,892 for gene 2. Because these DNA sequences correspond to those that would code for a part of the carboxyl-terminal end of protein S, Met-Asn-Asn-Asn-Thr, the amino acid sequences of the gene 1 and the gene 2 products were deduced from the DNA sequences by extending the reading frames in both the amino- and the carboxyl-terminal directions. As shown in Fig. 2, the gene 1 and the gene 2 products thus were assigned as a protein of 175 amino acid residues (M_r, 19,235) and a protein of 173 amino acid residues, (M_r, 18,792), respectively.

Assignment of Gene 1 and Gene 2 Products. The amino-terminal structure of protein S purified from differentiating cells was determined by sequential Edman degradation to be Ala-[1]

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: kb, kilobase(s).

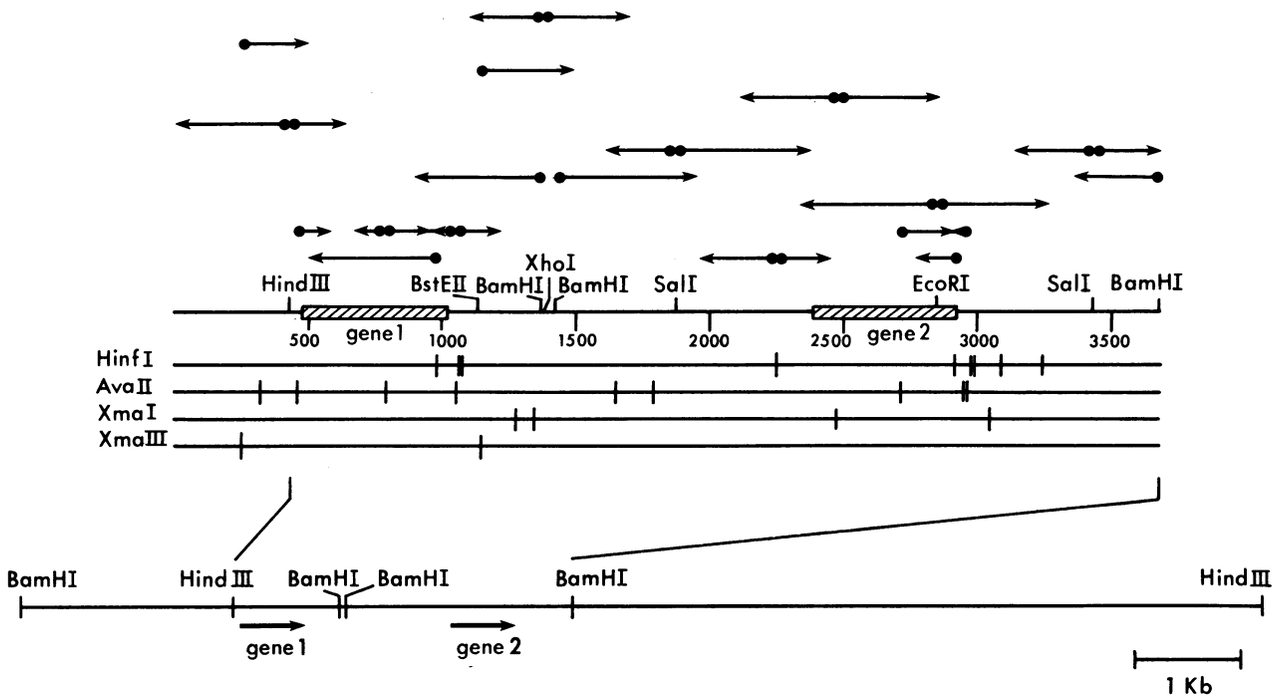

FIG. 1. Restriction enzyme map and sequence determination strategy. ●, 3′-end-labeled positions; the solid lines with arrowheads indicate the regions whose sequence was determined.

Asn-Ile-Thr-Val-Phe-Tyr-Asn-Glu-Asp^{10}-Phe-Gln-Gly-Lys-Gln-Val-Asp-Leu-Pro-X^{20}-Gly-X-Tyr-Thr-X-Ala-Gln-Leu-Ala^{29} (unpublished data). This result clearly indicates that the major protein S product during development is derived from gene 2 and that translation of gene 2 is initiated from the methionine residue immediately before the amino-terminal alanine (Fig. 2). The conclusion that protein S is the gene 2 product is supported also by the carboxyl-terminal structure of protein S, which has been determined to be either isoleucine or serine (4), while the carboxyl-terminal end of the gene 1 protein is proline (Fig. 2).

Structural Similarities with Calmodulin. The primary structure of the gene 2 product (protein S) appears to be composed of four unit structures or domains, which share homologous structures. In particular, there are extensive homologies between the first domain (residues 6–43; domain 1) and the third domain (residues 95–132; domain 3) and between the second domain (residues 45–84; domain 2) and the fourth domain (residues 133–172; domain 4) (Fig. 3). Domains 1 and 3 each consist of 38 amino acid residues, 26 of which are identical and 4 of which are functionally identical (79% homology; see Fig. 3). Domains 2 and 4 each consist of 40 amino acid residues, 19 of which are identical and 7 of which are functionally identical (65% homology; see Fig. 3). At the level of the DNA sequence, 88 of 114 bases are homologous between domains 1 and 3 (77% homology). Among 38 codons in these domains, 13 codons have a single-base substitution (7 of them are silent mutations), 5 have two mismatches, and 1 has three mismatches. Between domains 2 and 4, 78 of 120 bases are homologous (65% homology). Among 40 codons in these domains, 12 codons have a single-base substitution (only two of them are silent mutations), 9 have two mismatches [1 of them is a silent mutation, CGG (Arg^{83}) and AGA (Arg^{171})], and 4 have three mismatches.

Similar four-domain structures as those described above have been found for calmodulin from various eukaryotic sources from mammalians to invertebrates, plants, protozoa, and slime molds (for reviews, see refs. 9 and 10). Calmodulin is one of the most highly conserved proteins during evolution and is composed of four domains consisting of ≈40 amino acid residues, each of which has a Ca^{2+}-binding site. It is of great interest to find that the sequence of 9 amino acid residues in this region of domains 1 and 3, Glu (or Asp)-Asn-Asn-Thr-Ile-Ser-Ser-Val-Lys, shows a high level of homology to one of the proposed Ca^{2+}-binding sequences in bovine brain calmodulin, Asp*-Gly-Asn*-Gly-Thr*-Ile-Thr*-Thr-Lys (11) (underlines indicate residues identical or functionally identical to the protein S sequence, and asterisks indicate putative Ca^{2+}-binding residues).

There are other similarities that exist between calmodulin and protein S. (*i*) Calmodulin is an acidic protein with a pI of about 4.2, whereas the pI of protein S has been determined to be 4.5 (4). (*ii*) Both calmodulin and protein S are heat-stable proteins (4). (*iii*) The contents of hydrophobic amino acids for gene 1 and gene 2 products are 37% and 39%, respectively, in comparison with 37% for bovine brain calmodulin (11) and *Tetrahymena* calmodulin (12). (*iv*) There are no cysteine residues. However, both gene 1 and gene 2 products have much higher proline contents (10 and 12 residues, respectively) in contrast to only two residues in bovine brain and *Tetrahymena* calmodulins. Furthermore, trimethyllysine was not detected in protein S (unpublished results). Calmodulins from *Dictyostelium discoideum* and from barley shoots and fungi (13) have been reported to have no trimethyllysine residue (14).

It has been shown that Ca^{2+} is specifically required for the assembly of protein S on the surface of the myxospores and that EDTA can solubilize protein S from the myxospores (3). These facts clearly indicate that protein S is a Ca^{2+}-binding protein that can associate with a specific site on the myxospores in the presence of Ca^{2+} (1 mM) to become an insoluble form. These properties of protein S are similar to those of calmodulin, which interacts with other proteins by changing its conformation in the presence of Ca^{2+}. However, calmodulin appears to have a much higher affinity for Ca^{2+} (K_d around 10^{-5}–10^{-6} M; for reviews, see refs. 9 and 10) than does protein S. From the com-

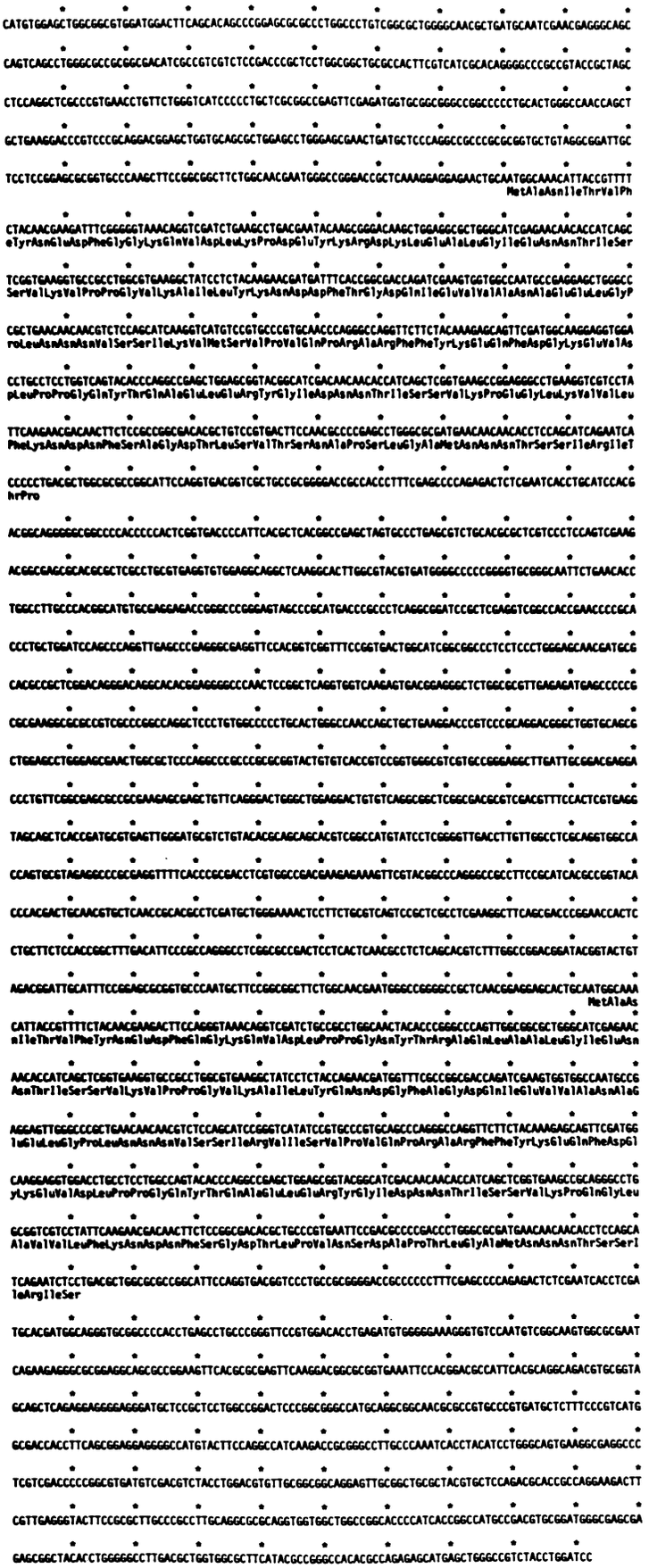

FIG. 2. DNA sequence of the fragment encompassing gene 1 and gene 2. The DNA sequence of the 3,692-bp fragment is shown. The first 160-bp sequence was determined only in one direction, leaving some ambiguity at the regions of very high G+C content (between residues 118 and 122 and between 152 and 156).

puter analysis of the amino acid sequence of protein S (data not shown) the putative Ca^{2+}-binding sites of protein S appear to be sandwiched between α-helical structures similar to that of the EF-hand structure on parvalbumin (15).

In prokaryotic cells, it has been reported that *Escherichia coli* (13) and *Bordetella pertussis* (16) have no calmodulin.

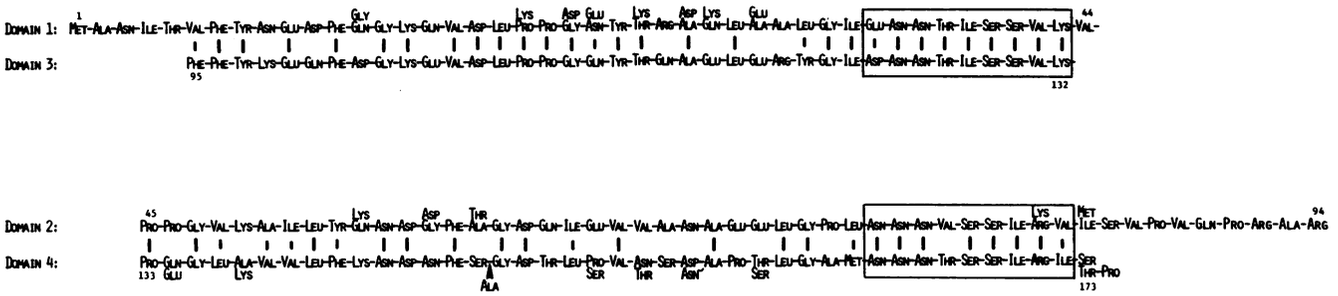

FIG. 3. Four-domain structure of protein S. Putative Ca^{2+}-binding sites are boxed. Numbers indicate the residue numbers from the amino-terminal methionine residue. Only those amino acid residues that are substituted in the gene 1 product are shown above or below the sequence of protein S. The alanine residue with an arrow in domain 4 and the carboxyl-terminal proline residue are extra amino acid residues found in the gene 1 product.

However, the existence of a calmodulin-like protein in *E. coli* has been demonstrated (17). At present, it is not known whether protein S or the gene 1 product or both have calmodulin-like functions.

Tryptic Core Structure. It is interesting to note that domains 1 and 2 and domains 3 and 4 of protein S are separated by two proline residues and one proline residue, respectively, and that between domains 2 and 3 there is a sequence of 10 amino acids, including two proline residues. There are also two proline residues in the middle of domains 1 and 3. These proline residues are possibly playing an important role in the conformation of protein S. In particular, the sequence of 10 amino acid residues between domains 2 and 3 may serve as a hinge in the center of the protein S molecule. In this regard, it can be concluded that the tryptic core peptide of molecular weight 10,000 derived from protein S, which was still functional for assembly (4), resulted from the tryptic cleavage between Arg^{92} and Ala^{93} at the putative hinge region. This conclusion is based on the following data: (*i*) the carboxyl-terminal structure of the core peptide has been determined to be Pro-Arg, and (*ii*) the amino acid composition of the core peptide is almost identical to that of the sequence from amino acid residue 2 to 92. These results indicate that the putative hinge region from residue 85 to 94 is more exposed to tryptic digestion than are domains 1 and 2 and that the amino-terminal half of protein S, consisting of domains 1 and 2, is able to serve as a functional unit for assembly on the surfaces of myxospores.

Homology Between Gene 1 and Gene 2. It is not certain at present whether gene 1 is expressed. However, the putative primary structure of the gene 1 product is highly homologous to that of protein S, the gene 2 product. There are 20 amino acid substitutions and 2 amino acid insertions in the gene 1 product (see Fig. 3). Thus, there are 88% and 93% homologies at the level of the amino acid sequences and the DNA sequences, respectively, between gene 1 and gene 2. It is interesting to note that the gene 1 product has 8 amino acid substitutions in domain 1, whereas there is no amino acid substitution in domain 3, resulting in 68% homology between domains 1 and 3 in contrast to 79% in the case of the gene 2 product. On the other hand, amino acid substitutions in domains 2 and 4 of the gene 1 product made homology between the two domains increase from 65% (gene 2) to 73% (gene 1). It should be noted that there is only one amino acid substitution in the putative Ca^{2+}-binding sites as shown in the boxed sequences in Fig. 3. At the level of DNA, there are 31 base substitutions and 6 base insertions in gene 1. Among 31 base substitutions, 5 are silent mutations and the remaining 26 result in amino acid substitutions (14 single- and 6 double-base substitutions in the codons).

Codon Usage. Codon usage in gene 1 and gene 2 is shown in Table 1. As expected from the high G+C content of the *M. xanthus* chromosomal DNA (68%; ref. 18), 86% and 87% of codons used in gene 1 and gene 2, respectively, have G or C at the third positions. As a result, the G+C contents of the coding regions of gene 1 and gene 2 are 57% and 60%, respectively. It is also interesting to note that 21 codons (gene 1) and 22 codons (gene 2) are not used at all.

Noncoding Regions. Besides the coding regions, there are extensive homologies in the 5′-end and the 3′-end noncoding regions between gene 1 and gene 2. At the 5′-end regions, the sequence of 100 bp immediately before the initiation codon of gene 1 (from residue 381 to 480; see Fig. 2) has 92% homology with the sequence of 101 bp immediately before the initiation codon of gene 2 (from residue 2,292 to 2,392; see Fig. 2). Similarly, at the 3′-end regions, the sequence of 114 bp immediately after the termination codon of gene 1 has 94% homology with the sequence of 114 bp immediately after the termination codon of gene 2. Beyond these regions, homologies abruptly disappear.

From the homology with gene 2, the initiation codon for gene 1 was assumed to be at residues 481–483. However, gene 1 can be substantially extended beyond the initiation codon without hitting a termination codon within the same reading frame. It is interesting to note, however, that the G+C content along the DNA sequence determined in Fig. 2 is distinctively low only in the assumed coding region of gene 1 (57%) as shown in Fig. 4. At the 5′-end side of the initiation codon, the G+C content abruptly increased to 71% (from residue 1 to 480). Immediately

Table 1. Codon usage in gene 1 and gene 2

First nucleotide	Second nucleotide								Third nucleotide
	T		C		A		G		
	a/b		a/b		a/b		a/b		
T	Phe	0/0	Ser	0/0	Tyr	0/0	Cys	0/0	T
	Phe	8/8	Ser	6/6	Tyr	6/6	Cys	0/0	C
	Leu	0/0	Ser	0/0	Term	0/0	Term	1/1	A
	Leu	0/2	Ser	2/2	Term	0/0	Trp	0/0	G
C	Leu	0/0	Pro	4/4	His	0/0	Arg	0/0	T
	Leu	1/1	Pro	3/3	His	0/0	Arg	0/0	C
	Leu	1/1	Pro	0/0	Gln	1/0	Arg	0/0	A
	Leu	10/8	Pro	4/5	Gln	5/10	Arg	2/3	G
A	Ile	1/1	Thr	1/0	Asn	1/2	Ser	0/0	T
	Ile	9/9	Thr	7/7	Asn	16/16	Ser	5/4	C
	Ile	0/1	Thr	0/0	Lys	2/2	Arg	1/1	A
	Met	3/2	Thr	1/1	Lys	11/5	Arg	2/2	G
G	Val	1/1	Ala	1/1	Asp	5/3	Gly	2/2	T
	Val	5/5	Ala	6/7	Asp	7/7	Gly	9/11	C
	Val	0/0	Ala	1/1	Glu	3/2	Gly	0/0	A
	Val	10/10	Ala	2/4	Glu	9/7	Gly	1/0	G

a, Gene 1; b, gene 2.

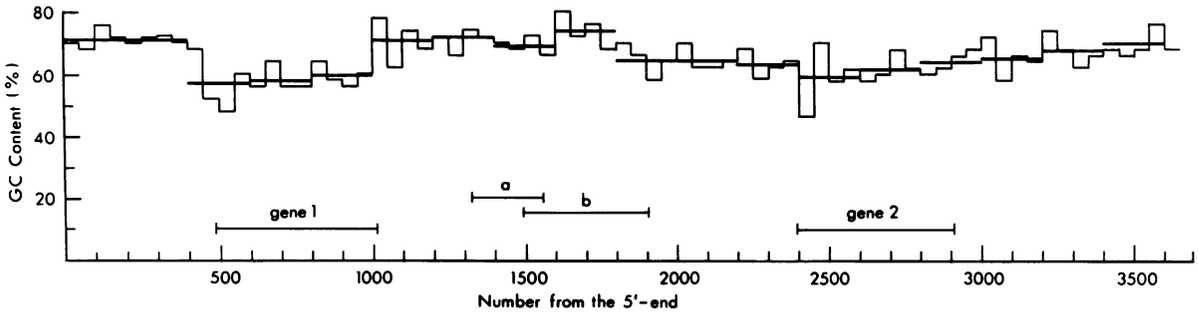

FIG. 4. Variation of the G+C content in the DNA sequence determined in Fig. 2. The G+C contents were calculated for every 50 nucleotides from the 5'-end of the DNA fragment shown in Fig. 2 and plotted along the DNA sequence. Numbers of the DNA sequence correspond to those in Fig. 2. Bars indicate the average G+C contents for every 200 nucleotides. The coding regions for gene 1 and gene 2 are shown by bars, and two regions with open reading frames between gene 1 and gene 2 are also shown by bars with letters a and b (see the text).

after the termination codon of gene 1, the G+C content also suddenly increases to 70%, and this high G+C content is maintained in a long sequence of a little more than 1 kb (from residue 1,009 to 2,050) (see Fig. 4). The G+C content then again drops to 63% in the region of ≈350 bp immediately upstream of the initiation codon of gene 2 (from residue 2,051 to 2,392). The G+C content of the gene 2 coding region is 60% as described previously, and the G+C content again increases to 67% after the termination codon of gene 2 (see Fig. 4). In Fig. 4, the G+C contents are plotted for every 50 and 200 bases, and one can observe clearly the unusual distribution of the G+C contents as discussed above.

The lower G+C content in the 350-bp region immediately before the gene 2 initiation codon may reflect the existence of the promoter for gene 2 in this region because stronger promoters appear to have higher A+T contents in *E. coli* (19). At present it is not possible to identify the promoters for both gene 1 and gene 2. However, in a preliminary experiment, a mRNA of ≈600 bases was detected only during development, which hybridized with a gene 2-specific probe but not with a gene 1-specific probe (unpublished data), indicating that the promoter for gene 2 resides between gene 1 and gene 2 and that the transcription of gene 2 starts within the 100-bp 5'-end noncoding region. At present, it is not known whether gene 1 is expressed. It is possible that gene 1 may be expressed only during vegetative growth or during a very specific period of development.

OTHER FEATURES

It should be pointed out that, in the spacer region between gene 1 and gene 2 (1,383 bp), there are two relatively long sequences of an open reading frame—one protein of 82 amino acid residues (a in Fig. 4; from residue 1,317 to 1,562) and another protein of 135 amino acid residues (b in Fig. 4; from residue 1,496 to 1,900). There are also three relatively long sequences of an open reading frame in the opposite direction in the space region—one protein of 65 amino acid residues from residue 2,224 (see Fig. 2), another protein of 111 amino acid residues from residue 2,088 (see Fig. 2), and the third protein of 103 amino acid residues from residue 1,468 (see Fig. 2).

It is a great surprise that two extremely homologous genes, gene 1 and gene 2, are stably maintained in the *M. xanthus* chromosome even if they are very closely linked to each other by being separated only by 1,383 bp. This spacer region between the two genes may be playing an important role for preventing one of the genes from deletion.

We examined whether there is another region in the *M. xanthus* chromosome that has homology with the gene for protein S by using a DNA fragment covering almost the entire gene 1 (residues 420–995; see Fig. 2) as a hybridization probe. We were not able to detect any other DNA fragments besides the DNA fragments carrying gene 1 and gene 2 that can hybridize with the probe. Therefore, it was concluded that there is no other gene homologous to the structural gene for protein S except for gene 1.

One of the surprising findings is that protein S is not produced from a secretory precursor with a signal peptide. Because protein S is found at the outermost layer of the myxospores, protein S is considered to be secreted across the cell envelope. Protein S may have the secretory signal inside the protein structure or may be translocated through a specific channel by an unknown mechanism.

The authors are grateful to Dr. R. Sarma and Dr. M. Teintze for their critical reading of this manuscript. This work was supported by United States Public Health Service Grant GM-26843 from the National Institutes of Health.

1. Kaiser, D., Manoil, C. & Dworkin, M. (1979) *Annu. Rev. Microbiol.* **33**, 595–639.
2. Inouye, M., Inouye, S. & Zusman, D. R. (1979) *Dev. Biol.* **68**, 579–591.
3. Inouye, M., Inouye, S. & Zusman, D. R. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 209–213.
4. Inouye, S., Harada, W., Zusman, D. R. & Inouye, M. (1981) *J. Bacteriol.* **148**, 678–683.
5. Inouye, S., Inouye, M., McKeever, B. & Sarma, R. (1980) *J. Biol. Chem.* **255**, 3713–3714.
6. Inouye, S., Ike, Y. & Inouye, M. (1983) *J. Biol. Chem.* **258**, 38–40.
7. Maxam, A. & Gilbert, W. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 560–564.
8. Sakano, H., Huppi, K., Heinrich, G. & Tonegawa, S. (1979) *Nature (London)* **280**, 288–294.
9. Lin, Y. M. (1982) *Mol. Cell. Biochem.* **45**, 101–112.
10. Wang, J. H. & Waisman, D. M. (1979) *Curr. Top. Cell. Regul.* **15**, 45–107.
11. Watterson, D. M., Sharief, F. & Vanaman, T. C. (1980) *J. Biol. Chem.* **255**, 962–975.
12. Yazawa, M., Yagi, K., Toda, H., Kondo, K., Narita, K., Yamazaki, R., Sobue, K., Kakiuchi, S., Nagao, S. & Nozawa, Y. (1981) *Biochem. Biophys. Res. Commun.* **99**, 1051–1075.
13. Grand, R. J. A., Nairn, A. C. & Perry, S. V. (1980) *Biochem. J.* **185**, 755–760.
14. Bazari, W. L. & Clarke, M. (1981) *J. Biol. Chem.* **256**, 3598–3603.
15. Krestinger, R. H. & Nockolds, C. E. (1973) *J. Biol. Chem.* **248**, 3313–3326.
16. Wolff, J., Cook, G. H., Goldhammer, A. R. & Berkowitz, S. A. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 3841–3844.
17. Iwasa, Y., Yonemitsu, K., Matsui, K., Fukunaga, K. & Miyamoto, E. (1981) *Biochem. Biophys. Res. Commun.* **98**, 656–660.
18. Geisselsoder, J. O., Campos, J. M. & Zusman, D. R. (1978) *J. Mol. Biol.* **119**, 179–189.
19. Nakamura, K. & Inouye, M. (1979) *Cell* **18**, 1109–1117.