The Nucleotide Sequence of the $M_r = 28,500$ Flagellin Gene of *Caulobacter crescentus*

Paul R. Gill and Nina Agabian

From the Department of Biochemistry SJ 70, University of Washington, Seattle, Washington 98195

(Received for publication, February 15, 1983)

The DNA sequences which encode the $M_r = 28,500$ flagellin polypeptide of *Caulobacter crescentus* CB15 have been determined. The size of the protein, deduced from its DNA sequence (276 amino acids), is in agreement with its apparent molecular weight as measured by sodium dodecyl sulfate-polyacrylamide gel electrophoresis. The distribution of arginine residues within the protein sequence encoded by the gene correlates with their relative location as predicted by peptide alignment analysis (Gill, P. R., and Agabian, N. (1982) *J. Bacteriol.* 150, 925–933).

DNA sequences 5′ and 3′ to the coding sequence were also determined. In the 5′ region, DNA sequences homologous to consensus sequences associated with RNA polymerase recognition and transcription initiation sites in *Escherichia coli* (Pribnow box) are found. These are centered around 60, 80, and 120 base pairs upstream from the ATG codon at the beginning of the structural gene. Sequences 3′ to the coding region were identified which might signal transcription termination. A typical *E. coli* 16 S ribosomal binding site (Shine-Dalgarno sequence) is located just 5′ to the coding sequence, and for most of the amino acids there is a strong codon usage preference. Although this protein is exported from the cell (Gill, P. R., and Agabian, N. (1982) *J. Bacteriol.* 150, 925–933), the encoded NH$_2$-terminal amino acid sequence is not different from the mature product.

Cell division in the prokaryote *Caulobacter crescentus* produces two progeny cells which are each functionally differentiated with respect to surface structures (1–4) and developmental potential (5, 6). One of the more obvious manifestations of differentiation in this system is the appearance and disappearance of a single polar flagellum on one of the progeny cells during a restricted interval of the cell cycle (1, 3).

Unlike most bacterial flagella, the filament portion of the *Caulobacter* flagellum consists mainly of two distinct polypeptide monomer types: $M_r = 25,000$ and 27,500 (7). In addition, at least one ($M_r = 25,800$) and perhaps two ($M_r = 24,500$) other flagellin monomers have been identified by immunologic (3) or by genetic (8) techniques.

A comparison of the peptide maps and tryptic peptide sequences obtained from the $M_r = 25,000$ and 27,500 flagellins was carried out by Weissborn et al. (9); their studies indicated that these two proteins are encoded by distinct structural genes. In a concurrent study, we showed that the $M_r = 25,000$ flagellin and a polypeptide with an apparent $M_r = 27,500$ flagellin were derived from distinct structural genes by NH$_2$-terminal amino acid sequence analysis, peptide mapping, and a peptide alignment technique (10). Our flagellin protein isolation conditions were novel and involved the use of strong denaturants, but did not impose a criterion of acid solubility and reassociation into filaments for assay of flagellin polypeptides as described in (9). The DNA sequence reported in this paper, in conjunction with protein sequence analysis (9, 10) of the various flagellins, together indicate that the protein sequence reported by Gill and Agabian (10) was in fact that of the $M_r = 28,500$ flagellin and not the $M_r = 27,500$ flagellin analyzed by Weissborn et al. (9). These results were unexpected since no previous evidence existed for production of significant amounts of $M_r = 28,500$ flagellin outside the cell.

In fact, based on the kinetics of its appearance in synchronous cell populations (3), it had been postulated that the $M_r = 28,500$ flagellin was a precursor either to the $M_r = 27,500$ or the $M_r = 25,000$ flagellin. The combined evidence thus indicates that the $M_r = 25,000, 27,500,$ and 28,500 flagellins are each encoded by distinct structural genes.

The flagellins are the most abundant components of the flagellar apparatus and thus serve as indicators in studying the regulation of expression of this organelle. The regulation of flagellin expression during *Caulobacter* development is primarily regulated at the translational level (11, 12). Sequences flanking the flagellin coding sequence identified in this study represent possible regulatory regions which function in the transcriptional regulation of this gene. Since this is the first gene isolated and sequenced for *Caulobacter*, the analysis of codon preference for this gene provides a percentage index of codon usage in these G + C-rich organisms.

**EXPERIMENTAL PROCEDURES**

**Bacterial Strains, Viruses, and Plasmids**—Recombinant plasmid pCA110 contains a fragment of the *C. crescentus* CB15 (ATCC 19089) genome, which was identified as containing flagellin gene sequences (13). For transformation and propagation of M13mp8 and M13mp9 (14) subclones of this genomic fragment, *Escherichia coli* JM 103 was used.

**Reagents**—Restriction enzymes and T$_4$ DNA ligase were purchased from Bethesda Research Laboratories (Gaithersburg, MD) and from New England Biolabs (Beverly, MA). *E. coli* DNA polymerase I was a generous gift from the laboratory of Dr. Lawrence Loeb (University of Washington, Seattle, WA). *E. coli* DNA polymerase I large fragment (Klenow enzyme) was obtained from either New England Nuclear (Boston, MA) or from Boehringer Mannheim (Indianapolis, IN). Deoxy- and dideoxyribonucleotide triphosphates were obtained from P-L Biochemicals (Milwaukee, WI). [α-32P]dCTP was obtained from New England Nuclear (>600 Ci/mmol). The M13-specific sequencing primers (12-mer and 17-mer) were purchased from Collaborative Research, Inc. (Lexington, MA). In general, all reagents were used and stored as suggested by the respective manufacturer. Plasmid DNA and M13 double-stranded replicative form DNA were prepared by a modification of the alkaline extraction technique (15). M13

---

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

‡ Recipient of National Institutes of Health Grant ROI GM 25527.
single-stranded (+ strand) DNA was prepared by the method of\footnote{The abbreviations used are: bp, base pair; kb, kilobase pair.}.

**M13 Subcloning and DNA Sequencing Strategy**—The chain termination method of Sanger et al.\footnote{19} as adapted to the M13 system\footnote{18} was used to determine DNA sequences. This general method relies on the ability to generate overlapping clones at about 200-bp\footnote{1} intervals for both orientations of the fragment to be sequenced. These clones were obtained using subcloning strategies described below.

The complete coding sequence of the $M_r = 28,500$ flagellin gene is contained within the 2.2-kb SalI fragment of pCA110 (see Fig. 1). 200 µg of purified pCA110 plasmid DNA was digested with SalI and electrophoresed on a 1.2-cm thick 0.7% agarose gel (Seakem), and then the 2.2-kb SalI fragment was electroeluted using DEAE-paper (Whatman DE81). Polyacrylamide gels of the 2.2-kb SalI fragment produced a 0.37-kb and a 0.93-kb PsI-SalI fragment and a 0.81-kb PsI fragment. These were cloned into SalI-PstI-cut M13mp8 and mp9 vectors or into a PstI-cut M13mp8 vector, respectively. The orientation of the three PstI subfragments of the 2.2-kb SalI fragment was determined using the asymmetric HpaI, PvuII, and CiaI sites (see Fig. 1). DNA sequence analysis of one of the ends of the 0.81-kb PstI fragment located the NH2-terminal coding region and sequences 5’ to the gene. Sequence analysis of the ends of the 0.93-kb SalI-PstI fragment identified some internal sequences and the sequences in the 3’ region of the gene.

Random subclones of the 0.93-kb SalI-PstI fragment were constructed in order to obtain overlapping sequences within the gene. In one approach, AccI-digested M13mp8 vector was used to clone TaqI or HpaII partial digestion products, and a majority of the sequences in the 0.93-kb SalI-PstI fragment were determined using these subclones. Alternatively, Sau3A partial digestion products of this fragment were cloned into a Sall-BamHI-cut M13mp8 vector. The recombinants generated by this approach contained variable length inserts starting at different Sau3A sites within the 0.93-kb SalI-PstI fragment and ending at the SalI site of this fragment.

Fig. 1 shows the sequence strategy and indicates that both strands of ~80% of the gene have been sequenced. All nucleotide residues presented have been determined from at least two and usually three independent sequencing reactions.

The fragments cloned into M13 vectors for DNA sequence analysis were routinely analyzed by hybridization, tests of complementarity with other M13 clones, the size of the inserted sequence, and single lane sequence analysis\footnote{16}. Not all of these tests were used for each M13 clone.

**RESULTS AND DISCUSSION**

Both genomic and plasmid mapping analyses were used for deducing the restriction map of the 2.2-kb SalI fragment of pCA110 (Fig. 1). Shown in Fig. 2 is the DNA sequence analysis that was used to identify the coding sequence for the NH2-terminal amino acids and sequences 5’ to the $M_r = 28,500$ flagellin gene. The gene that encodes the $M_r = 28,500$ flagellin protein has been located within the Ffa E complementation group\footnote{13}.

**Transcriptional Signals**—The entire sequence of a 1200-bp segment which includes the coding sequence of the $M_r = 28,500$ flagellin gene, is presented in Fig. 3. In Fig. 4, sequences which contain potential transcription initiation regions are shown in more detail. Preceding the translation start signals at -5 to -9 (discussed below) are three regions which contain sequences that typically identify transcription start sites in \textit{E. coli} (20, 21). The sequences associated with the $M_r = 28,500$ flagellin gene are centered at about positions -60, -90, and -120 relative to the initiator Met codon sequence. In \textit{E. coli}, the consensus sequence TATAAT is centered about 10 bp upstream from the start site of transcription, and the transcripts usually begin with either a G or an A residue. In such TATAAT sequences, the last T is invariant, the beginning TA is highly conserved, and the remaining TAA is subject to greater variation (20). At positions -90 and -120 are clusters of sequences which are consistent with the TATAAT sequence, whereas at the -60 position, a single such sequence is evident (see also Fig. 4).

In many \textit{E. coli} transcription start regions, there is also a consensus sequence about 35 bp upstream from the start site; this consensus sequence is TTAGAC (20). Among the three TATAAT-like regions of the $M_r = 28,500$ flagellin gene, only the -60 region has a similar sequence (TCTTGACA) which is approximately 35 bp upstream. Since the presence of this sequence in \textit{E. coli} transcription start regions may not be absolutely required, its significance is not known (20). A comparison between the two flagellin promoter-like sequences at -90 and -120 with the two promoters associated with the developmentally regulated 0.4-kb sporulation gene of \textit{Bacillus subtilis} (22) reveal significant homology, especially in the -95 and -115 region. As shown in Fig. 4, the start position at -99, TAATCC, in the \textit{ Caulobacter} gene is very similar to the "110" promoter of the \textit{B. subtilis} 0.4-kb gene: TAATGCTT (22).

Sequence which may signal the termination of $M_r = 28,500$ flagellin gene transcription have been more difficult to identify on the basis of comparisons with known prokaryotic termination signals. The sequence homology of transcription termination signals is less conserved than those of basic promoter sequences. Furthermore, the difficulties in identifying transcription termination signals are increased for termination signals which have increased $p$ dependence; nothing is known about $p$-like factors in \textit{C. crescentus}. Nevertheless, where information is available, sequences which are correlated with transcription termination include those with regions of hyphenated dyad symmetry, containing a G+C-rich region and a T-rich region 3’ to the dyad (20). A region of hyphenated dyad symmetry 3’ to the coding region of the $M_r = 28,500$ flagellin gene exists from sequences 840 to 883 bp (Fig. 3). The T-rich region found in other systems, however, seems to be absent in the 3’-untranslated region of the \textit{ Caulobacter} gene. Whether any of these \textit{Caulobacter} sequences are functional in transcription termination must await further analysis.

In addition to the sequences shown in Fig. 3, we have sequenced 345 bp upstream from the coding region and have not identified any sequences which have homology with the
FIG. 2. Nucleotide sequence analysis used to identify the $M_e = 28,500$ flagellin gene. A, the nucleotide sequences of the 5′ region of the $M_e = 28,500$ flagellin gene is shown. Indicated are the Shine-Dalgarno consensus sequences 5′ to the coding sequences (S.D.) and the NH$_2$-terminal amino acids encoded by the gene and previously identified by protein sequencing (10). B, the dideoxy sequence analysis gel autoradiogram which shows the sequence of the sense strand (shown in A). The gel is read 5′ to 3′ from the bottom to the top. Note the CAT sequence indicated (underlined) in A and (dots) in B.

consensus TATAAT. These results might be interpreted to suggest that the transcript of the $M_e = 28,500$ flagellin gene should be about 950 nucleotides if the transcription initiation and termination signals are the same as those suggested. A transcript of about 950 nucleotides was detected by Northern analysis using a 2.2-kb Safl hybridization probe. Several factors must be considered however, before any conclusions can be made regarding the analysis of flagellin transcripts. The structural relatedness of the flagellin polypeptides (3, 9, 10, 23) is reflected at the DNA level since the genes encoding the $M_e = 28,500$ and 25,000 flagellins specifically cross-hybridize. Thus, one might expect a hybridization probe prepared for a given flagellin gene to cross-hybridize with mRNA molecules encoded by the other flagellin genes. The different flagellin proteins appear to be synthesized in vastly different amounts, and clearly the mRNA for the $M_e = 25,000$ flagellin would be expected to be the most abundant mRNA in the cell, although this may not necessarily be the case at early cell cycle times of flagellin gene expression. Further experiments are in progress to determine the appearance of flagellin mRNAs during the cell cycle using gene probes from the regions of each of the flagellin genes. The flagellin genes of other bacteria are adjacent to transcription initiation signals (24), and usually constitute a single operon. Thus, the $M_e = 28,500$ flagellin gene in Caulobacter would be novel in its regulation if the gene were a downstream gene in a polysaccharide operon.

The confirmation of sequences as being promoters of terminators of transcription must await further genetic and biochemical analysis. Techniques such as S1 mapping (25) of the ends of the transcript and in vitro transcription will enable a more comprehensive analysis of these putative regulatory regions. It must be borne in mind, however, that the cross-hybridization of the flagellin genes may be a persistent problem in these analyses, and the construction of the appropriate mutant strains may be required before meaningful conclusions can be drawn.

Translation Signals—Centered at the −7-bp position relative to the coding sequence of the $M_e = 28,500$ flagellin gene (Figs. 3 and 4) is the sequence GGAG, which corresponds to a transcribed consensus sequence found 5′ to E. coli genes. This sequence is thought to hybridize to the 3′ segment of 16 S ribosomal RNA during translation initiation (26). The appropriate complementary sequence has been found at the 3′ end of the 16 S ribosomal RNA of C. crescentus (26). The ribosomal genes of C. crescentus also appear to be similar to those of E. coli (17). A methionine codon at the 5′ end of the coding sequence for the $M_e = 28,500$ flagellin gene precedes the known NH$_2$-terminal amino acid sequence (10), suggesting that the initiator N-formylmethionine residue, which is encoded as an ATG at the 5′ end of the gene, is cleaved from the NH$_2$ terminus of the newly synthesized polypeptide. At least two termination codons exist in every reading frame within 200 bp 5′ to the coding sequence (Fig. 3). Furthermore, the occurrence of the ribosome binding site centered at −7 indicates that there is no signal sequence which is cleaved from the NH$_2$ terminus of the $M_e = 28,500$ flagellin polypeptide during export and assembly. The Salmonella typhimurium H2 antigen (flagellin) also does not contain a signal sequence (28), perhaps indicating that at least the flagellins of Gram-negative bacteria are representative of a class of proteins which do not require an NH$_2$-terminal signal sequence for externalization. Although the exported proteins of E. coli are thought to be secreted at distinct sites in the membrane (29), the processes which result in flagellin export and assembly may be quite different and may involve an interplay between flagellin and other polypeptides which make up the components of the flagellar organelle.

The translation termination codon of the gene coding sequence is UAA. This codon is probably the most efficient of the terminator codons (30) and the major terminator used in E. coli (31). Several other terminator sequences are found in all three reading frames 3′ to this termination codon.

The distribution of arginine residues in the protein is in agreement with the peptide alignment data used to character-

---

P. Gill and N. Agabian, unpublished observation.

P. Gill and N. Agabian, manuscript in preparation.
**Fig. 3.** The nucleotide sequence of the \( M_r = 28,000 \) flagellin gene and its flanking sequences. The encoded protein is preceded by a consensus sequence as predicted by Shine and Dalgarno (S.D.) (26). Dyad symmetry is indicated by the **horizontal arrows**. The arginine residues within the protein that were predicted by the peptide alignment (10) are indicated by the **asterisks**. The polypeptide region indicated by the **hatched line** contains residues with significant antigenic potential (46), as determined by computer analysis. Some restriction endonuclease sites are indicated for orientation.

The gene sequence is \( M_r = 19,600 \) and 13,200 (Fig. 3). The distribution of the glutamic acid and aspartic acid residues is also consistent with the peptides generated in peptide mapping experiments. **Codon Usage**—There is a strong codon preference for most of the amino acids encoded in the \( M_r = 28,500 \) flagellin gene. For a number of highly expressed genes of *E. coli* (32) and the budding yeast *Saccharomyces cerevisiae* (33), there is a strong codon preference; about 70% of the codons that are preferred by the \( M_r = 28,500 \) flagellin gene are the same ones preferred by the highly expressed *E. coli* genes (Table 1). This is some-
FIG. 4. Nucleotide sequences 5’ to the coding region of the $M_r = 28,500$ flagellin gene. Arrows indicate regions of inverted repeat sequences. The Shine-Dalgarno consensus sequence is boxed and the brackets at -60, -90, and -120 bp indicate possible promoter sequences. The 5’ region of the $M_r = 28,500$ flagellin bears some homology with the dual promoters of $B. subtilis$ (32) whose sequence is included for comparison.

what surprising since the G + C percentage for the Caulobacter genome is about 62.5% (34), whereas that of the E. coli genome is about 50.1% (35). The differences in the codon preference between Caulobacter and E. coli are usually wobble position changes involving a change of an A:T bp (E. coli) to a G:C bp (C. crescentus).

In E. coli, the codons that are preferred are directly correlated with the relative abundance of tRNA isoacceptor species, in the cell (32, 36). Changes in the relative abundance of isoacceptor pools for given tRNA species have been correlated with distinct differentiated states in a number of cell types (37). It will be interesting to compare the codon preferences for a number of Caulobacter genes expressed at different times in development.

Dyad Symmetry 5’ to the Gene—5’ to the coding region of the $M_r = 28,500$ flagellin gene are several sequence domains which exhibit dyad symmetry; the most obvious are indicated with horizontal lines in Figs. 3 and 4. The largest domain of dyad symmetry consists of a 24-bp repeat centered around position -60 (Fig. 4). Other regions of dyad symmetry are present which overlap the 5’ arm of this large domain and several sequences 3’ to this domain. Another dyad overlaps the 3’ arm of the large domain and the sequences which encode the first few amino acids in the coding sequence. In addition, two other overlapping regions of dyad symmetry are centered around positions -20 and -12 relative to the initiation codon of the coding sequence. One of these (-12) overlaps the proposed ribosomal binding site (Shine-Dalgarno sequence).

The extent of dyad symmetry in the 5’ region suggests the potential formation of secondary structure which might have functional significance in the regulation of mRNA expression. The largest domains of dyad symmetry in the 5’ region of the gene suggest the possibility that specific regulatory proteins might interact with DNA sequences in a manner analogous to the interaction of the lac repressor and the lac operator (38). The function of such a DNA binding protein might be to repress the transcription of this gene during a specific period of the cell cycle. The occurrence of stage-specific DNA binding proteins has been noted in Caulobacter (5, 6, 39).

A (dC-dG) sequence block is located just 3’ to the large dyad centered at -60 bp. When the negative superhelical density of DNA containing alternating dC-dG sequences is greater than 0.072, these sequences can undergo a transition in helical state (from right-handed to left-handed) (40). The dC-dG block located between -60 and the structural gene might thus act to facilitate the formation of a cruciform structure by the large dyad. If the formation of such a cruciform structure is required for regulating transcription, the degree of supercoiling in this region may influence flagellin gene expression. The highly condensed state of the newly replicated swarmer chromosome probably reflects a change in its superhelical density as compared with that of stalked cell chromosome (41). A discrete interval of DNA replication is required for flagellin synthesis (42) which further suggests that localized changes in supercoiling as occurs during DNA replication may influence flagellin gene expression.

Although nothing is known about the stability of the $M_r = 28,500$ flagellin mRNA, the $M_r = 25,000$ flagellin mRNA has been shown to have a relatively long functional half-life (2, 11). The mRNA species encoded by the lipoprotein gene (43) and the ompA gene (44) of E. coli also have relatively long functional half-lives and predictions of secondary structure have suggested that a number of stem-loop structures could be formed to stabilize the mRNA of these genes in vivo. A similar kind of argument may be made for the $M_r = 28,500$ flagellin mRNA based on possible secondary structure of the mRNA. If transcription was initiated at -90 (Fig. 4), then a particularly stable mRNA structure might be formed centered at position -60. These stem-loop structures in mRNA may interact specifically with other cellular components, perhaps to sequester the mRNA at defined sites in the cell membrane during flagellin biosynthesis.

In contrast to the secondary structures proposed for ompA and lipoprotein mRNAs, the $M_r = 28,500$ flagellin gene has regions of dyad symmetry which include the ribosomal binding sites. This raises the possibility that secondary structure in the region can be important in regulating the initiation of flagellin mRNA translation. In support of this possibility, Inserentant and Fiers (45) have shown that the efficiency of translation initiation is directly dependent on the degree of

| Table I: $M_r = 28,500$ flagellin gene codon usage |
|-----------------------------------------------|
| Codon | Phe | Ser | Tyr | Cys |
|------------------------|-----|-----|-----|-----|
| 0 UUU | 1 UCU | 0 UAU | 1 UGU |
| 5 UUC | 0 UCC | 0 UAC | 0 UGC |
| 0 UUA | 0 UCA | 0 UAA | 0 UGA |
| 2 UUG | 9 UCG | 0 UAG | 1 UGG |

| Codon | Pro | His | Cys |
|---------------------|-----|-----|-----|
| 2 CUU | 2 CCU | 1 CAU | 4 CGU |
| 9 CUC | 1 CCC | 1 CAC | 6 CGC |
| 0 CUU | 2 CCA | 2 CAA | 0 CGA |
| 22 CUG | 2 CCG | 16 CAG | 1 CGG |

| Codon | Asn | Arg |
|-------------------|-----|-----|
| 0 AUA Il | 0 ACU | 6 AUA | 0 AGU |
| 13 AUG | 13 ACC | 13 AAC | 7 AGC |
| 0 AUA Il | 0 ACA | 0 AAA | 0 AGA |
| 5 AUG Met | 9 ACG | 10 AAG | 1 AGG |

| Codon | Ala | Asp | Gly |
|-----------------|-----|-----|-----|
| 1 GUU Val | 0 GCU | 6 GUA | 5 GGU |
| 7 GUC Val | 26 GCC | 17 GAC | 13 GGC |
| 0 GUA Val | 4 GCA | 2 GAA | 2 GGA |
| 9 GUG Val | 10 GCG | 5 GAG | 4 GGG |

Downloaded from http://www.jbc.org/ by guest on March 24, 2020
secondary structure in the region 5′ to the structural gene.

Antigenicity of Protein as Predicted from Sequence—The organization of the \( M_r = 28,500 \) flagellin protein encoded in the DNA sequence is in good agreement with a structural model of flagellins proposed by Iino (24). In this model, the NH₂-terminal sequences, and to a lesser extent the carboxyl-terminal sequences, are essential for flagellin function. The central portion of a bacterial flagellin monomer appears to be quite variable and is usually the site of mutations which alter antigenicity. Immunochemical and chemical analysis (24) of the central portion of the protein sequence of other flagellin monomers indicates that this region typically contains the antigenic determinants for the protein (24). We have found significant sequence homology between \( M_r = 25,000 \) and 28,500 flagellin genes in both the NH₂-terminal and carboxyl-terminal regions of the protein as would be predicted by lino’s model. The encoded protein of \( M_r, 28,500 \) flagellin gene was tested for antigenicity using an antigenic site predictor (46) computer program; the central portion of the protein encoded by the \( M_r = 28,500 \) flagellin gene contains a region of significant antigenic potential (see Fig. 3).

The relative position of the tryptic peptide sequenced by Weissborn et al. (9) from the \( M_r = 25,000 \) polypeptide was aligned with the \( M_r = 28,500 \) flagellin gene sequence by comparison of the COOH-terminal regions of the \( M_r = 25,000 \) and 28,000 genes. The sequence of the A16 peptide from \( M_r = 25,000 \) and the homologous sequence of B13 peptide from the \( M_r = 27,000 \) gene fall within the central region of the polypeptide. Thus, as predicted by lino (24), peptides which differentiate between flagellin subunits are located within the central portion of the protein in the region of greatest antigenic variability as predicted by the computer program. Our results indicate then that there should be two classes of flagellin epitopes: one class residing in the NH₂-and/or carboxyl-terminal sequences should be common to all the monomer types, and the other class should be found in the central, variable portion of the polypeptide and should vary between flagellins. Lagenaun and Agabian (23) have noted representatives of the former class in showing that antibody prepared against sodium dodecyl sulfate gel-purified \( M_r = 25,000 \) flagellin cross-reacts completely with \( M_r = 27,500 \) flagellin. Koyasu et al. (47) have recently shown that antibodies can be made against flagellins prepared under non denaturation conditions which distinguish between the different flagellins. Taken together, the immunological and sequence data would suggest that association of flagellin monomers into a filament structure conceals NH₂- and carboxyl-terminal portions of the peptide. In fact, it is not possible to decorate intact Caulobacter flagella with our \( M_r = 25,000 \) flagellin cross-reacting antibody except in a discrete region at the filament hook junction.⁴ The high degree of sequence homology found in the NH₂-terminal region of flagellins from both Gram-positive and Gram-negative, enteric and nonenteric bacteria (10) suggests that these regions of the molecule are essential for polymerization. This is in fact borne out by mutational analysis (24).

Relationship to Other Flagellin Genes—Lagenaun and Agabian (3) first observed in C. crescentus CB13 that in addition to the two major filament monomers, \( M_r = 25,000 \) and 27,500 flagellins, two other polypeptides of similar molecular weight and periodicity of expression during the cell cycle were immunoprecipitated with anti-flagellin antibody. These other polypeptide species have apparent \( M_r = 26,500 \) and 24,500. It was suggested that the \( M_r = 28,500 \) flagellin was a precursor form of one of the major polypeptide species because it was produced in small amounts and it appeared transiently during the cell cycle, just preceding the appearance of \( M_r = 27,500 \) and 25,000 flagellins. Several lines of evidence, discussed below, suggest that the gene described in this paper encodes the \( M_r = 25,000 \) flagellin polypeptide described by Lagenaur and Agabian and the sequence data provided in this report formally exclude the possibility that the \( M_r = 28,500 \) flagellin is a precursor to either \( M_r = 25,000 \) or 27,000 flagellin.

Flagellin monomer purification procedures reported by others (9, 48) and those previously described from this laboratory (23) used intact flagellar filaments as starting material. The more recent technique used by us to prepare flagellin polypeptides involves their isolation from insoluble aggregates which form in the medium under certain growth conditions; these aggregates contain large bundles of intact flagellar filaments and much greater quantities of flagellin are obtained under these conditions as compared with those used previously (10). The purification of flagellin monomers from these aggregates requires their denaturation in urea and Triton X-100 and these denatured flagellins were used as a starting material for purification. Although the protein we purified and used for structural and NH₂-terminal amino acid sequence analysis had an apparent \( M_r = 27,500 \) and had the same ionic exchange elution properties of the \( M_r = 27,500 \) flagellin as described by Fukuda et al. (48), its actual molecular weight is closer to 28,500 as deduced by DNA sequence analysis. The small differences in molecular weight between this polypeptide and the \( M_r = 27,500 \) flagellin, coupled with its ion exchange behavior after denaturation in urea and Triton X-100, were not recognized as unusual at the time.

After determination of the sequence of the \( M_r, 28,500 \) flagellin (then thought to be the \( M_r = 27,500 \) polypeptide), we scanned the DNA sequence for the corresponding B13 peptide fragment sequence determined by Weissborn et al. (9). This sequence is not encoded in this flagellin gene although the NH₂-terminal amino acid sequence determined in our previous studies corresponds with that predicted at the DNA sequence level. Subsequently, the \( M_r = 25,000 \) flagellin gene has been located and partially sequenced. Within this sequence, we have found the Weissborn et al. (9) A16 peptide sequence and the NH₂-terminal amino acid sequence determined by Gill and Agabian (10). Genomic restriction endonuclease mapping has also located two other chromosomal regions large enough to encode additional flagellin polypeptides. These regions are candidates for encoding the \( M_r = 27,500 \), and perhaps the \( M_r = 24,500 \), flagellin polypeptide. The results of these experiments will be presented in detail elsewhere.

The unexpected abundance of this \( M_r = 29,500 \) flagellin polypeptide in our protein preparations could indicate that other methods for preparing flagellin proteins which utilize acid dissociation and in vitro reassociation as a functional criterion for flagellin purification may have selected against obtaining this monomer type. Nevertheless, the mature flagellar filaments extruded into the medium during stalked cell morphogenesis (3) do not have significant quantities of this polypeptide. The large amount of \( M_r = 28,500 \) flagellin obtained from extracellular material in our new purification protocol (10) suggests that this monomer type is exported from the cell in significant amounts in wild type organisms. Certain flagellin-minus mutants of Caulobacter crescentus, although lacking both \( M_r = 28,500 \) and 27,500 flagellin, can assemble a \( M_r = 28,500 \) flagellin into short stubby filaments (8). These data together with the strong sequence homology among members of the flagellin family suggest that the \( M_r = 28,500 \) flagellin can function as a flagellin monomer under certain circumstances and may perhaps have a transient role.

---

⁴ J. Smit and N. Agabian, unpublished data.
in filament assembly. This is in agreement with the kinetics of its expression during the cell cycle in that its synthesis precedes that of the \( M_1 = 27,500 \) flagellin (3). Alternatively, a very small proportion of the flagellin filaments may substitute \( M_2 = 28,500 \) for \( 27,500 \) during assembly. Resolution of the functional role of each of these polypeptides in flagellar structure and assembly must await further genetic and biochemical characterizations which are in progress.

Acknowledgments—We thank Tom Holland and Dr. Mike Parker for the computer programs used in the analysis of the DNA sequences, and Pam Horbett for careful typing of this manuscript.

REFERENCES
1. Shapiro, L. (1976) *Annu. Rev. Microbiol.* 30, 377-407
2. Agabian, N., Evinger, M., and Parker, G. (1979) *J. Cell Biol.* 81, 123-136
3. Lagenaur, C., and Agabian, N. (1978) *J. Bacteriol.* 135, 1062-1069
4. Smit, J., Hermodson, M., and Agabian, N. (1981) *J. Biol. Chem.* 256, 3092-3097
5. Evinger, M., and Agabian N. (1979) *Proc. Natl. Acad. Sci. U. S. A.* 76, 175-178
6. Milhausen, M., and Agabian, N. (1981) *J. Bacteriol.* 148, 163-173
7. Lagenaur, C., and Agabian, N. (1977) *J. Bacteriol.* 132, 731-733
8. Johnson, R. C., Ferber, D. M., and Ely, B. (1983) *J. Bacteriol.* 150, 925-933
9. Weissborn, A., Steinman, H., and Shapiro, L. (1982) *J. Bacteriol.* 143, 317-328
10. Gill, P. R., Parker, G., and Agabian, N. (1982) *Proc. Natl. Acad. Sci. U. S. A.* 79, 6847-6851
11. Birnboim, H. C., and Doly, J. (1979) *Nucleic Acids Res.* 7, 1513-1521
12. Birnboim, H. C., and Doly, J. (1979) *Methods Enzymol.*, in press
13. Sanger, F., Nicklen, S., and Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U. S. A.* 74, 5463-5467
14. Messing, J., Crea, R., and Seeburg, P. H. (1981) *Nucleic Acids Res.* 9, 309-321
15. Dretzen, C., Bellard, M., Sassone-Corsi, P., and Chamber, P. (1981) *Anal. Biochem.* 122, 295-298
16. Rosenberg, M., and Court, D. (1979) *Annu. Rev. Genet.* 13, 319-358
17. Sibbenlist, U., Simpson, R. B., and Gilbert W. (1980) *Cell* 20, 269-281
18. Moran, C. P., Lang, N., Banner, C. D. B., Haldenang, W. G., and Losick, R. (1981) *Cell* 25, 783-792
19. Lagenaur, C., and Agabian, N. (1976) *J. Bacteriol.* 128, 435-444
20. Iino, T. (1977) *Annu. Rev. Genet.* 11, 161-182
21. Berk, A. J., and Sharp, P. A. (1975) *Cell* 12, 712-729
22. Shino, J., and Dalgar, N. (1975) *Eur. J. Biochem.* 57, 221-230
23. Ohta, N., and Newton, A. (1982) *J. Mol. Biol.* 153, 291-303
24. Simon, M., Zieg, J., Silverman, M., Mandel, M., and Dolittle, R. (1986) *Science (Wash. D. C.)* 209, 1370-1374
25. Davis, B. D., and Tai, C. P. (1980) *Nature (Lond.)* 282, 433-438
26. Whittfield, H. (1972) in *The Mechanism of Protein Synthesis and Its Regulation* (Bosch, L., ed) pp. 243-283, Elsevier, New York
27. Brenner, M. S., Streton, A. O., and Kaplan, S. (1965) *Nature (Lond.*) 206, 994-998
28. Ikemura, T. (1981) *J. Mol. Biol.* 146, 1-21
29. Bennetzen, J. L., and Hall, B. (1982) *J. Biol. Chem.* 257, 3026-3031
30. Agabian-Keshishian, N., and Shapiro, L. (1970) *J. Virol.* 5, 798-800
31. Belozersky, A. N., and Spire, A. S. (1960) in *The Nucleic Acids* (Chargoff, E., and Davidson, J. N., eds) Vol. 3, p. 147, Academic Press, New York
32. Post, L., and Normura, M. (1980) *J. Biol. Chem.* 255, 4550-4666
33. Cortese, R. (1979) in *Biological Regulation and Development* (Goldberger, R. F., ed) pp. 401-432, Plenum Press, New York
34. Barley, M. D., Riggs, A., Hobe, A., and Bourgeois, S. (1975) *Biochemistry* 14, 1700-1710
35. Cheung, K. K., and Newton, A. (1977) *J. Biol. Chem.* 252, 417-425
36. Singleton, C. K., Klysik, J., Stirividant, S. M., and Wells, R. D. (1982) *Nature (Lond.)* 302, 312-316
37. Ewing, M., and Agabian, N. (1979) *Proc. Natl. Acad. Sci. U. S. A.* 76, 175-176
38. Sheffery, M., and Newton, A. (1981) *Cell* 24, 49-57
39. Nakamura, K., Pirtle, I. L., Takeishi, K., and Inouye, M. (1981) *J. Biol. Chem.* 255, 210-216
40. Rao Morra, N., Nakamura, K., and Inouye, M. (1980) *J. Mol. Biol.* 143, 317-328
41. Ewing, M., and Agabian, N. (1979) *Proc. Natl. Acad. Sci. U. S. A.* 76, 175-176
42. Shiff, M., and Newton, A. (1981) *Cell* 24, 49-57
43. Nakamura, K., Pirtle, I. L., Takeishi, K., and Inouye, M. (1981) *J. Biol. Chem.* 255, 210-216
44. Hiroshi, M., and Agabian, N. (1980) *Nature (Lond.*) 289, 312-316
45. Milhausen, M., and Agabian, N. (1983) *Nature (Lond.)* 302, 620-622
46. Milhausen, M., Gill, P. R., Parker, G., and Agabian, N. (1982) *Proc. Natl. Acad. Sci. U. S. A.* 79, 6847-6851
47. Messing, J., and Vierira, J. (1982) *Gene* 19, 269-276
48. Sanger, F., Nicklen, S., and Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U. S. A.* 74, 5463-5467
49. Messing, J., Crea, R., and Seeburg, P. H. (1981) *Nucleic Acids Res.* 9, 309-321
50. Dretzen, C., Bellard, M., Sassone-Corsi, P., and Chamber, P. (1981) *Anal. Biochem.* 122, 295-298
The nucleotide sequence of the Mr = 28,500 flagellin gene of Caulobacter crescentus.
P R Gill and N Agabian

J. Biol. Chem. 1983, 258:7395-7401.

Access the most updated version of this article at http://www.jbc.org/content/258/12/7395

Alerts:
• When this article is cited
• When a correction for this article is posted

Click here to choose from all of JBC's e-mail alerts

This article cites 0 references, 0 of which can be accessed free at http://www.jbc.org/content/258/12/7395.full.html#ref-list-1