THE JOURNAL OF BIOLOGICAL CHEMISTRY
Vol. 262, No. 20, Issue of July 15, pp. 9724-9729, 1987
Printed in U.S.A.

The Primary Structure of a Procaryotic Glycoprotein

CLONING AND SEQUENCING OF THE CELL SURFACE GLYCOPROTEIN GENE OF HALOBACTERIA*

(Received for publication, February 17, 1987)

Johann Lechner and Manfred Sumper‡

From the Institute für Biochemie, Genetik und Mikrobiologie, Universität Regensburg, Universitätsstrasse 31, 8400 Regensburg, Federal Republic of Germany

The hexagonally patterned surface layer of halobacteria consists of a true glycoprotein. This procaryotic glycoprotein has recently been shown to exhibit novel features with respect to saccharide structure and saccharide biosynthesis.

The primary structure and the location of glycosylation sites were determined by cloning and sequencing of the glycoprotein gene of *Halobacterium halobium*. According to the predicted amino acid sequence, the glycoprotein is synthesized with a N-terminal leader sequence of 34 amino acid residues reminiscent of eucaryotic and procaryotic signal peptides. A hydrophobic stretch of 21 amino acid residues at the C terminus probably serves as a transmembrane domain. 14 threonine residues are clustered adjacent to this membrane anchor and linked to these threonines are all the disaccharides of the cell surface glycoprotein. 12 N-glycosylation sites are distributed over the polypeptide chain.

As early as 1956 Houwink (1) demonstrated that the surface of halobacteria are covered by a hexagonally patterned macromolecular monolayer (surface layer, S-layer). Mescher and Strominger (2) were the first to demonstrate the occurrence of a true glycoprotein as the main constituent of the S-layer in this procaryotic organism. Structural work following this discovery lead to a rather detailed view of the saccharide structure and of the biosynthesis of this glycoprotein from *Halobacterium halobium* (3–10).

The cell surface glycoprotein (CSG)[1] of halobacteria contains three types of protein-linked saccharides. 1) A single high molecular weight saccharide which is composed of repeats of sulfated pentasaccharides. It is linked to the polypeptide chain via the recently described new linkage Asn-GalNAc (10). 2) About 10 sulfated oligosaccharides (hexuronic acid 1–4)$_{2-3}$Glc are connected to the polypeptide via Asn-Glc, another novel type *N*-glycosidic linkage (6). 3) About 20 disaccharides Glc1–2Gal are *O*-glycosidically attached to threonine residues of the polypeptide chain (2, 5).

In this paper we describe the isolation and characterization of the CSG gene and correlate the deduced CSG primary structure with the established structural data of the protein-linked saccharides.

MATERIALS AND METHODS

Strains

H. halobium strain R_1M_1 was grown in "complex medium" as described in Ref. 11.

The following *Escherichia coli* strains were used for transformation experiments: HB 101 for pIN-III recombinants, JM 103 for M13/mp18 recombinants, and JM 109 for pUC 18 recombinants. *E. coli* NM 538 was used to grow λ EMBL 4, NM 539 was used to grow recombinants of λ EMBL 4. The λ phage EMBL 4 is described in Ref. 12.

Plasmids

pUC 18 (13) was obtained from Bethesda Research Laboratories. pIN-III (14) was obtained from Prof. Inouye and has the following features. The *E. coli* lipoprotein promoter and the 95-bp UV 5 promoter-operator region are tandemly inserted, so that a cloned gene is not expressed in the absence of lac inducer. A nucleotide sequence of 22 bp which contains *Eco*RI, *Hin*dIII, and *Bam*HI sites is inserted at the position coding for the third amino acid of the prolipoprotein, resulting in the three vector variants pIN-III-A1, pIN-III-A2, and pIN-III-A3, where A1, A2, and A3 stand for the three possible reading frames. Any expression of cloned DNA should result in a fusion protein consisting of the amino-terminal prolipoprotein amino acids followed by the insert-coded amino acids.

DNA Preparations

Halobacterial DNA—40 ml of bacterial suspension in complex medium (A_{578} = 1) were lysed and digested by Pronase as described (15). Starting with phenol extraction, a protocol was followed for DNA isolation from cells grown in tissue culture as described in Ref. 16. The final purification step was a CsCl density gradient centrifugation as described (16).

Plasmid and Phage DNA—These recombinants were isolated by the boiling method described in Ref. 16. pUC 18 recombinants were isolated as described in Ref. 17. λ Phage EMBL 4 DNA was isolated essentially according to Ref. 16.

Halobacterial RNA Preparation

Cells (10 g) were lysed as described in Ref. 18 and RNA was prepared by pelleting through a CsCl cushion following the protocol for the guanidinium/CsCl method described in Ref. 16.

Construction and Screening of the pIN-III Library

Halobacterial DNA was partially digested with *Mbo*I restriction endonuclease and fractionated by sucrose density gradient centrifugation (SW40 Beckman, 5–20% sucrose, 15 h at 36,000 rpm, 20 °C). The DNA fraction containing fragments of 1–4 kb was collected and ligated into the *Bam*HI site of plasmids pIN-III-A1, pIN-III-A2, and pIN-III-A3. The resulting plasmids were used to transform *E. coli* HB 101 by the procedure of Ref. 16 and selected for ampicillin resistance.

For immunological screening of the pIN-III library replica filters

* This work was supported by the Deutsche Forschungsgemeinschaft (SFB 43). 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.

The nucleotide sequence(s) reported in this paper has been submitted to the GenBank™/EMBL Data Bank with accession number(s) J02767.

‡ To whom correspondence should be addressed.

[1] The abbreviations used are: CSG, cell surface glycoprotein; bp, base pair; kb, kilobase pair; ORF, open reading frame; SDS, sodium dodecyl sulfate; IPTG, isopropyl-1-thio-β-D-galactopyranoside.

(Schleicher & Schüll, nitrocellulose) carrying bacterial colonies were transferred to H-agar plates (16) containing IPTG (480 μg/ml) and ampicillin (40 μg/ml) and incubated for 1.5 h. The filters were transferred to Whatman 3MM paper which was soaked in lysozyme solution (2 mg/ml) in 10 mM Tris-HCl, pH 8, 1 mM EDTA and kept there for 10 min. This step was repeated once. The filters were then transferred to Whatman 3MM paper, soaked in 0.5% *N,N*-dimethyldodecylamine-*N*-oxide in 10 mM Tris-HCl, pH 8, and incubated there for 30 min. The filters were kept on the paper and brought into a $CHCl_3$ chamber for 30 min. After air drying, the filters were washed in TNK (10 mM Tris-HCl, pH 8, 200 mM NaCl, 200 mM KCl) for 10 min, then in TNK containing 5% calf serum for 30 min, and finally twice in TNK for 10 min. Thereafter filters were incubated in antiserum solution (5% in Tris) and washed successively in TNK for 10 min and twice in calf serum (5% in TNK) for 10 min. An affinity column carrying proteins from an *E. coli* lysate was used to remove corresponding antibodies from the antiserum. Finally, filters were incubated with fluorescein isothiocyanate anti-rabbit antibody (Sigma; 1.25% in TNK) for 1 h in the dark, washed as above, and air dryed.

Construction and Screening of the λ EMBL 4 Library

Halobacterial DNA was digested with *Sau*IIIA under conditions optimized for the generation of 12–16-kb fragments. 14–16-kb fragments were prepared by preparative agarose gel electrophoresis and were treated with calf intestine alkaline phosphatase (Boehringer Mannheim). λ EMBL 4 DNA, digested with *Bam*HI, was cleaved by *Sal*I and precipitated by propanol-2 to remove the short cohesive ends of the middle fragment. The DNA precipitate was dissolved and ligated in the presence of the 14–16-kb halobacterial DNA fragments. Recombinant DNA was packaged *in vitro* (16) yielding a total of $1.5 \cdot 10^7$ recombinant phages and used to infect *E. coli* NM 539. Since NM 539 is a host lysogene for P1, transformants were selected for the Spi^- phenotype, and thus for the absence of the λ EMBL 4 middle fragment.

Hybridization of the recombinant plaques was done essentially as described in Ref. 16, but using a "hybridization incubator" (Bachhofer, Reutlingen, Federal Republic of Germany). Prehybridization was done for 4 h in 3 ml of prehybridization buffer/five filters (nitrocellulose, Schleicher & Schüll, 82 mm) at the hybridization temperature. Prehybridization solution consisted of 50% formamide, 5 × Denhardt's solution (Ficoll, 1 mg/ml; polyvinylpyrrolidone, 1 mg/ml; bovine serum albumin, 1 mg/ml), 5 × SSPE (43.5 mg/ml NaCl, 6.9 mg/m! $NaH_2PO_4 \cdot 2H_2O$, 1.85 mg/ml EDTA, pH 7.4) (1 × SSPE: 8.7 mg/ml NaCl, 1.38 mg/ml $NaH_2PO_4 \cdot 2H_2O$, 0.37 mg/ml EDTA, pH 7.4), 0.1% SDS, and 100 μg/ml denatured salmon sperm DNA. If DNA fragments labeled by nick translation with ^{35}S were used as probes, the prehybridization solution was adjusted to 100 mM dithiothreitol and the probe was added to give 10^6 cpm/ml. Hybridization was performed for 48 h at 42 °C and thereafter filters were washed 3 times in 2 × SSC (17.5 mg/ml NaCl, 8.8 mg/ml Na_3 citrate, pH 7.0) (1 × SSC: 8.76 mg/ml NaCl, 4.4 mg/ml Na_3 citrate; pH 7.0), 0.1% SDS (10 min; 25 °C), and once in 1 × SSC, 0.1% SDS (1 h, 68 °C).

Oligonucleotides used as hybridization probes were labeled by [γ-^{32}P] ATP and polynucleotide kinase to a specific activity of $6 \cdot 10^6$ cpm/pmol and were added to the prehybridization solution to give $3 \cdot 10^6$ cpm/ml. Hybridization was performed for 48 h at 37 °C, and filters were washed twice in 2 × SSC, 0.1% SDS (10 min; 25 °C) and 3 times in 1 × SSC, 0.1% SDS (10 min; 37 °C).

DNA Sequencing

DNA sequencing was performed by the dideoxy chain termination method (19) using ^{35}S-labeled α-thio-dATP. M13 clones were prepared as described in Ref. 20 and pUC 18 clones as described in Ref. 17.

Synthesis of Oligonucleotides

Oligonucleotides were synthesized on an Applied Biosystems DNA synthesizer.

Primer Extension

1 μg of total RNA and 25 fmol of single-stranded primer were incubated in 10 μl of RT buffer (50 mM Tris-HCl, pH 8.0, 8 mM $MgCl_2$, 50 mM KCl) at 55 °C for 15 min and at room temperature for 40 min. Dithiothreitol was added to a concentration of 40 mM and dGTP, dCTP, and dTTP to 1 mM each. After addition of ^{35}S-labeled α-thio-dATP and 1 μl of 10 × RT buffer to a final volume of 20 μl, the reaction was initiated with 10 units of reverse transcriptase (ASV, Pharmacia, Sweden) and allowed to proceed for 30 min at 42 °C. The reaction was terminated by addition of 20 μl of stop solution (90% formamide, 10 mM EDTA, pH 8.3, 3 mg/ml xylene cyanol, 3 mg/ml bromphenol blue). Prior to application to a 6% denaturing polyacrylamide gel, the sample was heated to 96 °C for 3 min and immediately cooled to 0 °C.

RESULTS

Construction of Libraries of the Halobacterial Genome—Two different genomic libraries of halobacterial DNA were used to isolate the complete CSG gene. Isolated total DNA from the halobacterial strain *H. halobium* R_1M_1 was partially digested with the endonuclease *Sau*IIIA and fragments of the size range 1–4 kb were inserted into the unique *Bam*HI site of the high level expression vector pIN-III, constructed by Masui and Inouye (14). Fragments of the partial digest in the size range of 14–15 kb were inserted into the phage λ vector EMBL 4.

Foreign DNA inserted into the expression vector pIN-III is under control of the lac UV 5 operator-promoter region which renders transcription from the insert DNA inducible by IPTG. Usage of a mixture of three variants of pIN-III (pIN-III A1, -A2, and -A3) ensures expression of insert DNA in all three possible reading frames.

After induction by IPTG, the library was screened with a polyclonal rabbit antibody raised against purified CSG. Positive clones were visualized by the indirect immunofluorescence staining technique. Out of 40,000 clones, a single immunopositive one was identified (clone 59). Cell lysates of clone 59, grown in the absence and presence of the inducer IPTG were compared on SDS-polyacrylamide gels. The protein pattern of induced clone 59 showed one additional prominent band. Upon Western blotting and immunostaining with CSG antibody, this inducible protein turned out to be the only one to give a positive signal. However, this inducible protein of clone 59 migrated on SDS-polyacrylamide gels much faster than authentic (hydrofluoric acid deglycosylated) cell surface glycoprotein. Consequently, the insert DNA of clone 59 carried only a fragment of the CSG gene.

To isolate the complete CSG gene, a second genomic library was screened with clone 59 insert DNA (labeled by nick translation). This library was constructed by ligating 14–15-kb fragments of halobacterial DNA to the arms of the λ EMBL 4 phage. Out of 3000 plaques, 18 clones gave a positive hybridization signal. These preselected clones were again screened with an oligonucleotide probe designed to detect the coding region for the N terminus of CSG. Since 21 positions of the N-terminal amino acid sequence of CSG were known from Edman degradation studies, a corresponding oligonucleotide mixture could be synthesized. Decisions regarding nucleotide assignment at redundant positions were guided by the codon usage data obtained from the bacterio-opsin gene from halobacteria (21). The oligonucleotide probe used was the following.

```
TTC AAC GAA AAC ACA AAC TAT A
            G       G       C
Phe Asn Glu Asn Thr Asn Tyr Thr
```

By screening with this probe, plaque λ-10/1 gave the strongest signal and was therefore used for the sequencing work. The halobacterial DNA insert of λ-10/1 was recovered by digestion with endonuclease *Eco*RI. The isolated 15-kb insert was further cleaved with endonuclease *Sst*I and a corresponding Southern blot was probed with the oligonucleotide as well as insert DNA of clone 59. A 3.1-kb fragment of λ-10/1 insert DNA hybridized to the oligonucleotide, whereas the 3.1-kb fragment and a 2.5-kb fragment gave a positive hybridization

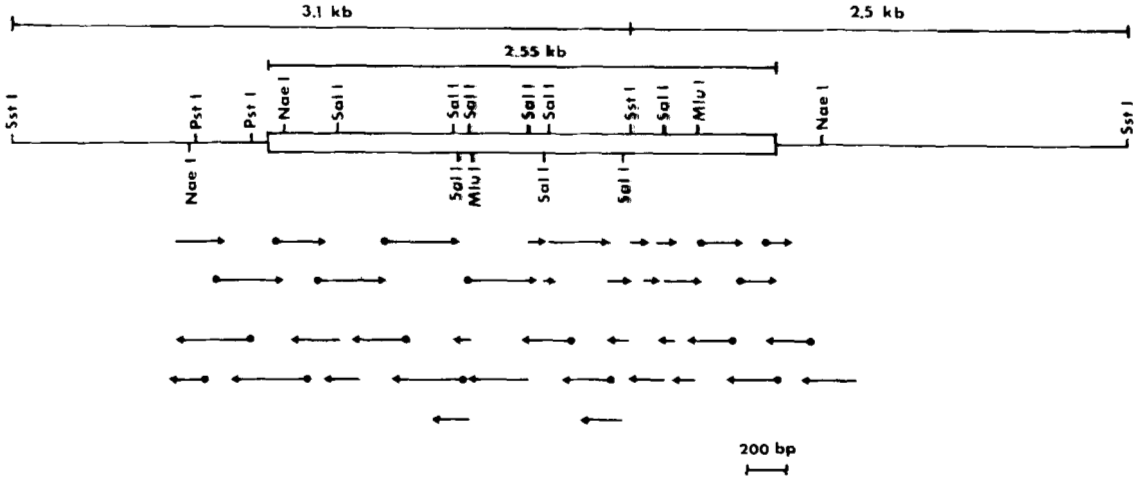

FIG. 1. **Restriction map of CSG gene and sequencing strategy.** The restriction map includes both the 3.1- and 2.5-kb *Sst*I fragments of the λ-10-1 insert DNA which contains the CSG coding region. *Arrows* indicate the strand and extent of DNA sequenced. Sequencing experiments which used synthetic primers are marked by ●. All other sequencing reactions used a 15-mer oligonucleotide universal primer.

signal with clone 59 insert DNA. Thus, the 3.1-kb fragment covers the coding region for the N-terminal portion of CSG and consequently the 2.5-kb fragment has to code for the more C-terminal portion of CSG.

DNA Sequence—The 3.1-kb as well as the 2.5-kb *Sst*I fragment of the λ-10/1 insert DNA was subcloned into the high copy number plasmid pUC 18 (13). In addition, digests with *Sau*IIIA and *Acc*I of both of these fragments were subcloned into the replicative form of M13mp18 bacteriophage. The strategy for sequencing the entire CSG gene is summarized in Fig. 1. The dideoxy chain termination method was used throughout. Sequencing experiments using the M13 vector were primed with a commercial 15-mer oligonucleotide primer, all sequencing experiments using the pUC 18 vector were primed with synthetic 17-mer oligonucleotides.

In Fig. 2 the results of all sequencing experiments are summarized. The 3167-bp segment shown was sequenced in both directions. An open reading frame (ORF) starts at position 457 with an ATG codon and ends at position 3015 with a TGA codon. The N-terminal amino acid sequence of CSG was determined earlier and found to be: Ala-Asn-Ala-Ser-Asp-Leu-Asn-Asp-Tyr-Gln-Arg-Phe-Asn-Glu-Asn-Thr-Asn-Tyr-Thr-Tyr-Ser- (10). This sequence starts with the 35th amino acid of the ORF (position 1 in Fig. 2). Three other stretches of the predicted amino acid sequence (*underlined* in Fig. 2) were confirmed by sequencing of pure peptides.[2] This allows the firm conclusion that the translation of the ORF indeed represents the primary structure of CSG.

The sequence downstream of the translational stop codon TGA is typical for terminator structures. A GC-rich self-complementary region (stem and loop, see *arrows* in Fig. 2) is followed by an AT-rich sequence.

Initiation Site of CSG mRNA Synthesis—The ORF starts 102 bp upstream of the N terminus of mature CSG, indicating the occurrence of a signal sequence on the nascent translation product. The open reading frame upstream of the N terminus includes two methionines: Met −34 and Met −18. Therefore, the initiation site of CSG mRNA synthesis was determined by a primer extension experiment.

A synthetic 17-mer oligonucleotide, TCGTTCAGGTCGCTTGC, complementary to the mRNA (covering the amino acid position 3–8 of ORF) was incubated together with a crude RNA fraction from halobacteria and primer extension was performed with avian myoblastosis virus reverse transcriptase. The same 17-mer oligonucleotide was used as a primer in a Sanger sequencing experiments with the 3.1-kb *Sst*I fragment as template. The products of the sequencing reaction and of the primer extension were compared on a sequencing gel (Fig. 3). Primer extension resulted in a homogenous DNA fragment and co-chromatographed with the DNA fragment generated in the Sanger sequencing reaction by chain termination at 346 bp in the 5′-untranslated region of the CSG gene (Fig. 2). CSG mRNA synthesis therefore starts at 346 bp with an adenosine nucleotide residue that is 111 bp upstream of Met −34.

Protein Structure—Assuming that the first methionine encoded by the mRNA is used to start translation, the protein chain is initiated with methionine −34. Thus, the CSG gene encodes for a leader peptide being absent in the mature glycoprotein. The sequence of this N-terminal extension is: Met-Thr-Asp-Thr-Thr-Gly-Lys-Leu-Arg-Ala-Val-Leu-Leu-Thr-Ala-Leu-Met-Val-Gly-Ser-Val-Ile-Gly-Ala-Gly-Val-Ala-Phe-Thr-Gly-Gly-Ala-Ala-Ala. The amino acid sequence around the potential cleavage site is Ala-Ala-Ala-Ala, a motif frequently used by signal peptidases (22).

Twelve potential *N*-glycosylation sites (sequons Asn-*X*-Ser(Thr); $X \neq$ Asp,Pro) are distributed throughout the polypeptide chain. The asparagine at position 2 is linked to the repetitive pentasaccharide structure occurring once per glycoprotein molecule (10). From the chemical data obtained previously it was concluded that about 10 sulfated saccharides of the type (hexuronic acid 1–4)$_{2-3}$Glc are linked *N*-glycosidically to the polypeptide chain (6). These data imply that most or all potential glycosylation sites are indeed linked to saccharides. Glycopeptide sequence analysis has directly confirmed glycosylation at Asn-2, Asn-479, and Asn-609. However, from N-terminal protein sequencing it was found that the sequon at position Asn-17 does not serve as a saccharide acceptor.

It was found previously, that all of the disaccharides linked *O*-glycosidically to CSG via threonine can be recovered in a single glycopeptide after Pronase digestion (5). This indicated a highly clustered arrangement of the *O*-linked disaccharides. The DNA sequencing data confirmed this arrangement: 14 Thr residues are clustered within a stretch of 19 amino acids. Since CSG carries about 20 disaccharide units per molecule,

[2] G. Paul, F. Lottspeich, and F. Wieland, unpublished results.

```
TGATCGGTGGCGAACGGAGCCCCCGCATGGATGTTGTTACCCGCGCCGTCGCTCCCCGCGACG   63

GCCGGCTGTGCAGCGACACCGCCTGCAGGGACGACGAGCCCGGCGATGAGAACACAGACGCCG  126

ACGACGAGGCACTTCCGTGGCCCACTCCCGCTCATTGCGGCTTCGGTTCGAACTGGGGCATCT  189

TTAACCCCCGTTTTCGCCGGACGGCACCGTGCGTGATGTCGCCGCTCCGCCATCCACGTTTCA  252

TGTGAGCAATACACACCCAAATCGGTGTTCTGACGGCGGCTGACGCGCGAAAAGGCAGAAAGC  315

ATTTACCAGTGGCCGGGTATAGTCTGGAGCACCCCTACCCGAAATGGCGGCTGCAGAAACCCA  378

CGATTCACCGTTTCGGCGGAATCAGGTGGATGCGGTCGTCTGGTTGGACTGACACCGTAGCTC  441
                    -34
AGTCACTCAGTAAAA ATG ACA GAC ACA ACA GGC AAA CTC CGC GCA GTC CTC   492
                Met Thr Asp Thr Thr Gly Lys Leu Arg Ala Val Leu

CTG ACG GCG CTG ATG GTC GGT TCC GTA ATC GGA GCC GGC GTC GCG TTC  540
Leu Thr Ala Leu Met Val Gly Ser Val Ile Gly Ala Gly Val Ala Phe
                        -1  1
ACG GGC GGG GCT GCT GCG GCG AAT GCA AGC GAC CTG AAC GAT TAT CAG  588
Thr Gly Gly Ala Ala Ala Ala Asn Ala Ser Asp Leu Asn Asp Tyr Gln

CGG TTC AAC GAA AAT ACA AAC TAC ACG TAT AGT ACC GCC TCA GAA GAC  636
Arg Phe Asn Glu Asn Thr Asn Tyr Thr Tyr Ser Thr Ala Ser Glu Asp

GGT AAA ACC GAA GGA AGT GTC GCC AGT GGC GCG ACC ATC TTC CAG GGC  684
Gly Lys Thr Glu Gly Ser Val Ala Ser Gly Ala Thr Ile Phe Gln Gly
                                    50
GAA GAG GAC GTT ACC TTC CGG AAG CTG GAC AAC GAG AAA GAG GTC AGT  732
Glu Glu Asp Val Thr Phe Arg Lys Leu Asp Asn Glu Lys Glu Val Ser

CCG GCG ACC CTC TCC CGC ACT GGC GGG TCT GAC GAG GGC GTT CCT CTC  780
Pro Ala Thr Leu Ser Arg Thr Gly Gly Ser Asp Glu Gly Val Pro Leu

CAG ATG CCG ATC CCC GAG GAC CAG TCG ACC GGT TCC TAC GAT AGC AAT  828
Gln Met Pro Ile Pro Glu Asp Gln Ser Thr Gly Ser Tyr Asp Ser Asn
                                        100
GGT CCA GAC AAC GAC GAG GCT GAC TTC GGC GTT ACG GTC CAG AGT CCA  876
Gly Pro Asp Asn Asp Glu Ala Asp Phe Gly Val Thr Val Gln Ser Pro

TCG GTG ACG ATG CTC GAA GTC CGC AAC AAC GCG GAC AAC GAC GTC ACC  924
Ser Val Thr Met Leu Glu Val Arg Asn Asn Ala Asp Asn Asp Val Thr

GGC GGT GTC CTG AAC ACA CAG CAG GAC GAG TTC TCG ATC GCC GTT GAC  972
Gly Gly Val Leu Asn Thr Gln Gln Asp Glu Ser Ser Ile Ala Val Asp
                                            150
TAC AAC TAC TAC GCT GCC GAG GAC CTC GAG CTG ACC GTC GAA GAC GAG 1020
Tyr Asn Tyr Tyr Ala Ala Glu Asp Leu Glu Leu Thr Val Glu Asp Glu

GAC GGT CTC GAC GTT ACG GAC GAG ATC CTC GCT GCC GAC CAG TCG GGC 1068
Asp Gly Leu Asp Val Thr Asp Glu Ile Leu Ala Ala Asp Gln Ser Gly

GGC GCG TAC GAA GAT GGC ACC GGA AAC AAC GGG CCC AAC ACG CTT CGC 1116
Gly Ala Tyr Glu Asp Gly Thr Gly Asn Asn Gly Pro Asn Thr Leu Arg
                                                    200
TTC GAC ATC GAC CCG AAC AAC GTT GAC GCG GGC GAC TAC ACG GTC TCG 1164
Phe Asp Ile Asp Pro Asn Asn Val Asp Ala Gly Asp Tyr Thr Val Ser

GTT GAA GGT GTC GAG GAC CTG GAC TTC GGT GAC GCC ACC GAG TCC GCC 1212
Val Glu Gly Val Glu Asp Leu Asp Phe Gly Asp Ala Thr Glu Ser Ala

TCC GTG ACG ATT TCC TCC TCG AAC AAG GCA TCG CTG AAC CTC GCC GAG 1260
Ser Val Thr Ile Ser Ser Ser Asn Lys Ala Ser Leu Asn Leu Ala Glu
                                                            250
GAC GAA GTC GTG CAG GGA GCG AAC CTC AAG TAC ACC ATC GAG AAC AGT 1308
Asp Glu Val Val Gln Gly Ala Asn Leu Lys Tyr Thr Ile Glu Asn Ser

CCG GAA GGC AAC TAC CAC GCT GTC ACC ATC GAC AGC AGC GAC TTC CGC 1356
Pro Glu Gly Asn Tyr His Ala Val Thr Ile Asp Ser Ser Asp Phe Arg

GAC AGC AGC AGC GGT GCT GAT GCC GCG AAA GTC ATG CGC AGC GTT GGT 1404
Asp Ser Ser Ser Gly Ala Asp Ala Ala Lys Val Met Arg Ser Val Gly

GAC ACT GTC GAC ACC GGT CTC GTC GTC GAC AAC GAC AGT ACC ACC GAA 1452
Asp Thr Val Asp Thr Gly Leu Val Val Asp Asn Asp Ser Thr Thr Glu
    300
ATT GTA GAC GAC TAT GAA AAC ACC TCG ATC TCG GAC GTC GAC TAC GCG 1500
Ile Val Asp Asp Tyr Glu Asn Thr Ser Ile Ser Asp Val Asp Tyr Ala

TAC GCC ATC GTC GAG ATC GAC GAC GGA AAC GGC GTC GGG TCC ATC GAG 1548
Tyr Ala Ile Val Glu Ile Asp Asp Gly Asn Gly Val Gly Ser Ile Glu

ACG CAG TAC CTC GAT GAC TCC AGC GCC GAC ATC GAC CTC TAC CCC GCA 1596
Thr Gln Tyr Leu Asp Asp Ser Ser Ala Asp Ile Asp Leu Tyr Pro Ala
            350
TCC GAC ACC GAA GAC GCC CCG GAT TAC GTC AAT AGC AAC GAA GAA CTC 1644
Ser Asp Thr Glu Asp Ala Pro Asp Tyr Val Asn Ser Asn Glu Glu Leu

ACA AAC GGC TCC GCC CTC GAC GGC GTC TCT ACC GAC GAC GAC ACT GAC 1692
Thr Asn Gly Ser Ala Leu Asp Gly Val Ser Thr Asp Asp Asp Thr Asp

TTC GAC GTC ACG CAG GGC GAC ATC ACC CTC GAC AAC CCG ACC GGC GCG 1740
Phe Asp Val Thr Gln Gly Asp Ile Thr Leu Asp Asn Pro Thr Gly Ala
                        400
TAC GTT GTC GGC TCG GAA GTC GAC ATC AAC GGG ACC GCC AAC GAG GGG 1788
Tyr Val Val Gly Ser Glu Val Asp Ile Asn Gly Thr Ala Asn Glu Gly

ACT GAC GAC GTC GTG CTG TAC GCT CGC GAC AAC AAC GAC TTC GAA CTC 1836
Thr Asp Asp Val Val Leu Tyr Ala Arg Asp Asn Asn Asp Phe Glu Leu

GTC ACC GTC GAC GGT GAG AAA TCG ATC GAA GTC GAC AGC GAC GAC ACG 1884
Val Thr Val Asp Gly Glu Lys Ser Ile Glu Val Asp Ser Asp Asp Thr
                                    450
TTC GAG GAG GAA GAC ATC ACG CTC TCC GAT GGA GAC AAG GGC GGT GAC 1932
Phe Glu Glu Glu Asp Ile Thr Leu Ser Asp Gly Asp Lys Gly Gly Asp

GAC ATC CTT GGT CTC CCC GGT ACG TAC CGC CTC GGC ATC ATC GCC AAG 1980
Asp Ile Leu Gly Leu Pro Gly Thr Tyr Arg Leu Gly Ile Ile Ala Lys

AGC GAC GCC GTC AAC TCG TCC GGT GGC GTG AAA GAC AAC ATT GAC ACG 2028
Ser Asp Ala Val Asn Ser Ser Gly Gly Val Lys Asp Asn Ile Asp Thr
                                            500
TCC GAC TTC AAC CAG GGC GTC AGC AGT ACG TCC TCC ATC CGT GTG ACC 2076
Ser Asp Phe Asn Gln Gly Val Ser Ser Thr Ser Ser Ile Arg Val Thr

GAC ACG GAA CTC ACC GCG TCC TTC GAG ACC TAC AAC GGG CAG GTC GCC 2124
Asp Thr Glu Leu Thr Ala Ser Phe Glu Thr Tyr Asn Gly Gln Val Ala

GAC GAC GAC AAC CAG ATC GAC GTT GAG GGG ACT GCC CCT GGG AAG GAC 2172
Asp Asp Asp Asn Gln Ile Asp Val Glu Gly Thr Ala Pro Gly Lys Asp
                                                550
AAC GTT GCC GCC ATC ATC ATC GGC AGC CGT GGC AAG GTC AAG TTC CAG 2220
Asn Val Ala Ala Ile Ile Ile Gly Ser Arg Gly Lys Val Lys Phe Gln

TCC ATC TCC GTC GAC AGC GAC GAC ACG TTC GAC GAG GAG GAC ATC GAC 2268
Ser Ile Ser Val Asp Ser Asp Asp Thr Phe Asp Glu Glu Asp Ile Asp

ATC TCG GAG CTC CGA CAG GGC AGT GCT TCC GCA CAC ATC CTC TCG TCG 2316
Ile Ser Glu Leu Arg Gln Gly Ser Ala Ser Ala His Ile Leu Ser Ser
                                                        600
GGT CGT GAC GGG AAG TTC GGT GAG GAC ACC GCC AAC AGC ATT AGC GAT 2364
Gly Arg Asp Gly Lys Phe Gly Glu Asp Thr Ala Asn Ser Ile Ser Asp

CTT GAG GAC GAA GTC GGT AAC TAC ACC TCG GGT CCG CCG ACT GGC GAC 2412
Leu Glu Asp Glu Val Gly Asn Tyr Thr Ser Gly Pro Pro Thr Gly Asp

CAG ATC CGC GAC CGC ATC CTC TCG AAC ACG GTC GAC GAC ACC GCC AGC 2460
Gln Ile Arg Asp Arg Ile Leu Ser Asn Thr Val Asp Asp Thr Ala Ser
                                                            650
GAC GAC CTC ATC GTC ACG CAG CAG TTC CGT CTC GTT GAC GGA CTC ACC 2508
Asp Asp Leu Ile Val Thr Gln Gln Phe Arg Leu Val Asp Gly Leu Thr

ACG ATC GAA GCC ACT GAG GGT GGC GAA GCG GGC GGC TCG GTC ACC GTC 2556
Thr Ile Glu Ala Thr Glu Gly Gly Glu Ala Gly Gly Ser Val Thr Val

ATG GGG ACG ACC AAC CGG AAG GCC GAC GAC AAC ACC ATC ACG GTT GAA 2604
Met Gly Thr Thr Asn Arg Lys Ala Asp Asp Asn Thr Ile Thr Val Glu

CTC CTC CAG GGC GAC GCG TCC ATC GAG ATC AAC AGC ACT GAT GAG TGG 2652
Leu Leu Gln Gly Asp Ala Ser Ile Glu Ile Asn Ser Thr Asp Glu Trp
    700
AAC AGC GAC GGC CAG TGG TCG GTT GAT GTC CCG CTC TCG AAC GTC GAG 2700
Asn Ser Asp Gly Gln Trp Ser Val Asp Val Pro Leu Ser Asn Val Glu

CCG GGC AAC TAC ACG GTC GAA GCT GAC GAC GGT GAC AAC ACC GAC CGT 2748
Pro Gly Asn Tyr Thr Val Glu Ala Asp Asp Gly Asp Asn Thr Asp Arg

CAG AAC GTC GAA ATC GTC GAG GAA CTC GAG GAG CCT GAT CAG ACG ACC 2796
Gln Asn Val Glu Ile Val Glu Glu Leu Glu Glu Pro Asp Gln Thr Thr

                750
GTC GAT CAG CCC GAG AAC AAC CAG ACG ATG ACG ACG ACG ATG ACC GAG 2844
Val Asp Gln Pro Glu Asn Asn Gln Thr Met Thr Thr Thr Met Thr Glu

ACG ACC ACC GAG ACG ACC ACC GAG ATG ACC ACC ACG CAG GAG AAC ACC 2892
Thr Thr Thr Glu Thr Thr Thr Glu Met Thr Thr Thr Gln Glu Asn Thr

ACC GAG AAC GGC TCG GAG GGC ACT TCC GAT GGC GAG TCA GGC GGC AGC 2940
Thr Glu Asn Gly Ser Glu Gly Thr Ser Asp Gly Glu Ser Gly Gly Ser
                                800
ATC CCC GGC TTC GGT GTC GGT GTT GCG CTC GTC GCG GTC CTC GGT GCG 2988
Ile Pro Gly Phe Gly Val Gly Val Ala Leu Val Ala Val Leu Gly Ala

GCG CTG CTG GCA CTC CGC CAG AAC TGA TTGACCCACTGAATCACGTCTGACCGC 3042
Ala Leu Leu Ala Leu Arg Gln Asn

GCGTACGCGGTCACTTGCGGTGCCGTTTTCTTTGTTACCGACGACCGACCAGCGACAGCCACC 3105

GCGCGCTCACTGCCACCAAAAGAGTCATATCACAGCCGACCAGTTTCTGGAACGTTCCCGAT  3167
```

FIG. 2. **DNA sequence of CSG gene.** The DNA sequence was determined in both directions as described under "Materials and Methods." The *numbers* on the *right* indicate nucleotide positions. *Numbers above* the nucleotide sequence designate amino acid positions. The N-terminal amino acid of the mature protein was marked *1*. Negative values indicate a leader sequence. Sequences which were confirmed by peptide sequencing are *underlined*. *Boxed* regions represent potential *N*-glycosylation sites and *arrows* indicate self-complementary nucleotide sequences.

all of these Thr residues within this cluster are likely to be engaged in *O*-glycosidic linkages.

As demonstrated by a hydropathy analysis (Fig. 4) according to Ref. 23 the entire polypeptide chain of the mature glycoprotein shows a single highly hydrophobic stretch of 21 amino acids (positions 795–815) which is only 3 amino acid positions away from the C terminus. Most probably, this hydrophobic peptide serves as a membrane anchor. All other regions of the polypeptide chain of the mature CSG mainly consist of polar amino acid residues and show a dominant negative net charge. From the predicted amino acid sequence for the mature CSG protein, a molecular mass of 86,538 daltons is calculated.

Fig. 5 schematically summarizes the structural features of CSG with respect to saccharide attachment sites and the hydrophobic peptide stretch.

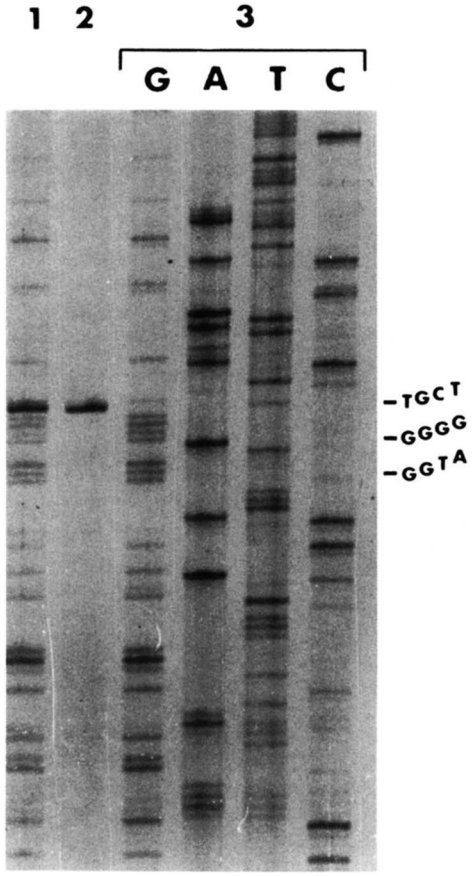

FIG. 3. **Mapping of the transcription initiation site of the CSG gene.** A primer extension experiment using halobacterial RNA as template was performed as described under "Materials and Methods." A synthetic 17-mer oligonucleotide served as primer (*lane 2*). The same oligonucleotide was used in a sequencing experiment (dideoxy chain termination) with the 3.1-kb fragment of the λ-10-1 insert DNA as template (*lane 3*). In *lane 1*, the primer extension product was co-electrophoresed with the DNA fragments shown in *lane 3G* of the sequencing experiment. The primer extension product ends with a thymidine nucleotide at 346 bp (Fig. 2). Therefore, RNA synthesis is initiated with an adenine nucleotide at position 346 of the DNA sequence.

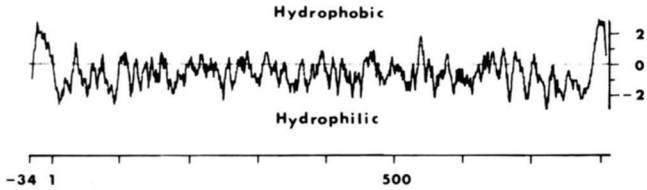

FIG. 4. **Hydropathy blot of CSG and its leader peptide according to Ref. 23.** The *curve* is the average of a hydrophobicity index for each residue over a window of 9 residues.

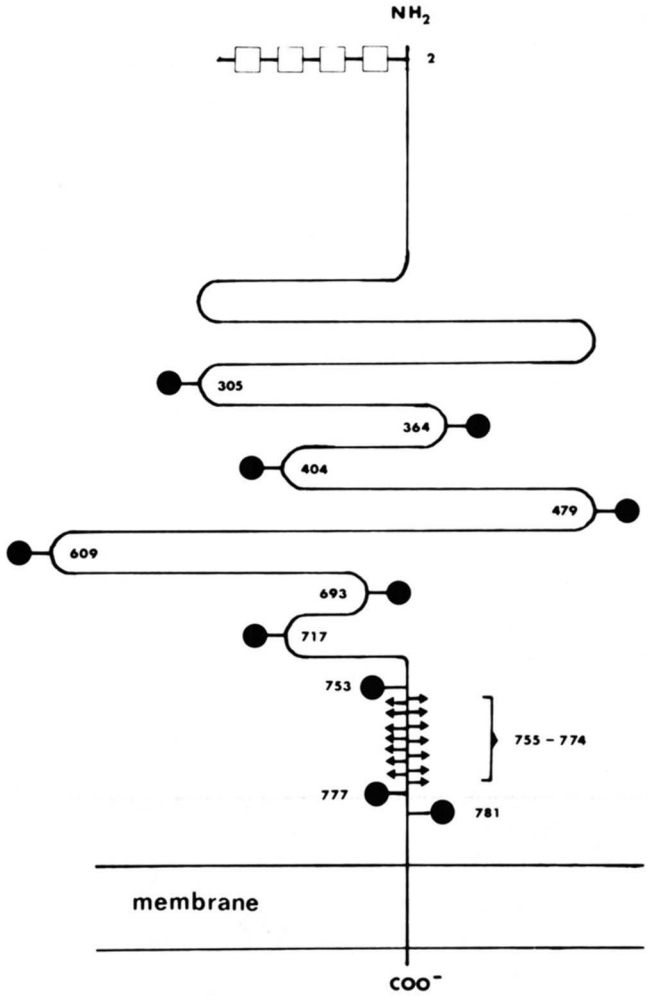

FIG. 5. **Schematic representation of CSG showing the different glycosylation sites and the membrane-binding domain.** Numbering indicates amino acid positions. □, the repeated unit saccharide; ●, *N*-linked oligosaccharides (hexuronic acid 1–4)$_{2-3}$Glc; ←, *O*-linked Glc1–2Gal disaccharides. Amino acids at position 2, 479, 609, 753, 755, 757, 758, 759, and 761 were confirmed to be glycosylated by sequencing of the corresponding glycopeptides. For further explanations see text. The extracellular surface is at the top of the membrane.

DISCUSSION

The gene of the cell surface glycoprotein of *H. halobium* was cloned and sequenced. It is the first procaryotic glycoprotein gene described.

As might be expected for a protein that has to traverse a membrane co- or post-translationally, the CSG gene codes for a signal peptide sequence which is absent in the mature glycoprotein. The signal sequence consists of 34 amino acid residues which is in the range of the longest leader sequences known so far. The signal sequence resembles eucaryotic and procaryotic signal peptides (22, 24) with respect to the following features. 1) It contains positively charged amino acids in the N-terminal region. 2) A stretch of at least 8 hydrophobic amino acid residues is localized 6 amino acid positions away from the cleavage site. 3) The sequence preceding the cleavage site is Ala-Ala-Ala, consistent with the proposed recognition sequence Ala-*X*-Ala for signal peptidases (23).

The CSG signal sequence, however, lacks negatively charged amino acids in the C-terminal region, described for most eucaryotic and procaryotic signal sequences. This may be related to the fact that the halobacterial membrane is devoid of lipids with positively charged head groups.

A second leader sequence is known for another halobacterial membrane protein, bacterio-opsin (21, 25). This sequence (Met-Leu-Glu-Leu-Leu-Pro-Thr-Ala-Val-Glu-Gly-Val-Ser-) lacks all of the characteristics summarized above. Possibly, quite different mechanisms are involved in the incorporation of the membrane protein bacterio-opsin (with seven transmembrane helices) and of CSG with a single transmembrane sequence.

The amino acid sequence as predicted from the CSG gene corresponds to a polypeptide chain with molecular mass of

86,538 daltons. The saccharide moieties (1 repetitive pentasaccharide with 10–15 repeats, 10 sulfated oligosaccharides, and about 20 *O*-glycosidically linked disaccharides) add to this value at best 30,000 daltons. Therefore the molecular mass of 200,000 daltons estimated for CSG on SDS-polyacrylamide gels is clearly too high an estimate. Most probably the unusual hydrophilic and acidic composition of CSG causes this aberrant migration behavior of CSG on SDS-polyacrylamide gels.

With respect to the protein-carbohydrate linkage units, the CSG exhibits novel features as yet not known from eucaryotic glycoproteins. Two different types of *N*-glycosidic linkages, Asn-Glc and Asn-GalNAc, are synthesized on the same polypeptide chain. A single Asn-GalNAc linkage is located at position 2 of the mature CSG polypeptide and this linkage is synthesized by transfer of the completed saccharide from a lipid pyrophosphate carrier (4). All the other *N*-glycosidic linkages are of the Asn-Glc type and are synthesized by transfer of a saccharide from a dolichyl monophosphate carrier (7). This unique situation of two different types of *N*-glycosidic linkages at defined positions of the polypeptide chain brings up the question as to how the (two?) saccharyl transferases involved discriminate between the glycosylation sites. A comparison of the surrounding amino acid sequences of the Asn-GalNAc site and of all the other sequons reveals a remarkable difference. All sequon sequences are preceded by 1 or even 2 negatively charged amino acid residues with the only exception of the unique Asn-GalNAc site. This latter sequon is preceded only by hydrophobic amino acid residues, provided glycosylation on Asn-2 occurs on the nascent polypeptide chain still being linked to the leader peptide. Possibly, the saccharyltransferase giving rise to the Asn-GalNAc linkage remains inactive on sequons being N terminally proximal to a negatively charged environment. Since oligosaccharyl transfer to synthetic oligopeptides was shown to work in the halobacterial system *in vivo* (8), this possibility can now be tested using appropriate model peptides as artificial acceptors.

Recently, much progress has been made in the establishment of three-dimensional structures of S-layers by image processing of electron microscopic pictures (26). Since the primary structure of the halobacterial S-layer glycoprotein is now available, this glycoprotein should become an attractive object for three-dimensional structural studies. However, detailed structural investigations are not yet available for the halobacterial S-layers, mainly because the high salt concentrations required to maintain the integrity of halobacterial S-layers hamper electron microscopic studies.

Acknowledgments—We wish to thank S. Stammler and U. Stöckl for excellent technical assistance and Prof. Tanner for reading the manuscript.

REFERENCES

1. Houwink, A. L. (1956) *J. Gen. Microbiol.* **15,** 146–150
2. Mescher, M. F., and Strominger, J. L. (1976) *J. Biol. Chem.* **251,** 2005–2014
3. Wieland, F., Dompert, W., Bernhardt, G., and Sumper, M. (1980) *FEBS Lett.* **120,** 110–114
4. Wieland, F., Lechner, J., Bernhardt, G., and Sumper, M. (1981) *FEBS Lett.* **132,** 319–323
5. Wieland, F., Lechner, J., and Sumper, M. (1982) *Zentralbl. Bakteriol. Mikrobiol. Hyg. 1 Abt. Orig. C* **3,** 161–170
6. Wieland, F., Heitzer, F., and Schaefer, W. (1983) *Proc. Natl. Acad. Sci. U. S. A.* **80,** 5470–5474
7. Lechner, J., Wieland, F., and Sumper, M. (1985) *J. Biol. Chem.* **260,** 860–866
8. Lechner, J., Wieland, F., and Sumper, M. (1985) *J. Biol. Chem.* **260,** 8984–8989
9. Wieland, F., Lechner, J., and Sumper, M. (1986) *FEBS Lett.* **195,** 77–81
10. Paul, G., Lottspeich, F., and Wieland, F. (1986) *J. Biol. Chem.* **261,** 1020–1024
11. Sumper, M., and Herrmann, G. (1978) *Eur. J. Biochem.* **89,** 229–235
12. Frischauf, A.-M., Lehrach, H., Poustka, A., and Murray, N. (1983) *J. Mol. Biol.* **170,** 827–842
13. Yanich-Perron, C., Vieira, J., and Messing, J. (1985) *Gene (Amst.)* **33,** 103–119
14. Masui, Y., Mizumo, T., and Inouye, M. (1984) *Biotechniques* **2,** 81–85
15. Vogelsang, H., Oertel, W., and Oesterhelt, D. (1983) *Methods Enzymol.* **97,** 226–241
16. Maniatis, T., Fritsch, E., and Sambrook, J. (1982) in *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY
17. Hattori, M., and Sakaki, Y. (1986) *Anal. Biochem.* **152,** 232–238
18. Chang, S. H., Majumdar, A., Dunn, R., Makabe, O., RajBhandary, U. L., Khorana, H. G., Ohtsuka, E., Tanaka, T., Taniyama, Y. O., and Ikehara, M. (1981) *Proc. Natl. Acad. Sci. U. S. A.* **78,** 3398–3402
19. Sanger, F., Nicklen, S., and Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U. S. A.* **74,** 5463–5467
20. Messing, J. (1983) *Methods Enzymol.* **101,** 20–78
21. Dunn, R., McCoy, J., Simsek, M., Majumdar, A., Majumdar, A., Chang, S. H., RajBhandary, U. L., and Khorana, H. G. (1981) *Proc. Natl. Acad. Sci. U. S. A.* **78,** 6744–6748
22. Perlman, D., and Halvorson, H. O. (1983) *J. Mol. Biol.* **167,** 391–409
23. Kyte, J., and Doolittle, R. F. (1982) *J. Mol. Biol.* **157,** 105–132
24. Michaelis, S., and Beckwith J. (1982) *Annu. Rev. Microbiol.* **36,** 435–465
25. Dellweg, H.-G., and Sumper, M. (1980) *FEBS Lett.* **116,** 303–306
26. Baumeister, W., Barth, M., Hegerl, R., Guckenberger, R., Hahn, M., and Saxton, O. W. (1986) *J. Mol. Biol.* **187,** 241–253