Proc. Natl. Acad. Sci. USA
Vol. 78, No. 11, pp. 6744–6748, November 1981
Biochemistry

The bacteriorhodopsin gene

(mRNA/S1 nuclease mapping/DNA sequence determination/codon usage/*Halobacterium halobium*)

ROBERT DUNN, JOHN MCCOY, MEHMET SIMSEK, ALOKES MAJUMDAR, SIMON H. CHANG*, UTTAM L. RAJBHANDARY, AND H. GOBIND KHORANA

Departments of Biology and Chemistry, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

Contributed by H. Gobind Khorana, July 28, 1981

ABSTRACT **The bacteriorhodopsin gene has been identified in a 5.3-kilobase restriction endonuclease fragment isolated from *Halobacterium halobium* DNA, using a cloned cDNA fragment as the probe. Of the 1229 nucleotides whose sequence was determined in the genomic fragment, 786 correspond to the structural gene of bacteriorhodopsin, 360 are upstream from the initiator methionine codon, and 83 are downstream from the COOH terminus. The bacteriorhodopsin gene codes for a precursor sequence of 13 amino acids at the NH_2 terminus, 248 amino acids that are present in the mature protein, and an additional aspartic acid at the COOH terminus. This determination of the DNA sequence for an archaebacterial gene reveals that the standard genetic code is used; however, there is a marked preference for either G or C in the third codon position. The gene does not contain any intervening sequences and no prokaryotic promoter can be identified in the region immediately upstream from the structural gene. The bacteriorhodopsin mRNA contains at the 5′ terminus only three nucleotides beyond the initiating AUG codon and this terminus can form a hairpin structure. Immediately downstream from this structure there is a sequence complementary to the 3′ terminus of *H. halobium* 16S rRNA.**

Bacteriorhodopsin, the only protein in the purple membrane of *Halobacterium halobium*, catalyzes the light-dependent translocation of protons and thus generates a transmembrane electrochemical gradient (1). The protein consists of a single polypeptide chain of 248 amino acids (2) and contains one molecule of retinaldehyde per protein linked as a Schiff's base to the ε-amino group of a lysine residue (1, 3). The amino acid sequence of the protein is known (2, 4), and a three dimensional model has been developed (5) that is compatible with this sequence and the diffraction data (6, 7). According to this model the polypeptide chain traverses the membrane seven times in the form of α-helical rods.

In view of these and other studies (1), bacteriorhodopsin is the most attractive energy-transducing system for mechanistic investigation. One versatile approach to such studies involves alterations of amino acids at predetermined sites in the molecule. Such an approach is now feasible by the application of general techniques of site-directed mutagenesis to the bacteriorhodopsin gene. The first requirement in developing this approach is the isolation and characterization of the gene from *H. halobium* DNA. Towards this goal, we have recently prepared and characterized a cDNA fragment, 74 nucleotides long, that corresponds to the 5′-proximal end of bacteriorhodopsin mRNA (8). This cDNA fragment has now been cloned and used as a probe for the bacteriorhodopsin gene. The latter has been identified in a 5.3-kilobase (kb) *Pst* I fragment of *H. halobium* DNA. Within this fragment the sequences of the structural gene, of 360 nucleotides upstream from the initiator codon, and of 83 nucleotides downstream from the COOH terminus have been determined. The amino acid sequence of the protein as deduced from the DNA sequence is in complete agreement with the sequence previously obtained by protein sequencing methods (2). In addition, the DNA sequence shows that the precursor to bacteriorhodopsin contains an extra 13 amino acids at its NH_2 terminus and a single extra amino acid residue at the COOH terminus. Further, some structural features of the 5′-terminal region of the bacteriorhodopsin mRNA are discussed.

MATERIALS AND METHODS

Materials. Deoxynucleoside triphosphates were from Sigma and P-L Biochemicals, α-^{32}P-labeled deoxynucleoside triphosphates were from Amersham. Carrier-free [γ-^{32}P]ATP was synthesized as described (9, 10). Restriction endonucleases were from Bethesda Research Laboratories (Rockville, MD), Boehringer Mannheim, and New England BioLabs. Phage T4 DNA polymerase, DNA polymerase I (large fragment), and S1 nuclease were from Bethesda Research Laboratories. DNA polymerase I was from Boehringer Mannheim. Dideoxynucleoside triphosphates were from P-L Biochemicals and Bethesda Research Laboratories. The 26-nucleotide-long primer specific for DNA sequences in M13mp7 was from Bethesda Research Laboratories. Synthetic *Bam*HI linkers were obtained from Collaborative Research. Reverse transcriptase (RNA-dependent DNA polymerase) from avian myeloblastosis virus was provided by J. W. Beard (Life Sciences, St. Petersburg, FL). Polynucleotide kinase was purified from phage T4-infected *Escherichia coli* cells as described (11). *H. halobium* strain S9 was obtained from W. Stoeckenius.

Synthesis and Cloning of the Double-Stranded cDNA (ds-cDNA) Fragment. (i) *Synthesis.* The single-stranded cDNA fragment (Fig. 1) was synthesized in a reverse transcriptase-catalyzed reaction by using the synthetic dodecanucleotide d(C-C-A-G-A-T-C-C-A-C-T-C) as the primer and partially purified bacteriorhodopsin mRNA as the template (8). The product was purified by electrophoresis on an 8% polyacrylamide gel in the presence of 8 M urea. After treatment with 0.3 M NaOH to remove residual RNA (25°C, 12 hr) the cDNA fragment (68 fmol) was dissolved in 30 μl of 50 mM Tris·HCl (pH 8.2)/70 mM KCl/10 mM $MgCl_2$/0.1 mM EDTA/1 mM dNTPs containing [α-^{32}P]GTP (5000 dpm/pmol). This reaction mixture was heated to 70°C for 3 min and chilled on ice, and to it were added dithiothreitol (10 mM) and 30 units of reverse transcriptase. After incubation at 42°C for 60 min, the reaction mixture was made 15 mM in EDTA and was heated to 90°C for 2 min together with 2 μg of carrier tRNA. Reaction products were chromatographed on a Sephadex G-100 column (1.0-ml plastic pipette) in 50 mM Tris·HCl (pH 8.0)/0.5 M NaCl/1 mM EDTA. The excluded radioactivity was precipitated with 3 vol of ethanol. The yield of ds-cDNA was estimated to be 15 fmol (22%).

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.

Abbreviations: kb, kilobase(s); bp, base pair(s); ds-cDNA, double-stranded cDNA.

* Permanent address: Dept. of Biochemistry, Louisiana State University, Baton Rouge, LA 70803.

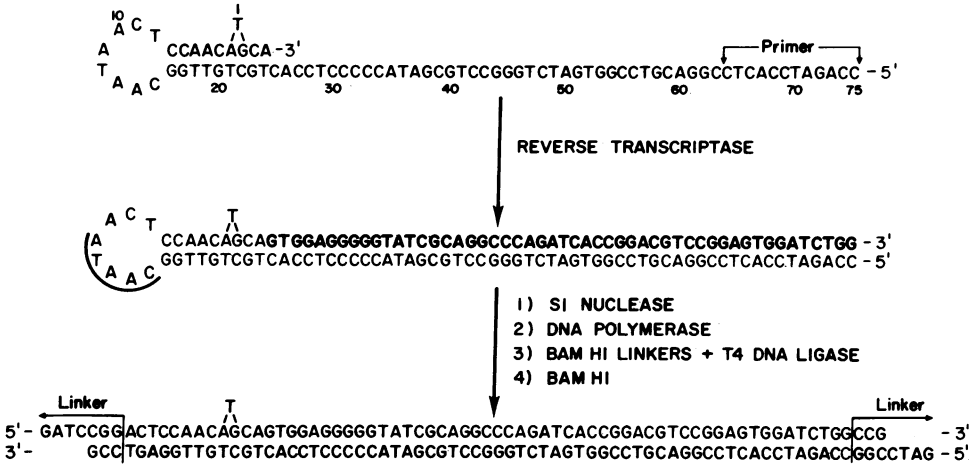

FIG. 1. Synthesis and cloning of double-stranded cDNA corresponding to the 5' terminus of the bacteriorhodopsin gene. Single-stranded cDNA was synthesized from partially purified mRNA as described (8). The second strand was synthesized by using avian myeloblastosis virus reverse transcriptase and the product was treated with S1 nuclease and DNA polymerase I before the addition of *Bam*HI linkers and ligation into the *Bam*HI site of pBR322.

(ii) *S1 nuclease treatment.* The ds-cDNA was incubated (25°C, 30 min) with S1 nuclease (0.75 unit) in 30 mM NaOAc (pH 4.5)/0.3 M NaCl/3 mM ZnOAc in a volume of 40 μl. After incubation, 10 mM EDTA and carrier tRNA (1 μg) were added and the mixture was extracted with phenol. Chromatography on Sephadex G-100 as described above yielded S1 nuclease-resistant ds-cDNA (10 fmol).

(iii) *Attachment of synthetic linkers.* The ds-cDNA was treated with DNA polymerase I and the flush-ended products were allowed to react with 4 pmol of phosphorylated d(C-C-G-G-A-T-C-C-G-G) and T4 DNA ligase essentially as published (12). The resulting mixture was treated with endonuclease *Bam*HI (1 unit) at 37°C for 1 hr before chromatography on a Bio-Gel A-1.5m column (1.0-ml pipette) in the Sephadex G-100 column buffer. The excluded material was precipitated with ethanol in the presence of 0.6 μg of plasmid pBR322 DNA that had previously been digested with *Bam*HI and treated with calf intestinal alkaline phosphatase (12).

(iv) *Cloning of the ds-cDNA fragment.* The ds-cDNA/pBR322 DNA mixture, containing 4 fmol of ds-cDNA, was dissolved in 20 μl of 50 mM Tris·HCl (pH 7.5)/10 mM $MgCl_2$/10 mM dithiothreitol/300 μM ATP. DNA ligase (0.5 unit) was added and the mixture was incubated at 14°C for 4 hr. The products were precipitated with ethanol and then incubated with calcium-shocked *E. coli* SF-8 (C600 rK^- mK^- $recBC^-$ *lop-11* lig^+) under transformation conditions (13). Recombinant colonies were selected for ampicillin (25 μg/ml) resistance and tetracycline (20 μg/ml) sensitivity. The ds-cDNA insert was inserted into the *Bam*HI site of M13mp7 and recloned (14). Digestion of the replicative form DNA from this phage with *Bam*HI, end-labeling of the resulting fragments with ^{32}P, and purification of the ds-cDNA insert [74 base pairs (bp)] on an 8% polyacrylamide gel yielded the probe for the colony hybridization.

Isolation of Bacteriorhodopsin Gene. *H. halobium* DNA (4 μg) was digested with *Pst* I and, separately, pBR322 DNA (1 μg) was also digested with *Pst* I and treated with calf intestinal alkaline phosphatase (12). Digests were mixed in 400 μl of 20 mM Tris·HCl (pH 7.6)/10 mM $MgCl_2$/10 mM dithiothreitol/0.25 mM ATP. T4 DNA ligase (1 unit) was added and the mixture was incubated at 4°C for various times. Portions (40 μl) were used to transform *E. coli* KH802 (SU_2^+ gal^- rK^- mK^+). Tetracycline-resistant colonies were screened by colony filter hybridization (15) with the ^{32}P-labeled ds-cDNA fragment described above.

DNA Sequence Analysis. DNA sequences were determined by the dideoxynucleotide chain termination procedures of Sanger and coworkers (16) on DNA cloned in phage M13mp7 (14). Additional sequences were determined by the chemical degradation methods of Maxam and Gilbert (17).

Bacteriorhodopsin mRNA Protection of 5'-^{32}P-Labeled DNA from S1 Nuclease Digestion. The 193-bp *Sau* 96 fragment, residues −148 to +45, was labeled with [γ-^{32}P]ATP and polynucleotide kinase and the strands were separated by denaturation in 30% (vol/vol) dimethyl sulfoxide followed by gel electrophoresis (17). The procedures used for DNA·RNA hybridization in the absence of formamide and subsequent S1 nuclease digestion were provided by U. Hansen and P. Sharp (personal communication). These are modifications of the procedures of Berk and Sharp (18). Single-stranded *Sau* 96 fragment (15 ng) and sucrose gradient-purified (8) bacteriorhodopsin mRNA (5 μg) were heated together at 70°C for 10 min in 25 μl of 50 mM Hepes (pH 7.4)/0.7 M NaCl/1 mM EDTA in the presence of 25 μg of yeast tRNA. The solution was then incubated at 52°C for 1 hr before dilution into 10 vol of ice-cold 30 mM NaOAc (pH 4.5)/0.3 M NaCl/3 mM ZnOAc. S1 nuclease (1000 units) was added and the reaction mixture was incubated at 37°C for 30 min. The products were precipitated with ethanol in the presence of 10 μg of carrier tRNA, extracted with phenol and reprecipitated with ethanol prior to analysis on a denaturing 8% polyacrylamide gel. In control experiments the mRNA was absent.

RESULTS

Cloning and Characterization of the Bacteriorhodopsin cDNA Fragment. ds-cDNA corresponding to the 5' end of bacteriorhodopsin mRNA was synthesized from single-stranded cDNA as described above. A 9-base imperfect inverted repeat at the 3' terminus of the single-stranded cDNA formed a hairpin loop that enabled the synthesis of the second strand as shown in Fig. 1. After addition of *Bam*HI linkers the ds-cDNA was inserted into the *Bam*HI site of pBR322. The sequence of the cDNA insert isolated from one such recombinant is shown in Fig. 1. The analysis showed the presence of two species of cDNA that differed only in the presence or absence of one additional A·T base pair at position 22. This is explained by the replication of both strands of the cDNA insert, one of which contained a mismatched T residue (Fig. 1).

Cloning and Characterization of the *Pst* I Fragment Containing the Bacteriorhodopsin Gene. The cDNA insert, labeled with ^{32}P, was used as the probe in colony filter hybridizations to cloned fragments of *H. halobium* DNA. A 5.3-kb *Pst* I DNA fragment was isolated that contained the bacteriorhodopsin gene. Digestion of this fragment with *Bam*HI and *Bst*EII yielded a 1.6-kb fragment that retained the entire gene. This was used for sequence determination. Some restriction sites found on this fragment, and the strategy for DNA sequence analysis is shown in Fig. 2. The DNA nucleotide sequence de-

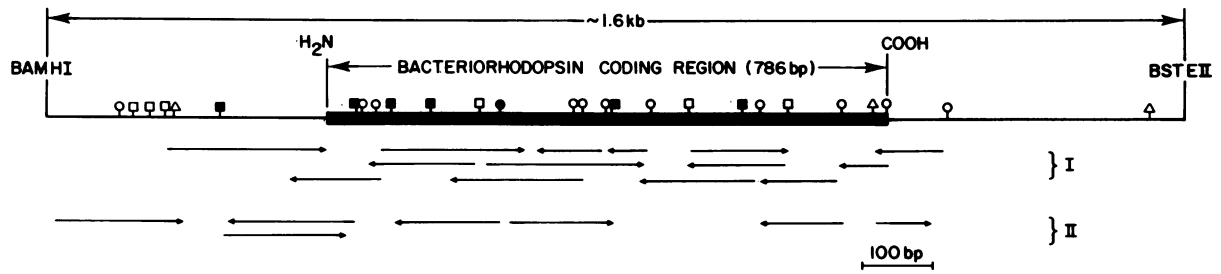

FIG. 2. Strategy for DNA sequence determination. The arrows indicate the direction and extent of DNA sequence determined by the dideoxynucleotide chain termination procedure (I) and the chemical degradation procedure (II). Restriction sites: ○, *Sau* 3A; ■, *Sau* 96; □, *Taq* I; ●, *Kpn* I; △, *Xma* I.

termined by a combination of the dideoxynucleotide chain termination procedure (16) and the chemical degradation procedure (17) is given in Fig. 3. The sequence includes the total structural gene, 360 nucleotides in the region upstream from the initiating AUG codon, and 80 nucleotides downstream from the termination codon.

S1 Nuclease Mapping of the mRNA 5′ Terminus. Previous work on the synthesis and characterization of the cDNA fragment corresponding to the 5′ terminus of bacteriorhodopsin mRNA indicated that the mRNA may extend only a few nucleotides beyond the AUG codon that initiates the synthesis of bacteriorhodopsin precursor (8). In order to determine more precisely the 5′ terminus of the mRNA, nuclease S1 protection experiments were performed. The separated strands of the *Sau* 96 fragment (193 bp, nucleotides −148 to +45) present at the 5′- terminal region of the bacteriorhodopsin gene were hybridized to bacteriorhodopsin mRNA and treated with nuclease S1. The results, which are described in Fig. 4, show strong protection up to the A residue corresponding to the U residue in the mRNA three nucleotides upstream from the initiator AUG. The protected fragment runs slower than the corresponding A band in the sequencing gel. This is because the S1 cleavage yields a 3′-hydroxyl group on the protected terminal A residue, whereas the chemical degradation in the sequencing reaction destroys this A residue and leaves a terminal 3′-phosphate group.

DISCUSSION

Structural Gene Sequence. We have reported on the cloning and sequence analysis of a *H. halobium* DNA fragment that includes the structural gene for bacteriorhodopsin and sequences flanking the NH_2 and COOH termini. The sequence of 248 amino acids for the mature protein derived from the DNA sequence (Fig. 3) is in complete agreement with the sequence that we obtained previously by using protein sequencing methods (2). It is recalled that there were differences in the assignments of six amino acids between our results and those of Ovchinnikov and coworkers (4). Three of the six differences could be due to strain-specific variations and can be accounted for by single nucleotide changes in the gene, but the other three require (*i*) the deletion of the three nucleotides corresponding to the Trp codon at amino acid position 138 and (*ii*) at least two nucleotide

```
GGGTGCAACCGTGAAGTCCGCCACGACCGCGTCACGACAGGAGCCGACCAGCGACACCCAGAAGGTGCGAACGGTTGAGTGCCGCAACGA
-360                                                                                    -271

TCACGAGTTTTTCGTGCGCTTCGAGTGGTAACACGCGTGCACGCATCGACTTCACCGCGGGTGTTTCGACGCCAGCCGGCCGTTGAACCA
-270                                                                                    -181

GCAGGCAGCGGGCATTTACAGCCGCTGTGGCCCAAATGGTGGGGTGCGCTATTTTGGTATGGTTTGGAATCCGCGTGTCGGCTCCGTGTC
-180                                                                                    -91

TGACGGTTCATCGGTTCTAAATTCCGTCACGAGCGTACCATACTGATTGGGTCGTAGAGTTACACACATATCCTCGTTAGGTACTGTTGC
-90                                                                                     -1

                                    ↓
Met Leu Glu Leu Leu Pro Thr Ala Val Glu Gly Val Ser Gln Ala Gln Ile Thr Gly Arg Pro Glu Trp Ile Trp Leu Ala Leu Gly Thr
ATGTTGGAGTTATTGCCAACAGCAGTGGAGGGGGTATCGCAGGCCCAGATCACCGGACGTCCGGAGTGGATCTGGCTAGCGCTCGGTACG
1                                                                                      90

Ala Leu Met Gly Leu Gly Thr Leu Tyr Phe Leu Val Lys Gly Met Gly Val Ser Asp Pro Asp Ala Lys Lys Phe Tyr Ala Ile Thr Thr
GCGCTAATGGGACTCGGGACGCTCTATTTCCTCGTGAAAGGGATGGGCGTCTCGGACCCAGATGCAAAGAAATTCTACGCCATCACGACG
91                                                                                     180

Leu Val Pro Ala Ile Ala Phe Thr Met Tyr Leu Ser Met Leu Leu Gly Tyr Gly Leu Thr Met Val Pro Phe Gly Gly Glu Gln Asn Pro
CTCGTCCCAGCCATCGCGTTCACGATGTACCTCTCGATGCTGCTGGGGTATGGCCTCACAATGGTACCGTTCGGTGGGGAGCAGAACCCC
181                                                                                    270

Ile Tyr Trp Ala Arg Tyr Ala Asp Trp Leu Phe Thr Thr Pro Leu Leu Leu Leu Asp Leu Ala Leu Leu Val Asp Ala Asp Gln Gly Thr
ATCTACTGGGCGCGGTACGCTGACTGGCTGTTCACCACGCCGCTGTTGTTGTTAGACCTCGCGTTGCTCGTTGACGCGGATCAGGGAACG
271                                                                                    360

Ile Leu Ala Leu Val Gly Ala Asp Gly Ile Met Ile Gly Thr Gly Leu Val Gly Ala Leu Thr Lys Val Tyr Ser Tyr Arg Phe Val Trp
ATCCTTGCGCTCGTCGGTGCCGACGGCATCATGATCGGGACCGGCCTGGTCGGCGCACTGACGAAGGTCTACTCGTACCGCTTCGTGTGG
361                                                                                    450

Trp Ala Ile Ser Thr Ala Ala Met Leu Tyr Ile Leu Tyr Val Leu Phe Phe Gly Phe Thr Ser Lys Ala Glu Ser Met Arg Pro Glu Val
TGGGCGATCAGCACCGCAGCGATGCTGTACATCCTGTACGTGCTGTTCTTCGGGTTCACCTCGAAGGCCGAAAGCATGCGCCCCGAGGTC
451                                                                                    540

Ala Ser Thr Phe Lys Val Leu Arg Asn Val Thr Val Val Leu Trp Ser Ala Tyr Pro Val Val Trp Leu Ile Gly Ser Glu Gly Ala Gly
GCATCCACGTTCAAAGTACTGCGTAACGTTACCGTTGTGTTGTGGTCCGCGTATCCCGTCGTGTGGCTGATCGGCAGCGAAGGTGCGGGA
541                                                                                    630

Ile Val Pro Leu Asn Ile Glu Thr Leu Leu Phe Met Val Leu Asp Val Ser Ala Lys Val Gly Phe Gly Leu Ile Leu Leu Arg Ser Arg
ATCGTGCCGCTGAACATCGAGACGCTGCTGTTCATGGTGCTTGACGTGAGCGCGAAGGTCGGCTTCGGGCTCATCCTCCTGCGCAGTCGT
631                                                                                    720

                                    ↓
Ala Ile Phe Gly Glu Ala Glu Ala Pro Glu Pro Ser Ala Gly Asp Gly Ala Ala Ala Thr Ser Asp ***
GCGATCTTCGGCGAAGCCGAAGCGCCGGAGCCGTCCGCCGGCGACGGCGCGGCCGCGACCAGCGACTGATCGCACACGCAGGACAGCCCC
721                                                                                    810

ACAACCGGCGCGGCTGTGTTCAACGACACACGATGAGTCCCCCACTCGGTCTTGTACTC
811                                                            869
```

FIG. 3. DNA sequence of the bacteriorhodopsin gene and adjacent regions. The amino acid sequence of the bacteriorhodopsin precursor is given; the NH_2-terminal methionine codon is at nucleotide residue 1. The arrows indicate the NH_2 and COOH termini of the mature protein.

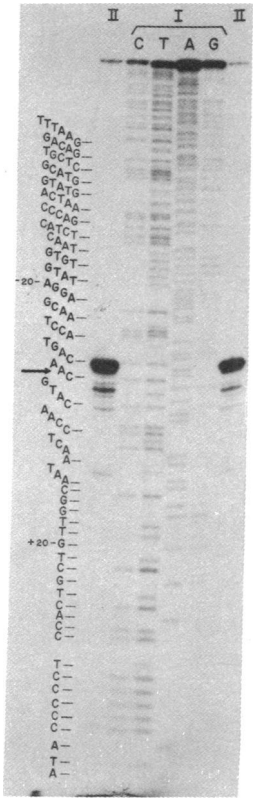

FIG. 4. Protection by bacteriorhodopsin mRNA of 5′-^{32}P-labeled single-stranded DNA from S1 nuclease digestion. The single-stranded *Sau* 96 fragment corresponding to nucleotides −148 to +45 labeled at the 5′ terminus was hybridized to mRNA and treated with S1 nuclease (outside lanes). The internal lanes are chemical degradation sequencing ladders from the *Sau* 96 fragment.

changes in each of the codons corresponding to amino acids 146 and 206 in bacteriorhodopsin.

The gene sequence (Fig. 3) encodes 13 extra amino acids at the NH_2 terminus that are not present in mature bacteriorhodopsin. This result confirms our previous conclusion, which was derived from the sequence of a cDNA fragment prepared *in vitro* from bacteriorhodopsin mRNA (8). The 13 amino acids must compose the total precursor, because there is no other ATG codon upstream from the NH_2-terminal methionine codon (nucleotides 1–3, Fig. 3) before an in-phase TAG termination codon (nucleotides −34 to −36).

The gene sequence also encodes an additional amino acid, aspartic acid, at the COOH terminus of the precursor. This finding has precedents in that α-tubulin has been reported to contain an extra tyrosine residue (19, 20) and the heavy chains of immunoglobulins can contain an extra lysine at their COOH termini (21).

The nucleotide sequence (64–75) within the gene explains our previous observation that, of the two dodecanucleotides tested as primers for cDNA synthesis, d(C-C-A-G-A-T-C-C-A-C-T-C) was efficient and specific, whereas d(C-C-A-G-A-T-C-C-A-T-T-C) was less efficient and gave multiple nonspecific elongation products (8). It is clear from the gene sequence that the first primer can form a perfect 12-bp match with bacteriorhodopsin mRNA, whereas the second primer has a G–T mismatch close to its 3′ end.

A comparison between the structural gene sequence and the amino acid sequence shows that the normal genetic code is used in *Halobacteria*. However, the present work provides no evidence for the assignments for His, Cys, and the AGA and AGG codons for arginine (Fig. 5). This finding, in agreement with previous studies (22), is of interest in that *Halobacteria* belong to a class of organisms that have been termed "archaebacteria" by Woese and coworkers (23) and that are proposed to constitute a third major line of descent besides eubacteria (previously prokaryotes) and urkaryotes (previously eukaryotes). There is a strong preference (about 82%) in the bacteriorhodopsin gene for the use of codons ending in G or C (Fig. 5). However, despite this bias, the overall G+C composition of the bacteriorhodopsin coding sequence (61.5%) is still lower than the G+C content (24) of *H. halobium* DNA (67%). Thus, the preferential use of codons ending in G or C could be more a result of the high G+C content of the DNA of the organism rather than a reflection of selective codon usage for optimizing translation of bacteriorhodopsin mRNA.

Bacteriorhodopsin mRNA. Bacteriorhodopsin mRNA contains only a few nucleotides at the 5′ terminus preceding the initiation codon. This was shown earlier by the sequence of the cDNA fragment (8) and has now been confirmed by the nuclease S1 mapping data shown in Fig. 4. Similarly, short 5′ termini have been found in human mitochondrial mRNAs (25).

Bacteriorhodopsin mRNA is significantly different from the prokaryotic mRNAs, which usually contain 5′ leader sequences of 26 or more nucleotides (26). An important feature of these 5′ leader sequences in prokaryotic mRNAs is a purine-rich region approximately 6–12 nucleotides upstream from the AUG codon that is presumed to interact with a complementary pyrimidine-rich sequence at the 3′ end of prokaryotic 16S rRNA during initiation of protein synthesis (27). Although the 3′ terminus of halobacterial 16S rRNA has a pyrimidine-rich sequence almost identical to that of *E. coli* (28), the type of interaction postulated to occur between leader sequences of prokaryotic mRNAs and 16S rRNA may not occur during bacteriorhodopsin synthesis. It is worth noting, however, that bacteriorhodopsin mRNA does contain the purine-rich sequences G-G-A-G (nucleotides 6–9) and G-G-A-G-G (nucleotides 27–31) which are complementary to the C-C-U-C-C sequence near the 3′ end of halobacterial 16S rRNA. However, these sequences are downstream from the initiating AUG codon (Fig. 6*A*).

Another interesting feature in the sequence at the 5′ end of bacteriorhodopsin mRNA is an interrupted inverted repeat sequence 9 bp in length, which could fold to form a stable hairpin loop structure (Fig. 6*B*). Such a structure has been found in several other mRNAs, including the *nifH* mRNA of *Klebsiella pneumoniae* (29), ovalbumin mRNA (30), and satellite tobacco necrosis virus RNA (31). Whether this feature plays a role in the functioning or stabilization of the mRNAs is not known.

Transcription Signals? No information is available regarding the promoters or other signals necessary for the specific initiation of transcription in *Halobacteria*. Consequently we cannot at present identify the bacteriorhodopsin promoter. There is no DNA sequence in the region proximal to the 5′ end of the

UUU	Phe 0	UCU	0	UAU	Tyr 3	UGU	Cys 0
UUC	13	UCC	Ser 3	UAC	8	UGC	0
UUA	Leu 2	UCA	0	UAA	Ter 0	UGA	Ter 1
UUG	6	UCG	5	UAG	0	UGG	Trp 8
CUU	2	CCU	0	CAU	His 0	CGU	3
CUC	Leu 12	CCC	Pro 3	CAC	0	CGC	Arg 3
CUA	2	CCA	3	CAA	Gln 0	CGA	0
CUG	15	CCG	6	CAG	4	CGG	1
AUU	0	ACU	0	AAU	Asn 0	AGU	Ser 1
AUC	Ile 15	ACC	Thr 7	AAC	3	AGC	5
AUA	0	ACA	2	AAA	Lys 3	AGA	Arg 0
AUG	Met 10	ACG	10	AAG	4	AGG	0
GUU	3	GCU	1	GAU	Asp 2	GGU	4
GUC	Val 8	GCC	Ala 8	GAC	8	GGC	Gly 10
GUA	3	GCA	5	GAA	Glu 4	GGA	4
GUG	9	GCG	16	GAG	7	GGG	8

FIG. 5. Frequency of codon usage in the bacteriorhodopsin gene. Ter, termination.

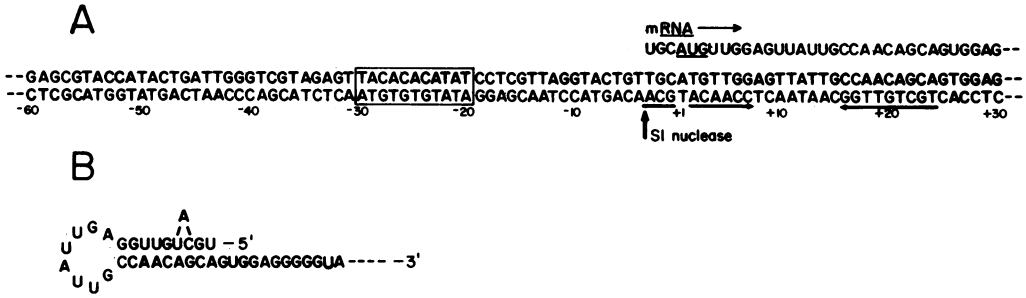

FIG. 6. Sequence of DNA corresponding to the 5'-terminal region of bacteriorhodopsin mRNA. The initiating AUG codon is underlined in the mRNA. Two horizontal arrows show the inverted repeat structure. The vertical arrow indicates the A residue protected from S1 nuclease by mRNA (Fig. 4). The boxed region is an 11-bp sequence of alternating purine-pyrimidine residues. (*B*) The hairpin structure at the 5' terminus of bacteriorhodopsin mRNA.

mRNA that is reminiscent of a prokaryotic promoter sequence. However, if the 5' terminus of the mRNA as isolated is produced by processing of a longer transcript the promoter would be expected to be further upstream.

As discussed above, the DNA corresponding to the 5' terminus of the bacteriorhodopsin mRNA contains an inverted repeat sequence. Such a structure is also present in the *E. coli* lactose operon, in which the *lac* operator DNA has been shown to contain an inverted repeat sequence at the mRNA start site (32, 33). Two other examples of such structures are the λ phage operators O_L and O_R (34). In view of these homologies, one possible function of this sequence in the bacteriorhodopsin gene could be to serve as the binding site for a regulatory protein.

Other features of the DNA sequence may play a significant role in transcription or other processes. There is a 10-nucleotide-long sequence of alternating pyrimidines and purines centered 20 nucleotides upstream from the 5' end of the mRNA (Fig. 6A). In addition, there are two other regions (−225 to −240 and +791 to +800) outside the coding sequences (Fig. 3) that contain similar alternating sequences. It is interesting to note that the salt concentration within *H. halobium* is very high (4 M KCl) and that under such ionic conditions DNA sequences that contain alternating purines and pyrimidines may adopt a Z-DNA structure (35, 36).

This investigation was supported by Grants GM28289 and AI11479 from the National Institutes of Health, and Grant PCM78-13713 from the National Science Foundation.

1. Stoeckenius, W., Lozier, R. H. & Bogomolni, R. A. (1979) *Biochim. Biophys. Acta* **505**, 215–278.
2. Khorana, H. G., Gerber, G. E., Herlihy, W. C., Gray, C. P., Anderegg, R. J., Nihei, K. & Biemann, K. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 5046–5050.
3. Bayley, H., Huang, K.-S., Radhakrishnan, R., Ross, A. H., Takagaki, Y. & Khorana, H. G. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 2225–2229.
4. Ovchinnikov, Y. A., Abdulaev, N. G., Feigina, M. Y., Kiselev, A. B. & Lobanov, N. A. (1979) *FEBS Lett.* **100**, 219–224.
5. Engelman, D. M., Henderson, R., McLachlan, A. D. & Wallace, B. A. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 2023–2027.
6. Henderson, R. & Unwin, P. N. T. (1975) *Nature (London)* **257**, 28–32.
7. Engelman, D. M. & Zaccai, G. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 5894–5898.
8. Chang, S. H., Majumdar, A., Dunn, R., Makabe, O., RajBhandary, U. L., Khorana, H. G., Ohtsuka, E., Tanaka, T., Taniyama, Y. O. & Ikehara, M. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 3398–3402.
9. Schendel, P. F. & Wells, R. D. (1973) *J. Biol. Chem.* **248**, 8319–8321.
10. Johnson, R. A. & Walseth, F. (1979) *Adv. Cyclic Nucleotide Res.* **10**, 135–167.
11. Panet, A., van de Sande, J. H., Loewen, P. C., Khorana, H. G., Raae, A. J., Lillehaug, J. R. & Kleppe, K. (1973) *Biochemistry* **12**, 5045–5050.
12. Goodman, H. M. & MacDonald, R. J. (1979) *Methods Enzymol.* **68**, 75–89.
13. Mandel, M. & Higa, A. (1970) *J. Mol. Biol.* **53**, 159–162.
14. Messing, J., Crea, R. & Seeburg, P. H. (1981) *Nucleic Acids Res.* **9**, 309–323.
15. Grunstein, M. & Hogness, D. S. (1975) *Proc. Natl. Acad. Sci. USA* **72**, 3961–3965.
16. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467.
17. Maxam, A. M. & Gilbert, W. (1980) *Methods Enzymol.* **65**, 499–560.
18. Berk, A. J. & Sharp, P. A. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 1274–1278.
19. Valenzuela, P., Quiroga, M., Zaldivar, J., Rutter, J. W., Kirschner, M. W. & Cleveland, D. W. (1981) *Nature (London)* **289**, 650–655.
20. Ginzburg, I., Behar, L., Givol, D. & Littauer, U. Z. (1981) *Nucleic Acids Res.* **9**, 2691–2697.
21. Tucker, P. W., Marcu, K. B., Slightom, J. L. & Blattner, F. R. (1979) *Science* **206**, 1299–1303.
22. White, B. M. & Bayley, S. T. (1972) *Can. J. Biochem.* **50**, 600–609.
23. Fox, G. E., Stackebrandt, E., Hespell, R. B., Gibson, J., Maniloff, J., Dyer, T. A., Wolfe, R. S., Balch, W. E., Tanner, R. S., Magrum, L. J., Zablen, L. B., Blakemore, R., Gupta, R., Bonen, L., Lewis, B. J., Stahl, D. A., Luehrsen, K. R., Chen, K. N. & Woese, C. R. (1980) *Science* **209**, 457–463.
24. Moore, R. L. & McCarthy B. J. (1969) *J. Bacteriol.* **99**, 248–254.
25. Montoya, J., Ojala, D. & Attardi, G. (1981) *Nature (London)* **290**, 465–470.
26. Steitz, J. A. (1979) in *Biological Regulation and Development*, ed. Goldberger, R. F. (Plenum, New York), Vol. 1, pp. 349–400.
27. Shine, J. & Dalgarno, L. (1974) *Proc. Natl. Acad. Sci. USA* **71**, 1342–1346.
28. Magrum, L. J., Luehrsen, K. R. & Woese, C. R. (1978) *J. Mol. Evol.* **11**, 1–8.
29. Scott, K. F., Rolfe, B. G. & Shine, J. (1981) *J. Mol. Appl. Genet.* **1**, 3–18.
30. McReynolds, L., O'Malley, B. W., Nisbet, A. D., Fothergill, J. E., Fields, S., Robertson, M. & Brownlee, G. G. (1978) *Nature (London)* **273**, 723–728.
31. Leung, D. W., Browning, K. S., Heckman, J. E., RajBhandary, U. L., & Clark, J. M. (1979) *Biochemistry* **18**, 1361–1366.
32. Gilbert, W. & Maxam, A. (1973) *Proc. Natl. Acad. Sci. USA* **70**, 3581–3585.
33. Dickson, R. C., Abelson, J., Barnes, W. M. & Reznikoff, W. S. (1975) *Science* **187**, 27–35.
34. Ptashne, M., Backman, K., Humayun, M. Z., Jeffrey, A., Maurer, R., Meyer, B. & Sauer, R. T. (1976) *Science* **194**, 156–161.
35. Wang, A. H., Quigley, G. J., Kolpak, F. J., Crawford, J. L., van Boom, J. H., van der Marel, G. & Rich, A. (1979) *Nature (London)* **282**, 680–686.
36. Arnott, S., Chandrasekaran, R., Birdsall, D. L., Leslie, A. G. W. & Ratliff, R. L. (1980) *Nature (London)* **283**, 743–748.