J Mol Evol (1985) 22:351–360

Journal of
Molecular Evolution

Sequence Divergence of an Archaebacterial Gene Cloned from a Mesophilic and a Thermophilic Methanogen

Paul T. Hamilton and John N. Reeve

Department of Microbiology, The Ohio State University, Columbus, Ohio 43210, USA

Summary. A 1.6-kb fragment of DNA from the thermophilic, methane-producing, anaerobic archaebacterium *Methanobacterium thermoautotrophicum* ΔH has been cloned and sequenced. This DNA complements mutations in both the $purE_1$ and $purE_2$ loci of *Escherichia coli*. The sequence of the *M. thermoautotrophicum* DNA predicts that complementation in *E. coli* results from the synthesis of a polypeptide with a molecular weight of 36,249. A polypeptide apparently of this molecular weight is synthesized in *E. coli* minicells containing recombinant plasmids that carry the cloned fragment of methanogen DNA. We have previously cloned and sequenced a purE-complementing gene from the mesophilic methanogen *Methanobrevibacter smithii*. The two methanogen-derived purE-complementing genes are 53% homologous and encode polypeptides that are 45% homologous in their amino acid sequences but would be 74% homologous if conservative amino acid substitutions were considered as maintaining sequence homology. The genome of *M. thermoautotrophicum* has a molar G + C content of 49.7%, whereas the genome of *M. smithii* is 30.6% G + C. Conservation of encoded amino acids while accommodating the very different G + C contents is accomplished by use of different codons that encode the same amino acid. The majority of base changes occur at the third codon position. The intergenic regions of the cloned *M. thermoautotrophicum* DNA contain sequences previously identified as ribosome binding sites and as putative methanogen promoters. Although the two purE-complementing genes are apparently derived from a common ancestor, only the gene from *M. smithii* maintains a codon usage that conforms to the RNY rule.

Key words: Methanogens — Archaebacteria — purE Complementation — DNA sequences — Divergence — Codon usage

Offprint requests to: J.N. Reeve

Introduction

Although all methanogens are procaryotic, are strict anaerobes, and gain energy by biosynthesis of methane, they are otherwise a very heterogeneous group (Balch et al. 1979). Molar DNA base compositions of methanogens range from 28 to 61% guanine plus cytosine (G + C). There are coccoid, rod-shaped, filamentous, and spiral methanogens. The current taxonomy of methanogens was developed on the basis of comparative studies of catalogues of oligonucleotides obtained by ribonuclease T1 digestion of 16S rRNAs. This taxonomy, while recognizing their diversity, maintains methanogens as a coherent group and places them with the extreme halophiles and thermoacidophiles as members of the archaebacterial kingdom (Fox et al. 1977; Woese and Fox 1977; Woese et al. 1978). We have cloned genes from several methanogens on the basis of the cloned DNAs being able to complement auxotrophic mutations of *Escherichia coli* or *Bacillus subtilis* (Reeve et al. 1982; Hamilton and Reeve 1984, 1985; Morris and Reeve 1984; Cue et al. 1985). By comparison of the sequences of genes cloned from different methanogens it should be possible to determine evolutionary and therefore taxonomic relationships among different methanogenic species.

Table 1. Microorganisms and plasmids used

Bacterial strains/plasmids	Genotype/phenotype	Source/reference
Strains		
Methanobacterium thermoautotrophicum ΔH	Wild type	D. Livingston, Dept. Chemistry, M.I.T.
Escherichia coli JA221	hsdR, trpΔE5, leuB, recA	Reeve (1979)
Escherichia coli DS410	minA, minB, λ^s	R. Curtiss III
Escherichia coli χ760	ara-1, leu-1, azi^r, tonA^r, lacY2, proC119, tsx, purE1, galK2, trp3, his4, argG36, rpsL, xyl-1, mtl-1, ilvA6, thi-1, met12	
Escherichia coli TX209	$purE_2$, Δlac	J. Gots, Dept. Microbiology, Univ. Pennsylvania
Escherichia coli TX257	$purE_1$, Δlac	J. Gots
Escherichia coli NK6051	$purE_1$::Tn10, Δlac-pro (phenotypically $purE_2$)	
Plasmids		
pUC8	Amp^r	Vieira and Messing (1982)
pET405	Amp^r, purE$^+$	Hamilton and Reeve (1985)
pET441	Amp^r, purE$^+$	1.6-kb PstI fragment of *M. thermoautotrophicum* DNA cloned into pUC8
pET445	Amp^r, purE$^-$	pET441 digested with Asp718 and religated

This approach was recently used to compare genes from two methanococcal species, *Methanococcus vannielii* and *Methanococcus voltae* (Cue et al. 1985), which have very similar DNA base compositions and morphologies (Balch et al. 1979). These species differ in that *M. voltae* is a marine isolate and requires NaCl, isoleucine, and valine for growth. Both are mesophiles. The sequence data confirm that these two species are related, but perhaps not so closely related as suggested in the published taxonomy of methanogens (Balch et al. 1979).

We have now extended this approach to two methanogens that, although placed in the same taxonomic family (Balch et al. 1979), have very different characteristics. *Methanobrevibacter smithii* is a mesophile with a G + C content of 30.6 mol%, whereas *Methanobacterium thermoautotrophicum* is a thermophile with a G + C content of 49.7 mol% (Balch et al. 1979). DNA fragments from *M. smithii* and *M. thermoautotrophicum* have been cloned that complement mutations at the purE locus of *E. coli* (Gots et al. 1977). Details of the cloning and sequencing of the *M. smithii* DNA have been presented (Hamilton and Reeve 1985). In this report we provide the *M. thermoautotrophicum* sequence and the results of comparisons of the two DNA sequences and the two encoded gene products.

Materials and Methods

Bacteria and Plasmids. Strains and plasmids used in this study are listed in Table 1. Plasmid pET441 was constructed by ligation of PstI-digested *M. thermoautotrophicum* ΔH DNA to PstI-digested pUC8. The ligation mixture was used to transform competent cells of *E. coli* JA221 and ampicillin-resistant transformants were selected. Plasmids were isolated from this population of transformants and used to transform *E. coli* χ760, and transformants were selected for purine-independent, ampicillin-resistant growth. One such isolate contained plasmid pET441. A Southern blot hybridization procedure, with ^{32}P-labeled pET441 as the probe against PstI-digested genomic *M. thermoautotrophicum* DNA, confirmed that the cloned DNA originated in *M. thermoautotrophicum* ΔH (results not shown). Plasmid pET445 was obtained by Asp718 (Boehringer Mannheim Biochemicals, Indianapolis, IN 46250) digestion of pET441 DNA, religation, and transformation of *E. coli* TX257 with selection for ampicillin-resistant transformants (Fig. 1). *E. coli* TX257 strains carrying plasmid pET445 retain the purine auxotrophy of *E. coli* TX257.

Media and Procedures. Media and facilities used to grow methanogens have been described by Hook et al. (1984). Media and culture conditions for *E. coli* strains were as described by Davis et al. (1980). References for or descriptions of the techniques used to isolate, clone, digest, and sequence DNAs from methanogens have been given previously (Hamilton and Reeve 1984, 1985). Plasmids of interest were transformed into *E. coli* DS410 and minicells produced by these transformants allowed to incorporate L-[^{35}S]-methionine. The techniques used to isolate minicells, label plasmid-encoded polypeptides in minicells, and characterize the labeled polypeptides by polyacrylamide gel electrophoresis and fluorography have been described in detail (Reeve 1979).

Results

Cloning of M. smithii and M. thermoautotrophicum Sequences That Complement purE Mutations of E. coli

Details of the cloning and sequencing of a 2.7 kilobase pair (kb) fragment of *M. smithii* DNA that complements *E. coli* $purE_1$ and $purE_2$ mutations have been published (Hamilton and Reeve 1984, 1985). A 1.6-kb PstI-generated fragment of *M. thermoautotrophicum* DNA was cloned into pUC8

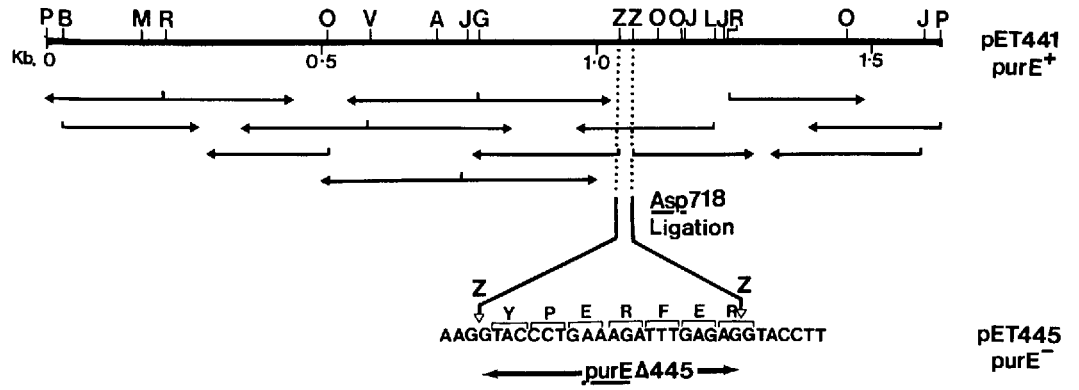

Fig. 1. Restriction map and strategy used to sequence the *M. thermoautotrophicum* DNA cloned in plasmid pET441. The heavy upper line represents the 1.6 kb of cloned DNA; restriction sites are AccI (A), BamHI (B), BglII (G), HpaII (J), BclI (L), CfoI (M), HaeIII (O), PstI (P), RsaI (R), EcoRV (V), and Asp718 (Z). The arrows below the restriction map indicate the extents of individually determined sequences obtained using the chemical cleavage method of Maxam and Gilbert (1980). A deletion of the indicated 21 bp was introduced into pET441, producing pET445, by digestion with Asp718 (which recognizes and cleaves 5'-G|GTACC), religation, and transformation of *E. coli* TX257. The codons deleted in purEΔ445 and the encoded amino acids whose removal inactivates purE complementation are shown

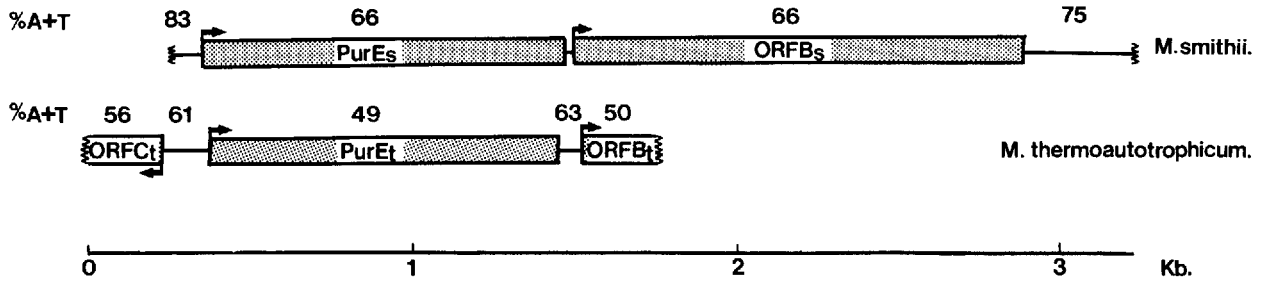

Fig. 2. Schematic representation of the cloned purE-complementing DNA fragments. The boxes represent open reading frames (ORFs). Identification of the purE-complementing ORF in the cloned *M. smithii* DNA (ORF-PurE$_s$) was previously reported (Hamilton and Reeve 1985). Identification of the purE-complementing ORF (ORF-PurE$_t$) from *M. thermoautotrophicum* is based on its sequence homology to ORF-PurE$_s$ and on the location of the deletion in pET445 that inactivates purE complementation (see Figs. 1 and 3). ORF-PurE$_t$ is the only intact ORF in the fragment of DNA cloned from *M. thermoautotrophicum*, capable of encoding a polypeptide containing more than 80 amino acid residues. The figures given above the ORFs and intergenic regions are the molar percentages of A + T in the indicated regions. In both DNAs, ORF-PurE and ORF-B are transcribed as indicated by the arrows from left to right, whereas ORF-C$_t$ would be transcribed from right to left

(Vieira and Messing 1982) to obtain plasmid pET441, which also complements the purE$_1$ and purE$_2$ mutations (Gots et al. 1977) present in the strains of *E. coli* listed in Table 1. Figure 1 shows a restriction map of the *M. thermoautotrophicum* DNA cloned in pET441 and the strategy used to obtain the sequence of this cloned DNA. The sequence obtained indicated the presence of two cleavage sites for the restriction enzyme Asp718 (an isoschizomer of KpnI) separated by only 21 bp. Digestion of pET441 with Asp718, religation, and transformation of *E. coli* TX257 produced a plasmid, pET445, that, although still capable of conferring ampicillin-resistant growth on *E. coli* TX257, was incapable of complementing the purE$_1$ mutation carried by *E. coli* TX257. Plasmid pET445 was subsequently also found to be incapable of complementing the mutation in purE$_2$ in *E. coli* TX209 (Table 1). DNA sequencing confirmed that construction of pET445 had resulted in a deletion of the 21 bp between the two Asp718 sites of pET441 (Fig. 1). This deletion (purEΔ445) removes seven in-frame codons from within the open reading frame (ORF) identified as the purE-complementing gene (see below). Loss of seven amino acids from within this *M. thermoautotrophicum*-encoded polypeptide apparently inactivates its ability to complement purE mutations in *E. coli*.

Figure 2 shows the overall organization of ORFs in the cloned DNAs from *M. smithii* and *M. thermoautotrophicum*. As previously reported, the *M. smithii* DNA contains two long ORFs (ORF-PurE$_s$ and ORF-B$_s$ in Fig. 2), separated by only 9 bp, that appear to be in the same transcriptional unit (Hamilton and Reeve 1985). The cloned *M. thermoautotrophicum* DNA contains one long, intact ORF (ORF-PurE$_t$ in Fig. 2) and the amino termini of two additional ORFs (ORF-B$_t$ and ORF-C$_t$ in Fig. 2).

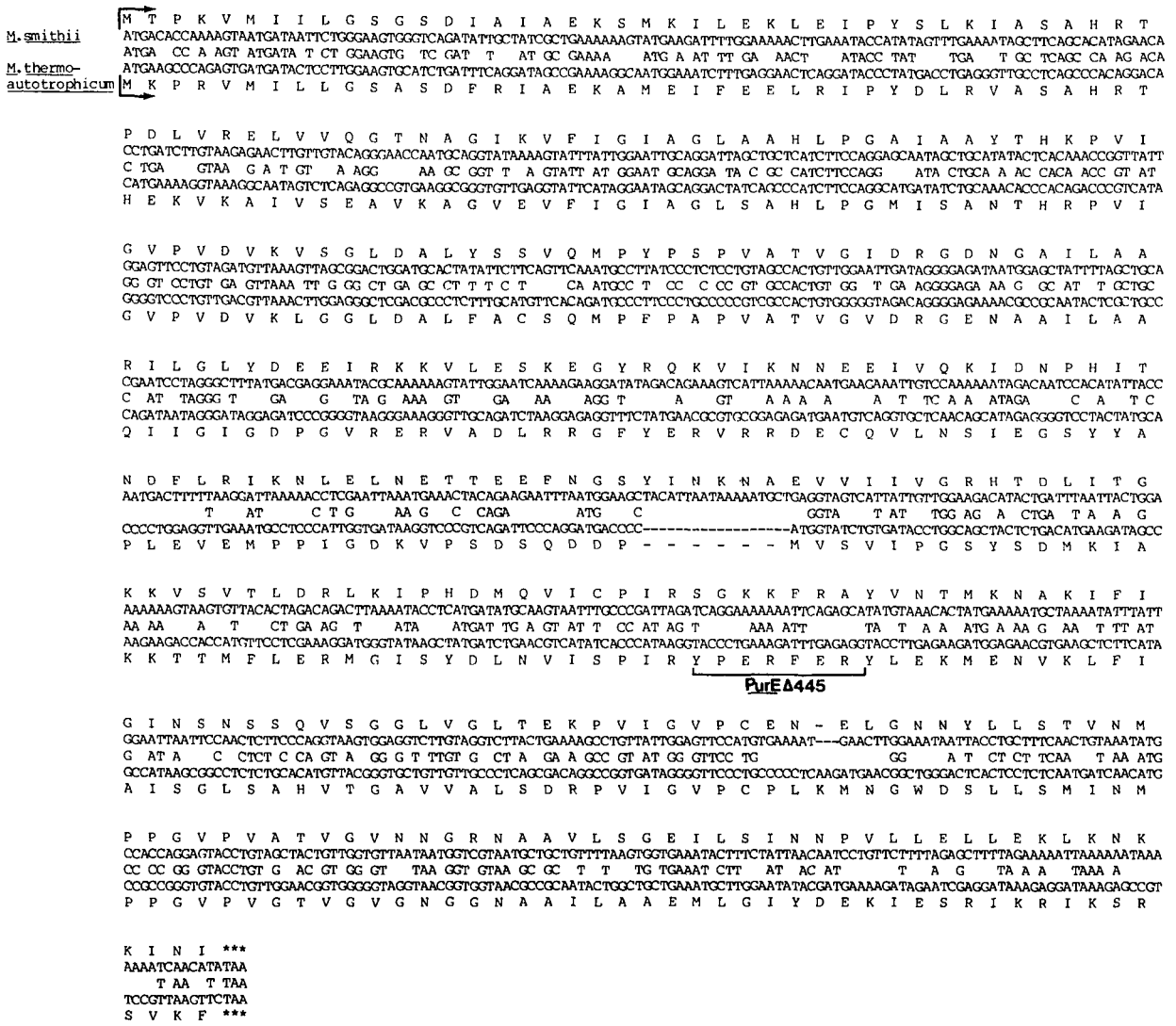

Fig. 3. DNA and amino acid sequences of the purE-complementing genes (ORF-PurE) and their encoded polypeptides. The upper, *M. smithii* sequence was reported previously (Hamilton and Reeve 1985). The *M. thermoautotrophicum* sequence was obtained using the chemical cleavage method of Maxam and Gilbert (1980). The sequencing strategy is shown in Fig. 1; all regions were sequenced at least twice, although the sequence of some regions was obtained from only one DNA strand. Positions at which the same base occurs in both sequences have been indicated by typing that base between the two DNA sequences. The amino acid sequences are indicated using the single-letter amino acid code above the first base of each amino acid-encoding codon. The gaps (indicated by dashes) were inserted by a homology-detecting and - maximizing program (ALIGN) purchased from DNASTAR (Madison, WI 53711). The location of the deletion, purEΔ445, present in pET445 is indicated

Construction of purEΔ445 deleted seven codons within ORF-PurE$_t$ (Fig. 1). ORF-B$_t$ is read in the same direction as ORF-PurE$_t$, and comparison of sequences (see below and Fig. 5) shows it clearly to be related to ORF-B$_s$. ORF-C$_t$ is read in the opposite direction from ORF-PurE$_t$ and is located in a region of DNA that has not been cloned from *M. smithii.*

Comparison of DNA Sequences

The DNA sequences demonstrate that the base compositions of the cloned purE genes correspond closely to the overall base compositions of the genomes of the two species. *M. smithii* genomic DNA is 69.4% A + T and ORF-PurE$_s$ of *M. smithii* is 66.2% A + T. *M. thermoautotrophicum* genomic DNA is 50.3% A + T and ORF-PurE$_t$ of *M. thermoautotrophicum* is 48.9% A + T. In both species the intergenic regions have significantly higher A + T contents than are found in the ORFs (Fig. 2).

Figure 3 is a comparison of the two methanogen-derived ORF-PurE DNA sequences and of the two encoded amino acid sequences. The two genes appear to have evolved from a common ancestor. ORF-PurE$_s$ is 1020 bp and ORF-PurE$_t$ is 1005 bp, predicting encoded polypeptides with molecular weights of 36,697 and 36,249, respectively. Polypeptides synthesized in minicells containing plasmid pET405 (which contains ORF-PurE$_s$ and ORF-B$_s$), pET441 (which contains ORF-PurE$_t$), or pET445 (which

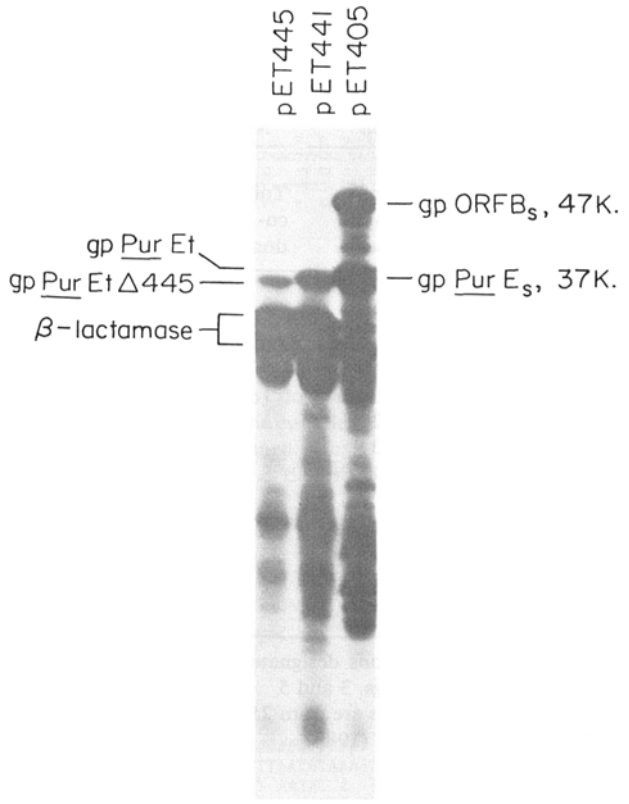

Fig. 4. Expression of purE-complementing plasmids in minicells of *E. coli*. Minicells containing pET405 (Hamilton and Reeve 1985), pET441, or pET445 were allowed to incorporate [^{35}S] methionine for 10 min at 37°C. Radioactively labeled polypeptides were separated by electrophoresis through a 10–20% gradient polyacrylamide gel. The locations of the radioactive polypeptides were detected by fluorography (Reeve 1979). The gene products of ORF-PurE$_s$ and ORF-B$_s$ (gpPurE$_s$ and gpORFB$_s$, respectively) encoded by pET405 have been previously identified. Their mobilities accurately reflect their molecular weights as calculated from the encoding DNA sequences and as demonstrated by coelectrophoresis with polypeptides of known molecular weights (Hamilton and Reeve 1985). The products of ORF-PurE$_t$ and ORF-PurE$_t$Δ445 (gpPurE$_t$ and gpPurE$_t$Δ445, respectively) encoded by pET441 and pET445 and the precursor and mature forms of β-lactamase encoded by pET405, pET441, and pET445 are indicated to the left of the figure. The additional polypeptides synthesized in minicells containing pET405 as compared with minicells containing pET441 or pET445 are products of pBR322 (the vector used in construction of pET405) not encoded by pUC8 (the vector used in construction of pET441 and pET445)

contains ORF-PurE$_t$Δ445) are compared in Fig. 4. The ORF-PurE$_s$ gene product (gpPurE$_s$) migrates slightly slower during electrophoresis than the ORF-PurE$_t$ gene product (gpPurE$_t$), which, in turn, migrates slightly slower than the ORF-PurE$_t$Δ445 gene product (gpPurE$_t$Δ445). This is consistent with the calculated molecular weights for the three gene products of 36,697, 36,249, and 35,273, respectively. The sequence alignment shown in Fig. 3 places the same base as in ORF-PurE$_s$ at 535 positions of the 1005-bp sequence of ORF-PurE$_t$, and 151 of the

Table 2. Base pair changes found in comparing ORF-purE$_s$ of *M. smithii* with ORF-purE$_t$ of *M. thermoautotrophicum*[a]

Base changes[b]	Codon position 1	2	3	Transition (Ts) or transversion (Tv)
A → T	10	9	24	Tv
A → C	13	10	36	Tv
A → G	30	24	38	Ts
T → A	4	9	31	Tv
T → C	6	6	47	Ts
T → G	11	2	28	Tv
C → G	5	8	7	Tv
C → T	4	7	7	Ts
C → A	13	5	8	Tv
G → A	13	7	5	Ts
G → T	8	0	3	Tv
G → C	9	15	5	Tv
Totals	126	102	239	Tv = 273 Ts = 194

[a] Sequences and alignment are given in Fig. 3
[b] Changes are given in the direction *M. smithii* → *M. thermoautotrophicum*

335 encoded amino acids of gpPurE$_t$ are found, as shown, in the same location in gpPurE$_s$. If amino acid substitutions that maintain charge, polarity, hydrophobicity, and approximate size are considered homologous, then the two polypeptides are 74% homologous in their amino acid sequences. There are base changes at 27%, 22%, and 51% of the first, second, and third positions in codons, respectively. Table 2 gives a detailed analysis of the base changes. It is clear that, as predicted from the overall A + T contents of the two genes, the majority of changes are A or T to G or C when evaluated in the direction of changes from the *M. smithii* sequence to the *M. thermoautotrophicum* sequence.

A summation of codon usages in the two ORF-PurE and ORF-B coding regions is given in Table 3. Examples of the replacement of A- or T-containing codons with G- or C-containing synonymous codons can be found throughout Table 3. A dramatic example is that 94% of lysine codons in *M. smithii* are AAA, whereas 86% of lysine codons in *M. thermoautotrophicum* are AAG. It is, however, also apparent from Table 3 that the overall *M. thermoautotrophicum* codon usage more closely resembles codon usage in *M. smithii* than in *E. coli*. There is frequent usage in both methanogens of codons such as AGA, AGG, and AUA, which are only very infrequently found in polypeptide-encoding genes of *E. coli* (Ikemura 1981; Konigsberg and Godson 1983). In contrast, codons containing the dinucleotide CG are infrequently used in the methanogen-derived genes, but are often found in *E. coli* genes. The rare occurrence of the CG dinucleotide is a well-established property of eucaryotic DNAs (Subak-Sharpe et al. 1967; Lennon and Fraser 1983; Nus-

Table 3. Codon usage in ORF-PurE and ORF-B[a] of *M. smithii* and *M. thermoautotrophicum* compared with codon usages in *E. coli*[b]

Residue and codon	*E. coli* Total codons	*E. coli* % Synonym use	*M. smithii* Total codons	*M. smithii* % Synonym use	*M. thermoautotrophicum* Total codons	*M. thermoautotrophicum* % Synonym use
Phe UUU	104	44	5	83	4	33
Phe UUC	135	56	1	17	8	67
Leu UUA	36	6	9	26	—	0
Leu UUG	51	8	4	11	—	0
Leu CUU	54	9	14	40	7	21
Leu CUC	41	7	1	3	17	52
Leu CUA	11	2	4	11	2	6
Leu CUG	432	69	3	9	7	21
Ile AUU	151	37	28	60	1	3
Ile AUC	252	62	5	10	3	9
Ile AUA	2	1	14	30	31	88
Met AUG	189	—	8	—	16	—
Val GUU	182	38	16	43	14	36
Val GUC	62	13	3	8	7	18
Val GUA	111	23	18	49	8	20
Val GUG	130	27	—	0	10	26
Ser UCU	86	27	5	21	5	17
Ser UCC	83	26	3	12	5	17
Ser UCA	27	8	7	29	10	33
Ser UCG	37	11	—	0	1	3
Ser AGU	21	6	6	25	2	6
Ser AGC	70	22	3	12	7	21
Pro CCU	24	9	9	45	6	24
Pro CCC	16	6	1	5	13	52
Pro CCA	53	20	7	35	2	8
Pro CCG	174	65	3	15	4	16
Thr ACU	76	24	11	58	1	11
Thr ACC	162	51	2	11	5	56
Thr ACA	19	6	6	31	1	11
Thr ACG	63	20	—	0	2	22
Ala GCU	202	28	15	63	5	15
Ala GCC	136	19	1	4	17	50
Ala GCA	166	23	8	33	11	32
Ala GCG	221	30	—	0	1	3
Tyr UAU	69	41	9	82	4	45
Tyr UAC	101	59	2	18	5	55
Ter UAA	22	88	1	100	1	100
Ter UAG	1	4	—	0	—	0
Ter UGA	2	8	—	0	—	0
His CAU	42	39	5	83	4	67
His CAC	66	61	1	17	2	33
Gln CAA	75	27	3	50	—	0
Gln CAG	207	73	3	50	4	100
Asn AAU	57	24	23	77	1	8
Asn AAC	179	76	7	23	12	92
Lys AAA	296	77	34	94	3	14
Lys AAG	90	23	2	6	18	86
Asp GAU	175	51	11	69	12	57
Asp GAC	168	49	5	31	9	43
Glu GAA	328	73	22	79	15	50
Glu GAG	119	27	6	21	15	50
Cys UGU	21	42	2	50	2	40
Cys UGC	29	58	2	50	3	60
Trp UGG	48	—	—	—	1	—
Arg CGU	201	58	1	7	1	4
Arg CGC	121	35	1	7	1	4
Arg CGA	8	2	1	7	—	0
Arg CGG	11	3	—	0	2	8
Arg AGA	4	1	8	56	7	27
Arg AGG	1	0.25	3	23	15	57
Gly GGU	231	48	6	20	9	29
Gly GGC	197	41	—	0	5	16
Gly GGA	22	5	22	73	8	26
Gly GGG	33	7	2	7	9	29

[a] Table contains the codons designated by the ORF-PurE and ORF-B sequences in Figs. 3 and 5

[b] The *E. coli* codon usages are from 25 *E. coli* genes as listed by Konigsberg and Godson (1983)

sinov 1984); results presented here and elsewhere (Cue et al. 1985; Hamilton and Reeve 1985) suggest that this eucaryotic property also extends to methanogenic archaebacteria.

Intergenic Regions

The intergenic region between ORF-PurE$_s$ and ORF-B$_s$ in *M. smithii* consists of only 9 bp (Hamilton and Reeve 1985). There are 57 bp in the analogous *M. thermoautotrophicum* region (Figs. 2 and 5). The sequence 5′-GGTGA, which has the potential to hybridize to the 3′ terminus of *M. thermoautotrophicum* 16S rRNA, is located as expected for a ribosome binding sequence, just preceding the ATG initiation codon for ORF-B$_t$. The origin and function, if any, of the 48 bp in the *M. thermoautotrophicum* intergenic region that are not present in the intergenic region of *M. smithii*, remain to be determined. There is, however, strong circumstantial evidence that these extra base pairs arose, at least in part, by duplication of an existing sequence. There is an 18-bp sequence located within ORF-PurE$_t$, but very close to the carboxyl terminus of ORF-PurE$_t$, that is directly repeated 30 bp downstream within the intergenic region. The first 9 bp of this duplicated 18 bp sequence are also found directly repeated a third time immediately preceding the 18-bp sequence within the ORF-PurE$_t$ coding region (Fig. 5).

The DNA sequences surrounding the initiation

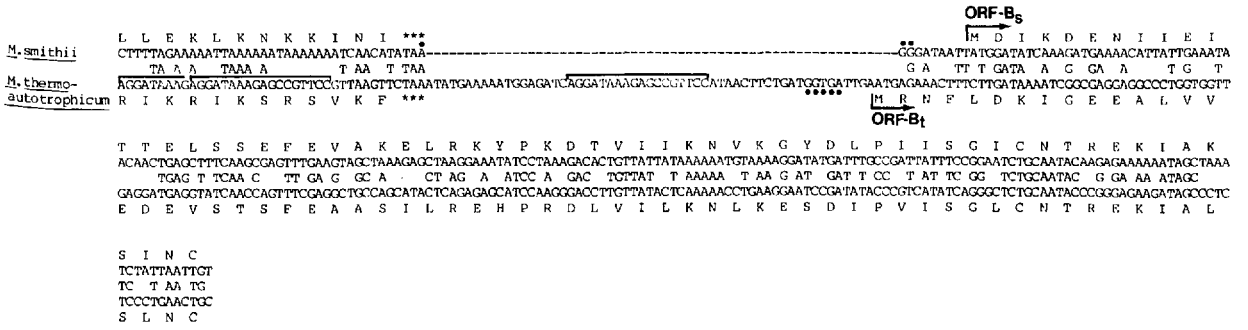

Fig. 5. Sequences of the intergenic regions between the ORF-PurE and ORF-B sequences. The termination codons (TAA) of the ORF-PurE sequences are indicated (***) and the initiation codons of the ORF-B sequences are shown by arrows. Ribosome binding sequences are indicated by dots, and the sequences found in both ORF-PurE$_t$ and in the intergenic region between ORF-PurE$_t$ and ORF-B$_t$ are overlined. The complete sequence of ORF-B$_s$ is available (Hamilton and Reeve 1985), whereas only the sequence shown in the figure has been cloned from *M. thermoautotrophicum*. Details of obtaining the *M. thermoautotrophicum* sequence and organization of the figure are as described in Fig. 3

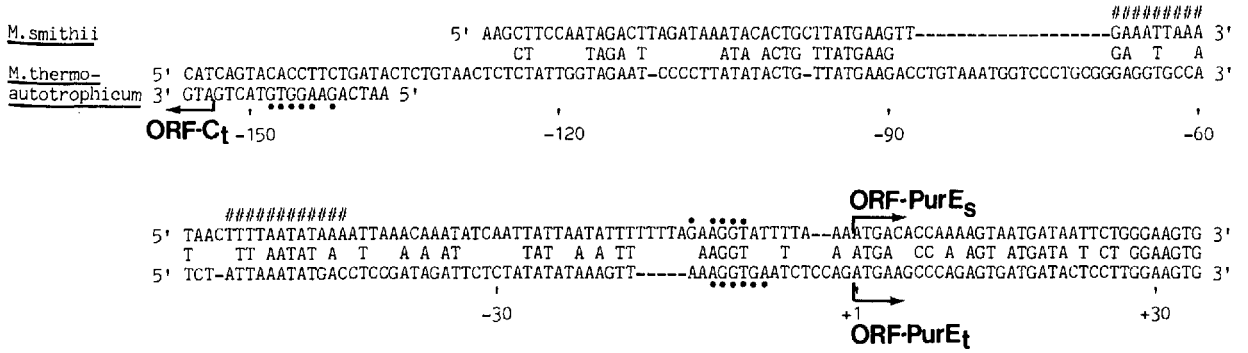

Fig. 6. Sequences of the regions surrounding and preceding the initiation codons of the ORF-PurE genes. The three upper rows of sequence information were obtained and used to construct the figure as described in Fig. 3. The bottom (fourth) row of sequence, where given, is the complementary strand of the *M. thermoautotrophicum* sequence. The initiation codons for ORF-PurE$_s$, ORF-PurE$_t$, and ORF-C$_t$ are indicated by arrows. The ORF-C$_t$ sequence continues, uninterrupted by nonsense codons, for 208 bp (sequence not shown) to the PstI site used in cloning the fragment of *M. thermoautotrophicum* DNA into pUC8. The sequences previously suggested as members of a family of conserved sequences that are possible promoters in *M. smithii* (Hamilton and Reeve 1985) are indicated (##). Ribosome binding sequences are indicated by dots. The bases in the sequence are numbered so that +1 is the A of the ORF-PurE$_t$ initiation codon in the *M. thermoautotrophicum* sequence

codons of the two ORF-PurE genes are compared in Fig. 6. Sequence homology is more conserved 3′ to the ATG initiation codons than it is 5′ to these codons (Fig. 3 and 6). The sequences that are conserved in the 5′ intergenic regions do, however, include the sequence previously suggested as being important as a ribosome binding sequence (5′-AGGT) and partially overlap with sequences suggested as possible promoters in *M. smithii* (Hamilton and Reeve 1985). Consensus promoter sequences for archaebacteria were suggested to be 5′-GAANTTTCA and an expanded "TATA" sequence, namely 5′-TTTTAATATAAA (Hamilton and Reeve 1985). Two variants of the former sequence precede ORF-PurE$_s$ and partially overlap with each other. These are −93 GAAGTTGAA and −68 GAAATTAAA, using the base-numbering system shown in Fig. 6. The corresponding sequence in the *M. thermoautotrophicum* DNA, −68 GAGGTGCCA, matches the 9-bp consensus sequence at six positions, although there are only 4 bp in this region that, in the alignment shown, are homologous with bases in the *M. smithii* sequence. The most extensive sequence homology (15 of 17 bases) occurs between positions −105 and −90 and terminates with the first four bases of the −93 GAAGTTGAA sequence of *M. smithii*. The same base occurs in the *M. smithii* sequence and in the *M. thermoautotrophicum* sequence at 8 of 12 positions in the expanded TATA sequence previously identified between bases −55 and −44 (Fig. 6).

As indicated in Fig. 2, there is an ORF (ORF-C$_t$) in the cloned *M. thermoautotrophicum* DNA on the opposite DNA strand from ORF-PurE$_t$. The initiation ATG codon and the bases immediately preceding this codon, including bases that, if transcribed, would form a ribosome binding site, are also shown in Fig. 6. If ORF-C$_t$ were a bona fide gene then the region between ORF-C$_t$ and ORF-PurE$_t$ would contain divergent transcriptional units.

Discussion

Structure and Activities of the Products of the Cloned Genes

DNA cloned from the mesophilic methanogen *M. smithii* or the thermophilic methanogen *M. thermoautotrophicum* has been shown to be expressed in aerobically grown *E. coli,* resulting in complementation of mutations in either or both of the $purE_1$ and $purE_2$ loci (Gots et al. 1977) of *E. coli.* Complementation was observed, in both cases, in *E. coli* grown over a wide range of temperatures, including temperatures as low as 11°C. There are no dramatic differences in the predicted amino acid sequences of the two methanogen polypeptides, such as additional opportunities to form disulfide bridges, that might readily correlate with the increased thermal stability expected of an enzyme from a thermophile. The amino acid changes that do occur are, however, consistent with previously observed preferred changes for increased thermostability of proteins. Argos et al. (1980) analyzed the amino acid sequences of several proteins and found that the most frequent exchanges are glycine to alanine, serine to alanine, serine to threonine, and lysine to arginine when comparing the composition of enzymes isolated from mesophiles with the composition of the same enzymes isolated from thermophiles, respectively. In comparing $gpPurE_s$ with $gpPurE_t$, there are eight glycine to alanine, five serine to alanine, two serine to threonine, and 11 lysine to arginine changes, in contrast to one alanine to glycine, two alanine to serine, three threonine to serine, and one arginine to lysine changes, respectively. Only the change of serine with threonine residues deviates from the correlations reported by Argos et al. (1980).

Expression of Methanogen DNA in E. coli

In a previous report (Hamilton and Reeve 1985) we identified conserved DNA sequences in cloned *M. smithii* DNAs that could act as ribosome binding sites and sequences that might be promoters in this methanogen. We argued, however, that transcription of methanogen DNA in *E. coli* probably resulted from the fortuitous presence in the *M. smithii* DNA of sequences that could act as promoters for *E. coli* RNA polymerase. Analysis of the sequences of the DNA cloned from *M. thermoautotrophicum,* as presented in Figs. 3, 5, and 6, supports our identification of 5'-AGGTGA as a consensus ribosome binding sequence. This sequence should facilitate mRNA binding to rRNA both in methanogens and in *E. coli* (Shine and Dalgarno 1974; Steitz 1978). The DNA cloned from *M. thermoautotrophicum* also contains appropriately positioned versions of the putative archaebacterial promoter sequence and several sequences that are sufficiently similar to the canonical *E. coli* promoter that they could presumably function as promoters in *E. coli.* There is, however, no value in further discussing the significance of these sequences until the sites of transcription initiation are accurately known. Research into this matter is currently in progress.

Sequence Divergence

The divergence of *M. smithii* and *M. thermoautotrophicum* as species has resulted in two ORF-PurE sequences that contain 53% homologous bases and that encode polypeptides with sequences containing 45% homologous amino acid residues. Comparison of oligonucleotides produced by RNase T1 digestion of 16S rRNAs of these two species showed 293 common nucleotides in oligonucleotides of hexamer length or longer (Balch et al. 1979). Based on these oligonucleotide catalogues, *M. smithii* and *M. thermoautotrophicum* were calculated to have an association coefficient (S_{AB}) of 0.49 and were assigned to different genera. The purE sequences reported here do not provide additional support for or detract from this assignment, but provide additional polynucleotide sequences that might be used to estimate evolutionary divergence. Ideally, one would like to correlate S_{AB} values for several pairs of methanogens, calculated by comparing RNase T1 oligonucleotide catalogues, with the extents of divergence of the purE sequences of the same methanogens. Such an extended correlation would require the determination of many methanogen purE sequences.

A less attractive, but more practical, possibility is correlation of the extents of sequence divergence of pairs of related genes from different methanogens with the corresponding oligonucleotide-based S_{AB} values. For example, the hisA genes of *Methanococcus vannielii* and *Methanococcus voltae* have recently been sequenced and were found to be 66% homologous in their DNA sequences (Cue et al. 1985). The RNase T1 oligonucleotide catalogues of the 16S rRNAs of these two species placed them in the same genus (*Methanococcus*) (Balch et al. 1979), with 352 common nucleotides in sequences of hexamer length or longer. This value was used to calculate a S_{AB} value of 0.60. Thus, the 53% sequence homology of the purE genes of *M. smithii* and *M. thermoautotrophicum* and the 66% sequence homology of the hisA genes of *M. vannielii* and *M. voltae* could be correlated with S_{AB} values for these two pairs of organisms of 0.49 and 0.60, respectively. The extent to which this type of correlation holds will become apparent only as more methanogen-derived sequences become available.

A major concern in evaluating such a correlation

Table 4. Frequencies of occurrence of RNY[a] codons in methanogen open reading frames (ORFs)

Methanogen	ORF[b]	Number of codons[c]	Frame[d] 0 RNY	0 Stop	1 RNY	1 Stop	2 RNY	2 Stop	Reference containing DNA sequence
M. smithii	$purE_s$	339	120	0	39	33	85	37	Hamilton & Reeve (1985)
M. thermoautotrophicum	$purE_t$	334	89	0	39	36	94	23	This report
M. smithii	B_s	418	155	0	58	33	95	52	Hamilton & Reeve (1985)
M. smithii	IS	401	98	0	91	32	94	47	Hamilton & Reeve (1985)
M. smithii	proC	251	84	0	37	23	58	25	Hamilton & Reeve (1985)
M. vannielii	1	502	141	0	85	36	121	52	Cue et al. (1985)
M. vannielii	3	76	25	0	9	5	23	9	Cue et al. (1985)
M. vannielii	hisA	238	81	0	26	18	58	27	Cue et al. (1985)
M. voltae	hisA	242	83	0	25	28	55	20	Cue et al. (1985)

[a] RNY codons defined by Shepherd (1981, 1983). R = purine; Y = pyrimidine; N = purine or pyrimidine

[b] ORF designations are given in the cited references. The genetic loci indicate that mutations in these genes of *E. coli* are complemented by the cloned methanogen gene

[c] Number of amino acid-encoding codons

[d] Frames 0, 1, and 2 begin with the A, U, and G of the AUG initiation codon, respectively. The number of termination codons (UAA, UAG, and UGA) in each reading frame is listed under "Stop"

is how to quantitate the effects of different evolutionary pressures on unrelated pairs of genes. In the examples described above the two purE genes have diverged under selective pressure so as to produce enzymes that can function in a thermophile (*M. thermoautotrophicum*) and in a mesophile (*M. smithii*), whereas the hisA genes have evolved to encode enzymes that can function in a marine organism (*M. voltae*) and in a freshwater organism (*M. vannielii*). Currently, there seems to be no accepted way to assess how these different selective pressures might differentially affect the rates of divergence of DNA sequences. Although there has been significantly more change in the overall base composition during the divergence of the genome of *M. smithii* (30.6% G + C) from the genome of *M. thermoautotrophicum* (49.7% G + C) as compared with the divergence of the genome of *M. vannielii* (31.1% G + C) from the genome of *M. voltae* (30.7% G + C), this is not reflected in a very substantial difference in the extent of divergence of the purE genes (53% homologous) as compared with the extent of divergence of the hisA genes (66% homologous).

The RNY Rule

Shepherd (1981, 1983) has proposed that polypeptide-encoding ORFs can be recognized by their preferential use of RNY codons (R = purine, Y = pyrimidine, N = purine or pyrimidine). According to this proposal, RNY codons occur most frequently in the correct reading frame, whereas nonutilized ORFs do not show preferential RNY codon usage. The only reported exception to this rule, other than the highly evolved overlapping genes of bacteriophages such as $\phi\chi174$, is the archaebacterial gene of *Halobacterium halobium* that encodes bacterio-opsin (Clarke 1983; Shepherd 1983).

We have analyzed the ORF-$PurE_s$ and ORF-$PurE_t$ sequences for RNY codons (Table 4). Surprisingly, whereas the $purE_s$ gene of *M. smithii* conforms to the RNY rule, the $purE_t$ gene of *M. thermoautotrophicum* does not. Nevertheless the evidence, in addition to the extensive sequence homology with ORF-$PurE_s$, that ORF-$PurE_t$ does encode the purE-complementing polypeptide is strong. The mutation $purE_t\Delta445$, which is located within ORF-$PurE_t$, inactivates complementation, and the electrophoretic mobility of the polypeptide encoded by the mutated ORF-$PurE_t$ sequence (gp$PurE_t\Delta445$, Fig. 4) is slightly increased. These are the phenotypes expected if, as shown by DNA sequencing, the mutation is an in-frame deletion that removes seven internal amino acids from gp$PurE_t$. If RNY codons do represent a primitive organization of the genetic code (Shepherd 1981, 1983) then it would appear that the evolutionary divergence of *M. smithii* from *M. thermoautotrophicum* has allowed only the *M. smithii* purE gene to maintain the preferential usage of RNY codons.

That archaebacterial genes might, in general, not conform to the RNY rule, as suggested by the DNA sequences of the bacterio-opsin gene (Clarke 1983; Shepherd 1983) and of ORF-$PurE_t$, does not seem to be the case. Several additional methanogen genes (Table 4) and the halobacterial brp gene (Betlach et al. 1984; C.H. Clarke, personal communication, 1985) have been analyzed and these archaebacterial genes all conform to the RNY rule.

Acknowledgments. This work was supported by contracts AC02-81ER10945 from the Department of Energy, 5083-260-0895 from the Gas Research Institute, and CR810340 from the Environmental Protection Agency. J.N.R. is the recipient of Research Career Development Award 5K04AG00108 from the National Institute on Aging. We thank Dr. C.H. Clarke for introducing us to the use of RNY analysis.

References

Argos P, Rossman MG, Grau UM, Zuber H, Frank G, Tratschin JD (1980) Thermal stability and protein structure. In: Sigman DS, Brazier MAR (eds) The evolution of protein structure and function. Academic Press, New York, pp 159–169

Balch WE, Fox GE, Magrum LJ, Woese CR, Wolfe RS (1979) Methanogens: reevaluation of a unique biological group. Microbiol Rev 43:260–296

Betlach M, Friedman J, Boyer HB, Pfeifer F (1984) Characterization of a halobacterial gene affecting bacterio-opsin gene expression. Nucleic Acids Res 12:7949–7959

Clarke CH (1983) Mutational evolution of an archaebacterial gene. Heredity (Edinburgh) 50:205

Cue D, Beckler GS, Reeve JN, Konisky J (1985) Structure and sequence divergence of two archaebacterial genes. Proc Natl Acad Sci USA 82:4207–4211

Davis RW, Botstein D, Roth JB (1980) Advanced bacterial genetics. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York

Fox GE, Magrum LJ, Balch WE, Wolfe RS, Woese CR (1977) Classification of methanogenic bacteria by 16S ribosomal RNA characterization. Proc Natl Acad Sci USA 74:4537–4541

Gots JS, Benson CE, Jochimsen B, Koduri KR (1977) Microbial models and regulatory elements in the control of purine metabolism. Ciba Found Symp 48:23–41

Hamilton PT, Reeve JN (1984) Cloning and expression of archaebacterial DNA from methanogens in *Escherichia coli*. In: Strohl WR, Tuovinen OH (eds) Microbial chemoautotrophy. Ohio State University Press, Columbus, pp 291–307

Hamilton PT, Reeve JN (1985) Structure of genes and an insertion element in the methane producing archaebacterium *Methanobrevibacter smithii*. Mol Gen Genet 200:47–59

Hook LA, Corder RE, Hamilton PT, Frea JI, Reeve JN (1984) Development of a plating system for genetic exchange studies in methanogens using a modified ultra-low oxygen chamber. In: Strohl WR, Tuovinen OH (eds) Microbial chemoautotrophy. Ohio State University Press, Columbus, pp 275–289

Ikemura T (1981) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. J Mol Biol 146:1–21

Konigsberg W, Godson GN (1983) Evidence for use of rare codons in the dnaG and other regulatory genes of *Escherichia coli*. Proc Natl Acad Sci USA 80:687–691

Lennon GG, Fraser NW (1983) CpG frequency in large DNA segments. J Mol Evol 19:286–288

Maxam AM, Gilbert W (1980) Sequencing end-labeled DNA with base-specific chemical cleavage. Methods Enzymol 65: 499–580

Morris CJ, Reeve JN (1984) Functional expression of an archaebacterial gene from the methanogen *Methanosarcina barkeri* in *Escherichia coli* and *Bacillus subtilis*. In: Crawford RL, Hanson RS (eds) Microbial growth on Cl compounds. American Society for Microbiology, Washington, DC, pp 205–209

Nussinov R (1984) Doublet frequencies in evolutionarily distinct groups. Nucleic Acids Res 12:1749–1763

Reeve JN (1979) Use of minicells for bacteriophage directed polypeptide biosynthesis. Methods Enzymol 68:493–503

Reeve JN, Trun NJ, Hamilton PT (1982) Beginning genetics with methanogens. In: Hollaender A, DeMoss RD, Kaplan S, Konisky J, Savage D, Wolfe RS (eds) Genetic engineering of microorganisms for chemicals. Plenum, New York, pp 233–244

Shepherd JCW (1981) Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. Proc Natl Acad Sci USA 78:1596–1600

Shepherd JCW (1983) From the primeval message to present-day gene. Cold Spring Harbor Symp Quant Biol 47:1099–1108

Shine J, Dalgarno L (1974) The 3'-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. Proc Natl Acad Sci USA 71:1342–1346

Steitz JA (1978) Methanogenic bacteria. Nature 273:100–101

Subak-Sharpe H, Burk RR, Crawford LV, Morrison JM, Hay J, Keir MH (1967) An approach to evolutionary relationships of mammalian DNA viruses through analysis of the pattern of nearest neighbour base sequences. Cold Spring Harbor Symp Quant Biol 31:737–751

Vieira J, Messing J (1982) The pUC plasmids, an M13mp7-derived system for insertion mutagenesis and sequencing with synthetic universal primers. Gene 19:259–268

Woese CR, Fox GF (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci USA 74:5088–5090

Woese CR, Magrum LJ, Fox GE (1978) Archaebacteria. J Mol Evol 11:245–252

Received June 7, 1985/Accepted August 20, 1985