J Mol Evol (1988) 27:65–76

Journal of Molecular Evolution

Nucleotide Sequence of Regions Homologous to *nifH* (Nitrogenase Fe Protein) from the Nitrogen-Fixing Archaebacteria *Methanococcus thermolithotrophicus* and *Methanobacterium ivanovii*: Evolutionary Implications

Nicole Souillard, Michel Magot,[1] Odile Possot, and Lionel Sibold

Unité de Physiologie Cellulaire, Département des Biotechnologies, Institut Pasteur, 25 rue du Dr. Roux, 75724 Paris Cedex 15, France

Summary. DNA fragments bearing sequence similarity to eubacterial *nifH* probes were cloned from two nitrogen-fixing archaebacteria, a thermophilic methanogen, *Methanococcus* (*Mc.*) *thermolithotrophicus*, and a mesophilic methanogen, *Methanobacterium* (*Mb.*) *ivanovii*. Regions carrying similarities with the probes were sequenced. They contained several open reading frames (ORF), separated by A+T-rich regions. The largest ORFs in both regions, an 876-bp sequence in *Mc. thermolithotrophicus* and a 789-bp sequence in *Mb. ivanovii*, were assumed to be ORFs*nifH*. They code for polypeptides of mol. wt. 32,025 and 28,347, respectively. Both ORFs*nifH* were preceded by potential ribosome binding sites and followed by potential hairpin structures and by oligo-T sequences, which may act as transcription termination signals. The codon usage was similar in both ORFs*nifH* and was analogous to that used in the *Clostridium pasteurianum nifH* gene, with a preference for codons ending with A or U. The ORF*nifH* deduced polypeptides contained 30% sequence matches with all eubacterial *nifH* products already sequenced. Four cysteine residues were found at the same position in all sequences, and regions surrounding the cysteine residues are highly conserved. Comparison of all pairs of methanogenic and eubacterial *nifH* sequences is in agreeement with a distant phylogenetic position of archaebacteria and with a very ancient origin of *nif* genes. However, sequence similarity between Methanobacteriales and Methanococcales is low (around 50%) as compared to that found among eubacteria, suggesting a profound divergence between the two orders of methanogens. From comparison of amino acid sequences, *C. pasteurianum* groups with the other eubacteria, whereas comparison of nucleotide sequences seems to bring *C. pasteurianum* closer to methanogens. The latter result may be due to the high A+T content of both *C. pasteurianum* and methanogens ORFs*nifH* or may come from an ancient lateral transfer between *Clostridium* and methanogens.

Key words: Nitrogenase Fe protein — *Methanococcus thermolithotrophicus* — *Methanobacterium ivanovii*

Introduction

Methanogens, a branch of the archaebacterial kingdom, form a homogeneous group, since all of them are strict anaerobes and synthesize methane. They are, however, diversified with respect to their G+C content (from 28 to 61%), morphology (coccoid, rod-shaped, filamentous, spiral bacteria), growth temperature (from 20 to 86°C), and substrate utilization (H_2/CO_2, formate, acetate, methanol, methylamines) (for a review, see Whitman 1985). Methanogens are divided into three orders: Methanomicrobiales, Methanococcales, and Methanobacteriales. S_{AB} values based on 16S rRNA cataloging data range from 0.22 to 0.28 between two orders

Offprint requests to: L. Sibold

[1] *Current address:* Elf Biorecherches, Labège, 31320 Castanet Tolosan, France

of methanogens, whereas they are about 0.15 between eubacteria and methanogens (Balch et al. 1979).

The ability to use N_2 as sole nitrogen source was demonstrated for archaebacteria only recently. In nitrogen-fixing eubacteria, nitrogenase is composed of two proteins, the MoFe protein and the Fe protein. The MoFe protein is an $\alpha_2\beta_2$ tetramer, the subunits of which are coded by *nifD* (α) and *nifK* (β). The Fe protein is a dimer comprised of two identical subunits coded by *nifH* (Eady and Smith 1979). In eubacteria, the three genes are highly conserved, and particularly *nifH* (Ruvkun and Ausubel 1980). In 1984, nitrogen fixation was reported in two methanogenic species: *Methanococcus* (*Mc.*) *thermolithotrophicus* (Belay et al. 1984) and *Methanosarcina* (*Ms.*) *barkeri* 227 (Murray and Zinder 1984). At the same time, sequence similarity with *nifHDK* was found in DNA of four different species (Sibold et al. 1985). Subsequently, three other species were characterized as Nif$^+$: *Ms. barkeri* Fusaro (Bomar et al. 1985), *Methanolobus tindarius* (König et al. 1985), and *Methanobacterium* (*Mb.*) *ivanovii* (Magot et al. 1986), and sequence similarity with eubacterial *nifH* probes was found in 14 methanogens tested (Possot et al. 1986).

In eubacteria, the *nifH* sequences were used to construct a phylogenetic tree, which appeared to be consistent with that deduced from 16S rRNA sequences (Hennecke et al. 1985). This suggested that *nifH* had evolved in eubacteria without recent lateral transfer. Souillard and Sibold (1986) found that the nucleotide sequence of a cloned region similar to *nifH* (ORF*nifH*) from *Mc. voltae* had lower S_{AB} values with all eubacterial *nifH* than the lowest value found between two eubacterial *nifH*. This was in agreement with the phylogenetic position of archaebacteria and with a very ancient origin of *nif* genes. However, attempts to derepress a nitrogenase activity in *Mc. voltae* have failed (Sibold, unpublished results). It thus appeared interesting to study *nifH* genes from methanogens that had been found to grow with N_2 as sole nitrogen source. We chose two such strains, *Mc. thermolithotrophicus* and *Mb. ivanovii*. *Methanococcus thermolithotrophicus* belongs to the Methanococcales, has a G+C content of 30%, an optimal growth temperature of 65°C, and is able to fix nitrogen at temperatures as high as 60–65°C (Belay et al. 1984). The generation time is about 1 h in the presence of ammonia and about 3 h under conditions of nitrogen fixation. *Methanobacterium ivanovii* belongs to the Methanomicrobiales, has a G+C content of 37%, and optimal growth at 37°C. Its generation time is about 9 h in the presence of ammonia and about 24 h under conditions of nitrogen fixation (Magot et al. 1986). At the DNA level, sequence similarity with *Anabaena* and *Klebsiella pneumoniae nifH* probes was found in *Mc. thermolithotrophicus* and *Mb. ivanovii* using the Southern blot technique (Possot et al. 1986). We report here the cloning and sequencing of regions exhibiting sequence similarity to *nif H* from both bacteria.

Materials and Methods

Bacterial Strains and Plasmids. *Methanobacterium ivanovii* (Belyaev et al. 1983; Jain et al. 1987) was obtained from Dr. J.G. Zeikus, and *Methanococcus thermolithotrophicus* (DSM 2095) from Dr. M. Thomm. *Escherichia coli* strain TG1 (Wain-Hobson et al. 1985) was used for M13 or pUC cloning. Strain VCS257 (Genofit) was used for bacteriophage λ transfection. Cloning vectors were pUC9 (Vieira and Messing 1982) or λEMBL3 (Frischauf et al. 1983). The intragenic *Anabaena nifH* probe was a 0.8-kbp *Hpa*I–*Pvu*I fragment isolated from plasmid pAn154.3 (Mevarech et al. 1980). The *Klebsiella pneumoniae nifH* probe was a 0.9-kbp *Eco*RI–*Kpn*I fragment of plasmid pSA30 (Cannon et al. 1979).

Routine Molecular Biology Techniques. DNA isolation, restriction endonuclease treatment, transformation, ligation, Klenow enzyme treatment, *Bal*31 treatment, calf intestine phosphatase treatment, Southern blot, and colony and plaque hybridizations were performed according to Maniatis et al. (1982) or as already described (Sibold et al. 1985).

Cloning of Mc. thermolithotrophicus nifH Region. A 2.8-kbp *Hin*dIII fragment was previously shown to carry homology with *Anabaena nifH* (Possot et al. 1986). In order to clone this fragment, plasmid pUC9 was digested by *Hin*dIII, dephosphorylated using calf intestine phosphatase, and ligated with *Mc. thermolithotrophicus* DNA digested by *Hin*dIII. After transformation of *E. coli* strain TG1, 4000 carbenicillin-resistant clones were obtained, among which 95% were Lac$^-$. Thirteen positive clones were detected by colony hybridization using the *Anabaena nifH* probe. Two clones, containing the 2.8-kbp *Hin*dIII fragment in both orientations, were termed pMCT100 and pMCT200 (Fig. 1A).

Cloning of Mb. ivanovii nifH Region. A 1.3-kbp *Hin*dIII fragment was previously shown to carry homology with *K. pneumoniae nifH* (Sibold et al. 1985). In order to clone the DNA region containing this fragment, *Mb. ivanovii* DNA was partially digested by *Sau*3A to obtain a majority of fragments of about 20 kbp, dephosphorylated using calf intestine phosphatase, and ligated with λEMBL3 DNA digested by *Bam*HI and *Eco*RI. Encapsidation was performed using Gigapack mix (Genofit) as recommended by the supplier, and *E. coli* strain VCS257 was transfected. Among 900 plaques tested by hybridization with a *K. pneumoniae nifH* probe, a phage was detected and termed λMIV1. Plasmid pMIV27 was obtained by subcloning a 5-kbp *Bam*HI fragment of phage λMIV1 into pUC9 (Fig. 1B).

DNA Sequencing. DNA sequencing was performed using the method of Sanger et al. (1977) with phage M13 derivatives. For *Mc. thermolithotrophicus*, sonicated DNA fragments of pMCT100 were cloned into M13mp9; for *Mb. ivanovii*, restriction fragments or *Bal*31-treated DNA fragments were cloned into M13mp8 and/or M13mp9. DNA regions indicated by arrows in Fig. 1 were sequenced on both strands. Nucleotide and amino acid sequence comparisons were performed with a computer using EMBL and Genbank data banks and a program adapted from Lipman and Wilbur (1983).

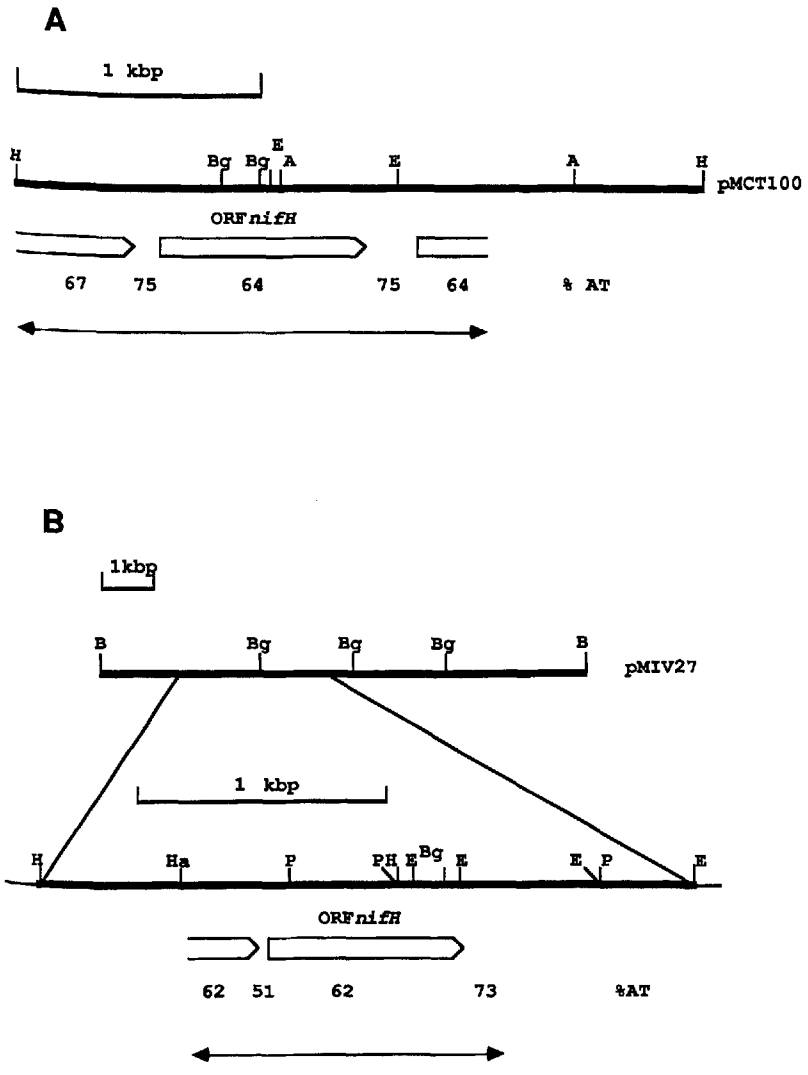

Fig. 1. Restriction map of *nifH*-containing regions of *Mc. thermolithotrophicus* and *Mb. ivanovii*. **A** Restriction map of the *Mc. thermolithotrophicus* DNA fragment with sequence similarity to *nifH* cloned in pMCT100. **B** Restriction map of the *Mb. ivanovii* DNA fragment with sequence similarity to *nifH* cloned in pMIV27. Thin double-headed arrows indicate the extent of the sequenced regions and open arrows indicate the open reading frames. The numbers indicate the mol% A+T of ORFs and noncoding regions. A, *Ava*II; B, *Bam*HI; Bg, *Bgl*II; E, *Eco*RI; Ha, *Hae*III; H, *Hin*dIII; P, *Pst*I.

Results

Nucleotide Sequence Analysis of Mc. thermolithotrophicus and Mb. ivanovii nifH Regions

A 1950-bp DNA region from pMCT100 containing sequence similarity with *Anabaena nifH* was sequenced (Fig. 1A). The nucleotide sequence and the predicted amino acid sequence are shown in Fig. 2. In the sequenced region, a large open reading frame (ORF) was found (Fig. 1A). Based on the sequence similarity (see below), this ORF was assigned to *nifH*. *Methanococcus thermolithotrophicus* ORF-*nifH* is 876 bp long, from a UUG codon to a UAA codon, with a coding capacity of 292 amino acid residues (mol. wt. 32,025). The UUG codon was predicted to be the translation initiation codon because it is preceded, 3 bp upstream, by the sequence 5′-AGGTGAT-3′, complementary to the 3′ end of the 16S rRNA from *Methanococcus* (Balch et al. 1979). No AUG or GUG codon was found in this region. Upstream and downstream from ORF*nifH*, two other ORFs were found, 500 and 200 bp long, respectively (Fig. 1A). The 200-bp ORF starts with an AUG codon preceded, 3 bp upstream, by the sequence 5′-GGTGA-3′, complementary to the 3′ end of the 16S rRNA from *Methanococcus*. The A+T content is not homogeneously distributed along the sequenced region. The ORFs have a mol% A+T of 65 whereas noncoding regions have a mol% A+T of 73–75. The average mol% A+T in the sequenced region is about 70, in agreement with the value of 69% for the genome of *Mc. thermolithotrophicus* (Huber et al. 1982).

A 1350-bp region from pMIV27 containing sequence similarity with *K. pneumoniae nifH* was sequenced (Fig. 1B). The nucleotide sequence and the predicted amino acid sequence are shown in Fig. 3. *Methanobacterium ivanovii* ORF*nifH* is 789 bp long, starting with an AUG and ending with a UAA colon, and codes for a polypeptide of 263 residues (mol.

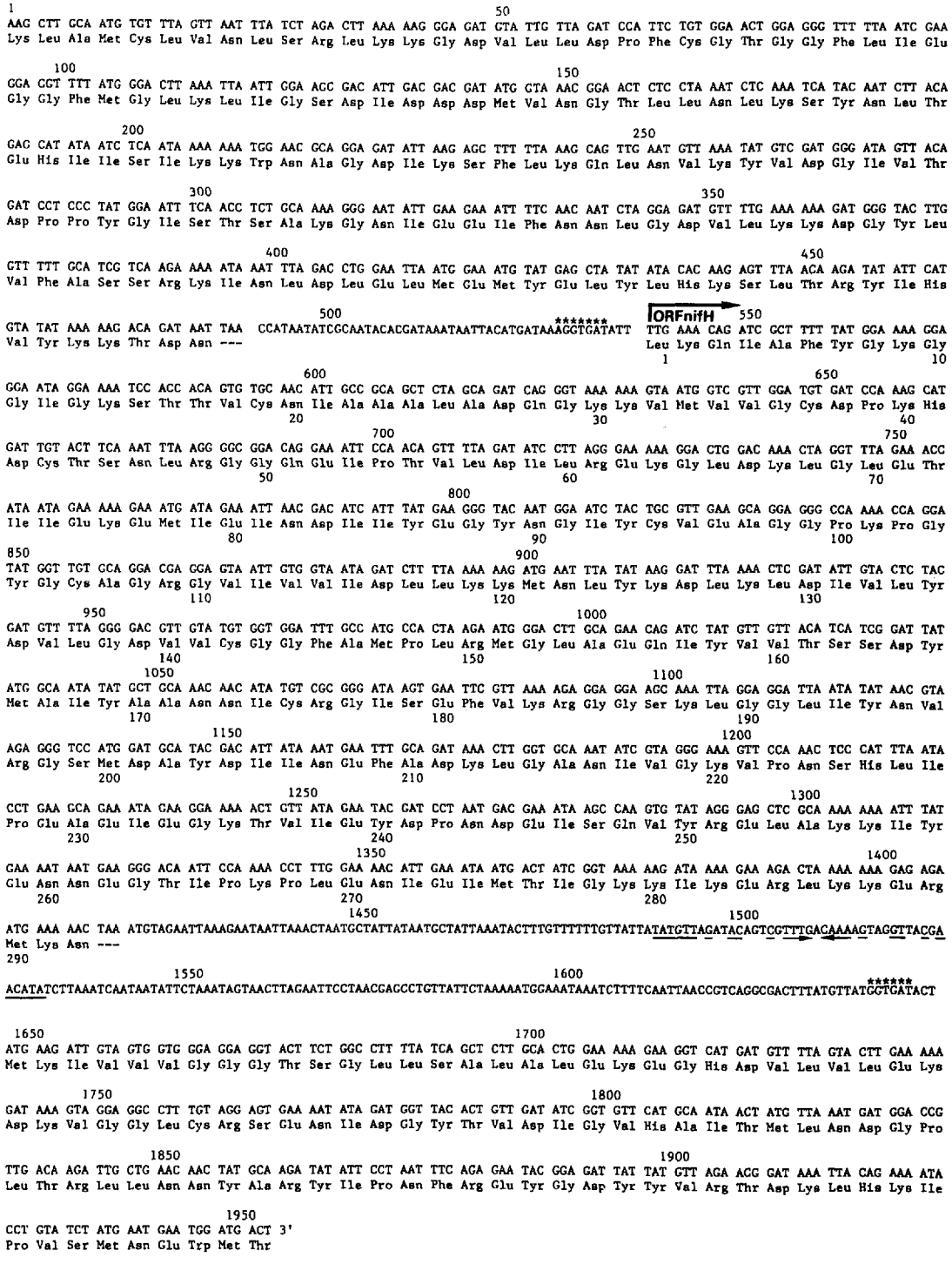

Fig. 2. Nucleotide sequence of *Mc. thermolithotrophicus* ORF*nifH* and surrounding regions. The noncoding DNA strand is shown. The deduced amino acid sequence of ORFs is shown below. Nucleotide positions are numbered above; amino acid positions of ORF*nifH* are numbered below. Potential ribosome-binding sites are indicated by stars. Palindromic sequences are indicated by arrows.

wt. 28,347). The sequence 5'-GGTGA-3', complementary to the 3' end of *Methanobacterium* 16S rRNA (Balch et al. 1979), was found 5 bp upstream from the putative translation initiation codon. Another 300-bp ORF was found upstream from ORF*nifH* (Fig. 1B). The A+T content is not homogeneously distributed along the sequenced region. The ORFs have a mol% A+T of 62 and 63 (Fig. 1B), whereas the noncoding regions have a mol% A+T of 73, except for the 50-bp stretch between the two ORFs, which has a mol% A+T of only 51. The average A+T content in the sequenced region is about 65%, which is consistent with the average A+T content of 63% for the genome of *Mb. ivanovii* (Belyaev et al. 1983).

Promoter consensus sequences in archaebacteria

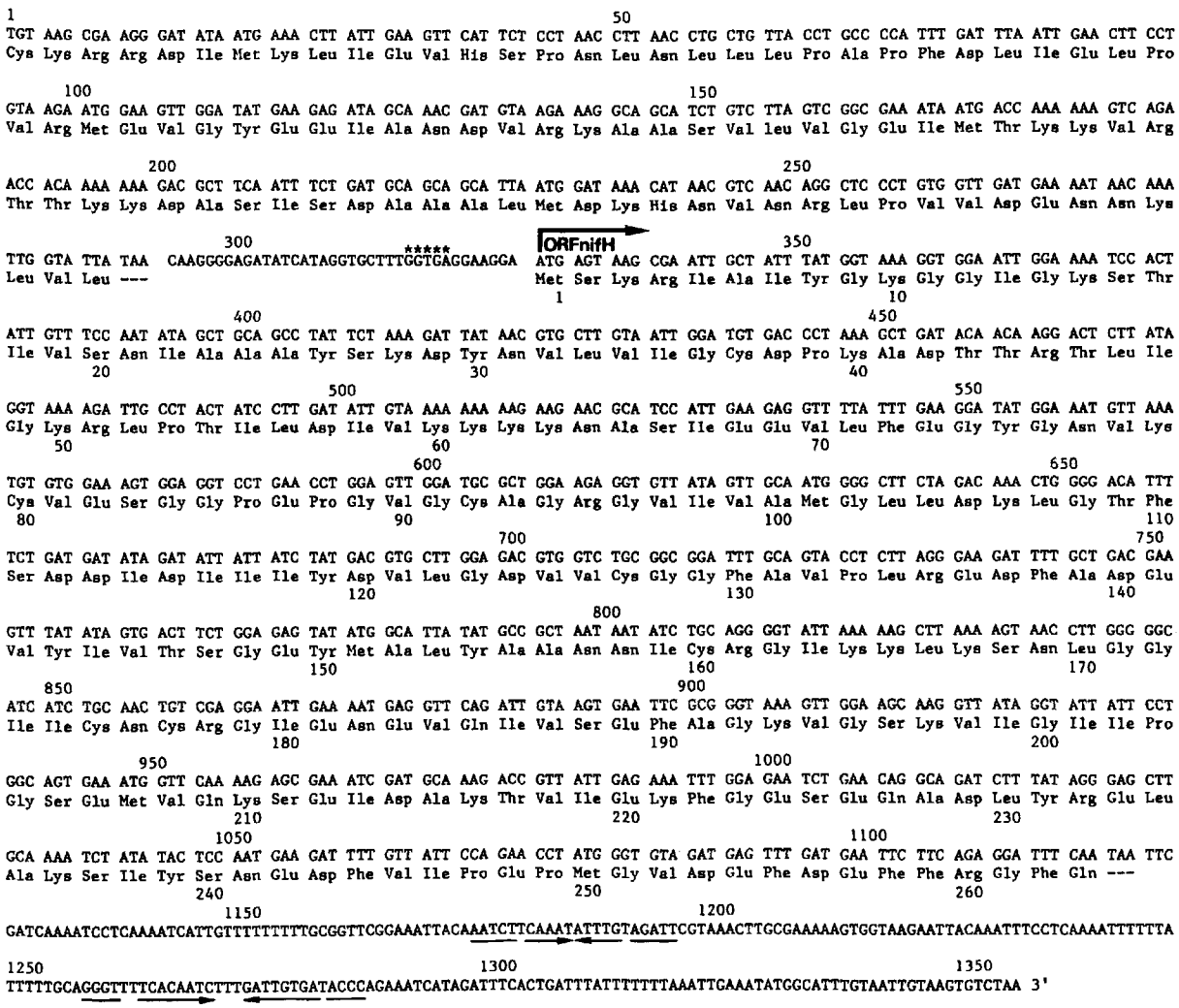

Fig. 3. Nucleotide sequence of *Mb. ivanovii* ORF*nifH* and surrounding regions. The noncoding DNA strand is shown. The deduced amino acid sequence of ORFs is shown below. Nucleotide positions are numbered above; amino acid positions of ORF*nifH* are numbered below. Potential ribosome-binding sites are indicated by stars. Palindromic sequences are indicated by arrows.

have been suggested to be 5′-GAANTTCA-3′ and 5′-TTTAATATAAA-3′ (Hamilton and Reeve 1985a). Such regions were not identified upstream from either ORF*nifH*. Another consensus sequence, 5′-ACCGAAAANTTTATATANTA(18–19 bases)TGCaagT-3′, found upstream from tRNA genes of *Methanococcus*, has been proposed to be a transcription signal (Wich et al. 1986). This consensus was not found in regions sequenced here. Similarly, the 5′-CTGGPyAPyPuNNNNTTGCA-3′ consensus sequence for *nifH* promoters (Dixon 1984) was not found.

Analysis of Amino Acid Sequences Deduced from ORFsnifH

The amino acid sequences deduced from ORFs*nifH* of *Mc. thermolithotrophicus* and *Mb. ivanovii* were aligned together with the amino acid sequence deduced from *Mc. voltae* ORF*nifH* (Fig. 4). To obtain a maximum sequence match, it was necessary to introduce a few gaps in the sequences. Highly conserved regions appear throughout the sequence, in particular up to position 180. Regions from positions 8 to 17, 106 to 114, 138 to 157, and 171 to 177 are entirely conserved in the three sequences. Cysteine residues 38, 95, 107, 143, and 176 are located at the same position in the three sequences. Interesting is the fact that the first four cysteine residues cited are found at the same position in all known *nifH* products (Fuhrmann and Hennecke 1984). By comparing all known *nifH* products, 30% of the residues were found to be conserved, especially regions surrounding the cysteine residues (Fuhrmann and Hennecke 1984). This is also valid for archaebacterial ORF*nifH*-deduced polypeptides, except for the *Mc. thermolithotrophicus nifH* product, in which the histidine, alanine, lysine, and tyrosine residues, occurring respectively at positions 42, 98, 102, and 105 (Fig. 4), are different from the

Table 1. Codon usage in *Mc. thermolithotrophicus* and *Mb. ivanovii* ORFs*nifH* compared to codon usage in *Mc. voltae, Anabaena,* and *C. pasteurianum nifH* genes

		Mc. thermolithotrophicus		*Mb. ivanovii*		*Mc. voltae*	*C. pasteurianum*	*Anabaena*
		Total	%	Total	%	%	%	%
Ala	***GCA***	12	70	8	47	59	46	30
	GCC	2	12	2	12	0	0	7
	GCG	0	0	1	6	12	0	4
	GCU	3	18	6	35	29	54	59
Arg	***AGA***	5	50	3	33	50	100	5
	AGG	3	30	4	45	42	0	5
	CGA	1	10	2	22	0	0	0
	CGC	1	10	0	0	0	0	0
	CGG	0	0	0	0	0	0	29
	CGU	0	0	0	0	8	0	61
Asn	AAC	8	50	4	40	31	41	87
	AAU	8	50	6	60	69	59	13
Asp	***GAC***	5	29	5	29	39	29	94
	GAU	12	71	12	71	61	71	6
Cys	UGC	2	29	4	57	0	8	66
	UGU	5	71	3	43	100	92	34
Gln	CAA	1	20	2	50	50	71	90
	CAG	4	80	2	50	50	29	10
Glu	GAA	21	91	15	71	77	94	88
	GAG	2	9	6	29	23	6	12
Gly	GGA	19	58	16	51	17	58	0
	GGC	1	3	3	10	13	3	10
	GGG	7	21	3	10	7	2	0
	GGU	6	18	9	29	63	37	90
His	CAC	0	0	0	0	100	25	100
	CAU	2	100	0	0	0	75	0
Ile	***AUA***	16	49	7	24	28	59	5
	AUC	7	21	6	21	20	29	68
	AUU	10	30	16	55	52	12	27
Leu	CUA	4	17	1	7	12	10	16
	CUC	3	12	0	0	12	0	8
	CUG	1	4	1	7	0	0	8
	CUU	4	17	10	66	18	35	8
	UUA	10	42	2	13	40	47	30
	UUG	2	8	1	7	18	8	30
Lys	AAA	24	86	14	67	83	67	43
	AAG	4	14	7	33	17	33	57
Met	AUG	9	100	5	100	100	100	100
Phe	***UUC***	1	25	3	27	33	73	100
	UUU	3	75	8	73	66	27	0
Pro	CCA	7	70	1	12	42	78	33
	CCC	0	0	0	0	8	0	23
	CCG	0	0	0	0	8	0	0
	CCU	3	30	7	88	42	22	44
Ser	AGC	2	22	2	13	0	12	17
	AGU	1	11	5	31	50	28	0
	UCA	2	22	0	0	0	44	8
	UCC	3	34	4	25	10	12	42
	UCG	1	11	0	0	0	4	0
	UCU	0	0	5	31	40	0	33
Thr	***ACA***	4	45	3	37	53	55	6
	ACC	2	22	1	13	20	4	88
	ACG	0	0	0	0	0	0	0
	ACU	3	33	4	50	27	41	6
Trp	UGG	0	0	0	0	0	0	0
Tyr	UAC	5	33	1	10	69	23	78
	UAU	10	64	9	90	31	77	22

Table 1. Continued

		Mc. thermolithotrophicus		*Mb. ivanovii*		*Mc. voltae*	*C. pasteurianum*	*Anabaena*
		Total	%	Total	%	%	%	%
Val	GUA	7	33	5	21	30	47	65
	GUC	1	5	1	4	8	3	0
	GUG	3	14	5	21	4	0	0
	GUU	10	48	13	54	58	50	35
Ter	UAA	1	100	1	100	100	100	0
	UAG	0	0	0	0	0	0	0
	UGA	0	0	0	0	0	0	100

References for *Mc. voltae, Anabaena,* and *C. pasteurianum* are Soillard and Sibold (1986), Mevarech et al. (1980), Chen et al. (1986), respectively. Ter: termination codon. Italicized bold codons indicate notable differences between methanogens and *Anabaena*

```
        1        10        20        30        40        50        60
Mct      LKQIAFYGKGGIGKSTTVCNIAAALADQGKKVMVVGCDPKHDCTSNLRGGQEIPTVLDI
Mbi     MSKRIAIYGKGGIGKSTIVSNIAAAYSKDYN-VLVIGCDPKADTTRTLIGKRL-PTILDI
Mcv     MRK-FCIYGKGGIGKSTNVGNMAAALAEDGKKVLVVGCDPKADSTRTLMHGK-INTVLDT

                70        80        90       100       110       120
Mct     LREKGLDKLGLETIIEKEMIEINDIIYEGYNGIYCVEAGGPKPGYGCAGRGVIVVIDLLK
Mbi     VKKKKNAS----------IEEVLFEGYGNVKCVESGGPEPGVGCAGRGVIVAMGLLD
Mcv     FRDKGPEY----------MKIEDIVYEGFNGVYCVESGGPEPGVGCAGRGVITAVDMLD

              130       140       150       160       170       180
Mct     KMNLYKDLKLDIVLYDVLGDVVCGGFAMPLRMGLAEQIYVVTSSDYMAIYAANNICRGIS
Mbi     KLGTFSDDI-DIIIYDVLGDVVCGGFAVPLREDFADEVYIVTSGEYMALYAANNICRGI-
Mcv     RLGVYDELKPDVVIYVILGDVVCGGFAMPLQKKLAEDYYIVTTCDPMAIYAANNICKGIK

              190       200       210       220       230       240
Mct     EFVKRGGSKLGGLIYNVRGSMDAYDIINEFADKLGANIVGKVPNSHLIPEAEIEGKTVIE
Mbi     ---KKLKSNLGGIICNCRGIENEVQIVSEFAGKVGSKVIGIIPGSEMVQKSEIDAKTVIE
Mcv     RYGNRGKIALGGIIYNGRSVVDEPEIIDKFVEGINSQVMGKVPMSNIITKAELRKQTTIE

              250       260       270       280       290
Mct     YDPNDEISQVYRELAKKIYENNEGTIPKPLENIEIMTIGKKIKERLKKERMKN
Mbi     KFGESEQADLYRELAKSIYSNEDFVIPEPMGVDEFDEFF----------RGFQ
Mcv     YAPDSEIANKFRELANSIYENKKTTIPTPLSEQGLDELTESIEELVR--RKYE
```

Fig. 4. Comparison of the amino acid sequences deduced from ORFs*nifH* from *Mc. thermolithotrophicus, Mc. voltae,* and *Mb. ivanovii*. Amino acids in bold characters are conserved in at least two of the three sequences. Mct, *Mc. thermolithotrophicus*; Mcv, *Mc. voltae*; Mbi, *Mb. ivanovii*.

alanine, serine, glutamate, and valine residues found in all other *nifH* polypeptide sequences. Another difference between the *Mc. thermolithotrophicus* ORF*nifH*-deduced polypeptide and all the other *nifH* products is the sequence found from positions 69 to 81, which does not match with any archaebacterial or eubacterial *nifH* product. Absence of histidine in the *Mb. ivanovii* ORF*nifH*-deduced polypeptide and of tryptophan in both polypeptides should also be noted.

The codon usage in both *nifH* genes is very similar (Table 1), and is in agreement with that found in *Mc. voltae* ORF*nifH*. There is a preference for codons ending in A or U and little use of codons containing GC or CG, as expected from the high A+T content. For example, ACA and ACU are more often used than ACC to specify threonine (ACG is never used), and CGG or CGC are rarely or never used to specify arginine. This unusual codon usage presents notable differences to that from the *Anabaena nifH* gene. For example, 61% of the arginine in the *Anabaena nifH* gene is coded by CGU, which is never used in *Mc. thermolithotrophicus* or *Mb. ivanovii* ORF*nifH*. Some other significant differences are observed for codons GCA, GCU, AGA, AGG, CGG, GAC, GAU, AUA, AUC, UUC, UUU, ACA, and ACC. On the contrary, codon usage appears to be better correlated with the very biased codon usage found in *C. pasteurianum nifH* sequences. This result may be explained by the low G+C content of the three organisms.

Relationship between Archaebacterial and Eubacterial nifH Sequences

Percent sequence matches were calculated for all eubacterial and archaebacterial *nifH* pairs of nu-

Table 2. Percent sequence matches calculated for *nifH* nucleotide sequences of different bacteria

	Mcv	Mbi	Cp1	Cp2	Av	Ac	Kp	An	Rj	RP	Rs	Rm	Rt	Rp
Mct	59	53	55	57	46	41	37	49	41	40	40	42	47	47
Mcv		58	57	58	51	48	49	51	50	48	44	48	47	49
Mbi			58	57	46	47	47	47	43	44	43	39	45	44
Cp1				90	57	52	54	58	53	53	51	51	52	50
Cp2					57	52	55	57	55	53	50	47	51	49
Av						88	78	68	72	72	72	70	69	72
Ac							80	67	70	71	72	72	70	74
Kp								65	70	71	72	67	68	70
An									63	61	63	63	62	64
Rj										88	78	70	70	72
RP											80	73	70	75
Rs 1 and 2												72	71	74
Rm													80	84
Rt														81

Organisms and references: *Mb. ivanovii* (Mbi), *Mc. thermolithotrophicus* (Mct), *Mc. voltae* (Mcv) (Souillard and Sibold 1986), *Clostridium pasteurianum* (Cp) (Chen et al. 1986), *Azotobacter vinelandii* (Av) (Brigle et al. 1985), *Azotobacter chroococcum* (Ac) (Robson et al. 1986), *Klebsiella pneumoniae* (Kp) (Scott et al. 1981; Sundaresan and Ausubel 1981), *Anabaena* (An) (Mevarech et al. 1980), *Bradyrhizobium japonicum* (Rj) (Fuhrmann and Hennecke 1984), *Rhizobium* sp. *Parasponia* (RP) (Scott et al. 1983a), *Rhizobium* ORS571 (Rs) (Norel and Elmerich 1987), *Rhizobium meliloti* (Rm) (Török and Kondorosi 1981), *Rhizobium trifolii* (Rt) (Scott et al. 1983b), *Rhizobium phaseoli* (Rp) (Quinto et al. 1985). The % sequence matches for two organisms, A and B, was: 2 × (number of identical nucleotides between A and B)/(number of total nucleotides in A + number of total nucleotides in B)

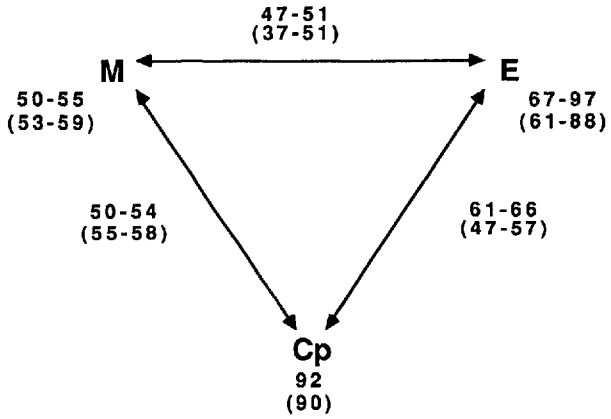

Fig. 5. Relationships between methanogens (M), *C. pasteurianum* (Cp), and the other eubacteria (E) as deduced from % sequence matches. Organisms and values are those of Tables 2 and 3. Numbers indicate extreme values based on amino acids found in the three groups and between different groups; numbers in brackets are values based on nucleotides.

cleotide sequences (Table 2) and amino acid sequences (Table 3). The organisms listed in Tables 2 and 3 fall into three groups: methanogens, *C. pasteurianum*, and the other eubacteria (Fig. 5). The main conclusions are as follows: (1) values based on amino acids are higher than values based on nucleotides, except for methanogen pairs and *C. pasteurianum*–methanogen pairs; (2) values between methanogens (50–59%) are lower than values found among eubacteria (61–97%) (*C. pasteurianum* excepted); (3) from values based on amino acids, *C. pasteurianum* appears closer to eubacteria (61–66%) than to methanogens (50–54%), but the difference between *C. pasteurianum* and methanogens is similar to differences within the methanogens themselves (50–55%). On the contrary, at the nucleotide level, *C. pasteurianum* appears to be more closely related to methanogens (55–58%) than to the other eubacteria (47%–57%).

Discussion

Nucleotide Sequence of nifH Regions with Sequence Similarity

The nucleotide sequences of DNA regions from *Mc. thermolithotrophicus* and *Mb. ivanovii* with sequence similarity to *nifH* confirm the sequence matches previously detected with *nifH* probes (Sibold et al. 1985; Possot et al. 1986). Expression of the genes at the mRNA level as well as the characterization of nitrogenase are under study and should definitely confirm the role of these sequences in nitrogen fixation. The putative translation initiation codons of *Mc. thermolithotrophicus* and *Mb. ivanovii* ORFs*nifH* were chosen because they are preceded by sequences complementary to the 3′ end of 16S rRNAs, which may act as ribosome-binding sites. These Shine and Dalgarno (1974)-like sequences were found in most sequenced genes from methanogens (for a review, see Reeve et al. 1986). For *Mc. thermolithotrophicus* ORF*nifH*, a UUG codon was predicted to be the translation initiation codon, because no AUG or GUG codon was found in the region following the Shine and Dalgarno-like sequence. In addition, the first conserved region in ORF*nifH* was found 15 bp downstream from the

Table 3. Percent sequence matches calculated for *nifH*-deduced amino acid sequences of different bacteria

	Mcv	Mbi	Cp1	Cp2	Av	Ac	Kp	An	Rj	RP	Rs	Rm	Rt	Rp
Mct	55	50	50	51	48	47	48	46	46	46	45	48	46	48
Mcv		50	52	54	50	50	51	50	49	49	49	49	49	49
Mbi			53	54	50	51	48	46	48	48	48	47	50	48
Cp1				92	66	65	64	61	61	61	61	62	61	62
Cp2					66	65	65	61	61	63	61	63	63	63
Av						89	88	71	75	74	73	68	69	70
Ac							85	72	74	71	71	71	73	73
Kp								70	74	74	75	67	68	70
An									70	70	72	69	70	71
Rj										97	92	77	78	79
RP											93	76	77	79
Rs 1 and 2												79	77	80
Rm													91	93
Rt														90

Same abbreviations as in Table 2. The % sequence matches for two organisms A and B was: 2 × (number of identical amino acids between A and B)/(number of total amino acids in A + number of total amino acids in B)

UUG codon, which is in agreement with the other *nifH* gene sequences. UUG translation initiation codons have already been found in other bacteria and may play a role in regulation of translation (for a review, see Marlière et al. 1987). Upstream and downstream from ORF*nifH*, no sequence matches were found between *Mc. thermolithotrophicus* and *Mb. ivanovii*. In both sequenced regions, ORFs surrounding ORF*nifH* were found. It is worth noting that several features of *nif* gene organization are different between eubacteria and methanogens. In eubacteria, *nifH* is often adjacent to *nifDK*, and the three genes are organized in a single operon transcribed from *nifH* to *nifK*. This is the case in *K. pneumoniae* (Elmerich et al. 1978; Merrick et al. 1978), *Azotobacter vinelandii* (Krol et al. 1981), *Azotobacter chroococcum* (Jones et al. 1984), *Azospirillum brasilense* (Perroud et al. 1985), *Rhodopseudomonas capsulata* (Avtges et al. 1983), *Gloeothece* (Kallas et al. 1983), and several *Rhizobium* species (Ruvkun et al. 1982; Schetgens et al. 1984; Donald et al. 1986; Denèfle et al. 1987). In *A. chroococcum*, a second copy of the *nifH* gene is found adjacent to a gene coding for a ferredoxin-like protein (Robson et al. 1986). For *Mc. thermolithotrophicus* and *Mb. ivanovii*, no sequence matches were detected between ORFs surrounding ORF*nifH* and *nifDK* or a gene coding for a ferredoxin. No sequence resembling the promoter consensus for archaebacteria was found upstream from either ORF*nifH*. This may indicate that the ORF immediately preceding ORF*nifH* and ORF*nifH* itself is cotranscribed in methanogens, whereas *nifH* genes from eubacteria are immediately preceded by a promoter. Oligo-T sequences and potential hairpin structures found in the 3′ noncoding regions following *Mc. thermolithotrophicus* and *Mb. ivanovii* ORFs*nifH* suggest that transcription is terminated immediately downstream to both ORFs*nifH*.

Comparison of Amino Acid Sequences

Many features common to previously described *nifH* gene products were found in the *Mc. thermolithotrophicus* and *Mb. ivanovii* ORF*nifH*-deduced polypeptides: (1) there is no tryptophan residue, like all *nifH* products except in *Rhizobium trifolii*; (2) positions of cysteine residues are conserved; (3) regions surrounding cysteine residues are highly conserved. Conservation of *nifH* products in eubacteria has been suggested to be due to the fact that the nitrogenase Fe protein interacts with electron donor proteins, with the MoFe protein, and with ATP (Eady and Smith 1979). Highly conserved regions may be necessary for all these interactions. Several regions of sequence similarity were found between nitrogenase Fe proteins and other ATP-binding proteins (Robson 1984). One of them contains the sequence GKGGIGKS, which is also found in the three ORF-*nifH* products of methanogens, from positions 9 to 16 (Fig. 4). This sequence is included in the G(four amino acid residues)GK(S/T) consensus sequence proposed as the ATP-binding site of ATP-binding proteins (Higgins et al. 1986).

There are no strong differences between predicted amino acid sequences of the two methanogenic polypeptides, despite the fact that *Mc. thermolithotrophicus* is a thermophile. Sequences of ORF*purE* from two other methanogens, *Mb. thermoautotrophicum* ΔH, a thermophilic bacterium, and *Methanobrevibacter smithii*, a mesophilic bacterium, have previously been compared (Hamilton and Reeve 1985b). The authors did not observe additional opportunities to form disulfide bridges in the enzyme

of the thermophile, which would increase thermal stability. However, they observed frequent changes of glycine to alanine, serine to alanine, and lysine to arginine when comparing the sequence of the mesophilic and the thermophilic enzymes. Such changes do not occur significantly in the predicted sequences of ORF*nifH* products from *Mb. ivanovii* and *Mc. thermolithotrophicus,* but % sequence matches between *Mc. thermolithotrophicus* and *Mb. ivanovii* or *Mc. thermolithotrophicus* and *Mc. voltae* are lower than between *Mb. ivanovii* and *Mc. voltae,* which are two mesophilic bacteria. More remarkable is the presence of a sequence of 12 amino acid residues in the predicted polypeptide from *Mc. thermolithotrophicus,* from positions 69 to 81 (Fig. 4), which does not appear in any other known *nifH* sequence. An interesting speculation is that this sequence could be an intron or the vestige of an intron. It is indeed surrounded by the sequences 5′-AG↑GTTTAG-3′ and 5′-TAG↑AA-3′, which are reminiscent of the consensus for splice sites in eukaryotic introns: 5′-AG↑GT(A/G)AGT-3′ and 5′-(C/T)AG↑G(G/T)-3′ (Breathnach and Chambon 1981). The putative splice sites indicated by arrows in the *Mc. thermolithotrophicus* sequence are located at positions corresponding to splice sites in the eukaryotic consensus sequences. Up to now, no introns have been found in methanogens. In other archaebacteria, several introns have been found, but not in genes coding for proteins (for a review, see Dennis 1986).

Sequence Divergence and Evolution

Percent sequence matches between *nifH*-deduced polypeptides from methanogens and eubacteria are consistent with the distant phylogenetic position of archaebacteria and with the absence of recent lateral transfer of *nif* genes between the two kingdoms. However, the low values found between methanogens and the particular position of *C. pasteurianum* deserve further comment. Methanogens appear to be as distant from each other as they are from eubacteria. This unexpected result can be explained by the fact that Methanomicrobiales and Methanococcales are relatively distant from each other. The S_{AB} values based on comparison of 16S rRNA between the two groups of methanogens are about 0.25, whereas they are about 0.1 between eubacteria and archaebacteria. According to this comparison, Methanomicrobiales and Methanococcales are as distant from each other as, for example, cyanobacteria from enterobacteria. It is therefore not surprising that the % sequence matches based on *nifH* comparison are relatively low between *Mc. thermolithotrophicus* and *Mb. ivanovii.* On the contrary, the S_{AB} values based on 16S rRNA comparison between *Mc. thermolithotrophicus* and *Mc. voltae* are about 0.45 (these two bacteria belong to the same genus), but their ORFs*nifH* are also relatively distant. However, as nitrogen fixation is not observed with *Mc. voltae,* the ORF*nifH* from *Mc. voltae* may not be expressed and would therefore not be under the same selective pressure as a functional *nifH* gene, and could have evolved differently. In order to compare ORFs*nifH* from the three branches of methanogens, it may be of interest to establish the sequence of the ORF*nifH* from the nitrogen-fixing Methanomicrobiale *Ms. barkeri.*

Clostridium pasteurianum seems to have a unique position, intermediate between the other eubacteria and methanogens. The % sequence matches at the amino acid level clearly group this bacterium with the other eubacteria. However, at the nucleotide level, *C. pasteurianum* seems to be closer to methanogens. Two biases could account for this result: (1) the high A+T content of both *C. pasteurianum* and methanogens; (2) the short length of the ORFs-*nifH* of *C. pasteurianum, Mc. voltae,* and *Mb. ivanovii.* However, one cannot exclude an ancient lateral transfer of *nif* genes between *Clostridium* and methanogens which are found in the same type of habitats. Hennecke et al. (1985) has suggested that the % sequence matches at the amino acid level between *C. pasteurianum* and *Anabaena* (60%) represented the lower limit of conservation necessary to preserve the properties of the Fe protein. It seems that this limit was overestimated, since the value of 50% was found in the case of methanogens.

Acknowledgments. We thank Professor J.-P. Aubert for helpful discussions, Dr. A. Edelman for critical reading of the manuscript, and Mr. M. Henriquet for technical assistance. This work was supported by research contracts from La Société Lyonnaise des Eaux et de l'Eclairage, Le Pecq, France, and from La Société Nationale Elf Aquitaine, Paris La Défense, France, by grant no. 85 T. 0704 from the Ministère de la Recherche et de la Technologie, and by research funds from the University Paris 7.

References

Avtges P, Scolnik PA, Haselkorn R (1983) Genetic and physical map of the structural genes (*nifHDK*) coding for the nitrogenase complex of *Rhodopseudomonas capsulata.* J Bacteriol 156:251–256

Balch WE, Fox GE, Magrum LJ, Woese CR, Wolfe RS (1979) Methanogens: reevaluation of a unique biological group. Microbiol Rev 43:260–293

Belay N, Sparling R, Daniels L (1984) Dinitrogen fixation by a thermophilic methanogenic bacterium. Nature 312:286–288

Belyaev SS, Wolkin R, Kenealy WR, De Niro MJ, Epstein S, Zeikus JG (1983) Methanogenic bacteria from the Bondyuzhskoe oil field: general characterization and analysis of stable-carbon isotopic fractionation. Appl Environ Microbiol 45:691–697

Bomar M, Knoll K, Widdel F (1985) Fixation of molecular

nitrogen by *Methanosarcina barkeri*. FEMS Microbiol Ecol Lett 31:47–55

Breathnach R, Chambon P (1981) Organization and expression of eukaryotic split genes coding for proteins. Annu Rev Biochem 50:349–383

Brigle KE, Newton WE, Dean DR (1985) Complete nucleotide sequence of the *Azotobacter vinelandii* nitrogenase structural gene cluster. Gene 37:37–44

Cannon FC, Riedel GE, Ausubel FM (1979) Overlapping sequences of *K. pneumoniae nif* DNA cloned and characterized. Mol Gen Genet 174:59–66

Chen KCK, Chen JS, Johnson JL (1986) Structural features of multiple *nifH*-like sequences and very biased codon usage in nitrogenase genes of *Clostridium pasteurianum*. J Bacteriol 166:162–172

Denèfle P, Kush A, Norel F, Paquelin A, Elmerich C (1987) Biochemical and genetic analysis of the *nifHDKE* region of *Rhizobium* ORS571. Mol Gen Genet 207:280–287

Dennis PP (1986) Molecular biology of archaebacteria. J Bacteriol 168:471–478

Dixon R (1984) The genetic complexity of nitrogen fixation. J Gen Microbiol 130:2745–2755

Donald RGK, Nees DW, Raymond CK, Loroch AI, Ludwig RA (1986) Characterization of three genomic loci encoding *Rhizobium* sp. strain ORS571 N_2 fixation genes. J Bacteriol 165: 72–81

Eady RR, Smith BE (1979) Physico-chemical properties of nitrogenase and its components. In: Hardy RWF, Bottomley F, Burns RC (eds) A treatise on dinitrogen fixation, sections I and II. Wiley and Sons, London, pp 399–490

Elmerich C, Houmard J, Sibold L, Manheimer I, Charpin N (1978) Genetic and biophysical analysis of mutants induced by bacteriophage Mu DNA integration into *Klebsiella pneumoniae* nitrogen fixation genes. Mol Gen Genet 165:181–189

Frischauf AM, Lehrach H, Poustka A, Murray N (1983) Lambda replacement vectors carrying polylinker sequences. J Mol Biol 170:827–842

Fuhrmann M, Hennecke H (1984) *Rhizobium japonicum* nitrogenase Fe protein gene (*nifH*). J Bacteriol 158:1005–1011

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

Hamilton W, Reeve JN (1985b) Sequence divergence of an archaebacterial gene cloned from a mesophilic and a thermophilic methanogen. J Mol Evol 22:351–360

Hennecke H, Kaluza K, Thöny B, Fuhrmann M, Ludwig W, Stackebrandt E (1985) Concurrent evolution of nitrogenase genes and 16S rRNA in *Rhizobium* species and other nitrogen fixing bacteria. Arch Microbiol 142:342–348

Higgins CF, Hiles ID, Salmond GPC, Gill DR, Downie JA, Evans IJ, Holland IB, Gray L, Buckel SD, Bell AW, Hermodson MA (1986) A family of related ATP-binding subunits coupled to many distinct biological processes in bacteria. Nature 323:448–453

Huber H, Thomm M, König H, Thies G, Stetter KO (1982) *Methanococcus thermolithotrophicus*, a novel thermophilic lithotrophic methanogen. Arch Microbiol 132:47–50

Jain MK, Thompson TE, Conway de Macario E, Zeikus JG (1987) Speciation of *Methanobacterium* strain Ivanov as *Methanobacterium ivanovii*, sp. nov. System Appl. Microbiol 9:77–82

Jones R, Woodley P, Robson R (1984) Cloning and organisation of some genes for nitrogen fixation for *Azotobacter chroococcum* and their expression in *Klebsiella pneumoniae*. Mol Gen Genet 197:318–327

Kallas T, Rebiere MC, Rippka R, Tandeau de Marsac N (1983) The structural *nif* genes of the cyanobacteria *Gloeothece* sp. and *Calothrix* sp. share homology with *Anabaena* sp. but the *Gloeothece* genes have a different arrangement. J Bacteriol 155:427–431

König H, Nusser E, Stetter KO (1985) Glycogen in *Methanolobus* and *Methanococcus*. FEMS Microbiol Lett 28:265–269

Krol ADM, Hondelez JGJ, Roozendaal B, van Kammen A (1981) On the operon structure of the nitrogenase genes of *Rhizobium leguminosarum* and *Azotobacter vinelandii*. Nucleic Acids Res 10:4147–4156

Lipman WJ, Wilbur DJ (1983) Rapid similarity searches of nucleic acid and protein data banks. Proc Natl Acad Sci USA 80:726–730

Magot M, Possot O, Souillard N, Henriquet M, Sibold L (1986) Structure and expression of *nif* (nitrogen fixation) genes in methanogens. In: Dubourguier HC, Albagnac G, Montreuil J, Romond C, Sautiere P, Guillaume J (eds) Biology of anaerobic bacteria. Elsevier, Amsterdam, pp 193–199

Maniatis T, Fritsch E, Sambrook J (1982) Molecular cloning, a laboratory manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor NY

Marlière P, Kofoid EC, Ganoza MC (1987) Can the T loop of initiator tRNA act as a second anticodon? Ann Inst Pasteur (Microbiology) (in press)

Mevarech M, Rice D, Haselkorn R (1980) Nucleotide sequence of a cyanobacterial *nifH* gene coding for nitrogenase reductase. Proc Natl Acad Sci USA 77:6476–6480

Merrick M, Filser M, Kennedy C, Dixon R (1978) Polarity of mutations induced by insertion of transposon Tn*5*, Tn*7* and Tn*10* into the *nif* gene cluster of *Klebsiella pneumoniae*. Mol Gen Genet 165:103–111

Murray PA, Zinder S (1984) Nitrogen fixation by a methanogenic archaebacterium. Nature 312:284–286

Norel F, Elmerich C (1987) Nucleotide sequence and functional analysis of the two *nifH* copies of *Rhizobium* ORS571. J Gen Microbiol 133:1563–1576

Perroud B, Bandhari SK, Elmerich C (1985) The *nifHDK* operon of *Azospirillum brasilense* Sp7. In: Klingmüller W (ed) Azospirillum III: genetics, physiology, ecology. Springer, Berlin, pp 10–19

Possot O, Henry M, Sibold L (1986) Distribution of DNA sequences homologous to *nifH* among archaebacteria. FEMS Microbiol Lett 34:173–177

Quinto C, de la Vega H, Flores M, Leemans J, Cevallos MA, Pardo MA, Azpiroz R, de Lourdes Girard M, Calva E, Palacios R (1985) Nitrogenase reductase: a functional multigene family in *Rhizobium phaseoli*. Proc Natl Acad Sci USA 82:1170–1174

Reeve JN, Hamilton PT, Beckler GS, Morris CJ, Clarke CH (1986) Structure of methanogens genes. Syst Appl Microbiol 7:5–12

Rice D, Mazur BJ, Haselkorn R (1982) Isolation and physical mapping of nitrogen fixation genes from the cyanobacterium *Anabaena* 7120. J Biol Chem 257:13157–13163

Robson RL (1984) Identification of possible adenine nucleotide-binding sites in nitrogenase Fe- and MoFe-proteins by amino acid sequence comparison. FEBS Lett 173:394–398

Robson R, Woodley P, Jones R (1986) Second gene (*nifH**) coding for a nitrogenase iron protein in *Azotobacter chroococcum* is adjacent to a gene coding for a ferredoxin-like protein. EMBO J 5:1159–1163

Ruvkun GB, Ausubel FM (1980) Interspecies homology of nitrogenase genes. Proc Natl Acad Sci USA 77:191–195

Ruvkun GB, Sundaresan V, Ausubel FM (1982) Directed transposon Tn*5* mutagenesis and complementation analysis of *Rhizobium meliloti* symbiotic nitrogen fixations genes. Cell 29:551–555

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain terminating inhibitors. Proc Natl Acad Sci USA 74: 5463–5467

Schetgens TMP, Bakkeren G, van Dun C, Hontelez JGJ, van Kammen A, van den Bos RC (1984) Identification and analysis of the expression of *Rhizobium leguminosarum* PRE symbiotic genes. In: Veeger C, Newton WE (eds) Advances in nitrogen fixation research. Martinus Nijhoff, Dr W Junk, The Hague, Netherlands, p 699

Scott KF, Rolfe GB, Shine J (1981) Biological nitrogen fixation: primary structure of the *Klebsiella pneumoniae nifH* and *nifD* genes. J Mol Appl Genet 1:71–81

Scott KF, Rolfe GB, Shine J (1983a) Nitrogenase structural genes are unlinked in the nonlegume symbiot Parasponia *Rhizobium*. DNA 2:141–148

Scott KF, Rolfe GB, Shine J (1983b) Biological nitrogen fixation: primary structure of the *Rhizobium trifolii* iron protein gene. DNA 2:149–155

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

Sibold L, Pariot D, Bhatnagar L, Henriquet M, Aubert JP (1985) Hybridization of DNA from methanogenic bacteria with nitrogenase structural genes (*nifHDK*). Mol Gen Genet 200: 40–46

Souillard N, Sibold L (1986) Primary structure and expression of a gene homologous to *nifH* (nitrogenase Fe protein) from the archaebacterium *Methanococcus voltae*. Mol Gen Genet 203:21–28

Sundaresan VK, Ausubel FM (1981) Nucleotide sequence of the gene coding for nitrogenase iron protein from *Klebsiella pneumoniae*. J Biol Chem 256:2808–2812

Török I, Kondorosi A (1981) Nucleotide sequence of the *R. meliloti* nitrogenase reductase (*nifH*) gene. Nucleic Acids Res 9:5711–5723

Vieira J, Messing J (1982) The pUC plasmids, an M13mp7-derived system for insertion mutagenesis and sequencing with synthetic universal primers. Gene 129:259–268

Wain-Hobson S, Sonigo P, Danos O, Cole S, Alizon M (1985) Nucleotide sequence of the AIDS virus, LAV. Cell 40:9–17

Wich G, Hummel H, Jarsch M, Bar U, Böck A (1986) Transcription signals for stable RNA genes in *Methanococcus*. Nucleic Acids Res 14:2459–2478

Whitman WB (1985) Methanogenic bacteria. In: Woese CR, Wolfe RS (eds) Archaebacteria. Academic Press, New York, pp 4–84

Received April 16, 1987/Revised July 22, 1987