Hermann, R. B. (1977) *Proc. Natl. Acad. Sci. U.S.A. 74*, 4144.
Hildebrand, J. H. (1947) *J. Chem. Phys. 15*, 225-228.
Hine, J., & Mookerjee, P. K. (1975) *J. Org. Chem. 40*, 292.
Huggins, M. L. (1941) *J. Chem. Phys. 9*, 440-449.
Kellis, J. T., Nyberg, K., & Fersht, A. R. (1989) *Biochemistry 28*, 4914-4922.
Kyte, J., & Doolittle, R. F. (1982) *J. Mol. Biol. 157*, 105.
Lee, B., & Richards, F. M. (1971) *J. Mol. Biol. 55*, 379-400.
Lee, B. K. (1985) *Biopolymers 24*, 813-823.
Lide, D. R., Ed. (1990) *CRC Handbook of Chemistry and Physics*, CRC Press, Boca Raton, FL.
Lim, W. A., & Sauer, R. T. (1989) *Nature 339*, 31-36.
Masterton, W. L. (1954) *J. Phys. Chem. 22*, 1830-1833.
Matsumura, M., Becktel, W. J., & Matthews, B. W. (1988) *Nature 334*, 406.
McAuliffe, C. (1966) *J. Phys. Chem. 70*, 1267.
McQuarrie, D. (1976) *Statistical Mechanics*, Harper & Row, New York.
Mezei, M., & Beveridge, D. L. (1986) *Computer Simulations and Biomolecular Systems, Ann. N.Y. Acad. Sci. 494*, 1-23.
Nakai, K., Kidera, A., & Kanehisa, M. (1988) *Protein Eng. 2*, 93.
Nicholls, A., Sharp, K. A., & Honig, H. (1991) *Proteins* (in press).
Nozaki, Y., & Tanford, C. H. (1971) *J. Biol. Chem. 246*, 2211.
Radzicka, A., & Wolfenden, R. (1988) *Biochemistry 27*, 1664-1670.
Richards, F. M. (1974) *J. Mol. Biol. 82*, 1-14.
Sandberg, W. S., & Terwilliger, T. C. (1991) *Proc. Natl. Acad. Sci. U.S.A. 88*, 1706-1710.
Sharp, K. A., & Nicholls, A. (1990) DELPHI 3.0, available from the Department of Biochemistry and Molecular Biophysics, Columbia University, New York, or from Biosym Corp., San Diego, CA.
Sharp, K. A., Nicholls, A., Fine, R. M., & Honig, B. (1991) *Science 252*, 106-109.
Shortle, D., Stites, W. E., & Meeker, A. K. (1990) *Biochemistry 29*, 8033-8041.
Tanford, C. H. (1980) *The Hydrophobic Effect*, John Wiley and Sons, New York.
Wolfenden, R., Andersson, L., Cullis, P. M., & Southgate, C. C. (1981) *Biochemistry 20*, 849-855.
Wood, R. H., & Thompson, P. T. (1990) *Proc. Natl. Acad. Sci. U.S.A. 87*, 946-949.

Primary Structure of Hydrogenase I from *Clostridium pasteurianum*[†]

Jacques Meyer*,[‡] and Jean Gagnon[§]
DBMS-Métalloprotéines and DBMS-Biologie Structurale, CNRS URA 1333, CENG 85X, 38041 Grenoble, France

Received May 2, 1991; Revised Manuscript Received July 1, 1991

ABSTRACT: Peptides obtained by cleavage of *Clostridium pasteurianum* hydrogenase I have been sequenced. The data allowed design of oligonucleotide probes which were used to clone a 2310-bp *Sau*3A fragment containing the hydrogenase encoding gene. The latter has been sequenced and was found to translate into a protein composed of 574 amino acids (M_r = 63 836), including 22 cysteines. *C. pasteurianum* hydrogenase is homologous to, but longer than, the large subunit of *Desulfovibrio vulgaris* (Hildenborough) [Fe] hydrogenase. It includes an additional N-terminal domain of ca. 110 amino acids which contains eight cysteine residues and which therefore could accommodate two of its postulated four [4Fe-4S] clusters. *C. pasteurianum* hydrogenase is most similar in length, cysteine positions, and sequence altogether to the translation product of a putative hydrogenase encoding gene from *D. vulgaris* (Hildenborough). Comparisons of the available [Fe] hydrogenase sequences show that these enzymes constitute a structurally rather homogeneous family. While they differ in the length of their N-termini and in the number of their [4Fe-4S] clusters, they are highly similar in their C-terminal halves, which are postulated to harbor the hydrogen-activating H cluster. Five conserved cysteine residues occurring in this domain are likely ligands of the H cluster. Possible ligation by other residues, and in particular by methionine, is discussed. The comparisons carried out here show that the H clusters most probably possess a common structural framework in all [Fe] hydrogenases. On the basis of the available data on these proteins and on the current developments in iron-sulfur chemistry, the H clusters possibly contain six to eight iron atoms. The large differences between the DNA compositions of *C. pasteurianum* (30% G + C) and *D. vulgaris* (65% G + C) result in significant differences, not only in codon usage for a given amino acid but also in the amino acid compositions of their respective hydrogenases.

Hydrogenases are iron-sulfur enzymes that catalyze the reaction $H_2 \leftrightarrow 2H^+ + 2e^-$. They are divided into two main groups, on the basis of metal content and sequence homology: those containing only iron ([Fe] hydrogenases) (Adams, 1990) and those containing iron and nickel ([NiFe] hydrogenases) (Fauque et al., 1988). A few members of the second group also contain selenium (Fauque et al., 1988). The widespread distribution of [NiFe] hydrogenases in the microbial world is reflected in the significant number of species from which hydrogenase genes have been cloned and sequenced (Voordouw et al., 1989a; Leclerc et al., 1988; Sayavedra-Soto et al., 1988; Reeve et al., 1989; Menon et al., 1990a,b; Uffen et al., 1990; Tran-Betcke et al., 1990; Alex et al., 1990; Rousset et al., 1990;

[†] The nucleotide sequence reported in this paper has been submitted to GenBank under Accession Number M62754.
[‡] DBMS-Métalloprotéines.
[§] DBMS-Biologie Structurale.

0006-2960/91/0430-9697$02.50/0

Deckers et al., 1990; Ford et al., 1990; Schneider et al., 1990). In contrast, [Fe] hydrogenases appear to be more scarcely distributed, and the only known sequences are those from two strains of *Desulfovibrio vulgaris* (Voordouw & Brenner, 1985; Voordouw et al., 1989b).

The anaerobe *Clostridium pasteurianum* contains two [Fe] hydrogenases, each consisting of a single polypeptide chain. Hydrogenase I, one of the first hydrogenases purified to homogeneity (Chen & Mortenson, 1974), has been thoroughly investigated by biochemical and spectroscopic methods (Adams, 1990). Its amino acid composition and iron–sulfur content have recently been reevaluated (Adams et al., 1989). Despite the host of structural data available on hydrogenase I, a most important piece of structural information, namely, the amino acid sequence, has so far remained missing. Its elucidation would piece together and substantiate a number of other data (Adams et al., 1989; Adams, 1990). It would also clarify uncertainties concerning the degree of similarity between the [Fe] hydrogenase from *D. vulgaris* and *C. pasteurianum* hydrogenase I, which are spectroscopically (Adams, 1990) and immunologically (Kovacs et al., 1989) similar, although their genes appear to be unrelated [quoted in Adams (1990)]. More generally, the sequence of a clostridial hydrogenase would allow an estimation of the structural diversity of [Fe] hydrogenases, as the available sequences would then encompass clostridia and sulfate reducers. Comparisons of [Fe] hydrogenases from diverse bacteria are also expected to highlight common sequence patterns which are likely to be functionally important. As a contribution toward the elucidation of these questions, we report here and discuss the sequence of the gene encoding hydrogenase I from *C. pasteurianum*.

Materials and Methods

Materials and Strains. *C. pasteurianum* W5 (ATCC 6013) was obtained from the American Type Culture Collection. Competent *Escherichia coli* DH5α cells, T4 DNA ligase, and T4 polynucleotide kinase were from Bethesda Research Laboratories. Other enzymes were purchased from Boehringer Mannheim. pUC18 plasmid was from Pharmacia. *E. coli* alkaline phosphatase, [γ-^{32}P]ATP, [α-^{35}S]dATPαS, and Hybond-N membranes were obtained from Amersham. Oligonucleotides were synthesized on a 381A Applied Biosystems machine, and those to be used as probes were end-labeled either with [γ-^{32}P]ATP or with Digoxigenin-11–dUTP (Boehringer Mannheim).

Hydrogenase Purification and Peptide Sequencing. *C. pasteurianum* cells were grown on N_2 (Rabinowitz, 1972), harvested, resuspended in 2 volumes of Tris-HCl (0.05 M), pH 8.5, and broken with lysozyme (0.3 mg/g wet weight, 1 h at 37 °C). After centrifugation at 20000*g* for 30 min, the supernatant was fractionated by poly(ethylene glycol) precipitation (Tsö et al., 1972). Hydrogenase I precipitated with 30% poly(ethylene glycol) and was further purified as described (Chen & Mortenson, 1974). Apohydrogenase was prepared by overnight dialysis against 0.01 M acetic acid, lyophilized, and carboxymethylated (Crestfield et al., 1963). Carboxymethyl-hydrogenase was freed from residual protein impurities by reverse-phase HPLC on an Aquapore Butyl BU300 (220 mm × 4.6 mm) column (Applied Biosystems) equilibrated with 0.1% trifluoroacetic acid in water and developed with an increasing concentration of acetonitrile containing 0.1% trifluoroacetic acid. The purified alkylated hydrogenase was cleaved with cyanogen bromide (100-fold molar excess of CNBr in 70% formic acid, 24 h at room temperature in the dark) or with trypsin. The resulting peptides were purified by reverse-phase HPLC on an Aquapore Octyl RP300 (220 mm × 4.6 mm) column (Applied Biosystems) equilibrated with 0.1% trifluoroacetic acid in water and developed with an increasing concentration of acetonitrile containing 0.1% trifluoroacetic acid. A Model 477A sequencer, equipped with on-line detection Model 120A (Applied Biosystems), was used for sequence analysis.

Cloning and DNA Sequencing. Genomic DNA of *C. pasteurianum* was purified from N_2 grown cells as described (Saito & Miura, 1963), with a slight modification (Graves et al., 1985). Digested DNA was electrophoresed through agarose gels in 0.04 M Tris–acetate/2 mM EDTA, pH 8.5, and stained with ethidium bromide (0.3 μg/mL). DNA transfer from agarose gels onto Hybond-N membranes was carried out with a Vacugene (LKB) equipment. Prehybridization was for a minimum of 1 h at 40 °C, in 0.1% sodium dodecyl sulfate/2× Denhardt's solution/0.75 M NaCl/0.075 M sodium citrate/0.1 mg/mL of sonicated and boiled herring sperm DNA. Hybridization was overnight at the same temperature, in the same solution supplemented with labeled oligonucleotide probe, and was followed by three 20-min washes at 40 °C in 0.3 M NaCl/0.03 M sodium citrate. Genomic *C. pasteurianum* DNA was digested and, without further fractionation, inserted (2.5:1 ratio of insert to vector) into pUC18 vector that had previously been digested with the same enzyme or with a compatible enzyme and treated with *E. coli* alkaline phosphatase. The recombinant plasmid was used to transform *E. coli* DH5α, and the cells were spread on LB plates supplemented with ampicillin (50 μg/mL), X-gal[1] (20 μg/mL), and IPTG (24 μg/mL). Tranformant colonies were transferred to Hybond-N membranes, by either replica plating or individual transfer with toothpicks. Membranes were treated as indicated by the manufacturer and screened by colony hybridization as described above for genomic DNA hybridization. Plasmid DNA was isolated by alkaline lysis (Kieser, 1984) and sequenced by the dideoxy method (Sanger et al., 1977) with the Sequenase 2.0 kit (U.S. Biochemicals). DNA and protein sequences were analyzed using the DNASTAR software package.

Results

Protein Sequencing and Design of Oligonucleotide Probes. The very different base compositions of *C. pasteurianum* (G + C = 30%; Tonomura et al., 1965) and *D. vulgaris* (G + C = 65%; Postgate, 1984) DNA suggested that the gene encoding *C. pasteurianum* hydrogenase I might not be straightforwardly cloned using the *D. vulgaris* [Fe] hydrogenase gene as a probe, even if the two proteins were highly similar. We therefore decided to use synthetic oligonucleotide probes designed on the basis of partial protein sequence.

Peptides were generated from carboxymethyl-hydrogenase either by cyanogen bromide or by trypsin cleavage and purified as described under Materials and Methods. Several of these peptides were sequenced, giving eight nonoverlapping sequence fragments (Figure 2), four from tryptic peptides and four from cyanogen bromide generated peptides. The initial sequencing yields of the peptides were in the 350–1200 pmol range, with 3–4 nmol of protein as starting material for each of the cleavage reactions. Thus, the combined yields of protein cleavage, peptide purification, and sequencing were in the 10–35% range. N-Terminal sequencing of the protein required

[1] Abbreviations: bp, base pair; IPTG, isopropyl 1-thio-β-D-galactopyranoside; kb, kilobase; X-gal, 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside.

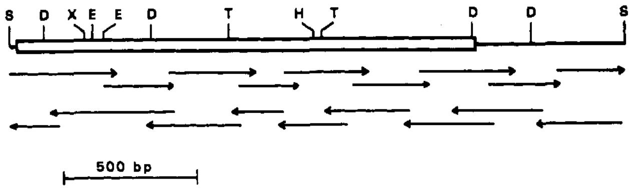

FIGURE 1: Restriction map and sequencing scheme of the 2310-bp *Sau*3A DNA fragment carrying the hydrogenase I gene (open rectangle) of *C. pasteurianum*. Restriction enzymes used: S, *Sau*3A; D, *Dra*I; X, *Xba*I; E, *Eco*RI; T, *Taq*I; H, *Hae*III. The 4.3-kb *Eco*RI fragment mentioned in the text starts at the second *Eco*RI site from the 5′-end (position 352) and extends ca. 2.4 kb beyond the 3′-end of the 2310-bp *Sau*3A fragment.

removal of the blocking group by treatment with trifluoroacetic acid (33% in water) for 1 h at 55 °C (Levy et al., 1981). This allowed unambiguous determination of the 25 N-terminal residues (Figure 2). The combined protein and peptide sequence data (159 residues altogether) opened a wide choice for oligonucleotide probes. Among these, a selection was made for low degeneracy and high specificity. The degeneracy of the oligonucleotide probes was lowered by taking into account the codon usage of *C. pasteurianum*, which is biased toward A or T in the wobble position (Graves et al., 1985; Chen et al., 1986; Hinton et al., 1987; Wang et al., 1988, 1990). Specificity was optimized by scanning data banks with potential probe sequences and eliminating those that were likely to hybridize with unwanted genes. At the outcome, the peptide sequence fragments GPFPMFT, GGVMEAA, and DFAENA (see Figure 2) were used to design the oligonucleotide probes 5′GT(AG)AACAT(AT)GG(AG)AA(AT)GG(AT)CC3′(P1), 5′GC(AT)GC(TC)TCCAT(AT)AC(AT)CC(AT)CC3′(P2), and 5′GC(AG)TT(TC)TC(AT)GC(AG)AA(AG)TC3′(P3), respectively. All three probes were made to hybridize with the nontranscribed DNA strand, the sequence of which is shown in Figure 2.

Cloning of the Hydrogenase I Gene. Genomic DNA of *C. pasteurianum* was digested with several enzymes cutting the multiple cloning site of pUC18 and analyzed by Southern blotting with the three oligonucleotide probes. Three restriction reactions yielded strongly hybridizing fragments of suitable size: *Xba*I, 7–8 kb; *Eco*RI, 4.3 kb; and *Sau*3A, 2.3 kb. For a given restriction reaction, the same pattern was obtained with all of the three probes, and only one hybridizing fragment was observed in each case. *Eco*RI/*Sau*3A double digestion produced a 1.9-kb hybridizing fragment, which showed that the 4.3-kb *Eco*RI and the 2.3-kb *Sau*3A fragments are partially overlapping. Total genomic *C. pasteurianum* DNA was digested with *Sau*3A and inserted into pUC18 cut with *Bam*HI. Cells transformed with the ligation mixture were plated and screened by colony hybridization with probe P1. One positive clone was obtained out of ca. 1500 transformants. Restriction analysis showed that the positive clone had an insert of the expected size and preliminary runs of sequencing with M13/pUC primers indicated that the hydrogenase gene was totally encompassed by the 2.3-kb *Sau*3A fragment (Figure 1). The latter fragment was therefore completely sequenced on both strands. Its nucleotide sequence and the translated amino acid sequence of hydrogenase I are shown in Figure 2.

DNA Sequence. The nucleotide sequence upstream of the hydrogenase gene includes a strong ribosome binding site AGGAGG, but it is too short (30 bp) to contain promoter regions. Downstream (ca. 20 bp) of the TAA termination codon, a clearcut transcription termination signal occurs, which consists of a 21-bp stem and 7-base loop (Figure 2). No additional open reading frames have been detected downstream of the hydrogenase gene. The G + C content of the sequence fragment was found to be 29%, in line with the low G + C content of *C. pasteurianum* DNA (Tonomura et al., 1965). As observed for other *C. pasteurianum* genes (Graves et al., 1985; Mathieu, Meyer, and Moulis, submitted for publication), the G + C content of the coding region (32%) is significantly higher than that of total DNA.

Table I: Amino Acid Composition of Hydrogenase I from *C. pasteurianum*

amino acid	analysis[a]	sequence[b]
N + D	79	76
T	33	30
S	23	27
Q + E	57	58
P	19	20
G	38	39
A	42	45
C	22	22
V	33	34
M	16	18
I	43	41
L	42	38
Y	13	16
F	25	26
H	9	10
K	50	52
R	21	21
W	9	1

[a] From Adams et al. (1989). Values originally given for 551 residues have been recalculated for 574 residues. [b] From this work.

Amino Acid Sequence. Hydrogenase I, as encoded by the sequenced gene, consists of 574 amino acids. Among these, 159 (28%) have been identified by amino acid sequencing (Figure 2). The molecular weight calculated from the sequence, 63 836 for the apoprotein, is in agreement with the value (61 944) derived from a combination of SDS–polyacrylamide gel electrophoresis and amino acid analysis (Adams et al., 1989). The calculated isoelectric point, 6.5, is very close to the experimental value of 6.8 (Hinton & Mortenson, 1985). The amino acid composition deduced from the gene sequence agrees quite well with the most recently reported analysis (Adams et al., 1989) except for tryptophan, which had been greatly overestimated (nine residues instead of one, in position 303) (Table I). The number of cysteine residues, 22, which is important with respect to iron–sulfur cluster ligation, is consistent with the analytical data (Adams et al., 1989). The sequence has been analyzed using various algorithms (see Materials and Methods) which have confirmed the hydrophilicity of hydrogenase I and predicted ca. 60% of a α-helical structure.

Discussion

The high A + T content (68%) of the hydrogenase I gene results in a very biased codon usage, as previously noted for other *C. pasteurianum* genes (Graves et al., 1985; Chen et al., 1986; Hinton et al., 1987; Wang et al., 1988, 1990). In particular, for amino acids encoded by four codons, the third base is A or T in most cases (>90%). This occurrence has been used for the design of the three oligonucleotide probes and resulted in no more than a single mismatch in P2, at position 1293 of the 2.31-kb *Sau*3A fragment (Figure 2). The differences between the hydrogenase gene sequences from *C. pasteurianum* and *D. vulgaris* (Voordouw & Brenner, 1985) were expected to be significant, due to the very distinct DNA compositions of the two bacteria (Tonomura et al., 1965; Postgate, 1984). This prediction has been borne out by sequence comparisons: the gene fragments encoding the over-

```
gatcttaaaattaaatctAGGAGGctagatatgaaaacaataattataaatggtgtacagtttaatactgat 72
                               M  K  T  I  I  I  N  G  V  Q  F  N  T  D

gaagacactactatattaaaatttgcacgagacaacaatattgatatatctgcactgtgtttttaaataat 144
 E  D  T  T  I  L  K  F  A  R  D  N  N  I  D  I  S  A  L  C  F  L  N  N

tgtaataatgacataaataagtgtgaaatatgtactgtagaggtagagggtactggattagtaacagcctgt 216
 C  N  N  D  I  N  K  C  E  I  C  T  V  E  V  E  G  T  G  L  V  T  A  C

gatacattaattgaggatggtatgattataaacacaaattccgatgctgtcaacgaaaaaattaaatctaga 288
 D  T  L  I  E  D  G  M  I  I  N  T  N  S  D  A  V  N  E  K  I  K  S  R

atatctcaattattagacatacatgaattcaaatgtggtccttgcaatagaagagaaaactgtgaattctta 360
 I  S  Q  L  L  D  I  H  E  F  K  C  G  P  C  N  R  R  E  N  C  E  F  L

aaacttgttataaaatataaagcaagagcttctaaaccattttacctaaagataagactgaatatgtagat 432
 K  L  V  I  K  Y  K  A  R  A  S  K  P  F  L  P  K  D  K  T  E  Y  V  D

gaaagaagtaaatcattaactgtagataggacaaaatgcttattatgtggaagatgtgttaatgcctgtgga 504
 E  R  S  K  S  L  T  V  D  R  T  K  C  L  L  C  G  R  C  V  N  A  C  G

aaaaatactgaaacctatgcaatgaaatttttaaacaaaaatggtaaaactataattggagcagaggatgaa 576
 K  N  T  E  T  Y  A  M  K  F  L  N  K  N  G  K  T  I  I  G  A  E  D  E

aaatgctttgatgatactaattgtctattatgtggtcaatgtataatcgcctgtccagtagcagcattatcg 648
 K  C  F  D  D  T  N  C  L  L  C  G  Q  C  I  I  A  C  P  V  A  A  L  S

gaaaaatcacacatggatagagtaaaaaatgccttaaatgcccctgaaaaacatgtaatagtagctatggct 720
 E  K  S  H  M  D  R  V  K  N  A  L  N  A  P  E  K  H  V  I  V  A  M  A

ccatctgtcagagcttctataggtgaactttttaatatgggatttggcgttgacgtaacaggaaaaatttat 792
 P  S  V  R  A  S  I  G  E  L  F  N  M  G  F  G  V  D  V  T  G  K  I  Y

actgctttaagacagcttggatttgataaaatattcgatataaacttcggagcagatatgacaattatggaa 864
 T  A  L  R  Q  L  G  F  D  K  I  F  D  I  N  F  G  A  D  M  T  I  M  E

gaggctacagaattagttcaaagaatagagaataatggacctttcccaatgtttacatcttgctgcccaggt 936
 E  A  T  E  L  V  Q  R  I  E  N  N  G  P  F  P  M  F  T  S  C  C  P  G
                                     ■  ■  ■  ■  ■  ■  ■
tgggtaagacaagctgaaaattattatcctgaattactaaataatctttcatcagctaaatcacctcaacaa 1008
 W  V  R  Q  A  E  N  Y  Y  P  E  L  L  N  N  L  S  S  A  K  S  P  Q  Q

atttttggtactgctagtaaaacttattatccttctatatctggtcttgacccaaagaatgtatttactgta 1080
 I  F  G  T  A  S  K  T  Y  Y  P  S  I  S  G  L  D  P  K  N  V  F  T  V

acagttatgccctgtacttcaaaaaaatttgaagcagatagaccacaaatggaaaaagacggcctaagagat 1152
 T  V  M  P  C  T  S  K  K  F  E  A  D  R  P  Q  M  E  K  D  G  L  R  D

atagatgctgttataactactcgagaattagcaaaaatgattaaagatgctaaaataccatttgctaaactt 1224
 I  D  A  V  I  T  T  R  E  L  A  K  M  I  K  D  A  K  I  P  F  A  K  L

gaagatagcgaagcagaccctgctatgggagaatacagcggtgctggtgccatatttggtgcaactggcgga 1296
 E  D  S  E  A  D  P  A  M  G  E  Y  S  G  A  G  A  I  F  G  A  T  G  G
                                                                   ■  ■
gttatggaagcagctttaagaagtgcaaaagactttgctgaaaacgctgaacttgaagatatagaatataag 1368
 V  M  E  A  A  L  R  S  A  K  D  F  A  E  N  A  E  L  E  D  I  E  Y  K
 ■  ■  ■  ■  ■                 ■  ■  ■  ■  ■  ■
caagttagaggattaaatggtataaaagaagctgaagtagaaataaataacaacaaatataatgtagctgtt 1440
 Q  V  R  G  L  N  G  I  K  E  A  E  V  E  I  N  N  N  K  Y  N  V  A  V

ataaatggtgcttcaaatttatttaagtttatgaaatctggtatgattaacgaaaaacaatatcatttcata 1512
 I  N  G  A  S  N  L  F  K  F  M  K  S  G  M  I  N  E  K  Q  Y  H  F  I

gaagtaatggcttgtcatggaggatgtgtaaatggtggtggacagcctcatgtaaacccaaaagatttagaa 1584
 E  V  M  A  C  H  G  G  C  V  N  G  G  G  Q  P  H  V  N  P  K  D  L  E

aaagtagacataaaaaaagtaagagcttctgtattgtataatcaggatgaacatctttccaagagaaaatct 1656
 K  V  D  I  K  K  V  R  A  S  V  L  Y  N  Q  D  E  H  L  S  K  R  K  S

catgaaaatactgcattagttaaaatgtatcaaaattattttggcaaaccaggtgaaggtcgtgcccatgaa 1728
 H  E  N  T  A  L  V  K  M  Y  Q  N  Y  F  G  K  P  G  E  G  R  A  H  E

atattacactttaaatataaaaaataaatttattatttgaaaataaaAATAAAAACAGCATTATGAAAAtat 1800
 I  L  H  F  K  Y  K  K  *                     ─────────────────────►

tgttTTTTCATAATGCTGTTTTTATTataaatgcaacaaaaactgtactatatctgttaaatttttatcaag 1872
•    ◄───────────────────
ctgtattataatacatttttgaaatacatcaatatttacacaatctaacaaaaatataatcctttaatttt 1944
ttataataactttttttaaataatactaaatttattatttaatatactattttcttactaatctttacaaat 2016
tattacttcaataataaaaaaagaaaatgtgcctaaagaaacagggactccttagggtacatttcttttttt 2088
tacaacaaaatatcataacataaattttatagattattttgaatcctcaatttttttaattaaattaaaatt 2160
atatttttctatattatctttaagttctccattaaaagacgaatttggatgataccatgtttttgattcaaa 2232
atattccttaaattcgggcgtattaaatatatatccatgtcttgcaaaaagttcatttcttgctaatataag 2304
ttgatc 2310
```

FIGURE 2: Nucleotide sequence of the 2310-bp *Sau*3A fragment. The translated amino acid sequence for hydrogenase I is written below the DNA coding region. The sequence corresponding to the ribosome binding site of mRNA is shown in underlined bold capital letters. The transcription termination signal (bold capital letters) is underlined by arrows, and a dot marks the center of symmetry. Those amino acids that have been identified by peptide sequencing are underlined. The segments of the protein that have been used for the design of oligonucleotide probes are underscored with heavy blocks.

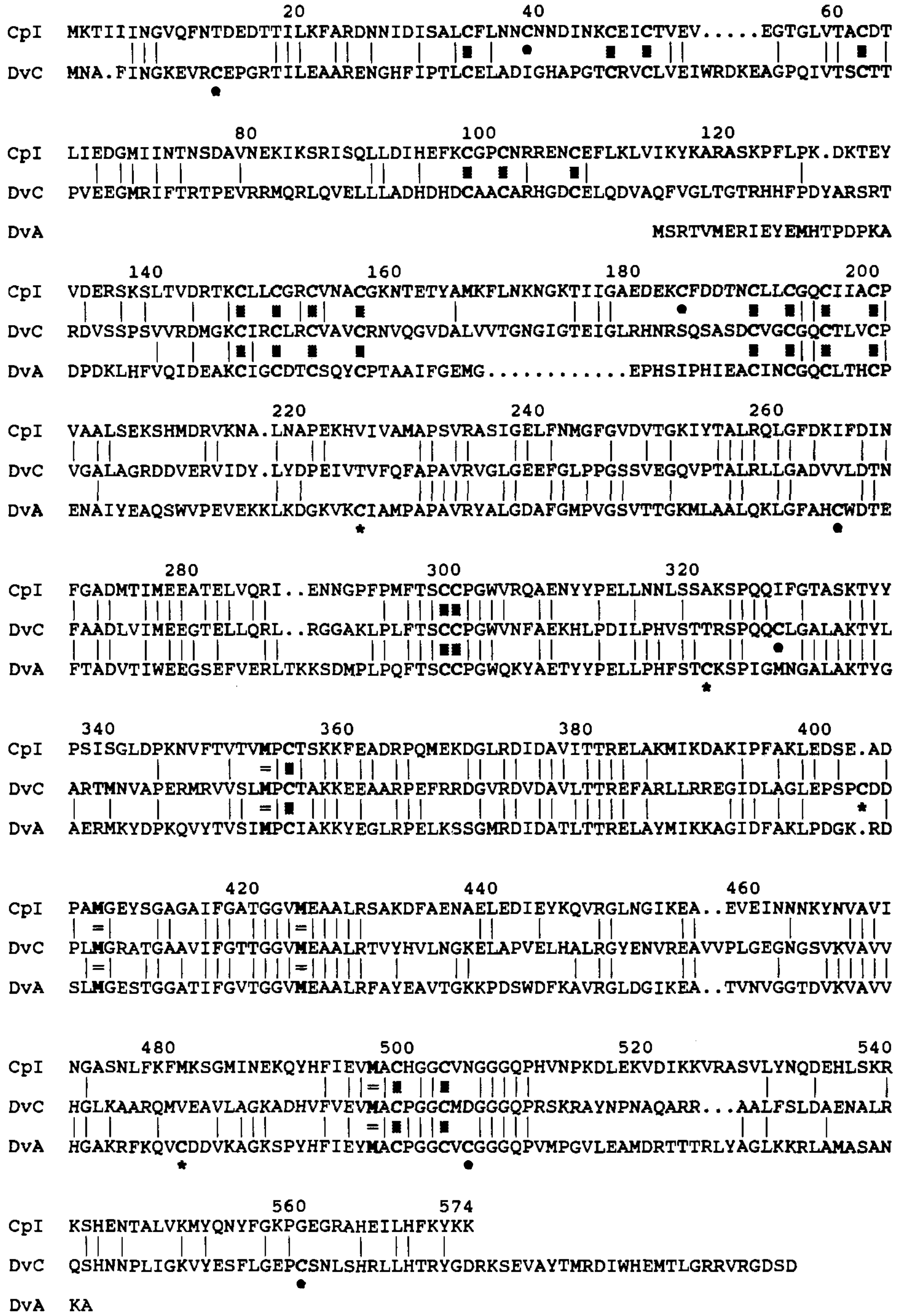

FIGURE 3: Comparison of the sequence of *C. pasteurianum* hydrogenase I with those of the translated hydγ gene from *D. vulgaris* (Hildenborough) (DvC; Stokkermans et al., 1989) and of the large hydrogenase subunit from the same strain (DvA; Voordouw & Brenner, 1985). The alignments were carried out using MULTALIN (Corpet, 1988). Matching residues are marked with vertical bars, conserved cysteines are marked with heavy blocks, conserved methionines are marked with a double underscore, and nonconserved cysteines are underscored with stars. Numbers are referring to the *C. pasteurianum* sequence only, with the designated residues standing below the last digit of each number.

lapping protein sequences (ca. 420 amino acids, Figure 3) display only 37% matching bases (not shown), despite the similarity of the protein sequences (Figure 3). Furthermore, no occurrence of more than eight consecutive matching bases was observed in the alignment. This explains why the two genes did not hybridize [quoted in Adams (1990)] and justifies our cloning strategy on the basis of the use of oligonucleotide probes designed from peptide sequences.

The considerable difference in A + T content between the DNA of *C. pasteurianum* and that of *D. vulgaris* is associated not only with a different codon usage but also with differences in amino acid compositions, as shown in Table II for *C. pasteurianum* hydrogenase and for the putative product of hydγ from *D. vulgaris* (see below): in these two homologous sequences (Figure 3), a strong bias favors those residues, within a group of physicochemically related amino acids, that have codon compositions closest to the overall DNA composition of the host bacterium (Table II). Of particular significance are the inverted contents of arginine and lysine, as well as the high levels of asparagine, isoleucine, phenylalanine, and tyrosine in the *C. pasteurianum* protein, as compared to the *D. vulgaris* one. Analogous discrepancies in amino acid compositions, though definitely smaller ones, had previously been observed between ribosomal proteins of *Mycoplasma capricolum* (G + C = 25%) and *E. coli* (G + C = 50%) (Yamao et al., 1989). They were shown to be associated with differences in the levels of various tRNAs, in the same fashion as codon usage biases (Yamao et al., 1989).

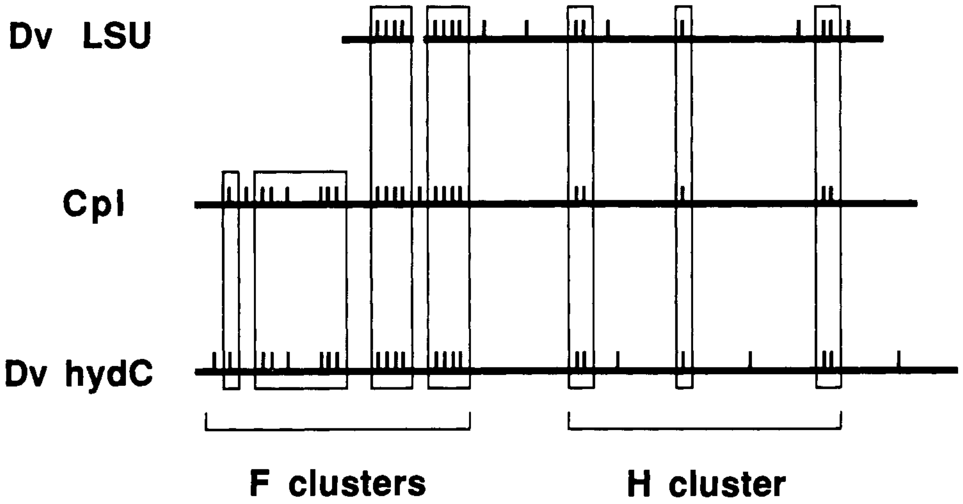

FIGURE 4: Schematic alignment of [Fe] hydrogenases, with emphasis on the cysteine residues (vertical bars). Conserved cysteines of the large subunit of *D. vulgaris* hydrogenase (Dv LSU), of *C. pasteurianum* hydrogenase I (CpI), and of the product of the hydγ gene from *D. vulgaris* (Dv hydC) are enclosed in rectangles. Note that the spacings between the cysteines are not exactly drawn to the same scale as the sequences. The blank in Dv LSU corresponds to a 12-residue gap in the sequence (see Figure 3). The sequence domains containing the ligands to the F and H clusters are indicated in the lower part of the figure.

Table II: Correlation between Codon Base Composition and Amino Acid Composition in Two Homologous Hydrogenases from *C. pasteurianum* (Hydrogenase I) and *D. vulgaris* (Putative Product of Gene hydγ) (Stokkermans et al., 1989)

		amino acid content of protein (mol %)	
residue	codon (two first bases only)	*C. pasteurianum* (G + C = 30%)	*D. vulgaris* (G + C = 65%)
R	AG C	3.7	9.2
K	AA	9.1	2.5
E	GA	7.7	6.9
Q	CA	2.4	2.8
D	GA	5.7	5.5
N	AA	7.5	3.0
G	GG	6.8	9.2
P	CC	3.5	5.3
A	GC	7.8	9.6
I	AT	7.1	3.0
F	TT	4.5	2.8
Y	TA	2.8	1.7

The amino acid sequence of *C. pasteurianum* hydrogenase I confirms that the sequence of [Fe] and [NiFe(Se)] hydrogenases are unrelated (Voordouw et al., 1989a). Further comparisons and discussions will mainly involve *D. vulgaris* (Hildenborough) [Fe] hydrogenase and, more specifically, its large subunit (46 kDa) which contains all of the 18 cysteine residues of the protein (Voordouw & Brenner, 1985; Voordouw et al., 1989b). Recently, a putative gene from *D. vulgaris* (Hildenborough), hydγ, has been sequenced (Stokkermans et al., 1989). Its potential translation product, which remains to be isolated and characterized, would be similar to, though significantly longer than, the large subunit of the dimeric [Fe] hydrogenase from the same organism (Voordouw & Brenner, 1985). The high similarity in size, cysteine topology, and sequence altogether between *C. pasteurianum* hydrogenase I and the product of hydγ (Figures 3 and 4) would suggest that the latter may be an actual hydrogenase encoding gene.

The amino acid sequences of the [Fe] hydrogenases from *C. pasteurianum* and *D. vulgaris*, and of the putative product of hydγ, have been aligned in Figure 3. A schematic alignment, which only shows the relative positions of the cysteine residues, is presented in Figure 4. Three domains are highlighted: an N-terminal segment of ca. 110 residues, present in the two longest sequences only, contains eight cysteine residues, seven of which occur in conserved positions. A second domain of ca. 80 residues (140–220) includes eight cysteines in a $[4Fe\text{-}4S]^{2+/+}$ ferredoxin-like pattern. In the third domain (290–510), which encompasses the larger part of their C-terminal halves, the three sequences are most similar. Of particular interest in this region is the occurrence of five matching cysteines (299, 300, 355, 499, and 503), which are themselves surrounded by conserved residues. The similarity between the three proteins decreases rather abruptly beyond residue 510, and the lengths of the C-termini vary greatly among the three sequences.

The sequence comparisons displayed in Figures 3 and 4 are most appropriately discussed in relation with available analytical and spectroscopic data on *C. pasteurianum* hydrogenase I (Adams, 1990) and *D. vulgaris* hydrogenase (Hagen et al., 1986; Patil et al., 1988). The latter enzyme has been proposed to accommodate two $[4Fe\text{-}4S]^{2+/+}$ ferredoxin-like (F) clusters in its N-terminal domain, which contains a characteristic run of eight cysteine residues, and a cluster of unknown structure, presumably the hydrogen activating site (H), in its C-terminal part (Hagen et al., 1986). *C. pasteurianum* hydrogenase I has been shown to contain an H cluster somewhat similar to the one present in the *D. vulgaris* enzyme but also to differ from the latter protein by its larger molecular weight (64 000 vs 48 000) and higher iron content (20–22 atoms/molecule vs 10–16) (Adams, 1990; Hagen et al., 1986; Patil et al., 1988). The difference in size is due to the presence of an additional N-terminal domain in the former enzyme (Figures 3 and 4). The difference in iron content has been tentatively rationalized, on the basis of spectroscopic data, by the presence of two additional (i.e., a total of four) [4Fe-4S] clusters in the *C. pasteurianum* protein (Adams, 1990). The sequence of the latter hydrogenase (Figure 3) now allows a reevaluation of the differences and similarities between the two hydrogenases. First, the inference that the H cluster of *D. vulgaris* hydrogenase is accommodated in the C-terminal domain of the protein (Hagen et al., 1986) most probably holds true for *C. pasteurianum* hydrogenase as well, in view of the great similarity of the two sequences in this region (Figure 3). Second, the two supplementary F clusters present in *C. pasteurianum* hydrogenase, as compared to the *D. vulgaris* one, are most likely harbored in the N-terminal extension of the former protein. This domain contains eight cysteines, which are likely ligands of the two [4Fe-4S] clusters, and the more so since seven of them are conserved between *C. pasteurianum* hydrogenase and the translation product of hydγ (Figure 4). The eight cysteine residues nearest to the N-terminus do not display the characteristic distribution observed in $[4Fe\text{-}4S]^{2+/+}$ ferredoxins, but the latter is not a prerequisite for the binding of such clusters: a number of studies have shown that the cysteine ligands of [4Fe-4S] clusters may display diverse sequence patterns (Hausinger & Howard, 1983; Cunningham et al., 1989; Beinert, 1990).

One of the most interesting features of the *C. pasteurianum* hydrogenase sequence is its great similarity to *D. vulgaris* hydrogenase in the H cluster binding domain, particularly in the segments containing likely ligands to this cluster (Figure 3, see below). It is thus most probable that the active sites of the two proteins are highly akin with respect to their core structure and amino acid coordination sphere. The sequence comparisons displayed in Figure 3 bear additional interest in that they allow more reliable predictions concerning the structure and the number of iron atoms of the H cluster. The latter has been the subject of a variety of proposals, ranging from three (Patil et al., 1988) to six (Hagen et al., 1986; Patil et al., 1988) iron atoms. In fact, if one considers the measured iron contents of the *C. pasteurianum* (Adams et al., 1989) and *D. vulgaris* (Hagen et al., 1986; Patil et al., 1988) hydrogenases and then subtracts the iron atoms ascribed to the F clusters, one obtains an even wider range of possible iron contents of the H cluster (two to eight atoms). According to a recent compilation and discussion of data at hand, a figure of six appeared to be more likely (Adams, 1990b), but this was by no means firmly established. Indeed, iron–sulfur chemistry has afforded a variety of structures containing three to eight iron atoms (Holm et al., 1984; Pohl & Saak, 1984; Noda et al., 1986; Snyder & Holm, 1988; Coucouvanis, 1991). Sequence comparisons (Figures 3 and 4) show, in the presumptive H cluster binding domain, five very conserved cysteines (residues 299, 300, 355, 499, and 503) which would provide a nearly complete set of ligands for a six-iron cluster. The coordination sphere of the H cluster might be contributed to by some of the nonconserved cysteine residues, which would then account for minor spectroscopic differences among the [Fe] hydrogenases (Adams, 1990).

An unusual and puzzling property of the sequences of the latter enzymes is the occurrence, in very conserved sequence segments, and in two cases very near to the probable cysteine ligands of the H cluster, of four conserved methionines (residues 353, 407, 424, and 497). At least some of these residues might thus be thought of as important agents in the vicinity of the H cluster, perhaps even as ligands. Methionine ligation to the H cluster, while unprecedented in iron–sulfur proteins, should not be ruled out a priori in the present case, since the H cluster is unusual by its spectroscopic properties and probably by its iron content as well. It is also worth mentioning that methionine is a metal ligand in c-type cytochromes (Dickerson & Timkovich, 1975) and plays a role in the close vicinity of the copper site of azurin and other related proteins (Karlsson et al., 1991).

At this point, the number of potential sulfur ligands to the H cluster appears to be sufficient, even for a six-to-eight-iron cluster. Thus, ligation by amino acid side chains containing no sulfur does not seem to be required on the basis of sequence data alone. However, since some spectroscopic data suggest the possibility of N or O ligation (Adams, 1990), we have inspected the sequences for such residues: several carboxylates occur in extremely conserved regions of the sequence (353–361, 370–383, 420–429, 493–503), but very few conserved potential nitrogen ligands are apparent except perhaps for the partially conserved histidine 492.

The number of hydrogenase genes possibly present in *C. pasteurianum* deserves some comment. The three probes used to clone the hydrogenase I gene have been hybridized with *C. pasteurianum* genomic DNA restricted by a number of enzymes. In no case was more than one strongly hybridizing fragment observed, which shows that only one hydrogenase I gene is present in the genome of *C. pasteurianum* but also that the gene encoding hydrogenase II has remained undetected. Hydrogenase II is smaller than hydrogenase I, has a lower iron content (Adams, 1990), and displays no immunological cross-reactivity with it (Kovacs et al., 1989). However, the spectroscopic signatures of their H clusters, while not identical, are similar in many aspects (Adams, 1990). Thus, from the comparisons made above (Figure 3) it may be inferred that the two proteins should contain some nearly identical patches in the C-terminal halves of their sequences. The apparent failure of our probes to hybridize with the hydrogenase II gene is not readily interpretable with the presently available data.

The [Fe] hydrogenases are diverse in subunit composition, iron content, and size (Adams, 1990). However, the sequence comparisons reported here show that they in fact constitute a rather homogeneous family of proteins. Not surprisingly, the greatest sequence conservation occurs in the region expected to bind the presumed hydrogen activating site, which a number of spectroscopic data suggest to possess a common architecture in all of these proteins (Adams, 1990). Another feature shared by [Fe] hydrogenases is the presence of at least two ferredoxin-like $[4Fe\text{-}4S]^{2+/+}$ clusters, although the sequences surrounding the corresponding ligands differ more than those forming the H cluster domain (Figure 3). Additional F clusters (two in the case of *C. pasteurianum* hydrogenase I) may occur in these enzymes, and they seem to be harbored in domains displaying even greater sequence variability. Interesting information regarding this point should be provided by the elucidation of the amino acid sequence of *Megasphaera elsdenii* hydrogenase, of which the size (Filipiak et al., 1989) appears to be intermediate between that of *C. pasteurianum* hydrogenase I and that of the large subunit from the *D. vulgaris* enzyme. The variability of the number and environment of the F clusters among [Fe] hydrogenases is consistent with their probable role as electronic shuttles between the redox partners of these enzymes and their H clusters.

Added in Proof

The N-termini of the HoxU polypeptide, a component of the NAD-reducing hydrogenase from *Alcaligenes eutrophus* (Tran-Betcke et al., 1990), and of the 75-kDa subunit of bovine mitochondrial NADH–ubiquinone reductase (Pilkington et al., 1991) have recently been shown to be homologous (Pilkington et al., 1991). The N-terminal region of the putative hydγ gene product (Stokkermans et al., 1989) is related to the two former sequences (van Dongen, personal communication), and so is, although to a lesser extent, the N-terminal domain (residues 1–130) of *C. pasteurianum* hydrogenase I. The functional and evolutionary implications of these homologies remain to be elucidated.

Acknowledgments

We thank J.-M. Moulis for continuous and stimulating advice throughout this work and for carefully reading the manuscript, P. Richaud for the synthesis of oligonucleotides and for sharing with us his experience in gene sequencing, N. Scherrer for determing the N-terminal sequence of hydrogenase, and I. Mathieu for discussions on gene cloning procedures.

References

Adams, M. W. W. (1990) *Biochim. Biophys. Acta 1020*, 115–145.

Adams, M. W. W., Eccleston, E., & Howard, J. B. (1989) *Proc. Natl. Acad. Sci. U.S.A. 86*, 4932–4936.

Alex, L. A., Reeve, J. N., Orme-Johnson, W. H., & Walsh, C. T. (1990) *Biochemistry 29*, 7237–7244.

Beinert, H. (1990) *FASEB J. 4*, 2483–2491.

Chen, J.-S., & Mortenson, L. E. (1974) *Biochim. Biophys. Acta 371*, 283–298.

Chen, K. C.-K., Chen, J.-S., & Johnson, J. L. (1986) *J. Bacteriol. 166*, 162–172.

Corpet, F. (1988) *Nucleic Acids Res. 16*, 10881–10890.

Coucouvanis, D. (1991) *Acc. Chem. Res. 24*, 1–8.

Crestfield, A. M., Moore, S., & Stein, W. H. (1963) *J. Biol. Chem. 238*, 622–627.

Cunningham, R. P., Asahara, H., Bank, J. F., Scholes, C. P., Salerno, J. C., Surerus, K., Münck, E., McCracken, J., Peisach, J., & Emptage, M. H. (1989) *Biochemistry 28*, 4450–4455.

Deckers, H. M., Wilson, F. R., & Voordouw, G. (1990) *J. Gen. Microbiol. 136*, 2021–2028.

Dickerson, R. E., & Timkovich, R. (1975) in *The Enzymes* (Boyer, P. D., Ed.) Vol. 11, pp 397–547, Academic Press, New York.

Fauque, G., Peck, H. D., Jr., Moura, J. J. G., Huynh, B. H., Berlier, Y., DerVartanian, D. V., Teixeira, M., Przybyla, A. E., Lespinat, P. A., Moura, I., & LeGall, J. (1988) *FEMS Microbiol. Rev. 54*, 299–344.

Filipiak, M., Hagen, W. R., & Veeger, C. (1989) *Eur. J. Biochem. 185*, 547–553.

Ford, C. M., Garg, N., Garg, R. P., Tibelius, K. H., Yates, M. G., Arp, D. J., & Seefeldt, L. C. (1990) *Mol. Microbiol. 4*, 999–1008.

Graves, M. C., Mullenbach, G. T., & Rabinowitz, J. C. (1985) *Proc. Natl. Acad. Sci. U.S.A. 82*, 1653–1657.

Hagen, W. R., van Berkel-Arts, A., Krüse-Wolters, K. M., Voordouw, G., & Veeger, C. (1986) *FEBS Lett. 203*, 59–63.

Hausinger, R. P., & Howard, J. B. (1983) *J. Biol. Chem. 258*, 13486–13492.

Hinton, S. M., & Mortenson, L. E. (1985) *J. Bacteriol. 162*, 477–484.

Hinton, S. M., Slaughter, C., Eisner, W., & Fisher, T. (1987) *Gene 54*, 211–219.

Holm, R. H., Hagen, K. S., & Watson, A. D. (1984) in *Chemistry for the Future* (Grünewald, H., Ed.) pp 115–124, Pergamon Press, New York.

Karlsson, B. G., Nordling, M., Pascher, T., Tsai, L.-C., Sjölin, L., & Lundberg, L. G. (1991) *Protein Eng. 4*, 343–349.

Kieser, T. (1984) *Plasmid 12*, 19–36.

Kovacs, K. L., Seefeldt, L. C., Tigyi, G., Doyle, C. M., Mortenson, L. E., & Arp, D. J. (1989) *J. Bacteriol. 171*, 430–435.

Leclerc, M., Colbeau, A., Cauvin, B., & Vignais, P. M. (1988) *Mol. Gen. Genet. 214*, 97–107; (1989) *215*, 368 (correction).

Levy, W. P., Rubinstein, M., Shively, J., Del Valle, U., Lai, C.-Y., Moschera, J., Brink, L., Gerber, L., Stein, S., & Pestka, S. (1981) *Proc. Natl. Acad. Sci. U.S.A. 78*, 6186–6190.

Menon, A. L., Stults, L. W., Robson, R. L., & Mortenson, L. E. (1990a) *Gene 96*, 67–74.

Menon, N. K., Robbins, J., Peck, H. D., Jr., Chatelus, C. Y., Choi, E.-S., & Przybyla, A. E. (1990b) *J. Bacteriol. 172*, 1969–1977.

Noda, I., Snyder, B. S., & Holm, R. H. (1986) *Inorg. Chem. 25*, 3851–3853.

Patil, D. S., Moura, J. J. G., He, S. H., Teixeira, M., Prickril, B. C., DerVartanian, D. V., Peck, H. D., Jr., LeGall, J., & Huynh, B.-H. (1988) *J. Biol. Chem. 263*, 18732–18738.

Pilkington, S. J., Skehel, J. M., Gennis, R. B., & Walker, J. E. (1991) *Biochemistry 30*, 2166–2175.

Pohl, S., & Saak, W. (1984) *Angew. Chem. 96*, 886–887.

Postgate, J. R. (1984) *The Sulphate-Reducing Bacteria*, 2nd ed., p 11, Cambridge University Press.

Rabinowitz, J. C. (1972) *Methods Enzymol. 24B*, 431–446.

Reeve, J. N., Beckler, G. S., Cram, D. S., Hamilton, P. T., Brown, J. W., Krzycki, J. A., Kolodziej, A. F., Alex, L., Orme-Johnson, W. H., & Walsh, C. T. (1989) *Proc. Natl. Acad. Sci. U.S.A. 86*, 3031–3035.

Rousset, M., Dermoun, Z., Hatchikian, C. E., & Belaich, J.-P. (1990) *Gene 94*, 95–101.

Saito, H., & Miura, K.-I. (1963) *Biochim. Biophys. Acta 72*, 619–629.

Sanger, F., Nicklen, S., & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U.S.A. 74*, 5463–5467.

Sayavedra-Soto, L. A., Powell, G. K., Evans, H. J., & Morris, R. O. (1988) *Proc. Natl. Acad. Sci. U.S.A. 85*, 8395–8399.

Schneider, C. G., Schmitt, H. J., Schild, Ch., Tichy, H. V., & Lotz, W. (1990) *Nucleic Acids Res. 17*, 5285.

Snyder, B. S., & Holm, R. H. (1988) *Inorg. Chem. 27*, 2339–2347.

Stokkermans, J., van Dongen, W., Kaan, A., van den Berg, W., & Veeger, C. (1989) *FEMS Microbiol. Lett. 58*, 217–222.

Tonomura, B. I., Malkin, R., & Rabinowitz, J. C. (1965) *J. Bacteriol. 89*, 1438–1439.

Tran-Betcke, A., Warnecke, U., Böcker, C., Zaborosch, C., & Friedrich, B. (1990) *J. Bacteriol. 172*, 2920–2929.

Tsö, M.-Y. W., Ljones, T., & Burris, R. H. (1972) *Biochim. Biophys. Acta 267*, 600–604.

Uffen, R. L., Colbeau, A., Richaud, P., & Vignais, P. M. (1990) *Mol. Gen. Genet. 221*, 49–58.

Voordouw, G., & Brenner, S. (1985) *Eur. J. Biochem 148*, 515–520.

Voordouw, G., Menon, N. K., LeGall, J., Choi, E.-S., Peck, H. D., Jr., & Przybyla, A. E. (1989a) *J. Bacteriol. 171*, 2894–2899.

Voordouw, G., Strang, J. D., & Wilson, F. R. (1989b) *J. Bacteriol. 171*, 3881–3889.

Wang, S.-Z., Chen, J.-S., & Johnson, J. L. (1988) *Nucleic Acids. Res. 16*, 439–454.

Wang, S.-Z., Chen, J.-S., & Johnson, J. L. (1990) *Biochem. Biophys. Res. Commun. 169*, 1122–1128.

Yamao, F., Andachi, Y., Muto, A., Ikemura, T., & Osawa, S. (1989) *Proc. Jpn. Acad. Ser. B 65*, 73–75.