Journal of Bacteriology, Apr. 1986, p. 162–172
0021-9193/86/040162-11$02.00/0

Vol. 166, No. 1

Structural Features of Multiple *nifH*-Like Sequences and Very Biased Codon Usage in Nitrogenase Genes of *Clostridium pasteurianum*

KATHERINE CHUAN-KAI CHEN,† JIANN-SHIN CHEN,* and JOHN L. JOHNSON

Department of Anaerobic Microbiology, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061

Received 20 November 1985/Accepted 16 January 1986

The structural gene (*nifH1*) encoding the nitrogenase iron protein of *Clostridium pasteurianum* has been cloned and sequenced. It is located on a 4-kilobase *Eco*RI fragment (cloned into pBR325) that also contains a portion of *nifD* and another *nifH*-like sequence (*nifH2*). *C. pasteurianum nifH1* encodes a polypeptide (273 amino acids) identical to that of the isolated iron protein, indicating that the smaller size of the *C. pasteurianum* iron protein does not result from posttranslational processing. The 5′ flanking region of *nifH1* or *nifH2* does not contain the *nif* promoter sequences found in several gram-negative bacteria. Instead, a sequence resembling the *Escherichia coli* consensus promoter (TTGACA-N_{17}-TATAAT) is present before *C. pasteurianum nifH2*, and a TATAAT sequence is present before *C pasteurianum nifH1*. Codon usage in *nifH1*, *nifH2*, and *nifD* (partial) is very biased. A preference for A or U in the third position of the codons is seen. *nifH2* could encode a protein of 272 amino acid residues, which differs from the iron protein (*nifH1* product) in 23 amino acid residues (8%). Another *nifH*-like sequence (*nifH3*) is located on a nonadjacent *Eco*RI fragment and has been partially sequenced. *C. pasteurianum nifH2* and *nifH3* may encode proteins having several amino acids that are conserved in other proteins but not in *C. pasteurianum* iron protein, suggesting a possible role for the multiple *nifH*-like sequences of *C. pasteurianum* in the evolution of *nifH*. Among the nine sequenced iron proteins, only the *C. pasteurianum* protein lacks a conserved lysine residue which is near the extended C terminus of the other iron proteins. The absence of this positive charge in the *C. pasteurianum* iron protein might affect the cross-reactivity of the protein in heterologous systems.

Biological nitrogen fixation is catalyzed by the enzyme nitrogenase, which is composed of two separable protein components: the iron protein (Fe protein, component II, or dinitrogenase reductase) and the molybdenum-iron protein (MoFe protein, component I, or dinitrogenase). Although nitrogen fixation appears to be under different modes of physiological control in different taxonomic groups (27, 39), active nitrogenase isolated from these organisms shows a remarkable similarity in component composition, enzymic properties, and the ability to form active heterologous complex (7, 14). Furthermore, a high degree of homology has been observed among gram-negative bacteria in the structural genes encoding the three polypeptides of nitrogenase: *nifH* for the iron protein and *nifD* and *nifK* for the α- and β-subunits of the MoFe protein (33, 36, 41).

We have been interested in *Clostridium pasteurianum* nitrogenase and its structural genes for several reasons. (i) *C. pasteurianum* is a gram-positive anaerobic bacterium with a low G+C content of 26 to 28% (11), which distinguishes *C. pasteurianum* from the rest of well-studied nitrogen-fixing organisms. (ii) The complete or partial amino acid sequence has been determined from isolated proteins (20, 21, 52), which facilitates the identification of functional genes and allows an examination for any posttranslational processing involving peptide bonds. (iii) The primary structure of *C. pasteurianum* nitrogenase components is significantly less related to that of nitrogenases from other organisms (9, 21, 23, 46, 54, 57, 59). (iv) *C. pasteurianum* nitrogenase has a high activity, but its components are distinctly ineffective in forming active heterologous complexes (14, 48, 55). (v) *C. pasteurianum* nitrogenase is less sensitive to H_2 as an inhibitor (19) and shows a higher specificity for nucleotides (58). Because of these intrinsic characteristics, the structural genes for *C. pasteurianum* nitrogenase are valuable for the investigation of two important properties of nitrogenase. The first concerns component interaction. By using distinct structures of compatible and incompatible *nifHDK* products as a clue, the cloned genes may be subjected to specific modifications to allow identification of regions of the component proteins that are crucial to the formation of an active enzyme complex. The second concerns the expression of nitrogenase genes in new host cells. By examining codon usage and regulatory features of nitrogenase genes from this gram-positive bacterium with a very low G+C content, we may gain clues as to the extent to which the efficiency of transcription and translation might limit the usefulness of transferring nitrogenase genes between certain organisms. The latter point is practically important because nitrogenase needs to be an abundantly expressed enzyme.

In this paper, we report the cloning and nucleotide sequencing of *nifH1*, *D* (partial) as well as additional *nifH*-related structures (*nifH2* and *nifH3*) from *C. pasteurianum*. The study provides the complete nucleotide sequence of a *nifH*-like structure (*nifH2*) and its exact genomic location in relation to the iron protein gene (*nifH1*). It also provides the first codon usage information for all 20 amino acids in a *Clostridium* sp. with a low G+C content. A comparison of the *nifH2*- and *nifH3*-encoded amino acid sequences with

* Corresponding author.

† Present address: Department of Biology, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061.

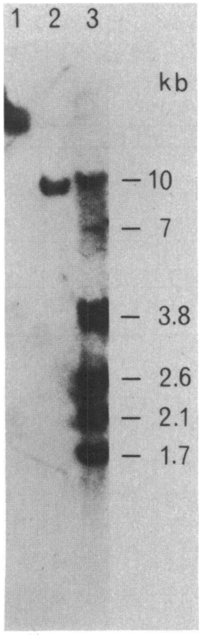

FIG. 1. Hybridization of ^{32}P-labeled *K. pneumoniae nif* fragment A3 (containing *nifDH*; 41) to *Hin*dIII-digested lambda DNA (1.5 μg, lane 1), *Eco*RI-digested *C. perfringens* DNA (10 μg, lane 2), and *Eco*RI-digested *C. pasteurianum* DNA (10 μg, lane 3). The autoradiogram was obtained after Southern transfer (49) of the fragments from agarose to nitrocellulose, and hybridization was performed at 42°C in the presence of 25% formamide. A sample of 10^7 cpm (in 0.24 μg) of *K. pneumoniae* fragment A3 DNA was used. Approximate sizes of bands are given in kilobase pairs.

that of iron proteins of eight other organisms suggests an evolutionary role for the multiple *nifH*-like sequences in *C. pasteurianum*.

MATERIALS AND METHODS

Plasmid pSA30 (8), containing the *Klebsiella pneumoniae nif* fragment A, was obtained from F. Ausubel. Subfragments A1 (*nifYK*), A2 (*nifKD*), and A3 (*nifDH*) (41) were cloned into pBR322. Bulk plasmid DNA was isolated by a variation of the alkaline lysing procedure of Birnboim and Doly (3). The plasmid preparations were further purified by two buoyant density centrifugations in ethidium bromide-CsCl. For the preparation of probe DNA, the fragment or subfragments, after appropriate restrictive digestion, were separated from the vector DNA by preparative agarose gel electrophoresis. The *nif* fragments were recovered by binding to NA 45 membrane (Schleicher & Schuell Co.). After the fragments were eluted from the membrane, a second electrophoresis and binding to NA 45 membrane were carried out. Contaminating vector DNA could not be detected in these probe preparations by ethidium bromide staining (4 μg of DNA), although some undoubtedly was present (but must be less than 1%). The isolated fragments were labeled with [^{32}P]dATP with the Bethesda Research Laboratories, Inc., nick translation kit.

High-molecular-weight DNA was isolated from late-log-phase cells of *C. pasteurianum* W5 by the Marmur procedure (31). A 5- to 10-μg sample of restriction endonuclease-digested DNA was electrophoresed in a 0.7% agarose gel. The DNA in the gels was transferred to nitrocellulose (Schleicher & Schuell; type BA85) by the method of Southern (49). The hybridization reaction mixtures contained 5× SSPE (0.9 M NaCl, 0.05 M phosphate buffer [pH 7.4], 5.0 mM EDTA), 5× Denhardt preincubation mixture (12), 0.1% sodium dodecyl sulfate, 100 μg of denatured salmon sperm DNA per ml, and 20 to 50% deionized formamide. The hybridizations were carried out at 42°C for 16 to 24 h. The 50% formamide concentration represents an equivalent hybridization temperature of 72°C (about 20°C below the melting point of *K. pneumoniae* DNA), whereas the 20% formamide concentration represents an equivalent hybridization temperature of 54°C (about 38°C below the melting point of *K. pneumoniae* DNA). The size of *C. pasteurianum* DNA fragments likely containing *nifH* or *nifD* genes was estimated by using subfragment A3 as a probe and *Hin*dIII-digested lambda DNA fragments as molecular weight markers.

After preparative electrophoresis of an *Eco*RI digest of *C. pasteurianum* DNA, fragments in the desirable size ranges were isolated by using NA 45 membranes. The fragments were ligated to *Eco*RI-digested and phosphatase-treated vector DNA (pBR322 or pBR325) and used to transform *Escherichia coli* HB101 (30). For direct hybridization screening, transformants were isolated, and plasmid DNA was isolated from small cultures (5 ml) as described above. Each plasmid preparation was then digested with *Eco*RI nuclease, electrophoresed on agarose gel, transferred to a nitrocellulose membrane, and then probed with labeled DNA fragment A3.

The DNA fragments were sequenced by the dideoxy chain termination method (42) and M13mp18 and M13mp19 phages. In addition, synthetic oligonucleotides were used as a primer to allow overlapping sequencing in regions where direct cloning was unsuccessful. The Bethesda Research Laboratories sequencing kit was used, except that 100 mM Tris–100 mM $MgCl_2$ (pH 8.5) was used as the 10× primer hybridization buffer. The ^{35}S-labeled dATP was obtained from either Amersham Corp. or New England Nuclear Corp. Electrophoresis was in polyacrylamide gradient gels as described by Biggin et al. (2). The sequences were analyzed with the Pustell and Kafatos DNA sequencing program (37). The similarity coefficient (S_{AB}) between two protein sequences (A, B) is defined as S_{AB} = (2 × number of identical residues between A and B)/[(number of total residues in A) + (number of total residues in B)].

RESULTS AND DISCUSSION

Cloning of *C. pasteurianum nifH* and *nifD* genes. The cloned *K. pneumoniae nifHDK* genes have been a very useful probe for the cloning of nitrogenase genes from organisms in which genetic manipulations are not yet as practical as in *K. pneumoniae*. The G+C content of the DNA from *C. pasteurianum* and from *K. pneumoniae* differs by about 30 mol%. Therefore, one would not expect to find extensive sequence similarity between homologous genes from the two organisms, if the G+C content of the genes reflects the average G+C content for the genome. Indeed, under more stringent hybridization conditions (50% formamide, 42°C), *Eco*RI-digested *C. pasteurianum* DNA showed only one very faint band around 3.8 kilobases (kb) when *K. pneumoniae* fragment A (*nifHDK*) was used as the probe. Under less stringent conditions (25% formamide), *K. pneumoniae* fragment A2 (*nifKD*) detected a very faint band around 7 kb, which might correspond to the 6.2-kb band reported earlier (41). When *K. pneumoniae* fragment A was used at 10% formamide or *K. pneumoniae* fragment A3

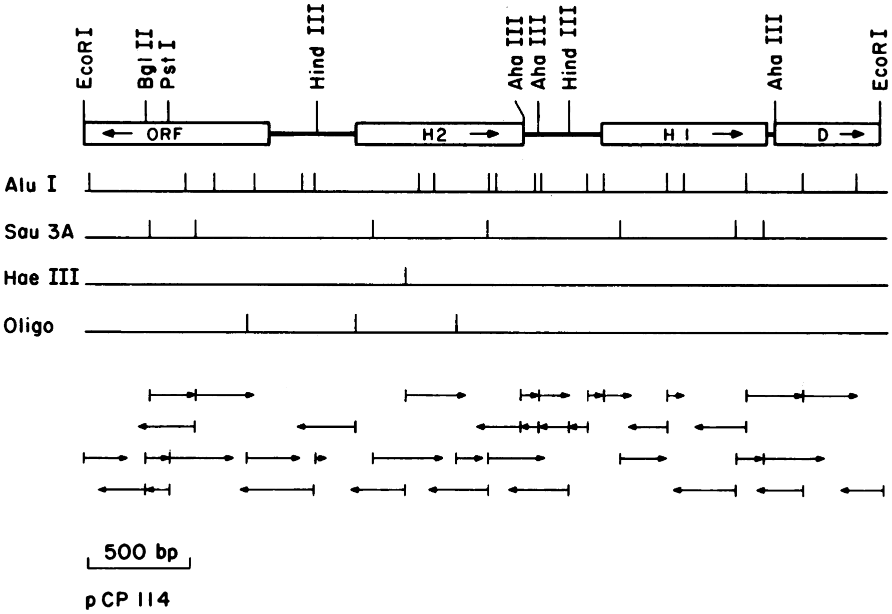

FIG. 2. Maps of restriction sites and *nif* gene locations on the 4-kb *Eco*RI fragment of *C. pasteurianum* DNA (cloned in pCP114) and the sequencing strategy. *nifH1* encodes the isolated iron protein, whereas *nifH2* encodes an amino acid sequence very similar to that of the isolated iron protein. *nifD* encodes the α-subunit of the MoFe protein. ORF, Open reading frame. Arrows in the boxes show directions of transcription. Oligo, Synthetic oligonucleotides used as specific primers in the sequencing of particular regions. The arrows indicate the extent of sequencing from each site and the strand on which the sequencing was performed. bp, Base pairs.

(*nifDH*) was used at 25% formamide, we detected six hybridizing bands (~10, ~7, 3.8, 2.6, 2.1, and 1.7 kb) (Fig. 1). Under the latter hybridization conditions, DNA from lambda phage and the non-N_2-fixing *Clostridium perfringens* each gave a false-positive band (Fig. 1). Thus, all of those *C. pasteurianum* bands could not be positively assigned as having *nifDH*-homologous sequences without further analyses. Because the 3.8-kb fragment was the strongest among the weakly hybridizing *C. pasteurianum* bands, it was selected for cloning.

Initial cloning experiments with colony hybridization as a method for detecting positive clones did not work because the nonspecific association between the G+C-rich probe DNA and the G+C-rich vector and host DNAs was much greater than any specific duplex. Low levels of contaminating vector DNA in the probe also contributed to the high background. Therefore, fragments from the size range of interest were cloned into pBR325 (for clear separation of the vector and insert DNA) and the inserts (from isolated plasmid DNA) probed with labeled fragment A3 DNA. Several clones showed weak hybridization with the probe DNA. Insert DNA from all of these clones hybridized strongly with each other and weakly to fragment A3. Based on this evidence, one of the clones (pCP114) was selected for further study. The insert DNA (ca. 3.8 kb) was sequenced with the sequencing strategy shown in Fig. 2. The complete sequence of this "3.8-kb" *Eco*RI fragment shows 3,987 base pairs (Fig. 3 and 4), and it is hereafter referred to as the 4-kb fragment. This 4-kb fragment hybridized to the "3.8-kb" band in *Eco*RI-digested *C. pasteurianum* DNA.

Identification of *nifH* in this fragment was based on a perfect match between the deduced amino acid sequence from an open reading frame (*nifH1*; Fig. 2 and 3; see below) and the known amino acid sequence of *C. pasteurianum* iron protein (52). We identified another sequence (*nifH2*; Fig. 2 and 3) very similar to *nifH1* and also located the N-terminal portion of *nifD* in the fragment. The sizes and locations of these genes are shown in Fig. 2. Another open reading frame (>296 amino acid residues) upstream of *nifH2* but in the opposite direction of translation was also identified (Fig. 2 and 4).

Nucleotide sequence of *C. pasteurianum nifH1* and the flanking regions. The complete nucleotide sequence of *C. pasteurianum nifH1* and its flanking regions is shown in Fig. 3. *nifH1* encodes 273 amino acids identical to that determined from the isolated Fe protein (52). A putative ribosome-binding site (AGGAGGA, underlined) was present between −14 and −8 nucleotides from the initiation codon AUG. A similar site is present between −16 and −11 nucleotides of the *C. pasteurianum* ferredoxin gene (18) and between −14 and −8 nucleotides of the *Clostridium thermocellum* cellulase gene (1). This sequence is assigned as the putative translational start signal (50), with the assumption that the nucleotide sequence at the 3′ terminus of the clostridial 16S rRNA is similar to that of *E. coli*.

Between −340 and −300 nucleotides, a potential stem-and-loop structure with a stem of 15 base pairs might be formed (Fig. 3, inverted repeats underlined by arrows). Whether it serves as a transcription termination signal (40) for the preceding operon is yet to be determined. Between −300 and −14 nucleotides, no sequence similar to known *nif* promoters, CTGG-N_{10}-GC(A/T) (13), TCTAC (56), or TGGCA-N_4-GGTTGC (59), was found. However, a TATAAT sequence (Fig. 3, underlined) (40) was present in the −250 region, and the entire region was abundant in long stretches of A and T (the noncoding region was 83% A + T, whereas the coding region was 64% A + T).

Since the amino acid sequence deduced from the DNA agreed completely with that of the isolated protein, there must be no posttranslational processing involving peptide

```
ORF ←                                          50                                                100
TATGGTGTAAAGGA-5'
ATACCACATTTCCTTCCATGATTGATTTATATTTTTTAACATTAGTAATAGTTAATTAGTATAATGCTTAAAACTTTAATATTATTCATATTGTAAGTAATATTTATTATACACATTATA

                             150                                               200
GATAAATAATATCATATAAATAATTATAAATATATTATGAATAAACTACAATAAGCTATAAATAAATTATATATCTTTATTTATATTGAATTTTTATCATATACAATATTTCAGCTTTGT

         250                                               300                                               350
AAGCTTTACGATTATAAACATTATATCATATAGAACTGAATAATCTATAAAAAATTAATGGTGTATCATAAATGGAGAAAATTTGATATATTGATATGTTATATAAATAAAAATTAATA

                                        400                                               450
ATTTAATAGTAAATATGGTATATTTGTTGACAAGTACTAAATTAAGGAATATAATGAAAAACGAAGTATAAAGCATAGAGATGTGGAAAATAATCATTCCATGTTAGACATCAAAGGGA

                     500                                                                550
                        nifH2
CGTAATTTAAGGAGGAATATTAA ATG AGA CAG TTG GCT ATT TAC GGA AAA GGT GGA ATA GCA AAA TCA ACT ACA ACA CAA AAC CTT ACA GCA GGT
                        Met Arg Gln Leu Ala Ile Tyr Gly Lys Gly Gly Ile Ala Lys Ser Thr Thr Thr Gln Asn Leu Thr Ala Gly

                                 600                                                                650
TTA GTT GAA AGA GGA AAT AAA ATA ATG GTA GTT GGT TGT GAT CCT AAG GCA GAT TCA ACA AGA TTA TTG TTA GGA GGA CTT GCT CAA AAG
Leu Val Glu Arg Gly Asn Lys Ile Met Val Val Gly Cys Asp Pro Lys Ala Asp Ser Thr Arg Leu Leu Leu Gly Gly Leu Ala Gln Lys

                                               700                                                            750
ACA GTT CTT GAT ACC TTG AGA GAA GAG GGA GAA GAC GTT GAA TTA GAT TCG ATA TTA AAA ACT GGA TAT GCT GGA ATC AGA TGC GTC GAA
Thr Val Leu Asp Thr Leu Arg Glu Glu Gly Glu Asp Val Glu Leu Asp Ser Ile Leu Lys Thr Gly Tyr Ala Gly Ile Arg Cys Val Glu

                                                           800
TCC GGT GGC CCA GAA CCA GGA GTA GGG TGT GCA GGA AGA GGA ATA ATC ACT TCA ATA AAT ATG CTT GAA CAA CTT GGA GCT TAT ACA GAC
Ser Gly Gly Pro Glu Pro Gly Val Gly Cys Ala Gly Arg Gly Ile Ile Thr Ser Ile Asn Met Leu Glu Gln Leu Gly Ala Tyr Thr Asp

       850                                                             900
GAT TTG GAT TTT GTA TTC TAC GAT GTA CTT GGA GAC GTT GTT TGT GGT GGA TTT GCA ATG CCA ATC AGA GAA GGA AAA GCT CAG GAA ATA
Asp Leu Asp Phe Val Phe Tyr Asp Val Leu Gly Asp Val Val Cys Gly Gly Phe Ala Met Pro Ile Arg Glu Gly Lys Ala Gln Glu Ile

                   950                                                                 1000
TAT ATA GTA GCA AGT GGA GAA ATG ATG GCA CTA TAT GCT GCT AAT AAC ATA TCA AAA GGT ATC CAA AAA TAT GCT AAG AGC GGT GGA GTT
Tyr Ile Val Ala Ser Gly Glu Met Met Ala Leu Tyr Ala Ala Asn Asn Ile Ser Lys Gly Ile Gln Lys Tyr Ala Lys Ser Gly Gly Val

                               1050                                                                1100
AGA CTT GGT GGT ATC ATC TGT AAC AGT AGA AAA GTT GCA AAT GAA TAT GAA TTA CTT GAT GCT TTC GCA AAA GAA TTA GGA AGT CAA TTA
Arg Leu Gly Gly Ile Ile Cys Asn Ser Arg Lys Val Ala Asn Glu Tyr Glu Leu Leu Asp Ala Phe Ala Lys Glu Leu Gly Ser Gln Leu

                                               1150                                                            1200
ATA CAC TTC GTA CCA AGA AGT CCA TCA GTA ACA AAG GCT GAA ATA AAT AAG AAA ACA GTT ATA GAA TAT GAT CCT ACT TGT GAA CAA GCT
Ile His Phe Val Pro Arg Ser Pro Ser Val Thr Lys Ala Glu Ile Asn Lys Lys Thr Val Ile Glu Tyr Asp Pro Thr Cys Glu Gln Ala

                                                           1250
AAT GAG TAC AGA GAA CTA GCT AGA AAA GTA GAG GAA AAT GAC ATG TTC GTT ATA CCA AAG CCA ATG ACT CAA GAA AGA TTA GAA CAA ATA
Asn Glu Tyr Arg Glu Leu Ala Arg Lys Val Glu Glu Asn Asp Met Phe Val Ile Pro Lys Pro Met Thr Gln Glu Arg Leu Glu Gln Ile

       1300                                                     1350                                          1400
TTA ATG GAA CAT GGT CTT ATT GAT TAA GATAGTATTAAATGTAAAACTATAATTTTAAAAATAAATAATTTGGAAACTTTTATACATGAAATTTTACTATAGAATAAAGGA
Leu Met Glu His Gly Leu Ile Asp ***

                                            1450                                              1500
TAGCTTAAAAGTTATCCTTTATTCTATTTAATTTTTAAATATAAGCTATATATGAATTTATTAATAAATACTTTGTGATTTTTTATAATTAATATTAATACATTGCTAAATTATCAATA
→          ←

                          1550                                              1600
TATTAATGCCTGTATCCATAAATACTTTGATAAAATTTATCTGAAATATCGTTTTCATTAAGCTTTTGATAAAAAATACAAAAAAACTTTATGAAATTATCAAATTTTCTATTGTATTA

      1650                                              1700                            nifH1                  1750
ATTAAATTATTGTGATATATTTTCATTAAGCTAAAAAAACAACATAGCAAAACGTAAATTACTTTTAATTTTTAGGAGGAATGTTTA ATG AGA CAG GTA GCT ATT TAT GGA
                                                                                        Met Arg Gln Val Ala Ile Tyr Gly

                                                           1800
AAA GGT GGA ATA GGA AAA TCA ACT ACA ACA CAA AAC TTA ACA TCA GGT CTT CAT GCA ATG GGT AAG ACT ATA ATG GTA GTA GGT TGT GAT
Lys Gly Gly Ile Gly Lys Ser Thr Thr Thr Gln Asn Leu Thr Ser Gly Leu His Ala Met Gly Lys Thr Ile Met Val Val Gly Cys Asp

       1850                                                            1900
CCT AAG GCA GAT TCA ACA AGA TTA TTA CTT GGA GGT CTT GCA CAG AAA TCA GTT CTT GAT ACA TTA AGA GAA GAA GGA GAA GAC GTT GAA
Pro Lys Ala Asp Ser Thr Arg Leu Leu Leu Gly Gly Leu Ala Gln Lys Ser Val Leu Asp Thr Leu Arg Glu Glu Gly Glu Asp Val Glu

                   1950                                                                2000
TTA GAT TCC ATA TTA AAA GAA GGA TAT GGC GGA ATT AGA TGT GTT GAA TCC GGT GGT CCA GAA CCA GGA GTA GGA TGT GCA GGA AGA GGA
Leu Asp Ser Ile Leu Lys Glu Gly Tyr Gly Gly Ile Arg Cys Val Glu Ser Gly Gly Pro Glu Pro Gly Val Gly Cys Ala Gly Arg Gly

                               2050                                                                2100
ATA ATC ACT TCA ATA AAC ATG CTT GAA CAA TTA GGA GCT TAT ACA GAC GAT TTA GAC TAT GTA TTC TAC GAT GTA CTT GGA GAC GTT GTT
Ile Ile Thr Ser Ile Asn Met Leu Glu Gln Leu Gly Ala Tyr Thr Asp Asp Leu Asp Tyr Val Phe Tyr Asp Val Leu Gly Asp Val Val

                                               2150                                                            2200
TGT GGT GGA TTC GCA ATG CCA ATC AGA GAA GGA AAA GCT CAG GAA ATA TAT ATA GTA GCA AGT GGA GAA ATG ATG GCA CTA TAT GCT GCT
Cys Gly Gly Phe Ala Met Pro Ile Arg Glu Gly Lys Ala Gln Glu Ile Tyr Ile Val Ala Ser Gly Glu Met Met Ala Leu Tyr Ala Ala

                                                           2250
AAT AAC ATA TCA AAA GGT ATC CAA AAA TAT GCT AAG AGC GGT GGA GTT AGA CTT GGT GGT ATC ATC TGT AAC AGT AGA AAA GTT GCA AAT
Asn Asn Ile Ser Lys Gly Ile Gln Lys Tyr Ala Lys Ser Gly Gly Val Arg Leu Gly Gly Ile Ile Cys Asn Ser Arg Lys Val Ala Asn
```

```
        2300                                                            2350
GAA TAT GAA TTA CTT GAT GCT TTT GCT AAA GAA CTA GGA AGT CAA TTA ATA CAT TTC GTA CCA AGA AGC CCA ATG GTT ACA AAA GCA GAA
Glu Tyr Glu Leu Leu Asp Ala Phe Ala Lys Glu Leu Gly Ser Gln Leu Ile His Phe Val Pro Arg Ser Pro Met Val Thr Lys Ala Glu

                    2400                                                2450
ATC AAT AAG CAA ACT GTT ATT GAA TAT GAT CCT ACT TGT GAA CAG GCT GAA GAA TAC AGA GAA TTA GCT AGA AAA GTA GAT GCA AAT GAA
Ile Asn Lys Gln Thr Val Ile Glu Tyr Asp Pro Thr Cys Glu Gln Ala Glu Glu Tyr Arg Glu Leu Ala Arg Lys Val Asp Ala Asn Glu

                                2500                                                                           2550
TTA TTC GTT ATA CCA AAG CCA ATG ACT CAA GAA AGA CTT GAA GAA ATA TTA ATG CAA TAT GGT TTA ATG GAT CTA TAA GATTTAATAAAAGT
Leu Phe Val Ile Pro Lys Pro Met Thr Gln Glu Arg Leu Glu Glu Ile Leu Met Gln Tyr Gly Leu Met Asp Leu ***

                            nifD    2600                                                            2650
ATTTAATTTTGATGAGGGGTGAATTTC GTG AGC GAA AAT TTA AAA GAC GAG ATT TTA GAA AAA TAT ATA CCT AAA ACT AAA AAG ACT AGA AGT GGT
                            Met Ser Glu Asn Leu Lys Asp Glu Ile Leu Glu Lys Tyr Ile Pro Lys Thr Lys Lys Thr Arg Ser Gly

                                                2700                                                                2750
CAT ATA GTT ATA AAA ACT GAA GAA ACA CCA AAT CCT GAA ATA GTT GCT AAC ACA AGA ACA GTG CCA GGA ATA ATC ACA GCT AGA GGT TGT
His Ile Val Ile Lys Thr Glu Glu Thr Pro Asn Pro Glu Ile Val Ala Asn Thr Arg Thr Val Pro Gly Ile Ile Thr Ala Arg Gly Cys

                                                                2800
GCT TAT GCA GGA TGT AAA GGT GTT GTT ATG GGA CCA ATA AAG GAT ATG GTT CAC ATC ACA CAC GGA CCT ATA GGA TGT TCA TTC TAT ACA
Ala Tyr Ala Gly Cys Lys Gly Val Val Met Gly Pro Ile Lys Asp Met Val His Ile Thr His Gly Pro Ile Gly Cys Ser Phe Tyr Thr

        2850                                                                    2900
TGG GGT GGA AGA AGA TTT AAG TCT AAA CCA GAA AAC GGT ACT GGA TTA AAT TTT AAT GAA TAT GTA TTC TCT ACT GAT ATG CAG GAA AGT
Trp Gly Gly Arg Arg Phe Lys Ser Lys Pro Glu Asn Gly Thr Gly Leu Asn Phe Asn Glu Tyr Val Phe Ser Thr Asp Met Gln Glu Ser

                        2950                                                                3000
GAC ATA GTT TTT GGT GGA GTT AAT AAA TTA AAA GAT GCT ATA CAT GAA GCA TAT GAA ATG TTC CAT CCA GCA GCT ATA GGT GTT TAT GCA
Asp Ile Val Phe Gly Gly Val Asn Lys Leu Lys Asp Ala Ile His Glu Ala Tyr Glu Met Phe His Pro Ala Ala Ile Gly Val Tyr Ala

                                    3050
ACA TGT CCA GTT GGT CTT ATC GGT GAT GAT ATA CTA GCA GTT GCT GCA ACA GCA AGC AAA GAA ATT GGA ATT C
Thr Cys Pro Val Gly Leu Ile Gly Asp Asp Ile Leu Ala Val Ala Ala Thr Ala Ser Lys Glu Ile Gly Ile
```

FIG. 3. Nucleotide sequence of the region containing *nifH2*, *nifH1*, and *nifD* (partial) of *C. pasteurianum*. The DNA strand shown is that identical to mRNA. Sequences discussed in the text are underlined. The inverted repeats are indicated by arrows. ORF, Open reading frame.

bonds of the iron protein in *C. pasteurianum*. In *K. pneumoniae* (22, 44, 51) and *Azotobacter vinelandii* (6, 23), the mature iron protein lacks the N-terminal methionine found in the deduced sequence.

Nucleotide sequence and the encoded amino acid sequence of *C. pasteurianum nifH2*. The presence of multiple *nif* sequences in *C. pasteurianum* was suggested by the number of *Eco*RI fragments detected by *K. pneumoniae nifHD* (Fig. 1). The existence of multiple *nifH*-like sequences was conclusively shown by nucleotide sequence data. *nifH2* had an open reading frame of 816 nucleotides and a potential ribosome-binding site (AGGAGGA) between −14 and −8 nucleotides from the putative initiation codon AUG (Fig. 3). In the 272 amino acids possibly encoded by *nifH2* (Fig. 5), only 23 amino acids differed from those of the *nifH1*-encoded Fe protein. This gave a similarity coefficient (S_{AB}) of 0.92 between the putative *nifH2* product and the iron protein (*nifH1* product). At the nucleotide level, the homology was only slightly lower (S_{AB} = 0.90) between *nifH1* and *nifH2*.

For the 23 different residues between *C. pasteurianum nifH1*- and *H2*-encoded proteins, 13 occur in regions where either conserved secondary structures among iron proteins are predicted or the *C. pasteurianum* iron protein contains distinct features. Four of them (residues 13, 23, 222, and 263 of the putative *nifH2*-encoded protein) may cause some changes in the secondary structure based on predictions by the Chou and Fasman methods (10).

Additional *nifH*-related sequences were also obtained in separate clones from *C. pasteurianum*. One of them (designated *nifH3*) was located on a 2.6-kb *Eco*RI fragment and was cloned into pBR322 as pCP3. The cloned portion of *nifH3* was sequenced. Nucleotide sequence data (not shown) indicate that *nifH3* is not in proximity to *nifH1* and *nifH2*. Of the 194 deduced amino acid residues of *C. pasteurianum nifH3* (Fig. 5), 64 were different from the *nifH1*-encoded iron protein.

Unexpectedly, a sequence resembling the *E. coli* consensus promoter (TTGACA-N_{17}-TATAAT) (40) was found between −116 and −88 nucleotides from the initiation codon of *nifH2* (Fig. 3, underlined; Fig. 4 shows a similar sequence before the open reading frame).

Partial nucleotide and amino acid sequences of *C. pasteurianum nifD*. The portion of *nifD* cloned in pCP114 has been sequenced. The deduced amino acid sequence (166 residues) matches that of the α-subunit of the *C. pasteurianum* MoFe protein (21), except that residue 94 was asparagine according to the nucleotide sequence instead of aspartate as reported from the protein analysis. In addition, residue 41 was arginine (21) instead of lysine (20). The initiation codon for *C. pasteurianum nifD* was assigned to the GUG (*N*-formylmethionine) which preceded the N-terminal residue (Ser) of the isolated protein, indicating posttranslational processing of the polypeptide. *C. pasteurianum nifH1* and *nifD* were separated by 41 nucleotides (Fig. 3); a potential ribosome-binding site (GAGG, underlined) was located between −14 and −11 nucleotides from the putative initiation codon (GUG) of *nifD*.

Open reading frame upstream of *nifH2* in the 4-kb *Eco*RI fragment. Figures 2 and 4 show the open reading frame located upstream of *nifH2* but on the complimentary strand of DNA. The general location of this open reading frame and its opposite direction of transcription in relation to *nifH1* made it similar to *nifJ* in *K. pneumoniae* (13, 47). However, it is not known whether *C. pasteurianum* has a *nif*-regulated pyruvate:ferredoxin/flavodoxin oxidoreductase. This open reading frame had a potential ribosome-binding site (AGGA; Fig. 4, underlined) at −14 to −11 nucleotides from the postulated translation start. There were two other nearby in-phase AUG codons up- and downstream, but the putative

```
nifH2 ←
AATTATAAGGAGGAATTTAATGCAGGGAAA.....5'           50                                                100
                                                 |                                                  |
TTAATATTCCTCCTTAAATTACGTCCCTTTGATGTCTAACATGGAATGATTATTTTCCACATCTCTATGCTTTATACTTCGTTTTTCATTATATTCCTTAATTTAGTACTTGTCAACAAA

                            150                                               200
                             |                                                 |
TATACCATATTTACTATTAAATTATTAATTTTTATTTATATAACATATCAATATATCAAATTTTCTCCATTTATGATACACCATTAATTTTTTATAGATTATTCAGTTCTATATGATATA

        250                                               300                                               350
         |                                                 |                                                 |
ATGTTTATAATCGTAAAGCTTACAAAGCTGAAATATTGTATATGATAAAAATTCAATATAAATAAAGATATATAATTTATTTATAGCTTATTGTAGTTTATTCATAATATATTTATAAT

                                       400                                               450
                                        |                                                 |
TATTTATATGATATTATTTATCTATAATGTGTATAATAAATATTACTTACAATATGAATAATATTAAAGTTTTAAGCATTATACTAATTAACTATTACTAATGTTAAAAAATATAAATC

                    500 ORF                                                         550
                      |                                                               |
AATCATGGAAGGAAATGTGGTAT ATG GAT ATG GAT ACA GCT CTC ACT CCT CAA GAG GTT GCG GAT ATA TTA AAA ATT TCA AAA AGT ACC GTA TAT
                        Met Asp Met Asp Thr Ala Leu Thr Pro Gln Glu Val Ala Asp Ile Leu Lys Ile Ser Lys Ser Thr Val Tyr

                               600                                             650
                                |                                               |
GAT TTA ATT AAG AAA AAA GAA ATA AAC TCT TAC AGA GTA GGT AAA AAA GTT CGA GTT GAC TTA AAG GAT GTA GAA GCC TAC AAA AAT AAA
Asp Leu Ile Lys Lys Lys Glu Ile Asn Ser Tyr Arg Val Gly Lys Lys Val Arg Val Asp Leu Lys Asp Val Glu Ala Tyr Lys Asn Lys

                                               700                                     750
                                                |                                       |
ACC AAA AAT ATA AAA TCT AAT ATT TTT GTT CCT AGT AAT TCA GTA GTT ATT AAT TCT TCA TCT TTA TAT GAT GGA ATG ACA CCA AAG GAA
Thr Lys Asn Ile Lys Ser Asn Ile Phe Val Pro Ser Asn Ser Val Val Ile Asn Ser Ser Ser Leu Tyr Asp Gly Met Thr Pro Lys Glu

                                                           800
                                                            |
GAG GTG TTA GAG GAT AGC TTT GTT ATA TCT GGT CAA GAT ACA ATT TTA GAT ATT TTA TGC CGT TAT CTC GAT TCC TAC CCT CAT GGT TCT
Glu Val Leu Glu Asp Ser Phe Val Ile Ser Gly Gln Asp Thr Ile Leu Asp Ile Leu Cys Arg Tyr Leu Asp Ser Tyr Pro His Gly Ser

       850                                                             900
        |                                                               |
ATG CGA GTT TTG AGA TCC TAT GAA GGC AGT TAT AAT GGT ATA TAT GAA TTA TAT TGT GGA AAA GTT CAA ATA GCT ACA GCA CAT ATT TGG
Met Arg Val Leu Arg Ser Tyr Glu Gly Ser Tyr Asn Gly Ile Tyr Glu Leu Tyr Cys Gly Lys Val Gln Ile Ala Thr Ala His Ile Trp

                   950                                            1000
                    |                                               |
GAT GGA AAA ACT GGT GAA TAT AAT GTT CCT TAT ATT GAA AGA ATG CTC CCT GGG ACA TCT GCA GTT ATA GTC CGT TTT GTT GGA AGA ATG
Asp Gly Lys Thr Gly Glu Tyr Asn Val Pro Tyr Ile Glu Arg Met Leu Pro Gly Thr Ser Ala Val Ile Val Arg Phe Val Gly Arg Met

                              1050                                            1100
                                |                                               |
CAG GGA TTC TAT GTT GCA AAG GGG AAT CCA AAG GGA ATA AAA GAT TGG AAT GAT CTT TCA AGA TCT GAC ATA GTT ATT GTA AAT AGA GAG
Gln Gly Phe Tyr Val Ala Lys Gly Asn Pro Lys Gly Ile Lys Asp Trp Asn Asp Leu Ser Arg Ser Asp Ile Val Ile Val Asn Arg Glu

                                              1150                                   1200
                                                |                                       |
AAA GGT AGT GGA ACT CGA ATT TTA TTG GAT GAA CAT TTA CGC CTA ATG AAT ATT TTA GGC AAA GAT ATA AAA GGT TAC AAT AAG GAA TGT
Lys Gly Ser Gly Thr Arg Ile Leu Leu Asp Glu His Leu Arg Leu Met Asn Ile Leu Gly Lys Asp Ile Lys Gly Tyr Asn Lys Glu Cys

                                                          1250
                                                            |
ACC TCT CAT TTA GCA ACT GCC AGT GTG ATT GCC CGC GGT AAT GCC GAC CTA GGT ATA GGA AAT GAA AAA GCA TGT TCT CAA GTA CAA GGT
Thr Ser His Leu Ala Thr Ala Ser Val Ile Ala Arg Gly Asn Ala Asp Leu Gly Ile Gly Asn Glu Lys Ala Cys Ser Gln Val Gln Gly

      1300                                                            1350
        |                                                               |
GTT GAC TTT ATA CCT ATA CAA CAA GAA AAA TAT GAT TTA GTC ATA AAA AAG GAA GAT ATA AAT CAT CCT ACT ACA AGA GCT ATT TTA GAT
Val Asp Phe Ile Pro Ile Gln Gln Glu Lys Tyr Asp Leu Val Ile Lys Lys Glu Asp Ile Asn His Pro Thr Thr Arg Ala Ile Leu Asp

ATT CTG AAT TC
Ile Leu Asn
```

FIG. 4. Nucleotide and amino acid sequences of the open reading frame upstream of *C. pasteurianum nifH2* (see Fig. 2 for map position). An overlapping region of the complementary strand is shown to complete (with Fig. 3) the sequence of the 4-kb (3,987-base-pair) *Eco*RI fragment. Sequences discussed in the text are underlined.

start codon was assigned because it would show a similar relative position to the potential ribosome-binding site as seen in the other clostridial genes. There were five TATAAT (Fig. 4, underlined) sequences upstream of the open reading frame. One of them had a sequence of TTG-N_{20}-TATAAT (−171 to −143 nucleotides), which is similar to the sequence of TTGACA-N_{17}-TATAAT before *nifH2* (Fig. 3). However, the function of these sequences remains to be determined.

The amino acid sequence of the open reading frame has one interesting feature. Between residues 17 and 59, one-third (14 of 43) of the residues were either Arg or Lys, which outnumber Asp plus Glu (5 residues) and would make this region highly positive in charge. Codon usage (53 being used) in this open reading frame is not as biased as in nitrogenase genes (see below).

Comparison of amino acid sequences encoded by *nifH1*, *nifH2*, and *nifH3* of *C. pasteurianum* and *nifH* of other organisms. An intriguing feature was noticed when the amino acid sequence encoded by *C. pasteurianum nifH2* and *nifH3* was compared with that of iron proteins from *C. pasteurianum* and eight other organisms (Fig. 5). Among the 23 differences between *C. pasteurianum nifH2* and *nifH1*, 11 (circled or boxed) of the *nifH2* residues matched the corresponding residues in at least one of the other eight Fe proteins. For residues 23 and 55 (circled), *C. pasteurianum nifH2* actually matched all of the other eight Fe proteins; residues 23, 34, 55, and 187 (circled) of *C. pasteurianum nifH3* showed a similar phenomenon. Thus, amino acid residues 23, 34, 55, and 187 are conserved in all other eight sequenced iron proteins plus *C. pasteurianum nifH2* or *nifH3* or both. Only the *C. pasteurianum* iron protein (*nifH1* product) is different. In the sequenced portion, *nifH3* encoded 14 residues (boxed or circled) that differed from both *nifH1* and *nifH2* (these residues were conserved between *nifH1* and *nifH2*) but matched the corresponding residue in at least one of the other eight iron proteins. The evolutionary implication of the highly conserved amino acid residues found in *C. pasteurianum nifH2* and *nifH3* (but not in *nifH1*),

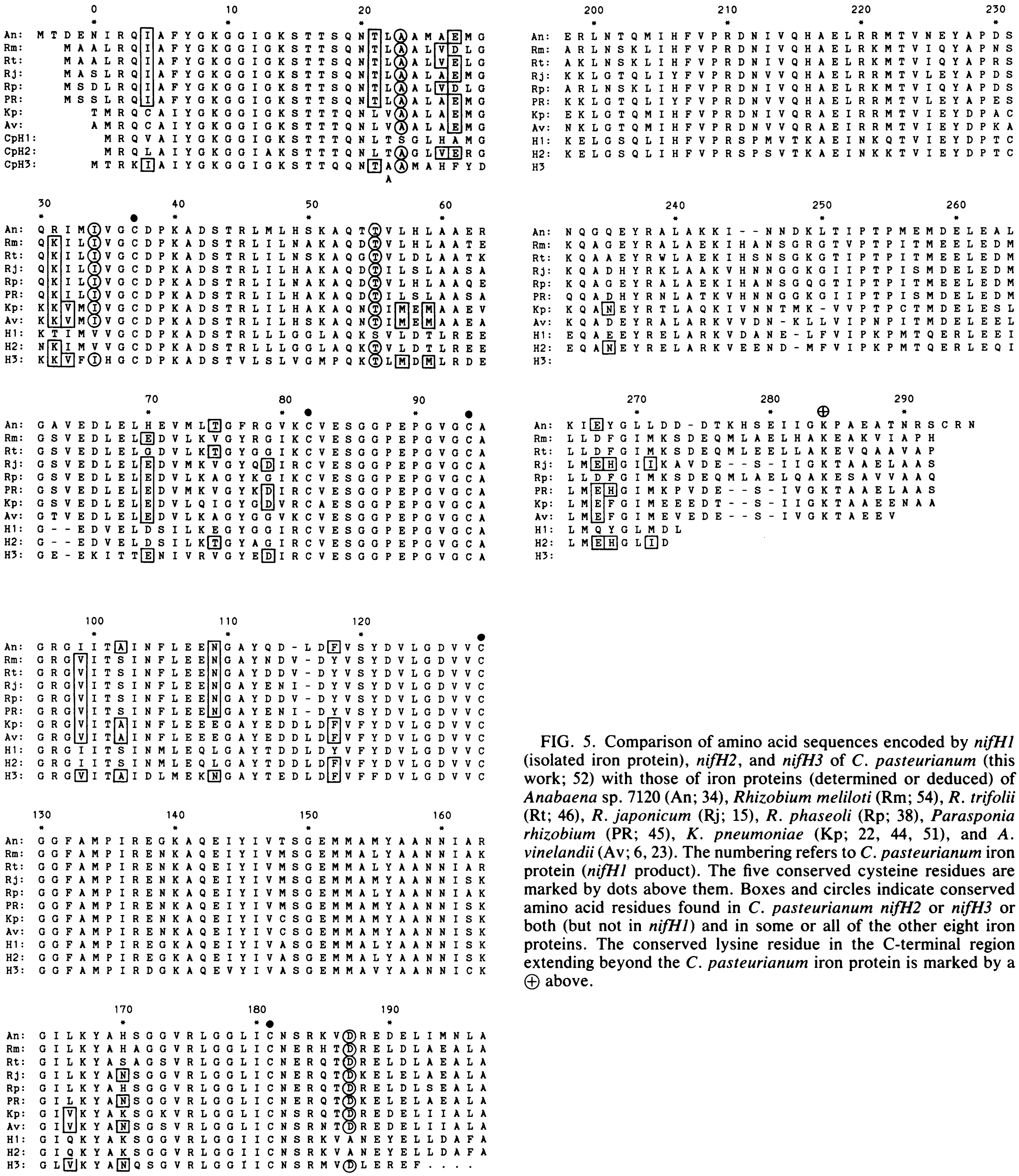

FIG. 5. Comparison of amino acid sequences encoded by *nifH1* (isolated iron protein), *nifH2*, and *nifH3* of *C. pasteurianum* (this work; 52) with those of iron proteins (determined or deduced) of *Anabaena* sp. 7120 (An; 34), *Rhizobium meliloti* (Rm; 54), *R. trifolii* (Rt; 46), *R. japonicum* (Rj; 15), *R. phaseoli* (Rp; 38), *Parasponia rhizobium* (PR; 45), *K. pneumoniae* (Kp; 22, 44, 51), and *A. vinelandii* (Av; 6, 23). The numbering refers to *C. pasteurianum* iron protein (*nifH1* product). The five conserved cysteine residues are marked by dots above them. Boxes and circles indicate conserved amino acid residues found in *C. pasteurianum nifH2* or *nifH3* or both (but not in *nifH1*) and in some or all of the other eight iron proteins. The conserved lysine residue in the C-terminal region extending beyond the *C. pasteurianum* iron protein is marked by a ⊕ above.

and *nifH* of other organisms merits further studies. Whether *C. pasteurianum nifH2* and *nifH3* function under certain specific growth conditions will be investigated.

Multiple copies of *nifH* or *nifH*-related sequences have also been found in other organisms (24, 35, 43). The nucleotide sequence of those found in *Rhizobium phaseoli* is identical in their coding regions (38), which would have different implications as compared with the different *nifH*-like sequences found in *C. pasteurianum*. The *nifH*-like sequence found in the photosynthetic gene cluster of *Rhodopseudomonas capsulata* (24) is interesting because it implicates either a current or a past electron transfer function, other than that in nitrogen-fixation, for the putative protein encoded by the *nifH*-like sequence. The presence of a nucleotide sequence similar to the *E. coli* consensus promoter before *C. pasteurianum nifH2* also raises the

TABLE 1. Comparison of codon usage in *nifH1* and *nifH2* of *C. pasteurianum* in *S. cerevisiae* mitochondria (mt; 4), and in *nifH* of *K. pneumoniae* (44)

Amino acid	Codon	No. of times codon appears in gene(s)			
		nifH1	*nifH2*	mt	*nifH*
Arg	CGA	0	0	0	0
	C	0	0	0	9
	G	0	0	0	0
	U	0	0	0	4
	AGA	12	13	28	0
	G	0	0	0	0
Leu	CUA	3	2	12	0
	C	0	0	0	7
	G	0	0	0	11
	U	9	9	2	1
	UUA	14	10	164	0
	G	0	4	2	0
Ser	UCA	6	5	46	1
	C	2	1	0	6
	G	0	1	0	2
	U	0	0	23	0
	AGC	2	1	0	1
	U	3	4	12	0
Thr	ACA	7	8	34	0
	C	0	1	1	12
	G	0	0	0	4
	U	6	5	27	0
Pro	CCA	7	7	24	1
	C	0	0	1	3
	G	0	0	0	4
	U	2	2	26	0
Ala	GCA	10	9	32	1
	C	0	0	2	14
	G	0	0	1	13
	U	10	12	47	1
Gly	GGA	18	18	15	1
	C	1	1	0	20
	G	0	1	3	1
	U	13	10	64	5
Val	GUA	9	9	54	0
	C	0	1	1	8
	G	0	0	3	12
	U	10	9	38	2
Lys	AAA	11	11	23	11
	G	5	6	1	5
Asn	AAC	4	3	4	11
	U	4	7	49	1
Gln	CAA	7	8	22	2
	G	4	2	3	8
His	CAC	0	1	0	2
	U	2	1	36	0
Glu	GAA	24	21	31	18
	G	0	3	2	11
Asp	GAC	4	4	1	11
	U	10	10	31	5
Tyr	UAC	2	3	6	8
	U	10	7	56	1
Cys	UGC	0	1	1	7
	U	6	5	11	2
Phe	UUC	4	4	38	4
	U	1	2	34	2
Ile	AUA	11	13	6	0
	C	6	6	21	17
	U	3	2	108	7
Met	AUG	11	9	45	15
Total		273	272	1,191	292

possibility that *nifH2* is not under *nif* control and that its protein product might be an electron carrier or reductase which serves or once served functions other than as a component of nitrogenase. It should be interesting to find out whether the putative *nifH2* product is synthesized in NH_3-grown cells.

Codon usage in *C. pasteurianum nifH1* and *nifH2*. Codon usage in Cp *nifH1* and *nifH2* is very biased, which is most prominent in amino acids with four to six synonymous codons (Table 1). Among the six codons for arginine, only AGA was used. Five amino acids were coded by single codons: AGA (Arg), GAA (Glu), UGU (Cys), CAU (His), and AUG (Met) (tryptophan is absent in *C. pasteurianum* iron protein). In total, 38 of the 61 codons were used in *C. pasteurianum nifH1*. The codon usage pattern is clearly different between *C. pasteurianum* and *K. pneumoniae* iron proteins (Table 1). Such a difference was also observed between the sequenced portion of *C. pasteurianum nifD* and *K. pneumoniae nifD*. Because nitrogenase components are abundantly expressed proteins in *C. pasteurianum* (60), it may be assumed that the codon usage pattern of *C. pasteurianum nifH1* reflects the distribution of isoaccepting tRNA species in this gram-positive anaerobe (25).

In *nifH1*, A and U were used more frequently at the third position of all codons. For codons of the (C/G)(C/G)(X) type, the third position was always A or U; the only exception was GGC, which was used once. However, GCC is used in the *C. pasteurianum* ferredoxin gene (18). Also, CAC and GAG were used in *C. pasteurianum nifD*, although not in *nifH1*. There is a homology of 67% between *C. pasteurianum* and *K. pneumoniae* iron proteins. At the triplet codon level, the homology was only 20% between *C. pasteurianum nifH1* and *K. pneumoniae nifH*. The low homology in nucleotide sequence affected the efficiency of *K. pneumoniae* fragment A3 (see reference 44 for its nucleotide sequence) as a probe for *C. pasteurianum nifH1*, especially because most of these homologous triplets were scattered throughout the gene. There was only one stretch each of 11, 10, and 9 nucleotides that were homologous between the *C. pasteurianum nifH1* and *K. pneumoniae nifH* genes. One stretch each of 14 and 10 homologous nucleotides was found between the pertinent portions of *C. pasteurianum* and *K. pneumoniae nifD* genes. However, there were stretches of triplets in which the first two bases matched between the *C. pasteurianum* and *K. pneumoniae* genes. The lack of longer homologous sequences between the *nif* genes of the two species explained the difficulties we encountered during the cloning of the clostridial genes. The presence of *nifH2*, which contained another set of short homologous nucleotides, on the same 4-kb *Eco*RI fragment may have enhanced hybridization between this fragment and *K. pneumoniae nifH* and facilitated detection of pCP114.

It is interesting to note that the codon usage pattern of *C. pasteurianum nifH1* is most similar to that in *Saccharomyces cerevisiae* mitochondria (4) (Table 1). The G+C content (18 to 21%) of *S. cerevisiae* mitochondria DNA (5) is close to that of *C. pasteurianum* (26 to 28%), which may explain the similar codon usage pattern. In addition, these mitochondrial genes may also be highly expressed.

The codon usage information derived from *C. pasteurianum nifH1* is so far the most complete for a *Clostridium* species with a G+C content below 30% (the *C. pasteurianum* ferredoxin gene [18] is much smaller in size and lacks several amino acids). The codon usage information could facilitate the use of more probable synthetic oligonucleotides as a probe for the cloning of genes which encode abundantly expressed proteins in *C. pasteurianum* or other clostridia of a similarly low G+C content. Organisms in the latter category include a number of industrially and medically important anaerobes such as the solvent-producing *Clostridium acetobutylicum* and *C. beijerinckii* (*C. butylicum*) and the toxin-producing *C. botulinum*, *C. difficile*, *C. perfringens*, and *C. tetani* (16, 17).

Distinct structural features of *C. pasteurianum* nitrogenase. The amino acid sequences either deduced from *nifH* (6, 15, 34, 38, 44–46, 51, 54) or determined from the iron protein (23, 52) of different organisms show a significant degree of homology, particularly in the N-terminal region (based on the *C. pasteurianum* sequence) and in the region spanning the five conserved cysteines (marked by dots, Fig. 5). Extensive regions of conserved secondary structure are predicted from the amino acid sequence of iron proteins (23). Nevertheless, the *C. pasteurianum* iron protein is uniquely inactive in heterologous combinations. A close examination of its structure may reveal regions pertinent to component interaction in nitrogenase.

The significantly different cross-reactivity between *C. pasteurianum* and other nitrogenase components (14, 48, 55) must reside in those unique amino acid sequences which give species-dependent secondary structure or surface charges or both that affect component interaction. (However, this does not exclude certain homologous sequences from being a part of the interacting regions.) Some clues were obtained by comparing the *A. vinelandii*, *K. pneumoniae*, and *C. pasteurianum* iron proteins. Although the *A. vinelandii* and *K. pneumoniae* iron proteins are highly homologous, they are not equivalent in terms of their interaction with the *C. pasteurianum* MoFe protein because the *K. pneumoniae* iron protein has some activity, whereas the *A. vinelandii* iron protein has no activity, with *C. pasteurianum* MoFe protein (14, 48).

Between the *A. vinelandii* and *K. pneumoniae* iron proteins, the main differences are (i) the N-terminal residue (Ala versus Thr); (ii) the sequence between residues 75 and 82, where a β-turn was predicted for the *A. vinelandii* (but also in *C. pasteurianum*) protein (23); and (iii) the sequence of the C-terminal region. The difference in the C-terminal region is by far more extensive, where the chain length, helical content, and charge locations are different between *A. vinelandii* and *K. pneumoniae* iron proteins. In this regard, the polypeptide chain of the *C. pasteurianum* iron protein is the shortest (shorter by 16 to 26 residues or about 6 to 10% of the total length) among the nine iron proteins sequenced so far (Fig. 5). Whether the mature iron proteins retain this size difference is yet to be shown by further protein sequence analyses, but it is now known that the C-terminal region of the *A. vinelandii* iron protein is not processed (6, 23). This study shows that the shorter polypeptide of the *C. pasteurianum* iron protein does not result from posttranslational processing. The apparent size difference among iron proteins is mainly in the C-terminal region. We thus postulate that size and charge differences in the C-terminal region have a major influence on the interaction between Fe and MoFe proteins.

Although the C-terminal region extending beyond the *C. pasteurianum* iron protein is not highly homologous, we have noticed a conserved lysine residue (Fig. 5) near the C terminus of all eight "elongated" Fe proteins. This region also contains an α-helix of various lengths in *Azotobacter*, *Klebsiella*, and *Rhizobium* species (23). The *C. pasteurianum* iron protein is thus unique in its lack of any positive residue within ten residues from its C terminus. (However, the Arg at position 260 is unique to *C. pasteurianum* and is in an α-helical region followed by a β-turn, which might serve a similar but not equivalent function as the Lys residue in the other iron proteins.)

It was postulated (51) that the GAA codon (for the Glu residue of the *K. pneumoniae* and *A. vinelandii* proteins immediately beyond the C terminus of the *C. pasteurianum* protein) in *K. pneumoniae nifH* might have been changed to TAA (stop codon) in *C. pasteurianum* to terminate translation and result in a shortened *C. pasteurianum* iron protein. The stop codon for *C. pasteurianum nifH1* is indeed TAA (Fig. 3); interestingly, there could be an Asp residue (conserved in *K. pneumoniae* and *A. vinelandii*) following TAA in *C. pasteurianum*. However, the remaining nucleotides between *nifH1* and *nifD* are not sufficiently long, and there is no homology beyond Asp between the speculated *C. pasteurianum* sequence (data not shown) and the *K. pneumoniae* and *A. vinelandii* sequences. Therefore, the distinct difference in the C-terminal region of the *C. pasteurianum* and the *K. pneumoniae* and *A. vinelandii* iron proteins is not caused by processing or by a simple conversion of a GAA into TAA.

Interestingly, the length of the N-terminal region of the α- and β-subunits (*nifD* and *nifK* products) of MoFe proteins seems proportional to that of the C-terminal region of the iron proteins in *Anabaena* sp. 7120 (28, 32, 34), *Rhizobium meliloti* (54), *Rhizobium japonicum* (26, 53), *Rhizobium trifolii* (46), *Parasponia Rhizobium* (45, 57), *K. pneumoniae* (44), *A. vinelandii* (6, 29), and *C. pasteurianum* (21; this work). The *C. pasteurianum nifD* and *nifK* proteins are the shortest in the N-terminal region, whereas the C-terminal region of the *C. pasteurianum* iron protein is also the shortest. Because of the seemingly correlated size and charge differences in the C-terminal regions of iron proteins and in the N-terminal regions of the α- and β-subunits of MoFe proteins, these regions may be examined to see whether they are sterically and electrostatically important to component interaction. Other investigators (22, 51) also postulated the involvement of the C-terminal region of the iron proteins in component interaction. At present, the *C. pasteurianum* MoFe protein shows the highest specificity for a compatible iron protein, for which the *C. pasteurianum* iron protein uniquely fits. Unique amino acid sequences, which might contribute to the specificity of the MoFe protein, have also been identified in the internal regions of the α-subunit of the *C. pasteurianum* MoFe protein (21). Through a comparison of nitrogenase proteins from *C.*

pasteurianum and other organisms and with the availability of their genes, it should be possible to carry out site-specific modifications to allow conclusive identification of regions of nitrogenase that are critical to component interaction and other functions.

ACKNOWLEDGMENTS

This work was supported by U.S. Department of Agriculture grant 84-CRCR-1-1417, by a grant from Exxon (during the initial phase of this study), and by project 6124400 from the Commonwealth of Virginia.

We thank F. Ausubel and F. Cannon for sending plasmids pSA30 and pACYC184. We also thank L. E. Mortenson for encouragement.

LITERATURE CITED

1. **Beguin, P., P. Cornet, and J.-P. Aubert.** 1985. Sequence of a cellulase gene of the thermophilic bacterium *Clostridium thermocellum*. J. Bacteriol. **162:**102–105.
2. **Biggin, M. D., T. J. Gibson, and G. F. Hong.** 1983. Buffer gradient gels and ^{35}S label as an aid to rapid sequence determination. Proc. Natl. Acad. Sci. USA **80:**3963–3965.
3. **Birnboim, H. C., and J. Doly.** 1979. A rapid alkaline extraction procedure for screening recombinant plasmid DNA. Nucleic Acids Res. **7:**1513–1523.
4. **Bonitz, S. G., R. Berlani, G. Coruzzi, M. Li, G. Macino, F. G. Nobrega, M. P. Nobrega, B. E. Thalenfeld, and A. Tzagoloff.** 1980. Codon recognition rules in yeast mitochondria. Proc. Natl. Acad. Sci. USA **77:**3167–3170.
5. **Borst, P., and R. A. Flavell.** 1975. Properties of mitochondrial DNAs, p. 363–374. *In* F. D. Fasman (ed.), Handbook of biochemistry and molecular biology, 3rd ed., Secion B: Nucleic acids, vol. 2. CRC Press, Inc., Boca Raton, Fla.
6. **Brigle, K. E., W. E. Newton, and D. R. Dean.** 1985. Complete nucleotide sequence of the *Azotobacter vinelandii* nitrogenase structural gene cluster. Gene **37:**37–44.
7. **Burgess, B. K.** 1984. Structure and reactivity of nitrogenase—an overview, p. 103–114. *In* C. Veeger and W. E. Newton (ed.), Advances in nitrogen fixation research. Klower Boston, Hingham, Mass.
8. **Cannon, F. C., G. E. Riedel, and F. M. Ausubel.** 1979. Overlapping sequences of *Klebsiella pneumoniae nif* DNA cloned and characterized. Mol. Gen. Genet. **174:**59–66.
9. **Chen, J. S., J. S. Multani, and L. E. Mortenson.** 1973. Structural investigation of nitrogenase components from *Clostridium pasteurianum* and comparison with similar components of other organisms. Biochim. Biophys. Acta **310:**51–59.
10. **Chou, P. Y., and G. D. Fasman.** 1978. Prediction of the secondary structure of proteins from their amino acid sequence. Adv. Enzymol. **47:**45–148.
11. **Cummins, C. S., and J. L. Johnson.** 1971. Taxonomy of the clostridia. Wall composition and DNA homologies in *Clostridium butyricum* and other butyric acid clostridia. J. Gen. Microbiol. **67:**33–46.
12. **Dendhardt, D. T.** 1966. A membrane-filter technique for the detection of complementary DNA. Biochem. Biophys. Res. Commun. **5:**641–646.
13. **Dixon, R. A.** 1984. The genetic complexity of nitrogen fixation. J. Gen. Microbiol. **130:**2745–2755.
14. **Emerich, D. W., and R. H. Burris.** 1978. Complementary functioning of the component proteins of nitrogenase from several bacteria. J. Bacteriol. **134:**936–943.
15. **Fuhrmann, M., and H. Hennecke.** 1984. *Rhizobium japonicum* nitrogenase Fe protein gene (*nifH*). J. Bacteriol. **158:**1005–1011.
16. **George, H. A., J. L. Johnson, W. E. C. Moore, L. V. Holdeman, and J.-S. Chen.** 1983. Acetone, isopropanol, and butanol production by *Clostridium beijerinckii* (syn. *Clostridium butylicum*) and *Clostridium aurantibutyricum*. Appl. Environ. Microbiol. **45:**1160–1163.
17. **Gottschalk, G., J. R. Andreesen, and H. Hippe.** 1981. The genus *Clostridium* (nonmedical aspects), p. 1767–1803. *In* M. P. Starr, H. Stolp, H. G. Truper, A. Balows, and H. G. Schlegel (ed.), The prokaryotes, vol. 2. Springer-Verlag, Berlin.
18. **Graves, M. C., G. T. Mullenbach, and J. C. Rabinowitz.** 1985. Cloning and nucleotide sequence determination of the *Clostridium pasteurianum* ferredoxin gene. Proc. Natl. Acad. Sci. USA **82:**1653–1657.
19. **Guth, J. H., and R. H. Burris.** 1983. Inhibition of nitrogenase-catalysed NH_3 formation by H_2. Biochemistry **22:**5111–5122.
20. **Hase, T., T. Nakano, H. Matsubara, and W. G. Zumft.** 1981. Correspondence of the larger subunit of the MoFe protein in clostridial nitrogenase to the *nifD* gene products of other N_2-fixing organisms. J. Biochem. (Tokyo) **90:**295–298.
21. **Hase, T., S. Wakabayashi, T. Nakano, W. G. Zumft, and H. Matsubara.** 1984. Structural homologies between the amino acid sequence of *Clostridium pasteurianum* MoFe protein and the DNA sequences of *nifD* and *K* genes of phylogenetically diverse bacteria. FEBS Lett. **166:**39–43.
22. **Hausinger, R. P., and J. B. Howard.** 1980. Comparison of the iron proteins from the nitrogen fixation complexes of *Azotobacter vinelandii*, *Clostridium pasteurianum*, and *Klebsiella pneumoniae*. Proc. Natl. Acad. Sci. USA **77:**3826–3830.
23. **Hausinger, R. P., and J. B. Howard.** 1982. The amino acid sequence of the nitrogenase iron protein from *Azotobacter vinelandii*. J. Biol. Chem. **257:**2483–2490.
24. **Hearst, J. E., M. Alberti, and R. F. Doolittle.** 1985. A putative nitrogenase reductase gene found in the nucleotide sequences from the photosynthetic gene cluster of *R. capsulata*. Cell **40:**219–220.
25. **Ikemura, T.** 1981. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. J. Mol. Biol. **151:**389–409.
26. **Kaluza, K., and H. Hennecke.** 1984. Fine structure analysis of the *nifDK* operon encoding the α and β subunits of dinitrogenase from *Rhizobium japonicum*. Mol. Gen. Genet. **196:**35–42.
27. **Kanemoto, R. H., L. Saari, M. Pope, T. D. Paul, R. Lowery, S. Murrell, and P. W. Ludden.** 1985. Nitrogen fixation in *Rhodospirillum rubrum*: the regulation of Fe protein and its activating enzyme, p. 253–260. *In* P. W. Ludden and J. E. Burris (ed.), Nitrogen fixation and CO_2 metabolism. Elsevier Science Publishing, Inc., New York.
28. **Lammers, P. J., and R. Haselkorn.** 1983. Sequence of the *nifD* gene coding for the α subunit of dinitrogenase from the cyanobacterium *Anabaena*. Proc. Natl. Acad. Sci. USA **80:**4723–4727.
29. **Lundell, D. J., and J. B. Howard.** 1978. Isolation and partial characterization of two different subunits from the molybdenum-iron protein of *Azotobacter vinelandii* nitrogenase. J. Biol. Chem. **253:**3422–3426.
30. **Maniatis, T., E. F. Fritsch, and J. Sambrook.** 1982. Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
31. **Marmur, J.** 1961. A procedure for isolation of deoxyribonucleic acid from microorganisms. J. Mol. Biol. **3:**208–218.
32. **Mazur, B. J., and C.-F. Chui.** 1982. Sequence of the gene coding for the β-subunit of dinitrogenase from the blue-green alga *Anabaena*. Proc. Natl. Acad. Sci. USA **79:**6782–6786.
33. **Mazur, B. J., D. Rice, and R. Haselkorn.** 1980. Identification of blue-green algal nitrogen fixation genes by using heterologous DNA hybridization probes. Proc. Natl. Acad. Sci. USA **77:**186–190.
34. **Mevarech, M., D. Rice, and R. Haselkorn.** 1980. Nucleotide sequence of a cyanobacterial *nifH* gene coding for nitrogenase reductase. Proc. Natl. Acad. Sci. USA **77:**6476–6480.
35. **Norel, F., N. Desnoues, and C. Elmerich.** 1985. Characterization of DNA sequences homologous to *Klebsiella pneumonial nifH*, *D*, *K* and *E* in the tropical *Rhizobium* OR S 571. Mol. Gen. Genet. **199:**352–356.
36. **Nuti, M. P., A. A. Lepidi, R. K. Prakash, R. A. Schilperoort, and F. C. Cannon.** 1979. Evidence for nitrogen fixation (*nif*) genes on indigenous *Rhizobium* plasmids. Nature (London) **282:**533–535.

37. **Pustell, J., and F. C. Kafatos.** 1982. A convenient and adaptable package of DNA sequence analysis programs for microcomputers. Nucleic Acids Res. **10:**51–59.

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

39. **Roberts, G. P., and W. J. Brill.** 1981. Genetics and regulation of nitrogen fixation. Annu. Rev. Microbiol. **35:**207–235.

40. **Rosenberg, M., and D. Court.** 1979. Regulatory sequences involved in the promotion and termination of RNA transcription. Annu. Rev. Genet. **13:**319–353.

41. **Ruvkun, G. B., and F. M. Ausubel.** 1980. Interspecies homology of nitrogenase genes. Proc. Natl. Acad. Sci. USA **77:**191–195.

42. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain terminating inhibitors. Proc. Natl. Acad. Sci. USA **74:**5463–5467.

43. **Scolnik, P. A., and R. Haselkorn.** 1984. Activation of extra copies of genes coding for nitrogenase in *Rhodopseudomonas capsulata*. Nature (London) **307:**289–292.

44. **Scott, K. F., B. G. Rolfe, and J. Shine.** 1981. Biological nitrogen fixation: primary structure of the *Klebsiella pneumoniae nifH* and *nifD* genes. J. Mol. Appl. Genet. **1:**71–81.

45. **Scott, K. F., B. G. Rolfe, and J. Shine.** 1983. Nitrogenase structural genes are unlinked in the nonlegume symbiont *Parasponia Rhizobium*. DNA **2:**141–148.

46. **Scott, K. F., B. G. Rolfe, and J. Shine.** 1983. Biological nitrogen fixation: primary structure of the *Rhizobium trifolii* iron protein gene. DNA **2:**149–155.

47. **Shen, S.-C., Z.-T. Xue, Q.-T. Kong, and Q.-L. Wu.** 1983. An open reading frame upstream from the *nif*H gene of *Klebsiella pneumoniae*. Nucleic Acids Res. **11:**4241–4250.

48. **Smith, B. E., R. N. F. Thorneley, R. R. Eady, and L. E. Mortenson.** 1976. Nitrogenases from *Klebsiella pneumoniae* and *Clostridium pasteurianum*. Kinetic investigations of cross reactions as a probe of the enzyme mechanism. Biochem. J. **157:**439–447.

49. **Southern, E. M.** 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol. **98:**503–517.

50. **Stormo, G. D., T. D. Schneider, and L. M. Gold.** 1982. Characterization of translational initiation sites in *E. coli*. Nucleic Acids Res. **9:**2971–2996.

51. **Sundaresan, V., and F. M. Ausubel.** 1981. Nucleotide sequence of the gene coding for the nitrogenase iron protein from *Klebsiella pneumoniae*. J. Biol. Chem. **256:**2808–2812.

52. **Tanaka, M., M. Haniu, K. T. Yasunobu, and L. E. Mortenson.** 1977. The amino acid sequence of *Clostridium pasteurianum* iron protein, a component of nitrogenase. J. Biol. Chem. **252:**7093–7100.

53. **Thöny, B., K. Kaluza, and H. Hennecke.** 1985. Structural and functional homology between the α and β subunits of the nitrogenase MoFe protein as revealed by sequencing the *Rhizobium japonicum nif*K gene. Mol. Gen. Genet. **198:**441–448.

54. **Torok, I., and A. Kondorosi.** 1981. Nucleotide sequence of the *R. meliloti* nitrogenase reductase (*nif*H) gene. Nucleic Acids Res. **9:**5711–5723.

55. **Tsai, L. B., and L. E. Mortenson.** 1978. Interaction of the nitrogenase components of *Anabaena cylindrica* with those of *Clostridium pasteurianum*. Biochem. Biophys. Res. Commun. **81:**280–287.

56. **Tumer, N. E., S. J. Robinson, and R. Haselkorn.** 1983. Different promotors for the *Anabaena* glutamine synthetase gene during growth using molecular or fixed nitrogen. Nature (London) **306:**337–342.

57. **Weinman, J. J., F. F. Fellows, P. M. Gresshoff, J. Shine, and K. F. Scott.** 1984. Structural analysis of the genes encoding the molybdenum-iron protein of nitrogenase in the *Parasponia rhizobium* strain ANU 289. Nucleic Acids Res. **12:**8329–8344.

58. **Weston, M. F., S. Kotake, and L. C. Davis.** 1983. Interaction of nitrogenase with nucleotide analogs of ATP and ADP and the effect of metal ions on ADP inhibition. Arch. Biochem. Biophys. **225:**809–817.

59. **Yun, A. C., and A. A. Szalay.** 1984. Structural genes of dinitrogenase and dinitrogenase reductase are transcribed from two separate promotors in the broad host range cowpea *Rhizobium* strain IRc78. Proc. Natl. Acad. Sci. USA **81:**7358–7362.

60. **Zumft, W. G., and L. E. Mortenson.** 1973. Evidence for a catalytic-center heterogeneity of molybdoferredoxin from *Clostridium pasteurianum*. Eur. J. Biochem. **35:**401–409.