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 (nifH) encoding the nitrogenase iron protein of *Clostridium pasteurianum* has been cloned and sequenced. It is located on a 4-kilobase EcoRI fragment (cloned into pBR325) that also contains a portion of nifD and another nifH-like sequence (nifH2). *C. pasteurianum* nifH 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 nifH 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-N7-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 273 amino 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 EcoRI 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 H₂ 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 hosts. 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
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 (nif/YK), A2 (nif/KD), and A3 (nif/DH) (41) were cloned into pBR322. Bulk plasmid DNA was isolated by a variation of the alkaline lysis 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 [32P]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 HindIII-digested lambda DNA fragments as molecular weight markers.

After preparative electrophoresis of an EcoRI digest of C. pasteurianum DNA, fragments in the desirable size ranges were isolated by using NA 45 membranes. The fragments were ligated to EcoRI-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 EcoRI 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 MgCl₂ (pH 8.5) was used as the 10× primer hybridization buffer. The 32P-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 (SAB) between two protein sequences (A, B) is defined as SAB = (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), EcoRI-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
(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-N2-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" EcoRI 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 EcoRI-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-N10-GC(A/T) (13), TCTAC (56), or TGGA-N4-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
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 EcoRI 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 *nifD*-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 EcoRI 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-N47-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 EcoRI 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 *nif* 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

---

**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.
VOL.

genes (see acid Sequences one in this outnumber highly positive (14 of 1986 4-

position overlapping region in start and open of the other TATACCATATTACTATTAAATTATTAATTTTTATTATATAACATATCAMTATATCAAATTTTCTCCATTATGATACACCATTAMTTTTTTATAGATTATTCAGTTCTATATOATATA

ATGTTATAMTCGTAAAGCTTACAAAGCTGAAATATTGTATATGATAAAATTCAMTATAAATAAAGAATATAATTMATTATAOCTTATTGTAGTTATTCATAATATATTATAAT

TTAATATTCCTCCTTAAATTACGTCCCTTTGATGT,CTAACATOGAATGATTATTTTCCACATCTCTATOCTTTATAC,rIAMTTTTTCATTATATTOCTTAATTTAGTACTTOTCAACAAA

GAT TTA

Arg GAC Asp MA GAG GOT ACC TCT CAT CTG MT ATG CGA GTT

Gly Lys Asp Gly Leu Asn Thr Ser Ala Val

ATT GTT

Glu Ser Tyr Arg Val Gly Lys Val Glu Ala Thr Arg Lys

GTG GAG AAA ACT GAT TAA GCT TAT ATT CAG TCC CTA TGG

Asp Lys Arg Leu Arg Ser Tyr Glu Ser Tyr Arg Glu Lys Ala Thr Ala His Leu Trp

AGC GCA AAA ACT GTG TAT ATT GAT ATT GCT ATT GAT ATT GTC TAT GAC TGA

Asp Lys Arg Arg Arg Asp Glu Arg Arg Glu Arg Thr Glu Thr Thr Glu Arg Thr Glu Thr Glu Arg

AAA GAT ATT CTA TGG GAA CAT TTA GCA AGA ATT ATT ATT


ACC TCT CAT TCA ACT GCT AGC ATT GCC GCT AAT GCA AAA CAT CTG TCT CAT GAA AAA CAT CTG TCT CAT GAA AAA CAT CTG TCT CAT GAA AAA CAT

Asp Leu Ala Thr Ser His Ala Thr Ser Val Ile Ala Arg Ala Asp Leu Arg Gly Ile Ala Cys Ser Gln Val Gly Lys Glu

GTT GAC TTT ATA CCT ATA CAA CAA GAA AAA TAT GAT TTA TGC ATA AAA AAG GAA GAT ATA CAT OCT ACA AGA GAT OCT TTA GAT

Val Asp Phe Ile Pro Ile Gin Gin Lys Tyr Asp Leu Ile Ile Lys Glu Asp Ile Asn His Pro Thr Arg Ala Ile Leu

ATT CTG AAT TC

Ile Leu Aam

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) EcoRI 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_{14}-TATAAT (171 to 143 nucleotides), which is similar to the sequence of TTGACA-N_{14}-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),
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

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 *Anaebaena* sp. 7120 (An; 34), *Rhizobium melloti* (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.
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)

<table>
<thead>
<tr>
<th>Amino acid</th>
<th>Codon</th>
<th>nifH1</th>
<th>nifH2</th>
<th>mt</th>
<th>nifH</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arg</td>
<td>CGA</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>U</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>AGA</td>
<td>12</td>
<td>13</td>
<td>28</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Leu</td>
<td>CUA</td>
<td>3</td>
<td>2</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>U</td>
<td>9</td>
<td>9</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>UUA</td>
<td>14</td>
<td>10</td>
<td>164</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>0</td>
<td>4</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Ser</td>
<td>UCA</td>
<td>6</td>
<td>5</td>
<td>46</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>U</td>
<td>0</td>
<td>0</td>
<td>23</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>AGC</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>U</td>
<td>3</td>
<td>4</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td>Thr</td>
<td>ACA</td>
<td>7</td>
<td>8</td>
<td>34</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>U</td>
<td>6</td>
<td>5</td>
<td>27</td>
<td>0</td>
</tr>
<tr>
<td>Pro</td>
<td>CCA</td>
<td>7</td>
<td>7</td>
<td>24</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>U</td>
<td>2</td>
<td>2</td>
<td>26</td>
<td>0</td>
</tr>
<tr>
<td>Ala</td>
<td>GCA</td>
<td>10</td>
<td>9</td>
<td>32</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>U</td>
<td>10</td>
<td>12</td>
<td>47</td>
<td>1</td>
</tr>
<tr>
<td>Gly</td>
<td>GGA</td>
<td>18</td>
<td>18</td>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>U</td>
<td>13</td>
<td>10</td>
<td>64</td>
<td>5</td>
</tr>
<tr>
<td>Val</td>
<td>GUA</td>
<td>9</td>
<td>9</td>
<td>54</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>U</td>
<td>10</td>
<td>9</td>
<td>38</td>
<td>2</td>
</tr>
<tr>
<td>Lys</td>
<td>AAA</td>
<td>11</td>
<td>11</td>
<td>23</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>5</td>
<td>6</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Asn</td>
<td>AAC</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>U</td>
<td>4</td>
<td>7</td>
<td>49</td>
<td>1</td>
</tr>
<tr>
<td>Gln</td>
<td>CAA</td>
<td>7</td>
<td>8</td>
<td>22</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>His</td>
<td>CAC</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>U</td>
<td>2</td>
<td>1</td>
<td>36</td>
<td>0</td>
</tr>
<tr>
<td>Glu</td>
<td>GAA</td>
<td>24</td>
<td>21</td>
<td>31</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>Asp</td>
<td>GAC</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>U</td>
<td>10</td>
<td>10</td>
<td>31</td>
<td>5</td>
</tr>
<tr>
<td>Tyr</td>
<td>UAC</td>
<td>2</td>
<td>3</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>U</td>
<td>10</td>
<td>7</td>
<td>56</td>
<td>1</td>
</tr>
</tbody>
</table>

TABLE 1—Continued

<table>
<thead>
<tr>
<th>Amino acid</th>
<th>Codon</th>
<th>nifH1</th>
<th>nifH2</th>
<th>mt</th>
<th>nifH</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cys</td>
<td>UGC</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>U</td>
<td>6</td>
<td>5</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>Phe</td>
<td>UUC</td>
<td>4</td>
<td>4</td>
<td>38</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>U</td>
<td>1</td>
<td>2</td>
<td>34</td>
<td>2</td>
</tr>
<tr>
<td>Ile</td>
<td>AUA</td>
<td>11</td>
<td>13</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>6</td>
<td>6</td>
<td>21</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>U</td>
<td>3</td>
<td>2</td>
<td>108</td>
<td>7</td>
</tr>
<tr>
<td>Met</td>
<td>AUG</td>
<td>11</td>
<td>9</td>
<td>45</td>
<td>15</td>
</tr>
</tbody>
</table>

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 NH₃-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 GCC, 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 EcoRI 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 so far the most complete for a Clostridium species with a G+C content below 30% (the C. pasteurianum ferredoxin gene 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 A. vinelandii) 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 mellioti (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. In the present study, 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-CRRC-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 E. L. Mortenson for encouragement.

LITERATURE CITED


