Proc. Natl. Acad. Sci. USA
Vol. 79, pp. 6782–6786, November 1982
Biochemistry

Sequence of the gene coding for the β-subunit of dinitrogenase from the blue-green alga *Anabaena*

(MoFe protein/FeS proteins/cyanobacteria/ribulosebisphosphate carboxylase/endosymbiotic hypothesis)

BARBARA J. MAZUR AND CHOK-FUN CHUI

Central Research and Development Department, E. I. du Pont de Nemours and Company, Wilmington, Delaware 19898

Communicated by H. E. Simmons, Jr., July 23, 1982

ABSTRACT **The nitrogen fixation *nifK* gene of the blue-green alga *Anabaena*, which codes for the β-subunit of dinitrogenase, has been subjected to sequence analysis. The *nifK* protein is predicted to be 512 amino acids long, to have a M_r of 57,583, and to contain six cysteine residues. Three of these cysteines are within peptides homologous to FeS cluster-binding cysteinyl peptides from ferredoxins and from a high potential iron protein and, thus, may be ligands to which FeS clusters bind in dinitrogenase. The sequences surrounding the cysteine residues are 70% homologous to the corresponding cysteinyl tryptic peptides of the *Azotobacter vinelandii* dinitrogenase, although the positions of the cysteine residues are not always conserved between the two proteins. A 15-amino acid coding sequence precedes *nifK* on its transcript. Amino acid codon usage is highly asymmetric and parallels that found for the *Anabaena* dinitrogenase reductase gene (*nifH*). Putative promoter and ribosome binding site sequences were identified for *nifK*. These regulatory sequences are homologous to sequences preceding *nifD*; *nifD* codes for the α-subunit of dinitrogenase but is separated from *nifK* on the chromosome by 11,000 nucleotides. The *nifK* promoter also is virtually identical to a promoter-like sequence that immediately precedes the start of the transcript for the large subunit of ribulosebisphosphate carboxylase from maize chloroplasts. This homology appears to support the theory that chloroplasts evolved from blue-green algae.**

Nitrogenase catalyzes the reduction of atmospheric nitrogen to ammonia. The nitrogenase complex is an association between two enzymes—the dinitrogenase, or MoFe protein, and the dinitrogenase reductase, or Fe protein. Dinitrogenase is itself multimeric, being composed of two α-subunits of M_r ≈56,000 and two β-subunits of M_r ≈60,000. Complexed to it are four Fe_4S_4 clusters and two molecules of a MoFe cofactor; the latter is the active site for nitrogen reduction. Dinitrogenase is reduced by dinitrogenase reductase, a dimer of M_r 33,000 subunits (1).

Seventeen genes are required for nitrogen fixation in *Klebsiella pneumoniae*, the only organism for which a detailed genetic map of the nitrogen fixation (*nif*) genes has been constructed. The *nifK* and *nifD* genes code for the β- and α-subunits of dinitrogenase, respectively, whereas the *nifH* gene codes for the nitrogenase reductase. The remaining genes are required for activation of these enzymes, for formation of the MoFe cofactor, for electron transport, and for regulation of the *nif* genes (2).

We have been studying the regulation of nitrogen fixation in *Anabaena* 7120, a photosynthetic blue-green alga that can fix simultaneously both carbon and nitrogen. No methods for genetic transformation have been identified for *Anabaena*, hampering construction of a genetic map. This difficulty was circumvented in the case of the *nif* genes by exploiting a homology between *Anabaena* and *Klebsiella nif* genes. By hybridizing cloned *Anabaena* and *Klebsiella nif* genes to each other, we were able to construct a limited map of the *Anabaena nif* genes (3). Surprisingly, the *Anabaena* and *Klebsiella nif* genes are rearranged relative to each other. Particularly striking are the different locations of the three nitrogenase structural genes, which are linked in a single operon in *Klebsiella* but not in *Anabaena*. Instead, in *Anabaena*, *nifH*—which codes for the dinitrogenase reductase—is linked to *nifD*—which codes for the α-subunit of dinitrogenase—whereas *nifK*—which codes for the β-subunit of dinitrogenase—is separated from *nifD* by 11 kilobase pairs (kbp) (4).

The *nifK* sequence that we report here reveals one way in which these unlinked genes can be coordinately regulated. The putative promoter sequences for the *nifK* and *nifH–nifD* transcripts are homologous, as are their ribosome binding sites. Surprisingly, the putative *nifK* promoter also is virtually identical to a promoter-like sequence that immediately precedes the start of transcription of the gene for the large subunit of ribulosebisphosphate carboxylase (RuP_2Case) from maize chloroplasts (5), lending support to the theory that chloroplasts evolved from blue-green algae (6).

MATERIALS AND METHODS

DNA Sequence Analysis. Chemical sequence analysis was performed according to the procedures of Maxam and Gilbert (7). Chain termination sequence analysis was performed according to the methods of Sanger *et al.* (8).

RESULTS AND DISCUSSION

DNA Sequence of *nifK*. A map of *nifK* is shown in Fig. 1. The gene was defined through a series of hybridizations between cloned fragments of *Klebsiella nifK* (9) and *Anabaena* DNA. A 1-kbp region of homology was found on three adjacent *Hin*dIII subfragments of a 17-kbp *Anabaena Eco*RI clone (4). These three *nifK*-containing subfragments, 1, 0.7, and 3 kb in size, were cloned into pBR322 (10) and into M13 mp5 (11) and were subjected to sequence analysis by the chemical and chain termination methods, respectively (7, 8). Overlapping DNA sequences were obtained at all restriction sites, including the *Hin*dIII sites that separate the subclones. Sequences were confirmed either by sequence analysis of the complementary strand or by using the complementary sequence analysis technique.

The 2,200 base pairs (bp) that flank and include the *nifK* gene are shown in Fig. 2. The direction of transcription of the gene was determined by hybridizing each of the M13 subclones with

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U. S. C. §1734 solely to indicate this fact.

Abbreviations: *nif*, nitrogen fixation; kb, kilobase(s); kbp, kilobase pair(s); bp, base pair(s); RuP_2Case, ribulosebisphosphate carboxylase.

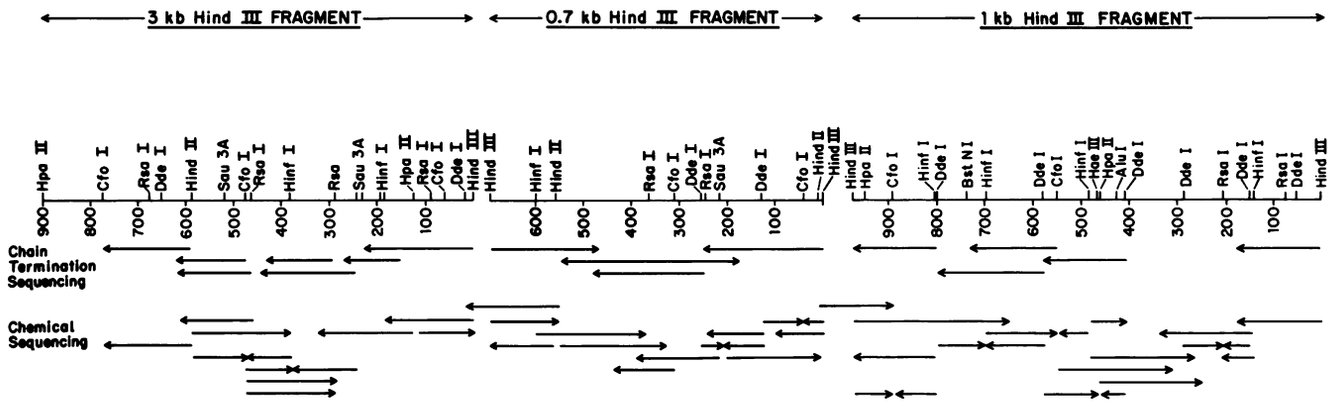

FIG. 1. Restriction endonuclease map of *nifK* and the sequence analysis strategy. The empirically derived restriction enzyme maps of the three *Hin*dIII subclones of pAn207 that encompass *nifK* are shown. The arrows designate the direction and extent of sequence analysis reactions. kb, Kilobase(s).

a pAn207 probe labeled at one of its 5′ termini; the direction of transcription within An207 had been established through DNA heteroduplex experiments between An207 and *Klebsiella nifK* (4). The direction of transcription derived from these experiments contains the only open reading frame in either direction. All other reading frames have multiple termination codons interspersed throughout the sequence.

We examined the DNA sequence 5′ to *nifK* and 5′ to *Anabaena nifH* for common regulatory elements. The *Anabaena nifH* gene had been subjected to sequence analysis by Mevarech *et al.* (12). The most highly conserved sequence in this region is the sequence (C/A)C-T-A-A-T-T-C, 200 bp upstream from the *nifK* coding sequence (at positions 337–344) and 140 bp upstream from the *nifH* coding sequence. We believe this sequence is part of the RNA polymerase binding site, as it appears to correspond to the −10 region of the consensus *Escherichia coli* promoter sequence, T-A-T-A-A-T (13). The other critical region for RNA polymerase binding, located 35 bp upstream from the site of transcription initiation, also bears homologies between the putative *nifK* and *nifH* sequences. The *nifK* sequence has C-T-G-C-A-A-A, whereas that for *nifH* has C-T-A-C-A-T-A. The identification of this sequence as a *nifH* promoter, in fact, has been recently confirmed by S. J. Robinson and R. Haselkorn (personal communication), who have found that the *in vivo* initiation site for the *nifH–nifD* transcript is a guanosine, 7 or 8 bp downstream from the putative −10 sequence. These homologous *nifK* and *nifH* promoter-like sequences are shown in Fig. 3. We expect that the homology in these sequences plays a role in the coordinate control of the *nifK* and *nifH–nifD* genes.

We also compared the *nifK* sequence with the 5′ sequence of the gene for the large subunit of RuP_2Case from maize chloroplasts. We were prompted to do so by the hypothesis that chloroplasts arose from a symbiosis between blue-green algae and the progenitors of eukaryotic cells (6). The *in vivo* transcription initiation site for the maize RuP_2Case gene has been identified by McIntosh *et al.* (5). Based upon this identification, the sequence C-C-T-A-A-T precedes the initiation site by three nucleotides and, thus, could be part of the −10 recognition region of the carboxylase promoter. This sequence is identical to the putative −10 sequence for *nifK*, identified by its homology with *nifH*. The −35 sequences of these two genes, T-G-C-A-A-A-G-G, also are identical, as shown in Fig. 3. However, the RuP_2Case sequence is lacking three bases between the −10 and −35 regions, so that if this is the actual site of polymerase binding, different spatial constraints would apply to the chloroplast polymerase than apply to the algal and *E. coli* polymerases. Interestingly, the noncoding region of the spinach large subunit RuP_2Case gene also has this same sequence (14). However, the spinach gene has a 4-bp insertion within the conserved −35 sequence, as shown in Fig. 3, and transcription does not appear to initiate downstream from it. It is tempting to speculate that this insertion hinders chloroplast RNA polymerase recognition.

The conservation of these 5′ noncoding sequences, for two unrelated genes, is striking. When combined with the evidence that these sequences may play a role in transcription initiation, this homology appears to support the hypothesis that chloroplasts and blue-green algae are evolutionarily related. This homology also appears to be fortuitous with regard to attempts to express nitrogenase genes in higher plants. Because RuP_2Case is the most abundant soluble leaf protein, the *Anabaena nif* genes may have the potential to be heavily transcribed in plant chloroplasts.

The Shine–Dalgarno (15) sequences of the ribosome binding sites also are largely homologous between *nifK* and *nifH*, with 10 of 16 bases preceding the initial A-T-G identical between the two. Again, this homology may contribute to the coordinated expression of the separated nitrogenase subunits.

The 5′ noncoding region of the *nifK* message is capable of considerable secondary structure. The 32-bp sequence from positions 357–388 is an inverted repeat that can form a highly stable hairpin, with a free energy change, $\Delta G°$ (25°C), of −27 kcal (1 cal = 4.184 J), based on the rules of Tinoco *et al.* (16). Other inverted repeats capable of forming stable hairpins are the 61-base stretch from positions 392–450 and the 17-base stretch from positions 468–484. Two tandem direct repeats, one of 6 bp and one of 7 bp, also are present in this region at positions 321–332 and positions 402–415, respectively.

The 5′ transcribed region also encodes a small polypeptide from positions 397–441. This sequence may be translated, as there is a Shine–Dalgarno sequence located upstream from it. Its ribosome binding site would be included in the aforementioned 32-bp hairpin; the *Klebsiella nifH* ribosome binding site also appears to be in a hairpin structure (17). The resulting polypeptide would contain 15 amino acids with the sequence Met-Val-Asn-Ala-Glu-Cys-Trp-Met-Lys-Ile-Phe-Ile-His-Tyr-Ser. What role, if any, this polypeptide would have remains unknown; neither it nor the adjoining sequence can form structures that resemble bacterial attenuators. Because two in-phase UGA termination codons intercede between it and the designated start of *nifK*, we consider it unlikely that read-through to *nifK* occurs. It is possible that the β-subunit once may have initiated at this site, to form a protein with the canonical β-subunit of M_r 60,000 (18–20). Alternatively, this polypeptide could

```
GAGAAAACTG CCGCTGGTAG ACGAAAGTGG CTCTAAGTCT GCAAAGGCTT GTCGATATTT GTCTTGACCC TGATTTTGCA TCGCTGTGGT ATTAGCCTAT
        10         20         30         40         50         60         70         80         90        100
ATTTAGCCTA AAAATTAATG TGTTATCAGC AAACAATGTT CATCACTAAC ACTGCTCAGT GCAAACATTA AGCTGTTGAA AGCCATTAAA CCACAAAAAG
       110        120        130        140        150        160        170        180        190        200
GATTACTCCG GCCCTTATCA CGGTTACCAC GGATTCGCTA TCTTCGCCCG TGACATGGAT TTAGCCCTCA ACAGCCCAAC TTGGAGCTTG ATTGGCGCTC
       210        220        230        240        250        260        270        280        290        300
CTTGGAAGAA AGCGGCTGCA AAGGCTAAGG CTGCTGCCTA ATTCCAGGTA AATAGAGGGA TAGCCTGGGG TTGTTGCCCC AGGAACCCAG GGAAGAATGG
       310        320        330        340        350        360        370        380        390        400
TGAATGCTGA ATGCTGGATG AAGATTTTTA TTCATTATTC ATGATTCATC ACTCGTTACT ACCTTGAGGG GGAGTGAACC TCCCAGGCTA TCCTCACTCA
       410        420        430        440        450        460        470        480        490        500
                                                  Met Pro Gln Asn Pro Glu Arg Thr Val Asp His Val Asp Leu Phe Lys Gln
TCACTTACAA ACCAACCAGC AAGCGTAGAG AGATACAACA ATG CCT CAG AAT CCA GAA AGA ACT GTA GAC CAC GTT GAT CTA TTC AAA CAG
       510        520        530        540          550        560        570        580        590

Pro Glu Tyr Thr Glu Leu Phe Glu Asn Lys Arg Lys Asn Phe Glu Gly Ala His Pro Pro Glu Glu Val Glu Arg Val Ser Glu
CCA GAA TAC ACC GAG CTA TTT GAA AAC AAG AGA AAG AAC TTT GAA GGC GCT CAT CCT CCT GAA GAA GTT GAA AGA GTG TCT GAA
       600        610        620        630        640        650        660        670

Trp Thr Lys Ser Trp Asp Tyr Arg Glu Lys Asn Phe Ala Arg Glu Ala Leu Thr Val Asn Pro Ala Lys Gly Cys Gln Pro Val
TGG ACA AAA TCT TGG GAC TAC CGG GAA AAG AAC TTC GCT CGT GAA GCT TTA ACC GTT AAC CCT GCT AAA GGT TGC CAA CCT GTA
   680        690        700        710        720        730        740        750

Gly Ala Met Phe Ala Ala Leu Gly Phe Glu Gly Thr Leu Pro Phe Val Gln Gly Ser Gln Gly Cys Val Ala Tyr Phe Arg Thr
GGC GCG ATG TTC GCT GCT TTG GGT TTT GAA GGT ACT CTA CCT TTC GTA CAA GGT TCT CAA GGT TGC GTT GCT TAC TTC CGT ACA
          770        780        790        800        810        820        830        840

His Leu Ser Arg His Tyr Lys Glu Pro Cys Ser Ala Val Ser Ser Ser Met Thr Glu Asp Ala Ala Val Phe Gly Gly Leu Asn
CAC CTC AGC CGT CAC TAC AAA GAG CCT TGC TCC GCA GTA TCT TCT TCC ATG ACA GAA GAT GCA GCA GTA TTC GGT GGT TTG AAC
      850        860        870        880        890        900        910        920

Asn Met Ile Glu Gly Met Gln Val Ser Tyr Gln Leu Tyr Lys Pro Lys Met Ile Ala Val Cys Thr Thr Cys Met Ala Glu Val
AAC ATG ATC GAA GGT ATG CAG GTT TCA TAC CAA CTG TAC AAG CCT AAG ATG ATT GCT GTT TGC ACC ACC TGT ATG GCG GAA GTT
930        940        950        960        970        980        990       1000       1010

Ile Gly Asp Asp Leu Gly Ala Phe Ile Thr Asn Ser Lys Asn Ala Gly Ser Ile Pro Gln Asp Phe Pro Val Pro Phe Ala His
ATC GGA GAT GAC TTG GGC GCG TTC ATC ACC AAC TCC AAG AAC GCT GGT TCT ATT CCT CAA GAT TTC CCC GTA CCC TTT GCT CAC
       1020       1030       1040       1050       1060       1070       1080       1090

Thr Pro Ser Phe Val Gly Ser His Ile Thr Gly Tyr Asp Asn Met Met Lys Gly Ile Leu Ser Asn Leu Thr Glu Gly Lys Lys
ACA CCT AGC TTT GTT GGT TCC CAC ATC ACT GGC TAC GAC AAC ATG ATG AAG GGT ATT CTG TCT AAC TTG ACA GAA GGT AAG AAG
  1100       1110       1120       1130       1140       1150       1160       1170

Lys Ala Thr Ser Asn Gly Lys Ile Asn Phe Ile Pro Gly Phe Asp Thr Tyr Val Gly Asn Asn Arg Glu Leu Lys Arg Met Met
AAA GCT ACC AGC AAC GGC AAA ATT AAC TTC ATT CCT GGT TTT GAT ACC TAT GTA GGT AAC AAC CGC GAA TTG AAG CGG ATG ATG
          1190       1200       1210       1220       1230       1240       1250       1260

Gly Val Met Gly Val Asp Tyr Thr Ile Leu Ser Asp Ser Ser Asp Tyr Phe Asp Ser Pro Asn Met Gly Glu Tyr Glu Met Tyr
GGT GTA ATG GGT GTT GAC TAC ACC ATC CTG TCT GAC AGC AGC GAC TAC TTT GAT TCA CCT AAC ATG GGT GAA TAC GAA ATG TAC
      1270       1280       1290       1300       1310       1320       1330       1340

Pro Ser Gly Thr Lys Leu Glu Asp Ala Ala Asp Ser Ile Asn Ala Lys Ala Thr Val Ala Leu Gln Ala Tyr Thr Thr Pro Lys
CCA AGT GGT ACA AAG CTG GAA GAT GCG GCT GAT TCT ATC AAC GCT AAA GCA ACT GTT GCT CTC CAA GCT TAC ACC ACA CCT AAG
          1360       1370       1380       1390       1400       1410       1420       1430

Thr Arg Glu Tyr Ile Lys Thr Gln Trp Lys Gln Glu Thr Gln Val Leu Arg Pro Phe Gly Val Lys Gly Thr Asp Glu Phe Leu
ACC CGC GAA TAC ATC AAA ACC CAG TGG AAG CAA GAA ACA CAA GTA TTG CGC CCC TTC GGT GTT AAG GGT ACT GAC GAG TTC TTG
       1440       1450       1460       1470       1480       1490       1500       1510

Thr Ala Val Ser Glu Leu Thr Gly Lys Ala Ile Pro Glu Glu Leu Glu Ile Glu Arg Gly Arg Leu Val Asp Ala Ile Thr Asp
ACT GCT GTT TCT GAA TTG ACC GGT AAA GCT ATT CCT GAA GAA TTG GAA ATC GAA CGC GGT CGT TTA GTT GAT GCT ATC ACT GAC
  1520       1530       1540       1550       1560       1570       1580       1590

Ser Tyr Ala Trp Ile His Gly Lys Lys Phe Ala Ile Tyr Gly Asp Pro Asp Leu Ile Ile Ser Ile Thr Ser Phe Leu Leu Glu
TCC TAC GCT TGG ATT CAT GGT AAG AAG TTC GCT ATC TAC GGC GAT CCA GAT TTG ATC ATC TCC ATC ACC AGC TTC TTG TTA GAA
          1610       1620       1630       1640       1650       1660       1670       1680

Met Gly Ala Glu Pro Val His Ile Leu Cys Asn Asn Gly Asp Asp Thr Phe Lys Lys Glu Met Glu Ala Ile Leu Ala Ala Ser
ATG GGT GCT GAA CCA GTA CAC ATC CTC TGC AAC AAC GGT GAT GAC ACC TTC AAG AAA GAA ATG GAA GCT ATC CTC GCT GCT AGC
      1690       1700       1710       1720       1730       1740       1750       1760

Pro Phe Gly Lys Glu Ala Lys Val Trp Ile Gln Lys Asp Leu Trp His Phe Arg Ser Leu Leu Phe Thr Glu Pro Val Asp Phe
CCA TTT GGT AAA GAA GCT AAA GTC TGG ATT CAA AAA GAC TTG TGG CAC TTC CGT TCC TTG TTG TTC ACC GAG CCT GTA GAC TTC
          1780       1790       1800       1810       1820       1830       1840       1850

Phe Ile Gly Asn Ser Tyr Gly Lys Tyr Leu Trp Arg Asp Thr Ser Ile Pro Met Val Arg Ile Gly Tyr Pro Leu Phe Asp Arg
TTC ATC GGT AAC TCC TAC GGT AAG TAC CTG TGG CGC GAT ACC AGC ATC CCA ATG GTG CGG ATT GGT TAT CCT CTC TTC GAT CGC
       1860       1870       1880       1890       1900       1910       1920       1930

His His Leu His Arg Tyr Ser Thr Leu Gly Tyr Gln Gly Gly Leu Asn Ile Leu Asn Trp Val Val Asn Thr Leu Leu Asp Glu
CAC CAC TTA CAC CGC TAT TCT ACC CTC GGC TAC CAA GGT GGT CTA AAT ATC CTC AAC TGG GTT GTT AAC ACC CTG TTG GAT GAA
  1940       1950       1960       1970       1980       1990       2000       2010

Met Asp Arg Ser Thr Asn Ile Thr Gly Lys Thr Asp Ile Ser Phe Asp Leu Ile Arg
ATG GAT CGC AGC ACC AAC ATC ACT GGT AAG ACC GAT ATC TCC TTT GAC TTG ATC CGC TAGA AATTAATGCA GCGTGCCATT
          2030       2040       2050       2060       2070       2080       2090       2100

GAAAGGTAGA ACTTAGGGAC TGGGGATTGG GTATTGGGTA CTAGGAATAT TATCTTCCCA GTCCCTTCCA GTCCCCGAGA CCCTTTG
      2110       2120       2130       2140       2150       2160       2170       2180
```

FIG. 2. The nucleotide sequence of *nifK* of *Anabaena*. The derived amino acid sequence of the β-subunit of dinitrogenase is also shown above residues 323–1,861.

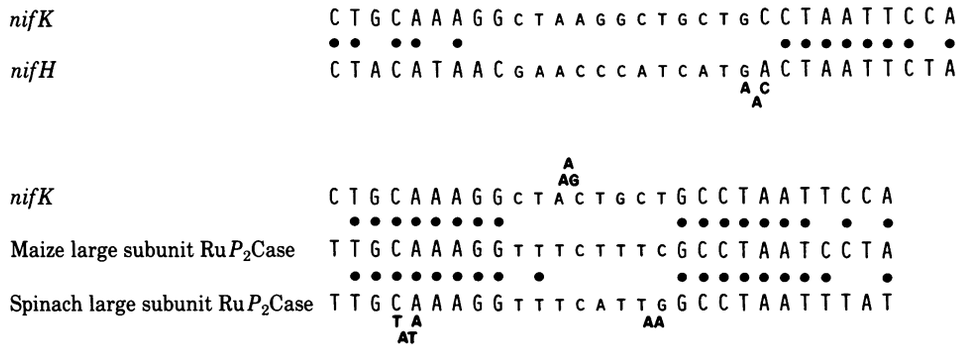

FIG. 3. Comparison of the putative *nifK* and *nifH* promoter sequence from *Anabaena* and comparison of this *Anabaena nifK* sequence with 5′ sequences that precede the maize and spinach chloroplast RuP_2Case large subunit genes. Identical bases are denoted by filled circles.

be involved in synthesis of the FeMo cofactor or in *nif* regulation; clarification of this must await a better system of genetic analysis for *Anabaena*.

A transcription termination sequence, G-C-C-T-A-A-A-A-T-T-A-A-T, occurs 230 bases 5′ to the *nifK* promoter sequence, at position 107. This sequence is similar to the ϕX174 termination sites identified by McMahon and Tinoco, which are characterized by a most stable nearest neighbor pair, G·G or C·C, followed by a region in which five of the next seven nearest neighbors are the least stable pairs, A·A or T·T (21). It differs from the classical transcription termination sites identified in *E. coli*, in which G + C-rich hairpins precede a string of Ts (13). We did not identify any transcription termination sites in the 100 bp 3′ to *nifK*. The identity of the genes flanking *nifK* is unknown.

Amino Acid Sequence of the β-Subunit of Dinitrogenase. The *nifK* polypeptide is 512-amino acid residues long, forming a protein of M_r 57,583. Its amber termination codon is immediately preceded by termination codons in the other two reading frames, so that a frameshift sequence analysis error could not have altered the molecular weight of the protein. The protein has six cysteines, five of which are clustered within an 80-residue stretch towards the amino terminus (at residues 70, 95, 111, 150, and 153), with the remaining cysteine at residue 391, towards the COOH terminus. Among the other amino acids, a number of the histidine and tryptophan residues are clustered at the COOH terminus of the protein. The central third of the protein is devoid of histidine residues and contains only one tryptophan. Tyrosines are arrayed throughout the protein. The protein has 9 hydrophobic regions of 10 or more residues, and 3 of these, at residues 179–197, 371–380, and 471–491, have particularly high hydrophobicity indices (1.95, 3, and 1.76, respectively) (22).

The cysteinyl tryptic peptides of the *Azotobacter vinelandii* dinitrogenase recently have been subjected to sequence analysis by Lundell and Howard (23). The *Azotobacter* β-subunit (α in their nomenclature) has eight cysteines in six tryptic peptides; five of the tryptic peptides are 72% homologous at the amino acid level to the corresponding *Anabaena* peptides. Two-thirds of the amino acid substitutions within these peptides can be accounted for by single base changes, and only one substitution requires a triple base change in the codon. Yet, despite the high homology between the cysteinyl peptides, two of the *Azotobacter* cysteines are not found in the *Anabaena* protein,

		Nif K	Nif H			Nif K	Nif H			Nif K	Nif H			Nif K	Nif H
UUU	Phe	9	0	UCU		11	4	UAU	Tyr	3	2	UGU	Cys	1	2
UUC		19	6	UCC	Ser	9	5	UAC		18	7	UGC		5	4
UUA	Leu	4	7	UCA		2	1	UAA	ochre	0	0	UGA	opal	0	0
UUG		16	7	UCG		0	0	UAG	amber	1	1	UGG	Trp	8	0
CUU		0	2	CCU		16	4	CAU	His	2	0	CGU		5	11
CUC	Leu	7	2	CCC	Pro	3	2	CAC		10	7	CGC	Arg	9	5
CUA		4	4	CCA		7	3	CAA	Gln	10	9	CGA		0	0
CUG		6	2	CCG		0	0	CAG		4	1	CGG		3	0
AUU		9	6	ACU		8	1	AAU	Asn	2	2	AGU	Ser	1	0
AUC	Ile	21	15	ACC	Thr	20	14	AAC		22	14	AGC		9	2
AUA		0	1	ACA		8	1	AAA	Lys	13	6	AGA	Arg	3	1
AUG	Met	18	15	ACG		0	0	AAG		19	8	AGG		0	1
GUU		15	6	GCU		25	16	GAU	Asp	17	1	GGU		32	26
GUC	Val	1	0	GCC	Ala	0	2	GAC		13	16	GGC	Gly	7	3
GUA		11	11	GCA		4	8	GAA	Glu	32	22	GGA		1	0
GUG		2	0	GCG		4	1	GAG		4	3	GGG		0	0

FIG. 4. Codon usage in *Anabaena nifK* and *nif*H^{12}.

and two more are at different positions. One of the missing cysteines is within a short peptide that is not homologous to the *Anabaena* protein, and the other is within a homologous region of the *Anabaena* protein but is an alanine in *Anabaena* (residue 384, nucleotide 1,690). The *Anabaena* cysteine at residue 111 (nucleotide 871) is a valine in *Azotobacter*, whereas *Azotobacter*, instead, has a cysteine two residues downstream, again at the position of an *Anabaena* alanine. Likewise, the first cysteine in the peptide Cys-Thr-Thr-Cys (Cys-150, nucleotide 988) is missing in *Azotobacter*, whereas *Azotobacter* once again has a cysteine at the position of an *Anabaena* alanine (residue 78, nucleotide 772).

The absence of this Cys-X-X-Cys peptide in *Azotobacter* is somewhat surprising, in that the cysteines of such peptides ligand FeS clusters in ferredoxins and in a high potential iron protein (24, 25) and also might be expected to ligand FeS clusters in dinitrogenase. One other *Anabaena* tetrapeptide surrounding a cysteine also is homologous to an FeS cluster binding cysteine in a high potential iron protein, the peptide Lys-Gly-Cys-Gln (Cys-70, nucleotide 748). The cysteine in this peptide is conserved in *Azotobacter*, although the adjacent glycine is instead an alanine. The adjacent glutamine is followed by Pro-Val; Lundell and Howard have postulated that such Pro-Val residues should provide an important secondary structure for ligand folding around FeS centers (23). It will be of interest to learn whether either of the translocated cysteines can perform equivalent functions in the *Anabaena* and *Azotobacter* proteins and to learn which cysteines in fact ligand FeS clusters. Unlike the cysteinyl peptides, the NH_2- more and COOH-terminal peptides are not homologous between the *Anabaena* and *Azotobacter* β-subunit proteins (20).

Codon Usage in the *nifK* Gene. The codon usage for the *nifK* protein is strikingly asymmetric. Fig. 4 tabulates the usage and compares it to the only other blue-green algal protein that has been subjected to sequence analysis, *nifH* (12). There are nine codons that are never used in *nifK*, and five of these also are never used in *nifH*. An additional eight codons are used but once or twice in *nifK* and *nifH*. Many of the remaining codons are used only a few times, or else heavily. This bias is not peculiar to the *nif* genes, as *nifH* from *Klebsiella* exhibits a different pattern of codon usage (17). Instead, it may reflect the distribution of tRNA populations in algal cells or in their heterocysts (26).

We thank Steven Robinson and Robert Haselkorn for permission to cite their unpublished data.

1. Mortenson, L. E. & Thorneley, R. N. F. (1979) *Annu. Rev. Biochem* **48**, 387–418.
2. Roberts, G. P. & Brill, W. J. (1981) *Annu. Rev. Microbiol.* **35**, 207–235.
3. Mazur, B. J., Rice, D. & Haselkorn, R. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 186–190.
4. Rice, D., Mazur, B. J. & Haselkorn, R. (1982) *J. Biol. Chem.* **257**, in press.
5. McIntosh, L., Poulsen, C. & Bogorad, L. (1980) *Nature (London)* **288**, 556–560.
6. Ris, H. & Plaut, W. (1962) *J. Cell Biol.* **13**, 383–391.
7. Maxam, A. M. & Gilbert, W. (1980) *Methods Enzymol.* **65**, 499–560.
8. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467.
9. Riedel, G. E., Ausubel, F. M. & Cannon, F. C. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 2866–2870.
10. Bolivar, F., Rodriguez, R., Greene, P., Betlach, M., Heyneker, H., Boyer, H., Crosa, J. & Falkow, S. (1977) *Gene* **2**, 95–113.
11. Messing, J. (1979) *Recomb. DNA Tech. Bull.* **2**, 43–48.
12. Mevarech, M., Rice, D. & Haselkorn, R. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 6476–6480.
13. Rosenberg, M. & Court, D. (1979) *Annu. Rev. Genet.* **13**, 319–353.
14. Zurawski, G., Perrot, B., Bottomley, W. & Whitfeld, P. R. (1981) *Nucleic Acids Res.* **9**, 3251–3270.
15. Shine, J. & Dalgarno, L. (1974) *Proc. Natl. Acad. Sci. USA* **71**, 1342–1346.
16. Tinoco, I., Jr., Borer, P., Dengler, B., Levine, M. D., Uhlenbeck, O. C., Crothers, D. M. & Gralla, J. (1973) *Nature (London)* **246**, 40–41.
17. Scott, K. F., Rolfe, B. G. & Shine, J. (1981) *J. Mol. Appl. Genet.* **1**, 71–81.
18. Fleming, H. & Haselkorn, R. (1974) *Cell* **3**, 159–170.
19. Kennedy, C., Eady, R. R., Kondorosi, E. & Rekosh, D. K. (1976) *Biochem. J.* **155**, 383–389.
20. Lundell, D. J. & Howard, J. B. (1978) *J. Biol. Chem.* **253**, 3422–3426.
21. McMahon, J. E. & Tinoco, I., Jr. (1978) *Nature (London)* **271**, 275–277.
22. Segrest, J. P. & Feldmann, R. J. (1974) *J. Mol. Biol.* **87**, 853–858.
23. Lundell, D. J. & Howard, J. B. (1981) *J. Biol. Chem.* **256**, 6385–6391.
24. Adman, E. T., Sieker, L. C. & Jensen, L. H. (1973) *J. Biol. Chem.* **248**, 3987–3996.
25. Carter, C. W., Kraut, J., Freer, S. T., Xuong, N.-H., Alden, R. A. & Bartsch, R. G. (1974) *J. Biol. Chem.* **249**, 4212–4225.
26. Ikemura, T. (1981) *J. Mol. Biol.* **146**, 1–21.