Proc. Natl. Acad. Sci. USA
Vol. 77, No. 11, pp. 6476–6480, November 1980
Biochemistry

Nucleotide sequence of a cyanobacterial *nifH* gene coding for nitrogenase reductase

(nitrogen fixation/*Anabaena*/protein sequence homologies)

MOSHE MEVARECH, DOUGLAS RICE, AND ROBERT HASELKORN

Department of Biophysics and Theoretical Biology, The University of Chicago, Chicago, Illinois 60637

Communicated by Martin D. Kamen, August 4, 1980

ABSTRACT **The nucleotide sequence of *nifH*, the structural gene for nitrogenase reductase (component II or Fe protein of nitrogenase) from the cyanobacterium *Anabaena* 7120 has been determined. Also reported are 194 bases of the 5′-flanking sequence and 170 bases of the 3′-flanking sequence. The predicted amino acid sequence was compared with that determined for the complete nitrogenase reductase of *Clostridium pasteurianum* and the cysteine-containing peptides of the protein from *Azotobacter vinelandii*. Amino acid sequences around five cysteines, located in the NH_2-terminal two-thirds of the protein, are highly conserved in all three species. Codon usage in the first gene from a cyanobacterium to be sequenced shows striking asymmetries for eight amino acids.**

The biological fixation of nitrogen is catalyzed by an enzyme complex that throughout the microbial world appears to be highly conserved in terms of the general properties of its proteins (1). Purified nitrogenase (component I) and nitrogenase reductase (component II) from *Klebsiella pneumoniae* (2), *Clostridium pasteurianum* (3), *Azotobacter vinelandii* (4), and the cyanobacterium *Anabaena* (5) share the following features: nitrogenase contains two pairs of two dissimilar subunits of molecular weights 60,000 and 56,000, a number of Fe_4S_4 clusters, and a cofactor containing Fe, S, and Mo; nitrogenase reductase is a dimer containing two identical subunits of molecular weight 28,000–35,000 and one Fe_4S_4 cluster.

Conservation of structure of nitrogenase components has been demonstrated by extensive *in vitro* complementation between purified components from diverse sources, including cyanobacteria (1, 5–7). Conservation of nucleotide sequences among genes from nitrogen-fixing bacteria has been shown by using cloned DNA containing the *Klebsiella* nitrogen fixation (*nif*) genes coding for nitrogenase structural components, which hybridize with DNA from other nitrogen-fixing bacteria, including *Anabaena* (8, 9).

Genetic analysis of nitrogen fixation in *Klebsiella pneumoniae* has revealed thus far 15 linked genes arranged in seven transcriptional units (10–13). Three of these, *nifH*, *nifD*, and *nifK*, are the structural genes for nitrogenase reductase and the two subunits of nitrogenase, respectively. These three genes form one transcriptional unit, with the promoter at the start of the *nifH* gene. *nifE*, *nifN*, *nifB*, and *nifQ* are involved in the synthesis of the FeMo cofactor. Genes *nifM*, *nifV*, and *nifS* are apparently required to convert the *nifH* gene product into an active nitrogenase reductase (14), but the details of this conversion are unknown.

Genes *nifF* and *nifJ* are believed to be involved in electron transport (14), whereas *nifA* is a positive regulatory protein required for expression of all other *nif* genes (15). The *nif* genes are regulated, presumably through *nifA*, by the level of NH_4^+ in the cell.

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.

In filamentous cyanobacteria such as *Anabaena*, nitrogen fixation under aerobic conditions occurs in specialized cells called heterocysts (16). Very little is known about the regulation of *nif* genes in these cells. We have used recombinant DNA techniques to study the organization and regulation of the *nif* genes in *Anabaena* 7120. We previously reported that a recombinant plasmid containing *K. pneumoniae nifH*, *nifD*, and *nifK* genes annealed to restriction fragments of *Anabaena* 7120 DNA (9). Subsequent cloning and restriction mapping of these two fragments, together with electron-microscope analysis of heteroduplex DNA molecules and hybridization to DNA restriction fragments immobilized on nitrocellulose filters, has shown that a 10-kilobase (kb) *Anabaena Eco*RI fragment carried in the recombinant phage λgt7-An154 contains the *Anabaena nifH* gene and approximately half of the *nifD* gene. A 17-kb *Eco*RI fragment of *Anabaena* DNA carried in the phage λCharon4-An207 contains the remainder of the *nifD* gene and all of *nifK* (unpublished data). A 1.8-kb *Hin*dIII fragment of the 10-kb *Eco*RI fragment of λgt7-An154 contains the entire coding portion of the *Anabaena nifH* gene and approximately 0.75 kb of DNA upstream from the gene. We present below the nucleotide sequence of the *Anabaena* 7120 *nifH* gene and 194 bases of the 5′-flanking sequence, and we compare the predicted amino acid sequence for nitrogenase reductase of *Anabaena* 7120 with the available sequences of the corresponding proteins from *Clostridium pasteurianum* (17) and *Azotobacter vinelandii* (18).

MATERIALS AND METHODS

The identification of a 10-kb *Eco*RI fragment of *Anabaena* 7120 DNA containing sequences homologous to the *Klebsiella nifH* gene and its cloning in a phage λ vector have been described (9). The 10-kb *Eco*RI insert of λgt7-An154 was cloned into the plasmid pBR322; the resulting plasmid was designated pAn154 (unpublished data).

Four hundred micrograms of pAn154 (Fig. 1) was digested with 300 units of *Hin*dIII endonuclease, and the resulting fragments were separated by electrophoresis in a 1% agarose horizontal slab gel as described (9). The 1.8-kb fragment was recovered by electroelution and purified by passage over DEAE-cellulose (19). Total recovery was 18.5 μg of DNA.

In general, the methods employed for labeling and sequence determination were those of Maxam and Gilbert (20). The choice of restriction endonucleases used to create fragments for sequence determinations was based on restriction mapping by the method of Smith and Birnstiel (21). In some cases this method failed to resolve the pair of bands created by two closely spaced restriction sites [e.g., the *Hpa* II sites near residue 300 (Fig. 1)]. These were discovered when the sequences of the overlapping *Hin*cII and *Alu* I fragments were determined.

Abbreviations: kb, kilobase pairs; bp, base pairs.

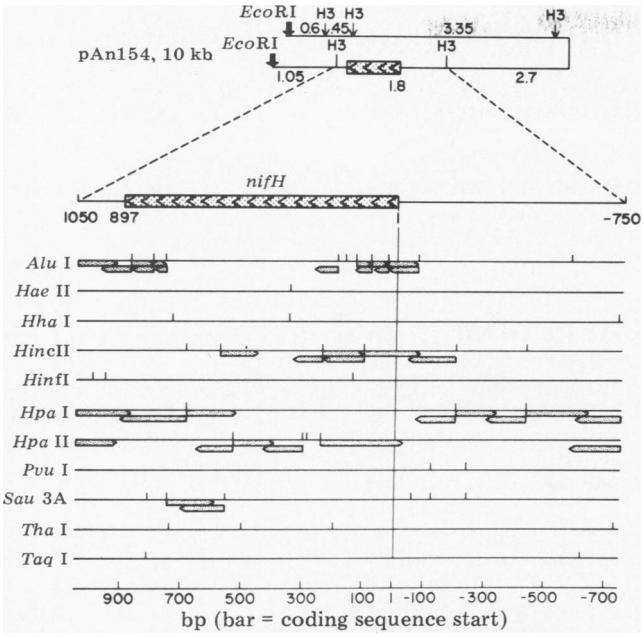

FIG. 1. Physical map of the *Anabaena* 7120 *nifH* (nitrogenase reductase) gene. The upper part of the figure shows the location of endonuclease *Hin*dIII (H3) restriction sites on the *Eco*RI 10-kb fragment cloned in λgt7An154 (9). Filled arrows on the restriction maps indicate the extent and direction of DNA-fragment sequence determinations.

After restriction of the 1.8-kb *Hin*dIII fragment, 5′-phosphates were removed with calf intestine alkaline phosphatase, and the resulting 5′-OH termini were labeled with ^{32}P using [γ-^{32}P]ATP and polynucleotide kinase. The ATP (>1000 Ci/mmol; 1 Ci = 3.7×10^{10} becquerels) was prepared as described (20). The labeled fragments were separated by electrophoresis on 5% (wt/vol) polyacrylamide gels, eluted, denatured in 30% (vol/vol) dimethyl sulfoxide, and rerun on 5% polyacrylamide gels to separate the denatured complementary strands. Nearly all of the fragments shown with sequences determined in Fig. 1 were strand-separable. In those cases where the complementary strands did not separate, a second restriction cut was used to create singly labeled fragments.

Polynucleotide kinase and restriction endonucleases *Hpa* I and *Hpa* II were gifts of Kan Agarwal, University of Chicago. Endonuclease *Hin*dIII and calf intestine phosphatase were obtained from Boehringer Mannheim, *Alu* I and *Hin*cII from Bethesda Research Laboratories, Rockville, MD, and *Sau* 3A from New England BioLabs.

RESULTS AND DISCUSSION

The location of the *Anabaena* 7120 gene for nitrogenase reductase (*nifH*) is indicated in Fig. 1. We selected the 1.8-kb *Hin*dIII restriction fragment, which contains the gene, for sequence determination because the region of homology between a cloned *K. pneumoniae* DNA probe containing *nifH*, *nifD*, and *nifK* and the *Anabaena* 10-kb fragment cloned on λgt7-An154 was located in the leftmost 2 kb of the 10-kb *Anabaena* insert (9). Additional evidence that λgt7-An154 actually contained the *Anabaena nifH* gene was provided by experiments in which UV-irradiated *Escherichia coli* were infected with λgt7-An154 and subsequently labeled with radioactive amino acids. Among the labeled products was a protein of molecular weight 33,000 that was selectively precipitated by antibody raised against *Anabaena* nitrogenase reductase (unpublished data).

A detailed restriction map of the 1.8-kb fragment and the strategy used for sequence determination are shown in Fig. 1. Approximately 70% of the sequence was determined on both strands. Every restriction site within the gene was overlapped, in some cases several times. With the exception of two small gaps, starting at 194 base pairs (bp) and at approximately 435 bp upstream from the *nifH* gene, the entire sequence of the 1.8-kb *Hin*dIII fragment was determined. Only the 194 bases before and 170 bases after the *nifH* gene are reported here.

The complete sequence of the *nifH* gene complementary to the coding strand (i.e., identical to the mRNA) is given in Fig. 2 with the flanking sequences that are positioned unambiguously. The presumptive coding region begins with ATG and ends with TAG 900 bp away. Within the gene, there is only one open reading frame longer than 10 amino acids.

Beyond the *nifH* gene terminator at residue 900 there are two ATG initiation codons at residues 920 and 1016. Both are in the same reading frame, which is not terminated before the end of the *Hin*dIII fragment. Either of these is a potential initiator of the *nifD* gene, which codes for the smaller subunit of nitrogenase (10–12).

The 750-bp sequence preceding the *nifH* structural gene was searched for regions that might be analogous to consensus sequences of bacterial control regions: the ribosome-binding sequence, transcription-initiation regions, and transcription-termination regions. A good "Shine–Dalgarno" ribosome-binding sequence, A-G-G-A, is present at residues −14 to −11. Of course, the significance of this sequence is unknown because the sequence of nucleotides at the 3′ end of *Anabaena* 16S ribosomal RNA is unknown. There is no sequence that corresponds closely to the consensus promoter T-A-T-A-A-T found about 10 bp upstream from transcription starts in *E. coli* (22). The closest fit is the sequence T-G-A-G-A-T at −57 to −52; the sequence at −81 to −77, C-T-C-A-C-A, gives a reasonable fit to the bacterial sequence centered at approximately −30—namely, T-T-G-A-C-A (22). The absence of a good "Pribnow–Schaller" box is consistent with our own observation that the *Anabaena nifH* gene is not expressed at high levels from the cloned 10-kb fragment in λgt7-An154-infected *E. coli*, unless the fragment is oriented correctly for readthrough of transcription from the λ leftward promoter and λ immunity is lifted (unpublished data).

The sequence beginning C-C-A . . . at residue −157 and ending . . . T-G-G at residue −133, followed by five T's, can be folded to form a seven-base-paired stem-and-loop configuration that looks like a bacterial transcription terminator (22). However, we do not yet know whether this sequence functions as a transcription terminator in *E. coli* cells that have not been irradiated with UV light.

Further upstream, in the 550-bp sequence not shown, there are two open reading frames in the same direction as *nifH*. One starts at residue −429 and runs for 68 amino acids to −225. Another starts at an unknown point to the right of the *Hin*dIII site and runs for at least 92 amino acids to residue −473. We do not yet know if either of these regions is actually transcribed and translated in *Anabaena*. None of the *Klebsiella nif* gene products described thus far has a M_r as low as 7500, the expected M_r of the product closest to *nifH*.

The amino acid sequence of *Anabaena* 7120 nitrogenase reductase, deduced from the nucleotide sequence, is shown in Fig. 3. The molecular weight predicted by the sequence, 33,000, is in agreement with that observed for a major oxygen-sensitive *Anabaena* protein synthesized in heterocysts (23) or in vegetative cells under inducing anaerobic conditions (unpublished data). Nitrogenase reductase has been partially purified from an *Anabaena* strain by Tsai and Mortenson (5). Insufficient

```
          -180                -160                -140                -120                -100                 -80
TAACACCCAAAAGAACTTTCACAACTACATAACGAACCCATCATGAACACTAATTCTACTGGTTTTTCTGTGGAGCGATCGCCCCCTCTTCGGCGACTGTTCTACATAACCCCTCACAG

          -60                 -40                 -20                  1                   20
CCATAGCTCAAACAGGCGTGAGATCCAAACACAAAGACCGACCAACTAACCAACCAATTGCAGGAAAAGAGAACA ATG ACT GAC GAA AAC ATT AGA CAG ATA GCT TTC

      40                  60                  80                  100                 120
TAC GGT AAA GGC GGT ATC GGT AAA TCT ACC ACC TCC CAA AAC ACC CTT GCA GCT ATG GCA GAA ATG GGT CAA CGC ATC ATG ATT GTA GGT

                 140                 160                 180                 200
TGC GAC CCT AAA GCT GAC TCC ACC CGT CTG ATG CTT CAC TCC AAA GCT CAA ACC ACC GTA CTA CAC TTA GCT GCT GAA CGC GGT GCA GTA

      220                 240                 260                 280                 300
GAA GAC TTA GAA CTC CAC GAA GTA ATG TTG ACC GGT TTC CGT GGC GTT AAG TGC GTA GAA TCT GGT GGT CCA GAA CCC GGT GTA GGT TGC

                 320                 340                 360                 380
GCC GGT CGT GGT ATC ATC ACC GCC ATT AAC TTC TTA GAA GAA AAC GGC GCT TAC CAA GAC CTA GAC TTC GTA TCC TAC GAC GTA TTG GGT

      400                 420                 440                 460                 480
GAC GTT GTA TGT GGT GGT TTC GCT ATG CCT ATC CGT GAA GGT AAA GCA CAA GAA ATC TAC ATC GTT ACC TCT GGT GAA ATG ATG GCG ATG

                 500                 520                 540                 560
TAT GCT GCT AAC AAC ATC GCT CGC GGT ATT TTG AAA TAT GCT CAC TCC GGT GGT GTA CGT TTA GGT GGT TTG ATC TGT AAC AGC CGT AAG

      580                 600                 620                 640                 660
GTT GAC CGT GAA GAC GAG TTA ATC ATG AAC TTG GCT GAA CGT TTG AAC ACC CAA ATG ATT CAC TTC GTA CCT CGT GAC AAC ATC GTT CAA

                 680                 700                 720                 740
CAC GCA GAA TTG CGC CGT ATG ACC GTT AAC GAG TAC GCA CCA GAC AGC AAC CAA GGT CAA GAG TAC CGC GCA TTA GCT AAG AAG ATC AAC

      760                 780                 800                 820                 840
AAC GAC AAG CTC ACC ATT CCT ACA CCA ATG GAA ATG GAT GAA CTA GAA GCT CTG AAG ATC GAA TAC GGT CTA TTA GAC GAC GAC ACC AAG

                 860                 880                 900                 920                 940
CAC TCT GAA ATC ATC GGT AAG CCC GCA GAA GCT ACC AAT AGG TCA TGC CGT AAT TAG GAGACACGGAGACAGGAGATGAGGAGCAATTCCTCTTCCCACT

           960                 980                 1000                1020                1040                1060
CTCCCTTCCCGACTCCTCACTCTCCCAAATATACTTCTATTCCCCCATTCGTAAGAGTCACTGAGGCAGATATGACACCTCCTGAAAACAAGAATCTTGTAGATGAAAATAAGGAACTT

ATTCAAGAAGTTCTG
```

FIG. 2. Nucleotide sequence of the *Anabaena* 7120 *nifH* gene and flanking regions. The sequence shown is that of the strand identical to the *nifH* mRNA. Residues within the coding portion of the gene are shown as triplets for comparison with the amino acid sequence in Fig. 3. Restriction sites can be located by reference to Fig. 1.

material was available to determine the NH_2-terminal amino acid sequence or the amino acid composition, but the molecular weight was determined by acrylamide gel electrophoresis to be 33,000. Fig. 3 also shows the complete sequence of nitrogenase reductase from *C. pasteurianum* determined by Tanaka *et al.* (17) and the sequences of NH_2-terminal, COOH-terminal, and cysteine-containing peptides of nitrogenase reductase from *A. vinelandii* determined by Hausinger and Howard (18). The three sequences were aligned on the basis of extensive homology, indicated by boxed areas in Fig. 3. At several points, it was necessary to delete one or two amino acid residues in one sequence to maximize sequence homology.

The *Clostridium* protein begins with methionine. The *Anabaena* gene sequence indicates isoleucine at the corresponding position. The only methionine in phase with the rest of the *Anabaena* gene, upstream from the first region of homology, is the one indicated. When the cloned *Anabaena* gene was transcribed and translated in UV-irradiated *E. coli*, the M_r 33,000 product had methionine at position 1 and isoleucine at positions 6, 9, and 17 (unpublished data), indicating that the sequence shown corresponds to the gene product, at least in UV-irradiated *E. coli*. We have not yet determined the NH_2-terminal sequence of mature nitrogenase reductase made in *Anabaena* itself. The mature *Azotobacter* protein begins with alanine, so at least one residue must be removed from the NH_2 terminus following translation in *Azotobacter*.

The COOH termini of the three proteins are also different. If the termination codon in the *Clostridium* gene follows immediately after the COOH-terminal leucine, its position corresponds to glutamate in *Azotobacter* and aspartate in *Anabaena*. These three codons can be related by single base substitution (18). However, the position of the presumptive termination codon in *Azotobacter* corresponds to asparagine in *Anabaena*. The Asn codons require transversions at the first and third positions to become termination codons.

Conservation of amino acid sequence around the cysteine residues is striking. The *Clostridium* protein has six cysteines, *Azotobacter* seven, and *Anabaena* six. Five of these, at *Anabaena* positions 43, 89, 101, 135, and 187, are fully conserved. Moreover, these cysteines occur in highly conserved regions clustered in the central third of the protein. However, it is still impossible to determine which cysteines are involved in binding the Fe_4S_4 cluster or MgATP. Since each dimer binds one Fe_4S_4 cluster and has two sites for MgATP, only three of the five conserved cysteines appear to be needed directly for ligand binding (1).

It is possible to estimate the total number of nucleotide changes relating the nonconserved residues among the three proteins for the regions where the sequence data are complete. Between *Anabaena* and *Clostridium*, full conservation of amino acid sequence in those regions (around the cysteines)

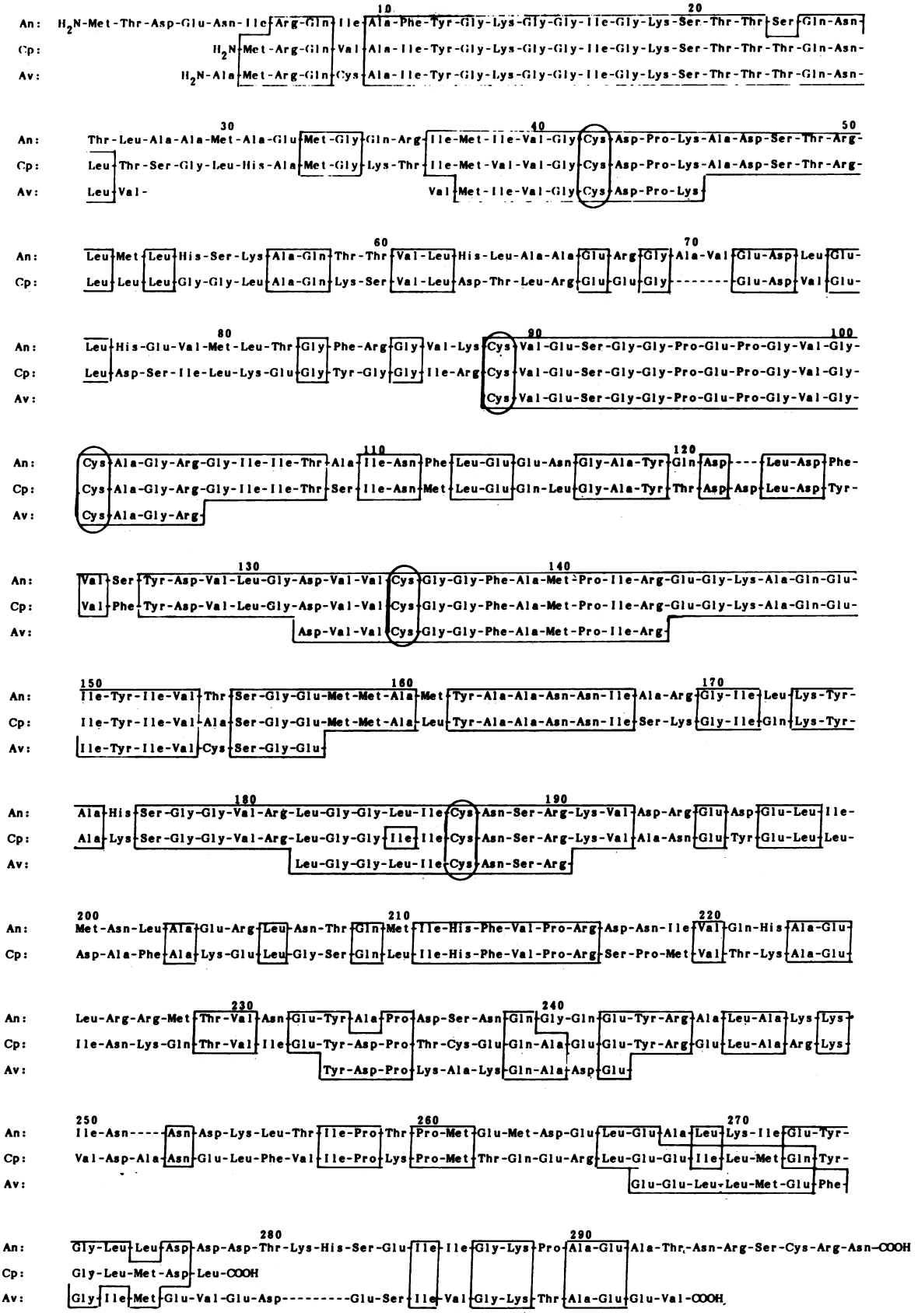

FIG. 3. Comparison of amino acid sequences of nitrogenase reductase from *Anabaena* 7120 (An), *C. pasteurianum* (Cp) (17), and *A. vinelandii* (Av) (18). Numbering refers to *Anabaena* amino acid residues. Conserved residues are enclosed in boxes; conserved cysteines are encircled. Dotted lines indicate amino acids deleted in one sequence relative to another.

Table 1. Codon utilization in the *Anabaena nifH* gene

UUU	Phe	0	UCU	Ser	4	UAU	Tyr	2	UGU	Cys	2
UUC		6	UCC		5	UAC		7	UGC		4
UUA	Leu	8	UCA		1	UAA	ochre	0	UGA	opal	0
UUG		6	UCG		0	UAG	amber	1	UGG	Trp	0
CUU	Leu	1	CCU	Pro	4	CAU	His	0	CGU	Arg	12
CUC		2	CCC		2	CAC		7	CGC		5
CUA		4	CCA		3	CAA	Gln	9	CGA		0
CUG		2	CCG		0	CAG		1	CGG		0
AUU	Ile	6	ACU	Thr	1	AAU	Asn	2	AGU	Ser	0
AUC		15	ACC		14	AAC		14	AGC		2
AUA		1	ACA		1	AAA	Lys	6	AGA	Arg	1
AUG	Met	15	ACG		0	AAG		8	AGG		1
GUU	Val	7	GCU	Ala	16	GAU	Asp	1	GGU	Gly	26
GUC		0	GCC		2	GAC		16	GGC		3
GUA		10	GCA		8	GAA	Glu	22	GGA		0
GUG		0	GCG		1	GAG		3	GGG		0

would require 6 transitions and 14 transversions; between *Clostridium* and *Azotobacter*, it would require 3 transitions and 15 transversions; and between *Azotobacter* and *Anabaena*, it would require 4 transitions and 15 transversions. Thus, each of the three proteins appears to be equidistant from the other two in terms of the number of separate mutational events needed to derive one from another.

The distribution of codons utilized in the *Anabaena nifH* gene is shown in Table 1. There are rather striking asymmetries in codon usage for valine, threonine, histidine, asparagine, aspartate, glutamate, arginine, and glycine. If codon usage is strongly correlated for different genes within an organism, as has been suggested recently (24), we may expect to find a distribution of tRNA in *Anabaena* that is quite different from that of *E. coli*.

We are indebted to Prof. Kan Agarwal for gifts of enzymes and advice on sequence determinations and to Pamela Keim for sequenator analysis of radioactive proteins. This work was supported by Grant GM 21823 from the U.S. Public Health Service and Grant 5901-0410 from the U.S. Department of Agriculture/Science and Education Administration through the Competitive Research Grants Office. D.R. was supported by U.S. Public Health Service Training Grants GM 780 and GM 07183.

1. Mortenson, L. E. & Thorneley, R. N. F. (1979) *Annu. Rev. Biochem.* **48**, 387–418.
2. Eady, R. R., Smith, B. E., Cook, K. A. & Postgate, J. R. (1972) *Biochem. J.* **128**, 655–675.
3. Huang, J. C., Zumft, W. & Mortenson, L. E. (1973) *J. Bacteriol.* **113**, 884–890.
4. Swisher, R. H., Landt, M. & Reithel, F. (1975) *Biochem. Biophys. Res. Commun.* **66**, 1476–1482.
5. Tsai, L.-B. & Mortenson, L. E. (1978) *Biochem. Biophys. Res. Commun.* **81**, 280–287.
6. Emerich, D. W. & Burris, R. H. (1978) *J. Bacteriol.* **134**, 936–943.
7. Nagatani, H. & Haselkorn, R. (1978) *J. Bacteriol.* **134**, 597–605.
8. Ruvkun, G. & Ausubel, F. M. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 191–195.
9. Mazur, B. J., Rice, D. & Haselkorn, R. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 186–190.
10. MacNeil, T., MacNeil, D., Roberts, G. P., Supiano, M. A. & Brill, W. J. (1978) *J. Bacteriol.* **136**, 253–266.
11. Merrick, M., Filser, M., Kennedy, C. & Dixon, R. (1978) *Mol. Gen. Genet.* **165**, 181–189.
12. Elmerich, C., Houmard, J., Sibold, L., Manheimer, I. & Charpin, N. (1978) *Mol. Gen. Genet.* **165**, 181–189.
13. Merrick, M., Filser, M., Dixon, R., Elmerich, C., Sibold, L. & Houmard, J. (1980) *J. Gen. Microbiol.* **136**, 253–266.
14. Roberts, G. P., MacNeil, T., MacNeil, D. & Brill, W. J. (1978) *J. Bacteriol.* **136**, 267–279.
15. Dixon, R., Kennedy, C., Kondorosi, A., Krishnapillai, V. & Merrick, M. (1977) *Mol. Gen. Genet.* **157**, 189–198.
16. Haselkorn, R. (1978) *Annu. Rev. Plant Physiol.* **29**, 319–344.
17. Tanaka, M., Hainu, M., Yasunobu, K. T. & Mortenson, L. E. (1977) *J. Biol. Chem.* **252**, 7093–7100.
18. Hausinger, R. P. & Howard, J. B. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 3826–3830.
19. Clarkson, S. G., Kurer, V. & Smith, H. O. (1978) *Cell* **14**, 713–724.
20. Maxam, A. M. & Gilbert, W. (1980) *Methods Enzymol.* **65**, 499–560.
21. Smith, H. O. & Birnstiel, M. L. (1976) *Nucleic Acids Res.* **3**, 2387–2398.
22. Rosenberg, M. & Court, D. (1979) *Annu. Rev. Genet.* **13**, 319–353.
23. Fleming, H. & Haselkorn, R. (1973) *Proc. Natl. Acad. Sci. USA* **70**, 2727–2731.
24. Grantham, R., Gautier, C., Gouy, M., Mercier, R. & Pavé, A. (1980) *Nucleic Acids Res.* **8**, 49–62.