CHAPTER 14

THE PHYLOGENY AND EVOLUTION OF NITROGENASES

J. P. W. YOUNG
Dept. of Biology, University of York, York, Y0105YW, UK

1. INTRODUCTION

The ability to fix N_2 has a very wide taxonomic distribution, but a patchy one. It is found in most major groups of bacteria and in the methanogenic archaea, but only in a minority of species. In a comprehensive survey published over a decade ago, reports of nitrogen fixation were found for almost a hundred genera (Young, 1992). The number has increased since then through the discovery of nitrogen fixation in new organisms, but also because advances in microbial taxonomy have led to the recognition of more genera. Strikingly though, nitrogen fixation has never been reported in eukaryotes.

The nitrogenase enzymes of all these nitrogen-fixing organisms are so similar (with one exception) that they are clearly derived from a common ancestor. We can, therefore, rule out multiple independent inventions of the process as a possible explanation for its patchy distribution. The remaining possibilities are either horizontal gene transfer or independent loss in many lineages. We can hope to distinguish between these by examining the phylogeny of the nitrogenases themselves in relation to that of the organisms that carry them. If the trees match, this would be persuasive evidence that horizontal gene transfer had not played a significant role. In fact, the trees are not the same, but they do reveal another complication, namely that many organisms have more than one nitrogenase system and these can be highly divergent. Once the possibility of gene duplication is included, discordance between phylogenies could be explained by the inclusion of genes that were not orthologs (derived from the same copy in the common ancestor) but paralogs (derived from different members of a gene family).

These issues have been discussed repeatedly in the literature. Postgate (1974; 1982) challenged the earlier view that nitrogen fixation was present in the common ancestor of life and had frequently been lost since. Instead, he suggested that

R. Palacios and W. E. Newton (eds.), Genomes and Genomics of Nitrogen-fixing Organisms, 221-241.

diazotrophy arose relatively recently, less than 1,500 million years ago, and was transferred laterally to achieve its present distribution. Molecular sequences, as they became available, were interpreted by some to support a phylogeny for nitrogenase that matched that of the ribosomal genes and so were consistent with an ancestral origin (Hennecke *et al*., 1985, Postgate and Eady 1988), whereas others saw evidence for lateral transfer (Normand and Bousquet 1989). Once the existence of deep paralogous branches was taken into account, the data could be seen as consistent with multiple losses, with multiple transfers, or with some combination of both processes (Young, 1992). As DNA sequencing proceeds apace and the complete genomes of many diazotrophs have now been determined, new groups of nitrogenase homologs are discovered and the web of relationships grows more tangled. Although horizontal gene transfer has probably played a role, there are few clear-cut examples and our understanding of the evolutionary history of nitrogenases is far from complete.

Within the nitrogenases, there are a number of distinct clades, but the overall subunit organisation is almost invariant, and there are enough conserved features to guide an alignment of the sequences of each of the polypeptide subunits. The only known exception is a pathway in *Streptomyces thermoautotrophicus* in which N_2 reduction is coupled to the oxidation of carbon monoxide by enzymes that are unrelated to other nitrogenases (Ribbe *et al*., 1997). It appears to be a true "independent invention" but, as it has not so far been reported in other organisms, it will not be considered further here.

2. THE GENETIC ORGANISATION OF NITROGENASE GENES

The widespread molybdenum-containing forms of the nitrogenase enzyme are composed of two alpha and two beta subunits (together forming component I, dinitrogenase, MoFe protein), encoded by the *nifD* and *nifK* genes, respectively, in association with a dimer encoded by *nifH* (component II, dinitrogenase reductase, Fe protein). It is the relationships of these catalytic proteins that I will discuss in this chapter, together with the products of the *nifE* and *nifN* genes, which are homologs of *nifD* and *nifK*. In addition, I will include the nitrogenases that either have vanadium rather than molybdenum or that have only iron and which have similar subunit components plus additional delta subunits encoded by either *vnfG* or *anfG*. There are many other gene products necessary to support the nitrogen-fixation process, but the work needed to piece together the history of this whole network lies in the future.

Surprisingly, there is a "standard arrangement" of the nitrogenase genes that can be discerned despite the large evolutionary distances that separate them. An operon arranged as either *nifHDKEN* or *vnfHDGKEN* or *anfHDGKEN* could be seen as the archetypal form. However, this arrangement is clearly not critical for function because many variants are found. For example, either *nifH* or *nifEN* may be separated from *nifDK* and individual genes may be duplicated (especially *nifH* or *nifK*). Some examples of variant arrangements are illustrated by Dean and Jacobson (1992) and Kaminski *et al*. (1998) and others are listed in Table 1.

Table 1. The arrangement of nitrogenase gene clusters in diazotroph genomes

Organism	B-type	C-type	A-type Vnf	A-type Anf	D-type
PROTEOBACTERIA					
Bradyrhizobium japonicum	**H, DKEN**[1] N768409, 383-6[2]				
Mesorhizobium loti	**HDKEN** N106489-93				
Sinorhizobium meliloti	**HDKE, N** N435696-8, 724				
Rhodopseudomonas palustris	**HDKEN** Z09392-88		**NE--H-DGK** Z10950-58	**HDGK** Z11017-14	**HEN** Z12151-3 **HEN** Z12171-3 **H----EN** Z11901-895
Rhodospirillum rubrum	**HDK** Z14415-3 **EN** Z15686-7			**HDGK** Z15420-18	**HEN** Z14613-5 **EN** Z14636-7
Rhodobacter sphaeroides	**HDKEN** Z07624-28				
Burkholderia fungorum	**HDK--EN** Z34963-57				
Azotobacter vinelandii	**HDK----EN** Z90773-65		**EN------H-DGK** Z89155-63	**HDGK** Z92169-72	
Magnetococcus sp. MC-1	**HDK-----EN** Z44347-56				
CYANOBACTERIA					
Nostoc sp. PCC7120	**H, DKEN** N485497, 84-80				
Nostoc punctiforme	**HD, KEN** Z112319-41 **HEN** Z111243-45				
Trichodesmium erythraeum	**HDK[EN]** Z73548-45				
CLOSTRIDIA					
Desulfitobacterium hafniense	**HDKEN** Z99588-92				**HNE** Z98133-5
Clostridium acetobutylicum		**H—DKEN** N346894-900			
CHLOROBI					
Chlorobium tepidum		**H—DKEN** N662417-23			
ARCHAEA					
Methanosarcina acetivorans		**H—DKEN** N618766-72	**H--DGKEN** N616152-9	**DGK--H** N616149-4	**ENH** N616561-3 **NH** N618503-2 **[HE]N** N616955-6
Methanosarcina barkeri		**H—DKEN** Z78739-33	**H--DGK-EN** Z78063-71	**DGK** Z77947-8	**NH** Z78936-7 **[HE]N?** Z76604-5
Methanosarcina mazei		**H—DKEN** N632743-49			**NH** N632538-9
Methanothermobacter thermautotrophicus		**H—DKEN** N276673-79			

[1] *– intervening gene*, *<30 intervening genes,* **[ ]** *fused genes, _ reverse orientation.*
[2] *database accession numbers, NP_ abbreviated to N, ZP_000 to Z.*

As gene sequences became available, it was realised that *nifE* is a homolog of *nifD* and *nifN* is a homolog of *nifK* (Brigle *et al*., 1987; Aguilar *et al*., 1987). Hence, it seems certain that *nifDK* and *nifEN* are the products of ancient gene duplication. Furthermore, at a deeper level, *nifD* and *nifK* are also homologous (Holland *et al*., 1987), so we can explain the *nifDKEN* genes, and their order, as a result of two successive duplications. The arguments for this have been discussed in detail by Fani *et al*. (2000).

3. NITROGENASE GENES FROM GENOME SEQUENCING PROJECTS

There are now hundreds of published nitrogenase gene sequences from many different species, especially for *nifH*, because this gene has been widely used for diversity studies. In most cases, these were obtained by cloning or amplifying sequences from organisms in which nitrogen fixation had been demonstrated experimentally. One advantage of the current availability of complete genome sequences is that one can make a complete list of the relevant genes in an organism, including those that might be cryptic because they are either seldom expressed or defective. This approach might reveal unsuspected abilities in some organisms, but it also allows us, for the first time, to state categorically that some organisms lack nitrogenase genes and are, therefore, incapable of nitrogen fixation under any circumstances (unless by a radically novel process that we know nothing about).

To provide a consistent set of sequences, the phylogenies presented here are based primarily on the genomes that can be searched in the database available at NCBI (http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi). A comprehensive search of the complete genomes was made using PSI-Blast (Altschul *et al*., 1997) to find all homologs of NifH, NifDKEN, and NifG. The predicted peptide sequences were aligned using ClustalX (Thompson *et al*., 1997) and phylogenies constructed using Mega 2.1 (Kumar *et al*., 2001). All trees shown were based on the neighbor-joining method with rate variation corrected by a gamma distribution with a parameter value of 2 (approximating Dayhoff distances).

At the time of writing, the complement of proteins had been predicted for 143 bacterial and 18 archaeal genomes. Of these, 16 bacterial and 4 archaeal genomes had identifiable sets of *nif* genes (including at least *H*, *D*, and *K*). However, *Magnetospirillum magnetotacticum* was not analysed further because the assembly was too fragmentary to give full-length genes and *Rhodospirillum rubrum* could only be included after careful scrutiny. These genomes are listed in Table 1 and form the main basis for this review. Nine are fully annotated and published (Deppenmeier *et al*., 2002; Eisen *et al*., 2002; Galagan *et al*., 2002; Galibert *et al*., 2001; Kaneko *et al*., 2000; 2001; 2002; Nölling *et al*., 2001; Smith *et al*., 1997). Incomplete DNA sequences of an additional 87 genomes were available, but without gene predictions. Six of these had close matches to *nifH*, *D*, and *K* genes; they were from *Methylococcus capsulatus*, *Acidithiobacillus ferrooxidans*, *Rhizobium leguminosarum*, *Erwinia chrysanthemi*, *E. carotovora*, and *Geobacter sulfurreducens*, all of which are proteobacteria.

Although this survey does not cover a completely representative sample of prokaryotes, it does emphasise that the potential for nitrogen fixation is confined to

a minority of species. The only sequenced organism that has *nif* genes but was not previously known to be capable of fixing N_2 (or closely related to a fixer) is *Desulfitobacterium hafniense*. A preliminary report, prompted by the discovery of *nif* genes in the genome, has now demonstrated that this organism can use N_2 gas as sole nitrogen source (Tiedje and Davis, 2002). There is no evidence for cryptic nitrogenase systems in organisms that are not known to fix N_2. It is clear that there is a patchy distribution of the genes, not just of the phenotype.

4. ORGANISATION OF THE NITROGENASE GENES

Genes encoding the different subunits of nitrogenase are clustered together in the genome, usually in a characteristic order (see Table 1). There are three types of nitrogenase, A, B, and C, that will be described in detail in the next section. In the case of the common B type of enzyme, the *nifHDKEN* genes are typically adjacent, without intervening genes, as in *Mesorhizobium loti*, for example. However, this arrangement is clearly not essential for function because this group can be fragmented in many different ways. In fact, of the thirteen complete examples in Table 1, only five are unbroken, and all possible breaks between the genes are found in the remaining examples. The intervening genes are sometimes directly relevant to nitrogenase, *e.g*., *nifT* and *nifY* following *nifK* in *Azotobacter vinelandii*, but in other cases, the gap may be longer and include seemingly unrelated genes.

In contrast, the C-type systems have a very consistent organisation, even though they are found in both bacteria and archaea. Between the *nifH* and *nifD* genes lie two small genes (either *nifI* or *glnB*) that encode P_{II} proteins, which are implicated in regulation of nitrogenase by ammonia in *Methanococcus maripaludis* (Kessler and Leigh 1999). In all six examples, the order is *nifHIIDKEN*.

In the Vnf systems, there is much more diversity of gene arrangement, although there is always a *vnfG* gene between *vnfD* and *vnfK*. Similarly, the Anf systems have an *anfG* gene, although this was not always detected by the automated gene-finding software that is used to deduce the protein complement from the genome sequence, for example, in *Rhodospirillum rubrum*. *Methanosarcina barkeri* has *anfDGK* genes, but there is no *nifH* either preceding or immediately following them. However, the genome is incompletely assembled and these genes are at the end of a contig, so it is possible that a *nifH* will eventually be located a few genes beyond *anfK* as seen in *Methanosarcina acetivorans*. The Nif and Vnf systems of *M. barkeri* have been described and studied in some detail (Chien *et al*., 2000), but there is as yet no experimental evidence for a functional Anf system.

5. EVOLUTIONARY RELATIONSHIPS OF THE NITROGENASE GENES

5.1. Phylogeny of NifH

Figure 1 shows a phylogeny based on NifH proteins in the completely sequenced nitrogen-fixers and their recognisable homologs in the same species. The first thing to note is that the true NifH sequences form a clearly defined clade, but there are

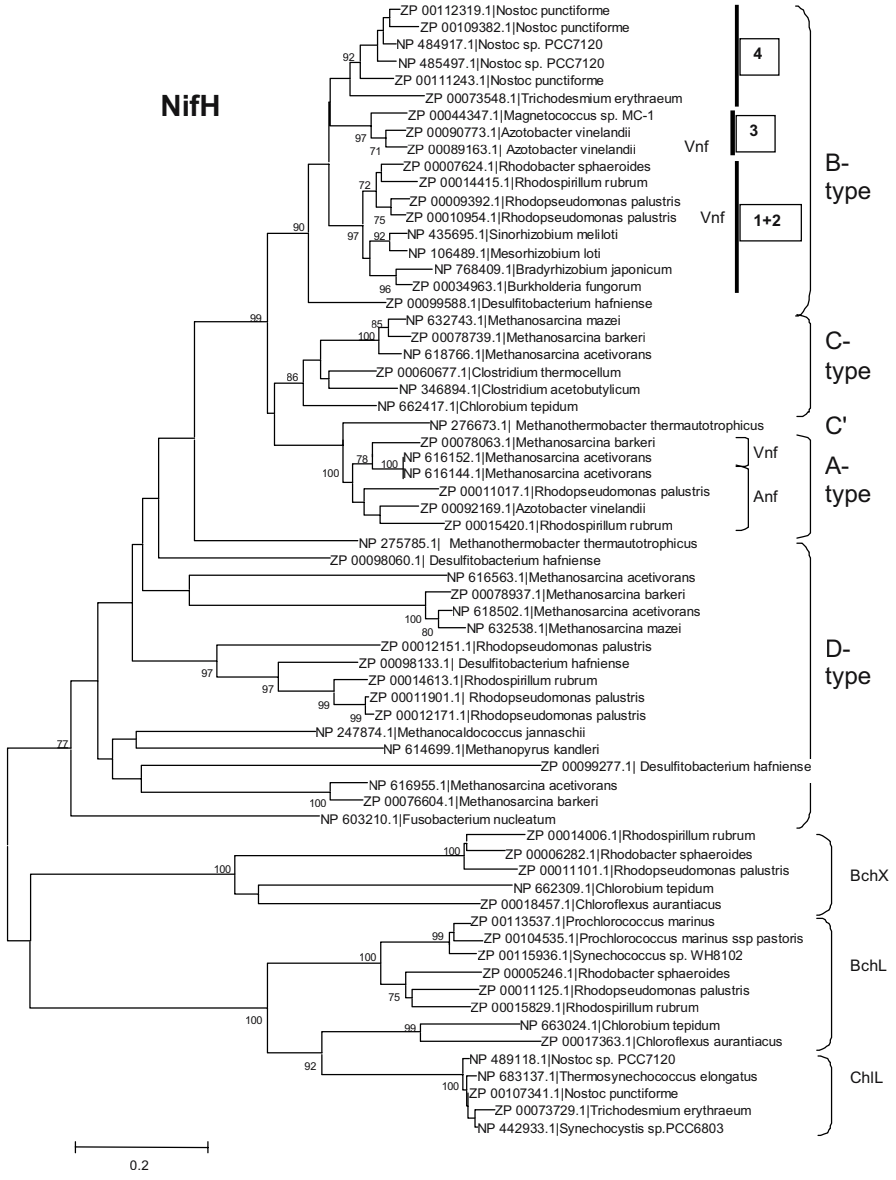

Figure 1. Phylogeny of NifH and homologous proteins from diazotroph genome sequences.

also a large number of more distant relatives. Some of these are well-known proteins involved in the synthesis of photosynthetic pigments, namely protochlorophyllide reductase (either BchL or ChlL) and chlorin reductase (BchX). The similarity between these proteins and NifH was analysed and discussed by Burke *et al*. (1993), who argued that nitrogen fixation probably originated before photosynthesis, so the photosynthesis enzymes would have been derived from NifH rather than the other way round. Nevertheless, modern NifH sequences are more tightly conserved, suggesting that this protein is subject to more rigorous structural constraints.

The true NifH proteins can be divided into three types (Young, 1990; 2000). Type B ("bacterial") is the best represented and includes enzymes from the proteobacteria, cyanobacteria, and firmicutes (the subclusters 1-4 within B are discussed later). Type C ("clostridial") is found in the firmicute bacterium *Clostridium*, the green sulfur bacterium *Chlorobium*, and also in the archaeon *Methanosarcina*.

Type A is associated with the "alternative" nitrogenases, which do not contain molybdenum, and is found in both archaea and proteobacteria. There is a group that is phylogenetically related to A-type, but is associated with Mo-containing enzymes in some archaea. It is represented among the sequenced genomes only by *Methanothermobacter thermautotrophicus*, but is also found in some other methanogens, including *Methanococcus maripaludis*, in which the Mo-dependence of the enzyme has been demonstrated (Kessler *et al*., 1997). The organisation of the genes in these systems is similar to that in the C-type found in other methanogens (*Methanosarcina*) and clostridia, even though they are phylogenetically remote, so this class is identified here as C'-type.

A more distant and diverse group was designated type D (Young 1990), but it is now clear that these proteins are not part of nitrogenase enzymes as their genes are not associated with *nifD* or *nifK*, although some are adjacent to rather distant homologs of the *nifDKEN* family.

5.2. Phylogeny of the NifDKEN family

The overall relationships among the NifD, K, E, and N proteins are shown in Figure 2. As with NifH, there are homologs involved in photosynthetic-pigment synthesis. These are not related to any particular Nif protein, but all the *bch* and *chl* genes cluster together. This suggests that the photosynthesis and nitrogen-fixation systems separated very early on, and that the two rounds of duplication that gave rise to the NifK, D, E, and N proteins was mirrored by a similar, but separate, succession of duplications to create the photosynthetic-pigment synthesis system. The NifD, NifK, NifE, and NifN proteins each form clearly distinct groups, as expected (Fig. 2). However, there are some exceptions, such as the putative *nifN* in the C-type systems, which is discussed in detail later. The archaeal VnfE and VnfN proteins are so diverged that they do not cluster with anything else on the tree.

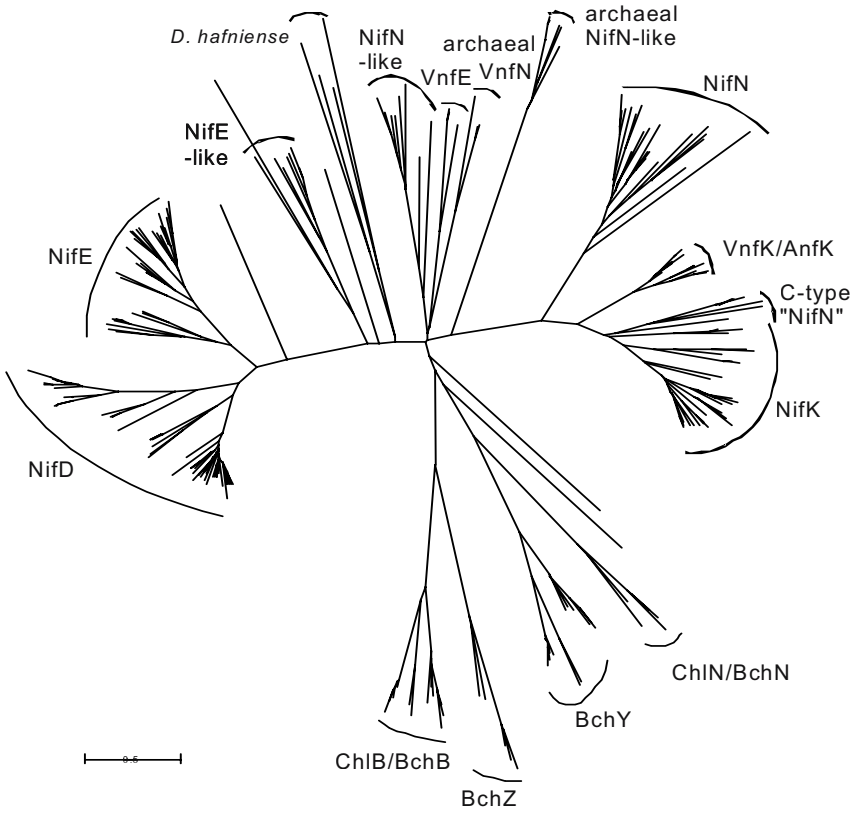

Figure 2. Overall phylogeny of NifD, K, E, N and homologous proteins from diazotroph genome sequences.

The diazotroph genomes contain a number of other homologs of the NifDKEN family. In both the photosynthetic proteobacteria and in *Desulfitobacterium*, there are adjacent pairs of genes that encode vaguely NifE-like and NifN-like proteins, mostly accompanied by a D-type NifH homolog. Their function is at present completely unknown and may have nothing to do with nitrogen fixation. Related genes are found in *Clostridium thermocellum*, which does not have nitrogenase. The archaeal genomes also have genes encoding NifN-like proteins adjacent to D-type NifH homologs, which in two cases are fused to NifE homologs. Again, these may not be relevant to nitrogen fixation. The non-fixers, *Methanopyrus kandleri* and *Methanococcus jannaschii,* also have them.

These very distant homologs are omitted from the phylogenies of the individual NifD, K, E and N proteins (Figures 3-6). Each of these phylogenies can be partitioned into types corresponding to those identified for NifH.

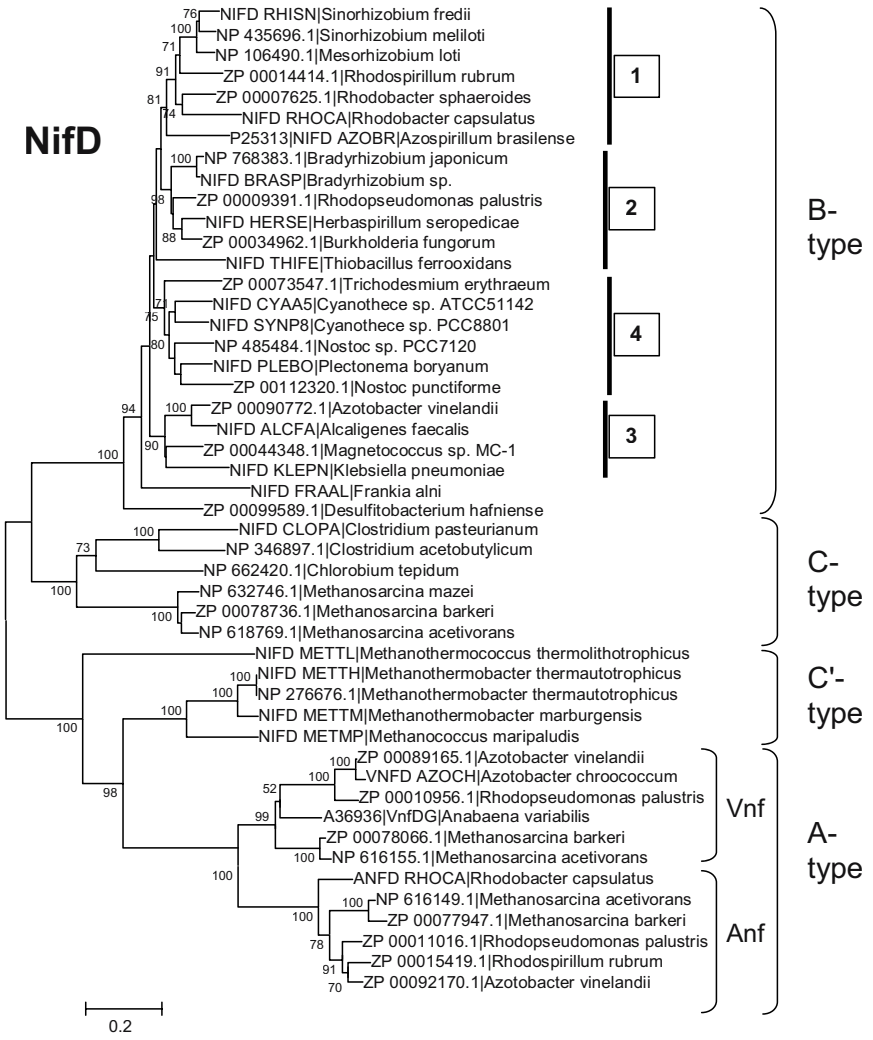

Figure 3. Phylogeny of NifD and homologous proteins from diazotroph genome sequences.

In most cases, a consistent phylogenetic clustering is seen for all the linked genes encoding a system, but there are exceptions. The VnfD and VnfK proteins of *Rhodopseudomonas* and *Azotobacter* are A-type, related to their archaeal counterparts, but their VnfH, VnfE, and VnfN are B-type. On the other hand, their Anf systems are consistently A-type. The C-type systems have a gene following *nifE*, where one would expect to find *nifN*, which encodes a protein that is quite

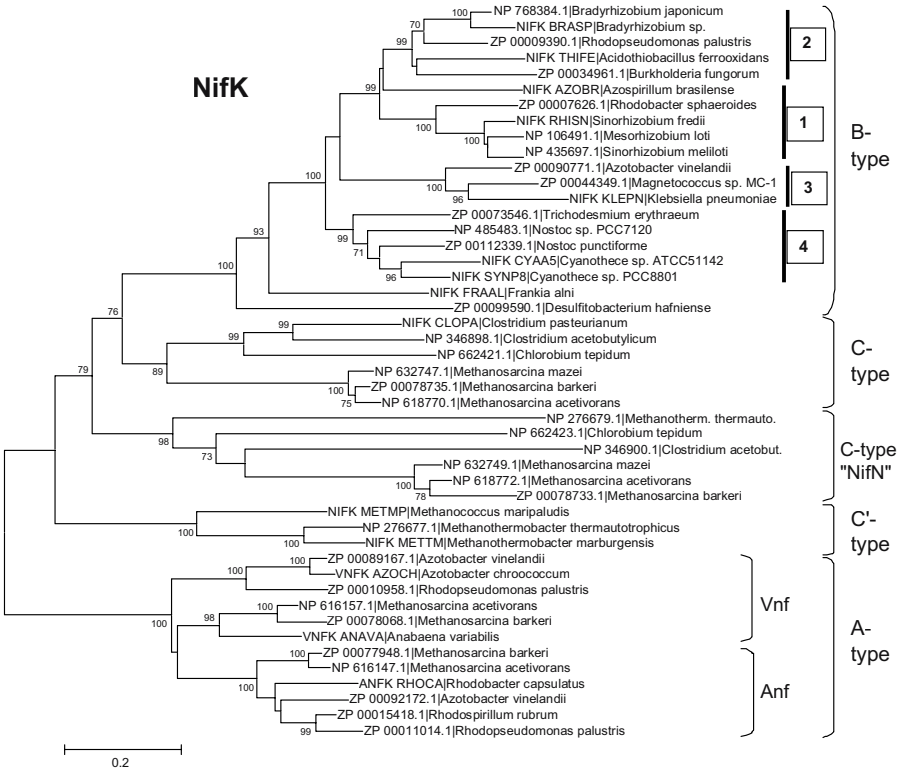

Figure 4. Phylogeny of NifK and homologous proteins from diazotroph genome sequences.

unrelated to other NifN proteins but is much closer to NifK. In fact, it is closer than both the VnfK and AnfK proteins. These putative NifN genes are shown on the NifK phylogeny (Figure 4). A somewhat related NifN is found in the C'-type system of *Methanothermococcus*, even though its other genes are quite unrelated to those of C-type systems.

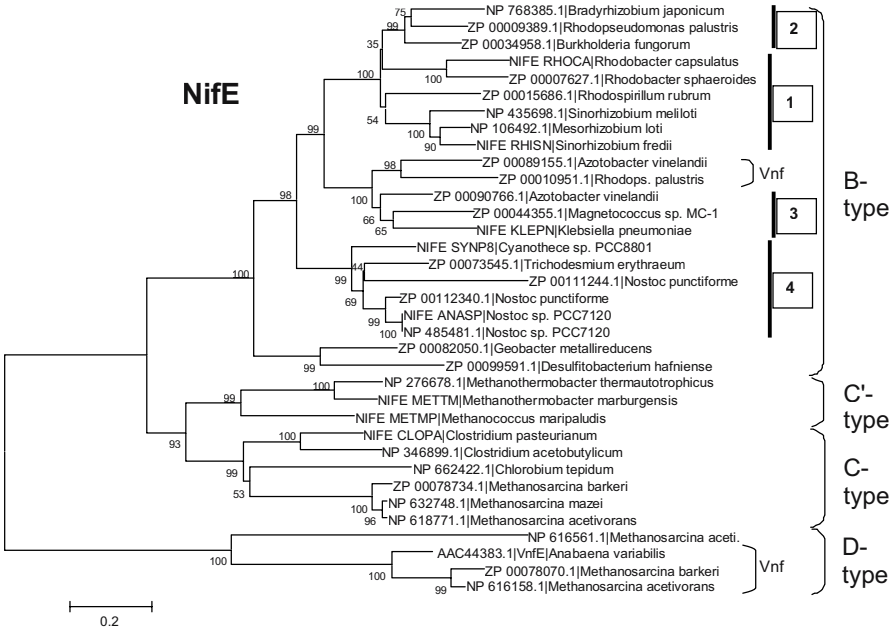

Figure 5. Phylogeny of NifE and homologous proteins from diazotroph genome sequences.

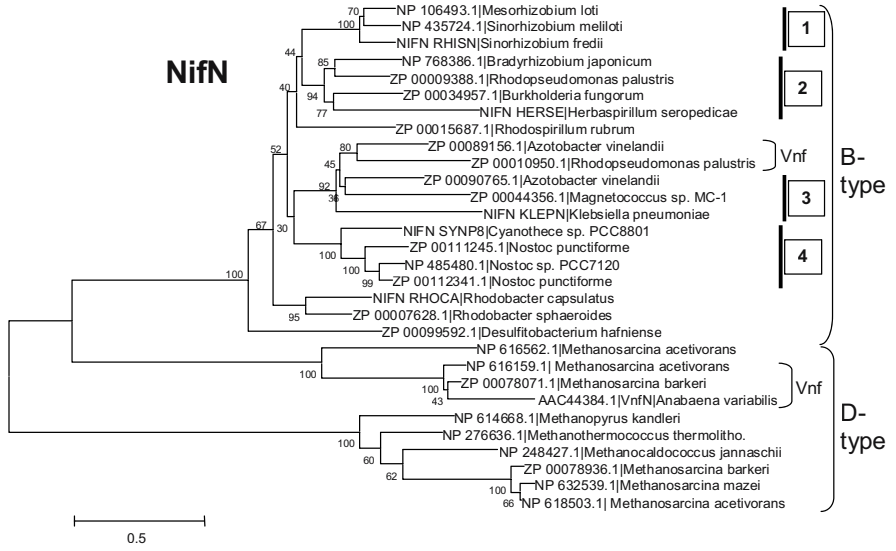

Figure 6. Phylogeny of NifN and homologous proteins from diazotroph genome sequences.

The Mo-independent nitrogenases have an extra delta subunit. This small protein is encoded by either *vnfG* or *anfG* and is located between the *D* and *K* genes. The VnfG proteins of the vanadium enzymes and the AnfG proteins of the iron-only enzymes form two separate groups (Figure 7). They are related to each other but show no significant homology to any other proteins.

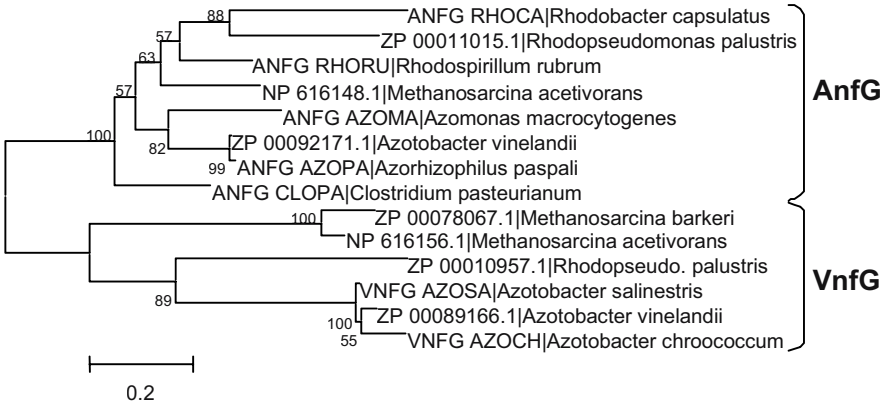

Figure 7. Phylogeny of VnfG and AnfG proteins from diazotroph genome sequences.

6. NITROGENASE PHYLOGENY *VERSUS* ORGANISM PHYLOGENY

Phylogenetic relationships among distantly related organisms are most often described using sequences of their small-subunit ribosomal RNA (SSU rRNA) genes. Such a phylogeny is shown in Figure 8 for the 19 diazotrophs whose nitrogenase proteins we have just discussed. Despite being based on very few taxa, this tree is largely consistent with more comprehensive SSU trees (Cole *et al*., 2003), although a specific relationship is seen between cyanobacteria and Gram-positive bacteria, such as clostridia, when more species are included and *Rhodobacter* moves deeper in the alpha-proteobacteria branch than it appears in Figure 8. There has been considerable discussion as to whether phylogenies based on rRNA really reflect the relationships of the genomes as a whole, but a recent analysis indicates that there is a large core of hundreds of genes, including rRNA, that does show a consistent phylogeny in most comparisons of bacteria (Daubin *et al*., 2003). It is, therefore, reasonable to compare the phylogenies of nitrogenase sequences with that of SSU rRNA and to interpret discrepancies as resulting from events in the evolution of nitrogenase. Each of the major types of nitrogenase will be discussed first, before considering the relationship between the types.

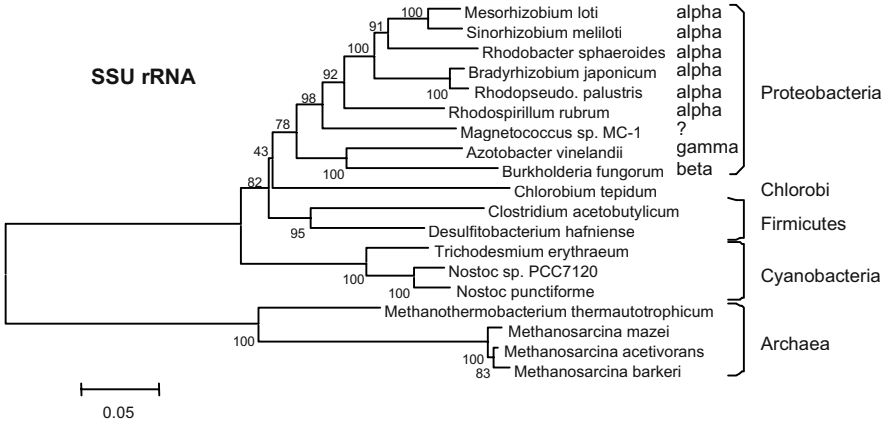

Figure 8. Phylogeny of small subunit ribosomal RNA from diazotroph genome sequences.

6.1. The history of B-type nitrogenases

B-type enzymes are widely distributed in bacteria but have not been found in archaea. Among those included in this survey, the enzyme from *Desulfitobacterium hafniense*, a firmicute or low-GC Gram-positive bacterium related to clostridia, is consistently the most divergent in every subunit (Figures 1, 3-6). The genome of *Frankia*, an actinobacterium or high-GC Gram-positive bacterium, has not been sequenced yet, but its Nif proteins are also outliers within the B-type clade (Figures 3 and 4). This situation makes some phylogenetic sense because the remainder of the B-type enzymes are from two other bacterial phyla, the proteobacteria and the cyanobacteria.

For each protein, the cyanobacterial sequences form a single clade, labelled "4" in the figures. The anomaly is that the cyanobacterial sequences are close to, and in some cases embedded within, the groups of proteobacterial sequences. This location is certainly inconsistent with the phylogenetic position of cyanobacteria which, based on SSU and other criteria, are closer to the firmicutes and certainly not particularly close to the proteobacteria. When this anomaly for NifH, D, and K was last discussed (Young, 2000), it was commented that either gene duplication or gene transfer were both plausible explanations. For example, the unexpectedly distant position of the -proteobacteria might reflect paralogy: an ancient duplication created two families of *nif* genes, one of which was retained in the γ-proteobacteria and the other in the remaining proteobacteria and in the cyanobacteria. This explanation is less plausible now that we have complete genome sequences. None of the relevant genomes have traces of an ancient duplication of the B-type nitrogenase system. It now seems more likely that there has been a transfer of nitrogenase genes between cyanobacteria and proteobacteria, although the direction is not immediately obvious. The cyanobacteria have a special control system

involving an excision element in *nifD* (Golden *et al*., 1985), but this is presumably a fairly recent acquisition since the transfer (in whichever direction).

The groups labelled "1", "2" and "3" correspond approximately to the α, □ and γ proteobacteria, respectively, and are fairly consistently maintained across the different proteins. However, *Bradyrhizobium japonicum* and *Rhodopseudomonas palustris*, two closely related α-proteobacteria, have nitrogenases that group with those of □-proteobacteria (group 2). On the whole, the different Nif proteins tell the same story, with varying degrees of certainty, but *B. japonicum* and *R. palustris*, closest neighbours for all other Nif sequences and for SSU, have NifH proteins that cluster with different proteobacterial groups (Figure 1). It is tempting to connect this situation with the fact that the *nifH* gene of *B. japonicum* is distant from the rest of the operon (Table 1), suggesting that it might have been acquired by horizontal gene transfer. It has to be said, though, that it is the *R. palustris* sequence, which is further from its expected place. Hurek *et al*. (1997) postulated lateral gene transfer when they found that some *Azoarcus* species had NifH sequences in group 2 and others in group 3. Interestingly, when Moulin *et al*. (2001) described the first example of a □-proteobacterium that nodulated legumes, the partial NifH sequence they determined was extremely similar to that of *Burkholderia fungorum* and, therefore, close to that of *Bradyrhizobium*. It seems that, while there is an overall phylogenetic pattern in the nitrogenases of proteobacteria, some horizontal transfers within and between the α, □ and γ orders have occurred. A detailed reconstruction is beyond our scope here, but mismatches between *nifH* and SSU phylogenies have been described previously, particularly among the rhizobia (Haukka *et al*., 1998).

6.2. The history of C-type nitrogenases

The second major type of Mo-dependent nitrogenase has a quite distinct distribution. There is, so far, no report of an organism with both B-type and C-type nitrogenase. The type-C nitrogenases are found in archaea and in the bacteria, *Clostridium* and *Chlorobium*. The sequences are, in general, more divergent than those of B-type systems and it could be argued that they were consistent with direct inheritance from the common ancestor of bacteria and archaea. This suggestion would require that the C-type genes were completely lost from other bacteria; there is certainly no sign of them now. The alternative idea, that genes have been transferred between archaea and bacteria, is much more acceptable now than it would have been a few years ago. The *Methanosarcina* genomes are the largest yet determined for any archaea and Deppenmeier *et al*. (2002) point out that 30% of the genes in *M. mazei* are closer to bacterial than to archaeal homologs (in fact, half of these have no significant match in any other archaeon). They ascribe this situation to massive horizontal gene transfer from the Bacteria to the Archaea, even emphasising this transfer in the title of their article.

This theme of extensive gene transfer between distant organisms (or, at least, organisms with distant rRNA sequences) has been taken up by many authors recently, including notably Ford Doolittle (*e.g*., Doolittle *et al*., 2002), who argue that there may in fact be almost no "core" genome that is immune to horizontal

transfer. An example of the kind of analysis that can lead to such a conclusion is provided by Raymond *et al.* (2002), who found that different sets of genes supported different topologies for photosynthetic bacteria. In an interesting parallel with our discussion of nitrogen fixation, they concluded that photosynthesis genes had been subject to repeated horizontal transfer during the evolution of prokaryotes. The extreme view that there is no substantial core of genes with a consistent phylogeny is countered by studies such as that of Daubin *et al*. (2003) but, even if half of the genome were "core", that would leave plenty of scope for transfer. As far as nitrogenase genes are concerned, they do not appear to be part of the core within the proteobacteria because there is evidence for transfer (see previous section), so wider-scale transfers are also plausible.

Although Deppenmeier *et al*. (2002) characterise the *M. mazei* genome as the recipient of massive transfer from bacteria, it is by no means certain that the nitrogenase genes moved in this direction. In fact, Eisen *et al*. (2002) point out that 12% of *Chlorobium tepidum* genes are closer to archaeal than to bacterial homologs, and argue that, in certain cases, the transfer has been from archaea because their phylogenetic origin is within archaea. Eisen *et al*. specifically suggest that the nitrogenase genes were transferred between kingdoms, but without defining the direction. Nölling *et al*. (2001) make similar comments about the nitrogenase genes in *Clostridium acetobutylicum*.

The direction of transfer is not obvious because C-type nitrogenases have a restricted taxonomic distribution in both Bacteria and Archaea. In Bacteria, they are known from the Firmicutes (several species of *Clostridium*) and the unrelated Chlorobi or green sulfur bacteria (*Chlorobium tepidum*). In Archaea, they are restricted to two groups of methanogens, *Methanosarcina* and *Methano-thermobacter*. The bacterial proteins are related to those in *Methanosarcina*, but the *Methanothermobacter* genes are much more divergent (the C'-type) and might be thought unrelated except that they share both the gene organisation (two P_{II} proteins encoded after *nifH*) and the NifK-like "NifN". Did these systems originate early in the Archaea as C and C'-types, with C-type later spreading to bacteria (perhaps separately to clostridia and chlorobi because their genes are not closely related)? Or is C'-type the true archaeal version, with *Methanosarcina* having recently adopted a bacterial C-type version?

6.3. The history of Vnf nitrogenases

The vanadium-dependent nitrogenases are also found in both the Archaea and the Bacteria and, in this case, there is much clearer evidence for horizontal gene transfer as well as for reassortment of the genes within the operon(s). The two archaeal examples, from *Methanosarcina acetivorans* and *M. barkeri*, are closely similar in all respects, which is not surprising because these bacteria are close relatives. Their genes are consistently A-type, related to Anf genes. The two bacterial examples from sequenced genomes are both proteobacteria, but a Vnf system is also described in the cyanobacterium, *Anabaena variabilis* (Thiel, 1993; 1996) and this has been included in the phylogenies (Figures 3-6). Each bacterial system shows peculiarities that suggest a mixed origin.

In *Azotobacter vinelandii*, the VnfD, G, and K sequences (Joerger *et al*., 1990) are A-type, related to the corresponding archaeal Vnf genes. However, VnfH is of B-type and very similar to the NifH in the same organism. It has clearly been recruited by a recent duplication of the NifH. The *vnfE* and *vnfN* genes (Wolfinger and Bishop, 1991) are not after *vnfK*, instead they are few genes upstream of *vnfH* and, like it, they are B-type and related to their *nif* equivalents in the same species. Vnf (and sometimes also Anf) was found in several other γ-proteobacteria by Loveless *et al*. (1999). They used primers in *vnfD* and *vnfK* to amplify *vnfG*, thus, verifying the gene order *vnfDGK*, but did not investigate the other *vnf* genes.

The gene organisation in *R. palustris* is even more disrupted, with *vnfH* inverted relative to *vnfDGK*. Again, VnfD, G, and K are A-type, VnfN and E are B-type and specifically related to the *A. vinelandii* equivalents, *i.e*., they are in the γ-proteobacterial cluster 3 rather than cluster 2 with the Nif proteins of *R. palustris*. This situation suggests that the α-proteobacterium, *R. palustris*, has acquired the Vnf system from a γ-proteobacterium. The *R. palustris* VnfH, on the other hand, is extremely similar to NifH in the same organism and has clearly been recruited locally in a similar fashion to the VnfH of *A. vinelandii*, but independently.

There is as yet no genome sequence for *Anabaena variabilis* ATCC 29413 as distinct from *Anabaena* sp. PCC1720, which is called *Nostoc* at NCBI and herein. However, the Vnf system of *A. variabilis* is sufficiently distinctive to be worth mentioning here. The arrangement of the genes is *vnf[DG]KEN*, where *D* and *G* are fused (Thiel, 1993; 1996). No *vnfH* has been found. The VnfD part of the VnfDG fused protein is A-type and roughly equidistant between the homologous bacterial (*A. vinelandii* and *R. palustris*) and archaeal (*Methanosarcina*) sequences (Figure 3) and the same is true of the VnfG-like part of this protein (not shown). The *Anabaena* VnfK is unambiguously related to the archaeal rather than the other bacterial sequences (Figure 4). Even more surprisingly, the VnfE and VnfN are very close to the *Methanosarcina* sequences (Figures 5 and 6) and, therefore, completely unlike the B-type proteins found in the other bacterial systems. In other words, all the known components of the *A. variabilis* Vnf system are similar to those in archaea and, especially in the case of VnfE and N, much too similar to be explained by descent from the common ancestor of bacteria and archaea.

There has clearly been horizontal transfer of the *vnf* genes between archaea and bacteria, but can we be sure of the direction? Vnf systems have been found in several taxonomic groups of bacteria, and their organisation is diverse and rather loose, whereas we only know of Vnf in two closely related archaea, which have a compact, self-contained operon that might have been delivered as a package from some bacterium. On the other hand, the bacterial examples of Vnf look like *bricolage*, rather amateur attempts to piece together a working system around a chance acquisition of the core *vnfDGK* genes. Further support for the idea that Vnf comes from archaea is provided by the organisation of the *Methanosarcina acetivorans vnf* operon, which is back-to-back with its *anf* operon in a manner that suggests the possibility of joint control. The arrangement of the genes is

anfH - - K G D I2 I1 vnfH I1 I2 D G K E N

where underlining indicates that the genes are in reverse orientation. The *I1* and *I2* genes encode GlnB-like P_{II} proteins, the *vnf* and *anf* equivalents are related but distinct. On the other hand, the gene labelled *anfH* is 97% identical to the *vnfH* gene and encodes an identical polypeptide; however, there is no evidence about its activity in either the Anf or Vnf systems. The two ORFs that separate it from the *anf* operon have no obvious connection with nitrogen fixation.

6.4. The history of Anf nitrogenases

The distribution of Anf systems among our sequenced genomes is correlated with that of Vnf. *Rhodopseudomonas palustris*, *Azotobacter vinelandii* and *Methanosarcina acetivorans* have both and the incomplete *M. barkeri* genome certainly has part of an Anf system (*anfDGK*, but as yet no *anfI1I2* or *H*). In addition, *Rhodospirillum rubrum* has Anf, although it does not have Vnf. Anf has also been found in *Rhodobacter capsulatus* (Masepohl and Klipp 1996) and in *Clostridium pasteurianum* (Zinoni *et al*., 1993), but the sequenced *C. acetobutylicum* does not have it. Loveless and Bishop (1999) reported an *anfD*-like sequence in *Heliobacterium gestii*, another firmicute.

Phylogenetically, the Anf systems are much more coherent than Vnf. In all cases, the components are A-type and most closely related to other Anf systems. Only *anfD* and *anfG* have been sequenced from *C. pasteurianum*, but both fall towards the edge of the Anf clade as might be expected (for *anfG* see Figure 7; *anfD* not shown). For each gene, the variation among the proteobacteria is roughly comparable with that seen for the equivalent B-type gene, while the distance to the archaeal sequences is not much greater than this. Unless there are extraordinary functional constraints preventing further divergence of the Anf genes, the conclusion must be that the Anf genes in *Methanosarcina* shared a common ancestor with the bacterial examples early in the history of the proteobacteria, and certainly long after the common ancestor of archaea and bacteria. Until more systems are characterised, it is hard to say whether this represents an old transfer in either direction or a much more recent transfer to *Methanosarcina* from a bacterium that has not yet been examined.

6.5. The origin of nitrogenases

It is clear that there are three major types of nitrogenase, each with phylogenetically distinct set of H, D, K, E and N proteins. The A-type systems are Mo-independent (either Vnf or Anf) and are widespread among bacteria and known from the archaeon, *Methanosarcina*. B-type is common among the proteobacteria and cyanobacteria and is found in *Desulfitobacterium*, a relative of *Clostridium*, but not so far in *Clostridium* itself, nor in any archaeon. C-type is found in various archaea (*Methanosarcina* and *Methanothermobacter*) and in *Clostridium* and *Chlorobium*, which represent two related bacterial groups. All these systems share related components and even a conserved gene order, implying that they must have a common ancestry. They are, however, very divergent, which is consistent with the widely held view that nitrogenase is a very ancient system that was already present

in the last universal common ancestor of all modern life. Given that there is evidence for interdomain transfer of both A-type and C-type, it not easy to determine whether the early development of these systems occurred in bacteria or in archaea.

7. NITROGENASE GENES IN THEIR GENOMIC CONTEXT

In addition to the sequence and gene organisation information that has already been discussed, complete genome sequences provide other potential clues about the origin of the nitrogenase genes in them. The location of the genes may indicate whether they are either long-term residents or recent intruders, as may their base composition relative to the surrounding genome.

In the rhizobia, the *nif* genes are on either plasmids or genomic islands, which are, in both cases, generally lower in G+C content than the basic chromosomal genome (Galibert *et al*., 2001; Kaneko *et al*., 2000; 2002). These parts of the genome are prone to transfer, at least between related bacteria, and include many genes that are specific to particular isolates or species. It can be said, therefore, that in these bacteria the *nif* genes form part of the accessory genome. Although their context is a relatively low G+C part of the genome, the *nifH*, *D*, *K,* and *E* genes have extremely high G+C (Young, unpublished analysis). The reason is unknown and it is not true in general of other *nif*-related genes in these organisms.

In the other sequenced organisms, the nitrogenase genes are chromosomally located, even when, as in *Nostoc* sp. PCC 7120, the genome also includes large plasmids. However, these genomes have not been systematically examined for genomic islands, so it is not clear whether the context of the nitrogenase genes is truly in the basic chromosome or in the accessory genome. It is possible that a careful analysis would reveal that each of the nitrogenase types was truly "at home" in certain organisms. On the other hand, we know of no major group of organisms in which nitrogen fixation is universal, so perhaps nitrogenase lives a perpetually nomadic existence as a life-long member of the horizontal gene pool.

8. CONCLUSIONS AND PROSPECTS

It is too soon for conclusions. Our access to genome information is still very limited, but some things are already becoming clearer. Now that there are hundreds of microbial genome sequences, we can see that nitrogenase genes are only present in a minority of them (and, incidentally, absent from eukaryotes). These data confirm the observation that most organisms do not fix N_2 and remove the nagging doubt that we just had not found the right conditions. The availability of more complete sets of nitrogenase sequences has reinforced the distinctness of the A, B and C types; the distinction is recognisable in each of the components. Phylogenetic evidence makes it virtually certain that nitrogenase genes have been transferred horizontally, even between bacteria and archaea, but it is difficult to be certain of the direction.

Analysis of the genomic context of the genes is still in its infancy, but promises to provide new insights into the distribution of nitrogenase genes and their

integration into the metabolism of their host organisms. Of course, genome-gazing is no substitute for experimental evidence, but it can tie together a mass of observations and can also suggest interesting questions for the future.

REFERENCES

Aguilar, O. M., Reilander, H., Arnold, W., and Pühler, A. (1987). *Rhizobium meliloti nifN* (*fixF*) gene is part of an operon regulated by a NifA-dependent promoter and codes for a polypeptide homologous to the *nifK* gene product. *J. Bacteriol., 169*, 5393-5400.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J. H., Zhang, Z., Miller, W., *et al*. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res., 25*, 3389-3402.

Brigle, K. E., Weiss, M. C., Newton, W. E., and Dean, D. R. (1987). Products of the iron-molybdenum cofactor-specific biosynthetic genes, *nifE* and *nifN*, are structurally homologous to the products of the nitrogenase molybdenum-iron protein genes, *nifD* and *nifK*. *J. Bacteriol., 169*, 1547-1553.

Burke, D. H., Hearst, J. E., and Sidow, A. (1993). Early evolution of photosynthesis - clues from nitrogenase and chlorophyll iron proteins. *Proc. Natl. Acad. Sci. USA, 90*, 7134-7138.

Chien, Y. T., Auerbuch, V., Brabban, A. D., and Zinder, S. H. (2000). Analysis of genes encoding an alternative nitrogenase in the archaeon *Methanosarcina barkeri* 227. *J. Bacteriol., 182*, 3247-3253.

Cole, J. R., Chai, B., Marsh, T. L., Farris, R. J., Wang, Q., Kulam, S. A., *et al*. (2003). The Ribosomal Database Project (RDP-II): Previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. *Nucleic Acids Res., 31*, 442-443.

Daubin, V., Moran, N. A., and Ochman, H. (2003). Phylogenetics and the cohesion of bacterial genomes. *Science, 301*, 829-832.

Dean, D. R., and Brigle, K. E. (1985). *Azotobacter vinelandii nifD*-encoded and *nifE*-encoded polypeptides share structural homology. *Proc. Natl. Acad. Sci. USA, 82*, 5720-5723.

Dean, D. R., and Jacobson, M. R. (1992). Biochemical genetics of nitrogenase. In G. Stacey and R. H. Burris and H. J. Evans (Eds.), *Biological nitrogen fixation.* (pp. 763-834). New York, NY: Chapman and Hall.

Deppenmeier, U., Johann, A., Hartsch, T., Merkl, R., Schmitz, R. A., Martinez-Arias, R., *et al*. (2002). The genome of *Methanosarcina mazei*: Evidence for lateral gene transfer between bacteria and archaea. *J. Mol. Microbiol. Biotechnol., 4*, 453-461.

Doolittle, W. F., Boucher, Y., Nesbø, C. L., Douady, C. J., Andersson, J. O., and Roger, A. J. (2003). How big is the iceberg of which organellar genes in nuclear genomes are but the tip? *Phil. Trans. Royal Soc. Series B, 358*, 39-57.

Eisen, J. A., Nelson, K. E., Paulsen, I. T., Heidelberg, J. F., Wu, M., Dodson, R. J., *et al*. (2002). The complete genome sequence of *Chlorobium tepidum* TLS, a photosynthetic, anaerobic, green-sulfur bacterium. *Proc. Natl. Acad. Sci. USA, 99*, 9509-9514.

Fani, R., Gallo, R., and Lio, P. (2000). Molecular evolution of nitrogen fixation: The evolutionary history of the *nifD*, *nifK*, *nifE*, and *nifN* genes. *J. Mol. Evol., 51*, 1-11.

Galagan, J. E., Nusbaum, C., Roy, A., Endrizzi, M. G., Macdonald, P., FitzHugh, W., *et al*. (2002). The genome of *M. acetivorans* reveals extensive metabolic and physiological diversity. *Genome Res., 12*, 532-542.

Galibert, F., Finan, T. M., Long, S. R., Pühler, A., Abola, P., Ampe, F., *et al*. (2001). The composite genome of the legume symbiont *Sinorhizobium meliloti*. *Science, 293*, 668-672.

Golden, J. W., Robinson, S. J., and Haselkorn, R. (1985). Rearrangement of nitrogen fixation genes during heterocyst differentiation in the cyanobacterium *Anabaena*. *Nature, 314*, 419-423.

Haukka, K., Lindström, K., and Young, J. P. W. (1998). Three phylogenetic groups of *nodA* and *nifH* genes in *Sinorhizobium* and *Mesorhizobium* isolates from leguminous trees growing in Africa and Latin America. *Appl. Environ. Microbiol., 64*, 419-426.

Hennecke, H., Kaluza, K., Thöny, B., Fuhrmann, M., Ludwig, W., and Stackebrandt, E. (1985). Concurrent evolution of nitrogenase genes and 16S ribosomal RNA in *Rhizobium* species and other nitrogen-fixing bacteria. *Arch. Microbiol., 142*, 342-348.

Holland, D., Zilberstein, A., Zamir, A., and Sussman, J. L. (1987). A quantitative approach to sequence comparisons of nitrogenase MoFe protein alpha subunits and beta subunits including the newly sequenced *nifK* gene from *Klebsiella pneumoniae*. *Biochem. J*., *247*, 277-285.

Hurek, T., Egener, T., and Reinhold-Hurek, B. (1997). Divergence in nitrogenases of *Azoarcus* spp., *Proteobacteria* of the beta subclass. *J. Bacteriol., 179*, 4172-4178.

Joerger, R. D., Loveless, T. M., Pau, R. N., Mitchenall, L. A., Simon, B. H., and Bishop, P. E. (1990). Nucleotide sequences and mutational analysis of the structural genes for nitrogenase 2 of *Azotobacter vinelandii*. *J. Bacteriol., 172*, 3400-3408.

Kaminski, P. A., Batut, J., and Boistard, P. (1998). A survey of symbiotic nitrogen fixation by rhizobia. In H. P. Spaink, A. Kondorosi and P. J. J. Hooykaas (Eds.), *The Rhizobiaceae* (pp. 431-460). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Kaneko, T., Nakamura, Y., Sato, S., Asamizu, E., Kato, T., Sasamoto, S., *et al*. (2000). Complete genome structure of the nitrogen-fixing symbiotic bacterium *Mesorhizobium loti*. *DNA Res., 7*, 331-338.

Kaneko, T., Nakamura, Y., Sato, S., Minamisawa, K., Uchiumi, T., Sasamoto, S., *et al*. (2002). Complete genomic sequence of nitrogen-fixing symbiotic bacterium *Bradyrhizobium japonicum* USDA110. *DNA Res., 9*, 189-197.

Kaneko, T., Nakamura, Y., Wolk, C. P., Kuritz, T., Sasamoto, S., Watanabe, A., *et al*. (2001). Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium *Anabaena* sp. strain PCC 7120. *DNA Res., 8*, 205-213.

Kessler, P. S., and Leigh, J. A. (1999). Genetics of nitrogen regulation in *Methanococcus maripaludis*. *Genetics, 152*, 1343-1351.

Kessler, P. S., McLarnan, J., and Leigh, J. A. (1997). Nitrogenase phylogeny and the molybdenum dependence of nitrogen fixation in *Methanococcus maripaludis*. *J. Bacteriol., 179*, 541-543.

Kumar, S., Tamura, K., Jakobsen, I. B., and Nei, M. (2001). MEGA2: Molecular evolutionary genetics analysis software. *Bioinformatics, 17*, 1244-1245.

Loveless, T. M., and Bishop, P. E. (1999). Identification of genes unique to Mo-independent nitrogenase systems in diverse diazotrophs. *Can. J. Microbiol., 45*, 312-317.

Loveless, T. M., Saah, J. R., and Bishop, P. E. (1999). Isolation of nitrogen-fixing bacteria containing molybdenum-independent nitrogenases from natural environments. *Appl. Environ. Microbiol., 65*, 4223-4226.

Masepohl, B., and Klipp, W. (1996). Organization and regulation of genes encoding the molybdenum nitrogenase and the alternative nitrogenase in *Rhodobacter capsulatus*. *Arch. Microbiol., 165*, 80-90.

Moulin, L., Munive, A., Dreyfus, B., and Boivin-Masson, C. (2001). Nodulation of legumes by members of the beta-subclass of *Proteobacteria*. *Nature, 411*, 948-950.

Nölling, J., Breton, G., Omelchenko, M. V., Makarova, K. S., Zeng, Q. D., Gibson, R., *et al*. (2001). Genome sequence and comparative analysis of the solvent-producing bacterium *Clostridium acetobutylicum*. *J. Bacteriol., 183*, 4823-4838.

Normand, P., and Bousquet, J. (1989). Phylogeny of nitrogenase sequences in *Frankia* and other nitrogen-fixing microorganisms. *J. Mol. Evol., 29*, 436-447.

Postgate, J. R. (1974). Evolution within nitrogen-fixing systems. In M. J. Carlile and J. J. Skehel (Eds.), *SGM Symposium 24: Evolution in the microbial world* (pp. 263-292). Cambridge, UK: Cambridge University Press.

Postgate, J. R. (1982). *The fundamentals of nitrogen fixation*. Cambridge, UK: Cambridge University Press.

Postgate, J. R., and Eady, R. R. (1988). The evolution of biological nitrogen fixation. In H. Bothe, F. J. De Bruijn and W. E. Newton (Eds.), *Nitrogen fixation: Hundred years after* (pp. 31-40). Stuttgart, Germany: Gustav Fischer.

Raymond, J., Zhaxybayeva, O., Gogarten, J. P., Gerdes, S. Y., and Blankenship, R. E. (2002). Whole-genome analysis of photosynthetic prokaryotes. *Science, 298*, 1616-1620.

Ribbe, M., Gadkari, D., and Meyer, O. (1997). N_2 fixation by *Streptomyces thermoautotrophicus* involves a molybdenum dinitrogenase and a manganese superoxide oxidoreductase that couple N_2 reduction to the oxidation of superoxide produced from O_2 by a molybdenum-CO dehydrogenase. *J. Biol. Chem., 272*, 26627-26633.

Smith, D. R., Doucette-Stamm, L. A., Deloughery, C., Lee, H. M., Dubois, J., Aldredge, T., *et al*. (1997). Complete genome sequence of *Methanobacterium thermoautotrophicum* Delta H: Functional analysis and comparative genomics. *J. Bacteriol., 179*, 7135-7155.

Thiel, T. (1993). Characterization of genes for an alternative nitrogenase in the cyanobacterium *Anabaena variabilis*. *J. Bacteriol., 175*, 6276-6286.

Thiel, T. (1996). Isolation and characterization of the *vnfEN* genes of the cyanobacterium *Anabaena variabilis*. *J. Bacteriol., 178*, 4493-4499.

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., and Higgins, D. G. (1997). The CLUSTAL_X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res., 25*, 4876-4882.

Tiedje, J. M., and Davis, J. K. (2002). *Co-occurrence of dehalogenation and nitrogen fixation in reductively dechlorinating bacteria.* Paper presented at the US-Japan Joint Workshop on Systems Biology of Useful Microorganisms, Keio University, Japan.

Wolfinger, E. D., and Bishop, P. E. (1991). Nucleotide sequence and mutational analysis of the *vnfENX* Region of *Azotobacter vinelandii*. *J. Bacteriol., 173*, 7565-7572.

Young, J. P. W. (1990). The phylogeny of *nifH*: Gene duplication or lateral transfer? In P. M. Gresshoff, L. E. Roth, G. Stacey and W. E. Newton (Eds.), *Nitrogen fixation: Achievements and objectives* (pp. 840). New York, NY: Chapman and Hall.

Young, J. P. W. (1992). Phylogenetic classification of nitrogen-fixing organisms. In G. Stacey, R. H. Burris and H. J. Evans (Eds.), *Biological nitrogen fixation.* (pp. 43-86). New York, NY: Chapman and Hall.

Young, J. P. W. (2000). Molecular evolution in diazotrophs: Do the genes agree? In F. O. Pedrosa, M. Hungria, M. G. Yates and W. E. Newton (Eds.), *Nitrogen fixation: From molecules to crop productivity* (pp. 161-164). Dordrecht, The Netheralnds: Kluwer.

Zinoni, F., Robson, R. M., and Robson, R. L. (1993). Organization of potential alternative nitrogenase genes from *Clostridium pasteurianum*. *Biochim. Biophys. Acta, 1174*, 83-86.