Minireview

Miraculous catch of iron–sulfur protein sequences in the Sargasso Sea

Jacques Meyer*

Laboratoire de Biophysique Moléculaire et Cellulaire (UMR 5090 CEA-CNRS-Université J. Fourier), DRDC, CEA-Grenoble, 38054 Grenoble, France

Received 23 April 2004; accepted 10 June 2004

Available online 22 June 2004

Edited by Takashi Gojobori

Abstract Recent shotgun sequencing of filtered Sargasso Sea water samples has yielded data in astounding amount and diversity. Iron–sulfur proteins, which are ancient, diverse and ubiquitous, have been implemented here to further probe the sequence diversity of the Sargasso Sea database (SSDB). Sequence searches and comparisons confirm that the SSDB by and large equals in diversity the combined currently available databases. The data thus suggest that microbial diversity has so far been underestimated by orders of magnitude.

Keywords: Biodiversity; Evolution; Ferredoxin; Hydrogenase; Nitrogenase; Metalloprotein

1. Introduction

Genome sequencing on a production scale has hugely increased the rate of DNA and protein sequence data acquisition. A further step has been taken through retrieval of genomic sequences directly from the environment, thus overcoming the limits set by the cultivability of organisms [1,2]. A massive effort of that kind has recently resulted in the generation, from ca. 2 m^3 of Sargasso Sea water, of over 1 billion basepairs of non-redundant sequence, i.e., ca. 10% of all sequence data collected to date. Even more impressive is the realization that the diversity of these new sequences may be comparable to that of all previously known ones [3,4].

We describe hereafter an assessment of the diversity of the Sargasso Sea sequence database (SSDB) by using iron–sulfur (Fe–S) proteins as benchmarks. Several reasons suggest these metalloproteins as appropriate tools for such a task. They are collectively ubiquitous and involved in various essential cellular functions [5]. They include numerous phylogenetically distinct families, of which the populations vary greatly in numbers. Many Fe–S proteins, most notably ferredoxins (Fd) [6,7] and rubredoxins (Rd) [8], are small (less than 130 amino acids) and therefore less likely to occur as truncated sequences in the short DNA fragments (1–2 kb) composing a large part of the SSDB [3]. In addition, the emergence of Fe–S proteins may have been contemporaneous with the origin of life, and perhaps even incipient in prebiotic chemistry [9]. The evolution of Fe–S proteins may thus span the full history of life on this planet.

2. BLAST searches

The presence of Fe–S protein sequences in the SSDB was first assessed by running BLAST searches [10] with several types of Fe–S protein sequences as baits. The data are summarized in Table 1, and compared with those of similar searches performed on the non-redundant NCBI database (http://www.ncbi.nlm.nih.gov/), hereafter designated NRDB. The latter is in fact a combination of the GenBank, EMBL, DDJB, and PDB databases. With few exceptions, to be discussed below, the numbers of hits in the two databases differ by less than an order of magnitude. Typically, in the first six lines of Table 1 (Rd and Fd), the ratio is within the 0.64–4.33 range. Complex I (NADH:ubiquinone oxidoreductase) of respiratory chains in its prokaryotic version is composed of 14 subunits [11], five of which contain Fe–S clusters. Only the data for subunit NuoE are shown here. Those for other subunits of the complex are quite similar (not shown) and indicate that this enzyme is better represented in the SSDB than in the NRDB. In contrast, sequences of both types of hydrogenases, the [NiFe] and [Fe] ones [12], are very rare in the SSDB. A similar trend, though less extreme, is noted for NifH, one of the nitrogenase components [13]. The IscU/NifU proteins belong to a very conserved family of Fe–S proteins that are involved in the biosynthesis of Fe–S active sites [14]. Because of their essentiality and ubiquity, their inventory is likely to provide a realistic picture of the biodiversity in any biotope. The data in Table 1 indicate that the SSDB includes twice as many IscU/NifU sequences as the NRDB.

BLAST searches were also run with a metalloprotein outside the realm of Fe–S proteins, namely cytochrome *c*, a benchmark hemoprotein. The numbers of hits were in the upper range of those observed for Fe–S proteins. Hits were more numerous in the NRDB database when a eukaryotic cytochrome *c* was used as the bait, but conversely more numerous in the SSDB when the bait was a bacterial cytochrome *c*. Similar observations were made with the "plant- and mammalian-type" [2Fe–2S]Fd (Table 1, lines 2 and 3), which in fact, notwithstanding their designation, occur in bacteria as

* Fax: +33-4-38-78-54-87.
E-mail address: jacques.meyer@cea.fr (J. Meyer).

Abbreviations: SSDB, Sargasso Sea database; NRDB, NCBI non-redundant database; Fd, ferredoxin; Rd, rubredoxin

0014-5793/$22.00 © 2004 Published by Elsevier B.V. on behalf of the Federation of European Biochemical Societies.
doi:10.1016/j.febslet.2004.06.030

Table 1
Numbers of hits in the SSDB and in the NRDB generated with various Fe–S protein sequences as baits

Bait sequence			*E*-value cut-off	Hits in SSDB	Hits in NRDB	Ratio SSDB/NRDB
Protein	Length (amino-acid residues)	Database entry				
Rd	54	AAA23279	10^{-6}	84	109	0.77
Plant-type Fd (spinach)	97	FESP1	10^{-6}	118	183	0.64
Plant-type Fd (*Aquifex aeolicus*)	96	P59799	10^{-6}	60	36	1.67
Thioredoxin-like [2Fe–2S]Fd	102	P07324	10^{-6}	49	35	1.40
2[4Fe–4S]Fd	55	P00195	10^{-6}	212	49	4.33
High potential [4Fe–4S]Fd	86	P00260	10^{-4}	14	13	1.08
Complex I (subunit E)	160	AAC06799	10^{-6}	315	152	2.07
[NiFe] hydrogenase (large subunit)	597	P15284	10^{-10}	11	157	0.07
[Fe] hydrogenase	497	CAC83731	10^{-10}	2	118	0.02
Nitrogenase (NifH = Fe protein)	290	P00459	10^{-10}	42	>500	<0.08
NifU/IscU	81	NP_897778	10^{-6}	302	158	1.91
Cytochrome *c* (mammalian)	104	P00004	10^{-6}	259	>1000	<0.25
Cytochrome *c* (bacterial c4)	210	AAA87314	10^{-6}	137	49	2.80

The TBLASTN [10] program was used in all cases, except for the high potential [4Fe–4S]Fd, where hits from TBLASTN and BLASTP [10] were combined. In each line the numbers of hits are those with *E*-values lower than the indicated cut-off.

well. These discrepancies arise from the relatively high abundance of eukaryotic sequences in the NRDB, while eukaryotes have mostly been excluded from the SSDB by filtration of the sea water samples [3].

3. Sequence alignments

A more detailed analysis was performed on some protein families. Comprehensive sets of sequences were retrieved, aligned [15], and dendrograms [16] were derived from the alignments. Partial sequences, a common occurrence in the small-sized DNA fragments composing much of the SSDB, were discarded. In case amino-acid sequences were identical, a single one was retained, regardless of differences in the encoding gene sequences. Sequences that did not include all ligands of the Fe–S active site in the appropriate positions were also rejected.

For Rd, 33 sequences were retrieved from the SSDB and 62 from the NRDB. The alignment yielded the dendrogram shown in Fig. 1. A majority (21) of the SSDB sequences form a single cluster branching off *Pseudomonas*-related sequences, and most of the remainder are related to sequences from other aerobes or cyanobacteria (Fig. 1). In contrast, no SSDB sequences are nested within those of Archae or anaerobes (a single exception is related to *Thermotoga maritima* Rd).

The SSDB includes 103 sequences of plant- and mammalian–type Fd, while 179 such sequences were recently retrieved from the NRDB [18]. The SSDB sequences are widely distributed (not shown) across all major functional groups (photosynthesis, hydroxylation reactions, Fe–S cluster biosynthesis) of these proteins [19].

Thioredoxin-like [2Fe–2S]Fd sequences were found in equal numbers (30) in the SSDB and NRDB. Half of the former are related to the cyanobacterial ones (Fig. 2), while the remainder are found among those of aerobes. None of the SSDB sequences is a close relative of any of the three biochemically characterized proteins of that family [6].

A vast majority of high-potential [4Fe–4S]Fd (21 sequences in the NRDB) are found in photoauxotrophic bacteria [7]. Their very narrow distribution probably explains the rarity of these proteins. The 11 SSDB sequences are clustered in two groups (Fig. 3) of undetermined affiliation.

4. Conclusions

The data collected in Table 1 indicate that the sequence diversity of the SSDB is overall similar to that of the NRDB, even though the former is ca. 10 times smaller than the latter. Furthermore, there are indications that the relative diversity of the SSDB might be underestimated. First, the non-redundancy of the NRDB is imperfect: indeed, a number of sequences have been obtained independently in more than one laboratory, are represented in more than one database entry, and therefore yield more than one hit in BLAST searches. Second, the filters used for sample collection eliminated free DNA, viruses, and virtually all eukaryotes, including microscopic ones, from the SSDB [3,4]. Even prokaryotic diversity has probably been underestimated, through elimination of symbiotic and particle-associated organisms [4]. Thus, the actual microbial diversity of the SSDB will possibly dwarf that of the NRDB.

Fe–S protein sequence alignments further highlight the SSDB diversity (Figs. 1–3). For the four independent families of proteins that were analyzed, the numbers of SSDB sequences are 50–100% of those in the NRDB. Expectedly, some of the SSDB sequences appear in clusters of closely similar units. However, most of them differ significantly from known sequences and thus expand the known sequence space. As stated in previous reports [6,18], it is difficult to derive taxonomic information from these short and promiscuous sequences. Nevertheless, sequence similarities within each family of Fe–S proteins in the SSDB suggest that the host microorganisms are mostly cyanobacteria, photosynthetic bacteria, or aerobes. Archaea and anaerobes appear to be barely represented in the SSDB. The large predominance of aerobes over anaerobes in the SSDB is further confirmed by the presence of numerous complex I (an aerobic enzyme) sequences and very few nitrogenase and hydrogenase (mostly found in anerobes) sequences (Table 1). These observations agree with those made for other sequence families [3] and are in keeping with the

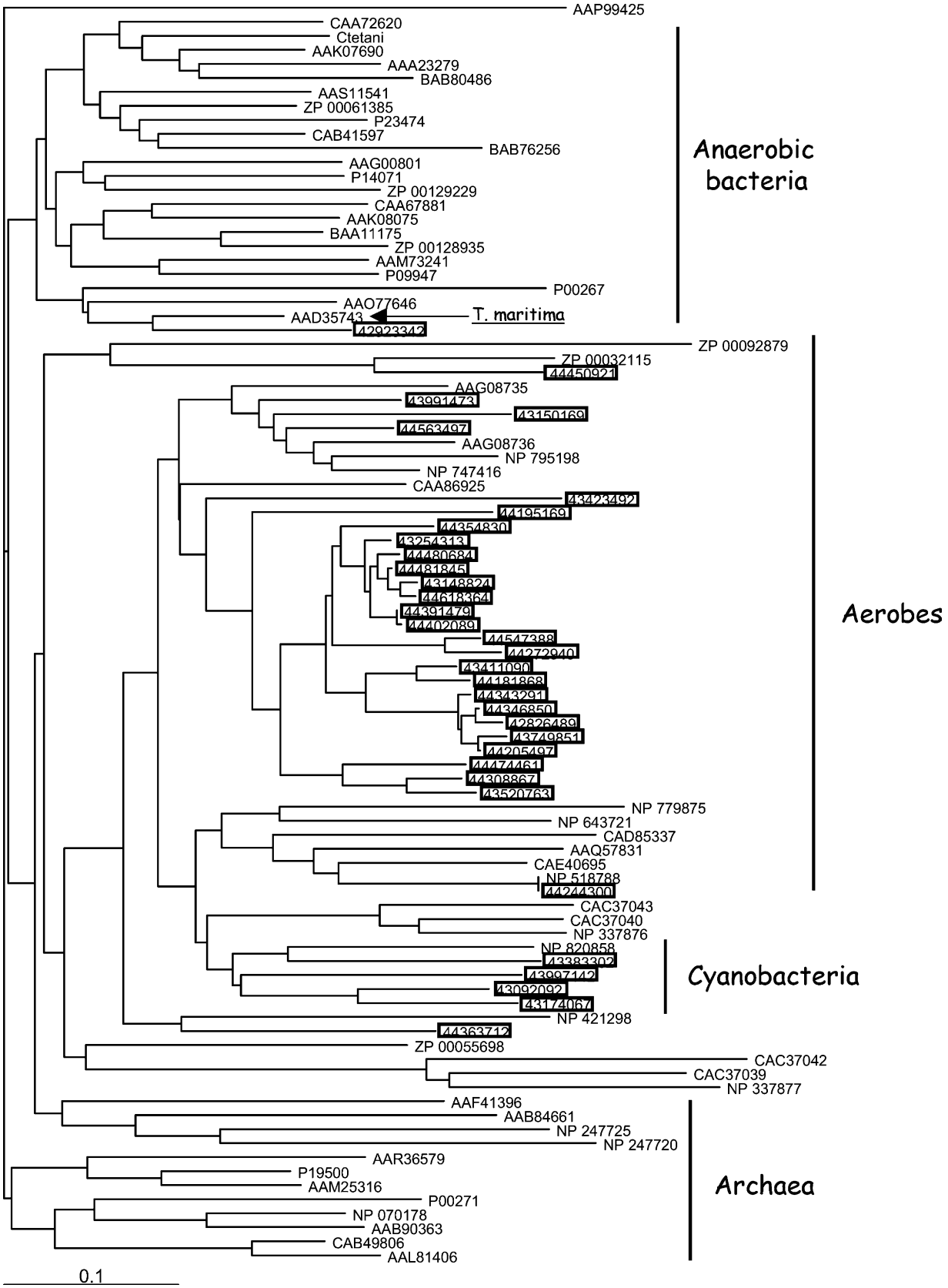

Fig. 1. Dendrogram [16] derived from sequence alignments [15] of Rd. All sequences are given as database access numbers, except for the *Clostridium tetani* protein (Ctetani), which was not annotated in the genome sequence [17]. Sequence entries from the SSDB [3] are composed of eight digits (no letters) and are framed. Categories of organisms are indicated to show that nearly all SSDB sequences are likely from aerobes or cyanobacteria. Anaerobes (with the possible exception of a *T. maritima*-related sequence) and Archaea are apparently not represented.

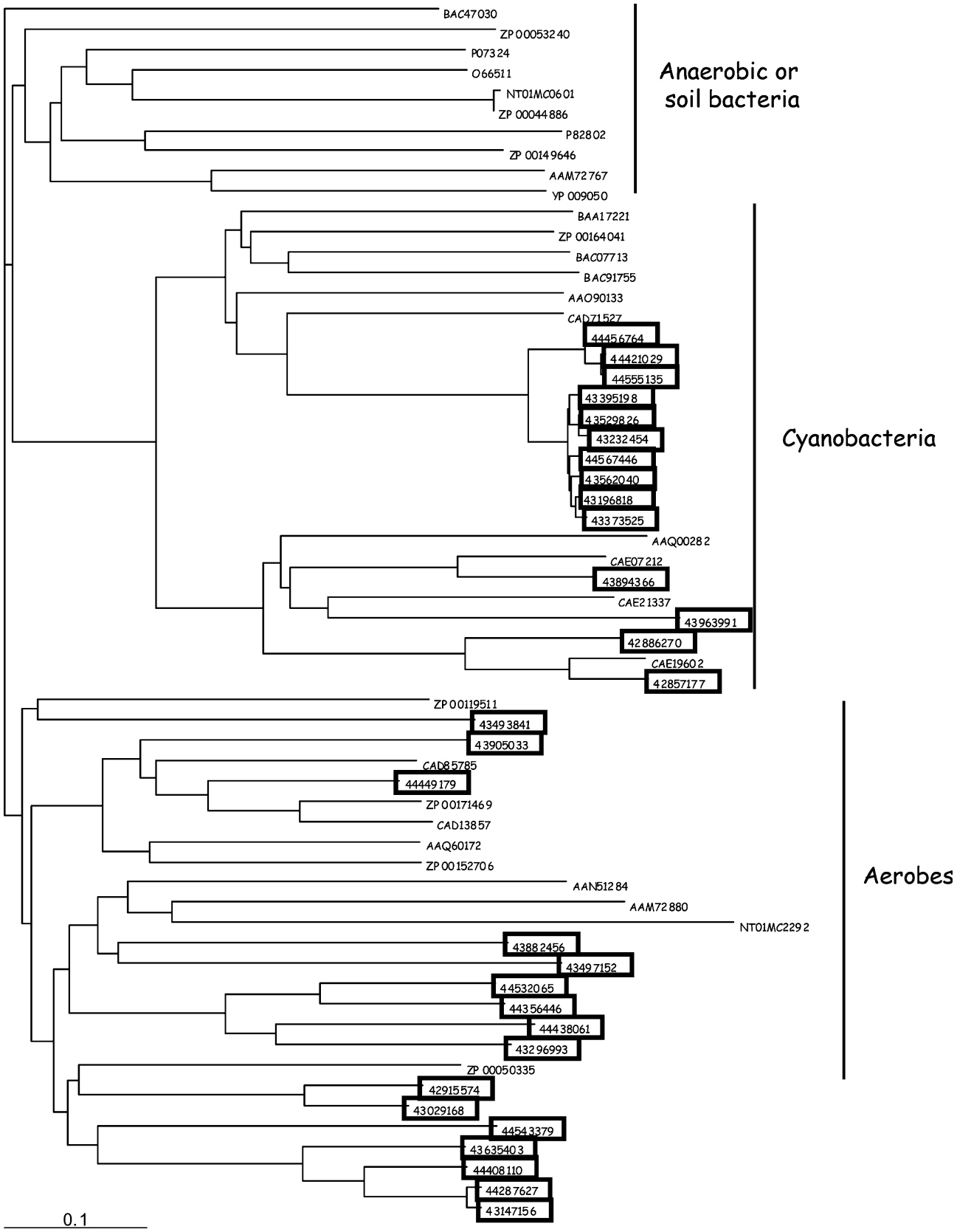

Fig. 2. Dendrogram [16] derived from sequence alignments [15] of [2Fe–2S] thioredoxin-like Fd. All sequences are given as database access numbers. Sequence entries from the SSDB [3] are composed of eight digits (no letters) and are framed. Most SSDB sequences appear to be from aerobes or cyanobacteria.

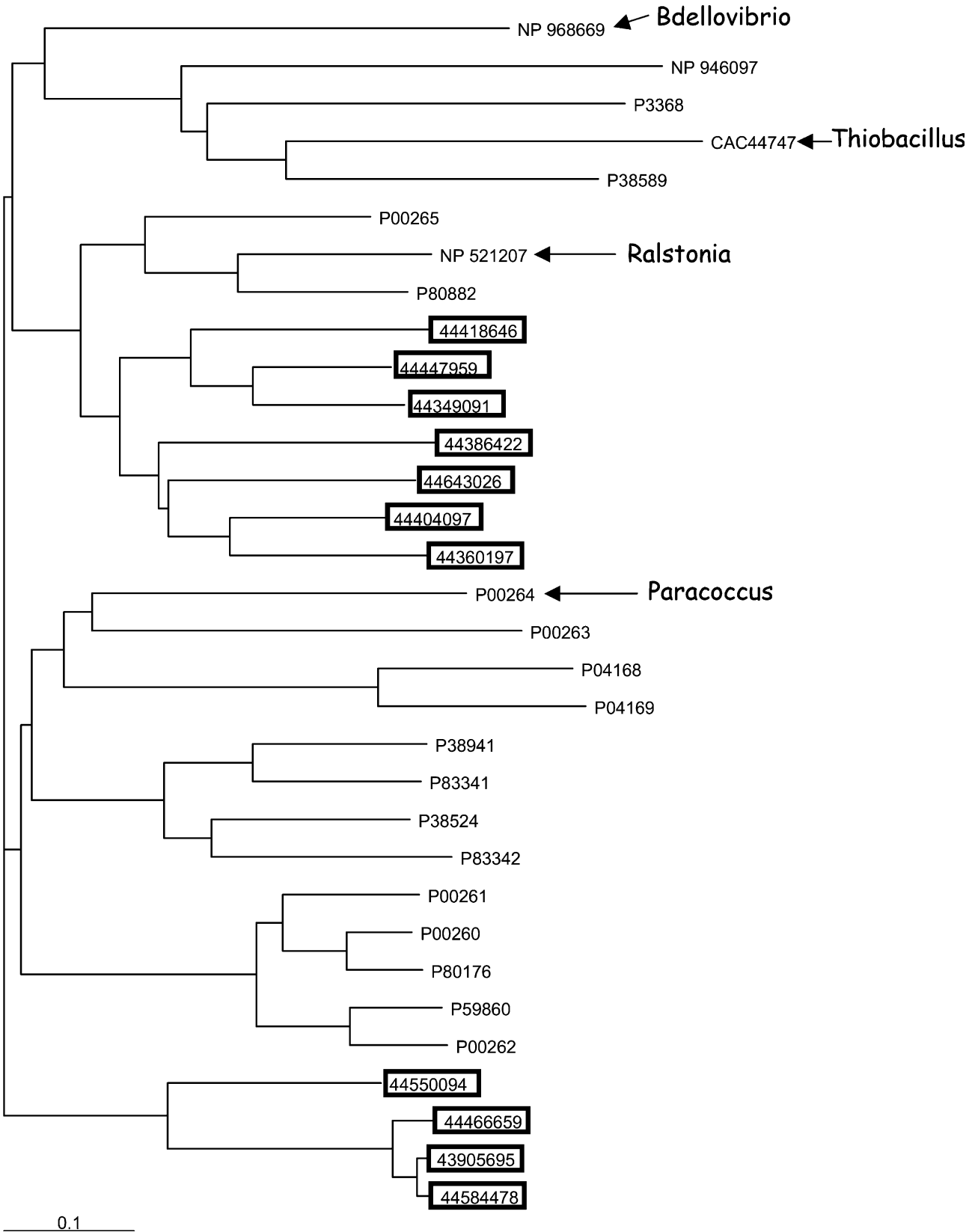

Fig. 3. Dendrogram [16] derived from sequence alignments [15] of [4Fe–4S] high-potential Fd. All sequences are given as database access numbers. Sequence entries from the SSDB [3] are composed of eight digits (no letters) and are framed. With few exceptions (indicated) the known sequences are from purple photosynthetic bacteria [7].

physico-chemical properties of the sites – surface sea water – of sample collection.

The retrieval of such an astounding sequence diversity from a nutrient-poor environment [3] opens nearly boundless perspectives for the nutrient-rich niches that are targeted for exploration by similar approaches in the near future. While the sequence searches and alignments reported here were aimed at known classes of Fe–S proteins, yet unidentified novel classes will most certainly also be found. Altogether novel Fe–S protein folds, as well as novel spatial distributions of Fe–S ligands within known folds, feature among the expected discoveries. Similar developments may be anticipated for most classes of presently known proteins. Thus, the misleading impression of a ''miraculous catch'' of sequences merely reflects that microbial and protein diversity has so far been vastly underestimated [4].

References

[1] De Long, E.F. (2002) Curr. Opin. Microbiol. 5, 520–524.
[2] Tyson, G.W., Chapman, J., Hugenholtz, P., Allen, E.E., Ram, R.J., Richardson, P.M., Solovyev, V.V., Rubin, E.M., Rokhsar, D.S. and Banfield, J.F. (2004) Nature 428, 37–43.
[3] Venter, J.C., Remington, K. and Heidelberg, J., et al. (2004) Science 304, 66–74.
[4] Falkowski, P.G. and deVargas, C. (2004) Science 304, 58–60.
[5] Beinert, H., Holm, R.H. and Münck, E. (1997) Science 277, 653–659.
[6] Meyer, J. (2001) FEBS Lett. 509, 1–5.
[7] Carter Jr., C.W. (2001) in: Handbook of Metalloproteins (Messerschmidt, A., Huber, R., Poulos, T. and Wieghardt, K., Eds.), pp. 602–609, Wiley, Chichester, UK.
[8] Meyer, J. and Moulis, J.-M. (2001) in: Handbook of Metalloproteins (Messerschmidt, A., Huber, R., Poulos, T. and Wieghardt, K., Eds.), pp. 505–517, Wiley, Chichester, UK.
[9] Martin, W. and Russell, M.J. (2003) Phil. Trans. R. Soc. Lond. B 358, 59–83.
[10] Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Nucleic Acids Res. 25, 3389–3402.
[11] Yagi, T. and Matsuno-Yagi, A. (2003) Biochemistry 42, 2266–2274.
[12] Vignais, P.M., Billoud, B. and Meyer, J. (2001) FEMS Microbiol. Rev. 25, 455–501.
[13] Rees, D.C. and Howard, J.B. (2000) Curr. Opin. Chem. Biol. 4, 559–566.
[14] Frazzon, J., Fick, J.R. and Dean, D.R. (2002) Biochem. Soc. Trans. 30, 680–685.
[15] Higgins, D., Thompson, J., Gibson, T., Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) Nucleic Acids Res. 22, 4673–4680.
[16] Page, R.D.M. (1996) Comput. Appl. Biosci. 12, 357–358.
[17] Brüggemann, H., Bäumer, S., Fricke, W.F., Wiezer, A., Liesegang, H., Decker, I., Herzberg, C., Martinez-Arias, R., Merkl, R., Henne, A. and Gottschalk, G. (2003) Proc. Natl. Acad. Sci. USA 100, 1316–1321.
[18] Bertini, I., Luchinat, C., Provenzani, A., Rosato, A. and Vasos, P.R. (2002) Proteins 46, 110–127.
[19] Meyer, J., Clay, M.D., Johnson, M.K., Stubna, A., Münck, E., Higgins, C. and Wittung-Stafshede, P. (2002) Biochemistry 41, 3096–3108.