Nucleic Acids Research, 1995, Vol. 23, No. 4 565–570

Novel protein families in archaean genomes

Christos Ouzounis*, Nikos Kyrpides[1] and Chris Sander

European Molecular Biology Laboratory, Meyerhofstrasse 1, Postfach 10.2209, D-69012 Heidelberg, Germany and [1]Institute for Molecular Biology and Biotechnology, Heraklion, Greece

Received December 7, 1994; Revised and Accepted January 9, 1995

ABSTRACT

In a quest for novel functions in archaea, all archaean hypothetical open reading frames (ORFs), as annotated in the Swiss-Prot protein sequence database, were used to search the latest databases for the identification of characterized homologues. Of the 95 hypothetical archaean ORFs, 25 were found to be homologous to another hypothetical archaean ORF, while 36 were homologous to non-archaean proteins, of which as many as 30 were homologous to a characterized protein family. Thus the level of sequence similarity in this set reaches 64%, while the level of function assignment is only 32%. Of the ORFs with predicted functions, 12 homologies are reported here for the first time and represent nine new functions and one gene duplication at an acetyl-coA synthetase locus. The novel functions include components of the transcriptional and translational apparatus, such as ribosomal proteins, modification enzymes and a translation initiation factor. In addition, new enzymes are identified in archaea, such as cobyric acid synthase, dCTP deaminase and the first archaean homologues of a new subclass of ATP binding proteins found in fungi. Finally, it is shown that the putative laminin receptor family of eukaryotes and an archaean homologue belong to the previously characterized ribosomal protein family S2 from eubacteria. From the present and previous work, the major implication is that archaea seem to have a mode of expression of genetic information rather similar to eukaryotes, while eubacteria may have proceeded into unique ways of transcription and translation. In addition, with the detection of proteins in various metabolic and genetic processes in archaea, we can further predict the presence of additional proteins involved in these processes.

INTRODUCTION

The definition of three major domains in the tree of life and the emergence of archaea as a separate group (1), distinct to and separate from eubacteria and eukaryotes has stirred much interest in the past years. Studies of deep phylogenies and the definition of a hypothetical universal common ancestor from which all branches of life may have emerged independently are based on the analysis of molecular families of extant species. In addition, there are direct implications not only for the origins of life and the definition of the common ancestor, but also for the understanding of the origins and distribution of contemporary molecular systems, such as replication, transcription and translation.

Some of the recent successes in this field were the identification of general transcription factors in archaea (2), for example TFIIB (3), TFIID (4,5) and TFIIS (6,7), RNA polymerases (8) and enzymes, such as members of the arginase (9) and methionine aminopeptidase (10) families.

As a test case for our analysis tools, enriched by new methods and larger and better sequence databases, we have analysed a set of sequences, those of hypothetical ORFs of archaea. The new functional assignments provide new insights into universal protein families present before the divergence of the prime domains into the contemporary ones.

MATERIALS AND METHODS

As a prime dataset, 95 archaean ORFs annotated as hypothetical ORFs from Swiss-Prot (11) were used as query sequences to search databases for the identification of homologues.

The sequence databases that were searched were Swiss-Prot, GenPept (translated GenBank, courtesy of Mark Gunnell), PirOnly (a subset of PIR, non-overlapping with Swiss-Prot, courtesy of Peter Rice) and TrEBML (translated EMBL, courtesy of Thure Etzold). The versions used were as of October 10, 1994. Database searches were exclusively performed using BLAST (12), using the BLOSUM62 substitution matrix and default parameters.

The database searches were controlled, parsed and visualized using GeneQuiz (13), a system for large scale sequence analysis that allows semi-automatic interactive evaluation. Multiple alignments were performed using MaxHom (14) and ClustalW (15).

Access to bibliographic databases was greatly facilitated using Nentrez (NCBI, NLM) and SRS (16). A side goal of the analysis was to perform as many actions of the process without recourse to paper, using only a workstation and the public databases available, something impossible only a few years ago (17).

RESULTS

Of the 95 hypothetical ORFs, as many as 61 had a homologue in the databases: 25 of those were similar to another hypothetical ORF in archaea and remain unassigned, while the other 36 were homologous to proteins from eubacteria and eukaryotes (Fig. 1). Of these, 30 proteins were homologous to 22 characterized families and thus their functions can be predicted. Twelve of these

* To whom correspondence should be addressed

a

Class	Number	Percent
No homology	34	36%
Homology	61	64%
Archaea	25	26%
Eubacteria only	14	15%
Eukaryotes only	10	10.5%
Universal	12	12.5%
Total	95	100%

b

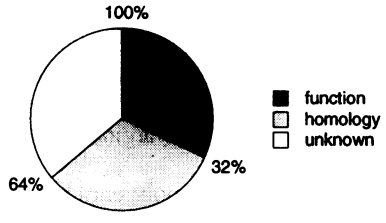

Figure 1. Summary of results and patterns in the analysis of the current set. (**a**) Numbers and percentages of proteins without homology and with homology to the three domains. Note that as many as one third of the ORFs with homology to non-archaean genomes are universal. (**b**) A pie diagram showing the levels of function prediction by homology (32%) and the detection of homology without functional assignment (another 32%). While the latter number is comparable to values obtained from the analysis of other model genomes, the function prediction is significantly lower.

proteins represent the discovery of 10 functions, nine of which are novel for the archaean domain (Table 1). The remaining 18 represent 12 functions and have recently been reported, but not adequately annotated (Table 2).

An interesting pattern emerging from the present analysis is the rather equal representation of relatives of archaean proteins in the two other major domains, eubacteria and eukaryotes. of the 61 sequences with homology to any protein, 25 were confined to the archaean domain only (all of unknown function) and the other 36 were distributed as follows: 14 were homologous to eubacteria only (of which the functions of four remain unknown), 10 were homologous to eukaryotes only (of which two are of unknown function) and finally 12 were ubiquitous (and all assigned a function) (Fig. 1). The 12 novel function assignments belonging to 10 families are presented below in some detail.

Archaea and eubacteria only

Cobyric acid synthase. The partial ORF230 downstream of the glutamine synthetase gene of *Methanococcus voltae* (18) is highly similar (45% sequence identity) to *CobQ* from *Pseudomonas denitrificans* (19) and homologous to dethiobiotin synthases (data not shown). Cobalamin synthesis is known to occur in various eubacterial species. The pathway is not fully understood, but cloning of the locus revealed the presence of five different genes, named *cobN–cobQ* and *cobW* (19). *CobQ* was identified as cobyric acid synthase (19). Therefore, the detected sequence similarity establishes the presence of an enzyme which may have a similar function in a methanogenic archaeon and enables us to predict the presence of additional genes in archaea for cobalamin biosynthesis.

Table 1. The 12 ORFs in archaea for which novel functions are predicted on the basis of sequence similarity

Query[a]	Closest homologue	*P*(N)	Predicted function
Archaea and eubacteria only			
ygln_metvo	swiss|P29932|cobq_psede	1.1e-64	Cobyric acid synthase
ylg5_desam	swiss|P28248|dcd_ecoli	1.3e-06	Deoxycytidine triphosphate deaminase
ydo3_sulso	trembl|Z21677|tmribprsa	0.0014	50S Ribosomal protein L24
yni1_mettl	swiss|P29749|mtt8_theth	0.0047	Modification methylase
Archaea and eukaryotes only			
yrp2_theac	gp|L18960|humeif4c_1	3.8e-15	Protein synthesis factor eIF-4C
yory_pyrwo	swiss|P05744|rl37_yeast	1.6e-13	60S Ribosomal protein L37
yorz_pyrwo	swiss|P15565|trm1_yeast	2.1e-07	N^2,N^2-dimethylguanosine tRNA methyltransferase
Ubiquitous families			
yac2_metso	swiss|P27095|acua_metso	1.3e-51	Acetyl-coenzyme A synthetase
yac1_metso	swiss|P27095|acua_metso	4.2e-23	Acetyl-coenzyme A synthetase
yeno_halma	swiss|P32905|nab1_yeast	2.5e-31	NAB1, putative laminin receptor,
RS2			
ypv1_mettf	trembl|L16793|spcdc18xg	8.5e-07	cdc18, ATP binding protein
ypz1_mettf	trembl|L16793|spcdc18xg	8.7e-07	cdc18, ATP binding protein

Query: Swiss-Prot name; closest homologue: unique identifier of closest homologue in databases (format: dbase\|accession number\|identifier); *P*(N): probability value from BLAST; predicted function: function of closest homologue, tentative assignment.

[a]The 12 ORFs fall into 10 functional classes, one of which has been previously found in archaea (acetyl-CoA synthetase has been identified before; upstream of the two ORFs, the closest homologue is also an archaean protein).

Table 2. The 18 ORFs in archaea for which their functions have been predicted and published recently

Query	Closest homologue	*P*(N)	Predicted function
ys91_halma	swiss|P22139|rpbx_yeast	1.1e-18	DNA-directed RNA polymerases RPB10[b]
ys92_halma	pironly|S38627|S38627	3.7e-06	DNA-directed RNA polymerases RPB6[b]
y9k_halha	pironly|S23795|S23795	3.1e-12	DNA-directed RNA polymerase H[c]
yrp1_sulac	swiss|P31815|rpoh_thece	1.1e-21	DNA-directed RNA polymerase H[c]
y14k_halha	swiss|P15738|y14k_halmo	7.6e-50	NusA transcription regulator[d]
y14k_halmo	swiss|P15739|y14k_halha	3.4e-50	NusA transcription regulator[d]
yrp1_theac	swiss|P14026|yrp7_metva	7.4e-16	NusA transcription regulator[d]
yrp3_sulac	swiss|P29157|yrpl_thece	4.2e-22	NusA transcription regulator[d]
yrp7_metva	swiss|P11523|yrp3_sulac	1.7e-17	NusA transcription regulator[d]
yrpl_thece	swiss|P11523|yrp3_sulac	2.0e-22	NusA transcription regulator[d]
yrp1_metva	swiss|P08245|ycih_ecoli	2.9e-09	Protein translation factor Sui1[e]
yrp2_sulac	swiss|P11522|rl3e_sulac[a]	1.6e-49	Ribosomal protein L30[f]
y36k_metsm	trembl|X03250|mtcomppu	1.9e-64	AIR carboxylases[g]
yhmf_metfe	swissnew|P37819|speb_strcl	2.5e-14	Agmatinase/arginase family[h]
yhsh_halma	swiss|P36132|yk18_yeast	2.5e-27	Glycoproteinase[i,j]
yrb1_halcu	swiss|P32235|gtp1_schpo	4.3e-40	GTP binding protein 1[k]
yrp8_metva	swissnew|P37742|rfk7_ecoli	0.0072	Phosphomannomutase/phosphoglucomutase family[l]
ysyn_metfe	swissnew|P38062|i2a6_rat	1.5e-11	IF2-associated Met aminopeptidase[m]

Annotation as in Table 1. References are given to the publications where the prediction has been made. The 18 ORFs fall into 12 functional classes.
[a]Renamed in the latest Swiss-Prot release. The table is split in four sections (from top to bottom) according to function: transcription, regulation, translation, metabolism. Phylogenetic distribution data (as in Table 1) is available from the authors on request.
[b]K. McKune, N. A. Woychik (1994) *J. Bacteriol.* **176**, 4754–4756.
[c]H. P. Klenk, P. Palm, F. Lottspeich, W. Zillig (1992) *Proc. Natl. Acad. Sci. USA* **89**, 407–410.
[d]T. J. Gibson, J. D. Thompson, J. Heringa (1993) *FEBS Lett.* **324**, 361–366.
[e]C. Fields, M. D. Adams (1994) *Biochem. Biophys. Res. Commun.* **198**, 288–291.
[f]A. Bairoch, B. Boeckmann (1991) *Nucleic Acids Res.* **19**, 2247–2249.
[g]D. J. Ebbole, H. Zalkin (1987) *J. Biol. Chem.* **262**, 8274–8287.
[h]C. A. Ouzounis, N. C. Kyrpides (1994) *J. Mol. Evol.* **39**, 101–104.
[i]K. M. Abdullah, R. Y. Lo, A. Mellors (1991) *J. Bacteriol.* **173**, 5597–5603.
[j]E. Arndt, C. Steffens (1992) *FEBS Lett.* **314**, 211–214.
[k]J. D. Hudson, P. G. Young (1993) *Gene* **125**, 191–193.
[l]K. Robison, W. Gilbert, G. M. Church (1994) *Nature Genet.* **7**, 205–214.
[m]J. F. Bazan, L. H. Weaver, S. L. Roderick, R. Huber, B. W. Matthews (1994) *Proc. Natl. Acad. Sci. USA* **91**, 2473–2477.

dCTP deaminase. The ORF present in the opposite strand of a locus coding for an ATP-dependent eukaryotic-like DNA ligase (ORF3) in *Desulfurolobus ambivalens* (20) is 33% identical to the dCTP deaminase from *Escherichia coli* (21). This enzyme converts dCTP to dUTP, which is then converted to dUMP and used as substrate for thymidylate synthase (21). This reaction was only known to be present in *E.coli*, while in eukaryotes dUMP is produced by dCMP deaminase (22). It is therefore unlikely that dCTP deaminase will ever be found in other eukaryotes.

50S ribosomal protein L24 (L24P family). An unidentified ORF (ORF3) between the *DPa* gene and ribosomal proteins L11, L1, L10 and L12 (in that order) of *Sulfolobus solfataricus* (23) is similar to the 50S ribosomal protein L24 (L24P family) of eubacteria (24) (38% sequence identity with the closest homologue). Ribosomal proteins known to be common only to archaea and eubacteria are very few (25) and it can be predicted that L24P may also be found in eukaryotes.

Modification methylase. A partial ORF (C-terminal part only) at the *nifH* locus of *Methanococcus thermolithotrophicus* (26) is similar to a modification methylase from *Thermus aquaticus* (27) and to the methyltransferase coded by the *bcgIA* gene from *Bacillus coagulans* (28). The sequence identity between the two enzymes is 26% and, in addition, the PROSITE pattern for N-6 adenine-specific DNA methylases (29) is present in this particular ORF (positions 91–97). The presence of modification methylases and related restriction systems has been previously confirmed in archaea (30), yet this case represents the detection of an archaean methylase of a distinct type.

Archaea and eukaryotes only

Translation initiation factor eIF-4C. An ORF at the locus coding for DNA-dependent RNA polymerase subunits H, B, A1 and A2 in *Thermoplasma acidophilum* (ORF125) (31) has 31% sequence identity with eukaryotic translation initiation factor eIF-4C (32). Translation factor eIF-4C is involved in the initiation pathway, enhancing ribosome dissociation and stabilizing the interaction of initiator tRNA with the 40S ribosomal subunit (32).

60S ribosomal protein L37 (L35AE family). The first upstream ORF (ORFy) in the antisense strand of the locus coding for glyceraldehyde-3-phosphate dehydrogenase from *Pyrococcus woesei* (33) is 34% identical to tRNA binding 60S ribosomal protein L37 from yeast (L35AE family) (34). It is unclear how many ribosomal proteins are common only between archaea and eukaryotes (35).

N^2,N^2-dimethylguanosine tRNA methyltransferase. The second partial ORF (ORFz) at the same locus as the previous case (33) is 50% identical to the protein coded by the *TRM1* gene of *Saccharomyces cerevisiae* (36). This enzyme is necessary for the N^2,N^2-dimethylguanosine modification of tRNA (36). Previous observations have already established the presence of modification mechanisms for rRNA and tRNA in archaea (37).

Archaea, eubacteria and eukaryotes

A sequential duplication of acetyl-CoA synthase? The acetyl-CoA synthase gene (*acs*) has been cloned from *Methanothrix soehngenii* (38) and was reported to be a single copy gene in this methanogenic archaeon. The two ORFs downstream of *acs* (ORF1, ORF2-partial) were reported not to show any significant similarity to proteins in the databases (38). However, the two ORFs have high sequence similarity to the upstream *acs* gene (ORF1 55% identity, ORF2 73% identity) and in consecutive order, a fact that may be interpreted as a sequencing error resulting in a frameshift. Thus, interestingly, the two ORFs downstream of *acs* in *M.soehngenii* represent a gene duplication. The functional significance of this tandem arrangement is not clear.

A 'laminin receptor'-like protein is the archaean RS2. An ORF downstream of the enolase gene (*MSG*) at the S9 locus of the archaeon *Haloarcula marismortui* (39) is 40% identical to the putative laminin receptor of eukaryotes. This was noted in the original publication, but the observation was offered without further explanations (40). In two recent reports, ribosomal proteins S2 have been found to share sequence similarities with the putative laminin receptors (41,42). Now this observation is confirmed by the discovery of a 'missing link': as shown here (Fig. 2), this family includes the archaean and eukaryotic homologues of the ribosomal protein family S2 from eubacteria (40). Profile searches (43) with the putative laminin receptors identify the eubacterial ribosomal proteins S2 unambiguously. The sequences are of comparable length, but seem to evolve rapidly, thus making the identification of homology possible only by using sensitive and selective methods of similarity detection.

Two CDC18 homologues. Two homologous ORFs (ORF1) coded by two plasmids (pfZ1 and pfV1) in *M.thermoformicicum* (30) are in turn 21% identical to a recently identified protein from the yeast *Schizosaccharomyces pombe*, named CDC18$^+$ (44). This protein has been identified as a control factor that allows cells to enter the S phase of cell division, preventing mitosis until the S phase is completed (44). Another protein from the yeast *Saccharomyces cerevisiae*, CDC6, which controls replication and nuclear division (45), also belongs to this family (Fig. 3). It is not clear what the exact function of the archaean proteins is for the replication of the plasmids. This protein family contains the typical motifs A and B present in purine NTP binding protein families (46) (Fig. 3) and it seems to be distantly related to a large family of ATP binding proteins, which includes FtsH, PAS1 and CDC48 (data not shown) (47).

Archaea only

There still remain in the dataset 25 families of ORFs of unknown function, unique to archaea. Their functions cannot be currently predicted, but the presence of homologues in various related species implies that these ORFs represent genuine proteins. Sensitive methods such as pattern and profile searches (43,48) failed to detect homology to proteins of known function.

DISCUSSION

Some general patterns in sequence analysis emerging from genome projects (17,49) were reproduced in this effort. In particular, the level of detection for sequence similarity was 64% (61 of 95 sequences have a homologue in the databases, 34 sequences do not), while the function assignment was rather low, at 32% (30 sequences have a homologue of known function). Yet 30 interesting cases were found where function can be assigned to the corresponding ORFs, of which 12 represent 10 new functions in archaea. Of these novel protein families discovered, some are not surprising (e.g. ribosomal protein L37), while some are most remarkable (e.g. CDC18 family).

Another implication of this particular study is the realization of the value of updating datasets of interest regularly. With better tools and larger and richer databases, the task of homology identification has become easier and the probability of detecting homologies to proteins of known function has reached an average level of ~50%. The challenge of collecting all hypothetical ORFs from *E.coli*, yeast, *C.elegans* or any other model organism and submitting them to regular database searches to update our knowledge about their possible function by homology is immense.

The biological implications of the present and other recent findings reinforce the idea that archaea are indeed the domain which holds the secrets about the origin of the cell in general and the eukaryotic domain in particular. The sharing of components from both eubacteria and eukaryotes is most puzzling.

The most convincing demonstration of this duality emerges from the study of genomic organization in relationship to gene expression. Curiously enough, archaea display the typical eubacterial mode of gene organization in operons, while their various levels of gene expression seem to be related to eukaryotes (e.g. transcription and translation factors). Whether the eukaryotic genome composition is a primitive characteristic or a derived one, amplified by long evolution and natural selection, cannot be decided at present.

Even more enigmatic is the observation that even within archaean operons, there are certain discrepancies as far as sequence similarities with the other two domains are concerned.

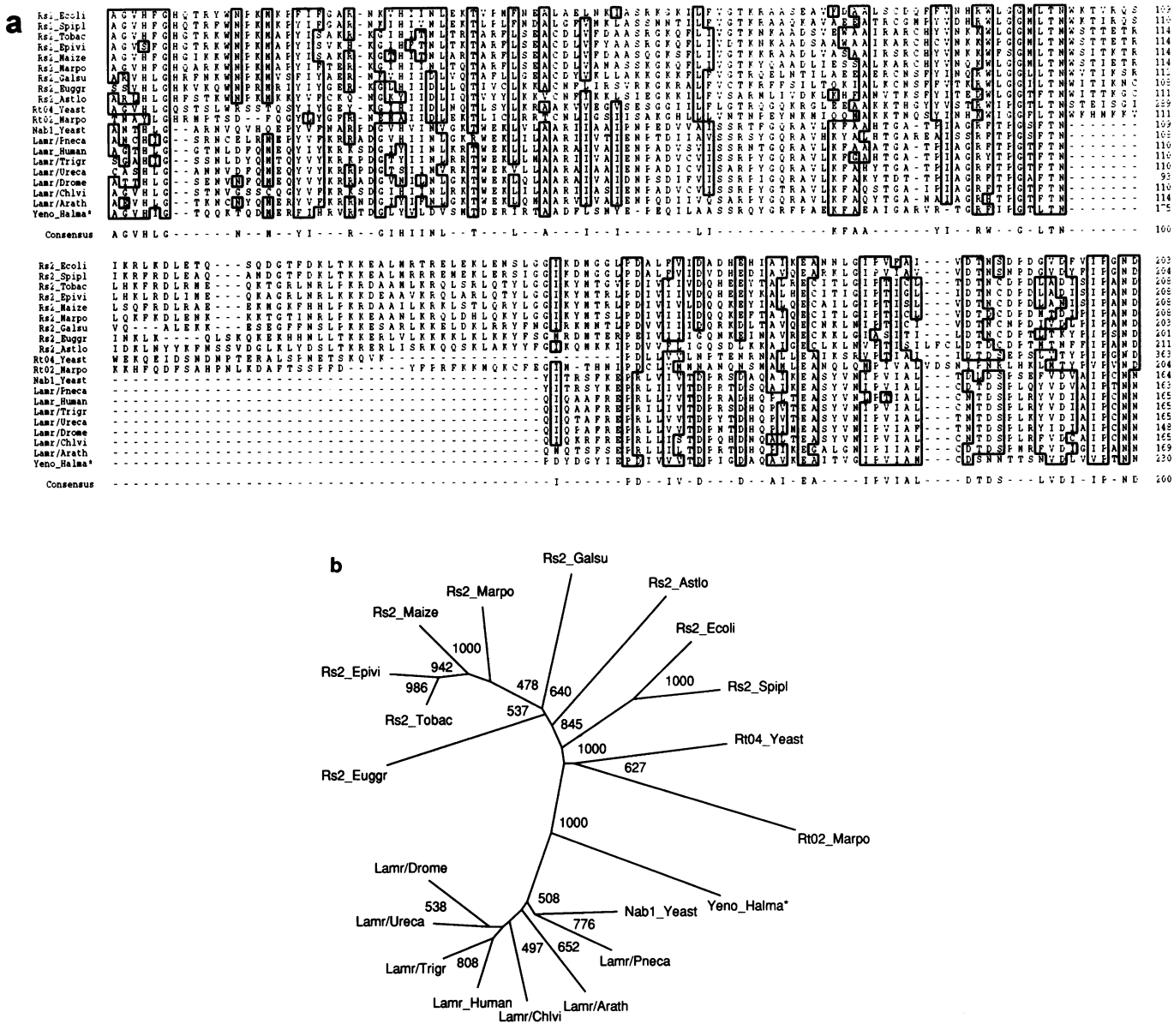

Figure 2. Alignment (**a**) and tree (**b**) of the eubacterial ribosomal family S2, the putative laminin receptor from eukaryotes (which is predicted to be the eukaryotic homologue of ribosomal protein S2) and the ORF from *Haloarcula marismortui* (39), marked with an asterisk. The tree displays bootstrap values from 1000 simulations. It is evident that eubacterial S2 shows a much greater variability than the eukaryotic homologues. The archaean sequence is reliably placed (bootstrap value 1000) at a location equidistant from the other two domains. The alignment figure was generated by PrettyPlot (written by Peter Rice) and the tree by TreeTool (University of Illinois). Sequence identifiers are from Swiss-Prot; provisional names are marked accordingly (name/species). Species names conventions as in Swiss-Prot. Sequence positions are marked.

First, the S9 operon of *Haloarcula marismortui* (39), which with the present work has been completely characterized, contains the genes for the ribosomal proteins L29, L13, S9, the two subunits of the eukaryotic RNA polymerase RPB10 and RPB6 (Table 2), an enolase gene and the ribosomal protein S2. This operon is an example where some of its proteins only have eukaryotic homologues, others only eubacterial ones and some homologues to both domains. Secondly, the locus coding for DNA-dependent RNA polymerase subunits H, B, A1, A2 and eIF-4C in *Thermoplasma acidophilum* (31) contains mostly ORFs with unique or highest similarity to eukaryotes, yet its structure is reminiscent of eubacterial operons.

Finally, it is expected that the more our knowledge of archaean sequences increases, the more evidence of a mosaic genome organization will be obtained. An interesting goal is the collection and documentation of proteins common to all three domains, allowing us to speculate about the origins and nature of the universal common ancestor (37), a hypothetical genetic or pre-cellular entity present before the divergence of the three prime domains of life. It seems that the common intersection of

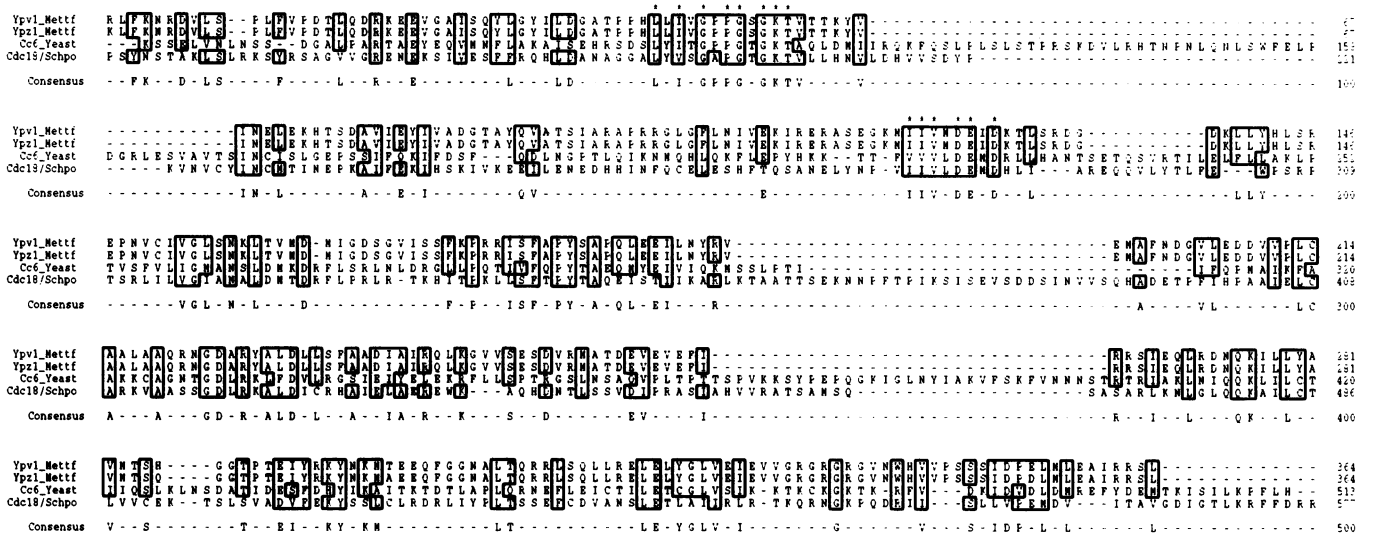

Figure 3. Alignment of the two homologous ORFs (ORF1) coded by plasmids (pfZ1 and pfV1) in *M.thermoformicicum* (30) with their closest relatives, from fungi. The ATP binding residues and the conserved hydrophobic positions preceding them are marked with an asterisk. Annotation as in Figure 2. Figure also generated by PrettyPlot.

the protein sets present in all organisms today contains only some components of the gene expression machinery, such as ribosomal proteins, as well as the series of enzymes participating in basic metabolic pathways.

REFERENCES

1 Woese, C.R., Kandler, O., Wheelis, M.L. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 4576–4579.
2 Klenk, H.-P., Doolittle, W.F. (1994) *Curr. Biol.* **4**, 920–922.
3 Ouzounis, C., Sander, C. (1992) *Cell* **71**, 189–190.
4 Marsh, T.L., Reich, C.I., Whitelock, R.B., Olsen, G.J. (1994) *Proc. Natl. Acad. Sci. USA* **91**, 4180–4184.
5 Rowlands, T., Baumann, P., Jackson, S.P. (1994) *Science* **264**, 1326–1329.
6 Langer, D., Zillig, W. (1993) *Nucleic Acids Res.* **21**, 2251–2251.
7 Kaine, B.P., Mehr, I.J., Woese, C.R. (1994) *Proc. Natl. Acad. Sci. USA* **91**, 3854–3856.
8 Huet, J., Schnabel, R., Sentenac, A., Zillig, W. (1983) *EMBO J.* **2**, 1291–1294.
9 Ouzounis, C.A., Kyrpides, N.C. (1994) *J. Mol. Evol.* **39**, 101–104.
10 Bazan, J.F., Weaver, L.H., Roderick, S.L., Huber, R., Matthews, B.W. (1994) *Proc. Natl. Acad. Sci. USA* **91**, 2473–2477.
11 Bairoch, A., Boeckmann, B. (1991) *Nucleic Acids Res.* **19**, 2247–2249.
12 Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J. (1990) *J. Mol. Biol.* **215**, 403–410.
13 Scharf, M., Schneider, R., Casari, G., Bork, P., Valencia, A., Ouzounis, C., Sander, C. (1994) In: *Second International Conference on Intelligent Systems for Molecular Biology*, Altman, R., Brutlag, D., Karp, P., Lathrop, R., Searls, D. (eds). AAAI Press, Stanford, CA, pp. 348–353.
14 Sander, C., Schneider, R. (1991) *Proteins* **9**, 56–68.
15 Higgins, D.G., Bleasby, A.J., Fuchs, R. (1992) *Comput. Appl. Biosci.* **8**, 189–191.
16 Etzold, T., Argos, P. (1993) *Comput. Appl. Biosci.* **9**, 49–57.
17 Bork, P., Ouzounis, C., Sander, C., Scharf, M., Schneider, R., Sonnhammer, E. (1992) *Protein Sci.* **1**, 1677–1690.
18 Possot, O., Sibold, L., Aubert, J.P. (1989) *Res. Microbiol.* **140**, 355–371.
19 Crouzet, J., Levy-Schil, S., Cameron, B., Cauchois, L., Rigault, S., Rouyez, M.C., Blanche, F., Debussche, L., Thibaut, D. (1991) *J. Bacteriol.* **173**, 6074–6087.
20 Kletzin, A. (1992) *Nucleic Acids Res.* **20**, 5389–5396.
21 Wang, L., Weiss, B. (1992) *J. Bacteriol.* **174**, 5647–5653.
22 Mathews, C.K., van Holde, K.E. (1990) *Biochemistry*. Benjamin–Cummings, Redwood City, CA, pp. 763–764.
23 Ramírez, C., Matheson, A.T. (1991) *Mol. Microbiol.* **5**, 1687–1693.
24 Ohama, T., Muto, A., Osawa, S. (1989) *J. Mol. Evol.* **29**, 381–395.
25 Otaka, E., Hashimoto, T., Mizuta, K. (1993) *Protein Sequence Data Anal.* **5**, 285–300.
26 Souillard, N., Magot, M., Possot, O., Sibold, L. (1988) *J. Mol. Evol.* **27**, 65–76.
27 Barany, F., Danzitz, M., Zebala, J., Mayer, A. (1992) *Gene* **112**, 3–12.
28 Kong, H., Morgan, R.D., Maunus, R.E., Schildkraut, I. (1994) *Nucleic Acids Res.* **21**, 987–991.
29 Bairoch, A. (1992) *Nucleic Acids Res.* **20**, 2013–2018.
30 Nölling, J., van Eeden, J.M., Eggen, R.I.L., de Vos, W.M. (1992) *Nucleic Acids Res.* **20**, 6501–6507.
31 Klenk, H.-P., Renner, O., Schwass, V., Zillig, W. (1992) *Nucleic Acids Res.* **20**, 5226–5226.
32 Dever, T.E., Wei, C.L., Benkowski, L.A., Browning, K., Merrick, W.C., Hershey, J.W. (1994) *J. Biol. Chem.* **269**, 3212–3218.
33 Zwickl, P., Fabry, S., Bogedain, C., Haas, A., Hensel, R. (1990) *J. Bacteriol.* **172**, 4329–4338.
34 Santangelo, G.M., Tornow, J., McLaughlin, C.S., Moldave, K. (1991) *Gene* **105**, 137–138.
35 Otaka, E., Hashimoto, T., Mizuta, K., Suzuki, K. (1993) *Protein Sequence Data Anal.* **5**, 301–313.
36 Ellis, S.R., Hopper, A.K., Martin, N.C. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 5172–5176.
37 Woese, C.R. (1987) *Microbiol. Rev.* **51**, 221–271.
38 Eggen, R.I.L., Geerling, A.C.M., Boshoven, A.B.P., de Vos, W.M. (1991) *J. Bacteriol.* **173**, 6383–6389.
39 Krömer, W.J., Arndt, E. (1991) *J. Biol. Chem.* **266**, 24573–24579.
40 An, G., Bendiak, D.S., Mamelak, L.A., Friesen, J.D. (1981) *Nucleic Acids Res.* **9**, 4163–4172.
41 Davis, S.C., Tzagoloff, A., Ellis, S.R. (1992) *J. Biol. Chem.* **267**, 5508–5514.
42 Tohgo, A., Takasawa, S., Munakata, H., Yonekura, H., Hayashi, N., Okamoto, H. (1994) *FEBS Lett.* **340**, 133–138.
43 Gribskov, M., Luethy, R., Eisenberg, D. (1990) *Methods Enzymol.* **183**, 146–159.
44 Kelly, T.J., Martin, G.S., Forsburg, S.L., Stephen, R.J., Russo, A., Nurse, P. (1993) *Cell* **74**, 371–382.
45 Bueno, A., Russell, P. (1992) *EMBO J.* **11**, 2167–2176.
46 Ouzounis, C.A., Blencowe, B.J. (1991) *Nucleic Acids Res.* **19**, 6953–6953.
47 Gorbalenya, A.E., Koonin, E.V. (1993) *Curr. Opin. Struct. Biol.* **3**, 419–429.
48 Hodgman, T.C. (1989) *Comput. Appl. Biosci.* **5**, 1–13.
49 Bork, P., Ouzounis, C., Sander, C. (1994) *Curr. Opin. Struct. Biol.* **4**, 393–403.