Proc. Natl. Acad. Sci. USA
Vol. 96, pp. 8545–8550, July 1999
Evolution

Transcription in Archaea

NIKOS C. KYRPIDES* AND CHRISTOS A. OUZOUNIS†‡

*Department of Microbiology, University of Illinois at Urbana-Champaign, B103 Chemistry and Life Sciences, MC 110, 407 South Goodwin Avenue, Urbana, IL 61801; and †Computational Genomics Group, Research Programme, The European Bioinformatics Institute, European Molecular Biology Laboratory, Cambridge Outstation, Wellcome Trust Genome Campus, Cambridge CB10 1SD, United Kingdom

Edited by Norman R. Pace, University of California, Berkeley, CA, and approved May 11, 1999 (received for review December 21, 1998)

ABSTRACT **Using the sequences of all the known transcription-associated proteins from Bacteria and Eucarya (a total of 4,147), we have identified their homologous counterparts in the four complete archaeal genomes. Through extensive sequence comparisons, we establish the presence of 280 predicted transcription factors or transcription-associated proteins in the four archaeal genomes, of which 168 have homologs only in Bacteria, 51 have homologs only in Eucarya, and the remaining 61 have homologs in both phylogenetic domains. Although bacterial and eukaryotic transcription have very few factors in common, each exclusively shares a significantly greater number with the Archaea, especially the Bacteria. This last fact contrasts with the obvious close relationship between the archaeal and eukaryotic transcription mechanisms *per se*, and in particular, basic transcription initiation. We interpret these results to mean that the archaeal transcription system has retained more ancestral characteristics than have the transcription mechanisms in either of the other two domains.**

Although homologous in their most basic componentry, the transcription machineries in Bacteria and Eucarya are highly diverged from one another, each having a variety of domain-specific elements (1). This divergence can be seen to some extent in terms of the transcription complex itself but more in terms of transcription initiation and regulation, which share essentially nothing in the two cases.

The bacterial RNA polymerase identifies promoters with the aid of σ-factors, which are first bound to the polymerase molecule and then facilitate promoter recognition (2). In Eucarya, the general transcription initiation factors mediate promoter recognition and guide RNA polymerase into the preinitiation complex (3, 4). In Bacteria, the specialized σ-subunits confer recognition and specificity to freely accessible promoters. In Eucarya, nucleosome structures block access to promoters, which the transcription factors such as TATA box-binding protein (TBP) and transcription factor (TF)IIB (5) then overcome. Although the eukaryotic preinitiation complexes permit multiple rounds of transcription, the bacterial σ-factors, attached (transiently) to the polymerase itself, function only once before recycling (5).

These mechanisms are effected by very different sets of molecules. In Bacteria, there are a large number of activator/repressor systems (6) acting together with the general σ^{70} transcription factors and the RNA polymerase holoenzyme for promoter recognition and activation (e.g., see refs. 7 and 8, and refs. therein). In Eucarya, the three RNA polymerases (I, II, and III) are assisted by the TBP, TBP-associated factors (9–12), and a large variety of regulators. There are only a few common elements across these domains, including the bacterial RNA polymerase core enzyme (13). For activation of transcription, the common elements between the two domains are even more scarce, with the cold-shock domain (14) being one of the few examples.

Following the discovery of Archaea as the third primary phylogenetic domain (15, 16), the archaeal RNA polymerase core enzyme was found to have a complexity similar to that of the Eucarya (consisting of up to 15 components) (17). Subsequently, the sequence similarity between the large (universal) subunits of archaeal and eukaryotic polymerases was demonstrated (18). This discovery was followed by the first unambiguous identification of transcription factor TFIIB in an archaeon, *Pyrococcus woesei* (19). Since then, we have witnessed a growing body of evidence confirming the presence of key eukaryotic-type transcription initiation factors in Archaea (5, 20). Therefore, the prevailing view has become that Archaea and Eucarya share a transcription machinery that is very different from that of Bacteria (5, 21, 22). The presence of bacterial-type regulators in Archaea (23–25), however, suggests that the evolutionary picture is somewhat more complex (and interesting) than such a simple formulation would suggest.

In this work, we quantify the phylogenetic extent of the predicted transcription-associated proteins in Archaea. Our analysis is based on the complete genome sequences of four archaeal species: *Methanococcus jannaschii* (MJ) (26), *Archaeoglobus fulgidus* (AF) (27), *Methanobacterium thermoautotrophicum* (MTH) (28), and *Pyrococcus horikoshii* (PH) (29). The above sequences were compared against the full range of over 4,000 known, annotated transcription-associated proteins in the databases. The results strongly support the notion that archaeal transcription is the least derived of the three types and indicate how the transcriptional apparatus may have evolved.

MATERIALS AND METHODS

All protein sequences from SWISS-PROT version 35 (March 1998) with the keyword "transcription" and TREMBL (May 1998) with the word "transcription" in the description line were extracted by using SRS (30). From a total of 4,147 sequences, 1,444 were bacterial (1,244 in SWISS-PROT and 200 in TREMBL) and 2,703 were eukaryotic (1,925 in SWISS-PROT and 778 in TREMBL). All proteins were compared against the four complete archaeal genomes (8,104 sequences, 2,255,809 residues), by using BLAST(P) (31), after correcting for composition bias (C.A.O., unpublished data). The same analysis was performed between the bacterial and the eukaryotic sequences (1,444 bacterial sequences were compared against 2,703 eukaryotic ones). Homologs with p values $<1 \times 10^{-6}$ were extracted, and manual annotation for all 5,591 runs eliminated false positives in (*i*) annotation (not transcription-associated query sequences) or (*ii*) homology (transcription-associated query sequences but spurious hits in Archaea). Given the extent of annotation and the homology relationships between transcription-associated proteins, the false negatives in annotation (not identified by keyword or description lines) appear to be minimal [e.g., eukaryotic queries form a superset of the TRANSFAC database (32)]. In Tables 1–3, italics refer to ORFs not detected at the given threshold by the query sequences, yet they identify at least one member of the family with the same criteria; parentheses represent the number of corresponding protein families. Asterisks signify that the corresponding ORFs map to a single protein and are not necessarily homologous. Individual

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

PNAS is available online at www.pnas.org.

This paper was submitted directly (Track II) to the *Proceedings* office.

‡To whom reprint requests should be addressed. e-mail: ouzounis@ebi.ac.uk.

references are too numerous to be included. All results, along with additional material, are available at the http://www.ebi.ac.uk/research/transcription/machinery/, and comments and corrections are welcome at transcription@ebi.ac.uk.

RESULTS AND DISCUSSION

Transcription-Related Proteins in Archaea. In total, we have identified 280 homologs of transcription-related proteins in the four archaeal genomes (Tables 1–3). The principal result of this analysis is the abundance of bacterial-type transcription-related proteins in Archaea. Of the 280 proteins, 168 are found elsewhere only in Bacteria, whereas 51 are associated otherwise only with Eucarya and 61 are universally distributed proteins (Fig. 1). Despite the well known difficulties of limited characterization for the full genomes (33), with only half of the proteins having a predicted function and some paralogy within the four species, it becomes clear that archaeal transcription is not solely similar to the eukaryotic process but instead bears elements present in Bacteria, Eucarya, or both.

Bacterial-Type Transcription-Related Proteins in Archaea. Archaeal transcription-associated proteins that have only bacterial homologs belong to a number of well known families, only two of which—AsnC/Lrp (23) and NusA (34)—had been identified in Archaea before the advent of genome projects. All of the archaeal genomes contain members of these two families, with the AsnC/Lrp being the most abundant in *P. horikoshii* (Table 1).

The newly identified homologs of bacterial factors include "atypical" activators involved in heavy metal-dependent regulation, such as ArsR/CadC (arsenical and cadmium), Fur (iron), ModE (molybdenum), and MerR (mercury). In addition, homologs to the following regulatory families are identified: LysR, TetR (tetracycline-inducible repressor), HypF (hydrogenase regulator), PhoU (phosphate transport regulator), NagC/XylR (repressor of nagE–BACD and xylose-utilization operons), DtxR (diphtheria toxin repressor), DegT/DnrJ/EryC1 (pleiotropic sensory transduction), MarR (multiple antibiotic-resistance operon repressor), Xre (PBSX repressor), MoxR (methanol dehydrogenase regulator), PspC (phage shock protein operon activator), PurR (purine nucleotide synthesis repressor–LacI family), and RpiR (RpiB gene repressor) (Table 1) (35). Three other families involved in transcription are the helix–turn–helix-containing phage integrase family, arylsulfatase activator (36), and a protease/sporulation regulatory protein PAI (Table 1). The presence of 63 sensor kinase-response regulators (two-component regulatory systems) (37) in *A. fulgidus*, *M. thermoautotrophicum* and *P. horikoshii* (three members being most similar to CheA, CheB, and CheY), but not *M. jannaschii*, has been previously noted (27). Most of these archaeal homologs are characterized here, whereas they have been previously classified as regulatory or hypothetical proteins (26, 28).

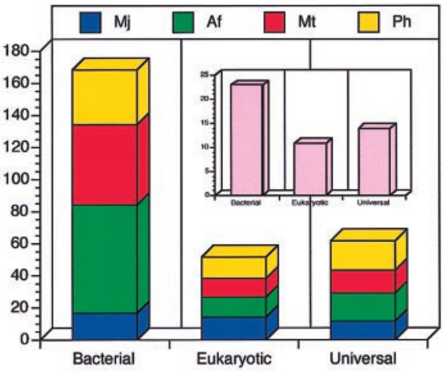

FIG. 1. Distribution of bacterial-, eukaryotic-, and universal transcription-associated homologs in the four complete archaeal genomes of *M. jannaschii* (blue), *A. fulgidus* (green), *M. thermoautotrophicum* (red), and *P. horikoshii* (yellow). A two-way ANOVA 3 (domain) × 4 (species) (df = 12–1 = 11) for normalized genome compositions (data not shown) of the transcription-associated homologs listed here, suggests that the variance arises mainly from the domain differences ($F^{d}_{2,11} = 7.76 > F_{2,11} = 7.21$) and not the species differences ($F^{s}_{3,11} = 0.73 < F_{3,11} = 6.22$) at 99% significance level. The *inset* represents the distribution of transcription-associated protein families. For ORF identifiers and species distribution of particular families, see Tables 1–3.

In total, these archaeal genomes contain homologs for 168 transcription-associated proteins found elsewhere only in Bacteria. These fall into 23 protein families, with considerable variation across the four archaeal species (Table 1). There are seven families that are found in all four genomes, two of which (NusA and HypF) have exactly one member per species and the other five having at least one member per species (AsnC/Lrp, PbsX, ArsR, DtxR, and PAI).

From the remaining 16 families, there are six cases (Fur, MerR, MarR, NagC, PspC, PurR) where the family is present in only one of the four genomes (Table 1). It should be noted that some of these cases might be the result of horizontal transfer events between Bacteria and Archaea. Progressively, as more genome data become available, it is possible that a number of any of the archaeal/bacterial transcription families listed above may eventually be identified also in Eucarya and thus classified as universal.

Apart from the 168 homologs, there exist an additional large number of proteins containing helix–turn–helix or other short domains, characteristic of various bacterial-type transcription regulator families (data not shown) but without a specific functional assignment (http://geta.life.vivc.edu/~nikos/mjannotations.html). However, these cannot be readily classified by using strict significance criteria and therefore are not included in the present analysis. For instance, the prior identification of the DNA-binding domains of σ^{70} transcription factors in *M. jannaschii* (24) is now confirmed with additional members in all genomes (data not shown). These observations are consistent with parallel work that underlines the bacterial-like genome properties of *M. jannaschii* (38).

Finally, the absence of certain bacterial transcription-associated proteins from Archaea is notable, for example σ^{54} factors (39), MetJ, NusB, and the Rho terminator. However, the possibility that some of these factors may be identified in other archaeal genomes in the future cannot be ruled out.

Eukaryotic-Type Transcription-Related Proteins in Archaea. The archaeal homologs of the factors confined otherwise to Eucarya are mainly those previously identified (Table 2). These are (*i*) the basic initiation factors TFIIB (19) and TFIID (20); (*ii*) eight "small" subunits of the eukaryotic RNA polymerase itself (5, 40, 41); and (*iii*) the archaeal histone family, which contains the core histone fold also found in the eukaryotic CAAT-binding factor subunits A (42) and C (http://www.ebi.ac.uk/~ouzonis/cbfc.html). It should be emphasized that this structural motif is absent in Bacteria (43).

In total, these proteins amount to 51 homologs of transcription-associated proteins found elsewhere only in Eucarya, which belong to 11 families. Virtually all of them have a similar distribution in all four archaeal genomes. In contrast to the archaeal/bacterial factors, the distribution of archaeal/eukaryotic factors within the four species is relatively homogeneous (Table 2). Here, every archaeal genome has at least one homolog for each of the protein families reported, with the exception of the RNA polymerase subunit RPB12, apparently present only in *A. fulgidus* and *P. horikoshii* (Table 2). There are eight families with exactly one homolog per genome, and the remaining two (TFIIB and archaeal histones) have at least one instance of a paralogous gene pair (Table 2).

It is remarkable that from the five small subunits shared by the three eukaryotic RNA polymerases (RPB5, RPB6, RPB8, RPB10, and RPB12) (12), only one of them (RPB8) is not present in any of the archaeal genomes, whereas RPB12 is found only in the genomes of the two nonmethanogens (Table 2). Moreover, of the three subunits unique to the eukaryotic RNA polymerase II (RPB4, RPB7 and RPB9) (12), two are found in (all

Table 1. Transcription-associated proteins in the four complete archaeal genomes with homologs present only in Bacteria

Protein family	*M. jannaschii*	*A. fulgidus*	*M. thermoautotrophicum*	*P. horikoshii*
AsnC/Lrp transcriptional activators	MJ0151 MJ0723	*AF0439* AF0474 AF0584 *AF1121* AF1148 AF1404 AF1448 AF1622 AF1723 AF1743	MTH1193	PH0045|PHBE036 PH0061|PHBE020 PH0140|PHBN040 PH1013|PHAQ008 PH1045|PHAJ008 PH1055|PHAI003 PH1519|PHCV018 PH1592|PHBQ026 PH1692|PHAV003 PH1916|PHBT016 PHS023|PHS045|
PbsX (XRE) family of transcriptional regulators	MJ0272	AF1627 AF1793	MTH659 MTH700 MTH1328	*PH0803|PHCH016* PH1748|PHAM003
Methanol dehydrogenase regulator MoxR		AF2425	MTH1814	PH0329|PHAY001 PH0385|PHAZ007 PH0776|PHC1011
Transcription repressors, "phase" integrase family	MJ0367		MTH893	PH1826|PHCY045
TetR/AcrR transcriptional regulators		AF1817	MTH1063 MTH1787	
LysR transcriptional activators	MJ0300 MJ1120	AF2127	MTH1545	
Mod operon regulation and molybdenym transporter ModE		*AF1022*	MTH1470	
Ferric uptake regulation, global negative regulator Fur		AF2232		
MerR transcriptional regulators		AF0673		
ArsR transcriptional regulators	MJ1325 MJ1553	AF1270 *AF1298* AF1697 AF2136	MTH899 MTH1795	PH0062|PHBE019 PH1101|PHCM034 PH1744|PHAM006 PH1930|PHBT030 *PH1932|PHBT032*
MarR transcriptional regulators			MTH313	
Transcription termination factor NusA	*MJ1045*	AF1891	*MTH1054*	*PH1543|PHCB022*
NagC/XyIR transcriptional regulators		AF1968		
Diphteria toxin gene iron-binding repressor DtxR	MJ0568	AF0245 AF1785 AF1984 AF2395	MTH214 MTH936	PH1163|PHBU004
HypF transcriptional regulators	MJ0713	AF1366	MTH1287	PH0897|PHAK022
PhoU regulators	MJ1009 *MJ1011*	AF1360	MTH1732 MTH1734	
Phage shock protein C (PspC) transcriptional regulator				PH1080|PHCM013
Pur operon repressor PurR				PH1691|PHAV004
RpiR regulatory protein involved in RpiB gene repression			MTH1546	PH1938|PHBT038
Arylsulfatase activator	MJ0907	AF2204	MTH114	
DegT/DnrJ family transcription regulators	MJ1066		MTH334 MTH1188	
Protease synthase & sporulation negative regulatory protein PAI family	MJ1207	AF0521 AF0739	MTH336 *MTH999*	*PH0296|PHBL005 PH1933|PHBT033*
Sensory transduction regulatory protein superfamily		AF0021 AF0208 AF0277 AF0410 AF0448 AF0449 AF0450 *AF0579* AF0770 AF0893 *AF1035 AF1036* AF1040 AF1041 AF1042 AF1045 AF1063 AF1184 AF1256 AF1384 AF1452 AF1467 AF1472 AF1473 AF1483 AF1515 AF1620 AF1639 AF1721 AF1898 AF2032 AF2109 AF2249 AF2419 AF2420	*MTH123 MTH174 MTH292* MTH356 *MTH360* MTH440 MTH444 MTH445 MTH446 MTH447 MTH457 *MTH459* *MTH468* MTH548 MTH549 MTH619 MTH786 MTH823 MTH901 MTH902 *MTH985 MTH1124 MTH1260* MTH1607 MTH1764	PH0482|PHBH003 PH0483|PHBH002 PH0484|PHBG001
Sum 168 (23)	16 (12)	68 (17)	50 (18)	34 (13)

Table 2. Transcription-associated proteins in the four complete archaeal genomes with homologs present only in Eukarya

Protein Family	*M. jannaschii*	*A. fulgidus*	*M. thermoautotrophicum*	*P. horikoshii*
Core histone fold (histone/CBF-A/CBF-C families)	MJ0168 MJ0932 *MJ1258* *MJECL17* MJECL29	AF0337 AF1493	MTH254 MTH821 MTH1696	*PHS046| PHS051|*
RNA polymerase subunits RPB3/RPC5 [RpoD]	MJ0192	AF2282	MTH37	PH1637|PHLE020
RNA polymerase subunits RPB5/XAP4 [RpoH]	MJ1039	AF1885	MTH1048	PHS044|
RNA polymerase subunits RPB6 [RpoK]	MJ0197	AF1131	MTH42	PH-orf: + 1434542-1450104
RNA polymerase subunits RPB10 [RpoN]	MJ0196	AF1130	MTH40	PH1632|PHLE015
RNA polymerase subunits RPB11/RPC19 [RpoL]	MJ0387	AF0207	MTH1317	PH-orf: −37346_67391
RNA polymerase subunits RPB7/RPCY [RpoE1]	MJ0397	AF1117	MTH264	PH1908|PHBT008
RNA polymerase subunits (RPB12)		AF0056		PHS056|
RNA polymerase subunits (RPA12/RPB9), TFIIS [RpoM]	MJ1148	AF1235	MTH1314	PH0664|PHLA007
TFIID (TBP)	MJ0507	AF0373	MTH1627	PH1009|PHAQ004
TFIIB	MJ0782	AF1299	MTH885	PH0864|PHAL038, PH1482|PHCC031
Sum 51 (11)	14 (10)	12 (11)	12 (10)	13 (11)

of) the archaeal genomes (RPB7/RpoE and RPB9/RpoM) (Table 2).

Although some similarity exists between the archaeal/eukaryotic RNA polymerase subunit (RpoD/RPB3) and the bacterial RNA polymerase subunit-α (RpoA) (5), it is clear that the bacterial version is highly modified *vis a vis* the archaeal and eukaryotic versions (which are quite alike). Since this distant relation scores below the threshold used in the present analysis, we have chosen not to classify this protein as universal (Table 2).

The eukaryotic TBP-interacting protein TIP49 (44), has a definite counterpart in *A. fulgidus* (AF1813) and *P. horikoshii* (PH1804|PHCY023) genomes (not shown in Table 2, because it is not annotated as transcription-associated protein; see *Materials and Methods*).

Finally, the eukaryotic transcription factor TFIIE-α may also be present in Archaea (omitted from Table 2, similarity below threshold). In each of the four archaeal genomes, the protein has a single counterpart (MJ0777, MTH1669, AF0757, and PH0619|PHAE005) that exhibits similarity to the N-terminal region of TFIIE-α (below threshold). The C-terminal region of this protein is not present in these archaeal genomes.

In addition to the well defined eukaryotic-type transcription factors described above, there exist a number of putative metal-binding, zinc-finger-like motifs (of the C_2C_2 type). Similar to the case of the helix–turn–helix archaeal proteins, these proteins cannot be readily classified functionally, and therefore are not included in the present analysis.

The absence in the Archaea of the eukaryotic-type transcription factor domains, such as the MADS box or homeoboxes (45) underscores the fact that only the basic transcription initiation machinery is shared between Archaea and Eucarya (1, 5). The present analysis confirms that no eukaryotic regulators are present in any completely sequenced archaeal genome.

Comparison of Eukaryotic- Versus Bacterial-Type Transcription Factors. In total, the comparison of eukaryotic- versus bacterial-type transcription factors in Archaea reveals some interesting contrasting patterns: (*i*) the number of the bacterial-type transcription-associated proteins in the four archaeal genomes is significantly larger; (*ii*) the distribution of eukaryotic-type archaeal factors within the four archaeal genomes is relatively homogeneous, with 10 of 11 families present in all of them (Table 2), in contrast to the bacterial-type archaeal factors where only 7 of 23 families have at least one member per genome (Table 1); and (*iii*) whereas the eukaryotic-type archaeal factors are RNA polymerase subunits or initiation factors, the bacterial-type archaeal factors are mostly regulators (repressors or activators).

Following this last point, and given that Archaea do have operonic genome organization, it appears reasonable that they seem to share three times as many transcription factors with Bacteria that they do with Eucarya. With the recent indications for the presence of operons in Eucarya (46), some of the archaeal/bacterial factors may be classified as universal if family members are discovered in eukaryotic genomes. The same is not expected for the archaeal/eukaryotic factors, given the already large existing bacterial genome repertoire.

Universal-Type Transcription-Related Proteins in Archaea. Only a limited range of the transcription-associated proteins are universal in distribution. In total, there exist 61 homologs composing 14 families. Only six of these contain exactly a single member per species. These include the second largest RNA polymerase B RpoB (B′/B″), the transcription elongation factor SPT5/NusG, the AcuC/AphA/histone deacetylase family, and three metabolic enzymes with some role in transcription (BirA, NDK, and enolase) (Table 3).

The second largest RNA polymerase RpoB is split into two subunits (B′/B″) except for *P. horikoshii.* The transcription elongation family that comprises the eukaryotic SPT5 and the bacterial NusG factors also is present in Archaea. The archaeal members of this family appear to be somewhat closer to their eukaryotic counterparts. This family is related to a number of ribosomal proteins through the common presence of the KOW domain (47).

The relationship between histone deacetylases, acetoin utilization proteins (AcuC), and acetylpolyamine aminohydrolases (AphA) has been previously observed (48). These three enzyme families belong to an ancient superfamily, and it has been suggested that a reversible acetylation and deacetylation of an aminoalkyl group of DNA-binding proteins might have been an ancestral gene regulatory mechanism (48).

The relationship of two of the metabolic enzymes, BirA and nucleoside diphosphate kinase (NDK), with the regulation of transcription, has previously been described (49): the biotin protein ligase BirA is a repressor of the biotin biosynthesis genes, whereas the NDK gene (also called nm23-H2) has been identified as the c-myc-binding protein PuF. The third metabolic enzyme family is enolase (phosphoglycerate dehydratase), which is highly similar to the transcription factor MBP1, which acts as a negative regulator for the human c-myc gene (50). Because the primary function of these proteins may not be gene regulation and their involvement in transcription may be species- or tissue-specific, they are reported here for completeness. It remains to be seen whether their role in transcription is conserved across large phylogenetic distances (49).

Two families have at least one duplicated copy in one of the four archaeal organisms: the RNA polymerase RpoA (A′/A″) with a duplicated A′ subunit (reported as possible pseudogene) split in the genes MTH297–MTH298–MTH299; and the DNA2/NAM7 helicase. The latter family has two members in the genomes of *A. fulgidus* and in *P. horikoshii* (only the C terminus of PH0109), and one member in the other two Archaea. Whereas DNA2 helicase

Table 3. Transcription-associated proteins in the four complete archaeal genomes with homologs present in Bacteria and Eukarya

Protein Family	*M. Jannaschii*	*A. fulgidus*	*M. thermoautotrophicum*	*P. horikoshii*
RPA2-RPB2-RPC2/RpoB/(B′/B′′)*	MJ1040/MJ1041	AF1886/AF1887	MTH1049/MTH1050	PH1546|PHCB019 PH1544|PHCB021/
RPA1-RPB1-RPC1/RpoC/(A′/A′′)*	MJ1042/MJ1043	AF1888/AF1889	MTH297-MTH298-MTH299/†MTH1051/MTH1052	PH1545|PHCB020
Thiamine phosphate phosphorylase/ThiE antagonist TenI		AF2074		PH1156|PHBU011
Transcriptional activator TenA				PH1160|PHBU007, PH1161|PHBU006
Biotin operon repressor BirA Acetyl-CoA:CO2 ligase	MJ1619	AF0074	MTH1916	PH0147|PHDC003
SNF2/RAD54 DNA helicase family, transcription regulator		AF2350		PH0900|PHAK019
SIR2 family		AF0112, AF1676		PH0947|PHAR015
Enolases, C-myc promoter-binding repressor (MBP1) homologs	MJ0232	AF1132	MTH43	PH1942|PHBT042
Nucleoside diphosphate kinase, C-myc transcriptional activator PuF	MJ1265	AF0767	MTH258	PH0698|PHCF002
Histone deacetylase/AcuC/AphA family	MJ0535	AF0130, AF1286, AF2290	MTH1194	PH1267|PHBJ007
DNA2/NAM7 helicase	MJ0104	AF1388, AF1960	MTH487, MTH1634	PH0109|PHBN009, PH0909|PHAK010
Helicase RAD25/XPB (TFIIH 90 kDa)		AF0358		PH0450|PHCJ004, *PH0210|PHBW005*
Helicase RAD3/XPD (TFIIH)/DinG	MJ0942			PH0697|PHCF001
Transcription elongation factor SPT5/NusG	MJ0372	AF0537	MTH1678	*PH0002|PHBC038*
Sum 61 (14)	11 (9)	18 (12)	14 (8)	18 (14)

*, Different subunits of one factor and not necessarily homologous; †, Second gene for A′ Polymerase subunit

is involved in DNA replication and NAM7 (UPF1) in mRNA turnover, they are both homologous to the transcription factor SMUBP-2 (51). However, neither DNA2 nor NAM7 family members contain the DNA-binding domain of SMUBP-2, and their direct role in transcription remains questionable.

Finally, the remaining six families are absent from at least one of the four genomes: TenI, TenA, SIR2, and the helicases SNF2/RAD54, RAD25/XPB and RAD3/XPD. One peculiar case here is the repressor TenI, an antagonist of TenA, found in *A. fulgidus* and *P. horikoshii*. TenA, although found both in Eucarya (where it is fused with ThiD) and Bacteria (52), is present in Archaea as a duplication, but only in *P. horikoshii*. Both TenI and TenA have been characterized as regulators (repressor and activator, respectively) for the production of several extracellular degradative (*deg*) enzymes (53).

SIR2 is a transcriptional silencer that has also been observed to participate in suppression of rDNA recombination and in regulation of histone deacetylation (54). However, the identification of a conserved SIR2 gene family in yeast, together with a homologous hypothetical protein family in Bacteria, has previously suggested that these proteins may have general functions in cell-cycle progression and genomic integrity (54). It is interesting, therefore, that this conserved protein family is not universally present among the Archaea, with members found only in *A. fulgidus* and in *P. horikoshii*.

The presence of the SNF2 protein family members in *A. fulgidus* and *P. horikoshii* (as well as *Halobacterium*) is very intriguing. This family consists of viral, bacterial, and eukaryotic proteins, with a variety of roles in different cellular processes, such as cell cycle control, transcriptional regulation, DNA repair, mitotic recombination, and chromatin remodeling (55). The archaeal subfamily seems to be closer to the bacterial member HepA helicase, which has recently been reported to be an RNA polymerase-associated protein (56). It is interesting that this family has undergone an extensive duplication and diversification in Eucarya, as opposed to the low degree of paralogy observed in Bacteria and Archaea.

Members of the helicase family RAD25/XPB are present only in *A. fulgidus* and *P. horikoshii*. This family, although originally identified for its DNA-repair properties, was subsequently shown to be identical to the basal transcription factor BTF2 (TFIIH subunit) (57). This family has so far a unique member among Bacteria, in *Mycobacterium tuberculosis* (hypothetical protein Rv0861c), that seems closer to the archaeal members of this family. Should this protein in *M. tuberculosis* represent a case of a horizontally transferred gene, then this family should be regarded as one more case of archaeal/eukaryotic transcription factors. The helicase family RAD3/XPD has been shown to participate in both DNA repair and basic eukaryotic transcription (58) and has DinG as its bacterial counterpart.

Despite their importance for deciphering the evolution of transcription machinery (59), some of the universally distributed proteins (e.g., BirA, NDK, enolase, SIR2) cannot be unambiguously assigned to a specific transcription-related function, because of the vast phylogenetic distances involved. With more sequence

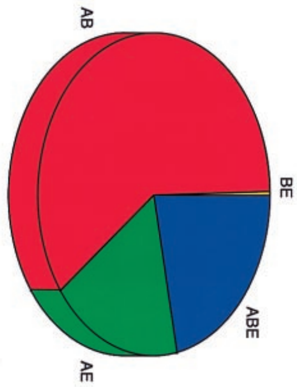

FIG. 2. Distribution of universal (ABE), bacterial/eukaryotic (BE), archaeal/bacterial (AB), and archaeal/eukaryotic (AE) families of transcription-associated proteins.

data, it is expected that this set of universal transcription proteins can only increase. Remarkably, all families considered to be of the bacterial/eukaryotic type are now known to be universal (Table 3), with the single exception being the cold-shock domain (14). This observation argues against possible patterns of horizontal gene transfer (60) from Bacteria to Eucarya, at least for the transcriptional process (Fig. 2).

The Mixed Character of Archaea Points to Their Ancient Nature. The existence of archaeal transcription factors that have no homologs in the other two domains, such as GvpE (61), points to the fact that archaeal transcription may contain elements that are unique. Yet, it is clear from the present analysis that the archaeal transcription machinery also contains a multitude of components that is distinctly found either in the bacterial or the eukaryotic domain. It is remarkable how two different types of transcription systems, i.e., bacterial regulators and eukaryotic initiators, actually coexist in Archaea.

There can be three explanations for the mixed character of archaeal transcription, depending on which phylogenetic domain is regarded as closest to the ancestral state. (*i*) If Bacteria are ancestral, then Archaea may have retained some bacterial-type transcriptional components, while inventing the eukaryotic-type transcription initiation machinery. (*ii*) If Eucarya are ancestral, then Archaea may have retained elements of the eukaryotic-type transcription and acquired various bacterial-like regulators by subsequent horizontal gene-transfer events. (*iii*) If Archaea are ancestral, then archaeal transcription may be considered as the source from which both bacterial- and eukaryotic-type transcription developed.

The problem with the first two alternatives is that they do not sufficiently explain the emergence of two preexisting incompatible sets of transcription-related proteins in Archaea: the former cannot fully account for the emergence of eukaryotic-like factors within a preexisting bacterial-type transcription system in Archaea but not in Bacteria, whereas the latter cannot account for any selective advantage of multiple horizontal transfer events of transcriptional regulators from Bacteria to Archaea, and not to Eucarya. Possibly, the only viable alternative is that archaeal transcription existed before the invention of the bacterial- and eukaryotic-like cellular entities during evolution.

The fact that Archaea may be considered the evolutionary source of transcription (and possibly other) components is consistent with the view that Archaea and Eucarya are sister groups, under the notion that Eucarya is a slowly evolving domain and probably more ancient than previously thought.

Assuming that the three phylogenetic domains are monophyletic, the single most important conclusion from the present analysis is that transcription, a fundamental process at the core of cellular physiology, could not have been reinvented twice in Archaea. Therefore, the frequent characterization of Archaea as "mosaic" (62, 63) should be discarded: "mosaicity" implies a derived state. On the contrary, the mixed character of archaeal transcription, similarly to translation (64), may be the primitive state, and by implication, the archaeal domain may be closer to the ancestral state (65).

Elements of Archaeal Transcription Present in Bacteria and Eucarya. Which are the new aspects that these findings bring forward? It is an historical accident that Archaea were discovered last, and their mixed nature suggests to many a polyphyletic character (66). Were they discovered first, it would have been easier to identify their components in Bacteria and Eucarya. In a very real sense, eukaryotic transcription is archaeal-like and not the other way around. At the same time, bacterial transcription may also be considered archaeal-like, with components that have significantly diversified after the major split of the bacterial domain. The present work forms a basis for an objective and thorough understanding of the evolution of the transcriptional machinery.

The authors thank Michael Ashburner for comments, Gillian Adams for help, and the European Bioinformatics Institute Visitors Programme for travel support. We are indebted to Professor Carl R. Woese for his invaluable comments on the manuscript. This work was supported by grants from the National Aeronautics and Space Administration (NAG 54500) and the Department of Energy (DEFG C02-95ER61963) to Carl R. Woese. C.O. acknowledges support by the European Molecular Biology Laboratory and the European Commission (DGXII—Science, Research and Development).

1. Ouzounis, C. & Kyrpides, N. (1996) *FEBS Lett.* **390,** 119–123.
2. Busby, S. & Ebright, R. H. (1994) *Cell* **79,** 743–746.
3. Gralla, J. D. (1996) *Methods Enzymol.* **273,** 99–110.
4. Nicolov, T. B. & Burley, S. K. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 15–22.
5. Langer, D., Hain, J., Thuriaux, P. & Zillig, W. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 5768–5772.
6. Gralla, J. D. (1996) *Curr. Opin. Genet. Dev.* **6,** 526–530.
7. Reznikoff, W. S., Siegele, D. A., Cowing, D. W. & Gross, C. A. (1985) *Annu. Rev. Genet.* **19,** 355–387.
8. Gallegos, M.-T., Schleif, R., Bairoch, A., Hofmann, K. & Ramos, J. L. (1997) *Microbiol. Mol. Biol. Rev.* **61,** 393–410.
9. Sheldon, M. & Reinberg, D. (1995) *Curr. Biol.* **5,** 43–46.
10. Roeder, R. G. (1996) *Trends Biochem. Sci.* **21,** 327–335.
11. Zawel, L. & Reinberg, D. (1995) *Annu. Rev. Biochem.* **64,** 533–561.
12. Hampsey, M. (1998) *Microbiol. Mol. Biol. Rev.* **62,** 465–503.
13. Allison, L. A., Moyle, M., Shales, M. & Ingles, C. J. (1985) *Cell* **42,** 599–610.
14. Wolffe, A. P. (1994) *BioEssays* **16,** 245–251.
15. Woese, C. R., Magrum, L. J. & Fox, G. E. (1978) *J. Mol. Evol.* **11,** 245–251.
16. Woese, C. R., Kandler, O. & Wheelis, M. L. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 4576–4579.
17. Huet, J., Schnabel, R., Sentenac, A. & Zillig, W. (1983) *EMBO J.* **2,** 1291–1294.
18. Zillig, W., Palm, P., Reiter, W. D., Gropp, F., Puhler, G. & Klenk, H.-P. (1988) *Eur. J. Biochem.* **173,** 473–482.
19. Ouzounis, C. & Sander, C. (1992) *Cell* **71,** 189–190.
20. Marsh, T. L., Reich, C. I., Whitelock, R. B. & Olsen, G. J. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 4180–4184.
21. Klenk, H.-P. & Doolittle, W. F. (1994) *Curr. Biol.* **4,** 920–922.
22. Thomm, M. (1996) *FEMS Microbiol. Rev.* **18,** 159–171.
23. Kyrpides, N. C. & Ouzounis, C. A. (1995) *Trends Biochem. Sci.* **20,** 140–141.
24. Kyrpides, N. C. & Ouzounis, C. A. (1997) *J. Mol. Evol.* **45,** 706–707.
25. Cohen-Kupiec, R., Blank, C. & Leigh, J. A. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 1316–1320.
26. Bult, C. J. White, O. Olsen, G. J. Zhou, L. Fleischmann, R. D. Sutton, G. G. Blake, J. A. FitzGerald, L. M. Clayton, R. A. Gocayne, J. D., *et al.* (1996) *Science* **273,** 1058–1073.
27. Klenk, H.-P. Clayton, R. A. Tomb, J.-F. White, O. Nelson, K. E. Ketchum, K. A. Dodson, R. J. Gwinn, M. HIckey, E. K. Peterson, J. D., *et al.* (1997) *Nature (London)* **390,** 364–370.
28. Smith, D. R. Doucette-Stamm, L. A. Deloughery, C. Lee, H. Dubois, J. Aldredge, T. Bashirzadeh, R. Blakely, D. Cook, R. Gilbert, *et al.* (1997) *J. Bacteriol.* **179,** 7135–7155.
29. Kawarabayasi, Y. Sawada, M. Horikawa, H. Haikawa, Y. Hino, Y. Yamamoto, S. Sekine, M. Baba, S. Kosugi, H. Hosoyama, A., *et al.* (1998) *DNA Res.* **5,** 147–155.
30. Etzold, T., Ulyanov, A. & Argos, P. (1996) *Methods Enzymol.* **266,** 114–128.
31. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215,** 403–410.
32. Heinemeyer, T., Wingender, E., Reuter, I., Hermjakob, H., Kel, A. E., Kel, O. V., Ignatieva, E. V., Ananko, E. A., Podkolodnaya, O. A., Kolpakov, F. A., *et al.* (1998) *Nucleic Acids Res.* **26,** 362–367.
33. Andrade, M. A. & Sander, C. (1997) *Curr. Opin. Biotechnol.* **8,** 675–683.
34. Ouzounis, C. A., Kyrpides, N. C. & Sander, C. (1995) *Nucleic Acids Res.* **23,** 565–570.
35. Huerta, A. M., Salgado, H., Thieffry, D. & Collado-Vides, J. (1998) *Nucleic Acids Res.* **26,** 55–59.
36. Cheng, Q., Hwa, V. & Salyers, A. A. (1992) *J. Bacteriol.* **174,** 7185–7193.
37. Pao, G. M. & Saier, M. H. J. (1995) *J. Mol. Evol.* **40,** 136–154.
38. Watanabea, H., Gojobori, T. & Miura, K. (1997) *Gene* **205,** 7–18.
39. Haldenwang, W. G. (1995) *Microbiol. Rev.* **59,** 1–30.
40. Kaine, B. P., Mehr, I. J. & Woese, C. R. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 3854–3856.
41. Rodriguez-Monge, L., Ouzounis, C. A. & Kyrpides, N. C. (1998) *Trends Biochem. Sci.* **23,** 169–170.
42. Ouzounis, C. A. & Kyrpides, N. C. (1996) *J. Mol. Evol.* **42,** 234–239.
43. Ouzounis, C. A. & Kyrpides, N. C. (1996) *J. Mol. Evol.* **43,** 541–542.
44. Kanemaki, M., Makino, Y., Yoshida, T., Kishimoto, T., Koga, A., Yamamoto, K., Yamamoto, M., Moncollin, V., Egly, J. M., Muramatsu, M. & Tamura, T. (1997) *Biochem. Biophys. Res. Comm.* **235,** 64–68.
45. Ouzounis, C. A. & Papavassiliou, A. G. (1997) in *Transcription Factors in Eukaryotes*, ed. Papavassiliou, A. G. (Landes, Austin, TX), pp. 1–21.
46. Blumenthal, T. (1998) *BioEssays* **20,** 480–487.
47. Kyrpides, N. C., Woese, C. R. & Ouzounis, C. A. (1996) *Trends Biochem. Sci.* **21,** 425–426.
48. Leipe, D. D. & Landsman, D. (1997) *Nucleic Acids Res.* **25,** 3693–3697.
49. Kyrpides, N. C. & Ouzounis, C. A. (1995) *J. Mol. Evol.* **40,** 564–569.
50. Ray, R. & Miller, D. M. (1991) *Mol. Cell. Biol.* **11,** 2154–2161.
51. Fukita, Y., Mizuta, T. R., Shirozu, M., Ozawa, K., Shimizu, A. & Honjo, T. (1993) *J. Biol. Chem.* **268,** 17643–17470.
52. Ouzounis, C. A. & Kyrpides, N. C. (1997) *J. Mol. Evol.* **45,** 708–711.
53. Pang, A. S., Nathoo, S. & Wong, S. L. (1991) *J. Bacteriol.* **173,** 46–54.
54. Baker-Brachmann, C., Sherman, J. M., Devine, S. E., Cameron, E. E., Pillus, L. & Boeke, J. D. (1995) *Genes Dev.* **9,** 2888–2902.
55. Eisen, J. A., Sweder, K. S. & Hanawalt, P. C. (1995) *Nucleic Acids Res.* **23,** 2715–2723.
56. Muzzin, O., Campbell, E. A., Xia, L., Severinova, E., Darst, S. A. & Severinov, K. (1998) *J. Biol. Chem.* **273,** 15157–15161.
57. van Vuuren, A. J., Vermeulen, W., Ma, L., Weeda, G., Appeldoorn, E., Jaspers, N. G., van der Eb, A. J., Bootsma, D., Hoeijmakers, J. H., Humbert, S., Schaeffer, L. & Egly, J.-M. (1994) *EMBO J.* **13,** 1645–1653.
58. Wang, Z., Svejstrup, J. O., Feaver, W. J., Wu, X., Kornberg, R. D. & Friedberg, E. C. (1994) *Nature (London)* **368,** 74–76.
59. Kyrpides, N. C., Overbeek, R. & Ouzounis, C. A. (1999) *J. Mol. Evol.*, in press.
60. Syvanen, M. (1994) *Annu. Rev. Genet.* **28,** 237–261.
61. Krüger, K., Hermann, T., Armbruster, V. & Pfeifer, F. (1998) *J. Mol. Biol.* **279,** 761–771.
62. Koonin, E. V., Mushegian, A. R., Galperin, M. Y. & Walker, D. R. (1997) *Mol. Microbiol.* **25,** 619–637.
63. Bell, S. D. & Jackson, S. P. (1998) *Trends Microbiol.* **6,** 222–228.
64. Kyrpides, N. C. & Woese, C. R. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 224–228.
65. Woese, C. R. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 6854–6859.
66. Doolittle, W. F. & Logsdon, J. M. J. (1998) *Curr. Biol.* **8,** R209–R211.