Proc. Natl. Acad. Sci. USA
Vol. 91, pp. 4180–4184, May 1994
Evolution

Transcription factor IID in the Archaea: Sequences in the *Thermococcus celer* genome would encode a product closely related to the TATA-binding protein of eukaryotes

(transcription initiation/molecular evolution/gene duplication/maximum likelihood and parsimony/least-squares distance)

TERRY L. MARSH, CLAUDIA I. REICH, ROBERT B. WHITELOCK*, AND GARY J. OLSEN†

Department of Microbiology, University of Illinois, Urbana, IL 61801

Communicated by Carl R. Woese, January 7, 1994

ABSTRACT **The first step in transcription initiation in eukaryotes is mediated by the TATA-binding protein, a subunit of the transcription factor IID complex. We have cloned and sequenced the gene for a presumptive homolog of this eukaryotic protein from *Thermococcus celer*, a member of the Archaea (formerly archaebacteria). The protein encoded by the archaeal gene is a tandem repeat of a conserved domain, corresponding to the repeated domain in its eukaryotic counterparts. Molecular phylogenetic analyses of the two halves of the repeat are consistent with the duplication occurring before the divergence of the archaeal and eukaryotic domains. In conjunction with previous observations of similarity in RNA polymerase subunit composition and sequences and the finding of a transcription factor IIB-like sequence in *Pyrococcus woesei* (a relative of *T. celer*) it appears that major features of the eukaryotic transcription apparatus were well-established before the origin of eukaryotic cellular organization. The divergence between the two halves of the archaeal protein is less than that between the halves of the individual eukaryotic sequences, indicating that the average rate of sequence change in the archaeal protein has been less than in its eukaryotic counterparts. To the extent that this lower rate applies to the genome as a whole, a clearer picture of the early genes (and gene families) that gave rise to present-day genomes is more apt to emerge from the study of sequences from the Archaea than from the corresponding sequences from eukaryotes.**

Woese and Fox (1) first recognized that there exist two groups of prokaryotes whose molecular features are as distinct from one another as either is from the eukaryotes. These two prokaryotic domains are now called the Bacteria (referring to the "typical" bacteria) and the Archaea (a group consisting of the methanogens, extreme halophiles, and a diverse array of "extreme thermophiles") (2). Only slowly has the fundamental nature of this division been accepted and its biological significance widely recognized. In part, this lag can be traced to the originally proposed names, "Eubacteria" and "Archaebacteria" (1), which seem to connote a specific relationship between the two prokaryotic groups. Furthermore, because the phylogenetic trees produced by studies of ribosomal RNAs (e.g., refs. 3 and 4) and proteins (e.g., ref. 5) were unrooted, it remained possible that the Archaea and Bacteria were, indeed, specifically related.

Now that a rooting for the molecular phylogenetic tree has been inferred from analyses of early gene duplications (6, 7), it appears that the Archaea and the lineage giving rise to the original nuclear genes of the Eucarya (the eukaryotes) are specifically related—that is, the Archaea and Eucarya share a more recent common ancestor than either domain does with the Bacteria. If this were true, then it is expected that there will be biological innovations shared by the Eucarya and the Archaea but not present in the Bacteria. This is a testable hypothesis.

Ouzounis and Sander (8) reported that the genome of the archaeon *Pyrococcus woesei* includes sequences that would encode a protein similar to transcription factor IIB (TFIIB) of eukaryotes. This fact, combined with previous observations that archaeal gene promoters include sequences similar to the TATA box of eukaryotic promoters (9), led them to suggest that the mechanism of transcription initiation in Archaea is more like that of the eukaryotes than like that of typical bacteria. TFIIB plays a role early in transcription initiation by RNA polymerase II (Pol II); first transcription factor IID (TFIID) binds to the promoter (TATA) region of the DNA, and then TFIIB joins the complex. Consequently, Ouzounis and Sander (8) predicted that a homolog of TFIID is present in Archaea.

TFIID is a multisubunit protein, of which the best-characterized component is the TATA-binding protein (TBP, or TFIIDτ). Although TFIID was originally characterized in terms of its recognition of the TATA sequence in Pol II promoters (for review, see ref. 10), its role is more general. (*i*) It was found that TFIID (TBP, in particular) is required for transcription from most or all Pol II promoters, even those without a recognizable TATA box (11). (*ii*) *In vitro* studies showed that TBP is also required for specific transcription initiation by RNA polymerases I and III (12). Thus, it appears that TBP is a general transcription factor (13).

The sequence of TBP from several eukaryotes has now been inferred from corresponding cDNA sequences (e.g., refs. 14–23; GenBank accession nos. M64861 and L16957). The protein is composed of three distinct regions: an amino-terminal region of variable length and sequence, followed by two conserved domains (e.g., 80% amino acid identity between yeast and human). The latter two domains are the products of an ancient direct repeat, for they display a residual amino acid identity of 26–32%.

We have been exploring the genome of *Thermococcus celer* (a close relative of *P. woesei*). This genome is < 2 Mbp (24), ≈40% the size of that of *Escherichia coli*. Thus, it provides a relatively distilled look at the genetic basis of life. We now report the identification of sequences that would produce a protein very similar to the TBP found in eukaryotes.‡

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.

Abbreviations: TFIIB and TFIID, transcription factor IIB and IID, respectively; TBP, TATA-binding protein; Pol II, RNA polymerase II.

*Present address: L-452, Lawrence Livermore National Laboratory, 700 East Avenue, Livermore, CA 94550.
†To whom reprint requests should be addressed.

‡The sequence reported in this paper has been deposited in the GenBank data base (accession no. U04932).

MATERIALS AND METHODS

Preparation of *T. celer* Genomic DNA. *T. celer* strain Vu 13 (DSM 2476) was cultured anaerobically on a mineral salts medium (24) at 88°C. Typically, cells grew to 2×10^8 cells per ml and then began to autolyse (25). At room temperature, stock cultures remain viable for 3 or more months.

Cells from overnight cultures were pelleted by centrifugation at $5000 \times g$ at 4°C. The cell pellet was resuspended in TE (10 mM Tris·HCl/1 mM Na·EDTA, pH 8.0) containing 1% $NaDodSO_4$. Cells were lysed by three to five freeze–thaw cycles. Nucleic acids were extracted twice with phenol, once with chloroform, and concentrated by EtOH precipitation (26). The pellet was resuspended in TE, RNase A was added to 40 units/ml, and the solution was incubated at 37°C for 3 hr. The nuclease was removed by phenol and chloroform extractions, and the DNA was precipitated with EtOH, as above.

Preparation of a Sheared DNA Library. A *T. celer* DNA-containing solution was adjusted to 1 M NaCl/1.7 mM Tris·HCl/0.17 mM Na·EDTA/50% (vol/vol) glycerol. A volume of 0.5 ml was placed in a nebulizer that had been arranged so that most of the mist produced collected on the walls of a length of Tygon tubing and ran back down into the nebulizer. Nitrogen gas (20 psi; 1 psi = 6.9 kPa) was passed through the nebulizer for 150 sec. The solution was collected, and the DNA was concentrated by EtOH precipitation. The average size of the fragments produced was estimated by agarose gel electrophoresis to be 400 bp.

The ends of the sheared DNA were evened by treatment with T4 DNA polymerase (26). Plasmid pGEM-4Z (Promega) was linearized by digestion with *Hin*cII and dephosphorylated with calf intestinal phosphatase (Promega). Ligations were performed by using T4 DNA ligase (United States Biochemical), following the manufacturer's instructions. A 10-μl reaction contained 1 μg of sheared DNA and 1 μg of plasmid DNA.

E. coli strains JM109 and DH5α were grown on Luria broth (LB). Electrocompetent cells were prepared (26) and stored at −70°C until needed. Cells were transformed by electroporation in a Gene Pulser (Bio-Rad), according to the manufacturer's instructions. Transformants were diluted in LB, incubated for 45 min, and plated on LB plates containing ampicillin at 100 μg/ml. Colonies were picked and inoculated into microtiter plate wells containing 75 μl of LB. After incubation at 37°C for 2–3 hr, an equal volume of sterile glycerol was added to each well, and the plates were frozen on dry ice. The plates were stored at −70°C.

Inverse PCR. *T. celer* genomic DNA was digested with *Apa* I restriction enzyme, diluted, and then incubated with T4 DNA ligase to produce a heterogeneous population of circles. Primers within the known DNA sequence were added to a concentration of 1 mM, and the solution was subjected to 30 cycles of PCR (denature 60 sec at 93°C, anneal 30 sec at 52°C, and extend 25 sec at 72°C). The products of the reaction were ligated with *Xba* I linkers, digested with *Xba* I to remove concatemers, and cloned into dephosphorylated *Xba* I-cut pGEM-4Z.

Restriction Fragment Cloning. Restriction fragments containing the carboxyl-terminal sequences of the presumptive TBP gene were derived from clone IIb6. ^{32}P-radiolabeled probes, prepared using multiprime DNA labeling (Amersham), bound to a 5-kbp fragment of *Hin*dIII-digested *T. celer* genomic DNA in Southern blot analyses (26). A library of 5-kbp *Hin*dIII restriction fragments was prepared in Blue-Script II KS$^-$ (Stratagene), and the desired recombinants were identified by colony hybridization (26).

DNA Sequencing. Plasmid DNAs were purified and sequenced with Sequenase version 2.0 (United States Biochemical), according to the manufacturer's instructions. DNA sequences were determined from both strands by extension from vector-specific priming sites and primer walking. When sequencing cloned PCR products, three clones were examined.

Sequence Analysis. The initial characterization of possible translation products of the randomly cloned sequences used the FASTA program (27) to find similar sequences in a local copy of the Swiss-prot data base (28). Later analyses used the electronic-mail-based and Internet BLAST (29) servers at the National Center for Biotechnology Information to search the nonredundant protein data base. Related sequences were collected and aligned manually in the SEQEDIT program [available through the Ribosomal Database Project (30)].

Maximum-likelihood estimates of (observed) distances between pairs of amino acid sequences were found using the Dayhoff PAM matrix (31) option of the PROTDIST program in version 3.5 of the PHYLIP package (32). The statistical uncertainties of the distances were estimated by calculating the root-mean-square deviations (from the observed distances) of 10 additional distances based on bootstrap resamplings (generated by the SEQBOOT program in the PHYLIP package) of the sequences. Least-squares fits of phylogenetic trees to

```
Hom.sapi  ...PATPASESSGIVPQLQNIVSTVNLGCKLDLKTIALRARN-AEYNPKRFAAVIMRIR---EPRTTALIFSSGKMVCTGAKSEEQSRLAARKYARVVQKLGFP-  247
Dro.mela  ...PATPGSADPGIVPQLQNIVSTVNLCCKLDLKKIALHARN-AEYNPKRFAAVIMRIR---EPRTTALIFSSGKMVCTGAKSEDDSRLAARKYARIIQKLGFP-  261
Sch.pomb  ...EVSKNEGVSGIVPTLQNIVATVNLDCRLDLKTIALHARN-AEYNPKRFAAVIMRIR---EPKSTALIFASGKMVVLGGKSEDDSKLASRKYARIIQKLGFN-  140
Sac.cerv  ...SEKDTSATSGIVPTLQNIVATVTLGCRLDLKTVALHARN-AEYNPKRFAAVIMRIR---EPKTTALIFASGKMVVTGAKSEDDSKLASRKYARIIQKIGFA-  148
Aca.cast  ...DVDRTKHPSGIVPTLQNIVSTVNLGCKLDLKNIALHARN-AEYNPKRFAAVIMRIR---EPKTTALIFASGKMVCTGAKSEEASRLAARKYARIIQKLGFA-  167
Zea mays  ...PVDLSKHPSGIVPTLQNIVSTVNLDCKLDLKAIALQARN-AEYNPKRFAAVIMRIR---EPKTTALIFASGKMVCTGAKSEQQSKLAARKYARIIQKLGFP-  107
Dic.disc  ...NVDLSKHPSGIIPTLQNIVSTVNMATELYLKAIALGARN-AEYNPKRFAAVIMRIR---EPKTTALIFKSGKMVCTGAKSEDASRFAARKYARIIQKLDFP-  109
Tet.ther  ...ETMDKSISEDLYPKLQNIVSTVNLSTKLDLKQIALRARN-AEYNPKRFAAVIMRLR---DPKTTALIFASGKMVCTGAKTEEDSNRAARKYAKIIQKIGFP-  135
Pla.falc  ...DLKNKLVHKNISLNIHNIISSANLCIDINLRLVAVSIRN-AEYNPSKINTLIIRLN---KPQCTALIFKNGRIMLTGTRTKKDSIMGCKKIAKIIKIVTKDK  135
Tc.celer          MSNVKLRIENIVASVDLLTQLNLERVIEMCPH-SKYNPEEFPGIICRFD---EPKVALLIFSSGKLVVTGAKSVEDIERAVNKLIQMLKKIGAKF   91

Hom.sapi          AKFLDFKIQNMVGSCDVKFPIRLEGLVLTHQQFSSYEPELFPGLIYRMI---KPRIVLLIFVSGKVVLTGAKVRAEIYEAFENIYPILKGFRKTT  339
Dro.mela          AKFLDFKIQNMVGSCDVKFPIRLEGLVLTHCNFSSYEPELFPGLIYRMV---RPRIVLLIFVSGKVVLTGAKVRQEIYDAFDKIFPILKKFKKQS  353
Sch.pomb          AKFTDFKIQNIVGSCDVKFPIRLEGLAYSHGTFSSYEPELFPGLIYRMV---KPKVVLLIFVSGKIVLTGAKVREEIYQAFEAIYPVLSEFRKH  231
Sac.cerv          AKFTDFKIQNIVGSCDVKFPIRLEGLAFSHGTFSSYEPELFPGLIYRMV---KPKIVLLIFVSGKIVLTGAKQREEIYQAFEAIYPVLSEFRKM  239
Aca.cast          AKFLDFKIQNIVGSCDVRFPIRLEGLAFAHNHYCSYEPELFPGLIYRMV---QPKIVLLIFVSGKIVLTGAKVREEIYEAFENIYPVLTEYKKT  258
Zea mays          AKFKDFKIQNIVGSCDVKFPIRLEGLAYSHGAFSSYEPELFPGLIYRMK---QPKIVLLIFVSGKIVLTGAKVREETYTAFENIYPVLSEFRKIQQ  200
Dic.disc          ARFTDFKIQNIVGSCDVKFPIKLELLHNAHTSFTNYEPEIFPGLIYKMI---QPKVLLLIFVSGKIVLTGAKVREYIYEAFENIYPVLSAFKKVNAITQ  205
Tet.ther          VQFKDFKIQNIVGSTDVKFPINLDHLEQDHKKFVQYEPEIFPGKIYREF---NTKIVLLIFVSGKIVLTGAKTRENINKAFQKIYWVLYNYQKKDYRGA...  246
Pla.falc          VKFCNFKIENIIASANCNIPIRLEVLAHDHKEYCNYEPELFAGLVYRYKPTSNLKSVILIFVSGKIIITGCKSVNKLYTVFQDIYNVLIQYKN  228
Tc.celer          SRAPQIDIQNMVFSGDIGMEFNLDAVALSLPN-CEYEPEQFPGVIYRVK---EPRAVILLFSSGKIVCSGAKSEHDAWEAVRKLLRELEKYDLIGEGEEEEW  189
```

FIG. 1. Alignment of inferred *T. celer* TBP sequence with those from diverse eukaryotes. The variable amino-terminal portion of the eukaryotic sequences is not shown. The first block of the alignment shows the first repeat of the conserved domain, and the second block shows the second repeat. The blocks are aligned to show the similarity of the two repeats. The full organism names are as follows: *Hom.sapi, Homo sapiens* (18); *Dro.mela, Drosophila melanogaster* (20); *Sch.pomb, Schizosaccharomyces pombe* (17, 19); *Sac.cerv, Saccharomyces cerevisiae* (14–16); *Aca.cast, Acanthamoeba castellanii* (22); *Zea mays* (21); *Dic.disc, Dictyostelium discoideum* (GenBank accession no. M64861); *Tet.ther, Tetrahymena thermophila* (GenBank accession no. L16957); *Pla.falc, Plasmodium falciparum* (23); and *Tc.celer, T. celer.*

Table 1. The fraction of identical amino acids (lower left) and amino acid replacements per position (upper right) between pairs of TBP sequences

	Hom.sapi	*Dro.mela*	*Sch.pomb*	*Sac.cerv*	*Aca.cast*	*Zea mays*	*Dic.disc*	*Tet.ther*	*Pla.falc*	*T.celer*
Hom.sapi	—	0.123	0.235	0.222	0.171	0.175	0.281	0.415	1.051	1.224
Dro.mela	0.893	—	0.237	0.228	0.199	0.213	0.312	0.434	1.056	1.201
Sch.pomb	0.798	0.803	—	0.070	0.173	0.130	0.278	0.449	0.979	1.272
Sac.cerv	0.803	0.803	0.933	—	0.153	0.142	0.291	0.413	0.963	1.206
Aca.cast	0.843	0.826	0.843	0.860	—	0.122	0.226	0.386	0.939	1.247
Zea mays	0.843	0.826	0.882	0.871	0.888	—	0.219	0.381	0.939	1.251
Dic.disc	0.764	0.747	0.775	0.764	0.809	0.815	—	0.413	1.010	1.337
Tet.ther	0.691	0.691	0.680	0.697	0.719	0.725	0.691	—	0.915	1.194
Pla.falc	0.383	0.400	0.417	0.422	0.439	0.433	0.394	0.433	—	1.483
T.celer	0.382	0.388	0.376	0.399	0.388	0.376	0.354	0.388	0.294	—

The distances were calculated using the PAM option of the PROTDIST program (32). For abbreviations, see Fig. 1 legend.

the observed pairwise distance data (33) were performed using the FITCH program in version 3.5 of the PHYLIP package (32). Each term in the sum of squares was weighted by one over the variance of the observed distance (i.e., inversely as the mean square deviation of the bootstrap distance estimates from the observed distance) (see ref. 34) by using the subreplicates option of the FITCH program. The order of sequence addition was repeatedly varied until the same best tree was found at least three times.

Parsimony-based phylogenetic analyses were performed by using version 3.0s of the PAUP program (35). The costs of changing amino acids were based upon the BLOSUM62 matrix (36). The uncertainties in trees were examined by using bootstrap analysis (37, 38).

Maximum-likelihood phylogenetic analyses of the protein sequences were done with PROTML version 1.00b (39). In addition to the default tree search, the most parsimonious trees found in a PAUP branch and bound search were supplied as user trees to PROTML. The confidence with which one tree is preferred over another was evaluated by using the paired-sites test of PROTML.

RESULTS

We have undertaken the sequence characterization of cloned fragments of randomly sheared genomic DNA from *T. celer* to gain insights into the fundamental machineries used by members of the Archaea, the least-studied domain of life. Comparison of potential translation products of these sequences with those in the GenPept and Swiss-prot data bases showed that the clone designated IIb6 could encode a product with substantial similarity to the TBP of eukaryotes. This clone was selected for further sequence determination and analysis.

Analysis of clone IIb6 showed that it contained sequences corresponding to one copy of the conserved repeat of the eukaryotic TBP followed by a termination codon. Because no upstream sequences were present in this clone, it could not be determined whether it was the second half of a tandem repeat, or whether the archaeal version of this protein was a dimer of half-sized molecules. This ambiguity was accentuated by the nearly equal similarities of the inferred *T. celer* protein sequence to both the first and second repeats of the eukaryotic proteins (e.g., 42% and 38% identity to the human sequence repeats). Intriguingly, these values are higher than all observed similarities between the two halves of the molecules from eukaryotes (see below), indicating greater conservation of the sequence in *T. celer*.

The amino-terminal portion of the gene was recovered both by inverse PCR and by cloning size-selected restriction fragments (see *Materials and Methods*). The resulting clones permitted the complete DNA-sequence determination of the open reading frame with similarity to TBP genes.

The inferred translation product of the presumptive *T. celer* TBP sequence is presented in Fig. 1, where it is aligned with the corresponding parts of eukaryotic TBPs. An in-frame termination codon 15 nt upstream of the presumptive translation initiation site suggests that the *T. celer* protein starts precisely at the beginning of the first of the tandem repeats. The resulting translation product (including the initiating methionine) would be composed of 189 amino acids, giving a molecular mass of 21,313 Da, the lowest of any reported TBP sequence. The inferred carboxyl-terminal sequence is slightly longer than most other TBPs, although not unprecedented (GenBank accession no. L16957).

Comparing the *T. celer* sequence to those from eukaryotes, the high level of amino acid conservation is evident. Table 1 presents the fraction of identical amino acids and the evolutionary distances (amino acid replacements per position) between pairs of TBP sequences.

Fig. 2 presents the maximum-likelihood phylogenetic tree inferred from the protein sequences. The only features of the tree that are supported with high confidence by all of the phylogenetic analysis methods (see *Materials and Methods*)

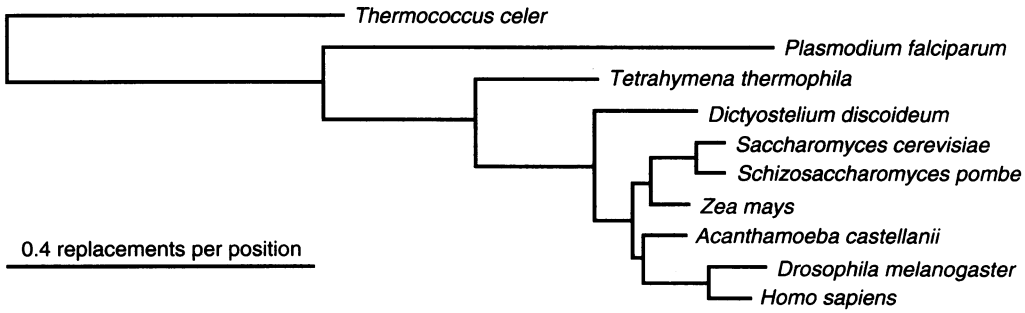

FIG. 2. A maximum-likelihood phylogenetic tree of TBP sequences. Trees were evaluated with the PROTML program (39), which uses the PAM (percent accepted mutations) model of change of Dayhoff (31). The tree was found by evaluating the 77 most parsimonious branching orders from a branch and bound search and a cost matrix based on BLOSUM62 (36).

Table 2. The fraction of identical amino acids between the first and second copies of the direct repeat in TBP sequences

Organism	Amino acid identity of halves
T. celer	0.420
Drosophila melanogaster	0.316
Tetrahymena thermophilus	0.316
Acanthamoeba castellanii	0.305
Saccharomyces cerevisiae	0.294
Zea mays	0.282
Homo sapiens	0.271
Schizosaccharomyces pombe	0.271
Plasmodium falciparum	0.267
Dictyostelium discoideum	0.260
Average among eukaryotes	0.287

are the grouping of *Drosophila* with human, the grouping of *Saccharomyces* with *Schizosaccharomyces*, and *Plasmodium* as the most deeply branching of these eukaryotes. The lack of finer resolution is not surprising, given the relatively small size of the TBP and the intentionally diverse collection of species sampled. In spite of the uncertainty in the details of the radiation of the plant, animal, fungal, *Acanthamoeba*, and *Dictyostelium* lineages, there is good agreement between the methods. Out of the >2,000,000 possible branching orders, the maximum-likelihood tree is the sixth most parsimonious. Similarly, the best tree by least-squares analysis is the 24th most parsimonious and the fifth best by the maximum-likelihood criterion.

The alignment in Fig. 1 also allows comparison of the two halves of the direct repeat. Table 2 presents the fraction identity between the first and second copies of the repeated sequence. The two are most similar in the case of the *T. celer* TBP. The higher similarity of the halves of the *T. celer* sequence could be the result of greater sequence conservation (lower average rate of change) in the case of the archaeal sequence, or it could reflect a more recent (and hence *independent*) duplication event. These possibilities can be distinguished by a phylogenetic analysis of the two halves of the molecule. In the former case, the first half of the archaeal sequence will be specifically related to the first half of the eukaryotic sequences. In the latter case the two halves of the archaeal sequence will be specifically related to each other. The inferred phylogeny of the halves (Fig. 3) supports the single duplication event, with a lower average rate of change in the archaeal lineage.

In all TBP sequences from eukaryotes, the first repeat contains one more amino acid than the second repeat. This difference could be the result of either an insertion into the first repeat or a deletion from the second repeat. In the archaeal protein, the repeats are both of the same length as the second repeat of the eukaryotic sequences. Thus, the archaeal sequence suggests that the length difference results from the insertion of 3 nt into the first repeat, subsequent to the divergence of the Archaea and Eucarya.

DISCUSSION

Identification of a TBP in a Member of the Archaea. The recognition of a sequence related to the TFIIB gene of eukaryotes in *P. woesei*, a member of the Archaea, led Ouzounis and Sander (8) to predict that a homolog of the TBP would also be found, a prediction we have confirmed.

Early History of the Transcription Apparatus. The phylogeny and structure of this archaeal gene strongly suggest that the invention of the TBP and the evolution of its current tandem-repeat configuration predate the separation of the Archaea and Eucarya. The discovery of a TFIIB-like protein in both the Archaea and Eucarya further defines the picture of this early transcription apparatus, indicating that the aboriginal version probably was quite similar to the one presently used by Pol II. However, the Archaea possess only *one* identified RNA polymerase, which is related to all *three* of the RNA polymerases observed in eukaryotes (5, 40). Thus, all three eukaryotic transcription machineries appear to be derived from a system resembling that of Pol II–a conjecture that is consistent with the discovery that all three eukaryotic polymerases require TBP (13).

Consistency of the Archaeal TBP with the Structure of the Eukaryotic Proteins. The crystal structure of TBP complexed to DNA has been reported recently (41, 42), allowing us to examine the inferred archaeal sequence for consistency with these structures. In brief, all of the features found conserved among the eukaryotic sequences and attributed function in the structural studies are found conserved in the archaeal sequence as well. In particular, in 22 of the 25 positions of yeast TBP that contact the DNA (41), *T. celer* has the same amino acid as one or more of the eukaryotic sequences in Fig. 1. In contrast, when the same test is applied to the *Plasmodium* sequence, only 15 of the 25 positions match another sequence.

The Archaeal Genome as a Window into the Primordial Eukaryotic Make-Up. The rooting of the universal phylogenetic tree inferred from ancestral gene duplications indicates that the Archaea and Eucarya are specific relatives (6, 7). Because the common ancestry of Eucarya and Archaea is more recent than that of either of these groups with the Bacteria, as our knowledge of the Archaea increases, so does our understanding of the nature and genetic complement of the "urkaryote" (43), the ancestor to the eukaryotes before the acquisition of organelles. In this context, the greater

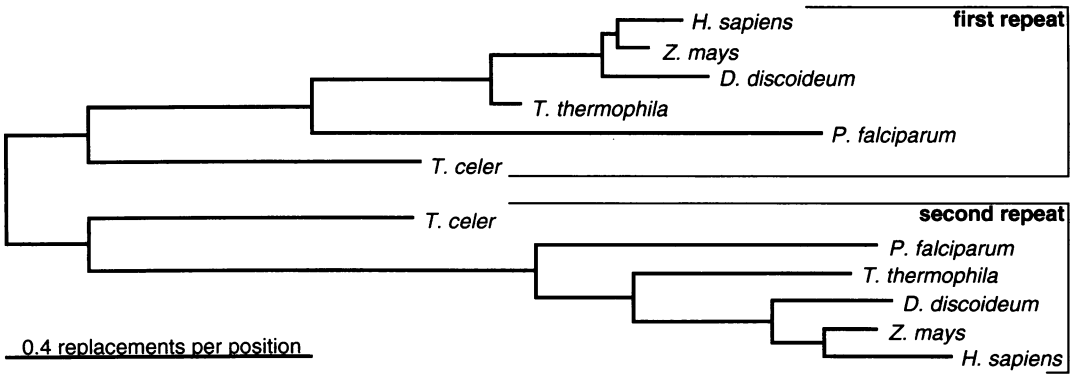

FIG. 3. A maximum-likelihood phylogenetic tree relating the two repeats of the TBP from *T. celer* and those of selected eukaryotes (see Fig. 2 for genera). The tree was inferred with the PROTML program (39). Alternative placements of the halves of the *T. celer* sequence were tested and found less favorable. The relationships inferred from each half of the TBP sequence precisely match those in Fig. 2.

conservation of the two halves of the *T. celer* protein relative to those of the eukaryotes indicates that some of the very distant molecular relationships, such as those sought in the identification of ancient gene (or domain, or exon) families, will be more easily recognized in the archaeal sequences than in the corresponding sequences from eukaryotes.

The genetic complement of the most recent common ancestor of extant life is not just an issue of academic interest; it also has practical consequences. In particular, it limits the genetic building blocks from which more complex present-day eukaryotic genomes were built. Although the rate at which truly new protein structures are being invented is not known, it is often assumed that it is easier to recruit old proteins (or domains, or exons) into new functions than to invent a functional protein *de novo* from random sequences in the genome. To the extent that this is so, the finite primordial repertoire ensures that we will continue to discover new members of a relatively limited number of protein families (31) and that we will be able to leverage our knowledge about individual proteins in these families to make predictions about other members, whose structure and function are otherwise unknown.

The order of the first two authors was determined randomly. We are grateful to C. R. Woese for encouragement, suggestions, and useful comments on the manuscript. We also thank those individuals and organizations that make fine data bases, computer programs, and computing services available to the scientific community. T.L.M. is the recipient of a National Institutes of Health Senior Research Fellowship and received additional support from National Aeronautics and Space Administration Grant NAGW-2554 to C. R. Woese. R.B.W. was supported, in part, by a grant from Exxon Education Foundation. G.J.O. is the recipient of a National Science Foundation Presidential Young Investigator Award (DIR 89-57026).

1. Woese, C. R. & Fox, G. E. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 5088–5090.
2. Woese, C. R., Kandler, O. & Wheelis, M. L. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 4576–4579.
3. Woese, C. R. (1987) *Microbiol. Rev.* **51,** 221–271.
4. Kjems, J. & Garrett, R. A. (1985) *Nature (London)* **318,** 675–677.
5. Pühler, G., Leffers, H., Gropp, F., Palm, P., Klenk, H.-P., Lottspeich, F., Garrett, R. A. & Zillig, W. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 4569–4573.
6. Gogarten, J. P., Kibak, H., Dittrich, P., Taiz, L., Bowman, E. J., Bowman, B. J., Manolson, M. F., Poole, R. J., Date, T., Oshima, T., Konishi, J., Denda, K. & Yoshida, M. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 6661–6665.
7. Iwabe, N., Kuma, K., Hasegawa, M., Osawa, S. & Miyata, T. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 9355–9359.
8. Ouzounis, C. & Sander, C. (1992) *Cell* **71,** 189–190.
9. Reiter, W.-D., Hüdepohl, W. & Zillig, W. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 9509–9513.
10. Greenblatt, J. (1991) *Cell* **66,** 1067–1070.
11. Pugh, R. F. & Tjian, R. (1991) *Genes Dev.* **5,** 1935–1948.
12. White, R. J., Jackson, S. P. & Rigby, P. W. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 1949–1953.
13. White, R. J. & Jackson, S. P. (1992) *Trends Genet.* **8,** 284–288.
14. Hahn, S., Buratowski, S., Sharp, P. A. & Guarente, L. (1989) *Cell* **58,** 1173–1181.
15. Horikoshi, M., Wang, C. K., Fujii, H., Cromlish, J. A., Weil, P. A. & Roeder, R. G. (1989) *Nature (London)* **341,** 299–303.
16. Schmidt, M. C., Kao, C. C., Pei, R. & Berk, A. J. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 7785–7789.
17. Fikes, J. D., Becker, D. M., Winston, F. & Guarente, L. (1990) *Nature (London)* **346,** 291–294.
18. Hoffmann, A., Sinn, E., Yamamoto, T., Wang, J., Roy, A., Horikoshi, M. & Roeder, R. G. (1990) *Nature (London)* **346,** 387–390.
19. Hoffmann, A., Horikoshi, M., Wang, C. K., Schroeder, S., Weil, P. A. & Roeder, R. G. (1990) *Genes Dev.* **4,** 1141–1148.
20. Muhich, M. L., Iida, C. T., Horikoshi, M., Roeder, R. G. & Parker, C. S. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 9148–9152.
21. Haass, M. M. & Feix, G. (1992) *FEBS Lett.* **301,** 294–298.
22. Wong, J. M., Liu, F. & Bateman, E. (1992) *Gene* **117,** 91–97.
23. McAndrew, M. B., Read, M., Sims, P. F. & Hyde, J. E. (1993) *Gene* **124,** 165–171.
24. Noll, K. M. (1989) *J. Bacteriol.* **171,** 6720–6725.
25. Zillig, W., Holz, I., Janekovic, D., Schäfer, W. & Reiter, W. D. (1983) *Syst. Appl. Microbiol.* **4,** 88–94.
26. Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Lab. Press, Plainview, NY), 2nd Ed.
27. Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 2444–2448.
28. Bairoch, A. & Boeckmann, B. (1992) *Nucleic Acids Res.* **20,** 2019–2022.
29. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215,** 403–410.
30. Larsen, N., Olsen, G. J., Maidak, B. L., McCaughey, M. J., Overbeek, R., Macke, T. J., Marsh, T. L. & Woese, C. R. (1993) *Nucleic Acids Res.* **21,** 3021–3023.
31. Dayhoff, M. O., ed. (1978) *Atlas of Protein Sequence and Structure* (Natl. Biomed. Res. Found., Silver Spring, MD), Vol. 5, Suppl. 3.
32. Felsenstein, J. (1989) *Cladistics* **5,** 164–166.
33. Cavalli-Sforza, L. L. & Edwards, A. W. F. (1967) *Evolution* **21,** 550–570.
34. Olsen, G. J. (1988) *Methods Enzymol.* **164,** 793–812.
35. Swofford, D. L. (1993) PAUP, Phylogenetic Analysis Using Parsimony (The Smithsonian Institution, Washington, DC), Version 3.0.
36. Henikoff, S. & Henikoff, J. G. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 10915–10919.
37. Efron, B. (1982) *The Jackknife, the Bootstrap, and Other Resampling Plans* (Soc. Indust. Appl. Math., Philadelphia).
38. Felsenstein, J. (1985) *Evolution* **39,** 783–791.
39. Adachi, J. & Hasegawa, M. (1992) *MOLPHY: Programs for Molecular Phylogenetics I—PROTML: Maximum Likelihood Inference of Protein Phylogeny*, Computer Science Monographs, No. 27 (Inst. Stat. Math., Tokyo).
40. Iwabe, N., Kuma, K., Kishino, H., Hasegawa, M. & Miyata, T. (1991) *J. Mol. Evol.* **32,** 70–78.
41. Kim, Y., Geiger, J. H., Hahn, S. & Sigler, P. B. (1993) *Nature (London)* **365,** 512–520.
42. Kim, J. L., Nikolov, D. B. & Burley, S. K. (1993) *Nature (London)* **365,** 520–527.
43. Woese, C. R. & Fox, G. E. (1977) *J. Mol. Evol.* **10,** 1–6.