Nucleic Acids Research, 1996, Vol. 24, No. 11 **2125–2132**

recA-like genes from three archaean species with putative protein products similar to Rad51 and Dmc1 proteins of the yeast *Saccharomyces cerevisiae*

Steven J. Sandler, Leslie H. Satin, Hardeep S. Samra and Alvin J. Clark

Department of Molecular and Cell Biology, Division of Genetics, 401 Barker Hall #3202, University of California at Berkeley, Berkeley, CA 94720-3202, USA

Received January 25, 1996; Revised and Accepted April 8, 1996

GenBank accession nos U45310–U45312 (incl.)

ABSTRACT

The process of homologous recombination has been documented in bacterial and eucaryotic organisms. The *Escherichia coli* RecA and *Saccharomyces cerevisiae* Rad51 proteins are the archetypal members of two related families of proteins that play a central role in this process. Using the PCR process primed by degenerate oligonucleotides designed to encode regions of the proteins showing the greatest degree of identity, we examined DNA from three organisms of a third phylogenetically divergent group, Archaea, for sequences encoding proteins similar to RecA and Rad51. The archaeans examined were a hyperthermophilic acidophile, *Sulfolobus sofataricus* (*Sso*); a halophile, *Haloferax volcanii* (*Hvo*); and a hyperthermophilic piezophilic methanogen, *Methanococcus jannaschii* (*Mja*). The PCR generated DNA was used to clone a larger genomic DNA fragment containing an open reading frame (orf), that we refer to as the *radA* gene, for each of the three archaeans. As shown by amino acid sequence alignments, percent amino acid identities and phylogenetic analysis, the putative proteins encoded by all three are related to each other and to both the RecA and Rad51 families of proteins. The putative RadA proteins are more similar to the Rad51 family (~40% identity at the amino acid level) than to the RecA family (~20%). Conserved sequence motifs, putative tertiary structures and phylogenetic analysis implied by the alignment are discussed. The 5′ ends of mRNA transcripts to the *Sso radA* were mapped. The levels of *radA* mRNA do not increase after treatment with UV irradiation as do *recA* and *RAD51* transcripts in *E.coli* and *S.cerevisiae*. Hence it is likely that *radA* in this organism is a constitutively expressed gene and we discuss possible implications of the lack of UV-inducibility.

INTRODUCTION

Homologous recombination is a process whereby two duplexes of DNA interact to either exchange or rearrange segments of their DNA. Cells use this process for DNA repair (1,2), generation of genetic diversity, gene regulation (3) and cell cycle-dependent events (4). Homologous recombination has been observed in bacteria, eucaryotes and now in hyperthermophilic archaeans (D. Grogan, personal communication).

Much effort has been put into understanding the process of homologous recombination at the molecular level. The central step in homologous recombination is bringing together the two interacting DNA molecules, searching for the homology and then exchanging the DNA strands between them. In the bacterium *Escherichia coli*, it is believed a single gene product, called RecA, is responsible for this activity. Known *in vitro* activities of RecA critical for its biological function include making RecA–ssDNA helical filaments, joint molecule formation between ssDNA and homologous dsDNA, the ability to promote branch migration and the ability to hydrolyze ATP (reviewed in 1,2,5,6). *recA* mutant strains are extremely deficient in homologous recombination, sensitive to DNA damaging agents (i.e., UV irradiation) and show an inability to process regulatory and enzymatic proteins used in DNA repair and mutagenesis. Currently 65 *recA* genes from different bacteria have been cloned and sequenced (6,7) (Roca and Cox, personal communication). Phylogenetic analysis of the bacterial *recA* genes has been done (7–9).

Putative homologs of RecA from eucaryotes have been identified based on amino acid sequence similarity (10). Some of these proteins in *Saccharomyces cerevisiae* are Rad51, Dmc1, Rad57 (11) and Rad55 (12). The amino acid sequence similarity for RecA is greatest for Rad51 and Dmc1. Correlations between RecA and Dmc1 have also been predicted based on amino acid alignments and the X-ray crystal structure of a form of *E.coli* RecA combined with ADP (13). Biochemical comparisons between RecA and Rad51 have been discussed elsewhere (1,2,5,14). Physiologically, mutants of *RAD51*, *RAD55* and *RAD57* also appear to be like bacterial *recA* mutants in that they are deficient in types of recombinational DNA repair (11,12,15,16). Mutants of *DMC1* are deficient in meiotic recombination, in forming synaptonemal complexes and in completing meiotic prophase (4). Recently Rad51 and Dmc1 proteins have been shown to colocalize in the nucleus prior to meiotic chromosome synapsis (17,18). *RAD51* gene expression in *S.cerevisiae*, like *recA*, is inducible at the level of transcription in response to DNA damage (16).

Rad51-like proteins have been found in the genomes of other eucaryotes: *Schizosaccharomyces pombe* (fission yeast) (19–21),

* To whom correspondence should be addressed

Homo sapiens (human) (20), *Mus musculus* (mouse) (20,22), *Gallus gallus* (chicken) (23), *Xenopus laevis* (South African clawed frog) [(24) GenBank D38488], *Drosophila melanogaster* (fruit fly) (25), *Neurospora crassa* (26) and *Lycopersicon esculentum* (tomato) (GenBank U22441). Dmc1-like proteins have been found in *H.sapiens* (human) (GenBank D64108), *M.musculus* (mouse) (GenBank D64107), *Candida albicans* (a fungus) (GenBank U39808) and *Lilium longiflorum* (lily) (27,28).

Genes for synaptase proteins like RecA and Rad51 have not yet been isolated from the third organismal domain, Archaea. It is of interest, therefore, to see if organisms from this domain have such genes and, if so, whether the protein they encode more resembles the RecA or the Rad51 family. Since phylogenetic trees based upon aminoacyl-tRNA synthetase (29), H^+-ATPase (30) and elongation factors genes (31) show that the Archaea and Eucarya share a more recent common ancestor than they do with bacteria, it would be predicted that archaeal RecA-like proteins should more closely resemble the eucaryal Rad51 proteins than the RecA bacterial proteins. We found putative *recA* and *RAD51* homologues in three archaeans and we tested one of these genes for UV inducibility.

MATERIALS AND METHODS

Archaean strains and bacterial strains

One *E.coli* strain was used in this work, namely DH5α. The genotype is *supE44* Δ*lacU169* (Φ*80 lacZ* Δ*M15*) *hsdR17 recA1 endA1 gyrA96 thi-1 relA1*. *Sulfolobus solfataricus* P2, obtained from D. Grogan, was cultured in medium similar to the type used to grow *Sulfolobus acidocaldarious* (32) with the exception that 0.2% sucrose was used instead of 0.2% xylose. Tryptone (0.2%) was used as the nitrogen source. Standing overnight cultures grown at 75°C were used to innoculate larger cultures. Cultures for DNA and RNA isolation were grown at 75°C with mild aeration provided by shaking in a rotating water bath.

DNA isolation

Sulfolobus solfataricus cells were grown as described above to an OD_{600} of 1.0. Cells of *Methanococcus janaschii* were the generous gift of D. Clark. Cells of each species were pelleted and resuspended in 1/10 vol 10 mM Tris–HCl pH 8, 1 mM EDTA, 100 mM NaCl. 1/10 vol of a 10% solution of sodium lauryl sarcosine was added. Then an equal volume of phenol–chloroform–isoamyl alcohol (25:24:1) was added. The mixture was vortexed and then centrifuged to separate the phases. The aqueous phase was removed and extracted two more times. The DNA was precipitated by adding 1/10 vol 3 M sodium acetate pH 7 and 2 1/2 vol of ethanol. After centrifugation, the DNA pellet was washed once with 70% ethanol, centrifuged and the pellet resuspended in water. This was then treated with RNase, re-extracted with the phenol–chloroform solution and precipitated with ethanol and 3 M sodium acetate. The final pellet was washed with 70% ethanol and resuspended in water.

DNA from *Haloferax volcanni* strain WFD11 was the generous gift of M. Dyall-Smith.

PCR primers, reactions and cloning

PCR primers (5′→3′) used were: GGWCCWGARTCWTCWGGWAARAC and ACWGARTCWACWACWATWAC (*Sso*), ACWGAATTTTTTGGWGAATTTGGWTCWGGRAA and CTRAAWGTWCCYTCWGTRTCWATRTA (*Mja*), and ATCACSGARSTSTTCGGSGARTT and SCGGAASGTSCCYTCSGTGTCGATRTA (*Hvo*) where W = A or T, S = G or C, R = G or A, and Y = C or T.

An MJ Research thermocycler PTC-150 was used for PCR. The reaction mixture contained 50 µl total. DNA (500 µg) was mixed with 250 pmol of each redundant primer mix. An aliquot of 10 µl of the Buffer 'C' from the Invitrogen PCR Optimizer Kit (Cat# K1220-01) was used (5× buffer solution containing 300 mM Tris–HCl, 75 mM $(NH_4)_2SO_4$, 12.5 mM $MgCl_2$ at pH 8.5). Two units of *Taq* polymerase purchased from Perkin Elmer was added. The samples were initially heated to 94°C for 5 min. Then the 5 µl of 2.5 mM solution of the dNTPs was added to final concentration of 0.25 mM. Finally, a layer of mineral oil was added. Then the samples were heated at 94°C for 1 min before the temperature was reduced to 37°C for 2 min. After this the temperature was increased at a rate of 1°/10 s to a final temperature of 72°C. The samples were then incubated at 72°C for 2 min. This cycle was repeated an additional two times. The reaction was then cycled for 30 times through 94°C for 2 min; 43°C for 2 min; 72°C for 2 min, where the temperatures were reached as quickly as possible. Lastly the reactions were incubated at 72°C for 10 min before storage at 4°C.

PCR fragments were first cloned so that they could be analyzed by DNA sequencing and then were used for a probe in Southern blot experiments. The PCR reaction mixture was electrophoresed in a 2% agarose gel and a band of the appropriate size was excised. The DNA was extracted from the gel using Qiaex Gel Extraction Kit (Cat# 20020). This DNA was cloned using SureClone Kit (Pharmacia). This mixture was used to transform DH5α cells to Amp^R. After genomic fragments were identified by Southern Blot analysis, *Hin*dIII–*Sac*II (*Sso*), *Eco*RV (*Mja*) and *Kpn*I–*Sph*I (*Hvo*) fragments were cloned into pUC118 using standard methods.

RNA isolation

Cells were grown as described above to an OD_{600} of 0.25–0.35. Cells were centrifuged and resuspended in 1× minimal salts buffer [(in g/l distilled water): K_2SO_4, 3.0; NaH_2PO_4, 0.5; $MgSO_4 \cdot 7H_2O$, 0.3; $CaCl_2 \cdot 2H_2O$, 0.1 and the pH adjusted to 3.5 with H_2SO_4] (32). If the cultures were to be irradiated, this was done at 0.5 J/m^2/s for 20 s. The cultures were then centrifuged and resuspended in growth medium and incubated at 75°C with mild aeration for the appropriate time before extraction. Total RNA was extracted from 10 ml of culture using the Qiagen RNEasy Total RNA Kit (Cat# T4103).

Alignment of protein sequences

All the protein sequences were aligned using PILEUP from the GCG Suite of programs. A comparison region was defined as the minimal amino acid sequence which all the proteins had in common and was called mCor for minimal common overlap region. The percent identities in mCor were then calculated using the number of amino acids in the mCor region as the denominator for each protein. Adjustments were made at the boundaries of gaps in the sequences as described in the Results.

Protein nomenclature

Seven different proteins will be compared in this paper: two RecA proteins from *E.coli* and *Bacillus subtilis*, Rad51 and Dmc1 from *S.cerevisiae* and the three new proteins from the archaeans. In the course of describing this work, all will be referred to as RecA-like proteins. The three archaean proteins and the yeast Rad51 and Dmc1 will also be referred to as Rad51-like proteins.

Gene nomenclature

The *recA*-like genes in the Archaeans were called *radA*. This was not intended to indicate any homology to the *radA* gene of *E.coli* (33).

RESULTS

Choice of archaeans

Genomes from three different types of archaeans were screened for *RecA*- or *RAD51*-like genes. Two major groups of *Archaea*, *Crenarchaeota* and *Euryarchaeota*, have been defined by 16S rRNA phylogenies. The *Crenarchaeota* consist mainly of hyperthermophiles. One of these, *S.solfataricus*, is a sulfur-oxidizing acidophilic hyperthermophilic aerobe (34). The *Euryarchaeota* consist mainly of two groups, the methanogens and the halophiles. *Methanococcus jannaschii* and *H.volcanii* are respectively members of these two groups of *Euryarchaeota* (35,36). *Methanococcus jannaschii* was isolated from a submarine hydrothermal vent (37). It is a piezophilic, hyperthermophilic methanogen which grows 5-fold faster at 750 atm of pressure than at 8 atm at 86°C (38). It is also a strict anaerobe (37) and may have some halophilic tendencies given its habitat. *Haloferax volcanii* grows aerobically in media containing 20% NaCl at temperatures ranging from 28 to 56°C (39). A survey of GenBank reveals that genes isolated from *Sulfolobus* and *M.jannaschii* are AT-rich (>60% AT) while the genes from *H.volcanii* are GC-rich (>60% GC).

Rationale and design of PCR primers

Initially we synthesized degenerate oligonucleotide PCR primers to encode two highly conserved regions of the Rad51, Dmc1 and RecA proteins. These correspond to codons 66–73 (GPESSGKT) and codons 140–146 (VIVVDSV) of *E.coli* RecA protein. These two regions are thought to interact with ATP (13,40). The GPESSGKT sequence is the Walker 'A' motif and is thought to form the phosphate binding hole. The VIVVDSV sequence is the Walker 'B' motif and is thought to interact with an Mg^{2+} ion via a water molecule. The selection of bases for the third position of the codons in the first set of primers was adjusted for an AT-rich organism to detect a *recA*- or *RAD51*-like gene in *S.solfataricus* (*Sso*). After the sequence of the putative gene from *Sso*, which we called *radA*, was determined, highly conserved regions among the Rad51, Dmc1 and *Sso* RadA proteins which were not conserved in the *Eco* RecA protein could be seen. These sequences were chosen as the basis for PCR primers to screen the other two archaean genomes. In each case the pairs of primers were adjusted for the GC content of the organism—*M.jannaschii* (AT-rich) or *H.volcanii* (GC-rich)—by appropriate selection of bases in the third position of each codon. Both upstream (+) primers of these pairs are similar to those used for screening the *Sso* genome; they encompassed part of the Walker 'A' motif as well as upstream sequences. The amino acid sequence used to design both downstream (–) primers was YIDTEGTFR. This amino acid sequence was present in *RAD51*, *DMC1* and *Sso radA*. Based on the alignment of Story *et al.*, the corresponding region in RecA protein is FIDAEHALD and is thought to be important for catalysis (13,40).

Identification of the *radA* genes

The DNA sequences for the three *radA* genes and ~300 bp upstream and downstream of the gene have been deposited in GenBank with accession numbers U45310 (*Sso*), U45311(*Mja*) and U45312 (*Hvo*). An open reading frame analysis of the sequences in all six potential open reading frames was carried out (data not shown). With *Sso*, *Hvo* and *Mja*, only one orf is seen which encodes putative proteins similar to RecA and Rad51. We call that orf the *radA* gene. In the sequences from *Sso* and *Mja*, no other orf that could encode a protein >10 kDa overlaps the *radA* gene in either direction and register. In the *Hvo* sequence, however, a sizeable orf in the direction opposite to that of *radA* was seen. Whether or not this overlapping gene is expressed and whether or not overlapping orfs are common in the genome of *H.volcanii* remain to be determined.

We assumed that the first ATG in the open reading frame with the homology to *RAD51* to be the start codon of the *radA* gene. In the *Sso* and *Hvo radA* sequences the next ATG codon is >200 nt distant (data not shown). Using this codon would eliminate regions similar to *DMC1* and *RAD51*. In the *Mja* sequence, however, there was a cluster of three ATG codons ~100 nt distant from the first (Fig. 1). Using any one of those would have made the length of the three archaean genes more similar. Because the sequence around the most downstream ATG has an adequate RBS (see below), we chose to translate the orf beginning with this ATG for our analysis.

General features of archaeal gene expression

It is thought that promoters of archaea and eucarya are similar to each other in that they bind transcriptional and regulatory factors prior to binding RNA polymerase for transcriptional initiation (41). In the bacteria promoters bind directly to RNA polymerase which contains a component needed for promotor recognition (42). A consensus sequence for archaeal promoters has been reached through the characterization of promoters from several species. This sequence, 5′-T/C T T A T/A A-3′, is centered 26 bp upstream of the transcriptional start site and is called BoxA (42,43). Interestingly, it is the same AT-rich sequence in all archaean species regardless of their overall GC content. It has been shown that a purine-rich region just upstream of Box A may also be critical for function of some promoters (42). Archaeal transcription usually initiates with a purine (42). Archaeal translational initiation is thought to require ribosomal binding sites like that of bacteria (44,45).

Upstream region of the *S.solfataricus radA* gene: identification of transcriptional start sites by primer extension analysis

The tentative start codon for the *Sso radA* gene is shown in Figure 1. This was the first ATG in the orf whose putative protein is similar to Rad51 protein. Several nonsense codons block translation from upstream ATG codons. The sequence of nucleotides at the 3′ end of the *Sso* 16S rRNA was used to screen for a complementary upstream sequence. Such a sequence was identified as a potential ribosome binding site (Fig. 1). This six nucleotide sequence, 5′-GGTGAT-3′, is centered 9 nt upstream of the first ATG in the orf.

Two putative promoters have been located at –52 and –73 nt upstream of the start codon by their similarity to the BoxA consensus sequence (Fig. 1). Each also is located downstream of a purine-rich region. We calculated that potential transcriptional start sites would be found at –24 and –48 respectively by counting 26 bp downstream from the middle of the Box A sequence and then identifying the nearest purine.

We tested whether or not initiation of transcription occurred *in vivo* from either of these two presumptive promoters by looking

Sequences of 5' upstream regions of the three archaean *radA* genes

```
                -80        -70        -60        -50        -40        -30        -20        -10        +1
                                                  *                     *
                                                  *                   * *
Sso radA aggar ggtatttaag gcgcggagg ataaatataa aaattagaaa gtagaggtat ttttctagta taagtgactG GTGATaatag A
                  p2                          p1

                          -70        -60        -50        -40        -30        -20        -10        +1
Hvo radA        ccccacc acgtctcccg ttttcacttt catcttgctg ttaacgaggg tttatgaggg tatAGGcGGa acgactgggt A

                          -150       140        -130       -120       -110       -100       -90
Mja radA        tttgaaac atactgtagt tataaaccaa aatctttata tagtcaaata tgacaattga tgtATGataa cctcatata

                -80        -70        -60        -50        -40        -30        -20        -10        +1
                tttctttgg caatatcatc tattacctcc caatttattc atatattaac aattacaaat ttATGcttAT GGTGAtaata A
```

Figure 1. Sequences shown are the upstream regions of the *radA* genes from *S.solfataricus* (*Sso*), *H.volcanii* (*Hvo*) and *M.jannaschii* (*Mja*). The A in the putative ATG start codon of each *radA* gene is designated +1. Three other potential start codons in *Mja radA* are shown captitalized and in bold print. Potential ribosome binding sites are in capital letters and normal print. The putative promoter sequences are shown by underlining potential BoxA sequences and italicizing the upstream purine rich regions. In the case of *S.solfataricus*, 5′ ends of mRNA have been mapped (see text) and are indicated by bold face lower case letters. The number of asterisks (*) above the line is indicative of the abundance of the RNA. The consensus for the BoxA sequence is 5′-T/C T T A T/A A-3′ (42,43) and the 3′ end region of the 16S rRNA that is thought to be functionally similar to the 3′ end of the 16S rRNA gene from *E.coli* of the three archaeans are: *Sso* 3′-ACUCCACUAGG... (52), *Hvo* 3′-UCCUCCACUAGG... (53) and *Mja* 3′-CCUCCACUAG... (GenBank MVORR16SY).

for 5′ ends of mRNAs that were complementary to the *radA* gene of *S.solfataricus*. The results in Figure 2 (summarized in Fig. 1) show 5′ ends at positions –24, –26 and –45. The bands at –24 and –45 show much greater intensities than the one at –26. The two extension products corresponding to –24 and –45 show a purine at the terminal position. The 5′ end of two major extension products that end at –24 and –45 are both 28 nt downstream of the two putative BoxA sequences at –52 and –73 respectively. We therefore tentatively conclude that the two primer extension products at –23 and –45 identify nucleotides where transcriptional initiation occurs. Either of these two transcription initiation points would provide transcripts that contain the RBS and have the first ATG codon on the mRNA as the putative start codon. The other primer extension product of lower intensity terminating at position –26 may represent an alternative start site for the BoxA sequence centered at –52 nt or a breakdown product of transcripts initiating at –45 nt.

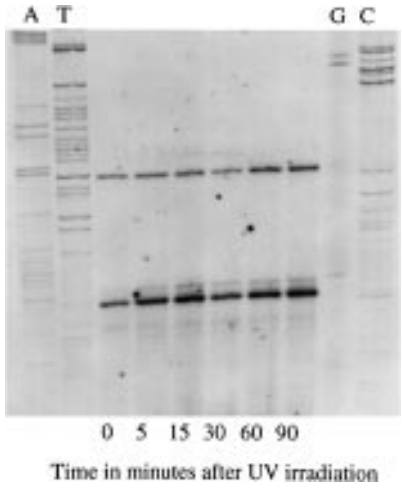

Figure 2. Primer extension analysis of mRNA isolated from *S.sofataricus*. RNA was isolated from UV-irradiated cells as a function of time after irradiation as described in Materials and Methods. DNA was generated from this RNA using an oligonucleotide primer complementary to part of the *radA* gene and reverse transcriptase. The same primer was used to generate the DNA sequencing ladders shown in the columns labelled A, T, G and C. The 5′ end of the extension products are shown in Figure 1. Methods are given in (54).

Upstream region of the *Hvo radA* gene

Using the same criteria and reasoning described for the *Sso radA* gene, we identified a potential ATG start codon for the *Hvo radA* gene (Fig. 1). Using the 3′ end of the cognate 16S rRNA as a visual probe, we identified a complementary sequence, 5′-AGGcGG-3′, centered 14 nt upstream of the ATG as a putative RBS. There are at least two sequences that have similarity to Box A. They are located at –27 and –56. The –27 Box A sequence is downstream of a purine-rich region and the –56 sequence is not. The –27 Box A sequence, although having a better match to the Box A consensus sequence than the –56 sequence and having an upstream purine-rich region, may not be the *in vivo radA* promoter because transcripts from this promoter would most likely initiate transcription in between the putative RBS and start codon. We hypothesize, therefore, that the –56 Box A sequence or another sequence further upstream may constitute the true promoter for *Hvo radA*.

Upstream region of the *Mja radA* gene

The assignment of the start codon for the *Mja radA* is complicated by the fact that there are four ATG codons in the 100 nt upstream region of the orf whose putative protein is similar to Rad51 protein. Scanning 5–15 nt upstream of each ATG reveals that only one, the fourth (at 1 nt in Fig. 1), has a potential RBS sequence 5′-TGGTGA-3′ centered 10 nt upstream of the potential start codon. A potential promoter sequence has been identified at –123 nt (Fig. 1). This sequence TTTATA is an exact match to the

Box A sequence and has a purine-rich sequence immediately upstream. A potential transcriptional start site would be at –101 nt and allow the fourth ATG codon to be the site for translational initiation.

Transcriptional initiation from the *Sso radA* promoters is not inducible by UV-irradiation

Transcriptional initiation of the *recA* and *RAD51* genes is inducible by UV irradiation (16). This regulation is mediated in a negative fashion by the LexA protein for *recA* and in a positive fashion for the *RAD51* gene by an upstream promoter enhancer sequence (16,46). Since archaeal transcriptional regulation is thought to resemble eucaryal more closely than bacterial transcriptional regulation, the regions upstream of the three putative promoters were scanned for the presence of a conserved sequence that could be regulatory in function. No obvious sequence was found. Nonetheless, since UV-inducible gene expression has been demonstrated by induction of the lytic cycle of an integrated SSV1 DNA phage of *Sulfolobus shibatae* (47), the inducibility of *radA* gene transcription by UV irradiation for *S.solfataricus* was tested. An increase in the basal level of primer extension products analyzed above after treatment with UV light would be indicative of an increase in the amount of mRNA. Using a level of UV irradiation that gave 50% survival for *S.acidcaldarius* (data not shown), the results shown in Figure 2 reveal that treatment with UV irradiation produced no or very little increase in the amount of primer extension product seen. We tentatively conclude that the gene expression of the *radA* gene of *S.solfataricus*, unlike *RAD51* or *recA*, does not increase at the level of transcription after UV irradiation.

Alignment of the putative RadA proteins with RecA and Rad51 proteins

Alignment of the three predicted RadA proteins from *S.solfataricus*, *M.jannaschii* and *H.volcanii* with the Rad51 and Dmc1 proteins from *S.cerevisiae* and the RecA proteins from *E.coli* and *B.subtilus* is shown in Figure 3. Inspection of the alignment reveals a minimal region in which all the sequences overlap. We call this the mCor. Two subgroups are evident based on whether or not the proteins have an N- or a C-terminal extension from the mCor region. The first subgroup consists of the RecA proteins from the two bacteria. These two proteins have C-terminal ends that extend beyond the mCor. The second subgroup consists of the three archaean proteins and the two yeast proteins. These contain N-terminal extensions beyond the mCor. Analysis of the spaces introduced by the computer program to align the seven sequences also shows this grouping among the seven proteins (see below).

Percent amino acid identities between the mCor regions of pairs of proteins were calculated and used as one measure of relatedness between the seven proteins (data not shown). The mCors of the RecA proteins are 65% identical and the mCors of the yeast proteins are 49% identical. The mCors of the three archaean proteins range from 42 to 49% identical with each other. With the mCors of the yeast proteins, the archaean proteins range from 37 to 46% identical and with the mCors of the bacterial proteins, they range from 16 to 20% identical. We conclude that the archaean proteins are more closely related to the proteins from eucaryotes than they are to the proteins from bacteria.

Phylogenetic analysis of the RecA-like proteins

From the topology of the tree for 16S rRNA, we expect that the RadA proteins from *Hvo* and *Mja*, both from the kingdom *Euryarchaeota* should be more similar to each other than to *Sso* RadA protein from the kingdom, *Crenarchaeota* (35). This was tested for the alignment in Figure 3 and for two variants by using the PROTPARS program of Phylip Suite 3.57. For the mCor plus N- and C-terminal sequences, or for the mCor sequences from which ambiguously aligned amino acids were eliminated as suggested by Eisen (7), we found an unrooted tree which corresponds to the 16S rRNA tree (figure not shown). When we used the mCor regions containing the ambiguously aligned amino acids, we found an unrooted tree in which *Hvo* and *Mja* are not grouped in a clade like the 16S rRNA tree (figure not shown).

DISCUSSION

Highly conserved structural motifs from the aligned sequence

The occurrence of *radA* genes whose putative proteins resemble RecA, Rad51 and Dmc1 proteins in representatives of both known archaean kingdoms implies that these organisms may carry out recombination by mechanisms similar to those of bacteria and eucaryotes. The putative proteins encoded by these *radA* genes are more similar in sequence to predicted Rad51 and Dmc1 proteins than they are to the two bacterial RecA proteins. This is consistent with the information accumulated from a wide variety of protein and RNA sequences. It establishes that in DNA metabolism, RNA metabolism and protein synthesis, archaeans are closer relatives of the eucaryotes than they are of bacteria.

In our analysis the similarity of proteins depends upon the way they are aligned. We chose the alignment produced by the PILEUP program but we actually used two other methods to align the seven proteins in Figure 3. One was the CLUSTAL3 program. The second was agreement with the alignment recently proposed by Roca and Cox (personal communication) supplemented by an alignment of part of the *Sso* RadA sequence supplied by A. Roca (personal communication). Both of the alignments so produced, however, failed to satisfy our two criteria for assessing the validity of an alignment: preservation of the continuity of putative secondary structure elements and consistency with a phylogenetic relationship mimicking that of 16S RNAs (data not shown).

α and β secondary structure elements have been proposed for *Eco* RecA protein based on its X-ray crystal structure of an ADP–protein complex (40,48). These are indicated on the alignment of proteins in Figure 3 (40,48). Only two of those elements (α helices C and D) are interrupted by gaps inserted by PILEUP to optimize the alignment. These two helices are on the outside of the molecule in the region of the protein involved with ATP hydrolysis (40,48). In the alignment produced by the two other methods, up to eight of the secondary structure elements were interrupted by gaps. In these alignments we moved amino acids at the boundaries of the gaps to restore continuity to the secondary structural elements one by one. After each move a phylogenetic tree based on the resulting alignment was drawn by PROTPARS. In neither case were we able to restore continuity to all but two elements without altering the tree from one that matched the 16S RNA tree to one that did not match. Consequently, we chose the PILEUP alignment to present.

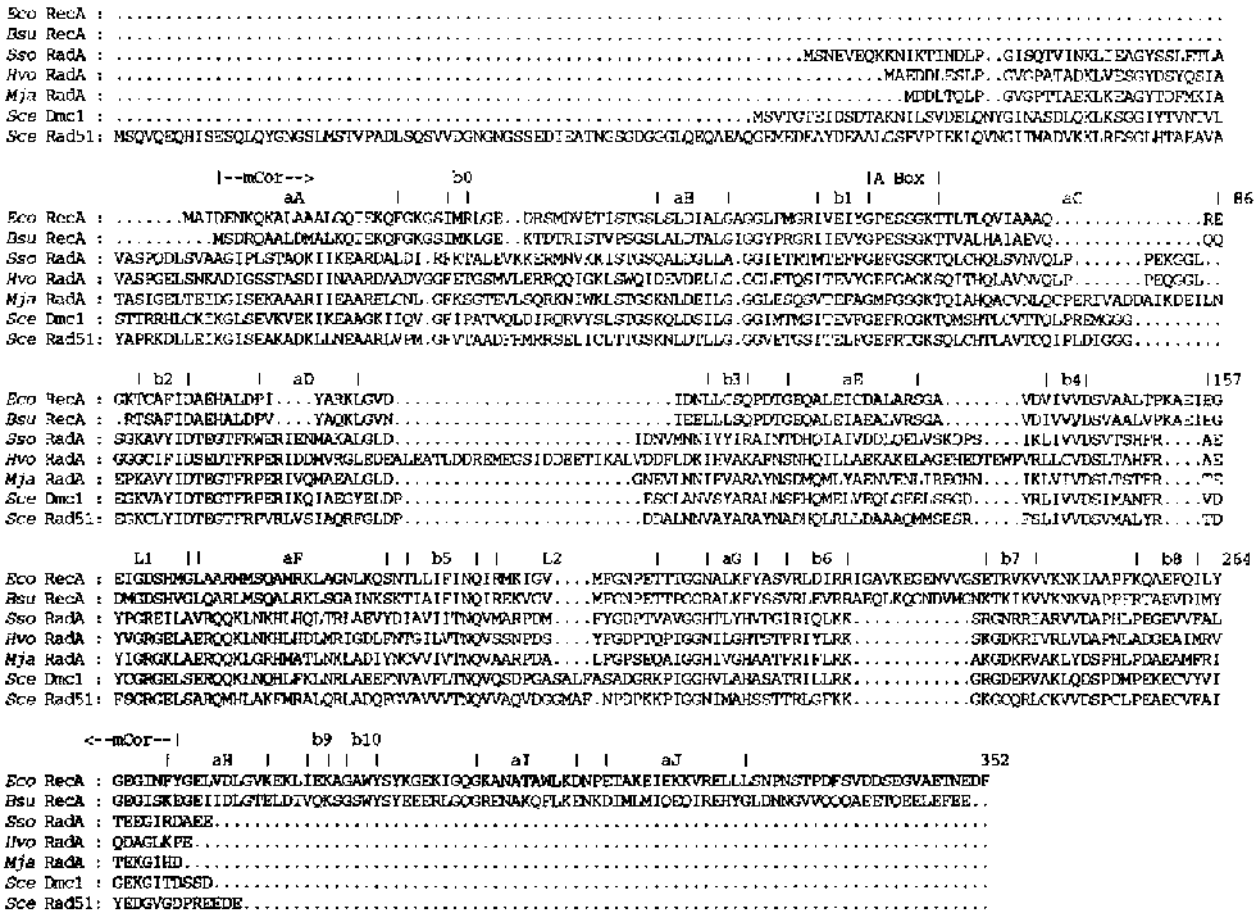

Figure 3. Alignment of the primary sequence of two bacterial RecA proteins, three archaean RadA proteins and two proteins (Dmc1 and Rad51) from the yeast *S.cerevisiae*, all of which are predicted from their corresponding DNA sequences. The alignment was produced by the program PILEUP of the GCG suite. The vertical lines above the *Eco* RecA sequence mark the boundaries of the mCor and secondary structural elements for *Eco* RecA as taken from Story *et al.* (13). Secondary structural elements are lettered and numbered as follows: aA = α helix A, b0 = β strand 0, etc. Dots indicate the absence of amino acids. The numbers at the end of the top line refer to the last amino acid position in that line for the *Eco* RecA protein.

Phylogenetic predictions of the alignment

There are nine places in the alignment where there are gaps of two or more amino acids in some of the sequences (Fig. 4). These are potential signature sequences. For example, there are two places (numbered 2 and 4 in Fig. 4) where the two bacterial sequences lack amino acids contained by the other five sequences. There are another two places (numbered 7 and 9 in Fig. 4) where the bacterial sequences have amino acids not possessed by the other five. Thus, there are potentially four sequence regions which serve to distinguish a bacterial sequence from that of an archaean or eucaryote. Scrutiny of the 63 other bacterial RecA protein sequences listed by Roca and Cox (personal communication) and Eisen (7) show that all are similar to *Eco*-RecA and *Bsu*-RecA in these four locations. There are two regions (numbered 1 and 8 in Fig. 4) which may serve to distinguish archaean from eucaryote sequences. Two places (numbered 5 and 6 in Fig. 4) seem to mark sequences that may serve to distinguish halophilic from other archaean sequences. We have sequenced the mid-part of the *radA* genes from two other halophile species, *Halobacterium halobium* and *Haloarcula hispanica*, and find sequences very similar to that of *H.volcanii* (unpublished data). Finally, region 3 (in Fig. 4) may serve as a signature for methanogen sequences or for the *Methanococcus* genus since *Mja* RadA has nine amino acids in this region not contained by the other two archaean sequences. The mid-parts of three other methanogen *radA* genes have also been sequenced to test this prediction and the data show that region 3 is a signature sequence for two other *Methanococcus* species but not for a *Methanosarcina* species (unpublished data).

Structural predictions of the alignment

We have explored the possibility that the alignment shown in detail in Figure 3 and diagrammatically in Figure 4 might reveal structural features of the Rad51-like proteins. One way to evaluate this possibility is to mark the three dimensional structure proposed for the crystalline form of *Eco* RecA complexed with ADP (40,48) with differences in primary structure shown in the alignment. We have done this in two ways in Figure 5. First, there are three regions of the primary structure of *Eco* RecA that are shown as missing from the Rad51-like proteins in Figures 3 and 4. These 'missing' amino acids are shown in red in Figure 5. Secondly, there are four regions where the Rad51-like proteins are shown in Figures 3 and 4 to have more amino acids in their primary structure than *Eco* RecA. These are marked by highlighting the amino acid on each side of the gaps in *Eco* RecA primary

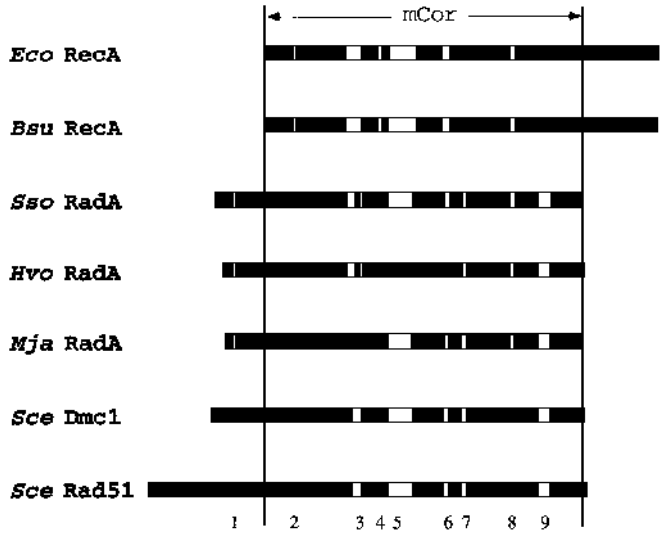

Figure 4. Diagrammatic representation of the alignment shown in Figure 3. Rectangles represent the amino acid sequences. White areas indicate the absence of amino acids and correspond to the dots in Figure 3. Below the rectangles each region of discontinuity is numbered with the exception of number 3. In this case the two gaps in the *Sso* and *Hvo* RadA sequences are considered to be a single region.

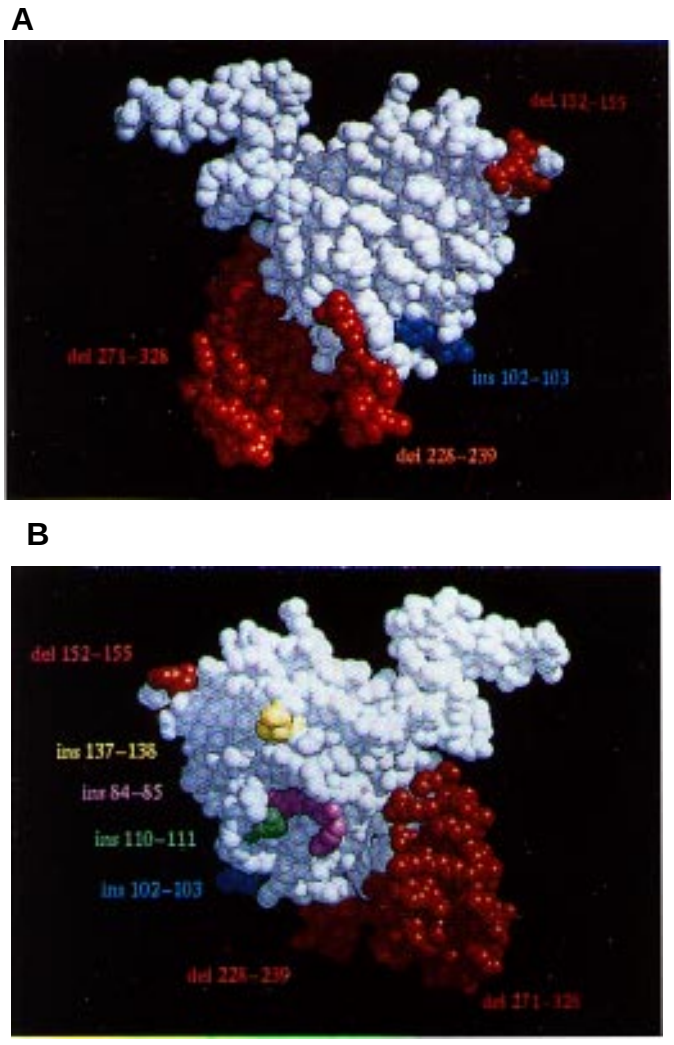

Figure 5. A space filling model of the atoms of the *Eco* RecA protein as defined by the crystal structure (48) is shown. This figure was drawn using Midasplus v2.0 software. The two parts of the figure show two faces of the RecA protein that are rotated about the vertical axis 180° from one another. The numbering system is that of the *Eco* RecA protein. In (**A**), the parts of the RecA protein not in the primary structure of the Rad51-like proteins are shown in red. They are, inclusive of amino acids 152–155, 228–239 and 271 to 328 (note: this is the last amino acid visible in the crystal structure of *Eco* RecA). These correspond to regions 4, 7 and the C-terminal region respectively of Figure 4. In (**B**), the amino acids present in the Rad51-like proteins and not in the RecA structure are illustrated by highlighting the two amino acids bordering the gaps in the *Eco* RecA sequence in Figure 3. The bordering amino acids are: magenta 84–85, blue 102–103, green 110–111 and yellow 137–138. These correspond to regions 3, 4, 5 and 6 respectively of Figure 4.

structure shown in Figures 3 and 4. We will comment briefly on features of the resulting molecular model in Figure 5 and then evaluate briefly the validity of this technique.

One of the main features of the model is the red color given to the C-terminal region indicating its absence in the Rad51-like proteins. The C-terminal region of the bacterial RecA proteins inhibits their binding to dsDNA (6,49) and is a region of interfilament interaction (48). Since the yeast Rad51 protein lacks the C-terminal domain and binds to dsDNA better than *Eco* RecA protein (14,50), we predict that the archaean proteins will do the same. Furthermore we suggest that the N-terminal domain of the yeast and archaean proteins takes the place of the C-terminal domain in interfilament interactions. Figure 3 also shows the loop region between the sixth and seventh β strands to be missing from the Rad51-like proteins (region 9 in Figs 4 and 5). A second major feature of the model in Figure 5A is that the region between the sixth and seventh β strands is physically close to the C-terminal region suggesting that its absence may be functionally correlated to the absence of the C-terminal region.

The four regions where the Rad51-like proteins have more amino acids than *Eco* RecA (numbered 3–6 in Fig. 4) lie between amino acids that are on the surface of the *Eco* RecA–ADP structure (Fig. 5B). This region is predicted to lie on the outside of the helical *Eco* RecA–DNA filament (40,48). Thus, this is a region where extra amino acids might be added without interfering with contacts with DNA and without altering the secondary structural core of the protein. These extra amino acids might be involved in interactions between Rad51-like proteins and other proteins (51).

The analysis shown in Figure 5 assumes that RecA and Rad51-like proteins are homologues with similar tertiary structures and that the alignment shown in Figure 3 has physical validity. The X-ray crystal structure of a Rad51-like protein remains to be determined, but there seems no present reason to question the assumption of similar tertiary structure. Validity of the alignment between RecA and Rad51-like proteins shown in Figure 3 is a different matter because we know of three other alignments that have been proposed for *Eco* RecA and *Sce* Rad51 proteins (10,13) (Roca and Cox, personal communication). Ignoring the N- and C-terminal regions, we find that the four alignments agree with each other in six regions consisting of 105 amino acids (data not shown). These are the unambiguously aligned amino acids which Esien (7) recommends for phylogenetic analysis. All of the numbered regions in Figure 4 lie outside of these consensus regions. We trust our alignment because it considers the archaeal RadA proteins, which show twice the

degree of identity with *Sce*Rad51 than do the RecA proteins, and the other alignments do not. However, we recognize that our model in Figure 5 is only one hypothesis and a way of thinking about the primary structural differences between RecA and Rad51-like proteins.

Function and regulation of *radA* genes

Lastly, this research has been based on the premise that homologous proteins in different organisms will have similar functions. Others have shown that a deletion mutation in the *radA* gene in *H.volcanii* leads to a UV-sensitive phenotype (Woods and Dyall-Smith, personal communication). Hence it is likely that the *radA* genes in Archaea will have a role in DNA repair and possibly also in homologous recombination. We have demonstrated that the *radA* orf in *S.solfataricus* is transcribed and that this transcription is initiated from promoter sequences that are similar to other promoters identified in Archaeans (Fig. 2). The outstanding result, however, is that the transcription of *radA* is not UV-inducible (Fig. 2). This is different from the transcriptional regulation pattern of both eucaryotic and bacterial *recA*-like genes. There are several possible reasons why we might have observed this. First, the dose of irradiation may not have been great enough to elicit a response or the response may be slower than we allowed for in our experiment. Secondly, since UV-inducible gene expression has been detected in *S.shibatae* (47), it is possible that *S.solfataricus* is mutant for this type of regulation. Lastly, it is possible that *S.solfataricus* is constitutive for *radA* expression. If the last possibility is true, and if the common ancestor of all exant life was both hyperthermophilic and constitutive for primordial *recA* expression (like *S.sofataricus*), then this raises the possibility that regulated UV-inducible *recA* (and *RAD51*) gene expression is a mesophilic adaptation. If so, then the types of UV-inducible DNA repair regulation that now appear in bacteria and eucaryotes are examples of convergent evolution.

ACKNOWLEDGEMENTS

We would like to thank the following people: Mike Dyall-Smith and Wayne Woods for supplying the *H.volcanii* genomic DNA and sharing their data before publication, Doug Clark for growing and supplying the *M.jannaschii* cells, Dennis Grogan for supplying us with an innoculum of *S.solfataricus* and sharing his results before publication, Aberto I. Roca and Mike Cox for a copy of their manuscript before publication and Ray Jacobson for instruction and help on the Midasplus program used to create the diagrams in Figure 5. This research was supported by grant AI05371 from the National Institutes of Health.

REFERENCES

1 Camerini-Otero, R.D. and Hsieh, P. (1995) *Annu. Rev. Genet.* **29**, 509–552.
2 Kowalczykowski, S.C., Dixon, D.A., Eggleston, A.K., Lauder, S.D. and Rehrauer, W.M. (1994) *Microbiol. Rev.* **58**, 401–465.
3 White, C.I. and Haber, J.E. (1990) *EMBO J.* **9**, 663–673.
4 Bishop, D.K., Park, D., Xu, L. and Kleckner, N. (1992) *Cell* **69**, 439–456.
5 Kowalczykowski, S.C. and Eggleston, A.K. (1994) *Annu. Rev. Biochem.* **63**, 991–1043.
6 Roca, A.I. and Cox, M.M. (1990) *Crit. Rev. Biochem. Mol. Biol.* **25**, 415–456.
7 Eisen, J.A. (1995) *J. Mol. Evol.* **41**,
8 Karlin, S. and Brocchieri, L. (1996) *J. Bacteriol.* **178**, 1881–1894.
9 Karlin, S., Weinstock, G.M. and Brendel, V. (1995) *J. Bacteriol.* **177**, 6881–6893.
10 Shinohara, A., Ogawa, H. and Ogawa, T. (1992) *Cell* **69**, 457–470.
11 Kans, J.A. and Mortimer, R.K. (1991) *Gene* **105**, 139–140.
12 Lovett, S.T. (1994) *Gene* **142**, 103–106.
13 Story, R.M., Bishop, D.K., Kleckner, N. and Steitz, T.A. (1993) *Science* **259**, 1892–1895.
14 Sung, P. and Robberson, D.L. (1995) *Cell* **82**, 453–461.
15 Johnson, R.D. and Symington, L.S. (1995) *Mol. Cell. Biol.* **15**, 4843–4850.
16 Basile, G., Aker, M. and Mortimer, R. (1992) *Mol. Cell. Biol.* **12**, 3235–3246.
17 Bishop, D.K. (1994) *Cell*, **79**, 1081–1092.
18 Haaf, T., Golub, E.I., Reddy, G., Radding, C.M. and Ward, D.C. (1995) *Proc. Natl Acad. Sci. USA* **92**, 2298–2302.
19 Jang, Y.K., Jin, Y.H., Kim, E.M., Fabre, F., Hong, S.H. and Park, S.D. (1994) *Gene* **142**, 207–211.
20 Shinohara, A., Ogawa, H., Matsuda, Y., Ushio, N., Ikeo, K. and Ogawa, T. (1993) *Nature Genet.* **4**, 239–243.
21 Muris, D.F., Vreekem, K., Carr, A.M., Broughton, B.C., Lehmann, A.R., Lohman, P.H. and Pastink, A. (1993) *Nucleic Acids Res.* **21**, 4586–4591.
22 Morita, T., Yoshimura, Y., Yamamoto, A., Murata, K., Mori, M., Yamamoto, H. and Matsushiro, A. (1993) *Proc. Natl Acad. Sci. USA* **90**, 6577–6580.
23 Bezzubova, O., Shinohara, A., Mueller, R.G., Ogawa, H. and Buerstedde, J.M. (1993) *Nucleic Acids Res.* **21**, 1577–1580.
24 Maeshima, K., Morimatsu, K., Shinohara, A. and Horii, T. (1995) *Gene* **160**, 195–200.
25 Akaboshi, E., Inoue, Y. and Ryo, H. (1994) *Jap. J. Genet.* **66**, 663–670.
26 Cheng, R., Baker, T.I., Cords, C.E. and Radloff, R.J. (1993) *Mutation Res.* **294**, 223–234.
27 Kobayashi, T., Hotta, Y. and Tabata, S. (1993) *Mol. Gen. Genet.* **237**, 225–232.
28 Kobayashi, T., Kobayashi, E., Sato, S., Hotta, Y., Miyazima, N., Tanaka, A. and Tabata, S. (1994) *DNA Res.* **1**, 15–26.
29 Brown, J.R. and Doolittle, W.F. (1995) *Proc. Natl Acad. Sci. USA* **92**, 2441–2445.
30 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. and Yoshida, M. (1989) *Proc. Natl Acad. Sci. USA* **86**, 6661–6665.
31 Iwabe, N., Kuma, K.-i., Hasegawa, M., Osawa, S. and Miyata, T. (1989) *Proc. Natl Acad. Sci. USA* **86**, 9355–9359.
32 Grogan, D.W. and Gunsalus, R.P. (1993) *J. Bacteriol.* **175**, 1500–1507.
33 Bachmann, B.J. and Low, K.B. (1980) *Microbiol. Reviews* **44**, 1–56.
34 Brock, T.D., Brock, K.M., Belly, R.T. and Weiss, R.L. (1972) *Arch. Microbiology*, **84**, 54–68.
35 Woese, C.R. (1987) *Microbiol. Rev.* **51**, 221–271.
36 Woese, C.R., Kandler, O. and Wheelis, M.l. (1990) *Proc. Natl Acad. Sci. USA* **87**, 4576–4579.
37 Jones, W.J., Leigh, J.A., Mayer, F., Woese, C.R. and Wolfe, R.S. (1983) *Arch. Microbiology* **136**, 254–261.
38 Miller, J.F., Shah, N.N., Nelson, C.M., Ludlow, J.M. and Clark, D.S. (1988) *Appl. Environ. Microbiol.* **54**, 3039–3042.
39 Scoarughi, G.L., Cimmino, C. and Donini, P. (1995) *J. Bacteriol.* **177**, 82–85.
40 Story, R.M. and Steitz, T.A. (1992) *Nature* **355**, 374–376.
41 Baumann, P., Qureshi, S.A. and Jackson, S.P. (1995) *TIG* **11**, 279–283.
42 Palmer, J.R. and Daniels, C.J. (1995) *J. Bacteriol.* **177**, 1844–1849.
43 Hain, J., Reiter, W.-D., Hudepohl, U. and Zillig, W. (1992) *Nucleic Acids Res.* **20**, 5423–5428.
44 Jones, J.G., Young, D.C. and DasSarma, S. (1991) *Gene* **102**, 117–122.
45 Zillig, W., Palm, P., Reiter, W.-D., Gropp, F., Puhler, G. and Klenk, H.-P. (1988) *Eur. J. Biochem.* **173**, 473–482.
46 Cole, G.M. and Mortimer, R.K. (1989) *Mol. Cell Biol.* **9**, 3314–3322.
47 Palm, P., Schleper, C., Grampp, B., Yeats, S., McWIlliam, P., Reiter, W.-D. and Zillig, W. (1991) *Virology* **185**, 242–250.
48 Story, R.M., Weber, I.T. and Steitz, T.A. (1992) *Nature* **355**, 318–325.
49 Benedict, R.C. and Kowalczykowski, S.C. (1988) *J. Biol. Chem.* **263**, 15513–15520.
50 Ogawa, T., Yu, X., Shinohara, A. and Egelman, E.H. (1993) *Science* **259**, 1896–1899.
51 Hays, S.L., Firmenich, A.A. and Berg, P. (1995) *Proc. Natl Acad. Sci. USA*, **92**, 6925–6929.
52 Olsen, G.J., Pace, N.R., Nuell, M., Kaine, B.P., Gupta, R. and Woese, C.R. (1985) *J. Mol. Evol.* **22**, 301–307.
53 Gupta, R., Lanter, J.M. and Woese, C.R. (1983) *Science* **221**, 656–659.
54 Clark, A.J., Satin, L.H. and Chu, C.C. (1994) *J. Bacteriol.* **176**, 7024–7031.