JOURNAL OF BACTERIOLOGY, Feb. 1999, p. 907–915 Vol. 181, No. 3
0021-9193/99/$04.00+0

Diversity of *radA* Genes from Cultured and Uncultured *Archaea*: Comparative Analysis of Putative RadA Proteins and Their Use as a Phylogenetic Marker

STEVEN J. SANDLER,[1]† PHILIP HUGENHOLTZ,[2] CHRISTA SCHLEPER,[3] EDWARD F. DELONG,[3]‡ NORMAN R. PACE,[1,2] AND ALVIN J. CLARK[1]*

Department of Molecular and Cell Biology, University of California, Berkeley, California 94720-3202[1]; Department of Plant and Microbial Biology, University of California, Berkeley, California 94720-3102[2]; and Marine Science Institute, University of California, Santa Barbara, California 93106[3]

Received 8 June 1998/Accepted 20 November 1998

***Archaea*-specific *radA* primers were used with PCR to amplify fragments of *radA* genes from 11 cultivated archaeal species and one marine sponge tissue sample that contained essentially an archaeal monoculture. The amino acid sequences encoded by the PCR fragments, three RadA protein sequences previously published (21), and two new complete RadA sequences were aligned with representative bacterial RecA proteins and eucaryal Rad51 and Dmc1 proteins. The alignment supported the existence of four insertions and one deletion in the archaeal and eucaryal sequences relative to the bacterial sequences. The sizes of three of the insertions were found to have taxonomic and phylogenetic significance. Comparative analysis of the RadA sequences, omitting amino acids in the insertions and deletions, shows a cladal distribution of species which mimics to a large extent that obtained by a similar analysis of archaeal 16S rRNA sequences. The PCR technique also was used to amplify fragments of 15 *radA* genes from uncultured natural sources. Phylogenetic analysis of the amino acid sequences encoded by these fragments reveals several clades with affinity, sometimes only distant, to the putative RadA proteins of several species of *Crenarcheota*. The two most deeply branching archaeal *radA* genes found had some amino acid deletion and insertion patterns characteristic of bacterial *recA* genes. Possible explanations are discussed. Finally, signature codons are presented to distinguish among RecA protein family members.**

DNA repair and recombination are fundamental molecular processes that were most likely present in the earliest life. Supporting this view is the observation that *Bacteria*, *Archaea*, and *Eucarya* all have phylogenetically related DNA repair and recombination genes that encode a crucial protein involved in synapsing two parental DNA molecules. The first and archetypal member of this protein family was identified by mutations in the *recA* gene of the bacterium *Escherichia coli* (5); therefore, the protein it encodes is called RecA. Homologues of the *E. coli recA* gene have now been found in all bacterial divisions in which they have been sought (7, 12, 20). Two budding-yeast genes, *RAD51* and *DMC1*, have been recognized to be homologues of bacterial *recA* genes (22, 26). Since then, homologues of these two genes have been found in all *Eucarya* species tested (24). Finally, Sandler et al. (21) identified genes from three archaeal genera whose putative proteins are similar to RecA proteins but are even more similar to the eucaryal Rad51 and Dmc1 proteins. These genes and proteins are called *radA* and RadA, respectively.

There were two objectives to this research: a protein structure-function objective and a phylogenetic objective. The protein structure-function objective was to determine how consistent the differences are between the RecA group of proteins and the RadA-Rad51-Dmc1 group observed by Sandler et al. (21) and Brendel et al. (3). These authors noted that the archaeal RadA proteins provided additional information that allowed a better alignment of the core regions of eucaryl Rad51 and Dmc1 proteins with the bacterial RecA proteins. Existence of a homologous core region in all of the RecA, Rad51, and Dmc1 proteins was already known (22). Also known was the fact that, relative to the core, the Rad51 and Dmc1 proteins have longer amino-terminal ends and the bacterial RecA proteins have longer carboxy-terminal ends (22). Unclear, however, was the alignment of amino acids within the core. Sandler et al. (21) and Brendel et al. (3) contended that the cores of RadA-Rad51-Dmc1 proteins have four highly conserved insertions and one deletion relative to the cores of the RecA proteins. Furthermore Sandler et al. (21) stated that there might be functional significance in the locations of the insertions. They located these insertions in the context of the X-ray crystal structure of *E. coli* RecA protein complexed with ADP and found them to be on the outside of the protein, away from the putative DNA binding site. This is consistent with the insertions not compromising the synaptase function of RadA while adding features that potentially might interact with accessory proteins.

Phylogenetic relationships among RecA proteins from greater than 65 different species of bacteria have been the focus of several studies (7, 12, 13, 17). The trees are robust and correlate well with the trees formed by 16S rRNAs from the corresponding bacteria, leading to the conclusion that RecA is a useful genetic marker for reconstructing bacterial phylogeny. Others have found that the eucaryal Dmc1 and Rad51 proteins are not good phylogenetic markers because they appear to have unequal rates of evolution (24). Thus, we wanted to

* Corresponding author. Present address: Lawrence Berkeley National Laboratory, Life Science Division, 1 Cyclotron Road, Building 74-157, Berkeley, CA 94720. Phone: (510) 486-5196. Fax: (510) 486-6690. E-mail: AJClark@LBL.gov.

† Present address: Department of Microbiology, University of Massachusetts, Amherst, MA 01003.

‡ Present address: Monterey Bay Aquarium Institute, Moss Landing, CA 95039.

TABLE 1. Abbreviations and accession numbers of sequences

Protein	Species or isolate	Abbreviation in Fig. 1–4	Accession no.	
			RecA-like protein	16S RNA
	Cultured			
	Euryarchaeota			
RadA	*Methanococcus jannaschii*	*Mc.jan*	U45311	M59126
	Methanococcus voltae[b]	*Mc.vol*	AF090200	U38461
	Methanococcus maripaludis[b,e]	*Mc.mar*	AF090204	U38486
	Methanococcus vannielii[b]	*Mc.van*	AF090203	M32222
	Methansarcinia mazei[b]	*Ms.maz*	AF090201	U20151
	Methanothermus fervidus[b,e]	*Mt.fer*	AF090202	M32222
	Picrophilius torridus[b]	*Pp.tor*	AF090205	Unknown
	Picrophilus oshimae[b,e]	*Pp.osh*	AF090206	X84901
	Haloferax volcanii	*Hf.vol*	U45312	K00421
	Halobacterium halobium[c]	*Hb.hal*	AF090196	M11583
	Haloarcula hispanica[a]	*Ha.his*	AF090199	U68541
	Archeoglobus fulgidus[b]	*Ag.ful*	AF090198	Y00275
	Crenarcheota			
	Sulfolobus solfataricus	*Sul.sol*	U45310	X90478
	Sulfolobus shibatae[b,f]	*Sul.shi*	AF090207	M32504
	Sulfolobus acidocaldarius[b,f]	*Sul.aci*	AF090208	D14876
	Pyrobacculum aerophilium[b]	*Pb.aer*	Unknown	L07510
	Cenarchaeum symbiosum[a]	*Ca.sym*	AF090197	U51469
	Uncultured			
	Norris Geyser Basin 1[b]	NGB#1	AF090209	na[g]
	Norris Geyser Basin 4[b]	NGB#4	AF090210	na
	Norris Geyser Basin 6[b]	NGB#6	AF090211	na
	Norris Geyser Basin 8[b]	NGB#8	AF090212	na
	Norris Geyser Basin 9[b]	NGB#9	AF090213	na
	Norris Geyser Basin 13[b]	NGB#13	AF090214	na
	Norris Geyser Basin 14[b]	NGB#14	AF090215	na
	Norris Geyser Basin 16[b]	NGB#16	AF090216	na
	White Creek 28[b]	WC#28	AF090217	na
	Obsidian Pool 1[b]	OP#1	AF090218	na
	Obsidian Pool 3[b]	OP#3	AF090219	na
	Obsidian Pool 4[b]	OP#4	AF090220	na
	Obsidian Pool 6[b]	OP#6	AF090221	na
	Obsidian Pool 9[b]	OP#9	AF090222	na
	Obsidian Pool 10[b]	OP#10	AF090223	na
	Antarctica 17[d,f]	Ant#17	AF090224	na
	Outgroups			
	Bacteria			
RecA	*Escherichia coli*	*E. coli*	V00328	E05005
	Deinococcus radiodurans	*Dc.rad*	U01876	M21413
	Thermotoga maritima	*Tt.mar*	L23425	M21774
	Eucarya			
RAD51	*Homo sapiens*	*H.sap*	D13804	U13369
RAD51	*Saccharomyces cerevisiae*	*S.cer*	D10023	Z75578
DMC1	*Homo sapiens*	*H.sap*	D64108	U13369
DMC1	*Saccharomyces cerevisiae*	*S.cer*	U18922	Z75578

[a] Isolated in this study by using prSJS252 and prSJS253 as oligonucleotide primers for PCR.
[b] Isolated in this study by using prSJS254 and prSJS255 as oligonucleotide primers for PCR.
[c] prSJS152 and prSJS153 were the oligonucleotides used for PCR. These primers were defined by Sandler et al. (21). The gene was cloned and sequenced from a genomic fragment by a process similar that described by Sandler et al. (21).
[d] Isolated in this study by using prSJS247 and prSJS248a as oligonucleotide primers for PCR.
[e] DNA fragments were amplified by the touchdown PCR method defined in Materials and Methods.
[f] Touchdown PCR, as described in Materials and Methods, was used, except the starting hybridization temperature was 47°C instead of 50°C.
[g] na, not applicable.

determine if the RadA proteins would be useful in deciphering the phylogeny of the *Archaea*. In addition, we wanted to test the hypothesis of Sandler et al. (21) that there might be taxonomic significance in the number of amino acids present in one or more of the insertions.

MATERIALS AND METHODS

Archaeal genomic DNA. DNA from cultured archaea was obtained for the strains listed in Table 1. Environmental DNAs were obtained from three different hot springs in Yellowstone National Park—Obsidian Pool, hot pool N10 in the Norris Geyser Basin, and hot pool O1 in the White Creek area—as previously described (11).

TABLE 2. Oligonucleotide primers for PCR

Primer	Sequence (5′ to 3′)[a]	Amino acid sequence	Avg GC content (%)
prSJS247	ACMGARTTCTWCGGMGARTTCGGMTCKGGMAA	TE F F GEFGSGK	51.6
prSJS252	ACSGARKTSTWCGGSGARTTCGGSKCSGGSAA	TE V/F Y/F GEFGSGK	62.5
prSJS254	ACWGARTTYKYWGGWGARTTYGGWWSYGGWAA	TE F F/A GEFGSGK	43.8
prSJS248a	RTCKGGKYTKGCKGAKACYTGRTTKGT	TNQV S A N/R PD	51.9
prSJS253	GTCSGGGTTSGMSAMSACCTGGTTSGT	TNQV A/S A/S N PD	63.0
prSJS255	RTCWGGYCTWGCWGCWACYTGRTTWGT	TNQV A A R PD	48.1

[a] Conforms to the standard DNA alphabet as follows: W, A or T; Y, T or C; S, G or C; M, A or C; K, T or G; R, A or G; B, T or C or G; H, A or T or C; D, A or T or G; V, A or C or G; N, A or T or C or G.

Antarctic bacterioplankton samples were collected in nearshore surface waters off of Anvers Island, Antarctica. Samples were filtered and DNA was extracted as previously described (16). A fosmid library was prepared from the archaeal symbiont *Cenarchaeum symbiosum*. Symbiont cells were initially dissociated from host tissues of the sponge *Axinella mexicana* and enriched by differential and density gradient centrifugation, as previously described (18). DNA extraction, preparation of fosmid libraries, and PCR-based screening were performed as previously described (18, 25). Individual fosmids which tested positive with RadA-specific primers were purified and used for further sequencing.

PCR, cloning, and sequencing. The primers used to amplify portions of genomic DNA in this work are listed in Table 2. For most PCR amplifications, we followed the procedure of Sandler et al. (21), which specifies an initial three cycles in which hybridization is carried out at temperatures increasing at 1°C per 10 s from 37 to 72°C. These cycles are followed by 26 cycles in which hybridization is carried out at 43°C. Six DNA preparations, however, did not yield amplification products by this procedure. For them (see Table 1) we used an alternative called "touchdown PCR" (9). In this procedure, hybridization is performed during the first cycle at 50°C for 1.5 min. After that cycle, the hybridization temperature is decreased by 0.5°C to 40°C in each of 19 successive cycles. An additional 20 cycles in which the hybridization temperature was 43°C were performed. In every cycle there was a denaturation step of 1 min at 94°C preceding and an extension step of 1 min at 71°C following hybridization. Lastly, the reactions were incubated at 71°C for 10 min before storage at 4°C. An MJ Research thermocycler, PTC-150, was used for PCR.

DNA fragments generated by PCR amplification were cloned by using the Pharmacia SureClone kit and pUC18 as a vector. Individual plasmids with single inserts were identified and subjected to automated DNA sequence analysis with vector-specific primers. The sequences of both strands of each insert were determined.

Sequence alignment and phylogenetic inference. The RadA alignment used in this study is based on our previous alignment (21). One hundred thirty-two unambiguously alignable amino acid positions, excluding the N and C termini, or 264 first- and second-codon positions of the corresponding nucleotide alignment were used in all RadA analyses. A total of 1,206 unambiguously alignable nucleotide positions were used in 16S rRNA analyses.

Maximum-likelihood (ML) analyses were conducted on the *radA* nucleotide and 16S rRNA datasets by using fastDNAml (version 1.1.1a [15]) with empirical base frequencies, optimized transition/transversion ratios (T-0.8 and 1/1 for RadA and 16S rRNA datasets, respectively), random sequence input order, and global branch swapping. Rate-corrected ML analyses were performed by using DNA rates (15) to estimate site-to-site rate variations, which were then reincorporated into the ML analysis.

Maximum-parsimony (MP) analyses were conducted on RadA amino acid and nucleotide datasets and the 16S rRNA dataset by using test version 4.0d55 of PAUP*, written by David L. Swofford. Default parameters were used in all analyses with the exception of random sequence addition with 10 repetitions per addition, TBR, and the steepest descent tree-building option with a heuristic search.

Evolutionary-distance (ED) trees were constructed from the *radA* nucleotide and 16S rRNA datasets by using PAUP* test version 4.0d55 (Kimura two-parameter or log Det distance matrix algorithms/neighbor joining) and from the RadA amino acid dataset by using PHYLIP version 3.57c (8) (Protdist [Dayhoff PAM or Kimura algorithms]/neighbor).

Transversion analysis also was performed on the 16S rRNA dataset for one set of MP and ED trees to compensate for known G+C bias in archaeal 16S rRNA datasets (29).

Bootstrap resampling (100 replicates) of the ML, MP, and ED trees was performed in all analyses to provide confidence estimates for the inferred topologies.

Nucleotide sequence accession numbers. GenBank accession numbers for the nucleotide sequences determined in this study are listed in Table 1.

RESULTS

Design of primers. Sandler et al. (21) used one set of primers to isolate a fragment of the *radA* gene of *Sulfolobus solfataricus* and another set to isolate fragments of *radA* genes of two other genera. Only 18 amino acids of the *E. coli* RecA protein lay between the equivalent positions of the primers in the *E. coli recA* gene. Our goal in this study required that we maximize the number of amino acids between the conserved regions used to design primers so that meaningful phylogenetic data could be extracted from the fragments isolated. Consequently, we chose two regions separated by 120 *E. coli* RecA amino acids for our primer design. The upstream region consists of portions of beta strand b1 and the Walker phosphate binding hole motif. The downstream region consists of portions of beta strand b5 and loop L2. The region in between comprises an integral domain of the *E. coli* RecA protein (27).

Amplification of *radA* gene fragments and alignment of their putative protein products. Three sets of primers were designed to be used in conjunction with DNAs containing high, low, and medium levels of G+C, as shown in Table 2. We used these primers to clone fragments of *radA* genes from 13 cultured species of *Archaea*. Alignments of the amino acids that are encoded by the fragments are shown in Fig. 1A and C. Evident from these panels is the occurrence of inserted or deleted amino acids in the RadA, Rad51, and Dmc1 sequences in comparison to the RecA sequences. Inserted amino acids occur between residues equivalent to residues 84 and 85, 102 and 103, 110 and 111, and 137 and 138 of *E. coli* RecA; these are designated "Indel sites" 1 to 4, respectively. Deleted are amino acids equivalent to residues 152 to 155 of *E. coli* RecA; this is called Indel site 5. The numbers of amino acids inserted at Indel site 2 and deleted at Indel site 5 are the same in all of the RadA, Rad51, and Dmc1 sequences shown in Fig. 1A and C. The numbers of amino acids at the other Indel sites differ, which indicates that multiple events may have occurred. Four sequences show 15- or 16-amino-acid insertions between *E. coli* RecA residues 84 and 85 rather than the 6 amino acids characteristic of the others (Indel site 1); we call these four sequences EL1 insertions (for extra length at site 1). These four are all from members of the genus *Methanococcus*. Four RadA sequences have between 11 and 33 amino acids between *E. coli* RecA residues 110 and 111 (Indel 3) rather than the 4 amino acids characteristic of the others; we call these sequences EL3 insertions. Three of these are from extreme halophiles. The same four sequences also have eight amino acid residues at Indel site 4 (EL4), thus differing from the three to five amino acids present in the others, except the sequence for *Archaeoglobus fulgidus*. Further information on the way these patterns of insertions and deletions are distributed phylogenetically is presented below.

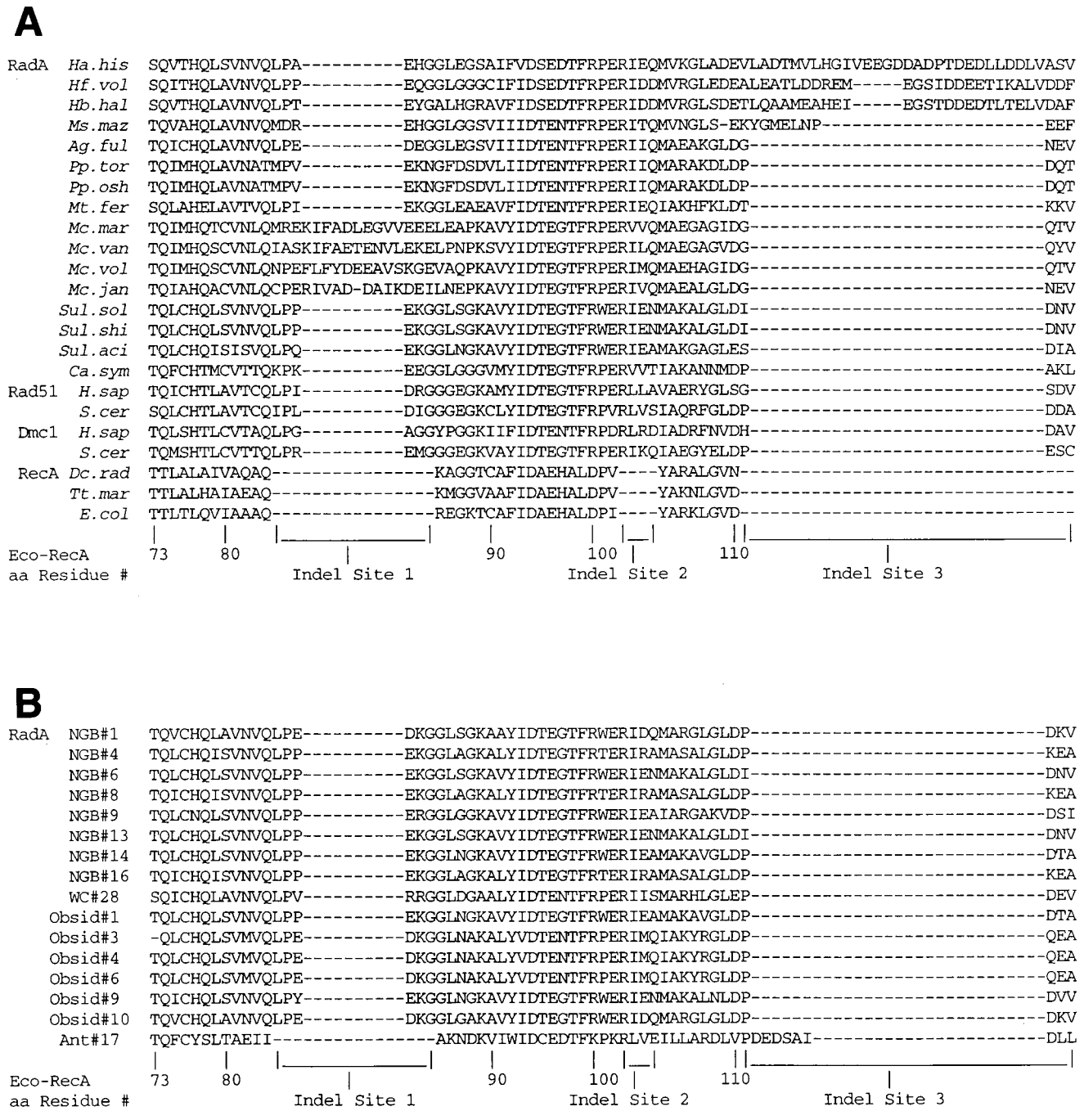

FIG. 1. Alignment of amino acids encoded by PCR-amplified fragments of *radA* genes. (A) Residues 1 to 90 of *radA* fragments from 15 cultivated and 1 enriched archaeal species. Thirteen fragments were isolated by the PCR technique and cloned. Those from *H. volcanii* (*Hf. vol*), *M. jannaschii* (*Mc. jan*), and *S. solfataricus* (*Sul. sol*) were taken from cognate parts of the sequences published by Sandler et al. (21). Cognate parts of Rad51, Dmc1, and RecA have previously been published (2, 10). (B) Residues 1 to 90 of *radA* fragments from environmental samples. These residues are shown in their proper alignment to the residues of the RecA, Rad51, and Dmc1 fragments in panel A. (C) Residues 91 to 180 of *radA* fragments from 15 cultivated and 1 enriched archaeal species. See the legend to panel A for further information. (D) Residues 91 to 180 of *radA* fragments from environmental samples shown in appropriate alignment to residues of the RecA, Rad51, and Dmc1 fragments in panel C. Abbreviations are defined in Table 1.

Also shown in Fig. 1 (panels B and D) are the sequences of 16 clones isolated from environmental DNA samples. These show much more uniformity with regard to the numbers of amino acids at Indel sites 1 and 3 than the RadA sequences from cultured and enriched species. The only exception is the sequence isolated from the Antarctic Ocean DNA sample. In this case, there is no insertion at Indel site 1, thereby distinguishing this sequence from all other RadA, Rad51, and Dmc1 sequences. On the other hand, the Antarctic Ocean sample has a longer insertion at Indel site 3, like the extreme halophiles. The phylogenetic significance of these characteristics is discussed below.

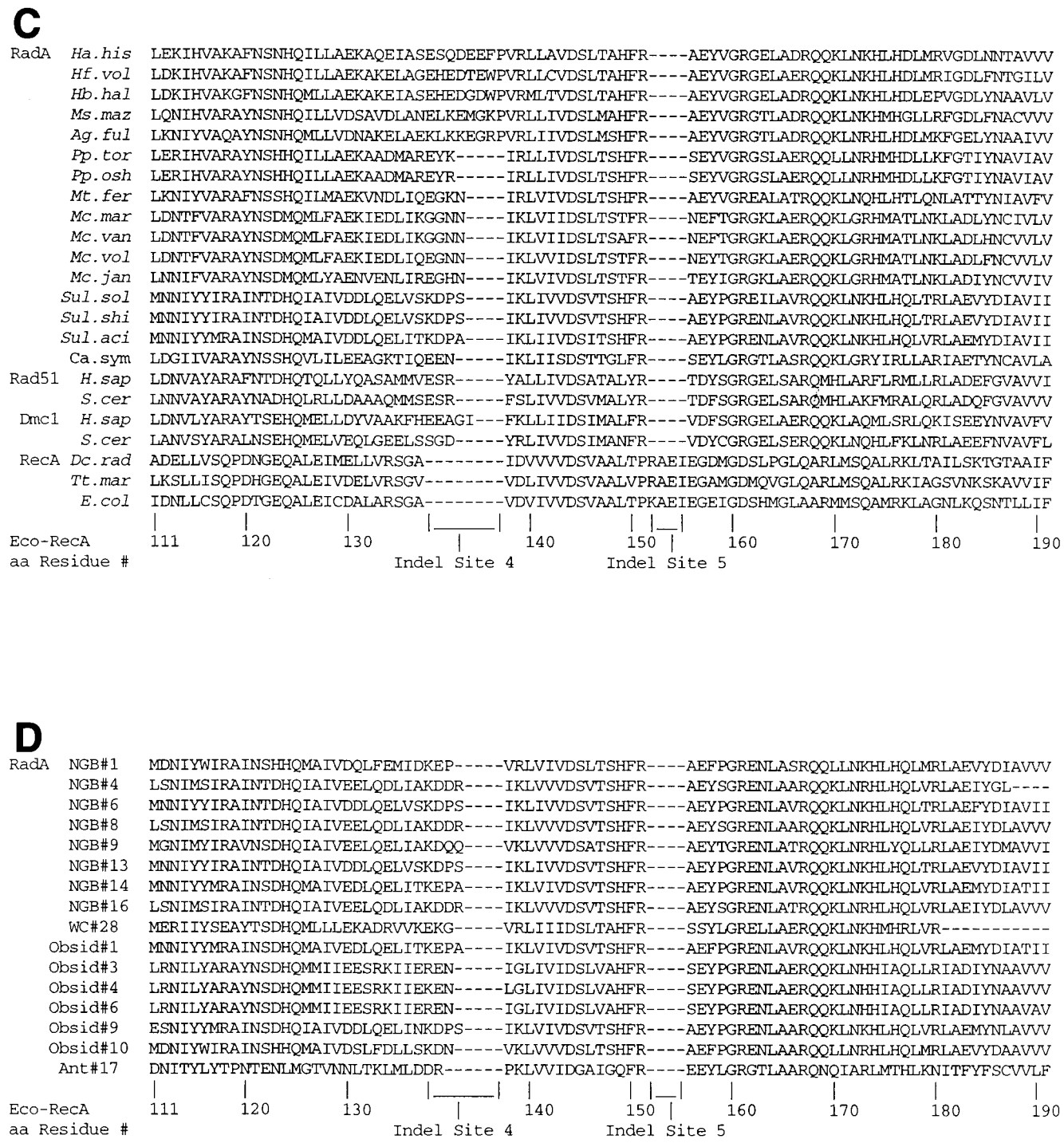

FIG. 1—*Continued.*

Sequences of complete RadA genes. Using cloned *radA* fragments (21) as hybridization probes, we cloned the entire *radA* genes of *Halobacterium halobium* and *C. symbiosum*. The protein sequences are presented in Fig. 2. Portions of the sequences, called "domains" by Brendel et al. (3), are indicated. A distinctive feature of the *C. symbiosum* sequence is that it does not have a complete domain A sequence. This distinguishes it from the *H. halobium* sequence and all other RadA, Rad51, and Dmc1 sequences so far examined (Fig. 2) (3). Another distinctive feature is the existence of a carboxy-terminal sequence similar in length, but not in sequence, to that of *E. coli* RecA protein and not possessed by any other RadA, Rad51, or Dmc1 sequence so far studied (3, 24).

Phylogeny of the Sequences: RadA versus 16S rDNA. The sequences of the RadA fragments shown in Fig. 1A and C were analyzed phylogenetically, and the phylogeny obtained was

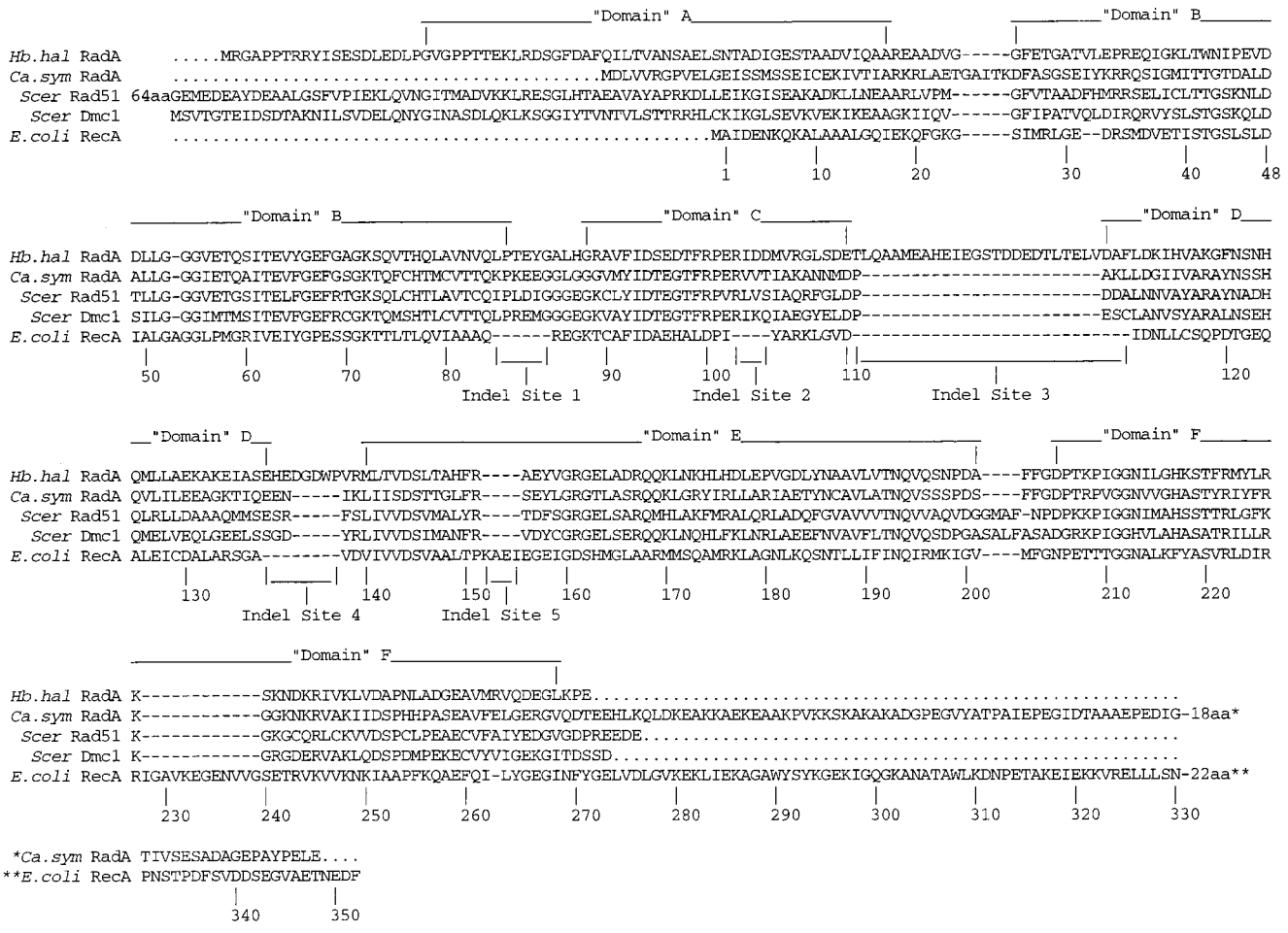

FIG. 2. Sequences of two complete putative RadA proteins (from *H. halobium* [*Hb. hal*] and *C. symbiosum* [*Ca. sym*]) aligned with *E. coli* RecA, *S. cerevisiae* Rad51, and *S. cerevisiae* Dmc1 proteins.

compared to that obtained with cognate 16S ribosomal DNA (rDNA) sequences (Fig. 3). In general there is a high degree of consistency between the two phylogenies. The shaded regions in this figure show that the two molecules reveal similar clades of halophiles, *Methanococcus* spp. and *Sulfolobus* spp., with minor branching-order discrepancies. In performing this analysis, we excluded the extra amino acids at Indel sites 1 to 5 (Fig. 1). Nonetheless, certain features of the inserts are consistent with the clade structure based on the RadA fragments. In Fig. 3, the numbers in circles refer to corresponding extra-length (EL) numbers and show the common phylogenetic ancestry of these extra-long insertions. The most notable inconsistency between the RadA and 16S rDNA phylogenies is the position of *C. symbiosum* in the *Crenarchaeota* by 16S phylogeny but in an independent lineage by RadA phylogeny. The implications of this are discussed below.

Phylogeny of environmental samples. We determined phylogenetic relationships of the *radA* nucleotide sequences extracted from environmental samples (Fig. 1B and D) with the *radA* nucleotide sequences from known archaeal species (Fig. 4). The results show that 14 of the 16 sequences cluster unambiguously with *Crenarchaeota* sequences. Two of the 14 (NGB#6 and NGB#13) are probably from representatives of the genus *Sulfolobus* because they occur within the radiation of reference *Sulfolobus* species in Fig. 4. Nine others are monophyletic with the *Sulfolobales* representatives, but at present there are insufficient reference RadA sequences to identify their generic affiliations. However, based on a proposed generic lower limit of 78% identity (data not shown), clones OP#1, NGB#14, and OP#9 may represent *Sulfolobus* species, as indicated by the dashed line in Fig. 4. Three others (OP#3, OP#4, and OP#6) are monophyletic with *Pyrobacculum aerophilium*, but again there are insufficient reference sequences at this time to conclude their generic affiliations.

Of the two sequences which do not belong to the *Crenarchaeota*, one, WC#28, is marginally supported (ca., 70% bootstrap values) as a member of the *Euryarchaeota* by evolutionary distance analysis. The final sequence, Ant#17, is very different, appears to branch independently, and shows only weak similarity to the other RadA sequence from a low-temperature crenarchaeotan, *C. symbiosum*.

Signature amino acid codons. Signature amino acids derived from aligned RecA family sequences are useful for deducing the subfamily to which each protein belongs without resort to formal phylogenetic analysis. Comparative analysis of the sequences in Fig. 1 reveals that nine amino acids can be used to distinguish five groups of RecA-like proteins. The first set of three amino acids (Table 3, group A) separates the family into three groups consisting of RecA in one group, RadA, Rad51, and Dmc1 in another, and RadB in a third. The next set of two

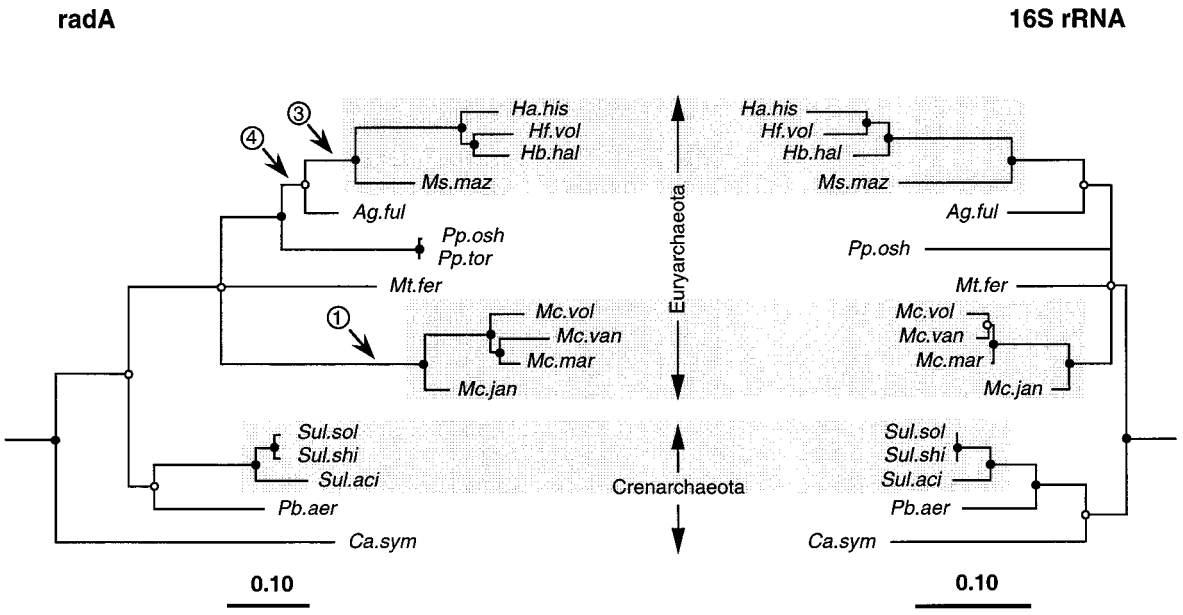

FIG. 3. Comparison of 16S rRNA and radA trees generated by rate-corrected ML analyses of nucleotide sequences (first and second codon positions only) for *radA* encoding the amino acid sequences shown in Fig. 1A and C. One additional unpublished sequence, that of *P. aerophilium*, was generously provided by S. Fitz-Gibbon (see reference 28). Branch points supported (bootstrap values, ≥75%) by most or all phylogenetic analyses (see Materials and Methods) are indicated by filled circles. Open circles indicate branch points supported by some analyses but marginally supported (bootstrap, 50 to 74%) or unsupported (bootstrap, <50%) by others (see Results and Discussion). Strongly supported monophyletic groups consistent between the two molecules are indicated by shading. Three bacterial outgroup sequences (from *E. coli*, *Deinococcus radiodurans* and *Thermotoga maritima*) were used for all 16S rRNA analyses, and four eucaryal outgroup sequences (*Homo sapiens RAD51* and *DMC1* and *S. cerevisiae RAD51* and *DMC1*) were used for all *radA* analyses. Three secondary structural features of *radA* (see Fig. 1) are indicated at branch points; that is, all sequences to the right of the indicated branch point share the secondary structural feature. 1, 3, and 4 represent extra-long inserts EL1, EL3, and EL4 at Indel sites 1, 3, and 4 respectively (see Fig. 1 and text). Abbreviations are defined in Table 1.

amino acids (Table 3, group B) separates RadA, Rad51, and Dmc1 from each other and increases the definition of RecA and RadB from each other and the other three family members. Finally, there are four additional amino acids that enhance the definitions still further.

We have applied this signature amino acid analysis to the *C. symbiosum* and Ant#17 sequences because they differ so markedly from the other RadA sequences. For example, the predicted *C. symbiosum* protein has N- and C-terminal sequences whose lengths differ from the other RadA, Rad51, and Dmc1 proteins and resemble the lengths of the N- and C-terminal sequences of bacterial RecA proteins. The *C. symbiosum* sequence, however, possesses the signature amino acids at *E. coli* RecA codon positions 74, 98, 100, 150, and 157 that all other RadA sequences possess (Table 3). The Ant#17 sequence lacks the six-or-more amino acid insertion at Indel site 1, which is characteristic of all other RadA proteins. Nonetheless, it has four of the five signature amino acids that characterize RadA sequences and at the fifth, *E. coli* RecA codon position 100, it has the basic amino acid lysine in place of the basic amino acid arginine. Thus, we think both of these sequences belong to the RadA subfamily, although we cannot rule out the possibility that they belong to a new subfamily paralogous to RadA.

TABLE 3. Codon signatures distinguishing recognized groups of *recA*-like sequences

Group	*E. coli* codon position	Codon signature				
		RecA (bacterial)[a]	RadA (archaeal)[b]	Rad51 (eucaryal)[c]	Dmc1 (eucaryal)[c]	RadB (archaeal)[d]
A	74	T	Q	Q	Q	N/T
	98	A(S)	T	T	T	G
	100	D(E)	R(K)	R	R	S
B	150	T(V)	F	Y	F	Y
	157	G	E(S)	D	D	E/K
C	122	G	S(T)	T(S/A)	S(Y)	F
	160	G(S/K)	G	G	G	N/D/R
	169	R	Q(N)	M(N/T)	Q(K)	L/K/R/A
	174	A	H(Y/L)	F	M(H)	Q

[a] Based on sequences presented by Eisen (7).
[b] Based on sequences presented in this study.
[c] Based on sequences presented by Stassen et al. (24).
[d] Based on five available RadB sequences (4, 6, 14, 19, 23).

DISCUSSION

Sandler et al. (21) proposed that three putative archaeal RadA proteins, although about 20% identical to bacterial RecA proteins, possess certain primary structural features that set them apart. We have found evidence to support this proposal by sequencing 2 complete and 11 partial *radA* genes from different archaeal species. In particular, we found that all archaeal RadA species have four regions of inserted and one region of deleted amino acids relative to bacterial RecA sequences (Fig. 1A and C). In addition, we found that the lengths of three of the inserts have taxonomic significance and distinguish particular phylogenetic clades determined by either RadA or 16S rRNA sequences (Fig. 3). Furthermore, all 13 new sequences share these diagnostic features with Rad51 and Dmc1 sequences from representative eukaryotes. This strengthens the proposal of Sandler et al. (21) that RadA is orthologous to the common ancestor of Rad51 and Dmc1.

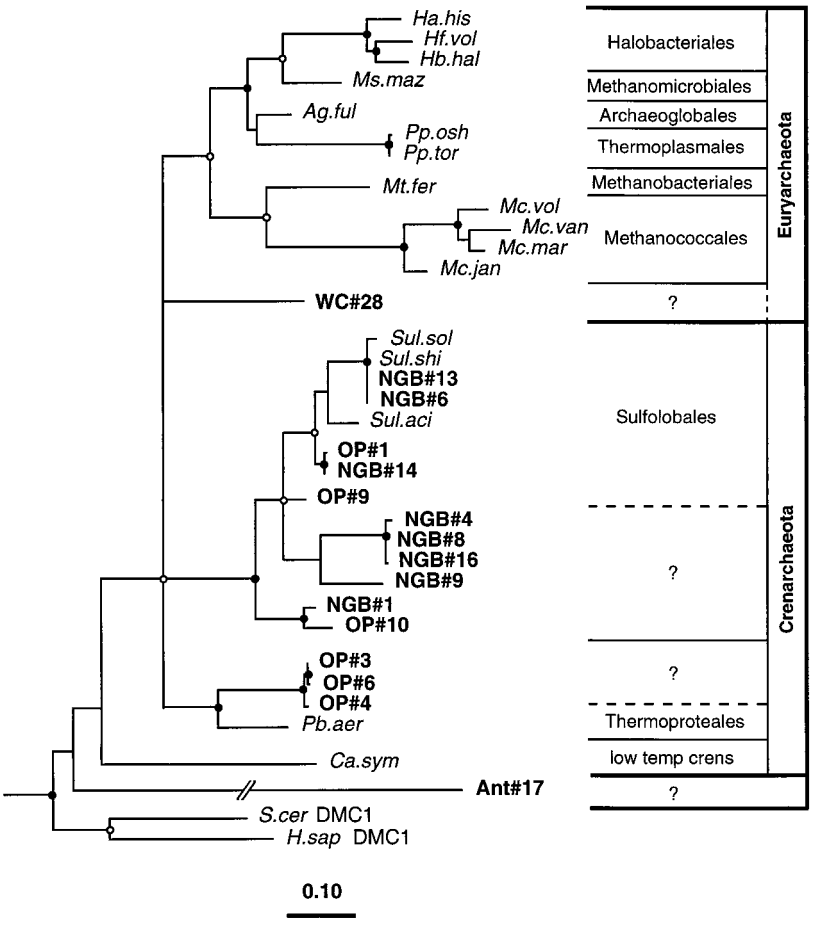

FIG. 4. Phylogenetic relationships of the 17 *radA* sequences from cultivated archaeal species and 16 environmental *radA* sequences determined by using rate-corrected ML analyses of first- and second-position nucleotide sequences. The *radA* nucleotide sequences used were those encoding the amino acid sequences in Fig. 1 with the exception of the *P. aerophilium* sequence (see the legend to Fig. 3). Eucaryal outgroup sequences were as described for Fig. 3. Filled and open circles indicating phylogenetic integrity of branch points are as in Fig. 3. Branch points without circles were not resolved (bootstrap, <50%) as specific groups in different analyses. Abbreviations are defined in Table 1.

Close correspondence of the phylogenies obtained for archaeal species with either the RadA or 16S rRNA sequences (Fig. 3) is also noteworthy because it suggests that RadA sequences can be used as an independent measure of archaeal diversity. Indeed, RadA clearly resolves the phylogenetic position of *A. fulgidus* as monophyletic with *Halobacteriales* and *Methanomicrobiales*. Furthermore, a secondary structural feature (the length of the insertion of Indel site 4) corroborates this association. This analysis independently supports the conclusion of Woese et al. (29), who based their finding on compensation for G+C bias in the 16S rDNA sequences. However, the failure of the *C. symbiosum* fragment to be monophyletic with the *Crenarchaeota* may indicate that distant phylogenetic relationships cannot be resolved by using this small RadA fragment. An alternative explanation is that we have sequenced a RadA paralog from *C. symbiosum*. We can test this alternative by screening more completely the *C. symbiosum* genome for additional RadA-like sequences.

As another test of the phylogenetic utility of the RadA molecule, we examined DNAs obtained from four different natural sources: three ecologically distinct hot pools in Yellowstone National Park (11) and the Antarctic Ocean. Fourteen of the 15 hot pool sequences cluster with members of the *Crenarcheota*. Six of these were from Obsidian Pool, whose analysis by 16S rDNA phylogeny had also shown a dominance of crenarchaeotan sequences (1). The other eight were from a pool in the Norris Geyser Basin for which there is no equivalent 16S rDNA analysis. However, we would predict that such an analysis would reveal a majority of crenarchaeotan sequences.

Two of the sequences, for *C. symbiosum* and Ant#17, stand out by their unique characteristics. Figure 4 demonstrates the novelty of these sequences by their independent branching in the archaeal tree. Although they may represent proteins paralogous to RadA, we are encouraged to anticipate the discovery of more diversity in RecA family sequences, with others possibly showing mixtures of the features of RecA and RadA proteins and perhaps representing evolutionary intermediates.

The RecA family actually contains two different member classes in the *Archaea*: *radA* and *radB*. In this study, we have been concerned mainly with *radA. radB*, however, was originally found in the genome sequence of *Methanococcus jannaschii* (4), where it was identified as a Rad51 relative. Other members of the *radB* subfamily have been found in *A. fulgidus* (14), *Pyrococcus* sp. KOD1 (19), *Pyrococcus furiosus* (6), and *Methanobacterium thermoautotrophicum* (23). Putative RadB proteins differ from RadA in two major features. First, they are only about 70% as big, having neither an N- nor a C-terminal extension. Second, their sequences differ appreciably, and we

have shown in Table 3 the signature sequences that distinguish RadA from RadB. The sequences are different enough that special primers would have to be designed to amplify them by the PCR technique. The primers that we used to amplify *radA* sequences (Table 1) would have had a very low probability of amplifying *radB* sequences. It seems worthwhile in the future to search for *radB* sequences, given the similarity of their proteins in sequence and possibly in function to the Rad55 and Rad57 proteins of *Saccharomyces cerevisiae* (17a).

ACKNOWLEDGMENTS

We thank Mike Dyall-Smith, Ken Jarrell, Shil DasSarma, Everly Conway de Macario, Patricia Hartzell, Wolfram Zillig, and Dennis Grogan for strains or purified genomic DNA used in this study. We also thank David Swofford for permission to publish results from a PAUP test version.

S.J.S. and A.J.C. were supported by grant AI05371 from the National Institutes of Health (NIH). P.H. and N.R.P. were supported by grants from the U.S. Department of Energy and NIH. E.F.D. was supported by NSF grants OCE95-29804 and OPP94-18442. C.S. was supported by a fellowship from the Deutsche Forschungsgemeinschaft.

REFERENCES

1. **Barns, S. M., C. F. Delwiche, J. D. Palmer, and N. R. Pace.** 1996. Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. Proc. Natl. Acad. Sci. USA **93:**9188–9193.
2. **Bishop, D. K., D. Park, L. Xu, and N. Kleckner.** 1992. DMC1: a meiosis-specific yeast homolog of *E. coli recA* required for recombination, synaptonemal complex formation, and cell cycle progression. Cell **69:**439–456.
3. **Brendel, V., L. Brocchieri, S. J. Sandler, A. J. Clark, and S. Karlin.** 1997. Evolutionary comparisons of RecA-like proteins across all major kingdoms of living organisms. J. Mol. Evol. **44:**528–541.
4. **Bult, C., O. White, G. J. Olsen, L. Zhou, R. D. Fleischmann, J. D. Gocayne, A. R. Kerlavage, B. A. Dougherty, J. R. Tomb, and M. D. Adams, et al.** 1996. Complete genome sequence of the methanogenic archaeon *Methanococcus jannaschii*. Science **273:**1058–1073.
5. **Clark, A. J., and A. D. Margulies.** 1965. Isolation and characterization of recombination-deficient mutants of *Escherichia coli* K-12. Proc. Natl. Acad. Sci. USA **53:**451.
6. **DiRuggiero, J., J. R. Brown, A. P. Bogert, and F. T. Robb.** DNA repair systems in archaea: momentos from the last universal common ancestor? J. Mol. Evol., in press.
7. **Eisen, J. A.** 1995. The RecA protein as a model molecule for molecular systematic studies of bacteria: comparison of RecAs and 16S rRNAs from the same species. J. Mol. Evol. **41:**1105–1123.
8. **Felsenstein, J.** 1993. PHYLIP (Phylogeny Inference Package), version 3.5c. Department of Genetics, University of Washington, Seattle.
9. **Hecker, K. H., and K. H. Roux.** 1996. High and low annealing temperatures increase both specificity and yield in touchdown and stepdown PCR. BioTechniques **20:**478–485.
10. **Horii, T., T. Ogawa, and H. Ogawa.** 1980. Organization of the *recA* gene of *Escherichia coli*. Proc. Natl. Acad. Sci. USA **77:**313–317.
11. **Hugenholtz, P., C. Pitulle, K. L. Hershberger, and N. R. Pace.** 1998. Novel division level bacterial diversity in a Yellowstone hot spring. J. Bacteriol. **180:**366–376.
12. **Karlin, S., and L. Brocchieri.** 1996. Evolutionary conservation of *recA* genes in relation to protein structure and function. J. Bacteriol. **178:**1881–1894.
13. **Karlin, S., G. M. Weinstock, and V. Brendel.** 1995. Bacterial classifications derived from *recA* protein sequence comparisons. J. Bacteriol. **177:**6881–6893.
14. **Klenk, H. P., R. A. Clayton, J. F. Tomb, O. White, K. E. Nelson, K. A. Ketchum, R. J. Dodson, M. Gwinn, E. K. Hickey, and J. D. Peterson, et al.** 1997. The complete genome sequence of the hypertheromophilic sulphate-reducing archaeon *Archeoglobus fulgidus*. Nature **390:**364–370.
15. **Maidak, B. L., G. J. Olsen, N. Larsen, R. Overbeck, M. J. McCaughey, and C. R. Woese.** 1997. The RDP (Ribosomal Database Project). Nucleic Acids Res. **25:**109–110.
16. **Massana, R., A. E. Murray, C. M. Preston, and E. F. DeLong.** 1997. Vertical distribution and phylogenetic characterization of marine planktonic archaea in the Santa Barbara Channel. Appl. Environ. Microbiol. **63:**50–56.
17. **Miller, R. V., and T. A. Kokjohn.** 1990. General microbiology of *recA*: environmental and evolutionary significance. Annu. Rev. Microbiol. **44:**365–394.

17a. **Olsen, G.** Personal communication.

18. **Preston, C. M., K. Y. Wu, T. F. Molinski, and E. F. DeLong.** 1996. A psychrophilic crenarchaeon inhabits a marine sponge: *Cenarchaeum symbiosum*, gen. nov., sp. nov. Proc. Natl. Acad. Sci. USA **93:**6241–6246.
19. **Rashid, N., M. Morikawa, and T. Imanaka.** 1996. A RecA/RAD51 homologue from a hyper-thermophilic archaeon retains the major RecA domain only. Mol. Gen. Genet. **253:**397–400.
20. **Roca, A. I., and M. M. Cox.** 1997. RecA protein: structure, function, and role in recombinational DNA repair. Prog. Nucleic Acid Res. Mol. Biol. **56:**129–223.
21. **Sandler, S. J., L. H. Satin, H. S. Samra, and A. J. Clark.** 1996. *recA*-like genes from three archaean species with putative protein products similar to Rad51 and Dmc1 proteins of the yeast *Saccharomyces cerevisiae*. Nucleic Acids Res. **24:**2125–2132.
22. **Shinohara, A., H. Ogawa, and T. Ogawa.** 1992. Rad51 protein involved in repair and recombination in *S. cerevisiae* is a RecA-like protein. Cell **69:**457–470.
23. **Smith, D. R., L. A. Doucette-Stamm, C. Deloughery, H. Lee, J. Dubois, T. Aldredge, R. Rashirzadeh, D. Blakely, R. Cook, and K. Gilbert, et al.** 1997. Complete genome sequence of *Methanobacterium thermoautotrophicum* Δ H: functional analysis and comparative genomics. J. Bacteriol. **179:**7135–7155.
24. **Stassen, N. Y., J. M. Logsdon, Jr., G. J. Vora, H. H. Offenberg, J. D. Palmer, and M. E. Zolan.** 1997. Isolation and characterization of rad51 orthologs from *Coprinus cinercus* and *Lycopersicon esculentum* and phylogenetic analysis of eukaryotic *recA* homologs. Curr. Genet. **31:**144–157.
25. **Stein, J. L. T. L. M., K. Y. Wu, H. Shizuya, and E. F. DeLong.** 1996. Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. J. Bacteriol. **178:**591–599.
26. **Story, R. M., D. K. Bishop, N. Kleckner, and T. A. Steitz.** 1993. Structural relationship of bacterial RecA proteins to recombination proteins from bacteriophage T4 and yeast. Science **259:**1892–1896.
27. **Story, R. M., and T. A. Steitz.** 1992. Structure of the recA protein-ADP complex. Nature **355:**374–376.
28. **Volkl, P., P. Markiewicz, C. Baikalov, S. Fitz-Gibbon, K. O. Stetter, and J. H. Miller.** 1996. Genomic and cDNA sequence tags of the hyperthermophilic archaean Pyrobaculum aierophilum. Nucleic Acids Res. **24:**4373–4378.
29. **Woese, C. R., L. Achenbach, P. Rouviere, and L. Mandelco.** 1991. Archaeal phylogeny: reexamination of the phylogenetic position of *Archaeoglobus fulgidus* in light of certain composition-induced artifacts. Syst. Appl. Microbiol. **14:**364–371.