JOURNAL OF BACTERIOLOGY, Dec. 1995, p. 6881–6893
0021-9193/95/$04.00+0
Copyright © 1995, American Society for Microbiology
Vol. 177, No. 23

Bacterial Classifications Derived from RecA Protein Sequence Comparisons

SAMUEL KARLIN,[1]* GEORGE M. WEINSTOCK,[2] AND VOLKER BRENDEL[1]

Department of Mathematics, Stanford University, Stanford, California 94305-2125,[1] and Department of Biochemistry & Molecular Biology, University of Texas Medical School, Houston, Texas 77225[2]

Received 27 June 1995/Accepted 29 September 1995

RecA protein sequences from 62 eubacterial sources were compared with one another and relative to one archaebacterial RecA-like and a number of eukaryotic RecA-like sequences. Pairwise similarity scores were determined by a novel method based on significant segment pair alignment. The sequences of different species were grouped on the basis of mutually high similarity scores within groups and consistency of score ranges in comparison to other groups. Following this protocol, the γ-proteobacteria can be subclassified into two major groups, those of mostly vertebrate hosts and those of mostly soil habitat. The α-proteobacterial sequences also divide into two distinct groups, whereas classification of the β-proteobacteria is more complex. The gram-positive bacterial sequences split into three groups of low and three groups of high G+C genome content. However, neither the combined low-G+C-content nor the combined high-G+C-content group nor the aggregate of all gram-positive bacteria form homogeneous groups. The mycoplasma sequences score best with the *Bacillus subtilis* sequence, consistent with their presumed origin from a gram-positive ancestor. The eukaryotic RAD proteins generally show a single high-scoring segment pair with the proteobacterial RecA sequences around the ATP-binding domain. The bacteriophage T4 UvsX protein aligns best with RecA sequences on two segments disjoint from the ATP-binding domain. The distribution of the most highly conserved regions shared between RecA and noneubacterial RecA-like sequences suggests a mosaic character and evolution of RecA. The discussion considers some questions on the validity and consistency of bacterial classifications derived from RecA sequence comparisons.

Prokaryotic molecular taxonomy is predominantly derived from sequence comparisons among 16S and 5S rRNAs (7, 13, 39, 40). The principal prokaryotic subdivision separates eubacteria and archaebacteria. The largest categories of eubacteria consist of the gram-negative [gram(−)] proteobacteria (purple bacteria) and the gram-positive [gram(+)] bacteria, both of which are generally further subclassified. Thus, on the basis of rRNA sequence comparisons the proteobacteria are divided into α, β, γ, δ, and ε types, and the gram(+) eubacteria are commonly split into two classes, characterized by genomes of high and low G+C content (39). Other distinct groups and unclassified cases among the eubacteria include the cyanobacteria, spirochetes, mycoplasmas, deinococci, planctomycetes, green sulfur bacteria, green nonsulfur bacteria, cytophagas, and thermotogas. The archaea incorporate principally extreme thermophiles, extreme halophiles, and methanogens. An important question concerns the extent to which these groupings are coherent. For example, the monophyletic character of the archaea is controversial (3, 11, 30).

Recently, evolutionary studies of prokaryotes have increasingly focused on sequence alignments of common proteins. Along these lines, Gupta and Golding (10) and Gupta and Singh (11) compared the prokaryotic 70-kDa heat shock proteins (*Escherichia coli* DnaK homologs) and inferred a phylogeny substantially different from the Woese organization of the archaea. Sequence comparisons of HSP60 (*E. coli* GroEL homologs) by Viale and Arakaki (38) and Gupta (9), of elongation factor EF-Tu by Rivera and Lake (30), of glutamate dehydrogenase by Benachenhou-Lahfa et al. (2), and of glutamine synthetase (3, 35) also gave results in conflict with the Woese scheme.

This paper has two objectives: (i) to describe a new method for assessing sequence similarity by means of calculation of significant segment pair alignment (SSPA) scores and (ii) to compare RecA sequences from 62 bacterial sources by means of the SSPA method. The SSPA scores are used to derive groupings of sequences. A group is established if the within-group SSPA scores almost invariably exceed the SSPA scores with sequences not in the group and if the scores with sequences of other groups are consistent for all members of the group.

RecA is ubiquitous in eubacteria and is among the most conserved proteins across bacterial organisms (5, 20, 31). It is a multifunctional protein contributing to homologous recombination, DNA repair, and the SOS response. Specifically, RecA binds stretches of single-stranded DNA, unwinds duplex DNA, and finds regions of homology between chromosomes in homologous recombination. In this context, RecA is required for synapsis and strand transfer. Moreover, RecA acts as an allosteric effector with respect to proteolysis of the proteins LexA and UmuD of *E. coli* and *c*I of λ phage. In particular, RecA can engender cleavage of the repressor protein LexA and induce activity of more than 15 SOS response genes. The RecA-like UvsX gene of phage T4 is involved in the initiation of DNA replication (22).

Lloyd and Sharp (24) compared 25 bacterial RecA sequences using neighbor-joining and maximum-parsimony protocols and concluded that there is "strong concordance" between their phylogenies and those derived from 16S rRNA sequences. The SSPA analysis of 63 presently available RecA sequences provides a means of discriminating among the proteobacteria and among the gram(+) bacterial sequences. In particular, the SSPA evaluations suggest division of the gram(+) bacterial sequences into six groups, three of which are low-G+C-content groups represented by the genera *Bacil-*

* Corresponding author.

Downloaded from http://jb.asm.org/ on October 28, 2020 by guest

lus, *Clostridium*, and *Lactococcus* and the other three of which are high-G+C-content groups represented by the genera *Streptomyces*, *Mycobacterium*, and *Corynebacter*. The relationships among the gram(−) purple bacteria separate the γ-proteobacteria of vertebrate hosts (e.g., *E. coli* and *Haemophilus influenzae*) from certain soil γ-proteobacteria (e.g., *Azotobacter vinelandii* and *Pseudomonas aeruginosa*). The α-proteobacteria split into two primary classes, one that includes bacteria interacting with plants (e.g., rhizobia) and the other found in free-living photosynthetic bacteria (e.g., *Rhodobacter capsulatus* and *Rhodobacter sphaeroides*). The β-proteobacteria appear to be a less coherent collection. In all SSPA comparisons of RecA the within-group scores of the conglomerate proteobacteria (α, β, γ, δ, and ε types) are invariably higher than the between-group scores of proteobacteria compared with any other group. This stands in contrast to the more diverse gram(+) set, where in some cases the sequences most similar to those of some gram(+) organisms are not from gram(+) organisms. The mycoplasmas contain three sequences which align best with those of the low-G+C-content gram(+) group represented by *Bacillus subtilis* and *Staphylococcus aureus* and score significantly lower against the low-G+C-content gram(+) group of *Lactococcus lactis* and *Streptococcus pneumoniae*, with the proteobacteria showing intermediate scores.

MATERIALS AND METHODS

Data. Protein sequences were taken from the latest updates of the SWISS-PROT and GenBank databases (1, 4), supplemented by six additional sequences received as personal communications as indicated in Table 1. The displayed groupings were defined on the basis of pairwise sequence comparisons as described in Results.

SSPA method. All sequences were pairwise compared by first identifying high-scoring segment pairs (HSSPs) according to the method of Karlin and Altschul (14) (at the 1% significance level) using the BLOSUM62 amino acid substitution scoring matrix of Henikoff and Henikoff (12). HSSPs are combined into a consistent optimal global alignment as described in the legend to Fig. 1. The gaps between consecutive HSSPs correspond to insertions or deletions or to freely varying regions that are likely not essential to the protein function and/or structure (15).

SSPA values comparing two sequences are then calculated by adding up the substitution scores in the aligned HSSPs and normalizing the resulting sum S in one of six different ways: (i) the global comparison values gmax and gmin are obtained by dividing S by the maximum and minimum of each of the two sequences scored against itself over its entire length (maximum and minimum normalizations, respectively); (ii) the local comparison values lmax and lmin are obtained by dividing S by the maximum and minimum of each of the two sequences scored against itself only over the region from the first HSSP to the last HSSP (i.e., residues a to f of Seq 1 and A to F of Seq 2 in Fig. 1); (iii) the matching segment comparison values smax and smin are obtained by dividing S by the maximum and minimum of each of the two sequences scored against itself only over the HSSPs (i.e., residues a to b′, c′ to d, and e to f of Seq 1 in Fig. 1 and similarly for Seq 2).

The global comparison values take into account different lengths of the proteins, giving lower scores to sequences not matching at the ends. The local comparison values ignore differences in the sequences beyond the region defined by the first and last HSSPs. In either case differences between the maximum and minimum normalizations generally derive from length differences at the ends or between HSSP regions. The segment comparison values indicate the quality of the HSSPs; maximum and minimum normalizations will give similar values unless the two sequences are of very different compositions in the nonmatching residues within the HSSPs. No prior alignments are needed in SSPA. The method effectively produces a gapped alignment (sometimes called bubble alignment) based on HSSP blocks (Fig. 2). These alignments can also be used for analysis of motifs, as will be exemplified in our comparisons of the RecA-like gene product UvsX of bacteriophage against the bacterial RecA sequences.

Unless indicated otherwise, the data reported here are based on gmax. All sequences being of comparable lengths and compositions, the different normalizations give essentially the same results in the case of the RecA proteins. It should be emphasized that the alignment score is based entirely on the significantly HSSPs, which represent regions of the proteins that can be most reliably aligned. Residues outside of the HSSPs are not scored at all, and in particular, no gap penalties are assigned to regions between HSSPs. The RecA proteins are highly conserved, and generally from 50 to 90% of the residues can be aligned in HSSPs.

The SSPA method produces an alignment and an assessment of the similarity for each pair of protein sequences. Groupings of sequences were developed according to the following criteria: (i) the range a to b (here a is generally at least 70) of the within-group SSPA scores for a given group (i.e., all pairwise scores among the sequences in the group) exceeds the between-group score ranges a_i to b_i (the ranges of scores between sequences in the given group and sequences in group i, where i indexes all remaining groups), that is, $b_i \leq a$ (i.e., the groups are essentially nonoverlapping); (ii) $b_i - a_i \leq \tau$, where τ is some reasonably small number (here 10); this condition expresses consistency of the scores. It should be noted that one can hardly expect to meet these criteria precisely when they are applied to real data. Nonetheless, these rules provide guidelines that help to establish an ordering involving only minimal subjective adjustments.

The pcompare program for pairwise protein sequence alignment and calculation of SSPA values may be obtained upon electronic mail request to V.B. (volker@gnomic.stanford.edu).

Interpreting HSSPs and SSPA values. SSPA values necessarily range between 0 and 1. For convenience of presentation we give all values in percent. A zero value obviously entails no HSSPs. An augmented or higher-quality HSSP generally increases the SSPA score at least four points. A global SSPA value of 60 corresponds roughly to 60% matching, allowing for conservative amino acid replacements. SSPA levels exceeding 90 reflect almost perfect identity. From examination of the HSSPs, especially alignment of their coordinate ranges, it is possible to discern segments common to some sequences and missing from other sequences. These may reflect amino acid insertions or deletions. Such events apparently have occurred in sequences of the HSP70 family (10), elongation factor EF-Tu (30), and glutamine synthetase (3).

The SSPA scores among RecA sequences are moderate to high, ranging from 43 to 98. Comparing the human RAD51 protein (putative human RecA homolog which also functions in repair and recombination) with eubacterial RecA sequences, we obtain SSPA values mostly of magnitude 3 corresponding to a single HSSP containing the classical ATP-binding motif (see Discussion). SSPA analysis comparing the *Sulfolobus solfataricus radA* gene (4a) with its eubacterial sequence homologs produces scores of 3 to 4 (one HSSP), 6 to 7 (two HSSPs), or 8 to 10 (three or, in one case, four HSSPs).

The notation G_1/G_2: x to y is used below to signify that SSPA scores between sequences of group G_1 versus sequences of group G_2 range between x and y.

RESULTS

Analysis of all pairwise SSPA evaluations reveals 15 major distinct groups, such that sequences of each group score mutually higher and consistently with sequences outside the group (Tables 1 and 2). The following classification is proposed: γ-proteobacteria group C is C1 plus C2, where C1 subsumes the subgroups C1e (enterobacteria), C1v (*Vibrio* strains), and C1h (*Haemophilus* strains) and C2 includes *Pseudomonas* and *Azotobacter* strains (subgroup C2p) as well as *Acinetobacter* strains (subgroup C2a); the β-proteobacteria constitute a rather diverse collection; the α-proteobacteria split into A1 (including soil bacteria mainly interacting with plants [e.g., rhizobia]), A2 (encompassing anoxygenic photosynthetic bacteria), and A3 (rickettsiae) sequences; the δ- and ε-proteobacteria remain distinct groups (D and E, respectively); the gram(+) bacterial sequences split into six groups; other groups comprise cyanobacteria, mycoplasmas, the *Deinococcus-Thermus* group, and others.

Gram(−) proteobacteria. SSPA analysis of RecA sequences essentially establishes the proteobacteria as an agglomerate composed of 14 subgroups.

(i) Subdivision of γ-proteobacteria. On the basis of DNA-DNA hybridizations and comparing attributes of metabolism, de Ley (7) partitioned the γ-proteobacteria into two superfamilies: I, including clusters of enterobacteria and members of the family *Vibrionaceae* largely specialized as pathogens across a variety of vertebrates, and II, consisting of strictly aerobic strains, including the (authentic) genus *Pseudomonas* and the genera *Azotobacter* and *Azomonas*. Most of superfamily II are found in soil and marine environments.

Consistent groups with respect to SSPA scores are as follows (for complete names, see Table 1): C1e includes ENTAG, SERMA, YERPE, ECOLI, ERWCA, PROMI, and PROVU (mostly enterobacteria); C1v includes VIBAN and VIBCH (genus *Vibrio*); C1h includes HAEIN (genus *Haemophilus*); C2p includes AZOVI, PSEAE, PSEFL, and PSEPU (genera

Downloaded from http://jb.asm.org/ on October 28, 2020 by guest

TABLE 1. RecA proteins

Group, subgroup, and bacterial source (*n*)[a]	SWISS-PROT abbreviation[b]	Sequence length (no. of residues)	Comment or reference
Gram(−) C (15)			Purple γ-proteobacteria
C1 (10)			Mostly associated with vertebrate hosts
C1e (7)			
Enterobacter agglomerans	ENTAG	354	
Serratia marcescens	SERMA	354	
Yersinia pestis	YERPE	356	
Escherichia coli	ECOLI	353	Identical to *Shigella flexneri*
Erwinia carotovora	ERWCA	342	Plant pathogen
Proteus mirabilis	PROMI	355	
Proteus vulgaris	PROVU	325	
C1v (2)			
Vibrio anguillarum	VIBAN	348	
Vibrio cholerae	VIBCH	354	
C1h (1), *Haemophilus influenzae*	HAEIN	354	
C2 (5)			Primarily of soil habitat
C2p (4)			
Pseudomonas aeruginosa	PSEAE	346	
Azotobacter vinelandii	AZOVI	349	
Pseudomonas putida	PSEPU	355	
Pseudomonas fluorescens	PSEFL	352	
C2a (1), *Acinetobacter calcoaceticus*	ACICA	349	
Gram(−) B1 (6)			Purple β-proteobacteria
B1m (3)			Aligns well with the C2 group
Methylomonas clara	METCL	342	99% identical to *Methylophylus methylotrophus*
Methylobacillus flagellatum	METFL	344	
Legionella pneumophila	LEGPN	348	
B1x (1), *Xanthomonas oryzae*	XANOR	355	26a
B1b (2)			
Burkholderia cepacia	BURCE	347	Also named *Pseudomonas cepacia*
Bordetella pertussis	BORPE	352	
Other gram(−) Bs (2)			Also purple β-proteobacteria
Thiobacillus ferrooxidans	THIFE	346	
Neisseria gonorrhoeae	NEIGO	348	
Gram(−) A (10)			Purple α-proteobacteria
A1 (7)			Soil habitat, some interacting with plants
Agrobacterium tumefaciens	AGRTU	363	
Rhizobium leguminosarum subsp. *phaseoli*	RHILP	360	
Rhizobium leguminosarum subsp. *viciae*	RHILV	351	
Rhizobium meliloti	RHIME	348	
Brucella abortus	BRUAB	360	Animal pathogen (bovine)
Aquaspirillum magnetotacticum	AQUMA	344	
Acetobacter polyoxogenes	ACEPO	331	
A2 (2)			Free living, photosynthetic
Rhodobacter sphaeroides	RHOSH	343	Has two chromosomes
Rhodobacter capsulatus	RHOCA	355	
A3 (1), *Rickettsia prowazekii*	RICPR	340	Satellite member of group A2
Gram(−) D (2)			Purple δ-proteobacteria
Myxococcus xanthus RecA1	MYXXA	342	13b
Myxococcus xanthus RecA2	MYXXA	358	13b
Gram(−) E (2)			
Campylobacter jejuni	CAMJE	343	
Helicobacter pylori	HELPY	347	Also named *Campylobacter pylori*
Gram(+) P1 (2)			Low G+C content
Staphylococcus aureus	STAAU	347	
Bacillus subtilis	BACSU	347	
Gram(+) P2 (1), *Clostridium perfringens*	CLOPE	352	Low G+C content (31a)
Gram(+) P3 (2)			Low G+C content
Lactococcus lactis	LACLA	365	
Streptococcus pneumoniae	STRPN	388	About 40 aa[c] longer than the typical RecA protein
Gram(+) P4 (3)			High G+C content; all three sequences about 20 aa longer than the typical RecA protein
Streptomyces ambofaciens	STRAM	372	
Streptomyces lividans	STRLI	374	
Streptomyces vinaceus	STRVI	377	
Gram(+) P5 (2)			High G+C content
Mycobacterium leprae	MYCLE	346	Length after posttranslational self-splicing
Mycobacterium tuberculosis	MYCTU	350	Length after posttranslational self-splicing

Continued on following page

Downloaded from http://jb.asm.org/ on October 28, 2020 by guest

TABLE 1—*Continued*

Group, subgroup, and bacterial source (*n*)[a]	SWISS-PROT abbreviation[b]	Sequence length (no. of residues)	Comment or reference
Gram(+) P6 (1), *Corynebacterium glutamicum*	CORGL	376	High G+C; about 20 aa longer than the typical RecA protein
Mycoplasmas (M) (3)			
Acholeplasma laidlawii	ACHLA	331	About 20 aa shorter than the typical RecA protein
Mycoplasma mycoides	MYCMY	345	
Mycoplasma pulmonis	MYCPU	339	
Cyanobacteria (S) (3)			
Anabaena variabilis	ANAVA	358	
Synechococcus sp.	SYNP7	361	5a
Synechococcus sp.	SYNP2	348	
Deinococcus-Thermus (DT) (3)			
Deinococcus radiodurans	DEIRA	363	Gram(+)
Thermus aquaticus	THEAQ	340	Gram(−)
Thermus thermophilus	THETH	340	Gram(−)
Others (6)			
Borrelia burgdorferi	BORBU	365	Spirochete (13a)
Thermotoga maritima	THEMA	356	Thermophile
Aquifex pyrophilus	AQUPY	348	Thermophile
Acidiphilium facilis	ACIFA	354	Gram(−)
Bacteroides fragilis	BACFR	318	Gram(−); shortest RecA sequence
Chlamydia trachomatis	CHLTR	352	Gram(−)

[a] Entries are listed in groups as defined in Table 2.
[b] See reference 1.
[c] aa, amino acids.

Azotobacter and [authentic] *Pseudomonas*); and C2a includes ACICA (genus *Acinetobacter*). This grouping is consistent with that of de Ley, who assigned groups C1e, C1v, and C1h to superfamily I (= C1) and combined groups C2p and C2a into superfamily II (= C2). All members of C1 are facultative aerobes and rod shaped. The C2 bacteria are free living and predominantly associated with a soil or aquatic ambience. The within-C1e-group SSPA scores range over the interval 83 to 95 (mostly 85 to 90), and between-group comparison C1e/C1v gives a range of 78 to 82 (Fig. 3; recall that the notation C1e/C1v refers to the set of all SSPA values obtained by comparing sequences of C1e with sequences of C1v). Note the next level of SSPA values, HAEIN/(C1e plus C1v): 73 to 77, suggesting HAEIN as a satellite sequence to C1e and C1v. The ACICA sequence compared with group C2p yields alignment scores of 74 to 76, considerably lower than the 85 to 91 range for within-C2p comparisons.

The alignment scores between all pairings of C1 versus C2 traverse the interval from 69 to 73, lower than the within-C1 and within-C2 SSPA values. These comparisons, in agreement with the results of de Ley (7), support the proposition that γ-proteobacteria should be divided into at least two major subgroups.

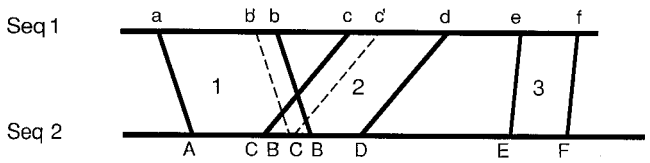


FIG. 1. SSPA. The two sequences share three HSSPs as defined by the method of Karlin and Altschul (14): residues a to b of sequence 1 (Seq 1) align with residues A to B of Seq 2, and similarly, residues c to d and e to f match up with residues C to D and E to F, respectively. The overlap of the first two HSSPs is resolved by shortening the segments to residues b′ and B′ and residues c′ and C′, respectively. The positioning of the new endpoints is determined by maximizing the sum of substitution scores for the two HSSPs combined.

(ii) Subdivision of β-proteobacteria. Examination of the SSPA values suggests the following groupings: B1m includes METCL, METME, METFL, and LEGPN; B1b includes BURCE and BORPE; B1x includes XANOR; and B1 is B1m plus B1b plus B1x. METME aligns almost perfectly with METCL and is omitted from our analysis below without a loss of information.

The within-B1m SSPA values are 78, 80, and 87; the within-B1b value is 82; and between-B1m/B1b values are 75 to 81 with a single pairing at 72, BORPE/LEGPN. Note that LEGPN is classified as a γ-proteobacterium by Woese (39). However, it scores only in the range from 68 to 71 against C1 and 73 to 76 against C2, considerably lower than the within-B1m scores. The XANOR sequence aligns with group B1 at the levels 73 to 74 and only XANOR/BORPE: 70, whereas XANOR against the C2p sequences gives the reduced scores 69 to 71. In most pairings, XANOR aligns consistently with B1, and therefore, XANOR is appended to B1. We have joined B1m with the β-proteobacterium type B1b, although one should notice that the sequences of group B1m align to the sequences of group C2 at about the same level, 73 to 76 (Fig. 3). Comparisons of B1m and C1 show scores of 68 to 74 (mostly 70 to 72), B1b/C1 alignments are diminished to 67 to 71, and XANOR/C1 scores are diminished to 65 to 70.

To maintain consistent SSPA determinations, it is appropriate to separate the β-proteobacteria NEIGO and THIFE. THIFE invariably scores higher with B1 sequences (71 to 75) than with C2 (69 to 71), providing further evidence that B1m is more β-like than γ-like. Similar inequalities apply to the NEIGO sequence, to wit, NEIGO/B1: 70 to 73 > NEIGO/C2: 68 to 69.

Detailed examination of Table 3 reveals that the B1 β-proteobacteria align with the γ- and α-proteobacteria in the 64 to 76 range and score at reduced levels relative to the gram(+) groups, especially with respect to P3 (54 to 59). The alignments with the cyanobacterial sequences cover the range from 57 to

Downloaded from http://jb.asm.org/ on October 28, 2020 by guest

```
PCOMPARE_output

Matrix: BLOSUM62;   Significance level: 0.01;   Threshold score: 49.

Upper sequence: RECA_BRUAB (360 residues)
Lower sequence: RECA_AQUMA (344 residues)

OPTIMIZED CONSISTENTLY ORDERED HSSPs:

 A) Length: 106    Score: 428

   14  VDKTKALDAA LSQIERAFGK GSIMRLGQND QVVEIETVST GSLSLDIALG VGGLPKGRIV
       +D+ KAL+AA +SQIERAFGK GSIM+LG  D QVVE E VST   L LD+ALG +GG+P+GRI+
    1  MDRQKALEAA VSQIERAFGK GSIMKLGGKD QVVETEVVST RILGLDVALG IGGVPRGRII

   74  EIYGPESSGK TTLALHTIAE AQKKGGICAF VDAEHALDPV YARKLG   119
       E+YGPESSGK TTLALH IAE AQKKGG CAF VDAEHALDP  YARKLG
   61  EVYGPESSGK TTLALHIIAE AQKKGGTCAF VDAEHALDPS YARKLG   106


 B) Length:  53    Score: 203

  122  LENLLISQPI TGEQALEITD TLVRSGAIDV LVVDSVAALT PRAEIEGEMG DSH   174
       L+ LLIS+P   GEQALEI D TLVR GA+DV LVVDSVAAL  PR E+EGEMG D+H
  108  LDELLISEPD AGEQALEIAD TLVRPGAVDV LVVDSVAALV PRGELEGEMG DNH   160


 C) Length: 168    Score: 707

  175  GLQARLMSQA VRKLTGSISR SNCMVIFINQ IRMKIGVMFG SPETTTGGNA LKFYASVRLD
       GL ARLMSQA +RKLTGS+S+ S  +VIFINQ IRMKIGVMFG +PETTTGGNA LKFYASVR++
  162  GLHARLMSQA LRKLTGSVSK SKTIVIFINQ IRMKIGVMFG NPETTTGGNA LKFYASVRME

  235  IRRIGSIKER DEVVGNQTRV KVVKNKLAPP FKQVEFDIMY GAGVSKVGEL VDLGVKAGVV
       IRR+G+IK+R DEVVGNQTRV KVVKNKLAPP FK V+FDIMY G G+SK+GEL +DLGVKA VV
  222  IRRVGAIKDR DEVVGNQTRV KVVKNKLAPP FKVVDFDIMY GEGISKMGEL IDLGVKANVV

  295  EKSGAWFSYN SQRLGQGREN AKQYLKDNPE VAREIETTLR QNAGLIAE   342
       +KSGAWFSYN S R+GQGREN AKQ+L+DNP  +A EIE  +R QNAGLI+E
  282  KKSGAWFSYN STRIGQGREN AKQFLRDNPA MAAEIEGAIR QNAGLISE   329

Bubble alignment:            A       B       C
                         13_    _2_     _0_     _18
                            \  /   \   /   \   /
                             106     53     168
                            _/  \_  _/  \_  _/  \_
                           0       1       1       15

RECA_BRUAB:   selfscore= 1766
              matching region:   14 to  342 (selfscore= 1611)
              HSSPs only selfscore= 1600
RECA_AQUMA:   selfscore= 1688
              matching region:    1 to  329 (selfscore= 1616)
              HSSPs only selfscore= 1607
Aggregate HSSP score:    1338

SSPA values:   0.76/0.79   0.83/0.83   0.83/0.84
```

FIG. 2. Illustration of the SSPA method comparing the RecA sequences of BRUAB and AQUMA. Three HSSPs can be consistently aligned. The center line of the bubble alignment gives the length of the three HSSPs and the number of unaligned residues in each sequence as shown above (BRUAB) and below (AQUMA). The six SSPA values are calculated as 1,338/1,766 = 0.76, 1,338/1,688 = 0.79, 1,338/1,616 = 0.83, 1,338/1,611 = 0.83, 1,338/1,607 = 0.83, and 1,338/1,600 = 0.84 (see text).

61. The SSPA values of β-proteobacteria against the unclassified sequences are relatively low, 53 to 60.

Overall, the β-proteobacteria have few common features and appear to be quite heterogeneous (7).

(iii) Subdivision of α-proteobacteria. A1 includes RHIME, RHILV, RHILP, BRUAB, AGRTU, ACEPO, and AQUMA, and A2 includes RHOSH and RHOCA. The A1 bacterial sequences are all in the α-proteobacterial division (39) with predominantly a soil habitat and generally interacting with plants, especially the *Rhizobium* family. BRUAB, a soil bacterium, is a known pathogen of cattle. Several fix nitrogen. *Agrobacterium tumefaciens* is a free-living plant pathogen with wide host range and does not fix nitrogen. The within-group A1 SSPA values record the high scores 73 to 93. ACEPO aligns less well in this group at the level of 75 to 78 but scores consonantly with A1 relative to sequences outside the group. AQUMA (prodigious in iron uptake from soil), like ACEPO, scores with other members of A1 at the level of 73 to 79. Comparisons of the sequences of A1 against the sequences of C yield the reduced SSPA values 64 to 69.

The members of the α-proteobacterial group A2, consisting of RHOSH and RHOCA, both free living, photosynthetic, and capable of nitrogen fixation, align mutually strongly (SSPA value, 82). The between-A1 and -A2 determinations have the relatively high scores of 70 to 76. The α-proteobacterial sequence RICPR can be regarded as an independent satellite sequence of both A1 and A2, since the corresponding SSPA values achieve the next levels—RICPR/A1: 69 to 71 (except RICPR/BRUAB is higher at 74) and RICPR/A2: 72 to 74. The A sequences (A1 plus A2 plus RICPR) align consonantly to sequences of C (SSPA values, 64 to 69).

(iv) δ-Proteobacterium *Myxococcus xanthus* RecA sequences.

Downloaded from http://jb.asm.org/ on October 28, 2020 by guest

TABLE 2. Classification of RecA protein sequences based on SSPA similarity scores

SSPA-derived[a]		No. of available sequences	Within-group SSPA score(s)	Description	Other classifications (reference)
Group	Subgroup				
C1		10	73–95		Purple γ-proteobacteria (39)
	C1e	7	83–95	Enterobacteria	γ1 (7)
	C1v	2	87	*Vibrionaceae* strains	γ2 (7)
	C1h	1	—[b]	*H. influenzae*	γ3 (7)
C2		5	74–91		
	C2p	4	85–91	*Pseudomonas* and *Azotobacter* strains	Purple γ-proteobacteria (39), γ4 (7)
	C2a	1	—	*Acinetobacter* strain	γ5 (7)
B1		6	70–87		
	B1m	3	78–87	Methylobacteria and *Legionella* strain	γ2 (*Legionella* strain) (39)
	B1x	1	—	*Xanthomonas* strain	
	B1b	2	82	*Burkholderia* and *Bordetella* strains	
A1		7	73–93	Rhizobacteria and agrobacteria	Purple α-proteobacteria (39)
A2		2	82	*Rhodobacter* strains	Purple α-proteobacteria (39)
A3		1	—	*Rickettsia* strain	Purple α-proteobacteria (39)
E		2	80	*Campylobacter* and *Helicobacter* strains	Purple ε-proteobacteria
P1		2	73	*Staphylococcus* and *Bacillus* strains	Low-G+C-content gram(+) bacteria (39)
P2		1	—	*C. perfringens*	Low-G+C-content gram(+) bacterium (39)
P3		2	71	*Lactococcus* and *Streptococcus* strains	Low-G+C-content gram(+) bacteria (39)
P4		3	91–95	*Streptomyces* strains	High-G+C-content gram(+) bacteria (39)
P5		2	90	*Mycobacterium* strains	High-G+C-content gram(+) bacteria (39)
P6		1	—	*Corynebacterium* strain	High-G+C-content gram(+) bacterium (39)
DT		3	64–93	*D. radiodurans* and *Thermus* sp.	*Deinococcus-Thermus* group
S		3	74–79	Cyanobacteria	

[a] Groups and subgroups were formed on the basis of mutually high SSPA scores (about 70 or greater) and consistency in terms of scoring against other groups.
[b] —, not applicable.

Intriguingly, two independent RecA sequences (RecA1 and RecA2) were cloned from *M. xanthus* (29). The mutual SSPA score between these two sequences is only 65. Alignments with all proteobacterial sequences give RecA2/proteobacteria: 62 to 68 > RecA1/proteobacteria: 59 to 66, with individual differences of 1 to 8 (mostly 3 to 5). The highest scores (around 68) are attained relative to the β-proteobacterium group B1 and α-proteobacterium group A1. With respect to gram(+) bacterial sequences, the SSPA assessments tend to be less. Alignments of the *M. xanthus* sequences with cyanobacterial sequences are rather reduced (levels of 53 to 59). The SSPA scores relative to unclassified sequences fall in the range of 60 to 62 for both *M. xanthus* sequences.

(v) ε-Proteobacteria. CAMJE is classified as a prototype ε-proteobacterium. This sequence produces SSPA values relative to most classical (proteobacterial) bacterial sequences in

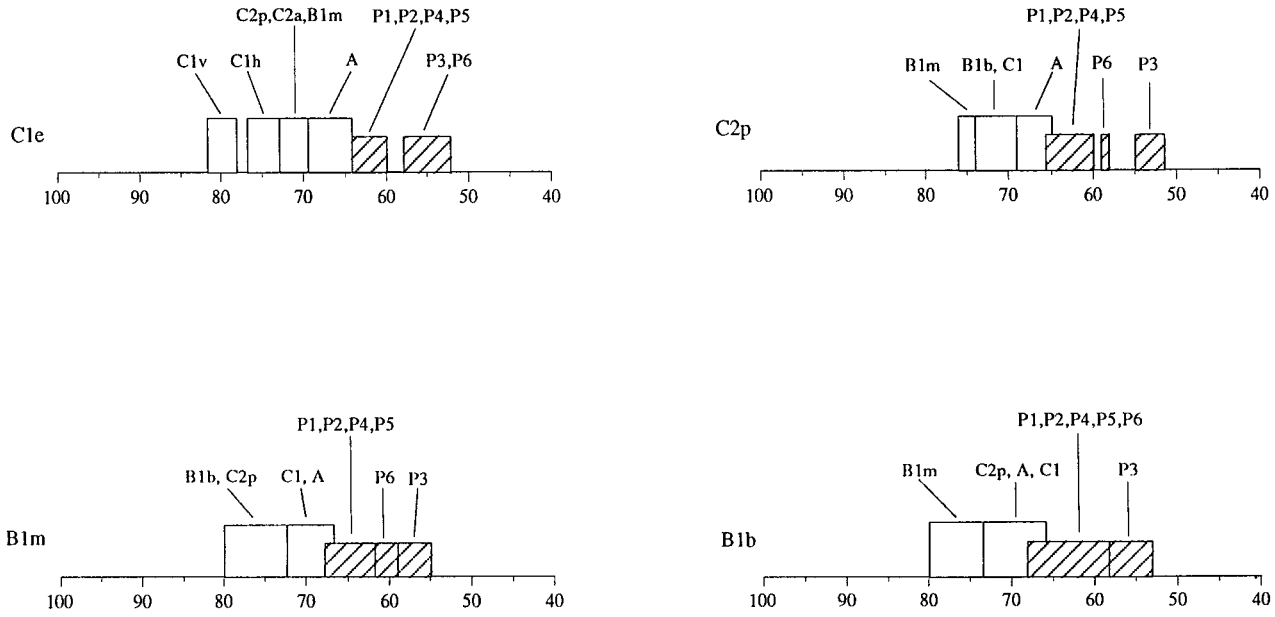


FIG. 3. Schematic presentation of between-group SSPA value ranges: comparisons relative to the γ-proteobacterial group C1e and among the gram(−) groups C2p, B1m, and B1b. The ranges are given relative to the groups indicated on the left as standards (see Table 3). The B subgroup B1m could also be put together with the C subgroup C2p.

Downloaded from http://jb.asm.org/ on October 28, 2020 by guest

TABLE 3. Within- and between-group SSPA value ranges[a]

Group	SSPA value range (no. of sequences) for group:													
	C1 (10)	C2 (5)	B1 (6)	A1 (7)	A2 (2)	E (2)	P1 (2)	P2 (1)	P3 (2)	P4 (3)	P5 (2)	P6 (1)	S (3)	DT (3)
C1	73–95	69–73	65–74	63–69	64–68	59–63	60–64	63–66	52–58	59–62	60–63	55–58	55–60	56–61
C2		74–91	69–76	65–69	67–70	60–66	60–66	65–66	52–56	62–64	64–65	58–59	58–63	54–62
B1			70–87	64–72	66–72	60–68	58–68	65–68	54–59	61–64	63–67	56–62	57–61	58–65
A1				73–93	70–76	63–67	62–66	65–70	54–58	61–65	63–68	58–60	54–62	57–63
A2					82–82	65–68	62–66	67–69	55–59	61–63	63–64	57–58	56–60	58–62
E						80–80	63–66	65–65	52–55	59–61	63–64	58–59	56–58	54–63
P1							73–73	67–70	60–63	62–64	66–67	60–63	57–62	56–62
P2								—[b]	57–59	62–63	67–67	59–59	59–60	58–63
P3									71–71	57–59	55–61	56–58	51–56	53–55
P4										91–95	74–76	69–70	59–61	55–58
P5											90–90	71–71	61–64	57–61
P6												—	56–58	55–57
S													74–79	50–58
DT														64–93

[a] Based on gmax. Groups are abbreviated according to the notation in Table 2.
[b] —, not determined.

the range of 60 to 70. The highest SSPA scores occur with B1m. The recently sequenced *Helicobacter pylori* RecA is highly similar to the CAMJE sequence (SSPA score, 80) and gives consistent scores against the other groups (Table 3).

Gram(+) organisms. On the basis of the SSPA values, the gram(+) bacterial sequences are divided into six groups: P1 includes BACSU (G+C content, 43%) and STAAU (36%); P2 includes CLOPE (30%); P3 includes LACLA (36%) and STRPN (40%); P4 includes STRAM (70%), STRLI (72%), and STRVI (72%); P5 includes MYCLE (65%) and MYCTU (65%); and P6 includes CORGL (55%). Those in groups P1 to P3 are low-G+C-content genomes; those in groups P4 to P6 are high-G+C-content genomes.

This partition conforms to the criterion of high versus low G+C content, which by itself seems to be too crude but warrants division into further subgroups. These subdivisions of the gram(+) strains are also appropriate with respect to evolutionary comparisons of other proteins, including glutamine synthetase, DnaA, and glutamate dehydrogenase (data not shown).

The within-group gram(+) SSPA values are as follows: within P1, 73; within P3, 71; within P4, 91 to 95; and within P5, 90. The alignments are substantial, with comparisons between the high-G+C-content gram(+) groups yielding P4/P5: 74 to 76 and P6/(P4 plus P5): 69 to 71. The sequence alignments of the low-G+C-content groups are somewhat erratic, with the sequences of P3 most deviant: P1/P2: 67 and 70; P1/P3: 60 to 63; P2/P3: 57 and 59. The low versus high between-group SSPA scores show the inequalities P1/P5: 66 to 67 $\gg$ P1/P4: 62 and 63 and P1/P6: 60 to 63 $>$ P2/P6: 59 and the reverse orderings P2/P5: 67 $\gg$ P1/P4: 62 to 64. The P3 sequences align relatively weakly with the G+C-rich groups P4 plus P5 plus P6 at the levels 55 to 61 (Fig. 4).

In comparisons with the proteobacterial sequences, the alignment values (Table 3) satisfy the dominance orderings P3/proteobacteria: 52 to 59 $<$ P1/proteobacteria: 58 to 68 (mostly 62 to 66) and P2/proteobacteria: 63 to 70. Thus, the gram(+) P1 group shows significantly improved alignments (5 to 15 SSPA units higher) with all proteobacterial sequences compared with the gram(+) P3 group. For P1 versus P2, the individual differences of SSPA values with gram(−) bacterial sequences are 2 to 6. Moreover, LACLA/proteobacterial values are higher than STRPN/proteobacterial values (individual differences, generally 2 to 7), indicating that LACLA sequences align consistently better than STRPN with proteobacterial sequences. The strength of alignments of the proteobacterial sequences against P6 is on the low scale of RecA sequence conservation: P6/proteobacteria: 55 to 62 (mostly 59 and 60).

The comparisons show that the gram(+) groups P1, P2, P4, and P5 have better alignments with proteobacterial sequences than do P3 and P6 (Table 3). In contrast, RecA sequence comparisons involving the gram(−) proteobacteria as a conglomerate align in a coherent manner relative to other groups.

Mycoplasmas. The SSPA analysis places the mycoplasma sequence ACHLA closest to P1 (Fig. 5; SSPA score with BACSU, 69), consistent with the origin of mycoplasmas from a gram(+) organism, especially *B. subtilis* (25). Specifically, values were as follows: P1/ACHLA: 65 and 69 $>$ P2/ACHLA: 63, P5/ACHLA: 61 and 62, and P4/ACHLA: 59 to 60 $>$ P3/ACHLA: 56 and 58 and P6/ACHLA: 55. The orderings are qualitatively similar for the mycoplasmas (MYCMY and MYCPU) but with substantially reduced SSPA scores.

With all gram(−) proteobacterial groups we have SSPA values of ACHLA/proteobacteria: 57 to 64 and even lower values with the cyanobacterial sequences (55 to 57). For ACHLA compared with the unclassified sequences, the scores are $\leq$61 (exception: BACFR, 64), falling to the relatively low score of ACHLA/ACIFA: 48.

Cyanobacteria. There are three RecA cyanobacterial sequences (group S), ANAVA, SYNP2, and SYNP7. The mutual SSPA scores of S sequences are relatively high, 74 to 79. The alignment values of cyanobacterial sequences relative to all proteobacteria have the range S/proteobacteria: 54 to 61 (one value of 63). In comparisons with the gram(+) sequences, values are S/P1: 60 to 62 (one 57), S/P2: 59 to 60, S/P3: 51 to 56, S/P4: 59 to 61, S/P5: 61 to 64, and S/P6: 56 to 58, placing the cyanobacteria generally closest to the gram(+) mycobacterial P5 group (Fig. 5).

Overall, the cyanobacterial sequences align with gram(−) bacterial sequences with scores about 5 to 20 lower than the within-gram(−) alignment scores. With one exception the alignments of the cyanobacterial sequences with the singular cases have principal SSPA scores of 54 to 58, similar to the alignment scores of cyanobacteria versus gram(−) proteobacteria. The lowest SSPA value among RecA sequences is S/ACIFA: 46 to 48.

DEIRA. DEIRA originally was classified as a gram(+) bacterium (high-G+C-content group) but now is relegated to an

Downloaded from http://jb.asm.org/ on October 28, 2020 by guest

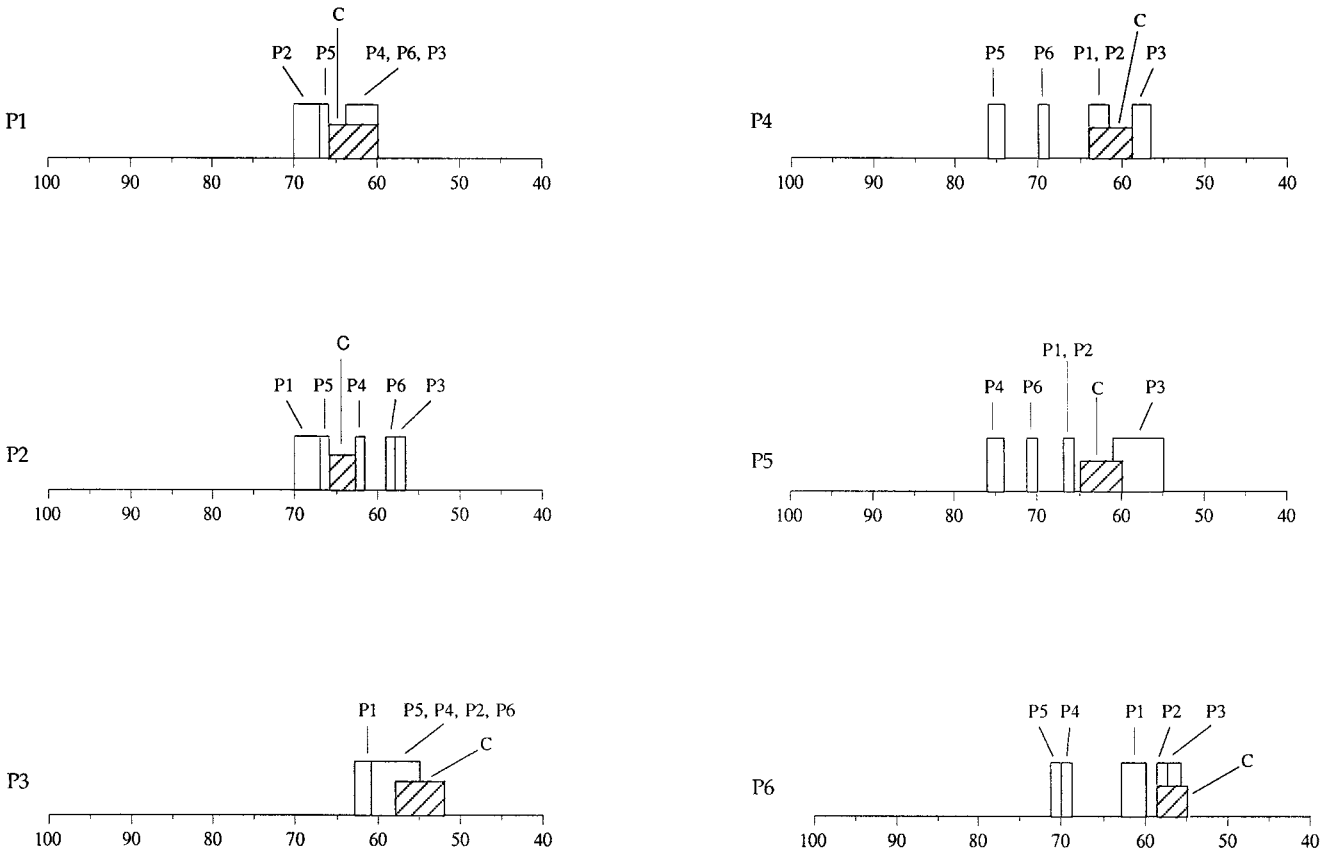


FIG. 4. Schematic presentation of between-group SSPA value ranges: comparisons among gram(+) bacteria. The ranges are given relative to the groups indicated on the left as standards (see Table 3). Notice that the gram(+) bacteria are very heterogeneous and cover a wide range of SSPA scores. In particular, the low-G+C-content groups P1, P2, and P3 are not separated from the high-G+C-content groups P4, P5, and P6. This is most evident comparing P1 and P3; whereas P1 is the closest group to P3, relative to P1, P2 is the closest group and P3 is the most distant.

outlier status as the *Deinococcus* genus. SSPA scores of DEIRA against all sequences do not exceed 60 (mostly 53 to 58) with the prominent exceptions of DEIRA/(THEAQ plus THETH): 64 to 65. By our criteria, these three sequences form the *Deinococcus-Thermus* group.

BORBU. In many respects, the spirochete BORBU is an outlier bacterium (39). This is also reflected in SSPA assessments of its RecA sequence, as the alignment scores of BORBU with other bacterial strains of eubacteria fall in the relatively low range of 48 to 58.

Thermophiles. There are four thermophile sequences among the available RecA sequences: THEMA, AQUPY, THEAQ, and THETH (Table 1). The last two are of the same genus (*Thermus*) with a very high SSPA value of 93. They are grouped with DEIRA, as discussed above, but also score moderately with the proteobacterial group B1m (SSPA score, 62 to 65). AQUPY aligns with THEAQ at the moderate level of 64, the highest SSPA attainment in all sequence pairings with AQUPY (ACEPO gives 63). THEMA scores highest with the gram(+) mycobacteria (SSPA value, 62).

Others. We summarize the alignment scores for the unclassified bacteria ACIFA and BACFR and for CHLTR. SSPA values of the ACIFA sequence against proteobacteria are 52 to 60 (mostly 54 to 56) with two exceptions, ACIFA/THIFE: 65 and ACIFA/BRUAB: 46. ACIFA versus gram(+) bacterial and cyanobacterial sequences gives the lowest alignment scores among all bacterial RecA comparisons.

When BACFR is compared with the proteobacteria, the

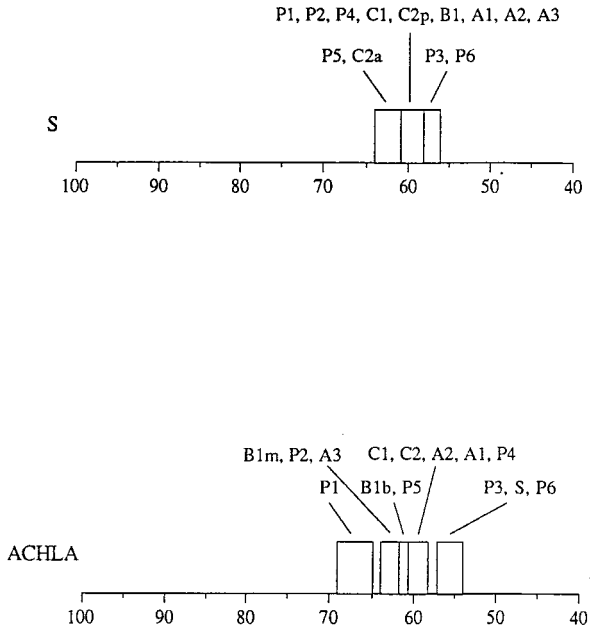


FIG. 5. Schematic presentation of between-group SSPA value ranges: comparisons between cyanobacteria and acholeplasmas. The ranges are given relative to the groups indicated on the left as standards (see Table 3). The cyanobacteria are placed closest to the gram(+) mycobacterial P5 group. The mycoplasma sequence ACHLA is closest to the gram(+) P1 group.

Downloaded from http://jb.asm.org/ on October 28, 2020 by guest

results are BACFR/C: 56 to 61, BACFR/B: 60 to 66, BACFR/A: 61 to 64, and BACFR/D: 60 to 62, where B1m provides the best matches to BACFR. Alignments with the gram(+) bacterial sequences yield BACFR/P1: 64, BACFR/P5: 63 to 64 ⩾ BACFR/P3: 56 and 58, and BACFR/(P2 plus P4 plus P6): 59 and 61, with the P1 and P5 groups giving the top alignments to BACFR. The mycoplasma ACHLA relative to BACFR also scores at 64.

For CHLTR, all SSPA determinations are ≤61. P1 and P5 share the top alignments at the level of 60 to 61.

DISCUSSION

RecA genes have been cloned and sequenced from a large number of eubacteria. Sixty-three RecA sequences are available from 62 bacterial sources (two separate RecA genes were sequenced from *M. xanthus* [29]). The sequences derive from 37 proteobacterial strains (α, β, γ, δ, and ε classifications), 11 gram(+) strains, three mycoplasma strains, three cyanobacterial strains, three representatives of the *Deinococcus-Thermus* group, and six unclassified cases (Table 1). The protein lengths are nearly constant, approximately 350 residues. The extent of similarity among the eubacterial RecA sequences (SSPA values) is moderate to high, ranging from 43 to 98% on the amino acid level.

The X-ray crystallographic structure of the *E. coli* RecA monomer-ADP complex has been determined (34). The ATP-binding domain (A site) extends approximately over positions 66 to 74, GPESSGKTT, in *E. coli* sequence coordinates, maintaining almost total identity across all 63 RecA sequences (exceptions are P-67, modified to N in *M. xanthus* RecA1, V in AQUPY, and Q in THEMA, and S-70, altered to G throughout the *Deinococcus-Thermus* group DT [DEIRA, THEAQ, and THETH]). The segment from 144 to 149, which overlaps the B-hydrolysis site associated with ATP binding, is also perfectly conserved with respect to all 63 RecA sequences. The L2 loop segment from 194 to 210 is proposed as one of two possible DNA binding domains largely conserved among the γ-, β-, and α-proteobacterial sequences. The segments from 6 to 30, 110 to 120, 127 to 139, 148 to 156, 213 to 222, and 247 to 257, considered to be involved in monomer-monomer interactions fomenting polymerization of RecA monomers, are also substantially conserved. Detailed analysis of identical positions and similar positions (those allowing conservative amino acid changes) contrasted with highly or restricted variable positions and relations with structure and genetic mutational studies is compiled elsewhere (16).

Sequence alignments among proteins seek to identify highly conserved segments of functional or structural importance which can also help to elucidate evolutionary relationships. There is the caveat that different phylogenies may result for the same set of organisms on the basis of analysis of different protein sequences, dependent on details of the alignment program (e.g., alignment biases, treatment of gaps, and sequence ambiguities [23, 28]). Moreover, evolutionary developments within and between proteins can reflect evolutionary rate differences, intrinsic nucleotide or amino acid incorporation biases, DNA repair systems, and general influences of ecological, environmental, and various stress conditions (e.g., UV irradiation, osmolarity gradients, temperature extremes, pH pressures, habitat variants, energy systems, and interacting fauna and flora). The genome in question may be a genetic mosaic, having acquired genes from different sources and undergone lateral transfer, transposition, and recombination events in the course of evolution.

The SSPA method produces an alignment and an assessment of the similarity for each pair of protein sequences. The SSPA determinations separate consistent groups. Criteria for group ascertainment are twofold: (i) within-group SSPA scores generally exceed SSPA scores with sequences not in the group, and (ii) alignment scores with other groups and singular sequences are reasonably congruent for all members of the group. Across RecA sequences there appear to be at least 15 groups in which group members score consistently within and between groups.

We highlight several key conclusions from the SSPA RecA sequence comparisons and classifications.

Proteobacteria. (i) The γ-proteobacteria comprise two major groups: inhabitants of vertebrate hosts (enterobacteria, members of the family *Vibrionaceae*, and *Haemophilus* strains) and those found primarily in soil or aquatic environments, including *Pseudomonas*, *Azotobacter*, and *Acinetobacter* strains (7).

A consonant division emerges from SSPA analysis of the prokaryotic glutamine synthetase I homologs (data not shown). This division is also supported by genomic analysis of extreme dinucleotide relative abundances. It was observed that γ-proteobacteria of mammalian hosts persistently carry significantly high GpC dinucleotide relative abundances whereas soil γ-proteobacteria have GpC representations at normal relative abundance levels (17, 19).

(ii) The β-proteobacterial sequences persistently show better alignments to the γ-proteobacterial sequences than to the α-proteobacterial sequences (Fig. 3).

(iii) The β-proteobacteria constitute a quite diverse class within the proteobacteria. In particular, NEIGO and THIFE are not very close (SSPA score, 69), and they generally do not score consistently with other groups.

(iv) The α-proteobacteria divide into two major groups: bacteria primarily interacting with plants and free-living, photosynthetic bacteria. The two α-proteobacterial groups A1 and A2 relate to each other about to the same extent as the two γ-proteobacterial groups C1 and C2 relate to each other.

(v) The unclassified genera *Methylomonas* and *Methylobacillus* align significantly and about equally with the *Pseudomonas* γ types (the C2 group) and the genus *Bordetella* (the B1b group), and with respect to other groups they score more like the sequences of B1b. Therefore, we identify them as β types.

(vi) On the basis of SSPA RecA sequence alignments the following cases in the classification from reference 39 could be altered: *Legionella pneumophila* from γ to β, *Acinetobacter calcoacetius* from β to γ, and *Xanthomonas oryzae* from γ to β.

(vii) The ε class consisting of *Campylobacter jejuni* and *Helicobacter pylori* align best with groups B1 and A2.

(viii) The two *M. xanthus* sequences are most similar to the β-proteobacterial sequences of group B1m and about equally similar to the A1 group sequences (see below for further discussion).

(ix) The γ-, β-, and α-proteobacterial sequences constitute a coherent group in that their mutual SSPA scores are never separated by SSPA scores of any gram(+) bacterial, cyanobacterial, or singular sequences.

Gram(+) bacteria. The bare subclassification of gram(+) organisms in terms of high or low G+C composition appears inadequate.

(i) The gram(+) bacterial sequences divide into three groups of low G+C content and three groups of high G+C content. However, the low-G+C-content gram(+) bacteria do not form a group distinct from the high-G+C-content gram(+) bacteria, nor can all gram(+) bacteria be grouped together in a consistent way. For example, relative to P2, P1 and P5 are closest, whereas P6 and P3 are less similar than

Downloaded from http://jb.asm.org/ on October 28, 2020 by guest

even the C group (Fig. 4). The low-G+C-content group P1 (*B. subtilis* and *S. aureus*) compared with the low-G+C-content group P3 (*L. lactis* and *S. pneumoniae*) shows better alignments relative to all proteobacterial, cyanobacterial, mycoplasma, and high-G+C-content gram(+) bacterial sequences (Fig. 3 and 4).

(ii) Compared with other groups, including proteobacteria, cyanobacteria, and mycoplasmas, P1, P2, P4, and P5 score significantly higher than do P3 and P6 (Fig. 3).

Mycoplasmas. Each of the mycoplasma sequences records its highest SSPA score with the *B. subtilis* sequence and generally scores highest with the P1 group. By contrast, with respect to G+C content, the ACHLA genome estimated value of about 38% deviates most from that of *B. subtilis* (43%) versus those of the P3 group of *L. lactis* (36%) and *S. pneumoniae* (40%). The SSPA scores of the RecA ACHLA sequence as a standard separate the six gram(+) groups (except P3 and P6, the farthest) (Fig. 5).

Cyanobacteria. The cyanobacterial sequences, considered distantly related (39) to gram(+) bacterial sequences, tend to score highest with the mycobacterial group P5, at reduced levels with all unclassified sequences, and at intermediate levels with proteobacterial and the other gram(+) bacterial sequences.

Thermophiles. The genus *Thermus* can be grouped with the genus *Deinococcus* on the basis of the RecA SSPA values. The best alignments of THEMA, considered a deeply divergent bacterium, occur with the gram(+) mycobacteria at an SSPA level of 62. This is another example of an extreme living bacterial strain which has the greatest similarity with a subgroup of the gram(+) bacteria (3, 10, 11).

How valid are these groupings for other protein sequence comparisons? SSPA analysis of the RecA proteins produces essentially the same groups and orderings as do sequence comparisons of glutamine synthetase, 70-kDa heat shock protein (*E. coli* DnaK homologs), and heat shock protein 60 (GroEL homologs), but other gram(+) bacterial subdivisions emerge for protein comparisons of glutamate dehydrogenase, dihydrofolate reductase, and thymidylate synthase (data to be published elsewhere).

Why two RecA genes in *M. xanthus?* *M. xanthus* is a soil bacterium which forms fruiting bodies under nutrient starvation (32). Two independent unlinked RecA genes (*recA1* and *recA2*) have been cloned and sequenced (29). The RecA2 protein is active in both vegetative and developmental growth, whereas transcripts of *recA1* have not been detected during either period. However, both gene products could complement *E. coli* RecA functions (29). A genetically engineered disrupted *recA1* gene was isolated by homologous recombination in *M. xanthus*, while corresponding attempts to replace a disrupted *recA2* gene were unsuccessful (29).

Surprisingly, the SSPA score between RecA1 and RecA2 is only 65. The best alignments for RecA1 and RecA2 are achieved with the B1 and A1 proteobacterial groups of primary soil habitat, with maximal SSPA scores of 66 for RecA1 and 68 for RecA2. We venture some thoughts on the role and evolution of the two *M. xanthus* RecA genes.

(i) *M. xanthus* has two life stages, vegetative growth and developmental growth (fruiting body). Similarly, *M. xanthus* encodes several pairs of putatively duplicated proteins expressed during different life stages (for example, Lon [36, 37] and serine-threonine kinase [27, 36, 37]). Similarly, DNA repair may be different at each stage, relying on distinct RecA proteins. However, as noted above, RecA2 functions at both life stages, while it is unknown whether RecA1 is functional in *M. xanthus*. Parenthetically, there are other bacteria with separate growing and sporulation life stages (e.g., *B. subtilis* and streptomycetes), and no duplicate RecA genes appear extant in these cases.

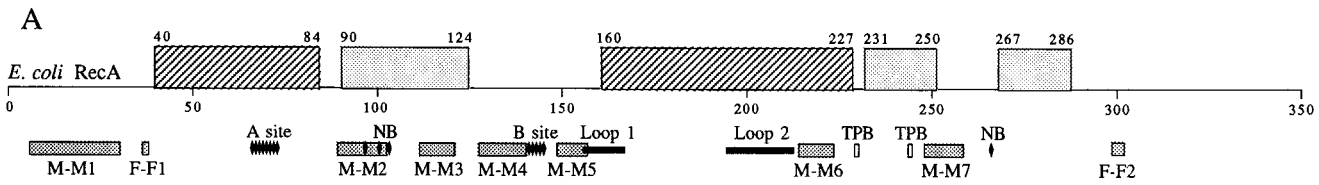
(ii) *M. xanthus* lives primarily in topsoil and is sensitive to UV light damage which putatively requires diligent DNA repair. In this context, *M. xanthus* harbors an abundance of pigments and carotinoid molecules that are quite photoprotective. Presumably, the presence of two genes avail greater RecA expression, mitigating problems attendant to UV irradiation. However, again, there are numerous topsoil bacterial species apparently expressing a single RecA gene.

(iii) The *M. xanthus* genome is among the largest known bacterial genomes (≈9.6 Mb). Shimkets (33) attributes the evolution of the large genome size of myxobacteria to events of phage and plasmid integration and tandem duplications. Could the two RecA genes have arisen by an endogenous duplication event with subsequent divergence? This seems unlikely since the RecA1/RecA2 mutual SSPA score of 65 is neither high enough to support a relatively recent duplication event nor low enough to suggest divergence to a nonfunctional gene. In this context, there are issues of time scale which are almost impossible to resolve. It is of note that also on the DNA level there is no evidence for a duplication event. Of the 331 residues aligned by the SSPA method, 220 are identical and 111 are substitutions. The codons corresponding to the identical residues display 13.5% sequence divergence at the DNA level, and the residue substitutions correspond to 66.4% sequence divergence, with overall sequence divergence at 31.2%. These numbers are very similar to those for other protein comparisons of similar SSPA values, e.g., RecA1/METFL (SSPA score, 66; DNA sequence divergence, 32.5%), RecA1/RHOSH (63; 33.7%), RecA2/METFL (68; 30.6%), RecA2/RHOSH (68; 29.1%), and METFL/RHOSH (70; 29.2%).

(iv) We propose a different hypothesis, to the effect that one of the two *M. xanthus* RecA genes was acquired via horizontal gene transfer by the following mechanism. *M. xanthus*, a proficient bacterial predator, acquired one of the RecA sequences (probably *recA1*) as a remnant from one of its bacterial prey. In this context, the myxobacteria secrete many digestive enzymes and through swarming behavior absorb prey bacterial lipids, peptides, and nucleic acid segments, etc. It is conceivable that from such an ingestation, intact DNA was integrated into the *M. xanthus* genome. Relics of other (especially soil) bacterial genomes may be expected to be distributed over the genome. An extended sequence containing *recA1*, the alleged transferred DNA sequence, allows more facile recombination in *E. coli*. Is it possible that the *recA1* gene transfer sequence does not carry the appropriate regulatory sequences to allow expression in cellular *M. xanthus*? However, in view of the relatively high SSPA score of RecA1 with several proteobacterial sequences, it seems likely that RecA1 is functional in *M. xanthus* in some circumstances.

Interestingly, both RecA *M. xanthus* sequences score highest (in SSPA values) with the B1 and A1 proteobacterial groups (SSPA scores, 64 to 68), which are principally soil and plant-related bacteria. Myxobacteria have a long ecological association with soil bacteria. From this perspective, genomic accretions stimulated by nucleic acid transfer events concomitant to *M. xanthus* predation activities may be an important factor underlying the expansion of the *M. xanthus* genome. In comparisons of genomic DNA, *M. xanthus* is closest to the C2-group soil bacteria *P. aeruginosa* and *A. vinelandii* (measured in terms of dinucleotide relative abundance distances [18]). The same intragenome distance measure when applied to different samples of 20-kb segments versus the intragenome distances for other proteobacterial sequences reveals relatively high in-

Downloaded from http://jb.asm.org/ on October 28, 2020 by guest



B

```
RECA_BACSU     38- 82   vpsGSlaLD (ta) LgiGGyprgrIiEvyGpesSGKTtvalhaiaevQ
RECA_ECOLI     40- 84   IsTGSlsLD (ia) LgaGGlpmgrIvEiyGpesSGKTtltlqviaaaQ
RADA_SULSO     89-132   IsTGSqaLD ( g) LlaGGietrtmtEffGefgSGKTqlchqlsvnvQ
RA51_HUMAN    102-145   ItTGSkeLD ( k) LlqGGietgsItEmfGefrtGKTqichtlavtcQ
RA51_LILLO    113-156   ItTGSqaLD ( e) LlgGGietlqItEafGefrSGKTqiahtlcvstQ
DMC1_YEAST    113-139   lsTGSkqlD ( s) ilgGGimtmsItEvfGefrcGKTqmshtlcvttQ

UVSX_BPT4      86-120   clFyDsEfgitPaYlrsmGVdpervihtpvqslEQ
RECA_BACSU     87-121   saFiDaEhaldPvYaqklGVnieelllsqpdtgEQ
RECA_ECOLI     90-124   caFiDaEhaldPiYarklGVdidnllcsqpdtgEQ

RECA_BACSU    157-224   GdshvglqarlmsqalrkLsgainksktiaifiNQi (  rekv  ) gvmfgnPettpGGraLkfyssvRLevRr
RECA_ECOLI    160-227   GdshmglaarmmsqamrkLagnlkqsntllifiNQi (  rmki  ) gvmfgnPetttGGnaLkfyasvRLdiRr
RADA_SULSO    222-289   GreilavrqqklnkhlhqLtrlaevydiaviitNQv (  marp  ) dmfygdPtvavGGhtLyhvpgiRiqLkK
RA51_HUMAN    234-304   GrgelsarqmhlarflrmLlrladefgvavvitNQv ( vaqvdga) amfaadPkkpiGGniiahasttRLyLRK
RA51_LILLO    245-314   GrgelaerqqklaqmlsrLtkiaeefnvavymtNQv (  iadpgg) gmfisdPkkpaGGhvLahaatvRLmLRK
DMC1_YEAST    228-299   GrgelserqqklnqhlfkLnrlaeefnvavfltNQv (qsdpgasa) lfasadgrkpiGGhvLahasatRilLRK

UVSX_BPT4     229-248   iKdGsdlqGyqfvlnVeKsr
RECA_BACSU    228-247   lKqGndvmGnktkikVvKnk
RECA_ECOLI    231-250   vKeGenvvGsetrvkVvKnk

UVSX_BPT4     265-284   GIdpysgllDmalelgfvvK
RECA_BACSU    264-283   GIskegeiiDlgteldivqK
RECA_ECOLI    267-286   GInfygelvDlgvkeklieK
```

FIG. 6. Sequence similarities between RecA, RAD proteins, and UvsX. The five regions of similarity were initially identified by overlapping HSSPs from the various pairwise comparisons and subsequently refined to focus on the regions of highest multiple similarity (for example, the last two motifs are the two dominant regions of similarity contained in the long HSSP shared between UvsX and the RecAs). (A) Regions of similarity between *E. coli* RecA and RAD proteins (striped) and between *E. coli* RecA and UvsX (dotted). Known and putative structural domains of RecA are indicated below the line. The A-site and B-site ATP-binding regions are associated with positions 66 to 74 and 140 to 144, respectively. Individual nucleotide binding (NB) positions occur at positions 96, 100, 103, and 265. M-M2 (positions 89 to 102) refers to the second monomer-monomer interface, and M-M3 refers the third monomer-monomer interface, etc. The hypothesized disordered DNA binding domains are assigned to loop 1 (155 to 165) and loop 2 (194 to 210). The two positions TBP-229 and TPB-243 are putative LexA (or UmuD) binding sites. Proposed filament-filament interaction sites (F-F) are positions 37, 38, and 298 to 301. (B) Blocks of sequence similarity corresponding to the boxes in panel A. Highly conserved residues are indicated in capital letters and by underlining.

tragenomic variability in *M. xanthus* (data not shown). We note that these observations would not be unexpected if indeed the *M. xanthus* genome were to carry remnants of incorporated DNA from other soil bacterial origin.

Comparison of RecA with other sequences. (i) RAD. Human RAD51 protein and yeast RAD51 protein (mutual SSPA score of 58) are considered functional homologs of RecA and are crucially involved in both mitotic and meiotic recombination and in repair of double-strand DNA breaks. Other members of this family include RAD51 of the trumpet lily and yeast proteins DMC1, RAD55, and RAD57.

SSPA comparisons of the human RAD51 sequence with the 63 RecA sequences yield in 52 cases a single HSSP (corresponding to *E. coli* residues 54 to 75), which includes the ATP-binding motif (20, 21). The low-G+C-content gram(+) bacterial RecA sequences, with the exception of CLOPE, share no HSSP with human RAD51. Yeast RAD51 does not score with any of the gram(+) bacterial RecAs, and the similarity in the ATP-binding domains scores significantly with only about half of the proteobacteria. For example, there are no HSSPs of yeast RAD51 versus any enterobacterial RecA sequences. Yeast RAD55 and RAD57 score weakly with most RecAs in the same region. DMC1, on the other hand, scores with very few of the RecAs, the strongest hits being with the cyanobacteria (as is the case for human and yeast RAD51). Parenthetically, the cyanobacteria are close to yeast also in genomic comparisons (18).

A. J. Clark and S. Sandler (4a) determined a RAD-like sequence from the archaebacterium *S. solfataricus*, designated RAD-A (324 amino acids [aa]). Its SSPA score with the human RAD51 sequence is 39, with the alignment extending essentially over the entire proteins. SSPA comparisons of RAD-A with C and B proteobacterial RecA sequences yield three HSSPs corresponding in *E. coli* to positions 35 to 51, 53 to 84, and 160 to 227, with total SSPA scores in the low range of 7 to 9. The second HSSP contains the ATP-binding motif, and the

Downloaded from http://jb.asm.org/ on October 28, 2020 by guest

third HSSP intersects one of the hypothesized DNA binding domains (34). Alignments of RAD-A with gram(+) bacterial RecA sequences are among the weakest and score exclusively with the last HSSP. Paradoxically, the best alignment is achieved with the mycoplasma ACHLA sequence involving four HSSPs (SSPA score of 10). Thus, the *Sulfolobus* sequence is much more similar to the eukaryotic RADs than it is to the eubacterial RecAs, but among the RADs, it is the most similar to the RecAs. This finding is consistent with both the Woese (39) and Lake (23, 30) views on bacterial evolution, which place *Sulfolobus* among the prokaryotes closest to eukaryotes (see also reference 18). Such associations hold true especially with respect to proteins of the replication and transcription machinery (11).

(ii) UvsX of phage T4. The RecA-like protein UvsX encoded by phage T4 is known to interact with the T4 replication machinery, for example, in creating primers for replication, in D-loop formation, and in establishing recombination intermediates (22). The effectiveness of UvsX requires a complex with protein UvsY and polypeptide gene 32 of T4. It is unclear whether the analogs of UvsY and gene 32 in *E. coli* exist (the Ssb protein of *E. coli* is a functional analog of gene 32; however, they have no sequence similarity).

SSPA comparison of UvsX sequence with the 63 RecA sequences reveals an almost invariant HSSP connecting positions 205 to 284 of UvsX with, for example, positions 207 to 286 of the *E. coli* RecA sequence. Another HSSP occurs with about half of the RecA sequences, aligning segment 71 to 120 of UvsX with, for example, segment 75 to 124 of *E. coli* RecA. The latter segment is immediately carboxyl to the standard ATP-binding motif. The highest SSPA scores (range, 6 to 7) in these comparisons occur with the α-proteobacteria, with the gram(+) bacterial sequences of groups P1 and P4 to P6, and with the mycoplasma ACHLA. The lowest scores (about 3) occur with the β-proteobacterial RecA sequences (except NEIGO).

Comparison of the UvsX sequence with the eukaryotic RAD sequences and RAD-A of *S. solfatarius* yields no HSSPs. Consistently, the regions of similarity (HSSPs) between the RecAs and the RADs on the one hand and between the RecAs and UvsX on the other hand are disjoint (Fig. 6).

The partition of the RecA sequence into different motifs, each of which is shared with other distinct protein families, suggests a mosaic composition and evolution of RecA and its relatives. In this context it is interesting that motif 2 of Fig. 6 is most similar between UvsX and the A1 proteobacteria, whereas motifs 4 and 5 score highest against the P4 group of gram(+) bacteria. Assuming genetic transfer between phage and bacterial host as part of the corresponding protein evolution, one might inquire as to the possibility of the T4 host range including α-proteobacteria and even gram(+) bacteria.

ACKNOWLEDGMENTS

We thank B. E. Blaisdell, A. M. Campbell, A. J. Clark, D. Kaiser, A. I. Roca, and S. J. Sandler for comments and valuable discussions on various aspects of the manuscript.

This work was supported in part by NIH grants 2R01GM10452-31 and 5R01HG00335-07 and NSF grant DMS 9403553 to S.K. and V.B.

REFERENCES

1. **Bairoch, A., and B. Boeckmann.** 1994. The SWISS-PROT protein sequence data bank: current status. Nucleic Acids Res. **22:**3578–3580.
2. **Benachenhou-Lahfa, N., P. Forterre, and B. Labedan.** 1993. Evolution of glutamate dehydrogenase genes: evidence for two paralogous protein families and unusual branching patterns of the archaebacteria in the universal tree of life. J. Mol. Evol. **36:**335–346.
3. **Brown, J. R., Y. Masuchi, F. T. Robb, and W. F. Doolittle.** 1994. Evolutionary relationships of bacterial and archaeal glutamine synthetase genes. J. Mol. Evol. **38:**566–576.
4. **Burks, C., et al.** 1990. GenBank: current status and future directions. Methods Enzymol. **183:**1–22.

4a. **Clark, A. J., and S. Sandler.** Personal communication.

5. **Clark, A. J., and S. J. Sandler.** 1994. Homologous genetic recombination: the pieces begin to fall into place. Crit. Rev. Microbiol. **20:**125–142.

5a. **Coleman, J.** Personal communication.

6. **Davis, E. O., H. S. Thangary, P. C. Brooks, and M. J. Colston.** 1994. Evidence of selection for protein introns in the recAs of pathogenic mycobacteria. EMBO J. **13:**699–703.
7. **de Ley, J.** 1992. Introduction to the proteobacteria, p. 2110–2140. *In* H. G. Trüper, M. Dworkin, W. Harder, and K. H. Schleifer (ed.), The prokaryotes. Springer-Verlag, Berlin.
8. **Golding, B. G., and R. S. Gupta.** 1995. Protein-based phylogenies support a chimeric origin for the eukaryotic genome. Mol. Biol. Evol. **12:**1–6.
9. **Gupta, R. S.** 1995. Evolution of the chaperonin families (HSP60, HSP10 and TCP-1) of proteins and the origin of eukaryotic cells. Mol. Microbiol. **15:**1–11.
10. **Gupta, R. S., and B. G. Golding.** 1993. Evolution of HSP70 gene and its implication regarding relationships between archaebacteria, eubacteria and eukaryotes. J. Mol. Evol. **37:**573–582.
11. **Gupta, R. S., and B. Singh.** 1994. Phylogenetic analysis of 70 kD heat shock protein sequences suggests a chimeric origin for the eukaryotic cell nucleus. Curr. Biol. **4:**1104–1114.
12. **Henikoff, S., and J. G. Henikoff.** 1992. Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. USA **89:**10915–10919.
13. **Hori, H., and S. Osawa.** 1987. Origin and evolution of organisms as deduced from 5S ribosomal RNA sequences. Mol. Biol. Evol. **4:**445–472.

13a. **Huang, W. M.** Personal communication.

13b. **Inouye, M.** Personal communication.

14. **Karlin, S., and S. Altschul.** 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc. Natl. Acad. Sci. USA **87:**2264–2268.
15. **Karlin, S., V. Brendel, and P. Bucher.** 1992. Significant similarity and dissimilarity in homologous proteins. Mol. Biol. Evol. **9:**152–167.
16. **Karlin, S., and L. Brocchieri.** Unpublished data.
17. **Karlin, S., and C. Burge.** 1995. Dinucleotide relative abundance extremes: a genomic signature. Trends Genet. **11:**283–290.
18. **Karlin, S., and L. R. Cardon.** 1994. Computational DNA sequence analysis. Annu. Rev. Microbiol. **48:**619–654.
19. **Karlin, S., I. Ladunga, and B. E. Blaisdell.** 1994. Heterogeneity of genomes: measures and values. Proc. Natl. Acad. Sci. USA **91:**12837–12841.
20. **Kowalczykowski, S. C., D. A. Dixon, A. K. Eggleston, S. D. Lauder, and W. M. Rehrauer.** 1994. Biochemistry of homologous recombination in *Escherichia coli*. Microbiol. Rev. **58:**401–465.
21. **Kowalczykowski, S. C., and A. K. Eggleston.** 1994. Homologous pairing and DNA strand-exchange proteins. Annu. Rev. Biochem. **63:**991–1043.
22. **Kreutzer, K. N., and S. W. Morrical.** 1994. Initiation of DNA replication, p. 28–42. *In* J. D. Karan (ed.), Molecular biology of bacteriophage T4. American Society for Microbiology, Washington, D.C.
23. **Lake, J. A.** 1994. Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. Proc. Natl. Acad. Sci. USA **91:**1455–1459.
24. **Lloyd, A. T., and P. M. Sharp.** 1993. Evolution of the *recA* gene and the molecular phylogeny of bacteria. J. Mol. Evol. **37:**399–407.
25. **Maniloff, J., R. N. McElhaney, L. R. Finch, and J. B. Basemann (ed.).** 1992. Mycoplasmas: molecular biology and pathogenesis. American Society for Microbiology, Washington, D.C.
26. **McCleary, W. R., B. Esmon, and D. R. Zusman.** 1991. *Myxococcus xanthus* protein C is a major spore surface protein. J. Bacteriol. **173:**2141–2145.

26a. **Mongkolsuk, S.** Personal communication.

27. **Munoz-Dorado, J., S. Inouye, and M. Inouye.** 1991. A gene encoding a protein serine/threonine kinase is required for normal development of *M. xanthus*, a gram-negative bacterium. Cell **67:**995–1006.
28. **Nei, M.** 1987. Molecular evolutionary genetics. Columbia University Press, New York.
29. **Norioka, N., M.-Y. Hsu, S. Inouye, and M. Inouye.** 1995. Two *recA* genes in *Myxococcus xanthus*. J. Bacteriol. **177:**4179–4182.
30. **Rivera, M. C., and J. A. Lake.** 1992. Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. Science **257:**74–76.
31. **Roca, A. I., and M. M. Cox.** 1990. The RecA protein: structure and function. Crit. Rev. Biochem. Mol. Biol. **25:**415–456.

31a. **Rood, J.** Personal communication.

32. **Shimkets, L. J.** 1990. Social and developmental biology of the myxobacteria. Microbiol. Rev. **54:**473–501.
33. **Shimkets, L. J.** 1993. The myxobacterial genome, p. 85–107. *In* M. Dworkin and D. Kaiser (ed.), Myxobacteria II. American Society for Microbiology, Washington, D.C.
34. **Story, R. M., and T. A. Steitz.** 1992. Structure of the RecA protein-ADP complex. Nature (London) **355:**318–325.

Downloaded from http://jb.asm.org/ on October 28, 2020 by guest

35. **Tiboni, O., P. Cammarano, and A. M. Sanangelantoni.** 1993. Cloning and sequencing of the gene encoding glutamine synthetase I from the archaeum *Pyrococcus woesei*: anomalous phylogenies inferred from analysis of archaeal and bacterial glutamine synthetase I sequences. J. Bacteriol. **175:**2961–2969.
36. **Tojo, N., S. Inouye, and T. Komano.** 1993. Cloning and nucleotide sequence of the *Myxococcus xanthus lon* gene: indispensability of *lon* for vegetative growth. J. Bacteriol. **175:**2271–2277.
37. **Tojo, N., S. Inouye, and T. Komano.** 1993. The *lonD* gene is homologous to the *lon* gene encoding an ATP-dependent protease and is essential for the development of *Myxococcus xanthus*. J. Bacteriol. **175:**4545–4549.
38. **Viale, A. M., and A. K. Arakaki.** 1994. The chaperone connection to the origins of the eukaryotic organelles. FEBS Lett. **341:**146–151.
39. **Woese, C. R.** 1987. Bacterial evolution. Microbiol. Rev. **51:**221–271.
40. **Woese, C. R., O. Kandler, and M. L. Wheelis.** 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc. Natl. Acad. Sci. USA **87:**4576–4579.

Downloaded from http://jb.asm.org/ on October 28, 2020 by guest