Molecular Microbiology (2002) **43**(6), 1565–1575

Identification of genes that are associated with DNA repeats in prokaryotes

Ruud. Jansen,[1]* Jan. D. A. van Embden,[2] Wim. Gaastra[1] and Leo. M. Schouls[2]
[1]*Department of Infectious Diseases and Immunology, Bacteriology Division, Veterinary Faculty, Utrecht University, Yalelaan 1, 3584 CL Utrecht, The Netherlands.*
[2]*Laboratory of Bacteriology of the Research Laboratory of Infectious Diseases, National Institute of Public Health and Environmental Protection, A. van Leeuwenhoeklaan 1, 3720 BA Bilthoven, The Netherlands.*

Summary

Using *in silico* analysis we studied a novel family of repetitive DNA sequences that is present among both domains of the prokaryotes (Archaea and Bacteria), but absent from eukaryotes or viruses. This family is characterized by direct repeats, varying in size from 21 to 37 bp, interspaced by similarly sized non-repetitive sequences. To appreciate their characteristic structure, we will refer to this family as the clustered regularly interspaced short palindromic repeats (CRISPR). In most species with two or more CRISPR loci, these loci were flanked on one side by a common leader sequence of 300–500 b. The direct repeats and the leader sequences were conserved within a species, but dissimilar between species. The presence of multiple chromosomal CRISPR loci suggests that CRISPRs are mobile elements. Four CRISPR-associated (*cas*) genes were identified in CRISPR-containing prokaryotes that were absent from CRISPR-negative prokaryotes. The *cas* genes were invariably located adjacent to a CRISPR locus, indicating that the *cas* genes and CRISPR loci have a functional relationship. The *cas3* gene showed motifs characteristic for helicases of the superfamily 2, and the *cas*4 gene showed motifs of the RecB family of exonucleases, suggesting that these genes are involved in DNA metabolism or gene expression. The spatial coherence of CRISPR and *cas* genes may stimulate new research on the genesis and biological role of these repeats and genes.

Accepted 10 December, 2001. *For correspondence. E-mail R.jansen@vet.uu.nl; Tel. (+31) 30 253 4791; Fax (+31) 30 254 0784.

Introduction

Repetitive sequences are common in the genomes of prokaryotic organisms, and their identification is increasingly facilitated by the availability of sequences of complete bacterial genomes. The length, sequence and position of these sequences in the genome are highly variable and often unique for a single strain (Lupski *et al*., 1996; Bachellier *et al*., 1997; van Belkum *et al*., 1998). Two main classes of short sequence repeats (SSRs) can be distinguished, contiguous repeats and interspersed repeats (van Belkum *et al*., 1998). The number of units of contiguous repeats usually varies from strain to strain, and this genetic heterogeneity in bacterial populations may lead to phenotypic differences due to differential gene transcription or translation. These SSRs are often part of open reading frames or promoter regions, and variation in the number of repeat units may lead to the switch of expression of surface-exposed components (Dybvig, 1993; Belland *et al*., 1997; van Belkum *et al*., 1998). Thus, the SSR-mediated variation is thought to provide a bacterial population with the diversity needed to adapt to changing environments.

The group of interspersed SSRs is also commonly present in bacteria. Their unit length is usually less than 200 bp; they are non-coding, intercistronic and widely dispersed throughout the genome. Many interspersed SSRs have been disclosed in bacterial species. Examples are the REP and ERIC seqsuences of the Enterobacteriaceae, the BOX element in *Streptococcus pneumoniae* and the SSRs in *Neisseria gonorrhoeae* and *Haemophilus influenzae* (Correia *et al*., 1988; Hulton *et al*., 1991; Martin *et al*., 1992; Fleischmann *et al*., 1995). The SSR sequences of different bacterial phyla differ considerably in sequence, and consensus sequences differ from genus to genus. These repeat structures have been implicated in mRNA stabilization, transcription termination and genetic rearrangements. In *H. influenzae*, the SSRs are involved in the DNA exchange by transformation (Fleischmann *et al*., 1995).

A distinct class of interspaced SSRs was recognized in 1987 in *E. coli* K12 (Ishino *et al*., 1987; Nakata *et al*., 1989), and later in other bacterial and archaeal species such as *Haloferax mediterranei, Streptococcus pyogenes, Anabaena* sp. PCC 7120 and *Mycobacterium tuberculosis* (Groenen *et al*., 1993; Mojica *et al*., 1995; Masepohl *et al*., 1996; Hoe *et al*., 1999).

Species carrying CRISPR loci

Organism		CRISPR sequence	No. of CRISPR loci	No. of repeats	*cas1*	*cas2*	*cas3*	*cas4*	EMBL/GenBank accession number
Aeropyrum pernix	A	GAATCT-CCGAGATAGAATTGCAAG	2	1-19	ape1240	n	ape1232	ape1239	AP000059, AP000060, AP000062
		.....-T.....	1	27					
		.C..A.C..T.A.GG....A.A...	1	42					
Anabaena		GTTTCAATCCCTGATAGGGATTTTTGTTAGTTAAAAC	1	16					X87270
Aquifex aeolicus		GTTCCTAATGTACCGT-GTGGAGTTGAAAC	4	2-4	aq0369	y	aq0371	aq0370	AE000667, AE000686, AE000680, AE000688, AE000754, AE000765, AE000768
		...T.....-.....A....	1	5					
		...A.GCCCTA.TA..GAG...-...CG..	2	2-3					
		...T.A.C.CC..TACG..AC.T.A.G...	1	5					
		...ATGCCCT...AA.AAG...-...CG..	1	4					
		...T.A.C.CC..-ACG..AC.T.A.G...	1	5					
		.....A.A.A.....	1	2					
		...T.....-.....	1	2					
		...T.....A.A.....	1	4					
		...T.....-.....G	1	3					
		No repeat sequence, only leader	1	0					
Archaeoglobus fulgidus	A	GTTGAAATCAGACCAAAATGG-GATTGAAAG	2	48-60	af1878	af1876	af1874	af1877	AE000973, AE001074, AE001107
		TAAGAAAG.G.GG.TCCTG.....A.....	1	44	af2435	af2434	n	af2436	
Bacillus halodurans		GTCGCACTCTTCATGGGTGCGTGGATTGAAAT	3	14-36	bh0341	bh0342	bh0336	bh0340	BA000004
		.....A....A.....	1	14					
		.....AT.....	1	11					
Bacillus strearothermophilus		GTTTCAATCCCTCATAGGTACGATAAAAAC	4	8-21	y		y	y	UOKNOR 1422 at EN
Campylobacter jejuni		TTTTAGTCCCTTTTTAAATTTCTTTATGGTAAAAT	1	5	cj1522c	n	n	n	AL111168
Carboxydothermus hydrogenoformans		CAATCCCAGAATGGTTCGATTAAAAC	1	60	y		y	y	chydro 2346
		.....T.....	4	26-60					
Chlorobium tepidum		GTTTCAATCCACGCGCCCGCGCGGGGCGCGAC	1	22	y				C.tepidum gct18 at EN
Clostridium difficile		ATTTACACCACTTAGTTAATATAAAAC	4	5-18	y		y	y	Contig930,914,890,845.1 at EN
Corynebacterium diphteriae		GAAGTCTATCAGGGTTTTTGAGAAGTCAACCCCAGT	1	8	y				gnl|Sanger_1717|cdiph at EN
Escherichia coli K-12		CGGTTTATCCCCGCTGGCGCGGGGAACTC	1	14	b2755	n	b2761	n	AE000359, M27059, M27060, U29579 U29580
		.....A.	1	7					
Escherichia coli O157 H7 sakai		CGGTTTATCCCCGCTGGCGCGGGGAACTC	1	4	Z4064		ZygcB		AE005174
		.....A.	1	5					
Fibrobacter succinogenes		CCCTGAAAAGCATTCCTCTCGCAAGAGAGAT	5	3-13	y				fsuccin 1224 at EN
Geobacter sulfurreducens		GTATTCCGGGGCCATGATGCCCCGGCCTCATTGAAGC	1	38	y		y		gsulf_2947 at EN
Haloferax mediterranei	A	GTTACAGACGAACCCTAGTTGGGTTGAAGC	3	22	?				X73453
Methanobacterium thermoautotrophicum	A	GTTAAAATCAGACCAAAATGGGATTGAAAT	2	47-124	mth1084	mth1083	mth1087	mth1085	AE000878, AE000920
Methanococcus jannaschii	A	AATTAAAATCAGACCGTTTCG-GAATGGAAA	2	16-27	mj0378	mj0386	mj0376	mj0377	U67459, U67463, U67467, U67468 U67470, U67480, U67491, U67498 U67505, U67506, U67516, U67536 U67540, U67552, U67553, U67568 U67572, U67578, U67581, U67589 U67599, U67600
		G.....-.....T	1	3					
		.....A.....TC..G.A.G.....	1	4					
		.....A....-.....	1	2					
		.....-.....T	1	24					
		G.....T.CC.CG.G.....T	1	12					
		G.....A....-.....C	1	16					
		G.....TC..G.A.G.....	7	7-14					
		G.....-.....	1	13					
		.....TC..G.A.G.....	1	2					
		....A.....-.....C	1	12					
		.....A.....-.....	1	13					
		No repeat sequence, only leader	1	0					
Methylococcus capsulatus		GTTTCAATCCACTCCCGGCTATTTAGCCGGGAGATAC	1	64	y				mcapsul_bmc_69 at EN
Mycobacterium avium		TCATCCCCGCTTGCGCGGGGAGCA	1	13					M.avium_339 and 23 at EN
Mycobacterium tuberculosis and Mycobacterium bovis		GTTTCCGTCCCCTCTCGGGGTTTTGGGTCTGACGAC	1	42		Rv2817	Rv2816	n	n	AD000007, X57835, Y14045, Y14046, Y14047, Y14048, Y14049, Z48304, Z81331, U47864
Neisseria meningitidis serogroup A		ATTGTAGCACTGCGAAATGAGAAAGGGAGCTACAAC	1	16	nma0630	n	n	n	AL162759
Pasteurella multocida		ATTGTAGCACTGCGAAATGAGAGAGGGAACTACAAC	1	7	pm1126	n	n	n	AE004439
		GTTCACCATCGTGTAGATGGCTAAGAAAG	1	33					
Porphyromonas gingivalis		GTTTTAATTCCTGTATGGTGCAATTGAAAT	1	24	y	y	y	y	gnl|TIGR|P.gingivalis_GPG.con
Pyrococcus furiosus	A	GTTCCAATAAGACTAAAATAGAATTGAAAG	1	52	y	y	y	y	NC 002968 at EN
		...A.....C.....	1	31					
		...A.....	1	46					
Pyrococcus horikoshii	A	GTTTCCGTAGAACTA-AATAGTGTGGAAAG	2	19-67	ph0173	y	ph0176	ph0175	AP000001, AP000004
		.....T-.G.....	1	26					
		...C.AA..AG....T..G..AA.T.....	2	7-18	ph01245	y	ph1246	S033	
Pyrococcus abyssii	A	GTTCCAATAAGACTATAAGAGAATTGAAAG	3	50	y	y	y	n	AJ248283, AJ248284 AJ248286, AJ248288
		GTTTCCGTAGAACTAAATAGTGTGGAAAG	2	9					
Salmonella enteritidis		CGGTTTATCCCCGCTGGCGCGGGGAAC	2	10	y		y		senteritdis_477_10.21, 1184_10.21at EN
Salmonella paratyphi		CGGTTTATCCCCGCTGGCGCGGG	2	4-6	y		y		spara_B_SPA.0.26636 at EN
Salmonella typhi		CGGTTTATCCCCGCTGGCGCGGGGAACAC	1	9	y		y		S.typhi_Salmonella at EN
Salmonella typhimurium		CGGTTTATCCCCGCTGGCGCG	1	40	y		y		stmlt2_contig2 LT2 at EN
Streptococcus pyogenes M1		GTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAAC	1	7	SPy1562	SPy1561	SPy1567	SPy1563	AE004092
		ATTTCAATCCACTCACCCATGAAGGGTGAGACT	1	4	SPy1047				
Streptococcus mutans		GTTTTAGAGCTGTGTTGTTTCGAATGGTTCCAAAAC	1	7	y		y		S.mutans_Contig4 at EN
Sulfolobus solfataricus	A	GATAATCTACTATAGAATTGAAAG	1	7	SSo1405	SSo1404	SSo1402	SSo1392	AE006641
		.C.....	1	89					
		.....CT.....	1	128					
Thermoplasma volcanium	A	CTTCCATACTAACTAGTACATCTTAAAC	2	16-19	TVn0106	TVn0105	n	n	BA000011
Thermotoga maritima		GTTTCAATACTTCCTTAGAGGTATGGAAAC	3	4-41	Tm1797	Tm1796	Tm1799	Tm1798	AE001689, AE001715, AE001716, AE001718, AE001719, AE001793, AE001800, AE001806, AE001817, and AE001818
		.....A.....	5	3-25					
Thermus thermophilus		AATCCCCTTACGGGGCTCAATCCCTTGCAAC	1	11					M33159
Yersinia pestis		TTTCTAAGCTGCCTGTGCGGCAGTGAAC	3	4-9	y				gnl|Sanger_632 at EN

Species without CRISPR loci

Actinobacillus actinomycetemcomitans	*Chlamydia trachomatis* MoPn	*Mycobacterium leprae*
Bacillus subtilis	*Deinococcus radiodurans*	*Mycoplasma genitalium*
Bordetella bronchiseptica	*Enterococcus faecalis*	*Mycoplasma pneumoniae*
Bordetella pertussis	*Haemophilus influenzae*	*Neisseria gonorrhoeae*
Borrelia burgdorferi sensu stricto	*Helicobacter pylori*	*Pseudomonas aeruginosa*
Buchnera sp.	*Klebsiella pneumoniae*	*Rickettsia conorii*
Caulobacter crescentus	*Legionella pneumophila*	*Rickettsia prowazekii*
Chlamydia pneumoniae AR39	*Mesorhizobium loti*	*Staphylococcus aureus*
Streptococcus pneumoniae type 4		
Synechocystis sp.		
Thermoplasma acidophylum A		
Thiobacillus ferrooxidans		
Treponema pallidum		
Ureaplasma urealyticum		
Vibrio cholerae O1 El Tor		
Xylella fastidiosa		

Only recently, this class of repeats was recognized as one family with members in many prokaryotic species (Mojica *et al*., 2000). Each member of this family of repeats was designated differently by the original authors, leading to a confusing nomenclature. To acknowledge the joining of this class of repeats as one family and to avoid confusing nomenclature, Mojica *et al*. and our research group have agreed to use in this report and future publication the acronym CRISPR, which reflects the characteristic features of this family of clustered regularly interspaced short palindromic repeats.

A characeristic of the CRISPRs, not seen in any other class of repetitive DNA, is that the repeats of the CRISPRs are interspaced by similarly sized non-repetitive DNA. The direct repeats varies in size from 21 bp in *Salmonella typhimurium* to 37 bp in *Streptococcus pyogenes,* and they are clustered in one or several loci on the chromosome.

The CRISPR loci in clinical isolates of *M. tuberculosis* and *S. pyogenes* are extremely polymorphic, and this strain-dependent polymorphism has been exploited for epidemiological and taxonomic purposes (Kamerbeek *et al*., 1997; Hoe *et al*., 1999). The short repeat sequence itself is remarkably well conserved in clinical isolates; however, the number of repeats and spacers differs from strain to strain (van Embden *et al*., 2000).

Apart from circumstantial evidence of the involvement of CRISPRs in replicon partitioning in *H. mediterranei*, it is unclear whether the CRISPR loci have a biological function (Mojica *et al*., 1995). In this study, the prevalence of CRISPR loci among organisms was determined, and by comparing the genetic environment of the CRISPR loci we identified four genes that are associated with the CRISPR loci. The possible interaction of these genes and the CRISPR loci will be discussed.

Results

Disclosure of CRISPR motifs in sequence databases

Initially, we tried to identify CRISPR sequences using the sequence similarity program NBLAST and the repeat sequence of *M. tuberculosis* (Groenen *et al*., 1993) as a query to search the EMBL/GenBank databases. However, no sequences similarities were identified in species other than *M. tuberculosis* and *M. bovis*, suggesting that this direct repeat sequence is present only in these closely related species. This was consistent with the observation that CRISPR sequences in different species, such as *E. coli* and *M. tuberculosis*, are very dissimilar (Mojica *et al*., 1995; 2000; Kamerbeek *et al*., 1997).

We considered the possibility that CRISPR loci might exist in other bacterial species with the characteristic motif of alternating short repeats and similarly sized unique sequences, but without sequences related to known CRISPRs. The PATSCAN program allows disclosure of sequence-independent motifs in DNA, and we used this program to search the EMBL/GenBank database for CRISPR motifs. The algorithm used recognizes motifs of at least four direct repeats, 15–70 bp in size, interspaced by sequences with the same size variation.

Several hundred hits were obtained, but most of these matching sequences consisted of large, mostly imperfect, direct repeats, without interspersing non-repetitive DNA. Therefore, the characteristic CRISPR motifs were selected manually from the sequences obtained using the PATSCAN algorithm. We searched for CRISPR motifs in the completely sequenced and partially sequenced genomes of prokaryotic species and the full database of EMBL/GenBank. This led to the disclosure of CRISPR motifs in more than 40 prokaryotic species, which are listed in Table 1. The CRISPR motifs were not found in approximately half of the bacterial species whose genome has been sequenced or almost completely sequenced (see Table 1). CRISPRs were not found in virus or eukaryotic sequences. The sizes of the prokaryotic repeat sequences were all within the narrow size range of 21–37 bp. The spacer sequences within a given CRISPR locus were similar to the size of the repeat sequences and varied slightly within a given CRISPR locus, similar to what has been observed in *E. coli* and *M. tuberculosis*. The number of CRISPR loci varied from 1 in some species, e.g. *M. tuberculosis* and *Neisseria meningitidis*, to 20 in *M. jannaschii*. The number of repeats within the CRISPR loci varied greatly, from two repeats to as many as 124 repeats in *M. thermoautotrophicum* (see Table 1).

Comparison of the repeat and spacer sequences in the various prokaryotes

The repeat sequences in the CRISPR loci of closely

Table 1. Listing of CRISPR and *cas* genes by species. Column 1 indicates the species names in alphabetical order. Archaeal species are indicated by 'A' after the name. The repeat sequences of the CRISPR loci are indicated In column 2. If applicable for the species, the variants of repeat sequences are aligned. The different nucleotides are indicated and identical bases are shown by a dot. In column 4 the lowest and the highest number of repeats in a CRISPR locus are indicated. Columns 5–8 give the identification numbers of *cas* genes from the genome projects. If a *cas* gene was not present in a completely sequenced and published genome, this is indicted by 'n' if the *cas* gene was found in an unfinished genome or if the *cas* gene was not annotated in a published genome, its presence is indicated by 'y'. The far right column contains the accession numbers of the EMBL/GenBank entry that contains the sequences of the CRISPR and *cas* genes. If the genome has not yet been published, the entrez nucleotide (EN) contig number is given. The lower panel shows the names of the published completed genomes that do not contain CRISPR sequences.

related species were similar or identical, e.g. between the streptococcal and the pyrococcal species and the closely related species *E. coli* and *Salmonella* (Table 1). CRISPRs of distantly related species were in general dissimilar. However, one exception was found in *N. meningitidis* and *P. multocida*, both of which carry a CRISPR with an identical repeat sequence. In addition, nearly identical repeat sequences differing only at a few nucleotide positions were found in the CRISPRs of *M. thermoautotrophicum* and *A. fulgidus*, and in *E. coli* and *M. avium* (Table 1).

Despite the absence of sequence similarity between the repeat sequences of the species, the repeat sequences share some common features. In most of the repeat sequences a lose dyad symmetry can be recognized (Mojica *et al.*, 2000). The nucleotides involved in the dyad symmetry are mainly located at the termini of the repeats and often include the complementary sequences GTT and AAC. The majority of the 76 different repeats identified have the sequence GTT at the terminus (see Table 1), and one-third of the repeats have AAC or AAG at the other terminus. Furthermore, the repeat sequences often contain stretches of three or four identical bases, mainly A or T residues, but stretches of C and G residues also occur. The stretches of identical bases and the dyad symmetry may give the DNA a particular secondairy structure. Using the program GeneQuest the bending of the DNA strand was predicted. The majority of CRISPR loci did not show a regular bending pattern at the repeats, but some species, in particular the thermophilic archaeal species, showed a strong bending of 72° at the repeat sequences that might result in a regular secondary structure of the CRISPR loci.

The repeat sequences within a given CRISPR locus were generally identical with rare single-basepair substitutions. Although the majority of the repeats in different CRISPR loci of the same organism were highly similar or identical, some were very dissimilar, such as those in *Aquifex aeolicus, Archaeoglobus fulgidus*, *Pasteurella multocida*, *Pyrococcus horikoshii* and *Streptococcus pyogenes.* Each of these species carried several CRISPR loci and the repeat sequences of the various CRISPR loci within the species may differ completely from each other. Such species were considered to carry two different CRISPR classes in their genome.

In most organisms, the last repeat or the few last repeats of a CRISPR locus contained mutations. In about one-third of the CRISPR loci the last repeat was truncated. Remarkably, in CRISPR loci with a leader sequence (see next paragraph) this was found only in the repeats distal to the leader sequence. Apparently, deviation from the consensus CRISPR sequence starts at the terminus and is unidirectional. This phenomenon was most apparent in *Aquifex aeolicus*, in which terminal degradation of the CRISPRs was found in 9 of the 14 CRISPR loci.

The DNA stretches between the repeats (the spacers) were present as a single copy in a CRISPR locus with very few exceptions. Exceptions were found in a CRISPR locus of *M. thermoautotrophicum*, in which small clusters of repeats and spacers had duplicates in the CRISPR locus and in *M. tuberculosis* that carried two identical spacers in the CRISPR region (Smith *et al.*, 1997; van Embden *et al.*, 2000). None of the spacers of a given organism shared sequence similarity with spacers of other organisms. This also holds true when the spacer sequences of closely related species were compared, e.g. *E. coli* and *S. enterica* or the *Streptococcus* and the *Pyrococcus* species (Table 1).

Identification of common sequences flanking CRISPR loci

Previous studies showed the presence of common sequences flanking the multiple CRISPR loci in *M. jannaschii*, *A. fulgidus* and *M. thermoautotrophicum* (Bult *et al.*, 1996; Klenk *et al.*, 1997; Smith *et al.*, 1997) and these common sequences have been designated as long repeats (LRs). In addition to these four species, we identified such common flanking leader sequences in *A. aeolicus*, *Bacillus halodurans, Thermoplasma volcanium*, the *Pyrococcus* species*, Thermotoga maritima* and *Yersinia pestis* (see Table 1). These leader sequences were several hundred basepairs long. They were located on one side of the CRISPR loci and their orientation with regard to the repeat sequence was invariably the same. The nucleotide sequences of the leaders within a given species shared approximately 80% sequence identity, however no homology was found among the leaders of taxonomically unrelated species. These leader sequences did not have an open reading frame, suggesting that they do not code for proteins. The only common feature of the leader sequences, except for their location, appeared to be the frequent presence of stretches of identical nucleotides and a high AT content. In general, leader sequences were only found near a CRISPR locus and not elsewhere in the genome. Only in *M. jannaschii* and *A. aeolicus* did we find a leader sequence without CRISPR. In *M. jannaschii* two truncated leaders were present. In *E. coli* and *Sulfolobus sulfataricus* no leader sequences were found despite the multiple CRISPR loci in these species.

Identification of CRISPR associated genes

Comparison of the genes that flank the CRISPR loci in the genomes of different prokaryotic species showed a clear homology among four genes. We designated these

genes the CRISPR-associated genes, *cas1* to *cas4*. The four *cas* genes were not present in all species with CRISPR loci. The presence of each of the four *cas* genes in the prokaryotic species is indicated in Table 1. For species for which the complete genome has not yet been published, the presence or absence of the *cas* genes is indicated. For those species whose genomes have been published the gene numbers of the *cas* genes are indicated. The positions of the *cas* genes relative to the CRISPR locus in the species that carry all four *cas* genes is depicted in Fig. 1.

The *cas2* genes of *P. horikoshii* and *A. aeolicus* were not annotated in the genome projects and were identified by searching these genomes using the TBLASTN program. All CRISPR-harbouring species for which the complete genome sequence was available shared the *cas1* homologue and one or more of the other three *cas* genes. In contrast, *cas* homologues were absent from any of the completely sequenced CRISPR-negative genomes, illustrating the strong association of *cas* genes and CRISPR loci (see Table 1). No CRISPR loci were identified in eukaryotic genomes and, as expected, no homologues of the *cas* genes were found in these genomes. The homology between the Cas proteins is illustrated by the alignments shown in Fig. 2. The amino acid sequences of the Cas proteins show some highly conserved amino acid residues or functional domains. The relationship between the four groups of Cas proteins was also recognized by the 'cluster of orthologous groups of proteins' (COG) of the NCBI and the 'pinned orthologue regions' of IGwit. Each of the four groups of Cas proteins almost perfectly matches a COG. The COG identification number of the Cas1 to Cas4 proteins are COG 1518, 1343, 1203 and 1468 respectively. The COGs 1343 and 1203 exactly match the groups of Cas2 and Cas3 proteins. The COG 1203 also contained additional orthologues beside the Cas3 proteins. Comparison of the genes in the pinned orthologue regions in which the four *cas* genes are located did not reveal additional common genes in these regions, indicating that the group of *cas* genes has no more than four members.

A query with Cas2 protein sequences using TBLASTN revealed two *cas2* genes in *P. horikoshii* and one *cas2* gene in *A. aeolicus* that were not annotated in these genome projects. The *cas2* genes of *P. horikoshii* were located at genome positions 153 147–153 401 next to Ph0173 (a *cas1* gene) and at 1 119 454–1 119 708 next to Ph1245 (a *cas1* gene). The *A. aeolicus cas2* gene is located at position 245 340–245 062 of the genome, next to the *cas1* gene Aq0369 (Fig. 1).

Comparison of the cas *genes and Cas proteins*

The *cas* genes were found to be located on either side of the CRISPR locus, and no preference for direction of the *cas* reading frames was observed. The *cas* genes of most species are present in the genome in the order *cas3–cas4–cas1–cas2* (Fig. 1), which might indicate transcriptional organization of the *cas* genes in these species. The *cas* gene cluster generally was found to be located within a few hundred of basepairs of the CRISPR locus. The largest distance (9 kbp) was found in *M. thermoautotrophicum*.

Species with multiple CRISPR loci with the same repeat sequence harbour *cas* genes adjacent to only one of the CRISPR loci. The species *Archaeoglobus fulgidus, Pyrococcus horikoshii, Pasteurella multocida* and *Streptococcus pyogenes* carry two classes of CRISPR loci with two dissimilar repeat sequences. In these species we found two different sets of *cas* genes, indicating that each class of CRISPR has its own set of *cas* genes. A comparison of the Cas proteins using CLUSTAL indicated that the two homologues of these four species did not cluster in four pairs, indicating that these Cas homologues are not more closely related to each other than to the Cas proteins of other species. This indicated that the *cas* genes of these species and the acompanying CRISPR loci have evolved separately, which is consistent with the idea that CRISPRs and the accompanying *cas* genes are functionally related. The absence of relatedness of the two Cas proteins from *A. fulgidus, P. horikoshii, P. multocida* and *S. pyogenes* is illustrated by the phylogenetic trees of each of the four families of Cas proteins at the website for the COGs (http://www.ncbi.nlm.nih.gov).

The function of the Cas proteins is not clear: only for Cas3 and Cas4 can a function be predicted. Searches with these two groups of Cas proteins in the Prosite database revealed conserved domains in their amino acid sequences. The predicted Cas3 proteins showed the seven functional domains of the superfamily 2 of helicases, indicating that the Cas3 homologues are members of this group of DNA-modifying proteins (Hall, 1999). Two of the seven functional domains are indicated in the alignment of Fig. 2. The Cas4 proteins showed similarity to the family of RecB exonucleases. In Cas4 most prominent were the three cysteine residues and the tyrosine at the carboxy terminus of most of the Cas4 proteins. It was suggested that the cysteine residues are involved in DNA binding, whereas the tyrosine residue might be involved in the formation of a covalent intermediate of the protein with the cleavaged DNA (Aravind *et al.*, 1999).

No similarity or conserved functional domains were identified for the Cas1 and Cas2 proteins. Of note is the fact that the predicted CasI proteins had strongly basic isoelectric points (pI) of between 9 and 10. The only exception was the Cas1 homologue of *M. tuberculosis*, which had an almost neutral pI of 7.6.

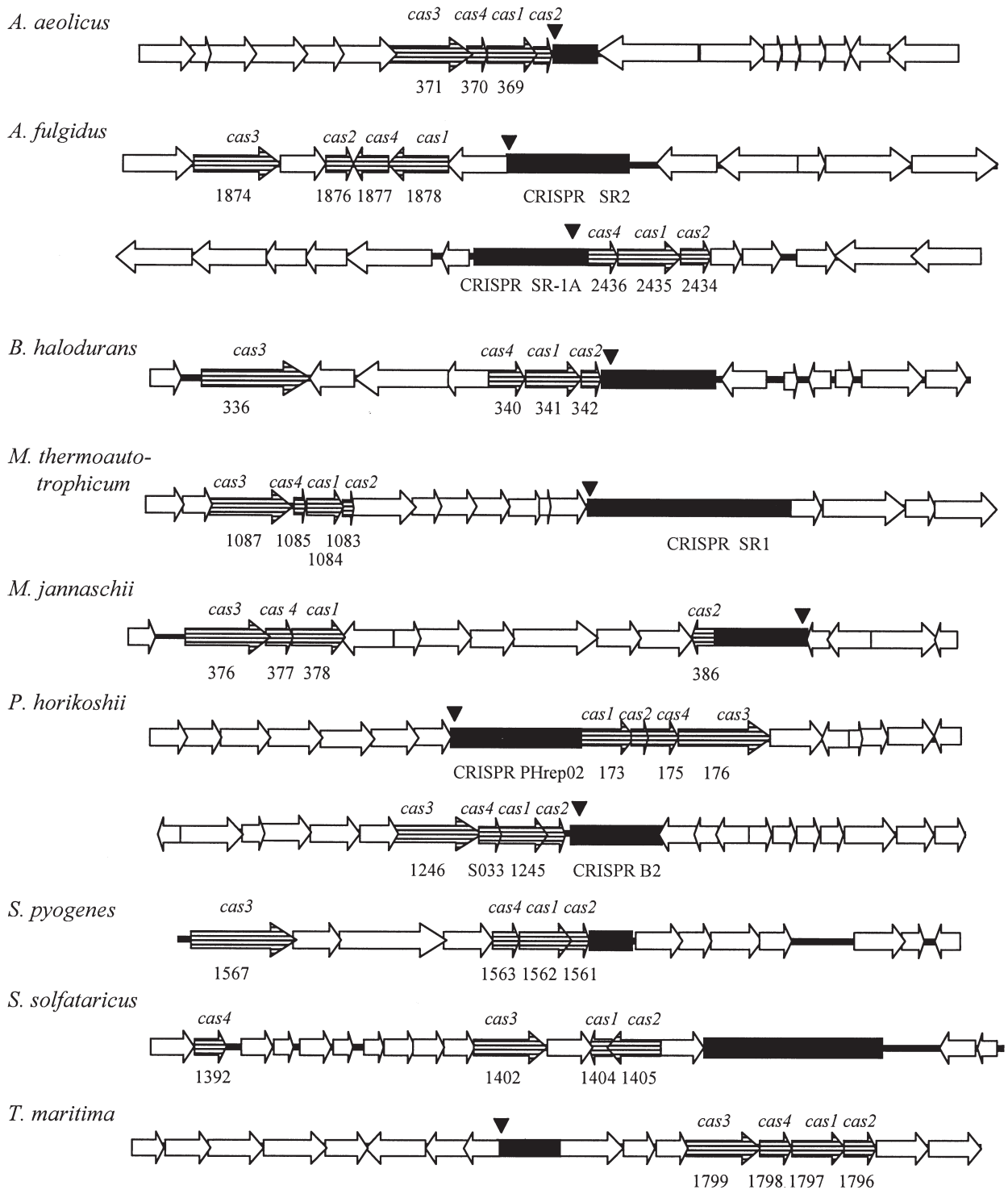

Fig. 1. Genetic organization of the CRISPR locus and the *cas* genes of the published genomes that contain all four *cas* genes. The CRISPR loci are depicted as black boxes. The position of the leader is indicated by a black triangle. The putative genes *cas1*, *cas 2, cas3* and *cas4* are indicated by dashed arrows. Other putative open reading frames (ORFs) are depicted as white arrows. The numbers below the ORFs indicate the gene numbers as assigned by the individual genome projects. The *cas2* genes of *P. horikoshii* and *A. aeolicus* have no gene number because these were not annotated by these genome projects.

Cas1 homologues

```
A. pernix                ..R.--------.......NA.L..G...LY......L....L.P.LG..H.......SL.LDA.E.FR..IVD..-..L...    270
A. aeolicus              ..R-------.........NA.I..G.S..Y......I....L.P.VGYLH.....R.SL.LDV.E.FK...VD.....LI..    237
A. fulgidus (Af187)      ..R........-.......NA.L..G.S.L.......V....L.P.AGFLH.....R.SL.IDL.E.FR..VVD.....LI..    271
A. fulgidus (Af2435)     ..R-------.........NA.I..G.S.LY......I....L.P.ISYLH.....R.SL.LDI.E.FK..VVD.....LV..    245
B. halodurans            ..R---------.......NA.L....S.LY......L....L...VGF.H.....R.SL.LDL.E..R....D.....LI..    266
C. jejuni                ..R-------.........N..L..G...........V....L.P.VG..H........L..DL.E.FR-..VD.....L...    239
E. coli                  ...................N..I....S.LY......I......P.ISFVH.......S...DI.D.....VV.......I..    238
E. coli O157             ...................N..I....S.LY......I......P.ISFIH.......S...DI.D.....VV.......I..    238
M. thermoautotrophicum   ..R-------.........NA.I..G.S.LY......L....L.P.ISYLH.....R.SL.LDL.E.FK..LID.....LI..    258
M. jannaschii            ..R-------.........NA.I....S.LY......L....L.P.VSYLH.....R.SL.LDL.E.FK..I.D.....LV..    246
M. tuberculosis          ..R-------.........NS.V..G.S.LY......I....L...IGFLH.....L..DL.E.WK..IID.....LI..    252
N. meningitidis          ..R---------.......NA.L......L.......L....L.P.LG..H........L..D..E..R....D.....L...    235
P. multocida             ..R---------.......NA.L......L.......L....L.P.LG..H........L..D..E..R..LVD.....L...    274
P. horikoshii (Ph0173)   ..R-------.........NA.I..G.S.LY......I....L.P.I.YLH.....R.SL.LDL.E.FK..I.D.....LV..    246
P. horikoshii (Ph1245)   ..R-------.........NA.I....S.LY......I....L.P.I.YLH.....R.SL.LDL.E.FK..VV......LV..    242
S. pyogenes   (Spy1047)  ..R-------.........NA.L..G...V.......L.........G..H...........DI.E.FR..LVD.........    236
S. pyogenes   (Spy1562)  ..R-------.........NA.L..G.S.L.......L....L...VG..H.....R.SL.LDL.E.FR..IVD.....LI..    265
S. solfataricus          ..R-----...........N..L..G...L.......V....L.P.IGFLH.....R.SL..DL.E.FR...VD.....L...    306
T. volcanium             ..R-------.........NA.I..G...LY......I....L.P.ISFLH.......SL.LDI.D.FK..IVE.....LV..    248
T. maritima              ..R-------.........N..I..G.S.LY......I....L.P.IGYLH.....R.SL.LDI.E.FK..VVD.....LV..    246
```

Cas2 homologues

```
A. aeolicus              .VILVYDV.........R..K..K.....-I..VQ.SVFEGEIT......L...L...I....D.V.IY..............-D   84
A. fulgidus (Af1876)     ..LVVYDI....-----R..RL.K.L....L..VQNSAF.GEL.......L...V.K.V....D.I.L..............GVE   84
A. fulgidus (Af2434)     YVIVAYDV...------R..RV.K.L...-L..VQNSLFEGELS......V...L...I....D.V.IY..............GIE   82
B. halodurans            Y.V........------R..KV.K.....-...VQNSVFE..V.......L...L...I....D.L.IY..............GIE   85
M. thermoautotrophicum   YLLIVYDV...------R..RV...L...-L..VQNSVFEGEVT......I...L.R.I....D.V.IY..............GLE   81
M. jannaschii            YVIIVYDV...------R..KI...L...-L..VQNSVFEGEVT......I...I.R.I....D.V.IY..............GLE  104
M. tuberculosis          .VLVIYDI....-----R...L.K.L.....-.VQ.SAFE..LT......L...I.R.A---.D.I.IY..............G-.  103
P. horikoshii(Phrep02)   YIVVVYDI....-----R..KV.K.L...-L..VQNSVFEGEVT......I...L.K.I....D.V.IY..............GIE   77
P. horikoshii(B2)        YVIVVYDV....-----R..R..K.L...-L...QNSVFEGELS......L...L.R.V...-D.V.IY..............G.D   77
S. pyogenes              YDV........------R...V.K.....-...VQNSLFEGELS......I...L...I....D.I..Y..............G..   85
S. solfataricus          YLI-.YDI...-----.R-.RV...L...-L..IQ.SVF.GDL.......V...L...I....E...I..............I.   90
T. maritima              YVI.VYDV...------R..KI.K.A...-L..VQNSVFE..VT......L...V.R.I....D.V..Y..............GVE   77
```

Cas3 homologues

```
                                                  motif V                    motif VI
A. pernix                ....................I.V.TQVLE..VDL.......TE...ID.VIQR.GR..R.......................    385
A. aeolicus              .................----V.V.TQLAE..LDL.......TE...ID.LIQR.GR..R.......................    556
A. fulgidus              ..........----------V.V.TQVIE..VDI.......TE......LIQR.GR..R.......................    382
B. halodurans            ....................I.V.TQLIE..VDV.........LD.I.Q..GR..R  .......................    600
E. coli                  ....................I.V.TQVVE..LDV.......T.....D.L.QR.GR..R.......................    691
E. coli O157             ....................V.I.TQVLE..VD..........VD.LIQR.GR..R.......................    695
M. thermoautotrophicum   ..........--........I..TQIIE..VDI.......TE...ID..IQR.GR..R.......................    653
M. jannaschii            ....................V.V.TQVIE..LD.......SE....D.LGQR.GR..R.......................    565
P. horikoshii (Ph0176)   -.................----I.V.TQVVE..LDI.......TE...ID.LIQR.GR..R.......................    535
P. horikoshii (Ph1246    ..................----..V.TQVIE..LDI.......TE...LD.LIQR.GR..R.......................    571
S. pyogenes              ....................I.L.TQLIE..VDV.........ID.IVQ..GR..R.......................    624
S. solfataricus          ....................I.I.TQVIE..V.I.......TE...I..IVQR.GR..R.......................    400
T. maritima              ..........----------L.V.TQVVEA.VDI.......D...VD.IVQ..GR..R.......................    565
```

Cas4 homologues

```
A. pernix                .....I........L....G.......I......K...V.........................PP.P...----C..C.....C..  212
A. aeolicus              ....QV.YYL.--.L...-GV....G.I.YPK......VEL.........I.............PP........--C..C.YYE.C..  177
A. fulgidus (Af2436)     ....QL.YYL.--YL...-GV.-..G.I.YPK......VEL.........L.............P-.P......--C..C.YYE.C..  167
A. fulgidus (Af1877)     ....QL..Y..--.V....-.V.-..G.L…….…..V.I.........L.............P........--C.YC...E.C..  189
B. halodurans            .....I...V..........I.--.G.I.Y........V.I........V.............P........--C..C.....C..  157
M. thermoautotrophicum   ....QL.YYL...YL....GI....G.I.YP.......V.L........L.............-PP.P.......C..C.Y.E.C..  181
M. jannaschii            ....QV.YYI.--YL....-GI.-....L.YPK......IEL.........I.............PP.P......--C..C.YYE.C..  170
P. horikoshii(Ph0175)    ....QL.YYL.--YL...-GI.-..G.I.YPK......IEL.........I.............PP.P......--C..C.Y.D.C..  162
P. horikoshii(s033)      ....Q..YYL.--YL...-GI.-......YPK......IEL........-V.............PP.P......--C..C.Y.E.C..   83
S. pyogenes              .....I...V.........-.I.-..G.L.Y........V..........V.............P........--C..C...D.C..  202
S. solfataricus          ....Q..YYL---YL....-GI.-..G.l.Y........LE.........V.............---CY.C....YC..  194
T. maritima              ....QL.YYL---Y.....GV.-..G.I..PK......VEL.........I.............PP.P.....--C..C.Y...C..  159
```

Fig. 2. Alignment of the amino acid sequences of putative Cas proteins 1–4 from published genomes. Only the regions with the highest homology are presented. The full alignments are available at the COG database. The number indicates the last amino acid residue of the alignment. Only matching residues of identical or chemically related amino acid residues are given. Non-matching residues are indicated with a dot. Gaps are indicated by a hyphen. The boxed regions in the Cas3 homologues are characteristic of helicases. In the Cas4 alignment all tyrosine residues in the cystein cluster at the carboxy termini of the proteins are indicated.The alignments were made with CLUSTAL using the neighbour-joining method. The identification numbers of the genes are indicated in Table 1.

CRISPR loci and *cas* genes are widely distributed among the prokaryotic species, and one might expect that the Cas proteins of closely related species to share greater amino acid sequence similarity than unrelated species. However, a neighbour-joining multiple alignment of the Cas proteins using CLUSTAL did not reveal such a relationship. The multiple alignment did not even group the Cas proteins of bacterial and archaeal origin.

The GC content of the *cas* genes did not differ significantly from the rest of the genome in which they were present. For example, the *cas* genes of *M. tuberculosis* had a relatively high GC content, like the genome, 59% and 65% respectively, whereas for the low GC content species *M. jannaschii* the corresponding figures were 30% and 31%.

Discussion

The first CRISPR locus was described 14 years ago in *E. coli* K12 as a sequence bordering the *iap* gene (Ishino *et al*., 1987). Since then CRISPR sequences of several other species have been published (Groenen *et al*., 1993; Mojica *et al*., 1995; Masepohl *et al*., 1996; Hoe *et al*., 1999), and recently the CRISPR loci were recognized as a family of repeats (Mojica *et al*., 2000).

By systematically searching the public DNA database specifically for the common structural motifs of CRISPR loci, we identified CRISPR loci in more than 40 prokaryotic species. The common structural characteristics of CRISPR loci are: (i) the presence of multiple short direct repeats, which show no or very little sequence variation within a given locus; (ii) the presence of non-repetitive spacer sequences between the repeats of similar size; (iii) the presence of a common leader sequence of a few hundred basepairs in most species harbouring multiple CRISPR loci; (iv) the absence of long open reading frames within the locus; and (v) the presence of the *cas1* gene accompanied by the *cas2, cas3* or *cas4* genes in CRISPR-containing species.

CRISPR regions were found in the genomes of species belonging to the superkingdoms of the Archaea and the Bacteria, but not in the Eukarya or in viral genomes. Thus, CRISPRs appear to be an exclusive prokaryotic feature. Nevertheless, Mojica *et al*. (2000) described a CRISPR-like motif in the *Vicia faba* mitochondrial genome. This CRISPR-like sequence consisted of five similar but non-identical repeats of 43 bp that were interspaced by four small spacers, one each of 31 and 25 bp and two of 16 bp. We did not consider the *Vicia faba* CRISPR-like motif to be a true CRISPR, because the spacer sequences were closely related to each other. The two 16-bp spacer sequences were identical for 62% and the two longer spacer sequences were identical for 50%. Thus, the *Vicia faba* sequence should be considered an imperfect direct repeat and not a CRISPR.

We found CRISPR regions in most completely sequenced archaeal genomes and in half of the sequenced bacterial genomes. This suggests that CRISPRs are more prevalent in the Archaea, but it should be noted that these archaeons are mostly thermo-extremophilic organisms. There is a tendency for thermo-extremophilic organisms such as these archaeons, but also the bacteria *A. pernix, A. aeolicus* and *T. maritima,* to have more and larger CRISPRs than mesophilic members of the Bacteria (Nelson *et al*., 1999).

The presence or absence of CRISPRs is not characteristic for a particular taxon among the Bacteria (see Table 1). For instance, *M. tuberculosis* and *M. leprae* belong to the same genus, but a CRISPR locus is present only in *M. tuberculosis*. The same holds true for the family of the *Bacillus* species *B. subtilis* and *B. halodurans*, and the Enterobacteriaceae, among which the species *E. coli* and *S. enterica* contain CRISPR loci whereas other species of this family, such as *Klebsiella pneumoniae*, do not contain CRISPR loci

The frequently observed presence of multiple CRISPR loci in many prokaryotes at different locations in the genome may suggest that these genetic elements are mobile. Furthermore, we found that taxonomically unrelated species such as *E. coli* and *M. avium* harbour identical or nearly identical CRISPR sequences. CRISPRs may have been disseminated among genetically unrelated microorganisms by lateral DNA transfer. In support of lateral DNA transfer of the *cas* genes and CRISPRs was the absence of a phylogenetic relationship between the Cas proteins that corresponds to the phylogenetic relationship of the species. Also in support of lateral gene transfer of CRISPR loci is the location of the *cas* genes of *T. maritima* on an archaeal island that was suggested to be subjected to lateral gene transfer (Nelson *et al*., 1999) The finding that the GC content of the *cas* genes is similar to the rest of the genomes in which they reside is not evidence against lateral gene transfer because it has been demonstrated that GC content is a poor indicator of lateral gene transfer (Koski *et al*., 2001).

Two mechanisms, transposition and recombination, might be involved in the spread of CRISPR loci. In favour of transposition within the genome was the finding of multiple CRISPR loci in genomes, each with its own set of unique spacer sequences. This indicates that each CRISPR locus has evolved independently. A possible mechanism might be that the leader sequence is translocated in the genome in which it grows by duplicating repeats and generating its own unique set of spacer sequences. This is in agreement with the finding that the sequences flanking the CRISPR loci and their leader

sequences show no sequence similarity, suggesting that the CRISPR loci were inserted at that position

A common leader sequence could be recognized in most species that harboured two or more CRISPR loci. The leader sequence was located directly adjacent to the CRISPR locus, indicating that the leader and the CRISPR constitute a single genetic entity that behave as a mobile element. Curiously, the spacers interspersing the CRISPRs may not be part of these putative mobile elements, as the spacer sequences within different loci generally do not share any homology. It is possible that the spacer sequences evolve separately for each of the CRISPR loci in a genome by an as yet unknown mechanism that may involve the action of the Cas proteins.

In this study we identified four CRISPR-associated genes, *cas1* to *cas4*. The presence of the *cas* genes was invariably associated with the presence of CRISPR loci in the genome. The *cas1* gene was found in all species with CRISPR loci, whereas the other three *cas* genes were present in most but not all CRISPR-containing genomes. The strict association between CRISPRs and *cas* homologues in sequenced genomes is highly suggestive for a functional relationship between these genes and the CRISPRs. The amino acid sequences of the Cas3 proteins contain the seven motifs that are characteristic for the superfamily 2 of the helicases, proteins that power DNA unwinding. Homologues of this superfamily may have functions in DNA repair, transcriptional regulation, chromosome segregation, recombination and chromatin remodelling (Hall and Matson, 1999). The Cas4 proteins showed similarity to RecB exonucleases. In *E. coli* the RecB nuclease is part of the RecBCD complex, which has several activities, including recombinase activity (Jockovich and Myers, 2001). The cysteine and tyrosine residues of Cas4 might be involved in DNA binding, which strongly suggests a direct interaction with DNA. In addition, the Cas1 proteins have a remarkably high pI, which is characteristic of proteins that bind to DNA, such as histone-like proteins and transposases. The highly basic nature of the Cas1 proteins suggests affinity of these proteins for nucleic acid, and because the *cas1* gene is invariably associated with the presence of CRISPR loci one may speculate that the Cas1 proteins could have the CRISPR DNA sequences as target. This is consistent with the findings of Mojica *et al.* (2000), who suggested that the DNA repeat sequenses of the CRISPR loci have characteristics that are exhibited by recognition sites for DNA-binding proteins.

Very little experimental information is available about the biological function of CRISPRs. Mojica *et al*. (1995) introduced a multicopy CRISPR-containing plasmid into *Haloferax mediterranei*, an archaeal species that carries CRISPR loci in the chromosome and on a resident megaplasmid. The resulting recombinant showed growth retardation, which was accompanied by unequal chromosome partitioning, suggestive of a defective replicon partitioning upon cell division. In contrast, no effect on *in vitro* growth was observed when the CRISPR locus of *M. tuberculosis* was introduced on a multicopy plasmid, or when the single CRISPR locus of *M. tuberculosis* was deleted from the genome. This suggests that the CRISPR of *M. tuberculosis* is not essential for growth under laboratory conditions (R. Jansen and K. Kremer, unpublished data). Because of the scarcity of these observations, the biological function of CRISPRs remains obscure.

The few insights into the function of the Cas proteins were derived from comparative studies. In addition, little is known about the expression of the genes. Recently, transcription of the *cas1* and *cas3* gene has been demonstrated in *E. coli* K12 by DNA arrays (presented on the website of the *E. coli* genome project of the Univeristy of Wisconsin-Madison; http://www.genome.wisc.edu).This is the first indication that the Cas proteins are expressed under laboratory growth conditions.

Repetitive DNA generally is associated with a high degree of DNA polymorphism (van Belkum *et al*., 1998). A high degree of strain-dependent DNA polymorphism within the CRISPR loci has been observed in *M. tuberculosis* and in *S. pyogenes* (Kamerbeek *et al*., 1997; Fang *et al*., 1998; Hoe *et al*., 1999), and preliminary investigations indicate that similar polymorphisms exist in other species, such as *E. coli* and *S. enterica* (R. Jansen, unpublished observations). Recently, the nature of rearrangements in the CRISPR locus of *M. tuberculosis* has been studied by comparison of the sequences of complete CRISPR loci in clinical isolates (van Embden *et al*., 2000). This study indicated that the observed polymorphism can be explained by successive deletions of single or contiguous multiple repeats of a CRISPR and its adjacent spacer from a primordial CRISPR locus. These deletions are probably mediated by homologous recombination between neighbouring or distant CRISPRs, by slippage during DNA replication and by transposase activity mediated by insertion elements.

Although variation by deletion from a primordial CRISPR locus may explain the genetic variation among bacterial isolates (van Embden *et al*., 2000), the genesis of the hypothetical primordial CRISPR locus is enigmatic, and the origin of the spacer sequences remains unknown. Spacer sequences do not occur in the genome beyond the CRISPR locus (van Embden *et al*., 2000). Furthermore, duplications of spacers in the CRISPR loci are very rare. Only in *M. tuberculosis* and *M. thermoautotrophicum* were a few duplications found. The finding in this study that the CRISPR loci were strictly associated with a set of homologous genes, one of which has nucleic acid helicase motifs (the Cas3 homologues), one of which has exonuclease activity (the Cas4 homologues) and one

of which has a high pI (the CasI homologues), as is often found for DNA-binding proteins, may be suggestive of a role for the Cas proteins in the genesis of CRISPR loci.

Experimental procedures

Computer programs

Nucleotide sequence databases were seached for CRISPR motifs using the PATSCAN program at the server of the Mathematics and Computer Science Division of the Argonne National Laboratory, Argonne, IL USA, at website http://wwwunix.mcs.anl.gov/compbio/PatScan/HTML/patscan.html and at the server of the RIVM for searches in the EMBL/GenBank database and unfinished genome databases of specific species. The algorithm used for disclosing CRISPR motif was p1 = *a* . . . *b c* . . . *d* p1 *c* . . . *d* p1 *c* . . . *d* p1, where *a* and *b* are the lower and upper size limit of the repeat p1 and *c* and *d* are the lower and upper size limit of the spacer sequences. The values of *a*, *b*, *c* and *d* were varied from 15 to 70 bp at increments of 5 bp.

The NBLAST and TBLASTN programs for comparison of nucleotide sequences and protein sequences with the EMBL/GenBank database and the unfinished genome database and searches in the Cluster of Orthologous Groups of Proteins (COG) database was used at the website http://www.ncbi.nlm.nih.gov/cgi-bin/blast of the National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD, USA. The collection of 'pinned orthologue regions' from the WIT project was used at website http://wit.integratedgenomics.com/IGwit.

The Prosite database was searched for function domains in proteins at website http://www.expasy.ch/prosite.

The DNA STAR software package, which includes Meg-align for sequence comparison and Seq-man for alignment of contigs, was purchased from DNASTAR Inc., Madison, WI, USA. The alignments were done by the neighbour-joining method using the CLUSTAL program (Higgins and Sharp, 1989). The penalty settings for the multiple alignment were: gap penalty 10 and gap length 10.

Nucleotide sequences

Sequences were obtained from the EMBL/GenBank data-base and from the sequence data from genome projects in progress that are accessible through the NCBI website (http://www.ncbi.nlm.nih.gov/Microb_BLAST/unfinishedgenome.html). The EMBL/GenBank accession numbers of the sequences that contained CRISPR sequences are listed in Table 1. The identification numbers of the genome projects in progress are indicated with their entrez nucleotides (EN) number.

Acknowledgements

We enjoyed pleasant discussions with Francisco Mojica of the University of Alicante, Spain, about the renaming of CRISPRs. This work was financially supported by the Dutch Foundation for Technical Sciences and the European Union project on the development of novel methodology and nomenclature for the identification of *M. bovis* strains.

References

Aravind, L., Walker, D.R., and Koonin, E.V. (1999) Conserved domains in DNA repair proteins and evolution of repair systems. *Nucleic Acids Res* **27:** 1223–1242.

Bachellier, S., Clement, J.M., Hofnung, M., and Gilson, E. (1997) Bacterial interspersed mosaic elements (BIMEs) are a major source of sequence polymorphism in *Escherichia coli* intergenic regions including specific associations with a new insertion sequence. *Genetics* **145:** 551–562.

van Belkum, A., Scherer, S., van Alphen, L., and Verbrugh, H. (1998) Short-sequence DNA repeats in prokaryotic genomes. *Microbiol Mol Biol Rev* **62:** 275–293.

Belland, R.J., Morrison, S.G., Carlson, J.H., and Hogan, D.M. (1997) Promoter strength influences phase variation of neisserial opa *genes. Mol Microbiol* **23:** 123–135.

Bult, C.J., White, O., Olsen, G.J., Zhou, L., Fleischmann, R.D., Sutton, G.G., *et al*. (1996) Complete genome sequence of the methanogenic archaeon *Methanococcus jannaschii. Science* **273:** 1058–1073.

Correia, F.F., Inouye, S., and Inouye, M. (1988) A family of small repeated elements with some transposon-like properties in the genome of *Neisseria gonorrhoeae. J Biol Chem* **263:** 12194–12198.

Dybvig, K. (1993) DNA rearrangements and phenotypic switching in prokaryotes. *Mol Microbiol* **10:** 465–471.

van Embden, J.D., van Gorkom, T., Kremer, K., Jansen, R., van der Zeijst, B.A., and Schouls, L.M. (2000) Genetic variation and evolutionary origin of the direct repeat locus of *Mycobacterium tuberculosis* complex bacteria. *J Bacteriol* **182:** 2393–2401.

Fang, Z., Morrison, N., Watt, B., Doig, C., and Forbes, K.J. (1998) IS6110 transposition and evolutionary scenario of the direct repeat locus in a group of closely related *Mycobacterium tuberculosis* strains. *J Bacteriol* **180:** 2102–2109.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., *et al*. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269:** 496–512.

Groenen, P.M., Bunschoten, A.E., and van Soolingen, D., and van Embden, J.D. (1993) Nature of DNA polymorphism in the direct repeat cluster of *Mycobacterium tuberculosis:* application for strain differentiation by a novel typing method. *Mol Microbiol* **10:** 1057–1065.

Hall, M.C., and Matson, S.W. (1999) Helicase motifs: the engine that powers DNA unwinding. *Mol Microbiol* **34:** 867–877.

Higgins, D.G., and Sharp, P.M. (1989) Fast and sensitive multiple sequence alignments on a microcomputer. *Comput Appl Biosci* **5:** 151–153.

Hoe, N., Nakashima, K., Grigsby, D., Pan, X., Dou, S.J., Naidich, S., *et al*. (1999) Rapid molecular genetic subtyping of serotype M1 group A *Streptococcus* strains. *Emerg Infect Dis* **5:** 254–263.

Hulton, C.S., Higgins, C.F., and Sharp, P.M. (1991) ERIC sequences: a novel family of repetitive elements in the

genomes of *Escherichia coli, Salmonella typhimurium* and other enterobacteria. *Mol Microbiol* **5:** 825–834.

Ishino, Y., Shinagawa, H., Makino, K., Amemura, M., and Nakata, A. (1987) Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in *Escherichia coli,* and identification of the gene product. *J Bacteriol* **169:** 5429–5433.

Jockovich, M.E., and Myers, R.S. (2001) Nuclease activity is essential for RecBCD recombination in *Escherichia coli. Mol Microbiol* **41:** 949–962.

Kamerbeek, J., Schouls, L., Kolk, A., van Agterveld, M., van Soolingen, D., Kuijper, S., *et al*. (1997) *Simultaneous detection and strain differentiation of* Mycobacterium tuberculosis *for diagnosis and epidemiology. J Clin Microbiol* **35:** 907–914.

Klenk, H.P., Clayton, R.A., Tomb, J.F., White, O., Nelson, K.E., Ketchum, K.A., *et al*. (1997) The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus. Nature* **390:** 364–370.

Koski, L.B., Morton, R.A., and Golding, G.B. (2001) Codon bias and base composition are poor indicators of horizontally transferred genes. *Mol Biol Evol* **18:** 404–412.

Lupski, J.R., Roth, J.R., and Weinstock, G.M. (1996) Chromosomal duplications in bacteria, fruit flies, and humans. *Am J Hum Genet* **58:** 21–27.

Martin, B., Humbert, O., Camara, M., Guenzi, E., Walker, J., Mitchell, T., *et al*. (1992) A highly conserved repeated DNA element located in the chromosome of *Streptococcus pneumoniae. Nucleic Acids Res* **20:** 3479–3483.

Masepohl, B., Gorlitz, K., and Bohme, H. (1996) Long tandemly repeated repetitive (LTRR) sequences in the filamentous cyanobacterium *Anabaena* sp. PCC 7120. *Biochim Biophys Acta* **1307:** 26–30.

Mojica, F.J., Ferrer, C., Juez, G., and Rodriguez-Valera, F. (1995) Long stretches of short tandem repeats are present in the largest replicons of the Archaea *Haloferax mediterranei* and *Haloferax volcanii* and could be involved in replicon partitioning. *Mol Microbiol* **17:** 85–93.

Mojica, F.J., Diez-Villasenor, C., Soria, E., and Juez, G. (2000) Biological significance of a family of regularly spaced repeats in the genomes of *Archaea, Bacteria* and mitochondria. *Mol Microbiol* **36:** 244–246.

Nakata, A., Amemura, M., and Makino, K. (1989) Unusual nucleotide arrangement with repeated sequences in the *Escherichia coli* K-12 chromosome. *J Bacteriol* **171:** 3553–3556.

Nelson, K.E., Clayton, R.A., Gill, S.R., Gwinn, M.L., Dodson, R.J., Haft, D.H. *et al*. (1999) Evidence for lateral gene transfer between *Archaea* and *Bacteria* from genome sequence of *Thermotoga maritima. Nature* **399:** 323–329.

Smith, D.R., Doucette-Stamm, L.A., Deloughery, C., Lee, H., Dubois, J., Aldredge, T. *et al*. (1997) Complete genome sequence of *Methanobacterium thermoautotrophicum* deltaH: functional analysis and comparative genomics. *J Bacteriol* **179:** 7135–7155.