Genomic rearrangements during evolution of the obligate intracellular parasite *Rickettsia prowazekii* as inferred from an analysis of 52015 bp nucleotide sequence

Jan O. Andersson and Siv G. E. Andersson

Author for correspondence: Siv G. E. Andersson. Tel: +46 18 174379. Fax: +46 18 557723.
e-mail: siv@skydancer.bmc.uu.se

Department of Molecular
Biology, Biomedical
Center, Uppsala University,
Uppsala, S-751 24, Sweden

In this study a description is given of the sequence and analysis of 52 kb from the 1.1 Mb genome of *Rickettsia prowazekii*, a member of the α -Proteobacteria. An investigation was made of nucleotide frequencies and amino acid composition patterns of 41 coding sequences, distributed in 10 genomic contigs, of which 32 were found to have putative homologues in the public databases. Overall, the coding content of the individual contigs ranged from 59 to 97%, with a mean of 81%. The genes putatively identified included genes involved in the biosynthesis of nucleotides, macromolecules and cell wall structures as well as citric acid cycle component genes. In addition, a putative identification was made of a member of the regulatory response family of two-component signal transduction systems as well as a gene encoding haemolysin. For one gene, the homologue of *metK*, an internal stop codon was discovered within a region that is otherwise highly conserved. Comparisons with the genomic structures of *Escherichia coli*, *Haemophilus influenzae* and *Bacillus subtilis* have revealed several atypical gene organization patterns in the *R. prowazekii* genome. For example, *R. prowazekii* was found to have a unique arrangement of genes upstream of *dnaA* in a region that is highly conserved among other microbial genomes and thought to represent the origin of replication of a primordial replicon. The results presented in this paper support the hypothesis that the *R. prowazekii* genome is a highly derived genome and provide examples of gene order structures that are unique for the *Rickettsia*.

Keywords: genome sequencing, *Rickettsia prowazekii*

INTRODUCTION

Rickettsia prowazekii is a member of the α -Proteobacteria and is thought to have a particularly close phylogenetic affiliation with the mitochondria (Gray, 1995; Gupta, 1995; Olsen *et al.*, 1994; Viale & Arakaki, 1994). Within the α -Proteobacteria, genome sizes vary by almost an order of magnitude. At one extreme, we find the *Rickettsia* with genome sizes in the 1 Mb range (Eremeeva *et al.*, 1993; Roux & Raoult, 1993), and at the other extreme *Bradyrhizobium japonicum* with a genome size of 8.7 Mb (Kündig *et al.*,

1993). It is thought that *R. prowazekii* evolved from an ancestor with a much larger genome (Andersson & Kurland, 1995).

A first survey of the *R. prowazekii* genome centred on a set of small random fragments representing in aggregate more than 200 kb of sequence data, from which more than 150 genes were identified (Andersson *et al.*, 1996). Additional studies have revealed several unusual features of the genomic architecture of *R. prowazekii*: First, genes normally present in multiple copies in larger genomes, such as the rRNA genes (*rrs* and *rrl*) and the gene encoding elongation factor Tu (*tuf*), are present in single copies in the genome of *R. prowazekii* (Andersson *et al.*, 1995; Pang & Winkler, 1993; Syvänen *et al.*, 1996). In addition, each of these single copy genes are located within gene arrangements that are unique for

The EMBL accession numbers for the sequences reported in this paper are Y11777–Y11786.

Rickettsia spp. As a consequence, several ancient genomic motifs common to other bacteria are absent from this genome (Andersson *et al.*, 1995; Syvänen *et al.*, 1996). Finally, there is a relatively large amount of non-coding sequence that makes up 30–40% of the genome (Andersson & Sharp, 1996; Andersson *et al.*, 1996).

The analysis of genomic fragments carried out to date suggest that non-coding sequences have a G + C content significantly lower on average (ca 22%) than that of coding sequences (ca 32%) in *R. prowazekii* (Andersson & Sharp, 1996; Andersson *et al.*, 1996). Accordingly, this genome appears to be a mosaic of relatively high G + C coding sequences scattered in a background of lower G + C non-coding sequences. Furthermore, the biased composition of the previously sequenced genes of *R. prowazekii* is reflected in a bias towards codons that are A + T-rich, with the consequence that amino acid sequences in proteins are enriched for the corresponding amino acids (Andersson & Sharp, 1996).

These previously characterized features of coding sequences in *R. prowazekii* have been applied here to new genome sequence data to identify ORFs as genes. The aim of the sequence analysis presented was to serve as a pilot experiment for our systematic large-scale genome sequencing project and to provide information about nucleotide frequencies, sequence similarities, gene organization patterns and coding content of the *R. prowazekii* genome.

METHODS

DNA sequencing. Genomic DNA was prepared from the Madrid E strain of *Rickettsia prowazekii* which was originally isolated in Madrid from a patient who died in 1941 with epidemic typhus. Propagation of *R. prowazekii* in the yolk sacs of embryonated hen eggs and purification of DNA was according to standard procedures (Winkler, 1976). Genomic clones were isolated from a library of *R. prowazekii* constructed in λ Zap II as described previously (Andersson *et al.*, 1995). Plasmids were recovered by excision *in vivo*. Plasmid DNA was isolated using a Jetstar plasmid maxi kit (Genomed). DNA sequences were determined by double-stranded dideoxy sequencing using modified T7 DNA polymerase (Sequenase) and fluorescently labelled forward and reverse vector primers as well as specific oligomers with fluorescent dATP (Pharmacia). PCR amplification of the *metK* gene was carried out under standard conditions using 0.5 pg genomic DNA μl^{-1} and 2 μM primers (nt 313–965). The PCR products were purified using the Wizard PCR Preps DNA Purification System (Promega) and they were sequenced by an autocycle sequencing protocol using the Thermo Sequenase Fluorescent Labelled Primer Cycle Sequencing Kit (Amersham). In each reaction 0.1 μg of PCR product and 2 pmol fluorescent primer were used. The products of the sequencing reactions were separated and analysed with the aid of an ALF Sequencer (Pharmacia). On average, raw sequence data corresponding to a 2.5-fold coverage on each strand was collected.

Southern blot hybridization. As hybridization probe, we used 100 ng of *R. prowazekii metK* PCR product prepared as described above. Labelling of the PCR product with horseradish peroxidase, Southern blot hybridization and detection

using chemiluminescence were performed using the ECL direct nucleic acid labelling and detection system according to the instructions of the manufacturer (Amersham).

Computer analysis. Searches of public databases such as GenBank/EMBL, NBRF-PIR and SWISS-PROT for homologous sequences were performed at the National Center for Biotechnology Information BLAST network server (<http://www.ncbi.nlm.nih.gov/>) using the BLAST program (Altschul *et al.*, 1990). Sequence alignments were performed using CLUSTAL W (Higgins *et al.*, 1992; Thompson *et al.*, 1994). Base frequencies and codon usage statistics were calculated using the programs CODONS (Lloyd & Sharp, 1992) and XBBTOOLS (T. Sicheritz-Pontén, Uppsala, unpublished). Percentage amino acid identities were calculated with the help of the algorithm ALIGN (Myers & Miller, 1988).

RESULTS AND DISCUSSION

DNA sequence analysis

We analysed sequences derived from a total of 10 fragments from the *R. prowazekii* genome covering 52015 bp and corresponding to about 4.7% of the genome (Eremeeva *et al.*, 1993). The overall G + C content of these fragments was 0.295 (Table 1), in good agreement with the estimated genomic G + C content of 0.29 (Schramek, 1972; Tyeryar *et al.*, 1973). The possibility that any of the sequence examined was derived from contaminating chicken genomic DNA was excluded on the basis of low overall G + C content of all sequences and the observation that, in all cases but one, the most similar protein sequences to the ORFs within the contigs were derived from bacterial genomes. Within these sequence fragments, we identified 41 ORFs of which 32 were complete with both start and stop signals. One of these, identified as *metK*, was found to contain an internal UAG stop codon (for discussion, see below). Four ORFs lacked a stop signal and five lacked a start codon because they were located at the 3' or 5' end of the contigs, respectively. The lengths and orientations of the ORFs within their respective sequence contigs is shown schematically in Fig. 1.

Nucleotide frequencies and amino acid composition patterns

The nucleotide frequencies of all of the ORFs identified here were compared with the nucleotide frequencies determined previously for a reference set of coding and non-coding DNA of *R. prowazekii* (Andersson & Sharp, 1996). We found that the G + C content of the ORFs ranged from 24% (RP001) to 38% (RP005) with a mean of 30.5% (Table 1). These values were within the range observed for the reference set of coding sequences: 25–42% with a mean of 32.5% (Andersson & Sharp, 1996). Furthermore, in each ORF we observed the characteristic gradient of G + C content values at the three codon positions, with mean values of 41.4, 31.4 and 18.6%, respectively. The G + C content at synonymous third codon positions (GC_3s) varied between the ORFs from 11 to 21% (Table 1), but this variation was also within the range (12–25%) determined pre-

Table 1. Base composition statistics of *R. prowazekii* gene sequences

Accession no.	ORF	No. of codons	GC _C *	GC _{3s} *	AT/GC†	Stop codon
Y11777	RP001	141	0.24	0.13	4.7	UAA
	RP002	128	0.25	0.15	5.1	UAG
	RP003	212	0.29	0.21	4.4	UAA
	RP004	386	0.34	0.17	1.4	UAA
	RP005	292	0.38	0.18	1.0	UAG
	RP006	266	0.26	0.15	4.2	UAA
	RP007	152	0.29	0.17	3.1	UAA
	RP008	294	0.31	0.14	1.8	UGA
	RP009	290	0.31	0.14	2.3	UAG
	RP010	145	0.27	0.12	2.9	
Y11778	RP011	212	0.31	0.19	2.5	UAA
	RP012	85	0.35	0.21	1.5	UAG
	RP013	215	0.32	0.19	2.1	UGA
	RP014	420	0.35	0.21	1.6	UAA
	RP015	297	0.34	0.16	1.6	UAA
	RP016	180	0.26	0.14	5.3	UAA
	RP017	303	0.30	0.17	3.0	UAA
	RP018	177	0.34	0.20	1.9	UGA
Y11785	RP019	434	0.31	0.15	2.0	UAA
	RP020	780	0.34	0.19	1.7	
Y11786	RP021	601	0.30	0.16	2.7	UGA
	RP022	295	0.27	0.17	4.1	UAG
	RP023	268	0.31	0.18	2.6	UAA
Y11779	RP024	1011	0.33	0.16	1.8	UAA
	RP025	297	0.24	0.14	4.5	UGA
	RP026	60	0.27	0.12	2.7	UAG
	RP027	180	0.29	0.13	2.7	UAA
	RP028	216	0.34	0.19	2.0	
Y11780	RP029	287	0.33	0.17	1.6	UAG
	RP030	292	0.30	0.17	3.2	UAA
	RP031	412	0.31	0.17	2.4	UAA
	RP032	236	0.29	0.14	3.1	UAA
	RP033	635	0.35	0.20	1.8	UAA
Y11781	RP034	250	0.30	0.16	2.5	UGA
Y11782	RP035	353	0.30	0.14	2.2	UGA
Y11783	RP036	95	0.28	0.11	3.0	
	RP037	480	0.30	0.18	2.5	UAG
	RP038	989	0.29	0.17	2.9	UAA
Y11784	RP039	303	0.30	0.14	2.4	UAA
	RP040	867	0.30	0.17	2.7	UAA
	RP041	380	0.36	0.19	1.5	UAA

* GC_C, GC_{3s}, G + C content in coding regions and in synonymous third codon positions, respectively.

† AT/GC, relative amino acid usage.

viously for the reference set of coding sequences (Andersson & Sharp, 1996). The total number of termination codons identified in these ORFs was 37 and of these, 22 were UAA (59%), 7 were UGA (19%) and 8 were UAG (22%). This distribution of termination codons was in good agreement with previous estimates of termination codon frequencies for the reference set of genes from *R. prowazekii*. These were 70% UAA, 15% UGA and 15% UAG (Andersson & Sharp, 1996).

It has previously been shown that the amino acid composition of *R. prowazekii* proteins has been influenced by A + T mutation bias to an extent that is inversely correlated with the degree of sequence conservation for individual gene families (Andersson & Sharp, 1996). The ratio of the frequency of amino acids encoded by A + T-rich codons (Asn, Ile, Lys, Phe, Tyr) divided by the frequency of amino acids encoded by G + C-rich codons (Ala, Gly, Pro), AT/GC, is a simple

Fig. 1. Schematic representation showing the organization of ORFs in the sequence contigs Y11777, Y11778, Y11785, Y11786, Y11779, Y11780, Y11781, Y11782, Y11783 and Y11784. Thin lines indicate non-coding regions; boxes indicate coding regions. The lengths and designations of the clones that were sequenced are shown. The genomic proximity of clones J25 and A235, J20 and A497, J35 and A393, and A393 and A308 was verified by PCR analysis using genomic DNA as template (data not shown).

measure of the effect of mutation bias on amino acid composition. In *Escherichia coli* the AT/GC values for different genes tend to cluster around 1.0 (Andersson & Sharp, 1996). In contrast, the mean value for the reference set of *R. prowazekii* genes is 1.8. The ORFs identified in this study encoded putative proteins with

amino acid compositions that were typical of *R. prowazekii* proteins, with AT/GC ratios ranging from 1.0 to 5.1 with a mean of 2.7. In summary, the 41 annotated ORFs described here resembled the reference set of *R. prowazekii* genes with respect to their idiosyncratic nucleotide distribution patterns and with

Table 2. Comparison of coding and non-coding genomic sequence of *R. prowazekii*

Accession no.	L_T^*	$GC_T^\dagger$	L_C^*	$GC_C^\dagger$	L_N^*	$GC_N^\dagger$	Coding fraction
Y11777	8781	0.29	7118	0.30	1663	0.21	0.81
Y11778	7653	0.30	5682	0.32	1971	0.23	0.74
Y11785	3811	0.32	3647	0.32	164	0.23	0.96
Y11786	3908	0.29	3503	0.30	405	0.23	0.90
Y11779	5433	0.31	5287	0.31	146	0.23	0.97
Y11780	9460	0.29	5601	0.32	3859	0.25	0.59
Y11781	945	0.29	882	0.31	63	0.13	0.93
Y11782	1230	0.30	1126	0.30	104	0.28	0.92
Y11783	4930	0.29	4704	0.30	226	0.25	0.95
Y11784	5864	0.30	4659	0.31	1205	0.24	0.79
Total	52015	0.30	42209	0.31	9806	0.24	0.81

* L_T , L_C , L_N , Length of sequences in complete fragments, coding and non-coding regions, respectively.

† GC_T , GC_C , GC_N , G + C content of sequences in complete fragments, coding and non-coding regions, respectively.

respect to the characteristically biased amino acid compositions of their translation products. Accordingly, we suggest that all 41 ORFs are genes.

In addition, we analysed sequences located at the terminal regions of the sequence contigs to determine whether or not they were coding or non-coding sequences. The putative, short ORFs thus identified have not been included in Table 1. In contig Y11781, we found a short ORF of 129 bp immediately downstream of RP034 which, because of its relatively high G + C content and typical base composition pattern ($GC3_s = 0.22$), was likely to correspond to the 5' region of a gene. Similarly, we observed a short ORF of 172 bp located upstream of RP001 but on the opposite strand. Also, for this ORF, base frequencies and composition patterns were as found in other genes of *R. prowazekii* ($GC3_s = 0.11$). Contig Y11782 overlapped with the previously sequenced contig U02603. There was an annotated ORF of unknown function in contig U02603 located immediately downstream of RP035 in the region overlapping the two contigs. The sequence corresponding to this ORF was annotated as coding DNA in contig Y11782.

The G + C content of DNA segments between genes (intergenic regions) varied from 12.4% (the spacer of RP002 and RP003) to 31.4% (the spacer of RP009 and RP010) with a mean of 23.5%. This value was similar to, albeit a little higher than, the estimated mean of 21.6% for the intergenic reference dataset (Andersson & Sharp, 1996). Nevertheless, the mean G + C content of these sequences was clearly distinguishable from that of coding sequences, which was on average over 30% G + C. Overall, the mean G + C content of the regions without identifiable ORFs was found to be 23.6%. The lack of ORFs of any significant length, the low G + C content values and the absence of homologous sequences in the public databases suggested that these sequences correspond to non-coding DNA.

Coding content

Of 52015 bp of sequence data analysed, 42209 bp were found to be located within coding regions and 9806 within non-coding regions (Table 2). This corresponded to a coding fraction of 81%, which is somewhat higher than the value estimated for the genome as a whole, 60–70% (Andersson & Sharp, 1996; Andersson *et al.*, 1996). However, individual contigs displayed a high degree of variation in their coding fractions, from 59.2% in contig Y11780 (9460 bp in total) to 97.3% in contig Y11779 (5157 bp in total). The low coding fraction of segment Y11780 was associated with intergenic regions of 1609 and 1720 bp that bracketed the *thrS* gene. Despite extensive searches, we have been unable to find any homologies between any coding sequences in the databases and these two intergenic sequences. Intergenic DNA of more than 1000 bp in length, lacking ORFs and having base composition patterns characteristic of non-coding DNA have also been observed in other parts of the genome (Andersson *et al.*, 1995). This suggests that the *R. prowazekii* genome may be a coding mosaic, with long stretches of non-coding DNA interwoven with regions of high coding density.

Coding information

In a previous comparison of a set of 21 homologous and functionally defined proteins from *R. prowazekii* and *E. coli*, we found that the sequence identity ranged from 25% for the most poorly conserved protein pairs to 75% for the most highly conserved protein pairs (Andersson & Sharp, 1996). Based on these results we have used a minimum of 25% amino acid sequence identity as a criterion for gene homology.

Contig Y11777: *R. prowazekii* *rbfA*, *sucC*, *sucD*, *com1*, *smpB*, *dapA* and *ilvE* genes and a gene encoding a transcriptional regulator. We associated 8 out of the 10 putative genes in

Table 3. Similarity of *R. prowazekii* ORF products to proteins in public databases

ORF	Gene	Product	Species	Accession no.*	Identity (%)	Reference†
RP002	<i>rbfA</i>	Ribosomal binding factor A	<i>Bacillus subtilis</i>	SP:P32731	30.2	1
RP004	<i>sucC</i>	Succinyl-CoA synthetase, β subunit	<i>Escherichia coli</i>	SP:P07460	51.8	2
RP005	<i>sucD</i>	Succinyl-CoA synthetase, α subunit	<i>Escherichia coli</i>	SP:P07459	69.5	2
RP006	<i>com1</i>	Outer-membrane protein	<i>Coxiella burnetii</i>	PIR:A49232	26.0	3
RP007	<i>smpB</i>	Small protein	<i>Escherichia coli</i>	SP:P32052	41.6	4
RP008	<i>dapA</i>	Dihydrodipicolinate synthetase	<i>Bacillus subtilis</i>	SP:Q04796	37.0	5
RP009	<i>ilvE</i>	Branched chain amino acid aminotransferase	<i>Escherichia coli</i>	SP:P00510	33.9	6
RP010	<i>petR</i>	Transcriptional regulatory protein	<i>Rhodobacter capsulatus</i>	SP:P31079	40.9	7
RP011	<i>nth</i>	Endonuclease III	<i>Haemophilus influenzae</i>	SP:P44319	50.9	8
RP012	<i>gltA</i>	Glutaredoxin like protein	<i>Legionella pneumophila</i>	GP:L46863	49.4	Unpublished
RP013		Esterase	<i>Spirulina platensis</i>	PIR:S43880	25.6	9
RP014	<i>glyA</i>	Serine hydroxymethyltransferase	<i>Methylobacterium extorquens</i>	SP:P50435	59.2	10
RP015	<i>acoS</i>	Lipoic acid synthetase	<i>Pelobacter carbinolicus</i>	PIR:E36953	47.6	11
RP016		Hypothetical protein H10004	<i>Haemophilus influenzae</i>	PIR:A64140	27.7	8
RP017	<i>tlyC</i>	Haemolysin	<i>Haemophilus influenzae</i>	PIR:G64060	31.1	8
RP018	<i>tlc</i>	ATP/ADP translocase	<i>Rickettsia prowazekii</i>	SP:P19568	38.7	12
RP019	<i>udg</i>	UDP-glucose dehydrogenase	<i>Bos taurus</i>	PIR:A54926	32.1	13
RP020	<i>dnaE</i>	DNA polymerase III, α subunit	<i>Salmonella typhimurium</i>	SP:P14567	42.6	14
RP023	<i>ksgA</i>	Dimethyladenosine transferase	<i>Escherichia coli</i>	SP:P06992	35.2	15
RP025	<i>ygiT</i>	Hypothetical protein	<i>Escherichia coli</i>	SP:P42601	33.6	Unpublished
RP029	<i>cydB</i>	Cytochrome oxidase <i>d</i> , subunit II	<i>Azotobacter vinelandii</i>	PIR:B38170	28.8	16
RP030		Hypothetical protein H10653	<i>Haemophilus influenzae</i>	SP:P44029	25.4	8
RP031	<i>ymxG</i>	Hypothetical processing peptidase	<i>Bacillus subtilis</i>	SP:Q04805	35.3	5
RP032	<i>purC</i>	SAICAR synthetase	<i>Streptococcus pneumoniae</i>	SP:Q07296	35.0	17
RP033	<i>thrS</i>	Threonyl-tRNA synthetase	<i>Haemophilus influenzae</i>	SP:P43014	49.8	8
RP035		Hypothetical protein	<i>Rickettsia tsutsugamushi</i>	GP:L11697	39.5	Unpublished
RP036	<i>murF</i>	UDP-MurNAc-pentapeptide synthetase	<i>Escherichia coli</i>	SP:P11880	30.7	18
RP037	<i>murE</i>	UDP-MurNAc-tripeptide synthetase	<i>Haemophilus influenzae</i>	SP:P45060	35.0	8
RP038	<i>mfd</i>	Transcription-repair coupling factor	<i>Haemophilus influenzae</i>	SP:P45128	32.7	8
RP039	<i>ftsY</i>	Cell division protein FtsY homologue	<i>Bacillus subtilis</i>	SP:P51835	38.5	19
RP040	<i>polA</i>	DNA polymerase I	<i>Bacillus caldotenax</i>	SP:Q04957	36.9	20
RP041	<i>metK</i>	S-Adenosylmethionine synthetase	<i>Escherichia coli</i>	SP:P04384	64.8	Unpublished

* GP, GENPEPT; SP, SWISS-PROT.

† 1, Shazand *et al.* (1993); 2, Buck *et al.* (1985); 3, Hendrix *et al.* (1993); 4, Miczak *et al.* (1991); 5, Chen *et al.* (1993); 6, Kuramitsu *et al.* (1985); 7, Tokito & Daldal (1992); 8, Fleischmann *et al.* (1995); 9, Salvi *et al.* (1994); 10, Chistoserdova & Lidstrom (1994); 11, Oppermann & Steinbüchel (1994); 12, Williamson *et al.* (1989); 13, Hempel *et al.* (1994); 14, Lancy *et al.* (1989); 15, van Buul & van Knippenberg (1985); 16, Moshiri *et al.* (1991); 17, Hui & Morrison (1993); 18, Parquet *et al.* (1989); 19, Oguro *et al.* (1995); 20, Uemori *et al.* (1993).

contig Y11777 with homologous sequences in the public databases (Table 3). For example, we observed an amino acid identity of 30.2% between RP002 and the *Bacillus subtilis* protein p15B, which is located immediately downstream of the *nusA-infB* operon (Shazand *et al.*, 1993). In *E. coli* the gene encoding p15B, *rbfA*, is located at the end of the *nusA-infB* operon (Shazand *et al.*, 1993) and the p15B protein has been shown to function as a ribosome-binding factor that is essential for maximal growth rates (Dammel & Noller, 1995). However, RP002 was only 22.1% identical to the gene product of the *E. coli rbfA* gene and was not located near genes homologous to *nusA* and *infB*.

We observed the highest levels of amino acid identities for RP005 (69.5%) and RP004 (51.8%) to the *E. coli* α and β subunits of succinyl-CoA synthetase, respectively. In *E. coli* the genes encoding these two proteins, *sucC* and *sucD*, are located within a cluster of genes encoding citric acid cycle enzymes in the order: *gltA-sdhCDAB-sucABCD* (Buck *et al.*, 1985). Four of these, *gltA* and *sdhCDA*, have been sequenced previously in *R. prowazekii* and *gltA* was found to be located distant from the *sdhCDA* gene cluster (Wood *et al.*, 1987; Aliabadi *et al.*, 1993). None of these were found in the vicinity of RP004 or RP005. The *sucAB* genes have not yet been encountered in *R. prowazekii*.

We conclude that these citric acid cycle component genes are split into at least three different clusters in the genome of *R. prowazekii*, one of which we have identified here. In the genomes of *E. coli*, *Haemophilus influenzae* and *Thermus flavus*, the termination codon of *sucC* overlaps the initiation codon of *sucD* (Buck *et al.*, 1985; Nishiyama *et al.*, 1991; Fleischmann *et al.*, 1995). In *R. prowazekii* there was a 162 bp spacer sequence between these two genes.

For RP006, we observed an amino acid identity of 26.0% with an outer-membrane protein of *Coxiella burnetii* encoded by the gene *com1* (Hendrix *et al.*, 1993). Two cysteine residues that have been shown to be conserved between the gene product of *com1* and disulfide oxidoreductases (Hendrix *et al.*, 1993) were also conserved in the RP006 protein sequence. RP007 showed 41.6% amino acid identity to the *E. coli* *smpB* gene product (Miczak *et al.*, 1991), and between 35 and 45% identity to the corresponding homologues in a variety of species such as *Enterococcus hirae* (Shibata *et al.*, 1992), *Neisseria meningitidis* (Jennings *et al.*, 1995), *H. influenzae* (Fleischmann *et al.*, 1995), *Synechocystis* sp. strain PCC6803 (Kaneko *et al.*, 1996), *Mycobacterium tuberculosis* (accession no. X70031) and *Mycoplasma genitalium* (Fraser *et al.*, 1995). Although the *smpB* gene has been shown to be expressed in *E. coli*, its function remains unknown (Miczak *et al.*, 1991). The presence of the *smpB* gene in microbial genomes of different phylogenetic affiliations suggests that it encodes an essential protein.

We have previously identified two of the genes involved in the biosynthesis of diaminopimelate (DAP), *asd* and *dapE* (Andersson *et al.*, 1996). In this study, we have identified the gene which encodes the first enzyme in the diaminopimelate biosynthetic pathway, *dapA* (RP008). We have previously shown data to suggest that there may be a low frequency of genes involved in amino acid biosynthetic pathways in the *R. prowazekii* genome (Andersson *et al.*, 1996). Here, we observed a 33.9% amino acid identity between RP009 and the *E. coli* *ilvE* gene product, a branched-chain amino acid aminotransferase. This gene is part of the operon *ilvGMEDA* in *E. coli* (Lawther *et al.*, 1979; Kuramitsu *et al.*, 1985). It remains to be determined whether the genes *ilvG*, *M*, *D* and *A* are located elsewhere or have been lost from the *R. prowazekii* genome.

We have putatively identified a gene encoding a transcription regulator (RP010). For this protein we observed the highest level of amino acid identity to a functionally undefined regulator protein from *Rhodobacter capsulatus* (40.9%). RP010 also showed homology to regulatory components of sensory transduction systems, such as the regulatory protein for alkaline phosphatase and phosphodiesterase in *B. subtilis* (Seki *et al.*, 1987). The molecular mechanism behind two-component signal transduction pathway systems involves transfer of phosphoryl groups from histidine residues in histidine protein kinases (HPK) to aspartic acid residues in their associated response regulators

(RR). There is a conserved region of about 100 aa in the response regulator family, with the residues corresponding to Asp-13, Asp-57 and Lys-109 in the CheY sequence being conserved among all the proteins in the family (Stock *et al.*, 1989). Each of these three key residues were found to be present in RP010 (data not shown), suggesting that it is a member of the response regulator family of proteins. The presence of such a gene in *R. prowazekii* is interesting, given the lack of genes for two-component signal transduction systems in the small genome of *M. genitalium* (Fraser *et al.*, 1995).

Contig Y11778: *R. prowazekii* *nth*, *grlA*, *glyA*, *lipA* and *tldD* genes and a gene encoding haemolysin. RP011 showed homology to a gene encoding endonuclease III in *H. influenzae*, with an overall amino acid identity of 50.9%. The alignment revealed segments that were strongly conserved in *H. influenzae*, *E. coli*, *B. subtilis* and *Micrococcus luteus* which lent additional support to this putative identification (data not shown). RP013 had some sequence similarity (25.6%) to an esterase of *Spirulina platensis*. In addition, the consensus motif, GX SXG, found at the catalytic site of many serine proteases, lipases and esterases (Salvi *et al.*, 1994) is GFSQG in both RP013 and in the *Spirulina platensis* esterase.

During a random sequencing survey of the *R. prowazekii* genome (Andersson *et al.*, 1996) we recovered a gene segment encoding an incomplete protein with sequence similarity to serine hydroxymethyltransferases (SHMT) from a variety of different species. In this study we determined the complete gene sequence (RP014). Our analysis confirmed its identity as *glyA*, with an overall amino acid identity of 59.2% to the *glyA* gene product of *Methylobacterium extorquens*. This gene is also the sole amino acid biosynthetic gene identified in the genome of *M. genitalium* (Fraser *et al.*, 1995). Since SHMT is indirectly involved in the synthesis of dTMP from dUMP, its primary function in *M. genitalium* and *R. prowazekii* may be related to nucleotide biosynthesis rather than to amino acid biosynthesis.

The *acoS* gene of *Pelobacter carbinolicus* is the homologue of the *E. coli* *lipA* gene, which encodes lipoic acid synthetase involved in the biosynthesis of lipoic acid, a sulfur-containing coenzyme. RP015 showed 47.6 and 43.5% identity to the products of *acoS* and *lipA*, respectively.

For RP017 we observed an amino acid identity of 26.6% to haemolysin of *Serpulina hyodysenteriae* and of 31.1% to the corresponding homologue in *H. influenzae*. The region corresponding to amino acid positions 125–205 was particularly well conserved, with 35 out of 83 residues identical to the *Serpulina hyodysenteriae* protein sequence. Crude suspensions of typhus-infected yolk sacs were found almost 50 years ago to cause the lysis of erythrocytes of several animal species (Clarke & Fox, 1948). Haemolysis is initiated by adsorption of the *Rickettsia* to the erythrocytes in a process that requires the generation of energy from the metabolism of glutamate (Ramm & Winkler, 1973a, b). The plasma

```

RP028 -----MIKKTKNIS-----YDEVVYPPTFSHTYPPVLTIG
RP034 MSLKSTSSLTNNHKTINSQSLVNGTGTADHNPYDEVYSPYAITNPHLSTLA
      * * * * *
RP028 KLFGLNPPPLETAKILDIGCGVGNLNFNAETPKSQSLGVDLSKTQIEIGKKTISDAKI
RP034 TLPGINAPEVENSKILELGAAGNLIPHAVLYPNAHFVGVLDLSKVQIDEANKVRLGL
      * * * * *
RP028 KNVRLKALSILDDESFGKFDYIVCHGVYSWVSKEVQDKILEVLNKLNPNGIAFISYNT
RP034 KNIEFHCSITDIDDSFGKFDYIICHGVISWVPKIVRDKIFKVCNRLNSTNGIAYISYNT
      * * * * *
RP028 LFGWMMQNTIREMMHSEFSNTSHDKLQSKLLKLPINDSLENSTTPYANFLREEAKLI
RP034 LFGWMMVRTIRDMLVHSSFTNIRDRIASRLLE-----
      * * * * *
RP028 STYADSVYLHEYLGEINTGTYPHQFIEKAQKKPKLLR
RP034 -----

```

Fig. 2. Comparison of the deduced amino acid sequences of RP028 and RP034 from *R. prowazekii*. Symbols beneath the aligned sequences indicate identical residues (*) and sites with conservative replacements (.).

membrane of the erythrocytes is subsequently lysed by phospholipase A, which is thought to stimulate internalization as well as release of the rickettsias from the phagosomes immediately after internalization in competent cells (Winkler & Miller, 1980, 1982, 1989). It has been shown that the haemolysin protein from *Serpulina hyodysenteriae*, when expressed in *E. coli*, exhibits both haemolytic and cytotoxic activity *in vitro* (ter Huurne *et al.*, 1994). The *Rickettsia typhi* homologue of the putative *R. prowazekii* haemolysin gene has been expressed in *E. coli* and analysis of the haemolytic activity is underway (S. Radulovic, J. M. Troyer, B. H. Noden, S. G. E. Andersson & A. Azad, unpublished data).

In metabolically damaged rickettsia, the addition of ATP can supply the energy required to restore haemolysis (Winkler, 1974). Transport of exogenously added ATP is mediated by an unusual ATP/ADP translocase, which is encoded by the *tlc* gene (Winkler, 1976; Winkler & Daugherty, 1984; Williamson *et al.*, 1989). The 3' terminal part of a gene with homology to the *tlc* gene in *R. prowazekii* was located downstream of the haemolysin gene, but it was oriented in the opposite direction. In total, we have identified four new putative *tlc* genes (Andersson *et al.*, 1996; R. Podowski & S. G. E. Anderson, unpublished data), all of which show high degrees of homology to each other and to the previously characterized *tlc* gene.

Contig Y11785: *R. prowazekii* *udg* and *dnaE* genes. For RP019 we observed sequence similarity to both UDP-glucose 6-dehydrogenase from cattle (*Bos taurus*) and GDP-mannose 6-dehydrogenase from bacteria (*Pseudomonas aeruginosa* and *Azotobacter vinelandii*). The amino acid sequence surrounding the catalytically active cysteine, GFFSC, is conserved in all four sequences. RP020 corresponded to *dnaE*, which encodes the α subunit of DNA polymerase III holoenzyme.

Contig Y11786: *R. prowazekii* *ksgA* gene and two ORFs. Contig Y11786 contained three ORFs, of which only one had significant homology to other proteins in the public databases. RP023 was 35.2% identical to a 16S rRNA methyltransferase encoded by *ksgA*. Inactivation of the *ksgA* gene in *E. coli* results in kasugamycin resistance (Blanchin-Roland *et al.*, 1986).

Contig Y11779: *R. prowazekii* *ygjT*-like gene and four ORFs. Five ORFs were detected in contig Y11779, but only one of these was similar to sequences in the public databases. For RP025 we observed weak similarity to a hypothetical protein from *E. coli* (33.6%) and to a tellurium resistance protein encoded by *terC* from *Alcaligenes* sp. (27.7%). We observed an amino acid identity of 46.0% between RP028 and a second ORF within our dataset, RP034 (Fig. 2). Neither of these two genes showed any similarity to sequences in the public databases, suggesting that they may be duplicated versions of an ancestral gene that was unique to the *Rickettsia*.

Contig Y11780: *R. prowazekii* *cydB*, *ymxG*, *purC* and *thrS* genes. We observed an amino acid identity of 28.8% between RP029 and the cytochrome *d* ubiquinol subunit II from *Azotobacter vinelandii*. The similarity was especially good at the termini of the proteins, with identities of 50.8% and 46.2% within the region from aa 51 to 109 and from 341 to 379 in *Azotobacter vinelandii*, respectively. However, the sequence homology in the middle section of the protein was less impressive with 20.7% identity for aa 110–340 as well as numerous gaps in the middle portion of the alignment. RP031 was homologous to a number of processing proteases from mitochondria as well as from bacteria. Several conserved segments were observed within the proteins from these distantly related genomes, indicating similar gene functions. RP032 was putatively identified as *purC*, which encodes phosphoribosylaminoimidazole-succinocarboxamide (SAICAR) synthetase involved in nucleotide biosynthesis. There was 49.8% amino acid sequence identity between RP033 and threonyl-tRNA synthetase of *H. influenzae*. Furthermore, an alignment revealed very well conserved blocks between threonyl-tRNA synthetase from several prokaryotic and eukaryotic species (data not shown). Accordingly, we suggest that RP033 encodes threonyl-tRNA synthetase.

Contigs Y11781 and Y11782: *R. prowazekii* RP028-like gene. Contig Y11781, which was the shortest sequence fragment within our dataset, contained only one complete ORF (RP034). As previously mentioned, we observed a strong homology to another protein within our dataset, RP028 (Fig. 2). Contig Y11782 partially overlapped with contig U02603, which was previously deposited in the databases. Within this contig we observed one ORF, RP035, which had an amino acid identity of 39.5% to a putative serine protease from *Rickettsia* (*Orientia*) *tsutsugamushi*.

Contig Y11783: *R. prowazekii* *murF*, *murE* and *mfd* genes located upstream of *dnaA*. Contig Y11783 overlapped with the previously sequenced contig U55213, which contained the gene *dnaA*. A cluster of genes with the order *rnpA*–*rpmH*–*dnaA*–*dnaN*–*recF*–*gyrB* is found in the genomes of *E. coli* and *B. subtilis* and is thought to represent the origin of replication in *B. subtilis* (Ogasawara *et al.*, 1985). In *H. influenzae* the same relative order of genes is observed, although four additional genes are located between *rnpA*–*rpmH* and *dnaA* (Fleischmann *et al.*, 1995). In *M. genitalium* and

M. pneumoniae the genes *rnpA*, *rpmH*, *dnaN* and *gyrB* are located near to *dnaA* but in the reverse orientation relative to the order seen in the other bacterial genomes (Fraser *et al.*, 1995; Hilbert *et al.*, 1996). In *R. prowazekii* the genes *rnpA* and *rpmH* were not found upstream of *dnaA*, nor did we find *dnaN* or *gyrB* in the vicinity of *dnaA*. The DnaA protein-binding sequence is repeated four times at the origin of replication in *E. coli* and is present in nine and four copies upstream and downstream of the *dnaA* gene of *B. subtilis*, respectively (Ogasawara *et al.*, 1985). No similarities to the consensus DnaA-box could be detected upstream of *dnaA* in *R. prowazekii*.

Immediately upstream but in the opposite orientation to *dnaA*, we found a gene with the highest level of sequence similarity to the *mfd* gene product of *H. influenzae*. The 5' terminal region of *mfd* is contained within the previously sequenced contig U55213, although it was not annotated as an ORF in the sequence database. Further upstream of *dnaA* we found the genes RP036 and RP037 which showed sequence similarity to the *murF* and *murE* gene products of *E. coli* (Table 3). In *E. coli*, *murE* and *murF* are located in a large gene cluster, *murE-murF-mraY-murD-ftsW-murG-murC-ddlB-ftsQ-ftsA-ftsZ-envA*, that encodes proteins involved in cell wall biosynthesis and cell division (Yura *et al.*, 1992). An identical stretch of genes is seen in the genome of *H. influenzae* (Fleischmann *et al.*, 1995). A similar arrangement is also found in *B. subtilis* (*murE-mraY-murD-spoVE-murG*), although in this species the *murE* gene is not linked to the *murF* gene (Daniel & Errington, 1993). In *Synechocystis* sp. the linkage of *murE* and *murF* also appears to be disrupted (Kaneko *et al.*, 1995; Malakhov *et al.*, 1995). In *R. prowazekii* the *murE* gene was linked to *murF* gene with an intervening spacer region of 106 nt. In *E. coli*, *murE* overlaps *murF* by 4 nt, while the two genes are separated by a spacer of only 17 nt in *H. influenzae*. In no bacterial genomes are the gene clusters containing the *murE* and *murF* genes located near to *dnaA* or the origin of replication. We conclude that *R. prowazekii* has a unique arrangement of genes upstream of *dnaA*. It remains to be determined whether the region surrounding the *dnaA* gene in *R. prowazekii* corresponds to the functional origin of replication.

Contig Y11784: *R. prowazekii* *ftsY*, *polA* and *metK* genes. Within contig Y11784 we found two ORFs (RP039 and RP040) with sequence similarities to the cell division protein FtsY (Oguro *et al.*, 1995) and DNA polymerase I of *Bacillus caldopenax* (Uemori *et al.*, 1993), respectively. The last ORF within this contig had all the characteristics of a gene that was predicted to be actively transcribed and translated as it had a high G + C content (36%), a low GC3_s value (18.7%) and a low AT/GC ratio (1.5). The relatively small influence of biased mutation pressure on its amino acid composition suggested that this protein corresponded to a gene that was highly conserved among bacteria. Indeed, BLAST searches identified this gene as a homologue of the *E. coli* gene *metK* which encodes S-adenosylmethionine

Fig. 3. Restriction enzyme analysis of the region of the *R. prowazekii* chromosome containing the *metK* gene. Hybridization signals generated with the *metK* probe are shown. Lanes: a-c and d-f, single digests of 0.5 µg (a-c) and 0.02 µg (d-f) of clone A500 with the enzymes *DraI* (a, d), *HpaII* (b, e) and *SspI* (c, f); g-i, single digests of 2.0 µg chromosomal DNA with the enzymes *DraI* (g), *HpaII* (h) and *SspI* (i). Molecular markers were based on double digests of λ DNA with the enzymes *HindIII* and *EcoRI*.

synthetase. The two homologues had an amino acid identity of 64.8%, which is the second highest within this dataset.

Internal termination codon in the *metK* gene

Surprisingly, sequence analysis of the *metK* gene from two plasmid clones of different sizes and isolated independently of each other (A500 and A217) revealed the presence of an internal termination codon (UAG) at nt 421–423. To exclude the possibility that this termination codon was the result of a cloning artefact, we amplified the region surrounding the UAG codon directly from genomic DNA by PCR. Similar size PCR products were obtained when two primers from within the *metK* gene (nt 313–965) were used in combination with plasmid DNA from clone A500 and with genomic DNA from *R. prowazekii* (data not shown). Sequence analysis of the PCR product obtained from genomic DNA confirmed the presence of a UAG termination codon within the *metK* gene of *R. prowazekii*.

To determine whether the potentially defective *metK* gene located in contig Y11784 was functionally replaced by a duplicate gene located elsewhere in the *R. prowazekii* genome, we performed a Southern blot analysis of *R. prowazekii* genomic DNA digested with the enzymes *DraI*, *HpaII* and *SspI* (Fig. 3). To serve as positive control, we also included plasmid DNA from clone A500 digested with the same set of enzymes. In the plasmid digests, the *metK* probe hybridized to single fragments corresponding to 0.8, 1.2 and 1.9 kb in the

*Hpa*II, *Dra*I and *Ssp*I digests, respectively. These sizes are in accordance with the fragment sizes expected from the nucleotide sequence data (870, 1224 and 1949 bp, respectively). In the genomic digests, we observed hybridization to single fragments of identical sizes to those obtained in the plasmid digests. We conclude that the *metK* gene is present as a single copy gene located within the fragment corresponding to Y11784 in the *R. prowazekii* genome.

Alignment of the *metK* gene product from *R. prowazekii* with those from six other bacterial and three other nuclear genomes is shown in Fig. 4. Here, it can be seen that the termination codon was located within a region that was highly conserved among these sequences. For example, the amino acid which was positioned second to the left of the termination codon was isoleucine in seven out of nine protein sequences. Similarly, the alanine next to the termination codon was conserved in eight of these proteins. Further up- and downstream, blocks of conserved sequences were easily recognized (Fig. 4). The high degree of amino acid conservation adjacent to the termination codon makes it unlikely that two or more frameshift events would account for translational readthrough. Similarly, the high degree of amino acid conservation downstream of the stop codon suggests that it is unlikely that a truncated, functional protein product is being made from the *R. prowazekii metK* homologue. Furthermore, we found no sequence resembling to a Shine–Dalgarno sequence upstream of the termination codon, nor any indications of strong secondary structures typical of transcriptional terminators downstream of the UAG stop codon. Thus, the termination codon is either by-passed by an unusual mechanism of translational readthrough, or it has been generated in a recent evolutionary event which has started this gene on a path of devolution. A preliminary survey of other strains of *R. prowazekii* and other rickettsial species suggests that the *metK* gene is exposed to mutational meltdown in different species independently of each other (J. O. Andersson, unpublished data). Further studies will help clarify the importance of the termination codon in the *metK* gene of *R. prowazekii*.

Conclusions

The analysis presented in this study has confirmed that codon-position-dependent nucleotide frequency statistics provide a useful quantitative measure to distinguish coding from non-coding DNA. The finding of long stretches of non-coding DNA was consistent with previous estimates of an overall low coding content in the *R. prowazekii* genome (Andersson & Sharp, 1996; Andersson *et al.*, 1996). The low overall G + C content of the non-coding regions suggests that they are not functionally constrained and that they are therefore

unlikely to encode RNA genes. RNA-encoding genes are highly constrained and characteristically have G + C values in the range of 45–60% (Muto & Osawa, 1987). Several examples of disrupted gene order structures were observed in regions that are otherwise highly conserved among microbial genomes, suggesting that the *R. prowazekii* genome has a highly rearranged gene organization pattern (Andersson & Kurland, 1995; Syvänen *et al.*, 1996). The finding of an internal stop codon in the *metK* gene raises the intriguing possibility that mutational meltdown is an ongoing process in the *R. prowazekii* genome. In summary, the results presented provide additional support for the hypothesis that the genome of *R. prowazekii* has evolved from bacteria with larger size genomes that have been reduced in size and altered in architecture by mutation, recombination and deletion events.

ACKNOWLEDGEMENTS

We thank Charles Kurland and Herbert Winkler for discussions and helpful comments on the manuscript. This work was financed by grants from the Natural Sciences Research Council, the Swedish Cancer Society, the Knut and Alice Wallenberg Foundation and the Foundation for Strategic Research to Siv Andersson and Charles Kurland.

REFERENCES

- Aliabadi, Z., Winkler, H. H. & Wood, D. O. (1993). Isolation and characterization of the *Rickettsia prowazekii* gene encoding the flavoprotein subunit of succinate dehydrogenase. *Gene* **133**, 135–140.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol* **215**, 403–410.
- Andersson, S. G. E. & Kurland, C. G. (1995). Genomic evolution drives the evolution of the translation system. *Biochem Cell Biol* **73**, 775–787.
- Andersson, S. G. E. & Sharp, P. M. (1996). Codon usage and base composition in *Rickettsia prowazekii*. *J Mol Evol* **42**, 525–536.
- Andersson, S. G. E., Zomorodipour, A., Winkler, H. H. & Kurland, C. G. (1995). Unusual organization of the rRNA genes in *Rickettsia prowazekii*. *J Bacteriol* **177**, 4171–4175.
- Andersson, S. G. E., Eriksson, A.-S., Näslund, A. K., Andersen, M. S. & Kurland, C. G. (1996). The *Rickettsia prowazekii* genome: a random sequence analysis. *Microb Comp Genomics* **4**, 293–315.
- Blanchin-Roland, S., Blanquart, S., Schmitter, J.-M. & Fayat, G. (1986). The gene for *Escherichia coli* diadenosine tetraphosphate is located immediately clockwise to *folA* and forms an operon with *ksgA*. *Mol Gen Genet* **205**, 515–522.
- Buck, D., Spencer, M. E. & Guest, J. R. (1985). Primary structure of the succinyl-CoA synthetase of *Escherichia coli*. *Biochemistry* **24**, 6245–6252.
- van Buul, C. P. J. J. & van Knippenberg, P. H. (1985). Nucleotide sequence of the *ksgA* gene of *Escherichia coli*: comparison of methyltransferases effecting dimethylation of adenosine in ribosomal RNA. *Gene* **38**, 65–72.

Fig. 4. Alignment of *metK* protein sequences from *R. prowazekii*, *E. coli* (P04384), *H. influenzae* (P43762), *B. subtilis* (U52812), *Staphylococcus aureus* (U36379), *M. genitalium* (P47293), *Pisum sativum* (P49613), *Homo sapiens* (P31153) and *Saccharomyces cerevisiae* (P19358). Identical residues in the alignment are indicated by boxes. The region surrounding the internal termination codon in the *R. prowazekii metK* gene is indicated by bold letters.

- Chen, N.-Y., Jiang, S.-Q., Klein, D. A. & Paulus, H. (1993). Organization and nucleotide sequence of the *Bacillus subtilis* diaminopimelate operon, a cluster of genes encoding the first three enzymes of diaminopimelate synthesis and dipicolinate synthase. *J Biol Chem* 268, 9448–9465.
- Chistoserdova, L. V. & Lidstrom, M. E. (1994). Genetics of the serine cycle in *Methylobacterium extorquens* AM1: cloning, sequence, mutation, and physiological effect of *glyA*, the gene for serine hydroxymethyltransferase. *J Bacteriol* 176, 6759–6763.
- Clarke, D. H. & Fox, J. P. (1948). The phenomenon of *in vitro* hemolysis produced by the rickettsiae of typhus fever, with a note on the mechanism of rickettsial toxicity in mice. *J Exp Med* 88, 24–41.
- Dammel, C. S. & Noller, H. F. (1995). Suppression of a cold-sensitive mutation in 16S rRNA by overexpression of a novel ribosome-binding factor, RbfA. *Genes Dev* 9, 626–637.
- Daniel, R. A. & Errington, J. (1993). DNA sequence of the *murE*–*murD* region of *Bacillus subtilis* 168. *J Gen Microbiol* 139, 361–370.
- Eremeeva, M. E., Roux, V. & Raoult, D. (1993). Determination of genome size and restriction pattern polymorphism of *Rickettsia prowazekii* and *Rickettsia typhi* by pulsed field gel electrophoresis. *FEMS Microbiol Lett* 11, 105–112.
- Fleischmann, R. D. and others (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269, 496–512.
- Fraser, C. M. and others (1995). The minimal gene complement of *Mycoplasma genitalium*. *Science* 270, 397–403.
- Gray, M. W. (1995). Mitochondrial evolution. In *Molecular Biology of Plant Mitochondria*, pp. 635–659. Edited by C. S. Levings III & I. K. Vasil. Dordrecht: Kluwer Academic Publishers.
- Gupta, R. S. (1995). Evolution of the chaperonin families (Hsp60, Hsp10 and Tcp-1) of proteins and the origin of the eukaryotic cells. *Mol Microbiol* 15, 1–11.
- Hempel, J., Perozich, J., Romovavek, H., Hinich, A., Kuo, I. & Feingold, D. S. (1994). UDP-glucose dehydrogenase from bovine liver: primary structure and relationship to other dehydrogenases. *Protein Sci* 3, 1074–1080.
- Hendrix, L. R., Mallavia, L. P. & Samuel, J. E. (1993). Cloning and sequencing of *Coxiella burnetii* outer membrane protein gene *com1*. *Infect Immun* 61, 470–477.
- Higgins, D. G., Bleasby, A. J. & Fuchs, R. (1992). CLUSTAL V: improved software for multiple sequence alignment. *CABIOS* 8, 189–191.
- Hilbert, H., Himmelreich, R., Plagens, H. & Herrmann, R. (1996). Sequence analysis of 56 kb from the genome of the bacterium *Mycoplasma pneumoniae* comprising the *dnaA* region, the *atp* operon and a cluster of ribosomal protein genes. *Nucleic Acid Res* 24, 628–639.
- Hui, F. M. & Morrison, D. A. (1993). Identification of a *purC* gene from *Streptococcus pneumoniae*. *J Bacteriol* 175, 6364–6367.
- ter Huurne, A. A. H. M., Muir, S., van Houten, M., van der Zeijst, B. A. M., Gastra, W. & Kusters, J. G. (1994). Characterization of three putative *Serpulina hyodysenteriae* hemolysins. *Microb Pathog* 16, 269–282.
- Jennings, M. P., Bisercic, M., Dunn, K. L. R., Virji, M., Martin, A., Wilks, K. E., Richards, J. C. & Moxon, E. R. (1995). Cloning and molecular analysis of the *lsi1* (*rfaF*) gene of *Neisseria meningitidis* which encodes a heptosyl-2-transferase involved in LPS biosynthesis: evaluation of surface exposed carbohydrates in LPS mediated toxicity for human endothelial cells. *Microb Pathog* 19, 391–407.
- Kaneko, T., Tanaka, A., Sato, S., Kotani, H., Sazuka, T., Miyajima, N., Sugiura, M. & Tabata, S. (1995). Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. I. Sequence features in the 1 Mb region from map position 64% to 92% of the genome. *DNA Res* 2, 153–166.
- Kaneko, T. and others (1996). Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res* 3, 109–136.
- Kündig, C., Hennecke, H. & Göttfert, M. (1993). Correlated physical and genetic map of the *Bradyrhizobium japonicum* 110 genome. *J Bacteriol* 175, 613–622.
- Kuramitsu, S., Ogawa, T., Ogawa, H. & Kagamiyama, H. (1985). Branched-chain amino acid aminotransferase of *Escherichia coli*: nucleotide sequence of the *ilvE* gene and the deduced amino acid sequence. *J Biochem* 97, 993–999.
- Lancy, E. D., Lifshits, M. R., Munson, P. & Maurer, R. (1989). Nucleotide sequence of *dnaE*, the gene for the polymerase subunit of DNA polymerase III in *Salmonella typhimurium*, and a variant that facilitates growth in the absence of another polymerase subunit. *J Bacteriol* 171, 5581–5586.
- Lawther, R. P., Nichols, B., Zurawski, X. & Haffield, G. W. (1979). The nucleotide sequence preceding and including the beginning of the *ilvE* gene of the *ilvGEDA* operon of *Escherichia coli* K12. *Nucleic Acids Res* 7, 2289–2301.
- Lloyd, A. T. & Sharp, P. M. (1992). CODONS: a microcomputer program for codon usage analysis. *J Hered* 83, 239–240.
- Malakhov, M. P., Los, D. A., Wada, H., Semenenko, V. E. & Murata, N. (1995). Characterization of the *murF* gene of the cyanobacterium *Synechocystis* sp. PCC 6803. *Microbiology* 141, 163–169.
- Miczak, A., Chauhan, A. K. & Apirion, D. (1991). Two new genes located between 2758 and 2761 kilobase pairs on the *Escherichia coli* genome. *J Bacteriol* 173, 3271–3272.
- Moshiri, F., Chawla, A. & Maier, R. (1991). Cloning, characterisation, and expression in *Escherichia coli* of the genes encoding the cytochrome *d* oxidase complex from *Azotobacter vinelandii*. *J Bacteriol* 173, 6230–6241.
- Muto, A. & Osawa, S. (1987). The guanine and cytosine content of genomic DNA and bacterial evolution. *Proc Natl Acad Sci USA* 84, 166–169.
- Myers, E. & Miller, W. (1988). Optimal alignments in linear space. *CABIOS* 4, 11–17.
- Nishiyama, M., Horinouchi, S. & Beppu, T. (1991). Characterization of an operon encoding succinyl-CoA synthetase and malate dehydrogenase from *Thermus flavus* AT-62 and its expression in *Escherichia coli*. *Mol Gen Genet* 226, 1–9.
- Ogasawara, N., Moriya, S., von Meyerbur, K., Hansen, F. G. & Yoshikawa, H. (1985). Conservation of genes and their organization in the chromosomal replication origins of *Bacillus subtilis* and *Escherichia coli*. *EMBO J* 4, 3345–3350.
- Oguro, A., Kakeshita, H., Honda, K., Takamatsu, H., Nakamura, K. & Yamane, K. (1995). *srb*: a *Bacillus subtilis* gene encoding a homologue of the alpha-subunit of the mammalian signal recognition particle receptor. *DNA Res* 2, 95–100.
- Olsen, G. J., Woese, C. R. & Overbeek, R. (1994). The winds of (evolutionary) change: breathing new life into microbiology. *J Bacteriol* 176, 1–6.

- Oppermann, F. B. & Steinbüchel, A. (1994). Identification and molecular characterization of the *aco* genes encoding the *Pelobacter carbinolicus* acetoin dehydrogenase enzyme system. *J Bacteriol* 176, 469–485.
- Pang, H. & Winkler, H. H. (1993). Copy number of the 16S rRNA gene in *Rickettsia prowazekii*. *J Bacteriol* 175, 3893–3896.
- Parquet, C., Flouret, B., Mengin-Lecreulx, D. & van Heijenoort, J. (1989). Nucleotide sequence of the *murF* gene encoding the UDP-MurNAc-pentapeptide synthetase of *Escherichia coli*. *Nucleic Acids Res* 17, 5379.
- Ramm, L. E. & Winkler, H. H. (1973a). Rickettsial hemolysis: adsorption of rickettsiae to erythrocytes. *Infect Immun* 7, 93–99.
- Ramm, L. E. & Winkler, H. H. (1973b). Rickettsial hemolysis: effect of metabolic inhibitors upon hemolysis and adsorption. *Infect Immun* 7, 550–555.
- Roux, V. & Raoult, D. (1993). Genotypic identification and phylogenetic analysis of the spotted fever group rickettsiae by pulsed-field gel electrophoresis. *J Bacteriol* 175, 4895–4909.
- Salvi, S., Trinei, M., Lanfalconi, L. & Pon, C. L. (1994). Cloning and characterization of the gene encoding an esterase from *Spirulina platensis*. *Mol Gen Genet* 243, 124–126.
- Schramek, S. (1972). Deoxyribonucleic acid base composition of members of the typhus group of rickettsiae. *Acta Virol* 16, 447.
- Seki, T., Yoshikawa, H., Takahashi, H. & Saitou, H. (1987). Cloning and nucleotide sequence of *phoP*, the regulatory gene for alkaline phosphatase and phosphodiesterase in *Bacillus subtilis*. *J Bacteriol* 169, 2913–2916.
- Shazand, K., Tucker, J., Grunberg-Manago, M., Rabinowitz, J. C. & Leighton, T. (1993). Similar organization of the *nusA-infB* operon in *Bacillus subtilis* and *Escherichia coli*. *J Bacteriol* 175, 2880–2887.
- Shibata, C., Ehara, T., Tomura, K., Igarashi, K. & Kobayashi, H. (1992). Gene structure of the *Enterococcus hirae* (*Streptococcus faecalis*) F_1F_0 -ATPase, which functions as a regulator of cytoplasmic pH. *J Bacteriol* 174, 6117–6124.
- Stock, J. B., Ninfa, A. J. & Stock, A. M. (1989). Protein phosphorylation and regulation of adaptive responses in bacteria. *Microbiol Rev* 52, 450–490.
- Syvänen, A.-C., Amiri, H., Jamal, A., Andersson, S. G. E. & Kurland, C. G. (1996). A chimeric disposition of elongation factor genes in *Rickettsia prowazekii*. *J Bacteriol* 178, 6192–6199.
- Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22, 4673–4680.
- Tokito, M. K. & Daldal, F. (1992). *petR*, located upstream of the *fbcFBC* operon encoding the cytochrome *bc₁* complex, is homologous to bacterial response regulators and necessary for photosynthetic and respiratory growth of *Rhodobacter capsulatus*. *Mol Microbiol* 6, 1645–1654.
- Tyeryar, F. J., Weiss, E., Millar, D. B., Bozeman, F. M. & Ormsbee, R. A. (1973). DNA base composition of rickettsiae. *Science* 180, 415–417.
- Uemori, T., Ishino, Y., Fujita, K., Asada, K. & Kato, I. (1993). Cloning of the DNA polymerase gene of *Bacillus caldopenax* and characterization of the gene product. *J Biochem* 113, 401–410.
- Viale, A. & Arakaki, A. K. (1994). The chaperone connection to the origins of the eukaryotic organelles. *FEBS Lett* 341, 146–151.
- Williamson, L. R., Plano, G. V., Winkler, H. H., Krause, D. C. & Wood, D. O. (1989). Nucleotide sequence of the *Rickettsia prowazekii* ATP/ADP translocase gene. *Gene* 80, 269–278.
- Winkler, H. H. (1974). Inhibitory and restorative effects of adenine nucleotides on rickettsial adsorption and hemolysis. *Infect Immun* 9, 119–126.
- Winkler, H. H. (1976). Rickettsial permeability: an ATP/ADP transport system. *J Biol Chem* 251, 389–396.
- Winkler, H. H. & Daugherty, R. M. (1984). Regulatory role of phosphate and other anions in transport of ADP and ATP by *Rickettsia prowazekii*. *J Bacteriol* 160, 76–79.
- Winkler, H. H. & Miller, E. T. (1980). Phospholipase A activity in the hemolysis of sheep and human erythrocytes by *Rickettsia prowazekii*. *Infect Immun* 29, 316–321.
- Winkler, H. H. & Miller, E. T. (1982). Phospholipase A and the interaction of *Rickettsia prowazekii* and mouse fibroblasts (L-292 cells). *Infect Immun* 38, 109–113.
- Winkler, H. H. & Miller, E. T. (1989). Phospholipase A activity associated with the growth of *Rickettsia prowazekii* in L929 cells. *Infect Immun* 57, 36–40.
- Wood, D. O., Williamson, L. R., Winkler, H. H. & Krause, D. C. (1987). Nucleotide sequence of the *Rickettsia prowazekii* citrate synthase gene. *J Bacteriol* 169, 3564–3572.
- Yura, T., Mori, H., Nagai, H., Nagata, T., Ishihama, A., Fujita, N., Isono, K., Mizobuchi, K. & Nakata, A. (1992). Systematic sequencing of the *Escherichia coli* genome: analysis of the 0–2.4 min region. *Nucleic Acids Res* 20, 3305–3308.

Received 25 November 1996; revised 10 April 1997; accepted 6 May 1997.