Impact of the Excision of an Ancient Repeat Insertion on *Rickettsia conorii* Guanylate Kinase Activity

Chantal Abergel, Guillaume Blanc,* Vincent Monchois,*[1] Patricia Renesto,†*
Cécile Sigoillot, Hiroyuki Ogata,* Didier Raoult,† and Jean-Michel Claverie**

*Information Génomique & Structurale, CNRS UPR 2589, IBSM, Marseille cedex, France; and †Unité des Rickettsies, Faculté de Médecine, Université de la Méditerranée, CNRS UMR 6020, IFR 48, Marseille cedex, France

The genomic sequencing of *Rickettsia conorii* revealed a new family of *Rickettsia*-specific palindromic elements (RPEs) capable of in-frame insertion in preexisting open reading frames (ORFs). Many of these altered ORFs correspond to proteins with well-characterized or essential functions in other microorganisms. Previous experiments indicated that RPE-containing genes are normally transcribed and that no excision of the repeat occurs at the mRNA level. Using mass spectrometry, we now confirmed the retention of the RPE-derived amino acid residues in 4 proteins successfully expressed in *Escherichia coli*, raising the general question of the consequences of this common insertion event on the fitness of *Rickettsia* enzymes. The predicted guanylate kinase activity of the *R. conorii gmk* gene product was measured both on the RPE-containing and RPE-excised recombinant proteins. We show that the 2 proteins are active but exhibit substantial differences in their affinity for adenosine triphosphate, guanosine monophosphate, and catalytic constants. The distribution of the RPE_{gmk} insert among *Rickettsia* species indicates that the insertion event is ancient and occurred after the divergence of *Rickettsia felis* and *R. conorii* but before that of *Rickettsia helvetica* and *R. conorii*. We found no evidence that the *gmk* gene fixed adaptive changes to compensate the RPE peptide insertion. Furthermore, the analysis of the rates of divergence in 23 RPE-containing genes indicates that coding RPE repeats tend to evolve under weak selective constraint, at a rate similar to intergenic noncoding RPE sequences. Altogether, these results suggest that the insertion of RPE-encoded "selfish peptides," although respecting the original fold and activity of the host proteins, might be slightly detrimental to the enzyme efficiency within limits tolerable for slow-growing intracellular parasites such as *Rickettsia*.

Introduction

Rickettsia are obligate intracellular small gram-negative proteobacteria of the α-subdivision associated with different arthropod hosts. Many *Rickettsia* species infect human beings and are responsible for mild to severe diseases. When transmitted to human through tick bites, *Rickettsia conorii* causes Mediterranean spotted fever (Raoult and Roux 1997). The comparison of the *R. conorii* genome sequence (Ogata et al. 2001) with its close relative *Rickettsia prowazekii* (Andersson et al. 1998) revealed a new type of "coding" mobile element (*Rickettsia*-specific palindromic element, RPE) frequently found inserted in frame within open reading frames (ORFs) (Ogata et al. 2000), whereas all previously described bacterial palindromic repeats appeared exclusively located within noncoding regions. RPEs have been identified in all *Rickettsia* species sequenced so far (Ogata, Renesto, et al. 2005; Ogata et al. 2006) and in 2 *Wolbachia* species (Ogata, Suhre, et al. 2005), making them apparently specific to the clade *Rickettsiaceae* of α-proteobacteria. Besides their unusual ability to parasite protein–coding regions, the 150-bp-long RPE resembles usual bacterial mobile elements (such as RSA and intergenic repeat unit) in respect of frequency, size, and palindromic structure (Claverie and Ogata 2003). In line with this view, a study has shown that transposable elements have contributed to thousands of human proteins (Britten 2006). It has been further hypothesized that RPEs could have fostered the association between the parasitic α-proteobacteria and their host (Ogata, Suhre, et al. 2005). Others have argued that the insertions of RPEs in genes were neutral or might have been tolerated in *Rickettsia* even if it had a negative effect on protein function because of recurrent bottlenecks and/or small population sizes (Amiri et al. 2002). Reverse transcriptase–polymerase chain reaction (PCR) experiments have previously demonstrated the retention of the RPE sequences within the transcript of at least 3 genes (*gltX*, *hemC*, and *ubiG*) (Ogata et al. 2000). However, there are yet no experimental evidences about the existence and activity of RPE-containing gene products due to the difficulty in expressing the low (30%) G + C content *Rickettsia* genes. For instance, there was still a possibility that RPE-derived segments could be spliced out from precursor proteins as described for inteins (Paulus 2000) or that the RPE-containing proteins would not be correctly folded and functional. As additional cases of unrelated palindromic coding elements have recently been identified in *Methanocaldococcus jannaschii* (Suyama et al. 2005), investigating the possible evolutionary consequence of the insertion of coding mobile elements in preexisting genes (Claverie and Ogata 2003) has become of more general interest.

We report here the first direct evidence that several RPE nucleotide sequences are translated into peptides that are retained in a folded and/or functional protein product. The analysis of *gmk* PCR amplicon sequences generated from a panel of *Rickettsia* at various evolutionary distances allowed us to determine that the insertion of RPE in the *gmk* gene is ancient. The predicted guanylate kinase activity of the RPE-containing Gmk protein, 47% identical to the *Escherichia coli* K-12 guanylate kinase outside the RPE insert, was experimentally measured (Gentry et al. 1993). The RPE element was then excised from the *R. conorii* gene that was in turn successfully expressed and resulted into a correctly folded functional protein. The enzymatic activity of the RPE-less *gmk* gene product could then be

[1] Present address: Protein'eXpert SA, 15 rue des martyrs, 38000 Grenoble, France.

Key words: *Rickettsia conorii*, palindromic coding repeat, guanylate kinase, enzyme fitness, recombinant protein expression.

E-mail: chantal.abergel@igs.cnrs-mrs.fr.

Mol. Biol. Evol. 23(11):2112–2122. 2006
doi:10.1093/molbev/msl082
Advance Access publication August 4, 2006

© 2006 The Authors
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

compared with the extant *R. conorii* Gmk thus revealing substantial differences in affinity for adenosine triphosphate (ATP) and guanosine monophosphate (GMP) to the disadvantage of the RPE-containing enzyme. Interestingly, comparison of divergence rates of the Gmk sequences with and without the RPE insert suggests that the insertion did not trigger adaptive changes in the rest of the Gmk protein. In addition we show that coding RPE repeats tend to evolve under weaker selective constraints than their recipient genes, at a rate similar to that measured for intergenic noncoding RPE sequences.

Our results support the previously proposed hypothesis (Ogata et al. 2002; Claverie and Ogata 2003) that the insertion of the RPE-encoded domain does not perturbate significantly the normal folding of the original protein while altering substantially the enzyme's activity. Following the notion of "selfish DNA," this validates RPEs as the first example of "selfish peptides" at the protein level.

Materials and Methods

Plasmid Construction

Four full-length *R. conorii* ORFs containing RPE and one *R. conorii* (*gmk*) ORF without the coding RPE nucleotide sequence were cloned into pDEST17 (Invitrogen, Cergy Pontoise, France) using the GATEWAY system (Walhout et al. 2000). Briefly, direct and reverse PCR primers were designed to amplify the target ORFs from *R. conorii* genomic DNA using HIFI *Taq* polymerase (Invitrogen). A 25-bp attB1 sequence was fused at the 5′ end of the direct primer, changing the start codon to leucine. For the RPE-excised *gmk* gene, the attB1 sequence was fused at the 5′ end corresponding to the first residue after the RPE nucleotide sequence. At the 5′ end of the reverse primer, a 25-bp attB2 sequence was fused to the stop codon of the ORF. The PCR products were purified by polyethylene glycol precipitation, inserted by homologous recombination into the pDONR201 donor vector, then transferred into pDEST17 (Invitrogen's 1-step cloning protocol), allowing the expression of each ORF in phase with a N-terminal His_6 tag under the control of a T7 promoter. After transformation into *E. coli* DH5α cells, the different purified plasmids were used for the overexpression of the fusion proteins both in a cell-free system and in *E. coli* Rosetta (DE3) plysS cells from Novagen, Madison, Wis. (Novy et al. 2001) to overcome the high A + T content (67.6%) of the *R. conorii* genome sequence (Ogata et al. 2001).

Protein Production and Purification

Proteins were expressed in *E. coli* Rosetta (DE3) plysS (Novagen, Madison, WI) carrying the relevant plasmids in 4 ml 2× YT + 2% (w/v) glucose supplemented with ampicillin (100 μg/ml) and chloramphenicol (34 μg/ml) at 30°C. Protein expression was induced when the optical density at 600 nm reached 0.4–0.6 by adding 0.5 mM isopropyl-L-thio-β-galactoside. After 3 h induction, cells were harvested and frozen at −80°C before subsequent treatments. Cells from 200-ml cultures were resuspended respectively in 10 ml buffer A (sodium phosphate 50 mM, pH 8.0, 300 mM NaCl) containing 0.1% (v/v) Triton X-100 and 5% (v/v) glycerol before sonication. In-parallel purification of proteins was performed by nickel affinity using 5 ml Ni-nitrilotriacetate resin (Qiagen, Hilden, Germany) on the AKTA Explorer System 10S (GE Healthcare Life Sciences) in 45 min. Briefly, the cleared lysate was applied on the column at a flow rate of 2.5 ml/min. The column was then washed with 100 ml of buffer A containing 1 mM imidazole, followed by 200 ml of buffer A containing 25 mM imidazole at a flow rate of 10 ml/min. Elution was then performed with buffer A containing 400 mM imidazole. Eluted fractions in a 96-well format were analyzed by sodium dodecyl sulfate–polyacrylamide gel electrophoresis before being pooled and desalted on a Fast Desalting Column HR 10/10 (GE Healthcare Life Sciences, Saclay, France) at a flow rate of 10 ml/min. Protein concentration was determined by UV absorption at 280 nm using extinction coefficients calculated on the basis of tyrosine and tryptophan contents (Gill and von Hippel 1989). After assessment of protein purity, proteins were characterized by mass spectrometry (MALDI-TOF, Voyager DE-RP, Perceptive Biosystem, Boston, MA) and by N-terminal Edman sequencing (Applied Biosystems [Foster City, CA] 473A).

Guanylate Kinase Activity

The purified *R. conorii* Gmk-like (Gmk) and RPE-excised Gmk (Gmk_{del}) were desalted against Tris–HCl 20 mM, pH 7.0, NaCl 50 mM. The reversible reaction catalyzed by Gmk in the presence of Mg^{2+} was measured spectrophotometrically through reduced form of nicotinamide adenine dinucleotide (NADH) oxidation by using the following coupled assay (Li et al. 1996):

$$\mathrm{GMP} + \mathrm{ATP} \overset{\mathrm{GMK}\,(\mathrm{Mg}^{2+})}{\leftrightarrow} \mathrm{GDP} + \mathrm{ADP}$$

$$\mathrm{ADP} + \mathrm{PEP} \overset{\mathrm{PK}\,(\mathrm{Mg}^{2+},\mathrm{K}^{+})}{\leftrightarrow} \mathrm{ATP} + \mathrm{Pyruvate}$$

$$\mathrm{Pyruvate} + \mathrm{NADH} \overset{\mathrm{LH}}{\leftrightarrow} \mathrm{Lactate} + \mathrm{NAD}^{+}.$$

In this assay, the adenosine 5′-diphosphate (ADP) released from the Gmk activity is used by the pyruvate kinase/L-lactate dehydrogenase to convert NADH in NAD^+. The decrease in absorbance at 340 nm over time is then used to follow the reaction. The assay was performed using the 2 *R. conorii* recombinant proteins in a 1-ml reaction solution containing 100 mM Tris–HCl, pH 7.5, 10 mM $MgCl_2$, 5 mM of GMP and 5 mM of ATP, and 4.5 units of pyruvate kinase/L-lactate dehydrogenase (Roche Biochemicals, Mannheim, Germany) in the presence of 1.5 mM phospho(enol)pyruvate mono potassium salt (FLUKA #79415) and 0.2 mM NADH.

For the 2 enzymes, kinetic constants for GMP and ATP were determined by fitting the Michaelis–Menten equation on the data obtained using the same reaction medium at 5 mM ATP concentration and increasing amounts of GMP (0.003 to 5 mM) or at 5 mM GMP and increasing ATP concentrations (0.25 to 10 mM). The enzymes' concentrations were 335 μg/ml for the Gmk and 196 μg/ml for the Gmk_{del} (Bradford protein assay).

Table 1
PCR Detection of the RPE Insertion in *gmk* Gene in Various *Rickettsia* Strains

Strain	*Gmk* Detection	RPE Length (aa)
Rickettsia aeschlemani	Present	36
Rickettsia africae	Present	36
Rickettsia akari	Present	0
Rickettsia australis	Present	0
Rickettsia canadensis	Not amplified	0
Rickettsia conorii	Present	36
Rickettsia felis	Present	0
Rickettsia helvetica	Present	48
Rickettsia japonica	Present	36
Rickettsia massiliae	Present	35
Rickettsia prowazekii	Present	0
Rickettsia rickettsii	Present	36
Rickettsia typhi	Present	0

Activity measurements for kinetic analysis were determined for each substrate concentration at least twice with values of duplicate measurements varying by less than 5%.

gmk Gene Insertion Site Amplification in Other *Rickettsia* Species

Based on the genomic nucleotide sequences of *R. conorii* and *R. prowazekii*, 2 sets of degenerated primers were designed in order to amplify the *gmk* gene as well as the potential RPE1 sequence in other *Rickettsia* genomes. Genomic DNA was purified from the *Rickettsia* strains listed in table 1.

The following primers were purchased from Eurogentec (Seraing, Belgium):

Ugmk: attgattggcaaggagYMag
Rgmk: tttgtgccRactgcatacg
Urpe1: tgagagcRgaagctaatcgt
Rrpe: gacttaccWgtacctgaaggRg
U, forward and R, reverse; Y, C, or T; M, A, or C; W, A, or T.

DNA samples were amplified in a reaction mixture (12.5 μl) that contained 0.5 μl genomic DNA (10–30 μg/ml), 1.25 μl 10× buffer (50% glycerol, 50 mM Tris–HCl, pH 8.0, containing 0.1 mM ethylenediaminetetraacetic acid, 50 mM KCl, 1 mM dithiothreitol, 0.1% Tween-20, and 0.1% Nonidet P-40), 1.25 μl deoxynucleoside triphosphates (2 mM each) (deoxyadenosine triphosphate, deoxyguanosine triphosphate, deoxycytosine triphosphate, and deoxythreonine triphosphate), 0.75 μl forward primer (5 μM), 0.75 μl reverse primer (5 μM), and 0.5 μl template DNA and 0.25 μl KOD (polymerase from *Pyrococcus kodakaraensis*) Dash DNA polymerase (2.5 units/μl) (TOYOBO Co Ltd, Osaka, Japan). Following a first denaturation step (94°C for 2 min), a 3-step cycle of 94°C for 30 s, 53°C for 30 s, and 72°C for 1 min was repeated 35 times. The PCR program was ended by a single 2-min extension step at 72°C (Peltier thermal cycler model PTC 200; MJ Research Inc., Watertown, MA). Amplicons were then resolved by 1% agarose gel electrophoresis and visualized by staining with ethidium bromide. The QIAquick PCR purification kit (Qiagen) was used to prepare amplicons for automated sequencing, which was performed on an Applied Biosystems model ABI 310 automatic DNA sequencer (PerkinElmer, Boston, MA) with dRhodamine Terminator Cycle Sequencing Ready Reaction Buffer (DNA sequencing kit; PerkinElmer) and primers Ugmk and Rgmk. In order to sequence the RPE segment, the Urpe1 and Rrpe primers were used.

Sequence Data

The GenBank accession numbers corresponding to the *Rickettsia* genome records are as follows: *Rickettsia bellii* (CP000087), *Rickettsia typhi* (NC_006142), *R. prowazekii* (NC_000963), *Rickettsia felis* (NC_007109), *Rickettsia akari* (NZ_AAFE00000000), *Rickettsia massiliae* (GenBank accession number to be assigned), *Rickettsia rickettsii* (NZ_AADJ00000000), *R. conorii* (NC_003103), *Rickettsia sibirica* (NZ_AABW00000000), and *Rickettsia africae* (GenBank accession number to be assigned). The *R. massiliae* and *R. africae* genome sequences will be available at http://www.igs.cnrs-mrs.fr/mgdb/Rickettsia/rig/ until the final GenBank accessions are assigned. The GenBank accession numbers corresponding to the *Rickettsia helvetica* gene records are the following: *16S RNA* (L36212), *gltA* (U59723), *sca1* (AY355363), *sca4* (AF163009), *sca5* (AF123725), *tuf* (AF502184), and *fusA* (AF502175). The corresponding genes in the other *Rickettsia* species were recovered from the genome records.

Phylogenetic Distances and Reconstruction

The phylogenetic tree presented in figure 3 was constructed from the concatenated alignments of *16S RNA*, *gltA*, *sca1*, *sca4*, *sca5*, *tuf*, and *fusA*. We aligned each gene family individually using the MUSCLE program (Edgar 2004) and discarded the poorly aligned regions using the GBLOCKS program (Castresana 2000). The alignments were then concatenated to produce a single multiple alignment of 12,082 nucleotide sites. This alignment was analyzed by the Modeltest program (Posada and Crandall 1998) to identify the model of substitution that best fit the data. The maximum likelihood tree was searched using the General Time Reversible + Γ + I model of substitution using the PHYML program (Guindon and Gascuel 2003). The neighbor-joining tree was constructed using the BIONJ program (Gascuel 1997). The Puzzle program (Schmidt et al. 2002) was used to compute pairwise distances between sequences using GTR + Γ + I model. Parameters for Puzzle were as estimated by PHYML from the data (GTR rate parameters, α shape parameter, and the fraction of invariable sites).

Estimation of K_a and ω

The level of substitutions at nonsynonymous sites (K_a) and the ratio $\omega = K_a/K_s$ were estimated with the CODEML program (Yang 1997). We used 2 models of codon substitution implemented by Yang (1998): the 1-ratio model M0 assumes the same ω ratio for all branches in the phylogeny, and the 2-ratio model M2 allows to estimate ω ratios in a specified branch and the rest of the tree independently. Models M0 and M2 were compared with the likelihood ratio test (LRT) (Felsenstein 1981), where twice the log likelihood difference between the models follows a χ^2 distribution with a number of degrees of freedom (df) being the difference of free parameters between the models (i.e.,

df = 1). To determine if the ω ratios were significantly different from 1, we applied the LRT that compares the 1-ratio model with ω estimated from the data and the same model with ω set to 1. The likelihood ratio statistics ($2\Delta \ln L$) was compared with a χ^2 distribution with df = 1.

Sequence Simulation

To simulate "average" *Rickettsia* gene sequences, we first assembled the genes annotated in the 10 *Rickettsia* species sequenced to date into orthologous groups using the reciprocal best BlastP hit criterion (*e* value $< 1 \times 10^{-5}$). We identified 704 core genes conserved in the 10 *Rickettsia* species. For each set of core genes we aligned the predicted translations using the MUSCLE program and generated the corresponding codon alignment. These alignments were then concatenated into a single alignment of 211,423 codons. The topology of the phylogenetic tree of the 10 *Rickettsia* species was determined with the neighbor-joining (Puzzle/BIONJ) and maximum likelihood (PHYML) methods using the concatenated core protein alignment and the Jones, Taylor, and Thornton + Γ substitution matrix. Both methods recovered the same tree topology with maximal bootstrap support for all branches and are in agreement with the tree recovered in figure 3. We used the CODEML program to estimate the level of nonsynonymous substitutions for each branch (i.e., branch length) of the tree from the concatenated codon alignment using the 1-ratio model. The proportions of the branch lengths served as reference to simulate the average gene sequences using the EVOLVER program (Yang 1997). Additional parameters for simulations were as follows: we used the transition/transversion ratio as estimated by CODEML from the concatenated gene alignment. All other parameters (i.e. K_a/K_s ratio, tree length, sequence length, and codon frequency) were estimated by CODEML from the *gmk* alignment (excluding the RPE segment).

Divergence among Orthologous RPE

RPE sequences were identified in the 10 *Rickettsia* genomes using the BlastN program (Altschul et al. 1997), and those located at syntenic location (i.e., inserted in the same gene or surrounded by the same 2 genes) were classified into orthologous groups. We obtained 56 orthologous RPE groups: 23 for RPEs inserted into genes and 32 for RPEs located in intergenic regions. The global level of nucleotide divergence among orthologous sequences was measured by the sum of the branch lengths of the corresponding tree (i.e., referred as the tree length hereafter). The topology of the corresponding tree was deducted from the *Rickettsia* tree reconstructed from the concatenated core proteins (see Sequence Simulation). We used the BASEML program (Yang 1997) to estimate the tree length using the GTR model. The different sets of orthologous RPEs did not contain the same species, and so, the tree lengths were not comparable directly. We therefore normalized the RPE tree lengths by the tree lengths estimated from the 47,594 four-fold–degenerated sites in 612 *Rickettsia* core genes for the same set of species. Four-fold–degenerated sites are the positions in codon sequences where any possible nucleotide change results in a synonymous substitution and so are expected to be free from selection and to evolve at a rate similar to the mutation rate (Kimura 1983).

Structural Modeling

We first reconstructed the Gmk protein sequence of the last common ancestor of *R. conorii* and *R. felis*. The ancestral sequence was inferred from the Gmk protein alignment and the 10-species tree topology (fig. 3*A*) using the marginal reconstruction method (Yang et al. 1995) implemented in the CODEML program. The average accuracy of the reconstructed sequence was estimated at 99.53% per site. In order to build a model of both the reconstructed ancestor and the *R. conorii* GMK structures, a structural alignment was performed combining the information coming from available GMK structures with the sequence information coming from a set of reference GMK sequences sharing at least 33% identity with the last common ancestor and the *R. conorii* GMK sequences. The multiple structure–sequence alignments produced with 3DCoffee (Poirot et al. 2004) is then followed by homology model building using MODELLER (Sali and Blundell 1993) to generate the 2 GMK models (fig. 5).

Reference GMK structures (Protein Data Bank accession numbers) are as follows: 1Z6G, *Plasmodium falciparum*; 1S4Q, *Mycobacterium tuberculosis*; 1LVG, *Mus musculus*; 1EX7, *Saccharomyces cerevisiae*; and 2ANB, *E. coli*.

Reference GMK sequences (TrEMBL) are as follows: KGUA_RICCN (Q92GC9), *R. conorii*; KGUA_HAEIN (P44310), *Haemophilus influenzae*; KGUA_CLOPE (Q8XJK8), *Clostridium perfringens*; KGUA_RICFE (Q4UK55), *R. felis*; KGUA_RICTHY (Q68VY3), *R. typhi*; KGUA_THEET (Q3CFM9), *Thermoanaerobacter ethanolicus*; KGUA_ALTAT (Q3IJH8), *Pseudoalteremonas atlantica*; KGUA_ECOLI (P60546), *E. coli*; KGUA_ZYMMO (Q5NQE8), *Zymomonas mobilis*; KGUA_HAEDU (Q7VKP3), *Haemophilus ducreyi*; KGUA_ANC, reconstructed *Rickettsia* ancestor (supplementary fig. S2, Supplementary Material online); KGUA_Rprowazekii (Q9ZCH7), *R. prowazekii*.

Results

Rickettsia genes have proven very challenging to express (Alexeyev and Winkler 1999), probably due in part to the anomalous codon usage resulting from their low G + C content (30%). However, it is worth to notice that our attempt to work from synthetic genes (*alr* and *ubiG*) conforming to *E. coli* optimal coding rules did not succeed, suggesting the influence of other properties perhaps depending on the amino acid sequence/composition of the tentative protein products. After *R. prowazekii* ADP/ATP translocase (Alexeyev and Winkler 1999) and the *in vitro* translated phospholipase D (Renesto et al. 2003), *R. conorii* Gmk is only the third enzymatic protein from *Rickettsia* to be functionally characterized.

Direct Characterization of 4 *R. conorii* RPE-Containing ORFs

One goal of this study was to achieve the isolation and characterization of several *R. conorii* RPE–containing

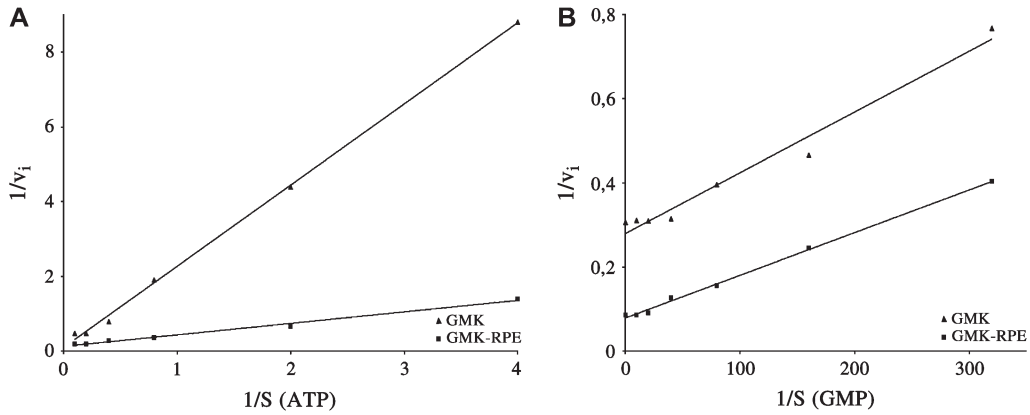

FIG. 1.—Gmk and Gmk_{del} affinity. Double reciprocal plot of a Lineweaver–Burk plot for 196 μg Gmk and Gmk_{del} enzymes. (*A*) The GMP concentration was held constant at 5 mM, and the concentration of ATP ranged from 0.25 to 10 mM. (*B*) The ATP concentration was held constant at 5 mM, and the concentration of GMP ranged from 0.003 to 5 mM.

ORFs to confirm the presence of the RPE-encoded amino acids in the final translation product. The same purification protocol was applied in parallel to the 4 ORFs expressed in 200-ml cultures (see Materials and Methods). The 4 ORFs (*era*, *gltX*, *gmk*, and RC0209) led to amounts of eluted proteins suitable for mass spectrometry analysis. A single molecular species was observed in each case where the measured versus predicted molecular weight is 12,940/13,193 for RC0209; 28,586/28,715 for Gmk; 40,944/41,109 for Era; and 60,833/60,980 for GltX (including the His_6 tag and the attB linker). The small observed differences (RC0209: 253 Da; Gmk: 129 Da; Era: 165 Da; and GltX: 147 Da) correspond to the loss of N-terminal methionine (131 Da) plus a residual difference of 121, 34, 15, and 2 Da. Given that the excision of the RPE-encoded peptides would have led to a molecular weight change of 3–5 kDa, we can confidently conclude that the RPE-derived peptides are not excised in these 4 proteins (supplementary fig. S1, Supplementary Material online). These results validate earlier predictions that RPE-containing ORFs should produce proteins containing RPE-derived amino acid sequences (Ogata et al. 2000).

Following the analysis of the location of the RPE element with respect to the protein 3-dimensional structure, Ogata et al. (2000) also proposed that the inserted element would neither affect the fold of the host protein nor interfere with its function. The successful production of the *R. conorii* Gmk protein, homologous to *E. coli* guanylate kinase, offered the possibility to assay its integrity (i.e., proper folding and enzymatic activity). Here we demonstrate that the *R. conorii gmk* ORF indeed encodes a bona fide guanylate kinase, the activity of which is proportional to the protein concentration (data not shown). This confirms the annotation of the gene from its sequence homology with other guanylate kinases. We thus compared the enzyme activity of the two Gmk and Gmk_{del} enzymes using an assay coupled to lactate dehydrogenase activity (Li et al. 1996). Initial kinetic measurements were made using 335 μg of Gmk (11.6 nmol) and 196 μg of Gmk_{del} (8.2 nmol) per reaction. Pseudo–first-order kinetics were used to determine the apparent K_m and V_{max} for ATP and GMP (see Materials and Methods) (figs. 1 and 2). The excision of the RPE domain in the Gmk protein improves the affinity for ATP by an order of magnitude (Gmk K_m = 23.6 mM

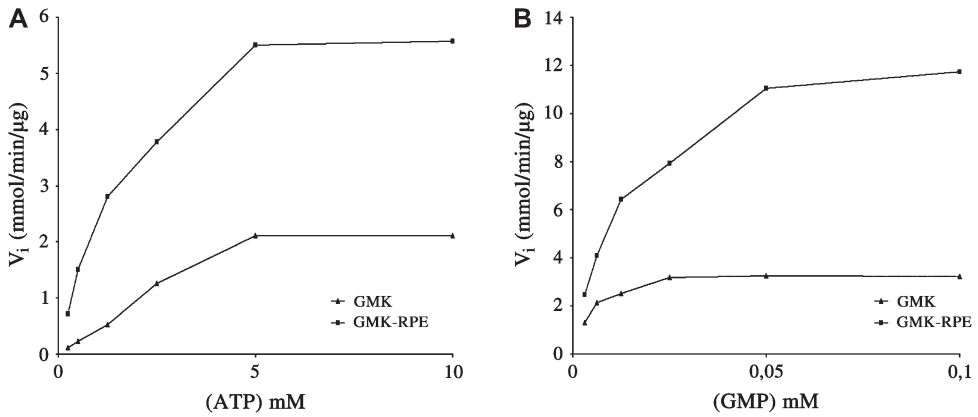

FIG. 2.—Kinetic analysis of the Gmk and Gmk_{del} enzymes. Gmk and Gmk_{del} initial velocity measurements were performed using, respectively, 11.6 nmol and 8.2 nmol of enzyme in each reaction. The phosphoryl transfer of ATP to GMP was coupled to pyruvate kinase–lactate dehydrogenase enzymes such that absorbance change at 340 nm, detecting the $NADH^+$ consumption (coupled to the release of ADP), revealed the guanylate kinase activity. V_i and $1/V_i$ values were computed for 196 μg of Gmk and Gmk_{del} enzymes. (*A*) The GMP concentration was held constant at 5 mM, and the concentration of ATP ranged from 0.25 to 10 mM. (*B*) The ATP concentration was held constant at 5 mM, and the concentration of GMP ranged from 0.003 to 5 mM.

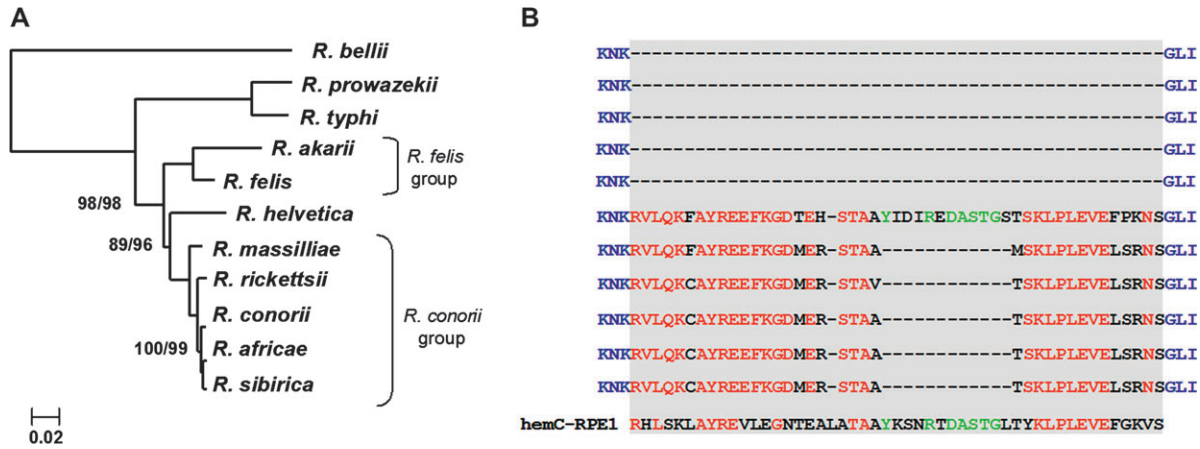

FIG. 3.—Distribution of the RPE_{gmk} insert among *Rickettsia*. (*A*) Maximum likelihood phylogenetic tree of the *Rickettsia* species reconstructed from the concatenated alignments of 7 genes (*16S RNA*, *gltA*, *fusA*, *tuf*, *sca1*, *sca4*, and *sca5*; 12,082 nucleotide sites). The GTR + Γ + I model of substitution was used for the reconstruction. All branches were supported by 100% bootstrap support by both maximum likelihood (ML) and neighbor-joining (NJ) methods except the branches for which the bootstrap values are indicated beside (ML/NJ). (*B*) Multiple alignment of the Gmk protein around the RPE insert. The shaded box corresponds to the RPE region. The complete RPE sequence of the hemC protein is given for reference. Strictly conserved residues are color coded in red. Strictly conserved residues between the Gmk 10 residues' longer repeat and the corresponding 10 residues' segment in hemC are colored in green. The blue residues correspond to the Gmk region in *Rickettsia* surrounding the RPE insert.

and Gmk_{del} K_m = 2.56 mM), whereas the V_{max} values of ATP consumption remained very close at 0.29 mM/min/mg and 0.21 mM/min/mg for the Gmk and Gmk_{del} enzymes, respectively. On the other hand, the affinity for GMP of the RPE-deleted recombinant enzyme was lowered (K_{mGmk} = 0.0051 mM, $K_{mGmkdel}$ = 0.013 mM) while we observed a 3-fold increase in the V_{max} of guanosine diphosphate production for Gmk_{del} (0.34 mM/min vs. 0.092 mM/min/mg for Gmk). Overall, the removal of the RPE insert is thus associated with a 10-fold increase in enzymatic efficiency (as crudely expressed by the V_{max}/K_m ratio).

Phylogenetic Position of the RPE Insertion in *gmk*

We determined the partial sequence of the *gmk* gene across a panel of available species, using PCR amplification, to infer the date of the original RPE insertion event in the *Rickettsia gmk* gene. Amplification products were obtained for all tested *Rickettsia* strains. All amplicons were then sequenced at least in triplicate, except for *Rickettsia canadensis* where the *gmk* gene is absent. When primers Urpe1 and Rrpe were used (see Materials and Methods), fragment lengths ranged from 450 to about 300 bp depending on the presence or absence of the RPE repeat (fig. 3*B*). Superposition of the RPE distribution onto the *Rickettsia* phylogenetic tree indicates that the insertion of the repeat in the *gmk* gene occurred after the divergence of *R. felis* and *R. conorii* but before the divergence of *R. helvetica* and *R. conorii* (fig. 3*A*). Thus, the acquisition of the RPE_{gmk} is relatively ancient and predates the divergence of the *R. conorii* group (as defined in fig. 3*A*). Evidence for ancient insertions of two other RPE repeats in *Rickettsia* was also reported by Amiri et al. (2002). Comparison of the translated RPE sequences from the various *Rickettsia* strains that contained the repeat in the *gmk* gene revealed sequence identity going from 94% to 100%. The RPE sequence inserted in the Gmk protein is 10 residues shorter than the standard RPE1 (Claverie and Ogata 2003) in most tested *Rickettsia* species (fig. 3*B*), except for the *gmk* gene of *R. helvetica*, suggesting that the original insertion event involved a full-length RPE that was subsequently shortened in the other species.

Divergence of the Gmk Protein

The insertion of the RPE sequence may have triggered adaptive changes from the *gmk* gene to account for the presence of the extra peptide in the protein. A classical strategy to detect adaptive evolution is to compare the level of nonsynonymous substitutions (K_a) with the rate of mutation, which can be approximated with the level of synonymous substitution (K_s). The ratio $\omega = K_a/K_s$ is used to measure the magnitude and direction of selective constraints acting on protein sequences, with $\omega = 1$, < 1, and > 1 indicating neutral evolution, purifying selection, and positive selection, respectively (Li 1997). For practical reasons, we restricted our analysis to the 10 *Rickettsia* species for which the *gmk* sequence is available in public databases (fig. 4*A*). We first estimated ω as an average over all branches of the *gmk* 10-species tree using the model M0. The ratio was substantially less than 1 (ω_0 = 0.1506), indicating that the *gmk* evolution was mostly driven by purifying selection. Estimation of ω when the branch corresponding to the RPE insertion and the rest of the tree are considered separately (2-ratio model M2) suggests that an increase of ω occurred in the branch of the RPE insertion ($\omega_{insertion}$ = 0.2279 vs. ω_0 = 0.1470). However, there is no evidence of positive selection as $\omega_{insertion}$ is less than one. In addition, the LRT indicated that the model M2 provided no significant better fit to the data than model M0 ($2\Delta \ln L = 0.36$, P = 0.55). Thus, we cannot reject the hypothesis that the ω ratios in the branch of the RPE insertion and the rest of the tree are homogeneous.

We used another approach that can provide evidence of adaptation by detecting an increased accumulation of nonsynonymous substitutions in the branch to the insertion

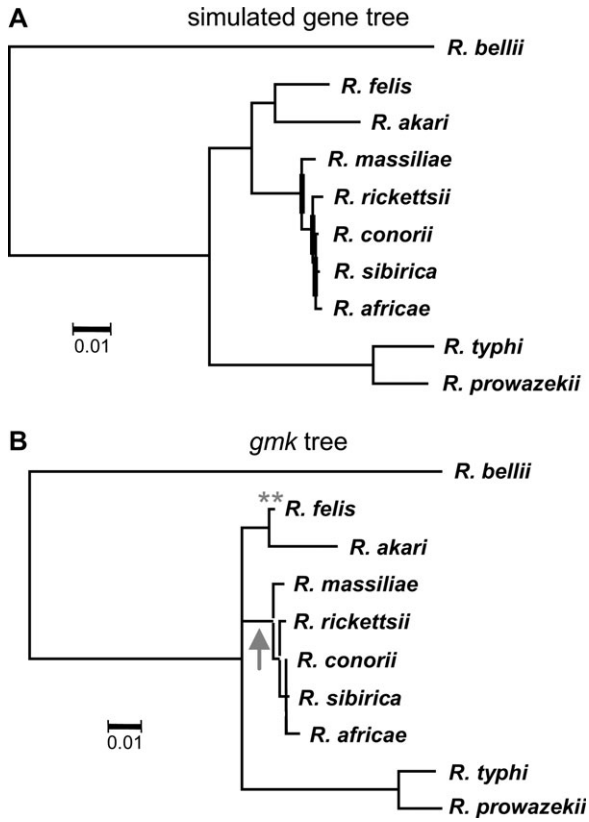

FIG. 4.—Branch lengths of the simulated gene and gmk trees. Branch lengths are expressed in number of substitutions per nonsynonymous site (K_a). (*A*) The tree topology was recovered from the concatenated alignment of 612 core proteins conserved in the 10 *Rickettsia* species. The branch lengths correspond to the mean K_a values estimated from 1,000 data sets of 10 simulated sequences. (*B*) Branch lengths estimated from the *gmk* alignment. The arrow shows the branch corresponding to the insertion of the RPE repeat. The double asterisk indicates the branch significantly shorter ($P < 0.01$) than the corresponding branch in the simulated gene tree. See supplementary table S1 (Supplementary Material online) for further details on the branch length values.

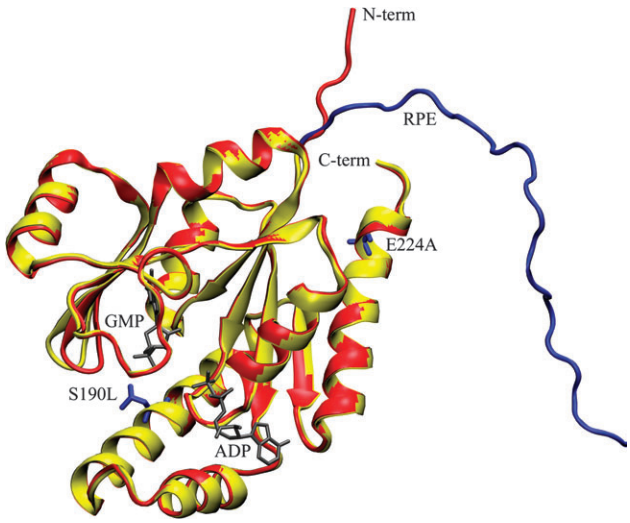

FIG. 5.—Structural models of the *Rickettsia conorii* and ancestral Gmk proteins. Model of the *R. conorii* RPE–containing Gmk structure (yellow) superimposed on the ancestral Gmk model structure (red). The GMP and ATP appear in gray to help visualize the active site of the Gmk structure. The 2 variable positions (S190L and E224A) in the RPE-containing Gmk as well as the RPE insert are mapped onto the Gmk model (blue). See supplementary figure S2 (Supplementary Material online) presenting the multiple alignment of all available Gmk sequences. The variable positions, between RPE-containing and RPE-free sequences, relative to the ancestral Gmk sequence are color coded.

event. Standard methods for identifying acceleration of divergence rest on the molecular clock assumption. These tests could not be applied to the *gmk* sequences due to the substantial variations of mutation rates between different *Rickettsia* lineages therefore violating the molecular clock assumption de facto. To circumvent this problem, we compared the branch lengths of the real *gmk* tree, measured by K_a, with those expected for an average *Rickettsia* gene retaining the same global proprieties of the rickettsial *gmk* (i.e., length, nucleotide composition, level of divergence, and mean ω). We generated 1,000 data sets of 10 artificial sequences using a model that simulates the evolution of the average gene along the *Rickettsia* tree. For each data set, we estimated the K_a value per branch and then derived the average length and standard deviation for each branch of the tree (fig. 4*A*). The branch corresponding to the RPE insertion did not present any significant increase of K_a when compared with the same branch in the average gene tree (fig. 4*B*; supplementary table S1, Supplementary Material online). No significant increase of K_a was found for the branches after the insertion either. These results reinforce the hypothesis that the insertion of the RPE repeat did not trigger adaptive changes from the rest of the Gmk protein. Intriguingly, we found that the *R. felis* branch in the gmk tree is significantly shorter than that of the simulated data ($P = 0.0015$; fig. 4*B*; supplementary table S1, Supplementary Material online). The *gmk* gene has accumulated only one nonsynonymous substitution in this branch, whereas the expected number for the average gene is 7.0 ± 2.0. The *R. felis gmk* gene may have fixed a smaller number of nonsynonymous substitutions than expected as a result of an increased level of purifying selection. However, the reason as to why the level of selective constraint acting on the *gmk* gene would have shifted upward in the *R. felis* lineage is unclear.

To further investigate the evolution of the Gmk protein, we reconstructed the ancestral sequence as it existed before the RPE insertion event (i.e., in the last common ancestor of *R. felis* and *R. conorii*) (supplementary fig. S2, Supplementary Material online). The reconstructed ancestral sequence exhibits 11 amino acid differences with the *R. conorii* Gmk protein, excluding the RPE region. Nine of the 11 changing sites (supplementary fig. S2, Supplementary Material online) are also variable among the RPE-less Gmk (i.e., those of *R. bellii*, *R. prowazekii*, *R. typhi*, *R. felis*, and *R. akari*). For the 2 remaining variable positions (S190L and E224A; supplementary fig. S2, Supplementary Material online), some RPE-containing sequences retained the ancestral residue. Thus, the amino acid changes that occurred since the insertion of the RPE repeat occurred mostly at sites that are naturally variable and presumably under weaker functional constraint. The predicted spatial structures of the 2 enzymes are almost perfectly superposable except for the N-terminal part containing the RPE peptide in *R. conorii*. The 11 amino acid changes and the RPE insert

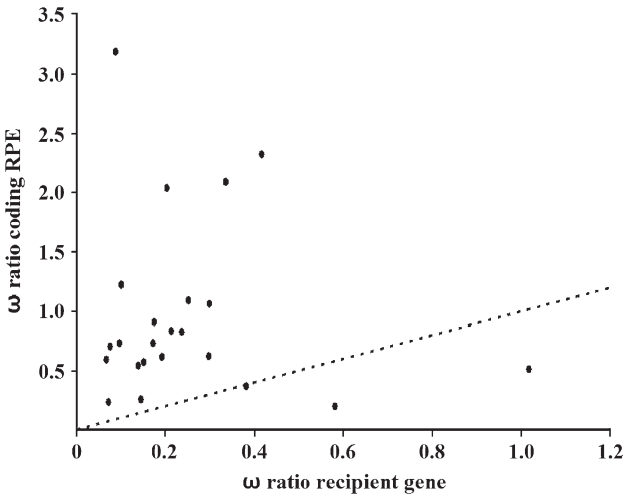

FIG. 6.—ω ratio for RPE repeats and their recipient genes. The dashed line represents the $x = y$ curve. See supplementary table S2 (Supplementary Material online) for further details on values.

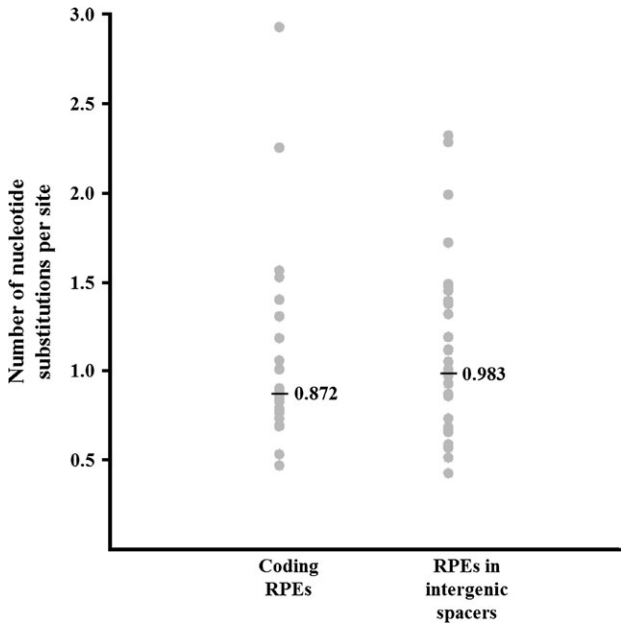

FIG. 7.—Rates of sequence divergence for coding and intergenic orthologous RPE groups. Dots represent the overall rates of sequence divergence of orthologous RPEs' families calculated by dividing the global level of nucleotide divergence among orthologous RPE sequences by the global level of divergence in the core genes at 4-fold degenerated sites (see Materials and Methods). Black horizontal bars indicate the median rate value for the category.

did not result in any significant structural difference in the domains common to both proteins (fig. 5). Most amino acid changes are located at the surface exposed to solvent, and none occurred inside or in the vicinity of the ligand-binding sites. No structural model could be generated for the *R. conorii* RPE peptide, but earlier predictions indicated that it most likely adopts an alpha-helix conformation (Ogata et al. 2000). Because the RPE peptide folds probably into a distinct domain at the surface of the enzyme we cannot rule out that it could reduce the substrates' accessibility to the binding sites. This reduced accessibility may be the primary determinant of the lower affinity for ATP of the native *R. conorii* Gmk compared with the RPE-deleted recombinant enzyme.

Evolution of RPE Sequences

To determine if the RPE sequence evolved under similar constraint as the rest of the *gmk* gene, we estimated the ω ratio of the 2 regions separately. As a result, the estimated ω ratio is higher in the RPE region ($\omega_{RPE} = 0.616$) than in the rest of the *gmk* coding sequence ($\omega_{gmk} = 0.192$), suggesting that the repeat evolved under lower levels of purifying selection. The LRT indicates that the ω_{RPE} ratio is not significantly different from 1 ($2\Delta\ln L = 1.67$; $P = 0.196$). Consequently, the RPE sequence appears to have evolved under very limited constraint. A similar analysis of 22 other RPE-containing genes indicates that most (19) RPE regions also evolved under lower levels of purifying selection than their recipient coding sequences (fig. 6; supplementary table S2, Supplementary Material online), this overall trend being highly significant ($\chi^2 = 8.5$, $P = 0.003$). In 7 cases, the estimated ω_{RPE} ratios were >1, suggesting that these RPE regions could have been subject to positive selection. However, these ω ratios were not significantly higher than 1 (LRT; supplementary table S2, Supplementary Material online). Only one gene encoding a putative glutamine amidotransferase contains an RPE with ω significantly <1 ($\omega = 0.26$; $P = 0.0.001$; supplementary table S2, Supplementary Material online) indicating that the evolution of this repeat has been markedly constrained by purifying selection. In summary, our data show that the process of nucleotide substitution in coding RPEs is generally only weakly constrained: coding RPEs might have tolerated most changes provided they do not result in premature stop codons. It could be argued that these RPE sequences evolved neutrally because the insertion of the repeats could have turned the recipient gene into a pseudogene. This hypothesis is unlikely because many of these genes encode proteins with essential functions in other microorganisms. In addition, the ω ratios estimated from the coding sequences (excluding the RPE regions) were significantly <1 for 19 out of these 23 genes (supplementary table S2, Supplementary Material online), indicating that these genes evolved under purifying selection and therefore are probably functional.

To further characterize the evolution of the RPE repeats, we compared the rates of divergence of coding RPEs with those lying in intergenic sequences. In this analysis, the divergence rate was defined as the global level of nucleotide divergence among orthologous RPE sequences normalized by the global level of divergence in the core genes at 4-fold–degenerated sites (see Materials and Methods). The divergence rates are distributed over similar ranges in the 2 RPE categories: from 0.43 to 2.31 for the coding RPEs and from 0.47 to 2.92 for the intergenic RPEs (fig. 7). The RPE_{gmk} exhibits a divergence rate (0.863) close to the median value for the coding RPE category (0.872). The median divergence rate for intergenic RPE (0.983) is close to 1 as expected for sequences presumably diverging under neutral evolution (Amiri et al. 2003).

Although the median divergence rate is slightly higher for intergenic RPEs than for coding RPEs, the difference is not statistically significant (Wilcoxon test, $P \leq 0.4513$). Hence, there is no evidence that coding RPEs tend to evolve under significantly higher selective constraints than their intergenic homologs. This result further confirms that nucleotide substitutions in coding RPEs were generally weakly constrained.

Discussion

The presence of the RPE insert in many different genes encoding important enzymes in all *Rickettsia* genomes sequenced so far remains a mystery. These coding palindromic repeats appear strictly evolutionary linked to the emergence of intracellular parasitisms in the α-proteobacteria division (Ogata, Suhre, et al. 2005). Their study is made troublesome by the difficulties encountered to express recombinant proteins of *Rickettsia* origin, a problem usually attributed to their high A + T content and anomalous codon usage. Using a parallel attempt on a set of RPE-containing proteins, we now demonstrated the presence of the expected RPE-derived peptide in the recombinant products of 4 genes: *era*, *gltX*, *gmk*, and RC0209. The RPE-derived peptide (45 to 50 amino acid long) is thus not spontaneously excised (as would be an intein). Examination of evolutionary parameters indicates that once inserted in genes, the RPE repeats evolved under weak constraints, similar to those acting on neutrally evolving intergenic RPEs. Inversely, the recipient coding sequences were still evolving under selective constraints. Furthermore, analyses of the rates of nonsynonymous substitutions and the location of the amino acid changes in the Gmk protein suggest that the repeat insertion neither triggered adaptive changes in the rest of the protein nor significantly modified the core enzyme structure. The influence of the repeat insertion on the enzyme activity was examined in detail for the Gmk gene product, using a construct from which the RPE-derived sequence was removed. The largest observed variation in enzymatic parameters was a 9-fold increased affinity for ATP (K_{mGmkdel} = 2.6 mM vs. K_{mGmk} = 23.6 mM) exhibited by the RPE-free enzyme. This effect was not unexpected given that the N-terminal RPE (position 7 to 42) of the Gmk protein is located at the immediate proximity of the ATP-binding site (position 51 to 58, with a reactive serine at position 53, by analogy with the *E. coli* homologue). The presence of the RPE-derived peptide might thus slightly hinder the accessibility of the ATP-binding site. This effect is partially compensated by a 38% increase of the V_{max} of ATP consumption for the RPE-containing enzyme. Interestingly, the deletion of the RPE insert exhibited inverse consequence on the GMP kinetic parameters: the affinity for GMP was decreased 2.5-fold (K_{mGmk} = 0.0051 mM, K_{mGmkdel} = 0.013 mM), and the V_{max} of GDP production increased 3.7-fold. Overall, the RPE-free enzyme appears 10-fold more efficient than its *Rickettsia* counterpart.

Now, how can we place these results in the context of *Rickettsia* evolutionary history? First, we have to be careful when interpreting the details of the changes induced in the various kinetic parameters, as their relevance to the *in vivo* situation (intracellular parasitism) is difficult to assess. The presence of the RPE globally appears to be slightly detrimental to the efficiency of the Gmk enzyme. We expect this to be true for the ancestral Gmk protein that acquired the RPE insert and for all the other RPE-containing proteins. It is generally thought that the fixation of slightly deleterious mutations is the consequence of genetic drift and is detrimental to the organism fitness (Ohta 1992). In line with this view, Amiri et al. (2002) have hypothesized that insertions of RPEs in genes were neutral or might have been tolerated in *Rickettsia* even if it had a negative effect on protein function because of recurrent bottlenecks and/or small population sizes. However, the picture might be slightly more complicated in the case of parasitic organisms. A loss of transcriptional regulator appears to be a general trend in the evolution of obligate intracellular bacteria, and most of their genes are likely to be constitutively expressed (Wernegreen 2002; Foster et al. 2005). In this context, modifying the kinetic parameters of metabolic enzymes might become a useful mechanism by which to convey subtle adaptive evolutionary changes. Energy parasitism by the capture of ATP from the host cell is a trademark of obligate intracellular bacteria belonging to Rickettsiales and Chlamydiales (Schmitz-Esser et al. 2004). This is achieved by means of an ATP/ADP translocase that exhibits an affinity for ATP in the 0.1 mM range (Dunbar and Winkler 1997). Assuming an *in vivo* concentration of ATP comparable to the ATP/ADP translocase K_{m} value, the insertion of the RPE in the original Gmk could have caused a 20-fold slowdown in the overall production of GDP. However, decreasing the efficiency of this particular enzyme might represent a fitness improvement of the intracellular parasite as a whole, by optimizing the utilization of a limiting quantity of ATP to be divided among many other metabolic pathways. From this, we may hypothesize that the insertion of RPE in the various *Rickettsia* enzymes might always be slightly detrimental to the molecular functions (i.e., slightly diminish their efficiency) while still increasing the fitness of the organism by slowing down its metabolism to a rate compatible with its parasitic/symbiotic way of life within slowly replicating arthropod cells. Comparable examples of adaptation by functional impairment are the selection of less virulent strains of myxoma virus, which was initially highly lethal in Australian rabbits (Fenner 2000) or the adaptive mutation resulting in a 10-fold reduced activity of the listeriolysin O-enzyme that led to a 100-fold increase in virulence in the mouse listeriosis model (Glomski et al. 2002). Obviously, the question as to which of the 2 hypotheses (i.e., genetic drift vs. adaptation) best explains the fixation of coding RPEs is still unsettled and will require further investigation.

Supplementary Material

Supplementary figures S1 and S2 and tables S1 and S2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was partially supported by the French National Génopole Network. We gladly acknowledge the use

of the Marseille-Nice Génopole bioinformatics and proteomics platform. We wish to thank Dr Jacques Bonicel and Dr Régine Lebrun for their dedication and expertise for the mass spectrometry analyses. We also wish to thank the associate editor and the referees for helpful comments and suggestions.

Funding to pay the Open Access publication charges for this article was provided by CNRS (Centre National de la Recherche Scientifique).

Literature Cited

Alexeyev MF, Winkler HH. 1999. Gene synthesis, bacterial expression and purification of the *Rickettsia prowazekii* ATP/ADP translocase. Biochim Biophys Acta 1419: 299–306.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–402.

Amiri H, Alsmark CM, Andersson SG. 2002. Proliferation and deterioration of *Rickettsia* palindromic elements. Mol Biol Evol 19:1234–43.

Amiri H, Davids W, Andersson SG. 2003. Birth and death of orphan genes in *Rickettsia*. Mol Biol Evol 20:1575–87.

Andersson SG, Zomorodipour A, Andersson JO, Sicheritz-Ponten T, Alsmark UC, Podowski RM, Naslund AK, Eriksson AS, Winkler HH, Kurland CG. 1998. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. Nature 396:133–40.

Britten R. 2006. Transposable elements have contributed to thousands of human proteins. Proc Natl Acad Sci USA 103:1798–1803.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17:540–52.

Claverie JM, Ogata H. 2003. The insertion of palindromic repeats in the evolution of proteins. Trends Biochem Sci 28:75–80.

Dunbar SA, Winkler HH. 1997. Increased and controlled expression of the *Rickettsia prowazekii* ATP/ADP translocase and analysis of cysteine-less mutant translocase. Microbiology 143(Pt 11):3661–9.

Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368–76.

Fenner F. 2000. Adventures with poxviruses of vertebrates. FEMS Microbiol Rev 24:123–33.

Foster J, Ganatra M, Kamal I, et al. (26 co-authors). 2005. The *Wolbachia* genome of *Brugia malayi*: endosymbiont evolution within a human pathogenic nematode. PLoS Biol 3:e121.

Gascuel O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol 14:685–95.

Gentry D, Bengra C, Ikehara K, Cashel M. 1993. Guanylate kinase of *Escherichia coli* K-12. J Biol Chem 268:14316–21.

Gill SC, von Hippel PH. 1989. Calculation of protein extinction coefficients from amino acid sequence data. Anal Biochem 182:319–26 [erratum in Anal Biochem 1990, 189:283].

Glomski IJ, Gedde MM, Tsang AW, Swanson JA, Portnoy DA. 2002. The *Listeria* monocytogenes hemolysin has an acidic pH optimum to compartmentalize activity and prevent damage to infected host cells. J Cell Biol 156:1029–38.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52:696–704.

Kimura M. 1983. The neutral theory of molecular evolution. New York: Cambridge University Press.

Li WH. 1997. Molecular evolution. Sunderland, MA: Sinauer Associates.

Li Y, Zhang Y, Yan H. 1996. Kinetic and thermodynamic characterizations of yeast guanylate kinase. J Biol Chem 271:28038–44.

Novy R, Drott D, Yaeger K, Mierendorf R. 2001. Overcoming the codon bias of. *E. coli* for enhanced protein expression. Innovations 12:1–3.

Ogata H, Audic S, Abergel C, Fournier PE, Claverie JM. 2002. Protein coding palindromes are a unique but recurrent feature in *Rickettsia*. Genome Res 12:808–16.

Ogata H, Audic S, Barbe V, Artiguenave F, Fournier PE, Raoult D, Claverie JM. 2000. Selfish DNA in protein-coding genes of *Rickettsia*. Science 290:347–50.

Ogata H, Audic S, Renesto P, et al. (11 co-authors). 2001. Mechanisms of evolution in *Rickettsia conorii* and *R. prowazekii*. Science 293:2093–8.

Ogata H, La Scola B, Audic S, Renesto P, Blanc G, Robert C, Fournier PE, Claverie JM, Raoult D. 2006. Genome sequence of *Rickettsia bellii* illuminates the role of amoebae in gene exchanges between intracellular pathogens. PLoS Genetics 2:e76.

Ogata H, Renesto P, Audic S, Robert C, Blanc G, Fournier PE, Parinello H, Claverie JM, Raoult D. 2005. The genome sequence of *Rickettsia felis* identifies the first putative conjugative plasmid in an obligate intracellular parasite. PLoS Biol 3:e248.

Ogata H, Suhre K, Claverie JM. 2005. Discovery of protein-coding palindromic repeats in *Wolbachia*. Trends Microbiol 13:253–5.

Ohta T. 1992. The nearly neutral theory of molecular evolution. Annu Rev Ecol Syst 23:263–86.

Paulus H. 2000. Protein splicing and related forms of protein autoprocessing. Annu Rev Biochem 69:447–96.

Poirot O, Suhre K, Abergel C, O'Toole E, Notredame C. 2004. 3DCoffee@igs: a web server for combining sequences and structures into a multiple sequence alignment. Nucleic Acids Res 32:W37–40.

Posada D, Crandall KA. 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics 14:817–8.

Raoult D, Roux V. 1997. Rickettsioses as paradigms of new or emerging infectious diseases. Clin Microbiol Rev 10: 694–719.

Renesto P, Dehoux P, Gouin E, Touqui L, Cossart P, Raoult D. 2003. Identification and characterization of a phospholipase D-superfamily gene in *rickettsiae*. J Infect Dis 188: 1276–83.

Sali A, Blundell TL. 1993. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234: 779–815.

Schmidt HA, Strimmer K, Vingron M, von Haeseler A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18: 502–4.

Schmitz-Esser S, Linka N, Collingro A, Beier CL, Neuhaus HE, Wagner M, Horn M. 2004. ATP/ADP translocases: a common feature of obligate intracellular amoebal symbionts related to *Chlamydiae* and *Rickettsiae*. J Bacteriol 186: 683–91.

Suyama M, Lathe WC III, Bork P. 2005. Palindromic repetitive DNA elements with coding potential in *Methanocaldococcus jannaschii*. FEBS Lett 579:5281–6.

Walhout AJ, Temple GF, Brasch MA, Hartley JL, Lorson MA, van den Heuvel S, Vidal M. 2000. GATEWAY recombinational cloning: application to the cloning of large numbers of open reading frames or ORFeomes. Methods Enzymol 328:575–92.

Wernegreen JJ. 2002. Genome evolution in bacterial endosymbionts of insects. Nat Rev Genet 3:850–61.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555–6.

Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15:568–573.

Yang Z, Kumar S, Nei M. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. Genetics 141:1641–50.

Herve Philippe, Associate Editor

Accepted August 1, 2006