Available online at www.sciencedirect.com

SCIENCE DIRECT

Gene 352 (2005) 109 – 117

GENE
SECTION
EVOLUTIONARY GENOMICS

www.elsevier.com/locate/gene

Gene expression levels influence amino acid usage and evolutionary rates in endosymbiotic bacteria

Jörg Schaber[a,*], Claude Rispe[b], Jennifer Wernegreen[c], Andreas Buness[d], François Delmotte[a,e], Francisco J. Silva[a], Andrés Moya[a]

[a]*Institut Cavanilles de Biodiversitat i Biologia Evolutiva, Universitat de Valencia, A.C. 22085, 46071 Valencia, Spain*
[b]*UMR Biologie des organismes et des populations appliquée à la protection des plantes, INRA, BP35327, 35653 Le Reu cedeux, France*
[c]*Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine, Biological Laboratory, 7 MBL Street, Woods Hole, MA 02543, USA*
[d]*Deutsches Krebsforschungszentrum DKFZ, Heidelberg, Germany*
[e]*UMR Santé Végétale (INRA-ENITAB), BP81, 33883 Bordeaux, France*

Received 21 September 2004; received in revised form 25 January 2005; accepted 1 April 2005

Received by F.G. Alvarez-Valin

Abstract

Most endosymbiotic bacteria have extremely reduced genomes, accelerated evolutionary rates, and strong AT base compositional bias thought to reflect reduced efficacy of selection and increased mutational pressure. Here, we present a comparative study of evolutionary forces shaping five fully sequenced bacterial endosymbionts of insects. The results of this study were three-fold: (i) Stronger conservation of high expression genes at not just nonsynonymous, but also synonymous, sites. (ii) Variation in amino acid usage strongly correlates with GC content and expression level of genes. This pattern is largely explained by greater conservation of high expression genes, leading to their higher GC content. However, we also found indication of selection favoring GC-rich amino acids that contrasts with former studies. (iii) Although the specific nutritional requirements of the insect host are known to affect gene content of endosymbionts, we found no detectable influence on substitution rates, amino acid usage, or codon usage of bacterial genes involved in host nutrition.

Keywords: *Blochmannia*; *Wigglesworthia*; *Buchnera*; Substitution rates; Insects

1. Introduction

The distinct lifestyle of obligatory endosymbiotic bacteria imposes population bottlenecks upon transmission to host offspring and restricts recombination with genetically distinct bacteria, both of which reduce the effective population size of these bacteria compared to free-living relatives. According to the nearly neutral theory of evolution (Ohta, 1987, 1992), small populations will experience reduced efficacy of selection and thus elevated rates of deleterious changes due to genetic drift. The study of complete genomes is valuable to weigh the respective effects of mutational pressures and selective pressures, both possibly acting at different levels, e.g. evolution of gene repertoires, optimization of translation, changes in base composition. Five complete genomes from endosymbotic

Abbreviations: AT, Adenine–Tymidine; GC, Guanine–Cytosine; BAp, *Buchnera Acrythosiphon pisum*; BSg, *Buchnera Schizaphis graminum*; BBp, *Buchnera Baizongia pistaciae*; Wgl, *Wigglesworthia glossinidia brevipalpis*; Bfl, *Blochmannia floridanus*; EK12, *Escherichia coli* K-12; Ypo, *Yersinia pestis* strain CO92; HEG, high expression genes; LEG, low expression genes; PLEG, putative low expression genes; PHEG, putative high expression genes; META, genes of different metabolic pathways; RAUU, relative amino acid usage; Arg, Arginie; Asn, Asparaginie; Asp, Aspatic acid; Cys, Cysteine; Gln, Glutamine; Glu, Glutamine acid; Gly, Glycine; His, Histidine; Ile, Isoleucine; Leu, Leucine; Lys, Lysine; Met, Methionine; Phe, Phenylalanine; Pro, Proline; Ser, Serine; Thr, Threonine; Trp, Tryptophan; Tyr, Tyrosine; Val, Valine.

* Corresponding author. Max-Planck Institut for Molecular Genetics, Ihnestr 63-73, 14195 Berlin, Germany.
E-mail address: schaber@molgen.mpg.de (J. Schaber).

0378-1119/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2005.04.003

bacteria have been sequenced to date, all belonging to a common clade, the γ3-proteobacteria: these consist of three genomes of *Buchnera aphidicola*, obligatory symbionts of aphids, which are named after their host species throughout this paper (*Acrythosiphon pisum*, BAp; *Schizaphis graminum*, BSg; *Baizongia pistaciae*, BBp), *Wigglesworthia glossinidia brevipalpis* (Wgl), the primary endosymbiont of tsetse flies, and *Blochmannia floridanus* (Bfl), the primary endosymbiont of carpenter ants.

All five genomes show several signs of deleterious evolution, including accelerated evolutionary rates and extreme AT-enrichment (Moran, 1996; Wernegreen and Moran, 1999; Tamas et al., 2002; Moran, 1996; Tamas et al., 2002; Woolfit and Bromham, 2003). Genome-level analyses have shed light on processes that partly explain the codon and amino acid composition in endosymbionts. The most important findings of previous studies were:

a) The strand specific mutational bias is the dominant source of codon usage variation in *Buchnera* (Rispe et al., 2004) and Bfl (Banerjee et al., 2004).
b) The majority of putatively highly expressed genes are situated on the leading strand in *Buchnera* (Rispe et al., 2004) and Bfl (Banerjee et al., 2004). Thus, a residual effect of replicational or translational selection can be hypothesized.
c) AT-enrichment is mitigated in a fraction of genes thought to be under greater selective constraint. This pattern seems to reflect greater conservation of these genes since divergence from free-living relatives (Herbeck et al., 2003; Rispe et al., 2004; Banerjee et al., 2004). The alternative explanation, i.e. selection against AT rich amino acids, could not be verified until the current study.
d) An effect of selection against the use of costly, aromatic amino acids in important proteins could be found (Palacios and Wernegreen, 2002; Rispe et al., 2004).

From the above, it is clear that while codon usage is largely unbiased in all these five endosymbionts (apart from the global mutational AT-bias, and from strand specific biases), amino acid usage could still be shaped by selectional forces at least in putative high expression genes. One objective of this study was to better weigh the influence of the different factors like GC content, aromaticity, and hydrophobicity on the amino acid composition of genes under high selective constraints, and to better characterize the selective forces (and their direction) that operate at the amino acid level.

Analysis of amino acids substitutions between endosymbionts and an inferred ancestor can shed light on selective processes that shape these genomes, and has been performed for Wgl (Herbeck et al., 2003) and Bfl (Banerjee et al., 2004). However, in these studies, only few genes were used possibly lacking statistical power. This and the fact that this kind of analysis was not done before for *Buchnera* motivated a new study of amino acid substitution patterns between all five endosymbionts and a common ancestor within a common framework.

The recent availability of the Bfl (Gil et al., 2003) and Wgl (Akman et al., 2002) genomes allows a broader comparison of codon usage and amino acid usage across endosymbionts of divergent insect host groups that have distinct nutritional requirements. While aphids (the host of *Buchnera*) feed on plant sap, tsetse flies (host of Wgl) feed on mammalian blood, and *Camponotus* and related ant genera (host of Bfl) are thought to have a more complex diet. Distinct selective pressures relating to host diet are reflected in the gene contents of endosymbionts, which retain abilities to biosynthesize specific nutrients that the host requires (Gil et al., 2003). Whereas *Buchnera* and Bfl provide their hosts primarily with essential amino acids that lack from the host's diet, Wgl provides cofactors to the host to enrich the host's diet with vitamins. Thus, we expect selection pressures to vary among functional categories of endosymbiont genes, particularly among biosynthetic categories that relate to host nutrition. We hypothesize that amino acid biosynthetic genes are more conserved than cofactor biosynthetic genes in Bfl and *Buchnera*, while the opposite would be true for Wgl. Indeed, using relative rates tests, Canback et al. (2004) showed that rate variability among genes in the *Buchnera* and *Wigglesworthia* genomes correlates with host-associated metabolic dependencies. Specifically, they found that host-level selection acting on *Buchnera* and *Wigglesworthia* has slowed both the loss of particular biosynthetic genes (for essential amino acids, or fatty acids and cofactors, respectively) and has slowed acceleration of sequence evolution in these gene categories.

This motivated to extend the abovementioned analysis of differential selective pressures within and among endosymbionts to other gene categories, including metabolic groups of genes.

2. Methods

2.1. Sequences

All sequences were extracted from GenBank. Accession numbers of whole annotated genomes are BA000003, AE013218, AE016826, BA000021, NC_005061, U00096, AL590842 for BAp, BSg, BBp, Wgl, Bfl, *Escherichia coli* K-12 (EK12), and *Yersinia pestis* strain CO92 (Ypo), respectively. The latter two genomes serve for comparison, because they are believed to be shaped by different evolutionary forces (see below). Genes containing less than 50 amino acids and genes with internal stop codons were discarded, but we included hypothetical proteins. Final analyzed data sets included 581, 603, 562, 543, 502, 4249, and 3852 genes for Bfl, Wgl, BAp, BSg, BBp, EK12, and Ypo, respectively. Top scores of a reciprocal blast were used (cut-off 0.0001) (Altschul et al., 1997) to identify putatively homologous sequences between pairs of species. Putatively

homologous amino acid sequences were aligned using T-Coffee (Notredame et al., 2000) with standard parameter settings.

2.2. Gene categories

Genes were categorized by aromaticity, i.e. relative frequency of aromatic amino acids (Phe, Trp, Tyr) per gene, and hydrophobicity, i.e. arithmetic mean of the hydrophobicity index of all amino acids in the respective gene (Kyte and Doolittle, 1982). High and low 'aromatic' and 'hydrophobic' genes were identified by selecting genes in the lower and upper quartiles (25%) of their respective score. In addition, we also analyzed other discrete categories: (a) high (HEG) and low expression genes (LEG), (b) putative high (PHEG) and putative low expression genes (PLEG), and (c) genes of different metabolic pathways (META), particularly genes involved in synthesis of essential amino acids and genes involved in the synthesis of cofactors (Gil et al., 2003, Table S3 of the Supplementary Material). The HEG category was derived from heat stress microarray experiments of BSg (Wilcox et al., 2003). We used expression levels of 530 protein coding genes of the reference sample (nonheat stressed) (GEO sample accession numbers GSM2470-2473). The analysis was done with the *R* statistical computing environment (http://www.r-project.org, Ihaka and Gentleman, 1996). We used the R-package vsn to calibrate and transform the raw spot signal intensities while accounting for dye, slide, and hybridization effects (Huber et al., 2002). Linear regression was used to adjust for the gene length effect on the normalized signal intensities. A single adjusted mean signal per gene was calculated by averaging other both among slide and within slide replicates while equally weighting the contribution of the two dyes. As suggested in Wilcox et al. (2003), slide 2 was considered as outlier and removed from the analysis. We found our results to be consistent to those presented in Fig. 5 of Wilcox et al. (2003). We defined as HEG and LEG the upper and lower 5% quantile of the mean normalized signal intensity ranking, respectively, and identified their corresponding homologous genes in the other species (see Table S1 and S2 of the Supplementary Material). This resulted in 17, 22, 26, 26, and 24 HEG and 18, 17, 25, 26, and 18 LEG for Bfl, Wgl, BAp, BSg, and BBp, respectively.

Because experimental data were only available for one species, we decided to use a second classification of high and low expression genes which is considered to be of a wider applicability, especially when species of different biological background are considered (see Discussion). The PHEG category was defined as in Palacios and Wernegreen (2002) and Herbeck et al. (2003). We included ribosomal proteins and heat shock proteins (GroE/L) as PHEG, resulting in 54 homologous genes for all endosymbionts except Wgl (53 genes). Notably, the HEG and PHEG categories overlapped, as many genes experimentally determined to be high expression were also ribosomal proteins and heat shock proteins (see Table S1 of the Supplementary Material). We defined PLEG as those genes that are in the lower 10% quantile of the distribution of CAI of homologous EK12 genes per species, excluding *ilvH* (Palacios and Wernegreen, 2002). This resulted in an approximately equal number of PLEG and PHEG, i.e. we obtained 58, 59, 54, 52, and 49 PLEG for Bfl, Wgl, BAp, BSg, and BBp, respectively. The maximum CAI of homologous EK12 genes of PLEG across all five species was 0.32 (Table S4 of the Supplementary Material).

2.3. Tests of difference in substitution rates between gene categories

We tested for a significant difference in substitution rates (Ka, Ks) (see below) between genes on different strands, aromatic and nonaromatic genes, hydrophobic and nonhydrophobic genes, and the three gene categories HEG/LEG, PHEG/PLEG, and META mentioned above, using the Wilcoxon rank sum test (*U*-test). Additionally, significantly different relative amino acid usage (RAAU) per amino acid between HEG/LEG and PHEG/PLEG per species was determined by a two-sided permutation test (5000 samples) (Sokal and Rohlf, 1995) (Table 2). For all multiple comparisons, the significance values were Bonferroni-corrected after Dunn-Sidak (Sokal and Rohlf, 1995).

2.4. Substitution rate estimation

We used three different methods to estimate rates of synonymous (Ks) and nonsynonymous substitutions (Ka) of coding sequences of the five genome pairs Bfl–Wgl, BAp–BSg, BAp–BBp, and BSg–BBp, respectively, i.e. Li's method (Li, 1993), using the diverge function from the GCG 10.2 package, Ina's method 1 (Ina, 1995, implementation downloaded from ftp://nig.ac.jp), and a maximum-likelihood method (Goldman and Yang, 1994; Yang and Nielsen, 2000) as implemented in the PAML package (Yang, 1997).

The results for Li's method did not substantially differ from Ina's method. Therefore, we only report results for Ina's method. PAML estimations reached internal implementation thresholds in about 50% of all genes for the BAp–BBp and the BSg–BBp pair and in about 70% of genes for the Bfl–Wgl pair. We concluded that the distances of the analyzed sequences pairs fall out of the application window of PAML (Anisimova et al., 2002, Yang, personal communication). Therefore PAML results were not further considered.

2.5. Comparison of amino acid substitution patterns

While the overall frequency of amino acid changes is known to be lower at PHEG because these genes are more conserved, it remains unclear whether PHEG and PLEG have the same configuration of changes (Herbeck et al.,

2003; Banerjee et al., 2004). To more clearly distinguish the effects of conservation and selection on particular amino acids, we extended the previous analyses to include additional loci, other endosymbiont genomes, and amino acids that are AT-rich in EK12 as well and for HEG and LEG. We used the EK12 genome as proxy for ancestral sequences, based on its close phylogenetic position to endosymbionts and its relatively slow evolutionary rate (Tamas et al., 2002; Gil et al., 2003).

We distinguished changes among three amino acid categories: those encoded by GC-rich, unbiased, and AT-rich codons (classified as Herbeck et al., 2003), and compared substitution configurations between PHEG and PLEG with the G-test adjusted by the Williams correction (Sokal and Rohlf, 1995).

3. Results

3.1. Different amino acid usage between expression levels

Investigating the specific amino acid differences between PHEG/PLEG as well as for HEG/LEG categories, we tested the significance of the difference in RAAU by a permutation test for each amino acid (Table 1). The results supported and extended those of other authors for *Buchnera* (Palacios and Wernegreen, 2002; Rispe et al., 2004), Wgl (Herbeck et al., 2003), and Bfl (Banerjee et al., 2004). The general pattern was the same using PHEG/PLEG or HEG/LEG categories, with one exception. In PHEG, Lys was overrepresented in all endosymbionts, whereas in HEG, it was underrepresented or there was no significant difference (results not shown). We will only report results for the PHEG/PLEG category in the following.

Table 1
Significance of the difference of RAAU between putative high (PHEG) and low (PLEG) expression genes for each amino acid and bacterial species considered: ++/+|−−/−: significantly higher|lower RAAU in PHEG compared to PLEG ($P<0.01/0.05$)

AA	Bfl	Wgl	BAp	BSg	BBp	EK12	Ypo
Ala	++	++	++	++	++	++	++
Arg	++	++	++	++	++	−	ns
Asn	ns	−−	−−	−−	−−	−−	+
Asp	ns	ns	ns	ns	ns	++	++
Cys	−−	−−	−−	−−	−−	−−	−−
Gln	−−	ns	ns	ns	ns	ns	−−
Glu	++	++	++	++	++	++	++
Gly	++	++	++	++	++	++	++
His	ns	ns	ns	ns	ns	−−	−−
Ile	−−	−−	−−	−−	−−	−−	−
Leu	−−	−−	−−	−−	−−	−−	−−
Lys	++	++	+	+	++	++	++
Met	ns	ns	ns	ns	ns	ns	ns
Phe	−−	−−	−−	−−	−−	−−	ns
Pro	ns	ns	ns	ns	ns	ns	−−
Ser	ns	−	ns	ns	ns	−−	−−
Thr	−−	+	ns	ns	−	ns	++
Trp	−−	−−	−−	−−	−−	−−	−−
Tyr	−−	−−	−−	−−	−−	−−	ns
Val	++	++	++	++	++	++	++

ns: not significant.

Differences in amino acid usage between PHEG and PLEG were very similar among endosymbionts and between endosymbionts and two related free-living bacteria (Table 1). Costly aromatic amino acids and AT-rich codons were underrepresented in PHEG compared to PLEG (Herbeck et al., 2003; Rispe et al., 2004; Banerjee et al., 2004). However, there were slight differences to a former study, probably due to the usage of different sets of genes. For example, the essential, high energy, and hydrophobic amino acids Phe and Ile were found to be significantly reduced in our study but showed no significantly different frequency between high and low expression genes in Banerjee et al. (2004).

However, under-representation of aromatic and AT rich amino acids did not completely explain the observed patterns. For example, Cys was avoided in PHEG, but is neither aromatic, nor is it AT-rich (50% GC). On the other hand, Lys is AT-rich (17% GC) but was overrepresented in PHEG. This phenomenon was most readily explained by conservation as the related free-living bacteria EK12 and Ypo showed the same pattern (Table 1).

To weigh the influences of GC content, aromaticity, and hydrophobicity of amino acids on differential amino acid usage between PHEG and PLEG, we stepwise fitted a linear model to our data:

$$\mathrm{d}F = a + C_{\mathrm{GC}} + H + M + e,$$

where dF is the difference of the relative frequency of an amino acid between PHEG and PLEG, a is a constant, C_{GC} is the average GC content of the respective amino acid (number of G and C nucleotides over all codons coding the respective amino acid divided by the number all of nucleotides in these codons), H is the hydropathy index of the respective amino acid (Kyte and Doolittle, 1982), M is a proxy for the metabolic cost of the respective amino acid (Akashi and Gojobori, 2002), and e is an error term.

In Table 2, we list results of the regression of the respective single parameter model and the results of the stepwise regression where in each step the parameter with the largest explanatory power was added. dF was pooled over all amino acids and endosymbionts.

The influence of the considered parameters on differential amino acid usage between PHEG and PLEG decreased in the order C_{GC}, H, and M. The ensemble of all linearly combined parameters explained about 37% of the variation of dF for all endosymbionts.

3.2. Substitution rates

The majority of the Ks estimations were saturated, i.e. median Ks>1 except for the BAp–BSg pair. However, the majority of the Ks estimates was well below two, and the median standard error was below 20% of the corresponding

Table 2
Regression analysis of the influence of various parameters P of the difference of relative frequencies of amino acids between PHEG and PLEG (dF)

P	One parameter		Stepwise regression	
	Adj. R^2	P	Cum. adj. R^2	P
C_{GC}	0.241	**	0.241	**
H	0.170	**	0.350	**
M	0.105	**	0.368	< 0.1

One parameter model: $dF = a + P + e$, where P is one of the following parameters: C_{GC}, GC-richness; H, hydrophobicity; M, metabolic cost. Adj. R^2: adjusted coefficient of determination. Stepwise regression: in each step the parameter that added most to the explanatory paper of the model was added to the model. Cum. adj. R^2: cumulative coefficient of determination. All parameters (including the intercept a) are highly significant (**: $P<0.01$).

median Ks rate. Therefore, we considered most Ks estimates to be in an acceptable range. Most Ka estimates were below 0.5 and more reliable. The median standard error of Ka estimations was around 10% (Table 3).

Ka and Ks calculated between Bfl–Wgl and the three *Buchnera* pairs (BAp–BSg, BAp–BBp, BSg–BBp) showed different density distributions for PHEG and PLEG (Fig. 1) as well as for HEG/LEG. Ka was always significantly different ($P<0.01$) between putative and nonputative expression levels, respectively. Significant differences in Ks were found for the *Buchnera* pairs (except BAp–BBp using HEG/LEG) but not for Bfl–Wgl (Table 4).

All endosymbiont pairs showed significant Ka differences between aromatic and nonaromatic genes (Table 4).

There was no effect of gene orientation on substitution rates for Bfl and Wgl. For the *Buchnera*, however, a significant effect for Ks estimates was observed (Table 4). To test whether this was an artifact of the fact that PHEG are mostly situated on the leading strand, we tested whether there was a difference in substitution rates between leading and the lagging strand for all expression levels separately (results not shown). Ks rates were significantly lower in the leading strand than in the lagging strand irrespective of the expression level.

Concerning hydrophobicity, a weakly significant ($P<0.05$) difference for Ka estimates between hydrophobic and nonhydrophobic genes for Bfl was found. All other investigated species showed no correlation between substitution rates and hydrophobicity.

Endosymbionts showed no difference in substitution rates between genes coding for essential amino acids and genes coding for cofactors (Table 4).

In sum, the endosymbionts showed strong differential selection pressures at nonsynonymous sites between PHEG and PLEG but also synonymous sites evolved at different rates. Because PHEG are mostly situated on the leading strand, are nonaromatic and hydrophilic (Palacios and Wernegreen, 2002; Herbeck et al., 2003; Rispe et al., 2004; Banerjee et al., 2004) residual effects of differential evolutionary pressures could be detected within those categories as well. The strand orientation as such also seems to affect mutation rates.

3.3. Amino acid substitution patterns

To further study the processes that lead to the overrepresentation of GC-rich amino acids in PHEG, we compared homologous amino acid sites between endosymbionts and EK12 of PHEG and PLEG, respectively. In Table 5, we show results of an analysis analogous to Herbeck et al. (2003).

As expected for strongly AT-enriched genomes, the most frequent mutational events were substitutions from GC-rich to AT-rich residues, both in PLEG and PHEG. However, we found significant deviations of the relative frequencies of GC-rich amino acid substitutions between PHEG and PLEG. In Bfl, substitutions of GC-rich to GC-rich amino acids were significantly higher in PHEG than PLEG, and substitutions from GC-rich to AT-rich were significantly reduced in PHEG compared to PLEG ($P<0.05$). In Wgl, substitutions of GC-rich to unbiased amino acids were significantly higher in PHEG, and substitutions of GC-rich to AT-rich amino acids were significantly decreased ($P<0.01$). These findings contrast with those reported by Banerjee et al. (2004) and Herbeck et al. (2003). *Buchnera* showed the same pattern as Wgl. Substitutions at amino acid residues that were AT-rich in EK12 also showed different configurations between PHEG and PLEG. Relative frequencies of AT-rich to GC-rich amino acid substitution were significantly higher in PHEG for Bfl, Wgl, and BBp. AT-rich to AT-rich substitutions were decreased in PHEG in all endosymbionts and AT-rich to unbiased substitutions were increased in PHEG in all endosymbionts except BBp. Substitution patterns for unbiased amino acids were not

Table 3
Number of used aligned sequence pairs (#) and quartiles (Q) of the resulting estimation distribution of synonymous and nonsynonymous (Ina's method) substitution for the considered species pairs

	25% Q	Med±Med of S.E.	75% Q	#[a]
BAp–BSg				
Ks	0.584	0.658±0.09	0.731	525
Ka	0.116	0.171±0.02	0.226	526
BAp–BBp				
Ks	1.056	1.202±0.21	1.463	448
Ka	0.251	0.333±0.03	0.418	477
BSg–BBp				
Ks	1.030	1.198±0.200	1.406	426
Ka	0.156	0.337±0.03	0.294	462
Bfl–Wgl				
Ks	1.127	1.284±0.23	1.552	375
Ka	0.312	0.416±0.03	0.507	415

The median (Med) values and their corresponding median of standard errors (S.E.) are given.

[a] In some cases, estimates are not available because internal thresholds were reached, so that different numbers of sequence pairs are used to estimate synonymous and nonsynonymous substitution rates, respectively.

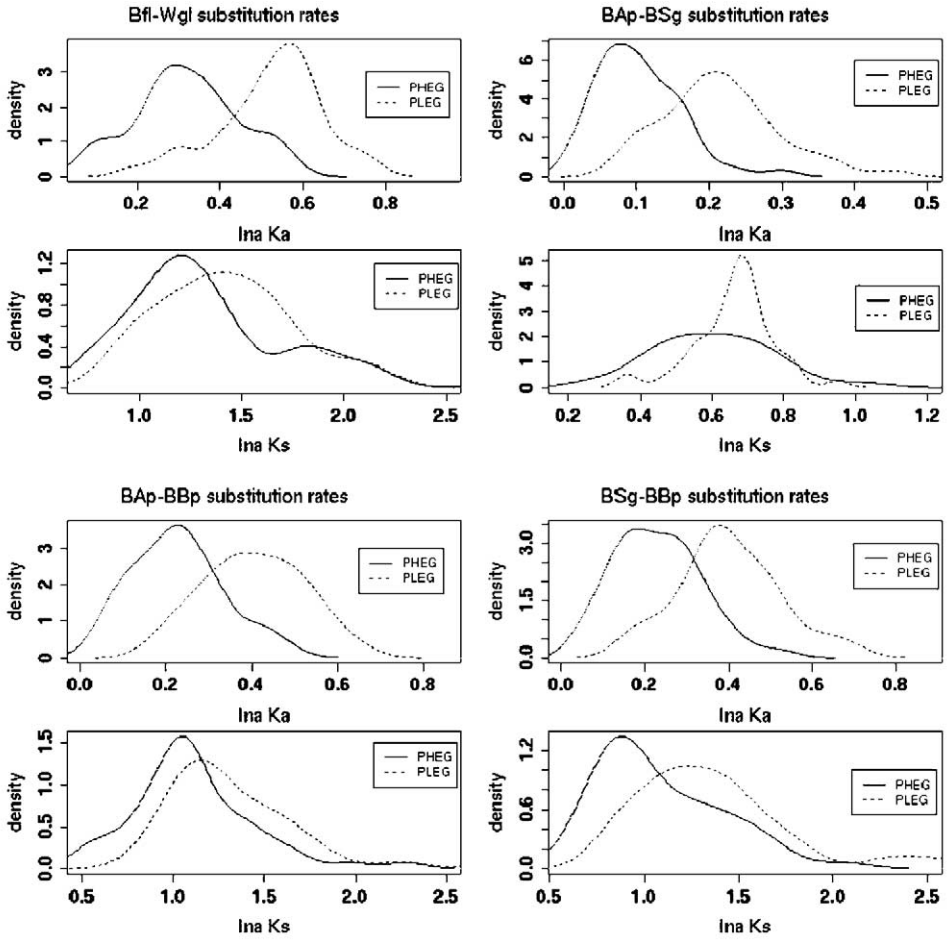

Fig. 1. Estimated density distributions (Gaussian kernel) of Ka and Ks substitution rates (Ina's method) for putative high expression genes (PHEG) and putative low expression genes (PLEG).

significantly different between PHEG and PLEG (results not shown). As expected from the overall reduction in amino acid substitutions at PHEG, these genes were clearly more conserved on the amino acid level compared to PLEG (Table 5), as is further demonstrated below. For HEG/LEG category, the general pattern was similar (results not shown).

Table 4
Significant differences in substitution rates (Ka and Ks, Ina's method) between several discrete gene categories using the Wilcoxon rank sum test

Sequences	HEG/LEG	PHEG/PLEG	AROMA	Strand	HYDRO	META
Bfl–Wgl	Ka**	Ka**	Ka**	–	Ka*	–
BAp–BSg	Ka**, Ks**	Ka**, Ks*	Ka**	Ks**	–	–
BAp–BBp	Ka**, Ks*	Ka**, Ks**	Ka**	Ks**	–	–
BSg–BBp	Ka**	Ka**, Ks**	Ka**	Ks*	–	–

HEG/LEG: high versus low expression genes, PHEG/PLEG: putative high versus putative low expression genes, AROMA: aromatic versus nonaromatic genes (see Method section), Strand: leading versus lagging strand, HYDRO: hydrophobic versus nonhydrophobic genes (see Methods), META: gene coding for essential amino acids versus gene codon for cofactors (see Table S1 of the Supplementary Material). **: $P<0.01$, *: $P<0.05$.

4. Discussion

This study represents a comprehensive comparison of amino acid usage and evolutionary rates across endosymbiont genomes. We included multiple genomes to detect similarities and differences among bacterial species that inhabit quite distinct insect hosts, and that have evolved independently from each other. Analyzing all available complete insect endosymbiont genomes reveals new patterns of differential amino acid usage as well as substitution rates among and within endosymbiont genomes and further elucidates evolutionary forces that have shaped them.

Classifying ribosomal proteins (together with the two chaperones mopA and mopB) as putative high expression genes and the genes determined from expression experiments as (nonputative) high expression genes might be misleading, because assuming the same expression ranks

Table 5
Pair wise uncorrected number of amino acid changes for homologous positions for different amino acid categories and endosymbionts

Type of AA change		Bfl			Wgl			BAp	BSg	BBp
EK12	Endo	PHEG	PLEG	*P*	PHEG	PLEG	*P*	*P*	*P*	*P*
GC-rich	GC-rich	74	138	+	44	156				
GC-rich	AT-rich	350	1106	–	465	1896	––	––	––	––
GC-rich	Unbiased	424	1054		399	1142	++	++	++	++
Expected frequency		0.27	0.73		0.22	0.78				
AT-rich	GC-rich	77	136	++	35	115	+			++
AT-rich	AT-rich	229	958	––	283	1701	–	––	–	–
AT-rich	Unbiased	273	715	+	163	716	+	++	++	
Expected frequency		0.24	0.76		0.16	0.84				
% conserved AA		65	46		65	42		74/47	73/47	69/45

PHEG: putative high expression genes, PLEG: putative low expression genes. *P*: *P*-values of G-test (+/++|––/–: significantly higher|lower frequency in PHEG ($P<0.05/0.01$) compared to PLEG in the same category). Absolute values are only displayed for Bfl and Wgl.

for all species is also a putative approach. However, we maintained this nomenclature because it is used in the literature. Incorporating more or less ranks in the classification of HEG/LEG did not change results. Given the differences in endosymbiont metabolisms and host requirements, we feel that in this study the PHEG/PLEG categories are of a more general applicability. Thus, we concentrate on the PHEG/PLEG category in the following.

Specific nutritional needs of the insect host have clearly affected gene content of the bacterial partner. Bfl and *Buchnera* provide their hosts with essential amino acids, whereas Wgl provisions vitamins and cofactors (Shigenobu et al., 2000; Akman et al., 2002; Gil et al., 2003). As expected, these small endosymbionts genomes retain the specific biosynthetic functions that the host requires. The ratio of number of genes coding for essential amino acids to the number of genes coding for cofactors is 46/22 for Bfl and 40/19 for *Buchnera*, but is 12/41 in Wgl, which obtains many amino acids from its host. Given the clear effect of host-level selection on gene content, we tested its effects on protein divergences. Specifically, we hypothesized that amino acid biosynthetic genes would be more conserved than cofactor biosynthetic genes in Bfl and *Buchnera*, while the opposite would be true for Wgl. Canback et al. (2004) found this to be true for relative rate comparisons between *Buchnera* and *Wigglesworthia*. However, we found no differences in codon usage (results not shown), amino acid usage, nor in substitution rates between the different classes of these biosynthetic functions. Our results are therefore contrary to our expectation and, at first sight, do not support the findings of Canback et al. (2004). One possible reason for our different results is that Canback et al. (2004) compared the lengths of branches from the ancestral node of BSg and Wgl to each of these endosymbiont taxa. These deep branches included the initial changes that occurred upon establishment of each endosymbiosis. In *Buchnera*, these early changes involved severe gene loss (Moran and Mira, 2001) and an exceptionally strong effect of AT bias on amino acid changes (Clark et al., 1999). In our rate comparisons, we focused on sequence changes that occurred after the divergence of *Buchnera* lineages, and after the divergence of the closely-related Wgl and Bfl. We found no differences in rates among functional categories since these more recent divergences. Thus, our results suggest that some of the rate differences that Canback et al. observed reflect differences in the early evolution of both *Buchnera* and Wgl. The question of when endosymbionts experienced differential rates among functional categories is a question that deserves further investigation. Additional complete genomes for closely-related endosymbionts would help to address this question.

Previous studies attributed relative GC-richness of high expression genes to stronger conservation of high expression genes of a presumably GC-rich ancestor and avoidance of aromatic amino acids (Palacios and Wernegreen, 2002; Herbeck et al., 2003; Rispe et al., 2004; Banerjee et al., 2004). In this study, we confirmed a greater conservation of high expression genes by (a) demonstrating their higher percentage of conserved homologous amino acids (Table 5), likely a sign of functional conservation, and (b) reduced nonsynonymous substitution rates (Table 4) compared to not highly expressed genes. Avoidance of costly amino acids is also confirmed to contribute significantly to shaping high expression genes amino acid content even though it is of minor importance (Table 2). The most prominent factor shaping amino acid usage differences in high expression genes is indeed preferred usage of GC-rich amino acids and secondly avoidance of hydrophobic amino acids (Table 2).

As GC content is the most significant factor shaping amino acid usage differences in each endosymbiont genome, this pattern may be caused by overall conservation of PHEG since their divergence from a relatively GC-rich ancestor, as well as active selection in endosymbionts to maintain GC-rich (and/or avoid AT-rich) amino acids at PHEG. Herbeck et al. (2003) and Banerjee et al. (2004) explored these alternatives by comparing the configuration of amino acid changes at PHEG and PLEG genes. Their analyses did not reject the null hypothesis of identical configurations of amino acid changes, and thus did not rule out the possibility that GC-richness of PHEG is an artifact of the last free-living ancestor and reflects overall

conservation of these proteins in endosymbionts. Herbeck et al. (2003) and Banerjee et al. (2004) used 10 or 38 genes, respectively, for both PHEG and PLEG, possibly lacking statistical power. Based on a larger sample of genes, we find significantly different substitution patterns of PHEG and PLEG of endosymbionts, thus rejecting the null hypothesis that PHEG and PLEG show the same relative substitution rates among GC-rich, AT-rich, and unbiased amino acids. The specific differences suggest selection to specifically maintain GC-rich and/or avoid AT-rich amino acids in PHEG of endosymbionts (Table 5).

Consideration of the biochemical properties of amino acids may help to explain the above pattern. Amino acid "neighbors" in the genetic code (differing by one base, e.g. Gly and Ala) are often biochemically similar. Thus, changes from a GC-rich amino acid to another GC-rich (or unbiased) amino acid often correspond to conservative changes between amino acids that are biochemically close. The selective consequences of those changes will be lower than changes from GC-rich to AT-rich amino acids, the latter of which may be more constrained PHEG. In addition, a switch from GC-rich to AT-rich amino acids requires more nucleotide changes. PHEG with fewer nonsynonymous substitutions (Table 5, Fig. 1) will have fewer radical amino acid changes such as this. Thus, a higher relative substitution frequency from GC-rich amino acids to GC-rich or unbiased amino acids is not necessarily an indication of selection on GC content per se. Rather, it may also reflect an overall greater conservation of PHEG. Examining only sites that are ancestrally GC-rich (here, GC-rich amino acids in *E. coli*) does not distinguish the two hypotheses. However, the process above would predict the same pattern for AT-rich amino acids in EK12. That is, we would expect PHEG to show higher relative frequencies of conservative changes between AT-rich to AT-rich or unbiased amino acids, and lower frequencies of radical changes from AT-rich to GC-rich amino acids. Notably, AT-rich amino acids in EK12 show exactly the opposite pattern. PHEG of Bfl, Wgl, and BBp show higher AT-rich to GC-rich substitutions and lower AT-rich to AT-rich substitutions (Table 5). This pattern is inconsistent with selection against radical changes in PHEG, and better supports the hypothesis of selection against AT-rich amino acids in these proteins.

The phenomenon of accelerated substitution rates in endosymbionts compared to free-living bacteria is well-documented (Moran, 1996; Tamas et al., 2002). However, reduced substitution rates (and higher GC content) at PHEG compared to PLEG suggest that selection may counteract the genome wide AT mutational bias. Our data clearly support reduced rates of nonsynonymous substitutions at high expression genes in endosymbionts (Table 5, Fig. 1). Moreover, synonymous substitution rates (Ks) are also reduced in high expression genes compared to low expression genes in *Buchnera* (Table 5). This is a novel result. It is generally believed that greater conservation of high expression genes at synonymous sites suggests selection on codon usage, a force previously thought to be absent or ineffective in endosymbiont genomes. Indeed, the high and almost uniform AT mutational bias within and among endosymbionts, and results from correspondence analyses of codon usage where no effect of expression level on codon usage could be found (Rispe et al., 2004, own analysis) contradict the notion of selection on codon usage. Alternatively, there might be a general mutational variability across the genome, irrespective of the strand, with a tendency of the important genes located in those areas, or tandem mutations that are negatively selected in important genes might also influence the synonymous substitution rate. However, differential synonymous substitution rates between strands of replication irrespective of the expression level indicate that strand orientation itself might influence AT mutational bias towards higher conservation and thus higher GC content of genes on the leading strand. One possible explanation is that genes on the leading strand are generally more conserved than genes on the lagging strand. However, it is not clear why for the majority of genes this is only true for synonymous and not for nonsynonymous substitution rates.

The estimated substitution rates should be interpreted cautiously. We used these values to compare rates among genes, with the understanding that their absolute values may be unreliable. Generally, Ina's method is not suitable for genomes with high compositional bias because it does not take this bias into account. Unfortunately, the method that does take this bias into account, i.e. PAML, relies on a large number of sequence comparisons due to its probabilistic approach (Yang, 1998; Anisimova et al., 2002), while our analysis of individual genes was based pairs of sequences. Moreover, a sensitivity analysis showed that good PAML estimates can only be expected when the distance between the considered sequences does not exceed a certain limit (roughly 10 or 50 nucleotide substitutions per nucleotide site along the tree (PAML FAQ http://abacus.gene.ucl.ac.uk/software/pamlFAQs.html, Anisimova et al., 2002), which is not the case neither for the analyzed endosymbionts nor for the distance between endosymbionts and *E. coli*, as used in a recent study (Banerjee et al., 2004). Here additional complete genomes for closely-related endosymbionts would greatly improve the reliability substitution rate estimations. Although estimates of the Ina method can be expected to be biased, comparisons of their values among genes and sequence pairs show consistent trends.

5. Conclusion

Despite reduced efficacy of selection in small endosymbiont populations, this study shows that different forces influence genome composition in these symbionts. While AT mutational bias is very strong in *Buchnera*, *Wigglesworthia*, and *Blochmannia*, such that codon usage is very uniform across all these genomes, slight differences at the codon level are still detectable between different

categories of genes. We have found indeed that genes that are characterized by higher levels of expression have lower synonymous rates of evolution than genes with low or average levels of expression. We found significant differences in profiles of amino acid usage between high and low expression genes in the five genomes, that interestingly almost always go in the same direction, with similar sets of residues preferred or avoided in each category of genes. We have also shown on a large set of sequences that highly expressed genes are in fact more conserved at the amino acid level, retaining therefore a composition closer the more GC-rich ancestral state. But in addition, we analyzed the types of amino acid substitutions in the different symbionts and found systematically the same following trend, i.e. an excess of GC rich to GC rich or unbiased substitutions, suggesting some degree of selection against AT-rich amino acids. In sum, these results indicate that endosymbiont genomes are not completely shaped by deleterious mutation and drift, but rather show signatures of selection at various levels.

Acknowledgments

J.S. is funded by the Marie Curie Host Fellowship of the European Union Nr. MCFI-1999-01055. The project has been supported by grants from the US National Institutes of Health (R01 GM62626-01), the US National Science Foundation (DEB 0089455), and the NASA Astrobiology Institute (NCC2-1054) to J.W. and by grants BMC2003-00305 from MiCyT (Spain) and Grupos03/204 (Generalitat Valenciana, Spain) to A.M.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.gene.2005.04.003.

References

Akashi, H., Gojobori, T., 2002. Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. Proc. Natl. Acad. Sci. 99, 3695–3700.

Akman, L., et al., 2002. Genome sequence of the endocellular obligate symbiont of tsetse flies, *Wigglesworthia glossinidia*. Nat. Genet. 32, 402–407.

Altschul, S.F., et al., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Anisimova, M., Bielawski, J.P., Yang, Z., 2002. Accuracy and power of bayes prediction of amino acid sites under positive selection. Mol. Biol. Evol. 19, 950–958.

Banerjee, T., Basak, S., Gupta, S.K., Gosh, T.C., 2004. Evolutionary forces in shaping the codon and amino acid usages in *Blochmannia floridanus*. J. Biomol. Struct. Dyn. 22, 1–11.

Canback, B., Tamas, I., Andersson, S.G., 2004. A phylogenomic study of endosymbiotic bacteria. Mol. Biol. Evol. 21 (6), 1110–1122 (June).

Clark, M.A., Moran, N.A., Baumann, P., 1999. Sequence evolution in bacterial endosymbionts having extreme base compositions. Mol. Biol. Evol. 16, 1586–1598.

Gil, R., et al., 2003. The genome sequence of *Blochmannia floridanus*: comparative analysis of reduced genomes. Proc. Natl. Acad. Sci. 100, 9388–9393.

Goldman, N., Yang, Z.H., 1994. Codon-based model of nucleotide substitution for protein-coding DNA-sequences. Mol. Biol. Evol. 11, 725–736.

Herbeck, J.T., Wall, D.P., Wernegreen, J.J., 2003. Gene expression level influences amino acid usage, but not codon usage, in the tsetse fly endosymbiont *Wigglesworthia*. Microbiology 149, 2585–2596.

Huber, W., Von heydebreck, A., Sultmann, H., Poustka, A., Vingron, M., 2002. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics 18 (Suppl. 1), S96–S104.

Ihaka, I., Gentleman, R., 1996. R: a language for data analysis and graphics. J. Comput. Graph. Stat. 5, 299–314.

Ina, Y., 1995. New methods for estimating the numbers of synonymous and nonsynonymous substitutions. J. Mol. Evol. 40, 190–226.

Kyte, J., Doolittle, R.F., 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157, 105–132.

Li, W.H., 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J. Mol. Evol. 36, 96–99.

Moran, N.A., 1996. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. Proc. Natl. Acad. Sci. 93, 2873–2878.

Moran, N.A., Mira, A., 2001. The process of genome shrinkage in the obligate symbiont *Buchnera aphidicola*. Genome Biol. 2 (RESEARCH0054).

Notredame, C., Higgins, D.G., HeringA, J., 2000. T-Coffee: a novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. 302, 205–217.

Ohta, T., 1987. Very slightly deleterious mutations and the molecular clock. J. Mol. Evol. 26, 1–6.

Ohta, T., 1992. The nearly neutral theory of molecular evolution. Ann. Rev. Ecolog. Syst. 23, 263–286.

Palacios, C., Wernegreen, J.J., 2002. A strong effect of AT mutational bias on amino acid usage in *Buchnera* is mitigated at high-expression genes. Mol. Biol. Evol. 19, 1575–1584.

Rispe, C., Delmotte, F., Van ham, R.C., Moya, A., 2004. Mutational and selective pressures on codon and amino acid usage in *Buchnera*, endosymbiotic bacteria of aphids. Genome Res. 14, 44–53.

Shigenobu, S., Watanabe, H., Hattori, M., Sakaki, Y., IshikawA, H., 2000. Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp APS. Nature 407, 81–86.

Sokal, R.R., Rohlf, F.J., 1995. Biometry. Freeman, New York.

Tamas, I., et al., 2002. 50 million years of genomic stasis in endosymbiotic bacteria. Science 296, 2376–2379.

Wernegreen, J.J., Moran, N.A., 1999. Evidence for genetic drift in endosymbionts (*Buchnera*): analyses of protein-coding genes. Mol. Biol. Evol. 16, 83–97.

Wilcox, J.L., Dunbar, H.E., Wolfinger, R.D., Moran, N.A., 2003. Consequences of reductive evolution for gene expression in an obligate endosymbiont. Mol. Microbiol. 48, 1491–1500.

Woolfit, M., Bromham, L., 2003. Increased rates of sequence evolution in endosymbiotic bacteria and fungi with small effective population sizes. Mol. Biol. Evol. 20, 1545–1555.

Yang, Z.H., 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13, 555–556.

Yang, Z.H., 1998. On the best evolutionary rate for phylogenetic analysis. Syst. Biol. 47, 125–133.

Yang, Z.H., Nielsen, R., 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol. Biol. Evol. 17, 32–43.