Codon and Amino Acid Usage in Two Major Human Pathogens of Genus Bartonella — Optimization Between Replicational-Transcriptional Selection, Translational Control and Cost Minimization

Sabyasachi Das, Sandip Paul, Sanjib Chatterjee, and Chitra Dutta*

Bioinformatics Centre, Indian Institute of Chemical Biology, Kolkata, 700032, India
(Received 19 October 2004; revised 20 November 2004)

Abstract

Intra-genomic variation in synonymous codon and amino acid usage in two human pathogens Bartonella henselae and B. quintana has been carried out through multivariate analysis. Asymmetric mutational bias, coupled with replicational-transcriptional selection, has been identified as the prime selection force behind synonymous codon selection — a characteristic of the genus Bartonella, not exhibited by any other alpha-proteobacterial genome. Distinct codon usage patterns and low synonymous divergence values between orthologous sequences of highly expressed genes from the two Bartonella species indicate that there exists a residual intra-strand synonymous codon bias in the highly expressed genes, possibly operating at the level of translation. In the case of amino acid usage, the mean hydropathy level and aromaticity are the major sources of variation, both having nearly equal impact, while strand-specific mutational pressure and gene expressivity strongly influence the inter-strand variations. In both species under study, the highly expressed gene products tend not to contain heavy and/or aromatic residues, following the cost-minimization hypothesis in spite of their intracellular lifestyle. The codon and amino acid usage in these two human pathogens are, therefore, consequences of a complex balance between replicational-transcriptional selection, translational control, protein hydropathy and cost minimization.

Key words: synonymous codon usage; relative amino acid usage; correspondence analysis; strand asymmetry; gene expression

1. Introduction

The complete genome sequences of two human pathogens, Bartonella henselae and B. quintana, have recently been reported.1 Both of these facultative intracellular bacteria are transmitted by insect vectors using mammalian reservoirs, and infect endothelial cells and erythrocytes, causing vasculoproliferative changes in immunocompromised hosts.1,2 B. quintana uses only humans as a reservoir, but B. henselae infects both humans and cats. Genome comparison of the two organisms1 revealed a high degree of overall similarity with major differences being that B. henselae contains the genomic islands coding for filamentous hemagglutinin, and B. quintana exhibits evidence of extensive genome reduction, similar to that observed in Rickettsia prowazekii.3

B. henselae and B. quintana both exhibit strand-specific mutation biases with a large excess of G and T on the leading strands,1 and both are characterized by low genomic coding content (72.7% and 72.3%, respectively) a typical characteristic of Rickettsia and Bartonella the two α-proteobacteria lineages following vector-borne intracellular lifestyles.1,3,4 Strong compositional asymmetries between the genes lying on the leading versus lagging strands were observed earlier in many other prokaryotic organisms at the level of nucleotides, codons and even in amino acids.5–7 For some species, the bias is so high that the sole knowledge of a protein sequence often allows one to predict whether the gene is transcribed from leading or lagging strand.5 For example, in the spirochaetes Borrelia burgdorferi and Treponema pallidum, the strand-specific mutational bias significantly differentiates codon and amino acid usage between the genes encoded on the two strands.8 There is no evidence of any influence of translational selection on synonymous codon usage in highly expressed genes in these two species and asymmetrical replication, coupled with tran-
Table 1. Non-parametric tests of association between first axis of COA and GC$_3$ and GT$_3$ content of corresponding genes in alpha-proteobacterial genomes under study.

| Taxonomic position (Order) | Organism                  | Accession No. | Variation explained by Axis 1 of COA on RSCU (%) | GC$_3$ | GT$_3$ |
|---------------------------|---------------------------|---------------|-------------------------------------------------|--------|--------|
| Caulobacterales           | *Caulobacter crescentus*  | AE005673      | -0.92                                           | 0.25   |        |
| Rhizobiales               | *Agrobacterium tumefaciens* | AE008688      | -0.78                                           | 0.23   |        |
|                           | *Bartonella henselae*     | BX897699      | 0.21                                            | -0.92  |        |
|                           | *Bartonella quintana*     | BX897700      | 0.17                                            | -0.93  | -0.24  |
|                           | *Bradyrhizobium japonicum* | BA000040      | 0.95                                            | -0.45  |        |
|                           | *Brucella melitensis*     | AE008917      | -0.82                                           | 0.08   |        |
|                           | *Brucella suis*           | AE014291      | -0.81                                           | 0.12   |        |
|                           | *Mesorhizobium loti*      | BA000012      | -0.92                                           | 0.26   |        |
|                           | *Rhodopseudomonas palustris* | BX571963     | 0.90                                            | -0.16  |        |
|                           | *Sinorhizobium meliloti*  | AL591688      | -0.83                                           | 0.23   |        |
| Rickettsiales             | *Rickettsia conorii*      | AE006914      | -0.32                                           | -0.14  |        |
|                           | *Rickettsia prowazekii*   | AJ235269      | -0.24                                           | -0.12  |        |
|                           | *Rickettsia typhi*        | AE017197      | 0.15                                            | 0.07   |        |
|                           | *Wolbachia pipiensis*     | AE017196      | 0.28                                            | -0.25  |        |

Notable significant relationships at $p<0.0001$ are indicated in bold type.

scriptional selection, is the major source of codon usage variation. In the endosymbiotic bacteria *Blochmannia floridanus* too, the replicational-transactional selection is the major factor governing the synonymous codon usage pattern, while the conservation of GC-rich amino acids in the highly expressed genes and the hydrophobicity of the gene products are the major sources of variation in amino acid usage.

However, the keto (GT) versus amino (AC) bias may not be the only governing factor in codon usage in the genomes exhibiting strong strand-specific mutational bias. In *Chlamydia trachomatis*, synonymous codon usage seems to be the result of a complex balance among different factors like replicational-transactional selection, selection for translation, the hydropathy of each encoded protein and the degree of amino acid conservation. In *Buchnera*, the endosymbiotic bacteria of aphids, synonymous codon selection is not only influenced by the strand asymmetry, but there is also a slight residual codon bias within each strand. Amino acid usage in *Buchnera* is also strongly biased in putative high-expression genes, characterized by low levels of aromatic amino acids and by greater conservation and resistance to AT enrichment. From the analysis of 15 bacterial chromosomes, Rocha et al. has shown that though the leading strands are, in general, biased towards codons starting or ending with G or T, not all such codons are selected for and some codons are chosen universally. It is, therefore, intriguing to examine how, in *B. henselae* and *B. quintana*, the strand-specific mutational bias could influence the synonymous codon usage patterns. It is also worth investigating whether the strand-specific nucleotide composition and/or the intracellular lifestyles could influence the amino acid selection in the proteins encoded by these two genomes. Multivariate analyses of relative synonymous codon usage and amino acid usage indicated that the codon and amino acid usage in these two human pathogens might be an outcome of a complex balance between replicational-transactional selection, translational control and other physicochemical properties of the gene products.

2. Materials and Methods

The complete genome sequences of *B. henselae*, *B. quintana* and 12 other fully sequenced alpha-proteobacteria (listed in Table 1) have been downloaded from NCBI GenBank. To minimize sampling error, we have taken only those open reading frames (ORFs), which are greater than or equal to 300 bp and have correct initiation and termination codons. The presumed duplicates, transposed genes and genes with
3. Results and Discussion

3.1. Variation in strand-specific compositional bias at three-codon positions

The GC-skew analysis shows a clear shift at two points, around 1931 kb and 975 kb in the *B. henselae* chromosome and around 1581 kb and 723 kb in the chromosome of *B. quintana*, suggesting that the putative origin and termination point of the replication of the chromosomes might be located in these regions. Comparisons with the *Escherichia coli* consensus sequence for the DnaA box (TTATCCACA)\(^\text{18}\) also reveal that a non-translatable region at 1,930,851–1,930,964 bp in *B. henselae* and 1,581,187–1,581,300 bp in *B. quintana* possess clusters of three putative DnaA boxes, thereby confirming the possible locations of the functional chromosomal origins between two genes, BH16700 (rho) and BH000110 for *B. henselae* and between BQ13350 (rho) and BQ60010 in *B. quintana*. As already observed in other α-proteobacteria,\(^\text{18}\) however, a translocation of the *dnaA* gene is observed from the putative oriC to 158 kb in *B. henselae* and to 146 kb in *B. quintana*.

Since the organization of genes usually remains conserved at the region of replication origin,\(^\text{22}\) additional evidence for the location of the putative origins of replication in the *Bartonella* species under study came from a comparison with a conserved gene cluster around the oriC regions of other α-proteobacteria like *C. crescentus*, *R. prowazekii*, *B. japonicum*, *S. meliloti* and *A. tumefaciens*.\(^\text{13–16}\) The analysis reveals that a cluster of genes including *parB*, *parA*, *gidB*, *gidA*, *thdF* and *rho* homologs is conserved at flanking regions of the predicted oriC, thus validating our prediction and confirming the locations of oriC at the non-translatable region near the rho gene in both species of *Bartonella*. The replication termini are expected to be at positions coinciding with the second shift (around 975 kb and 723 kb in *B. henselae* and *B. quintana*, respectively) in GC-skew. In *E. coli*, a 28-mer sequence (GGTGCGCATAATGTATATTGTTAAAT) known as a *dif* site lying in this region is required to convert circular chromosome dimers to monomers at the end of a replication cycle.\(^\text{23}\) As in other organisms,\(^\text{24}\) 28-mer sequences similar to that of *E. coli* *dif* have been found in both genomes of *Bartonella* near the regions of the second shift in GC-skew.

On the basis of the predicted oriC and terminus in each *Bartonella* genome, genes transcribed on the leading strand and those on the lagging strand are differentiated. Table 2 shows the mean G+T contents at three codon positions of the genes transcribed on the leading and lagging strands of the *B. henselae* and *B. quintana* genomes. The inter-genomic differences in the mean G+T contents of individual codon positions of the leading or lagging strand genes of the two organisms are small, but for both organisms, the mean G+T content of...
each codon position of the genes present in leading strand is significantly different (Mann-Whitney test, $P<0.0001$) from that on the lagging strand of replication. This indicates that the strand-specific compositional bias has significant influence on nucleotide selection not only at the synonymous sites but also at the first and second codon positions, which, in turn, may influence the amino acid usage in respective gene products. Therefore, in Bartonella DNA strand-specific mutational bias is an important force shaping codon usage and as well as amino acid content of the respective proteins. The inter-strand variation in G+T content is highest in the third and lowest in the second codon positions, while the intra-strand variation is highest in GT$_2$ (i.e., highest standard deviation) followed by GT$_3$ and GT$_1$ respectively.

3.2. Asymmetrical mutational bias, coupled with replicational-transcriptional selection on synonymous codon usage

Multivariate statistical analysis was used to study the codon usage variation among the genes in the two species of genus Bartonella. Figure 1a and 1b represent positions of the individual sequences on the plane defined by first and second principal axes generated by COA on RSCU values for B. henselae and B. quintana genes respectively. In each plot, the genes transcribed on the leading and the lagging strands of replication are segregated in two discrete clusters with little overlap along the first axis, which accounted for 12.5% of total variation for B. henselae (Fig. 1a) and 12.6% for B. quintana (Fig. 1b). When the coordinates of individual genes along axis 1 are plotted against the chromosomal locations of the corresponding genes, it is found that the coordinates of the genes transcribed rightward direction (and those of the genes transcribed leftwards) exhibit a transition from high to low (and low to high) values at the predicted oriC and ter regions (Fig. 2a, 2b). Thus, in both species, the primary influence on synonymous codon usage is whether a gene is transcribed in the same direction as replication, or opposite to it. No other axis accounted for significant quantities of total variance for both organisms under study. Furthermore, in both species of Bartonella, the number of predicted ORFs is significantly higher on the leading strand (58.9% for B. henselae and 57.8% for B. quintana) than on the lagging strand, and the distribution of highly expressed genes (including ribosomal proteins, dnaK, dnaJ, ssb, tuf, tsf, fusA and sub-units of RNA polymerase) are also significantly skewed, with most of the potential highly expressed genes (>70%) being transcribed from the leading strand. So, the discrepancies between leading and lagging strands indicate the combined effects of replication and transcription.
induced mutation. Replication-induced mutation is responsible for the higher number of genes on the leading strand, while mutations induced by transcription may be responsible for the enrichment of highly expressed genes on this strand.\textsuperscript{25,26} Replicational-transcriptional selection coupled with asymmetrical mutational bias is, therefore, primary cause of intra-genomic variations in codon usage pattern in \textit{Bartonella}. A similar observation was found previously in organisms belong to different taxonomic classes like the spirochaetes \textit{B. burgdorferi}\textsuperscript{8,9} and \textit{T. pallidum},\textsuperscript{8} the chlamydiae \textit{C. trachomatis},\textsuperscript{11} and the gamma-proteobacteria \textit{B. aphidicola}\textsuperscript{12} and \textit{B. floridanus}.\textsuperscript{10} However, to our knowledge, this is for the first report that the strand-specific biases is the prime selection force (i.e., variation represented by first axis of COA) behind the synonymous codon usage in any \textit{α}-proteobacteria.

In order to see whether the strand-specific codon usage is a general characteristic of \textit{α}-proteobacteria or it is a specific feature of \textit{Bartonella} species only, synonymous codon usage patterns in all \textit{α}-proteobacteria genomes sequenced so far have been analyzed in the present study. Table 1 represents the comparison of codon usage patterns (as determined by COA on RSCU values of genes) of 14 completely sequenced \textit{α}-proteobacteria. In \textit{C. crescentus} (order Caulobacterales) and seven other organisms belonging to Rhizobiales, the position of each gene along the first axis exhibits far stronger correlation with the respective GC\textsubscript{3} content than the corresponding GT\textsubscript{3}. \textit{B. henselae} and \textit{B. quintana} (characterized by extensive genome reduction) are the exceptions, although they belong to the Rhizobiales lineage of \textit{α}-proteobacteria. In both species of \textit{Bartonella}, a notable significant negative correlation exist ($r=-0.92$ for \textit{B. henselae} and $r=-0.93$ for \textit{B. quintana} at p<0.0001) between the positions of genes in the first major axis and GT\textsubscript{3} content of the sequences indicating the presence of strong DNA strand-specific compositional asymmetry in these organisms as the leading strand of replication is relatively enriched in G over C and T over A. On the other hand, in Rickettsiales (\textit{R. prowazekii}, \textit{R. conorii}, \textit{R. typhi} and \textit{W. pipientis}), the amount of variation explained by the first axis is very low (< 7% of total variation) and neither the GC\textsubscript{3} and nor the GT\textsubscript{3} contents of genes exhibit a notable significant relationship with it, suggesting that there is little heterogeneity in synonymous codon usage. These results reflect that codon usage bias due to pronounced strand-specific base compositional asymmetry might be a \textit{Bartonella}-specific property in \textit{α}-proteobacteria lineage.

The cumulative codon usage patterns corresponding to the genes located in leading and lagging strands of \textit{B. henselae} and \textit{B. quintana} are shown in Table 3. To estimate the statistical significance of the variations in codon usage between the genes of two strands, the chi-square test has been performed taking $P<0.01$ as the level of significance. The differences are found to be significant for 57 and 56 of 59 synonymously degenerate codons in \textit{B. henselae} and \textit{B. quintana}, respectively. The codons CCA (Pro), ACU (Thr) for both organisms and AGA (Arg) only for \textit{B. quintana} are the exceptions, which are almost equally used by genes of the leading and lagging strands. Previously it was observed in \textit{B. burgdorferi}\textsuperscript{9} that the codons, equally used for both strands are the rarely used codons for that organism. However, in \textit{Bartonella}, CCA (Pro) and ACU (Thr) are not among the rarely used codons (Table 3). Among the codons differentially used by two strands, 27 codons are used at the highest frequency on the leading strand (all G- or U-ending, except UUA which codes for Leu), whereas 30 and 29 codons are used most frequently on the lagging strand of \textit{B. henselae} and \textit{B. quintana}, respectively (all C- or A-ending, except CUU and CUG which code for Leu). Despite the anomalous behavior of the three codons for Leu, the biological interpretation of which is not clear, it can be concluded that in these two organisms, there are two distinct and significantly different codon usages in genes transcribed.
Table 3. Codon usage of the leading and lagging strand genes and preferred codons in highly expressed sequences for the *B. henselae* and *B. quintana* genomes.

| Amino Acid | Codon | Leading strand genes | Lagging strand genes | Leading strand genes | Lagging strand genes |
|------------|-------|-----------------------|----------------------|----------------------|----------------------|
|            |       | RSCU                  | N                    | RSCU                  | N                    |
| *B. henselae* |       |                       |                      | *B. quintana*        |                       |
| Phe        | UUU   | 1.77*                 | 10181                | 1.75*                 | 8775                 |
|            | UUC   | 0.23                  | 1315                 | 0.43*                 | 1239                 |
| Leu        | UUA   | 1.68*                 | 7077                 | 1.63*                 | 5897                 |
|            | UUG   | 1.50*                 | 6316                 | 1.56*                 | 5643                 |
|            | CUU   | 1.84                  | 7732                 | 1.84                  | 6676                 |
|            | CUC   | 0.41                  | 1738                 | 0.40                  | 1456                 |
|            | CUA   | 0.22                  | 942                  | 0.23                  | 842                  |
|            | CUG   | 0.34                  | 1415                 | 0.34                  | 1215                 |
| Ile        | AUU   | 2.04*                 | 12178                | 2.05*                 | 10500                |
|            | AUC   | 0.45                  | 2705                 | 0.47                  | 2404                 |
|            | AUA   | 0.51                  | 3040                 | 0.57+                 | 2474                 |
| Met        | AUG   | 1.00                  | 6447                 | 1.00                  | 5582                 |
| Val        | GUU   | 2.14*                 | 9974                 | 2.16*                 | 8759                 |
|            | GUC   | 0.40                  | 1844                 | 0.38                  | 1557                 |
|            | GUA   | 0.67                  | 3114                 | 0.64                  | 2586                 |
|            | GUG   | 0.80*                 | 3715                 | 0.81*                 | 3301                 |
| Ser        | UCU   | 2.01*                 | 5685                 | 1.99*                 | 4659                 |
|            | UCC   | 0.42                  | 1180                 | 0.75+                 | 937                  |
|            | UCA   | 1.24                  | 3523                 | 1.46+                 | 3089                 |
|            | UCG   | 0.56*                 | 1597                 | 0.33                  | 1462                 |
| Pro        | CUC   | 1.74*                 | 4124                 | 1.70*                 | 3510                 |
|            | CCC   | 0.39                  | 919                  | 0.77+                 | 830                  |
|            | CCA   | 1.34                  | 3170                 | 1.35                  | 2730                 |
|            | CCG   | 0.54*                 | 1278                 | 0.30                  | 1186                 |
| Thr        | ACU   | 1.00                  | 3135                 | 0.96                  | 2565                 |
|            | ACC   | 0.51                  | 1610                 | 0.78+                 | 1256                 |
|            | ACA   | 1.69                  | 5295                 | 1.87+                 | 4465                 |
|            | ACG   | 0.80*                 | 2523                 | 0.40                  | 2077                 |
| Ala        | GCU   | 1.62*                 | 8922                 | 1.62*                 | 7633                 |
|            | GCC   | 0.36                  | 1982                 | 0.70+                 | 1600                 |
|            | GCA   | 1.38                  | 7574                 | 1.57+                 | 6539                 |
|            | GCG   | 0.64*                 | 3514                 | 0.34                  | 3037                 |

from the leading and lagging strands of replication.

To see how the different codons are contributing towards codon usage variations among the genes transcribed on the leading and lagging strands of replication, we have plotted the distribution of codons on the first two major axes generated from COA on RSCU values (Fig. 3a, b). The G- and U-ending codons are distributed on the negative coordinate of axis 1, whereas C- and A-ending codons on the positive coordinate, the two codons for Leu (CUG and UUA) in both organisms, ACU (codes for Thr) in *B. henselae* and CUU (codes for Leu) in *B. quintana* are the exceptions. It is also notable that the triplets displaying the most extreme and opposite values on the first axis are NNG and NNC, while the variations in the NNU and NNA codons are relatively less in magnitude but opposite in direction with respect to the coordinate (Fig. 3). These observations suggest that G- and C-ending codons have greater contribution than their U- and A-ending counterparts to imparting strand-specific biases in the synonymous codon usage in *B. henselae* and *B. quintana* genes.

3.3. Influence of gene expressivity on synonymous codon usage

In an attempt to examine if translational selection could influence the codon usage in *Bartonella*, we have performed a COA on RSCU values of the genes transcribed in leading strand of replication, as most of the highly expressed genes are located in that strand. The position of genes along the first axis of COA is related to expressivity, since potential highly expressed genes (i.e. genes encoding ribosomal proteins, transcription trans-
Table 3. Continued.

| Amino Acid | Codon | B. henselae Leading strand genes | B. henselae Lagging strand genes | B. quintana Leading strand genes | B. quintana Lagging strand genes |
|------------|-------|---------------------------------|---------------------------------|---------------------------------|---------------------------------|
|            | RSCU  | N                               | RSCU  | N                               | RSCU  | N                               | RSCU  | N                               |
| Tyr        | 1.74* | 6146                            | 1.58  | 3819                            | 1.72* | 5036                            | 1.56  | 3391                            |
|            | 0.26  | 919                             | 0.42+ | 1020                           | 0.28  | 821                             | 0.44+ | 966                             |
| ter        | 1.56  | 366                             | 2.08  | 343                             | 1.67  | 340                             | 1.98  | 295                             |
| ter        | 0.45  | 107                             | 0.29  | 47                              | 0.44  | 89                              | 0.24  | 36                              |
| His        | CAU  1.68* | 4869                          | 1.46  | 3324                            | 1.67* | 3994                            | 1.47  | 3022                            |
|            | CAC  0.32  | 913                            | 0.54+ | 1231                            | 0.33  | 801                             | 0.53+ | 1099                            |
| Gln        | CAA  1.27  | 6285                           | 1.67+ | 6119                            | 1.23  | 5089                            | 1.63+ | 5456                            |
|            | CAG  0.73* | 3622                           | 0.33  | 1211                            | 0.77* | 3185                            | 0.37  | 1246                            |
| Asn        | AAU  1.67* | 8427                           | 1.44  | 5903                            | 1.68* | 6892                            | 1.41  | 5229                            |
|            | AAC  0.33  | 1666                           | 0.56+ | 2288                            | 0.32  | 1302                             | 0.59+ | 2173                            |
| Lys        | AAA  1.39  | 10336                          | 1.68+ | 8974                            | 1.38  | 8411                            | 1.67+ | 7968                            |
|            | AAG  0.61* | 4546                           | 0.32  | 1710                            | 0.62* | 3803                            | 0.33  | 1600                            |
| Asp        | GAU  1.80* | 12078                          | 1.59  | 6231                            | 1.78* | 10066                           | 1.57  | 5762                            |
|            | GAC  0.20  | 1321                           | 0.41+ | 1622                            | 0.22  | 1226                             | 0.43+ | 1569                            |
| Glu        | GAA  1.40  | 10916                          | 1.70+ | 8514                            | 1.40  | 9327                             | 1.69+ | 7487                            |
|            | GAG  0.60* | 4681                           | 0.30  | 1484                            | 0.60* | 3958                            | 0.31  | 1384                            |
| Cys        | UGU  1.55* | 2098                           | 1.26  | 1105                            | 1.53* | 1839                             | 1.23  | 1003                            |
|            | UGC  0.45  | 615                            | 0.74+ | 654                             | 0.47  | 572                             | 0.77+ | 622                             |
| ter        | UGA  0.99  | 233                            | 0.63  | 104                             | 0.90  | 183                             | 0.78  | 117                             |
| Trp        | UGG  1.00  | 2604                           | 1.00  | 1809                            | 1.00  | 2265                             | 1.00  | 1691                            |
| Arg        | CGU  2.93* | 6875                           | 2.40  | 3325                            | 2.84* | 5908                            | 2.35  | 3021                            |
|            | CGC  1.00  | 2340                           | 1.61+ | 2226                            | 1.01  | 2102                             | 1.67+ | 2147                            |
|            | CGA  0.54  | 1274                           | 0.77+ | 1065                            | 0.57  | 1178                             | 0.76+ | 978                             |
|            | CGG  0.58* | 1368                           | 0.31  | 427                             | 0.69* | 1440                             | 0.41  | 527                             |
| Ser        | AGU  1.19* | 3375                           | 0.91  | 1682                            | 1.14* | 2668                             | 0.93  | 1544                            |
|            | AGC  0.57  | 1621                           | 0.77+ | 1428                            | 0.53  | 1243                             | 0.82+ | 1370                            |
| Arg        | AGA  0.64  | 1502                           | 0.74+ | 1022                            | 0.58  | 1212                             | 0.65  | 839                             |
|            | AGG  0.31* | 716                            | 0.18  | 250                             | 0.30* | 622                             | 0.16  | 207                             |
| Gly        | GGU  1.68* | 8238                           | 1.42  | 3696                            | 1.69* | 6894                            | 1.40  | 3322                            |
|            | GGC  0.48  | 2372                           | 0.74+ | 1929                            | 0.50  | 2022                             | 0.80+ | 1903                            |
|            | GGA  1.10  | 5381                           | 1.46+ | 3792                            | 1.06  | 4302                             | 1.41+ | 3343                            |
|            | GGG  0.73* | 3581                           | 0.38  | 976                             | 0.75* | 3062                            | 0.39  | 932                             |

N, total number of codons; RSCU, relative synonymous codon usage.

*or + Indicate corresponding codons are significantly over-represented among leading or lagging strand genes (p<0.01). Bold letters represent corresponding codons are more frequent (p<0.05) in highly expressed sequences.

Translation processing factors, heat-shock proteins etc.) are clustered at one extreme. Axis 1 represents 8.5% of total variance for B. henselae leading strand genes and 7.6% of total variance for the leading strand genes of B. quintana. Consistent with this, when taking the pooled data from either extreme of the first axis generated by the correspondence analysis were compared, it was found that there are 18 codons for B. henselae and 19 codons for B. quintana whose usage is significantly higher among the highly expressed genes (Table 3) than among the genes in other extreme of axis 1. This observation suggests that apart from the replicational-transcriptional selection, a residual intra-strand codon bias is exhibited by the highly expressed genes encoded from the leading strands of B. quintana and B. henselae.

In order to ascertain the influence of the residual selection on synonymous codon bias in highly expressed genes, we have estimated the $d_S$ between 50 putative highly and 50 putative lowly expressed orthologous genes taken from the above mentioned pooled data of the two extremes of axis 1 generated by COA on RSCU for leading strand genes of both organisms. The genes with significantly low values of $d_S$ are the potential highly expressed genes (i.e., ribosomal proteins, transcription and translation processing factors, heat-shock proteins, etc.). The mean $d_S$ value for high-expression genes is 0.294,
3.4. Major sources of variation in protein composition

3.4.1. Influence of mean hydropathy level, aromaticity and strand-specific biases

The first four principle axes generated by COA of the amino acid frequencies in individual proteins explain 45.8% of the total variability for *B. henselae* and 45.4% for *B. quintana*. In both organisms, the first major axis shows a strong correlation with hydropobicity as well as with the aromaticity of encoded proteins. In *B. henselae* there are strong negative correlations between the positions of the genes along the first axis (represents 16.7% of total variance) and their corresponding hydropathy level ($r=-0.82$, $p<0.0001$) as well as aromaticity ($r=-0.74$, $p<0.0001$) of encoded protein; while in *B. quintana*, positions of genes on the first axis (represents 17.7% of total variance) positively correlated with hydropobicity ($r=0.84$, $p<0.0001$) and simultaneously with aromaticity ($r=0.75$, $p<0.0001$) of their protein product. This observation is surprising in view of the fact that the first axis of COA on amino acid frequencies in proteins of different organisms examined so far was found to be strongly correlated with a single parameter, hydropobicity, in most cases.²⁸,²⁹ But in the *Bartonella* species under examination, the mean hydropathy level and aromaticity — both seem to have nearly equal influence on the amino acid usage. The proteins present at the negative end of axis 1 in *B. henselae* or at the positive end in *B. quintana* are rich in hydrophobic residues, mostly represented by membrane proteins involved in signal transduction and several transports.

The second major axis discriminates the genes located on the leading and lagging strands with some overlapping between them. In *B. henselae*, 69.7% of genes on the leading strand present on the positive side and 70.8% of genes on the lagging strand present on the negative side of axis 2 (Fig. 4a) whereas in *B. quintana* 70.3% of the leading strand genes display negative values and 74.5% of the lagging strand genes display positive values on axis 2 (Fig. 4b). It is also important to note that there is a strong negative correlation ($r=-0.51$, $p<0.0001$) for *B. henselae* and a positive correlation ($r=0.60$, $p<0.0001$) for *B. quintana* between the coordinate of each sequence along that axis with the MMWs of the amino acids used in the respective protein. Therefore, in both organisms, the leading strand genes, in general, have lower MMW than the lagging strand genes. In other words, the smaller and energetically less expensive amino acids are preferred by leading strand gene products than that of the lagging strand. Since the highly expressed genes are mostly located on the leading strand of replication, this observation indicates that proteins encoded by highly expressed genes are preferentially constructed with smaller residues because these residues are energetically less expensive, as were also reported for the proteins in *Giardia lamblia*³⁰ and *Thermotoga maritime*.²⁹ This indicates that *B. henselae* and *B. quintana*, both in spite of their intracellular lifestyle, follow the cost-minimization hypothesis.³¹

To understand the influence of replication-associated mutational pressure on the amino acid composition of proteins, we have compared the RAAU values of proteins encoded by leading and lagging strand genes for both or-
ganisms. There are significant (p<0.001) interstrand differences for 14 amino acids in *B. henselae* and 13 amino acids in *B. quintana* (Table 4). As G and T bases are prevalent on the leading strand and C and A bases are over-represented on the lagging strand, the amino acids that were expected to be influenced by asymmetric base composition are those encoded by either G+T-rich or A+C-rich codons. There are 11 amino acid residues that fall into this category. Among these, Phe (TTY), Cys (TGY), Trp (TGG), Gly (GGN), and Val (GTN) are expected to be present at an increased frequency on the leading strand, whereas Lys (AAR), Asn (AAY), Pro (CCN), Gln (CAR), His (CAY) and Thr (ACR) are expected to be over-represented on the lagging strand. In both species of *Bartonella*, variation in amino acid usage almost always follows the expected trend towards increasing G+T content in the leading and A+C content in the lagging strand of replication. However, the amino acids Phe, Cys and Trp overruled this trend and occur with almost equal frequencies in proteins encoded by both strands (Table 4). It is also noticeable that Asp (GAY), Glu (GAR), Ala (GCN) and Arg (CGN/AGR) are over-represented in the leading strand genes whereas Leu (UUR/CUN) and Ile (AUH) are over-represented in the lagging strand genes. If the strand-specific mutational bias had been the sole source of inter-strand variation in amino acid composition of proteins encoded on the two strands, then the residues encoded by the codons having equal frequencies of G/T and A/C would exhibit a similar distribution in both strands. This, however, is not the case. Therefore, it can be concluded that, apart from the replication-associated mutational pressure, some other selection force(s) might influence the strand-specific variation in amino acid usage, and the gene expressivity could be a plausible origin of such force, as more than 70% of highly expressed genes are present on the leading strand.

### 3.4.2 Relation between gene expression, protein conservation and amino acid usage

In order to examine the possible effect of gene expression on amino acid usage, we compared the RAAU of each amino acid in putative highly and lowly expressed gene products. In both organisms, usage of Glu, Asp, Lys, Asn, Gln, Arg and Ala increase significantly (p<0.001)
in highly expressed gene products, while Phe, Leu, Ile, Ser, Tyr, Cys, Met and Trp occur at higher frequencies in lowly expressed proteins (Fig. 5). These two groups of amino acids have different features, both at the level of codons and in their molecular weight. The amino acids over-represented in highly expressed genes are mostly coded by those codons, which have a purine (R) in the first codon position (the exceptions are Gln and the quartet CGN of Arg). On the contrary, frequently used residues in lowly expressed genes are mostly coded by pyrimidine (Y) starting codons except Ile, Met and the duets (AGY) coding for Ser. Furthermore, the residues more abundant in highly expressed gene-products are comparatively small and energetically less expensive. In both species of Bartonella the highly expressed genes seem to avoid aromatic amino acids and other large residues in their products as seen in other microbial organisms.\textsuperscript{28,29,34} As already mentioned, this observation is compatible with the cost-minimization hypothesis of amino acid usage.\textsuperscript{31}

Finally, to explore the relationship between protein conservation and gene expression, we have compared the estimated pairwise non-synonymous divergences ($d_{\text{N}}$) between all orthologues of putative highly and lowly expressed genes pre-identified from the two extremes of axis 1 generated by COA on RSCU of the leading strand genes of two organisms. The mean $d_{\text{N}}=0.021$ for highly expressed genes is significantly lower than that of lowly expressed genes ($d_{\text{N}}=0.057$), suggesting that the amino acids of highly expressed genes are more conserved in both species of Bartonella. A similar inference was drawn previously in several organisms like E. coli and S. typhimurium.\textsuperscript{35} Therefore, we inferred that in Bartonella, the overall protein conservation in highly expressed genes is much higher and it also influences the discrimination in amino acid usage between highly and lowly expressed genes.

In conclusion, it can be said that though the most important source of variation in synonymous codon usage in B. henselae and B. quintana is ascribable to the asymmetrical mutational bias coupled with replicational-transcriptional selection, there is also a residual intra-strand bias in synonymous codon selection by the highly expressed genes, putatively operating at the level of translation. A question may arise at this point: why do the strand-specific asymmetries in codon usage patterns exist only in Bartonella among \textalpha-proteobacteria? Prokaryotic genomes exhibit strand compositional asymmetries to varying extents depending on the relative kinetics of genomic rearrangements and mutational bias.\textsuperscript{36} If the mutational bias is faster, then the corresponding genome is likely to exhibit observable asymmetry in strand composition. On the contrary, in genomes with high plasticity, a relatively high frequency of genomic rearrangement is likely to upset compositional skews induced by strand-specific mutational bias. In general, the genetic rearrangements relative to nucleotide substitutions occur at much lower frequencies in intracellular bacteria (usually have a reduced genome) than in free-living bacterial species (usually have an expanded genome);\textsuperscript{37} therefore, a low frequency of genomic rearrangement could be the plausible reason for pronounced influence of strand specific compositional asymmetries on codon usage in Bartonella. In some other members of \textalpha-proteobacteria such as in Rickettsia, unique genomic rearrangements are known to occur at a higher rate during evolution of obligate intracellular lifestyle.\textsuperscript{4,38} In Wolbachia too, intragenomic rearrangements associated with many DNA repeats\textsuperscript{39} might have upset the compositional asymmetry in two replicating strands.

In the case of protein composition, the mean hydropathy level and aromaticity seem to be the major sources of variation, both having nearly equal influence. Protein compositions in B. henselae and B. quintana also ex-
hhibit inter-strand biases in amino acid preferences under the influence of strand-specific mutational pressure and gene expressivity. In spite of the intracellular lifestyle, both species under study have been shown to follow the cost-minimization hypothesis. This is, however, not a unique feature of *Bartonella* among α-proteobacteria. Free-living members of this lineage follow high level of cost-minimization, even the intracellular bacteria like *Brucella* also exhibit cost-minimization at a lower level. As proposed by Peixoto et al., the existence of the cost-minimization effect in intracellular pathogens might be due to a genome-level adaptation to utilize less expensive and small residues from the host in the highly expressed genes. This might provide an evolutionary advantage by minimizing host energy exhaustion for maintaining sustained infection and the least chance of elimination by the host.

Acknowledgements: This work was supported by the Department of Biotechnology, Government of India (Grant Numbers BT/BI/01/017/97-Vol-II and BT/BI/04/055-2001).

References

1. Alsmark, C. M., Frank, A. C., Karlberg, E. O. et al. 2004, The louse-borne human pathogen *Bartonella quintana* is a genomic derivative of the zoonotic agent *Bartonella henselae*, *Proc. Natl. Acad. Sci. USA*, 101, 9716–9721.
2. Koehler, J. E. 1996, In: Aronoff, S. C., Hughes, W. T., Kohl, S., Speck, W. T., and Wald, E. R. (eds) Advances in Pediatric Infectious Diseases, Vol. 11. Mosby-Year Book, Inc., Chicago, pp. 1–27.
3. Andersson, S. G., Zomorodipour, A., Andersson, J. O. et al. 1998, The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria, *Nature*, 396, 133–140.
4. Ogata, H., Audic, S., Renesto-Audiffren, P. et al. 2001, Mechanisms of evolution in *Rickettsia conorii* and *R. prowazekii*, *Science*, 293, 2093–2098.
5. Rocha, E. P., Danchin, A., and Viari, A. 1999, Bacterial DNA strand compositional asymmetry: response, *Trends Microbiol.*, 7, 305–308.
6. Tillier, E. R. and Collins, R. A. 2000, The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes, *J. Mol. Evol.*, 50, 249–257.
7. Lobry, J. R. and Sueoka, N. 2002, Asymmetric directional mutation pressures in bacteria, *Genome Biol.*, 3, research0058.1–0058.14.
8. Lafay, B., Lloyd, A. T., McLean, M. J., Devine, K. M., Sharp, P. M., and Wolfe, K. H. 1999, Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases, *Nucleic Acids Res.*, 27, 1642–1649.
9. Mclnerney, J. O. 1998, Replication and transcriptional selection on codon usage in *Borrelia burgdorferi*, *Proc. Natl. Acad. Sci. USA*, 95, 10698–10703.
10. Banerjee, T., Basak, S., Gupta, S. K., and Ghosh, T. C. 2004, Evolutionary forces in shaping the codon and amino acid usages in *Buchnera floridanus*, *J. Bimol.*
11. Romero, H., Zavala, A., and Musto, H. 2000, Codon usage in *Chlamydia trachomatis* is the result of strand-specific mutational biases and a complex pattern of selective forces, *Nucleic Acids Res.*, 28, 2084–2090.
12. Rispe, C., Delmotte, F., van Ham, R. C., and Moya, A. 2004, Mutational and selective pressures on codon and amino acid usage in *Buchnera*, endosymbiotic bacteria of aphids, *Genome Res.*, 14, 44–53.
13. Brassinga, A. K., Siam, R., and Marczynski, G. T. 2001, Conserved gene cluster at replication origins of the alpha-proteobacteria *Caulobacter crescentus* and *Rickettsia prowazekii*, *J. Bacteriol.*, 183, 1824–1829.
14. Kaneko, T., Nakamura, Y., Sato, S. et al. 2002, Complete genomic sequence of nitrogen-fixing symbiotic bacterium *Bradyrhizobium japonicum* USDA110, *DNA Res.*, 9, 189–197.
15. Capela, D., Barloy-Hubler, F., Gouzy, J. et al. 2001, Analysis of the chromosome sequence of the legume symbiont *Sinorhizobium meliloti* strain 1021, *Proc. Natl. Acad. Sci. USA*, 98, 9877–9882.
16. Goodner, B., Hinkle, G., Gattung, S. et al. 2001, Genome sequence of the plant pathogen and biotechnology agent *Agrobacterium tumefaciens* C58, *Science*, 294, 2323–2328.
17. Lobry, J. R. 1996, Asymmetric substitution patterns in the two DNA strands of bacteria, *Mol. Biol. Evol.*, 13, 660–665.
18. Mackiewicz, P., Zakrzewska-Czerwinska, J., Zawilak, A., Dudek, M. R., and Cebrat, S. 2004, Where does bacterial replication start? Rules for predicting the oriC region, *Nucleic Acids Res.*, 32, 3781–3791.
19. Sharp, P. M. and Li, W. H. 1987, The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications, *Nucleic Acids Res.*, 15, 1281–1295.
20. Kyte, J. and Doolittle, R. F. 1982, A simple method for displaying the hydrophatic character of a protein, *J. Mol. Biol.*, 157, 105–132.
21. Nei, M. and Gojobori, T. 1986, Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions, *Mol. Biol. Evol.*, 3, 418–426.
22. Ogasawara, N. and Yoshikawa, H. 1992, Genes and their organization in the replication origin region of the bacterial chromosome, *Mol. Microbiol.*, 6, 629–634.
23. Blakely, G. W., Davidson, A. O., and Sherratt, D. J. 2000, Sequential strand exchange by XerC and XerD during site-specific recombination at dif, *J. Biol. Chem.*, 275, 9930–9936.
24. Yen, M. R., Lin, N. T., Hung, C. H., Choy, K. T., Weng, S. F., and Tseng, Y. H. 2002, oriC region and replication termination site, dif, of the *Xanthomonas campestris* pv. *campestris* 17 chromosome, *Appl. Environ. Microbiol.*, 68, 2924–2933.
25. Francino, M. P. and Ochman, H. 1997, Strand asymmetries in DNA evolution, *Trends Genet.*, 13, 240–245.
26. McLean, M. J., Wolfe, K. H., and Devine, K. M. 1998, Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes, *J. Mol. Evol.*, 47, 691–696.
27. de Miranda, A. B., Alvarez-Valin, F., Jabbari, K.,
Degrave, W. M., and Bernardi, G. 2000, Gene expression, amino acid conservation, and hydrophobicity are the main factors shaping codon preferences in *Mycobacterium tuberculosis* and *Mycobacterium leprae*, *J. Mol. Evol.*, **50**, 45–55.

28. Lobry, J. R. and Gautier, C. 1994, Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes, *Nucleic Acids Res.*, **22**, 3174–3180.

29. Zavala, A., Naya, H., Romero, H., and Musto, H. 2002, Trends in codon and amino acid usage in *Thermotoga maritima*, *J. Mol. Evol.*, **54**, 563–568.

30. Garat, B. and Musto, H. 2000, Trends of amino acid usage in the proteins from the unicellular parasite *Giardia lamblia*, *Biochem. Biophys. Res. Commun.*, **279**, 996–1000.

31. Seligmann, H. 2003, Cost-minimization of amino acid usage, *J. Mol. Evol.*, **56**, 151–161.

32. Mackiewicz, P., Gierlik, A., Kowalczyk, M., Dudek, M. R., and Cebrat, S. 1999, How does replication-associated mutational pressure influence amino acid composition of proteins?, *Genome Res.*, **9**, 409–416.

33. Morton, B. R. 1999, Strand asymmetry and codon usage bias in the chloroplast genome of *Euglena gracilis*, *Proc. Natl. Acad. Sci. USA*, **96**, 5123–5128.

34. Herbeck, J. T., Wall, D. P., and 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.

35. Sharp, P. M. 1991, Determinants of DNA sequence divergence between *Escherichia coli* and *Salmonella typhimurium*: codon usage, map position, and concerted evolution, *J. Mol. Evol.*, **33**, 23–33.

36. Lopez, P. and Philippe, H. 2001, Composition strand asymmetries in prokaryotic genomes: mutational bias and biased gene orientation, *C. R. Acad. Sci. III.*, **324**, 201–208.

37. Boussau, B., Karlberg, E. O., Frank, A. C., Legault, B. A., and Andersson, S. G. 2004, Computational inference of scenarios for alpha-proteobacterial genome evolution, *Proc. Natl. Acad. Sci. USA*, **101**, 9722–9727.

38. Andersson, J. O. and Andersson, S. G. 1997, Genomic rearrangements during evolution of the obligate intracellular parasite *Rickettsia prowazekii* as inferred from an analysis of 52015 bp nucleotide sequence, *Microbiology*, **143**, 2783–2795.

39. Wu, M., Sun, L. V., Vamathevan, J. et al. 2004, Phylogenomics of the reproductive parasite *Wolbachia pipientis* wMel: a streamlined genome overrun by mobile genetic elements, *PLoS Biol.*, **2**, e327–e341.

40. Peixoto, L., Fernandez, V., and Musto, H. 2004, The effect of expression levels on codon usage in Plasmodium falciparum, *Parasitology*, **128**, 245–251.