Trends in amino acid usage across the class Mollicutes

Andrés Iriarte, Juan Diego Baraibar, Leticia Diana, Susana Castro-Sowinski, Héctor Romero, and Héctor Musto*

*Laboratorio de Organización y Evolución del Genoma, Facultad de Ciencias, UDELAR. Iguá 4225, Montevideo, 11400 Uruguay; Laboratorio de Evolución, Facultad de Ciencias, UDELAR. Iguá 4225, Montevideo, 11400 Uruguay; Área Genética, Depto. de Genética y Mejora Animal, Facultad de Veterinaria, UDELAR. Av. A. Lasplaces s/n, Montevideo, CP 11600, Uruguay; Sección Bioquímica, Facultad de Ciencias, UDELAR. Iguá 4225, Montevideo, 11400 Uruguay; Centro Universitario Regional Este (C.U.R. E.), UDELAR, Maldonado, Uruguay

Communicated by Ramaswamy H. Sarma

(Received 3 November 2012; final version received 8 November 2012)

Many studies have explored the mechanisms involved in relative amino acid usage (RAAU) in a variety of prokaryotes and eukaryotes. A strong bias was observed in highly expressed genes (HEGs) of endosymbiotic bacteria. By means of correspondence analysis, we studied the major trends affecting internal variability of RAAU in Mollicutes. The principal trend is related to the usage of smaller, less aromatic, and GC-rich coded amino acids in HEGs. Given the nature of the genetic code, these properties are linked among them. Expectedly, we found a slow evolutionary rate of HEGs, which is likely driven by purifying selection. On the other hand, the rest of the genes accumulate rapid changes as a result of the extreme mutational bias toward A+T of the genomes and genetic drift, increasing internal variability. Amino acid changes across the phylogeny of the group were traced in order to estimate the mean molecular weight and aromaticity trends in each branch. Finally, we compared amino acid usage bias within and between Mollicutes and the free-living Firmicutes.

Keywords: Mycoplasma; complete genome; amino acid usage; highly expressed genes; evolution

Introduction

Understanding the main factors affecting the relative amino acid usage (RAAU) in different genomes is one of the major goals in molecular evolution, and is usually studied using multivariate approaches like correspondence analysis (COA). In the first report of this kind, (Lobry & Gautier, 1994) found that in Escherichia coli, the three most important factors regarding RAAU were hydrophobicity, expressivity, and aromaticity of the proteins. They proposed that translational constraints, which are stronger in highly expressed genes (HEGs), might affect the global amino acid composition. Using the same approach, Garat and Musto (2000) found that in Giardia lamblia, the most relevant factor in RAAU is related to the mechanisms of defense against reactive oxygen species. Further, these authors found a clear tendency of HEGs to use smaller amino acids, so the cell economy seemed to be another prominent factor. From then on, many authors have explored the possible effect of selective pressures over RAAU in a variety of prokaryotes and eukaryotes (e.g. Akashi & Gojobori, 2002; Banerjee & Ghosh, 2006; Basak, Banerjee, Gupta, & Ghosh, 2004; McDonald, 2010; Naya, Zavala, Romero, Rodriguez-Maseda, & Musto, 2004; Palacios & Wernegreen, 2002; Tekaia & Yeramian, 2006; Tekaia, Yeramian, & Dujon, 2002; Sabbia et al., 2007; Swire, 2007; Wang & Lercher, 2010; Zavala, Naya, Romero, & Musto, 2002).

Considering the particular case of endosymbiotic bacteria (Das, Paul, Chatterjee, & Dutta, 2005; Herbeck, Wall, & Wernegreen, 2003; Palacios & Wernegreen 2002; Rispe, Delmotte, van Ham, & Moya, 2004; Schaber et al., 2005), some authors postulated that RAAU bias in HEGs is driven by two factors: the avoidance of aromatic residues and the resistance to an enrichment in A+T (Rispe et al., 2004). According to this view, greater conservation in HEGs since the divergence of the group from the last free-living common ancestor may explain the observed resistance to A+T enrichment (Banerjee, Basak, Gupta, & Ghosh, 2004; Banerjee & Ghosh, 2006; Herbeck, Wall, & Wernegreen, 2003; Rispe et al., 2004).

On the other hand, it has been argued that G+C-rich amino acids are favored by natural selection (Schaber
et al., 2005). In the case of Mollicutes, Yamao, Andachi, Muto, Ikemura, and Osawa, 1991 suggested that the amount of tRNAs is affected by the corresponding amino acid usage in ribosomal proteins. Under this scenario, neutral evolution cannot explain the observed RAAU bias.

Protein evolutionary rates are usually quantified by the estimation of the number of nonsynonymous nucleotide changes per nonsynonymous site (dN), and expression levels might be its best indicator (Duret & Mouchiroud, 2000; Pal, Papp, & Hurst, 2001, 2003; Rocha & Danchin, 2004; Subramanian & Kumar, 2004). This is reinforced by the fact that in several species, HEGs are systematically more conserved in comparison to lowly expressed genes (LEGs). The influence of this phenomenon on intragenomic RAAU variability should be particularly evident for organisms that experience substitutions due to selective or neutral forces, the latter expected to be less effective in HEGs.

Mollicutes, as well as other endosymbiotic bacteria, go through recursive bottlenecks. They typically have reduced genomes, accelerated evolutionary rates and a strong A+T mutational bias, mainly reflecting an increased effect of genetic drift and mutational pressure. Biosynthetic capabilities are reduced in these organisms so several nutrients must be acquired from their hosts (Razin, Yogev, & Naot, 1998). Based on phylogenies inferred either from 16S rRNA and/or conserved protein coding genes, Mollicutes are known to be a branch of the Firmicutes, as well as Bacilli and Clostridia (Woese, Maniloff, & Zablen, 1980; Wolf, Muller, Dandekar, & Pollack, 2004). Although the taxonomy of the group is still a matter of controversy (Boone, Castenholz, & Garrity, 2001), the phylogeny of Mollicutes shows two main groups: the AAA branch (Astereoplasma, Anaeroeroplasma, Phytoplasma, and Acholeplasma) and the SEM branch (Spiroplasma, Entomoplasma, Mesoplasma, Mycoplasma, and Ureaplasma) (Razin, Yogev, & Naot, 1998). Considering only the complete sequenced Mollicutes, studies usually deal with four monophyletic clusters (Oshima & Nishida, 2007): hominis, pneumoniae, and spiroplasma which belong to the SEM branch while phytoplasma-like cluster concerns all organisms in the AAA branch.

At the moment of writing this manuscript, there are enough complete sequenced genomes of Mollicutes and, as far as we know, no study has focused exclusively on RAAU in this class under a comparative and phylogenetic framework. On the other hand, many articles discussed the influence of expression level on RAAU in organisms with similar population dynamics and genomic features.

In this study, our focus is to analyze factors shaping amino acid composition of proteins within Mollicutes taking into consideration some key properties of genes (like expression levels and GC content) and features of amino acids (like molecular weight, aromaticity, and hydrophobicity). We consider and discuss the possible role of natural selection, mutational bias and genetic drift taking an intragenomic and intergenomic perspective. We also make inference of aromaticity and mean molecular weight (MMW) trends based on orthologous amino acid changes traced among the class.

Materials and methods

Sequences

Complete genome and coding sequences were downloaded from NCBI ftp site (ftp://ftp.ncbi.nih.gov/ genomes/Bacteria). Finally, 19 complete Mollicutes and 14 other Firmicutes (outgroup) were studied (Table 1).

Phylogenetic reconstruction

The reconstruction procedure is fully described elsewhere (Iriarte, Baraibar, Romero, & Musto, 2011). The tree was rooted between Clostridia and the rest of the Firmicutes based on the interactive tree of life online tool (http://itol.embl.de/).

Amino acid usage analyses

The major sources of variation among proteins were studied through COA (Greenacre, 1984) on RAAU. Base composition, amino acid usage, gravy score, aromaticity, and COA were calculated with CodonW 1.4.2 (written by John Peden and available from http://sourceforge.net/projects/codonw/).

RAAU bias ($\beta$) in HEGs was calculated in relation to the total proteome, for each residue in each species, according to the following equation:

$$\beta_i = \frac{F_{R_i}^{\text{HEGs}} - F_{R_i}^{\text{total}}}{F_{R_i}^{\text{total}}}$$

where, $F_{R_i}$ is the relative frequency of amino acid $i$ calculated for HEGs and for the rest of the genes in each genome, $F_{R_i}^{\text{HEGs}}$ and $F_{R_i}^{\text{total}}$, respectively. HEGs were defined in two different ways: (1) genes encoding ribosomal proteins + elongation factors, $\beta^{\text{REF}}$, and (2) as the top 10% of the HEGs identified by COA-RAAU for Mollicutes, $\beta^{\text{COA}}$.

A hierarchical cluster analysis of studied species was done based on the estimated $\beta^{\text{REF}}$ following the method developed by Ward. The analysis was done by means of the “pvclust” package (Suzuki & Shimodaira, 2006) from R Development Core Team (2012). Pvclust assesses the uncertainty in hierarchical cluster analysis using bootstrap resampling techniques ($n=1000$) and provides two types of $p$-values: Approximately Unbiased and Bootstrap Probability.
Estimation of expression levels

For each gene, we computed MELP values as an expression-level predictor (Supek & Vlahovicek, 2005) using INCA2.1 (Supek & Vlahovicek, 2004). For a description of how it was used in this study, see Supek and Vlahovicek (2005). MELP is valid as an expression-level predictor for organisms in which natural selection for translation has been shown to shape the codon usage: eight Mollicutes (Iriarte et al., 2011; Yamao et al., 1991) and the outgroup (Musto, Romero, & Zavala, 2003; Sharp, Bailes, Grocock, Peden, & Sockett, 2005). Recently, Supek, Skunca, Repar, Vlahovicek, & Smuc, 2010 proved that translational selection in prokaryotes is practically universal. Therefore, the clustering of HEGs (including the reference set and other HEGs) at one extreme of the distribution of an axis generated by COA and the correlation of the position of the sequences along this axis with the respective MELP values were taken as evidence for the existence of a trend associated to expression level.

Energetic cost

As an index of relative cost, we calculated the MMW for each protein, and was calculated as in Zavala et al. (2002).

Trace of amino acid changes and molecular distances

Eighty-five putative orthologous sequences were identified among Mollicutes, as previously described in Iriarte et al., 2011. Protein sequences were aligned using...
ClustalW (Thompson, Higgins, & Gibson, 1994). Maximum likelihood method, using REV (nr = 189) model was used to infer amino acid changes. Trace, reconstruction and global dN divergence for orthologs were estimated by means of Codeml program, included in PAML 4.0 (Yang, 2007) package.

Putative ortholog proteins were identified as in (Iriarte et al., 2011). Pairwise nonsynonymous (dN) distances among hominis, pneumoniae, spiroplasma, and “phytoplasma-like” groups were also calculated using Codeml (see Table 1).

Trends in molecular weight (ΔMW) were estimated as the differences in MW of the involved amino acids inferred to have changed in the phylogeny. Protein average molecular weight changes were calculated in each branch for orthologous sequences, 23 HEGs, and 62 LEGs (Supplementary Table 1). In order to estimate trends in aromaticity among these proteins, the number of changes to (positive) Phe, Tyr, and Trp were calculated. Finally, GC content variation associated with each an amino acid change (AGC) was estimated as the difference in the average GC content between triplet coding for the involved residues.

**Results and discussion**

The phylogenetic tree was in agreement with previous reconstructions for this group (Oshima & Nishida, 2007; Woese et al., 1980; Wolf et al., 2004). The bootstrap values displayed in Supplementary Figure 1 are high in almost all nodes. Therefore, amino acid changes and global dN estimations were traced on it, since we assumed that this tree reflects correctly the real relationships among the analyzed species.

**Correspondence analysis on RAAU**

This analysis showed that *M. synoviae, M. agalactiae, M. pulmonis, M. hyopneumoniae, Ureaplasma parvum, Mesoplasma florum, M. capricolum, M. mycoides, Candidatus Phytoplasma mali, Onion yellows phytoplasma*, and *Aster yellows phytoplasma* present a common pattern. Indeed, the first trend (which, depending on the species, accounted between 18.7 and 25.6% of the total variability) was related to expression levels, as shown by the fact that HEGs tended to cluster at one side of the distribution (Supplementary Figure 2). Furthermore, MMW and GC content at the first two codon positions [GQ(1–2)] were linked with this principal trend (Table 2). With the exception of *M. synoviae*, the second axis (between 14.1 and 16.9% of the variability) was correlated with the aromaticity and hidropathy level of each protein. As hydrophobic and aromatic amino acids are energetically expensive, it was not surprising to find that this trend was often also related with MMW. In *M. synoviae*, the second axis was associated with aromaticity.

![Figure 1](image-url)

Figure 1. Trends in MW (A) and Aromaticity and (B) traced on the Mollicutes phylogenetic tree. Trends in MW were estimated for each branch as the differences in MW of the involved amino acids inferred to have changed. Trends in aromaticity were estimated as the number of changes to (positive) and from (negative) Phe, Tyr, and Trp. Averages are shown in each branch for orthologous sequences, 23 HEGs (left) and 62 LEGs (right) (Supplementary Table 1).
Pairwise dN distances were estimated within phylogenetic groups, as a result, the distributions of the sequences according to dN showed that the most heavily expressed genes display the lower values. This fact held true for all the species analyzed here. Moreover, as expected, there were strong correlations within each species between the position of the genes along the “expression” axes and dN (Table 2 and Supplementary Table 2).

*M. mobile*, *M. pneumoniae*, and *M. gallisepticum* fit neither in pattern 1 nor 2. The first trend in *M. mobile* was not associated with expression levels, though it was significantly correlated with MMW. In addition, the second trend correlated with the usage of Phe, Asn, and Trp. In the case of *M. pneumoniae* and *M. gallisepticum*, the third trend (~11.5% of the variability) was marginally related to the expression level. We stress, however, that the correlation with dN was significant (0.4 < r < 0.63, p < 0.001).

Interestingly, COA on RAAU did not find any trend associated with expression levels in the species used as outgroup, even in those whose GC content was in the range of Mollicutes. It seems reasonable to assume that expression levels was an important source of amino acids usage early in the evolution of Mollicutes, and this can be interpreted as a by-product of the evolutionary path toward a parasitic life style. Indeed, simple mutational bias is not enough to explain this pattern, since, as mentioned above, some organisms in the outgroup display genomic GC levels that fall within the range found in Mollicutes. Thus, genetic drift and/or a high evolutionary rate might (at least partially) explain these patterns and the similar ones observed in other endosymbiotic bacteria. We propose that, among Mollicutes, due to genetic drift, genes that do not experience a highly constrained amino acid sequence evolution (LEGs) may have suffered rapid accumulation of AT-rich amino acids generating significant internal variation. On the contrary, under this scenario, the same genes in free-living species may have more time to accumulate compensatory changes in response to mutational bias, which in turn keeps intragenomic variability at a low level.

To sum up, COA on RAAU showed that there were differences among the principal trends in amino acid usage among Mollicutes. However, two factors were widely distributed in this class. First of all, HEGs were more conserved at nonsynonymous positions, and second, expression, MMW, aromaticity, and GC content at the first and second codon positions played a primary role when explaining intragenomic RAAU variation. Similar situations have been reported in *Blochmannia floridanus* (Banerjee et al., 2004) and several endosymbiont γ-3-proteobacteria (Herbeck, Wall, & Wernegreen, 2003; Palacios & Wernegreen, 2002; Rispe et al., 2004).

**Amino acids usage bias in highly expressed genes (β)**

The direction of the relative amino acid biases in the reference set in comparison with the whole set of genes in each species (see Material and methods) were conserved across all studied organisms (Supplementary Table 3). We note that almost identical results were obtained when the highly expressed set of genes was
Interestingly, there are several differences between Mollicutes and Firmicutes (Table 3). As can be seen, 13 residues display statistically significant differences between the two groups ($U$-test, $p < 0.01$). In other words, although the direction of the biases in the Mollicutes and Firmicutes are identical, the strength of the biases are different. Two nonmutually exclusive explanations can account for this fact. The first is the existence of a strong phylogenetical inertia (see above and Supplementary Figure 3), while the second can be related to differences in lifestyles (parasitic vs. free-living organisms). This may be partly explained by the important reduction of genes involved in amino acid metabolism in the common
Table 3. Comparison of amino acid usage bias of HEGs ($P_{(ref)}$) among groups.\(^{ab}\)

| Amino acid | Mollicutes | No-Mollicutes | Significance\(^{(c)}\) | SEM group | AAA group | Significance\(^{(c)}\) |
|------------|------------|---------------|-------------------------|-----------|-----------|-------------------------|
| Ala        | 0.54 (±0.20) | 0.26 (±0.15)  | −0.47 (±0.03) | −0.47 (±0.02)  |
| Arg        | 1.45 (±0.21) | 1.15 (±0.20)  | −0.25 (±0.03) | −0.25 (±0.02)  |
| Asn        | −0.31 (±0.08) | −0.15 (±0.13) | −0.27 (±0.08) | −0.24 (±0.07)  |
| Asp        | −0.29 (±0.07) | −0.19 (±0.09) | 0.34 (±0.10)  | 0.12 (±0.04)   |
| Cys        | 0.15 (±0.48)  | −0.12 (±0.23) | 0.31 (±0.10)  | 0.55 (±0.19)  |
| Gin        | 0.00 (±0.13)  | −0.16 (±0.21) | −0.16 (±0.06) | −0.06 (±0.09)  |
| Glu        | −0.08 (±0.06) | 0.01 (±0.09)  | 0.12 (±0.11)  | 0.07 (±0.09)   |
| Gly        | 0.41 (±0.14)  | 0.19 (±0.06)  | 0.14 (±0.10)  | 0.08 (±0.02)   |
| His        | 0.39 (±0.24)  | 0.02 (±0.22)  | 0.51 (±0.21)  | 0.64 (±0.11)  |
| Ile        | −0.26 (±0.07) | −0.28 (±0.07) | −0.40 (±0.05) | −0.41 (±0.04)  |
| Leu        | −0.25 (±0.03) | −0.31 (±0.05) | 0.49 (±0.19)  | 0.13 (±0.18)   |
| Lys        | 0.33 (±0.12)  | 0.50 (±0.20)  | 0.02 (±0.12)  | −0.08 (±0.14)  |
| Met        | 0.28 (±0.15)  | −0.04 (±0.09) | −0.32 (±0.07) | −0.27 (±0.06)  |
| Phe        | −0.47 (±0.03) | −0.43 (±0.05) | 0.36 (±0.11)  | 0.25 (±0.07)   |
| Pro        | 0.11 (±0.11)  | 0.10 (±0.15)  | −0.31 (±0.05) | −0.24 (±0.06)  |
| Ser        | −0.14 (±0.08) | −0.17 (±0.07) | −0.08 (±0.06) | −0.09 (±0.02)  |
| Thr        | 0.12 (±0.09)  | 0.03 (±0.08)  | 0.32 (±0.44)  | −0.33 (±0.25)  |
| Trp        | −0.48 (±0.07) | −0.55 (±0.08) | −0.51 (±0.05) | −0.41 (±0.08)  |
| Tyr        | −0.40 (±0.05) | −0.33 (±0.06) | 1.40 (±0.19)  | 1.57 (±0.20)   |
| Val        | 0.37 (±0.17)  | 0.30 (±0.08)  | 0.38 (±0.13)  | 0.47 (±0.12)   |

\(^{a}\)Average $P_{(ref)}$ for each amino acid in each phylogenetic group. In parenthesis is shown the standard deviation of the distribution.

\(^{b}\)Amino acid usage comparisons were made between Mollicutes vs. Non-Mollicutes and between the AAA branch (Phytoplasma and Acholeplasma) vs. the SEM branch (Mesoplasma, Mycoplasma and Ureaplasma).

\(^{c}\)The statistically significant differences are marked with ** (U-test, p < 0.01).

ancestor of Mollicutes (Chen, Chung, Lin, & Kuo, 2012).

Species which belong to the branches AAA (Phytoplasm and Acholeplasma) and SEM (Mesoplasma, Mycoplasma and Ureaplasma) were also compared (Table 3). Results showed similar biases in all amino acids, and the differences became statistically significant only for Cys, His, Met, and Trp (U-test, p < 0.01). No particular feature of these amino acids seems to explain the observed pattern. We think that these results can be partially associated with differences in the main hosts (animal vs. plants) and/or the independent pathway toward parasitic lifestyle of each group (Razin, Yogev, & Naot, 1998).

We identified some particular cases that did not match the expected phylogenetic pattern (Supplementary Figure 3). This is the case of species of the Clostridium genus, Mycoplasma mycoides, and A. laidlawii. The idiosyncratic behavior observed in Clostridium may be linked with their basal phylogenetic position respective to the rest of the organisms analyzed. Similarly, Acholeplasma are believed to have retained many ancestral characteristics in comparison to phytplasmas and other highly diverged Mollicutes (Bai et al., 2006; Iriarte et al., 2011; Lazarev et al., 2011) No particular feature seems to explain the observed result for M. mycoides.

Once again, the underrepresentation of AT$_{12}$-rich codons, and aromatic and heavy amino acids may explain the trends detected in HEGs, but none of these properties could explain the whole trend by itself. These properties, which are deeply interdependent and superimposed, might affect amino acid usage evolution in Mollicutes but also in the other Firmicutes analyzed (see the estimated $P_{(ref)}$ in Supplementary Table 3). Given that HEGs biases (estimated with $P_{(ref)}$) are mainly conserved in the outgroup and in several other free-living species, there is no link between these trends in HEGs and distinctive selection pressures in Mollicutes. Besides, when the MMW were estimated without considering Tyr, Phe and Trp, the significance of the correlations were lost. This suggests that the avoidance of these three aromatic residues is a main determinant for the MMW biases in HEGs and supports the interdependence among the properties studied.

As previously reported (Herbeck, Wall, & Wernegreen, 2003; Palacios & Wernegreen, 2002; Rispe et al., 2004), purifying selection acting at the amino acid usage level on HEGs in parasitic species is expected to be less effective than drive and mutational biases in relation to free-living relatives. On the other hand, we found that in Mollicutes (as well as in other symbiotic species), expression levels (plus GC$_{12}$, MW, etc) is a main factor, often the first one, explaining an important part of the internal variability on the usage of amino acids (11% < inertia of the “expression” axis < 25%). Examples can be found in Palacios and Wernegreen (2002), Herbeck, Wall, and Wernegreen (2003), and Rispe et al. (2004) for parasitic endosymbiotic bacteria. Considering unicellular eukaryotic parasites, similar results were reported for Leishmania (Chauhan, Vidyarthi, & Poddar, 2011) and for Plasmodium falciparum (Chanda, Pan, & Dutta,
In opposition to the “symbiotic” pattern, for the rest of the firmicutes, no axis was related with expression levels. In *E. coli* and *Thermotoga maritima*, the “expression” axes only explained between 11.3 and 13% of the internal variability (Lobry & Gautier, 1994; Zavala et al., 2002). This is also true for *Saccharomyces cerevisiae* (Kahali, Basak, & Ghosh, 2007). Once again, this result is quite puzzling for parasitic bacteria, where a lower effect of purifying selection is expected. We believe that the conserved biases of HEGs is widespread in a variety of taxa, but that in many parasitic species it emerges as a very important factor explaining intragenomic variability. Note that this statement does not rule out the possibility that some other features, such as genome contraction, may be partially responsible of the observed pattern.

**Molecular evolutionary rate, mutational bias and amino acids changes**

Many years ago, Karlin and Bucher (1992) described correlated amino acid frequencies based on strong (GC-rich) and weak (AT-rich) codons, among other features. Nowadays, it is widely accepted that changes in nucleotide biases affect global amino acid composition and probably molecular weight and cost. Singer and Hickey (2000), among others, reported that Phe, Tyr, Met, Ile, Asp, and Lys frequencies correlate negatively with GC composition, while the opposite occurs with Gly, Ala, Arg, and Pro. Mutational bias and genetic drift effects may be always less pronounced in HEGs that are believed to evolve slower at nonsynonymous and synonymous sites (Pal, Papp, & Hurst, 2001, 2003). In these sequences, conserved amino acids biases are driven by purifying selection and may have a functional, structural, and/or energy cost basis. Nevertheless, our results show that the relative contribution of each factor is not clear, which one-by-one cannot completely explain the observed patterns.

Twenty-three (23) out of 85 putative orthologous sequences were considered HEGs (elongation factors tufA, fusA, and tsf plus 20 ribosomal proteins genes). The other 62 putative orthologous sequences were considered LEGs (tRNA synthetases, kinases, DNA-ligase, among others) (see Supplementary Table 1 for a complete list of all orthologs). This classification was conservative since all the latter group was certainly not composed only by sequences expressed at low levels.

The degree of conservation between these two groups was compared as the sum of the dN of each branch across the tree. Results showed that HEGs were significantly more conserved than LEGs (*U*-test, *p* < 0.0001). Note that if, estimated only for this group of genes, showed essentially the same trend as comparisons made considering the whole proteome (data not shown). Both results validated working with orthologous sequences and suggest that the effect of horizontal gene transfer events on the general patterns is rather marginal.

The trends in molecular weight and aromaticity, derived from the amino acid changes traced in the phylogeny, were averaged for both groups of orthologs in each branch (Figures 2 and 3). Although the reconstruction of the ancestral compositional biases was not done, a simple inspection of the tree suggested that the observed trends may be explained by the differences in genomic GC. As expected, changes in compositional biases toward an increased GC content were associated with a decreasing MMW tendency, and *vice versa* (see Figure 2). HEGs usually experience the same bias than LEGs but with a lesser degree, which may be mainly due to their higher degree of conservation. In agreement with previous results, the aromatic trends strongly correlated with average MW changes (see above), both for HEGs (*r* = 0.74) and LEGs (*r* = 0.83) (Figure 1).

As previously stated, and as can be deduced from the structure of the genetic code, there is an expected negative correlation between the GC content of a set of genes and the molecular weight of the corresponding proteins. However, since not all amino acid changes are equally frequent, it is interesting to analyze “top 20%” of the most frequent ones, which accounts for more than 70% of the inferred total changes in the phylogeny. The results confirm the idea that positive ΔGC are associated with negative ΔMW trends (*r* = −0.49, *p* < 0.0001).

The global amino acid substitution matrix of HEGs and LEGs was compared. A contingency *χ*² test was used to establish significant differences between both groups of genes. A total of 14 changes were recognized as significantly more frequent in HEGs (*χ*² *p* < 0.01), while 12 changes were more frequent in LEGs (*χ*² *p* < 0.01) (Supplementary Figure 4). Results showed that, on an average, the more frequent amino acid changes in HEGs were also characterized by a decreasing MW (ΔMW = −8.10), while the opposite was true for LEGs (ΔMW = −1.85). The average AMW and aromatic trend per inferred change in each branch confirmed this observation. Taken together, these results suggest that HEGs are characterized not only by a reduced nonsynonymous substitution rate but also by more conservative changes.

Identical directional trends in HEGs and LEGs were observed in some lineages (Figure 1). This probably reflects the action of a strong nucleotide compositional bias throughout these lineages. In addition, a negative correlation between the observed MW bias of HEGs (the difference between MW of HEGs in relation to the rest of the genes in each genome) and the genomic GC was observed for all the species analyzed in the present study (*r*² = 0.62). Note that the observed bias in MW was slightly more pronounced in Mollicutes than in the outgroup when considering the same GC content variation (compare the slopes in Figure 2). This supports
the idea that under the same mutational bias, the intragenomic difference between HEGs and LEGs may increase faster in Mollicutes than in the outgroup. We believe that this is probably attributable to the noteworthy effect of genetic drift and the characteristic high mutation rate. This observation does not undermine the idea that intragenomic MMW difference between HEGs and LEGs is mainly driven by the frequency of aromatic residues, which are among the biggest amino acids. Therefore, even thought that this “aromatic” bias may also have a cell economy explanation, other functional properties cannot be completely discarded, for instance, residue-specific interactions or functions (Dougherty, 2007; Gallivan & Dougherty, 1999).

In summary, we propose that among Mollicutes, several pressures affect amino acid usage. The main factors are mutational bias, genetic drift, and purifying selection. The first two should be quantitatively very important in these organisms given their way of life, which in most cases is characterized by recursive bottlenecks and should be more effective in LEGs. On the other hand, negative selection should be very important for HEGs. Thus, the cumulative effect of these evolutionary forces acting differently among genes, might explain the observed increase of the internal variability in the genomes of Mollicutes.

Abbreviations

| Abbreviation | Definition |
|--------------|------------|
| GC           | molar content of guanine + cytosine |
| HEGs         | highly expressed genes |
| LEGs         | lowly expressed genes |
| COA          | correspondence analysis |
| RAUU         | relative amino acid usage |
| MELP         | MILC-based expression level predictor |
| dN           | Nonsynonymous substitution rate |
| β            | RAUU bias |
| MW           | Molecular weight |

Acknowledgment
We thank Rosina Piovani for fruitful discussions. This work was partially supported by Fondo Clemente Estable 2007, FCE 722 (ANII, Uruguay) and PEDECIBA (Uruguay).

Supplementary Material
The supplementary material for this paper is available online at http://dx.doi.org/10.1080/07391102.2012.748636.

References
Akashi, H., & Gojobori, T. (2002). Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. Proceedings of the National Academy of Sciences, 99, 3695–3700.

Bai, X., Zhang, J., Ewing, A., Miller, S. A., Jancso Radek, A., Shevchenko, D. V., …, Hogenhout, S. A. (2006). Living with genome instability: The adaptation of phytoplasmas to diverse environments of their insect and plant hosts. Journal of Bacteriology, 188, 3682–3696.

Banerjee, T., Basak, S., Gupta, S. K., & Ghosh, T. C. (2004). Evolutionary forces in shaping the codon and amino acid usages in Blochmannia floridanus. Journal of Biomolecular Structure & Dynamics, 22, 13–23.

Banerjee, T., & Ghosh, T. C. (2006). Gene expression level shapes the amino acid usages in Prochlorococcus marinus MED4. Journal of Biomolecular Structure & Dynamics, 23, 547–554.

Basak, S., Banerjee, T., Gupta, S. K., & Ghosh, T. C. (2004). Investigation on the causes of codon and amino acid usages variation between thermophilic Aquifex aeolicus and mesophilic Bacillus subtilis. Journal of Biomolecular Structure & Dynamics, 22, 205–214.

Boone, D. R., Castenholz, R. W., & Garrity, G. M. (2001). Bergey’s manual of systematic bacteriology (George M. Garrity, editor-in-chief). New York, NY: Springer.

Chanda, I., Pan, A., & Dutta, C. (2005). Proteome composition in Plasmodium falciparum: Higher usage of GC-rich nonsynonymous codons in highly expressed genes. Journal of Molecular Evolution, 61, 513–523.

Chauhan, N., Vidyarthi, A. S., & Poddar, R. (2011). Comparative multivariate analysis of codon and amino acid usage in three Leishmania genomes. Genomics Proteomics Bioinformatics, 9, 218–228.

Chen, L. L., Chung, W. C., Lin, C. P., & Kuo, C. H. (2012). Comparative analysis of gene content evolution in phytoplasmas and mycoplasmas. PLoS ONE, 7, e34407.

Das, S., Paul, S., Chatterjee, S., & Dutta, C. (2005). Codon and amino acid usage in two major human pathogens of genus Bartonella—optimization between replicational-transcriptional selection, translational control and cost minimization. DNA Research, 12, 91–102.

Dougherty, D. A. (2007). Cation-pi interactions involving aromatic amino acids. Journal of Nutrition, 137, 1504S–1508S discussion 1516S–1517S.

Duret, L., & Mouchiroud, D. (2000). Determinants of substitution rates in mammalian genes: Expression pattern affects selection intensity but not mutation rate. Molecular Biology and Evolution, 17, 68–74.

Gallivan, J. P., & Dougherty, D. A. (1999). Cation-pi interactions in structural biology. Proceedings of the National Academy of Sciences, 96, 9459–9464.

Garat, B., & Musto, H. (2000). Trends of amino acid usage in the proteins from the unicellular parasite Giardia lamblia. Biochemical and Biophysical Research Communications, 279, 996–1000.

Greenacre, M. J. (1984). Theory and applications of correspondence analysis. London: Academic Press.

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.

Iriarte, A., Bariabau, J. D., Romero, H., & Musto, H. (2011). Selected codon usage bias in members of the class Mollicutes. Gene, 473, 110–118.

Kahali, B., Basak, S., & Ghosh, T. C. (2007). Reinvestigating the codon and amino acid usage of S. cerevisiae genome: A new insight from protein secondary structure analysis. Biochemical and Biophysical Research Communications, 354, 693–699.
Karlin, S., & Bucher, P. (1992). Correlation analysis of amino acid usage in protein classes. *Proceedings of the National Academy of Sciences*, 89, 12165–12169.

Lazarev, N. V., Leviatskii, S. A., Basovskii, Y. I., Chukin, M. M., Akopian, T. A., Veshchagin, V. V., ... Govorun, V. M. (2011). Complete genome and proteome of Acholeplasma laidlawii. *Journal of Bacteriology*, 193, 4943–4953.

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

McDonald, J. H. (2010). Temperature adaptation at homologous sites in proteins from nine thermophile-mesophile species pairs. *Genome Biology and Evolution*, 2, 267–276.

Musto, H., Romero, H., & Zavala, A. (2003). Translational selection is operative for synonymous codon usage in *Clostridium perfringens* and *Clostridium acetobutylicum*. *Microbiology*, 149, 855–863.

Naya, H., Zavala, A., Romero, H., Rodriguez-Maseda, H., & Musto, H. (2004). Correspondence analysis of amino acid usage within the family Bacillaceae. *Biochemical and Biophysical Research Communications*, 325, 1252–1257.

Oshima, K., & Nishida, H. (2007). Phylogenetic relationships among mycoplasmas based on the whole genomic information. *Journal of Molecular Evolution*, 65, 249–258.

Pal, C., Papp, B., & Hurst, L. D. (2001). Highly expressed genes in yeast evolve slowly. *Genetics*, 158, 927–931.

Pal, C., Papp, B., & Hurst, L. D. (2003). Genetic function: Rate of evolution and gene dispensability. *Nature*, 421, 496–497; discussion 497–498.

Palacios, C., & Wernegreen, J. J. (2002). A strong effect of AT mutational bias on amino acid usage in Buchnera, endosymbiotic bacteria of aphids. *Genome Research*, 14, 44–53.

Rocha, E. P., & Danchin, A. (2004). An analysis of determinants of amino acids substitution rates in bacterial proteins. *Molecular Biology and Evolution*, 21, 108–116.

Sabbia, V., Piovani, R., Naya, H., Rodriguez-Maseda, H., Romero, H., & Musto, H. (2007). Trends of amino acid usage in the proteins from the human genome. *Journal of Biomolecular Structure & Dynamics*, 25, 55–59.

Schaber, J., Rispe, C., Wernegreen, J., Buness, A., Delmotte, F., Silva, F. J., & Moya, A. (2005). Gene expression levels influence amino acid usage and evolutionary rates in endosymbiotic bacteria. *Gene*, 352, 109–117.

Sharp, P. M., Bailes, E., Grocock, R. J., Peden, J. F., & Sneath, R. E. (2005). Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Research*, 33, 1141–1153.

Singer, G. A., & Hickey, D. A. (2000). Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. *Molecular Biology and Evolution*, 17, 1581–1588.

Subramanian, S., & Kumar, S. (2004). Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. *Genetics*, 168, 373–381.

Supek, F., Skunca, N., Repar, J., Vlahovicek, K., & Smuc, T. (2010). Translational selection is ubiquitous in prokaryotes. *PLoS Genetics*, 6, e1001004.

Supek, F., & Vlahovicek, K. (2004). INCA: synonymous codon usage analysis and clustering by means of self-organizing map. *Bioinformatics*, 20, 2239–2230.

Supek, F., & Vlahovicek, K. (2005). Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity. *BMC Bioinformatics*, 6, 182.

Suzuki, R., & Shimodaira, H. (2006). Pvclust: An R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, 22, 1540–1542.

Swire, J. (2007). Selection on synthesis cost affects interprotein amino acid usage in all three domains of life. *Journal of Molecular Evolution*, 64, 558–571.

Tekaia, F., & Yeramian, E. (2006). Evolution of proteomes: Fundamental signatures and global trends in amino acid compositions. *BMC Genomics*, 7, 307.

Tekaia, F., Yeramian, E., & Dujon, B. (2002). Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: A global picture with correspondence analysis. *Gene*, 297, 51–60.

Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22, 4673–4680.

Wang, G. Z., & Lercher, M. J. (2010). Amino acid composition in endothermic vertebrates is biased in the same direction as in thermophilic prokaryotes. *BMC Evolutionary Biology*, 10, 263.

Woese, C. R., Maniloff, J., & Zablen, L. B. (1980). Phylogenetic analysis of the mycoplasmas. *Proceedings of the National Academy of Sciences*, 77, 494–498.

Wolf, M., Muller, T., Dandekar, T., & Pollack, J. D. (2004). Phylogeny of Firmicutes with special reference to Mycoplasma (Mollicutes) as inferred from phosphoglycerate kinase amino acid sequence data. *International Journal of Systematic and Evolutionary Microbiology*, 54, 871–875.

Yamao, F., Andachi, Y., Muto, A., Ikemura, T., & Osawa, S. (1991). Levels of tRNAs in bacterial cells as affected by amino acid usage in proteins. *Nucleic Acids Research*, 19, 6119–6122.

Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24, 1586–1591.

Zavala, A., Naya, H., Romero, H., & Musto, H. (2002). Trends in codon and amino acid usage in Thermotoga maritima. *Journal of Molecular Evolution*, 54, 563–568.