An assessment of the impacts of molecular oxygen on the evolution of proteomes.

Sara Vieira-Silva, Eduardo P. C. Rocha

An Assessment of the Impacts of Molecular Oxygen on the Evolution of Proteomes

Sara Vieira-Silva† and Eduardo P. C. Rocha*†*

*Atelier de BioInformatique, Université Pierre et Marie Curie-Paris 6, Paris, France; and †Microbial Evolutionary Genomics, Institut Pasteur, CNRS, URA2171, Paris, France

Oxygen is not only one of life's essential elements but also a source of protein damage, mutagenesis, and ageing. Many proteome adaptations have been proposed to tackle such stresses and we assessed them using comparative genomics in a phylogenetic context. First, we find that aerobiosis is a trait with important phylogenetic inertia but that oxygen content in proteins is not. Instead, oxygen content is close to the expected values given the nucleotide composition. Accordingly, we find no evidence of oxygen being a scarce resource for protein synthesis even among anaerobes. Second, we searched for counterselection of amino acids more prone to oxidation among aerobes. Only cysteine follows the expected trend, whereas tryptophan follows the inverse one. When analyzing composition in the context of protein structures and residue accessibility, we find that all oxidable residues are avoided at the surface of proteins. Yet, there is no difference between aerobes and anaerobes in this respect, and the effect might be explained by the hydrophobicity of these residues. Third, we revisited the hypothesis that atmospheric enrichment in molecular oxygen led to the development of the communication capabilities of eukaryotes. With a larger data set and adequate controls, we confirm the trend of longer oxygen-rich outer domains in transmembrane proteins of eukaryotes. Yet, we find no significant association between oxygen concentration in the environment and this trait within prokaryotes, suggesting that this difference is clade specific and independent of oxygen availability. We find that genes involved in cellular responses to oxygen are much more frequent among aerobes, and we suggest that they erase most expected differences in terms of proteome composition between organisms facing high and low oxygen concentrations.

Introduction

Comparative studies have uncovered compositional biases in protein sequences, which have been attributed to a wide range of causes. First, amino acid composition is strongly constrained by the nucleotide composition of the genome (Sueoka 1962; Lobry 1997; Singer and Hickey 2000). Genomic G + C contents vary between less than 20% and more than 70% resulting in the differentiated use of amino acids due to the structure of the genetic code (Muto and Osawa 1987). Although some amino acid changes are neutral, very few are adaptive and many, or most, are deleterious. Therefore, amino acid composition results from mutation-selection balance caused by the antagonism between mutational biases and the selective pressure to maintain protein function and structural stability. Apart from mutational patterns driven by the replication and repair machineries, a number of systematic amino acid compositional biases have been described and interpreted as signatures of selective pressures. 1) Amino acids have very different biosynthetic costs, and expensive amino acids are selectively purged from proteomes, especially among highly expressed genes (Akashi and Gojobori 2002; Hurst et al. 2006). 2) Proteins involved in the biosynthetic pathways of amino acids are activated in times of amino acid starvation, and they are selectively purged of these cognate amino acids to enhance metabolic efficiency (Baudouin-Cornu et al. 2001; Alves and Savageau 2005). 3) Amino acids differ in the amount and type of elements they contain. Some elements are scarce and hard to scavenge, like sulfur, and their frequency in proteins depends on their availability in the environment (Mazel and Marliere 1989; Bragg et al. 2006; Bragg and Wagner 2007). 4) Growth temperatures vary between less than 0 °C and more than 100 °C. Because proteins are sensitive to denaturation at high temperatures, adaptation to high temperature involves the selection of some amino acids over others (Kreil and Ouzounis 2001; Tekaia and Yeramian 2006; Zeldovich et al. 2007). Many of these effects highlight the emerging view that the proteome amino acid composition is influenced by external factors. Knowing such factors is essential to achieve a better understanding of the evolution of genomes and proteins.

Oxygen is the most abundant element on earth and is a major constituent of water, carbon hydrates, lipids, proteins, and nucleic acids. Although atmospheric molecular oxygen kills some organisms, it is essential to the life of many others. This, along with the chemical reactivity of many of its derivatives, makes oxygen an ideal candidate to study constraints on protein composition both from the point of view of inducer of protein damage as well as a potential scarce resource to some organisms. Although molecular oxygen makes more than 20% of today's atmosphere, the ancestral atmosphere lacked oxygen and was reducing for the early part of the life's history (Goldblatt et al. 2006). The so-called Great Oxidation Event is thought to have resulted from the invention of photosystem II by cyanobacteria between 2.7 and 1.9 billion years ago and raised the atmospheric level of oxygen to the present level from less than 10^{-5} times that value (Bekker et al. 2004). Since then, some organisms have remained anaerobes and typically very sensitive to molecular oxygen, whereas most have developed aerobic respiration.

The switch from a reductive to an oxidative atmosphere after the evolution of oxygenic photosynthesis must have been a massive source of environmental stress. Indeed, most contemporary obligate anaerobes are incapacitated by oxidative stress because their metabolic pathways still rely on enzymes that react with oxidants. On the other hand, aerobes have adapted them to mitigate the toxicity of oxygen derivatives, mostly H_2O_2 and O_2 radicals (Imlay 2002).

Key words: oxidative stress, cysteine, protein evolution, hydrophobicity, evolution.

E-mail: sara@abi.snv.jussieu.fr.

Mol. Biol. Evol. 25(9):1931–1942. 2008
doi:10.1093/molbev/msn142
Advance Access publication June 25, 2008

Toxicity results from lipid peroxidation, oxidative damage of DNA bases resulting in mutations and, in what proteins are concerned, oxidative cleavage of the polypeptide chain, modification of amino acid side chains, generation of protein–protein cross-linkage, and formation of derivatives sensitive to proteolytic degradation (Stadtman 2006). Molecular oxygen diffuses freely through membranes, and the concentrations of H_2O_2 and O_2 radicals increase with the concentration of O_2 (Imlay 2002). Therefore, there is a direct positive correlation between atmospheric oxygen concentration and oxidative stress intensity inside the cells. One might then expect that the amino acids most susceptible to oxidation (histidine, tryptophan, methionine, tyrosine, and cysteine; Davies and Truscott 2001) would be avoided in oxidizing circumstances. This effect should be more prominent on amino acids subject to irreversible oxidation (histidine, tryptophan, and tyrosine) than on amino acids capable of reversible oxidation (methionine and cysteine). However, cysteine participates in Fe–S complexes that are very susceptible to oxidation (Imlay 2006), and its excess contributes to DNA damage by creating hydroxyl radicals (Park and Imlay 2003). On the other hand, it has been speculated that surface-exposed methionines could serve as a pool of targets to protect functionally essential residues from oxidation and limit the effects of oxygen on ageing (Levine et al. 2000). Therefore, the adaptation of proteomes to oxygen-rich environments should lead to adaptive cysteine avoidance and possibly methionine overrepresentation at the surface of proteins.

Because oxygen diffuses through the membranes to enter the cell, membrane proteins are expected to show signs of adaptation to high oxygen concentrations. Transmembrane helices of proteins of eukaryotes have longer oxygen-rich outer domains than those of prokaryotes (Acquisti et al. 2007). This led to suggestions that the rise of molecular oxygen concentration allowed the expansion of these domains and led to the development of communication-related transmembrane proteins among eukaryotes (Acquisti et al. 2007). The ancestral reducing atmosphere would have made long oxygen-rich outer domains structurally unstable and expensive, given oxygen scarcity (Acquisti et al. 2007). It has been further hypothesized that the rise in atmospheric oxygen concentration should have inversed the situation and favored the use of oxygen-rich amino acids because they are metabolically less expensive. This hypothesis sustains that oxygen was a source of adaptive opportunities that was used to the advantage of the arising eukaryotic lineages but not the extant prokaryotic ones (Acquisti et al. 2007). This would only be possible if the oxygen content of protein had such high evolutionary inertia that anciently acquired proteomic compositional signatures could be maintained until the present day. If one accepts the frequently held idea that eukaryotes derive from common ancestors with prokaryotes, this hypothesis leaves unsolved how oxygen-rich proteins in eukaryotes arose in the first place.

The availability of hundreds of genomes that can be classed according to aerobiosis, photosynthesis capacity, or growth rates, allows testing the different hypotheses put forward for the adaptation of proteomes to oxygen-rich environments (McCord et al. 1971; Imlay 2002; Towe 2003; Hedges et al. 2004; Zeilstra-Ryalls and Kaplan 2004; Acquisti et al. 2007): 1) Is oxygen content in proteins a trait with important phylogenetic inertia or does it closely follow mutational biases? 2) Is aerobiosis itself a trait frequently lost? 3) Is there evidence for oxygen being a scarce resource? 4) Are oxidable amino acids counterselected in aerobes? 5) Is the exposition of the oxidable amino acids in the protein structure important in the adaptation to aerobiosis? 6) Is there an association between oxygen content in transmembrane helices and oxygen availability? For this purpose, we classed 306 different species of prokaryotes and eukaryotes according to capacity for aerobiosis, photosynthesis, and rapid growth. We then analyzed their complete proteomes controlling for the influence of mutational biases and phylogenetic nonindependence.

Material and Methods

Genome and Proteome Data

We retrieved 295 prokaryotic genomes, 1 per species, from GenBank Genomes (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). We also retrieved the genomes of 11 eukaryotic species from Ensembl release 44 (http://www.ensembl.org) and PlasmoDB (http://plasmodb.org/) (see supplementary table 1 [Supplementary Material online] for a comprehensive listing). Genes and proteins were extracted from annotation data and pseudogenes were ignored. Organism's classification by oxygen requirements (anaerobic, aerobic, and photosynthetic) and minimal doubling times were taken from the literature or obtained by personal communication from researchers in the field (Couturier and Rocha 2006). From the 295 prokaryotic species, 231 were classed as aerobes (including aerobes, microaerophiles, or facultative aerobes), 53 as anaerobes, and 11 as photosynthetic (see supplementary fig. 1, Supplementary Material online).

Transmembrane Protein Prediction and Protein Topology Prediction

The transmembrane proteins (tm-proteins) were extracted from the full proteome of each organism using TMHMM, a method for prediction of protein topology based on a hidden Markov model (transmembrane vs. other domains) (Krogh et al. 2001). From the results of TMHMM, we retrieved the topology of each tm-protein for every organism. Internal, outer, and transmembrane domains were then extracted from each tm-protein and separately concatenated in 3 sets corresponding to the 3 localizations (in, out, and tm) for further analysis.

G + C Content, Oxygen Content, and Amino Acid Frequencies

The G + C content for the 306 genomes was calculated using only coding sequences. This allows comparing similar biological objects, that is, genes, thus accounting for different gene density in genomes. It also reduces the problem of heterogeneous G + C composition within genomes, which albeit present in bacteria (Daubin and Perriere 2003) is particularly important in eukaryotes (Bernardi et al. 1985).

The relative oxygen content of the proteome was calculated as the ratio between the number of oxygen atoms in the side chains and the number of residues in the proteins. Based on the prediction of topology by TMHMM, the oxygen content of the internal, outer, and transmembrane domains was also calculated for each tm-protein. Each amino acid frequency was calculated as the percentage of that amino acid in the total proteome. Methionine frequency was calculated excluding methionines on the first position of protein sequences. Amino acids were classified according to several chemical properties for subsequent analysis: 1) oxygen-rich amino acids (aspartic acid D, glutamic acid E, asparagine N, glutamine Q, serine S, threonine T, and tyrosine Y) and 2) amino acids prone to oxidation (cysteine C, histidine H, tryptophan W, methionine M, and tyrosine Y).

Gene Expression Levels

There is no available data for the levels of gene expression in the vast majority of prokaryotes. Thus, we identified highly expressed genes using codon usage bias. Because this bias is only a good proxy of gene expression for fast-growing bacteria (Couturier and Rocha 2006), we restricted such analyses to these genomes. This resulted in a sample of 57 prokaryotic species with a minimal doubling time lower or equal to 2.5 h, 47 of which aerobic and 10 anaerobic (see supplementary table 1, Supplementary Material online). For these species, we used the codon adaptation index (CAI) to evaluate codon usage bias, as frequently done to assess gene expression (Sharp and Li 1987; Coghlan and Wolfe 2000). The CAI was calculated using the ribosomal genes as a reference set for highly expressed genes. We retrieved from the genome the 5%, 5–10%, and 10–15% of highest CAI values as relevant categories of the top 5%, 5–10%, and 10–15% most highly expressed genes in the genome.

Prediction of Accessibility

We built a data set of proteins that are ubiquitous among a group of genomes so that we could analyze the relevance of residue accessibility in terms of adaptation to the presence of molecular oxygen. We defined a set including 28 anaerobes and 108 aerobes (including *Escherichia coli* MG1655), selected to achieve a maximum number of ubiquitous orthologs along with a minimal reduction of the number of genomes in the data set (particularly among anaerobes that are rarer). The putative ubiquitous orthologs were defined as bidirectional best hits (BBH) between the proteome of *E. coli* as a pivot and the other proteomes, showing more than 40% sequence similarity and less than 20% difference in protein length. We then excluded ribosomal proteins because their accessibility depends on the large ribosomal complex and cannot be meaningfully computed for the subunit alone. We identified 34 such ubiquitous orthologs within these 136 species, which we aligned with MUSCLE (Edgar 2004). We assumed that such putative orthologs, because they have more than 40% sequence similarity, should have similar 3-dimensional (3D) structures. Indeed, for sequences with more than 20% identity, a high score of structure similarity is observed (Chothia and Lesk 1986; Krissinel 2007). Therefore, even if our definition of orthology by BBH, which implies transitivity, is occasionally violated, the observed sequence similarity is enough to consider that they have similar structures. Solvent accessibility was predicted using the SABLE software (Adamczak et al. 2005) on the 34 *E. coli* orthologs, and the predictions were attributed to the positions in the 34 multiple alignments, thus extending the predictions to the orthologs in all the other species. This is justified because at this level of sequence similarity, structural features are typically conserved. Also, because there are many proteins crystallized in *E. coli*, we could use this information to benchmark SABLE. To assess the accuracy of the predictions, we analyzed among the 34 orthologs the 29 for which there was an available 3D structure in the PDB database for *E. coli* (for a listing, see supplementary table 2, Supplementary Material online). The accessibility to solvent (SAobserved) of the residues of these 29 proteins was calculated using the "sasa" function of VMD software with a probe of 1.4 Å (radius of the water molecule) (Humphrey et al. 1996). If SAobserved < 60 Å^2, the residue was considered buried, and if SAobserved > 60 Å^2, the residue was considered exposed (Sasidharan and Chothia 2007). We made a discretization of the solvent accessibility results (SApredicted) of SABLE setting the boundary at 3, to respect the proportion of total accessible amino acids in the 29 proteins. If SApredicted < 3, the residue was considered buried, and if SApredicted ≥ 3, the residue was considered exposed. We then computed the association between the sasa and the SABLE results. One must emphasize that most of these structures are only partial 3D structures (although we excluded structures when they corresponded to less than 70% of the protein). Partial structures will suggest that some regions are exposed, when in fact, they are buried by parts of the protein that are not included in the crystal. This increases the difference between the 2 methods and thus underestimates accuracy for exposed residues. A total of 9,710 residues were defined as "buried" or "exposed" by both SAobserved and SApredicted classification and 76% of those were matches (53% expected randomly from buried and exposed proportions). Contrary to exposed residuals, if a residual is buried in the incomplete PDB crystal, then it is also buried in the complete crystal. If we restrict the evaluation only to these buried amino acids in the PDB file, the SABLE prediction shows an accuracy of 81%. Thus, SABLE seems to be a reliable predictor of solvent accessibility.

The average amino acid composition of proteomes strongly depends on genome GC composition. To correct for this effect, the frequency of exposed amino acids was calculated as the ratio of the number of exposed amino acid over the total number of that amino acid in the concatenation of all the 34 orthologs of each species.

Phylogenetic Analyses

The 16S rDNA Global Tree

We extracted the sequences of the 16S rDNA genes from the 295 prokaryotic genomes and aligned them using

MUSCLE (Edgar 2004). Poorly aligned regions were removed using Gblocks (Talavera and Castresana 2007), allowing a maximum of 8 nonconserved contiguous positions and minimum block size of 40. The phylogenetic tree was then reconstructed by maximum likelihood using PHYML (Guindon and Gascuel 2003), with the HKY + Γ(8) model. The pairwise distances, that is, the expected number of substitutions per site separating 2 sequences, were then extracted from the phylogenetic tree using the "ape" package from R (Paradis et al. 2004; R Development Core Team 2007) (see supplementary fig. 1, Supplementary Material online).

Accessibility Data Tree

The 34 alignments of orthologs of the data set for the prediction of accessibility (136 prokaryotes) were concatenated into one long alignment. Poorly aligned regions were removed using Gblocks (Talavera and Castresana 2007), allowing a maximum of 8 nonconserved contiguous positions and minimum block size of 40. The distance matrix was then computed by maximum likelihood using Puzzle (Schmidt et al. 2002), with the JTT + Γ(4) model. The phylogenetic tree was reconstructed from the distance matrix using the Neighbour-Joining method QuickTree (Howe et al. 2002).

Phylogenetic Inertia—Generalized Estimation Equations

Although most variables analyzed in this work have a weak phylogenetic inertia, some of the genomes are very close whereas others are very distant. This phylogenetic dependency poses a statistical inference bias. We therefore performed comparative analysis of amino acid frequencies in anaerobes and aerobes taking into account their phylogenetic relationship. This was performed using generalized estimation equations (GEE) computed in R and implemented in the ape package (Paradis and Claude 2002).

Estimation of the Parameters of a Mixture of 2 Normal Distributions

The bimodal distribution of the oxygen content of the tm-proteins was previously observed but not precisely deconvoluted (Acquisti et al. 2007). We assumed that the bimodal distribution of the oxygen content of the tm-proteins is a mixture of 2 normal distributions. We then extracted the parameters (mean and standard deviation [SD]) of the 2 Gaussians and the proportion of the mixture using maximum likelihood estimators and the Broyden–Fletcher–Goldfarb–Shanno method of optimization, computed in R (Venables and Ripley 2002).

Results and Discussion

Aerobiosis Is a Trait with Important Phylogenetic Inertia

We assessed the phylogenetic inertia associated to oxygen requirement for respiration. For this purpose, we analyzed the association between the phylogenetic distance between pairs of prokaryotic species and the frequency with which a pair had the same or opposite oxygen requirements,

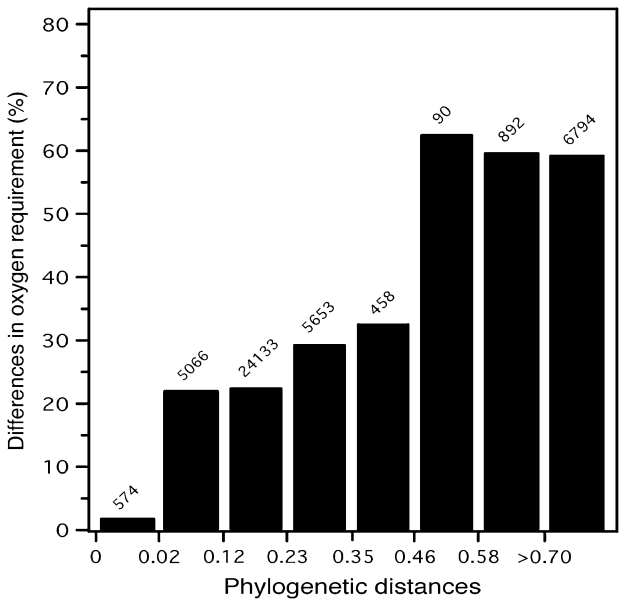

FIG. 1.—Difference in oxygen requirement between species in function of their phylogenetic distance. For each bin of phylogenetic distances among pairs of genomes, we plot the frequency with which both elements of the pair have different oxygen requirements. A pair was regarded as different if one was an anaerobe and the other an aerobe. With the exception of the first bin, designed to be upper limited by the phylogenetic distance between *Escherichia coli* and *Yersinia pestis*, the bins were selected to represent equally spaced phylogenetic distances. Hence, the number of pairs (*N*) in each bin is not equal. *N* is given on top of each bin.

that is, if the pair was similar or not in terms of aerobiosis/anaerobiosis (fig. 1). There is a very clear difference between very closely related species, with an evolutionary distance under 0.02 and distant ones over 0.46. The first threshold corresponds to the evolutionary distance between *E. coli* and *Yersinia pestis* and shows that closely related pairs are more concordant in oxygen requirement (<2% difference), compared with intermediate pairs (<30% difference) and the distant ones (>50% difference). This last group corresponds to very distant comparisons, mostly between bacteria and archaea. Therefore, phylogenetic inertia of this trait is important and leaves an imprint at very distant phylogenetic scales. There are few cases of pairs of closely related bacteria being differently classed relative to aerobiosis. Among these, one finds among the γ-proteobacteria *Haemophilus ducreyi* (anaerobe) and *Haemophilus influenzae* (facultative aerobe), the α-proteobacteria *Zymomonas mobilis* subsp. *mobilis* (anaerobe) and *Erythrobacter litoralis* (aerobe), and the Chlorobi *Chlorobium chlorochromatii* (anaerobe) and *Chlorobium phaeobacteroides* (facultative aerobe). The small number of available closely related genomes showing different oxygen requirements precludes for the moment the detailed analysis of changes associated with this major biochemical transition.

Oxygen Content in Proteins Is a Trait with Weak Phylogenetic Inertia

Historical hypotheses establishing adaptive links between the rise of atmospheric molecular oxygen and protein

composition in oxygen assume that the latter trait has a strong phylogenetic inertia. Otherwise, the initial conditions, that is, the amino acid compositions of proteomes at the time, are of little importance and adaptation can be done in parallel in each lineage. To test this hypothesis, we turned to the standard neutral theory result that amino acid composition will follow closely the genomic G + C content except if selection is strong (Sueoka 1988). We thus analyzed the correlation between the proteomic oxygen content in a set of 306 prokaryotic and eukaryotic species with their respective genomic G + C contents (fig. 2). As expected under a trait that is evolving largely by drift, we found that the 2 variables are highly correlated with higher genomic G + C values associated with less oxygen in proteins (Pearson coefficient $r = -0.82$, $P < 0.0001$). G + C content is a labile trait, often with very weak phylogenetic inertia. For example, within γ-proteobacteria, it varies from less than 25% to over 65%. Its excellent correlation with proteomic oxygen content strongly suggests that the latter follows closely the genomic compositional bias, frequently thought to reflect mutational bias. It is therefore important to be careful in making assumptions based on the comparative analysis of the proteomic oxygen content of a few species, which could present significant differences not due to evolutionary pressures, but to nucleotide composition. This does not preclude selection on protein oxygen content, but it disproves the claim that this trait evolves very slowly.

Because G + C content is such a strong determinant of oxygen content in proteins, we controlled for this effect throughout our work. Although we found no association between G + C content and aerobiosis in our data set ($P = 0.08$, ANOVA and $P > 0.3$, GEE), it has been suggested that aerobic prokaryotes display significantly higher genomic G + C content than anaerobic ones (Naya et al. 2002). Controlling by the G + C has the advantage of controlling for such an effect. For each genome, we computed observed/expected values of oxygen content, where the expected value was calculated as expected = 0.56 − 0.0022 %GC ($R^2 = 0.68$, $P < 0.0001$). An equivalent correction was done for the frequency of each amino acid in proteomes respecting their individual association with G + C content (supplementary table 3, Supplementary Material online).

Five outliers stand out in figure 2. The first one, *Plasmodium falciparum*, has one of the GC-poorest genomes, leading to a proteome enriched in amino acids encoded by AT-rich codons. The most frequent ones are lysine (K) and asparagine (N), which both represent 20% of all amino acids in that proteome (Musto et al. 1995). Asparagine is an oxygen-rich amino acid and is responsible for the higher oxygen content of the proteome of *P. falciparum* in relation to the other eukaryotes. The other 4 outliers are all halophilic bacteria and archaea. Aspartic acid enrichment is largely responsible for the higher oxygen content of their proteomes in comparison to the other prokaryotes ($R^2 = 0.28$ and $P < 0.0001$, ANOVA). This confirms a previously described compositional bias in halophilic archaea that might be associated with the superior water-binding abilities of aspartic acid (Fukuchi et al. 2003). It is noticeable from the data in figure 2 that eukaryotes do have more oxygen in their proteins than prokaryotes. Indeed, there

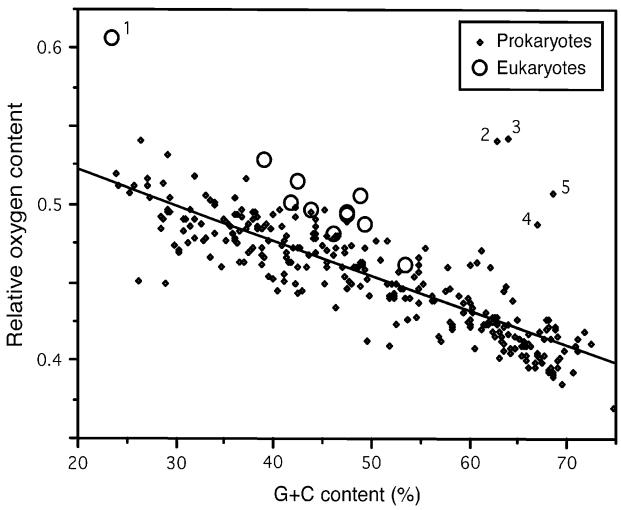

FIG. 2.—Side-chain oxygen atoms per residue versus genome G + C percentage. The relative oxygen content correlates linearly and negatively with G + C content of the genome of 11 eukaryotic species (open circles) and 295 prokaryotic species (dots). Linear regression: $Y = 0.56 - 0.0022X$, $R^2 = 0.68$, $P < 0.0001$. Outliers: 1, *Plasmodium falciparum*; 2, *Haloarcula marismortui*; 3, *Natronomonas pharaonis*; 4, *Salinibacter ruber*; and 5, *Halobacterium* sp.

is a significant difference between the mean observed/expected oxygen content of eukaryotes and prokaryotes ($R^2 = 0.12$ and $P < 0.0001$, both for ANOVA and Wilcoxon test). This confirms, within a much larger data set, previous findings that had been interpreted as a signature of adaptation of eukaryotes to the rise of atmospheric oxygen concentrations (Acquisti et al. 2007). However, we cannot subscribe this interpretation of the results because the reactivity of oxygen content to changes in G + C content suggests that such an old effect would have been erased by drift in a trait with such weak phylogenetic inertia. Furthermore, aerobic prokaryotes could have also profited from the same adaptive opportunities and thus should have higher oxygen content than anaerobes. Instead, there is no significant difference in terms of average observed/expected oxygen content between aerobes and anaerobes of prokaryotic origin ($P > 0.15$, ANOVA; supplementary fig. 2, Supplementary Material online).

It has been suggested that the oxygen content of tm-proteins evolves slowly in response to oxygen availability in the environment (Acquisti et al. 2007). Using the tm-protein topology prediction of TMHMM, we analyzed the correlation between the proteomic oxygen content and the genomic G + C content separately for tm-proteins and nontransmembrane proteins (nontm-proteins). We found that for internal (in), outer (out), and transmembrane (tm) helices domains of tm-proteins (supplementary fig. 3, Supplementary Material online), there were similar linear and negative high correlations between oxygen composition and G + C content (Pearson coefficients: $r(\text{in}) = -0.82$, $r(\text{out}) = -0.80$, $r(\text{tm}) = -0.89$; $P < 0.0001$). Thus, both in general and in the particular case of tm-proteins, the frequency of atomic oxygen follows closely the expected patterns from genome composition.

Oxygen Is Not a Scarce Resource

Practically, all chemical elements composing amino acids have been proposed to be a scarce resource under some circumstances. As a consequence, their availability might limit protein composition (Baudouin-Cornu et al. 2001; Bragg and Hyder 2004; Bragg et al. 2006; Acquisti et al. 2007; Bragg and Wagner 2007). Yet, evidence for scarce oxygen has not been tested, and it is important to know if even the most abundant element on the planet can be regarded as scarce in some circumstances. If, as proposed (Acquisti et al. 2007), molecular oxygen is a scarce resource for the biosynthesis of amino acids, then anaerobes should avoid oxygen-rich amino acids more than aerobes. Indeed, aerobes have slightly more oxygen-rich amino acids (DENQSTY) than anaerobes. The difference between the 2 groups when accounting for expected values given by G + C content is significant, but only barely so ($P = 0.048$, Wilcoxon test), and it accounts for less than 2% of overall variance ($R^2 = 0.018$, ANOVA). Furthermore, this difference proves not to be significant ($P = 0.83$) after controlling for phylogenetic dependence using GEE with a phylogenetic tree based on 16S rDNA (see Material and Methods). Thus, the enrichment of oxygen in proteins of aerobes is very small and cannot be distinguished from a sampling artifact due to phylogenetic dependencies. These results do not provide substantial evidence for the hypothesis that oxygen is a scarce resource among anaerobes.

Most oxygen-rich amino acids are also among the least expensive to synthesize (Akashi and Gojobori 2002). Oxygen is not expected to be a scarce resource in aerobes. Therefore, these amino acids are expected to be overrepresented in highly expressed genes, as is generally the case for inexpensive amino acids (Akashi and Gojobori 2002; Swire 2007). In anaerobes either oxygen is scarce and the trend should be inversed to save oxygen in highly expressed genes or it is not scarce and the trend should be similar to the one of aerobes after accounting for mutational biases and phylogenetic dependency. Therefore, we made a complementary analysis of the less costly amino acids (AG, both oxygen poor) and the following 6 less costly amino acids (DENQST, all oxygen rich). We excluded tyrosine (Y) from the oxygen-rich amino acids because it has a much higher biosynthetical cost (Akashi and Gojobori 2002) (supplementary fig. 4, Supplementary Material online). We then used the CAI to select for highly expressed genes (Sharp and Li 1987). Because codon usage bias only reflects selection for translation of highly expressed genes in fast-growing bacteria (Rocha 2004), we restricted our analyses to these (see Material and Methods). As expected, the relative frequency of the less costly amino acids (AG) shows a monotonic relation to gene expressivity, with these amino acids being overrepresented in highly expressed genes and then progressively rarer for both aerobes ($R^2 = 0.24$, $P < 0.0001$, $N = 47$) and anaerobes ($R^2 = 0.20$, $P = 0.04$, $N = 10$). On the contrary, the relative frequency of oxygen-rich amino acids (DENQST) shows no monotonic relation to gene expressivity, with the most highly expressed genes being less biased than intermediately expressed genes and not much more biased than the least expressed genes (fig. 3). Strikingly, and contradicting the hypothesis that oxygen is a scarce resource, the trend is exactly the same in both aerobes and anaerobes. Although these oxygen-rich amino acids are cheaper than the average amino acid, there is no significant overrepresentation of these amino acids in highly expressed genes of aerobes, relative to anaerobes. Therefore, there is no significant evidence of differences in atomic oxygen composition of proteins between organisms with ample access to molecular oxygen (aerobes) and organisms that have practically no contact with molecular oxygen (anaerobes). Oxygen does not seem to be a scarce resource in any case.

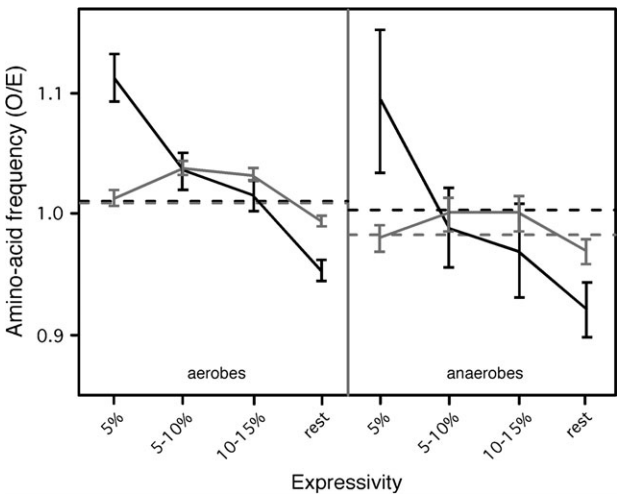

FIG. 3.—Relative frequency of amino acids in proteomes (after G + C content correction) versus expressivity. Expressivity was measured using the CAI and the bins correspond to the top 5%, 5–10%, 10–15% most highly expressed genes and the remaining genes. The 10 anaerobes and 47 aerobes with maximal growth rates under 2.5 h were analyzed. Dashed lines represent mean proteomic content of AG and DENQST for anaerobes and aerobes separately. These values are not significantly different after phylogenetic correction. Bars represent standard errors.

The Frequency of Cysteine Depends on Molecular Oxygen Concentration

Oxidative stress is proportional to the concentration of O_2 in the immediate surrounding environment. It is therefore expected that anaerobic species would be subject to less oxidative stress than aerobic species (McCord et al. 1971). Therefore, the most oxidable amino acids (in decreasing order: cysteine, histidine, tryptophan, methionine, and tyrosine; Davies and Truscott 2001) should be rarer in aerobic species. An association between cysteine depletion and oxidative stress has previously been shown without controlling for genomic G + C content (Moosmann and Behl 2008). Yet, the establishment of a causal association between selection against oxidable amino acids and oxygen abundance requires a common pattern among all oxidable amino acids, not just one. Otherwise, avoidance of cysteine might be ascribed to any other property of this chemically unusual amino acid, notably its unique potential to form disulphide bonds, which is both fundamental for protein stability and damage.

For each one of the 20 amino acids, we did a comparative analysis of each amino acid's proteomic frequency (after G + C content correction) between anaerobes and aerobes using the data set of 295 prokaryotes. GEE were used in order to correct for phylogenetic dependence in the sample with the 16S rDNA tree. To correct for multiple testing, we used a sequential Bonferroni correction. Only tryptophan (W), cysteine (C), and threonine (T) show a significant difference in distribution between anaerobes and aerobes (fig. 4, $P(\mathrm{W}) < 0.0001$, $P(\mathrm{C}) < 0.001$, and $P(\mathrm{T}) < 0.01$). Because cysteine and tryptophan are oxidable amino acids, we expect them to be avoided in aerobes. Cysteine does follow this expected trend and is more frequent in anaerobes ($R^2 = 0.07$ and $\mathrm{P} < 0.0001$, ANOVA). However, tryptophan follows the opposite trend, being rarer in anaerobes ($R^2 = 0.05$ and $P < 0.0001$, ANOVA). Tryptophan is the most expensive amino acid (Akashi and Gojobori 2002), and this result may reflect the lower energetic efficiency of nonoxygenic respiration, which may impose stronger selection against highly expensive amino acids. Overall, there is no systematic trend for oxidable amino acids to be avoided in aerobes as a result of adaptation to high molecular oxygen concentrations ($P > 0.6$, GEE).

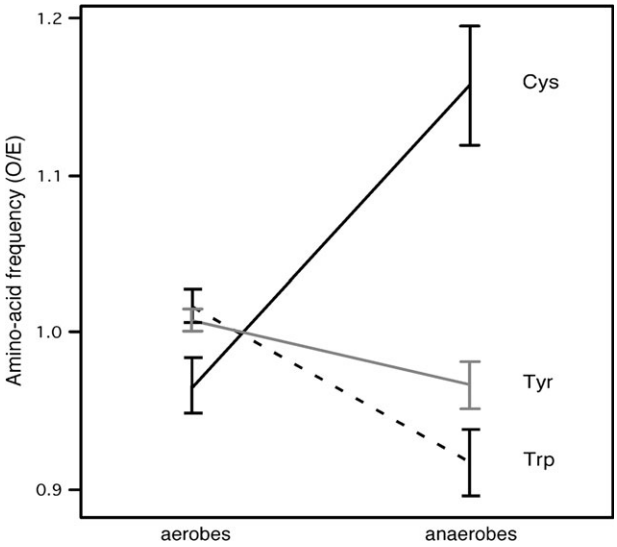

FIG. 4.—Average relative proteomic frequency of tryptophan, cysteine, and threonine in anaerobes and aerobes (after G + C content correction). Bars indicate standard errors.

The Strong Avoidance of Oxidable Amino Acids on the Surface of Proteins Is Unrelated to Aerobiosis

In the previous analyses, we found little evidence for a general avoidance of highly oxidable amino acids. Yet, we neglected the position of the residuals in the protein structure. Only residues accessible to reactive oxygen species (ROS) are expected to show avoidance of oxidable amino acids. It was further suggested that although most oxidable amino acids should be avoided at the surface of proteins, methionine might follow the inverse trend. Accordingly, methionine might protect other important and sensitive residues from oxidation (Levine et al. 2000). To investigate the validity of these theories, we analyzed the distribution of amino acids on the surface of proteins with orthologs in a large set of genomes (28 anaerobes and 108 aerobes). The choice of genomes was done in order to retrieve a maximum number of ubiquitous orthologs while preserving the largest possible number of genomes. Thus, we could compare the same proteins in all the different proteomes, and this allows controlling for functional bias. This is particularly important for the reason that anaerobes and aerobes have different metabolic pathways and functional repertoires, which may bias the overall proteomic frequency of amino acids. We made 34 multiple alignments of the orthologs of the 136 different species. For each position (residue) on each alignment, solvent accessibility was predicted using SABLE software (Adamczak et al. 2005). We then classed each residue as exposed or buried (see Material and Methods). For any given amino acid, we counted the number of exposed and buried residues (Xe and Xb). We then controlled for genomic compositional biases by calculating for each species the ratio $X\mathrm{e}/(X\mathrm{e} + X\mathrm{b})$.

In this data set of 34 ubiquitous proteins in 136 prokaryotic species, the usage of cysteine, threonine, and tryptophan followed the same qualitative trend as in the larger data set (data not shown). As predicted by the hypothesis that only accessible residues are sensible to oxidation, we find a very strong overall underrepresentation of oxidable residues in the surface of proteins (fig. 5). Cysteine is the most underrepresented amino acid, being present 5.4 times less frequently than expected considering the global percentage of exposed residues ($P < 0.0001$, Mann–Whitney test). Tryptophan, tyrosine, and methionine are also between 2 and 4 times less represented than expected ($P < 0.0001$, Mann–Whitney test), and only histidine shows moderate underrepresentation among accessible residues ($P < 0.0001$, Mann–Whitney test). However, it comes to mind that all these amino acids are hydrophobic, as well as oxidable, and may be avoided on exposed positions on that account. Indeed, we found a very strong positive correlation between the representation of amino acids on the surface of proteins and their hydrophobicity (Pearson $r = 0.89$, $P < 0.0001$) (fig. 6). Cysteine is the amino acid that strikes out for being more underrepresented than expected by its hydrophobicity. The other oxidable amino acids show underrepresentation in line with the expected values given their hydrophobicity.

It is therefore necessary to test more direct predictions of the theory that oxidable residues are avoided at the surface of proteins, notably that the underrepresentation of oxidable amino acids at the surface of proteins is more important in aerobes than in anaerobes. To control for phylogenetic dependence, we built a new phylogenetic tree of the 136 prokaryotes based on the concatenated amino acid sequence of the 34 ubiquitous proteins, in order to obtain a closer reconstruction of the evolutionary history of these proteins as possible (see Material and Methods). The global topology of this accessibility data tree is concordant with the 16S rDNA tree. Although cysteine, methionine, tryptophan, and tyrosine follow the expected trend, being more represented at the surface of proteins of anaerobes in comparison to aerobes (fig. 5), the effects are never significant after the control for phylogeny and Bonferroni sequential

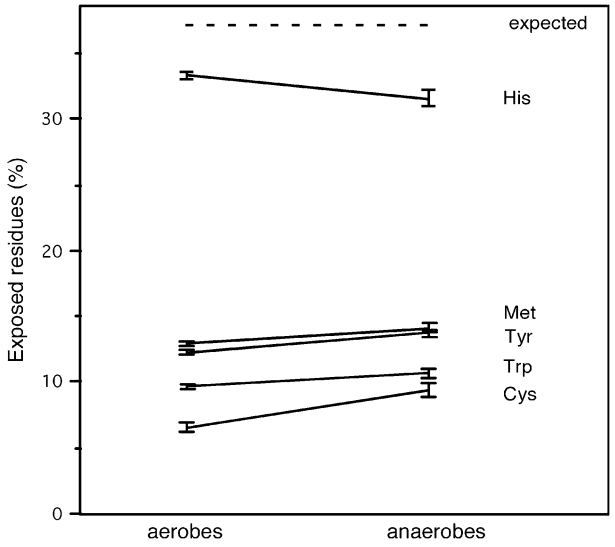

FIG. 5.—Percentage of exposed residues for each oxidable amino acid in anaerobes and aerobes. Bars indicate standard errors. The dashed line represents the percentage of exposed residues among all residuals in the set of proteins.

correction (all $P > 0.2$). Cysteine shows the most pronounced difference between aerobes and anaerobes, but this is mostly because cysteine content varies considerably between phylogenetic groups. For example, Firmicutes have an average observed/expected ratio of cysteine content of 0.76, whereas Proteobacteria have 1.09 (supplementary fig. 5, Supplementary Material online). Histidine, which overall is by far the least underrepresented amino acid in the set among exposed positions, displays the opposite behavior to the other oxidable residues, but its frequency is also not significantly different between the 2 groups. Thus, the differences observed between anaerobes and aerobes are more correlated to phylogeny than to oxygen requirement.

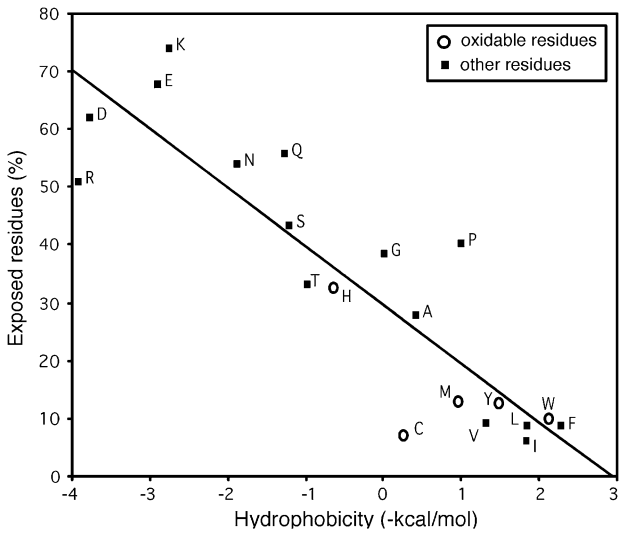

FIG. 6.—Percentage of exposed residues for each amino acid versus its hydrophobicity. The black line was obtained by linear regression: $Y = 0.30 + 0.10X$ ($R^2 = 0.79$, $P < 0.0001$). Circles represent oxidable amino acids. The hydrophobicity values were retrieved from Roseman (1988).

Consequently, there is no significant evidence for underrepresentation of oxidable amino acids on the surface of proteins of aerobic prokaryotes caused by selection to avoid the effects of oxidative stress. The overall underrepresentation of methionines at surfaces together with its nonsignificant difference in usage between aerobes and anaerobes further shows that, at least for these proteins, there is no evidence of methionines being selected to serve as a "target pool" to protect other more important residues from oxidation.

No Association between Oxygen-Rich tm-proteins and Aerobiosis

Previous research based on the analysis of 19 genomes without correcting amino acid compositions for G + C content has suggested that the evolution of communication-related tm-proteins was constrained by atmospheric oxygen availability (Acquisti et al. 2007). Transmembrane helices are usually present in the inner cell membrane, and their oxygen content shows a bimodal distribution that can be modeled by a mixture of 2 Gaussian distributions $N_A(\mu_A,\sigma_A)$ and $N_B(\mu_B,\sigma_B)$, with N_B having higher oxygen content than N_A ($\mu_B > \mu_A$). The outer domains of these proteins are typically enriched in oxygen-rich residues. Because tm-proteins in eukaryotes have longer outer domains ($R^2 = 0.24$ and $P < 0.005$, ANOVA; supplementary fig. 6, Supplementary Material online), they are also oxygen richer. Also, compared with prokaryotes, in eukaryotes the 2 Gaussians tend to merge ($\mu_B \approx \mu_A$) and the proportion of N_B is higher. Because N_B includes more communication-related tm-proteins, it was speculated that ancestral oxygen-poor atmosphere constrained the use of oxygen-rich residues in proteomes and that the switch to an oxidative atmosphere enabled eukaryotic species to increase the number of communication-related tm-proteins and the length of their outer domains (Acquisti et al. 2007). This hypothesis gives a preeminent role to oxygen availability in the appearance of complex multicellular life forms. Yet, we showed that oxygen does not seem to be a scarce resource for anaerobes and therefore should not be limiting to the evolution of tm-proteins.

A correlation between oxygen content in the outer domains of tm-proteins and the atmospheric oxygen concentration at the time of appearance of each group (archaea, prokaryotes and eukaryotes) implies that ancient constraints provoked signatures at the proteomic level that are still visible nowadays. Thus, it implies a strong phylogenetic inertia associated to proteomic oxygen level, which we have shown not to be true, and different proteomic oxygen composition of aerobes compared with anaerobes, which we have also shown not to be true. We analyzed 11 eukaryotes, 53 anaerobes, 231 aerobes, and 11 oxygenic photosynthetical prokaryotes. For such a large data set, we had to develop statistically rigorous ways of analyzing this data to apply automatic procedures. We first extracted the mean and SD of each Gaussian of tm-proteins and the proportion of each in the mixture using maximum likelihood and after controlling for G + C content (see Material and Methods). We then calculated delta (Δ) as the difference between the means of the 2 sets N_A and N_B ($\Delta = \mu_B - \mu_A$). We did

a logistic regression of μ_A, μ_B, and Δ to see which best discriminates between eukaryotes, bacteria, and archaea. The variables Δ ($R^2 = 0.27$, $P < 0.0001$) and μ_A ($R^2 = 0.23$, $P < 0.0001$) are the best discriminants, whereas μ_B is a very poor discriminant ($R^2 = 0.03$, $P = 0.03$). A logistic regression using both discriminants μ_A and Δ only slightly increases the accuracy of discrimination ($R^2 = 0.29$, $P < 0.0001$) relative to using Δ alone. Hence, we used only Δ to characterize the 2 sets of tm-proteins, grouping our data into eukaryotes, anaerobes, aerobes, and oxygenic photosynthetical prokaryotes. The results (fig. 7) fit the previously published study (Acquisti et al. 2007) showing a lower Δ in eukaryotes compared with prokaryotes. We controlled for phylogenetic dependency using GEE and did a contrast analysis of the Δ between anaerobes (AN), aerobes (AE), and oxygenic photosynthetical prokaryotes (PO). Eukaryotes were excluded from this analysis because they are monophyletic and were grouped together by phylogenetic criteria. This analysis shows that there is no significant difference between the Δ of any of the prokaryotic groups (P value AE–PO = 0.43, P value AE–AN = 0.39, P value AN–PO = 0.38). Although we cannot disprove the significance of the difference of Δ between eukaryotes and prokaryotes, the results show that the difference in terms of oxygen contents in the 2 Gaussians of tm-proteins, which is much more pronounced in prokaryotes, is not correlated to atmospheric oxygen availability during the evolution of the species studied. Indeed, oxygenic photosynthetical prokaryotes have been evolving in an environment enriched in oxygen resulting from their own photosynthesis. Ever since they invented that process, possibly before the appearance of the first eukaryote, they have never reverted to a nonaerobe state. Therefore, if the effect of oxygen enrichment in outer domains of tm-proteins was the result of oxygen availability, then cyanobacteria should be the most biased of all organisms, which they are not.

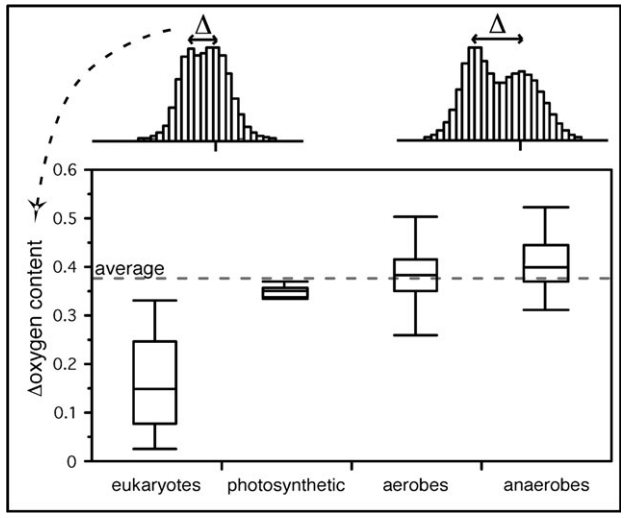

FIG. 7.—Association between the difference in oxygen content of the 2 populations of tm-proteins (Δoxygen) and 4 distinct groups of genomes. Given a mixture of 2 Gaussian $N_A(\mu_A,\sigma_A)$ and $N_B(\mu_B,\sigma_B)$, Δ is calculated as $\Delta = \mu_B - \mu_A$. Species are grouped into eukaryotes, photosynthetic, aerobic, and anaerobic prokaryotes. For each group, the box-and-whiskers plots are shown. The central line of the box is the median, the edges of the box are the quartiles, and the whiskers extend from the ends of the box to the outermost data point that falls within the distances computed: quartiles $\pm$ 1.5 $\times$ interquartile range. After correcting for phylogenetic dependence between the 3 prokaryotic groups, none is significantly different from another.

Conclusions

Oxygen is an essential constituent of living organisms and played a major role in their evolution. On one hand, the rise of molecular oxygen offered the opportunity for respiration to evolve, which is energetically more efficient than fermentation. On the other hand, the switch to an oxidative atmosphere was a source of massive environmental stress on existing species, forcing them to either adapt or retract to specific nonoxidative habitats. Throughout this work, we exploited the differential exposure to molecular oxygen of aerobes and anaerobes to analyze the role of oxygen in the evolution of the proteomes. This type of studies necessarily involves certain simplifications, notably about lifestyles, metabolic backgrounds, and toxic effects. Yet, if the effects of oxygen are important, one would expect that a sample of ~300 genomes should be enough to find a significant imprint of selection in proteomes. This is not the case for most tested hypotheses in this work, and the small differences between aerobes and anaerobes are most often the result of phylogenetic dependence.

First, we have found no evidence for the underrepresentation of oxygen-rich amino acids in the proteomes of anaerobes, not even among the highly expressed proteins. Hence, although oxygen is one of the fundamental elements of biomolecules, and contrary to previous proposals (Acquisti et al. 2007), there is no evidence that it is a scarce resource even for anaerobes. Oxygen is the most abundant chemical element on earth and can be incorporated in amino acids by other ways including reactions with many organic compounds and such abundant molecules as water. Indeed, amino acid biosynthesis rarely uses directly O_2, an exception being the synthesis of tyrosine from phenylalanine.

Second, there is no clear evidence of a global influence of oxidative stress on the amino acid composition of proteins: 1) Most oxidable amino acids are not avoided in aerobes relative to anaerobes. 2) Although most oxidable amino acids are avoided in exposed residuals in proteins, this trend is not different among aerobes and anaerobes, and it may be parsimoniously explained by their hydrophobicity. 3) Methionine behaves just like the average oxidable amino acid, showing no evidence for a protecting role on more important catalytical or structural residues. Besides disproving previously held ideas, this may uncover an important biological finding that the mechanisms of protein repair in aerobes fully compensate for the increase in oxidative stress they endure. In fact, molecular oxygen makes little more than 20% of today's atmosphere, but 300 MYA it was much more frequent having peaked at over 35% (Berner 1999). It is likely that this period of very high oxygen concentration has lead to the evolution of very efficient mechanisms to deal with oxidative stress among aerobes.

The genome of *E. coli* contains at least 27 proteins associated with oxidative stress response, which enables this facultative anaerobe to promptly withstand frequent and

high variations in the redox potential in the surrounding environment (Storz and Imlay 1999). We retrieved the orthologs of these 27 genes among our set of prokaryotic genomes and found an average of 15.4 orthologs in aerobic prokaryotes and 9.5 in anaerobic ones. This shows a clear trend of adaptation to aerobic/anaerobic lifestyles, which could largely counterweigh oxidative stress. It also shows that contrary to a frequently held view, the genomes of anaerobes sometimes code for extensive responses to ROS. Indeed, recent studies have shown that even Clostridia, which are obligate anaerobes routinely studied under strictly anaerobic conditions, have the potential to respond to stress induced by oxygen (Hillmann et al. 2008). Furthermore, other strategies have been found in bacteria to deal with the potential toxicity of derivatives of molecular oxygen. For example, *Borrelia burgdorferi* cells lack iron and this prevents the formation of oxygen radicals by the Fenton reaction (Posey and Gherardini 2000). *Pseudoalteromonas haloplanktis* lacks a series of activities leading to the formation of ROS, notably the nearly ubiquitous molybdopterin-dependent metabolism (Medigue et al. 2005). This bacterium grows at very low temperatures for which oxygen solubility in water is high. Individual cells may be stressed in the presence of unusual levels of oxygen. There is ample biochemical evidence that oxygen can induce protein damage. Yet, our results suggest that organisms have learned to cope with the oxygen levels they typically face by changing their metabolism, by avoiding certain highly susceptible proteins, or by developing adequate responses to damage. This matches very recent studies showing that adequate responses can turn anaerobes surprisingly tolerant to oxygen (Imlay 2008).

Third, we have reexamined the hypothesis that atmospheric enrichment in molecular oxygen might have led to the development of the communication capabilities of eukaryotes (Acquisti et al. 2007). We confirmed the trend of longer oxygen-rich outer domains in tm-proteins of eukaryotes compared with prokaryotes and thus oxygen richer tm-proteins. However, comparing photosynthetic, aerobic, and anaerobic prokaryotes, we found no significant association between oxygen availability in the species environment and oxygen distribution in outer domains of tm-proteins. Therefore, the results suggest that the effect is clade specific and independent of oxygen availability. Eukaryotes, even unicellular ones, have highly compartmentalized cells relative to prokaryotes. This may have led to the early development of tm-proteins involved in communication between compartments which were unnecessary in prokaryotes and which by some biochemical reason may involve unusual amounts of oxygen-rich amino acids. If so, these domains arose to cope with the necessity for communication between compartments not from an opportunity to introduce oxygen in proteins arising from the rise in oxygen concentrations in the atmosphere.

On the whole, the fact that available evidence leads to refuting most theories regarding the impact of molecular oxygen in proteins composition is a further example of the importance of rigorous testing of evolutionary hypotheses with adequate controls and within a phylogenetic framework (Gould and Lewontin 1979; Felsenstein 1985). It also shows the importance of using the full set of available genomes in evolutionary studies as some of the features previously found to be significant ceased being so simply by the increase in sample sizes.

Supplementary Material

Supplementary tables 1–3 and figures 1–6 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

Sara Vieira-Silva is supported by a grant from Fundação para a Ciência e a Tecnologia—Portugal (SFRH/BD/32968/2006). We thank Isabelle Gonçalves, Antoine Danchin, and Guillaume Santini for criticisms, comments, and suggestions on previous versions of this manuscript and Antoine Danchin for pointing to us the hypothesis that accessible residues might be adapted to face oxygen toxicity.

Literature Cited

Acquisti C, Kleffe J, Collins S. 2007. Oxygen content of transmembrane proteins over macroevolutionary time scales. Nature. 445:47–52.

Adamczak R, Porollo A, Meller J. 2005. Combining prediction of secondary structure and solvent accessibility in proteins. Proteins. 59:467–475.

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

Alves R, Savageau MA. 2005. Evidence of selection for low cognate amino acid bias in amino acid biosynthetic enzymes. Mol Microbiol. 56:1017–1034.

Baudouin-Cornu P, Surdin-Kerjan Y, Marliere P, Thomas D. 2001. Molecular evolution of protein atomic composition. Science. 293:297–300.

Bekker A, Holland HD, Wang PL, Rumble D 3rd, Stein HJ, Hannah JL, Coetzee LL, Beukes NJ. 2004. Dating the rise of atmospheric oxygen. Nature. 427:117–120.

Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J, Cuny G, Meunier-Rotival M, Rodier F. 1985. The mosaic genome of warm-blooded vertebrates. Science. 228:953–958.

Berner RA. 1999. Atmospheric oxygen over Phanerozoic time. Proc Natl Acad Sci USA. 96:10955–10957.

Bragg JG, Hyder CL. 2004. Nitrogen versus carbon use in prokaryotic genomes and proteomes. Proc Biol Sci. 271(Suppl 5):S374–S377.

Bragg JG, Thomas D, Baudouin-Cornu P. 2006. Variation among species in proteomic sulphur content is related to environmental conditions. Proc Biol Sci. 273:1293–1300.

Bragg JG, Wagner A. 2007. Protein carbon content evolves in response to carbon availability and may influence the fate of duplicated genes. Proc Biol Sci. 274:1063–1070.

Chothia C, Lesk AM. 1986. The relation between the divergence of sequence and structure in proteins. Embo J. 5:823–826.

Coghlan A, Wolfe KH. 2000. Relationship of codon bias to mRNA concentration and protein length in *Saccharomyces cerevisiae*. Yeast. 16:1131–1145.

Couturier E, Rocha EP. 2006. Replication-associated gene dosage effects shape the genomes of fast-growing bacteria

but only for transcription and translation genes. Mol Microbiol. 59:1506–1518.
Daubin V, Perriere G. 2003. G+C3 structuring along the genome: a common feature in prokaryotes. Mol Biol Evol. 20:471–483.
Davies MJ, Truscott RJW. 2001. Photo-oxidation of proteins and its role in cataractogenesis. J Photochem Photobiol B. 63:114–125.
Edgar RC. 2004. Muscle: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32: 1792–1797.
Felsenstein J. 1985. Phylogenies and the comparative method. Am Nat. 125:1–15.
Fukuchi S, Yoshimune K, Wakayama M, Moriguchi M, Nishikawa K. 2003. Unique amino acid composition of proteins in halophilic bacteria. J Mol Biol. 327:347–357.
Goldblatt C, Lenton TM, Watson AJ. 2006. Bistability of atmospheric oxygen and the great oxidation. Nature. 443:683–686.
Gould SJ, Lewontin RC. 1979. The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. Proc R Soc Lond B Biol Sci. 205:581–598.
Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 52:696–704.
Hedges SB, Blair JE, Venturi ML, Shoe JL. 2004. A molecular timescale of eukaryote evolution and the rise of complex multicellular life. BMC Evol Biol. 4:2.
Hillmann F, Fischer RJ, Saint-Prix F, Girbal L, Bahl H. 2008. PerR acts as a switch for oxygen tolerance in the strict anaerobe clostridium acetobutylicum. Mol Microbiol. 68: 848–860.
Howe K, Bateman A, Durbin R. 2002. Quicktree: building huge neighbour-joining trees of protein sequences. Bioinformatics. 18:1546–1547.
Humphrey W, Dalke A, Schulten K. 1996. Vmd: visual molecular dynamics. J Mol Graph. 14:33–38 27–38.
Hurst LD, Feil EJ, Rocha EP. 2006. Protein evolution: causes of trends in amino-acid gain and loss. Nature. 442:E11–E12.
Imlay JA. 2002. How oxygen damages microbes: oxygen tolerance and obligate anaerobiosis. Adv Microb Physiol. 46:111–153.
Imlay JA. 2006. Iron-sulphur clusters and the problem with oxygen. Mol Microbiol. 59:1073–1082.
Imlay JA. 2008. How obligatory is anaerobiosis? Mol Microbiol. 68:801–804.
Kreil DP, Ouzounis CA. 2001. Identification of thermophilic species by the amino acid compositions deduced from their genomes. Nucleic Acids Res. 29:1608–1615.
Krissinel E. 2007. On the relationship between sequence and structure similarities in proteomics. Bioinformatics. 23: 717–723.
Krogh A, Larsson B, von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol. 305:567–580.
Levine RL, Moskovitz J, Stadtman ER. 2000. Oxidation of methionine in proteins: roles in antioxidant defense and cellular regulation. IUBMB Life. 50:301–307.
Lobry JR. 1997. Influence of genomic G+C content on average amino-acid composition of proteins from 59 bacterial species. Gene. 205:309–316.
Mazel D, Marliere P. 1989. Adaptive eradication of methionine and cysteine from cyanobacterial light-harvesting proteins. Nature. 341:245–248.
McCord JM, Keele BB Jr., Fridovich I. 1971. An enzyme-based theory of obligate anaerobiosis: the physiological function of superoxide dismutase. Proc Natl Acad Sci USA. 68:1024–1027.
Medigue C, Krin E, Pascal G, et al. (24 co-authors). 2005. Coping with cold: the genome of the versatile marine Antarctica bacterium *Pseudoalteromonas haloplanktis* tac125. Genome Res. 15:1325–1335.
Moosmann B, Behl C. 2008. Mitochondrially encoded cysteine predicts animal lifespan. Aging Cell. 7:32–46.
Musto H, Rodriguez-Maseda H, Bernardi G. 1995. Compositional properties of nuclear genes from *Plasmodium falciparum*. Gene. 152:127–132.
Muto A, Osawa S. 1987. The guanine and cytosine content of genomic DNA and bacterial evolution. Proc Natl Acad Sci USA. 84:166–169.
Naya H, Romero H, Zavala A, Alvarez B, Musto H. 2002. Aerobiosis increases the genomic guanine plus cytosine content (GC%) in prokaryotes. J Mol Evol. 55:260–264.
Paradis E, Claude J. 2002. Analysis of comparative data using generalized estimating equations. J Theor Biol. 218:175–185.
Paradis E, Claude J, Strimmer K. 2004. Ape: analyses of phylogenetics and evolution in R language. Bioinformatics. 20:289–290.
Park S, Imlay JA. 2003. High levels of intracellular cysteine promote oxidative DNA damage by driving the fenton reaction. J Bacteriol. 185:1942–1950.
Posey JE, Gherardini FC. 2000. Lack of a role for iron in the Lyme disease pathogen. Science. 288:1651–1653.
R Development Core Team. 2007. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing.
Rocha EP. 2004. Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. Genome Res. 14:2279–2286.
Roseman MA. 1988. Hydrophilicity of polar amino acid sidechains is markedly reduced by flanking peptide bonds. J Mol Biol. 200:513–522.
Sasidharan R, Chothia C. 2007. The selection of acceptable protein mutations. Proc Natl Acad Sci USA. 104: 10080–10085.
Schmidt HA, Strimmer K, Vingron M, von Haeseler A. 2002. Tree-puzzle: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics. 18:502–504.
Sharp PM, Li WH. 1987. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 15:1281–1295.
Singer GA, Hickey DA. 2000. Nucleotide bias causes a genome-wide bias in the amino acid composition of proteins. Mol Biol Evol. 17:1581–1588.
Stadtman ER. 2006. Protein oxidation and aging. Free Radic Res. 40:1250–1258.
Storz G, Imlay JA. 1999. Oxidative stress. Curr Opin Microbiol. 2:188–194.
Sueoka N. 1962. On the genetic basis of variation and heterogeneity of DNA base composition. Proc Natl Acad Sci USA. 48:582–591.
Sueoka N. 1988. Directional mutation pressure and neutral molecular evolution. Proc Natl Acad Sci USA. 85:2653–2657.
Swire J. 2007. Selection on synthesis cost affects interprotein amino acid usage in all three domains of life. J Mol Evol. 64:558–571.
Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol. 56:564–577.
Tekaia F, Yeramian E. 2006. Evolution of proteomes: fundamental signatures and global trends in amino acid compositions. BMC Genomics. 7:307.

Towe KM. 2003. Evolution of protein amino acids. Science. 300:1370–1371.
Venables WN, Ripley BD. 2002. Modern applied statistics with S, 4th. New York: Springer.
Zeilstra-Ryalls JH, Kaplan S. 2004. Oxygen intervention in the regulation of gene expression: the photosynthetic bacterial paradigm. Cell Mol Life Sci. 61:417–436.
Zeldovich KB, Berezovsky IN, Shakhnovich EI. 2007. Protein and DNA sequence determinants of thermophilic adaptation. PLoS Comput Biol. 3:e5.

Jennifer Wernegreen, Associate Editor

Accepted June 19, 2008