Average oxidation state of carbon in proteins

Jeffrey M. Dick

1Department of Chemistry and Department of Applied Geology, Curtin University, Perth, WA, Australia
email: j3ffdick@gmail.com

Summary

The degree of oxidation of carbon atoms in organic molecules depends on the covalent structure. In proteins, the average oxidation state of carbon (Z_C) can be calculated as an elemental ratio from the chemical formula. To investigate oxidation-reduction (redox) patterns, groups of proteins from different subcellular locations and phylogenetic divisions were selected for comparison. Extracellular proteins of yeast have a relatively high oxidation state of carbon, corresponding with oxidizing conditions outside of the cell. However, an inverse relationship between Z_C and redox potential occurs between the endoplasmic reticulum and cytoplasm; this trend is interpreted as resulting from overall coupling of protein turnover to the formation of a lower glutathione redox potential in the cytoplasm. In Rubisco homologues, lower Z_C tends to occur in organisms with higher optimal growth temperature, and there are broad changes in Z_C in whole-genome protein compositions in microbes from different environments. Energetic costs calculated from thermodynamic models suggest that thermophilic organisms exhibit molecular adaptation to not only high temperature but also the reducing nature of many hydrothermal fluids. A view of protein metabolism that depends on the chemical conditions of cells and environments raises new questions linking biochemical processes to changes on evolutionary timescales.

Keywords: oxidation state, redox potential, subcellular location, protein metabolism, protein evolution
1 Introduction

Chemical reactions involving the transfer of electrons, known as oxidation-reduction or redox reactions, are ubiquitous in cellular and environmental systems [1, 2]. In the cell, the oxidation of thiol groups in proteins to form disulfides has the potential to regulate (activate or inhibit) enzymatic function [3]. Because these reactions are reversible on short timescales, a regulatory network known as redox signalling is made possible by reactions of small-molecule metabolites, including glutathione (GSH) and reactive oxygen species [4]. On timescales of metabolism, complex oxidation-reduction reactions are required for the formation (anabolism) and degradation (catabolism) of proteins and other biomolecules. Although many individual steps in biomass synthesis are irreversible, much biomass is ultimately recycled through endogenous metabolism [5]. On longer timescales, forces outside of individual cells and organisms sustain the redox disequilibria between inorganic and/or organic species that provide the energy source for metabolisms suited to a multitude of environments [6]. In turn, the actions of organisms can alter the redox conditions on Earth; the oxygenation of the atmosphere and oceans over geological time has a biogenic origin, and changed the course of later biological evolution [7].

Through evolution, the sequences of genes, and their protein products, are progressively altered. The elemental stoichiometry (chemical formula) and standard Gibbs energy of the molecules have a primary impact on metabolic requirements for energy and elemental resources. The energetic cost for synthesis of biomass is a function not only of the composition of the biomass, but also of environmental parameters including temperature and the concentrations of metabolic precursors. Temperature and oxidation-reduction potential have profound effects on the relative energetic costs of formation of different amino acids [8] or proteins [9]. These energetic costs are sensitive to differences in the elemental compositions of biomolecules. To a first approximation, a shift to a more reducing environment alters the energetics of reactions in a direction that favours the formation of relatively reduced chemical compounds. In a field test of this principle, metagenomic sequences for the most highly reduced proteins were found in the hottest and most reducing zones of a hot spring [10, 11].

The purpose of this study is to investigate a particular stoichiometric quantity, the average oxidation state of carbon (\(Z_C\), defined below), as a comparative tool for identifying compositional patterns at different levels of biological organization. By comparing a quantity derived from the elemental compositions of proteins, this study addresses one aspect of biochemical evolution. However, the questions raised here differ in important respects from conventionally defined biochemistry and evolutionary biology. Biochemical studies are most often concerned with the functions of molecules [12], including enzymatic catalysis and non-covalent interactions involved in the structural conformation of proteins and binding of ligands. Studies in molecular evolution often place emphasis on the historical relationships between sequences, but not their physical properties [12]. Combining these viewpoints, most current work assumes that structural stability of proteins is the primary criterion for molecular adaptation to high temperature [13]. In contrast, in this study, more attention is given to the stoichiometric and energetic demands of the reactions leading to protein formation. Because material replacement of proteins depends on metabolic outputs [14], the compositional differences among proteins have significant consequences for cellular organization and metabolism.

The following questions have been identified: 1) How does the relationship between \(Z_C\) of amino acids and corresponding codons relate to the origin or form of the genetic code? 2) How do the differences in \(Z_C\) between membrane proteins and others compare with properties of amino acids, e.g. hydrophobicity, known to favour localization to membranes? 3) How are the differences in \(Z_C\) of proteins among eukaryotic subcellular compartments related to differences in redox potential? 4) How are the differences in \(Z_C\) in families of redox-active proteins related to the standard reduction potentials of the proteins? 5) How are the differences in \(Z_C\) among different organisms, both in terms of bulk (genome-derived) protein composition and for homologues of a single family (Rubisco), related to environmental conditions, especially temperature and redox potential?
In the Results each problem is briefly introduced, the empirical distribution of $Z_C$ is described, and a discussion is developed to explore how the patterns reflect biochemical and evolutionary constraints. These short discussions, corresponding to topics (1)-(4), should be regarded as preliminary, and probably incomplete interpretations. The discussions are limited because there is no complete conceptual framework that links the biochemical reactions with the evolutionary processes that are implicit in all of the comparisons. The final section of the Results goes into more detail for the phylogenetic comparisons (topic 5) by examining the relative Gibbs energies of formation of proteins in environments of differing redox potential.

2 Methods

Throughout this study, “reducing” and “oxidizing” are used in reference to oxidation-reduction potential, tied to a particular redox couple or to environmental conditions, often expressed as millivolts on the Eh scale. “Reduced” and “oxidized” are used to refer to variations in the oxidation state of carbon.

The formal oxidation state of a carbon atom in an organic molecule can be calculated by counting -1 for each carbon-hydrogen bond, 0 for each carbon-carbon bond, and -1 for each bond of carbon to O, N or S [8, 15]. In photosynthesizing organisms, and autotrophs in general, the carbon source is CO$_2$, having the highest oxidation state of carbon (+4). The products of photosynthetic reactions include proteins and other biomolecules with a lower oxidation state of carbon. Even if the molecular structure is unknown, analytical elemental compositions can be used in calculations of the average oxidation state of carbon in biomass [16,17]. Because any gene or protein sequence corresponds to a definite canonical (nonionized, unphosphorylated) chemical formula, the average oxidation state of carbon in these biomolecules is easily calculated.

In amino acids and proteins, the average oxidation state of carbon ($Z_C$) can be calculated using

$$Z_C = \frac{-n_H + 3n_N + 2n_O + 2n_S + Z}{n_C},$$

where $Z$ is the charge on the molecule, and $n_C$, $n_H$, $n_N$, $n_O$ and $n_S$ are the numbers of the subscripted elements in the chemical formula of the molecule. The coefficients on the terms in the numerator derive from formal charges of atoms other than C, as follows: H (+1), N (-3), O (-2), S (-2). Negative formal charges reflect greater electronegativities of these elements compared to carbon. If two thiol groups react to form a disulfide bond, the oxidation states of the two affected sulfur atoms change from -2 to -1. Although H$_2$S is produced in this reaction, the oxidation state of carbon in the protein remains constant. It follows that equation (1) is applicable only to chemical formulas of proteins in which the N, O, and S are all fully reduced (bonded only to H and/or C).

The $Z$ in equation (1) ensures that ionization by gain or loss of a proton, having an equal effect on $Z$ and $n_H$, does not change the $Z_C$. Likewise, gain or loss of H$_2$O, which affects equally the values $2n_H$ and $n_O$, does not alter the average oxidation state of carbon [15]. Accordingly, the $Z_C$ of a peptide formed by polymerization of amino acids (a dehydration reaction) is a weighted average of the $Z_C$ in the amino acids, where the weights are the number of carbon atoms in each amino acid. As an example, the $Z_C$ of hen egg white lysozyme, having a chemical formula of C$_{613}$H$_{959}$N$_{193}$O$_{185}$S$_{10}$, is 0.016. This protein is oxidized compared to many other proteins, which commonly have negative values of $Z_C$.

To aid in reproducibility, data files of protein sequences or amino acid composition, except large files available from public databases, and computer program files for the calculations are provided in the Supporting Information. The calculations and figures were generated using the R software environment [18] together with the CHNOSZ package [19].
Figure 1: Average oxidation state of carbon ($Z_C$) in amino acids compared with (a) $Z_C$ in first two bases of the corresponding RNA codons and (b) hydropathy index of the amino acids taken from Kyte and Doolittle, 1982 [24]. Standard one-letter abbreviations for the amino acids are used to identify the points. In (a), the different codon compositions for serine (S) and arginine (R) are indicated by letters below the symbols, and some amino acid labels are shifted for readability. In (b), labels for asparagine (N) and glutamine (Q) are omitted for clarity; they plot at the same positions as aspartic acid (D) and glutamic acid (E), respectively.

3 Results

3.1 Comparison of $Z_C$ of amino acids with hydropathy and properties of codons

In contemplating the ancient origin of the genetic code, the chemical similarities of respective codons and amino acids have been used to argue for coevolution (shared biosynthetic pathways) [20] or a tendency toward similar physicochemical properties. The possible advantages that were identified for similar physicochemical properties include enhancing the steric interactions between amino acids and codons [21], or increasing the similarity between different amino acids resulting from a single DNA base mutation in order to maintain protein structure [20].

In the genetic code, the first two bases (a “doublet”) are more indicative of the amino acid than the third position of the codon [21]. The $Z_C$ of amino acids are compared with the values calculated for the corresponding RNA nucleobase doublets in figure 1a. Some of the doublets, e.g. UU (phenylalanine, leucine), CU (leucine), UC (serine) and CC (proline) have identical $Z_C$ (in this case, 1.5), leading to only 5 possible values of $Z_C$ for the doublets. The overall relationship suggested by figure 1a is loose correlation between the $Z_C$ of amino acids and of the RNA doublets. The most highly reduced amino acids, leucine (L) and isoleucine (I), are coded for by doublets having the two lowest $Z_C$ values.

The increase in $Z_C$ going from leucine to alanine to glycine (figure 1) is reflected in the metastability fields of these amino acids, which occur in order of increasing oxidation potential, or oxygen fugacity [22]. Metastable equilibrium refers to the equalization of the energies of reactions to form the amino acids; it is a partial equilibrium because the amino acids generally remain unstable with respect to inorganic species. Likewise, the relative Gibbs energies of formation reactions of amino acids differ considerably between hydrothermal (hot, reducing) and surface (cool, oxidizing) environments [8]. These patterns
Figure 2: Average oxidation state of carbon shown in histograms and normal probability plots for (a-b) all human proteins and (c-d) human membrane proteins. Only proteins of sequence length greater than or equal to 50 amino acids are considered. In the normal probability plots the lines are drawn through the 1st and 3rd quartiles, indicated by the crosses.

Support the possibility of metastable equilibria in hydrothermal environments between amino acids and nucleobase sequences that are paired in the genetic code. Tests of the potential for these states, carried out using Gibbs energies available at high temperature [9, 23], could reveal thermodynamic constraints on the energetics of abiotic or early biosynthesis independent of the arguments based on similarities in protein structure and biosynthetic pathways [20, 21].

The hydropathy index, based on the relative hydrophobicity and hydrophilicity of amino acids [24], is commonly used for identifying probable membrane-spanning domains of proteins. In figure 1b, $Z_C$ is compared with the hydropathy values for individual amino acids. The three most hydrophobic amino acids, isoleucine, leucine and valine, are also the three with the lowest $Z_C$. Therefore, membrane proteins with hydrophobic domains are likely to be more reduced than other proteins. The following sections examine the actual differences in human and yeast proteins.

3.2 Differences in $Z_C$ of membrane proteins

The lipid (fatty acid) components of membranes are reduced relative to many other biomolecules, including amino acids, nucleotides, and saccharides (see figure 1 of Amend et al., 2013 [25]). Proteins that are embedded in membranes tend to contain more hydrophobic amino acids, which enhance solubility of proteins in the membrane environment [24] and generally are relatively reduced (figure 1b).

To compare membrane proteins with other human proteins, sequences for all human proteins were taken from the UniProt database [26], and sequences for predicted membrane proteins were taken from all FASTA sequence files provided in Additional File 2 of Almén et al., 2009 [27]. Only sequences at least 50 amino acids in length were considered. The distribution of $Z_C$ of all human proteins (figure 2a;
For some model organisms, including *Saccharomyces cerevisiae* (yeast), the identities of proteins associated with subcellular compartments are now available in databases. Here, calculations of $Z_C$ of proteins and a comparison with independent measurements of redox potential are used to investigate the oxidation-reduction features and dynamics of cellular structure.

In a previous study, the limiting conditions for chemical transformations among proteins in subcellular compartments were quantified theoretically as a function of redox potential and hydration state [29]. In that study, the locations of proteins were taken from the "YeastGFP" study of Huh *et al.* 2003 [30]. That dataset has the advantage that relative abundances of many of the proteins are available, but it is limited to 23 named locations in the cell. In order to consider more cellular components (including the membranes), a more extensive reference proteome is used in this study. This proteome is based on the current *Saccharomyces* Genome Database (SGD) [31] annotations combined with the Gene Ontology (GO) [32] vocabulary for the "cellular component" aspect, which describes many organelles and membranes within the cell. Major cell components were selected for comparison, and the $Z_C$ was calculated for protein products of the genes, as summarized in table 1. The median values are also portrayed in the drawing of a yeast cell in figure 3.

It is apparent that the membrane proteins are highly reduced. However, not all membranes are equal; the proteins in the nuclear and inner and outer mitochondrial membranes are less reduced than those in the plasma membrane, and the endoplasmic reticulum (ER) membrane has very highly reduced proteins. Among the organelar proteins considered, the ER has the most reduced proteins, followed by vacuoles and mitochondria; the cytoplasmic proteins are moderately reduced. The proteins in the nucleus, bud neck and bud are more oxidized than in the other compartments. The most oxidized proteins in the system are the extracellular ones. The relative $Z_C$ of the proteins in the ER, mitochondrion, cytoplasm, and nucleus are consistent with the previous study in which the calculations took account of the abundances of proteins [29].

For comparison with $Z_C$ of proteins, the values of subcellular redox potential (Eh, in mV) listed in table 2 were compiled from literature sources [1, 33–40]. Measurements of reduced and oxidized glutathione...
Table 1: Summary of $Z_C$ of proteins in subcellular locations of yeast. Numbers of proteins ($n$) in SGD associated with the indicated GO terms are listed. The numerators of the fractions denote membrane-associated proteins that are also listed as “integral to membrane” (GO:0016021); only these proteins were used in the calculations of $Z_C$.

<table>
<thead>
<tr>
<th>cellular component</th>
<th>GO term</th>
<th>$n$</th>
<th>median $Z_C$</th>
<th>mean $Z_C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>cytoplasm</td>
<td>GO:0005737</td>
<td>2245</td>
<td>-0.136</td>
<td>-0.127</td>
</tr>
<tr>
<td>nucleus</td>
<td>GO:0005634</td>
<td>2073</td>
<td>-0.129</td>
<td>-0.121</td>
</tr>
<tr>
<td>mitochondrion</td>
<td>GO:0005739</td>
<td>1077</td>
<td>-0.164</td>
<td>-0.159</td>
</tr>
<tr>
<td>endoplasmic reticulum</td>
<td>GO:0005783</td>
<td>435</td>
<td>-0.191</td>
<td>-0.192</td>
</tr>
<tr>
<td>nucleolus</td>
<td>GO:0005730</td>
<td>263</td>
<td>-0.137</td>
<td>-0.128</td>
</tr>
<tr>
<td>Golgi apparatus</td>
<td>GO:0005794</td>
<td>215</td>
<td>-0.160</td>
<td>-0.167</td>
</tr>
<tr>
<td>cellular bud neck</td>
<td>GO:0005935</td>
<td>153</td>
<td>-0.111</td>
<td>-0.108</td>
</tr>
<tr>
<td>vacuole</td>
<td>GO:0005773</td>
<td>175</td>
<td>-0.164</td>
<td>-0.163</td>
</tr>
<tr>
<td>extracellular region</td>
<td>GO:0005766</td>
<td>95</td>
<td>-0.096</td>
<td>-0.098</td>
</tr>
<tr>
<td>cellular bud tip</td>
<td>GO:0005934</td>
<td>96</td>
<td>-0.113</td>
<td>-0.110</td>
</tr>
<tr>
<td>endoplasmic reticulum membrane</td>
<td>GO:0005789</td>
<td>283/338</td>
<td>-0.209</td>
<td>-0.211</td>
</tr>
<tr>
<td>plasma membrane</td>
<td>GO:0005886</td>
<td>224/427</td>
<td>-0.200</td>
<td>-0.188</td>
</tr>
<tr>
<td>mitochondrial inner membrane</td>
<td>GO:0005743</td>
<td>143/218</td>
<td>-0.184</td>
<td>-0.184</td>
</tr>
<tr>
<td>vacuolar membrane</td>
<td>GO:0005774</td>
<td>100/145</td>
<td>-0.205</td>
<td>-0.196</td>
</tr>
<tr>
<td>Golgi membrane</td>
<td>GO:0000139</td>
<td>76/121</td>
<td>-0.203</td>
<td>-0.200</td>
</tr>
<tr>
<td>nuclear membrane</td>
<td>GO:0031965</td>
<td>54/67</td>
<td>-0.161</td>
<td>-0.139</td>
</tr>
<tr>
<td>mitochondrial outer membrane</td>
<td>GO:0005741</td>
<td>51/92</td>
<td>-0.176</td>
<td>-0.177</td>
</tr>
</tbody>
</table>

(GSH and GSSG) in whole-cell extracts have been interpreted as reflecting cytoplasmic redox potential, but redox-sensitive green fluorescent protein (roGFP) probes [33] provide more specific data for subcellular locations. The data are not in all cases acquired from yeast, but it has been noted that cytoplasmic Eh values based on roGFP are similar for different model organisms [36].

The redox potentials in the vacuole and extracellular space are less well constrained than other locations. Under stress response, high amounts of GSSG, but not GSH, are sequestered in vacuoles [41]. A conservative lower range for the Eh of vacuoles (-160 to -130 mV) was calculated by taking a value of 80% GSSG and computing Eh from the GSH-GSSG equilibrium at concentrations of 1–10 mM GSH (see equation 21 and figure 4 of Ref. [35]). The redox potential would be higher if the GSSG/GSH ratio were in fact greater than 80/20. A high redox potential is also implicated by the presence of ferric iron species in vacuoles [42]. Extracellular redox state can vary greatly, but in aerobic organisms and laboratory culture it is likely to be generally oxidizing compared to subcellular compartments.

The values in table 2 are not comprehensive, and should be taken as a rough guide, but even with the uncertainties, comparison with the interquartile range of $Z_C$ of the proteins reveals some trends (figure 4a). The difference in both $Z_C$ and Eh is positive going from any subcellular compartment, except for vacuoles, to the extracellular space. This pattern has an intuitive explanation: by evolutionary adjustment to optimize proteins for their environment, the inside of the cell, which is more reducing, would be expected to have more reduced proteins compared to the outside.

The more surprising trend in figure 4a is an inverse relationship between $Z_C$ and Eh of the cytoplasm and ER. Does this contrast have any biochemical significance? The ER is a component of the secretory pathway, which transports proteins to membranes and to outside the cell [38]. Let us conjecture that the populations of proteins in the ER and cytoplasm are connected through common metabolic intermediates – their formation and degradation are part of the recycling of biomass through endogenous metabolism [5], also implied by metabolic closure [14]. It follows that the formation of proteins of a higher $Z_C$ in the cytoplasm entails the loss of electrons from proteins into metabolic pathways. Perhaps these pathways ultimately transfer these electrons to the formation of GSH in the cytoplasm.
Figure 3: A schematic drawing of a yeast cell showing median values of the average oxidation state of carbon in proteins from selected subcellular locations listed in table 1. All cellular components listed in the table are represented in the drawing. The colour scale is adjusted so that the cytoplasm has a neutral hue (white), and locations with relatively oxidized and reduced proteins are depicted by blue and red colours, respectively. Darker red colours are used for the more reduced groups of proteins in some of the membranes such as the ER, Golgi, vacuolar and plasma membranes. The nuclear and inner and outer mitochondrial membranes are shown with lighter reds because their proteins are relatively oxidized compared to those in the other membranes. Components not separated by membranes (or with membranes not shown) include the nucleolus, bud neck and bud tip.

Table 2: Values of Eh compiled from literature sources, for yeast cells or culture except as noted. The ranges account for variation among different cell types, experimental techniques and published values, and are used to construct figure 4.

<table>
<thead>
<tr>
<th>location</th>
<th>range (mV)</th>
<th>references</th>
</tr>
</thead>
<tbody>
<tr>
<td>mitochondrion</td>
<td>-360 to -255</td>
<td>Ref. [33]: roGFP probe (-360 mV; human HeLa).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ref. [34]: rxYFP in matrix (-296 mV) and intermembrane space (-255 mV).</td>
</tr>
<tr>
<td>cytoplasm</td>
<td>-320 to -240</td>
<td>Ref. [35]: GSH in proliferating mammalian cells (-240 mV).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ref. [36]: roGFP probe of GSH (-320 mV).</td>
</tr>
<tr>
<td>endoplasmic reticulum</td>
<td>-208 to -133</td>
<td>Ref. [37]: NYTC peptide (-185 to -133 mV; murine hybridoma CRL-1606).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ref. [38]: roGFP (-208 mV; human HeLa)</td>
</tr>
<tr>
<td>vacuole</td>
<td>-160 to &gt;-130</td>
<td>See text.</td>
</tr>
<tr>
<td>extracellular</td>
<td>-150 to &gt;160</td>
<td>Ref. [1]: Aerobic (160 mV) and anaerobic (90 mV) cultures.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ref. [39]: Very high-gravity fermentation (-150 mV).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ref. [40]: <em>H. sapiens</em> plasma (-140 mV).</td>
</tr>
</tbody>
</table>
Figure 4: The plot (a) compares average oxidation state of carbon in proteins from different subcellular locations of yeast with Eh values taken from glutathione (GSH-GSSH) or other redox indicators (see table 2). The heights of the boxes indicate the interquartile ranges of $Z_C$ values, and the widths represent the ranges of Eh listed in table 2. The scheme (b) invokes electron transfer to account for the contrasts in redox potential of GSH/GSSG and $Z_C$ of proteins between ER and cytoplasm (see text).

This scheme is depicted in figure 4b. The vertical dashed line represents a physical (but not impermeable) boundary between ER and cytoplasm; the curved lines represent reactions and transport within glutathione and protein systems that cross the compartments, and the arrow represents a linkage between glutathione and protein systems, which is effectively a coupled oxidation-reduction reaction. The scheme represents a redox mass-balance interpretation of the overall stoichiometric relationships, not a mechanism applied to individual GSH or protein molecules; the complete picture of metabolic connectivity in the cell is certainly more complex. Note that this scheme refers to the relative oxidation states of carbon of the proteins, not the oxidation of protein thiol groups to form disulfides. Disulfide bond formation takes place during folding and secretion of proteins, and may also contribute to glutathione metabolism [4]. Although there is growing detail of the pathways of glutathione metabolism in the cell, including compartmentalization between ER and cytoplasm, it is not known how they connect with non-thiol systems (e.g. [40]). Integrating the oxidation-reduction requirements of protein metabolism into existing metabolic models may help to complete the balance sheet of redox interactions in the cell.

Experiments on the connections between redox conditions and protein metabolism at the subcellular level can help elucidate the possible effects of coupling of protein metabolism to the glutathione redox system. If the net transfer of high-$Z_C$ proteins to the cytoplasm is stopped, then a decrease in GSH/GSSG redox potential in the ER relative to the cytoplasm would be expected based on the scheme shown in figure 4. This prediction is consistent with the outcome of experiments showing that puromycin-induced halting of protein synthesis causes a decrease in the redox potential monitored by roGFP in the ER [43]. A further untested implication of this hypothesis is that in ER-stress experiments the $Z_C$ of the protein population in the ER would increase. Metabolism of proteins might also interact with redox pathways other than the oxidation and reduction of glutathione. A linkage of this type has been documented in plants, where degradation of aromatic and branched-chain amino acids was identified as a source of electrons for the mitochondrial electron transport chain [44].

---

223 This scheme is depicted in figure 4b. The vertical dashed line represents a physical (but not impermeable) boundary between ER and cytoplasm; the curved lines represent reactions and transport within glutathione and protein systems that cross the compartments, and the arrow represents a linkage between glutathione and protein systems, which is effectively a coupled oxidation-reduction reaction. The scheme represents a redox mass-balance interpretation of the overall stoichiometric relationships, not a mechanism applied to individual GSH or protein molecules; the complete picture of metabolic connectivity in the cell is certainly more complex. Note that this scheme refers to the relative oxidation states of carbon of the proteins, not the oxidation of protein thiol groups to form disulfides. Disulfide bond formation takes place during folding and secretion of proteins, and may also contribute to glutathione metabolism [4]. Although there is growing detail of the pathways of glutathione metabolism in the cell, including compartmentalization between ER and cytoplasm, it is not known how they connect with non-thiol systems (e.g. [40]). Integrating the oxidation-reduction requirements of protein metabolism into existing metabolic models may help to complete the balance sheet of redox interactions in the cell.

236 Experiments on the connections between redox conditions and protein metabolism at the subcellular level can help elucidate the possible effects of coupling of protein metabolism to the glutathione redox system. If the net transfer of high-$Z_C$ proteins to the cytoplasm is stopped, then a decrease in GSH/GSSG redox potential in the ER relative to the cytoplasm would be expected based on the scheme shown in figure 4. This prediction is consistent with the outcome of experiments showing that puromycin-induced halting of protein synthesis causes a decrease in the redox potential monitored by roGFP in the ER [43]. A further untested implication of this hypothesis is that in ER-stress experiments the $Z_C$ of the protein population in the ER would increase. Metabolism of proteins might also interact with redox pathways other than the oxidation and reduction of glutathione. A linkage of this type has been documented in plants, where degradation of aromatic and branched-chain amino acids was identified as a source of electrons for the mitochondrial electron transport chain [44].
In plants, the oxidation and reduction of the iron-sulfur cluster in ferredoxin and the thiol/disulfide groups in ferredoxin/thioredoxin reductase and thioredoxin are coupled to form the ferredoxin/thioredoxin system [45] ("system" here refers to the set of interacting proteins, and is not a system in the thermodynamic sense). Ferredoxin has the lowest standard reduction potential (midpoint potential, $E^{\circ'}$ at 25 °C) in this system, and its iron-sulfur cluster is reduced by light energy through photosystem I. The oxidation of ferredoxin coupled to the reduction of thioredoxin is catalysed by ferredoxin/thioredoxin reductase. Reduced thioredoxin can disseminate the redox signal via reduction of disulfide groups in other proteins, activating their enzymatic functions. Glutaredoxins are another group of thiol/disulfide proteins that also interact with disulfide bonds in proteins and, unlike thioredoxin, are reduced by glutathione [46].

The standard reduction potentials of proteins in the ferredoxin/thioredoxin system in spinach [47] and of thioredoxin and the glutaredoxin system in *E. coli* [48] are compared with $Z_C$ in figure 5. In the ferredoxin/thioredoxin chain of spinach, there is a strong decrease in $Z_C$ with increasing $E^{\circ'}$. In the glutaredoxin system of *E. coli* and the associated proteins protein disulfide isomerase (PDI) and thiol:disulfide interchange protein (DsbA), there is a smaller decrease in $Z_C$ with increasing $E^{\circ'}$. Thioredoxin in *E. coli*, which has $Z_C$ and $E^{\circ'}$ that are similar to thioredoxin in spinach, does not follow the trend apparent in the glutaredoxin and related proteins.

From figure 5 it appears that the $Z_C$ of proteins involved in some parts of the redox signalling networks are inversely correlated with their standard reduction potential. Do systematic changes in amino acid composition implied by $Z_C$ impact the chemistry of the active site? The atomic environment surrounding the redox-active sites affects reduction potential [49], and changing very few residues in the vicinity of...
the active site can affect the function [50]. However, these or similar proximal effects may not provide a complete explanation for the trends in $Z_C$ in figure 5, which apply to the entire protein sequences. A speculative explanation is offered here. In general, any oxidation of a protein molecule may involve loss of an electron from the active site or from a covalent bond distal to the active site. In proteins with higher $Z_C$, the degree of oxidation of amino acids is greater, making the further loss of electrons from covalent bonds more difficult. The covalent oxidation of proteins ultimately leads to their degradation, disrupting the function of redox signalling networks. Therefore, it may be advantageous for low-$E^\prime\prime$ active sites (those that have a greater potential to lose electrons on signalling timescales), to be associated with high-$Z_C$ proteins (those in which the covalent bonds have lost a greater number of electrons on evolutionary timescales). The break in the pattern by thioredoxin in E. coli, and the scattered distribution of $Z_C$ and $E^\prime\prime$ of many other redox-active proteins that are not shown, presumably indicate that these hypothetical relations are applicable only to closely interacting chains of redox-active proteins and not the entire cell.

There are dual hypotheses here: first, that high-$Z_C$ proteins have a lower tendency for irreversible oxidative degradation, and second, that reversible reduction potential and tendency for irreversible oxidation, which are different biochemical properties of the same molecule, are jointly tuned by evolution. One implication of the first hypothesis is that thioredoxin in spinach, which has a relatively low $Z_C$, would be more easily covalently oxidized, and therefore may have a higher turnover rate than ferredoxin.

### 3.5 Phylogenetic variation in $Z_C$ of proteins and comparison with optimal growth temperature

A comparison of $Z_C$ of the combined proteins from selected microbial genomes is shown in figure 6. The sets of proteins shown on the left-hand side of figure 6 correspond to those organisms whose scientific names contain the indicated substring. In many cases, the names of the organisms reflect their environments and/or metabolic strategies. Examples of the matching genus names are *Natronobacterium*, *Haloferax*, *Rhodobacter*, *Acidovorax*, *Methyllobacterium*, *Chlorobium*, *Nitrosomonas*, *Desulfovibrio*, *Geobacter*, *Methanococcus*, *Thermococcus*, *Pyrobaculum*, *Sulfolobus*. Most terms, however, match more than one genus (e.g. *Pyrobaculum* and *Pyrococcus*). On the right-hand side of figure 6 are shown genera containing many groups with clinical and technological relevance; by the numbers of points it is apparent that their representation in RefSeq is greater than that of the environmental microbes.

A general trend toward lower $Z_C$ in proteins in organisms from hot environments (e.g. represented by Thermo and Pyro in the names) is apparent. Organisms with the highest $Z_C$ inhabit saline evaporative waters (Nat, Halo), while other aquatic organisms (e.g. Rhodo) have less highly oxidized proteins. Within a given genus (right-hand side of plot), the clusters of $Z_C$ tend to be tighter, reflecting conserved compositional trends. *Streptomyces*, common in soils, has the highest $Z_C$ of the genera shown here. *Buchnera* is notable because its proteins are highly reduced, including one example (*Buchnera aphidicola* BCc) with the lowest $Z_C$ of proteins within the entire dataset. *B. aphidicola* BCc is the primary endosymbiont of the cedar aphid (*Cinara cedri*) and, at the time of sequencing, had one of the smallest known bacterial genomes [51]. Being relatively closely related [52], *Mycoplasma* and *Clostridium* also have relatively low $Z_C$. *Mycoplasma* are known for their small genomes and dependence on metabolic products of the host; the low $Z_C$ may be a constraint imposed by growth in reducing intracellular or intraorganismal environments.

We now turn to a comparison of homologues of a specific protein. Rubisco is an essential enzyme for carbon fixation. Sequence comparison of homologues (related sequences that appear in different organisms) has provided the basis for many phylogenetic studies [53]. Major divergent forms of the enzyme are Forms I and II, found in aerobic organisms, and Form III and “Rubisco-like proteins”, found in anaerobic organisms [54]. The organisms listed in table 3 have in common the occurrence of Rubisco in their genome. Forms I, II or III were included in this comparison, but Rubisco-like proteins were
Figure 6: Average oxidation state of carbon in total combined proteins from sequenced microbial genomes. Sequences of microbial proteins were taken from NCBI RefSeq release 61. Only organisms with total sequenced protein length greater than 100,000 amino acids were used, leaving 6323 organisms. The group on the left-hand side is identified by substring matches in the scientific name of the organism; the terms were chosen to emphasize environmental variation. The group on the right-hand side consists of the indicated genera, emphasizing organisms of clinical and biotechnological relevance. The final category represents total proteins from all microbial genomes meeting the minimal size requirement, many of which are not shown in the other categories.
Table 3: Names of species, optimal growth temperatures ($T_{\text{opt}}$) and UniProt [26] accession numbers (IDs) for the large subunit of ribulose bisphosphate carboxylase (Rubisco). Literature references for $T_{\text{opt}}$ are indicated in brackets. Abbreviations: A – Archaea; B – Bacteria; E – Eukaryota. The numbers are used to identify the points in figure 7 (duplicated numbers occur in different temperature ranges).

<table>
<thead>
<tr>
<th>number</th>
<th>domain</th>
<th>species</th>
<th>$T_{\text{opt}}, ^{\circ} \text{C}$</th>
<th>reference</th>
<th>ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>E</td>
<td><em>Phaeocystis antarctica</em></td>
<td>0–6</td>
<td>[55]</td>
<td>G9FID8</td>
</tr>
<tr>
<td>2</td>
<td>B</td>
<td><em>Octadecabacter antarcticus 307</em></td>
<td>4–10</td>
<td>[56]</td>
<td>M9R7V1</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td><em>Methanolobus psychrophilus</em></td>
<td>18</td>
<td>[57]</td>
<td>K4MAK9</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td><em>Methanococcoides burtonii</em></td>
<td>23</td>
<td>[58]</td>
<td>Q12TQ0</td>
</tr>
<tr>
<td>5</td>
<td>B</td>
<td><em>Bradyrhizobium japonicum</em></td>
<td>25–30</td>
<td>[59]</td>
<td>Q9ZI34</td>
</tr>
<tr>
<td>6</td>
<td>B</td>
<td><em>Thiobacillus ferrooxidans</em></td>
<td>28–30</td>
<td>[60]</td>
<td>P0C916</td>
</tr>
<tr>
<td>7</td>
<td>E</td>
<td><em>Zea mays</em></td>
<td>30</td>
<td>[61]</td>
<td>P00874</td>
</tr>
<tr>
<td>8</td>
<td>B</td>
<td><em>Mariprofundus ferrooxydans</em></td>
<td>30</td>
<td>[62]</td>
<td>Q0EX22</td>
</tr>
<tr>
<td>9</td>
<td>B</td>
<td><em>Desulfovibrio hydrothermalis</em></td>
<td>35</td>
<td>[63]</td>
<td>L0RHZ1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>number</th>
<th>domain</th>
<th>species</th>
<th>$T_{\text{opt}}, ^{\circ} \text{C}$</th>
<th>reference</th>
<th>ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td><em>Methanosarcina acetivorans</em></td>
<td>35–40</td>
<td>[64]</td>
<td>Q8TGH2</td>
</tr>
<tr>
<td>2</td>
<td>B</td>
<td><em>Acidithiobacillus caldus</em></td>
<td>45</td>
<td>[65]</td>
<td>F9ZLP0</td>
</tr>
<tr>
<td>3</td>
<td>E</td>
<td><em>Cyanidium caldarium</em></td>
<td>45</td>
<td>[66]</td>
<td>P37393</td>
</tr>
<tr>
<td>4</td>
<td>B</td>
<td><em>Sulfobacillus acidophilus</em></td>
<td>45–50</td>
<td>[67]</td>
<td>P72383</td>
</tr>
<tr>
<td>5</td>
<td>B</td>
<td><em>Pseudomonas hydrogenothermophila</em></td>
<td>52</td>
<td>[68]</td>
<td>Q51856</td>
</tr>
<tr>
<td>6</td>
<td>B</td>
<td><em>Synechococcus sp. (strain JA-2-3B’aw2(1-3))</em></td>
<td>50–55</td>
<td>[69]</td>
<td>Q2JJIP3</td>
</tr>
<tr>
<td>7</td>
<td>A</td>
<td><em>Methanoseta thermophila</em></td>
<td>55–60</td>
<td>[70]</td>
<td>A0B9K9</td>
</tr>
<tr>
<td>8</td>
<td>B</td>
<td><em>Thermosynechococcus elongatus</em></td>
<td>57</td>
<td>[71]</td>
<td>Q8DIS5</td>
</tr>
<tr>
<td>9</td>
<td>B</td>
<td><em>Clostridium clariflavum</em></td>
<td>60</td>
<td>[72]</td>
<td>G8LZL2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>number</th>
<th>domain</th>
<th>species</th>
<th>$T_{\text{opt}}, ^{\circ} \text{C}$</th>
<th>reference</th>
<th>ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>B</td>
<td><em>Bacillus acidocaldarius</em></td>
<td>65</td>
<td>[73]</td>
<td>F8IIID7</td>
</tr>
<tr>
<td>2</td>
<td>B</td>
<td><em>Thermotoga lettingae</em></td>
<td>65</td>
<td>[74]</td>
<td>A8F7V4</td>
</tr>
<tr>
<td>3</td>
<td>B</td>
<td><em>Thermomicrobium roseum</em></td>
<td>70</td>
<td>[75]</td>
<td>B9KXE5</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td><em>Archaeoglobus fulgidus</em></td>
<td>76</td>
<td>[76]</td>
<td>O28635</td>
</tr>
<tr>
<td>5</td>
<td>A</td>
<td><em>Methanocaldococcus jannaschii</em></td>
<td>85</td>
<td>[77]</td>
<td>Q58632</td>
</tr>
<tr>
<td>8</td>
<td>A</td>
<td><em>Pyrococcus horikoshii</em></td>
<td>98</td>
<td>[80]</td>
<td>O58677</td>
</tr>
<tr>
<td>9</td>
<td>A</td>
<td><em>Pyrococcus furiosus</em></td>
<td>100</td>
<td>[81]</td>
<td>Q8U1P9</td>
</tr>
</tbody>
</table>

excluded; however, some that were tested were found to have considerably lower $Z_C$. The selection of organisms was made in order to represent a variety of optimal growth temperatures ($T_{\text{opt}}$) as reported in the studies cited in the table [55–81].

A comparison between $T_{\text{opt}}$ and $Z_C$ of Rubisco is presented in figure 7. The $Z_C$ of Rubisco are somewhat higher than the bulk protein content of the organisms; compare for example the values for *Pyrococcus horikoshii* and *P. furiosus* (the highest-temperature points labelled 8 and 9 in figure 7a) with the range of values for “Pyro” in figure 6. At lower temperatures (0 to 50 °C), the differences between domains of life are most apparent; Rubiscos of the Bacteria in this sample set are more oxidized than those of Archaea and Eukaryota. There is a tendency for the Rubiscos of the Archaea to have lower $Z_C$; this appears to be characteristic of anaerobic methanogenesis and Form III Rubiscos. An interesting exception is the high-$Z_C$ Rubisco of *Methanoseta thermophila*; this organism grows on acetate to produce both CH$_4$ and CO$_2$ [70]. The major pattern that emerges is that higher temperatures are associated with a lower average oxidation state of carbon in proteins. As outlined below, a decrease in oxidation state of carbon in the covalent structure of the proteins confers energetic savings in hot, reducing environments.
3.6 Beyond \( Z_C \): energetics of protein formation as a function of environmental redox potential

To a first approximation, energetic considerations predict that more reducing conditions tend to favour formation of proteins with relatively lower \( Z_C \), and vice versa. To assess the directionality and magnitude of chemical forces on the evolutionary transformations of proteins, the energetics of reactions can be calculated using thermodynamic models. Because the timescales of evolution are much longer than transformations of biomolecules during metabolism, a discussion of the assumptions underlying the application of thermodynamic theory to biochemical evolution is warranted.

The calculation of equilibrium provides a quantitative description of the state of a system in an energy minimum. The assumption of equilibrium is the foundation for many models of inorganic processes in geochemistry. Although the metabolic formation of proteins and other biochemical constituents proceeds in a non-equilibrium manner, using the equilibrium state as a frame of reference makes it possible to compare the energetics of living systems quantitatively (“how far from equilibrium”).

In energetic terms, adaptation can be defined as a problem of optimal efficiency, with trade-offs between energy utilization and power [82, p. 140]. Overall minimization of biomass synthesis costs can be expected from the energy utilization standpoint. The links between energetics and evolutionary outcome implicitly depend on the proposal that greater fitness is associated with mutations that lower the synthesis costs of proteins for a given function (e.g. [83]).

It has been argued previously that the propensity for some evolutionary changes can be modelled using
the related concepts of equilibration, energy minimization, and maximum entropy. In an evolution-
ary context, these concepts have often been defined by analogy to their definitions in thermodynam-
ics [82, 84]. The current discussion instead considers the possibility of direct application of chemical
thermodynamics (the geochemical approach) to formulate a quantitative description of patterns of protein
composition. The major assumption used in the following discussion is that energetic demands of pro-
tein formation depend not only on the composition of the protein, but also the environmental conditions.
Therefore, it is expected that environmental adaptation has left an imprint on chemical compositions of
phylogenetically distinct proteins.

An example calculation is carried out here for selected Rubiscos from organisms with optimal tempera-
tures in the range of 23–30 °C (table 3, numbers 4–7). The basis species (representing inorganic starting
materials) and their chemical activities used for this example are CO₂ (10⁻³), H₂O (1), NH₃ (10⁻⁴), H₂S
(10⁻⁷) and H⁺ (10⁻⁷, i.e. pH = 7). The activity of the electron, representing the effects of the redox
variable (Eh) through the Nernst equation, is left to vary. The Gibbs energies of the reactions to form
the proteins were calculated as described previously [9, 19] and are plotted in figure 7b. In figure 7b it
can be seen that the Gibbs energies of the reactions to form the proteins (∆G), normalized by protein
length, steadily increase (become less favourable) with increasing Eh, but the differences between the
proteins can not be discerned easily. In figure 7c, the values of ∆G are shown relative to the reaction
for T. ferrooxidans, giving a difference in Gibbs energy of formation (ΔΔG) that can be used to assess
the relative energetics of the reactions. Lower energies indicate the most stable protein (in the sense of
chemical formation, not structural conformation), and these relative energies depend on the chemical
conditions of the environment, measured in part by the oxidation-reduction potential.

By using the Gibbs energy calculations such as shown in figure 7c, one can make an assessment of
which protein in a given redox environment demands lower energy for overall synthesis (i.e. is more
stable) than other possibilities. Where two lines cross in figure 7c, the energies to form the two proteins
are equal, representing a metastable (partial) equilibrium. In metastable equilibrium, the coexistence of
proteins with equal energies of formation corresponds to a local energy minimum. This interpretation
does not preclude the non-equilibrium character of the overall biosynthetic process, because the energies
of formation remain non-zero (figure 7b).

The three least costly proteins, going from low to high Eh, are those from M. burtonii, T. ferrooxidans,
and Z. mays. Methanogenic archaea such as M. burtonii [58] inhabit reducing environments, where Eh
values as low as at least -400 mV have been documented [2]. In contrast, the metal-leaching activity of
T. ferrooxidans is associated with an increase in oxidation-reduction potential (ORP, which carries the
same units as Eh) [85] and, compared to M. burtonii, its Rubisco is more oxidized and is consequently
more stable at higher Eh. The correspondence between redox conditions and lower ∆G of formation of
the proteins provides thermodynamic evidence for environmental adaptation of the proteins.

The protein from T. ferrooxidans, which has the highest ZC of the four considered, does not have the
lowest value of ΔΔG in the most oxidizing (highest Eh) conditions. Instead, the Rubisco from Z. mays,
even though it has lower ZC, is calculated to relatively more stable at the highest Eh values considered.
This result shows that comparing values of ZC provides only an approximation of the dependence of the
relative energetics of protein formation on redox changes; chemical thermodynamic models integrate
more information about reaction stoichiometry and the effects of multiple environmental variables.

Previously, comparison of Rubisco from hot-spring organisms adapted to higher temperatures revealed an
increase in frequency hydrophobic amino acids, interpreted as increasing the conformational stability of
proteins at high temperature [86]. However, another basis for interpretations of thermophilic adaptation
depends on the relative energetic costs of synthesis of amino acids [87]. The finding made in this study
is that the high-temperature Rubiscos exhibit a shift toward lower ZC (figure 7a). Higher temperatures
are often associated with more reducing environments. For example, compared to seawater, activities
of dissolved hydrogen in hydrothermal fluids are higher, and mixed hydrothermal-seawater fluids have a reducing potential that favours formation of relatively reduced amino acids [8]. Redox potential is a major variable affecting the energetics of protein formation at different temperatures; therefore, adaptation to minimize biosynthetic costs in high-temperature environments is likely to have more than just a thermophilic (temperature dependent) aspect.

Although the stoichiometric comparisons are distinct from sequence-based phylogenetic analyses which are used to test for positive selection, it is conceivable that Z_C could in the future be incorporated into tree-based models of evolution that take account of physicochemical properties of amino acids in proteins [88]. However, as noted above, thermodynamic comparisons have greater power than compositional comparisons such as Z_C. The computation of metastable equilibrium takes account simultaneously of temperature, redox potential (expressed as Eh, activity of hydrogen, or oxygen fugacity) and other variables [19]. In another study, analysis of metagenomic and geochemical data led to a predicted metastable succession of proteins (with generally increasing Z_C) that could be aligned with a gradient of increasing oxidation potential and decreasing temperature in a flowing hot spring [10]. When grouped by taxonomic similarity, the Z_C in the hot-spring proteins, while becoming lower at high temperature, also spread over a broader range, leading to tighter constraints on the redox conditions suitable for metastable coexistence of the organisms [11].

Not only differences in the oxidation states of present-day environments, but also the oxygenation of Earth’s atmosphere and oceans through geological time could have profound impacts on the energetics of biomass synthesis [7]. Adaptation to reduce these costs likely would lead to divergences in Z_C that are apparent across different taxa, while closer phylogenetic relationships should confer a similarity in Z_C. In common with the Rubiscos, comparison of the total proteins of microbes reveals a tendency for proteins in organisms associated with hot, as well as sulfidic and methanogenic environments, to be more reduced (figure 6).

4 Conclusions

Proteins are products of metabolism; their synthesis and degradation are part of the network of chemical reactions that sustains the living cell. Comparisons of the average oxidation state of carbon in proteins have provided a starting point for visualizing the compositional diversity of proteins in relation to redox chemistry in subcellular compartments and external environments. The large differences in Z_C of proteins in locations such as the ER and cytoplasm likely have consequences for the dynamics of oxidation-reduction reactions involving glutathione and other metabolites. Further insight may be gained by including the formation and degradation of proteins in kinetic and stoichiometric models of metabolic networks. Extension of these concepts to other phenomena entailing changes in both redox potential and protein expression, such as stress response to oxidizing agents and the cell cycle, can be envisioned. The deepest significance of the observed patterns lies in their emergence over evolutionary timescales. The inverse trend relating Z_C with standard reduction potentials in chains of redox-active proteins is a case where the chemical composition of the proteins may be tuned with the electron-transfer chemistry of the active sites. Compositional divergences among proteins are also apparent in phylogenetic comparisons, and here it is reasonable to conclude that correlations between oxidation state of carbon in proteins and the redox potential of the environment indicate some degree of energy savings conferred by evolution. The natural history of protein evolution is a result of processes that are both unpredictable (mutation events) and, to some extent, deterministic (selection for fitness in a given environment). By describing protein molecules in terms of chemical composition and energetics it will be possible to identify some of the forces that help to shape the occurrences of proteins in different cells and environments.
5 Supporting information

The supporting information is provided in a ZIP archive containing the following code and data files:

**prep.R** This file contains code used to prepare the data files for easier handling by the plotting functions.

**plot.R** This file contains code used to make the figures appearing in this paper. The functions, in the order of the figures (1–7), are amino(), human(), yeast(), potential(), midpoint(), phylo(), and rubisco(). The code is written in R [18] and depends on version 1.0.2 of the CHNOSZ package [19], available from the Comprehensive R Archive network (http://cran.r-project.org).

**data/SGD_associations.csv** For yeast genes, this table lists the accessions, SGDID, and the association to cellular components in the Gene Ontology, derived from gene_association.sgd.gz, protein_properties.tab and go_terms.tab downloaded from http://www.yeastgenome.org on 2013-08-24. All gene associations with the NOT qualifier were removed, as were those without a matching entry in protein_properties.tab (e.g. RNA-coding genes).


**data/codons.csv** In the first column, the three-letter abbreviations for each of the RNA codons; in the second column, the names of the corresponding amino acids.

**data/midpoint.csv** List of protein names, UniProt IDs and standard midpoint reduction potentials used to make figure 5. Start and stop positions, taken from UniProt, identify the protein chain excluding initiator methionines or signal peptides.

**data/protein_refseq.csv** Amino acid compositions of total proteins in 6758 microbial genomes from RefSeq release 61, dated 2013-09-09. The gene identifier (gi) numbers of the sequences were assigned taxonomic IDs (taxids) using the RefSeq release catalogue. The amino acid compositions of the total proteins were calculated by averaging the compositions of all proteins for each taxid. The “organism” column contains the taxid used in NCBI databases, the “ref” column contains the names of the RefSeq files from which the amino acid sequences were taken (with start and end positions in parentheses) followed by the scientific name of the organisms in brackets, and the “abbrv” column contains the number of amino acids for that organism. Scientific names for the taxids at the species level were found using the names.dmp and nodes.dmp files downloaded from ftp://ftp.ncbi.nih.gov/pub/taxonomy/taxdump.tar.gz on 2013-09-18.

**data/rubisco.csv** UniProt IDs for Rubisco and optimal growth temperatures of organisms (see table 3).

**cell/*.png** PNG images for each of the cellular components used to make figure 3.

**fasta/midpoint/*.fasta** FASTA sequence files for proteins shown in figure 5.

**fasta/rubisco/*.fasta** FASTA sequence files for each Rubisco identified in table 3.

Acknowledgements

Thanks to Svenja Tulipani and Katy Evans for their comments on an earlier version of the manuscript.
References

[1] Hewitt LF. 1950 Oxidation-Reduction Potentials in Bacteriology and Biochemistry, 6th edn. Edin-
burgh, Scotland: E. & S. Livingstone Ltd.


[7] Sleep NH, Bird DK. 2008 Evolutionary ecology during the rise of dioxygen in the Earth’s atmos-


[34] Hu JJ, Dong LX, Outten CE. 2008 The redox environment in the mitochondrial intermembrane space is maintained separately from the cytosol and matrix. *J. Biol. Chem.* 283, 29126–29134. (doi:10.1074/jbc.M803028200)


