npg

ORIGINAL ARTICLE

The significance of nitrogen cost minimization in proteomes of marine microorganisms

Joseph J Grzymski and Alex M Dussaq

Division of Earth and Ecosystem Sciences, Desert Research Institute, Reno, NV, USA

Marine microorganisms thrive under low levels of nitrogen (N). N cost minimization is a major selective pressure imprinted on open-ocean microorganism genomes. Here we show that amino-acid sequences from the open ocean are reduced in N, but increased in average mass compared with coastal-ocean microorganisms. Nutrient limitation exerts significant pressure on organisms supporting the trade-off between N cost minimization and increased average mass of amino acids that is a function of increased A + T codon usage. N cost minimization, especially of highly expressed proteins, reduces the total cellular N budget by 2.7–10%; this minimization in combination with reduction in genome size and cell size is an evolutionary adaptation to nutrient limitation. The biogeochemical and evolutionary precedent for these findings suggests that N limitation is a stronger selective force in the ocean than biosynthetic costs and is an important evolutionary strategy in resource-limited ecosystems.

Introduction

Biosynthetic cost minimization in microorganisms is a reduction in amino-acid biosynthetic costs to maximize central metabolic costs and is thought to be a necessary, adaptive strategy under resource limitation in natural environments (Seligmann, 2003). Thus, microorganisms generally first use amino acids with lower molecular weights and lower assimilation costs, especially in less critical protein domains (Barrai *et al.*, 1995; Dufton, 1997). Much of the ocean is resource limited, especially in nitrogen (N) and iron (Fe). Fe availability limits nitrate utilization and primary productivity in the ocean (Martin *et al.*, 1994; Price *et al.*, 1994). Low or fluctuating N availability results in multiple N-saving strategies in *Prochlorococcus* (Lv *et al.*, 2008; Gilbert and Fagan, 2010). Here we show that in oligotrophic, open-ocean regions of low N or Fe, there is a preferential reduction in N side chains compared with carbon (C) in amino-acid sequences, with a concomitant increase in the average mass of proteins. These findings, derived from 41 Global Ocean Survey (GOS) sites (more than 20 million protein fragments) and whole-genome sequences, suggest the importance of N limitation as a selective force on molecular evolution of codon usage and protein atomic composition in microorganisms that are well adapted to oligotrophic ocean conditions.

Why might N cost minimization be so important in the ocean with myriad adaptational strategies available to organisms? The dominant form of N in the modern ocean is nitrate, and mean annual concentrations in the open ocean are low (Figure 1). The ratio of available N to phosphorus (P) in the oceans is modulated across geological timescales by N fixation, which, in turn, is controlled by the oxidation state of the ocean and supply of trace metals—especially Fe (Falkowski, 1997); P availability is a function of continental weathering and fluvial discharge (Broecker, 1982). In areas not impacted by significant N discharge, net N fixation is less than denitrification, causing the dissolved pool of inorganic nutrients in the ocean to be slightly enriched in P compared to N (Falkowski, 1997). Thus, it is assumed that most of the modern ocean is oligotrophic and in a perpetual state of N deficiency—areas of exception include upwelling zones and large parts of the Southern Ocean (Figure 1). The tropical Pacific, for example, is warm, stratified and oligotrophic, with low levels of nitrate and Fe (Behrenfeld *et al.*, 2006). The Sargasso Sea in the Atlantic Ocean is an oligotrophic subtropical gyre where P and N at various times of the year are limiting (Wu *et al.*, 2000). Areas of excess inputs of N include coastal margins where nutrient runoff results in eutrophication (Vitousek *et al.*, 1997).

In general, organisms in the open-ocean surface struggle to acquire enough N and have acquired or

Correspondence: JJ Grzymski, Division of Earth and Ecosystem Studies, Desert Research Institute, 2215 Raggio Parkway, Reno, NV 89512, USA.
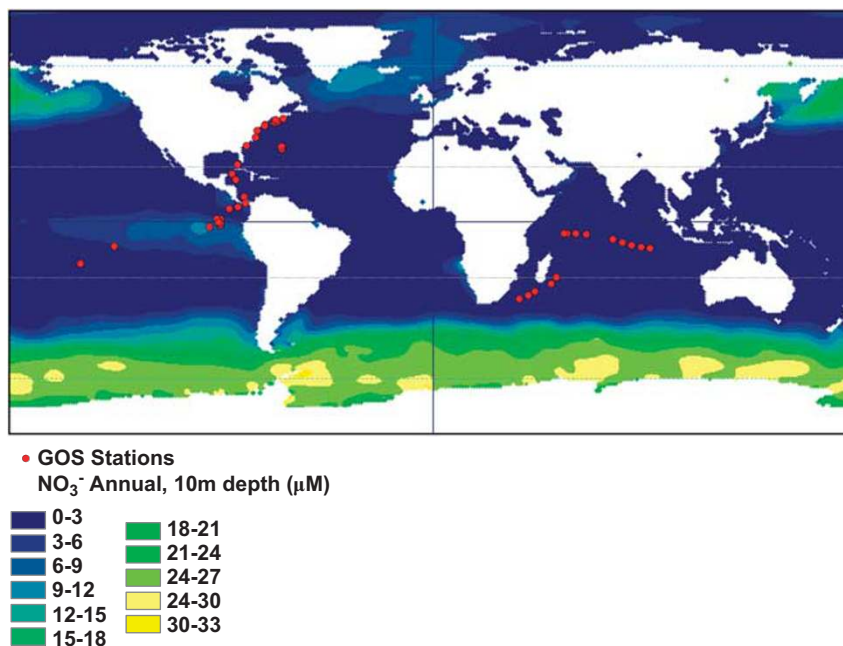E-mail: joeg@dri.edu

Figure 1 Annual 10 m nitrate concentration for world oceans. Red dots are GOS sampling sites used in our analyses.

evolved metabolic adaptations such as the ability to transport amino acids into the cell with high affinity, the use of various inorganic N sources or, in the case of diazotrophs, the ability to reduce N_2 to ammonia. These strategies are most effective when coupled with rapid, energy-saving adaptations—organisms assimilate highest-energy substrates first or store excess nutrients. When ammonium is available, for example, the diazotroph *Trichodesmium* limits N fixation, saving ~ 16 mol of ATP per mol of N_2 fixed (Mulholland *et al.*, 1999). Under N limitation, *Prochlorococcus* CCMP1986 ($=$MED4) quickly downregulates N-rich translation proteins, which differs from the strategy employed by the MIT9313 strain (Gilbert and Fagan, 2010). Regardless of the individual nutrient acquisition and assimilatory strategy, the longer-term evolutionary strategy and selective pressure that time in low-nutrient environments exerts is for organisms that face even transient nutrient stress to minimize usage of those compounds in proteins. This phenomenon should be particularly evident in marine microorganisms that inhabit oligotrophic waters. *Escherichia coli* and *Saccharomyces cerevisiae*, for example, have reduced C, sulfur and N in respective assimilatory proteins; this is an imprint on amino-acid composition from evolutionary nutritional constraints (Baudouin-Cornu *et al.*, 2001). We hypothesized the extent of N cost minimization to be pervasive in genomes of open-ocean microorganisms—proteins vital for survival in these oligotrophic conditions would be especially limited in N side-chain usage.

Open-ocean (oligotrophic) and coastal-ocean (copiotrophic) communities of microorganisms have adapted to their respective environments through a combination of genomic architecture differences (Lauro *et al.*, 2009) and metabolic (gene-specific) differences that confer advantages under different environmental conditions (Rocap *et al.*, 2003; Giovannoni *et al.*, 2005). For example, closely related *Prochlorococcus* species are differentially adapted to surface-ocean/high-light and deeper-water/low-light conditions (Rocap *et al.*, 2003), temperature (Johnson *et al.*, 2006) and Fe (Rusch *et al.*, 2010). Oligotrophic organisms tend to use broad-specificity, high-affinity transporters in comparison with copiotrophs that often contain myriad high-specificity transporters (Lauro *et al.*, 2009). Competition for resources makes high-affinity ABC transporters some of the most abundant proteins in nutrient-poor waters of the Sargasso Sea (Sowell *et al.*, 2008).

Biosynthesis of amino acids makes up a considerable portion of the cell energy budget (Akashi and Gojobori, 2002), yet few data are available on cost minimization in oligotrophic environments. The atomic composition of proteins reflects the assimilatory costs of nutrients. Organisms that are persistently resource-limited should have protein atomic compositions that reflect this selection pressure. Thus, the success of oligotrophic organisms—as varied as the diazotroph, *Trichodesmium*; the abundant, diverse cyanobacteria, *Prochlorococcus*; or the ubiquitous ultramicrobacterium, *Candidatus Pelagibacter ubique*—is due to a combination of multiple, short-term adaptational mechanisms for dealing with changing environmental conditions coupled with long-term genomic changes to enhance metabolic efficiency, streamline metabolism and reduce cellular costs. In systems that are N-limited over long timescales, this involves efficient N cost minimization of amino acids in proteins and thus favors low $G+C$ codon usage (Bragg and

Hyder, 2004). This effect is exaggerated compared with the organisms from environments with moderate relief from resource limitation, because these adaptations must come at a cost to other aspects of fitness.

Materials and methods

Marine microbial genomes
The Moore Marine Microbial data table, containing a list of marine microbes and pertinent information regarding them, was downloaded on 17 August 2010. The 139 available draft genomes from the Moore table and *Bermanella marisrubri* were downloaded from the National Center for Biotechnology Information on 18 August 2010. Nine available completed reference genomes from the Moore table were downloaded from GenBank on 1 July 2010: *Ca. Pelagibacter ubique* HTCC1062, *Prochlorococcus marinus* MIT9215, *P. marinus* MIT9312, *P. marinus* MIT9313, *P. marinus* NATL1A, *P. marinus* NATL2A, *P. marinus* CCMP1375, *P. marinus* CCMP1986 and *Trichodesmium erythraeum* IMS101. Not all sequenced marine microbes are found in the Moore table, but it is a good representation of cultivable marine microorganisms.

PathoSystems Resource Integration Center genomes
The 890 available complete genomes from the PathoSystems Resource Integration Center (Snyder *et al.*, 2007) database were downloaded on 4 October 2010. Complete genomes were identified by cross referencing with the genomes online database (Liolios *et al.*, 2006) downloaded concurrently.

Environmental data
All individual GOS for metagenomes and corresponding information were downloaded from the Community Cyberinfrastructure for Advanced Microbial Ecology Research & Analysis database on 9 June 2010. All collection sites that were described as coastal or open ocean were used for analysis, with the exception of GS001a and GS001b given potential contamination, and GS117b and GS000b–d were removed owing to different filter sizes being used in collection. Sites were then redefined as open- or coastal-ocean sites based on the distance from major landmasses, with the furthest coastal distance being approximately 104 km. In addition, the metagenome from the Phosphorus Removing (EBPR) Sludge Community collection in Australia was downloaded on 2 September 2010 from Cyberinfrastructure for Advanced Microbial Ecology Research & Analysis. Mean annual nitrate concentrations from 10 m depth were downloaded from the National Oceanographic Data Center (http://www.nodc.noaa.gov) and processed in ArcMAP 9.3.1 (ESRI Inc). G + C nucleotide data were obtained from Cyberinfrastructure for Advanced Microbial Ecology Research & Analysis curators.

Protein atomic composition analysis
Protein sequences from input files were analyzed based on the atomic composition of side chains. Each protein was analyzed separately for N, C, sulfur, hydrogen and oxygen content using the metric of number of x atoms per residue side chain (ARSC) as in Baudouin-Cornu *et al.* (2001). In addition, the metric of percent amino-acid usage was utilized as in Karlin and Brendel (1992). To compare different samples in these metrics, we determined the quantile distribution for all metrics and used $Q(x_{0.5})$, the median value, for a given collection of proteins.

Atomic composition analysis: PathoSystems Resource Integration Center versus Moore
All protein sequences from each database were collected and analyzed as above. Using pattern recognition, all proteins with 'transporter' contained within the protein name were extracted from each database and analyzed as above with all residues and with only non-hydrophobic residues. Analysis of variance in conjunction with the multiple comparison test was used for statistical comparison. The comparison of genes related to urea and glutamate transport and assimilation in *Prochlorococcus* were done on the following genes: acetylglutamate kinase, bifunctional ornithine acetyltransferase/N-acetylglutamate synthase protein, ferredoxin-dependent glutamate synthase, glutamate-1-semialdehyde aminotransferase, glutamine amidotransferase class-I, putative aspartate and glutamate racemase:glutamate racemase, putative ATP-binding subunit of urea ABC transport system, putative ATP-binding subunit of urea ABC transport system, putative bifunctional dihydrofolate/folylpolyglutamate synthase, putative urea ABC transporter, UDP-N-acetylmuramoyl-L-alanyl-D-glutamate synthetase urease accessory proteins UreD,E,F,G and urease subunits alpha, beta and gamma. To test for correlation, Spearman's rank correlation coefficient was calculated.

Circos plots
Gene locations were found on both leading and lagging strands using nucleotide annotations from GenBank and verified with pattern matching in simple cases, or an alignment algorithm in more difficult cases. In addition, amino-acid sequences were aligned with the corresponding nucleotide sequences based on name, file location and verified by translation of nucleotides. Using these data in combination with the N ARSC for each protein, Circos software circular plots were created with the following parameters: (1) a high (green) N content is anything greater than 0.408 N ARSC, (2) a low (red) N content is anything less than 0.318 N ARSC and (3) only proteins in the range (0.14, 0.59) N ARSC are shown (this remove 76/3986 data points but drastically improves visualizing the data). Values

were chosen based on the mean and standard deviation of N ARSC across both *P. marinus* MIT 9313 and *P. marinus* CCMP 1986 strains 0.363 (± 0.090) N ARSC. Subtraction plots were created using the *P. marinus* MIT 9313−*P. marinus* CCMP 1986 values of the non-hypothetical core proteins. Coloring was based on positive (green) versus negative (red) values.

Addition of distance from land data
The distance from major landmass was estimated using the Perl math package, specifically the great_circle_distance function with an average Earth radius of 6371 km. Site latitude and longitude coordinates were obtained from Cyberinfrastructure for Advanced Microbial Ecology Research & Analysis, and major landmass coordinates were determined using $\frac{1}{4}°$ cells that are mapped as continental landmasses in the ESRI ArcData data sets. These data were then used to redefine coastal- and open-ocean collection sites and plotted against the median N ARSC, showing a negative correlation.

Creation of overlapping genomes
All-against-all BLAST comparisons between two *Prochlorococcus* strains (MIT9313 vs CCMP 1986) were performed using protein sequences, and reciprocal genome-specific best hits were identified (*E*-value $<10^{-5}$). Core genome was determined from genes that were the best hit for the other.

Estimation of N requirements
Three parameters defined the model—G + C content (G_i), cell size (S_i) and genome size (l_i). Normal G + C content was based on the G + C content of the GOS data used in these analyses (34% ± 4%), and high G + C content was based on *P. marinus* MIT 9313 (50% ± 4%). Cell size was based on the work of Gundersen *et al.* (2002), which determined the geometric mean of size was 0.048 μm³ for bacteria from a surface water Bermuda Atlantic Time Series Station sample, with an approximate lognormal distribution. Cell sizes were converted to a log (base 10) scale: average was −1.32 (± 0.02) (0.048 μm³), small cell size was −2 (± 0.02) (0.010 μm³) and large cell size was −1.21 (± 0.02) (0.062 μm³). Genome size (2 ± 0.2 Mb) of ocean samples was based on the estimates done by Frank and Sorensen (2011). Large genome size was set as the average of all genomes in Supplementary Table S2 (9 ± 0.2 MB).

Given these parameters, N content in one copy of the genome was calculated as follows:

$$\left[\begin{array}{l} G_i \cdot l_i \text{ number of G + C bp} \cdot \frac{1\,\text{mol G+C bp}}{6.022\times10^{23}\,\text{G+C bp}} \cdot \frac{7\,\text{mol N}}{1\,\text{mol G+C bp}} + \\ (1-G_i) \cdot l_i \text{ number of G + C bp} \cdot \frac{1\,\text{mol A+T bp}}{6.022\times10^{23}\,\text{A+T bp}} \cdot \frac{8\,\text{mol N}}{1\,\text{mol A+T bp}} \end{array} \right] \cdot \frac{10^{15}\,\text{mol}}{1\,\text{mol}} \quad (1)$$

The relationship between cell size and protein content for ocean microorganisms was calculated

based on published data (Simon and Azam, 1989) as follows, with m_{tp_i} being total mass of cellular proteins for the *i*th cell and S_i being cell size:

$$m_{tp_i}\,(\text{fg}) = 88.6 S_i\left(\mu m^3\right)^{0.59} (R^2 = 0.67) \quad (2)$$

The relationships between G + C content and N_{ARSC_i}, N ARSC for the *i*th cell, and $\bar{m}_{aa_i}$, average mass per residue for the *i*th cell, were calculated based on Supplementary Table S2:

$$\text{N}_{ARSC_i} = 0.0009 G_i + 0.3071 (R^2 = 0.356) \quad (3)$$

$$\frac{\bar{m}_{aa_i}(\text{g})}{\text{mol res}} = -0.1514 G_i + 118.18 \,(R^2 = 0.875) \quad (4)$$

Finally, these values were used to calculate the amount of N needed for the growth of one cell with the following stoichiometric ratios $(1 + N_{ARSC_i})$ accounts for protein backbone N content:

$$\begin{array}{l} m_{tp_i} \frac{1\text{g}}{10^{15}\text{fg}} \cdot \frac{1\,\text{mol res}}{\bar{m}_{aa_i}\text{g}} \cdot \frac{6.022\times10^{23}\,\text{res}}{1\,\text{mol res}} \\ \cdot \frac{1 + \text{N}_{ARSC_i}\,\text{N atoms}}{1\,\text{res}} \cdot \frac{1\,\text{mol N}}{6.022\times10^{23}\,\text{N atoms}} \\ \cdot \frac{10^{15}\text{fmol}}{1\,\text{mol}} \end{array} \quad (5)$$

Ten populations of 200 theoretical cells (parameters based on normal (lognormal for size) distributions with given mean and standard deviation as above) were averaged to determine approximate changes in N requirements for cells of different sizes and G + C contents.

Estimation of ocean diversity
In order to verify the most abundant organisms in GOS, both coastal- and open-ocean, GOS protein fragments downloaded above were grouped by ocean location (coastal 8 606 573/open 6 995 092) and 100 random files of 20 000 protein fragments of at least 50 residues were compared with concatenated marine microbial genomes described above using BLAST. The top hit corresponding to each organism for each protein fragment (*E*-value $<10^{-4}$) was counted. This was used to calculate average percentages and standard deviations for organism occurrence.

Results and discussion

Recent environmental genome sequencing from ocean habitats (Venter *et al.*, 2004) makes it possible to test if N cost-minimization is imprinted on genomes of microorganisms that inhabit oligotrophic environments where a majority of the surface water is depleted in NO_3^- (Figure 1). The atomic composition of more than 20 million protein
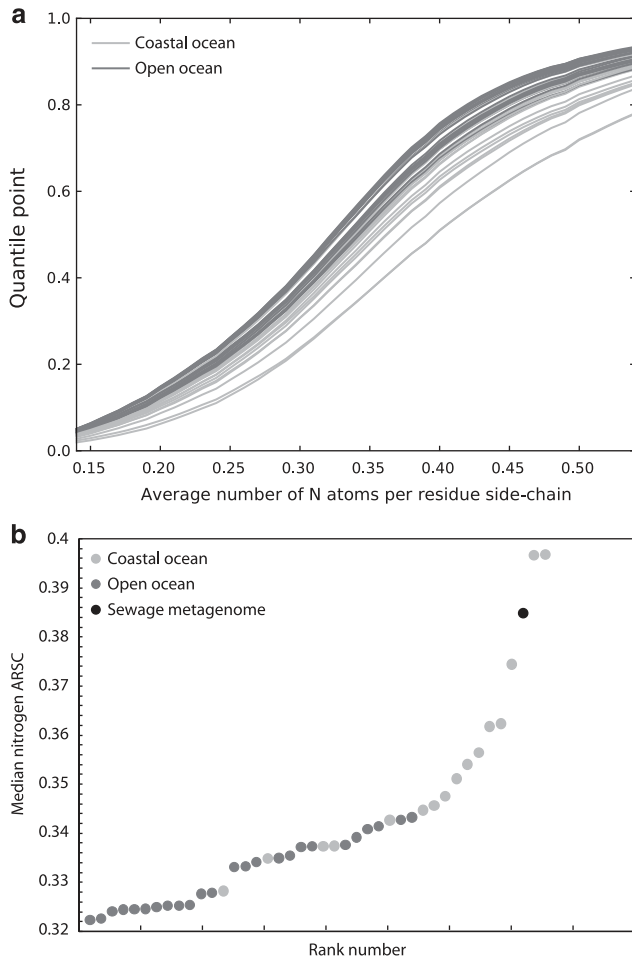
Figure 2 Nitrogens per amino-acid residue side chain for GOS environmental genome data. (**a**) Quantile plots of the average number of nitrogen ARSC for GOS stations. Red lines represent coastal-ocean sites from the GOS data set as defined in Materials and methods. Blue lines are from open-ocean sites. (**b**) The median N ARSC for each GOS site and one from a sewage metagenome. The color reproduction of this figure is available at the *ISME Journal* online.

Figure 3 Median nitrogen ARSC as it correlates to proximity of landmass. The correlation between the nearest land mass and median N ARSC for each GOS metagenome station. Spearman correlation coefficients are presented in Supplementary Table S1.

fragments from 15 coastal-ocean and 26 open-ocean sites from the GOS data set were analyzed (Rusch *et al.*, 2007). For each site, we calculated the average number of N and C ARSC of each predicted protein. From these data, quantile distributions for N and C:N were calculated (Baudouin-Cornu *et al.*, 2001). We show a selective reduction of N (and a concomitant increase in C:N) in side chains of amino acids from open-reading frames (ORFs; predicted proteins) of open-ocean microorganisms compared with organisms found in coastal-ocean environments (Figure 2a). The medians for all 41 stations are plotted in Figure 2b, where it is clearly evident that below 0.34 N ARSC the samples are predominantly open-ocean GOS sites. A sewage metagenome site—where there is presumably no N limitation—is plotted in Figure 2b for comparison. Similarly, median N ARSC is significantly and negatively correlated with distance from land and
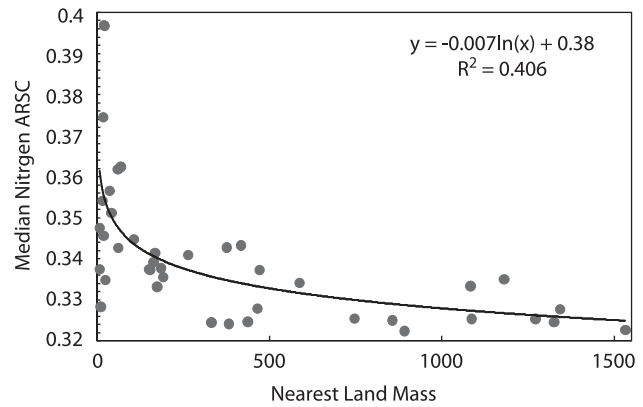
positively correlated with 10 m annual nitrate concentration (Figure 3 and Supplementary Table S1). These data suggest that organisms living in open-ocean environments have reduced N content in their protein biomass, which can account for up to 50% of the cellular mass (*cf.* Cox (2004) for an in-depth analysis).

There is a strong relationship between the increased N ARSC as microbial communities are sampled closer to coastlines (Supplementary Table S1), but environmental genomes sampled for the GOS are only an estimate of the population through DNA sequencing. More than likely, this has little relationship to *in situ* physiology and the pathways and proteins that are active and abundant in these cells. A compilation of data on marine microbes (curated from the Moore Marine microbe sequencing initiative), focusing on atomic properties of the predicted protein, reveals the diversity of amino acid usage, G+C content and genome size of organisms found in coastal and open oceans (Supplementary Table S2). These organisms are ranked based on bootstrapped estimates of their presence in GOS coastal- and open-ocean samples. Not surprisingly, there are significant relationships between the G+C content of genomes and amino-acid usage and N ARSC (Supplementary Table S1). For example, the commonly substituted (biochemically similar; Bordo and Argos, 1991) amino acids lysine (K) and arginine (R) are significantly correlated with G+C (Supplementary Table S1). K is inversely related while R is positively correlated. Thus, low G+C-containing organisms use K over R preferentially and save 2 mol N for each mol of K used over R. This significantly impacts the N budget of the cell—the modeled amount of N needed in proteins and nucleic acids is positively correlated with G+C (Table 1). This results from codon preference and amino-acid substitutions over evolutionary time favoring less nitrogenous amino acids.

Table 1 Protein requirements in theoretical marine microorganism populations determined by modeling

Description	Cell size (μm³)	Genome size (Mb)	G+C (%)	Total protein mass (fg cell⁻¹)	N ARSC	Average mass (g mol⁻¹)	Nitrogen requirement for protein (fmol cell⁻¹)	Nitrogen requirement for DNA (fmol cell⁻¹)	Nitrogen requirement for protein+DNA (fmol cell⁻¹)
Average open ocean	0.0488 ± 0.0008	2.01 ± 0.01	34.0 ± 0.2	14.833 ± 0.145	0.3377 ± 0.0002	113.03 ± 0.04	34.443 ± 0.337	4.802 ± 0.029	39.245 ± 0.346
Open-ocean high G+C	0.0495 ± 0.0007	2.00 ± 0.01	49.9 ± 0.2	14.941 ± 0.130	0.3521 ± 0.0002	110.62 ± 0.03	35.831 ± 0.308	4.898 ± 0.036	40.729 ± 0.304
Open-ocean high cell volume	0.0634 ± 0.0011	2.00 ± 0.01	34.1 ± 0.3	17.282 ± 0.180	0.3378 ± 0.0003	113.02 ± 0.05	40.138 ± 0.425	4.785 ± 0.033	44.923 ± 0.435
Open-ocean large genome size	0.0491 ± 0.0008	3.90 ± 0.01	34.1 ± 0.2	14.875 ± 0.130	0.3377 ± 0.0002	113.02 ± 0.04	34.543 ± 0.313	9.332 ± 0.027	43.875 ± 0.306
Ideal open ocean	0.0103 ± 0.0002	2.01 ± 0.02	34.1 ± 0.3	5.929 ± 0.060	0.3378 ± 0.0003	113.02 ± 0.05	13.769 ± 0.142	4.796 ± 0.044	18.565 ± 0.169
'Coastal' ocean	0.0635 ± 0.0009	3.90 ± 0.02	49.9 ± 0.3	17.303 ± 0.140	0.3521 ± 0.0002	110.62 ± 0.04	41.495 ± 0.330	9.526 ± 0.040	51.021 ± 0.318

Abbreviations: ARSC, atoms per residue side chain; N, nitrogen.
Parameters and calculations based on observed relationships and previously published data as described in Materials and methods.

Nitrogen costs, on average, decrease as G + C content, cell size and genome size decrease (Table 1). These changes are in addition to short-term adaptive mechanisms that cells utilize to conserve N during N limitation (Gilbert and Fagan, 2010).

Three important organisms in the surface open ocean—Ca. P. ubique, P. marinus (specifically, strains MIT9312 and CCMP1986) and T. erythaeum—all have low G + C content, despite broad phylogenetic and metabolic differences. The SAR11 clade of alphaproteobacteria (of which Ca. P. ubique is a member) are some of the most abundant organisms in the ocean (Giovannoni et al., 2005). These genomes are characterized by reduced N ARSC and increased average mass of amino acids. Low G + C codon usage and lower N ARSC come at a cost: increased average mass of amino acids (Baudouin-Cornu et al., 2004; Supplementary Tables S1 and S2). There is a paradox if we consider cost minimization based on ATP, NADH or other energetic units: these costs streamline metabolic efficiency and decrease the average mass of amino acids, especially in highly expressed genes (Akashi and Gojobori, 2002). Our results from analyzing the GOS data set (Supplementary Tables S1 and S3) and the success of Pelagibacter, Prochlorococcus and Trichodesmium in the open ocean (Campbell et al., 1994; Capone et al., 1997; Morris et al., 2002) suggest that N cost minimization is crucial in microbes inhabiting marine surface waters despite potential decreases in fitness from the increased average mass of proteins. Not one simple mechanism—loss of DNA-repair enzymes as hypothesized in Prochlorococcus CCMP1986 (Rocap et al., 2003) or other non-repair-based mechanisms in Pelagibacter (Giovannoni et al., 2005)—can explain G + C reduction in genomes of dominant marine microbes. Thus, the convergence of outcomes of many successful open-ocean microbes with low G + C content adds to a parsimonious explanation of N conservation. Regardless of the mechanism, the net effect is a trade-off between the decreased N content in proteins (and only slightly so in nucleic acids, <10% of the change from 50% to 34% G + C is attributable to nucleotides, given all other things held constant (Table 1)) and an increase in biosynthetic cost (for example, ATP and NADH) of larger amino acids.

Differences in N content of proteins and average mass of amino acids between P. marinus MIT9313 (higher G + C contents, low-light adapted) and P. marinus CCMP1986 (lower G + C contents, high-light adapted; Figure 4) are not a function of gene content. The average N ARSC for each protein is significantly different between the two strains (Figure 4), predominantly due to amino-acid usages based on G + C content. However, Prochlorococcus strains share a large core genome. We compared only the 1286 ORFs in common among the two ecotypes and homologous proteins involved in urea and glutamate metabolism and assimilation
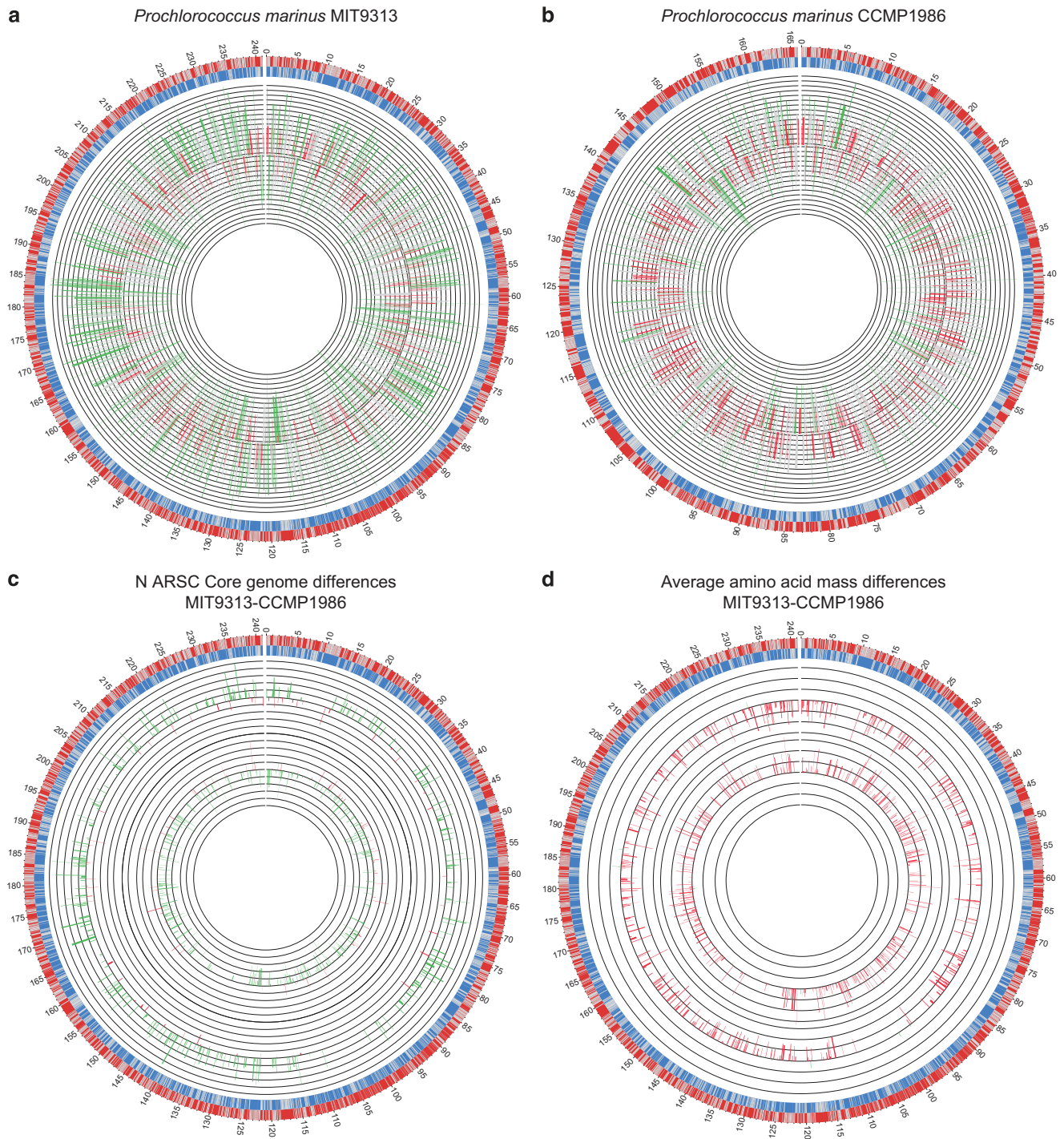
a *Prochlorococcus marinus* MIT9313

b *Prochlorococcus marinus* CCMP1986

c N ARSC Core genome differences
MIT9313-CCMP1986

d Average amino acid mass differences
MIT9313-CCMP1986

Figure 4 Circular genome plots of *Prochlorococcus marinus* ecotypes. (**a**) Coding (gray) and non-coding regions (red = leading strand, blue = lagging strand) for the *P. marinus* ecotype MIT 9313. Additional data are available in Supplementary Tables S4 and S5c. Averaged nitrogen ARSC for each ORF is plotted as a colored bar graph, with green indicating high nitrogen ARSC and red indicating lower as in Materials and methods. The concentric rings scale represents an increase of 0.03 N ARSC. (**b**) As in (**a**), but for the *P. marinus* ecotype CCMP 1986. (**c**) As in **a**, but bar graphs represent subtraction of strain MIT9313—CCMP1986 N ARSC values for all genes in the core genome (1286) excluding hypothetical proteins. Green indicates positive values and red indicates negative values. The concentric rings scale represents changes of 0.03 N ARSC. (**d**) Same as **c** except representing the average amino-acid mass difference for strain MIT9313—CCMP1986. The concentric rings scale represents changes of $3 \, g \, mol^{-1}$.

(Supplementary Table S4; Figure 4). The high-light strain (CCMP 1986) that only grows on ammonium has significantly fewer N atoms in proteins but must build on average, larger proteins. Proteins in the core genome are almost 3% larger by mass ($P < 0.01$, Supplementary Table S5c).

Since there is a fitness trade-off between the reduced N content of proteins and metabolic cost increases of making large amino acids in those proteins, we expect there to be a more pronounced balance of N and energy cost in proteins that are highly expressed and highly conserved across phylogenetic boundaries. The metaproteome of SAR11-related organisms from the oligotrophic Sargasso Sea is dominated by transport proteins, chaperones and membrane proteases (Sowell *et al.*, 2008). These data suggest that microbial community function in resource-limited environments is focused on resource acquisition, assimilation and protein maintenance compared with resource-rich environments dominated by ribosome synthesis and other activities to increase the growth rate (Lindahl and Zengel, 1986; Cox, 2004).

We compared transport, chaperones and N-assimilation proteins as subsets of the whole genome N ARSC, hypothesizing that the link between resources and cost-minimization would be substantiated in abundant and highly conserved proteins that were further N-cost minimized. The median N ARSC value for 44 transporter-related proteins in *Ca. P. ubique* is 15% lower (*t*-test $P < 0.01$) than the median N ARSC for the rest of the genome. This is partially a result of a high frequency of hydrophobic amino-acid usage in membrane-spanning regions, but there is also tremendous plasticity in transporter N ARSC values ($10 \times$ variation across all marine microbe transporters). Lower N ARSC values in transporters for *Ca. P. ubique* represent a savings of 14 mol N per mol avg. transporter protein compared with a mole of proteins of other function in an already cost-minimized genome (assuming 300 residues per protein). These proteins only nominally affect the N-budget and can be abundantly synthesized even under severe N stress; they are among the most abundant proteins identified in oligotrophic waters. Given the selective pressure to get energetic materials into the cell, especially in resource-limited environments, and the low N content of these proteins, we hypothesize that transporters evolved and diversified rapidly in these low-nutrient environments. Transporters are cost-minimized with respect to the whole proteome, but this is only partially due to usage of hydrophobic residues. They are significantly cost-minimized for N when hydrophobic residues are excluded and also when marine transporters are compared with a broader group of microorganisms (Supplementary Table S5). Thus, cells investing heavily in transport functions under N-limited conditions would have even lower N budgets for the cell than estimated in Table 1.

The membrane protease HflKC and chaperone proteins GroEL, GroES and DnaK also were abundant in the environmental proteome of the Sargasso Sea, indicating the probable importance of protein refolding and proteolysis under constant environmental stress (Sowell *et al.*, 2008). The median N

savings in this group of proteins in marine microorganisms versus non-marine organisms is 6 mol N per mol avg. protein. Previous analysis of *E. coli* and *S. cereviseae* genomes suggested that selection pressure to reduce the respective atoms in C, N and S assimilatory protein was high (Baudouin-Cornu *et al.*, 2001). N cost minimization in the ocean is important and not limited to assimilatory proteins. Genes affected by this selective pressure should include environmentally relevant (and highly expressed) genes. N cellular budgets should be closely examined in the light of these findings.

What is the evolutionary/geological time context for N cost minimization? The divergence of *Synechococcus* and *Prochlorococcus* approximately 150 mya and subsequent divergence of the high-light ecotype of *Prochlorococcus* from the more diverse low-light clade (Dufresne *et al.*, 2005) ~ 80 mya are a result of the rise in dominance of eukaryotic phytoplankton and changing biogeochemical conditions affecting marine microbes (Falkowski *et al.*, 2004). The Mesozoic Era (~ 250–65 mya) saw a diversification of red plastid-containing phytoplankton in the ocean (Falkowski *et al.*, 2004), fundamentally altering the ecology of the surface ocean. This shift to stochastic, opportunistic resource utilization and greater episodic export events put significant pressure on cyanobacteria by changing the nutrient profile for silica, nitrate and dissolved Fe in the upper water column. The decrease in N and Fe had to be a significant selective pressure on cyanobacteria and helped cause differentiation of *Prochlorococcus* from *Synechococcus* ~ 150 mya (Dufresne *et al.*, 2005). These ancestral *Prochlorococcus* cells were probably very similar to strains MIT9313 and MIT9303, were low-light adapted and found success away from the increasing area of continental margins that occurred in the late Cretaceous. Low-light-adapted *Prochlorococci* used dissolved free amino acids, NO_2^- and ammonium, as sources of N. Genome reduction and changes to DNA repair mechanisms possibly involved in $G + C$–$A + T$ mutations further streamlined the genomes, caused ultra-specialization, and maximized metabolic efficiency. A consequence of decrease in inorganic N and dissolved Fe (either from increased diatom productivity or from specialization in oligotrophic water) was pressure to further reduce metabolic costs in the cell associated with these elements. Thus, transporters, N-assimilation proteins and chaperones are all reduced in N compared with already streamlined genomes. Diversification into high-light environments inherently reduced the Fe quota in the cell, thereby offsetting the move to an even lower dissolved Fe environment (the upper 10 m of the water column). Further adaptations to Fe-depleted waters were revealed in two newly identified high-light, low-Fe clades of *Prochlorococcus* from parts of the Pacific and Indian Oceans (Rusch *et al.*, 2010). A combination of genome reduction, use of $A + T$-rich codons

and accelerated rates of evolution in proteins (Dufresne *et al.*, 2005) that clearly favored reduced N ARSC are hallmarks of the evolution of *Prochlorococcus*. The combination of high surface area-to-volume ratio and minimized N in proteins makes the high-light-adapted clade of *Prochlorococcus* one of the most abundant organisms in warm, oligotrophic surface waters (Partensky *et al.*, 1999). In fact, this combination contributes to the success of *Trichodesmium*, *Ca. P. ubique*, certain *Prochlorococcus* and other microorganisms that seek refuge in oligotrophic oceans. Streamlined genomes, A+T-rich codon usage and N-cost-minimized proteins—especially in highly expressed genes—make these organisms well adapted to oligotrophic conditions. More work clearly needs to be done on the intriguing links between N ARSC, G+C content and Fe quotas, as they are integral to oligotrophic ocean photoautotrophy.

These results suggest a rule of success in oligotrophic oceans: minimize N costs in proteins. Nitrogen cost minimization represents a potentially stronger selective pressure on protein atomic composition than biosynthetic cost minimization to the lower average mass of amino acids.

References

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

Barrai I, Volinia S, Scapoli C. (1995). The usage of oligopeptides in proteins correlates negatively with molecular weight. *Int J Pept Prot Res* **45**: 326–331.

Baudouin-Cornu P, Schuerer K, Marliere P, Thomas D. (2004). Intimate evolution of proteins—proteome atomic content correlates with genome base composition. *J Biol Chem* **279**: 5421–5428.

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

Behrenfeld MJ, Worthington K, Sherrell RM, Chavez FP, Strutton P, McPhaden M *et al.* (2006). Controls on tropical Pacific Ocean productivity revealed through nutrient stress diagnostics. *Nature* **442**: 1025–1028.

Bordo D, Argos P. (1991). Suggestions for 'safe' residue substitutions in site-directed mutagenesis. *J Mol Biol* **217**: 721–729.

Bragg JG, Hyder CL. (2004). Nitrogen versus carbon use in prokaryotic genomes and proteomes. *P Roy Soc Lond B Bio* **271**: S374–S377.

Broecker WS. (1982). Glacial to interglacial changes in ocean chemistry. *Prog Oceanogr* **11**: 151–197.

Campbell L, Nolla HA, Vaulot D. (1994). The importance of *Prochlorococcus* to community structure in the central north Pacific Ocean. *Limnol Oceanogr* **39**: 954–961.

Capone DG, Zehr JP, Paerl HW, Bergman B, Carpenter EJ. (1997). *Trichodesmium*, a globally significant marine cyanobacterium. *Science* **276**: 1221–1229.

Cox RA. (2004). Quantitative relationships for specific growth rates and macromolecular compositions of *Mycobacterium tuberculosis*, *Streptomyces coelicolor* A3(2) and *Escherichia coli* B/r: an integrative theoretical approach. *Microbiol* **150**: 1413–1426.

Dufresne A, Garczarek L, Partensky F. (2005). Accelerated evolution associated with genome reduction in a free-living prokaryote. *Genome Biol* **6**: R14.

Dufton MJ. (1997). Genetic code synonym quotas and amino acid complexity: cutting the cost of proteins? *J Theor Biol* **187**: 165–173.

Falkowski PG. (1997). Evolution of the nitrogen cycle and its influence on the biological sequestration of CO_2 in the ocean. *Nature* **387**: 272–275.

Falkowski PG, Katz ME, Knoll AH, Quigg A, Raven JA, Schofield O *et al.* (2004). The evolution of modern eukaryotic phytoplankton. *Science* **305**: 354–360.

Frank J, Sorensen SJ. (2011). Quantitative metagenomic analyses based on average genome size normalization. *Appl Environ Microbiol* **77**: 2513–2521.

Gilbert JDJ, Fagan WF. (2010). Contrasting mechanisms of proteomic nitrogen thrift in *Prochlorococcus*. *Mol ecol* **20**: 92–104.

Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D *et al.* (2005). Genome streamlining in a cosmopolitan oceanic bacterium. *Science* **309**: 1242–1245.

Gundersen K, Heldal M, Norland S, Purdie Da, Knap AH. (2002). Elemental C, N, and P cell content of individual bacteria collected at the Bermuda Atlantic Time-series Study (BATS) site. *Limnol Ocenogr* **47**: 1525–1530.

Johnson ZI, Zinser ER, Coe A, McNulty NP, Woodward EM, Chisholm SW. (2006). Niche partitioning among Prochlorococcus ecotypes along ocean-scale environmental gradients. *Science* **311**: 1737–1740.

Karlin S, Brendel V. (1992). Chance and statistical significance in protein and DNA sequence analysis. *Science* **257**: 39–49.

Lauro FM, McDougald D, Thomas T, Williams TJ, Egan S, Rice S *et al.* (2009). The genomic basis of trophic strategy in marine bacteria. *Proc Natl Acad Sci USA* **106**: 15527–15533.

Lindahl L, Zengel JM. (1986). Ribosomal genes in Escherichia coli. *Annu Rev Genet* **20**: 297–326.

Liolios K, Tavernarakis N, Hugenholtz P, Kyrpides NC. (2006). The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide. *Nucleic Acids Res* **34**: D332–D334.

Lv J, Li N, Niu D-K. (2008). Association between the availability of environmental resources and the atomic composition of organismal proteomes: evidence from *Prochlorococcus* strains living at different depths. *Biochem Biophys Res Comm* **375**: 241–246.

Martin JH, Coale KH, Johnson KS, Fitzwater SE, Gordon RM, Tanner SJ *et al.* (1994). Testing the iron hypothesis in ecosystems of the equatorial Pacific Ocean. *Nature* **371**: 123–129.

Morris RM, Rappe MS, Connon SA, Vergin KL, Siebold WA, Carlson CA *et al.* (2002). SAR11 clade dominates ocean surface bacterioplankton communities. *Nature* **420**: 806–810.

Mulholland MR, Ohki K, Capone DG. (1999). Nitrogen utilization and metabolism relative to patterns of N-2 fixation in cultures of *Trichodesmium* NIBB1067. *J Phycol* **35**: 977–988.

Partensky F, Hess WR, Vaulot D. (1999). *Prochlorococcus*, a marine photosynthetic prokaryote of global significance. *Microbiol Mol Biol Rev* **63**: 106–127.

Price N, Ahner BA, Morel FMM. (1994). The equatorial Pacific Ocean: Grazer-controlled phytoplankton populations in an iron-limited ecosystem. *Limnol Oceanogr* **39**: 520–534.

Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, Ahlgren NA *et al.* (2003). Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* **424**: 1042–1047.

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S *et al.* (2007). The sorcerer II global ocean sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol* **5**: e77.

Rusch DB, Martiny AC, Dupont CL, Halpern AL, Venter JC. (2010). Characterization of *Prochlorococcus* clades from iron-depleted oceanic regions. *Proc Natl Acad Sci USA* **107**: 16184–16189.

Seligmann H. (2003). Cost-minimization of amino acid usage. *J mol evol* **56**: 151–161.

Simon M, Azam F. (1989). Protein content and protein synthesis rates of planktonic marine bacteria. *Mar ecol prog ser* **51**: 201–213.

Snyder EE, Kampanya N, Lu J, Nordberg EK, Karur HR, Shukla M *et al.* (2007). PATRIC: The VBI PathoSystems Resource Integration Center. *Nucleic Acids Res* **35**: D401–D406.

Sowell SM, Wilhelm LJ, Norbeck AD, Lipton MS, Nicora CD, Barofsky DF *et al.* (2008). Transport functions dominate the SAR11 metaproteome at low-nutrient extremes in the Sargasso Sea. *ISME J* **3**: 93–105.

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA *et al.* (2004). Environmental genome shotgun sequencing of the Sargasso sea. *Science* **304**: 66–74.

Vitousek PM, Aber JD, Howarth RW, Likens GE, Matson PA, Schindler DW *et al.* (1997). Human alteration of the global nitrogen cycle: Sources and consequences. *Ecol Applicat* **7**: 737–750.

Wu JF, Sunda W, Boyle EA, Karl DM. (2000). Phosphate depletion in the western North Atlantic ocean. *Science* **289**: 759–762.

Supplementary Information accompanies the paper on The ISME Journal website (http://www.nature.com/ismej)