

Nitrogen versus carbon use in prokaryotic genomes and proteomes

Jason G. Bragg^{1*} and Charles L. Hyder²

¹Department of Biology, University of New Mexico, Albuquerque, NM 87131, USA

²412 Alvarado SE, Apartment 415 West, Albuquerque, NM 87108, USA

* Author for correspondence (jbragg@unm.edu).

Recd 24.01.04; Accptd 03.03.04; Published online 22.04.04

There is growing recognition that the elemental composition of genomes and proteins can be related to resource limitation. We examine the possibility that the elemental composition of nucleic acids and the amino acids (and proteins) they encode are correlated. We report a positive association between the stoichiometric ratio of N/C content of individual amino acids and their codons. Potentially, this is an outcome of chemical interactions between amino acids and anticodons that influenced the evolution of the genetic code. We also find a strong, positive relationship between N/C values of whole genomes and proteomes, across 94 prokaryotic species. This relationship is part of a spectrum in nitrogen versus carbon use across genomes and proteomes, which is correlated with genomic GC content. GC content is correlated positively with average nitrogen use, and negatively with average carbon use, across both genomes and proteomes.

Keywords: nitrogen; carbon; genome; proteome; GC content; ecological genomics

1. INTRODUCTION

The five nucleotides used in nucleic acids and 20 amino acids used in proteins contain different quantities of important nutrients. Therefore, the composition of nucleic acids (McEwan *et al.* 1998) and proteins (Mazel & Marlière 1989; Baudouin-Cornu *et al.* 2001) influence nutrient use, and may be related to the availability of nutrients in the environment. In *Escherichia coli* and *Saccharomyces cerevisiae*, proteins used for assimilating elements (carbon and sulphur, as well as nitrogen in *S. cerevisiae* only) tend to be relatively poor in those elements (Baudouin-Cornu *et al.* 2001). This probably helps organisms to use those elements sparingly when they are scarce (Baudouin-Cornu *et al.* 2001).

Among different species of prokaryotes, the proportion of guanine plus cytosine base pairs in genomes (GC content) varies widely (*ca.* 25–75%; Sueoka 1961). Elevated GC content has been related to atmospheric nitrogen fixation across species from eight genera of prokaryotes, and within genera of aerobic prokaryotes (but not within anaerobic genera) (McEwan *et al.* 1998). This was attributed to GC bonds having eight nitrogen atoms whereas AT bonds contain only seven (McEwan *et al.* 1998). However, greater use of nitrogen in GC bonds is probably not

sufficient to drive this pattern alone, as a large proportion of whole-organism nitrogen is in proteins (Sterner & Elser 2002).

This leads us to predict that prokaryotes with nitrogen-rich (GC rich) genomes have nitrogen-rich proteomes. Among prokaryotes, GC content influences amino acid composition (Sueoka 1961), such that amino acids with GC-rich codons are used more frequently by organisms with GC-rich genomes (Singer & Hickey 2000; Knight *et al.* 2001). Therefore, a relationship between nitrogen content of whole genomes and proteomes would probably result from a correlation between the nitrogen content of amino acids and their codons. Associations between chemical properties of amino acids and their anticodons have been reported previously, and related to the evolution of the genetic code (Jungck 1978; Lacey & Mullins 1983; see Knight *et al.* (1999) for a recent review of genetic code evolution). If these chemical properties of amino acids and anticodons are correlated with their atomic composition, this may lead to associations between amino acids and codons in their content of ecologically important elements.

We test the prediction that nitrogen use is correlated between whole genomes and proteomes, across 94 species of prokaryotes. We also examine the relationship between nitrogen content of individual amino acids and their codons. In examining these relationships we adopt the approach of ecologists (Sterner & Elser 2002) in considering the stoichiometric ratios of nitrogen : carbon atoms in these different molecules.

2. MATERIAL AND METHODS

(a) Codons and amino acids

We counted nitrogen and carbon atoms in the five nucleotides: A, C, G, T and U. These counts were added to obtain counts of nitrogen and carbon atoms in codons. This was carried out for three versions of codons: (i) single-stranded RNA (three nucleotides); (ii) single-stranded RNA, considering only nucleotides in the first two codon positions; and (iii) double-stranded DNA (six nucleotides). N/C values of codons (ν_C) were calculated as the ratio of nitrogen atoms to carbon atoms for the codon. Nitrogen and carbon atoms were counted for each of the 20 coded amino acids, and N/C values (ν_{AA}) were calculated as the ratio of nitrogen atoms to carbon atoms for each amino acid. We tested relationships between N/C values of amino acids and their codons (for single-stranded RNA codons with three nucleotides, and for double-stranded DNA codons; $n = 61$) using Spearman's rank correlation (SPSS v. 11.0).

We obtained average hydrophobicity rankings for amino acids and for anticodon dinucleoside monophosphates from Lacey & Mullins (1983). We then tested relationships between N/C values of amino acids and their hydrophobicity rankings ($n = 20$), and between N/C values of codons (calculated based on the first two codon positions) and the hydrophobicity rankings of their anticodon dinucleoside monophosphates ($n = 16$), using Spearman's rank correlation (SPSS v. 11.0).

(b) Genomes and proteomes

Counts of nucleotides A, C, G and T in genomes were obtained from the GenBank FTP site (<ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria>). Accession numbers are provided in electronic Appendix A. These counts were summed for multiple chromosomes, but plasmids were excluded to avoid uncertainties of copy number. Total numbers of nitrogen atoms and carbon atoms were then calculated for each genome. These are given by $N_G = l_G(P_G + 7)$ and $C_G = l_G(-P_G + 20)$, where l_G is the length of the genome in base pairs, and P_G is the guanine plus cytosine content of the genome, expressed as a proportion. Genomic N/C values (ν_G) were calculated as the ratio of total nitrogen atoms to total carbon atoms for the genome, which is given by $\nu_G = (P_G + 7)/(20 - P_G)$. Average nitrogen and carbon use per base pair were also calculated (N_G/l_G and C_G/l_G , respectively), and are given by $N_G/l_G = P_G + 7$ and $C_G/l_G = 20 - P_G$. Further, $N_G/l_G + C_G/l_G = 27$.

For the same prokaryotes, primary amino acid sequences were obtained for all open reading frames from The Institute of Genomic

Research (TIGR) Web site (http://www.tigr.org/tigr-scripts/CMR2/batch_download.dbi). For each proteome, we counted the number of times each amino acid was used, summing amino acids in different chromosomes, and excluding plasmids. The total number of nitrogen atoms and carbon atoms in each proteome was then calculated (N_p and C_p , respectively) by summing atoms in all amino acids in the proteome. N/C values for proteomes were calculated as $\nu_p = N_p/C_p$. Average use of nitrogen and carbon per amino acid (N_p/l_p and C_p/l_p , respectively) were calculated, where l_p is the total number of amino acids in the proteome. These calculations do not take into account differential levels of expression of proteins, or the fact that not all open reading frames that are identified will be transcribed and translated. Inferring nutrient use by organisms in proteins from these data carries the assumption that amino acids are expressed in similar proportions to their representation in the proteome.

We tested for relationships between variables: ν_G and ν_p ; N_G/l_G and N_p/l_p ; C_G/l_G and C_p/l_p ; N_p/l_p and C_p/l_p , using Spearman's rank correlation (SPSS v. 11.0). These analyses were performed among the 94 species common to the NCBI and TIGR databases, as at 1 October 2003. Where multiple strains were available for a single species, one strain only was included in analyses, by selecting strains for exclusion at random. Prokaryotes were classified taxonomically as Archaea or Eubacteria, and into habitats as free living, or capable of living as animal and/or plant symbionts (where symbionts may be pathogens, commensals or mutualists). For most species, this information was taken from Dworkin (2003). Primary literature sources were used for species not found in Dworkin (2003) (see electronic Appendix A).

3. RESULTS AND DISCUSSION

N/C values of codons (ν_C) and amino acids (ν_{AA}) are correlated positively, where ν_C is calculated for single-stranded RNA codons ($n = 61$, $r_s = 0.53$, $p < 0.001$; figure 1a). There is a strong, positive correlation between the N/C values of codons and the average hydrophobicity rankings of their anticodon dinucleoside monophosphates ($n = 16$, $r_s = 0.97$, $p < 0.001$), and between N/C values of amino acids and their average hydrophobicity rankings ($n = 20$, $r_s = 0.76$, $p < 0.001$). Taken together, these three relationships are consistent with the positive correlation between average hydrophobicity rankings of amino acids and their anticodon dinucleoside monophosphates (Lacey & Mullins 1983). This suggests the correlation between the N/C content of codons and their amino acids may be an outcome of chemical interactions between amino acids and anticodons that influenced the evolution of the genetic code (Lacey & Mullins 1983). The correlation between N/C values of amino acids and their codons also holds significantly, but is less strong, for the double-stranded DNA version of the codons ($n = 61$, $r_s = 0.38$, $p = 0.002$; figure 1b).

There is a strong, positive relationship between N/C values of whole genomes (ν_G) and proteomes (ν_p) across species ($n = 94$, $r_s = 0.90$, $p < 0.001$; figure 2a). This is probably a result of the relationship between N/C values of amino acids and codons, and the influence of GC content on amino acid frequencies (Sueoka 1961; Singer & Hickey 2000; Knight *et al.* 2001).

Further, for both nitrogen and carbon atoms, average atomic count per base pair and average atomic count per amino acid are positively related ($n = 94$, $r_s = 0.63$, $p < 0.001$; figure 2b; and $n = 94$, $r_s = 0.88$, $p < 0.001$; figure 2c, respectively). However, average nitrogen use in genomes and proteomes are both correlated positively with GC content, whereas average carbon use in genomes and proteomes are correlated negatively with GC content (figure 2b,c). This suggests that there is a negative relationship between nitrogen and carbon use among genomes and among proteomes. For the genome, this

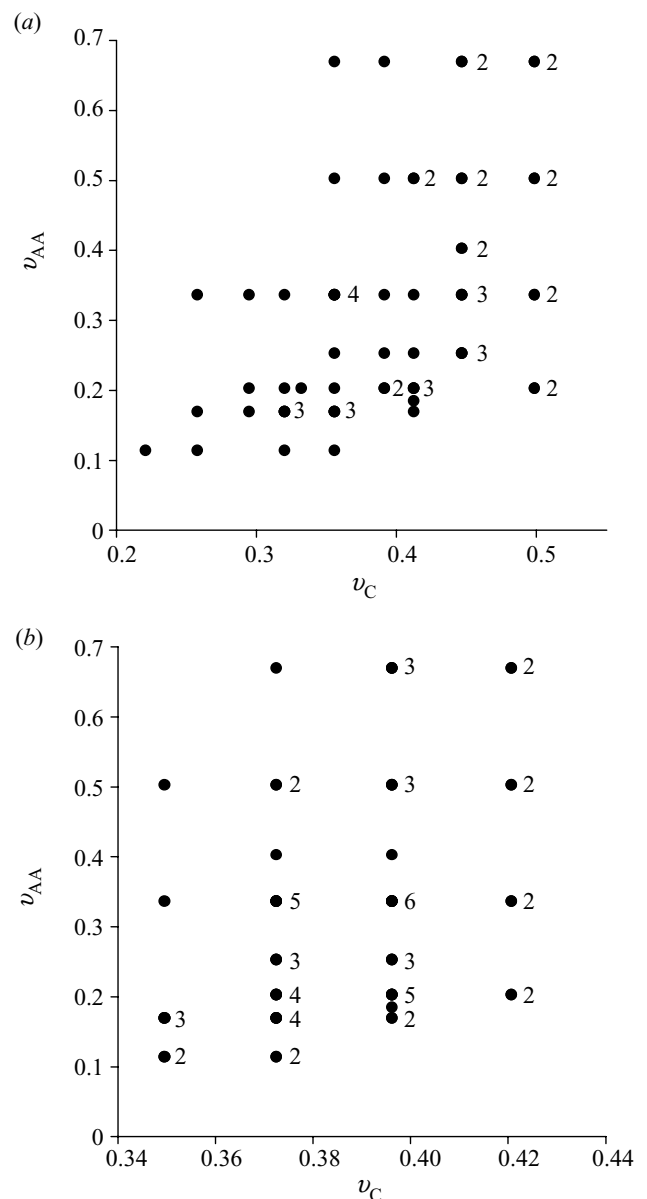


Figure 1. Relationships between N/C content of amino acids (ν_{AA}) and their codons (ν_C), calculated for (a) single-stranded RNA codons, and (b) double-stranded DNA codons. In both cases, N/C content of amino acids and codons are correlated positively. Numerical labels denote the number of data points that lie exactly in the same position.

relationship is given by $N_G/l_G = -C_G/l_G + 27$ (figure 2d). Note that no variation around this relationship is possible. Among proteomes, there is also a negative relationship about average use of nitrogen and carbon, but deviation about this relationship is possible ($n = 94$, $r_s = -0.50$, $p < 0.001$; figure 2e).

These results describe a spectrum among prokaryote genomes and proteomes in nitrogen versus carbon use. Across this spectrum, one might predict that animal symbionts would have greater N/C in both genomes and proteomes than plant symbionts, as the tissues of animals often have greater N/C content than plant tissues (Sterner & Elser 2002). However, there is no evidence to support this prediction and, if anything, the mean N/C values in genomes and proteomes are higher for plant symbionts than animal symbionts (mean \pm s.e.; animal symbiont $\nu_G =$

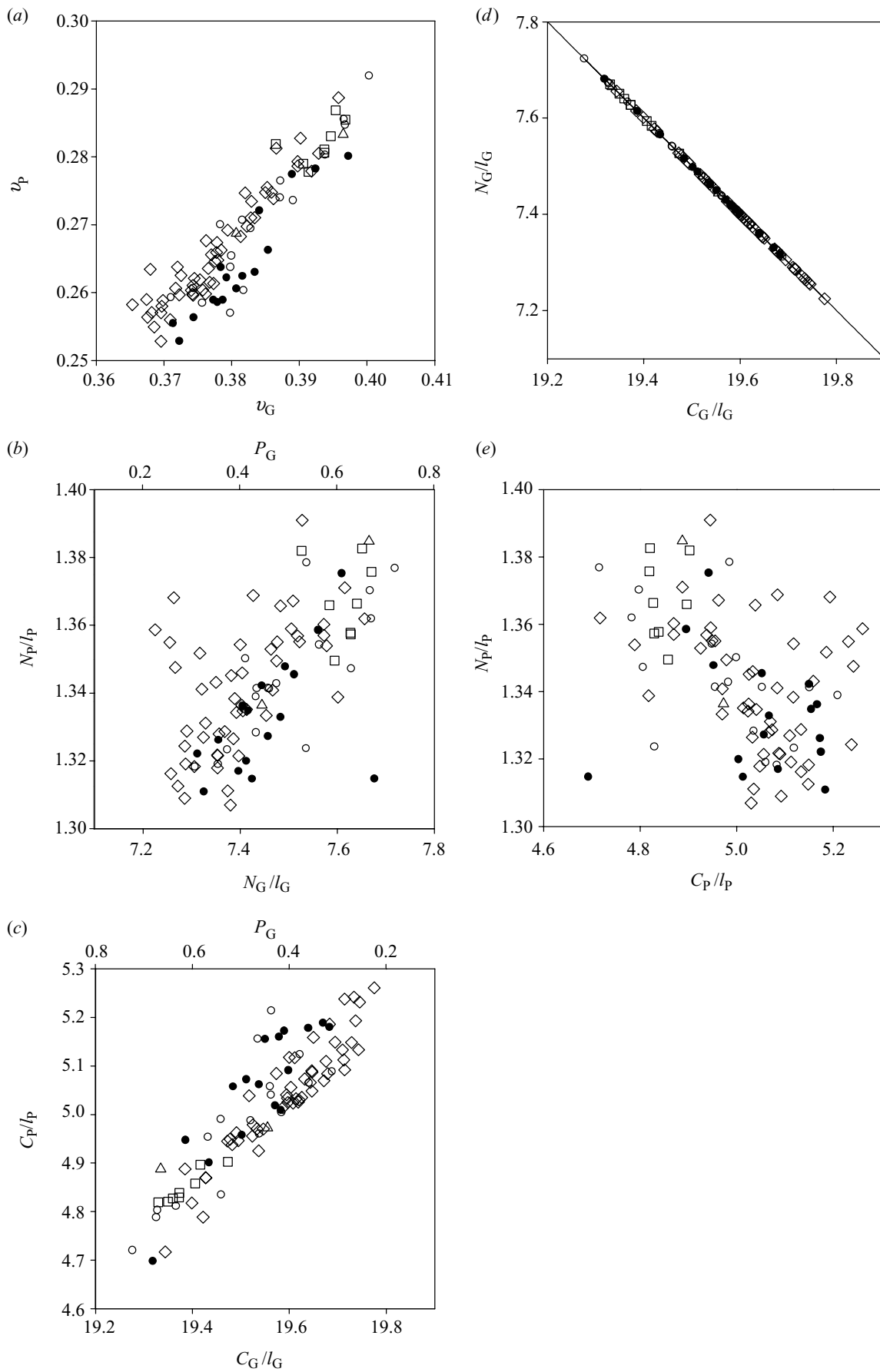


Figure 2. (Caption opposite.)

Figure 2. Relationships of the elemental composition of genomes and proteomes among 94 prokaryotic species. (a) N/C content of whole genomes (ν_G) and proteomes (ν_P) are correlated positively. For (b) nitrogen atoms and (c) carbon atoms, average atomic counts per base pair in genomes (N_G/l_G and C_G/l_G , respectively) and average atomic counts per amino acid in proteomes (N_P/l_P and C_P/l_P , respectively) are correlated positively. However, (b) average nitrogen content of genomes and proteomes are associated positively with GC content (P_G), whereas (c) average carbon content of genomes and proteomes are associated negatively with GC content (P_G). Average nitrogen atom counts and average carbon atom counts are associated negatively among (d) genomes (per base pair, N_G/l_G and C_G/l_G , respectively) and (e) proteomes (per amino acid, N_P/l_P and C_P/l_P , respectively). For the genome, this relationship is given by $N_G/l_G = -C_G/l_G + 27$. Filled circles denote Archaea. Open symbols denote Eubacteria: free-living only, circle; capable of living as animal symbionts, diamond; capable of living as plant symbionts, square; capable of living as animal and plant symbionts, triangle.

0.378 ± 0.001 and $\nu_P = 0.267 \pm 0.001$, $n = 53$; plant symbiont $\nu_G = 0.392 \pm 0.002$ and $\nu_P = 0.281 \pm 0.002$, $n = 10$; see also figure 2a). It is therefore unclear whether variation in GC content among prokaryotes is related adaptively to constraints on the availability of nitrogen and carbon in the environment. Previously, variation in GC content among prokaryotes has been related to pressures including mutation bias (Sueoka 1962, 1988; Muto & Osawa 1987), exposure to ultraviolet radiation (Singer & Ames 1970), aerobiosis (Naya *et al.* 2002) and atmospheric nitrogen fixation (McEwan *et al.* 1998). Atmospheric nitrogen fixation is related to elevated GC content, which was attributed to the greater expenditure of nitrogen in GC bonds (McEwan *et al.* 1998). We provide support for this hypothesis by showing that species with a high use of nitrogen in their genomes are also likely to have a high use of nitrogen in their proteomes (figure 2). This emphasizes the notion that variation in GC content among species influences nitrogen and carbon use in genomes and proteomes, and may have significant implications for the ecology and evolution of prokaryotes, whether or not it reflects an evolutionary response to constraints in nitrogen or carbon availability.

Note added in proof. Following acceptance of this paper, it came to our attention that Baudouin-Cornu *et al.* (2004) independently studied relationships between genomic GC content and the atomic content of whole proteomes. We find it encouraging that a strong association between GC

content and average carbon use in proteomes was found in both investigations, despite differences in the methods used to calculate average atomic use in proteomes.

Acknowledgements

The authors thank L. Barton, M. Gilchrist, D. Haig, H. Morowitz, D. Natvig and R. Plunkett for discussion of this work. A. Allen, J. Brown, B. Fox, J. Gillooly, L. Schwanz, E. White and two anonymous reviewers provided valuable comments on earlier drafts of this manuscript. J.G.B. was supported by an NSF Biocomplexity fellowship.

- Baudouin-Cornu, P., Surdin-Kerjan, Y., Marlière, P. & Thomas, D. 2001 Molecular evolution of protein atomic composition. *Science* **293**, 297–300.
- Boudouin-Cornu, P., Schuerer, K., Marlière, P. & Thomas, D. 2004 Intimate evolution of proteins: proteome atomic content correlates with genome base composition. *J. Biol. Chem.* **279**, 5421–5428.
- Dworkin, M. (ed.) 2003 *The prokaryotes: an evolving electronic resource for the microbiological community*, 3rd edn. New York: Springer.
- Jungck, J. R. 1978 The genetic code as a periodic table. *J. Mol. Evol.* **11**, 211–224.
- Knight, R. D., Freeland, S. L. & Landweber, L. F. 1999 Selection, history and chemistry: the three faces of the genetic code. *Trends Biochem. Sci.* **24**, 241–247.
- Knight, R. D., Freeland, S. L. & Landweber, L. F. 2001 A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biol.* **2**, research0010.1–0010.13.
- Lacey Jr, J. C. & Mullins Jr, D. W. 1983 Experimental studies related to the origin of the genetic code and the process of protein synthesis—a review. *Orig. Life* **13**, 3–42.
- McEwan, C., Gatherer, D. & McEwan, N. 1998 Nitrogen-fixing aerobic bacteria have higher genomic GC content than non-fixing species within the same genus. *Hereditas* **128**, 173–178.
- Mazel, D. & Marlière, P. 1989 Adaptive eradication of methionine and cysteine from cyanobacterial light-harvesting proteins. *Nature* **341**, 245–248.
- Muto, A. & Osawa, S. 1987 The guanine and cytosine content of genomic DNA and bacterial evolution. *Proc. Natl Acad. Sci. USA* **84**, 166–169.
- Naya, H., Romero, H., Zavala, A., Alvarez, B. & Musto, H. 2002 Aerobiosis increases the genomic guanine plus cytosine content (GC%) in prokaryotes. *J. Mol. Evol.* **55**, 260–264.
- Singer, C. E. & Ames, B. N. 1970 Sunlight ultraviolet and bacterial DNA base ratios. *Science* **170**, 822–826.
- Singer, G. A. C. & Hickey, D. A. 2000 Nucleotide bias causes a genome-wide bias in the amino acid composition of proteins. *Mol. Biol. Evol.* **17**, 1581–1588.
- Sternner, R. W. & Elser, J. J. 2002 *Ecological stoichiometry*. Princeton University Press.
- Sueoka, N. 1961 Compositional correlation between deoxyribonucleic acid and protein. *Cold Spring Harbor Symp. Quant. Biol.* **26**, 35–43.
- Sueoka, N. 1962 On the genetic basis of variation and heterogeneity of DNA base composition. *Proc. Natl Acad. Sci. USA* **48**, 582–592.
- Sueoka, N. 1988 Directional mutation pressure and neutral molecular evolution. *Proc. Natl Acad. Sci. USA* **85**, 2653–2657.

Visit www.journals.royalsoc.ac.uk and navigate to this article through *Biology Letters* to see the accompanying electronic appendix.