

Gene amelioration demonstrated: the journey of nascent genes in bacteria

Pradeep Reddy Marri and G. Brian Golding

Abstract: Gene amelioration is the hypothesis that genes acquired via lateral gene transfer will, over time, acquire the molecular characteristics of the host genome. Species for which multiple strains have been sequenced permit a demonstration that this hypothesis is correct. We use 7 sequenced genomes of *Streptococcus pyogenes* and 6 sequenced genomes of *Staphylococcus aureus* to illustrate the action of amelioration on these genomes.

Key words: *Streptococcus pyogenes*, *Staphylococcus aureus*, lateral gene transfer, unique genes, phylogeny, molecular evolution.

Résumé : L'amélioration génique est une hypothèse selon laquelle des gènes acquis via transfert latéral en viendront, au fil du temps, à acquérir les caractéristiques moléculaires de leur génome-hôte. Des espèces chez lesquelles plusieurs souches ont été séquencées permettent d'en faire la démonstration. Les auteurs ont employé sept génomes séquencés du *Streptococcus pyogenes* et six génomes séquencés du *Staphylococcus aureus* pour illustrer l'impact de l'amélioration sur ces génomes.

Mots-clés : *Streptococcus pyogenes*, *Staphylococcus aureus*, transfert latéral de gènes, gènes uniques, phylogénie, évolution moléculaire.

[Traduit par la Rédaction]

Bacterial genomes are under constant selective pressures to adapt to their surroundings. Bacterial genomes can evolve by modifying their gene repertoire, either by mutating their existing genes (Sokurenko et al. 1998; Feldgarden et al. 2003), by gene loss (Cole et al. 2001; Ogata et al. 2001; Foster et al. 2005), and (or) by acquisition of new genes by lateral gene transfer. Lateral gene transfer is now understood to be one of the predominant forces of bacterial evolution (Dutta and Pan 2002; Lawrence 1999; Lawrence and Ochman 2002). Within a genome, the GC content and codon bias of the genes are determined by the selectional and mutational pressures acting on the genome (Sueoka 1988). These pressures should result in a relatively homogeneous base composition of the entire bacterial chromosome. Genes acquired through lateral gene transfer may originate from genomes with a different GC composition and codon bias (Grantham et al. 1980; Karlin and Burge 1995) and can be recognized by their abnormal properties. Lawrence and Ochman (1997) hypothesized that following their acquisition, newly acquired genes will be subject to the same genome-wide mutational pressures as native genes and that, as their residence time increases, the acquired genes will be amelio-

rated and grow to resemble the native genes. There have been few documented cases of this process in action and amelioration is not a foregone conclusion in bacterial genomes, since many factors can affect GC content and codon bias (for example, translational robustness; Drummond et al. 2005). The availability of multiple sequenced genomes from a single species permits a demonstration of this important process in bacterial genome evolution. In this note we demonstrate the process of amelioration using the sequenced genomes of *Streptococcus pyogenes* and *Staphylococcus aureus*.

We have analyzed indices based on codon usage and GC content in 7 sequenced strains of *S. pyogenes* and 6 sequenced strains of *S. aureus* and demonstrate that these indices are useful to show the process of amelioration in action in closely related species. To provide a timeline, the genes in each genome were divided into 3 categories: strain-specific genes, species-specific genes, and core genes. The strain-specific genes are thought to represent the nascent genes that were acquired most recently, the species-specific genes represent a relatively older set of genes that have been resident in the genome longer than the strain-specific

Received 29 March 2007. Accepted 17 October 2007. Published on the NRC Research Press Web site at genome.nrc.ca on 29 January 2008.

Corresponding Editor: P.B. Moens.

P.R. Marri and G.B. Golding.¹ Department of Biology, McMaster University, Hamilton, ON L8S 4K1, Canada.

¹Corresponding author (e-mail: Golding@McMaster.CA).

genes, and the core genes represent the oldest genes in a genome. Although there may be some multiply deleted genes or other reasons for some genes to be placed in the wrong category, on the whole these misplacements should be comparatively infrequent and do not have a significant bearing on the results presented here.

For each category of genes we measured 5 variables: codon adaptation index (CAI; Sharp and Li 1987), number of effective codons (Nc; Wright 1990), GC content at the third codon position (GC3), GC content at the first and second codon positions (GC12), and length of the gene. In addition, we examined the correlation of GC12 with GC3 (relative neutrality plot; Sueoka 1995) and Nc with GC12 (relative codon efficiency).

Genome sequences were downloaded from the NCBI Entrez Genome Project database at <http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi>. All the protein sequences in each of 7 *S. pyogenes* strains were compared with all 14 sequenced *Streptococcus* genomes using BLASTP (Altschul et al. 1997) with an *E* value cutoff set at 1.0×10^{-10} and the minimum match length set at 85% of the query sequence. All the single-copy genes present across the 14 genomes were considered core genes. Since highly expressed genes might be considered to have abnormal codon and GC characteristics and might obscure the effects of amelioration, the analysis was repeated with the genes coding for ribosomal proteins and chaperones and those involved in transcription and translation (Karin and Mrazek 2000) removed from the core genes (see Fig. S0²). Genes present in all the *S. pyogenes* genomes but not in any other congeneric genome were considered species-specific genes and the genes that were uniquely present in a strain of *S. pyogenes* were considered strain-specific genes. A similar procedure was followed for identifying the core, species-specific, and strain-specific genes of *S. aureus*. A list of all the genomes used is given in Table 1.

The GC content, GC3, CAI, and Nc for each gene were computed using CodonW (<http://bioweb.pasteur.fr/seqanal/interfaces/codonw.html>). To calculate the CAI, the codon usage of the corresponding ribosomal proteins was used as a reference set. GC12 for each gene was obtained using a perl script. Statistical analyses were performed using the R statistical package (<http://www.r-project.org/>) and figures were generated using Gnuplot (<http://www.gnuplot.info/>).

The CAI is correlated with the relative expression level of a gene (Sharp and Li 1987). Analysis of the 3 categories of genes clearly indicated that the average CAI for core genes is higher than that for species-specific genes, which in turn is higher than that for strain-specific genes (Fig. 1).

A similar but opposite trend is seen for Nc (Fig. 1). The strain-specific genes, representing newly acquired genes, use a significantly higher number of codons compared with the core genes (Mann–Whitney test; $p < 0.0001$). This is expected because these genes represent those that have been most recently acquired by lateral gene transfer from divergent sources having their own preferences for codon usage. On their initial entry into the new genome these genes tend

Table 1. List of genomes used.

Organism	GenBank accession No.
<i>Streptococcus</i>	
<i>Streptococcus pyogenes</i> M1 GAS	AE004092
<i>Streptococcus pyogenes</i> MGAS10394	CP000003
<i>Streptococcus pyogenes</i> MGAS315	AE014074
<i>Streptococcus pyogenes</i> MGAS5005	CP000017
<i>Streptococcus pyogenes</i> MGAS6180	CP000056
<i>Streptococcus pyogenes</i> MGAS8232	AE009949
<i>Streptococcus pyogenes</i> SSI-1	BA000034
<i>Streptococcus agalactiae</i> 2603V/R	AE009948
<i>Streptococcus agalactiae</i> NEM6	AL732656
<i>Streptococcus mutans</i> UA159	AE009948
<i>Streptococcus pneumoniae</i> R6	AE007317
<i>Streptococcus pneumoniae</i> TIGR4	AE005672
<i>Streptococcus thermophilus</i> CNRZ1066	CP000024
<i>Streptococcus thermophilus</i> LMG18311	CP000023
<i>Staphylococcus</i>	
<i>Staphylococcus aureus</i> COL	CP000046
<i>Staphylococcus aureus</i> MRSA252	BX571856
<i>Staphylococcus aureus</i> MSSA476	BX571857
<i>Staphylococcus aureus</i> MW2	BA000033
<i>Staphylococcus aureus</i> Mu50	BA000017
<i>Staphylococcus aureus</i> N315	BA000018
<i>Staphylococcus epidermidis</i> RP62A	CP000029
<i>Staphylococcus saprophyticus</i> ATCC 15305	AP008934
<i>Staphylococcus haemolyticus</i> JCSC1435	AP006716

to use the same set of codons as their donor, but as their residence time in the recipient increases, they ameliorate to resemble the native genes, since they are under the same selective pressures. The core genes have had the longest residence time and the greatest influence from selection and mutation biases. As indicated by the results shown in Fig. 1, and as expected in the amelioration hypothesis, the species-specific genes form an intermediate group between the strain-specific genes and the core genes.

The GC content analysis also showed differences between the 3 categories of genes consistent with the amelioration hypothesis. The average GC content and the GC content at the non-synonymous codon positions (GC12) were significantly higher in the core genes compared with the strain-specific genes (Mann–Whitney test; $p < 0.0001$), with the species-specific genes having an intermediate average. However, it was surprising that the GC content of the core genes at the third codon position was significantly lower than that of the strain-specific genes (Mann–Whitney test; $p < 0.0001$). This is in contrast to earlier reports that have shown a higher GC content at synonymous codon positions for core genes compared with laterally acquired genes (Daubin et al. 2003; dos Reis et al. 2003). This could be a result of the AT richness of the laterally acquired genes. The strain-specific genes constitute mostly ORFan genes (genes that have no known homologues in sequenced genomes, as defined by Fischer and Eisenberg 1999), which

²Supplementary data for this article are available on the journal Web site (<http://genome.nrc.ca>) or may be purchased from the Depository of Unpublished Data, Document Delivery, CISTI, National Research Council Canada, Building M-55, 1200 Montreal Road, Ottawa, ON K1A 0R6, Canada. DUD 3706. For more information on obtaining material refer to http://cisti-icist.nrc-cnrc.gc.ca/irm/unpub_e.shtml.

Fig. 1. Average GC content (GC), GC content at the first and second codon positions (GC12), GC content at the third codon position (GC3), effective number of codons (Nc), and codon adaptation index (CAI) for strain-specific (black), species-specific (gray), and core (white) genes for *Streptococcus pyogenes* MGAS315 (A) and *Staphylococcus aureus* MU50 (B).

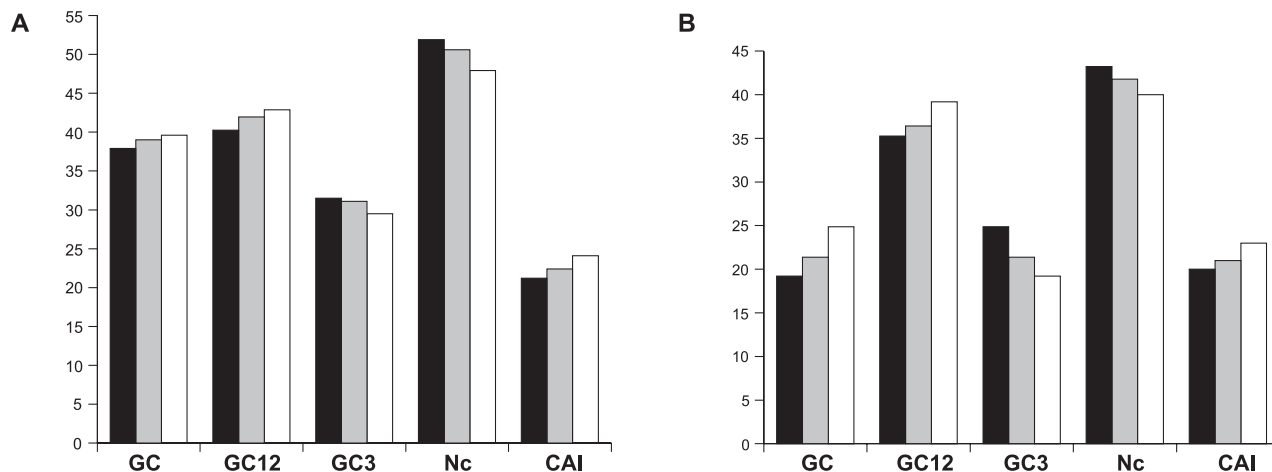


Table 2. Relative frequency of codons used in the core and strain-specific genes of *Streptococcus pyogenes* MGAS315 and *Staphylococcus aureus* MU50 for codons with C/G, C/G, A/T nucleotides.

Codon	<i>Streptococcus pyogenes</i>		<i>Staphylococcus aureus</i>	
	Core	Strain-specific	Core	Strain-specific
CCT	1.33	0.96	1.12	1.13
CCA	1.60	1.19	1.69	1.00
GCT	3.97	2.41	2.10	1.71
GCA	2.23	1.75	3.09	1.72
CGT	2.28	1.37	1.54	0.91
CGA	0.44	0.64	0.51	0.52
GGT	2.93	2.14	2.58	2.01
GGA	1.78	1.68	1.41	1.49

have been shown to have a lower GC composition compared with native genes (Daubin et al. 2003). On their entry into the new genome, the new genes will be driven by the AT→GC mutational bias to homogenize with the recipient genome. As the synonymous sites are amenable to substitutions (Lawrence and Ochman 1997), this might result in the mutation of A's and T's in the silent sites to G's and C's, resulting in a relatively higher GC3 compared with the core genes. Moreover, a comparison of the frequency of the codons used by the core and strain-specific genes revealed that core genes use a relatively higher frequency of codons of type C/G C/G A/T (at the first, second, and third codon positions) compared with the strain-specific genes (Table 2), probably resulting in their relatively higher GC12 and lower GC3.

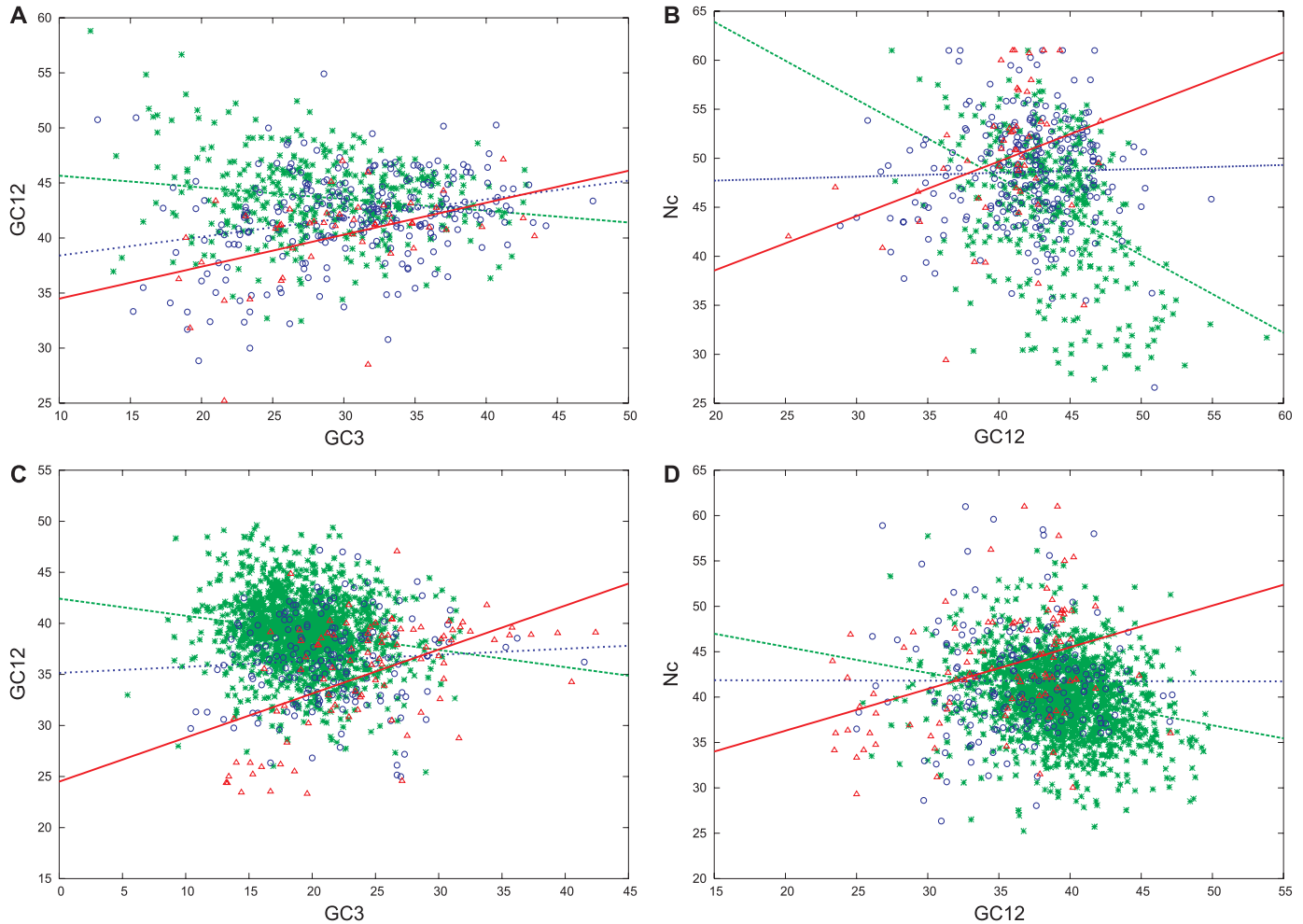
The relative neutrality plot (plot of GC3 vs. GC12) gives a measure of selective constraints acting on the first/second or third codon position. A slope of unity of a linear regression would indicate no selective constraint and a decreasing slope would indicate an increasing selective constraint. The relative neutrality plots for the core, species-specific, and strain-specific genes of *S. pyogenes* MGAS315 had slopes of -0.11, 0.17, and 0.22, respectively (Fig. 2A). The core genes had a significantly lower slope (ANCOVA, $p <$

0.0001) compared with the species-specific and strain-specific genes, indicating the expected strong selective constraints on the core genes. The strain-specific genes had the highest slope, indicating that they are under relaxed selective constraints. If the highly expressed genes coding for ribosomal proteins and chaperones and those involved in transcription and translation are removed from the list of core genes, a very similar pattern is observed (see Fig. S0²), with slopes of 0.01, 0.13, and 0.37, respectively. The same trend was observed for *S. aureus* MU50, where the core genes had the lowest slope (-0.15), the strain-specific genes had a significantly higher slope (0.32; ANCOVA, $p <$ 0.0001), and the species-specific genes had an intermediate slope (0.06) (Fig. 2C), again confirming that the core genes are under higher selective constraints compared with the species-specific genes and that the strain-specific genes are under relaxed selective constraints. The trend described here was similar in all the strains of *S. pyogenes* and *S. aureus* analyzed (see Figs. S1–S11²).

A plot of Nc versus GC12 (Figs. 2B and 2D) indicates that the newly acquired genes are undergoing synonymous as well as non-synonymous substitutions during the process of amelioration. The newly acquired genes, which represent a diverse pool of genes, are able to use a higher number of codons effectively. These genes tend to have a relatively lower GC12 and, as a result, an initial increase in GC12 results in the use of a larger number of codons (positive correlation for strain-specific genes), but a further increase in GC12 due to amelioration would restrict the number of codons used (and no correlation for species-specific genes is found). As the genes are ameliorated to an even higher GC12 (core genes), they tend to use a minimum set of codons and any further rise in GC12 would decrease the number of accessible codons (causing a negative slope for core genes). A similar trend was observed for *S. aureus* MU50 and all the strains of *S. pyogenes* and *S. aureus* analyzed (see Figs. S1–S11²).

By differentiating genes into their relative residency times within a genome, these data demonstrate that strain-specific genes (those most recently acquired), once in a recipient ge-

Fig. 2. Average GC12 versus average GC3 (relative neutrality plot; right) and average Nc versus average GC12 (relative codon efficiency; left) for *Streptococcus pyogenes* MGAS315 (A and B) and *Staphylococcus aureus* MU50 (C and D). The core genes are shown as green stars, the species-specific genes are shown as blue circles, and the strain-specific genes are shown as red triangles. The lines show the linear regressions.



nome, are subject to the same selective pressures as the native genes and over a period of time ameliorate to resemble the native genes. This provides a demonstration of the amelioration process hypothesized by Lawrence and Ochman (1997). Additionally, the data confirm that the strain-specific genes are under relaxed selective constraints compared with the species-specific and core genes, and hence are evolving faster, while the core genes are under stronger selective constraints and are evolving under a comparatively greater influence of purifying selection.

References

- Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**: 3389–3402. doi:10.1093/nar/25.17.3389. PMID:9254694.
- Cole, S.T., Eiglmeier, K., Parkhill, J., James, K.D., Thomson, N.R., Wheeler, P.R., et al. 2001. Massive gene decay in leprosy bacillus. *Nature (London)*, **409**: 1007–1011. doi:10.1038/35059006. PMID:11234002.
- Daubin, V., Lerat, E., and Perriere, G. 2003. The source of laterally transferred genes in bacterial genomes. *Genome Biol.* **4**: R57. doi:10.1186/gb-2003-4-9-r57. PMID:12952536.
- dos Reis, M., Wernisch, L., and Savva, R. 2003. Unexpected correlations between gene expression and codon usage bias from microarray data for the whole *Escherichia coli* K-12 genome. *Nucleic Acids Res.* **31**: 6976–6985. doi:10.1093/nar/gkg897. PMID:14627830.
- Drummond, D.A., Bloom, J.D., Adami, C., Wilke, C.O., and Arnold, F.H. 2005. Why highly expressed proteins evolve slowly. *Proc. Natl. Acad. Sci. U.S.A.* **102**: 14338–14343. doi:10.1073/pnas.0504070102. PMID:16176987.
- Dutta, C., and Pan, A. 2002. Horizontal gene transfer and bacterial diversity. *J. Biosci.* **27**: 27–33. doi:10.1007/BF02703681. PMID:11927775.
- Feldgarden, M., Byrd, N., and Cohan, F.M. 2003. Gradual evolution in bacteria: evidence from *Bacillus* systematics. *Microbiology*, **149**: 3565–3573. doi:10.1099/mic.0.26457-0. PMID:14663088.
- Fischer, D., and Eisenberg, D. 1999. Finding families for genomic ORFans. *Bioinformatics*, **15**: 759–762. doi:10.1093/bioinformatics/15.9.759. PMID:10498776.
- Foster, J., Ganatra, M., Kamal, I., Ware, J., Makarova, K., Ivanova,

- N., et al. 2005. The *Wolbachia* genome of *Brugia malayi*: endosymbiont evolution within a human pathogenic nematode. *PLoS Biol.* **3**: e121. doi:10.1371/journal.pbio.0030121. PMID: 15780005.
- Grantham, R., Gautier, C., Gouy, M., Mercier, R., and Pavé, A. 1980. Codon catalog usage and the genome hypothesis. *Nucleic Acids Res.* **8**: r49–r62. PMID:6986610.
- Karlin, S., and Burge, C. 1995. Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet.* **11**: 283–290. doi:10.1016/S0168-9525(00)89076-9. PMID:7482779.
- Karlin, S., and Mrazek, J. 2000. Predicted highly expressed genes of diverse prokaryotic genomes. *J. Bacteriol.* **182**: 5238–5250. doi:10.1128/JB.182.18.5238-5250.2000. PMID:10960111.
- Lawrence, J.G. 1999. Gene transfer, speciation, and the evolution of bacterial genomes. *Curr. Opin. Microbiol.* **2**: 519–523. doi:10.1016/S1369-5274(99)00010-7. PMID:10508729.
- Lawrence, J.G., and Ochman, H. 1997. Amelioration of bacterial genomes: rates of change and exchange. *J. Mol. Evol.* **44**: 383–397. doi:10.1007/PL00006158. PMID:9089078.
- Lawrence, J.G., and Ochman, H. 2002. Reconciling the many faces of lateral gene transfer. *Trends Microbiol.* **10**: 1–4. doi:10.1016/S0966-842X(01)02282-X. PMID:11755071.
- Ogata, H., Audic, S., Renesto-Audiffren, P., Fournier, P.E., Barbe, V., Samson, D., et al. 2001. Mechanisms of evolution in *Rickettsia conorii* and *R. prowazekii*. *Science* (Washington, D.C.), **293**: 2093–2098. doi:10.1126/science.1061471. PMID:11557893.
- Sharp, P.M., and Li, W.H. 1987. The codon adaptation index — a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15**: 1281–1295. doi:10.1093/nar/15.3.1281. PMID:3547335.
- Sokurenko, E.V., Chesnokova, V., Dykhuizen, D.E., Ofek, I., Wu, X.R., Krogfelt, K.A., et al. 1998. Pathogenic adaptation of *Escherichia coli* by natural variation of the FimH adhesin. *Proc. Natl. Acad. Sci. U.S.A.* **95**: 8922–8926. doi:10.1073/pnas.95.15.8922. PMID:9671780.
- Sueoka, N. 1988. Directional mutation pressure and neutral molecular evolution. *Proc. Natl. Acad. Sci. U.S.A.* **85**: 2653–2657. doi:10.1073/pnas.85.8.2653. PMID:3357886.
- Sueoka, N. 1995. Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. *J. Mol. Evol.* **40**: 318–325. doi:10.1007/BF00163236. PMID:7723058.
- Wright, F. 1990. The ‘effective number of codons’ used in a gene. *Gene*, **87**: 23–29. doi:10.1016/0378-1119(90)90491-9. PMID: 2110097.