Coupling Between Protein Level Selection and Codon Usage Optimization in the Evolution of Bacteria and Archaea

Wenqi Ran, David M. Kristensen, Eugene V. Koonin

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA

ABSTRACT **The relationship between the selection affecting codon usage and selection on protein sequences of orthologous genes in diverse groups of bacteria and archaea was examined by using the Alignable Tight Genome Clusters database of prokaryote genomes. The codon usage bias is generally low, with 57.5% of the gene-specific optimal codon frequencies (F_{opt}) being below 0.55. This apparent weak selection on codon usage contrasts with the strong purifying selection on amino acid sequences, with 65.8% of the gene-specific d*N*/d*S* ratios being below 0.1. For most of the genomes compared, a limited but statistically significant negative correlation between F_{opt} and d*N*/d*S* was observed, which is indicative of a link between selection on protein sequence and selection on codon usage. The strength of the coupling between the protein level selection and codon usage bias showed a strong positive correlation with the genomic GC content. Combined with previous observations on the selection for GC-rich codons in bacteria and archaea with GC-rich genomes, these findings suggest that selection for translational fine-tuning could be an important factor in microbial evolution that drives the evolution of genome GC content away from mutational equilibrium. This type of selection is particularly pronounced in slowly evolving, "high-status" genes. A significantly stronger link between the two aspects of selection is observed in free-living bacteria than in parasitic bacteria and in genes encoding metabolic enzymes and transporters than in informational genes. These differences might reflect the special importance of translational fine-tuning for the adaptability of gene expression to environmental changes. The results of this work establish the coupling between protein level selection and selection for translational optimization as a distinct and potentially important factor in microbial evolution.**

IMPORTANCE **Selection affects the evolution of microbial genomes at many levels, including both the structure of proteins and the regulation of their production. Here we demonstrate the coupling between the selection on protein sequences and the optimization of codon usage in a broad range of bacteria and archaea. The strength of this coupling varies over a wide range and strongly and positively correlates with the genomic GC content. The cause(s) of the evolution of high GC content is a long-standing open question, given the universal mutational bias toward AT. We propose that optimization of codon usage could be one of the key factors that determine the evolution of GC-rich genomes. This work establishes the coupling between selection at the level of protein sequence and at the level of codon choice optimization as a distinct aspect of genome evolution.**

Received 14 February 2014 **Accepted** 21 February 2014 **Published** 25 March 2014

Citation Ran W, Kristensen DM, Koonin EV. 2014. Coupling between protein level selection and codon usage optimization in the evolution of bacteria and archaea. mBio 5(2): e00956-14. doi:10.1128/mBio.00956-14.

Editor Richard Lenski, Michigan State University

Address correspondence to Eugene V. Koonin, koonin@ncbi.nlm.nih.gov.

The amino acid sequences of the great majority of proteins evolve under the pressure of purifying selection that can be measured through the ratio of the rates of nonsynonymous and synonymous substitutions (*dN/dS*) in protein-coding sequences (1–4). The strength of purifying selection shows broad variation between sites within a protein-coding gene, between genes within an evolving genome, and between evolving genomes in different organismal lineages (5–8). Generally, purifying selection is strong in organisms with large effective population sizes, such as bacteria, but substantially weaker in organisms with small effective population sizes, such as multicellular eukaryotes (9, 10). Within a bacterial or archaeal genome, which typically encompasses between 1,000 and 10,000 protein-coding genes, the d*N*/d*S* ratio varies within approximately 2 orders of magnitude, from ~0.01 to ~1.00, with the mean and median of the distribution being close to 0.1 (11–13). Furthermore, comparative analysis of the d*N*/d*S* ratios across a broad range of bacterial and archaeal genomes that were collected in the database of Alignable Tight Genome Clusters (ATGC) (14) has shown that the median d*N*/d*S* ratio is stable within each ATGC but differs between ATGC, with the implication that this ratio is a robust, lineage-specific gauge of purifying selection (12).

The use of d*N*/d*S* as a measure of selection on protein sequences is based on the assumption that synonymous substitutions are neutral. This assumption can be a reasonable approximation inasmuch as selection affecting nonsynonymous sites is substantially stronger than that affecting synonymous sites. However, it is well established that synonymous sites in protein-coding sequences actually are subject to selection driven by at least two factors, RNA secondary structure and codon usage (15–18). The study of codon usage bias (CUB) is a long-standing direction in molecular evolution. Two fundamentally different but not mutu-

ally exclusive types of explanations for the existence of CUB have been explored, namely, mutational (neutral) and selectional origins. The important contribution of neutral mutational processes is suggested by the observations that GC content is the variable that best explains the interspecies differences in codon usage (19, 20). Moreover, it has been shown that CUB in bacteria could be predicted from the nucleotide composition of intergenic regions (20). However, there are also multiple strong indications of the important role of selection in the evolution of codon usage. The key early observations, made primarily on classical model organisms, the bacterium *Escherichia coli*, and the yeast *Saccharomyces cerevisiae*, are compatible with the selectionist but not the neutral hypothesis: CUB is particularly strong in highly expressed genes, and the usage of a particular codon strongly correlates with the abundance of the cognate tRNAs (21–25).

Subsequent research in this area to a large extent concentrated on the nature and strength of the selection that affects CUB (26, 27). It has been reported that in enterobacteria, CUB is strongly and negatively correlated with the synonymous evolution rate, i.e., genes with strong CUB typically evolve slowly; in contrast, little correlation was detected between CUB and the rate of protein evolution (28, 29). However, a more recent analysis of a range of model organisms, including *E. coli* and several eukaryotes, has revealed roughly the same strength of inverse correlation between CUB and d*N* compared to d*S* (30). Hartl and colleagues applied population genetic theory to estimate the selection coefficient on synonymous codon positions in enterobacteria and arrived at values on the order of 10^{-9}, indicative of weak selection that, however, could be consequential in large bacterial populations (31).

Two major factors underlying selection for CUB have been considered, namely, accuracy and rate of translation. The importance of translation accuracy was first suggested by experimental data indicating that codon choice strongly affected the rate of amino acid misincorporation during translation (32–34). Subsequently, it has been demonstrated that evolutionarily conserved amino acid sites showed a significantly stronger CUB than variable sites, as one would expect if selection acted to minimize the effect of amino acid misincorporation (35, 36). However, there are also substantial indications that selection for an increased rate or, more precisely, efficiency of translation contributes to the evolution of CUB. Indeed, optimal codons appear to be translated faster than suboptimal ones (37). Although this difference might not substantially affect the actual rate of protein production, which appears to be determined primarily by the rate of translation initiation (38), acceleration of elongation increases the supply of free ribosomes, a growth rate-limiting parameter in bacteria (26, 39). Indeed, a strong inverse correlation between codon bias and bacterial generation time has been detected, suggesting that the use of optimal codons is essential for fast growth (40–42). tRNA modifications also enhance translation speed and/or accuracy in different codon groups (43). A recent analysis of the codon usage of yeast took advantage of ribosome profiling data to show that optimal codons were actually not translated faster than suboptimal codons *in vivo* (44). Instead, it has been shown that, under conditions of tRNA shortage, the primary determinant of translation efficiency was the usage of codons proportional to the abundance of the cognate tRNAs (44). Analysis of codon usage in diverse bacteria by a recently developed statistical method yielded indications that selection for translation efficiency made a substantially greater contribution to the evolution of CUB than selection for translation accuracy (45).

The regulatory effects of CUB on cellular processes are likely to be multifaceted and remain only partially explored. For example, a recent study of the expression of bacterial operons that encode protein complexes with uneven subunit stoichiometry has shown that CUB is a key factor that provides for higher expression of the more abundant subunits (46).

Overall, the current view of CUB evolution centers around the selection-mutation-drift model, according to which there is (relatively) weak selection for preferred (major or optimal) codons but nonpreferred codons persist owing to mutational bias and genetic drift (26, 47–49). The strength of selection on CUB appears to vary broadly both across genes and across species, and translation accuracy and translation efficiency are both subject to selection, although the relative contributions of these two factors remain a matter of debate.

We were interested in exploring the connection between selection on CUB and selection on protein sequences. Generally, one would expect that the selective pressures at the two levels are coupled, given that high-expression genes, on the one hand, show a greater CUB than low-expression genes (21–25) and, on the other hand, on average evolve slowly (50–53).

However, previous studies have not resulted in certainty with regard to the existence and strength of this coupling, largely because CUB (measured as the fraction of optimal codons, F_{opt}) has been shown to depend similarly on d*S* and d*N*, with both dependencies thought to be uniformly gauged by the effective population size of an organism (30).

We performed a broad survey of the correlations between d*N*/d*S* and F_{opt} in bacteria and archaea, taking advantage of the database of ATGC, which encompasses groups of closely related genomes across the diversity of bacteria and archaea (14). We find that there is a nearly universal inverse correlation between these two variables; i.e., the two levels of selection are coupled. The strength of this coupling depends on the genomic GC content, suggesting that fine-tuning of translation efficiency and fidelity, especially in highly expressed genes, is an important factor in the evolution of the GC content of microbial genomes away from mutational equilibrium.

RESULTS

Universal coupling between selection on codon usage and selection on amino acid sequences and its dependence on genomic GC content. We first calculated the F_{opt} value and the d*N*/d*S* ratio of each pair of orthologous genes in a randomly selected pair of genomes from each ATGC (or the only pair for the ATGC consisting of two genomes; see Materials and Methods for details). The CUB was found to be relatively low, with F_{opt} being below 0.55 for 57.5% of the genes (Fig. 1). This relatively weak selection on codon usage contrasts with the typically strong purifying selection on amino acid sequences, with 65.8% of the d*N*/d*S* ratios being below 0.10 (Fig. 1).

For the substantial majority of the 120 ATGC analyzed, a statistically significant negative correlation between the gene-specific F_{opt} value and the d*N*/d*S* ratio was detected (Fig. 1). As shown previously, the genome-wide median d*N*/d*S* ratio is a stable characteristic of an ATGC (12). Therefore, we used the median d*N*/d*S* ratios and F_{opt} values of all of the genes in each ATGC (that is, of a random pair of genomes in the case of a large ATGC) as ATGC

	R	p	dN/dS	F	GS	GC%	deltaGC	Sample species
ATGC066	-0.606	0.000	0.058	0.587	4.600	60.3	4.566	Geobacter_bemidjiensis_Bem_uid58749
ATGC126	-0.567	0.000	0.094	0.592	2.260	60.1	4.372	Bifidobacterium_longum_NCC2705_uid57939
ATGC062	-0.517	0.000	0.057	0.573	7.747	60.9	3.774	Rhizobium_leguminosarum_bv__viciae_3841_uid57955
ATGC065	-0.510	0.000	0.106	0.547	3.771	63.3	4.077	Desulfovibrio_vulgaris_Hildenborough_uid57645
ATGC057	-0.501	0.000	0.038	0.632	8.495	64.8	5.059	Bradyrhizobium_ORS278_uid58941
ATGC061	-0.488	0.000	0.051	0.538	6.708	62.2	3.817	Sinorhizobium_meliloti_1021_uid57603
ATGC055	-0.472	0.000	0.037	0.627	5.471	68.2	5.194	Methylobacterium_extorquens_PA1_uid58821
ATGC042	-0.464	0.000	0.055	0.653	3.800	66.8	5.740	Acidovorax_ebreus_TPSY_uid59233
ATGC034	-0.457	0.000	0.060	0.599	5.240	63.6	4.457	Xanthomonas_oryzae_PXO99A_uid59131
ATGC091	-0.443	0.000	0.094	0.394	3.593	52	2.149	Geobacillus_kaustophilus_HTA426_uid58227
ATGC040	-0.441	0.000	0.052	0.607	9.735	62.6	4.344	Burkholderia_xenovorans_LB400_uid57823
ATGC025	-0.436	0.000	0.042	0.618	6.264	66.6	5.953	Pseudomonas_aeruginosa_PAO1_uid57945
ATGC124	-0.419	0.000	0.067	0.603	5.200	69.5	5.361	Salinispora_tropica_CNB_440_uid58565
ATGC052	-0.411	0.000	0.040	0.623	4.468	69	4.863	Rhodobacter_sphaeroides_ATCC_17029_uid58449
ATGC129	-0.409	0.000	0.033	0.686	3.404	72.4	6.685	Clavibacter_michiganensis_sepedonicus_uid61577
ATGC077	-0.408	0.000	0.046	0.554	2.434	59.4	3.176	Synechococcus_WH_8102_uid61581
ATGC122	-0.407	0.000	0.041	0.630	6.500	67.8	5.195	Mycobacterium_vanbaalenii_PYR_1_uid58463
ATGC121	-0.407	0.000	0.071	0.648	4.830	69.3	5.578	Mycobacterium_avium_paratuberculosis_K_10_uid57699
ATGC008	-0.400	0.000	0.085	0.416	3.828	53.6	1.849	Erwinia_amylovora_CFBP1430_uid46839
ATGC027	-0.400	0.000	0.039	0.491	6.112	57.9	3.309	Pseudomonas_syringae_phaseolicola_1448A_uid58099
ATGC116	-0.399	0.000	0.067	0.675	2.116	69.5	6.618	Thermus_thermophilus_HB8_uid58223
ATGC123	-0.398	0.000	0.076	0.621	5.921	68.4	5.096	Mycobacterium_MCS_uid58465
ATGC032	-0.395	0.000	0.035	0.659	4.574	66.3	5.798	Stenotrophomonas_maltophilia_R551_3_uid58657
ATGC039	-0.387	0.000	0.050	0.668	7.715	66.8	5.675	Burkholderia_cenocepacia_HI2424_uid58369
ATGC112	-0.368	0.000	0.047	0.649	5.029	74.7	6.967	Anaeromyxobacter_dehalogenans_2CP_1_uid58989
ATGC038	-0.362	0.000	0.093	0.661	7.308	68	5.524	Burkholderia_pseudomallei_1710b_uid58391
ATGC058	-0.351	0.000	0.066	0.635	5.744	64.9	4.937	Rhodopseudomonas_palustris_TIE_1_uid58995
ATGC006	-0.340	0.000	0.049	0.421	5.500	57.6	2.909	Klebsiella_variicola_At_22_uid42113
ATGC053	-0.322	0.000	0.051	0.464	4.740	57.3	1.427	Roseobacter_litoralis_Och_149_uid54719
ATGC051	-0.320	0.000	0.138	0.380	1.200	49.8	1.107	Anaplasma_marginale_Maries_uid57629
ATGC060	-0.317	0.000	0.094	0.565	2.623	38.8	1.328	Bartonella_tribocorum_CIP_105476_uid59129
ATGC041	-0.311	0.000	0.051	0.643	5.800	67	5.682	Ralstonia_solanacearum_MolK2_uid32085
ATGC119	-0.311	0.000	0.205	0.561	5.802	65.4	4.068	Mycobacterium_ulcerans_Agy99_uid62939
ATGC007	-0.305	0.000	0.040	0.415	4.620	57.3	3.344	Cronobacter_turicensis_z3032_uid40821
ATGC021	-0.305	0.000	0.045	0.453	5.046	58.2	3.580	Aeromonas_salmonicida_A449_uid58631
ATGC011	-0.304	0.000	0.054	0.408	4.800	54.5	2.065	Dickeya_zeae_Ech1591_uid59297
ATGC026	-0.303	0.000	0.036	0.526	6.200	61.5	4.795	Pseudomonas_putida_KT2440_uid57843
ATGC036	-0.300	0.000	0.067	0.431	4.100	54.2	2.310	Janthinobacterium_Marseille_uid58603
ATGC117	-0.299	0.000	0.110	0.428	3.300	53.8	2.146	Corynebacterium_glutamicum_ATCC_13032_uid61611
ATGC005	-0.295	0.000	0.082	0.388	5.176	50.6	1.493	Escherichia_coli_UTI89_uid58541
ATGC114	-0.286	0.000	0.072	0.504	5.723	60.7	2.511	Roseiflexus_castenholzii_DSM_13941_uid58287
ATGC043	-0.284	0.000	0.105	0.656	4.774	68.1	5.507	Bordetella_parapertussis_12822_uid57615
ATGC029	-0.266	0.000	0.062	0.430	3.101	42.2	0.603	Psychrobacter_cryohalolentis_K5_uid58373
ATGC059	-0.264	0.000	0.047	0.498	3.300	57.3	2.625	Brucella_suis_1330_uid57927
ATGC013	-0.263	0.000	0.067	0.418	3.844	59.7	3.619	Edwardsiella_tarda_EIB202_uid41819
ATGC125	-0.261	0.000	0.228	0.494	2.500	60.1	3.137	Propionibacterium_acnes_SK137_uid48071
ATGC104	-0.260	0.000	0.037	0.522	2.000	34.6	1.356	Lactobacillus_johnsonii_NCC_533_uid58029
ATGC107	-0.260	0.000	0.031	0.488	2.905	38	0.994	Listeria_monocytogenes_serotype_4b_F2365_uid57689
ATGC047	-0.258	0.000	0.160	0.415	0.859	41.1	0.172	Neorickettsia_sennetsu_Miyayama_uid57965
ATGC010	-0.253	0.000	0.049	0.408	5.100	50.5	1.190	Pectobacterium_wasabiae_WPP163_uid41297
ATGC023	-0.249	0.000	0.043	0.379	5.175	44.7	0.016	Shewanella_pealeana_ATCC_700345_uid58705
ATGC096	-0.229	0.000	0.164	0.419	1.839	38.5	0.856	Streptococcus_pyogenes_MGAS5005_uid58337
ATGC009	-0.228	0.000	0.136	0.350	4.770	47.5	0.357	Yersinia_pseudotuberculosis_PB1__uid59153
ATGC098	-0.225	0.000	0.206	0.550	2.031	36.8	1.190	Streptococcus_mutans_UA159_uid57947
ATGC035	-0.222	0.000	0.120	0.381	2.145	51.9	2.657	Neisseria_meningitidis_alpha14_uid61649
ATGC100	-0.219	0.206	0.222	0.398	2.000	40	0.546	Streptococcus_suis_P1_7_uid32235
ATGC109	-0.218	0.000	0.098	0.485	2.930	32.9	1.611	Staphylococcus_aureus_JH9_uid58455
ATGC080	-0.217	0.000	0.073	0.561	3.257	28.4	1.883	Clostridium_perfringens_ATCC_13124_uid57901
ATGC016	-0.217	0.000	0.089	0.480	1.830	38.2	1.254	Haemophilus_influenzae_Rd_KW20_uid57771
ATGC014	-0.212	0.000	0.120	0.443	2.345	41.2	1.295	Actinobacillus_pleuropneumoniae_serovar_7_AP76_uid59231
ATGC063	-0.212	0.000	0.106	0.415	2.056	46.3	0.864	Zymomonas_mobilis_ZM4_uid58095
ATGC094	-0.211	0.000	0.074	0.444	2.598	35.8	1.291	Lactococcus_lactis_cremoris_SK11_uid57983
ATGC002	-0.208	0.000	0.041	0.559	1.740	31.3	1.308	Methanocaldococcus_jannaschii_DSM_2661_uid57713
ATGC086	-0.201	0.000	0.067	0.625	0.880	29.7	1.216	Mycoplasma_agalactiae_uid46679
ATGC115	-0.201	0.000	0.106	0.440	1.470	48.9	1.502	Dehalococcoides_ethenogenes_195_uid57763
ATGC102	-0.198	0.000	0.075	0.386	3.034	46.7	0.512	Lactobacillus_rhamnosus_Lc_705_uid59315
ATGC089	-0.196	0.000	0.051	0.360	4.200	43.5	0.633	Bacillus_subtilis_168_uid57675
ATGC131	-0.195	0.000	0.187	0.388	2.343	48.3	0.918	Porphyromonas_gingivalis_W83_uid57641
ATGC110	-0.194	0.000	0.074	0.456	1.824	46.1	1.604	Thermotoga_petrophila_RKU_1_uid58655
ATGC022	-0.192	0.000	0.049	0.354	4.799	47.9	0.583	Shewanella_MR_7_uid58343
ATGC017	-0.191	0.000	0.035	0.429	4.480	38.2	1.118	Vibrio_fischeri_MJ11_uid58907
ATGC037	-0.185	0.000	0.032	0.426	2.159	44.8	0.498	Polynucleobacter_necessarius_asymbioticus_QLW_P1DMWA_1_uid58611
ATGC064	-0.184	0.000	0.131	0.603	3.960	66.3	4.060	Gluconacetobacter_diazotrophicus_PAl_5_uid61587
ATGC018	-0.184	0.000	0.096	0.361	4.250	47.3	0.550	Vibrio_cholerae_MJ_1236_uid59387
ATGC076	-0.182	0.000	0.151	0.369	2.400	50.7	0.710	Prochlorococcus_marinus_MIT_9313_uid57773

FIG 1 Quantitative characteristics of genome evolution of the 120 ATGC analyzed. *R* is the Spearman rank coefficient of correlation between F_{opt} and d*N*/d*S* in an ATGC. Column P includes the *P* values of *R* for each ATGC. F_{opt} is the median F_{opt} in an ATGC. d*N*/d*S* is the median d*N*/d*S* ratio in an ATGC. GS is the genome size of sample species. GC% is the GC content of sample species. deltaGC stands for ΔGC (see Results). The color code is explained at the bottom.

ATGC093	-0.176	0.000	0.066	0.440	5.400	35.3	1.242	Bacillus_cereus_B4264_uid58757
ATGC097	-0.171	0.000	0.152	0.430	2.128	35.6	1.278	Streptococcus_agalactiae_A909_uid57935
ATGC135	-0.163	0.000	0.061	0.507	1.174	39.3	0.787	Chlamydophila_felis_Fe_C_56_uid57971
ATGC133	-0.161	0.001	0.056	0.498	1.039	41.3	0.782	Chlamydia_trachomatis_L2b_UCH_1_proctitis_uid61635
ATGC019	-0.153	0.000	0.086	0.356	5.260	46.7	0.588	Vibrio_vulnificus_YJ016_uid58007
ATGC105	-0.152	0.000	0.230	0.386	3.340	44.4	0.540	Lactobacillus_plantarum_WCFS1_uid62911
ATGC088	-0.139	0.000	0.149	0.402	5.728	47.4	1.072	Desulfitobacterium_hafniense_Y51_uid58605
ATGC108	-0.138	0.000	0.114	0.576	2.627	32.1	1.483	Staphylococcus_epidermidis_RP62A_uid57663
ATGC003	-0.138	0.000	0.054	0.515	1.744	33.4	1.411	Methanococcus_maripaludis_C6_uid58947
ATGC095	-0.136	0.000	0.115	0.380	2.150	41.5	0.781	Streptococcus_equi_zooepidemicus_uid59261
ATGC015	-0.134	0.000	0.128	0.583	2.264	37.4	1.308	Haemophilus_somnus_2336_uid57979
ATGC099	-0.128	0.000	0.259	0.453	1.800	39.1	0.984	Streptococcus_thermophilus_LMG_18311_uid58219
ATGC101	-0.127	0.000	0.206	0.412	2.100	39.7	0.699	Streptococcus_pneumoniae_P1031_uid59123
ATGC137	-0.124	0.000	0.164	0.613	1.237	27.5	1.641	Borrelia_recurrentis_A1_uid58793
ATGC028	-0.110	0.000	0.068	0.416	3.761	39	1.047	Acinetobacter_baumannii_AB307_0294_uid59271
ATGC138	-0.104	0.000	0.104	0.585	0.920	29.1	1.414	Borrelia_turicatae_91E135_uid58311
ATGC012	-0.103	0.000	0.073	0.672	0.655	26.4	1.661	Buchnera_aphidicola_APS__Acyrthosiphon_pisum__uid57805
ATGC069	-0.103	0.000	0.065	0.443	3.954	38.9	0.855	Leptospira_biflexa_serovar_Patoc__Patoc_1__Paris__uid58993
ATGC078	-0.091	0.000	0.099	0.454	7.200	41.3	0.591	Nostoc_PCC_7120_uid57803
ATGC092	-0.078	0.000	0.103	0.397	5.523	37.9	0.593	Bacillus_megaterium_QM_B1551_uid15862
ATGC024	-0.070	0.000	0.151	0.586	1.900	32.3	1.423	Francisella_tularensis_WY96_3418_uid58811
ATGC130	-0.068	0.000	0.191	0.395	5.311	43.2	0.539	Bacteroides_fragilis_YCH46_uid58195
ATGC004	-0.057	0.047	0.237	0.507	2.842	35.3	0.782	Sulfolobus_islandicus_Y_N_15_51_uid58825
ATGC033	-0.049	0.059	0.184	0.358	2.521	51.8	0.487	Xylella_fastidiosa_Temecula1_uid57869
ATGC079	-0.046	0.008	0.167	0.533	4.804	39.8	1.075	Cyanothece_PCC_8802_uid59143
ATGC082	-0.042	0.000	0.092	0.650	4.018	28.3	1.655	Clostridium_botulinum_F_Langeland_uid58929
ATGC020	-0.031	0.002	0.033	0.438	5.166	45.4	0.881	Vibrio_parahaemolyticus_RIMD_2210633_uid57969
ATGC136	-0.018	0.252	0.095	0.612	0.983	28.1	1.359	Borrelia_garinii_PBi_uid58125
ATGC087	-0.016	0.637	0.040	0.737	0.752	25.5	1.987	Ureaplasma_parvum_serovar_3_ATCC_700970_uid57711
ATGC083	-0.014	0.420	0.105	0.632	4.200	28.8	1.759	Clostridium_difficile_R20291_uid40921
ATGC081	0.008	0.574	0.051	0.654	3.660	27.4	1.955	Clostridium_botulinum_E3_Alaska_E43_uid59157
ATGC044	0.011	0.679	0.112	0.614	1.112	28.9	1.499	Rickettsia_typhi_Wilmington_uid58063
ATGC049	0.024	0.403	0.119	0.586	1.500	27.5	1.443	Ehrlichia_ruminantium_Welgevonden_uid58243
ATGC030	0.042	0.000	0.111	0.507	3.636	38.3	0.688	Legionella_pneumophila_Paris_uid58211
ATGC073	0.055	0.002	0.115	0.543	1.843	35.1	1.092	Prochlorococcus_marinus_NATL2A_uid58359
ATGC045	0.065	0.000	0.139	0.584	1.258	32.5	1.161	Rickettsia_rickettsii__Sheila_Smith__uid58027
ATGC067	0.073	0.000	0.147	0.545	1.609	39.1	1.355	Helicobacter_pylori_HPAG1_uid58517
ATGC084	0.081	0.000	0.082	0.553	2.363	34.5	0.933	Thermoanaerobacter_pseudethanolicus_ATCC_33223_uid58339
ATGC050	0.089	0.001	0.055	0.582	1.176	30.1	1.346	Ehrlichia_chaffeensis_Arkansas_uid57933
ATGC048	0.143	0.000	0.207	0.521	1.446	35.2	0.767	Wolbachia_wRi_uid59371
ATGC068	0.161	0.000	0.083	0.633	1.778	30.3	1.617	Campylobacter_jejuni_RM1221_uid57899
ATGC111	0.189	0.000	0.035	0.561	1.856	34	1.401	Dictyoglomus_turgidum_DSM_6724_uid59177
ATGC085	0.266	0.000	0.126	0.646	0.900	28.5	1.254	Mycoplasma_hyopneumoniae_J_uid58059
ATGC075	0.268	0.000	0.055	0.606	1.658	30.8	1.423	Prochlorococcus_marinus_pastoris_CCMP1986_uid57761
ATGC074	0.285	0.000	0.069	0.600	1.642	31.3	1.333	Prochlorococcus_marinus_MIT_9301_uid58437

R	p	dN/dS	F	GS	GC%	deltaGC
<=-0.6		<=0.05	<=0.39	<=1	<=30	<=0.7
<=-0.5	<=0.05	<=0.07	<=0.43	<=2	<=35	<=1.4
<=-0.4		<=0.09	<=0.47	<=3	<=40	<=2.1
<=-0.3		<=0.11	<=0.51	<=4	<=45	<=2.8
<=-0.2		<=0.13	<=0.55	<=5	<=50	<=3.5
<=-0.1		<=0.15	<=0.59	<=6	<=55	<=4.2
<=0		<=0.17	<=0.63	<=7	<=60	<=4.9
<=0.1		<=0.19	<=0.67	<=8	<=65	<=5.6
<=0.2	>0.05	<=0.21	<=0.71	<=9	<=70	<=6.3
<=0.3		<0.26	<=0.75	<=10	<=75	<=7

FIG 1 (Continued)

properties; for the sake of simplicity, here we use dN/dS and F_{opt} to denote these median values. The dN/dS and F_{opt} values of the 120 ATGC showed a limited but statistically significant negative correlation (Fig. 2; Spearman's $\rho = -0.251$, $P = 0.0058$ [Spearman test]; here we denote this correlation coefficient R). Thus, the generally expected coupling between selection at the level of protein sequences and selection at the level of codon usage indeed seems to exist across a broad range of bacterial and archaeal genomes.

Having established the existence of the coupling between the two levels of selection, we sought to identify its possible underlying causes. Given that codon usage depends strongly on genomic GC content, which itself is strongly positively correlated with genome size (GS) (54) (Fig. 3, Spearman's $\rho = 0.661$, $P < 2.2e\text{-}16$), we turned to principal-component analysis (PCA) with five variables, dN/dS, F_{opt}, R, GS, and GC content (GC%). The first principal component explained more than half of the variation in the data, with the main contributions, with opposite signs, coming from GC content and R (Fig. 4). However, GS also makes a sub-

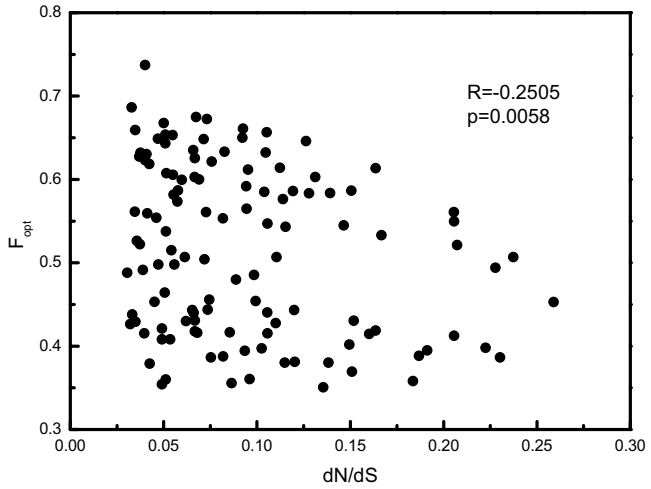

FIG 2 Correlation between the genomic median d*N*/d*S* and F_{opt} values of the 120 ATGC.

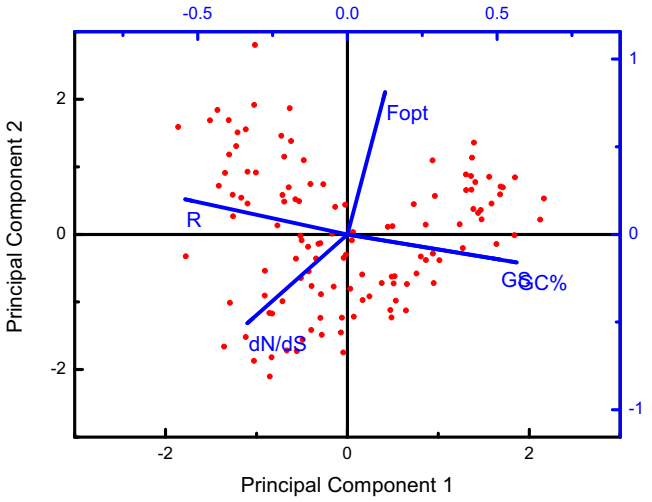

FIG 4 PCA of the 120 ATGC in the space of five variables, d*N*/d*S*, F_{opt}, *R*, GC content, and GS.

stantial contribution to principal component 1, conceivably because of the strong correlation between GC content and GS (Fig. 3). Principal component 2, which explains 22.5% of the data variance, reflects primarily the opposite contributions of *Fopt* and d*N*/d*S*, in agreement with the observed negative correlation (Fig. 2 and 3).

Pairwise correlation analysis showed that by far the strongest correlation exists between GC content and *R* (Fig. 5a), followed by the correlation between GS and *R* (Fig. 5b). Notably, the median d*N*/d*S* showed a relatively weak, albeit significant, negative correlation with GC content and GS (Fig. 5c and d), whereas there was no significant correlation between F_{opt} and either of these genomic characteristics (Fig. 5e and f). The peculiar, U-shaped dependence of F_{opt} on GC content most likely reflects the paucity of codon choices in extremely AT-rich and extremely GC-rich genomes, resulting in an inflation of F_{opt} values that does not reflect selective processes. Thus, the strong dependence of *R* on GC content appears to be a distinct phenomenon, with the implication that coupling between selection on protein sequence and selection on codon usage is a selectable trait in itself.

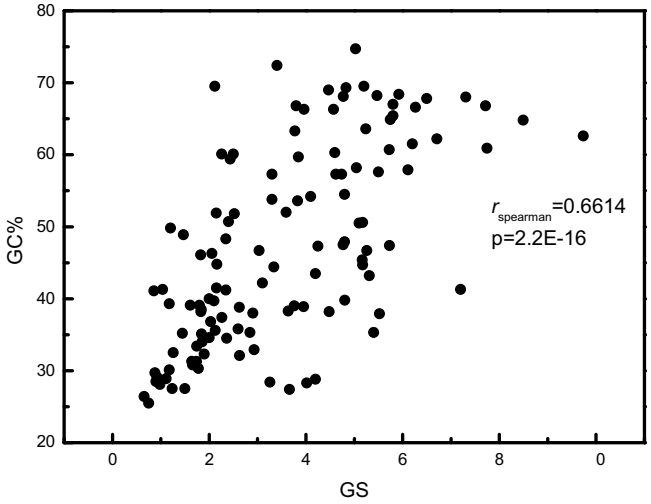

FIG 3 Correlation between the GC contents and GSs of the 120 ATGC. GSs are in megabases.

To further explore the potential biological underpinning of the strong connection between *R* and genomic GC content, we used the parameter ΔGC, which was defined as the difference in GC content between optimal and nonoptimal codons as follows:

$$\Delta GC = \sum_{i=1}^{18} \left(f_{opt}^{i} \times GC_{opt}^{i} - \overline{f_{nonopt}^{l} \times GC_{nonopt}^{l}} \right)$$

Here the sum is taken over 18 amino acids with more than one codon for all of the orthologous gene pairs in a given ATGC; f_{opt}^{i} and f_{nonopt}^{l} are the frequencies of the optimal and nonoptimal codons of amino acid *i*, respectively; GC_{opt}^{i} and GC_{nonopt}^{l} are, respectively, the GC contents of the optimal and nonoptimal codons of amino acid *i*; $f_{nonopt}^{l} \times GC_{nonopt}^{l}$ is the mean of all nonoptimal codes for amino acid *i* This parameter was designed to reflect the strength of selection for increased GC content in the optimal codon that could underlie the strong correlation between *R* and GC content. When the ΔGC values were plotted against the GC content for the 120 ATGC, a peculiar, nonmonotonic dependence was observed (Fig. 6). Whereas for low-GC genomes, ΔGC slightly decreased with the GC content, upward of ~45% GC, a steady increase in ΔGC was observed (Fig. 6). The small effect at a low GC content is likely to be purely statistical, caused by the strong bias toward AT. In contrast, at a high GC content, there seems to be strong selection for increased GC content of the optimal codons. Thus, the selection on codon bias indeed appears to be particularly pronounced in bacteria and archaea with GC-rich genomes.

Dependence of coupling between the two levels of selection on lifestyle, biological function, and taxonomy of prokaryotes. We further investigated possible connections of the coupling between the selection on amino acid sequences and on codon usage with various biological features of prokaryotes, including optimal growth temperature, cell shape, sporulation capacity, motility, and oxygen requirement. None of these biological properties showed a significant link with *R* (data not shown). It appeared particularly plausible that the coupling between the two levels of selection would be linked to the optimal growth rate (time between cell divisions under optimal growth conditions) of a mi-

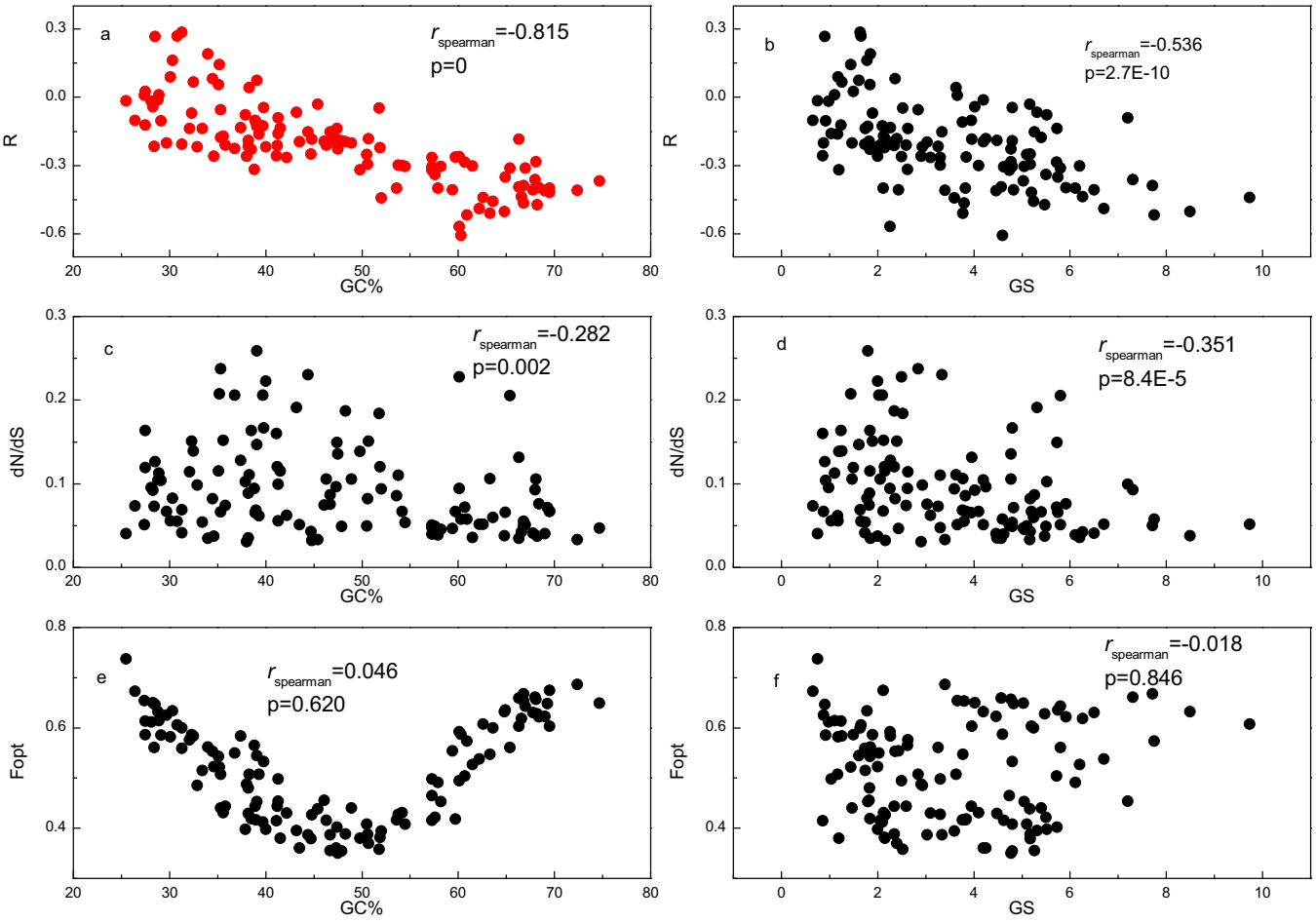

FIG 5 Pairwise correlation analysis of the five variables in the 120 ATGC. (a) *R* versus GC content. (b) *R* versus GS. (c) d*N*/d*S* versus GC content. (d) d*N*/d*S* versus GS. (e) F_{opt} versus GC content. (f) F_{opt} versus GS. GSs are in megabases.

crobe (55). Again, however, no connection between this parameter and *R* was found to exist (Fig. 7).

In contrast, the partitioning of *Proteobacteria* (the most extensively sequenced bacterial phylum) into pathogens and nonpathogens revealed a significantly stronger coupling among nonpathogens (Fig. 8). No such connection of d*N*/d*S* or F_{opt} values was detected. In contrast, a significant difference between pathogenic and non-pathogenic bacteria was observed also with respect to GC content, with a higher GC content in nonparasites (Fig. 8). These observations are compatible with the conclusion that (i) coupling of the selective processes at the protein and codon levels and (ii) GC content are subject to the same or related selective pressures.

We further explored the coupling in different functional classes of genes by using the coarse-grain classification implemented in the Clusters of Orthologous Groups (COG) system (56, 57). Although the differences between functional classes of genes were small in magnitude, genes that encode proteins related to metabolic activities (enzymes and transporters) consistently showed stronger coupling than informational genes encoding components of the translation, transcription, and replication systems (Fig. 9a). The difference was found to be statistically significant when the metabolic genes were pooled and collectively compared to informational genes (Fig. 9b).

Finally, we compared the strengths of the coupling between different bacterial and archaeal phyla (Fig. 10). Significant differences were detected, with *Actinobacteria* showing particularly strong coupling, in contrast to the weak coupling in *Cyanobacteria* and *Firmicutes*. Among the two most extensively sequenced phyla, *Proteobacteria* showed significantly stronger coupling than *Firmicutes*.

DISCUSSION

The results of the present analysis demonstrate the coupling between selection forces that affect protein sequences and codon usage. This relationship could be readily anticipated from previous observations on the relationships between gene expression level and protein sequence conservation on the one hand and CUB on the other (30). The coupling between selection on protein sequence and selection on codon usage can be interpreted as a fine-tuning of translation via CUB that depends on the "status" of a gene in an organism. "High-status" genes that are highly expressed tend to occupy central positions in various biological networks and typically evolve slowly (65), the selection for translational fine-tuning apparently is measurably stronger than it is in lower-status genes, resulting in the observed negative correlations between d*N*/d*S* and F_{opt}.

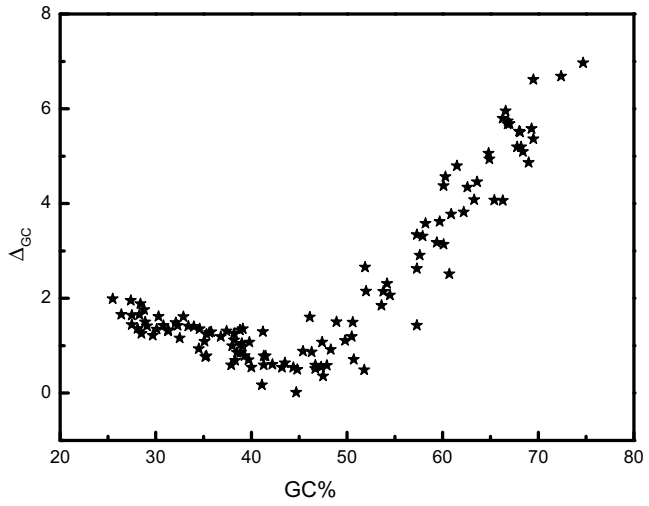

FIG 6 Dependence of ΔGC on the genomic GC contents of the 120 ATGC.

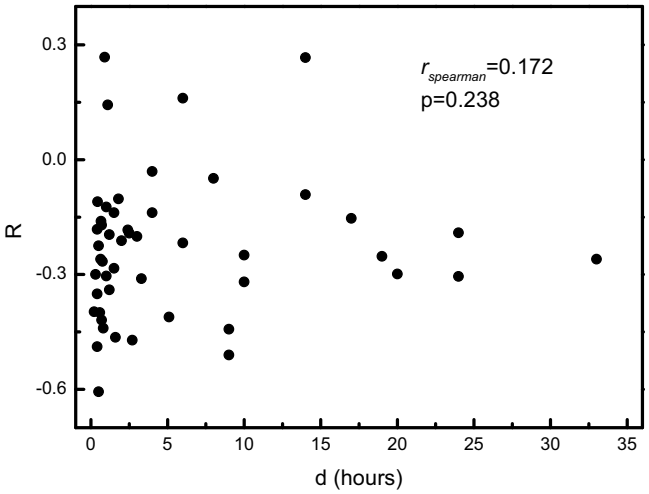

FIG 7 Dependence of *R* on optimal growth rates of bacteria and archaea. The optimal growth rate data (time between divisions under optimal growth conditions) are from reference 55.

The main, nontrivial observation in this work is that the strength of the coupling between the two levels of selection is effectively determined by the genomic GC content. It has been shown that mutational processes in all organisms are biased toward AT accumulation, so by inference, high GC content results from selection (58). The nature of this selection is not fully understood, but apparently, CUB is an important optimization criterion, as demonstrated by the finding that in bacteria, CUB tracks the nucleotide composition of the intergenic regions and in particular, that in sufficiently GC-rich genomes, the optimal codons typically contain G or C in synonymous positions. Moreover, the bias toward GC-rich codons is the strongest in highly expressed genes, such as those encoding translation system components (59). A subsequent, updated analysis indicates that the GC content in synonymous positions of codons tends to be higher than that in intergenic regions and that GC enrichment in synonymous positions without changing protein sequences results in increased fitness of bacteria expressing the respective genes (60). The results of the present study add an extra dimension to these observations by showing that the dependence of the translational fine-tuning on gene status is strongly correlated with the genomic GC content. In other words, in GC-rich genomes, the difference between the levels of translational fine-tuning in high- and low-status genes is

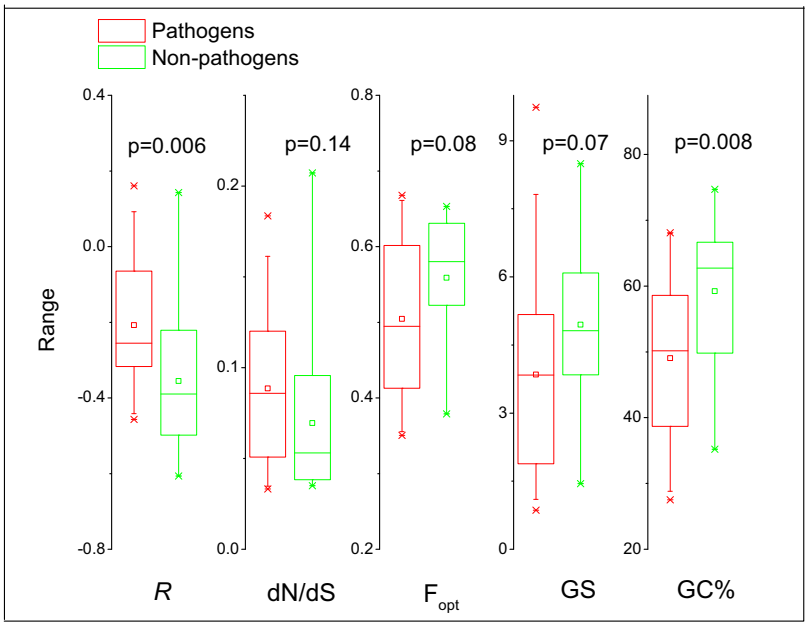

FIG 8 Comparison of the *R* values, d*N*/d*S* ratios, F_{opt} values, GSs, and GC contents of pathogens and nonpathogens in the phylum *Proteobacteria*. Of the 61 proteobacteria, 38 were classified as pathogens and 16 were classified as nonpathogens (data are from the GOLD [http://www.genomesonline.org/] and PATRIC [http://patric.vbi.vt.edu/] databases; the remaining 7 species were not classified in either of the two categories; see Table S1 in the supplemental material). The *R* values and GC contents of pathogens differ significantly from those of nonpathogens.

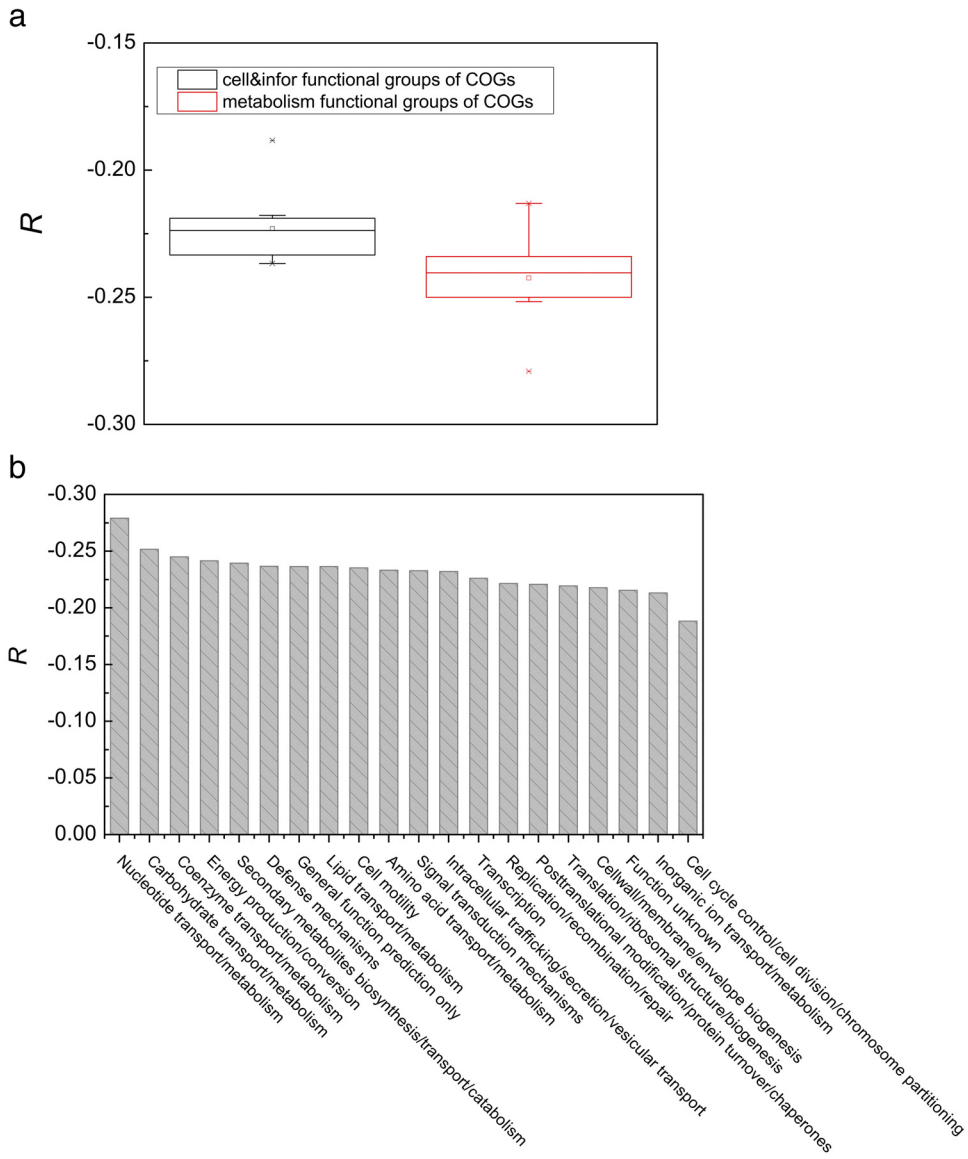

FIG 9 Correlation of the F_{opt} values and d*N*/d*S* ratios (*R* values) of different functional classes of genes of prokaryotes. The data for all 120 ATGC were pooled. (a) Comparison of *R* values of two broad categories of genes, those encoding proteins involved in information processing and cellular functions (e.g., cell division) (cell&infor) and those encoding proteins involved in metabolism (enzymes and transporters). The mean *R* value of the metabolic class (red box) is −0.242, which is significantly greater than the mean *R* value of −0.223 of the informational class ($P < 0.02$). (b) Comparison of *R* values of individual functional categories of genes. The functional categories are from the COG classifications.

greater than it is in AT-rich genomes, resulting in the observed strong correlation between strength of coupling and GC content.

This conclusion is clearly supported by the dependence of ΔGC on the genomic GC content (Fig. 6).

Although historically it is customary to speak of GC content determining other features of genomes, the causality could be reversed, with the selection for G and C at synonymous sites, which is particularly strong in high-status genes, driving the evolution toward high GC content (60). Furthermore, the overall optimization of the translational landscape of a microbial genome might enable the accumulation of genes via horizontal gene transfer and duplication, resulting in the strong positive correlation between GC content and GS (54) (Fig. 3).

In addition to demonstrating the coupling between protein level selection and CUB and its dependence on GC content, we observed that the strength of this coupling differs for parasites versus nonparasites and for genes encoding metabolic proteins versus those encoding informational proteins. Although subtle, these differences were found to be significant and did not appear to be by-products of the GC content connection. One might hypothesize that translational fine-tuning shows a stronger dependence on gene status in organisms and genes that are involved in frequent adaptation to changing environments and that this fine-tuning is particularly important in genes directly involved in such adaptation. Furthermore, some of the prokaryotic phyla significantly differ in the strength of coupling, which is suggestive of additional links with lifestyle and physiology.

The overall outcome of this analysis identifies the coupling

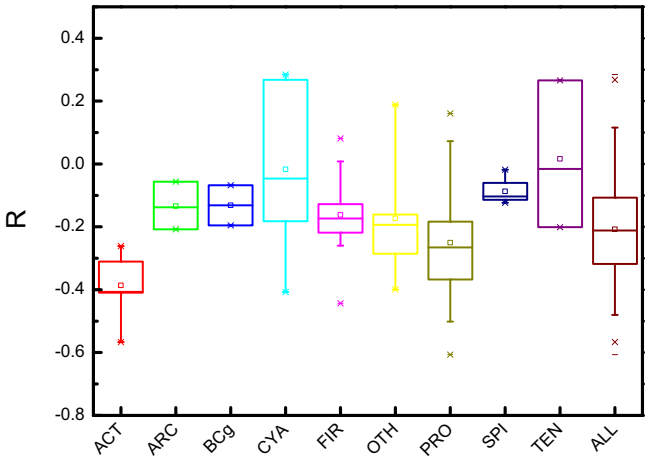

FIG 10 Comparison of *R* values of major prokaryotic taxa. The taxa shown belong to *Actinobacteria* (ACT), *Archaea* (ARC), the *Bacteroidetes-Chlorobi* group (BCg), *Cyanobacteria* (CYA), *Firmicutes* (FIR), other phyla (OTH), *Proteobacteria* (PRO), *Spirochaetes* (SPI), *Tenericutes* (TEN), and all phyla (ALL). One-way analysis of variance was used to test if any two groups differ significantly from each other. At the 0.05 level, the correlation in *Actinobacteria* is significantly stronger than those in the *Bacteroidetes-Chlorobi* group, *Cyanobacteria*, *Firmicutes*, *Proteobacteria*, *Spirochaetes*, *Tenericutes*, and other phyla, and the correlation in *Archaea* and *Proteobacteria* is significantly stronger than those in *Cyanobacteria*, *Firmicutes*, and *Tenericutes*.

between selection processes that act at the level of proteins and at the level of codon usage as a distinct characteristic of prokaryotic genome evolution. The strength of this coupling is tightly linked to genomic GC content and could be an important determinant of the nucleotide composition of genomes, the evolution of which remains poorly understood.

The biological factors behind the wide range of the strengths of coupling between the two levels of selection, from very strong negative correlation in many groups of microbes to a positive correlation in a few groups (Fig. 1 and 2), remain unclear. Explaining the nature of this variance and connecting it to specifics of microbial biology is a challenge for further research. Given the stronger coupling observed for operational genes than for informational genes (Fig. 9), it appears plausible that fine-tuning of CUB is subject to stronger selection in microbes whose lifestyle includes adaptation to changing environments that requires rapid protein dosage adjustment via translational regulation.

MATERIALS AND METHODS

The ATGC database and genome sequences. The ATGC database was built in 2009 and included 446 prokaryotic genomes and 104 ATGC (14). We updated the data set to include 1,390 genomes and 120 ATGC; *Firmicutes* and *Proteobacteria* account for 63.3% of the genomes (880/1,390) because of the relative paucity of sequenced genomes from other phyla. All of the pairs of orthologous genes in this database are synteny-supported bidirectional best hits (12, 14, 61). Altogether, 2,817,540 orthologous gene pairs were analyzed.

Selection of genome pairs for analysis. Suppose there are m species in an ATGC with the same gene number, n, which is the simplest case. The number of orthologous gene pairs is then $n \times \{m!/[2!(m-2)!]\}$. Obviously, the number of orthologous gene pairs rapidly increases with the number of species in an ATGC. If all of the orthologous gene pairs from all ATGC were taken into account in this analysis, the results would have been strongly biased toward large ATGC. Thus, we randomly chose a pair of species from each ATGC containing more than two species and used the orthologous gene pairs from these two species as a representative sample of the given ATGC.

Calculation of parameters. For each orthologous gene pair in an ATGC, protein sequences were aligned with MUSCLE (62), and the protein alignment was used to generate the alignment of the respective nucleotide sequences extracted from the genomic sequences by using a custom script. Maximum-likelihood approximation (codeML) was used to calculate dS and dN (63). In order to eliminate those orthologous gene pairs for which the estimates of the parameters was deemed unreliable either because of the small number of substitutions or conversely because of extreme divergence, the gene pairs with a dN value of <0.0002, a dS value of ≤ 0.0002, a dS value of >3, or a dN/dS ratio of >3 were discarded. The orthologous gene pairs in which the lengths of the two genes differed by more than 20% (presumably because of gene misannotation) were discarded as well. F_{opt} is a widely used measure of CUB (64). The F_{opt} values of genes in an orthologous gene pair are very close in most cases (data not shown). Thus, the mean F_{opt} value of two orthologous genes was taken as the F_{opt} value for that gene pair.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at http://mbio.asm.org/lookup/suppl/doi:10.1128/mBio.00956-14/-/DCSupplemental.

Table S1, PDF file, 0.1 MB.

ACKNOWLEDGMENTS

We thank Yuri Wolf for many helpful suggestions and critical reading of the manuscript and Koonin group members for useful discussions.

Our research is funded through the intramural fund of the U.S. Department of Health and Human Services (to the National Library of Medicine).

REFERENCES

1. **Kimura M.** 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge.
2. **Li WH.** 1997. Molecular evolution. Sinauer, Sunderland, MA.
3. **Yang Z, Bielawski JP.** 2000. Statistical methods for detecting molecular adaptation. Trends Ecol. Evol. **15:**496–503. http://dx.doi.org/10.1016/S0169-5347(00)01994-7.
4. **Hurst LD.** 2002. The Ka/Ks ratio: diagnosing the form of sequence evolution. Trends Genet. **18:**486. http://dx.doi.org/10.1016/S0168-9525(02)02722-1.
5. **Grishin NV, Wolf YI, Koonin EV.** 2000. From complete genomes to measures of substitution rate variability within and between proteins. Genome Res. **10:**991–1000. http://dx.doi.org/10.1101/gr.10.7.991.
6. **Larracuente AM, Sackton TB, Greenberg AJ, Wong A, Singh ND, Sturgill D, Zhang Y, Oliver B, Clark AG.** 2008. Evolution of protein-coding genes in Drosophila. Trends Genet. **24:**114–123. http://dx.doi.org/10.1016/j.tig.2007.12.001.
7. **Nielsen R.** 2005. Molecular signatures of natural selection. Annu. Rev. Genet. **39:**197–218. http://dx.doi.org/10.1146/annurev.genet.39.073003.112420.
8. **Koonin EV, Wolf YI.** 2010. Constraints and plasticity in genome and molecular-phenome evolution. Nat. Rev. Genet. **11:**487–498. http://dx.doi.org/10.1038/nrg2810.
9. **Lynch M, Conery JS.** 2003. The origins of genome complexity. Science **302:**1401–1404. http://dx.doi.org/10.1126/science.1089370.
10. **Lynch M.** 2007. The origins of genome architecture. Sinauer Associates, Sunderland, MA.
11. **Jordan IK, Rogozin IB, Wolf YI, Koonin EV.** 2002. Microevolutionary genomics of bacteria. Theor. Popul. Biol. **61:**435–447. http://dx.doi.org/10.1006/tpbi.2002.1588.
12. **Novichkov PS, Wolf YI, Dubchak I, Koonin EV.** 2009. Trends in prokaryotic evolution revealed by comparison of closely related bacterial and archaeal genomes. J. Bacteriol. **191:**65–73. http://dx.doi.org/10.1128/JB.01237-08.
13. **Takeuchi N, Wolf YI, Makarova KS, Koonin EV.** 2012. Nature and intensity of selection pressure on CRISPR-associated genes. J. Bacteriol. **194:**1216–1225. http://dx.doi.org/10.1128/JB.06521-11.

14. **Novichkov PS, Ratnere I, Wolf YI, Koonin EV, Dubchak I.** 2009. ATGC: a database of orthologous genes from closely related prokaryotic genomes and a research platform for microevolution of prokaryotes. Nucleic Acids Res. **37:**D448–D454. http://dx.doi.org/10.1093/nar/gkn684.
15. **Katz L, Burge CB.** 2003. Widespread selection for local RNA secondary structure in coding regions of bacterial genes. Genome Res. **13:**2042–2051. http://dx.doi.org/10.1101/gr.1257503.
16. **Stoletzki N.** 2008. Conflicting selection pressures on synonymous codon use in yeast suggest selection on mRNA secondary structures. BMC Evol. Biol. **8:**224. http://dx.doi.org/10.1186/1471-2148-8-224.
17. **Shabalina SA, Spiridonov NA, Kashina A.** 2013. Sounds of silence: synonymous nucleotides as a key to biological regulation and complexity. Nucleic Acids Res. **41:**2073–2094. http://dx.doi.org/10.1093/nar/gks1205.
18. **Resch AM, Carmel L, Mariño-Ramírez L, Ogurtsov AY, Shabalina SA, Rogozin IB, Koonin EV.** 2007. Widespread positive selection in synonymous sites of mammalian genes. Mol. Biol. Evol. **24:**1821–1831. http://dx.doi.org/10.1093/molbev/msm100.
19. **Knight RD, Freeland SJ, Landweber LF.** 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–research0010.13. http://dx.doi.org/10.1186/gb-2001-2-4-research0010.
20. **Chen SL, Lee W, Hottes AK, Shapiro L, McAdams HH.** 2004. Codon usage between genomes is constrained by genome-wide mutational processes. Proc. Natl. Acad. Sci. U. S. A. **101:**3480–3485. http://dx.doi.org/10.1073/pnas.0307827100.
21. **Gouy M, Gautier C.** 1982. Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res. **10:**7055–7074. http://dx.doi.org/10.1093/nar/10.22.7055.
22. **Ikemura T.** 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. **2:**13–34.
23. **Bulmer M.** 1987. Coevolution of codon usage and transfer RNA abundance. Nature **325:**728–730. http://dx.doi.org/10.1038/325728a0.
24. **Kurland CG.** 1991. Codon bias and gene expression. FEBS Lett. **285:**165–169. http://dx.doi.org/10.1016/0014-5793(91)80797-7.
25. **Karlin S, Mrázek J, Campbell AM.** 1998. Codon usages in different gene classes of the *Escherichia coli* genome. Mol. Microbiol. **29:**1341–1355. http://dx.doi.org/10.1046/j.1365-2958.1998.01008.x.
26. **Hershberg R, Petrov DA.** 2008. Selection on codon bias. Annu. Rev. Genet. **42:**287–299. http://dx.doi.org/10.1146/annurev.genet.42.110807.091442.
27. **Plotkin JB, Kudla G.** 2011. Synonymous but not the same: the causes and consequences of codon bias. Nat. Rev. Genet. **12:**32–42. http://dx.doi.org/10.1038/nrg2899.
28. **Sharp PM, Li WH.** 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. J. Mol. Evol. **24:**28–38. http://dx.doi.org/10.1007/BF02099948.
29. **Sharp PM, Li WH.** 1987. The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol. Biol. Evol. **4:**222–230.
30. **Drummond DA, Wilke CO.** 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell **134:**341–352. http://dx.doi.org/10.1016/j.cell.2008.05.042.
31. **Hartl DL, Moriyama EN, Sawyer SA.** 1994. Selection intensity for codon bias. Genetics **138:**227–234.
32. **Parker J, Johnston TC, Borgia PT, Holtz G, Remaut E, Fiers W.** 1983. Codon usage and mistranslation. *In Vivo* basal level misreading MS 2 coat protein message. J. Biol. Chem. **258:**10007–10012.
33. **Precup J, Parker J.** 1987. Missense misreading of asparagine codons as a function of codon identity and context. J. Biol. Chem. **262:**11351–11355.
34. **Sørensen MA, Kurland CG, Pedersen S.** 1989. Codon usage determines translation rate in *Escherichia coli*. J. Mol. Biol. **207:**365–377. http://dx.doi.org/10.1016/0022-2836(89)90260-X.
35. **Akashi H.** 1994. Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. Genetics **136:**927–935.
36. **Stoletzki N, Eyre-Walker A.** 2007. Synonymous codon usage in *Escherichia coli*: selection for translational accuracy. Mol. Biol. Evol. **24:**374–381.
37. **Sørensen MA, Pedersen S.** 1991. Absolute *in vivo* translation rates of individual codons in *Escherichia coli*. The two glutamic acid codons GAA and GAG are translated with a three-fold difference in rate. J. Mol. Biol. **222:**265–280. http://dx.doi.org/10.1016/0022-2836(91)90211-N.
38. **Marintchev A, Wagner G.** 2004. Translation initiation: structures, mechanisms and evolution. Q. Rev. Biophys. **37:**197–284. http://dx.doi.org/10.1017/S0033583505004026.
39. **Kudla G, Murray AW, Tollervey D, Plotkin JB.** 2009. Coding-sequence determinants of gene expression in *Escherichia coli*. Science **324:**255–258. http://dx.doi.org/10.1126/science.1170160.
40. **Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE.** 2005. Variation in the strength of selected codon usage bias among bacteria. Nucleic Acids Res. **33:**1141–1153. http://dx.doi.org/10.1093/nar/gki242.
41. **Higgs PG, Ran W.** 2008. Coevolution of codon usage and tRNA genes leads to alternative stable states of biased codon usage. Mol. Biol. Evol. **25:**2279–2291. http://dx.doi.org/10.1093/molbev/msn173.
42. **Sharp PM, Emery LR, Zeng K.** 2010. Forces that influence the evolution of codon bias. Philos. Trans. R. Soc. Lond. B Biol. Sci. **365:**1203–1212. http://dx.doi.org/10.1098/rstb.2009.0305.
43. **Ran W, Higgs PG.** 2010. The influence of anticodon-codon interactions and modified bases on codon usage bias in bacteria. Mol. Biol. Evol. **27:**2129–2140. http://dx.doi.org/10.1093/molbev/msq102.
44. **Qian W, Yang JR, Pearson NM, Maclean C, Zhang J.** 2012. Balanced codon usage optimizes eukaryotic translational efficiency. PLoS Genet. **8:**e1002603. http://dx.doi.org/10.1371/journal.pgen.1002603.
45. **Ran W, Higgs PG.** 2012. Contributions of speed and accuracy to translational selection in bacteria. PLoS One **7:**e51652. http://dx.doi.org/10.1371/journal.pone.0051652.
46. **Quax TE, Wolf YI, Koehorst JJ, Wurtzel O, van der Oost R, Ran W, Blombach F, Makarova KS, Brouns SJ, Forster AC, Wagner EG, Sorek R, Koonin EV, van der Oost J.** 2013. Differential translation tunes uneven production of operon-encoded proteins. Cell Rep. **4:**938–944. http://dx.doi.org/10.1016/j.celrep.2013.07.049.
47. **Bulmer M.** 1991. The selection-mutation-drift theory of synonymous codon usage. Genetics **129:**897–907.
48. **Akashi H.** 1995. Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in Drosophila DNA. Genetics **139:**1067–1076.
49. **Duret L.** 2002. Evolution of synonymous codon usage in metazoans. Curr. Opin. Genet. Dev. **12:**640–649. http://dx.doi.org/10.1016/S0959-437X(02)00353-2.
50. **Pál C, Papp B, Hurst LD.** 2001. Highly expressed genes in yeast evolve slowly. Genetics **158:**927–931. (Letter.)
51. **Krylov DM, Wolf YI, Rogozin IB, Koonin EV.** 2003. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. Genome Res. **13:**2229–2235. http://dx.doi.org/10.1101/gr.1589103.
52. **Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH.** 2005. Why highly expressed proteins evolve slowly. Proc. Natl. Acad. Sci. U. S. A. **102:**14338–14343. http://dx.doi.org/10.1073/pnas.0504070102.
53. **Drummond DA, Raval A, Wilke CO.** 2006. A single determinant dominates the rate of yeast protein evolution. Mol. Biol. Evol. **23:**327–337. http://dx.doi.org/10.1093/molbev/msj038.
54. **Nakabachi A, Yamashita A, Toh H, Ishikawa H, Dunbar HE, Moran NA, Hattori M.** 2006. The 160-kilobase genome of the bacterial endosymbiont Carsonella. Science **314:**267. http://dx.doi.org/10.1126/science.1134196.
55. **Couturier E, Rocha EP.** 2006. Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes. Mol. Microbiol. **59:**1506–1518. http://dx.doi.org/10.1111/j.1365-2958.2006.05046.x.
56. **Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA.** 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics **4:**41. http://dx.doi.org/10.1186/1471-2105-4-41.
57. **Tatusov RL, Koonin EV, Lipman DJ.** 1997. A genomic perspective on protein families. Science **278:**631–637. http://dx.doi.org/10.1126/science.278.5338.631.
58. **Hershberg R, Petrov DA.** 2010. Evidence that mutation is universally biased towards AT in bacteria. PLoS Genet. **6:**e1001115. http://dx.doi.org/10.1371/journal.pgen.1001115.
59. **Hershberg R, Petrov DA.** 2009. General rules for optimal codon choice. PLoS Genet. **5:**e1000556. http://dx.doi.org/10.1371/journal.pgen.1000556.
60. **Raghavan R, Kelkar YD, Ochman H.** 2012. A selective force favoring increased G+C content in bacterial genes. Proc. Natl. Acad. Sci. U. S. A. **109:**14504–14507. http://dx.doi.org/10.1073/pnas.1205683109.
61. **Kristensen DM, Wolf YI, Mushegian AR, Koonin EV.** 2011. Computa-

tional methods for gene orthology inference. Brief. Bioinform. **12:** 379–391. http://dx.doi.org/10.1093/bib/bbr030.
62. **Edgar RC.** 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32:**1792–1797. http://dx.doi.org/10.1093/nar/gkh340.
63. **Yang Z.** 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. **24:**1586–1591. http://dx.doi.org/10.1093/molbev/msm088.
64. **Ikemura T.** 1981. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. J. Mol. Biol. **151:**389–409. http://dx.doi.org/10.1016/0022-2836(81)90003-6.
65. **Wolf YI, Carmel L, Koonin EV.** 2006. Unifying measures of gene function and evolution. Proc. Biol. Sci. **273:**1507–1515.