Supplementary material for Lee et al., Translationally optimal codons associate with aggregation-prone sites in proteins

Table S1: Odds ratio of optimal codon usage between aggregation-prone and not-aggregation-prone sites for each amino acid in *Halobacterium* sp. NRC-1.

AA	Optimal codon	Odds ratio
Ala	GCG	1.43^{***}
Arg	AGA	0.86
Asn	–	–
Asp	GAC	0.83^{**}
Cys	UGU	1.14
Gln	–	–
Glu	–	–
Gly	–	–
His	–	–
Ile	AUC	1.28^{**}
Leu	UUG	1.28^{***}
Lys	AAA	$0.73^{*(*)}$
Phe	UUC	0.92
Pro	CCG	1.42
Ser	AGU	0.95
Thr	–	–
Tyr	UAC	1.06
Val	GUG	1.19^{***}
Overall		1.22^{***}

Note.—AA: amino acid; -: no optimal codon. Significance levels: $^{***}P < 0.001$; $^{**}P < 0.01$; $^{*}P < 0.05$. Significance levels in parentheses disappear after correction for multiple testing. We determined optimal codons based on codon usage, as follows. We calculated the adjusted effective number of codons (ENC') for each gene, according to the method developed by Novembre [Mol. Biol. Evol. 19:1390, 2002], which corrects for nucleotide content. We then compared the codon usage pattern between the genes showing the lowest 10% and highest 10% ENC'. We defined codons as *optimal* if they showed a statistically significant increase in frequency in the genes with low ENC', as determined by a χ^2 test. Amino acids Asn, Gln, Glu, Gly, His, and Thr had no optimal codons according to this definition and were omitted from the analysis.

Table S2: Odds ratio of optimal codon usage between aggregation-prone and not-aggregation-prone sites and between buried and exposed sites only for genes with structural data.

	E. coli		*S. cerevisiae*		*D. melanogaster*		*M. musculus*	
AA	A/N	B/E	A/N	B/E	A/N	B/E	A/N	B/E
Ala	1.01	0.92	0.87	1.19	1.21$^{(*)}$	1.14$^{(*)}$	–	–
Arg	1.00	1.08	1.38	0.89	1.29	1.02	1.29$^{*(*)}$	1.14$^{*(*)}$
Asn	1.21**	1.34***	0.99	1.10	1.39**	1.07	1.16*	1.20$^{**(*)}$
Asp	1.15*	1.08	0.91	0.85	1.55***	0.91	1.15*	1.09
Cys	0.99	0.85	0.67	1.82$^{(*)}$	1.02	0.95	–	–
Gln	1.00	1.21**	1.16	1.06	0.85	0.91	1.17$^{(*)}$	1.23**
Glu	0.98	1.10	1.10	0.92	0.93	1.02	0.92	1.04
Gly	1.27***	1.07	0.98	1.30$^{*(*)}$	1.06	1.01	1.19***	1.08
His	1.28**	1.00	1.22	0.85	1.26	0.75$^{(*)}$	1.21*	1.05
Ile	1.14$^{*(*)}$	0.99	0.77$^{(*)}$	1.14	0.94	1.09	1.14$^{*(*)}$	1.09
Leu	0.84***	1.14$^{**(*)}$	1.03	1.10	1.03	1.08	0.84***	1.02
Lys	–	–	0.75$^{(*)}$	1.29$^{*(*)}$	0.72$^{*(*)}$	0.91	1.17*	1.06
Phe	1.02	0.99	1.04	1.18	0.89	1.03	1.14$^{*(*)}$	1.07
Pro	0.92	1.07	0.98	1.24$^{(*)}$	0.58	1.10	0.71	1.01
Ser	1.26***	1.42***	1.22$^{(*)}$	1.13	0.95	1.10	1.16$^{*(*)}$	1.22***
Thr	1.26***	1.22***	0.99	1.11	1.24$^{*(*)}$	1.09	1.26***	1.14**
Tyr	1.23**	0.85*	1.07	0.90	1.30$^{(*)}$	0.94	–	–
Val	0.85***	0.88$^{*(*)}$	1.29$^{(**)}$	1.16	1.05	1.23$^{(*)}$	0.96	1.07
Overall	1.07***	1.08***	1.02	1.10$^{**(*)}$	1.09$^{**(*)}$	1.04	1.09***	1.09***

Note.—AA: amino acid; A/N: odds ratio between aggregation-prone and not-aggregation-prone sites; B/E: odds ratio between buried and exposed sites; -: no optimal codon. Significance levels: $^{***}P < 0.001$; $^{**}P < 0.01$; $^{*}P < 0.05$. Significance levels in parentheses disappear after correction for multiple testing.

Table S3: Odds ratio of optimal codon usage between aggregation-prone and not-aggregation-prone sites for each amino acid (aggregation propensity cutoff = 0)

AA	*E. coli*	*S. cerevisiae*	*D. melanogaster*	*M. musculus*
Asn	1.13***	1.05***	1.23***	1.04***
Ala	1.02	1.03*	1.33***	–
Arg	1.12***	1.01	1.26***	1.18***
Asp	1.07$^{**(*)}$	1.00	1.24***	1.11***
Cys	0.92	1.04	1.25***	–
Gln	1.10***	1.04$^{(*)}$	0.97	1.04***
Glu	1.06**	1.03$^{(*)}$	1.03	0.98*
Gly	1.20***	1.11***	1.20***	1.17***
His	1.20***	1.03	1.32***	1.03*
Ile	1.06**	1.06***	1.14***	1.11***
Leu	0.98	1.01	1.04**	0.92***
Lys	–	0.9***	0.94**	0.97**
Phe	0.98	1.00	1.09***	1.09***
Pro	1.06	1.01	1.44***	0.98
Ser	1.10***	1.08***	1.17***	1.15***
Thr	1.35***	1.06***	1.44***	1.13***
Tyr	1.20***	1.04$^{(*)}$	1.29***	–
Val	0.85***	1.1***	1.08***	0.99
Overall	1.07***	1.03***	1.17***	1.05***

Note.—AA: amino acid; -: no optimal codon. Significance levels: $^{***}P < 0.001$; $^{**}P < 0.01$; $^{*}P < 0.05$. Significance levels in parentheses disappear after correction for multiple testing.

Table S4: Odds ratio of optimal codon usage between exposed-aggregation-prone and exposed-not-aggregation-prone sites, between buried-aggregation-prone and buried-not-aggregation-prone sites, between exposed-aggregation-prone and buried-not-aggregation-prone sites, and between buried-aggregation-prone and exposed-not-aggregation-prone sites for each amino acid (aggregation propensity cutoff = 0)

	E. coli				*S. cerevisiae*				*D. melanogaster*				*M. musculus*			
AA	EA-EN	BA-BN	EA-BN	BA-EN	EA-EN	BA-BN	EA-BN	BA-EN	EA-EN	BA-BN	EA-BN	BA-EN	EA-EN	BA-BN	EA-BN	BA-EN
Ala	1.07	1.02	1.13	0.97	1.15	0.97	0.77	1.19	1.15	1.12	1.01	$1.31^{*(*)}$	–	–	–	–
Arg	1.04	1.14	0.97	1.12	1.25	1.20	1.24	0.96	1.50^{***}	1.10	$1.38^{*(*)}$	1.26	1.22^{**}	1.07	1.02	1.26^{**}
Asn	$1.22^{*(*)}$	1.05	0.85	1.58^{***}	1.09	1.28	1.03	1.20	0.99	$1.67^{**(*)}$	1.20	$1.33^{(*)}$	0.96	1.16	0.87	1.21^{*}
Asp	1.05	1.16	1.06	$1.18^{*(*)}$	0.95	0.99	1.07	0.79	1.29^{**}	$1.64^{**(*)}$	1.67^{***}	1.21	1.07	1.08	0.98	$1.15^{(*)}$
Cys	3.14	0.88	0.92	1.04	1.33	0.70	$0.08^{*(*)}$	0.93	0.94	1.29	1.40	1.23	–	–	–	–
Gln	1.08	1.00	0.89	$1.21^{(*)}$	1.23	1.30	1.25	1.16	0.92	1.14	1.24	0.99	1.04	1.03	$0.80^{(*)}$	1.12
Glu	1.10	1.03	1.02	1.16	1.04	1.30	1.16	1.07	0.93	0.82	0.96	0.88	$0.90^{(*)}$	0.95	0.88	1.01
Gly	$1.25^{**(*)}$	$1.28^{**(*)}$	$1.22^{*(*)}$	1.28^{***}	1.29	1.19	0.91	1.58^{***}	1.06	1.20	1.16	1.08	1.05	$1.18^{*(*)}$	1.05	$1.13^{(*)}$
His	1.15	1.60^{***}	$1.32^{(*)}$	1.18	$1.64^{(*)}$	0.99	$1.86^{(*)}$	1.22	1.02	$1.58^{*(*)}$	2.19^{***}	1.03	1.08	1.01	1.05	1.09
Ile	1.07	1.10	$1.21^{(*)}$	1.10	0.88	0.82	0.86	1.25	1.05	1.03	0.94	1.20	1.23	1.06	1.07	$1.28^{*(*)}$
Leu	0.92	$0.88^{**(*)}$	$0.80^{**(*)}$	1.04	0.90	0.99	0.83	1.06	0.94	0.99	0.86	1.08	0.88	$0.90^{*(*)}$	$0.86^{(**)}$	0.91
Lys	–	–	–	–	0.84	0.76	$0.63^{*(*)}$	1.14	$0.83^{(*)}$	0.73	0.95	0.80	0.93	$1.23^{(*)}$	0.99	1.11
Phe	0.76	0.97	0.92	0.92	1.23	0.93	0.93	1.10	1.10	1.05	0.96	1.05	0.98	1.08	1.02	1.01
Pro	1.10	1.10	1.09	1.15	1.31	1.22	1.15	1.33	1.61^{*}	1.24	1.22	1.41	1.09	0.85	1.10	0.86
Ser	1.03	1.24^{**}	$0.78^{*(*)}$	1.56^{***}	1.10	$1.42^{(**)}$	1.14	$1.32^{(*)}$	1.01	1.01	0.98	1.19	$1.16^{(*)}$	1.08	0.87	1.39^{***}
Thr	1.36^{***}	1.36^{***}	1.11	1.70^{***}	$1.39^{(*)}$	1.04	1.14	$1.45^{(*)}$	$1.34^{*(*)}$	$1.45^{*(*)}$	$1.35^{(*)}$	$1.36^{*(*)}$	$1.27^{**(*)}$	1.10	0.99	1.34^{***}
Tyr	1.06	1.23^{*}	$1.35^{*(*)}$	1.03	1.26	1.17	1.23	1.02	1.45	1.23	1.37	1.27	–	–	–	–
Val	0.91	0.79^{***}	0.91	0.73^{***}	0.90	$1.24^{(*)}$	1.00	1.17	1.20	1.08	0.98	$1.39^{*(*)}$	1.02	0.97	0.89	1.05
Overall	1.09^{***}	1.03	1.00	1.16^{***}	1.07	$1.07^{(*)}$	0.98	1.17^{***}	$1.11^{**(*)}$	1.13^{***}	1.13^{***}	1.17^{***}	1.03	1.02	$0.95^{(**)}$	1.12^{***}

Note.—AA: amino acid; EA: exposed and aggregation-prone sites; BN: buried and not-aggregation-prone sites; BA: buried and aggregation-prone sites; EN: exposed and not-aggregation-prone sites; -: no optimal codon. Significance levels: $^{***}P < 0.001$; $^{**}P < 0.01$; $^{*}P < 0.05$. Significance levels in parentheses disappear after correction for multiple testing.

Table S5: Odds ratio of optimal codon usage between sites that are sensitive ($S_i > 0$) and not sensitive ($S_i \leq 0$) to amino-acid substitutions. Analysis was carried out for 845 genes in fly only.

AA	Odds ratio
Ala	1.30^{***}
Arg	$1.16^{*(*)}$
Asn	1.10
Asp	1.10
Cys	$0.77^{(*)}$
Gln	1.02
Glu	0.97
Gly	1.01
His	1.06
Ile	0.98
Leu	1.00
Lys	$0.87^{*(*)}$
Phe	$0.86^{(*)}$
Pro	1.26^{***}
Ser	1.01
Thr	$0.85^{*(*)}$
Tyr	$0.84^{(*)}$
Val	1.03
Overall	$1.03^{(*)}$

Note.—AA: amino acid. Significance levels: $^{***}P < 0.001$; $^{**}P < 0.01$; $^{*}P < 0.05$. Significance levels in parentheses disappear after correction for multiple testing.