Codon usage in *E. coli*: an evolutionary approach

Fanny POUYET[1], Julien JACQUEMETTON[1], Marc BAILLY-BECHET[1] and Laurent GUÉGUEN[1]

Lab. Biométrie et Biologie Évolutive, UMR5558 CNRS, 43 Bd du 11 nov. 1918, 69622 Villeurbane Cedex, France
{fanny.pouyet, julien.jacquemetton, marc.bailly-bechet, laurent.gueguen}@univ-lyon1.fr

Abstract *We develop a codon-based evolutionary model, based on previous works by Yang and Nielsen [11], with the capacity to distinguish between selective pressures acting specifically on codon usage or more generally on nucleotidic content. Our model, implemented in Bio++ [5] is multi-layered and allows to infer: i) the equilibrium frequencies for the nucleotidic mutational process; ii) the strengh of codon usage between the synonymous codons of each amino acid; and iii) the amino acid preferences. We apply this model in an homogeneous, non-stationary context on a dataset of three close* E. coli *strains [9] and show that codon usage and nucleotidic mutational process are counteracting each other, mutational process tending to increase dramatically the AT content in the equilibrium frequencies, while selection for codon usage acts towards a more balanced GC content.*

Keywords Evolutionary model, codon usage, mutational bias, *E. coli*.

Usage du code chez *E. coli*: une approche évolutive

Abstract *Nous avons développé un modèle évolutif à l'échelle des codons, basé sur des travaux de Yang et Nielsen [11], qui permet de distinguer entre des pressions évolutives agissant sur l'usage du code ou sur la composition nucléotidique. Notre modèle, implémenté en Bio++ [5], est multi-couche et permet d'inférer: i) les fréquences d'équilibre du processus mutationnel à l'échelle nucléotidique, ii) l'intensité du biais d'usage du code entre codons synonymes pour chaque acide aminé, et iii) les préférences évolutives pour chaque acide aminé. Nous appliquons ce modèle à un jeu de données de gènes provenant de 3 souches proches de* E. coli *[9], et montrons que le biais d'usage du code génétique et les tendances mutationnelles à l'échelle nucléotidique sont en opposition: les processus mutationnels tendent vers une augmentation du contenu en AT des gènes, tandis que la sélection sur l'usage du code favorise un contenu plus équilibré en GC.*

Keywords Modèle évolutif, usage du code, biais mutationnels, *E. coli*

The genetic code is degenerated, allowing for different codons – said nsynonymous – to code for the same amino acid. Codon usage bias is the non-random preference, in a gene or more generally a genome, for using a particular codon over synonymous ones. Thanks to the accumulation of genomic sequences, codon bias has been documented in all living organisms – see [6,10] for a review. Various causes have been advanced to explain the existence of this bias; amongst them, two main hypotheses have emerged: *mutational bias* and *translational selection*. The *mutational bias* hypothesis postulates that codon usage is due to global or local biases in the mutation patterns, that affect in particular unselected positions in a sequence, such as the third position of codons in amino acids four times degenerated, where all changes have no consequence on the protein sequence. Conversely, the *translational selection* hypothesis says that the choice of a particular codon can benefit the organism, by making the translation of this codon – and more generally the translation of the entire gene – more accurate or more efficient. This hypothesis is supported by the observed correlation between codon frequencies and cognate tRNA frequencies in various bacteria [3], correlation that has been predicted in models of translational selection [2].

In bacterial genomes, both forces seem to act, with translational selection being very strong mainly for highly expressed proteins such as ribosomal proteins. In order to disentangle these two forces, we developed an evolutionary model, inspired from the works of Yang and Nielsen [11], in which we can infer a preference parameter for each codon, relative to its synonymous ones. These preferences, denoted $\phi_{aa}(i)$ for codon i inside

amino acid aa, are the relative equilibrium frequencies of the codons, inside their amino acids, if only selection for codon usage was acting on the sequences. Moreover, our model includes classical equilibrium nucleotidic frequencies to account for mutational biases. In a more mathematical way, one can write the generator q_{ij} of our model, i.e. the probability to observe a substitution between codon i and codon j, as:

$$q_{ij} \propto \begin{cases} 0 & \text{if more than one nucleotide change} \\ \pi_{j_p} \cdot \frac{-\log\left(\frac{\phi_{aa}(i)}{\phi_{aa}(j)}\right)}{1-\frac{\phi_{aa}(i)}{\phi_{aa}(j)}} & \text{synonymous transversion} \\ \pi_{j_p} . \kappa . \frac{-\log\left(\frac{\phi_{aa}(i)}{\phi_{aa}(j)}\right)}{1-\frac{\phi_{aa}(i)}{\phi_{aa}(j)}} & \text{synonymous transition} \\ \pi_{j_p} . \omega . \frac{-\log\left(\frac{\psi_{aa_i}\phi_{aa_i}(i)}{\psi_{aa_j}\phi_{aa_j}(j)}\right)}{1-\frac{\psi_{aa_i}\phi_{aa_i}(i)}{\psi_{aa_j}\phi_{aa_j}(j)}} & \text{non synonymous transversion} \\ \pi_{j_p} . \kappa . \omega . \frac{-\log\left(\frac{\psi_{aa_i}\phi_{aa_i}(i)}{\psi_{aa_j}\phi_{aa_j}(j)}\right)}{1-\frac{\psi_{aa_i}\phi_{aa_i}(i)}{\psi_{aa_j}\phi_{aa_j}(j)}} & \text{non synonymous transition} \end{cases},$$

with π_{j_p} being the equilibrium frequency of nucleotide j_p (p indicating the position in codon j), κ the transition/transversion ratio, ω the non-synonymous/synonymous ratio, and ψ_{aa_j} a preference towards amino acid j relative to other amino acids. We then have a three-layered model, with:

– the nucleotidic layer with parameters π_{j_p} and κ,
– the codon usage layer with parameters $\phi_{aa}(i)$,
– the amino acid layer with parameters ψ_{aa_j} and ω.

This formulation is a reparametrization of the previous works of [11]; the main difference in our version of the model is the complete separation between the codon usage parameters and amino acid parameters, ensuring that pressures acting on both biological scales can be measured separately. Then, we have 67 parameters: 5 for the nucleotidic layer, 41 for the codon layer and 21 for the amino acid layer. This is a quite high number, that can however be dealt with, as shown in [11]. All these three layers can be infered jointly from the data, and given a set of parameters, one can compute the equilibrium frequencies of all codons (and then nucleotides or amino acids) under the effect of all three of them, or only one by one. Computationally, this model has been implemented in the Bio++ suite [5].

The model was applied to a dataset composed of 3353 orthologuous genes in 3 closely related strains of *E. coli*, coming from [9], in order to study the relative importance of mutational biases and codon preferences. As a single gene alignment does not contain enough information to estimate all parameters of the model, we built gene concatenates in the following way. Genes were first sorted by increasing Fop values; Fop is the fraction of optimal codons in a gene, one of the common measure of codon bias strength [8]. The gene list was segmented in 33 consecutive groups of 102 genes each (except for the last one, 89 genes), and all genes in the group were concatenated to be studied as a single sequence. By definition of the Fop, the last concatenate mostly contains ribosomal proteins and factors involved in translation, *i.e* highly expressed genes.

Another way to group genes, instead of using Fop values measuring codon usage intensity, would have been to cluster together genes sharing the same codon bias, using e.g. the method by [1]; here we preferred grouping by Fop in order to keep close to the standard literature of codon usage bias in bacteria and make comparisons easy. Descriptive statistics of our 33 concatenates can be found in TABLE 1.

In this paper, we will only focus on the analysis of the codon and nucleotidic parameters, the amino acid content evolution in these three closely related strains being negligible in first approach. The obtained parameters are reported in TABLE 1. We first verified that inferring codon usage parameters did not modify strongly the value of known parameters, such as ω. Indeed, it is known that, due to purifying selection, ω values tend to decrease in genes with strong codon usage bias (and a high Fop). We checked that this was still true in our analyses: indeed, the concatenate mean Fop and ω correlates fairly well (Spearman $\rho = -0.782$, p$< 10^{-7}$).

Then, we looked for the values obtained for codon preferences. As one can see in FIG. 1, these values are quite consistent from one concatenate to another, often giving the same trend of the entire genome: as

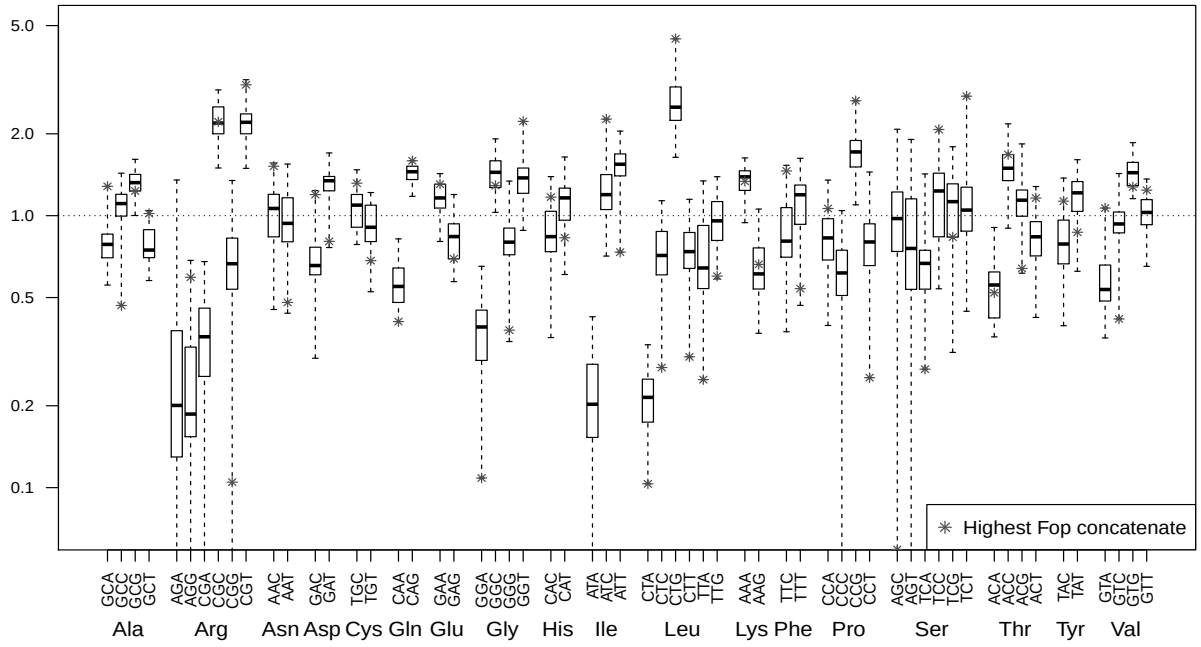

Figure 1. Codon preferences; x-axis, the codons, organized by amino acid; y-axis, the normalized preferences, ie preferences multiplied by the degeneracy of the amino acid, such that a normalized preference of 1 means uniform preferences for synonymous codons in the amino acid. The boxplot extends to the extreme values; codon preferences in the highest Fop concatenate, containing ribosomal proteins, are shown by a star.

an example, for Glutamine, codon CAG is always prefered over codon CAA. Note that this is not a trival consequence of the nucleotidic content of the amino acid, as in Lysine, coded for by AAA and AAG, the A-ending codon is almost always prefered over the G-ending one. In some cases, such as Serine, preferences are much more volatile, with no clear-cut rule. This confirms *a posteriori* one of the potentiel interests of the model, *i. e.* to be able to discriminate on which amino acids selection on codon usage is visible, and to go from a single measure of codon bias intensity, such as Fop or CAI, to a more detailed, amino-acid dependent characterization. Finally, one can see on this graph that the concatenate including genes with the highest Fop, ie. mostly ribosomal proteins, often shows a more marked preference towards a codon relative to the others, as indicated by the extreme star positions in the boxplots: see by example the Leucine or Isoleucine plots. This marked preference is in agreement with what was expected for these genes, which are used as a reference for the codon usage bias of the whole organism in measures like Fop.

Finally, we studied the GC content at equilibrium (GC*) in the 33 concatenates that would result from selection on codon usage only, or mutational bias only. In reality, both layers influence sequence evolution, and global equilibrium frequencies in this dataset are much closer to those given by the nucleotidic layer than by the codon usage layer (data not shown); but, aside from the question of the relative strength of both pressures on equilibrium frequencies – which is still under study – one can ask how these two pressures relate. As shown in FIG. 2, the codon and nucleotidic layers are counteracting each other, the mutational bias decreasing dramatically the GC content, while the codon layer tends towards a more balanced GC*. Both these facts could be independently supposed, as a global mutational bias towards AT has been documented in bacteria [7], and the codon layer, by definition, can not have a strong effect on the average nucleotide composition; but the negative correlation observed, implying that stronger mutational biases are countered by stronger selection on codon usage, was unexpected. Moreover, one can note on this graph that genes with a stronger Fop (darker points) tend to cluster above the regression line, showing that they are significantly more affected by the codon layer than the low Fop genes (Pearson correlation test between Fop and difference between GC* of the codon layer and the regression line , $R = 0.79$, $p < 2.10^{-8}$). These results implicate that selection on codon usage is higher on sequences submitted to high mutational biases; one could hypothesize that codon usage bias is then

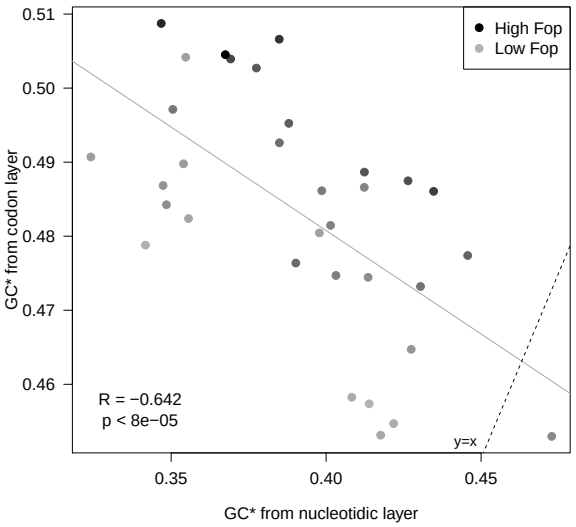

Figure 2. Comparison of GC* from the nucleotidic (x-axis) and codon (y-axis) layers. The Spearman correlation coefficient and p-value of the corresponding test are given. Note the difference in range between the two axis. The $y = x$ line is shown in dots to locate where the points should align if the GC* of both layers were the same.

a way to prevent a fast sequence degradation due to AT biased mutational processes. A next step in this study could be to analyze the set of genes for which both evolutive layers tend to close values (points on bottom right of FIG. 2), and those which are far from it (top left), and see if any qualitative differences exist between them.

Thus, we have developed a codon-based, multi-layered evolutionary model, for the study of codon usage bias, in the Bio++ suite. Our first analyses, on a simple dataset, show that our results are reliable and allow an easy intepretation of the different forces acting on sequence evolution, by example in shedding light on the antagonists effects of mutational bias and selection for codon usage in *E. coli*. These tools could be adapted and used in various phylogenetic contexts, to help untangle evolutionary effects that can hardly be distinguished based on sequence analysis alone.

References

[1] M. Bailly-Bechet, A. Danchin, M. Iqbal, M. Marsili and M. Vergassola. Codon usage domains over bacterial chromosomes. *PLoS Comp. Biol.*, 2(4):e37, 2006.

[2] M. Bulmer. Coevolution of codon usage and transfer rna abundance. *Nature*, 325(6106):728–730, 1987.

[3] H. Dong, L. Nilsson, and C. G. Kurland. Co-variation of trna abundance and codon usage in escherichia coli at different growth rates. *J Mol Biol*, 260(5):649–663, Aug 1996.

[4] J. Dutheil, S. Gaillard, E. Bazin, S. Glémin, V. Ranwez, N. Galtier, and K. Belkhir. Bio++: a set of c++ libraries for sequence analysis, phylogenetics, molecular evolution and population genetics. *BMC Bioinformatics*, 7:188, 2006.

[5] L. Guéguen, S. Gaillard, B. Boussau, M. Gouy, M. Groussin, N.C. Rochette, T. Bigot, D. Fournier, F. Pouyet, V. Cahais, A. Bernard, C. Scornavacca, B. Nabholz, A. Haudry, L. Dachary, N. Galtier, K. Belkhir, J.Y. Dutheil Bio++: efficient, extensible libraries and tools for computational molecular evolution *Mol Biol Evol*, 2013.

[6] R. Hershberg and D. A. Petrov. Selection on codon bias. *Annu Rev Genet*, 42:287–299, 2008.

[7] R. Hershberg and D. A. Petrov. Evidence that mutation is universally biased towards at in bacteria. *PLoS Genet*, 6(9), Sep 2010.

[8] T. Ikemura. Correlation between the abundance of *Escherichia coli* tRNAs and the occurence of the respective codons in its protein genes. *J Mol Biol* 146:1–21

[9] I. K. Jordan, F. A. Kondrashov, I. A. Adzhubei, Y. I. Wolf, E. V. Koonin, A. S. Kondrashov, and S. Sunyaev. A universal trend of amino acid gain and loss in protein evolution. *Nature*, 433(7026):633–638, Feb 2005.

[10] P. M. Sharp, L. R. Emery, and K. Zeng. Forces that influence the evolution of codon bias. *Philos Trans R Soc Lond B Biol Sci*, 365:1203–1212, Apr 2010.

[11] Z. Yang and R. Nielsen. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Mol Biol Evol*, 25(3):568–579, Mar 2008.

Concatenate	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Size (kb)	212.68	232.31	255.00	276.19	285.91	289.30	286.07	300.20	279.43	280.17	306.01
GC%	0.48	0.51	0.50	0.51	0.51	0.51	0.52	0.52	0.52	0.52	0.52
Identity%	0.83	0.90	0.84	0.91	0.88	0.87	0.92	0.92	0.90	0.90	0.92
Mean Fop	0.34	0.35	0.36	0.37	0.38	0.38	0.39	0.39	0.40	0.40	0.41
					Estimated Parameters						
ω	0.27	0.11	0.18	0.14	0.18	0.10	0.14	0.14	0.08	0.12	0.13
π_A	0.34	0.38	0.36	0.35	0.36	0.35	0.33	0.36	0.43	0.41	0.32
π_C	0.22	0.16	0.24	0.21	0.21	0.17	0.18	0.17	0.17	0.20	0.22
π_G	0.20	0.18	0.18	0.20	0.21	0.18	0.21	0.18	0.15	0.15	0.21
π_T	0.25	0.28	0.22	0.25	0.22	0.30	0.27	0.29	0.24	0.24	0.25
Codon preferences $\phi_{aa}(i)$											
GCA	0.21	0.18	0.17	0.19	0.18	0.20	0.21	0.17	0.17	0.16	0.20
GCC	0.27	0.32	0.22	0.23	0.28	0.30	0.29	0.32	0.31	0.25	0.26
GCG	0.38	0.30	0.40	0.31	0.31	0.32	0.33	0.36	0.35	0.34	0.32
GCT	0.14	0.21	0.21	0.26	0.23	0.18	0.16	0.15	0.16	0.24	0.22
AGA	0.07	0.03	0.03	0.06	0.04	0.06	0.06	0.03	0.03	0.01	0.06
AGG	0.08	0.05	0.04	0.08	0.07	0.04	0.03	0.02	0.03	0.03	0.05
CGA	0.08	0.07	0.06	0.09	0.06	0.06	0.10	0.08	0.08	0.04	0.07
CGC	0.34	0.43	0.41	0.30	0.33	0.40	0.36	0.48	0.33	0.33	0.33
CGG	0.18	0.16	0.10	0.13	0.12	0.15	0.12	0.13	0.19	0.22	0.14
CGT	0.25	0.26	0.36	0.35	0.38	0.29	0.33	0.26	0.34	0.36	0.34
AAC	0.24	0.42	0.23	0.31	0.32	0.52	0.57	0.62	0.43	0.47	0.39
AAT	0.76	0.58	0.77	0.69	0.68	0.48	0.43	0.38	0.57	0.53	0.61
GAC	0.20	0.32	0.15	0.24	0.20	0.37	0.30	0.49	0.31	0.30	0.31
GAT	0.80	0.68	0.85	0.76	0.80	0.63	0.70	0.51	0.69	0.70	0.69
TGC	0.40	0.39	0.44	0.43	0.50	0.45	0.62	0.57	0.60	0.65	0.44
TGT	0.60	0.61	0.56	0.57	0.50	0.55	0.38	0.43	0.40	0.35	0.56
CAA	0.37	0.38	0.32	0.37	0.36	0.26	0.41	0.30	0.26	0.23	0.35
CAG	0.63	0.62	0.68	0.63	0.64	0.74	0.59	0.70	0.74	0.77	0.65
GAA	0.48	0.40	0.52	0.53	0.61	0.54	0.54	0.52	0.47	0.42	0.63
GAG	0.52	0.60	0.48	0.47	0.39	0.46	0.46	0.48	0.53	0.58	0.37
GGA	0.15	0.15	0.11	0.15	0.11	0.11	0.11	0.09	0.07	0.07	0.11
GGC	0.36	0.28	0.42	0.26	0.32	0.38	0.43	0.40	0.33	0.27	0.31
GGG	0.27	0.33	0.18	0.22	0.20	0.18	0.19	0.21	0.30	0.34	0.18
GGT	0.22	0.24	0.29	0.37	0.37	0.32	0.27	0.30	0.30	0.33	0.40
CAC	0.27	0.25	0.18	0.37	0.25	0.52	0.30	0.46	0.44	0.37	0.37
CAT	0.73	0.75	0.82	0.63	0.75	0.48	0.70	0.54	0.56	0.63	0.63
ATA	0.12	0.09	0.06	0.14	0.10	0.12	0.13	0.08	0.07	0.06	0.06
ATC	0.24	0.40	0.26	0.32	0.30	0.37	0.32	0.45	0.40	0.33	0.36
ATT	0.65	0.51	0.68	0.54	0.60	0.52	0.54	0.47	0.53	0.60	0.58
CTA	0.03	0.05	0.02	0.04	0.03	0.04	0.04	0.03	0.02	0.03	0.05
CTC	0.10	0.16	0.08	0.13	0.09	0.18	0.12	0.17	0.19	0.10	0.08
CTG	0.28	0.36	0.28	0.27	0.30	0.34	0.42	0.44	0.38	0.42	0.39
CTT	0.14	0.17	0.18	0.16	0.19	0.19	0.13	0.12	0.12	0.13	0.12
TTA	0.22	0.12	0.22	0.22	0.18	0.10	0.11	0.09	0.09	0.09	0.19
TTG	0.23	0.15	0.22	0.18	0.21	0.15	0.19	0.16	0.19	0.23	0.17
AAA	0.58	0.58	0.60	0.57	0.71	0.68	0.64	0.65	0.52	0.47	0.70
AAG	0.42	0.42	0.40	0.43	0.29	0.32	0.36	0.35	0.48	0.53	0.30
TTC	0.26	0.36	0.19	0.24	0.23	0.36	0.35	0.47	0.40	0.32	0.30
TTT	0.74	0.64	0.81	0.76	0.77	0.64	0.65	0.53	0.60	0.68	0.70
CCA	0.17	0.16	0.14	0.21	0.17	0.20	0.24	0.14	0.10	0.16	0.20
CCC	0.17	0.26	0.16	0.19	0.22	0.24	0.17	0.26	0.22	0.18	0.17
CCG	0.37	0.35	0.38	0.32	0.27	0.39	0.42	0.46	0.45	0.43	0.35
CCT	0.29	0.23	0.33	0.29	0.34	0.17	0.16	0.14	0.23	0.23	0.28
AGC	0.16	0.27	0.19	0.12	0.20	0.17	0.22	0.14	0.12	0.16	0.14
AGT	0.22	0.32	0.27	0.17	0.30	0.14	0.12	0.13	0.11	0.19	0.09
TCA	0.19	0.10	0.11	0.12	0.09	0.15	0.12	0.12	0.12	0.09	0.15
TCC	0.12	0.09	0.10	0.23	0.10	0.22	0.19	0.24	0.27	0.12	0.16
TCG	0.16	0.15	0.22	0.19	0.13	0.17	0.21	0.21	0.22	0.27	0.28
TCT	0.15	0.07	0.12	0.17	0.18	0.16	0.12	0.16	0.17	0.18	0.18
ACA	0.21	0.17	0.14	0.15	0.16	0.21	0.11	0.17	0.10	0.10	0.16
ACC	0.26	0.35	0.22	0.34	0.32	0.27	0.43	0.36	0.33	0.24	0.32
ACG	0.34	0.31	0.36	0.28	0.32	0.31	0.29	0.36	0.31	0.46	0.33
ACT	0.18	0.17	0.27	0.24	0.21	0.21	0.17	0.11	0.26	0.20	0.19
TAC	0.20	0.36	0.20	0.26	0.22	0.31	0.32	0.53	0.32	0.37	0.34
TAT	0.80	0.64	0.80	0.74	0.78	0.69	0.68	0.47	0.68	0.63	0.66
GTA	0.13	0.10	0.10	0.13	0.14	0.13	0.16	0.11	0.09	0.09	0.17
GTC	0.22	0.36	0.21	0.26	0.23	0.25	0.25	0.26	0.27	0.23	0.22
GTG	0.39	0.29	0.37	0.29	0.33	0.41	0.36	0.46	0.36	0.44	0.32
GTT	0.26	0.25	0.32	0.33	0.30	0.20	0.23	0.16	0.28	0.24	0.29

Concatenate	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**
Size (kb)	295.73	315.13	298.40	291.13	272.28	290.82	337.11	267.95	291.40	315.53	344.65
GC%	0.52	0.53	0.52	0.53	0.52	0.53	0.53	0.53	0.53	0.53	0.54
Identity%	0.94	0.93	0.93	0.91	0.92	0.94	0.94	0.94	0.91	0.93	0.94
Mean Fop	0.41	0.42	0.42	0.43	0.43	0.43	0.44	0.44	0.45	0.45	0.46
Estimated Parameters											
ω	0.10	0.08	0.13	0.11	0.11	0.09	0.09	0.05	0.04	0.07	0.06
π_A	0.40	0.40	0.32	0.28	0.33	0.33	0.34	0.30	0.33	0.31	0.34
π_C	0.18	0.19	0.20	0.24	0.19	0.20	0.21	0.19	0.17	0.23	0.19
π_G	0.17	0.16	0.21	0.23	0.22	0.20	0.19	0.21	0.18	0.20	0.19
π_T	0.26	0.25	0.26	0.25	0.26	0.27	0.25	0.30	0.31	0.26	0.27
Codon preferences $\phi_{aa}(i)$											
GCA	0.16	0.14	0.18	0.25	0.20	0.19	0.18	0.23	0.14	0.21	0.18
GCC	0.28	0.29	0.31	0.26	0.29	0.24	0.25	0.21	0.36	0.25	0.32
GCG	0.37	0.39	0.29	0.31	0.34	0.34	0.33	0.38	0.33	0.36	0.34
GCT	0.20	0.18	0.21	0.19	0.17	0.22	0.25	0.18	0.18	0.18	0.16
AGA	0.03	0.01	0.05	0.11	0.01	0.04	0.05	0.07	0.01	0.07	0.03
AGG	0.04	0.01	0.02	0.01	0.03	0.03	0.03	0.03	0.03	0.02	0.01
CGA	0.04	0.06	0.09	0.11	0.05	0.05	0.04	0.08	0.06	0.07	0.04
CGC	0.45	0.39	0.35	0.25	0.43	0.35	0.32	0.37	0.42	0.37	0.46
CGG	0.14	0.15	0.11	0.10	0.11	0.12	0.14	0.14	0.11	0.10	0.09
CGT	0.30	0.38	0.39	0.42	0.38	0.41	0.42	0.30	0.37	0.37	0.38
AAC	0.55	0.42	0.46	0.35	0.59	0.39	0.45	0.69	0.56	0.41	0.57
AAT	0.45	0.58	0.54	0.65	0.41	0.61	0.55	0.31	0.44	0.59	0.43
GAC	0.33	0.27	0.35	0.23	0.33	0.35	0.24	0.31	0.32	0.29	0.38
GAT	0.67	0.73	0.65	0.77	0.67	0.65	0.76	0.69	0.68	0.71	0.62
TGC	0.55	0.66	0.52	0.44	0.59	0.52	0.49	0.42	0.55	0.59	0.64
TGT	0.45	0.34	0.48	0.56	0.41	0.48	0.51	0.58	0.45	0.41	0.36
CAA	0.24	0.21	0.29	0.41	0.27	0.27	0.28	0.29	0.21	0.32	0.25
CAG	0.76	0.79	0.71	0.59	0.73	0.73	0.72	0.71	0.79	0.68	0.75
GAA	0.47	0.53	0.58	0.66	0.67	0.60	0.56	0.59	0.61	0.59	0.58
GAG	0.53	0.47	0.42	0.34	0.33	0.40	0.44	0.41	0.39	0.41	0.42
GGA	0.07	0.11	0.16	0.14	0.12	0.13	0.09	0.08	0.07	0.10	0.08
GGC	0.37	0.31	0.31	0.28	0.40	0.36	0.34	0.32	0.44	0.30	0.41
GGG	0.22	0.26	0.23	0.23	0.20	0.20	0.20	0.21	0.18	0.20	0.14
GGT	0.34	0.32	0.30	0.34	0.28	0.32	0.38	0.38	0.31	0.39	0.37
CAC	0.39	0.30	0.40	0.36	0.42	0.51	0.37	0.46	0.42	0.40	0.49
CAT	0.61	0.70	0.60	0.64	0.58	0.49	0.63	0.54	0.58	0.60	0.51
ATA	0.11	0.05	0.09	0.09	0.08	0.06	0.10	0.08	0.05	0.14	0.04
ATC	0.39	0.35	0.38	0.35	0.44	0.41	0.35	0.44	0.55	0.38	0.43
ATT	0.50	0.60	0.52	0.56	0.48	0.53	0.55	0.48	0.39	0.48	0.53
CTA	0.03	0.03	0.04	0.05	0.04	0.04	0.04	0.04	0.06	0.04	0.04
CTC	0.11	0.11	0.12	0.09	0.15	0.09	0.12	0.11	0.12	0.10	0.16
CTG	0.37	0.42	0.37	0.31	0.38	0.50	0.40	0.51	0.49	0.39	0.47
CTT	0.12	0.16	0.13	0.12	0.17	0.10	0.14	0.07	0.11	0.12	0.12
TTA	0.15	0.10	0.16	0.21	0.11	0.10	0.11	0.12	0.09	0.20	0.08
TTG	0.20	0.18	0.17	0.21	0.15	0.17	0.19	0.14	0.13	0.14	0.14
AAA	0.62	0.59	0.75	0.73	0.70	0.73	0.70	0.79	0.66	0.76	0.75
AAG	0.38	0.41	0.25	0.27	0.30	0.27	0.30	0.21	0.34	0.24	0.25
TTC	0.35	0.36	0.42	0.32	0.40	0.45	0.36	0.51	0.54	0.43	0.40
TTT	0.65	0.64	0.58	0.68	0.60	0.55	0.64	0.49	0.46	0.57	0.60
CCA	0.12	0.15	0.24	0.23	0.21	0.23	0.19	0.24	0.24	0.24	0.21
CCC	0.18	0.19	0.15	0.15	0.14	0.14	0.11	0.14	0.23	0.16	0.22
CCG	0.42	0.49	0.43	0.40	0.48	0.42	0.46	0.45	0.33	0.40	0.37
CCT	0.28	0.17	0.17	0.21	0.16	0.21	0.23	0.17	0.19	0.20	0.20
AGC	0.24	0.11	0.20	0.16	0.12	0.15	0.35	0.20	0.32	0.12	0.26
AGT	0.22	0.09	0.19	0.21	0.09	0.13	0.31	0.06	0.15	0.10	0.15
TCA	0.09	0.09	0.09	0.18	0.10	0.12	0.06	0.15	0.08	0.12	0.08
TCC	0.15	0.23	0.14	0.12	0.13	0.15	0.09	0.21	0.23	0.23	0.24
TCG	0.18	0.21	0.18	0.16	0.30	0.22	0.11	0.22	0.10	0.23	0.12
TCT	0.12	0.26	0.20	0.18	0.26	0.24	0.09	0.16	0.12	0.20	0.15
ACA	0.10	0.10	0.12	0.23	0.13	0.15	0.14	0.11	0.13	0.16	0.16
ACC	0.45	0.34	0.39	0.28	0.47	0.35	0.37	0.38	0.39	0.34	0.46
ACG	0.30	0.29	0.24	0.25	0.26	0.28	0.30	0.30	0.31	0.28	0.25
ACT	0.16	0.28	0.25	0.25	0.14	0.21	0.18	0.21	0.18	0.22	0.14
TAC	0.33	0.35	0.43	0.31	0.38	0.42	0.41	0.39	0.48	0.39	0.44
TAT	0.67	0.65	0.57	0.69	0.62	0.58	0.59	0.61	0.52	0.61	0.56
GTA	0.14	0.10	0.15	0.13	0.10	0.15	0.12	0.17	0.13	0.17	0.12
GTC	0.21	0.29	0.21	0.24	0.29	0.24	0.24	0.23	0.26	0.22	0.26
GTG	0.39	0.33	0.30	0.35	0.32	0.39	0.39	0.41	0.39	0.37	0.32
GTT	0.26	0.28	0.34	0.28	0.29	0.22	0.26	0.19	0.23	0.24	0.30

Concatenate	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**
Size (kb)	308.05	321.78	341.66	310.31	333.66	370.78	326.68	324.28	329.25	339.14	268.67
GC%	0.53	0.53	0.54	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.52
Identity%	0.94	0.93	0.93	0.94	0.95	0.96	0.96	0.96	0.94	0.95	0.97
Mean Fop	0.46	0.47	0.48	0.49	0.50	0.51	0.53	0.54	0.57	0.60	0.68
Estimated Parameters											
ω	0.05	0.12	0.05	0.04	0.04	0.04	0.03	0.03	0.03	0.02	0.03
π_A	0.36	0.32	0.33	0.33	0.26	0.31	0.33	0.28	0.26	0.28	0.29
π_C	0.20	0.24	0.20	0.18	0.19	0.23	0.19	0.21	0.18	0.16	0.21
π_G	0.19	0.21	0.19	0.20	0.23	0.20	0.18	0.22	0.21	0.19	0.15
π_T	0.25	0.24	0.28	0.29	0.33	0.26	0.30	0.29	0.35	0.37	0.34
Codon preferences $\phi_{aa}(i)$											
GCA	0.20	0.22	0.21	0.18	0.25	0.20	0.17	0.26	0.26	0.25	0.32
GCC	0.32	0.26	0.28	0.29	0.29	0.25	0.31	0.27	0.24	0.24	0.12
GCG	0.25	0.33	0.34	0.35	0.31	0.34	0.33	0.27	0.27	0.32	0.31
GCT	0.23	0.19	0.17	0.18	0.15	0.22	0.19	0.20	0.23	0.20	0.25
AGA	0.03	0.02	0.02	0.08	0.23	0.02	0.01	0.10	0.02	0.18	0.00
AGG	0.03	0.01	0.01	0.06	0.06	0.11	0.03	0.03	0.11	0.03	0.10
CGA	0.09	0.05	0.04	0.06	0.07	0.05	0.04	0.03	0.02	0.03	0.01
CGC	0.34	0.33	0.40	0.37	0.33	0.28	0.47	0.35	0.48	0.42	0.37
CGG	0.11	0.07	0.11	0.09	0.04	0.06	0.08	0.01	0.00	0.02	0.02
CGT	0.39	0.53	0.42	0.34	0.27	0.48	0.36	0.48	0.37	0.33	0.50
AAC	0.49	0.53	0.54	0.69	0.60	0.61	0.60	0.63	0.74	0.78	0.76
AAT	0.51	0.47	0.46	0.31	0.40	0.39	0.40	0.37	0.26	0.22	0.24
GAC	0.31	0.34	0.45	0.48	0.49	0.38	0.48	0.38	0.51	0.62	0.60
GAT	0.69	0.66	0.55	0.52	0.51	0.62	0.52	0.62	0.49	0.38	0.40
TGC	0.45	0.54	0.55	0.57	0.50	0.60	0.65	0.69	0.55	0.74	0.66
TGT	0.55	0.46	0.45	0.43	0.50	0.40	0.35	0.31	0.45	0.26	0.34
CAA	0.29	0.24	0.25	0.23	0.35	0.26	0.21	0.23	0.23	0.32	0.20
CAG	0.71	0.76	0.75	0.77	0.65	0.74	0.79	0.77	0.77	0.68	0.80
GAA	0.57	0.59	0.56	0.56	0.71	0.65	0.70	0.71	0.70	0.68	0.65
GAG	0.43	0.41	0.44	0.44	0.29	0.35	0.30	0.29	0.30	0.32	0.35
GGA	0.08	0.09	0.12	0.09	0.10	0.06	0.05	0.10	0.06	0.04	0.03
GGC	0.32	0.37	0.35	0.38	0.39	0.37	0.42	0.38	0.42	0.48	0.32
GGG	0.23	0.19	0.20	0.25	0.13	0.18	0.16	0.16	0.14	0.09	0.09
GGT	0.37	0.36	0.33	0.28	0.37	0.39	0.37	0.36	0.37	0.39	0.56
CAC	0.42	0.33	0.53	0.54	0.57	0.44	0.66	0.65	0.70	0.68	0.59
CAT	0.58	0.67	0.47	0.46	0.43	0.56	0.34	0.35	0.30	0.32	0.41
ATA	0.07	0.03	0.05	0.05	0.09	0.05	0.03	0.05	0.05	0.01	0.00
ATC	0.32	0.36	0.47	0.49	0.53	0.52	0.52	0.47	0.68	0.68	0.76
ATT	0.62	0.61	0.47	0.46	0.38	0.43	0.45	0.47	0.27	0.31	0.24
CTA	0.02	0.02	0.03	0.04	0.04	0.03	0.04	0.03	0.04	0.03	0.02
CTC	0.11	0.07	0.15	0.19	0.12	0.12	0.16	0.12	0.13	0.09	0.05
CTG	0.40	0.47	0.50	0.49	0.56	0.50	0.50	0.43	0.58	0.66	0.74
CTT	0.13	0.11	0.11	0.09	0.07	0.10	0.11	0.16	0.08	0.05	0.05
TTA	0.11	0.15	0.08	0.07	0.10	0.09	0.08	0.13	0.06	0.07	0.04
TTG	0.22	0.18	0.14	0.12	0.11	0.16	0.12	0.13	0.10	0.10	0.10
AAA	0.70	0.61	0.68	0.78	0.81	0.69	0.69	0.76	0.82	0.73	0.67
AAG	0.30	0.39	0.32	0.22	0.19	0.31	0.31	0.24	0.18	0.27	0.33
TTC	0.41	0.38	0.54	0.52	0.56	0.55	0.56	0.60	0.75	0.77	0.73
TTT	0.59	0.62	0.46	0.48	0.44	0.45	0.44	0.40	0.25	0.23	0.27
CCA	0.17	0.25	0.18	0.22	0.28	0.30	0.16	0.27	0.31	0.34	0.27
CCC	0.11	0.10	0.14	0.13	0.13	0.07	0.13	0.05	0.06	0.03	0.01
CCG	0.36	0.48	0.38	0.46	0.45	0.49	0.50	0.59	0.56	0.47	0.66
CCT	0.36	0.16	0.29	0.19	0.14	0.14	0.21	0.08	0.07	0.16	0.06
AGC	0.07	0.15	0.16	0.17	0.32	0.09	0.33	0.12	0.18	0.19	0.01
AGT	0.07	0.10	0.09	0.08	0.14	0.07	0.15	0.05	0.10	0.04	0.00
TCA	0.12	0.11	0.12	0.14	0.08	0.24	0.08	0.14	0.10	0.10	0.05
TCC	0.31	0.16	0.27	0.21	0.21	0.17	0.20	0.24	0.32	0.31	0.35
TCG	0.20	0.22	0.19	0.23	0.15	0.22	0.12	0.14	0.11	0.05	0.14
TCT	0.24	0.25	0.17	0.17	0.11	0.21	0.12	0.30	0.20	0.31	0.46
ACA	0.15	0.14	0.14	0.10	0.14	0.10	0.11	0.13	0.09	0.09	0.13
ACC	0.46	0.40	0.36	0.36	0.49	0.42	0.41	0.39	0.54	0.45	0.42
ACG	0.16	0.25	0.30	0.29	0.19	0.26	0.27	0.15	0.17	0.24	0.16
ACT	0.23	0.21	0.21	0.25	0.17	0.22	0.21	0.32	0.20	0.22	0.29
TAC	0.55	0.35	0.48	0.62	0.60	0.46	0.50	0.48	0.50	0.69	0.57
TAT	0.45	0.65	0.52	0.38	0.40	0.54	0.50	0.52	0.50	0.31	0.43
GTA	0.13	0.15	0.16	0.13	0.24	0.16	0.12	0.22	0.18	0.20	0.27
GTC	0.26	0.17	0.21	0.31	0.23	0.17	0.27	0.19	0.23	0.22	0.10
GTG	0.35	0.40	0.44	0.39	0.33	0.42	0.31	0.31	0.34	0.31	0.32
GTT	0.25	0.27	0.19	0.17	0.21	0.25	0.29	0.27	0.25	0.27	0.31

Table 1. Table showing descriptive statistics of the gene concatenates along with the value of estimated parameters of the model. Identity% is the fraction of exactly identical positions in the alignments. Codons are organized by synonymous groups for easier reading.