Supporting Information

Vetsigian and Goldenfeld 10.1073/pnas.0810122106

SI Text

Analytic Solution of the Model. Restricting ourselves to gradual changes of $\{c\}$, the condition for adaptor pool optimality is $\partial_{c_i} f(c, u) = 0$, which leads to

$$\frac{u_i}{c_i^2 \alpha_i} = \frac{T}{L}(-G'/G), \qquad [s1]$$

generalizing a scaling relationship between letter usage and adaptor concentrations noted for identical α_i (1) and tested in *Escherichia coli* (2).

The equilibrium n_gs correspond to the eigenvector with largest eigenvalue of the matrix $\bar{M}_{gg'} f_{g'}$ (3). In the large-genome limit, the synthesis rate can be expressed as a product of independent fitness contributions for the different sites

$$\frac{1}{T} \propto \prod_s \prod_i \{e^{-c_i^{-1}/T}\}^{L_s u_{si}} \equiv \prod_s \prod_i F_i^{L_s u_{si}}, \qquad [s2]$$

and the letter usage relaxation problem reduces to finding the eigenvector u_{si} corresponding to the largest eigenvalue of the matrix $W_{ij}^{(s)} = M_{ij} F_j(T) R_{sj}$ for each site type s, self-consistently with T.

We now solve the one site type, two synonymous letter case $T = \Theta + L(u_1/c_1 + u_2/c_2)$, $R_{11} = R_{12} = 1$. We derive the adaptor bias, $\psi_c = c_2/c_1$ as a function of $\mu \equiv M_{21}$, L, Θ, the mutational bias $m = (M_{21} - M_{12})/M_{21}$, and the adaptor cost bias $\alpha = \alpha_2/\alpha_1$. The letter usage bias is then simply $\psi_u \equiv u_2/u_1 = \alpha \psi_c^2$.

Combining the optimality condition with $u_1 + u_2 = 1$, and assuming $G = \exp(-\Sigma \alpha_i c_i)$ for algebraic convenience, we obtain

$$F(\psi_c) = \frac{1 - \psi_c}{2\psi_c \theta_1}(1 + \alpha\psi_c)\left\{\sqrt{1 + \frac{4(1 + \alpha\psi_c^2)\theta_1}{(1 + \alpha\psi_c)^2}} - 1\right\}. \qquad [s3]$$

where $F(\psi_c) \equiv -L \ln F_2/F_1$ and $\theta_1 = \Theta/(L\alpha_1)$. F_2/F_1 is the relative fitness of the two letters. For $\Theta = 0$, the result is the $\theta_1 \to 0$ limit of the above but is valid for any decreasing G. (The expression for $F(\psi_c)$ is also valid for two-letter models with many site types.)

Finding the eigenvalues in the limit $\mu \to 0, L \to \infty$ and solving for μL, we obtain

$$\mu L = F(\psi_c) \frac{\alpha\psi_c^2}{1 + \alpha\psi_c^2} \frac{1}{1 + \alpha\psi_c^2(m - 1)}. \qquad [s4]$$

Combining Eq. **s4** and Eq. **s3**, we get the equilibrium solutions $\mu L(\psi_c)$ for any Θ, α, and m. For values of μL for which we have three ψ_c, the middle solution is always unstable, as shown by the red line on Fig. 5. The maximum μL for which bistability is possible (the transition point) is determined from $d(\mu L)/d\psi_c = 0$ as a function of m and α.

For $\Theta = 0$, $\alpha = 1$, $m = 0$ the expression simplifies to $\mu L = \psi_c/(1 + \psi_c)^2$. The symmetric solution $\psi_c = 1$ is stable above the transition point $\mu L = 1/4$ (Fig. 5, blue line). For extreme mutational bias, $m \to 1$, and $\Theta = 0$, the transition point approaches from above $(\mu L)^*_{min} = (2 + \alpha - 2\sqrt{1 + \alpha})/\alpha$. Thus, for $\alpha = 1$, it shifts only from 1/4 to $3 - 2\sqrt{2} \approx 0.17$.

Consistency Condition for the Two-Stranded Model. Selection on the replication speed for the two-stranded DNA model can be modeled with

$$T = \Theta_0 + \max\left(\Delta_{LAG} + \sum_i \frac{U_i}{c_i}, \sum_i \frac{U_i}{c_{\bar{i}}}\right), \qquad [s5]$$

where $\Delta_{LAG} > 0$ is the additional waiting time for lagging strand replication (coming from primer synthesis, etc.), U_i is the usage on the lagging strand, and $\bar{i}$ is the Watson–Crick complement of a letter i. This model reduces to the generic single-template model with $\Theta = \Theta_0 + \Delta_{LAG}$ if the following consistency condition is satisfied

$$\Theta > \Delta_{LAG} > \sum_i U_i\left(\frac{1}{c_{\bar{i}}} - \frac{1}{c_i}\right).$$

Dynamic Mutation Model. Let $M_{ji}^{(0)}$ be the probability that nucleotide j is accepted, once it has arrived at the active site of the polymerase, given a template letter i. Then, the mutation rate coming from replication is

$$M_{ji} = M_{ji}^{(0)} c_j / \sum_k M_{ki}^{(0)} c_k. \qquad [s6]$$

The matrix $M^{(0)}$ is a molecular property of the replication machinery, whereas M is an observable property of the cell.

The equilibrium behavior of this model as a function of $M^{(0)}$ is generally very similar to that of the static model as a function of M, except for the neutral and symmetric case $M_{ji}^{(0)} = (1 - \varepsilon)\delta_{ij} + \varepsilon(1 - \delta_{ij})$. In this regime, the solution is equivalent to that of the static model, if we replace μL with $\varepsilon^2 L$, leading to $\sqrt{L}$ scaling, rather than independence of L, of the transition point expressed as mutations per genome per generation. This scaling disappears in the presence of asymmetries or nonneutral sites, but the transition is typically shifted to the right, extending the region of bistability, as shown by the black curves in Fig. S2.

A statistical signature of selection is the deviation of $D \equiv (M_{21}u_1)/(M_{12}u_2)$ from unity, the value expected for purely mutational equilibrium. We plot this quantity for the static and dynamic mutational models in the symmetric case, and in the presence of 1% fixed sites (half of them belonging to 1, and the other half to 2). Whereas in the static model the statistical signature of selection rises sharply immediately below the transition (Fig. S2, solid red), in the dynamic model it stays close to unity well into the symmetry broken phase (Fig. S2, dashed red).

1. Kurland C, Ehrenberg M (1987) Growth-optimizing accuracy of gene expression. *Annu Rev Biophys Biophys Chem* 16:291–317.
2. Berg O, Kurland C (1997) Growth rate-optimised tRNA abundance and codon usage. *J Mol Biol* 270:544–550.
3. Sella G, Ardell D (2002) The impact of message mutation on the fitness of a genetic code. *J Mol Evol* 54:638–651.

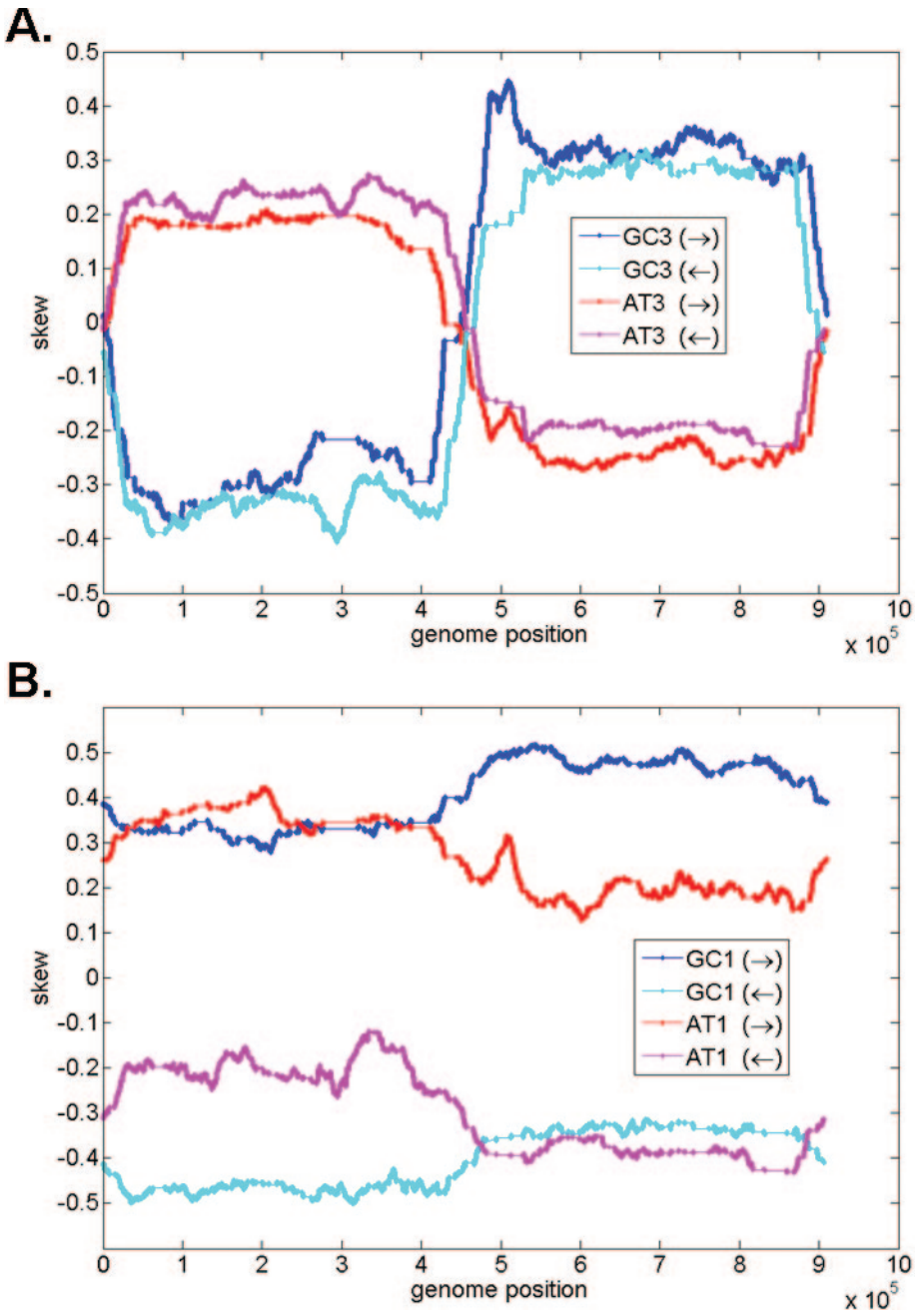

Fig. S1. Third- and first-codon position skews of *Borrelia burgdorferi* obtained by sliding a 10-kb window. To separate replication-related bias from codon usage and gene orientation bias, we compare the nucleotide skews of genes encoded to the left with those encoded to the right. (*A*) Third-codon position skews are almost the same for the two gene orientations (blue curve is close to light blue curve, and red curve is close to magenta curve), indicating that strand rather than codon bias is the primary factor shaping nucleotide usage at 3rd-codon positions. Skews switch sign at the origin of replication located near the middle of the linear chromosome. (*B*) First-codon position skew curves are offset vertically for the two gene orientations, indicating that functional constraints (e.g., typical amino acid usage) contribute significantly to nucleotide usage. Strand bias has a contribution as well, as seen from the steps in the curves near the middle of the genomes.

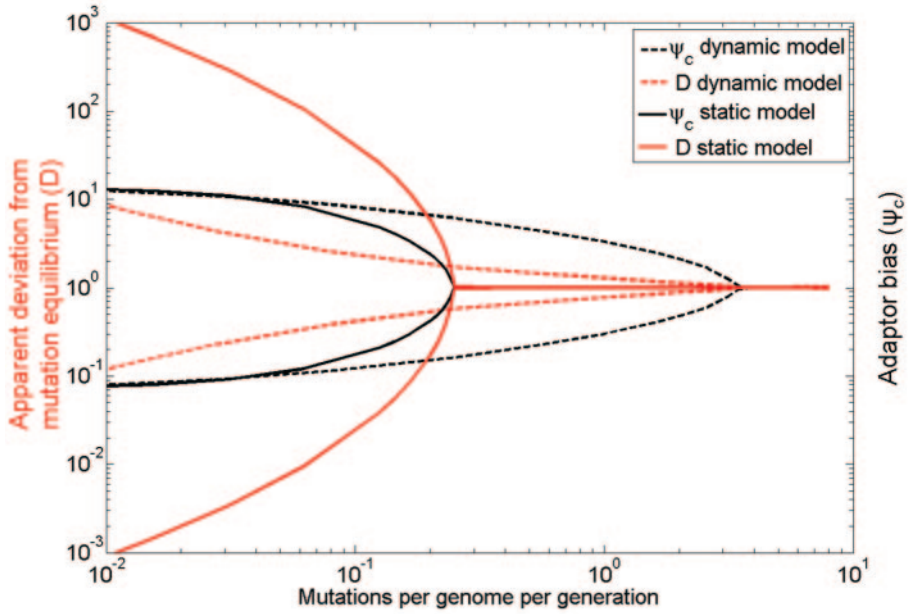

Fig. S2. Apparent deviations from mutation equilibrium for the static and dynamic mutation models. The horizontal axis is $LM_{21}u_1 + LM_{12}u_2$, which is the expected number of mutations per genome per generation. The vertical axis is the adaptor bias ψ_c for the black lines and the deviation from mutation equilibrium $D \equiv (M_{21}u_1)/(M_{12}u_2)$ for the red lines. The solid and dashed curves correspond to the static and dynamic mutation models, respectively. Whereas for the static model the apparent deviation from mutation equilibrium (a statistical signature of selection) rises sharply immediately below the transition (solid red curve), for the dynamic model it stays close to unity well into the symmetry broken phase (dashed red curve). Thus, the selection on efficiency responsible for the emergence of the bias is masked in the dynamic mutation model. In addition, for the dynamic mutation model the transition is typically shifted to the right, extending the region of bistability (solid and dashed black curves). For both the static and dynamic mutation models we use $\alpha = 1$, $\Theta = 0$, 1% of fixed sites, and symmetric M and $M^{(0)}$ correspondingly.