Supporting Information

Greenbaum et al. 10.1073/pnas.1402285111

SI Text

I. Entropy of Sequences in the Absence of Selective Force

In the absence of selective force, our model for random codon sequences is very simple. Consider a sequence of L amino acids $A=\{a_1,a_2,\ldots,a_L\}$. The probability of the ith codon, c_i, in the associated nucleotidic sequence is given by $p_i(c_i)=p(c_i|a_i)$, where $p(c|a)$ is the (Human or segment) codon bias. The probability of the sequence $C=\{c_1,c_2,\ldots,c_L\}$ is simply the product of the probabilities of its codons c_i. We readily compute the entropy σ_0 of sequences with this model:

$$\begin{aligned}\sigma_0 &= -\sum_{i=1}^{L}\sum_{c_i} p_i(c_i)\log p_i(c_i)\\ &= \sum_a N_a\left(-\sum_c p(c|a)\log p(c|a)\right)\end{aligned} \qquad \textbf{[S1]}$$

where N_a is the number of occurrences of amino acid a in A. Note that, by definition, σ_0 coincides with the average entropy $\sigma_{av}(x_s=0)$, and is the height of the maximum of the entropy curve $\sigma_{av}(x_s)$.

A simple upper bound to σ_0 is $\sigma_0^{upper}=L\cdot\log 6$, as amino acids are at most sixfold degenerate. A slightly more complicated upper bound would maximize the entropy expression for σ_0 for the same amino acid sequence but with a codon bias where all codon for a given amino acid are equiprobable. In that case it is straightforward to show that $\sigma_0^{upper}=\sum_a N_a\log(\deg(a))$, where, as before, $\deg(a)$, is the degeneracy of amino acid a. For instance, for the influenza B isolate analyzed in Fig. 1, the real value maximum entropy is 4,342.6. The upper bound for this sequence is 7,869.4 by the first method and 4,913.3 by the second.

II. Transfer Matrix Method

We calculate the normalization constant $Z(x_s)$, Eq. **2**, using the transfer matrix formalism. We call K the number of nucleotides in motif m: $m=\{m_1,m_2,\ldots,m_K\}$. Let $C=\{c_1,c_2,\ldots,c_L\}$ be a sequence of L codons; equivalently, C can be seen as a sequence of $3\times L$ nucleotides. Let $c_{i,\ell}$ denote the ℓth nucleotide in codon i, with $\ell=1,2,3$. We denote by $C[n:n+K-1]$ the subsequence of K nucleotides in C, starting at position n and ending up at position $n+K-1$. The number of occurrences of the motif m in C can be written as the following sum:

$$N_m(C)=\sum_{n=1}^{3L-K+1}\delta\big(C[n:n+K-1],m\big) \qquad \textbf{[S2]}$$

where δ denotes the Kronecker function: $\delta(X,X)=1$, and $\delta(X,Y)=0$ if $X\neq Y$.

The subsequence $C[n:n+K-1]$ spreads over at most $K_c=\mathrm{Int}((K+1)/3)+1$ contiguous codons c_i in C, where Int denotes the integer part. Consider for instance the case of dinucleotide motifs m, for which $K=2$ and $K_c=2$ according to the formula above. The two nucleotides of such a motif can indeed be found

- at positions 1,2 of a single codon, say, c_i; then we have $m_1=c_{i,1}$, $m_2=c_{i,2}$.
- at positions 2,3 of codon c_i; then we have $m_1=c_{i,2}$, $m_2=c_{i,3}$.
- at position 3 of codon c_i and position 1 of codon c_{i+1}; then we have $m_1=c_{i,3}$, $m_2=c_{i+1,1}$.

For the sake of simplicity, we start by assuming that $K=2$; the case of longer motifs will be briefly discussed later on. According to the discussion above we can write

$$N_m(C)=\sum_{i=1}^{L-1}F(m,c_i,c_{i+1}), \qquad \textbf{[S3]}$$

where

$$\begin{aligned}F(m,c_i,c_{i+1})&=\delta(m_1,c_{i,1})\delta(m_2,c_{i,2})+\delta(m_1,c_{i,2})\delta(m_2,c_{i,3})\\&\quad+\delta(m_1,c_{i,3})\delta(m_2,c_{i+1,1})\end{aligned} \qquad \textbf{[S4]}$$

for all $i=1,\ldots,L-2$ and

$$\begin{aligned}F(m,c_{L-1},c_L)&=\delta(m_1,c_{L-1,1})\delta(m_2,c_{L-1,2})\\&\quad+\delta(m_1,c_{L-1,2})\delta(m_2,c_{L-1,3})\\&\quad+\delta(m_1,c_{L-1,3})\delta(m_2,c_{L,1})\\&\quad+\delta(m_1,c_{L,1})\delta(m_2,c_{L,2})\\&\quad+\delta(m_1,c_{L,2})\delta(m_2,c_{L,3}).\end{aligned} \qquad \textbf{[S5]}$$

The expression for F in the bulk of the sequence $(i\leq L-1)$ avoids double counting of the motif occurrences.

We now rewrite $Z(x_s)$ as a sum over the possible codons corresponding to the same amino acids as in the viral sequence C_0:

$$Z(x_s)=\sum_C\left(\prod_{i=1}^{L}p_i(c_i)\right)\exp\left[x_s\sum_{i=1}^{L-1}F(m,c_i,c_{i+1})\right] \qquad \textbf{[S6]}$$

$$=\sum_C\prod_{i=1}^{L-1}\big(p_i(c_i)\ \exp\left[x_s\ F(m,c_i,c_{i+1})\right]\big)\ p_L(c_L), \qquad \textbf{[S7]}$$

where $p_i(c_i)$ is the codon bias for codon c_i (synonymous to the ith codon of sequence C_0). Let us now define L transfer matrices $\mathbf{M_i}$, $i=1,\ldots,L$. The dimension of matrix $\mathbf{M_i}$ is $\deg(a_i)\times\deg(a_{i+1})$, where $\deg(a)$ is the codon degeneracy for amino acid a. The entries of $\mathbf{M_i}$ are given by, for all $i=1,\ldots,L-2$,

$$M_i(c_i,c_{i+1})=p_i(c_i)\exp\left[x_s\ F(m,c_i,c_{i+1})\right], \qquad \textbf{[S8]}$$

and

$$M_{L-1}(c_{L-1},c_L)=p_i(c_{L-1})\exp\left[x_s\ F(m,c_{L-1},c_L)\right]p(c_L). \qquad \textbf{[S9]}$$

Then, we observe that

$$\begin{aligned}Z(x_s)&=\sum_{c_1,c_2,\ldots,c_{L-2},c_{L-1}}M_1(c_1,c_2)M_2(c_2,c_3)\ldots M_{L-2}(c_{L-2},c_{L-1})\\&\quad\times M_{L-1}(c_{L-1},c_L)\\&=\sum_{c_1,c_L}(M_1\times M_2\times\ldots\times M_{L-2}\times M_{L-1})(c_1,c_L),\end{aligned} \qquad \textbf{[S10]}$$

where × denotes the matrix product in the formula above. This formula shows that Z can be computed in a time growing linearly with L only. This is huge gain compared with the original expression of Z, Eq. **2**, which sums up an exponentially large-in-L number of codon configurations.

In practice we define the deg(a_L)-dimensional vector $\mathbf{v_L}$, with entries $v_L(c_L) = 1$ for all codons c_L coding for amino acid a_L. Then we compute the vector

$$v_{L-1}(c_{L-1}) = \sum_{c_L} M_{L-1}(c_{L-1}, c_L) v_L(c_L). \qquad \textbf{[S11]}$$

Then, we sum over all possible values for the $(L-1)$th codon, c_{L-1}:

$$v_{L-2}(c_{L-2}) = \sum_{c_{L-1}} M_{L-2}(c_{L-2}, c_{L-1}) v_{L-1}(c_{L-1}). \qquad \textbf{[S12]}$$

The process is iterated until the first codon

$$v_1(c_1) = \sum_{c_2} M_1(c_1, c_2) v_2(c_2). \qquad \textbf{[S13]}$$

Finally, we obtain the value of the normalization constant through

$$Z(x_s) = \sum_{c_1} v_1(c_1). \qquad \textbf{[S14]}$$

For large values it is easier, and often practically necessary, to work with the logarithm of the partition function, rather than with the partition function itself.

When the motif is of longer length, and overlap with K_c contiguous codons, expression **S3** has to be modified. In general one can write

$$N_m(C) = \sum_{i=1}^{L-1} F(m, c_i, c_{i+1}, \ldots, c_{i+K_c-1}), \qquad \textbf{[S15]}$$

where function F is an obvious extension of **[S4]** and **[S5]**. The transfer matrix method exposed above can still be used, but at a price of introducing larger transfer matrices $\mathbf{M_i}$.

III. Numerical Computation of the Legendre Transform

An important problem is to find the value of the selective force x_s, corresponding to the number $N_m(C_0)$ of occurrences of the motif m in the virus sequence C_0. Let us call $x_s(C_0)$ this force. One way to find $x_s(C_0)$ is to compute the average number of occurrences, $N_{av}(x_s)$, for many values of x_s on a grid and try to be as close as possible to the data, i.e., choose x_s such that $N_{av}(x_s) \simeq N_m(C_0)$. A much faster procedure is the following.

Consider the function (for a given C_0)

$$G(x_s) = \log Z(x_s) - x_s N_m(C_0). \qquad \textbf{[S16]}$$

Two important facts about G are

- the first derivative of G vanishes when x_s takes the value $x_s(C_0)$ we are looking for, because

$$\frac{d}{dx_s} G(x_s) = N_{av}(x_s) - N_m(C_0) \qquad \textbf{[S17]}$$

- G is a convex function of x_s, as its second derivative is positive:

$$\frac{d^2}{dx_s^2} G(x_s) = \frac{d}{dx_s} N_{av}(x_s) = \sum_C P(C|x_s) N_m(C)^2 - \left(\sum_C P(C|x_s) N_m(C)\right)^2$$
$$= \sum_C P(C|x_s)(N_m(C) - N_{av}(x_s))^2 \geq 0. \qquad \textbf{[S18]}$$

Hence, G has a single minimum in $x_s = x_s(C_0)$, and we can find it very quickly with standard optimization techniques, e.g., the Newton–Raphson algorithm. The procedure is here below.

i) Start with $x_s = 0$.

ii) Compute the first and second derivatives of G in x_s, that is, respectively $d_1 = N_{av}(x_s) - N_m(C_0)$ and $d_2 = \sum_C P(C|x_s) N_m(C)^2 - N_{av}(x_s)^2$.

iii) Compute the new value of x_s [which would be equal to $x_s(C_0)$ if G were a parabolic function]

$$x_s \to x_s - \frac{d_1}{d_2}. \qquad \textbf{[S19]}$$

iv) Iterate step *ii* until convergence is achieved.

As the parabolic approximation is generally good, the procedure generally converge very fast, in a few iterations.

IV. Illustrations on Very Short Sequences of Amino Acids

We illustrate the notion of entropy on two simple ad hoc sequences with $L = 2$ amino acids, Pro-Pro and Pro-Cys, and one sequence with $L = 3$ amino acids, Pro-Pro-Cys. For all three sequences the motif considered is m = CT.

A. Case of Pro-Pro. Proline is a fourfold degenerate amino acid, corresponding to codons c = CCA, CCC, CCG, CCT. For the sake of simplicity we assume that each codon has probability $1/4$. The entropy of the random codon model in the absence of force is $\sigma_0 = \log 16 = 4 \log 2$. The transfer matrix $\mathbf{M_1}$ is given by **[S9]**, with the result

$$\mathbf{M_1} = \frac{1}{16} \begin{pmatrix} 1 & 1 & 1 & e^{x_s} \\ 1 & 1 & 1 & e^{x_s} \\ 1 & 1 & 1 & e^{x_s} \\ e^{x_s} & e^{x_s} & e^{x_s} & e^{2x_s} \end{pmatrix}. \qquad \textbf{[S20]}$$

The normalization constant is (refer to **[S10]**),

$$Z(x_s) = \sum_{c_1, c_2} M_1(c_1, c_2) = \frac{1}{16}\left(9 + 6e^{xs} + e^{2xs}\right) = \frac{1}{16}(3 + e^{x_s})^2. \qquad \textbf{[S21]}$$

The average number of motifs and the entropy of sequences are therefore given by

$$N_{av}(x_s) = \frac{d}{dx_s} \log Z(x_s) = \frac{2e^{x_s}}{3 + e^{x_s}}$$
$$\sigma_{av}(x_s) = \sigma_0 + \log Z(x_s) - x_s N_{av}(x_s) = 2\log(3 + e^{x_s}) - \frac{2x_s e^{x_s}}{3 + e^{x_s}}. \qquad \textbf{[S22]}$$

In Fig. S1 we plot the entropy σ_{av} vs. the number N_{av} of occurrences of CT. The maximum of the entropy, $\sigma_{av} = 4 \log 2$, always corresponds to the unconstrained case $x_s = 0$ (there are indeed $e^{4\log 2} = 16$ possible nucleotidic sequences); the corresponding average number of occurrences of the motif $m = \mathrm{CT}$ is 0.5 as expected, as each one of the two codons can contain CT with probability $1/4$.

By varying the parameter x_s, equal to minus the slope of σ_{av} as function of N_{av}, we scan the entire entropy curve. Note that for $N_{av} = 0$, i.e., $x_s \to -\infty$, we obtain $\sigma_{av} = 2\log 3$; there are indeed $e^{2\log 3} = 9$ nucleotidic sequences compatible with Pro-Pro without CT. For $N_{av} = 2$, i.e., $x_s \to +\infty$, we obtain $\sigma_{av} = 0$; there is $e^0 = 1$ sequence compatible with Pro-Pro and including the motif twice, namely CCTCCT.

Remark that for $N_{av}=1$ we obtain $\sigma_{av}\simeq 2.472$; $e^{\sigma_{av}}$ is larger than 6, the number of sequences compatible with Pro-Pro with one CT. This is because our calculation gives the entropy of sequences that contain on average (and not exactly) N_{av} repetitions of the motif m. For large values of L we expect that N_{av} will coincide with $N_m(C)$ (up to small relative fluctuations). For extreme (minimal or maximal) values of the number of occurrences of the motifs fluctuations vanish even for small L. For instance, if the number of motifs vanishes on average then all sequences C with nonzero probability $P(C)$ must be free of the motif. This is why the entropies of sequences containing the motif exactly 0 or 2 times coincide with the outcome of our calculation.

B. Case of Pro-Cys. Cysteine is twofold degenerate, with corresponding codons TGT and TGC. The motif CT can now be found in the second and third positions of the Pro codon, or at the third position of the Pro codon and the first position of the Cys codon. We assume that there all four Pro codons are equally likely, and so are the two Cys codons. The entropy of the random codon model in the absence of force is $\sigma_0=\log 8=3\log 2$. The transfer matrix is a 4×2 matrix, given by

$$\mathbf{M_1}=\frac{1}{8}\begin{pmatrix}1 & 1\\ 1 & 1\\ e^{x_s} & e^{x_s}\\ e^{x_s} & e^{x_s}\end{pmatrix}. \tag{S23}$$

The normalization constant is (refer to **[S10]**)

$$Z(x_s)=\sum_{c_1,c_2}M_1(c_1,c_2)=\frac{1}{2}\left(1+e^{x_s}\right). \tag{S24}$$

The average number of motifs and the entropy of sequences are therefore given by

$$N_{av}(x_s)=\frac{e^{x_s}}{1+e^{x_s}},\qquad \sigma_{av}(x_s)=2\log 2+\log(1+e^{x_s})-\frac{x_s e^{x_s}}{1+e^{x_s}}. \tag{S25}$$

The entropy σ_{av} when plotted vs. the average number of motifs N_{av} is a bell-shaped curve with maximum in $\sigma_{av}=\log 8$, equal to the logarithm of the total number of nucleotidic sequences as expected. The corresponding average number of motifs is 0.5, as four sequences (CCTTGT, CCTTGC, CCCTGT, CCCTGC) contain the motif once, whereas the four remaining sequences are free of the motif. The latter four sequences are selected when $x_s\to-\infty$, corresponding to $\sigma_{av}=\log 6$ and $N_{av}=0$. Conversely, for $x_s\to+\infty$, we select the four sequences with one motif, and obtain $\sigma_{av}=\log 2$ and $N_{av}=1$.

C. Case of Pro-Pro-Cys. The entropy of sequences in now $\sigma_0=\log 32=5\log 2$ (all codons compatible with A are assumed to be equally likely). There are two transfer matrices, defined according to **[S8]** and **[S9]**:

$$\mathbf{M_1}=\frac{1}{4}\begin{pmatrix}1 & 1 & 1 & 1\\ 1 & 1 & 1 & 1\\ 1 & 1 & 1 & 1\\ e^{x_s} & e^{x_s} & e^{x_s} & e^{x_s}\end{pmatrix},\qquad \mathbf{M_2}=\frac{1}{8}\begin{pmatrix}1 & 1\\ 1 & 1\\ e^{x_s} & e^{x_s}\\ e^{x_s} & e^{x_s}\end{pmatrix}. \tag{S26}$$

Note that matrix the $\mathbf{M_1}$ above is different from its counterpart **[S20]** defined for the sequence Pro-Pro, due to the difference between F in the bulk of the sequence and at its end, compare **[S4]** and **[S5]**.

The product of the two transfer matrices is given by

$$\mathbf{M_1}\times\mathbf{M_2}=\frac{1+e^{x_s}}{16}\begin{pmatrix}1 & 1\\ 1 & 1\\ 1 & 1\\ e^{x_s} & e^{x_s}\end{pmatrix}, \tag{S27}$$

and the normalization constant is

$$Z(x_s)=\sum_{c_1,c_2}(M_1\times M_2)(c_1,c_2)=\frac{(1+e^{x_s})(3+e^{x_s})}{8}. \tag{S28}$$

The average number of motifs and the entropy of sequences are therefore given by

$$\begin{aligned}N_{av}(x_s)&=\frac{e^{x_s}}{1+e^{x_s}}+\frac{e^{x_s}}{3+e^{x_s}}\\ \sigma_{av}(x_s)&=2\log 2+\log(1+e^{x_s})+\log(3+e^{x_s})-\frac{x_s e^{x_s}}{1+e^{x_s}}-\frac{x_s e^{x_s}}{3+e^{x_s}}.\end{aligned} \tag{S29}$$

The entropy σ_{av} is plotted vs. the average number of motifs N_{av} in Fig. S2. There are $e^{\sigma_{av}(-\infty)}=12$ sequences with no copy of the motif ($N_{av}=0$): those corresponds to three codons CCA, CCC, CCG for the first Pro amino acid, the two codons CCA, CCG for the second Pro, and the two codons for Cys. We also see that there are $e^{\sigma_{av}(+\infty)}=4$ sequences with two copies of the motifs, which start with CCT followed by one of the four sequences coding for Pro-Cys with one CT listed above.

V. Case of Multiple Motifs

To calculate the entropy associated with the number of occurrences of several motifs, one can extend the preceding definitions. As an example, for two dinucleotides the partition function will vary over two parameters $(x_s^{(1)},x_s^{(2)})$ corresponding to dinucleotide motifs $m^{(1)}=(m_1^{(1)},m_2^{(1)})$ and $m^{(2)}=(m_1^{(2)},m_2^{(2)})$. The partition function naturally becomes

$$\begin{aligned}Z\left(x_s^{(1)},x_s^{(2)}\right)=\sum_C\left(\prod_{i=1}^{L}p_i(c_i)\right)\exp\Bigg[&x_s^{(1)}\sum_{i=1}^{L-1}F\left(m^{(1)},c_i,c_{i+1}\right)\\ &+x_s^{(2)}\sum_{i=1}^{L-1}F\left(m^{(2)},c_i,c_{i+1}\right)\Bigg].\end{aligned} \tag{S30}$$

This normalization constant can be calculated using the transfer matrix method as in the single motif case. The transfer matrices are defined through

$$\begin{aligned}M_i(c_i,c_{i+1})=p_i(c_i)\exp\Bigg[&x_s^{(1)}\sum_{i=1}^{L-1}F\left(m^{(1)},c_i,c_{i+1}\right)\\ &+x_s^{(2)}\sum_{i=1}^{L-1}F\left(m^{(2)},c_i,c_{i+1}\right)\Bigg],\end{aligned} \tag{S31}$$

for all $i=1,\ldots,L-2$, and

$$\begin{aligned}M_{L-1}(c_{L-1},c_L)=p_{L-1}(c_{L-1})\exp\Bigg[&x_s^{(1)}\sum_{i=1}^{L-1}F\left(m^{(1)},c_i,c_{i+1}\right)\\ &+x_s^{(2)}\sum_{i=1}^{L-1}F\left(m^{(2)},c_i,c_{i+1}\right)\Bigg]p_L(c_L).\end{aligned} \tag{S32}$$

Once Z has been calculated, we obtain the entropy through a Legendre transform with respect to the two forces $x_s^{(1)}$ and $x_s^{(2)}$:

$$\sigma_{av}\left(x_s^{(1)},x_s^{(2)}\right)=\sigma_0+\log Z\left(x_s^{(1)},x_s^{(2)}\right)-x_s^{(1)}N_{av}^{(1)}\left(x_s^{(1)},x_s^{(2)}\right) -x_s^{(2)}N_{av}^{(2)}\left(x_s^{(1)},x_s^{(2)}\right) \quad \textbf{[S33]}$$

where

$$N_{av}^{(1)}\left(x_s^{(1)},x_s^{(2)}\right)=\frac{\partial}{\partial x_s^{(1)}}\log Z\left(x_s^{(1)},x_s^{(2)}\right) \quad \textbf{[S34]}$$

and likewise for $N_{av}^{(2)}(x_s^{(1)},x_s^{(2)})$. Then

$$N_{av}^{(1)}\left(x_s^{(1)},x_s^{(2)}\right)=\frac{\partial}{\partial x_s^{(1)}}\log Z\left(x_s^{(1)},x_s^{(2)}\right), \quad \textbf{[S35]}$$

together with a similar expression for the average number of motifs $m^{(2)}$. The second derivatives of Z give access to the covariance matrix of $\mathbf{N^{(1)}}$ and $\mathbf{N^{(2)}}$.

The above formula can be straightforwardly extended to the case of more than two forces and motifs. Assume we have $K_m\geq 2$ motifs, $m^{(j)}$, with $j=1,\ldots,K_m$. Then $\mathbf{x_s}$ is a K_m dimensional vector, and so is $\mathbf{N_{av}}(\mathbf{x_s})$. In particular the entropy of sequences is now given by

$$\sigma_{av}(\mathbf{x_s})=\sigma_0+\log Z(\mathbf{x_s})-\mathbf{x_s}\cdot\mathbf{N_{av}}(\mathbf{x_s}) \quad \textbf{[S36]}$$

where $\cdot$ denotes the dot product over the K_m components of $\mathbf{x_s}$ and $\mathbf{N_{av}}$, and

$$\mathbf{N_{av}}(\mathbf{x_s})=\frac{\partial}{\partial \mathbf{x_s}}\log Z(\mathbf{x_s}). \quad \textbf{[S37]}$$

The partition function Z can be computed with the transfer matrix as in the $K_m=2$ case above. In addition, the numerical procedure of *SI Text*, section III to calculate x_s can be extended to the multidimensional case of more than one force parameter as follows. We now define G through

$$G(\mathbf{x_s})=\log Z(\mathbf{x_s})-\sum_{j=1}^{K_m}x_s^{(j)}N_{m^{(j)}}(C_0). \quad \textbf{[S38]}$$

The gradient of G in $\mathbf{x_s}$ is a K_m- dimensional vector $\mathbf{d_1}$, and its Hessian matrix $\mathbf{d_2}$ is the $K_m\times K_m$ semidefinite positive matrix of the second derivatives. The only change in the algorithm of *SI Text*, section III is the updated rule for the forces:

$$\mathbf{x_s}\to\mathbf{x_s}-\mathbf{d_2}^{-1}\times\mathbf{d_1}, \quad \textbf{[S39]}$$

where $\mathbf{d_2}^{-1}$ denotes the matrix inverse of $\mathbf{d_2}$.

VI. Local Density of Motifs

Let us call $p_i(c_i|x_s)$ the probability that the ith codon on a randomly drawn sequence under force x_s is c_i. This quantity can be computed with the transfer matrix formalism of *SI Text*, section II. For simplicity we restrict ourselves to the case of motifs with two nucleotides $(K=K_c=2)$.

To do so we first apply the procedure described by formulae **S11** and **S12**. We start from $v_L(c_L)=1$ for all $\deg(a_L)$ codons at site L, and calculate $v_{L-1}(c_{L-1})$ using transfer matrix $\mathbf{M_{L-1}}$ and Eq. **S11**. Through successive applications of the transfer matrices $\mathbf{M_{L-2}},\ldots,\mathbf{M_{i+1}}$ we obtain the vector $v_i(c_i)$ at site i.

Next the same procedure is followed, starting from site $i=1$, through successive multiplications by the transfer matrices from left to right. More precisely, we define $w_1(c_1)=1$ for all $\deg(a_L)$ codons at site 1. We then compute

$$w_2(c_2)=\sum_{c_1}w_1(c_1)M_1(c_1,c_2). \quad \textbf{[S40]}$$

This procedure is iterated until we compute

$$w_i(c_i)=\sum_{c_{i-1}}w_{i-1}(c_{i-1})M_{i-1}(c_{i-1},c_i). \quad \textbf{[S41]}$$

Finally we obtain the probability of codon c_i through

$$p_i(c_i|x_s)=\frac{w_i(c_i)v_i(c_i)}{Z(x_s)}. \quad \textbf{[S42]}$$

This probability is correctly normalized, according to **[S10]**. Special care must be brought to the cases $i=1, i=L$, that is, to the extremities of the sequence to ensure a proper counting of the number of motif occurrences in the sequence.

The generalization to the joint probability $p_{i,i+1}(c_i,c_{i+1}|x_s)$ of contiguous codons c_i,c_{i+1} is straightforward. The outcome is

$$p_{i,i+1}(c_i,c_{i+1}|x_s)=\frac{w_i(c_i)M_i(c_i,c_{i+1})v_{i+1}(c_{i+1})}{Z(x_s)}. \quad \textbf{[S43]}$$

To compute the probability $p_b(m|x_s)$ that motif m appears in the sequence, starting on base b, two cases must be considered:

- If b is a multiple of 3, plus 1 or 2, then the motif is in positions 1,2 or 2,3 of a codon, say, c_i. We can use the single-codon probability $p_i(c_i|x_s)$ to calculate $p_b(m|x_s)$, e.g., for $b=3(i-1)+1$,

$$p_b(m|x_s)=\sum_{\nu}p_i(c_i=\{m_1,m_2,\nu\}|x_s), \quad \textbf{[S44]}$$

where the sum runs over all nucleotides ν such that $\{m_1,m_2,\nu\}$ is a valid codon (synonymous to the ith codon of C_0).

- If b is a multiple of 3, then the motif is in position 3 of codon c_i, and in position 1 of c_{i+1} for some i. We can use the two-codon probability $p_{i,i+!}(c_i,c_{i+1}|x_s)$ to calculate $p_b(m|x_s)$:

$$p_b(m|x_s)=\sum_{\nu_1,\nu_2,\mu_1,\mu_2}p_{i,i+1}(c_i=\{\nu_1,\nu_2,m_1\},c_{i+1}=\{m_2,\mu_1,\mu_2\}|x_s), \quad \textbf{[S45]}$$

where $b=3i$.

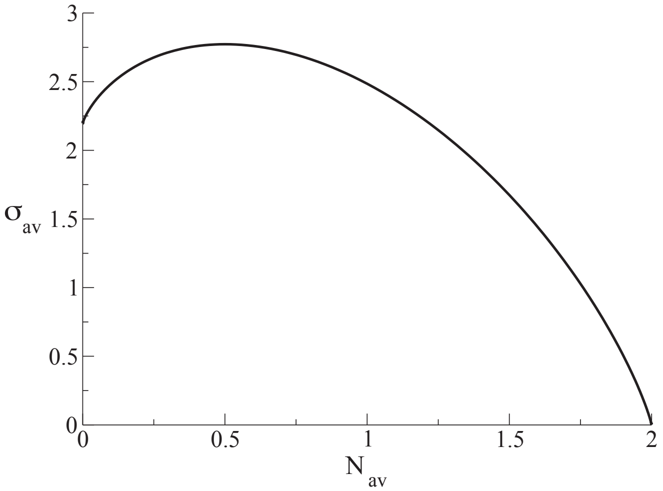

Fig. S1. Entropy σ_{av} of sequences with amino acid sequence Pro-Pro vs. average number N_{av} of occurrences of the motif m = CT. The curve was obtained from a parametric representation $(N_{av}(x_s),\sigma_{av}(x_s))$, and by varying x_s from $-\infty$ to $+\infty$.

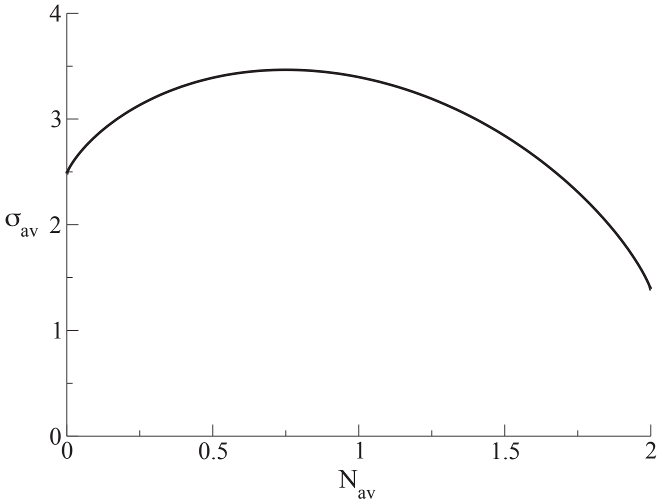

Fig. S2. Entropy σ_{av} of sequences with amino acid sequence Pro-Pro-Cys vs. average number N_{av} of occurrences of the motif m = CT. The curve was obtained from a parametric representation $(N_{av}(x_s),\sigma_{av}(x_s))$ (refer to **[S29]**) and by varying x_s from $-\infty$ to $+\infty$.

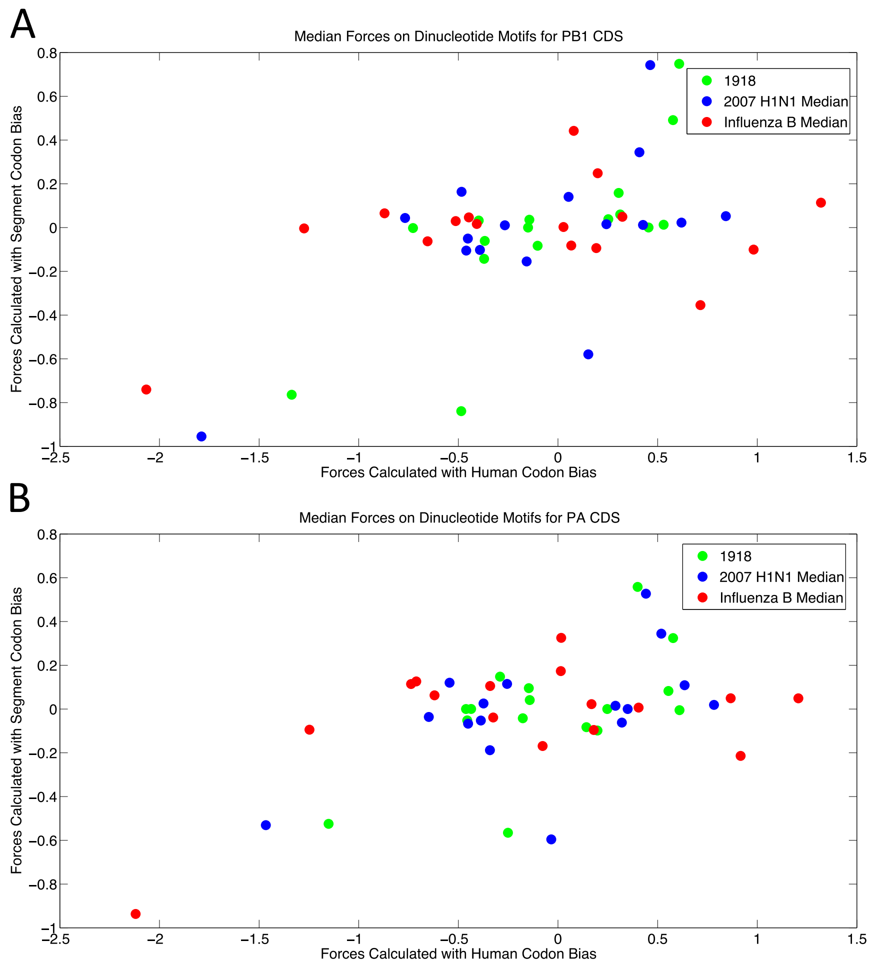

Fig. S3. A comparison of the selective forces when calculated using the segment and human codon biases for the 16 dinucleotides for the (*A*) PB1 and (*B*) PA genes in influenza. These quantities are calculated for the 1918 H1N1 segments, and the H1N1 segments from 2007 and for influenza B. In the later two cases the median values are shown.

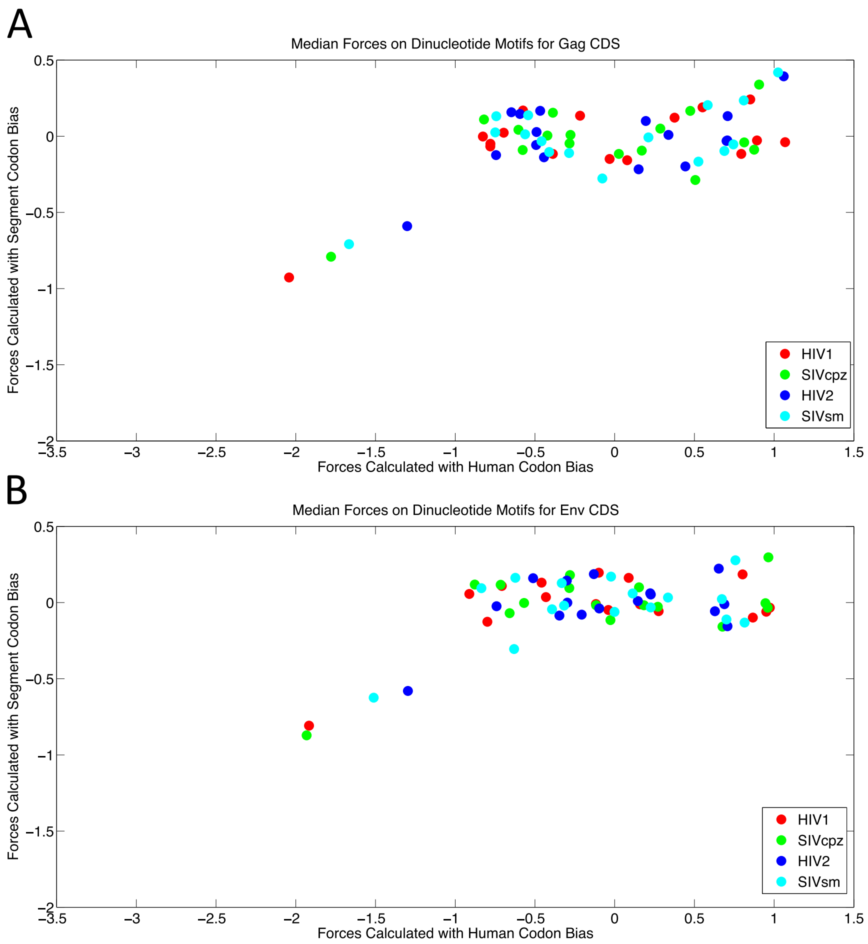

Fig. S4. A comparison of the median selective forces when calculated using the segment and human codon biases for the 16 dinucleotides for the (*A*) gag and (*B*) env genes. These quantities are calculated for HIV1, SIV chimpanzee (SIVcpz), HIV2, and SIV sooty mangabee (SIVsm).

Other Supporting Information Files

Table S1 (DOCX)
Table S2 (DOCX)
Table S3 (DOCX)
Dataset S1 (XLSX)