Quantitative theory of entropic forces acting on constrained nucleotide sequences applied to viruses

Benjamin D. Greenbaum[a,b], Simona Cocco[c], Arnold J. Levine[a,1], and Rémi Monasson[d]

[a]The Simons Center for Systems Biology, Institute for Advanced Study, Princeton, NJ 08540; [b]Departments of Medicine and Pathology, Division of Hematology and Medical Oncology, and the Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai, New York, NY 10029; [c]Laboratoire de Physique Statistique de l'Ecole Normale Supérieure, Unité Mixte de Recherche 8550, Associé au Centre National de la Recherche Scientifique et aux Universités Paris VI et Paris VII, 75005 Paris, France; and [d]Laboratoire de Physique Théorique de l'Ecole Normale Supérieure, Unité Mixte de Recherche 8549, Associé au Centre National de la Recherche Scientifique et à l'Université Paris VI, 75005 Paris, France

Contributed by Arnold J. Levine, February 11, 2014 (sent for review November 21, 2013)

We outline a theory to quantify the interplay of entropic and selective forces on nucleotide organization and apply it to the genomes of single-stranded RNA viruses. We quantify these forces as intensive variables that can easily be compared between sequences, outline a computationally efficient transfer-matrix method for their calculation, and apply this method to influenza and HIV viruses. We find viruses altering their dinucleotide motif use under selective forces, with these forces on CpG dinucleotides growing stronger in influenza the longer it replicates in humans. For a subset of genes in the human genome, many involved in antiviral innate immunity, the forces acting on CpG dinucleotides are even greater than the forces observed in viruses, suggesting that both effects are in response to similar selective forces involving the innate immune system. We further find that the dynamics of entropic forces balancing selective forces can be used to predict how long it will take a virus to adapt to a new host, and that it would take H1N1 several centuries to adapt to humans from birds, typically contributing many of its synonymous substitutions to the forcible removal of CpG dinucleotides. By examining the probability landscape of dinucleotide motifs, we predict where motifs are likely to appear using only a single-force parameter and uncover the localization of UpU motifs in HIV. Essentially, we extend the natural language and concepts of statistical physics, such as entropy and conjugated forces, to understanding viral sequences and, more generally, constrained genome evolution.

The nucleotide sequence of a genome is composed of a variety of sequence motifs whose organization is influenced by many forces. Most prominently, amino acid coding sequences are restricted by the genetic code and codon use patterns for a particular organism or tissue (1–3). Likewise, a variety of *cis*-acting nucleotide sequences control gene expression profiles, regulating factors such as timing, quantity, and responses to environmental cues.

Karlin et al. first showed that the relative abundance of dinucleotides in viral genomes could elucidate evolutionary relationships between groups of viruses, and viruses and their hosts (4). Likewise, Rabadan et al. (5) and Greenbaum et al. (6, 7) first demonstrated that in influenza genomes, nucleotide sequence-specific evolutionary changes occur over decades and reflect viral transitions from avian to human hosts. These changes are not driven by amino acid alterations or codon preference—they largely reduce CpG containing nucleotide sequence motifs by third codon position changes that have no impact on amino acid composition of the viral proteins. It was posited that this effect was due to differences between the human and avian innate immune systems, which would recognize (in humans) or not recognize (in birds) CpG dinucleotides in the RNA of these viruses, possibly via a Toll-like receptor (TLR) (8). Hence, influenza viruses moving into humans would adapt their genome sequence motifs to avoid detection and inhibition by the host immune system. Other patterns of host genome mimicry have been demonstrated between viruses and their hosts (9, 10). In viruses such as HIV, host enzyme activity creates biased nucleotide composition in viral RNA and DNA (11, 12). Additional examples, such as secondary structures of RNA genomes and bacterial restriction enzymes, exert analogous selective forces on sequence motifs (13–15). Thus, there are many different forces under which a genome's information content may be optimized for a particular environmental advantage.

There has not been a general quantitative theory designed to characterize the forces that directly affect nucleotide sequence organization and how they can change over evolutionary time when a genome is introduced into a new environment. To accomplish this we use an approach from statistical physics (16, 17). We apply our method to the genomes of single-stranded RNA (ssRNA) viruses, quantifying the degree to which avoidance or enhancement of a nucleotide motif causes a virus to alter its sequence organization relative to a given background distribution. The magnitude of the effect is captured by a selective force, conjugated to the number of times a motif occurs. In much the same way as with thermodynamic forces, acting on the volume or the number of particles constrain a system, the presence of selective forces constrains the diversity of viral genomes. By contrast, the high rates of mutation and replication for these RNA viruses provide a great deal of sequence diversity, creating "entropic forces" opposing the selective forces minimizing sequence diversity. The larger of these two forces then drives the evolution of the virus until, eventually, the two forces balance each other and an evolutionary equilibrium state is reached.

Many viral genomes, such as those for ssRNA viruses, are largely devoted to protein coding. In the absence of selective forces on motifs, and fixing the amino acid sequence for a given protein, codon use patterns would dictate the diversity of genome sequences. An "ideal" virus, in the absence of other outside forces on motifs, would evolve to have the number of motifs one would expect given its amino acid sequence and a codon use bias for the tissue in which it replicates. We can then derive the

Significance

This paper proposes a simple theory, inspired by statistical mechanics, of the interplay of entropic forces and selective forces on dinucleotide frequencies in viral genomes that occur when a virus migrates to a new host. The theory is quantitatively developed, and leads to many predictions about statistical features of viral evolution. The approach is general, and could be easily extended to other genomic data and have wider applications. For example, an analysis of avian influenza entering the human genome has identified selection against CpG dinucleotides, which have been shown to trigger a response of the innate immune system and interferon production.

Author contributions: B.D.G., S.C., A.J.L., and R.M. designed research; B.D.G., S.C., and R.M. performed research; B.D.G., S.C., A.J.L., and R.M. analyzed data; and B.D.G., S.C., A.J.L., and R.M. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. E-mail: alevine@ias.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1402285111/-/DCSupplemental.

selective force on a motif in a virus by calculating the degree to which the viral sequence is in a lower probability state than this ideal virus, given the number of times that motif occurs. In this work we examine these forces on dinucleotide motifs using both the codon use bias of the protein sequence under consideration and the average codon use bias of its host, interpreting the cases where they differ.

Materials and Methods

Sequence Data. The influenza sequences used in this study were taken from the National Center for Biotechnology Information Influenza Database (18). Only those sequences containing complete coding regions were then used in the analysis. The HIV sequences were taken from the Los Alamos HIV database and the same controls were applied (19). The list of all sequences used appears in Dataset S1. The human genomes used for the codon bias calculation was Consensus-Coding DNA Sequence (CCDS) Build Hs36.3. The data were obtained from the University of California, Santa Cruz Genome Browser (20–22).

Distribution over Sequence Space. We want to quantify the constraints acting on a nucleotide motif m in a (viral) DNA or RNA sequence, hereafter called C_0. We introduce a model over the set of all codon sequences $C = (c_1, c_2, \ldots, c_L)$, differing from C_0 through synonymous changes only. In the absence of constraints the probability of a sequence C in our model is simply the product of the probabilities of its codons, $p_i(c_i)$, where p_i is the codon bias of the ith codon in C_0. In the presence of a constraint over a nucleotide motif m, the probability of a sequence C becomes

$$P(C|x_s) = \frac{1}{Z(x_s)} \prod_{i=1}^{L} p_i(c_i) \exp\left(x_s\ N_m(C)\right) \qquad [1]$$

where $N_m(C)$ is the number of occurrences of the motif m in C, and the denominator

$$Z(x_s) = \sum_{\text{sequences } C} \prod_{i=1}^{L} p_i(c_i) \exp\left(x_s\ N_m(C)\right) \qquad [2]$$

ensures that the probability P is correctly normalized. Parameter x_s, hereafter called the "selective" force, introduces a bias over P. Positive values for x_s push the distribution toward sequences with large N, whereas negative x_s favor sequences with a small N.

The choice of the exponential dependence on N in **[1]** is justified by information-theoretic arguments: P defined above is the least constrained distribution (with minimal information, or with maximal entropy), whose average number of motifs is

$$N_{av}(x_s) = \sum_{\text{sequences } C} P(C|x_s) N_m(C) = \frac{\partial \log Z}{\partial x_s}(x_s). \qquad [3]$$

The value of the selective force x_s can then be chosen such that $N_{av}(x_s)$ is equal to the number of motifs m in the original sequence, $N_m(C_0)$. The formalism above can be easily extended to the case of multiple selective forces, acting on multiple motifs. Details can be found in *SI Text*.

Entropy of Sequences as a Function of the Number of Motifs. Let $\sigma(N)$ be the logarithm of the number of sequences C having N repetitions of m, hereafter called "entropy." $\sigma(N)$ is bounded from above by σ_0, the total entropy of the distribution of sequences in the absence of selective force ($x_s = 0$). σ_0 is equal to the sum over all 20 amino acids a of the number of codons coding for a in the sequence C_0, multiplied by the entropy of the codon bias distribution for this amino acid a. See *SI Text* where bounds on its value are also derived.

Classical equilibrium thermodynamic relations show that $\sigma(N)$ is the Legendre transform of $\log Z(x_s)$ (23):

$$\log Z(x_s) = \max_N \left(\sigma(N) - \sigma_0 + x_s N\right). \qquad [4]$$

The maximization condition over N expresses the balance between the selective force x_s and the "entropic" force

$$x_e(N) = \frac{d\sigma}{dN}(N), \qquad [5]$$

equal to the derivative of the entropy. At equilibrium, x_s and x_e sum to zero. However, selective and entropic force need not always compensate each other, as when out-of-equilibrium dynamical effects are present (*Dynamical Modeling*).

The Legendre formalism **[4]** provides a parametric representation of the entropy curve $(N,\sigma(N))$ under the form $(N_{av}(x_s),\sigma_{av}(x_s))$, which yields $N_{av}(x_s)$ as given by **[3]**, and

$$\sigma_{av}(x_s) = \sigma_0 + \log Z(x_s) - x_s N_{av}(x_s). \qquad [6]$$

As x_s spans the set of real numbers, the entropy curve is obtained; its maximum is reached in $(N_{av}(0),\sigma_{av}(0) = \sigma_0)$, corresponding to vanishing force, $x_s = x_e = 0$.

We illustrate the notions above with a very simple example of a sequence C_0 coding for one alanine ($L = 1$). We assume for simplicity that all four codons $c =$ GCn, with n = A, U, C, and G, coding for alanine have equal probabilities $p(c) = 1/4$ (uniform codon bias). The entropy of sequences in the absence of selective force is $\sigma_0 = \log 4$, which is the logarithm of the degeneracy of alanine. In the presence of a selective force x_s and for the motif $m =$ CG, we readily obtain $Z(x_s) = 3/4 + e^{x_s}/4$. The average number of motifs is $N_{av}(x_s) = e^{x_s}/(3 + e^{x_s})$ according to **[3]**, and the entropy is $\sigma_{av}(x_s) = \log(3 + e^{x_s}) - x_s e^{x_s}/(3 + e^{x_s})$ according to **[5]**. The corresponding entropy curve is plotted parametrically in Fig. 1*A* (see legend for further explanations).

In the generic case of a sequence C_0 of length L, the sum defining $Z(x_s)$ in **[2]** runs over an exponentially large-in-L number of sequences C. It can, however, be computed very efficiently, in a time growing linearly with L only. The method, called "transfer matrix" in statistical physics or "dynamic programming" in computer science, allows us to compute the entropy even for very long sequences in a short time. This method is useful for understanding the properties of a large system based on the interactions between its subsystems, which in our case are neighboring codons. Simple examples of the transfer matrix method (Figs. S1 and S2), and details about its implementation are found in *SI Text*.

Given the selective force x_s the number N of motifs m in a random sequence C fluctuates around the average value $N_{av}(x_s)$, with a variance $\mathrm{var}(N|x_s)$. Reciprocally the value of the force such that $N_{av}(x_s)$ is equal to the number N of motifs in the real sequence C_0 may fluctuate around its average value with a variance $\mathrm{var}(x_s|N)$. Both variances can be computed from the uncertainty relation

$$\mathrm{var}(N|x_s) = \frac{1}{\mathrm{var}(x_s|N)} = \frac{\partial^2 \log Z}{\partial x_s^2}(x_s). \qquad [7]$$

The variance in the force on a motif is useful to estimate whether two values of the forces computed for two sequences are statistically distinct.

Dynamical Modeling. We model the time evolution of N from an initial value N_0 to its equilibrium value $N_{av}(x_s)$ under the action of a selective force x_s through a simple relaxation dynamics

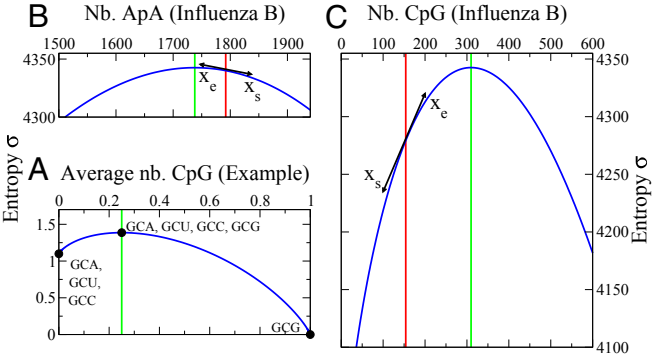

Fig. 1. Entropy curve $\sigma(N)$ as a function of the number N (Nb.) of occurrences of a motif. (*A*) Toy example of a single-codon sequence, coding for alanine (derived in *Materials and Methods*). The entropy for an influenza B isolate (B/Cordoba/2979/1991) is derived for motifs (*B*) ApA and (*C*) CpG. Green and red lines show, respectively, the zero-force and real values of the numbers of motifs, with the arrows indicating the balance of selective and entropic forces at the real value. The ApA (*B*) entropy is flatter than the CpG (*C*) entropy around the maximum $\sigma_0 = 4{,}342.6$.

$$\tau \frac{dN(t)}{dt} = x_e(N(t)) + x_s, \quad [8]$$

where the entropic force, x_e, is defined in **[5]**. The value of N will evolve until the imposed selection force x_s balances the entropic force $x_e(N)$, resulting from the loss of entropic diversity of the sequences. Parameter τ is a measure of the time scale on which the number of motifs diminishes by one unit, when the difference between the forces is of the order of the unity, e.g., at the beginning of the evolution. As the difference between the forces gets smaller and smaller with time, the relaxation time to equilibrium is much larger than τ.

Results

Forces Relative to Viral and Human Codon Biases. *Materials and Methods* presents a procedure to compute the entropy, that is the logarithm of the number of sequences, as a function of the number of repetitions of a given motif. The background distribution is derived from either viral or human codon biases. We also work through a simple example of how to use these methods, which is illustrated in Fig. 1*A*.

When the H1N1 influenza A virus entered the human population from a likely avian host in 1918, the CpG dinucleotide content of the genome was lowered from levels typically associated with avian viruses toward levels more associated with human viruses (6–8). For the genomes of influenza B isolates, a virus for which humans have been a natural host for many centuries, the number of CpG dinucleotides varies little over time. Fig. 1*B* shows the entropy curve for ApA. The curve is flat and symmetric, and the slope of the curve at the value of ApA in the real virus is close to zero (the maximum entropy value). The occurrence of ApA dinucleotides to a large degree may therefore vary randomly. The number of CpG dinucleotides corresponds to a location on an entropy curve of high slope, as shown in Fig. 1*C*. We define the entropic force as the slope at the actual value of these motifs in the viral genome. Unlike ApA, the selective force acting on CpG, opposite to the entropic force, is very different from the zero value corresponding to maximum entropy. Both curves have the same maximum value, as they have the same entropy when no force is applied. An expression for the maximum value, and how it is bounded, appears in *SI Text*.

One can use either the virus or host codon bias to generate the sequence background distribution relative to which these forces may be inferred, and the resulting forces must therefore be interpreted relative to that choice. For the human codon bias, we use the coding regions of the whole human genome. This is an average codon use bias that may not reflect more restricted biases that occur in particular gene families or cell types. Fig. 2 compares the selective forces calculated for all 16 dinucleotides derived relative to both the host and virus segment codon use biases, for the longest genes of influenza polymerase basic 2 (PB2) and HIV polymerase (pol). In Fig. 2*A* the median force values are given for influenza A H1N1 in 1918 (green) and 2007 (blue), along with those for influenza B (red). Relative to the host codon use bias, and unlike the viral segment codon bias, the forces acting on dinucleotides are often nonzero, with CpG being the only large standout in magnitude. The dispersion of forces over many dinucleotide motifs, relative to the average host codon use bias, typically increases with time, and is greatest in influenza B, the virus adapted to humans for the longest time. UpG and CpA, the mutational outcomes of CpG avoidance, have a positive force relative to the host codon bias, and UpA has a negative force (two mutational events are a less likely path). Most of the forces on dinucleotides are smaller relative to the viral codon bias than relative to the human codon bias. Therefore, actual sequences in a viral genome are closer to ideal viruses generated by the viral codon use than those generated by human codon use. Thus, there is a limit to the host mimicry observed in these viruses. The influenza A PB1 and PA genes are similarly analyzed and can be found in Fig. S3.

The polymerase gene (pol) from HIV-1 was analyzed in the same fashion and the results are shown in Fig. 2*B*, where the same quantities are calculated for the pol gene from HIV-1, SIV chimpanzee (SIVcpz), HIV-2, and SIV sooty mangabee (SIVsm), all related viruses. The selective force on dinucleotides for viruses changes less between these hosts, which are more closely related (humans and simians), than in influenza (humans and avians). One difference between HIV and influenza is that the dinucleotide ApG in HIV genomes stands out as having a positive force not observed with influenza virus segments. ApG motifs have been associated with the action of RNA-editing enzymes on the HIV genome (11, 12).

Despite their very different genome replication cycles, most motifs of dinucleotides from HIV and influenza have no force acting on them relative to the viral codon bias, whereas there are

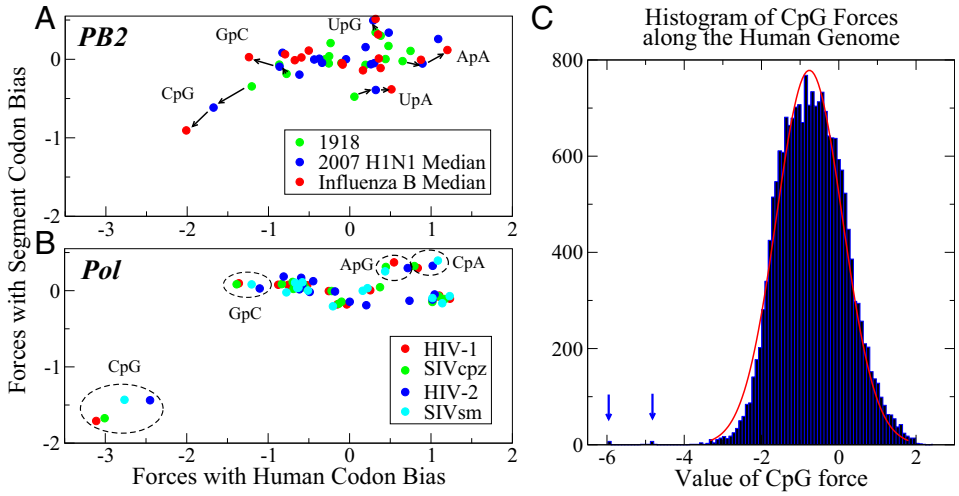

Fig. 2. Comparison of selective forces using both segment and human codon biases for all dinucleotides. Forces are derived for (*A*) influenza PB2 (showing the 1918 H1N1 segment, and the median values for all 2007 H1N1 and influenza B segments), and for (*B*) HIV pol (showing median values for HIV-1, SIVcpz, HIV-2, and SIVsm). Dinucleotides under large forces are indicated. For PB2, arrows indicate the direction from 1918 to influenza B. In the HIV and SIV, ellipses contain outlying dinucleotides. (*C*) Histogram of forces on CpG for all human CDS regions, with a Gaussian fit to the bulk of the distribution. Far left outliers contain many type I IFNs.

dinucleotide motifs with forces acting on them relative to the human codon bias. A parallel analysis of gag and env is presented in Fig. S4. The forces on HIV may reflect a functional significance, as shown by the studies of Vabret et al. (24, 25), who showed that the HIV-1 virus did not replicate as well when third position codon nucleotides were changed (with no amino acid changes) in the gag gene.

Host Gene Mimicry. An intriguing result was obtained in Greenbaum et al. (7), when many of the genes that compose the innate immune system were examined, particularly type I IFN genes in the human genome. These genes also had very low numbers of CpG dinucleotides, as was observed with influenza viruses evolving in human populations. Based on those observations it was hypothesized that a subset of genes in the innate immune system are most subject to the forces acting on CpG motifs. The quantitative theory developed here now allows us to calculate and quantify those forces. It permits us to test the idea that forces acting to change CpG content are gene- or function-specific in the human genome. We show in Fig. 2*C* the histogram of the selective forces on CpG for all coding regions in the human genome. The distribution can be fitted with a Gaussian, with mean $\mu = -0.7611$ and SD $\sigma = 0.8551$ apart from genes falling well outside the distribution, with values less than −4. According to standard extreme value theory the expected minimum value from the normal distribution is equal to $\mu - \sigma\sqrt{2\log(N)}$, where N is the number of normally drawn samples (26). For this case, the expected minimum value if −4.5674 and all outliers less than −4 also fall outside of that value. A table of the median values of the forces and their variances, another means of assessing outlier significance, is shown in Table S1.

Many type I IFN genes appear as outliers on the left of Fig. 2*C*. The value of the CpG forces these genes are under, along with other information is shown in Table S2. One would predict that the effect of such forces on these genes could be used as a discovery tool for human genes regulated in a similar fashion with similar functions. This would be a quantitative definition of a subset of genes that populate the human innate immune system. As an experiment, we explored all genes whose CpG force values were less than that observed in PB2 for influenza B. The results are depicted in Table S3. Strikingly, not only are the type I IFN genes depicted, but other innate immune genes are present in this group.

For the same force to be causing these effects on both a host and viral gene set, causing the virus to mimic the very genes that respond to it, many mutational events must occur. The force could be driven by a receptor that observes and interacts with the RNA CpG motif, leading to the transcription of genes that limit viral replication (8). The force would act on a set of host response genes, as well as viral genes, so the mRNA of the host genes would minimize CpG content to prevent a positive feedback loop from occurring. Innate immune recognition of CpG in DNA is known to occur via TLR9, and TLR7 and -8 recognize ssRNA (8, 27). CpG methylation occurs in the DNA of host genes protecting them from these innate responses, whereas methylation is not observed in RNA viral genomes lacking a DNA intermediate step.

A Dynamical Model for the Influenza A Virus. In H1N1 human influenza viruses, the force on CpG levels declines over time approaching levels seen within influenza B viral genomes. The effect is strongest in the three longest genome segments, and is less noticeable in the HA gene, presumably due to strong selection from the adaptive immune system on the HA protein. As seen in the previous section, these dynamical changes which occur when influenza viruses switch from avian to human hosts, are not observed when HIV and SIV are compared, likely reflecting the fact that HIV came into the human population from a more closely related simian reservoir.

A dynamical model was used to better understand how forces and motifs evolve with H1N1 influenza viruses with time. In this model, the number of motifs evolves under a (negative) selective force, which increases the magnitude of the entropic force (reducing sequence complexity) until both the selective and entropic forces compensate one another, and equilibrium is reached (*Materials and Methods*). For PB2, PB1, and PA, we first determine the selective force under the assumption that the influenza B genome represents the equilibrium force value for that segment, as it has evolved in humans for many years. The equilibrium force x_B is estimated by the mean value of the selective forces computed for all influenza B sequences (Table 1). For the initial condition x_0 we choose the corresponding force value for the H1N1 sequence from 1918 when H1N1 was first introduced into humans (Table 1). Our dynamical model then gives the entropic force as a function of time, $x_e(t)$, where t measures years of evolution. The opposite of the entropic force interpolates between $-x_e(0) = x_0$, and $-x_e(t \to \infty) = x_B$.

The outcomes of this analysis are shown in Fig. 3. The model includes a single time scale, τ, which represents the elementary time for motif loss, and is fitted to make $-x_e(t)$ best coincide with the H1N1 data over the available time range. Because the H1N1 virus disappeared from the human population in the early 1950s, and a nearly identical virus reappeared in 1977, the 27 y that H1N1 was not circulating in the human population are not included in the time that this virus evolved in humans (28). The three rates $(1/\tau)$ for evolutionary change range from 0.17 per year for PA, to about 0.4 per year for PB1 and PB2. Those estimates are 2–5 times larger than the average time for a synonymous substitution to happen in the corresponding genes (of comparable lengths), about 0.07 per year (29). This would imply that one in two synonymous substitutions in PA and one in five in PB1 and PB2 result in the loss of a CpG motif. In addition, according to the model it would take about five centuries for the PA segment to reach equilibrium (Fig. 3). Remarkably all of these sequence data fall within one SD (calculated according to **[7]** and **[8]**) forming a narrow strip around the model prediction as seen in Fig. 3.

Motif Localization. To visualize how these forces affect where the motifs are likely or unlikely to be found in a viral genome, we examined the local motif density, as described in *SI Text*. To do so, we calculated the probability that a motif appears at a given position along the genome, both with the viral sequence and human codon bias. Compared with the positions of dinucleotides from sequence data this then becomes a test of the validity of the approach. Locations with high probabilities can be directly compared with the real locations of motifs from the viral sequence data.

To get a sense of what these distributions might look like in an equilibrium setting, the case of influenza B was examined. Both CpG and UpG dinucleotides were calculated with the viral segment codon bias. In the former case the motif has a negative force, and is therefore suppressed. In the latter the force is positive and the motif is enhanced.

The local probability landscape for CpG dinucleotide motifs with C in the third codon position is shown in Fig. 4*A*. The locations of CpGs determined from sequence data clearly tend to coincide with peaks in the probability landscape. To better visualize this effect, in Fig. 4*B*, the occupation probability predicted by our model,

Table 1. Dynamical parameters in H1N1 CpG force evolution

	Segment		
Parameters	PB2	PB1	PA
Equilibrium force x_B	−1.99	−2.04	−2.2
Initial force x_0	−1.21	−1.34	−1.15
Time scale τ	2.3	2.4	6.0

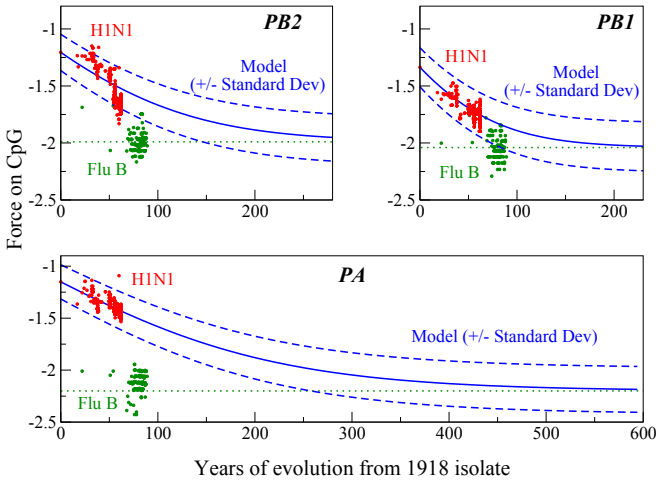

Fig. 3. Dynamical simulation of force equilibration using real values from H1N1 for PB2, PB1, and PA, with the human codon bias. PA has the longest time scale. Red dots show the selective force and year for one isolate; multiple isolates may come from a given year. The negative average entropic force in the model is shown as a function of time in blue (continuous line), with ±1 SD (dashed lines). Selective forces for influenza B for each segment are in green (each dot corresponds to a virus sequence). Dotted green lines indicate the equilibrium force x_B.

averaged over the positions at which CpGs occur in the viral sequence, is plotted along with the same probability averaged over the positions where the motif is absent from the RNA sequence. The average probability associated with CpG occurrences from the sequence data are consistently higher than the one corresponding to locations with no CpG in the sequence data. Even though the values of the force on CpG dinucleotides declines over time (*Dynamical Modeling*), the ratio of the average probability associated with actual CpG occurrences in a sequence to the probabilities associated with sites where no real CpG occurs remains essentially fixed.

Next, local probability landscapes for two examples of motifs under positive selective forces were examined. Fig. 4*C* examines the dinucleotide UpG, which undergoes a meaningful positive force with respect to the viral segment bias. Unlike with CpG, UpG is not a rare dinucleotide, so the fact that most sequence occurrences of the dinucleotide come at high probability "hotspots" is clear.

Finally, we note that in HIV-1, a retrovirus with a very different replication cycle than influenza, localization of motifs also occurs. Fig. 4*D* shows the HIV-1 gag dinucleotide probability landscape with respect to the human genome bias for the dinucleotide UpU. There is a clear cluster toward the end of the gag gene. This cluster is located in the region shown by Pavlakis and colleagues to be a regulatory feature for the timing of the gag gene expression, and is more precisely defined here (30, 31). Here the selective forces for optimal replication of the virus limits the sequence motif entropy. UpU motifs have also been associated with the activation of a TLR (TLR8) (32).

Discussion

We have developed a quantitative method for the analysis of the entropic and selective forces that act to shape the distribution of nucleotide motifs in a genome. Although the genetic code and codon use clearly shape nucleotide sequence motifs in a genome, forces such as motif specific receptors and enzymes also play a role. Our approach quantifies these forces, shows their effect on sequence entropy, and allows direct comparisons between genomes as intensive quantities, meaning that their value is not reflective of the size of the genome considered, and one can thus compare their value between sequences with different origins. In addition to providing a more formal theoretical framework, this approach is computationally far faster than other attempts to measure similar parameters (6), where a set of randomized viruses had to be created to infer the number of significant motifs.

By far the strongest repressive forces altering viral genome landscapes act on CpG dinucleotides. This can be observed in very diverse viral genomes such as human influenza strains, HIV, and SIV, taking as a reference both the viral codon bias and the human codon bias. The observation is consistent with previous observations (6), where CpG was found to be underrepresented in influenza with respect to the viruses own codon bias, as well as in all other mammalian ssRNA viruses.

Likewise, in the H1N1 strain of influenza A, UpG and CpA are enhanced and UpA is repressed. With respect to the viral codon bias, those forces are essentially zero for HIV, SIV, and influenza B, but they differ from zero for the large majority of dinucleotides when using the human codon bias as reference. The fact that this occurs in viruses that have been replicating in humans (or simians) for a very long time indicates that, at equilibrium, the entropic and selective forces described in this paper generically reflect a divergence away from the typical human codon bias and the viral codon bias, limiting genome sequence motif mimicry. However, in the human genome a subset of genes have evolved very similar motif distributions to those observed for the viruses studied here. Many genes with strongly negative forces on CpG, correspond to genes of the innate immune system, providing a quantitative definition for detecting a gene that is part of the innate immune response to infections by these viruses. Indeed it is likely that diverse classes of viruses trigger different innate immune responses and we will detect different host genes by mimicry using different viruses.

Although innate immunity is one source of explanation for the observed effects, it is not the only one. Common RNA structural motifs could also provide an explanation for why some nucleotide motifs are subject to positive and negative forces, as well as other

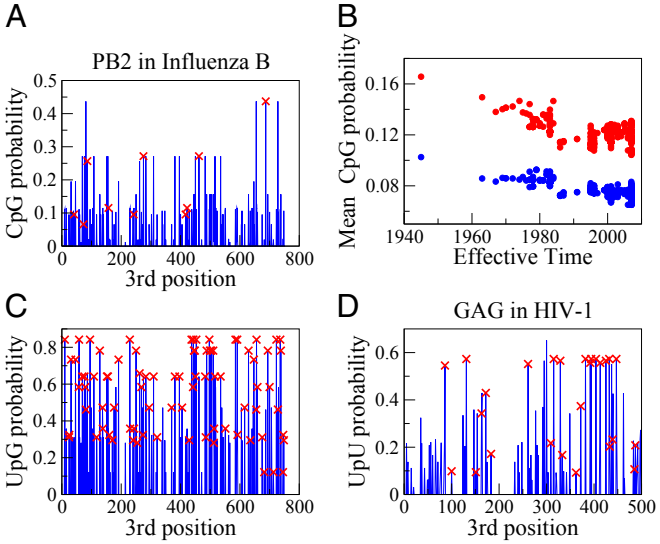

Fig. 4. (*A*) Probability of finding a CpG motif at a given third position along a sequence for PB2 in an influenza B isolate (blue lines), whereas a red × indicates where a real motif occurs. (*B*) Occupation probability for third codon positions computed from our model as a function of year of evolution, averaged over sites where a real motif occurs (red) and sites with no real motif but a nonzero probability of occurrence (blue). The former is higher by a ratio of about 1.6. (*C*) Same as *A* for PB2 under a force on UpG and using its segment codon bias. (*D*) Same as *A* for gag with a force on UpU using the human codon bias.

protein–RNA interactions involved in other processes besides innate immunity. There is a growing body of evidence that codon use, tRNA concentrations, and other rate-limiting translational factors will impact the sequence motifs used by a virus (33–36). As these methods are applied to other genomes besides viruses, a wider set of phenomena may underlie these forces.

Moreover, there is the issue of the larger set of forces that appear relative to the human codon bias, as opposed to the viral bias. We offer two possible explanations for this effect. First, it may be that codon use for the different cell types in which a given virus replicates is not the same as the average human codon bias. Certainly, tRNA use may vary between cell types. To remedy this, one would ideally use a cell-specific codon bias in future applications. Moreover, some viruses are known to manipulate host tRNA use to their advantage, which would induce selective forces, and may be responsible for the general dispersion of forces relative to the human codon bias, but not the viral codon bias. For instance, HIV-1 was postulated to modulate actively the tRNA pool under which it replicates, to maximize replication efficiency for its A-rich genome (33). Indeed relative to the human codon bias, ApA is the motif in HIV-1 with the greatest positive force on it. An enhancement of A-rich dinucleotides relative to the human codon bias may well reflect such a tendency, and the observance of a similar dispersion of dinucleotide forces in influenza B may show that such phenomena are a common viral strategy.

The model permits one to predict locations where a dinucleotide is more or less likely to occur along a sequence due to a given force, and to demonstrate regions where many occurrences of a motif localize, as was the case for UpU motifs in gag, associated with the timing and levels its protein. In addition the model can be constructed to show how these forces evolve. With only a time-scale parameter, we can fit through the dynamical model the evolution of CpG forces during the history of influenza A H1N1 segments and use this parameter to predict how long it will take for the virus to attain the level of force found in influenza B. The model showed an excellent fit of the predicted data points to the actual results as the H1N1 virus evolved between 1918 and 2007. It therefore provides an estimate for how long it may take an avian strain to equilibrate in a human host, as well as provides an estimate for the degree to which CpG forces contribute to its overall evolutionary rate.

The ideas presented here give a language for taking into account nonprotein coding features in a quantitative evolutionary theory. The approach is very general, can be used to study longer motifs, and can be generalized for many other types of sequence constraints. In doing so we hope to uncover other forms of latent information hidden beneath known constraints in genomes, and use this information as a tool for biological discovery. The generality of the approach comes from statistical physics, where forces describing the ordering of a system have a natural framework for characterization.

ACKNOWLEDGMENTS. B.D.G. thanks Nicholas F. Parrish (University of Pennsylvania) for his assistance in collecting the HIV and SIV sequence data used here and Olivier Manches (Icahn School of Medicine at Mount Sinai) for helpful discussions. S.C. and R.M. acknowledge the hospitality of the Simons Center for Systems Biology (Institute for Advanced Study) where this work was initiated, and the authors thank Suzanne Christen for her assistance. The authors thank the Rita Allen Foundation for its support of this work. B.D.G. was the Eric and Wendy Schmidt Member at the Simons Center for Systems Biology and thanks the Center for their support.

1. Sharp PM, Li WH (1987) The codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* 15(3): 1281–1295.
2. Powell JR, Moriyama EN (1997) Evolution of codon usage bias in Drosophila. *Proc Natl Acad Sci USA* 94(15):7784–7790.
3. Rocha EP (2004) Codon usage bias from tRNA's point of view: Redundancy, specialization, and efficient decoding for translation optimization. *Genome Res* 14(11): 2279–2286.
4. Karlin S, Doerfler W, Cardon LR (1994) Why is CpG suppressed in the genomes of virtually all small eukaryotic viruses but not in those of large eukaryotic viruses? *J Virol* 68(5):2889–2897.
5. Rabadan R, Levine AJ, Robins H (2006) Comparison of avian and human influenza A viruses reveals a mutational bias on the viral genomes. *J Virol* 80(23):11887–11891.
6. Greenbaum BD, Levine AJ, Bhanot G, Rabadan R (2008) Patterns of evolution and host gene mimicry in influenza and other RNA viruses. *PLoS Pathog* 4(6):e1000079.
7. Greenbaum BD, Rabadan R, Levine AJ (2009) Patterns of oligonucleotide sequences in viral and host cell RNA identify mediators of the host innate immune system. *PLoS ONE* 4(6):e5969.
8. Jimenez-Baranda S, et al. (2011) Oligonucleotide motifs that disappear during the evolution of influenza virus in humans increase alpha interferon secretion by plasmacytoid dendritic cells. *J Virol* 85(8):3893–3904.
9. Pezda AC, Penn A, Barton GM, Coscoy L (2011) Suppression of TLR9 immunostimulatory motifs in the genome of a gammaherpesvirus. *J Immunol* 187(2):887–896.
10. Elde NC, Malik HS (2009) The evolutionary conundrum of pathogen mimicry. *Nat Rev Microbiol* 7(11):787–797.
11. Liddament MT, Brown WL, Schumacher AJ, Harris RS (2004) APOBEC3F properties and hypermutation preferences indicate activity against HIV-1 in vivo. *Curr Biol* 14(15): 1385–1391.
12. Wood N, et al. (2009) HIV evolution in early infection: Selection pressures, patterns of insertion and deletion, and the impact of APOBEC. *PLoS Pathog* 5(5):e1000414.
13. Qian L, Kussell E (2012) Evolutionary dynamics of restriction site avoidance. *Phys Rev Lett* 108(15):158105.
14. Lin WH, Kussell E (2012) Evolutionary pressures on simple sequence repeats in prokaryotic coding regions. *Nucleic Acids Res* 40(6):2399–2413.
15. Brower-Sinning R, et al. (2009) The role of RNA folding free energy in the evolution of the polymerase genes of the influenza A virus. *Genome Biol* 10(2):R18.
16. Nelson P (2007) *Biological Physics: Energy, Information, and Life* (WH Freeman, New York), 1st Ed, Chap 7.
17. Baxter RJ (1982) *Exactly Solved Models in Statistical Mechanics* (Academic, Amsterdam), Chap 1-2.
18. Bao Y, et al. (2008) The influenza virus resource at the National Center for Biotechnology Information. *J Virol* 82(2):596–601.
19. HIV Sequence Compendium (2012) eds Kuiken C, et al. (Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, NM), LA-UR 12-24653.
20. Kent WJ, et al. (2002) The human genome browser at UCSC. *Genome Res* 12(6):996–1006.
21. Karolchik D, et al. (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* 32(Database issue):D493–D496.
22. Lander ES, et al.; International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* 409(6822):860–921.
23. Zia RKP, Redish EF, McKay SR (2009) Making sense of the Legendre transform. *Am J Phys* 77:614.
24. Vabret N, et al. (2012) The biased nucleotide composition of HIV-1 triggers type I interferon response and correlates with subtype D increased pathogenicity. *PLoS ONE* 7(4):e33502.
25. Vabret N, et al. (2014) Large-scale nucleotide optimization of simian immunodeficiency virus (SIV) reduces its capacity to stimulate type-I IFN in vitro. *J Virol*, 10.1128/JVI.03223-13.
26. Gumbel EJ (1958) *Statistics of Extremes* (Columbia Univ Press, New York), 1st Ed.
27. Hemmi H, et al. (2000) A Toll-like receptor recognizes bacterial DNA. *Nature* 408(6813): 740–745.
28. Nakajima K, Desselberger U, Palese P (1978) Recent human influenza A (H1N1) viruses are closely related genetically to strains isolated in 1950. *Nature* 274(5669):334–339.
29. Hanada K, Suzuki Y, Gojobori T (2004) A large variation in the rates of synonymous substitution for RNA viruses and its relationship to a diversity of viral infection and transmission modes. *Mol Biol Evol* 21(6):1074–1080.
30. Schwartz S, Felber BK, Pavlakis GN (1992) Distinct RNA sequences in the gag region of human immunodeficiency virus type 1 decrease RNA stability and inhibit expression in the absence of Rev protein. *J Virol* 66(1):150–159.
31. Schwartz S, et al. (1992) Mutational inactivation of an inhibitory sequence in human immunodeficiency virus type 1 results in Rev-independent gag expression. *J Virol* 66(12):7176–7182.
32. Chang JJ, Altfeld M (2009) TLR-mediated immune activation in HIV. *Blood* 113(2): 269–270.
33. van Weringh A, et al. (2011) HIV-1 modulates the tRNA pool to improve translation efficiency. *Mol Biol Evol* 28(6):1827–1834.
34. Mueller S, Papamichail D, Coleman JR, Skiena S, Wimmer E (2006) Reduction of the rate of poliovirus protein synthesis through large-scale codon deoptimization causes attenuation of viral virulence by lowering specific infectivity. *J Virol* 80(19): 9687–9696.
35. Coleman JR, et al. (2008) Virus attenuation by genome-scale changes in codon pair bias. *Science* 320(5884):1784–1787.
36. Mueller S, et al. (2010) Live attenuated influenza virus vaccines by computer-aided rational design. *Nat Biotechnol* 28(7):723–726.

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:

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

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:

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

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

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

where

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

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

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

The expression for F in the bulk of the sequence $(i\le L-1)$ 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:

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

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

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 $i=1,\ldots,L-2$,

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

and

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

Then, we observe that

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

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

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

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

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

The process is iterated until the first codon

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

Finally, we obtain the value of the normalization constant through

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

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

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

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)

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

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

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

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

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

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]

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

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

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

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

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

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

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

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 = 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

$$\mathbf{M_1}=\frac{1}{8}\begin{pmatrix}1 & 1\\ 1 & 1\\ e^{x_s} & e^{x_s}\\ e^{x_s} & e^{x_s}\end{pmatrix}. \qquad \textbf{[S23]}$$

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

$$Z(x_s)=\sum_{c_1,c_2}M_1(c_1,c_2)=\frac{1}{2}(1+e^{x_s}). \qquad \textbf{[S24]}$$

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

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

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]**:

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

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

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

and the normalization constant is

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

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

$$N_{av}(x_s)=\frac{e^{x_s}}{1+e^{x_s}}+\frac{e^{x_s}}{3+e^{x_s}}$$
$$\sigma_{av}(x_s)=2\log 2+\log(1+e^{x_s})+\log(3+e^{x_s})-\frac{x_s e^{x_s}}{1+e^{x_s}}-\frac{x_s e^{x_s}}{3+e^{x_s}}. \qquad \textbf{[S29]}$$

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

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

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

$$M_i(c_i,c_{i+1})=p_i(c_i)\exp\left[x_s^{(1)}\sum_{i=1}^{L-1}F\left(m^{(1)},c_i,c_{i+1}\right)+x_s^{(2)}\sum_{i=1}^{L-1}F\left(m^{(2)},c_i,c_{i+1}\right)\right], \qquad \textbf{[S31]}$$

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

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

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

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

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

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

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

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

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:

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

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 $b=3i$.

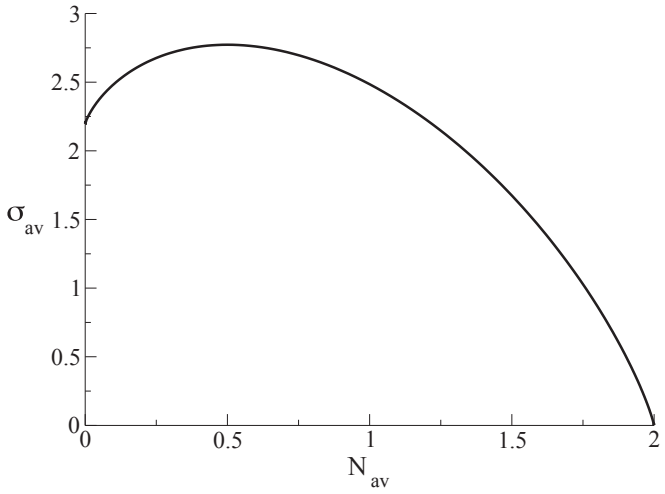

Fig. S1. Entropy σ_{av} of sequences with amino acid sequence Pro-Pro vs. average number N_{av} of occurrences of the motif m = CT. The curve was obtained from a parametric representation $(N_{av}(x_s),\sigma_{av}(x_s))$, and by varying x_s from $-\infty$ to $+\infty$.

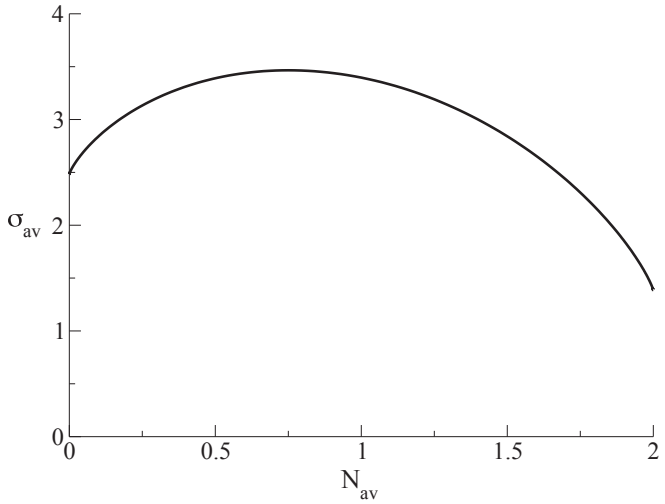

Fig. S2. Entropy σ_{av} of sequences with amino acid sequence Pro-Pro-Cys vs. average number N_{av} of occurrences of the motif m = CT. The curve was obtained from a parametric representation $(N_{av}(x_s),\sigma_{av}(x_s))$ (refer to **[S29]**) and by varying x_s from $-\infty$ to $+\infty$.

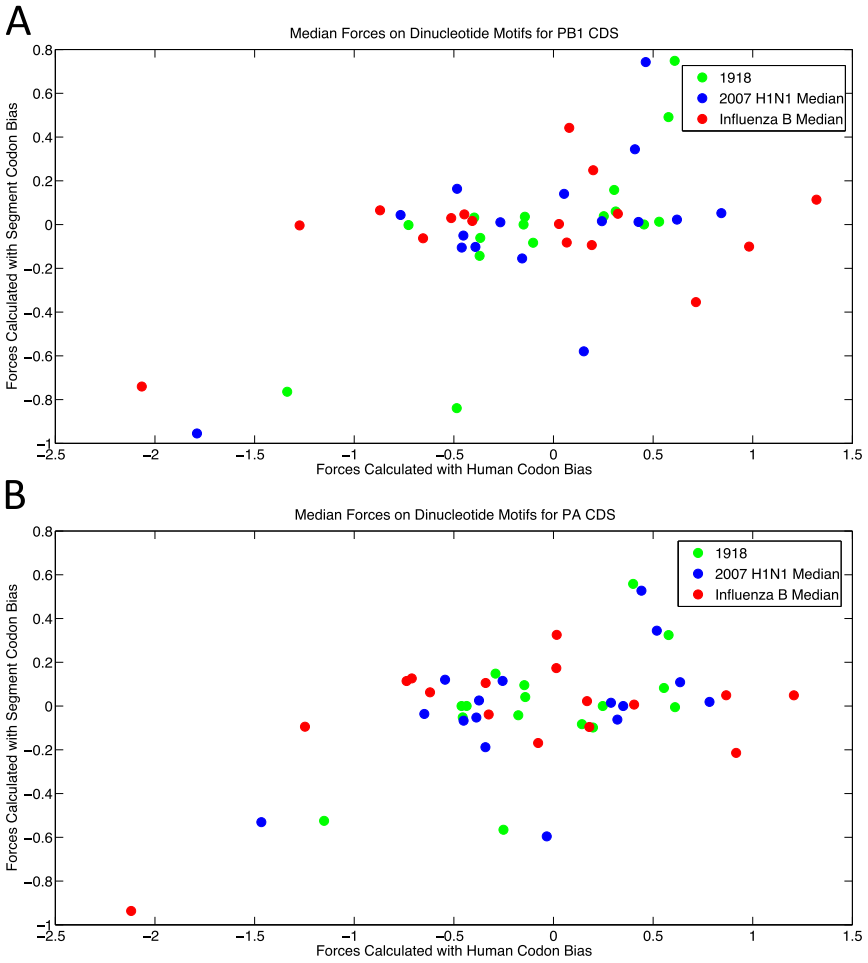

Fig. S3. A comparison of the selective forces when calculated using the segment and human codon biases for the 16 dinucleotides for the (*A*) PB1 and (*B*) PA genes in influenza. These quantities are calculated for the 1918 H1N1 segments, and the H1N1 segments from 2007 and for influenza B. In the later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

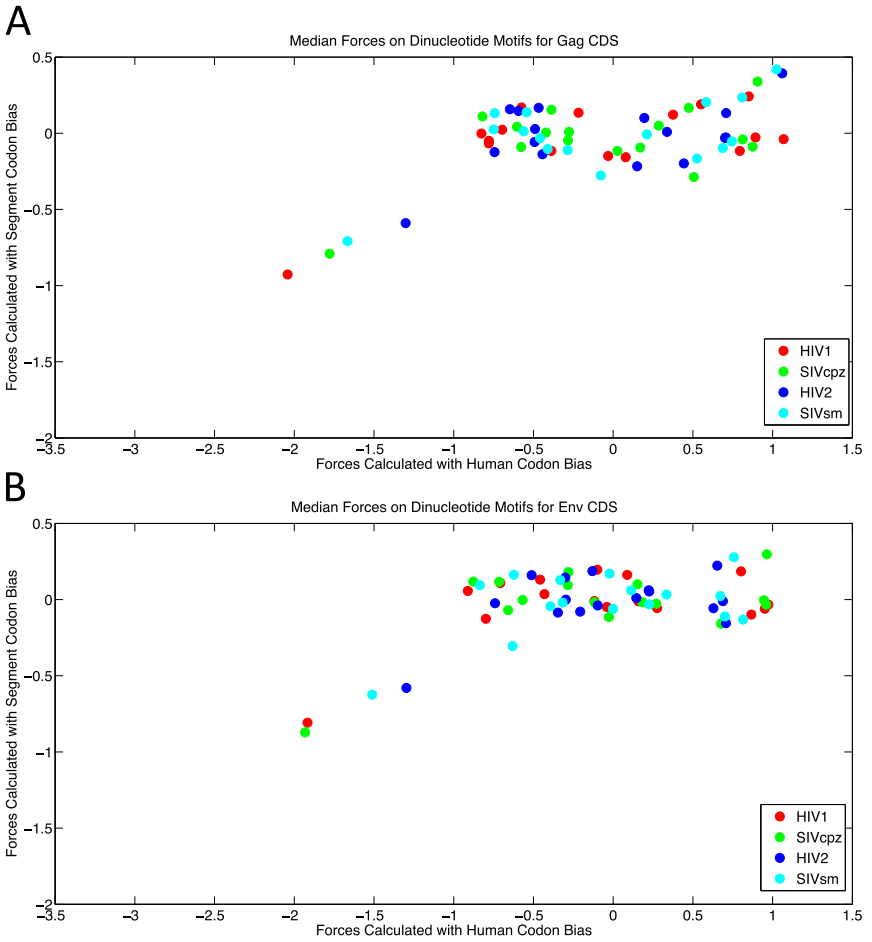

Fig. S4. A comparison of the median selective forces when calculated using the segment and human codon biases for the 16 dinucleotides for the (*A*) gag and (*B*) env genes. These quantities are calculated for HIV1, SIV chimpanzee (SIVcpz), HIV2, and SIV sooty mangabee (SIVsm).

Other Supporting Information Files

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