Entropy and codon bias in HIV-1

Aakash Pandey
Department of Biotechnology, Kathmandu University
gammadelta99@gmail.com

Abstract

HIV is rapidly evolving virus with high mutation rate. For heterologous gene expression system, the codon bias has to be optimized according to the host for efficient expression. Although DNA viruses show some correlation on codon bias with their hosts, HIV genes show very low correlation for both nucleotide composition and codon usage bias with its human host which limits the efficient expression of HIV genes. Despite this variation, HIV is efficient in infecting hosts and multiplying in large number. In this study, I have performed information theoretic analysis of nine genes of HIV-1 based on codon statistics of the whole HIV genome, individual genes and codon usage of human genes. For the HIV-1 whole genome sequence analyzed, it has been observed that the codon statistics of the third reading frame has the highest bias representing minimum entropy and hence maximum information. Similarly, despite being poorly adapted to the codon usage bias of human hosts, I have found that the Shannon entropies of nine genes based on overall codon statistics of HIV-1 genome are very similar to the entropies calculated from codon usage of human genes probably suggesting co-evolution of HIV-1 along with human genes.

Keywords: HIV-1 genome, Shannon entropy, codon usage bias, codon adaptation index, expected codon adaptation index

Introduction

Every organism has its own pattern of codon usage. All the synonymous codons for particular amino acid are not used equally. Some synonymous codons are highly expressed, whereas the uses of others are limited. The use depends upon species (Grantham et. al 1980) (Grantham et. al. 1981). Difference in codon usage has also been observed among genes of a same organism (Sharp et. at. 1988). Codon bias has been linked to specific tRNA levels that are mainly determined by the number of tRNA genes that code for a particular tRNA(Hershberg R, Petrov DA. 2008). The choice of codon has been shown to affect the expression level of genes. This can be seen in the expression pattern of transgenes. Gustafsson *et. al.* showed that use of particular codon can increase the expression of transgene by over 1,000 fold(Gustafsson et. al 2004). In bacteria it has been shown that gene expressivity correlates with codon usage (Gouy M., and C. Gautier. 1982). Although bacteriophages have been shown to have codons that are preferred by their hosts (Lucks et. al. 2008), however, codon usage pattern of RNA virus seems to differ with that of host (Jenkins G., Holmes E. 2003). Despite this variation HIV virus can effectively multiply on human T cells. Although codon usage of early genes (*tat, rev, nef*) show some correlation with that of human codon usage (van Weringh et. al. 2011), however, late genes show very little correlation. It raises a question of how such variation in codon usage still allows for efficient viral replication. *van Weringh et. al.* showed that there is difference in the tRNA pool of HIV-1 infected and uninfected cells. Although they speculated that HIV-1 modulates the tRNA pool of the host making it suitable for its genome to translate efficiently, however, the extent to which such modulation helps in efficient translation is yet to account for.

After Shannon published his ground breaking paper "A Mathematical theory of Communication" (Shannon, 1948), there has been several attempts in utilizing information theory in the context of living systems. Shannon used the term *information* differently than classical information theorists would have used. Here we have tried to use the information as mentioned by Shannon. According to Shannon, for a possible set of events with probability distribution given by $\{p_1, p_2, p_3, \ldots, p_n\}$ the entropy or uncertainity is given by,

$$H = -\sum_{1=1}^{n} p_i * log(p_i)$$

This is in fact observed entropy of a sequence with given probability distribution. H is the maximum when all p_i are equally likely. In this condition the information content is zero. The amount of information or 'negentropy' in a sequence then can be given as,

$$I = H_{max} - H_{obs}$$

where H_{obs} is the entropy obtained from given probability distribution (Brillouin, L. 1956). DNA consists of 4 nucleotides A, G, C, T whose distribution pattern varies among different species. Gatlin deduced information content based on this distribution pattern (Gatlin 1966) and using transition probability values obtained from

neighbor data (Josse et. al. 1961). Lately the information theoretic value of a given DNA sequence has been obtained using the Shannon formula as double sum (Zeeberg 2002),

$$H = \sum_{i=1}^{i=n_{aa}} \left(-\sum_{j=1}^{j=n_{syncod(i)}} P_{(i,j)} log_2 \left(P_{(i,j)} \right) \right)$$

where n_{aa} is the number of distinct amino acids, $n_{syncod(i)}$ is the number of synonymous codons (or microstates) for amino acid *i* (or macrostate) whose value range from 1 to 6, and $P_{(i,j)}$ is the probability of synonymous codon *j* for amino acid *i*. Also, we know that information is not absolute. It depends upon the environment. This means that the same sequence of DNA may represent different amount of information depending on what environment it is in or on the machinery that interprets the sequence. We exploit this fact to calculate Shannon entropy for 9 genes of HIV-1 based on codon distribution of viral genome, individual genes and that of its host – human codon usage frequency. Information is calculated based on the codon distribution for three possible reading frames. To the best of my knowledge, I believe such study has not been carried out yet. Movitation for this calculation is the fact that viruses show overlapping genes which are generally read by shifting the reading frame and are speculated to be present to increase the density of genetic information (Lamb and Horvath 1991). These genes are read by ribosomal frameshifting (Wilson et. al. 1988). For those nine genes, I have also calculated intrinsic entropy of the sequence which can be defined as the entropy based on its own codon usage (i.e. codon usage within the same gene) to compare with other entropy values.

For heterologous expression system such as viruses use host translational machinery for their replication. They are under evolutionary pressure to adapt to the host tRNA pool. To estimate a degree of evolutionary adaptiveness of host and viral codon usage, Codon Usage Index (CAI) can be used (Sharp & Li 1987). But for sequences with highly biased nucleotide composition, interpretation of CAI can be tricky (Grocock & Sharp 2002). So to know whether the given value of CAI are statistically significant and has arisen from codon preferences or they are merely artifacts of nucleotide composition bias, expected CAI (eCAI) can be used to set a threshold for comparison (Pugibo et. al. 2008).

Methodology:

All the DNA sequences were obtained from NCBI database in FASTA format. For each sequence, codon statistics was obtained by entering the given sequence on online Sequence Manipulation Suite (http://www.bioinformatics.org/sms2/codon_usage.html) and using standard genetic code as the parameter. The number and fraction of each possible codons were noted. The first nucleotide was deleted to shift the reading frame by +1 to include other possible codon patterns and again the number and frequency was noted. The process was repeated for +2 reading frame. Now as any of the reading frames can contain the gene of interest, all three reading frame statistics were used to calculate the Shannon entropy independently. The assumption made for the calculation is that reading the message occurs in a linear fashion without slippage of the reading frame (RF). Fraction was normalized for a particular amino acid but not with total number of codons. Thus,

$$\sum_{j=1}^{j=n_{syncod(i)}} P_{(i,j)} = 1$$

First a row matrix was constructed with fractions of synonymous codons used.

$$codon0 = [p_1, p_2, \ldots, p_{64}]$$

The fractions in the matrix thus constructed were treated as microstate to calculate Shannon entropy and thus another matrix was constructed consisting of Shannon entropy of each fraction distribution.

$$H = [-codon0(1) * log_2 codon0(1), \cdots, -codon0(64) * log_2 codon0(64)]$$

Total Shannon entropy for the sequence is then calculated as:

$$H_{gene} = \sum_{i=1}^{64} N_i * H_i$$

Here N_i is the total number of a particular codon present in the gene of interest. Such calculation was performed for all three RFs.

Correlation coefficient for each gene's codon statistics was calculated with human codon usage statistics. Correlation coefficients for two genes *vpr* and *vpu* were calculated again removing the codon data for which no amino acid is present in that gene. Then Shannon entropy was calculated for all nine genes using the human codon usage statistics. Intrinsic entropy which is the entropy based on own codon statistics of each gene was also calculated. Again the assumption is that there is no slippage of reading frame during translation of the message. Thus codon statistics for single reading frame starting with start codon was used to calculate intrinsic entropy. Similarly average entropy was calculated by averaging the fractions of synonymous codons for all three RFs. For the calculation of percentage overall GC content and position specific GC content of codons of nine genes and CAI values, http://genomes.urv.cat/CAIcal/ (Pugibo et. al. 2008) online site was used again using standard genetic code as the parameter. For calculation of expected codon adaptation index was performed in E-CAI server (http://genomes.urv.es/CAIcal/) using Markov chain and standard genetic code as the parameters. Human codon usage statistics was obtained from the http://genomes.urv.cat/CAIcal/CU_human_nature.html online site. Computations were performed in R.

Result and Discussion:

We see that there is some correlation between codon present in HIV-1 early genes (*tat, rev, nef*) and human codon usage (fig 1), but correlation is very low for other genes: *env, gag, pol, vpr, vif, vpr* (table 1). *vpu* and *vif* genes have the lowest correlation with human codon usage. The degree of correlation also differs for 3 RFs codon statistics and nine genes: high correlation probably suggesting the location of the particular gene at that particular RF.

Table 1. Correlation coefficients calculated among human codon usage, codon usage of HIV-1 genes and codon statistics for all three reading frames of HIV-1 genome

Genes	Correlation coefficients			
	Human codon	RF 0	RF +1	RF +2
env	0.48	0.78	0.74	0.87
gag	0.55	0.84	0.76	0.83
tat	0.72	0.88	0.96	0.91
rev	0.62	0.91	0.87	0.94
vpu	0.26 / 0.41	0.61	0.54	0.68
vpr	0.48 / 0.55	0.61	0.66	0.65
vif	0.44	0.79	0.74	0.76
nef	0.73	0.78	0.76	0.73
pol	0.57	0.84	0.82	0.89
Human	-	0.74	0.78	0.66

Table showing correlation between codon fractions of nine gene with human codon usage fraction (column 2) and also with three different reading frames. RF 0 represents the codon fractions of initial reading frame of HIV-1 sequence, RF +1 represents a codon fractions of single frame shift and RF +2 represents codon fractions of +2 frame shifts of the HIV-1 genome sequence. Human codon usage statistics is also compared with the three possible reading frames of HIV-1 genome where +1 shifted reading frames shows highest correlation. For *vpu* and *vpr* gene, number after backlash is calculated removing the data for which no codons are present for certain amino acids.

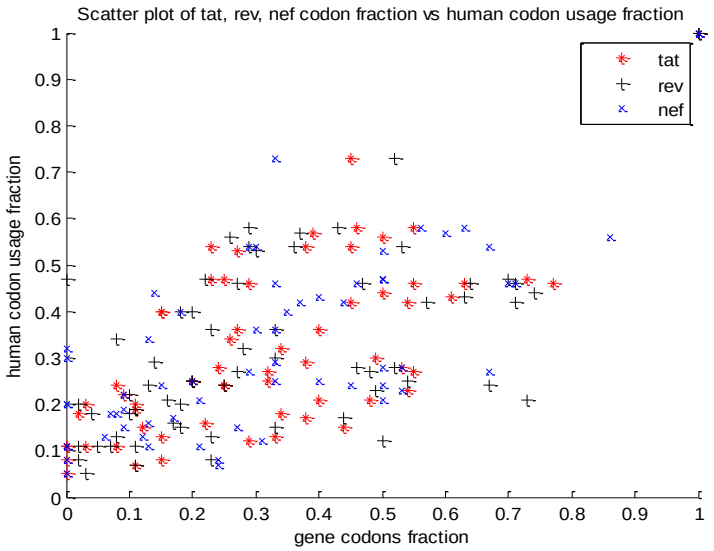

Fig 1: Scatter plot between codons fractions of *tat, rev* and *nef* genes of HIV-1 and human codon usage fraction with correlation coefficient of 0.72, 0.62 and 0.73 respectively. These three genes have higher correlation with human codon usage fraction and are also the early genes.

Entropic calculation shows a general trend for the sequences analyzed. First is that +2 frame shifted reading frame shows lower entropy as compared to two other reading frames. This marked distinction of the third reading frame among three possible reading frames of the sequence analyzed is surprising. As it has the lowest entropy among three reading frames, sequence with the codon usage pattern of third RF represents the highest information (in Shannon's sense). This probably suggests that there is genome wide conservation of codon usage for that reading frame but the reason for that is unclear. *env*, *rev*, *pol and vpu* genes have the highest correlation with the third reading frame as compared to other two reading frames. Similarly, *gag* and *vpr* also have a very high correlation coefficient. If we use codon statistics of third RF to calculate Shannon entropy, we get the minimum entropy and hence maximum information. But then again we run into problem as this third RFs shows the lowest correlation coefficient with human codon usage pattern. So there has to be a balance between these contrasts: maximizing information or maximizing correlation. Take *gag* for example, it shows high correlation with RF0 and RF2 both of which has lower correlation with human codon usage. This means that the expression of gag is affected by this choice of codons. It has been shown that ratio of native and optimized codons determine the HIV-1 *gag* expression (Kofman et. al. 2003). This also supports the speculation that codon bias leads to sub-optimal expression in infected cells. There is in fact good evidence that HIV-1 gene expression is not maximum-but is fine tuned to allow regulation of diverse processes (Marzio et. al. 2002). More evidence of sub-optimal expression is shown by the fact that when codon optimized genes that are better adapted to host tRNA pool were introduced, it led to higher expression(Haas et. al. 1996)(Anson & Dunning 2005)(Ngumbela et. al. 2008). From the entropic calculation based on human codon usage, we can see that *vpu* and *vif* have lowest entropy but as they have low correlation with human codon usage, their expression is limited. Codon optimization of these genes results in the increase of expression level (Nguyen et. al. 2004). However, high correlation does not imply that in fact the gene is in that reading frame. It is possible that such bias may or may not affect biological function, but it is likely that such distinction of lower entropy have some evolutionary importance.

Table 2: Entropies of HIV-1 genes based on various codon distributions

Genes	***pol***	***env***	***nef***	***vif***	***vpr***	***vpu***	***rev***	***tat***	***gag***
Entropy H	259.60	129.72	93.84	84.72	43.00	36.17	398.83	400.92	225.52
Entropy 0	264.11	131.32	94.79	87.81	42.86	38.53	399.91	399.71	229.41
Entropy +1	262.34	129.77	94.26	87.79	42.76	38.83	398.85	403.42	227.55
Entropy +2	256.70	125.54	92.12	85.16	41.27	37.15	388.22	390.52	222.10
Entropy I	229.54	110.10	91.23	75.25	38.21	27.32	378.79	397.98	216.69
Average entropy	264.48	130.58	94.79	87.95	42.77	38.69	400.59	402.66	229.09

Shannon Entropic values in bits for nine genes based of different codon statistics. Entropy H denotes entropic value based on human codon usage statistics. Entropy 0, Entropy +1 and Entropy +2 represent entropic values based on first reading frame, +1 shifted reading frame and +2 shifted reading frame of HIV-1 genome sequence respectively. Entropy I represents intrinsic entropy. Average entropy is calculated by averaging the codon statistics for three possible reading frames.

Intrinsic entropy differs greatly with other entropic values as it shows lowest values. Such low intrinsic entropies may have significance for free living organisms as lower entropies suggest higher bias. But for heterologous expression systems such as HIV-1, entropy H probably represents the best entropic values for the genes analyzed as host (human) gene usage codon statistics was used for the calculation. Average entropy, which is closer to entropy H, rather than intrinsic entropy gives a better representation for entropic value and hence for the amount of information a gene contain inside human host. Although there is great variation in the synonymous codon usage statistics between HIV-1 genes and human genes, we observe that the entropic values for the HIV-1 genes based on overall codon distribution of the HIV genome shows almost similar values as compared to the calculation based on human codon usage statistics (Table 2). Even for *vpu* gene which has a very low correlation coefficient (0.26), the entropic values based on overall codon statistics of HIV-1 genome and human codon usage statistics show similarity: 36.17 and 38.69 bits respectively. Even if we remove the data for which there is no single codon for a certain amino acid in that gene, correlation coefficient is still low. In *vpu* gene, codons for Cysteine, Threonine and Phenylalanine are completely absent. If we remove that data, we get a correlation coefficient of 0.41 which is still low. However, this removal does not affect entropic calculation. Similarly, for *vpr* gene, codons for Cysteine are completely absent. Removing that data new correlation coefficient obtained is 0.55 and average entropic values and entropy H are very close: 42.77 and 43.00 bits respectively. This might suggest that despite having a different nucleotide composition with human, HIV-1 viruses have co-evolved with human genes to represent the same level of information. HIV is a highly variable virus which undergoes rapid mutation. In order to maintain the overall codon statistics, it has to maintain the nucleotide composition which is determinant of codon bias (Bronson & Anderson, 1993).It has infact been shown that despite this variability, the biased nucleotide composition of HIV is constant over time (van der Kuyl & Berkhout 2012).

Besides having similar entropy, we can note that all the genes have high CAI values (Table 3) although varying degree of GC content. All the CAI values are greater than 0.6 with *vpu* having the lowest value of 0.62 whereas *tat* has the highest value of 0.77. Generally CAI well above 0.5 are considered to be showing good level of adaptation towards the host, however, care should be given while interpreting these values as they may not reveal the level of adaptation themselves. Such values may be due to the bias in nucleotide composition. So to know whether these values actually represent the adaptation we need to set a threshold set by bias in nucleotide composition so that we may say that CAI do indeed represent level of adaptation. For that expected CAI (eCAI) is calculated and compared (Pugibo et. al. 2008). From these comparisons, none of the genes seem to be well adapted to human codons usage pattern. We can note that the GC content of all the genes is below average. Also GC content of second and third nucleotides of codons show greatest variability. Thus we can conclude that nucleotides of these positions are the determinant of codon bias in HIV genes rather than the selection pressure for codons.

Table 3: Codon adaptation index (CAI) and GC content of HIV-1 genes.

Gene	Length	CAI	%GC	%GC1	%GC2	%GC3	eCAI ($p<0.05$)
gag	1503	0.73	44.0	50.3	43.5	38.3	0.74
nef	621	0.76	49.4	58.0	44.4	45.9	0.76
tat	2592	0.77	39.8	38.8	41.0	39.7	0.78
pol	1746	0.71	38.3	48.8	36.4	29.6	0.72
rev	2682	0.73	40.5	41.5	39.7	40.2	0.74
vif	579	0.72	42.0	46.6	41.5	37.8	0.74
vpr	291	0.73	45.0	54.6	34.0	46.4	0.74
vpu	249	0.62	37.8	54.2	31.3	27.7	0.66

CAI and eCAI values with overall GC percentage and position specific GC percentage for nine HIV-1 genes is reported. %GC 1, %GC2 and %GC3 represent GC percentage of first, second and third position of the codons respectively. Second and third position of the codon shows greatest bias in GC content as compared to first position (except for *tat* and *rev* gene which shows almost no bias for all three positions)

Conclusion:

Despite lots of studies, HIV viral genome still possesses several mysteries. It is true that HIV is evolving along with its human host. However, it is not clear why its nucleotide composition and synonymous codon usage bias differ greatly from its host. From comparison of CAI with eCAI, we can conclude that HIV genes are poorly adapted to

the tRNA pools of human. So it can be inferred that selection pressure on HIV to adapt to tRNA pools is minimal as compared to the rapid mutation it has on its genome (Jenkins & Holmes 2003). Because of this, HIV tends to evolve as a separate entity although there is selection pressure on different levels. It is not clear whether nucleotide composition bias can give rise to the asymmetry in the observed information content along three possible reading frames. However, despite having large differences in nucleotide composition and synonymous codon usage bias, HIV genes are seem to have evolved to represent the same level of information as obtained by the codon bias of human genes. How HIV is able to attain such uniformity despite being different from its host is yet another mystery this study has surfaced. Much work is needed in the future which can bring together the differences in one place to give a clear picture of evolution of HIV viral genome.

References:

Anson DS, Dunning KR. (2005). Codon-optimized reading frames facilitate high-level expression of the HIV-1 minor proteins. *Mol Biotechnol*, 31:85–88.

Brillouin, L. (1956). "Science and Information Theory". New York: Academic Press.

Bronson, E. C., & Anderson, J. N. (1994). Nucleotide composition as a driving force in the evolution of retroviruses. *Journal of molecular evolution*,*38*(5), 506-532.

Gouy M. and Gautier C. (1982). Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res*, 10: 7055-7074

Grantham, R., Gautier, C., Gouy, M., Mercier, R. & Pave, A. (1980) Codon catalog usage and the genome hypothesis**.** *Nucleic Acids Res*, 8r49–r62.

Grantham, R., Gautier, C., Gouy, M., Jacobzone, M., Mercier, R. (1981) Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Res*, 8r43–r74

Grocock RJ, Sharp PM. (2002). Synonymous codon usage in Pseudomonas aeruginosa PA01. *Gene*, 289: 131–139. 10.1016/S0378-1119(02)00503-6

Gustafsson C., Govindarajan S., Minshull J. (2004). Codon bias and heterologous protein expression. *Trends Biotechnol.* 22 346–35310.1016/j.tibtech.2004.04.006

Haas J, Park EC, Seed B. (1996). Codon usage limitation in the expression of HIV-1 envelope glycoprotein. *Curr Biol*, 6:315–324.

Hershberg R, Petrov DA (2008) Selection on codon bias. *Annu Rev Genet*, 42:287–299.

Ikemura, T. 1985 Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.* 213–34

Jenkins G, Holmes E (2003) The extent of codon usage bias in human rna viruses and its evolutionary origin. Virus Res 92: 1–7.

J. Josse, A.D. Kaiser, A. Kornberg (1961). Enzymatic synthesis of deoxyribonucleic acid: VIII. frequencies of nearest neighbor base sequences in deoxyribonucleic acid. *Journal of Biological Chemistry*, 236, 864-875.

Kofman A, Graf M, Bojak A, Deml L, Bieler K, et al. (2003) HIV-1 gag expression is quantitatively dependent on the ratio of native and optimized codons. Tsitologiia 45: 86–93.

Lamb RA, Horvath CM (1991). Diversity of coding strategies in influenza viruses. *Trends Genet* 7:261–266

Lucks JB, Nelson DR, Kudla GR, Plotkin JB (2008) Genome Landscapes and Bacteriophage Codon Usage. PLoS Comput Biol 4: e1000001. doi: 10.1371/journal.pcbi.1000001.

L.L. Gatlin (1966) The information content of DNA. *Journal of Theoretical Biology*, 10,281-300.

Marzio G, Vink M, Verhoef K, de RA, Berkhout B. (2002). Efficient human immunodeficiency virus replication requires a fine-tuned level of transcription. *J Virol*, 76:3084–3088

Ngumbela KC, Ryan KP, Sivamurthy R, Brockman MA, Gandhi RT, Bhardwaj N, et al. (2008) Quantitative effect of suboptimal codon usage on translational efficiency of mRNA encoding HIV-1 gag in intact T cells. *PLoS One*,3:e2356.

Nguyen, K. L. *et al.* (2004). Codon optimization of the HIV-1 *vpu* and *vif* genes stabilizes their mRNA and allows for highly efficient Rev-independent expression. *Virology* **319**, 163–175.

Puigbo P, Bravo IG and Garcia-Vallve S. (2008) CAIcal: a combined set of tools to assess codon usage adaptation. *Biology Direct, 3:38.*

Puigbo P, Bravo IG and Garcia-Vallve S. (2008) E-CAI: a novel server to estimate an expected value of Codon Adaptation Index (eCAI). *BMC Bioinformatics*, 9:65.

Shannon, C. (1948). A Mathematical Theory of Communication. *Bell Systems Technical Journal,* 27: 279-423, 623-656.

Sharp, P.M., Cowe, E., Higgins, D.G., Shields, D.C., Wolfe, K.H., Wright, F. (1988) Codon usage in *Escherichia coli*, *Bacillus subtilis*, *Saccharomyces cerevisiae*,*Schizosaccharomyces pombe*, *Drosophila melanogaster* and *Homo sapiens*; a review of the considerable within-species diversity. *Nucleic Acids Res,* 168207–8211

Sharp PM, Li WH. (1987) The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res,* 15:1281–1295.

van der Kuyl, A. C., & Berkhout, B. (2012). The biased nucleotide composition of the HIV genome: a constant factor in a highly variable virus.*Retrovirology*, *9*(1), 1-14.

van Weringh A, Ragonnet-Cronin M, Pranckeviciene E, Pavon-Eternod M, Kleiman L, Xia X (2011) HIV-1 modulates the tRNA pool to improve translation efficiency. *Mol Biol Evol*, 28:1827–1834.

Wilson W, Braddock M, Adam SE, Rathjen PD, Kingsman SM, Kingsman AJ (1988). HIV expression strategies: ribosomal frameshifting is directed by a short sequence in both mammalian and yeast systems. *Cell* 55:1159–1169

Zeeberg, B. (2002). Shannon information theoretic computation of synonymous codon usage biases in coding regions of human and mouse genomes. *Genome research*, *12*(6), 944-955.