Evolution of Relative Synonymous Codon Usage in Human Immunodeficiency Virus Type-1

Peter L. Meintjes, Allen G. Rodrigo

Bioinformatics Institute and School of Biological Sciences
University of Auckland
Private Bag 92019, Auckland, New Zealand

a.rodrigo@auckland.ac.nz

Abstract

Mutation in HIV-1 is extremely rapid, a consequence of a low-fidelity viral reverse transcription process. The envelope gene has been shown to accumulate substitutions at a rate of approximately 1% per year and can frequently spend a long time in the host (approximately 10 years). The relative synonymous codon usage (RSCU) in HIV-1 is known to be different from that of the human host. However, by reengineering the protein coding sequences of HIV-1 to reflect the RSCU patterns observed in humans, a large increase in protein expression is observed. It is reasonable to suggest that within a host there may be a selective drive for change in the RSCU of HIV-1 towards human RSCU.

To test this hypothesis we analysed HIV-1 partial envelope sequences from 8 patients sampled serially in time. For each sequence, an RSCU table was constructed. Sequences were labelled as "early" or "late" depending on whether they were sampled before or after the mid-point of the study. Using the RSCU values as descriptor variables, a Principal Components Analysis (PCA) was performed. The first three components clearly discriminated between early and late sequences. We also constructed pooled groupwise RSCU tables for early and late sequences. The viral RSCU values of each of the groups were correlated with human RSCU. If there is selection for host-adaptation in RSCU, we expect that "late" viral RSCUs would tend to be more highly correlated with human RSCU than "early" viral RSCUs. In fact, tests of significance suggest that this is the case. However, closer examination of the data revealed that the apparent trend towards human RSCU can be attributed to the homogenisation of the codon usage by mutation pressure rather than host adaptation.

Keywords: Codon Usage, RSCU, HIV-1, Principle Components Analysis

Introduction

1.1 Hypervariability of HIV-1

Human Immunodeficiency Virus type 1 (HIV-1) is a member of the lentivirus subfamily of retroviruses. These viruses are characterised by a post-infection asymptomatic period that can exceed 10 years. In the case of HIV, infection typically leads to Acquired Immune Deficiency Syndrome (AIDS). Envelope glycoproteins mediate important aspects of host-virus relationships and lentivirus pathogenesis (Dettin et al. 2003; Si et al. 2003). Amino acid sequences from envelope glycoproteins between different HIV-1 variants consist of a succession of variable and conserved regions. In some cases, as many as 30% of the amino acids may differ among variants (Gaschen et al. 2002). It is fairly clear that the error prone nature of the viral reverse transcriptase provides the biochemical basis to this variation (Zhang et al. 1997). Furthermore there is evidence that suggests HIV-1 is subject to a positive selection pressure imposed by the immune system (Williamson et al. 2003; Ross & Rodrigo 2002). Thus, by diversifying, the virus is able to escape detection.

1.2 Early versus late HIV-1 sequences

According to current literature, infection of a host is believed to be initiated by a single virion or by a small homogeneous population of virions (Learn et al. 2002). HIV-1 population kinetics studies have estimated that as many as 10^{10} new virions are produced daily within an infected individual (Perelson et al. 1996). Post-infection mutation has been estimated at approximately 1% per year at a relatively consistent linear rate (Shankarappa et al. 1999). The rapid mutation of HIV-1 leads to an accumulation of diversity, as much as 10% within an individual (Delwart et al. 1994; Shankarappa et al. 1999). This rapid mutation rate has grabbed the attention of evolutionary biologists, as it is possible within measurable time to follow the process of molecular evolution of the viral population. Such populations are termed measurably evolving populations or MEPs (Drummond et al. 2003), and to attain full value from them, we take multiple samples at regular intervals. By doing this we capture a sample of the changing diversity within the host population of viruses. We can visualise this accumulation of diversity of HIV-1 within a host using phylogenetic analysis. In Figure 1 for instance, we can clearly see a clustering of HIV-1 sequences from

early time points near the base and the positioning of sequences from later time points towards the "canopy". This illustrates that as infection progresses, the viral population diverges further from the founder sequence.

Figure 1: A Neighbour-Joining tree of partial envelope sequences from Patient 1, rooted with one of the early timepoints. Early timepoints are labelled with filled ovals and late timepoints are labelled with unfilled rectangles. We observe that the early timepoints cluster at the base, while late timepoints fan out towards the "canopy".

1.3 The RSCUs in humans and HIV are different from one another

There is a large element of redundancy within the genetic code. A direct result of the redundancy is the observation that codons that code for the same amino acid (synonymous codons) are very rarely used with equal frequency (Grantham et al. 1980). The patterns of codon usage varies among species, and also among genes from within the genome of a single species (Ikemura 1985, Sharp et al. 1988, 1995). Relative Synonymous Codon Usage (RSCU) measures the relative frequency that each codon is used to encode a particular amino acid. Through multivariate analyses on budding yeast *Saccharomyces cerevisiae*, it has been shown that the single major trend in its RSCU is its correlation to expression level (Sharp & Crowe 1991). Genes with similar expression levels all have similar codon biases. While there are a number of factors that contribute to translational efficiency, one would expect a virus that does not encode its own tRNAs (e.g. HIV) to share similar codon usage preferences with its host since they utilise the same set of tRNAs (i.e. host tRNAs). However, the RSCUs from HIV-1 are known to be quite different from that of the human (Grantham & Perrin 1986; Kypr & Mrazek 1987). Table 1 gives a pairwise comparison of all codons except the stop codons TAG, TGA, TAA and the unique codons ATG (methionine) and TGG (tryptophan). Relative frequencies (expressed as percentages) were taken from our data and from the codon table presented in the publication of the human genome (International Human Genome Sequencing Consortium 2001).

In contrast to yeast, the patterns of codon usage in different human genes have not been related (directly or indirectly) to any aspects of their expression (Sharp et al. 1995). However, experiments conducted by Haas et al. (1996) have shown, that not only are the RSCUs of HIV-1 and humans very different, but that reengineering of the protein coding sequences of both *gag* and *env* genes to be more like highly expressed human genes leads to an increase in protein expression.

AA	Codon	HIV-1	Human	AA	Codon	HIV-1	Human
Phe	**TTT**	66	46	**Ala**	**GCT**	16	26
	TTC	34	54		**GCC**	20	40
Leu	**TTA**	32	8		**GCA**	63	23
	TTG	3	13		**GCG**	1	11
	CTT	3	13	**Tyr**	**TAT**	71	44
	CTC	12	19		**TAC**	29	56
	CTA	16	7	**His**	**CAT**	54	41
	CTG	33	40		**CAC**	46	59
Ile	**ATT**	34	36	**Gln**	**CAA**	61	26
	ATC	13	48		**CAG**	39	74
	ATA	53	16	**Asn**	**AAT**	71	47
Val	**GTT**	13	18		**AAC**	29	53
	GTC	11	24	**Lys**	**AAA**	87	43
	GTA	73	12		**AAG**	13	57
	GTG	3	47	**Asp**	**GAT**	50	47
Ser	**TCT**	11	19		**GAC**	50	53
	TCC	10	22	**Glu**	**GAA**	77	43
	TCA	31	15		**GAG**	23	57
	TCG	2	6	**Cys**	**TGT**	81	45
	AGT	37	15		**TGC**	19	55
	AGC	9	24	**Arg**	**CGT**	1	8
Pro	**CCT**	32	29		**CGC**	0	19
	CCC	27	32		**CGA**	1	11
	CCA	36	28		**CGG**	1	21
	CCG	5	11		**AGA**	89	20
Thr	**ACT**	26	24		**AGG**	8	20
	ACC	13	36	**Gly**	**GGT**	15	17
	ACA	55	28		**GGC**	2	34
	ACG	6	12		**GGA**	50	25
					GGG	32	24

Table 1: A pairwise table of the RSCU values for the human and the pooled viral RSCU. Displayed as percentages, we observe some substantial differences.

Given,

(1) that there are difference in RSCUs between HIV-1 and its human host,

(2) that it is possible to obtain a several-fold increase in viral protein expression when HIV-1 sequences are reengineered to reflect human codon usage and,

(3) that rapid viral mutation generates significant viral mutation on which selection can act,

it is reasonable to hypothesise that within the relatively long infection, selection will favour a measurable shift in HIV-1 RSCU towards human RSCU. In this study we test this hypothesis.

Material and Methods

2.1 Sequence data

The eight patients in the study were all homosexual men infected with HIV-1 enrolled prospectively in the Multicentre AIDS Cohort Study (MACS). The sequencing methods and evolutionary genetics of the sample have been described by Shankarappa et al. (1999) and Ross and Rodrigo (2002). Serial-sampled sequences were obtained at approximately 8 month intervals with approximately 10 sequences at each timepoint. Samples were taken post-seroconversion (the point at which antibodies against HIV-1 are detectable) until death, the onset of AIDS or the conclusion of the study. To differentiate between early and late sequences, we define the mid-point as half the duration of the study, for that patient. All sequences are from the C2-V5 region of the *env* gene coding for the Gp120 coat protein, and were extracted from the blood plasma of the 8 patients. The total information from each of the 8 patients is summarised in Table 2.

2.2 Phylogenetics

Phylogenetic trees were constructed using the Neighbour-Joining method as implemented in PAUP* (Swofford 1996) using the default settings with the exception of DNA/RNA distances which were set to General time-reversible. Tree files were rooted by using an early timepoint as the outgroup and displayed using Treeview (Page 1996).

	Sequences	Number of Time-Points
Patient 1	87	10
Patient 2	132	14
Patient 3	106	10
Patient 5	160	16
Patient 6	98	10
Patient 7	107	9
Patient 8	119	13
Patient 9	113	12
Total	922	94 at an average of 9.8

Table 2: The available raw data: Genbank accession numbers AF137629 to AF137715, AF137766 to AF138163, AF138166 to AF138263 and AF138305 to AF138643.

2.3 RSCU extraction

We ran the sequence data through a Perl script to extract the RSCU values for every amino acid. Not all of the 64 possibilities are necessary in the context of the analysis. Stop codons are eliminated from analysis as well as ATG (coding for methionine) and TGG (coding for tryptophan). Methionine and tryptophan are only associated with a single codon, and thus do not exhibit any codon bias.

2.4 Statistical Analysis

All statistical analyses were performed using JMP v3.2.2.

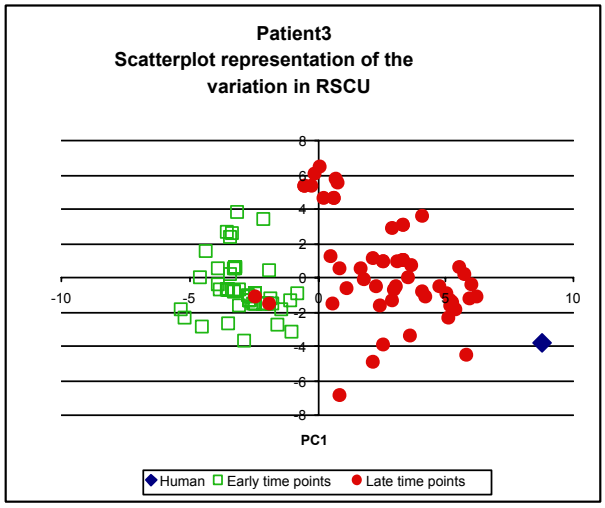

Figure 2: Early sequences and late sequences were differentially labelled with unfilled squares and filled circles respectively to display the results of the PCA. From the figure we can see that the filled circles (late sequences) and unfilled squares (early sequences) are spatially separate from one another. The diamond is the position of the Human RSCU value.

Results

3.1 Exploring the data

We used Principle Components Analysis (PCA) to explore the patterns of variation in the RSCU values of every sequence. PCA generates a linear combination of the descriptor variables (RSCU values) that best summarise the variation in the data. We used the first three principle components as axes and plotted the values associated with each sequence in three dimensions. Each point in three dimensions represents the variation of a particular viral sequence. The first three principle components described approximately 30% of the variation in each patient. The sequences were divided into early or late, depending on whether they were sampled before or after the mid-point of the study. Figure 2 shows the results of the PCA for patient 3, with the early and late sequences clearly distinct from one another. This pattern of variation was observed in all eight patients. Furthermore, when the human RSCU (diamond) is added to assess the direction of the selective drive, it appears

that the late RSCUs correlate more closely to the human RSCUs than do early RSCUs.

3.2 Can we quantify this trend?

To address the apparent increase in correlation of RSCUs from early to late towards the human, we constructed pooled groupwise RSCU tables for early and late sequences. For each patient, and for both early and late grouped RSCU values, we performed a Pearson correlation with the human RSCU to give two correlation coefficients. These results, displayed in Table 3, show that in 6 out of 8 patients the correlation coefficient increased from early to late. We performed two statistical tests on the correlation to assess the significance. Performing a *t*-test gave us a *p*-value of 0.0685 while performing a Wilcoxon Signed-Rank test gave us a *p*-value of 0.055. While not conclusive, both *p*-values provide suggestive evidence that there is a movement towards the human RSCU from early to late sequences.

Patient	Early	Late	Direction
1	0.342	0.3357	rev
2	0.3157	0.3264	**fwd**
3	0.282	0.3102	**fwd**
5	0.3093	0.3343	**fwd**
6	0.2896	0.273	rev
7	0.3412	0.3499	**fwd**
8	0.3388	0.3483	**fwd**
9	0.3235	0.4039	**fwd**

Table 3: The correlation coefficient values from a Pearson's correlation of the human RSCU to the early and late RSCUs respectively. An increase in correlation coefficient is indicated as "forward".

3.3 Is this correlation a statistical or biological phenomenon?

We are concerned that the trend we observed may be a statistical artifact. To examine in more detail the origin of the apparent change in correlation coefficient over time, we took a closer look at the formula for the correlation coefficient. For two variables, *x* and *y*, (HIV-1 RSCUs and human RSCUs respectively), *r*, the correlation coefficient, is proportional to the covariance of *x* and *y* (S_{xy}), and inversely proportional to the individual standard deviations of *x* and *y* (S_x and S_y).

$$r = \frac{S_{xy}}{(S_x)(S_y)}$$

If the data support our hypothesis of a directional change in the RSCU the increase in correlation coefficient from early to late must be due to the covariance, because this is a measure of association. We tested for equality of covariances with an *F* test (df 7) and obtained a *p*-value 0.398. This does not allow us to reject the null hypothesis of equal covariances indicating that the increase in correlation coefficient must be due to other factors.

	std dev early	std dev late	std dev human
Patient 1	0.2544	0.2498	0.1668
Patient 2	0.2564	0.2423	0.1668
Patient 3	0.2695	0.2585	0.1668
Patient 5	0.2703	0.2655	0.1668
Patient 6	0.2681	0.2740	0.1668
Patient 7	0.2539	0.2525	0.1668
Patient 8	0.2719	0.2689	0.1668
Patient 9	0.2508	0.2451	0.1668

Table 4: In each of the rows, the standard deviation of the viral RSCUs decreased except in patient 6.

If there is no significant difference between the covariances, but there is a significant difference in the correlation coefficients, that significant difference must be due to S_x, as S_y is constant for both correlations. A decrease in S_x for the viral RSCU would increase the correlation coefficient. However, a decrease in S_x would also imply that the codons in the viral RSCU are being used more evenly (i.e. the codon usage is becoming more homogeneous). On inspection of standard deviations, we observe that 7 out of the 8 patients showed a decrease (Table 4). Furthermore, the null hypothesis that the standard deviations were equal can be rejected under a *t*-test with a *p*-value of 0.038.

Amino Acid	Codon	Early	Late	Human
ARG	**CGT**	0.000	0.018	0.08
	CGC	0.000	0.024	0.19
	CGA	0.006	0.034	0.11
	CGG	0.009	0.051	0.21
	AGA	0.893	0.779	0.2
	AGG	0.091	0.095	0.2
	Std. Dev.	0.3576	0.30117	0.0554

Table 5: The RSCU values for the six arginine codons, and the standard deviation are presented for early, late and human sequences.

Using arginine as an example, we can see how the extreme bias observed in early sequences is homogenised with time to give us RSCU values with a lower standard deviation (Table 5). The RSCU values of each codon within the amino acid arginine show how, with time, we see a change in the RSCUs. Two codons that were never used before are now being used (CGT and CGC) and small increases in the RSCU values for a further 3 codons (CGA, CGG and AGG)) are observed. The final codon (AGA) showed a decrease from its extremely high relative value of 0.89 to 0.78. This shows an overall narrowing of the distribution of values that covers the RSCU for arginine as illustrated in the row of standard deviations. If we consider that random mutation acts on the viral RNA sequence, we would expect by chance alone that these sequences would accumulate synonymous mutation. This leads to a decrease in the bias. This homogenisation of the codon usage seems now to be the best explanation of the observed data.

Discussion

Our results show that there is a change in RSCU between early and late sequences and in most cases in the direction of the human RSCU. Codon usage in HIV-1 is quite different from that of human genes (Grantham & Perrin 1986; Kypr & Mrazek 1987) and three rules have been suggested to explain the patterns observed in HIV-1 (Haas et al. 1996). First, preferred codons maximize the number of A residues in the viral RNA. Second, T is preferred over C whenever there is degeneracy of pyrimidines at the third residue. Third, the dinucleotide CG is highly under-represented. It is a characteristic of lentiviruses, but not of all retroviruses to have a striking under-representation of CpG duplets. CpG duplets have been implicated in methylation (Eden & Cedar 1994). It has been suggested that the low levels of CpG may be a mechanism of avoiding silencing through methylation, but it would seem that this is at the cost of lowering the translational efficiency. Together, these ideas seem plausible and may account for the long period of asymptomatic infection observed not only in HIV-1 but in all lentiviruses.

Arginine best serves to illustrate the RSCU patterns observed in HIV-1. We see that the most highly expressed codon for arginine in the early sequences is AGA (89%) while codons containing the dinucleotide CG are largely unused. This illustrates rules one and three as this is the only codon with two A residues and also does not contain CG. In the human sequences we observe a relatively even spread of the codon usage for arginine. While the late sequences have a codon usage bias that adheres to the rules set out by Haas et al. (1996), we also note that in each of the codons, the codon usage bias is less severe than for the early sequences. We have two possible explanations for our observations.

The first explanation is that the RSCU of HIV-1 has become more like that of the human RSCU through a selective drive to increase translational efficiency later in infection, but that we have simply failed to detect it. In favour of this idea is the observed increase in expression noted by Haas et al. (1996) when the codons of HIV-1 were reengineered to resemble those of the human. In addition it has also been suggested that codon bias may allow the expression of viral proteins such as the *env* gene to be suppressed in order to minimise the antigenic profile. This may be a necessary step in the lentiviral disease progression. Perhaps, if the virus is too antigenic before there is sufficient variation amongst the viral population, the immune system would be able to control the disease. By extension, we might say that the necessity of maintaining a low antigenic profile may eventually cease to exist and thus lead to highly pathogenic viral forms completely overwhelming the immune system.

The second explanation is that the observed change in RSCU from early to late sequences is merely a result of mutation pressure on the viral sequence. We observe this through the close analysis of the correlation coefficient. This informs us that while there is a suggestion of directional change, the lowering of the standard deviation of the RSCUs is the cause of the observed increase in correlation coefficient. Thus it appears that the best explanation of our observation of directional change is most likely a statistical artefact caused by the homogenisation of the RSCU.

Acknowledgements

We acknowledge valuable comments from Alexei Drummond and grants from the United States Public Health Service.

References

International Human Genome Sequencing Consortium, (2001) Initial sequencing and analysis of the human genome *Nature,* **409,** 860-921.

Delwart, E., Sheppard, H., Walker, B., Goudsmit, J. and Mullins, J. (1994) Human immunodeficiency virus type 1 evolution in vivo tracked by DNA heteroduplex mobility assays *Journal of Virology,* **68,** 6672-6683.

Dettin, M., Zanchetta, M., Pasquato, A., Borrello, M., Piatier-Tonneau, D., Di Bello, C. and De Rossi, A. (2003) CCR5 N-terminus peptides enhance X4 HIV-1 infectioni by CXCR4 up-regulation *Biochemical and Biophysical Research Communications,* **307,** 640-646.

Drummond, A., Pybus, O., Rambaut, A., Forsberg, R. and Rodrigo, A. (2003) Measurably Evolving Populations *TRENDS in Ecology and Evolution,* **18,** 481-488.

Eden, S. and Cedar, H. (1994) Role of DNA methylation in the regulation of transcription *Current Opinion in Genetics and Developement,* **4,** 255-259.

Gaschen, B., Taylor, J., Yusim, K., Foley, B., Gao, F., Lang, D., Novitsky, V., Haynes, B., Hahn, B., Bhattacharya, T. and Korber, B. (2002) Diversity Considerations in HIV-1 Vaccine Selection *Science,* **296,** 2354-2360.

Grantham, P. and Perrin, P. (1986) AIDS virus and HTLV-I differ in codon choices *Nature,* **319,** 727-728.

Grantham, R. (1980) Codon usage and the genome hypothesis *Nucleic Acids Research,* **8,** 49-62.

Ikemura, T. (1985) Codon usage and tRNA content in unicellular organisms *Molecular Biology and Evolution,* **2,** 13-34.

Kypr, J. and Mrazek, J. (1987) Unusual codon usage in HIV *Nature,* **327,** 20.

Learn, G., Muthui, D., Brodie, S., Zhu, T., Diem, K., Mullins, J. and Corey, L. (2002) Virus population homogenization following acute human immunodeficiency virus type 1 infection *Journal of Virology,* **76,** 11953-11959.

Page, R. (1996) TREEVIEW: An application to display phylogenetic trees on personal computers *Computer Applications in the Biological Sciences,* **12,** 357-358.

Perelson, A., Neumann, A., Markowitz, M., Leonard, J. and Ho, D. (1996) HIV-1 dynamics in vivo: virion clearance rate, infected cell lifespan, and viral generation time *Science,* **271,** 1582-1586.

Ross, H. and Rodrigo, A. (2002) Immune-Mediate Positive Selection Drives Human Immunodeficiency Virus Type 1 Molecular Variation and Predicts Disease Duration *Journal of Virology,* **76,** 11715-11720.

Shankarappa, R., Margolick, J., Gange, S., Rodrigo, A., Upchurch, D., Farzadegan, H., Gupta, P., Rinaldo, C., Learn, G., He, X., Huang, X.-L. and Mullins, J. (1999) Consistent Viral Evolutionary Changes Associated with the Progression of Human Immunodeficiency Virus Type 1 Infection *Journal of Virology,* **73,** 10489-10502.

Sharp, P., Crowe, E., Higgins, D., Shields, D., Wolfe, K. and Wright, F. (1988) Codon usage patterns in *Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster* and *Homo sapiens*: a review of the considerable within species diversity. *Nucleic Acids Research,* **16,** 8207-11.

Sharp, P., Averof, M., Lloyd, A., Matassi, G. and Peden, J. (1995) DNA sequence evolution: the sounds of silence *Phil. Trans. R. Soc. Lond. B,* **349,** 241-247.

Sharp, P. and Crowe, E. (1991) Synonymous codon usage in *Saccharomyces cerevisiae Yeast,* **7,** 647-678.

Si, Z., Phan, N., Kiprilov, E. and Sodroski, J. (2003) effects of HIV Type 1 Envelope Glycoprotein Proteolytic Processing on Antigenicity *AIDS RESEARCH and HUMAN RETROVIRUSES,* **19,** 217-226.

Swofford, D. (1996), PAUP*4b10 Phylogenetic Analysis Using Parsimony and other methods. Sinauer, Sunderland, Massachusetts.

Williamson, C., Morris, L., Maughan, M., Ping, L., Dryga, S., Thomas, R., Reap, E., Cilliers, T., Van Harmelen, J., Pascual, A., Ramjee, G., Gray, G., Johnston, R., Karim, S. and Swanstrom, R. (2003) Characterisation and Selection of HIV-1 Subtype C Isolates for Use in Vaccine Development *AIDS RESEARCH and HUMAN RETROVIRUSES*

Zhang, L., Diaz, R., Ho, D., Mosley, J., Busch, M. and Mayer, A. (1997) Host-Specific Driving Force in Human Immunodeficiency Virus Type 1 Evolution In Vivo *Journal of Virology,* **71,** 2555-2561.