JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 14, Number 8, 2007

Pp. 1105–1114
DOI: 10.1089/cmb.2007.0073

Estimating the Relative Contribution of dNTP Pool Imbalance and APOBEC3G/3F Editing to HIV Evolution *In Vivo*

KOEN DEFORCHE,[1] RICARDO CAMACHO,[2] KRISTEL VAN LAETHEM,[1]
BETH SHAPIRO,[3] YVES MOREAU,[1] ANDREW RAMBAUT,[4]
ANNE-MIEKE VANDAMME,[1] and PHILIPPE LEMEY[3]

ABSTRACT

The human immunodeficiency virus (HIV) has a genome that is rich in adenine, and its rapid evolution shows an observed bias of guanine (G) to adenine (A) mutations. Two mechanisms have been proposed to explain these properties: (1) an imbalance in dNTP pool concentrations which drives the misincorporation process during reverse transcription, and (2) cytidine deamination by the APOBEC3G/3F restriction factor, causing G to A mutations most notably in specific dinucleotide contexts. Although crucial to understanding HIV evolution, current estimates on misincorporation bias during the replication cycle are based on scarce *in vitro* measurements. In this work, HIV partial *pol* sequences obtained for drug resistance testing purposes are analyzed using likelihood methods to estimate various models of HIV misincorporation bias *in vivo*. The technique is robust to selection on the amino acid sequence and selection against CpG dinucleotides. A model where misincorporations are explained only by an imbalance in dNTP pool concentrations, together with a preference for transitions versus transversions, explained 98% (95% confidence interval [C.I.] 93–100) of the observed variation in freely estimated misincorporation rates. Although dinucleotide context was responsible for variation in misincorporation probabilities, this variation was not specific for G to A mutations implying that the footprint of APOBEC3G/3F editing could not be detected. These results indicate that an imbalance in dNTP pool concentrations explains most of the bias in HIV nucleotide misincorporations, while the effect of editing by APOBEC3G/3F on HIV evolution, based on its dinucleotide specificity, could not be observed in this study.

Key words: algorithms, evolution, sequence analysis, viruses.

[1]Rega Institute for Medical Research, Katholieke Universiteit Leuven, Leuven, Belgium.
[2]Molecular Biology Laboratory, Centro Hospitalar de Lisboa Ocidental, Lisbon, Portugal.
[3]Institute for Evolutionary Biology, Oxford University, Oxford, United Kingdom.
[4]Institute for Evolutionary Biology, University of Edinburgh, Edinburgh, United Kingdom.

1. INTRODUCTION

THE RAPID EVOLUTION of the human immunodeficiency virus (HIV) is caused by a rapid turnover of the large intra-host population with a high mutation rate per generation. This is a major obstacle for designing effective therapeutic strategies (Rambaut et al., 2004) as it has caused a large genetic diversity in the HIV pandemic, complicating vaccine design, and as it causes drug resistance to evolve quickly during antiviral therapy.

Evolution can be considered as a two-step process of misincorporation followed by selection. For example in the context of evolution of antiviral drug resistance within a HIV-infected patient, the rate at which resistance mutations are expected to appear depends on the probability of generating the required nucleotide mutations (van de Vijver et al., 2006), as well as on the fitness gain of these amino acid mutations in presence of treatment.

HIV is a retrovirus that encodes its genomic information as diploid single stranded RNA (ssRNA), which is copied during the replication process by the viral reverse transcriptase (RT) to ds-cDNA and integrated in the host genome. Subsequently, the proviral DNA is transcribed by the host RNA pol II into messenger ssRNA, which is used as a template for protein translation or serves as the RNA genome in new virus particles. The high mutation rate can be partly explained by the the negligible proof reading capability of the two enzymes involved (O'Neil et al., 2002), leading to many misincorporations during the replication process. While G $\rightarrow$ A has been described as the most frequently observed substitution (Vartanian et al., 1994), no reliable quantitative estimates exist for the 12 different misincorporation probabilities. Current estimates have been drawn from scarce *in vitro* measurements of the fidelity during the HIV replication cycle, mainly from the study by Mansky and Temin (1995). In this study, only 42 nucleotide changes were observed in total and some mutations were even not observed, leading to rather unreliable esimates of the 12 probabilities. Furthermore, the *in vitro* infidelity of replication may not be representative for the *in vivo* environment, for example due to a potentially different composition of the *in vivo* dNTP pools (Jamburuthugoda et al., 2006), which are affected by cell metabolism and transcriptional activity.

Like many lentiviruses, HIV has an unusual genome nucleotide composition that is highly rich in adenine, but poor in guanine or cytosine (Berkhout and Hemert, 1994). Besides intrinsic properties of the RT and RNA pol II enzymes, which may cause a bias in their misincorporation rates, two hypotheses have been proposed to explain the adenine richness of lentiviral genomes, in view of the large number of G $\rightarrow$ A substitutions. On the one hand, an imbalance in intracellular dNTP concentrations, with in particular a low amount of dCTP and high amount of dTTP, was suggested to cause frequent misincorporation by RT of dTTP instead of dCTP in the minus strand, resulting in a G $\rightarrow$ A mutation in the plus strand (Vartanian et al., 1994; Balzarini et al., 2001). On the other hand, the cellular proteins APOBEC3G/3F have been shown to cause long stretches of G $\rightarrow$ A hypermutation in the HIV genome and is assumed to act as an antiviral defense mechanism (Lecossier et al., 2003). These enzymes are sensitive to the nucleotide context to carry out their editing, and cause preferentially GpA $\rightarrow$ ApA and GpG $\rightarrow$ ApG mutations.

Our work stems from the need for an accurate estimate of the misincorporation rates, to reverse engineer the fitness impact of observed patterns of substitutions during treatment (Deforche et al., 2007). We present a new likelihood-based computational method to estimate relative misincorporation rates from substitutions in a protein coding gene region, under the assumption that selection acts mostly on the amino acids, and against CpG dinucleotides (Shpaer and Mullins, 1990). We use the method to estimate several models of HIV-1 misincorporation probabilities from a large set of longitudinal *pol* population sequences collected for antiviral resistance testing.

2. METHODS

Notation

A codon C is defined given its three bases C_1, C_2, and C_3. The amino acid, into which the codon is translated, is given by $A(C)$. A nucleotide substitution from base X to base Y is denoted by $s(X \rightarrow Y)$, and the misincorporation rate from base X to base Y is denoted by $r(X \rightarrow Y)$. A codon C that is modified by a substitution m at codon position i, $s(C_i \rightarrow C_i')$ is denoted by $C' = s(C)$. Two non-synonymous nucleotide substitutions $s_1(C_i \rightarrow C_i')$ and $s_2(C_j \rightarrow C_j')$ at codon C are defined to be *equivalent* if

$A(s_1(C)) = A(s_2(C))$. In addition, all synonymous nucleotide substitutions are considered mutually *equivalent*, since they have no effect on the amino acid sequence. For serially sampled sequence data, C^1 and C^2 were used to denote the same codon in the baseline and follow-up sequence. When codons C^1 and C^2 differed by only one nucleotide substitution at position i, then we defined $s(C^1, C^2) = s(C_i^1 \to C_i^2)$. We may distinguish between substitutions and misincorporation rates not only by the nucleotides involved, X and Y, but also by other nucleotide sequence context information (such as changes in CpG dinucleotide content or the downstream nucleotide). Where that is the case, we use the notations s^* and r^*. For a base X, dXTP denotes the relative dNTP pool concentration for that base.

Likelihood functions

Observed nucleotide substitutions were considered as the outcome of probabilistic experiments, when (1) they were the sole observed nucleotide substitution at a codon and (2) there were no ambiguities at the codon. The first condition avoids to deal with the uncertainty of the intermediate amino acids when the multiple nucleotide substitutions (which were less than 2% of all observed nucleotide substitutions) were not simultaneous. Ambiguities, denoted using IUPAC ambiguity notation, are not uncommon in these sequences, and reflect nucleotides polymorphisms in the population. The second condition reflects that only fixation events are analyzed within the HIV intra-host population.

For each non-synonymous nucleotide substitution, the probabilistic experiment that was considered at codon C had, as its set of different outcomes, S_C, all nucleotide substitutions that were equivalent with the observed nucleotide substitution (Fig. 1):

$$S_C = \{s : A(s(C^1)) = A(C^2)\}.$$

If there were no equivalent nucleotide substitutions possible and thus only a single mutation can be observed for the relevant amino acid substitution ($|S_C| = 1$), the experiment had only one possible outcome, and provided no information to infer misincorporation rates. Otherwise, the probability $P_C^{\text{ns}}(s)$ of observing each of the equivalent nucleotide substitutions s at codon C was considered to be proportional to the misincorporation rate of such a mutation:

$$P_C^{\text{ns}}(s^*(X \to Y)) = K\, r^*(X \to Y)$$

with K a rescaling constant to make P_C^{ns} a proper distribution.

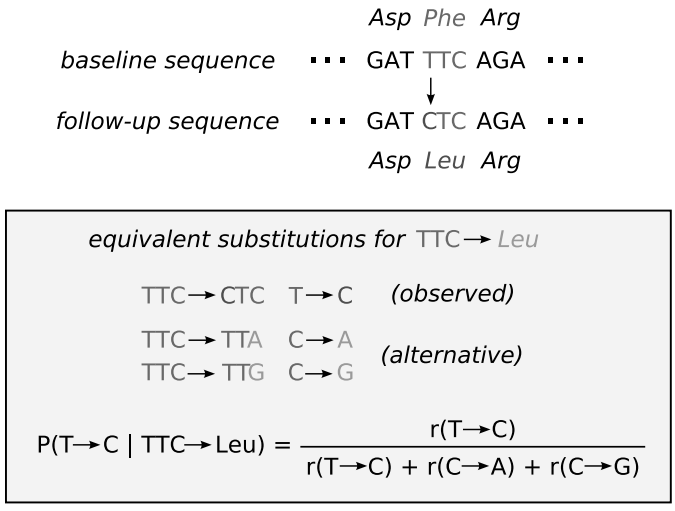

FIG. 1. An observed non-synonymous substitution (T → C) provides information of the underlying misincorporation rates, by considering the set of nucleotide substitutions that are *equivalent* at the amino acid level. In the illustrated case, next to the observed substitution, two alternative nucleotide substitutions would have generated the same amino acid substitution (TTC → *Leu*). Assuming that selection acts mostly on the amino acid level, the observed nucleotide substitution was the outcome of a probabilistic experiment, where the probability of observing any of the equivalent nucleotide substitutions is proportional to the misincorporation rate.

Similarly, since all synonymous substitutions were considered equivalent, each observed synonymous nucleotide substitution was the outcome of a probabilistic experiment with as set of outcomes all twelve nucleotide substitutions. The probability for each synonymous substitution s^* to be observed at codon C was defined as

$$\begin{aligned} P_C^{\mathrm{s}}(s^*(X \to Y)) &= P^{\mathrm{s}}(s^*(X \to Y)) \\ &= K'\, r^*(X \to Y)\; N^*(X \to Y) \end{aligned}$$

with $N^*(X \to Y)$ the number of nucleotides X in the baseline sequence (within the same context) for which a substitution to Y would have been synonymous, and K' a scaling constant to make P_C^{s} a proper distribution.

In general, given a number of model parameters $p_1, \ldots, p_n$, from which the probability for each misincorporation in every considered context may be derived, we may write the likelihood function of the data given the model:

$$\begin{aligned} \mathcal{L}(D|\theta) &= \mathcal{L}(D|p_1, \ldots, p_n) \\ &= \prod P_C^{\mathrm{ns}}(s^*(C^1, C^2)) \prod P_C^{\mathrm{s}}(s^*(C^1, C^2)) \end{aligned}$$

where the first product iterates over all observed non-synonymous nucleotide substitutions, and the second product iterates over all observed synonymous nucleotide substitutions.

We used this technique to discriminate between different mutations, not only based on the two nucleotides involved ($X \to Y$), but also taking into account other contextual features of the nucleotide sequence. In this way, we distinguished between mutations which did not affect CpG dinucleotide content of the nucleotide sequence, and those that added or removed a CpG dinucleotide. We also determined variation in misincorporation probability for a particular mutation $X \to Y$ based on the downstream nucleotide Z ($X\mathrm{p}Z \to Y\mathrm{p}Z$). The following models were estimated (with n the number of parameters in the model):

- *free ($n = 13$)*: 11 relative misincorporation rates ($r(X \to Y)$), and 2 independent factors which adjust the rate for adding (CpG^+) or removing (CpG^-) a CpG dinucleotide. The twelfth misincorporation rate was computed from the additional constraint that $\prod_i r_i = 1$.

$$r^*(X \to Y) = Dr(X \to Y)$$

  with $D = 1$ when CpG dinucleotide content is unchanged, $D = \mathrm{CpG}^+$ when adding a CpG dinucleotide, and $D = \mathrm{CpG}^-$ when removing a CpG dinucleotide.
- *dNTP ($n = 6$)*: 3 relative dNTP concentrations, transition/transveration bias (κ), CpG^+ and CpG^-. The fourth dNTP concentration was computed from the additional constraint that $\prod_X \mathrm{d}X\mathrm{TP} = 1$. From the 4 relative dNTP concentrations and κ, the 12 relative misincorporation rates are computed based on the principle that they are driven by the relative abundance of the corresponding dNTPs during positive strand synthesis:

$$r^*(X \to Y) = Dk\frac{\mathrm{d}Y\mathrm{TP}}{\mathrm{d}X\mathrm{TP}}$$

  where $k = \kappa$ for a transition, and $k = 1$ for a transversion.
- *free-dinucleotide ($n = 16$)* for all 12 mutations $U \to V$: 10 misincorporation rates (excluding $U \to V$), 4 relative misincorporation rates for $U\mathrm{p}Z \to V\mathrm{p}Z$ depending on the downstream nucleotide Z, and CpG^+ and CpG^-.

$$\begin{aligned} r^*(X\mathrm{p}Z \to Y\mathrm{p}Z) &= Dr(X \to Y); && \text{for } X \neq U, Y \neq V \\ &= Dr(U\mathrm{p}Z \to V\mathrm{p}Z) && \text{for } X = U, Y = V \end{aligned}$$

Maximum likelihood analysis

The parameters of the likelihood functions were estimated by Maximum Likelihood (ML) analysis, performed in the R statistical package (R Development Core Team, 2004) and maximum likelihood estimates with 95% likelihood intervals were reported.

Markov Chain Monte Carlo analysis

The likelihood function for the *dNTP* model was also used in a Bayesian Markov Chain Monte Carlo (MCMC) framework, to model the joint posterior distributions of the parameters with flat priors. The MCMC samples were used to computed also the posterior distributions of the model parameters and the derived relative rates. Each posterior distribution was summarized using its posterior mean and 95% Highest Probability Density (HPD) interval, using the R statistical package (R Development Core Team, 2004).

Data

Clinical data were pooled from the Stanford HIV Drug Resistance Database (Kantor et al., 2001), from the University Hospitals, Leuven, Belgium, and from Hospital Egas Monis, Lisbon, Portugal, to create a large data set of 5614 pairs of consecutive partial *pol* population sequences with an average length of 1084 bps. These sequences reflect the consensus sequence of the intra-host population. When more than 2 longitudinal sequences were available for a patient, $n-1$ consecutive pairs were derived from the n longitudinal sequences. The data set was curated by creating a phylogenetic tree for these sequences with PAUP (Swofford, 2000), excluding codons at resistance positions 30, 46, 48, 50, 54, 71, 82, 84, 88, 90 in protease and positions 41, 44, 65, 67, 70, 74, 75, 77, 100, 103, 106, 108, 115, 116, 151, 181, 184, 188, 190, 210, 215, 219 in RT (Johnson et al., 2005) to avoid the confounding effect of convergent evolution on the tree reconstruction (Lemey et al., 2005). Sequence pairs that did not cluster together in the tree could be an indication of contamination or super-infection, and therefore 450 sequence pairs were excluded from the analysis. The 5164 remaining sequence pairs had a median time of 469 days (25–75% quantile range: 231–894) between baseline and follow-up sample, and a mean genetic distance of 0.017 nucleotide substitutions per site. Protease positions 1 to 13 were excluded because this region overlaps with the *gag* open reading frame. In these pairs, a total of 27,155 nonambiguous nucleotide substitutions were observed. Of these, 2269 non-synonymous and 9522 synonymous mutations provided information to estimate misincorporation rates.

3. RESULTS

A method is presented to estimate misincorporation rates from observed substitutions in serially sampled sequences (Fig. 1). Nucleotide substitutions were considered *equivalent* when they had exactly the same effect at the amino acid level, i.e., either changing to the same amino acid or both a silent substitution. The observed nucleotide substition was considered only against alternative equivalent substitutions. Because selection is assumed to act on phenotypic changes at the amino acid level, and is therefore blind for equivalent substitutions, the method was robust with respect to selection at the amino acid level. Each observed substitution in a codon at which only a single nucleotide substitution was observed, was considered as the outcome of a probabilistic experiment with probabilities proportional to the misincorporation rates of each of the equivalent substitutions. A misincorporation model is then estimated by applying this process for many observed nucleotide substitutions in a maximum likelihood or Markov Chain Monte Carlo framework.

Several models for HIV-1 *in vivo* misincorporation rates were estimated using 5164 partial *pol* population sequence pairs, obtained for resistance testing.

A first model allowed unrestricted estimates of the misincorporation rates (model *free* in Fig. 2). A large variation is observed: the most frequent mutation (G to A) is approximately 30 times more probable than the least frequent mutation (C to G). In general, transitions were more frequent than transversions. The most common transversions were C $\rightarrow$ A and T $\rightarrow$ A. When combining rates per template base, we obtain

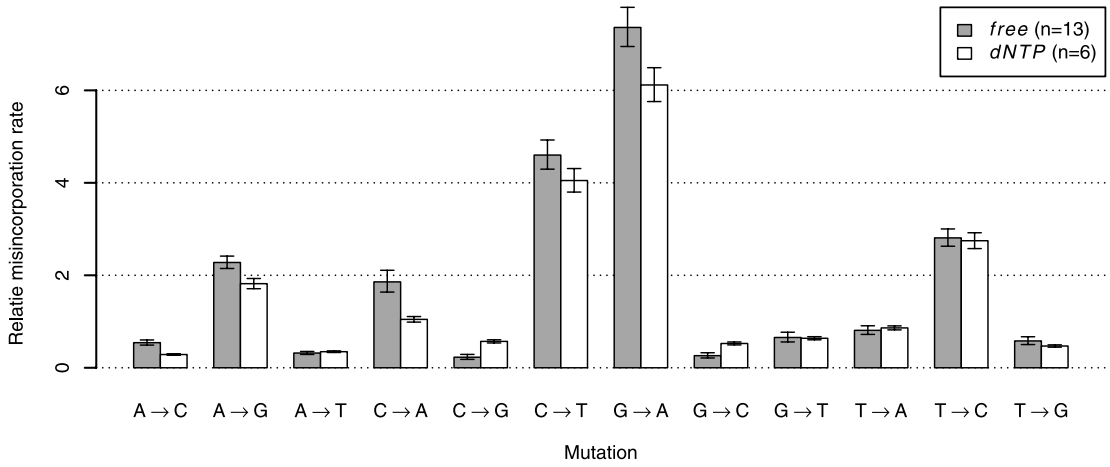

FIG. 2. Maximum likelihood estimates of relative HIV misincorporation rates estimated with two models from observed *in vivo* substitutions, cancelling for selection on the amino acid sequence, and including correction for selection against CpG dinucleotides. The *free* model allows unrestricted estimates of the misincorporation rates, while the *dNTP* model allows variation only based on dNTP concentrations with a bias for transitions. n, number of degrees of freedom.

the estimated fidelity of maintaining the correct bases in the HIV-1 genome template during one replication cycle (Fig. 3, left). When combining rates per misincorporated base, we obtain the bias in mutation towards each of the four bases (Fig. 3, right). A striking symmetry may be observed between fidelity for maintaining a base and mutation for that base (Pearson correlation: $R^2 = 0.94[0.02 - 1.00]$, $p = 0.03$). Most notably, adenine was at the same time the base most reliably maintained during a replication cycle, and most frequently mutated to.

In a *dNTP* model, we let the mutation rate $X \to Y$ to be proportional to the concentration of the misincorporated base, and inversionally proportional to the concentration of the correct base, with a preference for transitions (model *dNTP* in Fig. 2). In this way, the model forces the symmetry that was observed between fidelity for maintaining a base and mutation for that base in the *free* model. Although this model ($n = 6$) has 7 degrees of freedom less than the *free* model ($n = 13$), it can explain most of the estimated variation in misincorporation rates (Pearson correlation, $R^2 = 0.98[0.93 - 1.00]$, $p < 10^{-9}$). Nevertheless, a comparison of the two models using AIC indicates that the *free* model (AIC = 44909) is still preferred over the *dNTP* model (AIC = 45460). Table 1 shows the estimates for the parameters in the dNTP model.

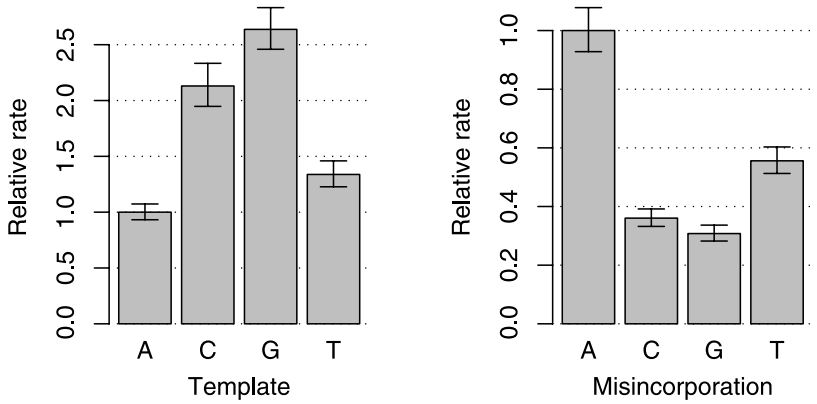

FIG. 3. Combined estimated HIV-1 relative misincorporation rates. Combined by template **(left)** shows the relative infidelity for maintaining each base in the genome (infidelity from), normalized by $\sum_Y r_{A\to Y} = 1$. Combined by generated mutation **(right)** shows the relative rate of misincorporation to each base, normalized by $\sum_X r_{X\to A} = 1$.

TABLE 1. MODEL ESTIMATES OF PARAMETERS IN THE dNTP MODEL

Parameter	*Estimate (95% HPD interval)*
dATP	1.53 (1.49–1.57)
dCTP	0.80 (0.78–0.83)
dGTP	0.84 (0.81–0.86)
dTTP	0.97 (0.95–1.00)
κ	6.10 (5.75–6.44)
CpG^+	0.56 (0.52–0.61)
CpG^-	1.47 (1.33–1.61)

To investigate variation in estimated misincorporation rates, depending on the downstream nucleotide, we estimated 12 models corresponding to every mutation $X \to Y$ where we allowed independent estimates of the probability of that misincorporation for every downstream nucleotide (Fig. 4). For every mutation, substantial variation in misincorporation probability was found depending on the downstream nucleotide.

All estimated models include two parameters to correct for known selection against CpG dinucleotides. Estimates for these parameters were similar for all models and confirmed a bias against CpG dinucleotides in the genome. In the *free* model, estimates for these parameters were $CpG^+ = 0.55$ (0.50–0.60) and $CpG^- = 1.43$ (1.29–1.59).

If no selection would counter balance the effect of misincorporation, then the misincorporation rate matrix Q estimated by the *free* model (Fig. 2) would result in HIV genome equilibrium base frequencies Π by solving $\Pi Q = 0$, which results in $\Pi = (0.51, 0.11, 0.21, 0.16)$. Observed base frequencies in the HIV dataset were $f = (0.39, 0.16, 0.21, 0.23)$.

4. DISCUSSION

We estimated several models of HIV misincorporation rate bias from *in vivo* data. Unlike substitutions, which may be observed directly in offspring, or estimated using phylogenetic methods from a cross-sectional sample of a population, generated mutations are not proportionally observed in offspring due to selection and little is known about biases in their generation. The presented method uses observed substitutions in genomic regions that code for proteins, and is robust to the selective pressures that acts on the amino acid sequence level, by only observing preferences between mutations that are *equivalent* on the

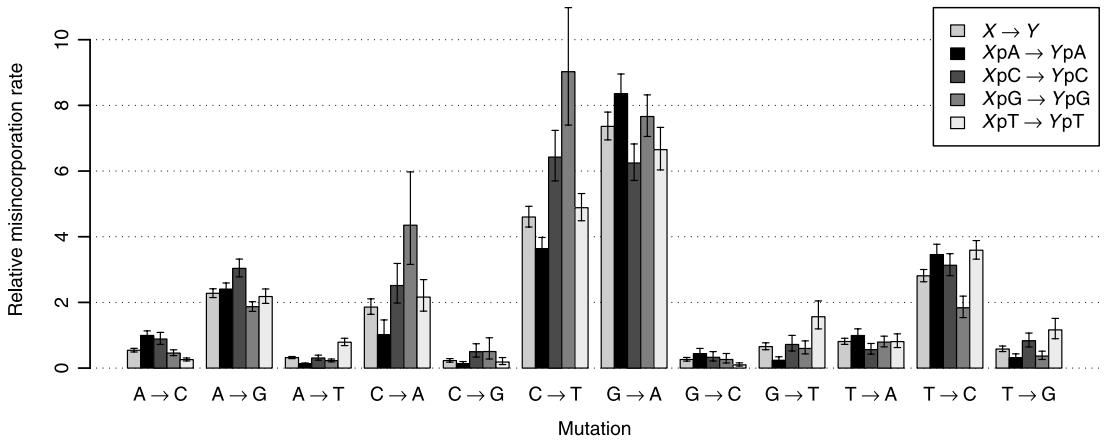

FIG. 4. Misincorporation rate variation depending on the downstream nucleotide. For each mutation X, we compare the *free* misincorporation rate estimate ($X \to Y$) with the four estimates depending on the downstream nucleotide (XpA, XpC, XpG, XpT) taken from a *free-dinucleotide* model for that base (four right bars).

amino acid sequence, since they generate the same amino acid sequence. The method is not confounded by hitchhiking effects where a substitution is fixed by a positive selection of a mutation on a different position in the the same strain, or by sequential evolution at multiple codons, because the method infers from an observed bias for a particular substitution over other substitutions that would have had exactly the same effect on the amino acid sequence, including at intermediate time points.

The method generates accurate estimates of misincorporation bias only when selection acts solely on the amino acid sequence. However, selection against CpG dinucleotides in HIV is both observed and its mechanism well understood (Shpaer and Mullins, 1990). Therefore we attempted to correct for this effect in a pragmatic way by including a factor that modifies the rate when a CpG dinucleotide is added or removed by the mutation. While estimates of these factors indeed show a consistent correction, the selection against CpG dinucleotides competes with selection on the amino acid sequence (unless for neutral mutations), and therefore is not necesarilly uniform for all mutations. This may explain why in Figure 4 estimated rates dependent on the downstream nucleotide still reveal variation involving CpG (such as deflated rates for TpG $\rightarrow$ CpG and inflated rates for CpG $\rightarrow$ ApG and CpG $\rightarrow$ TpG), although this was corrected for using two general factors. We did not consider other factors that provide selective pressure directly on the nucleotide sequence, such as selection to preserve RNA secondary structure which has been described for certain transcriptional areas in the HIV genome (Hofacker et al., 2004).

Our findings suggest a general tendency against misincorporation of a base instead of adenine (A $\rightarrow$ *Y* mutation), and at the same time the tendency towards misincorporation of adenine instead of any other base (*X* $\rightarrow$ A mutation; Fig. 3). The latter was not solely caused by the often cited G $\rightarrow$ A transition, but also by C $\rightarrow$ A and T $\rightarrow$ A which were the most frequent transversions (Fig. 2). This finding is compatible with the hypothesis that imbalance in dNTP pools is the mechanism that has shaped the observed genome nucleotide composition of HIV-1, as this would explain the observed symmetry between fidelity and misincorporation for each base: higher abundance of a specific dNTP would cause both the higher fidelity in copying and higher tendency of misincorporation of the corresponding base. To verify this hypothesis, we estimated a *dNTP* model, which assumed that misincorporations were solely caused by an imbalance in dNTP concentrations with a bias towards transitions. Since our method cannot distinguish between mutations that are generated during the minus or plus strand synthesis, we modelled only misincorporations during plus strand synthesis. Rates obtained in this way explained most of the variation observed in the free estimates, and thus confirm that dNTP pool concentrations may indeed be the source for misincorporation during the reverse transcription process. The estimated relative concentrations of dNTPs in Table 1 should not be interpreted as biological estimates, since they depend critically on the model assumption that all misincorporations are generated during plus strand synthesis. For example, if we instead had assumed that they are all generated during minus strand synthesis, the values would have been swapped with their complement (dCTP $\leftrightarrow$ dGTP, and dATP $\leftrightarrow$ dTTP). The dNTP model explains the observation that G $\rightarrow$ A is the most frequently estimated misincorporation (and most observed substitution) as a consequence of the fact that it is the only transition towards adenine.

Because APOBEC3G/3F causes mostly GpA $\rightarrow$ ApA and GpG $\rightarrow$ ApG mutations, we investigated dependence of the misincorporation rate on the downstream nucleotide (Fig. 4). We discovered considerable variation for all mutations with respect to the downstream nucleotide, which suggests either a true dinucleotide context-dependent generation or context-dependent selection on the nucleotide sequence which confounds the analysis, or both. Variation for G $\rightarrow$ A with respect to the downstream nucleotide was not more pronounced than variation for other mutations. Other context-dependent influences on the misincorporation of other mutations appear to have a larger effect than the effect APOBEC3G/3F possibly may have on the G $\rightarrow$ A mutation rate, and it cannot be excluded or confirmed that there is any effect of APOBEC3G/3F at all.

A comparison of the equilibrium frequencies of the four bases implied by the misincorporation matrix of the *free* model, with observed frequencies of the four bases in HIV isolates, implies that a further enrichment of the viral genome towards more adenine is still possible and perhaps ongoing unless the genome has already reached an equilibrium where selection balances this intrinsic bias in misincorporation.

Computational methods to reconstruct the phylogeny from a nucleotide sequence alignment usually co-estimate a nucleotide substitution rate model. Estimated substitution rates are inherently influenced by selection. However, the third codon position is degenerate for many amino acids and substitutions at this position are more likely to be synonymous. Therefore, analyses which are confounded by selection are

frequently done using data from the third codon position only. However, substitution rates estimated from a third codon position will be strongly biased towards those substitutions (mostly transitions) which are synonymous, and thus also cannot be used to reliably estimate misincorporation rates. Codon substitution models consider the entire codon as a single state, and provide in this way information on preference for a specific nucleotide substitution over other equivalent substitutions (those that generate the same amino acid). Therefore, the described technique could in principle be adapted to such a model to estimate misincorporation rates. This could be useful for comparative analyses among different viruses or other organisms.

In summary, we presented a computational method to estimate misincorporation rates from substitutions. We applied the method to estimate several models for misincorporation during the HIV replication cycle, from which we could conclude that imbalance in dNTP concentrations provide a reasonable explanation for the estimated bias in misincorporation rates. We believe the estimated misincorporation models will be useful for modelling virus evolution in forward-time simulators, for example to predict evolution of drug resistance and treatment response, or to model the process of CTL escape.

ACKNOWLEDGMENTS

We wish to thank the people that maintain the Stanford HIV Drug Resistance Database, all authors that made data available, and Anneleen Hombrouck for critical reading of the manuscript. K.D. was funded by a Ph.D grant of the Institute for the Promotion of Innovation through Sciences and Technology in Flanders (IWT). P.L. was supported by a long-term EMBO fellowship. Y.M. is a post-doctoral researcher with the FWO-Vlaanderen; his research is supported by KULeuven GOA-Mefisto-666 and GOA-Ambiorics, Belspo IUAP V-22, and EU FP6 NoE Biopattern. This work was supported by a Marie Curie GeneTime grant, by a FWO-Vlaanderen grant (G.0266.04), and by a Katholieke Universiteit Leuven grant (OT/04/43).

REFERENCES

Balzarini, J., Camarasa, M.J., Pérez-Pérez, M.J., et al. 2001. Exploitation of the low fidelity of human immunodeficiency virus type 1 (HIV-1) reverse transcriptase and the nucleotide composition bias in the HIV-1 genome to alter the drug resistance development of HIV. *J. Virol.* 75, 5772–5777.

Berkhout, B., and Hemert, F.J. 1994. The unusual nucleotide content of the HIV RNA genome results in a biased amino acid composition of HIV proteins. *Nucl. Acids Res.* 22, 1705–1711.

Deforche, K., Camacho, R., Van Laethem, K., et al. 2007. Estimating the in vivo HIV-1 fitness landscape to predict evolution during antiviral drug treatment (submitted).

Hofacker, I.L., Stadler, P.F., and Stocsits, R.R. 2004. Conserved RNA secondary structures in viral genomes: a survey. *Bioinformatics* 20, 1495–1499. Available at: *http://dx.doi.org/10.1093/bioinformatics/bth108*. Accessed August 1, 2007.

Jamburuthugoda, V.K., Chugh, P., and Kim, B. 2006. Modification of human immunodeficiency virus type 1 reverse transcriptase to target cells with elevated cellular dNTP concentrations. *J. Biol. Chem.* 281, 13388–13395. Available at: *www.jbc.org/cgi/reprint/281/19/13388/pdf*. Accessed August 1, 2007.

Johnson, V.A., Brun-Vénizet, F., Onaventura, C., et al. 2005. Update of the drug resistance mutations in HIV-1: fall 2005. *Topics HIV Med.* 13, 125–131.

Kantor, R., Machekano, R., Gonzales, M.J., et al. 2001. Human Immunodeficiency Virus Reverse Transcriptase and Protease Sequence Database: an expanded data model integrating natural language and sequence analysis programs. *Nucleic Acids Res.* 29, 296–299.

Lecossier, D., Bouchonnet, F., Clavel, F., et al. 2003. Hypermutation of HIV-1 DNA in the absence of the Vif protein. *Science* 300, 1112.

Lemey, P., Derdelinckx, I., Rambaut, A., et al. 2005. Molecular footprint of drug-selective pressure in a human immunodeficiency virus transmission chain. *J. Virol.* 79, 11981–11989.

Mansky, L., and Temin, H. 1995. Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. *J. Virol.* 69, 5087–5094. Available at: *http://jvi.asm.org/cgi/reprint/69/8/5087.pdf*. Accessed August 1, 2007.

O'Neil, P.K., Sun, G., Yu, H., et al. 2002. Mutational analysis of HIV-1 long terminal repeats to explore the relative contribution of reverse transcriptase and RNA polymerase II to viral mutagenesis. *J. Biol. Chem.* 277, 38053–38061.

R Development Core Team. 2004. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rambaut, A., Posada, D., Crandall, K.A., et al. 2004. The causes and consequences of HIV evolution. *Nat. Rev. Genet.* 5, 52–61.

Shpaer, E.G., and Mullins, J.I. 1990. Selection against CpG dinucleotides in lentiviral genes: a possible role of methylation in regulation of viral expression. *Nucleic Acids Res.* 18, 5793–5797.

Swofford, D. 2000. *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4*. Sinauer Associates, Sunderland, MA.

van de Vijver, D.A., Wensing, A.M.J., Angarano, G., et al. 2006. The calculated genetic barrier for antiretroviral drug resistance substitutions is largely similar for different HIV-1 subtypes. *J. Acquir. Immune. Defic. Syndr.* 41, 352–360.

Vartanian, J., Meyerhans, A., Sala, M., et al. 1994. G → A hypermutation of the human immunodeficiency virus type 1 genome: evidence for dCTP pool imbalance during reverse transcription. *Proc. Natl. Acad. Sci. USA* 91, 3092–3096. Available at: *www.pnas.org/cgi/repriint/91/8/3092.pdf*. Accessed August 1, 2007.

Address reprint requests to:
Dr. Koen Deforche
Rega Institute for Medical Research
Katholieke Universiteit Leuven
Minderbroedersstraat 10
Leuven, 3000, Belgium

E-mail: koen.deforche@uz.kuleuven.ac.be