

A Model of Directional Selection Applied to the Evolution of Drug Resistance in HIV-1

Cathal Seoighe,* Farahnaz Ketwaroo,† Visva Pillay,‡ Konrad Scheffler,* Natasha Wood,* Rodger Duffet,* Marketa Zvelebil,§ Neil Martinson,|| James McIntyre,¶ Lynn Morris,‡ and Winston Hide†

*Computational Biology Group, Institute of Infectious Disease and Molecular Medicine, University of Cape Town, Rondebosch, South Africa; †South African National Bioinformatics Institute, University of the Western Cape, Bellville, South Africa; ‡AIDS Virus Research Unit, National Institute for Communicable Diseases, Johannesburg, South Africa; §Dept of Biochemistry & Molecular Biology, University College London, London, United Kingdom; ||School of Medicine, Johns Hopkins University, Baltimore, Maryland; and ¶Perinatal HIV Research Unit, University of the Witwatersrand, Johannesburg, South Africa

Understanding how pathogens acquire resistance to drugs is important for the design of treatment strategies, particularly for rapidly evolving viruses such as HIV-1. Drug treatment can exert strong selective pressures and sites within targeted genes that confer resistance frequently evolve far more rapidly than the neutral rate. Rapid evolution at sites that confer resistance to drugs can be used to help elucidate the mechanisms of evolution of drug resistance and to discover or corroborate novel resistance mutations. We have implemented standard maximum likelihood methods that are used to detect diversifying selection and adapted them for use with serially sampled reverse transcriptase (RT) coding sequences isolated from a group of 300 HIV-1 subtype C-infected women before and after single-dose nevirapine (sdNVP) to prevent mother-to-child transmission. We have also extended the standard models of codon evolution for application to the detection of directional selection. Through simulation, we show that the directional selection model can provide a substantial improvement in sensitivity over models of diversifying selection. Five of the sites within the RT gene that are known to harbor mutations that confer resistance to nevirapine (NVP) strongly supported the directional selection model. There was no evidence that other mutations that are known to confer NVP resistance were selected in this cohort. The directional selection model, applied to serially sampled sequences, also had more power than the diversifying selection model to detect selection resulting from factors other than drug resistance. Because inference of selection from serial samples is unlikely to be adversely affected by recombination, the methods we describe may have general applicability to the analysis of positive selection affecting recombining coding sequences when serially sampled data are available.

Introduction

Probabilistic models of evolution are frequently applied to detect positive selection in coding sequences (Yang et al. 2000). In one common approach to this problem, the likelihood of a codon alignment, given an estimated phylogeny, is calculated under a model of evolution that constrains all sites to evolve neutrally or under purifying selection and then compared with the likelihood under a model in which a subset of sites are permitted to evolve with nonsynonymous substitution rate greater than the synonymous rate (Nielsen and Yang 1998). Under the assumption that synonymous sites evolve neutrally, positive selection is inferred when the more general model is favored over the neutral model. Strategies such as this, as well as other methods to compare nonsynonymous with synonymous substitution rates, have been used to investigate the evolution of resistance to drug treatment in pathogens including HIV-1 (Sa-Filho et al. 2003; de S Leal et al. 2004; Lemey et al. 2005; Doron-Faigenboim and Pupko 2007) and have also been used to study escape from immune responses (Zanotto et al. 1999; Beaumont et al. 2001).

Resistance to antiretroviral drugs frequently involves mutations in the HIV-1 protease and reverse transcriptase (RT) enzymes (Rhee et al. 2003). Consistent mutation to a specific amino acid can provide strong evidence of selection even in the absence of a generalized increase in the nonsynonymous substitution rate, but this signal is not

detected by the standard methods of inferring positive selection. In order to make use of the evidence for selection contained in the consistent replacement of one amino acid for another Chen et al. (2004) compared counts of mutations to a specific amino acid, Y, at a site (which they refer to as N_Y) with the number of synonymous mutations at that site (N_S) and then normalized this ratio by dividing by the ratio expected if mutations to Y occur at the neutral rate (given the transition/transversion rate ratio). By comparing the rate at which mutations to a specific amino acid occur with the neutral rate this approach has the potential to detect positive selection for drug resistance even when the overall rate of nonsynonymous substitution is substantially below the neutral rate.

We have developed an extension to the standard codon models of evolution in order to adopt a model-based approach for the detection of directional selection to a specific amino acid or set of amino acids, for example an amino acid associated with drug resistance. The models we propose are nonreversible and are designed with sets of paired sequences in mind, in which, for example, one sequence of the pair is isolated before antiretroviral exposure and the second is a postexposure sample from the same individual. We applied these models to pairs of RT coding sequences from 300 South African women infected with HIV-1 subtype C to investigate the relationship between diversifying and directional selection and the evolution of drug resistance and as a proof-of-concept for the use of tests of directional selection to identify novel drug resistance mutations. We also discuss ways in which models of directional selection such as we propose here can be further developed for application to a wider range of questions and types of data.

Key words: HIV-1, nevirapine, drug resistance, positive selection.

E-mail: cathal.seoighe@uct.ac.za.

Mol. Biol. Evol. 24(4):1025–1031. 2007

doi:10.1093/molbev/msm021

Advance Access publication February 1, 2007

© 2007 The Authors.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/2.0/uk/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Data and Methods

Paired sequences from 300 HIV-1 subtype C–infected women at baseline and after single-dose nevirapine (sdNVP) were obtained from 2 research studies. The first conducted at the Chris Hani Baragwanath Hospital in Soweto, and the King Edward Hospital in Durban, South Africa was aimed at assessing the selection of resistant variants following the use of intrapartum and postpartum nevirapine (NVP) therapy for the reduction of mother-to-child transmission of HIV-1. A second study conducted at the same site in Soweto examined the effectiveness of NVP in preventing mother-to-child transmission of HIV in a second pregnancy. The median time between first and second samples across both studies was 12 weeks (interquartile range: 7 weeks). The region sequenced spanned a 1.7-kb fragment of the polymerase gene including both protease and two-thirds of RT from plasma viral RNA. Sequences were manually edited using Sequencher version 4.5 (Gene Codes, Ann Arbor, MI). The data were generated through direct polymerase chain reaction (PCR) sequencing and included ambiguity symbols representing polymorphic nucleotide positions. Cases where a polymorphic site could be interpreted unambiguously to imply a mutation that has risen to a sufficiently high frequency to be detected as a polymorphism were included in the analysis. For example, if the codon TTG was observed at a given position in a given patient at the first time point and TTR (R represents a purine, i.e., A or G) was observed at the second time point, we infer that a mutation from TTG to TTA has risen to sufficiently high frequency to be observed as a polymorphism. The rationale for treating ambiguities in this way is presented in the Discussion. RT sequences from this study have been submitted to GenBank under accession numbers EF381747–EF382346.

Models

We used models of codon evolution similar to Muse and Gaut (1994) and Goldman and Yang (1994) to test for diversifying selection. Individual codon sites were tested separately (see Implementation) and sites with ω , the ratio of nonsynonymous to synonymous substitution rates, significantly greater than one were inferred to be evolving under diversifying selection. We have also extended the standard models of codon evolution to allow for directional selection to a specific amino acid following drug exposure. The instantaneous rate matrix describing this model is

$$q_{ij} = \begin{cases} 0 & \text{for codons differing at more than one position} \\ \pi_j & \text{for a synonymous transversion} \\ \kappa\pi_j & \text{for a synonymous transition} \\ \omega\pi_j & \text{for a nonsynonymous transversion to an amino acid other than Y} \\ \omega\kappa\pi_j & \text{for a nonsynonymous transition to an amino acid other than Y} \\ \omega_T\pi_j & \text{for a nonsynonymous transversion to amino acid Y} \\ \omega_T\kappa\pi_j & \text{for a nonsynonymous transition to amino acid Y} \end{cases}$$

where π_j is a parameter describing the frequency of codon j , κ is the transition/transversion rate ratio and ω is the ratio of nonsynonymous to synonymous substitution rates. This model is identical to the standard model of codon evolution except for the addition of a parameter ω_T , describing the selection coefficient affecting nonsynonymous mutations from codon i to codon j , where codon j encodes the “target” amino acid Y. The addition of a specific selective coefficient acting on mutations to a given amino acid can be used to test whether the rate of substitutions to codons encoding that amino acid is greater than the neutral rate, as approximated by the rate of synonymous substitution. Note that the instantaneous rate matrix given above is not time reversible. Correspondingly, the data is inherently time directional in that it includes information about which sequences were obtained before and which were obtained after exposure to the drug.

Implementation

Let $D_{ij}^{(1)}$ be the i th codon for patient j at the first sampling point and $D_{ij}^{(2)}$ the corresponding codon at the second sampling point. We estimated the parameters ω and κ (transition–transversion rate ratio) using the data at all sites by optimizing the following expression over κ and ω :

$$L(\mathbf{D}) = \prod_{ij} P(D_{ij}^{(1)} \rightarrow D_{ij}^{(2)} | \omega, \omega_T = \omega, \kappa).$$

Codon equilibrium frequencies were estimated empirically, again using all of the data and the $F3 \times 4$ model (Goldman and Yang 1994). Optimization was carried out using the Nelder and Mead method as implemented in the optim function in R (R Development Core Team 2005). Let t_j be the estimated length of the branch separating the sequences from patient j and κ' , the optimal value of κ .

To test for diversifying selection at site i , we compared the maximum value of

$$L(\mathbf{D}_i) = \prod_j P(D_{ij}^{(1)} \rightarrow D_{ij}^{(2)} | \omega, \omega_T = \omega, t = t_j, \kappa = \kappa'),$$

with $\omega \leq 1$, to the maximum value obtained when ω was unconstrained and used the likelihood ratio test to determine whether there was significant support for the more general (unconstrained) model. We used the method of Benjamini and Hochberg (1995) to calculate false discovery rates given the P values obtained from independent hypothesis tests carried out at each site in the sequence alignment.

To test for directional selection to amino acid Y at a site i , we optimized

$$L(\mathbf{D}_i) = \prod_j P(D_{ij}^{(1)} \rightarrow D_{ij}^{(2)} | \omega, \omega_T, t = t_j, \kappa = \kappa'),$$

where ω_T is the rate of nonsynonymous substitution to codons encoding amino acid Y, and ω is the rate of nonsynonymous substitution to codons that encode amino acids other than Y. In this case, we compared the maximum likelihood obtained when ω_T was constrained to be less than or equal to 1 with the likelihood obtained when ω_T was

unconstrained using the likelihood ratio test. P values for the null hypothesis (no directional selection to amino acid Y) obtained from the likelihood ratio test were multiplied by 20 for comparison with P values obtained from the diversifying selection model, to account for the fact that, a priori, the amino acid associated with drug resistance (i.e., the target of directional selection) is unknown and 20 separate model comparisons are performed, one for each target amino acid Y. This applies the conservative Bonferroni correction to correct for the multiple model comparisons that are carried out in the directional selection case. We also applied the method of Benjamini and Hochberg (1995) to calculate false discovery rates, considering all of the P values obtained from each of the 20 hypothesis tests at each site along the codon alignment.

An R workspace that includes the sequence data as well as an implementation of the diversifying and directional selection models for serially sampled data is available from the authors on request.

Results

We used the codon models for directional selection described in Data and Methods to model the evolution of HIV-1 coding sequences from the RT gene following exposure to sdNVP. The data consisted of pairs of sequences, with the first sequence of each pair obtained shortly before women received sdNVP and a second sequence obtained after sdNVP. We first tested for diversifying selection as described in Data and Methods. The 8 sites for which the rate of nonsynonymous substitution was significantly ($P < 0.05$) greater than the synonymous rate along the branches separating the first and second sampling points are shown in table 1. Mutations that confer high levels of resistance to NVP have previously been reported at 4 of these 8 sites. There are 7 sites in total in the RT gene listed as associated with high-level resistance to NVP in the Stanford Drug Resistance Database (Rhee et al. 2003). Following correction for multiple testing at codon sites (see Data and Methods) 4 of these sites remained significant (with a false discovery rate (fdr) < 0.05).

To test for directional selection favoring amino acid Y at position i of the RT gene after exposure to NVP, we compared the likelihood of the data for the case in which ω_T in our model of codon evolution is constrained to be less than or equal to 1 with the likelihood of a model in which ω_T is unconstrained (see Data and Methods). Sites inferred to be evolving under directional selection after exposure to NVP are shown in table 2. For ease of comparison with table 1 only sites that remain significant ($P < 0.05$) after Bonferroni correction was applied to correct for the fact that 20 model comparisons were carried out at each site (one for each putative target amino acid) are shown. All but one of these sites remained significant when we applied an fdr approach to correct for multiple testing, considering simultaneously multiple hypotheses tested at each site and across all sites in the alignment. Sites at which there is evidence of selection to regain the consensus amino acid are not shown in table 2 (see Discussion). The top 3 most strongly selected sites from the table (103, 181, and 188) are known to be involved in resistance to NVP. Furthermore the amino acids

Table 1
Sites in the Amino Acid Alignment Showing Evidence of Diversifying Selection in Serially Sampled Data

Position	$\Delta \ln L^a$	P	FDR ^b	ω	Known Drug Resistance
103	110.3	$< 10^{-16}$	$< 10^{-16}$	7.5	Highly resistant
181	37.5	$< 10^{-16}$	$< 10^{-16}$	4.2	Highly resistant
123 ^c	24.9	8×10^{-13}	2×10^{-10}	3.5	None
188	18.9	4×10^{-10}	6×10^{-8}	3.0	Highly resistant
211 ^c	4.3	2×10^{-3}	0.2	1.8	None
286	2.6	0.01	1	1.8	None
106	2.3	0.01	1	1.6	Highly resistant
4	1.9	0.02	1	1.8	None

^a Logarithm of the likelihood under the diversifying selection model minus the logarithm of the likelihood under the null model.

^b False discovery rate.

^c Sites that are evolving under diversifying selection in treatment naïve sequences.

targeted by selection at these sites after NVP exposure all correspond to the amino acids that were previously known to be associated with resistance to NVP. Two further known resistance mutations (190 and 106) appear further down the table and again an amino acid associated with resistance is correctly identified by the directional selection model in both cases. For all of the sites known to be associated with NVP resistance, just one resistance mutation is selected in our clade C data set even when there are several mutations that are known to be associated with similar levels of drug resistance. For example at position 103, one of the most important sites for the evolution of resistance to NVP, the only mutation that we detect under selection is the mutation to asparagine, even though in the Stanford Drug Resistance Database mutations to serine and threonine are also listed as causing high levels of resistance to NVP. Although these alternative mutations are known to confer resistance, they occur extremely rarely in our cohort (once in the case of threonine and not at all in the case of serine).

For the sites in RT that are known to confer resistance to NVP but do not appear in table 1 or 2, it is possible to use the directional selection model to test whether there is any evidence of accelerated evolution toward the amino acid or set of amino acids associated with resistance. In this case, the targeted amino acid (or set of amino acids) can be specified a priori, resulting in an increase in power. The only mutations that are associated with high-level resistance but do not appear in table 2 are 101P and 230L. Neither of these mutations appears in any of the 300 postexposure sequences, suggesting that these resistance mutations are extremely rare in NVP-treated South African individuals infected with HIV-1 subtype C, if they occur at all. At codon 101, glutamic acid is associated with low-level resistance to NVP and this amino acid appears in 6 of the 300 postexposure sequences. The directional selection model allows us to assess the evidence for selection to glutamic acid at this site and in the case of this mutation the null model cannot be rejected in favor of the directional selection model. We conclude that 6 mutations from lysine to glutamic acid (each involving a single transition) from the 300 paired sequences are not inconsistent with neutral evolution.

Table 2
Sites in the Amino Acid Alignment Showing Evidence of Directional Selection

Position	S ^a	F1 ^b	F2 ^c	$\Delta\ln L^d$	<i>P</i>	FDR ^e	ω	ω_T	Known Drug Resistance
103	N	0.00	0.48	341.7	$<10^{-16}$	$<10^{-16}$	0.20	44.4	Highly resistant
181	C	0.01	0.19	132.1	$<10^{-16}$	$<10^{-16}$	0.17	26.8	Highly resistant
188	C	0.01	0.14	86.96	$<10^{-16}$	$<10^{-16}$	0.16	19.6	Highly resistant
123 ^f	S	0.19	0.22	25.42	5×10^{-13}	7×10^{-10}	1.00	11.9	None
190	A	0.00	0.06	23.29	4×10^{-12}	5×10^{-9}	0.03	7.33	Highly resistant
162	C	0.12	0.13	19.64	2×10^{-10}	2×10^{-7}	0.85	21.1	None
36	E	0.23	0.25	11.76	6×10^{-7}	4×10^{-4}	0.40	4.58	None
174 ^f	K	0.46	0.44	11.75	6×10^{-7}	4×10^{-4}	0.89	4.71	None
211 ^f	R	0.37	0.36	10.41	3×10^{-6}	2×10^{-3}	1.00	4.96	None
106	M	0.00	0.06	8.35	2×10^{-5}	0.01	0.81	3.04	Highly resistant
286	T	0.21	0.23	7.13	8×10^{-5}	0.04	0.97	4.07	None
277 ^f	K	0.42	0.46	7.07	8×10^{-5}	0.04	0.46	2.85	None
177	D	0.25	0.27	6.19	2×10^{-4}	0.10	0.15	3.92	None

^a The selected amino acid.^b Frequency of the selected amino acid in pretreatment sequences.^c Frequency of the selected amino acid in the posttreatment sequences.^d Logarithm of the likelihood under the directional selection model minus the logarithm of the likelihood under the null model.^e False discovery rate.^f Sites that are evolving under diversifying selection in treatment naïve sequences.

Simulations

We simulated serially sampled data at a single neutrally evolving codon site to test the false positive rate of inference of selection using our methods. The simulation was modeled on amino acid position 103 of the RT gene, a site associated with high levels of resistance to NVP. Each simulated data set consisted of 300 codons encoding lysine (corresponding to the 300 sequences from the first sampling time), which we evolved neutrally for a genetic distance equal to the mean distance between the pre- and postexposure sequences in our data set using *evolver* (Yang 1997). We then applied the models of diversifying and directional selection to the resulting simulated data sets. As expected, comparison of the likelihood ratio test statistic obtained to the Chi-bar-square distribution with one degree of freedom (df) (i.e., the equal mixture of a Chi-square distribution with one df and a point mass at zero, appropriate for a one-sided likelihood ratio test) provides a sound basis on which to reject the null hypothesis (neutral evolution or purifying selection) in the case of both the diversifying selection and directional selection models (fig. 1). In 1,000 simulated data sets the null hypothesis was rejected in 4.3% of cases for the diversifying selection model and in 2.6% of cases for the directional selection model.

To test the power of the methods to identify selection for drug resistance we again modeled our simulations on site 103. We wished to establish the minimum number of resistance amino acids required for rejection of the null hypothesis in the case of the diversifying and directional selection models. For each value of this number from 0 to 30, we randomly selected a subset of the patients to encode the resistance amino acid (asparagine). Asparagine codons in excess of this number at the second time point were mutated to the closest lysine codon. We then plotted the logarithm of the *P* value of the null hypothesis as a function of the number of drug resistance codons introduced (fig. 2). For the case of position 103 in the alignment there is only one nucleotide difference (a transversion) between the consensus codon and the drug resistance codon. The

number of occurrences of the codon associated with drug resistance required to reject the null hypothesis is far lower for the directional selection model than for the diversifying selection model. For the directional selection model, we required just 7 resistance codons, whereas for the diversifying selection model we required 17 resistance codons to reject the null hypothesis. For mutations involving a transition the number of resistance mutations required to reject the null hypothesis is significantly greater.

Discussion

The strong selective pressure that drives the evolution of drug resistance in HIV should be discernable at the sequence level, but the classic approach of calculating ω , the nonsynonymous to synonymous substitution rate ratio, has not always given the expected result. Crandall et al. (1999) measured mean ω values less than one (characteristic of purifying selection) in drug-treated HIV sequences despite the presence of parallel evolution of identical resistance mutations in sequences from several individuals, strongly indicative of natural selection. Detection of selection in HIV associated with drug treatment can be made more sensitive by the application of methods that detect a subset of sites with ω greater than one even when the mean value of ω is less than one, because only a minority of sites are likely to experience selective pressure to evolve drug resistance. Further increase in sensitivity can be gained by application of phylogenetic methods that allow different values of ω along different branches of the tree because drug resistance affects only a clearly defined portion of the evolutionary history of a set of HIV sequences. The branch-site models of Yang and Nielsen (2002) have been applied for this purpose (Lemey et al. 2005) and should give similar results to the method used here to identify diversifying selection. Our method considers only the branches that connect pre-NVP exposure to postexposure sequences and is thus a far simpler way to detect sites with $\omega > 1$ from serially sampled coding sequences. It also considers each site independently and thus avoids making assumptions about the distribution

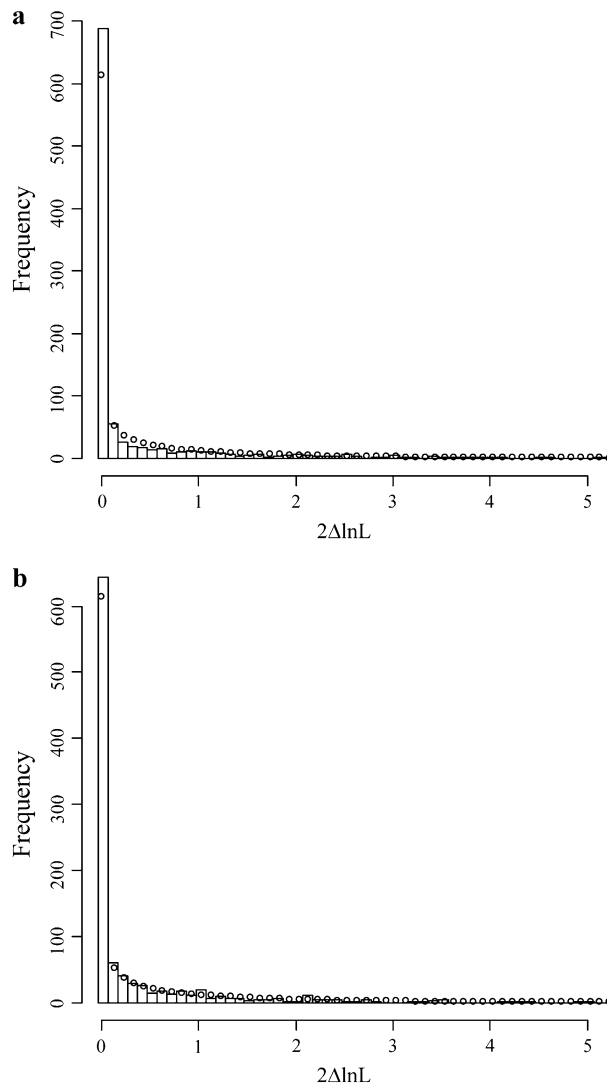


FIG. 1.—Histograms of the likelihood ratio test statistic obtained from simulated data under neutral evolution for (a) the diversifying selection and (b) the directional selection model (with Bonferroni correction for multiple testing as described in Data and Methods).

of ω over sites. By introducing the additional option of comparing rates of specific mutations to the expected rate under neutrality the directional selection model offers a substantial further increase in sensitivity.

The diversifying selection model identified 4 of the 7 sites in the RT gene that are known to confer high levels of resistance to NVP. However, only 3 of the 4 sites detected remained significant after correction for multiple testing using an *fdr* cutoff of 0.05. For each of these 3 known sites, the statistical support for the alternative model was far higher with the directional selection model, both before and after correction for multiple testing. Because HIV has an exceptionally high mutation rate, many virions in the population have suboptimal fitness and may even be entirely nonviable. If an amino acid associated with relatively low fitness becomes fixed in the population, it has a high probability of being replaced by fitter viruses and this can result in directional selection to the consensus amino acid. For this reason none of the sites that involve

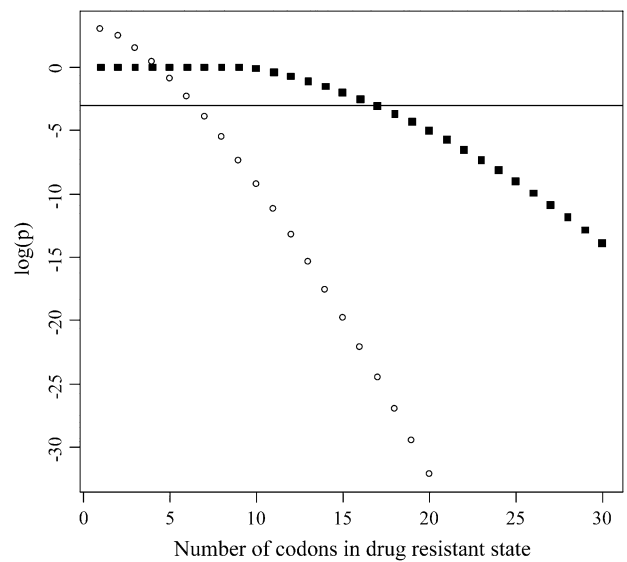


FIG. 2.—Logarithm of the *P* value for the null hypothesis plotted against the simulated number of drug-resistant codons. The horizontal line indicates the 0.05 significance threshold.

directional selection in favor of the most common preexposure amino acid can be considered as good candidate sites for the evolution of NVP resistance. After removing cases in which the support for the directional selection model could result from purifying selection to regain the consensus amino acid, we retained 12 sites that are candidates for NVP resistance. Five of these sites are among the 7 sites associated with high-level resistance to NVP. All 5 are retained after application of an *fdr* cutoff of 0.05. The presence of 5 of the known sites of high-level NVP resistance among the 12 sites undergoing directional selection represents significant enrichment for sites associated with drug resistance (Fisher's Exact Test: $P = 8 \times 10^{-7}$).

Although our method provides good evidence for selection along the branches of the phylogeny connecting the pre- and post-NVP exposure sequences, the observed selection can have causes other than the evolution of drug resistance. To test whether there may be other causes for the selection pressure observed at the sites not previously reported to be involved in NVP resistance, we established a data set consisting of only preexposure sequences. We also removed from the data set all sequences from mothers who had received NVP to prevent mother-to-child transmission of HIV-1 in a previous pregnancy. We constructed a phylogenetic tree from the 150 sequences that remained and inferred sites evolving under diversifying selection using the selection model (M2a) from Wong et al. (2004), implemented in HyPhy (Pond et al. 2005). Five sites (123, 174, 207, 211, and 277) showed evidence of positive selection using this method and 4 of the 5 are found in table 2. Although recombination can cause false inference of positive selection in HIV-1 sequences, this is thought to be less of a problem for site-specific inference of selection (Anisimova et al. 2003). Our inference of selection based only on branches connecting serially sampled sequences is unlikely to be severely affected by recombination. Because only 2 sequences are used per patient there is no possibility of error in the inpatient tree topology (the inpatient

trees consist of a single branch). When we used the mean branch lengths (i.e., mean distance between pairs of sequences from the same patient) instead of the estimated inpatient branch lengths, the results obtained were virtually unchanged (not shown). This illustrates that the method is not very sensitive to the individual branch lengths used. Poor estimates of the branch lengths that apply at specific sites, resulting from inpatient recombination, are therefore unlikely to give rise to the high probability of false detection of positive selection that has been reported for phylogenetic methods. The presence of sequences from individuals infected with more than one strain of HIV in the data can cause false inference (due to multiple counting of the same substitution events). Dual infection is evident when sequences from the same patient fail to cluster on a phylogenetic tree, and our method requires that data from such patients be removed prior to the analysis.

The only candidate novel resistance mutations in table 2 following the removal of sites evolving under positive selection in the preexposure sequences are 36, 162, 177, and 286. All of the known NVP resistance sites showed strong evidence of purifying selection in the drug-naive data set, but each of these 4 novel sites was placed in the neutral evolution class by HyPhy. It is possible that these sites are evolving under positive selection in the drug-naive data set but that the selection was too weak for them to be classified in the selection class (for which ω was 2.7). If that is the case, the selection detected from the serially sampled data is not necessarily indicative of an involvement in drug resistance. To investigate this possibility, we ran a more sensitive phylogenetic test of selection again using HyPhy, but this time with a model with discrete ω components at 0, 0.33, 1.5, and 3.0. All of the novel candidate resistance sites were placed in the $\omega > 1$ categories with posterior probability (summed over the 2 categories with $\omega > 1$) greater than 0.5 except site 286 for which the sum of the posteriors was marginally larger for the neutral and purifying categories compared with the positive selection categories. Interestingly, after correction for multiple testing, the diversifying selection model applied to the serially sampled data detected only one of the sites evolving under positive selection in the drug-naive sequences compared with 4 sites using the directional selection model. This suggests that the directional model is also likely to have more power to detect other forms of selection than the diversifying model when applied to serially sampled data. It may well be that other forms of selection, such as selection to escape from host immune responses, are also better described by a directional model that allows specific mutations to occur more frequently than the neutral rate than by the diversifying model that measures a generalized increase in ω .

In order to evaluate further the sites in table 2 that represent possible novel NVP resistance mutations, it is useful to combine evidence of directional selection with a comparison of amino acid frequency between the pre- and postexposure sequences. Comparison of amino acid frequency between drug-resistant and drug-susceptible sequences is typically used to detect novel drug resistance mutations. None of the 4 novel candidate resistance amino acids (36, 162, 177, and 286) had significantly higher frequency in the post-NVP exposure sequences compared with the

preexposure sequences. Furthermore analysis of amino acids present at these sites in published sequences from drug-treated and drug-naive individuals using HIVseq (Rhee et al. 2006) provided no indication that these amino acids are associated with resistance to non-nucleoside reverse transcriptase inhibitors in other studies. Therefore, further data or experimentation will be required to determine what involvement, if any, mutations at these sites have in the development of NVP resistance.

The model of directional selection that we propose is nonstationary—that is, we do not expect the codon frequencies to remain constant over time. Instead, for codon positions that confer drug resistance, we expect the frequency of codons associated with the resistance amino acid to increase following the initiation of drug treatment. We retained the codon frequency parameters, which were estimated empirically over the entire sequence length, in order to continue modeling possible codon frequency biases such as differences in translational efficiency between codons. However, unlike the codon models on which the model of directional selection is based, the empirical codon frequency parameters in the model can no longer be interpreted as the limiting frequencies of the Markov process described by the instantaneous rate matrix. Including the codon frequency parameters has the advantage of causing the directional selection model to reduce exactly to the more standard codon model when ω_T is constrained to be equal to ω .

Typically, the codon associated with resistance following sdNVP does not supplant the wild-type codon completely and instead the viral population consists of a mixture of resistant and susceptible viruses. The data for this study were generated through direct PCR sequencing and included ambiguity symbols representing nucleotide positions that appear to be polymorphic. Our analysis considered mutations that have reached fixation as well as mutations that have reached a high enough frequency to be detected as polymorphisms, provided there was exactly one such polymorphism in the pair of codons from a patient at a specific site. If there was more than one polymorphism in a pair of codons, or a nucleotide substitution in addition to the polymorphism, then the codon pair was omitted from the analysis (because of the complexity introduced by having multiple possible paths between the codon pair). Under neutrality we expect the rate at which new mutations rise to sufficiently high frequency to be detected as polymorphisms to be the same for synonymous and nonsynonymous mutations. This rate should be lower for nonsynonymous mutations compared with synonymous mutations at sites evolving under purifying selection and higher at sites evolving under positive selection. Therefore, the inference of positive selection from a higher rate of nonsynonymous than synonymous substitution remains valid when we include mutations that have not yet reached fixation as described above.

One of the most useful practical applications of the directional selection model we present is likely to be for evaluating specific hypotheses about the evolution of drug resistance in a serially sampled cohort. Association of mutations with drug treatment or drug resistance is frequently used to identify novel resistance mutations but this association can be an artifact of linkage or genetic drift. Evidence of directional selection in a serially sampled cohort provides

powerful support for candidate resistance mutations identified through association. Additionally, as we see in this cohort, not all resistance mutations are actually selected in a given cohort and the model of directional selection presented here can be used to evaluate, which of the possible resistance mutations are actually selected in a given set of individuals.

Recently, Chen and Lee (2006) proposed a conditional selection model to test for interactions between sites involved in the evolution of drug resistance. This method compares ω at site j between sequences with a mutation at site i and sequences with the wild-type amino acid at site i . This is a natural application of the model-based approach to serially sampled coding sequences. Comparison of the likelihood of a model with separate ω values at site j depending on the character at site i provides a more powerful and direct way of evaluating the evidence for conditional selection that could help to uncover dependencies between sites involved in evolution of drug resistance. However, this would probably require larger data sets of serially sampled sequences than the data set investigated here.

Acknowledgments

We would like to thank Allen Rodrigo and 3 anonymous reviewers for their comments on the manuscript. This work was supported by the South African National Bioinformatics Network (K.S. and R.D.), the National Institutes of Health through the Centre for the AIDS Programme of Research in South Africa (grant number 1U19AI51794; C.S. and W.H.). F.K. was supported by a training grant under the Stanford-South Africa Biomedical Informatics Training Program, which is supported by the Fogarty International Center, part of the National Institutes of Health (grant number 5D43 TW006993). The studies that generated the sequences were funded by the South African National Department of Health and the United States Agency for International Development (USAID) and the President's Emergency Plan for AIDS Relief (grant numbers 674-0320-G-00-5053 and 674-A-00-05-00003-00). The views expressed are those of the authors and do not necessarily reflect those of USAID. We thank Johanna Ledwaba for assembling the sequences. We are also grateful for access to the Linux cluster at the University of Cape Town Linux Competency Centre.

Funding to pay the Open Access publication charges for this article was provided by the South African National Bioinformatics Network.

Literature Cited

Anisimova M, Nielsen R, Yang Z. 2003. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics*. **164**:1229–1236.

Beaumont T, van NA, Broersen S, Blattner WA, Lukashov VV, Schuitemaker H. 2001. Reversal of human immunodeficiency virus type 1 IIIB to a neutralization-resistant phenotype in an accidentally infected laboratory worker with a progressive clinical course. *J Virol*. **75**:2246–2252.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B*. **57**:289–300.

Chen L, Lee C. 2006. Distinguishing HIV-1 drug resistance, accessory, and viral fitness mutations using conditional selection

pressure analysis of treated versus untreated patient samples. *Biol Direct*. **1**:14.

Chen L, Perlina A, Lee CJ. 2004. Positive selection detection in 40,000 human immunodeficiency virus (HIV) type 1 sequences automatically identifies drug resistance and positive fitness mutations in HIV protease and reverse transcriptase. *J Virol*. **78**:3722–3732.

Crandall KA, Vasco DA, Posada D, Imamichi H. 1999. Advances in understanding the evolution of HIV. *AIDS*. **13**(Suppl A): S39–S47.

de S Leal, Holmes EC, Zanotto PM. 2004. Distinct patterns of natural selection in the reverse transcriptase gene of HIV-1 in the presence and absence of antiretroviral therapy. *Virology*. **325**:181–191.

Doron-Faigenboim A, Pupko T. 2007. A combined empirical and mechanistic codon model. *Mol Biol Evol*. **24**:388–397.

Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol*. **11**:725–736.

Lemey P, Derdelinckx I, Rambaut A, Van LK, Dumont S, Vermeulen S, Van WE, Vandamme AM. 2005. Molecular footprint of drug-selective pressure in a human immunodeficiency virus transmission chain. *J Virol*. **79**:11981–11989.

Muse SV, Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol*. **11**:715–724.

Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics*. **148**:929–936.

Pond SL, Frost SD, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. *Bioinformatics*. **21**:676–679.

R Development Core Team. 2005. R: a language and environment for statistical computing.

Rhee SY, Gonzales MJ, Kantor R, Betts BJ, Ravela J, Shafer RW. 2003. Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res*. **31**:298–303.

Rhee SY, Kantor R, Katzenstein DA, Camacho R, Morris L, Sirivichayakul S, Jorgensen L, Brigido LF, Schapiro JM, Shafer RW. 2006. HIV-1 pol mutation frequency by subtype and treatment experience: extension of the HIVseq program to seven non-B subtypes. *AIDS*. **20**:643–651.

Sa-Filho DJ, Costa LJ, de Oliveira CF, Guimaraes AP, Accetturi CA, Tanuri A, Diaz RS. 2003. Analysis of the protease sequences of HIV-1 infected individuals after Indinavir monotherapy. *J Clin Virol*. **28**:186–202.

Wong WS, Yang Z, Goldman N, Nielsen R. 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics*. **168**:1041–1051.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci*. **13**:555–556.

Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol*. **19**:908–917.

Yang Z, Nielsen R, Goldman N, Pedersen AM. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*. **155**:431–449.

Zanotto PM, Kallas EG, de Souza RF, Holmes EC. 1999. Genealogical evidence for positive selection in the nef gene of HIV-1. *Genetics*. **153**:1077–1089.

Edward Holmes, Associate Editor

Accepted January 26, 2007