Sequence analysis

Estimation of an in vivo fitness landscape experienced by HIV-1 under drug selective pressure useful for prediction of drug resistance evolution during treatment

K. Defochre1, R. Camacho2, K. Van Laethem1, P. Lemey1,3, A. Rambaut4, Y. Moreau5 and A.-M. Vandamme1,*

1Rega Institute for Medical Research, Katholieke Universiteit Leuven, Leuven, Belgium, 2Centro Hospitalar de Lisboa Ocidental, Lisbon, Portugal, 3Department of Zoology, Oxford University, Oxford, 4Institute for Evolutionary Biology, University of Edinburgh, Edinburgh, UK and 5ESAT, Katholieke Universiteit Leuven, Leuven, Belgium

Received on January 4, 2007; revised on October 19, 2007; accepted on October 22, 2007
Advance Access publication November 17, 2007
Associate Editor: Burkhard Rost

ABSTRACT
Motivation: HIV-1 antiviral resistance is a major cause of antiviral treatment failure. The in vivo fitness landscape experienced by the virus in presence of treatment could in principle be used to determine both the susceptibility of the virus to the treatment and the genetic barrier to resistance. We propose a method to estimate this fitness landscape from cross-sectional clinical genetic sequence data of different subtypes, by reverse engineering the required selective pressure for HIV-1 sequences obtained from treatment naive patients, to evolve towards sequences obtained from treated patients. The method was evaluated for recovering 10 random fictive selective pressures in simulation experiments, and for modeling the selective pressure under treatment with the protease inhibitor nelfinavir.

Results: The estimated fitness function under nelfinavir treatment considered fitness contributions of 114 mutations at 48 sites. Estimated fitness correlated significantly with the in vitro resistance phenotype in 519 matched genotype-phenotype pairs ($R^2 = 0.47 (0.41 - 0.54)$) and variation in predicted evolution under nelfinavir selective pressure correlated significantly with observed in vivo evolution during nelfinavir treatment for 39 mutations (with FDR = 0.05).

Availability: The software is available on request from the authors, and data sets are available from http://jose.med.kuleuven.be/~kdeforc0/nfv-fitness-data/.

Contact: annemie.vandamme@uz.kuleuven.be
Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION
HIV antiviral drugs interfere with viral proteins resulting in the inhibition of HIV replication. In many cases, HIV escapes the inhibition of these drugs by selection of drug resistance mutations, leading to treatment failure (Vandamme, 1999). To combine drugs in an effective treatment therefore requires taking into account the presence of resistance mutations, and resistance testing has become a standard of care (Vandamme et al, 2004). Different viral mechanisms may be distinguished that affect short-term versus long-term response to antiviral treatment. In addition to the impact of other factors such as adherence, potency of therapy and pharmacokinetics, the short-term response to treatment is mainly determined by the susceptibility of the virus to the drugs. In the long-term, susceptible virus may evolve to acquire resistance mutations, and the expected time needed for the virus to evolve the necessary resistance mutations is related to the number of nucleotide substitutions required, which can be quantified as the genetic barrier. Several bioinformatics methods have been used successfully in the field of antiviral drug resistance, including methods that predict in vitro phenotypic resistance from the genetic sequence, and methods that describe qualitative relationships between different mutations selected during treatment (Beerenwinkel et al, 2005b). Recently, these techniques were combined to compare the genetic barrier for individual drugs versus drug combinations (Beerenwinkel et al, 2005c).

Due to technical shortcomings, problems with the interpretation of the results, and the lack of a genetic barrier concept, in vitro phenotypic assays display limitations in their capacity of predicting therapy outcome (Van Laethem and Vandamme, 2006). Therefore, the usefulness of machine-learning approaches to predict resistance phenotype from genotype may be limited. On the other hand, the success of attempts to directly learn genotypic patterns responsible for reduced treatment response from clinical data has been limited by the lack of sufficient data, and the confounding effect of many other factors (DiRienzo and DeGruttola, 2002). The in vivo fitness of the virus in presence of treatment, which reflects both effects of drug resistance and replication capacity,1 determines the immediate treatment response, but cannot be measured directly.

1In this article, the universal meaning of fitness as the capacity to replicate in a given environment is used, rather than as a synonym for replication capacity in a drug-free environment as is often the case in the HIV drug resistance community.
However, HIV tries to recover its ability to replicate efficiently in presence of treatment by accumulating resistance mutations, thus exploring sequence space in the immediate neighborhood of the current sequence. Therefore, observed evolution in clinical sequences at treatment failure provides information about the fitness landscape, but only in the immediate neighborhood of the current sequence. In this article, we present a method to reverse engineer this fitness landscape experienced by HIV-1 in presence of treatment as a function of the genetic sequence, from observed selection during treatment in clinical sequences. The method searches for a fitness landscape, which explains how an observed population of treated sequences could have evolved from a population of untreated sequences under selective pressure. After showing that random, but known fitness functions could be successfully estimated in this way, we applied the method to model the fitness landscape of HIV-1 in presence of the protease inhibitor (PI) nelfinavir (NFV).

2 MATERIALS AND METHODS

2.1 Clinical data set

To estimate the fitness function under NFV selective pressure, clinical data was pooled from the Stanford HIV Drug Resistance Database (Kantor et al., 2001), from the University Hospitals, Leuven, Belgium, and from Hospital Egas Moniz, Lisbon, Portugal, to create a treated population $P^T$ of 10732 sequences from patients with experience to NFV as sole PI, and a naive population $P^N$ of 7774 sequences from PI naive patients. At most one treated and one naive sequence per patient was used, and duplicate sequences (that were present in the hospital database but also published in the Stanford database) were identified and removed. The treated population consisted mostly of HIV-1 subtype B sequences, but included also a large number of subtype G and subtype C sequences (Fig. 1), as determined from the protease and partial reverse transcriptase sequences using the REGA HIV-1 Subtype tool v2.0 (de Oliveira et al., 2005). Nucleotide ambiguities that occur commonly in the population sequences were resolved by randomly substituting the mixture with a suitable pure nucleotide. Using a threshold of 0.5% prevalence at variable sites, 114 mutations at 48 protease positions were included in the fitness function models, and are listed in Supplementary Material A. To remove redundancy, the most prevalent mutation at each position was considered the ‘wild type’ and was omitted from the fitness function model, as its presence was implied in the absence of any mutation.

2.2 Method to estimate a fitness function

We present a method with objective to learn a function $F(A_1, \ldots, A_n)$, where $A_i$ presents presence or absence of a mutation, that represents the fitness landscape of HIV under drug selective pressure. To learn $F$, we find a function that fits with the evolution of the virus in a naive population of patients $P^N$ to a treated population $P^T$, and is closest to neutrality (minimizing $|F - 1|$). The fitness function $F$ incorporates interactions indicated using Bayesian Network (BN) learning, and its parameters are estimated using an iterative procedure where evolution for $P^T$ over the current fitness function estimate is simulated, and compared to $P^T$.

2.2.1 Fitness function structure

The protease amino acid sequences from the treated population $P^T$ were used to learn interactions between mutations as described before (Deforche et al., 2006). Briefly, a data set was created where a boolean variable indicated the presence of each included mutation. BN structure learning (Myllymäki et al., 2002) on this boolean data was used to discover relationships between these mutations that may indicate epistatic fitness effects. By assuming conditional independencies, the Bayesian Network refactors the Joint Probability Distribution (JPD) in a product of Conditional Probability Distributions (CPD), leading to a reduction in number of parameters to model the JPD. Formally, for $n$ variables $A_1, \ldots, A_n$ (representing amino acid mutations), we would write:

$$P(A_1, \ldots, A_n) = \prod_{i=1}^{n} P(A_i | parents(A_i))$$

with $P(A | B)$ the conditional probability of $A$ given $B$, and parents($A_i$) the parents in the BN structure of variable $A_i$. We denote the most probable network of the amino acid sequences of the treated population $P^T$ with structure $S^T$ and CPD parameters $\theta^T$ as $BN^T(\theta^T, S^T)$.

We model the relative fitness function $F(A_1, \ldots, A_n)$ in the same way as $BN^T(\theta^T, S^T)$ refactors the JPD:

$$F(A_1, \ldots, A_n) = \prod_{i=1}^{n} F(A_i | parents(A_i))$$

with parents($A_i$) the parents in $S^T$, and $F(A | B)$ the Conditional Fitness Contribution (CFC) of the presence of $A_i$ depending on the presence of $B$. The assumption here is that if two mutations are synergistic for example, they would occur more often together than not, and a dependency should be visible in the JPD too. See Supplementary Material B for an example.

The CPDs are modeled by specifying the probability for a mutation $A_i$ given any pattern of parent mutations $k$, in Conditional Probability Tables (CPTs): $\theta_{ik} = P(A_i = 1 | parents(A_i) = k)$. Similarly, we used Conditional Fitness Tables (CFTs) to model the CFCs for each mutation $A_i$, which specify a different fitness contribution of the presence of a mutation $A_k$ for every pattern of parent mutations: $\phi_{ik} = F(A_i = 1 | parents(A_i) = k)$.

2.2.2 Model of evolution

Both for estimating the fitness function parameters, and for prediction of sequence evolution during treatment, the same stochastic model of HIV evolution was used. Evolution was considered as an accumulation of fixations of nucleotide mutations in the HIV intra-host population, as reflected in the consensus sequence, under the selective pressure of an arbitrarily complex fitness function. This corresponds roughly to how HIV resistance evolution is observed in population sequences obtained by genotypic resistance tests (Van Laethem and Vandamme, 2006).

The HIV intra-host population was modeled by a finite ideal Wright–Fisher population with selection and mutation, using empirical estimates of the HIV intra-host effective population size, mutation rate and
mutation rate biases derived from literature, and selection coefficients derived from the fitness function \( F \). Analytical results lack for fixation time distributions of mutations in the Wright–Fisher model for all but the simplest cases (Evens, 1979). Therefore, to sample from these distributions, for a Wright–Fisher model with multiple loci and a complex fitness function with epistatic interactions, an approximate simulation of this model was implemented. For a detailed description of the implemented model and approximations see Supplementary Material C.

### 2.2.3 Fitness function parameters
The parameters \( \phi_{i,k} \) of the function \( F \) are estimated so that evolution over the fitness landscape of a naive population \( p^N \) resembles the treated population \( p^T \). Therefore, evolution is simulated for sequences sampled from the naive population \( p^N \) using the fitness function, to obtain an evolved population \( p^E \). The difference between the sequence populations \( p^T \) and \( p^E \), which must thus be minimized, is measured by comparing the parameters of \( BN(\theta^T, S^T) \) of the treated data set, with \( BN(\theta^P, S^P) \), a BN estimated from the simulated population using the structure that was learned from the treated data set. Thus, we measure and minimize the difference in prevalence of each mutational pattern that is modeled by the BN, and for which the fitness function specifies a separate fitness contribution.

Given this minimization objective, the parameters \( \phi_{i,k} \) are not necessarily unique, since we cannot quantify how unfit unobserved mutational patterns are. Indeed, we can only determine how unfit these patterns must be at least to explain their lack of evolution. We constrain \( \phi_{i,k} \) to be a function of two parameters: the unfitness \( d \) of the sequence and the parameter \( \alpha_k \) of the fitness function. Therefore, the structure of \( \phi_{i,k} \) is estimated so that evolution over the fitness landscape of a naive population \( p^N \) resembles the treated population \( p^T \). The fitness function parameters \( \phi_{i,k} \) are subsequently adjusted based on the difference in the sufficient statistics (reflect the counts) related to the BN parameters \( \theta^P \) and \( \theta^T \). The fitness function parameters \( \phi_{i,k} \) are then used to simulate the evolution of sequences in the treated data set.

### 2.2.4 Phylogenetic guide tree
The sampling of sequences from \( p^N \) was guided by a phylogenetic tree, and more weight was given to sequences from the naive population that were epidemiologically linked to the treated population. This assures that the sampled population had a similar epidemiological background to the treated population avoiding that mutations linked to epidemiologies with a different distribution among these population were assigned as arising during treatment [an improvement compared to stratifying according to epidemiology (Deforche et al., 2006; Kantor et al., 2006)]. The protease and partial reverse transcriptase nucleotide sequences were used to reconstruct a neighbor-joining phylogenetic tree including all naive and treated isolates used in the training data. The tree was built using PAUP (Swofford, 2000) using the HKY-+\( \gamma \) substitution model, and codons representing IAS resistance associated positions (Johnson et al., 2005) were excluded to avoid problems of convergent evolution.

Each sequence \( n \) from the treated population \( p^T \) added a contribution \( k e^{-\frac{1}{2\sigma^2} (\mu_n^T, n^T)^2} \) to the sampling weight of a sequence \( n^T \), with \( d(n_1, n_2) \) the tree distance between two taxa \( n_1 \) and \( n_2 \), \( r \) a decay factor and \( k \) a normalizing coefficient so that \( \sum_{nT} e^{-\frac{1}{2\sigma^2} (\mu_n^T, n^T)^2} = |p^T|^{-1} \).

### 2.2.5 Constants
For the HIV simulation model, a constant intra-patient effective population size \( N_e = 10^4 \) was assumed, a value previously estimated from in vivo observations during treatment (Nijhuis et al., 1998; Rouzine and Coffin, 1999), and an average mutation rate \( \mu = 2.17 \times 10^{-3} \) mutations/site/generation (Mansky and Temin, 1995) was used. Furthermore, we used base-dependent mutation rates \( \mu_i = \mu(b_{n_{00}}, b_{n_{11}}) \) that were estimated from in vivo longitudinal data (Deforche et al., 2007). For the estimation, we used \( G_{\max} = 200 \), corresponding to about a year of evolution, given an estimated generation turnover time of \( \pm 1.5 \) days; \( L^E = |p^E| = 10 \times |p^T| \) and \( e = 10^{-7} \).

### 2.3 Validation experiments

#### 2.3.1 Correlation with nelfinavir resistance phenotype
Using a public data set of matched genotype–phenotype pairs for subtype B sequences [from the Stanford HIV Drug Database (Kantor et al., 2001)], estimated fitness was compared to in vitro resistance fold change phenotype. Sequences with unknown amino acid mutations (‘Z’ or ‘X’) were removed, as were sequences with a fold change at the upper detection range of the assay. For each of the remaining 519 amino acid sequences \( j \), fitness \( f_j \) was estimated from resistance fold change \( R_j \), using (Holford and Sheiner, 1982):

\[
\hat{f}_j = \frac{e_j}{1 + e_j} \]

with \( e_j \) the replication capacity of the virus and \( D \) the effective drug concentration. The values \( e_j \) are generally unknown, and \( e_j = 1 \) was assumed for all strains when computing \( \hat{f}_j \). The fitness estimated from the phenotypes was then compared with the fitness computed using the estimated fitness landscape by computing the correlation coefficient, which was indifferent to the value of \( D \).

#### 3 RESULTS

#### 3.1 Simulation experiments
To illustrate convergence properties of the method, the method was tested in two series of simulation experiments: (i) to recover...
coefficients showed considerable variation with values for random and estimated NFV fitness functions. Correlation the same way as the training data. Results were similar for the polymorphisms. The accuracy of the estimated fitness functions was evaluated by correlation of known and estimated fitness for the fitness function, included 271 arcs (of which 33 indicated antagonisms between different mutations at a single position, which is an artifact caused by the boolean data representation), and the corresponding fitness interactions were included in the fitness model. The network was similar to the one described previously (Deforche et al., 2006), but here we allowed for more putative interactions because no bootstrap procedure was run to reduce the entire model. Bootstrap support for an arc reflects the robustness of the arc against sampling effects in the data set, and is only important when inferring conclusions from the presence or absence of arcs.

3.2 Nelfinavir fitness function

3.2.1 Nelfinavir Bayesian network A BN was estimated from the treated sequences to estimate epistatic fitness interactions between the included mutations. The network with highest a posteriori probability, that served as a blueprint for the fitness function, included 271 arcs (of which 33 indicated antagonisms between different mutations at a single position, which is an artifact caused by the boolean data representation), and the corresponding fitness interactions were included in the fitness model. The network was similar to the one described previously (Deforche et al., 2006), but here we allowed for more
estimated fitness landscape reflects the selective pressures on HIV to evolve the necessary mutations to explain the change in prevalence of single mutations or patterns of mutations in HIV isolated from patients failing a specific treatment, compared to treatment naive sequence data. The estimation follows two consecutive steps. First, epistatic fitness interactions between mutations are estimated. Since an interaction between two mutations is expected to lead to a different observed prevalence of one mutation depending on the presence of the other, observed associations in prevalence may indicate such fitness interactions. BN learning was used to search for interactions between mutations as described in Deforche et al. (2006). These interactions were included in a multiplicative fitness function (Sanjuan et al., 2004), which describes fitness as a product of independent contributions of presence of amino mutations, augmented with independent contributions for combinations of interacting mutations. Second, the fitness contributions were estimated using an iterative procedure so that simulated evolution over the fitness landscape of treatment naive sequences resulted in sequences comparable to the sequences from treated patients. Therefore, the fitness function models the part of sequence space bounded by these circulating sequences and does not extrapolate to sequences that are not observed in the epidemic, and for which there is little interest in an improved resistance interpretation.

The estimation was not based on in vitro experimental data, such as resistance fold change assays and fitness assays, or on in vivo correlates of viral fitness such as viral load, but instead was estimated from observed evolution in clinical sequences. Fitness was estimated based on the evolutionary principle that substitutions observed in the consensus sequence of a population under strong selective pressure are mostly fixed to increase the fitness of the population. As such, the increase in prevalence of a particular mutation in the population of sequences after NFV failure, compared to the population of sequences that were NFV naive, reflects the consecutive fixation of mutations in a population that acquires increased fitness under NFV selective pressure. Not only increase in prevalence of individual mutations was considered, but also of patterns of mutations, since epistatic fitness interactions alter the fitness impact of mutations depending on a context of other mutations. The fitness model included n-ary epistatic effects, which were estimated using BN structure learning that has demonstrated its ability to learn epistatic interactions between different protein residues in general (Klingler and Brutlag, 1994) and applied to HIV drug resistance mutations in particular (Deforche et al., 2006).

Experiments with random fitness landscapes showed that the method, at least within the assumptions of the model, could be used to accurately estimate intra-subtype variation in fitness, but in general does not allow learning of inter-subtype variation. This may be explained by the fact that the method relies on observed evolution to distinguish less fit and more fit mutational patterns. Even if, for example, subtype C viruses
would be less susceptible to NFV treatment, and therefore more fit under treatment, a subtype B virus will never evolve to become subtype C. Since only intra-subtype evolution under drug selective pressure is observed, the method cannot estimate the fitness impact of conserved patterns of polymorphisms in different subtypes that are responsible for the fitness difference. This explains the nature of the shifted bands in the scatter plots (Figs 1 and 3 in Supplementary Material E), which disappear when adding subtype into the linear regression model (Figs 2 and 4 in Supplementary Material E). Still, the method will use differences in observed evolution under treatment in different subtypes, to model fitness interactions between some of these polymorphisms and certain resistance pathways.

Since fitness during treatment depends on the susceptibility of the virus to the drug, the fitness estimate can be considered an in vivo resistance phenotype, and should therefore correlate with the in vitro resistance phenotype. While the correlation was found to be highly significant (Fig. 2), discordance was higher than with the random fitness landscapes. Uncertainty about parameters, assumptions and simplifications made by the model may explain part of the observed discordance. However, in this evaluation we may question which one is the better estimate, our estimate, since the method will use differences in observed evolution under treatment in different subtypes, to model fitness interactions between some of these polymorphisms and certain resistance pathways.

Fig. 5. Predicted evolution graph for a subtype G sequence (bold), for which evolution of mutations M891 L90M was observed during NFV treatment (dashed). Legend as in Figure 3.

Fig. 6. Predicted evolution graph for a subtype C sequence (bold), for which evolution of mutations E35D L63P was observed during NFV treatment (dashed). Legend as in Figure 3.

Table 1. Predicted mutations during NFV treatment

<table>
<thead>
<tr>
<th>m</th>
<th>N</th>
<th>n</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>10F</td>
<td>390</td>
<td>17</td>
<td>1.11E-04</td>
</tr>
<tr>
<td>10V</td>
<td>388</td>
<td>11</td>
<td>2.21E-02</td>
</tr>
<tr>
<td>13V</td>
<td>323</td>
<td>34</td>
<td>8.89E-14</td>
</tr>
<tr>
<td>15V</td>
<td>302</td>
<td>28</td>
<td>4.68E-02</td>
</tr>
<tr>
<td>20F</td>
<td>356</td>
<td>13</td>
<td>2.56E-08</td>
</tr>
<tr>
<td>20R</td>
<td>388</td>
<td>12</td>
<td>2.37E-07</td>
</tr>
<tr>
<td>20T</td>
<td>387</td>
<td>15</td>
<td>8.92E-02</td>
</tr>
<tr>
<td>20V</td>
<td>402</td>
<td>4</td>
<td>2.53E-09</td>
</tr>
<tr>
<td>30N</td>
<td>317</td>
<td>35</td>
<td>1.14E-07</td>
</tr>
<tr>
<td>33F</td>
<td>396</td>
<td>5</td>
<td>3.12E-02</td>
</tr>
<tr>
<td>33I</td>
<td>400</td>
<td>5</td>
<td>3.50E-17</td>
</tr>
<tr>
<td>35D</td>
<td>254</td>
<td>29</td>
<td>1.26E-04</td>
</tr>
<tr>
<td>35N</td>
<td>398</td>
<td>6</td>
<td>9.29E-16</td>
</tr>
<tr>
<td>36I</td>
<td>262</td>
<td>32</td>
<td>2.78E-04</td>
</tr>
<tr>
<td>36V</td>
<td>390</td>
<td>7</td>
<td>7.68E-03</td>
</tr>
<tr>
<td>39S</td>
<td>396</td>
<td>2</td>
<td>8.12E-10</td>
</tr>
<tr>
<td>45R</td>
<td>389</td>
<td>9</td>
<td>3.00E-02</td>
</tr>
<tr>
<td>46I</td>
<td>361</td>
<td>41</td>
<td>4.73E-04</td>
</tr>
<tr>
<td>46L</td>
<td>376</td>
<td>15</td>
<td>1.29E-02</td>
</tr>
<tr>
<td>54V</td>
<td>369</td>
<td>10</td>
<td>2.37E-03</td>
</tr>
</tbody>
</table>

Mutations for which predicted variation in rate of selection, based on the genetic context, correlated significantly with observed selection in patients during NFV treatment (FDR < 0.05). N: the number of baseline sequences without the mutation, of which n developed the mutation during treatment. p: P-value for correlation between predicted probability for selection of the mutation and observed selection, after correcting for multiple comparison with Benjamini and Hochberg.

Intrinsic for the recombinant virus assays used, replication capacity of the recombinant virus, and other possible artifacts from the in vitro test environment. The in vivo estimate on the other hand, while based on indirect information, does take into account the influence of replication capacity, and interaction with the immune system through epitopes.

As a model for evolution in the HIV intra-host population, an ideal Wright–Fisher model was assumed, seeded with empirical parameters for HIV intra-host evolution from literature. The model however did not include recombination, assumed a constant population size and no other effects of selection besides the treatment-related fitness function. Each of these assumptions
are unlikely for HIV and this will impact the accuracy of the estimated fitness landscape. Moreover, analytical results lack for fixation time distributions in a Wright–Fisher model. An accurate implementation of this model therefore requires a full simulation, which was avoided because of the high computational cost, through several approximations that are detailed in Supplementary Material C. Because the estimate was optimized to predict evolution, however, the accuracy of predicted evolution should be less affected (since the fitness function will be distorted to predict observed evolution with an approximate evolutionary model). These simplifications and approximations, including the lack of recombination in the model, may be avoided with availability of a more accurate, but also more computationally demanding simulator.

4.2 Predicting evolution during treatment
The fitness landscape together with the evolution simulator were used to predict evolution during treatment, and depending on variation in the baseline sequences (presence of polymorphic and resistance mutations), variation in rate of selection of mutations was predicted that correlated significantly with observed variation in selection for 39 mutations (Table 1). The predictability of selection of a mutation that is more prevalent in isolates after treatment implies the involvement of that mutation in improving fitness in the presence of treatment. However, a mutation whose selection does not depend on genetic context may equally well be an important resistance mutation, while such a mutation would not yield any predictable variation of selection for that mutation. Therefore, predictability should not be interpreted as a quantification of the fitness gain. For example, a common polymorphism such as I64V for which there were 22 occurrences was about as significantly predicted as L90M (which is non-polymorphic) for which there were 59 occurrences. This only indicates that selection of I64V is more dependent on the genetic background than L90M. The 39 mutations included most of the described NFV resistance mutations (Johnson et al., 2005): 10F, 30N, 36I, 46I/L, 71T/V, 77I, 82A, 88D/S and 90M. In addition, there were a number of resistance mutations that have not been associated with NFV but with other protease inhibitors: 10V, 13V, 20I/R, 33I/F, 36V, 54V, 62V, 63P, 85V and 93L, including mutations that have been described only for the more recently introduced PIs atazanavir or tipranavir (such as 13V, 20I, 33I/F, 36V, 62V, 85V and 93L). These novel protease inhibitors may require more mutations before losing clinical utility, but seem to be affected by the same set of mutations that are selected by older inhibitors such as NFV, and therefore cross-resistance may be underestimated. For these novel mutations that are often selected and well predicted, such as 13V, 35D, 62V, 64V or 93L, the fitness landscape contains knowledge about the genetic context that influences the fitness contributions of these mutations. This knowledge could be used for further investigation of the biological role and mechanism through in vitro mutagenesis experiments.

A convenient representation of HIV resistance evolution is as a probabilistic-ordered accumulation of resistance mutations. Such models has been inferred previously from cross-sectional data (Beerenwinkel et al., 2005), describing evolution starting from a 'wild-type' sequence. We extended this approach by creating individualized evolution graphs that predict evolution for any sequence (Figs 3–6), wild type of whatever subtype or recombinant, or partially resistant sequences that just enter higher in the landscape. The examples illustrate how variation in wild-type HIV-1 sequences influences the predicted evolution, for some sequences major or primary mutations arising first, and for other sequences first minor or secondary mutations, all in an ordered stochastic but predictable fashion. Our predictions imply that resistance evolution is highly individual, depending on the baseline sequence of the virus, and thus could be used for guiding individualized treatment choices. For example, the evolution graph in Figure 6 implies a higher predicted genetic barrier, in expected number of mutations before becoming resistant, than the evolution graph in Figure 5, under the assumption that mutations E35D and L63P do not cause resistance to NFV (Van Laethem et al., 2002).

The technique presented here was developed to particularly take into account the large natural diversity of HIV-1, and estimates a fitness landscape that may be used across subtypes. As further discussed in Supplementary Material F, no assumptions were made about wild type and mutation (a common source for a subtype bias when using the subtype B reference strain), and the method deals in a natural way with sequences from other subtypes that have polymorphisms that are considered resistance mutations in a subtype B genetic environment (such as M36I in subtype C). Since access to antiviral medication is expanding beyond patients infected with HIV-1 subtype B, which is the most prevalent subtype in the Western World but almost absent in other parts of the world, resistance development in other subtypes requires more attention.

4.3 Related work and other applications
Beerenwinkel et al. (2005c) used a combination of phenotypic data and a model of ordered resistance evolution learned from cross-sectional data to estimate the genetic barrier against zidovudine and lamivudine. A Poisson process for selection of mutations was assumed to derive the expected selection time from the observed selection probabilities along the branches of mutagenic trees. Each mutagenic tree defines possible evolutionary pathways on a lattice structure (Beerenwinkel et al., 2005a) and the informative trees in the mixture therefore constrain evolution a priori. In contrast, we did not constrain evolution according to a limited number of pathways, but let an evolutionary model, using the shape of the fitness landscape, decide the probabilities of selecting particular mutations.

While the method for estimating a fitness function was developed to obtain a better understanding in HIV drug resistance evolution with the objective of improving prediction of treatment response, it could be applied to other selective pressures for HIV (such as adaptation to immune response) or other organisms. The main requirement is that convergent evolution is observed in a number of isolated populations experiencing the same selective pressure. This makes the technique particularly well suited for fast evolving viruses, which by their nature form relatively isolated intra-host populations.
Other than genotypic sequence data, the method requires estimates of effective population size and mutation rates.

**ACKNOWLEDGEMENTS**

The authors wish to thank T. Silander for review of the Methods, the people who maintain the Stanford HIV Drug Resistance Database, and researchers and clinicians who made data available. These include, amongst others, Van Wijngaerden E., Carvalho A.P., Cabanas J., Valadas M.E., Aguas M.J., Vera J., Rosado L., Batista T., Bezerra V., Soares I., Branco T., Mouzinho A., Teófilo E., Faria T. and Mansingo K. 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 funded by an EMBO long-term fellowship. Y.M. is a post-doctoral researcher with the FWO-Vlaanderen; his research is supported by KU-Leuven GOA-Mefisto-666 and GOA-Ambiorics, Belspo IUAP V-22, and EU FP6 NoE Biopattern. This work was supported by FWO-Vlaanderen grant G.0266.04, by the Katholieke Universiteit Leuven through Grant OT/04/43, and by the VIROlab project (EU IST STREP Project 027446).

**Conflict of Interest:** none declared.

**REFERENCES**


