Phylogenetic surveillance of viral genetic diversity and the evolving molecular epidemiology of human immunodeficiency virus type 1

BIOINFORMATICS ORIGINAL PAPER

Vol. 24 no. 1 2008, pages 34–41
doi:10.1093/bioinformatics/btm540

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. Deforche[1], R. Camacho[2], K. Van Laethem[1], P. Lemey[1,3], A. Rambaut[4], Y. Moreau[5] and A.-M. Vandamme[1,*]

[1]Rega Institute for Medical Research, Katholieke Universiteit Leuven, Leuven, Belgium, [2]Centro Hospitalar de Lisboa Ocidental, Lisbon, Portugal, [3]Department of Zoology, Oxford University, Oxford, [4]Institute for Evolutionary Biology, University of Edinburgh, Edinburgh, UK and [5]ESAT, 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/~kdforc0/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.

*To whom correspondence should be addressed.

[1]In 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 Monis, Lisbon, Portugal, to create a treated population $\mathcal{P}^{\mathrm{T}}$ of 1026 sequences from patients with experience to NFV as sole PI, and a naive population $\mathcal{P}^{\mathrm{N}}$ 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 $\mathcal{P}^{\mathrm{N}}$ to a treated population $\mathcal{P}^{\mathrm{T}}$, 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 $\mathcal{P}^{\mathrm{N}}$ over the current fitness function estimate is simulated, and compared to $\mathcal{P}^{\mathrm{T}}$.

2.2.1 Fitness function structure The protease amino acid sequences from the treated population $\mathcal{P}^{\mathrm{T}}$ 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:

$$P(A_1, \ldots, A_n) = \prod_i^n P(A_i \mid \text{parents}(A_i))$$

with $P(A \mid B)$ 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 $\mathcal{P}^{\mathrm{T}}$ with structure S^{T} and CPD parameters θ^{T} as $\mathrm{BN}^{\mathrm{T}}(\theta^{\mathrm{T}}, S^{\mathrm{T}})$.

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

$$F(A_1, \ldots, A_n) = \prod_i^n F(A_i \mid \text{parents}(A_i))$$

with parents (A_i) the parents in S^{T}, and $F(A \mid B)$ the *Conditional Fitness Contribution (CFC)* of the presence of A, 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): $\theta_{i,k} = P(A_i = 1 \mid \text{parents}(A_i) = k)$. 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_i for every pattern of parent mutations: $\phi_{i,k} = F(A_i = 1 \mid \text{parents}(A_i) = k)$.

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

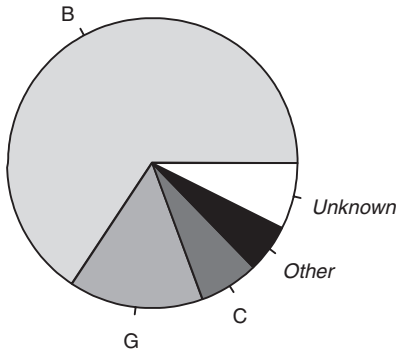

Fig. 1. HIV-1 subtype distribution of NFV treated population. *Other* subtypes were mostly CRFs. Sequences whose subtype could not be determined either because they were recombinant, or because only the protease was available were classified as *Unknown*.

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 (Ewens, 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 $\mathcal{P}^{\mathrm{N}}$ resembles the treated population $\mathcal{P}^{\mathrm{T}}$. Therefore, evolution is simulated for sequences sampled from the naive population $\mathcal{P}^{\mathrm{N}}$ using the fitness function, to obtain an evolved population $\mathcal{P}^{\mathrm{E}}$. The difference between the sequence populations $\mathcal{P}^{\mathrm{E}}$ and $\mathcal{P}^{\mathrm{T}}$, which must thus be minimized, is measured by comparing the parameters of $\mathrm{BN}^{\mathrm{T}}(\theta^{\mathrm{T}}, S^{\mathrm{T}})$ of the treated data set, with $\mathrm{BN}^{\mathrm{E}}(\theta^{\mathrm{E}}, S^{\mathrm{T}})$, 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 the search to a unique solution by minimizing $|\phi_{i,k} - 1|$ for each parameter $\phi_{i,k}$, as a secondary objective (with a low weight compared to the first objective).

An iterative algorithm was used to estimate the parameters. The algorithm starts initially from a flat fitness function (i.e. $F(A_1, \ldots, A_n) = 1$), and in each iteration this function is updated in a step-wise fashion. A population $\mathcal{P}^{\mathrm{E}}$ is computed using the current estimate of the fitness function F. The fitness function parameters $\phi_{i,k}$ are subsequently adjusted based on the difference in the sufficient statistics (which reflect the counts) related to the BN parameters $\theta^{\mathrm{E}}_{i,k}$ and $\theta^{\mathrm{T}}_{i,k}$: $\phi_{i,k}$ is increased with a small multiplicative factor $(1+\delta_{i,k})$ when there is too few of mutation A_i for parent combination k in the evolved versus treated population, or vice versa if there is too much of mutation A_i. By using the sufficient statistics instead of the actual CPD parameters, uncertainty on these parameters is taken into account. The step sizes $\delta_{i,k}$ are dynamically adjusted depending on the convergence of the corresponding $\phi_{i,k}$. Details, pseudo-code, and convergence properties of the algorithm are presented in Supplementary Material D.

The amount of evolution experienced by HIV under treatment depends on many factors such as the baseline viral load, the potency of the combination therapy to suppress residual replication (which depends on the amount of resistance), patient adherence and duration of the therapy. The probability distribution for the effective number of generations under drug pressure between sequences from drug naive and treated patients $P(G^{\mathrm{T}})$, which is used to create $\mathcal{P}^{\mathrm{E}}$, is in general unknown and was assumed to be uniform in the interval $[0, G_{max}]$, with G_{max} a maximum limit for the number of generations. Variation of G_{max} does not affect the shape of the landscape, but its steepness (lower values result in a steeper landscape).

2.2.4 Phylogenetic guide tree The sampling of sequences from $\mathcal{P}^{\mathrm{N}}$ 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-γ substitution model, and codons representing IAS resistance associated positions (Johnson *et al.*, 2005) were excluded to avoid problems of convergent evolution.

Each sequence n^{T} from the treated population $\mathcal{P}^{\mathrm{T}}$ added a contribution $k\, e^{-r\, d}(n^{\mathrm{T}}, n^{\mathrm{N}})$ to the sampling weight of a sequence n^{N}, 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 $\sum_{n^{\mathrm{N}}} k e^{-rd(n^{\mathrm{T}}, n^{\mathrm{N}})} = |\mathcal{P}^{\mathrm{T}}|^{-1}$.

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 $\mu = 2.17\times10^{-5}$ mutations/site/generation (Mansky and Temin, 1995) was used. Furthermore, we used base-dependent mutation rates $\mu_i = \mu(b_{\mathrm{from}}, b_{\mathrm{to}})$ 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 ± 1.5 days; $L^{\mathrm{E}} = |\mathcal{P}^{\mathrm{E}}| = 10 \times |\mathcal{P}^{\mathrm{T}}|$ and $\epsilon = 10^{-7}$.

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 $\hat{f}_j$ was estimated from resistance fold change R_j, using (Holford and Sheiner, 1982):

$$\hat{f}_j = \frac{e_j}{1 + \frac{D}{R_j}}$$

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.

2.3.2 Correlation with observed nelfinavir resistance evolution In 404 patients for which a baseline and consecutive follow-up sequence during NFV treatment was available, the accuracy of the model to predict observed resistance evolution was evaluated. These pairs were independent from the cross-sectional training data. For each wild type and mutation included in the fitness function, correlation of observed evolution (0 or 1) with predicted evolution ($0<p<1$) was evaluated. Correlation with observed evolution was analyzed with a linear model, which included next to the predicted evolution a non-linear correction for the number of observed substitutions for each sequence. Correction for multiple testing was done using the Benjamini and Hochberg method with FDR = 0.05. An observed mixture (such as L63LPA) was evaluated against predicting evolution of the mutations (for this example L63P or L63A). Prediction of loss of mutations in mixtures (such as LPA63P) was not considered.

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

10 random but known fitness functions for HIV, each corresponding to a random fictive selective pressure on protease, and (ii) to re-estimate the fitness function that was estimated for HIV in presence of NFV, from four training data sets of varying size.

Each of the 10 random fitness functions used the same 114 mutations as those considered for the NFV fitness function (listed in Supplementary Material A), and was generated with 100 random interactions (BN arcs), and random fitness contributions for presence of different mutations and patterns of mutations sampled from the distribution $1+U(0,1)^6$. For each random fitness function, a training data set of 1000 sequences was generated by evolving PI naive sequences using the fitness function for g generations, sampled from distribution $P(G^T)$. Similarly, the estimated NFV fitness function was used to create training data sets of size 513, 1026, 2052 and 10 260. The generated training sequences together with the PI naive sequences were then used to estimate the original (random or estimated NFV) fitness function.

A weight w was defined for each arc in a random BN as the change in likelihood of that BN in the corresponding generated data set, when removing the arc, and reflects the strength of the interaction in the generated sequences. For around 60% of arcs, a value of $w<1$ indicated that the implied interaction was not considered during evolution. These arcs were in general not present in the learned BN. Of arcs with $w>1$, over 60% were present in the learned BN, with higher probability for arcs with higher w. Around 12% of arcs in the learned BNs indicated the obvious antagonism between different mutations at a single position (a consequence of the chosen data representation), and other artefact arcs mostly indicated associations between polymorphisms. The accuracy of the estimated fitness functions was evaluated by correlation of known and estimated fitness for 1000 independent validation sequences that were generated in the same way as the training data. Results were similar for the random and estimated NFV fitness functions. Correlation coefficients showed considerable variation with values for R^2 between 0.47 and 0.84 (see Supplementary Material E). Low correlation values were associated with a banding pattern in the scatter plots. When adding subtype as an explanatory variable to the linear regression, these banding patterns disappeared, and correlation improved with R^2 ranging between 0.79 and 0.94. This indicates that the estimated fitness landscapes performed well to explain intra-subtype variation of fitness under treatment, but not inter-subtype variation, which is indeed not related to fitness changes under treatment.

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 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.

3.2.2 Nelfinavir fitness function By comparing the PI naive population with the NFV treated population, the parameters for the fitness function during NFV treatment were estimated using the iterative procedure described in Methods section, and using a phylogenetic guide tree to correct for different epidemiology of naive and treated viruses.

The fitness function modeled contributions and epistatic interactions of well-described resistance mutations. For example, in absence of the well-described D30N major resistance mutation, the model predicted no contribution of the N88D mutation to fitness under NFV selective pressure. In presence of a D30N mutation, the model predicted a contribution of 58% for N88D, and in presence of both D30N and V77I, a contribution of 75%.

3.2.3 Correlation with in vitro *resistance phenotype* Correlation with *in vitro* resistance fold change data for NFV was investigated for 519 sequences (not included in the training data set), assuming no effects of replication capacity and a constant drug concentration. The log estimated *in vitro* fitness showed a reasonable correlation with the log estimated *in vitro* fitness ($R^2 = 0.47$ $(0.41-0.54)$, $p<10^{-15}$, Fig. 2) with no clear trend in the differences.

3.2.4 Correlation with observed evolution The ability of the model for evolution using the estimated fitness function, to predict observed evolution during NFV treatment, was evaluated by comparing predicted evolution with observed evolution in 404 patients treated with NFV. In Figures 3–6 the predicted evolution is shown for some examples of sequences from patients for which the prediction was in agreement with observed evolution, meaning the observed sequence after NFV failure was the sequence with the highest predicted probability compared to the other probable sequences of the same generation. These graphs illustrate variation in prediction, caused by variability in baseline sequence. In Table 1, the mutations were listed for which variation in predicted evolution showed a significant positive correlation with observed evolution, after correcting for multiple testing. Negative correlations were not found for any mutation, and thus we made no overall wrong predictions. For some patients under NFV treatment, only *minor* (or *secondary* mutations) (Shafer, 2002) were predicted and observed in the virus, in absence of any *major* mutation. This may provide an explanation why the initial classification of primary and secondary mutations were untenable (now major and minor) since in reality for some patients secondary mutations were observed first.

4 DISCUSSION

4.1 Estimating *in vivo* fitness during treatment

In this study, we presented a computational method to estimate the *in vivo* fitness function of HIV-1 during treatment. The

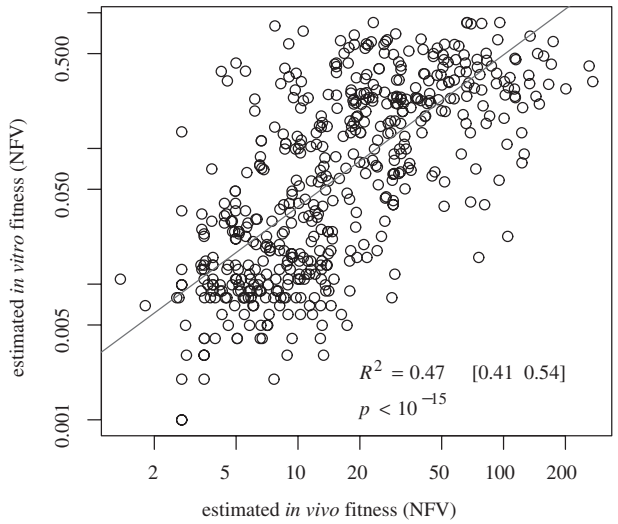

Fig. 2. Estimating the *in vivo* fitness of HIV-1 during NFV treatment. Comparison of estimated fitness from *in vitro* phenotypic resistance data for 519 matched genotype–phenotype pairs, assuming no effects of replication capacity and effective drug concentration $D = 100$.

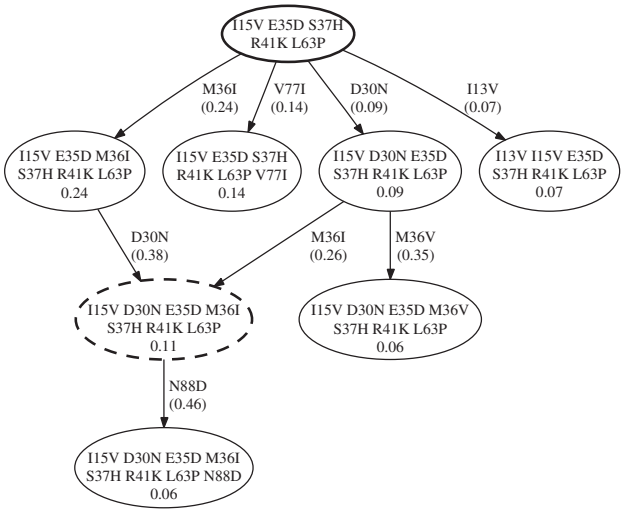

Fig. 3. Predicted evolution graph for a subtype B sequence (bold). Each node corresponds to a sequence that is predicted to evolve from the baseline sequence with the estimated probability given. An arc from sequence X to sequence Y corresponds to a mutation with given probability to evolve sequence X into sequence Y. Only sequences with predicted probability $p > 0.05$ are shown. For this baseline sequence, evolution of mutations D30N M36I was observed during NFV treatment, corresponding to the predicted sequence with two mutations with highest probability (dashed).

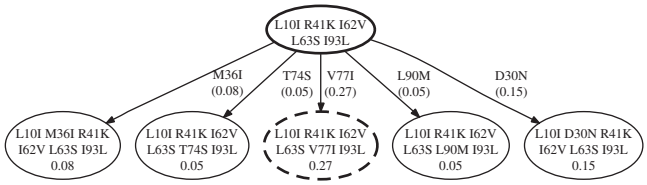

Fig. 4. Predicted evolution graph for a subtype B sequence (bold), for which evolution of mutation V77I was observed during NFV treatment (dashed). Legend as in Figure 3.

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

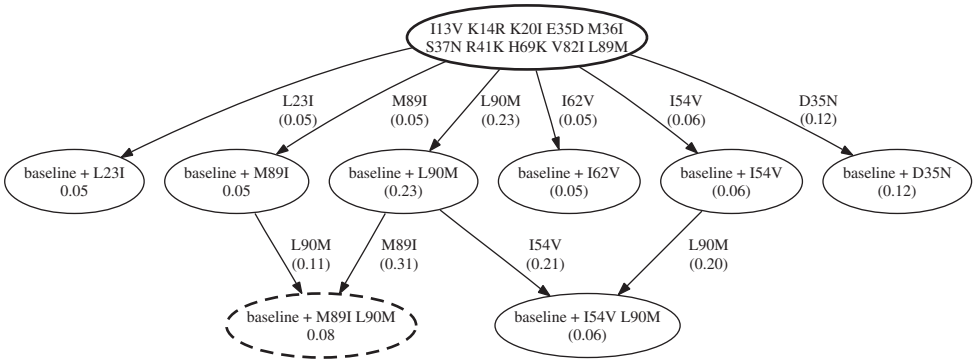

Fig. 5. Predicted evolution graph for a subtype G sequence (bold), for which evolution of mutations M89I L90M was observed during NFV treatment (dashed). Legend as in Figure 3.

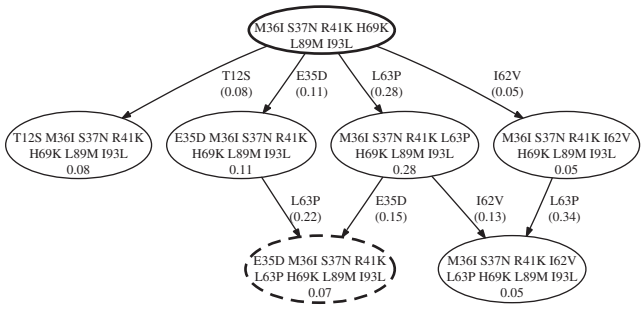

Fig. 6. Predicted evolution graph for a subtype C sequence (bold), for which evolution of mutations E35D L63P was observed during NFV treatment (dashed). Legend as in Figure 3.

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 computational approach taking into account only *in vivo* parameters, or the *in vitro* phenotypic data. One could argue in favor of the *in vivo* estimate, since the *in vitro* data suffers from problems with the *in vitro* resistance assays, such as the reproducibility at low fold changes, recombination artifacts

Table 1. Predicted mutations during NFV treatment

m	*N*	*n*	*p*	*m*	*N*	*n*	*p*
10F	390	17	1.11*E*–04	57R	50	7	1.28*E*–09
10V	388	11	2.21*E*–02	62V	322	42	4.56*E*–07
13V	283	34	8.89*E*–14	63P	133	27	4.38*E*–03
15V	331	18	4.68*E*–02	64V	329	22	5.30*E*–07
20I	356	13	2.56*E*–08	70R	390	7	1.75*E*–04
20R	388	12	2.37*E*–07	71T	366	27	7.45*E*–03
20T	387	15	8.92*E*–07	71V	332	37	6.37*E*–07
20V	402	4	2.53*E*–09	74S	378	21	7.68*E*–04
30N	317	55	5.01*E*–07	77I	277	31	1.27*E*–05
33F	396	5	3.12*E*–02	82A	369	11	1.45*E*–07
33I	400	5	5.30*E*–17	85V	395	6	6.67*E*–07
35D	254	29	1.26*E*–04	88D	355	39	8.95*E*–12
35N	398	6	9.29*E*–16	88S	391	16	3.20*E*–02
36I	262	32	2.78*E*–04	89I	394	7	1.08*E*–11
36V	390	7	7.68*E*–03	89L	74	2	1.15*E*–02
39S	396	2	8.12*E*–10	89V	397	3	4.73*E*–03
45R	389	9	3.00*E*–02	90M	325	59	1.14*E*–07
46I	361	41	4.73*E*–04	92K	400	5	3.98*E*–02
46L	376	15	1.29*E*–02	93L	262	23	1.50*E*–04
54V	369	10	2.37*E*–03				

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., Águas 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 KULeuven 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

Beerenwinkel,N. *et al.* (2005) Learning multiple evolutionary pathways from cross-sectional data. *J. Comput. Biol.*, **12**, 584–598.

Beerenwinkel,N. *et al.* (2005a) Evolution on distributive lattices. URL doi:10.1016/j.jtbi.2006.03.013.

Beerenwinkel,N. *et al.* (2005b) Computational methods for the design of effective therapies against drug resistant HIV strains. *Bioinformatics*, **21**, 3943–3950.

Beerenwinkel,N. *et al.* (2005c) Estimating HIV evolutionary pathways and the genetic barrier to drug resistance. *J. Infect. Dis.*, **191**, 1953–1960.

de Oliveira,T. *et al.* (2005) An automated genotyping system for analysis of HIV-1 and other microbial sequences. *Bioinformatics*, **21**, 3797–3800.

Deforche,K. *et al.* (2006) Analysis of HIV-1 pol sequences using Bayesian Networks: implications for drug resistance. *Bioinformatics*, **22**, 2975–2979.

Deforche,K. *et al.* (2007) Estimating *in vivo* Bias in HIV Mutagenesis Rates. *J. Comput. Biol.*, **14**, 1105–1114.

DiRienzo,G. and DeGruttola,V. (2002) Collaborative HIV resistance-response database initiatives: sample size for detection of relationships between HIV-1 genotype and HIV-1 RNA response using a non-parametric approach. *Antivir. Ther.*, **7**, S71.

Ewens,W.J. (1979) *Mathematical Population Genetics. Biomathematics*. Vol. 9. Springer-Verlag, Berlin, Heidelberg, New York.

Holford,N. and Sheiner,L. (1982) Kinetics of pharamcologic response. *Pharamacol. Ther.*, **16**, 143–166.

Johnson,V.A. *et al.* (2005) Update of the drug resistance mutations in HIV-1: fall 2005. *Top. HIV Med.*, **13**, 125–131.

Kantor,R. *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.

Kantor,R. *et al.* (2006) Role for geographical location and founder effects in describing hiv-1 subtype-specific polymorphisms (author's reply). *PLoS Med.*, **2**, e112.

Klingler,T.M. and Brutlag,D.L. (1994) Discovering structural correlations in α-helices. *Protein Sci.*, **3**, 1847–1857.

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.

Myllymäki,P. *et al.* (2002) B-Course: a web-based tutorial for Bayesian and causal data analysis. *Int. J. Art Intell. Tools*, **11**, 396–387.

Nijhuis,M. *et al.* (1998) Stochastic processes strongly influence HIV-1 evolution during suboptimal protease-inhibitor therapy. *Proc. Natl Acad. Sci. USA*, **95**, 14441–14446.

Rouzine,I.M. and Coffin,J.M. (1999) Linkage disequilibrium test implies a large effective population number for HIV in vivo. *Proc. Natl Acad. Sci. USA*, **96**, 10758–10763.

Sanjuan,R. *et al.* (2004) The contribution of epistasis to the architecture of fitness in an RNA virus. *Proc. Natl Acad. Sci. USA*, **101**, 15376–15379.

Shafer,R.W. (2002) Genotypic testing for human immunodeficiency virus type 1 drug resistance. *Clin. Microbiol. Rev.*, **15**, 247–277.

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

Van Laethem,K. and Vandamme,A.-M. (2006) Interpreting resistance data for HIV-1 therapy management – know the limitations. *AIDS Revi.*, **8**, 37–43.

Van Laethem,K. *et al.* (2002) A genotypic drug resistance interpretation algorithm that significantly predicts therapy response in HIV-1 infected patients. *Antivir. Ther.*, **7**, 1359–6535.

Vandamme,A.-M. (1999) Managing resistance to anti-hiv drugs: an important consideration for effective disease management. *Drugs*, **57**, 337–361.

Vandamme,A.-M. *et al.* (2004) Updated European recommendations for the clinical use of HIV drug resistance testing. *Antivir. Ther.*, **9**, 829–848.