Resistant pathways of human immunodeficiency virus type 1 against the combination of zidovudine and lamivudine

K. Theys,1 K. Deforche,2 P. Libin,2 R. J. Camacho,3 K. Van Laethem1 and A.-M. Vandamme1

1Rega Institute for Medical Research, Katholieke Universiteit Leuven, Leuven, Belgium
2MyBioData, Rotselaar, Belgium
3Instituto de Higiene e Medicina Tropical, Universidade Nova de Lisboa, Lisbon, Portugal

A better understanding of human immunodeficiency virus type 1 drug-resistance evolution under the selective pressure of combination treatment is important for the design of long-term effective treatment strategies. We applied Bayesian network learning to sequences from patients treated with the reverse transcriptase inhibitor combination of zidovudine (AZT) and lamivudine (3TC) to identify the role of many treatment-selected mutations in the development of resistance. Based on the Bayesian network structure, an in vivo fitness landscape was built, reflecting the necessary selective pressure under treatment, to evolve naive sequences to sequences obtained from patients treated with the combination. This landscape, combined with an evolutionary model, was used to predict resistance evolution in longitudinal sequence pairs. In our analysis, mutations 41L, 70R, 184V and 215F/Y were identified as major resistance mutations to the combination of AZT and 3TC, as they were associated directly with treatment experience. The network also suggested a possible role in resistance development for a number of novel mutations. Estimated fitness, using the landscape, correlated significantly with in vitro resistance phenotype in genotype–phenotype pairs \( R^2 = 0.70 \). Variation in predicted evolution under selective pressure correlated significantly with observed in vivo evolution during AZT plus 3CT treatment. In conclusion, we confirmed current knowledge on resistance development to the combination of AZT and 3CT, but additional novel mutations were identified. Moreover, a model to predict resistance evolution during AZT and 3CT treatment has been built and validated.

INTRODUCTION

The goal of highly active antiretroviral therapy is to reduce morbidity and mortality associated with human immunodeficiency virus type 1 (HIV-1) infection. Current therapy consists of a combination of antiviral drugs from at least two different classes and aims to suppress virus replication maximally in order to achieve immunological recovery, to prevent resistance evolution and to prolong life expectancy. The effectiveness of antiviral therapy is compromised by the emergence of drug resistance. For each available drug, considerable efforts have been made to identify individual mutations as well as complex mutational patterns associated with in vitro resistance and reduced clinical response. Genotypic interpretation algorithms use this knowledge to make predictions about phenotypic resistance or expected treatment outcome (Van Laethem & Vandamme, 2006).

Resistance interpretation is complex, as the role of many mutations and the influence of HIV-1 natural variation on resistance development remain insufficiently known. Additionally, these algorithms suffer from their inability to incorporate the evolutionary potential of HIV-1 to develop resistance, i.e. the genetic barrier towards resistance. The order and rate at which HIV-1 evolves over resistance pathways is unknown for most drugs. Computational methods have been proposed to better understand such mutational pathways (Beerenwinkel et al., 2005a; Deforche et al., 2008b). During treatment failure, numerous mutations may accumulate that are not involved directly in resistance, but improve replicative capacity in the presence of resistance mutations. The in vivo fitness, as the combined effect of phenotypic resistance and intrinsic replication capacity, is the driving force of evolution during treatment. A better understanding of the origins and mechanisms of antiviral resistance, and more generally the in vivo viral fitness in the presence of a therapy combination, may lead to improved strategies that will maximize the usefulness of antiviral drugs.
The nucleoside reverse transcriptase inhibitors (NRTIs) zidovudine (AZT) and lamivudine (3TC) inhibit the activity of reverse transcriptase (RT) and are often used as a fixed-dose combination. AZT was the first drug approved for clinical use of HIV-1 treatment. Consequently, extended knowledge is available on the development of resistance against AZT in monotherapy. The most common RT mutations developing during AZT treatment are the well-known thymidine analogue mutations (TAMs) 41L, 67N, 70R, 210W, 215Y/F and 219Q/E. These mutations accumulate in a stepwise manner along two distinct pathways, defined as TAM1 and TAM2 (Cozzi-Lepri et al., 2005). It is still unclear whether differentiation of these pathways has implications for clinical disease progression. Effectiveness of 3TC is compromised by a single mutation, 184V. Interaction between mutations selected by both drugs has been reported. Low-level resistance to AZT can be reversed with the emergence of the 184V mutation (Wainberg et al., 2005). This resensitization effect may explain why this dual NRTI backbone has been found to be both potent and slow in resistance development. Whilst resistance development towards each of these drugs separately has been well studied, it is not sufficiently known how the combination affects resistance pathways. As there is a strong interaction between both resistance pathways, mapping resistance development towards the combination may improve our prediction systems significantly.

Bayesian network learning (BNL) has previously been shown to be useful to map epistatic interactions important for antiviral resistance development (Deforche et al., 2006, 2007a). A Bayesian network (BN) is a probabilistic model that describes statistical independencies between variables (Pearl, 1988). Conditional independences make it possible to elucidate a role for mutations selected during treatment, to identify resistance pathways and to investigate influences of background polymorphic positions. Furthermore, observed associations in prevalence between mutations may reflect epistatic fitness interactions between mutations and therefore can be used to model a fitness landscape. Deforche et al. (2008b) recently developed a mathematical model to estimate an in vivo fitness function, reflecting the required selective pressure for HIV-1 to evolve the necessary mutations in order to explain the change in prevalence of mutations or mutational patterns before and after treatment experience. Simulation of evolution over a fitness landscape makes it possible to predict evolution during treatment and to define the genetic barrier to resistance.

In this paper, we investigated resistance pathways against the combination of AZT + 3TC as sole NRTIs in first-line therapy, using computational techniques. Epistatic interactions between RT amino acids were identified by using BNL and were used as a template to model the in vivo fitness landscape experienced by HIV-1 under AZT + 3TC selective pressure. The technique takes into account the large natural diversity of HIV-1, in order to develop a model that could be used across subtypes.

RESULTS

BNL

The dataset consisted of 1124 sequences from AZT + 3TC-experienced patients, of which 420 (37 %) had additional non-nucleoside RT inhibitor (NNRTI) experience, and 2307 sequences from RT inhibitor (RTI)-naive patients sampled from a dataset of 4950 sequences. The subtype distribution of the entire dataset was B (49 %), G (16 %), C (10 %), A (10 %), 02_AG (5 %), F (2 %) and other (8 %). Statistical analysis identified 64 treatment-associated mutations with a prevalence >0.5 %, of which seven were polymorphic (36E, 39T, 98S, 135T, 162S, 173K and 207Q). Also identified were known NNRTI-resistance mutations 90I, 103N, 106M, 108I, 181C, 188L, 190A, 190S, 221Y and 225H, as NNRTIs frequently accompany an NRTI backbone. For each included treatment-associated mutation, excluding known NNRTI mutations, the prevalence and subtype distribution is shown in Fig. 1 (see Supplementary Table S1, available in JGV Online, for subtype distribution for all mutations included in the network). Many of these mutations have been reported previously as being selected by AZT or 3TC. The final dataset included 106 variables: treatment-associated mutations, mutations anti-associated with treatment and wild-type amino acids (including polymorphisms), together with an AZT + 3TC- and an NNRTI-experience node.

BNL discovered many robust interactions between the variables: the fraction of arcs with bootstrap support >65 % was high. The consensus network shown in Fig. 2 includes 235 arcs, of which 153 have bootstrap support >65 %. Mutations 41L, 70R and 184V were connected with AZT + 3TC treatment with high bootstrap support (>80 %); the antagonistic association with wild-type 215T indicated an association with mutations 215F and 215Y (94 % bootstrap). Postulated semantics make it possible to consider these mutations as major mutations in AZT + 3TC-resistance pathways. It has been documented widely that AZT resistance follows two distinct pathways, TAM1 and TAM2, characterized by the initial appearance of mutations 215Y and 70R, respectively (Beerenwinkel et al., 2005c; Cozzi-Lepri et al., 2005). These two pathways are not exclusive, as we observed a combination of mutations 215Y and 70R.

All TAM mutations were present in the network. TAM1 mutations 41L, 210W and 215Y were associated with mutations 43R, 43Q, 44D, 98G, 118I and 208Y. TAM2 mutations 67N, 70R, 215F and 219Q/E were associated with minor mutations 20R, 214L, 218E and 228H. Mutation 184V was connected to 31L, 62V, 98G, 135L, 142V and 203K. The multi-NRTI-resistance pathway associated with insertions at positions 69 was not significantly present in our dataset (P =0.56, Fisher’s exact test). The 151M pathway with mutations 75I, 77L and 116Y connected directly to position 219 and indirectly to position 190 through 132L. NNRTI-resistance pathways
were marked by mutations 103N, 106A/M, 181C, 188L, and 190A/S (see Supplementary Fig. S1, available in JGV Online). The 103N pathway was further associated with mutations 101Q, 108I, 138Q, 221Y, and 225H, the 181C pathway with 108I and 221Y, and the 190A/S pathway with mutations 101E, 132L, and 138Q. Interactions between NNRTI- and NRTI-resistance pathways exist, but are rare (Deforche et al., 2008a). AZT + 3TC-resistance pathways were not influenced substantially by genetic diversity, as associations between resistance mutations and polymorphisms were not observed frequently. TAM1 mutation 208Y showed an interaction with polymorphism 211K, 184V with 135L, and NNRTI mutation 90I with 162A/C.

**Fitness landscapes**

A BN was learned from 890 sequences from AZT + 3TC-treated patients to estimate epistatic interactions between mutations. The network with the highest a posteriori probability included 100 mutations and 230 arcs, and the corresponding epistatic fitness interactions were included in the fitness model (see Supplementary Table S2, available in JGV Online, for list of mutations included in the fitness function). The network was more complex than the one described in the previous paragraph; we allowed for more putative interactions because no bootstrap procedure was run to reduce the entire model. Fitness function parameters were estimated by using an iterative procedure by comparing 4950 sequences from RTI-naive patients with sequences from treated patients, using a phylogenetic guide tree to correct for different epidemiology.

Correlation of estimated fitness with *in vitro* resistance fold change for AZT + 3TC was investigated by using 687 sequences from an independent dataset, assuming no effects of replication capacity and a constant drug concentration. The log estimated *in vivo* fitness showed a good correlation with the log estimated *in vitro* fitness ($R^2 = 0.70$), with no clear trend in the differences (Fig. 3).

The ability of the model, using the estimated fitness function, to predict observed evolution during AZT + 3TC treatment was evaluated by comparing predicted evolution with observed evolution in 447 patients treated with AZT + 3TC. In Figs 4 and 5, 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 failure was the sequence with the highest predicted probability compared with the other probable sequences. For Fig. 6, the most probable sequence does not match with the observed sequence after failure. However, our model predicted the emergence of TAM2, and observed evolution showed two out of four TAM2 mutations. These graphs highlight variation in prediction, caused by variability in baseline sequence. Table 1 lists the mutations for which variation in predicted evolution showed a significant positive correlation with observed evolution (see Supplementary Table S3, available in JGV Online, for results of all mutations). Negative correlations were not found for any mutation, and thus we made no overall wrong predictions. The prediction of mutation M184V was not statistically significant ($P = 0.2$), probably because of its high prevalence, lacking predictive potential in the dataset.
Fig. 2. Annotated AZT + 3TC experience BN showing direct influences between therapy-associated mutations, polymorphisms and AZT + 3TC experience. Influences between mutations and NNRTI treatment experience are not shown (see Supplementary Fig. S1 for the complete BN figure). An arc represents a direct dependence between the corresponding variables, and thickness is proportional to bootstrap support. Arc colour indicates whether it is a direct influence between resistance mutations (black), an influence from background polymorphisms on resistance mutations (blue), a direct influence between treatment (purple), associations between background polymorphisms (green) or other interactions. Arc direction does not reflect either cause and effect or order of accumulations of mutations, but may indicate a non-additive multivariable effect, which requires analysis of the quantitative components of the BN. Arcs were coloured according to their function to improve reading the graph, but colouring is only indicative.
DISCUSSION

The use of BNL has proved to be a valuable tool to untangle complex mutational patterns selected by protease inhibitors and NNRTIs in clinical sequences from diverse subtypes (Deforche et al., 2007a, 2008a). Here, we applied BNL to HIV-1 sequence data exposed to the NRTI combination of AZT and 3TC to identify treatment-associated amino acid changes and epistatic interactions between these mutations. Based on higher prevalence in sequences from treated versus untreated patients, together with the robust association with a drug or known resistance mutation, we confirmed the selection of many known mutations involved in AZT+3TC-resistance pathways, but also provided support for novel mutations.

Resistance pathways to AZT and 3TC have been well studied. A group of RT mutations, known as TAMs, are associated with AZT resistance. HIV-1 develops TAMs by one of two distinct pathways, characterized by mutations 41L, 210W and 215Y (TAM1) or by 67N, 70R, 215F and 219Q/E (TAM2). Other mutations seem to cluster with these pathways; 44D and 118I tend to cluster with TAM1 and are believed to be compensatory mutations that improve replication capacity (Cozzi-Lepri et al., 2005). Selection of 184V leads to high-level 3TC resistance.
(Boucher et al., 1993). Dual therapy can lead to virus strains dually resistant to both drugs (Miller et al., 1998).

Our results indicated that 41L, 70R, 184V and 215Y/F are major mutations in AZT+3TC resistance, as they were linked directly to the drug node (Fig. 2). This is in agreement with current knowledge (Johnson et al., 2008) and they represent the major pathways TAM1, TAM2 and 184V to AZT+3TC resistance (Boucher et al., 1992). These pathways were associated with a number of additional, novel mutations that can be regarded as minor mutations given the postulated semantics, i.e. dependent on the presence of major resistance mutations instead of treatment. Their biological role could be to increase resistance further or to compensate for the impaired replication capacity after emergence of a major mutation. In particular, mutations 20R, 43R/Q, 44D, 118I, 135L, 203K, 208Y, 214L, 218E and 228H were associated with AZT+3TC treatment.

Two other pathways to AZT resistance are known, but are rather rare (Van Vaerenbergh et al., 2000). The multi-NRTI-resistance 151M complex usually occurs alone, but can be associated with the TAM2 profile and mutation 219E in particular (Cozzi-Lepri et al., 2005; Rhee et al., 2007). In our network, 151M was associated with position 219. The appearance of the 151M complex is associated with long-term antiretroviral treatment and associated

Fig. 5. Predicted evolution graph 2. Legend as for Fig. 4.

Fig. 6. Predicted evolution graph 3. Legend as for Fig. 4.
Table 1. Mutations for which predicted evolution correlated with observed substitution

Predicted mutations during AZT + 3TC treatment were based on 429 longitudinal sequence pairs. The number of baseline sequences without the mutation (N), number of observed substitutions (n) and adjusted P value are given. All mutations stated in the drug-resistance mutations list of the International AIDS Society for AZT and 3TC are present. Variations in predicted probabilities are based solely on baseline sequence and thus because of epistatic interactions between residues. The prediction of 184V was not statistically significant.

<table>
<thead>
<tr>
<th>Mutation</th>
<th>N</th>
<th>n</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>20R</td>
<td>398</td>
<td>16</td>
<td>1.9 × 10^{-02}</td>
</tr>
<tr>
<td>35T</td>
<td>365</td>
<td>7</td>
<td>2.1 × 10^{-03}</td>
</tr>
<tr>
<td>39A</td>
<td>420</td>
<td>15</td>
<td>4.6 × 10^{-03}</td>
</tr>
<tr>
<td>41L</td>
<td>349</td>
<td>41</td>
<td>6.3 × 10^{-16}</td>
</tr>
<tr>
<td>43E</td>
<td>429</td>
<td>6</td>
<td>2.5 × 10^{-05}</td>
</tr>
<tr>
<td>43Q</td>
<td>431</td>
<td>8</td>
<td>3.8 × 10^{-08}</td>
</tr>
<tr>
<td>43R</td>
<td>436</td>
<td>2</td>
<td>1.6 × 10^{-02}</td>
</tr>
<tr>
<td>44D</td>
<td>431</td>
<td>8</td>
<td>2.1 × 10^{-03}</td>
</tr>
<tr>
<td>49R</td>
<td>409</td>
<td>10</td>
<td>1.0 × 10^{-07}</td>
</tr>
<tr>
<td>67N</td>
<td>390</td>
<td>27</td>
<td>1.2 × 10^{-17}</td>
</tr>
<tr>
<td>69D</td>
<td>427</td>
<td>5</td>
<td>1.1 × 10^{-04}</td>
</tr>
<tr>
<td>69N</td>
<td>426</td>
<td>7</td>
<td>5.0 × 10^{-02}</td>
</tr>
<tr>
<td>70R</td>
<td>381</td>
<td>39</td>
<td>2.3 × 10^{-09}</td>
</tr>
<tr>
<td>102K</td>
<td>28</td>
<td>5</td>
<td>4.8 × 10^{-02}</td>
</tr>
<tr>
<td>118I</td>
<td>418</td>
<td>10</td>
<td>6.1 × 10^{-18}</td>
</tr>
<tr>
<td>135L</td>
<td>424</td>
<td>14</td>
<td>4.0 × 10^{-08}</td>
</tr>
<tr>
<td>135M</td>
<td>437</td>
<td>9</td>
<td>1.1 × 10^{-02}</td>
</tr>
<tr>
<td>162Y</td>
<td>432</td>
<td>6</td>
<td>8.7 × 10^{-08}</td>
</tr>
<tr>
<td>174K</td>
<td>400</td>
<td>4</td>
<td>4.1 × 10^{-02}</td>
</tr>
<tr>
<td>208Y</td>
<td>435</td>
<td>8</td>
<td>2.2 × 10^{-08}</td>
</tr>
<tr>
<td>210W</td>
<td>393</td>
<td>18</td>
<td>1.7 × 10^{-19}</td>
</tr>
<tr>
<td>211A</td>
<td>444</td>
<td>2</td>
<td>1.9 × 10^{-07}</td>
</tr>
<tr>
<td>211K</td>
<td>241</td>
<td>15</td>
<td>3.4 × 10^{-02}</td>
</tr>
<tr>
<td>211S</td>
<td>422</td>
<td>2</td>
<td>4.8 × 10^{-04}</td>
</tr>
<tr>
<td>214F</td>
<td>66</td>
<td>6</td>
<td>2.2 × 10^{-03}</td>
</tr>
<tr>
<td>215F</td>
<td>424</td>
<td>15</td>
<td>1.3 × 10^{-02}</td>
</tr>
<tr>
<td>215I</td>
<td>444</td>
<td>4</td>
<td>2.9 × 10^{-03}</td>
</tr>
<tr>
<td>215Y</td>
<td>361</td>
<td>25</td>
<td>4.8 × 10^{-16}</td>
</tr>
<tr>
<td>218E</td>
<td>438</td>
<td>9</td>
<td>1.1 × 10^{-04}</td>
</tr>
<tr>
<td>219E</td>
<td>438</td>
<td>10</td>
<td>2.2 × 10^{-03}</td>
</tr>
<tr>
<td>219Q</td>
<td>421</td>
<td>7</td>
<td>3.1 × 10^{-17}</td>
</tr>
<tr>
<td>228H</td>
<td>433</td>
<td>9</td>
<td>6.3 × 10^{-03}</td>
</tr>
</tbody>
</table>

inversely with 3TC administration (Zaccarelli et al., 2004), which could explain the low prevalence observed. As a subset of patients received NNRTIs in combination with AZT + 3TC, an NNRTI experience node was added to correct for false interactions. Major mutations in NNRTI-resistance pathways were 103N, 106A/M, 181C, 188L and 190A/S. Additional mutations were 101E/Q, 108I, 132L, 138Q, 221Y, 225H and 228R.

Several minor mutations identified here have been associated with NRTI treatment. The presence of mutations 43Q and 208Y has been linked to TAM1, whilst mutations 20R and 218E have been associated with TAM2 (Svicher et al., 2006; Gonzales et al., 2003; Rhee et al., 2005; Saracino et al., 2006). The exact role of other minor mutations in the regulation of resistance is less clear. We observed a robust interaction between 228H and the TAM2 pathway, suggesting a possible role in resistance. Mutations at position 228 have been linked to NRTI treatment (Deforche et al., 2008a; Gonzales et al., 2003; Rhee et al., 2005; Shahriar et al., 2009). Others have suggested a role in NNRTI resistance for mutations at this position (Ceccherini-Silberstein et al., 2007; Saracino et al., 2006). However, Saracino et al. (2006) did not differentiate between mutations 228H and 228R, and their univariate analysis of the impact on susceptibility cannot exclude influences from other mutations present, whilst the study of Ceccherini-Silberstein et al. (2007) only provides weak evidence for mutation 228H and rather suggests an involvement in NRTI resistance. In our BN, mutation 228R is associated weakly with NNRTI treatment and position 215, so a dual role cannot be excluded. Mutation 31L has been linked to resistance to the new NNRTI etravirine (Vingerhoets et al., 2005), but also associated with NRTI treatment (Shahriar et al., 2009). An interaction between 31L and 184V was reported previously (Deforche et al., 2008a), and this association is seen in our network. We could not confirm some other previously reported interactions. Ceccherini-Silberstein et al. (2007) reported a synergistic interaction between polymorphism 135T and mutation 103N for efavirenz resistance; however, in our analysis, 135T is linked to 60I. Mutation 60I has been found in patients treated with AZT (Shafer et al., 1995).

An association of a mutation with treatment by comparing its prevalence in the treated and naive population, although useful for initial screening, does not indicate the relative role of the mutation. For example, a higher prevalence of a mutation in the treated population may be of little clinical predictive value if the mutation further increases resistance only in the presence of other mutations already compromising clinical response. Pairwise covariation provides a first indication of possible antagonistic or synergistic interactions between mutations and clustering techniques allow grouping of mutations into distinct resistance pathways. By simultaneously investigating interactions between treatment and selection of major mutations, between major and minor mutations and between background polymorphisms and resistance mutations, BNL enables a better definition of mutational pathways.

Cozzi-Lepri et al. (2005) reported that mutation 67N seemed to have comparable chances of being selected irrespective of whether 215Y or 70R had been previously selected. This finding is confirmed by the robust interaction between positions 215 and 67, indicating that the TAM1 and TAM2 pathways are not mutually exclusive. The association of 208Y with polymorphism 211K was previously observed together with TAM1 (Sturmer et al., 2003). Although interactions between NRTI- and NNRTI-resistance pathways have been observed, they are rare in our network. The low genetic barrier of 3TC resulted in a
very high prevalence (59%) of 184V. Consequently, 184V will act as a substitute for early therapy failure and be selected at failure, together with a large variety of other mutations. As we model associated prevalences of mutations, mutations may be linked to 184V without there necessarily being a biological interaction in the development of resistance. An important conclusion from this BN profile is that resistance to AZT + 3TC does not seem to be subtype-dependent, as robust interactions between resistance mutations and polymorphisms are lacking.

Deforche et al. (2008b) have developed a method to model drug selective pressure on HIV-1 to evolve the necessary mutations in sequences from untreated patients to sequences from treatment-experienced patients. The model relates differences in prevalence of mutations and mutational patterns between naive and experienced patients with the selective advantage of these mutations or patterns. The resulting fitness function was combined with an evolutionary model to predict resistance evolution during treatment. Predicted variation in the selection rate of mutations, depending on the presence of polymorphic and resistance mutations in the baseline sequence, correlated significantly with observed variation in selection for 32 mutations. The predictability of selection of a mutation that is more prevalent after treatment implies the involvement of that mutation in improving fitness during 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 (Deforche et al., 2008b). For example, there was no significant prediction for 184V, meaning that selection is not dependent on the genetic background. Previously, the performance of this fitness landscape has been evaluated retrospectively to predict virological outcome in a clinical cohort of HIV-1 patients starting on AZT + 3TC plus nelfinavir (NFV). A higher genetic barrier, quantified per additional mutation required to develop AZT + 3TC resistance, was associated significantly with higher log viral load reduction in the short term and with lower odds of virological failure in the long term (Deforche et al., 2008c).

A qualitative model of AZT resistance, describing the order of mutation accumulation as a directed graph, has been generated using longitudinal data (Boucher et al., 1992). Evolutionary pathways during first-line AZT + 3TC dual therapy have been reconstructed quantitatively from cross-sectional data using probabilistic graphical models (Beerenwinkel et al., 2005b). Their model of the evolutionary process describes evolution as the ordered accumulation of permanent genetic changes, starting from a uniform wild type, following a fixed number of possible trees (Beerenwinkel et al., 2005c). A restricted BN is used to model a specific type of dependence between mutations, each BN is a tree structure where nodes are mutations, and a ‘child’ mutation only develops in presence of the ‘parent’ mutations. A strict ordering of resistance mutations, however, is not always appropriate to describe the stochastic effects that apply to HIV-1 evolution. The model developed by Deforche et al. (2008b) does not restrict simulation of evolution to a limited number of pathways, and includes not only major resistance mutations but also treatment-associated mutation (including polymorphisms). An evolutionary model, using the fitness function shape, decides the probabilities of selecting particular mutations. In addition, we applied this technique to a much larger dataset than the one used by Beerenwinkel et al. (2005c).

Studying resistance to NRTIs from clinical data is often complicated by the presence of NNRTIs with resistance mutations in the same region. As a result, declaring a potential role in resistance development for a particular inhibitor must be done with care. Secondly, acquired resistance to one of the drugs in the combination therapy will facilitate emergence of mutations conferring resistance to the other drugs, given that virus replication is less suppressed. Therefore, associated prevalence of resistance mutations may not necessarily imply a biological interaction. This study is limited by the restriction of only including mutations up to RT position 230, whilst it has been known that mutations in the RNase H domain affect susceptibility towards AZT treatment (Gotte, 2007). Overall, the associations observed in this study provide neither conclusive evidence for a biological role in pathways to AZT + 3TC resistance, by having either a phenotypic effect or an impact on replication capacity, nor information on the clinical relevance. The knowledge gained, however, could guide the design of in vitro mutagenesis experiments in order to confirm the hypothesized role of a particular mutation within its suggested context of other mutations.

In conclusion, by studying the resistance pathways to the combination of AZT + 3TC using computational techniques, we confirmed the role of many mutations, but we also identified additional mutations involved in resistance development. Using mutational interactions identified by BNL, we estimated a fitness landscape of HIV-1 under the selective pressure of AZT + 3TC. By simulating evolution over this landscape, resistance accumulation in longitudinal sequence pairs was predicted, and we previously reported the usefulness of this fitness landscape in predicting therapy response in patients receiving NFV + AZT + 3TC. By obtaining a better understanding of HIV-1 drug-resistance evolution, prediction of treatment response can be improved.

**METHODS**

**Clinical dataset.** Clinical data were pooled from the Stanford HIV Drug Resistance Database, from the University Hospital Leuven, Belgium, and from Hospital Egas Moniz, Lisbon, Portugal (Deforche et al., 2006; Kantor et al., 2001). Sequences from patients naive to RTIs and from patients treated with the combination AZT + 3TC, as
sole NRTI components within a first-line therapy, were included in the analysis. Sequences in the untreated population showing evidence of transmitted NRTI resistance, defined by Bennett et al. (2009), were excluded. At most one sequence per treated patient and one sequence per untreated patient were considered, and duplicate sequences within each population were removed. Subtyping was done by using REGA HIV-1 subtyping tool V2.0 (de Oliveira et al., 2005). Protease-inhibitor–experienced patients were included in the RTI-naive patient population, assuming that selective pressure on protease will not affect resistance evolution in RT. Mutations up to position 230 were considered in the analysis. Nucleotide ambiguities that occur commonly in the population sequences were resolved by randomly substituting the mixture with a parent pure nucleotide.

Phylogenetic guide tree. Correcting for the confounding effect of different epidemiologies when identifying resistance mutations, both for BNL and for estimating a fitness landscape, was done by inferring a neighbour-joining tree including all sequences from the untreated and treated population, using PAUP (Swofford, 2000) with the HKY+ substitution model. Codons representing IAS resistance-associated positions (Johnson et al., 2008) were excluded to avoid problems of convergent evolution when estimating the phylogenetic tree. A naive dataset is created by sampling sequences from the naive population in the phylogenetic tree, giving more weight to sequences from the naive population that are epidemiologically closest to the treated population (Deforche et al., 2008b). See Supplementary Methods S2 (available in JGV Online) for more information on the use of phylogeny to obtain equal epidemiological distributions.

BNL to uncover resistance pathways. We applied BNL to identify resistance pathways to AZT+3TC, following closely the method described by Deforche et al. (2006). The dataset was composed by combining sequences from treated patients with a sample of sequences from the naive population, according to a ratio of one-third treated and two-thirds untreated (as sequences from treatment naive patients were abundantly available). Amino acid changes associated with AZT+3TC experience were identified by testing for conditional independence between a mutation and treatment using the Cochran–Mantel–Haenszel $\chi^2$ test, correcting for multiple comparisons using Benjamini–Hochberg with a false discovery rate (FDR) of 0.05. Boolean variables included were beside identified mutations also amino acids at polymorphic positions with a prevalence $>15\%$ in the untreated population, and a drug-experience node ($\text{AZT+3TC}$). As for some patients the RT is also under selective pressure of NNRTIs, an NNRTI-experience node was added to correct for false-positive associations. The most prevalent mutation at each position was considered the wild-type amino acid and this was omitted from the analysis when only one other mutation was present at that position, as its presence was implied by the absence of any mutation.

BNL was done using the B-course software (Myllymäki et al., 2002), searching for the network that explains a maximum of the observed correlations in the data using a minimum number of direct influences. Dependences were visualized in a directed acyclic graph and formed the qualitative component of a BN. Each node corresponded to a variable, and an arc between nodes encoded an unconditional dependence, which represented a direct influence. The robustness of the network was assessed with a non-parametric bootstrap using 100 replicates. In the network graph, obvious strong antagonistic direct influences between different amino acids at a single position were not shown. Network features (presence or absence of an arc) with bootstrap support $>65\%$ were considered robust and shown as solid arcs. Dashed arcs have bootstrap support $>35\%$, and lack of arcs was also considered robust. Known resistance mutations were those defined in the drug-resistance mutations list of the International AIDS Society (IAS) (Johnson et al., 2008) or included in the resistance score in at least one of the latest versions of public resistance interpretation systems (Rega V8.0.1, ANRS V18, HIVDB V5.1.2s).

The biological role of treatment-associated mutations is to confer lowered drug susceptibility and/or to restore replication-capacity defects caused by drug-resistance mutations. A major mutation confers phenotypic resistance on its own, whilst a minor mutation further increases drug resistance only in the presence of a major mutation or compensates for a possible fitness impact of other mutations, and is therefore only selected in the presence of these other mutations, regardless of the drug used. A semantic meaning of the BN with respect to drug resistance can be postulated for the presence of arcs in the network between amino acids and for the network structure around the drug node. As a minor mutation interacts epistatically with a corresponding major mutation, the BN indicates this relationship by an arc between these mutations. The presence of a minor mutation is dependent on the presence of the corresponding major mutation, and thus expected to be unconnected to the treatment node in the BN. In contrast, the selection of a major mutation, usually as a first mutation, is unconditionally dependent on treatment, and thus expected to be connected to treatment. Arc colouring reflected the estimated semantic meaning, in order to improve the interpretation of the graph.

Fitness landscapes. A second analysis in this study comprised the construction of an HIV-1 fitness function ($F$) under AZT+3TC selective pressure from cross-sectional data, as described previously in detail (Deforche et al., 2008b). To learn $F$, we find a function that fits with the evolution of the virus from a naive population to a treated population. This function incorporated interactions indicated using BNL and its parameters were estimated by using an iterative procedure where evolution of sequences from a naive population is simulated, and the evolved sequences are compared with sequences from a treated population. 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.

The BNL analysis described earlier in this study was designed to identify amino acid changes associated with treatment experience using a statistical test. Differently, BNL was now only applied to sequences from treated patients in order to learn epistatic fitness interactions between amino acids. As variables, mutations with a prevalence $>1.5\%$ in the treated population were included, excluding the wild-type amino acid at each position and mutations at positions involved in NNRTI resistance. These positions were defined according the IAS list and interpretation of the BN described earlier in this study, which resulted in exclusion of mutations at positions 90, 101, 103, 106, 108, 181, 188, 190, 221, 225 and 228. First, epistatic fitness interactions between mutations are estimated. As 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. BNL was used to search for interactions between mutations. These interactions were included in a multiplicative fitness function, which describes fitness as a product of independent contributions for combinations of interacting mutations. Secondly, the fitness contributions were estimated by using an iterative procedure so that simulated evolution over the fitness landscape of sequences from naive patients resulted in sequences comparable to the sequences from treated patients. This naive population was compiled by using a sampling process similar to that described above, to avoid a bias from epidemiological dependences. Starting from a flat fitness function, a simulator of HIV-1 intra-host evolution makes the connection between sequences from untreated patients, treatment selective pressure and sequences from
patients failing treatment by evolving a sequence from the naïve population over the current fitness function estimate. Fitness parameters were increased or decreased when too low or too high prevalence of mutations was observed, thereby minimizing differences in prevalence of mutational patterns between the predicted and the observed population.

The stochastic model of HIV-1 evolution that was used for estimating the fitness function will also be used, once the model has been built, to predict evolution during treatment of a particular sequence. Briefly, the HIV-1 intra-host population was modelled by a finite ideal Wright–Fisher population based on biological principles of mutation and selection, using empirical estimates of effective population size and mutation rates, added with selection coefficients based on the fitness function. Evolution by the model considered single nucleotide mutations steps only, except at positions 215 and 151, as double nucleotide mutations are required for these residues to evolve to resistance. An extensive description of fitness landscapes and the mathematical model of evolution is available in Supplementary Methods S1 (available in JGV Online) and Deforche et al. (2007b, 2008b).

Validation experiments. Estimated fitness was compared with in vitro resistance fold change phenotype by using an independent, public set of matched genotype–phenotype pairs for subtype B sequences (Kantor et al., 2001). Sequences with ambiguities resulting in unknown amino acid mutations were removed, as well as sequences from virus isolates for which the phenotypic fold change was at the upper detection range of the assay. For each of the remaining sequences, a fitness value $f_j$ was estimated from resistance fold change $R_{AZT}$ and resistance fold change $R_{3TC}$ using $f_j = e_j [1 + (D/R_{AZT}) + (D/R_{3TC})]$, with $e_j$ representing the replication capacity of the virus and D the effective drug concentration (Holford & Sheiner, 1982). The values $e_j$ are generally unknown, and $e_j = 1$ was assumed for all strains when computing $f_j$. 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.

The performance of the model to predict observed evolution during treatment was evaluated on a dataset of longitudinal sequence pairs, sampled before, during or after the AZT + 3TC regimen. These sequences were not included in the cross-sectional training dataset. For each wild type and mutation included in the fitness function, sequences were not included in the cross-sectional training dataset. Estimated fitness was compared with $RAZT$ and $R3TC$ using $f_j = e_j [1 + (D/R_{AZT}) + (D/R_{3TC})]$, with $e_j$ representing the replication capacity of the virus and D the effective drug concentration (Holford & Sheiner, 1982). The values $e_j$ are generally unknown, and $e_j = 1$ was assumed for all strains when computing $f_j$. 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.

The stochastic model of HIV-1 evolution that was used for estimating the fitness function will also be used, once the model has been built, to predict evolution during treatment of a particular sequence. Briefly, the HIV-1 intra-host population was modelled by a finite ideal Wright–Fisher population based on biological principles of mutation and selection, using empirical estimates of effective population size and mutation rates, added with selection coefficients based on the fitness function. Evolution by the model considered single nucleotide mutations steps only, except at positions 215 and 151, as double nucleotide mutations are required for these residues to evolve to resistance. An extensive description of fitness landscapes and the mathematical model of evolution is available in Supplementary Methods S1 (available in JGV Online) and Deforche et al. (2007b, 2008b).

Validation experiments. Estimated fitness was compared with in vitro resistance fold change phenotype by using an independent, public set of matched genotype–phenotype pairs for subtype B sequences (Kantor et al., 2001). Sequences with ambiguities resulting in unknown amino acid mutations were removed, as well as sequences from virus isolates for which the phenotypic fold change was at the upper detection range of the assay. For each of the remaining sequences, a fitness value $f_j$ was estimated from resistance fold change $R_{AZT}$ and resistance fold change $R_{3TC}$ using $f_j = e_j [1 + (D/R_{AZT}) + (D/R_{3TC})]$, with $e_j$ representing the replication capacity of the virus and D the effective drug concentration (Holford & Sheiner, 1982). The values $e_j$ are generally unknown, and $e_j = 1$ was assumed for all strains when computing $f_j$. 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.

The performance of the model to predict observed evolution during treatment was evaluated on a dataset of longitudinal sequence pairs, sampled before, during or after the AZT + 3TC regimen. These sequences were not included in the cross-sectional training dataset. 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 assessed with a linear model, which included, apart from 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–Hochberg method with an FDR of 0.05. An observed mixture was evaluated by predicting evolution of the individual mutations of the mixture.

ACKNOWLEDGEMENTS

K. T. was funded by a PhD grant from the Institute for the Promotion of Innovation through Sciences and Technology in Flanders (IWT). This work was supported in part by the Virolab project (EU IST STREP Project 027446) and the Interuniversity Attraction Poles (IAP) P6/41, CHAIN 7FP, 223131, EURISYS 6FP (IST-2004-027173), FWO grant G.0611.09N, by the AIDS Reference Laboratory of Leuven which receives support from the Belgian Ministry of Social Affairs through a fund within the Health Insurance System. We wish to thank Robert Shafer and Soo-Yon Rhee who maintain the Stanford HIV Drug Resistance Database and researchers and clinicians that made data available. These include, amongst others, Éric Van Wijngaarden, Paul De Munter, Kamal Mansinho, Francisco Antunes, José Vera Dinis, Teresa Branco, Isabel Germano, Eugénio Teófilo, Lino Rosado, Maria João Águas, Paula Fonseca, Domitilia Faria, Telo Faria, António de Sousa, Luis Tavares, António Silva Graça, Cristina Guerreiro, Fausto Roxo, Graça Amaro, Isabel Mota, Margarida Bentes de Jesus, José Poças and Teresa Cabral.

REFERENCES


Deforche, K., Camacho, R. J., Grossman, Z., Soares, M. A., Van Laethem, K., Katzenstein, D. A., Harrigan, P. R., Kantor, R., Shafer, R. W.,
& other authors (2008a). Bayesian network analyses of resistance pathways against efavirenz and nevirapine. AIDS 22, 2107–2115.


