Computational methods for the design of effective therapies against drug resistant HIV strains

Niko Beerenwinkel1,*, Tobias Sing2, Thomas Lengauer2, Jörg Rahnenführer2, Kirsten Roopp2, Igor Savenkov2, Roman Fischer3, Daniel Hoffmann4, Joachim Selbig4, Klaus Korn5, Hauke Walter5, Thomas Berg6, Patrick Braun7, Gerd Fätkenheuer8, Mark Oette10, Jürgen Rockstroh11, Bernd Kupfer12, Rolf Kaiser9 and Martin Daemer9

1Department of Mathematics, University of California, Berkeley, CA, USA, 2Max Planck Institute for Informatics, Saarbrücken, Germany, 3Center of Advanced European Studies and Research, Bonn, Germany, 4Max Planck Institute of Molecular Plant Physiology and University of Potsdam, Germany, 5Institute of Clinical and Molecular Virology, University of Erlangen-Nürnberg, Erlangen, Germany, 6Medical Laboratory, Berlin, Germany, 7PZB, Aachen, Germany, 8Department of Internal Medicine and 9Institute of Virology, University of Cologne, Germany and 10Department of Gastroenterology, University of Düsseldorf, Germany and 11Department of Internal Medicine and 12Institute of Medical Microbiology and Immunology, University of Bonn, Germany

Received on June 8, 2005; revised on July 27, 2005; accepted on August 30, 2005
Advance Access publication September 6, 2005

ABSTRACT

Summary: The development of drug resistance is a major obstacle to successful treatment of HIV infection. The extraordinary replication dynamics of HIV facilitates its escape from selective pressure exerted by the human immune system and by combination drug therapy. We have developed several computational methods whose combined use can support the design of optimal antiretroviral therapies based on viral genomic data.

Contact: niko@math.berkeley.edu

1 INTRODUCTION

Persons infected with human immunodeficiency virus type 1 (HIV-1) are highly susceptible to develop the acquired immunodeficiency syndrome (AIDS), a major global threat to human health. HIV-1 is a retrovirus with a 9.2 kb genome coding for 15 viral proteins. Currently, 19 drugs targeting three distinct steps in the viral replication cycle are available for antiretroviral therapy. These drugs can be grouped into four different classes, according to their target and mechanism of action. Nucleoside and nucleotide analogs act as chain terminators in reverse transcription of RNA to DNA. Non-nucleoside reverse transcriptase inhibitors bind to and inhibit reverse transcriptase (RT), a viral enzyme that catalyzes reverse transcription. Protease inhibitors target the HIV protease, which is involved in maturation of released viral particles by cleaving precursor proteins. Finally, entry inhibitors block the penetration of HIV virions into their target cells.

Cell entry is a complex process mediated by sequential interactions of the viral proteins gp120 (envelope) and gp41 (transmembrane) with the cellular CD4 receptor and a co-receptor, usually CCR5 or CXCR4, depending on the individual virion. Consequently, different types of entry inhibitors have been proposed: fusion inhibitors prevent merging of viral and host cell membranes by binding to the transmembrane protein gp41. In contrast, co-receptor antagonists bind to the host protein prior to membrane fusion.

The available antiretroviral agents are applied in combination therapies—so-called highly active antiretroviral therapy (HAART), typically comprising two nucleoside analogs and either a protease inhibitor or a non-nucleoside RT inhibitor. However, therapeutic success, even of HAART, is limited. Antiretroviral therapy is not able to eradicate HIV, and durable suppression of virus replication below detectable limits is achieved in only a fraction of patients. Drug resistance can be the cause of treatment failure and is almost always a consequence of it (Clavel and Hance, 2004; DeGruttola et al., 2000).

1.1 Drug resistance

The intrapatient virus population is a highly dynamic system, characterized by high virus production and turnover rates and a high mutation rate. These evolutionary dynamics are the basis for a large and diversified virus population that predisposes or quickly generates resistance mutations. In a replicating population escape mutants with a selective advantage under therapy become dominant and lead to increased virus production and eventually to therapy failure. A number of mutations in protease, RT and gp41 have been associated with resistance to different antiviral agents (Shafer et al., 2000).
Each drug has its own characteristic resistance profile reflecting its chemical properties and mechanism of action. Nevertheless, cross-resistance (i.e., resistance against an unused drug) is common between drugs from the same class. Therefore, HAART advocates the use of two different drug classes in order to reduce the likelihood of a mutant to resist all drugs in the combination and to suppress viral replication more effectively (Jordan et al., 2002).

After treatment failure, the shifted population may be hit with a new drug combination, but finding such a potent regimen is challenging. Cross-resistance severely limits the remaining treatment options and the success of subsequent regimens is further impaired. The interplay between development of drug resistance and insufficient suppression of virus replication can eventually lead to situations in which the currently available drugs cannot at all control replication any longer. In the United States, as many as 50% of patients receiving HAART carry a virus that is resistant to at least one of the approved drugs (Richman et al., 2004). Furthermore, transmission of drug-resistant viruses is estimated to occur in ~15% of persons newly diagnosed with HIV infection in the United States (Bennett et al., 2005).

Since cross-resistance is frequent, treatment changes cannot be based on the assumption that the virus will remain susceptible to the unused drugs. Therefore, resistance testing has become an important diagnostic tool in the management of HIV-infections (Perrin and Telenti, 1998). Resistance testing can be performed either by measuring viral activity in the presence and absence of a drug (phenotypic resistance testing), or by sequencing the viral genes coding for the drug targets (genotypic resistance testing). Genotypic assays are much faster and cheaper, but sequence data provide only indirect evidence of resistance.

The Arevir project is a collaborative effort between clinicians, virologists and computational biologists to exploit genotype data from genotypic resistance tests for the individual selection of optimal drug combinations. We have developed several computational methods for the analysis of integrated genotypic, phenotypic and clinical data. Our goal was to provide tools for supporting personalized genotype-driven treatment decisions.

### 1.2 Challenges

The following questions have been addressed and approaches to their solutions will be described in the following sections.

1. **Data integration.** A prerequisite for any attempt to use genotypic data in a clinical setting is to provide this information at the right time and place. Resistance testing is often performed in specialized virological laboratories separated from the clinical department. Furthermore, most clinical data management systems are not prepared to handle sequence data. Thus, our first task is to collect, organize, and integrate all relevant patient data.

2. **Phenotype prediction from genotypes.** The first step in interpreting genotypic data is to understand the effect of single mutations and to relate mutational patterns to the *in vitro* phenotype. We have addressed predicting phenotypic drug resistance from the viral drug targets as well as prediction of co-receptor usage from gp120. Both models can augment the cheaper and faster genotypic test with a prediction of the phenotype, namely the susceptibility to each of the drugs and the co-receptor in use. This piece of information is important for the choice of therapy.

3. **Evolution of drug resistance.** Understanding the mutational pathways that lead to resistant strains is important for two reasons. First, this knowledge allows for estimating the distance of a virus population to escape from drug pressure, a quantity referred to as the genetic barrier. Second, the prediction of mutational pathways makes it possible to design sequences of therapies rather than one regimen at a time. We have addressed the problem of estimating evolutionary pathways from sequence data.

4. **Therapy optimization.** Our ultimate goal is to determine optimal drug combinations on the basis of genotypic information. For this task, we need to estimate the *in vivo* effect of a drug combination on a given viral genotype and to identify the regimen that maximizes clinical response. In addressing these problems we make use of both the *in vitro* phenotype predictions and the estimated evolutionary pathways.

For each of the four challenges we present computational approaches and indicate the biological or clinical impact. We show how the developed tools can be linked together in order to support the selection of effective therapies against drug-resistant HIV strains.

### 2 DATA MANAGEMENT

In order to meet the data integration and management challenge we have developed the Arevir database, a secure electronic platform for collaborative research aimed at optimizing anti-HIV therapies. This system is designed to facilitate data exchange, improve diagnostics, support medical decisions and provide the basis for data analysis.

#### 2.1 Database schema

In managing HIV-infected patients a number of different types of data arise, including personal patient data, therapy histories, numerous virologic, immunologic and other clinical test results derived from patient samples from different tissues, and sequence data, e.g. from genotypic resistance tests. Our database schema captures these data types in different modules, consisting of a few tables each (Beerenwinkel, 2004).

There is an important relationship between sequences and therapies via the drug targets. The compounds making up a combination therapy target specific viral proteins. DNA segments coding for these proteins are sequenced in order to gain information on the level of resistance that has been developed by the virus. Thus, given the values of clinical markers the data model allows for asking for outcomes of therapy types versus mutational patterns within the drug targets. This is the central question of the Arevir project. It will be revisited in a later section.

#### 2.2 Implementation

The data model has been implemented in the open source relational database management system MySQL. A secured client/server architecture allows for remote access to the centralized database. Since sensitive patient data are involved, this setting needs to meet the security demands imposed by state and national law. In addition, we have developed a web interface to the database for
clinicians and virologists. For these users the appropriate view on the data is through a single patient or a single patient sample. Thus, treating physicians as well as lab personnel get access to an integrated view onto all relevant data for one patient. For example, they can evaluate a genotypic resistance test result in the context of the patient’s medical history and current immunologic status. Moreover, applying the developed computational tools yields phenotypic interpretations of the genotypes. As of 2005, the Arevir database comprises 5720 patients, 9685 therapies, 5065 DNA sequences, and 146 539 laboratory test results from seven different institutions including three clinical centers and two virological laboratories.

2.3 Public databases
In addition to our pooled cohort data, public datasets can also provide valuable information on sequence variation and response to therapy. A major resource for clinical trials data is the AIDS Clinical Trials Group (http://aactg.org). The Los Alamos National Laboratories maintain databases of annotated HIV sequence data, drug-resistance mutations, HIV epitopes and vaccine trials results (http://www.hiv.lanl.gov). The Stanford HIV Drug Resistance Database contains sequences coding for the drug targets of antiretroviral therapy, drug susceptibility data and therapy histories where publicly available (http://hivdb.stanford.edu).

3 FROM GENOTYPE TO PHENOTYPE
Genotype–phenotype relations are much easier to study if the phenotype is determined by a well-defined lab experiment than for in vivo phenotypes that depend on many factors, which can confound the analysis. Therefore, predicting in vitro phenotypes from HIV genotypes is a good starting point for sequence interpretation.

3.1 Drug susceptibility
Prediction of phenotypic drug resistance from genotypes is based on matched genotype–phenotype pairs derived from patients failing antiretroviral therapy. For each drug, phenotypic resistance is determined in a recombinant virus assay (Kellam and Larder, 1999). In this experiment the replication capacity of the virus is measured as a function of drug concentration. The drug–response relationship is summarized by the resistance factor (Walter et al., 1999). However, determination of genotypes by cycle-sequencing is highly reproducible, but the common population sequencing strategy detects only those variants that are present in at least 20% of viruses in the population. For drug-resistance testing, the full protease (99 amino acids), the 5’ part of the RT (typically the first 250–300 residues), and possibly parts of gp120 are sequenced. To predict the resistance phenotype from the genotype means to solve, for each drug, the regression problem with predictors being the sequence positions of the drug target and response being the resistance factor. Alternatively, we may consider the related binary classification problem induced by choosing a drug-specific cutoff to define a susceptible and a resistant class of viruses.

A number of machine learning approaches to resistance phenotype prediction from genotypes have been proposed including neural networks (Drăghici and Potter, 2003; Wang and Larder, 2003), recursive partitioning (Sevin et al., 2000), linear stepwise regression (Wang et al., 2004), and more elaborate statistical models (Foulkes and DeGruttola, 2002; DiRienzo et al., 2003). We discuss in more detail support vector machine (SVM) regression (Beerenwinkel, 2001) and decision tree classification (Beerenwinkel, 2002), which serve as the engine for a widely used web-based prediction tool (cf. Section 5.2).

For SVM regression, sequences are mapped into an Euclidean vector space by introducing 20 indicator variables for each amino acid position of the multiple sequence alignment. The SVM learning strategy is suitable for this type of high-dimensional noisy data. Table 1 summarizes the performance of the regression models on a set of 650 genotype–phenotype pairs (Beerenwinkel et al., 2002).

<table>
<thead>
<tr>
<th>Drug</th>
<th>N</th>
<th>SV</th>
<th>MSE</th>
<th>Std error</th>
<th>r²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zidovudine</td>
<td>649</td>
<td>387</td>
<td>0.554</td>
<td>0.040</td>
<td>0.62</td>
</tr>
<tr>
<td>Didanosine</td>
<td>649</td>
<td>437</td>
<td>0.101</td>
<td>0.009</td>
<td>0.42</td>
</tr>
<tr>
<td>Zalcitabine</td>
<td>534</td>
<td>325</td>
<td>0.122</td>
<td>0.013</td>
<td>0.30</td>
</tr>
<tr>
<td>stavudine</td>
<td>649</td>
<td>401</td>
<td>0.145</td>
<td>0.015</td>
<td>0.33</td>
</tr>
<tr>
<td>Lamivudine</td>
<td>648</td>
<td>408</td>
<td>0.332</td>
<td>0.019</td>
<td>0.72</td>
</tr>
<tr>
<td>Abacavir</td>
<td>637</td>
<td>405</td>
<td>0.075</td>
<td>0.011</td>
<td>0.60</td>
</tr>
<tr>
<td>Tenofovir</td>
<td>321</td>
<td>206</td>
<td>0.091</td>
<td>0.005</td>
<td>0.50</td>
</tr>
<tr>
<td>Nevirapine</td>
<td>649</td>
<td>418</td>
<td>0.638</td>
<td>0.056</td>
<td>0.55</td>
</tr>
<tr>
<td>Delavirdine</td>
<td>648</td>
<td>403</td>
<td>0.476</td>
<td>0.033</td>
<td>0.55</td>
</tr>
<tr>
<td>Efavirenz</td>
<td>634</td>
<td>437</td>
<td>0.354</td>
<td>0.026</td>
<td>0.60</td>
</tr>
<tr>
<td>Saquinavir</td>
<td>652</td>
<td>394</td>
<td>0.204</td>
<td>0.022</td>
<td>0.71</td>
</tr>
<tr>
<td>Indinavir</td>
<td>652</td>
<td>387</td>
<td>0.197</td>
<td>0.017</td>
<td>0.73</td>
</tr>
<tr>
<td>Ritonavir</td>
<td>652</td>
<td>383</td>
<td>0.176</td>
<td>0.017</td>
<td>0.79</td>
</tr>
<tr>
<td>Nelfinavir</td>
<td>651</td>
<td>391</td>
<td>0.207</td>
<td>0.011</td>
<td>0.71</td>
</tr>
<tr>
<td>Amprenavir</td>
<td>464</td>
<td>303</td>
<td>0.173</td>
<td>0.013</td>
<td>0.65</td>
</tr>
<tr>
<td>Lopinavir</td>
<td>307</td>
<td>210</td>
<td>0.169</td>
<td>0.016</td>
<td>0.73</td>
</tr>
<tr>
<td>Atazanavir</td>
<td>305</td>
<td>187</td>
<td>0.262</td>
<td>0.034</td>
<td>0.61</td>
</tr>
</tbody>
</table>

Predictive performance was estimated from N samples by 10-fold cross-validation, and is reported as the mean squared error (MSE), its standard error (std error) and the squared correlation coefficient (r²) between predicted and observed log_{10}-resistance factors. SV denotes the number of support vectors.

---

Table 1. SVM regression models

A number of machine learning approaches to resistance phenotype prediction from genotypes have been proposed including neural networks (Drăghici and Potter, 2003; Wang and Larder, 2003), recursive partitioning (Sevin et al., 2000), linear stepwise regression (Wang et al., 2004), and more elaborate statistical models (Foulkes and DeGruttola, 2002; DiRienzo et al., 2003). We discuss in more detail support vector machine (SVM) regression (Beerenwinkel, 2001) and decision tree classification (Beerenwinkel, 2002), which serve as the engine for a widely used web-based prediction tool (cf. Section 5.2).

For SVM regression, sequences are mapped into an Euclidean vector space by introducing 20 indicator variables for each amino acid position of the multiple sequence alignment. The SVM learning strategy is suitable for this type of high-dimensional noisy data. Table 1 summarizes the performance of the regression models on a set of 650 genotype–phenotype pairs (Beerenwinkel et al., 2002).

SVMs are among the best performing machine-learning methods in terms of prediction accuracy. However, other methods are advantageous if interpretation of the learned model is intended. We have applied decision trees to the classification problem described above in order to elucidate the effect of mutational patterns on the resistance phenotype. This analysis has revealed concise models incorporating only 4–7 sequence positions as compared with some 10–20 positions that are associated with resistance (Johnson et al., 2004). Moreover, decision trees can model the effect of a mutation in the context of other mutations. In particular, some decision trees display resensitization or hypersusceptibility effects. For example, zidovudine resistance induced by mutation T215Y in the RT may be reverted by mutations L74V/I and M184I/V. The latter substitution can also resensitize tenofovir resistant strains (Wolf et al., 2003). Likewise, mutation N85S in the protease gene has been found to increase susceptibility to amprenavir.
3.2 Co-receptor usage

The effective use of co-receptor antagonists that target a particular co-receptor depends on the ability of determining prior to drug application the type of co-receptor used by the virus for cell entry. In fact, careful monitoring of viral co-receptor usage is mandatory during such treatment, because few mutations in the envelope protein gp120 of HIV are sufficient for switching to another co-receptor. In addition, a switch from CCR5 to CXCR4 has been associated with accelerated progression towards AIDS. Since experimental determination of co-receptor usage is costly, the availability of sequence-based methods would be advantageous for routine clinical practice with upcoming CCR5 antagonists.

We have analyzed this genotype–phenotype relation in 1100 sequences of the third hypervariable (V3) region of gp120 for which co-receptor usage had been determined experimentally. To accommodate for the extraordinary genetic variability within this region, sequences were aligned to a fixed reference multiple alignment containing representatives of all HIV-1 subtypes. We compared decision trees, SVMs (Pillai et al., 2003), neural networks (Resch et al., 2001), position-specific scoring matrices (Jensen et al., 2003) and a classical rule based on charge of amino acids at positions 11 and 25 in the V3 loop (Fouchier et al., 1995). Using ROCR (Sing et al., 2005b), a comprehensive tool for evaluating classifier performance, we found SVMs to outperform the other methods. In an effort to attain this current gold standard in performance with a model that lends itself more readily for interpretation, we have suggested mixtures of localized rules (Sing et al., 2004), a novel weighted voting strategy for rules-based classifiers. Rules, describing specific mutational patterns, are localized in the sense that their associated weights are modulated in an instance-dependent manner based on the genetic background in which the pattern occurs. This method significantly outperformed classical decision tree building, thus representing an alternative for knowledge extraction.

4 EVOLUTIONARY PATHWAYS

Under suboptimal therapy the virus population continuously replicates and acquires new resistance mutations. This process occurs in a non-uniform, stochastic fashion and gives rise to co-existing evolutionary pathways. Understanding this evolutionary process is important for estimating the proximity of a virus to escape from drug pressure. We use mutagenetic trees, a family of probabilistic graphical models, to estimate rate and order of occurrence of resistance-associated mutations in the viral drug targets.

4.1 Mutagenetic trees

We consider a set of $n$ specific amino acid changes (mutations) that develop under drug treatment. A mutagenetic tree for these $n$ mutations is a connected branching on $\{0, \ldots, n\}$ rooted at 0 (Fig. 1). Each vertex $v \neq 0$ represents the binary random variable $X_v$ that indicates the occurrence of mutation $v$. We associate probability parameters $\theta_v$ with the tree edges to obtain a directed acyclic graph model with conditional probability matrices

\[
\Pr(X_v = b | X_{\text{pa}(v)} = a)_{a,b=0,1} = \begin{pmatrix} 1 & 0 \\ 1 - \theta_v & \theta_v \end{pmatrix},
\]

where $\text{pa}(v)$ denotes the parent of $v$ in the tree. The first row of this matrix imposes the constraint that a mutation can occur only if all of its ancestor mutations have already occurred. A mutagenetic tree defines a probability distribution on the set of all possible mutational patterns. In particular, this model family includes linear path models (chains) and the model of complete independence given by the star topology. It is possible to characterize the complete family of mutagenetic tree models by their algebraic invariants, which turn out to have a simple combinatorial structure (Beerenwinkel et al., 2005). Mutagenetic trees can be reconstructed from observed cross-sectional data by Edmonds' maximum weight branching algorithm involving only pair-wise probabilities (Desper et al., 1999).

We have extended the single tree model to mixture models of mutagenetic trees that combine several weighted trees (Beerenwinkel et al., 2005c). The first tree component is a star with uniform probabilities that models the spontaneous and independent occurrence of mutations. All other components represent dependencies between mutations and are estimated from the data. The mixture model is learned by an Expectation–Maximization Algorithm that iteratively estimates the expected values of the missing data (i.e. the association of samples to the trees) and the structure and parameters of the trees. For model selection (choosing the number of tree components) we either use cross-validation or a modified Bayesian Information Criterion that includes an estimate of the structural redundancy between tree components (J. Yin, N. Beerenwinkel, J. Rahnenfuhrer and T. Lengauer submitted for publication). Mrvtreemtix, a software package for statistical inference with mutagenetic trees and mixtures of these, is described in Beerenwinkel et al. (2005a).

Assuming independent Poisson processes for the occurrence of mutations and for the observed sampling times (i.e. the time on therapy) with rates $\lambda_\lambda$ and $\lambda_\lambda$, respectively, we find

\[
\theta_v = \frac{\lambda_v}{\lambda_\lambda + \lambda_\lambda}.
\]

This relation allows for translating the estimated conditional probabilities between mutations into the expected waiting time for the mutation to occur. Furthermore, the probabilities of occurrence of any mutational pattern can be computed for any fixed mean waiting time. Hence, using these timed mutagenetic trees we can compare models that have initially been estimated from datasets sampled after different mean waiting times.

Figure 1 shows a mutagenetic tree and the corresponding timed mutagenetic tree for the development of drug resistance in the HIV RT under therapy with zidovudine, the first anti-HIV drug approved.
Table 2. Genetic barriers of the wild-type virus to resistance to zidovudine (ZDV) under the three regimens zidovudine monotherapy, zidovudine + lamivudine (3TC) double therapy, and zidovudine + didanosine (ddI) double therapy.

<table>
<thead>
<tr>
<th>Therapy</th>
<th>Genetic barrier to ZDV resistance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10-fold</td>
</tr>
<tr>
<td>ZDV</td>
<td>0.51</td>
</tr>
<tr>
<td>ZDV + 3TC</td>
<td>0.66</td>
</tr>
<tr>
<td>ZDV + ddI</td>
<td>0.30</td>
</tr>
</tbody>
</table>

The genetic barrier was defined as the probability of the virus population not reaching any escape mutant, after 96 weeks of therapy, with a fold change in susceptibility to ZDV of 10 (low level resistance) and 100 (high level resistance).

The tree has been estimated from 364 genotypes derived from previously untreated patients under zidovudine mono-therapy (Beerenwinkel et al., 2005b). This dataset is publicly available at the Stanford HIV Drug Resistance Database (Rhee et al., 2003). The model displays two characteristic pathways, namely the 70-219 and the 215-41 pathways (cf. Boucher et al., 1992).

4.2 Genetic barrier

Suppose we have estimated a mutagenetic tree model for the development of resistance to a certain drug. In particular, this model can be used to compute transition probabilities between mutational patterns. As described in the previous section we can predict the resistance phenotype from the genotype. Using a classifier restricted to the set of n mutations we predict each mutational pattern to be either susceptible or resistant. Now, for a given virus we may ask what the transition probability to any resistant state is. In fact, this question is crucial for minimizing the risk of resistance development with the next regimen. We refer to the genetic barrier as the probability of not reaching any resistant state after a fixed time period under therapy. This quantity can be calculated as the sum of the probabilities of all mutational patterns predicted as susceptible. Thus, a higher genetic barrier indicates that the virus is less likely to become resistant.

For example, Table 2 shows the genetic barriers to both low level and high level zidovudine resistance of the wild type virus under three different regimens, namely zidovudine monotherapy, double therapy with zidovudine plus lamivudine, and double therapy with zidovudine plus didanosine. The underlying mutagenetic tree model is the tree displayed in Figure 1 scaled to a mean sampling time of 96 weeks. As expected, the genetic barrier to zidovudine is always higher under the combination of zidovudine plus lamivudine than under zidovudine alone, because these drugs do not share any resistance mutations. More surprisingly, we find that zidovudine resistance appears to develop faster under zidovudine plus didanosine than under zidovudine monotherapy. This effect may be explained by the stronger selective pressure exerted by the double therapy and the cross-resistance profile of zidovudine and didanosine (Beerenwinkel et al., 2005b; Brun-Vezinet et al., 1997). Thus, the genetic barrier is a useful concept for designing effective treatment strategies.

5 THERAPY OPTIMIZATION

The computational task of identifying optimal antiretroviral drug combinations with respect to a given viral genotype is a typical bioinformatics problem (such as sequence alignment) in the sense that the objective function of the optimization problem is not known. In fact, we need to know the in vitro effect of any drug combination on any mutational pattern in order to find the best regimen. Typical clinical parameters of interest are the virus load (the amount of plasma HIV RNA) and the number of CD4+ cells (T-lymphocytes). Estimating these response functions is much more challenging than actually selecting the optimal drug therapy. Indeed, the number of drug combinations is only on the order of thousands, and hence they can be enumerated. By contrast, HIV’s high genetic diversity induces a much higher number of mutational patterns over all drug targets. Furthermore, clinical response is influenced by several factors other than resistance, including patient adherence, immunological status and baseline virus load.

One way to estimate the activity of a therapeutic regimen against a viral strain is to learn this effect from an observational clinical database such as the Areviv database. This is straightforward, if we fix a combination therapy or a narrowly defined type of therapy. In this case, machine-learning approaches similar to those presented in a previous section can be used to predict clinical response. However, if the drug combination is not fixed, direct learning from cohort data is limited by the amount of data necessary to derive useful models, because now the complexity of the problem depends on both mutational patterns and drug combinations (DiRienzo and DeGruttola, 2002). Furthermore, the distribution of drug combinations in clinical databases is heavily skewed, reflecting approval times and treatment strategies over time (Beerenwinkel, 2004). Thus, training on such datasets is likely to result in models that capture the features of only a few frequently observed combinations, but are not appropriate to explore the product space of all mutational patterns and drug combinations.

5.1 Scoring functions

An alternative approach to general response prediction is to score drug combinations on the basis of single drug effects. This implies assuming a functional dependency of the effect of a drug combination on the single drug effects. The simplest way of doing this is to use a classifier for resistance phenotype prediction and to count the number of active drugs in a combination, i.e. the number of drugs for which the virus is predicted susceptible. For example, De Luca et al. (2003) have separated 332 patients according to viral genotype and therapy. SVM based phenotype predictions were used to define one group of patients with two or fewer drugs predicted as active and another group with three or more active drugs. Using a Cox proportional hazards model they show that patients in the group with at most two active drugs are at significantly higher risk of virological failure (Fig. 2). As compared with 11 other interpretation systems that are based on expert rules, only this data-driven approach yielded significant predictions of virological response.

A natural refinement of this scoring scheme is to sum over the real-valued predicted resistance factors instead of the binary resistance predictions. However, the dynamic range of resistance factors varies by as much as two orders of magnitude between different drugs. In order to normalize these values we estimate their distribution over a large random sample of 2000 genotypes. Since bimodality is a common feature for all drugs, we model this density by a Gaussian mixture model,

$$\lambda \times N(x; \mu_1, \sigma_1) + (1 - \lambda) \times N(x; \mu_2, \sigma_2), \quad \lambda \in [0, 1],$$

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/21/21/3943/226842 by guest on 13 March 2019
whose parameters can be estimated by the Expectation–Maximization Algorithm. This two-state model provides a data-derived definition of susceptible and resistant. By linearizing the log-likelihood ratio between these two classes, we obtain the activity score, which approximates the conditional probability of membership in the susceptible class given the viral genotype (Beerenwinkel et al., 2003a). Thus, the activity score provides a normalized and comparable measure of resistance, and we can extend it to multi-drug therapies by summing over all drugs in the combination.

Similarly, we can use the genetic barrier of the virus to resistance to each of the compounds of the regimen (Fig. 3). Summing these values provides an estimate of how easy it is for the virus to escape from the selective pressure of the combination therapy. As demonstrated in Section 4.2 this genetic barrier score can be different from the genetic barrier of the drug combination. We confine ourselves with this approximation, because estimating the genetic barrier for all drug combinations would again require, for each combination, many samples derived from patients under the respective regimen. Despite these simplifications both the activity score and the genetic barrier score are predictive of virological response. Figure 4 shows their performance of classifying genotype–therapy pairs on a special and instructive dataset consisting of 64 sequences, each paired with one successful and one failing regimen. The genotype alone does not provide any useful information for classifying these pairs. Similarly, by randomizing the genotype data, we see that the therapy data alone do not give rise to a competitive classifier either. The noticeably best performance is obtained on the combined genotype–therapy data. Thus, the learned concept is specific for the combined effect of drug combination and mutational pattern. The genetic barrier score, which makes use of three different types of datasets (Fig. 3), performs best.

In a related approach we have estimated the proximity of the virus to an escape state more conservatively. Applying a heuristic greedy search, we explore the mutational neighborhood of the viral sequence by successively introducing point mutations and following the in silico mutants that reduce the activity of the regimen most.

The estimated ‘worst case’ activities were used in a regression model to predict the expected drop in virus load (Beerenwinkel et al., 2003b).

5.2 Geno2pheno

We have implemented the web server geno2pheno (http://www.genafor.org) that provides interpretations of genotypic test results in terms of phenotype predictions (Beerenwinkel et al., 2003a; Sing et al., 2005a). The system predicts co-receptor usage from submitted HIV-1 V3 loop sequences as well as phenotypic resistance to 17 antiretroviral agents from protease and RT sequences. The output also includes activity scores rendering predictions comparable between drugs. An additional software tool, theo, for selecting and evaluating drug combinations on the basis of the different scoring functions discussed above is currently validated and tested by virologists and clinicians. Since December 2000, geno2pheno has made 35 000 online resistance predictions and since June 2004...
>1000 co-receptor predictions. The system is used worldwide by virologists performing genotypic resistance tests as well as by clinicians seeking effective drug combinations.

6 CONCLUSIONS

In order to support clinical decision making on the basis of viral genomic data, we have developed and applied several computational methods and tools. Specifically, we have addressed data integration and management (Arevir database), prediction of drug resistance and co-receptor usage from genotypes (geno2pheno), modeling of the evolution of drug resistance and the genetic barrier by mutagenetic trees (mtreemix), and selection of optimal drug combinations (theo). The integration of various types of genomic, phenotypic and clinical data as well as the coupling of different computational models yields predictive models of therapy outcome that may support the design of combination therapies.

6.1 Future work

Further factors of therapeutic outcome, involving pharmacological, viral and host factors, need to be accounted for in future work. For example, pharmacokinetic properties of drugs and their specific realization in different patients (pharmacogenomics) are important predictors. The amount of drug actually present in infected cells may yield more accurate predictions of the development of resistance. Besides resistance, replication capacity (fitness) is another viral property currently investigated. It depends on phenotypic properties of many viral proteins, such as protease cleavage rate or RT error rate. It may be expected that a fitness estimate based on integrating these predictions into a model of the viral replication cycle will lead to improved predictions. Finally, there is strong evidence that viral evolution is, in part, also host-dependent. In particular, we have started to study the impact of the host HLA genotype on the development of viral escape mutations (Rooimp et al., 2005).

ACKNOWLEDGEMENTS

The Arevir project, including Open Access publication charges for this article, has been funded by Deutsche Forschungsgemeinschaft (DFG) under Grant No. HO 1582/1-3 and KA 1569/1-3. N.B. acknowledges funding from DFG under Grant No. BE 3217/1-1. J.R. has been funded by BMBF under Grant No. 01GR0453. The work at the Max-Planck Institute of Computer Science has been performed partly in the context of the BioSapiens Network of Excellence (EU Grant No. LSHG-CT-2003-503265).

Conflict of Interest: none declared.

REFERENCES

De Luca,A. et al. (2003) The prognostic value to predict virological outcomes of 14 distinct systems used to interpret the results of genotypic HIV-1 drug resistance testing in untreated patients starting their first HAART. HIV Med., 4, 20.

HIV drug resistance

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/21/21/3943/226842 by guest on 13 March 2019


