Bioinformatics Advance Access published September 6, 2005

BIOINFORMATICS

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

Niko Beerenwinkel[1,*], Tobias Sing[2], Thomas Lengauer[2], Jörg Rahnenführer[2], Kirsten Roomp[2], Igor Savenkov[2], Roman Fischer[3], Daniel Hoffmann[3], Joachim Selbig[4], Klaus Korn[5], Hauke Walter[5], Thomas Berg[6], Patrick Braun[7], Gerd Fätkenheuer[8], Mark Oette[9], Jürgen Rockstroh[10], Bernd Kupfer[11], Rolf Kaiser[12], Martin Däumer[12]

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

Received on …; revised on …; accepted on …

Advance Access publication . . .

ABSTRACT

Motivation: 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.

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*kbp* 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 analogues 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 coreceptor, 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, coreceptor 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 analogues 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 et al., 2004, DeGruttola et al., 2000).

1.1 Drug resistance

The intra-patient 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., 2004).

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.

*To whom correspondence should be addressed.

The interplay between development of drug resistance and insufficient suppression of virus replication can eventually lead to situations in which the currently available drugs can no longer control replication at all. 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 about 15% of persons newly diagnosed with HIV infection in the US (Bennett et al., 2005).

Because 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 et al., 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 labs 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 coreceptor 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 coreceptor in use, respectively. 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 to 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, 2004a).

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 5,720 patients, 9,685 therapies, 5,065 DNA sequences, and 146,539 laboratory test results from 7 different institutions including 3 clinical centers and 2 virological labs.

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 & Larder, 1994; Walter et al., 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* (or the *fold-change in susceptibility*), defined as the ratio between the amount of drug necessary to inhibit replication of the virus by 50% and the corresponding value for a standardized wild type virus. Coefficients of variation between 10% and 60% have been reported for the resistance factor (Walter et al., 1999). On the other hand, 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 gp41 and 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 et al., 2003; Wang et al., 2003), recursive partitioning (Sevin et al., 2000), linear stepwise regression (Wang et al., 2004), and more elaborate statistical models (Foulkes et al., 2002; DiRienzo et al., 2003). We discuss in more detail support vector machine (SVM) regression (Beerenwinkel, 2001a) and decision tree classification (Beerenwinkel, 2002a), 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., 2002a).

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 to 7 sequence positions as compared to some 10 to 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 N88S in the protease gene has been found to increase susceptibility to amprenavir.

3.2 Coreceptor usage

The effective use of coreceptor antagonists that target a particular coreceptor depends on the ability of determining prior to drug application the type of coreceptor used by the virus for cell entry. In fact, careful monitoring of viral coreceptor usage is mandatory during such treatment, because few mutations in the envelope protein gp120 of HIV are sufficient for switching to another coreceptor. In addition, a switch from CCR5 to CXCR4 has been associated with accelerated progression towards AIDS. Since experimental determination of coreceptor 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 coreceptor 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. 2). Each vertex $v \neq 0$ represents the binary random variable X_v that indicates the occurrence of mutation v. We associate probability parameters θ_v with the tree edges to obtain a directed acyclic graphical model with conditional probability matrices

$$\left(\Pr(X_v = b \mid X_{pa(v)} = a)\right)_{a,b=0,1} = \begin{pmatrix} 1 & 0 \\ 1-\theta_v & \theta_v \end{pmatrix},$$

where $pa(v)$ 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., 2005c). Mutagenetic trees can be reconstructed from observed cross-sectional data by Edmonds' maximum weight branching algorithm involving only pair-wise probabilities (Desper at al., 1999).

We have extended the single tree model to mixture models of mutagenetic trees that combine several weighted trees (Beerenwinkel et al., 2005d). 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 (Yin et al., 2005). *Mtreemix*, 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 λ_v and λ_S, respectively, we find

$$\theta_v = \frac{\lambda_v}{\lambda_v + \lambda_S}.$$

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 data sets 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. 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 pathway (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 mono-therapy, 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 mono-therapy. 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, for example) in the sense that the objective function of the optimization problem is not known. In fact, we need to know the *in vivo* 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 Arevir 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 et al., 2002). Furthermore, the distribution of drug combinations in clinical databases is heavily skewed, reflecting approval times and treatment strategies over time (Beerenwinkel et al., 2004a). Thus, training on such data sets 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 2 or fewer drugs predicted as active and another group with 3 or more active drugs. Using a Cox proportional hazards model they show that patients in the group with at most 2 active drugs are at significantly higher risk of virological failure (Figure 2). As compared to 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,

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

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 (Figure 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 data sets (Figure 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 those *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 coreceptor 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 more than 1,000 coreceptor predictions. The system is used world-wide 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 coreceptor 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 (Roomp et al., 2005).

ACKNOWLEDGEMENTS

The *Arevir* project 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)

REFERENCES

Beerenwinkel,N., Schmidt,B., Walter,H., Kaiser,R., Lengauer,T., Hoffmann,D., Korn,K. and Selbig,J. (2001a) Geno2pheno: interpreting genotypic HIV drug resistance tests. *IEEE Intelligent Systems*, **16**, 35–41.

Beerenwinkel,N., Schmidt,B., Walter,H., Kaiser,R., Lengauer,T., Hoffmann,D., Korn,K. and Selbig,J. (2002) Diversity and complexity of HIV-1 drug resistance: a bioinformatics approach to predicting phenotype from genotype. *Proc. Natl. Acad. Sci. U. S. A.*, **99**, 8271–8276.

Beerenwinkel,N., Däumer,M., Oette,M., Korn,K., Hoffmann,D., Kaiser,R., Lengauer,T., Selbig,J. and Walter,H. (2003a) Geno2pheno: Estimating phenotypic drug resistance from HIV-1 genotypes. *Nucl. Acids Res.*, **31**, 3850–3855.

Beerenwinkel,N., Lengauer,T., Däumer,M., Kaiser,R., Walter,H., Korn,K., Hoffmann,D. and Selbig,J. (2003b) Methods for optimizing antiviral combination therapies. *Bioinformatics,* **19**, i16–i25.

Beerenwinkel,N. (2004a) *Computational Analysis of HIV Drug Resistance Data.* Shaker, Aachen, Germany.

Beerenwinkel,N., Rahnenführer,J., Kaiser,R., Hoffmann,D., Selbig,J. and Lengauer,T. (2005a) Mtreemix: a software package for learning and using mixture models of mutagenetic trees. *Bioinformatics,* **21**, 2106-2107.

Beerenwinkel,N., Däumer,M., Sing,T., Rahnenführer,J., Lengauer,T., Selbig,J., Hoffmann,D. and Kaiser,R. (2005b) Estimating HIV evolutionary pathways and the genetic barrier to drug resistance. *J. Infect. Dis.*, **191**, 1953-1960.

Beerenwinkel,N. and Drton,M. (2005c) Mutagenetic tree models. In L. Pachter and B. Sturmfels, editors, *Algebraic Statistics for Computational Biology*, chapter 14. Oxford University Press, Oxford, UK, pp. 278-290, *in press*.

Beerenwinkel,N., Rahnenführer,J., Däumer,M., Hoffmann,D., Kaiser,R., Selbig,J. and Lengauer,T. (2005d) Learning multiple evolutionary pathways from cross-sectional data. *J. Comput. Biol.,* **12**, *in press*.

Bennett,D., McCormick,L., Kline,R., Wheeler,W., Hemmen,M., Smith,A., Zaidi,I., Dondero,T. and The HIV Drug Resistance ARVDRT/VARHS Surveillance Group (2005) U.S. surveillance of HIV drug resistance at diagnosis using HIV diagnostic sera [abstract 674]. *12th Conference on Retroviruses and Opportunistic Infections*.

Boucher,C., O'Sullivan,E., Mulder,J. *et al.* (1992) Ordered appearance of zidovudine resistance mutations during treatment of 18 human immunodeficiency virus-positive subjects. *J. Infect. Dis.*, **165**, 105-110.

Brun-Vezinet,F., Boucher,C., Loveday,C., Descamps,D., Fauveau,V., Izopet,J., Jeffries,D., Kaye,S. et al. (1997) HIV-1 viral load, phenotype, and resistance in a subset of drug-naïve participants from the Delta trial. *Lancet*, **350**, 983-90.

Clavel,F. and Hance,A.J. (2004) HIV drug resistance. *N. Engl. J. Med.*, **350**, 1023-35.

DeGruttola,V., Dix,L., D'Aquila,R., Holder,D., Phillips,A., Ait-Khaled,M., Baxter,J., Clevenbergh,P., Hammer,S. and Harrigan,R., *et al.* (2000) The relation between baseline HIV drug resistance and response to antiretroviral therapy: re-analysis of retrospective and prospective studies using a standardized data analysis plan. *Antivir. Ther.*, **5**, 41–48.

De Luca,A., Cozzi-Lepri,A., Perno,C., Balotta,C., Di Giambenedetto,S., Orani,A., Mussini,C., Toti,M. and d'Arminio Monforte,A. (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.

Desper.R., Jiang,F. Kallioniemi,O.-P., Moch,H. Papadimitriou,C. and Schäffer,A. (1999) Inferring tree models for oncogenesis from comparative genome hybridization data. *J. Comput. Biol.*, **6**, 37-51.

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.

DiRienzo,A.G., DeGruttola,V., Larder,B. and Hertogs,K. (2003) Nonparametric methods to predict HIV drug susceptibility phenotype from genotype. *Stat. Med.,* **22**, 2785-2798.

Drăghici,S. and Potter,R. (2003) Predicting HIV drug resistance with neural networks. *Bioinformatics,* **19**, 98-107.

Fouchier,R.A., Brouwer,M., Broersen,S.M. and Schuitemaker,H. (1995) Simple determination of human immunodeficiency virus type 1 syncytium-inducing V3 genotype by PCR. J. Clin. Microbiol., 33, 906-911.

Foulkes,A.S. and DeGruttola,V. (2002) Characterizing the relationship between HIV-1 genotype and phenotype: prediction based classification. *Biometrics*, **58**, 145-156.

Jensen,M.A., Li,F.S., van 't Wout,A.B., Nickle,D.C., Shriner,D., He,H.X., McLaughlin,S., Shankarappa,R. et al. (2003) Improved coreceptor usage prediction and genotypic monitoring of R5-to-X4 transition by motif analysis of human immunodeficiency virus type 1 env V3 loop sequences. *J. Virol.*, **77**, 13376-13388.

Johnson,V.A., Brun-Vézinet,F., Clotet,B., Conway, B., D'Aquila,R.T., Demeter,L.M., Kuritzkes,D.R., Pillay,D. et al. (2004) Update of the Drug Resistance Mutations in HIV-1: 2004. *Topics in HIV Medicine*, **12**, 2004.

Jordan,R., Gold,L., Cummins,C. and Hyde,C. (2002) Systematic review and meta-analysis of evidence for increasing numbers of drugs in antiretroviral combination therapy. *British Med. J.*, **324**, 1-10.

Kellam,P. and Larder,B. (1994) Recombinant virus assay: a rapid, phenotypic assay for assessment of drug susceptibility of human immunodeficiency virus type 1 isolates. *Antimicrob. Agents and Chemother.,* **38**, 23-30.

Perrin,L. and Telenti,A. (1998) HIV treatment failure: testing for HIV resistance in clinical practice. *Science*, **280**, 1871–1873.

Pillai,S., Good,B., Richman,D. and Corbeil,J. (2003) A new perspective on V3 phenotype prediction. *AIDS Res. Hum. Retroviruses*, **19**, 145-149.

Resch,W., Hoffman,N. and Swanstrom,R. (2001) Improved success of phenotype prediction of the Human Immunodeficiency Virus type 1 from envelope variable loop 3 sequence using neural networks. *Virology*, **288**, 51-62.

Rhee,S.-Y., Gonzales,M., Kantor,R., Betts,B., Ravela,J. and Shafer,R. (2003) Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucl. Acids Res.*, **31**, 298-303.

Richman,D.D., Morton,S.C., Wrin,T., Hellmann,N., Berry,S., Shapiro,M.F. and Bozzette,S.A. (2004) The prevalence of antiretroviral drug resistance in the United States. *AIDS,* **18**, 1393-1401.

Roomp,K., Ahlenstiel,G., Beerenwinkel,N., Rockstroh,J., Däumer,M., Spengler,U. and Lengauer,T. (2005) HLA profiles predict known and novel HIV-1 escape mutations at a population level. *2nd Int. Immunoinformatics Symposium*.

Sevin,A.D., DeGruttola,V., Nijhuis,M., Schapiro,J.M,, Foulkes,A.S., Para,M.F. and Boucher,C.A. (2000) Methods for investigation of the relationship between drug-susceptibility phenotype and human immunodeficiency virus type 1 genotype with applications to AIDS clinical trials group 333. *J. Infect. Dis.*, **182**, 59-67.

Shafer,R.W., Kantor,R. and Gonzales,M.J. (2000) The genetic basis of HIV-1 resistance to reverse transcriptase and protease inhibitors. *AIDS Reviews*, **2**, 211–228.

Sing,T., Beerenwinkel,N. and Lengauer,T. (2004) Learning mixtures of localized rules by maximizing the area under the ROC curve. In *Proc. 16th European Conference on Artificial Intelligence, Workshop on ROC Analysis in AI*.

Sing, T., Beerenwinkel, N., Kaiser, R., Hoffmann, D., Däumer,M., and Lengauer, T. (2005a) Geno2pheno[coreceptor]: a tool for predicting coreceptor usage from genotype and for monitoring coreceptor-associated sequence alterations. *3rd European HIV Drug Resistance Workshop*.

Sing,T., Sander,O., Beerenwinkel,N. and Lengauer,T. (2005b) ROCR: Visualizing classifier performance, *in press*.

Wang,D. and Larder,B. (2003) Enhanced prediction of lopinavir resistance from genotype by use of artificial neural networks. *J. Infect. Dis.,* **188**, 653-660.

Wang,K., Jenwitheesuk,E., Samudrala,R. and Mittler,J.E. (2004) Simple linear model provides highly accurate genotypic predictions of HIV-1 drug resistance. *Antivir. Ther.*, **9**, 343-352.

Walter,H., Schmidt,B., Korn,K., Vandamme,A.M., Harrer,T. and Überla,K. (1999) Rapid, phenotypic HIV-1 drug sensitivity assay for protease and reverse transcriptase inhibitors. *J. Clin. Virol.*, **13**, 71–80.

Wolf,K., Walter,H., Beerenwinkel,N., Keulen,W., Kaiser,R., Hoffmann,D., Lengauer,T., Selbig,J., Vandamme,A.-M., Korn,K. and Schmidt,B. (2003) Tenofovir resistance and resensitization. *Antimicrob. Agents and Chemother.*, **47**, 3478–3484.

Yin,J., Beerenwinkel,N., Rahnenführer,J. and Lengauer,T. (2005) Model selection for mixtures of mutagenetic trees, *submitted*.

TABLES

Table 1. SVM regression models.

Drug	N	SV	MSE	Std error	r^2
Zidovudine	649	387	0.554	0.040	0.62
Didanosine	649	437	0.101	0.009	0.42
Zalcitabine	534	325	0.122	0.013	0.30
Stavudine	649	401	0.145	0.015	0.33
Lamivudine	648	408	0.332	0.019	0.72
Abacavir	637	405	0.075	0.011	0.60
Tenofovir	321	206	0.091	0.005	0.50
Nevirapine	649	418	0.638	0.056	0.55
Delavirdine	648	403	0.476	0.033	0.55
Efavirenz	634	437	0.354	0.026	0.60
Saquinavir	652	394	0.204	0.022	0.71
Indinavir	652	387	0.197	0.017	0.73
Ritonavir	652	383	0.176	0.017	0.79
Nelfinavir	651	391	0.207	0.011	0.71
Amprenavir	464	303	0.173	0.013	0.65
Lopinavir	307	210	0.169	0.016	0.73
Atazanavir	305	187	0.262	0.034	0.61

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 (r^2) between predicted and observed $\log_{10}$-resistance factors. SV denotes the number of support vectors

Table 2. Genetic barriers of the wild type virus to resistance to zidovudine (ZDV) under the three regimens zidovudine mono therapy, zidovudine + lamivudine (3TC) double therapy, and zidovudine + didanosine (ddI) double therapy.

	Genetic barrier to ZDV resistance	
Therapy	10-fold	100-fold
ZDV	0.51	0.80
ZDV + 3TC	0.66	0.93
ZDV + ddI	0.30	0.72

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), respectively.

FIGURES

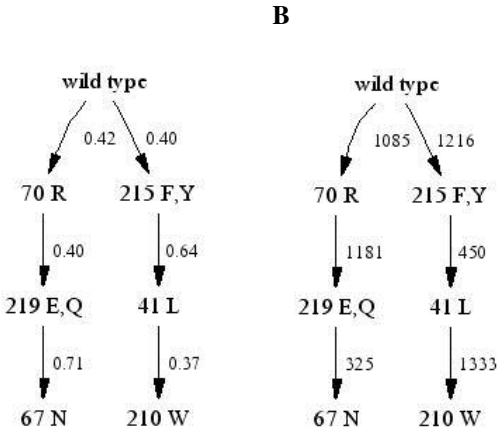


Figure 1. Mutagenetic tree for the development of zidovudine resistance. Vertices denote amino acid changes from the wild type, edges are labeled with conditional probabilities (A) and expected waiting times in days (B), respectively.

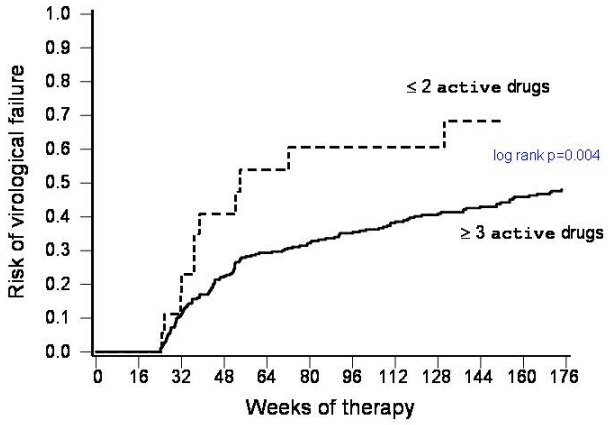
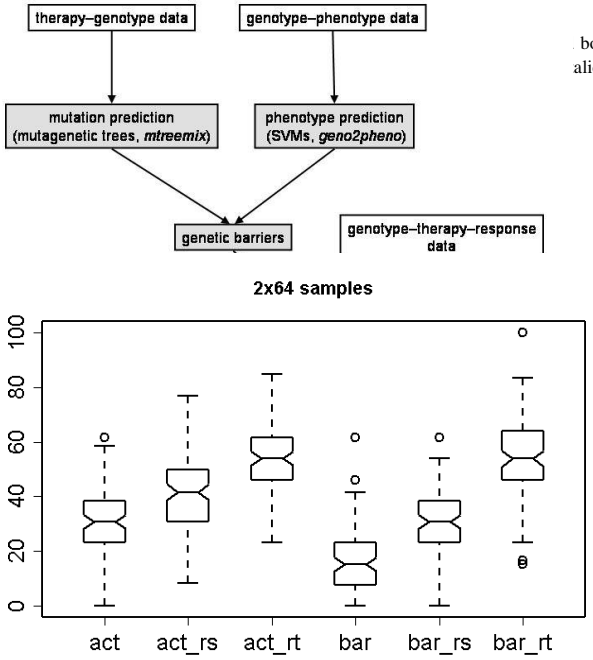


Figure 2. Risk of virological failure (two consecutive virus load values of more than 500 cps/ml after 24 weeks of therapy) as a function of the number of weeks on therapy. Two patient groups are distinguished according to whether the number of drugs scored as active is smaller than 3 or not. The two groups experience a significantly different risk of virological failure. (Data kindly provided by Andrea De Luca, Catholic University, Rome.)



boxes
alics).

Figure 4. Error rates for different scoring functions on a set of 128 genotype-therapy pairs in which each genotype occurs exactly twice, once with a drug combination resulting in a successful therapy (defined as undetectable virus load), and once with another drug combination resulting in therapy failure (defined as virus load above 1000 cps/ml). From left to right: activity scores (act), with sequences randomized (act_rs), with therapies randomized (act_rt), genetic barrier scores (bar), with sequences randomized (bar_rs), with therapies randomized (bar_rt).