Improved prediction of response to antiretroviral combination therapy using the genetic barrier to drug resistance

André Altmann[1], Niko Beerenwinkel[2], Tobias Sing[1], Igor Savenkov[1], Martin Däumer[3], Rolf Kaiser[3], Soo-Yon Rhee[4], W Jeffrey Fessel[5], Robert W Shafer[4] and Thomas Lengauer[1]*

[1]Max-Planck-Institute for Informatics, Saarbrücken, Germany
[2]Department of Mathematics, University of California, Berkeley, CA, USA (current address: Program for Evolutionary Dynamics, Harvard University, Cambridge, MA, USA)
[3]Institute of Virology, University of Cologne, Germany
[4]Division of Infectious Diseases, Stanford University, Stanford, CA, USA
[5]Kaiser-Permanente Medical Care Program Northern California, San Francisco, CA, USA

*Corresponding author: Tel: +49 681 9325 308; Fax +49 681 9325 399; E-mail: altmann@mpi-inf.mpg.de

Background: The outcome of antiretroviral combination therapy depends on many factors involving host, virus, and drugs. We investigate prediction of treatment response from the applied drug combination and the genetic constellation of the virus population at baseline. The virus's evolutionary potential for escaping from drug pressure is explored as an additional predictor.
Methods: We compare different encodings of the viral genotype and antiretroviral regimen including phenotypic and evolutionary information, namely predicted phenotypic drug resistance, activity of the regimen estimated from sequence space search, the genetic barrier to drug resistance, and the genetic progression score. These features were evaluated in the context of different statistical learning procedures applied to the binary classification task of predicting virological response. Classifier performance was evaluated using cross-validation and receiver operating characteristic curves on 6,337 observed treatment change episodes from the Stanford HIV Drug Resistance Database and a large US clinic-based patient population.
Results: We find that the choice of appropriate features affects predictive performance more profoundly than the choice of the statistical learning method. Application of the genetic barrier to drug resistance, which combines phenotypic and evolutionary information, outperformed the genetic progression score, which uses exclusively evolutionary knowledge. The benefit of phenotypic information in predicting virological response was confirmed by using predicted fold changes in drug susceptibility. Moreover, genetic barrier and predicted phenotypic drug resistance were found to be the best encodings across all datasets and statistical learning methods examined.
Availability: THEO (THErapy Optimizer), a prototypical implementation of the best performing approach, is freely available for research purposes at http://www.geno2pheno.org.

Introduction

Today's treatment options against HIV type 1 (HIV-1) involve about 20 antiretroviral agents targeting three different stages of the viral replication cycle: cell entry, reverse transcription and virion maturation. The drugs are grouped into four classes according to their molecular target and mechanism of action. The first, nucleoside and nucleotide reverse transcriptase inhibitors (NRTIs), are chemically modified versions of deoxynucleosides that interfere with reverse transcription by blocking chain elongation after their incorporation into newly synthesized DNA. The second, non-nucleoside reverse transcription inhibitors (NNRTIs), bind to the viral reverse transcriptase and block DNA polymerization by impairing the mobility of particular RT domains. Protease inhibitors (PIs) constitute the third class of antiretroviral drugs. They occupy the active site of the viral protease and prevent the maturation of released virions. Finally, the entry inhibitors (EIs) block virus entry into the host cell by binding to the viral transmembrane protein gp41. Besides these approved drugs, several additional compounds, some with novel modes of action, are being developed or have already entered clinical trials [1].

The goal of highly active antiretroviral therapy (HAART) is to combine three or more drugs, typically from at least two different classes, in order to suppress HIV replication and to prevent progression to AIDS and death. However, despite the use of drug cocktails, treatment failure is not uncommon. Virological failure, which refers to a rebound in viral load (VL), is distinguished from immunological failure, which is manifested by a decline in $CD4^+$ T-cell levels. Usually, virological failure precedes immunological failure, but discordant responses have been observed in some patients [2].

The rationale for HAART is to maximally suppress virus replication and to avoid (or at least retard) the development of drug resistance. The emergence of drug resistance mutations in the virus population is both a major cause and a consequence of therapy failure [3]. Established resistance mutations lead to therapy failure and result in the loss of future treatment options. Moreover, resistance to a drug in a regimen often confers reduced susceptibility to other drugs in the same class, a phenomenon known as cross-resistance. The speed and the extent of resistance development depend on many factors, including drug treatment history, the current regimen, patient adherence, plasma drug levels, the present virus population, and the immune status of the host. Combining drugs from different classes can slow down the emergence of resistant variants substantially, because mutants that are resistant to all components are unlikely to pre-exist, and new variants need to generate several escape mutations while retaining the ability of effective replication. A typical first-line HAART regimen would consist of two NRTIs and one NNRTI or PI [4].

With about 20 drugs available and novel drugs being approved almost every year, it becomes increasingly difficult for the treating physician to select an optimal drug combination. There are a variety of different therapeutic goals, including VL reduction, increase in $CD4^+$ T-cell counts, minimization of adverse effects and preservation of future drug options. Importantly, different optimization criteria will tend to favour different therapies [5]. To date, most methods predict virological response to therapy based on the baseline genotype and the compounds in the applied combination. Specifically, artificial neural networks [6] and fuzzy rules combined with a genetic algorithm [7] were used in this manner to predict the change in VL. Related approaches include the application of case-based reasoning [8] and combinatorial optimization based on expert rules [9].

In this paper, we report new ways of analysing treatment change episodes (TCE) and demonstrate how these new approaches might be more useful than current methods for prediction of response to therapy. The improvement results from incorporating genetic analysis, phenotypic prediction, and a prediction of the probability that further evolution of resistance will occur. The applied methodologies involve various techniques of statistical learning. We dichotomize virological response and compare the performance of several classifiers that predict the therapeutic success or failure of each genotype–therapy pair. The novel features derived from genotype and drug combination encode information about the evolutionary potential of the virus and the predicted level of phenotypic drug resistance. Specifically, we consider predicted phenotypes [10], the activity score based on a heuristic search over *in silico* mutants [11], the genetic barrier [12] and the genetic progression score [13]. Analysing over 6,300 TCEs observed in a clinical setting, we show that all new descriptors significantly improve prediction of therapy outcome, especially if they combine evolutionary and phenotypic information.

Methods

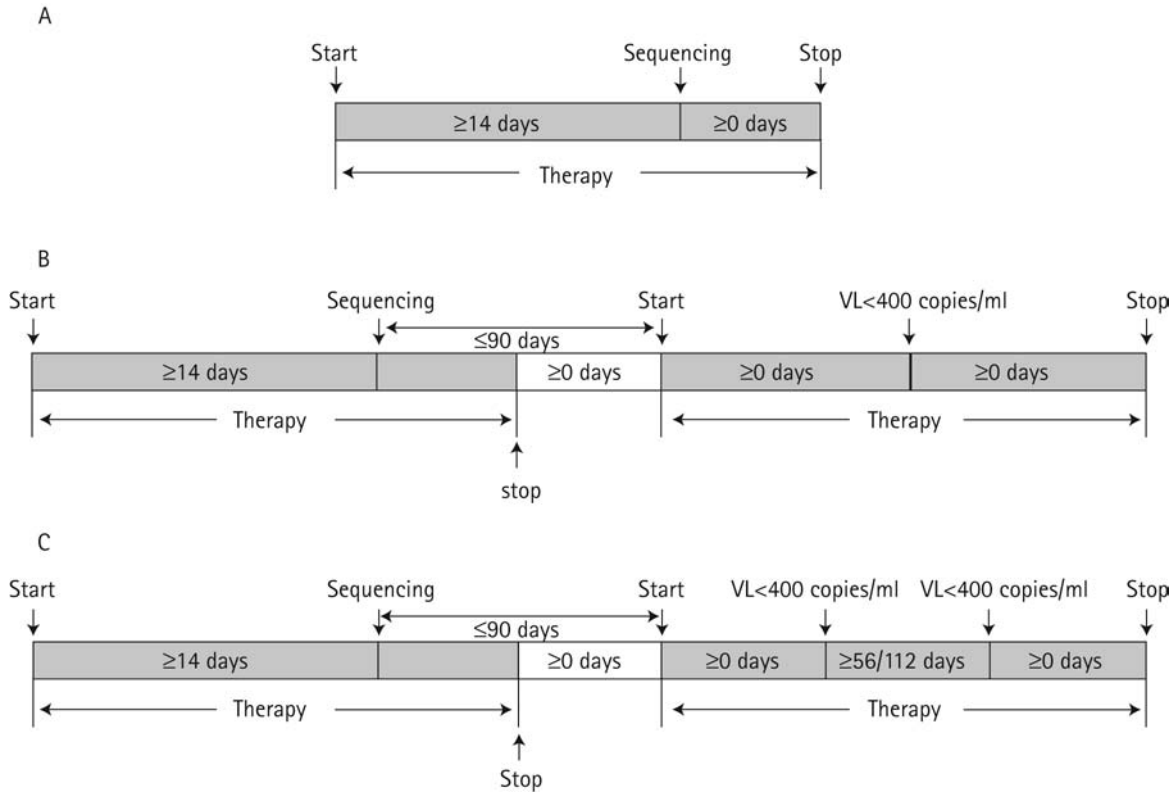
Treatment change episodes

A TCE [14] consists of a baseline genotype, a drug combination, and a binary outcome indicating success or failure of the regimen. For our analysis, valid successful or failing TCEs were defined as follows (Figure 1): any available genotype is considered as evidence of a failing regimen, because, in general, sequencing can only be performed if the VL exceeds ~1,000 copies/ml. Successful regimens are defined by inspecting therapies that follow a genotype measurement. When multiple genotypes are available, the most recent sequence sample before the onset of therapy was used. If the VL decreases below 400 copies/ml at least once during the course of the follow-up therapy and if genotyping was performed no earlier than 3 months before starting the therapy, then the respective treatment is considered a success. This definition of success focuses on initial response. Sustained response is also investigated by using an alternative definition of success that requires a second follow-up VL value below the threshold at least 8 weeks (or 16 weeks) after the first (Figure 1C).

Datasets

We analysed data obtained from the Stanford HIV Drug Resistance Database (comprising data from clinical studies ACTG 320, ACTG 364, GART and HAVANA) and from two Northern California clinic populations undergoing genotypic resistance testing at Stanford University. From a total of 25,717 therapies, 10,288 sequences, and 6,706 patients, we extracted 6,337 TCEs, including 4,776 failures and 1,561 successes, according to the definition based on initial

Figure 1. Definition of failing or successful treatment change episodes



A treatment change episode (TCE) resulting in sequencing is always considered to be a failure **(A)**. If the virus is undetectable during the treatment following a failure, the TCE is called a success **(B)**. In addition, the alternative definition for successful TCEs requires two subsequent viral load (VL) measurements below the threshold with at least 8 (or 16) weeks in between **(C)**.

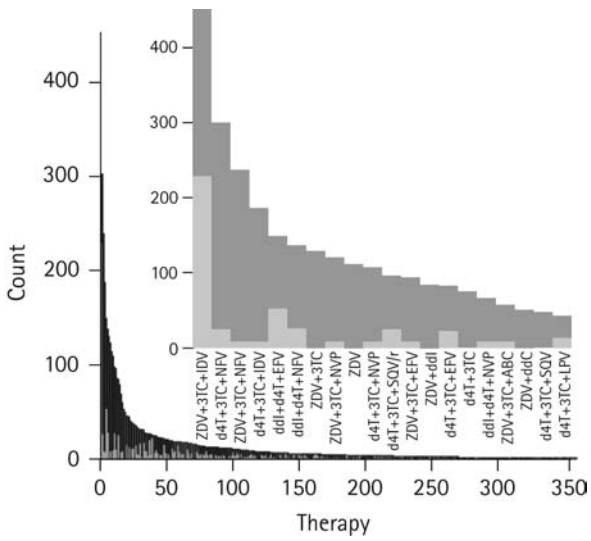
response. The median time period between treatment change and the first follow-up VL measurement <400 copies/ml is 38 weeks (interquartile range: 16–78 weeks). Figure 2 shows the distribution of drug combinations for this dataset (denoted A). The alternative definitions resulted in 1,082 and 900 successes for 8 and 16 weeks sustained response, respectively. Because of lack of data on the drug enfuvirtide (currently the only approved EI), we considered only regimens consisting of NRTIs, NNRTIs and PIs.

From dataset A, we generated two special subsets that are balanced with respect to sequences (BS) and with respect to therapies (BT), respectively. The dataset BS contains one successful and one failing regimen for each of 1,364 sequences exactly. Each genotype in this dataset gives rise to two TCEs, defined by the regimens before and after a successful treatment change. Similarly, in the set BT each regimen is paired with two different sequences, giving rise to exactly one successful TCE and one failure. This selection resulted in a total of 2,436 TCEs comprising 321 different regimens. If, for a fixed drug combination, there were more sequences resulting in failure than in success, then failures have been selected randomly to match the number of successes (or *vice versa*).

Dataset A was split into two groups denoted A1 and A2 according to the clinical centre in which the TCE was observed. Group A1 comprises 816 successes and 2,098 failures from one health care provider, and group A2 includes the remaining 745 successes and 2,678 failures from clinical studies and other hospitals. By using, in turn, one of these subsets for training and the other for testing, we seek to identify possible biases that might be related to the specifics of individual clinics, such as preferred treatment protocols. Datasets BS and BT were split in the same manner.

In addition to the clinical data, we also used *in vitro* data on phenotypic drug resistance. For each drug, 880 sequences with fold change (FC) determined by a recombinant virus assay [15] were available. These data were used to regress phenotype on genotype and to identify resistant states in defining the genetic barrier (described below).

Figure 2. Distribution of drug combinations in the complete dataset A



The 6,337 analysed TCEs comprise a total of 875 distinct combination therapies. The histogram summarizes all 354 drug combinations that occur at least three times in the dataset. Another 116 combinations appeared only twice and 405 only once. Grey bars indicate successful therapies defined as initial virological responses below 400 copies/ml. The inlet histogram shows the 20 most abundant drug combinations.

Features

The baseline approach for predicting TCE outcome uses one indicator variable for each resistance-associated mutation and one for each drug. For this approach and approaches based on mutagenetic trees (described below), we considered the resistance mutations presented in [16] resulting in 66 binary variables (49 mutation indicators, 17 drug indicators). We refer to this encoding of genotypes and therapies as the indicator representation. All other encodings contain these straightforward covariates and additional, more elaborate, features.

The phenotype representation includes, for each drug in the respective regimen, the predicted FC in susceptibility. These predictions are based on a linear support vector machine that has been trained on the 880 matched genotype–phenotype pairs as described previously [10].

In the activity representation, the indicator vector is extended by the estimated activity of the drug cocktail against the virus population. The notion of single-drug activity is based on the distribution of predicted FCs and reflects the probability that the sample is phenotypically susceptible given the predicted FC. Evolutionary information is included by introducing mutations into the considered genotype and searching sequence space by following mutants of least activity. The activities of the worst-case mutants at each level of search depth are then combined into a single activity score. We refer to [11] for a detailed description of this procedure.

The genetic barrier representation also adds both phenotypic and evolutionary information to the indicator representation, and it can be regarded as an advancement over the activity score. More precisely, we used mutagenetic trees, a family of probabilistic graphical models, to estimate the order and rate of occurrence of resistance mutations. Using the Mtreemix software (http://mtreemix.bioinf.mpi-sb.mpg.de/) [17], for each drug, a mixture model of mutagenetic trees was learned from sequences derived under regimens comprising that drug. A validation of mutagenetic tree models in terms of tree stability and goodness of fit has been presented in [18]. Based on these evolutionary models, we defined the genetic barrier as the probability that the virus will not escape from drug pressure by developing further mutations [12]. Here, viral escape is approximated by exceeding a predefined level of phenotypic resistance. These levels are defined by the following cut-offs for the FC in susceptibility: zidovudine 30.0; zalcitabine 2.2; didanosine 2.4; stavudine 2.0; lamivudine 15.4; abacavir 3.4; tenofovir disoproxil fumarate 2.1; nevirapine 9.0; delavirdine 9.7; efavirenz 7.0; saquinavir 4.5; indinavir 4.6; ritonavir 2.6; nelfinavir 5.8; amprenavir 12.0; lopinavir 10.0 and atazanavir 4.2. Unlike the sequence space search for low-activity mutants, the genetic barrier accounts for the fact that not all mutations are equally likely to occur. This is also an advantage over simple counting of resistance mutations, a frequently employed approximation to the genetic barrier.

Finally, the genetic progression score (GPS) involves only evolutionary information that is extracted from the mutagenetic tree models. The GPS of a genotype is defined as the expected waiting time for the mutational pattern to occur [13]. Thus, the GPS also accounts for different probabilities of different mutations, but it does not include any phenotypic information. We emphasize that the GPS is not intended to estimate waiting times on the real time scale. Rather it provides a dimension-free measure of genetic progression that allows for comparing mutational patterns.

Statistical learning methods

The five feature sets defined the input to several different machine-learning techniques, which were used to predict treatment response. We selected several standard classification methods [19], including linear discriminant analysis (LDA), which was used in [11], together with the activity representation, least-squares regression, linear support vector machines (SVMs), decision trees (C4.5 software, http://www.rulequest.com) and logistic regression. We also included the more

Table 1. Classifier performance on the full dataset A

	LDA	LSR	SVM	C4.5	LOGR	LMT	Mean
Indicator	0.825 (0.005)	0.825 (0.005)	0.819 (0.009)	0.845 (0.005)	0.820 (0.007)	0.868 (0.004)	0.834 (0.008)
+ Phenotype	0.912 (0.004)	0.911 (0.004)	0.910 (0.004)	0.839 (0.005)	0.912 (0.002)	0.905 (0.004)	0.898 (0.012)
+ Activity	0.868 (0.005)	0.870 (0.005)	0.864 (0.006)	0.842 (0.007)	0.868 (0.003)	0.898 (0.004)	0.868 (0.007)
+ Genetic barrier	0.891 (0.005)	0.892 (0.003)	0.891 (0.005)	0.875 (0.005)	0.891 (0.002)	0.916 (0.005)	0.893 (0.005)
+ GPS	0.856 (0.006)	0.856 (0.005)	0.864 (0.005)	0.861 (0.005)	0.868 (0.005)	0.899 (0.004)	0.867 (0.007)
Mean	0.870 (0.015)	0.871 (0.015)	0.870 (0.015)	0.852 (0.007)	0.872 (0.015)	0.897 (0.008)	

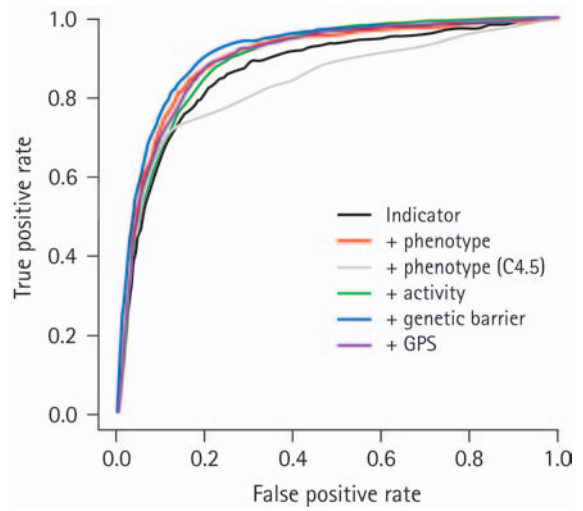
The table displays, for all combinations of feature encodings (rows) and learning techniques (columns), the resulting area under the receiver operating characteristic (ROC) curve (AUC) and its standard error (in parentheses) computed using 10-fold cross-validation. C4.5, C4.5 software; GPS, genetic progression score; LDA, linear discriminant analysis; LMT, logistic model trees; LOGR, logistic regression; LSR, least-squares regression; SVM, linear support vector machines.

recent method of logistic model trees (LMT), which combine decision trees with logistic regression in the leaves of the tree [20].

Receiver operating characteristic analysis

We used receiver operating characteristic (ROC) curves to compare the predictive power of classifiers. ROC curves arise from varying a parameter of the classifier that controls the trade-off between sensitivity and specificity. Each point on the ROC curve represents one classifier, and the curve allows for reading off its false positive and true positive rate. For example, the point (0.1, 0.8) represents a classifier that will falsely predict 10% of failing regimens as 'success', but correctly detect 80% of the successful regimens. The comparison of ROC curves is preferred to comparing error rates, because it corrects for skewed class distributions (as in dataset A) and controls both sensitivity and specificity of the classifier [21]. The area under the ROC curve (AUC) is used as a summary performance measure. The trivial classifier that makes random predictions produces a linear ROC curve with an AUC of 0.5. The maximum AUC a classifier can achieve is 1 and the larger this value, the better its prediction performance. AUC values were compared using Wilcoxon rank-sum tests. ROCR software (http://bioinf.mpi-sb.mpg.de/projects/rocr/) as used for ROC analysis [22].

Figure 3. ROC curves for the complete dataset A using LMT (unless stated otherwise in the legend) and 10-fold cross validation



Every feature encoding is represented by a receiver operating characteristic (ROC) curve, namely the baseline indicator representation (indicator), and the following additional features: predicted phenotypes, activity score, genetic barrier, and genetic progression score (GPS). Each point on the curve represents a classifier and allows determining its true positive rate and false positive rate. For example, the point (0.355, 0.9) on the black line represents a logistic model trees (LMT) classification model trained on the plain indicator representation with an expected 35.5% of false positives and 90% of true positives. C4.5, C4.5 software.

Results

Table 1 shows the performance of all learning techniques in combination with all feature encodings for the complete dataset A using the initial response definition of therapeutic success. Classifier performance was estimated by 10-fold cross-validation and is reported as the AUC. All of the proposed extensions of the indicator representation improved predictive power (Table 1, last column). This improvement was observed across all statistical learning methods. The additional features activity score and GPS yield the same predictive power on average. The largest improvement is achieved by the genetic barrier and by phenotype predictions, which both use phenotypic information for prediction. Compared with the feature encoding, the choice of the learning method has only a small effect. On average, the AUC differs by as much as 0.064 (7.7%) between different feature encodings, but only by 0.027 (3.2%) between different learning methods. The AUC of the phenotype representation decreases if combined with decision trees. However, as illustrated in Figure 3, the ROC curves reveal that this difference is mainly due to

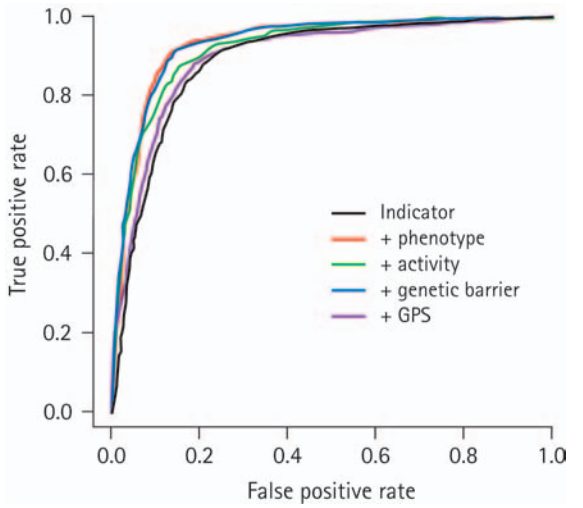
lower true positive rates at very high false positive rates, which might not be relevant for practical purposes. We restrict the further analysis of ROC curves to the LMT learning method, which, on average, outperformed all other techniques (Table 1).

In Figure 3, the ROC curves for the five different feature encodings used with LMT on dataset A are shown. The indicator representation (black line) is improved significantly by all four additional features (P<0.0002). This advancement is most prominent for the genetic barrier that incorporates phenotypic information indirectly (P=0.0002), followed by the predicted phenotype, the activity score and the GPS. If we accept a false positive rate of 10%, then the indicator representation will detect 65% of the successful TCEs correctly (represented by the point [0.1, 0.65] on the black line in Figure 3), whereas the genetic barrier representation achieves an accuracy of 76.6%. The accuracy can be further increased by accepting more false positives. For example, if 90% of the successes were to be detected, we would have to accept 19.5% false positives for the genetic barrier or the phenotype encoding, and 35.5% for the indicator representation. The differences between the four encodings are even more strongly articulated in the analysis of the two balanced subsets of the complete dataset.

In the dataset BS, each viral genotype is paired with a drug combination that gave rise to a successful TCE and with another TCE that resulted in failure. Thus, the genotype alone does not provide any information on the outcome of these TCEs. As might have been expected, the GPS does not improve the predictive power on this dataset, because this feature is derived only from the genotype (Figure 4). By contrast, the remaining three encodings, namely activity, genetic barrier, and phenotype, enhance the performance significantly (P<0.004). The genetic barrier encoding outperforms the activity encoding, and the phenotype representation provides the best encoding on this dataset. However, the difference between genetic barrier and phenotype is negligible (P>0.9).

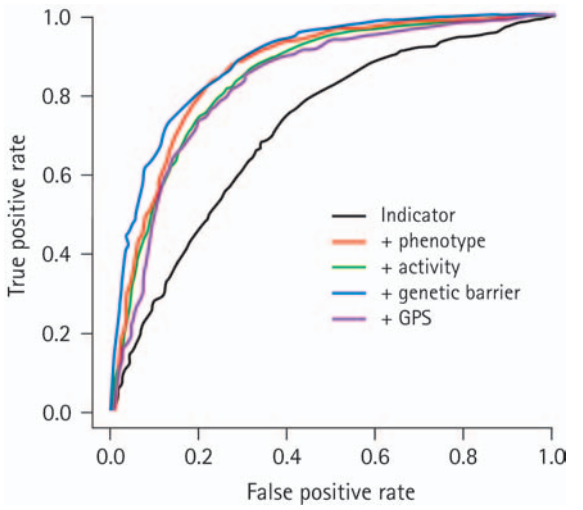
In the dataset BT, the therapies (rather than the sequence, as in BS) are balanced. For every therapy there exists the same number of genotypes that gave rise to successful and to failing TCEs. Here, the drug combination alone does not provide any information on the outcome of the TCEs. Usage of the GPS on this dataset increases the performance of the indicator representation to the level reached with the activity representation (Figure 5). Similarly to dataset BS, application of the phenotypic and the genetic barrier representation results in maximal performance. Here, the genetic barrier encoding outperforms the phenotype encoding, but the difference does not reach statistical significance (P=0.1655). In the interesting region of low false positive rates (<30%), however, the difference can be substantial. For example, at a false positive rate of 10% the indicator approach recognizes 28% of the successful TCEs correctly, GPS yields a true positive rate of 54.7%, using the activity score increases true positives to 54.5%, the phenotype representation achieves almost 58%, and the

Figure 4. ROC curves for dataset BS using LMT and 10-fold cross validation



For more details see legend of Figure 3. BS, balanced with respect to sequence; GPS, genetic progression score; LMT, logistic model trees; ROC, receiver operating characteristic.

Figure 5. ROC curves for dataset BT using LMT and 10-fold cross validation



For more details see legend of Figure 3. BT, balanced with respect to therapy; GPS, genetic progression score; LMT, logistic model trees; ROC, receiver operating characteristic.

genetic barrier recognizes as many as 64.9% of successes correctly. This rate more than doubles the precision of the indicator encoding alone.

In order to investigate possible biases in the estimated models resulting from the specific clinical centres which collected the TCE data, we used the split of dataset A into A1 and A2. The TCEs in A1 originate from a single health care provider, but those in A2 stem from several different clinical studies and hospitals and hence are expected to be less homogeneous. When A1 and A2 are used separately in the same cross-validation procedure described above for the pooled dataset A, then the predictive performance is similar in both cases (AUC of 0.898 for A1 and 0.906 for A2). If A1 (A2) is used for training and A2 (A1) for testing, we estimate an AUC of 0.837 (0.875) The performance loss due to separation of the data by clinics was slightly more pronounced for the balanced dataset BS than for BT (data not shown).

All of the reported results remain qualitatively unchanged when therapeutic success is defined by a more sustained response over at least 8 or 16 weeks instead of by initial response. Using the alternative definitions we re-calculated the AUC values for all feature encodings with LMT on all datasets. None of the derived performance measures showed any significant difference compared with the initial definition.

Discussion

Given the increasing number of possible drug combinations and the genetic diversity of HIV, it is unlikely that simple hand-crafted rules will capture the complex interplay between drug cocktails and mutational patterns that determine response to antiretroviral therapy. Thus, statistical and computational approaches are required for optimal use of the available drugs in each individual patient. Here we have analysed the ability of various statistical learning techniques in combination with different feature encodings to predict therapy outcome from the baseline genotype and the applied drug combination.

The viral genotype is only one of many patient-specific characteristics, such as immune status or genetic predisposition, and it is straightforward to extend our approach to situations where additional parameters are available. Nevertheless, the viral genotype has a prominent role among those covariates as it encodes the structure of the target proteins. The challenge in using these genetic data is to predict the evolution of the virus population under drug pressure and to understand the complex relationship between mutational patterns and *in vivo* drug resistance. We have addressed these issues by considering, in addition to drug and mutation indicators, features that make use of phenotypic and evolutionary information. Our computational experiments have identified the predicted phenotype and the genetic barrier to drug resistance as the most beneficial features, significantly boosting the predictive power of all classifiers. The choice of the learning technique had less impact, but LMT consistently showed an advantage over the other methods. In general, the representations that incorporate *in vitro* phenotype predictions yielded better performance than the GPS, a purely evolutionary feature, but both sources of information led to improvements among all tested classifiers and datasets.

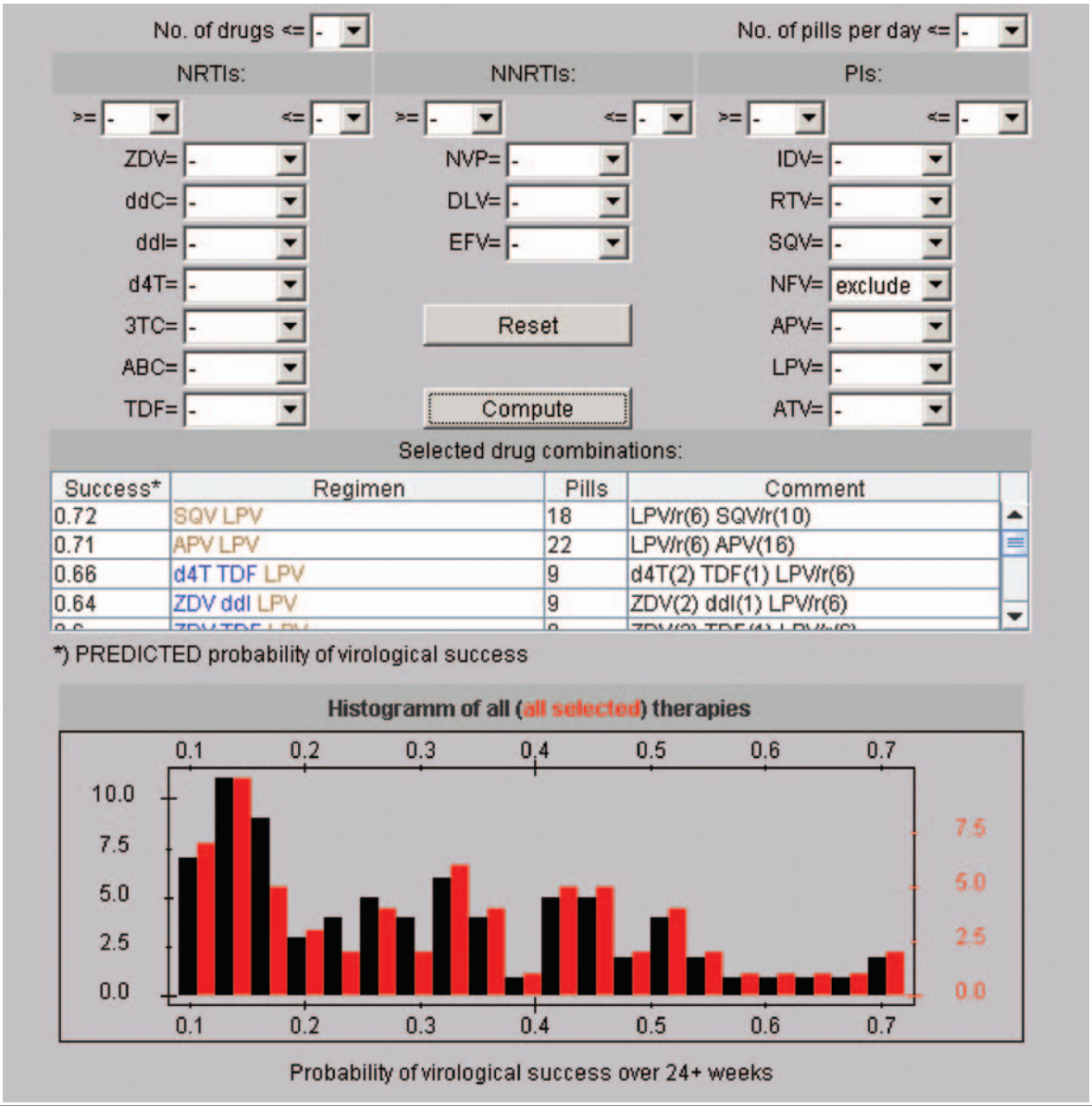
The two subanalyses of the complete TCE dataset A showed opposing results. Whereas the performance was increased compared with dataset A for the balanced sequences (BS), it was decreased for the balanced therapies (BT). In the case of the BS data, classifiers need to learn the differences between drug combinations conditioned on the genotype. However, in effect, with this dataset the dependence on genotype is largely masked by the differing application profiles of regimen use accumulated in the dataset. For example, lopinavir appeared in only 38 failing regimens and in 284 successful follow-up therapies. Likewise, the combination of zidovudine and didanosine defined 60 failures, but not a single success. This is because the underlying clinical cohort data reflects the historical approval and use of drugs. For example, the combination of zidovudine and didanosine has been available for some time and many patients have received it until failure. This is easy to learn from the data, but the generalization that this combination always fails, although strongly supported by the data, is certainly wrong and thus misleading when evaluating treatment options containing zidovudine and didanosine. This problem is eliminated with the other subset, BT, where classifiers learn differences between mutational patterns conditioned on the regimen. Because, in learning therapy outcome, we aim to generalize the mutation–drug interactions and not to reconstruct historical drug-use patterns, we regard the performance using dataset BT as a more genuine estimate of our ability to predict treatment response.

We have addressed the concern that the learned models might be biased by the specific sampling of patients by analysing TCEs from different clinics separately. We found a small decrease in predictive performance that was more pronounced when the more homogeneous subset A1 was used for training. This finding highlights the importance of including data from many different clinical centers. The performance loss was greater for the BS dataset than for the BT dataset indicating that the patterns of drug use can vary among clinics. Larger studies that include

sufficient data from several clinics need to investigate this source of bias in more detail in the future.

In the present study, we have dichotomized therapy response into 'success' and 'failure'. However, with response defined as the change in VL after a certain time, regression methods can be used for learning in a similar way, and this is unlikely to affect the results presented here. Furthermore, most classifiers predict a score (in fact, often a probability) rather than only the class label. Thus, for a given genotype, they can be used for ranking drug combinations with respect to the expected therapeutic success. In the future, ranking systems could evolve into valuable tools for supporting the complex decision-making process of clinicians. However, such systems will have to provide an interface that allows physicians to incorporate their prior

Figure 6. The THEO applet for selecting and evaluating drug combinations



The applet allows for limiting the number of drugs that can be part of a therapy (No. of drugs). Furthermore, the daily burden (No. of pills per day) can be limited. The number of compounds from each drug class can be set (for example, 1≤NRTIS≤3), and the use of single drugs can be enforced (for example, 3TC = include) or excluded. Pushing the 'Compute' button ranks all remaining therapies according to the underlying prediction model. In the resulting table, the components of the regimens are listed (Regimen), the number of pills, how many pills of every compound are included in the regimen (Comment), and the score calculated by the model (Success). Additionally, the predictions are compared in a histogram between all selected (according to the constraints specified by the user; red bars) and all possible therapies (black bars). The figure shows the result of a THEO analysis of a genotype from a patient failing treatment after first line therapy with abacavir, lamivudine and efavirenz. The sequence (subtype B) contained the following mutations (compared with the reference strain HXB2): PR V3I, E35D, S37N, L63P and A71T and RT L74V, K102N, K103N, Y115F, E122K, D123E, D177E, I178V, V179D, M184V, G190A, T200A, R211K, L214F, H221Y, T286A, E297R and S322T.

knowledge on a given case, and to constrain the range of possible regimens by explicitly excluding or including specific drugs. In this way, it is possible to spare drugs or drug classes, limit total pill burden, or react to adverse effects, while being able to identify the most promising options from the remaining regimens. The development of improved therapy rankers will crucially depend on their public availability, which allows experts to identify strengths and weaknesses of different approaches. For this reason, we have implemented the prototypical therapy ranker THEO (THErapy Optimizer, Figure 6). It is freely accessible for research purposes as part of the geno2pheno web site (http://www.geno2pheno.org). In order to produce a ranking of therapies, THEO applies the classifier trained for predicting therapy response to the sequences of PR and RT provided by the user and a predefined set of therapies (consisting of any combination of either two PIs or two NRTIs plus one NNRTI or one PI). The score produced by the classifier for every combination of drugs is the expected therapeutic success, which is used to rank the considered therapies. In particular, THEO applies the LMT classifier, which was trained on dataset BT using the genetic barrier encoding as input, to derive the scores needed for the ranking. This selection of feature encoding and statistical learning method was based on the cross-validation results obtained on dataset BT. Figure 6 depicts the results of a THEO analysis of a genotype from a patient failing treatment after first-line therapy with abacavir, lamivudine and efavirenz (for a list of mutations, see legend of Figure 6). The first two regimens proposed by THEO are double PI therapies (saquinavir and lopinavir/ritonavir, and amprenavir and lopinavir/ritonavir) with a predicted probability of virological success of over 70%.

As discussed above, the statistical model used by THEO has some limitations. Thus, suggestions made by THEO must be used with care. Moreover, the ranking displayed by THEO might appear counter-intuitive, because it is based on confidence scores provided by the statistical learning method and the single criterion for therapy success is reduction of VL below a threshold. Hence, regimens that are highly active tend to occur among the top-ranked therapies. Although these regimens are most likely to reduce the VL below the threshold necessary for classifying a TCE as a success, less active regimens might be a better choice for preserving future drug options and might therefore be ranked higher by clinicians [5]. Thus, we emphasize that the purpose of therapy rankers will always be to support, and not to replace, the complex decision-making process of clinicians.

In future work, further improvement of the genetic barrier representation might be achieved by estimating mutagenetic trees for combinations of drugs instead of single drugs. Such trees would model evolutionary pathways to multi-drug resistance and handle the interplay of drugs within the applied regimen. However, this approach is currently infeasible due to the lack of sufficient data. We also emphasize that the drug-wise computation of the genetic barrier, as employed here, facilitates the incorporation of a new drug, because only one additional mutagenetic tree needs to be learned. Moreover, to a first approximation, this tree can be learned from *in vitro* and clinical study data even prior to approval. The most promising avenues to increased prediction accuracy and improved therapy ranking are via larger datasets, the use of additional parameters (such as $CD4^+$ T-cell counts, plasma levels of drugs, viral replication capacity, and host genetic factors), analysis of previous sequences of the viral population that represent important background information, and a profound understanding of viral evolutionary escape from drug pressure.

Acknowledgements

Part of the work at Max-Planck-Institute for Informatics was supported by the Eu*Resist* project (IST -4- 027173-STP). N Beerenwinkel was supported by Deutsche Forschungsgemeinschaft (DFG grant BE 3217/1-1).

Parts of this work have been presented at two international conferences [23,24].

References

1. Reeves JD, Piefer AJ. Emerging drug targets for antiretroviral therapy. *Drugs* 2005; **65:**1747–1766.
2. Grabar S, Le Moing V, Goujard C, Leport C, Kazatchkine MD, Costagliola D, *et al*. Clinical outcome of patients with HIV-1 infection according to immunologic and virologic response after 6 months of highly active antiretroviral therapy. *Ann Intern Med* 2000; **133:**401–410.
3. Clavel F, Hance AJ. HIV drug resistance. *N Engl J Med* 2004; **350:**1023–1035.
4. Dybul M, Fauci AS, Bartlett JG, Kaplan JE, Pau AK. Panel on Clinical Practices for Treatment of HIV Infection, Guidelines for the Use of Antiretroviral Agents in HIV-1–Infected Adults and Adolescents. *Ann Intern Med* 2002; **137:**381–433.
5. Jiang H, Deeks SG, Kuritzkes DR, Lallemant M, Katzenstein D, Albrecht M, *et al*. Assessing resistance costs of antiretroviral therapies via measures of future drug options. *J Infect Dis* 2003; **188:**1001–1008.
6. Wang D, Larder BA, Revell A, Harrigan R, Montaner J. A neural network model using clinical cohort data accurately predicts virological response and identifies regimens with increased probability of success in treatment failures. *Antiv Ther* 2003; **8:**U99–U99.
7. Prosperi M, Di Giambenedetto S, Trotta MP, Cingolani A, Ruiz L, Baxter JD, *et al*. A fuzzy relational system trained by genetic algorithms and HIV-1 resistance genotypes/virological response data from prospective studies usefully predicts treatment outcomes. *Antiv Ther* 2004; **9:**U89–U89.

8. Prosperi M, Zazzi M, Perno CF, Di Giambenedetto S, Baxter J, Ruiz L, *et al*. 'Common law' applied to treatment decisions for drug resistant HIV. *Antiv Ther* 2005; **10:**S62–S62.
9. Lathrop RH, Pazzani MJ. Combinatorial optimization in rapidly mutating drug-resistant viruses. *J Comb Optim* 1999; **3:**301–320.
10. Beerenwinkel N, Daumer M, Oette M, Korn K, Hoffmann D, Kaiser R, *et al*. Geno2pheno: Estimating phenotypic drug resistance from HIV-1 genotypes. *Nucleic Acids Res* 2003; **31:**3850–3855.
11. Beerenwinkel N, Lengauer T, Daumer M, Kaiser R, Walter H, Korn K, *et al*. Methods for optimizing antiviral combination therapies. *Bioinformatics* 2003; **19 Suppl 1:**i16–25.
12. Beerenwinkel N, Daumer M, Sing T, Rahnenfuhrer J, Lengauer T, Selbig J, *et al*. Estimating HIV evolutionary pathways and the genetic barrier to drug resistance. *J Infect Dis* 2005; **191:**1953–1960.
13. Rahnenfuhrer J, Beerenwinkel N, Schulz WA, Hartmann C, von Deimling A, Wullich B, *et al*. Estimating cancer survival and clinical outcome based on genetic tumor progression scores. *Bioinformatics* 2005; **21:**2438–2446.
14. Larder BA, Wang D, Revell A, Lane C. Neural network model identified potentially effective drug combinations for patients failing salvage therapy. *2nd IAS Conference on HIV Pathogenesis and Treatment* 13–17 July 2003, Paris, France. Poster LB39.
15. Walter H, Schmidt B, Korn K, Vandamme AM, Harrer T, Uberla K. Rapid, phenotypic HIV-1 drug sensitivity assay for protease and reverse transcriptase inhibitors. *J Clin Virol* 1999; **13:**71–80.
16. Johnson VA, Brun-Vezinet F, Clotet B, Conway B, Kuritzkes DR, Pillay D, *et al*. Update of the drug resistance mutations in HIV-1: Fall 2005. *Top HIV Med* 2005; **13:**125–131.
17. Beerenwinkel N, Rahnenfuhrer J, Kaiser R, Hoffmann D, Selbig J, Lengauer T. Mtreemix: a software package for learning and using mixture models of mutagenetic trees. *Bioinformatics* 2005; **21:**2106–2107.
18. Beerenwinkel N, Rahnenfuhrer J, Daumer M, Hoffmann D, Kaiser R, Selbig J, *et al*. Learning multiple evolutionary pathways from cross-sectional data. *J Comput Biol* 2005; **12:**584–598.
19. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*, 2001; pp.79–114. New York: Springer.
20. Landwehr N, Hall M, Frank E. Logistic model trees. *Mach Learn* 2005; **59:**161–205.
21. Brun-Vezinet F, Costagliola D, Khaled MA, Calvez V, Clavel F, Clotet B, *et al*. Clinically validated genotype analysis: guiding principles and statistical concerns. *Antivir Ther* 2004; **9:**465–478.
22. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. *Bioinformatics* 2005; **21:**3940–3941.
23. Savenkov I, Beerenwinkel N, Sing T, Däumer M, Kaiser R, Lengauer T. HAART outcome prediction using statistical learning methods. *Antiv Ther*, 2005; **10:**S60 (Abstract #55).
24. Beerenwinkel N. The evolutionary potential of HIV predicts response to antiretroviral therapy. Workshop on Quantitative Methods for Research on Antiviral Resistance. 11–12 May, 2006, Boston, MA, USA.

Accepted for publication 15 September 2006