History-alignment models for bias-aware prediction of virological response to HIV combination therapy

Jasmina Bogojeska [1], **Daniel Stöckel** [2], **Maurizio Zazzi** [3]
Rolf Kaiser [4], **Francesca Incardona** [5], **Michal Rosen-Zvi** [6] **and Thomas Lengauer** [1]
[1]Max Planck Institute for Informatics, Germany; [2]Saarland University, Germany;
[3]University of Siena, Italy; [4]University of Cologne, Germany; [5]Informa, Italy; [6]IBM Research Labs, Israel

Abstract

The relevant HIV data sets used for predicting outcomes of HIV combination therapies suffer from several problems: different treatment backgrounds of the samples, uneven representation with respect to the level of therapy experience and uneven therapy representation. Also, they comprise only viral strain(s) that can be detected in the patients' blood serum. The approach presented in this paper tackles these issues by considering not only the most recent therapies but also the different treatment backgrounds of the samples making up the clinical data sets when predicting the outcomes of HIV therapies. For this purpose, we introduce a similarity measure for sequences of therapies and use it for training separate linear models for predicting therapy outcome for each target sample. Compared to the most commonly used approach that encodes all available treatment information only by specific input features our approach has the advantage of delivering significantly more accurate predictions for therapy-experienced patients and for rare therapies. Additionally, the sample-specific models are more interpretable which is very important in medical applications.

1 Introduction

Causing the acquired immunodeficiency syndrome (AIDS), with no cure or vaccine available, the human immunodeficiency virus (HIV) is among the deadliest pathogens in the history of mankind. It has claimed more than 27 million lives since its discovery in 1981 and the current number of infected people worldwide is larger than 33 million (UNAIDS/WHO, 2010). HIV patients are customarily treated with administration of combinations of several antiretroviral drugs. Although compared to individual drugs such a drug cocktail prolongs the time until HIV has acquired resistance to the therapy, it is eventually defeated by the evolution of HIV to resistance and needs to be replaced by a different drug combination. Finding a successful combination therapy on such an occasion, while keeping future therapy options open, is the central problem with which the physician is faced when treating HIV patients. However, selecting an appropriate therapy is quite hard, mainly because of the very large number of putative drug combinations (hundreds to thousands) and the large number of resistance-relevant mutations accumulated in the latent virus population in several tissues and organs. Such hidden mutations have to be taken into account because they are quickly accessed if this is beneficial for the virus after a therapy change.

The large amount of available clinical data combined with the use of advanced statistical learning methodology offer an automated computational approach to utilizing the available knowledge for predicting the outcome of a potential antiretroviral therapy. Such technology can therefore assist physicians in choosing a successful regimen for an HIV patient. This is not so central for first-line therapies, as there are specific guidelines for administering initial therapies, but it becomes highly relevant for therapy-experienced patients.

However, there are several important issues affecting the HIV clinical data sets. First of all, they comprise therapy samples that originate from patients with different treatment backgrounds. Also the specific treatment histories for the majority of these therapy-experienced samples are unique. Second, the various levels of therapy experience ranging from therapy-naïve to heavily pretreated are represented with different sample abundances – especially samples stemming from patients with higher therapy-experience levels are

Appearing in Proceedings of the 15^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume XX of JMLR: W&CP XX. Copyright 2012 by the authors.

underrepresented. Third, in most clinical data sets and in practice only the genotype of the most abundant viral strain in the patient's blood serum is available and considered when making a decision on the future therapy, while the remaining latent virus variants whose genomes are stored in different organs of the patient do not leave a mark. However, the information regarding the latent virus population is important for making accurate predictions for therapy-experienced HIV patients. Finally, the HIV clinical data sets contain data on different combination therapies with widely differing frequencies. In particular, many therapies are only represented with very few data points. All this creates what we refer to as *treatment bias* in the data sets which propagates to the derived statistical models and influences their predictions and usefulness.

This paper presents an interpretable statistical method for predicting outcomes of HIV combination therapies that deals with the treatment bias pertaining to the HIV clinical data sets. For this purpose, our approach takes not only the most recent (target) therapy but also available information on preceding therapies into account. First, we adapt techniques from sequence alignment to the problem of aligning sequences of therapies and use them to introduce a quantitative notion of pairwise similarity of therapy sequences. The similarity measure also incorporates information on the similarity of the corresponding genomic fingerprints in the latent virus population of the compared therapy sequences. Then, for each sample of interest, which is associated with a corresponding target therapy sequence, we train an individual model for predicting therapy outcome. The model utilizes the similarities between the target therapy sequence and the training therapy sequences in order to quantify the influence of the respective training samples on the model's prediction. In this way the model, incorporates information on the latent virus population, the specific therapies previously given to a patient and the order in which they were administered, on the one hand, and uses this information and all available data to deal with the treatment bias present in the clinical data, on the other hand.

1.1 Related work

Over the years a wide range of statistical learning methods, including artificial neural networks, decision trees, random forests, support vector machines (SVMs) multi-task learning and logistic regression (Wang et al., 2003; Larder et al., 2007; Deforche et al., 2008; Rosen-Zvi et al., 2008; Bickel et al., 2008; Altmann et al., 2009; Prosperi et al., 2009; Bogojeska et al., 2010; Revell et al., 2010), have emerged for tackling the problem of predicting the virological response to HIV combination therapies. Some of these approaches (Bickel et al., 2008; Rosen-Zvi et al., 2008) incorporate information on the previous therapies administered in a patient's history and thereby demonstrate the value of such knowledge. In the aforementioned publications the information on treatment history has been flattened to the set of different drugs that have been administered in any of the therapies that comprise the relevant treatment history record. While this simple approach can easily be incorporated in every statistical learning method, it neglects the information on the specific makeup of drug combinations comprising the patient's treatment history, their resulting viral genomic fingerprints in the latent viral population and the order in which they were administered. Saigo et al. (2010) present an approach denoted as sequence boosting for predicting therapy effectiveness targeted at therapy-experienced patients with completely recorded treatment history. This method incorporates information on the order in which the therapies were administered and shows the importance of such information for treatment-experienced patients. However, in the available HIV clinical data the information on whether the treatment information is complete or not is missing for the majority of the samples.

None of the approaches mentioned above tackles the bias introduced by the different treatment backgrounds of the samples and their sparse representation in the clinical data sets.

2 Methods

In this paper we present an approach, referred to as *history-alignment model*, that tackles the treatment bias in the HIV clinical data by introducing a notion of treatment similarity which considers not only information on the current therapy but also detailed information on the treatment history. More specifically, it considers two treatments as similar if they have similar treatment patterns and their genomic fingerprint in the latent viral population is similar. Our approach trains a separate model for each sample of interest by using all available training samples, each with a specific weight, that reflects the similarity of the corresponding treatment pattern to the treatment pattern of the target sample. In this way we address the different treatment backgrounds of the clinical samples, their differing sample abundances, the latent virus population and the uneven therapy representation in the clinical data sets. In what follows we first describe the problem setting and then provide detailed description of the similarity measure of therapy sequences and the history-alignment model.

2.1 Problem setting

Let $\mathbf{x}$ denote the viral genotype represented as a binary vector indicating the occurrence of a set of resistance-relevant mutations, let $\mathbf{z}$ denote the therapy combination encoded as a binary vector that indicates the individual drugs comprising the current therapy and let $\mathbf{h}$ denote a binary vector representing the drugs administered in all known previous therapies for the specific therapy example. The label y indicates the success (1) or failure (-1) of each therapy sample. Let $D = \{(\mathbf{x}_1, \mathbf{z}_1, \mathbf{h}_1, y_1), \ldots, (\mathbf{x}_m, \mathbf{z}_m, \mathbf{h}_m, y_m)\}$ denote the training set and let $\mathbf{t}$ denote the therapy sample of interest. Let $start(\mathbf{t})$ denote the point of time when the therapy $\mathbf{t}$ was started and $pat(\mathbf{t})$ denote the patient identifier corresponding to therapy sample $\mathbf{t}$. Then:

$$r(\mathbf{t}) = \{\mathbf{z} \mid (start(\mathbf{z}) \leq start(\mathbf{t})) \ and \ (pat(\mathbf{z}) = pat(\mathbf{t}))\}$$

denotes the complete treatment record associated with the therapy sample $\mathbf{t}$ and is referred to as *therapy sequence.* It contains all known therapies administered to $pat(\mathbf{t})$ not later than $start(\mathbf{t})$ ordered by their corresponding starting times, from older to newer and will be referred to as the therapy sequence. We point out that each therapy sequence also contains the current therapy, *i.e.,* the most recent therapy in the therapy sequence $r(\mathbf{t})$ is $\mathbf{t}$. Our goal is to train a model $f(\mathbf{x}, \mathbf{t}, \mathbf{h})$ that correctly predicts the outcome of the target therapy $\mathbf{t}$ for given viral genotypes by utilizing the information from its associated therapy sequence.

2.2 Similarity of therapy sequences

Our main objective when quantifying the similarity of therapy sequences is to consider two therapy sequences similar if they consist of similar drug combinations administered in a similar order and producing similar genomic fingerprints in the latent viral population.

We first quantify the pairwise similarity between different drug combinations and then use it together with the order in which the therapies were administered to compute the overall similarity between two therapy sequences. Since we lack primary data on the latent virus population, the pairwise therapy similarity measure considers the genomic fingerprint the therapies leave in the viral genome as a surrogate. This fingerprint comprises resistance-relevant mutations of the drugs making up the therapy.

For quantifying the pairwise similarities between different therapy combinations we use the *resistance mutations kernel*, which uses the table of resistance-associated mutations of each drug afforded by the International AIDS society (Johnson et al., 2008). The kernel assumes that the similarity between different drug groups is additive. This is a reasonable assumption since drugs belonging to different groups have different targets and/or modes of action and thus can be assumed to act independently (Beerenwinkel et al., 2003). Formally the kernel is defined as follows. Let G denote the set of different drug groups. In our clinical data set we have three drug groups: NRTIs (Nucleoside Reverse Transcriptase Inhibitors), NNRTIs (Non-Nucleoside Reverse Transcriptase Inhibitors) and PIs (Protease Inhibitors). Let $\mathbf{u}_{zg}$ and $\mathbf{u}_{z'g}$ be binary vectors indicating the resistance-relevant mutations for the set of drugs occurring in drug group $g \in G$ of the therapies $\mathbf{z}$ and $\mathbf{z}'$, respectively. The similarity between the drug-g mutations of the two therapies $\mathbf{z}$ and $\mathbf{z}'$ is then calculated by:

$$sim_g(\mathbf{z}, \mathbf{z}') = \frac{\mathbf{u}_{zg}^{\mathsf{T}} \mathbf{u}_{z'g}}{\max(\|\mathbf{u}_{zg}\|^2, \|\mathbf{u}_{z'g}\|^2)},$$

where $\mathbf{x}^{\mathsf{T}}\mathbf{y}$ denotes the scalar product of the vectors $\mathbf{x}$ and $\mathbf{y}$, and $\| \cdot \|$ is the L_2-norm. We derive the similarity $k_m(\mathbf{z}, \mathbf{z}')$ between the therapies $\mathbf{z}$ and $\mathbf{z}'$ by averaging the similarities of their corresponding drug groups:

$$k_m(\mathbf{z}, \mathbf{z}') = \sum_{g \in G} \frac{sim_g(\mathbf{z}, \mathbf{z}')}{|G|}.$$

Since the group similarities $sim_g(\mathbf{z}, \mathbf{z}')$ lie in the interval $[0, 1]$, the values of the resistance mutations kernel are also within $[0, 1]$. Intuitively, the higher the number of common resistance relevant mutations associated with the corresponding sets of drugs making up the two therapies of interest, the higher their similarities. In this way the therapy similarity also accounts for the similarity of the genomic fingerprints of the potential latent virus populations of the compared therapies. Furthermore, our kernel represents drugs in terms of their mutation profile and, by doing so, allows for high group similarity for non-identical drugs that have very similar resistance mutation profiles. In this way we take the high levels of cross resistance within the same drug classes into account.

Once we have determined the pairwise similarities of different drug combinations, we will use them to quantify the pairwise similarities between complete therapy sequences. We need a similarity score that accounts for both the similarity of the different therapies comprising the therapy sequences and the order in which they were administered. Thus we can adapt the score commonly used for assessing the quality of an alignment of protein or nucleic acid sequences. In what follows we give the details of this adaptation.

Let $X = [x_1, \ldots, x_{|X|}]$ and $Y = [y_1, \ldots, y_{|Y|}]$ be two therapy sequences defined over a finite alphabet Σ with lengths $|X|$ and $|Y|$, respectively. Let the pair of sequences (X', Y') defined over the alphabet $\{\Sigma \cup \text{"}-\text{"}\}$

that includes the gap character "$-$" denotes their sequence alignment.

Each alignment can be associated with a score that determines its quality:

$$S(X', Y') = \sum_{i=1}^{|X'|} s(x'_i, y'_i),$$

where s is a similarity function that quantifies all pairwise similarities of all letters in the alphabet $\{\Sigma \cup$ "$-$"$\}$. Of course only good alignments with as few gaps as possible are of interest. In this sense an optimal alignment (X^*, Y^*) is the one that maximizes the alignment score S:

$$(X^*, Y^*) = \arg\max_{(X', Y')} S(X', Y').$$

The maximization problem above can be solved with the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970).

The alphabet used for the therapy sequence alignment comprises all distinct drug combinations making up the clinical data set. The mutations kernel determines the pairwise similarities s between its letters. Each therapy sequence ends with the current (most recent) therapy – the one that determines the label of the sample. Therefore, we adapt the sequence alignment such that the rightmost (most recent) therapies (characters) are always matched, i.e. we do not allow for gaps at the right end of an alignment. In this way we also address the problem of uneven representation of the different therapies. We apply linear gap cost penalty. The parameter specifying the gap cost is selected in the model selection procedure. The score of such an optimal alignment quantifies the pairwise similarity of therapy sequences and is referred to as *alignment similarity.* It should also be pointed out that since it is a sum that uses the mutations kernel values, the alignment similarity also reflects the similarity of the accumulated mutations (genomic fingerprints) of the latent virus population of the compared therapy sequences.

2.3 Method

The history-alignment model utilizes the alignment similarity to train a separate model for every sample of interest. The details of this method for a given target sample are summarized in Algorithm 1. The first step utilizes the alignment similarity: the therapy sequence of the target sample $r(\mathbf{t})$ is aligned to the corresponding therapy sequences of all training samples $\{r(\mathbf{z}_i), i = 1, \ldots, m\}$ and the resulting alignment scores $\{S(r(\mathbf{z}_i), r(\mathbf{t})), i = 1, \ldots, m\}$ are the weights for the training samples.

```
Algorithm 1: History-alignment method
Input: Target sample with corresponding current therapy $\mathbf{t}$ and therapy sequence $r(\mathbf{t})$.
  1. Calculate the weights for all training samples
     $\{S(r(\mathbf{z}_i), r(\mathbf{t})), i = 1, \ldots, m\}$.
  2. Apply linear rescaling to normalize the
     alignment similarity weights to the range of $[0, 1]$:
     $S(r(\mathbf{z}_i), r(\mathbf{t})) = \frac{S(r(\mathbf{z}_i), r(\mathbf{t})) - \min_i S(r(\mathbf{z}_i), r(\mathbf{t}))}{\max_i S(r(\mathbf{z}_i), r(\mathbf{t})) - \min_i S(r(\mathbf{z}_i), r(\mathbf{t}))}$.
  3. Use the weights $\{S(r(\mathbf{z}_i), r(\mathbf{t})), i = 1, \ldots, m\}$ to
     estimate the final model for the target sample -
     minimize weighted loss on training data.
```

Once the sample weights are available we can proceed to step two and train the final model that predicts the therapy response for the sample of interest. For this purpose we use regularized logistic regression model (Evgeniou et al., 2000) that minimizes the loss with respect to $\mathbf{w}_t$ over the weighted training samples:

$$\frac{1}{|D|} \sum_{D} S(r(\mathbf{z}_i), r(\mathbf{t}))^{\gamma} \cdot \ell(f(\mathbf{x}_i, \mathbf{z}_i, \mathbf{h}_i, \mathbf{w}_t), y_i) + \sigma \mathbf{w}_t^T \mathbf{w}_t$$

where σ is the regularization parameter, γ is the smoothing parameter and $\mathbf{w}_t$ is the model parameter. In the minimization above we use all available training samples, from therapy-naïve to heavily pretreated, to produce a separate model for each sample of interest or, if we have a specific test set, for each test sample. Intuitively, the history-alignment approach estimates a model tailored towards the sample of interest such that it up-weights those samples that are relevant for the target sample and down-weights the remaining samples. In this manner the method accounts for the various treatment backgrounds associated with the samples making up the clinical data sets, the different abundances of the levels of therapy experience, the latent virus population and the sparse therapy representation. Note also that using the alignment similarity kernel which allows for gaps enables our method to utilize information from samples with incomplete treatment histories.

As an important aspect in every biomedical application interpretability should be one of the properties of our prediction models. We thus use linear logistic regression and the loss function in the formula above is given by:

$$\ell(f(\mathbf{x}, \mathbf{z}, \mathbf{h}, \mathbf{w}_t), y) = \ln(1 + \exp(-y \mathbf{w}_t^T [\mathbf{x}, \mathbf{z}, \mathbf{h}])).$$

Our approach of training a separate model for each target sample demands an efficient method for minimizing the loss function. The choice for linear models and the sparse input feature space, provided by the binary input features, offer the possibility to use the trust region Newton method for training linear logistic regression (Lin et al., 2008). In this way we ensure real-time model fitting (in the range of few milliseconds) and time-efficient model selection. Section 3 of the Supplementary material provides more detailed information on the running time of the history-alignment models.

3 Experiments and results

3.1 Validation setting

Data set. The data source for our models is the EuResist clinical database that contains information on 93014 antiretroviral therapies administered to 18325 HIV (subtype B) patients from several countries in the period from 1988 to 2008. The information employed by our models includes the consensus sequences of the predominant viral strains in the patients' blood, the individual drugs that comprise a therapy, the virus load measurements (copies of viral RNA per *ml* blood plasma, *cp*/*ml*) at different time points during therapy, and all available (known) therapies administered to each patient before some specific point of time. We point out that the clinical data do not necessarily have the complete information on all administered HIV therapies for all patients. Furthermore, the information on whether all administered therapies for a given patient is available or not is also missing. Therefore, the statistical methods utilize only the available information. The viral sequence assigned to each therapy sample is obtained shortly before the respective therapy was started (up to 90 days before). The response to a given therapy is quantified with a label (success or failure) based on the virus load values measured during its course. The label assignment is identical to the one described in Bogojeska et al. (2010). The information on the viral genotype is given in terms of a binary vector indicating the presence (1) or absence (0) of a set of predefined resistance-relevant mutations derived from the list given in Johnson et al. (2008). The currently administered therapy is also encoded by a binary vector that indicates the presence or absence of all drugs appearing in the data set. The set of drugs administered in all available therapies preceding the current therapy is represented in the same manner. Finally, our training set comprises all samples providing a viral sequence and a label; it includes 6537 labeled therapy samples from 690 distinct therapy combinations.

Time-oriented validation scenario. The trends of treating HIV patients change over time as a result of the gathered practical experience with the drugs and the introduction of new antiretroviral drugs. As in Bickel et al. (2008); Bogojeska et al. (2010), our evaluation scenario accounts for this phenomenon by using a time-oriented split when selecting the training and the test set. In this way, our models are trained on the data from the more distant past, while their performance is measured on the data from the more recent past. This scenario, referred to as time-oriented scenario, is more realistic than other scenarios since it captures how a given model would perform on the recent trends of combining the drugs. For our clinical data set we realize it as follows. First, we order all available training samples by the starting dates of their corresponding therapies. We then make a time-oriented split by selecting the most recent 20% of the samples as the test set and the rest as the training set. For the model selection we split the training set further in a similar manner. We use the most recent 25% of the training set for selecting the best model parameters and refer to this set as the tuning set. Figure 1 in the Supplementary material depicts the different treatment trends in the training, tuning and test sets, defined as explained in the text above and thereby illustrates how treatment trends change over time. One can observe that, unlike the treatment trends in the training set, the treatment trends in the tuning set closely resemble those in the test set. This justifies the choice of the tuning set.

The therapy samples gathered in the HIV clinical data sets are associated with patients whose treatment histories differ in length: while some patients receive their first antiretroviral treatment, others are heavily pretreated. Moreover, these different sample groups, from treatment-naïve to heavily pretreated, are represented with different abundances in the HIV clinical data. Figure 2 in the Supplementary material depicts a histogram of the frequencies of the previously mentioned sample groups in the training data set: the number of samples stemming from patients in early stages of HIV treatment is much higher than the number of samples from therapy-experienced patients (with more than five or more than ten previously administered therapies). The numbers are based on the therapy-history data in the data set. We should also point out that most of the therapy sequences associated with patients in the mid or late stages of HIV treatment are unique, *i.e.,* the representation of specific longer therapy sequences in the clinical data sets is very sparse.

In our computational experiments we want to assess the predictive power of the models in dependence on the level of therapy experience. We therefore group the therapy samples in the test set into different bins based on the number of therapies administered prior to the therapy of interest – the current therapy. Note that for some patients some therapy information might be missing. Thus, with the sample binning we make sure that the samples in the treatment-experienced bin (denoted by > 5) originate from patients that had at least five previous therapies. Another important property of our approach is its ability to address the uneven and sparse representation of the different therapies as depicted in Figure 3 in the Supplementary material. This property arises from the definition of similarities of therapy sequences where the current therapies are always matched. In order to consider the uneven representation of the different therapies when assessing the performance of our models we adopt the validation scenario from Bogojeska et al. (2010): the therapies in the test set are grouped based on the number of samples they have in the training set, and then the model performance on each of the groups is measured. We thereby assess the performance of the models for the rare and the abundant therapies, separately. Note that due to the lack of data and practical experience for the rare HIV combination therapies, predicting their efficiency is more challenging compared to estimating the efficiency of the frequent therapies. Some details on each of the bins for both groupings are given in Table 1 and Table 2 in the Supplementary material. We assess the quality of a given target model by reporting its performance for each of the bins.

In order to be able to assist the selection of a potential combination therapy for HIV patients our method should provide a good ranking based on the probability of therapy success. For this reason, we carry out the model selection based on AUC (Area Under the ROC Curve) results and use AUC to assess the model performance. The standard errors of the AUC values and the significance of the difference of two AUCs used for the pairwise method comparison are estimated as described in Hanley and McNeil (1983).

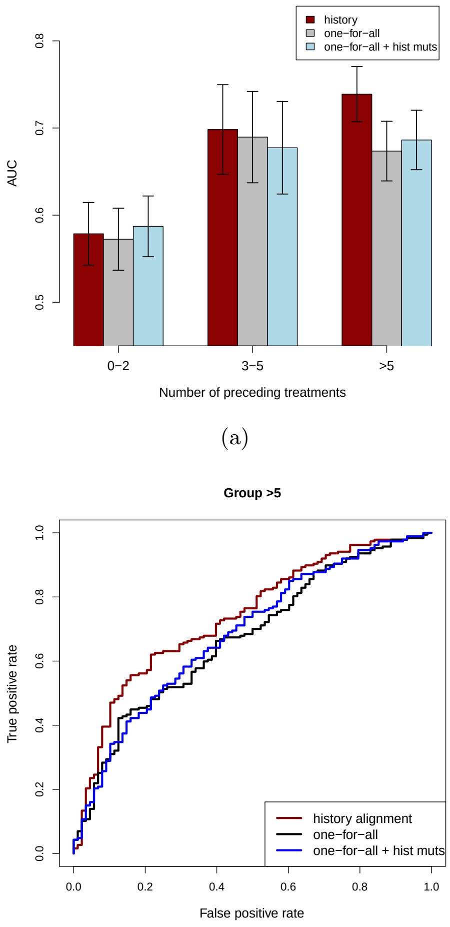
Reference methods. In our computational experiments we compare the results of our history-alignment approach, denoted as *history-alignment validation scenario*, to those of the *one-for-all validation scenario* and the *one-for-all + hist mutations validation scenario*, which are used as reference methods. The *one-for-all* reference method mimics the most common approach in the field where a single linear logistic regression model is trained on all available therapy samples in the data set. The information on the individual drugs comprising the target (most recent) therapy and the drugs administered in all its available preceding therapies are encoded in a binary vector and supplied as input features. We should also point out that removing the similarity score weights from the history-similarity approach yields the *one-for-all* method. The *one-for-all + hist mutations* approach is a modified version of *one-for-all* approach where the drugs from the drug indicator representation of the treatment history are replaced with their respective cumulative resistance-mutation profiles. In this way the accumulated mutations of the latent virus population are encoded in the input feature space.

When assessing the ability of our history-alignment model to address the uneven representation of the different therapies in the clinical data sets we also consider the *therapy-specific model* as a second reference method. It represents the approaches that deal with the uneven, sparse therapy representation by training a separate model for each combination therapy by using not only the samples from the target therapy but also the available samples from similar therapies with appropriate sample weights. It implements the drugs kernel therapy similarity model as described in Bogojeska et al. (2010) on the input feature space defined in the previous section of this paper.

3.2 Experimental results

In what follows, we first present the results of the validation experiments of the time-oriented validation scenario stratified for the length of treatment history, followed by the results stratified for the abundance of the different therapies.

The experimental results for the history-alignment method and the two *one-for-all* reference methods stratified for the length of treatment history are summarized in Figure 1. For samples with a small number of previously administered therapies (less than six), i.e. with short treatment histories, all considered models have comparable performance. The low AUC values of all methods for the group of samples with very short history lengths (less than three) are to be expected. Based on the information available in our clinical data this group comprises samples from therapy-naïve patients ($\sim 75\%$) and samples from patients who had only one or two previous HIV therapies. Therefore, most of them are successful – the success rate is 89%. The main reason for ineffectiveness of initial therapies is lack of adherence. An additional reason for observing failing therapies in the bin of samples with short treatment histories is the wrong assignment of treatment history lengths due to the incomplete patient histories in the database. All these issues may be causes for the low AUCs for the samples with short treatment history. One should also point out that there are specific guidelines for both treating therapy-naïve patients with first-line therapy and administering the first couple of follow-up therapies, which normally are successfully applied. This is also reflected in the high success rate in our clinical data for this group of therapy samples (see Table 1 in the Supplementary material). Thus assistance is mainly necessary for therapy-experienced patients. According to the paired difference test described in Hanley and McNeil (1983) the history-alignment model that incorporates knowledge on the specific therapies comprising the treatment history, their latent virus population and the order in which they were applied significantly improves the performance for the test samples stemming from patients with longer treatment histories (> 5) over the two reference models with *p-value*$=$ 0.001 for the *one-for-all* and *p-value*$=$ 0.005 for the *one-for-all + hist mutations* model. Figure 1 (b) depicts the ROC curves for this group.



(a)

(b)

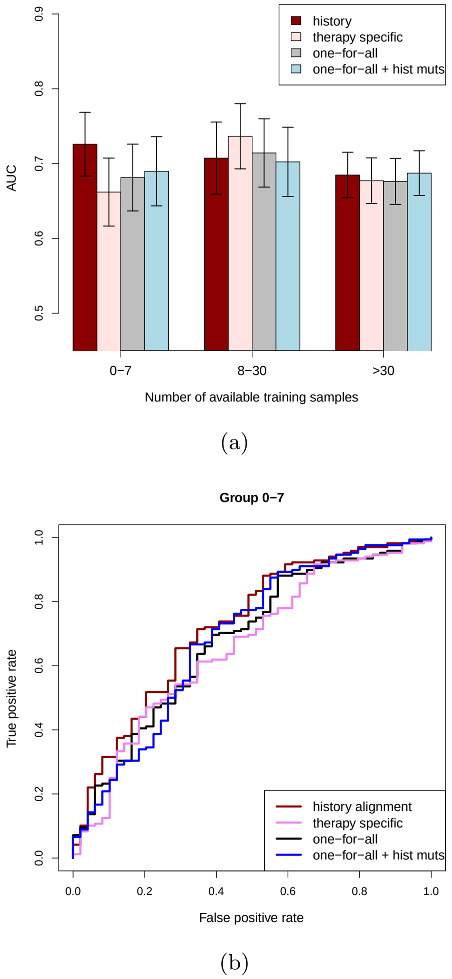
Figure 1: Experimental results stratified for the length of treatment history. (a) Barplot representing the AUC values with their corresponding standard errors for the history-alignment approach and the reference (one-for-all, one-for-all + hist mutations) models. The test samples are grouped by their corresponding number of available previously administered therapies – length of treatment history; and (b) ROC curves displaying the performance of all methods for the group of test samples with more than five previously administered therapies (> 5).

The experimental results stratified for the abundance of the therapies are summarized in Figure 2 (a). As can be observed, the history-alignment method achieves better results than the three reference methods for the test therapies with $0-7$ available training samples. According to the paired difference test described in Hanley and McNeil (1983), the improvement is significant with estimated *p-value*$=$ 0.018 for the *one-for-all*, *p-value*$=$ 0.050 for the *one-for-all + hist mutations*, and *p-value*$=$ 0.008 for the *therapy-specific* model. Considering the test therapies with $8 - 30$ and more than 30 training samples all considered approaches deliver comparable results with no significant differences. The relevant ROC curves for the rare test therapies are shown in Figure 2 (b).

4 Discussion

This paper presents the history-alignment learning approach for predicting the outcome of combination therapies that trains individual model for each target sample. Each of these models weights different training samples differently: the more similar the respective therapy sequences to the target therapy sequence, the higher their importance for the respective model. The similarity of the therapy sequences is quantified by means of sequence alignment which incorporates information on the resistance-relevant mutations. In this way we account for the bias imposed by the sparse sample representation of the various treatment histories in the clinical data and we extract information on the genomic fingerprint of the latent virus population. According to the experimental results this approach significantly outperforms the reference methods for test therapies associated with treatment-experienced patients (with at least five previous treatments) and exhibits comparable performance for the rest of the test therapies. Considering the available guidelines for choosing the several initial HIV treatments and their high success rates, on the one hand, and the difficulty of choosing successful therapies for heavily pretreated patients, on the other hand, availability of statistical methods that focus on providing high-quality models for treatment-experienced patients is becoming increasingly important. Furthermore, our model also addresses the uneven therapy representation in the clinical data sets and outperforms the reference methods for rare test therapies. This is an important feature because the rare therapies comprise 61% of the different therapies in the test set.

An example of the ability of the history-alignment approach to tackle the bias in the clinical data introduced from the sparse therapy-history representation



(a)

(b)

Figure 2: Experimental results stratified for the abundance of therapies. (a) Barplot representing the AUC values with their corresponding standard errors for the history-alignment approach and the three reference models: the one-for-all, the one-for-all + hist mutations, and the therapy-specific model. The test samples are grouped based on the number of training examples for their corresponding therapy combinations; and (b) ROC curves displaying the performance of all models for the rare therapies (with $0 - 7$ training samples) of the test set.

is illustrated in Figure 3. From the image of the therapy sequence corresponding to the sample of interest (Figure 3 a)) we can observe that the target model predicts the outcome of the therapy *ZDV 3TC SQV TDF RTV LPV* – this is the most recent therapy in the therapy sequence, and the therapy sequence has a length of nine. Furthermore, Figure 3 b) depicts the three most relevant therapy sequences for this specific model. Here the relevance is reflected in the similarity of the training therapy sequences to the target therapy sequence. One can observe that the most recent therapies in these sequences are similar to the most recent target therapy. Moreover, the corresponding training samples originate from pretreated patients. Also the average length of the therapy sequences for the 100 most relevant training samples for the considered model is 11. In this way the target model assigns the highest relevance to the training samples originating from therapy-experienced patients with therapy sequences similar to the target therapy sequence and thereby compensates for the bias caused by the different treatment backgrounds of the training samples and the sparse representation of therapy sequences. Furthermore, the available information on the contribution of each training combination therapy to predicting the outcome of the sample of interest is an important aspect of model interpretability. Such information details the most relevant training therapy sequences for a given target therapy sequence and thus enables access to the argumentative basis of the predictions. An additional contribution to model interpretability is achieved by assessing the relevance of the different input features is presented in Section 2 of the Supplementary material.

a)

ZDV > ZDV DDI > ZDV > ZDV 3TC SQV > DDI D4T RTV LPV > DDI D4T EFV > DDI TDF EFV > ZDV 3TC TDF > ZDV 3TC TDF SQV RTV LPV

b)

ZDV > ZDV DDI > ZDV 3TC > ZDV 3TC IDV > D4T 3TC IDV > D4T ABC EFV > D4T ABC TDF EFV > D4T TDF > D4T 3TC TDF RTV LPV

ZDV > ZDV DDI > D4T 3TC > D4T 3TC SQV > D4T SQV RTV > D4T 3TC SQV > D4T SQV RTV > DDI D4T NFV > D4T 3TC EFV > 3TC TDF EFV > ZDV 3TC RTV APV > 3TC TDF RTV APV

ZDV DDI > ZDV DDC > ZDV 3TC > ZDV 3TC IDV > D4T 3TC IDV > D4T 3TC SQV RTV > DDI D4T 3TC NFV > DDI D4T EFV > DDI D4T NVP > ZDV 3TC ABC > DDI ABC NVP > D4T 3TC RTV LPV

Figure 3: a) target therapy sequence; b) therapy sequences of the three most relevant training therapies for the given target therapy sequence. The therapies comprising each therapy sequence are given from older to newer where the current (latest) therapy is depicted in red and > denotes a treatment change.

To summarize, training an individual model for each sample when predicting outcomes for HIV combination therapies enhances the model interpretability. Additionally, these models incorporate detailed information on the treatment history which contributes information on the genomic fingerprint of the latent virus population, addresses the uneven therapy representation and deals with the various treatment backgrounds of the samples making up the clinical data sets. This results in significant improvement of the predictions for the rare test therapies and the therapy samples associated with patients in the mid or late stages of HIV treatment.

Acknowledgements

We gratefully acknowledge the EuResist EEIG for providing the clinical data. This work was funded by the Cluster of Excellence (Multimodal Computing and Interaction).

References

Altmann, A., Däumer, M., Beerenwinkel, N., Peres, Y. Schülter, E., Büch, A., Rhee, S., Sönnerborg, A., Fessel, W., Shafer, W., Zazzi, M., Kaiser, R., and Lengauer, T. (2009). Predicting response to combination antiretroviral therapy: retrospective validation of geno2pheno-THEO on a large clinical database. *Journal of Infectious Diseases*, 199:999–1006.

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

Bickel, S., Bogojeska, J., Lengauer, T., and Scheffer, T. (2008). Multi-task learning for HIV therapy screening. In *Proceedings of the International Conference on Machine Learning*.

Bogojeska, J., Bickel, S., Altmann, A., and Lengauer, T. (2010). Dealing with sparse data in predicting outcomes of HIV combination therapies. *Bioinformatics*, 26:2085–2092.

Deforche, K., Cozzi-Lepri, A., Thays, K., Clotet, B., Camacho, R., Kjaer, J., Van Laethem, K., Phillips, A., Moreau, Y., Lundgren, J., and Vandamme, A. (2008). Modelled in vivo HIV fitness under drug selective pressure and estimated genetic barrier towards resistance are predictive for virological response. *Antiviral Therapy*, 13:399–407.

Evgeniou, T., Pontil, M., and Poggio, T. (2000). Regularization networks and support vector machines. *Advances in Computational Mathematics*, 13:1–50.

Hanley, J. and McNeil, B. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148:839–843.

Johnson, V., Brun-Vezinet, F., Clotet, B., Günthrad, H., Kuritzkes, D., Pillay, D., Schapiro, J., and Richman, D. (2008). Update of the drug resistance mutations in HIV-1: December 2008. *Topics in HIV Medicine*, 16:138–145.

Larder, B., Wang, D., Revell, A., Montaner, J., Harrigan, R., De Wolf, F., Lange, J., Wegner, S., Ruiz, L., Prez-Elas, M., Emery, S., Gatell, J., DArminio Monforte, A., Torti, C., Zazzi, M., and Lane, C. (2007). The development of artificial neural networks to predict virological response to combination HIV therapy. *Antiviral Therapy*, 12:15–24.

Lin, C., Weng, R., and Keerthi, S. (2008). Trust region Newton method for large-scale logistic regression. *Journal of Machine Learning Research*, 9:627–650.

Needleman, S. and Wunsch, C. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453.

Prosperi, M., Altmann, A., Rosen-Zvi, M., Aharoni, E., Borgulya, G., Bazso, F., Sönnerborg, A., Schülter, E., Struck, D., Ulivi, G., Vandamme, A., Vercauteren, J., and Zazzi, M. (2009). Investigation of expert rule bases, logistic regression, and non-linear machine learning techniques for predicting response to antiretroviral treatment. *Antiviral Therapy*, 14:433–442.

Revell, A., Wang, D., Harrigan, R., Hamers, R., Wensing, A., DeWolf, F., Nelson, M., Geretti, A., and Larder, B. (2010). Modelling response to HIV therapy without a genotype: an argument for viral load monitoring in resource-limited settings. *Journal of Antimicrobial Chemotherapy*, 65:605–607.

Rosen-Zvi, M., Altmann, A., Prosperi, M., Aharoni, E., Neuvirth, H., Sönnerborg, A., Schülter, E., Struck, D., Peres, Y., Incardona, F., Kaiser, R., Zazzi, M., and Lengauer, T. (2008). Selecting anti-HIV therapies based on a variety of genomic and clinical factors. *Proceedings of the ISMB.*

Saigo, H., Altmann, A., Bogojeska, J., Müller, F., Nowozin, S., and Lengauer, T. (2010). Learning from past treatments and their outcome improves prediction of in vivo response to anti-HIV therapy. *Statistical Applications in Genetics and Computational Biology*, 10.

UNAIDS/WHO (2010). Report on the global aids epidemic: 2010.

Wang, D., Larder, B., Revell, A., Harrigan, R., and Montaner, J. (2003). A neural network model using clinical cohort data accurately predicts virological response and identifies regimens with increased probability of success in treatment failures. *Antiviral Therapy*, 8:U99–U99.