BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 17 2010, pages 2085–2092
doi:10.1093/bioinformatics/btq361

Genome analysis

Advance Access publication July 12, 2010

Dealing with sparse data in predicting outcomes of HIV combination therapies

Jasmina Bogojeska[1,*], Steffen Bickel[2], André Altmann[1] and Thomas Lengauer[1]

[1]Max Planck Institute for Informatics, Campus E1 4, 66123, Saarbrücken and [2]Nokia Gate 5, Berlin, Germany

Associate Editor: John Quackenbush

ABSTRACT

Motivation: As there exists no cure or vaccine for the infection with human immunodeficiency virus (HIV), the standard approach to treating HIV patients is to repeatedly administer different combinations of several antiretroviral drugs. Because of the large number of possible drug combinations, manually finding a successful regimen becomes practically impossible. This presents a major challenge for HIV treatment. The application of machine learning methods for predicting virological responses to potential therapies is a possible approach to solving this problem. However, due to evolving trends in treating HIV patients the available clinical datasets have a highly unbalanced representation, which might negatively affect the usefulness of derived statistical models.

Results: This article presents an approach that tackles the problem of predicting virological response to combination therapies by learning a separate logistic regression model for each therapy. The models are fitted by using not only the data from the target therapy but also the information from similar therapies. For this purpose, we introduce and evaluate two different measures of therapy similarity. The models are also able to incorporate phenotypic knowledge on the therapy outcomes through a Gaussian prior. With our approach we balance the uneven therapy representation in the datasets and produce higher quality models for therapies with very few training samples. According to the results from the computational experiments our therapy similarity model performs significantly better than training separate models for each therapy by using solely their examples. Furthermore, the model's performance is as good as an approach that encodes therapy information in the input feature space with the advantage of delivering better results for therapies with very few training samples.

Availability: Code of the efficient logistic regression is available from http://www.mpi-inf.mpg.de/%7Ejasmina/fastLogistic.zip

Contact: jasmina@mpi-inf.mpg.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on March 3, 2010; revised on June 18, 2010; accepted on June 30, 2010

1 INTRODUCTION

The human immunodeficiency virus (HIV) targets the human immune system and leads to the acquired immunodeficiency syndrome (AIDS) that has a high mortality rate. Being a disease that claimed more than 25 million lives since its discovery in 1981, AIDS is highly ranked on the chart of most destructive world epidemics. Currently, there is no vaccine against HIV and there is no way of ridding HIV patients of the virus.

Therapy administration to patients is hampered by the high evolutionary dynamics of the virus which eventually escapes to resistance against any drug therapy. In order to prolong the time for which a therapy is effective, combinations of drugs are administered to patients. According to the WHO (UNAIDS/WHO, 2009) more than 30 million people were infected and approximately 4 million people were receiving antiretroviral therapy at the end of 2008. Since every therapy eventually fails, finding a successful regimen remains a major challenge for HIV treatment. Whenever a therapy regimen fails a physician decides on a new regimen based on the presence of a set of resistance-relevant mutations in the viral strands sampled from the patient's blood serum. The large number of available drugs renders the number of potential therapy combinations too large for manual assessment. Thus, in recent years statistical models for therapy outcome prediction have been developed on the basis of clinical data. These data comprise samples of many different drug combination therapies taken from many patients over a couple of decades. Different therapies are represented in different abundance of samples: while for some therapies many samples exist, for others there are very few. Furthermore, due to the large number of possible combination therapies and the introduction of new antiretroviral agents, for many combinations no samples are available at all. In a medical setting, we strive for every therapy to be effective. Thus, it is important to improve the quality of the prediction on these difficult to assess therapies that have very few training samples.

The work we present in this article is inspired by the uneven therapy representation in the clinical datasets. We propose a method that tackles the problem of predicting the outcome of HIV drug combination therapies as follows. A separate model is trained for each distinct therapy combination by using not only the samples from the target therapy but also the available samples from similar therapies with appropriate sample weights. In order to quantify therapy similarity, we introduce two different similarity measures: the *drugs kernel* and the *groups additivity kernel*. Utilizing knowledge from similar therapies is particulary important for the therapy models that have very few or no examples in the available dataset. Since we learn separate models for each drug combination in order to provide efficient model fitting and model selection, we choose the *trust region Newton method* for training linear logistic regression (Lin *et al.*, 2008) that takes advantage of the sparsity of the data. Furthermore, our method has the facility of integrating prior knowledge on therapy outcomes through a Gaussian prior.

*To whom correspondence should be addressed.

In the article, we show how to integrate phenotypic models that give information on the *in vitro* effectiveness of each drug as prior knowledge.

Being collected over two decades the clinical data cannot be representative for any given time point: the treatment trends change with the introduction of new drugs and the viral sequences evolve over time. We take this phenomenon into account by a time-oriented evaluation which uses the most recent data samples to assess the quality of the models and learns the models on the remaining samples.

The article is structured as follows. After reflecting on related work we give the details of the method, the similarity measures and the prior knowledge in Section 2. In Section 3, we describe the datasets, the experimental setting and present and discuss the results of the computational experiments. Section 4 concludes our article.

1.1 Related work

Machine learning methods, such as artificial neural networks, decision trees, random forests, support vector machines (SVMs) and logistic regression (Altmann *et al.*, 2007, 2009b; Deforche *et al.*, 2008; Larder *et al.*, 2007; Lathrop and Pazzani, 1999; Prosperi *et al.*, 2005, 2009; Rosen-Zvi *et al.*, 2008; Wang *et al.*, 2003), have been used to tackle the problem of predicting virological response to a given therapy combination by fitting a single model for all therapies using both the viral genotype and the drugs in the applied treatment as features. Bickel *et al.* (2008) use a multi-task learning approach that learns a separate model for each combination therapy from all available samples with properly derived sample weights. The weights are computed by matching the distribution of all samples to the distribution of the samples of each individual therapy. This method computes the similarities between therapies and the prediction model in a single integrated procedure. While being statistically sound, the method is also quite compute intensive, as it involves a multi-class logistic regression with as many classes as there are therapies (usually several hundred). A somewhat more heuristic but much more efficient and configurable alternative is to separate the computation of the similarities in a preprocessing step from the learning of the model. This reduces the computation time of the therapy similarities to a few seconds and has the added advantage that the resulting similarities can be reconfigured at the users discretion and then used to train a new model.

The results from Altmann *et al.* (2007, 2009a) have demonstrated that phenotypic information improves the predictive performance of the response to antiretroviral combination therapies. While these papers describe methods that use the predictions of phenotypic models in the form of additional features in the input feature space, our method incorporates the phenotypic models directly via a Gaussian prior. We will show that this is not only an attractive modeling approach, but is also a way to better utilize the additional information from the phenotypic prior knowledge compared with approaches that rely on the predictions of the phenotypic models.

2 METHODS

Our approach to the problem of predicting the outcome of HIV drug combination therapies learns a separate model for each therapy by using the genotypic information from similar therapies. This is done in the fitting procedure by using precomputed weights that up-weigh samples originating from therapies similar to the therapy of interest. In this way the separate models are tuned to focusing on information coming from therapies akin to the target therapy. The mathematical concept of therapy similarity is governed by a specific predefined understanding of what similar therapies are. This article describes and evaluates two different approaches for quantifying pairwise therapy similarity. The therapy-specific models that we introduce also offer the possibility of incorporating phenotypic information on the effectiveness of the individual drugs comprising the therapy of interest. All this sums up to easy-to-interpret linear logistic regression models fitted for each individual therapy with a learning procedure that takes advantage of additional information (similar therapies and phenotypic information) and therefore can deal with therapies that have only few training samples available.

Let $\mathbf{x}$ denote the input features that comprise the genetic information encoded as a binary vector indicating the occurrence of a set of resistance-relevant mutations (Johnson *et al.*, 2008). The drug combinations are denoted by z—each of them is represented by a binary vector that indicates the individual drugs given in the combination. The binary class label y marks each therapy sample with *success* (1) or *failure* (−1). Let $D = \{(\mathbf{x}_1,z_1,y_1),\ldots,(\mathbf{x}_m,z_m,y_m)\}$ denote the training set and t denote the therapy of interest.

We model the problem of predicting the outcome of the therapy t with weighted linear logistic regression using a logarithm of a Gaussian prior (Evgeniou *et al.*, 2000) with mean μ_t and isotropic covariance matrix $\sigma^2\mathbf{I}$ shared by all therapies. The model parameters $\mathbf{w}_t$ for therapy t are obtained by solving the optimization problem given as follows.

Optimization Problem 1. *Over parameters* $\mathbf{w}_t$, *minimize*

$$\frac{1}{|D|}\sum_{(\mathbf{x}_i,z_i,y_i)\in D} k(z_i,t)^{\gamma}\cdot\ell(f(\mathbf{x},\mathbf{w}_t),y)+\frac{(\mu_t-\mathbf{w}_t)^T(\mu_t-\mathbf{w}_t)}{2\sigma^2}. \qquad (1)$$

$k(z_i,t)$ is a function that provides sample-specific weights that quantify the similarity of the therapy of interest t with the therapy z_i from the i-th sample (Section 2.1), and γ is its smoothing parameter. The expression:

$$\ell(f(\mathbf{x},\mathbf{w}_t),y)=\ln(1+\exp(-y\mathbf{w}_t^T\mathbf{x})) \qquad (2)$$

is the loss of linear logistic regression. μ_t is phenotypic prior knowledge on the outcome of therapy t as explained in Section 2.2. We will refer to this model as *therapy similarity model*. The large number of distinct therapies and our approach of training a separate model for each of them demand an efficient method for solving Optimization Problem 1. To do this, we apply a *trust region Newton method* for training logistic regression (Lin *et al.*, 2008) that takes advantage of the sparseness of our feature space. This enables fast model fitting which results in efficient model selection. In what follows, we give a detailed description of the therapy similarity kernels and the prior phenotypic knowledge on the therapy outcomes.

2.1 Therapy similarity kernels

We quantify the pairwise similarity between the different drug combinations with two kernels: the *drugs kernel* and the *groups additivity kernel*. The method also allows for alternative definitions of pairwise therapy similarity that can include different types of additional information, e.g. expert knowledge or information obtained from phenotypic drug resistance tests.

The drugs kernel similarity is based on the number of common drugs that two combination therapies share. Let $\mathbf{u}_z$ and $\mathbf{u}_{z'}$ be binary vectors indicating the distinct drugs comprising the therapies z and z', respectively. The similarity $k_d(z,z')$ between the combination therapies z and z' is given by:

$$k_d(z,z')=\frac{(\mathbf{u}_z^T\mathbf{u}_{z'})}{\max(\mathbf{u}_z^T\mathbf{1},\mathbf{u}_{z'}^T\mathbf{1})}. \qquad (3)$$

where $\mathbf{x}^T\mathbf{y}$ is the scalar product of the vectors $\mathbf{x}$ and $\mathbf{y}$. According to this kernel the more drugs therapies z and z' have in common the higher their similarity. Its values are in the $[0,1]$ interval.

The groups additivity 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.*, 2003b). Let G denote the set of different drug groups. In our dataset, 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 set of drugs occurring in drug group $g \in G$ of the therapies z and z', respectively. The similarity between the group-g drugs of the two therapies z and z' is then calculated by:

$$\mathrm{sim}_g(z,z') = \frac{\mathbf{u}_{zg}^T \mathbf{u}_{z'g}}{\max(\mathbf{u}_{zg}^T \mathbf{1}, \mathbf{u}_{z'g}^T \mathbf{1})}. \tag{4}$$

Intuitively, the larger the number of common drugs making up a specific drug group of the two therapies of interest, the higher their group similarity.

We derive the similarity $k_a(z,z')$ between the therapies z and z' by averaging the similarities of their corresponding drug groups:

$$k_a(z,z') = \frac{\sum_{g \in G}(\mathrm{sim}_g(z,z'))}{|G|}. \tag{5}$$

Since the group similarities $\mathrm{sim}_g(z,z')$ lie in the interval $[0,1]$, $k_a(z,z')$ also has values within $[0,1]$.

Note that computing the kernel values for all therapies takes only several seconds. In the article, we only focus on the results obtained with the drugs kernel. The corresponding results obtained with the groups additivity kernel are very similar. We present them in Section 4 of the Supplementary Material.

2.2 Phenotypic prior knowledge on therapy outcome

Phenotypic resistance tests are laboratory experiments that produce continuous values, referred to as resistance factors, that measure the effectiveness of individual drugs against a given viral strain. Genotype–phenotype pairs (GPP) are sequences with the associated resistance factor measured in a phenotypic test using the virus defined by the sequence. The models trained on GPP data aim at predicting the resistance factors for each single drug for unknown genotypes. We will refer to these models as *phenotypic models*. One such model is described in Beerenwinkel *et al.* (2003a). Furthermore, this article reveals the bimodal nature of the distribution of the resistance factors common to all drugs. Such a distribution can be approximated with a two-component Gaussian mixture model. We derive drug-specific resistance cutoffs from the intersection of the two mixture components. The cutoffs can then be used to infer the effectiveness of each drug against a given genotype: a drug is *effective* when its resistance value is smaller than its resistance cutoff, otherwise it is *ineffective*.

Unlike other approaches that add predictions facilitated by phenotypic models as additional features in their input feature space (Altmann *et al.*, 2007, 2009a), we incorporate the models themself via a logarithm of a Gaussian prior on the model parameters for each therapy combination. We choose a Gaussian prior for two reasons. First, it is easy to integrate into regularized logistic regression as can be seen from Optimization Problem 1. Second, the *trust region Newton method* (Lin *et al.*, 2008) which affords an efficient solution of the problem requires this prior. For a given therapy t, we do this as follows:

- consider the subset of the GPP data comprising the virus genotypes that have an associated resistance factor for all individual drugs that appear in the clinical data;
- label each virus sequence based on the effectiveness of the best performing drug from the drugs comprising therapy t: *success* if effective, *failure* if ineffective;
- fit a logistic regression model to the labeled data with model parameters (weights) p_t; and
- use the model parameters p_t from the fitted logistic regression as means $\mu_t = p_t$ for the Gaussian prior ($N(\mu_t, \sigma^2\mathbf{I})$) on model parameters $\mathbf{w}_t$ in Optimization Problem 1.

We repeat the procedure described above for every individual combination therapy. Instead of using the effectiveness of the best performing drug to label a therapy combination, one can also use other quantities such as, for example, the average of the effectiveness of the drugs comprising the therapy of interest (for more details on this see Section 5 in the Supplementary Material). We will show that with the procedure described above one can better utilize the phenotypic knowledge compared with using the prediction of the phenotypic model as additional input feature in a single logistic regression model fit for all therapies.

GPP data provide knowledge on the efficiency of individual drugs against HIV that is especially valuable for assessing therapies for which not many clinical samples are available. For example, while there can be a considerable amount of available GPP data for newly introduced drugs, clinical data for therapies that include these newly introduced drugs may be very sparse. This is the case simply because after the approval of a new antiretroviral agent it is spared as an option for highly treatment-experienced patients.

3 RESULTS AND DISCUSSION

We developed a model that targets the problem of predicting the outcome of HIV combination therapies from the genotype of the most abundant virus strain in the patient's blood serum. In the next sections we describe the datasets, the details of the computational study that assesses the quality of our model and its results.

3.1 Datasets

We use the EuResist database (Rosen-Zvi *et al.*, 2008) that incorporates information of 88 469 antiretroviral therapies administered to 18 255 HIV (subtype B) patients from several countries in the period from 1988 through 2008. This information includes the combination of drugs given to the patients, the sequence of the predominant viral variant and virus load measurements at different time points during a therapy.

The dataset we use to train our models is derived from the EuResist database as follows.

Each sample of the data corresponds to a therapy given to a patient and contains information describing the viral sequence obtained shortly before the respective therapy was administered. The virus genotype is represented by a binary vector indicating the occurrence or absence of a set of predefined resistance-relevant mutations based on the mutation list reported in Johnson *et al.* (2008). Each therapy is denoted by a binary vector indicating the presence or absence of the individual drugs comprising it. We label each therapy sample as success or failure based on the virus load (copies of viral RNA per ml blood plasma, cp/ml) measured in the course of the therapy. If the virus load drops bellow 400 cp/ml in the period from 21 days after the start of the therapy to the end of the therapy we label it with *success* (1); otherwise with a *failure* (−1). Figure 1 shows a schema of the labeling procedure. In this way we create a labeled dataset that includes 6336 therapy samples with 638 distinct therapy combinations.

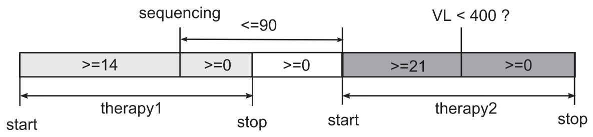

Fig. 1. Assigning a label and a viral sequence to *therapy2*, where *therapy1* and *therapy2* are two consecutive therapy administered to a patient.

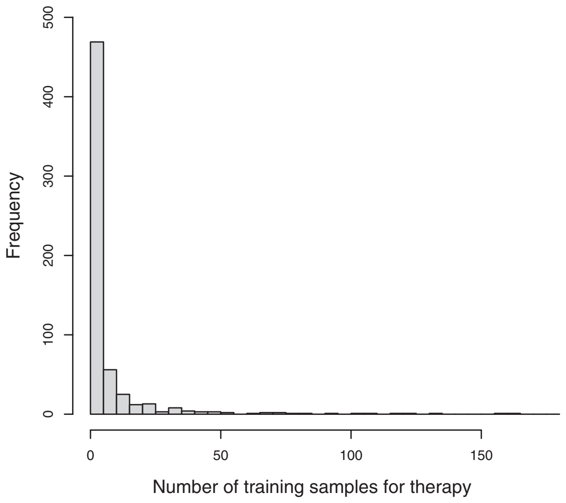

Fig. 2. Histogram that groups the HIV combination therapies based on the number of samples present in the clinical dataset.

As we mentioned above the different drug combinations are not evenly represented in the clinical datasets: while some therapies are present with many samples, others have only few. The histogram in Figure 2, shows that for most therapies that make up our dataset we have fewer than 50 samples available. Additionally, almost 500 antiretroviral therapies are represented with fewer than five examples.

We take the GPP data from the Arevir database (Roomp *et al.*, 2006). As described in Section 2.2, the GPP data comprise drug resistance factors that characterize the effectiveness of individual drugs on specific viral variants. We consider only the virus sequences that have an associated resistance factor value in the database for all individual drugs that appear in our clinical dataset. This is necessary because the construction of the prior knowledge on the therapy outcome for a specific therapy requires resistance factors for all drugs comprising the respective therapy. We label the GPP viral sequences as *success* or *failure* with respect to a given therapy of interest based on the resistance factor of the most effective drug. After the filtering we end up with 200 samples. For representing the genotypic information describing the virus, we use the same binary encoding as for the clinical data.

3.2 Evaluation setting and reference methods

The HIV sequences evolve under drug pressure over time. The treatment trends also change with time as a result of both the introduction of new drug agents and the practical experience with the drugs acquired over time. Being collected over two decades the clinical data cannot be representative for any given time point. Therefore, we work with an evaluation setting that takes the evolution of both the viral sequences and the treatment trends into account. This is done by using a time-oriented split when choosing the training and the test set. First, we order all the therapy samples by their starting dates. We then make the split by selecting the most recent 20% of the samples (from June 2006 to January 2008) as a test set and using the rest as a training set. We also do the model selection by splitting the training set in a similar manner. We use the most recent 25% of the training set for the model selection process. In this way our models learn from data seen in the more distant past and their performance is measured on unseen data from the more recent past. Such an evaluation scenario reflects the actual situation in HIV therapy outcome prediction: the outcomes of unseen future therapy samples are estimated by utilizing the information from past treatment records. We refer to this setting as *time-oriented scenario*. More details supporting such an evaluation scenario and plots depicting the changing therapy trends over time can be found in Section 1 of the Supplementary Material.

In the two following subsections, we report the results of computational experiments conducted with our method and several other reference methods used for comparison.

One reference method consists of training a separate logistic regression model for each combination therapy by using only the samples from the target therapy. If we had enough data for each therapy combination this would be the best choice as the separate model captures the characteristics specific to the corresponding therapy and therefore can also make the best predictions for it. In our case we fit the separate models by using the data available for each individual therapy in the clinical database. The therapies with no samples available are either randomly classified as *success* or *failure* both with equal probability of 50%, or the phenotypic model (Section 2.2) is used to assign their labels. For therapies represented only with successful (failing) examples, we assign success probabilities of 1 (0). We will refer to this approach as *partitioned evaluation scenario*.

The second reference method, referred to as *transfer evaluation scenario*, implements the distribution matching approach by Bickel *et al.* (2008). In order to provide a fair comparison, we use a linear instead of a nonlinear logistic regression for the separate models learned for each therapy. The sample-specific weights for each therapy are obtained as described in Bickel *et al.* (2008) by using a nonlinear multi-class logistic regression.

The last reference method, referred to as *one-for-all evaluation scenario*, fits a single logistic regression model to the viral genotypes of all therapies. The information about the drugs comprising the corresponding therapies is added to the input feature space. This is the most common approach in the field.

3.3 Experiments and discussion

In this subsection, we present and discuss the results of the validation experiments. The therapy similarity models and each of the different reference models, as explained in the previous subsection, are trained on the EuResist clinical dataset by using the *time-oriented scenario*. The goal of each model is to predict the outcomes of the combination therapies for the most recent therapy samples in the dataset. We are primarily interested in the accuracy as a measure of the quality of the different models. However, sometimes one is not only interested in the absolute results, but also in the quality of the ranking of the therapies based on their success probability. This is especially important when choosing a future therapy for a patient. Therefore, we also carry out model selection based on AUC [area under the receiver operating characteristic (ROC) curve] performance and report AUC results. Table 1 summarizes the classification accuracies (ACCs) and the AUCs for the different methods: *drugs kernel* denotes our therapy similarity models with the drugs kernel therapy similarity measure as sample-specific weights; *partitioned* denotes the reference models fitted for each distinct therapy by

Table 1. Classification accuracies (ACCs) and AUCs with their corresponding SEs for our therapy similarity model (drugs kernel) and the three reference models (one-for-all, partitioned and transfer) that predict the outcomes of drug combination therapies

Method	Drugs kernel		One-for-all		Partitioned		Transfer
	No prior	With prior	No prior	With prior	No prior	With prior	
ACC ± SE	0.850±0.010	0.848±0.010	0.850±0.010	0.856±0.010	0.830±0.011	0.822±0.011	0.848±0.010
AUC ± SE	0.703±0.022	0.695±0.023	0.700±0.023	0.703±0.022	0.624±0.025	0.608±0.026	0.625±0.024

using only the samples belonging to the target therapy; the label *with prior (no prior)* can refer to any of the previously described models with (without) additional phenotypic knowledge encoded as a Gaussian prior; *one-for-all* refers to the reference method that fits a single logistic regression model to the samples from all combination therapies where the therapy information is encoded in the input feature space; the label *with prior* for the *one-for-all* approach refers to encoding the prediction of the phenotypic model described in Section 2.2 as an additional feature; *transfer* refers to the linear version of the transfer model by Bickel *et al.* (2008). As we mentioned before, efficient model fitting is important for the approaches that fit a separate model for each combination therapy. By using the *trust region Newton method* for training logistic regression (Lin *et al.*, 2008), we fit a single model in a split of a second. For example, fitting 154 separate models for the different therapies in the test set takes only 13 s on a normal desktop computer. In what follows, we will first discuss the performance of the different models with respect to the accuracy and then continue with a similar discussion for the AUC. Resampling techniques (e.g. bootstrap) to estimate the standard errors (SEs) of these measurements are not readily applicable in the *time-oriented scenario* in which the data samples are ordered by the starting times of their corresponding therapies. Therefore, we resort to calculating SEs of the accuracies based on the paired t-test detailed in Section 2 of the Supplementary Material and SEs of the AUCs are computed as described in DeLong *et al.* (1988).

As shown in Table 1, our approach (*drugs kernel*) of utilizing information of similar therapies, the *transfer model* and the model from the *one-for-all scenario* perform significantly better than training separate models by solely using the samples from the target therapies (*partitioned scenario*). We assess the significance of the accuracy with a paired t-test where we observe p-values≤ 0.01 for all pairwise comparisons between the models from the *partitioned scenario* and the other models. The performance of the *partitioned models* is worse because for many therapies there are only few samples in the dataset (Fig. 2). The therapy similarity kernels in our approach and the sample-specific weights in the *transfer scenario* compensate for this lack of data by using the samples from the similar therapies for making the predictions.

In order to further investigate the performance of the models, we take the uneven representation of the different therapies into account. We do this by grouping the therapies in the test set based on the number of samples they have in the training set and then computing the accuracies of all the groups. The details on both the test used for pairwise method comparisons and the computations of the SE bars are given in Section 2 of the Supplementary Material. The results are depicted in Figure 3. All models, including the derived in the *partitioned scenario*, deliver very good predictions for therapies for which there is a reasonable number of samples ($\geq$15) available in the training dataset. In such cases the models have enough samples to capture the characteristics of each different combination therapy.

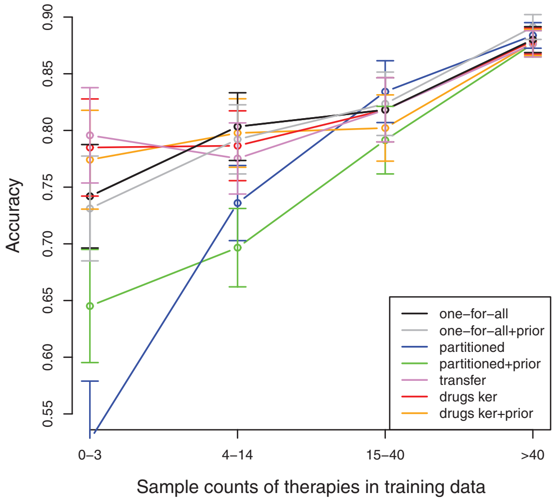

Fig. 3. Classification accuracy of the different models over groups of test samples grouped by the number of training examples for the therapy combinations. Error bars indicate the SEs of each model.

As can be anticipated the models derived in the *partitioned scenario* achieve much worse performance compared with the other models for the therapies that have fewer samples in the training set $(0-14)$. Our model (*drugs kernel*) and the *transfer model* that utilize therapy similarity in the learning process significantly outperform the *one-for-all model* for the therapies that have only very few $(0-3)$ samples in the training set. This group is small. It comprises only about 8% of the test set. However, it contains 61 of the 154 different drug combination in the test set which is around 40% of all drug combinations occurring among the test therapies. We verified the significance of the improvements with paired t-test: p-value$=0.04$ for the *drugs kernel* and p-value$=0.09$ for the *transfer model.*

For the group of therapies with 4–14 samples, the *therapy similarity model* and the *transfer model* perform slightly worse than the *one-for-all model*. However, according to paired t-test this difference is only significant (p-value$=0.03$) for the *transfer model*. The p-value for the comparison with the *drugs kernel* model is 0.31.

The results of the *partitioned models* for therapies with only very few (0–3) training samples significantly improve by incorporating the phenotypic prior knowledge: the paired t-test shows significant

Table 2. AUCs pertaining to the therapy similarity models (drugs kernel) with their corresponding SEs for two groups of test therapies: with 0–20 and >20 available training samples

Method	Drugs kernel		One-for-all		Transfer
	No prior	With prior	No prior	With prior	
0–20 (SE)	0.659 (0.041)	0.690 (0.039)	0.641 (0.041)	0.642 (0.041)	0.608 (0.043)
>20 (SE)	0.694 (0.028)	0.681 (0.028)	0.697 (0.029)	0.700 (0.029)	0.637 (0.029)

improvement of the accuracy (p-value$=0.006$). However, including this prior knowledge makes the results worse for the therapies with 14–40 training samples. It also does not improve the accuracy results for the *therapy similarity* or the *one-for-all models*. The reason for this may be that with respect to accuracy the samples from the similar therapies in the clinical data probably carry at least as much relevant information as the prior itself. It is encouraging to see that the added phenotypic information does not deteriorate the similarity models, either.

Section 3 of the Supplementary Material depicts and discusses the accuracy results of several different groupings of the group of test therapies with few training samples. We give this additional detail as this is the group of therapies that is most in the focus of our investigation. The results in the Supplementary Material confirm the advantage of our *therapy similarity model* regarding the accuracy performance for the therapies with very few training samples.

Inspecting the overall AUC performance in Table 1 we can observe that all models except for the *transfer model* and the *partitioned models* have comparable performance. The reason for the poor performance of the *partitioned models* is that they very often assign probabilities of 1 or 0 for therapies with very few training samples because they will be very often either all successful or all failing. That is why we do not look at the AUC performance of the *partitioned models* into more detail.

We take the uneven therapy representation into account by splitting the data into two groups: one with 0–20 training samples and another with more than 20 samples. The AUC results for the different methods are shown in Table 2. In this case our *therapy similarity model* with phenotypic prior knowledge significantly improves the AUC results compared with those of the *transfer model* (p-value$=0.002$) for the therapies with 0–20 available training samples. This group of therapies comprises about 25% of the test data. Integrating the prediction of the phenotypic model as additional input feature in the *one-for-all model* does not boost its AUC performance for the therapies with fewer training samples. Our *therapy similarity model* that incorporates the model parameters of the phenotypic model via a Gaussian prior outperforms the *one-for-all model* with prior knowledge for the therapies with 0–20 available training samples (p-value$=0.04$). This demonstrates the ability of our approach to better integrate the additional information provided with the phenotypic model. The differences between the AUCs of the *one-for-all model* (with and without prior) and the *therapy similarity model* (with and without prior) for the therapies with >20 available training samples were not significant (all p-values>0.149). We compute the significance of the difference of the AUCs and its SE as described in DeLong *et al.* (1988).

To summarize, with respect to measured accuracies our approach that trains separate models for each therapy by utilizing the data from similar therapies has its prime advantage for therapies with few (<4) training samples. The *transfer model* also achieves (slightly less) significant results for this group of therapies. However, it has lower accuracy for the group of therapies with 4–14 available training samples and has a significantly worse AUC performance. Another disadvantage of this method is the compute-intensive calculation of the sample-specific weights: fitting a single multi-class logistic regression model for a large number of classes (in our case several hundreds) and training samples takes several days; in order to update such model with new data one has to fit many multi-class logistic regression models since they also have a tuning parameter.

The phenotypic prior knowledge added to the *therapy similarity model* significantly improves the AUC performance for therapies with 0–20 available training samples. Here, the added phenotypic information is essential in bringing the performance of the model to a level, which is also achieved for the abundant therapies.

Our method also allows for alternative definitions of therapy similarity. For example, one can derive a similarity measure that includes phenotypic knowledge. Another advantage of our approach is that one can analyze the profile of similar therapies chosen by the model of a given target therapy that harbors relevant information. An interesting example is illustrated in Figure 4. This figure shows a heatmap that depicts the magnitude of similarity of each of six randomly chosen test therapies from the group of therapies with no available training samples with all therapies from the training set with similarity values greater than 0.5 according to the *drugs kernel*. In this way we can easily see how much each of the drug combinations contributed to predicting the outcome for the test treatments with no training samples available. As an example, let us consider the model for the combination therapy *ZDV 3TC ABC TDF RTV ATV*. From the heatmap, we can easily see that the model assigns the highest weights to the examples from the therapies: *ZDV 3TC TDF RTV LPV*, *3TC ABC TDF RTV ATV*, *ZDV 3TC ABC TDF SQV RTV ATV* and *ZDV 3TC ABC RTV ATV*.

Furthermore, we can assess how different mutations in the viral genome contribute to predicting the outcome of a given target therapy. Since we have a separate model for the target therapy, we can simply do this by quantifying the importance of the model features. One way to do this is to calculate z-scores for each of the model coefficients, which corresponds to a test of the null hypothesis that the coefficient of interest is zero, while all the others are not. The coefficients with the highest z-scores are the most significant ones. The table of the z-scores for the coefficients of the model for the combination therapy *ZDV 3TC ABC TDF RTV ATV* are given in Section 6 of the Supplementary Material. According to them, for this therapy the three most important positions in the protease sequence are 54, 90 and 10; and in the reverse transcriptase sequence the most relevant positions are 215, 210 and 41.

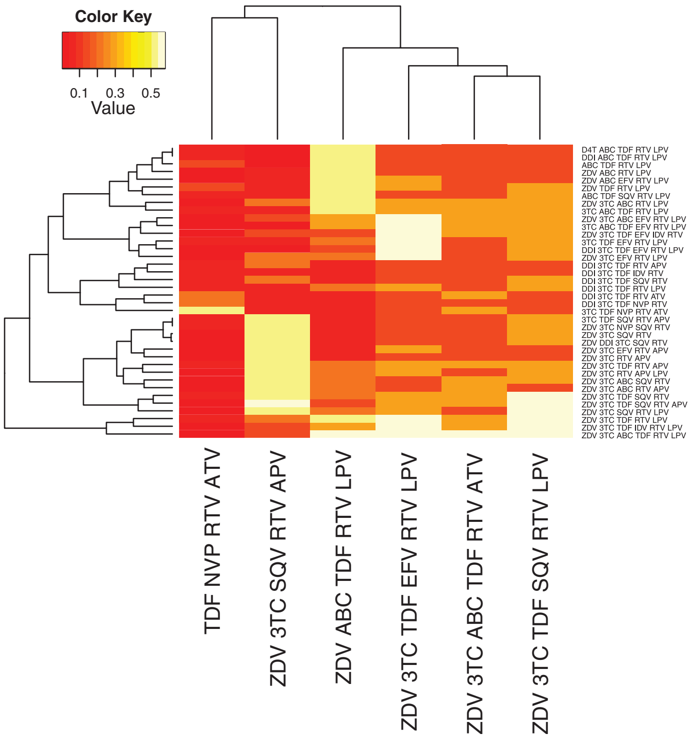

Fig. 4. Heatmap of the similarity profile of six test therapies with no training samples available. The profile considers only the training therapies with similarity value >0.5. The test therapies are depicted on the horizontal axis and the training therapies are depicted on the vertical axis. The similarity values are derived according to the drugs kernel.

4 CONCLUSIONS

This article presents an approach that tackles the problem of predicting virological response to combination therapies by learning a separate logistic regression model for each different combination therapy. Each model is fitted by using not only the data from the target therapy but also the information from therapies similar to it. For this purpose, we introduce and evaluate two different measures of pairwise therapy similarity which are used as weights in the logistic regression models. The model is also able to incorporate phenotypic knowledge on the therapy outcomes through a Gaussian prior. With such an approach we balance the uneven therapy representation in the datasets and produce higher quality models for the therapies with very few training samples. Our model is not only advantageous for therapies with few training samples, but also for all other therapies. Having a separate model for each therapy increases the interpretability of the fitted models in that users have access to the argumentative basis of the prediction. On the one hand, the scores of the mutations contributing to therapy effectiveness, which result from the linear predictions are derived in a therapy-specific manner and can, therefore, be considered more informative than for a general model. On the other hand, the therapy similarity kernel affords information on which similar therapies were most informative for the prediction. It has to be stressed that interpretability of the prediction is a prime requirement for use of a prediction method in a medical setting. Finally, the use of an efficient optimization method that takes advantage of the sparseness of our input data ensures very fast model fitting and model selection, although we train a separate model for each combination therapy.

In terms of accuracy, our *therapy similarity model* performs significantly better (at the 1% level) than training separate models for each therapy by using solely its samples. Furthermore, the *therapy similarity model* significantly outperforms (at the 4% level for the *drugs kernel* and the 8% level for the *groups additivity kernel*), the *one-for-all scenario* for the group of therapies with few (fewer than four) training samples. Although this group comprises only about 8% of the test data, it contains around 40% of the different drug combinations in the test set. For therapies with a sizeable number of samples (above four) both the *similarity* and the *one-for-all models* have comparable performance. Our model achieves similar accuracy and significantly better AUC performance than the *transfer model*, which uses a compute-intensive distribution matching approach to quantify the therapy similarities. This demonstrates the quality of our therapy similarity measures.

The phenotypic prior knowledge included via a Gaussian prior does not improve the accuracies of the similarity models, but it significantly improves (at the 4% level) the AUC of the test therapies that have 0–20 available training samples. This group comprises about 25% of the test data and contains around 77% of the different drug combinations in the test set.

ACKNOWLEDGEMENT

We gratefully acknowledge the EuResist project [contract number EU-STREP IST-2004-027173], the Virolab project [contract number INFSO-IST-027446] and the Arevir project [grant number HO 1582/1-3] for providing the data.

Funding: This work was funded by the Cluster of Excellence (Multimodal Computing and Interaction).

Conflict of Interest: none declared.

REFERENCES

Altmann,A. *et al.* (2007) Improved prediction of response to antiretroviral combination therapy using the genetic barrier to drug resistance. *Antivir. Ther.*, **12**, 169–178.

Altmann,A. *et al.* (2009a) Advantages of predicted phenotypes and statistical learning models in inferring virological response to antiretroviral therapy from hiv genotype. *Antivir. Ther.*, **14**, 273–283.

Altmann,A. *et al.* (2009b) Predicting response to combination antiretroviral therapy: retrospective validation of geno2pheno-theo on a large clinical database. *J. Infect. Dis.*, **199**, 999–1006.

Beerenwinkel,N. *et al.* (2003a) Geno2pheno: Estimating phenotypic drug resistance from HIV-1 genotypes. *Nucleic Acids Res.*, **31**, 3850–3855.

Beerenwinkel,N. *et al.* (2003b) Methods for optimizing antiviral combination therapies. *Bioinformatics*, **19**, i16–i25.

Bickel,S. *et al.* (2008) Multi-task learning for HIV therapy screening. In *Proceedings of the 25th Conference on Machine Learning*. Omnipress, pp. 56–63.

Deforche,K. *et al.* (2008) Modelled in vivo hiv fitness under drug selective pressure and estimated genetic barrier towards resistance are predictive for virological response. *Antivir. Ther.*, **13**, 399–407.

DeLong,E. *et al.* (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, **44**, 837–845.

Evgeniou,T. *et al.* (2000) Regularization networks and support vector machines. *Adv. Comput. Math.*, **13**, 1–50.

Johnson,V. *et al.* (2008) Update of the drug resistance mutations in HIV-1: December 2008. *Top. HIV Med.*, **16**, 138–145.

Larder,B. *et al.* (2007) The development of artificial neural networks to predict virological response to combination HIV therapy. *Antivir. Ther.*, **12**, 15–24.

Lathrop,R. and Pazzani,M. (1999) Combinatorial optimization in rapidly mutating drug-resistant viruses. *J. Comb. Optim.*, **3**, 301–320.

Lin,C. *et al.* (2008) Trust region newton method for large-scale logistic regression. *J. Mach. Learn. Res.*, **9**, 627–650.

Prosperi,M. *et al*. (2005) 'Common law' applied to treatment decisions for drug resistant HIV. *Antivir. Ther.*, **10**, S62.

Prosperi,M. *et al.* (2009) Investigation of expert rule bases, logistic regression, and non-linear machine learning techniques for predicting response to antiretroviral treatment. *Antivir. Ther.*, **14**, 433–442.

Roomp,K. *et al.* (2006) Arevir: a secure platform for designing personalized antiretroviral therapies against HIV. In *Lecture Notes in Computer Science: Data Integration in the Life Sciences*, Vol. 4075, Springer, pp. 185–194.

Rosen-Zvi,M. *et al.* (2008) Selecting anti-hiv therapies based on a variety of genomic and clinical factors. In *ISMB 2008 Conference Proceedings, Bioinformatics*, Vo. 24, pp. i399–i406.

UNAIDS/WHO (2009) AIDS Epidemic Update: December 2009. Available at http://data.unaids.org/pub/Report/2009/JC1700_Epi_Update_2009_en.pdf

Wang,D. *et al.* (2003) A neural network model using clinical cohort data accurately predicts virological response and identifies regimens with increased probability of success in treatment failures. *Antivir. Ther.*, **8**, S112.