Combining Kernel and Model Based Learning for HIV Therapy Selection

Sonali Parbhoo, MSc.[1], Jasmina Bogojeska, PhD[3], Maurizio Zazzi, PhD[4], Volker Roth, PhD[1] and Finale Doshi-Velez, PhD[2]

[1]University of Basel, Department of Mathematics and Computer Science, Switzerland.
[2]Harvard University, John A. Paulsen School of Engineering and Applied Sciences, Cambridge, MA.
[3]IBM Research - Zurich, Switzerland.
[4]University of Siena, Department of Medical Biotechnology, Italy.

Abstract

We present a mixture-of-experts approach for HIV therapy selection. The heterogeneity in patient data makes it difficult for one particular model to succeed at providing suitable therapy predictions for all patients. An appropriate means for addressing this heterogeneity is through combining kernel and model-based techniques. These methods capture different kinds of information: kernel-based methods are able to identify clusters of similar patients, and work well when modelling the viral response for these groups. In contrast, model-based methods capture the sequential process of decision making, and are able to find simpler, yet accurate patterns in response for patients outside these groups. We take advantage of this information by proposing a mixture-of-experts model that automatically selects between the methods in order to assign the most appropriate therapy choice to an individual. Overall, we verify that therapy combinations proposed using this approach significantly outperform previous methods.

Introduction

Human Immunodeficiency virus (HIV-1) affects over 36 million people worldwide [1]. To date, the only practical treatment for HIV is life-long administration of combinations of antiretroviral drugs targeting different phases of viral replication. However, repeated use of the same combinations of drugs encourages the formation of drug-resistant viral strains. These drug-resistant viral strains are stored and render the future administration of the same or similar therapy combinations ineffective. The large number of potential therapy combinations and the high mutation rate of the virus make searching for an effective therapy particularly challenging, especially for patients at advanced stages of infection.

Several computational approaches have been developed to address this challenge. Regression-based approaches map elements of a patient's history directly to some output, such as virological response. For example, Altmann et al. [2] use tree-based learning to predict a patient's virological response using evolutionary information, while others [3, 4] predict virological response based on the patient's treatment history. Treatment recommendations can then be made based on what therapy combination is expected to have the best response. More recently, Bogojeska et al. [5] present a kernel-based approach that predicts whether a particular therapy choice will be successful given information about a patient and their treatment history—where success is defined by the viral load dropping below 400 copies/mL after at least 21 days of treatment under the therapy. The premise here is that patients with similar treatment histories are likely to respond to treatment in a similar way.

Unfortunately, none of these regression-based approaches directly address the sequential nature of the therapy selection process: that a choice of combination now might result in drug-resistant viral strains that may be hard to control later. Approaches based on reinforcement learning [6] make this sequential nature explicit: they output a *treatment policy* that selects therapy combinations not only to optimise virological response in the present, but also in the future. While reinforcement learning has been used to optimise HIV treatment strategies in simulation studies [7, 8], reasoning about possible futures from limited observational data can make these approaches brittle in practice.

In this work, we present a mixture-of-experts approach [9] that combines the strengths of both kernel-based regression and reinforcement-learning approaches for therapy selection in HIV. Kernel-based regression approaches excel when there are clusters of similar patients; they can model the idiosyncrasies in viral response specific to those patients. However, their prediction quality drops for patients that are not part of a tight cluster. In contrast, model-based reinforcement learning approaches first build a model of how the patient is expected to respond and then use that model to reason about how well a series of therapy selections will perform. These approaches tend to find simpler but

more robust patterns of response, a better alternative for patients outside of clusters. Our mixture-of-experts approach automatically chooses between these two options.

Specifically, our work makes the following contributions:

- We modify the kernel-based approach [5] to predict not just whether a therapy combination will be immediately successful, but whether it will control HIV in the long term. We demonstrate that optimising for immediate viral load reduction does not control viral loads or mutations in the long term.
- We train a model-based learner and combine it with the kernel-based approach via mixture-of-experts. On a hold-out set of 3 000 patients, we demonstrate that the therapy combinations proposed by our treatment policy outperform previous approaches.

Our post-hoc investigation also support our hypothesis that the kernel-based approach is used when a patient lies as part of a cluster, whereas the model-based approach is used for patients that have no nearby neighbours; this observation suggests that more nuanced approaches are needed to make optimal treatment recommendations for patients with HIV.

Background & Related Work

Kernel-based history-alignment models for HIV therapy prediction It is well-established that a patient's prior history is a key factor for predicting the efficacy of HIV treatment [3, 4]. Bogojeska et al. [5] use this fact in a history alignment model that measures the similarity between two therapy sequences[1]. Two therapy sequences are considered to be similar if they consist of similar drug combinations, are administered in a similar order, and produce similar genomic fingerprints in the viral population. If two patients have similar histories, Bogojeska et al. [5] demonstrate that they often respond similarly to treatment. Specifically, the history-alignment model first uses a *resistance mutations kernel* to quantify the pairwise similarities between different therapy combinations. Resistance mutations are the Single Nucleotide Polymorphisms (SNPs) in the HIV viral genome identified by the International AIDS Society as being associated with drug resistance [10]. The kernel assumes that similarity between the different drug groups is additive, which is a reasonable assumption since drugs belonging to different groups have different therapeutic targets and/or modes of action, and thus can be assumed to act independently.

Formally, the resistance mutations kernel is defined as follows. Let G denote the set of different drug groups. In our clinical data set we have five different drug groups: Nucleoside Reverse Transcriptase Inhibitors (NRTIs), Non-Nucleoside Reverse Transcriptase Inhibitors (NNRTIs), Protease Inhibitors (PIs), Integrase Inhibitors (IIs) and Entry Inhibitors (EIs). Let $\mathbf{u}_{zg}$ and $\mathbf{u}_{z'g}$ be the sets of resistance-relevant mutations for the drugs occurring in drug group $g \in G$ of the therapies $\mathbf{z}$ and $\mathbf{z}'$, respectively. The pairwise similarity between the drug-g mutations of the drug combinations $\mathbf{z}$ and $\mathbf{z}'$ is then calculated using the Jaccard index:

$$sim_g(\mathbf{z}, \mathbf{z}') = \frac{|\mathbf{u}_{zg} \bigcap \mathbf{u}_{z'g}|}{|\mathbf{u}_{zg} \bigcup \mathbf{u}_{z'g}|}, \tag{1}$$

where $|\cdot|$ denotes set cardinality. We then 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:

$$k_m(\mathbf{z}, \mathbf{z}') = \sum_{g \in G} \frac{sim_g(\mathbf{z}, \mathbf{z}')}{|G|}. \tag{2}$$

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 similarity. In this way the resistance mutations kernel also accounts for the similarity of the genetic fingerprint of the potential latent virus populations of the compared therapies. Furthermore, this kernel represents drugs in terms of their mutation profile and, by doing so,

[1] The combinations of drugs a patient takes at a particular time is defined as a therapy. A therapy sequence refers to a sequence of such combinations over time.

allows for high group similarity for non-identical drugs that have very similar resistance mutation profiles and thus takes the high level of cross resistance within the same drug class into account.

The therapy sequence alignment kernel uses the resistance mutations kernel and in what follows, we describe it in detail. Let $start(\mathbf{t})$ denote the point of time when the therapy $\mathbf{t}$ was started and $patient(\mathbf{t})$ denote the patient identifier corresponding to therapy sample $\mathbf{t}$. Then:

$$\rho(\mathbf{t}) = \{\mathbf{z} \mid (start(\mathbf{z}) \leq start(\mathbf{t})) \textit{ and } (patient(\mathbf{z}) = patient(\mathbf{t}))\} \tag{3}$$

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 $patient(\mathbf{t})$ not later than $start(\mathbf{t})$ ordered by their corresponding starting times, from older to newer. Note that each therapy sequence also contains the current therapy, the most recent therapy in the therapy sequence $\rho(\mathbf{t})$ is $\mathbf{t}$. The main objective is to quantify the similarity of therapy sequences by considering two therapy sequences as similar if they consist of similar drug combinations administered in a similar order and producing similar genomic fingerprints in the latent viral population. For this purpose, first the pairwise similarity between different drug combinations is quantified using the resistance mutations kernel and then the overall similarity between two therapy sequences of interest is computed using the kernel and the information on the order in which the therapies were administered. As the treatment history similarity score needs to reflect both the similarity of the different therapies comprising the therapy sequences and the order in which they were administered, it is computed by adapting the the Needleman-Wunsch score commonly used for assessing the quality of an alignment of a protein or nucleic acid sequences [11]. The alphabet used for the therapy sequence alignment consists of all distinct drug combinations making up the clinical data set. The resistance mutations kernel in Equation 2 determines the pairwise similarities s between its letters (therapies). As each therapy sequence ends with the current (most recent) therapy - the one that determines the label of the sample, the sequence alignment is adapted such that the rightmost (most recent) therapies (alphabet letters) are always matched, i.e. no gaps are allowed at the right end of an alignment.

The score from the alignment kernel, together with the viral genotype and drug history information, can then be used to train a regression model for predicting the outcome of a therapy in terms of success or failure. A therapy is defined as successful if that patient's viral load falls below 400 copies/mL after 21 days of treatment under the therapy [5]. In our work, we will replace this success criteria with the potential long-term value of a therapy choice.

Bayesian Model-Based Reinforcement Learning. Many problems, including HIV therapy selection, involve making a sequence of decisions with long-term consequences. The reinforcement learning framework formalises the sequential decision-making process as a series of repeated exchanges between an agent and its environment. At each time step, the agent selects an action a (such as a drug combination) and the environment returns some observations o (e.g. CD4 counts, viral loads, mutations) as well as an immediate reward r (e.g. did the viral load drop below a desired threshold). Given a history of length t, $h = \{a_1, o_1, r_1 \ldots, a_t, o_t, r_t\}$, the agent's goal is to choose the subsequent action such that it maximises discounted sum of its expected rewards, $\mathbb{E}[\sum_t \gamma_t r_t]$, where $\gamma \in [0, 1)$ trades off between current and future rewards.

This decision-making task may be formulated as a Partially Observable Markov Decision Process (POMDP) [12]. A POMDP m is defined by a finite set of hidden states $\mathcal{S}$ (e.g. a patient's true physiological state), actions $\mathcal{A}$ and observations $\mathcal{O}$. A transition function $T(s'|s, a)$ specifies the probability of transitioning from state s to s' when taking an action a. Similarly, an observation function $\Omega(o|s, a)$ specifies the probability of observing o from state s when taking action a. The reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ specifies the immediate reward that an agent receives upon performing an action from a particular state.

When the agent has full knowledge of its environment, the current state s_t describes everything about the past that is necessary to predict the future. However, in a POMDP the state is never directly observed. Instead, an agent perceives an observation at each time step. This observation, unlike the state, does not capture all the information from the past that may be necessary to make future decisions [13]. In particular, the entire history h_t may still be necessary to make predictions about the future. To overcome this problem, upon perceiving an observation at each time step, the agent must maintain a belief state $b \in \Delta\mathcal{S}$ that specifies the probability of being in each state given the series of actions performed and observations received thus far, from some initial belief b_0. The belief state may be updated at each time

step according to Bayes' rule [14] :

$$b_{t+1}(s) = \Omega(o|s,a) \sum_{s' \in \mathcal{S}} \frac{T(s'|s,a) b_t(s')}{Pr(o|b,a)}, \tag{4}$$

where $Pr(o|b,a) = \Omega(o|s,a) \sum_{s \in \mathcal{S}} T(s'|s,a) b_t(s)$. Optimal decisions are subsequently made on the basis of these beliefs.

The Bayesian reinforcement learning framework involves two steps. First, the agent learns a distribution over possible models $p(m)$ based on data—that is, we must first infer patterns of response based on the treatment histories that we have. Next, we must use these models to determine what would be the most optimal set of treatment decisions, known as a treatment policy. Reasoning about outcomes given a distribution over models ensures that we account for the uncertainty of not knowing a patient's true physiological dynamics. We summarise the core ideas for each of these steps below; for more details we refer the reader to [13, 14, 15].

If we choose to model the patient's hidden physiological state as a discrete variable, then we can model the uncertainty over the transitions $T(s'|s,a)$ and observations $\Omega(o|s,a)$ with Dirichlet distributions. Sampling a model given this Dirichlet prior is relatively straightforward; we only need to keep track of the number of times we believe a physiological state may have been visited [15]:

$$T(s'|s,a) = \frac{\phi^a_{s,s'}}{\sum_{s'' \in \mathcal{S}} \phi^a_{s,s''}} \qquad \Omega(o|s,a) = \frac{\psi^a_{s,o}}{\sum_{o' \in \Omega} \phi^a_{s,o'}}. \tag{5}$$

where $\phi^a_{s,s'} \forall s'$ and $\psi^a_{s',o} \forall o$ monitor the statistics of how often particular transitions and observations were observed. We sample a set of such models here. Given a set of models, we can find a (near) optimal treatment policy via forward search over the beliefs [15]. This process involves simulating a forward-looking tree of possible outcomes of a particular treatment policy, and using this to search for the actions that optimise these outcomes.

Methods

We now propose our mixture-of-experts approach for HIV therapy selection. First, we provide the specifics of the data we have used for experimentation. Next, we explain the structure of our mixture-of-experts model in its specific components. For our first expert, we modify the history alignment model of [5] to predict whether a drug combination will optimise long-term outcomes rather than immediate outcomes. Our second expert is a Bayesian POMDP. In our results, we will demonstrate that combining the Bayesian POMDP and history alignment kernel in our mixture-of-experts approach succeeds at delivering the most appropriate therapy choice for an individual that is tailored to their specific situation.

Cohorts The data used for this research comes from the EuResist database [16]. The entire EuResist database is among the largest available HIV data sources with records of HIV genotypes and response to antiretroviral therapy for more than 65 000 patients. We focus on a subset of this data set, where we make use of the genotypic and treatment response data of 32 960 patients, together with their corresponding $CD4^+$ and viral load measurements, gender, age, risk group, and number of past treatments recorded. The data set is a heterogenous set with patients of various ethnicities coming from over 100 countries, ranging from heavily pre-treated to having short treatment histories. In the modelling below, we limit ourselves to the 312 most common drug combinations that occur in the cohort. These drug combinations span 20 total drugs. The database has previously been used to build models such as the therapy alignment model, to predict the outcome of a particular therapy [17, 18]. We are however, specifically interested in optimising the therapy choice for a patient.

Mixture-of-experts architecture for choosing between models. We present a mixture-of-experts approach for HIV therapy selection. The idea is primarily based on the 'divide and conquer' principle, where a complex problem is solved by dividing it into simpler problems whose solutions can then be combined to provide a reasonable overall solution. The mixture-of-experts [9] is a technique that allows us to overcome the problems associated with choosing a particular

model, by automatically assigning regions of the input space to different models. In this way, if two models exhibit complementary properties and operate well in different regions of the input space, we can take advantage of both. In our case, we hypothesise that kernel-based methods do best for patients that are part of a cluster—that is, when there are many similar patients to draw inference from. However, they can perform poorly for patients that are farther from any cluster because they are not particularly similar to any other patient in the cohort. In contrast, the model-based POMDP approach tends to over-simplify: clearly, patients cannot be fully modelled as having discrete physiological states. However, an over-simplified model of HIV response may be preferable to the completely inaccurate prediction that we may make if we map a patient to dissimilar patient. However, the question remains of how to choose between the two experts in a formal way. The mixture-of-experts model provides an automatic solution to this problem of choosing between these methods.

The general architecture for a mixture-of-experts model consists of a set of expert classifiers, followed by a gating network. The gating network is typically not a classifier, but a combination rule to determine which classifier to select. In this particular case, we consider the policies produced by both the POMDP and alignment kernel approaches and would like to select between policies depending on a patient's particular situation. The model takes as input a set of viral load measurements, gender, age, risk group, resistance mutations, number of past treatments recorded, the lower quantile of the distance between a patient and their neighbours, and the length of the treatment history, as well as the policy parameters from both the POMDP and alignment kernel models. The outputs consist of the learned policies or therapy combinations under each model. We train classification models using these features to determine which policies should ultimately be applied for each patient. In principle, any classification model can be used for this purpose (as is the case in all mixture-of-expert models). We specifically used random forest models here, where hyperparameters such as the number of trees etc were chosen by experimenting on a separate hold-out set. In this case the number of trees used is 500. The overall outcome of the mixture-of-experts approach is thus a decision of which treatment policy is optimal for a patient (in terms of either the POMDP or alignment kernel models).

Long-term Reward Criterion Following [7, 8], we propose the following immediate reward function:

$$r_t = \begin{cases} -0.7 \log V_t + 0.6 \log T_t - 0.2|M|, & \text{if } V_t \text{ is above detection limits} \\ 5 + 0.6 \log T_t - 0.2|M|, & \text{if } V_t \text{ is below detection limits,} \end{cases}$$

where V_t is the viral load (in copies/mL), T_t is the CD4$^+$ count (in cells/mL), and $|M|$ is the number of mutations at time t respectively. This function penalise instances where a patient's viral load increases and rewards instances where a patient's CD4$^+$ count increases (more weight is placed on the viral load, as it is an earlier indicator of whether a therapy is working). We also penalise on the basis of the number of mutations a patient has at a particular time, as these may ultimately contribute to resistance and therapy failure. There is also a bonus for if the viral load is below detectable limits as this is something we would like to sustain over time. The long-term reward criterion sums these immediate rewards over the patient's history.

A kernel-based approach for HIV therapy selection that optimises long-term rewards. The history alignment model for predicting therapy outcome presented in the background was trained to predict whether a drug combination would be immediately successful. In the following, we call the treatment policy that optimises for immediate success the 'Short-Term History Alignment' model. We adapt this model in two ways: first, convert the binary problem of 'Will this drug combination succeed?' into a multi-class problem of 'Which drug combination should I choose?' (we consider the 312 most common drug combinations). Second, and more importantly, this prediction is made by summing over *all* the time-steps in the patient's future history. We call this treatment policy the 'Long-Term History Alignment' model.

A Bayesian POMDP for HIV therapy selection The RL framework introduced in Section 2 provides a natural way of modelling the sequential decision-making process of therapy selection.

As before, we limit our actions to the 312 most common drug combinations in the cohort and learn a model with 7 hidden physiological states. Our observation space consists of (a) binning the values of the viral load using a log scale of [0.0, 1.0, 10.0, 1000, 10000, 10M, 100M] copies/ML[2], (b) the 70 resistance mutations that may occur as a result

[2]One could have equally well modelled the data as continuous values with a Gaussian emission model.

of a particular therapy together with a patient's $CD4^+$ count, gender, risk group. Overall the procedure of interaction between an agent and its environment is illustrated in Figure 1.

To learn the parameters for the transitions and emissions, one option would be to fix the parameters on the basis of historical data. However, this approach would assume that the parameters are known with certainty. In reality, these quantities are highly uncertain, and should thus be learned through the process of monitoring and managing patients. Hence we treat the parameters in a Bayesian manner (as outlined in the background section), where the agent can explicitly track its uncertainty and refine its knowledge as a result. We model time in discrete increments of 6 months, and perform a forward search for therapy choices that optimise outcomes over a 5 year horizon (10 total steps).

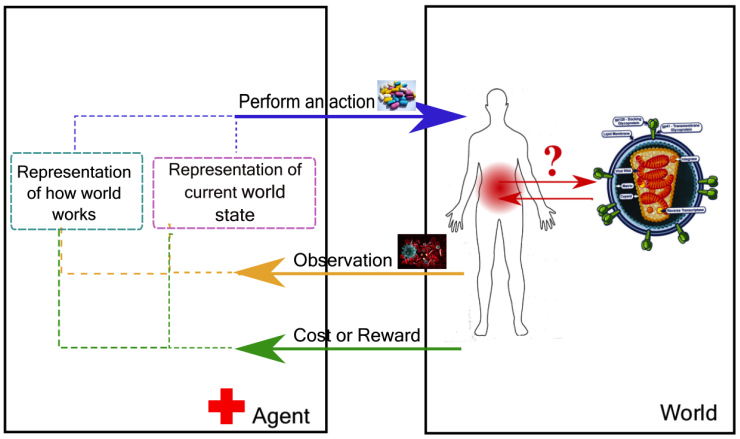

Figure 1. HIV Therapy Selection as a POMDP problem.

Evaluation Procedure So far, we have presented several approaches to tackling the HIV therapy selection problem: the history alignment models, the model-based approaches, and the mixture-of-experts. How can we know which has better performance, given only retrospective data in which completely different treatment decisions were made? A series of off-policy evaluation strategies have been proposed to address this question. In our results, we compare our approaches using three different schemes to show that our results are robust to the choice of off-policy evaluation:

- *Importance Sampling* [19] corrects the mismatch between the treatment policy to be evaluated and the treatment decisions made for the cohort using importance sampling. It is unbiased but can have high variance.
- *Weighted Importance Sampling* [20] attempts to reduce this variance by weighting the importance scores.
- *Doubly Robust Evaluation* [21] further reduces variance and acheives consistency by coupling the importance score with a regression estimate of the value of the treatment policy. The procedure works well if importance ratio between the treatment policy to be evaluated and the actual decisions is balanced or the regression estimate is accurate.

Results

Table 1 compares the performance of the history alignment method, the POMDP and the mixture-of-experts against a random policy where a completely random therapy choice is made. The evaluation is performed for a test set of 3 000 randomly selected patients from the cohorts. The performance is measured in terms of the average accumulated long-term rewards over a period of 5 years with their standard deviations. A higher value indicates a better performing treatment policy.

	Doubly Robust	Importance Sampling	Weighted Importance
Random Policy	–2.31 $\pm$ 1.42	–3.48 $\pm$ 1.36	–2.80 $\pm$ 1.27
Short-term History Alignment	2.17 $\pm$1.47	2.14 $\pm$1.22	2.15 $\pm$1.16
Long-term History Alignment	9.48 $\pm$ 1.90	5.42 $\pm$ 1.93	6.74 $\pm$ 1.89
POMDP	6.34 $\pm$ 2.15	4.36 $\pm$ 2.38	6.76 $\pm$ 2.24
Mixture-of-experts	**11.47 $\pm$ 1.38**	**12.25 $\pm$ 1.41**	**11.23 $\pm$ 1.40**

Table 1. Off-Policy evaluation using importance sampling, weighted importance sampling and doubly robust methods for different therapy selection models.

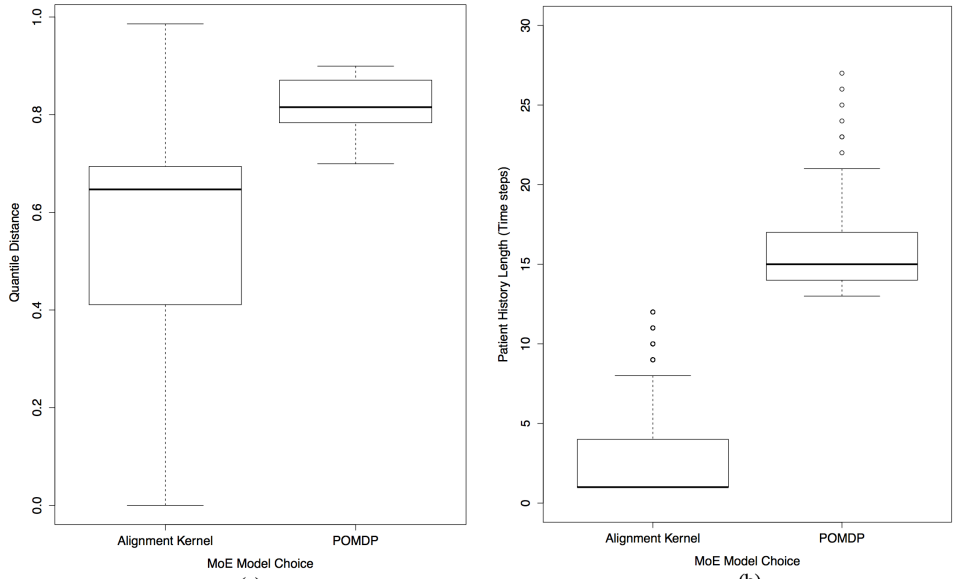

Figure 2. Mixture-of-experts model choice over (a) distances to closest neighbour and, (b) varying history lengths.

Optimising for long-term health produces different treatment policies than predicting the most common next therapy Table 1 shows that the short-term history alignment model achieves significantly worse rewards than the long-term history alignment model (and the POMDP). These results suggest that treatments which may initially appear attractive may result in poor patient outcomes at a later stage—unsurprising to many in HIV. Specifically, resistance against a particular drug may lead to cross-resistance against another, leading to long-term dependencies in therapy response.

The mixture-of-experts produces the best treatment policies. The mixture-of-experts approach outperforms the other approaches across all evaluation schemes. While the POMDP performs worse than the long-term alignment kernel in general, the fact that the mixture of experts approach outperforms both the POMDP and alignment kernel suggests that these models are making mistakes in different places—and thus we can do better by choosing between the two. A post-hoc investigation reveals that the mixture-of-experts chooses the POMDP model approximately 26% of the time in comparison to the alignment kernel.

The mixture-of-experts chooses experts based on clustering characteristics. We follow up on our clustering hypothesis: when is the long-term history alignment method being chosen, and when is the POMDP being chosen? Specifically, we consider role that a patient's history and the lower quantile of their distance to other patients plays here. Figure 2 (a) and (b) provide box plot illustrations of the values of the latter features in relation to the choice of model selected by the mixture-of-experts. As the lower quantile of the distance between a patient and their neighbours increases, the POMDP is more likely to become the model of choice. Moreover, the length of a patient's therapy history seems to play a defining role in the choice of expert. The mixture-of-experts selects the POMDP for patients with longer history lengths and the alignment kernel for the others. One possible explanation for this would be that as a patient's treatment history increases in length, they become more unique and hence have smaller similarity values relative to other patients in the kernel. In these instances the POMDP is the model of choice, possibly because of the fact that it is able to incorporate this rich history into its belief state, while the alignment kernel cannot capture the same level of information.

Discussion

In this paper, we demonstrated how kernel methods and model-based RL can be combined for HIV therapy selection using a mixture-of-experts approach. The mixture-of-experts enabled us to account for heterogeneity in patient data, that typically makes it difficult for a single model to provide reasonable therapy predictions for individuals. Kernel methods attempt to group patients on the basis of similarity, and use this as a means of identifying optimal therapies, while RL models treat patients as individuals and use past experience to reason about suitable therapies and outcomes. By combining these approaches, we were able to take advantage of the strengths of each method in different situations. Specifically, we showed that the kernel approach is optimal for patients with short treatment histories; these are likely patients that are at early stages of infection, or those that are treatment-inexperienced. On the other hand, the model-based method proved more suited to patients with long treatment histories and rare therapy combinations. These are patients that have been heavily pre-treated. We attribute this difference directly to the way in which each model uses a patient's history: the POMDP incorporates knowledge about a patient's history implicitly through its beliefs and actions, each influenced by past observations, treatments and mutations, while the history alignment method only uses patient's treatment history from the kernel, and does not account for observations that occur further back in time. Overall, the mixture-of-experts approach outperformed each method on its own as a result of its adaptability in various patient situations.

There are many opportunities for future research. An interesting direction would be to incorporate the kernel-based predictions directly into the reinforcement-learning models. Another potential aspect of interest would be to use such a model to rank existing clinical policies for different scenarios.

Acknowledgements

This research was partially supported by the Swiss National Science Foundation, project 51MRP0_158328. The first author gratefully acknowledges C. Van Alten and P. Ranchod at University of the Witwatersrand, Johannesburg, for many helpful discussions that ultimately contributed to this work.

References

1. UNAIDS. AIDS by the numbers; 2015. Available from: http://www.unaids.org/en/resources/documents/2015/AIDS_by_the_numbers_2015.

2. Altmann A, Beerenwinkel N, Sing T, Savenkov I, Däumer M, Kaiser R, et al. Improved prediction of response to antiretroviral combination therapy using the genetic barrier to drug resistance. Antiviral therapy. 2007;12(2):169.

3. Revell A, Wang D, Harrigan R, Hamers R, Wensing A, Dewolf F, et al. Modelling response to HIV therapy without a genotype: an argument for viral load monitoring in resource-limited settings. Journal of antimicrobial chemotherapy. 2010;p. dkq032.

4. Prosperi MC, Rosen-Zvi M, Altmann A, Zazzi M, Di Giambenedetto S, Kaiser R, et al. Antiretroviral therapy optimisation without genotype resistance testing: a perspective on treatment history based models. PloS one. 2010;5(10):e13753.

5. Bogojeska J, Stöckel D, Zazzi M, Kaiser R, Incardona F, Rosen-Zvi M, et al. History-alignment models for bias-aware prediction of virological response to HIV combination therapy. In: AISTATS; 2012. p. 118–126.

6. Sutton RS. Introduction to reinforcement learning. vol. 135;.

7. Ernst D, Stan GB, Goncalves J, Wehenkel L. Clinical data based optimal STI strategies for HIV: a reinforcement learning approach. In: Proceedings of the 45th IEEE Conference on Decision and Control. IEEE; 2006. p. 667–672.

8. Parbhoo S. A Reinforcement Learning Design for HIV Clinical Trials. University of the Witwatersrand, Faculty of Science, School of Computer Science; 2014.

9. Jordan MI, Jacobs RA. Hierarchical mixtures of experts and the EM algorithm. Neural computation. 1994;6(2):181–214.

10. Wensing AM, Calvez V, Günthard HF, Johnson VA, Paredes R, Pillay D, et al. Special contribution 2014 Update of the drug resistance mutations in HIV-1. Topics in antiviral medicine. 2014;22(3):642.

11. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of molecular biology. 1970;48(3):443–453.

12. Kaelbling LP, Littman ML, Cassandra AR. Planning and acting in partially observable stochastic domains. Artificial intelligence. 1998;101(1):99–134.

13. Spaan MT. Partially observable Markov decision processes. In: Reinforcement Learning. Springer; 2012. p. 387–414.

14. Vlassis N, Ghavamzadeh M, Mannor S, Poupart P. Bayesian reinforcement learning. In: Reinforcement Learning. Springer; 2012. p. 359–386.

15. Ross S, Pineau J, Chaib-draa B, Kreitmann P. A Bayesian approach for learning and planning in partially observable Markov decision processes. Journal of Machine Learning Research. 2011;12(May):1729–1770.

16. Zazzi M, Incardona F, Rosen-Zvi M, Prosperi M, Lengauer T, Altmann A, et al. Predicting response to antiretroviral treatment by machine learning: The euresist project. Intervirology. 2012 1;55(2):123–127.

17. Zazzi M, Kaiser R, Sönnerborg A, Struck D, Altmann A, Prosperi M, et al. Prediction of response to antiretroviral therapy by human experts and by the EuResist data-driven expert system (the EVE study). HIV medicine. 2011;12(4):211–218.

18. Prosperi MC, Altmann A, Rosen-Zvi M, Aharoni E, Borgulya G, Bazso F, et al. Investigation of expert rule bases, logistic regression, and non-linear machine learning techniques for predicting response to antiretroviral treatment. Antivir Ther. 2009;14(3):433–42.

19. Precup D, Sutton RS, Singh S. Eligibility traces for off-policy policy evaluation. Computer Science Department Faculty Publication Series. 2000;p. 80.

20. Rubinstein RY, Kroese DP. Simulation and the Monte Carlo method. vol. 707. John Wiley & Sons; 2011.

21. Jiang N, Li L. Doubly Robust Off-policy Evaluation for Reinforcement Learning. arXiv preprint arXiv:151103722. 2015;.