Improved Prediction of HIV Resistance In-Vitro by Biochemically-Driven Models

Hani Neuvirth[1,2]**, Michal Rosen-Zvi**[2]**, Nathan Srebro**[2,3]**, Ehud Aharoni**[2]**,**
Maurizio Zazzi[4] and **Naftali Tishby**[1]
[1] School of Engineering and Computer Science, Hebrew University of Jerusalem, Jerusalem 91904, Israel
[2] IBM Research Laboratory in Haifa, Haifa University, Mount Carmel, Haifa 31905, Israel
[3] Toyota Technological Institute at Chicago , Chicago IL 60637, USA
[4] Section of Microbiology,Department of Molecular Biology, University of Siena, Italy

Abstract

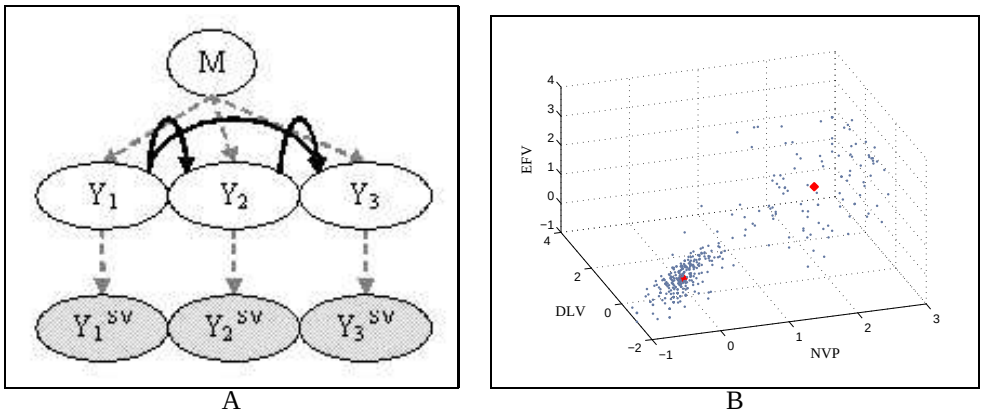
A crucial aspect in anti-HIV therapy is choosing the best treatment among the many available. To this aim studies try to extract information from the HIV virus' genotype that can be used to computationally predict its in-vitro susceptibility to the available drugs. This paper combines the generative and discriminative approaches in a novel model that improves the utilization of the biological structure of the system. Two aspects are considered. First is the prior distribution of the data which is a mixture of two Gaussians, representing viruses that are either susceptible or resistant to the drugs. Second are the strong dependencies between drugs of the same mechanism. The performance is remarkably improved showing error rates of up to 38% of the error achieved by previous studies.

1 Introduction

The Acquired Immunodeficiency Syndrome (AIDS) is a fatal disease considered to be the global epidemic of our time. It is caused by the HIV virus, which directly attacks the cells of the immune system, until it fails to respond against simple everyday invaders. A major obstacle to anti-HIV treatment is the high rate at which this virus replicates and mutates. At a lack of appropriate treatment new viral generations become less susceptible to the available drugs. Hence, confronting the virus with the best available treatment at its earliest stages is crucial.

There exist two test schemes searching for the best treatment. One method is phenotypic testing, in which a sample of the virus is tested in the lab against each of the existing drugs. These tests measure the fold-resistance (FR) namely, the change in susceptibility of the virus to the drug relative to the baseline susceptibility of the wild-type sequence. This method is expensive, and thus not accessible to large part of the HIV-infected population. An alternative method is genotypic testing in which the relevant gene of the virus is sequenced, and computational tools are used to predict its susceptibility to each of the drugs.

Several tools exist that aim to predict phenotypic resistance from genotypic data. Rule based systems such as those of the Agence Nationale de Recherches sur le Sida [3], Rega Institute [7], HIVdb (Stanford University, [5]) and AntiRetroScan (University of Sienna, [8]) typically provide a three-level prediction, classifying the virus as resistant, susceptible, or intermediate. A different approach is applied by the Geno2pheno system which uses decision trees, information theoretic analysis and Support Vector Regression (SVR) to provide full range predictions [2, 1]. A new paper examines five different statistical learning techniques, namely decision trees, neural networks, SVR, least square regression and least angle regression for the prediction task [6]. All these methods were found to have similar performance.



A B

Figure 1: **A. The SVR-net Model.** The lower nodes of the model represent the observed sequence's features, extracted by the SVR. These can also be regarded as noisy measures of the "true" FRs that are represented by the nodes in the intermediate layer. These nodes are inter-connected with a full DAG (black edges) to form an N-dimensional Gaussian capturing the cross-resistance between the drugs. Finally, the root node imposes the overall structure of the sequences that is a mixture of Gaussians. **B. The FR Distribution of Sequences Tested against the Drugs in the NNRTI Class.** The blue dots represent different RT sequences. Each axis represents the log(FR) against one drug. The dense Gaussian has coordinates lower than 1, thus these are the susceptible sequences. The wide Gaussian belongs to resistant sequences. The red dots are the centers of the Gaussians learnt by the SVR-net model.

Whilst these tools use state-of-the-art Machine Learning techniques, the elementary relations in the data are overlooked. The utilization of the underlying biochemical processes should significantly improve the prediction success rate.

In this study we design a hierarchical Bayesian model that is based on some basic biochemical insights. In a cost of slightly increasing the complexity of the model, the predictions are remarkably improved.

2 The SVR-net Model

Existing HIV-phenotypic-resistance modeling techniques ignore several biological aspects. The most important one is the classification of the drugs according to the mechanism of the reaction. Phenotypic data is available for drugs that belong to one of three classes: Protease Inhibitors (PI), Nucleoside Reverse Transcriptase Inhibitors (NRTI) and Non-NRTI (NNRTI). Newer drugs exist, but resistance data is not yet available for them. All these drugs inhibit viral proteins that are crucial for the virus' replication. The first class, as implied by its name, inhibits a protein named Protease while the two other classes inhibit the protein Reverse Transcriptase (RT). The NRTI and NNRTI classes differ by the location of the binding site of the drug on the RT. Thus, the effects of drugs within the same class are highly correlated and result in multidrug resistance. Therefore, a unified model combining the separate predictors within a class would reduce the parameters space size, and better exploit the information in the available data.

An additional fact that has been reported, but not expressed in the models is the prior distribution of the available sequences. Beerenwinkel et al. [1] show that the distribution of the FR of each drug is not uniform rather it is a mixture of two Gaussians. The first Gaussian which has a small variance and low FR values is associated with sequences that are susceptible to the drugs. The Second Gaussian having a large variance and high mean FR is associated with viruses that developed resistance to the drug.

These biological insights are the motivation for the SVR-net model suggested here. The SVR learners used in previous studies serve here as a preprocessing feature extraction stage, with a Bayesian network built on top of them to impose the biologically relevant structure. The Bayesian network (figure 1**A**) can be interpreted as a generating process of the sequence's features. The equivalence to

the biological system is problematic, as the biologically-relevant order of things starts from a new mutated sequence showing a certain level of resistance. Nevertheless, this model is in some sense equivalent to the evolutionary selection pressure applied on the sequences by the drug. Specifically, at the low resolution the selection process distinguishes between resistant and susceptible sequences, each having different probability of overcoming the drug, and becoming the dominant sequence. The following selection is at a higher resolution, refining the resistance (and the probability of the sequence to dominate) according to specific features of the sequence.

Reading from the model, there are two classes of sequences, susceptible and resistant. First, the sequence's class is chosen at random from a Multinomial distribution, $P(M)$. The particular susceptibilities per drug in the particular group of drugs is jointly chosen from an N dimensional Gaussian distribution $P(\mathbf{Y}|M) \sim \mathcal{N}(\mu, \mathbf{\Sigma})$, note that the covariance matrix, $\mathbf{\Sigma}$, of this multivariable Gaussian is not diagonal, as we expect correlations between drugs from the same group. Then the specific features that represent the sequence are randomly chosen, each feature from its corresponding susceptibility component, Y_i where $i = 1...N$ with N being the number of drugs in the group. These features are regarded as the "true" susceptibilities. The ones observed are a noisy version of the true ones, $P(Y_i^{SV}|Y_i) \sim \mathcal{N}(Y_i, \Sigma_i^{SV})$. In figure 1**A** we illustrate this model for a class of N=3 drugs. The Y_i^{SV} variables at the bottom are always observed. This means that the sequences are observed, and hence also their features. Thus, we aim in learning the model parameters $\theta = \{P(M), \mu_i^{SV}, \Sigma_i^{SV}, \mathbf{\Sigma}, \mu\}$ from examples of $\{\mathbf{Y}, \mathbf{Y}^{SV}\}$ and predicting the values of the variables $\{\mathbf{Y}\}$ given a previously unseen example, of $\{\mathbf{Y}^{SV}\}$ Formally, the probabilistic model induced by the network is: $P(\mathbf{Y}, \mathbf{Y}^{SV}, M|\theta) = P(M)P(\mathbf{Y}|M)\prod_{i=1}^{N} P(Y_i^{SV}|Y_i)$ and the maximum likelihood optimization problem is: $\arg\max_\theta \log P(\mathbf{Y}, \mathbf{Y}^{SV}|\theta)$.

The parameter estimation procedure of this model is composed of two phases. First, is the feature extraction phase implemented as separate SVRs for each drug. Second, the Bayesian network is trained, when the prediction of the SVR for each sequence serves as the observed variables of the network. Since the "true" FR values, $\mathbf{Y}$, are only partially observed, the optimization is done using EM.

3 Results

We used the SVR-net model to predict the FR for the HIVdb dataset [5]. This dataset includes 621 protease sequences and 418 RT sequences. The results for each drug class are presented in figure 2. The height of the bars is the mean squared error (MSE) achieved. The error bars represent the standard error (SE) over the 10-fold cross validation (CV). The white (leftmost) bars represent the state of the art results using independent SVR for each drug, and the dark (rightmost) bars represent the results achieved by the SVR-net model. The improvement over the state-of-the-art prediction is immense showing mean squared errors that are 80-38% of the current errors. For the NNRTI class, due to the small number of drugs in this class, the improvement is less remarkable, though still evident.

The SVR-net model introduces two novel elements over predictions with independent SVRs: a Gaussian-mixture prior distribution on the drug resistance, and correlations between the drugs, modeled by a non-diagonal multivariate Gaussian. To asses the effect of these elements, we experimented with predicting the drug resistance of viral sequences in the PI dataset using two intermediate models: An SVR-net model with a single Gaussian component (i.e. M is fixed and constant) and an SVR-net model where the prior distribution over each Y_i is an independent mixture of two univariate Gaussians (i.e. a separate and independent SVR-net model for each drug). The results, shown in Figure 2(a), indicate that the role of the Gaussian-mixture prior is stronger than that of the correlation between the drugs.

Figure 1**B** shows the 3-dimensional data of the NNRTI class. The axis represent the FR of the three drugs of this class in a log scale. The red dots mark the center of the Gaussians that were learnt. This figure exemplifies the motivations as well as the suitability of the SVR-net model to the data. The two aspects utilized here are observed: the prior distribution of the two Gaussians, as well as the fact that the coordinates of these Gaussians support their interpretation as either susceptible or resistant, thus expressing the correlation between the drugs.

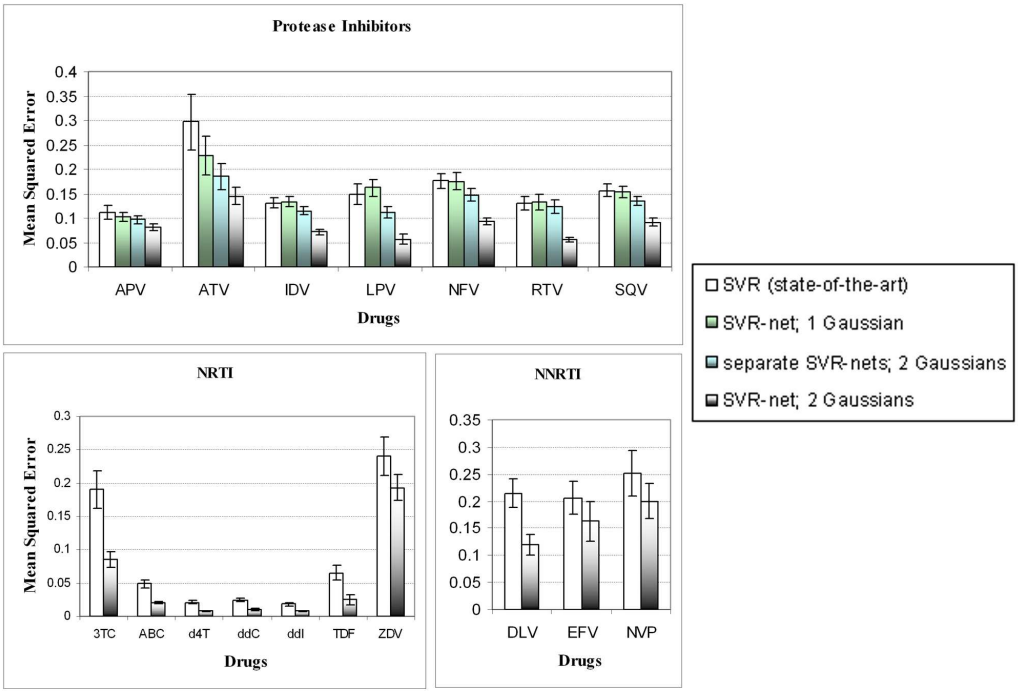


Figure 2: **The Success Rate on the Three Drug Classes.** The drug names appear on the X axis. The Y axis shows the mean squared error. The error bars are the standard error. The state-of-the-art (separate linear SVR; leftmost bars) is compared with the new SVR-net model (rightmost). For Protease inhibitors (A) a comparison of the contribution of the two different aspects incorporated in our model: the correlation between the drugs (green; second leftmost) and the miture of two Gaussians as a prior (cyan; second rightmost) is presented.

4 Discussion and Conclusions

This paper presents a model that focuses on exploiting the known biochemically-driven structure of the data. Two main aspects are considered. First is the Prior distribution of the data. This has been reported before to fit a mixture of two Gaussians independently for every drug [1]. We show that by considering this fact the prediction is significantly improved. Theoretically, having 2 Gaussians in each of the one dimensional spaces, might result with many Gaussians in the higher space. Considering that each Gaussian is associated with either resistant or susceptible sequences, having only two Gaussians at the high dimension would mean that the multidrug resistance is inherent. Due to limited data, we could not test the SVR-net model with a larger number of Gauissians (4, 6 etc.). Inspecting the NNRTI data (figure 1**B**) supports the 2 Gaussians model. Yet, when data will be available, examining its distribution at the higher space will reveal more insights regarding the relations between the drugs.

These relations are the second aspect considered in this study. Using the classification of the drugs according to their mechanism of action the number of models is reduced from 17 independent models to 3. Thus, the noise in the data is better overcome. This allows to slightly increase the complication of the models over the independent linear SVR, which are evidently very simplifying.

The described model brings up a novel approach also from the Machine Learning point of view. The utilization of SVR as a feature extraction stage is a simple way of combining the advantages of discriminative learning together with the generative approach. The Bayesian network models the inherent structure of the data, but it would be intractable to directly add a layer of hundreds of nodes necessary to represent the sequence directly. This is overcome by the discriminative power of the SVR. We are currently examining an alternative approach formalizing the same model under the framework of conditional random fields.

References

[1] Niko Beerenwinkel, Martin Daumer, Mark Oette, Klaus Korn, Daniel Hoffmann, Rolf Kaiser, Thomas Lengauer, Joachim Selbig, and Hauke Walter. Geno2pheno: Estimating phenotypic drug resistance from HIV-1 genotypes. *Nucleic Acids Res*, 31(13):3850–3855, Jul 2003.

[2] Niko Beerenwinkel, Barbara Schmidt, Hauke Walter, Rolf Kaiser, Thomas Lengauer, Daniel Hoffmann, Klaus Korn, and Joachim Selbig. Diversity and complexity of HIV-1 drug resistance: a bioinformatics approach to predicting phenotype from genotype. *Proc Natl Acad Sci U S A*, 99(12):8271–8276, Jun 2002.

[3] Francoise Brun-Vezinet, Dominique Costagliola, Mounir Ait Khaled, Vincent Calvez, Francois Clavel, Bonaventura Clotet, Richard Haubrich, Dale Kempf, Marty King, Daniel Kuritzkes, Randall Lanier, Michael Miller, Veronica Miller, Andrews Phillips, Deenan Pillay, Jonathan Schapiro, Janna Scott, Robert Shafer, Maurizio Zazzi, Andrew Zolopa, and Victor DeGruttola. Clinically validated genotype analysis: guiding principles and statistical concerns. *Antivir Ther*, 9(4):465–478, Aug 2004.

[4] Thorsten Joachims. Making large-scale svm learning practical. In Bernhard Schölkopf, Chris Burges, and Alex Smola, editors, *Advances in Kernel Methods – Support Vector Learning*, pages 169–185. MIT Press, 1998.

[5] Soo-Yon Rhee, Matthew J Gonzales, Rami Kantor, Bradley J Betts, Jaideep Ravela, and Robert W Shafer. Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res*, 31(1):298–303, Jan 2003.

[6] Soo-Yon Rhee, Jonathan Taylor, Gauhar Wadhera, Asa Ben-Hur, and Robert W. Brutlag, Douglas L. Shafer. Genotypic predictors of human immunodeficiency virus type 1 drug resistance. *Proc Natl Acad Sci U S A*, Epub ahead of print, Oct 2006.

[7] Kristel Van Laethem, Andrea De Luca, Andrea Antinori, Antonella Cingolani, Carlo Federico Perna, and Anne-Mieke Vandamme. A genotypic drug resistance interpretation algorithm that significantly predicts therapy response in HIV-1-infected patients. *Antivir Ther*, 7(2):123–129, Jun 2002.

[8] Maurizio Zazzi, Laura Romano, Giulietta Venturi, Robert W Shafer, Caroline Reid, Federico Dal Bello, Cristina Parolin, Giorgio Palu, and Pier E Valensin. Comparative evaluation of three computerized algorithms for prediction of antiretroviral susceptibility from HIV type 1 genotype. *J Antimicrob Chemother*, 53(2):356–360, Feb 2004.

5 Supplementary Material - Training and Testing

An SVR was trained using SVMLight [4] in 10-fold cross-validation separately of each drug on its labeled data. The tube width was set to 0.1. The cost was optimized using binary search. The final cost values for protease inhibitors are APV: 0.09; ATV: 0.06; IDV: 0.05; LPV: 0.12; NFV: 0.04; RTV: 0.01; SQV: 0.08; for NRTIs: 3TC: 0.11; ABC: 0.01; D4T: 0.07; DDC: 0.04; DDI: 0.18; TDF: 0.01; AZT: 0.06; d4T: 0.07; ddI: 0.18; ddC: 0.04; ZDV: 0.06; and for NNRTIs: DLV: 0.16; EFV: 0.36; NVP: 0.30;

These costs were used both for the baseline prediction estimation, and for the new SVR-net model suggested here. For the SVR-net model, the labeled data (i.e. data with available phenotypic fold-resistance measured) of each drug was repartition to 10 folds. At the first phase, an SVR was trained using 9 folds. In the following phase, the network was trained using EM on the same 9 folds and the available unlabeled data of this drug.

Given a new sequence of the virus, predictions are extracted from each SVR, and then the network is used to evaluate the most probable FR value of the relevant drug node. Note, that available FR values of neighboring drugs were not given as input to the network.