Semi-Supervised Learning for Improving Prediction of HIV Drug Resistance

Juliane Perner, André Altmann, Thomas Lengauer

Computational Biology and Applied Algorithmics
Max Planck Institute for Informatics
Campus E1.4
D-66123 Saarbrücken
{jperner,altmann,lengauer}@mpi-inf.mpg.de

Abstract: Resistance testing is an important tool in today's anti-HIV therapy management for improving the success of antiretroviral therapy. Routinely, the genetic sequence of viral target proteins is obtained. These sequences are then inspected for mutations that might confer resistance to antiretroviral drugs. However, interpretation of the genomic data is challenging. In recent years, approaches that employ supervised statistical learning methods were made available to assist the interpretation of the complex genetic information (e.g. *geno2pheno* and *VircoTYPE*). However, these methods rely on large amounts of labeled training data, which are expensive and labor-intensive to obtain. This work evaluates the application of semi-supervised learning (SSL) for improving the prediction of resistance from the viral genome.

1 Introduction

The Human Immunodeficiency Virus (HIV) is causing one of the most challenging infectious diseases. HIV is a retrovirus that mainly infects cells of the human immune system. Today there are about 25 antiretroviral drugs approved by the US Food and Drug Administration for treating HIV infections[1]. These drugs can be divided into different classes by their mechanism of action and the viral proteins they target. Reverse transcriptase inhibitors aim at prohibiting the synthesis of DNA from viral RNA by the viral protein reverse transcriptase (RT). This can currently be accomplished by nucleos(t)id analogs that lead to abortion of DNA synthesis after their incorporation. In constrast to these nucleoside reverse transcriptase inhibitors (NRTIs), non-nucleoside reverse transcriptase inhibitors (NNRTIs) bind to the viral RT and impair its flexibility. Integrase inhibitors prevent the integration of the viral DNA into the host genome by blocking the viral enzyme integrase. Finally, protease inhibitors (PIs) bind to the active site of the viral protease that cleaves precursor proteins into functionally units. The large number of drugs that are on the marketplace is required because the process of reverse transcription is error prone and therefore HIV eventually develops mutations in the targeted proteins that confer resistance against the applied drugs. These mutations enable the virus to replicate in the presence

[1] `http://www.fda.gov/oashi/aids/virals.html`

of a drug and are therefore selected evolutionarily. Unfortunately, these resistance mutations also confer drug resistance to drugs of the same class that were not applied yet, this phenomenon is termed *cross-resistance*. Resistance testing is an important tool in therapy management for choosing an appropriate drug regimen for the patient and consequently slow down disease progression to AIDS and death. There are two approaches to resistance testing. The first approach, *phenotyping*, affords a lab test that compares the viral replication of the virus of a patient with that of a wild type virus in the presence of the drug [Wa99]. The quotient of dosages of the drug that are required to cut the replication rate of the patient sample and the wild type, respectively, in half is called the *resistance factor*. The second approach, *genotyping*, amounts to sequencing the genes of the viral drug targets harbored by the virus variant predominating in the patient. These sequences have to be inspected for mutations that are related to drug resistance. Phenotyping is expensive and labor-intensive but delivers a single number per drug that is easy to interpret. Genotyping on the other hand is fast, cheap, and standardized, but the correct interpretation of the genetic sequence poses a major challenge. One way to address this problem is provided by knowledge-based approaches (expert systems) that apply classification rules. These rules are hand-crafted by experts based on literature, *in vitro* results, and clinical experience. Rule sets can be found, e.g. in Stanford's HIVdb [Rh99]. More systematic approaches employ supervised statistical learning methods to predict the resistance of a virus against drug based on the sequences of the genes coding for the target proteins, e.g. *geno2pheno* [Be03] and *VircoTYPE* [Ve07]. These supervised learning methods are trained on viral samples for which both, a genotypic test and a phenotypic test has been performed. However, for achieving a good performance a sufficient number of training samples is required (at least several hundred), which is in general expensive and labor-intensive to collect. Thus, especially, at the time shortly after the approval of a novel drug usually only a small number of genotype-phenotype pairs is available and consequently prediction methods lag behind in providing an assessment of these drugs. Since relevant parts of the HIV genome are routinely sequenced for diagnostic reasons, ample genotypic data without phenotypic measurements are available in clinical databases. This work focuses on the use of semi-supervised learning (SSL) for improving the prediction of drug resistance based on genotype-phenotype data together with available routinely collected sequence data. Recently, an SSL approach using unlabeled data from clinical routine for improved dimensionality reduction was applied to predict *in vivo* response to antiretroviral combination therapies [RAS09]. Section 2 provides a brief overview over the available data as well as supervised and semi-supervised methods that were applied. Section 3 presents the results, and section 4 gives a conclusion and an outlook.

2 Materials and Methods

2.1 Data

The genome sequences of the target proteins were available as amino acid sequences that had been aligned to the reference sequence HXB2. For the protease all 99 amino acids and

	NRTI						NNRTI	
drug name	ZDV	3TC	ddI	d4T	ABC	TDF	EFV	NVP
cutoff	0.9	1.37	0.37	0.25	0.54	0.25	0.7	0.67
susceptible (%)	49	57	50	53	44	42	61	50
$\|S_{labeled}\|$	1055	740	882	881	871	598	1037	880
$\|S_{drug}\|$	2717	5143	2329	3047	1225	1668	1264	1237
$\|S_{class}\|$	7887						2502	
	PI							
drug name	APV	ATV	IDV	LPV	NFV	SQV	DRV	TPV
cutoff	0.59	1.13	0.72	0.6	0.67	0.98	1.14	0.61
susceptible (%)	54	60	48	66	43	59	50	48
$\|S_{labeled}\|$	645	523	721	682	725	725	55	60
$\|S_{drug}\|$	290	320	756	1442	1075	687	0	0
$\|S_{class}\|$	4435							

Table 1: Description of the data. $|S_{labeled}|$ indicates the number of available genotype-phenotype pairs. The row *cutoff* lists the log_{10}(resistance factor) cutoff values used to dichotomize the continuous value into the categories *susceptible* (below the cutoff) and *resistant* (above the cutoff). The row susceptible (%) indicates the percentage of labeled data that was considered susceptible after dichotomization. The rows $|S_{drug}|$ and $|S_{class}|$ list the numbers of sequences that were obtained during exposure to the specific drug and drug class, respectively. Drugs: zidovudine (ZDV), lamivudine (3TC), didanosine (ddI), stavudine (d4T), abacavir (ABC), tenofovir disoproxil fumerate (TDF), efavirenz (EFV), nevirapine (NVP), (fos-)amprenavir (APV), atazanavir (ATV), indinavir (IDV), lopinavir (LPV), nelfinavir (NFV), saquinavir (SQV), darunavir (DRV), tipranavir (TPV).

for the RT only the first 220 amino acids were considered. The genotype-phenotype pairs were provided by the Arevir database [Ro06]. For every drug a different number of measured resistance factors (RFs) with corresponding genotype was available (see: Table 1). Unfortunately, most SSL approaches work for classification only, thus the continuous RFs were dichotomized to *susceptible* and *resistant* using a drug-specific cutoff. This cutoff was defined by the intersection of two Gaussian distributions, which the RFs display when plotted on logarithmic scale. The two Gaussian distributions represent the susceptible and resistant subpopulation as described in [Be03]. The cutoffs derived in this way for each drug are listed in Table 1. Sequences generated in diagnostic routine were taken from the EuResist database [Ro08] and constitute the unlabeled data used by the SSL methods. Sequences were categorized as to whether they were exposed to a specific drug (S_{drug}) or to a specific drug class (S_{class}) at the time the sample was obtained (see Table 1).

2.2 Statistical Methods

Semi-supervised learning methods operate on a labeled set $S_{labeled} = \{(x_1, y_1), (x_2, y_2)\ , ..., (x_n, y_n)\}$ and a set of unlabeled data $S_{unlabeled} = \{x_1^*, x_2^*, ..., x_m^*\}$, where x_i and y_i denote feature vector and corresponding label, respectively. The unlabeled data $S_{unlabeled}$

reveals information about the underlying data density. This knowledge can be exploited by SSL methods for generating improved prediction models compared to supervised methods. We can expect that SSL improves the prediction only, if labels show a tendency to be locally constant in input data space. This assumption is termed *smoothness assumption* and states that: if two points are located closely in data space, then their corresponding output is more likely to be similar (regression) or identical (classification). Consequently, the decision boundary derived by a SSL classification method should not cut through regions of high data density. Most of the semi-supervised methods perform transductive learning, i.e. the learner has to predict a set of labels $\{y_1^*, y_2^*, ..., y_m^*\}$ for the given unlabeled data $S_{unlabeled} = \{x_1^*, x_2^*, ..., x_m^*\}$. These unlabeled samples have to be available while training the method. According to the definition of transductive learning in [Zh07], transductive methods cannot handle unseen data. Thus, if a prediction for a new unlabeled sample x_{m+1}^* is needed, a new model using $x_{m+1}^* \cup S_{unlabeled}$ has to be trained for computing the label y_{m+1}^*. In contrast, inductive learners (e.g. classic supervised methods) yield a prediction function on the whole input space. Thus, inductive learners can also handle previously unseen data.

This section gives a brief overview over the SSL methods used in this work. The large number of different SSL approaches (for an overview see [CSZ06, Zh07]) was restricted to methods that are easily accessible (e.g. in the form of command line tools or available source code). As reference supervised methods support vector machines (SVMs) [CL01] were used for classification and regression, whereas regularized least-squares regression (RLSR) [SGV98] was used for regression only.

Transductive Support Vector Machine (tSVM) The standard soft margin SVM optimizes the following function:

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\| + C\sum_{i=1}^{N}\xi_i, \text{subject to } \xi_i \geq 0, y_i(\mathbf{w}x_i + b) \geq 1 - \xi_i, \forall_i \quad (1)$$

where $\mathbf{w}$ and b define the hyperplane, ξ_i are the slack variables that allow for misclassification and C is the cost parameter for misclassified examples. The tSVM aims at determining a separating hyperplane under consideration of the unlabeled samples, therefore equation (1) is extended in the following way:

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\| + C\sum_{i=1}^{n}\xi_i + C^*\sum_{j=1}^{m}\xi_j^*, \text{subject to}$$
$$\xi_i, \xi_j^* \geq 0, y_i(\mathbf{w}x_i + b) \geq 1 - \xi_i, y_j^*(\mathbf{w}x_j^* + b) \geq 1 - \xi_j^*, \forall_{i,j} \quad (2)$$

where the additional parameters ξ_j^* and C^* are the slack variables and the misclassification cost parameter for the unlabeled instances, respectively. Thus, the optimization problem in (2) differs from (1) in that the tSVM has to find a labeling $y_1^*, ..., y_m^*$ for the unlabeled data and a hyperplane $< \mathbf{w}, b >$ simultaneously. An approximative optimization procedure, which is required due to the complexity of the optimization problem, has been implemented in the software library SVM^{light} by Joachims [Jo99]. The approach begins with

a labeling of $x_1^*, ..., x_m^*$ based on the classification of an inductive SVM and a low weight C^* for the penalty for misclassified unlabeled data points. Then the labels of two randomly selected samples (one positive and one negative) are swapped. If the objective function is improved by that exchange of labels, then the switch is made permanent. This process is repeated until there are no more switches possible that yield an improved objective function. At this point the penalty for misclassified unlabeled data points C^* is increased and further labels are swapped to greedily improve the objective function. The iterative procedure stops when C^* exceeds a user defined value. Notice that applying the definition of transductive learning stated above, tSVMs are in fact inductive learners. However, the name tSVM originated from the intention to work only on the observed data [Zh07].

Low Density Separation The Low Density Separation (LDS) approach introduced in [CZ05] is a combination of a tSVM and a kernel based on graph distances that takes advantage of unlabeled data. The main idea of LDS is the construction of a density-sensitive kernel. This is achieved by representing the feature vectors x_i and x_j^* of labeled and unlabeled samples as nodes in a graph. Each node is connected to its k nearest neighbors by weighted edges, with the weight of an edge corresponding to the Euclidean distance of its endpoints. For all paths between two points the largest edge weight on the path is computed. The similarity of two points is then defined as a function of the minimum of these largest edge weights. The main idea behind the density-sensitive kernel for SSL is to enlarge the distance between points that are separated by regions of low data density. This kernel is used by a tSVM that applies gradient descent for finding a solution of a slightly modified version of equation (2) and is therefore termed ∇SVM. For a detailed description of the approach see [CZ05].

Co-Regularized Least-Squares Regression (coRLSR) In comparison with semi-supervised classification, semi-supervised regression is largely under-studied. However, in [Br06] an efficient semi-supervised regression method is introduced that is based on the idea of co-learning. Briefly, the approach assumes the existence of multiple views, i.e. distinct sets of features, which are equally well suited for predicting the outcome. CoRLSR trains one regularized least-squares regression (RLSR) for each view on the labeled data and the available unlabeled data are used to measure the disagreement of the models. By the optimization process the disagreement of models for different views is minimized. CoRLSR with two views has the following optimization function:

$$Q(\mathbf{c}) = \sum_{v=1}^{2} \left[\|y_v - L_v c_v\|^2 + \nu_v c_v^t L_v c_v \right] + \lambda_v \sum_{u,v=1}^{2} \|U_u c_u - U_v c_v\|^2 \qquad (3)$$

where $\mathbf{c} = (c_1, c_2) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ represents the trained model for each view, n_v is the number of training samples in each view, ν_v and λ_v control the influence of the regularization term $c_v^t L_v c_v$ and the penalty for disagreement between views, respectively. Furthermore, $L_v \in \mathbb{R}^{n_v \times n_v}$ is the kernel matrix for all labeled samples and the matrix $U_v \in \mathbb{R}^{m \times n_v}$ comprises the inner products of all combinations of unlabeled and labeled instances. The first term of the sum represents the optimization criterion for fitting a regularized least-squares model, the second part of the sum calculates the disagreement of two views on the

unlabeled samples. In the setting under study different views were not available. However, results in [BS04] demonstrated that for many problems the feature set can be randomly split into different views and, together with co-classification approaches, still outperform traditional single-view learning algorithms. Thus, in the experiments the amino acid positions of protease and RT were randomly distributed among two views.

2.3 Evaluation setup

The labeled data that were used to train the methods are denoted by L, where L is a subset of $S_{labeled}$. Method performance was then assessed on the remaining labeled data $S_{labeled} - L$. From this subset only the genome sequences were used in the training procedure of SSL approaches and those are referred to as $U_{lab} = \{x_i|(x_i, y_i) \in (S_{labeled} - L)\}$. The genetic sequences from routine diagnostic used by the SSL methods are referred to as S_{drug} and S_{class} for sequences exposed to the same drug and to the same drug class as the drug for which a prediction model is trained, respectively. The training data of the SSL methods comprised $L \cup U_{lab} \cup S_{drug}$ or $L \cup U_{lab} \cup S_{class}$ while the training data for the standard supervised methods were restricted to L. For each drug listed in Table 1 separate models were trained. Performance was computed by using 10-fold cross-validation, which means that for each cross-validation fold 90% of $S_{labeled}$ were attributed to L and the remaining 10% to U_{lab}. In addition to evaluating the usefulness of SSL methods (using different sets of unlabeled data) over standard supervised methods, the influence of the size of available labeled data on the prediction performance was studied. To this end, only a randomly chosen subset of L was actually used during training. The size of that subset was either 2.5%, 5%, 10%, 20%, 40%, 60%, 80%, or 100% of the size of L. The remaining samples from L were excluded from the respective analysis. All learning approaches except LDS applied a linear kernel. The amino acid sequences were encoded as described in [Be03]: one amino acid position was represented by 20 indicator variables, i.e. one indicator for each amino acid. Classification performance was assessed by calculating the the area under the receiver operating characteristics (ROC) curve (AUC). Regression performance was measured as mean squared error (MSE) between predicted log_{10}(RF) and measured log_{10}(RF). The model parameters of the methods (see section 2.2) were optimized during the 10-fold cross-validation. Sets of different parameters were tested for each fold and the set performing best was used for performance computation.

3 Results and Discussion

3.1 Classification

Figure 1 summarizes the classification results by depicting the performance of all methods for all drugs when 10% and 100% of L were used during training, respectively. These fractions of L were selected for reflecting the amount data typically available shortly after

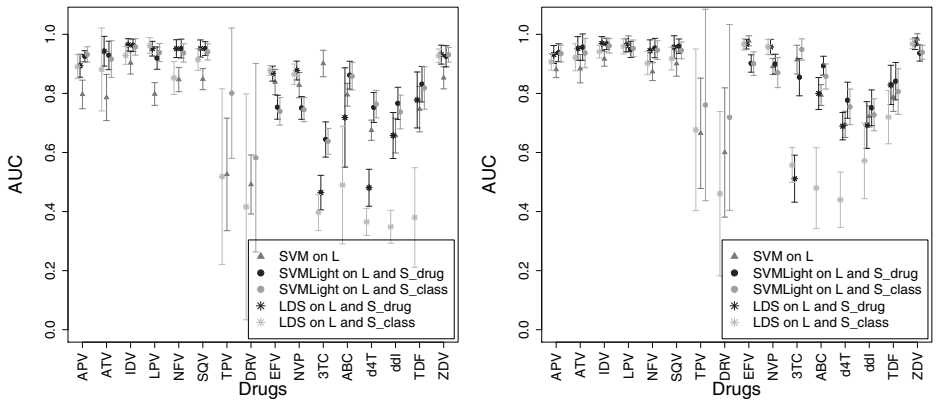

Figure 1: Mean area under the ROC curve (AUC) for 10% (left) and 100% (right) of the labeled data for the reference method and both SSL classification methods trained with the additional unlabeled sets S_{drug} and S_{class}, respectively. Whiskers indicate the standard deviation computed via 10-fold cross-validation.

approval of a novel drug and the amount of maximally available data. Figure 2 shows the AUC for varying volume of labeled data for three drugs representing the three drug classes. For protease inhibitors SSL brought a consistent benefit over supervised learning. With only 10% of L used during training all SSL methods performed at least as well as the supervised SVM for all PIs. Usage of the smaller unlabeled set S_{drug} brought a slight benefit over S_{class}. When 100% of L were used the gain in performance of SSL methods over the SVM was less pronounced. TPV and DRV are novel drugs in the class of PIs. The amount of labeled data is small and none of the available sequences were ever exposed to these drugs (Table 1). For both drugs SSL classification models did not show an improvement over the supervised SVM classification. However, this lack of observable improvement might be a consequence of the low number of instances available for assessing the performance. This assumption is supported by the large standard deviation of the AUC. For the two NNRTIs the results were less consistent. For EFV the SSL version in SVM^{light} did not show an improvement over classical supervised learning for any fraction of labeled data (Figure 2). For NFV the use of the SSL routine in SVM^{light} resulted in a clearly lower performance only when a small volume (10%) of labeled data was used. LDS performed for both drugs as well as or slightly better than the supervised SVM. For the group of NRTIs the results were even more diverse. While for ZDV and a small volume of labeled data the SSL methods displayed an improvement over the supervised SVM, for 3TC both SSL approaches drastically corrupted the performance. This difference might be explained by the different resistance profiles of the drugs. For 3TC one amino acid exchange is sufficient to confer complete resistance, while for ZDV several mutations are necessary. NRTIs are usually given in pairs to the patients, thus viruses that were exposed to 3TC were also exposed to other NRTIs with more complicated resistance patterns (e.g. ZDV). As a consequence, the data density does not reflect the labeling of 3TC resistance, which is a violation of the smoothness assumption. This finding is supported by the fact

that LDS with its density-sensitive kernel performs worse than SVM^{light} for 3TC. For the remaining four NRTIs the classification performance was worse compared to the remaining drugs. This is related to the small ranges of resistance factors that are observed for these drugs. Consequently, the Gaussian densities for the susceptible and resistant subpopulations are heavily overlapping, and therefor the computations of an appropriate cutoff is difficult. However, SVM^{light}performed better than the supervised SVM.

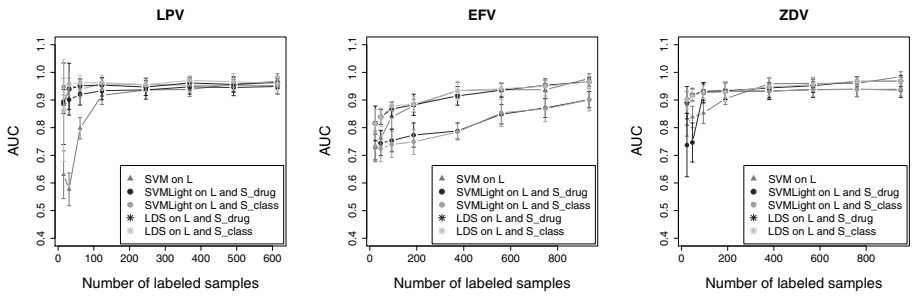

Figure 2: Development of the area under the ROC curve (AUC) for different volumes of labeled data for the reference method and both SSL classification methods trained with the additional unlabeled sets S_{drug} and S_{class}, respectively, for three drugs. Lopinavir (left), efavirenz (middle), and zidovudine (right). Whiskers indicate the standard deviation computed via 10-fold cross-validation.

3.2 Regression

Figure 3 depicts the performance of SVM, RLSR, and coRLSR for all drugs when 10% and 100% of L were used during training, respectively. Figure 4 shows the detailed development of the mean squared error for increasing volume of labeled data for LPV, EFV, and ZDV. CoRLSR, the only semi-supervised regression method tested in this study, did not improve the performance over RLSR or support vector regression. Moreover, the set of unlabeled data used during training (S_{drug} or S_{class}) did not play any substantial role in the performance of coRLSR. CoRLSR performed worse than RLSR for 3TC. As a consequence of dividing the amino acid positions among the two views for coRLSR, only one view had access to the single amino acid position that causes 3TC resistance. This fact violates the assumption that both views are sufficient for correct predictions and therefore lead to a significantly decreased performance compared to RLSR.

4 Conclusion and Outlook

Semi-supervised learning has the capability to improve the prediction of drug resistance from important regions in the HIV genome. The classification methods displayed a clear benefit over classical supervised learning for most drugs when only few labeled training

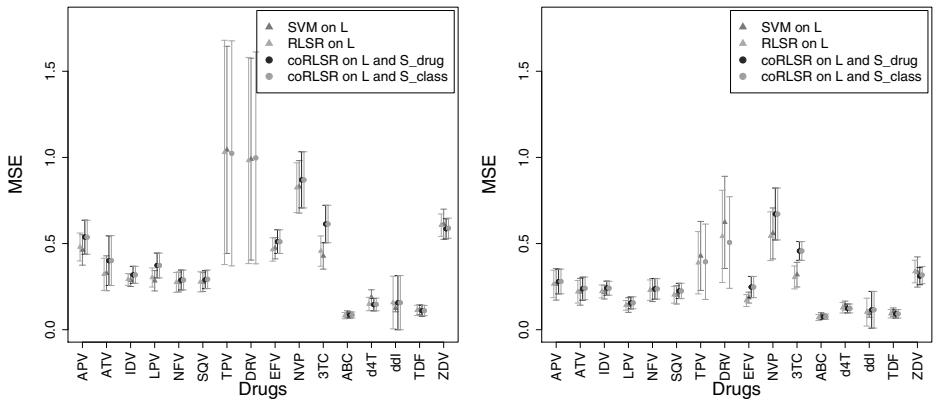

Figure 3: Mean squared error for 10% (left) and 100% (right) of the labeled data for the reference methods and coRLSR trained with the unlabeled sets S_{drug} and S_{class}, respectively. Whiskers indicate the standard deviation computed via 10-fold cross-validation.

samples were available. PIs, a drug class with strong cross-resistance between drugs, benefited the most from the use of SSL. The results support that SSL methods are suitable for improving prediction of drug resistance for novel drugs in established drug classes, such as daraunavir and tipranavir. Generally, it is not clear whether SSL is helpful for drugs belonging to novel drug classes (e.g. integrase inhibitors), because only few sequences harboring resistance mutations are available and SSL can also corrupt the classification results as seen for 3TC. The only semi-supervised regression model coRLSR could not improve the performance over the supervised methods.

Acknowledgments

The authors thank the Arevir project for providing training data in form of genotype-phenotype pairs, the EuResist project (IST-4- 027173-STP) and its coordinators Francesca Incardona and Maurizio Zazzi for providing a large number of sequences from routine diagnostics, and Ulf Brefeld for sharing the coRLSR code.

References

[Be03] Beerenwinkel, N. *et al.*: Geno2pheno: estimating phenotypic drug resistance from HIV-1 genotypes. In: Nucleic Acids Research, Vol. 31, No. 13, 2003. Oxford University Press, 2003; pp. 3850-3855.

[Br06] Brefeld, U. *et al.*: Efficient Co-Regularized Least Squares Regression. In: Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, 2006. pp.137 - 144.

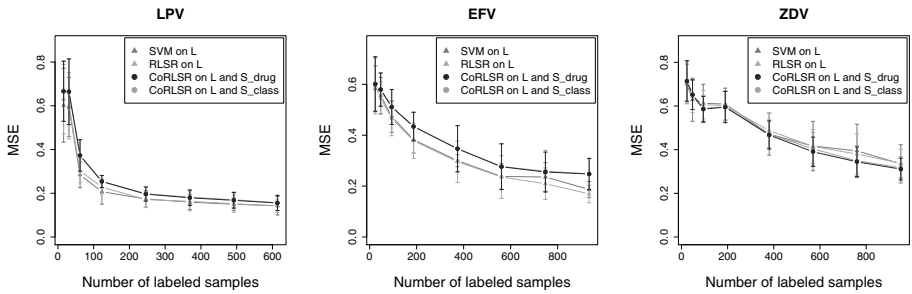

Figure 4: Development of the mean squared error for different volumes of labeled data for the reference methods and coRLSR trained with the unlabeled sets S_{drug} and S_{class}, respectively, for three drugs. Lopinavir (left), efavirenz (middle), and zidovudine (right). Whiskers indicate the standard deviation computed via 10-fold cross-validation.

[BS04] Brefeld, U.; Scheffer, T.: Co-EM Support Vector Learning. In: Proceedings of the 21st International Conference on Machine Learning, Banff, 2004. pp.121-128.

[CL01] Chang C.-C.; Lin C.-J.: LIBSVM: a library for support vector machines. 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`

[CSZ06] Chapelle, O.; Schölkopf, B.; Zien, A.: Semi-Supervised Learning. The MIT Press, Cambridge, Massachusetts, 2006.

[CZ05] Chapelle, O.; Zien, A.: Semi-Supervised Classification by Low Density Separation. In: Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics, Barbados, 2005. pp.57-64.

[Jo99] Joachims, T.: Transductive Inference for Text Classification using Support Vector Machines. In: Proceedings of the 16th International Conference on Machine Learning, Bled, 1999. pp.200-209.

[RAS09] Rosen-Zvi, M.; Aharoni, E.; Selbig, J.: HIV-1 Drug Resistance Prediction and Therapy Optimization: A Case Study for the Application of Classification and Clustering Methods. M. Biehl *et al.* (Eds.): Similarity-Based Clustering, LNAI 5400, 2009. pp.185-201.

[Rh99] Rhee, SY *et al.*: Human immunodeficiency virus reverse transcriptase and protease sequence database. In: Nucleic Acids Research, 2003, Vol. 31, No. 1. Oxford Press, 2003. pp. 298-303.

[Ro06] Roomp, K. *et al.*: Arevir: a secure platform for designing personalized antiretroviral therapies against HIV. In: Lecture Notes in Computer Science, Vol. 4075, Berlin/Heidelberg. Springer, 2006. pp.185-194.

[Ro08] Rosen-Zvi, M. *et al.*: Selecting anti-HIV therapies based on a variety of genomic and clinical factors. In: Bioinformatics, Vol. 24, Issue 13, 2008. Oxford Journals, 2008. pp. i399-i406.

[SGV98] Saunders, C.; Gammerman, A.; Vovk, V.: Ridge regression learning algorithm in dual variables. In: Proceedings of the 15th International Conference on Machine Learning, San Francisco, 1998. Morgan Kaufmann, 1998. pp. 515-521.

[Ve07] Vermeiren, H. *et al.*: Prediction of HIV-1 drug susceptibility phenotype from the viral genotype using linear regression modeling. In: Journal of Virological Methods, Vol. 145, Issue 1, 2007. Elsevier Science, 2007. pp. 47-55.

[Wa99] Walter, H. *et al.*: Rapid, phenotypic HIV-1 drug sensitivity assay for protease and reverse transcriptase inhibitors. In: Journal of Clinical Virology, Vol. 13, Issue 1, 1999. Elsevier Science, 1999. pp. 7180.

[Zh07] Zhu, X.: semi-supervised learning Literature Survey. Webpage: `http://pages.cs.wisc.edu/~jerryzhu/research/ssl/semireview.html`, accessed Mai 2008.