92

Keeping models that predict response to antiretroviral therapy up-to-date: fusion of pure data-driven approaches with rules-based methods

Andre Altmann[1], Alexander Thielen[1], Dineke Frentz[2], Kristel Van Laethem[3], Laura Bracciale[4], Francesca Incardorna[5], Anders Sönnerborg[6], Maurizio Zazzi[7], Rolf Kaiser[8], Thomas Lengauer[1]

1 Max Planck Institute for Informatics, Saarbrücken, Germany; 2 Erasmus Medical Centre Rotterdam, Rotterdam, the Netherlands; 3 Katholieke Universiteit Leuven, Leuven, Belgium; 4 Catholic University of Rome, Rome, Italy; 5 Informa srl, Rome, Italy; 6 Department of Medicine, Karolinksa Institute, Stockholm, Sweden; 7 Department of Molecular Biology, University of Siena, Siena, Italy; 8 Institute of Virology, University of Cologne, Cologne, Germany

Background

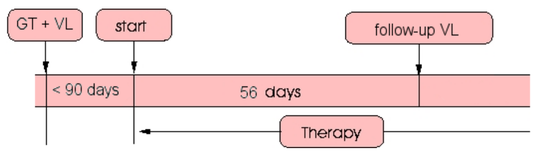
Purely data-driven approaches predicting response to antiretroviral combination therapy (ART) are frequently criticized for not being easily applicable to novel compounds. This drawback originates from lack of treatments containing novel compounds in clinical databases that constitute the source of training data for statistical-learning methods. Here we present a method that integrates rules-based methods into data-driven approaches to overcome this limitation.

Material & Methods

Treatment Data

4,538 treatment change episodes (TCEs) were extracted from the EuResist and Violab databases as training data for the approach. Response to ART was dichotomized to success and failure based on the 8 (4-12) week follow-up viral load (VL), i.e. reduction below 500 cp/ml or a more than 100-fold reduction compared to the baseline VL defined a success.

An additional 119 TCEs containing new drugs (DRV n=59, TPV n=52, ETR n=4, DRV+ETR n=4) were extracted from the databases for independently assessing performance of the new approach.

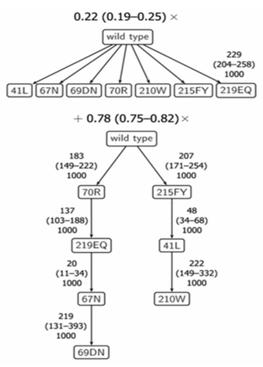


Novel drug GSS (ndGSS)

The 4,538 training TCEs did not contain novel drugs according to today's definition. However, for training of the new approach all drugs in a treatment that were approved by FDA at most 4 years prior to treatment start were treated as novel drugs.

	ZDV	ddI	d4T	3TC	ABC	TDF	NVP	EFV	IDV	SQV	NFV	LPV	APV	ATV
Year FDA approval	1987	1991	1994	1995	1998	2001	1996	1998	1996	1995	1997	2000	1999	2003
Year until *novel*	1991	1995	1998	1999	2001	2003	2000	2002	1999	1999	2000	2002	2003	2004
Drug used in TCEs	1355	1085	900	3123	752	1852	395	823	237	245	359	1726	314	423
# TCEs changed	0	0	60	133	153	413	75	243	66	54	128	394	69	110

Standard Encoding



The information contained in one TCE was encoded as a vector comprising binary and continuous values. The drugs used in the treatment and the mutations present in the baseline genotype were encoded by binary variables: representing presence (1) and absence (0). Furthermore, the genetic barrier to drug resistance for every drug was computed. Interaction terms between drugs and drugs as well as between drugs and mutations were introduced. This set of covariates represents the baseline encoding.

The extended encoding augmented the baseline encoding by indicators for previous use of a drug, indicators for exposure to NRTIs, NNRTIs, and PIs, respectively. Additionally the log10(baseline VL) and interaction terms between previously used drugs and currently used drugs were considered.

	baseline	extended
Genetic barrier	yes	yes
indicators for drugs	yes	yes
indicators for mutations	yes	yes
drug x drug	yes	yes
mutation x mutation	yes	yes
drug x mutation	yes	yes
indicators for previous drugs	no	yes
drug x previous drug	no	yes
baseline VL	no	yes

ndGSS Encoding

In contrast to the standard encoding, drugs that were treated as being novel were neither considered in the binary variables nor in the genetic barrier, thus they were not part of the treatment anymore according to the standard encoding. Instead, the HIVdb 5.0.0 tool was used to compute the resistance for these drugs, and the GSS of all novel drugs was used as an additional covariate, the novel drug GSS (ndGSS).

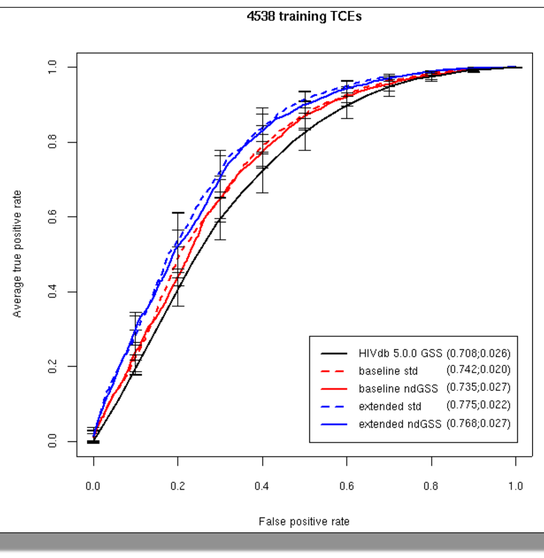
Example:

Year treatment	*standard encoding*	*ndGSS encoding*	*ndGSS*
2001ZDV+3TC+LPV/r	ZDV+3TC+LPV/r	ZDV+3TC	1.0
2006ZDV+3TC+LPV/r	ZDV+3TC+LPV/r	ZDV+3TC+LPV/r	0.0

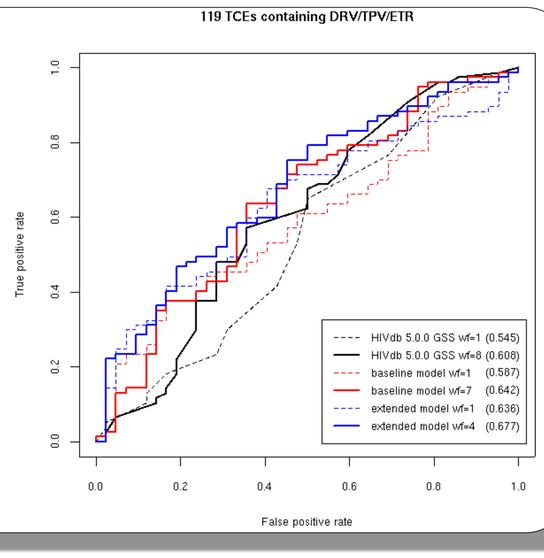
For differentiating between the lack of novel drugs and all novel drugs being resistant (both cases result in an ndGSS of 0.0), the ndGSS was set to -1.0 if all novel drugs were classified as resistant. Moreover, a weighting factor (wf) was introduced for adjusting the influence of ndGSS in all predictions.

Results

Logistic regression models were trained for both encodings and tested in a 10-fold cross validation setting. Results were compared to an HIVdb 5.0.0 based GSS. There was no statistical significant difference between the standard and ndGSS encoding (paired Wilcox-test: p>0.13). The baseline encoding outperformed HIVdb, and the extended encoding outperformed the baseline encoding (both: p=0.02).



The Logistic regression models were trained on all 4,538 TCEs and tested on the 119 TCEs containing novel drugs. Moreover, the wf was optimized by 5-fold cross validation. Results were compared to a HIVdb 5.0.0 based GSS. For comparison, the baseline (extended) model without ndGSS achieves an AUC of 0.553 (0.601).



Conclusions

No statistically significant difference in predicting response to regimens comprising only old drugs between standard and ndGSS encoding was observed. The ndGSS encodings could predict response to regimens comprising novel drugs better than the GSS alone. Boosting the influence of ndGSS improved classification performance for all methods. That effect was probably related to the increased potency of the novel drugs. Further improvements might be achievable by applying time-matched versions of HIVdb for computing the ndGSS during training, since this would better reflect understanding of novel drugs at that time.

Acknowledgements

The work was supported by the Eu*Resist* project (IST -4- 027173-STP). We thank the Virolab project for contributing data to this study.

References

1. Michal Rosen-Zvi , Andre Altmann, Mattia Prosperi, Ehud Aharoni, Hani Neuvirth, Anders Sönnerborg, Eugen Schülter, Daniel Struck, Yardena Peres, Francesca Incardona, Rolf Kaiser, Maurizio Zazzi and Thomas Lengauer. Selecting anti-HIV therapies based on a variety of genomic and clinical factors. Bioinformatics, 2008, 24 (13):i399-406.
2. Andre Altmann, Michal Rosen-Zvi , Mattia Prosperi, Ehud Aharoni, Hani Neuvirth, Anders Sönnerborg, Eugen Schülter, Joachim Büch, Daniel Struck, Yardena Peres, Francesca Incardona, Rolf Kaiser, Maurizio Zazzi and Thomas Lengauer. Comparison of Classifier Fusion Methods for Predicting Response to Anti HIV-1 Therapy. PLoS ONE, 2008, 3(10): e3470.