



mpi max planck institut
informatik

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



André Altmann

Department of Computational Biology and
applied Algorithmics

Max Planck Institute for Informatics

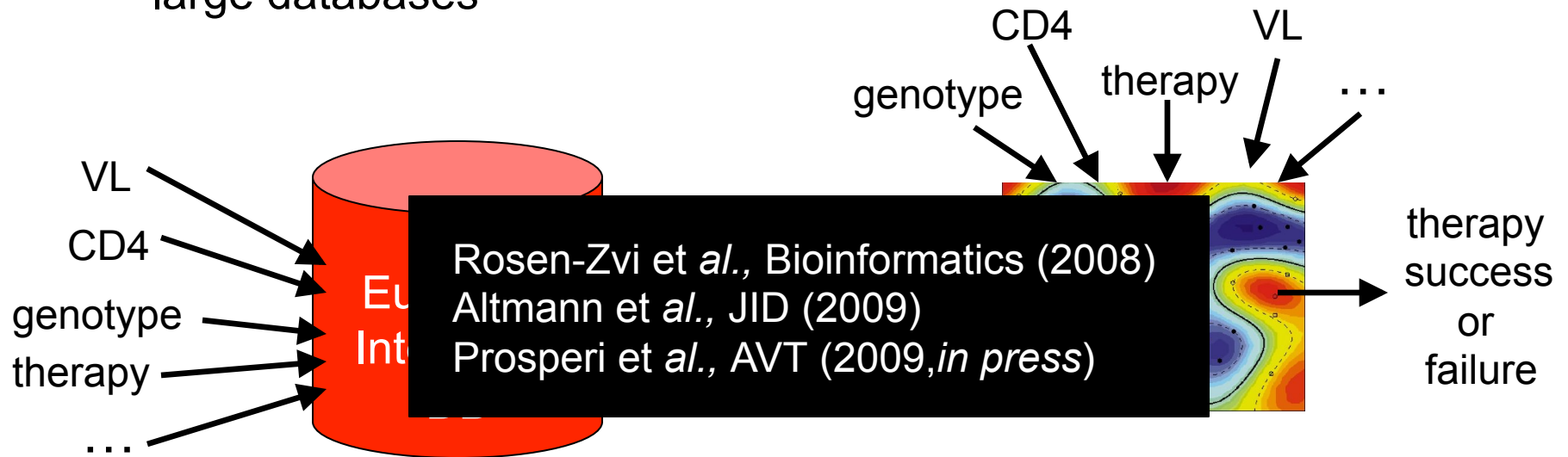
D-66123 Saarbrücken Germany

7th European HIV Drug Resistance Workshop

25-27.03.09, Stockholm, Sweden



- Data-driven models for predicting response to ART learn from large databases



- Due to delayed update of databases, information on treatments with novel drugs is scarce

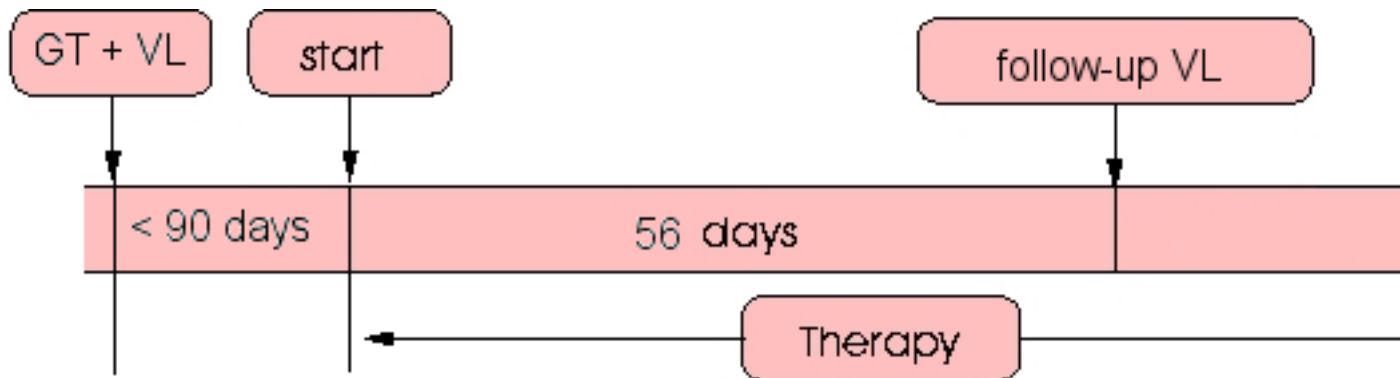
It is hard to extend data-driven models in order to predict response to treatments containing novel drugs



- Idea: Include rules-based rating (e.g. HIVdb) for novel drugs as additional covariate in the data-driven approach
- Problem: Database has still too few observation about novel drugs for training statistical models
- Solution: Treat all drugs in the regimen that were approved by FDA at most 4 years prior to treatment start as novel drugs

	ZDV	ddl	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 <i>novel</i>	1991	1995	1998	1999	2001	2003	2000	2002	1999	1999	2000	2002	2003	2004

- 4,538 treatment change episodes (TCEs) not containing novel drugs were extracted from the *EuResist* and *Virolab* databases



- For validation 119 TCEs containing novel drugs were extracted

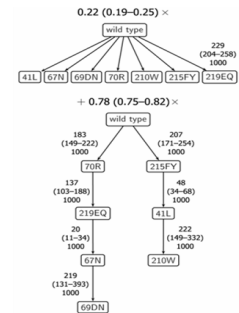
- DRV (n=59)
- TPV (n=52)
- ETR (n=4)
- DRV+ETR (n=4)



■ For every TCE extract information on genotype and treatment

■ Baseline encoding

- Applied drug combination (binary)
- Mutations in viral genotype (binary)
- Genetic Barrier to drug resistance
- Interactions between binary variables



$(1, 0, \dots, 1)$ $(0, 1, \dots, 0, 1, 1)$ $(0.2, 0.0, \dots, 0.9)$

■ Extended encoding

- Baseline encoding
- Treatment history (binary)
- Baseline viral load

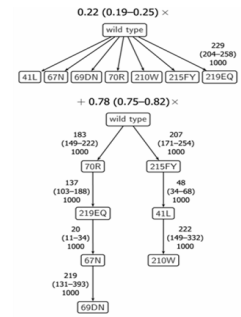
■ Logistic Regression is trained for predicting treatment response



■ For every TCE extract information on genotype and treatment

■ Baseline encoding

- Applied drug combination (binary)
- Mutations in viral genotype (binary)
- Genetic Barrier to drug resistance
- Interactions between binary variables
- **GSS for novel drugs (HIVdb 5.0.0)**



↓ (1,0,...,1) (0,1,...,0,1,1) (0.2,0.0,...,0.9) (1.5)

■ Extended encoding

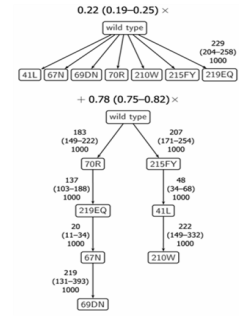
- Baseline encoding
- Treatment history (binary)
- Baseline viral load





■ Logistic Regression is trained for predicting treatment response



Baseline encoding

- Applied drug combination (binary)
- Mutations in viral genotype (binary)
- Genetic Barrier to drug resistance
- Interactions between binary variables
- **GSS for novel drugs (HIVdb 5.0.0)**



 $(1, 0, \dots, 1)$ $(0, 1, \dots, 0, 1, 1)$ $(0.2, 0.0, \dots, 0.9)$ (1.5)

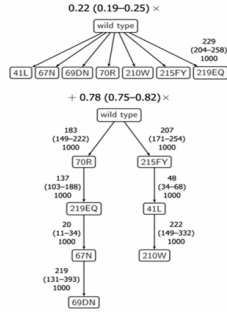
Example:

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



ndGSS encodings - example

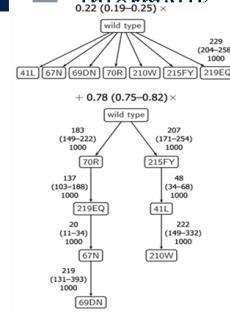
ZDV+3TC+LPV/r treatment in 2006:



$(1, 1, \dots, 1)$ $(0, 1, \dots, 0, 1, 1)$ $(0.2, 0.1, \dots, 0.9)$

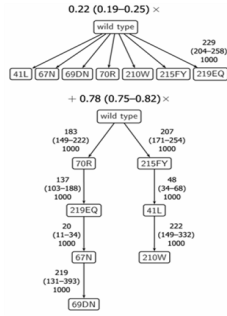


max planck institut
informatik

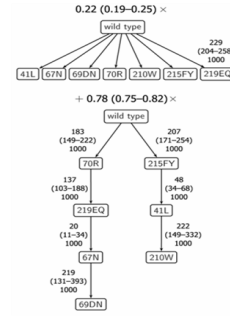


$(1, 1, \dots, 1)$ $(0, 1, \dots, 0, 1, 1)$ $(0.2, 0.1, \dots, 0.9)$ (0.0)

in 2001:



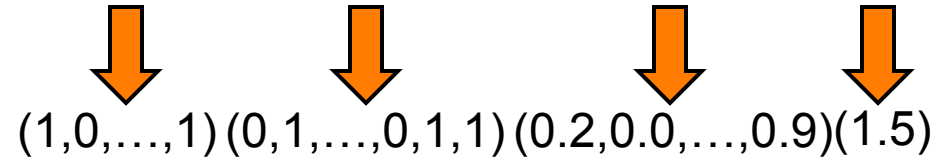
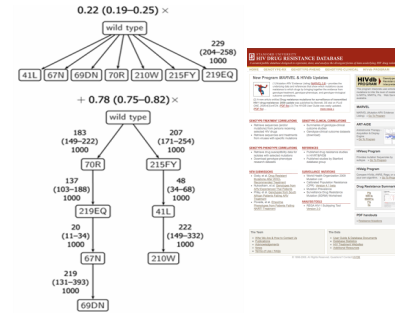
$(1, 1, \dots, 1)$ $(0, 1, \dots, 0, 1, 1)$ $(0.2, 0.1, \dots, 0.9)$



$(1, 1, \dots, 0)$ $(0, 1, \dots, 0, 1, 1)$ $(0.2, 0.1, \dots, 0.0)$ (1.0)

Computation of ndGSS

- Predict activity of all novel drugs with rules-based algorithm
- Compute sum of activities (GSS)

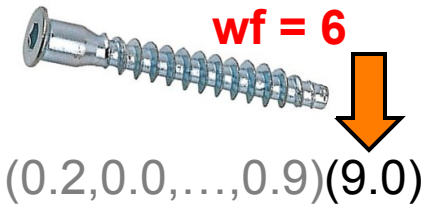
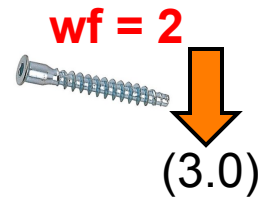


Differentiate between:

- no novel drugs in regimen \Rightarrow 0.0
- all novel drugs are resistant \Rightarrow -1.0

Introduced weighting factor (wf)

- for increasing influence of ndGSS in predictions



Questions:

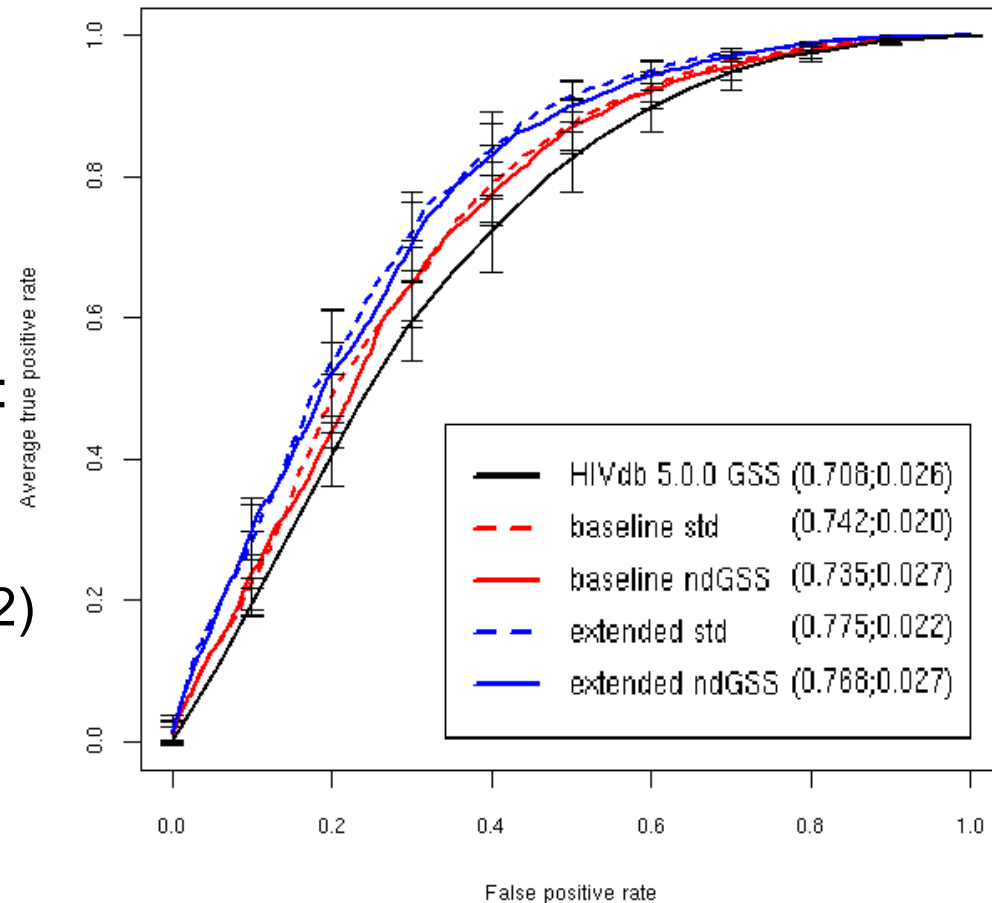
1. Is there a loss in performance when using the ndGSS encoding?
 - Compare cross-validation performance of standard and ndGSS encoding on the training data
2. How reliable is the prediction of regimens containing real novel drugs?
 - Compute performance on the 119 TCEs containing novel drugs
 - Compare results to a rules-based approach (HIVdb 5.0.0)
3. Can performance be improved by increasing influence of novel drugs?
 - Compute performance with different weighting factors



Results

4538 training TCEs

- Performance was computed on 4,538 TCEs using 10-fold cross-validation
- Difference between the standard and ndGSS encoding (paired Wilcoxon-test: $P > 0.13$) not significant
- Extended > Baseline ($P = 0.02$)
- Baseline > HIVdb 5 ($P = 0.02$)



Results

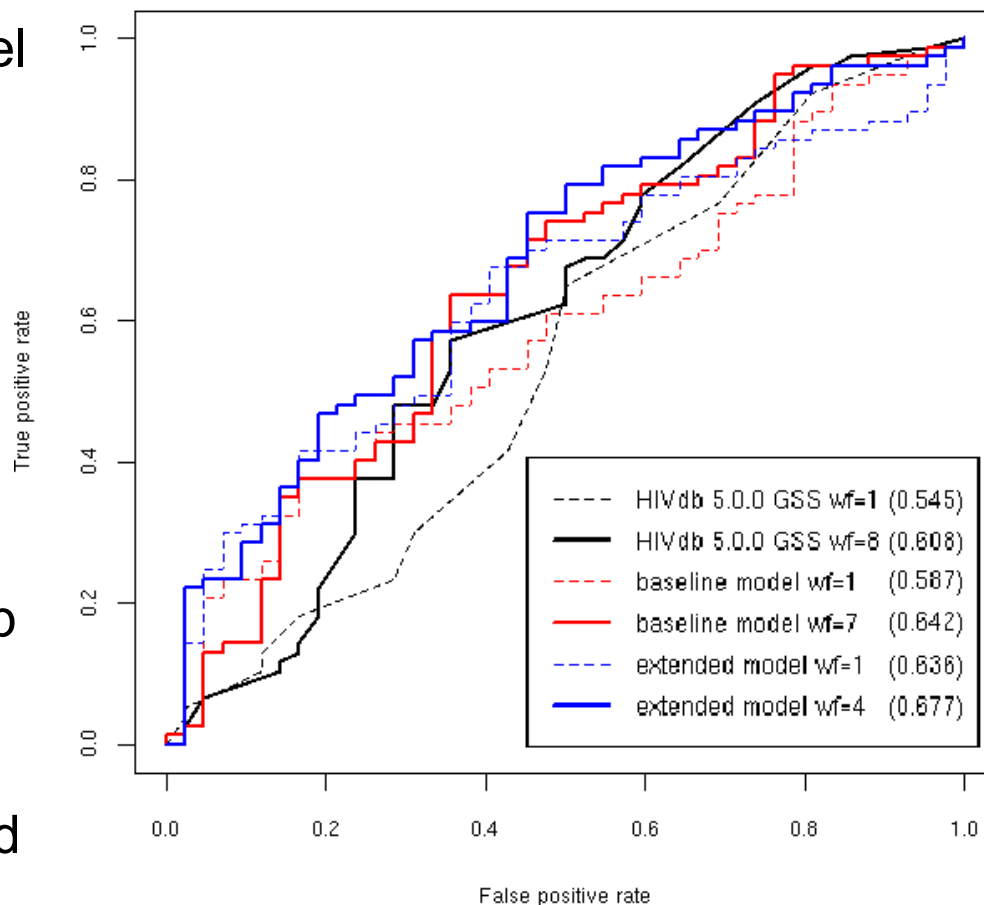
119 TCEs containing DRV/TPV/ETR

■ Performance was computed on 119 TCEs containing novel drugs (42 failures, 77 successes)

■ Performance is moderate
■ but w/o ndGSS: 0.553 / 0.601
i.e. average gain of 0.03 AUC

■ Extended > Baseline > HIVdb

■ Optimization of the wf by 5-fold cross-validation improved AUC by approx 0.06



- It is possible to enable data-driven methods to predict response to ART containing novel drugs
- No statistically significant difference in predicting response to regimens comprising only old drugs
- Boosting influence of ndGSS improved classification performance for all approaches
- Further improvements might be achievable by time-matched versions of HIVdb for computing ndGSS during training



Acknowledgements



Thomas Lengauer

Joachim Büch

Alexander Thielen



Eugen Schülter

Rolf Kaiser



Universität zu Köln

Francesca Incardona



Anders Sönnnerborg



Maurizio Zazzi



Members of
for sharing
data with

