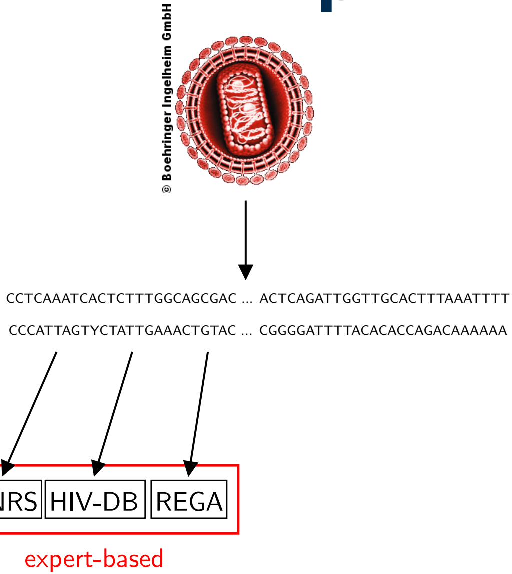
Validation of geno2pheno and THEO on a large independent clinical dataset

André Altmann
Department of Computational Biology
and Applied Algorithmics
Max Planck Institute for Informatics
D-66123 Saarbrücken Germany

5th European HIV Drug Resistance Workshop,
Cascais, Portugal, 28-30 March, 2007

Background

- Genotyping is "standard of care"
- Various methods provide help to genotype interpretation
- Expert-based algorithms
  - ANRS V2006.07
  - HIV-DB
  - Rega V6.4.1
  - ...



© Boehringer Ingelheim GmbH

CCTCAAATCACTCTTTGGCAGCGAC ... ACTCAGATTGGTTGCACTTTAAATTTT

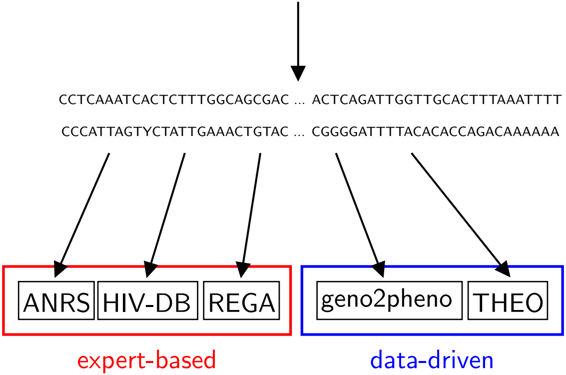
CCCATTAGTYCTATTGAAACTGTAC ... CGGGGATTTTACACACCAGACAAAAAA

expert-based

Background

- Genotyping is “standard of care”
- Various methods provide help to genotype interpretation
- Expert-based algorithms
  - ANRS V2006.07
  - HIV-DB
  - Rega V6.4.1
  - …
- Data-driven algorithms
  - geno2pheno
  - THEO
- Data-driven methods have to be validated on clinical data


© Boehringer Ingelheim GmbH


CCTCAAATCACTCTTTGGCAGCGAC … ACTCAGATTGGTTGCACTTTAAATTTT

CCCATTAGTYCTATTGAAACTGTAC … CGGGGATTTTACACACCAGACAAAAAA

ANRS | HIV-DB | REGA

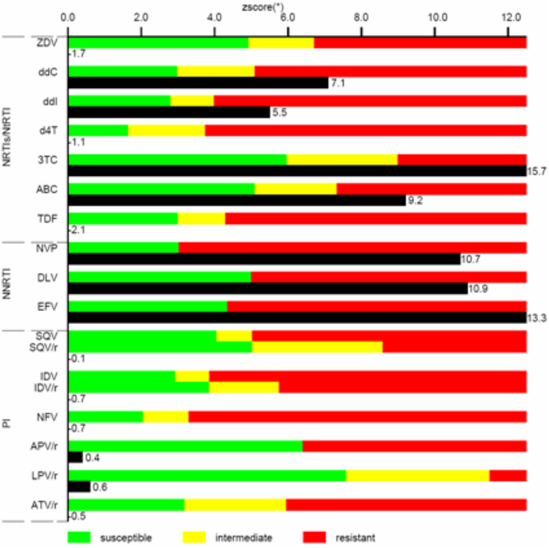
geno2pheno | THEO

expert-based

data-driven

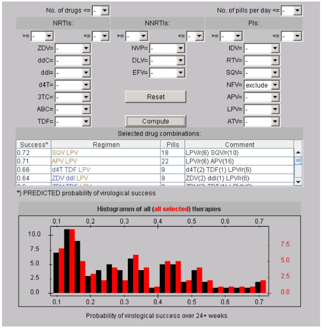
Background - geno2pheno

- geno2pheno$_{\text{[resistance]}}$ predicts phenotypic resistance from genotype
- Trained on ≈ 800 genotype-phenotype pairs per drug
- Expert derived clinical cut-offs
- www.geno2pheno.org
- Details:
  Presentation of H. Walter



Background - THEO

- THErapy Optimizer (THEO) predicts *in vivo* response to combination therapies
- Trained on data from the Stanford HIV Drug Resistance Database and from two Northern California clinic populations
- Computes genetic barrier to drug resistance for every drug in regimen
- Part of the geno2pheno web-service
- *Antiviral Therapy 2007* (**12**) p 169-178:



No. of drugs <= -
No. of pills per day <= -

NRTIs: >= - <= -
NNRTIs: >= - <= -
PIs: >= - <= -

ZDV= -
ddC= -
ddI= -
d4T= -
3TC= -
ABC= -
TDF= -
NVP= -
DLV= -
EFV= -
IDV= -
RTV= -
SQV= -
NFV= exclude
APV= -
LPV= -
ATV= -

Selected drug combinations:

Success*	Regimen	Pills	Comment
0.72	SQV LPV	18	LPV/r(6) SQV/r(10)
0.71	APV LPV	22	LPV/r(6) APV(16)
0.66	d4T TDF LPV	9	d4T(2) TDF(1) LPV/r(6)
0.64	ZDV ddI LPV	9	ZDV(2) ddI(1) LPV/r(6)

*) PREDICTED probability of virological success

Improved prediction of response to antiretroviral combination therapy using the genetic barrier to drug resistance

André Altmann[1], Niko Beerenwinkel[2], Tobias Sing[1], Igor Savenkov[1], Martin Däumer[3], Rolf Kaiser[3], Soo-Yon Rhee[4], W Jeffrey Fessel[5], Robert W Shafer[4] and Thomas Lengauer[1]*

Background - THEO

CCTCAAATCACTCTTTGGCAGCGAC ...

CCCATTAGTYCTATTGAAACTGTAC ...

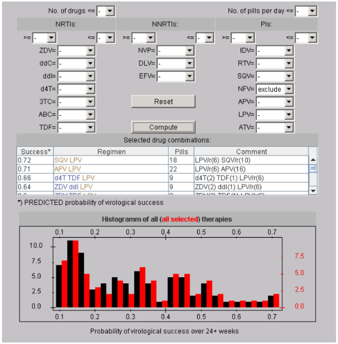
+

d4T TDF LPV

APV LPV

ZDV ddI LPV

SQV LPV



No. of drugs <= -

No. of pills per day <= -

NRTIs: >= - <= -
ZDV= -
ddC= -
ddI= -
d4T= -
3TC= -
ABC= -
TDF= -

NNRTIs: >= - <= -
NVP= -
DLV= -
EFV= -

PIs: >= - <= -
IDV= -
RTV= -
SQV= -
NFV= exclude
APV= -
LPV= -
ATV= -

Selected drug combinations:

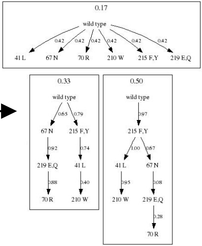
Success*	Regimen	Pills	Comment
0.72	SQV LPV	18	LPV/r(6) SQV/r(10)
0.71	APV LPV	22	LPV/r(6) APV(16)
0.66	d4T TDF LPV	9	d4T(2) TDF(1) LPV/r(6)
0.64	ZDV ddI LPV	9	ZDV(2) ddI(1) LPV/r(6)

*) PREDICTED probability of virological success

Histogramm of all (all selected) therapies

Probability of virological success over 24+ weeks

0.72	SQV LPV
0.71	APV LPV
0.66	d4T TDF LPV
0.64	ZDV ddI LPV



statistical learning

Methods - Data

- Database: EuResist
  Integration of viral genomics with clinical data
  to predict response to anti-HIV treatment
  - ARCA
  - Arevir
  - Karolinska
- Two views to extract data:
  - classic: typical treatment situation
  - alternative: less requirements ⇒ generates more data
- See also:
  - Presentation of M. Zazzi
  - Poster Eu*Resist*: #56

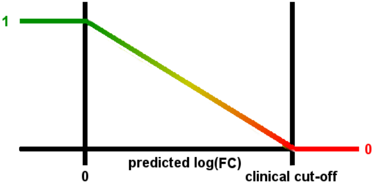
Methods - Treatment scores

Computation of treatment scores:

- Expert-based methods:
  - calculate rating for single drugs
  - sum of single drugscores used

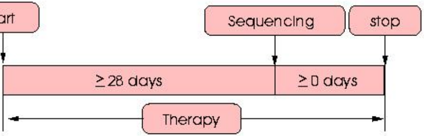
resistant ⇒ 0.0
intermediate ⇒ 0.5
susceptible ⇒ 1.0

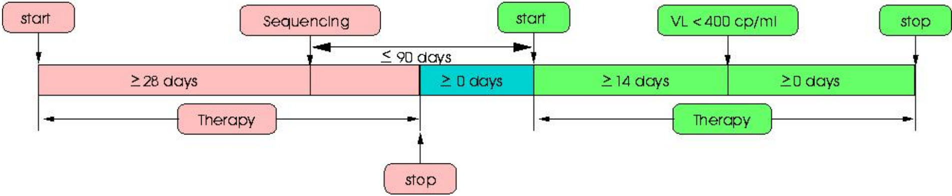
- geno2pheno:
  - calculate fold-change (FC) in IC_{50}
  - normalize using linear interpolation
  - sum of single drugscores used



- THEO directly generates scores for a combination therapy!

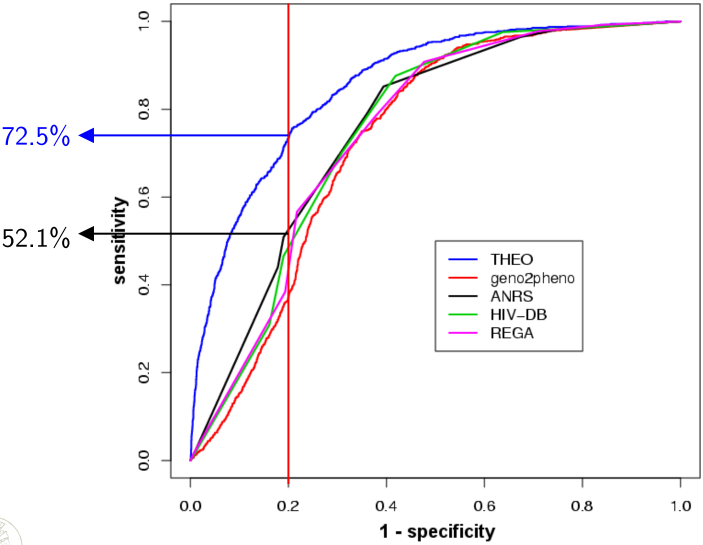
Results - “alternative” Standard Datum





- 5224 treatment-sequence pairs (Eu*Resist* database)
  - 904 successes
  - 4320 failures

Results - “alternative” Standard Datum

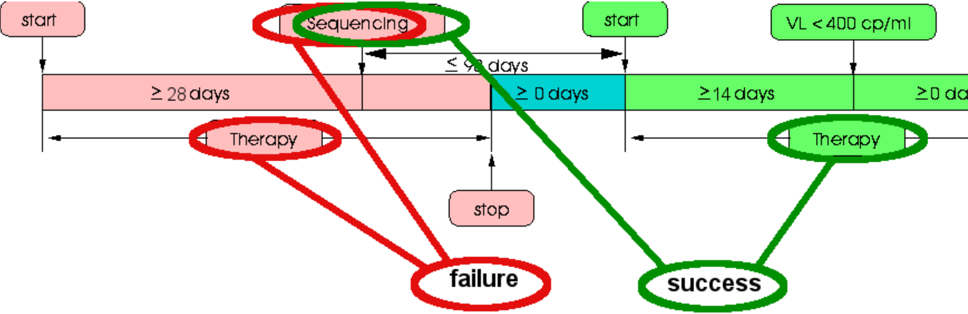


ROC reminder:

Y-Axis: proportion of successes classified as successes

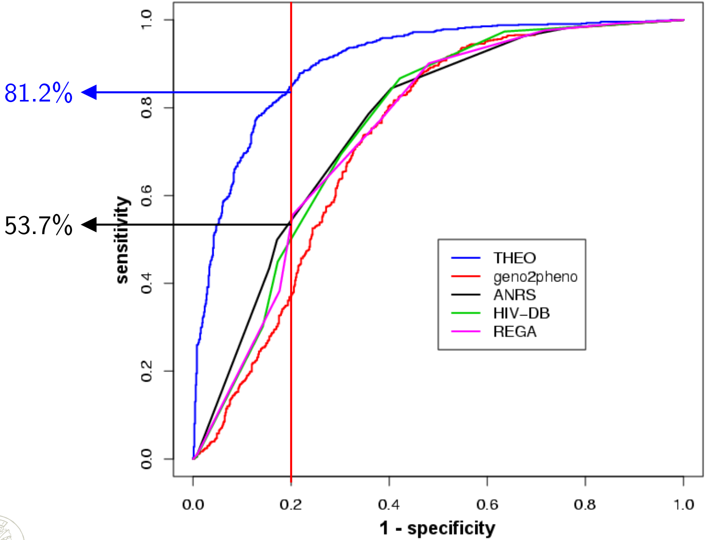
X-Axis: proportion of failures classified as successes

Results - Balanced subset

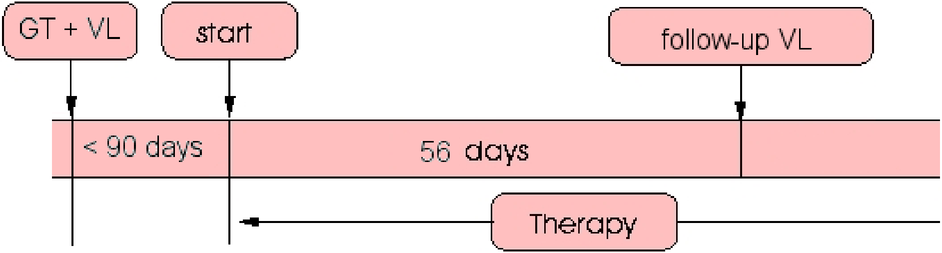


- 1518 treatment-sequence pairs (Eu*Resist* database)
  - 759 sequences
  - 2 therapies per sequence: one failure & one success (same patient)
- perfect dataset to compare algorithms

Results - Balanced subset



Results - “classic” Standard Datum

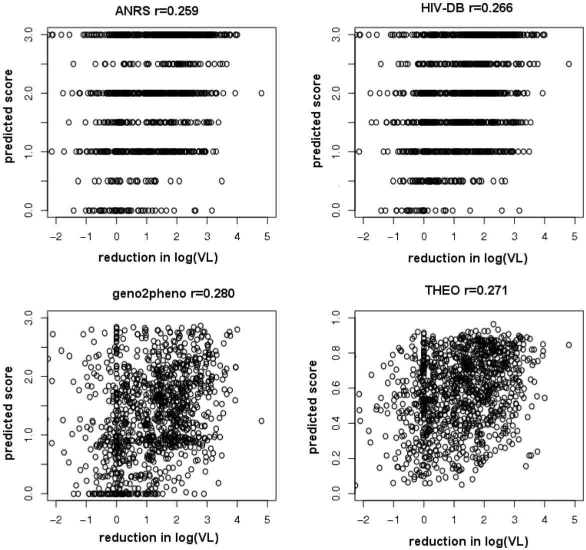


- Focus on *initial response* under *typical treatment* situation
- Success: undetectable (400 cp/ml) or VL reduction of two log
- 1064 treatment-sequence pairs

  ⇒ 614 successes and 450 failures
- Two different tasks:
  - Classification
  - Prediction of $\Delta \log(VL)$

Results - Correlation

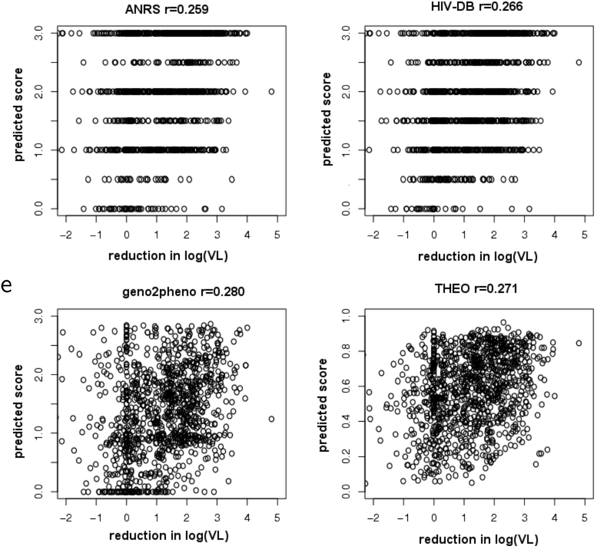
- Method score vs. $\Delta \log(VL)$
- Important information missing:
  $\Rightarrow$ baseline viral load



Results - Correlation

- Method score vs. $\Delta \log(VL)$
- Important information missing:
  $\Rightarrow$ baseline viral load
- Predict $\Delta \log(VL)$ with simple linear model:

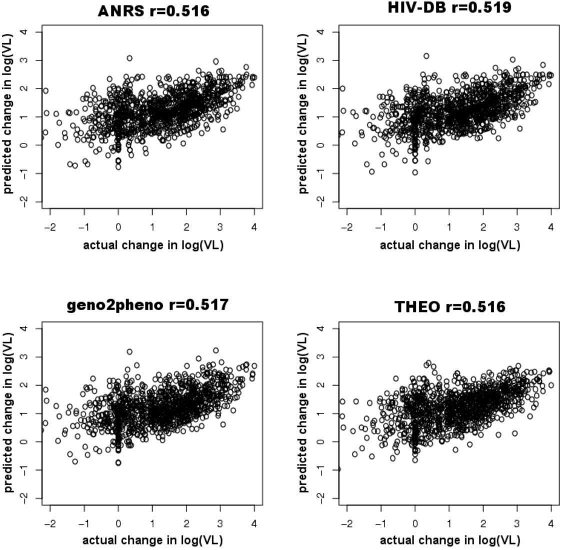
$$\Delta \log(VL) = x_1 * VL + x_2 * \text{score}$$



Results - Correlation

- Method score vs. $\Delta \log(VL)$
- Important information missing:
  $\Rightarrow$ baseline viral load
- Predict $\Delta \log(VL)$ with simple linear model:

$$\Delta \log(VL) = x_1 * VL + x_2 * \text{score}$$

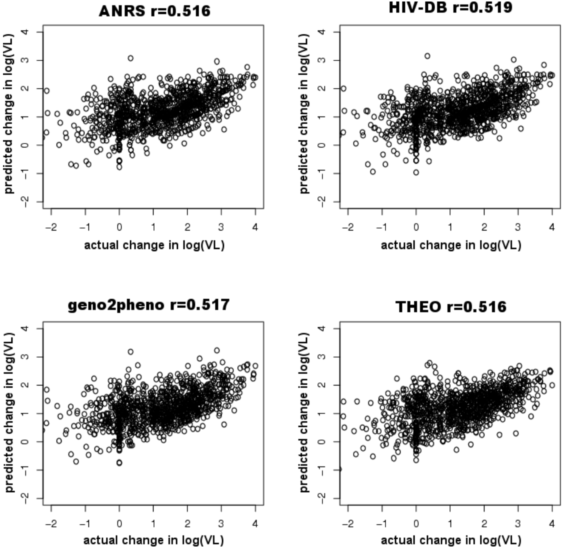


Results - Correlation

- Method score vs. $\Delta \log(VL)$
- Important information missing:
  $\Rightarrow$ baseline viral load
- Predict $\Delta \log(VL)$ with simple linear model:

$$\Delta \log(VL) = x_1 * VL + x_2 * \text{score}$$

- THEO allows to include log(VL) as an additional feature into the statistical model

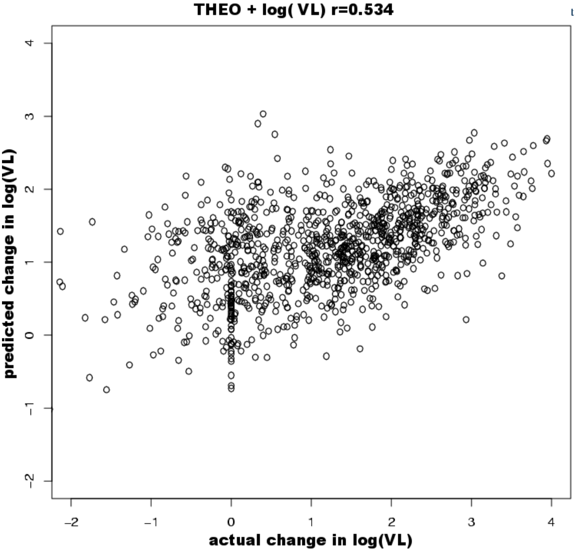


Results - Correlation

- Method score vs. $\Delta \log(VL)$
- Important information missing: $\Rightarrow$ baseline viral load
- Predict $\Delta \log(VL)$ with simple linear model:

$$\Delta \log(VL) = x_1 * VL + x_2 * \text{score}$$

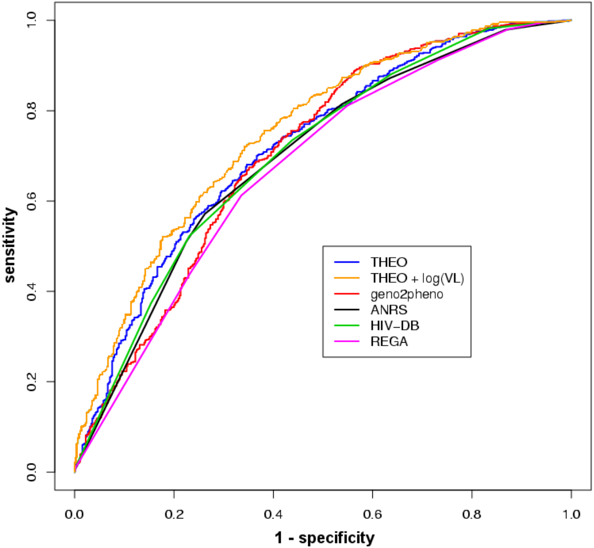
- THEO allows to include log(VL) as an additional feature into the statistical model



$$\boxed{\Delta \log(VL) = 0.58 * \log(VL) + 1.77 * THEO \; - 2.22}$$

Results - Classification

- Back to classification
- Methods perform equally well
- Slight advantage for THEO+log(VL)
- significantly better than ANRS ($p = 0.04883$) REGA ($p = 0.01953$) HIV-DB ($p = 0.04883$)
- pair-wise rank sum test



Conclusions

- PSS guided treatment selection works as well as GSS guided treatment selection
- No difference between methods in *typical treatment* situation
- Completely data-driven approach outperforms GSS/PSS in classification task
- THEO shows ≈ 20%-points improvement in sensitivity at same specificity (80%)
- On balanced subset ≈ 28%-points improvement
- THEO is part of the geno2pheno service and freely available for research purposes at www.geno2pheno.org

Acknowledgments

Tobias Sing Joachim Büch Thomas Lengauer	Max-Planck-Institute for Informatics, Saarbrücken, Germany
Martin Däumer Eugen Schülter Rolf Kaiser	University of Cologne, Germany
Niko Beerenwinkel	Harvard University, Cambridge, MA, USA
Hauke Walter	National Reference Center for Retroviruses, Erlangen, Germany
Rober W Shafer Soo-Yon Rhee	Stanford University, CA, USA
W Jeffrey Fessel	Kaiser-Permanente Medical Care Program Northern California San Francisco, USA
Maurizio Zazzi	University Siena, Italy
Anders Sonnerborg	Karolinska Institute, Sweden
Yardena Peres	IBM Isreal, Haifa, Israel
Francesca Incadorna	Informa S.r.l., Rome, Italy