Clinical relevance of HIV evolutionary pathways and the genetic barrier to drug resistance

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

GMDS 2006, Leipzig
September 11, 2006

Outline

1 The Task

2 Feature Generation

3 Datasets

4 Results

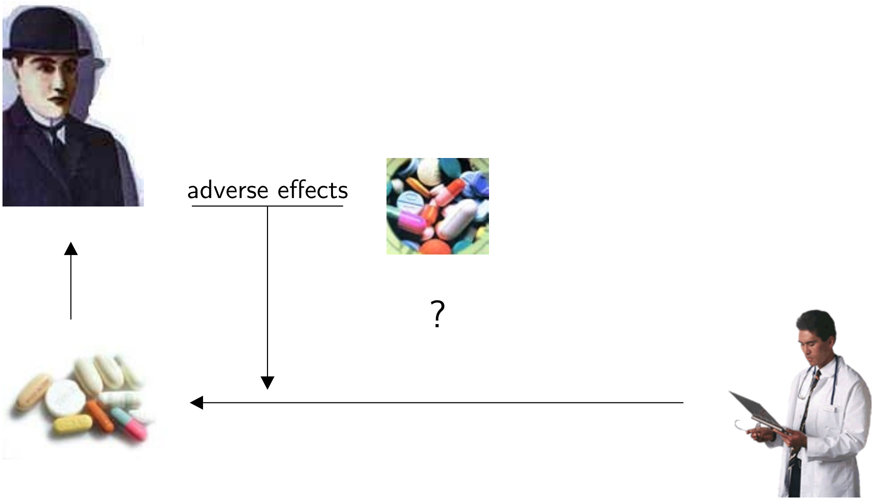
5 Case Study

The Task

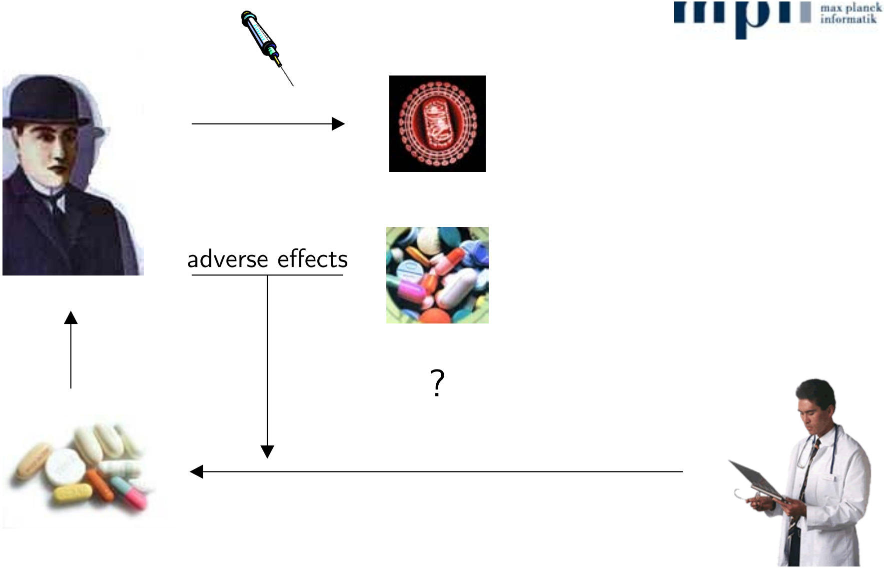
The Task

The Task

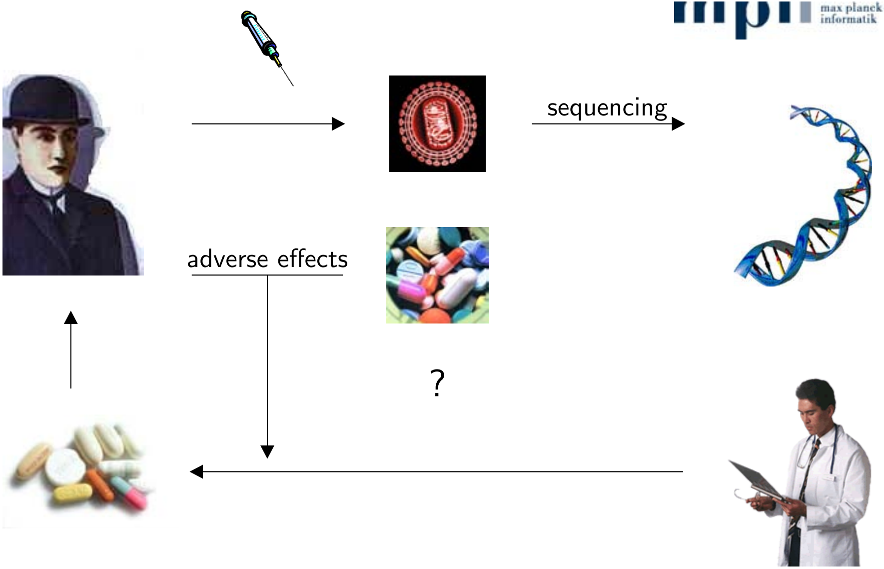
?


adverse effects
?

The Task

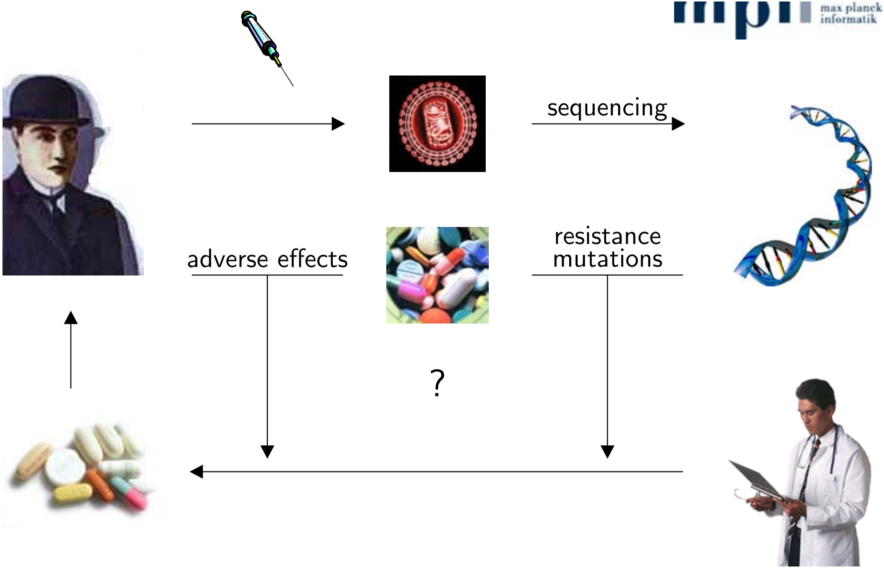


The Task

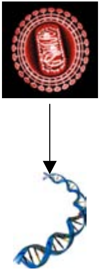


The Task

max planck institut informatik

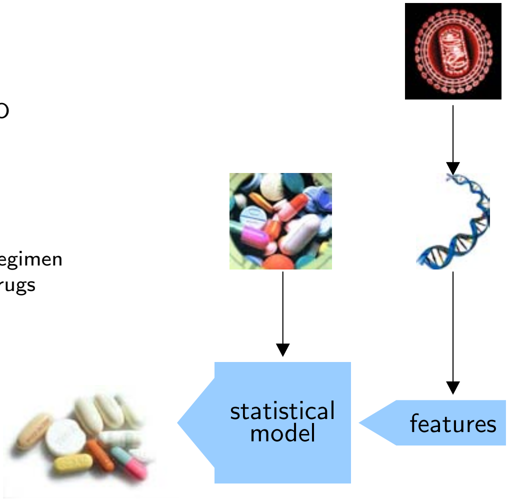


- Optimize therapy outcome
  - given
    - sequences of RT and PRO
    - set of therapies
  - “optimal”
    - therapy success
  - additional knowledge
    - application pattern of a regimen
    - include/exclude certain drugs



The Task

- Optimize therapy outcome
  - given
    - sequences of RT and PRO
    - set of therapies
  - “optimal”
    - therapy success
  - additional knowledge
    - application pattern of a regimen
    - include/exclude certain drugs



The Task

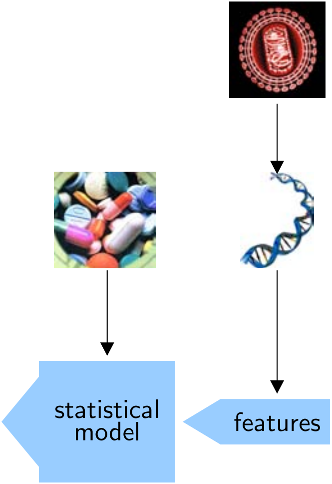
- Optimize therapy outcome
  - given
    - sequences of RT and PRO
    - set of therapies
  - “optimal”
    - therapy success
  - additional knowledge
    - application pattern of a regimen
    - include/exclude certain drugs
- How to create a model?

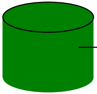
statistical model

features

The Task

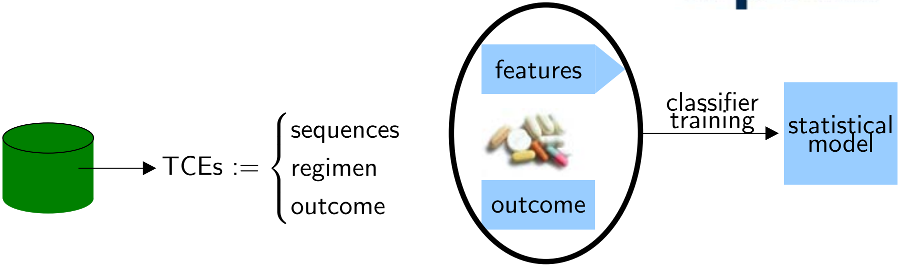
- Optimize therapy outcome
  - given
    - sequences of RT and PRO
    - set of therapies
  - “optimal”
    - therapy success
  - additional knowledge
    - application pattern of a regimen
    - include/exclude certain drugs
- How to create a model?
- How to validate results?



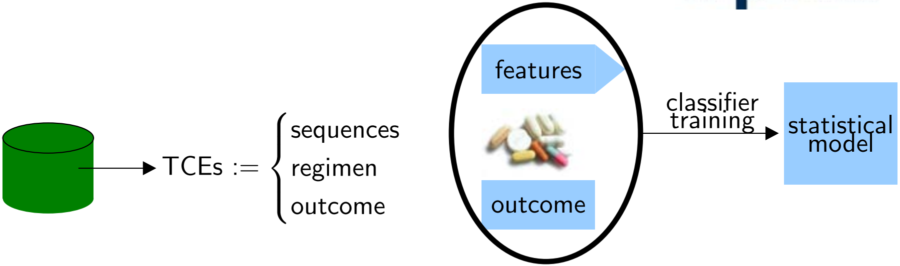


$$\text{TCEs} := \begin{cases} \text{sequences} \\ \text{regimen} \\ \text{outcome} \end{cases}$$

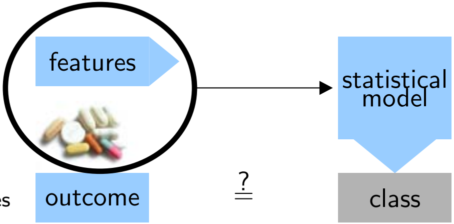
The Task


TCEs := sequences
regimen
outcome
features
outcome
classifier training
statistical model

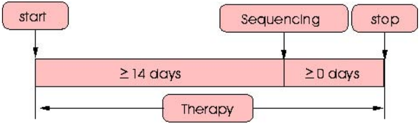
The Task



- Validation
  - count misclassification
    - outcome $\neq$ class
  - true positives
    - recognized successes
  - false positives
    - failures recognized as successes

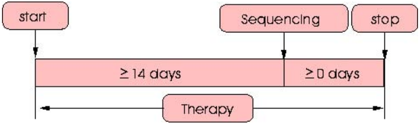


- Definition of therapy failure and success
  - failure

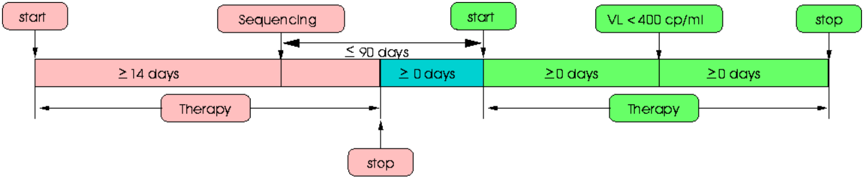


The Task

- Definition of therapy failure and success
  - failure



  - success



- Indicators for
  - Resistance associated Mutations [Johnson, V.A., et al. (2005) Top. HIV Med.]

- Indicators for
  - Resistance associated Mutations [Johnson, V.A., et al. (2005) Top. HIV Med.]

CCTCAAATCACTCTTTGGCAGCGAC … ACTCAGATTGGTTGCACTTTAAATTTT

PRO

- Indicators for
  - Resistance associated Mutations [Johnson, V.A., et al. (2005) Top. HIV Med.]

PRO

CCTCAAATCACTCTTTGGCAGCGAC ... ACTCAGATTGGTTGCACTTTAAATTTT

10 F,I,R,V   16 E   ...   90 M   93 L

$\Downarrow$

- Indicators for
  - Resistance associated Mutations [Johnson, V.A., et al. (2005) Top. HIV Med.]

CCTCAAATCACTCTTTGGCAGCGAC ... ACTCAGATTGGTTGCACTTTAAATTTT

	10 F,I,R,V	16 E	...	90 M	93 L	
PRO →	1	0	...	1	0	pro I

max planck institut informatik

- Indicators for
  - Resistance associated Mutations [Johnson, V.A., et al. (2005) Top. HIV Med.]

PRO →

CCTCAAATCACTCTTTGGCAGCGAC		...	ACTCAGATTGGTTGCACTTTAAATTTT		
10 F,I,R,V	16 E	...	90 M	93 L	
		⇓			
1	0	...	1	0	pro I

RT →

CCCATTAGTYCTATTGAAACTGTACC		...	GGGGATTTTACACACCAGACAAAAAA		
41 L	44 D	...	215 Y,F	219 Q,E	
		⇓			
1	0	...	1	1	rt I

  - 49 indicator variables

- Indicators for
  - Resistance associated Mutations [Johnson, V.A., et al. (2005) Top. HIV Med.]

PRO →

CCTCAAATCACTCTTTGGCAGCGAC		...	ACTCAGATTGGTTGCACTTTAAATTTT	
10 F,I,R,V	16 E	...	90 M	93 L
		⇓		
1	0	...	1	0

pro I

RT →

CCCATTAGTYCTATTGAAACTGTACC		...	GGGGATTTTACACACCAGACAAAAAA	
41 L	44 D	...	215 Y,F	219 Q,E
		⇓		
1	0	...	1	1

rt I

- 49 indicator variables

- Drugs

ZDV	ddC	ddI	d4T	3TC	ABC	TDF
1	0	0	0	1	0	0
NVP	DLV	EFV				
0	0	0				
SQV	IDV	RTV	NFV	APV	LPV	ATV
0	0	0	0	0	1	0

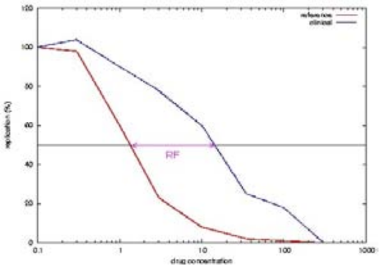
d I

- 17 indicator variables

Feature Generation - Phenotype

- Resistance factor (RF)
  - fold increase in drug concentration needed to cut the replication rate in half (compared to WT)

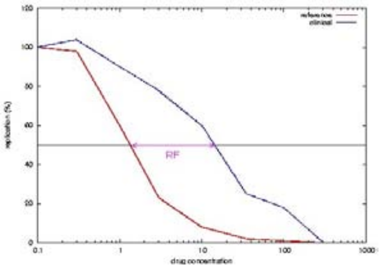
$$RF = \frac{IC_{50}(\text{clinical sample})}{IC_{50}(\text{wild type})}$$



max planck institut informatik

- Resistance factor (RF)
  - fold increase in drug concentration needed to cut the replication rate in half (compared to WT)

$$RF = \frac{IC_{50}(\text{clinical sample})}{IC_{50}(\text{wild type})}$$



- Predicted resistance factor

phenotype

  - use geno2pheno$_{[resistance]}$

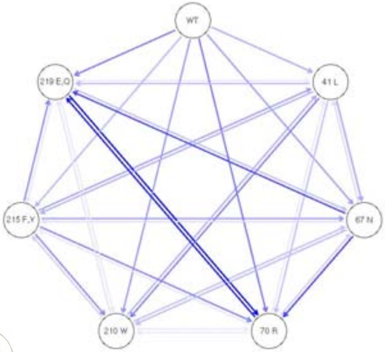
ZDV	ddC	ddI	d4T	3TC	ABC	TDF
1.92572	0.23869	0.49570	0.39141	0.87097	0.54309	0.68443
NVP	DLV	EFV				
1.15549	1.03159	0.74470				
SQV	IDV	RTV	NFV	APV	LPV	ATV
1.51242	1.24985	1.32110	1.35764	0.51961	0.59543	1.06091

max planck institut informatik

- Estimate model from cross-sectional data:

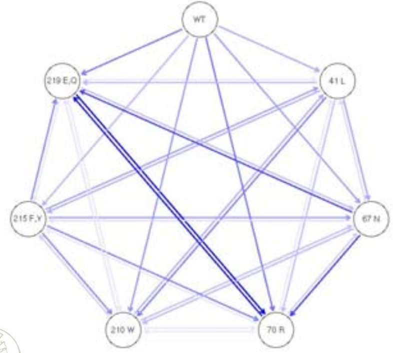
- Estimate model from cross-sectional data:
- Build complete digraph with weights:

$$w(u,v) = \log\left(\frac{p(u,v)}{p(u)p(v)} \times \frac{p(u)}{p(u)+p(v)}\right)$$

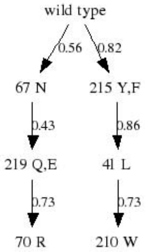


- Estimate model from cross-sectional data:
- Build complete digraph with weights:

$$w(u, v) = \log \left(\frac{p(u,v)}{p(u)p(v)} \times \frac{p(u)}{p(u)+p(v)} \right)$$



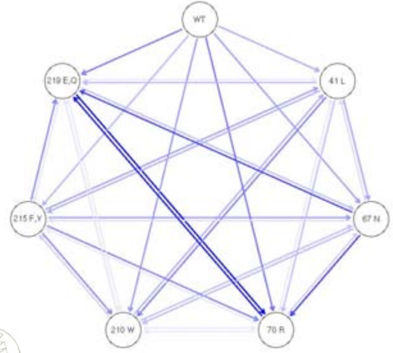
- Find maximum weighted branching (Edmond's algorithm)



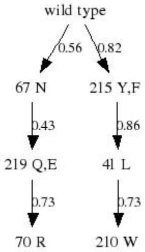
Feature Generation - Evolutionary Pathways

- Estimate model from cross-sectional data:
- Build complete digraph with weights:

$$w(u,v) = \log\left(\frac{p(u,v)}{p(u)p(v)} \times \frac{p(u)}{p(u)+p(v)}\right)$$



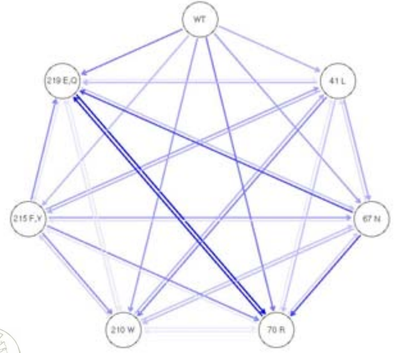
- Find maximum weighted branching (Edmond's algorithm)



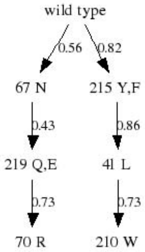
- Problem:
  - many patterns with probability 0

- Estimate model from cross-sectional data:
- Build complete digraph with weights:

$$w(u,v) = \log\left(\frac{p(u,v)}{p(u)p(v)} \times \frac{p(u)}{p(u)+p(v)}\right)$$

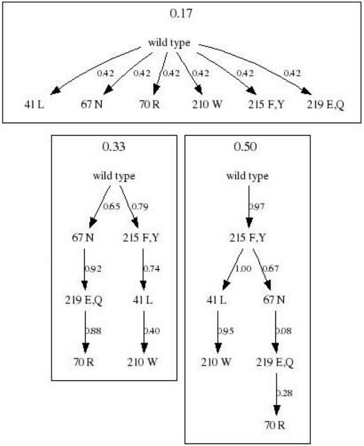


- Find maximum weighted branching (Edmond's algorithm)

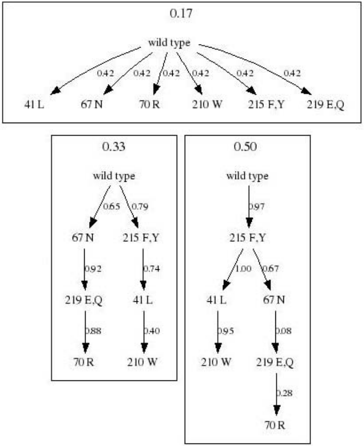


- Problem:
  - many patterns with probability 0
- Solution:
  - mixtures of mutagenetic trees
  - EM-like learning algorithm

- Mutagenetic trees model evolution of drug resistance in HIV
  - estimated on mutational patterns derived from sequences in failing TCEs
  - Beerenwinkel, N. *et al.* (2005) *J. Comput. Biol.*



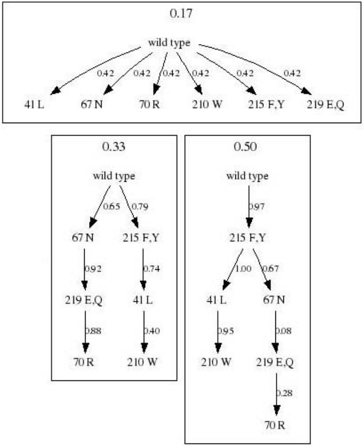
- Mutagenetic trees model evolution of drug resistance in HIV
  - estimated on mutational patterns derived from sequences in failing TCEs
  - Beerenwinkel, N. *et al.* (2005) *J. Comput. Biol.*



- Genetic Progression Score GPS
  - estimate waiting time for a mutational pattern

max planck institut informatik

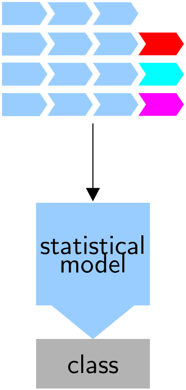
- Mutagenetic trees model evolution of drug resistance in HIV
  - estimated on mutational patterns derived from sequences in failing TCEs
  - Beerenwinkel, N. *et al.* (2005) *J. Comput. Biol.*



- Genetic Progression Score **GPS**
  - estimate waiting time for a mutational pattern
- Genetic Barrier **barrier**
  - probability of NOT reaching any resistant state

  = probability of remaining susceptible to a drug in future
  - Beerenwinkel, N. *et al.* (2005) *J. Infect. Dis.*

- Classification is based on:
  - baseline features
  - baseline + predicted RF (phenotype)
  - baseline + genetic barrier
  - baseline + GPS



- Various methods for classification
  - Linear Discriminant Analysis (LDA)
  - Support Vector Machines (SVM)
  - Linear Logistic Regression
  - Decision Trees (C4.5)
  - Logistic Model Trees (LMT)

- Arevir
  - provides genotype-phenotype data
    - used to train SVR (g2p)
    - identify resistant states in mutagenetic trees
  - contains 776 TCEs
    - 552 failures
    - 224 successes (121 treatment naïve)

Datasets

Arevir

- provides genotype-phenotype data
  - used to train SVR (g2p)
  - identify resistant states in mutagenetic trees
- contains 776 TCEs
  - 552 failures
  - 224 successes (121 treatment naïve)

Stanford (A) - all

- no genotype-phenotype data!
- contains 6, 337 TCEs
  - 1, 561 successes
  - 4, 776 failures (used to estimate mutagenetic trees)

- Arevir
  - provides genotype-phenotype data
    - used to train SVR (g2p)
    - identify resistant states in mutagenetic trees
  - contains 776 TCEs
    - 552 failures
    - 224 successes (121 treatment naïve)

- Stanford (A) - all
  - no genotype-phenotype data!
  - contains 6,337 TCEs
    - 1,561 successes
    - 4,776 failures (used to estimate mutagenetic trees)

- Stanford (BS) - balanced sequences
  - contains 2,728 TCEs
    - one failure and one success for the sample sequence

Arevir

- provides genotype-phenotype data
  - used to train SVR (g2p)
  - identify resistant states in mutagenetic trees
- contains 776 TCEs
  - 552 failures
  - 224 successes (121 treatment naïve)

Stanford (A) - all

- no genotype-phenotype data!
- contains $6,337$ TCEs
  - $1,561$ successes
  - $4,776$ failures (used to estimate mutagenetic trees)

Stanford (BS) - balanced sequences

- contains $2,728$ TCEs
  - one failure and one success for the sample sequence

Stanford (BT) - balanced therapies

- contains $2,436$ TCEs
  - same number of successes and failures for 321 different therapies

Arevir

- provides genotype-phenotype data
  - used to train SVR (g2p)
  - identify resistant states in mutagenetic trees
- contains 776 TCEs
  - 552 failures
  - 224 successes (121 treatment naïve)

Stanford (A) - all

- no genotype-phenotype data!
- contains 6,337 TCEs
  - 1,561 successes
  - 4,776 failures (used to estimate mutagenetic trees)

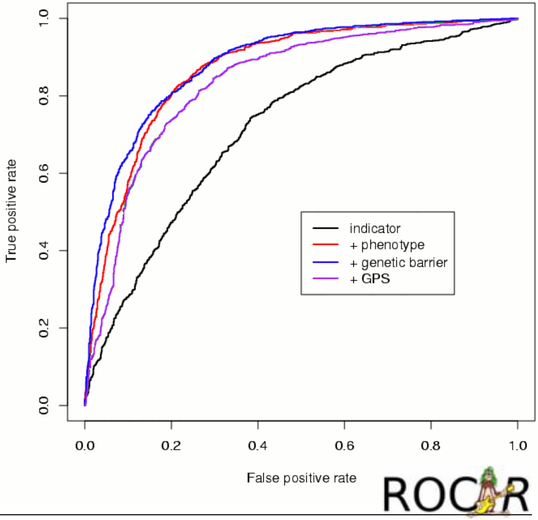
Stanford (BS) - balanced sequences

- contains 2,728 TCEs
  - one failure and one success for the sample sequence

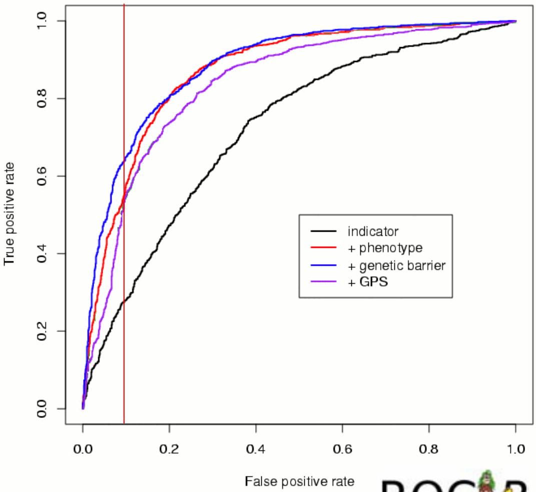
Stanford (BT) - balanced therapies

- contains 2,436 TCEs
  - same number of successes and failures for 321 different therapies

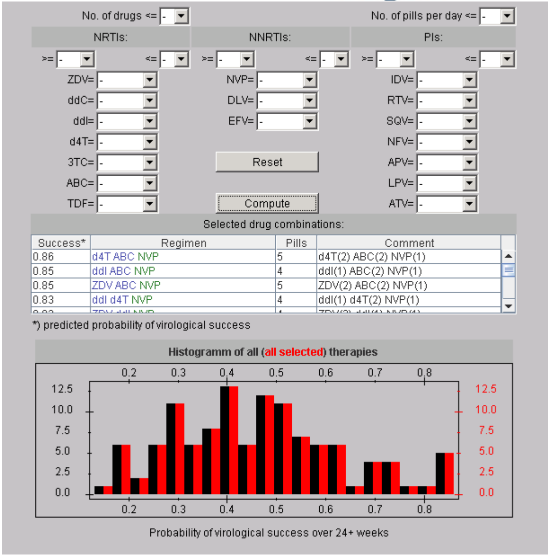
- Performance analyses based on receiver operator characteristics (ROC) curves
- "The more in the upper left corner, the better the model"
- LMT classifier estimated on Stanford (BT)
- Using 10-fold cross validation
- ALL extensions of the baseline improve performance significantly ($p < 0.02$ for GPS)



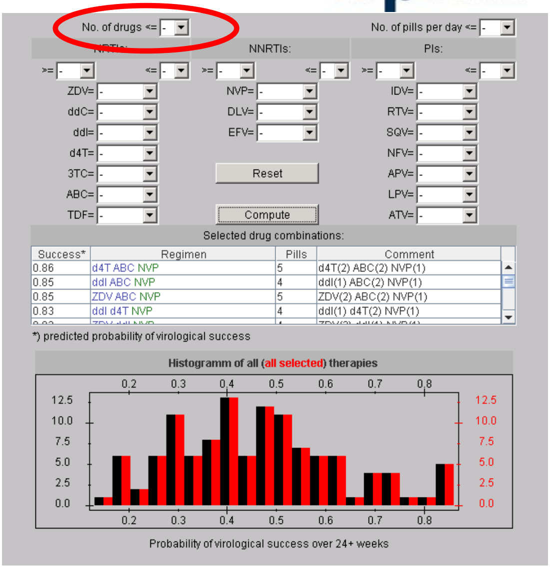
- Performance analyses based on receiver operator characteristics (ROC) curves
- "The more in the upper left corner, the better the model"
- LMT classifier estimated on Stanford (BT)
- Using 10-fold cross validation
- ALL extensions of the baseline improve performance significantly ($p < 0.02$ for GPS)



- THErapy Optimization

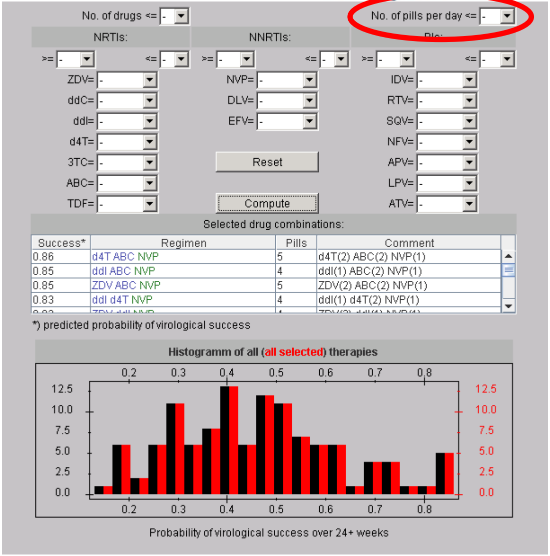


- THErapy Optimization
  - limit no. of drugs

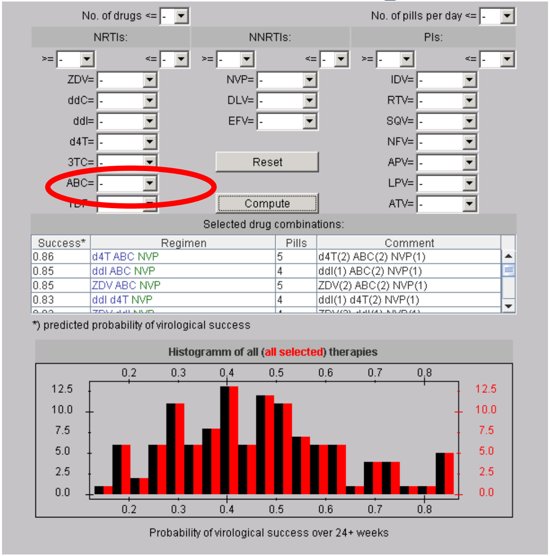


THErapy Optimization

- limit no. of drugs
- limit daily burden

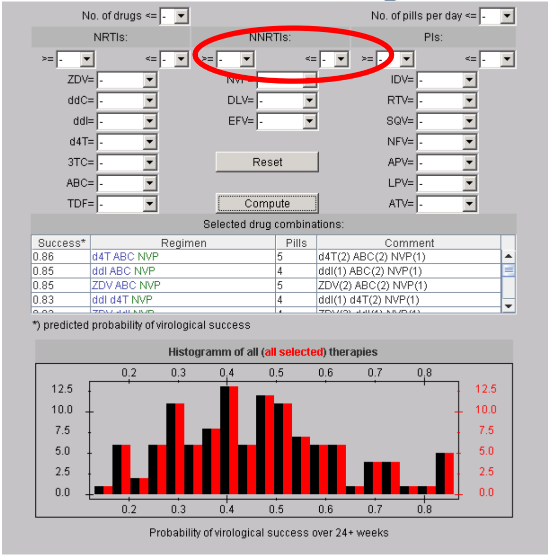


- THErapy Optimization
  - limit no. of drugs
  - limit daily burden
  - include/exclude drugs



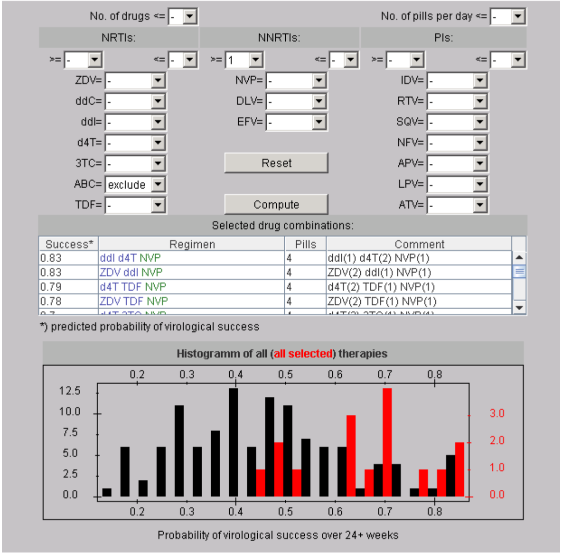
Results - THEO applet

- **THE**rapy **O**ptimization
  - limit no. of drugs
  - limit daily burden
  - include/exclude drugs
  - set number of drugs per class



Results - THEO applet

- THErapy Optimization
  - limit no. of drugs
  - limit daily burden
  - include/exclude drugs
  - set number of drugs per class



max planck institut informatik

- L-V *1957 (teacher)
- HIV pos. first diagnosed in 1988
- ART since NOV 1995
  - Compliance: highly motivated
  - Side effects: none

- L-V *1957 (teacher)
- HIV pos. first diagnosed in 1988
- ART since NOV 1995
  - Compliance: highly motivated
  - Side effects: none
- Therapy history
  - since NOV 1995:
    AZT+ddC, AZT+3TC, AZT+3TC+IDV, d4T+DLV+NFV, d4T+ddI+3TC+NVP+IDV+RTV, d4T+ddI+3TC+ABC+NFV+DMP, AZT+3TC+ABC+EFV, d4T+ddI+3TC+ABC+IDV/r
  - since JAN 2001: d4T + ddI + 3TC + ABC + APV/r

- L-V *1957 (teacher)
- HIV pos. first diagnosed in 1988
- ART since NOV 1995
  - Compliance: highly motivated
  - Side effects: none
- Therapy history
  - since NOV 1995:
    AZT+ddC, AZT+3TC, AZT+3TC+IDV, d4T+DLV+NFV, d4T+ddI+3TC+NVP+IDV+RTV, d4T+ddI+3TC+ABC+NFV+DMP, AZT+3TC+ABC+EFV, d4T+ddI+3TC+ABC+IDV/r
  - since JAN 2001: d4T + ddI + 3TC + ABC + APV/r
- Viral Load (RNA cp/ml):
  - 22.01.2003: 1.851
  - 02.04.2003: 1.705
  - 10.09.2003: 8.751 ⇒ Resistance testing

Resistance Testing

PRO	RT
L10F	M41L
M46I	E44D
M46L	S68G
I54M	K103N
L63P	V118I
A71V	184V
V82A	L210W
V3I	T215Y
I15V	D121H
S37N	I135T
R57K	D177E
D60E	I178L
Q61E	E203D
I62V	Q207E
I72T	L214F
L76V	R211K
I93L	I293V

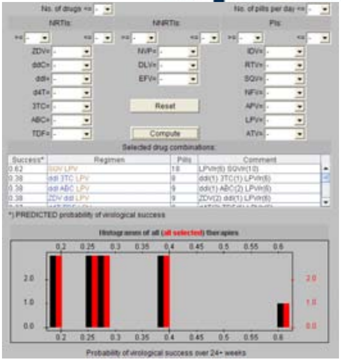
Resistance Testing

PRO	RT
L10F	M41L
M46I	E44D
M46L	S68G
I54M	K103N
L63P	V118I
A71V	184V
V82A	L210W
V3I	T215Y
I15V	D121H
S37N	I135T
R57K	D177E
D60E	I178L
Q61E	E203D
I62V	Q207E
I72T	L214F
L76V	R211K
I93L	I293V

Resistance Testing

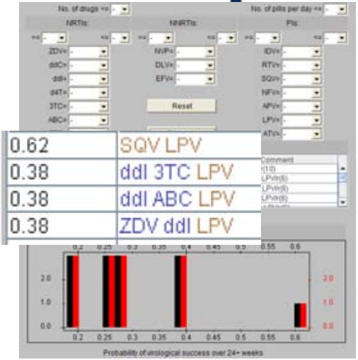
PRO	RT
L10F	M41L
M46I	E44D
M46L	S68G
I54M	K103N
L63P	V118I
A71V	184V
V82A	L210W
V3I	T215Y
I15V	D121H
S37N	I135T
R57K	D177E
D60E	I178L
Q61E	E203D
I62V	Q207E
I72T	L214F
L76V	R211K
I93L	I293V

input →



Resistance Testing

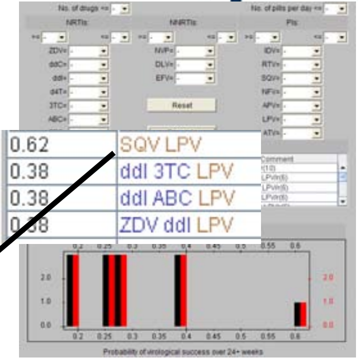
PRO	RT
L10F	M41L
M46I	E44D
M46L	S68G
I54M	K103N
L63P	V118I
A71V	184V
V82A	L210W
V3I	T215Y
I15V	D121H
S37N	I135T
R57K	D177E
D60E	I178L
Q61E	E203D
I62V	Q207E
I72T	L214F
L76V	R211K
I93L	I293V

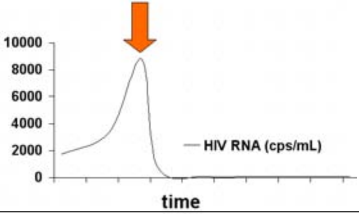


Case Study

Resistance Testing

PRO	RT
L10F	M41L
M46I	E44D
M46L	S68G
I54M	K103N
L63P	V118I
A71V	184V
V82A	L210W
V3I	T215Y
I15V	D121H
S37N	I135T
R57K	D177E
D60E	I178L
Q61E	E203D
I62V	Q207E
I72T	L214F
L76V	R211K
I93L	I293V





Viral Load (RNA cp/ml):

10.09.2003	8.751
31.10.2003	729
16.01.2004	<50
15.04.2004	<50
13.08.2004	<50
28.10.2004	<50

Acknowledgments

Niko Beerenwinkel	Program for Evolutionary Dynamics, Harvard University
Tobias Sing Jörg Rahnenführer Thomas Lengauer	MPI für Informatik, Saarbrücken
Rolf Kaiser Martin Däumer Melanie Balduin Saleta Sierra-Aragon Dörte Hammerschmidt	Institute of Virology, University of Cologne
Daniel Hoffmann	ZMB, Bioinformatik, University Duisburg-Essen
Joachim Selbig	MPI of Molecular Plant Physiology, Golm
Barbara Schmidt Hauke Walter Klaus Korn	Institute of Clinical and Molecular Virology, German National Reference Center for Retrovriuses, University of Erlangen-Nürnberg
Marc Oette	Dept. of Gastroenterology, University of Düsseldorf
Gerd Fätkenheuer	Dept. of Internal Medicine I, University of Cologne
Jürgen Rockstroh	Dept. of Internal Medicine I, University of Bonn
Thomas Berg	Medical Laboratory, Berlin
Patrick Braun	PZB, Aachen

END