

IST-2004-027173

Multiple modeling techniques for prediction of response to treatment in HIV patients

Francesca Incardona (Informa s.r.l.)



EuResist objective: to support clinicians treating HIV patients

The *EuResist* project aims at developing an integrated system for prediction of response to antiretroviral treatment

Novel approach: viral genotype data integrated with clinical data.

Focus is on genotype -
response correlation

A critical amount of
resistance data is needed.

An integrated and comprehensive genotype-response database has been created.

Several distinct prediction engines are under study to be combined in the *EuResist* Prediction System.



Started: January 1st 2006
Will end: June 30th 2008



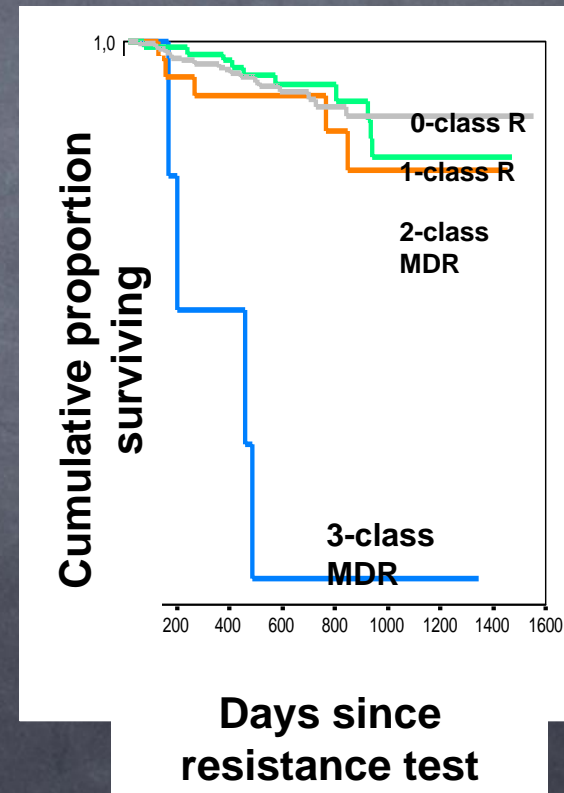
The nature and dimension of the problem

Effective treatment available since 1996: **HAART**
but the virus is not eradicated →

DRUG RESISTANCE

(virus variants with reduced susceptibility to anti-HIV drugs)

- Resistance to drugs of a single class:
NRTI 50-70%, NNRTI 35-45%,
PI 35-55%.
- Resistance to drugs of three classes (MDR): 5-25%.
- Resistance to any drug in newly infected subjects ("transmitted resistance"): 5-25%



Broad resistance to 3 classes is an independent predictor of death

Different approaches

- A genotype is the array of mutations found (direct sequencing of the relevant parts of HIV genome: PR and RT):

PR: 10I 35D 36I 37D 46I 54V 57K 63P 71V 82A 84V 90M

RT: 67N 70R 177Q 196E 207E 215F 219Q

some are involved in resistance, others are 'polymorphisms'

We want to know when a mutation is involved in resistance :

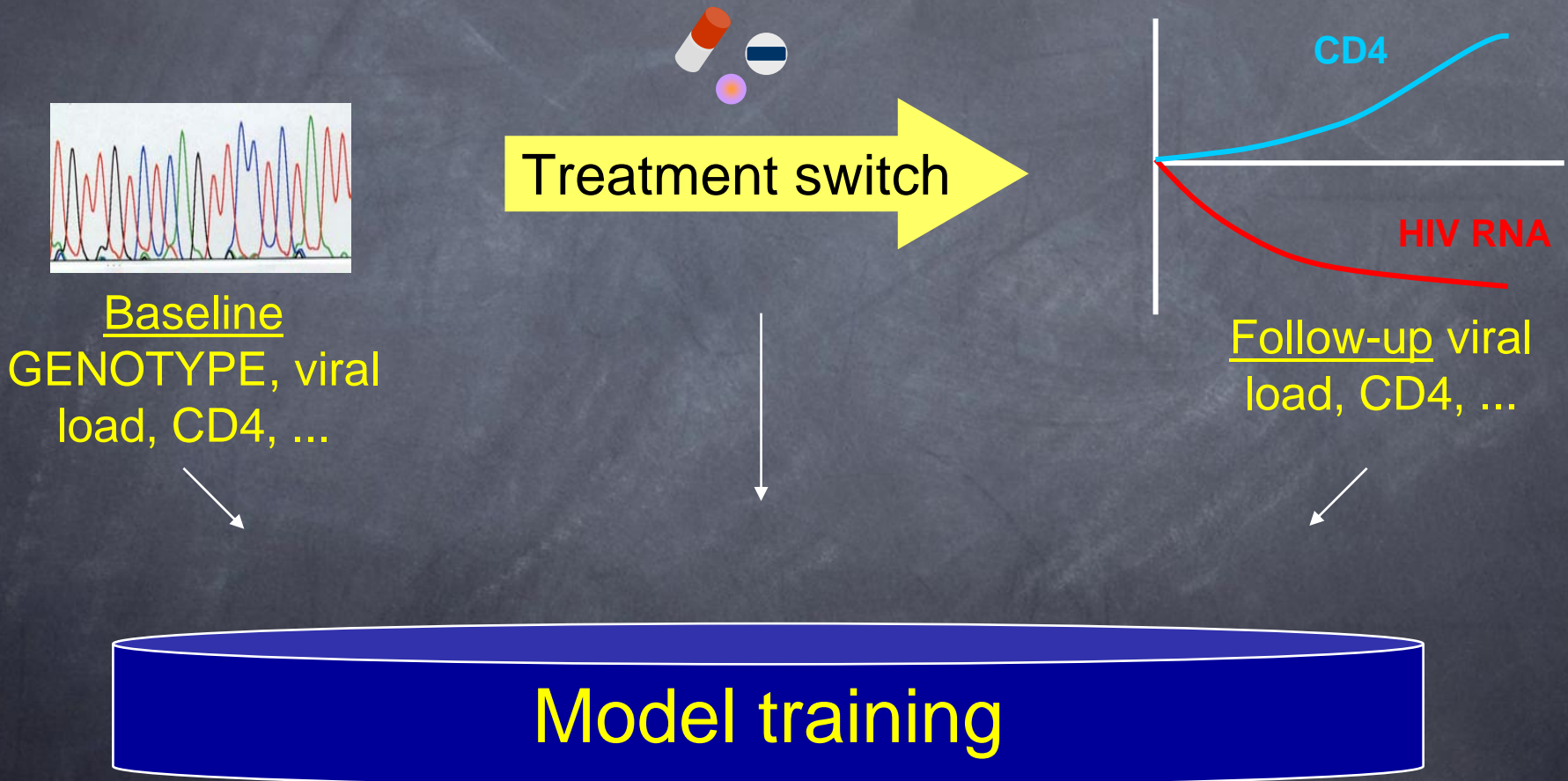
- **Correlations between genotype and phenotype.** We can observe the behaviour of mutant virus in the presence of drug in vitro.
- **Correlations between genotype and treatment history.** We can find that certain mutations are more frequent in vivo following exposure to certain drugs.

- **Correlations between genotype and response** to treatment. We can find that certain mutations decrease the effectiveness of certain drugs in vivo.

Viral Load (HIV RNA plasma level) and CD4 counts (size of the residual target cell population) are the reference 'surrogate markers' for monitoring response to treatment

EuResist - goals

To develop treatment response models from experience



The Integrated *EuResist* DB

Integrates **clinical** and **virological** data from ARCA, AREVIR, Karolinska

THERAPY

PatientID
Treatment regimen
Date of start
Date of stop
Reason for change/stop

AIDS EVENTS

PatientID
Event
Date

GENOTYPE

PatientID
Date
Sequence
Method

PATIENTS

Patient ID
Gender
Year of birth
Country of origin
HCV status
HBV status

HIV RNA

PatientID
Date
Copies/ml
<LLD (undetectable)
Method

STATUS

PatientID
Date
Followed
Lost
Died

CD4

PatientID
Date
CD4/mmc
CD4%

Standard datum

Two different definitions of the standard datum to be used for model training have been set:

- **Classical**: based on measured viral load changes around a treatment change event,
- **Alternative**: based on matched therapy-genotype pairs (allowing a large increase in the number of training instances).

Both binary (treatment success/failure) and continuous (viral load change) prediction have been pursued

HIV sequence is the minimal information required
Extra patient data (e. g. **treatment history**, **CD4 counts**)
and sequence-derived features (i. e. **inferred phenotypic resistance**, **genetic progression** and **genetic barrier scores**) have been considered

The prediction engines

An array of independent prediction engines based on different models:

- **Case Based Reasoning (CBR)**: local fitting procedure which selects compact subsets of predictive variables: large amount of data is crucial
- **Generative-discriminative engine**: global fitting method employs first a generative model that uses all data and then applies Kernel method (or Support Vector Machines) for prediction
- **Graph model**: partition method used to group the data in the features space minimising the information loss.
- **Evolutionary model**: includes genetic evolutionary information into derived features (not in Standard Datum) and uses different machine learning techniques for prediction
- **Fuzzy logic**: an existing predictor retrained on the *EuResist* IDB to generate features for the CBR engine.

The Integrated *EuResist* DB (January 21 Release)

	ARCA	Arevir	Karolinska	<i>EuResist</i>
Patients	8651	953	4207	13.811
Therapies	25222	4787	14211	44.220
Therapy Compounds	73498	14221	22482	110.201
CD4 Isolates	100948	24742	85035	210.725
Viral Load Isolates	77236	22171	58718	158.125
Raw Sequences	13784	1274	1184	16.242
Protease Sequences	12545	1274	1054	14.873
Protease Mutations	109544	11827	8823	130.194
Reverse Transcriptase Sequences	12609	1273	841	14.723
Reverse Transcriptase Mutations	288032	28161	27186	343.379

Data	Train -Positive	Train Negative	Test Positive	Test negative
Classic	1522	747	170	84
Alternative	1271	5673	143	632

Results

Depending on the different modeling techniques and attributes, treatment success/failure was correctly predicted in 70-78% of cases in the validation set :

GROUP	Best Model	Features	Precision	Area Under ROC Curve
IBM	Linear Regression+ Logistic Regression+ Naive Bayes	Genotype-indicator + treatment-indicator + in-vitro prediction + genotype history + naive Bayes on drug history, treatment indicator and age + baseline VL + number of treatments + age	0.762	0.786
RM3 - Informa	Logistic Model Trees	All filtered attributes (including derived)	0.7791	0.798
RMKI	Logistic Regression	-	0.771	0.788
MPI	Logistic Regression	Genotype-indicator + treatment-indicator + Genetic Barrier II + Baseline VL + Number of treatments + Class History	0.768	0.775

Conclusions

Availability of a large aggregated database containing clinical and virological data makes it possible to train multiple models for predicting response to treatment at a fast pace.

While HIV genotype is a recognized key factor, inclusion of additional patient data and sequence-derived features can improve the accuracy of the prediction.

Thanks to:

- Maurizio Zazzi (University of Siena)
- Andre Altmann (Max Plank Inst.)
- Mattia Prosperi (Informa s.r.l.- University of Roma3)
- Michal Rosen Zvi (IBM Haifa lab.)
- Yardena Peres (IBM Haifa lab.)
- All *EuResist* team

Thank you

Francesca Incardona

f.incardona@informacro.info – www.euresist.org