

Probabilistic Graphical Models for Prediction of HIV-1 Drug Resistance Mutations Based on Antiretroviral Treatment History.

J. Ravela¹, R. Raina¹, S.Y. Rhee¹, J.M. Schapiro¹ and R.W. Shafer¹
¹Stanford University. Stanford, CA, USA

Robert Shafer
Stanford University
Dept. of Infectious Diseases
rshafer@stanford.edu
http://hivdb.stanford.edu

INTRODUCTION

Algorithms that predict the presence of HIV drug-resistance mutations based on past antiretroviral treatment may be useful in parts of the world without access to drug susceptibility testing.

Such algorithms could also be used to supplement drug susceptibility test results in persons with complicated treatment histories for whom susceptibility testing often underestimates the extent of drug resistance present within an individual's virus population.

Drug	41 L	44 DA	62 V	65 R	67 NGE	69 DN	69 ins	70 R	74 VI	75 TS	75 MA	75 I	77 L	115 F	116 Y	118 I	151 M	184 VI	210 W	215 YF	219 QENR
3TC	0.71	0.92	0.97	1.00	0.75	0.89	1.00	0.86	0.95	1.00	0.97	1.00	1.00	1.00	1.00	0.88	1.00	0.45	0.83	0.65	0.83
ABC	0.86	0.98	1.00	1.00	0.84	0.98	1.00	1.00	0.84	1.00	1.00	1.00	1.00	1.00	1.00	0.91	1.00	0.92	0.90	0.90	0.88
AZT	0.67	0.91	0.98	1.00	0.72	0.85	0.98	0.75	0.96	1.00	0.98	0.97	0.97	1.00	0.99	0.85	0.98	0.63	0.79	0.58	0.80
D4T	0.70	0.93	0.99	1.00	0.74	0.92	1.00	0.91	0.95	0.98	0.97	1.00	1.00	1.00	1.00	0.87	0.98	0.67	0.82	0.66	0.85
DDC	0.69	0.89	1.00	1.00	0.75	0.90	1.00	0.94	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.87	1.00	0.87	0.76	0.68	0.86
DDI	0.69	0.91	1.00	0.99	0.74	0.92	1.00	0.88	0.91	1.00	0.99	0.99	0.98	1.00	0.96	0.85	0.96	0.84	0.79	0.62	0.84
TDF	1.00	1.00	1.00	0.85	1.00	1.00	1.00	1.00	0.97	1.00	1.00	1.00	1.00	1.00	1.00	0.96	1.00	0.94	0.95	1.00	1.00
Bias	0.66	0.90	0.94	0.94	0.70	0.85	0.98	0.73	0.91	0.96	0.96	0.93	0.95	0.96	0.95	0.84	0.94	0.57	0.79	0.56	0.77

Table 1: Maximum likelihood inhibition matrix for the NRTIs. The (i, j)-th entry in the above matrix gives the probability of a drug 'i' not causing mutation 'j'. Predicting mutations using the above matrix had an accuracy of 78%. In addition to the drugs, each inhibition matrix also contains a row for a "Bias". The "Bias" is used to account for all unmodeled influences (background polymorphism and drug-class effects).

RESULTS

4350, 5750 and 2600 genotype-treatment correlations were available for PI, NRTI, and NNRTI drugs, respectively.

PI-, NRTI-, and NNRTI-resistance positions were correctly classified in 77%, 78%, and 78% of the test sequences, respectively.

Sensitivity (correctly predicting ≥ 0.5 probability of a mutation) was 77% for the PIs, 76% for NRTIs and 76% for the NNRTIs. Specificity (correctly predicting < 0.5 probability of a mutation) was 76% for the PIs, 78% for NRTIs and 78% for the NNRTIs.

METHODS

Graphical models were implemented based on genotype-treatment correlations in the Stanford HIV RT and Protease Sequence database to predict the probability of specific drug resistance mutations (DRMs) from past treatment. Individual drugs and DRMs were treated as nodes and the probability of the absence of a mutation given a drug was estimated from the data set (Figure 1).

For each drug-mutation pair, a parameter controls how much of the drug's effect on the mutation is inhibited. Tables 1,2 and 3 show these parameters. The initial estimate for these parameters was the observed probability of a position being wildtype given that a drug was administered. The initial estimate is then optimized using a Maximum Likelihood (ML) estimation algorithm. To counter the skew between the observed mutations and the wild types, we optimize a modified likelihood estimate to give more weight to correctly predicting mutations than to correctly predicting wild types.

The parameters were estimated using 70% of the dataset and the models were evaluated based on their ability to predict mutations on the unseen sequences in the remaining 30% of the dataset.

Separate models were created for NRTIs, PIs, and NNRTIs.

Drug	10 FIRV	20 IRM	24 I	30 N	32 I	33 FI	36 ILV	46 ILV	47 AV	48 MV	50 L	50 V	53 L	54 VTAS	54 LM	71 VTIL	73 ASTC	82 AFTS	84 V	88 D	88 ST	90 M
APV	0.47	0.81	1.00	1.00	0.97	0.82	0.75	0.59	0.99	1.00	1.00	0.96	1.00	0.77	0.83	0.58	0.81	0.78	0.82	1.00	1.00	0.61
IDV	0.52	0.81	0.91	1.00	0.95	0.99	0.70	0.65	0.99	0.96	1.00	1.00	0.96	0.75	1.00	0.62	0.84	0.67	0.90	1.00	1.00	0.69
LPV	0.56	0.72	1.00	1.00	0.98	0.86	0.60	0.69	0.98	1.00	1.00	1.00	1.00	0.65	0.96	0.60	0.83	0.75	0.86	1.00	1.00	0.66
NFV	0.65	0.82	1.00	0.82	1.00	0.99	0.62	0.75	1.00	1.00	1.00	1.00	0.97	0.86	1.00	0.67	0.89	0.88	0.93	0.90	0.97	0.70
RTV	0.54	0.79	0.99	1.00	0.99	0.96	0.67	0.74	1.00	0.94	1.00	1.00	0.93	0.69	0.98	0.56	0.83	0.66	0.80	1.00	1.00	0.62
SQV	0.52	0.81	1.00	1.00	1.00	0.98	0.72	0.75	1.00	0.90	1.00	1.00	0.91	0.75	0.99	0.54	0.81	0.76	0.82	1.00	1.00	0.56
Bias	0.55	0.76	0.93	0.90	0.94	0.97	0.58	0.70	0.96	0.94	1.00	0.99	0.95	0.78	0.98	0.60	0.86	0.72	0.89	0.93	0.94	0.68

Table 2: Maximum likelihood inhibition matrix for PIs. Predicting mutations using the above matrix had an accuracy of 77%.

Drug	98 G	100 I	101 P	101 EQ	103 NS	106 A	106 M	108 I	179 DE	181 C	188 CH	188 L	190 A	190 S	190 E	225 H	227 L	230 L	236 L	238 TN
DLV	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
EFV	1.00	0.84	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.96	1.00	1.00	0.97	0.94	0.80	1.00	1.00	1.00	1.00
NVP	1.00	1.00	1.00	1.00	1.00	0.93	1.00	1.00	1.00	0.94	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Bias	0.59	0.70	0.88	0.37	0.06	0.81	0.80	0.54	0.73	0.19	0.96	0.50	0.30	0.73	0.94	0.78	0.81	0.85	0.93	0.78

Table 3: Maximum likelihood inhibition matrix for the NNRTI drugs. Predicting mutations using the above matrix had an accuracy of 78%

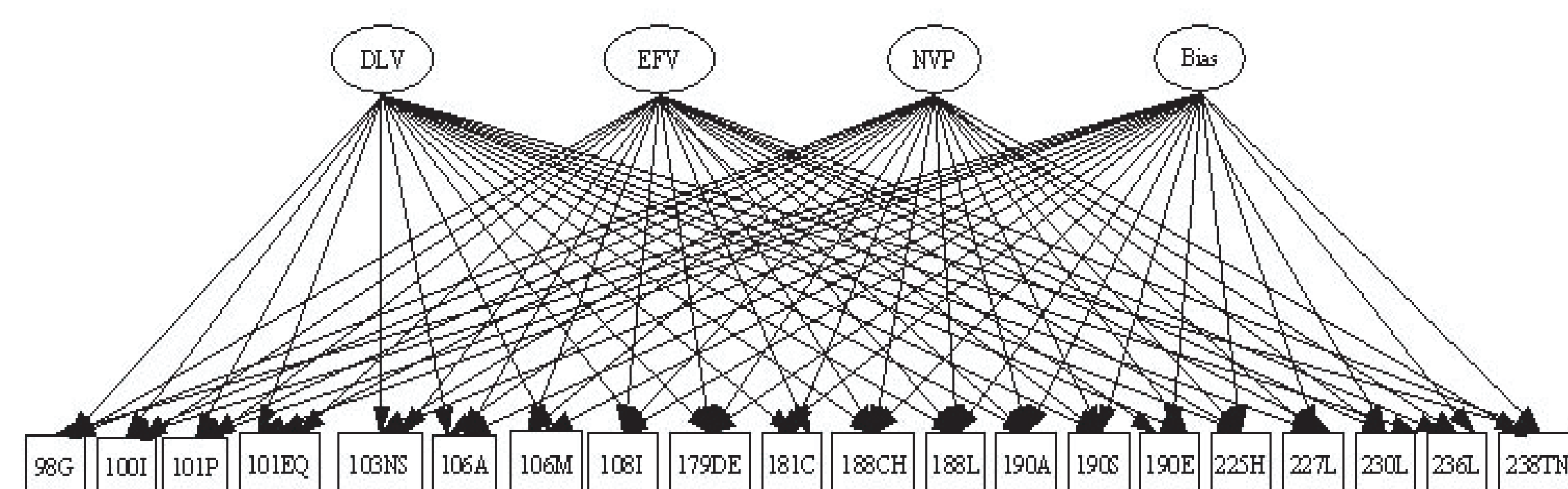


Figure 1: The graphical model showing the relationships between the NNRTI drugs and mutations. The weight of an edge from a Drug to a Mutation is read from the inhibition matrix. Similar models were constructed for PIs and NRTIs. This model is also known as a noisy-OR model.

Computing the probability of a mutation: The probability of a mutation M_i is given by

$$P(M_i | d_1, d_2, \dots, d_n) = 1 - \prod_{j=1}^n (\text{Inhibit}(j, i))^{d_j}$$

where d_1, d_2, \dots, d_n are the drugs received by the person from whom the sequenced virus was obtained, $\text{Inhibit}(j, i)$ were obtained from the inhibit matrix shown above, and d_j is 0 or 1 depending on treatment history.

DISCUSSION and CONCLUSIONS

The proposed probabilistic graphical models provide an intuitive framework with drugs as causes leading to mutations as effects. The models attempt to "understand" the relationship between drugs and mutations rather than the relation between regimens and mutations and thus should have good generalization properties.

This simple model predicts the presence (probability ≥ 0.5) of drug-resistance mutations with an accuracy of $>75\%$. This suggests that the association between mutations and drugs is highly nonrandom.

Models that contain interactions among drugs and among mutations are under development. Such models would take into account additional drug resistance knowledge and would likely perform better. For example a model that has different mutation parameters for amprenavir used as the only PI compared with amprenavir used for late salvage would be more likely to accurately detect amprenavir-specific mutations. A model that takes into account that certain mutations such as D30N and N88D are correlated whereas other mutations (e.g. D30N and L90M) are anti-correlated is also likely to perform better.