XVI International HIV Drug Resistance Workshop

Prediction of phenotypic susceptibility to the three major HIV drug classes from physiochemical properties of the primary enzymatic structure using artificial neural networks (ANNs)

J Kjær[1,2], L Høj[2], Z Fox[1] and JD Lundgren[1]

[1]Copenhagen HIV Programme (CHIP), Denmark; [2]University of Copenhagen, Bioinformatics Centre; Denmark

Jesper Kjær
Copenhagen HIV Programme
Hvidovre University Hospital
Kettegaards Alle 30
DK-2650 Hvidovre, Denmark
Tel: +45 36 32 30 15
Fax:+45 36 47 33 40
E-mail:jkj@cphiv.dk

BACKGROUND

Genotypic interpretation algorithms are empiric and based on several sources of information incl. extrapolation of associations in datasets between different patterns of mutations and phenotypic susceptibility results and/or viral response. We feel there is a lack of chemical and biological thinking in these systems.

Here we present the development and validation of quantitative structure-activity relationship models (QSAR, **Figure 1**) as artificial neural networks (ANNs) using descriptors for physiochemical properties for mutations in HIV-1 protease (PR) and reverse-transcriptase (RT) for predicting phenotypic susceptibility to drugs in the NRTI, NNRTI, and PI classes.

METHOD

We extracted datasets containing pairs of unique gene sequences (PR and RT respectively) and their corresponding exact phenotype values from the publicly available Stanford HIV Drug Resistance Database.

We extracted 544 different chemical and structural descriptors from the AAindex (Amino acid indices and similarity matrices; http://www.genome.ad.jp/dbget/aaindex.html) and applied a series of unsupervised feature selection (UFS) data mining techniques to obtain a set of relevant physiochemical descriptors for each of the three drug classes (see box below) .

The physiochemical descriptors were used to translate the sequence data into a vector of values containing the physiochemical properties for every position in the PR and RT sequences (see **Figure 2** for an example). The physiochemical descriptor values used were delta values, defined as the difference between each physiochemical descriptor value at each amino acid position in the sequences and the HXB2 sequence.

For each of the drugs we used internal validation (10 fold cross) in a 1:9 split to identify the best average correlation coefficient across 10 ANNs while optimizing the number of neurons in the hidden layer (from 1 to 15) and the number of training iterations (maximum of 25) (**Figure 2**). To avoid over fitting to the test data we used the mean square error of the training iterations to determine the point of optimum training.

A subset of the data was withheld for an external validation. All sequences in the external validation set were unique compared to the set used for training and internal cross-validation. The 10 ANNs for each drug were used in ensemble and the obtained prediction is the average predicted IC50 fold change value across these 10 ANNs.

Figure 1 Example of the drug and enzyme interaction that is modelled in the QSAR models

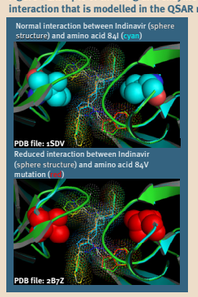
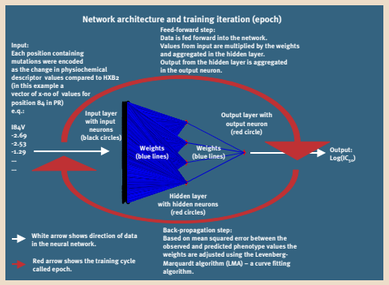


Figure 2 Schematic of the neural network architecture and training



IDENTIFICATION OF RELEVANT CHEMICAL AND STRUCTURAL DESCRIPTORS – DETAILED STEPS:

1. For each of the descriptors and for every position in the PR and RT sequences we calculated the difference in the physiochemical descriptor between the wildtype amino acid (as defined by the HXB2 reference sequence) and the observed amino acid.
2. For every sequence we summed all the descriptor values into a single value per sequence. Step 1 and 2 was repeated for all 544 descriptors in the AAindex generating 8,704 datasets (16 drugs x 544 descriptors) containing two values per sequence:the IC50 fold change value and the summed descriptor value.
3. For each of the 544 different physiochemical properties we calculated the correlation coefficients between the observed IC50 fold change values and the values for the summed physiochemical descriptor for all sequences for each drug at a time. We used the correlation coefficients to rank the physiochemical descriptors for the relevance to each drug and selected around 20 descriptors with the numerically* highest correlation coefficient for further analysis.
4. In order to generate general models we merged the 20 descriptors for each drug together into a single set for each of the three drugs classes (n=number of physiochemical descriptors): NRTI (n=67), NNRTI (n=34) and PI (n=46) respectively.
5. Descriptors showing high inter-correlation as indicated by the AAindex were removed, only the highest ranking descriptor was kept. Descriptors that based on their content had no biochemical relevance for the interaction between PR/RT and the drugs were also removed.
6. The remaining descriptor sets were further optimised using backwards elimination of one descriptor at a time.

* We considered strong negative correlations equally important to strong positive correlations. Several descriptors showed a strong negative correlation coefficient with an increasing IC50 fold change value indicating a clear correlation between increasing resistance and loss of a specific physiochemical property.

RESULT DETAILS

Figures 3 - 18; Correlation plots between observed and predicted log(IC_{50} fold change) for the internal (10 fold cross) (shown with black dots) and the external validation (shown with red stars). Histograms and tables show the proportion observed and predicted IC_{50} fold changes > x3 and the positive predictive value (PPV) for these predictions.

RESULTS

Phenotypic results in the datasets were obtained with PhenoSense (n=10,286) associated with sequences for PR- and RT-genotypes (n=1,507) primarily from subtype B (98.5%) clinical isolates.

The correlation coefficients (r_{mean} = mean coefficient across the ten ANNs from internal validation; range: lower-upper) between the observed and the predicted phenotype values for each drug ranged for:

NRTIs from:

tenofovir: r_{mean} = 0.75 (0.67-0.89) to:
lamivudine: r_{mean} = 0.94 (0.90-0.97);

NNRTIs from:

efavirenz: r_{mean} = 0.82 (0.77-0.89) to:
nevirapine: r_{mean} = 0.83 (0.77-0.89);

PIs from:

atazanavir: r_{mean} = 0.83 (0.76-0.90) to:
ritonavir: r_{mean} = 0.92 (0.88-0.94);

Table 1 Correlation coefficients between observed and predicted values for the internal 10 fold cross and external validation

	Generic drug name (Figure number)	Internal 10 fold cross validation r_{mean} (range: lower-upper) (n=number of genotype-phenotype pairs)	External validation (n=number of genotype-phenotype pairs)
NRTI	Abacavir (3)	0.84 (0.80 - 0.91) (n=637)	0.90 (n=63)
	Didanosine (4)	0.87 (0.84 - 0.92) (n=644)	0.80 (n=62)
	Emtricabine (5)	0.91 (0.77 - 0.97) (n=93)	- (n=2)
	Lamivudine (6)	0.94 (0.90 - 0.97) (n=403)	0.97 (n=52)
	Stavudine (7)	0.88 (0.83 - 0.93) (n=642)	0.84 (n=62)
	Tenofovir (8)	0.75 (0.67 - 0.89) (n=333)	0.95 (n=11)
	Zidovudine (9)	0.84 (0.77 - 0.90) (n=621)	0.89 (n=62)
NNRTI	Efavirenz (10)	0.82 (0.77 - 0.89) (n=605)	0.78 (n=72)
	Nevirapine (11)	0.83 (0.77 - 0.89) (n=575)	0.82 (n=69)
PI	Amprenavir (12)	0.86 (0.82 - 0.91) (n=776)	0.91 (n=116)
	Atazanavir (13)	0.83 (0.76 - 0.90) (n=273)	0.90 (n=28)
	Indinavir (14)	0.87 (0.81 - 0.92) (n=758)	0.95 (n=134)
	Lopinavir (15)	0.89 (0.84 - 0.93) (n=519)	0.95 (n=55)
	Nelfinavir (16)	0.89 (0.84 - 0.92) (n=783)	0.96 (n=138)
	Ritonavir (17)	0.92 (0.90 - 0.94) (n=677)	0.95 (n=127)
	Saquinavir (18)	0.91 (0.88 - 0.93) (n=771)	0.95 (n=123)

The correlation coefficients between observed and predicted phenotype values in the dataset used for validation (external) ranged for:

NRTIs from: didanosine: r=0.80 to: lamivudine: r = 0.97;
NNRTIs from: efavirenz: r=0.78 to: nevirapine: r = 0.82;
PIs from: atazanavir: r=0.90 to: nelfinavir: r = 0.96.

The proportion of pairs with observed (using Stanford HIV-DB) and predicted (using ANNs) IC_{50} >3-fold change, are:

NRTIs from: abacavir: 39.1% vs. 45.4% (n=637) to: didanosine: 92.6% vs. 93.2% (n=644);
NNRTIs from: nevirapine: 60.4% vs. 60.2% (n=575) to: efavirenz: 63.3% vs. 62.8% (n=605);
PIs from: nelfinavir: 30.9% vs. 30.8% (n= 783) to: amprenavir: 57.2% vs. 59.3% (n=776);

Details for each drug is shown in **Figures 3 – 18**

LIMITATION

These analyses have consistently validated correlations of *in vitro* susceptibility to antiretroviral drugs. The extent that the predictions of drug susceptibility from the ANNs can predict virological response to antiretroviral therapy in persons remains to be determined.

CONCLUSION

Based on the physiochemical properties of the PR and RT amino acid sequences, ANNs predict the *in vitro* phenotypic susceptibility to drugs inhibiting these viral enzymes to an extent comparable to that obtained from routine phenotypic susceptibility testing. An advantage of this approach is that the ANNs can interpolate between the various physiochemical properties it was trained on, and hence do not require updating to predict susceptibility from novel mutational patterns.

These results provide a basis for developing drug resistance predictors for HIV-1 PR and RT mutations using chemical and structural property descriptors.

CHIP COPENHAGEN HIV PROGRAMME