

Prediction of phenotypic susceptibility to seven protease inhibitors from chemical and structural descriptors of the protease gene's amino acid sequence using artificial neural networks.

J Kjær^{1,2}, L Høj², Z Fox¹ and JD Lundgren¹

¹Copenhagen HIV Programme (CHIP), ²University of Copenhagen, Bioinformatics Centre

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.

Here we examine if a quantitative structure-activity relationship models (QSAR) using descriptors for chemical and structural properties (lipophilicity, steric properties and electro properties) for the mutations in the HIV-1 protease enzyme (**Figure 1**) can predict phenotypic susceptibility to individual protease inhibitors.

This was investigated by supervised training of artificial neural networks (ANNs) with datasets containing protease-gene sequences and their corresponding levels of phenotypic susceptibility to each protease inhibitor.

METHOD

ANNs for the following protease inhibitors were created: atazanavir, amprenavir, indinavir, lopinavir, nelfinavir, ritonavir and saquinavir.

We extracted datasets of pairs of unique protease-gene sequences and their corresponding exact phenotype values from the publicly available Stanford HIV Drug Resistance Database.

Amino acid mutations can be described chemically and structurally in several ways. We have so identified best way to be the three principal components of amino acid properties, as identified by Hellberg et al. in 1987 [1], which represent: lipophilicity, steric properties and electro properties. E.g. mutation 84V becomes a vector of 3 values:

Lipophilicity	-2.69
Steric properties	-2.53
Electro properties	-1.29

Output was transformed to log2 values of IC50 fold change from wildtype.

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.

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

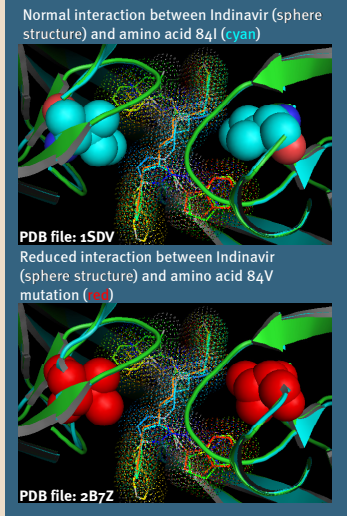
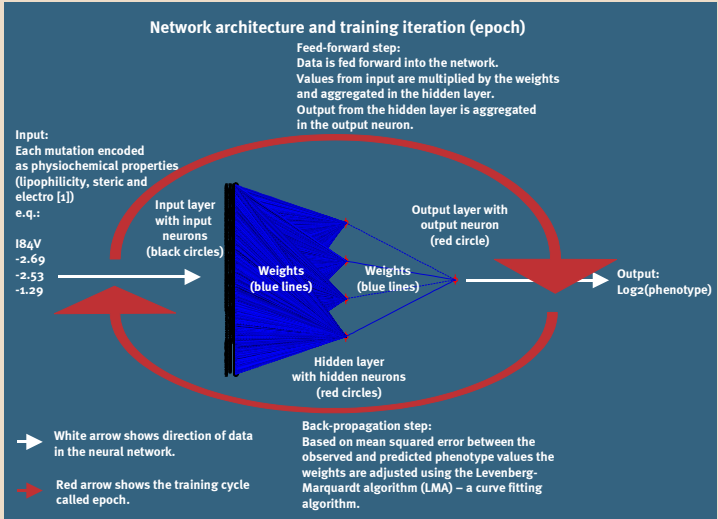


Figure 2 Schematic of the neural network architecture and training



RESULTS

Phenotypic results in the datasets were obtained with PhenoSense (n=4,557), Antivirogram (n= 2,261) and other unspecified assays (n=2,761) associated with protease-gene sequences of primarily clinical (n=1,817) and only few laboratory (n=150) isolates.

A total of 9,579 data pairs were used after removal of 299 sequences and their corresponding exact phenotype value as these pair wise had an exact identical genotypic sequence. These were removed to avoid training and testing with the same genotypic sequence. Any phenotypic results listed as > or < were also excluded. Genotype sequences included amino acid mixture codes.

The proportion of pairs with observed (as per data in Stanford database) /predicted (from ANNs) IC₅₀ fold change >3-fold and >10-fold can be seen in **Figures 3 – 9**.

The obtained correlation coefficients (r_{mean} = mean coefficient across the 10 ANNs; range: lower – upper) between the observed and the predicted log₂(IC₅₀ fold change) values for each drug are listed in **Table 1**.

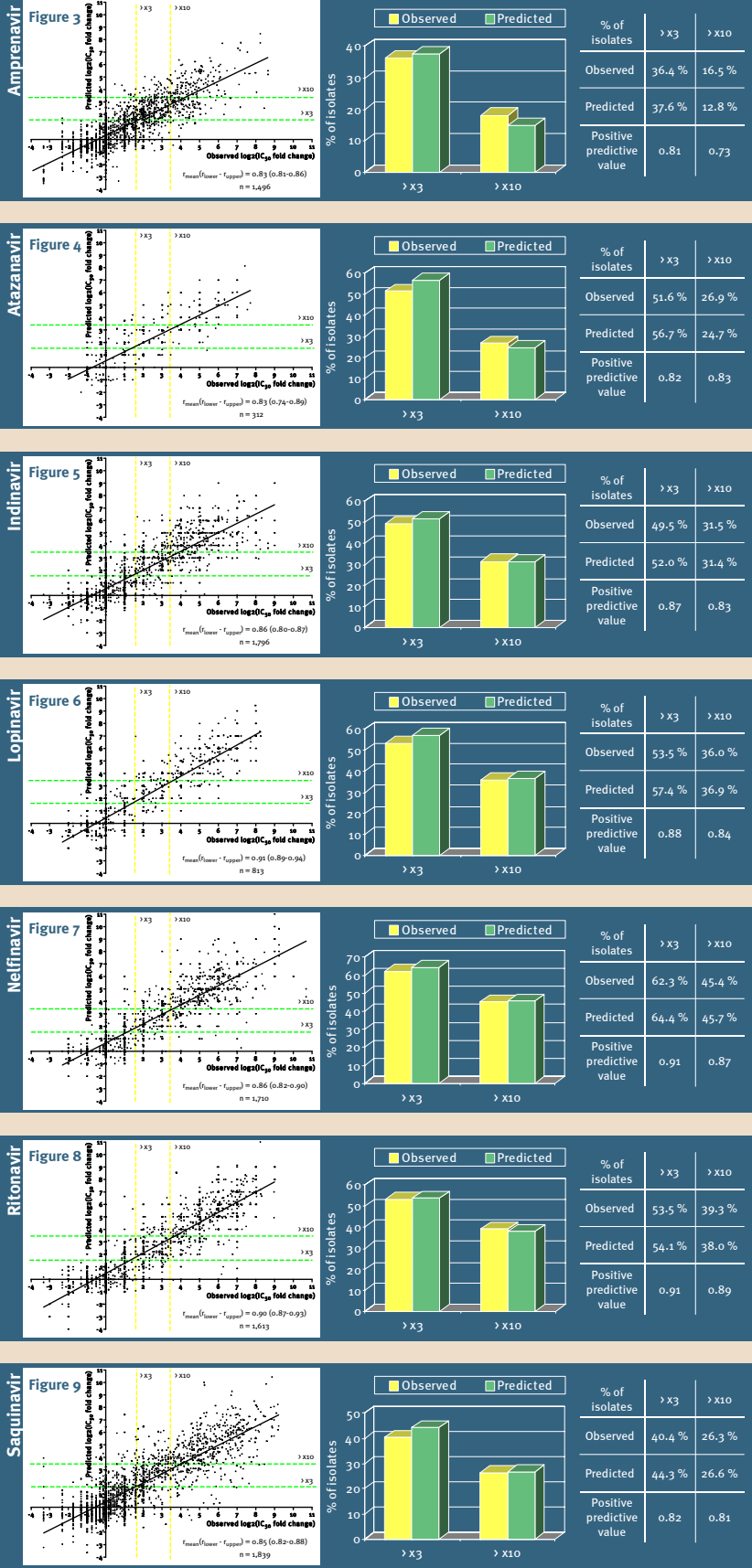
Correlation plots for each drug and the 10 ANNs can be seen in **Figures 3 – 9**.

We used the 299 data pairs with identical genotype sequence profiles and their corresponding exact phenotype value to asses reproducibility of a phenotype test by another phenotype test as found in the dataset. Correlation coefficients between two sets of observed log₂(IC₅₀ fold change) values were calculated for each drug (**Table 2**).

The correlations coefficient for these sets indicate that the reproducibility of phenotype testing is comparable to the reproducibility of the best ANNs.

RESULT DETAILS

Figures 3 - 9: Correlation plots between observed and predicted log₂(IC₅₀ fold change) for the 10 ANNs for each of the seven protease inhibitors. Histogram and table show the proportion observed and predicted above IC₅₀ fold change > x3 and > x10 and the positive predictive value of the predictions above these clinical cut-offs.



LIMITATION

Despite a very diverse dataset with measurements from PhenoSense (n=4,557), Antivirogram (n= 2,261) and other unspecified assays (n=2,761) the neural networks still need to be validated on independent datasets (external validation).

CONCLUSION

Our results show that artificial neutral networks predict *in vitro* susceptibility to protease inhibitors comparable to what can be obtained from routine phenotypic susceptibility testing of the gene.

These results provide a basis for developing drug resistance predictors for HIV-1 protease mutations using chemical and structural property descriptors. The potential development of resistance predictors for reverse transcriptase mutations using similar methods may also warrant further investigation.