**BACKGROUND**

Genetic interpretation algorithms are empirical 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 have been in a rush of chemical and biological thinking in these systems. Here we present the development and validation of quantitative structure-activity-relationship models (QSAR, Figure 3) as artificial neural networks (ANNs) using descriptors for physicochemical properties for resistance to 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 public available Stanford HIV Drug Resistance Database.

We extracted 344 different chemical and structural descriptors from the Akinski amino acid indices and similarity matrices; http://www.genome.ad.jp/targer/rasned.html and applied a series of unsupervised feature selection (UFS) data mining techniques to obtain a set of relevant physicochemical descriptors for each of the three drug classes (see box below).

The physicochemical descriptors were used to translate the sequence data into a sector of values containing the physicochemical properties for every position in the PR and RT sequences (see Figure 3 for an example). The physicochemical descriptor values used were data values, defined as the difference between each physicochemical descriptor value at each amino acid position in the sequences and the baseline sequence.

For each of the drugs we used internal validation (10 fold cross in a 1:1 split) to identify the best average correlation coefficient across ANNs while optimizing the number of neurons in the hidden layer (from 1 to 15) and the number of training iterations (maximum of 100) (Figure 1). 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 external validation. All sequences in the external validation set were unique compared to the set used for training and internal cross validation. The 35 ANNs for each drug were used in ensemble and the obtained prediction is the average predicted IC50 fold change value across these 35 ANNs.

**RESULTS**

Phenotypic results in the datasets were obtained with PhenoSense (n=28) and associated with sequences for PR- and RT genotypes (n=20) primarily from subtypes B (43%) clinical isolates.

The correlation coefficients $r_{mean}$ mean coefficient across the ten ANNs from internal validation; range: lower-upper)

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

<table>
<thead>
<tr>
<th>Drugs</th>
<th>NRTI (n=67)</th>
<th>NNRTI (n=34)</th>
<th>PI (n=46)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rmean</td>
<td>0.40 (0.70-0.73)</td>
<td>0.83 (0.70-0.85)</td>
<td>0.81 (0.75-0.89)</td>
</tr>
<tr>
<td>Positive isolates</td>
<td>31.9%</td>
<td>39.7%</td>
<td>Observed</td>
</tr>
<tr>
<td>Predicted</td>
<td>55.0%</td>
<td>75.0%</td>
<td>Observed</td>
</tr>
<tr>
<td>Observed</td>
<td>22.5%</td>
<td>42.8%</td>
<td>Observed</td>
</tr>
<tr>
<td>Validated</td>
<td>37.7%</td>
<td>65.7%</td>
<td>Observed</td>
</tr>
</tbody>
</table>

**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 physicochemical properties of the PR and RT amino acid sequences, ANNs predict the in vitro 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 physicochemical properties it was trained on, and hence do not need to update 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.