XVII International HIV Drug Resistance Workshop

In silico identification of physiochemical properties at mutating positions relevant to reduced susceptibility to amprenavir.

L Høj[1], J Kjær[1], O Winther[2], A Cozzi-Lepri[3] and JD Lundgren[1,4]

[1]Copenhagen HIV Programme (CHIP), Denmark; [2]University of Copenhagen, Bioinformatics Centre Denmark; [3]Department of Infection & Population Health, Royal Free Campus, University College London, UK; and [4]Centre for Viral Diseases/KMA, Rigshospitalet, Copenhagen, Denmark

Leif Høj and Jesper Kjær
Copenhagen HIV Programme
University of Copenhagen,
Faculty of Health Sciences
The Panum Institute/Building 21.1 Blegdamsvej 3B
2200 Copenhagen N Denmark
Tel: +45 35 45 57 57
Fax: +45 35 45 57 58
E-mail:leh@cphiv.dk / jkj@cphiv.dk

BACKGROUND

We have previously presented a high performing artificial neural network capable of prediction in vitro IC50 fold change (FC) values based on genotypic datasets described by the physiochemical properties of the amino acid mutations.

Here we as an example use amprenavir to find the physiochemical properties that best describe the mutations in the protease enzyme in relation to reduced susceptibility.

METHOD

We used physiochemical properties derived from the Amino Acid index (544 properties) (Amino acid indices and similarity matrices; http://www.genome.ad.jp/dbget/aaindex.htm) with additional properties found in literature (21 properties) to describe the mutational change from the HXB2 sequence in a dataset of amino acid sequences.

Each sequence is paired with an IC50 FC value for amprenavir. Data was extracted from the Stanford HIV Drug Resistance Database.

The dataset used for analysis contained 99 positions for protease x 565 properties and no interaction variables. We performed regression analysis using Least Angle Regression (LARS) with Least Absolute Shrinkage Selection Operator (LASSO) to derive a model by 10 fold cross-validation in a 1:9 split on our dataset (please see the box below for details on the method).

The model with the lowest mean square error (MSE) of the predicted vs. observed IC50 FC values was used to identify the most relevant amino acid positions and their physiochemical properties.

Least angle regression (LARS) with Least Absolute Shrinkage Selection Operator (LASSO)

Least angle regression is a model selection algorithm that like *All Subset regression*, *Forward Selection* and *Backwards Elimination* choose a linear model based on data presented to the model.

Typically the dataset contain a large collection of covariates from which we hope to select a set of covariates that provide us with an efficient predictions of response variables. Modifying the LARS algorithm to implement a LASSO solution, we not only achieve that goal but also derive a space model without covariates.

Math of a LARS LASSO solution:

Given that regression coefficients $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3 \ldots \hat{\beta}_m)$ provides a prediction vector $\hat{\mu}$ for a given LARS model the optimisation problem is as follows:

$$\hat{\mu} = \sum_{j=1}^{m} x_j \hat{\beta}_j = X\hat{\beta} \quad [X_{nxm} = (x_1, x_1, \ldots x_m)]$$ M=covariates N= sample size X o dataset

with a total square error :

$$S(\hat{\beta}) \equiv \|y - \hat{\mu}\|^2 = \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2$$

we let $T(\hat{\beta})$ be the absolute norm of $\hat{\beta}$:

$$T(\hat{\beta}) = \sum_{j=1}^{m} |\hat{\beta}_j|$$

By implementing the LASSO solution, we choose $\hat{\beta}$ by minimizing $S(\hat{\beta})$ subject to a bound t on $T(\hat{\beta})$

Ref: Efron B. et. Al. Least Angle Regression 2003

RESULT DETAILS

Figure 1: Shows the calculated log(IC50 FC) for each amino acid (labelled with the corresponding letter in figure) at each of the 10 positions and for each of the 22 physiochemical properties as identified in the LARS/LASSO model. A log(IC50 FC) of 0 equals an IC50 FC of 1. Most of the mutations plotted are not observed in the Stanford dataset. Only for position 88 was the amino acids with a log(IC50 FC) below zero also observed in the dataset.

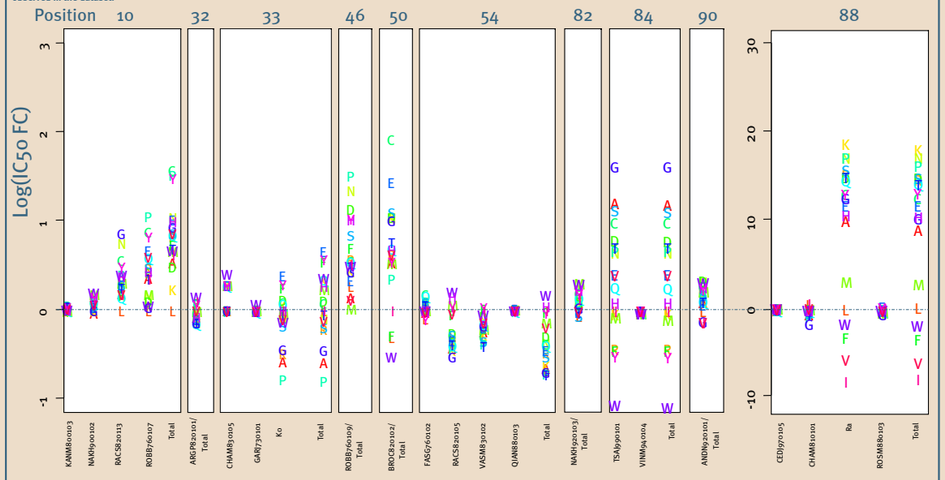

Table 1: Shows the detailed information on the identified 22 physiochemical properties. Listing position, name of descriptor, short description and physiochemical category.
Position numbers in **bold** are listed as major mutations in the IAS 2008 drug resistance list, numbers in normal font are minor mutations and numbers in *italic* are currently not in the IAS list.

Position	Descriptor Name	Description	Category
10	KANM800103	Average relative probability of inner helix (Kanehisa-Tsong, 1980)	Hydrophilicity
	NAKH900102	SD of AA composition of total proteins (Nakashima et al., 1990)	
	ROBB760107	Information measure for extended without H-bond (Robson-Suzuki, 1976)	Conformational Properties
	RACS820113	Value of theta(i) (Rackovsky-Scheraga, 1982)	
32	ARGP820101	Hydrophobicity index (Argos et al., 1982)	Hydrophobicity
33	CHAM830105	The number of atoms in the side chain labelled 3+1 (Charton-Charton, 1983)	Secondary Structure
	GARJ730101	Partition coefficient (Garel et al., 1973)	Polarity
	Ko	Accessibility (Caballero, 2006)	Accessibility
46	ROBB760109	Information measure for N-terminal turn (Robson-Suzuki, 1976)	Conformational Properties
50	BROC820102	Retention coefficient in HFBA (Browne et al., 1982)	Hydrophobicity
54	FASG760102	Melting point (Fasman, 1976)	Secondary Structure
	QIAN880103	Weights for alpha-helix at the window position of -4 (Qian-Sejnowski, 1988)	
	RACS820105	Average relative fractional occurrence in Eo(i) (Rackovsky-Scheraga, 1982)	Conformational Properties
	VASM830102	Relative population of conformational state C (Vasquez et al., 1983)	
82	NAKH920103	AA composition of mt-proteins (Nakashima et al., 1990)	Composition
84	TSAJ990101	Volumes including the crystallographic waters using the ProtOr (Tsai et al., 1999)	Accessibility
	VINM940104	Normalized flexibility parameters (B-values) for each residue surrounded by two rigid neighbours (Vihinen et al., 1994)	Conformational Properties
88	Ra	Compressibility (Caballero, 2006)	Accessibility
	CHAM810101	Steric parameter (Charton, 1981)	Conformational Properties
	CEDJ970105	Composition of amino acids in nuclear proteins (percent) (Cedano et al., 1997)	Composition
	ROSM880103	Loss of Side chain hydropathy by helix formation (Roseman, 1988)	Hydrophilicity
90	ANDN920101	Alpha-CH chemical shifts (Andersen et al., 1992)	Secondary Structure

RESULTS

The extracted dataset contained 776 IC50 FC values each paired a 99 amino acid long protease sequence. Each amino acid in the protease sequences was described with 565 different physiochemical properties.

We identified the best LARS/LASSO derived model (MSE=0.885, mean correlation coefficient=0.847) to be based on only 22 properties for 10 positions in the protease enzyme (10,32,33,46,50,54,82,84,88,90).

In **Figure 1** we have plotted the calculated effect of the relevant physiochemical properties for each possible amino acid mutation at the 10 positions identified by the model and the total effect by position on the IC50 FC value.

Eight of the ten identified positions are in the 2008 IAS drug resistance list for amprenavir (including the major mutations at positions 50 and 84). Positions 33 and 88 are currently in the IAS list. Position 88 which was found to be associated with hypersusceptibility as can be seen from **Figure 1.** Positions 47, 73 and 76 from the IAS list where not identified as relevant by the LARS/LASSO method.

Categorised into six groups the identified properties are:

- **Accessibility** – at positions: 33, 84 and 88
- **Conformational properties** – at positions: 10, 46, 54, 82, 84 and 88
- **Hydrophilicity** – at positions: 10 and 88
- **Hydrophobicity** – at positions: 32 and 50
- **Secondary structure** – at positions: 33, 54 and 90
- **Polarity** – at position: 33

Table 1 lists the details of the physiochemical descriptors identified.

LIMITATION

The LARS/LASSO method eliminates descriptors that have high collinearity and therefore it is possible that a larger number of properties than those shown here are significantly related to the phenotype as highly collinear physiochemical descriptors could have been eliminated. Further work on identifying these descriptors is needed.

CONCLUSION

We have mathematically identified a number of important physiochemical properties that seem to drive the change in IC 50 FC to amprenavir. The set of relevant properties are for some positions more complex than others (e.g. position 10 vs. 82).

Applying this approach to other antiviral inhibitors will provide cost effective in silico knowledge about the physiochemical changes to the affected enzymes and could serve as input for computer assisted drug designs and improvement of existing drug compounds.

CHIP COPENHAGEN HIV PROGRAMME