From Predicting to Analyzing HIV-1 Resistance to Broadly Neutralizing Antibodies

Anna Feldmann[1,2] **and Nico Pfeifer**[1]

[1]**Computational Biology and Applied Algorithmics Department, Max Planck Institute for Informatics, Saarbrücken, Germany**
[2]**Saarbrücken Graduate School of Computer Science, Saarland University, Saarbrücken, Germany**

ABSTRACT

Treatment with broadly neutralizing antibodies (bNAbs) has recently proven effective against HIV-1 infections in humanized mice, non-human primates, and humans. For optimal treatment, susceptibility of the patient's viral strains to a particular bNAb has to be ensured. Since no computational approaches are so far available, susceptibility can only be tested in expensive and time-consuming neutralization experiments. Here, we present well-performing computational models (AUC up to 0.84) that can predict HIV-1 resistance to bNAbs given the envelope sequence of the virus. Having learnt important binding sites of the bNAbs from the envelope sequence, the models are also biologically meaningful and useful for epitope recognition. Additional to the prediction result, we provide a motif logo that displays the contribution of the pivotal residues of the test sequence to the prediction. As our prediction models are based on non-linear kernels, we introduce a new visualization technique to improve the model interpretability. Moreover, we confirmed previous experimental findings that there is a trend towards antibody resistance for the subtype B population of the virus. While previous experiments considered rather small and selected cohorts, we were able to show a similar trend for the global HIV-1 population comprising all major subtypes by predicting the neutralization sensitivity for around 36,000 HIV-1 sequences - a scale-up which is very difficult to achieve in an experimental setting.

Keywords: broadly neutralizing antibody, HIV-1 antibody resistance, support vector machine visualization

INTRODUCTION

To date, there is neither a vaccine nor a cure available for infection with the human immunodeficiency virus type 1 (HIV-1). With an incidence rate of around 2 million each year and 1.6 million deaths in 2012 (WHO, 2014), HIV-1 infections continue to be a major global health issue. Since humans seem not to have natural immune mechanisms to clear the infection, infected individuals need to receive lifelong antiretroviral treatment (ART). Due to the high mutation rate of the virus, drug resistances emerge frequently, often requiring a change of drug targets. However, the number of available drug target classes remains limited; this is why there is still a high demand for drugs addressing new targets.

A currently investigated treatment option is the passive transfer of a combination of broadly neutralizing antibodies (bNAbs) to HIV-1 patients. The advantage of these antibodies is that they are very broad and potent. The potency of an antibody is defined as the antibody concentration needed to inhibit HIV-1 infectivity by 50% (IC50), while the neutralization breadth of an antibody is measured by the ability of the antibody to neutralize viruses from different subtypes. Upon the advent of new single-cell based methods, an abundance of these new bNAbs has been isolated and their higher neutralization potency and breadth have been shown in several studies (Walker et al., 2009; Mouquet et al., 2012). The target of these antibodies is the HIV-1 spike, a trimeric heterodimer of two viral envelope glycoproteins, gp120 and gp41. The successful binding of an antibody to this spike blocks the two main functions of the spike, namely mediating host cell fusion and viral entry. As a consequence, the corresponding virus cannot infect any new cell. So far, there are five known epitopes on the envelope glycoprotein, which are targeted by a variety of bNAbs (given in brackets): on gp120 the CD4 binding site (e.g., VRC01, VRC-PG04, 3BNC117, NIH45-46) (Falkowska et al., 2012; Wu et al., 2010; Scheid et al., 2011), the V1/V2 region (e.g., PG9 and PG16), and the V3

loop (e.g., PGT128, PGT121, 10-996, 10-1074) (Walker et al., 2011, 2009); the membrane proximal external region (MPER) on gp41 (e.g., 10E8) (Burton et al., 1994); and a newly identified epitope comprising parts of gp41 and gp120 (e.g., 35O22) (Huang et al., 2014).

Since these specific binding sites of bNAbs on the envelope protein are not accessed by any available drug, a therapy with bNAbs would offer a new effective treatment option for patients with resistance to all current therapies or boost therapy combinations with few active drugs. The efficacy of a treatment with a combination of these broad and potent neutralizing antibodies has been first shown in HIV-1 infected humanized mice and non-human primates (Klein et al., 2012; Barouch et al., 2013), followed by a phase 1 clinical trial this year that confirmed the effective suppression of viremia in HIV-1 infected humans treated with the bNAb 3BNC117 (Caskey et al., 2015). An advantage in comparison with the daily intake of existing drugs is the longer half-life of bNAbs, which can control viral load for more than 28 days in humans after administration. As the envelope protein is the sole target of bNAbs, high sequence variation of the viral envelope sequence together with a glycan shielding of more conserved regions on the envelope often allow the virus to escape immune recognition (Taylor et al., 2008). Thus, for treatment success, neutralization resistances of the patient's viral strains to the given bNAbs must be detected beforehand. Up to now, the neutralization sensitivity of a virus to an antibody can only be determined in time-consuming and expensive experiments, so-called neutralization assays.

In this study, we present prediction models for 11 different bNAbs (VRC01, VRC-PG04, 3BNC117, NIH45-46, PG9, PG16, PGT121, PGT128, 10-996, 10-1074, and 35O22) that automatically learn discriminant signals (amino acids or patterns of amino acids) in the envelope sequence, which influence the neutralization sensitivity to the particular antibody. For the learning process, the models were trained on data sets from three previously published neutralization assays (Doria-Rose et al., 2009; Mouquet et al., 2012; Huang et al., 2014) that in total contain neutralization sensitivity information (IC50 values) for 115 to 220 HIV-1 isolates covering all major HIV-1 subtypes. Having learnt the discriminant signals, the models can predict the neutralization sensitivity of an unseen viral sequence to the considered bNAbs. To predict resistance to a particular antibody, we used two different approaches. On the one hand, we built classifiers with support vector machines (SVM) based on the biological threshold that determines resistance by an IC50 value above 50 µg/mL. On the other hand, in order to provide more fine-grained information, we directly predicted the IC50 value using support vector regression. Since prediction models are often seen as black boxes, we traced back what each classifier learnt from the data and show that many of the learnt discriminant signals are known to play an important role for the binding success of the antibody. In addition, we introduce a new visualization technique that displays the interrelations between the train and test data in the potential high-dimensional feature space. For a better interpretation of the classification decision (resistant or susceptible), we offer motif logos that illustrate which and up to what extent amino acids in the tested sequence contributed to the particular classification result. Apart from their ability to support treatment decisions, we used the prediction models to analyze how neutralization sensitivity changes in the HIV-1 population over time.

Correlating neutralization sensitivity and the variation in the viral envelope sequence has so far only be used to identify potential epitope sites of bNAbs (West et al., 2013; Lacerda et al., 2013). After learning so-called epitope networks of bNAbs, Evans et al. (2014) also predicted neutralization sensitivity to validate these epitope networks. However, the prediction performance was assessed on the same data on which the epitope networks were learnt.

MATERIALS AND METHODS

Data

To learn the neutralization susceptibility of HIV-1 strains to broadly neutralizing antibodies (bNAbs), we trained our prediction models on data from three previously published neutralization assays (Doria-Rose et al., 2009; Mouquet et al., 2012; Huang et al., 2014). Depending on the neutralization assay, IC50 titers for 115 to 220 HIV-1 isolates were available for each of the analyzed 11 bNAbs (VRC01, VRC-PG04, 3BNC117, NIH45-46, PG9, PG16, PGT121, PGT128, 10-996, 10-1074, and 35O22). Since the sole target of antibodies is the envelope glycoprotein of HIV-1, we used for each considered HIV-1 isolate the corresponding viral envelope sequence from the Los Alamos HIV sequence database and aligned the sequences using HIVAlign (Foley et al., 2013).

Prediction models

Following neutralization assay protocols, we use an IC50 value above 50 µg/mL as a threshold to determine neutralization resistance of a virus towards a particular antibody. Applying this threshold,

we built binary classifiers to distinguish between HIV-1 resistance and susceptibility to a bNAb based on the envelope sequence of the virus. Since there are differences in potency between the antibodies, it is also of interest to know how strongly a bNAb neutralizes the virus. For this reason, we also built regression models that directly predict the IC50 value from the envelope sequence of the virus, thereby enabling more fine-grained results.

For the learning process we used kernels in conjunction with large-margin based methods: support vector machines (SVMs) for the classification, and support vector regression for the regression analysis. The underlying kernel for both tasks should preferably fulfill three properties: to allow for positional uncertainty to account for the high mutation rate of the virus, to be able to identify consecutive patterns of amino acids reflecting the shape of some epitopes, and to learn multivariate signals in order to model the fact that epitopes might consist of residues that are not consecutive in the sequence. String kernels that capture positional information such as the oligo kernel (Meinicke et al., 2004) or the weighted degree kernel with shifts (WDKS) (Rätsch et al., 2005), match these requirements and therefore might lead to better performances than conventional kernels like the polynomial (Poly) or the Gaussian RBF kernel. To validate this hypothesis, we compared the performances of models based on each of these kernels. The comparison was conducted by 10 runs of a 5-fold nested cross-validation using AUC and Pearson Correlation Coefficient as performance measure for the classification and regression task, respectively. For the polynomial and Gaussian RBF kernel the amino acid sequences have to be transformed to a real-valued input. We used one-hot encoding to represent the sequence information for the polynomial kernel, i.e., each amino acid $a_i, i \in \{1,\dots 20\}$ is transformed into a 20-dimensional vector, where only the i-th entry is 1, and the others are 0. For the Gaussian RBF kernel, we encoded the sequence information using physico-chemical properties based on Atchley et al. (2005) (RBF1) and Braun and Venkatarajan (2001) (RBF2).

Understanding the classifier

Visualizing the samples' interrelations in the reproducing kernel Hilbert space

Since non-linear dependencies in the data can exist, disregarding them by simply using a linear method might lead to worse performances. Transforming the data from the linear input space in to a space $\mathscr{H}$, in which those dependencies are better represented, can lead to a better separability of the data. Support vector machines that only need dot products of the samples can take advantage of those non-linear dependencies using kernels, which correspond to dot products in the space $\mathscr{H}$. However, the interpretation of the learnt non-linear models and the predicted results remains a challenge, which might explain why non-linear SVMs are less often used than advisable despite their good performances. So far, few methods exist that address this disadvantage of non-linear SVM classifiers using graphical representation (Caragea et al., 2001; Wang et al., 2006). These methods are usually neither generally applicable (only for certain kernels or restricting the data to be low-dimensional) nor simple (requiring additional optimization steps).

In this study, we propose a general method that displays the interrelations between the training and the test samples in the reproducing kernel Hilbert space (RKHS) without explicitly using the feature mapping function Φ. Euclidean distances between the samples in the RKHS, representing the interrelations, can be expressed with the help of the kernel function (Shawe-Taylor and Cristianini, 2004)

$$d(\Phi(x),\Phi(x')) = \|\Phi(x)-\Phi(x')\|^2 = k(x,x) - 2k(x,x') + k(x',x'). \quad (1)$$

To provide a user-friendly representation, we visualize the pairwise feature space distances in a three dimensional space using multi-dimensional scaling (MDS) (Kruskal and Wish, 1978). Multi-dimensional scaling preserves the between-samples distances to the magnitudes of the variables' interrelationships while projecting the data into a D-dimensional space. This way, highly similar variables are spatially closer. Analyzing the stress for MDS with D in a range from 1 to 10, reveals that a three-dimensional space is sufficient to represent the data for all bNAbs (data not shown).

Visualizing the feature space distances with MDS offers new information on the interrelations between the used training data in the RKHS. Furthermore, it can be used to investigate the representation of the test sample x with respect to the training samples x_i with $i \in \{1,\dots,N\}$ in the feature space. Upon the MDS step, we include the class label and the contribution of the training points to the classifier via a color scheme into the MDS visualization (cf. Fig. 2). The two classes were colored in two different colors (blue and orange) where the color intensity was assigned by $\alpha_i k(x_i,x)$ with α_i being the weight of sample i in the classifier. Thus, the color intensity of each training point increases with growing similarity to the test sample as well as with larger influence on the classification result.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1304v1 | CC-BY 4.0 Open Access | rec: 13 Aug 2015, publ: 13 Aug 2015

The labels of the close-by neighborhood of the test sample provide the user additional information on the prediction result.

We present not only a general meaningful graphical representation of the complex feature space, but also a tool to further investigate and interpret the prediction outcome.

Identifying learnt discriminant signals of the classifiers

In general, the learnt signals of a classifier can be traced back, if the kernel incorporates positional information such as the weighted degree kernel with shifts (WDKS) (Rätsch et al., 2005) or the oligo kernel (Meinicke et al., 2004). Both kernels compute the similarity between two sequences x and x' of same length L by comparing the co-occurrences of their substrings within a certain distance. While the oligo kernel considers only the substrings of length l with $1 \leq l \leq L$, the WDKS takes into account the co-occurrences of every substring up to a length l, thereby adjusting for potential overlapping signals of lower order ($\leq l$). For the WDKS, the significance of each oligomer can be identified using positional oligomer importance matrices (POIMs) (Sonnenburg et al., 2008). Due to better performances, we used in this study the oligo kernel to build the the prediction models (cf. Section Results and Discussion). The oligo kernel defines each sequence by the occurrences of its l-mers, which are encoded via so-called oligo functions. The oligo function μ, encoding all occurrences $p \in x_\omega$ of a particular l-mer ω in a sequence x, is defined as

$$\mu_\omega(t) = \sum_{p \in x_\omega} \exp\left(-\frac{1}{2\sigma^2}(t-p)^2\right), \tag{2}$$

with the continuous position variable $t \in [1, L]$ and σ^2 controlling the positional uncertainty. As described in Meinicke et al. (2004), the corresponding learnt weight of the classifier for each oligomer ω at each position t can be retrieved by

$$|w_\omega(t)| = |\sum_{i=1}^{N} \alpha_i y_i \mu_\omega^i(t)|, \tag{3}$$

where $i \in \{1, \ldots, N\}$ denotes the i-th training sample with $\alpha_i \geq 0$ and $y_i \in \{-1, 1\}$ being the learnt weight and classification label of the i-th sample.

Understanding the classification result

Additional knowledge on the classifiers such as the provided interrelationships of the training and test data in the kernel feature space or the discriminant signals learnt by the classifiers, leads to interpretable prediction models. The classification decision for each test sample remains however elusive. In this paper, we offer for each classification of a test sequence, a motif logo - a representation of the test sequence - that shows those residues in the test sequence that contributed the most to the classification result.

Using the kernel feature representation of the oligo kernel, we retrieve the contribution for each residue of the test sequence x^* to the classification result. Since the residue at a certain location t of the test sequence is fixed, there exists only one oligomer ω containing this residue as starting point whose contribution is calculated as

$$S_\omega^*(t) = \sum_{i=1}^{N} \alpha_i y_i < \mu_\omega^i(t), \mu_\omega^*(t) >, \tag{4}$$

with μ_ω^* being the oligo function of l-mer ω of the test sequence. For l-mers > 1 the computed contribution is assigned to all amino acids of the oligomer. Since showing the contribution of each residue of the test sequence might rather be confusing than improving the interpretability of the classification result, we limit the motif logos to only represent the most discriminant residues of the test sequence. This is possible, since we could show that models using only the strongest $p\%$ signals with $p \in \{1, 3, 5, 7, 10, 15, 20, 25\}$ do not have a significantly worse performance compared to models using the full envelope sequence as information (data not shown). To generate the motif logos we used Weblogo 3.0 (Crooks et al., 2004).

RESULTS AND DISCUSSION

Kernel preselection

To validate the hypothesis that string kernels such as the oligo kernel or the weighted degree kernel with shifts might be more suitable for this application than commonly used kernels such as the

Gaussian RBF or the polynomial kernel, we compared the corresponding prediction performances with each other for both, the classification and the regression task. Here, we show the performances of the classifiers for the bNAbs PG9, PG16, 10-996, 10-1074, PGT121, VRC01, and VRC-PG04 as listed in Table 1.

In general, the classifiers using the string kernels yield not a better performance compared to the Gaussian RBF or the polynomial kernel. Only for the VRC-PG04 bNAb the classifier using the oligo kernel performed better than all the other kernels. A reason for this might be the characteristic binding site of VRC-PG04 on the envelope protein. The binding site of VRC-PG04 is rather a large consecutive sequence than single residues as for the others bNAbs. While all kernels are good at identifying single residues, only the oligo and the weighted degree kernel with shifts are able to capture substrings of length l (l-mers). In contrast to the oligo kernel, the weighted degree kernel with shifts counts all co-occurrences of substrings of length $\leq l$, thereby, adding a lot of noise the higher the parameter l is set, if only the long k-mers are informative.

Table 1. Tested parameter ranges for each kernel together with the average AUC performance for each bNAb classifier in the nested cross-validation: For VRC-PG04 (in **bold**), the oligo kernel based classifier performed better than the others.

	Antibody						
	V1/V2 loop		V3 loop			CD4 binding site	
Kernel	PG9	PG16	PGT121	10-996	10-1074	VRC01	VRC-PG04
Poly	0.72	0.74	0.81	0.87	0.83	0.73	**0.54**
RBF1	0.69	0.70	0.79	0.85	0.82	0.70	**0.52**
RBF2	0.69	0.73	0.77	0.85	0.81	0.68	**0.63**
WDKS	0.69	0.71	0.78	0.84	0.79	0.78	**0.57**
Oligo	0.67	0.71	0.79	0.84	0.81	0.71	**0.69**

Parameter settings for the models

Upon kernel comparison, the oligo kernel was selected for all bNAbs to predict the neutralization susceptibility to each bNAb for unseen viral strains. Table 2 presents the final parameters settings for the classifiers for the bNAbs PG9, PG16, 10-669, 10-1074, PGT121, VRC01, and VRC-PG04 fitted by a 5-fold cross-validation (cf. Table 2) and the corresponding performances (cf. Figure 1).

For the PGT121 and VRC-PG04 classifier an l-mer of length 6 led to the best performance whereas the l-mer length for the other antibodies was comparatively small (2-mers for VRC01 and single positions for the remaining antibodies). The length differences of the l-mers for different epitope classes supports the knowledge gained from experimental findings. For the N-glycan dependent antibodies, a single glycan site is the most important residue for a successful binding. The N332-linked (V3-loop-directed) antibodies PGT121, 10-1074, and 10-996 need in the first instance an asparagine at position 332 for successful binding (Julien et al., 2013). The N160-linked antibodies PG9 and PG16 bind in a hammerhead-like way to the virus, building contacts with two glycans (160 and 156 or 171) (Louder et al., 2011). For the CD4 binding site (CD4bs), which forms a cavity, it is only known that it is sterically not easy to bind to for antibodies (West et al., 2014) . Longer l-mers led to the best prediction results for the CD4bs classifiers which is likely due to the fact that the CD4bs-directed bNAbs target a larger epitope compared to the other bNAbs.

Identified discriminant signals

Using the oligo kernel properties as described in the Methods section, we examined the 15% strongest learnt signals for each classifier. Among this set of signals, several residues were learnt by the classifiers that are also supported by literature (Lacerda et al., 2013; West et al., 2012). In Table 3 we present the confirmed learnt signals of the classifiers for the bNAbs PG9, PG16, 10-669, 10-1074, PGT121, VRC01, and VRC-PG04 as an example.

Most of the found discriminant signals for the N-glycan dependent antibodies, that is, for the V1/V2-loop- and V3-loop-directed antibodies, contain the amino acids asparagine (N), serine (S) and threonine (T). These amino acids are also part of the pattern N-X-[S or T], which defines potential N-glycosylation sites (PNGS) (Marshall, 1974). The classifiers for the CD4bs antibodies identified known required residues for CD4-binding as reported in West et al. (2012). The fact that all classifiers learnt some known discriminant position, further support the reliability of the prediction models in addition to the provided prediction performances. Additionally to the already known epitope sites, we also found other discriminant residues that might be interesting for follow-up structural studies.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1304v1 | CC-BY 4.0 Open Access | rec: 13 Aug 2015, publ: 13 Aug 2015

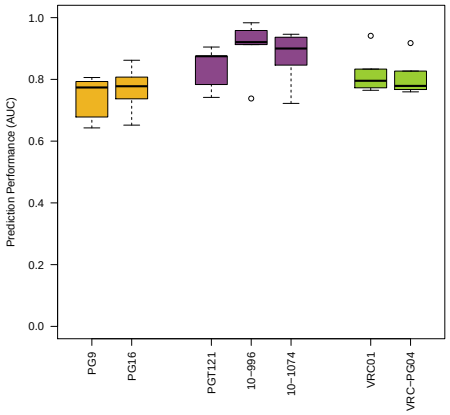

Figure 1. The AUC performances for the best parameter setting for each bNAb classifier using the oligo kernel.

Table 2. The selected parameter settings for the oligo kernel classifier for each bNAb.

Epitope	bNAb	degree	width
V1/V2 loop	PG9	1	1
	PG16	1	0.4
V3 loop	PGT121	6	1.6
	10-996	1	2.6
	10-1074	1	1.6
CD4bs	VRC01	2	3.6
	VRC-PG04	6	20

Table 3. Discriminant signals for each bNAb classifier that are supported by literature. Positive signals are colored in blue, negative in orange.

bNAb	Amino Acid: Positions
PG9	N: 160, 301, 624; S: 393, 613; D: 187; K: 168, 169, 171
PG16	N: 136, 141, 160, 186, 230, 234, 289, 356; S: 393; K: 169,171; D: 167; T: 138
VRC01	N: 186, 276, 279, 280; G: 459; K: 232
VRC-PG04	N: 186, 276,279, 280; G 459; K: 232; R: 456; D: 368
10-996	N: 332, 334; S: 334
10-1074	N: 332, 334; S:334; T: 388, 818
PGT121	QAHCN: 328-332; R: 332

Application of the visualization methods

To demonstrate the introduced visualization methods, we retrieved several HIV-1 envelope sequences from the Los Alamos HIV sequence database (Foley et al., 2013) serving as test input for the classifiers. We present here the test case for the sequence with the GenBank ID HM469973 and the PG9 classifier, which classified the sequence as susceptible.

Visualizing the samples' interrelations in the reproducing kernel Hilbert space (RKHS)

In general, the training samples form dense agglomerations (clusters) with respect to existent interrelations in the RKHS; exemplary shown for the PG9 classifier and the test sequence HM469973 in Figure 2 (fitted to 2-dimensional space for better representation). By adding subtype information to the plot, we observed that the clusters comprise mainly the sequences of the same subtype but not exclusively (data not shown). This is an expected finding, since the arrangement of the points is only based on the kernel similarities, which consider the whole sequence and not only the discriminant positions. Due to the coloring scheme, the most visible close-by point to the test sample (+) is also the most similar and influencing training sequence in the feature space offering the user more information about the representation of the training samples in the classifier as well as a better understanding of the classification result.

Motif logo for the test sequence

To provide the influence of each residue of the test sequence on its classification outcome, we proceeded as described in the Methods Section. In Figure 3 such a motif logo is presented for the test sequence HM469973 and the PG9 classifier.

It can be observed that the test sequence has an asparagine (N) at position 160, which is known to be decisive for a successful binding of the PG9 bNAb. In all three logos, this pivotal residue has the most influence on the classification outcome compared to the other contributions. Considering the contribution of the 1, 3 and 5% learnt discriminant signals to the classification outcome of the test sequence, it can be seen that more discriminant signals occurred in the test sequence that are linked with neutralization susceptibility than with neutralization resistance.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1304v1 | CC-BY 4.0 Open Access | rec: 13 Aug 2015, publ: 13 Aug 2015

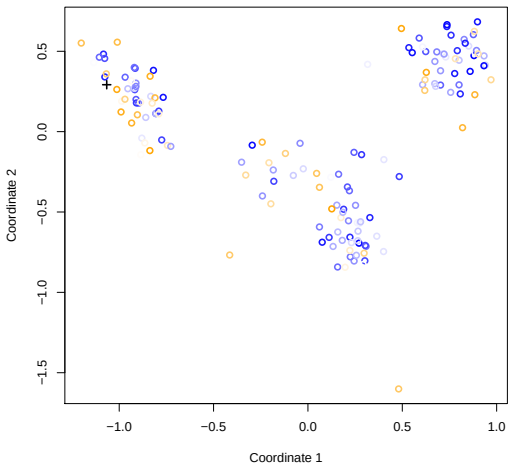

Figure 2. Visualization of the interrelationships between the training and test samples in the PG9 classifier with the test input sample HM469973 projected into two dimensions using MDS. Training samples that could be positively neutralized by PG9 are colored blue, while neutralization resistant training samples are colored orange. The test input sample is displayed as a black cross. The color intensities of the training points increase with growing similarity to the test sample as well as with larger importance of the training sample to the classification outcome. The arrangement of the feature space distances is based only on the distances of the support vectors.

(a) (b) (c)

Figure 3. Motif logo for the test sequence HM469973 using the PG9 classifier. The contribution of (a) 1%, (b) 3% and (c) 5% of the strongest discriminant signals are considered. The height of the letters depends on the proportional contribution. Amino acids of the test sequence that influence the classification outcome towards neutralization susceptibility are displayed in capital letters and blue color; lowercase letters and orange color if they contribute to neutralization resistance. For better interpretability, the corresponding positions of the amino acids in the envelope sequence of the HIV strain HXB2 are shown on the x-axis.

HIV-1 resistance trend analysis

Apart from predicting neutralization sensitivity of unseen viral strains, we used our models to investigate whether neutralization sensitivity of HIV-1 to bNAbs has changed over time. For subtype B variants, a continuous trend towards resistance has been already confirmed in certain cohorts of the French and Dutch HIV-1 population (Bunnik et al., 2008; Bouvin-Pley et al., 2014, 2013). Since evolving resistance towards antibody neutralization in the HIV-1 species would have major implications on the antibody selection for current vaccine development, it is important to know whether such a drift towards resistance also exists in the global HIV-1 population for all subtypes. In contrast to an experimental setting, where expensive neutralization assays need to be performed for a large number of viral strains, we can use our learnt prediction models to examine this question.

To model the global HIV-1 population over time, we used all available envelope sequences from the Los Alamos database (∼36,000) comprising viral isolates from all major subtypes over a time interval from 1981 to 2013. We divided the given time interval into 5 time periods to account for changes in HIV-1 treatment strategies: 1981-1986 before ART; 1987-1991 ART monotherapy; 1992-1995 ART combination therapy (cART); 1996-1999 cART with protease inhibitors; 2000-2005 cART with Lopinavir/Ritonavir; 2006-2013 cART with Maraviroc/Raltegravir. The neutralization sensitivity of the samples to the 11 considered bNAbs was determined using our support vector regression models to predict directly the IC50 value.

Taking all available samples into consideration, we observed a general trend to resistance over time to all bNAbs except PG9 and PG16, for which the virus seems to become more susceptible (cf. Fig. 4). Choosing the same time periods as in Bouvin-Pley et al. (2014, 2013) did not change the result. The existence and significance of a trend was tested using the umbrella test and a Bonferroni correction threshold of 0.05/22 adjusting for multiple testing. When considering only the subtype B variants, the predicted IC50 values show a trend towards resistance for every bNAb confirming the results from Bouvin-Pley et al. (2013). In the non-B subtype samples, a trend towards resistance was observed for

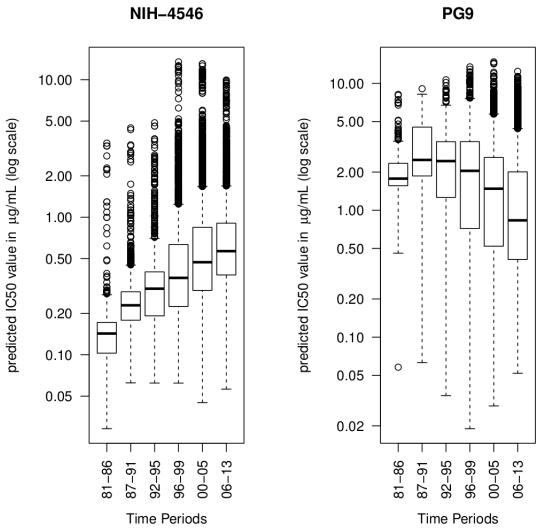

Figure 4. Predicted neutralization sensitivity (log scale) for the ∼36,000 HIV-1 sequences from the Los Alamos database to the PG9 and NIH-4546 bNAbs using the regression models. While there is a trend towards antibody resistance for the CD4 binding site targeting bNAb NIH-4546, the neutralization susceptibility seems to increase for PG9.

all bNAbs, but PG9 and PG16 again; for PGT121 and PGT128 the trend was not significant anymore. By predicting the coreceptor usage for all sequences with geno2pheno[coreceptor] (Lengauer et al., 2007), we detected an increasing ratio of R5/X4-capable viruses over the time periods. Together with the known R5-bias of PG9 and PG16 (Pfeifer et al., 2014) (better against R5, worse against X4), this might lead to the observed trend towards susceptibility. In general, we could confirm that HIV-1 variants of subtype B show a trend towards antibody resistance (Bunnik et al., 2008; Bouvin-Pley et al., 2014, 2013). By using our prediction models, we extended the analysis to the world wide HIV-1 population considering all major subtypes and thus, validating the hypothesis on a large scale, which is very difficult to achieve in an experimental setting.

CONCLUSION

In this study, we showed that neutralization sensitivity of new HIV-1 variants to broadly neutralizing antibodies (bNAbs) is predictable using existing neutralization assays. The performance of the prediction models for the 11 considered bNAbs motivate their use in the selection of a bNAb combination therapy as a recommendation tool. The credibility of the models were enhanced by the finding that the prediction models learnt important binding sites for the bNAbs only based on the envelope sequence. Hence, additional information such as structural binding site information is unlikely to boost the performance. We increased the interpretability of the models, by offering the user more information on the prediction outcome in form of a motif logo where the logo displays the contribution of the pivotal residues of the test sequence to the prediction. In addition, we introduced a new general method that enables to visualize the feature space interrelations of the SVM models, providing thereby more information on the SVM classifiers.

Apart from their potential use as recommendation tool, the models can be used to analyze the change in the neutralization sensitivity of HIV-1 over time. We could confirm previous results suggesting a trend towards antibody resistance in the subtype B population (Bunnik et al., 2008; Bouvin-Pley et al., 2014, 2013). Moreover, we scaled up the analysis to the global HIV-1 population, showing that there is a general drift towards antibody resistance in the world-wide HIV-1 population. These findings are relevant for the selection of suitable vaccine candidates; a combination of bNAbs is however still very potent in neutralizing HIV-1 (Bouvin-Pley et al., 2014).

REFERENCES

Atchley, W. R., Zhao, J., Fernandes, A. D., and Drüke, T. (2005). Solving the protein sequence metric problem. *Proc. Natl. Acad. Sci. U. S. A.*, 102(18):6395–6400.

Barouch, D. H., Whitney, J. B., Moldt, B., Klein, F., Oliveira, T. Y., Liu, J., Stephenson, K. E., Chang, H.-W., Shekhar, K., Gupta, S., Nkolola, J. P., Seaman, M. S., Smith, K. M., Borducchi, E. N., Cabral, C., Smith, J. Y., Blackmore, S., Sanisetty, S., Perry, J. R., and Beck, M. (2013). Therapeutic efficacy of potent neutralizing hiv-1-specific monoclonal antibodies in shiv-infected rhesus monkeys. *Nature*, 503(7475):224 – 228.

Bouvin-Pley, M., Morgand, M., Meyer, L., Goujard, C., Moreau, A., Mouquet, H., Nussenzweig, M., Pace, C., Ho, D., Bjorkman, P. J., Baty, D., Chames, P., Pancera, M., Kwong, P. D., Poignard, P.,

Barin, F., and Braibant, M. (2014). Drift of the hiv-1 envelope glycoprotein gp120 toward increased neutralization resistance over the course of the epidemic: a comprehensive study using the most potent and broadly neutralizing monoclonal antibodies. *Journal of Virology*, 88(23):13910–13917.

Bouvin-Pley, M., Morgand, M., Moreau, A., Jestin, P., Simonnet, C., Tran, L., Goujard, C., Meyer, L., Barin, F., and Braibant, M. (2013). Evidence for a continuous drift of the hiv-1 species towards higher resistance to neutralizing antibodies over the course of the epidemic. *PLoS Pathog*, 9(7):e1003477.

Braun, W. and Venkatarajan, M. S. (2001). New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical-chemical properties. *J. Mol. Model.*, 7(12):445–453.

Bunnik, E. M., Pisas, L., van Nuenen, A. C., and Schuitemaker, H. (2008). Autologous neutralizing humoral immunity and evolution of the viral envelope in the course of subtype b human immunodeficiency virus type 1 infection. *Journal of Virology*, 82(16):7932–7941.

Burton, D., Pyati, J., Koduri, R., Sharp, S., Thornton, G., Parren, P., Sawyer, L., Hendry, R., Dunlop, N., Nara, P., and et, a. (1994). Efficient neutralization of primary isolates of hiv-1 by a recombinant human monoclonal antibody. *Science*, 266(5187):1024–1027.

Caragea, D., Cook, D., and Honavar, V. (2001). Gaining insights into support vector machine pattern classifiers using projection-based tour methods. In *INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING*, pages 251–256. ACM.

Caskey, M., Klein, F., Lorenzi, J. C. C., Seaman, M. S., West Jr, A. P., Buckley, N., Kremer, G., Nogueira, L., Braunschweig, M., Scheid, J. F., Horwitz, J. A., Shimeliovich, I., Ben-Avraham, S., Witmer-Pack, M., Platten, M., Lehmann, C., Burke, L. A., Hawthorne, T., Gorelick, R. J., Walker, B. D., Keler, T., Gulick, R. M., Fatkenheuer, G., Schlesinger, S. J., and Nussenzweig, M. C. (2015). Viraemia suppressed in hiv-1-infected humans by broadly neutralizing antibody 3bnc117. *Nature*.

Crooks, G. E., Hon, G., Chandonia, J.-M., and Brenner, S. E. (2004). Weblogo: A sequence logo generator. *Genome Research*, 14(6):1188–1190.

Doria-Rose, N. A., Klein, R. M., Manion, M. M., O'Dell, S., Phogat, A., Chakrabarti, B., Hallahan, C. W., Migueles, S. A., Wrammert, J., Ahmed, R., Nason, M., Wyatt, R. T., Mascola, J. R., and Connors, M. (2009). Frequency and phenotype of human immunodeficiency virus envelope-specific B cells from patients with broadly cross-neutralizing antibodies. *J. Virol.*, 83(1):188–99.

Evans, M. C., Phung, P., Paquet, A. C., Parikh, A., Petropoulos, C. J., Wrin, T., and Haddad, M. (2014). Predicting HIV-1 broadly neutralizing antibody epitope networks using neutralization titers and a novel computational method. *BMC Bioinformatics*, 15(1):77.

Falkowska, E., Ramos, A., Feng, Y., Zhou, T., Moquin, S., Walker, L. M., Wu, X., Seaman, M. S., Wrin, T., Kwong, P. D., Wyatt, R. T., Mascola, J. R., Poignard, P., and Burton, D. R. (2012). PGV04, an HIV-1 gp120 CD4 binding site antibody, is broad and potent in neutralization but does not induce conformational changes characteristic of CD4. *J. Virol.*, 86(8):4394–403.

Foley, B., Leitner, T., Apetrei, C., Hahn, B., Mizrachi, I., Mullins, J., Rambaut, A., Wolinsky, S., and Korber, B. (2013). Hiv sequence compendium 2013.

Huang, J., Kang, B. H., Pancera, M., Lee, J. H., Tong, T., Feng, Y., Georgiev, I. S., Chuang, G.-Y., Druz, A., Doria-Rose, N. A., Laub, L., Sliepen, K., van Gils, M. J., de la Peña, A. T., Derking, R., Klasse, P.-J., Migueles, S. A., Bailer, R. T., Alam, M., Pugach, P., Haynes, B. F., Wyatt, R. T., Sanders, R. W., Binley, J. M., Ward, A. B., Mascola, J. R., Kwong, P. D., and Connors, M. (2014). Broad and potent HIV-1 neutralization by a human antibody that binds the gp41-gp120 interface. *Nature*, advance on.

Julien, J.-P., Cupo, A., Sok, D., Stanfield, R. L., Lyumkis, D., Deller, M. C., Klasse, P.-J., Burton, D. R., Sanders, R. W., Moore, J. P., Ward, A. B., and Wilson, I. A. (2013). Crystal structure of a soluble cleaved HIV-1 envelope trimer. *Science*, 342(6165):1477–83.

Klein, F., Halper-Stromberg, A., Horwitz, J. A., Gruell, H., Scheid, J. F., Bournazos, S., Mouquet, H., Spatz, L. A., Diskin, R., Abadir, A., Zang, T., Dorner, M., Billerbeck, E., Labitt, R. N., Gaebler, C., Marcovecchio, P. M., Incesu, R.-B., Eisenreich, T. R., Bieniasz, P. D., and Seaman, M. S. (2012). Hiv therapy by a combination of broadly neutralizing antibodies in humanized mice. *Nature*, 492(7427):118 – 122.

Kruskal, J. B. and Wish, M. (1978). *Multidimensional scaling*, volume 11. Sage.

Lacerda, M., Moore, P. L., Ngandu, N. K., Seaman, M., Gray, E. S., Murrell, B., Krishnamoorthy, M., Nonyane, M., Madiga, M., Wibmer, C. K., Sheward, D., Bailer, R. T., Gao, H., Greene, K. M., Karim, S. S., Mascola, J. R., Korber, B. T., Montefiori, D. C., Morris, L., Williamson, C., and Seoighe, C. (2013). Identification of broadly neutralizing antibody epitopes in the HIV-1 envelope glycoprotein using evolutionary models. *Virol. J.*, 10(1):347.

Lengauer, T., Sander, O., Sierra, S., Thielen, A., and Kaiser, R. (2007). Bioinformatics prediction of hiv coreceptor usage. *Nature Biotechnology*, 25(12):1407 – 1410.

Louder, M. K., Crump, J. a., Koff, W. C., Kwong, P. D., Kapiga, S. H., Phogat, S., Boyington, J. C., Haynes, B. F., Ward, A. B., Bonsignori, M., Khayat, R., O'Dell, S., Sastry, M., Burton, D. R., Wyatt, R., Arthos, J., Mascola, J. R., Pejchal, R., Julien, J.-P., Gorman, J., Wilson, I. a., Nabel, G. J., Zhou, T., Louder, R., Zhang, B., Yang, Z.-Y., Yang, Y., Chuang, G.-Y., Lee, D., McLellan, J. S., Diwanji, D., Do Kwon, Y., Shahzad-ul Hussan, S., Zhu, J., Dai, K., Schmidt, S. D., Sam, N. E., Carrico, C., Bewley, C. a., Schief, W. R., Moquin, S., Orwenyo, J., Pancera, M., Walker, L. M., Wang, L.-X., Georgiev, I., and Patel, N. (2011). Structure of HIV-1 gp120 V1/V2 domain with broadly neutralizing antibody PG9. *Nature*, 480(7377):336–343.

Marshall, R. D. (1974). The nature and metabolism of the carbohydrate-peptide linkages of glycoproteins. *Biochem. Soc. Symp.*, (40):17–26.

Meinicke, P., Tech, M., Morgenstern, B., and Merkl, R. (2004). Oligo kernels for datamining on biological sequences: a case study on prokaryotic translation initiation sites. *BMC Bioinformatics*, 5:169.

Mouquet, H., Scharf, L., Euler, Z., Liu, Y., Eden, C., Scheid, J. F., Halper-Stromberg, A., Gnanapragasam, P. N. P., Spencer, D. I. R., Seaman, M. S., Schuitemaker, H., Feizi, T., Nussenzweig, M. C., and Bjorkman, P. J. (2012). Complex-type n-glycan recognition by potent broadly neutralizing hiv antibodies. *Proceedings of the National Academy of Sciences*, 109(47):E3268–E3277.

Pfeifer, N., Walter, H., and Lengauer, T. (2014). Association between hiv-1 coreceptor usage and resistance to broadly neutralizing antibodies. *Journal of acquired immune deficiency syndromes (1999)*, 67(2):107—112.

Rätsch, G., Sonnenburg, S., and Schölkopf, B. (2005). RASE: recognition of alternatively spliced exons in C.elegans. *Bioinformatics*, 21 Suppl 1:i369–77.

Scheid, J. F., Mouquet, H., Ueberheide, B., Diskin, R., Klein, F., Oliveira, T. Y. K., Pietzsch, J., Fenyo, D., Abadir, A., Velinzon, K., Hurley, A., Myung, S., Boulad, F., Poignard, P., Burton, D. R., Pereyra, F., Ho, D. D., Walker, B. D., Seaman, M. S., Bjorkman, P. J., Chait, B. T., and Nussenzweig, M. C. (2011). Sequence and structural convergence of broad and potent hiv antibodies that mimic cd4 binding. *Science*, 333(6049):1633–1637.

Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge University Press, illustrated edition edition.

Sonnenburg, S., Zien, A., Philips, P., and Rätsch, G. (2008). Poims: positional oligomer importance matrices—understanding support vector machine-based signal detectors. *Bioinformatics*, 24(13):i6–i14.

Taylor, B. S., Sobieszczyk, M. E., McCutchan, F. E., and Hammer, S. M. (2008). The challenge of HIV-1 subtype diversity. *N. Engl. J. Med.*, 358(15):1590–602.

Walker, L. M., Huber, M., Doores, K. J., Falkowska, E., Pejchal, R., Julien, J.-P. P., Wang, S.-K. K., Ramos, A., Chan-Hui, P.-Y. Y., Moyle, M., Mitcham, J. L., Hammond, P. W., Olsen, O. A., Phung, P., Fling, S., Wong, C.-H. H., Phogat, S., Wrin, T., Simek, M. D., Protocol G Principal Investigators, Koff, W. C., Wilson, I. A., Burton, D. R., and Poignard, P. (2011). Broad neutralization coverage of HIV by multiple highly potent antibodies. *Nature*, 477(7365):466–470.

Walker, L. M., Phogat, S. K., Chan-Hui, P.-Y., Wagner, D., Phung, P., Goss, J. L., Wrin, T., Simek, M. D., Fling, S., Mitcham, J. L., Lehrman, J. K., Priddy, F. H., Olsen, O. a., Frey, S. M., Hammond, P. W., Kaminsky, S., Zamb, T., Moyle, M., Koff, W. C., Poignard, P., and Burton, D. R. (2009). Broad and potent neutralizing antibodies from an African donor reveal a new HIV-1 vaccine target. *Science*, 326(5950):285–289.

Wang, X., Wu, S., Wang, X., and Li, Q. (2006). Svmv-a novel algorithm for the visualization of svm classification results. In Wang, J., Yi, Z., Zurada, J., Lu, B.-L., and Yin, H., editors, *Advances in Neural Networks - ISNN 2006*, volume 3971 of *Lecture Notes in Computer Science*, pages 968–973. Springer Berlin Heidelberg.

West, A. P., Diskin, R., Nussenzweig, M. C., and Bjorkman, P. J. (2012). Structural basis for germ-line gene usage of a potent class of antibodies targeting the CD4-binding site of HIV-1 gp120. *Proc. Natl. Acad. Sci. U. S. A.*, 109(30):E2083–90.

West, A. P., Scharf, L., Horwitz, J., Klein, F., Nussenzweig, M. C., and Bjorkman, P. J. (2013). Computational analysis of anti-HIV-1 antibody neutralization panel data to identify potential functional epitope residues. *Proc. Natl. Acad. Sci. U. S. A.*, 110(26):10598–603.

West, A. P., Scharf, L., Scheid, J. F., Klein, F., Bjorkman, P. J., and Nussenzweig, M. C. (2014). Structural Insights on the Role of Antibodies in HIV-1 Vaccine and Therapy. *Cell*, 156(4):633–648.

WHO (2014). http://www.who.int/en/.

Wu, X., Yang, Z.-Y., Li, Y., Hogerkorp, C.-M., Schief, W. R., Seaman, M. S., Zhou, T., Schmidt, S. D., Wu, L., Xu, L., Longo, N. S., McKee, K., O'Dell, S., Louder, M. K., Wycuff, D. L., Feng, Y., Nason, M., Doria-Rose, N., Connors, M., Kwong, P. D., Roederer, M., Wyatt, R. T., Nabel, G. J., and Mascola, J. R. (2010). Rational design of envelope identifies broadly neutralizing human monoclonal antibodies to HIV-1. *Science*, 329(5993):856–861.