Bioinformatics, 34(15), 2018, 2575–2580
doi: 10.1093/bioinformatics/bty170
Advance Access Publication Date: 15 March 2018
Original Paper

OXFORD

Sequence analysis

SCOTCH: subtype A coreceptor tropism classification in HIV-1

Hannah F. Löchel[1], Mona Riemenschneider[2], Dmitrij Frishman[3,4] and Dominik Heider[1,*]

[1]Department of Mathematics and Computer Science, Philipps-University of Marburg, 35032 Marburg, Germany, [2]Department of Bioinformatics, TUM Campus Straubing, 94315 Straubing, Germany, [3]Department of Genome-Oriented Bioinformatics, Technical University of Munich, 85354 Freising, Germany and [4]Laboratory of Bioinformatics, St. Petersburg State Polytechnic University, St. Petersburg 195251, Russia

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on November 22, 2017; revised on March 6, 2018; editorial decision on March 13, 2018; accepted on March 14, 2018

Abstract

Motivation: The V3 loop of the gp120 glycoprotein of the Human Immunodeficiency Virus 1 (HIV-1) is considered to be responsible for viral coreceptor tropism. gp120 interacts with the CD4 receptor of the host cell and subsequently V3 binds either CCR5 or CXCR4. Due to the fact that the CCR5 coreceptor is targeted by entry inhibitors, a reliable prediction of the coreceptor usage of HIV-1 is of great interest for antiretroviral therapy. Although several methods for the prediction of coreceptor tropism are available, almost all of them have been developed based on only subtype B sequences, and it has been shown in several studies that the prediction of non-B sequences, in particular subtype A sequences, are less reliable. Thus, the aim of the current study was to develop a reliable prediction model for subtype A viruses.

Results: Our new model SCOTCH is based on a stacking approach of classifier ensembles and shows a significantly better performance for subtype A sequences compared to other available models. In particular for low false positive rates (between 0.05 and 0.2, i.e. recommendation in the German and European Guidelines for tropism prediction), SCOTCH shows significantly better prediction performances in terms of partial area under the curves and diagnostic odds ratios compared to existing tools, and thus can be used to reliably predict coreceptor tropism for subtype A sequences.

Availability and implementation: SCOTCH can be downloaded/accessed at http://www.heiderlab.de.

Contact: dominik.heider@uni-marburg.de

1 Introduction

Infection of the host cells with the Human Immunodeficiency Virus 1 (HIV-1) proceeds in several steps that include the binding of the gp120 surface protein of HIV-1 to the CD4 receptor and a coreceptor, namely one of the chemokine receptors CCR5 or CXCR4 (Lee *et al.*, 1999). Coreceptor tropism, i.e. the type of coreceptor that is used by an HIV-1 virus, has important clinical implications. First, patients with a CXCR4-tropic virus progress faster to AIDS compared to patients with CCR5-tropic viruses (Koot *et al.*, 1993). Second, entry inhibitors that bind to the coreceptor and thus inhibit viral entry, such as Maraviroc (Dorr *et al.*, 2005), are only available for the CCR5 coreceptor, and are thus ineffective against CXCR4-tropic viruses. Today, entry inhibitors are frequently used in antiretroviral treatment, thus the determination of coreceptor tropism has become crucial for patient therapy. The gold standard for determining coreceptor tropism is by cell-based assays (Whitcomb *et al.*, 2007). The main disadvantages of cell-based assays are that they can only be carried out by specialized laboratories and that these assays are expensive and time-consuming. It has been shown in several studies, that computational approaches for coreceptor tropism

prediction can be a viable alternative to cell-based assays. The main advantage of these predictive models is that the procedure is cheap and very fast, in particular when these algorithms are executed in a parallelized manner, e.g. on graphics cards (Olejnik *et al.*, 2014). Due to the fact that the third variable loop of the gp120 protein (V3) is considered to be responsible for coreceptor usage, these models are typically trained on a set of V3 sequences with known tropism, and subsequently applied to new, unseen V3 sequences in order to predict tropism. Several models have been proposed, from simple rules, such as the 11/25 rule (Fouchier *et al.*, 1992; Shioda *et al.*, 1992), to sophisticated machine learning models. For instance, geno2pheno[coreceptor] is based on a support vector machine (Lengauer *et al.*, 2007) trained on V3 sequences. T-CUP uses structural information for modeling the electrostatic potential and hydrophobicity of the V3 sequences in order to predict coreceptor tropism (Dybowski *et al.*, 2010a,b; Heider *et al.*, 2014). PhenoSeq makes use of sequence motifs and predicted charges of the sequences (Cashin *et al.*, 2015), while WebPSSM (Jensen *et al.*, 2003) uses scoring matrices. These models have been shown to give reliable predictions and can be used for clinical assessment of coreceptor tropism. However, HIV-1 can be subdivided into different subtypes that show different abundancies and different spatial distributions. HIV-1 subtype B is mainly found in North America, the Caribbean, Latin America, Western and Central Europe and Australia and makes up 11% of the infections worldwide (Hemelaar *et al.*, 2011). Almost all available computational models have been trained on subtype B data. Subtype C makes up 48% of worldwide infections and is mainly found in Africa. It has been shown in several studies that the available models for coreceptor tropism can also be applied for subtype C sequences with comparable prediction accuracy (Gupta *et al.*, 2015; Riemenschneider *et al.*, 2016). The third major subtype, namely subtype A, is responsible for around 12% of worldwide infections and can be mainly found in Eastern Europe and Central Asia. Unfortunately, it has been demonstrated that the available models are not reliable for tropism prediction of subtype A viruses. The performance of the existing tools drops to less than 50% accuracy when applied to subtype A sequences (Riemenschneider *et al.*, 2016). They proposed that there may be a slightly different underlying mode of binding, which could involve other parts of gp120. The potential involvement of the V2 loop, for instance, is also mentioned by others (Kitawi *et al.*, 2017; Pastore *et al.*, 2006). Moreover, there is an apparent selection for subtype A variants that are less glycosylated and with shorter V1-V2 loop sequences (Chohan *et al.*, 2005). The aim of the current study was the development of a reliable subtype A specific coreceptor prediction model.

2 Materials and methods

2.1 Dataset

We used the dataset of V3 sequences of subtype A collected by Riemenschneider *et al.* (2016). The V3 loop sequences of HIV-1 with assigned subtype A or CRFs with a V3 region originating from subtype A were downloaded from the Los Alamos HIV sequence database (http://hiv-web.lanl.gov) in March 2015. Sequences that occur with contradictory tropism annotation in the database or contain non-canonical amino acid symbols were removed. Duplicated sequences were included only once. Additionally, nine subtype A sequences that were collected at the Institute of Virology at the University of Cologne were used as well. The final dataset consists of 182 V3 sequences of subtype A from 147 R5-tropic and 35 X4-tropic viruses.

2.2 Phylogenetic analysis of the samples

We performed a phylogenetic analysis of the V3 sequences in order to confirm the assigned subtypes. To this end, a multiple sequence alignment (MSA) of the V3 sequences was computed with MUSCLE (Edgar, 2004). The MSA was used to generate phylogenetic trees with SeaView 4 (Gouy *et al.*, 2010) using Poisson distance and BioNJ (Gascuel, 1997). Gap sites were ignored and significance was estimated by bootstrapping with 100 replicates.

2.3 Feature encoding

It has been shown in several studies that the most crucial part in predictive modeling is the feature encoding, i.e. the encoding of the protein sequences. In order to improve ensemble diversity (Kuncheva and Whitaker, 2003), we made use of structural and sequence information of the V3 loop according to Dybowski *et al.* (2010a), which have been demonstrated to give reliable prediction on HIV tropism, when used as input in subsequent machine learning models. Introducing structural information into classification models has been demonstrated to improve overall prediction performance (Bozek *et al.*, 2013; Dybowski *et al.*, 2011; Sander *et al.*, 2007). To this end, we used the same overall architecture as for subtype B sequences (Dybowski *et al.*, 2010a), however, due to the higher flexibility of the subtype A V3 loop sequences, the distances needed to be adapted. We encoded the V3 sequences in two ways: (i) by building homology models of the V3 loop and calculating the electrostatic potential on the surface, and (ii) by using a hydrophobicity encoding of the V3 sequences (see Fig. 1). We employed Modeller (v 9.17) (Šali and Blundell, 1993) in order to generate homology models of the V3 sequences, using the X-ray structure of the gp120 protein (PDB: 2QAD) as a template. The V3 loop sequences were aligned pairwisely against the sequence of the template structure with the R package bio3d (Grant *et al.*, 2006). Upon visual inspection we found that the alignments were of good quality and there was no need to manually adjust them. Ten models were generated, and for each V3 sequence, we selected the structure with the highest DOPE-Score (Elias *et al.*, 1991) for subsequent analyses. Next, we calculated the electrostatic potential at the surface of the V3 structures using the AMBER force field and PDB2PQR (v 2.1.1) (Dolinsky *et al.*, 2004). The solvent accessible surface of the structures was determined by APBS (v 1.4.2.1) (Baker *et al.*, 2001) using a grid of 33^3 points with a spacing of 3 Å according to Dybowski *et al.* (2010a), which have been demonstrated to give reliable prediction on HIV resistance and tropism, and a radius of the solvent molecules of 1.4 Å. In order to find the best distance for our prediction model, we evaluated the electrostatic hull at distances of 0, 3, 6, 9 and 12 Å from the solvent accessible surface. The hydrophobicity encoding of the V3 sequences was generated with Interpol (v 1.3) (Heider and Hoffmann, 2011). V3 sequences were translated into their numerical hydrophobicity representation according to the Kyte-Doolittle hydropathy index (Kyte and Doolittle, 1982). Subsequently, the numerical hydrophobicity vectors were interpolated to a common length of 35, which represents the average length in our dataset. We also evaluated whether an additional feature encoded in Interpol (Heider and Hoffmann, 2011) could improve overall prediction performance. However, we found no significant improvements of the resulting models in terms of AUC, compared to the model based on hydrophobicity and electrostatics.

2.4 Machine learning

The dataset was not balanced prior to training. In imbalanced data, there are two opportunities in order the balance the dataset, namely

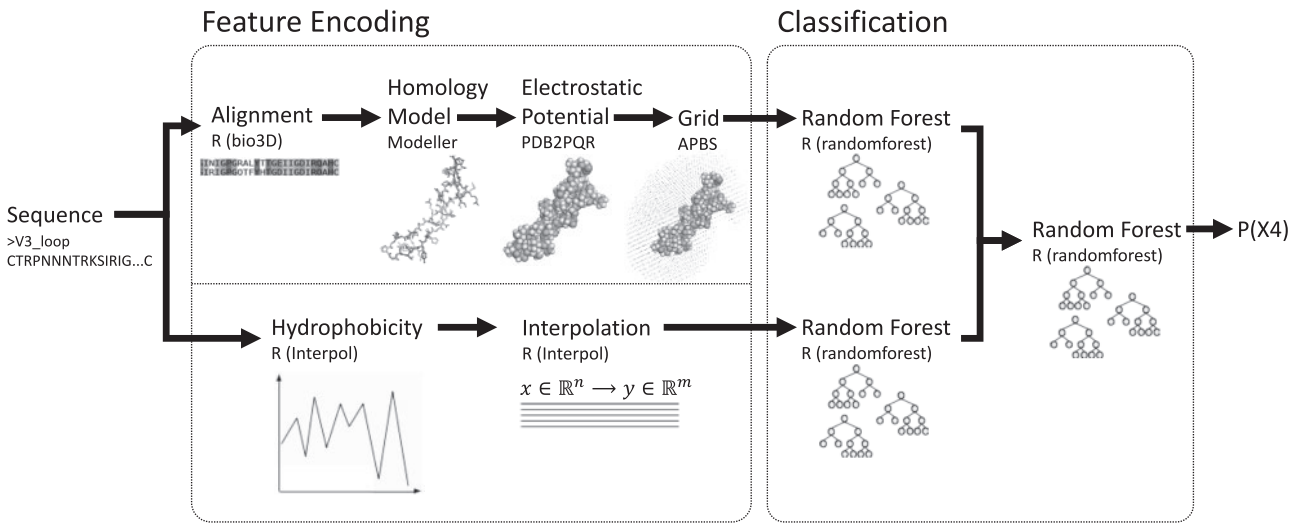

Fig. 1. Overview: V3 sequences are encoded in two ways: i) by building homology models and calculating the electrostatic potential with APBS, and ii) by using hydrophobicity encoding via Interpol. Two separate random forests were trained and combined via stacking

upsampling of the minority class or downsampling of the majority class. Although balancing has been shown to improve the overall performance metrics, e.g. AUC, upsampling may lead to overestimation of the AUC and downsampling may lead to a significant loss of information. Due to the fact that the amount of available subtype A V3 sequences with known tropism is limited, we used an imbalanced dataset in the current study.

In order to make use of the two encodings mentioned above, we employed a stacking approach (Wolpert, 1992). We trained two separate random forests (RFs) (Breiman, 2001), one based on the electrostatic potentials and another on the hydrophobicity-encoded V3 sequences using the randomForest package in R. For each approach, we trained ten RFs with 3000 trees. The RFs were evaluated using the internal out-of-bag estimation, which is based on bootstrapping. RFs have been shown in several studies to be highly accurate classifiers and less prone to overfitting compared to other machine learning approaches. Besides producing accurate predictions, RFs can also be used to estimate the importance of features. We used the Gini-index in order to estimate feature importance. The outputs of the electrostatic-RF and the hydrophobicity-RF, i.e. the RF trained on the electrostatic potentials and the hydrophobicity vectors, respectively, are combined by a stacking approach with a third RF. The RFs were evaluated by receiver operating characteristics (ROC) analyses using the R packages ROCR (Sing *et al.*, 2005) and pROC (Robin *et al.*, 2011). In ROC analysis, the true positive rate (TPR) is plotted against the false positive rate (FPR):

$$TPR = \frac{TP}{TP + FN} = sensitivity \tag{1}$$

$$FPR = \frac{FP}{FP + TN} = 1 - specificity \tag{2}$$

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{3}$$

with TP: true positives, FN: false negatives, FP: false positives, TN: true negatives. Besides the area under the curve (AUC), we also calculated corrected partial AUCs for low FPRs and the Diagnostic Odds Ratio (DOR) for FPRs of 0.05, 0.1, 0.15 and 0.2 in order to reflect the performance of the models with respect to current treatment guidelines (Vandekerckhove *et al.*, 2011) for entry inhibitors. The DOR (Glas *et al.*, 2003) is defined as

$$DOR = \frac{TP/FP}{FN/TN} \tag{4}$$

2.5 Comparison with other methods

We compared our novel subtype A prediction model SCOTCH with existing models, namely T-CUP (Heider *et al.*, 2014), geno2pheno [coreceptor] (Lengauer *et al.*, 2007), PhenoSeq (Cashin *et al.*, 2015), WebPSSM (Jensen *et al.*, 2003) using all available matrices (i.e. x4r5, sinsi and sinsi c), and the genotypic rules of Raymond *et al.* (2012) and Esbjörnsson *et al.* (2010).

3 Results

3.1 Overall approach

The aim of the study was to build a reliable coreceptor tropism prediction for HIV-1 subtype A. Our model SCOTCH is based on a stacking approach of two random forests (RFs) that were trained on different feature encoding in order to improve classifier diversity. The first RF was trained on the electrostatic potentials at the surface of the V3 structure models. The second RF was trained on the numerical hydrophobicity representations of the V3 sequences. The outputs of these RFs are combined via stacking. To this end, a third RF uses the outputs, i.e. pseudo-probabilities, and makes a final prediction whether a given V3 sequence belongs to a CCR5- or CXCR4-tropic virus. The performance of SCOTCH was compared with the existing models.

3.2 Electrostatics hull

The sequences were aligned using MUSCLE (Edgar, 2004) in a pairwise manner (all-against-all). 117 V3 sequences (64.3% of all sequences) share at least 90% identity of their sequence with at least one other sequence in the dataset. For 48 V3 sequences (i.e. 26.4%), there is at least one other V3 sequence with a similarity of $\geq$97.1%. The R5-tropic sequences show a higher average similarity compared to the X4 sequences (91.8 and 86.7%, respectively). We used

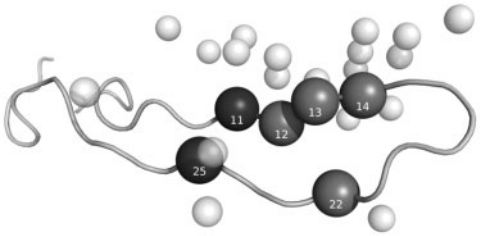

Fig. 2. Most important grid points in the template structure: The twenty most important grid points identified by the RF are plotted as white spheres around the V3 loop structure of the template. The Cα atoms of residues neighboring important grid points are shown in grey. Residues 11 and 25 are highlighted in black

Modeller (Šali and Blundell, 1993) in order to generate the V3 structures for the prediction of coreceptor tropism, and APBS Baker *et al.* (2001) in order to calculate the electrostatic hulls at different distances to the surface. For the discretization of the electrostatic hulls in order to be used in the subsequent classification models, we used a grid spacing of 3 Å, which was centered over each V3 loop structure. For the electrostatics hull at a distance of 0 Å, all 182 V3 loops are too close to the hull or even penetrate it, so they have at least one grid point that is not accessible by the solvent. For a distance of 3 and 6 Å, 177 and 65 sequences, respectively, still penetrate the hull. In contrast to our results with subtype B sequences (Dybowski *et al.*, 2010a), only the electrostatic hulls at a distance of 9 Å completely enclose all V3 loop structures. These findings might imply that, on average, V3 sequences from subtype A are more disordered compared to subtype B, which have been used in former studies. These findings are in line with the notion that subtype A sequences show biochemical differences compared to subtype B sequences (Chohan *et al.*, 2005).

3.3 Electrostatics-based classification

A homology model for each V3 loop was generated based on the template X-ray structure of the viral gp120 protein (PDB: 2QAD). Electrostatic potentials at discrete grid for each V3 loop structure were obtained by solving the non-linear Poisson-Boltzmann equation by APBS (Baker *et al.*, 2001). We used the RFs to estimate feature importance. Figure 2 shows the twenty most important positions on the V3 structure according to the RF importance analysis. The most important positions cluster around residues 11–14 and residue 25, which is in partial agreement with the 11/25 rule (Fouchier *et al.*, 1992; Shioda *et al.*, 1992). Residue 25 might be less important than residues 11 to 14, given the fact that only two out of the twenty most important grid points can be assigned to this residue.

3.4 Hydrophobicity-based classification

In addition to the RF trained on the electrostatic hulls, we also trained a model based on the hydrophobicity encoding. The V3 loop sequences were encoded with the Kyte-Doolittle hydropathy index (Kyte and Doolittle, 1982) using Interpol (Heider and Hoffmann, 2011) and normalized to the average V3 length, i.e. 35 residues. Figure 3 shows the importance of the residues of the normalized V3 sequences according to the hydrophobicity-based RF. Two important clusters can be identified at positions 10–14 and 22–25. Again, these findings are in agreement with the 11/25 rule. Due to the fact that the dataset contains sequences shorter than 35 amino acids, residue positions in the interpolated sequences are again slightly shifted to the right and do not necessarily correspond to the actual residues.

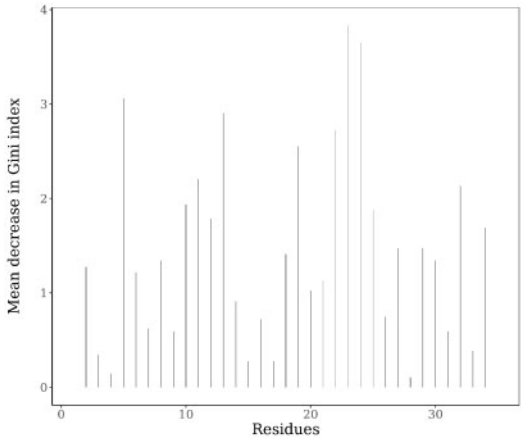

Fig. 3. Importance of the V3 loop residues: The importance has been estimated using the Gini-index

3.5 Stacking approach

We used a stacking approach in order to combine both models. To this end, the pseudo-probabilities of both models were used as an input for a third RF. Figure 4 shows the ROC curves of the electrostatics- and hydrophobicity-based models. The AUC of the RF trained on the electrostatic hulls is 0.7704 ± 0.0043, while the RF trained on hydrophobicity reaches an AUC of 0.7004 ± 0.0023. However, the difference in AUC is not significant ($P = 0.1899$). The electrostatics-based model reaches higher true TPRs for almost all FPRs, thus it is not obvious why the use of stacking could improve overall performance. Nevertheless, the third combined RF outperformed the other two RFs significantly at low FPRs. The AUC of the combined RF (i.e. our final method SCOTCH) is 0.7031 ± 0.0046, which is not significantly higher compared to the model trained solely on the electrostatic hull ($P = 0.2216$). However for low FPRs (<0.2) the combined model outperforms the single models (see Fig. 4). For instance, the European guidelines recommend the use of a 10% FPR cutoff. The corrected partial AUCs for the combined model is 0.7498 for an FPR between 0.05 and 0.2, which is significantly higher compared to the corrected partial AUCs of 0.6993 and 0.5969 for the electrostatics- and hydrophobicity-based model, respectively. The sensitivity and accuracy at a specificity of 95, 90 and 85% for the different models are listed in Table 1. For a specificity of 95%, the electrostatics-based model achieves a sensitivity of 40.0% and an accuracy of 84.4%. The combined method has a higher sensitivity (47.7%) and a higher accuracy (85.9%). The DOR is also higher for the combined model at low FPRs (see Table 1).

3.6 Comparison with other methods

The prediction performance of SCOTCH was compared with existing methods, namely T-CUP (Heider *et al.*, 2014), geno2pheno[coreceptor] (Lengauer *et al.*, 2007), PhenoSeq (Cashin *et al.*, 2015), WebPSSM (Jensen *et al.*, 2003) using all available matrices (i.e. x4r5, sinsi and sinsi c), and the genotypic rules of Raymond *et al.* (2012) and Esbjörnsson *et al.* (2010). In Table 2, the results of the comparison are shown. For all existing methods, except WebPSSM with sinsi.c, the sensitivity is less than 20% at a specificity between 93.94 and 99.49%. WebPSSM with sinsi.c only reaches a sensitivity of 37.8% at a specificity of 58.59%. SCOTCH outperforms all existing models in terms of sensitivity and accuracy at comparable specificities. For instance, PhenoSeq reaches a specificity of 94.74%, which is only slightly lower compared to SCOTCH (95%).

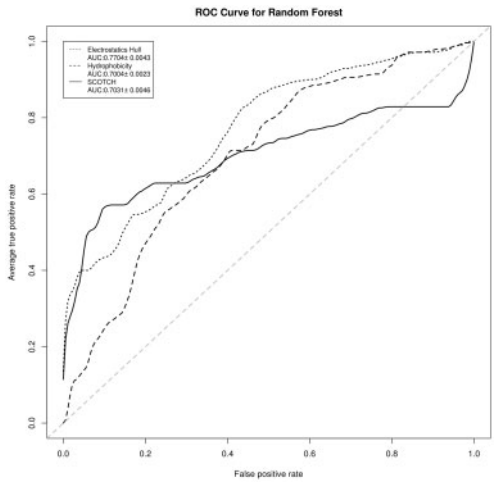

Fig. 4. ROC curves of all RF models: The combined SCOTCH method incorporates both the 'Electrostatics Hull' and the 'Hydrophobicity' RFs

Table 1. Performance of all methods

Method	Sensitivity	Specificity	Accuracy	DOR
Electrostatics	40.00	95.00	84.42	15.67
Hydrophobicity	14.00	95.00	79.42	3.85
Combined model (SCOTCH)	47.71	95.00	85.91	17.07
Electrostatics	43.43	90.00	81.04	7.92
Hydrophobicity	24.57	90.00	77.42	3.36
Combined model (SCOTCH)	56.57	90.00	83.57	20.08
Electrostatics	51.71	85.00	78.60	6.45
Hydrophobicity	31.43	85.00	74.70	2.76
Combined model (SCOTCH)	57.14	85.00	79.64	13.37

Note: Performance measures were calculated at 95, 90 and 85% specificity.

Table 2. Comparison of tropism prediction models on subtype A sequences

Method	Sensitivity	Specificity	Accuracy
T-CUP	18.18	96.32	55.39
geno2pheno	15.79	97.89	54.89
Phenoseq	17.70	94.74	54.39
WebPSSM-x4r5	15.31	93.94	53.56
WebPSSM-sinsi	11.54	97.98	53.69
WebPSSM-sinsi.c	37.80	58.59	47.91
Raymond	11.00	98.48	53.56
Esbjörnsson	13.40	99.49	55.28
SCOTCH	47.71	95.00	85.91

However, the resulting sensitivity of SCOTCH is 47.71%, which is significantly higher compared to the sensitivity of PhenoSeq, namely 17.7%. The accuracy of PhenoSeq is 54.39%, which is again significantly lower than the accuracy of our new model (85.91%).

4 Discussion

Almost all existing approaches for the prediction of HIV-1 coreceptor tropism are based on subtype B and have been shown to perform poorly on subtype A, which is responsible for approximately 12% of all HIV-1 infections worldwide. We therefore sought to develop a reliable subtype A specific model. To this end, two RF models were developed, one based on electrostatics and the other one based on a hydrophobicity index. These two models were combined by using a stacking approach. The resulting model SCOTCH shows significantly better performance for subtype A compared to all other methods that have been evaluated. Nevertheless, the sensitivity and accuracy of our new model still did not reach the same levels than those of the prediction methods for subtype B or C. In this study we developed a novel coreceptor tropism prediction algorithm which makes use of sequence and structural information of the V3 loop for subtype A. Combining structural and sequence information improves diversity in ensembles and thus leads to higher prediction performance compared to single models. We could demonstrate that SCOTCH outperforms existing approaches, but there is still room for improvement. Riemenschneider *et al.* (2016) already proposed that other regions beside the V3 loop, namely the V2 loop, might also be involved in coreceptor binding in subtype A. Involvement of V2 information might improve coreceptor tropism prediction of subtype A sequences in the future. However, although sequencing of the complete gp120 region might be useful in order to improve prediction accuracy for subtype A, it is not really practical due to length restrictions in the sequencing protocols at the moment. Nevertheless, there is little doubt that technological progress will soon make the sequencing of gp120, or even whole viral genomes, feasible for routine diagnostics.

5 Conclusion

Our new model SCOTCH was specifically developed for HIV-1 subtype A and it has been shown in our study to outperform the existing models, which have mostly been trained with subtype B sequences. However, lower prediction performance of SCOTCH compared to subtype B prediction tools implies the existence of a different binding mode of subtype A to the host cell, as already proposed by Riemenschneider *et al.* (2016). Other regions, in particular the V2 loop, might be also involved in this process. In the future we intend to improve our model by also incorporating other regions of the gp120 protein. However, incorporating longer reads may be impractical due to the distance between the V3 and V2 loops. In order to get V2 and V3 information simultaneously, a region of around 430 nucleotides within the env gene needs to be sequenced.

Acknowledgements

We thank Cedric Staniewski (TUM) for supporting the analyses.

Funding

This work was supported by a grant of the Federal Ministry of Education and Research (BMBF) and the German Academic Exchange Service (DAAD) under the ATN-DAAD Joint Research Cooperation Scheme to DH. DF acknowledges the support of the Russian Science Foundation (grant number 5-14-00026).

Conflict of Interest: none declared.

References

Baker,N.A. *et al.* (2001) Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. USA*, **98**, 10037–10041.

Bozek,K. *et al.* (2013) Analysis of physicochemical and structural properties determining hiv-1 coreceptor usage. *PLoS Comput. Biol.*, **9**, e1002977.

Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

Cashin,K. *et al*. (2015) Reliable genotypic tropism tests for the major hiv-1 subtypes. *Sci. Rep*., **5**, 8543.

Chohan,B. *et al*. (2005) Selection for human immunodeficiency virus type 1 envelope glycosylation variants with shorter v1-v2 loop sequences occurs during transmission of certain genetic subtypes and may impact viral rna levels. *J. Virol*., **79**, 6528–6531.

Dolinsky,T.J. *et al*. (2004) Pdb2pqr: an automated pipeline for the setup of Poisson–Boltzmann electrostatics calculations. *Nucleic Acids Res*., **32**, W665–W667.

Dorr,P. *et al*. (2005) Maraviroc (uk-427, 857), a potent, orally bioavailable, and selective small-molecule inhibitor of chemokine receptor ccr5 with broad-spectrum anti-human immunodeficiency virus type 1 activity. *Antimicrob. Agents Chemother*., **49**, 4721–4732.

Dybowski,J.N. *et al*. (2010) Prediction of co-receptor usage of hiv-1 from genotype. *PLoS Comput. Biol*., **6**, e1000743.

Dybowski,J.N. *et al*. (2011) Improved bevirimat resistance prediction by combination of structural and sequence-based classifiers. *BioData Min*., **4**, 26.

Dybowski,N. *et al*. (2010) Structure of hiv-1 quasi-species as early indicator for switches of co-receptor tropism. *AIDS Res. Ther*., **7**, 41.

Edgar,R.C. (2004) Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*., **32**, 1792–1797.

Elias,D. *et al*. (1991) Vaccination against autoimmune mouse diabetes with a t-cell epitope of the human 65-kda heat shock protein. *Proc. Natl. Acad. Sci. USA*, **88**, 3088–3091.

Esbjörnsson,J. *et al*. (2010) Frequent cxcr4 tropism of HIV-1 subtype a and crf02_ag during late-stage disease-indication of an evolving epidemic in west africa. *Retrovirology*, **7**, 23.

Fouchier,R. *et al*. (1992) Phenotype-associated sequence variation in the third variable domain of the human immunodeficiency virus type 1 gp120 molecule. *J. Virol*., **66**, 3183–3187.

Gascuel,O. (1997) Bionj: an improved version of the nj algorithm based on a simple model of sequence data. *Mol. Biol. Evol*., **14**, 685–695.

Glas,A.S. *et al*. (2003) The diagnostic odds ratio: a single indicator of test performance. *J. Clin. Epidemiol*., **56**, 1129–1135.

Gouy,M. *et al*. (2010) Seaview version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol*., **27**, 221–224.

Grant,B.J. *et al*. (2006) Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics*, **22**, 2695–2696.

Gupta,S. *et al*. (2015) Performance of genotypic tools for prediction of tropism in hiv-1 subtype c v3 loop sequences. *Intervirology*, **58**, 1–5.

Heider,D. and Hoffmann,D. (2011) Interpol: an r package for preprocessing of protein sequences. *BioData Min*., **4**, 16.

Heider,D. *et al*. (2014) A simple structure-based model for the prediction of hiv-1 co-receptor tropism. *BioData Min*., **7**, 14.

Hemelaar,J. *et al*. (2011) Global trends in molecular epidemiology of hiv-1 during 2000–2007. *AIDS (London, England)*, **25**, 679.

Jensen,M.A. *et al*. (2003) Improved coreceptor usage prediction and genotypic monitoring of r5-to-x4 transition by motif analysis of human immunodeficiency virus type 1 env v3 loop sequences. *J. Virol*., **77**, 13376–13388.

Kitawi,R.C. *et al*. (2017) Partial hiv c2v3 envelope sequence analysis reveals association of coreceptor tropism, envelope glycosylation and viral genotypic variability among kenyan patients on haart. *Virol. J*., **14**, 29.

Koot,M. *et al*. (1993) Prognostic value of hiv-1 syncytium-inducing phenotype for rate of cd4+ cell depletion and progression to aids. *Ann. Intern. Med*., **118**, 681–688.

Kuncheva,L.I. and Whitaker,C.J. (2003) Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Mach. Learn*., **51**, 181–207.

Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol*., **157**, 105–132.

Lee,M.K. *et al*. (1999) Identification of determinants of interaction between cxcr4 and gp120 of a dual-tropic hiv-1dh12isolate. *Virology*, **257**, 290–296.

Lengauer,T. *et al*. (2007) Bioinformatics prediction of hiv coreceptor usage. *Nat. Biotechnol*., **25**, 1407.

Olejnik,M. *et al*. (2014) gcup: rapid gpu-based hiv-1 co-receptor usage prediction for next-generation sequencing. *Bioinformatics*, **30**, 3272–3273.

Pastore,C. *et al*. (2006) Human immunodeficiency virus type 1 coreceptor switching: v 1/v2 gain-of-fitness mutations compensate for v3 loss-of-fitness mutations. *J. Virol*., **80**, 750–758.

Raymond,S. *et al*. (2012) Phenotyping methods for determining hiv tropism and applications in clinical settings. *Curr. Opin. HIV AIDS*, **7**, 463–469.

Riemenschneider,M. *et al*. (2016) Genotypic prediction of co-receptor tropism of hiv-1 subtypes a and c. *Sci. Rep*., **6**, 24883.

Robin,X. *et al*. (2011) proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, **12**, 77.

Šali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol*., **234**, 779–815.

Sander,O. *et al*. (2007) Structural descriptors of gp120 v3 loop for the prediction of hiv-1 coreceptor usage. *PLoS Comput. Biol*., **3**, e58.

Shioda,T. *et al*. (1992) Small amino acid changes in the v3 hypervariable region of gp120 can affect the t-cell-line and macrophage tropism of human immunodeficiency virus type 1. *Proc. Natl. Acad. Sci. USA*, **89**, 9434–9438.

Sing,T. *et al*. (2005) Rocr: visualizing classifier performance in r. *Bioinformatics*, **21**, 3940–3941.

Vandekerckhove,L. *et al*. (2011) European guidelines on the clinical management of hiv-1 tropism testing. *Lancet Infect. Dis*., **11**, 394–407.

Whitcomb,J.M. *et al*. (2007) Development and characterization of a novel single-cycle recombinant-virus assay to determine human immunodeficiency virus type 1 coreceptor tropism. *Antimicrob. Agents Chemother*., **51**, 566–575.

Wolpert,D.H. (1992) Stacked generalization. *Neural Netw*., **5**, 241–259.