SCIENTIFIC REPORTS

OPEN

HIV-1 tropism prediction by the XGboost and HMM methods

Xiang Chen[1], Zhi-Xin Wang[1] & Xian-Ming Pan[2]

Received: 23 November 2018
Accepted: 25 June 2019
Published online: 10 July 2019

Human Immunodeficiency Virus 1 (HIV-1) co-receptor usage, called tropism, is associated with disease progression towards AIDS. Furthermore, the recently developed and developing drugs against co-receptors CCR5 or CXCR4 open a new thought for HIV-1 therapy. Thus, knowledge about tropism is critical for illness diagnosis and regimen prescription. To improve tropism prediction accuracy, we developed two novel methods, the extreme gradient boosting based XGBpred and the hidden Markov model based HMMpred. Both XGBpred and HMMpred achieved higher specificities (72.56% and 72.09%) than the state-of-the-art methods Geno2pheno (61.6%) and G2p_str (68.60%) in a 10-fold cross validation test at the same sensitivity of 93.73%. Moreover, XGBpred had more outstanding performances (with AUCs 0.9483, 0.9464) than HMMpred (0.8829, 0.8774) on the Hivcopred and Newdb (created in this work) datasets containing larger proportions of hard-to-predict dual tropic samples in the X4-using tropic samples. Therefore, we recommend the use of our novel method XGBpred to predict tropism. The two methods and datasets are available via http://spg.med.tsinghua.edu.cn:23334/XGBpred/. In addition, our models identified that positions 5, 11, 13, 18, 22, 24, and 25 were correlated with HIV-1 tropism.

Human Immunodeficiency Virus 1 (HIV-1) is a retrovirus which mainly infects T-lymphocytes, macrophages and dendritic cells[1]. HIV-1 enters into those host cells by chronologically interacting with primary receptors and co-receptors[2]. Fourteen co-receptors have been detected *in vitro*[3]. However, *in vivo*, the major co-receptors are CCR5 and CXCR4[4,5]. Indeed, a vast majority of subtype B and probably all subtype C HIV-1 positive individuals are initially infected via CCR5[2]. Viruses using CCR5 are known as R5 tropic, whereas viruses using CXCR4 are called X4 tropic. R5X4 or dual tropic viruses as a third class can bind to either CCR5 or CXCR4[6]. For simplicity, X4 and dual tropic viruses are called X4-using tropic.

R5 tropic viruses start the HIV-1 infection[7]. This start is shown by the HIV-1 resistance in individuals where the function of CCR5 is disabled by a homozygous *ccr5*-Δ32 gene[4,8,9]. Besides, X4-using tropic viruses are associated with disease progression, since those viruses emerge at the later stage of an infection in about half of the infected individuals[2,8,10–12]. Furthermore, Miraviroc (MVC), a CCR5 antagonist and the only FDA-approved entry inhibitor, binds to the hydrophobic transmembrane helices of CCR5 so as to allosterically inhibit viruses from entering[13]. It has been proved that MVC cannot transform R5 viruses into X4-using viruses[14,15]. Consequently, it becomes clear that tropism testing is necessary for several reasons: (1) To determine the illness progression[2,11]; (2) To decide whether MVC can be used[10]; and (3) To monitor changes in viral quasispecies in order to modify regimens in time[4].

In the last decades, two kinds of tropism testing methods, phenotypic and genotypic, have been developed. The phenotypic methods, such as ES-Trofile, are expensive, time-consuming, poorly accessible due to requiring specialized centers, and cannot provide consistent results when the viral load is below 1000 copies/ml[16]. Thus, the application of these methods is limited in clinical routines in Europe[5,8]. Instead, the genotypic tropism testing is a preferred method due to low cost, reduced turnaround time and great accessibility, even when the viral load is below 1000 copies/ml[17]. In contrast to phenotypic methods, genotypic methods are based on statistics or machine learning. These methods analyze the third variable (V3) loop of the viral glycoprotein gp120, which predominantly determines its tropism[18]. The earliest proposed genotypic method for prediction of X4-using tropism is the 11/25 rule. This rule is based on the presence of a positively charged amino acid in positions 11 or 25 of the V3 sequence[19]. Other genotypic methods such as WebPSSM[20,21] and CM[22] predict tropism based on scores that are calculated from position specific score matrices (PSSMs). In detail, WebPSSM constructs ungapped PSSMs,

[1]Key Laboratory of Ministry of Education for Protein Science, School of Life Sciences, Tsinghua University, Beijing, 100084, China. [2]Key Laboratory of Ministry of Education for Bioinformatics, School of Life Sciences, Tsinghua University, Beijing, 100084, China. Correspondence and requests for materials should be addressed to X.-M.P. (email: pan-xm@mail.tsinghua.edu.cn)

Subtype	Number (R[a], X[b], D[c])	Percentage
B	1503 (1209, 93, 201)	50.13%
C	511 (460, 26, 25)	17.04%
D	233 (120, 52, 61)	7.77%
01_AE	213 (149, 45, 19)	7.10%
A	155 (140, 5, 10)	5.17%
02_AG	124 (50, 3, 71)	4.14%

Table 1. Distribution of the six major subtypes in the Newdb dataset. Notes: [a]The number of R5 tropic sequences. [b]The number of X4 tropic sequences. [c]The number of dual tropic sequences.

		X4-using		
Dataset	R5	X4	Dual	Sum
Newdb	2335	245	418	2998
G2p_str[23]	973	94	121	1188
Hivcopred[a] [24]	1768	246	321	2335
CM[22]	2354	277	48	2679
WebPSSM[21]	228[b] (47[c])	51[b] (24[c])		279[b] (71[c])

Table 2. Distribution of tropisms in the different datasets. Notes: [a]Removed 31 duplicated sequences from the original Hivcopred dataset which are marked as not only R5 tropism but also X4-using tropism. [b]Training set. [c]Validation set.

while CM constructs gapped PSSMs and takes the 11/25 rule and net charge into consideration. Recently, many genotypic methods based on machine learning have also been published. The method Geno2pheno[4] combines two machine learning approaches, support vector machine (SVM) and decision trees, and uses clinical information such as viral loads and CD4-cell counts if available. Another method from the same laboratory, G2p_str[23], combines SVM and Lasso regression and uses the amino acid structure feature. Hivcopred[24] is based on SVM^{light} with the split amino acid composition feature. T-CUP2[25] employs random forests (RFs) with the structural information of hydrophobicity and electrostatic potential. Currently, Geno2pheno is the most widely used method and the only genotypic method recommended for usage in clinical routines by the European Consensus Group[5,8].

Genotypic methods can predict R5 viruses (~90%) accurately, but are inaccurate in the prediction of X4-using viruses (~50–70%)[26]. Thus, more accurate tropism prediction methods are required. Here, we present two methods, XGBpred and HMMpred. We analyzed the HIV-1 tropism prediction ability of our methods and compared them with the Geno2pheno, G2p_str, Hivcopred, CM and WebPSSM methods. The results show that XGBpred is robust with the hard-to-predict dual tropic sequences.

Methods

Datasets.

To construct the Newdb dataset, we extracted 6790 R5 tropic, 590 X4 tropic and 1125 dual tropic sequences from the Los Alamos HIV sequence database (http://www.hiv.lanl.gov/, last update: 10 Sep 2017). The tropisms of the sequences from the Los Alamos HIV sequence database have been phenotypically determined, none of them have been inferred from sequences. Then we removed sequences containing non-canonical residues, reserved sequences with lengths between 31 and 39, and dislodged duplicated sequences to guarantee the high quality of genotype/phenotype pairs. This process finally generated 2335 R5 and 663 X4-using (245 X4 and 418 dual) tropic sequences. The distribution of the six major subtypes in the Newdb dataset is shown in Table 1. To compare our methods with the Geno2pheno, G2p_str, Hivcopred, CM and WebPSSM methods, we used the datasets constructed in these studies, respectively. These datasets can be accessed in Supplementary Spreadsheet S1. The distributions of tropisms in different datasets are shown in Table 2.

Machine learning method: XGBpred.

Extreme gradient boosting (XGboost), like RFs used by T-CUP2[25], is an ensemble algorithm of decision trees[27]. The ensemble works by combining a set of weaker machine learning algorithms to get an improved machine learning algorithm in overall. The main difference between XGboost and RFs is the way of sampling. RFs are based on uniform sampling with return. Instead, XGboost gives higher weights to the wrongly predicted samples in the current weaker leaner, and then these samples will be paid more attention when training the next weaker leaner. In addition, XGboost adds regularization to avoid overfitting. Therefore, XGboost is a more complicated algorithm than RFs, and thus always outperforms.

Because XGboost is designed for vectors, it is necessary to convert V3 loop string sequences of different lengths to numerical vectors. For this task, we used many kinds of features to describe the characteristics of protein sequences, such as split amino acid composition[24], dipeptide composition[28], and net charge or hydropath[29]. We also proposed an additional set of features: the alignment score. The 35-dimensional alignment scores were generated by scoring alignments using the block substitution matrices BLOSUM62, BLOSUM90 or BLOSUM100[30], and the alignments were generated by aligning sequences to the consensus sequence with

35 residues by the means of Needleman-Wunsch (Version: EMBOSS: 6.6.0)[31]. For the XGBpred method, we tested these different features and their combinations to find the optimal model to discriminate R5 and X4-using sequences.

Statistics method: HMMpred.

Hidden Markov model (HMM) is a finite model applied in time series and linear sequences. Just as the PSSM profile, HMM also can be used to describe protein families. The HMM profile described by state-transition and symbol-emission probabilities performs better than PSSM in terms of sequence alignment and homology recognition because it can deal with gaps in protein families better by hidden state chains[32].

HMM profile construction. We used the maximum likelihood estimation method to establish R5 and X4-using specific HMM profiles from R5 and X4-using tropic multiple sequence alignments generated by ClustalO[33], respectively. In addition, we simply assigned columns that had more than half gap characters as insertion states. The structure of HMM that we used was no transition allowed from D_j to I_j or from I_j to D_{j+1} (This kind of structure performed better than the full structure, as shown in Supplementary Table S1). M, D, and I denote match, deletion and insertion states, respectively.

$$\hat{a}_{kl} = \frac{A_{kl} + 1}{\sum_{l'} A_{kl'} + 3} \tag{1}$$

$$\hat{e}_k(a) = \frac{E_k(a) + 1}{\sum_{a'} E_k(a') + 21} \tag{2}$$

Where in, k and l are indices over states M, D, or I; a is an amino acid symbol or gap; $\hat{a}_{kl}$ means the estimated probability of transiting from state k to state l, $\hat{e}_k(a)$ means the estimated probability of emitting residue a at state k, and A_{kl} and $E_k(a)$ are the corresponding frequencies. In order to avoid the zero probability which represents it cannot happen in the future, we applied the Laplace's pseudo-count rule that added one to each frequency.

Sequence-profile alignment. We employed Viterbi algorithm[34], a dynamic programing algorithm, to get two alignment scores S_{R5} and $S_{non\text{-}R5}$. Those alignment scores represent the optimal state pathway scores from the R5 and X4-using HMM profiles, respectively. the final score was defined as:

$$\mathrm{S} = S_{R5} - S_{non-R5} \tag{3}$$

Then the given sequence would be classified as R5 tropic if the final score S is higher than a threshold, otherwise it would be classified as X4-using tropic.

Ten-fold cross validation.

The widely-used 10-fold cross validation was used to evaluate the performance of our methods in this study, where the sequences were divided into 10 subsets randomly, one subset was used as the testing set, and the others were used as the training set. After ten repetitions, the final performance was average of the performances of those ten subsets.

Evaluation parameters.

For evaluation, we used sensitivity, specificity, accuracy and Matthew's correlation coefficient (MCC). In particular, MCC is robust even when the size of classes varies widely[35]. An MCC value '0' corresponds to a completely random prediction, while '1' corresponds to a perfect perdition. These parameters were calculated using the following equations:

$$\text{Sensitivity} = \frac{\text{TP}}{TP + FN} \tag{4}$$

$$\text{Specificity} = \frac{\text{TN}}{FP + TN} \tag{5}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{TP + FP + TN + FN} \tag{6}$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{7}$$

where TP is the number of true positives, FP false positives, TN true negatives and FN false negatives. We regarded R5 tropic samples as positives in this study.

In contrast to the four threshold-dependent parameters, the receiver operating characteristic (ROC) curve, a threshold-independent parameter, illustrates the trade-off between sensitivity and specificity at various threshold settings. In this study, we used the area under the curve (AUC) to measure a predictive power, where 0.5 means a random method, and 1 means a perfect method[36].

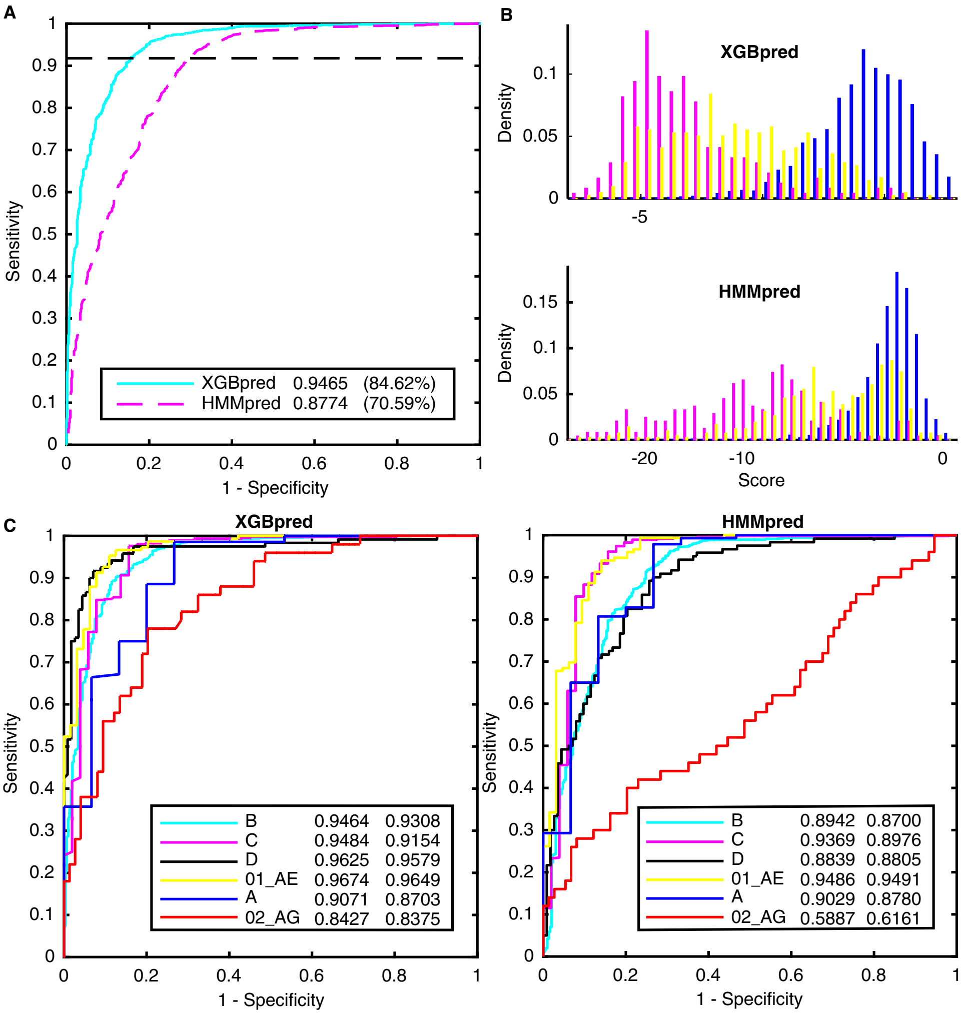

Figure 1. Performance of the XGBpred and HMMpred methods on the Newdb dataset. (**A**) ROC curves on the Newdb dataset in a same 10-fold cross validation test. The legend lists AUCs and specificities at the sensitivity of 91.78% which is plotted as the dashed black line. (**B**) Distribution of V3 loop sequence scores calculated from XGBpred and HMMpred on the Newdb dataset. The score distribution of the R5 tropic sequences is shown in blue, that of X4 is carmine and that of dual is yellow. (**C**) ROC curves of XGBpred and HMMpred for the six major subtypes. The legend lists AUCs and mAPs.

Results

Performance on the Newdb dataset.

The feature set and the model that gave the strongest predictive power for the XGBpred and HMMpred methods were found, respectively (Supplementary Tables S1 and S2). The performances of the two methods on the Newdb dataset in a same 10-fold cross validation test are shown in Fig. 1A and Table 3. XGBpred had a higher specificity, accuracy, MCC and AUC than HMMpred when having the same sensitivity. Furthermore, the specificity of XGBpred was higher than 80% (84.62%) at the sensitivity of 91.78%. Results from the two methods were highly consistent: they predicted same tropisms for 87.96% of total samples, and achieved 96.70% sensitivity, 83.39% specificity and 93.93% accuracy.

Considering the poorer performance of HMMpred, the score distributions of the two methods were plotted (Fig. 1B). As depicted, the scores of dual tropic sequences mostly placed in the middle of the scores of X4 and R5 tropic sequences. Furthermore, HMMpred generated higher scores for a considerable number of dual tropic samples than XGBpred. This phenomenon illustrates that it is hard for dual tropic sequences to be correctly classified, especially by HMMpred.

The performances of the two methods for the six major subtypes (subtypes B, C, D, 01_AE, A and 02_AG) in the Newdb dataset were analyzed due to the sequence divergence among different subtypes and the different number of sequences in each subtype (Fig. 1C). HMMpred for subtypes B and D showed much lower AUCs (0.8942 and 0.8839) than for subtypes C and 01_AE (0.9369 and 0.9486). The reason was that subtypes B and D contained more hard-to-predict dual tropic sequences (Table 1). This also resulted in a low AUC (0.5887) for subtype 02_AG, and a higher AUC (0.9029) for subtype A than for subtype D (0.8839) generated by HMMpred. In contrast, the performance of XGBpred was not so deeply influenced by dual tropic sequences. XGBpred had higher AUCs for the top four most common subtypes (subtypes B, C, D and 01_AE) than for subtypes A and

Dataset	Method	Specificity	Accuracy	MCC	AUC
Newdb	XGBpred	84.62%	90.19%	0.7310	0.9465
	HMMpred	70.59%	87.09%	0.6247	0.8774
G2p_str[23]	Geno2pheno[23]	61.6%	—	—	0.860
	G2p_str[23]	68.6%	—	—	0.892
	XGBpred	72.56%	89.90%	0.6605	0.8952
	HMMpred	72.09%	89.81%	0.6570	0.9002
Hivcopred[24]	Hivcopred[24]	81.44%	87.07%	0.67	0.904
	XGBpred	87.13%	88.52%	0.7154	0.9483
	HMMpred	71.08%	84.63%	0.5899	0.8829
CM[22]	CM[22]	92.92%	95.21%	0.885	0.97
	XGBpred	93.85%	95.33%	0.8106	0.9809
	HMMpred	89.54%	94.81%	0.7826	0.9635
WebPSSM[21]	WebPSSM[21]	83.3%	—	—	0.881
	XGBpred	83.33%	83.10%	0.6419	0.9043
	HMMpred	75.00%	80.28%	0.5693	0.8678

Table 3. Performance of the XGBpred and HMMpred methods on the different datasets. Performance of XGBpred and HMMpred on the Newdb, G2p_str, Hivcopred, CM and WebPSSM datasets at the sensitivities of 91.78%, 93.73%, 89.99%, 95.54% and 82.98%, respectively.

02_AG. In addition, The V3 loops of subtypes 01_AE and 02_AG come from subtypes E and A, respectively[37,38]. This can also further lead to the weaker predictive power for subtypes A and 02_AG as it is a trickier task to determine tropism for subtype A than the other subtypes[39,40]. Besides, the large biases existed between the number of R5 and X4-using samples for subtypes C and A (Table 1). Therefore, we also reported the mean average precision (mAP) of the two classes (Fig. 1C). AP is the areas under the precision-recall curve for a certain class. The bigger the mAP is, the better the method preforms. The mAPs and AUCs demonstrated the same tendency for the predictive power of our methods. Among all subtypes, just as AUCs, XGBpred and HMMpred showed the highest mAPs (0.9646, 0.9491) for subtype 01_AE. Moreover, for both XGBpred and HMMpred, the divergences between AUCs and mAPs for subtypes C and A were biggest. This may arise from the large biases between the amount of R5 and X4-using samples.

Comparison with other methods.

In this section, to evaluate our methods, we compared with the previously published methods Geno2pheno, G2p_str[23], Hivcopred[24], CM[22], and WebPSSM[21] by implementing our methods in a 10-fold cross validation test on the datasets used in these published methods, respectively. The exception was WebPSSM[21] where we used the training set from WebPSSM to model our methods in a 10-fold cross validation test and used the validation set from WebPSSM to test (Table 3).

First when comparing with the Geno2pheno and G2p_str methods[23], XGBpred and HMMpred achieved AUCs of 0.8952 and 0.9002, respectively. Our methods had higher AUCs than Geno2pheno (0.860) and G2p_str (0.892). In addition, XGBpred and HMMpred achieved specificities of 72.56% and 72.09% at the sensitivity of 93.73%. The specificities were obviously higher than the specificities of Geno2pheno (61.6%) and G2p_str (68.6%) at the same sensitivity. Second, when comparing with the Hivcopred method[24], XGBpred had a higher AUC (0.9483) than Hivcopred (0.904), but HMMpred had a low AUC (0.8829) as on the Newdb dataset. Third, when comparing with the CM method[22]. Our methods were as accurate as the CM method on the CM dataset which only contains a small amount of hard-to-predict dual tropic samples (Table 2). Finally, when comparing with the WebPSSM method[21], although the WebPSSM dataset is small, XGBpred had a higher AUC (0.9043) than WebPSSM (0.881), and HMMpred presented a similar AUC (0.8678) with WebPSSM.

Feature importance analysis.

Given the high performance of XGBpred presented in the previous subsections, we discussed which features XGBpred provided with its predictive power (Fig. 2). We did not analyze the feature importance on the WebPSSM dataset as it contains few training samples (Table 2). In the XGBpred method, the feature alignment score in the 5th position of the V3 loop appeared in the top three most important features on all datasets. Interestingly, amino acid Tyr in position 5 appeared more frequently in X4-using tropic than in R5 tropic sequences (Supplementary Fig. S1). Currently, X4-using tropism can be predicted by the 11/25 rule[19]. However, since position 5 was as same important as positions 11 and 25, the pragmatic 11/25/5 rule was proposed to predict a virus as X4-using tropic by the presence of a positively charged amino acid in positions 11 or 25, or by the presence of amino acid Tyr in position 5 of its V3 loop. Compared with the 11/25 rule, the 11/25/5 rule reduced sensitivities by 1.29%, 1.03%, 1.14% and 1.19% on the Newdb, G2p_str, Hivcopred and CM datasets while increasing specificities by 7.39%, 5.11%, 6.34% and 10.77%, respectively. The 11/25/5 rule also had higher accuracies and MCCs than the 11/25 rule on the four datasets, which indicates the influence of amino acid Tyr in position 5 with regard to viral tropism (Supplementary Table S3). In addition to positions 5, 11 and 25, positions 13, 18, 22 and 24 also ranked in the top ten most important features on the four datasets. Two exceptions were position 18 ranked 21st on the G2p_str dataset, and position 22 ranked 14th on the CM dataset. Indeed, all the positions that we identified as correlated with HIV-1 tropism are exactly in accordance with the results from

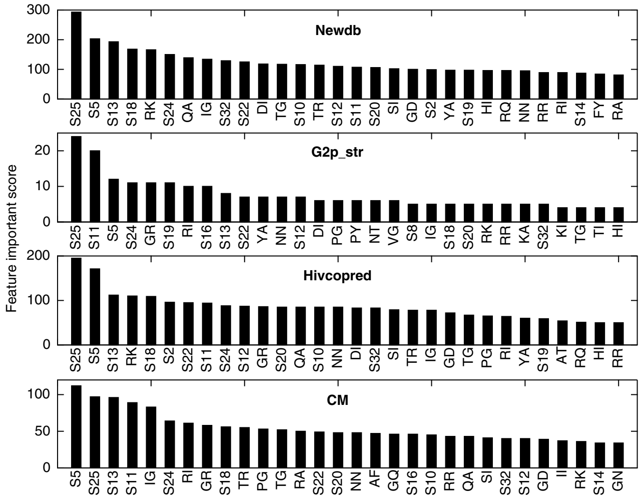

Figure 2. Distribution of feature importance scores. The top 30 most important features indicated by XGBpred on the Newdb, G2p_str, Hivcopred and CM datasets. S# means the alignment score in position #, and R1R2 represents a dipeptide.

Sander *et al.*[41] who point to the residues 298 (3), 302 (7), 306 (11), 308 (13), 315 (18), 317 (20), 319 (22), 321 (24), 322 (25) and 328 (32) are important for tropism. Furthermore, the feature importance distribution generated by XGBpred is a feasible method to judge whether a new discovered association pattern is of importance to co-receptor usage or not.

Discussion

In this study, we present two methods, XGBpred and HMMpred, for HIV-1 co-receptor usage prediction. XGBpred is based on machine learning, and HMMpred is based on statistics. XGBpred performed best on the Hivcopred and Newdb datasets containing larger proportions of hard-to-predict dual tropic samples in the X4-using samples, while HMMpred performed worst. In contrast, the predictive powers of the two methods were similar on the smaller G2p_str and CM datasets containing fewer dual tropic samples (Table 3). The poor ability of HMMpred to predict tropism stemmed from the high probability that HMMpred incorrectly predicted dual tropic samples as R5 tropic (Fig. 1B and Supplementary Fig. S2). The profiles used in HMMpred may not be meticulous enough. Several reasons may account for this phenomenon. Firstly, the two sequence families are highly similar since even one amino acid substitution may change their tropisms[42,43]. Secondly, the characteristics of dual tropic sequences may be overwhelmed by R5 and X4 tropic sequences. Finally, the unavailability of X4-using tropic samples makes it uncertain to learn its accurate HMM profile. Moreover, as the number of samples increased, the gap of predictive powers between XGBpred and HMMpred became large (Tables 2 and 3). This corresponds to the fact that the machine learning based Geno2pheno method is more widely used than the statistics based 11/25 rule and WebPSSM. As a result, a machine learning based method, in particular XGBpred, is recommended to predict co-receptor usage as the number of samples continues to expand.

In an effort to further increase the predictive power, we also generated three meta methods by the means of stacking[44]. The scores generated by XGBpred, Hivcopred (SVM^{light}) and HMMpred were added as additional features to the new stacking based XGBpred models. Compared with the original XGBpred method, the new stacking-based XGBpred methods had slightly higher AUCs on the G2p_str dataset but lower AUCs on the other datasets (Supplementary Table S4). The poor performances of the meta methods may due to the poorer predictive abilities of Hivcopred and HMMpred than the original XGBpred method, and/or the dependence of the results generated by XGBpred, Hivcopred and HMMpred (Supplementary Table S5). This may stem from the fact that the V3 loop is not the sole determinant of viral tropism. Moreover, V1, V2, C4 and the bridge sheet regions of gp120 also have an impact on co-receptor usage[45,46]. To predict tropism, several methods gain a higher accuracy by employing other information in addition to the V3 loop, such as clinical information[47], V2 loop sequences[48] and structure information[23,25,41]. Therefore, the stacking based method can be constructed to improve its predictive power by combining methods with different kinds of information.

In summary, the two methods we developed performed comparably on the datasets containing less hard-to-predict dual tropic sequences, but XGBpred performed much better on the datasets with more dual tropic sequences. This means XGBpred is more robust to predict dual tropic sequences than other methods. Thus, we strongly recommend to use XGBpred to predict viral tropism. Our two methods have been implemented as a freely available webserver under http://spg.med.tsinghua.edu.cn:23334/XGBpred/.

References

1. Hladik, F. *et al.* Initial events in establishing vaginal entry and infection by human immunodeficiency virus type-1. *Immunity* **26**, 257–270, https://doi.org/10.1016/j.immuni.2007.01.007 (2007).
2. Wilen, C. B., Tilton, J. C. & Doms, R. W. HIV: cell binding and entry. *Cold Spring Harb Perspect Med* **2**, https://doi.org/10.1101/cshperspect.a006866 (2012).

3. Koning, F., van Rij, R. & Schuitemaker, H. *Biological and Molecular Aspects of HIV1 Coreceptor Usage*. (2019).
4. Lengauer, T., Sander, O., Sierra, S., Thielen, A. & Kaiser, R. Bioinformatics prediction of HIV coreceptor usage. *Nature biotechnology* **25**, 1407–1410, https://doi.org/10.1038/nbt1371 (2007).
5. Vandekerckhove, L. P. R. *et al*. European guidelines on the clinical management of HIV-1 tropism testing. *The Lancet Infectious Diseases* **11**, 394–407, https://doi.org/10.1016/s1473-3099(10)70319-4 (2011).
6. Berger, E. A. *et al*. A new classification for HIV-1. *Nature* **391**, 240, https://doi.org/10.1038/34571 (1998).
7. Hoffmann, C. The epidemiology of HIV coreceptor tropism. *European journal of medical research* **12**, 385–390 (2007).
8. Panos, G. & Watson, D. C. Effect of HIV-1 subtype and tropism on treatment with chemokine coreceptor entry inhibitors; overview of viral entry inhibition. *Crit Rev Microbiol* **41**, 473–487, https://doi.org/10.3109/1040841X.2013.867829 (2015).
9. Huang, Y. *et al*. The role of a mutant CCR5 allele in HIV-1 transmission and disease progression. *Nature medicine* **2**, 1240–1243 (1996).
10. Gutiérrez, F., Carlos Rodríguez, J., García, F., Poveda, E. & Tropismo del, V. I. H. Técnicas disponibles y utilidad. *Enfermedades Infecciosas y Microbiología Clínica* **29**, 45–50, https://doi.org/10.1016/S0213-005X(11)70043-X (2011).
11. Naif, H. M. Pathogenesis of HIV Infection. *Infectious disease reports* **5**, e6, https://doi.org/10.4081/idr.2013.s1.e6 (2013).
12. Berger, E. A., Murphy, P. M. & Farber, J. M. Chemokine receptors as HIV-1 coreceptors: roles in viral entry, tropism, and disease. *Annual review of immunology* **17**, 657–700, https://doi.org/10.1146/annurev.immunol.17.1.657 (1999).
13. Tsamis, F. *et al*. Analysis of the mechanism by which the small-molecule CCR5 antagonists SCH-351125 and SCH-350581 inhibit human immunodeficiency virus type 1 entry. *Journal of virology* **77**, 5201–5208 (2003).
14. Nelson, M. & Panos, G. *Resistance to Chemokine (C-C Motif) Receptor 5 Antagonists HIV and AIDS CCR-5 Virus*. (2007).
15. Westby, M. *et al*. Emergence of CXCR4-using human immunodeficiency virus type 1 (HIV-1) variants in a minority of HIV-1-infected patients following treatment with the CCR5 antagonist maraviroc is from a pretreatment CXCR4-using virus reservoir. *Journal of virology* **80**, 4909–4920, https://doi.org/10.1128/jvi.80.10.4909-4920.2006 (2006).
16. Su, Z. *et al*. Response to vicriviroc in treatment-experienced subjects, as determined by an enhanced-sensitivity coreceptor tropism assay: reanalysis of AIDS clinical trials group A5211. *The Journal of infectious diseases* **200**, 1724–1728, https://doi.org/10.1086/648090 (2009).
17. Obermeier, M., Symons, J. & Wensing, A. M. HIV population genotypic tropism testing and its clinical significance. *Curr Opin HIV AIDS* **7**, 470–477, https://doi.org/10.1097/COH.0b013e328356eaa7 (2012).
18. Huang, W. *et al*. Coreceptor tropism can be influenced by amino acid substitutions in the gp41 transmembrane subunit of human immunodeficiency virus type 1 envelope protein. *Journal of virology* **82**, 5584–5593, https://doi.org/10.1128/jvi.02676-07 (2008).
19. Fouchier, R. A. *et al*. Phenotype-associated sequence variation in the third variable domain of the human immunodeficiency virus type 1 gp120 molecule. *Journal of virology* **66**, 3183–3187 (1992).
20. Jensen, M. A. *et al*. 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. *Journal of virology* **77**, 13376–13388, https://doi.org/10.1128/jvi.77.24.13376-13388.2003 (2003).
21. Jensen, M. A., Coetzer, M., van 't Wout, A. B., Morris, L. & Mullins, J. I. A reliable phenotype predictor for human immunodeficiency virus type 1 subtype C based on envelope V3 sequences. *Journal of virology* **80**, 4698–4704, https://doi.org/10.1128/JVI.80.10.4698-4704.2006 (2006).
22. Shen, H. S. *et al*. HIV coreceptor tropism determination and mutational pattern identification. *Sci Rep* **6**, 21280, https://doi.org/10.1038/srep21280 (2016).
23. Bozek, K., Lengauer, T., Sierra, S., Kaiser, R. & Domingues, F. S. Analysis of physicochemical and structural properties determining HIV-1 coreceptor usage. *PLoS Comput Biol* **9**, e1002977, https://doi.org/10.1371/journal.pcbi.1002977 (2013).
24. Kumar, R. & Raghava, G. P. Hybrid approach for predicting coreceptor used by HIV-1 from its V3 loop amino acid sequence. *PLoS One* **8**, e61437, https://doi.org/10.1371/journal.pone.0061437 (2013).
25. Heider, D., Dybowski, J. N., Wilms, C. & Hoffmann, D. A simple structure-based model for the prediction of HIV-1 co-receptor tropism. *BioData mining* **7**, 14, https://doi.org/10.1186/1756-0381-7-14 (2014).
26. Agwu, A. L. *et al*. Phenotypic Coreceptor Tropism in Perinatally HIV-infected Youth Failing Antiretroviral Therapy. *The Pediatric Infectious Disease Journal* **35**, 777–781, https://doi.org/10.1097/inf.0000000000001158 (2016).
27. Chen, T. & Guestrin, C. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (ACM, San Francisco, California, USA, 2016).
28. Bhasin, M. & Raghava, G. P. Classification of nuclear receptors based on amino acid composition and dipeptide composition. *J Biol Chem* **279**, 23262–23266, https://doi.org/10.1074/jbc.M401932200 (2004).
29. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology* **157**, 105–132 (1982).
30. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America* **89**, 10915–10919 (1992).
31. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology* **48**, 443–453 (1970).
32. Eddy, S. R. Profile hidden Markov models. *Bioinformatics (Oxford, England)* **14**, 755–763 (1998).
33. Sievers, F. *et al*. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular systems biology* **7**, 539, https://doi.org/10.1038/msb.2011.75 (2011).
34. David Forney, G. Jr. The Viterbi Algorithm: A Personal History. (2005).
35. Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et biophysica acta* **405**, 442–451 (1975).
36. Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29–36, https://doi.org/10.1148/radiology.143.1.7063747 (1982).
37. Gao, F. *et al*. The heterosexual human immunodeficiency virus type 1 epidemic in Thailand is caused by an intersubtype (A/E) recombinant of African origin. *Journal of virology* **70**, 7013–7029 (1996).
38. Carr, J. K. *et al*. Full genome sequences of human immunodeficiency virus type 1 subtypes G and A/G intersubtype recombinants. *Virology* **247**, 22–31, https://doi.org/10.1006/viro.1998.9211 (1998).
39. Riemenschneider, M. *et al*. Genotypic Prediction of Co-receptor Tropism of HIV-1 Subtypes A and C. *Sci Rep* **6**, 24883, https://doi.org/10.1038/srep24883 (2016).
40. Lochel, H. F., Riemenschneider, M., Frishman, D. & Heider, D. SCOTCH: subtype A coreceptor tropism classification in HIV-1. *Bioinformatics (Oxford, England)* **34**, 2575–2580, https://doi.org/10.1093/bioinformatics/bty170 (2018).
41. Sander, O. *et al*. Structural descriptors of gp120 V3 loop for the prediction of HIV-1 coreceptor usage. *PLoS Comput Biol* **3**, e58, https://doi.org/10.1371/journal.pcbi.0030058 (2007).
42. Shimizu, N. *et al*. Changes in and discrepancies between cell tropisms and coreceptor uses of human immunodeficiency virus type 1 induced by single point mutations at the V3 tip of the env protein. *Virology* **259**, 324–333, https://doi.org/10.1006/viro.1999.9764 (1999).
43. De Jong, J. J., De Ronde, A., Keulen, W., Tersmette, M. & Goudsmit, J. Minimal requirements for the human immunodeficiency virus type 1 V3 domain to support the syncytium-inducing phenotype: analysis by single amino acid substitution. *Journal of virology* **66**, 6777–6780 (1992).

44. Wolpert, D. H. Stacked generalization. *Neural Networks* **5**, 241–259, https://doi.org/10.1016/S0893-6080(05)80023-1 (1992).
45. Monno, L. *et al*. Impact of mutations outside the V3 region on coreceptor tropism phenotypically assessed in patients infected with HIV-1 subtype B. *Antimicrobial agents and chemotherapy* **55**, 5078–5084, https://doi.org/10.1128/aac.00743-11 (2011).
46. Dimonte, S. *et al*. Selected amino acid mutations in HIV-1 B subtype gp41 are associated with specific gp120v(3) signatures in the regulation of co-receptor usage. *Retrovirology* **8**, 33, https://doi.org/10.1186/1742-4690-8-33 (2011).
47. Brumme, Z. L. *et al*. Molecular and clinical epidemiology of CXCR4-using HIV-1 in a large population of antiretroviral-naive individuals. *The Journal of infectious diseases* **192**, 466–474, https://doi.org/10.1086/431519 (2005).
48. Thielen, A. *et al*. Improved prediction of HIV-1 coreceptor usage with sequence information from the second hypervariable loop of gp120. *The Journal of infectious diseases* **202**, 1435–1443, https://doi.org/10.1086/656600 (2010).

Acknowledgements

This work was partly supported by Grant 31770841 from the National Natural Science Foundation of China, and 2016YFA0502114 from the Ministry of Science and Technology of China. The funding source had no role in the design or conduct of the study, or in the collection, analysis, or interpretation of the data.

Author Contributions

X.C. designed the study, developed the method, implemented the data analysis and wrote the manuscript. Z.-X.W and X.-M.P. participated in study design and revision of the manuscript. All the authors read and approved the final manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-46420-4.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.