The Journal of Microbiology, October 2007, p. 441-446
Vol. 45, No. 5

Improved Prediction of Coreceptor Usage and Phenotype of HIV-1 Based on Combined Features of V3 Loop Sequence Using Random Forest

Shungao Xu, Xinxiang Huang, Huaxi Xu, and Chiyu Zhang*

Department of Biochemistry and Molecular Biology, Jiangsu University School of Medical Technology, Zhenjiang, Jiangsu 212013, P. R. China

(Received August 30, 2007 / Accepted October 17, 2007)

HIV-1 coreceptor usage and phenotype mainly determined by V3 loop are associated with the disease progression of AIDS. Predicting HIV-1 coreceptor usage and phenotype facilitates the monitoring of R5-to-X4 switch and treatment decision-making. In this study, we employed random forest to predict HIV-1 biological phenotype, based on 37 random features of V3 loop. In comparison with PSSM method, our RF predictor obtained higher prediction accuracy (95.1% for coreceptor usage and 92.1% for phenotype), especially for non-B non-C HIV-1 subtypes (96.6% for coreceptor usage and 95.3% for phenotype). The net charge, polarity of V3 loop and five V3 sites are seven most important features for predicting HIV-1 coreceptor usage or phenotype. Among these features, V3 polarity and four V3 sites (22, 12, 18 and 13) are first reported to have high contribution to HIV-1 biological phenotype prediction.

Keywords: HIV-1, V3 loop, random forest, cereceptor, phenotype, net charge

Human Immunodeficiency Virus Type 1 (HIV-1) infection requires two cellular receptors, CD4 and a chemokine receptor (known as coreceptor). *In vivo*, the most important coreceptors are CCR5 and CXCR4. According to the differential use of two major coreceptors, HIV-1 was classified into three biological variants, R5, R5X4, and X4 (Berger *et al.*, 1998). The R5 strains use CCR5, X4 strains use CXCR4, and R5X4 strains use both coreceptors. Furthermore, according to their replication rate and ability to induce syncytia in MT2 cells, HIV-1 is classified phenotyically into syncytium-inducing (SI) and non-syncytium-inducing (NSI) (Tersmette *et al.*, 1988). HIV-1 strains of NSI and R5 appear to have identical biological properties and are generally associated with slow replication rate and low virulence. SI and X4/R5X4 viruses also show similar phenotype and have rapid replication rate and high virulence (Bjorndal *et al.*, 1997). The disease progression of AIDS is generally associated with a switch in coreceptor usage from CCR5 (R5) to CXCR4 (X4) (Richman and Bozzette, 1994; Connor *et al.*, 1997). Therefore, predicting the emergence of X4 has potential value for understanding pathogenesis, monitoring disease progression and making treatment decisions.

The V3 loop of HIV-1 gp120, a disulfide-linked loop of approximately 35 amino acids, makes direct contact with the coreceptor and plays a dominant role in determinant of vial coreceptor usage and phenotype (Hwang *et al.*, 1991; Shioda *et al.*, 1991). Several amino acid substitutions in V3 loop (e.g. at positions 11 and 25) and total net charge of V3 loop frequently determine viral corecptor usage and phenotype (De Jong *et al.*, 1992; Fouchier *et al.*, 1992; Xiao *et al.*, 1998). Positively charged residues at positions 11 and 25 and an increasing net charge of the V3 are strongly associated with X4 strains (Brelot *et al.*, 1999). Therefore, bioinformatics approaches based on V3 sequences have been developed for predicting HIV-1 biological phenotype (Jensen and Van't Wout, 2003). Based on several special residues (especially amino acid profiles at V3 sites 11 and 25) or net charge of V3 loop, these approaches employed a multiple linear regression (Briggs *et al.*, 2000), a neural network strategy (Resch *et al.*, 2001), a machine-learning method (Pillai *et al.*, 2003), or a position-specific scoring matrix (PSSM) (Jensen *et al.*, 2003; Jensen *et al.*, 2006), to predict HIV-1 coreceptor usage and phenotype. PSSM approach appears to have better predictive power over other methods (Jensen and Van't Wout, 2003). Apart from the machine-learning method, however, they are limited to predict two major subtypes B and C.

Random forest (RF), a robust classifier developed by Breiman (2001), was demonstrated to have better performance over other machine learning approaches (Svetnik *et al.*, 2003). In this study, we employed RF to predict the coreceptor usage and phenotype of HIV-1, based on 37 random features, including 35 amino acid profiles, total net charge and polarity of V3 loop. In comparison with PSSM, our method obtained better performance especially for non-B non-C subtypes. Different from previous observation, the total net charge and polarity of V3 loop are most important contributors together with residues at five V3 sites 22, 25, 11, 12 and 13 to R5/X4 prediction and together with residues at another five V3 sites 11, 18, 13, 24 and 22 to NSI/SI prediction. Furthermore, four sites 22, 12, 18 and 13 are first reported to have high contribution to HIV-1 biological phenotype prediction.

* To whom correspondence should be addressed.
(Tel) 86-511-8503-8449; (Fax) 86-511-8503-8449
(E-mail) zhangcy1999@hotmail.com

Materials and Methods

Data sets

All available HIV-1 V3 sequences with known coreceptor usage or phenotype were retrieved from the Los Alamos HIV Sequence Database (http://www.hiv.lanl.gov/content/hiv-db/mainpage.html) in June 2007. To generate the R5/X4 data set, a total of 2684 V3 sequences associated with known coreceptor usage, including all available HIV-1 subtypes and recombinant forms, were downloaded from this Sequence Database. Of them, 1516 peptide sequences (including 1132 R5 and 384 X4 sequences) were randomly selected as training data set for training our RF prediction model. After deletion of all duplicate sequences from remainder 1168 V3 sequences, 651 unique sequences were used to test the validity of our RF predictor for coreceptor usage. For NSI/SI prediction, the data set including 1901 V3 sequences was used. A subset of 1073 sequences (including 735 NSI and 338 SI sequences) was randomly selected as training data set for training RF prediction model. Of remainder 828 V3 sequences, 432 unique sequences were used to test the validity of our RF method for phenotype prediction.

To compare with PSSM method, the same testing data sets were used. The PSSM scores were calculated using the online PSSM tool (http://ubik.microbiol.washington.edu/computing/pssm/) (Jensen *et al.*, 2003; Jensen *et al.*, 2006).

Random forest and random features

Random forest (RF) is a robust classifier consisting of an ensemble of unpruned classification or regression trees (Breiman, 2001). RF classifier uses bagging and random feature selection in tree induction. In bagging, each tree is trained on a bootstrap sample of the training data. To obtain a low-bias tree (unpruned tree), RF randomly selects a subset of features to split at each node. Then, prediction is made by aggregating (majority vote or averaging) the predictions of the ensemble. To assess the prediction performance, RF performs a type of cross-validation in parallel with the training step by using the so-called out-of-bag (OOB) samples.

The V3 loop contains approximately 35 amino acids (with a scope of 33-37 residues). In order to obtain consistent V3 alignment in length, each V3 peptide sequence was aligned with typical 35-residue V3 sequences using CLUSTAL W 1.83 (Thompson *et al.*, 1997), and residues representing insertions with respect to the typical sequences were discarded and gaps were retained. The amino acid profiles or gaps at 35 retained V3 sites were used as random features of RF predictor together with the net charge of V3 loop. In addition, we also introduced a new feature, the polarity of V3 loop into the RF model for improving prediction power. The net charge and polarity were calculated according to intact V3 loop using AAindex database(Kawashima *et al.*, 1999). Then, the importance of 37 features (including net charge, polarity, and 35 amino acid profiles) of V3 loop in prediction models of HIV-1 biological phenotypes was evaluated using RF.

Evaluation of the predictive performance

The final performance of RF method was determined by measuring the sensitivity (SE), specificity (SP), total prediction accuracy (ACC) and Matthew's Correlation coefficient (MCC). The SE, SP, ACC and MCC parameters were calculated using Eqs. (1), (2), (3) and (4), respectively. In addition, by Bayes' theorem, the positive predictive value (PPV), which is the probability that a sequence predicted to be X4(R5) is in fact X4(R5), was also calculated using Eq (5).

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Sp = \frac{TN}{TN + FP} \quad (2)$$

$$Se = \frac{TP}{TP + FN} \quad (3)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TN + FN) \times (TP + FN) \times (TN + FP)}} \quad (4)$$

$$PPV = \frac{P \times Se}{P \times Se + (1 - P) \times (1 - Sp)} \quad (5)$$

Where TP is true positives (e.g. X4 predicted as X4); FN is false negatives (e.g. X4 predicted as R5); TN is true negatives (e.g. non-X4 predicted as R5); FP is false positives (e.g. R5 predicted as X4) and P is the pre-test probability of X4 (R5).

The software package for HIV-1 coreceptor usage and phenotype prediction

Our RF predictor (named as R5/X4-pred) and the testing data set are freely available at http://yjxy.ujs.edu.cn/R5-X4 pred.rar and http://yjxy.ujs.edu.cn/Testing data set.rar, respectively.

Results and Discussion

The RF prediction performance and the comparison with PSSM method

Several bioinformatic approaches have been previously developed to predict HIV-1 receptor usage and phenotype, and PSSM method appeared to have the best performance (Jensen and Van't Wout, 2003). PSSM method is based on position-specific scoring matrices and is best applicable to HIV-1 subtypes B and C (Jensen *et al.*, 2003; Jensen *et al.*, 2006). So, we divided our testing set into three subsets, including B-subtype set, C-subtype set and non-B non-C data set, and compared our RF predictor with PSSM using the three subsets. When HIV-1 subtype-B and -C data sets were used, RF method for coreceptor usage was able to predict with 94% total prediction accuracy (ACC) and 0.83 MCC value for subtype B viruses and with 96.7% ACC and 0.86 MCC value for subtype C viruses (Table 1), both of which are better than that of PSSM method. In particular, when non-B non-C subtype data set was used, RF predictor appeared to have the best performance with 96.6% ACC and 0.93 MCC value, also significantly better than PSSM (Table 1), suggesting its wide applicability for different HIV-1 subtypes. For HIV-1 phenotype prediction, our RF method also appears more excellent performance than PSSM especially for non-B non-C HIV-1 subtypes (95.3% ACC and 0.87 MCC value) (Table 2). The excellent prediction accuracy of RF method indicates that extracting more information of

Table 1. Comparison of our RF method with PSSM method for HIV-1 coreceptor usage prediction

Subtype	Method	ACC (%)	Sp (%)	Se (%)	MCC	PVV (%)
B	RF	94	98.3	80	0.83	98.2
	PSSM	89.4	98.3	60	0.69	95.4
C	RF	96.7	100	76.5	0.86	100
	PSSM	92.5	95.1	76.5	0.7	90.5
Non-B, Non-C	RF	96.6	96.7	96.4	0.93	99.7
	PSSM	86.3	86.8	85.5	0.71	94.4
All	RF	95.1	98.4	85.2	0.87	98.9
	PSSM	89.2	95.5	70.4	0.7	92.9

Only the X4 prediction results were shown.

Table 2. Comparison of our RF method with PSSM method for HIV-1 phenotype prediction

Subtype	Method	ACC (%)	Sp (%)	Se (%)	MCC	PVV (%)
B	RF	90.6	90.4	90.9	0.8	97.4
	PSSM	89.6	91.1	87	0.78	96.5
C	RF	92	98.9	60	0.7	96.1
	PSSM	90.3	92.5	80	0.69	90.4
Non-B, Non-C	RF	95.3	96.3	92.3	0.87	98.7
	PSSM	90.7	91.4	88.5	0.76	95.6
All	RF	92.1	94.5	86.2	0.81	97.0
	PSSM	90	91.6	86.2	0.76	95.5

Only the SI prediction results were shown.

V3 sequences, rather than only dependent on one or two features (e.g. net charge or positively-charged residue at V3 sites 11 and 25), obviously enhances the prediction power of HIV-1 coreceptor usage and phenotype.

Estimating and ranking the feature importance

Decision tree is able to select important ones from many relevant and irrelevant features, and reveals the relationship between features and predictions by an explicit model. A measure of feature importance with respect to its contribution to the prediction performance of random forest can be provided by mean decrease accuracy, which is calculated in the course of training (Breiman, 2001; Svetnik *et al.*, 2003). In our RF method for predicting HIV-1 coreceptor usage and phenotype, the importance of each feature of V3 loop is shown in Table 3. The total net charge and polarity of V3 loop, as well as five V3 amino acid sites are the top seven features that mainly determine HIV-1 coreceptor usage or phenotype. As expected, total net charge is the most important contributor in our RF prediction method. The polarity of V3 loop firstly introduced by us into prediction method is the third most important feature influencing the prediction accuracy of HIV-1 coreceptor usage and phenotype. In addition, among these top features, three sites 11, 22 and 13 were observed in both RF prediction models, indicating their crucial roles in prediction of HIV-1 coreceptor usage and phenotype (Table 3). Surprisingly, however, the V3 sites 11 and 25, both of which were considered as very important factors in previous prediction approaches (De Jong *et al.*, 1992; Fouchier *et al.*, 1992; Xiao *et al.*, 1998; Briggs *et al.*, 2000; Resch *et al.*, 2001), are not ranked in top three features of RF predictor for HIV-1 receptor usage. Furthermore, in HIV-1 phenotype prediction model, V3 site 25 only ranks eighth in all 37 random features, also showing less importance for prediction (Table 3). These results provide a possible interpretation of low prediction accuracy of previous approaches based on the appearance of positively charged mutations at V3 sites 11 and 25 (Jensen and Van't Wout, 2003).

The characteristics of HIV-1 V3 loop sequences with different coreceptor usages and phenotypes

Because of the importance of net charge and polarity of V3 loop in viral phenotype and coreceptor usage, the characteristics of 2684 R5/X4 sequences and 1901 NSI/SI sequences were analyzed. The net charge distribution for both data sets is shown in Fig. 1A. The distributions of net charge appear to be consistent between R5 and NSI viruses and between X4 and SI viruses. Approximate 90% of R5 and NSI V3 sequences have low net charge scores of ≤ 4. Contrarily, more than 90% of X4 and SI V3 sequences have the scores of net charge more than 4. Furthermore, the X4 (5.17 ± 0.05) and SI (5.42 ± 0.06) viruses have significantly higher net charge scores than R5 (3.41 ± 0.02) and NSI (3.36 ± 0.02) viruses ($P<0.0001$) (Fig. 1A) (Xiao *et al.*, 1998; Jensen *et al.*, 2006). Increasing total net charge of V3 loop facilitates the recog-

nition of HIV-1 Gp120 to CXCR4 via an electrostatic interaction between V3 loop and the negatively charged CXCR4, which accelerates the coreceptor switch from R5 to X4 and phenotype switch from NSI to SI (Brelot *et al.*, 1999).

The polarity analysis of all V3 sequences shows that 88% of R5 and 92% of NSI sequences have the low polarity scores of <70. The proportions of high polarity scores (≥70) in X4 and SI V3 sequences increase to 47% and 63%, respectively (Fig. 1B). The V3 loops of X4 (69.97±0.54) and SI (75.77±0.60) viruses have significantly stronger polarity than R5 (60.12±0.21) and NSI (59.72±0.22) viruses ($P<0.0001$). Similar to net charge, enhancing polarity of V3 loop accounts for the emergence of X4 and SI viruses possible also by favoring the recognition of HIV-1 Gp120 to CXCR4 (Brelot *et al.*, 1999).

To display consensus sequence of each V3 data set, a graphical representation of these sequences using WebLogo (Crooks *et al.*, 2004) is shown in Fig. 2. Among variable V3 sites, sites 22, 25, 11, 12, and 13 are the top five sites for R5/X4 prediction model, and sites 11, 18, 13, 24, and 22 are the top five sites for NSI/SI prediction model (Table 3). Amino acid profiles at these sites are distinctly different between R5 and X4 types and between NSI and SI types, suggesting their importance for HIV-1 coreceptor usage and phenotype prediction (Fig. 2). In addition, it is worth pointing out that four sites 22, 12, 18 and 13 are first reported to have high contribution to HIV-1 biological phenotype pre-

Table 3. Estimating and ranking the relative importance of the features

Rank	R5/X4 Model		NSI/SI Model	
	Features	MDA (%)	Features	MDA (%)
1.	Net charge	8.41	Net charge	4.01
2.	V3 site: 22	5.85	V3 site: 11	3.73
3.	Polarity	4.03	Polarity	2.53
4.	V3 site:25	3.74	V3 site: 18	1.01
5.	V3 site:11	3.28	V3 site: 13	0.96
6.	V3 site:12	2.04	V3 site: 24	0.88
7.	V3 site:13	1.91	V3 site: 22	0.76
8.	V3 site:7	1.83	V3 site: 25	0.76
9.	V3 site:24	1.59	V3 site:7	0.59
10.	V3 site:5	1.57	V3 site:8	0.58
11.	V3 site:18	1.55	V3 site:19	0.55
12.	V3 site:10	1.39	V3 site:34	0.50
13.	V3 site:8	1.10	V3 site:32	0.48
14.	V3 site:19	1.05	V3 site:10	0.46
15.	V3 site:2	1.04	V3 site:23	0.45
16.	V3 site:32	1.03	V3 site:27	0.45
17.	V3 site:20	0.83	V3 site:14	0.44
18.	V3 site:14	0.82	V3 site:20	0.43
19.	V3 site:34	0.65	V3 site:16	0.41
20.	V3 site:29	0.57	V3 site:2	0.39
21.	V3 site:27	0.53	V3 site:12	0.30
22.	V3 site:16	0.42	V3 site:29	0.29
23.	V3 site:6	0.28	V3 site:5	0.24
24.	V3 site:23	0.24	V3 site:21	0.23
25.	V3 site:30	0.24	V3 site:30	0.15
26.	V3 site:26	0.18	V3 site:26	0.13
27.	V3 site:9	0.17	V3 site:9	0.08
28.	V3 site:21	0.16	V3 site:6	0.06
29.	V3 site:15	0.09	V3 site:31	0.03
30.	V3 site:33	0.02	V3 site:15	0.02
31.	V3 site:4	0.01	V3 site:33	0.02
32.	V3 site:1	0.01	V3 site:28	0.01
33.	V3 site:31	0.01	V3 site:17	0.00
34.	V3 site:28	0.00	V3 site:35	0.00
35.	V3 site:3	0.00	V3 site:1	0.00
36.	V3 site:17	0.00	V3 site:3	0.00
37.	V3 site:35	0.00	V3 site:4	-0.01

MDA, Mean decrease accuracy. MDA measures the feature importance in terms of the contribution to prediction accuracy. To obtain MDA, the values of the m^{th} feature of each tree were rearranged for the out-of- bag set. Then puts this permuted set down the tree, and gets new classifications for the forest. The importance of the m^{th} feature, that is MDA, is defined as the difference of the out-of-bag error rate between randomly permuted m^{th} feature and original feature.

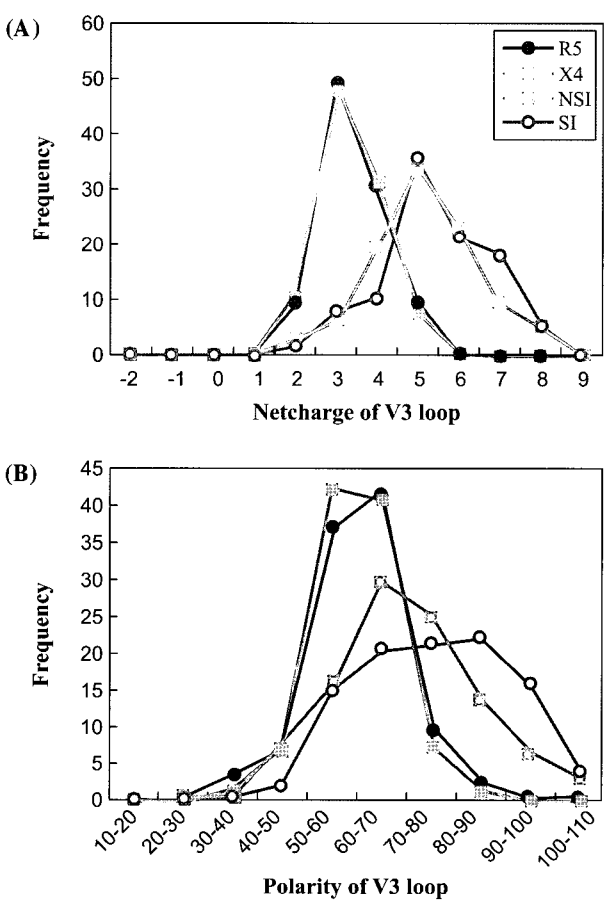

Fig. 1. Score distributions of net charge and polarity of all V3 sequences. (A) The distribution of total net charge. (B) Polarity distribution of V3 loop. The polarity of V3 loop is calculated by summing the polarity score of all amino acids in V3 loop.

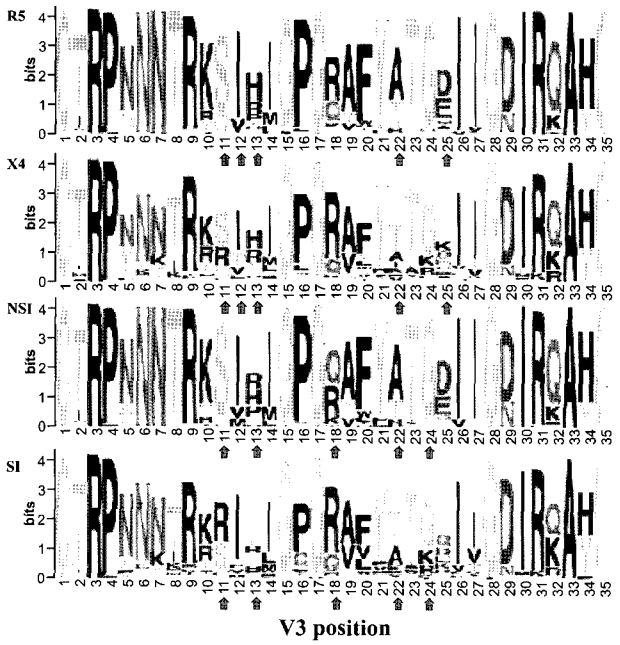

Fig. 2. Sequence logos of V3 sequences used in this study. The character and size of each logo represent the proportion of an amino acid at the specific site. The R5 and X4 data sets are represented by 2012 R5 and 672 X4 V3 sequences, respectively. The NSI and SI data sets correspond to 1306 NSI and 595 SI V3 sequences, respectively. The red arrows highlight the top five sites in V3 loop that determine HIV-1 coreceptor usage or phenotype.

diction (Fig. 2). Surprisingly, however, the well-recognized crucial site 25 ranks eighth, implying less contribution to NSI/SI prediction. Furthermore, other V3 sites, e.g. 19, 23 and 32, that were also demonstrated to play an important role in HIV-1 R5/X4 and NSI/SI prediction in previous studies (Milich *et al.*, 1997; Xiao *et al.*, 1998), are ranked in middle of the relative importance index of features in our RF prediction, suggesting less importance in HIV-1 biological phenotype prediction despite that they also have obvious amino acid variation between R5/NSI and X4/SI (Table 3, Fig. 2).

Conclusions

We devised an RF-based software (R5/X4-pred) for predicting HIV-1 coreceptor usage and phenotype. It has better performance over pervious methods especially for non-B non-C HIV-1 subtypes. The net charge, polarity of V3 loop and five V3 sites are seven most important features for predicting HIV-1 coreceptor usage or phenotype. Among these features, V3 polarity and four V3 sites (22, 12, 18 and 13) are first reported to have high contribution to HIV-1 biological phenotype prediction. Furthermore, the R5/X4-pred software is freely available at http://yjxy.ujs.edu.cn/R5-X4 pred.rar.

Acknowledgements

The authors would like to thank Miss Xiumei Sheng for her careful reading of our manuscript and many valuable suggestions. This work was supported by research grants from National Natural Science Foundation of China (No. 30600352) and Natural Science Foundation of Jiangsu Province, China (No. BK2006550), and the Startup Fund from Jiangsu University (No. 2281270002) to C.Z.

References

Berger, E.A., R.W. Doms, E.M. Fenyo, B.T. Korber, D.R. Littman, J.P. Moore, Q.J. Sattentau, H. Schuitemaker, J. Sodroski, and R.A. Weiss. 1998. A new classification for HIV-1. *Nature* 391, 240.

Bjorndal, A., H. Deng, M. Jansson, J.R. Fiore, C. Colognesi, A. Karlsson, J. Albert, G. Scarlatti, D.R. Littman, and E.M. Fenyo. 1997. Coreceptor usage of primary human immunodeficiency virus type 1 isolates varies according to biological phenotype. *J. Virol.* 71, 7478-7487.

Breiman, L. 2001. Random Forests. *Mach Learn.* 45, 5-32.

Brelot, A., N. Heveker, K. Adema, M.J. Hosie, B. Willett, and M. Alizon. 1999. Effect of mutations in the second extracellular loop of CXCR4 on its utilization by human and feline immunodeficiency viruses. *J. Virol.* 73, 2576-2586.

Briggs, D.R., D.L. Tuttle, J.W. Sleasman, and M.M. Goodenow. 2000. Envelope V3 amino acid sequence predicts HIV-1 phenotype (co-receptor usage and tropism for macrophages). *AIDS* 14, 2937-2939.

Connor, R.I., K.E. Sheridan, D. Ceradini, S. Choe, and N.R. Landau. 1997. Change in coreceptor use coreceptor use correlates with disease progression in HIV-1--infected individuals. *J. Exp. Med.* 185, 621-628.

Crooks, G.E., G. Hon, J.M. Chandonia, and S.E. Brenner. 2004. WebLogo: a sequence logo generator. *Genome Res.* 14, 1188-1190.

De Jong, J.J., A. De Ronde, W. Keulen, M. Tersmette, and J. Goudsmit. 1992. Minimal requirements for the human immunodeficiency virus type 1 V3 domain to support the syncytium-inducing phenotype: analysis by single amino acid substitution. *J. Virol.* 66, 6777-6780.

Fouchier, R.A., M. Groenink, N.A. Kootstra, M. Tersmette, H.G. Huisman, F. Miedema, and H. Schuitemaker. 1992. Phenotype-associated sequence variation in the third variable domain of the human immunodeficiency virus type 1 gp120 molecule. *J. Virol.* 66, 3183-3187.

Hwang, S.S., T.J. Boyle, H.K. Lyerly, and B.R. Cullen. 1991. Identification of the envelope V3 loop as the primary determinant of cell tropism in HIV-1. *Science* 253, 71-74.

Jensen, M.A., M. Coetzer, A.B. Van't Wout, L. Morris, and J.I. Mullins. 2006. A reliable phenotype predictor for human immunodeficiency virus type 1 subtype C based on envelope V3 sequences. *J. Virol.* 80, 4698-4704.

Jensen, M.A., F.S. Li, A.B. Van't Wout, D.C. Nickle, D. Shriner, H.X. He, S. McLaughlin, R. Shankarappa, J.B. Margolick, and J.I. Mullins. 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.

Jensen, M.A. and A.B. Van't Wout. 2003. Predicting HIV-1 coreceptor usage with sequence analysis. *AIDS Rev.* 5, 104-112.

Kawashima, S., H. Ogata, and M. Kanehisa. 1999. AAindex: Amino Acid Index Database. *Nucleic Acids Res.* 27, 368-369.

Milich, L., B.H. Margolin, and R. Swanstrom. 1997. Patterns of amino acid variability in NSI-like and SI-like V3 sequences and a linked change in the CD4-binding domain of the HIV-1 Env protein. *Virology* 239, 108-118.

Pillai, S., B. Good, D. Richman, and J. Corbeil. 2003. A new perspective on V3 phenotype prediction. *AIDS Res. Hum. Retroviruses* 19, 145-149.

Resch, W., N. Hoffman, and R. Swanstrom. 2001. Improved success of phenotype prediction of the human immunodeficiency virus type 1 from envelope variable loop 3 sequence using neural networks. *Virology* 288, 51-62.

Richman, D.D. and S.A. Bozzette. 1994. The impact of the syncytium-inducing phenotype of human immunodeficiency virus on disease progression. *J. Infect Dis.* 169, 968-974.

Shioda, T., J.A. Levy, and C. Cheng-Mayer. 1991. Macrophage and T cell-line tropisms of HIV-1 are determined by specific regions of the envelope gp120 gene. *Nature* 349, 167-169.

Svetnik, V., A. Liaw, C. Tong, J.C. Culberson, R.P. Sheridan, and B.P. Feuston. 2003. Random forest: a classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* 43, 1947-1958.

Tersmette, M., R.E. De Goede, B.J. Al, I.N. Winkel, R.A. Gruters, H.T. Cuypers, H.G. Huisman, and F. Miedema. 1988. Differential syncytium-inducing capacity of human immunodeficiency virus isolates: frequent detection of syncytium-inducing isolates in patients with acquired immunodeficiency syndrome (AIDS) and AIDS-related complex. *J. Virol.* 62, 2026-2032.

Thompson, J.D., T.J. Gibson, F. Plewniak, F. Jeanmougin, and D.G. Higgins. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* 25, 4876-4882.

Xiao, L., S.M. Owen, I. Goldman, A.A. Lal, J.J. deJong, J. Goudsmit, and R.B. Lal. 1998. CCR5 coreceptor usage of non-syncytium-inducing primary HIV-1 is independent of phylogenetically distinct global HIV-1 isolates: delineation of consensus motif in the V3 domain that predicts CCR-5 usage. *Virology* 240, 83-92.