AIDS RESEARCH AND HUMAN RETROVIRUSES
Volume 29, Number 00, 2013

DOI: 10.1089/aid.2012.0173

A Predictive Model for HIV Type 1 Coreceptor Selectivity

Chris A. Kieslich,[1] David Shin,[1] Aliana López de Victoria,[1] Gloria González-Rivera,[2] and Dimitrios Morikis[1]

Abstract

Despite its sequence variability and structural flexibility, the V3 loop of the HIV-1 envelope glycoprotein gp120 is capable of recognizing cell-bound coreceptors CCR5 and CXCR4 and infecting cells. Viral selection of CCR5 is associated with the early stages of infection, and transition to selection of CXCR4 indicates disease progression. We have developed a predictive statistical model for coreceptor selectivity that uses the discrete property of net charge and the binary coreceptor preference markers of the $N^6X^7[T/S]^8X^9$ glycosylation motif and 11/24/25 positive amino acid rule. The model is based on analysis of 2,054 V3 loop sequences from patient data and allows us to infer the most likely state of the disease from physicochemical characteristics of the sequences. The performance of the model is comparable to established sequence-based predictive methods, and may be used in combination with other methods as a supportive diagnostic for coreceptor selection. This model may be used for personalized medical decisions in administering coreceptor-specific therapies.

Introduction

The V3 loop of the HIV-1 glycoprotein 120 (gp120) is implicated in HIV-1 entry into host cells by interacting with coreceptors CCR5 or CXCR4, while the remainder of gp120 is anchored to receptor CD4 and viral surface glycoprotein 41 (gp41).[1–5] Given the sequence variability[6] and structural flexibility[7–9] of the V3 loop, persistent sequence, structural, and physicochemical patterns have been sought to describe the mechanism of viral recognition and entry at the molecular level. Charge and electrostatic potential have been proposed to be dominant factors in the recognition between the positive V3 loop and the negative N-terminal domain of CCR5.[10,11] Long-range electrostatic potential interactions are nonspecific, but are capable of steering the V3 loop toward CCR5. Clustering analysis of electrostatic potentials of V3 loop consensus sequences has revealed persistent electrostatic potential characteristics, which are more pronounced in the subtypes of Group M, despite sequence variability.[11] A further complication for understanding the molecular role of the V3 loop in viral entry arises from the fact that HIV-1 changes coreceptor as the disease progresses, with preference for CCR5 at the initial stages of infection and preference for CXCR4 as the patient's health deteriorates.[2–5,10–21]

The absence of the $N^6X^7[T/S]^8X^9$ glycosylation sequence motif has been proposed to favor binding to CXCR4,[22,23] and the presence of one or more positive amino acids at sequence positions 11, 24, or 25 has been proposed to also favor binding to CXCR4 (the 11/24/25 positive amino acid rule).[24] Glycosylation is also related to charge because of the presence of sialic acids, which carry negative charge, and affect the overall charge of the V3 loop, as discussed.[11] The presence of the glycosylation motif (and the charge glycosylation carries) can contribute to the evolutionary pressure for charge adjustments at other sites of the V3 loop sequence.

In this study we have analyzed V3 loop sequences with known coreceptor preference from patient samples, available at the Los Alamos HIV Databases.[25] Our analysis utilizes physicochemical information included in the sequences, such as net charge, the $N^6X^7[T/S]^8X^9$ glycosylation sequence motif, and the 11/24/25 positive amino acid rule, to develop a predictive statistical model for HIV-1 coreceptor selectivity.

Materials and Methods

We first retrieved 5,309 V3 loop sequences deposited at the Los Alamos HIV Databases[25] at the beginning of the study (June 27, 2011). The deposited sequences are derived from patient data and are associated with known coreceptor selection from experimental studies.[5,16,17,25] The sequence sample was reduced to 2,054 by filtering duplicate sequences belonging to the same patient and keeping only unique sequences per patient. Sequence analysis was performed using the amino acids within and including the disulfide bridge located at the base of the V3 loop. The sequences were 33–37 amino acids in length, with those associated with CCR5 having a length of 34–35 and those associated with CXCR4 or CCR5/CXCR4 (meaning dual or mixed coreceptor) showing larger length variability. Net

[1]Department of Bioengineering, University of California, Riverside, California.
[2]Department of Economics, University of California, Riverside, California.

charge was determined by counting the unit charges of positively and negatively charged amino acids. Arginines and lysines have charge +1, whereas aspartic and glutamic acids have charge −1. Given the high conformational variability (owed to lack of specific structure and solvent exposure of the V3 loop[9]), we consider that histidines have a pK_a close to that of free amino acids in solution (in the range of 6–6.5), and therefore they are neutral at physiological pH (at the range of 7–7.5). The presence or absence of the glycosylation motif and the 11/24/25 rule were determined as binary variables.

We used an ordered probit statistical model for quantitative estimation of coreceptor selectivity. Our underlying assumptions are (1) disease progression follows the coreceptor selection pattern in the order of CCR5 → CCR5/CXCR4 → CXCR4, and (2) that coreceptor selectivity can be inferred by the information found in the sequence of the V3 loop. The probit model depicts the coreceptor transition order, and can be used to predict probabilities for coreceptor selection given the properties of glycosylation motif, positive amino rule, and net charge. The model accounts for a discrete net charge integer variable and binary variables of 1 and 0 for the presence and absence, respectively, of the glycosylation motif and the 11/24/25 positive amino acid rule. These variables are not independent of each other, and they are all related to charge, as mentioned above. Let us call y_i^* the coreceptor state embedded in experimentally derived V3 loop sequence i, which is a latent continuous variable. Then, we define an ordered variable y_i such that

$$y_i = \begin{cases} 1 & y_i^* < \mu_1 \\ 2 & \mu_1 \le y_i^* < \mu_2 \\ 3 & y_i^* \ge \mu_2 \end{cases} \quad (1)$$

where 1, 2, and 3 refer to progression in coreceptor state (CCR5, CCR5/CXCR4, and CXCR4, respectively), and μ_1 and μ_2 are unknown thresholds. We model the coreceptor selection as a function of a set of the following physicochemical characteristics of the V3 loop sequence: (1) the $N^6X^7[T/S]^8X^9$ glycosylation motif (denoted as Motif); (2) the 11/24/25 positive amino acid rule (denoted as Rule); and (3) net charge (denoted as Charge). For each individual sequence i

$$y_i^* = \beta_1 Motif_i + \beta_2 Rule_i + \beta_3 Charge_i + \varepsilon_i = \beta' x_i + \varepsilon_i \quad (2)$$

where ε_i is a normal error term, independent and identically distributed (mean zero and variance 1). Under these assumptions, we obtain the probabilities of being in coreceptor state 1, 2 or 3, as follows:

$$\begin{aligned} P(y_i = 1) &= P(y_i^* < \mu_1) = P(\varepsilon_i < \mu_1 - \beta' x_i) = \Phi(\mu_1 - \beta' x_i) \\ P(y_i = 2) &= P(\mu_1 \le y_i^* < \mu_2) = P(\varepsilon_i < \mu_2 - \beta' x_i) \\ &\quad - P(\varepsilon_i < \mu_1 - \beta' x_i) = \Phi(\mu_2 - \beta' x_i) - \Phi(\mu_1 - \beta' x_i) \\ P(y_i = 3) &= P(y_i^* \ge \mu_2) = P(\varepsilon_i \ge \mu_2 - \beta' x_i) = 1 - \Phi(\mu_2 - \beta' x_i) \end{aligned} \quad (3)$$

where Φ is the cumulative standard normal distribution function.

The aforementioned model considers three coreceptor states, namely CCR5, CCR5/CXCR4, and CXCR4. But it can be argued that only two coreceptors physically exist, and therefore we can define an ordered variable y_i such that

$$y_i = \begin{cases} 1 & y_i^* < \mu \\ 2 & y_i^* \ge \mu \end{cases} \quad (4)$$

The variable y_i^* is defined as in Eq. (2), and the probabilities for coreceptor state 1 or 2 (CCR5 or CXCR4, respectively) are given by

$$\begin{aligned} P(y_i = 1) &= P(y_i^* < \mu) = P(\varepsilon_i < \mu - \beta' x_i) = \Phi(\mu - \beta' x_i) \\ P(y_i = 2) &= P(y_i^* \ge \mu) = P(\varepsilon_i \ge \mu - \beta' x_i) = 1 - \Phi(\mu - \beta' x_i) \end{aligned} \quad (5)$$

The profile of our dataset of 2,054 sequences is shown in Table 1. The (Motif, Rule) = (1, 0) combination is most abundant (79.1% of total sum of sequences), with 87.5% of these sequences showing preference for CCR5. The (Motif, Rule) = (1, 1) combination is the second most abundant (11.5% of total sum of sequences), with 43.9% of these sequences showing preference for CCR5/CXCR4. The (Motif, Rule) = (0, 1) combination is the third most abundant (5.6% of total sum of sequences), with 61.4% of these sequences showing preference for CXCR4.

The analysis for the three coreceptor model was performed using the dataset of Table 1, whereas the analysis for the two coreceptor model was performed using a reduced subset of the dataset, by excluding the 322 CCR5/CXCR4 entries.

To test the accuracy and robustness of the probit predictions, a second model was produced using the 1,368 sequences (of the 2,054 total sequences) that do not have an experimentally determined CD4 count assigned. The remaining 686 sequences with assigned CD4 counts were used

Table 1. Dataset Profile with Regard to (Motif, Rule) Binary Combinations and Coreceptor Selection

(Motif, Rule)	*CCR5*	*CCR5/CXCR4*	*CXCR4*	*Sum*[a]
0, 0	35 (44.3%)[b]	16 (20.3%)[b]	28 (35.4%)[b]	79 (3.8%)
0, 1	4 (3.5%)[b]	40 (35.1%)[b]	70 (61.4%)[b]	114 (5.6%)
1, 0	1,421 (87.5%)[b]	162 (10.0%)[b]	41 (2.5%)[b]	1,624 (79.1%)
1, 1	63 (26.6%)[b]	104 (43.9%)[b]	70 (29.5%)[b]	237 (11.5%)
Total[c]	1,523 (74.1%)	322 (15.7%)	209 (10.2%)	2,054 (100%)

Number of counts (percent values) for coreceptor selection and (Motif, Rule) binary combination.

[a]Refers to the sum of the three coreceptor selections for a given (Motif, Rule) combination. The percent value in parentheses refers to the specific (Motif, Rule) count with respect to the total count of 2,054.

[b]The percent value in parentheses refers to the specific (Motif, Rule)/Coreceptor count with respect to the sum of the three coreceptor selections for the specific (Motif, Rule) given in the last column.

[c]Refers to the total number of sequences showing preference for a given coreceptor selection. The percent value in parentheses refers to the specific coreceptor count with respect to the total count of 2,054.

as a test set for side-by-side comparisons with established methods, specifically geno2pheno$_{[coreceptor]}$[27] and webPSSM.[28] The webPSSM predictions were performed using the subtype B x4r5 matrix, while the geno2pheno$_{[coreceptor]}$ predictions were performed using the original g2p coreceptor model with optimized cutoffs based on clinical data. Receiver operating characteristic (ROC) curve analysis based on the probit coreceptor probabilities was performed using the CD4 count dataset. A ROC curve for webPSSM CXCR4 preference was generated using the assigned score, while r5.pct was used to produce an ROC curve for CCR5 selection. Similarly, an ROC curve for CXCR4 selection was also produced for the geno2pheno$_{[coreceptor]}$ predictions using the assigned percentile; however, no CCR5 ROC curve was produced since the geno2pheno$_{[coreceptor]}$ server does not readily provide CCR5 analysis.

A final subset of the complete dataset, consisting of 317 sequences with assigned CD4 count and patient health status, was also identified. According to the Los Alamos HIV Databases,[25] the following disease states can be assigned to each sequence: (1) acute infection, (2) asymptomatic, (3) symptomatic, (4) AIDS, and (5) death; however, only states 1–4 are relevant for our analysis since sequences with a patient health status of death were excluded from this aspect of our results. The patient health subset was selected to allow comparisons between disease state and predictions for coreceptor selectivity. Three degrees of disease advancement have been assigned based on the patient health status: passed acute infection (patients in the asymptomatic, symptomatic, or AIDS states), passed asymptomatic phase (patients in the symptomatic or AIDS states), and AIDS.

Results and Discussion

Net charge is a significant factor in showing preference for CCR5, CXCR4, or CCR5/CXCR4, as shown by the statistical distributions of Fig. 1. The distribution that peaks at net charge of $\sim$3 denotes preference for CCR5 whereas the distribution that peaks at net charge of $\sim$5.5 denotes preference for CXCR4. An intermediate distribution denotes CCR5/CXCR4 preference, and marks the transition from CCR5 to CXCR4.

We pursued further analysis that incorporates all known markers embedded in V3 loop sequences in order to develop a quantitative estimation model for coreceptor selectivity. We used the ordered probit statistical model to account for the discrete net charge data of Fig. 1, and coreceptor selectivity binary markers of the glycosylation motif and the 11/24/25 positive amino acid rule.

We consider that the preference for coreceptor selection is implicit in the observed V3 loop sequence. Our goal is to infer the most likely coreceptor selection from the physicochemical characteristics of the observed sequence. We constructed the predictive model based on the 2,054 V3 loop unique patient sequences with known coreceptor selections from experimental data deposited at the Los Alamos HIV Databases.[25]

The estimated parameters of the ordered probit model described by Eqs. (1)–(3) are summarized in Table 2. Interpretation of the $\hat{\beta}$ coefficients of Eq. (2) suggests that the binary markers Motif and Rule have opposite effects of about similar magnitudes, denoted by the opposite sign and similar absolute values. This means that when Motif and Rule are both present, (1, 1), charge is the defining parameter in coreceptor selection. Similarly, when both Motif and Rule are absent, (0, 0) in Eq. (2), charge is the only parameter that determines coreceptor selection.

Based on the analysis described above, the estimated model is

$$\hat{y}_i^* = -0.887 Motif_i + 1.081 Rule_i + 0.356 Charge_i \qquad (6)$$

and the probabilities of being in coreceptor state 1, 2, or 3 are calculated as follows:

$$\begin{aligned} P(y_i = 1) &= \Phi(1.474 - \hat{y}_i^*) \\ P(y_i = 2) &= \Phi(2.513 - \hat{y}_i^*) - \Phi(1.474 - \hat{y}_i^*) \\ P(y_i = 3) &= 1 - \Phi(2.513 - \hat{y}_i^*) \end{aligned} \qquad (7)$$

where Φ is the cumulative standard normal distribution function. Table 3 shows the comparisons of the sample and predicted data. The count of Column 3 corresponds to assigning a value of 1 for the coreceptor state (CCR5, CCR5/CXCR4, or CXCR4), predicted by the ordered probit model, and 0 for the other two states. The data show that the prediction accuracy for CCR5 coreceptor selection is higher than that for CXCR4 (98.2% versus 56%). Prediction for CCR5/CXCR4 selection is much lower (11.5%), as the model reclassifies 285 (out of 322) entries as showing preference for

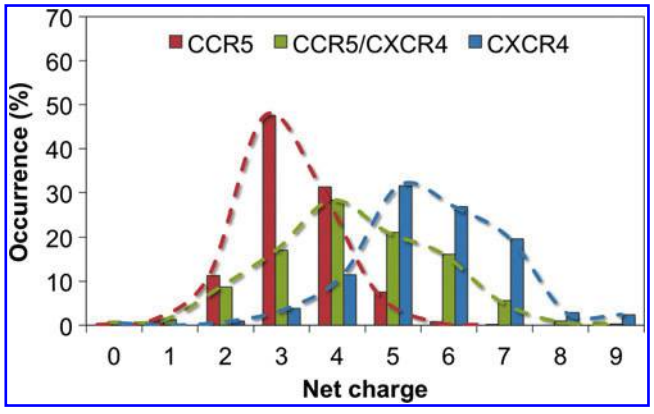


FIG. 1. Charge distributions of V3 loop sequences with known coreceptor preference.[25] Color images available online at www.liebertpub.com/aid

Table 2. Ordered Probit Model Estimated Parameters for the Three Coreceptor Model (2,054 Observations)

x_i	$\hat{\beta}$	$\sigma_{\hat{\beta}}$	$Z-\text{stat} = \frac{\hat{\beta}}{(\sigma_{\hat{\beta}})}$	p-*value*
Motif	−0.887	0.094	−9.463	0.000
Rule	1.081	0.083	13.062	0.000
Charge	0.356	0.035	10.198	0.000
		Limit points		
	$\hat{\mu}$	$\sigma_{\hat{\mu}}$	$Z-\text{stat} = \frac{\hat{\mu}}{(\sigma_{\hat{\mu}})}$	p-*value*
μ_1	1.474	0.173	8.498	0.000
μ_2	2.513	0.185	13.551	0.000

The ordered probit analysis was performed using the program EViews (Quantitative Micro Software, Irvine, CA; www.eviews.com).

Table 3. Prediction of Ordered Dependent Variable for the Three Coreceptor Model

y_i	*Dataset sample count*	*Correct count of observations*	*Incorrect count of observations*	*% Correct*	*% Incorrect*
1	1,523	1,496	27	98.227	1.773
2	322	37	285	11.491	88.509
3	209	117	92	55.981	44.019
Total	2,054	1,650	404	80.331	19.669

coreceptor CCR5 or CXCR4. This observation may reflect the fact that under the CCR5/CXCR4 category are placed both dual and mixed coreceptor selections. Use of the term "dual" implies the capability to bind to either CCR5 or CXCR4 coreceptors, whereas the term "mixed" implies a viral population that may contain combinations of CCR5-, CXCR4-, and/or dual-binding viral strains. Because of the physicochemical basis of our V3 loop analysis, it is likely that our predictive model can discriminate between dual and mixed coreceptor selection. This argument suggests that the predicted count for coreceptor state 2 (37 counts in Table 3) refers to dual CCR5/CXCR4 coreceptor selection.

Comparison of the predicted coreceptor assignments to the database assignments is shown in Table 4. A significant portion of the CCR5/CXCR4 and CXCR4 database assignments is reassigned (more frequently to CCR5), as discussed above. The origin of the reassignments is not understood at the moment, and possibly reflects the use of physicochemical properties (probit model) compared to the use of a variety of experimental methods by several different researchers in populating the database. This issue may be resolved in future work using self-consistent datasets derived from identical experimental methodologies, as well as time-dependent data from individual patients. Perhaps the most accurate experimental method to determine coreceptor selection depends on the use of cells that express CCR5 or CXCR4 only.

Table 4. Predictive Performance of the Probit Model for HIV-1 Coreceptor Selection

Count % Total % Row	*Predicted coreceptor*			*Total (database assignment)*
	CCR5	*CCR5/CXCR4*	*CXCR4*	
CCR5	*1,496*	18	9	**1,523**
	72.83	0.88	0.44	**74.15**
	98.23	1.18	0.59	**100.00**
	197	*37*	88	**322**
CCR5/CXCR4	9.59	*1.80*	4.28	**15.68**
	61.18	*11.49*	27.33	**100.00**
	49	43	*117*	**209**
CXCR4	2.39	2.09	*5.70*	**10.18**
	23.44	20.57	*55.98*	**100.00**
Total (Probit reassignment)	**1,742**	**98**	**214**	**2,054**
	84.81	**4.77**	**10.42**	**100.00**
	84.81	**4.77**	**10.42**	**100.00**

Count refers to predicted coreceptor assignments (reassignments) compared to the database assignments.

Italicized entries correspond to correct predictions. Boldfaced entries correspond to totals from the database assignment and probit reassignment. The rest of the entries correspond to lost (columns)/gained (rows) assignments.

Equations (6) and (7) can be used in predicting probabilities for coreceptor selectivity for a patient's experimentally derived V3 loop sequence, by simply assessing the presence or absence of the $N^6X^7[T/S]^8X^9$ glycosylation motif and 11/24/25 positive amino acid rule and by determining the net charge of the sequence (derived by summing the number of positively and negatively charged amino acids).

Figure 2 shows graphically the calculated probabilities as a function of net charge, glycosylation motif, and 11/24/25 positive amino acid rule. The calculated probabilities show that for (Motif, Rule) = (1, 0) there is a preference for CCR5 as charge decreases (Fig. 2A), whereas the opposite happens for (Motif, Rule) = (0, 1) for which there is preference for CXCR4 as charge increases (Fig. 2C). For (Motif, Rule) = (1, 1) charge is the dominant factor, and for (Motif, Rule) = (0, 0) charge is the only factor, in both cases favoring CCR5 as charge decreases and CXCR4 as charge increases (Fig. 2A and C). The probabilities for CCR5/CXCR4 preferences, marking the transition from CCR5 to CXCR4, are shown in Fig. 2B, and include the overlapping region between the probabilities for CCR5 and CXCR4 preferences.

The data of Fig. 2 suggest that the CCR5 preference when the glycosylation motif is present [case of (Motif, Rule) = (1, 0)] switches to CCR5/CXCR4 preference upon incorporation of a positive amino acid according to the 11/24/25 rule [and concurrent net charge increase, case of (Motif, Rule) = (1, 1)], and subsequently switches to CXCR4 preference upon loss of glycosylation capacity and concurrent net charge increase [case of (Motif, Rule) = (0, 1)]. Simultaneous loss of glycosylation capacity and positive charge at the 11/24/25 positions [least abundant combination of (Motif, Rule) = (0, 0)] shows no apparent preference for any of the coreceptor selections.

We have also performed the probit analysis using only two coreceptors, CCR5 and CXCR4, using Eqs. (4) and (5). The resulting estimated model is

$$\hat{y}_i^* = -1.294 Motif_i + 1.276 Rule_i + 0.621 Charge_i \qquad (8)$$

and the probabilities of being in coreceptor state 1 or 2 are

$$P(y_i = 1) = \Phi(2.961 - \hat{y}_i^*)$$
$$P(y_i = 2) = 1 - \Phi(2.961 - \hat{y}_i^*) \qquad (9)$$

The probit results are summarized in Fig. 3 and Supplementary Tables S1 and S2 (Supplementary Data are available online at www.liebertpub.com/aid), and they are on par with the results of the three coreceptor model. This is evidenced by the opposite sign, and about equal magnitude of the $\hat{\beta}$ coefficients, showing that the binary markers Motif and Rule have opposite effects of about similar magnitudes. It is also evidenced in Fig. 3, which shows that that the graphs for cases of (Motif, Rule) = (1, 1) or (0, 0) are overlapping. The critical role of charge for coreceptor selection in the cases of (Motif, Rule) = (1, 1) or (0, 0) is demonstrated in both the three and two coreceptor models. Indeed, charge is almost twice as strong a determining factor in the two coreceptor model compared to the three coreceptor model, given the magnitudes of the $\hat{\beta}$ coefficients. Supplementary Table S2 also shows that the prediction of correct CXCR4 assignments has

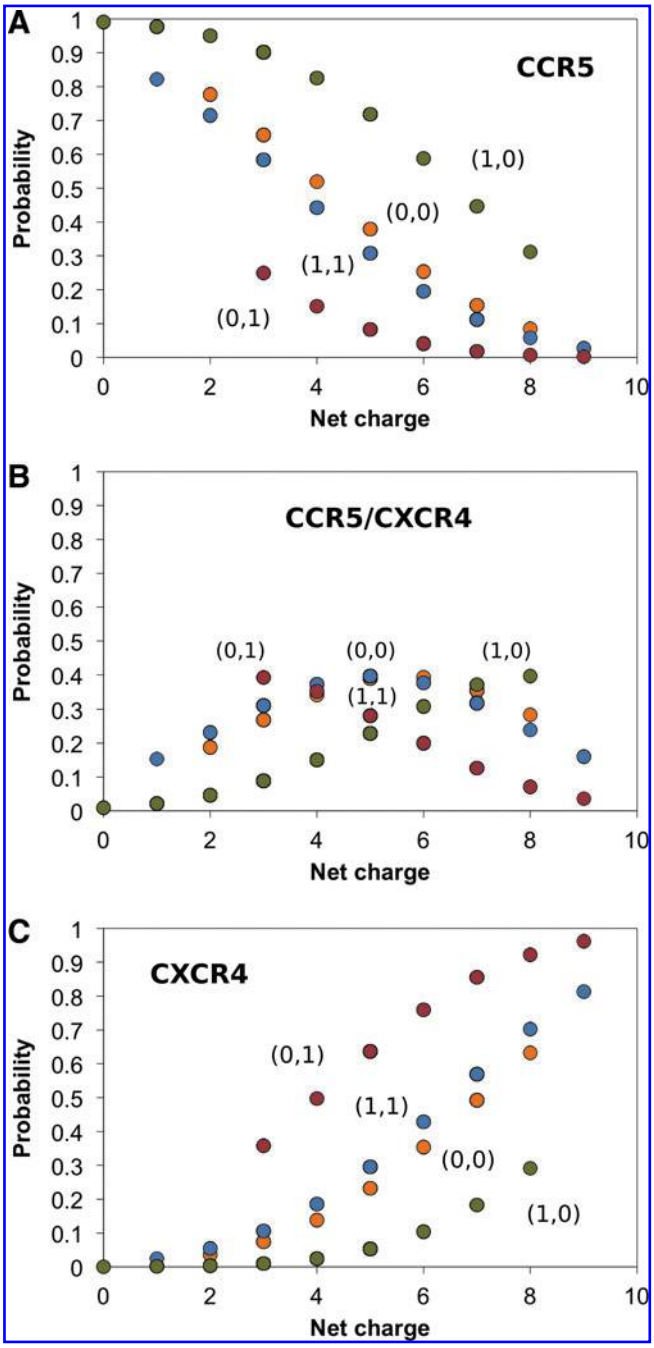


FIG. 2. Probit analysis of V3 loop sequences using the three coreceptor disease model. Probabilities for coreceptor preference taking into account the property of net charge in the range 0–9 and the binary (1 for presence and 0 for absence) coreceptor markers of the $N^6X^7[T/S]^8X^9$ glycosylation motif (Motif) and the 11/24/25 positive amino acid rule (Rule), marked as (Motif, Rule) pairs. **(A)** CCR5. **(B)** CCR5/CXCR4. **(C)** CXCR4. Combinations of (Motif, Rule) = (0, 0), (0, 1), (1, 0), (1, 1) are shown in orange, red, green, and blue, respectively. Color images available online at www.liebertpub.com/aid

increased, given the absence of the CCR5/CXCR4 state. In comparison, we consider the three coreceptor model more general than the two coreceptor model, because the CCR5/CXCR4 state is supported by experimental data, and it incorporates a transition state between viral selection of CCR5 and CXCR4.

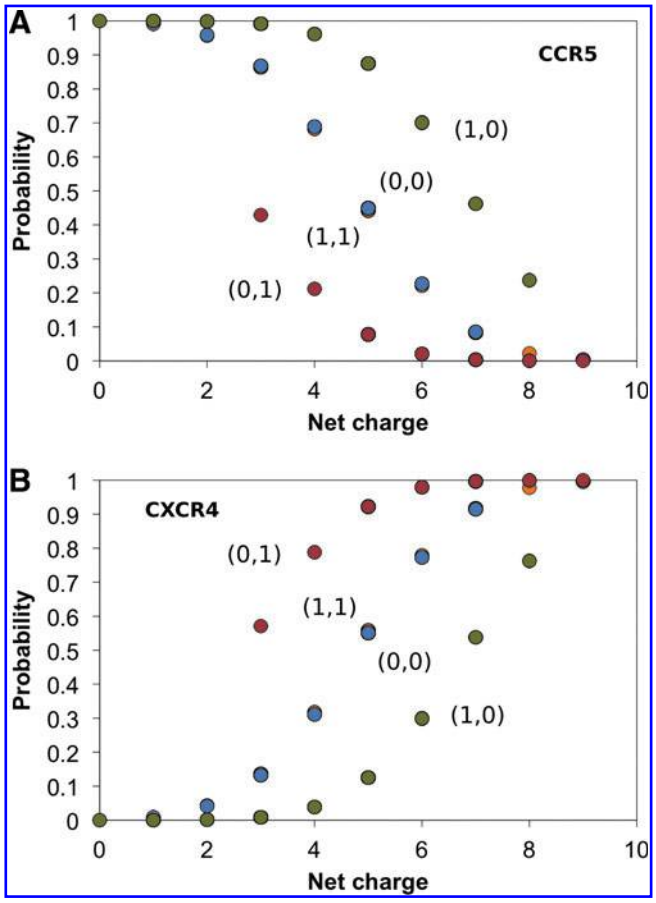


FIG. 3. Probit analysis of V3 loop sequences using the two coreceptor model. The presentation of the data is similar to the presentation of Fig. 2. **(A)** CCR5. **(B)** CXCR4. Color images available online at www.liebertpub.com/aid

The two phenotypic classes of HIV-1 strains are syncitia-inducing (SI) strains (infecting $CD4^+$ T cells), and non-syncitia-inducing (NSI) strains (infecting macrophages and $CD4^+$ T cells). The latter strains show preference for use of CCR5 for cell entry and are associated with the primary infection, whereas the former show preference for use of CCR5/CXCR4 or CXCR4 for cell entry and are associated with a rapid reduction in CD4 count and disease progression. Current laboratory tests to diagnose AIDS are CD4 counts, viral loads, and genotypic or phenotypic resistance tests.[26] With CD4 counts being a deposited parameter for a subset of the Los Alamos dataset, we have split the dataset of 2,054 sequences into a subset with associated experimental CD4 counts (686 entries) and a subset without available CD4 counts (1,368 entries). We used the subset without CD4 counts to test the robustness of the probit model and to train a model that could be used in blind prediction of the subset with CD4 counts.

The probit results for the three coreceptor model using the reduced dataset without CD4 counts are summarized in Supplementary Tables S3–S5, and they are similar to those with the complete dataset of 2,054 entries (Tables 2–4). Figure 4 shows ROC curves for prediction of coreceptor selection for the CD4 count dataset, demonstrating rather high predictive values for CCR5 and CXCR4, but lower predictive value for CCR5/CXCR4. The poor prediction for CCR5/CXCR4 is potentially due to vagueness in the experimental methods

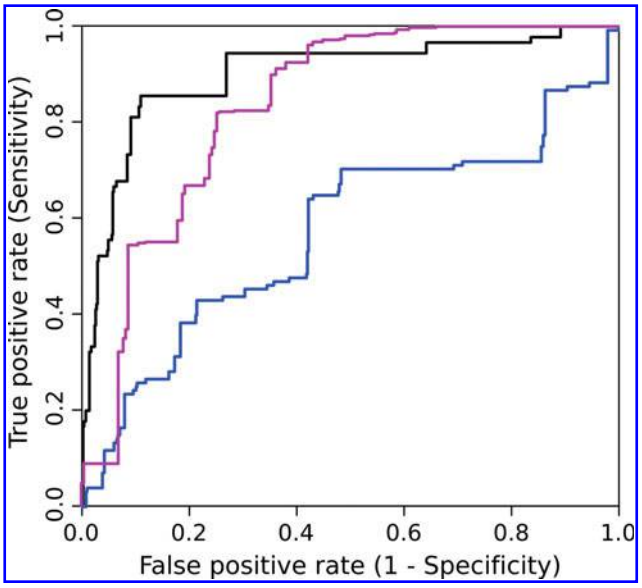


FIG. 4. Accuracy of probit coreceptor preference predictions. Receiver operating characteristic (ROC) curve analysis for the three coreceptor preferences [color, area under the curve (AUC)]: CCR5 (magenta, 0.833); CCR5/CXCR4 (blue, 0.571); CXCR4 (black, 0.900). Color images available online at www.liebertpub.com/aid

used in determining a dual tropic virus, as is suggested by the fact that the probit model reclassified most of the CCR5/CXCR4 sequences. We also used the probit model trained with the subset without CD4 counts to evaluate the probit performance in comparison to predictions from existing popular servers, geno2pheno$_{[coreceptor]}$[18,27] and webPSSM.[19,28] Figure 5A shows ROC curves for predictions of CXCR4 coreceptor selection for the CD4 count dataset, using probit, webPSSM, and geno2pheno$_{[coreceptor]}$. Probit and geno2pheno$_{[coreceptor]}$ perform comparably, whereas webPSSM performs slightly better for this dataset. Figure 5B shows a similar analysis of predictions of CCR5 coreceptor selection for the CD4 count dataset, using probit and webPSSM, while the geno2pheno$_{[coreceptor]}$ web server does not directly predict CCR5. Both methods perform equally well for CCR5 prediction.

Although it is known that as the infection/disease progresses a switch for coreceptor preference occurs, starting with selection of CCR5 and continuing with selection of CCR5/CXCR4 and CXCR4,[2–5,10–21] it is debatable if coreceptor selection may be predictive of disease state. To this end, we have analyzed the relationship between coreceptor selectivity and disease state, based on the probit predictions using the "patient health status" subset described in Materials and Methods. Figure 6 contains ROC curves illustrating the prediction of the AIDS patient health status based on preference for CXCR4, as assigned by probit, webPSSM, and geno2pheno$_{[coreceptor]}$. CXCR4 preferences calculated by all three methods perform comparably well at predicting the AIDS status, with areas under the curve (AUC) of $\sim$0.7. Additionally, Fig. 6 also contains an ROC curve for AIDS status prediction based on experimentally assigned CD4 count, as a comparison. Surprisingly, the computationally predicted CXCR4 preferences perform similarly to CD4 count in assigning the AIDS status at high specificity values ($>$0.8).

We have also performed analysis of the utility of probit predicted coreceptor preference in assigning degree of disease advancement. Supplemental Fig. S1 contains ROC curves for prediction of disease progression based on probit coreceptor preference, as well as CD4 count. CCR5 probability shows some predictive value for advancement passed the asymptomatic phase, and inversely predictive of the AIDS state. CCR5/CXCR4 probability has some predictive value for all three degrees of advancement, while CXCR4 probability shows the highest AUC for the prediction of the AIDS state. These results provide some evidence supporting the hypothesis that coreceptor selection may be indicative, but not a quantitative predictor, of disease state. Interestingly, despite these observed relationships between coreceptor selectivity and disease state, there is only a weak correlation between CD4 count and coreceptor selection. Supplemental Fig. S2 shows a graph of CD4 counts per coreceptor assignment (provided by the Los Alamos HIV Databases) for the CD4 count dataset of our sample, and Supplementary Table S6 shows the sample statistics. Although there is a trend in decreasing mean and median as we transition from selecting coreceptor CCR5 to CCR5/CXCR4 to CXCR4, there are many entries below the medically accepted threshold for AIDS diagnosis (200 counts) associated with CCR5 selection. The observations discussed above provide insight into the

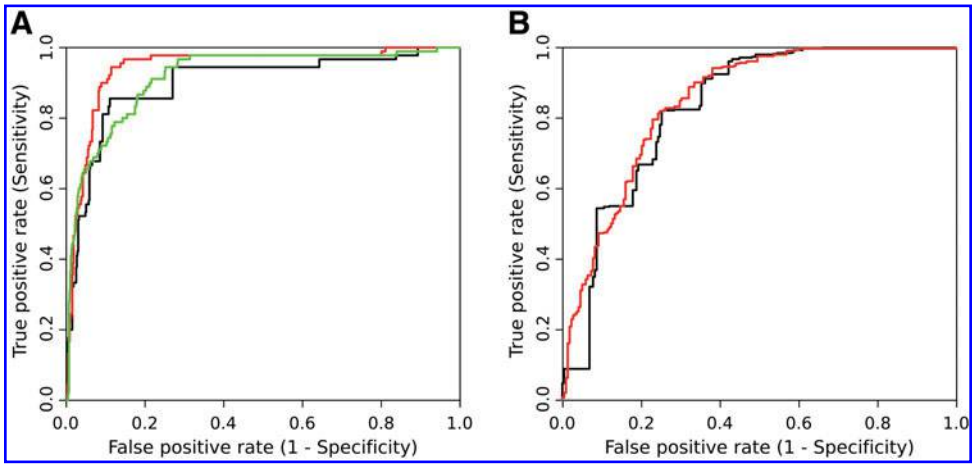


FIG. 5. Comparison of probit predictions with established methods. **(A)** ROC curve analysis for prediction of CXCR4 preference by (color, AUC) probit (black, 0.900); webPSSM (red, 0.943); geno2pheno$_{[coreceptor]}$ (green, 0.918). **(B)** ROC curve analysis for prediction of CCR5 preference by (color, AUC) probit (black, 0.833); webPSSM (red, 0.848). Color images available online at www.liebertpub.com/aid

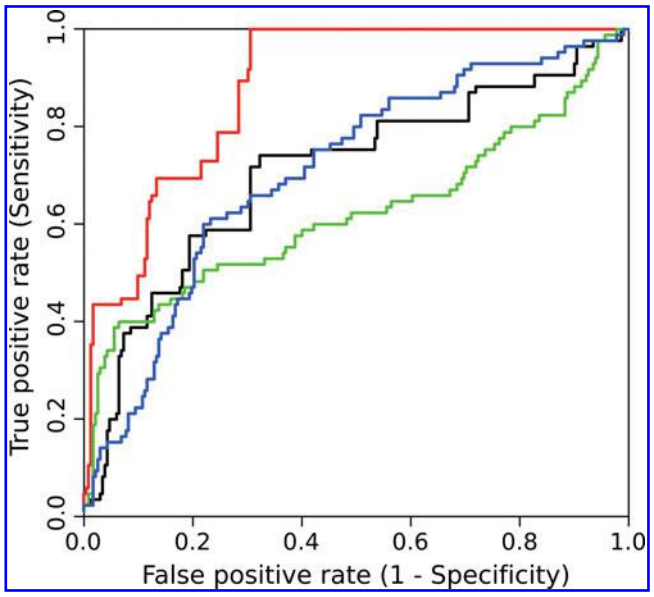


FIG. 6. Prediction of AIDS patient health status based on CXCR4 preference. ROC curve analysis for the prediction of AIDS patient health status based on CXCR4 preference, as predicted by (color, AUC) probit (black, 0.706); webPSSM (green, 0.620); geno2pheno$_{[coreceptor]}$ (blue, 0.708). For comparison a ROC curve for the prediction of the AIDS disease state based on CD4 count (red, 0.881) is also presented. Color images available online at www.liebertpub.com/aid

relationship between disease progression and coreceptor selection, and can serve as the foundation for the development of predictive models for HIV disease state progression.

The probit predictive model contributes to the available tools for the analysis of HIV sequence data and for the prediction of coreceptor selectivity,[18,19,29–40] including web server tools geno2pheno$_{[coreceptor]}$[27] and webPSSM.[28] What distinguishes probit from other methods is its simplicity. The probit model uses only three physicochemical characteristics that are embedded in the V3 loop genetic code, without dependencies of multiple adjustable parameters or heuristic arguments, and without the necessity for prior sequence alignments, nor a reliance on sequence templates.

Knowledge of coreceptor assignment provides information on the first contact point for viral entry, and therefore may be useful in determining medication targeting CCR5 or CXCR4 and at what ratio. Currently, there is one CCR5 entry drug clinically available and several CCR5 and CXCR4 entry drugs are in the pipeline.[41–43] The need for CXCR4 entry drugs is evident, considering that development of drug resistance against CCR5 entry drugs is manifested as coreceptor switch from CCR5 to CXCR4.[42] The use of probit and other methods will be beneficial once the option of having both CCR5 and CXCR4 entry drugs becomes clinically available.

Conclusions

In practical terms, the predictive use of our model is demonstrated in Table 5, and the graphic presentation of the predictions is shown in Fig. 7. The predicted probabilities for selecting coreceptor CCR5, CCR5/CXCR4, and CXCR4, calculated using Eqs. (6) and (7), are shown in the net charge range of 0–10, and at the four (Motif, Rule) binary combinations described above. Table 5 can be used for quick and efficient assessment of coreceptor selection, and associated HIV-1 tropism, for an unknown V3 loop sequence.

Although charge alone is a strong marker for coreceptor preference at extreme charge values, it is a less definitive marker at intermediate charge values, where combinations of $N^6X^7[T/S]^8X^9$ glycosylation motif and 11/24/25 positive amino acid rule become discriminating factors (Figs. 2 and 7). The ordered probit model is useful to predict probabilities for CCR5, CCR5/CXCR4, and CXCR4 selection, using information for coreceptor preference that is found in the V3 loop sequence. Given the nature of viral infection and the fact that numerous viral strains with different coreceptor preferences may be present in a patient, a probabilistic model is suitable to assign percent coreceptor preference based on observed V3 loop sequences. The sequence-based analysis presented here can be used to predict coreceptor selection and may potentially be used to make personalized medical decisions, in addition to existing tools, for administration of drugs or

Table 5. Predictive Value of the Probit Model for HIV-1 Coreceptor Selection

Net charge	*0.0*	*1.0*	*2.0*	*3.0*	*4.0*	*5.0*	*6.0*	*7.0*	*8.0*	*9.0*	*10.0*
(Motif, Rule) = (0, 0)											
CCR5	0.93	0.87	0.78	0.66	0.52	0.38	0.25	0.15	0.08	0.04	0.02
CCR5/CXCR4	0.06	0.12	0.19	0.27	0.34	0.39	0.39	0.35	0.28	0.20	0.13
CXCR4	0.01	0.02	0.04	0.07	0.14	0.23	0.35	0.49	0.63	0.76	0.85
(Motif, Rule) = (0, 1)											
CCR5	0.65	0.51	0.37	0.25	0.15	0.08	0.04	0.02	0.01	0.00	0.00
CCR5/CXCR4	0.27	0.34	0.39	0.39	0.35	0.28	0.20	0.13	0.07	0.04	0.02
CXCR4	0.08	0.14	0.24	0.36	0.50	0.64	0.76	0.86	0.92	0.96	0.98
(Motif, Rule) = (1, 0)											
CCR5	0.99	0.98	0.95	0.90	0.83	0.72	0.59	0.45	0.31	0.20	0.11
CCR5/CXCR4	0.01	0.02	0.05	0.09	0.15	0.23	0.31	0.37	0.40	0.38	0.32
CXCR4	0.00	0.00	0.00	0.01	0.02	0.05	0.10	0.18	0.29	0.42	0.57
(Motif, Rule) = (1,1)											
CCR5	0.90	0.82	0.71	0.58	0.44	0.31	0.20	0.11	0.06	0.03	0.01
CCR5/CXCR4	0.09	0.15	0.23	0.31	0.37	0.40	0.38	0.32	0.24	0.16	0.10
CXCR4	0.01	0.02	0.05	0.11	0.19	0.30	0.43	0.57	0.70	0.81	0.89

Probabilities for coreceptor selection, accounting for net charge in the range of 0–10 and the four (Motif, Rule) binary combinations.

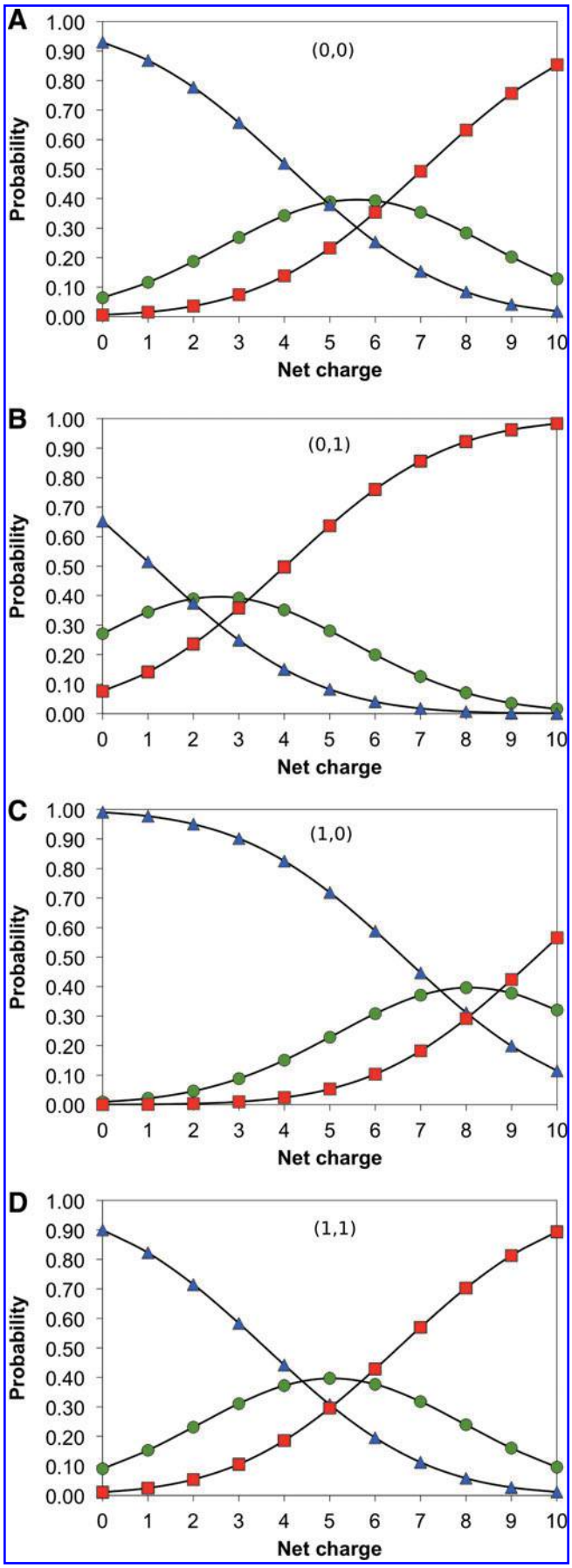


FIG. 7. The predictive use of the probit model. Graphic representation of predicted data corresponding to Table 5. **(A)** (Motif, Rule) = (0, 0). **(B)** (Motif, Rule) = (0, 1). **(C)** (Motif, Rule) = (1, 0). **(D)** (Motif, Rule) = (1, 1). Data for coreceptors CXCR5, CCR5/CXCR4, and CXCR4 are shown in blue, green, and red, respectively. These graphs can be used to predict probabilities for each coreceptor selection for a new sequence, taking into account the net charge of the sequence, and the presence or absence (1 or 0) of the $N^6X^7[T/S]^8X^9$ glycosylation motif and the 11/24/25 positive amino acid rule in the sequence. Color images available online at www.liebertpub.com/aid

combinations of HIV-1 entry drugs targeting CCR5 and/or CXCR4. Additional validation work with different experimental datasets, preferably using cells that express only one coreceptor, as well as predictive model refinement, will be necessary in reaching clinical applications.

Author Disclosure Statement

No competing financial interests exist.

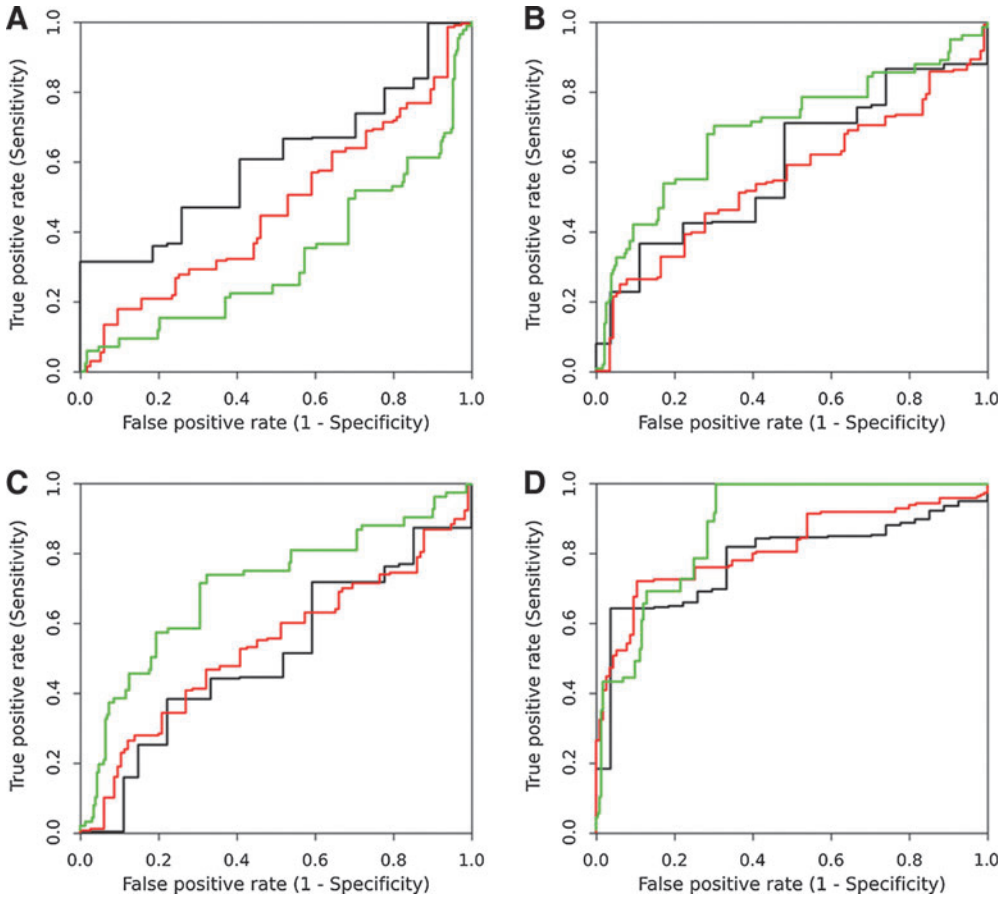
References

1. Emini EA and Koff WC: AIDS/HIV. Developing an AIDS vaccine: Need, uncertainty, hope. Science 2004;304:1913–1914.
2. Wyatt R and Sodroski J: The HIV-1 envelope glycoproteins: Fusogens, antigens, and immunogens. Science 1998;280: 1884–1888.
3. Berger EA, Doms RW, Fenyo EM, *et al.:* A new classification for HIV-1. Nature 1998;391:240–240.
4. Berger EA, Murphy PM, and Farber JM: Chemokine receptors as HIV-1 coreceptors: Roles in viral entry, tropism, and disease. Annu Rev Immunol 1999;17:657–700.
5. Broder CC and Jones-Trower A: Coreceptor use by primate lentiviruses. In: *Human Retroviruses and AIDS 1999* (Kuiken CL *et al.*, eds.). Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM, 1999, pp. 517–541.
6. Kuiken C, *et al.*, eds.: *HIV Sequence Compendium 2011,* Vol LA-UR 10-03684. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM, 2011.
7. Huang CC, Tang M, Zhang MY, *et al.:* Structure of a V3-containing HIV-1 gp120 core. Science 2005;310:1025–1028.
8. Huang CC, Lam SN, Acharya P, *et al.:* Structures of the CCR5 N terminus and of a tyrosine-sulfated antibody with HIV-1 gp120 and CD4. Science 2007;317:1930–1934.
9. López de Victoria A, Tamamis P, Kieslich CA, and Morikis D: Insights into the structure, correlated motions, and electrostatic properties of two HIV-1 gp120 V3 loops. PLoS One 2012;7:e49925.
10. Fouchier RAM, Groenink M, Kootstra NA, *et al.:* Phenotype-associated sequence variation in the third variable domain of the human immunodeficiency virus type 1 gp120 molecule. J Virol 1992;66:3183–3187.
11. López de Victoria A, Kieslich CA, Rizos AK, Krambovitis E, and Morikis D: Clustering of HIV-1 subtypes based on gp120 V3 loop electrostatic properties. BMC Biophys 2012;5:3.
12. Moore JP: Co-receptors for HIV-1 entry. Curr Opin Immunol 1997;9:551–562.
13. Connor RI, Sheridan KE, Ceradini D, Choe S, and Landau NR: Change in coreceptor use correlates with disease progression in HIV-1-infected individuals. J Exp Med 1997;185: 621–628.

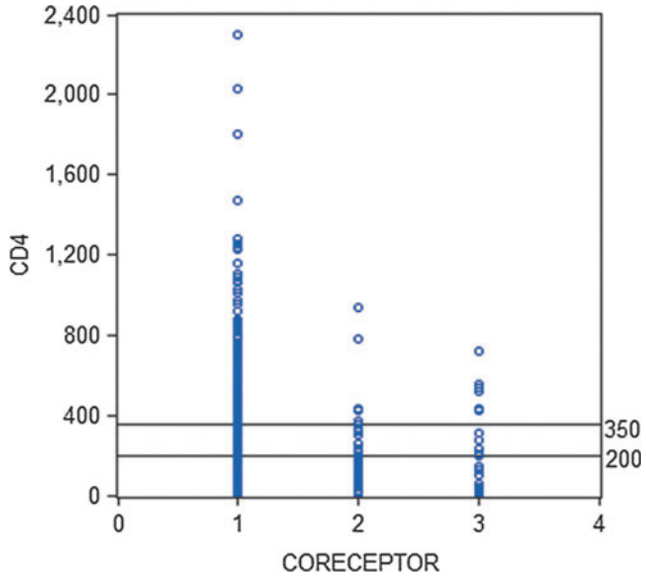
14. Maas JJJ, Gange SJ, Schuitemaker H, Coutinho RA, van Leeuwen R, and Margolick JB: Strong association between failure of T cell homeostasis and the syncytium-inducing phenotype among HIV-1-infected men in the Amsterdam Cohort Study. AIDS 2000;14:1155–1161.
15. Scarlatti G, Tresoldi E, Bjorndal A, *et al.:* In vivo evolution of HIV-1 co-receptor usage and sensitivity to chemokine-mediated suppression. Nature Med 1997;3:1259–1265.
16. Doms RW and Moore JP: HIV-1 Coreceptor use: A molecular window into viral tropism. In: *HIV Molecular Immunology Database 1997*, Vol. IV (Korber B *et al.,* eds.). Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM, 1997, pp. IV-25–36.
17. Koning F, van Rij R, and Schuitemaker H: Biological and molecular aspects of HIV-1 coreceptor usage. In: *HIV Sequence Compendium 2002* (Kuiken C *et al*., eds.). Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM, 2002, pp. 24–42.
18. Lengauer T, Sander O, Sierra S, Thielen A, and Kaiser R: Predicting disease outcomes in the clinic. Nature Biotechnol 2008;26:612–613.
19. Jensen MA, Li FS, van't Wout AB, *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. J Virol 2003;77:13376–13388.
20. Schuitemaker H, van't Wout AB, and Lusso P: Clinical significance of HIV-1 coreceptor usage. J Transl Med 2011; 9(Suppl 1):S5.
21. Poon AFY, Swenson LC, Bunnik EM, *et al.:* Reconstructing the dynamics of HIV evolution within hosts from serial deep sequence data. PLoS Comput Biol 2012;8:e1002753.
22. Pollakis G, Kang S, Kliphuis A, Chalaby MI, Goudsmit J, and Paxton WA: N-linked glycosylation of the HIV type-1 gp120 envelope glycoprotein as a major determinant of CCR5 and CXCR4 coreceptor utilization. J Biol Chem 2001;276:13433–13441.
23. The abbreviations of the amino acids of the glycosylation motif are N for asparagine (the glycosylation site), T for threonine, S for serine, and X for any amino acids except for proline.
24. Cardozo T, Kimura T, Philpott S, Weiser B, Burger H, and Zolla-Pazner S: Structural basis for coreceptor selectivity by the HIV type 1 V3 loop. AIDS Res Hum Retroviruses 2007;23:415–426.
25. http://hiv.lanl.gov/
26. http://aids.gov/hiv-aids-basics/
27. http://geno2pheno.org/
28. http://indra.mullins.microbiol.washington.edu/webpssm/.
29. Pillai S, Good B, Richman D, and Corbeil J: A new perspective on V3 phenotype prediction. AIDS Res Hum Retroviruses 2003;19:145–149.
30. Low AJ, Dong W, Chan D, *et al.:* Current V3 genotyping algorithms are inadequate for predicting X4 co-receptor usage in clinical isolates. AIDS 2007;21:F19–F26.
31. Jensen MA, Coetzer M, van 't Wout AB, Morris L, and Mullins JI: A reliable phenotype predictor for human immunodeficiency virus type 1 subtype C based on envelope V3 sequences. J Virol 2006;80:4698–4704.
32. Brumme ZL, Dong WWY, Yip B, *et al.:* Clinical and immunological impact of HIV envelope V3 sequence variation after starting initial triple antiretroviral therapy. AIDS 2004;18:F1–F9.
33. Masso M and Vaisman II: Accurate and efficient gp120 V3 loop structure based models for the determination of HIV-1 co-receptor usage. BMC Bioinform 2010;11:494.
34. Masso M and Vaisman II: AUTO-MUTE: Web-based tools for predicting stability changes in proteins due to single amino acid replacements. Protein Eng Des Sel 2010;23:683–687.
35. Prosperi MCF, Fanti I, Ulivi G, Micarelli A, De Luca A, and Zazzi M: Robust supervised and unsupervised statistical learning for HIV type 1 coreceptor usage analysis. AIDS Res Hum Retroviruses 2009;25:305–314.
36. Sander O, Sing T, Sommer I, *et al.:* Structural descriptors of gp120 V3 loop for the prediction of HIV-1 coreceptor usage. Plos Comput Biol 2007;3:555–564.
37. Sing T, Low AJ, Beerenwinkel N, *et al.:* Predicting HIV coreceptor usage on the basis of genetic and clinical covariates. Antivir Ther 2007;12:1097–1106.
38. Xu S, Huang X, Xu H, and Zhang C: Improved prediction of coreceptor usage and phenotype of HIV-1 based on combined features of V3 loop sequence using random forest. J Microbiol 2007;45:441–444.
39. Resch W, Hoffman N, and Swanstrom R: Improved success of phenotype prediction of the human immunodeficiency virus type 1 from envelope variable loop 3 sequence using neural networks. Virology 2001;288:51–62.
40. Briggs DR, Tuffle DL, Sleasman JW, and Goodenow MM: Envelope V3 amino acid sequence predicts HIV-1 phenotype (co-receptor usage and tropism for macrophages). AIDS 2000;14:2937–2939.
41. http://aidsmeds.com/list.shtml.
42. Lobritz MA, Ratcliff AN, and Arts EJ: HIV-1 entry, inhibitors, and resistance. Viruses-Basel 2010;2:1069–1105.
43. Tilton JC and Doms RW: Entry inhibitors in the treatment of HIV-1 infection. Antiviral Res 2010;85:91–100.

Address correspondence to:
Dimitrios Morikis
Department of Bioengineering
University of California
Riverside, California 92521

E-mail: dmorikis@ucr.edu



SUPPLEMENTARY FIG. S1. Prediction of disease progression based on probit-predicted coreceptor preference. ROC curves for prediction of the three states of disease progression [as defined in Materials and Methods (passed acute infection, passed asymptomatic phase, and AIDS)] were generated based on the subset of sequences with assigned Patient Health Status and CD4 count. **(A)** ROC curve analysis for the prediction of disease progression based on CCR5 preference (color, AUC): passed acute infection (black, 0.608); passed asymptomatic phase (red, 0.529); AIDS (green, 0.374). **(B)** ROC curve analysis for the prediction of disease progression based on CCR5/CXCR4 preference (color, AUC): passed acute infection (black, 0.593); passed asymptomatic phase (red, 0.557); AIDS (green, 0.698). **(C)** ROC curve analysis for the prediction of disease progression based on CXCR4 preference (color, AUC): passed acute infection (black, 0.510); passed asymptomatic phase (red, 0.536); AIDS (green, 0.706). **(D)** ROC curve analysis for the prediction of disease progression based on CD4 count (color, AUC): passed acute infection (black, 0.785); passed asymptomatic phase (red, 0.816); AIDS (green, 0.881). AUC, area under the curve.



SUPPLEMENTARY FIG. S2. Graph of CD4 count for each coreceptor assignment using the reduced dataset with CD4 counts (the number of $CD4^+$ T cells per μl). Coreceptors CCR5, CCR5/CXCR4, and CXCR4 are represented by 1, 2, and 3, respectively, in the horizontal axis. The horizontal lines correspond to a CD4 count of 350 and 200. A normal CD4 count is in the range of 500–1,000, and a CD4 count of <200 is used for AIDS diagnosis; a CD4 count of <350 suggests initiation of treatment (http://aids.gov/hiv-aids-basics/).

Supplementary Table S1. Ordered Probit Model Estimated Parameters for the Two Coreceptor Model (2,054 Observations)

x_i	$\hat{\beta}$	$\sigma_{\hat{\beta}}$	$Z-\text{stat}=\frac{\hat{\beta}}{(\sigma_{\hat{\beta}})}$	p-*value*
Motif	−1.294	0.156	−8.306	0.000
Rule	1.276	0.129	9.861	0.000
Charge	0.621	0.080	7.791	0.000
		Limit points		
	$\hat{\mu}$	$\sigma_{\hat{\mu}}$	$Z-\text{stat}=\frac{\hat{\mu}}{(\sigma_{\hat{\mu}})}$	p-*value*
μ	2.961	0.380	7.791	0.000

The ordered probit analysis was performed using the program EViews (Quantitative Micro Software, Irvine, CA; www.eviews.com).

Supplementary Table S2. Prediction of Ordered Dependent Variable for the Two Coreceptor Model

y_i	*Dataset sample count*	*Correct count of observations*	*Incorrect count of observations*	*% Correct*	*% Incorrect*
1	1,523	1,496	27	98.227	1.773
2	209	164	45	78.469	21.531
Total	1,732	1,660	72	95.843	4.157

Supplementary Table S3. Ordered Probit Model Estimated Parameters for the Three Coreceptor Model Using the Reduced Dataset Without CD4 Counts (1,368 Observations)

x_i	$\hat{\beta}$	$\sigma_{\hat{\beta}}$	$Z-\text{stat}=\frac{\hat{\beta}}{(\sigma_{\hat{\beta}})}$	p-*value*
Motif	−0.840	0.118	−7.123	0.000
Rule	1.253	0.105	11.905	0.000
Charge	0.338	0.046	7.318	0.000
		Limit points		
	$\hat{\mu}$	$\sigma_{\hat{\mu}}$	$Z-\text{stat}=\frac{\hat{\mu}}{(\sigma_{\hat{\mu}})}$	p-*value*
μ_1	1.598	0.219	7.284	0.000
μ_2	2.654	0.238	11.143	0.000

The ordered probit analysis was performed using the program EViews (Quantitative Micro Software, Irvine, CA; www.eviews.com).

Supplementary Table S4. Prediction of Ordered Dependent Variable for the Three Coreceptor Model Using the Reduced Dataset Without CD4 Counts

y_i	*Dataset sample count*	*Correct count of observations*	*Incorrect count of observations*	*% Correct*	*% Incorrect*
1	1,055	1,038	17	98.389	1.611
2	194	16	178	8.247	91.753
3	119	66	53	55.462	44.538
Total	1,368	1,120	248	81.871	18.129

Supplementary Table S5. Performance of the Three Coreceptor Probit Model Developed with the Reduced Dataset Without CD4 Counts, Used to Predict Coreceptor Selection for the Reduced Dataset With CD4 Counts

Count *% Total* *% Row*	*Predicted coreceptor*			*Total (database assignment)*
	CCR5	*CCR5/CXCR4*	*CXCR4*	
CCR5	*459*	8	1	**468**
	66.91	1.17	0.15	**68.22**
	98.08	1.71	0.21	**100.00**
	87	*14*	27	**128**
CCR5/CXCR4	12.68	*2.04*	3.94	**18.66**
	67.97	*10.94*	21.09	**100.00**
	31	19	*40*	**90**
CXCR4	4.52	2.77	*5.83*	**13.12**
	34.44	21.11	*44.44*	**100.00**
Total (Probit reassignment)	**577**	**41**	**68**	**686**
	84.11	**5.98**	**9.91**	**100.00**
	84.11	**5.98**	**9.91**	**100.00**

In the table entries "Count" refers to predicted coreceptor assignments (reassignments) compared to the database assignments.

Italicized entries correspond to correct predictions. Boldfaced entries correspond to totals from the database assignment and probit reassignment. The rest of the entries correspond to lost (columns)/gained (rows) assignments.

Supplementary Table S6. Statistics for the Reduced Dataset With CD4 Counts (686 Observations)

Coreceptor	*Mean*	*Median*	*Max*	*Min*	*Quant.*[a]	*Std. Dev.*	*Observ.*
1	387.833	331.5	2,300	6	409.0	297.270	468
2	183.391	147.5	940	2	178.0	142.505	128
3	207.822	140.0	720	2	198.5	199.522	90
All	326.070	260.5	2,300	2	333.1	278.264	686

[a]Quantiles computed for $p = 0.6$, using the Rankit (Cleveland) definition.