AIDS RESEARCH AND HUMAN RETROVIRUSES
Volume 19, Number 2, 2003, pp. 145–149
© Mary Ann Liebert, Inc.

Sequence Note

A New Perspective on V3 Phenotype Prediction

SATISH PILLAI,[1,*] BENJAMIN GOOD,[2,*] DOUGLAS RICHMAN,[1,2] and JACQUES CORBEIL[1,2]

ABSTRACT

The particular coreceptor used by a strain of HIV-1 to enter a host cell is highly indicative of its pathology. HIV-1 coreceptor usage is primarily determined by the amino acid sequences of the V3 loop region of the viral envelope glycoprotein. The canonical approach to sequence-based prediction of coreceptor usage was derived via statistical analysis of a less reliable and significantly smaller data set than is presently available. We aimed to produce a superior phenotypic classifier by applying modern machine learning (ML) techniques to the current database of V3 loop sequences with known phenotype. The trained classifiers along with the sequence data are available for public use at the supplementary website: http://genomiac2.ucsd.edu:8080/wetcat/v3.html

The entry of HIV-1 into a host cell is a two-stage process. First, the viral envelope glycoprotein binds to the cell surface molecule CD4, inducing a conformational change in the gp120 ectodomain of the protein. Second, the glycoprotein docks to a seven-transmembrane chemokine coreceptor on the cell surface, triggering the presentation of its gp41 transmembrane segment. This sequence of events results in membrane fusion and penetration of the virus into the cytosol.[1] The two principal coreceptors used by HIV-1 are CXCR4 and CCR5, members of the CXC and CC chemokine receptor families, respectively.[2]

The particular coreceptor used by a strain of HIV-1 (CXCR4 vs. CCR5) largely defines its replication kinetics and cytopathology *in vitro*. Moreover, coreceptor usage is indicative of the pathogenicity, tissue tropism, and transmissibility of a virus *in vivo*. Unsurprisingly, the determination of this viral phenotype is critical in a wide variety of HIV research contexts.

Several experiments have been conducted on HIV isolates to pinpoint the genetic basis underlying coreceptor preference. The generation and analysis of chimeric (recombinant) viruses have localized the primary determinant of coreceptor usage to the 35-amino acid V3 loop subregion of the HIV envelope glycoprotein.[3]

Earlier work involving statistical analysis of V3 loop amino acid sequences and their respective phenotypes suggested that the presence of a positively charged residue at positions 11 and/or 25 of the V3 loop (numbered according to the North American consensus; see Fig. 1) conferred the ability to dock with CXCR4, while CCR5 binding is the default condition.[4] To date, this "charge rule" is the most accepted method of sequence-based prediction. However, prediction based on this rule does not always align with experimental determination of coreceptor usage.[5] The inaccuracy of the charge rule is most likely due to the comparatively sparse and unreliable data that were available at the time of its creation. Since then, the number of sequences with known phenotype has increased substantially, and the laboratory-based assays used to generate the data have improved. Another possible candidate for a deficiency in this predictive scheme is the consideration of only 2 of the 35 available amino acid positions in the V3 loop.

Modern machine learning (ML) techniques for class prediction can provide advantages over traditional statistics in terms of their abilities to identify and exploit interactions between feature variables. In addition, the rules they generate can often be interpreted with relative ease.[6,7] ML has already proven extremely useful in segregating biological sequence data into

[1]University of California, San Diego, La Jolla, California 92093.
[2]Veterans Administration, San Diego Healthcare System, San Diego, California 92161.
[*]Satish Pillai and Benjamin Good contributed equally to this work.

```
ctrpnnntrksihigpgrafytageiigdirqahc
```

FIG. 1. North American V3 loop sequence with positions 11 and 25 (basis of the "charge rule") underlined.

functional classes.[8,9] We used a variety of ML approaches to develop a better classifier of coreceptor usage and to assess the impact of other V3 loop attributes on viral tropism.

All the V3 loop sequence entries containing documentation of experimentally determined coreceptor usage were downloaded from the Los Alamos National Laboratory HIV Sequence Database, and duplicate sequences were removed. V3 loops less than 34 or more than 36 amino acids in length were deleted from the data set in the interests of producing a relatively gap-free alignment (Table 1). CLUSTAL W[10] was used to generate an automated multiple sequence alignment of the remaining 271 sequences, using the default parameter settings.

Classifiers were trained to make the distinction between viruses capable of using CXCR4 as a coreceptor, versus those that were incapable. Dual-tropic (R5X4) viruses were therefore pooled into the X4 class. Each sample in the initial training set included the amino acid character (or gap) at each of the 40 positions in the V3 loop alignment.

All the experiments in this analysis were conducted using WEKA (the Waikato Environment for Knowledge Analysis), an open source collection of data-processing and machine-learning algorithms.[7] Written in Java, it runs on most platforms and is available for free download at http://www.cs.waikato.ac.nz/ml/Weka/. Of the many techniques for classification included in the WEKA package, we chose to focus on an implementation of the Quinlan C4.5 decision tree inducer called "j48," an algorithm that builds rules from partial decision trees constructed with C4.5, called "PART," and a sequential minimal optimization-based implementation of support vector machines (SVM).[11,12]

One hundred iterations of stratified 10-fold cross-validation were used to evaluate the different classifiers and training set compositions. For each of 100 trials, the data set was randomly divided into 10 groups of approximately equal size and class distribution. For each "fold," the classifier was trained using all but 1 of the 10 groups and then tested on the unseen group. This procedure was repeated for each of the 10 groups. The cross-validation score for 1 trial was the average performance across each of the 10 training runs. The reported score is the average across the 100 trials. The same divisions of the data set were used for each type of classifier (including the change rule) to allow for direct comparison.

In the first trial, we compared the abilities of four classifiers, the charge rule, SVM, C4.5, and PART, to accurately predict the coreceptor usage of HIV-1 V3 loop amino acid sequences.

TABLE 1. COMPOSITION OF DATA SET

R5	*X4*	*R5X4*
168 (62%)	103 (38%)	21 (8%)

TABLE 2. CLASSIFIER PERFORMANCE[a]

Classifier	*Full sequence (%)*	*Positions 11 & 25 removed(%)*
Charge rule	87.45	0
SVM	**90.86**	88.79
C4.5	**89.51**	84.54
PART	**89.37**	85.95

[a]Percent correct for 100 rounds of 10-fold cross-validation. Values in boldface indicate a statistically significant improvement over the charge rule. The default settings from WEKA were used in all cases.

The second trial compared the performance of these classifiers on the same data set but with positions 11 and 25 deleted. To reiterate, positions 11 and 25 of the V3 loop constitute the entire basis of prediction for the canonical charge rule predictor. We eliminated this information from the training data for the second trial in the interests of both informatics and biology; we aimed to assess the capacity of machine learning to unearth novel information content, while concomitantly identifying new areas within the loop that influence HIV coreceptor usage.

The results presented in Table 2 indicate that we can generate a more reliable sequence-based predictor of HIV coreceptor usage by employing a variety of ML techniques. In addition, classifiers trained on sequences lacking positions 11 and 25 produce results competitive to the conventional method and to classifiers constructed using the entire available feature set. Trials conducted with a variety of different sequence attributes resulted in fairly consistent construction of classifiers performing near 90% accuracy in cross-validation trials (data not shown*), suggesting that information regarding coreceptor usage is widely distributed throughout the V3 loop.

Throughout all our trials, the SVM was consistently the best phenotypic classifier. Table 3 summarizes its class-specific performance in cross-validation on the full sequence set.

An obvious benefit of the conventional charge rule is its simplicity. Implementing a simpler method may be desirable, especially if significantly more complex schemes provide only marginally better results. In the interest of succinctness, we constructed decision trees using only two V3 loop attributes. The tree in Fig. 2 was constructed with the two sequence attributes identified as having the highest information content with regard to coreceptor usage. Using only positions 7 and 11 (positions 8 and 12 in our alignment), we were able to automatically construct a theoretical framework that consistently outperformed the charge rule (Fig. 2). Moreover, this rule set had the second highest cross-validation score, 89.97%, and constituted the fourth-best classifier overall by correctly classifying 90.77% of the training set.

It is worth noting that when limiting the search to predictors formed from only two V3 positions, the combination of positions 11 and 25 does not provide the greatest information content. Position 11 is certainly the most predictive, but position

*See supplementary website: http://genomiac2.ucsd.edu:8080/wetcat/v3.html

TABLE 3. CLASS-BASED STATISTICS FOR SUPPORT VECTOR MACHINE ON FULL SEQUENCE SET[a]

Class	*True-positive rate*	*False-positive rate*	*Precision*
CXCR4	0.757	0.024	0.951
CCR5	0.976	0.243	0.868

[a] Precision (predictive power) is equal to the number of true-positive predictions for a class divided by the number of predicted positives.

25 does not contribute much additional information (Table 2). Surprisingly, rules generated using only position 11 result in slightly better classification performance than rules that include position 25 (data not shown*).

A small but significant subset of the isolates within the data set was consistently misclassified by most of the classifiers, even when they were included in the training set. The final step in our analysis concentrated on unearthing a common element between those isolates that were misclassified when using the entire data set for both training and testing. As mentioned earlier, in addition to R5 and X4 strains there is a third phenotypic class of "dual-tropic" R5X4 variants that can utilize either chemokine receptor to enter a target cell. It has been hypothesized that dual-tropic viruses may represent an intermediate evolutionary stage between full-fledged X4 strains and their CCR5-utilizing ancestors.[13] Our classifiers were trained to make the distinction between viruses capable of using CXCR4, dual-tropic viruses included, versus those that were incapable. It is likely that dual-tropic viruses use a different "sequence key" to unlock the X4 door than conventional (monotropic) CXCR4-using strains, because they have retained the capacity to bind CCR5 as well. Therefore, these isolates may be incorrectly labeled as "X4 incapable" by our classification methods, and hence contribute disproportionately to the various error sets.

We investigated this possibility by tallying the monotropic and dual-tropic sequences in each error set and comparing these numbers against the total in each category (250 monotropic and 21 dual-tropic isolates). The data in Table 4 unequivocally demonstrate that dual-tropic variants are significantly overrepresented in all error sets ($p < 0.01$, Fisher exact test), speaking to the biological uniqueness of the R5X4 class, and to the efficacy and sensitivity of the classifiers themselves.

To determine whether there was a consistent sequence pattern associated with this third phenotypic class, each of the classifiers was implemented to classify the data into R5, X4, and R5X4 (dual-tropic) isolates.* However, the classifiers could not satisfactorily perform this task, likely because of the inadequate number of available training cases in the dual-tropic category.

The two primary goals of this project were to create a better classifier of coreceptor usage based on V3 sequence and to identify new biologically meaningful positions within this region. Our results indicate a marginal improvement in performance over the established charge rule, and demonstrate conclusively that positions within V3 other than positions 11 and 25 can be substantially informative in determining HIV phenotype. Furthermore, examination of the linkage between these newly implicated positions and the two relied on by the conventional classifier suggests that they contain novel, independent information pertaining to coreceptor usage.[1]

*To use the SVM, the MultiClassClasifier method was invoked from WEKA.

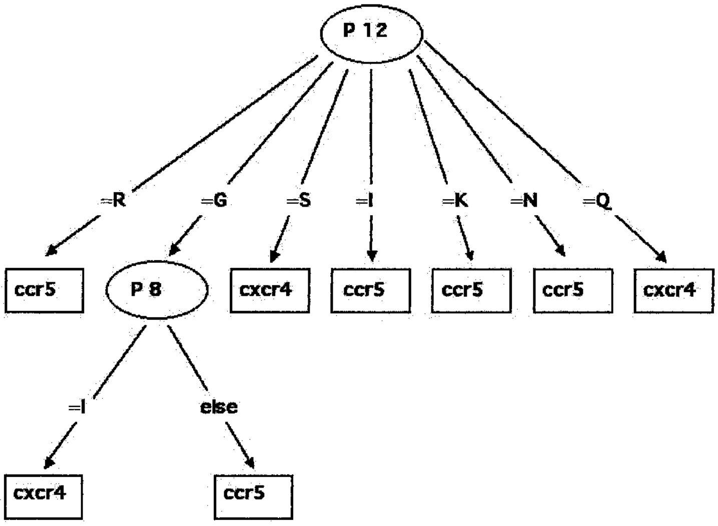

FIG. 2. Decision tree constructed using C4.5 trained on positions 7 and 11 of the V3 sequence (positions 8 and 12 in the alignment).

TABLE 4. MISCLASSIFICATION OF MONOTROPIC AND DUAL-TROPIC SEQUENCES

Classification method	*Monotropics misclassified (%)*	*Dual-tropics misclassified (%)*
Charge rule	7	33
C4.5	6	43
PART	5	38
PART—positions 11 and 25 deleted	6	19
SVM	0	5

A limitation to classifier performance stems from the minor subset of V3 loop sequences that violate the sequence–phenotype relationships exhibited by the vast majority of training cases. We determined that a large proportion of these cases represent a third, biologically distinct phenotypic class of dual-tropic isolates that can utilize either chemokine receptor to enter a target cell. It is likely, therefore, that these errors were reflections of a conflict between the two-way classification task and the tripartite structure of the phenotypic data, rather than a shortcoming in the classifiers themselves.

Our revisitation of the conventional prediction scheme suggests that the charge rule may in fact be somewhat obsolete. Predictions based on position 11 alone were on average more accurate than those based on positions 11 and 25. Considering that the statistical derivation of the charge rule was performed several years ago, it is possible that the inclusion of position 25 reflects a sampling bias in the significantly smaller data set available at that time.

Experiments involving mutagenesis of the HIV-1 envelope glycoprotein have introduced the possibility that positions outside the V3 loop may also influence viral tropism.[14] We would like to apply the same ML techniques to systematically determine how sequence positions within the HIV-1 envelope but outside the V3 loop subregion modulate coreceptor usage. This will depend on the large-scale generation of full-length envelope sequences with corresponding phenotypic data. In addition, as *in vitro* assays become more sophisticated, it should be possible to describe coreceptor usage on a continuous scale, rather than by categorizing the data into discrete, arbitrary classes. This information will allow for a high-resolution map of sequence against phenotype, whereby subtle changes in sequence could be predictive of minor effects on coreceptor preference.

The most significant contributions of this work are the elucidation of predictive V3 positions other than positions 11 and 25, and the demonstration of the power of machine learning for rapid knowledge discovery based on protein sequence. In an age when sequence data are being generated at an astonishing pace, machine learning is an invaluable tool for fluently bridging the gap between genotype and phenotype.

ACKNOWLEDGMENTS

We thank Brian Gaschen at the HIV Sequence Database (Los Alamos, NM) for assistance in consolidating the training data. We also extend our gratitude to Drs. Joe Wong, Simon Frost, and Andrew Leigh-Brown for critical reviews of this manuscript.

This work was supported by the National Institute of Allergy and Infectious Diseases (A146237 and AI47703; J.C.), the Center for AIDS Research Genomics Core Laboratory (AI36214), the Universitywide AIDS Research Program (IS99-SD213 and PH97-SD-201), and the San Diego Veterans Medical Research Foundation, as well as by NIH grants AI27670, AI38858, AI43638, and AI29164 (D.D.R.) and the San Diego Veterans Affairs Healthcare System.

REFERENCES

1. Wyatt R and Sodroski J: The HIV-1 envelope glycoproteins: Fusogens, antigens, and immunogens. Science 1998;280:1884–1888.
2. Fenyö EM, Schuitemaker H, Åsjö B, McKeating J, Sattentau Q, and EC Concerted Action on HIV Variability: The history of HIV-1 biological phenotypes past, present and future. *In*: *Human Retroviruses and AIDS 1997: A Compilation and Analysis of Nucleic Acid and Amino Acid Sequences* (Korber B, Hahn B, Foley B, Mellors JW, Leitner T, Myers G, McCutchan F, and Kuiken CL, eds. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, New Mexico, 1997, pp. III-13–III-18.
3. Cheng-Mayer C, Quiroga M, Tung JW, Dina D, and Levy JA. Viral determinants of human immunodeficiency virus type 1 T-cell or macrophage tropism, cytopathogenicity, and CD4 antigen modulation. J Virol 1990;64:4390–4398.
4. Fouchier RAM, Groenink M, Kootstra NA, Tersmette M, Huisman HG, Miedema F, and Schuitemaker H. Phenotype-associated sequence variation in the third variable domain of the human immunodeficiency virus type 1 gp120 molecule. J Virol 1992;66: 3183–3187.
5. McDonald RA, Chang G, and Michael NL. Relationship between V3 genotype, biologic phenotype, tropism, and coreceptor use for primary isolates of human immunodeficiency virus type 1. J Hum Virol 2001;4:179–187.
6. Mjolsness E and DeCoste D. Machine learning for science: State of the art and future prospects. Science 2001;293:2051–2055.
7. Witten IH and Frank E. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations.* Morgan Kaufmann, San Francisco, 2000.
8. Hua S and Sun Z. Support vector machine approach for protein subcellular localization prediction. Bioinformatics 2001;17: 721–728.
9. Resch W, Hoffman N, and Swanstrom R: Improved success phenotype prediction of the human immunodeficiency virus type 1

from envelope variable loop 3 sequence using neural networks. Virology 2001;288:51–62.

10. Thompson JD, Higgins DG, and Gibson TJ. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22:4673–4680.
11. Quinlan JR. *C4.5: Programs for Machine Learning.* Morgan Kaufmann, San Francisco, 1993.
12. Vapnik V. *The Nature of Statistical Learning Theory.* Springer-Verlag, New York, 1995.
13. Lu Z, Berson JF, Chen Y, Turner JD, Zhang T, Sharron M, Jenks MH, Wang Z, Kim J, Rucker J, Hoxie JA, Peiper SC, and Doms RW. Evolution of HIV-1 coreceptor usage through interactions with distinct CCR5 and CXCR4 domains. Proc Natl Acad Sci USA 1997;94:6426–6431.
14. Rizzuto CD, Wyatt R, Hernandez-Ramos N, Sun Y, Kwong PD, Hendrickson WA, and Sodroski J. A conserved HIV gp120 glycoprotein structure involved in chemokine receptor binding. Science 1998;280:1949–1953.

Address reprint requests to:
Satish Pillai
University of California, San Diego
9500 Gilman Drive
Stein Clinical Research Bldg. Room 327
Mail Code 0679
La Jolla, California 92093

E-mail: satish@biomail.ucsd.edu