Virology **288,** 51–62 (2001)
doi:10.1006/viro.2001.1087, available online at http://www.idealibrary.com on IDEAL®

Improved Success of Phenotype Prediction of the Human Immunodeficiency Virus Type 1 from Envelope Variable Loop 3 Sequence Using Neural Networks

Wolfgang Resch,*[,1] Noah Hoffman,†[,1] and Ronald Swanstrom*[,†,‡,2]

Department of Biochemistry and Biophysics, †Department of Microbiology and Immunology, and ‡UNC Center for AIDS Research, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599-7295

Received March 1, 2001; returned to author for revision June 15, 2001; accepted July 12, 2001

We have assembled two sets of HIV-1 V3 sequences with defined epidemiologic relationships associated with experimentally determined coreceptor usage or MT-2 cell tropism. These data sets were used for three purposes. First, they were employed to test existing methods for predicting coreceptor usage and MT-2 cell tropism. Of these methods, the presence of one basic amino acid at position 11 or 25 proved to be most reliable for both phenotypic classifications, although its predictive power for the X4 phenotype was less than 50%. Second, we used the sequence sets to train neural networks to infer coreceptor usage from V3 genotype with better success than the best available motif-based method, and with a predictive power equal to that of the best motif-based method for MT-2 cell tropism. Third, we used the sequence sets to reexamine patterns of variability associated with the different phenotypes, and we showed that the phenotype-associated sequence patterns could be reproduced from large sets of V3 sequences using phenotypes predicted by the trained neural network. © 2001 Academic Press

Key Words: HIV-1; V3; neural network; coreceptor; MT-2 cell tropism; phenotype; Env.

INTRODUCTION

Human immunodeficiency virus type 1 (HIV-1) isolates can be classified phenotypically according to their ability to replicate and induce syncytia in MT-2 cells, a transformed T-cell line (Tersmette *et al.,* 1988), or according to the primary coreceptor used to enter cells (Alkhatib *et al.,* 1996; Choe *et al.,* 1996; Deng *et al.,* 1996; Doranz *et al.,* 1996; Dragic *et al.,* 1996; Feng *et al.,* 1996). In the former case, the strains are termed non-syncytium-inducing (NSI) and syncytium-inducing (SI); in the latter case they are referred to as CXCR4-using (X4) and CCR5-using (R5). These biological phenotypes are related (Bjorndal *et al.,* 1997), though the extent of the correlation between the classifications and the biological significance of any discordance remain poorly defined. Primary infection is almost always limited to virus strains of R5/NSI phenotype (Roos *et al.,* 1992; Zhu *et al.,* 1993), and X4/SI variants generally evolve *de novo* during the chronic infection with HIV-1. The evolution of X4/SI variants coincides with a two- to fourfold accelerated loss of $CD4^+$ T cells (Koot *et al.,* 1993; Schellekens *et al.,* 1992; Tersmette *et al.,* 1989), which in turn leads to an increased probability of disease progression (Koot *et al.,* 1993; Richman and Bozzette, 1994).

The third variable region (V3) of the gp120 subunit of the HIV-1 envelope protein (Env), a 35 amino acid loop constrained by a disulfide bond, has been identified as a major determinant of coreceptor usage and MT-2 cell tropism (Chesebro *et al.,* 1991; Choe *et al.,* 1996; Cocchi *et al.,* 1996; Hwang *et al.,* 1991; Takeuchi *et al.,* 1991). A large body of work has led to the identification of putative sequence motifs distinguishing the X4/SI phenotypes from the NSI/R5 phenotypes (see references in Table 1). Rules derived from these motifs that have been employed to predict phenotypes are summarized in Table 1. These rules, particularly the presence of a positively charged amino acid at V3 position 11 or 25, have been used to predict phenotype from the genotype of a virus strain. However, the reliability of phenotype prediction using these rules, especially for coreceptor usage, has not been tested with large sets of epidemiologically unrelated sequences. A reliable prediction algorithm would be desirable for two reasons. First, a sequence-based method of phenotype prediction as an alternative to the more difficult experimental determination of phenotypes would be beneficial in the clinical setting. Second, the large number of existing HIV-1 sequences available in public databases could be classified by V3 genotype to discern potential linkage between X4-associated changes in V3 and positions elsewhere in gp120.

As an alternative to the design of motif-based rules for

[1] These authors contributed equally to this work.

[2] To whom correspondence and reprint requests should be addressed at UNC Center for AIDS Research, Lineberger Bldg., Rm 22-059, CB 7295, University of North Carolina, Chapel Hill, NC 27599-7295. Fax: +1-(919)-966-8212. E-mail: risunc@med.unc.edu.

0042-6822/01 $35.00
Copyright © 2001 by Academic Press
All rights of reproduction in any form reserved.

TABLE 1

Description of Sequence Motif-Based Rules Used to Predict Phenotypes from V3 Sequences.

Name	Ref.	x4/si	r5/nsi	f
Method 1	(1, 2)	Basic at 11 or 25	Not x4	1
Method 2	(3)	Not r5	S or G at 11 and D or E at 25 and GPG at 15–17	1
Method 3	(4)	Basic at one or more of 11, 13, 19, 23, 24, or 32 and no acidic at 25	Not x4	1
Method 4	(3, 4)	Basic at one or more of 11, 13, 19, 23, 24, or 32 and no acidic at 25 (Method 3)	S or G at 11 and D or E at 25 and GPG at 15–17 (Method 2)	<1
Method 5	(5)	Net Charge ≥ +5	Net Charge < +5	1

Note. Numbers refer to V3 amino acid positions starting with the first cysteine, letters represent the single letter amino acid code. Method 4 excludes approximately 30% of all sequences that are predicted to have different phenotypes by the two rules. Dual-tropic virus strains (R5X4) are considered to be X4 for the purposes of this analysis. Ref. 1, de Jong *et al.,* 1992; 2, Fouchier *et al.,* 1995; 3, Xiao *et al.,* 1998; 4, Milich *et al.,* 1997; 5, Fouchier *et al.,* 1992. r5/nsi, x4/si, rules defining the r5/nsi and x4/si phenotypes, respectively. f, fraction of sequences assigned an unequivocal phenotype by the prediction algorithm.

phenotype prediction, we generated neural networks to predict coreceptor usage or MT-2 cell tropism from the amino acid sequence using a subset of positions in V3. Neural networks are well suited for the detection of complex patterns in the input data and do not require any mechanistic assumptions about the nature of the link between sequence and phenotype. To implement such a neural network and to verify the reliability of the existing phenotype prediction algorithms, we assembled two distinct sets of sequences, one with known MT-2 cell tropism (NSI/SI set), and the other with known coreceptor usage (R5/X4 set). These sets include information about the epidemiologic relatedness of all sequences and thus represent an improvement over sequence sets described previously. Using these sequences, we determined that the presence of a positively charged amino acid at positions 11 or 25 is the most reliable predictor of the previously defined sequence motif-based rules for both phenotypes. However, the reliability of this rule for predicting the X4 phenotype was estimated here to be below 50%. We have generated a neural network that has improved reliability of coreceptor usage prediction, as well as a separate network that performs as well as the 11/25 rule for the prediction of the NSI/SI phenotype. Moreover, we present a more detailed statistical analysis of V3 sequence determinants associated with the two phenotypes.

RESULTS AND DISCUSSION

Construction of V3 sequence sets with experimentally determined coreceptor usage or MT-2 cell tropism and known epidemiologic relatedness

To test the performance of existing rules for the inference of phenotypes from V3 sequences and to generate neural networks for phenotype prediction, we assembled two sequence sets, one with known coreceptor usage (R5/X4 sequence set), and the other with known MT-2 cell tropism (NSI/SI sequence set). The ability to employ the CXCR4 coreceptor was treated as sufficient for the X4 phenotype (i.e., dual-tropic viruses are classified as X4). For both sets, the epidemiologic relationships among the sequences were determined. The sources and phenotypes of all unique sequences are summarized in Table 2. The R5/X4 set contains both published and unpublished sequences. For the unpublished sequences, we obtained descriptions of the phenotypes and the source of the indicated sequences. The sequence sets represent the largest collection of unrelated sequences with known phenotypes described to date, although the paucity of X4 and SI sequences is still a limitation. In this discussion we will use uppercase R5/X4/NSI/SI to refer to experimentally determined phenotypes, while lowercase letters (e.g., r5, nsi) indicate phenotypes inferred from V3 sequence.

To compare groups of sequences with predicted phenotypes to groups with experimentally determined phenotypes and to estimate the frequency of x4-associated sequence, we also assembled a database containing 1997 unique V3 peptide sequences isolated from different individuals. From this set we estimated the frequency of X4- or SI-associated sequences to be approximately 0.15 using all classification methods available to us; this value is used as an estimate for the frequency of X4-associated sequences in the calculations of reliability.

The presence of a positively charged amino acid at position 11 or 25 (method 1) is the best available motif-based predictor of coreceptor usage and MT-2 cell tropism

We first determined the performance of the previously used rules (Table 1) for predicting R5/X4 or NSI/SI phenotypes. To test these rules, sequences were chosen at random from each of the sequence sets such that no two sequences associated with the same phenotype were derived from the same patient, and the predicted phenotypes of these sequences were compared to the experimental phenotypes. This strategy minimizes the confounding effects of epidemiologic relatedness among the test sequences. To fully sample the set of sequences,

TABLE 2

Source and Associated Phenotype of Sequence Sets

(A) Unique R5/X4 sequences collected

Sequences				
total	X4	R5	Patients	Reference
83	31	52	59	(1)
17	0	17	13	(2)
11	4	7	3	(3)
6	6	0	3	(4), (5)
11	3	8	11	(6)
31	2	29	31	(7)
57	1	56	57	(8)
216	47	169	177	
UnrelatedR5/X4sequences				
181	31	150	177	

(B) Unique NSI/SI sequence set

Sequences				
total	SI	NSI	Patients	Reference
10	2	8	5	(1)
11	3	8	8	(9)
28	12	16	22	(10)
199	48	151	63	(11)
9	5	4	8	(12)
257	70	187	106	
UnrelatedNSI/SIsequences				
124	40	84	106	

Note. 1, GenBank; 2, Chan *et al.,* 1999; 3, Singh *et al.,* 1999; 4, Nelson *et al.,* 2000; 5, Michael *et al.,* 1998; 6, Lathey *et al.,* 2000; 7, Li *et al.,* 1999; 8, Zhang *et al.,* 1998; 9, Bhattacharyya *et al.,* 1996; 10, Cornelissen *et al.,* 1995; 11, Fouchier *et al.,* 1995; 12, Fouchier *et al.,* 1992.

this analysis was repeated 100 times to include each sequence in the test set at least once. The average 2 × 2 contingency table summarizing the performance of the five sequence rules is shown in Table 3, with several corresponding measures of performance. Here, the specificity of the rules is the most important measure of performance, as it represents the fraction of R5 or NSI sequences that are misclassified as X4 or SI. Since X4 or SI sequences are underrepresented in the set of all published sequences, even small deviations of specificity from 1 can lead to a considerable contamination of the predicted x4 or si sequence set with R5 or NSI sequences. In addition, the phi coefficient is informative as a measure of correlation between the algorithm results and the experimental phenotypes.

For the R5/X4 set, methods 1 and 2 display the greatest specificity, although method 1 is the more sensitive test (Table 3A). If specificity is considered the most important of the performance measures, method 1, using the presence of a positively charged amino acid at either position 11 or 25, should be considered the best sequence-motif-based method currently available for the determination of coreceptor usage. Note, however, that this method has an estimated reliability of X4 prediction of only 0.48 based on the estimated frequency of 0.15 of X4 associated sequences in the set of all available sequences. Method 1 also outperforms all other sequence rules for NSI/SI phenotype prediction as measured by both specificity and the phi coefficient (Table 3B), with a reliability of SI phenotype prediction of approximately 0.85. Although method 4 performs well for both phenotypes, it fails to classify 30% of all sequences and is thus a poor choice for phenotype prediction. Method 5, which requires a positive charge of at least 5 for x4 or si classification, shows a phi association coefficient of 0.77 with the NSI/SI phenotype. It should be noted that the overall performance of all sequence rules is lower for the R5/X4 sequence set than for the NSI/SI sequences.

The overall reliability of all sequence motif-based methods for phenotype inference, especially for coreceptor usage prediction, was limited. This indicates that simple sequence motifs encompassing a few amino acid positions in V3 may not be sufficient to determine coreceptor usage reliably, although there was a strong association between the accumulation of positive charges, particularly at positions 11 and 25, and MT-2 cell tropism. The differential performance of all five methods applied to the two data sets is an indication that the two phenotypes may not be synonymous. We could not formally exclude the possibility that selection bias contributed to the observed differences between the two phenotypic classifications, especially because of the small number of SI and X4 sequences. Nevertheless, we have chosen to treat these two phenotypes as only partially overlapping and analyze them separately. In the following analyses we selected method 1 as the standard to which we compared the performance of the neural networks.

Neural networks infer coreceptor usage from V3 sequence with greater reliability than sequence motifs

To devise a more reliable method for the prediction of coreceptor usage and the NSI/SI phenotype, we trained and tested neural networks on each of the two sequence sets. Neural networks were generated and trained using epidemiologically unrelated subsets of sequences randomly chosen from the sequence sets as described in the previous section. Training was done on half of the randomly chosen epidemiologically unrelated sequences, and performance was assessed on the other half. Thus, the set of test sequences did not overlap the sequences presented to the network during training.

TABLE 3

Performance of Sequence Motif-Based Phenotype Prediction Methods for the R5/X4 and the NSI/SI Sequence Sets

(A) R5/X4 sequence set

	2 × 2 Contingency table							
Rule		R5	X4	f	Sensitivity	Specificity	ϕ	X4 reliability
Method 1	r5	135	14	1	0.53	0.90	0.43	0.48
	x4	15	17					
Method 2	r5	135	17	1	0.45	0.90	0.36	0.44
	x4	15	14					
Method 3	r5	92	6	1	0.81	0.61	0.32	0.27
	x4	58	25					
Method 4	r5	92	6	0.7	0.71	0.86	0.49	0.47
	x4	15	14					
Method 5	r5	130	10	1	0.66	0.87	0.48	0.47
	x4	20	21					

(B) NSI/SI Sequence Set

	2 × 2 Contingency table							
Rule		NSI	SI	f	Sensitivity	Specificity	ϕ	SI reliability
Method 1	nsi	82	6	1	0.84	0.97	0.84	0.85
	si	2	34					
Method 2	nsi	79	17	1	0.58	0.94	0.58	0.63
	si	5	23					
Method 3	nsi	54	1	1	0.96	0.65	0.57	0.32
	si	30	39					
Method 4	nsi	54	1	0.7	0.94	0.92	0.83	0.67
	si	5	23					
Method 5	nsi	78	7	1	0.83	0.93	0.77	0.68
	si	6	33					

Note. Lower- and upper-case letters indicate predicted and experimentally determined phenotypes, respectively. Note the general trend to better performance for the NSI/SI sequences. Note also that the specificity is a major determinant for performance due to the abundance of R5 or NSI sequences. In the contingency table the underlined numbers indicate correct predictions; in the other columns underlined numbers indicate the best scores. f, fraction of sequences for which the method could provide an unambiguous classification; sensitivity, fraction of correctly classified X4 sequences; specificity, fraction of correctly classified R5 sequences; ϕ, phi coefficient of association; reliability, probability of a sequence with predicted x4 or si phenotype to actually be X4 or SI, respectively.

Some R5 sequences were discarded from the R5/X4 data set prior to training to increase the proportion of X4 sequences. Several networks were trained and the network with the best performance on the test set was chosen for further analysis. Performance of the best networks for both phenotypes was compared to the performance of method 1 on identical sequence subsets (Table 4). The best networks were termed "R5/X4 network" and "NSI/SI network" and are available upon request.

TABLE 4

Performance of Two Neural Networks on Their Respective Test and Training Sets Compared with the Performance of Method 1, Which Uses the Presence of a Basic Amino Acid at Either Position 11 or 25 of V3 to Classify X4 and SI Sequences

	R5/X4 Sequence Set								
	Training Set			Test Set			Overall		
Algorithm	Sens	Spec	ϕ	Sens	Spec	ϕ	Sens	Spec	ϕ
Method 1	0.60	0.89	0.51	0.53	0.87	0.42	0.53	0.90	0.43
R5/X4 network	0.80	0.98	0.82	0.80	0.89	0.67	0.75	0.94	0.68
	NSI/SI sequence set								
Method 1	0.85	1.00	0.89	0.85	0.90	0.75	0.84	0.97	0.84
NSI/SI network	0.95	1.00	0.96	0.95	0.93	0.89	0.90	0.96	0.87

Note. Sens, sensitivity; spec, specificity; ϕ, phi coefficient of association. Note that the critical parameters are the phi coefficient and the specificity of the prediction algorithm. Overall performance is measured as the mean of $n = 100$ subsets of unrelated sequences drawn at random from the respective data sets.

The best neural networks had greater specificities and phi coefficients than method 1 on training and test sets for both coreceptor usage and NSI/SI phenotypes (Table 4). The improvement over method 1 for the NSI/SI set, however, was marginal. Note that the performance of the networks on the test set was lower than the performance on the training set for networks trained on the R5/X4 sequences. This is most likely due to the limited number of available sequences and the greater relative abundance of R5 sequences. The mean reliability for X4 prediction of the R5/X4 neural network was 0.69 for 100 subsets of unrelated sequences, a considerable improvement over the reliability of 0.48 achieved by method 1. The mean reliability of the NSI/SI neural network was 0.80, comparable to the 0.84 achieved by the 11/25 motif of method 1. The lower performance of the networks on the test sets compared to the training sets was due to a reduction in specificity. Networks trained to infer coreceptor usage have a reduced, though still significant, ability to infer the NSI/SI phenotype and vice versa (data not shown).

We can conclude that the neural network approach achieves a reliability similar to method 1 for prediction of NSI/SI phenotypes, indicating that MT-2 cell tropism may be largely determined by changes at positions 11 and 25 in the V3 region of HIV-1 Env. However, the prediction of coreceptor usage from V3 sequence can be enhanced considerably by a neural network, suggesting that the sequence patterns underlying this phenotype may be more complex than the patterns characterizing the NSI/SI phenotype. The predictive power of the network for the X4 phenotype still remains below 0.7. This could

2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
T (87.6)	**R** (99.5)	**P** (97.9)	**N** (81.5)	**N** (97.1)	**N** (96.4)	**T** (96.9)	**R** (94.0)	**K** (81.4)	**S** (72.2)	**I** (95.4)	**H** (58.4)	**I** (69.6)	**G** (96.3)	**P** (91.3)	**G** (98.5)	**R** (77.3)
I (9.0)		L (1.4)	S (9.4)	K (0.8)	Y (0.8)	I (1.7)	S (2.3)	R (16.1)	G (19.6)	V (2.8)	P (16.2)	L (15.7)	A (2.8)	W (3.3)	R (0.6)	K (7.6)
V (1.2)			G (5.1)	T (0.7)	K (0.8)	K (0.6)	I (1.1)	Q (0.7)	R (6.3)	L (0.7)	N (8.7)	M (12.2)		L (2.0)		Q (6.6)
S (0.9)			H (2.0)		T (0.7)		K (1.0)	T (0.6)		M (0.6)	T (5.4)	V (1.3)		Q (1.4)		S (4.1)
A (0.5)			T (0.9)								S (5.0)			F (0.5)		G (3.5)
											R (2.5)					
											Y (1.7)					
											Q (1.2)					

19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
A (88.2)	**F** (74.2)	**Y** (89.5)	**T** (51.8)	**T** (93.3)	**G** (87.0)	**E** (34.9)	**I** (95.2)	**I** (88.5)	**G** (98.9)	**D** (84.9)	**I** (96.9)	**R** (98.2)	**Q** (87.7)	**A** (99.0)	**H** (89.4)
T (5.2)	W (13.1)	F (4.8)	A (46.0)	A (2.1)	g (5.8)	D (26.5)	V (3.4)	V (4.6)		N (12.8)	M (0.7)	K (1.2)	K (7.1)		Y (8.8)
V (3.9)	L (5.7)	H (3.0)	G (0.6)	g (1.5)	E (2.6)	Q (16.6)		T (4.0)			V (0.7)		R (3.1)		Q (0.6)
S (1.7)	I (2.9)	V (1.1)		I (0.8)	R (1.5)	R (5.1)		g (1.4)			T (0.5)		L (0.7)		
	V (2.2)			R (0.7)	K (1.0)	A (4.5)							H (0.5)		
	Y (1.0)			S (0.7)	D (0.8)	K (3.8)									
	M (0.6)				N (0.6)	G (3.3)									
						N (1.7)									
						g (1.4)									
						T (0.9)									
						S (0.6)									

FIG. 1. Distribution of amino acids at positions 2 to 34 in HIV-1 clade B V3. The sequence set used for this analysis contained $n = 1997$ epidemiologically unrelated sequences obtained from the Los Alamos HIV Sequence Database. Numbers denote the percentage of sequences containing the corresponding amino acid. Only amino acids that occurred at least 10 times (0.5% of all sequences) are shown. Residue numbers are shown above the line. Uppercase letters are single letter amino acid codes; g, gap.

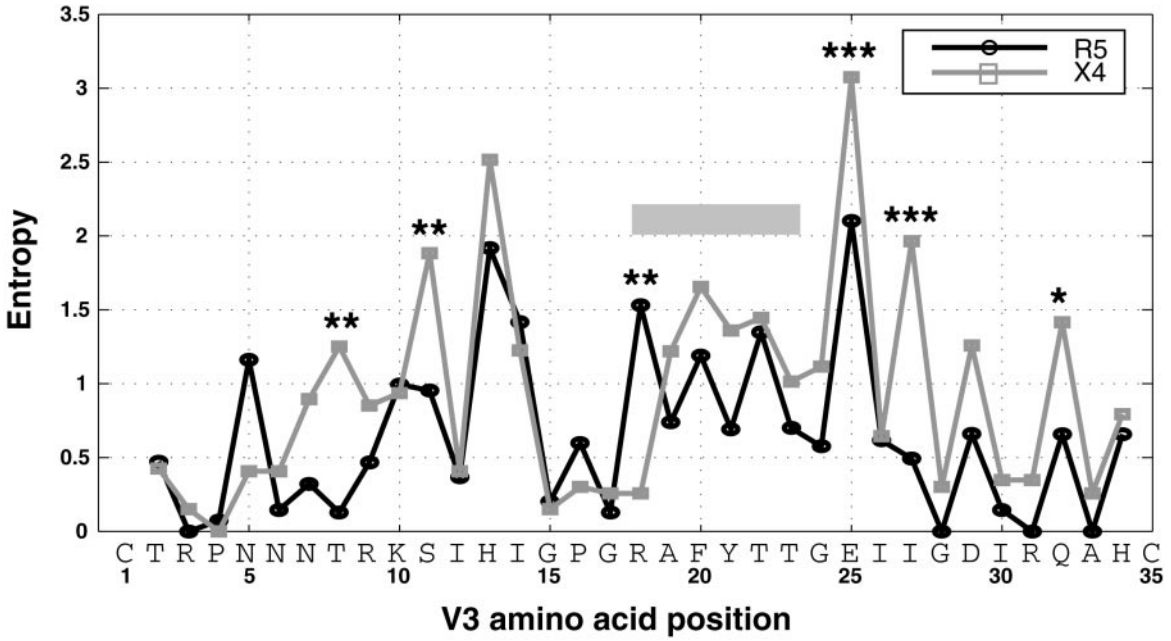

FIG. 2. Comparison of sequence entropy at positions 2 through 34 of V3 for the R5/X4 sequence set divided into X4 and R5 sequences. Note the striking change of pattern between positions 18 and 23 (indicated by a gray bar). Other significant changes were detected at positions 8, 11, 25, 27, and 32. The HIV-1 clade B consensus sequence of the V3 region is shown below the graph. The overall entropy is higher within the X4 sequences than within the R5 sequences. *$P \leq 0.01$; **$P \leq 0.005$; ***$P < 0.001$.

be due either to an insufficient number of sequences used to train the neural network or the existence of coreceptor usage determinants outside of V3. The latter possibility is supported by the reported correlations between other regions of Env with coreceptor usage, in particular, the V1/V2 and C4 regions (Andeweg *et al.,* 1993; Carrillo and Ratner, 1996; Cornelissen *et al.,* 1995; Groenink *et al.,* 1993; Ly and Stamatatos, 2000; Milich *et al.,* 1997; Shieh *et al.,* 2000; Sullivan *et al.,* 1993). The importance of regions outside of V3 is also suggested by the identification of identical V3 sequences found in viruses with different phenotypes (R. Collman, personal communication; Groenink *et al.,* 1992; Singh *et al.,* 1999; Singh and Collman, 2000).

Distinctive patterns of sequence entropy are found among sequences associated with R5 or X4 phenotypes and can be reproduced in larger data sets using phenotype prediction

Env regions classically described as "variable" contain both positions that are highly conserved and those that experience frequent amino acid substitution (for example, see Fig. 1 and Yamaguchi-Kabata and Gojobori, 2000). In addition it has previously been shown that the overall variability is lower in NSI/R5 sequences than SI/X4 sequences (Chesebro *et al.,* 1992; Milich *et al.,* 1993). We used sequences of known phenotype both to reexamine variability at each position of V3 and to show that phenotyping algorithms can reproduce similar patterns using V3 sequences retrieved from GenBank.

Entropy was used as a measure of the position-by-position variability of the complete R5/X4 data set ($n = 216$). The data set was divided into R5 and X4 sequences, and the entropy of the sequence alignment was determined for every position in V3 among sequences belonging to each of the two phenotypes (Fig. 2). Permutation tests were used to determine the significance of the difference in entropy between the sequences associated with the two phenotypes at each position. On average, R5 sequences had lower entropy at most V3 positions. Unexpectedly, position 18 was significantly less variable among X4 sequences, suggesting a functional role for this position in CXCR4 usage. Another difference between the two phenotypes is the disappearance of a pattern of alternating high and low entropies at positions 18 through 23 among X4 sequences. The increase in entropy in X4 sequences is greatest for positions 8, 11, 25, 32, and 27. From our analysis it is unclear whether the observed differences in entropy between the two phenotypes are directly linked to the coreceptor usage change or are confounded by the changes in the host environment accompanying the phenotypic switch.

To determine whether the phenotype prediction algorithms would reproduce the entropy patterns characteristic for each phenotype, 1997 epidemiologically unlinked sequences retrieved from GenBank were divided

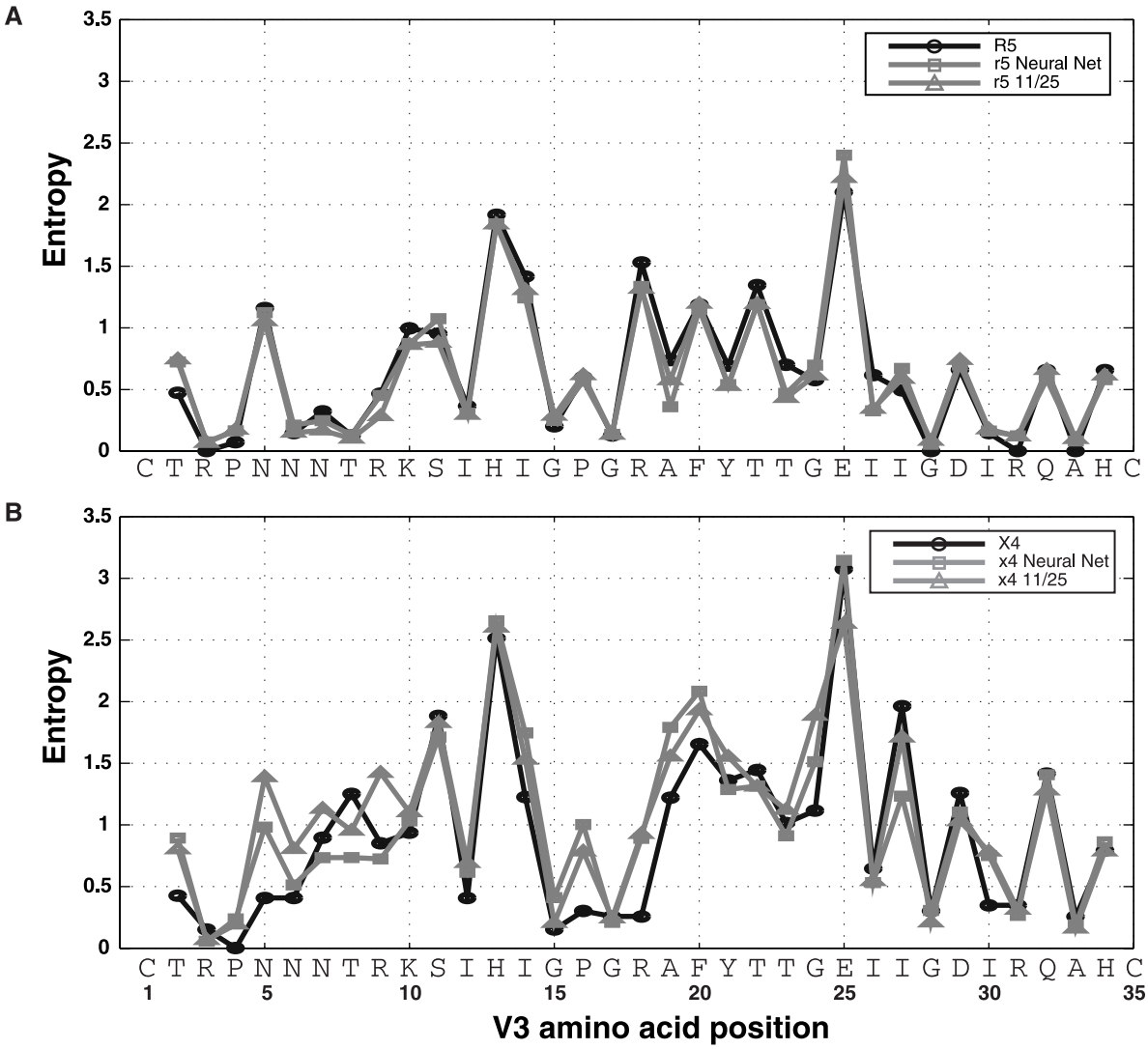

FIG. 3. Reproduction of entropy patterns in V3 sequences by the R5/X4 neural network and sequence motif-based method 1. There were $n = 1997$ sequences with predicted phenotypes. Of these, 362 were predicted to be x4 by the neural net and 298 sequences were classified x4 by method 1 (11/25). Note that lowercase letters indicate inferred phenotypes and uppercase letters indicate experimental phenotypes. (A) R5 and r5 sequences. (B) X4 and x4 sequences. Below the graphs the consensus sequence for HIV-1 clade B V3 is shown. At several positions both algorithms deviate from the entropy calculated for the coreceptor sequence set in a similar manner. Note that both algorithms show the same shift of pattern for the different phenotypes between positions 18 and 23.

into groups by their predicted phenotypes and analyzed as described above. We used the sequence motif-based method 1 and the R5/X4 neural network for coreceptor usage prediction. In Fig. 3, the entropies of the sequences with the predicted phenotypes are shown in comparison with the entropies measured in the R5/X4 sequence set. Overall, the entropy pattern observed among sequences of either phenotype is reproduced well by both algorithms. Thus, both phenotype prediction algorithms, despite their performance differences and their different approaches, generate similar entropy patterns. Note that the two different algorithms appear to result in similar deviations from the entropy of sequences of known phenotype (e.g., at position 5). This deviation points to a possible selection bias in the small set of sequences known to use the CXCR4 coreceptor.

The R5 to X4 and NSI to SI switches are associated with distinct but similar changes in amino acid composition at several positions in V3

Numerous studies have reported linkage between viral phenotype and changes in the amino acid composition at specific positions in V3 (Carrillo and Ratner, 1996; de Jong *et al.,* 1992; Hung *et al.,* 1999; Milich *et al.,* 1997; Shioda *et al.,* 1992). However, either the relatedness, limited number of sequences, or the unavailability of experimentally determined phenotypes limited the power of those studies to describe the significance of the re-

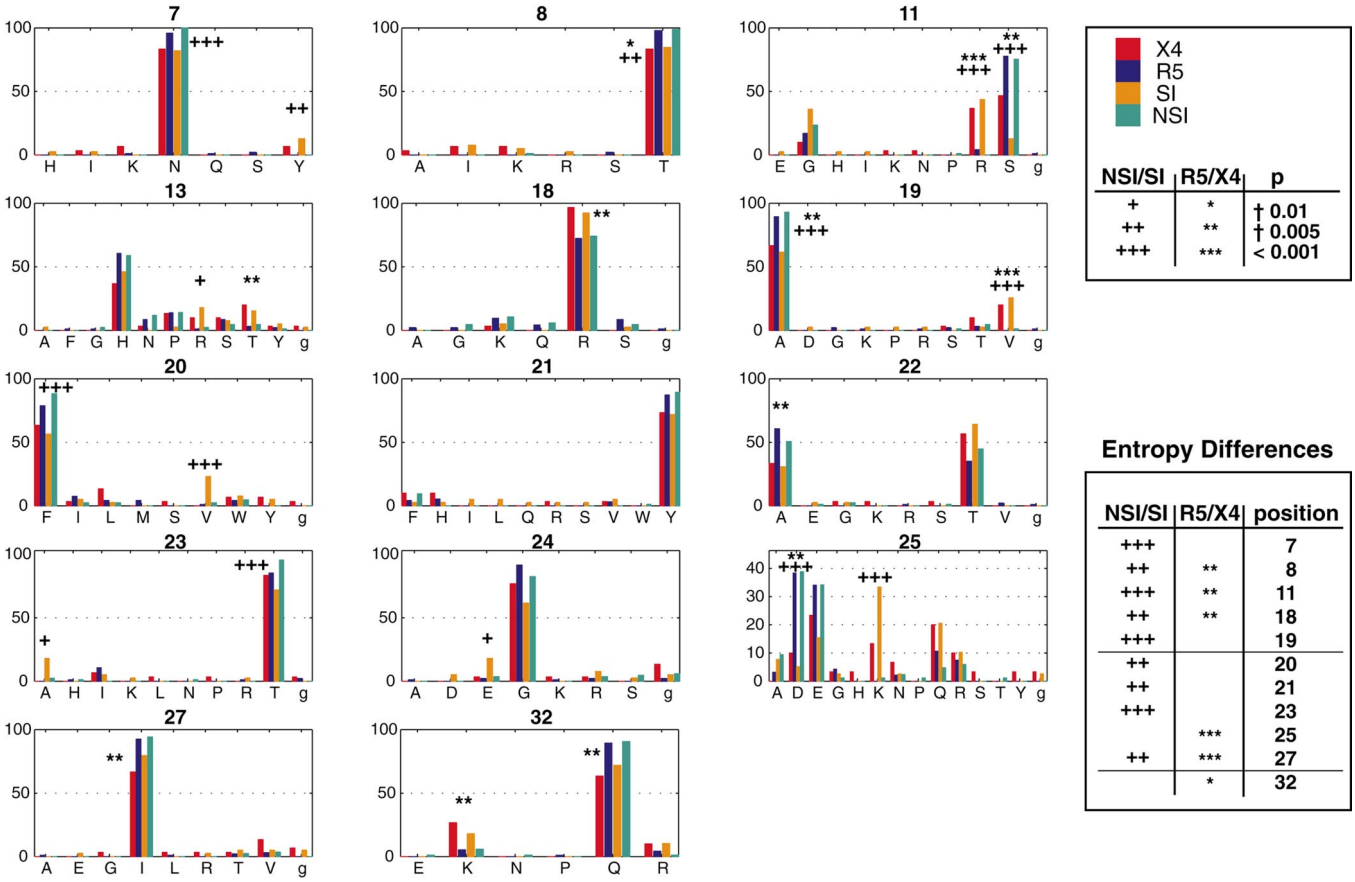

FIG. 4. Characteristic differences in amino acid composition associated with the R5/X4 and NSI/SI phenotypes. A single set of unrelated sequences was chosen at random from both the R5/X4 and the NSI/SI sequence sets and divided according to phenotype. The amino acid composition at positions that showed at least one significant difference between one of the phenotype pairs for either the distribution of an amino acid or the entropy are shown here. The significance level of the differences is indicated close to the bars that display the difference. Significance levels were determined by a permutation test. Differences assigned a significance level of $P < 0.001$ fell outside the support of the reference distribution for the permutation test. Note that the amino acid composition at position 25 is shown at a smaller scale. A–Y: amino acids; g: gap.

ported correlations. Therefore we reexamined the relationship between phenotype and amino acid composition using both the R5/X4 and the NSI/SI data sets. To minimize the confounding effects of epidemiologic relatedness, a single subset of unrelated sequences from each of the R5/X4 ($n = 124$) and NSI/SI ($n = 124$) data sets was analyzed to examine differences in amino acid composition between the X4/R5 and NSI/SI phenotypes, respectively.

The statistical significance of the phenotype-associated differences in the frequencies of amino acids at each position was determined using a permutation test, whereby the difference in abundance of each amino acid at each position was compared to the differences observed in 1000 random partitions of the same data set without respect to phenotype. The probability P that the observed difference is due to chance can be estimated as the number of times the same or a more extreme difference was observed in the random sequence partitions divided by the number $n = 1000$ of random partitions.

Figure 4 summarizes all V3 positions that showed a statistically significant ($P < 0.01$) difference for one of the phenotype pairs either in at least one amino acid or in the entropy of the sequence groups (see previous section). As expected, amino acids at positions 11 and 25 were present at significantly different proportions in each of the phenotype pairs. At position 11, R, but not K, was significantly enriched among SI and X4 sequences. At position 25, D was significantly underrepresented in SI and X4 sequences, whereas K was overrepresented. These observations indicate that the acidic (D and E) and basic (R and K) residues, respectively, are not entirely equivalent at these positions. The generally lower abundance of R and K at positions 11 and 25, respectively, in X4 sequences compared to SI sequences also helps to explain the reduced reliability of the 11/25 rule for coreceptor usage prediction.

Several other changes were either restricted to, or much more apparent in, a single data set. The frequency of 11S was higher among X4 sequences (41%) than SI sequences (13%). Significant shifts from F to V at position

20, from G to E at 24, and a trend from T to A at 23 were apparent only in the NSI/SI data set. Overall, the pattern changes accompanying a switch from R5 to X4 or NSI to SI are similar, though not identical (see below). This is further evidence for the disparity between the two phenotypic classifications. The less dramatic nature of changes accompanying the R5 to X4 switch compared to the NSI to SI switch in these data sets raises the possibility that the X4 phenotype could include the early events in a phenotypic change pathway in which the SI phenotype represents a late stage. Alternatively, this observation could be explained by the inclusion of V3 sequences associated with the ability to use both coreceptors in the X4 set. These sequences may fall along a continuum from more X4-like to more R5-like and the resulting contamination of the X4 sequence set may therefore obscure greater differences between the pure X4 and R5 sequences.

A common feature among several positions differing in composition between phenotypes was a shift away from the majority amino acid without a change in consensus among both X4 and SI sequences. This underrepresentation of the consensus amino acid was significant at a level of $P < 0.01$ in both the R5/X4 and the NSI/SI data sets for three positions (8, 11, and 19). Significance was achieved at other positions within either the R5/X4 data set alone (13, 27, 32), or the NSI/SI data set alone (7, 20, 24), although the overall trend at each of these positions was similar in both data sets. The opposite trend, a shift away from the consensus amino acid among R5/NSI sequences, was observed at position 18, though this trend achieved significance only in the R5/X4 data set. An actual change in the majority amino acid was observed at two positions. The consensus at position 22 changed from A among R5 and NSI sequences to T among X4 and SI sequences, though again the shift was only significant in the coreceptor set. At position 25, the consensus amino acid changed from D to K in the NSI/SI data set. Although this trend was present in the R5/X4 set, the magnitude of the change was smaller and did not result in a change in consensus.

The characteristic differences of amino acid composition between the R5 and X4 phenotype can also be employed as another indirect measure for the performance of the R5/X4 neural network. To this end, we randomly selected sequences with inferred x4 or r5 phenotypes from a large set of unrelated sequences obtained from GenBank and compared amino acid composition at each V3 position in these sets to sequences with experimentally determined phenotypes. This analysis showed that sequence sets with x4 phenotype inferred using the R5/X4 neural networks showed few statistically significant differences when compared to a set of sequences with experimentally determined phenotypes. However, using method 1 to infer the x4 phenotype resulted in more differences between the x4 and the X4 sets, especially at positions 11 and 25 (data not shown). This indicated that the R5/X4 neural network can extract sequences from the publicly available databases that, as a group, are largely indistinguishable from sequences with experimentally determined phenotypes.

Summary

To analyze and compare the reliability of existing, sequence motif-based algorithms and new, neural network-based algorithms for predicting phenotype, we have assembled sets of HIV-1 V3 sequences with known coreceptor usage or MT-2 cell tropism. Of the five rules examined, the presence of a basic amino acid at position 11 or 25 emerged as the best sequence motif-based predictor of both phenotypes, although its reliability for predicting X4 phenotypes fell below 0.50. We have developed a neural network that predicts CXCR4 coreceptor usage with a considerably higher reliability (0.69).

We can draw several general conclusions from this work. First, the reliability of even the best predictors of phenotype are probably not yet sufficient for use in a clinical setting. However, V3 genotype can be used to recapitulate patterns of amino acid variability linked to the phenotypes on a population level. In this respect, the neural network should be useful for examining large numbers of sequences to identify correlations between predicted phenotype and sequence changes outside of V3, and, as such, represents a substantial improvement over other algorithms. Finally, the difficulty of predicting coreceptor usage using V3 sequence alone reinforces the potential for the existence of phenotypic determinants that lie elsewhere in gp120 and highlights the need for their identification.

MATERIALS AND METHODS

Sequences

V3 sequences associated with known coreceptor usage (R5/X4 data set). To generate the R5/X4 data set, V3 sequences representing molecular or biological HIV-1 clones with known coreceptor usage were collected from the Los Alamos HIV Sequence Database (http://hiv-web.lanl.gov) and from published studies. Sequences derived from uncloned isolates were not considered. Unpublished sequences with phenotypes described in Singh *et al.* (1999), Lathey *et al.* (2000), and Zhang *et al.* (1998) were generously provided by the authors. Single representatives of entries with identical V3 peptide sequences were chosen arbitrarily and are described in Table 2. Information about the epidemiologic relationship between the sequences was either provided by the corresponding authors or inferred from cited literature and dendrograms (using the program PileUp in the Wisconsin program suite of the Genetics Computer Group) con-

structed using the largest common nucleotide fragment available from GenBank (255 nucleotides, data not shown). Closely related clusters of sequences from subjects sharing common sources of transmission were classified as originating from a single patient. Of the 47 sequences classified as X4 in this set, 18 used CXCR4 exclusively; the remainder retained some ability to utilize CCR5. Note that both sequence sets contain epidemiologically related sequences. However, all analyses were done using randomly chosen subsets of the R5/X4 set such that no two sequences of the same phenotype were derived from the same subject. This resulted in a subset of 181 of the original 216 R5/X4 sequences, which was used in most analyses. For some analyses a smaller subset was created by omitting some R5 sequences to increase the fraction of X4 sequences in the total set.

V3 sequences associated with known MT-2 cell tropism (NSI/SI data set). A separate set of V3 sequences from either molecular clones, biological clones, or viral isolates with known MT-2 cell tropism was also assembled (Table 2). Again, after all sequences were collected, only one of each set of identical entries was retained. Because in many cases only V3 peptide sequences were available, phylogenetic trees were not constructed. Instead, relatedness of sequences was inferred entirely from epidemiologic information available in GenBank records or in cited publications. Again, subsets of sequences of size $n = 124$ were chosen at random from the total of 257 sequences in the NSI/SI set such that no two sequences of identical phenotype were derived from the same subject.

V3 sequences with unknown phenotype. A set of $n = 1997$ V3 sequences in which no two sequence entries originated from the same patient was selected from over 11,000 V3 sequences retrieved from GenBank. This set was chosen using descriptions of V3 sequences composed largely of information provided by the Los Alamos HIV Sequence Database (B. Foley, personal communication).

All three sequence sets are available from the authors at http://cancer.med.unc.edu/swanstromlab/resources.html. Only HIV-1 clade B sequences were considered.

Sequence alignments and sequence selection

Alignments of V3 sequences were constructed by making serial pairwise alignments with the Clade B V3 consensus. For each aligned sequence, amino acids representing insertions with respect to the consensus were discarded and gaps were retained. Alignments were performed using a Java implementation of the Needleman–Wunsch algorithm (Needleman and Wunsch, 1970).

Analysis of algorithm performance

Uppercase letters indicate experimentally determined phenotypes (e.g., X4, SI), whereas lowercase letters denote inferred phenotypes (e.g., x4, si). The performance of the sequence classification methods was measured in terms of sensitivity, specificity, reliability, and the association coefficient phi (ϕ). The sensitivity of the algorithm can be determined as $P(x4 \mid X4) = [P(x4 \cap X4)]/P(X4)$ and the specificity as $P(r5 \mid R5) = [P(r5 \cap R5)]/P(R5)$. These measures reflect the fraction of correctly predicted sequences for each of the two phenotypes. Numbers range from 0 to 1, with 1 being complete agreement between the predicted and actual phenotypes. By Bayes' theorem, the reliability (positive predictive power), which is the probability that a sequence predicted to be *x*4 is in fact X4, was calculated as $P(X4 \mid x4) = [P(x4 \mid X4)P(X4)]/[P(x4 \mid X4)P(X4) + P(x4 \mid R5)P(R5)]$. The phi coefficient of association was calculated as $\phi = \sqrt{X^2/n}$, where X^2 is the test statistic from a chi-squared test of the 2×2 table and n is the total sample size.

Neural networks

Neural networks were fully connected feed-forward networks with 16 sigmoidal input nodes, three hidden sigmoidal nodes, and one linear output node. All nodes had a bias. The V3 amino acid positions 5, 7, 8, 10, 11, 13, 18–22, 24, 25, 27, and 32 plus the overall charge of the V3 loop were presented to the network. These positions were chosen based on both statistical analysis and empirical approaches. Amino acids and gaps were encoded numerically by consecutive numbers from 1 to 21 on the order of D, E, C, N, F, T, M, S, Y, Q, GAP(g), W, I, V, L, A, G, P, H, K, R. Several other number assignments were tried and did not improve network performance. Networks and analysis tools were implemented in Matlab (MathWorks). Training was done using a Bayesian regulation modification of backpropagation and started with random weights (MacKay, 1992). Learning rate was set to $\eta = 0.05$. The training target used values of 0 for R5/NSI and 1 for X4/SI. We assigned the x4 phenotype to any sequence that achieved a network score of greater than or equal to an empirically determined cutoff value of 0.4. The distribution of values at each amino acid position of V3 used as an input node was normalized using the mean and standard deviation of the corresponding position in a set of $n = 11{,}471$ numerically encoded V3 sequences from GenBank.

Statistical analysis

Amino acid (aa) composition at each position in V3 was compared between R5 and X4 sequences (in proportions of $P(R5)$ and $P(X4)$, respectively) in a subset of epidemiologically unrelated sequences from the R5/X4 data set using a permutation test. At each position *i*, a significance test based on the normal approximation of the binomial distribution was used to generate a Z score, $Z_{XR}(aa_{ij})$, describing the magnitude of the difference in the proportion of sequences in the X4 and R5 groups con-

taining a given amino acid, *j*. Only amino acids representing at least 10% of either R5 or X4 sequences at a position were considered. For each aa_{ij}, a reference distribution of $R = 1000$ Z scores was created by randomly casting sequences into groups A and B, such that $P(A_k) = P(X4)$, and $P(B_k) = P(R5)$ for all $k \in [1, \ldots, R]$, and calculating a new Z score, $Z_{AB}(aa_{ijk})$. The probability that a difference in the proportion of an aa *j* at position *i* is due to chance can thus be represented by $p \leq C/R$, where *C* is the smaller of the number of events in which $Z_{XR}(aa_{ij}) \leq Z_{AB}(aa_{ijk})$, or $Z_{XR}(aa_{ij}) \geq Z_{AB}(aa_{ijk})$, $1 \leq k \leq R$. If $Z_{XR}(aa_{ij})$ falls beyond the maximum or minimum of the reference distribution (i.e., $C = 0$), the probability is $p < 1/R$. The same procedure was used to compare the amino acid composition between SI and NSI sequences at each position in an alignment of epidemiologically unrelated sequences from the MT-2 data set.

Entropy

Entropy at each V3 position was calculated from an alignment of sequences according to Shannon as $H(i) = -\sum_{s=1}^{21} P(s_i)\log_2 P(s_i)$, with $P(s_i)$ being the frequency of each amino acid *s* at position *i* in the alignment (Shannon, 1948). Gaps were assigned a separate, twenty-first character. The maximum entropy of $H_{\max} = \log_2 21 \approx 4.4$ at any position is reached in the event of equal representation of all 21 characters. The minimal entropy of zero is reached for positions with 100% conservation.

Custom software

All custom programs, including the trained neural networks for phenotype inference, are available from the authors at http://cancer.med.unc.edu/swanstromlab/resources.html.

ACKNOWLEDGMENTS

Unpublished sequences with known coreceptor usage were generously provided by Ronald Collman and David Ho; our thanks also to Janet Lathey for providing sequences prior to publication. Brian Foley was instrumental in identifying epidemiologically unrelated V3 sequences from the Los Alamos HIV database. Francoise Seillier-Moiseiwitsch of the UNC CFAR biostatistics core provided valuable advice on statistical analysis. We are thankful to Jason Walker, who contributed the Java implementation of the Needleman–Wunsch alignment algorithm and other software. This work was supported through NIH Grant RO1-AI44667, the UNC Center for AIDS Research (P30-HD37260), and NIH Training Grant T32-AI07419 (N.H.).

REFERENCES

Alkhatib, G., Combadiere, C., Broder, C. C., Feng, Y., Kennedy, P. E., Murphy, P. M., and Berger, E. A. (1996). CC CKR5: A RANTES, MIP-1alpha, MIP-1beta receptor as a fusion cofactor for macrophage-tropic HIV-1. *Science* **272**(5270), 1955–1958.

Andeweg, A. C., Leeflang, P., Osterhaus, A. D., and Bosch, M. L. (1993). Both the V2 and V3 regions of the human immunodeficiency virus type 1 surface glycoprotein functionally interact with other envelope regions in syncytium formation. *J. Virol.* **67**(6), 3232–3239.

Bhattacharyya, D., Brooks, B. R., and Callahan, L. (1996). Positioning of positively charged residues in the V3 loop correlates with HIV type 1 syncytium-inducing phenotype. *AIDS Res. Hum. Retroviruses* **12**(2), 83–90.

Bjorndal, A., Deng, H., Jansson, M., Fiore, J. R., Colognesi, C., Karlsson, A., Albert, J., Scarlatti, G., Littman, D. R., and Fenyo, E. M. (1997). Coreceptor usage of primary human immunodeficiency virus type 1 isolates varies according to biological phenotype. *J. Virol.* **71**(10), 7478–7487.

Carrillo, A., and Ratner, L. (1996). Human immunodeficiency virus type 1 tropism for T-lymphoid cell lines: Role of the V3 loop and C4 envelope determinants. *J. Virol.* **70**(2), 1301–1309.

Chan, S. Y., Speck, R. F., Power, C., Gaffen, S. L., Chesebro, B., and Goldsmith, M. A. (1999). V3 recombinants indicate a central role for CCR5 as a coreceptor in tissue infection by human immunodeficiency virus type 1. *J. Virol.* **73**(3), 2350–2358.

Chesebro, B., Nishio, J., Perryman, S., Cann, A., O'Brien, W., Chen, I. S., and Wehrly, K. (1991). Identification of human immunodeficiency virus envelope gene sequences influencing viral entry into CD4-positive HeLa cells, T-leukemia cells, and macrophages. *J. Virol.* **65**(11), 5782–5789.

Chesebro, B., Wehrly, K., Nishio, J., and Perryman, S. (1992). Macrophage-tropic human immunodeficiency virus isolates from different patients exhibit unusual V3 envelope sequence homogeneity in comparison with T-cell-tropic isolates: Definition of critical amino acids involved in cell tropism. *J. Virol.* **66**(11), 6547–6554.

Choe, H., Farzan, M., Sun, Y., Sullivan, N., Rollins, B., Ponath, P. D., Wu, L., Mackay, C. R., LaRosa, G., Newman, W., Gerard, N., Gerard, C., and Sodroski, J. (1996). The beta-chemokine receptors CCR3 and CCR5 facilitate infection by primary HIV-1 isolates. *Cell* **85**(7), 1135–1148.

Cocchi, F., DeVico, A. L., Garzino-Demo, A., Cara, A., Gallo, R. C., and Lusso, P. (1996). The V3 domain of the HIV-1 gp120 envelope glycoprotein is critical for chemokine-mediated blockade of infection. *Nat. Med.* **2**(11), 1244–1247.

Cornelissen, M., Mulder-Kampinga, G., Veenstra, J., Zorgdrager, F., Kuiken, C., Hartman, S., Dekker, J., van der Hoek, L., Sol, C., Coutinho, R., *et al*. (1995). Syncytium-inducing (SI) phenotype suppression at seroconversion after intramuscular inoculation of a non-syncytium-inducing/SI phenotypically mixed human immunodeficiency virus population. *J. Virol.* **69**(3), 1810–1818.

de Jong, J. J., Goudsmit, J., Keulen, W., Klaver, B., Krone, W., Tersmette, M., and de Ronde, A. (1992). Human immunodeficiency virus type 1 clones chimeric for the envelope V3 domain differ in syncytium formation and replication capacity. *J. Virol.* **66**(2), 757–765.

Deng, H., Liu, R., Ellmeier, W., Choe, S., Unutmaz, D., Burkhart, M., Di Marzio, P., Marmon, S., Sutton, R. E., Hill, C. M., Davis, C. B., Peiper, S. C., Schall, T. J., Littman, D. R., and Landau, N. R. (1996). Identification of a major co-receptor for primary isolates of HIV-1. *Nature* **381**(6584), 661–666.

Doranz, B. J., Rucker, J., Yi, Y., Smyth, R. J., Samson, M., Peiper, S. C., Parmentier, M., Collman, R. G., and Doms, R. W. (1996). A dual-tropic primary HIV-1 isolate that uses fusin and the beta-chemokine receptors CKR-5, CKR-3, and CKR-2b as fusion cofactors. *Cell* **85**(7), 1149–1158.

Dragic, T., Litwin, V., Allaway, G. P., Martin, S. R., Huang, Y., Nagashima, K. A., Cayanan, C., Maddon, P. J., Koup, R. A., Moore, J. P., and Paxton, W. A. (1996). HIV-1 entry into CD4+ cells is mediated by the chemokine receptor CC-CKR-5. *Nature* **381**(6584), 667–673.

Feng, Y., Broder, C. C., Kennedy, P. E., and Berger, E. A. (1996). HIV-1 entry cofactor: Functional cDNA cloning of a seven-transmembrane, G protein-coupled receptor. *Science* **272**(5263), 872–877.

Fouchier, R. A., Brouwer, M., Broersen, S. M., and Schuitemaker, H. (1995). Simple determination of human immunodeficiency virus type 1 syncytium-inducing V3 genotype by PCR. *J. Clin. Microbiol.* **33**(4), 906–911.

Fouchier, R. A., Groenink, M., Kootstra, N. A., Tersmette, M., Huisman,

H. G., Miedema, F., and Schuitemaker, H. (1992). Phenotype-associated sequence variation in the third variable domain of the human immunodeficiency virus type 1 gp120 molecule. *J. Virol.* **66**(5), 3183–3187.

Groenink, M., Andeweg, A. C., Fouchier, R. A., Broersen, S., van der Jagt, R. C., Schuitemaker, H., de Goede, R. E., Bosch, M. L., Huisman, H. G., and Tersmette, M. (1992). Phenotype-associated env gene variation among eight related human immunodeficiency virus type 1 clones: Evidence for in vivo recombination and determinants of cytotropism outside the V3 domain. *J. Virol.* **66**(10), 6175–6180.

Groenink, M., Fouchier, R. A., Broersen, S., Baker, C. H., Koot, M., van't Wout, A. B., Huisman, H. G., Miedema, F., Tersmette, M., and Schuitemaker, H. (1993). Relation of phenotype evolution of HIV-1 to envelope V2 configuration. *Science* **260**(5113), 1513–1516.

Hung, C. S., Vander Heyden, N., and Ratner, L. (1999). Analysis of the critical domain in the V3 loop of human immunodeficiency virus type 1 gp120 involved in CCR5 utilization. *J. Virol.* **73**(10), 8216–8226.

Hwang, S. S., Boyle, T. J., Lyerly, H. K., and Cullen, B. R. (1991). Identification of the envelope V3 loop as the primary determinant of cell tropism in HIV-1. *Science* **253**(5015), 71–74.

Koot, M., Keet, I. P., Vos, A. H., de Goede, R. E., Roos, M. T., Coutinho, R. A., Miedema, F., Schellekens, P. T., and Tersmette, M. (1993). Prognostic value of HIV-1 syncytium-inducing phenotype for rate of CD4+ cell depletion and progression to AIDS. *Ann. Intern. Med.* **118**(9), 681–688.

Lathey, J. L., Brambilla, D., Goodenow, M. M., Nokta, M., Rasheed, S., Siwak, E. B., Bremer, J. W., Huang, D. D., Yi, Y., Reichelderfer, P. S., and Collman, R. G. (2000). Co-receptor usage was more predictive than NSI/SI phenotype for HIV replication in macrophages: Is NSI/SI phenotyping sufficient? *J. Leukoc. Biol.* **68**(3), 324–330.

Li, S., Juarez, J., Alali, M., Dwyer, D., Collman, R., Cunningham, A., and Naif, H. M. (1999). Persistent CCR5 utilization and enhanced macrophage tropism by primary blood human immunodeficiency virus type 1 isolates from advanced stages of disease and comparison to tissue-derived isolates. *J. Virol.* **73**(12), 9741–9755.

Ly, A., and Stamatatos, L. (2000). V2 loop glycosylation of the human immunodeficiency virus type 1 SF162 envelope facilitates interaction of this protein with CD4 and CCR5 receptors and protects the virus from neutralization by anti-V3 loop and anti-CD4 binding site antibodies. *J. Virol.* **74**(15), 6769–6776.

MacKay, D. J. C. (1992). Bayesian Interpolation. *Neural Comput.* **4**(3), 415–447.

Michael, N. L., Nelson, J. A., KewalRamani, V. N., Chang, G., O'Brien, S. J., Mascola, J. R., Volsky, B., Louder, M., White, G. C., 2nd, Littman, D. R., Swanstrom, R., and O'Brien, T. R. (1998). Exclusive and persistent use of the entry coreceptor CXCR4 by human immunodeficiency virus type 1 from a subject homozygous for CCR5 delta32. *J. Virol.* **72**(7), 6040–6047.

Milich, L., Margolin, B., and Swanstrom, R. (1993). V3 loop of the human immunodeficiency virus type 1 Env protein: Interpreting sequence variability. *J. Virol.* **67**(9), 5623–5634.

Milich, L., Margolin, B. H., and Swanstrom, R. (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**(1), 108–118.

Needleman, S. B., and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**(3), 443–453.

Nelson, J. A., Baribaud, F., Edwards, T., and Swanstrom, R. (2000). Patterns of changes in human immunodeficiency virus type 1 V3 sequence populations late in infection. *J. Virol.* **74**(18), 8494–8501.

Richman, D. D., and Bozzette, S. A. (1994). The impact of the syncytium-inducing phenotype of human immunodeficiency virus on disease progression. *J. Infect. Dis.* **169**(5), 968–974.

Roos, M. T., Lange, J. M., de Goede, R. E., Coutinho, R. A., Schellekens, P. T., Miedema, F., and Tersmette, M. (1992). Viral phenotype and immune response in primary human immunodeficiency virus type 1 infection. *J. Infect. Dis.* **165**(3), 427–432.

Schellekens, P. T., Tersmette, M., Roos, M. T., Keet, R. P., de Wolf, F., Coutinho, R. A., and Miedema, F. (1992). Biphasic rate of CD4+ cell count decline during progression to AIDS correlates with HIV-1 phenotype. *AIDS* **6**(7), 665–669.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Tech. J.* **27**, 379–423.

Shieh, J. T., Martin, J., Baltuch, G., Malim, M. H., and Gonzalez-Scarano, F. (2000). Determinants of syncytium formation in microglia by human immunodeficiency virus type 1: Role of the V1/V2 domains. *J. Virol.* **74**(2), 693–701.

Shioda, T., Levy, J. A., and Cheng-Mayer, C. (1992). Small amino acid changes in the V3 hypervariable region of gp120 can affect the T-cell-line and macrophage tropism of human immunodeficiency virus type 1. *Proc. Natl. Acad. Sci. USA* **89**(20), 9434–9438.

Singh, A., Besson, G., Mobasher, A., and Collman, R. G. (1999). Patterns of chemokine receptor fusion cofactor utilization by human immunodeficiency virus type 1 variants from the lungs and blood. *J. Virol.* **73**(8), 6680–6690.

Singh, A., and Collman, R. G. (2000). Heterogeneous spectrum of coreceptor usage among variants within a dualtropic human immunodeficiency virus type 1 primary-isolate quasispecies. *J. Virol.* **74**(21), 10229–10235.

Sullivan, N., Thali, M., Furman, C., Ho, D. D., and Sodroski, J. (1993). Effect of amino acid changes in the V1/V2 region of the human immunodeficiency virus type 1 gp120 glycoprotein on subunit association, syncytium formation, and recognition by a neutralizing antibody. *J. Virol.* **67**(6), 3674–3679.

Takeuchi, Y., Akutsu, M., Murayama, K., Shimizu, N., and Hoshino, H. (1991). Host range mutant of human immunodeficiency virus type 1: Modification of cell tropism by a single point mutation at the neutralization epitope in the env gene. *J. Virol.* **65**(4), 1710–1718.

Tersmette, M., de Goede, R. E., Al, B. J., Winkel, I. N., Gruters, R. A., Cuypers, H. T., Huisman, H. G., and Miedema, F. (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**(6), 2026–2032.

Tersmette, M., Lange, J. M., de Goede, R. E., de Wolf, F., Eeftink-Schattenkerk, J. K., Schellekens, P. T., Coutinho, R. A., Huisman, J. G., Goudsmit, J., and Miedema, F. (1989). Association between biological properties of human immunodeficiency virus variants and risk for AIDS and AIDS mortality. *Lancet* **1**(8645), 983–985.

Xiao, L., Owen, S. M., Goldman, I., Lal, A. A., deJong, J. J., Goudsmit, J., and Lal, R. B. (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**(1), 83–92.

Yamaguchi-Kabata, Y., and Gojobori, T. (2000). Reevaluation of amino acid variability of the human immunodeficiency virus type 1 gp120 envelope glycoprotein and prediction of new discontinuous epitopes. *J. Virol.* **74**(9), 4335–4350.

Zhang, L., He, T., Huang, Y., Chen, Z., Guo, Y., Wu, S., Kunstman, K. J., Brown, R. C., Phair, J. P., Neumann, A. U., Ho, D. D., and Wolinsky, S. M. (1998). Chemokine coreceptor usage by diverse primary isolates of human immunodeficiency virus type 1. *J. Virol.* **72**(11), 9307–9312.

Zhu, T., Mo, H., Wang, N., Nam, D. S., Cao, Y., Koup, R. A., and Ho, D. D. (1993). Genotypic and phenotypic characterization of HIV-1 patients with primary infection. *Science* **261**(5125), 1179–1181.