15. M. Koenig *et al.*, *Cell* **50**, 509 (1987).
16. For example, *neu* and nerve growth factor receptor sequences are at least several megabases apart and are resolvable only on more elongated chromosomes.
17. B. Trask, D. Pinkel, G. van den Engh, *Genomics* **5**, 710 (1989).
18. Evidence indicates that the chromatin fiber is packaged in loops, believed to contain on the order of 50 to 100 kb of DNA. See, for example, J. R. Paulson and W. K. Laemmli [*Cell* **12**, 817 (1977)] or (*19*).
19. B. Lewin, *Genes II* (Wiley, New York, 1985), p. 469.
20. In a cycling cell population that contains G_2 cells or in cultures of established cell lines, which generally contain tetraploid cells, a fraction of interphase nuclei will show two signals where one is expected. Replicated signals are generally closely spaced and are indistinguishable from the paired signals produced by simultaneous hybridization of two probes for two different sequences within the gene (see Fig. 3). This was confirmed by analysis of hybridizations with single probes in conjunction with BrdU labeling of S-phase and G_2 cells. This allowed us to determine that ≤11% of the normal human PBLs studied generated paired signals with a single probe and that these were S/G_2 cells. The X chromosome sequences studied here are apparently late replicating; however, this frequency can be much higher when using probes for other regions of the genome. When hybridizing two probes, a small fraction of S/G_2 cells will contain clusters of three to four signals because of the presence of replicated DNA. These cells are easily recognized and not included in the data. The consideration of replicated DNA becomes critical primarily in an analysis of very closely spaced sequences (less than 100 kb), in which resolution of two probes is not observed in most cells, but only 20 to 30%. Hence, replicated and tetraploid cells could constitute a major fraction of data points or actually mimic resolution of closely spaced sequences. We have found that this potential pitfall can be addressed either by BrdU labeling of cycling cells or by use of G_1-arrested cells (see WI-38 1 Mb data, Fig. 3B).
21. J. B. Lawrence and C. V. Johnson, unpublished observations.
22. L. J. Saez *et al.*, *Nucl. Acids Res.* **15**, 5443 (1987).
23. J. B. Lawrence and C. V. Johnson, unpublished observations.
24. C. V. Johnson and J. B. Lawrence, in preparation.
25. D. G. Harnden and H. P. Klinger, Eds. *An International System for Human Cytogenetic Nomenclature* (Karger, Basel, 1985), p. 48.
26. W. Vogel, M. Autenrieth, G. Speit, *Hum. Genet.* **72**, 129 (1986).
27. Additional information concerning the position of probes utilized was obtained from L. Kunkel (personal communication).
28. M. Burmeister and H. Lehrach, *Nature* **324**, 582 (1986); S. Kenwrick *et al.*, *Cell* **48**, 351 (1987).
29. Because single-probe hybridization efficiencies ranged from 80 to 90%, a fraction of nuclei when hybridized with two probes showed only one signal (10 to 30%). These were not included as "zero distance" data points, as they would introduce significant error. The histograms indicate that for distances of 675 and 750 kb, there is no spread of data points into the region where resolution becomes limiting (0.2 μm) and very little for the 375-kb distance. For smaller distances, the exclusion of single-signal data points may introduce some error into the determination of the PR (Table 1), but should still accurately discriminate larger and smaller distances.
30. We appreciate the generosity of L. Kunkel in providing a variety of probes and advice. We thank F. Ruddle, R. White, P. O'Connell, and L. Leinwand for providing the nerve growth factor receptor, the Erb-B2, and the myosin heavy chain probes, respectively. We are grateful to C. V. Johnson for technical assistance and M. Giorgio for photographic assistance. Supported by grants from the National Center for the Human Genome (HG 00251) and the National Institute of Child Health and Human Development (HD 18066) and a grant from the Muscular Dystrophy Association to J.B.L.

19 December 1989; accepted 17 May 1990

Conserved Sequence and Structural Elements in the HIV-1 Principal Neutralizing Determinant

GREGORY J. LAROSA, JOSEPH P. DAVIDE, KENT WEINHOLD, JULIE A. WATERBURY, ALBERT T. PROFY, JOHN A. LEWIS, ALPHONSE J. LANGLOIS, GORDON R. DREESMAN, R. NEAL BOSWELL, PHILLIP SHADDUCK, L. HOWARD HOLLEY, MARTIN KARPLUS, DANI P. BOLOGNESI, THOMAS J. MATTHEWS, EMILIO A. EMINI, SCOTT D. PUTNEY*

The principal neutralizing determinant (PND) of human immunodeficiency virus HIV-1 is part of a disulfide bridged loop in the third variable region of the external envelope protein, gp120. Analysis of the amino acid sequences of this domain from 245 different HIV-1 isolates revealed that the PND is less variable than thought originally. Conservation to better than 80 percent of the amino acids in 9 out of 14 positions in the central portion of the PND and the occurrence of particular oligopeptide sequences in a majority of the isolates suggest that there are constraints on PND variability. One constraining influence may be the structural motif (β strand—type II β turn—β strand—α helix) predicted for the consensus PND sequence by a neural network approach. Isolates with a PND similar to the commonly investigated human T cell lymphoma virus III_B (HTLV-III_B) and LAV-1 (BRU) strains were rare, and only 14 percent of sera from 86 randomly selected HIV-1 seropositive donors contained antibodies that recognized the PND of these virus isolates. In contrast, over 65 percent of these sera reacted with peptides containing more common PND sequences. These results suggest that HIV vaccine immunogens chosen because of their similarity to the consensus PND sequence and structure are likely to induce antibodies that neutralize a majority of HIV-1 isolates.

THE PROGRESS OF HIV-1 VACCINE development has been impeded by the amino acid sequence variability among different isolates of HIV-1. This variability is particularly high in the external envelope protein, gp120, which is the primary target for antibodies that neutralize virus infectivity (*1–3*). The PND of gp120 lies within a loop formed by a disulfide bridge between two invariant cysteines at positions 303 and 338 (*4–7*). Polyclonal antisera elicited by peptides of the PND, as well as monoclonal antibodies that bind the PND, neutralize virus infectivity (*4, 5, 7–9*), and PND peptides absorb most of the neutralizing antibodies elicited by gp120 or its precursor, gp160 (*4*). The PND is one of the more variable regions of the envelope and differs by as much as 50% among HIV-1 isolates (*10*). Because of this variability, neutralizing antibodies elicited by the PND from one isolate generally do not neutralize isolates with PNDs of different amino acid sequence (*4*). To develop a vaccine that elicits antibodies that neutralize a majority of HIV-1 isolates by binding to the PND, it is necessary to analyze PND sequences from a large number of HIV-1–infected individuals.

To obtain the PND sequences, we collected peripheral blood mononuclear cells (PBMCs) from 133 HIV-1–infected donors and cocultured these cells for 14 days with uninfected PBMCs (113 donors) or with H9 or CEM cell lines (20 donors) (*11*). The HIV-1–infected PBMCs were obtained from 95 randomly selected U.S. Air Force personnel from diverse geographic locations or from 38 infected persons in seven major U.S. cities and are likely to be representative of HIV-1 isolates from the United States. Donors were chosen irrespective of clinical symptoms.

Donor PBMCs were cultured with uninfected PBMCs to amplify infectious rather than defective virus genomes. (Similar sequence results were obtained when PND sequences were determined from unamplified virus cultures.) Total cellular DNA was extracted, and the region encoding the PND was amplified by the polymerase chain reaction (PCR) with oligonucleotide primers that hybridize with conserved flanking sequences (*11, 12*). The PCR product was

G. J. LaRosa, A. T. Profy, S. D. Putney, Repligen Corporation, Cambridge, MA 02139.
J. P. Davide, J. A. Waterbury, J. A. Lewis, E. A. Emini, Merck Sharp & Dohme Research Laboratories, West Point, PA 19486.
K. Weinhold, A. J. Langlois, P. Shadduck, D. P. Bolognesi, T. J. Matthews, Department of Surgery, Duke University Medical School, Durham, NC 27710.
G. R. Dreesman, BioTech Resources, Inc., San Antonio, TX 98249.
R. N. Boswell, Department of Medicine, Wilford Hall, U.S. Air Force Medical Center, Lackland Air Force Base, TX 78236.
M. Karplus and L. H. Holley, Department of Chemistry, Harvard University, Cambridge, MA 02138.

*To whom correspondence should be addressed.

Table 1. Frequently occurring amino acid sequences within the PND.

Sequence	Occurrence in 245 PND sequences (%)
SIHIGPGRAFY	41 (17)
SIHIGPG	46 (19)
IHIGPGRAFY	65 (27)
IHIGPGRAF	72 (29)
IHIGPGRA	84 (34)
IHIGPGR	90 (37)
IGPGRAFY	109 (44)
IGPGRAF	127 (52)
IGPGRA	156 (64)
GPGRAF	146 (60)

cloned into a plasmid vector and used to transform *Escherichia coli*; the sequence of the PND from one or more clones was determined. Each different PND sequence obtained from each PCR product was included in the analysis, resulting in a total of 222 PND sequences, to which were added 23 published sequences (*10*). None of the PND sequences contained termination codons, large deletions or insertions, or apparent frame shifts.

To analyze the sequences, we derived a consensus PND sequence by determining the amino acid occurring most often at each position (Fig. 1). Of the 245 sequences, 222 (91%) contained 35 amino acids from cysteine to cysteine; the remaining sequences contained one or two amino acid insertions or deletions. To accommodate an alignment of the sequences, gaps were introduced so that the consensus would contain 40 amino acid positions (Fig. 1).

The amino acids midway between the two cysteines (particularly the GPG tripeptide and the flanking amino acids) are the binding site for neutralizing antibodies (*13, 14*). The frequency of occurrence of amino acids at each position within the central 17 is shown in Fig. 2. The central proline and the two flanking glycines were highly conserved, and the GPG tripeptide was present in 237 of the 245 sequences. A single amino acid occurred in 80% or more sequences at six other positions (11, 14, 16, 22, 23, and 25). In addition, there was a tendency to conserve amino acid character (nonpolar, polar, or charged) at particular positions. This was particularly true in the positions

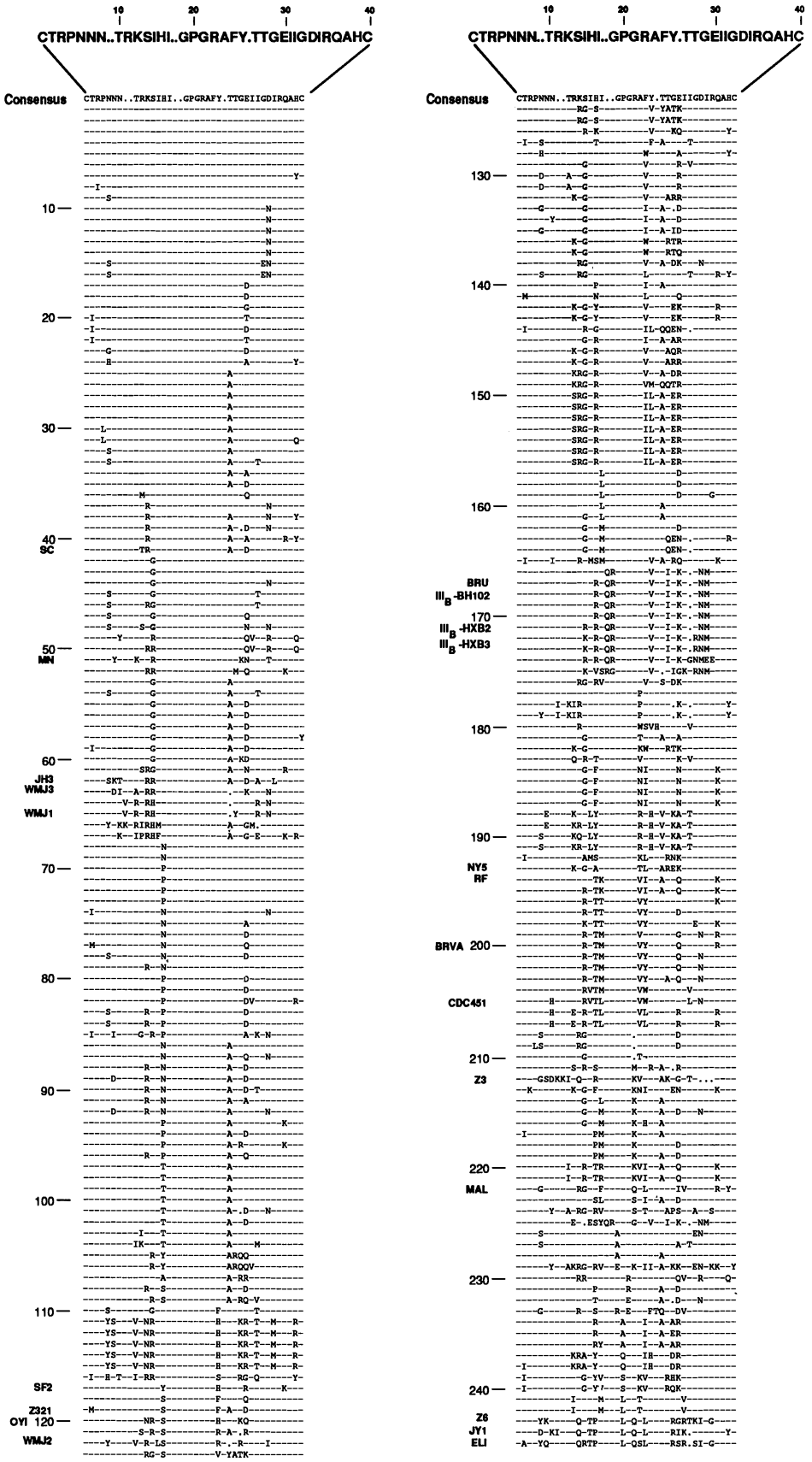

Fig. 1. Alignment of the 245 PND sequences with the consensus sequence. The entire C-C region of each sequence is shown in decreasing order of similarity to the consensus. In the comparison to the consensus, we first determined similarity at the position of the central proline, then determined similarity at the position of the flanking glycines, and so on. Isolates with substitutions relative to the consensus at the center of the PND were therefore considered less similar to the consensus than isolates with differences only near either end of the PND. The sequences (1 to 245) are shown with a dash where the residue is identical to the consensus, a single-letter amino acid code where the residue is different, and a period where there is a deletion (*13*). Published sequences (*10*) are indicated by name to the left of the sequence. Although similar, the sequences obtained from the virus isolates from the Air Force personnel tended to be more similar to the consensus than those from persons living in major U.S. cities. The reason for this difference is not understood.

COOH terminal to the GPG (for example, amino acids 22, 23, and 24 tended to be positively charged, nonpolar, and nonpolar, respectively). An aromatic amino acid tended to be present in the third position, or the fourth position, or both, COOH terminal to the GPG [in 232 of the 245 sequences (95%)]. We also observed conservation on the NH_2 terminal side of the GPG, with positions 14 and 16 usually occupied by hydrophobic residues and position 12 usually occupied by a positively charged residue. Even the more variable positions were occupied by a limited number of amino acids (five or less) in more than 80% of the sequences. In addition, the three positions at which insertions occurred within the central 17 positions (17, 18, and 26) were unoccupied in more than 95% of the sequences.

11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
R 200	K 177	S 124	I 230	H 114	I 202	• 234	• 234	G 240	P 233	G 241	R 223	A 204	F 176	Y 196	• 243	T 137
K 20	R 57	G 69	L 5	T 27	M 17	Q 10	R 10	A 3	L 5	E 2	K 11	V 18	I 29	V 17	H 1	A 92
S 11	N 6	R 40	M 3	R 27	L 10	R 1	G 1	E 1	A 3	R 2	Q 4	N 5	V 14	H 13	T 1	• 5
I 6	Q 5	H 4	T 3	P 22	V 5			R 1	Q 2		• 3	T 5	L 10	L 8		V 4
E 3		K 4	V 2	N 16	T 3				S 2		S 2	R 4	Y 7	F 5		Q 3
P 1		A 3	E 1	S 16	R 2						M 1	K 4	W 6	R 3		Y 3
Q 1		• 1	F 1	Y 13	K 2						G 1	P 3	T 1	S 1		S 1
G 1				F 5	Y 2							S 1	S 1	M 1		
M 1				A 2	F 1							W 1	H 1	I 1		
T 1				G 1	S 1											
				V 1												
				K 1												

Fig. 2. Frequency of occurrence of amino acids at positions 11 to 27 of the consensus PND. The 245 sequences were analyzed with a multiple-sequence alignment program to find the most frequently occurring amino acids and the extent of variability at positions in the central region of the PND. The amino acids occurring at positions 11 to 27 are shown with the number of times they occur at these positions. Periods indicate deletions.

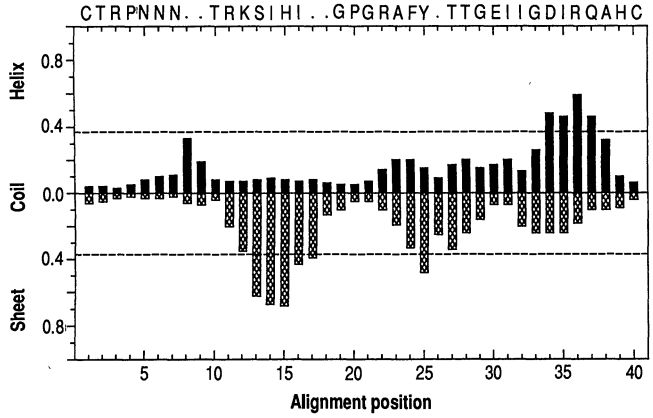

Fig. 3. Average neural network output for the 245 PND sequences. The neural network produces two outputs for each residue, the helix tendency and the sheet tendency. These are converted to predictions with the aid of a threshold, which is 0.37 for both helix and sheet. Average helix tendencies (solid bars) are plotted upward, and average sheet tendencies (hatched bars) are plotted downward. Dashed lines mark the thresholds within which coil is predicted. The consensus sequence is given at the top of the graph.

Table 2. Predicted secondary structure of 20 sample PNDs. H indicates α helix, E indicates β strand, and – indicates coil.

Isolate	Structure
1	------- --EEEEE ------- -------HHHHH--
12	------- --EEEEE ------E E------HHHH---
24	------- --EEEEE ------E E------HHHH---
36	------- -EEEEE- ------- ----------
48	------- ---EEEE -----EE E-----EE------
60	------- ---EEEE ------- -------HHHH---
72	------- ------- -----EE -------HHHHH--
84	------- ------- ----EEE -------HHHH---
96	------- ---EE-- ------- -------HHHH---
108	------- --EEEEE ------- -------HHHH---
119	------- ---EEE- ------- -------HHHH---
131	------- ---EEEE -----EE E----EEE------
143	------- ---EEEE ------E E------HHHH---
155	------- ---EEE- ----HHH HHHHH--HHHH---
167	------- ---EEEEE------EE EE--- --------
179	-------- -EEEEE- ------E EE -- --------
191	------- ---EEEE -----EE EEE----------
203	------- --EEEEE -----EE E------HHHH---
214	------- ---EEE- ------- -------HHHHH--
225	------- --- -EE-------EE EE--- --------
	10 20 30 40

Specific peptide sequences were present in a significant percentage of the virus isolates (Table 1). Given that an antibody bound to the PND will neutralize virus infectivity, and that six or seven amino acids can constitute a linear antibody binding site (*15, 16*), this suggests that, for example, an antibody that binds GPGRAF as contained in the gp120 glycoprotein will neutralize roughly 60% of randomly selected virus isolates. Antibodies have been elicited that bind the peptide GPGRAF (*17*). These antisera effectively neutralized all HIV-1 isolates we have assayed that contain the sequence GPGRAF (for example, those with PND isolate numbers 51, 86, 113, and 168) and did not neutralize those lacking GPGRAF (isolates 152 and 194).

These data suggest that, although the PND is variable (compared to conserved regions of gp120), there is selective pressure preserving particular types of amino acids and peptide sequences at most of the positions of the PND. One possible reason for this is the necessity to preserve a structural motif in the PND. To determine if structural elements are likely to be conserved in the PND, we generated secondary structure predictions for the 245 PNDs by means of the neural network method (*18*). We used a neural network that had been trained to "learn" the relation between amino acid sequence and secondary structure on a set of 62 proteins of known structure (*19*). The secondary structure predictions for a representative sample of the PNDs are given in Table 2. Of the 245 sequences, 93% are predicted to have a β strand around positions 13 to 16, even though this region contains two of the more variable positions (13 and 15). The sequences that do not have a β strand in positions 13 to 16 have a proline at position 15. Sixty-six percent of the sequences are predicted to have a short helix in positions 34 to 37, and 49% of the sequences show a β strand near position 25. The average helix and sheet tendencies for the 245 sequences are shown in Fig. 3. The conserved GPGR (positions 19 to 22) matches the most probable amino acids for a type II turn at positions $i + 1$ (P), $i + 2$ (G), and $i + 3$ (R) (*20*). Position i of the type II turn shows no strong residue preference. In the neural network method, the

Table 3. Frequency of reactivity of PND peptides with randomly selected sera. Eighty-six randomly selected adult and pediatric HIV-positive human sera were screened in a standard ELISA for binding activity to the peptides found in positions 6 to 33 of the PND. All sera were confirmed HIV-positive by protein immunoblot. PND peptides were plated in Linbro E.I.A. II Plus Microtitration plates (Flow Laboratories), at concentration determined to be optimum: 1 μg/ml for MN, SC, RF peptides, and 2 μg/ml for WMJ-2 and III_B peptides. Seropositive test and seronegative control human sera were incubated in duplicate wells for 90 min. The sera were tested a 1:200 dilution.

Isolate	Peptide sequence	Number of HIV sera reacting with PND peptides (%)
41	NNTTRSIHIGPGRAFYATGDIIGC	56 (65)
51	YNKRKRIHIGPGRAFYTTKNIIGC	57 (66)
122	NNVRRSLSIGPGRAFRTREIIGC	61 (71)
167	NNTRKSIRIQRGPGRAFVTIGKIGC	12 (14)
194	NNTRKSITKGPGRVIYATGQIIGC	19 (22)

magnitudes of helix and sheet output correlate with prediction accuracy (*18*). Given this result, we have the most confidence in the predictions for the GPG turn, the first β strand, the helix, and the second β strand, in that order. Taking these observations together, the most probable structure for the consensus PND is: C—β strand—type II turn—β strand—α helix—C.

The sequence comparison also reveals that the $HTLV\text{-}III_B$ and the LAV-1 (BRU) isolates have rare PND sequences that differ from the consensus sequence. The PND of both of these isolates contains a QR dipeptide NH_2 terminal to the GPG (*21, 22*); this dipeptide occurs in only 10 of the 245 sequences (4%). Because they were the HIV-1 isolates initially identified and characterized (*23, 24*), they have been widely used for vaccine development and other studies. Antibodies elicited by the PND from the III_B/LAV-1 isolates are type-specific and neutralize only a small percentage of virus isolates (*4, 14*). Thus, efforts at vaccine development that use envelope solely from the III_B/LAV-1 isolates should be expanded to include more representative HIV-1 isolates.

We also investigated the prevalence of virus isolates serologically by measuring the reactivity of HIV-positive human antisera to PND peptides. Sera from 86 randomly selected donors (not included in the sequence analysis) were assayed by direct enzyme-labeled immunosorbent assay (ELISA) for antibodies that bind to five peptides (Table 3). A majority of infected people appeared to be infected with viruses containing sequences similar to the consensus sequence. This serological analysis is in accord with the sequence data showing that consensus-like viruses are more common than those of the III_B/LAV-1 or RF type. The results also agree with an analysis of sera from 33 HIV-1–infected human donors in which a COOH terminal fragment of the gp120 protein or similar synthetic peptides were used (*25, 26*).

It is not known what function the PND plays in virus infection or what effects the differences in amino acid sequence have on that process. Because antibodies bound to the PND prevent virus infection, as with viruses such as influenza (*27, 28*), HIV-1 may undergo changes in the PND to escape selective pressure imposed by neutralizing antibody in vivo. HIV-1 isolates have been generated that are resistant to PND-directed antibody (*29, 30*).

The mechanism by which initial virus infection occurs is not well understood, nor is it known whether particular HIV-1 isolates are more infectious and should therefore be the major focus of vaccine efforts. However, our data are encouraging in that they indicate the restricted nature of the most prevalent PND sequences in the virus population and suggest that a limited set of peptide sequences will be able to elicit neutralizing antibodies against a large percentage of HIV-1 isolates.

REFERENCES AND NOTES

1. W. G. Robey *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **83**, 7023 (1986).
2. S. D. Putney *et al.*, *Science* **234**, 1392 (1986).
3. J. R. Rusche *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **84**, 6924 (1987).
4. J. R. Rusche *et al.*, *ibid.* **85**, 3198 (1988).
5. T. J. Palker *et al.*, *ibid.*, p. 1932.
6. J. Goudsmit *et al.*, *ibid.*, p. 4478.
7. S. Matsushita *et al.*, *J. Virol.* **62**, 2107 (1988).
8. M. A. Skinner *et al.*, *AIDS Res. Hum. Retroviruses* **4**, 187 (1988).
9. E. Kinney-Thomas *et al.*, *AIDS* **2**, 25 (1988).
10. G. Myers *et al.*, *Human Retroviruses and AIDS* (Los Alamos National Laboratory, Los Alamos, NM, 1989), pp. II–47.
11. PBMCs from infected donors were purified by Ficoll gradient: 10^7 PBMCs were cocultured with 10^7 phytohemagglutinin-blasted, uninfected, normal donor PBMCs in 10 ml of RPMI-1640, 20% fetal calf serum, and 5% interleukin-2. After 14 days, DNA was isolated from p24-positive cocultures (or from chronically infected cells) by lysis and proteolytic digestion of ~10^7 cells in 3 ml of 10 mM tris hydrochloride (pH 8.0), 100 mM NaCl, 25 mM EDTA, 0.5% SDS, Proteinase K (0.1 mg/ml) (Bethesda Research Laboratories) (50°C, 18 hours), followed by phenol/chloroform extraction and ethanol precipitation. The PND encoding sequence was amplified in a Techne PHC-1 thermal cycler by 30 cycles of PCR (*12*) (The cycles were 95°C, 1.5 min; 48°C, 1.5 min; 72°C, 1.5 min.) from 1 μg of infected-cell DNA in a reaction (50 to 100 μl) containing 50 mM tris-hydrochloride (pH 8.3), 50 mM KCl, 1.25 mM $MgCl_2$, 0.01% gelatin, 200 mM each deoxynucleotide triphosphate (dNTP) (Pharmacia), 0.5 μM PCR-1 or RP-Hpa primer, 0.5 μM PCR-2 or RP-Cla primer, *Taq* polymerase (0.02 U/ml) (Perkin-Elmer Cetus), or in a Perkin-Elmer Cetus thermal cycler as follows: 1 cycle at 94°C, 7 min; 2 cycles at 94°C, 5 min, at 55°C, 1.5 min, and at 72°C, 1.5 min; 28 cycles at 94°, 55°, and 72°C, each for 1.5 min; 1 cycle at 72°C, 8.5 min, followed by the addition of 50 μM fresh dNTPs and 2 units of the Klenow fragment of DNA polymerase and incubation at 37°C, 30 min. The amplified PND fragment was purified from a 3% NuSieve gel (FMC Corp.) or by ultrafiltration with an MC100 filter unit (Millipore Corp.) and was cloned blunt-end into a plasmid vector. *Escherichia coli* were transformed, and the PND insert of one to five clones was sequenced (usually three clones). The PCR primers were: PCR-1 5′-GCTAAAACCATAATAGTACAGCTG-3′ [corresponding to Ratner *et al.* positions 6642–6665 (*21*)] and PCR-2 5′-CAATTCTGGGTCCCCTCCTGAGG-3′ (positions 6915–6892), or RP-Hpa 5′-TCTGTTAACTTCACTGATAARGCTAAAACCATCATAGTACATCTCG-3′ (positions 6627–6675) and RP-Cla 5′-GCTATCGATTTGTCTTAAAGTGTCATTCCATTGTGC-3′ (positions 6842–6807).
12. R. K. Saiki *et al.*, *Science* **239**, 487 (1988).
13. The single-letter amino acid code is: nonpolar amino acids, A (Ala), V (Val), L (Leu), I (Ile), P (Pro), F (Phe), W (Trp), M (Met); polar amino acids, G (Gly), S (Ser), T (Thr), C (Cys), Y (Tyr), N (Asn), Q (Gln); negatively charged amino acids, D (Asp), E (Glu); positively charged amino acids, K (Lys), R (Arg), H (His).
14. K. Javaherian *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 6768 (1989).
15. H. M. Geysen, R. H. Meloen, S. J. Barteling, *ibid.* **81**, 3998 (1984).
16. R. A. Houghten, *ibid.* **82**, 5131 (1985).
17. K. Javaherian, D. Bolognesi, S. D. Putney, T. Matthews, in preparation.
18. L. H. Holley and M. Karplus, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 152 (1989).
19. W. Kabsch and C. Sander, *Biopolymers* **22**, 257 (1983).
20. C. M. Wilmont and J. M. Thornton, *J. Mol. Biol.* **203**, 221 (1988).
21. L. Ratner *et al.*, *Nature* **313**, 277 (1985).
22. S. Wain-Hobson, P. Sonigo, O. Danos, S. Cole, M. Alizon, *Cell* **40**, 9 (1985).
23. M. Popovic, M. G. Sarngadharan, E. Read, R. C. Gallo, *Science* **224**, 497 (1984).
24. F. Barré-Sinoussi *et al.*, *ibid.* **220**, 868 (1983).
25. G. Zuart *et al.*, *Lancet* **335**, 474 (1990).
26. Y. Devash *et al.*, *AIDS Res. Hum. Retroviruses* **6**, 307 (1990).
27. R. G. Webster, W. G. Laver, G. M. Air, G. C. Schild, *Nature* **296**, 115 (1982).
28. D. C. Wiley *et al.*, *ibid.* **289**, 373 (1981).
29. J. A. McKeating *et al.*, *AIDS* **3**, 777 (1989).
30. P. Nara, E. Emini, S. Putney, unpublished data.
31. The new, previously unpublished nucleotide sequences determined in this work have been deposited with GenBank under the accession number M37586.

32. We thank D. Tendler, M. Kaufmann, P. Leighton, B. Clark, D. Mann, T. Greenwell, C. McDanal, M. White, W. Schleif, J. Quintero, and H. McClure for technical assistance; K. Javaherian, G. Gray, W. Herlihy, F. Wong-Staal, and R. Gallo for helpful discussions; and T. Amirault for developing the multiple sequence alignment software. The analysis done at Harvard University was supported in part by a grant from the National Institutes of Health.

13 April 1990; accepted 3 July 1990