JOURNAL OF VIROLOGY, Feb. 1987, p. 570–578
0022-538X/87/020570-09$02.00/0

Vol. 61, No. 2

Computer-Assisted Analysis of Envelope Protein Sequences of Seven Human Immunodeficiency Virus Isolates: Prediction of Antigenic Epitopes in Conserved and Variable Regions

SUSANNE MODROW,[1] BEATRICE H. HAHN,[2] GEORGE M. SHAW,[2] ROBERT C. GALLO,[3] FLOSSIE WONG-STAAL,[3] AND HANS WOLF[1]*

Max von Pettenkofer-Institut, 8000 Munich 2, Federal Republic of Germany[1]; University of Alabama, Birmingham, Alabama 25294[2]; and National Institutes of Health, Bethesda, Maryland 20892[3]

Received 16 June 1986/Accepted 15 October 1986

Independent isolates of human immunodeficiency virus (HIV) exhibit a striking genomic diversity, most of which is located in the viral envelope gene. Since this property of the HIV group of viruses may play an important role in the pathobiology of the virus, we analyzed the predicted amino acid sequences of the envelope proteins of seven different HIV strains, three of which represent sequential isolates from a single patient. By using a computer program that predicts the secondary protein structure and superimposes values for hydrophilicity, surface probability, and flexibility, we identified several potential antigenic epitopes in the envelope proteins of the seven different viruses. Interestingly, the majority of the predicted epitopes in the exterior envelope protein (gp120) were found in regions of high sequence variability which are interspersed with highly conserved regions among the independent viral isolates. A comparison of the sequential viral isolates revealed that changes concerning the secondary structure of the protein occurred only in regions which were predicted to be antigenic, predominantly in highly variable regions. The membrane-associated protein gp41 contains no highly variable regions; about 80% of the amino acids were found to be conserved, and only one hydrophilic area was identified as likely to be accessible to antibody recognition. These findings give insight into the secondary and possible tertiary structure of variant HIV envelope proteins and should facilitate experimental approaches directed toward the identification and fine mapping of HIV envelope proteins.

Most patients with acquired immunodeficiency syndrome (AIDS) or AIDS-related complex show specific antibodies directed against proteins of human immunodeficiency virus (HIV), which have virus-neutralizing activity (3, 37, 40, 46) and are supposed to be directed against antigenic determinants located on the surface glycoprotein, as has been shown for other enveloped virus particles. However, the virus seems to have adopted properties which allow evasion of the immune surveillance mechanisms of the host. Differences among various isolates of HIV have primarily been analyzed at the nucleotide sequence level (35) of independent isolates and also recently in sequential viral isolates from the same patient (21). These variations seem to be concentrated in the envelope protein-encoding region of HIV (20) and may be fundamentally important for the biology and pathogenicity of HIV (40, 50). For this reason and for the development of viral antigens for diagnostic or vaccine use, it is important to identify and characterize antigenic determinants located in the glycoprotein complex of HIV and to define their possible functions.

In this study, we analyzed the amino acid sequences of the envelope protein complexes derived from the nucleotide sequences of seven AIDS virus isolates (21, 36, 39, 40, 43), three of which represented sequential isolates from the same patient. The present work is an extension of previously published reports on the genetic variability of the HIV envelope protein complex, which mainly focused on the DNA and primary amino acid sequences (9, 28, 43). By computer analysis we predicted the secondary structure of gp120 and gp41, the cleavage products of gp160 (11), and predicted potential antigenic sites by superimposing this secondary structure with the values for hydrophilicity, flexibility, surface probability, and glycosylation. Thus, 11 potential antigenic sites were identified, 9 of which were located in the exterior part (gp120) of the envelope protein and 2 in the membrane-bound portion (gp41). Five highly variable regions were characterized, all contained in gp120, coinciding with the predicted epitopes. In sequential isolates from a single patient, all alterations of secondary structures occurred in those regions which were identified as antigenic epitopes.

These results indicate that genomic variations of the AIDS virus seem to be manifested mainly in the extracellular portion of the envelope protein. The fact that those variations coincide with possible antigenic sites suggests that these regions may be immunogenic and may be of fundamental importance for the pathobiology of the virus.

MATERIALS AND METHODS

DNA and protein sequences. The following convention will be used to designate the HIV strains used. Strains HTLV-III(BH10), LAV(1A), HTLV-III(HAT3), HTLV-III(WMJ1), HTLV-III(WMJ2), and HTLV-III(WMJ3) are referred to as BH10, LAV1A, HAT3, WMJ1, WMJ2, and WMJ3, respectively.

The nucleotide sequences of the envelope open reading frames of HIV strains BH10, LAV1A, and ARV2 have been previously reported (36, 39, 45). HAT3 is the nucleotide sequence of a virus isolated in 1983 from a patient with AIDS (20, 33, 43); WMJ1 is a virus isolate from a child with AIDS born in 1982 and infected perinatally by her HIV-positive mother. WMJ2 and WMJ3 are sequential isolates from this same patient taken 3 (WMJ2) and 7 (WMJ3) months after the first (21).

* Corresponding author.

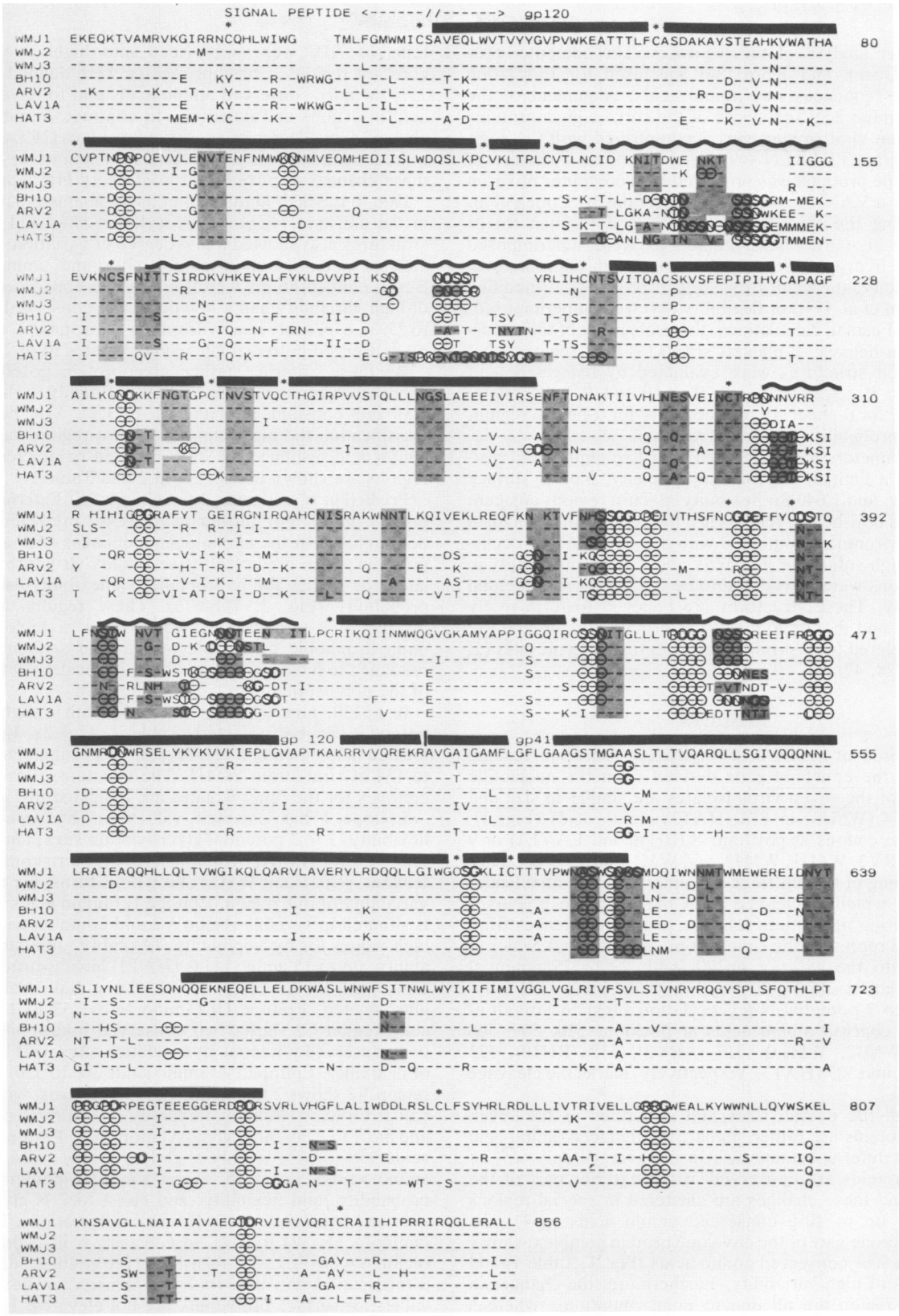

FIG. 1. Amino acid sequences of the entire envelope genes of seven HIV isolates. WMJ1, WMJ2, and WMJ3 are sequential isolates from the same patient; BH10, ARV2, LAV1A, and HAT3 represent independent virus isolates. Alignment of the sequences was done with the assistance of PRTLAN (49). Numbering of amino acid sequences is done from the first residue of the envelope protein open reading frame of WMJ1 up to the end. Dashes indicate amino acid identity with WMJ1, spaces indicate the absence of that amino acid. N-glycosylation sites are indicated as shaded areas, amino acids in β-turn regions are encircled. Symbols: *, cysteine residues; ▬, conserved regions; ⁓⁓, variable regions.

Computer analysis of the sequences. The computer program of Queen and Korn (34) was used for translating nucleotide sequences into amino acids. Alignment of the deduced amino acid sequences of the surface glycoproteins of the seven viral isolates was accomplished with the computer program PRTALN (49). The secondary structure of the envelope proteins was predicted by a computer program written for a VAX750, based on suggestions by Cohen et al. (8) by using the algorithms of Chou and Fasman (6) or Garnier et al. (16). These predictions were superimposed with local hydrophilicity values (23). As an alternative to hydrophilicity, the values for surface probability (modified from Emini et al. [13]) or flexibility (24) were superimposed. For Chou-Fasman calculations, the probability of the occurrence of α-helixes, β-pleated sheets, β-turn regions, and random coil structures were evaluated by using stringent conditions: $P_{boundary} \geq 1$, with $P_\beta > P$ and $P_t > P_\alpha$ (P_β = probability for β-sheet, P_α = probability for α-helical region, and P_t = probability for β-turn regions).

The parameters were averaged over five amino acid residues with a limit of 0.7 for hydrophilicity, 5.0 for surface probability, and 1.040 for flexibility. β-Turn regions adjacent to β-sheets or α-helical regions in a hydrophilic or nonhydrophobic environment combined with the simultaneous occurrence of high values for flexibility and surface probability in those regions were considered to be candidates for potential antigenicity. These structures are believed to form freely accessible and mobile loops at the protein surface and are thus considered to be prime candidates for antigenic sites (2, 6, 12, 23, 28, 47).

RESULTS

Comparison of amino acid sequences. The open reading frames of the envelope gene derived from the nucleotide sequence of the seven virus isolates are similar in size and encode 854 (WMJ3) to 873 (HAT3) amino acids (Fig. 1). Methionine codons at positions 8 (BH10 and LAV1A) or 9 (HAT3, ARV2, WMJ1, WMJ2, and WMJ3) supposedly mark the beginning of the envelope protein with a potential leader sequence, which is cleaved from the envelope precursor protein during maturation (1). The peptide sequences from position 38 represent the envelope precursor gp160, which is cleaved into the exterior gp120 (with 23 to 25 potential N-glycosylation sites) and the membrane-bound gp41 (containing 4 to 7 potential N-glycosylation sites). A stretch of positively charged amino acids at positions 510, 509, 508 (WMJ1, WMJ2, WMJ3), 517 (ARV2), 518 (BH10), 523 (LAV1A), and 527 (HAT3), respectively, marks the cleavage site (1, 11).

Although the overall sizes and structures of the seven surface proteins are rather similar, the deduced amino acid sequences differ substantially. On the average only 66% of the amino acids are conserved in the exterior part of the protein, and these changes are clustered in special regions with only up to 10% conserved amino acids. gp41, the transmembrane part of the envelope protein complex, shows more than 80% conserved amino acids (Fig. 1, Table 1) and no regions of high variability. Furthermore, the changes in the latter region are all due to point mutations, whereas changes in gp120 frequently result from insertions and deletions and appear as clustered mutations interspersed with segments which have a high content of conserved amino acids (89%).

According to their content of conserved amino acids, glycosylation sites, and β-turn regions, the envelope proteins of HIV were subdivided into highly variable and constant regions. Constant regions were defined to contain 75% or more conserved amino acids and to have no amino acid insertions and deletions. In contrast, variable regions showed a low degree of conserved residues (25% or less) and a great variability in length due to deletions or insertions; these changes occurred at at least every fifth amino acid.

Due to the high number of conserved amino acid residues in the constant regions, glycosylation sites and secondary structures also showed a low degree of variation; more than 50% of potential glycosylation sites and amino acids in β-turn configurations were found to be conserved. Regions of high variability had a low degree of conserved β-turns (0 to 25%), especially in a hydrophilic environment, and almost no conserved potential glycosylation sites (0 to 25%).

By these criteria, surface glycoprotein gp160 could be subdivided into a clustered pattern of highly variable (V1 to V5) and constant regions (C1 to C6). For purposes of classification the minimum length of a region was set at 10 amino acid residues; the locations and parameters of those regions are shown in Fig. 1 and 2 and Table 1.

Prediction of antigenic determinants. (i) Exterior envelope protein gp120. For further analysis we predicted antigenic determinants in the amino acid sequences with a computer program which predicted the secondary structure and calculated the values for hydrophilicity, flexibility, and surface probability (Fig. 2, Table 1). These regions are mainly located in β-turn regions which show a high degree of nonhydrophobic or flexible amino acid sequences or are predicted to have a high probability of location at the surface of the polypeptide (8, 12, 13, 19, 32, 44).

In gp120, nine epitopes (I to IX) with a high antigenic potential can be predicted (Table 2, Fig. 2). Epitope I is located in a region of high variability (V1) at amino acids 137 to 154 of viral strain WMJ1. The locations of the antigenic epitopes for the other isolates are indicated in Table 2.

Epitope I has elevated values for hydrophilicity and flexibility, 1 to 3 potential glycosylation sites, and a number (up to 11 in LAV1A) of amino acids in β-turn configurations. Isolates WMJ1 and WMJ3 show no β-turns due to the high variability of this region. Epitope II (amino acids 186 to 203) is situated in variable region V2 and contains residues with high values for hydrophilicity, flexibility, and surface probability; up to 15 amino acids (HAT3) have β-turn configurations (with the exception of ARV2), and all strains may be glycosylated. Epitope III (amino acids 232 to 246) in constant region C2 shows two conserved hydrophilic β-turns, two conserved potential glycosylation sites, and high values of flexibility. Epitope IV (amino acids 300 to 320) in variable region V3 shows 2 to 8 amino acids in β-turns, with elevated values for hydrophilicity, flexibility, and surface probability, and has 1 to 2 potential glycosylation sites. Epitope V (amino acids 358 to 375) is only slightly variable (72% conserved amino acids), has high values for hydrophilicity, surface probability, and flexibility, and has 1 to 2 N-glycosylation sites and 6 to 8 residues in β-turn configurations. Epitope VI (amino acids 394 to 412), in contrast, is in highly variable region V4, with 2 to 4 possible glycosylation sites, 4 to 11 β-turns, and high values for hydrophilicity and flexibility; the values for surface probability are not elevated. Epitope VII (amino acids 445 to 458) in constant region C3 has 4 to 7 β-turns. One glycosylation site is hydrophilic and flexible and is directly followed by epitope VIII (amino acids 459 to 469), which shows high variability (V5), is hydrophilic, flexible, and possibly glycosylated, and has one (ARV2) to six residues in β-turn configurations. Epitope IX (amino

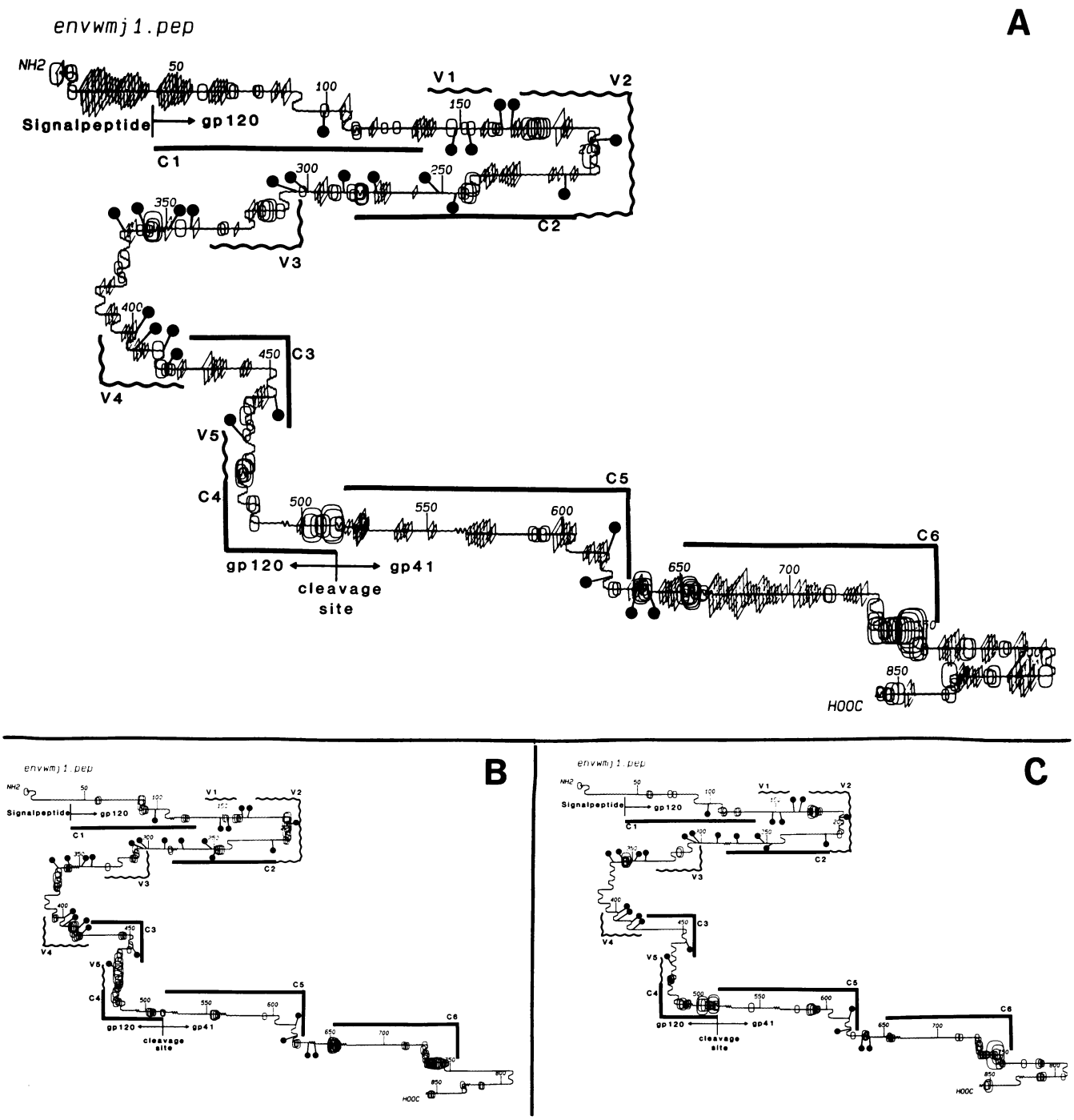

FIG. 2. Chou-Fasman prediction of the envelope protein gp160 derived from the sequence of WMJ1. The computer plots start at the methionine at residue 9 of the open reading frame. The probability of the occurrence of α-helices (〰), β-pleated sheets (〰), random coils (〰), and β-turn regions (ↄ) were evaluated by using stringent conditions. The parameters for hydrophilicity, flexibility, and surface probability were averaged over five amino acid residues, with a limit of 0.7 for hydrophilicity, 1.040 for flexibility, and 5.0 for surface probability. Symbols: ●—, N-glycosylation sites; ▬, conserved regions; 〰, variable regions. (A) Secondary structures superimposed with the values for hydrophilicity; O, hydrophilic regions, ◇, hydrophobic regions. (B) Secondary structures superimposed with the values for flexibility; O, flexible regions. (C) Secondary structures superimposed with the values for surface probability; O, high surface probability.

acids 470 to 483) in region C4 has high values for all parameters, four conserved β-turns, and no potential glycosylation sites. In fact, epitopes VII to IX are one continuous antigenic region which is subdivided because constant regions C3 and C4 are interrupted by V5, a rather short region with high variability.

There may be additional epitopes in singular virus isolates (e.g., amino acids 285 to 293 in ARV2) which were not classified because similar epitopes could not be found in the other virus strains.

Some of these epitopes must be considered of lesser antigenic quality, since they have only a few β-turns (epitope

TABLE 1. Parameters and values for conserved and variable regions, calculated from the sequences of seven HIV envelope proteins (gp 160)[a]

Region	Length	Amino acid deletions/ insertions	Conserved amino acids		Glycosylation sites	Conserved glycosylation sites		β-Turns	Conserved β-turns	
			No.	%		No.	%		No.	%
C1	97/97/97	0	84	87	1/1/1	1	100	4/4/4	2	50
38–134	97/97/97/97				1/1/1/1			2/4/2/4		
V1	20/20/20	0–15	3	10	2/2/2	0	0	0/2/0	0	0
135–154	24/23/29/25				1/2/3/3			8/5/11/7		
	8/8/8	0	7	87	1/1/1	1	100	0/0/0	0	100
155–162	8/8/8/8				1/1/1/1			0/0/0/0		
V2	41/41/41	0–18	11	21	2/3/2	1	20	5/6/5	0	0
163–203	39/44/39/52				2/4/3/5			2/2/2/17		
C2	76/76/76	0	68	89	3/3/3	2	50	2/2/2	2	50
204–279	76/76/76/76				4/3/4/3			2/4/2/4		
	25/25/25	0	19	76	3/3/3	3	100	0/0/0	0	0
280–304	25/25/25/25				3/3/3/3			0/3/0/0		
V3	26/26/26	0–5	7	25	0/0/0	0	0	4/2/4	2	25
305–330	27/26/27/26				1/1/1/1			8/6/8/8		
	65/65/65	0–1	48	72	5/6/6	4	67	11/9/10	8	72
331–395	66/66/66/65				5/5/5/6			11/10/11/10		
V4	19/19/19	0–11	4	17	3/2/3	0	0	4/6/4	0	0
396–414	23/18/23/19				2/1/2/3			11/4/11/9		
C3	44/44/44	0	38	86	1/1/1	1	100	7/7/7	4	57
415–458	44/44/44/44				1/1/1/1			7/7/7/4		
V5	11/11/11	0–2	2	16	1/1/1	0	0	5/5/5	1	18
459–469	11/12/11/12				1/1/1/1			5/1/6/1		
C4	41/41/41	0	36	87	0/0/0	0	100	4/4/4	4	100
470–510	41/41/41/41				0/0/0/0			4/4/4/4		
C5	106/106/106	0–1	95	90	2/2/2	2	100	6/8/8	6	75
511–616	105/106/105/106				2/2/2/2			6/6/6/8		
	38/38/38	0	17	44	2/2/2	2	100	2/0/0	0	0
616–653	38/38/38/38				2/2/2/2			2/0/2/2		
C6	92/92/92	0	84	85	0/0/1	0	0	6/6/6	6	66
654–745	92/92/92/92				1/0/1/0			6/8/6/10		
	111/111/111	0	77	69	0/0/0	0	0	5/5/5	4	57
746–856	111/111/11/111				2/1/2/2			5/4/5/7		

III) or no high values for surface probability (epitopes I, VI, VII, and VIII).

(ii) Transmembrane protein gp41. Only two antigenic sites could be identified in gp41. These contain amino acids 612 to 635 and 722 to 745 (Fig. 1 and 2, Tables 1 and 2). The first, epitope X, is located in a slightly variable region (56% conserved amino acids) and contains four conserved glycosylation sites, has high values for hydrophilicity, surface probability, and flexibility, has 4 to 6 β-turns, and is probably the only antigenic site in gp41 which is located outside the lipid bilayer and accessible to antibody reaction and recognition. The second epitope (amino acids 722 to 745) is located directly after a stretch of hydrophobic amino acids which is likely to be a transmembrane region of gp41 (TM3); these hydrophilic, flexible amino acids in β-turn regions with high surface probability probably represent the hydrophilic anchor sequence which has been identified in most membrane proteins (10). This region, however, should be inside the cell and thus is not the best epitope for antigenic response. The following stretch of about 100 amino acids has a further region which might be an antigenic determinant. This region, whose function is not known, might however be gradually cleaved off by proteases from the precursor before maturation of gp41. With antibodies against a synthetic peptide derived from this region (25), mainly the precursor gp160 could be identified; furthermore, the molecular size of a glycosylated gp41 should be about 52,000 to 54,000 daltons by calculation (42,000 daltons of the primary product plus the molecular size of the carbohydrate moiety) unless it is proteolytically processed.

Alterations in the envelope proteins of the sequential isolates. The envelope proteins of the sequential viral isolates from the same patient which were taken at intervals of 3 and 4 months are very similar in size and show only three amino acid deletions or insertions. Most changes are due to point mutations (21). Most of the alterations in the amino acid sequences which are due to those mutations are located in those regions which were found to be highly variable. Differences in the secondary structures (Fig. 1) were identified in regions V1, V2, V3, and V4, corresponding with antigenic epitopes I, II, IV, and VI, and in epitopes V and X, which are located in slightly variable regions (76 and 56% of conserved amino acids, respectively). A further β-turn alteration occurs in region C5 in a short hydrophilic environment which may be located between two transmembrane-spanning regions (Fig. 1, 2, and 3). That means that all alterations concerning the secondary structure of the envelope protein, with the exception of that in region C5, are located in the predicted antigenic determinants; epitopes III, VII, VIII, IX, and XI are conserved. Interestingly, some of the β-turn alterations identified in isolates WMJ1 and WMJ2 are reversed in isolate WMJ3 (I, II, IV, and VI), even if there are further variations in the amino acid sequence (I and IV).

Prediction of tertiary structures. After cleavage of the leader peptide from precursor protein gp160, gp120 represents the highly glycosylated exterior part of the envelope

TABLE 1—*(Continued)*

Region	Length	Hydrophilic β-turns	Conserved hydrophilic β-turns		Hydrophilicity		Surface probability		Flexibility	
			No.	%	No.	%	No.	%	No.	%
C1	97/97/97	2/2/2	2	100	10/8/7	7–10	7/7/9	7–10	11/1/11	8–10
38–134	97/97/97/97	2/2/2/2			8/9/9/9		8/9/7/10		11/11/8/12	
V1	20/20/20	0/0/0	0	0	5/5/7	13–33	0/0/0	0–13	6/7/6	28–82
135–154	24/23/29/25	3/0/1/0			9/9/4/4		2/3/3/0		14/19/17/14	
	8/8/8	0/0/0	0	100	0/0/0	0	0/0/0	0	0/0/0	0
	8/8/8/8	0/0/0/0			0/0/0/0		0/0/0/0		0/0/0/0	
V2	41/41/41	4/4/4	0	0	9/17/6	14–41	9/12/4	10–32	8/11/12	20–46
163–203	39/44/39/52	2/0/2/3			7/9/7/9		5/5/5/12		11/9/11/24	
C2	76/76/76	2/2/2	2	67	9/8/9	7–11	0/0/0	0	5/5/5	7–11
204–279	76/76/76/76	2/3/2/2			5/5/5/9		0/0/0/0		7/8/9/6	
	25/25/25	0/0/0	0	0	5/5/5	8–20	0/0/0	0	2/2/2	8–12
280–304	25/25/25/25	0/2/0/0			2/2/2/4		0/0/0/0		0/3/0/2	
V3	26/26/26	3/1/3	1	16	7/5/8	19–34	2/5/5	7–22	5/0/3	0–57
305–330	27/36/27/26	6/7/6/4			7/7/7/9		6/2/6/3		12/8/12/15	
	65/65/65	4/4/4	4	57	11/11/11	12–17	5/5/5	6–8	8/8/9	12–25
331–395	66/66/66/65	6/7/6/4			11/8/9/10		5/5/4/4		17/10/10/10	
V4	19/19/19	1/4/1	0	0	5/7/7	2–11	0/0/0	0–6	9/13/9	10–20
396–414	23/18/23/19	8/2/6/3			3/4/3/1		0/1/0/0		12/7/12/13	
C3	44/44/44	4/4/4	4	100	0/1/0	0–11	0/0/0	0	5/5/5	5–11
415–458	44/44/44/44	4/4/4/4			2/3/5/2		0/0/0/0		5/2/5/4	
V5	11/11/11	2/2/2	0	0	10/10/10	1–90	4/5/5	0–45	10/6/10	54–100
459–469	11/12/11/12	1/0/1/0			8/2/1/7		0/0/0/1		11/9/9/11	
C4	41/41/41	4/4/4	4	100	12/12/12	29–32	18/19/19	43–46	12/11/12	24–29
470–510	41/41/41/41	4/4/4/4			13/13/13/12		19/19/19/19		12/12/12/10	
C5	106/106/106	0/0/0	0	100	5/5/5	5	8/8/8	5–7	5/5/5	6
511–616	105/106/105/106	0/0/0/0			5/5/5/5		8/8/8/5		6/5/6/5	
	38/38/38	0/0/0	0	0	7/5/5	13–23	4/4/4	10–13	3/3/3	7–10
616–653	38/38/38/38	2/0/2/2			7/5/7/4		3/4/3/4		3/3/3/2	
C6	92/92/92	6/6/6	6	67	34/34/34	34	29/20/23	20–29	35/35/36	32–40
654–745	92/92/92/92	6/8/6/10			34/34/34/34		19/22/24/28		33/32/37/40	
	111/111/111	5/5/5	2	39	20/20/19	15–18	13/13/13	10–12	9/9/9	7–10
746–856	111/111/111/111	5/2/5/7			20/17/20/20		12/12/12/11		10/8/10/11	

[a]Values are arranged in the following mode: WMJ1/ WMJ2/WMJ3
BH10/ ARV2/ LAV1A/HAT3.

protein complex of HIV. This polypeptide part is likely connected only via the ionic interactions of 14 positively charged amino acids in the carboxy-terminal region of gp120 (region C4) adjacent to the cleavage site with the negatively charged phosphate groups in the bilayer membrane. There are 18 conserved cysteine residues dispersed over the gp120 sequence (Fig. 1) which might also be involved in the complex formation of gp120 and gp41. A report describing the frequent loss of gp120 from the particles during purification and immunoelectron microscopy argues, however, against a direct involvement of disulfite bonds in the complex formation of gp120 and gp41 (17).

The transmembrane polypeptide gp41 contains three stretches of hydrophobic amino acids. Directly after the cleavage site there are about 60 hydrophobic residues (amino acids 511 to 571) which are interrupted by a stretch of about 10 amino acids that are predicted to have a high surface probability and, in comparison with the surrounding residues, a higher hydrophilicity and flexibility and some alterations in β-turns. From the patterns of other viral (22, 32) or

TABLE 2. Antigenic epitopes in the various strains

Epitope no.	Region	Amino acid residues in virus strain:						
		WMJ1	WMJ2	WMJ3	BH10	ARV2	LAV1A	HAT3
I	V1	137–154	137–154	137–153	137–157	137–158	137–163	137–158
II	V2	186–203	186–203	185–202	188–203	189–209	195–209	189–217
III	C2	232–246	232–246	228–245	232–246	240–253	240–253	249–261
IV	V3	300–320	300–319	299–318	300–321	300–327	308–328	316–355
V		358–375	257–374	356–373	360–377	367–384	367–384	375–391
VI	V4	394–412	393–411	392–410	397–418	404–420	404–425	410–427
VII	C3	445–458	444–457	443–456	451–464	453–466	457–470	459–472
VIII	V5	459–469	458–468	457–467	465–475	467–478	471–481	473–484
IX	C4	470–483	469–482	468–481	476–489	478–492	482–496	485–499
X		611–637	610–636	609–635	616–643	620–646	623–649	627–653
XI	C6	724–745	723–744	722–743	729–750	733–754	736–757	740–761

membrane proteins, e.g., bacteriorhodopsin (14), acetylcholine receptor (15), and erythrocyte band III (26), such regions are known to represent sequential transmembrane-spanning regions (TM1 and TM2 [Fig. 3]). A domain of hydrophilic, flexible, and charged amino acids (amino acids to 572 to 594) linked to the hydrophobic region may represent a hydrophilic anchor sequence.

Transmembrane region TM3, which was suggested from amino acids 670 to 695, with its hydrophilic anchor has already been mentioned. In transmembrane regions TM2 and TM3 three charged amino acids (arginine at positions 557 and 707 and glutamic acid at position 560) are identifiable. Those charged amino acids were also reported in other transmembrane regions in multi-spanning proteins and do not lead to disturbance of the overall hydrophobic character of a transmembrane region (48). Between TM2 and TM3 hydrophobic regions from the N terminus of gp41 are 90 amino acid residues (580 to 670) which should be facing the outer side of the lipid bilayer. Within this region are four conserved glycosylation sites and a good antigenic probability (epitope X). These considerations were combined into a three-dimensional model of the HIV envelope protein complex (Fig. 3). Alternatively, TM1 and TM2 may represent an apolar hydrophobic stretch of amino acids located outside the viral membrane and may play a role similar to the paramyxovirus fusion protein by penetrating into the membrane of the fusion partner (31). The same could occur if the C-terminal part of gp41, including or not including TM3, would be directed to the outer membrane or be in equilibrium between those two orientations. This possibility is supported by the observation that antiserum against a peptide from amino acids 725 to 752 gives cell surface labeling and is neutralizing (5a). The predicted antigenic determinants are not affected by the discussed models.

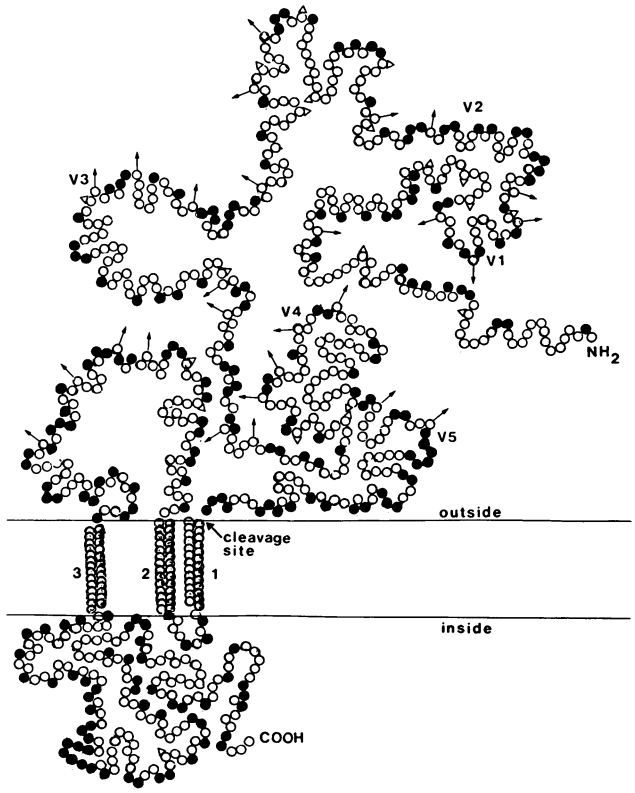

FIG. 3. Suggestion for the tertiary structure of envelope protein gp160 beginning at amino acid 38. Each circle represents one amino acid residue, shaded areas indicate transmembrane-spanning regions. Variable regions are indicated. ●, Polar amino acids; arrows, N-glycosylation sites; ⬡, cysteine residues.

DISCUSSION

Previous studies have shown that variability in the genomes of different HIV isolates is a prominent feature of this group of viruses (21, 40, 43, 50). In the present study we analyzed the deduced amino acid sequences of seven HIV isolates and predicted the secondary structures in combination with the values for hydrophilicity, flexibility, and surface probability, which resulted in the identification of potentially antigenic epitopes.

Analysis of the primary and secondary structure of polypeptide shows that antigenic determinants are often found in loop like structures on the surface of a molecule that have been identified as biologically important regions in several other viral membrane or capsid proteins (8, 12, 19, 28, 30, 32, 47). Antigenic sites formed by the folding of the amino acid chain to tertiary structures contribute to immunological activities; this cannot be predicted by calculations similar to those applied here. By our methods, 11 epitopes could be identified and characterized in the envelope protein complex of HIV, 9 of which are located in the exterior protein gp120. A comparison of the amino acid sequences of the seven viral strains led to the subdivision of gp120 into highly variable and conserved regions. Five highly variable regions could be identified, all of which coincided with the predicted epitopes I, II, IV, VI, and VIII. These epitopes, due to their concentration of β-turns and hydrophilic amino acids, have a very high potential for antigenicity. Genomic variation primarily occurs in the exterior envelope sequences and corresponds to predicted epitopes of different AIDS virus strains, which differ also in sequential isolates from the same patient (epitopes I, II, IV, and VI), suggesting that immune selection may play an important role in the generation of variant virus strains. Two related retroviruses, equine infectious anemia virus and visna virus (18, 41, 42), show similar progressive changes in their envelope genes. There is evidence that these changes do, in fact, lead to substantial changes in the antigenic properties of the envelope which may reflect immune selective pressure exerted by the host (7, 29, 38).

Whereas the variable regions of the exterior protein gp120 possess properties typical for antigenic sites, the conserved regions are generally hydrophobic and lack areas with a high number of β-turns. There are only three exceptions: epitope III in region C2, epitope VII in region C3, and epitope IX in region C4. Epitope III has only two amino acids predicted in β-turn configurations and must be considered of minor antigenic potential, in comparison with the other epitopes. Epitope VII shows many β-turns; these however are only slightly hydrophilic and have no high values for surface probability. Epitope XI in C4 (amino acids 470 to 483) consists of a stretch of hydrophilic, flexible amino acids in β-turn areas with high surface probability; this region is adjacent to the cleavage site, and a synthetic peptide corresponding to that region reacted with about 80% of the sera from HIV-positive individuals (S. Modrow, unpublished data). However, it is unclear if this epitope is conserved in all virus strains. It is also possible that conserved region C4 might contribute to biologically important functions of the virus particle such as virus cell adsorption and connection of

gp120 to the envelope membrane; both functions could protect this region from selective immune pressure.

The highly conserved protein structures together with 18 cysteine residues conserved in the seven strains may contribute to the stability of a constant core structure of gp120, with variable, highly antigenic regions looping out.

In contrast to the exterior protein gp120, transmembrane polypeptide gp41 contains no region of high variability and consists mainly of regions with hydrophobic areas which may be arranged into three transmembrane regions. Between these transmembrane regions is a glycosylated region with good predicted antigenicity (epitope X), which may be located outside the viral envelope and thus may be recognized by the immune system of the host. This region has only slight variability and may contribute to biologically important functions as well. It thus could represent a valuable diagnostic antigen. It has been shown (4, 5) that segments of the protein, including epitope X, when produced as recombinant gene products in bacteria, are recognized by HIV-positive sera in enzyme-linked immunosorbent assays and Western blots; preliminary results with a synthetic peptide from this region give similar results (Modrow, unpublished). The last epitope, epitope XI, consists of stretches of highly hydrophilic, flexible amino acids with high values for surface probability and several β-turns. This region may represent the hydrophilic anchor sequences located adjacent to transmembrane region TM3 and thus should be located inside the cell.

The observation that HIV binds to the CD4 molecule of T cells (27) which is supposed to bind to molecules of the major histocompatibility complex class II during stimulation of immune responses may suggest functional similarities of yet unidentified regions of gp120 or gp41 with those of the major histocompatibility complex class II molecules. Our analysis of constant and variable regions may well identify candidates for structures of dominant immunogenicity for humoral immune response. This does not, however, exclude the possibility that during biological degradation of viral structures other segments, including constant regions of the viral envelope proteins, lead to formation of specific antibody responses in patients. Such antibodies, when detected by in vitro antigens not presented in their native configuration, can be of diagnostic value. In addition, these antibodies may also have neutralizing activity, since the respective sequences recognized may be located in accessible areas of the envelope or be accessible during molecular mobility.

Cellular immune mechanisms have not been addressed so far and are not predictable by our protein analysis. However, in this case, sequential rather than complex antigens seem to be of particular importance. Selective priming of cell-mediated immunity with epitopes selected from detailed analysis of the viral envelope should open new approaches for the control of HIV-related disease.

LITERATURE CITED

1. **Allen J. S., J. E. Coligan, F. Barin, M. F. McLane, J. G. Sodroski, C. A. Rosen, W. A. Haseltine, T. H. Lee, and M. Essex.** 1985. Major glycoprotein antigens that induce antibodies in AIDS patients are encoded by HTLV-III. Science **228**:1091–1093.
2. **Atassi, M. Z.** 1978. Precise prediction of the entire antigenic structure of lysozyme. Molecular features of protein antigenic structures and potential of "surface stimulation" synthesis—a powerful new concept for protein binding sites. Immunochemistry **15**:909–936.
3. **Barin, F., M. F. McLane, J. S. Allan, T. H. Lee, J. E. Groopman, and M. Essex.** 1985. Virus envelope protein of HTLV-III represents major target antigen for antibodies in AIDS-patients. Science **228**:1094–1096.
4. **Cabradilla, C. D., J. E. Groopman, J. Lanigan, M. Renz, L. A. Laskey, and D. J. Capon.** 1986. Serodiagnosis of antibodies to the human AIDS retrovirus with a bacterially synthesized env polypeptide. Biotechnology **4**:128–133.
5. **Chang, T. W., K. Ikunoshin, S. McKinney, P. Chanda, A. D. Barone, F. Wong-Staal, R. C. Gallo, and N. T. Chang.** 1985. Detection of antibodies to human T-cell lymphotropic virus III (HTLV-III) with an immunoassay employing a recombinant *Escherichia coli*-derived viral antigenic peptide. Biotechnology **3**:905–909.

5a. **Chanh, T. C., G. R. Dreesman, P. Kanda, G. Linette, J. T. Sparrow, D. D. Ho, and R. C. Kennedy.** 1986. Induction of anti-HIV neutralizing antibodies by synthetic peptides. EMBO J. **5**:3065–3073.

6. **Chou, P. Y., and G. D. Fasman.** 1974. Prediction of protein conformation. Biochemistry **13**:222–245.
7. **Clements, Y. E., F. S. Pedersen, O. Narayan, and W. A. Haseltine.** 1980. Genomic changes associated with antigenic variation of visna virus during persistent infection. Proc. Natl. Acad. Sci. USA **77**:4454–4458.
8. **Cohen, G. H., B. Dietzschold, M. Ponce de Leon, D. Long, E. Golub, A. Varrichio, L. Pereira, and R. J. Eisenberg.** 1984. Localization and synthesis of an antigenic determinant of herpes simplex virus glycoprotein D that stimulates the production of neutralizing antibody. J. Virol. **49**:102–108.
9. **Crowl, R., K. Ganguly, M. Gordon, R. Conroy, M. Schaber, R. Kramer, G. Shaw, F. Wong-Staal, and E. P. Reddy.** 1985. HTLV-III *env* gene products synthesized in *E. coli* are recognized by antibodies present in the sera of AIDS-patients. Cell **41**:979–986.
10. **Davis, N. G., J. D. Boeke, and P. Model.** 1985. Fine structure of a membrane anchor region domain. J. Mol. Biol. **181**:111–121.
11. **DiMarzo-Veronese, F., A. L. deVico, T. D. Copeland, S. Oroszlan, R. C. Gallo, and M. G. Sarngadharan.** 1985. Characterization of gp41 as the transmembrane protein coded by the HTLV-III/LAV envelope gene. Science **229**:1402–1405.
12. **Eisenberg, R. J., D. Long, M. Ponce de Leon, J. T. Matthews, P. G. Spear, M. G. Gibson, L. A. Lasky, P. Berman, E. Golub, and G. H. Cohen.** 1985. Localization of epitopes of herpes simplex virus type 1 glycoprotein D. J. Virol. **53**:634–644.
13. **Emini, E. A., J. V. Hughes, D. S. Perlow, and J. Boger.** 1985. Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide. J. Virol. **55**:836–839.
14. **Engleman, D., A. Goldman, and T. Steitz.** 1982. The identification of helical segments in the polypeptide chain of bacteriorhodopsin. Methods Enzymol. **88**:81–98.
15. **Finer-Moore, J., and R. M. Stroud.** 1984. Amphipathic analysis and possible formation of the ion channel in acetylcholine receptor. Proc. Natl. Acad. Sci. USA **81**:155–159.
16. **Garnier, J., D. J. Osguthorpe, and B. Robson.** 1978. Analysis of the accuracy and implication of simple methods for predicting the secondary structure of globular proteins. J. Mol. Biol. **120**:97–120.
17. **Gelderblom, H. R., H. Reupke, and G. Pauli.** 1985. Loss of envelope antigens of HTLV-III/LAV, a factor in AIDS-pathogenesis? Lancet **ii**:1016–1017.
18. **Gonda, M. A., F. Wong-Staal, R. C. Gallo, J. E. Clements, O. Narayan, and R. V. Gilden.** 1985. Sequence homology and morphologic similarities of HTLV-III and visna virus, a pathogenic lentivirus. Science **227**:173–177.
19. **Gunn, P. R., F. Sato, K. F. H. Powell, A. R. Bellamy, J. R. Napier, D. R. K. Harding, W. S. Hancock, L. J. Siegman, and G. W. Both.** 1985. Rotavirus neutralizing protein VP7: antigenic determinants investigated by sequence analysis and peptide synthesis. J. Virol. **54**:791–797.
20. **Hahn, B. H., M. A. Gonda, G. M. Shaw, M. Popovic, J. Hoxie, R. C. Gallo, and F. Wong-Staal.** 1985. Genomic diversity of the AIDS virus HTLV-III: different viruses exhibit greatest divergence in their envelope genes. Proc. Natl. Acad. Sci. USA **82**:4813–4817.
21. **Hahn, B. H., G. M. Shaw, M. E. Taylor, R. R. Redfield, P. D.**

Markham, S. Z. Salahuddin, F. Wong-Staal, R. C. Gallo, E. S. Parks, and W. P. Parks. 1986. Genetic variation in HTLV-III/LAV over time in patients with AIDS or at risk of AIDS. Science **232**:1548–1554.

22. **Hennessey, K., S. Fennewald, M. Hummel, T. Cole, and E. Kieff.** 1984. A membrane protein encoded by Epstein-Barr virus in latent growth-transforming infection. Proc. Natl. Acad. Sci. USA **81**:7207–7211.
23. **Hopp, T. P., and K. R. Woods.** 1981. Prediction of protein antigenic determinants from amino acid sequences. Proc. Natl. Acad. Sci. USA **78**:3824–3828.
24. **Karplus, P. A., and G. E. Schulz.** 1985. Prediction of chain flexibility in proteins. Naturwissenschaften **72**:212–213.
25. **Kennedy, R. C., R. D. Henkel, D. Pauletti, Y. S. Allan, T. H. Lee, M. Essex, and G. R. Dreesman.** 1986. Antiserum to a synthetic peptide recognizes the HTLV-III envelope glycoprotein. Science **231**:1556–1559.
26. **Kopito, R., and H. F. Lodish.** 1985. Primary structure and transmembrane orientation of the murine anion and exchange protein. Nature (London) **316**:234–238.
27. **McDougal, J. S., M. S. Kennedy, J. M. Sligh, S. P. Cort, A. Mawle, and J. K. A. Nicholson.** 1986. Binding of HTLV-III/LAV to $T4^+$ T-cells by a complex of the 110kD viral protein and the T4 molecule. Science **231**:382–385.
28. **Modrow, S., and H. Wolf.** 1986. Characterization of two related Epstein-Barr virus-encoded membrane proteins that are differentially expressed in Burkitt lymphoma and in vitro transformed cell lines. Proc. Natl. Acad. Sci. USA **83**:5703–5707.
29. **Montelaro, R. C., B. Parekh, A. Orego, and C. J. Issel.** 1984. Antigenic variation during persistent infection by equine infectious anemia virus and retrovirus. J. Biol. Chem. **259**: 10539–10544.
30. **Motz, M., J. Fan, R. Seibl, W. Jilg, and H. Wolf.** 1986. Expression of the Epstein-Barr virus 138 kDa early protein in *Escherichia coli* for the use as antigen in diagnostic tests. Gene **42**:303–312.
31. **Paterson, R. G., T. J. R. Harris, and R. A. Lamb.** 1984. Fusion proteins of the paramyxovirus simian virus 5: nucleotide sequence of mRNA predicts a highly hydrophobic glycoprotein. Proc. Natl. Acad. Sci. USA **81**:6706–6710.
32. **Pellett, P. E., K. G. Kousoulas, L. Pereira, and B. Roizman.** 1985. Anatomy of herpes simplex virus 1 strain F glycoprotein B gene: primary sequence and predicted protein structure of the wild type and of monoclonal antibody-resistant mutants. J. Virol. **53**:243–253.
33. **Popovic, M., M. G. Sarugadharan, E. Read, and R. C. Gallo.** 1984. A method for detection, isolation and continuous production of cytophathic human T-lymphotropic retroviruses of the HTLV family (HTLV-III) from patients with AIDS and pre-AIDS. Science **224**:497–500.
34. **Queen, C. L., and L. J. Korn.** 1980. Computer analysis of nucleic acids and proteins. Methods Enzymol. **65**:595–609.
35. **Ratner, L., R. C. Gallo, and F. Wong-Staal.** 1985. HTLV-III, LAV, and ARC are variants of the same AIDS virus. Nature (London) **313**:636–637.
36. **Ratner, L., W. Haseltine, R. Patarca, K. Livak, B. Starcich, S. Josephs, R. E. Doran, J. A. Rafalski, E. A. Whitehorn, K. Baumeister, L. Ivanoff, S. R. Petteway, Jr., I. A. Lautenberger, T. S. Papas, J. Ghrayeb, J. Chang, R. C. Gallo, and F. Wong-Staal.** 1985. Complete nucleotide sequence of the AIDS virus, HTLV-III. Nature (London) **313**:277–284.
37. **Robert-Guroff, M., M. Brown, and R. C. Gallo.** 1985. HTLV-neutralizing antibodies in AIDS and ARC. Nature (London) **316**:72–74.
38. **Salinovich, O., S. L. Payne, R. C. Montelaro, K. A. Hussain, C. J. Issel, and K. L. Schnorr.** 1986. Rapid emergence of novel antigenic sites and genetic variants of equine infectious anemia virus during persistent infection. J. Virol. **57**:71–80.
39. **Sanchez-Pescador, R., M. D. Power, J. P. Barr, K. S. Steimer, M. M. Stempien, S. L. Brown-Shimer, W. W. Gee, A. Renard, A. Randolph, J. A. Levy, D. Dina, and P. A. Luciw.** 1985. Nucleotide sequence and expression of an AIDS-associated retrovirus (ARV-2). Science **227**:484–492.
40. **Shaw, G. M., B. H. Hahn, S. K. Arya, J. E. Groopman, R. C. Gallo, and F. Wong-Staal.** 1984. Molecular characterization of human T-cell leukemia (lymphoma) virus type III in the acquired immune deficiency syndrome. Science **226**:1165–1171.
41. **Shaw, G. M., M. E. Harper, B. H. Hahn, L. G. Epstein, C. D. Gaidusek, R. W. Price, B. A. Navia, C. K. Petito, C. J. O'Hara, E.-S. Cho, J. M. Oleska, F. Wong-Staal, and R. C. Gallo.** 1985. HTLV-III infection in brains of children and adults with AIDS-encephalopathy. Science **227**:177–182.
42. **Sonigo, P., M. Alizon, K. Staskus, D. Klatzmann, S. Cole, O. Danos, E. Retzel, P. Tiollais, A. Haase, and S. Wain-Hobson.** 1985. Nucleotide sequence of the visna lentivirus: relationship to the AIDS virus. Cell **42**:369–382.
43. **Starcich, B. R., B. H. Hahn, G. M. Shaw, P. D. McNeely, S. Modrow, H. Wolf, E. S. Parks, W. P. Parks, S. F. Josephs, R. C. Gallo, and F. Wong-Staal.** 1986. Identification and characterization of conserved and variable regions in the envelope gene of HTLV-III/LAV, the retrovirus of AIDS. Cell **45**:637–648.
44. **Tainer, J. A, E. D. Getzoff, H. Alexander, R. A. Houghten, A. J. Olson, R. A. Lerner, and W. A. Hendrickson.** 1984. The reactivity of anti-peptide antibodies is a function of the atomic mobility of sites in a protein. Nature (London) **312**:127–134.
45. **Wain-Hobson, S., P. Sonigo, O. Danos, S. Cole, and M. Alizon.** 1985. Nucleotide sequence of the AIDS virus, LAV. Cell **40**:9–17.
46. **Weiss, R. A., P. R. Chapham, R. Cheingson-Popov, A. G. Dayleish, C. A. Carne, I. V. D. Weller, and R. S. Tedder.** 1985. Neutralizing antibodies to human T-cell lymphotropic virus type III. Nature (London) **316**:69–72.
47. **Westhof, E., D. Altschuh, D. Moras, A. C. Bloomer, A. Mondragon, A. Klug, and M. H. V. van Regenmortel.** 1984. Correlation between segmental mobility and the location of antigenic determinants in proteins. Nature (London) **311**: 123–126.
48. **Wickner, W. T., and H. F. Lodish.** 1985. Multiple mechanisms of protein insertion into and across membranes. Science **230**:400–407.
49. **Wilbur, W. J., and D. J. Lipman.** 1983. Rapid similarity searches of nucleic acid and protein data banks. Proc. Natl. Acad. Sci. USA **80**:726–730.
50. **Wong-Staal, F., G. M. Shaw, B. H. Hahn, S. Z. Salahuddin, M. Popovic, P. D. Markham, R. Redfield, and R. C. Gallo.** 1985. Genomic diversity of human T-lymphotropic virus type III. Science **229**:759–762.