Volume 17 Number 8 1989 Nucleic Acids Research

Sequence divergence and open regions of RNA secondary structures in the envelope regions of the 17 human immunodeficiency virus isolates

Shu-Yun Le[1,3], Jih-H.Chen[2], Devjani Chatterjee[1] and Jacob V.Maizel[1]

[1]Laboratory of Mathematical Biology, Division of Cancer Biology and Diagnosis, National Cancer Institute, National Institutes of Health, Bldg 469, Rm 151, Frederick, MD 21701, [2]Advanced Scientific Computer Laboratory, Program Resources, Inc., NCI/FCRF, Frederick, MD 21701, USA and [3]Shanghai Institute of Biochemistry, Chinese Academy of Sciences, Shanghai 200031, China

Received August 26, 1988; Revised and Accepted March 8, 1989

ABSTRACT

Genetic variation during the course of infection of an individual is a remarkable feature of the acquired immune deficiency syndrome(AIDS) disease. This variation has been studied for the envelope protein encoding regions of seventeen different sequences from various isolates of human immunodeficiency virus (HIV) using multiple sequence comparison and calculation of variability. The open regions with little intramolecular base pairing in these envelope sequences are predicted by a recently developed statistical method. The minimum length L for a run of hypervariable sites, conserved sites, or open regions that gives significance at the 1%(or 0.1%) level is then determined by a scan statistical method. The results show that significant clusters of open regions predicted at the RNA level correlate with significant clusters of hypervariable sites in the HIV envelope gene. Those significant genomic variations in HIVs seem to be manifested mainly in the extracellular portion of the envelope protein. Twelve potential antigenic determinants are predicted using an antigenic index method. Interestingly, the majority of the significant hypervariable regions in the exterior envelope protein (gp120) were predicted potential epitopes.

INTRODUCTION

Human immunodeficiency virus(HIV) is the etiological agent for the acquired immunodeficiency syndrome(AIDS). Patients afflicted with AIDS or with other enveloped viruses demonstrate antibodies directed against the antigenic determinants present on the envelope glycoprotein. The HIV envelope region shows extreme variability in its sequence in independent isolates and also in sequential viral isolates from the same patient(1-5,11).

The correlation between hypervariable sites and open regions with little intramolecular base pairing in the ARV, LAV, and HTLV-III isolates has been noted(6). This mechanism probably allows the virus to elude immune surveillance of the host and is important for the biology and pathogenicity of HIV(7). Various workers have compared and analyzed the

sequence of the envelope region from a number of isolates of HIV(3-5). The variable regions in the sequence have been shown to coincide with antigenic sites(3) and thus play an important role in the immunogenicity and pathobiology of the virus. However, the sequence comparisons are performed on only seven or fewer than seven isolates, three of which are relatively closely related to each other(3). It is unclear if the coincidence is preserved in all virus strains of HIV. Moreover, the mechanism by which sequence variations occur in specific regions of the virus genome remains obscure.

This study tests the correlations of the sequence divergence with the potentially unstable local RNA folding regions in the envelope coding regions of seventeen isolates of HIV. An analysis of the nucleotide and deduced amino acid sequences of the complete envelope RNAs is presented. Using computer assisted methods, the open regions with little intramolecular base pairing are derived by a recent statistical method(8) for analyzing potential RNA folding based on the comparison of the predicted lowest free energy of a real biological sequence with the mean of a number of randomized sequences of the same nucleotide composition. The statistical method has been applied to 16S rRNAs and viroids whose secondary structures have been established and published, and this has provided validation of the method(9). The hypervariable and conserved sites among these HIV envelope regions are derived by multiple sequences comparison(10) of seventeen different isolates of human immunodeficiency virus and calculation(11,12) of variability at different sequences. The minimum length L for a run of hypervariable sites, consensus sites and open regions that gives significance at the 1%(or 0.1%) level is then determined by a scan statistical method(13). Thus, the significant runs of hypervariable sites, conserved sites and open regions can be identified in the envelope regions of seventeen HIV isolates. The results show that these significant clusters of hypervariable sites are found to correlate with the significant clusters of open regions of RNA secondary structure in the HIV envelope coding regions. These significant genomic variations of the HIV seem to be manifested mainly in the extracellular portion of the envelope protein.

METHOD

Calculations of the Sequence Variability and Significant Open Region

All analyzed RNA and protein sequences of various isolates of HIV are derived from available genomic DNA sequences in the HIV sequence

database(2). Their locus names in the database are hivhxb2, hivbh102, hivbh8, hivhxb3, hivpv22, hivbru, hivmal, hiveli, hivsf2, hivwmj22, hivcdc42, hivz6, hivz3, hivny5, hivmn, hivrf, and hivsc. In the calculation of the variability of these envelope protein sequences(or RNA sequences), first these seventeen sequences from different isolates of HIV are aligned using Martinez's GENALIGN (general multiple sequence alignment) program(10). For each position i of the alignments mentioned above, variability V(i) of the sequences is then determined as follows(12): $V(i)=d_i/f_i$, where d_i is the number of different residues and f_i is the fraction of the most preserved residues at position i. In the calculation, gaps in the alignments are considered as a special kind of residue which is dealt with in the same way as normal residues. Obviously, the hypervariable and conserved sites can be detected using the variability distribution of the sequences.

The term "significant open region" means that the structure of the biological segment sequence is very unstable, and the structure of the random permutation of the biological segment is more stable than that of the natural sequence. The significance of the structure (segment score) is measured by $score=(e_b-e_r)/std$, where e_b is the lowest free energy for the real segment sequence, e_r the average lowest free energy of the randomized segments and std the standard deviation of the free energies of the randomized sequences. In practice, we sort significant open regions by the following procedure: First, we calculate the lowest free energies of the segment with a suitable length(for example length=60) and calculate the segment scores along the testing sequence from the 5' to 3' end using the RANFOLD program(8). The frequency distribution of the lowest free energy of the segment thus can be determined. The threshold of the free energy estimating the open region of the testing sequence can be decided according to the frequency table and the significant open regions can be identified by sorting segment scores and the lowest free energies of the segment in the sequence.

Estimations of Mean Free Energies

We have found that large savings in computational time with acceptable precision can be obtained using a method in which the segment score is based on a calculation using the predicted free energy of the natural segment and an estimated mean free energy (E_r) and standard deviation(SD_r) based on the segment length and composition instead of performing a Monte Carlo simulation(15). That is

$$E_r = L(a + b/L + c/L^2) \qquad (1)$$

The constants a, b and c are derived by solving the equation,

$$HX = E \qquad (2)$$

where

$$H = \begin{bmatrix} 1 & \lambda_1 & \lambda_1^2 \\ 1 & \lambda_2 & \lambda_2^2 \\ 1 & \lambda_3 & \lambda_3^2 \end{bmatrix}, \quad E = \begin{bmatrix} e_r(1) \\ e_r(2) \\ e_r(3) \end{bmatrix}, \quad X = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

λ_1=1/50, λ_2=1/100, λ_3=1/150;
$e_r(1)$, $e_r(2)$, $e_r(3)$ corresponding to l = 50, 100, 150 are calculated by the formula:

$$e_r(j) = l_j (\Sigma d_i f_i) \qquad (3)$$

from i=1 to 10, where j=1, 2, 3; l_1=50, l_2=100, and l_3=150; $f_1=f_A$, $f_2=f_C$, $f_3=f_G$, $f_4=f_U$, $f_5=\min\{f_C,f_G\}$, $f_6=\min\{f_A,f_U\}$, $f_7=f_Cf_G$, $f_8=f_Af_U$, $f_9=\min\{f_U,f_G\}$, $f_{10}=f_Gf_U$ (f_x denotes the fraction of nucleotide x in the segment), and d_i are the empirical coefficients derived by the least squares method and listed elsewhere(15). The standard deviation (SD_r) of the random sample set is estimated by the same procedure as for mean free energy E_r. Thus, the calculation of the segment score is speeded up and completed in a few minutes.

Significant Runs

Let N denote the total number of residues in the sequence(amino acid sequence or nucleotide sequence) and M be a distinct pattern, either a hypervariable site or an open region with little intramolecular base pairing. According to a method known as scan statistics, the probability P_r of observing a run of length, L=x-lnN/lnf, of the pattern M, where N>=150, f is the corresponding probability of sampling M (in practice, f is considered as the fraction of occurrence of the pattern M in the sequence), is satisfied by the following inequality(13)

$$1-\exp\{-(1-f)f^{x+1}\} < P_r < 1-\exp\{-(1-f)f^x\}.$$

Here, we may identify significant runs by the following procedure(16) i) we solve the equation, $1-\exp\{-(1-f)f^x\} = 0.01$ for x, i.e. we have $x=\ln\{\ln(1-0.01)/(f-1)\}/\ln f$. ii) the minimum length L_{min} for a run of the

Table 1: Identified Open Regions with Little Intramolecular Base Pairing in the Envelope RNA of RF Isolate

Starting Position No.	*Lengths of Run*	*Significant Clusters*
208-211	2	
271-277	3	
391-412	8	391-471
487-493	3	
817-847	11	817-906
928-988	21	928-1047
1030-1060	11	1030-1119
1330-1363	12	1330-1422
1390-1417	10	1390-1476
1840-1843	2	
1849-1876	11	1849-1935
1906-1954	17	1906-2014
1969-1993	9	1969-2052
2041-2068	10	2041-2127
2215-2218	2	
2476-2479	2	

Table 1. The numbers in the column 1 denote the starting positions of the overlapping segments of 60 nucleotides length where the segment score is greater than 0 and the lowest free energy of the segment greater than -5 kcal/mol. The length in the column 2 means the observed length for a run of each distinct open region. In the calculation the data(energies and segment scores) have been averaged at three successive positions and then smoothed in overlapping groups of five.

pattern M that gives significance at the 1% level is determined as the smallest integer larger than x-lnN/lnf. For example, the alignment of the envelope protein derived by aligning seventeen amino acid sequences from different isolates of HIV consists of 924 residues(including gaps). In the alignment the fraction of the occurrence of the pattern M is 0.282, where the pattern M is the hypervariable cluster with variability larger than 4.0. For f=0.282, we calculate x to be 3.37 at the 1% level, and 3.37≃ln924/ln0.282=8.76. So, we shall consider a run of the hypervariable site for which variability is larger than 4.0 in those envelope protein sequences significant only if it exceeds length 9. The minimum length 9 which gives significance at the 1% level is thus determined for a run of the specific pattern M.

RESULTS and DISCUSSIONS

Significant Open Regions with Little Intramolecular Base Pairings

To estimate the significant open regions in seventeen sequences of the envelope regions of different isolates of HIV, RANFOLD simulations(8) are run, in which the window sizes of the overlapping segments are taken as 60

Table 2: Significant Clusters of the Open Region in Envelope RNAs of Seventeen HIV Isolates

Isolates	*Significant Clusters*						*Lmin of Runs*	
	I	II	III	IV	V	VI	.01	.001
HXB2	346-447	502-597	856-930	952-1095	1885-2031	2044-2121	7	8
			913-987		1990-2082			
BH8	346-447	508-597	913-987	1006-1092	1831-1905	1975-2067	6	7
					1870-2016			
SF2	280-351	431-509	895-1018	1153-1228	1852-2023		7	8
		466-553	985-1087		1996-2086			
CDC42	340-441	526-606	952-1047		1885-2016	2020-2121	7	8
			1015-1110		1972-2052			
WMJ22	244-442	505-589	781-860		1780-1870		8	9
			820-898		1828-2060			
			886-1063					
RF	391-471	817-906	928-1047	1330-1422	1849-1935	1969-2052	7	8
			1030-1119	1390-1476	1906-2014	2041-2127		
Z6	370-445	802-877	1159-1264	1852-1984	2041-2116	2440-2515	6	8
	430-568			1942-2023				
ELI	232-336	493-567		1006-1080	1843-1920	1945-2016	5	6
	391-468					2008-2079		
MAL	367-468	670-744	970-1053	1327-1422	1888-1983	2185-2259	6	7
			1021-1095		1954-2055			
					2011-2121			
NY5	235-339	430-510	784-855	883-1011	1288-1359	1831-1962	8	9
	337-414			970-1062		1924-2061		
PV22	346-447	508-597	850-927	952-1095	1849-1920	2044-2100	7	8
					1888-2025	1996-2085		
BH102	346-447	508-597	850-927	952-1092	1846-1923	2044-2121	7	8
					1885-2031	1990-2082		
Z3	235-331	385-453	778-860	991-1075	1390-1477	1837-1918	7	9
	25-120	446-523	889-964	1081-1170		1891-2017		
				1135-1219	1891-2017	2494-2572		
MN	431-500	673-753	901-978	1174-1275	1315-1395	1852-1980	7	8
			937-1086			1948-2031		
		m				1999-2088		
SC	278-350	421-494	916-1045	1138-1213	1339-1450	1849-2031	7	8
		482-556	1006-1087			1996-2086		
BRU	412-483	523-612	865-939	1021-1110	1861-1938	2011-2100	6	7
			928-1002		1897-1986	1936-2046		
HXB3	346-447	508-597	913-987	1006-1095	1885-2031	2044-2100	6	7
						1990-2082		

Table 2. The numbers in the right two columns denote the minimum length L_{min} for a run of the open region of 60 nucleotides length that gives significance at the 1% and 0.1% level. The nucleotide positions in Columns I-VI are numbered according to the envelope RNA sequence of each isolate.

bases. In the calculations the distribution of the segment score is derived empirically as described in the methods and (15). For each envelope RNA sequence we compute the distributions of the lowest free energy of the secondary structure and the segment score for a segment with 60 nucleotides along their RNA strand. As an example, the cumulative percent of the lowest free energy of the RF(hivrf) isolate larger than -5.0

kcal/mol in the frequency distribution is 16.6%. The threshold energy -5.0 kcal/mol is considered to be an indication for an unstable folding region. The open regions in the RF envelope sequence can be identified according to two conditions, score >0 and E(free energy) > -5.0 kcal/mol and are listed in Table 1(the others are not shown here). The minimum length L_{min} for a run of the significant open region mentioned above that gives significance at the 1% or 0.1% level is determined using a scan statistical method(see Method section). We have $L_{min}(0.01)=7$ or $L_{min}(0.001)=8$ for the RF envelope sequence. This means that these 7 or more successive overlapping segments with 60 bases can form a significant cluster of the open region with little intramolecular base pairing. Similarly, all minimum lengths for a run of the open region in seventeen RNAs of HIV envelope are listed in Table 2, which gives significance at the 1% or 0.1% level. The data of significant clusters of the open region in seventeen RNAs of HIV envelope are also listed in Table 2.

Sequence Divergences and Significant Clusters of Hypervariable and Consensus Sites

Two alignments of the envelope nucleotide and amino acid sequences of various isolates of BH10, BH8, BRU, CDC-451, ELI, HXB2, HXB3, MAL, MN, NY-5(1984), PV22, RF, SC, ARV-2/SF2, WMJ2, Zaire-3(Z3) and Zaire-6(Z6) of HIV are obtained independently using the GENALIGN program(10). In the alignment of the amino acid sequences an additional nucleotide in HIVZ3(amino acid 477) is removed, which causes a frame shift and an immediate termination of the open reading frame(2). Likewise, the amino acid alignment is done on the HIVNY5 sequence deduced after insertion of the deleted base(a T at position 438 of amino acid)(2). The graphical representations of the envelope variability are depicted in Figs.1a-b. Fig.1a is a plot of the variability of HIV envelope protein. For each position x of the alignment of seventeen amino acid sequences, variability V(x) is calculated(see Method section). The variability data are then smoothed in overlapping groups of five successive positions. In Fig.1b the variability data of envelope nucleotide sequences of seventeen different isolates of HIV are averaged at three successive positions and then smoothed in overlapping groups of five positions. From Figs.1a-b, we can clearly identify conserved sites, with little or no genetic variation, and hypervariable sites. It is clear that the mutations found in the nucleotide sequences are not silent substitutes. The distribution of the variability in nucleotide sequences tends towards that in amino acid

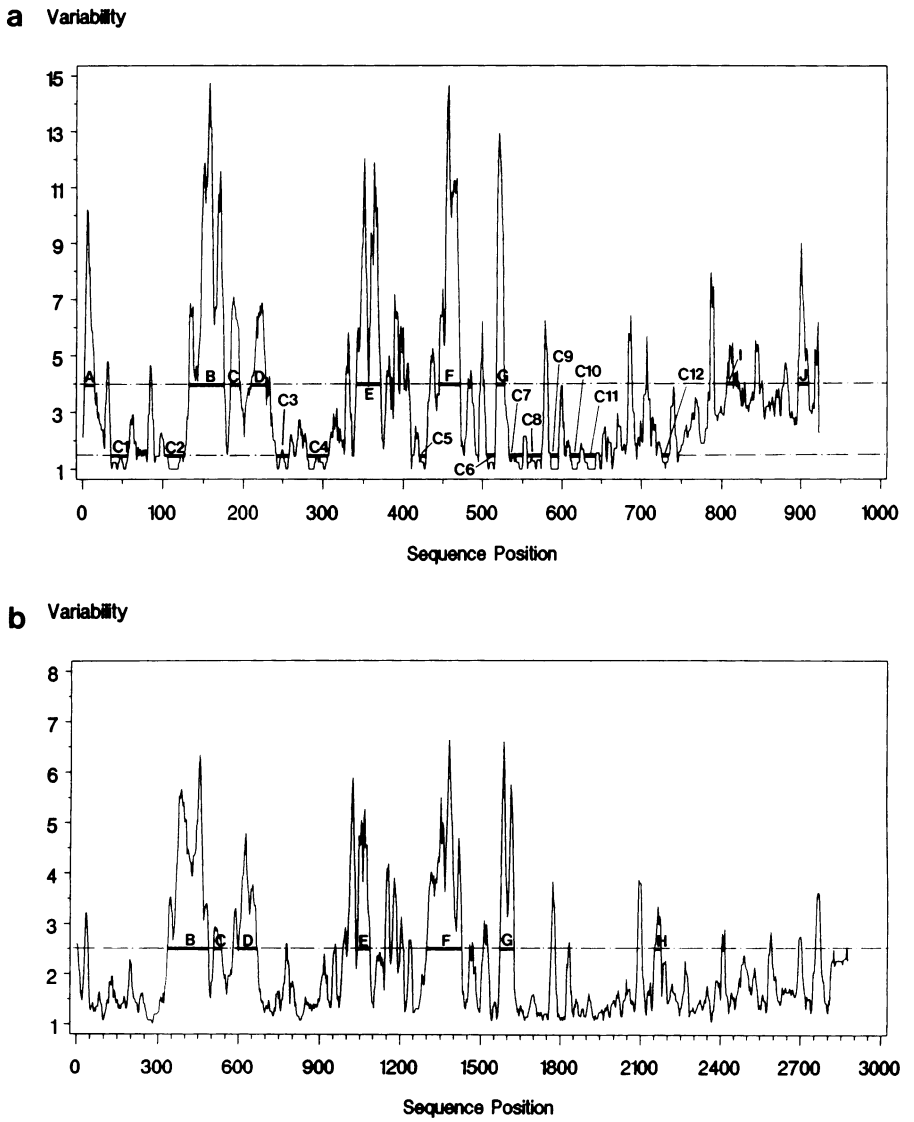

Figure 1. Variability map at different positions of envelope coding regions of seventeen isolates of HIV. Variabilities of the amino acid sequences is shown in (a), and variabilities of the nucleotide sequences is shown in (b). The data have been smoothed in overlapping groups of five for the amino acid sequence in (a), and averaged in groups of three successive positions and then smoothed in overlapping groups of five for the RNA sequences in(b). In (a) letters A-G, I AND J represent significant clusters of the hypervariable site, and C1-C12 represent significant clusters of the consensus site at amino acid sequence level. In (b) letters B-H represent significant clusters of the hypervariable site at RNA level.

Table 3: Significant Clusters of the Hypervariable Site in Envelope RNAs of Seventeen HIV Isolates

Isolates	*Significant Clusters*						
	B	C	D	E	F	G	H
NY5	325-392	405-431	475-499	819-902	1087-1153	1285-1310	1831-1853
MAL	325-417	435-460	503-529	849-929	1117-1177	1310-1338	1856-1878
ELI	325-393	412-438	480-517	837-917	1105-1171	1301-1322	1841-1863
Z6	325-396	414-440	484-521	840-923	1111-1174	1307-1328	1847-1869
RF	325-402	421-446	490-553	873-956	1144-1198	1331-1355	1877-1899
WMJ22	325-387	405-430	475-505	825-905	1093-1147	1280-1301	1823-1845
CDC42	331-420	438-463	508-547	867-950	1138-1210	1343-1364	1886-1908
SF2	325-396	414-440	484-523	843-926	1117-1168	1301-1325	1847-1869
BH8	328-402	420-446	490-515	834-920	1111-1162	1295-1316	1835-1857
HXB2	328-402	420-446	490-514	834-920	1111-1177	1310-1331	1850-1872
BH102	328-402	420-446	490-514	834-920	1111-1177	1310-1331	1850-1872
PV22	329-403	421-447	492-515	835-921	1112-1178	1311-1332	1851-1873
HXB3	328-402	420-446	490-514	834-920	1111-1177	1310-1331	1850-1872
BRU	328-417	435-461	505-529	849-935	1126-1192	1325-1346	1865-1887
SC	325-411	429-455	499-517	834-917	1105-1159	1292-1330	1850-1872
MN	325-417	435-461	505-529	849-932	1110-1177	1310-1339	1853-1874
Z3	325-384	396-422	466-505	825-905	1093-1159	1292-1317	1846-1857
Consensus Sequence	337-489	508-535	580-664	988-1087	1300-1429	1573-1624	2152-2176

sequences of the HIV envelope. The frequency of the variability in the alignments of amino acid and nucleotide sequences of these isolates have also been determined. The cumulative percent of the variability larger than 4.0 in Fig.1a is 28.2%, and that of the variability larger than 2.5 in Fig.1b is 27.5%. The two thresholds of variability(V_p=4.0 and V_n=2.5) are considered an indication of the hypervariable site in seventeen amino acid and nucleotide sequences of the HIV envelope. The minimal length L_{min} for a run of the significant hypervariable site that gives significance at the 1% or 0.1% level in the alignment of seventeen envelope proteins of HIV is determined. We have $L_{min}(0.01)=9$ and $L_{min}(0.001)=11$. The significant clusters of hypervariable sites are identified and labelled with characters A-G, I and J in Fig.1a. Similarly, we have $L_{min}(0.01)=9$ and $L_{min}(0.001)=11$ for f=27.5% in the alignment of these RNAs. The significant clusters of hypervariable sites identified at the RNA level are labelled with letters B-H and J in Fig.1b. The data are listed in Table 3. Following the same procedure we determine that the cumulative percent of the variability less than and equal to 1.5 in seventeen envelope protein sequences is 25.1%. The variability value 1.5 is considered an indication of consensus sites in seventeen amino acid sequences of HIV envelope. The minimal length L_{min} for a run of the conserved site that gives significance at the 1% or 0.1% level in the alignment of seventeen envelope proteins of HIV is determined.

Table 4: Significant Clusters of the Hypervariable, Conserved Site and Potential Antigenic epitopes in the Consensus Sequence of Envelope Proteins of Seventeen HIV Isolates

Hypervariable		*Antigenic Epitopes*		*Conserved*	
Cluster	Region	Epitope No.	Region	Cluster	Region
A	3-15				
				C1	36-57
				C2	102-125
B	133-175	I	134-158		
C	184-194				
D	210-228	II	210-230		
				C3	241-258
		III	270-282		
				C4	281-307
E	341-371	IV	337-348		
		V	373-388		
		VI	394-404		
				C5	420-428
F	446-471	VII	440-471		
				C6	505-515
G	518-527	VIII	516-527		
		IX	535-549	C7	533-550
				C8	556-573
				C9	586-595
				C10	611-622
				C11	628-642
		X	676-690		
				C12	726-735
I	807-818	XI	805-813		
J	897-909				
		XII	913-919		

We have $L_{min}(0.01)=9$ and $L_{min}(0.001)=10$. The significant clusters of the conserved site can be identified and labelled with characters C1-C12 in Fig.1a. The data of significant clusters of hypervariable and conserved sites in the 'consensus' amino acid sequence are listed in Table 4.

Correlations between Hypervariable Site Clusters and Open Region Clusters with Little Intramolecular Base Pairings

Fig.2 shows the correlation of significant hypervariable clusters and open region clusters with little intramolecular base pairings. In general, they are both located in three separated regions and 60% of the hypervariable clusters have at least one-third of nucleotides overlapped with clusters of open regions of size 60 nucleotides each. The correlation between these two highly non-random events perhaps suggests that the hypervariable regions have a strong correlation with regions of little intramolecular base pairing. If they do indeed correlate to each other, it perhaps implies that in these regions there are few base pairing constraints which limit the rate of sequence divergence. Alternatively,

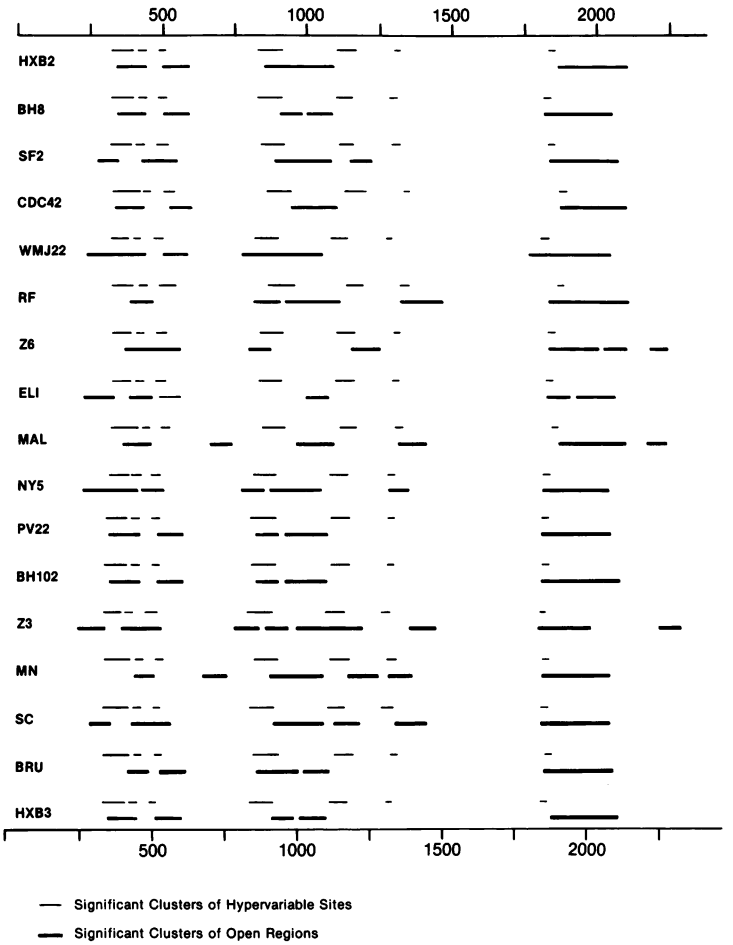

Figure 2. Correlations of significant clusters of hypervariable sites and open regions with little intramolecular base pairings in seventeen HIV isolates. The heavy line denotes the significant cluster of the open region. The light line denotes the significant cluster of the hypervariable site in the nucleotide sequence.

nucleotides of these regions would be more exposed to chemical or physical mutagens than those with stable secondary structures. Similar results have been obtained by analyzing HTLV-III(BH10), LAV and ARV-2 sequences(6). The results obtained in this study support our supposition proposed in the previous paper(6) although different energy rules are used in assessing the segment score(Salser and Tinoco energy rules(17,18) are used in this paper, and Freier energy rules(19) are used in the previous paper) .

Significant Clusters of Hypervariable and Consensus Sites and Potential Antigenic Epitopes

For further analysis the potential antigenic epitopes in the consensus sequence are predicted using a computer program PeptideStructure from the GCG Package(20) which predicted the secondary structure and calculated the values for hydrophilicity(21), flexibility(22), surface probability(23), and antigenicity(14). The significant conserved region clusters are generally hydrophobic, lack areas with a high number of β-turns and low degrees of surface probability, flexibility as well as antigenicity, however, the opposite results are observed in the significant clusters of hypervariable sites(data are not shown here).

Twelve potential antigenic determinants are identified, nine of them located in exterior envelope protein gp120 and three in transmembrane protein gp41. These potential epitope regions are listed in Table 4 and are located in predicted β-turn regions(24,25) which show high measures of nonhydrophobic and surface probability and chain flexibility. In gp120, there are six significant hypervariable regions (B to G in table 4) and five of them were predicted epitopes. There is only one epitope located in the significant cluster(C7) of the conserved site. The C7 fragment predicted to be antigenic is the consensus sequence GGGDMRDNWRSELYK. The fragment is adjacent to the cleavage site of gp120 and gp41. The predicted antigenic epitope is the same as that predicted from strains HTLV-III (BH10), LAV(1A), HTLV-III(HAT3), HTLV-III(WMJ1), HTLV-III(WMJ2), and HTLV-III(WMJ3) by Modrow et al.(3).

At the present time, little is known about the mechanism responsible for envelope gene heterogeneity. It is known that nucleotide substitutions in retroviral (RNA) genomes accumulate about a million times faster than in the nuclear DNA of higher organisms(26,27,28). In this study, we combined the sequence comparison and statistical analysis to identify significant hypervariable, conserved, and open regions among seventeen HIV isolates. Our analysis indicates that these hypervariable regions correlate with open regions and highly non-random variable regions in the exterior protein gp120 are potential antigenic epitopes. Is the variability dominated by need for antigenic variabilities and absense of secondary structures of RNAs or is there a structural or functional need to avoid secondary structure which coincidently leads to regions that are less constrained in the rate of evolution? Although it remains to be proven, it is possible that the non-random correlation between hypervariable and open regions may

provide a clue of the mutability for these rapidly evolving viruses. Two related retroviruses, equine infectious anemia virus and visna virus, 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(29,30,31).

Acknowledgements

Research sponsored, at least in part, by the National Cancer Institute, DHHS, under contract NO1-CO-74102 with Program Resources, Incorporated.

REFERENCES

1. Ratner, L. et al. (1985) Nature 313, 636-637.
2. Myers G., Rabson,A.B., Josephs, S.F., Smith, T.F., and Wong-Staal, F. (1987) A Compilation and Analysis of Nucleic Acid and Amino Acid Sequences. In Human Retroviruses and AIDS. Los Alamos, New Mexico, USA.
3. Modrow, S., Hahn, B.H., Shaw, G.M., Gallo, R.C., Wong-Stall, F. and Wolf, H. (1987) J.Virol. 61, 570-578.
4. Starcich, B.R., Hahn, B.H., Shaw, G.M., Mcneely, P.D., Modrow, S., Wolf, H., Parks, E.S., and Parks, W.P. (1986) Cell 45, 637-648.
5. Willey, R. et al. (1986) Proc. Natl. Acad. Sci. U.S.A. 83, 5038-5042.
6. Le, S.Y., Chen, Jih-H., Braun, M.J., Gonda, M.A. and Maizel, J.V.Jr. (1988) Nucl. Acids Res. 16, 5153-5168.
7. Wong-Stall, F. et al. (1985) Science 229, 759-762
8. Le, S.Y., Chen, J.H., Currey, K.M., and Maizel, J.V.Jr. (1988) Computer Applications in the Biosciences 4, 153-159.
9. Le, S.Y. and Maizel, J.V. (1988) Submitted.
10. Martinez, H,M. (1988) Nucl. Acids Res. 16, 1683-1691.
11. Alizon, M., Wain-Hobson, S., Montagnier, L. and Sonigo, P. (1986) Cell 46, 63-74.
12. Wu, T.T. and Kabat, E.A. (1970) J. Exp. Med. 132, 211-250.
13. Foulser, D.E. and Karlin,S. (1987) Stoch. Processes Appl. 24, 203-224.
14. Jameson, B.A. and Wolf, H. (1988) Computer Applications in the Biosciences 4, 187-192.
15. Chen, Jih-H., Le, S.Y., Shapiro, B., Currey, K.M. and Maizel, J.V.Jr. (1988) submitted.
16. Karlin, S., Blaisdell, B.E. and Brendel, V. (1988) submitted
17. Salser, W. (1977) Cold Spring Harbor Symp. Quant. Biol. 42, 985-1002.
18. Cech, T.R., Tanner, N.K., Tinoco, I.Jr., Weir, B.R., Zuker, M. and Perlman, P.S. (1983) Proc. Natl. Acad. Sci. USA 80, 3903-3907.
19. Freier, S.M., Kierzek, R., Jaeger, J.A., Sugimoto, N., Caruthers, M.H., Neilson, T. and Turner, D.H. (1986) Proc. Natl. Acad. Sci. USA 83, 9373-9377.
20. Devereux, J., Haeberli, P and Marquess, P. (1987) Introduction to the Sequence Analysis Software Package of the Genetics Computer Group. Version 5, University of Wisconsin Biotechnology Center;

Devereux, J., Haeberli, P. and Smithies, M. (1984) Nucl. Acids Res. 12, 387-395.
21. Kyte, J. and Doolittle, R.F. (1982) J. Mol. Biol. 157, 105-132.
22. Karplus, P.A. and Schulz, G.E. (1985) Naturwissenschaften 72, 212-213.
23. Emini, E.A., Hughes, J.V., Perlow, D.S. and Boger, J. (1985) J. Virol. 55, 836-839.
24. Chou, P.Y. and Fasman, G.D. (1974) Biochemistry 13,222-245; (1978) Adv. Enzymol. Relat. Areas Mol. Biol. 47, 45-148.
25. Garnier, J. Osguthorpe, D.J. and Robson, B. (1978) J. Mol. Biol. 120, 97-120.
26. Sharp, P.M. and Li, W-H (1988) Nature 336, 315.
27. Gojobori, T. and Yokoyama, S. (1985) Proc. natn. Acad. Sci. USA 82, 4198-4202.
28. Li, W-H, Tanimura, M. and Sharp, P.M. (1988) Molec. biol. Evol. 5, 313-330.
29. Clements, Y.E., Pedersen, F.S., Narayan, O. and Haseltine, W.A. (1980) Proc. Natl. Acad. Sci. USA 77, 4454-4458.
30. Montelaro, R.C., Parekh, B., Orego, A. and Issel, C.J. (1984) J. Biol. Chem. 259, 10539-10544.
31. Salinovich, O., Payne, S.L., Montelaro, R.C., Hussain, K.A., Issel, C.J. and Schnorr, K.L. (1986) J. Virol. 57, 71-80.