J Mol Evol (1993) 36:448–457

Journal of Molecular Evolution

Divergent Evolution May Link Human Immunodeficiency Virus GP41 to Human CD4

Antonio Facchiano,[1] Francesco Facchiano,[2] and Jos van Renswoude[1,3]

[1] Raggio Italgene S.p.A., Via delle Antille 29, 00040 Pomezia, Roma, Italy
[2] Laboratory of Molecular Neurobiology, Istituto Mario Negri Sud, S. Maria Imbaro, Chieti, Italy
[3] Department of Experimental Medicine, Universita' La Sapienza, Roma, Italy

Summary. A local sequence similarity of HIV envelope proteins (gp120 and gp41) to immunoglobulins suggests that a mimicry phenomenon may form the basis of the HIV–cell membrane interaction and of HIV-induced autoimmune reaction. We explored the hypothesis of any deeper relationship between HIV env proteins and immunoglobulin family members. An overall DNA sequence similarity between gp41 coding region of env gene and the HIV-receptor CD4 gene was observed and a 14-base-long oligonucleotide, almost unique in the GenBank, was found in gp41 and CD4 genes. The alignment of env gene to CD4 gene and to 84 different sequences showed a significantly higher homology score and a nonrandom similarity in the CD4-env alignment. A significant similarity was also found between the env protein and the sequence encoded by an alternate reading frame of CD4 gene. Our observations suggest that gp41 coding region might have a different origin than the gp120 coding region of the env gene, and that a divergent evolution might link gp41 to CD4 or immunoglobulin family members. In this study the analysis of alternate-reading-frame products is also proposed as a novel approach to investigate evolutionary links and structure-function relationships.

Key words: Retrovirus — HIV — CD4 — Minus strand — Alternate reading frame — Frameshift — Divergence — Evolution

Offprint requests to: A. Facchiano

An increasing body of evidence suggests that viruses are the most likely agents involved in the pathogenesis of autoimmunity (Schattner and Rager-Zisman 1990; Paque and Miller 1991; Prabhakar et al. 1988; Guldner et al. 1990). Therefore, understanding the viral strategies to overcome the host's defenses may help in defining new therapeutic approaches for viral and autoimmune diseases. Several mechanisms have been proposed for the induction of autoimmune diseases and self-tolerance, including the death of lymphocytes and their inactivation or suppression (Kronenberg 1991). A so-called mimicry evolution has been suggested to form, at least partially, the basis of the induction of autoimmune diseases (Oldstone 1987; Golding et al. 1989; Fitzpatrick et al. 1990). According to this model, some viruses synthesize antigens similar to the host histocompatibility complexes, giving rise to an autoimmune reaction. Mimicry has also been proposed to underlie, at least to some extent, persistent viral infections, especially those caused by retroviruses. HIV envelope proteins of gp120 and gp41 are encoded for by the env gene. They are currently being thoroughly investigated and limited homologies between them and immunoglobulins have been reported (Golding et al., 1989; Kieber-Emmons et al. 1989; Maddon et al. 1986), suggesting a mimicry evolution of env proteins toward the immunoglobulin epitopes. According to Golding (1989), this phenomenon may be partially responsible for the autoantibody synthesis occurring during the HIV infection (see Buskila and Gladman 1990 and references therein) and may contribute to the functional impairment of $CD4^+$ T cells. The local similarity between gp120/gp41 and immunoglobu-

lins is in line with the finding that gp120-binding site and major histocompatibility complex (MHC)-binding site on CD4 partially overlap (Maddon et al. 1986; Klatzmann et al. 1984; Dalgleish et al. 1984; Clayton et al. 1989).

We hypothesized that an early divergent event, rather than a convergent one, might have played an important role in the mimicry process of env proteins toward the immunoglobulins. To test this hypothesis we looked for any deeper sequence relationship between the HIV env and the CD4 receptor (an immunoglobulin family member) by means of DNA and protein sequence analyses.

Materials and Methods

The DNA sequences were taken from the GenBank version #64 (Burks et al. 1991). The protein sequences were from PIR version #25 (Barker et al. 1991). The sequence analyses were performed by using HIBIO DNASIS software (distributed by Pharmacia LKB Biotechnology, Uppsala, Sweden), GENEPRO software, version 5.0 (distributed by Riverside Scientific Enterprise, WA, USA), and PCGENE release 6.00 (by Intelligenetics Inc., Geneva, Switzerland). The dinucleotide usage was computed by GENEPRO software as follows: the observed occurrence of each dinucleotide was divided by the expected value, calculated from the base composition of the sequence. Therefore, a ratio of 1 indicated the presence of that dinucleotide exactly corresponding to chance, for the given base composition. The ratios of occurrence of all dinucleotides for the entire CD4, for the extracellular part of CD4 (from base 1 to 1,030), and for the intracellular portion of CD4 (from base 1031 to the end) were compared to the ratios of the entire env gene, of the gp120 coding portion (bases 1–1,530), and of the gp41 coding portion (bases 1,530–2,570). The ratios comparison was performed with a correlation analysis and r^2 and P values were calculated. The correlations showing P values ≤ 0.001 were considered the most significant ones.

The sequence alignment (gap insertion allowed) of the human CD4− and the HIV isolate BH10 env− DNA strands was carried out using a modified form of the Needleman and Wunsch algorithm (Needleman and Wunsch 1970). An additional "no-gap-allowed" comparison was performed according to Lipman and Pearson (1985) as follows: five DNA sequences (CD4 gene; env gene, the entire HIV isolate BH10 genome lacking the env portion, taken as control of the env portion; the human immunoglobulin G Fc receptor, randomly chosen among the immunoglobulins, as control of the CD4 gene; the entire human retrovirus Herv K10 genome, randomly chosen among the retroviruses, as control of the HIV genome sequence) were considered as "key sequences" and compared to 86 different DNA sequences (see legend of Table 3 for the complete list), here referred to as "target sequences," consisting of the coding strands of 43 genes and of their respective 43 complementary sequences. In total 405 comparisons were made. The 43 analyzed genes included the five used as key sequences and 38 more, randomly chosen among those encoding intracellular, extracellular, and membrane-bound proteins, from humans, other eukaryotes, and lower species. Each key sequence was aligned to each target sequence and homology scores were calculated. The score calculation was carried out as follows: +4 points were given to each identity and −2 to each mismatch, for the sequences being compared in the "no gap allowed" mode. Each DNA sequence comparison consisted of several different alignments (in an iterative way) of the same pair of DNA strands, yielding the 50 best alignments and the 50 highest scores. From the 50 highest scores of each comparison, a mean score was computed, always ranging from 30 to 50, with a standard deviation of about 10 in all cases. A value of 100 was chosen as the threshold value, because it was significantly higher than the mean values observed [i.e., $50 + (5 \times 10)$, the highest mean value plus 5 standard deviation units]. The alignments showing a score higher than 100 were considered as related in a nonrandom fashion (indicated with boxed dots in Table 3).

The known gene products are here referred to as the "real" proteins and are coded for by the known coding reading frame (RF-1) of the DNA plus strands. The "hypothetical" species encoded by the plus strand in reading frames 2 (RF-2) (i.e., coding + 1) and 3 (RF-3) (i.e., coding + 2) and those encoded by the minus strand in all three reading frames, translated in 5′–3′ direction, are considered here as "alternate-reading-frame products." The alignment of the "real" and "hypothetical" CD4 and env proteins was obtained according to Gotoh (1982) as modified by Myers and Miller (1988), whereas the homology scores were calculated according to Needleman and Wunsch (1970). The stop-codon codes present in the alternate-reading-frames products were not considered in the identity percentage computation, since it is not clear which "weight" has to be attributed to the match between the stop-codon code and any amino acid or gap insertion.

The DNA as well as the protein comparisons were subjected to Monte Carlo–like analysis, using the Needleman and Wunsch comparison algorithm as modified by Feng et al. (1984). Briefly, the maximum score (S) that can be achieved in the alignment of two sequences is compared with the maximum scores computed on 100 scrambled sequences (i.e., with the same length and composition). The mean (Sr), the standard deviation (SDr) of the randomized comparisons scores, and an alignment score (A) are computed. The A score is the number of standard deviations by which the maximum score for the real sequences exceeds the mean of the maximum scores of the shuffled sequences:

$$A = (S - Sr)/SDr$$

Scores with $A \geq 3$ are usually considered to indicate a nonrandom correlation and therefore point to a relatedness of the aligned sequences.

Results

DNA Comparison

The base composition of env and CD4 genes is reported in Table 1. The base composition of the gp120 coding region of the env gene [especially the AT, CG, and (AT)/(CG) values] appears quite different from that of the gp41 coding region, i.e., 62% vs. 55%, 38% vs. 45%, and 1.63 vs. 1.22. In addition, the base composition of gp41 coding region is more similar to the 5′ end of CD4 (coding for the extracellular portion of the CD4 protein) than the gp120 coding region. In fact, if one subtracts the percentage values of A, T, C, and G of the CD4 5′ end from those of gp120 and calculates the sum of the absolute values of the differences, a value of 32 is obtained. If the same values of the CD4 5′ end are

Table 1. Base composition of env and CD4 genes[a]

	env	env gp120	env gp41	CD4	CD4 5′ end	CD4 3′ end
A	35%	38%	30%	23%	26%	19%
T	24%	24%	25%	22%	20%	24%
G	24%	22%	27%	27%	27%	26%
C	17%	16%	18%	28%	27%	31%
AT	59%	62%	55%	45%	46.5%	43%
CG	41%	38%	45%	55%	53.5%	57%
(AT)\(CG)	1.44	1.63	1.22	0.81	0.87	0.75

[a] "Env" refers to the entire env gene sequence; "env gp120" refers to bases 1–1530 of the env gene, coding for the gp120 protein; "env gp41" refers to bases 1531–2570 of the env gene, coding for the gp41 protein; "CD4" refers to the entire CD4 gene; "CD4 5′ end" refers to bases 1–1030 of the CD4 gene, coding for the extracellular portion of the CD4 protein; "CD4 3′ end" refers to bases 1031–1742 of the CD4 gene, coding for the intracellular portion of CD4 protein

subtracted from those of gp41, and the sum of the absolute values is calculated, a value of 18 is obtained. The differences are always higher between gp120 and CD4 compositions, as compared to gp41 and CD4, except in T percentage composition, where the two differences are almost identical (12 vs 4; 4 vs 5; 5 vs 0; 11 vs 9). These observations indicate that gp41 base composition is closer to CD4 than that of gp120 is.

The dinucleotide usage comparison shows that, again, the two env portions behave differently (correlation with $r^2 = 0.34$) and that the gp41 region shares a better correlation to the CD4 ($r^2 = 0.54$) than the gp120 region ($r^2 = 0.27$). Table 2 reports the P values calculated for each correlation. The most significant (i.e., $P \leq 0.001$) are: env to gp41, env to gp120, CD4 to CD4 5′ end, CD4 to CD4 3′ end (all expected) and, unexpectedly, gp41 to CD4 5′ end. These observations support the hypothesis that gp41 might have a different origin than the gp120 coding region and might show some similarity to CD4 or immunoglobulins.

Therefore, the human CD4 gene and the HIV env gene were compared. They were first aligned in a "gaps allowed" mode. An overall 50% identity was found, for a stretch of more than 1,000 bases in both sequences (Fig. 1). The aligned bases of the CD4 gene encode the extracellular, N-terminal part of the CD4 protein; the aligned bases of the env gene code for gp41. Nine regions show a very high local identity, equal to or higher than 70% (underlined in Fig. 1), and an identical stretch of 14 consecutive bases (boxed in Fig. 1) was found in both CD4 and gp41 genes. This oligonucleotide occurs rarely in the entire GenBank. In fact, the "plus" sequence (5′-AGAAGAAGGTGGAG-3′) was found in only five genes — namely, in the genes coding for the env protein from HIV type 1; the human T-cell surface glycoprotein CD4; the human (and mouse) recombination activating protein RAG-1; the *Xenopus laevis* adult alpha I-globin; and the *Saccharomyces cerevisiae* STE12 protein. The complementary sequence (5′-CTCCACCTTCTTCT-3′), present on the minus strands of the HIV env gene and the CD4 gene, was also found in the plus strand of the *Trypanosoma cruzi* kinetoplast-associated protein (KAP) gene. Interestingly, despite the large variability of the env gene sequence, this 14-base-long oligonucleotide appears to be a conserved sequence. It is present in 23 different HIV-1 isolates; it shares 13 identities with the simian immunodeficiency virus (SIV isolate 155), with 3 more HIV type 1 isolates, and with the mouse Cas NS-1 retrovirus, and it shares 12 identical bases with an HIV type 2 gene (isolate ROD). The complementary "minus" sequence shares 12 identities with the human T-cell leukemia virus type II (clone lambda-H6.0) and with the human T-cell lymphotropic virus type 2 (both in the env coding region).

The CD4 and gp120/gp41 genes were also compared, using a "no gap allowed" procedure, to each other and to 85 other DNA sequences. Five DNA "key sequences" were compared to the "target sequences" consisting of the plus and minus strands of five key sequences and of the plus and minus strands of 38 other DNA sequences. (See Table 3 footnote for the complete list of the target sequences.) Table 3 reports the comparisons performed and indicates as boxed dots those showing significant homology scores (i.e., scores higher than 100). Out of 405 different comparisons carried out, less then 4% show significant scores; i.e., the comparison between the plus strand of the env− and the plus strand of the CD4− gene (pair #1-#3); the one between the env gene and the minus strand of the beta globin region on chromosome 11 (#3-#46); the one between the entire HIV genome lacking the env part and both strands of the beta globin region on chromosome 11 (#7-#45, #7-#46); the alignment between the HIV genome and the minus strand of the *Xenopus laevis* alpha 1 globin (#7-#68); the Fc receptor alignment to the plasmid pT48

Table 2. *P* values from the correlation analysis among the dinucleotide frequency ratios[a]

		env total	env gp120	env gp41	CD4 total	CD4 5′ end	CD4 3′ end
env	total						
env	gp120	≤0.001					
env	gp41	≤0.001	>0.01				
CD4	total	>0.001	>0.01	≤0.001			
CD4	5′ end	>0.001	>0.01	≤0.001	≤0.001		
CD4	3′ end	>0.001	>0.01	>0.001	≤0.001	≤0.001	

[a] "Env" refers to the entire env gene sequence; "env gp120" refers to bases 1–1530 of the env gene, coding for the gp120 protein; "env gp41" refers to bases 1531–2570 of the env gene, coding for the gp41 protein; "CD4" refers to the entire CD4 gene; "CD4 5′ end" refers to bases 1–1030 of the CD4 gene, coding for the extracellular portion of the CD4 protein; "CD4 3′ end" refers to bases 1031–1742 of the CD4 gene, coding for the intracellular portion of CD4 protein. See Material and Methods section for the procedure followed.

(#5-#75) and to the beta globin region on chromosome 11 (#5-#45, #5-#46); and the alignment of the Herv K10 retrovirus genome to the minus strand of the *Clostridium elegans* vinculin gene (#9-#74). The high homology scores found in the comparisons among viral and retroviral sequences (pairs #7-#9, #7-#79, #9-#79, #9-#85), as well as those between the CD4 and MHC beta chain genes (pair#1-#39), were expected. The Monte Carlo analysis confirmed that a threshold homology score of 100 provides a statistically meaningful cutoff point; in fact, the high homology scores found for pair #1-#3 and pair #1-#39 (in Table 3) were found to be nonrandom ($A = 3.6$ and $A = 4.2$, respectively). On the other hand, comparisons showing low homology scores consistently show A scores below the cutoff point. In fact, alignments of the CD4 gene, the env gene, and the entire HIV genome lacking the env coding region with the *Xenopus laevis* alpha 1 globin (pairs #1-#68, #3-#68, #7-#68, respectively, all indicated by dashes in Table 3) showed homology scores lower than 100, and A scores lower than three.

Protein Comparison

The CD4/env relationship was also investigated at the protein sequence level. The 6 CD4- and the 6 env-related sequences were compared to one another and all 66 comparisons were subjected to the Monte Carlo–like analysis. Table 4 shows that only one comparison exceeded the randomness threshold (A score = 3.5). It corresponds to the alignment between the real env gene product and the CD4 "hypothetical" protein coded for by reading frame 3 of the CD4 plus strand.

Their partial alignment is reported in Fig. 2. Five regions of the known env gene product, all located in the gp41 portion, match with high identity percentage to five regions of the CD4-related protein: namely, residues 488–517 of env (55% identity to residues 36–68 of CD4-related protein); residues 536–601 of env (40% identity to residues 139–188 of CD4-related protein); residues 602–640 (37% identity to residues 205–236 of CD4-related protein); residues 691–725 (28% identity to residues 278–311); residues 737–762 (35% identity to residues 319–347). The insert in Fig. 2 summarizes the positions and the percentages of the similarities found.

Discussion

In this report we show sequence homology of CD4 and env, at both the DNA and protein level, which was observed by using "gap allowed" and "no gap allowed" comparisons. The significance of the similarity was evaluated in terms of randomness (i.e., through a Monte Carlo analysis) and in terms of "specificity" (i.e., through a comparison of the CD4 and env genes to a large number of nonrandom DNA sequences). The analyses were carried out on DNA plus and minus strands and on the known gene product as well as on the alternate-reading-frame protein products.

The base composition and the dinucleotide usage of the gp41 coding portion of the env gene are different from the gp120 coding portion and are unexpectedly close to CD4. These observations may suggest a different origin and/or evolutionary pressure on the two regions of env gene and a possible relationship of gp41 to CD4. The comparison of gp41 and CD4 DNA sequences showed 50% identity. The Monte Carlo–like analysis indicated that it is nonrandom; hence, the possibility that the sequences appear similar because of the similar base composition can be ruled out. The similarity found was also evaluated by comparing the CD4 and env

Fig. 1. Alignment of the 3′ end of the env HIV type 1 (isolate BH10) gene (bases 1531–2570) to the 5′ end of the human CD4 gene (bases 1–1036). The aligned region of the env gene encodes the entire gp41 protein; the aligned region of the CD4 gene encodes the extracellular region of CD4 antigen. The 14-base sequence common to both strands is *boxed;* the regions showing high identity (i.e., equal to or higher than 70%, spanning at least 15 bases) are *underlined.*

Table 3. DNA comparison of key sequences (1, 3, 5, 7, 9) to target sequences (1 to 86)[a]

	1	3	5	7	9			1	3	5	7	9	
1		⊡	—	—	—		49	—	—	—	—	—	
2	—	—	—	—	—		50	—	—	—	—	—	
3	⊡		—	—	—		51	—	—	—	—	—	
4	—	—	—	—	—		52	—	—	—	—	—	
5	—	—		—	—		53	—	—	—	—	—	
6	—	—	—	—	—		54	—	—	—	—	—	E
7	—	—	—		⊡		55	—	—	—	—	—	U
8	—	—	—	—	—		56	—	—	—	—	—	K
9	—	—	—	⊡			57	—	—	—	—	—	A
10	—	—	—	—	—		58	—	—	—	—	—	R
							59	—	—	—	—	—	Y
11	—	—	—	—	—		60	—	—	—	—	—	O
12	—	—	—	—	—		61	—	—	—	—	—	T
13	—	—	—	—	—		62	—	—	—	—	—	E
14	—	—	—	—	—		63	—	—	—	—	—	S
15	—	—	—	—	—		64	—	—	—	—	—	
16	—	—	—	—	—		65	—	—	—	—	—	
17	—	—	—	—	—		66	—	—	—	—	—	
18	—	—	—	—	—		67	—	—	—	—	—	
19	—	—	—	—	—		68	—	—	—	⊡	—	
20	—	—	—	—	—								
21	—	—	—	—	—		69	—	—	—	—	—	
22	—	—	—	—	—		70	—	—	—	—	—	
23	—	—	—	—	—		71	—	—	—	—	—	
24	—	—	—	—	—		72	—	—	—	—	—	
25	—	—	—	—	—		73	—	—	—	—	—	
26	—	—	—	—	—	H	74	—	—	—	—	⊡	O
27	—	—	—	—	—	U	75	—	—	⊡	—	—	T
28	—	—	—	—	—	M	76	—	—	—	—	—	H
29	—	—	—	—	—	A	79	—	—	—	⊡	⊡	E
30	—	—	—	—	—	N	80	—	—	—	—	—	R
31	—	—	—	—	—	S	81	—	—	—	—	—	S
32	—	—	—	—	—		83	—	—	—	—	—	
33	—	—	—	—	—		83	—	—	—	—	—	
34	—	—	—	—	—		84	—	—	—	—	—	
35	—	—	—	—	—		85	—	—	—	—	⊡	
36	—	—	—	—	—		86	—	—	—	—	—	
37	—	—	—	—	—								
38	—	—	—	—	—								
39	⊡	—	—	—	—								
40	—	—	—	—	—								
41	—	—	—	—	—								
42	—	—	—	—	—								
43	—	—	—	—	—								
44	—	—	—	—	—								
45	—	—	⊡	⊡	—								
46	—	⊡	⊡	⊡	—								
47	—	—	—	—	—								
48	—	—	—	—	—								

[a] Boxed dots indicate the statistically most significant comparisons (i.e., those showing homology scores higher than 100, $P < 0.01$); dashes indicate the less-significant comparisons (i.e., those with scores equal to or lower than 100). Each key sequence was compared to each target sequence. For each comparison, homology scores were computed. (For more details see the Materials and Methods section.) The odd numbers indicate the plus strands; the even numbers indicate the corresponding complementary (minus) strands. The DNA sequences, from humans, nonhuman eukaryotes, and others species, were the following: #1: human CD4; #3: HIV type 1 env, isolate BH10; #5: human immunoglobulin G Fc receptor; #7: entire HIV type 1 genome, isolate BH10, lacking the env part corresponding to the sequence

genes to a large number of coding — and complementary — strands of genes coding for extracellular and intracellular proteins as well as cytoplasmic and membrane-bound proteins, from higher species (mostly human) and lower species. The significance threshold of the homology scores was chosen with highly stringent criteria. Some of the comparisons showed interesting relationships. For instance, the correlation between the human endogenous Herv retrovirus and the minus strand of the vinculin gene from *Cl. elegans* (#9-#74 in Table 3) was unexpected, although some functional and structural relationships between cytoskeletal proteins (to which vinculin belongs) and (retro)viral sequences have been proposed (Gooding et al. 1990; Brake et al. 1990). The 50% identity found between the CD4 and the gp41 genes is supported by many observations. According to Table 3, the homology between CD4 and env can be considered "specific," since most sequences (i.e., 96%) share much lower homology scores. Furthermore, the presence of the identical 14-base-long sequence in the CD4 and gp41 coding strands is very unlikely to be coincidental (Helene 1987) and occurs very rarely in the GenBank. The oligonucleotide is also conserved in other retroviral sequences (13 or 12 identities out of 14 bases). We believe that the presence of the common oligonucleotide is not significant by itself; however, along with the high degree of overall homology, it may support a common-ancestry hypothesis. It is also worthwhile to mention that the env gene portion found to be similar to the CD4 gene belongs to a

#3; #9: human endogenous retrovirus Herv K10; #11: human lysozyme; #13: human alpha-1-antichymotrypsin; #15: human m2 muscarinic acetylcholine receptor; #17: human adenosine deaminase; #19: human complement component C9; #21: human cytochrome P450; #23: human glucose-6-phosphate dehydrogenase; #25: human glucocorticoid receptor; #27: human histone H4; #29: human c-erb-B-2; #31: human thyroid hormone receptor; #33: human estrogen receptor; #35: human fibronectin; #37: human factor X; #39: human major histocompatibility complex (MHC) beta chain; #41: human recombination activating protein RAG-1; #43: human epidermal growth factor receptor; #45: human beta globin region on chromosome 11; #47: human plasminogen; 49: guinea pig preproinsulin; #51: sperm whale myoglobin; #53: *Drosophila melanogaster* sn-glycerol-3-phosphate dehydrogenase; #55: amoeba myosin 1B heavy chain; #57: bovine cation-independent mannose-6-phosphate receptor; #59: mouse pancreatic ribonuclease; #61: sea urchin complete mitochondrial genome; #63: bovine pancreatic phospholipase A_2; #65: sheep prostaglandin G/H synthase; #67: *Xenopus laevis* adult alpha 1 globin; 69: *Trypanosoma cruzi* kinetoplast-associated protein; #71: maize mitochondrial S-1 genome; #73: *Clostridium elegans* vinculin; #75: plasmid pT48 from *Staphylococcus aureus;* #77: entire genome of rhinovirus, strain 14; #79: rat endogenous retroviral sequence, Sprague-Dawley strain; #81: yeast Pep4 gene encoding the vacuolar proteinase A (PrA); #83: *Saccharomyces cerevisiae* STE12 protein; #85: bacteriophage T7, complete genome

Table 4. *A* scores, computed with the Monte Carlo analysis, for the alignment of the CD4- and env-related proteins[a]

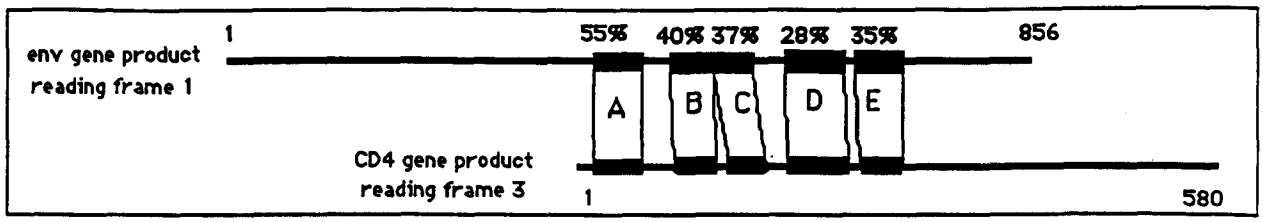
	A	B	C	D	E	F	G	H	I	L	M	N
A												
B	1.3											
C	−0.1	0.1										
D	1.6	1.5	−1.1									
E	−0.04	0.2	0.1	−1.1								
F	−0.4	0.1	0.03	−1.2	−0.2							
G	1.1	1.8	0.4	−0.9	−1.5	0.2						
H	−0.9	0.1	1.3	−0.5	0.5	−0.4	0.2					
I	3.5	0.9	−0.7	0.5	0.2	2.1	1.6	−0.7				
L	1.6	−0.7	−0.6	1.3	0.2	1.2	−0.1	−0.5	−1.2			
M	−1.4	−0.1	0.7	−1.4	0.04	0.8	−1.1	1.9	−0.2	0.1		
N	−2.0	−0.6	0.2	−0.7	1.5	−0.06	0.3	−1.2	−2.3	−0.5	2.1	

[a] The CD4- and env-related proteins are here referred to as the protein products of the three reading frames of the corresponding plus and minus DNA strands. The code for the sequences is the following: A, translated sequence from the env gene, plus strand, reading frame 1 (i.e., the known gene product); B, as A, but translated from reading frame 2; C, as A, but translated from reading frame 3; D, translated sequence from the env gene, minus strand, reading frame 1; E, as D, but translated from reading frame 2; F, as D, but translated from reading frame 3; G, translated sequence from the CD4 gene, plus strand, reading frame 1 (i.e., the known gene product); H, as G, but translated from reading frame 2: I, as G, but translated from reading frame 3; L, translated sequence from the CD4 gene, minus strand, reading frame 1; M, as L, but translated from reading 2; N, as L, but translated from reading frame 3. The statistically significant score (i.e., higher than 3) is boxed

so-called (almost) invariable region well conserved in different HIV isolates (Alizon et al. 1986; Anand et al. 1989). In light of the high env gene variability, the meaning of the found similarity is further strengthened. The presence of the 14-base-long oligonucleotide in CD4, gp41, and the RAG-1 gene, and mostly conserved in other retroviruses, is rather interesting because of the role RAG-1 plays in the recombination process of immunoglobulins and immunoglobulin-related proteins (Oettinger et al. 1990). Recombination and integration into the host genome are crucial steps in the retroviruses' life cycle (Bushman et al. 1990), as well as their evolution (Gojobori et al. 1990). In the absence of any significant overall homology, the common oligonucleotide may represent a common functional or regulatory site in possibly functionally related genes.

In most cases, similarity between DNA strands is expected to be more evident than between the corresponding protein sequences, since single insertions or deletions are able to change the reading frame and, therefore, the derived protein sequence. The known gene products of env and CD4 do not show any significant similarity. Nevertheless, the relationships observed between CD4 and gp41 genes were confirmed at the protein level; interestingly, high identity percentage and nonrandom relationship were found between the "real" env protein, in the gp41 region, and a CD4-related sequence, encoded by an alternate reading frame of the plus strand (i.e., coding +2).

The observed structural relationship can help in elucidating the still-unclear evolution and origin of the HIV genome (Gojobori et al. 1990; Doolittle et al. 1989; Argos 1989; Yokoyama 1988). Local similarities had been already observed between env proteins and immunoglobulins: namely, (1) between gp120 and the immunoglobulins [around 30% identity out of approximately 70 residues (Kieber-Emmons et al. 1989) and around 30% identity out of approximately 40 residues (Maddon et al. 1986)], and (2) between gp41 and the human HLA class II beta chain (four identities out of five consecutive residues) (Golding et al. 1989). Based on these findings, it had been proposed that the env gene is convergently evolving toward the immunoglobulins. We have observed homology between large parts of gp41 and CD4 gene and protein sequences. The presence of a common sequence in a virus and its host, provided it is not by chance, might be explained as being the result of the incorporation of a cellular DNA sequence into the viral genome, i.e., by a divergent process (Schattner and Rager-Zisman 1990; Argos 1989). This is especially possible in the case of retroviruses, which integrate into the host's genome. Our interpretation of the data presented here implies that divergence has occurred, either due to early common ancestry or, more likely, as a result of a recent gene capture process. The similarity observed between the real gp41 sequences and the product of an alternate reading frame of CD4 gene may suggest that frameshift phenomena may have occurred on originally in-frame genes. We propose a divergent evolutionary relationship between gp41 and CD4 (or immunoglobulins) genes for a number of reasons:



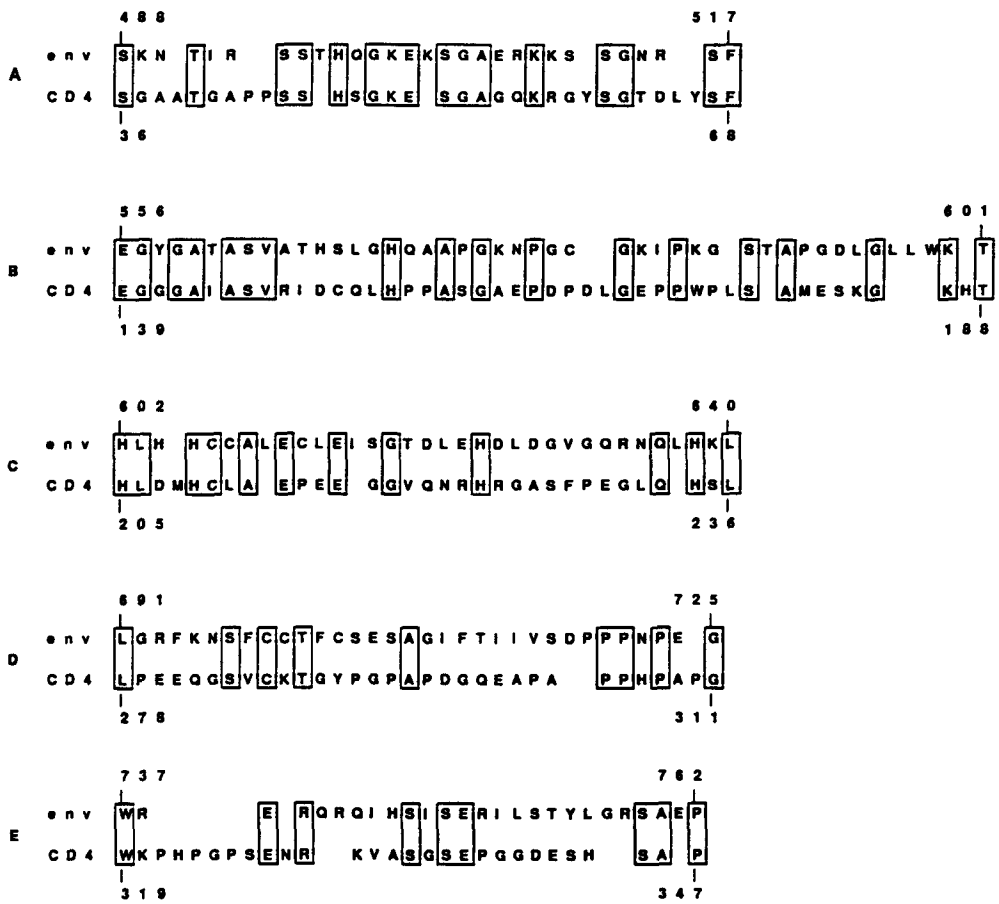


Fig. 2. Alignment of the known env gene product to the product of the CD4 gene plus strand, translated in reading frame 3. *A*, *B*, *C*, *D*, and *E* represent the regions showing the highest homology. The stop-codon codes were eliminated on the CD4-related sequence, since their match to gaps or any amino acid is a debated point. Identities are *boxed;* conservative substitutions are not highlighted. The *insert* summarizes the positions and the percentage of identity.

1. The different base composition and dinucleotide usage of gp120 and gp41 coding regions suggest that they might have had a different origin and/or evolution; the gp41 base usage appears to be closer to CD4 than gp120; the 14-base oligonucleotide common to the CD4 and gp41 DNA strands is found conserved in many HIV isolates and other retroviruses, and is very rare in the GenBank. Its presence also correlates to the nonrandom DNA and protein overall similarity between CD4 and gp41. All these data support the hypothesis of an evolutionary link between the HIV gp41 and the human CD4 (or the immunoglobulin family).
2. While a convergent evolution may explain the homology between known protein products (i.e., derived from plus strand and in-frame translation), a similarity involving out-of-frame products can be less easily explained as being due to a convergent mechanism. Therefore, the protein

similarity found between the gp41 gene product sequence and a CD4 alternate-reading-frame product may indicate that frameshift phenomena and point mutations took place on originally more-similar and in-frame DNA strands.

3. The convergent evolution explains homology of short regions well, as in the case of the similarity between the immunoglobulins and HIV envelope proteins reported in the literature (Golding et al. 1989; Kieber-Emmons et al. 1989; Maddon et al. 1986), whereas a divergent mechanism better explains homology spanning entire genes or large parts of them, or long protein sequences, as in the CD4-gp41 case reported here.
4. Since a divergent event may have occurred in a very recent past (i.e., few tens or hundreds of years ago) (Doolittle et al. 1989), one might expect a sequence homology between CD4 and gp41 higher than 50% identity. Nevertheless, the retroviral genome in general, and the env coding region in particular, is known to mutate at an extremely high rate, about 10^6 times higher than the rates observed in DNA-based genetic systems. This may explain why we found no more than 50% identity between the CD4 and gp41 DNA strands and why the env proteins of different HIV isolates share low overall identity (Alizon et al. 1986), despite their close relatedness. The high mutation rate of retroviruses has to be taken into account in tracing the origins of the HIV genome and in evaluating the divergent pressure on the env gene ancestor.
5. Finally, gene capture from cellular counterparts has been proposed for other viruses (Chee et al. 1990; Moore et al. 1990). Also, a divergent evolution of HIV aspartyl protease from a mammalian ancestor has been proposed (Navia et al. 1989). As recently pointed out (Katz and Skalka 1990; and references therein), cellular DNA is a possible source of retroviral genes, which might result from early gene capture events, especially in the gag- and env-gene cases. Single base insertion or deletion is a common way of gene remodelling. It yields frameshift mutations which are usually responsible for premature translational termination, resulting in nonactive protein species. However, they may also give rise to functional protein sequences, depending on the length of the open reading frame downstream of the mutation site. It has been reported that, in some cases, frameshift mutations do not impair gene function, even in the absence of any translational suppression of the frameshift (Polard et al. 1991; Fetten et al. 1991). Furthermore, retroviruses are known to possess polycistronic genes, with two different genes lying on overlapping nucleotide sequences, in different reading frames (Jacks et al. 1988), indicating that frameshifting may be compatible with gene function. Thus, it is not unreasonable to propose that a nonlethal frameshift mutation has occurred on the ancestral sequence of the gp41 gene after its divergence from the CD4 (or immunoglobulin) gene.

We have shown here that the analysis of alternate-reading-frame products may be an useful and novel approach for the investigation of evolutionary links and structure-function relationships. Divergent relations between gp41 and immunoglobulins, as proposed here, may open new perspectives in understanding strategies of viral genome evolution, virus-mediated autoimmunity, and HIV–T lymphocyte interactions.

Acknowledgments: We would like to thank Dr. Bradford Jameson for useful discussions during the preparation of the manuscript.

References

Alizon M, Wain-Hobson S, Montagnier L, Sonigo P (1986) Genetic variability of the AIDS virus: nucleotide sequence analysis of two isolates from African patients. Cell 46:63–74

Anand R, Thayer R, Srinivasan A, Nayyar S, Gardner M, Luciw P, Dandekar S (1989) Biological and molecular characterization of human immunodeficiency virus (HIV-1BR) from the brain of a patient with progressive dementia. Virology 168: 79–89

Argos P (1989) A possible homology between immunodeficiency virus p24 core protein and picornaviral VP2 coat protein: prediction of HIV p24 antigenic sites. EMBO J 8:779–785

Barker WC, George DG, Hunt L, Garavelli JS (1991) The PIR protein sequence database. Nucleic Acids Res 19:2231–2236 Suppl

Brake DA, Debouck C, Biesecker G (1990) Identification of an Arg-Gly-Asp (RGD) cell adhesion site in human immunodeficiency virus type 1 transactivation protein, tat. J Cell Biol 111:1275–1281

Burks C, Cassidy M, Cinkosky MJ, Cumella KE, Gilna P, Hayden JED, Kee GM, Kelley TA, Kelly M, Kristofferson D, Ryals J (1991) GenBank. Nucleic Acids Res 19:2221–2225 Suppl

Bushman FD, Fujiwara T, Craigie R (1990) Retroviral DNA integration directed by HIV integration protein in vitro. Science 249:1555–1558

Buskila D, Gladman D (1990) Musculoskeletal manifestations of infection with human immunodeficiency virus. Rev Infect Dis 12:223–235

Chee MS, Satchwell SC, Preddie E, Weston KM, Barrell BG (1990) Human cytomegalovirus encodes three G protein-coupled receptor homologues. Nature 344:774–777

Clayton LK, Sieh M, Pious DA, Reinherz EL (1989) Identification of human CD4 residues affecting class II MHC versus HIV-1 gp120 binding. Nature 339:548–551

Dalgleish AG, Beverley PCI, Clapham PR, Crawford DH, Greaves MF, Weiss RA (1984) The CD4 (T4) antigen is an essential component of the receptor for the AIDS retrovirus. Nature 312:763–767

Doolittle RF, Feng DF, Johnson MS, McClure MA (1989) Origins and evolutionary relationships of retroviruses. Q Rev Biol 64:1–30

Feng DF, Johnson MS, Doolittle RF (1984) Aligning amino acid sequences: comparison of commonly used methods. J Mol Evol 21:112–125

Fetten JV, Roy N, Gilboa E (1991) A frameshift mutation at the NH_2 terminus of the nucleoprotein gene does not affect generation of cytotoxic T lymphocyte epitopes. J Immunol 147: 2697–2705

Fitzpatrick DR, Snider M, McDougall L, Beskorwayne T, Babiuk LA, Zamb TJ, Bielefeldt-Ohmann HB (1990) Molecular mimicry: a herpes virus glycoprotein antigenically related to a cell-surface glycoprotein expressed by macrophages, polymorphonuclear leukocytes, and platelets. Immunology 70:504–512

Gojobori T, Mariyama EN, Ina Y, Ikeo K, Miura T, Tsujimoto H, Hayami M, Yokoyama S (1990) Evolutionary origin of human and simian immunodeficiency viruses. Proc Natl Acad Sci USA 87:4108–4111

Golding H, Shearer GM, Hillman K, Lucas P, Manischewitz J, Zajac RA, Clerici M, Gress RE, Boswell RN, Golding B (1989) Common epitope in human immunodeficiency virus (HIV)1-gp41 and HLA class II elicits immunosuppressive autoantibodies capable of contributing to immune dysfunction in HIV 1-infected individuals. J Clin Invest 83:1430–1435

Gooding LR, Sofola IO, Tollefson AE, Duerksen-Hughes P, Wold WS (1990) The adenovirus E3-14.7K protein is a general inhibitor of tumor necrosis factor-mediated cytolysis. J Immunol 145:3080–3086

Gotoh O (1982) An improved algorithm for matching biological sequences. J Mol Biol 162:705–708

Guldner HH, Netter HJ, Szostecki C, Yaeger E, Will H (1990) Human anti-p68 autoantibodies recognize a common epitope of U1 RNA containing small nuclear ribonucleoprotein and influenza B virus. J Exp Med 171:819–829

Helene C (1987) Chimie et modulation selective de l'expression des genes. La vie des Sciences Comptes rendus 4:17–37

Jacks T, Madhani HD, Masiarz FR, Varmus HE (1988) Signals for ribosomal frameshifting in the Rous Sarcoma Virus gag-pol region. Cell 55:447–458

Katz RA, Skalka AM (1990) Generation of diversity in retroviruses. Ann Rev Genet 24:409–445

Kieber-Emmons T, Jameson BA, Morrow WJ (1989) The gp120-CD4 interface: structural, immunological and pathological considerations. Biochim Biophys Acta 989:281–300

Klatzmann D, Champagne E, Chamaret S, Gruest J, Guetard D, Hercend T, Gluckman JC, Montagnier L (1984) T-lymphocyte T4 molecule behaves as the receptor for the human retrovirus. Nature 312:767–768

Kronenberg M (1991) Self-tolerance and autoimmunity. Cell 65: 537–542

Lipman DJ, Pearson WR (1985) Rapid and sensitive protein similarity searchers. Science 227:1435–1441

Maddon PJ, Dalgleish AG, Mc Dougal JS, Clapham PR, Weiss RA, Axel R (1986) The T4 gene encodes the AIDS virus receptor and is expressed in the immune system and in the brain. Cell 47:333–348

Moore KW, Vieira P, Fiorentino DF, Trounstine ML, Khan TA, Mosmann TR (1990) Homology of cytokine synthesis inhibitory factor (IL-10) to the Epstein-Barr virus gene BCRF1. Science 248:1230–1234

Myers EW, Miller W (1988) Optimal alignments in linear space. CABIOS 4:11–17

Navia MA, Fitzgerald PMD, McKeever BM, Leu CT, Heimbach JC, Herber WK, Sigal IS, Darke PL, Springer JP (1989) Three-dimensional structure of aspartyl protease from human immunodeficiency virus HIV-1. Nature 337:615–620

Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48:443–453

Oettinger MA, Schetz DG, Gorka C, Baltimore D (1990) RAG-1 and RAG-2 adjacent genes that synergistically activate V(D)J recombination. Science 248:1517–1523

Oldstone MB (1987) Molecular mimicry and autoimmune disease. Cell 50:819–820

Paque RE, Miller R (1991) Autoanti-idiotypes exhibit mimicry of myocyte antigens in virus-induced myocarditis. J Virol 65: 16–22

Polard P, Prere MF, Chander M, Fayet O (1991) Programmed translational frameshifting and initiation at an AUU codon in gene expression of bacterial insertion sequence IS911. J Mol Biol 222:465–477

Prabhakar BS, Srinivasappa J, Beisel KW, Notkins AL (1988) Virus-induced autoimmunity: cross reactivity of antiviral antibodies with self components. In: Schultheiss HP (ed) New concepts in viral heart disease. Springer-Verlag, Heidelberg, pp 168–178

Schattner A, Rager-Zisman B (1990) Virus-induced autoimmunity. Rev Infect Dis 12:204–222

Yokoyama S (1988) Molecular evolution of the human and simian immunodeficiency viruses. Mol Biol Evol 5:645–659

Received May 11, 1992/Revised October 2, 1992