II

Amino Acid Alignments

Construction of the HIV-1 Protein Alignments

The number of full-length gene sequences is still growing rapidly for all genes. The envelope master alignment now contains 285 full-length sequences. For the purposes of the printed alignments, we have had to limit the number of sequences dramatically. Here we list the criteria we have followed to make the selection.

First, we have decided to end the supremacy of the B clade sequences. More than half (144, to be precise) of the full-length envelope sequences are still subtype B, though the contribution of other subtypes is increasing. We have tried to balance the number of representatives of all subtypes in these alignments. For this, we had to make a heavy selection on subtype B sequences. We have tried to include as many 'classical' sequences as possible. A lot of follow-up work has been done based on lab strains such as HXB2, MN, SF2, and JR-CSF/JR-FL, so these strains are included in the alignments. Although the reference strain for this year's alignments is HXB-2, WEAU is still the reference strain for this Compendium's sister volume, the HIV Immunology Compendium, so this strain is included as well. Furthermore, within subtype B we have tried to represent sequences from diverse geographical origins, so as to represent a broad spectrum of variants. In the case of subtype B, this means that we have included African, Asian and Brazilian variants along with the 'Western' strains.

In a few cases, there were a lot of sequences from non-B subtypes. In these cases we have selected a few representative sequences from each dataset, again with an eye on maintaining geographical diversity. We have left all representatives of group O in the alignment, as these sequences are much more genetically diverse that the subtypes.

Sequences from isolates that are known recombinants have been included only when they belonged to an established 'Circulating Recombinant Form' (CRF). A CRF is a recombinant mosaic that is epidemiologically relevant, *i.e.*, one that has been found in two or more epidemiologically unrelated persons. There are presently four CRFs: AB(KAL153), represented by KAL153 and a series of AB sequences obtained from Kaliningrad IV drug users; AG(IbNG), represented by the isolates IbNG, DJ263, and DJ264; AGI(CY032), represented by the isolates CY032, PVMY and PVCH; and AE(CM240), consisting of all AE recombinants found so far. In the alignments, these sequences are indicated by a prefix of the letters indicating the subtypes they are recombinants of. All recombinants included in the alignments belong to one of the CRFs. Sequences of AB(KAL153) arrived too late to be included in these alignments.

Explanation of Symbols in Alignments

Symbol	Meaning
Alignment symbols	
`?` in consensus	no majority-rule consensus could be determined at this position
`x`	nucleotide missing from codon
`#`	frameshift, or codon contains N or illegal character
`$`	stop codon
Annotation symbols	
`|- -|`	domain boundaries
`/`	protein start point
`\`	protein end point
`\/`	splice site or exon join
`->`	start of overlapping coding region
`<-`	end of overlapping coding region
`*`	cysteine
`^^^ [NxS, NxT]`	glycosylation site
`^*^ [NCS, NCT]`	glycosylation site with cysteine
`CD4`	residue critical for CD4 binding
`cds`	coding sequence (indicates regions where two proteins overlap; the overlapping proteins use two different reading frames)
`MHR`	major homology region
`nls`	nuclear localization signal
`phos site`	phosphorylation site
`PKC`	protein kinase C binding
`Zn-motif`	Zinc finger binding motif

Sources of Annotation in the Alignments

Protein	Annotation	Reference
Gag	phos site Ser (111)	Yu, *J Biol Chem* **270**:4792 (1995)
Gag	MHR, (284-302)	Otteken, *J Virol* **70**:3407 (1996)
Gag	CyPa (205-241)	Braaten, *J Virol* **70**:4220 (1996)
Gag	vpr packaging domain LKSLFG (489-494)	Lu, *J Virol* **69**:6873 (1995) Kondo, *J Virol* **70**:159 (1996)
Nef	myristylation (1-7)	Huang, *J Virol* **69**:93 (1995)
Nef	(PxxP)3 (67-76)	Huang, *J Virol* **69**:93 (1995)
Nef	PKC (75-80)	Huang, *J Virol* **69**:93 (1995)
Nef	polypurine tract (89-97)	Huang, *J Virol* **69**:93 (1995)
Nef	Beta turn (128-131)	Huang, *J Virol* **69**:93 (1995)
Nef	PxxP (145-148)	Huang, *J Virol* **69**:93 (1995)
Vpr	alpha helix (16-34)	Cornelissen, *ARHR* **13**:247 (1997)
Vpr	H(S/F)RIG motifs (71-82)	Macreadie, *PNAS USA* **92**:2770 (1995)
Vpu	all annotations	Cornelissen, *ARHR* **13**:247 (1997)
Vpr	LR domain (60-82)	Wang, *Gene* **178**:7 (1996)