II

Amino Acid Alignments

Construction of the Protein Alignments

The number of full-length gene sequences is still growing rapidly for all genes. The envelope master alignment now contains almost 450 sequences, more than half of which are full length. For the purposes of the printed alignments, we have had to limit the number of sequences dramatically, although the full set of sequences is still available through our anonymous ftp site (ftp://ftp-t10.lanl.gov/pub/aids-db). Here we list the criteria we have followed to make the selection.

First, we have decided not to overrepresent the B clade sequences in the printed alignment. Almost half of the envelope sequences are still subtype B, though the contribution of other subtypes is increasing. We have decided to limit the number of sequences to 20 per subtype. Only subtype B frequently reached this limit, so a further selection had to be made to determine which 20 to include. For subtype B, we have tried to include as many ‘classical’ sequences as possible. Extensive 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. Furthermore, 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 Thai and Brazilian variants along with the ‘Western’ strains. We have not specifically attempted to include whole genomes in the alignments, because a separate section of this compendium is devoted to these.

Sequences from isolates that are known recombinants have been included in some cases when it has been established that they are not recombinant in the env gene, and there were only a small number of sequences (notably, subtypes E, F, and G). These mosaic sequences are indicated by a prefix of two letters indicating the subtypes that have contributed to their genome. An exception is the sequences of subtype E; all subtype E isolates found so far are recombinant and show the same mosaic pattern, so the prefix has been omitted. All subtype G sequences found so far are AG recombinants too, but since there exist many different mosaic patterns for subtype G, the prefix in this case has been maintained.

For subtypes other than B, there are less than 20 full-length sequences, and all available full-length sequences are included. In some cases (notably subtype H), we have used all available sequences, even if they are not full-length, because of the paucity of information on this subtype.

Recombinant (other than A/G and A/E) and unclassified sequences have been used in the alignment only when the number of sequences for a protein was very small and the sequence was known to be non-recombinant in that protein. Recombinants other than AE recombinants in subtype E are indicated with a prefix before the sequence name (for example, AD_MAL).

For reasons of standardization and synchronization with the HIV Immunology Database, the reference strain (shown on top of the alignments) is now WEAU instead of the overall consensus. WEAU is a primary isolate that has been dually sequenced and very carefully annotated, and it eliminates the disadvantages of working with an overall consensus (mainly, the existence of undefined positions).

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
Zn-motif	Zinc finger abinding 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)

Contents

PART II Amino Acid Alignments

A. HIV-1 Amino Acid Alignments

B. HIV-2/SIV Amino Acid Alignments

C. SIV AGM and SIV SYK Amino Acid Alignments