II

ALIGNMENTS

Organization

Part II of this compendium contains alignments of PV coding sequences and their corresponding protein sequences. The order of the alignments is E6, E7, E1, E2, E4, E5, L2, L1, and LCR. Fusion proteins such as E1^E4 are discussed in Part III. The LCR was operationally defined as the region after the end of L1 and before the first methionine of the E6 ORF. The LCR region is presented as a nucleotide alignment only, and contains only HPVs (no animal PVs). You can find your way in this section by looking at either the headline or the page number, both of which contain the gene name.

In previous years alignments have been generated progressively, beginning with pairwise alignments within the groups and then proceeding to inter-group comparisons. This procedure was largely carried out using the MASE sequence editor. In this compendium we have aligned the Supergroup A and Supergroup B sequences using a new approach—hidden Markov models (HMM), described in the article *Papillomavirus Alignments and Structure Predictions* by Farmer and Myers on page III-125 of this compendium. Supergroups C through E, and certain highly divergent viruses not assignable to a Supergroup were aligned by the old methods. This dual approach has, of necessity, made the presentation of this chapter slightly more complex than last year. The outline here shows the pattern of presentation for each gene.

First Gene (E6)

- Protein Alignment
  - HMM Alignment of Supergroup A
  - HMM Alignment of Supergroup B*
  - Alignment of Supergroups C–E and Others
- Nucleotide Alignment
  - HMM Alignment of Supergroup A
  - HMM Alignment of Supergroup B*
  - Alignment of Supergroups C–E and Others

Next Gene (E7)

⋮

etc.

**Note, because only a single Supergroup B E5 sequence exists (HPV5) no HMM alignment was run for E5 Supergroup B.*

How to Read the Alignments

The sequences have been grouped within the alignments according to the taxonomic system described in Chan et al., *J. Virol.* **69**:3074–83. Under this system, sets of relatively close sequences are termed “groups”, while sets of relatively close “groups” are termed “supergroups”. Each of the groups has at least two members, although in some cases, e.g. GroupA3 and GroupA11, only one member of the group has been sequenced over its complete genome, so that the group will have only one representative in many of the alignments. Some sequences clearly belong to a particular supergroup but are not related closely enough to any other sequence to justify including them in a group. In addition, a few PV sequences (FPV, MmPV, MnPV) are so distant from all other PVs that they cannot be considered as belonging to any of the supergroups. These sequences have been placed together at the very bottom of the alignments.

At the head of each group is a consensus-like sequence or most-likely sequence for that group. Each supergroup consensus sequence represents a consensus for all sequences in that supergroup (not including the individual group consensus sequences). Each supergroup consensus is placed at the head of the groups which are its members. Immediately following the supergroup consensus are sequences

belonging to the supergroup, but not assigned to any group, e.g. HPV54. A consensus sequence for those sequences belonging to no supergroup is given as "Unclass.con"

Each set of sequences is referenced to the consenusus sequence immediately above them. Agreement with the consensus sequence at any location is shown by a dash (-) while gaps are indicated by dots (. . .). Blank spaces within the alignment indicate lack of sequence information over that region. Occasionally, a nucleotide sequence will contain a percentage sign, (%), which indicates that the sequence appears to contain a frameshift indel at that position.

At the beginnings and ends of some of the alignments, some sequences may be separated by lines of arrows (➥). This indicates that these sequences are significantly longer than the rest, and that they continue below on the same page. The other exception is found in the E4, E5 and LCR alignments. These regions contain solid "separation bars" () throughout the alignment, to indicate that no significant similarity exists between sequences above and below the bar, and they could not be aligned for that reason. In some cases, it seems probable that this lack of similarity may be attributable to an absence of homology between the sequences.

PART II Alignments

This page intentionally left blank.