CONTENTS

Tables of contents are also to be found within the various parts of the compendium. The following provides an overview.

Contents

Acknowledgments

For the idea of a papillomavirus sequence database, as well as the support for its implementation, we thank Drs. Penelope Hitchcock and John LaMontagne, of the Division of Microbiology and Infectious Diseases of the National Institute of Allergy and Infectious Diseases, Bethesda, Maryland. Dr. Hitchcock is the Chief of the Sexually Transmitted Diseases Branch and Dr. LaMontagne is the Director of the Division.

We also thank Dr. Michelle Manos and coworkers for providing L1 sequences prior to publication, Dr. Tom Broker for supplying us with his extensive mailing list (that allowed us to initially contact each of you), Dr. Carl Baker for helpful discussions, and Ms. Anne Yesley for her technical assistance in the final weeks of the preparation of this compendium.

INTRODUCTION

This compendium and the accompanying floppy diskettes are the result of an effort to compile and rapidly publish all relevant molecular data concerning the human papillomaviruses (HPV) and related animal papillomaviruses. The scope of the compendium and database is best summarized by the five parts that it comprises: (I) HPV and animal PV Nucleotide Sequences; (II) Amino Acid and Nucleotide Sequence Alignments; (III) Analyses; (IV) Related Host Sequences; and (V) Database Communications. Information within all the parts is updated at least once a year, which accounts for the modes of binding and pagination in the compendium. In addition to the general descriptions below of the parts of the compendium, the user should read the individual introductions for each part.

Part I. HPV and Animal PV Nucleotide Sequences. Annotated nucleic acid sequences of HPV and related PVs are presented in a form close to that of the GenBank Sequence Library. Our few modifications of standard GenBank format were instituted to better serve the particular community for which this database is intended.

The LOCUS name or identifier of an entry may differ slightly from that found in the GenBank or EMBL libraries, but the ACCESSION numbers are identical for entries in all three databases. Thus each entry is universally and uniquely traceable. Large sets of PCR-derived sequences are represented using a single sequence entry followed by an alignment of all the sequences in the set. When available, the entire range of ACCESSION numbers for the PCR set is usually indicated within the representative entry. The SOURCE line provides information, when available, about the molecular clone from which a sequence has been derived. REFERENCES are limited to literature or personal communications having authority for the original sequence data; references that review sequence information, or that shed light upon the function or variation of coding and regulatory sequences, may be mentioned in the COMMENT and will be listed in Part V.

Entries in Part I are annotated within the sequence, while their GenBank or EMBL-formatted versions on the floppy diskettes make use of FEATURES tables. The hard-copy annotation includes coding regions, regulatory structures, splice sites, and other features of functional significance. The authority for this annotation is largely invariance, the recurrence of patterns such as `TATAA` and `AATAAA`. Although our practice has been to conservatively annotate, we caution the user against docility: sequence information regarding transcripts, for example, is far from certain or complete at this time. Part I is divided into nine subsections, A through I, concerned with HPV Genital-mucosal groups A-F, the cutaneous PVs, G, the EVs, H, and the animal PVs, I. The rationale for this grouping is presented in Part III.

Part II. Alignments. This section contains in alignment the amino acid and nucleic acid sequences of all known coding regions and open reading frames of HPVs and animal PVs. LCR (URR) alignments are also included. Consensus sequences and `AACC` consensus-like patterns are given with their respective alignments. Protein processing sites are annotated when known. The reader should consult the introduction to Part II for further explanation of the presentation and annotation of the alignments.

Part III. Analyses. This section is open-ended with the constraint that the sequence analyses and compilations be basic and of interest to the diversity of users. Currently the analyses include phylogenetic trees, synonymous/nonsynonymous substitution frequencies, and summaries of protein variability.

Part IV. Cellular Proteins. Part IV entries include coding sequences for cellular proteins involved with HPV regulation and pathogenesis. Entries are presented in the same format as Part I entries.

Part V. Communications. This part consists of i) information about the HPV database server

that is accessible through Internet; ii) a printed supplemental reference list and iii) diskettes. The supplemental reference list contains so-called “secondary references” that review sequence information and shed light upon the function or variation of coding and regulatory sequences. These references are captured from the molecular subset of MEDLINE as communicated through GENINFO at the National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health, Bethesda, Maryland.

The floppy diskettes contain the nucleic acid sequences from Part I and Part IV and the translated amino acid sequences of coding regions. As space permits, sequence alignments are also included to meet user requests. For the most current information regarding database files, see the READ.ME file on each diskette. Nucleotide entries are presented in GenBank format for North American users and in EMBL format for European users (unless otherwise requested). Similarly, amino acid sequences are in either PIR or Swiss-Prot format. The diskettes themselves are either 5.25” IBM-DOS format or 3.5” Macintosh format, depending upon what has been requested. If there is any trouble using these files with software designed to work with the format we have sent, please let us know the name of the program you are using and the file that it could not handle.

We are prepared to quickly enter both protein and nucleotide sequences into the Human Papillomaviruse sequence, and in the case of nucleotide sequences, oversee their entry into the GenBank and EMBL libraries. Submission of unpublished sequences is invited and encouraged. Sequence data or inquiries regarding the database should be addressed to

Charles Calef
Theoretical Division
T-10, MS K710
LANL
Los Alamos, NM 87545

(505)-665-1356; fax (505)-665-3493
e-mail: cxc@t10.lanl.gov

GLOSSARY

Sequence elements:

E1 The E1 orf encodes a 68–76 kD protein essential for plasmid DNA replication. The full-length E1 product is a phosphorylated nuclear protein that binds to the origin of replication in the LCR of BPV1. E1 has also been shown to bind ATP and to bind in vitro to the full length E2 protein called the E2 transcription transactivator (E2TA), thereby enhancing viral transcription. Binding to E2 also strengthens the affinity of E1 for the origin of DNA replication. In HPV-16, E1 has indirect effects on immortalization.

E2 The E2 orf (of BPV-1) encodes three proteins which regulate viral DNA transcription and replication. The full-length 40-58 kD E2 protein, the E2 transcription transactivator (E2TA), activates viral promoters by binding to E2-responsive enhancer elements. The function of this protein is repressed (perhaps by competitive binding) by two other E2 proteins, the E2 transcriptional repressor (E2TR) and E8/E2 transcriptional repressor (E8/E2TR). In HPV-16 and HPV-18 the E2 protein suppresses the promoter from which E6 and E7 (transforming) proteins are transcribed. When HPV-16 integrates into the host-cell chromosome, the integrity of the E1 and E2 orfs is disrupted, so that the normal repression of E6 and E7 is lost, with consequent overexpression of these transforming proteins.

E3 An E3 orf is present only in BPV1, BPV2, EEPV and BPV4. In all of these except BPV4, the orf overlaps both E2 and E4, whereas in BPV4 it overlaps E1. It is not known whether this orf is translated.

E4 The E4 orf is contained completely within the E2 orf. It often lacks an initiation codon, and is expressed from spliced transcripts. E4 gene products are found primarily in the cytoplasm of superficial keratinocytes, where they are extremely abundant. In the HPV1 genome, the E4 gene is expressed from the late promoter P(L).

E5 The E5 orf encodes a cell-transforming protein. It is one of the more poorly conserved orfs among the papillomaviruses, and often lacks an initiation codon. The E5 gene product is the smallest transforming protein yet identified. It may also cause proliferation of dermal fibroblasts in fibropapillomas.

E6 The E6 orf encodes a 16–19 kD cell-transforming protein. The E6 gene product contains four Cys–X–X–Cys motifs, indicating a potential for zinc binding; it may also act as a nucleic acid binding protein. In high-risk HPVs such as HPV-16, E6 and E7 proteins are necessary and sufficient to immortalize their hosts—squamous epithelial cells. The E6 gene products of high-risk HPVs have been shown to complex with p53, and to promote its degradation.

E7 The E7 orf encodes a 10–14 kD cell-transforming protein. The E7 gene product is a zinc-binding nuclear phosphoprotein. E7 binds pRB, p107 and p130.

E8 An E8 orf is present only among the bovine papillomaviruses and HPV6b. In BPV1, a 28 kD E8/E2 fusion product is involved in transcriptional regulation by repressing E2 transactivation. The E8 orfs of BPV3, BPV4 and BPV6 seem to be analogous to the E6 orf, which is missing in these three PVs.

L1 The L1 orf encodes the 56–60 kD major capsid protein. It is relatively well-conserved among all papillomaviruses. The carboxy-terminus of the L1 gene product is the site of two nuclear localization signals.

L2 The L2 orf encodes the 49–60 kD minor capsid protein.

L3 An L3 orf appears only in BPV4, DPV and HPV5b. It is not known whether this orf is translated.

L4 An L4 orf appears only in BPV4. It is not known whether this orf is translated.

LCR (Long Control Region–sometimes referred to as upstream regulatory region (URR) or non-coding region) Operationally defined as the region from the termination of the L1 orf to the first methionine of the E6 orf (some authors use the beginning of the E6 orf); contains various transcriptional regulatory motifs as well as the origin of replication.

Regulatory elements:

GRE (Glucocorticoid Responsive Element) Binds the glucocorticoid receptor, is a partial palindrome and consists of the consensus 5′-TGTACAN$_3$TGTCAT-3′ (Chan et al. *J Virol* **63**: 3261–9). In the presence of glucocorticoids, up-regulation of the HPV-16 P_{97} promoter is observed. The GRE is present in many of the mucosal papillomaviruses.

E2 binding site Partially palindromic sequence elements of the form 5′-ACCN$_6$G[GT]T-3′, these act as transcriptional regulatory elements through their interaction with various E2 products. The specific "core sequence" represented in the consensus by N_6 is largely determinant of the site's affinity for E2 products. These sites are well-conserved in the LCRs of most papillomaviruses.

YY1 binding site The transcriptional repressor YY1 has been shown to bind to an unusually diverse set of sequence elements. In papillomaviruses, the most common potential binding sites are similar either to the consensus 5′-AANATGGNS-3′ or to the sequence 5′-CTCCATTTT-3′. YY1 binding with sequence elements in HPV-16 and HPV-18 silences promoter activity, probably by interfering with transcriptional initiation.

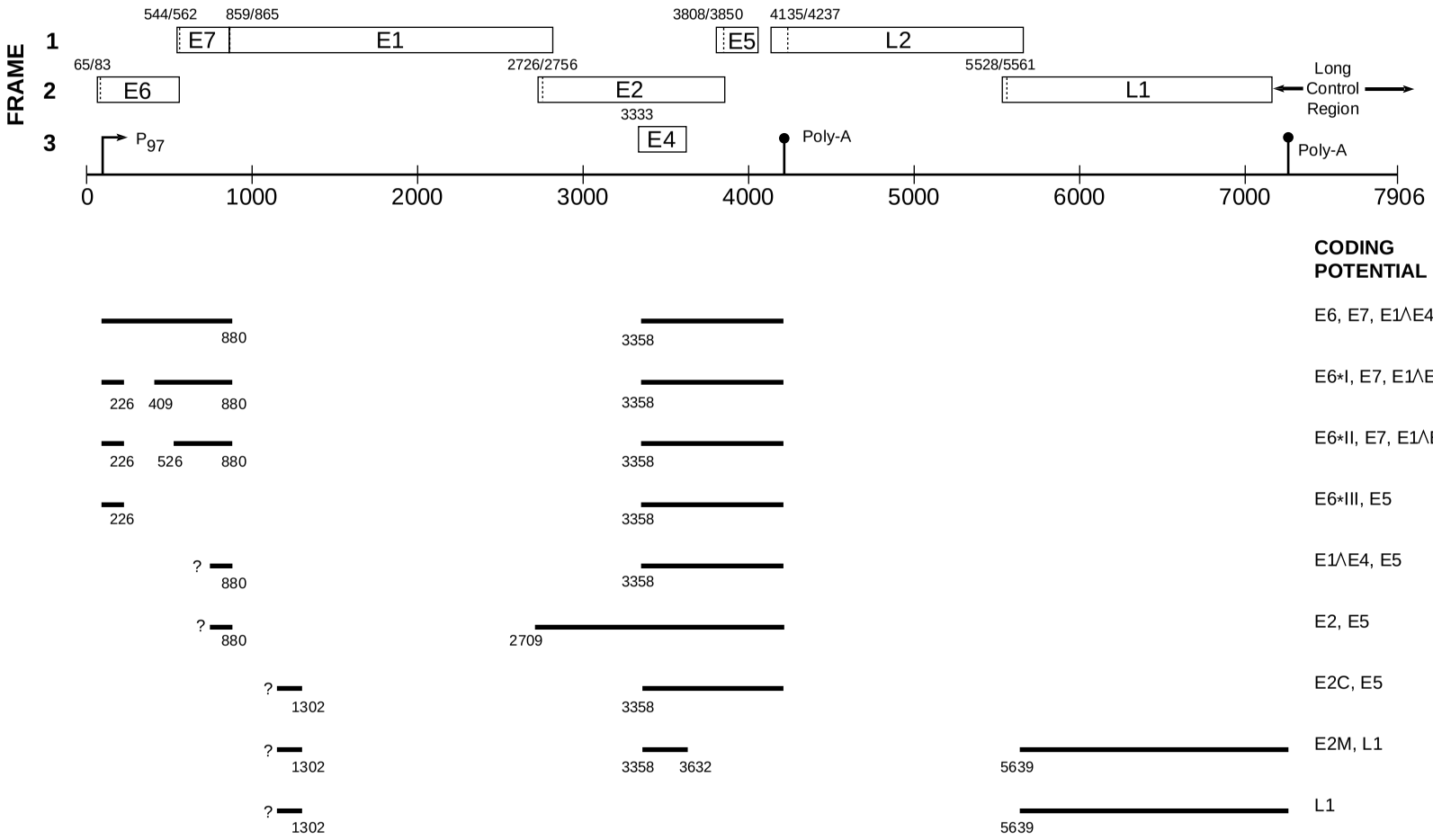
NF-1 In papillomaviruses, NF-1 binding sites have been discovered in the form 5′-TTGGC-3′. Sequences of the form 5′-TTGGC[AT]-3′ are much more frequently found than those of the form 5′-TTGGC[GC]-3′, and the latter seem to be considerably less effective in binding NF-1 in vitro. The "mucosal" papillomaviruses all have relatively well-conserved clusters of NF-1 binding sites in their LCRs which have been shown to be necessary for enhancer activation.

SP-1 SP-1 binding sites located in the promoters of diverse papillomaviruses have the consensus 5′-NGGNGN-3′. These have been shown to be involved in transcriptional regulation.

AP-1 AP-1 binding sites in papillomaviruses are typically similar to the sequence 5′-TGACTAA-3′. Involved in transcriptional regulation.

PVF (Papillomavirus enhancer associated factor) A transcription factor binding to a relatively well-conserved sequence in the LCRs of the mucosal papillomaviruses. This integrity of this sequence in the enhancer of HPV-16 (5′-AAGCACATAT-3′) has been shown to be vital to enhancer function.

Nuclear localization signal (NLS) specific sequences of basic amino acids direct the protein from the cytoplasm to the nucleus. Two NLS sequences have been reported for HPV16; both can be found at the carboxy terminal end of L1.



This diagram (compiled from Baker,C.C., The genomes of the papillomaviruses, in *Genetic Maps* Cold Spring Harbor Lab Press, and Sherman,L. et al., *Int. J. Cancer*, **50**:356–364) illustrates the open reading frames (orfs) and spliced mRNAs of HPV-16. The early (E) and late (L) orfs are shown as open rectangles in their proper reading frames. The numbers separated by a / at the 5′ end of each orf show the nucleotide number of the orf start and the number for the first ATG in each orf respectively. E4 contains no ATG. The dotted line near the left end of each orf marks the position of the first ATG. The P_{97} promoter and two polyadenylation sites are marked on the linearized genome scale. The nucleotide numbers assume that a nucleotide has been inserted at position 1137 of HPV16, i.e., at the position of the presumed deletion in this sequence. For this reason all numbers $>$ 1137 in this diagram are one greater than the corresponding positions in the HPV-16 sequence on pages I-A-3 to I-A-8.

Located below the genome scale are diagrams of spliced mRNAs. The exons are illustrated by heavy horizonatal lines, while the introns are indicated by gaps. The numbers written below the line indicate the 5′ and 3′ splice junction positions. The numbers give the position of the last nucleotide in the exon before the splice and the position of the first nucleotide of the exon following the splice. The potential coding regions are listed at the right. In that list a ^ symbol between two gene name (e.g. E1^E4) indicates a fusion product. The * symbol indicates different forms of the E6 product.