CONTENTS

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

Acknowledgments

The Division of Microbiology and Infectious Diseases of the National Institute of Allergy and Infectious Diseases, Bethesda, Maryland provides funding for the HPV Sequence Database and Analysis Project through an interagency agreement with the U.S. Department of Energy and Los Alamos National Laboratory. We thank Dr. Penelope Hitchcock, Chief of the Sexually Transmitted Diseases Branch in that Division, and Dr. Leigh Sawyer, the Project Officer for the database, for their encouragement and support.

The image on this year's cover was kindly provided by D. Belnap, N. Olson, and T. Baker, Purdue University. It shows the structure of human papillomavirus type 1 (HPV-1) as seen by cryoelectron microscopy (background) and three-dimensional image reconstruction (foreground).

- (Foreground, top) Surface view of HPV-1 oriented to show near the center of the image the two types of pentameric capsomeres that form the capsid: pentavalent (surrounded by five other capsomeres) and hexavalent (surrounded by six others).
- (Foreground, bottom) Views inside HPV-1, after the reconstruction was computationally 'sliced open' to expose the chromatin core.

References: Belnap et al. (1996) *J. Mol. Biol*. **259**:249–263 Baker et al. (1991) *Biophys. J*. **60**:1445–1456

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 usually differs 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. 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 annotate conservatively, we caution the user against docility.

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 with the exception of variant alignments that appear in Part I. 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 diverse users. The analyses presented in the 1995 compendium comprised maps of RNA transcripts, BLAST searches of conserved protein motifs, in-depth reviews of LCR, E6, E7, L1, and L2 sequences, presentations of methods for evaluating recombination events, and a discussion of PV variation. In '96 we present more maps of transcripts, in-depth reviews of the E2, E4, and E5 proteins, a summary of HPV type prevalences, and a discussion of PV sequence alignment and structure prediction.

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) instructions for accessing our data and compendia by computer through the World Wide Web; ii) a printed list of papillomavirus references for the years 1995 and '96, and iii) diskettes. The 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.

Introduction

The floppy diskettes contain the nucleic acid sequences released in 1996 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 3.5" 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 papillomavirus sequence database, 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 normally 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 of BPV1 and probably other PVs. E1 has also been shown to bind ATP, DNA, and to the full length E2 protein, called the E2 transcription transactivator (E2TA). Binding to E2 strengthens the affinity of E1 for the origin of DNA replication. In HPV-16, E1 has indirect effects on immortalization.

E2 The E2 orf encodes a 48 kDa highly phosphorylated protein. E2 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 and heterodimer formation) by two other E2 proteins, the E2 transcriptional repressor (E2TR) and E8/E2 transcriptional repressor (E8/E2TR). Partially palindromic sequences, 5'-ACCN6G[GT]T-3', are dimeric-E2 binding sites in PV promotors; BPV-1 may contain as many as seventeen such sites. 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, with the result that repression of E6 and E7 may be affected.

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. In the HPV-16 genome, E4 is expressed from the late promotor, P(L). E4 gene products are found primarily in the cytoplasm of superficial keratinocytes, where they are reported to induce collapse of the cytokeratin matrix.

E5 The E5 orf encodes a small hydrophobic protein typically found in membrane compartments including the Golgi. In BPV-1, it is a 7 kDa cell-transforming protein, which appears to exert its effect through stimulation of growth factor receptor signal transduction pathways. It binds to and activates the PDGF receptor (platelet-derived growth factor), and also binds the 16kDa pore protein (ATPase subunit) and a 120 kDa adaptin-related protein of fibroblasts. It is one of the more poorly conserved orfs among the papillomaviruses.

E6 The E6 orf encodes a 16–19 kDa multifunctional protein present at extremely low levels in both the nucleus and the cytoplasm. The E6 gene product contains four Cys–X–X–Cys motifs, indicating a potential for zinc binding; the relatively large fingers created by these motifs are characteristic of some transcriptional transactivator proteins, and E6 has this capability. E6 has transforming and immortalizing capabilities in high-risk HPVs such as HPV-16. The E6 gene products of high-risk HPVs have been shown to form a complex with p53, affecting the transcriptional regulatory activity of p53 as well as targeting its degradation. E6 also binds E6BP (or ERC-55), a calcium-binding protein found in the endoplasmic reticulum. Splicing internal to the E6 orf creates truncated E6 products, the meaning of which is unclear.

E7 The E7 orf encodes an acidic 10–14 kD phosphorylated protein with transforming and transcriptional regulatory functions similar to what is seen in adenovirus E1A protein and SV40 large T antigen. The E7 gene product contains two cysteine arrays capable of forming a zinc-finger. It is found in both the nucleus and the cytoplasm. E7 binds pRB (retinoblastoma-susceptibility protein), thereby modulating cell cycle control; E7 also binds 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. L1, the most antigenic of PV proteins, is weakly phosphorylated and does not bind DNA. It can be glycosylated and cross-linked through disulfides, but the implications of these changes are unclear. (Glycosylated forms have not been

reported in virions). It is relatively well-conserved among all papillomaviruses. L1 has self-assembly capacity and is the overwhelmingly predominant molecule in the viral capsid.

L2 The L2 orf encodes the 49–60 kD minor capsid protein, which is highly phosphorylated and binds DNA. L2 migrates in a gel as if it were a 73 kDa protein. Unlike L1, L2 does not self-assemble nor does it link to itself.

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 noncoding 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.

Landmarks of the Genome

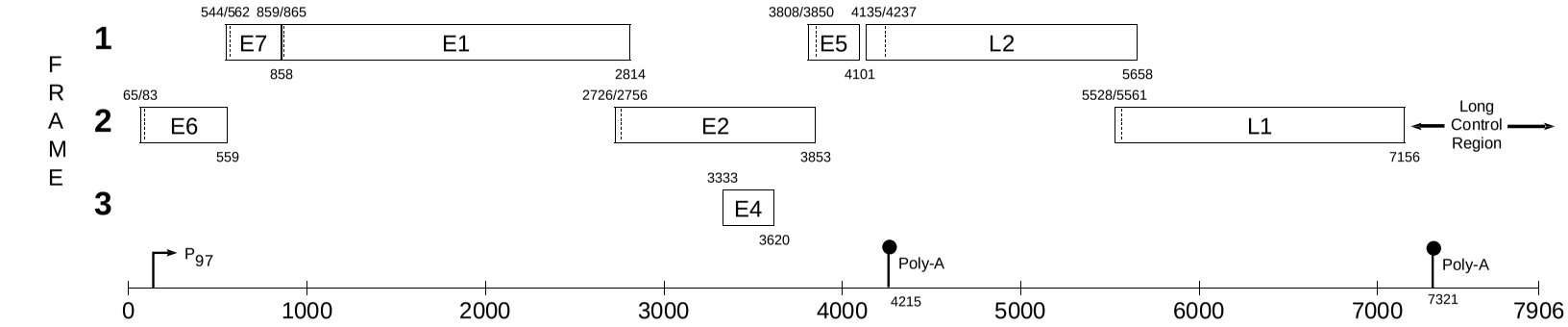
A map showing the open reading frames and mRNA transcripts of HPV-16 is presented on the facing page. The significant ORFs are shown in their proper reading frames as rectangles. At the upper left end of the rectangle are two numbers. The first corresponds to the nucleotide (nt) position of the ORF start, the first nucleotide following a stop codon. The second number records the nt position of the first ATG, which is also indicated by a dotted line within the rectangle. In the case of the E4 ORF in which no ATG occurs, only the nt position of the ORF start is present in the upper left corner. The position of the last nt in the stop codon of each ORF is printed below the lower right corner of the rectangles.

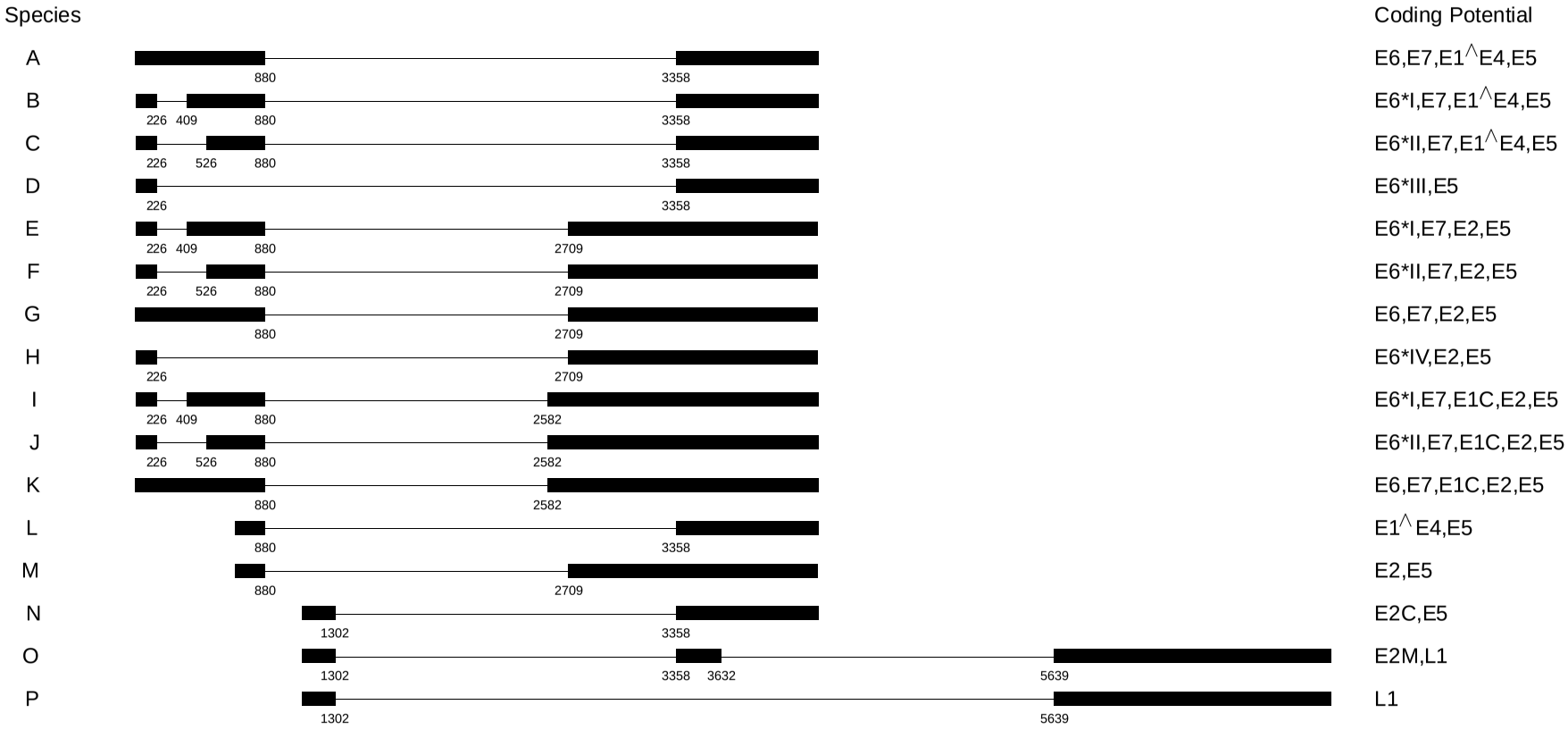
Below the ORFs is a scale of the genome divided into thousands. On the scale are placed the positions of promotors (represented by arrows) and the poly(A) signals. The exact position of the poly(A) signal is printed below the scale line.

Located below the genome scale are diagrams of mRNA species, most of which are spliced. The exons are illustrated by heavy black lines, while the introns are indicated by black hairlines between. The numbers printed below the lines indicate the 5′ and 3′ termini of the RNAs, and the 5′ and 3′ splice junction positions. The splice junction 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. Where 5′ or 3′ ends of the RNAs are uncertain, no nt position is given. The coding potential of the transcripts 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.

Maps of other papillomavirus genomes are presented in Part III.

HPV16R





Landmarks