Review

J Biomed Sci 1995;2:90–104

The Human Papillomavirus Database

Andrew D. Farmer
Charles E. Calef
Kim Millman
Gerald L. Myers

Theoretical Biology and Biophysics (T-10), Theory Division, Los Alamos National Laboratory, Los Alamos, N.M., USA

Key Words
Papillomaviruses
Genetic sequence database
Sequence analysis
Phylogenetic trees

Abstract
Papillomaviruses are responsible for a variety of diseases in humans and animals, ranging from harmless skin warts to lethal cancers. They also make up one of the most genetically diversified families of viruses known, and could represent a model system of DNA-virus evolution. A specialized genetic sequences database, The Human Papillomavirus Database and Analysis Project, was recently established in an effort to provide database services that are specific to papillomaviruses to the research community and to perform a variety of sequence-based analyses. This review is intended to present the scope of the information currently contained in the database and to outline some of the analyses that have been performed on the genetic sequences. These analyses will address issues including phylogenetic relationships, recombination events, selective pressures on different genes and the possibility of cross-species transmission in the case of the papillomaviruses.

A Brief Introduction to the Papillomaviruses

Papillomaviruses are members of the viral family Papovaviridae, whose name embodies the three distinct genera grouped together by classical taxonomy to form this family: *pa*pillomaviruses, *po*lyomaviruses and *va*cuolating viruses. The rationale for their inclusion in the same family is that the viruses of all three genera possess circular genomes composed of double-stranded DNA and enclosed in the mature virion by a nonenveloped, icosahedral capsid protein. The papillomaviruses are present in both human and animal hosts and have been shown to be responsible for a wide range of clinically distinct, proliferating lesions of epithelial tissues, including many kinds of warts and tumors, both malignant and benign. The primary stimulus for most research in the field is the strong association of papillomaviruses with diverse forms of human cancer; they are currently thought to be responsible for approximately 10% of all human cancers. Cervical cancer, with which they are especially associated, is the most prevalent form of cancer encountered in developing countries [28]. Papillomaviruses are considered the leading sexually transmitted pathogen; furthermore, infection by papillomaviruses appears to be linked with infection by other sexually transmitted viruses, in particular human immunodeficiency virus (HIV) [16].

Initially, it was believed that a single genetic form of the virus was responsible for all of the diverse growths associated with papillomavirus infection, and that the differences in clinical effects were caused by differences in the host cells and cofactors [22]. Subsequent research has revealed, however, that the genetic diversity of the papil-

Received:
November 29, 1994
Accepted:
December 12, 1994

Gerald. L. Myers
Theoretical Biology and Biophysics (T-10)
Theory Division
Los Alamos National Laboratory
Los Alamos, NM 87545 (USA)

lomaviruses is extraordinarily broad – at present, there are nearly seventy recognized 'types' of human papillomaviruses [8], close to twenty 'types' in other animals (including several nonhuman primate, rodent, ungulate and avian papillomaviruses) [24], and many recently discovered papillomaviruses with novel sequences that will probably be recognized as distinct 'types' in the near future [2]. For all this diversity, however, there is abundant evidence indicating that the papillomaviruses by themselves form a *monophyletic* group, i.e., the set of all descendants of the most recent common ancestor of all known papillomaviruses includes only sequences that would themselves be classified as papillomaviruses.

Monophyletic groups are most easily identified when all members of the group share some objective characteristic that originated with their common ancestor and which can serve in practice to define the group. In the case of the papillomaviruses, this defining characteristic is provided by certain broad similarities at the genomic level. The genomes of all known papillomaviruses are of comparable length (from 8–10 kb), display similarly skewed base compositions (31% A, 28% T, 22% G, 19% C), and possess a set of open reading frames (ORFs) distributed exclusively on one strand of the genome in a well-conserved manner. These ORFs are divided into two distinct sets according to their location in the genome, and corresponding to the time of their expression in the replicative cycle of the virus. The 'early' ORFs (so-called by analogy with the 'early' region found in the polyomaviruses) code for proteins responsible for replicative and transcriptional functions (E1 and E2), and transforming (E4–E7) functions; the 'late' ORFs code for the major and minor capsid proteins of the virus (L1 and L2, respectively). The majority of the papillomaviruses' ORFs display many conserved subsequences that allow one to generate relatively unambiguous alignments of the otherwise extremely divergent gene subsequences. In addition, a noncoding region (long control region, or LCR) that is found in all papillomaviruses between the late and early genes (in this order, reading 5′ to 3′) contains the origin of replication and many sequence elements, conserved to differing degrees, that are involved in transcriptional regulation, some of which seem to be specific to papillomaviruses [5].

The classification scheme of 'types' that has been adopted as a means of comprehending papillomavirus diversity is based upon their genotypic similarity, rather than serological reactivity, as in the case of some viruses. Originally, the criterion for considering two papillomaviruses as distinct types was that they should show less than 50% similarity in liquid hybridization experiments [6]. The results of these experiments, however, depended largely upon the distribution of regions of similarity within the genome and, we now know, did not always give accurate estimates for genetic distance. Thus, with the advent of sequencing technology, this criterion was considerably refined, requiring that two papillomaviruses differ by at least 10% at the nucleotide level over their E6, E7 and L1 ORFs in order for them to be considered distinct types [8]. The set of papillomaviruses sequenced to date includes only a few instances of sequences with any significant degree of divergence (greater than 2%) that are not also above this limit of 10%. (Sequences displaying a level of divergence below 2% are referred to as 'variants'; those diverging from 2–10% are often referred to as 'subtypes', although this term was originally defined in terms of restriction digest pattern polymorphisms [3].) In a sense, the size of the discontinuities observed in the genetic distances have made the arbitrarily defined limit of divergence for the types seem a fortunate choice – with such a criterion, it migth have been expected that many intermediate types would be found that could be classified as subtypes of more than one of the prototypic strains. Nevertheless, this classification scheme possesses some weaknesses both from practical and theoretical perspectives. From the former point of view, it is clear that a system encompassing over seventy distinct entities and lacking more general hierarchical groupings does not lend itself to easy assimilation. Further, as the type designations are simply numbers assigned sequentially with the isolation and genetic characterization of the sequences, the resulting nomenclature is informative only from a historical perspective and gives no insight into phylogenetic groupings or clinical associations. Who would be able to discern from their names alone, for example, that human papillomavirus (HPV)-4 and HPV-65 are phylogenetically linked and clinically similar? The definition of the types is also somewhat deficient from a theoretical point of view, as it takes no account of the important distinction between nucleotide differences that also correspond to differences at the amino acid level and those that do not. Thus, under the present scheme, it is entirely conceivable that two sequences could qualify as distinct types, yet show little or no difference in their gene products.

The current state of papillomavirus research is quite advanced from a molecular standpoint and seems to possess great potential for further gains. Already, recombinant DNA technology has overcome many of the difficulties imposed upon papillomavirus research by the lack of an effective in vitro system of viral propagation for most

of the papillomavirus types. Because the papillomaviruses are one of the most genetically diverse virus families, a deeper understanding of their relationships would probably make them a model system for the study of the evolution of DNA viruses. Further, their overwhelming clinical diversity makes it extremely likely that their study will lead to many important advances in our understanding of the biology of their hosts. The realization of this potential will depend to a large degree upon the extent to which research efforts may be carried out according to a rational framework.

The HPV Database – A Specialized Database

The HPV Database and Analysis Project was established in order to provide the services of a *specialized molecular-sequences database* (at no cost) to the papillomavirus research community. An analogous project was initiated in 1986 for HIV, and since that time, the number of sequences included in the HIV Sequence Database has doubled every year, due in part to the continuing variation of the virus, and in part to the widespread application of polymerase chain reaction (PCR) sequencing technology [18]. In the field of papillomavirus sequencing, although some authors have argued that it is unlikely that a large number of sequences that would qualify as novel types remain to be discovered [27], several factors suggest that the number of papillomavirus sequences may rise sharply in the near future. First, not all papillomaviruses that have been recognized as distinct types have been sequenced over their complete genomes, having been given their status as types on the basis of hybridization experiments. Second, comparatively little analysis has yet been performed upon intratypic variation of the papillomaviruses; what has been accomplished to date has been largely restricted to 'variants' of a few select types that are of particular interest because of their association with human cancer (e.g., HPV-16 and HPV-18, probably the most extensively studied of all the types) [4, 8, 21]. Although there seems to be little intratypic variation within these particular types, it would be of great interest to ascertain whether or not this is also the case with other papillomavirus types, for it could provide valuable information as to the relative genetic stability of the various types and give insight into the mechanisms of papillomavirus evolution. Lastly, the use of PCR technology for obtaining fragmentary sequences has been relatively limited in the field of papillomavirus research. Recent studies have demonstrated that sequence fragments under 200 bp in length obtained through the use of PCR can provide enough information in themselves to construct accurate phylogenies and indicate probable candidates for the status of novel types [2]. It seems likely, therefore, that the examination of subgenomic sequence fragments through the use of PCR techniques will be much more widespread in the papillomavirus community in the future. For all these reasons, we must expect that the number of papillomavirus sequences available to the community will increase dramatically in the near future. This projected increase in the amount of sequence information in the field is one of the most powerful arguments in favor of the present effort to create a specialized sequence database.

The fundamental ideas behind a specialized database require some preliminary explanation. Unlike the well-known comprehensive genetic database available to the community (GenBank, EMBL, DDBJ), a specialized database concentrates on a subset of all available sequence information, related in some way to the specific organism in question. This focus obviously frees it from some of the limitations imposed upon general databases both by the vast amount of information which they have to process and by the overwhelming genetic diversity of the sequences within their scope. Moreover, the association of the database with a specific research community makes possible a sort of dialogue between the two – not only can the database be responsive to the specific needs of the community, it can also act as a directive force to new avenues of exploration.

In its *curatorial* capacity, a specialized database can provide the basic services of the more general databases at a higher level of accessibility to the community. At present, sequence information can be obtained from the HPV database as hard copy in the Human Papillomaviruses compendium [19], on floppy diskettes in DOS or MAC format, or via anonymous FTP at atlas.lanl.gov (fig. 1). Common interactions between the research community and the database, such as searching for specific sequences, obtaining new sequences or updating those already in store, will be made considerably more efficient. Further, the specific manner in which the database presents its information can be tailored to the particular organism with which it is concerned. Sequence entries are provided with extensive comments concerning clinical information and prevalence data, as well as 'in site' annotation detailing functional sequence elements. In addition, sequence alignments comprehending all sequenced papillomavirus types have been created for each of the coding regions, as well as for the LCRs. Both the sequence

Instructions for Accessing ftp Server "atlas"

In the following instructions, what you actually type at your terminal is shown in **boldface** type [with an explanation shown in square brackets]. What the server responds with is shown indented in *italic* type.

ftp atlas.lanl.gov
Name (atlas.lanl.gov):
anonymous
Guest login ok, send ident as password.
Password:
[enter your e-mail address as ident]
Guest login ok. access restrictions apply.
cd pub [change to the directory called pub]
CWD command successful
cd papilloma
ls [this lists the files and directories inside of papilloma]
SWISS-PROT-files, RoadMap, Alignments, ReadMe, EMBL-files, GenBank-files
pwd [to show your location in the file hierarchy]
"/pub/papilloma" is current directory.
cd GenBank-files/Human-papilloma [move down 2 directories]
ls
[all the GenBank files are listed]
get HPV47.gb [this copies the file HPV47.gb to your home computer]
bye [this disconnects you from the server]

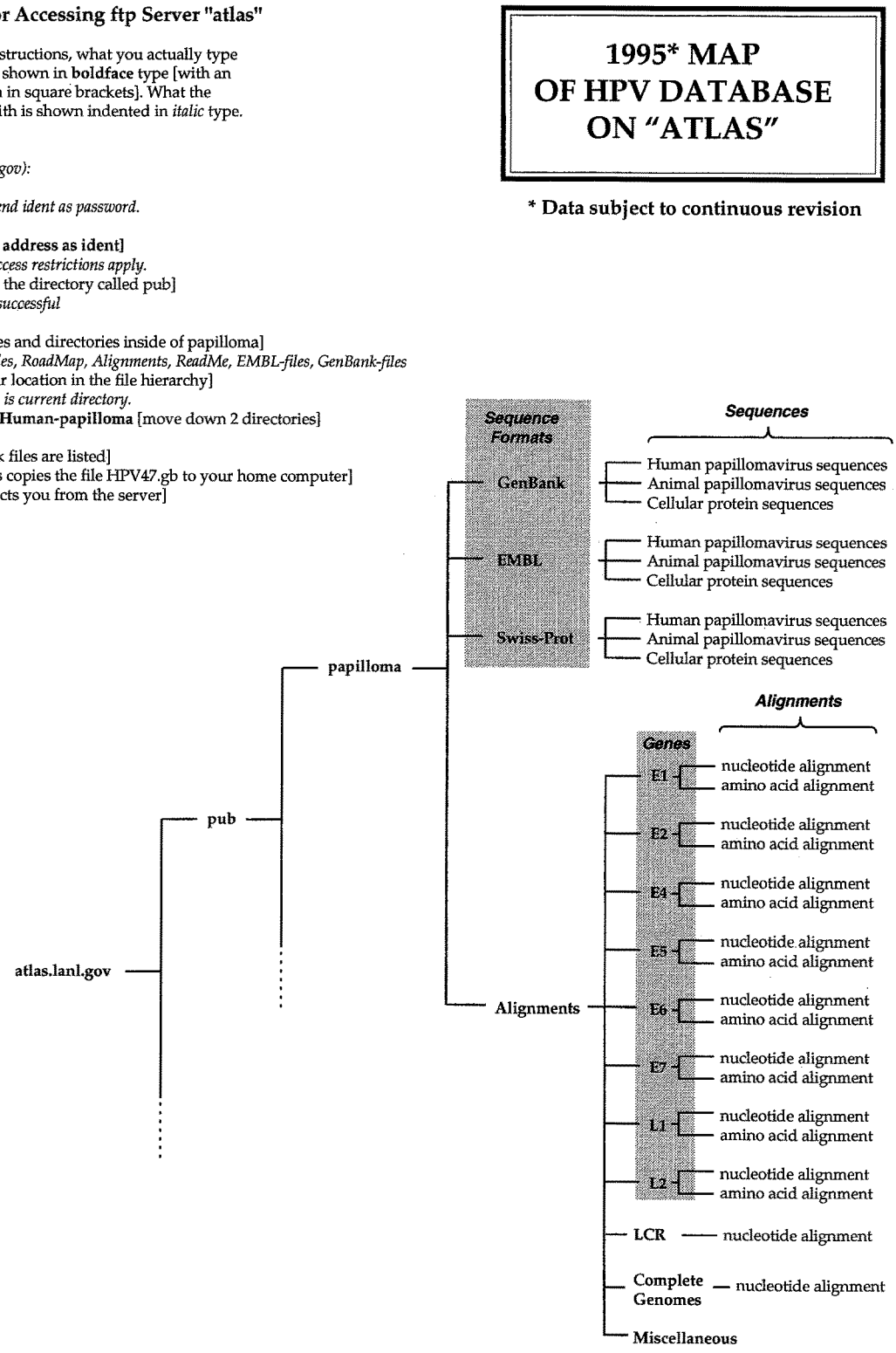

Fig. 1. 1995 map of HPV database on 'atlas'.

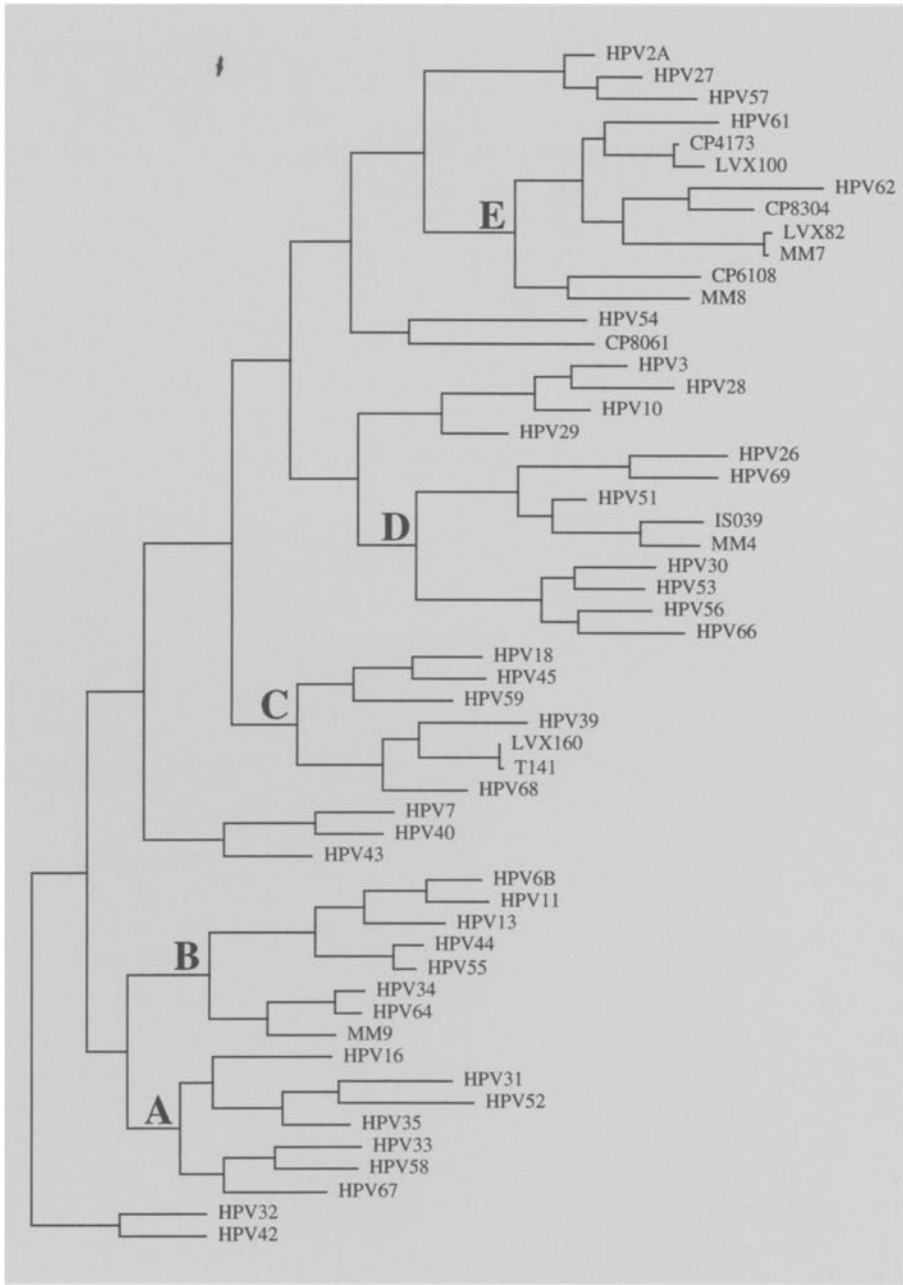

Fig. 2. A phylogenetic tree created using weighted parsimony on a set of 54 HPVs. The papillomavirus types included in the tree were sequenced over a 291-bp L1 fragment flanked by the My09/My11 primer pairs of Bernard et al. [1]. The groups adopted for the sequences in the tree for the purpose of presentation are indicated by letters placed at the node representing the most recent common ancestor of all sequences in each group. All sequences not belonging to these groups (A–E) were assigned to the 'catchall' Group F. Groups G–I are not represented in this tree, which includes only the 'genital/mucosal' HPVs and those HPVs which typically cluster with this group, although their clinical associations are more diverse. The data set used to construct the tree included 213 variable sites.

entries and the alignments have been organized according to 'groups' that were defined by phylogenetic analysis performed on a set of 54 partial L1 nucleic acid sequences (fig. 2). These groups have not been proposed as a new taxonomic level, but were adopted merely to enhance the intelligibility of presentation. In the database alignments, for example, sequences were clustered into their respective groups, and consensus sequences were generated for each of these groups.

While most of the curatioral functions of a specialized database are essentially only customized versions of functions provided by the large sequence libraries, a critical difference exists between the two sorts of database in another respect: with a specialized sequence database, compilation and analysis are intertwined. Analyses may be regarded as representing a higher order of sequence information, an order that is unattainable by the general databases because of their very comprehensiveness. Some

of these analyses may be performed from a naive perspective, making use of no information other than that provided by the raw sequence data. The sequence alignments and phylogenetic groupings mentioned earlier fall into this class. Other analyses may attempt to incorporate external sources of information, e.g., analyzing sets of sequences grouped together on the basis of their clinical associations. The remainder of this review will be devoted to a discussion of the various analyses that have already been published by the HPV database and of these that are planned for upcoming releases. Most of the analyses that have been performed so far have used the naive approach, attempting to come to an understanding of the evolutionary relationships between the papillomaviruses by concentrating exclusively on the 'text' that is represented by the available sequences and referring to the 'context' of their clinical associations only when this seems to be warranted by the results of the analysis.

Analyses

Phylogenetic Trees

One of the most basic forms of analysis that may be performed using genetic sequences is the attempt to reconstruct their evolutionary history in the form of phylogenetic trees. A great number of different methods have been proposed to address the problem of reconstructing the evolutionary history of organisms on the basis of the available genetic data [for a comprehensive review of these methods, see ref. 9]. Without exception, however, each of these methods depends upon a fundamental assumption of the existence of *homology* between the organisms under examination. When genetic sequences are used as the relevant data, the character states assumed to be homologous are nucleotide or amino acid positions. The evolutionary mechanism implied by the sequence differences includes not only transformations between characters (substitutions) but also insertions and deletions of characters (indels).

This assumption of homology is realized in the positioning of sequence arrays into an alignment matrix, and all different methods of phylogenetic analysis in some way make use of these sequence alignments to assess evolutionary relationships. In a multiple-sequence alignment, each row corresponds to a particular sequence, and each column is supposed to contain only homologous characters. Any two characters in a column are thus postulated to be related by the evolutionary process through an undetermined number of intermediate steps. Typically, it is necessary to introduce 'gap characters' into various sequences at certain columns to indicate that between those sequences that possess the gap character at this position and those that do not, one step in the evolutionary process was either an insertion or a deletion of the character, depending on which sequence is the more ancestral. In order to evaluate the quality of different possible alignments between sequences, it is necessary to utilize a scoring matrix that assigns different costs to the various relations between homologous characters in different sequences. Generally, these scoring matrices take into consideration only character substitutions, which may be differentially weighted to favor certain transformations over others (and insertion/deletion events), although other evolutionary mechanisms may be involved (e.g., duplications and inversions).

To create alignments for the papillomavirus coding regions, we first aligned the amino acid sequences of the gene products using the PIMA algorithm developed by Smith and Smith [23]. PIMA uses a scoring matrix based on physiochemical similarities between amino acids, assigning smaller penalties to substitutions that are 'conservative' from the perspective of protein function. The algorithm also involves a 'progressive' approach to the problem of multiple-sequence alignment. The basic idea behind a progressive alignment strategy is that once two relatively close sequences have been aligned to each other, the inferred homologies between the characters implicit in this alignment will not be altered when these sequences are compared to more distant sequences in the set. The amino acid alignments created with this approach were then used to generate nucleotide alignments, putting each nucleotide codon in the position of its corresponding amino acid; these were then manually edited to account for regions of nucleotide similarity not reflected at the protein level.

After an alignment has been created for the sequences under consideration, a phylogenetic tree representing their evolutionary history may be created using one of the many available tree construction algorithms. The performance of many of these methods has been tested by comparing the results given by each method using sequence sets (actual or simulated) with a priori known phylogenies [10]. The degree of accuracy possessed by different algorithms depends largely upon certain parameters of the evolutionary process, such as the rate of variation between sequences and the extent to which this rate varies in different parts of the tree. One method that has been found to perform extremely well under a wide range of conditions, particularly when divergence between all taxa

Table 1. Substitution frequency matrix for a papillomavirus L1 tree and the resultant weighting matrix created using the software packages PAUP and MacClade

		To				To			
		A	T	G	C	A	T	G	C
From	A		0.09	0.14	0.11		11	7	10
	T	0.06		0.04	0.08	16		23	13
	G	0.11	0.03		0.04	9	34		28
	C	0.11	0.16	0.04		9	6	24	

in the data set is large, is that of weighted parsimony. In unweighted-parsimony analysis, the sum of the branch lengths in the path between the nodes corresponding to any two taxa is equal to the number of substitutions in the postulated evolutionary path between the two (typically, regions containing gaps are removed from the data set, as alignments are generally most uncertain in these regions). The most parsimonious tree (or set of trees) is defined as that in which the sum of all branch lengths is minimized, thus representing the globally minimal evolutionary path. These most parsimonious trees are taken to be the most accurate representations of the true phylogeny, under the principle that nature tends to take the 'path of least action'. Weighted parsimony differs from simple parsimony in that it uses a nonuniformly weighted matrix to assign different costs to the various kinds of transformations between characters. Thus, the branch lengths on a weighted-parsimony tree take into account not only the number of substitutions between sequences, but the kind of substitutions involved as well. The weights assigned to the different kinds of character transformation are empirically determined, by first creating a tree based on unweighted parsimony, then calculating the frequencies with which the various substitutions take place in this tree. The actual cost assigned to each kind of transformation is the inverse of the frequency with which this transformation takes place along the branches of the original tree. Thus, the most common types of changes are given a small penalty, while rare transformation events are more costly. A representative substitution frequency matrix for a papillomavirus L1 tree and the resultant weighting matrix created using the software packages PAUP [25] and MacClade [17] are shown in table 1.

The most frequent types of change (in the table 1 matrix, C → T) will be responsible for the greatest amount of *homoplasy* in the sequence data – characters that are shared by two sequences but have separate origins in the true tree and do not therefore indicate common ancestry. Because weighted parsimony makes these the least significant to the determination of the overall branch length of the tree, it reduces the effect of homoplasies in the data upon the structure of the tree. (The large genetic distances between papillomavirus sequences make it highly probable that a large amount of homoplasy is present in the data, as we shall discuss below). Phylogenetic trees created using weighted parsimony over L1 and E6 are shown in figures 2 and 3.

The most obvious application of the information supplied by these trees is toward the development of a classification scheme to augment that already given by the present type designation scheme, which is based solely on simple distance measurements derived from nucleotide differences, so-called Hamming distances. To the extent that simple distances embody mutational noise (such as homoplasies), they lose value as a basis for database classification. As will be made clear below, HPV sequences are in a state of mutational saturation, and therefore the noise level is quite high in sequence comparisons. Weighted parsimony is perhaps the best method at this time for overcoming evolutionary noise.

A desirable characteristic from the perspective of classification is that trees constructed on the basis of different portions of the genome should display a similar structure; otherwise, the specific classification of a papillomavirus sequence might depend to a large degree on the precise region under examination. The topologies of trees created for papillomavirus sequences based on different portions of the genome are, in fact, reasonably consistent. One inference suggested by these results is that recombination between types has not played a significant role in the evolution of papillomaviruses. Another is that genetic distances between types are consistent across the genome, which would imply that it is largely irrelevant whether the definition of the types is based upon the E6, E7 and L1 ORFs or upon some other regions, although the exact level of divergence might have to be adjusted depending on which regions were chosen. We will examine this possibility at greater length in the section on linear correlations.

The consistency of the clades obtained from the phylogenetic analysis of different segments of the genome seems to justify the use of phylogenetic trees for classification of the papillomaviruses, if only for heuristic purposes. The HPV database has therefore defined a set of nine groups (A–I), in order to provide a general framework for the presentation of sequence information and to guide certain aspects of subsequent analyses. Groups A–F were defined on the basis of a weighted-parsimony analy-

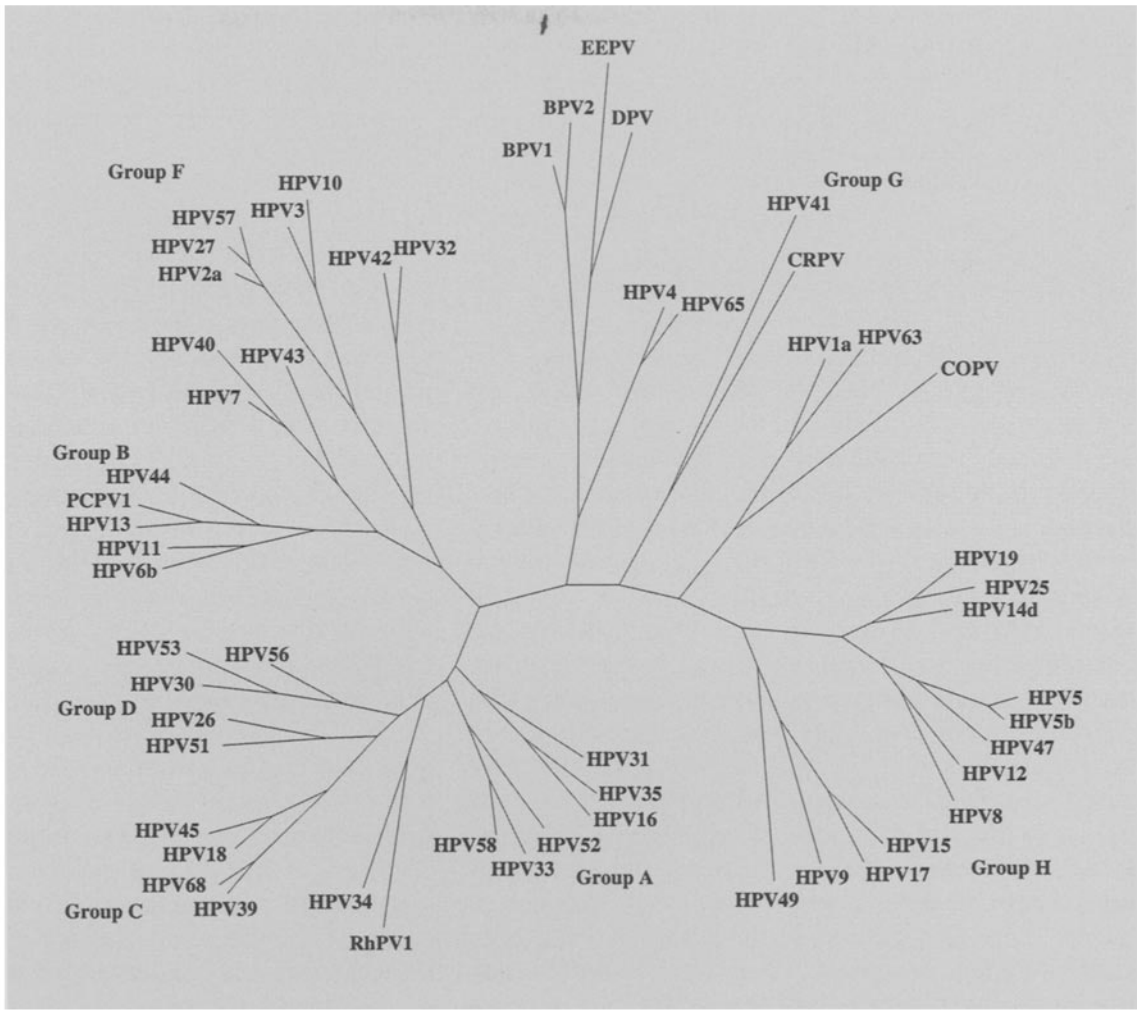

Fig. 3. A phylogenetic tree created using weighted parsimony on all papillomavirus types for which complete E6 sequences were available. The data set included 256 variable sites. The representation of this tree differs from the other figures only insofar as it has been given no hypothetical 'root', thus making no assumption concerning the direction of evolution along the branches. Note that group I is not identified on the tree, as its members (all the animal papillomaviruses) are dispersed in various locations throughout the tree – group I was not defined according to cladistic criteria. COPV and CRPV, for example, are members of group I, not group G.

sis of 54 sequences over a 291-bp fragment of L1 (fig. 2). Of these, group F represents a miscellaneous category comprising sequences not grouping with any of the larger clades in the tree. Since most of the members of these groups have a specific tropism for mucosal tissues, groups A–F will be referred to as the 'mucosal' groups, although some of their members have been found in both mucosal and cutaneous tissues, and a very small number have been found to date only in cutaneous tissues. Groups G and H were defined partly on the basis of their clustering in other trees and partly on the basis of their clinical associations: members of group G have a specific tropism toward cutaneous tissues and are generally associated with benign warts; members of group H are also specific to cutaneous tissues and are associated with the rare skin disease epidermodysplasia verruciformis (EV). Group I was defined solely on the basis of the fact that its members were isolated from nonhuman hosts – some of these in fact cluster extremely close to HPVs in phylogenetic analysis, e.g., the pygmy chimpanzee papillomavirus, PCPV-1. Arguments could be made against the definition of these latter groups on the basis of phenetic considerations, i.e., characteris-

tics of the viruses that are expressed as similarities, irrespective of their genetic makeup. However, because of their great diversity, if one were to treat these sequences in the same manner as the members of the first six groups, the number of groups would grow to an unmanageable size. Since current research tends to focus upon the mucosal types, because of the strong associations of some of these with extremely prevalent forms of cancer, it seemed natural to give more attention to these for purposes of presentation.

Ideally, a classification system for the papillomaviruses would not only reflect their evolutionary history, but would also provide insight into their clinical associations. Now, although we have used phenetic considerations to define certain of the groups, it should be made clear that the phenetic analysis of sequences is theoretically a mode of inquiry distinct from the cladistic approach taken in phylogenetic analysis. Even so, it is often (though by no means always) the case that the genetic characteristics causing certain sequences to be grouped together under the phylogenetic approach will also correspond to shared characteristics at the protein level that are manifested in phenotypically similar ways. Thus, phenotypic clustering analysis based on protein similarity scores will generally produce a branching structure quite similar to that generated by phylogenetic analysis. An important exception to this general rule occurs when phenetic analysis is carried out on short subsequences of the entire coding region, such as epitopic regions, in which evolutionary parallelism or convergence may play an important role in determining the clustering patterns. We will discuss the significance of this form of analysis further in the section on future directions of analysis planned by the database.

In the case of the papillomaviruses, some authors have attempted to demonstrate that the structure of phylogenetic trees created for papillomavirus sequences gives a fairly accurate picture of their phenotypic associations [26]. In their trees, sequences whose primary tropism is cutaneous tissue appear in different branches of the tree than those typically found in mucosal tissues; sequences associated with a high risk for malignant progression cluster apart from those associated with a low risk. Our trees (fig. 1, 2) bear these results out to a certain extent, although the lines are not so clearly drawn. As pointed out in the discussion of the groups, several sequences with ambiguous tissue specificities (e.g., HPV-2a) are distributed among primarily 'mucosal' sequences. Further, the two distinct groups of predominantly 'high-risk' types (groups A and C) do not always associate more closely together than with the 'low-risk' group B. On the whole, it is probably most accurate to say that clinical findings (e.g., tissue tropisms, association with malignant growth, and prevalence data) are not yet sufficiently well defined for most known HPV types to make any definite assertions concerning the correlation of cladistics and phenetics.

Linear Correlations

Linear correlation analysis provides an alternative to the tree construction method for examining the question of how the information contained in papillomavirus genetic sequences varies from region to region. In a sense, this mode of analysis is more in conformity with the accepted type designation scheme, since, like the definition criterion, it uses only the observed genetic distances between sequences.

In order to perform this type of analysis, alignments must be created for each of the regions of the genome under consideration. We have concentrated on the L1, E6 and E7 ORFs, as these are the regions used for designating papillomavirus types. Again, as in phylogenetic analysis, all positions containing gap characters are removed from the data. Pairwise comparison of the sequences gives a set of simple Hamming distances for each sequence pair, where the distance is equal to the percentage of nonidentical positions between the two sequences. The values for distances between sequence pairs over different regions is displayed graphically in figure 4a–c. Every different point on the graphs represents the results obtained for a particular sequence pair; the distance of this point along each coordinate axis represents the distance between the two sequences for the gene indicated along the axis. Thus, in figure 4a, the most distant sequences have diverged about 50% in L1, and by about 65% in E6.

One has the option of leaving these distance values in this uncorrected form, or using some sort of correction factor to take into account the effect of multiple mutations at the same nucleotide on the observed distances, e.g., the Jukes-Cantor equation [11], or the Kimura two-parameter model [12]. The effect of these corrections will always be to increase the value of the observed distances, by an amount that depends upon the size of the observed distance. Figure 4d shows distance values for the L1 and E6 ORFs that have been corrected for multiple hits using the Jukes-Cantor equation:

$$\text{distance}_{\text{corrected}} = 0.75 \times \ln (1 - 4/3 \times \text{distance}_{\text{uncorrected}}).$$

The theory behind this correction formula makes certain assumptions (equality of substitution rates at all sites, equal probabilities for all types of substitutions) which are not valid for most sequences. It is unclear to what extent

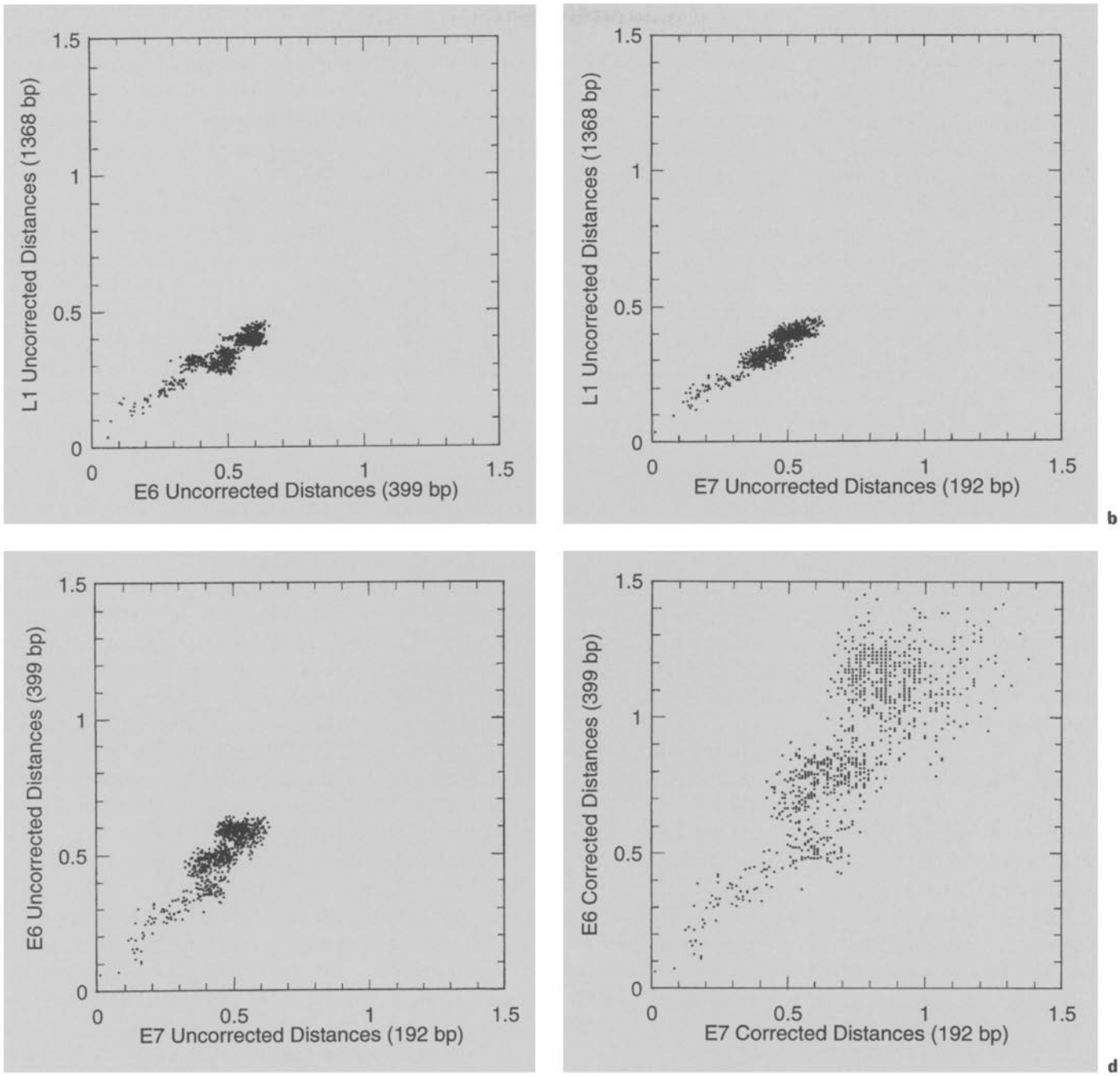

Fig. 4. Linear correlation analyses for genetic distances over various ORFs. **a** E6 and E7 ORFs, uncorrected distances: r = 0.85. **b** E7 and L1 ORFs, uncorrected distances: r = 0.91. **c** E6 and L1 ORFs, uncorrected distances: r = 0.90. **d** E6 and L1 ORFs, distances corrected using the Jukes-Cantor equation.

divergence from these assumptions affects the accuracy of its predictions – its main advantage is its extreme simplicity.

These graphs reveal a significant degree of homogeneity in the distance relationships between the papillomaviruses across these three regions of the genome. Similar results for partial L1 sequences of less than 300 bp in length have also been reported by another group [2]. The linearity of these graphs lends strength to the arguments based upon phylogenetic analysis against recombination

between the types, for any significant recombination events would appear as points significantly outside of the main 'swarm' of points. (A statistical measure of linearity, Pearson's r, is given in the figure legend. Values may range from −1 to +1: those near the extremes indicate strong negative and positive correlations, respectively, while values near zero indicate no correlation). Further, the slope of the line most nearly fitting the data set in each graph provides an indication of the relative selective pressures between the various coding regions. The most direct and well-supported inference that may be drawn from these results, however, is that the distance relationships between pairs of sequences over any of these regions are consistent with those that would be found upon examination of the other regions. Thus, the probability that a sequence will be considered a distinct type under the strict criterion of divergence from all known sequences over its E6, E7 and L1 ORFs may be accurately inferred on the basis of a subset of this sequence information. This is of particular utility to the discovery of candidates for designation as novel types, since it means that PCR fragments, which may be obtained and sequenced with comparative ease, can be used to provide a synopsis of the story that will be told by the complete genome. Some authors have already proposed using the region of L1 flanked by the My09/My11 primer pairs as a guide to the discovery of potentially novel types [2].

Synonymous versus Nonsynonymous Substitution Frequencies

Up to this point, we have discussed forms of analysis that are based solely upon nucleotide differences, without considering how these may relate to differences at the protein level. However, because of the degeneracy of the genetic code, an important distinction can be made between nucleotide changes that correspond to differences at the amino acid level and those that do not. The terms nonsynonymous and synonymous are used to distinguish these two classes of substitution events. Approximately 70 to 80% of the sites in a coding region will be nonsynonymous targets, depending on the base composition of the sequence. Nevertheless, the majority of substitutions observed in most sequences are synonymous, due to negative selection pressures. Because synonymous mutations are under minimal selective pressure, these typically accumulate at an approximately linear rate. They can therefore be used as a 'molecular clock' to gauge the temporal divergence of sequences, up to the point when these begin to approach the theoretical level of mutational saturation. Further, if the synonymous substitution rate is assumed to represent the frequency of mutations subject to no selective forces, the ratio between this and the nonsynonymous substitution rate will give a reasonable measure of the positive or negative selective pressures upon the proteins encoded by the different genes. Nonsynonymous substitutions will typically reach a limiting saturation value at a level much lower than the theoretical limit, because of negative Darwinian selection.

In order to perform this sort of analysis, nucleotide alignments must be made on a codon by codon basis; generally, the only regions in which standard nucleotide alignments will depart from this rule are those in which codons have been broken because of frameshift mutations in some sequences. Pairwise comparisons are then made between all the sequences according to one of the various methods that have been proposed for determining synonymous and nonsynonymous rates of substitution. One of the simplest of these is the Nei-Gojobori algorithm, which addresses the problem of comparing codons differing by more than one nucleotide by taking an unweighted average of the numbers of the two kinds of substitution involved in each of the possible evolutionary pathways between the two codons [20]. The calculations result in two ratios: ps is the ratio of the number of synonymous mutations observed between two sequences to the number of possible synonymous substitutions between the two; pn is the corresponding ratio for nonsynonymous substitutions. Again, these values may be corrected by means of the Jukes-Cantor equation to account for multiple mutations of the same nucleotide; the resulting values are termed ds and dn.

Figure 5 shows ps and pn values for all pairwise sequence comparisons across the L1, E6 and E7 genes. Each point on the graphs represents a particular sequence pair. The relatively small initial slope of the few sequence comparisons with ps values below 0.6 indicates that the sequences are subject to fairly stringent negative selection constraints. A corrected version of the L1 graph, using ds and dn values, is shown in figure 5d. Since most pn values are below 0.4, the effect of the correction factor is relatively insignificant in this dimension. On the other hand, most sequences are near the theoretical limit of 0.75 for synonymous substitutions and, accordingly, the correction greatly magnifies these distances. The fact that most HPV sequence relationships display mutational saturation suggests the need for more refined distance measurements than those provided by simple comparisons or unweighted parsimony analysis. Apparently, HPV sequence relationships go far back in time.

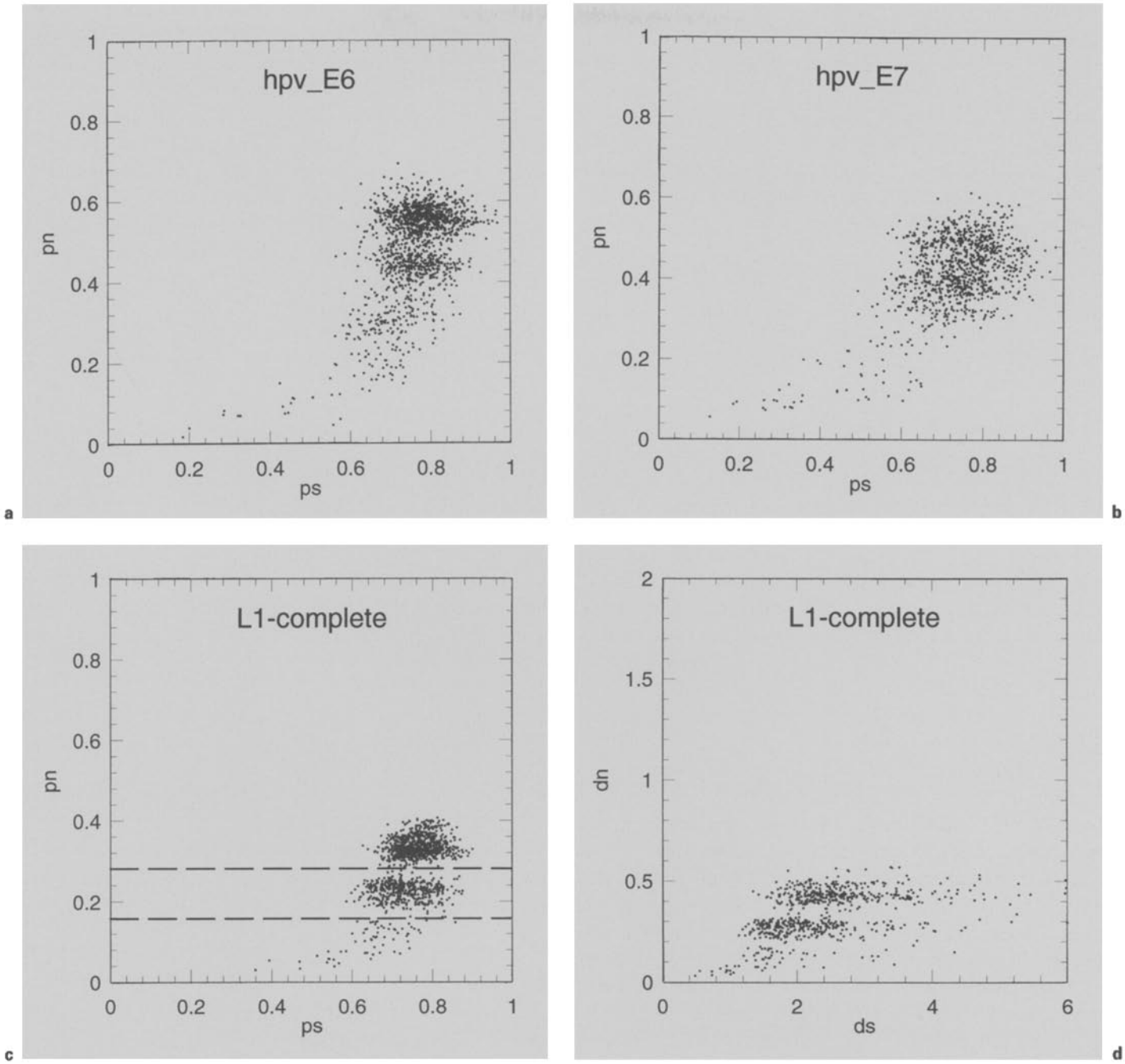

Fig. 5. Synonymous vs. nonsynonymous substitutions for E6, E7 and L1 ORFs. Substitution frequency values were obtained using the Nei-Gojobori algorithm. **d** represents the Jukes-Cantor-corrected version of **c**, showing ds and dn values rather than ps and pn.

The most striking feature shared by all the graphs is the vertical stratification of the data set into three distinctive clusters with virtually no overlap. This is most apparent in the L1 graph (fig. 4c), where the three strata are unambiguously defined by the limits: (1) $0 < pn < 0.16$; (2) $0.16 < pn < 0.28$, and (3) $0.28 < pn < 0.4$, as indicated by the dashed horizontal lines. Further analysis reveals that these strata correspond to: (1) comparisons within each of the groups (A–H); (2) comparisons between the groups, but within 'phenotypic categories', and (3) comparisons

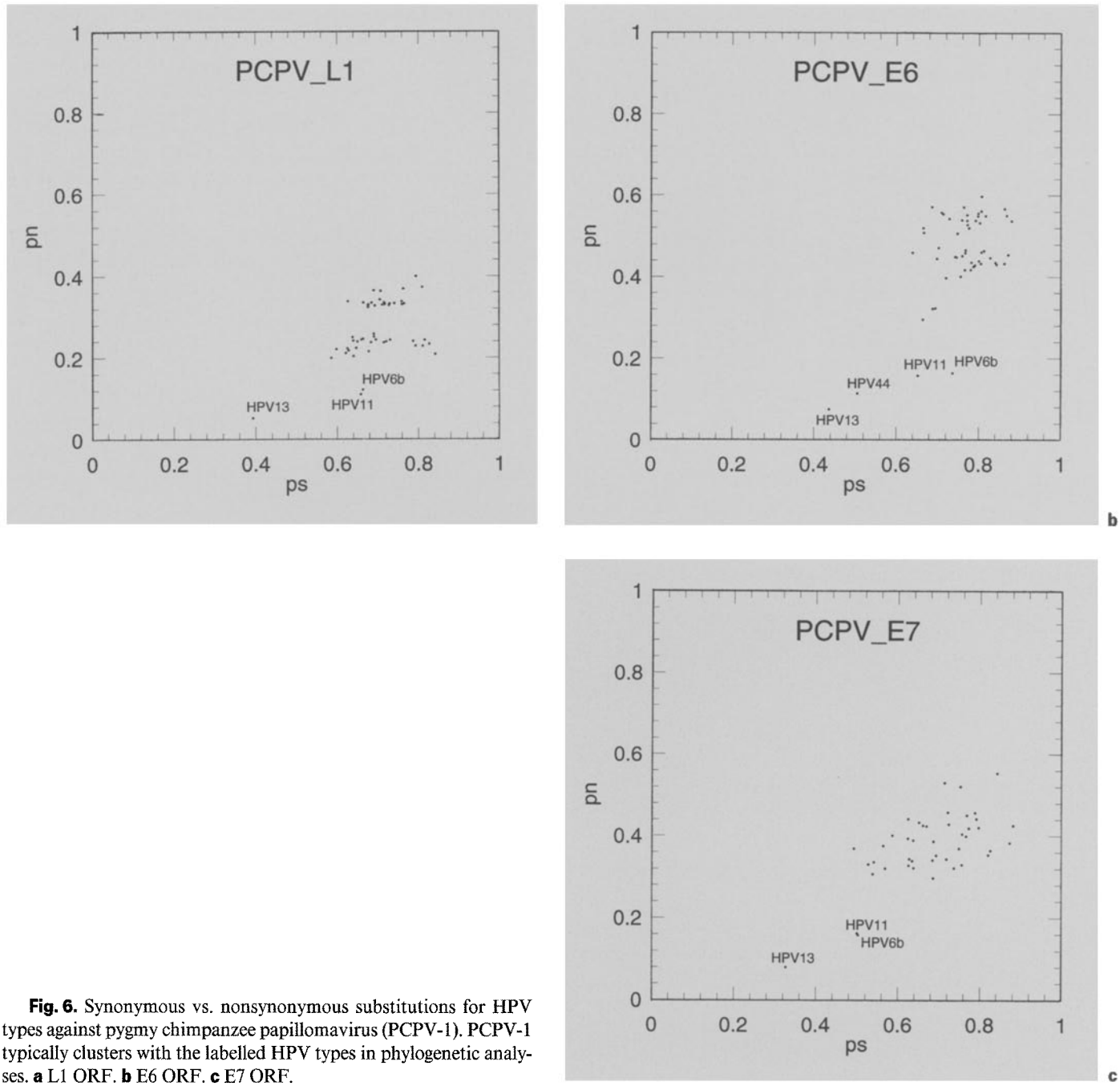

Fig. 6. Synonymous vs. nonsynonymous substitutions for HPV types against pygmy chimpanzee papillomavirus (PCPV-1). PCPV-1 typically clusters with the labelled HPV types in phylogenetic analyses. **a** L1 ORF. **b** E6 ORF. **c** E7 ORF.

between 'phenotypic categories', where 'phenotypic categories' refers to the division between the mucosal types (groups A–F), the non-EV-associated cutaneous types (group G) and the EV-associated cutaneous types (group H).

Since intragroup comparisons fall exclusively within the lowest stratum of the data set defined by nonsynonymous substitution rates, it is probable that these phylogenetically defined clusters are also relevant in terms of significant protein level similarities and differences. Further examination of the distribution of points in this stratum ($0 < pn < 0.16$ in figure 5c) reveals that a small number of sequence pair comparisons yield pn values significantly smaller ($pn < 0.1$) than most of the values obtained for PV types within the same group. The implication is that although these sequences do qualify to be recognized as

distinct types when considering only undifferentiated nucleotide changes (i.e., nucleotide differences that have not been further categorized into synonymous and nonsynonymous changes), the differences between the proteins encoded by these sequences are probably slight. We have designated such sequence pairs as 'close types' to indicate that this lack of nonsynonymous divergence should at least be taken into account when considering them as distinct types according to the standard criterion. Many of these 'close types' also display significantly low values for their synonymous substitution rates, indicating that these sequences have diverged from each other relatively recently. Sequence sets which may be regarded as 'close types' are especially interesting from the perspective of the processes underlying papillomavirus evolution. In addition to the category of intratypic 'variants' and that of 'subtypes', the category of 'close types' seems to represent another important level in the spectrum of papillomavirus diversity, which at first glance seems to be almost entirely discontinuous.

Interestingly, comparisons between the pygmy chimpanzee papillomavirus sequence PCPV-1 and the members of group B (most notably, HPV-13) have values for both types of substitutions that are among the lowest for any sequence comparisons (fig. 6). To date, no direct experimental evidence has been found for instances of cross-species papillomavirus transmission. Some authors have argued on the basis of this lack of evidence and from a putative correspondence between the phylogeny of papillomavirus sequences and that of their hosts, that papillomaviruses have strictly coevolved with their hosts. Under this hypothesis, minimal estimates for papillomaviral mutation rates have been calculated using the divergence of PCPV-1 and HPV-13 and the approximate time of divergence between their two host species [1]. Considering the extremely close relationship between these two sequences, however, especially from the perspective of the limited amount of divergence at the protein level, cross-species transmission seems a highly probable conjecture. At the very least, the possibility should not be ruled out without a much more extensive examination than has hitherto been conducted.

Future Directions

In future installments of the HPV database, in addition to further exploration of the aforementioned analyses, we plan to devote a great deal more attention to two fields of analysis which have been touched on only briefly here, viz., intratypic variation analysis and protein level analyses.

Several forms of protein level analysis are planned for future releases of the database. One such analysis would be to determine *signature patterns* of amino acids that are common to sequences sharing a certain property (e.g., high risk for malignant conversion) but rare among sequences not possessing that property [15]. This may be done by defining signatures by either a simple majority rule or by imposing some threshold value of conservation for amino acids in the pattern. Alternatively, one could examine patterns of amino acid positions that display high levels of variability in one sequence set relative to their variability in another sequence set, according to some formal measure of variability, such as the information theoretic quantity known as entropy [13]. Another possible approach to the analysis of protein level relationships is given by the tree-like structure known as a *phenogram.* Phenograms can be constructed using a variety of amino acid substitution matrices [14]; the branching pattern is determined by pairwise protein similarity scores, and the position along the x-axis of each node in the tree represents the score for the pattern generated according to the amino acid class hierarchy for all sequences below the node on the tree. Unlike phylogenetic trees, therefore, the phenogram is not directly concerned with evolutionary relationships based on assumed homologies, but rather with clusters based on similarities at the protein level that probably bear some relation to function. This sort of analysis is generally most interesting when performed on short subsequences of genes such as epitopic regions, for these will often display clustering patterns significantly unlike those observed in phylogenetic analysis. The discrepancies observed between the two clustering patterns may help to explain certain anomalies hitherto observed among the papillomaviruses, such as the presence of certain cutaneous-tissue-specific types in the phylogenetic cluster of the primarily mucosal types of groups A–F.

Acknowledgments

The HPV Sequence Database and Analysis Project is funded by the Division of Microbiology and Infectious Diseases of the National Institute of Allergy and Infectious Diseases, USA, through an interagency agreement with the US Department of Energy (Dr. Penelope Hitchcock, Project Officer).

References

1 Bernard HU, Chan SY, Delius H. Evolution of papillomaviruses. In: zur Hausen H, ed. Human Pathogenic Papillomaviruses. Berlin, Springer, 34–54; 1994.

2 Bernard HU, Chan SY, Ong CK, Villa LL, Delius H, Peyton CL, Bauer HM, Manos MM, Wheeler CM. Identification and assessment of known and novel human papillomaviruses by polymerase chain reaction, restriction digestion fingerprinting, nucleotide sequence, and phylogenetic algorithms. J Infect Dis 170: 1077–1085;1994.

3 Chan SY, Delius H, Bernard HU. Phylogenetic and taxonomic evaluation of 96 papillomavirus types. (submitted).

4 Chan SY, Ho L, Ong CK, Chow V, Drescher B, Durst M, ter Meulen J, Villa L, Luande J, Mgaya HN, Bernard HU. Molecular variants of human papillomavirus type 16 from four continents suggest ancient pandemic spread of the virus and its coevolution with humankind. J Virol 66:2057–2066;1992.

5 Chong T, Chan WK, Bernard HU. Transcriptional activation of human papillomavirus 16 by nuclear factor I, AP1, steroid receptors and a possibly novel transcription factor, PVF: A model for the composition of genital papillomavirus enhancers. Nucleic Acids Res 18:465–470;1990.

6 Coggins JR, zur Hausen H. Workshop on papillomaviruses and cancer. Cancer Res 39:545–546;1979.

7 Deau MC, Favre M, Orth G. Genetic heterogeneity among human papillomaviruses (HPV) associated with epidermodysplasia verruciformis: Evidence for multiple allelic forms of HPV5 and HPV8 E6 genes. Virology 184:492–503;1991.

8 de Villiers EM. Human pathogenic papillomavirus types: An update. In: zu Hausen H, ed. Human Pathogenic Papillomaviruses. Berlin, Springer, 1–12;1994.

9 Hillis DM, Allard MW, Miyamoto MM. Analysis of DNA sequence data: Phylogenetic inference. Methods Enzymol 224:456–487;1993.

10 Hillis DM, Huelsenbeck JP, Cunningham CW. Application and accuracy of molecular phylogenies. Science 264:671–677;1994.

11 Jukes TH, Cantor CR. Evolution of protein molecules. In: Munroe H, ed. Mammalian Protein Metabolism. New York, Academic Press, 21–132;1969.

12 Kimura M. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111–120;1980.

13 Korber BTM, Farber RM, Wolpert DH, Lapedes AS. Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: An information theoretic analysis. Proc Natl Acad Sci USA 90:9061–9065;1993.

14 Korber BTM, MacInnes K, Smith RF, Myers G. Mutational trends in V3 loop protein sequences observed in different genetic lineages of human immunodeficiency virus type 1. J Virol 68:6730–6744;1994.

15 Korber B, Myers G. Signature pattern analysis: A method for assessing viral sequence relatedness. AIDS Res Hum Retroviruses 8:1549–1560;1992.

16 Laga M, Diallo MO, Buve A. Inter-relationship of sexually transmitted diseases and HIV: Where are we now? AIDS 8 (suppl 1):S119–S124;1994.

17 Maddison WP, Maddison DR. MacClade: Analysis of Phylogeny and Character Evolution. Sunderland, Sinauer, 1992.

18 Myers G. Assimilating HIV sequences. AIDS Res Hum Retroviruses 9:697–702;1993.

19 Myers G, Bernard HU, Delius H, Favre M, Icenogel J, van Ranst M, Wheeler C. Human Papillomaviruses: A Compilation and Analysis of Nucleic Acid and Amino Acid Sequences. Los Alamos, Theoretical Biology and Biophysics, Los Alamos National Laboratory, 1994.

20 Nei M, Gojobori T. Simple methods for estimating the numbers of synonymous and nonsynonymous substitutions. Mol Biol Evol 2: 418–426;1986.

21 Ong CK, Chan SY, Campo MS, Fujinaga K, Mavromara-Nagos P, Labropoulou V, Pfister H, Tay SK, ter Meulen J, Villa LL, Bernard HU. Evolution of human papillomavirus type 18: An ancient phylogenetic root in Africa and intratype diversity reflect coevolution with human ethnic groups. J Virol 67:6424–6431; 1993.

22 Rowson KEK, Mahy BWJ. Human papova (wart) virus. Bacteriol Rev 31:110–131;1967.

23 Smith RF, Smith TF. Automatic generation of primary sequence pattern from sets of related sequences. Proc Natl Acad Sci USA 87:118–121;1990.

24 Sundberg JP. Papillomavirus infections in animals. In: Syrjaenen K, Gissmann L, Koss LG, eds. Papillomaviruses and Human Disease. Berlin, Springer, 40–103;1987.

25 Swofford DL. Paup: Phylogenetic Analysis Using Parsimony (Version 3.1). Computer program distributed by the Illinois Natural History Survey, Champaign, Illinois, 1991.

26 Van Ranst MA, Kaplan JB, Burk RD. Phylogenetic classification of papillomaviruses: Correlation with clinical manifestations. J Gen Virol 73:2653–2660;1992.

27 Van Ranst MA, Tachezy R, Burk RD. Human papillomavirus nucleotide sequences: What's in stock? Papillomavirus Rep 6:65–75;1994.

28 Zur Hausen H. Preface. In: zur Hausen H, ed. Human Pathogenic Papillomaviruses. Berlin, Springer, v–ix;1994.