III

ANALYSES

A. Phylogenetic Trees

Before embarking on a discussion of phylogenetic tree analysis, it will be appropriate to briefly discuss the distinction between *similarity* and *homology*. Fitch, Doolittle, and others have argued for more than two decades that the term *homology* should denote an inference, whereas the term *similarity* should be applied to an observation. By this diction, sequences are "70% similar" or "95% similar" . . . but *not* 70% or 95% homologous. Homologous sequences are either 100% homologous or they are nonhomologous (although subtle distinctions of *parology* are made). Homologous sequences can actually be less than 25% similar, and therefore below the level of chance similarity. Conversely, nonhomologous sequences can be, say, 65% similar; convergent evolution would be implied in such a case. The database will adhere to the distinction: specifically all PV sequences are inferred to be homologous from i) a shared genomic organization and ii) phylogenetic analysis. The extent of their similarity is another matter. The groups in Part I are inferred from phylogenetic analysis, but the definition of "close types" (see III-C) is an observation.

We should also distinguish *cladistic* analysis (phylogenetic analysis) from *phenetic* analysis. Phenetic analysis (the Greek verb *phaino* ($\phi\alpha\iota\nu\omega$) is associated with appearance, hence the English word phenomenon) emphasizes structural similarities irrespective of evolutionary relationships or pathways. Cladistic analysis (from the Greek noun *klados* ($\kappa\lambda\alpha\delta\text{o}\varsigma$), which means olive branch or young branch) is concerned with the network of evolutionary relationships. For a discussion of these concepts, see Li and Graur [1] and Myers and Korber [2]. In section III-A, the focus is on cladistic (i.e., phylogenetic) relationships. However, we anticipate that future releases of the database will be increasingly concerned with phenetic relationships, namely protein similarities irrespective of evolutionary origin.

Phylogenetic inferences are reached through many different analyses, which can be either *distance-based* or *character-based.* Many of the analyses in this edition of the HPV compendium are distance-based, for example those of III.B. Here we pursue character-based analysis.(For an excellent introduction to phylogenetic analysis of sequences, see Hillis et al. [3].) Parsimony analysis and maximum likelihood are the two most widely used character-based approaches, and parsimony is usually the more practical, or tractable. However, it is well known that parsimony analysis can lead to erroneous inferences when very different evolutionary rates are represented by a sequence data set [4]. Moreover, *homoplasy*, the chance occurrence of identical characters at homologous positions in sequences from different lineages, is always a problem with highly diverged sequences, such as the HPVs. In this section, and throughout the compendium, we have employed *weighted parsimony analysis* in order to overcome homoplasy as well as the deficiencies of ordinary parsimony in the event some lineages of PVs have radically different evolutionary rates.

Parsimony analysis looks for the minimum global evolutionary path: the tree or set of trees with the fewest overall changes (lowest sum of branch lengths) becomes the basis for a phylogenetic inference. The assumption underlying this analysis, to put it simply, is that nature takes the least path. Ordinary parsimony presupposes equal substitution frequencies, and therefore gives the same weight to every base change. In fact, the most common changes (e.g., A $\rightarrow$ G), will dominate in the analysis, and they will contribute most to homoplasy. For viruses with skewed base compositions, such as HIV and HPV, weighted parsimony will improve the analysis [4]. The first step in this procedure is to run ordinary parsimony analysis in order to determine the substitution biases; all things being equal, the biases will be in accord with the base composition. We use Macintosh versions of PAUP[5] and MacClade[6] to accomplish this step. A resulting substitution matrix is shown below.

	to: **A**	**T**	**G**	**C**
from: **A**		0.089	0.141	0.105
T	0.062		0.044	0.076
G	0.110	0.029		0.036
C	0.110	0.155	0.042	

The next step in the analysis is to re-execute PAUP using an inverse weighting table generated from the substitution matrix. Thus the least common changes are given the greatest weight and the most common changes are given the least weight; the virtue of this strategy for reduction of homoplasy should be obvious. In applying the inverse weighting rule, it may be necessary to "truncate" terms in order to satisfy "triangle inequality" (Euclidean distances, so-to-speak; for example, the cost of a direct G to T transformation in the following matrix is higher than that of the indirect path from G to A, then from A to T)[5].

	to:	**A**	**T**	**G**	**C**
from:	**A**		11	7	10
	T	16		23	13
	G	9	34		28
	C	9	6	24	

The phylogenetic tree on the cover of this compendium and at the head of sections in Part I was generated from partial L1 sequences (the MY09-MY11 region of Manos and colleagues) using the "stepwise weighting" procedure. The same L1 tree, now with complex branch lengths, is shown in Figure III.1. Because the branch lengths are made up of different weighted terms, they are not linearly proportional to the total number of single base changes, as they would be in ordinary parsimony. However, these lengths do provide accurate relative distances. For comparison, an E6 coding sequence tree generated by weighted parsimony is shown in Figure III.2b; its identical, star-like counterpart without actualized branch lengths is shown in III.2a. Figure III.3 was generated by weighted parsimony analysis of complete L1 coding sequences from a wider array of PVs.

As an alternative to weighted parsimony, ordinary parsimony based on second codon positions only will reduce evolutionary noise (homoplasy). The tree shown in Figure III.4 is based on second base positions of the L1 data set analyzed in Figure III.3. In III.4, the branch lengths are proportional to single base changes at the codon positions being analyzed.

A maximum likelihood analysis of the sequences analyzed in Figure III.1 is reported by Bernard et al. [6]. Subtle differences between the various methods are encountered, affecting inferences about the grouping of some types, but the overall topologies are the same.

---

[1] Li W.-H. and Graur D. *Fundamentals of Molecular Evolution* Sinauer Associates, Sunderland MA, 1991.

[2] Myers G and Korber B: The Future of Human Immunodeficiency Virus. In: *Evolutionary Biology of Viruses* S.S. Morse (Ed.), Raven Press, New York, 1994; pp. 211-232.

[3] Hillis DM, Allard MW, and Miyamoto MM; Analysis of DNA Sequence Data: Phylogenetic Inference. Methods in Enzymology 1993;224:456-487.

[4] Swofford DL. *PAUP: Phylogenetic Analysis Using Parsimony* (Version 3.1) Computer Program distributed by the Illinois Natural History Survey, Champaign, Illinois, 1991.

[5] Maddison WP and Maddison DR.*MacClade: Analysis of Phylogeny and Character Evolution* Sinauer Associates, Sunderland MA, 1992.

[6] Bernard H.-U., Chan S.-Y., Ong C.-K., Villa L.L., Delius H., Peyton C.L., Bauer H.M., Manos M.M., and Wheeler C.M.: Identification and assessment of known and novel human papilloma-viruses by polymerase chain reaction, restriction digestion fingerprinting, nucleotide sequence, and phylogenetic algorithms. *J. Infect. Dis.* 1994 (Nov issue).

Phylogenetic Trees

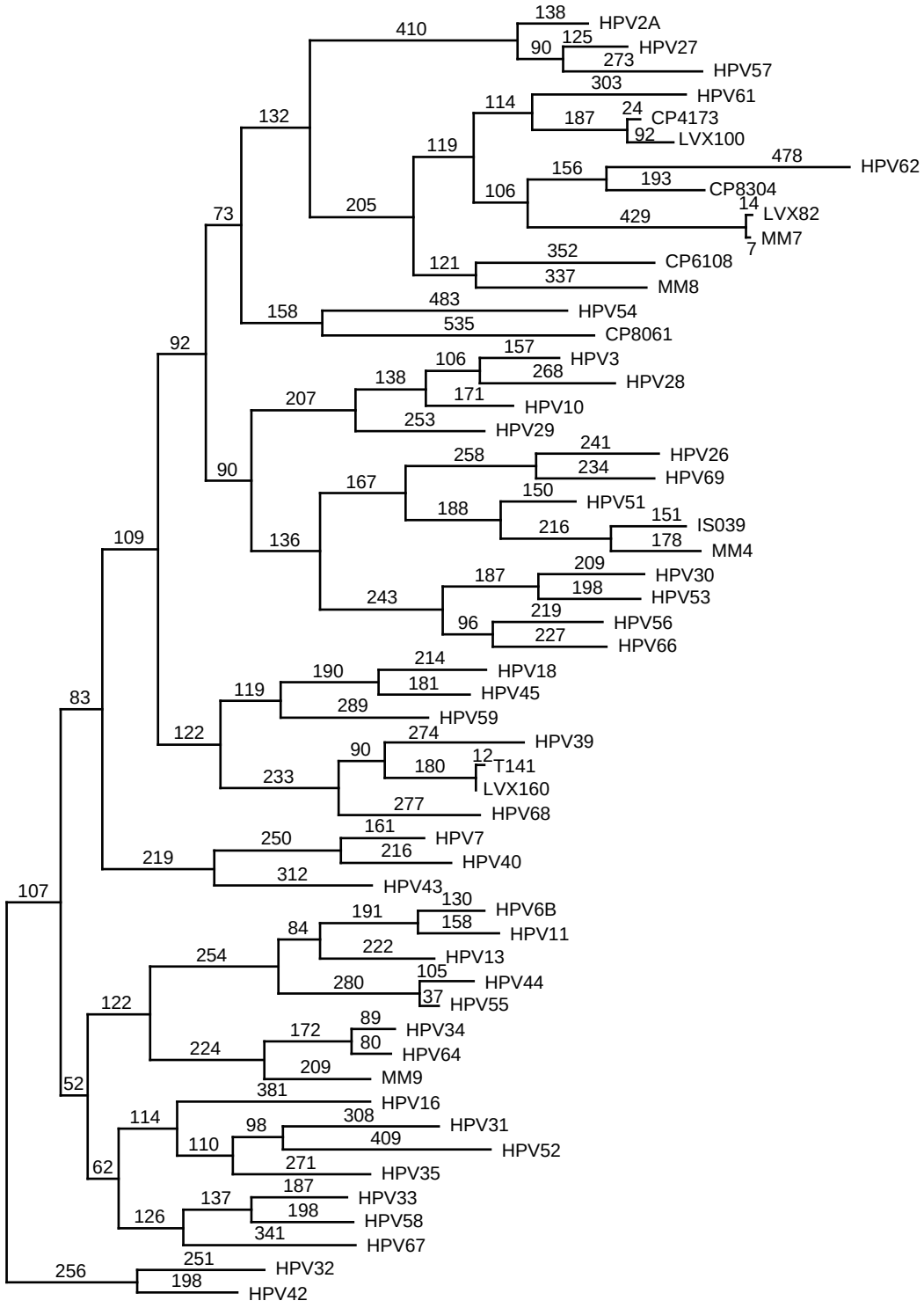

Fig. III.1 My09-My11 region weighted parsimony; 213 variable sites.

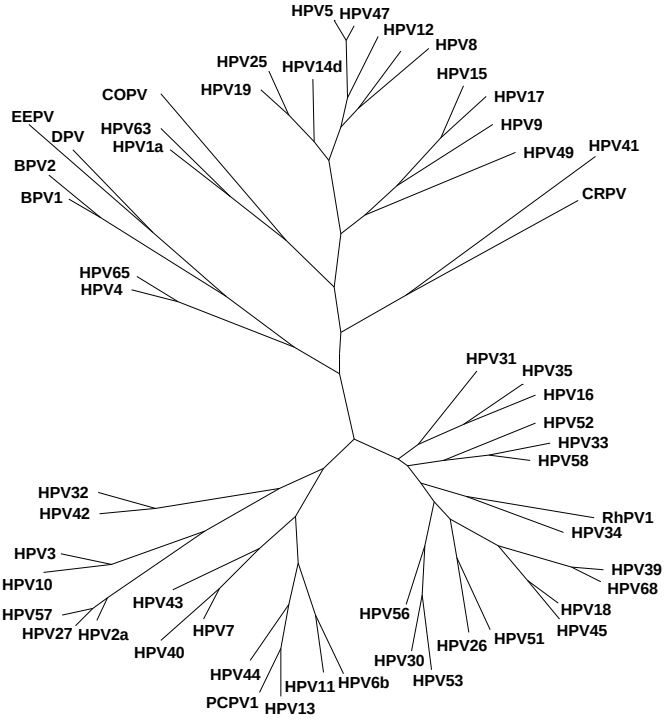

Figure III.2a E6 weighted parsimony tree; star configuration.

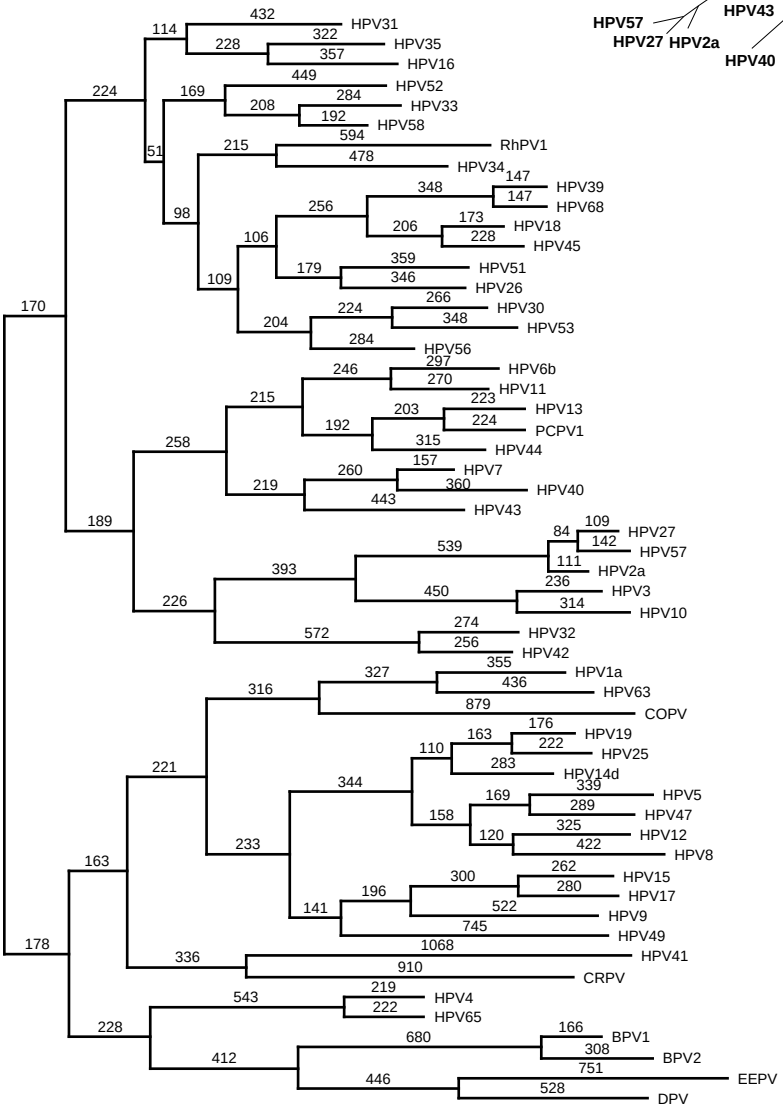

Figure III.2b E6 weighted parsimony with actualized branch lengths.

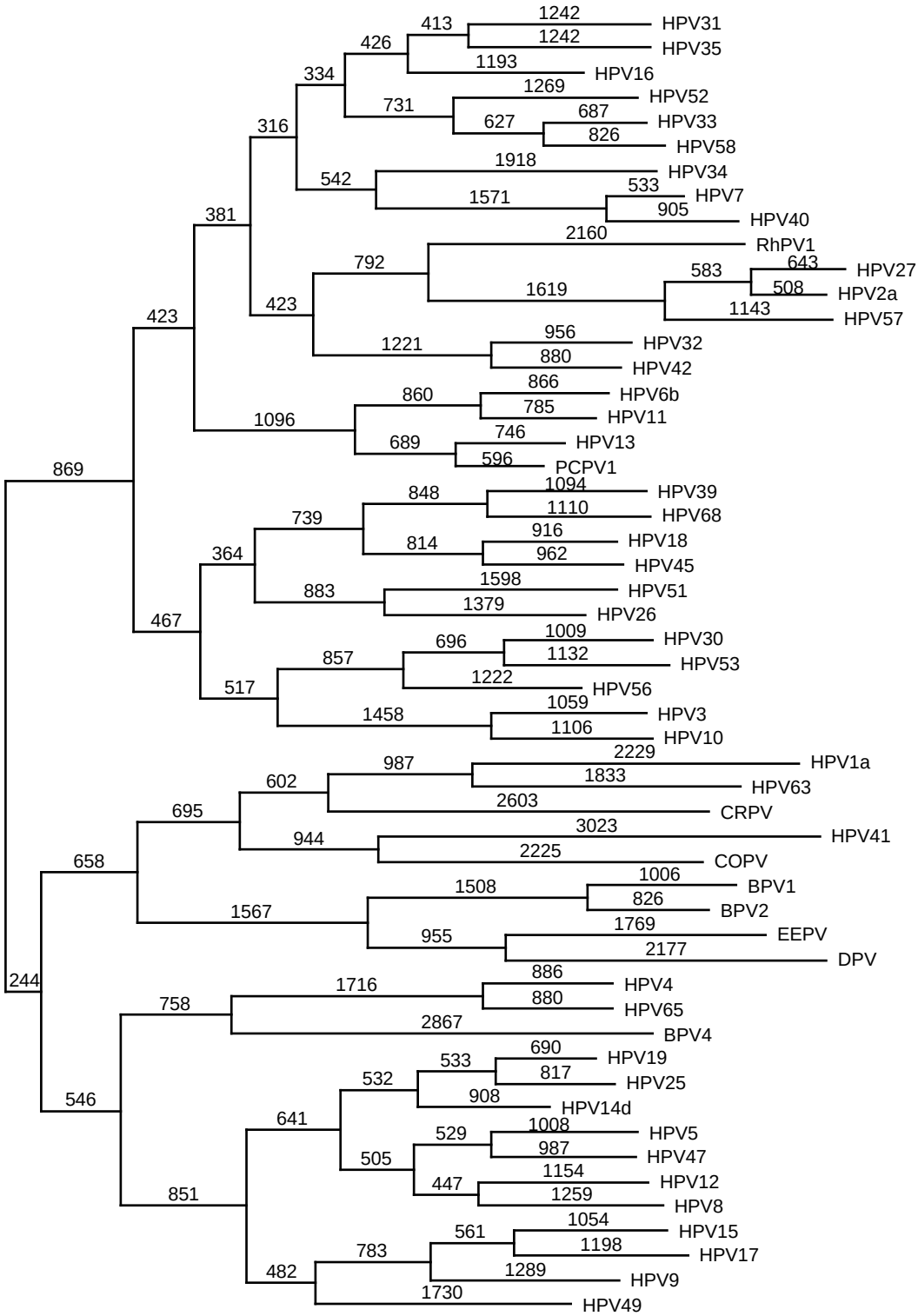

Fig. III.3 Complete L1 region weighted parsimony; 1058 variable sites.

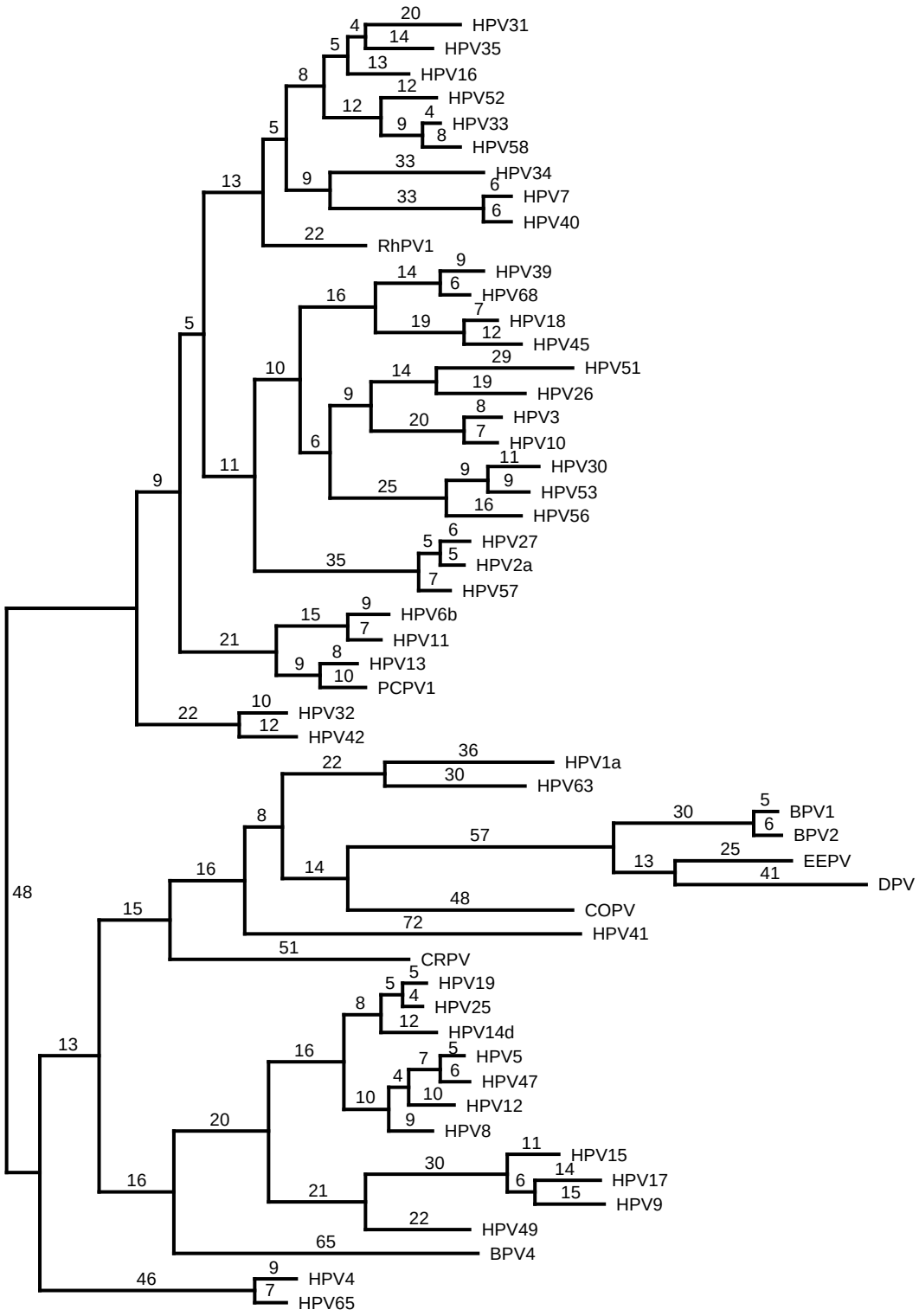

Fig. III.4 Complete L1 region unweighted parsimony using second base positions only; 276 variable sites. One of three topologically similar most parsimonious trees.

B. Linear Correlations

Figure III.5 illustrates the homogeneity of the relationships between the papillomaviruses across different regions of the genome. Each point on the graphs represents a comparison between two sequences; the set of all points in each graph represents the set of all pairwise comparisons between the sequences. The x-coordinate of the point gives the distance between those sequences for the gene indicated on the x-axis, while the y-coordinate gives the distance between them for the gene indicated on the y-axis. The uncorrected distances are simply similarity values between sequences calculated from global alignments after gapped regions had been removed. Figure III.5d shows the the data set of Figure III.5c after the Jukes-Cantor correction was applied to the original distances, in order to try to take into account the effect of multiple mutations of the same nucleotide on the observed distances (See also III-C).

The linearity of the correlation between the two measures of distance can be used as an argument against any significant recombination events. Further, the slope of the line approximated by the data points demonstrates relative selective pressures on the two different proteins. A standard statistical measure of linearity is "Pearson's *r*", which can range from –1 to +1, where the extreme values indicate perfect linear correlation in its negative and positive sense, while a value near zero indicates no correlation. The "Student's *t* probability" gives a measure of the significance of the correlation, where small values indicate significant correlation. A table showing these values for comparisons between E6 and E7, E6 and L1, and E7 and L1, for uncorrected distance measures is given below.

Genes Compared	Pearson's r	Student's t
E6-E7	0.854554	0
E6-L1	0.900692	0
E7-L1	0.914113	0

Similar results for partial L1 sequences are reported in Bernard, et al., *J. Infect. Dis.*, (in press, November 1994)

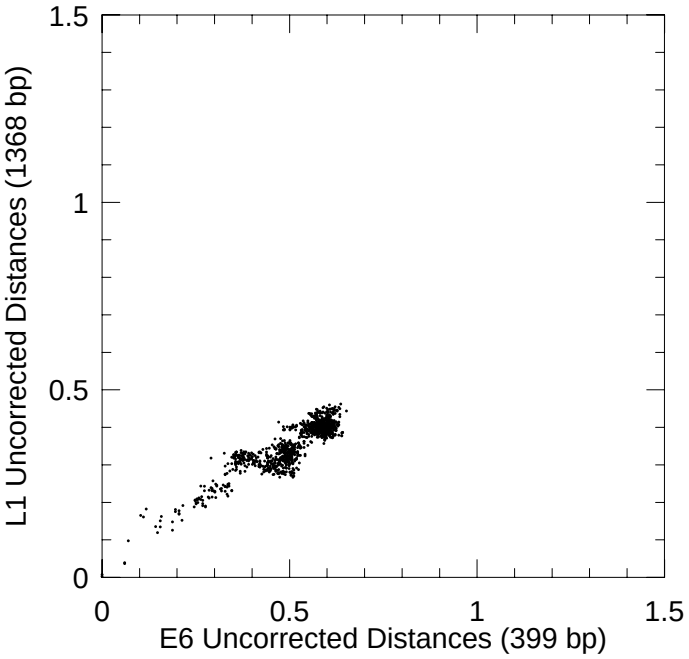

Fig. III.5a

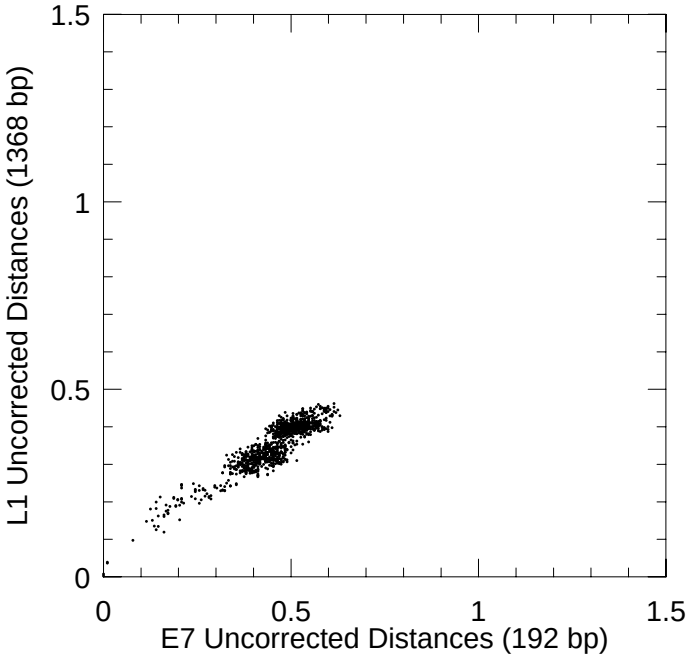

Fig. III.5b

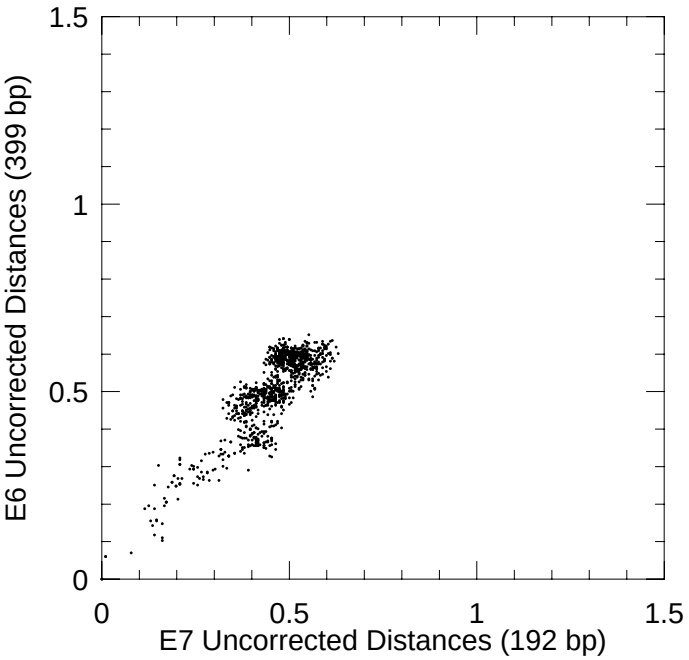

Fig. III.5c

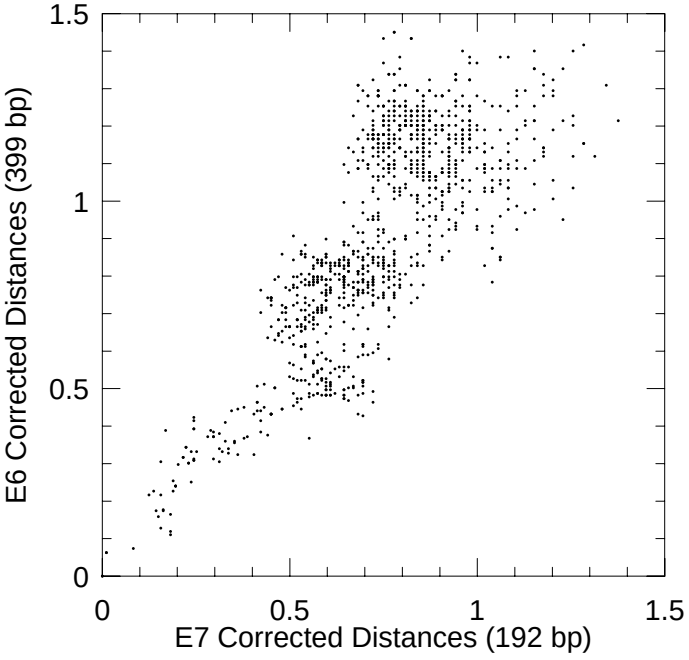

Fig. III.5d

This matrix was created in an attempt to analyze sequence relationships between the LCRs of the papillomaviruses, while at the same time circumventing the difficult problem of determining an unambiguous global alignment for this extremely divergent region. The question under examination was whether or not the relationships in the L1 phylogenetic tree (Figure III.1) which had been used to define the groups would hold for other regions of the genome as well. The strong linear correlations between L1, E6 and E7 shown in the previous section indicate that these regions are related in roughly the same way. In order to conduct these analyses, however, it was first necessary to generate an alignment of the coding sequences, to obtain an accurate measure of the simple distances plotted in Figures III.5a-d. To appreciate the difficulties that this requirement causes when working with the papillomavirus LCRs, it is first necessary to understand some of the principles of sequence alignment.

Virtually all methods of multiple sequence analysis require that the sequences first be aligned to one another, in order that the characters at the same position in different sequences be truly comparable ('positional homology'). Typically, multiple sequence alignment methods are based on algorithms that may be characterized as "globally-oriented" in two different senses, each of these senses corresponding to one of the two dimensions of a multiple sequence alignment. For the first of these, the "along the sequences" direction, gap characters are usually introduced within sequence strings in order to extend regions of similarity. For the second, the "between the sequences" direction, instead of providing a separate optimal alignment for every pair of sequences, one tries to optimize the sum of the pairwise alignment scores for just one alignment that includes all the sequences simultaneously.

In general, an alignment that is "global" in both of these senses is the ideal for multiple sequence analysis. In the first sense, having hypothesized that two sequences are related, it is reasonable to assume further that, barring recombination, they possess the same relationship along their entire lengths (e.g. if they are siblings for one gene, they will not be distant cousins for another). In the other sense of the term "global", it may be argued that under the hypothesis that all of the sequences in a set are related to one another, and given a set of alignment parameters that represent a reasonably accurate model of the evolutionary process, the optimal global alignment is more likely to represent the "true" evolutionary path between two sequences than is an optimal pairwise alignment, even though the actual score of the latter alignment is necessarily at least as good as that of the former.

In some cases, however, it is neither practical to create a globally optimal alignment, nor reasonable to assume that one's model of the evolutionary process is sufficiently accurate to make such an alignment meaningful. The LCRs of the papillomaviruses illustrate this thorny problem. Since this region is non-coding, one cannot begin by aligning the corresponding protein sequences, then use this alignment to establish the corresponding nucleotide alignment. Further, the LCR is one of the most divergent regions in the genome; with the exception of a few unusually close types, most of the sequences are significantly less than fifty percent similar. In addition to these practical difficulties to creating a global alignment for the LCRs, there are theoretical problems raised by the fact that while the LCRs seem to contain a relatively high number of complex evolutionary events such as indels, inversions, and repeats, the theory behind most alignment scoring matrices has been developed for substitution events only.

These problems may be avoided by using a measure of local similarity rather than a global one. As may be surmised from what was said concerning global alignments, local similarity may be measured without adding gaps to the sequences to extend regions of similarity and without taking into consideration all sequences at the same time. One possible measure of local similarity is used by BLAST (Basic Local Alignment Search Tool), a family of programs that are typically used to scan large databases quickly for matches to query sequences [1]. The BLAST algorithm generates a list of High-scoring Segment Pairs (HSPs) for each pair of sequences compared. These HSPs are simply regions in the two sequences which, when aligned together without introducing gaps into either of the segments, have an alignment score above a certain specified cutoff value. The ends of the HSP are defined so that the score is maximized for the region, and this region will be made as long as possible without lowering the score.

In order to generate this matrix, we first created a small database consisting only of the LCRs of all papillomavirus types. Then, using blastn (a BLAST program that matches a nucleotide query sequence against a nucleotide database), we generated the list of all Maximal-scoring Segment Pairs (MSPs) for the LCR of each papillomavirus type against that of every other type. Then, this output was put into a matrix, which was ordered "by group". Using the groups already defined by the weighted-parsimony-generated partial L1 tree (Figure III.1), the rows and columns were ordered so that HPV types in the same group appear consecutively (for discussion of the groups, see Part I). The diagonal of the matrix reports the score for the MSP of each sequence compared to itself, which is simply proportional to the length of the sequence. Since sequences in the same group are clustered together in the ordering of the rows and columns, if these groupings also apply to the LCR, we expect to find the highest scores clustered around the diagonal, and very few comparable scores elsewhere in the matrix. Scores greater than or equal to 200 are in bold type; boxes have been drawn around all the intragroup comparisons, and smaller boxes have been drawn inside these to indicate the subgroupings often used for analyses.

Many of the results found by other means of analysis (III-A, III-C, III-D) are duplicated in the matrix. Not only the strengths of the groupings are evident, but also some of their weaknesses, such as the inclusion of HPV-34 in Group B, and the tendency of HPV-51 and HPV-26 to be associated with Group C while being classified in Group D. The legitimacy of adopting subgroupings in certain cases is also quite apparent.

Figure III.6 is a linear correlation plot for the LCR against L1 that was created from an alignment of the few conserved regions in the LCRs, and only represents the members of Groups A–F. The Pearson's r-value for the data is 0.7459.

---

[1] Altschul,S.F., Gish,W., Miller,W., Myers,E.W., and Lipman, D.J., Basic Local Alignment Search Tool *J. Mol. Biol.* (1990)

	Group A								Group B				Group C					Group D				
																		Da		Db		
	31	52	35	35h	16	33	58	RhPV	6b	11	13	PCPV	34	39	68	18	45	51	26	30	53	56
HPV31	**4765**	134	195	**271**	**228**	172	141	66	141	183	146	137	173	115	178	121	121	144	124	88	132	159
HPV52	134	**4445**	169	192	**221**	**277**	**264**	96	154	145	126	153	174	138	114	**200**	177	161	122	116	112	110
HPV35	195	169	**4345**	**1577**	**202**	150	152	129	158	149	158	163	177	130	188	165	170	170	112	175	172	139
HPV35h	**271**	192	**1577**	**4395**	**300**	178	152	124	158	149	124	122	166	117	142	153	144	103	120	175	116	115
HPV16	**228**	**221**	**202**	**300**	**4160**	109	158	93	121	141	84	107	122	99	99	188	179	128	126	105	101	165
HPV33	172	**277**	150	178	**218**	**4620**	**506**	115	133	149	123	123	126	111	147	132	135	124	149	122	138	156
HPV58	141	**264**	152	152	158	**506**	**3970**	91	107	125	120	137	181	128	111	119	115	170	129	99	101	155
RhPV1	131	96	129	124	93	115	79	**3470**	75	94	67	84	74	71	81	103	78	99	106	70	73	84
HPV6b	141	154	158	158	115	133	107	84	**3560**	**1011**	**236**	**243**	108	142	121	129	133	140	172	139	117	129
HPV11	183	116	149	149	141	149	125	94	**1011**	**3780**	**235**	**234**	147	144	176	120	130	143	190	139	102	**246**
HPV13	146	126	158	124	84	123	120	67	**236**	**235**	**3710**	**698**	141	149	143	162	171	155	142	148	144	88
PCPV1	137	153	163	122	107	123	116	84	**243**	**234**	**698**	**3755**	153	187	124	158	138	126	139	107	86	103
HPV34	105	174	177	166	168	133	181	74	108	147	141	153	**4040**	171	**210**	**202**	**252**	**248**	**205**	186	181	165
HPV39	115	138	130	117	99	111	110	71	142	144	149	187	171	**3895**	**558**	**236**	**226**	195	180	106	119	122
HPV68	178	114	188	142	99	147	111	82	121	131	143	124	**210**	**558**	**4055**	**386**	**348**	**295**	172	118	113	152
HPV18	121	**200**	165	153	188	132	119	103	129	120	162	158	**202**	**236**	**386**	**4125**	**1368**	**268**	**211**	146	140	107
HPV45	121	177	170	144	179	135	113	78	133	130	171	138	**252**	**226**	**348**	**1368**	**4050**	**300**	**247**	171	136	133
HPV51	144	161	170	133	128	124	170	99	140	143	155	126	**248**	195	**295**	**268**	**300**	**4345**	**201**	163	159	151
HPV26	124	122	112	120	126	149	129	106	172	190	142	139	**205**	180	172	**211**	**247**	**201**	**4260**	134	113	112
HPV30	86	116	175	175	112	122	99	70	80	104	148	107	186	106	126	146	171	163	134	**3980**	**717**	**284**
HPV53	132	112	172	116	101	138	101	91	117	102	144	113	181	119	109	140	136	159	113	**717**	**4015**	**375**
HPV56	159	132	139	112	165	117	155	78	129	**246**	95	103	156	101	152	107	133	151	112	**284**	**375**	**4245**
HPV27	83	**206**	189	193	161	163	179	87	184	175	122	145	185	153	127	133	133	122	121	72	113	77
HPV57	101	192	172	172	138	162	178	64	161	152	158	172	119	150	119	143	140	159	114	71	81	104
HPV2a	136	188	185	198	87	88	161	77	193	184	131	145	184	162	150	133	133	84	121	91	83	109
HPV3	157	178	155	155	91	180	185	92	135	161	164	150	118	113	94	118	100	117	99	91	100	110
HPV10	118	109	69	77	115	92	84	128	120	134	101	97	123	140	145	87	161	101	109	119	132	118
HPV7	146	168	155	155	161	188	191	113	157	185	138	129	140	111	117	126	140	127	136	113	116	118
HPV40	108	181	140	167	155	170	184	119	**205**	**218**	125	102	137	107	102	138	129	110	138	98	136	103
HPV32	151	120	140	151	109	**243**	**219**	84	133	152	128	137	147	115	100	123	124	150	134	104	96	122
HPV42	103	186	103	128	189	**224**	199	66	146	147	119	123	114	146	111	138	152	105	122	113	100	115
HPV1a	83	67	67	96	68	70	75	70	90	72	79	81	88	85	65	92	57	76	99	66	72	75
HPV63	75	65	96	104	62	69	65	53	83	81	115	97	88	79	99	88	97	65	97	61	79	66
HPV41	66	86	62	62	62	82	51	65	71	62	75	56	65	84	78	66	94	84	59	60	50	62
HPV4	52	54	52	56	62	56	71	76	57	77	62	52	77	64	66	55	85	86	79	79	85	78
HPV65	60	55	68	68	66	64	77	61	69	61	60	74	64	57	60	56	95	76	79	81	81	80
HPV19	87	66	76	76	71	80	68	71	126	72	71	65	103	80	94	136	59	67	85	54	51	77
HPV25	139	65	83	74	63	55	77	45	104	76	64	64	50	55	95	108	56	69	60	68	68	85
HPV14d	72	64	68	68	61	70	56	57	131	107	68	135	75	56	87	115	75	66	56	68	55	70
HPV5	77	59	57	57	62	51	91	56	77	58	69	76	55	62	75	94	73	66	63	58	55	65
HPV5b	77	59	57	57	62	75	82	53	77	58	62	76	62	61	75	90	80	64	53	62	55	65
HPV47	72	70	70	70	68	78	66	64	61	67	77	80	60	70	75	79	60	60	68	61	62	75
HPV12	59	57	57	57	57	51	73	58	91	72	63	75	54	51	88	94	55	73	67	55	74	59
HPV8	68	116	79	79	63	68	80	53	82	78	68	69	96	70	97	86	92	71	84	97	97	89
HPV15	53	64	66	66	69	61	67	61	77	74	72	68	88	60	57	61	65	55	63	57	66	51
HPV17	68	56	62	62	55	60	55	58	62	62	67	68	56	51	62	52	61	59	70	66	56	56
HPV9	62	81	63	63	58	45	41	73	66	78	53	57	62	67	72	67	66	64	71	58	56	56
HPV49	73	61	84	75	54	72	76	48	80	91	67	70	101	80	69	71	62	62	84	65	58	78

	Group F									Group G					Group H											
	Fa			Fb		Fc		Fd		Ga			Gb		Ha								Hb			
	27	57	2a	3	10	7	40	32	42	1a	63	41	4	65	19	25	14d	5	5b	47	12	8	15	17	9	49
HPV31	83	101	136	157	96	146	98	151	103	83	75	66	61	60	87	139	51	77	77	84	81	84	45	57	62	73
HPV52	**206**	192	188	178	110	168	181	120	186	67	65	86	54	60	66	65	61	63	63	70	57	60	64	53	81	61
HPV35	189	172	185	155	88	155	140	140	103	67	96	62	52	68	74	83	69	62	62	70	57	72	69	53	63	84
HPV35h	193	172	198	155	88	155	167	151	128	96	104	62	56	68	70	74	68	62	62	70	57	72	66	62	68	75
HPV16	84	138	82	91	115	161	155	109	189	68	62	62	60	66	63	61	86	62	88	68	47	61	69	55	58	54
HPV33	163	162	88	180	105	188	170	**243**	**224**	94	69	82	67	64	105	52	61	51	75	68	75	68	61	60	49	72
HPV58	179	178	161	185	110	191	184	**219**	199	76	65	59	71	77	72	72	56	91	82	65	73	60	67	55	67	76
RhPV1	87	71	77	62	99	113	119	82	66	64	50	55	76	61	64	45	57	56	56	64	58	53	69	58	73	53
HPV6b	184	161	193	135	120	157	**205**	133	146	90	83	71	57	69	136	110	134	82	73	64	85	82	77	62	66	72
HPV11	175	152	184	112	123	185	**218**	152	147	72	81	62	77	74	119	70	57	57	57	67	69	72	72	62	78	91
HPV13	122	158	131	164	101	138	125	128	119	79	115	75	62	69	71	61	48	69	62	77	63	68	78	67	51	66
PCPV1	145	94	145	150	97	129	102	137	123	81	97	56	52	60	113	122	135	108	91	65	121	88	68	68	57	70
HPV34	185	117	184	118	123	140	137	147	114	88	88	63	77	78	103	52	75	55	62	77	54	96	88	61	63	101
HPV39	153	150	162	113	109	111	107	115	146	106	79	84	64	61	80	57	55	63	63	70	53	60	69	51	67	80
HPV68	127	119	150	94	145	115	102	100	111	68	99	78	66	51	106	95	87	75	75	75	88	78	57	54	72	69
HPV18	133	143	133	118	109	126	138	123	138	92	88	66	56	62	136	108	115	94	90	79	94	86	61	51	51	71
HPV45	133	140	133	100	161	140	129	124	152	92	97	94	68	61	46	56	71	58	80	60	55	76	65	62	68	62
HPV51	122	159	84	117	101	127	110	150	105	65	65	84	76	76	69	66	66	66	64	73	55	71	55	52	50	62
HPV26	121	114	121	99	109	136	138	134	122	99	97	58	79	79	85	82	57	64	62	69	67	69	63	70	71	84
HPV30	70	77	91	91	119	113	98	104	113	66	61	57	79	81	54	68	68	58	62	68	64	97	55	45	66	65
HPV53	113	81	71	100	132	116	101	96	100	62	79	72	85	81	51	68	55	65	65	65	74	97	66	57	57	55
HPV56	85	104	109	110	138	147	103	122	115	75	66	62	78	80	59	85	70	56	56	75	60	89	54	50	52	78
HPV27	**3500**	**874**	**1300**	158	87	158	167	82	96	92	85	87	62	60	74	104	55	69	60	65	83	87	65	52	-	74
HPV57	**874**	**3650**	**603**	152	159	167	149	139	143	80	85	55	48	80	62	77	55	50	48	70	43	61	50	70	45	70
HPV2a	**1300**	**603**	**3370**	149	87	131	140	89	87	78	88	54	52	46	50	60	61	69	68	70	61	60	45	75	45	70
HPV3	158	152	149	**3280**	**733**	180	184	175	170	95	60	84	84	84	94	52	67	51	54	67	62	71	55	47	52	66
HPV10	133	159	87	**733**	**3430**	156	159	86	152	88	51	65	74	75	111	47	76	59	59	73	68	72	66	46	74	55
HPV7	158	167	131	180	156	**4065**	**737**	170	**232**	68	87	75	79	84	59	63	64	60	62	82	67	75	61	60	60	71
HPV40	167	149	140	184	159	**737**	**3570**	123	**213**	87	77	55	78	78	53	56	68	95	86	65	57	73	61	61	73	77
HPV32	82	139	89	175	134	170	123	**3395**	**322**	72	66	64	52	56	63	69	54	70	70	73	52	68	61	57	65	62
HPV42	133	143	87	170	152	**232**	**213**	**322**	**3425**	76	55	75	62	62	73	78	61	73	64	79	71	78	67	65	74	61
HPV1a	80	80	78	95	88	68	87	74	76	**4895**	**203**	78	74	60	147	75	88	95	85	89	107	68	88	88	83	71
HPV63	85	85	88	60	43	87	72	66	55	**203**	**2775**	54	62	94	40	51	74	58	59	65	51	60	93	93	77	80
HPV41	87	62	66	84	46	75	70	60	75	78	54	**3140**	61	53	63	70	52	62	62	57	47	56	63	60	55	67
HPV4	62	42	45	84	74	79	78	52	62	69	58	61	**2795**	**816**	71	105	114	97	102	87	83	63	146	113	68	92
HPV65	60	80	52	84	74	84	64	56	62	58	94	53	**816**	**2665**	91	62	118	102	97	78	102	133	55	62	56	65
HPV19	53	54	56	94	111	62	53	56	66	147	64	63	71	91	**2350**	**918**	**278**	**256**	**206**	**220**	**283**	**257**	88	88	83	112
HPV25	104	77	53	48	42	55	56	69	65	106	51	70	105	68	**918**	**2350**	**270**	**253**	189	199	**308**	**232**	75	75	87	94
HPV14d	55	55	50	55	73	64	68	61	51	88	74	52	114	118	**278**	**270**	**2375**	**217**	**246**	**280**	**255**	**326**	70	67	77	77
HPV5	52	52	63	50	59	55	69	63	61	95	58	58	97	102	**256**	**253**	**217**	**2390**	**1443**	**297**	**461**	**452**	67	67	94	105
HPV5b	52	52	60	50	59	62	69	63	61	87	59	61	102	97	**206**	189	**246**	**1443**	**2430**	**318**	**461**	**434**	73	75	94	82
HPV47	65	70	70	60	73	82	65	60	58	79	69	77	87	78	**220**	199	**280**	**297**	**318**	**2430**	**310**	**344**	93	96	95	87
HPV12	45	53	61	50	50	55	53	46	71	110	63	56	83	102	**283**	**308**	**255**	**461**	**461**	**310**	**2375**	**717**	81	89	95	93
HPV8	121	78	115	114	72	94	74	69	78	76	60	79	65	133	**257**	**232**	**326**	**452**	**434**	**344**	**717**	**2270**	93	73	85	122
HPV15	65	51	52	55	66	61	61	61	67	88	93	63	52	55	88	75	70	67	73	93	81	93	**1925**	**686**	**219**	130
HPV17	63	56	55	45	48	57	61	53	60	88	93	60	113	62	67	60	67	61	61	66	89	63	**686**	**1895**	**204**	149
HPV9	48	-	45	45	74	60	73	65	74	83	77	55	61	56	83	87	77	94	94	95	95	85	**219**	**204**	**1825**	155
HPV49	74	70	70	66	46	71	77	62	61	71	80	67	60	75	112	94	77	105	82	87	93	122	130	149	155	**2095**

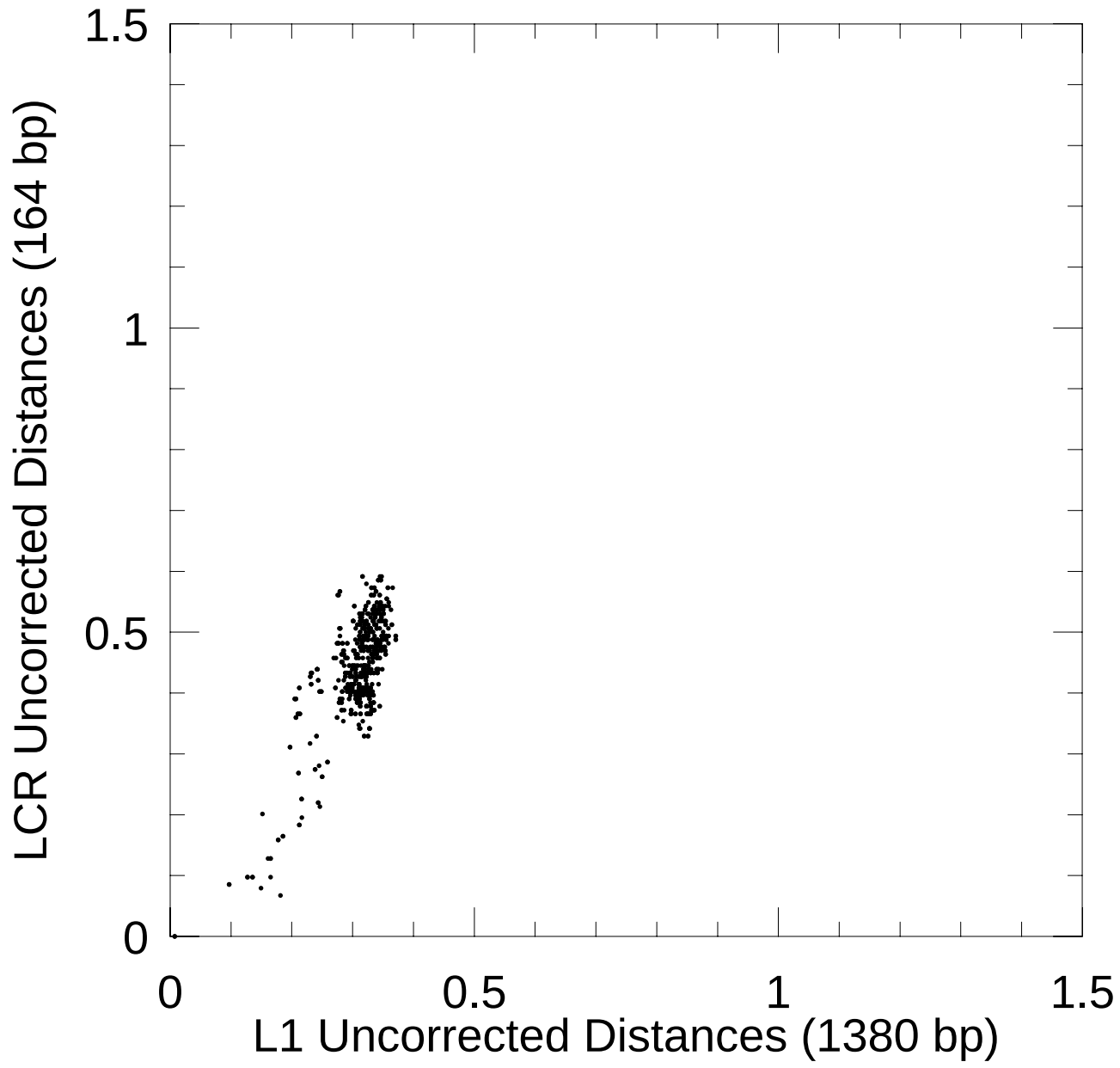
1.5
1
0.5
0
0
0.5
1
1.5
LCR Uncorrected Distances (164 bp)
L1 Uncorrected Distances (1380 bp)

Fig. III.6

C. Synonymous/Nonsynonymous Frequencies

Of the possible kinds of nucleotide substitutions in a typical viral coding sequence, synonymous substitutions, or so-called "silent" mutations, are the most common. Assuming they have not reached a state of mutational saturation, these can provide a linear measure of genetic variation under minimal selection pressure. Nonsynonymous substitutions, on the other hand (amino-acid replacing changes), tell us something about negative and positive Darwinian selection; under certain circumstances, they also serve as a metric. The majority of sites in a coding sequence will be nonsynonymous targets—approximately 70% to 80% depending on the base composition—however the majority of changes observed (except over short sequences that display positive selection or "overdominance") will be synonymous, because these produce the least negative effects.

Phylogenetic analysis is often indifferent to these frequencies, although as we have argued in Part III.A tree analysis can be based upon third base positions or, alternatively, second base positions in codons (Fig. III.4) as a way of capturing these different observables. With papillomaviral sequences it is especially informative to differentiate the frequencies and ratios of synonymous and nonsynonymous substitutions in order to begin to evaluate the temporal relationships and the selective forces implied by the sequence data. In this section, we will present evidence that

1) PCPV (of the pygmy chimp) and HPV-13, 11 and 6b are probably related by cross-species transmission (if PCPV is not a contaminant);
2) HPV-13, 11 and 6b, and certain other HPVs, are "close" types that have relatively recently diverged from one another;
3) Fitch's *covarion* hypothesis can help illuminate the differences between cutaneous and mucosal HPV L1 proteins.

The analyses that follow use the Nei-Gojobori (N-G) algorithm for determining synonymous and nonsynonymous substitution frequencies in PV sequence sets [1,2]. Nucleotide sequences must be aligned according to codons (the *positional homology* must be by codons) at the outset. Pairwise relationships are then determined; with N sequences, there will be $N(N-1)/2$ pairwise relationships. For synonymous substitutions in the range of 0.0 to approximately 0.3, multiple hits are unlikely. Beyond this range, the analysis underestimates the number of changes, therefore a correction term will be required. The N-G algorithm uses the Jukes-Cantor equation/correction for multiple hits [2]: the uncorrected frequency of synonymous substitutions—the ratio of those that have occurred to all those that could have occurred—is denoted *ps* and the corrected ratio is denoted *ds*. Although most PV relationships approach saturation and therefore require correction estimates, the uncorrrected *ps* values are broadly informative. The Jukes-Cantor correction procedure can be also employed for nonsynonymous changes, however many PV coding sequences reach saturation before 0.3, simply due to intense negative selection pressure.

The first set of analyses examines intertype comparisons over PV L1 coding sequences (Figures III.7a-d). For simplicity, only uncorrected frequencies, *pn* and *ps*, are shown. From Figure III.7a it is immediately apparent that a discrete pattern of evolutionary relationships exists among the PVs; this pattern can be dissected (Figures III.7b-d) to show that:

1) Most relationships have attained mutational saturation in synonymous changes, $ps > 0.6$ (theoretical saturation is 0.75).
2) Two discrete clusters of relationships exist with respect to nonsynonymous frequencies, *pn* around 0.25 and *pn* around 0.35.
3) One cluster represents intertypic comparisons either among mucosal PVs only, among the non-EV cutaneous PVs of Group G only, or among the EV-associated cutaneous PVs (Fig. III.7c); the other cluster (Fig. III.7d) represents intertypic comparisons between these three broad classes of PV sequences.
4) A small number of sequence relationships are not part of either of these two clusters ($pn <$ 0.2) and some of these display the lowest *ps* values for intertype comparisons (Fig. III.7b).

Intratype comparisons (not shown) would fall into the lower range of frequencies, $ps < 0.4$ and $pn < 0.1$. Intratype comparisons will be examined in later releases of the compendium when subtypes and variants are emphasized. The type relationships shown in Figure III.7b, noted in 4 above, we term "close" types. The undifferentiated nucleotide differences between these sequences (10% or more) qualifies them to be separate types, but their measured differences in nonsynonymous substitutions are small, < 0.1, leading to modest changes in protein sequences. "Close" types have probably diverged from one another relatively recently. Among them are the close relationships of PCPV (pygmy chimp PV), HPV-13, HPV-6b and HPV-11, analyzed in Figures III.9a-c from the point of view of the PCPV virus and across L1, E6 and E7 coding sequences. From these analyses, we argue that cross-species transmission between nonhuman and human primate PVs is highly likely: either PCPV is an HPV or HPV13, 11 and 6b stem from animal PVs. By this reasoning, lookback estimates of evolutionary rates based upon undifferentiated comparisons of PCPV and HPV13 are called into question.

In Figure III.9, more distant relationships of the PCPV sequences to mostly other HPVs ($pn > 0.2$) are in accord with the result in Figure III.7. We now turn to the observation that two distinct clusters of relationships characterize L1 sequences that are not "close." If saturation of both cutaneous and mucosal HPV nonsynonymous relationships is obtained at around 0.25, how are we to understand the cluster around 0.35 when cross-comparisons are made? One explanation invokes Fitch's *covarion* hypothesis (3). Approximately 25% of the nonsynonymous sites are available to change (due to selection) in the cutaneous and also in the mucosal type L1 sequences. However, the sites in the one group are not the sites in the other group; the set of codons available for change are the "coordinately variable codons," hence covarions. Protein sequence comparisons are currently in progress with the HPVs as a way to test this hypothesis.

Figure III.8 displays the data of Figure III.7 after the Jukes-Cantor correction for multiple hits has been applied. The maximum value of *pn* in Figure III.8 is not increased significantly; only the estimate of synonymous changes is affected. The findings in Figure III.10, based upon E6 coding sequences, are consistent with those of III.7, which was based upon L1.

All of the analyses in Figures III.7–III.10 manifest relatively shallow *initial* slopes, which signify low ratios of nonsynonymous to synonymous substitutions compared to what is seen with influenza (slope = 0.3 in the hemagglutinin gene) or HIV (slope = 0.4 in the envelope gene). The predominant picture is one of stringent Darwinian negative selection. Select subsequences of L1, E6 or E7, of course, could still manifest higher ratios indicative of positive selection.

---

[1] Nei M and Gojobori T: Simple methods for estimating the numbers of synonymous and non-synonymous substitutions. *Mol. Biol. Evol*. 1986;**3**:418–426.

[2] Korber BM, MacInnes K, Smith RF, and Myers G: Mutational trends in V3 loop protein sequences observed in different genetic lineages of human immunodeficiency virus type-1. *J.Virology* 1994;**68**: (Oct issue)

[3] Fitch WM and Markowitz E: An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochemical Genetics* 1970;**4**:579–593.

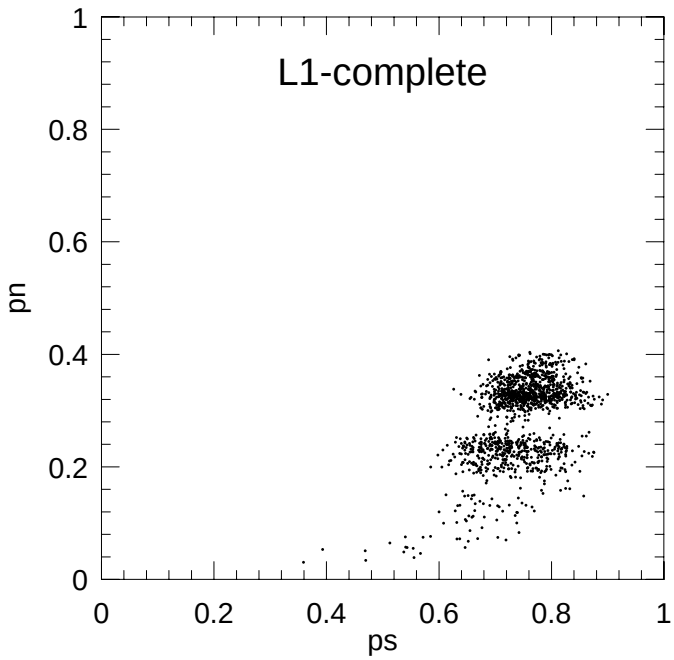

Fig. III.7a

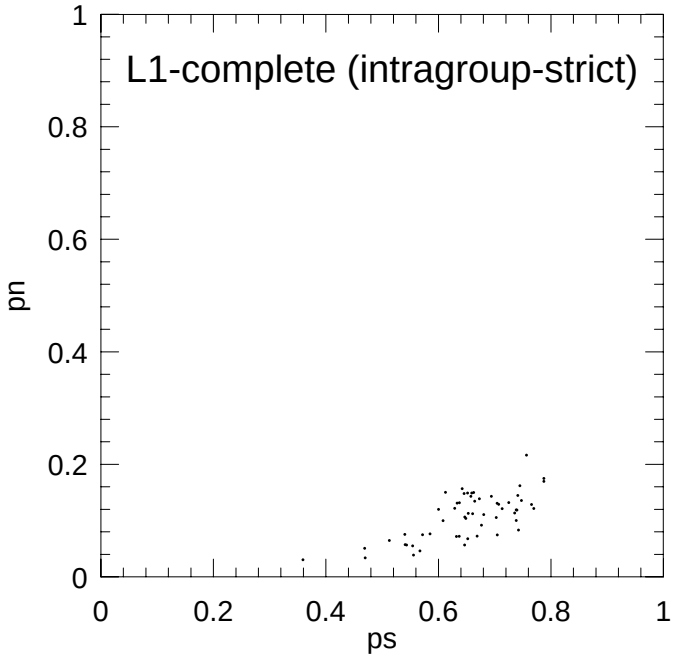

Fig. III.7b

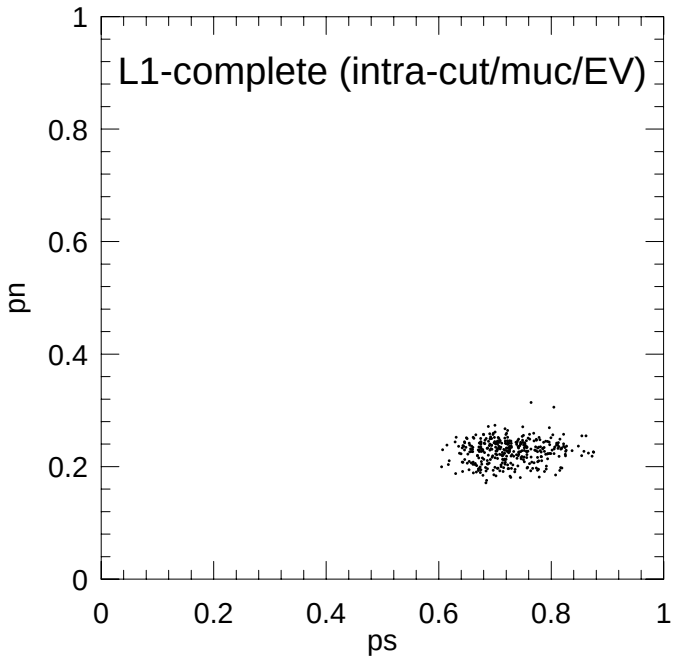

Fig. III.7c

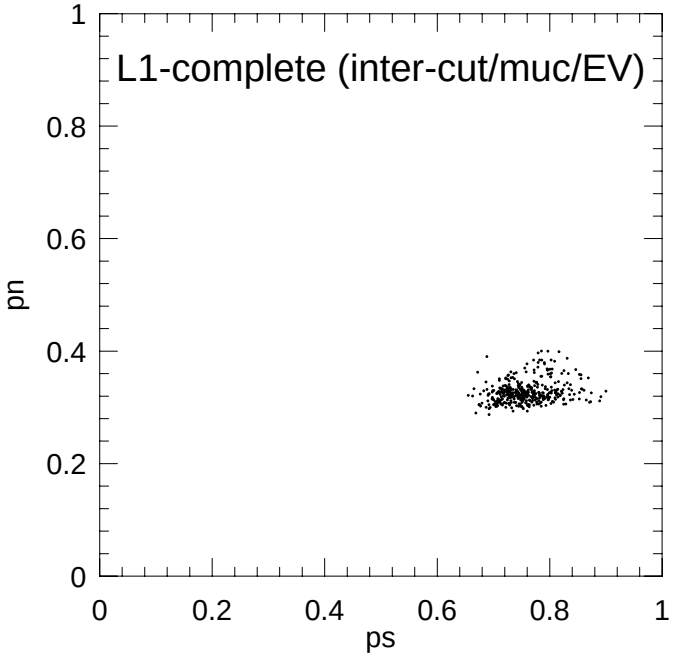

Fig. III.7d

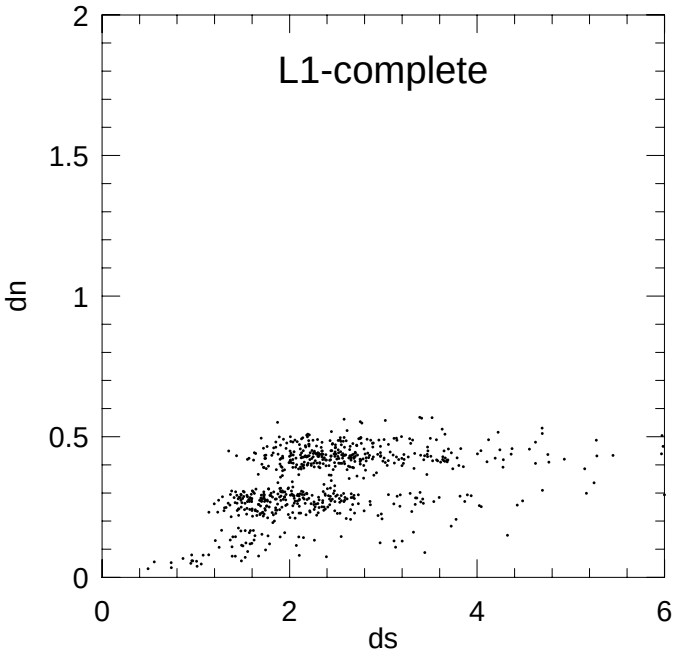

Fig. III.8a

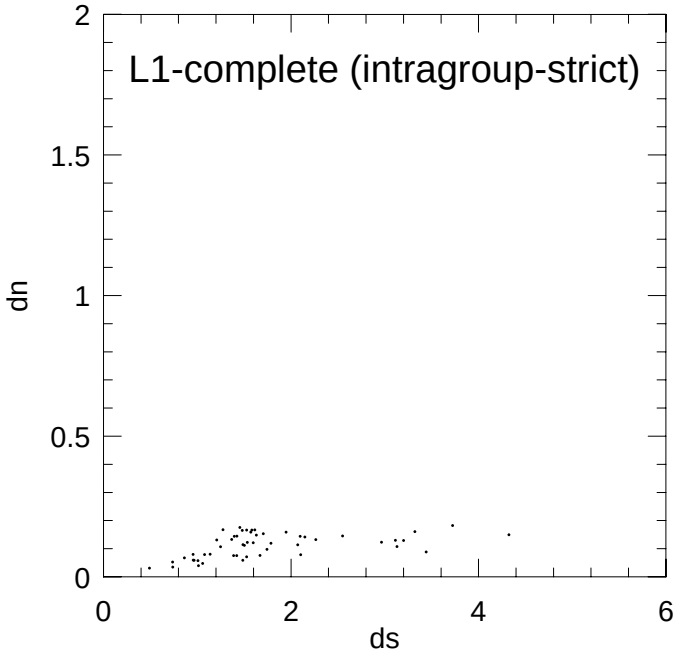

Fig. III.8b

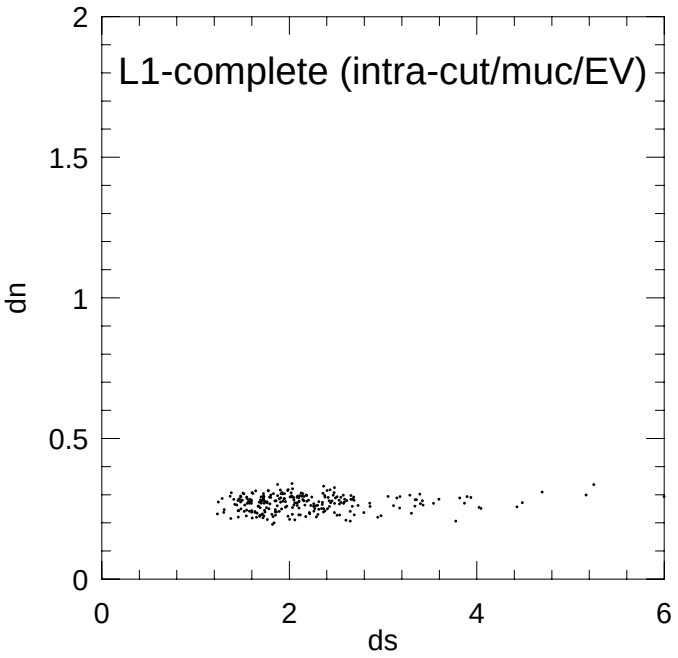

Fig. III.8c

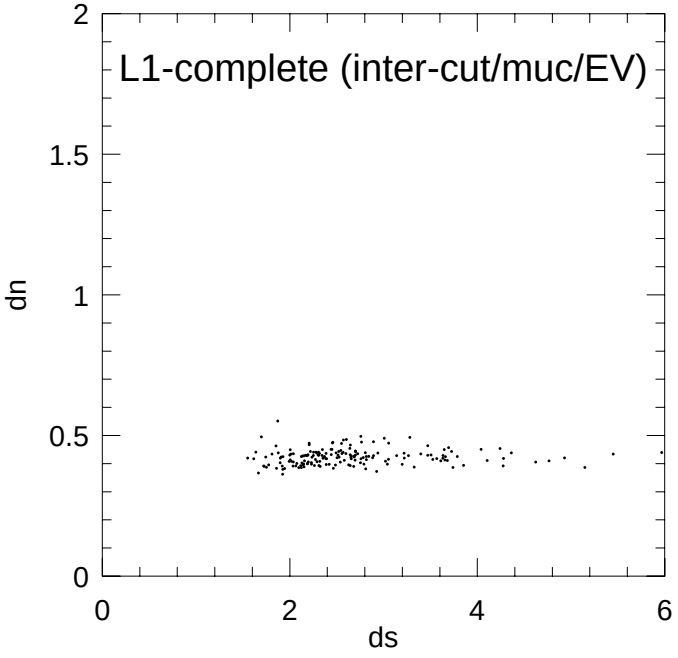

Fig. III.8d

Syn/Nonsyn Frequencies

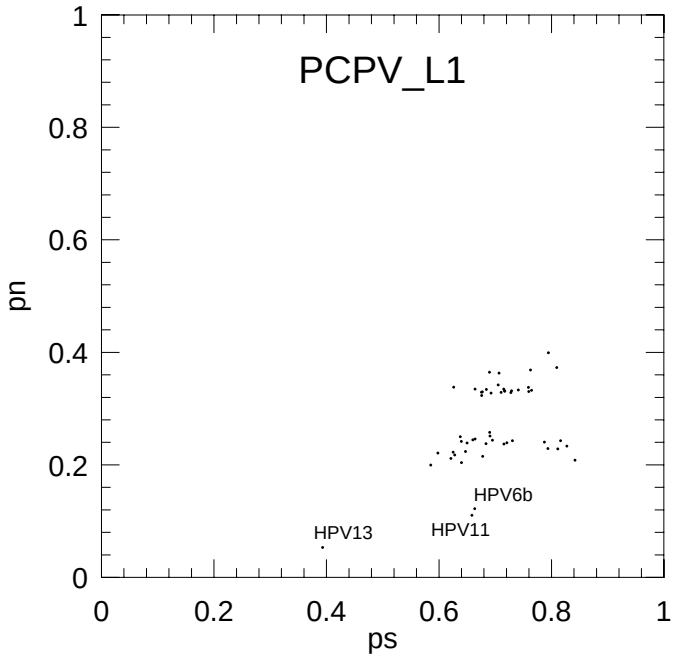

Fig. III.9a

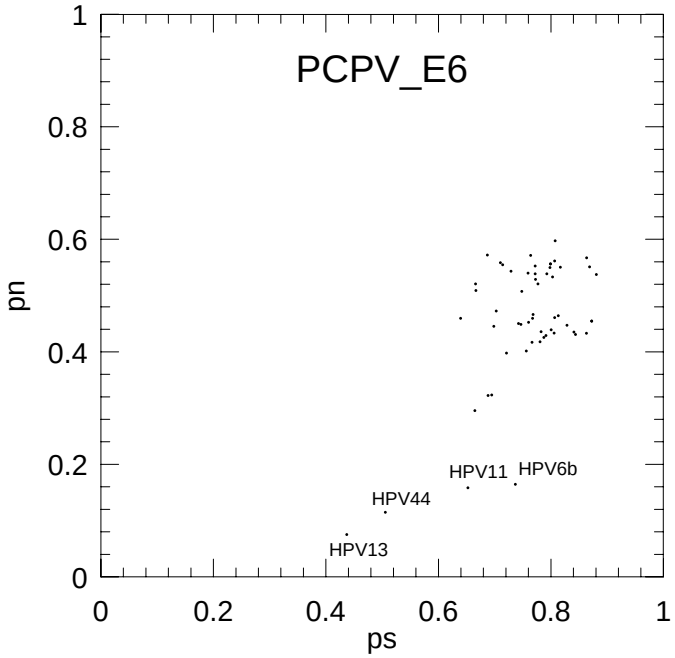

Fig. III.9b

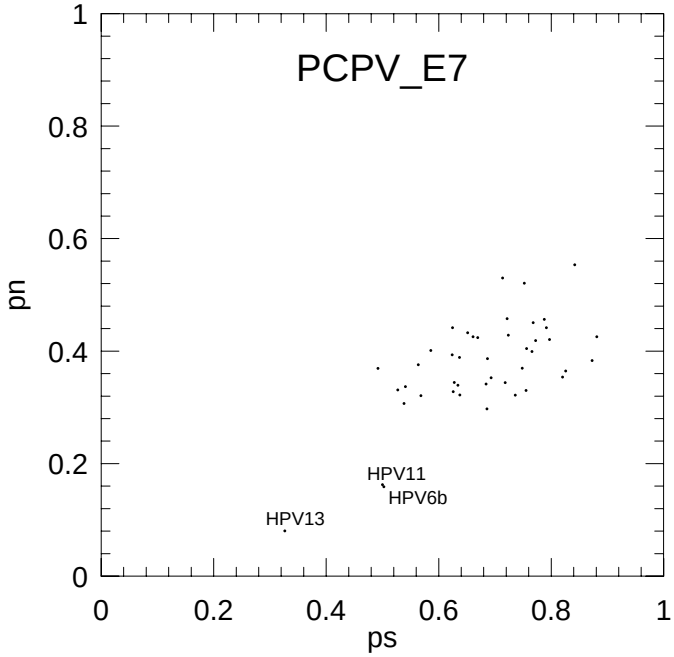

Fig. III.9c

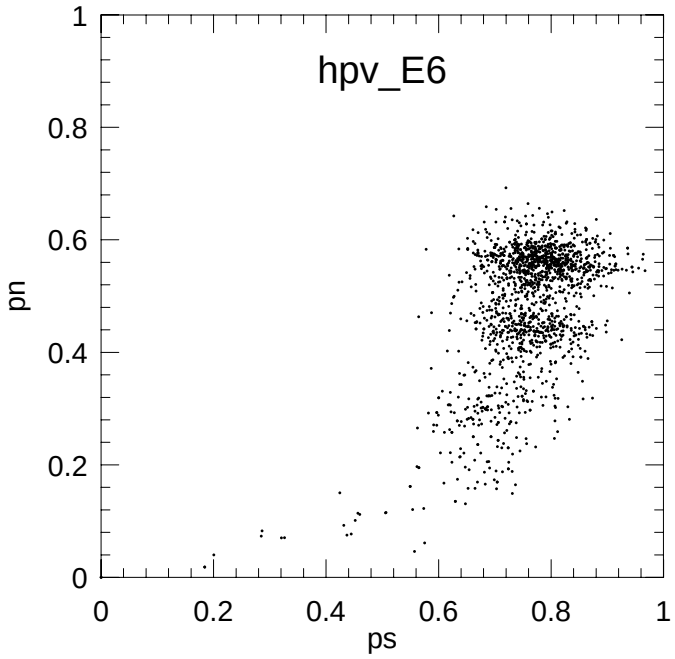

Fig. III.10a

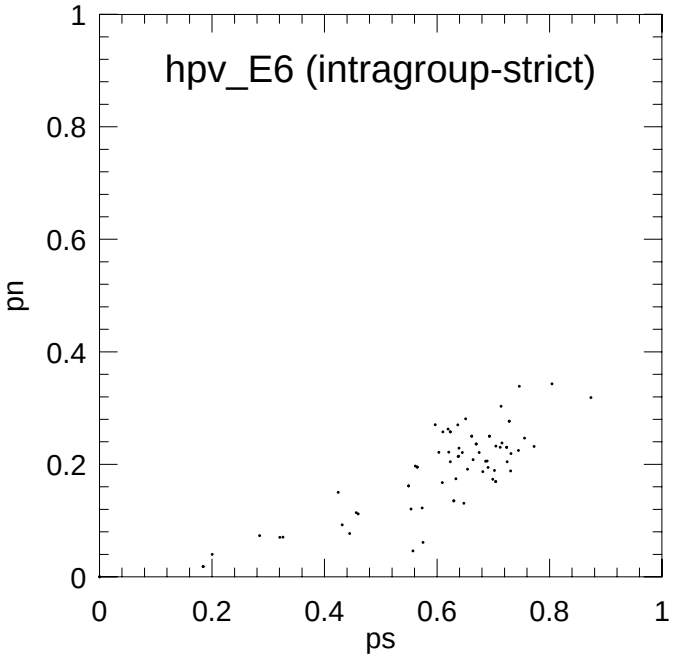

Fig. III.10b

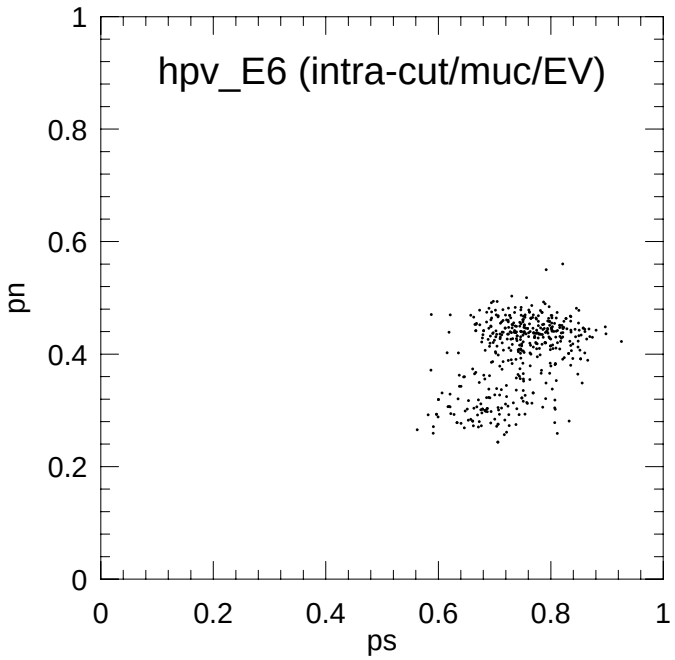

Fig. III.10c

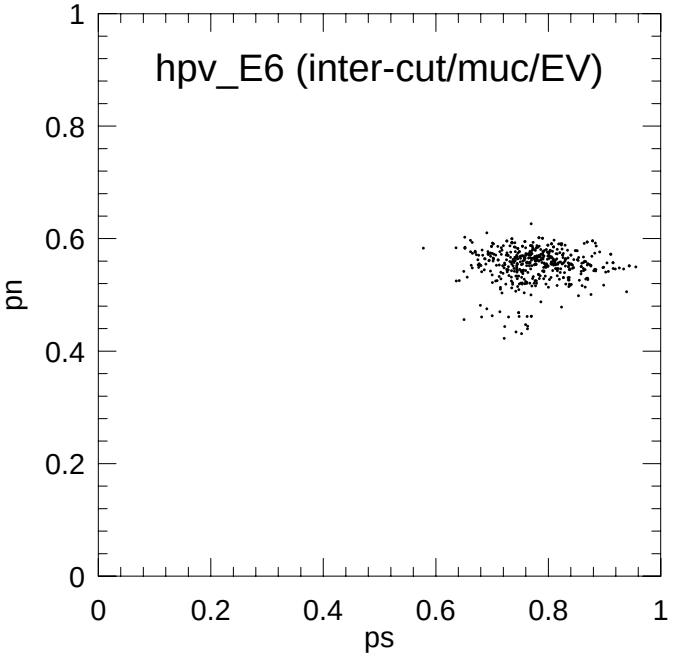

Fig. III.10d

D. Protein Information Content and Density

In Part III.C, HPV coding sequences are shown to be diverse and yet under intense negative selection pressure—the ratio of nonsynonymous substitutions (amino-acid replacing changes) to synonymous substitutions ("silent" changes) is low. Moreover, the assessment of nonsynonymous substitution frequencies does not take into consideration whether the changes are conservative (for example leucine for isoleucine) or nonconservative (for example tryptophan for glycine). We anticipate that amino acid sequence analyses will become a significant part of future database publications. In this section, we explore one approach to relative assessment of HPV protein conservation.

Various "reduced" amino acid codes have been proposed over the years for measurement of protein conservation – a popular code reduces the twenty amino acids to six amino acid classes, for example. The PAM matrices of Dayhoff and coworkers and the BLOSUM matrix of Henifkoff and Henikoff are further examples of substitution schemes. In this section, we utilize the PIMA amino acid similarity scheme of Smith and Smith [1], which was also employed for nucleotide and amino acid sequence alignments in Part II. In this analysis, the various amino acids are hierarchically grouped according to chemical similarities—altogether there are five levels running from perfect matches to a perfect "wild card" [1]. Using an algorithm based on information theory, identical matches are assigned a value of 1.0, i.e., one amino acid equivalent of information content; perfect wild cards are assigned a value of 0.0, i.e., no amino acid equivalents of information; and all other substitutions fall between 0.0 and 1.0, depending upon the extent of amino acid conservation.

For the moment, we are not arguing for the superiority of this analysis over other analyses, many of which are based upon information theoretics. The simplicity of this scheme is that it reports average protein information densities in terms of amino acid equivalents: to take an example, the average information density of six group A (Part I) HPV E1 proteins is 0.63; this implies that on the average there are 0.63 amino acid equivalents of information in the group A E1 sequences. In the following table, the average information densities for E1, E6, E7, L1 and L2 (reported as amino acid equivalents) are listed for the groups of sequences compiled in Part I. (In the cases of groups F and H, the subgroupings discussed in their respective introductions in Part I have been adopted. Group B* excludes the problematic sequence HPV-34 from the original group B.) By inspection of group A results in the first of the three tables, we see that L1 has the highest average and E6 has the lowest. L1 is usually the most conserved protein of the groups; however, E6 is not always the least conserved protein. Group G sequences (cutaneous HPVs) are highly diverse compared to the other groups; although the number of group G sequences is not high, they are clearly a qualitatively diverse group.

The actual fractions depend upon the number of sequences analyzed – with greater numbers of sequences, increasing variability is encountered down to some characteristic asymptotic value. Given the arbitrariness regarding the number and makeup of sequences to be analyzed, relative information densities take on greater meaning. Hence the quantities have been normalized to first E6 and then to L1, as a way of revealing selection differentials. We find, for example, that the ratio of L1 to E6 and E7 in the cutaneous HPVs (group G), 1.0 to 0.43 and 0.54, is dramatically different from an otherwise virtually homogeneous result (1.0 to around 0.8). Because E1 and L2 are also disproportionately variable in the group G sequences, relative to L1, we can conclude that the L1s in this group of highly divergent viruses are extraordinarily conserved. EV viruses, although often clustered with cutaneous viruses, display ratios (selection differentials) that are very similar to those seen in mucosal HPVs.

To determine an asymptote for the information densities, sequences were successively added, first within groups, then across groups, to assess the decline in average information; this is shown in Figure III.11. Throughout all HPVs, E7 appears to be the most variable protein, with an average information density below 0.1. At any given position in the E7 amino acid sequence, on the average there are fewer than 0.1 amino acid equivalents of information, when 1.0 represents perfect conservation. Human immunodeficiency virus mutates extremely rapidly compared to HPV, and the selective pressures on HIV proteins are not stringent; nevertheless, HPVs have apparently evolved over such a long time span that an asymptotic density of 0.1 is well below those of HIV [2] and

of many cellular proteins: 107 alpha hemoglobins have an average of 0.20 and 42 IG heavy chain precursors have an average of 0.21 [1].

---

[1] Smith RF and Smith TF: Automatic generation of primary sequence patterns from sets of related sequences. *Proc. Natl. Acad. Sci. U.S.A.* 1990; **87**:118–121.

[2] Myers G and Pavlakis GN: Evolutionary potential of complex retroviruses. In: *'The Retroviridae, Volume 1* JA Levy (Ed.). Plenum Press, New York 1992; pp. 51–105.

Average Information Densities:

		E1	E6	E7	L1	L2
GROUPA	(6 seqs)	0.631539	0.59463	0.633842	0.768479	0.636568
GROUPB*	(3 seqs)	0.860564	0.818773	0.729598	0.871207	0.798709
GROUPC	(4 seqs)	0.766316	0.730202	0.659497	0.807699	0.750392
GROUPD	(5 seqs)	0.648965	0.601662	0.511477	0.717874	0.621554
GROUPFa	(3 seqs)	0.94996	0.892294	0.84253	0.932946	0.90712
GROUPFb	(2 seqs)	0.894817	0.872043	0.881855	0.925489	0.927382
GROUPFc	(2 seqs)	0.926424	0.94656	0.887502	0.965249	0.928333
GROUPFd	(2 seqs)	0.932223	0.8589	0.795405	0.93098	0.819227
GROUPG	(5 seqs)	0.409478	0.28041	0.224722	0.520737	0.282329
GROUPHa	(6 seqs)	0.809768	0.677186	0.702447	0.826868	0.804279
GROUPHb	(4 seqs)	0.696488	0.573673	0.585423	0.790532	0.729516

Normalized to E6:

		E1	E6	E7	L1	L2
GROUPA	(6 seqs)	1.06207	1	1.06594	1.29237	1.07053
GROUPB*	(3 seqs)	1.05104	1	0.891087	1.06404	0.975495
GROUPC	(4 seqs)	1.04946	1	0.903171	1.10613	1.02765
GROUPD	(5 seqs)	1.07862	1	0.850107	1.19315	1.03306
GROUPFa	(3 seqs)	1.06463	1	0.944229	1.04556	1.01662
GROUPFb	(2 seqs)	1.02612	1	1.01125	1.06129	1.06346
GROUPFc	(2 seqs)	0.978727	1	0.937608	1.01974	0.980744
GROUPFd	(2 seqs)	1.08537	1	0.926074	1.08392	0.95381
GROUPG	(5 seqs)	1.46028	1	0.801405	1.85706	1.00684
GROUPHa	(6 seqs)	1.19578	1	1.0373	1.22104	1.18768
GROUPHb	(4 seqs)	1.21409	1	1.02048	1.37802	1.27166

Normalized to L1:

		E1	E6	E7	L1	L2
GROUPA	(6 seqs)	0.821804	0.773775	0.824801	1	0.828348
GROUPB*	(3 seqs)	0.987784	0.939815	0.837457	1	0.916784
GROUPC	(4 seqs)	0.948764	0.904052	0.816513	1	0.929049
GROUPD	(5 seqs)	0.90401	0.838116	0.712489	1	0.865826
GROUPFa	(3 seqs)	1.01824	0.956426	0.903085	1	0.972318
GROUPFb	(2 seqs)	0.966859	0.942251	0.952853	1	1.00205
GROUPFc	(2 seqs)	0.959777	0.980638	0.919454	1	0.961755
GROUPFd	(2 seqs)	1.00134	0.922576	0.854374	1	0.879962
GROUPG	(5 seqs)	0.786343	0.538487	0.431546	1	0.542172
GROUPHa	(6 seqs)	0.97932	0.818977	0.849527	1	0.972681
GROUPHb	(4 seqs)	0.881037	0.72568	0.740543	1	0.922817

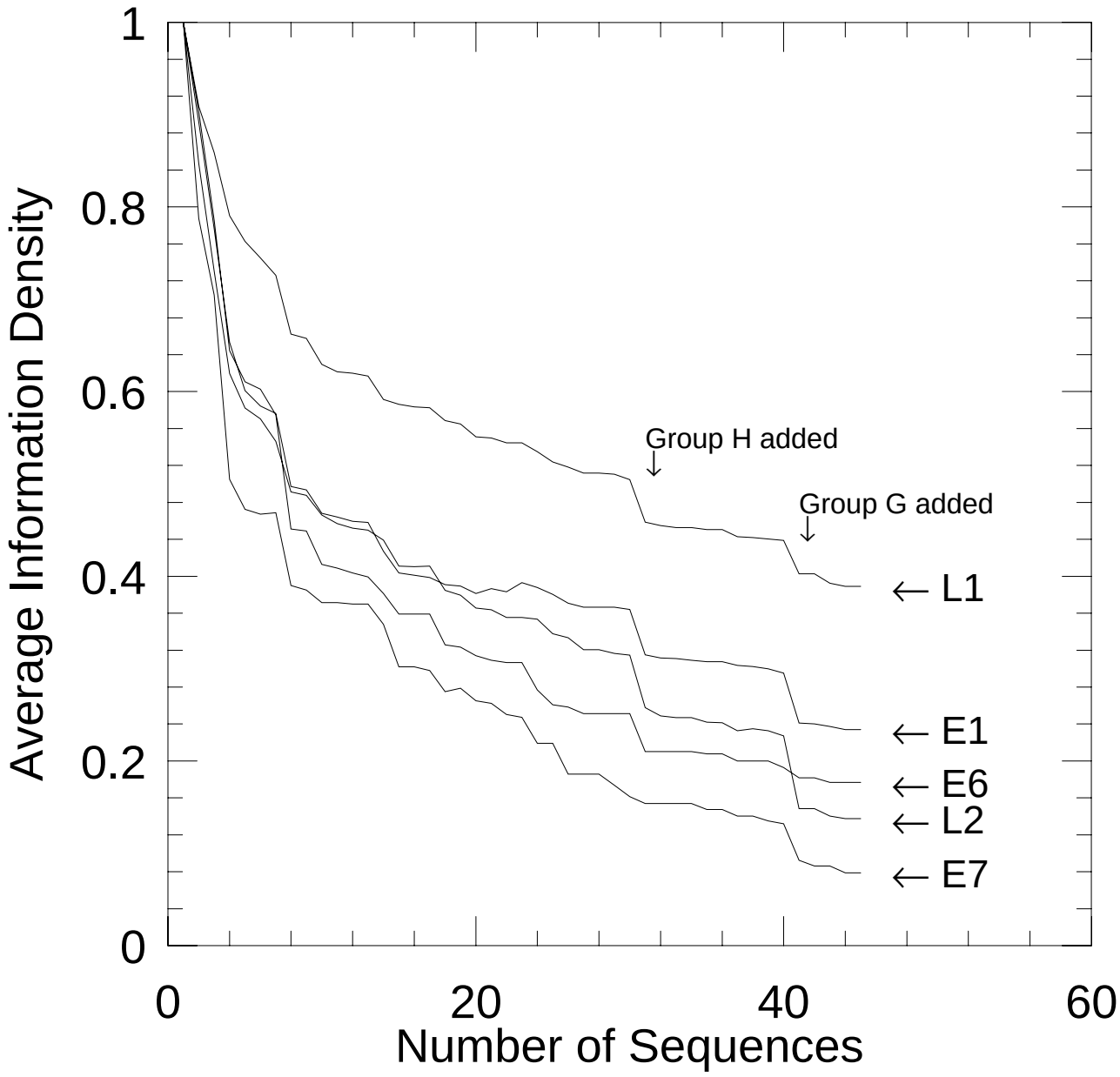
Average Information Density
1
0.8
0.6
0.4
0.2
0
0
20
40
60
Number of Sequences
Group H added
Group G added
← L1
← E1
← E6
← L2
← E7

Fig. III.11