Genome Sequence Comparison and Scenarios for Gene Rearrangements: A Test Case

SRIDHAR HANNENHALLI,* COLOMBE CHAPPEY,† EUGENE V. KOONIN,† AND PAVEL A. PEVZNER*,[1]

*Department of Computer Science and Engineering, Pennsylvania State University, University Park, Pennsylvania 16802; and †National Center for Biotechnology Information, National Library of Medicine, NIH, Bethesda, Maryland 20894

Received March 28, 1995; accepted September 7, 1995

As large portions of related genomes are being sequenced, methods for comparing complete or nearly complete genomes, as opposed to comparing individual genes, are becoming progressively more important. A major, widespread phenomenon in genome evolution is the rearrangement of genes and gene blocks. There is, however, no consistent method for genome sequence comparison combined with the reconstruction of the evolutionary history of highly rearranged genomes. We developed a schema for genome sequence comparison that includes three successive steps: (i) comparison of all proteins encoded in different genomes and generation of genomic similarity plots; (ii) construction of an alphabet of conserved genes and gene blocks; and (iii) generation of most parsimonious genome rearrangement scenarios. The approach is illustrated by a comparison of the herpesvirus genomes that constitute the largest set of relatively long, complete genome sequences available to date. Herpesviruses have from 70 to about 200 genes; comparison of the amino acid sequences encoded in these genes results in an alphabet of about 30 conserved genes comprising 7 conserved blocks that are rearranged in the genomes of different herpesviruses. Algorithms to analyze rearrangements of multiple genomes were developed and applied to the derivation of most parsimonious scenarios of herpesvirus evolution under different evolutionary models. The developed approaches to genome comparison will be applicable to the comparative analysis of bacterial and eukaryotic genomes as soon as their sequences become available. © 1995 Academic Press, Inc.

INTRODUCTION

Complete sequences of small genomes and large portions of the sequences of more complex genomes are rapidly accumulating. Hundreds of viral genome sequences and about 20 organellar sequences are currently available, and the genomes of several bacteria and yeast are expected to be completed within 2–3 years (Bork *et al.*, 1994). Accordingly, methods for sequence comparison on a genomic scale are becoming progressively more important. Comparing long, distantly related genomes directly is extremely computationally intensive and is not highly informative, as the conservation is manifest primarily at the level of protein sequences rather than that of nucleotide sequences as such. In addition, standard sequence comparison algorithms do not reveal gene rearrangements.

Comparison of related genomes shows that their evolution includes numerous changes in the gene order, or genome rearrangements. The analysis of rearrangements in genome evolution was pioneered in the 1930s by Dobzhansky and Sturtevant, who found a rearrangement scenario with 17 inversions for different *Drosophila* species (Dobzhansky and Sturtevant, 1938). With the advent of large-scale DNA mapping and sequencing, the number of attempts to analyze genome rearrangements is rapidly growing in different areas, including the evolution of mitochondrial genomes of plants (Palmer and Herbon, 1987, 1988; Atlan and Couvet, 1993), fungi (Bruns and Palmer, 1989), and animals (Hoffmann *et al.*, 1992; Sankoff *et al.*, 1992); chloroplast genomes (Milligan *et al.*, 1989; Knox *et al.*, 1993; Hoot and Palmer, 1994); genomes of lambdoid bacteriophages (Casjens *et al.*, 1992) and small viruses (Koonin and Dolja, 1993; Hull, 1992); and mammalian chromosomes (Nadeau and Taylor, 1984; Nadeau *et al.*, 1992; Zakharov *et al.*, 1992). In some of these cases, attempts have been made to construct evolutionary scenarios, i.e., to derive the shortest (most parsimonious) series of rearrangements explaining the existing diversity of genome organization (Koonin and Dolja, 1993; Palmer, 1992). These analyses, however, were based on heuristic approaches, and, as recently shown (Bafna and Pevzner, 1995a), in many complicated cases they have overlooked the most parsimonious scenario(s).

Analysis of genome rearrangements involves a combinatorial problem of finding the shortest series of evo-

[1] To whom correspondence should be addressed at present address: Departments of Computer Science and Mathematics, University of Southern California, Los Angeles, CA 90089-1113. Telephone: (213) 740-2407. Fax: (213) 740-2424. E-mail: ppevzner@hto.usc.edu.

0888-7543/95 $12.00

lutionary events that is required to transform one gene order into another. For genomes consisting of more than 10 conserved gene blocks, exhaustive searches over all potential solutions is extremely computationally intensive, and until recently, there were no algorithms for systematic analysis of genome rearrangements. The first steps toward a computational theory of genome rearrangements have been made recently (Sankoff *et al.*, 1992; Sankoff, 1993; Kececioglu and Sankoff, 1993, 1994; Bafna and Pevzner, 1993, 1995a,b; Kececioglu and Ravi, 1995; Hannenhalli and Pevzner, 1995). However, methods to analyze rearrangements that account for gene order diversity in more than two genomes remain largely unexplored (see Sankoff *et al.*, 1995, for an initial study in this area).

We are interested in developing a schema that would lead from genome sequences to ordered sets of homologous genes and from these gene orders to genome rearrangement scenarios. This requires a combination of methods for sequence comparison with genome rearrangement algorithms.

We chose the seven complete and three partially sequenced genomes of herpesviruses as our test case—the largest set of relatively long genomic sequences available to date (Table 1). Nucleotide sequences of herpesviruses are so divergent that direct comparisons are generally inadequate for evolutionary studies, and alternative approaches have been proposed (Schachtel *et al.*, 1991; Karlin *et al.*, 1994). On the other hand, a set of genes that encode proteins with significant amino acid sequence similarities and that form partially conserved clusters has been revealed (Davison and Taylor, 1986; McGeoch, 1989, 1992; Chee *et al.*, 1990; Albrecht *et al.*, 1992; Bublot *et al.*, 1992; McGeoch *et al.*, 1993; McGeoch and Cook, 1994). It has to be noted that these studies have been performed in a labor-consuming, case-by-case fashion.

We are unaware of an automated approach to genome sequence comparison, and this work is an attempt to fill this gap. A method to define the alphabet of conserved genes and gene blocks and to derive the most parsimonious evolutionary scenarios is proposed, and the application of this method to the analysis of herpesvirus evolution is described. It has to be stressed that the main goal of this work is the development of methods for genome sequence comparison, rather than the study of specific aspects of herpesvirus biology.

MATERIALS AND METHODS

Nucleotide sequences of herpesvirus DNA and the amino acid sequences of the encoded proteins were extracted from the GenBank database (Table 1).

Separate databases, each containing the proteins encoded by a single herpesvirus, were generated, and each protein from one genome was compared to each protein from the other genomes using the BLASTP program (Altschul *et al.*, 1990, 1994). The program GCT (Genome Comparison Tool) was developed to perform all the pairwise comparisons between the amino acid sequences from the different viruses automatically and to visualize the results of these comparisons. The input for the GCT program is the GenBank feature table, i.e., an ordered set of (putative) protein sequences encoded in a given genome. Under this program, the function GCPLOT produces a genomic dot plot that displays the pairs of proteins with BLAST scores exceeding a certain cutoff, and another function, called GCTAB, automatically generates a table of conserved genes among the different genomes and depicts their order in each genome. The source code in C and the executable version of the GCT program for Unix platforms are available by anonymous ftp at ftp.cse.psu.edu in the directory / pub/hannenha/koonin.

Multiple amino acid sequence alignments were constructed using the MACAW program (Schuler *et al.*, 1991). Tentative phylogenetic trees were generated from juxtaposed ungapped alignment sections using the program NEIGHBOR (neighbor-joining algorithm) implemented in the PHYLIP package Version 3.5 (Felsenstein, 1989).

RESULTS AND DISCUSSION

Alphabet of Conserved Genes and Gene Blocks

Sequence comparison of the complete sets of proteins encoded in the genomes under study is the first step in our schema. This is required for any study of genome rearrangements to define the "alphabet" of conserved genes and to identify conserved gene blocks. So far, such comparisons generally have been performed in a time-consuming, case-by-case fashion, and despite the recognized importance of this task (Cedergren *et al.*, 1990; O'Brien, 1991), we are unaware of programs for the automatic generation of gene orders from multiple DNA sequences. We developed such programs and applied them to herpesvirus genomes. Again, it has to be indicated that herpesviruses are considered only as a test case; sequence similarities between herpesvirus proteins have been studied in great detail (Davison and Taylor, 1987; Chee *et al.*, 1990; Albrecht *et al.*, 1992; Bublot *et al.*, 1992), and it is unlikely that a significant number of additional conserved genes will be revealed.

For each pair of herpesvirus genomes, the protein product of each putative gene from one genome was compared with the protein products of all putative genes in the other genome using the BLASTP program. BLAST scores above the chosen threshold of 75, which roughly corresponds to the probability of matching by chance of 10^{-4} (given the total length of the herpesvirus proteins), were used as an indicator that genes from two genomes are homologous. Figure 1, generated using the procedure GCPLOT of the GCT program, shows the genomic similarity plots for (a) two closely related and (b) two distantly related herpesvirus genomes. The plot in Fig. 1a shows a high level of conservation of the gene order, whereas that in Fig. 1b reveals the differences in the genome organization resulting from inversions and transpositions.

BLAST comparisons are not transitive; i.e., if A is related to B and B is related to C at a certain level of significance, this does not necessarily imply that A is related to C at the same level. Therefore, a group of genes was considered to be a family of homologs if the encoded proteins formed a connected component in the graph of significant pairwise similarities, not necessarily requiring that all pairs have a score above the cutoff.

TABLE 1

Sequencing of Herpesvirus Genomes

Virus	Acronym	Class	Genomize size (kb)	Status	Number of sequenced genes	Reference	GenBank Accession No.
Herpes simplex virus	HSV	α	152	Complete	76	McGeoch *et al.*, 1989	X14112
Varicella-zoster virus	VZV	α	124	Complete	71	Davison *et al.*, 1986	X04370
Equine herpesvirus	EHV	α	150	Complete	81	Telford *et al.*, 1992	M86664
Human cytomegalovirus	HCMV	β	229	Complete	206	Chee *et al.*, 1990	X17403
Epstein–Barr virus	EBV	γ	172	Complete	87	Baer *et al.*, 1984	v01555
Herpesvirus saimiri	HVS	γ	112	Complete	78	Albrecht *et al.*, 1990	X64346
Channel catfish virus	CCV	?	134	Complete	94	Davison, 1992	M75136
Human herpesvirus 6	HHV6	β	162	Incomplete	33	Neipel *et al.*, 1991	a
Bovine herpesvirus-4	BHV	γ	145	Incomplete	33	Bublot *et al.*, 1992	a
Murine herpesvirus	MHV	γ	115	Incomplete	18	Efstathiou *et al.*, 1990	a

As a result of such multiple comparisons, we identified families of genes that are conserved among the herpesviruses (procedure GCTAB). In several cases, we used the notion of positional relatedness (Albrecht *et al.*, 1992) as an aid to infer relationships between poorly conserved genes. Positionally related genes are those that are located between neighboring homologous genes. If the BLAST score for the protein products of such genes is higher than their scores with any other virus protein, this is an indication that they are likely to be homologs. Such weakly conserved but apparently homologous genes were manually incorporated in the preliminary version of the conserved gene table produced by GCTAB. In addition, for the proteins encoded by positionally related genes, multiple alignments were generated, and conserved motifs were sought. The list of the conserved herpesvirus genes is shown in Table 2. Together, 27 conserved genes were found; two of these, however, are lacking in cytomegalovirus. The results of our comparisons were compatible with those of previous analyses (Albrecht *et al.*, 1992; McGeoch *et al.*, 1993). A unique letter was assigned to each conserved gene across all the genomes, resulting in a conserved gene alphabet (Table 2).

Figure 2a shows the order of the conserved genes in herpesvirus genomes, with a sign associated with each gene indicating the direction of transcription. The genomes of alphaherpesviruses are comprised of L (large) and S (small) segments that undergo systematic inversion during the virus replication (reviewed in McGeoch, 1989). Therefore, in Fig. 2b, we show all the virus genomes in the same orientation, which does not necessarily correspond to the isomer represented in GenBank.

Herpesvirus DNA replication typically includes circular intermediates, and linear virion molecules apparently are produced by cleavage of head-to-tail concatemers (Poffenberger and Roizman, 1985; Hammerschmidt *et al.*, 1988; Fraefel *et al.*, 1993). Therefore, some of the dramatic differences in the conserved gene order of alpha- and betaherpesviruses on the one hand, and gammaherpesviruses on the other hand, may be readily accounted for by an alternative concatemer scission (albeit the frequency of this hypothetical event is not known). Figure 2b shows the comparison of the alpha-, beta-, and gammaherpesvirus genome organizations, with the latter modified according to the alternative scission scheme so as to maximize the similarity with the other two classes.

The conserved gene alphabet may be further reduced to an alphabet of conserved blocks. A conserved block is defined as a maximal substring of letters occurring in the same relative ordering, i.e., either as $\{X_i \cdots X_{i+n-1}\}$ or as $\{-X_{i+n-1} \cdots -X_i\}$ for a block of *n* genes, in all the genomes. We will assign "+" to $\{X_i \cdots X_{i+n-1}\}$ and "−" to $\{-X_{i+n-1} \cdots -X_i\}$. For example, the substring of letters O, P, and Q in Figs. 2a and 2b occurs as either {+O + P + Q} or {−Q − P − O} (inversion of }+O + P + Q}) in all the completely sequenced genomes and, hence, qualifies as a conserved block. There are seven conserved blocks in the herpesvirus genomes, and three distinct arrangements of these blocks correspond to the alpha-, beta-, and gammaherpesviruses (Fig. 2c). For further discussion, we consider HSV, CMV, and EBV as prototypes of each of these genome organizations, respectively. Eight of the conserved genes were found in CCV, but none of the conserved herpesvirus gene blocks could be detected in this virus, in accord with previous results indicating only a very distant relationship (Davison, 1992).

The representation of herpesvirus genomes in the conserved block alphabet is a gross oversimplification, as it completely leaves out the genes that are not conserved among alpha-, beta-, and gammaherpesviruses. There are a number of such genes; some of them are obviously homologous to known host genes and should have been acquired from the host genome relatively recently on the evolutionary scale (reviewed in McGeoch, 1989, 1992; McGeoch *et al.*, 1993). This apparent recombination between the viral and the host genomes is a very important aspect of herpesvirus evolution. Nevertheless, it can be ignored for the purpose of reconstructing the history of genome rearrangements.

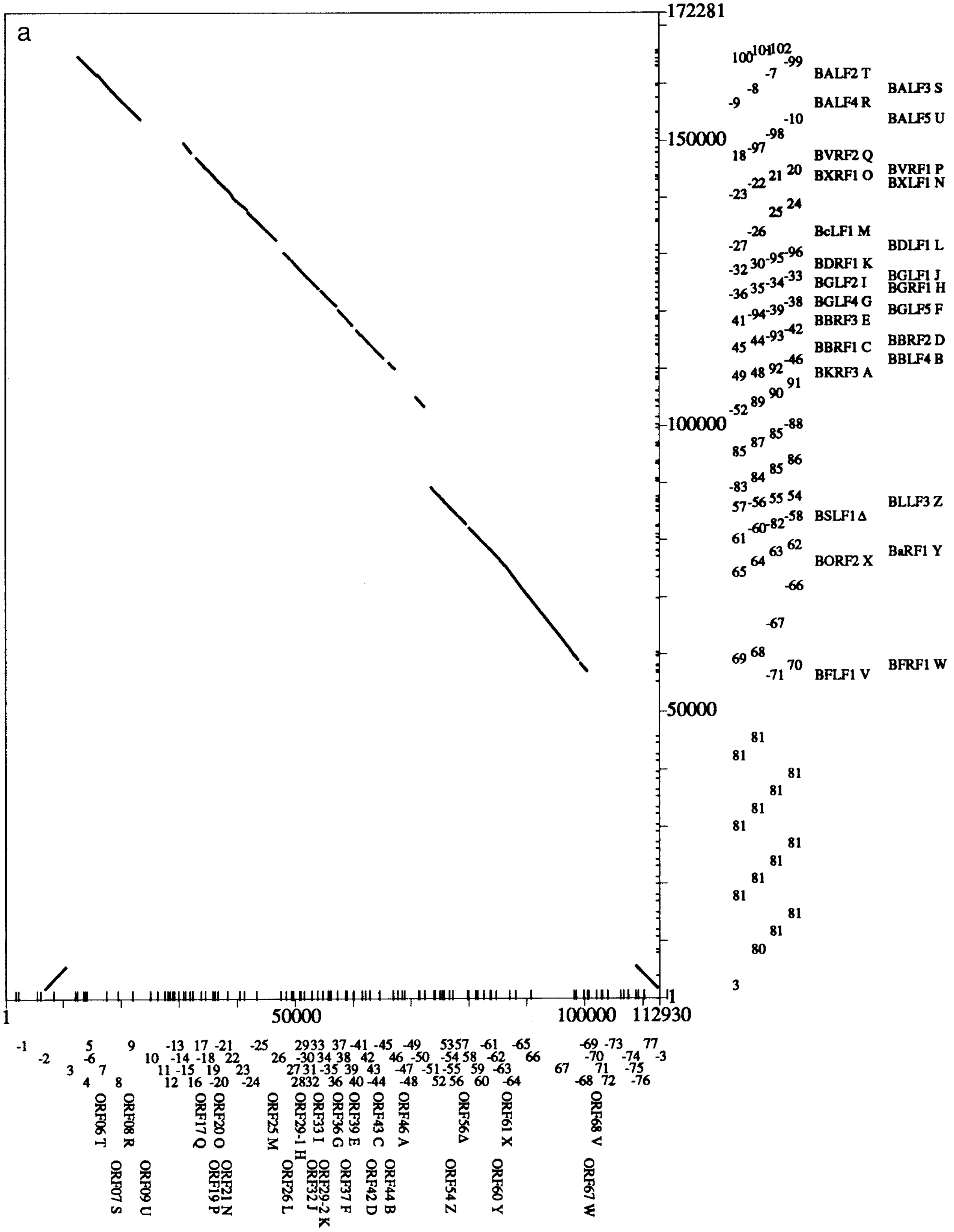

FIG. 1. Genomic similarity plots. The figure, generated by the GCPLOT program, shows gene by gene comparisons of pairs of viral genomes; each streak corresponds to a pair of genes whose proteins products scored over 75 in BLASTP comparisons. The signs indicate the direction of transcription. (**a**) Closely related genomes—herpesvirus saimiri vs Epstein–Barr virus. The HVS genome is shown as a sequence of 77 genes, most of which have homologs in EBV. Homologous genes are indicated by identical numbers, e.g., in the gene alphabet, 24 corresponds to ORF 23 in HVS and to BTRF1 in EBV. The conserved genes are indicated by letters as in Table 2. (**b**) Distantly related genomes—human cytomegalovirus vs Epstein–Barr virus. Only the genes on the borders of conserved gene blocks are shown.

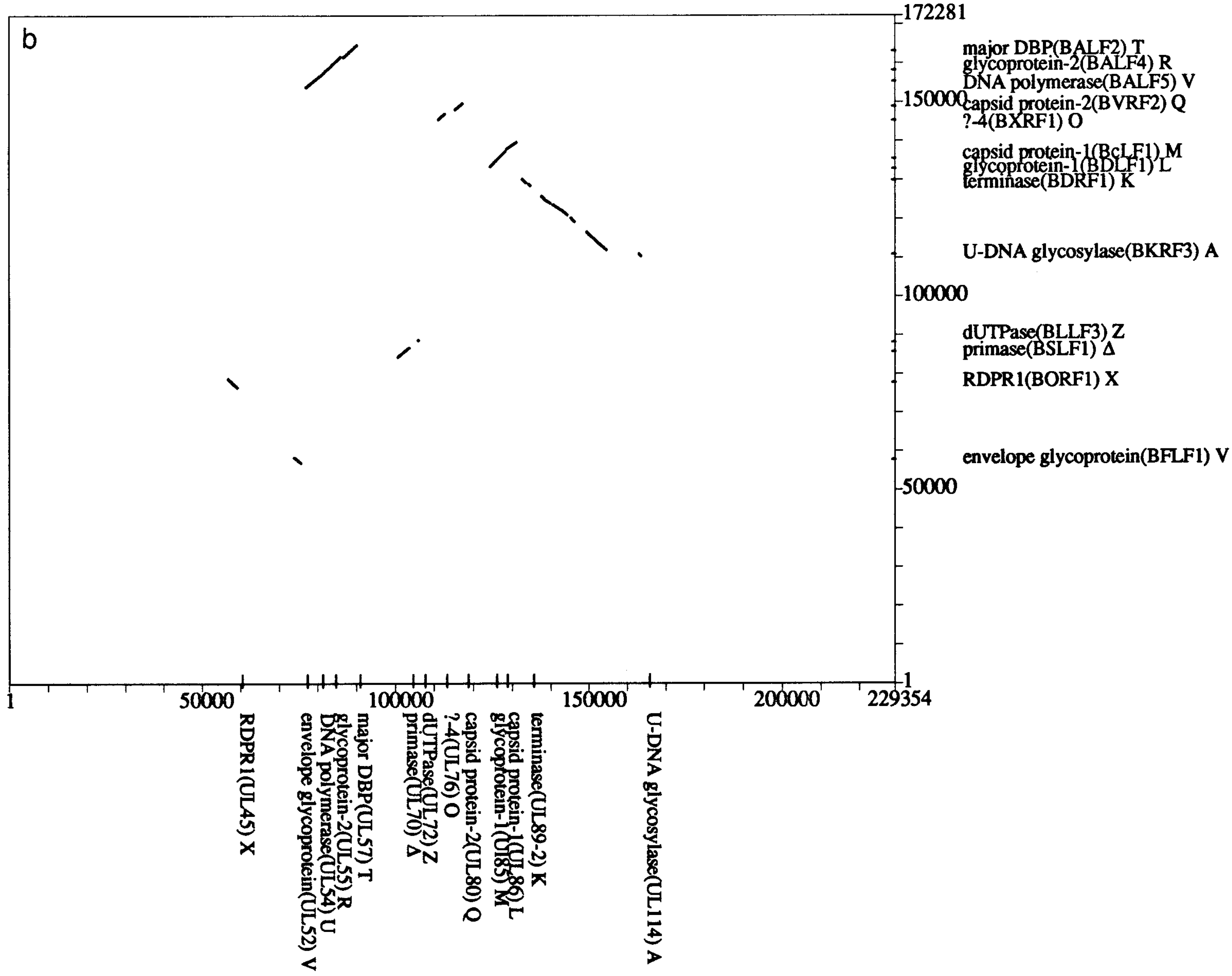

FIG. 1—*Continued*

Genome Rearrangements

The herpesvirus family is the largest set of completely sequenced, relatively long genomes available to date. To our knowledge, a most parsimonious scenario of genome rearrangements in herpesviruses has not been discussed so far. Below, we derive such scenarios for two models: (i) rearrangement by inversions only and (ii) rearrangement by inversions and transpositions.

The first model assumes that inversions of gene blocks containing an arbitrary number of genes constitute the only genome rearrangement mechanism. Despite obvious shortcomings, this model has proved to be useful for the analysis of rearrangements in mitochondrial and chloroplast DNAs. Algorithms for analyzing genome evolution by inversions have been developed by Kececioglu and Sankoff (1993, 1995). Below, we give a formal definition of sorting by inversions and describe the ideas leading to an efficient algorithm for the analysis of genome rearrangements by inversions (Hannenhalli and Pevzner, 1995).

We define rearrangement distance $d(\pi, \sigma)$ between genomes π $(\pi_1\pi_2 \cdots \pi_n)$ and σ $(\sigma_1\sigma_2 \cdots \sigma_n)$, containing a set of n homologous genes, as the minimum number of rearrangements that is required to transform the gene order of π into the gene order of σ. In the very first computational studies of genome rearrangements, Watterson *et al.* (1982) and Nadeau and Taylor (1984) introduced the notion of a "breakpoint" (disruption of gene order) and noticed some correlations between the rearrangement distance and the number of breakpoints [in fact, Sturtevant and Dobzhansky (1936) implicitly considered these correlations almost 50 years earlier]. However, the estimate of the inversion distance in terms of breakpoints is very rough and does not provide an accurate bound. Recently, another "hidden" parameter that allows one to determine inversion distances with much greater accuracy has been revealed (Bafna and Pevzner, 1993). As this parameter has been overlooked in previous studies of genome rearrangements, we define it below.

First, define a transformation from a signed n-ele-

TABLE 2

The "Alphabet" of the Conserved Herpesvirus Genes

Function	HSV	CMV	VZV	EHV	HVS	EBV	BHV	MHV	HHV6	Letter in the conserved gene alphabet
U-DNA glycosylase	UL2	HCMVUL114	ORF59	ORF61	ORF46	BKRF3	ORF20a			A
Helicase	UL5	HCMVUL105	ORF55	ORF57	ORF44	BBLF4	ORF19			B
Virion protein	UL6	HCMVUL104	ORF54	ORF56	ORF43	BBRF1				C
?$_1$	UL7	HCMVUL103	ORF53	ORF55	ORF42	BBRF2				D
Membrane protein	UL10	HCMVUL100	ORF50	ORF52	ORF39	BBRF3				E
Exonuclease	UL12	HCMVUL98	ORF48	ORF50	ORF37	BGLF5	ORF16		ORF16R	F
Protein kinase	UL13	HCMVUL97	ORF47	ORF49	ORF36	BGLF4			ORF15R	G
Terminase	UL15$_I$	HCMVUL89$_1$	ORF45	ORF47__44$_1$	ORF29$_1$	BGRF1	ORF15		ORF12L	H
?$_2$	UL16	HCMVUL94	ORF44	ORF46	ORF33	BGLF2				I
?$_3$	UL17	HCMVUL93	ORF43	ORF45	ORF32	BGLF1			ORF10R	J
Terminase	UL15$_2$	HCMVUL89$_2$	ORF42	ORF47__44$_2$	ORF29$_2$	BDRF1	ORF14a		ORF7L	K
Capsid protein	UL18	HCMVUL85	ORF41	ORF43	ORF26	BDLF1	ORF13b		ORF3L	L
Major capsid protein	UL19	HCMVUL86	ORF40	ORF42	ORF25	BcLF1	ORF13a	ORF8	ORF4L	M
Thymidine kinase	UL23	—	ORF36	ORF38	ORF21	BXLF1		ORF16		N
?$_4$	UL24	HCMVUL76	ORF35	ORF37	ORF20	BXRF1		ORF15		O
Virion protein	UL25	HCMVUL77	ORF34	ORF36	ORF19	BVRF1	ORF10,11	ORF13		P
Capsid protein	UL26	HCMVUL80	ORF33	ORF35	ORF17	BVRF2		ORF14		Q
Glycoprotein	UL27	HCMVUL55	ORF31	ORF33	ORF08	BALF4	ORF6			R
Processing	UL28	HCMVUL56	ORF30	ORF32	ORF07	BALF3	ORF5b			S
Major DBP	UL29	HCMVUL57	ORF29	ORF31	ORF06	BALF2	ORF2	ORF17		T
DNA polymerase	UL30	HCMVUL54	ORF28	ORF30	ORF09	BALF5	ORF7	ORF6	X1LF0	U
Env. glycoprotein	UL32	HCMVUL52	ORF26	ORF28	ORF68	BFLF1			X1RF1	V
Phosphoprotein	UL34	HCMVUL50	ORF24	ORF26	ORF67	BFRF1			X1LF2	W
RDPR1	UL39	HCMVUL45	ORF19	ORF21	ORF61	BORF2	ORF26	ORF3	P2LF2	X
RDRPs	UL40	—	ORF18	ORF20	ORF60	BaRF1	ORF25			Y
dUTPase	UL50	HCMVUL72	ORF8	ORF9	ORF54	BLLF3				Z
Primase	UL52	HCMVUL70	ORF6	ORF7	ORF56	BSLF1	ORF24			Δ

Note. Each of the partially sequenced herpesvirus genomes is represented by several sequences under different accession numbers that are not indicated. The complete sequence of HHV6 has become available after the submission of the present paper (Gompels *et al.*, 1995).

ment permutation $\pi(\pi_1 \cdots \pi_n)$ (i.e., a permutation with a "+" or "−" sign associated with each gene block as defined above) into a unsigned $2n$-element permutation. For each element i in permutation π, substitute i_a followed by i_b for i if i has an associated plus sign and i_b followed by i_a if i has an associated minus sign. As a result, element i is transformed into two elements, i_a and i_b, corresponding to the beginning and the end of the gene block i, respectively. In addition, the resulting $2n$-element permutation is extended by 0 in the beginning and by $n + 1$ in the end; the elements 0 and $n + 1$ correspond to the beginning and the end of the genome. As a result of this transformation, the 7-element permutation CMV(+1 −2 −3 +7 −4 +5 +6) is transformed into the 16-element permutation CMV′(0 1_a 1_b 2_b 2_a 3_b 3_a 7_a 7_b 4_b 4_a 5_a 5_b 6_a 6_b 8) (Fig. 3), and permutation EBV(+1 +2 +3 −5 +4 +6 +7) is transformed into EBV′(0 1_a 1_b 2_a 2_b 3_a 3_b 5_b 5_a 4_a 4_b 6_a 6_b 7_a 7_b 8).

Define a graph $G(\pi, \sigma)$ with $2n + 2$ vertices 0, 1_a, 1_b, 2_a, 2_b . . . , n_a, n_b, $n + 1$, where n is the number of conserved genes. We join vertices i and j, corresponding to two different genes, by an edge if i and j are neighbors either in π or in σ. For example, vertices 2a and 3b are joined by an edge in the graph G(CMV, EBV) since they are neighbors in CMV′, whereas the vertices 3b and 5b are joined by an edge since they are neighbors in EBV′ (Fig. 3b). $G(\pi, \sigma)$ is a collection of edge-disjoint cycles, i.e., cycles without common edges. Let $c(\pi, \sigma)$ be the number of cycles in this collection, or cycle decomposition of $G(\pi, \sigma)$. In our example, c(CMV, EBV) = 3. Cycle decompositions play an important role in estimating the inversion distances, since it has been proved (Bafna and Pevzner, 1993) that $d(\pi, \sigma) \geqslant n + 1 - c(\pi, \sigma)$. The estimate of the inversion distance in terms of cycle decomposition is much tighter than the estimate based on the number of breakpoints (Kececioglu and Sankoff, 1994). In fact, for all biological examples that we studied, $d(\pi, \sigma) = n + 1 - c(\pi, \sigma)$, thus reducing the inversion distance problem to the cycle decomposition problem (Bafna and Pevzner, 1995a). For example, d(CMV, EBV) = 7 + 1 − c(CMV, EBV) = 5. Moreover, very recently, a new algorithm has been proposed that allows a precise calculation of $d(\pi, \sigma)$ by introducing an additional hidden parameter, number of hurdles, which closes the gap between $d(\pi, \sigma)$ and $n + 1 - c(\pi, \sigma)$ (Hannenhalli and Pevzner, 1995). This algorithm also generates all the most parsimonious scenarios of genome rearrangements by inversions.

Figure 3 shows the arrangement of conserved gene

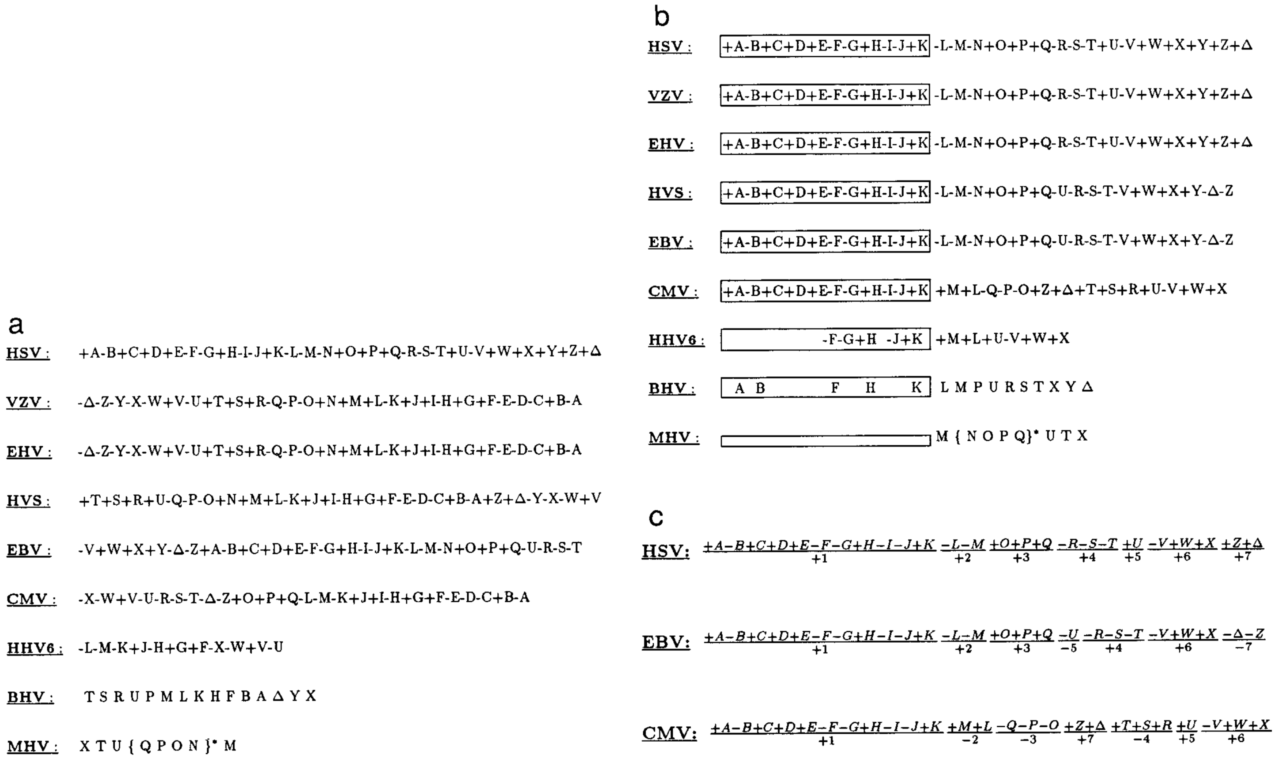
a
HSV : +A-B+C+D+E-F-G+H-I-J+K-L-M-N+O+P+Q-R-S-T+U-V+W+X+Y+Z+Δ
VZV : -Δ-Z-Y-X-W+V-U+T+S+R-Q-P-O+N+M+L-K+J+I-H+G+F-E-D-C+B-A
EHV : -Δ-Z-Y-X-W+V-U+T+S+R-Q-P-O+N+M+L-K+J+I-H+G+F-E-D-C+B-A
HVS : +T+S+R+U-Q-P-O+N+M+L-K+J+I-H+G+F-E-D-C+B-A+Z+Δ-Y-X-W+V
EBV : -V+W+X+Y-Δ-Z+A-B+C+D+E-F-G+H-I-J+K-L-M-N+O+P+Q-U-R-S-T
CMV : -X-W+V-U-R-S-T-Δ-Z+O+P+Q-L-M-K+J+I-H+G+F-E-D-C+B-A
HHV6 : -L-M-K+J-H+G+F-X-W+V-U
BHV : T S R U P M L K H F B A Δ Y X
MHV : X T U { Q P O N }* M
b
HSV : +A-B+C+D+E-F-G+H-I-J+K -L-M-N+O+P+Q-R-S-T+U-V+W+X+Y+Z+Δ
VZV : +A-B+C+D+E-F-G+H-I-J+K -L-M-N+O+P+Q-R-S-T+U-V+W+X+Y+Z+Δ
EHV : +A-B+C+D+E-F-G+H-I-J+K -L-M-N+O+P+Q-R-S-T+U-V+W+X+Y+Z+Δ
HVS : +A-B+C+D+E-F-G+H-I-J+K -L-M-N+O+P+Q-U-R-S-T-V+W+X+Y-Δ-Z
EBV : +A-B+C+D+E-F-G+H-I-J+K -L-M-N+O+P+Q-U-R-S-T-V+W+X+Y-Δ-Z
CMV : +A-B+C+D+E-F-G+H-I-J+K +M+L-Q-P-O+Z+Δ+T+S+R+U-V+W+X
HHV6 : -F-G+H -J+K +M+L+U-V+W+X
BHV : A B F H K L M P U R S T X Y Δ
MHV : M { N O P Q}* U T X
c
HSV: +A-B+C+D+E-F-G+H-I-J+K -L-M +O+P+Q -R-S-T +U -V+W+X +Z+Δ
+1 +2 +3 +4 +5 +6 +7
EBV: +A-B+C+D+E-F-G+H-I-J+K -L-M +O+P+Q -U -R-S-T -V+W+X -Δ-Z
+1 +2 +3 -5 +4 +6 -7
CMV: +A-B+C+D+E-F-G+H-I-J+K +M+L -Q-P-O +Z+Δ +T+S+R +U -V+W+X
+1 -2 -3 +7 -4 +5 +6

FIG. 2. Order of conserved genes and gene blocks in herpesviruses. (**a**) Virus genomes shown in the conserved gene alphabet. Information on gene order in BHV4, HHV6, and MHV was obtained from partial sequences and physical maps. The direction of transcription for BHV4 and MHV is not indicated because of the lack of the respective data. (**b**) The three types of conserved gene order compared after the gammaherpesvirus gene arrangement has been modified under the hypothesis of an alternative cleavage during concatemer maturation. The invariant block of genes is boxed. (**c**) The reduced alphabet of conserved gene blocks.

blocks in CMV and EBV (Fig. 3a) and possible shortest series of rearrangements transforming one of them into the other under a model allowing only inversions (Fig. 3c) or under a model allowing both inversions and transpositions (Fig. 3d). For the case of inversions only, the rearrangement (inversion) distances are

$$d(\text{HSV}, \text{CMV}) = 5, \quad d(\text{CMV}, \text{EBV}) = 5, \quad \text{and } d(\text{EBV}, \text{HSV}) = 3.$$

As we consider the three major herpesvirus classes and no *a priori* root in their phylogeny, only one intermediate species, which may be considered a hypothetical common ancestor, has to be defined (Figs. 4a and 5a). The most parsimonious evolutionary scenario is defined as the shortest series of rearrangements along the three branches of the graph. Let X be the gene order of a unknown hypothetical ancestor and let $d(X, \text{HSV})$, $d(X, \text{EBV})$, and $d(X, \text{CMV})$ be the rearrangement distances from X to HSV, EBV, and CMV, respectively. The overall number of rearrangements in a scenario is given by

$$D = d(X, \text{HSV}) + d(X, \text{EBV}) + d(X, \text{CMV}).$$

The triangle inequality implies

$$d(X, \text{HSV}) + d(X, \text{EBV}) \geqslant d(\text{HSV}, \text{EBV}) = 3$$
$$d(X, \text{HSV}) + d(X, \text{CMV}) \geqslant d(\text{HSV}, \text{CMV}) = 5$$
$$d(X, \text{EBV}) + d(X, \text{CMV}) \geqslant d(\text{CMV}, \text{EBV}) = 5$$

Adding up the above inequalities yields

$$\begin{aligned} D &= d(X, \text{HSV}) + d(X, \text{EBV}) + d(X, \text{CMV}) \\ &\geqslant [(d(\text{HSV}, \text{EBV}) + d(\text{HSV}, \text{CMV}) \\ &\quad + d(\text{CMV}, \text{EBV}))/2] \\ &= [(3 + 5 + 5)/2] = 7. \end{aligned}$$

Thus, the most parsimonious tree for HSV, EBV, and CMV should include at least seven inversions. Only three possible values of $(d(X, \text{HSV}), d(X, \text{EBV}), d(X, \text{CMV}))$, namely (1, 2, 4), (2, 1, 4), and (2, 2, 3), attain the equality $d(X, \text{HSV}) + d(X, \text{EBV}) + d(X, \text{CMV}) = 7$, while at the same time satisfying all of the above inequalities. Let a d-neighborhood of a gene order π be the set of all gene orders that can be derived from π by d rearrangements. Then, for each of the scenarios, the gene order of the ancestor species X corresponding to $(d1, d2, d3)$ was determined as the intersection of $d1$-, $d2$-, and $d3$-neighborhoods of the gene orders of HSV, EBV, and CMV,

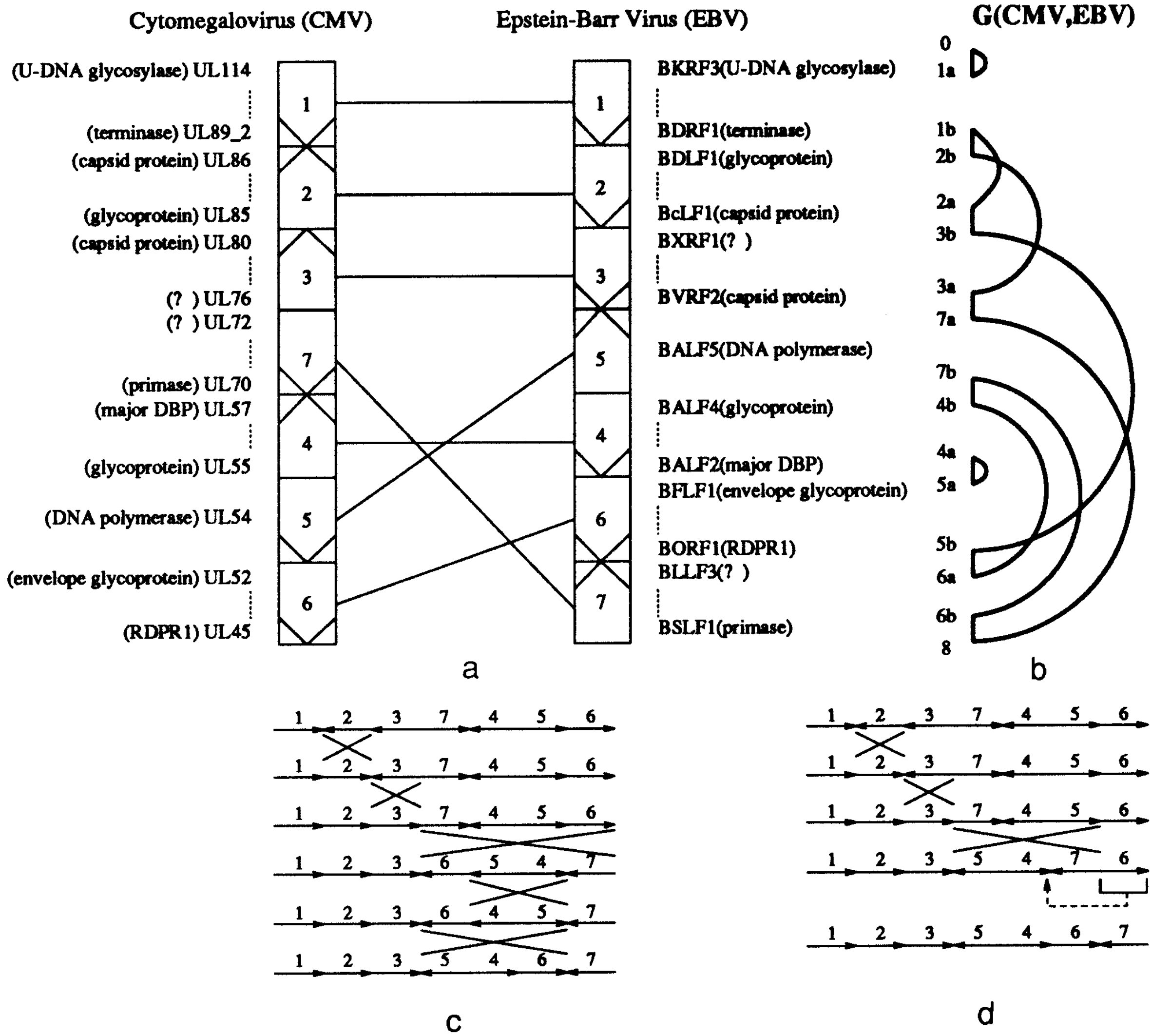

FIG. 3. Genome rearrangements between two distantly related herpesvirus genomes. (**a**) The relationship between the conserved gene blocks in CMV and EBV. (**b**) Graph *G* (CMV, EBV) with three edge-disjointed cycles. (**c**) A shortest series of inversions transforming CMV into EBV over the conserved block alphabet. (**d**) A shortest (with respect to the total count of operations) series of inversions and transpositions transforming CMV into EBV over the conserved block alphabet.

respectively (Fig. 4b). This simple analysis allows one to find a gene order of the putative ancestor species *X*, which can be transformed into the three extant gene organizations with $D = 7$, for the first two of the above three sets of inversion distances; the third did not yield any scenario with seven inversions. Figures 5b and 5c show the two most parsimonious scenarios, with seven inversions each.

Similarly, for the case of inversions combined with transpositions

$$d(X, \text{HSV}) + d(X, \text{EBV}) \geqslant d(\text{HSV}, \text{EBV}) = 3$$

$$d(X, \text{HSV}) + d(X, \text{CMV}) \geqslant d(\text{HSV}, \text{CMV}) = 4$$

$$d(X, \text{EBV}) + d(X, \text{CMV}) \geqslant d(\text{CMV}, \text{EBV}) = 4,$$

the above inequalities were obtained by exhaustive search and verified using the bounds described by Bafna and Pevzner (1995b). After adding up, we have

$$D = d(X, \text{HSV}) + d(X, \text{EBV}) + d(X, \text{CMV})$$
$$\geqslant [(3 + 4 + 4)/2] = 6.$$

The only values of $((d(X, \text{HSV}), d(X, \text{EBV}), d(X, \text{CMV}))$ that satisfy all of the above inequalities are (1, 2, 3), (2, 1, 3), and (2, 2, 2). For each of these possibilities, the unique gene order of the putative ancestor genome was generated in a manner similar to the previous case. Figure 5 shows all the three most parsimonious scenarios, with six rearrangements each; note that for two of these, the gene order of the putative ancestor is identical to that under the inversion only model (compare Figs. 4b, 4c, and 5b, and 5c). Interestingly, there always was only one ancestral gene order meeting all of the above inequalities.

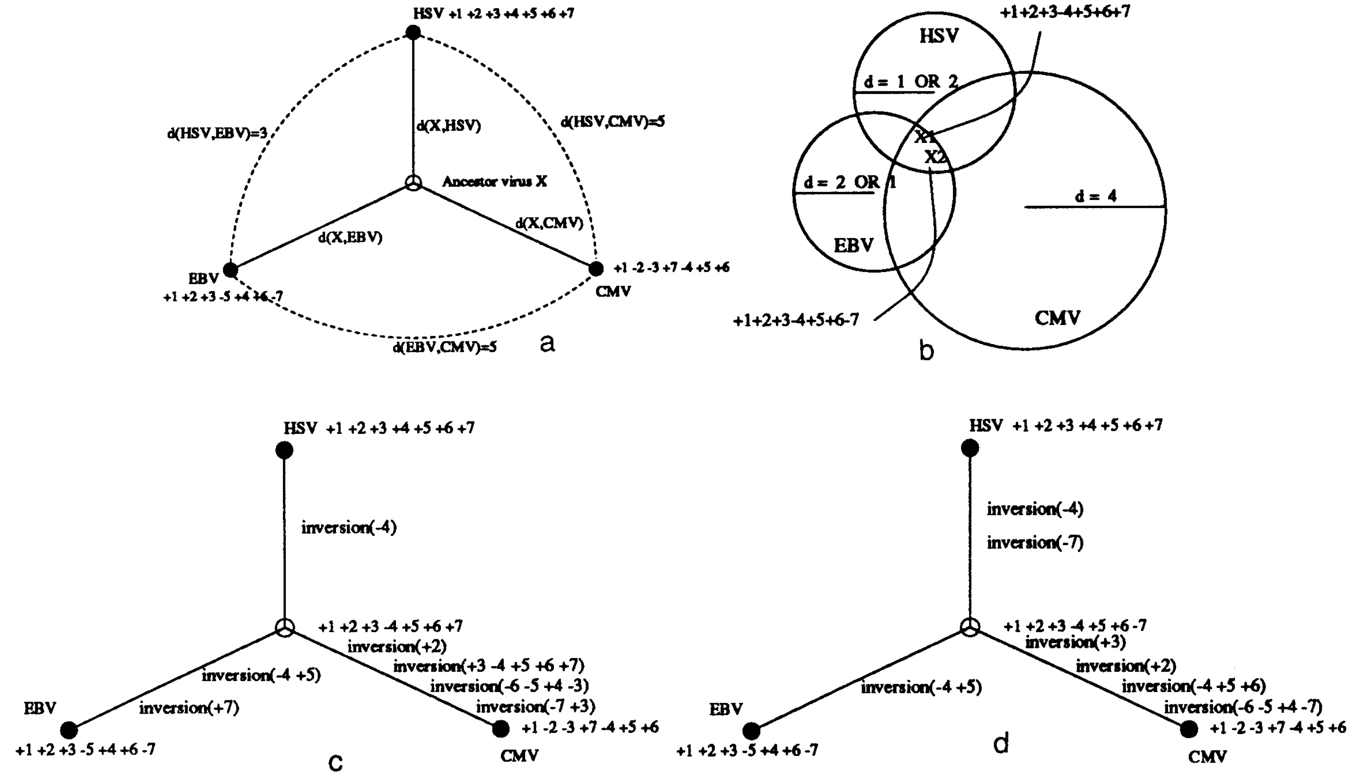

FIG. 4. Most parsimonious evolutionary scenarios for herpesviruses under the "inversions only" model. (**a**) Inversion distances. (**b**) *d*-neighborhoods of the gene orders of CMV, EBV, and HSV. The gene orders that belong to the intersection of all three *d*-neighborhoods and that may represent the gene order of the ancestral species *X* are shown. (**c** and **d**) The two possible scenarios with seven inversions each. HSV, CMV, and EBV represent the organization of gene blocks that is conserved in alpha-, beta-, and gammaherpesviruses, respectively.

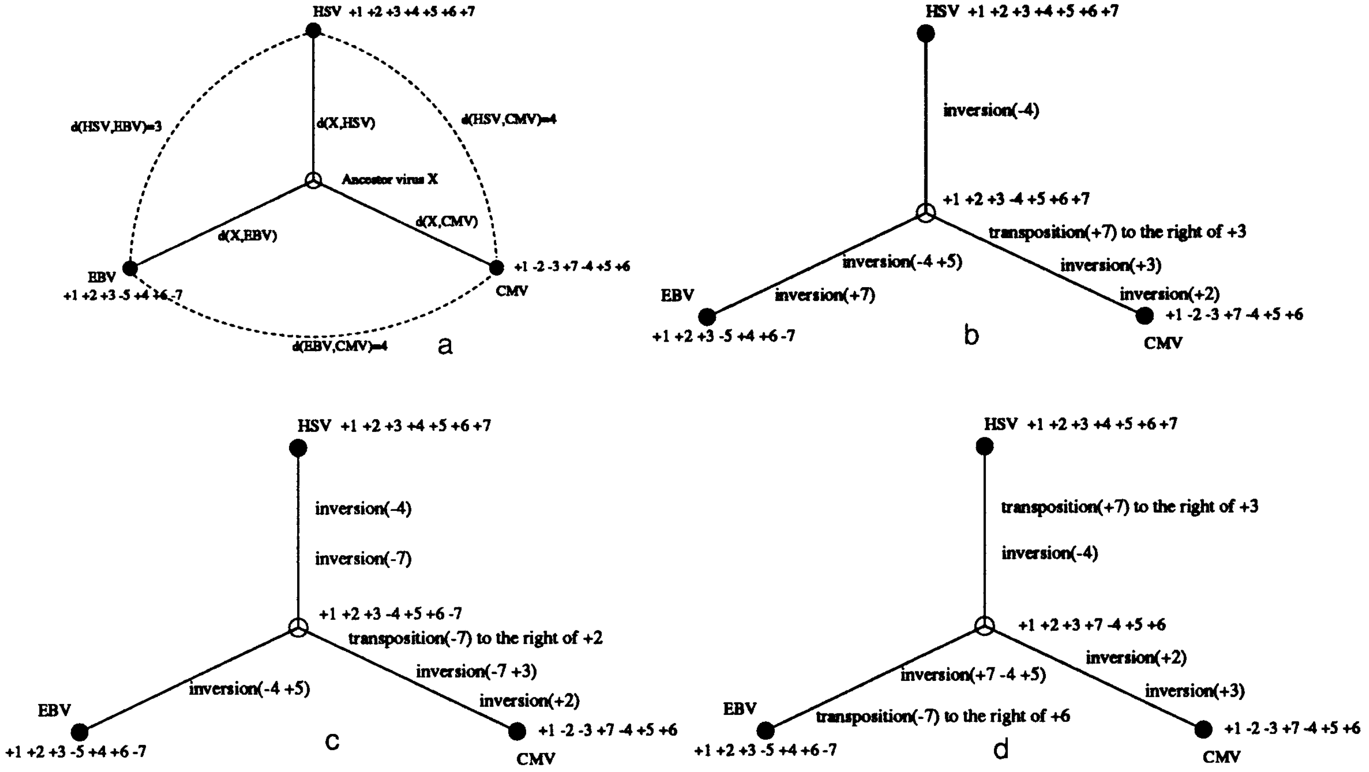

FIG. 5. Most parsimonious evolutionary scenarios for herpesviruses under the "inversions and transpositions" model. (**a**) Rearrangement distances. (**b–d**) The three possible scenarios with six rearrangements each.

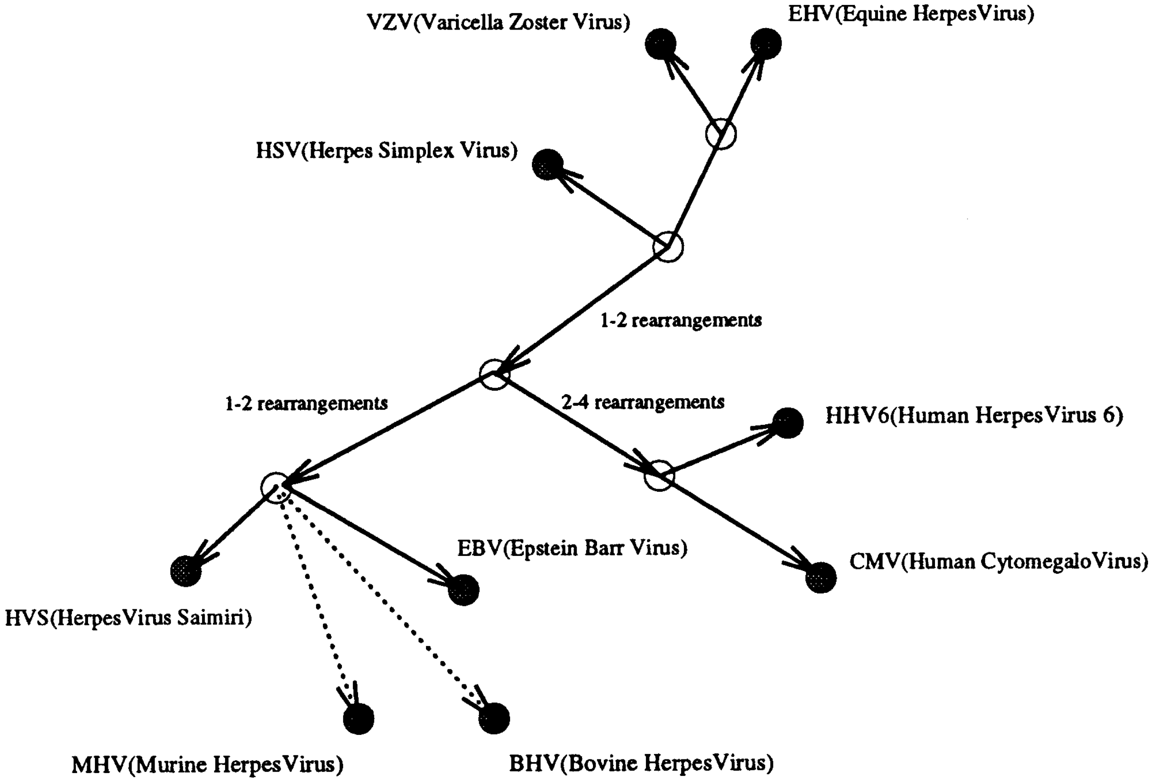

FIG. 6. Tentative phylogenetic tree of herpesviruses based on genome rearrangements. The range of possible rearrangement distances associated with each branch is shown.

Although the above analysis of multiple genome rearrangement was successful for the simple case of herpesviruses, its application for more complex rearrangement scenarios will require new algorithms. In particular, the neighborhood intersection approach may become prohibitively computationally intensive if the rearrangement distances are of the order of 15–20 inversions (see also Sankoff *et al.*, 1995, for an alternative approach to the derivation of the ancestral gene order in the case of three genomes).

Phylogenetic Trees Derived from Gene Comparison and Rearrangement Trees

Based on the most parsimonious evolutionary scenarios constructed, one can derive an unrooted tree depicting the evolution of herpesviruses in terms of genome rearrangements, with the three main branches corresponding to the alpha, beta, and gamma divisions (Fig. 6). However, the order of divergence of these groups remains uncertain. The distance (number of rearrangements) separating the hypothetical common ancestor from each of these groups depends on the chosen evolutionary scenario (compare Figs. 4 and 5), and the genome organization of the ancestor herpesvirus could not be reconstructed unambiguously. Elucidation of the complete genome structure of additional herpesviruses will help to solve this problem by eliminating some of the evolutionary scenarios. A detailed analysis of the arrangement of about 40 genes that are conserved among alphaherpesviruses showed that VZV and EHV are more closely related to one another than each of them is to HSV (data not shown). Note that information about gene order in each additional genome imposes restrictions on the number of consistent evolutionary scenarios. For example, there are 149 most parsimonious scenarios with five inversions for CMV and EBV. However, only 28 of these are consistent with one of the most parsimonious scenarios, with seven inversions each, for the three genomes CMV, EBV, and HSV.

Qualitatively, the conclusions from the genome rearrangement study are compatible with those suggested by the phylogenetic analysis of individual protein sequences. Figure 7 shows tentative phylogenetic trees constructed for three highly conserved herpesvirus proteins using the neighbor-joining method (Saitou and Nei, 1987). The homologous sequences from CCV were used as the outgroup to root the trees. Conspicuously, these trees display all the three possible topologies of the alpha, beta, and gamma branches, in accord with the notion that the three main divisions of the herpesviruses could have diverged from the common ancestor at roughly the same time (McGeoch, 1992). Thus the rearrangement trees and the trees based on sequence comparison involve more or less the same level of uncertainty in the way they describe the evolution of herpesviruses. An advantage of the rearrangement analysis is that it suggests possible evolu-

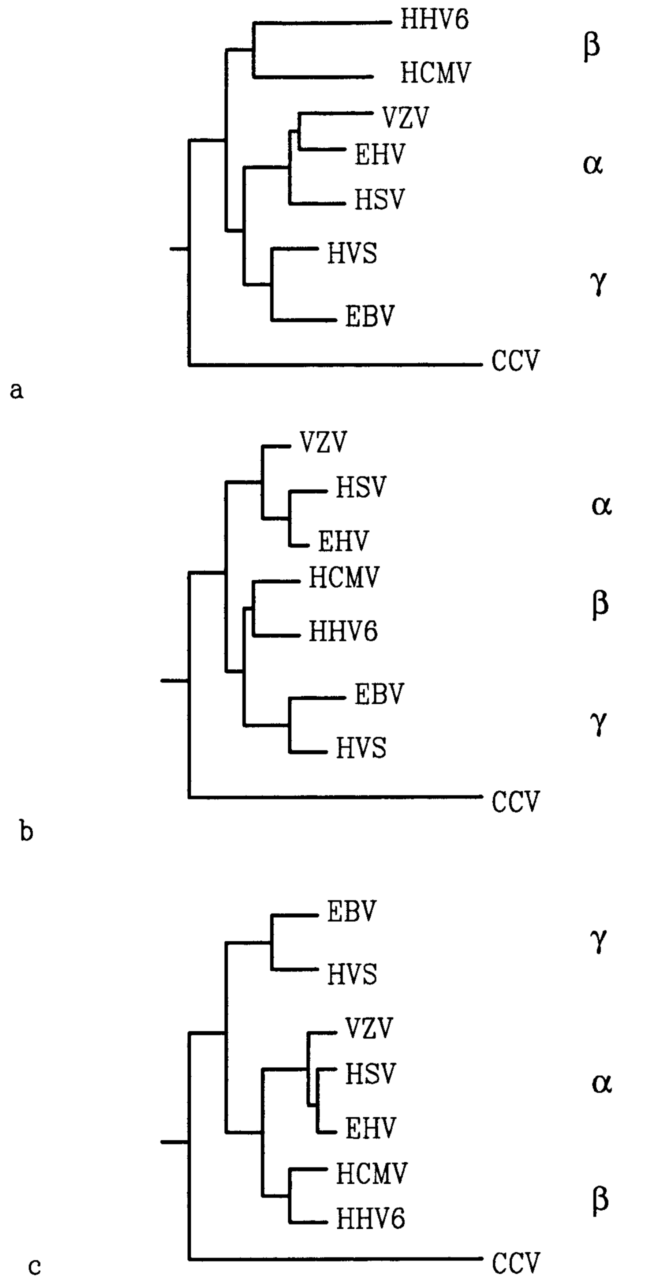

FIG. 7. Tentative phylogenetic trees derived by comparison of the amino acid sequences of conserved herpesvirus proteins. Trees were constructed using the neighbor-joining method for concatenated conserved alignment blocks of the DNA-dependent DNA polymerases (**a**); helicases (**b**); and terminases (**c**). The trees were rooted using the respective sequences of CCV as the outgroup.

tionary scenarios at the level of specific genes and gene blocks and defines a hypothetical ancestral genome organization.

CONCLUDING REMARKS

We described here a strategy for genome sequence comparison that allows one to derive genome rearrangement scenarios. For the chosen test case, which included seven complete and three partial sequences of herpesvirus genomes, in spite of the remarkable variability of the size, base composition, gene repertoire, and nucleotide sequences, the evolutionary radiation could be represented as a simple series of rearrangements of the seven conserved gene blocks. This is certainly not the complete picture of evolution, as that should include acquisition of all the unique genes and their subsequent rearrangements. Moreover, it is possible that the ancestor virus might have had genes that were essential for its replication but that have been subsequently deleted and complemented by the respective cellular functions in some of the viruses. Also, it has to be kept in mind that we were seeking the most parsimonious scenarios and that by the very nature of the maximum parsimony principle, the reconstruction of the "true" history cannot be guaranteed (Swofford and Olsen, 1990). These limitations notwithstanding, the rearrangement scenarios, even though they do not immediately offer new biological insights, seem to provide a useful framework to describe the evolution of herpesviruses.

The herpesvirus case is relatively simple. In fact, it is possible to perform an exhaustive analysis of rearrangements among the three types of organization of the seven conserved gene blocks and to verify directly that the versions we derived analytically were the most parsimonious ones. Clearly, the principal application of the described approach is in comparative analysis of more complex genomes, e.g., those of related bacteria, for which the exhaustive search will not be feasible. An important goal for further analytical work is a generalization of the genome rearrangement algorithms to include deletions and insertions.

ACKNOWLEDGMENTS

We are grateful to Drs. V. Bafna and W. Miller for providing their unpublished code and useful discussions and to Dr. J. C. Wootton for critical reading of an early version of the manuscript. The work of S.H. and P.A.P. was supported, in part, by the NSF Young Investigator Award, NSF Grant CCR-9308567, and NIH Grant 1R01 HG00987-01.

REFERENCES

Albrecht, J.-C., Nicholas, J., Biller, D., Cameron, K. R., Biesinger, B., Newman, C., Wittmann, S., Craxton, M. A., Coleman, H., Fleckenstein, B., and Honess, R. W. (1992). Primary structure of the herpesvirus saimiri genome. *J. Virol.* **66:** 5047–5058.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Altschul, S. F., Boguski, M. S., Gish, W., and Wootton, J. C. (1994). Issues in searching molecular sequence databases. *Nature Genet.* **6:** 119–129.

Atlan, A., and Couvet, D. (1993). A model simulating the dynamics of plant mitochondrial genomes. *Genetics* **135:** 213–222.

Baer, R., Bankier, A. T., Biggin, M. D., Deininger, P. L., Farrel, P. J., Gibson, T. J., Hatfull, G., Hudson, G. S., Satchwell, S. C., Seguin, C., Tuffnell, P. S., and Barrell, B. G. (1984). DNA sequence and expression of the B95-8 Epstein–Barr virus genome. *Nature* **310:** 207–211.

Bafna, V., and Pevzner, P. A. (1993). Genome rearrangements and sorting by reversals. *In* "34th IEEE Symposium on Foundation of Computer Science," pp. 148–157, IEEE Computer Society Press.

Bafna, V., and Pevzner, P. A. (1995a). Sorting by reversals: Genome rearrangements in plant organelles and evolutionary history of X chromosome. *Mol. Biol. Evol.* **12:** 239–246.

Bafna, V., and Pevzner, P. A. (1995b). Sorting by transpositions. *In* "Proceedings of 6th Annual ACM-SIAM Symposium on Discrete Algorithms," pp. 614–623, Assoc. Comput. Mach.

Bork, P., Ouzounis, C., and Sander, C. (1994). From genome sequence to protein functions. *Curr. Opin. Struct. Biol.* **4:** 393–403.

Bruns, T. D., and Palmer, J. D. (1989). Evolution of mushroom mitochondrial DNA: *Suillus* and related genera. *J. Mol. Evol.* **28,** 349–362.

Bublot, M., Lomonte, P., Lequarre, A., Albrecht, J., Nicholas, J., Fleckenstein, B., Pastoret, P., and Thiry, E. (1992). Genetic relationships between bovine herpesvirus 4 and the gammaherpesviruses Epstein–Barr virus and herpesvirus saimiri. *Virology* **190:** 654–665.

Casjens, S., Hatfull, G., and Hendrix, R. (1992). Evolution of dsDNA tailed-bacteriophage genomes. *Semin. Virol.* **3:** 383–397.

Cedergren, R., Sankoff, D., and Abel, Y. (1990). Relationships from gene sequences. *Nature* **345:** 484.

Chee, M. S., Bankier, A. T., Beck, S., Bohni, R., Brown, C. M., Cerny, R., Horsnell, T., Hutchison III, C. A., Kouzarides, T., Martignetti, J. A., *et al.* (1990). Analysis of the protein-coding content of the sequence of human cytomegalovirus strain AD169. *Curr. Top. Microbiol. Immunol.* **154:** 125–169.

Davison, A. J. (1992). Channel catfish virus: A new type of herpesvirus. *Virology* **186:** 9–14.

Davison, A. J., and Scott, J. E. (1986). The complete DNA sequence of varicella-zoster virus. *J. Gen. Virol.* **67:** 1759–1816.

Davison, A. J., and Taylor, P. (1987). Genetic relations between varicella-zoster virus and Epstein–Barr virus. *J. Gen. Virol.* **68:** 1067–1079.

Dobzhansky, T., and Sturtevant, A. H. (1938). Inversions in the chromosomes of *Drosophila pseudoobscura*. *Genetics* **23:** 28–64.

Efstathiou, S., Ho, Y. M., Hall, S., Styles, C. J., Scott, S. D., and Gompels, U. (1990). Murine herpesvirus 68 is genetically related to the gammaherpesviruses Epstein–Barr virus and herpesvirus saimiri. *J. Gen. Virol.* **71:** 1365–1372.

Felsenstein, J. (1989). PHYLIP—Phylogeny inference package (Version 3.2). *Cladistics* **5:** 164–166.

Fraefel, C., Wirth, U. V., Vogt, B., and Schwyzer, M. (1993). Immediate-early transcription over covalently joined genome ends of bovine herpesvirus 1: The circ gene. *J. Virol.* **67:** 1328–1333.

Gompels, U. A., Nicholas, J., Lawrence, G., Jones, M., Thomson, B. J., Martin, M. E., Efstathiou, S., Craxton, M., and Macaulay, H. A. (1995). The DNA sequence of human herpesvirus-6: Structure, coding content, and genome evolution. *Virology* **209:** 29–51.

Hammerschmidt, W., Ludwig, H., and Buhk, H.-J. (1988). Specificity of cleavage in replicative-form DNA of bovine herpesvirus 1. *J. Virol.* **62:** 1355–1362.

Hannenhalli, S., and Pevzner, P. A. (1995). Transforming Cabbage into Turnip (polynomial algorithm for sorting signed permutations by reversals). *In* "Proceedings of 27th Annual ACM Symposium on the Theory of Computing," pp. 178–189, Assoc. Comput. Mach.

Hoffmann, R. J., Boore, J. L., and Brown, W. M. (1992). A novel mitochondrial genome organization for the blue mussel, *Mytilus edulis*. *Genetics* **131:** 397–412.

Hoot, S. B., and Palmer, J. D. (1994). Structural rearrangements including parallel inversions within the chloroplast genome of Anemone and related genera. *J. Mol. Evol.* **38:** 274–281.

Hull, R. (1992). Genome organization of retroviruses and retroelements: Evolutionary considerations and implications. *Semin. Virol.* **3:** 373–382.

Karlin, S., Mocarski, E. S., and Schachtel, G. A. (1994). Molecular evolution of herpesviruses: Genomic and protein sequence comparisons. *J. Virol.* **68:** 1886–1902.

Kececioglu, J., and Ravi, R. (1995). Of mice and men: Evolutionary distances between genomes under translocations. *In* "Proceedings 6th Annual ACM-SIAM Symposium on Discrete Algorithms," pp. 604–613, Assoc. Comput. Mach.

Kececioglu, J., and Sankoff, D. (1993). Exact and approximation algorithms for the inversion distance between two permutations. *In* "Proceedings 4th Annual Symposium on Combinatorial Pattern Matching," pp. 87–105, Springer–Verlag, Berlin.

Kececioglu, J., and Sankoff, D. (1994). Efficient Bounds for Oriented Chromosome Inversion Distance. *In* "Proceedings 5th Annual Symposium on Combinatorial Pattern Matching," pp. 307–325, Springer–Verlag, Berlin.

Kececioglu, J., and Sankoff, D. (1995). Exact and approximation algorithms for the inversion distance between two permutations with applications to genome rearrangements. *Algorithmica* **13:** 187–210.

Knox, E. B., Downie, S. R., and Palmer, J. D. (1993). Chloroplast genome rearrangements and evolution of giant lobelias from herbaceous ancestors. *Mol. Biol. Evol.* **10:** 414–430.

Koonin, E. V., and Dolja, V. V. (1993). Evolution and taxonomy of positive-strand RNA viruses: Implications of comparative analysis of amino acid sequences. *Crit. Rev. Biochem. Mol. Biol.* **28:** 375–430.

Lawrence, G. L., Chee, M., Craxton, M. A., Gompels, U. A., Honess, R. W., and Barrell, B. G. (1990). Human herpesvirus 6 is closely related to human cytomegalovirus. *J. Virol.* **64:** 287–299.

McGeoch, D. J. (1989). The genomes of the human herpesviruses: Contents, relationships, and evolution. *Annu. Rev. Microbiol.* **43:** 235–265.

McGeoch, D. J. (1992). Molecular evolution of large DNA viruses of eukaryotes. *Semin. Virol.* **3:** 399–408.

McGeoch, D. J., Barnett, B. C., and MacLean, C. A. (1993). Emerging functions of alphaherpesvirus genes. *Semin. Virol.* **4:** 125–134.

McGeoch, D. J., and Cook, S. (1994). Molecular phylogeny of the alphaherpesvirinae subfamily and a proposed evolutionary timescale. *J. Mol. Biol.* **238:** 9–22.

McGeoch, D. J., Dalrymple, M. A., Davison, A. J., Dolan, A., Frame, M. C., McNab, D., Perry, L. J., Scott, J. E., and Taylor, P. (1988). The complete DNA sequence of the long unique region in the genome of herpes simplex virus type 1. *J. Gen. Virol.* **69:** 1531–1574.

Milligan, B. G., Hampton, J. N., and Palmer, J. D. (1989). Dispersed repeats and structural reorganization in subclover chloroplast DNA. *Mol. Biol. Evol.* **6:** 355–368.

Nadeau, J. H., and Taylor, B. A. (1984). Lengths of chromosomal segments conserved since divergence of man and mouse. *Proc. Natl. Acad. Sci. USA* **81:** 814–818.

Nadeau, J. H., Davisson, M. T., Doolittle, D. P., Grant, P., Hillyard, A. L., Kosowsky, M. R., and Roderick, T. H. (1992). Comparative map for mice and humans. *Mamm. Genome* **3:** 480–536.

Neipel, F., Ellinger, K., and Fleckenstein, B. (1991). The unique region of the human herpesvirus 6 genome is essentially collinear with the UL segment of human cytomegalovirus. *J. Gen. Virol.* **72:** 2293–2297.

O'Brien, S. J. (1991). Mammalian genome mapping: Lessons and prospects. *Curr. Opin. Genet. Dev.* **1:** 105–111.

Palmer, J. D. (1992). Mitochondrial DNA in plant systematics: Applications and limitations. *In* "Molecular Systematics of Plants" (P. A. Soltis, D. E. Soltis, and J. J. Doyle, Eds.), pp. 36–49, Chapman and Hall, New York, London.

Palmer, J. D., and Herbon, L. A. (1987). Unicircular structure of the *Brassica hirta* mitochondrial genome. *Curr. Genet.* **11:** 565–570.

Palmer, J. D., and Herbon, L. A. (1988). Plant mitochondrial DNA evolves rapidly in structure, but slowly in sequence. *J. Mol. Evol.* **28:** 87–97.

Poffenberger, K. L., and Roizman, B. (1985). A noninverting genome of a viable herpes simplex virus 1: Presence of head-to-tail linkages in packaged genomes and requirements for circularization after infection. *J. Virol.* **53:** 587–595.

Saitou, N., and Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4:** 406–425.

Sankoff, D. (1993). Analytical approaches to genomic evolution. *Biochimie* **75:** 409–413.

Sankoff, D., Leduc, G., Antoine, N., Paquin, B., Lang, B. F., and Cedergren, R. (1992). Gene order comparisons for phylogenetic inference: Evolution of the mitochondrial genome. *Proc. Natl. Acad. Sci. USA* **89:** 6575–6579.

Sankoff, D., Sundaram, G., and Kececioglu, J. (1995). Steiner points in the space of genome rearrangements. *In* "Technical Report 2203," Centre de Recherches Mathematiques, University of Montreal.

Schachtel, G. A., Bucher, P., Mocarski, E. S., Blaisdell, B. E., and Karlin, S. (1991). Evidence for selective evolution in codon usage in conserved amino acid segments of human alphaherpesvirus proteins. *J. Mol. Evol.* **33:** 483–494.

Schuler, G. D., Altschul, S. F., and Lipman, D. J. (1991). A workbench for multiple alignment construction and analysis. *Proteins Struct. Funct. Genet.* **9:** 180–190.

Sturtevant, A. H., and Dobzhansky, T. (1936). Inversions in the third chromosome of wild races of Drosophila pseudoobscura, and their use in the study of the history of the species. *Proc. Natl. Acad. Sci. USA* **22:** 448–450.

Swofford, D. L., and Olsen, G. J. (1990). Phylogeny reconstruction. *In* "Molecular Systematics" (D. M. Hillis and C. Moritz, Eds.), pp. 411–501, Sinauer, Sanderland, MA.

Telford, E., Watson, M., McBride, K., and Davison, A. (1992). The DNA sequence of equine herpesvirus-1. *Virology* **189:** 304–316.

Watterson, G. A., Ewens, W. J., Hall, T. E., and Morgan, A. (1982). The chromosome inversion problem. *J. Theor. Biol.* **99:** 1–7.

Zakharov, I. A., Nikiforov, V. S., and Stepaniuk, E. V. (1992). Homology and evolution of gene orders: Combinatorial measure of synteny group similarity and simulation of the evolution process. *Genetika* **28:** 77–81.