JOURNAL OF VIROLOGY, Apr. 2003, p. 4326–4344 Vol. 77, No. 7
0022-538X/03/$08.00+0 DOI: 10.1128/JVI.77.7.4326–4344.2003

In Silico Pattern-Based Analysis of the Human Cytomegalovirus Genome

Isidore Rigoutsos,[1]* Jiri Novotny,[2] Tien Huynh,[1] Stephen T. Chin-Bow,[1] Laxmi Parida,[1] Daniel Platt,[1] David Coleman,[3] and Thomas Shenk[3]

Bioinformatics and Pattern Discovery Group, IBM TJ Watson Research Center, Yorktown Heights, New York 10598[1]; Victor Chang Cardiac Research Institute, Darlinghurst, New South Wales 2010, Australia[2]; and Department of Molecular Biology, Princeton University, Princeton, New Jersey 08544[3]

Received 10 July 2002/Accepted 23 December 2002

More than 200 open reading frames (ORFs) from the human cytomegalovirus genome have been reported as potentially coding for proteins. We have used two pattern-based in silico approaches to analyze this set of putative viral genes. With the help of an objective annotation method that is based on the Bio-Dictionary, a comprehensive collection of amino acid patterns that describes the currently known natural sequence space of proteins, we have reannotated all of the previously reported putative genes of the human cytomegalovirus. Also, with the help of MUSCA, a pattern-based multiple sequence alignment algorithm, we have reexamined the original human cytomegalovirus gene family definitions. Our analysis of the genome shows that many of the coded proteins comprise amino acid combinations that are unique to either the human cytomegalovirus or the larger group of herpesviruses. We have confirmed that a surprisingly large portion of the analyzed ORFs encode membrane proteins, and we have discovered a significant number of previously uncharacterized proteins that are predicted to be G-protein-coupled receptor homologues. The analysis also indicates that many of the encoded proteins undergo posttranslational modifications such as hydroxylation, phosphorylation, and glycosylation. ORFs encoding proteins with similar functional behavior appear in neighboring regions of the human cytomegalovirus genome. All of the results of the present study can be found and interactively explored online (http://cbcsrv.watson.ibm.com/virus/).

The advent of DNA sequencing technology is generating vast amounts of sequences that are deposited in public databases. The rate at which genomes can be sequenced has now outpaced the rate at which a sequence's function can be determined through wet-lab experimentation, thus leading to increasing demand for automated (in silico) approaches to the elucidation of protein function. As more and more protein sequences and complete genomes become available in the public domain, in silico protein annotation is emerging as an inexpensive and effective approach for dealing with the flood of genomic data.

Of the numerous approaches that have been proposed over the years, the determination of regions of similarity between a novel protein of unknown function and one or more database proteins with known annotation has been the method of choice. Such a determination allows one to predict the common region in the protein of unknown function as exhibiting the functional characteristics of the respective region from the annotated database protein through what is frequently called a "guilty-by-association" approach. These methods are also known as homology-based methods, and they have led to significant advances in protein annotation (2, 22, 36).

During the latter half of the 1990s, pattern-based approaches have been steadily gaining ground as the methods of choice for solving various computational problems in molecular biology (28). One such algorithm is MUSCA, a multiple sequence alignment algorithm, which we described in an earlier study (23). MUSCA begins by using the Teiresias pattern discovery algorithm (25, 26) to identify patterns that are shared by k or more input sequences. During its second phase, MUSCA exploits the location of the discovered patterns to anchor and induce alignments of increasingly larger input fragments. Because of the manner in which it operates, MUSCA is uniquely suitable to handle inputs in which one or more domains are shared among the sequences to process. In a parallel effort, we also described a pattern-based approach to the problem of protein annotation (30). The approach is centered on the Bio-Dictionary, an exhaustive collection of amino acid patterns (heretofore referred to as seqlets) that completely covers the natural sequence space of proteins defined by the currently available sequences. The Bio-Dictionary is computed by carrying out pattern discovery with the Teiresias algorithm (25, 26) on very large databases of biological sequences such as SwissProt/TrEMBL (4). The seqlets contained in the Bio-Dictionary can capture functional and structural signals that have been reused during evolution both within and across families of related proteins (27, 28, 29). This new method uses the seqlets contained in the Bio-Dictionary to exhaustively annotate a query protein by using the information that is available in a well-maintained database, such as SwissProt/TrEMBL, and employs a weighted, position-specific scoring scheme that is not affected by the overrepresentation of well-conserved proteins and protein fragments, which exist in the public databases. As we showed elsewhere (30) and for several published genomes, this Bio-Dictionary-based approach matched

* Corresponding author. Mailing address: Bioinformatics and Pattern Discovery Group, IBM TJ Watson Research Center, PO Box 218, Yorktown Heights, NY 10598. Phone: (914) 945-1384. Fax: (914) 945-4104. E-mail: rigoutso@us.ibm.com.

the quality and sensitivity of the annotations that were obtained with semiautomated approaches while requiring only a very small investment of computational resources.

In an earlier study (21), we examined the annotation of human cytomegalovirus (HCMV; also known as human herpesvirus 5 [HHV-5]) by using ProCeryon (a program for fold recognition and protein structure analysis; ProCeryon Biosciences), a structure prediction program that is based on threading (16, 34). Each of the HHV-5 ORFs was threaded with the ProCeryon algorithm, and a structural and functional hypothesis was generated. As anticipated and due to the large number of membrane proteins coded for by this genome, the threading approach provided hypotheses for a little less than 50% of the coding regions. The desire to further push the annotation envelope for HHV-5 led us to the sequence-based work that we discuss here.

HHV-5 is a member of the betaherpesviruses, a subgroup of herpesviruses with common growth characteristics (1). Considered the prototypical betaherpesvirus, HHV-5 spreads to the majority of the population at an early age, causing asymptomatic infections in healthy individuals. However, it can produce life-threatening disease in immunosuppressed individuals and as a result of congenital infections (24).

The HHV-5 virion contains a linear, double-stranded DNA genome (~230 kbp) encased in an icosahedral capsid (8). The capsid is surrounded by a protein matrix and a lipid envelope with integral glycoproteins. Two major unique regions, denoted as long (U_L) and short (U_S), can be identified in the viral genome and are bracketed by repeated domains. The AD169 strain of HHV-5 was sequenced in 1990 (8), and 208 ORFs were predicted as coding for proteins ≥100 amino acids in length. More recently, an insertion that modifies ORFs 42 and 43 was identified in the AD169 strain (10, 20), and analysis of a cDNA sequence has revealed that UL101 does not exist, whereas UL102 is modified (35). Finally, several additional ORFs were found in the Towne and Toledo strains of HHV-5 (7). The repeated ORFs include J1L/J1I/J1S, which are partially related; TRL1 through TRL13, which at a second location are labeled IRL1 through IRL13; TRL14, which shares a N-terminal region with its IRL14 counterpart; and IRS1/TRS1, which are half repeated and half unique. The unique ORFs are UL1 to UL154 and US1 to US36, with some ORFs receiving fractional designations such as UL21.5, UL48.5, and UL80.5.

We describe below the application of the Bio-Dictionary to the in silico annotation of the HHV-5 genome. We have generated and processed a "composite" genome that is the union of the originally reported genes from AD169, the three modified ORFs in AD169, as well as the genes from the Towne and Toledo strains. The method we used to annotate the composite genome is described in detail elsewhere (30), and an implementation of it is available online (http://cbcsrv.watson.ibm.com/Tpa.html). The functional hypotheses, as well as numerous features that we identified in these proteins are also available online (http://cbcsrv.watson.ibm.com/virus/); at this website, one can find summaries of each protein's functional annotation and information about the nature and location of posttranslational modifications, active sites, identifiable domains (e.g., transmembrane), and alignments with other proteins from the public databases, as well as detailed information on the similarity of each annotated amino acid sequence to archaeal, bacterial, eukaryotic, and viral sequences. For completeness, we have also included the results from our earlier, threading-based annotation of HHV-5 (21). Finally, we used MUSCA to reanalyze the originally proposed HHV-5 ORF families and we present our conclusions here.

MATERIALS AND METHODS

For the analysis outlined in the present study, we used two computational approaches, both of which have previously been described. We briefly outline each of these two approaches below.

Bio-Dictionary-based automated protein annotation. The computational tool that we used for the annotation component of this work relies on the Bio-Dictionary: the latter was originally created by using the Teiresias pattern discovery algorithm (25, 26) to process the GenPept database as a whole (29); this computation has since been repeated at regular intervals on the increasingly larger installments of the SwissProt/TrEMBL database (4). The Bio-Dictionary is a very large collection of sequence patterns, referred to as seqlets. In other words, seqlets are strings of literals interspersed with zero or more wild cards: a unique amino acid, or a small set of permitted amino acids, can occupy the locations of each literal; the positions corresponding to the wild cards indicate locations that can be occupied by any amino acid. For example, the seqlet [KR].K[ILMV][AG]L describes all hexapeptides that begin with either a lysine or an arginine; followed by any one of the 20 amino acids; followed by a lysine; followed by a isoleucine, leucine, methionine, or valine; followed by alanine or glycine; and finally ending with a leucine. The Bio-Dictionary seqlets capture functional and structural signals that extend beyond protein family boundaries, which is not an unexpected result considering the manner in which the collection is produced.

An additional property of this collection is that it nearly completely covers the currently known sequence space of natural proteins and can thus be used in lieu of the original processed sequence database to solve a gamut of problems, including gene finding (33) and protein annotation (30). For the purposes of protein annotation, each seqlet is augmented with additional information pertaining to functional, structural, or other properties of the seqlet's known instances in proteins that have been studied computationally and experimentally.

To annotate a previously uncharacterized protein, instances of all of the seqlets in the Bio-Dictionary are sought in the sequence under consideration: for seqlets that are present in the sequence, their respective meanings are used to label the part of the sequence corresponding to the seqlet's instance in a straightforward "guilty-by-association" approach. The meanings of overlapping seqlets are subsequently accumulated and coalesced into hypotheses about the function of the processed protein, the presence of various domains and active sites, the nature and location of posttranslational modifications, etc. Details on the computational aspects of the Bio-Dictionary-based protein annotation are given elsewhere by Rigoutsos et al. (30).

MUSCA (pattern-based multiple sequence alignment). MUSCA is a two-phase algorithm for computing the multiple sequence alignment of a set of N sequences (23). During the first phase, MUSCA uses Teiresias to discover patterns that are common among K or more of the input's N sequences. These patterns are used in the second phase to generate and report the multiple sequence alignment. In particular, the motifs are first mapped to vertices of a directed graph. If the two motifs p_i and p_j do not occur simultaneously in any sequence, then there is no edge connecting the corresponding vertices of the graph. The vertices corresponding to p_i and p_j will be connected by an edge with direction from p_i to p_j if p_i occurs before p_j in all of the sequences where they both appear. The labels of the edges depend on three things: whether p_i and p_j are pairwise incompatible, whether they have overlapping instances, or whether they are pairwise compatible but do not overlap. Vertices that are joined by incompatible edges or participate in inconsistent cycles form the basic nonfeasible sets. After the vertices of the reduced graph were labeled with the help of a simple cost function, we used a greedy algorithm to obtain a solution to a weighted set-cover problem that essentially identifies the minimum number of motifs/vertices to be removed. The resulting graph was used to determine the blocks that involve overlapping feasible motifs. We obtained the final alignment by properly aligning the blocks and padding up the existing gaps.

The alignments that MUSCA generates are independent of the order in which the input sequences are given. The algorithm is uniquely suitable to process inputs where the various sequences share domains that are present in some of the sequences only or share inputs that comprise two or more subsets with high conservation within each subset but low conservation across subsets.

RESULTS

In this section, we present a summary of the results that we obtained from processing the composite HHV-5 genome. Additional information and details for each individual sequence can be found at http://cbcsrv.watson.ibm.com/virus/. From this site, the user has the option of using either a graphical or a textual interface for accessing the annotation that is available for each annotated ORF. In each case, we generated a file with statements that succinctly describe our findings. Corroborating local and/or global alignments are given when relevant and/or available; also given for each annotated ORF are plots that show in graphical form the nature and location of phylogenetic-domain-specific fragments, discovered local or global similarities, domains of interest, sites of interest, etc. Only a small subset of the information that is available through the website is included and discussed below.

Benchmarking our approach. In a recent study (30), we discussed and tested the Bio-Dictionary-based protein annotation method in detail and with the help of many diverse input sequences. We demonstrated that this method provides substantial benefits versus traditional approaches in terms of objective annotation capability, and the reader is referred to that discussion for more information. It is important to stress that our approach makes use of a weighted, position-specific scoring scheme that is not affected by the overrepresentation of well-conserved proteins and protein fragments that exist in the public databases. A given feature that has been associated with a region of the query is assigned a normalized score between 0 and 100; the resulting figure is an estimate of the percentage of the total number of the region's distinct instances that have also been annotated as sharing the feature. Empirically, we determined that scores between 90 and 100 correspond to good, conservative results: for the HHV-5 annotation presented below, we only considered features whose confidence estimates fell in this interval of values.

General findings. Table 1 summarizes our findings for each of the annotated ORFs of the composite genome. The ORFs are listed in the order in which they appear on the composite HHV-5 genome. In the first three columns, and for each ORF, we list the ORF's name, its accession number, and a functional hypothesis. The fourth column lists features of each predicted protein such as binding sites, posttranslationally modified sites, etc.

Several general observations can be made readily. As pointed out by Chee et al. (8), the importance of glycoproteins as surface antigens has generated interest in the identification and characterization of the members of this group. Chee et al. (8) predicted the presence of one or more glycosylation sites in proteins encoded by a total of 53 ORFs. Mokarski and Courcelle (19) revised this number to ca. 60. As seen in Table 1, our analysis predicts the presence of glycosylation sites in 125 of the annotated ORFs. Four of these ORFs—TRL12, IRL12, UL32, and UL132—contain O-linked glycosylation sites, and the remainder contain N-linked modifications. Of the 125 ORFs predicted to encode glycosylated proteins, the following have 10 or more potential glycosylation sites: TRL12/IRL12 (7 O-type and 15 N-type), UL1 (10 N-type, possibly 13), UL7 (11 N-type), UL18 (13 N-type), UL20 (12 N-type), UL37 (18 N-type), UL55 (19 N-type), UL74 (18 N-type), UL116 (14 N-type), and UL120 (10 N-type). UL32 has been proposed to contain a single O-type glycosylation site (19); however, our analysis indicates the presence of two such sites, at amino acid locations 921 and 952, respectively.

Our analysis predicts that as many as 144 proteins, approximately half of them glycoproteins, are probably integral membrane proteins. Also, for at least 49 of the analyzed proteins we find evidence for the presence of a signal peptide. These proteins are TRL2/IRL2, TRL10/IRL10, TRL11/IRL11, UL4, UL11, UL12, UL13, UL14, UL16, UL18, UL20, UL21, UL21.5, UL31, UL37, UL40, UL41, UL50, UL55, UL56, UL73, UL75, UL91, UL111.5, UL115, UL117, UL118, UL119, UL121, UL124, UL130, UL132, UL132/Toledo, UL139/Toledo, UL144/Toledo, UL147/Toledo, UL148/Toledo, UL149/Toledo, UL152/Towne, US3, US6, US7, US8, US9, US10, US11, US25, US30, and US31. These proteins would be expected to represent a collection of plasma membrane proteins, proteins that reside within intracellular membranous compartments, and secreted proteins.

We found evidence for one or more phosphorylation sites for the proteins encoded by the following nine ORFs: TRL5/IRL5, UL4, UL34, UL59, UL67, UL83, UL109, IRS1, and US36. Also, at least 17 of the coded proteins contain one or more hydroxylation sites: J1L, TRL9/IRL9, UL15, UL31, UL44, UL52, UL57, UL61, UL62, UL104, UL141/Toledo, UL150/Toledo, IRS1, US8, US32, TRS1, and J1S. Of these, UL61 has the largest number (i.e., nine) of such sites.

Thirty-one ORFs seem to be virus specific in the sense that they contain no notable, distinguishing features and no identifiable similarities to anything except for other viral sequences. In particular, at least 15 ORFs (J1L, TRL8/IRL8, UL60, UL62, UL64, UL66, UL81, UL90, UL106, UL110, UL137/Toledo, UL145/Toledo, UL148/Toledo, UL149/Toledo, US5, US33, and J1S) appear to be specific to the HCMV, at least 7 ORFs (UL25, UL29, UL36, UL47, UL88, UL91, and UL96) are specific to the herpesviruses, and at least 3 ORFs (UL53, UL85, and UL115) are virus specific, i.e., homologies can be found outside the herpesvirus family. The presence of many ORFs that are specific to HHV-5 but not other organisms is consistent with the view that HHV-5 is an ancient virus whose genome has evolved separately from its host for an extended period.

G-protein-coupled receptor homologues. Our analysis revealed, among the annotated ORFs, the presence of a group of 15 sequences that are likely to code for GPCR-like proteins. In particular, six ORFs, namely, UL33, UL78, US12, US14, US27, and US28, show strong homologies to members from well-understood GPCR families.

For an additional nine ORFs, our analysis clearly shows the presence of exactly seven transmembrane regions, thus implying a membership in the GPCR family: these nine ORFs are UL100, US13, US15, US16, US17, US18, US19, US20, and US21. It is notable that a substantial part of the US region contains ORFs that appear to be coding for proteins that have a GPCR-like sequence composition and GPCR-like characteristics. An interesting observation can be made with respect to the nine ORFs that our analysis places in this category: although there is support that each of these ORFs contains seven transmembrane regions, none of these ORFs show any notable global sequence similarity to members of known GPCR fami-

TABLE 1. List of brief annotations for the composite genome of HCMV[a]

ORF label	Accession no.	Protein class or functional hypothesis	Features
J1L	125058	HCMV-specific protein	Region 205-265 is eukaryote specific; remaining regions are similar to bacterial and eukaryotic sequences; weak support for hydroxylation site at locations 294, 295, 300, 301, 307, and 308; weak support for ATP-binding site at location 238; local similarity with Q9VU12 from *Drosophila melanogaster;* homology with J1S_HCMVA
TRL1/IRL1	124876	HCMV-specific transmembrane gp	Regions 32-40 and 300-310 are virus specific; remaining regions are shared among bacterial and eukaryotic sequences; N-glycosylation site at location 7; probable disulfide bridge involving the cysteines at locations 135 and 137; probable active site (of unknown nature) at location 180; support for transmembrane domain in regions 11-27, 65-85, 95-105, 140-168, and 183-195; similarity with the polycystin precursor (PKD1_HUMAN) from *Homo sapiens;* similarity with *trans*-acting transcriptional protein ICP4_HSV11
TRL2/IRL2	124877	HCMV-specific transmembrane gp	Support for a probable signal sequence in region 66-86; region 55-61 is virus specific (the remainder of the sequence is similar to eukaryotic sequences); N-glycosylation site at location 45; support for transmembrane domain in regions 1-33 and 100-115; similarity with a region from rat and fugu Huntingtin; similarity with the N terminus of gB precursor VGLB_HSV11; similarity with the human period circadian protein 1; similarity with a region from the middle of human and horse inhibin alpha precursor (IHA_HUMAN/IHA_HORSE)
TRL3/IRL3	124878	HCMV-specific transmembrane protein	Support for transmembrane domain in regions 40-60 and 77-end; similarity to an NADH-ubiquinone oxidoreductase chain 6 (NU6M_ASCSU); short region shared between this sequence and ATP-dependent helicase LHR_ECOLI
TRL4/IRL4	137816	Transmembrane gp	Support for transmembrane domain in regions 5-79 and 149-163; proline-rich in region 10-41; cytoplasmic domain in region 10-41; N-glycosylation site at locations 135 and 141; similarity with the human fas antigen ligand FASL_HUMAN
TRL5/IRL5	124879	Inconclusive	Region 45-57 is virus specific; probable phosphorylation site at location 23; C-terminus similarity with flavoprotein subunit A (PDB: 1efv); similarity with histone 1 (H1_CHITH); similarity with the mouse period circadian protein 2; similarity to a region of the DNA polymerase from the African swine fever virus
TRL6/IRL6	124880	HCMV-specific transmembrane gp	Region 50-65 is virus specific; N-glycosylation sites at locations 50, 56, and 62; probable transmembrane domain in region 50-94; similarity with the middle part of NADH-ubiquinone oxidoreductase chain 4 (NU4M_BALPH)
TRL7/IRL7	124881	CMV specific	Similarity with a region of gi_202456, a zinc finger protein (Zfx) from *Mus musculus;* homologous to 30-kDa major early protein from strain Eisenhardt
TRL8/IRL8	124882	HCMV specific	Serine-rich in region 85-115; glycine-rich in region 1-23
TRL9/IRL9	124883	Inconclusive	Probable hydroxylation site at locations 52, 62, and 65; probable triple-helical region in 53-72; probable coiled coil in region 85-118; similar to sperm protamine p1 (HSP1_ONRAN)
TRL10/IRL10	124884	HCMV-specific transmembrane gp	Support for a signal sequence in region 1-25; N-glycosylation site at locations 48, 49, 56, and 108; support for transmembrane domain in regions 80-101; homologous to TRL10; similarity to the C terminus of NADH-ubiquinone oxidoreductase chain 6 (NU6M_BRABR) from *Brachyramphus brevirostris* (Kittlitz murrelet)
TRL11/IRL11	124885	HCMV-specific transmembrane gp	Support for a signal sequence in region 5-20; support for N-glycosylation site at locations 56 and 110; support for transmembrane domain in regions 95-135 and 190-203; homologous to UL153 from the Towne strain
TRL12/IRL12	124886	HCMV-specific transmembrane gp	O-glycosylation sites at locations 30, 33, 35, 36, 38, 43, and 44; N-glycosylation sites at locations 32, 63, 80, 83, 127, 131, 137, 198, 206, 212, 276, 286, 320, 326, and 351; support for transmembrane domain in region 370-390; weak support for an active site of unknown nature at location 323; possibly homologous to UL154 from the Towne strain
TRL13/IRL13	124887	HCMV-specific transmembrane gp	N-glycosylation site at locations 35, 54, 89, 99, and 144; serine/threonine-rich in region 25-110; support for transmembrane domain in regions 5-18 and 125-145; good agreement with Q68408 also known as ORF UL153 from the Towne strain
TRL14	136113	HCMV-specific transmembrane gp	N-glycosylation site at locations 24, 64, and 72; support for transmembrane domain in region 136-165; local similarity with biliary gp 1 precursor (BGP1_HUMAN); shares its N-terminal region with IR14 from the same virus; homologous with UL153 from the same virus
UL1	136774	Transmembrane gp	N-glycosylation site at locations 6, 71, 92, 99, 110, 119, 123, 140, 157, and 161 and possibly at locations 151, 155, and 179 as well; support for transmembrane domain in regions 8-30 and 190-200; similarity to carcinoembryonic antigen
UL2	136779	HCMV-specific transmembrane protein	Support for transmembrane domain in region 35-55; weak support for cadherin 2 domain in region 5-21; very strong similarity to a region from *Homo sapiens* cadherin
UL3	136780	Transmembrane protein	Support for transmembrane domain in region 51-66; similarity to the C-terminal region of maturase O9828 from *Upuna bomeensis*
UL4	136785	HCMV-specific transmembrane gp	Support for a signal sequence in region 1-25; regions 35-50 and 135-150 are virus specific; N-glycosylation site at locations 48, 53, 61, 69, 108, 112, 122, 139, and 148; weak support for phosphorylation at location 40; virtually identical to the same protein from the Towne strain
UL5	136791	HCMV-specific transmembrane gp	Support for transmembrane domain in regions 25-35 and 115-140
UL6	136793	HCMV-specific transmembrane gp	Region 140-156 is virus specific; N-glycosylation site at locations 79, 102, 111, 147, 162, 174, 211, 228, and 234; support for transmembrane domain in region 235-265; similarity with protein E321_ADE1P from adenovirus; similarity with a transposase (TC1A_CAEEL) from *Caenorhabditis elegans*
UL7	136797	HCMV-specific transmembrane gp	N-glycosylation sites at locations 50, 56, 60, 105, 109, 125, 132, 147, 164, 168, and 189; support for transmembrane domain in regions 26-43, 85-102, 165-180, and 191-216

Continued on following page

TABLE 1—*Continued*

ORF label	Accession no.	Protein class or functional hypothesis	Features
UL8	136801	Virus-specific transmembrane gp	N-glycosylation site at locations 28 and 53; support for transmembrane domain in region 80-99; similarity with the C terminus of UL10 from the same genome
UL9	136805	Virus-specific transmembrane gp	Region 132-140 is virus specific; N-glycosylation sites at locations 41, 93, 100, 128, and 164; support for transmembrane domain in regions 3-17, 93-125, and 191-205; homology with a large number of ORFs from the same genome
UL10	136809	Virus-specific transmembrane gp	Region 85-95 is virus specific; N-glycosylation sites at locations 128, 172, 181, and 251; support for transmembrane domain in region 281-295; similarity with UL08 from the same genome
UL11	136814	Virus-specific transmembrane gp	Support for a signal sequence in region 1-25; N-glycosylation site at locations 42, 84, 92, 99, and 141; support for transmembrane domain in region 230-245; polythreonine stretches in region 145-190; homology with a number of ORFs from the same genome
UL12	136820	HCMV-specific gp	Support for a signal sequence in region 2-10; N-glycosylation site at location 38; similarity with splicing factor R/S-rich SFR3_HUMAN
UL13	136821	HCMV-specific transmembrane gp	Support for a signal sequence in region 1-32; N-glycosylation site at locations 146, 150, 226, 356, and 437; probable active site of unknown nature at location 386; support for transmembrane domain in region 450-470
UL14	136822	HCMV-specific transmembrane gp	Support for a probable signal sequence in region 15-30; N-glycosylation site (at locations 160 and 255; support for transmembrane domain in regions 15-30, 280-297, and 322-337; support for disulfide bridge involving the cysteine at position 156; homologous with UL141 from the same genome
UL15	136826	HCMV-specific transmembrane gp	Region 5-20 is virus specific; N-glycosylation site at location 285 (probably at locations 128 and 314 as well); support for transmembrane domain in regions 65-80, 130-145, 185-200, and 235-250; hydroxylation site at locations 299 and 301
UL16	136828	HCMV-specific transmembrane gp	Support for a signal sequence in region 6-25; N-glycosylation site at locations 35, 41, 68, 84, 95, 101, 132, and 145; support for transmembrane domain in region 190-205; substantial similarity with sequence P78996 from *Saccharomyces pastorianus*
UL17	136834	HCMV-specific transmembrane protein	Support for transmembrane domain in region 66-80
UL18	138157	Transmembrane gp; similarity to MHCI	Support for a signal sequence in region 3-20; support for transmembrane domain in region 320-350; domain alpha-1-like (HLA class I) in region 19-114; domain alpha-2-like (HLA class I) in region 115-208; domain alpha-3-like (HLA class I) in region 209-298; N-glycosylation site at locations 56, 66, 74, 95, 123, 127, 150, 167, 177, 193, 240, 282, and 291; similarity to MHC class I antigen
UL19	136839	Inconclusive	Probable active site at locations 31 and 97; similarity to meiotic-9 protein MEI9_DROME from *D. melanogaster;* similar to region from cyclolysin secretion ATP-binding protein of *Bordetella pertussis*
UL20	136840	HCMV-specific transmembrane gp	Support for a signal sequence in region 2-25; region 105-120 is virus specific; N-glycosylation sites at locations 27, 54, 57, 68, 72, 78, 83, 107, 118, 146, 173, and 180; support for transmembrane domain in region 222-245
UL21	136844	HCMV-specific transmembrane protein	Support for a signal sequence in region 160-174; support for transmembrane domain in region 85-105; C-terminus similarity with a region from a *Rhodopseudomonas palustris* phosphoenolpyruvate carboxylase (CAPP_RHOPA)
UL21.5	15808850	Inconclusive	Support for a signal sequence in region 1-20; probable coiled coil in region 86-103; homologous to the N terminus of a putative bacterial lipoprotein
UL22	136850	Transmembrane protein	Support for transmembrane domain in regions 50-78, 85-95, and 105-119; similarity to the C terminus of cytochrome *b;* weak similarity to cytochrome *c* oxidase polypeptide III
UL23	136851	HCMV-specific transmembrane gp	Transmembrane domain in regions 145-160 and 190-205; N-glycosylation site at location 137
UL24	136853	Herpesvirus specific transmembrane protein	Support for transmembrane domain in regions 60-98, 210-225, 287-305, and 317-340; shares a small region with bacterial 47.4-kDa protein O83469; similarity with ATP synthase a chain ATP6_MARPO
UL25	136862	Herpesvirus-specific virion protein	Region 260-272 is virus specific; N-glycosylation site at locations 146, 154, and 303; polyserine in regions 1-15 and 140-160
UL26	136868	HCMV-specific virion gp	N-glycosylation site at location 136; support for transmembrane domain in region 55-75; probable metal binding (magnesium) at location 42
UL27	136869	Herpesvirus specific	Moderate support for transmembrane domain in regions 125-145, 180-195, 390-408, and 565-580; similarity with other ORFs from herpesvirus genomes
UL28	136870	Virus-specific transmembrane protein	Moderate support for transmembrane domain in regions 35-45, 80-110, 130-145, 220-250, 260-275, and 291-310
UL29	136871	Herpesvirus specific	Region 1-30 is virus specific; probable disulfide bridge involving the cysteines at locations 103 and 125; polyserine in region 1-30; second half of gi_136871 is similar to the middle of gi_137158
UL30	136872	HCMV-specific transmembrane protein	Transmembrane domain in region 84-102; weakly similar to mitochondrial protein s14
UL31	136874	Herpesvirus-specific gp	Support for a signal sequence in regions 100-115 and 135-150; N-glycosylation sites at locations 275 and 296; support for hydroxylation site at location 381, 386, 390, and 393; sequence shares a small region with chmadrin Q9XS53; polyserine/threonine in region 100-125; polyglycine in region 380-400
UL32	130702	Virion gp	O-glycosylation site at locations 921 and 952; strong support for transmembrane domain in regions 130-150, 190-210, 305-318, 450-473, 580-600, and 895-913; polyalanine in region 345-366; polyserine in region 825-890; support for coiled coil in region 370-410
UL33	136882	GPCR homologue (chemokine receptor)	Support for transmembrane domain in regions 8-32, 49-73, 81-107, 117-138, 190-203, 219-245, and 270-283; similarity to galanin receptors type 3 (GPCRs)

Continued on facing page

TABLE 1—*Continued*

ORF label	Accession no.	Protein class or functional hypothesis	Features
UL34	136887	HCMV-specific transmembrane gp	N-glycosylation site at location 30; probable phosphorylation site at location 160; support for transmembrane domain in regions 60-75, 265-278, 315-335, and 350-370; homology with O12931 from atypical cytopathic virus and stealth virus; similarity to DNA polymerase
UL35	1136891	Herpesvirus-specific gp	N-glycosylation site at location 208 (possibly at location 515 as well); polyserine in region 495-530
UL36	136892	Herpesvirus specific	Polyaspartic acid/glutamic acid in region 325-335
UL37	138325	HCMV-specific transmembrane gp	Support for a signal sequence in region 2-22; region 95-110 is particular to eukaryotic sequences; N-glycosylation site at locations 206, 210, 219, 223, 242, 246, 275, 281, 294, 297, 306, 333, 337, 343, 379, 384, 390, and 391; support for transmembrane domain in region 435-460
UL38	136897	HCMV-specific transmembrane gp	Metal-binding site (iron) at location 144; N-glycosylation site at location 76; support for transmembrane domain in regions 25-45 and 130-160; probable coiled coil in region 160-195; probable disulfide bridge involving the cysteines at locations 64 and 66; C-terminal similarity to a pulative transcriptional regulator from *Streptomyces coelicolor* (O69988); similarity to *ski* oncogene; homology with ORFs from HHV-6 and HHV-7
UL39	136898	HCMV-specific transmembrane gp	N-glycosylation site at location 15; active site of unknown nature at location 38; support for transmembrane domain in regions 2-20 and 60-75
UL40	136899	HCMV-specific transmembrane gp	Support for a signal sequence in region 14-26; region 110-132 is virus specific; N-glycosylation site at locations 89, 110, and 143 (possibly at location 6 as well); support for transmembrane domain in region 175-190
UL41	136900	HCMV-specific transmembrane protein	Support for a probable signal sequence in region 125-141; region 1-7 is virus specific; local similarity with Q9ZV49 from *Arabidopsis thaliana*
UL42	136904 2454626	HCMV-specific transmembrane gp	Region 35-60 is eukaryote specific; N-glycosylation site at location 115; polyproline in region 27-50; support for transmembrane domain in region 88-108
UL43	136908 2454627	HCMV-specific transmembrane gp	N-glycosylation site at location 159; transmembrane domain in regions 106-115, 134-147, and 176-192
UL44	136913	DNA polymerase processivity factor	Region 325-400 is eukaryote specific; polyglycine in region 290-400; support for hydroxylation at locations 339 and 384; homologous to a mouse CMV phosphoprotein (Q69216)
UL45	132604	Ribonucleoside-diphosphate reductase large-chain protein	
UL46	136918	Herpesvirus-specific capsid protein	Support for transmembrane domain in regions 145-153 and 241-265
UL47	136919	Herpesvirus-specific virion protein	Region 955-981 is eukaryote specific; N-glycosylation site at locations 425 and 474
UL48	136925	Virion protein	Region 260-335 is eukaryote specific; support for coiled coil in regions 939-975, 985-1015, and 1292-1348; polyserine in region 330-335; homologous to Q9WJZ8, a human CMV F fragment DNA encoding DNA polymerase and gB
UL48.5	xxxxx	Capsid protein	
UL49	136926	Virus-specific transmembrane gp	Region 395-408 is virus specific; N-glycosylation site at location 237; polyalanine in region 304-323; probable transit peptide (mitochondrion) in region 1-14; probable active site of unknown nature at location 143; support for transmembrane domain in regions 138-160, 180-195, 305-320, 330-350, and 450-460; homology with a number of ORFs from other viruses
UL50	136931	Herpesvirus-specific transmembrane protein	Support for a probable signal sequence in region 355-390; support for transmembrane domain in regions 80-100, 225-240, and 275-300; proline-rich region 200-230; homology with ORFs from other herpesviruses
UL51	136932	Virus-specific transmembrane protein	Support for transmembrane domain in regions 45-55 and 133-145; similarity to a region of a DNA polymerase from herpes simplex virus (DPOL_HSV21); homology with ORFs from other viruses
UL52	136938	Herpesvirus-specific transmembrane gp	Region 555-574 is virus specific; N-glycosylation sites at locations 142, 145, 295, and 453; hydroxylation site at location 412; probable coiled coil in region 375-413; support for transmembrane domain in regions 582-598 and 609-625; homology with ORFs from other herpesviruses
UL53	136944	Virus specific	Regions 94-105, 144-155, and 204-215 are virus specific; polyserine in region 310-340; coiled coil in region 260-300
UL54	1169409	DNA polymerase	Region 645-680 is eukaryote specific; polyserine in region 640-680; N-glycosylation site at location 661 (possibly at location 345 as well); support for a disulfide bridge involving the cysteines at locations 304 and 313
UL55	138192	Virion gp; gB	Support for a signal sequence in region 1-25; N-glycosylation site at locations 37, 68, 73, 85, 208, 281, 286, 302, 341, 383, 405, 409, 417, 447, 452, 464, 465, 554, and 585; region 25-705 is extracellular; region 773-end is cytoplasmic; support for transmembrane domain in region 705-770
UL56	136950	Terminase DNA-packaging protein	Support for a signal sequence in region 1-5; region 440-490 is eukaryote specific; N-glycosylation site at locations 127, 446, 449, and 458; probable transmembrane domain in regions 240-265, 545-573, 595-626, 640-665, 700-720, and 780-790; polyserine in region 435-460; homology with many viral sequences
UL57	118745	Single-stranded DNA-binding protein	Region 535-600 is eukaryote specific; zinc finger (c2hc type) in region 467-481; polyglycine in region 540-587; probable transmembrane domain in regions 40-55, 125-152, 345-355,425-445, 455-480, 680-705, and 1170-1185; N-glycosylation site at location 999; probable triple helical domain in region 534-588; ATP binding in region 799-807; hydroxylation site at locations 541, 547, and 1195

Continued on following page

TABLE 1—*Continued*

ORF label	Accession no.	Protein class or functional hypothesis	Features
UL58	136953	Transmembrane protein	Region 62-70 is virus specific; probable transmembrane domain in region 17-33; similarity to the N-terminal region of nodulation protein from *Rhizobium fredii*
UL59	136954	HCMV-specific transmembrane protein	Probable phosphorylation site at locations 101 and 103; support for transmembrane domain in regions 50-65 and 83-95
UL60	136955	HCMV specific	Transmembrane domain in regions 30-60 and 140-160; this sequence shares similarity with an extended region of "spike" glycoprotein e2 from Simliki Forest virus (POLS_RRVT/POLS_SFV)
UL61	136956	HCMV-specific transmembrane protein	Regions 305-315 and 380-405 are eukaryote specific; possible cell attachment site in region 229-232; support for transmembrane domain in regions 140-160 and 205-220; hydroxylation site at locations 24, 26, 225, 227, 238, 264, 311, 383, and 391; probable triple-helical region in region 375-400; similarity to the N-terminal region of a (probable) nuclear antigen from pseudorabies virus (strain Kaplan) (VNUA_PRVKA)
UL62	138957	HCMV specific	Regions 1-15 and 160-175 are eukaryote specific; hydroxylation site at location 159 and 169; polyserine in region 1-20
UL63	136958	HCMV-specific transmembrane protein	Region 68-75 is virus specific; support for transmembrane domain in regions 7-25 and 45-65
UL64	136959	HCMV specific	Transmembrane domain in region 35-54; similarity to a region of a valyl-tRNA synthetase from *Synechocystis* sp. (SYV_SYNY3)
UL65	136960	Virion protein	Similarity to the N terminus of the 67-kDa tegument protein TEGU_HCMV
UL66	136961	HCMV specific	Short C terminus similarity with sperm protamine p1 from *Alouatta seniculus*
UL67	136962	HCMV specific gp	Region 50-64 is eukaryote specific; N-glycosylation site at locations 20 and 97; probable phosphorylation site at location 62
UL68	136963	HCMV-specific transmembrane gp	Region 60-80 is eukaryote specific; probable N-glycosylation site at location 92; support for transmembrane domain in region 7-24; similarity to the C terminus of a homeobox protein from *Xenopus laevis* (HMD4_XENLA)
UL69	136964	Transcriptional regulator	Regions 170-210 and 690-720 are eukaryote specific; polyproline in region 700-718; support for transmembrane domain in regions 480-491 and 525-550; ie63 homolog
UL70	136965	DNA helicase or primase	N-glycosylation site at location 730
UL71	136966	Herpesvirus specific	Region 310-340 is eukaryote specific; moderate support for transmembrane domain in region 215-240; probable coiled coil in region 80-105; 15 of the 20 amino acids of a DNA-binding (helix-tum-helix) motif are in region 279-293; similarity to the C-terminal region of DNA-directed RNA polymerase (RPOB_SALTY, RPOB_ECOLI, and RPOB_BUCAP)
UL72	118954	Deoxyuridine 5′-triphosphate nucleotide hydrolase	
UL73	136967	Herpesvirus-specific transmembrane protein	Support for a signal sequence in region 2-20; region 75-end is particular to viral sequences; support for transmembrane domain in region 100-122
UL74	136968	HCMV-specific transmembrane gp; gO	N-glycosylation site at locations 75, 83, 87, 103, 130, 157, 162, 171, 219, 242, 288, 292, 350, 385, 392, 399, 433, and 454; support for transmembrane domain in regions 10-28 and 190-212; probable coiled coil in region 240-272
UL75	138313	Virus specific transmembrane gp; GH	Support for a signal sequence in region 1-22; N-glycosylation site at locations 56, 63, 68, 193, 642, and 701; support for transmembrane domain in region 720-735
UL76	136969	Virus specific	Transmembrane domain in regions 40-60 and 73-95; probable ATP binding in region 2-9
UL77	136970	Virus specific virion protein	N-glycosylation site at locations 210, 246, and 263; support for transmembrane domain in regions 2-15, 115-125, 155-165, 430-450, 465-480, 518-530, 550-560, and 630-642
UL78	136971	GPCR homologue (purinoreceptor-like)	N-glycosylation site at location 105; transmembrane domain in regions 43-60, 74-95, 110-131, 155-180, 204-222, 236-255, and 280-300; there is support for similarity with NADH-ubiquinone oxidoreductase chains 4 and 6
UL79	136972	Virus specific	Transmembrane domain in regions 130-170 and 205-217; sequence appears to contain glimpses of a kinase domain; very good conservation of this protein across the herpesviruses
UL80 UL80a UL80.5	139232	Maturational proteinase; assembly scaffold protein	Cleavage by the protease in regions 256-257 and 643-644; active site (charge relay system) at locations 63, 132, and 157; support for transmembrane domain in regions 8-18, 45-60, 66-80, 340-365, and 670-687
UL81	136973	HCMV specific	
UL82	130716	Virion protein	Site involved in the formation of salt bridges at location 238; probable stromal domain in region 359-373; possible transmembrane domain in regions 120-145, 200-220, 323-345, 460-475, and 480-500
UL83	130714	Virion protein	Phosphorylation site at location 472; N-glycosylation site at location 93; active site of unknown nature at location 184; probable binding site (calcium) in region 398-409; similarity between its C terminus and the C terminus of gi_2352031_gb_AAC60369.1_, a Tuple1/HirA protein from *Fugu rubripes*
UL84	136974	HCMV-specific transmembrane protein	Polyglutamic acid/aspartic acid in region 150-183; polyarginine in region 9-19; support for transmembrane domain in regions 43-50 and 225-240
UL85	139190	Virus specific minor capsid protein	Region 1-12 is virus specific; N-glycosylation site at locations 67 and 297
UL86	137570	Major capsid protein	Probable calcium-binding site in region 520-530; site involved in the control of polyamine-mediated channel gating at location 637; probable N-glycosylation site at locations 341, 526, and 637; site acylated by penicillin at location 1256
UL87	136976	Virus-specific transmembrane gp	Regions 340-360 and 910-935 are virus specific; N-glycosylation site at locations 37, 591, 661, and 801; polyglycine in region 340-365; polyserine in region 900-end; support for transmembrane domain in regions 85-95, 570-590, and 813-825; probable triple-helical region in 341-362

Continued on facing page

TABLE 1—*Continued*

ORF label	Accession no.	Protein class or functional hypothesis	Features
UL88	136978	Herpesvirus-specific virion protein	
UL89	139645	Terminase DNA-packaging protein	Probable N-glycosylation site at location 329
UL90	136980	hcmv specific	
UL91	136981	Herpesvirus specific	Support for a probable signal sequence in region 1-19
UL92	136984	Herpesvirus-specific transmembrane protein	Region 129-141 is virus specific; support for transmembrane domain in regions 35-50, 90-105, and 144-162; very well conserved across many viruses
UL93	136986	Virus-specific transmembrane protein	Support for transmembrane domain in regions 125-135, 420-440, 475-490, and 529-550; GPI anchor at location 235; homologous to a sequence from rhesus CMV (strain 68-1) (Q9YUG2)
UL94	136987	Virus-specific transmembrane protein	Region 73-90 is virus specific; support for transmembrane domain in regions 95-115, 165-199, 205-228, 279-305, and 329-345; probable disulfide bridges involving the cysteines at locations 243, 250, 252, and 256; similarity with a number of viral sequences
UL95	136989	Herpesvirus-specific transmembrane gp	Region 50-100 is eukaryote specific; polyglycine in region 375-395; polyserine in region 45-100; support for transmembrane domain in region 375-395; N-glycosylation site at location 190
UL96	136991	Herpesvirus specific	Probable active site at location 43
UL97	136993	Phosphotransferase (ganciclovir kinase)	Regions 270-285 and 410-422 are virus specific; ATP binding in region 337-345; polyalanine in region 40-65; active site (of unknown nature) at locations 454, 456, and 461; weak support for the presence of a calcium-binding region in 696-707
UL98	119692	DNase	Probable transmembrane domain in regions 170-195, 240-270, 305-318, 390-415, and 525-535
UL99	130711	Virion protein	
UL100	136994	GPCR homologue; gM	Region 1-10 is virus specific; support for transmembrane domain in regions 15-35, 80-100, 150-172, 200-220, 240-260, 270-285, and 299-319; D/E-rich in region 359-end; N-glycosylation site at locations 56 and 113
UL102	136996, 603226	DNA helicase/primase	Polyserine in region 795-840
UL103	136997	Herpesvirus-specific transmembrane gp	Support for transmembrane domain in regions 57-70, 95-110, and 220-230; N-glycosylation site at location 20; similarity of a C-terminus region to an N-terminus region from cytochrome *b*
UL104	136998	Transmembrane gp	N-glycosylation sites at locations 354, 457, and 570; support for transmembrane domain in regions 61-75, 173-185, 215-240, 395-420, and 570-590; hydroxylation site at locations 684, 687, and 693 (possibly at 391 and 394 as well); similarity with sequences from many viruses
UL105	122808	DNA helicase/primase	ATP binding in region 120-127
UL106	136999	HCMV specific	Local similarity with a bacterial dehydrogenase (Q9ZHH7)
UL107	137000	HCMV-specific gp	N-glycosylation site at location 144
UL108	137001	HCMV-specific gp	N-glycosylation site at location 69; moderate support for transmembrane domain in regions 50-77 and 106-123; similarity with two regions from NU5M_CHOCR; similarity to the N terminus of DNA polymerase from woodchuck hepatitis virus (DPOL_WHV1/DPOL_WHVW6)
Y13K	139947	Transmembrane protein	Support for transmembrane domain in region 91-116; similarity to the C-terminal region of cytochrome *b*
Y9K	140150	Transmembrane gp	Probable N-glycosylation site at locations 3 and 27; support for transmembrane domain in regions 5-24 and 45-60; weak similarity to the C terminus of yeast protein phosphatase regulatory subunit
UL109	137002	Inconclusive	Probable active site at locations 69 and 72; probable phosphorylation site at locations 96 and 98; a small region of the C terminus of this sequence is shared with FACA_HUMAN (Fanconi anemia group A); similar to an N-terminal region of a putative dehydrogenase from *S. coelicolor*
UL110	137003	HCMV specific	
UL111	137004	HCMV-specific transmembrane protein	Weak support for transmembrane domain in regions 7-21 and 74-85; region 1-121 shows good similarity to the N-terminal region of glutamate gated chloride channel beta subunit precursor from *Haemonchus contortus* (P91730)
UL111.5	137012	IL-10-like protein	Support for a signal sequence in region 2-15
UL112	137005	Early phosphoprotein p34	Poly-glycine in region 200-220; homologous to the N terminus of EP84_HCMVA, O10418, and O10417 from the same genome
UL112-113	137006	Early phosphoprotein p84	Polyglycine in region 86-111; shares region 70-150 with O10417 from the same strain, O10418 from the same strain, and Q69215 (Pp43) from the Towne strain; probable N-glycosylation site at locations 140, 141, 410, 423, and 444
UL114	137036	Uracil glycosylase	Active site at location 91
UL115	2506510	Virus-specific gp; gL	Support for a signal sequence in region 1-35; N-glycosylation sites at locations 74 and 114; similarity with citrate synthase
UL116	137008	Virus-specific transmembrane protein	N-glycosylation sites at locations 58, 73, 87, 107, 132, 133, 139, 181, 200, 247, 253, 282, 285, and 298; support for transmembrane domain in regions 40-70, 290-300, and 307-320
UL117	137009	Herpesvirus-specific transmembrane gp	Support for a signal sequence in region 165-200; N-glycosylation site at locations 247, 309, and 396; support for transmembrane domain in regions 165-200, 230-250, 310-321, and 390-410; extensive similarity shared with a group of herpesvirus sequences
UL118	137010	HCMV-specific transmembrane gp	Support for a signal sequence in region 1-15; N-glycosylation sites at locations 12, 43, 62, 81, 89, 105, 108, and 124; support for transmembrane domain in region 160-180; similarity with vesicular inhibitory amino acid transporter; substantial similarity to NADH-ubiquinone oxidoreductase chain 2; similar to gi_1052835_gb_AAB00492.1 and gi_1616985_gb_AAB16887.1_; homologous to Q9PY29, a late transcript from HHV-5 (see MEDLINE; 92015512)

Continued on following page

TABLE 1—*Continued*

ORF label	Accession no.	Protein class or functional hypothesis	Features
UL119	137011	HCMV-specific gp	Support for a signal sequence in region 1-15; polyserine in region 23-74; N-glycosylation site at locations 34, 48, 95, and 104
UL120	137013	CMV-specific transmembrane gp	N-glycosylation sites at locations 46, 49, 55, 84, 95, 113, 122, 137, 144, and 187; support for transmembrane domain in regions 4-30 and 174-190; similar to UL120 from *Macaca mulatta* CMV and BOLF3 from simian CMV; C-terminal similarity to the VPU protein of human immunodeficiency virus
UL121	137014	CMV-specific transmembrane gp	Support for a signal sequence in region 5-25; N-glycosylation site at location 104; support for transmembrane domain in regions 10-24, 94-110, and 140-160; similarity with cytochrome *c* oxidase polypeptide ii precursor (COX2_BACFI); similarity with (i) (M93360) ORF UL121 *M. mulatta* CMV and (ii) (U38308) BOLF4 simian CMV
UL122	138482	Transcriptional regulatory protein, IE2	Zinc finger in region 258-274; polyglycine/serine in region 90-152
UL123	138476	Virus-specific, transcriptional regulatory protein, IE1	Polyglutamic acid/aspartic acid in region 420-472; support for transmembrane domain in regions 116-135, 180-195, and 300-310
UL124	137015	Transmembrane gp	Support for a signal sequence in region 2-25; support for transmembrane domain in region 82-100; N-glycosylation site at locations 33, 46, and 64; local similarity with an adenylate cyclase type V from rat (CYA5_RAT); local similarity with *trans*-acting transcriptional activator protein icp4 from Marek's disease herpesvirus
UL125	137016	HCMV-specific transmembrane protein	Support for transmembrane domain in region 5-16; similarity with a region of SN22_HUMAN, a possible global transcription activator; similarity with a region of a cyclin-dependent kinase inhibitor 1C from *H. sapiens* (CDNC_HUMAN)
UL126	137017	HCMV-specific gp	N-glycosylation site at locations 16 and 33; local similarity with a region from Q9VN01 from *D. melanogaster*; similarity with protamine 2 from rat (gi_56968); similarity with a piece of sperm histone P2 from *H. sapiens* (HSP2_HUMAN)
UL127	137018	HCMV-specific transmembrane gp	N-glycosylation site at location 97; probable transmembrane domain in regions 84-102 and 115-127; similarity to a bacterial helix-tum-helix in RP54_KLEPN; similarity to the C-terminal region of viral DNA polymerases
UL128	137019	HCMV-specific transmembrane gp	N-glycosylation site at location 55; support for transmembrane domain in regions 25-40 and 85-100; moderate support for nucleotide phosphate binding (GTP) in region 13-16
UL129	137020	HCMV-specific transmembrane protein	Support for transmembrane domain in regions 35-45 and 95-112
UL130	137021	CMV-specific gp	Support for a signal sequence in region 1-25; N-glycosylation site at locations 85, 118, and 201; methylation site at location 145; moderate local similarity with protein kinase C-like 1
UL131	137022	HCMV-specific transmembrane gp	N-glycosylation site at location 70; support for transmembrane domain in region 10-28; homology with Q69208 (transforming domain III from HCMV) and Q68827 (Towne strain); similarity to the middle of a RNA polymerase alpha subunit from *Haemophilus influenzae*
UL132	137023	HCMV-specific transmembrane gp	Support for a signal sequence in region 2-20; region 190-205 is virus specific; N-glycosylation site at locations 31, 61, and 245; polyserine/threonine in region 30-60; support for transmembrane domain in region 80-107; weak support for O-glycosylation site at locations 60, 63, 64, and 68
UL132/Toledo	1167934	HCMV-specific transmembrane gp	Support for a signal sequence in region 1-20; region 190-205 is specific to viral sequences; polyserine/threonine in region 30-60; support for transmembrane domain in region 80-107; N-glycosylation site at locations 31, 72, and 245; homologous to ULD2_HCMVA and to Q68830 from strain Towne
UL133/Toledo	1167918	HCMV-specific transmembrane protein	Region 1-12 is specific to viral sequences; support for a disulfide bridge involving the cysteines at locations 87 and 97; support for transmembrane domain in regions 15-30, 45-65, and 230-245
UL134/Toledo	1167919	HCMV-specific transmembrane protein	Moderate support for transmembrane domain in region 145-175
UL135/Toledo	1167920	HCMV-specific transmembrane protein	Polyserine in region 145-165; support for transmembrane domain in region 25-45
UL136/Toledo	1167921	HCMV-specific transmembrane protein	Support for transmembrane domain in region 60-90
UL137/Toledo	1167922	HCMV-specific	
UL138/Toledo	1167923	HCMV-specific transmembrane protein	Support for transmembrane domain in region 5-30; polyserine in region 95-130
UL139/Toledo	1167924	HCMV-specific transmembrane gp	Support for a signal sequence in region 1-15; transmembrane domain in region 55-72; N-glycosylation site at locations 24 and 90; polyserine in region 10-40
UL140/Toledo	1167925	HCMV-specific transmembrane protein	Support for transmembrane domain in regions 20-45 and 55-70
UL141/Toledo	1167926	Transmembrane gp	Hydroxylation site at location 412; possible disulfide bridge involving the cysteines at locations 154 and 230; N-glycosylation site at locations 204, 219, and 234; support for transmembrane domain in regions 210-220 and 365-387; local similarities shared with YDEC_SCHPO, a putative DNA repair protein from *Saccharomyces pombe*; similarity with UL14 from the same genome
UL142/Toledo	1167927	HCMV-specific transmembrane gp	Region 160-175 is virus specific; support for transmembrane behavior domain in regions 55-70 and 275-290; N-glycosylation site at locations 32, 166, 171, and 176
UL143/Toledo	1167928	HCMV-specific transmembrane gp	Support for transmembrane domain in region 70-83; probable nucleotide phosphate binding (FAD/ATP part) in region 32-46; N-glycosylation site at location 54
UL144/Toledo	1167929	Transmembrane protein related to tumor necrosis factor receptor	Support for a probable signal sequence in region 1-15; region 65-110 is virus specific; support for transmembrane domain in region 135-159; similarity to a eukaryotic tumor necrosis factor receptor (Q9UM65)

Continued on facing page

TABLE 1—*Continued*

ORF label	Accession no.	Protein class or functional hypothesis	Features
UL145/Toledo	1167930	HCMV specific	Moderate similarity with ACPM_SCHPO, a putative acyl carrier protein, mitochondrial precursor (NADH-ubiquinone oxidoreductase)
UL146/Toledo	1167931	HCMV-specific transmembrane protein	Region 51-56 is virus specific; support for transmembrane domain in region 3-15; homologous with Towne strain UL152_HCMVA
UL147/Toledo	1167932	HCMV-specific transmembrane protein	Support for a signal sequence in region 15-44; support for transmembrane domain in region 55-65
UL147/Towne	1167940	HCMV-specific chemokine	Support for transmembrane domain in regions 25-40 and 55-75; active site of unknown nature at location 4; probable N-glycosylation site at location 54
UL148/Toledo	1167933	HCMV specific	Support for a signal sequence in region 1-19; region 60-71 is virus specific
UL149/Toledo	1167936	HCMV specific	Support for a signal sequence in region 35-52; region 95-120 is eukaryote specific; active site of unknown nature at location 3; local similarity with Q13424 alpha-1 syntrophin from *H. sapiens*
UL150/Toledo	1167937	HCMV-specific transmembrane protein	Region 1-20 is eukaryote specific; hydroxylation site at location 215; support for transmembrane domain in regions 275-295, 330-350, and 595-611
UL151/Toledo	1167938	HCMV-specific transmembrane protein	Support for transmembrane domain in regions 50-60, 90-100, and 140-160
UL152/Towne	1167941	HCMV-specific chemokine	Signal peptide in region 1-15; homologous with UL146/Toledo
UL153/Towne	1167942	HCMV-specific transmembrane gp	Regions 120-130 and 265-277 are virus specific; support for transmembrane domain in region 235-260; polyserine/threonine in region 10-75; N-glycosylation site at locations 20, 168, and 233; homologous with TR14_HCMVA
UL154/Towne	1167943	CMV-specific transmembrane gp	Support for transmembrane domain in region 355-373; polyserine in region 90-110; moderate support for N-glycosylation site at locations 106, 119, 150, 162, 189, and 224
IRL14	124888	HCMV-specific gp	Region 1-35 is virus specific; region 70-95 is eukaryote specific; remaining regions are bacterium-eukaryote specific; N-glycosylation site at location 24; calcium-binding site at location 150; shares its N terminus with TRL14 from the same genome
IRL13 to IRL1		See TRL13 to TRL1	See TRL13 to TRL1
J1I	125057	HCMV-specific gp	N-glycosylation sites at locations 50 and 233; support for disulfide bridges involving the cysteines at locations 131, 134, 137, and 142; similarity with J1L_HCMVA and J1S_HCMVA from the same genome
IRS1	124910	HCMV-specific gp	Bacterium-eukaryote specific; N-glycosylation site at locations 76, 118, and 223; hydroxylation site at locations 12, 64, 68, and 71; acetylation site at location 820; phosphorylation site at location 721; similarity with NUFM_CAEEL, a putative NADH-ubiquinone oxidoreductase 17.3-kDa subunit; similarity with the immediate-early protein ie180 IE18_PRVIF (pseudorabies virus); there is also a similarity with the middle and the C terminus of a β-lactamase precursor from *E. coli* (AMPC_ECOLI); quasi-identical to gi_136375 of the same genome
US1	137122	HCMV-specific transmembrane gp	Region 140-170 is eukaryote specific; N-glycosylation sites at locations 48 and 208; polyserine in region 145-170; support for transmembrane domain in regions 12-30 and 60-75; similarity to a region of OL56_STRAT (*S. antibioticus*), an oleandomycin polyketide synthase (modules 5 and 6)
US2	137123	HCMV-specific transmembrane gp	Region 148-164 is virus specific; N-glycosylation sites at locations 68, 172, and 188; support for transmembrane domain in region 165-190; similarity with putative NADH-ubiquinone oxidoreductase chain 2 (NU2M_CHICK)
US3	137128	HCMV-specific transmembrane gp	Support for a signal sequence in region 1-15; region 45-65 is virus specific; N-glycosylation site at locations 60 and 68; support for transmembrane domain in region 156-186; good similarity with US02_HCMVA
US4	137129	HCMV-specific transmembrane protein	Region 45-60 is eukaryote specific; support for transmembrane domain in region 104-117; active site of unknown nature at location 17; similarity with the RNA polymerase RPOD_SULSO
US5	137130	HCMV specific	
US6	137133	HCMV-specific transmembrane gp	Support for a signal sequence in region 1-19; N-glycosylation site at location 52; support for transmembrane domain in region 155-170
US7	137134	Transmembrane gp	Support for a signal sequence in region 1-15; N-glycosylation site at locations 69 and 205; support for transmembrane domain in region 170-190; US07_HCMVA, US11_HCMVA, US08_HCMVA, and US09_HCMVA share various regions in various orders; similarity to the C-terminal region of a cytochrome *c* oxidase polypeptide i from *Leishmania tarentolae*
US8	137135	HCMV-specific transmembrane gp	Support for a signal sequence in region 1-21; regions 50-55 and 149-157 are virus specific; N-glycosylation site at location 61; support for transmembrane domain in region 180-200; hydroxylation site at location 90; similarity with US7 and US9 from the same genome; also, some similarity with regions from enolases
US9	137136	HCMV-specific transmembrane gp	Support for a signal sequence in region 6-26; region 20-60 is eukaryote specific; support for transmembrane domain in region 195-213; N-glycosylation site at locations 97 and 158; probable disulfide bridge involving the cysteines at locations 9 and 16; similarity with Huntingtin
US10	137137	HCMV-specific transmembrane gp	Support for a signal sequence in region 1-25; transmembrane domain in region 150-170; N-glycosylation site at locations 111 and 143
US11	137143	HCMV-specific transmembrane gp	Support for a signal sequence in region 2-15; N-glycosylation site at location 73; support for transmembrane domain in region 185-203; similarity with truncated testis-specific box 1-less prolactin receptor from *Gallus gallus* (gi_2295649); similarity with the C terminus of the β-galactosidase precursor from *Arthrobacter* sp. (BGAL_ARTSP); similarity with US09 and US07 from the same genome
US12	137144	GPCR homologue (delta opioid receptor)	Support for transmembrane domain in regions 65-75, 115-135, 145-167, 175-195, 200-220, 250-258, and 265-275

Continued on following page

TABLE 1—*Continued*

ORF label	Accession no.	Protein class or functional hypothesis	Features
US13	137145	HCMV-specific GPCR homologue	Strong support for transmembrane domain in regions 30-49, 70-88, 95-112, 120-145, 155-165, 180-195, and 210-235; extended similarity with cytochrome *b* from *Hansenula wingei* (CYB_HANWI); weaker similarity with mouse NADH-ubiquinone oxidoreductase chain 4 (NU4M_MOUSE)
US14	137146	GPCR homologue (calcitonin receptor)	Support for transmembrane domain in regions 63-90, 105-118, 130-145, 155-165, 185-198, 210-225, and 255-269; N-glycosylation site at locations 291 and 299; similarity with an ATP synthase (ATP6_MYCTU)
US15	137147	HCMV-specific GPCR homologue	Support for transmembrane domain in regions 95-115, 188-210, 275-292, 355-378, 390-405, 415-430, and 450-475
US16	137148	HCMV-specific GPCR homologue	N-glycosylation site at location 178; active site of unknown nature at location 117; support for transmembrane domain in regions 50-65, 90-103, 130-145, 155-175, 190-202, 215-230, and 250-265; similarity with ATP synthase from *Synechocystis* sp.
US17	137149	HCMV-specific GPCR homologue	N-glycosylation site at location 287; support for transmembrane domain in regions 55-70, 85-95, 110-128, 140-158, 166-182, 195-210, and 235-250; similarity with US18 from the same genome; similarity with NADH-ubiquinone oxidoreductase chain 2 (NU2M_SYMSY and NU2M_RHIUN)
US18	137150	HCMV-specific GPCR homologue	Support for transmembrane domain in regions 50-68, 80-97, 110-125, 140-155, 170-185, 195-210, and 230-250; N-glycosylation site at location 242; helix-tum-helix-like motif in region 190-230; similarity with an ATP synthase (ATP6_ARTSF); similarity with NADH-ubiquinone oxidoreductase chain 5 (NU5M_DROYA)
US19	137151	HCMV-specific GPCR homologue	Support for transmembrane domain in regions 25-45, 65-85, 95-115, 125-145, 155-170, 185-200, and 220-238; reasonable quality alignment with NADH-ubiquinone oxidoreductase chain 1 (NU1M_ANOGA and NU1M_DROME)
US20	549171	HCMV-specific GPCR homologue	N-glycosylation site at locations 149, 176, and 330; support for transmembrane domain in regions 119-140, 150-168, 180-193, 205-225, 235-250, 265-286, and 296-315; support for a disulfide bridge involving the cysteines at locations 287 and 292; local similarity with a region of AAF60456 from *C. elegans*; similarity with NADH-ubiquinone oxidoreductase chain 5 from *Anopheles quadrimaculatus* (NU5M_ANOQU); the ORF is embedded in the middle of a sequence from rhesus CMV
US21	137153	GPCR homologue	Support for transmembrane domain in regions 15-35, 55-65, 80-96, 105-125, 140-152, 162-176, and 190-210; similarity with a genome polyprotein (POL1_BAYMJ) that contains an RNA-directed RNA polymerase from Barley yellow mosaic virus (Japanese strain II-1) (BaYMV); similarity with an ATP synthase a chain from *A. quadrimaculatus* (mosquito) (ATP6_ANOQU); similarity with cytochrome *b* from *Erinaceus europaeus* (Western European hedgehog) CYB_ERIEU; homologous to gi_3348079 from rhesus CMV; similarity with the newly sequenced CGI-119 from *H. sapiens* (gi_4929707_gb_AAD34114.1_AF151)
US22	137154	Transmembrane gp	N-glycosylation site at locations 410, 411, and 507; support for transmembrane domain in regions 55-70 and 290-310; active site (charge relay system) at location 531; shares a small region with cg10951 from *D. melanogaster*
US23	137155	HCMV-specific transmembrane protein	Region 10-25 is virus specific; support for transmembrane domain in regions 65-73, 208-218, and 250-265; probable binding site for GlcNAc-1-p at location 121; homologous with Q9QJ47 from HHV-6B
US24	137156	HCMV-specific transmembrane gp	N-glycosylation site at location 200; support for transmembrane domain in regions 260-275 and 285-301; lipid GPI anchor site at location 2; similarity with AF078102 rhesus CMV and with (U10326) ORF HJ7 SwissProt accession number P09722 (mouse CMV 1)
US25	137157	HCMV-specific transmembrane protein	Support for probable signal sequence in region 1-14; region 150-end is eukaryote specific; support for transmembrane domain in region 36-59; amidation site at location 121; local similarity with Q9V4J6 from *D. melanogaster*
US26	137158	Herpesvirus specific, transmembrane gp	N-glycosylation site at locations 44, 478, 482, 517, and 521; polyglutamic acid/aspartic acid in region 490-525; support for transmembrane domain in regions 270-294 and 315-340
US27	137159	GPCR homologue (chemokine receptor)	Support for transmembrane domain in regions 35-58, 68-90, 105-126, 149-167, 194-213, 234-257, and 275-298; probable pyridoxal phosphate-binding site at location 61; N-glycosylation site at locations 7, 15, 18, and 22; similarity with galanin receptors type 2 and 3 (GPCRs)
US28	137160	GPCR homologue (chemokine receptor)	Region 290-322 is eukaryote specific; strong support for transmembrane domain in regions 28-55, 64-94, 105-125, 145-160, 180-205, 225-240, and 275-290; N-glycosylation site at location 30; similarity with galanin receptors type 2 and 3 (GPCRs)
US29	137161	HCMV-specific transmembrane gp	Region 40-75 is eukaryote specific; N-glycosylation site at locations 54, 98, and 182; support for transmembrane domain in regions 140, 173, 219-232, 260-280, and 363-390; probable NADP binding in region 225-231
US30	137162	HCMV-specific transmembrane gp	Support for a signal sequence in region 51-65; support for transmembrane domain in region 225-255; N-glycosylation site at locations 75, 98, and 140
US31	137163	HCMV-specific gp	Support for signal sequence in region 1-23; region 55-66 is virus specific; N-glycosylation site at location 182 (possibly at location 165 as well); poly-serine in regions 120-140 and 160-183; similarity with an RNA2 polyprotein from the tobacco black ring virus (POL2_TBRVS); similarity with the core of a DNA polymerase from the hepatitis B virus (DPOL_HPBVW); similarity with a region from an ATP-dependent helicase recg from *Synechocystis* sp. (RECG_SYNY3)
US32	137164	Inconclusive	Region 1215-140 is eukaryote specific; hydroxylation site at locations 129, 131, and 135; similarity to an ATP-dependent DNA helicase from *Thiobacillus ferrooxidans*
US33	137165	HCMV-specific metalloprotein	Probable metal-binding site (zinc) at location 78; similarity with 8-amino-7-oxononanoate synthase from *M. tuberculosis* (BIOF_MYCTU)
US34	137166	HCMV-specific gp	Region 148-end is eukaryote specific; N-glycosylation site at locations 55, 79, 133, and 152
US35	137167	HCMV-specific transmembrane gp	N-glycosylation site at location 51; support for transmembrane domain in region 38-50

Continued on facing page

TABLE 1—*Continued*

ORF label	Accession no.	Protein class or functional hypothesis	Features
US36	137168	HCMV-specific transmembrane gp	N-glycosylation site at locations 14 and 96; phosphorylation site at location 90; support for transmembrane domain in region 50-65
TRS1	136375	hcmv-specific gp	N-glycosylation site at locations 76, 118, and 223; polyalanine in region 360-460; hydroxylation site at location 12, 68, 71, and 756; the N-terminal half of this ORF is identical to gi_124910_sp_P09715_IRS1_HCMVA
J1S	125059	HCMV specific	Regions 90-150 and 175-195 are eukaryote specific; the remaining regions are bacterium-eukaryote specific; polyproline in regions 90-110 and 175-190; probable hydroxylation site at locations 177, 178, 184, 185, 189, and 190; probable ATP-binding site at location 121; homologous with gi_125058_sp_P17143_J1L_HCMVA 137 from the same genome

[a] The following terms are used in this table: gp, glycoprotein; eukaryote specific, used in the features column to characterize a region and indicating that the majority of the seqlets matching somewhere in the region had instances primarily within eukaryotic sequences; bacterium specific, used in the features column to characterize a region and indicating that the majority of the seqlets matching somewhere in the region had instances primarily within bacterial sequences; virus specific, used in the features column to characterize a region and indicating that the majority of the seqlets somewhere in the region matched primarily within viral sequences and used in the functional hypothesis column to indicate that any identifiable sequence similarities are with viral sequences only; herpesvirus specific, used in the functional hypothesis column to indicate that any identifiable sequence similarities are with herpesvirus sequences only; CMV specific, used in the functional hypothesis column to indicate that any identifiable sequence similarities are with cytomegalovirus sequences only; HCMV specific, used in the functional hypothesis column to indicate that any identifiable sequence similarities are with HCMV sequences only. Regions expressed in the form *N-M*, with *N* and *M* integers, indicate all amino acids of the respective ORF between positions *N* and *M* inclusive. Finally, we make a distinction between "homologous" and "similar" sequences: we use the first adjective to refer to groups of sequences when there is reason to believe that they are evolutionarily related and the second adjective to refer to groups of sequences that are related to one another as deduced with the help of a similarity search tool and a scoring matrix. CMV, cytomegalovirus; MHC, major histocompatibility complex; GPI, glycophosphatidylinositol; gB, glycoprotein B, etc.

lies. However, each of these ORFs is composed of transmembrane helices that appear to have been "borrowed" from distinct GPCR families and placed in an order that has not been previously encountered; in other words, the transmembrane helices of each of these nine ORFs have not appeared as a group in a single GPCR (sub)family before. We computationally verified the situation with the ORFs encoding newly predicted GPCR-like proteins (UL100, US13, US15, US16, US17, US18, US19, US20, and US21) as follows. By using each of these sequences as a query and employing standard similarity searching tools, we compared it with other sequences and in particular with those contained in the GPCRDB database (15); only weak conservation spanning part of the ORF sequences could be identified. Subsequently, we extracted the amino acid subsequences that our analysis indicated as corresponding to the seven transmembrane helices of the ORF under consideration and used these shorter regions as queries for a search of the GPCRDB database: in each case, we discovered notable similarities of these queries/helices to annotated transmembrane helices of known GPCRs, but each one of these "hits" came from a different functional subdivision of the GPCR superfamily, thus supporting our statement above.

The case of one ORF in particular, UL78, is discussed in detail in (30). Therein, we show local alignments between the transmembrane helices of UL78 to transmembrane helices of well-characterized transmembrane proteins. A PSI-BLAST (3) search of the public databases based on UL78 could only identify an ~70-amino-acid region of UL78 as weakly similar to the rhodopsin family but could not determine the local similarities mentioned above, a direct consequence of their short length. In our analysis (see data available online at the companion website), we also give an alignment of UL78 with P2Y7_HUMAN, a human purinoreceptor.

Revisiting the previously defined HHV-5 gene families. In the original study of the sequence of HHV-5, Chee et al. (8) defined and described several families that comprised subsets of the reported putative genes. This categorization into the families was based on the use of heuristics-based similarity searching approaches. As we have described in previous work (29, 30), this approach to family determination can lead to incorrect conclusions in a manner analogous to incorrectly annotating proteins by assuming that the "transitive closure" property applies: the fact that sequence A is similar to sequence B and that sequence B is similar to sequence C should not be used to imply that sequence A is similar to sequence C. An annotator will frequently exploit either the first or the best "hit" in the output of a database search carried out by using the FASTA (22), BLAST (2), or Smith-Waterman (36) tool: in the presence of small but well-conserved regions or of domains that are shared by distinct proteins (11) this choice is sometimes not optimal. In fact, the multidomain organization of proteins can lead to incorrectly annotated database entries and, by extension, to incorrect definitions of protein families.

Chee et al. (8) defined eight gene families, namely, UL25, UL82, RL11, US1, US2, US6, US12, and US22. We analyzed these HCMV families with the help of MUSCA (23), a pattern discovery-based multiple sequence alignment algorithm, and reevaluated their definitions. The MUSCA algorithm is described in some detail above in Materials and Methods. The use of patterns in inducing multiple sequence alignments is particularly appropriate in the presence of shared domains. In the alignments that are described below, amino acids that participated in the patterns that induced the respective alignment are capitalized and are also colored based on their hydropathy. For some of these cases, the rather involved relationship of the considered sequences necessitated the manual selection of the regions to be aligned.

US1 family. The first family we examined was the US1 family comprising ORFs US1, US31, and US32. The multiple sequence alignment for these three sequences is shown in Fig. 1A and supports the original definition of this family.

RL11 family. We next analyzed the RL11 family, which included 14 ORFs: IRL11, TRL11, IRL12, TRL12, TRL14, UL1, and UL4 through UL11. Each of these 14 sequences also was used as a query in a FASTA search against the remaining 13 members of the original family. In all searches the deter-

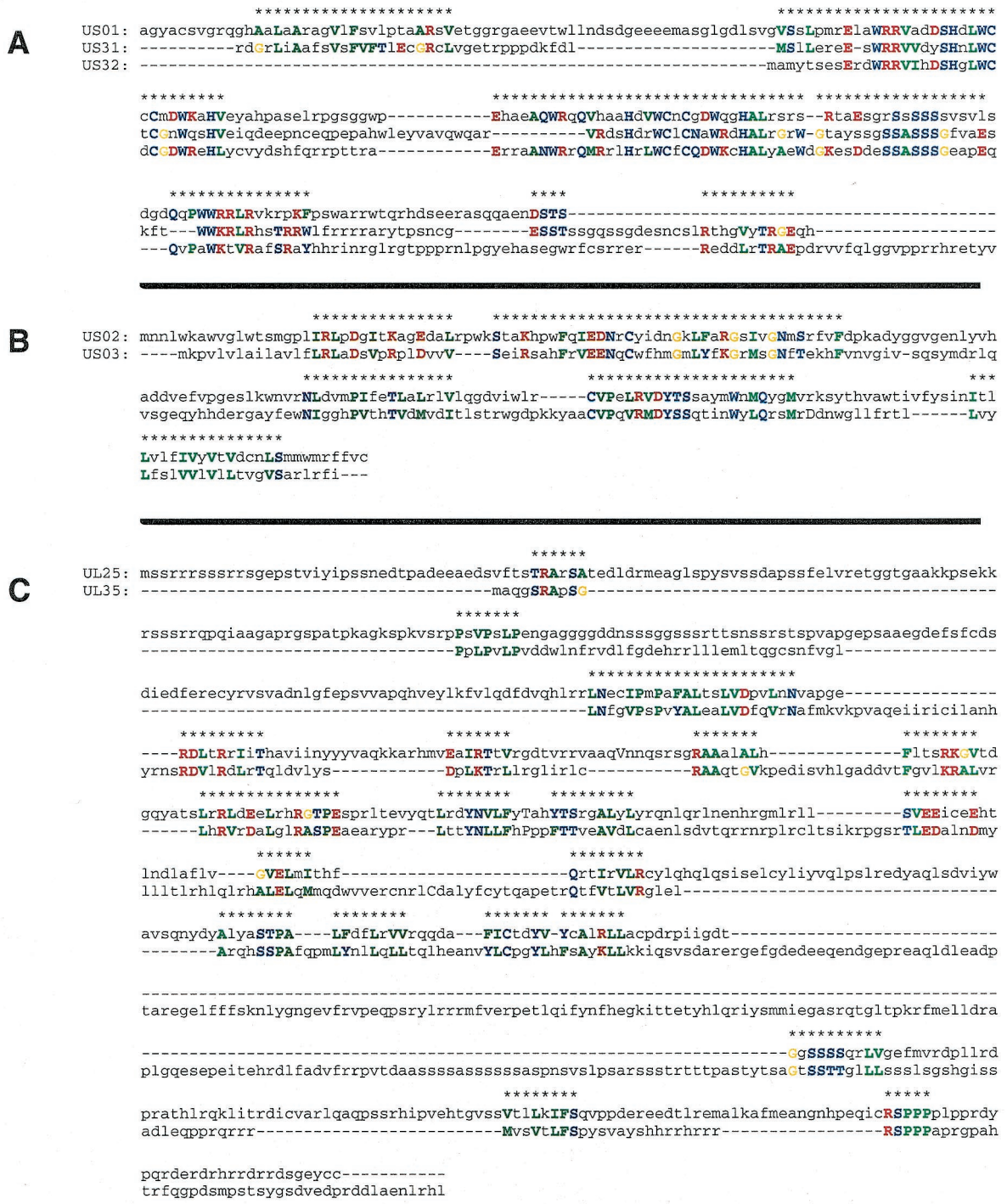

FIG. 1. Pattern-based alignments of the US1 family (A), the US2 family (B), and of UL25 and UL35 (C). In all cases, only the amino acids which participated in the patterns that induced the alignment are shown in color; the different colors represent different hydropathies.

mined local similarities scored below the moderately conservative threshold value of 200, essentially indicating the presence of only local, weak similarities among these 14 sequences and questioning the original family definition.

US2 family. The US2 family, as originally defined, comprised the proteins US2 and US3 whose alignment is shown in Fig. 1B and supports the original family definition.

UL25 family and UL82 family. The original definition of the UL25 family comprised the ORFs UL25 and UL35. We again used MUSCA to align the members of this family (Fig. 1C). As is evident, the similarity of these two sequences is rather weak, thus putting the original definition of the UL25 family into question. A similar situation exists in the case of the UL82 family that consists of UL82 and UL83: the re-

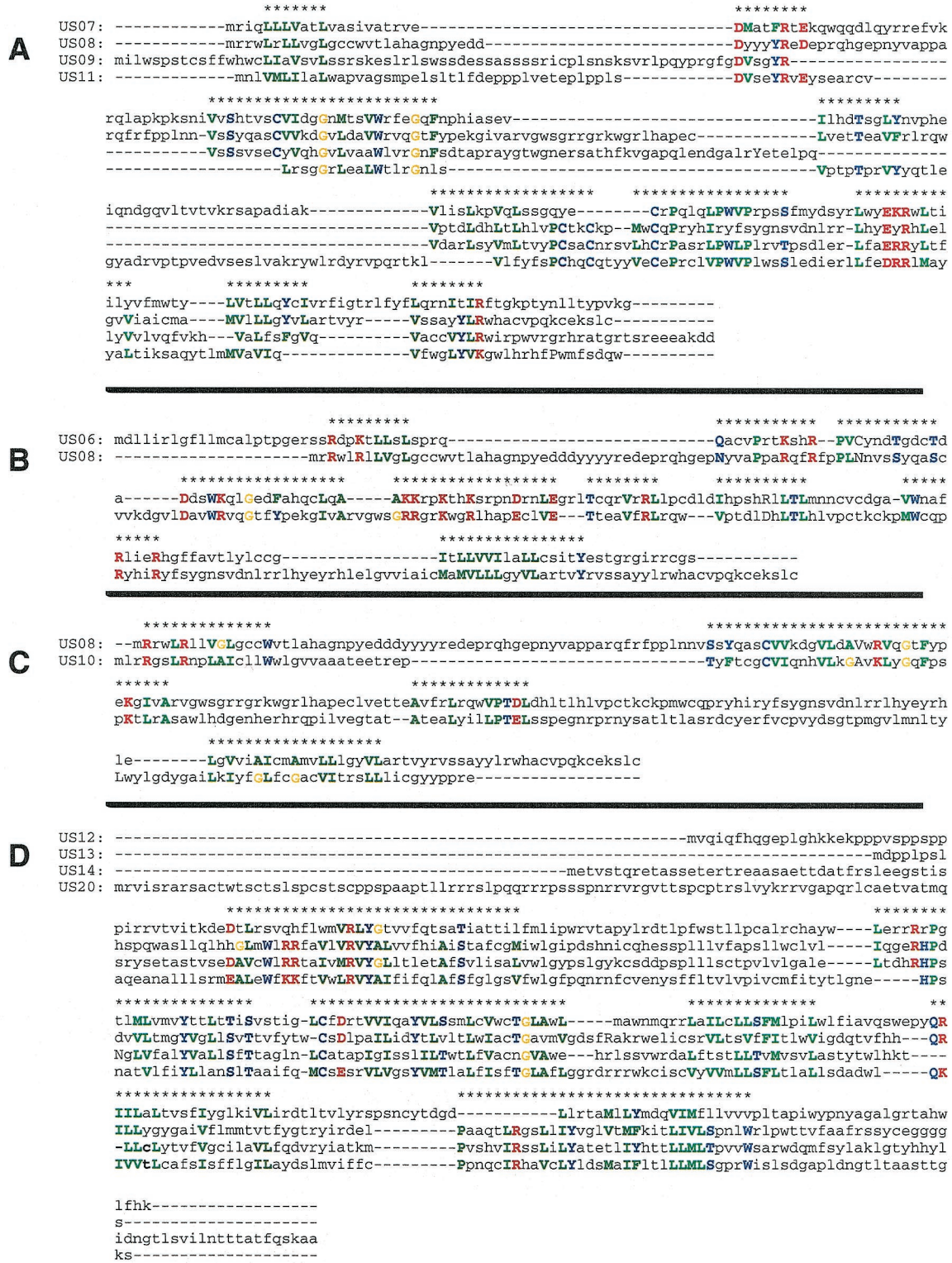

FIG. 2. Pattern-based alignments for four of the six members in the original US6 family (A); US6 with US8 from the original US6 family (B); US10 with US8 from the original US6 family (C); and US12, US13, US14, and US20 from the original US12 family (D). In all cases, only the amino acids which participated in the patterns that induced the alignment are shown in color based on their hydropathy.

maining sequence similarity is also weak (the alignment is not shown).

US6 family. The originally defined US6 family consists of the ORFs US6 through US11. Manual analysis of this sequence group indicates that US7, US8, US9, and US11 form a cluster and a multiple sequence alignment is shown in Fig. 2A. Fairly long regions are reasonably well conserved, thus supporting the hypothesis that these four sequences form a family. For the

remaining two sequences, US6 and US10, small regions appear to be shared among the pairs US6-US8 and US8-US10, but no unifying similarities are evident; see Fig. 2B and C. To recapitulate, our analysis suggests that the US6 family definition includes only US7, US8, US9, and US11.

US12 family. The original US12 family definition included the 10 ORFs US12 through US21. As in the case of the US6 family, manual analysis indicates that US12, US13, US14, and US20 form a cluster, and a multiple sequence alignment for them is shown in Fig. 2D. Similarly, US15 and US16 form a separate cluster, whereas US19 and US21 ought to be treated as singletons. The remaining two sequences, US17 and US18, form a weak cluster with US20 (cf. Fig. 2D), and an alignment of these three sequences, shown in Fig. 3A, indicates a rather low degree of sequence conservation. It is interesting that with the exception of the first conserved block, the regions shared among the group US17/US18/US20 are not the same as those shared by the group US12/US13/US14/US20. In summary, the original family definition ought to be revised and split into two groups: US12/US13/US14/US20 and US15/US16. MUSCA provides provides no support for inclusion of the remaining sequences: US17, US18, US19, and US21 should be removed from the family definition altogether.

US22 family. We, finally, examined the US22 family whose original definition comprised the ORFs UL23, UL24, UL28, UL29, UL36, UL43, US22, US23, US24, US26, IRS1, and TRS1. As above, we prescreened these sequences by using manual analysis and determined the presence of two sequence clusters: IRS1/TRS1/US22/US23 and US22/US23/US24/US26. US22, US23, US24, and US26 form a cluster that exhibits a much better degree of local conservation: indeed, as can be seen from the multiple sequence alignment for this group in Fig. 3B the conserved region spans their N-terminal regions but does not extend to their C termini. On the other hand, IRS1, TRS1, US22, and US23 form a cluster and a multiple sequence alignment is shown in Fig. 4: the degree of observed local conservation is moderate. It is difficult to say what the best recommendation is for this family; a conservative approach would reduce the members of this family to only the group US22/US23/US24/US26.

Table 2 summarizes the results of the analysis described above. For each of the originally defined families, we indicate how our results revise the family memberships of the respective ORFs.

DISCUSSION

Using pattern-based approaches, we analyzed the HHV-5 genome (strains AD169, Towne, and Toledo). A brief summary of the annotations that we generated is presented in Table 1, and the complete set of our findings can be explored at http://cbcsrv.watson.ibm.com/virus/. Several enhancements to the annotation of the HHV-5 genome have been made possible by this approach. In particular, we have revised the number of ORFs that code for transmembrane proteins, we have revised the number of ORFs that code for proteins containing glycosylation sites, we have found evidence for phosphorylation sites in at least 9 ORFs and for hydroxylation sites in at least 17 ORFs, we have found support for the existence of 15 ORFs that code for proteins with GPCR-like sequence composition and characteristics and, finally, 31 of the ORFs appear to be virus specific, adding support to the view that HHV-5 is an ancient virus that evolved separately from its host for an extended period.

One of the more intriguing predictions from this annotation exercise is the expectation that HHV-5 encodes a larger number of GPCRs than previously thought. UL33, UL78, US27, and US28 have been identified previously as encoding GPCRs (for a review, see reference 32). Our analysis strongly predicts that US12 and US14 encode GPCRs, and this is corroborated by good quality alignments with known GPCRs. An additional nine sequences, namely, the ORFs UL100, US13, and US15 to US21, each contain seven recognizable transmembrane domains; similarity searches indicate that the sequence fragments corresponding to each of the seven transmembrane regions of these nine ORFs match annotated transmembrane helices of known GPCRs but of distinct functional behavior. In other words, the transmembrane helices in each of these nine ORFs have not appeared as a group in a single GPCR (sub)family before. If these proteins prove to have GPCR activity, analysis of their physiological effects might provide important new insights to the function of this class of proteins since they are only distantly related in terms of their seqlet content to known GPCRs.

The possible hydroxylation of a set of HHV-5 proteins deserves further comment. Hydroxylation of proline residues influences protein assembly into a triple-stranded helix in the case of collagen (17) and directs the proteasome-dependent degradation of other proteins (6). It is possible, then, that the structure or half-life of one or more HHV-5 proteins is influenced by hydroxylation. Hydroxylation might influence the structure of UL61, which contains a domain that, like collagen, is predicted to form a triple helix and includes two possible hydroxylation sites. Further, since hydroxylation of hypoxia-inducible transcription factors controls the response of cells to changes in oxygen availability (6), one might speculate that hydroxylation of a viral protein might allow HHV-5 to sense and adapt to the oxygen environment of its host cell. For example, IRS1 and TRS1, two viral proteins with putative hydroxylation motifs, exhibit transcriptional activity in transfected cells (31, 38). Conceivably, their activity is modulated by hydroxylation.

Our analysis suggests numerous functional hypotheses based on similarities of domains within HHV-5 proteins to domains within other proteins of known function. For example, TRL4 is predicted to encode a transmembrane glycoprotein with similarity to the fas antigen ligand. Very little is known about this protein, although its mRNA is expressed in large quantities with early kinetics (14). There also is a report that an epitope-tagged (C-terminal) version of the protein is localized to the nucleolus within transfected cells (5). Our annotation does not anticipate a nucleolar localization, and it is possible that the reported localization was perturbed by the epitope tag. HHV-5 infection has been reported to induce the expression of the cellular fas ligand in certain cell types (9), and the expression of fas ligand has been proposed to assist in virus escape from immune surveillance (37). Consequently, the TRL4-encoded protein might prove to antagonize the host antiviral response by serving as a viral mimic of the cellular fas ligand. This

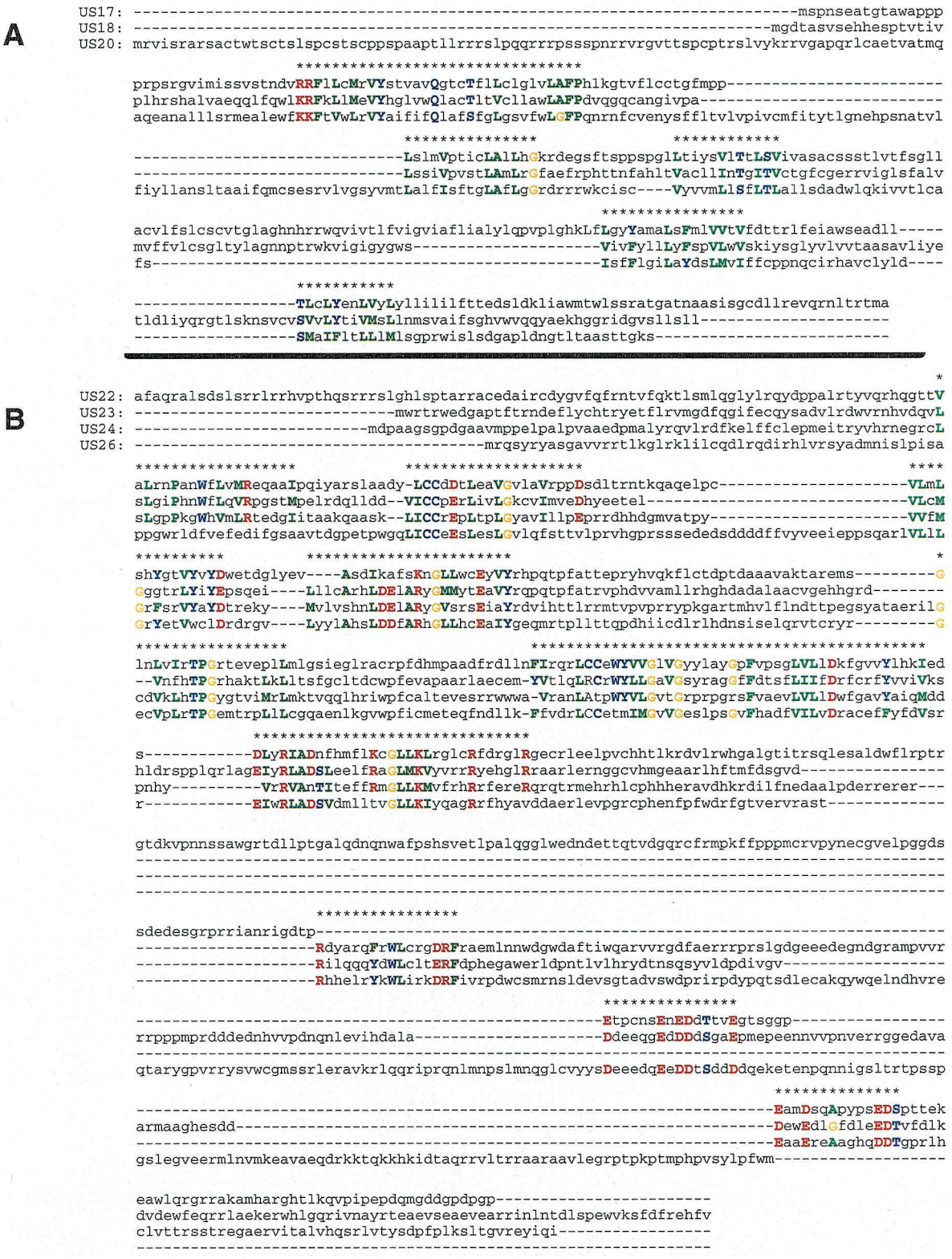

FIG. 3. Pattern-based alignments for US17, US18, and US20 from the original US12 family (A) and of US22, US23, US24, and US26 (B) from the original US22 family.

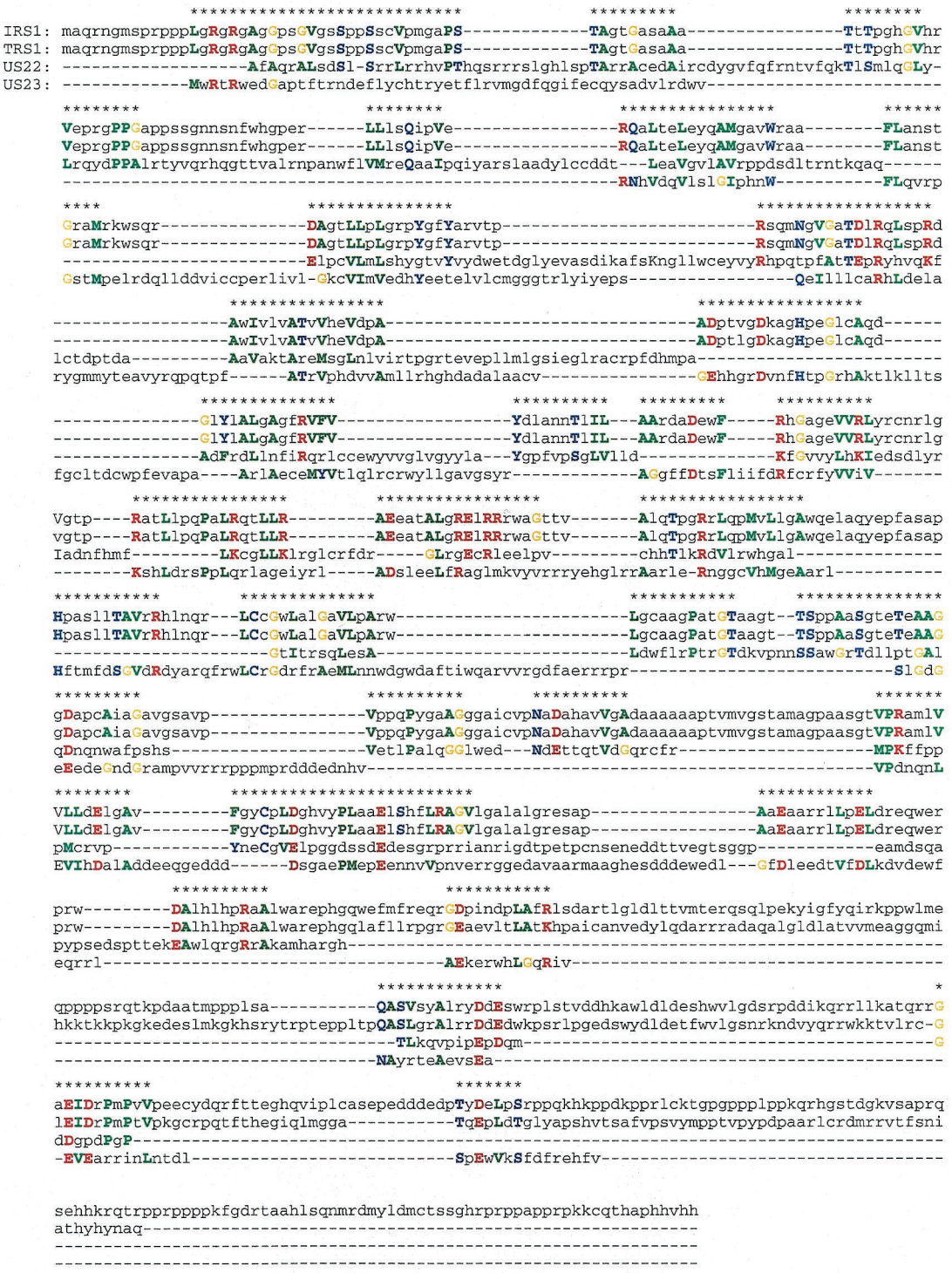

FIG. 4. Pattern-based alignment of IRS1, TRS1, US22, and US23 of the original US22 family. Only the amino acids which participated in the patterns that induced the alignment are shown in color based on their hydropathy.

hypothesis, as well as additional hypotheses generated by our annotation, can be tested experimentally.

In addition to reannotating the complete HCMV genome, we also reevaluated the original HHV-5 gene family definitions with the help of a pattern-based multiple sequence alignment algorithm and proposed new groupings for the original family members; we also identified cases in which the detectable sequence similarities observed in our analyses were not strong enough to support family membership. Our analysis with MUSCA of protein families was limited to a reevaluation of

TABLE 2. Revised groupings of the originally defined HCMV gene families

Original family name	Original family members	Revised sequence groupings
US1	US1, US31, US32	US1, US31, US32
RL11	IRL11, TRL11, IRL12, TRL12, TRL14, UL1, UL4 through UL11	No groupings
US2	US2, US3	US2, US3
UL25	UL25, UL35	No groupings
UL82	UL82, UL83	No groupings
US6	US6, US7, US8, US9, US10, US11	US7, US8, US9, US11
US12	US12, US13, US14, US15, US16, US17, US18, US19, US20, US21	US12, US13, US14, US20; and US15, US16
US22	UL23, UL24, UL26, UL28, UL29, UL36, US22, US23, US24, US26, IRS1, TRS1	US22, US23, US24, US26

the families proposed by Chee et al. (8). Additional families might exist that could be identified by using pattern-based approaches.

The present analysis focused on the original annotated set of HHV-5 ORFs (8). It is possible that the genome contains substantially more coding ORFs than we have analyzed here. We employed the MacVector genome analysis program (version 7.0; Oxford Molecular Group), utilizing a human codon bias, to identify all ORFs within the HHV-5 AD169 genome that encode polypeptides that are at least 50 amino acids long. We also analyzed the genome by using the TESTCODE algorithm (12). These two filters identified ca. 700 ORFs encoding polypeptides >50 amino acids in size and starting with an AUG, as well as many more when the requirement for an N-terminal AUG was dropped (E. Murphy, I. Rigoutsos, and T. Shenk, unpublished data). We are currently utilizing the BDGF algorithm (33), a Bio-Dictionary-based gene finder, to validate our expectation that the HHV-5 genome contains substantially more coding ORFs than predicted in the original annotation.

ACKNOWLEDGMENTS

We thank Edward Mocarski for insightful comments on the manuscript.

This work was supported in part by grants from the National Institutes of Health to T.S. (CA82396, CA85786, and CA87661).

REFERENCES

1. **Alford, C. A., and W. J. Britt.** 1996. Cytomegalovirus, p. 1981–2010. *In* B. N. Fields, D. M. Knipe, and P. M. Howley (ed.), Fields virology, 3rd ed. Lippincott-Raven Publishers, Philadelphia, Pa.
2. **Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman.** 1990. Basic local alignment search tool. J. Mol. Biol. **5:**403–410.
3. **Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25:**3389–3402.
4. **Bairoch, A., and R. Apweiler.** 2000. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 2000. Nucleic Acids Res. **28:**45–48.
5. **Bergammi, G., M. Reschke, M. C. Battista, M. C. Boccuni, F. Campanini, A. Ripalti, and M. P. Landini.** 1998. The major open reading frame of the β2.7 transcript of human cytomegalovirus: in vitro expression of a protein post-transcriptionally regulated by the 5′ region. J. Virol. **72:**8425–8429.
6. **Bruick, R. K., and S. L. McKnight.** 2002. Oxygen sensing gets a second wind. Science **295:**807–808.
7. **Cha, T. A., E. Tom, G. W. Kemble, G. M. Duke, E. S. Mocarski, and R. R. Spaete.** 1996. Human cytomegalovirus clinical isolates carry at least 19 genes not found in laboratory strains. J. Virol. **70:**78–83.
8. **Chee, M., A. Bankier, S. Beck, R. Bohni, C. Brown, R. Cerny, T. Horsnell, C. Hutchinson III, T. Kouzarides, J. Martignetti, et al.** 1990. Analysis of the protein-coding content of the sequence of human cytomegalovirus strain AD169. Curr. Top. Microbiol. Immunol. **154:**125–169.
9. **Cinatl, J., Jr., R. Blaheta, M. Bittoova, M. Scholz, S. Margraf, J.-U. Vogel, J. Cinatl, and H. W. Doerr.** 2000. Decreased neutrophil adhesion to human cytomegalovirus-infected retinal pigment epithelial cells is mediated by virus-induced upregulation of Fas ligand independent of neutrophil apoptosis. J. Immunol. **165:**4405–4413.
10. **Dargan, D. J., F. E. Jamieson, J. MacLean, A. Dolan, C. Addison, and D. J. McGeoch.** 1997. The published DNA sequence of human cytomegalovirus strain AD169 lacks 929 base pairs affecting genes UL42 and UL43. J. Virol. **71:**9833–9836.
11. **Doolittle, R.** 1995. The multiplicity of domains in proteins. Annu. Rev. Biochem. **64:**287–314.
12. **Fickett, J. W.** 1982. Recognition of protein coding regions in DNA sequences. Nucleic Acids Res. **10:**5303–5318.
13. **Floratos, A., I. Rigoutsos, L. Parida, G. Stolovitzky, and Y. Gao.** 1999. Sequence homology detection through large-scale pattern discovery, p. 164–173. *In* Proceedings of the Third Annual ACM International Conference on Computational Molecular Biology (RECOMB '99). ACM, Lyon, France.
14. **Greenaway, P. J., and G. W. Wilkinson.** 1987. Nucleotide sequence of the most abundantly transcribed early gene of human cytomegalovirus strain AD169. Virus Res. **7:**17–31.
15. **Horn, F., J. Weare, M. W. Beukers, S. Hörsch, A. Bairoch, W. Chen, Ø. Edvardsen, F. Campagne, and G. Vriend.** 1998. GPCRDB: an information system for G protein-coupled receptors. Nucleic Acids Res. **26:**277–281.
16. **Jones, D. T., W. R. Taylor, and J. M. Thornton.** 1992. A new approach to protein fold recognition. Nature **358:**86–89.
17. **Kivirikko, K. I., and T. Pihlajaniemi.** 1998. Collagen hydroxylases and the protein disulfide isomerase subunit of prolyl 4-hydroxylases. Adv. Enzymol. Relat. Areas Mol. Biol. **2:**325–400.
18. **Mocarski, E. S.** 1996. Cytomegaloviruses and their replication, p. 2447–2492. *In* B. N. Fields, D. M. Knipe, and P. M. Howley (ed.), Fields virology, 3rd ed. Lippincott-Raven Publishers, Philadelphia, Pa.
19. **Mocarski, E. S., and C. T. Courcelle.** 2001. Cytomegaloviruses and their replication, p. 2629–2673. *In* D. M. Knipe, P. M. Howley, D. E. Griffin, R. A. Lamb, M. A. Martin, B. Roizman, and S. E. Straus (ed.), Fields virology, 4th ed., vol. 2. Lippincott-Raven Publishers, Philadelphia, Pa.
20. **Mocarski, E. S., M. N. Prichard, C. S. Tan, and J. M. Brown.** 1997. Reassessing the organization of the UL42-UL43 region of the human cytomegalovirus strain AD169 genome. Virology **239:**169–175.
21. **Novotny, J., I. Rigoutsos, D. Coleman, and T. Shenk.** 2001. *In silico* structural and functional analysis of the human cytomegalovirus (HHV5) genome. J. Mol. Biol. **310:**1151–1166.
22. **Pearson, W. R., and D. J. Lipman.** 1998. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA **85:**2444–2448.
23. **Parida, L., A. Floratos, and I. Rigoutsos.** 1999. An approximation algorithm for alignment of multiple sequences using motif discovery. J. Comb. Optim. **3:**247–275.
24. **Pass, R. F.** 2001. Cytomegalovirus, p. 2675–2705. *In* D. M. Knipe, P. M. Howley, D. E. Griffin, R. A. Lamb, M. A. Martin, B. Roizman, and S. E. Straus (ed.), Fields virology, 4th ed., vol. 2. Lippincott-Raven Publishers, Philadelphia, Pa.
25. **Rigoutsos, I., and A. Floratos.** 1998. Combinatorial pattern discovery in biological sequences: the *Teiresias* algorithm. Bioinformatics **14:**55–67.
26. **Rigoutsos, I., and A. Floratos.** 1998. Motif discovery without alignment or enumeration, p. 221–227. *In* Proceedings of the Second Annual ACM International Conference on Computational Molecular Biology (RECOMB '98). ACM, New York, N.Y.
27. **Rigoutsos, I., Y. Gao, A. Floratos, and L. Parida.** 1999. Building dictionaries of 1D and 3D motifs by mining the unaligned 1D sequences of 17 archaeal and bacterial genomes, p. 223–233. *In* Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology (ISMB '99). AAAI Press, Heidelberg, Germany.

28. **Rigoutsos, I., A. Floratos, L. Parida, Y. Gao, and D. Platt.** 2000. The emergence of pattern discovery techniques in computational biology. Metabolic Eng. **2:**159–177.
29. **Rigoutsos, I., A. Floratos, C. Ouzounis, Y. Gao, and L. Parida.** 1999. Dictionary building via unsupervised hierarchical motif discovery in the sequence space of natural proteins. Proteins Struct. Funct. Genet. **37:**264–277.
30. **Rigoutsos, I., T. Huynh, A. Floratos, L. Parida, and D. Platt.** 2002. Dictionary-driven protein annotation. Nucleic Acids Res. **30:**3901–3916.
31. **Romanowski, M. J., and T. Shenk.** 1997. Characterization of the human cytomegalovirus IRS1 and TRS1 genes: a second immediate early transcription unit within IRS1 whose product antagonizes transcriptional activation. J. Virol. **71:**1485–1496.
32. **Rosenkilde, M. M., M. Waldhoer, H. R. Luttichau, and T. W. Schwartz.** 2001. Virally encoded 7TM receptors. Oncogene **20:**1582–1593.
33. **Shibuya, T., and I. Rigoutsos.** 2002. Dictionary-driven microbial gene finding. Nucleic Acids Res. **30:**2710–2725.
34. **Sippl, M., and S. Weitkus.** 1992. Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations. Proteins Struct. Funct. Genet. **13:**258–271.
35. **Smith, J. A., and G. S. Pari.** 1995. Human cytomegalovirus *UL102* gene. J. Virol. **69:**1734–1740.
36. **Smith, T. F., and M. S. Waterman.** 1981. Identification of common molecular subsequences. J. Mol. Biol. **147:**195–197.
37. **Smyth, M. J., and J. A. Trapani.** 1998. The relative role of lymphocyte granule exocytosis versus death receptor-mediated cytotoxicity in viral pathophysiology. J. Virol. **72:**1–9.
38. **Stasiak, P. C., and E. S. Mocarski.** 1992. Transcription of the cytomegalovirus *ICP36* gene requires the alpha gene product TRS1 in addition to IE1 and IE2. J. Virol. **66:**1050–1058.