Whole-proteome phylogeny of large dsDNA virus families by an alignment-free method

Guohong Albert Wu[a,b], Se-Ran Jun[a], Gregory E. Sims[a,b], and Sung-Hou Kim[a,b,1]

[a]Department of Chemistry, University of California, Berkeley, CA 94720; and [b]Physical Biosciences Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720

Contributed by Sung-Hou Kim, May 15, 2009 (sent for review February 22, 2009)

The vast sequence divergence among different virus groups has presented a great challenge to alignment-based sequence comparison among different virus families. Using an alignment-free comparison method, we construct the whole-proteome phylogeny for a population of viruses from 11 viral families comprising 142 large dsDNA eukaryote viruses. The method is based on the feature frequency profiles (FFP), where the length of the feature (*l*-mer) is selected to be optimal for phylogenomic inference. We observe that (*i*) the FFP phylogeny segregates the population into clades, the membership of each has remarkable agreement with current classification by the International Committee on the Taxonomy of Viruses, with one exception that the mimivirus joins the phycodnavirus family; (*ii*) the FFP tree detects potential evolutionary relationships among some viral families; (*iii*) the relative position of the 3 herpesvirus subfamilies in the FFP tree differs from gene alignment-based analysis; (*iv*) the FFP tree suggests the taxonomic positions of certain ''unclassified'' viruses; and (*v*) the FFP method identifies candidates for horizontal gene transfer between virus families.

alignment-free genome comparison | feature frequency profile | horizontal gene transfer | whole-genome phylogeny | virus phylogeny

Phylogenetic and taxonomic studies of viruses have become increasingly important as more and more whole viral genomes are sequenced (1–4). Knowledge of viral taxonomy and phylogeny is useful for understanding the diversity and evolution of viruses not only within a viral family, but also among different viral families that may have a common origin (5). They also provide useful information in drug design against virally induced diseases (6).

One of the unusual aspects of viral genomes is that they exhibit high sequence divergence due to high mutation rate, genetic recombination, reassortment, horizontal gene transfer (HGT), gene duplication, and gene gain/loss (7, 8). A direct consequence of the high sequence divergence and relatively small number of genes in viruses is that the number of highly conserved genes among different viral families is very small or, sometimes, undetectable. For example, the relationship among different families of eukaryote large DNA viruses (LDV) has often been studied based on multiple sequence alignment of a single gene, the DNA polymerase gene (9). Whether this single-gene based analysis can be used to properly infer viral species phylogeny is debatable.

Due to this and other limitations (10) of multiple sequence alignment comparison of 1 or a few selected viral genes, there has been a growing interest in alignment-free methods for whole-genome comparison and phylogenomic studies (11, 12). Alignment-free approaches have been used in the reconstruction of virus genome trees for individual virus families (13, 14) and across virus families. Examples of the latter include the composition vector method used to construct a genome tree for large dsDNA viruses (15), the average common substring approach used for phylogenomic analysis of the reverse-transcribing viruses and the negative-sense ssRNA viruses (16), and tetranucleotide usage patterns that have been found useful for inferring host-virus coevolution among bacteriophages and eukaryotic viruses (17). Besides genome trees, self-organizing maps (18) have also been used to understand the grouping of viruses.

In the previous alignment-free phylogenomic studies using *l*-mer profiles, 3 important issues were not properly addressed: (*i*) the selection of the feature length, *l*, appears to be without logical basis; (*ii*) no statistical assessment of the tree branching support was provided; and (*iii*) the effect of HGT on phylogenomic relationship was not considered. HGT in LDVs has been documented by alignment-based methods (19–22), but these studies have mostly searched for HGT from host to a single family of viruses, and there has not been a study of interviral family HGT among LDVs.

To address these issues, we have developed an alignment-free method using feature frequency profiles (FFPs) (23). In this work, we use the FFP method, supplemented by an HGT detection technique, to study the taxonomic grouping and phylogenomic relationship among subfamilies within each family, and phylogenomic relationship among 11 LDV families and 4 dsDNA insect viruses that have not yet been assigned to any virus family by the International Committee on the Taxonomy of Viruses (ICTV). Altogether, we analyze 142 complete LDV proteomes from National Center for Biotechnology Information's non-redundant RefSeq database (24).

Results and Discussion

We first present results on the whole-proteome tree reconstruction, including the choice of optimal feature length, and the identification of interviral-family HGT genes. To increase the sensitivity of the FFP method, we have applied 2 filtering schemes: the filtering of HGT candidate genes and the filtering of low-complexity features. Next, we describe the overall features of the LDV proteome tree, possible evolutionary relationship among families, and the differences between the FFP phylogeny and existing alignment-based phylogenies of several individual viral families. Finally, we compare the FFP tree to a previously published alignment-free analysis.

Optimal Feature Length. When whole proteomes are compared using *l*-mer FFP, different feature (*l*-mer) lengths can lead to different tree topologies. Thus, determining the optimal feature length is critical for phylogeny inference. Based on both cumulative relative entropy (CRE) and relative sequence divergence (RSD) analyses, the optimal feature length for LDV proteomes is determined to be 8 aa (see *Materials and Methods*). This estimate depends on the range of proteome sizes and the sequence divergence properties of the viruses (Fig. 1).

Horizontal Gene Transfer Between Viral Families. We use the Jensen–Shannon divergence (JS) (25) of pairwise FFPs to estimate the dissimilarity of 2 proteomes. JS provides a summary statistic of

Author contributions: G.A.W., G.E.S., and S.-H.K. designed research; G.A.W. performed research; G.A.W., S.-R.J., and G.E.S. contributed new reagents/analytic tools; G.A.W., S.-R.J., G.E.S., and S.-H.K. analyzed data; and G.A.W. and S.-H.K. wrote the paper.

The authors declare no conflict of interest.

[1]To whom correspondence should be addressed. E-mail: shkim@cchem.berkeley.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0905115106/DCSupplemental.

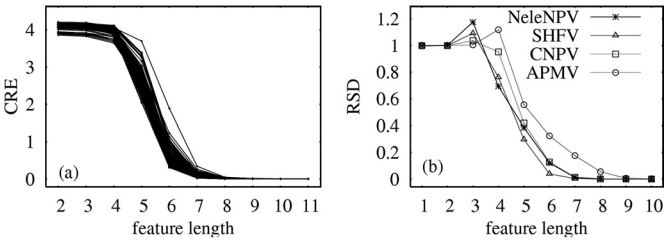

Fig. 1. Optimal feature (*l*-mer) length. (*A*) Cumulative relative entropy (CRE) curves for 142 large dsDNA virus proteomes. (*B*) Relative sequence divergence (RSD) values for 4 representative viral proteomes, the smallest (NeleNPV), the intermediate (SHFV and CNPV), and the largest (APMV). The optimal feature length for whole-proteome comparison and phylogeny inference is 8 and approximately corresponds to when both CRE and RSD fall to <10% of their maximum values.

given FFP pairs (see *Materials and Methods*), and to a first approximation, is a measure of the fraction of common features between 2 proteomes. Thus, JS can be dominated by 1 or more unusually similar genes as they may contribute the most number of shared features, and this can distort the tree topology. For viruses from different families, such genes can be considered as candidates for interfamily HGT and should be removed before constructing FFPs. The interfamily gene transfer may be the result of a direct viral gene transfer between 2 viruses while coinfecting the same host, or when 2 viruses capture the same cellular gene from their phylogenetically related hosts in 2 separate events. In either case, we assume that HGT events occurred more recently than viral speciation, thus, the HGT genes have much higher sequence similarity than other common genes between 2 compared viral families.

With our criteria for interfamily HGT detection (see *Materials and Methods* and Fig. 2), the total number of HGT instances is 164, consisting of 8 genes and distributed unevenly among viral families (Table S1). Six of the 8 genes are present in the poxviridae family, and all 6 have cellular homologues. Some of these 6 genes have been suspected to be captured from host (21, 22). The remaining two (*bro* and *hr* genes) are present in the insect-infecting baculoviruses and ascoviruses, and do not seem to have cellular homologues (26). None of the 8 genes is directly involved in the core viral activities of DNA replication and virus assembly. These 164 HGT proteins are excluded in FFP calculations and tree reconstruction.

Low Complexity Feature Filtering. Low complexity features are those 8-mers consisting of 1 or very few types of amino acids. They generally bear no or little phylogenetic signal and may lead to misleading phylogeny if not removed in the proteome tree reconstruction. For the LDV proteomes, 8-mers with $K_2 < 1.1$ are filtered out (see *Materials and Methods*).

FFP Proteome Tree of LDV Superfamily. After deleting the HGT candidate proteins and filtering out the low complexity features, the whole-proteome FFP tree is obtained for feature length 8 (Fig. 3). We use the invertebrate herpesvirus OsHV1 (the single member of Malacoherpesviridae) as the outgroup, because its proteome shows the greatest sequence divergence from the rest. A modified bootstrap resampling was used to estimate the robustness of the tree branching patterns (see *Materials and Methods*). Most viral families form monophyletic groups with high statistical support. One exception is that the mimivirus is mixed within phycodnaviruses and the 2 families form a monophyletic group with a moderate statistical support. Furthermore, the FFP tree shows subfamily divisions within a viral family, some of them do not agree with current alignment-based subdivisions (see below for individual families)

Relationship Among LDV Viral Families. A potential evolutionary relationship between families is also observed: The 2 families of iridovirus and ascovirus form a monophyletic group with high statistical support, in support of a gene-alignment based study (27); nudiviruses cluster with the baculovirus family with moderate support; and asfarvirus clusters with the poxvirus family with relatively weak support. Finally, the above-mentioned 6 viral families form a large monophyletic group with moderate statistical

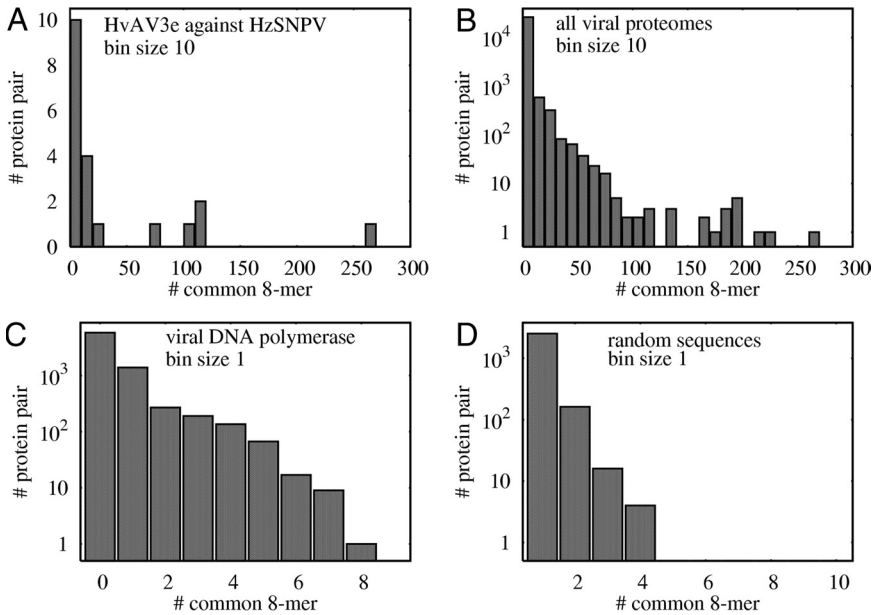

Fig. 2. Common 8-mers and HGT. The number of interviral-family protein pairs vs. the number of common 8-mers in a protein pair for LDVs. (*A*) The ascovirus HvAV3e proteome against the baculovirus HzSNPV proteome, suggesting that there are several protein pairs due to interfamily HGT events. (*B*) Interviral-family protein pairs from all LDV proteomes. (*C*) Interviral-family DNA polymerase pairs. (*D*) Same as in *B* but with each protein sequence subject to random permutation of its amino acids. Interfamily HGT candidates are identified when a protein pair shares unusually high number of common 8-mers relative to the most conserved LDV protein of DNA polymerase, with a maximum of eight 8-mers as shown in *C*. Randomized protein sequences share much fewer common 8-mers with a maximum of four 8-mers as shown in *D*.

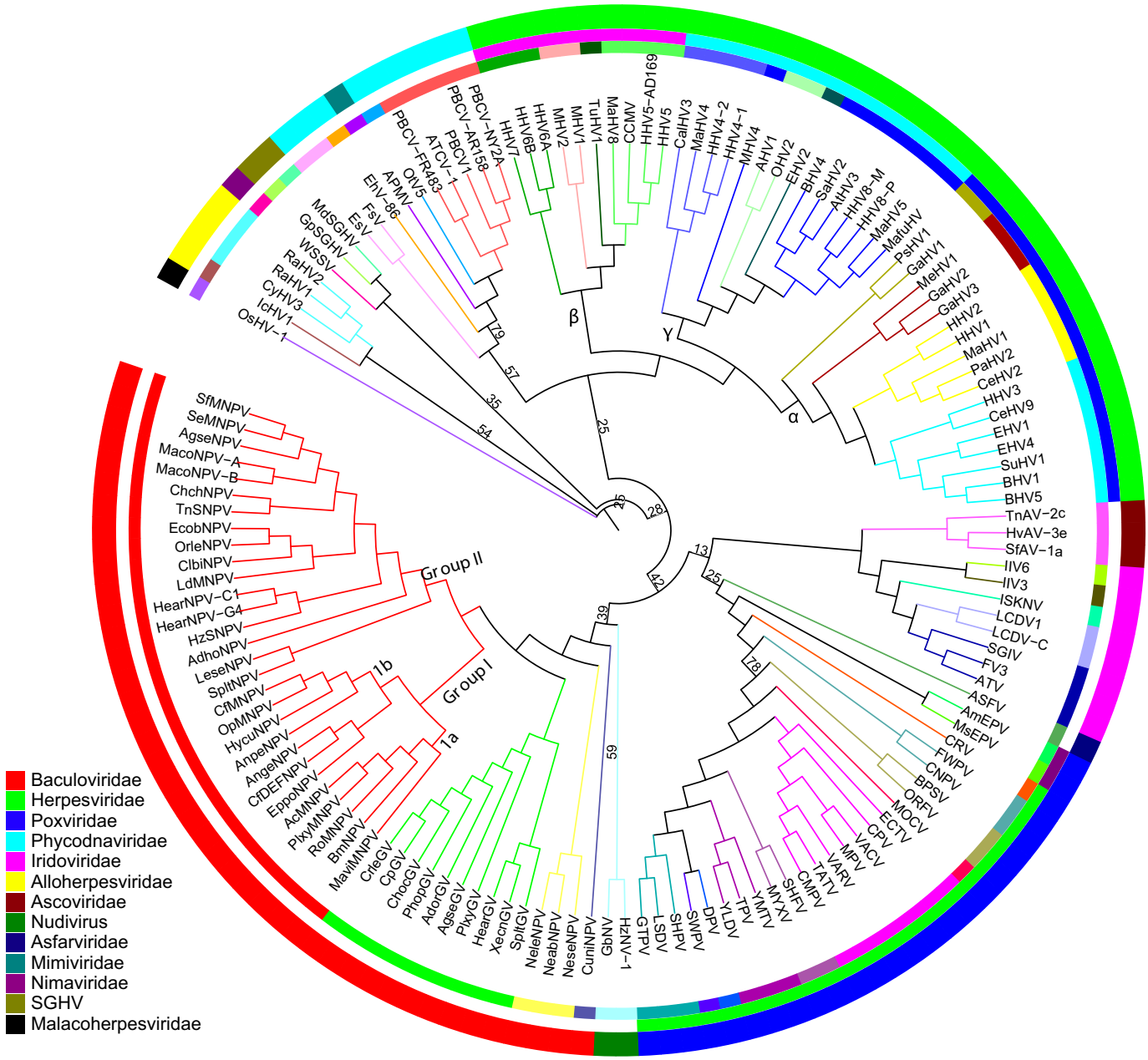

Fig. 3. The LDV whole-proteome tree. The FFP tree of large dsDNA viruses at feature length 8 after deleting horizontally transferred genes between viral families and filtering out low-complexity features. Modified bootstrap percentages <80% are shown and are based on 200 replicates. The tree is drawn using iTOL (48), and is not drawn to scale. Outer circle color-codes 11 viral families as per ICTV and 2 groups of viruses not assigned to any family: nudivirus and saliva gland hypertrophy virus (SGHV) (see key in the bottom left). The middle layer color-codes viral subfamilies of the poxviridae and herpesviridae. The different viral genera are color-coded by both the inner ring and tree leaves.

support. We also notice that the 3 herpesvirus families (herpesviridae, alloherpesviridae and malacoherpesviridae) are not related phylgenetically (see *Herpesviridae* below).

Below, we compare the FFP phylogeny of individual viral families to those based on sequence alignment.

Baculoviridae. The grouping of baculoviruses in the FFP tree (shown in red in the outer ring of Fig. 3) is consistent with the newly proposed 4-genera classification (28). Furthermore, the lepidopteran NPVs (shown in red in the inner ring of Fig. 3) can be divided into 2 monophyletic groups, the group I and group II NPVs, in agreement with a recent analysis based on sequence alignment of 29 core genes of the Baculoviridae (1). In particular, AcMNPV clusters with PlxyMNPV, RoMNPV and BmNPV within group I, in agreement with the 29-gene analysis (1). This grouping is in conflict with the analysis based on the single polyhedrin (*polh*) gene, which assigns AcMNPV to group II. This conflict was shown to result from recombination in the AcMNPV *polh* gene (29). At an even finer resolution, the division of group I NPVs into clade 1a and clade 1b also agrees with the 29-gene analysis (1). The remarkable agreement of the FFP-based baculovirus phylogeny with that of the 29-gene alignment-based analysis suggests that when a "large enough" number of genes are used, alignment-based and alignment-free methods converge for a given virus family. It is not clear what fraction of the genome/proteome can be considered "large enough" in alignment-based methods. Besides, when several viral families are compared, no or very few conserved genes may be common among them.

Herpesvirales. Herpesviruses are morphologically distinct from other viruses and they divide into 3 families under the recently established order Herpesvirales (30, 31), namely Herpesviridae, Alloherpesviridae, and Malacoherpesviridae. In the FFP tree, each family forms a clade, but the 3 families do not cluster to form a monophyletic group, indicating a lack of interfamily phylogenetic relationship at the sequence level despite of morphological similarities. The Herpesviridae clade further divides into 3 monophyletic subgroups corresponding to the α, β, and γ subfamilies with high statistical support. Of the 3 subfamilies, the β subfamily branches off first. This branching order is at variance with alignment-based analysis (31). The 4-member clade of the Alloherpesviridae shows moderate statistical support as a result of its great sequence divergence among the 4 viral proteomes, of which all but IcHV1 are currently not assigned at the genus level.

At the genus level, all except the rhadinovirus genus of the γ subfamily (shown in blue in inner ring of Fig. 3) are monophyletic. Within the rhadinovirus genus, the murid herpesvirus 4 (MHV4) proteome shows great sequence divergence and is separated from other members of the genus. Sequence alignment-based analysis also found that MHV4 has a particularly high level of sequence divergence, causing difficulties in determining its phylogenetic position unambiguously (32). The unclassified Tupaiid herpesvirus 1 (TuHV1) clusters with the cytomegalovirus genus of the β subfamily (shown in light green in the inner ring) in the FFP tree, although it may or may not be assigned to the same genus.

Phycodnaviridae and Mimiviridae. There are 9 phycodnaviruses and 1 mimivirus with complete proteomes in our dataset. Each multimember genus forms its own clade with high branch support. The recently sequenced marine green algae virus OtV5 (33) is not yet included in the ICTV 2008 Official Taxonomy, although sequence comparison of the DNA polymerase gene suggested that it belong to the genus prasinovirus (33). In the FFP tree, OtV5 is positioned next to the chlorovirus genus (shown in red in inner ring of Fig. 3), as is also the case with the DNA polymerase-based analysis.

The 9 phycodnaviruses do not form a monophyletic group in the FFP tree, because mimivirus (APMV) nests within them. However, all phycodnaviruses and the mimivirus together form a monophyletic group with moderate statistical support. Sequence alignment using the major capsid protein (34) or the DNA polymerase gene (35) found similar mixing between the mimivirus and phycodnaviruses. This is at variance with an earlier phylogenetic analysis suggesting that the mimivirus form a separate family (36). In the FFP tree, the mimivirus, OtV5, and the chlorovrius genus form a highly supported clade. Both the FFP tree and the recent sequence-alignment analyses show the high sequence divergences among the genera of Phycodnaviridae (37), suggesting a possible taxonomic revision of the Phycodnaviridae family (34, 38) and the mimivirus (35).

Poxviridae. The grouping of poxviruses in the proteome tree is consistent with the ICTV classification. The highly supported poxvirus clade falls into 2 monophyletic groups corresponding to the entomopoxvirinae and chordopoxvirinae subfamilies (middle ring, purple and green respectively), and the latter further divides into 3 monophyletic groups associated with reptilian, avian and mammalian hosts, respectively. Each genus forms a clade in the FFP tree. The branching order of different genera mostly agrees with an analysis based on alignment of a core set of 35 genes common to the chordopoxvirinae (39), although minor discrepancies also exist, for example, in the relative position of cervipoxvirus (DPV) and capripoxvirus (SHPV, GTPV, LSDV). In the FFP tree, the unclassified crocodile poxvirus (CRV) is the outgroup of the chordopoxvirinae clade and positioned next to the avipoxvirus genus (FWPV, CNPV). This suggests that CRV could be assigned to a new genus within the chordopoxvirinae subfamily.

Other viruses. There are 4 insect viruses that are not assigned to any viral family. Two (HzNV1 and GbNV) are nudiviruses, and they form a clade and cluster with the baculovirus family in the FFP tree, consistent with an analysis based on alignment of the DNA polymerase gene (40). The other two insect viruses causing salivary gland hypertrophy (MdSGHV and GpSGHV) form a clade with strong support, corroborating a recent finding that the two are related and form a distinct clade based on analysis of gene trees (41). They cluster with WSSV. The FFP tree also suggests that the 2 nudiviruses and the 2 SGHVs be separately assigned to 2 new viral families.

Comparison with Another Alignment-Free Method. In a previous report on the reconstruction of the whole-proteome phylogeny of large dsDNA viruses (15), the authors used an *l*-mer-based composition vector (CV) method with subtracted background "noise" modeled by a Markov chain estimator. Notable differences between the FFP tree and the CV tree are (*i*) the CV tree was based on *l*-mers of length 5, but the optimal feature length for FFP tree is 8; (*ii*) the CV tree did not explicitly deal with HGT among LDV families; (*iii*) the authors did not provide statistical assessment of branch support in the CV tree; (*iv*) neither baculoviruses nor iridoviruses are monophyletic in the CV tree; (*v*) the phycodnaviruses do not form a monophyletic group, with or without the mimivirus in the CV tree; and (*vi*) ascoviruses were not included in the CV tree, which could further distort the CV tree topology due to the extensive HGT between ascovirus and baculovirus.

FFP Method vs. Multiple Sequence Alignment (MSA) Method. MSA method has to select a set of highly conserved genes for alignment, and assumes that phylogeny of those selected genes represents species phylogeny. Thus, MSA can be applied only within individual families or for closely related families, and cannot be used for comparing diverse multiple families of LDVs. For inferring phylogeny of diverse families, FFP method has at least 3 advantages: (*i*) the whole genome/proteome is used to represent each species, (*ii*) it does not require selection of highly conserved genes common to all families, and (*iii*) it is not very sensitive to large-scale genome rearrangement and other changes including gene gain and loss.

On a more technical note, the presence of a common 8-mer between two proteins does not in general imply that the two proteins are homologous, and vice versa. This is illustrated in Fig. 2*D*, which shows that random sequences can have common 8-mers, and in Fig. 2*C*, which shows that there may be no common 8-mer between many protein pairs of DNA polymerase from different viral families. To make the distinction between distant and closely related viral species, we use 50 type species representing all of the LDV genera and find that only 53% of the interviral-family 8-mer-sharing protein pairs are homologous, after excluding HGT genes and low complexity features and using a blast E-value cutoff 0.01. In contrast, for intrafamily protein pairs, 8-mer conservation implies gene homology 95% of the time. However, even for the latter case, FFP and MSA, which use the whole proteome and a fraction of the proteome respectively, can give different phylogenies as exemplified by the branching order of the α, β, and γ subfamilies of the herpesviridae. These observations suggest that 8-mer conservation is not a useful measure for phylogenetic inference, but the profile of all 8-mers determines the FFP tree topology.

Conclusion

Using the alignment-free FFP method, we have studied the molecular phylogeny and horizontal gene transfer (HGT) between families for a broad population of large dsDNA eukaryote viruses consisting of 11 viral families. The unique aspects of this study include: (*i*) the selection of optimal feature length for phylogeny inference, (*ii*) a modified bootstrap support analysis of the branching orders in the FFP tree, and (*iii*) identification of interfamily HGT candidate genes and exclusion of the genes from the FFP tree reconstruction. The analysis of the FFP tree for the broad population of LDVs suggests that the method is suitable for grouping diverse families of virues, subgrouping within individual families,

finding possible evolutionary relationship among the families, and assigning "unclassified" species, even when there are no or few common genes among the broad population.

Materials and Methods

Dataset. The viral sequences were downloaded from National Center for Biotechnology Information's REFSEQ database (September 2008 release) (24). Protein sequences for large eukaryote dsDNA viruses are extracted from the .faa file. Polydnaviruses are excluded from consideration because they are a distinct group and hardly share any common genes with other virus families. The final dataset of 142 LDVs consists of 11 viral families and 4 insect viruses unassigned to any family. The list of viruses is included in Table S2.

Feature Frequency Profile (FFP) and Distance Matrix. A general description of FFP method is published in ref. 23. The feature frequency profile of a given sequence is obtained by counting all overlapping features of length *l* by sliding a window of width *l* along the sequence, advancing 1 letter at a time. The FFP of a proteome is the total sum of the FFPs for each protein sequence contained therein. In this work, we use the normalized FFP, i.e., the probability of occurrence of each word in a proteome. The dissimilarity between 2 FFPs can be estimated from the Jensen–Shannon divergence (JS) (25). For 2 probability distributions $P = (p_1, p_2, \ldots)$ and $Q = (q_1, q_2, \ldots)$, JS is given by

$$JS(P, Q) = \frac{1}{2} KL\left(P, \frac{P+Q}{2}\right) + \frac{1}{2} KL\left(Q \frac{P+Q}{2}\right), \quad [1]$$

where KL(*P*, *Q*) is the Kullback–Leibler divergence (42) or relative entropy

$$KL(P, Q) = \sum_i p_i \log_2 \frac{p_i}{q_i}, \quad [2]$$

and the summation is over all features. Note that JS is bounded between 0 and 1. Strictly speaking, JS is not a distance metric, because it does not satisfy the triangle inequality. However, this violation happens only for short feature lengths and is of no concern to us. For a given feature length *l*, the distance matrix for a collection of proteomes is constructed from all pairwise JSs.

Relative Sequence Divergence (RSD), Cumulative Relative Entropy (CRE), and Optimal Feature Length. Two methods exist for estimating the optimal feature length for whole-genome phylogeny. The first is related to information theory and makes use of cumulative relative entropy (CRE) of individual proteomes. By contrast, the second method estimates the relative sequence divergence (RSD) of a proteome relative to a random sequence of the same size by comparing their relatedness (in terms of FFP) to a group of proteomes. Both methods give the same estimate for LDVs.

CRE. This method estimates the minimal feature length for which the information content of a proteome can be approximated by its FFP. This is done by requiring the CRE between the FFP of a proteome and that of a Markov chain estimator to be small. Under a Markov chain model of order *l-2*, the expected *l*-mer frequencies of a sequence or proteome is given by frequencies of features of lengths *l-1* and *l-2* as follows (43),

$$\tilde{f}_{a_1 \ldots a_l} = \frac{f_{a_2 \ldots a_l} * f_{a_1 \ldots a_{l-1}}}{f_{a_2 \ldots a_{l-1}}}, \quad [3]$$

where *f* denotes observed feature-frequencies of a proteome, a_i denotes amino acid type at position *i* of a feature. The difference between the estimated and observed *l*-mer frequencies can be measured by the relative entropy KL(P_l, $\tilde{P}_l$), where $\tilde{P}_l$ and P_l are estimated and observed probability vectors of *l*-mers respectively. This difference as a function of feature length exhibits a peak, whose position can be estimated using random sequences (zero-order Markov chains) and is well approximated by

$$l_{peak} \cong \log_{20} N + 1, \quad [4]$$

where the base 20 is the number of amino acid types and *N* is the proteome size.

A monotonically decreasing function can be constructed for the cumulative relative entropy (CRE),

$$\mathrm{CRE}(l) = \sum_{1 \geq l} KL(P_i, \tilde{P}_i). \quad [5]$$

The minimal feature length at which CRE(*l*) approaches zero can be used iteratively to infer approximate frequencies of increasingly longer features, and is defined as the optimal feature length for phylogeny inference. For a group of divergent sequences like LDVs, this is approximately given by

$$l_{\mathrm{CRE}} \sim 2 \log_{20} N, \quad [6]$$

where *N* denotes the largest proteome size. For LDVs, the largest proteomes (i.e., mimivirus and phycodnaviruses) give $l_{\mathrm{CRE}} \approx 8$. This estimate is confirmed in Fig. 1*A*, where CRE values from Eq. **5** are plotted for all LDVs against feature length, and they all approach zero at feature length 8, with the largest proteome of the mimivirus (APMV) as the main determining factor.

RSD. This method requires that, on average, a biological sequence shares more features than a random sequence of the same length with a group of biosequences. For a group of *n* related biological sequences, the relative sequence divergence (*RSD*) for a biological sequence s_i at feature length *l* with $i = 1..n$ can be defined as

$$RSD(s_i, l) = \frac{\sum_{j \neq i} c(r_i, s_j, l)}{\sum_{j \neq i} c(c_i, s_j, l)}, \quad [7]$$

where $c(s_i, s_j, l)$ denotes the number of common feature of length *l* between sequences s_i and s_j. r_i denotes a random sequence of zero-order Markov chain with the same length as s_i. For short feature lengths ($l < l_{\mathrm{peak}}$), nearly all possible features are used by both the random sequence and viral proteomes, and the *RSD* is approximately 1. For longer feature lengths ($l > l_{\mathrm{peak}}$), the feature space is sparsely sampled, with all of the viral proteomes sampling one region and the random sequence a different region. As feature length increases, the overlap in feature space between the viral proteomes and random sequence becomes smaller and the RSD decreases to zero. Optimal feature length for phylogeny inference is obtained when *RSD* becomes much smaller than 1.

In Fig. 1*B*, the *RSD's* are plotted for 4 representative LDV proteomes including the smallest (NeleNPV), the largest (APMV), and intermediate (SHFV and CNPV), and they all fall <0.05 at feature length 8 and longer. Thus, both *RSD* and *CRE* analyses give $l = 8$ as the optimal feature length of the LDV proteomes. With longer feature lengths, *RSD* and *CRE* become even smaller, but the average number of shared features between viral proteomes (especially distantly related ones) becomes fewer and the resulting tree topology is less robust.

Interfamily HGT Candidates. HGT between viral families can cause some distortion of the tree topology, because JS can be biased by the few highly similar genes shared between 2 viruses as measured by the number of common 8-mers. For LDV proteomes at the optimal feature length $l = 8$, the distribution of common 8-mers in a protein pair is illustrated in Fig. 2. In particular, Fig. 2*B* shows the results from pairwise comparison of all proteins from different viral families, and Fig. 2*D* shows the same comparison after the amino acids in each protein sequence are randomly permuted. From Fig. 2*D* we infer that a protein-pair from our dataset can share up to 4 different 8-mers by chance. Fig. 2*C* plots the number of common 8-mers from DNA polymerase pairs between viral families, and the maximum number of shared 8-mers is 8. Thus, a protein pair from different viral families that share unusually high number of 8-mers relative to the DNA polymerase protein, which is common to all members, are candidates for HGT. For example, as shown in Fig. 2*A*, the unusually large number of common 8-mers present in protein pairs from the ascovirus HvAV3e and the baculovirus HzSNPV suggests direct or indirect HGT events between the 2 viruses.

To see the effect of using different HGT cutoffs (i.e., number of shared 8-mers) on LDV phylogeny, we compare tree topologies with cutoffs ranging from 6 to 40. We observe that the tree topology remains stable for a HGT cutoff in the range 13–31 (Fig. S1). For this work, we use a conservative HGT cutoff of 20, and identified 164 HGT instances consisting of 8 genes (Table S1).

Filtering Out Low Complexity Features. Features with low complexity generally bear no or little phylogenetic signal and could distort the tree topology if enough of them are present in the viral proteomes. One measure of feature complexity is the Shannon entropy

$$K_2 = -\sum_i \frac{n_i}{l} \log_2 \frac{n_i}{l}, \quad [8]$$

where *i* runs over the 20 aa types, n_i is the occurrence frequency of amino acid type *i* in a given feature, and *l* is the feature length. This and another closely related

EVOLUTION

complexity measure K_1 were used to detect and exclude regions of low complexity in amino acid sequences (44) during sequence alignment. For 8-mers, K_2 takes on values between 0 and 3, corresponding to using 1 and 8 aa types respectively.

The effect of using different low complexity cutoffs on phylogenetic tree reconstruction is illustrated in Fig. S2. Note that even excluding only the least complex features (i.e., homo 8-mers) causes appreciable change in the tree topology. For K_2 between 0 and 1.5, we observe that the tree topology is most stable for cutoffs 0.9, 1.1, and 1.3. Based on this analysis, we filter out 8-mer features with $K_2 < 1.1$ for this study. These features account for 0.3% of the viral proteomes on average, and up to a maximum of 2% for the EhV86 proteome. By way of comparison, for random sequences with equal usage of different amino acid types, the fraction of 8-mers with $K_2 < 1.1$ is $<10^{-5}$. The compositional types of these low complexity features include A_8, A_xB_{8-x} ($x = 1$–4), and $A_6B_1C_1$, where A, B, and C denote different amino acid types.

Phylogenetic Tree Reconstruction and Robustness Test. Phylogenetic trees are constructed from distance matrices using BIONJ (45). Robustness of the tree topology is estimated using a modified version of the bootstrap method (46), which works as follows. A table is first constructed with each row representing 1 viral proteome and each column representing 1 feature present in a viral proteome. Each table element indicates the feature frequency in a proteome (zero if absent). The bootstrap is applied to the columns of the table except that columns that are redrawn are treated as drawn only once (i.e., each column is either present or absent in the bootstrapped table). Thus, the resampled table has fewer columns but each feature maintains the same frequency as in the original table. This procedure is equivalent to a jackknife test deleting 1/e (i.e., 37%) of the features. A new distance matrix is then calculated for the resampled table. We use 200 replicates to estimate the branch support for the un-bootstrapped tree. For the LDV dataset, a significant proportion of the features are unique to only 1 proteome, thus the resampling is expected to underestimate the branch support. We have taken this and other factors (47) into consideration when making phylogenetic inferences.

ACKNOWLEDGMENTS. We thank Drs. B. Glausinger, L. Volkman, and M. Strand for their expert advice. This work was supported by National Institutes of Health Grant GM62412 and Korean Ministry of Education, Science and Technology World Class University Project Grant R31-2008-000-10086-0.

1. Herniou EA, Jehle JA (2007) Baculovirus phylogeny and evolution. *Curr Drug Targets* 8:1043–1050.
2. Montague MG, Hutchison CA, 3rd (2000) Gene content phylogeny of herpesviruses. *Proc Natl Acad Sci USA* 97:5334–5339.
3. McLysaght A, Baldi PF, Gaut BS (2003) Extensive gene gain associated with adaptive evolution of poxviruses. *Proc Natl Acad Sci USA* 100:15655–15660.
4. de Andrade Zanotto PM, Krakauer DC (2008) Complete genome viral phylogenies suggests the concerted evolution of regulatory cores and accessory satellites. *PLoS ONE* 3:e3500.
5. Iyer LM, Aravind L, Koonin EV (2001) Common origin of four diverse families of large eukaryotic DNA viruses. *J Virol* 75:11720–11734.
6. Marra MA, et al. (2003) The Genome sequence of the SARS-associated coronavirus. *Science* 300:1399–1404.
7. Shackelton LA, Holmes EC (2004) The evolution of large DNA viruses: Combining genomic information of viruses and their hosts. *Trends Microbiol* 12:458–465.
8. Duffy S, Shackelton LA, Holmes EC (2008) Rates of evolutionary change in viruses: Patterns and determinants. *Nat Rev Genet* 9:267–276.
9. Fauquet CM (2005) *Virus Taxonomy*: *Classification and Nomenclature of Viruses*: *Eighth Report of the International Committee on the Taxonomy of Viruses* (Elsevier, San Diego).
10. Wong KM, Suchard MA, Huelsenbeck JP (2008) Alignment uncertainty and genomic analysis. *Science* 319:473–476.
11. Vinga S, Almeida J (2003) Alignment-free sequence comparison—a review. *Bioinformatics* 19:513–523.
12. Hohl M, Ragan MA (2007) Is multiple-sequence alignment required for accurate inference of phylogeny? *Syst Biol* 56:206–221.
13. Stuart G, Moffett K, Bozarth RF (2004) A whole genome perspective on the phylogeny of the plant virus family Tombusviridae. *Arch Virol* 149:1595–1610.
14. Yang AC, Goldberger AL, Peng CK (2005) Genomic classification using an information-based similarity index: Application to the SARS coronavirus. *J Comp Biol* 12:1103–1116.
15. Gao L, Qi J (2007) Whole genome molecular phylogeny of large dsDNA viruses using composition vector method. *BMC Evol Biol* 7:41.
16. Ulitsky I, Burstein D, Tuller T, Chor B (2006) The average common substring approach to phylogenomic reconstruction. *J Comp Biol* 13:336–350.
17. Pride DT, Wassenaar TM, Ghose C, Blaser MJ (2006) Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses. *BMC Genomics* 7:8.
18. Gatherer D (2007) Genome signatures, self-organizing maps and higher order phylogenies: A parametric analysis. *Evol Bioinf* 3:211–236.
19. Monier A, Claverie JM, Ogata H (2007) Horizontal gene transfer and nucleotide compositional anomaly in large DNA viruses. *BMC Genomics* 8:456.
20. Filee J, Pouget N, Chandler M (2008) Phylogenetic evidence for extensive lateral acquisition of cellular genes by Nucleocytoplasmic large DNA viruses. *BMC Evol Biol* 8:320.
21. Hughes AL, Friedman R (2005) Poxvirus genome evolution by gene gain and loss. *Mol Phylogenet Evol* 35:186–195.
22. Bratke KA, McLysaght (2008) A Identification of multiple independent horizontal gene transfers into poxviruses using a comparative genomics approach. *BMC Evol Biol* 8:67.
23. Sims GE, Jun SR, Wu GA, Kim SH (2009) Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proc Natl Acad Sci USA* 106:2677–2682.
24. Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 35:D61–65.
25. Lin J (1991) Divergence measures based on the Shannon entropy. *IEEE T Inform Theory* 37:145–151.
26. Bideshi DK, et al. (2003) Phylogenetic analysis and possible function of bro-like genes, a multigene family widespread among large double-stranded DNA viruses of invertebrates and bacteria. *J Gen Virol* 84:2531–2544.
27. Stasiak K, et al. (2003) Evidence for the evolution of ascoviruses from iridoviruses. *J Gen Virol* 84:2999–3009.
28. Jehle JA, et al. (2006) On the classification and nomenclature of baculoviruses: A proposal for revision. *Arch Virol* 151:1257–1266.
29. Jehle JA (2004) The mosaic structure of the polyhedrin gene of the Autographa californica nucleopolyhedrovirus (AcMNPV). *Virus Genes* 29:5–8.
30. Davison AJ, et al. (2009) The order Herpesvirales. *Arch Virol* 154:171–177.
31. McGeoch DJ, Rixon FJ, Davison AJ (2006) Topics in herpesvirus genomics and evolution. *Virus Res* 117:90–104.
32. McGeoch DJ, Gatherer D, Dolan A (2005) On phylogenetic relationships among major lineages of the Gammaherpesvirinae. *J Gen Virol* 86:307–316.
33. Derelle E, et al. (2008) Life-cycle and genome of OtV5, a large DNA virus of the pelagic marine unicellular green alga Ostreococcus tauri. *PLoS ONE* 3:e2250.
34. Larsen JB, Larsen A, Bratbak G, Sandaa RA (2008) Phylogenetic analysis of members of the Phycodnaviridae virus family, using amplified fragments of the major capsid protein gene. *Appl Environ Microbiol* 74:3048–3057.
35. Monier A, et al. (2008) Marine mimivirus relatives are probably large algal viruses. *Virol J* 5:12.
36. Raoult D, et al. (2004) The 1.2-megabase genome sequence of Mimivirus. *Science* 306:1344–1350.
37. Dunigan DD, Fitzgerald LA, Van Etten JL (2006) Phycodnaviruses: A peek at genetic diversity. *Virus Res* 117:119–132.
38. Allen MJ, Schroeder DC, Holden MT, Wilson WH (2006) Evolutionary history of the Coccolithoviridae. *Mol Biol Evol* 23:86–92.
39. Lefkowitz EJ, Wang C, Upton C (2006) Poxviruses: Past, present and future. *Virus Res* 117:105–118.
40. Wang Y, Kleespies RG, Huger AM, Jehle JA (2007) The genome of Gryllus bimaculatus nudivirus indicates an ancient diversification of baculovirus-related nonoccluded nudiviruses of insects. *J Virol* 81:5395–5406.
41. Garcia-Maruniak A, et al. (2009) Two viruses that cause salivary gland hypertrophy in Glossina pallidipes and Musca domestica are related and form a distinct phylogenetic clade. *J Gen Virol* 90:334–346.
42. Kullback S, Leibler RA (1951) On information and sufficiency. *Ann Math Stat* 22:79–86.
43. Sadovsky MG (2003) Comparison of real frequencies of strings vs. the expected ones reveals the information capacity of macromoleculae. *J Biol Phys* 29:23–38.
44. Wootton JC, Federhen S (1993) Statistics of local complexity in amino acid sequences and sequence databases. *Comp Chem* 17:149–163.
45. Gascuel O (1997) BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol* 14:685–695.
46. Felsenstein J (1985) Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* 39:783–791.
47. Alfaro ME, Zoller S, Lutzoni F (2003) Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. *Mol Biol Evol* 20:255–266.
48. Letunic I, Bork P (2007) Interactive Tree Of Life (iTOL): An online tool for phylogenetic tree display and annotation. *Bioinformatics* 23:127–128.