Distinctive features of large complex virus genomes and proteomes

Jan Mrázek[†] and Samuel Karlin[‡§]

[†]Department of Microbiology and Institute of Bioinformatics, University of Georgia, Athens, GA 30602; and [‡]Department of Mathematics, Stanford University, Stanford, CA 94305

Communicated by Samuel Karlin, Stanford University, Stanford, CA, January 18, 2007 (received for review September 25, 2006)

More than a dozen large DNA viruses exceeding 240-kb genome size were recently discovered, including the "giant" *mimivirus* with a 1.2-Mb genome size. The detection of *mimivirus* and other large viruses has stimulated new analysis and discussion concerning the early evolution of life and the complexity and mechanisms of evolutionary transitions. This paper presents analysis in three contexts. (*i*) Genome signatures of large viruses tend to deviate from the genome signatures of their hosts, perhaps indicating that the large viruses are lytic in the hosts. (*ii*) Proteome composition within these viral genomes contrast with cellular organisms; for example, most eukaryotic genomes, with respect to acidic residue usages, select Glu over Asp, but the opposite generally prevails for the large viral genomes preferring Asp more than Glu. In comparing Phe vs. Tyr usage, the viral genomes select mostly Tyr over Phe, whereas in almost all bacterial and eukaryotic genomes, Phe is used more than Tyr. Interpretations of these contrasts are proffered with respect to protein structure and function. (*iii*) Frequent oligonucleotides and peptides are characterized in the large viral genomes. The frequent words may provide structural flexibility to interact with host proteins.

mimivirus | genome signature | frequent oligonucleotides and peptides

The last half decade has witnessed a surge of discovery and preliminary characterizations of complex large viruses. These include 11 dsDNA viruses (completely sequenced) exceeding 240-kb genome length, featuring the "giant" 1.2-Mb *mimivirus*, four *Phycodnaviridae* ranging in size from 300 kb to 410 kb, three bacterialphages, two poxviruses, and one chimpanzee herpesvirus (1, 2). All these viruses have no RNA stage. Table 1 describes the 11 viruses, indicating their five letter abbreviations, overall G + C content, full names, genomic DNA lengths, major viral hosts, and classifications based on the NCBI taxonomy database. More than 1,600 viruses and phages have been sequenced to date but the *mimivirus* is remarkable with a repertoire of 911 protein encoding genes, of which ≈80% are tentatively exclusive to *mimivirus* (1, 3). The *mimivirus* exceeds in genomic DNA size at least 20 established bacterial genomic sequences (1). The detection of *mimivirus* and other large viruses has raised challenging issues concerning the evolution of life and mechanisms of evolutionary transitions (2, 4).

A special issue of the journal *Virus Research* proffers several review and discussion articles on the evolution of complexity in the viral world. For example, it is proposed in the Forterre article (5), that RNA viruses arose early among living organisms and contributed critically to evolutionary developments including the creation of DNA viruses, DNA replication mechanisms, formations and bifurcations of the three major domains of life (Bacteria, Archaea and Eukaryotes), and the origin of the eukaryotic nucleus. The ssRNA nidoviruses of animals (6) and the closteroviruses of plants (7) are among the large viruses not studied in this article. Apart from the 11 viruses of Table 1, many more large viruses with complete genomic sequences are expected soon. The large eukaryotic viruses include the nucleocytoplasmic dsDNA viruses (NCDLV), which subsume the highly malleable poxviruses and the phycodnaviruses, the latter infecting algae and protists. The virus-host interactions are diverse, and it is speculated that many of these mechanisms were later transferred to cellular organisms. Most viruses and phages previously studied are viewed as "simple" (distinguished from "complex").

We investigate *in silico* the large viruses of Table 1 in three contexts: (*i*) genome signature comparisons and interpretations of dinucleotide and tetranucleotide relative-abundance representations among the viruses and hosts, (*ii*) contrasts of proteome composition and codon biases within these viral genomes and (*iii*) characterizations of frequent oligonucleotides and peptides in these viral genomes.

Dinucleotide Relative Abundance Values. Dinucleotide biases of a DNA sequence are evaluated through the odds ratio $\rho_{XY} = f_{XY}/f_X f_Y$, where f_{XY} is the frequency of the dinucleotide XY in the genome contig sequence under study and f_X is the frequency of the nucleotide X. For double-stranded DNA sequences, a symmetrized version $\{\rho^*_{XY}\}$ is calculated from frequencies of the sequence concatenated with its inverted complementary sequence. Our studies of DNA and genomic data have demonstrated that the dinucleotide relative abundance profiles $\{\rho^*_{XY}\}$ evaluated for disjoint 50-kb DNA contigs from the same organism are approximately constant throughout its genome and are generally more congruent to each other than they are to those from 50-kb contigs of different organisms (8–12). On this basis, we refer to the vector profile $\{\rho^*_{XY}\}$ of a given genome as its "genomic signature" diagnostic of different groups of organisms.

Based on statistical analysis and extensive sequence data, the dinucleotide XY is said to be underrepresented if $\rho^*_{XY} \leq 0.78$ and overrepresented if $\rho^*_{XY} \geq 1.23$ (13). Biochemical experiments measuring nearest-neighbor frequencies have established that the set of dinucleotide biases is a remarkably stable property of the DNA of an organism (e.g., 10, 12). An assessment of the genomic difference between two sequences f and g from different organisms (or from different regions of the same genome) is the average dinucleotide absolute relative-abundance difference calculated as

$$\delta^* = \delta^*(f, g) = (1/16)\ \Sigma_{XY} \mid \rho^*_{XY}(f) - \rho^*_{XY}(g) \mid$$

where the sum extends over all dinucleotides. This dinucleotide relative-abundance measure of distance between DNA sequences appears to provide meaningful measures of similarities. Several factors that can impact DNA structures and the genomic signature include dinucleotide stacking energies, curvature, superhelicity, methylation, oligonucleotide modifications, and context-dependent mutation biases (12). Mechanisms for the evolution and maintenance of the genomic signature are unknown, although data suggest that genome-wide processes, including DNA replication, recombination, and repair (11, 12, 14), contribute intrinsically to the genomic signature, either by prefer-

Author contributions: J.M. and S.K. wrote the paper.

The authors declare no conflict of interest.

[§]To whom correspondence should be addressed. E-mail: karlin@math.stanford.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0700429104/DC1.

Table 1. Eleven dsDNA large viruses (>240-kb genome length), no RNA stage

Abbreviation	Full name	G + C frequency	Size, bp	Possible hosts	
APMiV	Mimivirus	25.95	1,181,404	*Acanthamoeba polyphaga* *Entamoeba histolytica*	ENTHI
Nimaviridae					
SWSSV	Shrimp white spot Syndrome virus	41.01	305,107	Shrimp	
Phycodnaviridae					
EmHuV	Emiliania huxleyi virus 86	40.18	407,339	*Emiliania huxleyi*	
PBChV	Paramecium bursaria, Chlorella virus 1	39.96	330,743	Green algae	
ESi1V	Ectocarpus siliculosus virus 1	51.74	335,593	Brown algae	
Herpesviridae					
ChCMV	Cytomegalovirus	61.70	241,087	Chimpanzee	ptr22
Pox viridae					
CPoxV	Canarypox virus	30.39	359853	Chicken	gag15
FPoxV	Fowl virus	30.88	288539	Wild bird	
Bacterial Phage					
SSMph	Cyanophage P-SSM2	35.50	252,401	*Prochlorococcus*	pnmed
FKZph	Pseudomonasphage phi KZ	36.82	280,334	*Pseudomonas aeruginosa*	pseae
KVPph	Vibriophage KVP40	42.60	244,834	*Vibrio parahaemolyticus*	vpa1

entially generating or selecting specific dinucleotides in the DNA and RNA modifications and through local DNA and RNA structures (base-step conformational propensities).

It is useful for later comparisons to provide references of δ^* distances between different organisms. For random sequences, generally $\delta^* \leq 15$ (distances are multiplied by 1,000). Generally $20 < \delta^* < 40$ reflect "close" similarity, e.g. cow vs. pig; "moderate" similarity, $45 < \delta^* < 70$ (e.g., human vs. chicken and *Escherichia coli* vs. *Salmonella typhimurium*); "weak" similarity, $80 < \delta^* < 105$ (e.g., human vs. trout and human vs. sea urchin); "distantly" similar, $115 < \delta^* < 135$ (e.g., human vs. *Sulfolobus*); "far," $140 < \delta^* < 175$ (e.g., human vs. *Drosophila*); "very far," $\delta^* > 180$ (e.g., human vs. *E. coli*).

Viral Dinucleotide Relative Abundance Extremes. It was shown with few exceptions that small eukaryotic viral genomes (≤30-kb genome size) are CpG suppressed (significantly underrepresented), generally to the same extent as the host genome (8, 9, 15, 16). Small plant viruses are also inclined to be CpG suppressed. In contrast, no consistent CpG suppression is observed in large viral genomes (16). Interestingly, temperate phages, including λ, Mu, P1, P4, P22, exhibit high relative abundance values of GpC and TpA. Virtually all coliphages have low representations of the tetranucleotides CTAG and GATC (11). In the lytic coliphages (e.g., T_2, T_4, T_3, T_7) and the *Bacillus subtilis* phageϕ29, CGCG is underrepresented as is TATA in the filamentous (parasitic) phages (M13, I22, IKE, PF3). This suppression may be related to restriction system avoidance. Temperate phages tend to be moderately similar to their host or hosts (11). On the other hand, the δ^* distance of lytic phage to their host vary from weak similarity to far exemplified among the phage (T_2, T_4, T_3, T_7) where T_2 and T_4 are weakly similar to *E. coli* and T_3 and T_7 are far to very far, the phageϕ29 is very far from its host *B subtilus* (11). Temperate coexistence shows dinucleotide bias profiles similar to those of the host. The genome signatures of

Table 2. Average δ^* distances among the genomes of large viruses and selected hosts

APMiV	SWSSV	SSMPh	FKZPh	SD1Ph	EmHuV	PBChV	ESilV	ChCMV	CPoxV	FPoxV	KVPPh	pmmed	pseae	vpa_1	gag15	ptr22	enthi	Species
35	96	122	124	124	129	111	151	165	203	194	181	174	164	179	192	166	89	APMiV
	24	80	105	105	177	118	124	126	163	163	169	107	131	141	132	116	73	SWSSV
		46	92	93	173	154	178	147	139	147	207	139	178	187	139	119	91	SSMPh
			23	20	113	124	142	106	111	101	170	170	136	128	172	172	138	FKZPh
				24	113	124	141	105	111	101	169	170	136	128	173	172	138	SD1Ph
					43	103	137	130	169	163	146	242	151	131	212	221	187	EmHuV
						29	86	101	140	129	127	187	140	124	226	220	147	PBChV
							41	77	151	170	92	203	99	85	214	218	154	ESilV
								37	99	118	95	182	87	96	214	216	166	ChCMV
									36	58	186	211	161	171	242	237	209	CPoxV
										36	197	212	166	179	249	243	206	FPoxV
											35	236	117	93	231	246	192	KVPPh
												30	179	187	149	119	143	pmmed
													26	97	191	208	169	pseae
														31	179	201	176	vpa_1
															48	75	137	gag15
																39	123	ptr22
																	40	enthi

Average δ^* distances (multiplied by 1,000) between two disjoint 50-kb samples drawn from two different genomes (or the same genome, diagonal entries) are shown. Species name abbreviations are defined in Table 1. Only hosts with sufficient DNA sequences available are included. *Entamoeba histolytica* was used in place of the mimivirus apparent natural host *Acantamoeba polyphaga*, because the latter does not have sufficient sequences available. Genomes composed of multiple chromosomes are represented by a single chromosome.

Table 3. Di- and tetra-nucleotides extremes in large viral and several host genomes

Species	Underrepresentations	Overrepresentations
APMiV	GC 0.74	—
SWSSV	CG 0.64, TA 0.72	—
	CCGG 0.76	—
EmHuV	AG 0.70	—
PBChV	CTAG 0.76	—
ESi1V	TA 0.58	—
ChCMV	CTAG 0.59	—
CPoxV	—	—
FPoxV	—	—
SSMph	CGCG 0.00, GCGC 0.00	CCGA 1.39
	CCGG 0.00, CG 0.40	GGCA 1.34
	GGCC 0.42, AGCT 0.42	GCGA 1.30
	AGCC 0.58, CATG 0.63	AGGC 1.26
	TCGA 0.64, GTAC 0.76	ACGG 1.25
	ACGT 0.76	
PKZph	—	—
KVPph	GGCC 0.22, CCGG 0.55	GGGA 1.44
	CCCC 0.60, CC 0.64	GACC 1.29
	TA 0.70, CTCC 0.75	CG 1.24
	GATC 0.75	—
enthi	CG 0.35	
	CCGG 0.73	
	GC 0.73	
pmmed	CG 0.50	CC 1.27
	AC 0.72	
pseae	CTAG 0.23	CTAC 1.40
	TA 0.54	TTAA 1.38
		ATAG 1.31
vpa1	TA 0.71	GC 1.24
gag 15	CG 0.29	CA 1.30
	TA 0.66	
ptr 22	CG 0.23	CGCC 1.25
	TA 0.71	

Dashes indicates no cases. Species name abbreviations are defined in Table 1.

bacterial plasmids are also generally moderately similar to the host genome signature (17).

The ρ^* Vector Profile and δ^* Distances Among the Large dsDNA Viruses

Table 2 displays the δ^* distances among the large viruses and with the host genomes. The within-genome δ^* distances are pervasively <50 (mostly <40),underscoring the invariance of the genome signature. As expected, the two vertebrate hosts (chicken and chimpanzee genomes) score $\delta^* = 75$, indicating moderate similarity and the two γ proteobacterial hosts *Pseudomonas* (PSEAE) and VIBRIO are weakly similar. Table 3 shows the relatively few dinucleotide biases among the large viruses, a property valid for the bulk of small phages as set forth elsewhere (11). The paucity of extremes in the genome signature putatively allow the viruses to invade several types of hosts with few incompatibilities.

Here, we highlight and interpret several observations embedded in Table 2.

1. The two poxviruses verify a δ^* distance of 58, signifying they are moderately similar.
2. The shrimp white spot virus (SWSSV) genomic signature is closest to ENTHI among the hosts of Table 2. The *mimivirus* first found in *Acanthamoeba polyphaga* (1) is also closest ($\delta^* = 89$) to ENTHI. The phage SSMph also records a δ^* distance of 91 to ENTHI. All other large viruses are far or very far from ENTHI, with δ^* distances exceeding 138.
3. The *mimivirus* and SWSSV show a δ^* distance of 96, implying that the genomic sequences are weakly similar.
4. The two algae viruses ESi1V (host brown algae) and PBChV (host green algae) yield a δ^* distance of 86 of moderate to weak similarity. The brown algae virus ESi1V is also moderately similar to the chimpanzee cytomegalovirus ($\delta^* = 77$).
5. The *mimivirus* is very far from the two pox viruses ($\delta^* = 203$, 194) and from the Vibriophage KVP40 ($\delta^* = 181$).
6. As indicated previously, temperate phages are generally moderately similar to their respective host ($\delta^* < 70$) (cf. 11). On this basis, we would interpret the evaluations of Table 2 to imply that the large viruses are marginally to potently lytic relative to their respective hosts. From this perspective, *mimivirus* is expected to be lytic; and in fact, the infected amoeba is lysed within 24 h after infection (18), analogous to the relationship of phage T4 with *E. coli* (11).

Proteome Composition of the Large DNA Viruses

The average amino acid and nucleotide codon preferences in the *mimivirus* genome are provided in supporting information (SI) Tables 7–9. For each of the large viruses, Table 4 displays the highly abundant amino acid types, attaining $\geq 8\%$ usages on average. Three of the viruses show no abundant amino acid usages. The almost exclusive abundant residue in bacterial, archaeal and eukaryotic genomes is Leucine (data not shown), which is less valid in large viral genomes (Table 4). The least-used amino acids (Table 5) in large eukaryotic viruses tends to be Trp, as in most eukaryotic genomes contrasted with prokaryotic genomes where Cys is the least-used, on average, of the order 0.5% to 1.5%. In a consistent manner, Cys is the least-used amino acid in the three large phage genomes (Table 5). Thus,

Table 4. Amino acid frequency usages ≥8% (on average) and principal codon preferences

Species	Amino acid usage, %; principal codon preference				
APMiV	N 8.91; AAT	I 9.89; AAT	L 8.18; TTA	K 9.01; AAA	
SWSSV	L 9.06; TTG, TTA	S 9.98; TCT			
EmHuV	None >8		S (7.52); TCT		
PBChV	None >8		L (7.78); TTG		
ESi1V	L 8.06; CTC, CTG	S 8.20; TCG	A (7.73); GCC, GCG		
ChCMV	L10.18; CTG	A (7.93); GCC	R (7.76); CGC	S (7.80); TCG, TCC	
CPoxV	I 9.66; ATA	L 9.23; TTA	K 8.40; AAA	N (7.61); AAT	S (7.90); TCT
FPoxV	I 10.02; ATA	L 9.13; TTA	K 8.22; AAA	S (7.92); TCT	
SSMph	G 8.50; GGT				
FKZph	L 8.43; TTA				
KVPph	None >8		L (7.62); CTT, TTG		

Amino acids of usage between 7.50% and 8.00% are displayed in parentheses. Species name abbreviations are defined in Table 1.

MICROBIOLOGY

Table 5. Least amino acid usages among the large viruses

Species	W, %	C, %	M, %	H, %
APMiV	0.74	1.81	1.87	2.06
SWSSV	0.94	1.93	2.42	2.00
EmHuV	1.10	1.84	2.70	2.54
PBChV	1.11	1.86	2.21	2.15
ESi1V	0.97	2.06	2.44	2.44
ChCMV	1.43	2.42	1.90	3.01
CPoxV	0.67	2.26	2.28	1.99
FPoxV	0.68	2.26	2.25	2.02
FKZph	1.31	0.98	2.31	2.03
KVPph	1.37	1.26	2.35	2.32
SSMph	1.16	1.07	1.60	1.84

W, Trp; C, Cys; M, Met; H, His. Species name abbreviations are defined in Table 1.

amino acid usages of viruses and phages tend to mirror that of their primary host.

The prevalent codons are of the type RNY (R, purine; Y, pyrimidine; N, any nucleotide), as with most living organisms (19, 20). At codon site 1, we observe $R_1 > 60\%$ (the average purine frequency at codon site 1), except for ChCMV and SSMph, which report $R_1 = 53.9\%$ and $R_1 = 54.1\%$, respectively (SI Tables 8 and 9). The third codon site features nucleotide T for all G + C poor genomes and nucleotide C for G + C rich genomes as follows: APMiV $T_3$45.9, CpoxV $T_3$39.2, FpoxV T_3 37.7, EmHuV T_3 34.3, PBChV T_3 31.4, SWSSV T_3 31.4, FKZph T_3 43.7, KVPph T_3 32.7, SSMph T_3 46.6, ChCMV C_3 46.1, ESi1V C_3 33.3.

The nucleotide A is predominant at codon site 1. Two genomes (SSMph and ChCMV) do not use purines >60%. The third codon site mostly selects a pyrimidine nucleotide that is T, except for ChCMV and ESi1V, where C is strongly preferred. The second codon site selects a balanced purine or pyrimidine nucleotide with frequency ≈50% in the range 47% to 53%. The difference A2 − T2 (fractions of A and T at the second codon positions) is especially large for the *mimivirus* conveying a reduced hydrophobicity level for its proteome in contrast with the other large viruses.

The amino acid usage implications of Table 6 are as follows.

1. The average acidic (Glu + Asp) charge and the average basic (Lys + Arg) charge differ among the large DNA viruses (excepting phage) by at most 0.40%, indicating a near balance in their charge ambience. This balance presumably entails less conflict in charge interactions with the host, which may facilitate invasion and cohabitation of alternative eukaryotic cells; this also helps maintain capsid proteins near neutral. The large phage genomes contrast in this respect, carrying acidic charge dominating basic charge by almost 2% (Table 6). The phage FKZph and KVPph genomes are among the highest in average acidic residue usages exceeding most bacterial genomes in this respect (data not shown).
2. Unlike most bacterial and eukaryotic genomes with respect to acidic amino acid usages, the large viruses primarily select Asp over Glu. What does this mean? Glu is most favored in protein secondary structures (especially in α helices), whereas Asp is confined structurally largely to coils and loops. By contrast, Asp is prominent at protein active sites (e.g., in serine proteases), contributes more often than Glu in metal ligation (e.g., for Fe^{2+}, Cu^{2+}, Zn^{2+}, and Mn^{2+}) and functions decisively in two-component systems. We may interpret these contrasts to the effect that Asp in large viruses is generally preferred to Glu presumably for functional objectives, whereas Glu serves better in protein conformational secondary structures. A reduced number of protein α helices in large DNA viruses may facilitate their import and export into cells and host compartments.
3. The amino acid usages of Phe compared with Tyr in bacterial and eukaryotic genomes predominantly favor Phe over Tyr (data not shown). However, among archaeal genomes, Crenarchaea species favor Tyr over Phe with the preference for Phe over Tyr in most euryarchaeal genomes. Among current complete eukaryotic genomes, again Phe is generally of higher usage compared with Tyr (data not shown). By contrast, most of the large viruses, including *mimivirus*, prefer Tyr over Phe (Table 6). We interpret the usage of Phe over Tyr as better suited for structural objectives, whereas Tyr is better suited to ensure functional needs. In summary, the large viruses tend to prefer Asp over Glu and Tyr over Phe with protein function being the primary objective.
4. The mean hydrophobicity frequency Φ is reported in Table 6. The two poxviruses carry the highest overall hydrophobicity frequency among the large viruses. Most large viruses have an average hydrophobicity frequency Φ of 27–28%, which is the usual average hydrophobicity frequency in bacterial proteomes.

Frequent Oligonucleotides and Peptides of the Large Viral Genomes

Mimivirus. The annotation of the *mimivirus* genome predicts 911 protein coding genes (1), 450 putatively encoded in one strand and 461 encoded from the complementary strand. Suhre *et al.* (3) detected in the *mimivirus* genome, distributed intergenically, the highly frequent 8-bp oligonucleotide AAAATTGA in conjunction with its inverted complement TCAATTTT, totaling 403 of these sequence elements. Based on the distributions of these sequences with respect to consecutive genes, they propose that AAAATTGA

Table 6. Major charge and hydrophobic amino acid usage comparisons among large viruses

Species	E + D, %	R + K, %	E vs D usages, %	Y vs F usages, %	Median Φ frequency,* %
APMiV	12.3	12.3	D 6.8 > E 5.6	Y 5.4 > F 4.6	27.6
SWSSV	11.7	11.9	D 5.1 < E 6.6	Y 2.7 < F 4.9	27.9
EmHuV	10.8	10.5	D 5.6 > E 5.1	Y 4.2 > F 3.5	27.8
PBChV	10.1	12.2	D 5.1 ≈ E 5.1	Y 3.7 < F 5.5	29.9
ESi1L	12.3	12.1	D 6.3 > E 5.9	Y 2.8 ≈ F 2.8	26.9
ChCMV	10.5	10.7	D 5.0 < E 5.5	Y 3.4 < F 3.7	27.0
CPoxV	12.4	12.6	D 6.4 > E 5.9	Y 5.7 > F 3.9	31.1
FPoxV	12.0	12.4	D 6.4 > E 5.6	Y 5.1 > F 4.3	30.4
FKZph	13.1	11.1	D 6.7 > E 6.4	Y 4.4 > F 4.0	29.0
KVPph	14.4	11.7	D 6.8 < E 7.6	Y 4.2 < F 4.3	27.6
SSMph	12.8	10.3	D 6.7 > E 6.1	Y 4.2 > F 4.1	25.3

E, Glu; D, Asp; R, Arg; K, Lys; Y, Tyr; F, Phe. *, Φ = Ile + Val + Leu + Met + Phe, the total frequency of the major hydrophobic amino acids. Species name abbreviations are defined in Table 1.

is a core promoter motif analogous to the TATA-box of eukaryotic genomes. In this context, we apply our method for detecting significant frequent words and *r*-scan analysis for the ascertainment of clustering, over-dispersion, or evenness in the distribution of a marker array (21, 22). The markers at hand are the sequence elements AAAATTGA and its inverted complement.

The Method of Frequent Words and *r*-Scan Analysis. To identify frequent oligonucleotides and peptides, a ball-in-urn model is used (21, 23). Urns correspond to all DNA words of a given size, and balls refer to the observed words in the given sequence. For a sequence of length L, comprised of letters drawn from an alphabet of size A, there is a natural word length s defined by the inequality $A^{s-1} \leq L < A^s$, and a natural copy number r determined by the inequality $(r-1)/r < (\log L)/\log A^s \leq r/(r+1)$. For sufficiently large L of a random sequence, the number of words of length s that occur more than r times is approximately Poisson distributed (21). In the case of a strongly biased sequence, such as an A + T rich genome, the method can be generalized to words w that have individual frequencies p_w. The word size s is the same as defined previously, but the copy threshold, r_w, for a particular word is now determined to satisfy $(P_wL)^{r_w}\exp(-p_wL)/r_w! \leq 1/L$ (22). When the foregoing inequality is satisfied, at most one frequent word is expected in a random sequence of the same length. For a Markov model, with transition probabilities f_{ii+1} between the ith and $(i+1)$th letters, $p_w = f_1 f_{1,2} \ldots f_{s-1,s}$. For the *Haemophilus influenzae* genome, L $\approx$ 1.8 Mb, a nucleotide alphabet $A = 4$ gives a frequent word length of $s =$ 11 bp. Invoking the copy threshold determination reveals that, among others, the 11 mers containing the uptake signal sequence AAGTGCGGT and its inverted complement qualify as frequent words (22). For the cyanobacterial genome *Synechocystis*, the palindrome GGCGATCGCC is frequent and significantly evenly distributed (14). In *E. coli*, REP elements and Chi sites qualify as frequent words. On the amino acid level, the pentapeptides corresponding to the ATP binding domain distinguish frequent peptides. In eukaryotes, amino acid frequent words include peptides characteristic of zinc fingers (CGKAF, CEECG, one letter amino acid code used), of chymotrypsin proteases (LTAAH, GDSGGP), of kinases (ADFGL, FGQGT), and of homeobox proteins (FQNRR, HFNRY).

Frequent "Promoter" Elements of Mimivirus. Concentrating on the aggregate of the non coding regions of the *mimivirus* genome the natural length for a frequent word is $s = 9$. The most frequent DNA words with a length of 9 across the totality of noncoding sequence is TCAATTTTT (173 occurrences), and the next most frequent is its inverted complement AAAAATTGA (171). If we restrict attention to 8-mers as Suhre *et al.* (3) have done, we find that AAAATTGA (212) occurrences and its complement TCAATTTT (231) total 443 markers. Note that these 8-mers are not the most frequent in the genome as AAAAAATT and AATTTTTT register 252 and 235 copies, respectively but that the AAAATTGA/TCAATTTT words exhibit a stronger excess over the expected counts based on the Markov model.

The *r*-scan distribution of the markers TCAATTTT and AAAATTGA selected by Suhre *et al.* (3) as core promoter elements across the *mimivirus* noncoding sequences were examined. A significant cluster was revealed, embracing six close repeats of AAAATTGA preceded by a copy of TCAATTTT. The cluster for $r = 6$, comprised of seven markers, is contained in the segment 326665–326767 (the marker elements are underlined) as follows: 326641 ATATATTTAT ATGTATTATT TACATCAATT TTTCTACACA AAATTGATTC ACGAAAATTG ACTCACAAAA TTGATTAATA AAATTGATTC ACAAAATTTA TTAATAAAAA TTGATTCATA AAATTGATTC ACGAAAAACT TAAACAGACA. The flanking genes are MIMLL258 (thymidine kinase) starting at 326609 and translated backwards to 325932, and MIMLR259 (unknown function), starting at 326823 and translated forward to 327701; thus, the repeat cluster lies between divergent genes putatively controlled by the same cluster. A smaller significant cluster is contained in the interval 813246–813278. The sequence is 813241 AAAAAAAAAT TGAAAAAAAA AAATTGAAAA AAAATTGAAA AAAATAATCA. The flanking genes are MIMLL615 (unknown function) from 813154 translated backwards to 811049, and MIMLR616 (unknown function) from 813381 translated forward to 814442 (again a pair of divergent genes).

How should one interpret extant close multiple copies of a regulatory element? Multiple close promoter sequences may enhance or contrariwise reduce the influence and activity of the promoter elements. Multiple copies of the classical TATA-box sequence in eukaryotes are generally not present in situations of gene control.

Frequent Oligonucleotides in the Large Virus Genomes. The top frequent oligonucleotides over the complete mimivirus genome (relevant word size $s = 11$) are parts of imperfect tandem iterations of a 9-bp word GGTGATAAA and its inverted complement TTTATCACC. These frequent words occur exclusively in genes annotated "collagen triple helix repeat containing protein" and translate into imperfect tandem repeats of the tripeptide Gly-Asp-Lys (allowing some replacements of Asp and Lys with other residues including Glu, Asn, Ile and Thr, Val, Asp, respectively). Collagen repeats are not unusual in large viral genomes. For example, herpesvirus saimiri protein STP-C488 contains 18 tandem copies of the "classical" collagen motif Gly-Pro-Pro, which is important for the virus' capacity to transform the host cells (24). The collagen repeats in the *mimivirus* genome are more elaborate (often exceeding 50 copies), the repeated tripeptide differs from the "standard" collagen motif of the Gly-Pro-Pro consensus, and several *mimvirus* proteins contain multiple Gly-Asp-Lys repeats. The collagen repeat in the herpesvirus saimiri STP-C488 protein associates with host proteins (25), and possibly the repeats in the *mimivirus* proteins assume similar roles and interact with critical host proteins.

Chimpanzee Cytomegalovirus (ChCMV). Most frequent oligonucleotides are imperfect trinucleotide repeats, based mainly on the trinucleotides GCC and CGG. These trinucleotide repeats occur mostly in genes and translate into short runs of Gly (57 copies of G_4), Ser (59 copies of S_4), Ala (52 copies of A_4), and fewer, but many runs of Pro, Thr, Gln, Glu, His, and Asp. In human, proteins with multiple long amino acid runs are often associated with genetic diseases (26). The ChCMV possesses five proteins with multiple extended amino acid runs including UL44 (processivity subunit of DNA polymerase), UL69 (posttranscriptional regulator of gene expression), UL80 (protease and a minor scaffold capsid protein), UL112 (contributing in recruiting DNA replication proteins to a nuclear replication compartment), and UL122 (immediate-early transcriptional regulator). These refer primarily to regulatory proteins where the amino acid runs may provide structural flexibility for interactions with host proteins (cf. 27).

Canarypox Virus (CPoxV). The most frequent oligonucleotides are parts of two major motifs, TAACGAGTGTTACGAG and its inverted complement, which establish tandem repeats near both ends of the genome. Similar tandem repeats are also found in the fowlpox virus. The second motif, TCAGGACTCTT and its inverted complement, is distributed in several small clusters 2436–3517 (five copies), 191367–192011 (seven copies), 239648–239915 (four copies), 283909–283997 (three copies) and 357095–357147 (five copies).

Emiliania Huxleyi Virus 86 (EmHuV). Because the genome of EmHuV extends 407339 bp, the natural frequent oligonucleotide length is 10 bp. The most frequent 10-mers are CTCCACCGCC (97 copies), CGCCATCTCC (83), CCGCCATCTC (81), and TCCACCGCCA (83). The top frequent words arise from imperfect tandem repeats based on the trinucleotides CCN or NGG (several with lengths

in excess of 50 bp), which translate into proline-rich stretches. In analogy to the amino acid runs of ChCMV and collagen repeats in APMiV, these proline-rich sequences may contribute to structural flexibility facilitating interactions with several host proteins (27). Another conspicuous motif is CCAGTTCCCTAAA and its inverted complement, which combined occur 45 times exclusively in the region 222352–290187. These motifs correspond to a subclass of the "family A" repeats proposed to function as promoters (28).

Large Phage Genomes. The frequent words in the large phage genomes do not reach the same number of copies as in the viral genomes projecting lower "repetitiveness" (or higher complexity) of the phage genomes. The top frequent oligonucleotides of the SSMPh genome arise from TSS tandem repeats (S stands for G or C), which constitute several strong clusters in the region 21000–28500. This region contains two phage tail fiber-like proteins and several proteins of unknown function that contain proline and glycine rich segments encoded by the TSS repeats.

Frequent Peptides in the *Mimivirus* Genome. The natural peptide size in the amino acid alphabet over the 911 protein ensemble is s = 5. The highest frequent pentapeptides include DKGDK (one letter amino acid code used) 198 copies, GDKGD 335, KGDKG 422, and further frequent words have copy number 91 and fewer. Thus, the top peptide frequent words over the coding genome arise largely as imperfect tandem repeats of GDK, which also give rise to frequent oligonucleotides and were discussed previously.

In addition to the triple helix repeat motifs, the *mimivirus* is replete in other amino acid repeats. More than 25% of all proteins of *mimivirus* contain frequent peptides. We highlight several examples. The most frequent peptides apart from the collagen repeats include VKYLV of 91 occurrences and VVKYL of 69 occurrences. In the BroA protein (gene designation MIML_L37) we obtain the peptide ESSDN (three tandem copies). The MIML_L111 gene contains a translation to seven copies of the peptide HRDND. The gene MIML_L116 encodes seven copies of ERSRYRSP, six in a single tandem repeat and one separate. The genes MIML_L175, MIML_L176, MIML_L177, MIML_L179, MIML_L181, and MIML_L182 all encode in their midst the heptapeptide MAQLLID. The protein encoded by the gene MIML_L357 contains seven tandem copies of the hexapeptide QMNPMN. All these genes are of unknown function, and most contain other frequent peptides as well.

Frequent Peptides Among the Other Large Viruses. For genome sizes 240–450 kb, the natural peptide lengths have s = 4. The most frequent peptides emphasize runs of the amino acids G, P, and S, proline-rich and glycine-rich motifs, or tandem repeats as follows: ChCMV, top frequent peptides S_4 59 copies, G_4 57, A_4 52, P_4 39, T_4 39, and E_4 24; EmHuV, P_4 496 copies, P_3S 372, P_2SP 357, PSP_2 393, SP_3 329, and others <75; PBChV, PAPK 143, APKP 142, KPAP 139, PKPA 135, and others <32; SWSSV, S_4 423, E_4 192, Q_4 97, A_4 115 D_4 79, P_4 71, G_4 69, and L_4 73; CpoxV, TPLH 57, GADV 39, and GADI 37 (note that the top repeats in the two poxviruses are identical, attesting to the close similarity of the two genomes); FpoxV, TPLH 38, GADV 17, and GADI 22; ESi1V, G_4 79, Q_4 32, S_4 34, P_4 24, and E_4 24; and SSMph, GPTG 59, PTGP 48, P_2GP 38, GPAG 42, and GP_2G 49.

Many of these frequent peptides also qualify as frequent oligonucleotides (see above). The two phages FKZph and KVPph have essentially no frequent peptides.

Conclusion

The proteomes of the completely sequenced large DNA viruses (≥240 kb), including the "giant" *mimivirus* possess a near neutral charge distribution verifying, on average, acidic charge frequency and basic charge frequency approximately commensurate $|\,(\mathrm{Glu}+\mathrm{Asp})\% - (\mathrm{Lys}+\mathrm{Arg})\%\,| \leq 0.3\%$ [SI Table 7]. Also, a paucity of dinucleotide or tetranucleotide relative-abundance extremes (Table 3) is paramount. This regular genome signature could imply that the invasion and cohabitation of large viruses for multiple host species is viable. However, the genome signature distances (δ^* distances, see Table 2) among the viruses and their potential hosts tend to be strongly deviant, suggesting that the large viruses are lytic in their hosts (*cf.* 11).

Two anomalies of amino acid usages in the large viruses are prominent. (*i*) Whereas most eukaryotic genomes with respect to acidic residue usages select Glu over Asp, the opposite prevails for the large viral genomes generally preferring Asp over Glu (SI Table 8). (*ii*) In almost all bacterial and eukaryotic genomes, Phe is used more than Tyr, probably because Phe serves mostly in protein structural capacities, whereas Tyr contributes more to functional objectives. By contrast, in most of the large virus genomes, Tyr dominates Phe in residue usages.

Many of the frequent oligonucleotides and peptides may contribute to "disordered" segments in the associated protein structure. Such disordered segments frequently occur in proteins that function as hubs in protein interaction networks, i.e., that can interact with many different proteins (27). From this perspective, the viral and phage proteins containing such regions characterized by the presence of many frequent words may be structurally flexible proteins that interact with multiple host proteins and allow the virus to survive better and/or affect functions of the host cells. At variance with bacterial phage, most large viruses are abundant with repeats. The *mimivirus* is especially replete with frequent oligonucleotides and peptides.

We thank Prof. A. M. Campbell for valuable discussions on the topics of large viruses.

1. Raoult D, Audic S, Robert C, Abergel C, Renesto P, Ogata H, La Scola B, Suzan M, Claverie JM (2004) *Science* 306:1344–1350.
2. Koonin EV, Dolja VV (2006) *Virus Res* 117:1–4.
3. Suhre K, Audic S, Claverie JM (2005) *Proc Natl Acad Sci USA* 102:14689–14693.
4. Claverie, JM (2006) *Genome Biol* 7:110–115.
5. Forterre P (2006) *Virus Res* 117:5–16.
6. Gorbalenya AE, Enjuanes L, Ziebuhr J, Snijder EJ (2006) *Virus Res* 117:17–37.
7. Dolja VV, Kreuze JF, Valkonen JP (2006) *Virus Res* 117:38–51.
8. Subak-Sharpe H, Burk RR, Crawford LV, Morrison JM, Hay J, Keir HM (1966) *Cold Spring Harbor Symp Quant Biol* 31:737–748.
9. Morrison JM, Keir HM, Subek-Sharpe H, Crawford LV (1967) *J Gen Virol* 1:101–108.
10. Russel GJ, Walker PMB, Elton RA, Subek-Sharpe JH (1976) *J Mol Biol* 108:1–28.
11. Blaisdell BE, Campbell AM, Karlin S (1996) *Proc Natl Acad Sci USA* 93:5854–5859.
12. Karlin S (1998) *Curr Biol* 1:598–610.
13. Karlin S, Cardon LR (1994) *Ann Rev Microbiol* 48:619–654.
14. Mrázek J, Bhaya D, Grossman AR, Karlin S (2001) *Nucleic Acids Res* 29:1590–1601.
15. Shpaer EG, Mullins JI (1990) *Nucleic Acids Res* 18:5793–5797.
16. Karlin S, Doerfler W, Cardon LR (1994) *J Virol* 68:2889–2897.
17. Campbell AM, Mrázek J, Karlin S (1999) *Proc Natl Acad Sci USA* 96:9184–9189.
18. Suzan-Monti M, La Scola B, Raoult D (2006) *Virus Res* 117:145–155.
19. Shepherd JCW (1981) *Proc Natl Acad Sci USA* 78:1596–1600.
20. Kypr J, Mrázek J (1987) *Int J Biol Macromol* 9:49–53.
21. Karlin S, MY, Leung (1991) *Ann Appl Probability* 4:513–538.
22. Karlin S, Mrázek J, AM, Campbell (1996) *Nucleic Acids Res* 24:4263–4272.
23. Karlin S (2005) *Proc Natl Acad Sci USA* 102:13355–13362.
24. Jung JU, Desrosiers RC (1994) *Virology* 204:751–758.
25. Lee H, Choi, JK, Li M, Kaye K, Kieff E, Jung JU (1999) *J Virol* 73:3913–3919.
26. Karlin S, Brocchieri L, Bergman A, Mrázek J, Gentles AJ (2002) *Proc Natl Acad Sci USA* 99:333–338.
27. Dunker AK, MS, Cortese, P Romero, LM, Iakoucheva, VN, Uversky (2005) *FEBS J* 272:5129–5148.
28. Wilson WH, Schroeder DC, Allen MJ, Holden MTG, Parkhill J, Barrell BG, Churcher C, Hamlin N, Mungall K, Norbertczak H, *et al.* (2005) *Science* 309:1090–1092.