JOURNAL OF VIROLOGY, Oct. 2009, p. 10719–10736 Vol. 83, No. 20
0022-538X/09/$08.00+0 doi:10.1128/JVI.00595-09

Overlapping Genes Produce Proteins with Unusual Sequence Properties and Offer Insight into De Novo Protein Creation▽†

Corinne Rancurel,[1] Mahvash Khosravi,[2] A. Keith Dunker,[2] Pedro R. Romero,[2]* and David Karlin[3]*

Architecture et Fonction des Macromolécules Biologiques, Case 932, Campus de Luminy, 13288 Marseille Cedex 9, France[1]; Center for Computational Biology and Bioinformatics, 410 West 10th Street, Suite 5000, Indiana University-Purdue University, Indianapolis, Indiana 46202-5122[2]; and 19, rue Ornano, 69001 Lyon, France[3]

Received 23 March 2009/Accepted 23 July 2009

It is widely assumed that new proteins are created by duplication, fusion, or fission of existing coding sequences. Another mechanism of protein birth is provided by overlapping genes. They are created de novo by mutations within a coding sequence that lead to the expression of a novel protein in another reading frame, a process called "overprinting." To investigate this mechanism, we have analyzed the sequences of the protein products of manually curated overlapping genes from 43 genera of unspliced RNA viruses infecting eukaryotes. Overlapping proteins have a sequence composition globally biased toward disorder-promoting amino acids and are predicted to contain significantly more structural disorder than nonoverlapping proteins. By analyzing the phylogenetic distribution of overlapping proteins, we were able to confirm that 17 of these had been created de novo and to study them individually. Most proteins created de novo are orphans (i.e., restricted to one species or genus). Almost all are accessory proteins that play a role in viral pathogenicity or spread, rather than proteins central to viral replication or structure. Most proteins created de novo are predicted to be fully disordered and have a highly unusual sequence composition. This suggests that some viral overlapping reading frames encoding hypothetical proteins with highly biased composition, often discarded as noncoding, might in fact encode proteins. Some proteins created de novo are predicted to be ordered, however, and whenever a three-dimensional structure of such a protein has been solved, it corresponds to a fold previously unobserved, suggesting that the study of these proteins could enhance our knowledge of protein space.

Since their discovery (76), overlapping genes, i.e., DNA sequences simultaneously encoding two or more proteins in different reading frames, have exerted a fascination on evolutionary biologists. Among several mechanisms, they can be created by a process called "overprinting" (43), in which a DNA sequence originally encoding only one protein undergoes a genetic modification leading to the expression of a second reading frame in addition to the first one (Fig. 1). The resulting overlap encodes an ancestral, "overprinted" protein region and a protein region created de novo (i.e., not by duplication) called an "overprinting" or "novel" region (Fig. 1). At present, it is widely thought that the creation of proteins de novo is very rare, contrary to their emergence by gene duplication, which is thought to be the major factor (for reviews, see references 55 and 94). However, this belief might actually reflect the fact that proteins created de novo are in general very difficult to identify (55). Indeed, a long-standing question is whether a protein that has no detectable homolog in other organisms (called an "orphan" protein or "ORFan" [27] or "taxonomically restricted" [110]) represents a protein created de novo in a particular organism or merely a protein that is a member of a larger family whose other members have diverged beyond recognition or have become extinct (115). Proteins created de novo by overprinting provide a valuable opportunity to address these questions, and this constitutes one of the two strands of our study.

Practically all studies of overlapping genes have been focused on evolutionary constraints and informational characteristics at the DNA level (see, e.g., references 46, 71, 75, 84, 85, and 114). However, very little has been done to assess potential effects of the overlap on the corresponding protein products. Two studies reported that overlapping proteins are enriched in amino acids with a high codon degeneracy (arginine, leucine, and serine) (68) and that they often simultaneously encode a cluster of basic amino acids in one frame and a stretch of acidic amino acids in the other frame (66).

The other strand of the present study is based on earlier observations of the overlapping gene set of measles virus (41), which suggested that protein regions encoded by overlapping genes might have a propensity toward structural disorder.

Structural disorder is an essential state of numerous proteins, in which it is associated mostly with signaling and regulation roles (21, 96, 111). The key feature of intrinsically disordered proteins (also called "unstructured" or "natively unfolded") is that under physiological conditions, instead of a particular three-dimensional (3D) structure, they adopt ensembles of rapidly interconverting structural forms. Different degrees of disorder exist, from random coils to molten globules (100), and some disordered regions can become ordered under certain conditions (21, 96, 117). A variety of computer programs have been developed to predict these regions (19, 23, 101). Each predictor typically differs in what kind of "disorder"

* Corresponding author. Mailing address for Pedro R. Romero: Center for Computational Biology and Bioinformatics, 410 West 10th Street, Suite 5000, Indiana University-Purdue University, Indianapolis, IN 46202-5122. Phone: (317) 278-4101. Fax: (317) 278-9201. E-mail: promero@compbio.iupui.edu. Mailing address for David Karlin: 19, rue Ornano, 69001 Lyon, France. Phone: 44 (0) 755 194 5984. Fax: 44 (0) 207 611 8254. E-mail: karlin.david@gmail.com.

† Supplemental material for this article may be found at http://jvi.asm.org/.

▽ Published ahead of print on 29 July 2009.

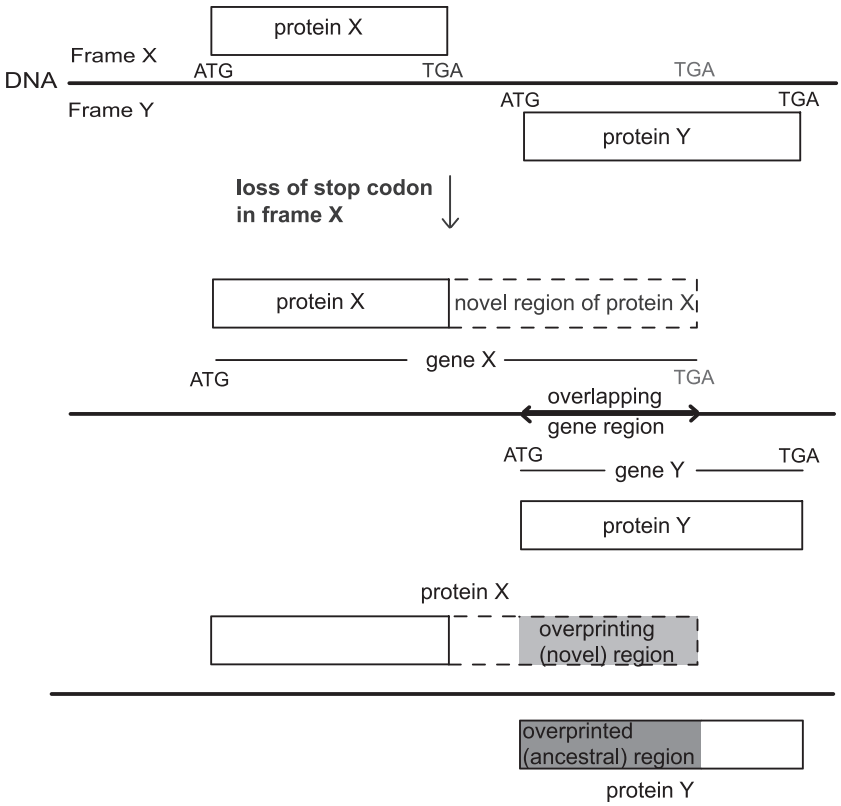

FIG. 1. Creation of a novel protein region (C-terminal extension) by overprinting. Top, a DNA sequence encodes two proteins in different reading frames. Notice the potential, unused stop codon downstream of protein X. Middle, a mutation abolishes the stop codon of protein X, causing its elongation ("overprinting") to the preexisting stop codon. This results in a gene overlap. Bottom, the overlap encodes an overprinted (ancestral) protein region (dark gray) and an overprinting (novel) protein region (light gray).

it identifies (23, 78), matching only some of the types of disorder mentioned above. Therefore, in order to choose a proper predictor, it was necessary to define precisely what kind of structural disorder we expected to find in proteins encoded by overlapping genes.

At least two nonexclusive hypotheses can explain why overlapping genes might encode disordered proteins: (i) the newly created (overprinting) protein of each overlap might tend to be disordered, and (ii) structural disorder in proteins encoded by overlapping genes might alleviate evolutionary constraints imposed on their sequence by the overlap. These hypotheses are clarified below.

Intuitively, the conditions required for a protein to fold into a stable 3D configuration, including sequence composition, periodicity, and complexity, are such that structurally ordered proteins represent a vanishingly small fraction of all possible amino acid sequences. Indeed, proteins artificially created from random nucleotide sequences generally have a low secondary structure content (107, 112). Hence our first hypothesis: novel, overprinting proteins are not expected to have a fixed 3D structure at birth, given the low probability of generating structure from a completely new sequence.

Disordered proteins are generally subject to less structural constraint than ordered ones (13). Hence our second hypothesis: the presence of disorder in one or both products of an overlapping gene pair could greatly alleviate evolutionary constraints imposed by the overlap, allowing both protein products to scan a wider sequence space without losing their function.

Both hypotheses suppose only the lack of a rigid structure, as opposed to a total lack of structure (e.g., some proteins created de novo from a random nucleotide sequence, though lacking secondary structure, have a certain degree of order [112]). For that reason, in this work, we use the widest possible definition of disorder, i.e., the lack of a rigid 3D structure, and we use a program whose predictions of disorder correspond to this definition, PONDR VSL2 (69) (see Results).

In this work, we collected a large number of experimentally proven cases of proteins encoded by overlapping genes in unspliced eukaryotic RNA viruses and analyzed their sequence properties.

MATERIALS AND METHODS

Selection and curation of the data set of viral overlapping gene products. We set out to find virus genomes containing overlapping genes whose existence was supported by experimental evidence. We first downloaded the file "Virus.ids," release 2 July 2004 (ftp://ftp.ncbi.nih.gov/genomes/IDS/Viruses.ids), containing accession numbers for all complete viral genomes (except those of bacteriophages) from the NCBI viral database (6). We then downloaded the 1,562 corresponding genomes or genome segments, corresponding to 1,098 viruses (some viruses have a segmented genome), and parsed all relevant information for each genome. Since the NCBI viral genome database (6) is not completely reliably annotated (62), we had to carefully select bona fide overlapping genes. We excluded from the analysis all files containing a "join" instruction (regardless

whether it reflected a splicing event, a frameshift, or a circular genome with genes crossing the genome map borders) because their manual curation would have been too time-consuming. We excluded from the analysis all DNA viruses and all viral genera in which at least one virus is known to make use of splicing, and we selected only overlaps longer than 90 nucleotides, corresponding to 30 amino acids (aa) (see Results). We considered only one prototype virus per genus. We kept overlaps only if there was biochemical evidence that both proteins they encoded existed (i.e., detection in infected cells or in in vitro translation experiments) or if such evidence was available for the protein products of a homologous gene overlap in a related virus.

Overlaps found only in one virus species might stem from a sequencing error resulting in an artifactual N-terminal or C-terminal extension. Therefore, we checked in the literature that the proteins expressed had the actual, predicted size or that several viral strains from that species also had a similar overlap. If we could not exclude a sequencing artifact, we discarded the overlap.

If the theoretical start or stop codon of an overlapping open reading frame (ORF) as described in the NCBI file was incorrect, it was manually corrected (for instance, VP5 of infectious pancreatic necrosis aquabirnavirus starts at nucleotide 113 and not at nucleotide 68 [108]). A few unspliced RNA viruses contain bona fide overlapping genes that are not described in the corresponding NCBI genome file. They were included in the analysis, and the missing proteins they encode were manually added: rice dwarf phytoreovirus OP-ORF (89), Theiler's cardiovirus protein L* (104), and vesicular stomatitis Indiana vesiculovirus protein C′ (47). We provide their sequences in File S1 in the supplemental material.

A few viruses make use of frameshifting to generate overlapping reading frames but (presumably by mistake) their genome file does not contain a "join" instruction (for instance, the mumps rubulavirus P/V overlap), and therefore they were included in the analysis. Among those, some frameshifts or editing events result in genes that are partially colinear (upstream of the frameshift) and that thus truly overlap only downstream of the frameshift. In these cases, we excluded the colinear part. For instance, in the case of the mumps rubulavirus P/V gene system we excluded the N-terminal part common to both P and V (41). Finally, in some cases an ORF (called "1") overlaps several ORFs (called 2, 2′, 2″, 2‴, etc.) that are colinear with each other because of alternative translation initiation sites, for instance, proteins C, C′, Y1, and Y2 in Sendai respirovirus (16). In that case we kept only the ORF 2 for which the overlap with ORF 1 is the longest (in that case the ORF C).

Viral taxonomy. Viral taxonomy changes quickly, and some names of viral taxons that are widely used by virologists are not officially recognized. Several of these taxons proved to be crucial for interpretation of our results in an evolutionary light (e.g., the proposed family *Tubiviridae* [97]). Therefore, in addition to the official taxonomy (58), we have also indicated proposed taxa, indicating the corresponding references. The interested reader can consult the website where proposals to the International Committee for the Taxonomy of Viruses are made, http://talk.ictvonline.org.

PONDR analysis of viral genes. The sequences of overlapping genes and their protein products were stored in a MySQL database for analysis. Protein intrinsic disorder was predicted using PONDR VSL2 (69), a neural network trained on a set of ordered and disordered sequences, which relies on attributes such as the composition of particular amino acids or hydropathy to predict disorder propensity along a protein sequence. PONDR predictions were also stored in the database.

Bootstrapping was used on the results to generate the confidence intervals shown. Ten thousand data sets of overlaps were randomly selected with replacement, and the calculations were repeated on each one of them. The 10,000 results were sorted and used to provide the boundary results for the appropriate confidence intervals.

The distribution of disordered regions in the overlapping regions was compared to the overall distribution of disorder in the entire data set. The null hypothesis tested was that the distribution of disorder in overlapping regions is the same as that in the entire data set; that is, we assume that there is no bias toward a greater concentration of disordered residues in overlapping regions. Using a chi-square test on sequence positions located 15 residues apart (which satisfies the assumption of independence), we obtain a P value that expresses the probability that our null hypothesis is correct.

Identification of putative ancestral, overprinted proteins. As a first screen, all proteins encoded by overlapping genes were subjected to SMART analysis (52), which includes prediction of PFAM and SMART domains, transmembrane and low-complexity regions, signal peptides, etc. The sequences of all overlapping protein regions were analyzed using (i) Psi-blast (2); (ii) sequence profile comparison methods, which automatically run a Psi-blast query on a single sequence, align the retrieved sequence hits, derive a profile from the corresponding multiple-sequence alignment, and search the library of sequence profiles in PFAM release 23 (25) for similar profiles (HHpred [86], Compass [74], and FFAS03 [39]); and (iii) fold recognition methods (Fugue [81] and Phyre [9]). Finally, we submitted the 3D structures of proteins, when available, to structural similarity searches using VAST (30) and SSM (49). Protein regions were considered ancestral if they had statistically significant sequence or structural similarity with at least another protein region from a different viral family (unclassified genera were counted as distinct families).

Prediction of structural organization of pairs of known ancestral/novel overlapping regions. The analyses described in the previous paragraph identified known domains, transmembrane segments, etc. Refined disorder prediction was carried out as follows (respecting the principles described in reference 23). We analyzed proteins containing novel or ancestral regions using the disorder predictor iPDA. For a conservative approach, we also used the predictors Prelink and Disopred, which have a very high specificity (113), when the presence of disorder in a certain region was dubious. If neither program predicted disorder within the region under scrutiny, we considered the whole region to be ordered. The boundaries of disordered regions were refined by visual inspection of hydrophobic cluster analysis plots (14). To find experimental evidence of disorder, all proteins were subjected to a Blastp similarity search (2) against the database of disordered proteins Disprot (82), and we also carried out extensive bibliographical searches.

Analysis of amino acid composition. Composition Profiler (102) allows comparison of the composition of a user-defined "query" data set (for instance, overlapping regions of proteins) with that of another user-defined "background" data set (for instance, nonoverlapping regions) or with that of a precompiled data set. The precompiled data sets we used are SwissProt 51 (4), which is most similar to the distribution of amino acids in nature; PDB Select 25, which is a subset of structures from the Protein Data Bank (10) with less than 25% sequence identity, biased toward the composition of proteins amenable to crystallization studies; and DisProt 3.4 (82), which is a set of sequences of experimentally determined disordered regions. Composition Profiler also allows the discovery of biases in certain groups of amino acids such as order-promoting amino acids or charged amino acids ("discover" option) (102) and the calculation of the relative entropy (RE) of two data sets, which roughly summarizes how dissimilar their compositions are. We used a significance value of 0.01 to identify composition biases.

Disorder content of differentially constrained overlapping genes. The disorder content of viral overlapping genes whose evolutionary rates are known was calculated using the PONDR VSL2 predictor. Protein sequences were taken from genome entries. The GenBank accession numbers of the genomes are as follows: hepatitis B virus, NC_003977; human T-lymphotropic virus, AF139170; simian immunodeficiency virus, U72748; human papillomavirus, AF293961; coliphage ϕX174, J02482; potato leafroll virus, AF453389; Sendai virus, AB039658; and cotton leaf curl virus, NC_004607.

RESULTS

Collection of a curated data set of overlapping genes from a wide range of eukaryotic RNA viruses. We carefully selected overlapping genes whose existence was supported by experimental evidence. Indeed, including an overlapping reading frame that is in fact not translated might introduce noise in our analyses, since such sequences are not subject to evolutionary pressure. Misannotated overlaps might stem from untranslated "hypothetical" genes, from a start codon wrongly assigned upstream of the true start codon, or from an undetected splicing event that results in an exon/intron overlap instead of an overlap of coding sequences. The last possibility prompted us to exclude all viruses that are known to make use of splicing. Curation of prokaryotic viruses (bacteriophages) and of DNA viruses proved to be too difficult. Therefore, we focused on unspliced, eukaryotic RNA viruses, which are either single stranded with a plus or minus genome polarity (respectively, +ssRNA and −ssRNA) or double stranded (dsRNA), and on unspliced retroid viruses, which use both DNA and RNA in their genome (for a review, see reference 5). Only one representative virus per genus was chosen.

The construction and curation of the data set are described

TABLE 1. Properties of the overlapping gene data set[a]

Property	No. of:				aa
	Families[b]	Genera[b]	Overlapping gene pairs[c]	Proteins affected by overlaps[d]	
Type of nucleic acid					
+ssRNA	13	27	30	58	
−ssRNA	4	8	12	20	
dsRNA	3	6	6	12	
Retroid	2	2	4	6	
All viruses	22	43	52	96	
No. of residues					
Total					42,656
Encoded by overlaps					16,175 (38%)
Length of protein region encoded by overlap					
Minimum					36 (*Arterivirus*)
Maximum					626 (*Tymovirus*)
Mean					138

[a] Repartition of collected viruses by taxonomy and various statistics.
[b] Distinct, unassigned genera or unassigned families are counted as bona fide genera or families.
[c] Some genera contain several overlapping gene pairs.
[d] Some genes overlap with more than one gene.

in Materials and Methods. We concentrated on overlaps longer than 90 nucleotides, corresponding to 30 aa, for two reasons: (i) shorter regions are unlikely to fold by themselves (87) and are thus expected to have a lesser structural impact, and (ii) the reliability of disorder prediction increases with length (65, 90). By taking all of these precautions, we built a very conservative, high-quality data set of 43 viral genomes containing bona fide overlapping genes.

Table 1 shows some statistics for the 43 viral genomes comprising our data set, which are presented in Tables 2 to 6. They are grouped by taxonomy, to which we have paid particular attention in order to make this work as informative as possible (see Materials and Methods).

Some viral genomes contain several pairs of overlapping genes (for instance, the *Arterivirus* GP2/GP3 and GP3/GP4 overlaps [Table 2]), while some genes overlap with more than

TABLE 2. Overlapping genes in unspliced viruses of the orders *Reovirales*, *Picornavirales*, and *Nidovirales*[a]

Order	Family	Genus	Virus species	Genome accession no.	Protein product[b]	Protein accession no.	Boundaries of overlap (aa)	
							Start	End
Reovirales (proposed)	*Birnaviridae*	*Aquabirnavirus*	*Infectious pancreatic necrosis virus*	NC_001915	VP5	NP_047195	3	133
					VP2 (capsid)	NP_047196	1	131
		Avibirnavirus	*Infectious bursal disease virus*	NC_004178	VP5	NP_690837	16	149
					VP2 (capsid)	NP_690838	1	134
	Reoviridae	*Orthoreovirus*	*Mammalian orthoreo virus* 1	NC_004267	Sigma-1a (hemagglutinin)	NP_694621	21	139
					Sigma-1bNS	NP_694622	1	119
		Oryzavirus	*Rice ragged stunt virus*	NC_003771	Replicase	NP_620541	160	485
					P4b	NP_620542	1	326
		Phytoreovirus	*Rice dwarf virus*	NC_003768	Pns12	NP_620538	91	182
					OP-ORF	—[c]	1	92
	Totiviridae	*Totivirus*	*Saccharomyces cerevisiae L-BC (La) virus*	NC_001641	Capsid (Gag)	NP_042580	649	697
					Replicase (Pol)	NP_042581	1	49
Picornavirales	*Picornaviridae*	*Cardiovirus*	*Theiler's virus*	NC_001366	L (polyprotein, VP4)	NP_040350	5	160
					L* ("L star")	—	1	156
Nidovirales	*Arteriviridae*	*Arterivirus*	*Lactate dehydrogenase-elevating virus*	NC_001639	GP2	NP_042574	184	227
					GP3	NP_042575	1	44
					GP3	NP_042575	156	191
					GP4	NP_042576	1	36

[a] Details of overlapping genes and the proteins they encode are shown. Common alternative names of proteins are given in parentheses. A comprehensive list of alternative names of viral proteins can be found in the database Virgen (http://bioinfo.ernet.in/virgen/virgen.html) (50).
[b] Abbreviations: GP, glycoprotein; L, large protein.
[c] —, several proteins are not mentioned in the NCBI genome file and thus have no accession numbers, although their existence has been proven (see Materials and Methods). We provide their sequences in File S1 in the supplemental material.

TABLE 3. Overlapping genes in unspliced retroid viruses[a]

Family	Genus	Virus species	Genome accession no.	Protein product[b]	Protein accession no.	Boundaries of overlap (aa)	
						Start	End
Caulimoviridae	*Badnavirus*	*Cacao swollen shoot virus*	NC_001574	Polyprotein	NP_041734	1721	1834
				ORF5	NP_041736	1	114
Hepadnaviridae	*Orthohepadnavirus*	*Arctic ground squirrel hepatitis B virus*	NC_001719	Capsid precursor (E antigen precursor)	NP_043862	166	217
				P (pol)	NP_043864	1	52
				P (pol)	NP_043864	188	614
				L	NP_043865	1	427
				P (pol)	NP_043864	793	877
				X protein	NP_043868	1	85

[a] See Table 2, footnote *a*.
[b] Abbreviations: L, large envelope protein; P, polymerase.

one gene (for instance, the *Orthohepadnavirus* P gene overlaps with three genes: L, X, and the capsid gene [Table 3]). Therefore, in total there are 52 gene overlaps (104 overlapping regions) in the data set, involving 96 protein products (Table 1). All overlaps in the data set are sense/sense, i.e., correspond to genes found on the same nucleic acid strand, and none encodes more than two proteins in different reading frames. The mean size of viral overlaps was 138 aa (Table 1), which corresponds to the typical size of a protein domain and is much longer than typical overlaps reported to exist in bacterial genomes (29, 71). No precise data are available for eukaryotes due to the difficulty in reliably predicting overlapping genes, but a significant number of overlaps with a comparable length has been reported (1, 70).

Examples of bona fide overlapping genes that have not been incorporated in this study because of the restrictions described above or because of technical limitations (see Materials and Methods) include the *Bornavirus* P/X gene overlap (109), which was removed because bornaviruses are known to make use of splicing (79), and the *Henipavirus* P/V and P/C overlaps

TABLE 4. Overlapping genes in unspliced +ssRNA viruses of the alphavirus-like supergroup[a]

Group or order[b]	Family	Genus	Virus species	Genome accession no.	Protein product[c]	Protein accession no.	Boundaries of overlap (aa)	
							Start	End
Group Altovirus	*Bromoviridae*	*Cucumovirus*	*Cucumber mosaic virus*	NC_002035	Replicase	NP_049324	778	857
					2b	NP_619631	1	80
		Ilarvirus	*Spinach latent virus*	NC_003809	Replicase	NP_620678	696	797
					2b	NP_620679	1	102
	Hepeviridae	*Hepevirus*	*Hepatitis E virus*	NC_001434	Capsid protein (ORF2)	NP_056787	1	110
					ORF3 (P)	NP_056788	14	123
	Proposed family *Tubiviridae*	*Hordeivirus*	*Barley stripe mosaic virus*	NC_003481	TGBp2 (beta C)	NP_604488	69	131
					TGBp3 (beta D)	NP_604489	1	63
		Pecluvirus	*Indian peanut clump virus*	NC_004730	P14 (TGBp2)	NP_835266	71	122
					P17 (TGBp3)	NP_835267	1	52
		Pomovirus	*Potato mop-top virus*	NC_003725	TGBp2	NP_620439	72	119
					TGBp3	NP_620440	1	48
Group Typovirus (proposed order *Tymovirales*)	*Tymoviridae*	*Tymovirus*	*Turnip yellow mosaic virus*	NC_004063	Movement protein (OP)	NP_663296	3	628
					Replicase	NP_663297	1	626
	Flexiviridae	*Capillovirus*	*Apple stem grooving virus*	NC_001749	Polyprotein	NP_044335	1584	1903
					Movement protein (36K)	NP_044336	1	320
		Carlavirus	*Blueberry scorch virus*	NC_003499	Coat protein	NP_612812	268	312
					NABP (16kD)	NP_612813	1	45
		Trichovirus	*Apple chlorotic leaf spot virus*	NC_001409	Movement protein	NP_040552	356	460
					Coat protein	NP_040553	1	105
		Mandarivirus	*Indian citrus ringspot virus*	NC_003093	Coat protein	NP_203557	226	325
					NABP (23kD)	NP_203558	1	100
		Potexvirus	*Cassava common mosaic virus*	NC_001658	TGBp2 (movement protein)	NP_042697	63	112
					TGBp3	NP_042698	1	50

[a] See Table 2, footnote *a*.
[b] Unofficial taxons (see "Collection of a curated data set of overlapping genes from a wide range of eukaryotic RNA viruses" in Results).
[c] Abbreviations: NABP, nucleic acid-binding protein; OP, overlapping protein; P, phosphoprotein; TGBp, protein encoded by the triple gene block.

TABLE 5. Overlapping genes in unspliced +ssRNA viruses which do not belong to any order or supergroup

Family	Genus	Virus species	Genome accession no.	Protein product	Protein accession no.	Boundaries of overlap (aa)	
						Start	End
Barnaviridae	*Barnavirus*	*Mushroom bacilliform virus*	NC_001633	ORF1	NP_042508	3	179
				Vpg-protease	NP_042509	1	177
				Vpg-protease	NP_042509	605	657
				Replicase	NP_042510	1	53
Unclassified	*Sobemovirus*	*Sesbania mosaic virus*	NC_002568	Polyprotein	NP_066392	900	962
				Capsid	NP_066394	1	63
Nodaviridae	*Alpha-nodavirus*	*Flock house virus*	NC_004146	Protein A (replicase)	NP_689444	900	998
				B2	NP_689446	1	99
	Beta-nodavirus	*Striped Jack nervous necrosis virus*	NC_003448	Protein A (replicase)	NP_599247	893	967
				B (B2)	NP_599248	1	75
	Unclassified	*Macro-brachium rosenbergii noda virus*	NC_005094	Replicase	NP_919036	901	1033
				B2	NP_919037	1	133
Tetraviridae	*Betatetravirus*	*Nudaurelia capensis beta virus*	NC_001990	Replicase	NP_048059	1316	1925
				Coat	NP_048060	1	610
	Omegatetravirus	*Dendrolimus punctatus tetravirus*	NC_005899	p17	YP_025095	32	158
				p71 (capsid)	YP_025096	1	127
Unclassified	*Umbravirus*	*Tobacco bushy top virus*	NC_004366	RNP (LDM)	NP_733849	6	237
				MP	NP_733850	1	232
Tombusviridae	*Aureusvirus*	*Pothos latent virus*	NC_000939	Movement protein (27K)	NP_051033	44	173
				14K	NP_051034	1	130
	Carmovirus	*Hibiscus chlorotic ringspot virus*	NC_003608	Replicase (p28/p81)	NP_619671	4	212
				p23	NP_619673	1	209
				Coat	NP_619676	5	228
				p25	NP_619677	1	224
	Machlomovirus	*Maize chlorotic mottle virus*	NC_003627	p31	NP_619720	130	279
				Coat	NP_619722	1	150
	Necrovirus	*Tobacco necrosis D virus*	NC_003487	$P7_1$	NP_608313	13	62
				P7a	NP_608314	1	50
	Tombusvirus	*Cymbidium ringspot virus*	NC_003532	MP	NP_613263	11	182
				p19	NP_613264	1	172

(106), which were excluded because the genome file contained a "join" instruction (see Materials and Methods), which is generally indicative of splicing but in this case is indicative of a frameshift.

In spite of these limitations, our data set still covers a wide evolutionary range. It consists mostly of ssRNA and dsRNA viruses, with only two retroid viruses (Table 3), because most retroid viruses are spliced and have thus been excluded. The data set includes at least one representative from several large viral orders or supergroups: the (unofficial) alphavirus-like supergroup (72, 103) (Table 4); the orders *Picornavirales*, *Nidovirales* (Table 2), and *Mononegavirales* (Table 6); and the proposed order *Reovirales* (58) (Table 2). Thus, our data set represents a good sampling of the diversity of overlapping genes in RNA viruses.

Proteins regions encoded by overlaps have a higher disorder content. We have chosen to use the PONDR VSL2 software for the automated analysis because it has consistently been found to have one of the best combinations of specificity and sensitivity (88) and because its definition of "disorder" is well suited to the biological question studied. Indeed, when PONDR VSL2 predicts a region to be "disordered," what it predicts, more precisely, is that it has no fixed 3D structure (69), which corresponds to our hypotheses about overlapping gene products (see the introduction). In addition to using PONDR, we also carried out in-depth analysis of selected proteins using a combination of structural prediction methods, as described in Materials and Methods and below. Our strategy is described in Fig. 2.

All proteins encoded by overlapping genes were subject to prediction of structural disorder using PONDR VSL2. As shown in Fig. 3, 29% of the amino acids of the whole data set are predicted to be in a disordered state. This is distributed in relation to overlapping as follows: 23% of the amino acids in nonoverlapping regions are predicted to be disordered, to be compared with 48% of the amino acids in overlapping regions. This difference in disorder content is highly significant (chi-square value = 254.4, one degree of freedom, $P = 2.7 \times 10^{-57}$) (see Materials and Methods). Thus, in our data set, protein regions encoded by overlapping genes show a significant bias toward structural disorder.

Identification of ancestral/novel protein pairs by their phylogenetic distribution. One of our hypotheses (see the introduction) was that novel proteins created by overprinting tend to be disordered. Therefore, we tried to identify overlaps encoding recognizable ancestral/novel protein pairs.

Finding which protein is the ancestral one and which is the novel one in an overlapping pair is a difficult problem. Methods include (i) comparison of the codon usage of each overlapping reading frame to that of nonoverlapping genes of the viral genome (67, 68) and (ii) assessing the phylogenetic distribution of each overlapping gene product, i.e., the extent to which they have homologs in other organisms (43, 71). In these methods, the ancestral reading frame is assumed to be, respec-

TABLE 6. Overlapping genes in unspliced −ssRNA viruses[a]

Order	Family	Genus	Virus species	Genome accession no.	Protein product[b]	Protein accession no.	Boundaries of overlap (aa)	
							Start	End
None	*Bunyaviridae*	*Orthobunyavirus*	*Bunyamwera virus*	NC_001927	N	NP_047213	7	107
					NSs	NP_047214	1	101
Mononegavirales	*Paramyxoviridae*	*Morbillivirus*	*Measles virus*	NC_001498	P	NP_056919	232	299
					V	—[c]	232	299
					P	NP_056919	8	193
					C	NP_056920	1	186
		Unclassified	*Tupaia paramyxovirus*	NC_002199	P	NP_054691	230	282
					V	NP_054692	230	282
					P	NP_054691	9	161
					C	NP_054693	1	153
		Unclassified	*Mossman virus*	NC_005339	P	NP_958049	244	295
					V	NP_958050	244	295
					P	NP_958049	11	162
					C	NP_958051	1	152
		Respirovirus	*Sendai virus*	NC_001552	C′	NP_056872	8	215
					P	NP_056873	1	208
					P	NP_056873	318	369
					V	—[c]	318	369
		Rubulavirus	*Mumps virus*	NC_002200	P	NP_054708	156	224
					V	NP_054709	1	224
	Rhabdoviridae	*Vesiculovirus*	*Vesicular stomatitis Indiana virus*	NC_001560	NS	NP_041713	25	91
					C′	—[c]	1	67
	Filoviridae	*Ebolavirus*	*Reston Ebola virus*	NC_004161	SGP	NP_690583.1	297	367
					sSGP	NP_690584.1	297	367

[a] See Table 2, footnote *a*.
[b] Abbreviations: N, nucleoprotein; P, phosphoprotein; NS, nonstructural protein (phosphoprotein); NSs, nonstructural protein produced from small RNA; SGP, structural glycoprotein; sSGP, soluble structural glycoprotein.
[c] Several proteins are not mentioned in the corresponding genome file and thus have no accession number, although their existence has been proven (see Materials and Methods). We provide their sequences in File S1 in the supplemental material.

tively, the one having the standard genome codon usage and the one with the widest phylogenetic distribution. Whenever possible, both methods should be used together, since they are complementary (43). However, implementing the first method with nearly 100 viral proteins is a large project in itself and is clearly outside the scope of this work. Therefore, we chose to examine the phylogenetic distribution of each overlapping gene product. We presumed that a protein region (>30 aa) involved in an overlap was ancestral only if it was conserved in at least two viral families. Given the high rate of evolution of RNA viruses (20), this is a very stringent, and thus very conservative, criterion.

Our strategy is described in Fig. 2 and in Materials and Methods. Briefly, protein regions were considered ancestral only if they had either statistically significant sequence similarity or structural similarity with at least another protein region from a different viral family. Sequence similarity was assessed using profile-profile comparison, and structural similarity was assessed using fold recognition methods or direct structural comparison.

We found 21 protein regions matching this criterion, coming from 20 proteins from 19 viral genera. They are presented in Table 7. Several viral families contain genera with homologous pairs of overlapping genes (i.e., both overlapping regions have homologs in another viral genus, which also overlap): the *Birnaviridae* VP5/VP2 overlap, the *Tubiviridae* TGB2/TGB3 overlap, and the *Tombusviridae* movement protein/p19 or p14 overlap (Table 7). In these cases we retained only one viral genus per family (*Avibirnavirus*, *Pomovirus*, and *Tombusvirus*, respectively). In the end we found 17 nonhomologous overlaps encoding ancestral regions, from 15 different genera corresponding to nine families of +ssRNA, dsRNA, and retroid viruses (Table 7).

All ancestral regions match at least one PFAM sequence family as shown using profile-profile comparison (see Materials and Methods); in other terms, no ancestral region was selected only on the basis of structural similarity. (Briefly, a PFAM family is a collection of sequences of homologous protein domains or regions [25]. Related PFAM families are grouped in "clans" [24].)

We found no gene overlap for which both protein products were presumed to be ancestral according to the phylogenetic distribution criterion. In other terms, all the overlaps selected by this method encoded, on the one hand, a protein region conserved in at least two viral families and, on the other hand, a protein region that was restricted to one family at most. This reinforces our working hypothesis that protein regions conserved in two viral families can be considered ancestral whereas the regions overlapping them are novel (see also Discussion). Table 7 presents novel protein regions together with the ancestral protein regions that they overlap.

Some ancestral regions have homologs in a very large number of viral families, and it would be highly impractical to mention all these viral families. Instead, we present in Table 7 the PFAM families (release 23) corresponding to ancestral regions. This allows the reader to visualize easily the taxonomic distribution of homologs of ancestral regions, thanks to a user-

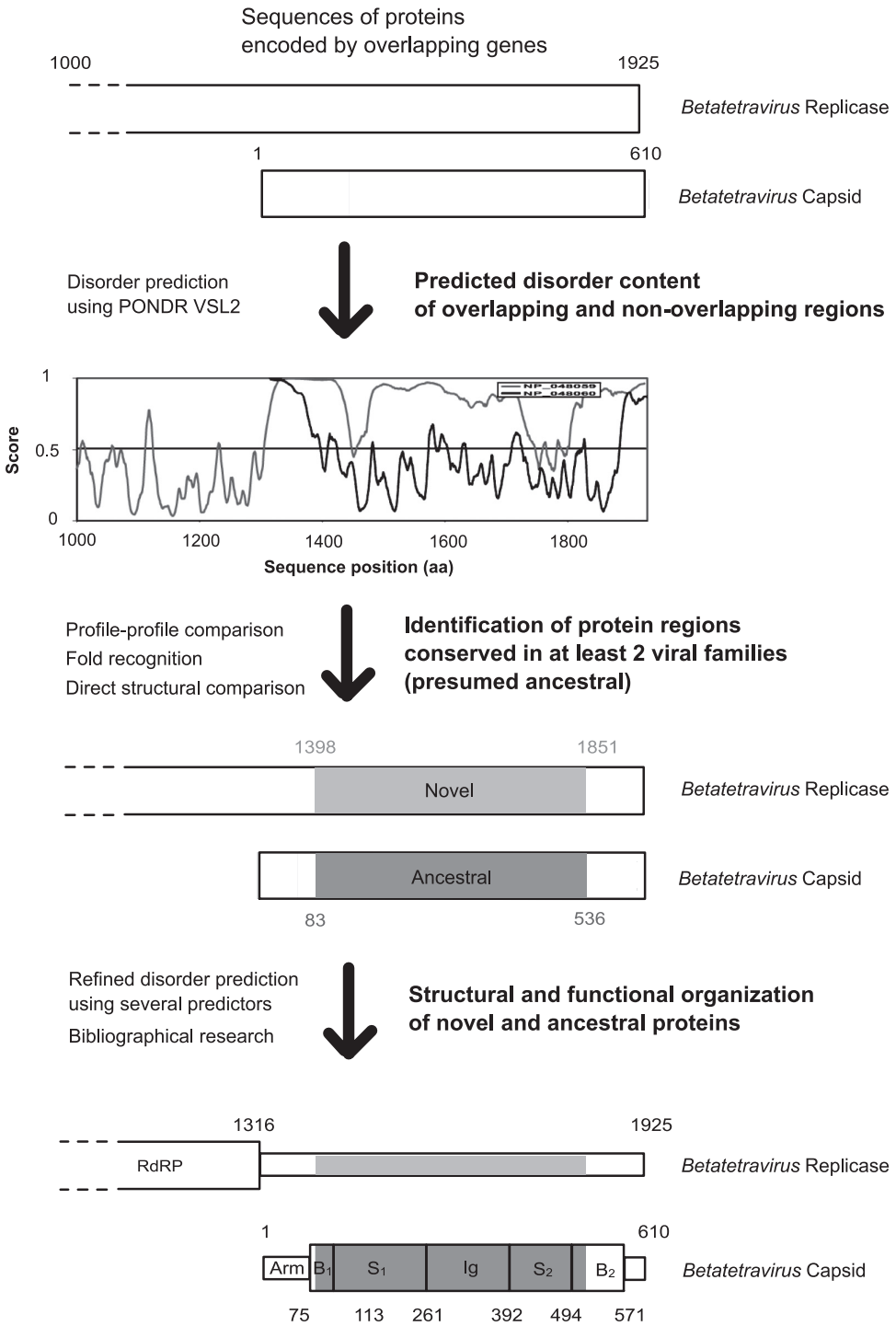

FIG. 2. Structural and functional prediction work flow, showing the *Betatetravirus* replicase/capsid overlap. Conventions are the same as in Fig. 1. Second panel, superimposed PONDR prediction for the capsid (dark gray) and replicase (light gray). Regions with a score of above 0.5 are predicted to be disordered. Third panel, predictions of the boundaries of ancestral and novel regions of the replicase and capsid (see text). Bottom, result of refined structural and functional analysis (see text). Wide and narrow boxes correspond, respectively, to predicted order and disorder. Domain names were obtained from the literature. Note the good agreement between automated PONDR predictions and the refined analysis.

friendly service called "species" available on the PFAM website as well as relevant bibliographical references (25).

During the analysis of this large data set, we uncovered evolutionary relationships between some viral proteins, using profile-profile comparisons (see Materials and Methods). In Table 7 we propose corresponding new PFAM families and clans (24). Two of these suggested clans correspond to distant sequence similarities unreported so far, to our knowledge. The

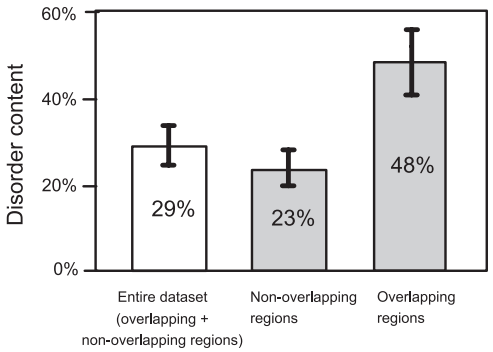

FIG. 3. Predicted disorder content of proteins encoded by overlapping genes. The prediction was made using PONDR VSL2. The error bars correspond to a 95% confidence interval.

first involves the nucleoproteins of the *Bunyaviridae* and of the unclassified genus *Tenuivirus*. The second involves the C-terminal moiety of the methyltransferase-guanylyltransferase (MT-GT) (72) of the *Altovirus* group, called the "Y region" (45). We found that it is also present in the *Typovirus* group and is thus conserved throughout the alphavirus-like supergroup (Table 4). This finding is consistent with experimental evidence that the MT-GTs of this viral supergroup have a common mechanism (56). This MT-GT is unique to these viruses and thus constitutes an important drug target for a number of human pathogens such as hepatitis E virus or chikungunya virus. Its structure has not been solved at present, and thus our finding might facilitate further protein expression studies or modeling studies.

Prediction of the structural organizations of ancestral proteins and of novel proteins. We then predicted the structural organization of each ancestral and novel protein using a combination of complementary methods (see Materials and Methods and Fig. 2) and plotted it in Fig. 4. All 17 ancestral protein regions are predicted to be ordered. Out of the 17 novel protein regions, 6 are predicted to be mostly ordered (*Carmovirus* p25, *Tombusvirus* p19, *Orthohepadnavirus* S domain, *Capillovirus* replicase, *Orthobunyavirus* nonstructural proteins, and *Carmovirus* p23), 1 is predicted to be about half ordered (the *Potexvirus* TGBp3), and the 10 others are predicted to be mostly disordered. Thus, these results suggest a greater tendency for intrinsic disorder in novel protein regions, which is compatible with the first hypothesis described in the introduction.

Biased sequence composition of protein regions encoded by overlaps. Earlier studies have suggested that overlapping protein regions have a biased sequence composition, being enriched in amino acids with the highest codon degeneracy (i.e., those encoded by six different codons) (68). We performed an exploratory analysis based on our larger data set. Using Composition Profiler (102), we first examined global biases in amino acid composition, represented by the "RE" (see below), and then examined biases in specific amino acids. We compared the sequence composition of all overlapping regions, or of novel or ancestral regions (Table 7 and Fig. 4), to that of reference sets, i.e., Swiss-Prot, PDB, and Disprot. Roughly, they correspond, respectively, to the mean composition of proteins in nature, to that of ordered proteins, and to that of disordered proteins (see Materials and Methods). To examine biases in global composition, we calculated the RE between each data set and Swiss-Prot, which is a rough measure of their difference in mean composition (102) (see Materials and Methods). The higher the RE of two data sets, the more they differ in composition. For instance, the REs of PDB and of Disprot relative to Swiss-Prot are, respectively, 0.002 and 0.07 (Fig. 5), which indicates that Swiss-Prot has a composition much closer to that of PDB than to that of Disprot.

Figure 5 clearly shows that overlapping regions (bar 4) have an important composition bias relative to Swiss-Prot (RE lower than that of Disprot but much higher than that of PDB). Considering the subset of ancestral/novel regions (listed in Table 7), we see that ancestral regions have an RE only slightly lower than that of all overlapping regions (compare bars 5 and 4) but that novel regions (bar 6) have a spectacular composition bias, with an RE more than twice that of Disprot. As a control, the RE of the "background" composition is much lower than that of the overlapping data sets (compare bar 3 and bars 4 to 6).

We then computed the relative enrichment or depletion in specific amino acids of our data sets with respect either to Swiss-Prot or to nonoverlapping regions (used as a "background" composition of viral proteins). The biases uncovered when comparing the data sets to the background were similar to those observed compared to Swiss-Prot but of lower magnitude (not shown). Consequently, in order to draw conservative conclusions, we present the composition bias of each amino acid relative to this background, instead of Swiss-Prot, in Fig. 6. Amino acids are arranged according to their codon degeneracy as described previously (68). We also examined whether the data sets were significantly ($P < 0.01$) biased in disorder-promoting or in order-promoting amino acids (listed in reference 102) using the "Discovery" option of Composition Profiler (see Materials and Methods) (Fig. 6).

Taken together, overlapping regions have a significant deviation in most amino acids (16 out of 20) and are significantly biased toward disorder, i.e., enriched in disorder-promoting amino acids and depleted in order-promoting amino acids (Fig. 6, top panel). The subsets of ancestral and of novel regions show distinct trends. Ancestral regions have a composition bias for three amino acids only (middle panel) and have no significant bias toward order or disorder. In contrast, novel regions (bottom panel) are heavily biased regarding both the number of amino acids involved (18) and the magnitude of the bias (on average more than twice that of overlapping regions taken globally [compare top and bottom panels]). Furthermore, they are biased toward disorder (bottom panel, right).

Finally, we examined Fig. 6 qualitatively, looking for a bias of overlapping regions with respect to codon degeneracy: for instance, enrichment in amino acids encoded by highly degenerate codons (as reported in reference 68) or depletion in amino acids encoded by low-degeneracy codons. This simple visual examination suggests that overlapping regions taken globally (top panel) are enriched in amino acids with a codon degeneracy of ≥ 4 and depleted in amino acids with a degeneracy of <4. However, the magnitude of this bias depends upon the data set chosen as background (Swiss-Prot or nonoverlapping regions [not shown]), and it should be taken with great care until validated by a rigorous statistical analysis of a

TABLE 7. Pairs of recognizable ancestral/novel overlapping protein regions[a]

Family	Genus	Protein pair	Novel region/ancestral region (aa)	Matching PFAM families (matching PFAM clans)	Suggested new families (Suggested new clans)	Common name of corresponding region	Function of novel full-length protein/ function of ancestral full-length protein (reference)	Function of novel region (reference)
Birnaviridae	*Avibirnavirus* (*Aquabirnavirus*)	VP5/VP2	28–149/13–134	Birna_VP5/Birna_VP2 (Viral_ssRNA_CP)		Capsid protein with a nucleoplasmin-like fold	Antiapoptosis/viral capsid	Antiapoptosis (36)
Bunyaviridae	*Orthobunyavirus*	NSs/N	60–98/66–104	Bunya_NS-S/Bunya_nucleocap, Tenui_N, Phlebovirus_N	—/(Bunyaviridae_N)	N-terminal moiety of nucleoprotein of *Bunyaviridae* and *Tenuivirus*	Suppressor of RNA silencing, inhibitor of interferon response, inhibitor of viral polymerase/ binds to and protects the viral genome	ND
Flexiviridae	*Potexvirus*	TGBp3/TGBp2	10–50/72–112	7kD_coat/Plant_vir_prot		Movement protein	Virus cell-to-cell movement/virus cell-to-cell movement	Virus cell-to-cell movement (48)
	Trichovirus	Movement protein/ coat protein	383–460/28–105	—/Flexi_CP, Clostero_coat, Tricho_coat, Poty_Coat	—/(Flexuous_coat)	Coat protein of flexuous viruses	Virus cell-to-cell movement/viral capsid	ND
	Mandarivirus	NABP/coat protein	1–53/226–278	—/Flexi_CP, Clostero_coat, Tricho_coat, Poty_Coat	—/(Flexuous_coat)	Coat protein of flexuous viruses	ND/viral capsid	ND
	Capillovirus	Polyprotein/movement protein	1593–1840/10–257	—/MP, 3A, TBSV_P22	—/(30K_MP)	Movement protein of the 30K superfamily	Multifunctional viral replicase/virus cell-to-cell movement	ND
Hepadnaviridae	*Orthohepadnavirus*	L/P	193–427/380–614	vMSA/RVT_1 [RVT]		Reverse transcriptase domain	Viral envelope/reverse transcriptase	Viral envelope (7)
		X/P	1–39/793–831	X/DNA_pol_viral_C, transposase 36	—/(RNAse_H)[b]	RNAse H	Multifunctional regulator of transcription, cell cycle, and apoptosis/DNA-RNA duplex endoribonuclease	ND
Tetraviridae	*Betatetravirus*	Replicase/coat protein	1398–1851/83–536	—/Peptidase_A21, Peptidase_A6 (Viral_ssRNA_CP)		Capsid protein	Multifuntional viral replicase/viral capsid, Self-cleaving peptidase	ND
	Omegatetravirus	P17/coat protein	119–158/88–127	—/Peptidase_A21, Peptidase_A6 (Viral_ssRNA_CP)		Capsid protein	ND/viral capsid	ND
Tombusviridae	*Tombusvirus* (*Aureusvirus*)	P19/P22	20–148/30–158	Tombus_p19/TBSV_P22, MP, 3A	—/(30K_MP)	Movement protein of the 30K superfamily	Suppressor of RNA silencing/cell-to-cell movement of viral RNA	Suppressor of RNA silencing (80)
	Carmovirus	P23/replicase	45–169/48–172	—/Tombus_P33, Luteo_P1–P2	—/(Tombus_Luteo_P33)	P33 auxiliary replication protein	Factor indispensable for host-specific replication (54)/ essential component of viral replicase	ND
		P25/coat	77–224/81–228	—/Viral_coat (Viral_ssRNA_CP)		Capsid protein with a Nucleoplasmin-like fold	Long distance (systemic) movement of viral RNA/encapsidation of virion	Long-distance (systemic) movement of viral RNA (119)
	Machlomovirus	P31/coat protein	175–279/46–150	—/Viral_coat (Viral_ssRNA_CP)		Capsid protein with a Nucleoplasmin-like fold	ND/Encapsidation of virion	ND

Tubiviridae	*Pomovirus* (*Pecluvirus*, *Hordeivirus*)	TGBp3/TGBp2	1–38/72–109	Viral_Beta_CD/Plant_vir_prot		Movement protein	Cell-to-cell movement of viral RNA/cell-to-cell movement of viral RNA	ND
Tymoviridae	*Tymovirus*	Movement protein/ Replicase	56–332/58–219 and 220–334	Tymo_45kd_70kd/Methyltrans_Typ, Vmethyltransf and Methyltrans_Typ, Vmethyltransf	—/Typovirus_MT_GT_N (Alphavirus like_MT_GT_N) and Typovirus_MT_GT_C (Alphavirus like_MT_GT_C)	N-terminal moiety of the MT-GT of these viruses and C-terminal moiety of the MT-GT of these viruses ("Y region")	Long-distance (systemic) movement of viral RNA/MT-GT/MT-GT	Long-distance (systemic) movement of viral RNA (12)
Unclassified	*Umbravirus*	RNP/movement protein	6–237/1–232	Umbravirus_LDM/3A, MP, TBSV_P22	—/(30K_MP)	Movement protein of the 30K superfamily	Long-distance (systemic) movement/virus cell-to-cell movement	Long-distance (systemic) movement (91)

[a] Abbreviations are the same as in Fig. 4. ND, not determined. For each pair of overlapping protein regions, we indicate the boundaries of the ancestral region and of the novel region. When several genera encode homologous overlaps, data (or) only one genus are presented and the other genus names are given in parentheses (e.g. pomoviruses, pecluviruses, and hordeiviruses encode a homologous TGBp2/TGBp3 overlap, but only the data for pomoviruses are presented). We indicate PFAM families and clans that match these region, and proposed ones suggested on the base of profile-profile comparisons. Note that the boundaries of ancestral regions might extend beyond those of the corresponding PFAM family, since the former have been determined by structured-based methods in addition to sequence-based methods. The "species" function at http://pfam.sanger.ac.uk can be accessed for the taxonomic distribution of each PFAM family. We indicate the function of full-length ancestral and novel proteins (bibliographical information can be found on the PFAM website mentioned above) and the specific function of novel regions, when available.

[b] We suggest that the related families DNA_pol_viral_C and Transposase_36 are part of the existing clan RNAse_H on the basis of significant similarity of Transposase_36 to a member of this clan, the endonuclease family DDE, established through profile-profile comparison. This is coherent with an earlier report (57).

larger data set. No clear bias with respect to codon degeneracy is visible for either the novel or ancestral regions (Fig. 6, middle and bottom panels).

In summary, the composition of overlapping protein regions is biased toward disorder-promoting amino acids. In particular, novel regions have a very large compositional bias. Overlapping regions seem to favor the use of amino acids with a high codon degeneracy ($\geq$4), as seen using a merely qualitative approach, but this observation should be taken with caution until validated by further studies.

Specific functions of overlapping proteins. In Table 7, we have compiled the known functions of overlapping proteins. In most cases, one function or several functions have been attributed to the full-length protein but the precise function of the novel region itself has not been determined. In cases where a function has been attributed specifically to the novel region, we included it with the associated bibliographical references. Table 7 and Fig. 4 show that all novel overprinting proteins with known function, except one (the *Orthohepadnavirus* L), are "accessory" proteins (i.e., neither structural nor enzymatic), most often overprinting a structural or enzymatic protein.

Proteins generated by overprinting homologous DNA sequences are extremely diverse. Several ancestral viral proteins of our data set, from different genera, are homologous to each other (i.e., they share statistically significant sequence similarity). They have been overprinted by proteins that show no distinguishable sequence or structural similarity to each other and thus might have been created independently in each genus. The identification of such proteins, which show a wide diversity both in function and in structure, offers an unprecedented insight into de novo protein creation by viruses. For instance, consider Fig. 4, panel 4, and the corresponding Table 7. Capilloviruses, tombusviruses, and umbraviruses encode a movement protein belonging to the "30K" superfamily, sharing a homologous central domain (61). In these genera, the movement protein has been overprinted, respectively, by an ordered domain of unknown function that is part of a polyprotein, by a mostly ordered suppressor of RNA silencing (105), and by a ribonucleoprotein (which also plays a role in long-distance movement) that is predicted to be disordered but might undergo a disorder-to-order transition upon binding to RNA (92). The case of mandariviruses, trichoviruses, and capilloviruses (same panel), which all encode a homologous coat protein (18, 44), is as striking. In the first two genera it has been overprinted, respectively, by the disordered N-terminal domain of an RNA-binding protein and by the disordered C-terminal domain of a 30K movement protein, while in capilloviruses it is not part of an overlap.

Finally, Fig. 4, panel 3, shows that regions homologous to the shell (S) domain of the superfamilies of capsids having the SCOP fold "nucleoplasmin-like/VP (viral coat and capsid proteins)" (3) have been overprinted in several taxonomically distant viruses by very diverse protein regions: the *Avibirnavirus* VP5, a disordered antiapoptosis protein (36); a disordered tail of the *Betatetravirus* replicase; a disordered tail of *Machlomovirus* p31; and a region of the *Carmovirus* p25 that contains a predicted transmembrane segment (the last three having an unknown function). These examples highlight the "creativity" of nature, which, although starting from a similar material

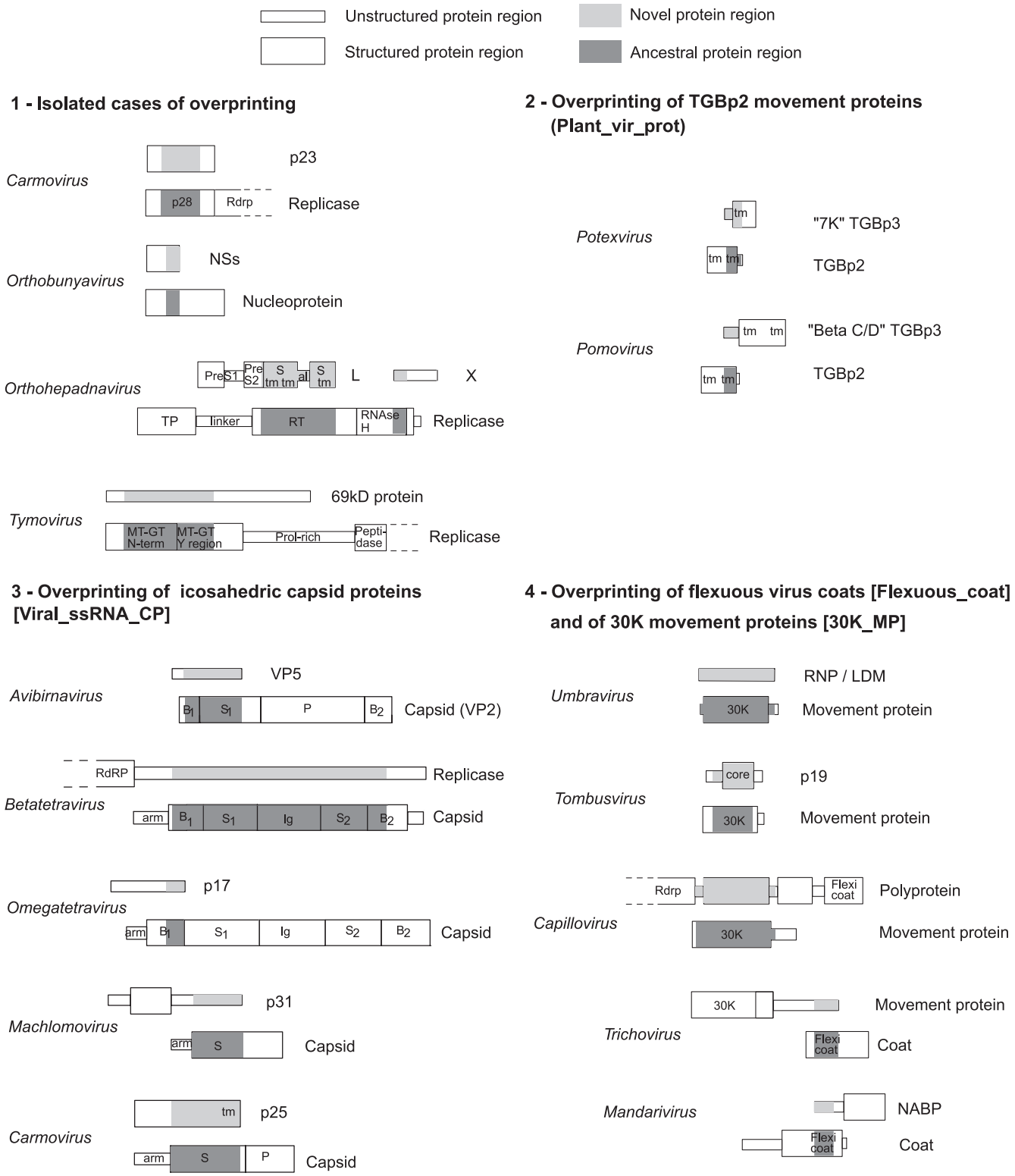

FIG. 4. Structural and functional organization of recognizable ancestral/novel overlapping protein regions. Proteins encoded by overlapping genes are represented to scale with the same conventions as in Fig. 1 and 2. Boundaries of ancestral and novel regions are given in Table 7. Each panel represents different cases of overprinting. For instance, the panel 3 represents all novel proteins that have overprinted homologous capsid proteins. The name of the panel refers to the PFAM family (in parentheses) or clan (in brackets), actual or proposed herein, to which ancestral protein regions belong (see text and Table 7). Ancestral regions within a given clan are aligned vertically (e.g., the 30K domain of *Umbra*-, *Tombus*-, and *Capillovirus* movement proteins, in panel 4). Note that domains bearing a similar name are not always homologous. For instance, in panel 2 the *Pomovirus* and *Potexvirus* TGBp2 proteins are homologous (they belong to the family Plant_vir_prot), whereas the *Pomovirus* and *Potexvirus* TGBp3 proteins are not (they belong, respectively, to the β C/D and 7K families) (Table 7). Likewise, there is no evidence that the RNA-binding

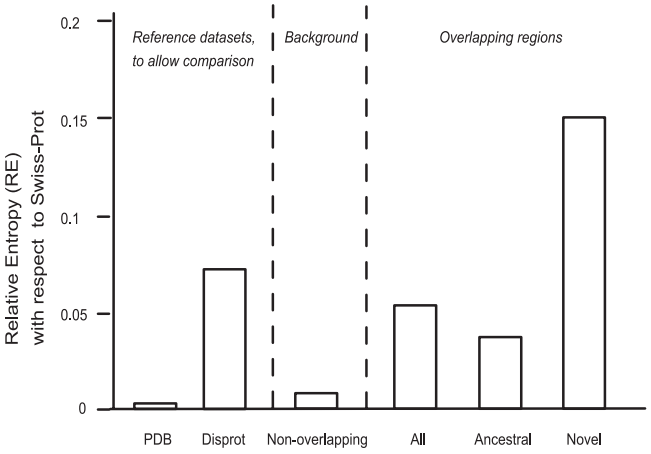

FIG. 5. REs of overlapping or nonoverlapping protein regions versus Swiss-Prot. The RE of two data sets is a rough measure of their difference in mean amino acid composition (see text). We have plotted, from left to right, the REs of biologically meaningful data sets (PDB and Disprot) with respect to Swiss-Prot; the RE of nonoverlapping regions (representative of viral proteins) with respect to Swiss-Prot; and the REs with respect to Swiss-Prot of either all overlapping regions, ancestral regions, or novel regions. Note that ancestral and novel regions form only a subset of all overlapping regions, since for some pairs of overlapping regions we could not determine which was the ancestral one and which was the novel one.

(homologous DNA sequences), did not "invent" similar proteins twice.

Disorder and sequence constraints on overlapping reading frames. Several studies have shown that overlapping genes often encode a protein heavily constrained in sequence and another one that is much less constrained (28, 32, 37, 59, 63, 64, 67, 77, 98). In these cases, we would expect the protein with the less constrained sequence to have the greater disorder content, since disordered proteins are less sensitive to sequence changes.

Measuring sequence constraints of overlapping reading frames is usually done by comparing the rate of synonymous substitutions to that of nonsynonymous substitutions for each frame, using closely related genome sequences; the frame for which this ratio is higher is considered the most constrained (38, 71). Performing such analyses on our entire data set was beyond the scope of this work, so, in order to provide some verification of the above hypothesis, we gathered from the literature all studies that provide information on the evolutionary rate differences between specific sets of viral overlapping genes (28, 32, 37, 59, 63, 64, 67, 77, 98). For each, we performed disorder predictions on the corresponding protein products using PONDR VSL2.

Figure 7 plots the predicted disorder content of both regions encoded by each overlap. It clearly shows that in 8 cases out of 10, the less constrained frame encodes the protein region with the greatest disorder content. In another case, that of human papillomavirus, the less constrained protein (E2) is only marginally less disordered than the more constrained (E4), i.e., 89% versus 100%, respectively, which in fact corresponds to both proteins being almost entirely disordered. The last overlap (ϕX174) corresponds to regions of proteins D and E predicted to be both ordered. Thus, this preliminary exploration supports the idea that the less constrained reading frame generally encodes the most disordered region. However, this is not an absolute rule, and overlapping frames can encode two ordered protein regions simultaneously (such ordered/ordered overlaps can also be found in our data set [Fig. 4]).

DISCUSSION

Our carefully curated data set and conservative analysis allow us to make a strong case for our prediction that proteins encoded by gene overlaps tend to be disordered and to offer unprecedented insight in their evolution.

Unfortunately, it was difficult to find experimental evidence relating to our predictions of disorder, in part because many proteins considered here are accessory ones, which are poorly characterized (see below). Examples of disorder predictions that are experimentally confirmed include the *Orthohepadnavirus* protein X (73), the N-terminal "arm" of the capsid proteins of omegatetraviruses (35) (Fig. 4) and sobemoviruses (51), and the N-terminal moieties of the P proteins of morbilliviruses (42) and vesiculoviruses (17). We could not find any evidence in the literature that would contradict our predictions, even though some regions predicted to be disordered can actually become partially ordered, e.g., the basic, N-terminal "arms" of the capsid proteins of a number of icosahedral viruses (51). However, this corresponds to the definition of disorder used in this work (see the introduction): proteins that do not have a unique, rigid 3D structure.

Regarding our prediction of ancestral protein regions (Fig. 4), there is good evidence for most that they are correct. For instance, the reverse transcriptases of orthohepadnaviruses belong to an ancient enzyme family (83); likewise, the S domains of capsid proteins (34), the 30K domains of movement proteins (61), and the MTs of the alphavirus-like supergroup (72) are each found in more than a dozen virus families. Furthermore, evolutionary studies of viruses from our data set that used complementary analyses, such as codon usage, are in agreement with our results: they predict that the *Tymovirus* polyprotein (68) and the *Birnavirus* VP2 are ancestral (93).

We hope to obtain further insights from other organisms. For instance, we noticed a few exciting examples of ancient proteins overprinted by proteins predicted or known to be disordered (in parentheses): the ankyrin domain of mamma-

"arms" of capsid proteins of different genera are homologous (panel 3). Abbreviations: 30K, conserved domain of the 30K family of movement proteins; al: antigenic loop; B (or B_1 or B_2), base domain (or subdomain); Flexi coat, central conserved region of flexuous viral coats; Ig, immunoglobulin-like domain; L, large envelope protein; LDM, long-distance movement protein; NABP, nucleic acid-binding protein; Prol-rich, proline-rich region; RNP, ribonucleoprotein; Rdrp: RNA-dependent RNA polymerase; RT, reverse transcriptase; S (or S_1 or S_2), shell domain (or subdomain); tm: transmembrane segment; TGBp2 and TGBp3: triple gene block proteins 2 and 3; TP, terminal protein.

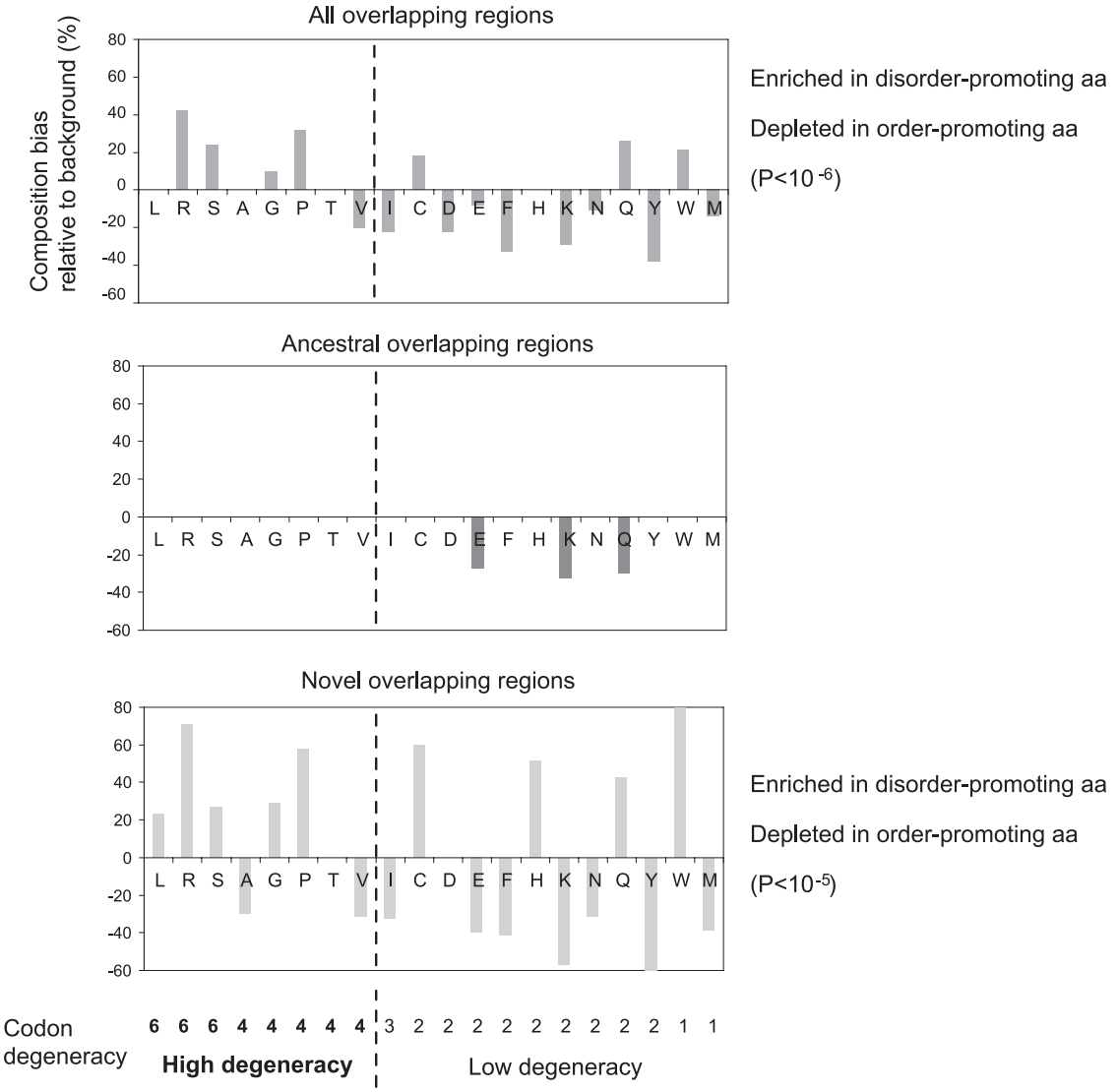

FIG. 6. Deviation in sequence composition of overlapping protein regions relative to the background composition of nonoverlapping regions. Relative enrichment (positive values) or depletion (negative values) in amino acids of each data set with respect to that of nonoverlapping regions is shown (see text). For easier visualization, we have plotted values only for the amino acids that show a statistically significant bias ($P < 0.01$). Amino acids are arranged according to their level of codon degeneracy, indicated below the lower panel (a codon degeneracy of 3 for isoleucine [I] means that three codons code for isoleucine). The dashed vertical lines separate amino acids with a high codon degeneracy (≥ 4) from those with a low degeneracy (≤ 3). Note that the data sets of novel and ancestral regions (2,280 aa each) represent only 22% of the amino acids contained in "all overlapping regions". Thus, the composition of all overlapping regions is not expected to correspond exactly to the mean composition of the ancestral and novel subsets.

lian $p16^{INK4}$ ($p19^{ARF}$) (15) and the bacterial ribosomal protein L34 (N-terminal extension of RNase P) (22).

Earlier observations on the properties of proteins encoded by overlapping genes. There have been earlier anecdotal observations of a connection between gene overlap and structural disorder. Jordan et al. suggested that the emergence of protein C in the P/C overlap of *Paramyxoviridae* (Table 6) was favored by the disordered nature of P (40). Likewise, Narechania et al. noticed that a disordered region of the *Papillomaviridae* protein E2 might have favored the overprinting of protein E4, also predicted to be disordered (64). However, these studies gave no reliable evidence that P and E2 were ancestral.

More recently, Meier et al. expressed ideas similar to those in this work, based on the analysis of a single overlap (60). They suggested that the abundant disorder observed in the crystal structure of the *Coronavirus* protein NSP9, most likely created by overprinting the nucleoprotein (N), may reflect its recent creation as well as constraints imposed by the N reading frame.

Prior to this article, there had been only one systematic study of overlapping genes at the protein level (68). It reported that proteins encoded by overlaps were enriched in amino acids with the highest codon degeneracy (R, L, and S). We found enrichment in R and S but not in L and no clear-cut influence of codon degeneracy. The difference might be due to the much

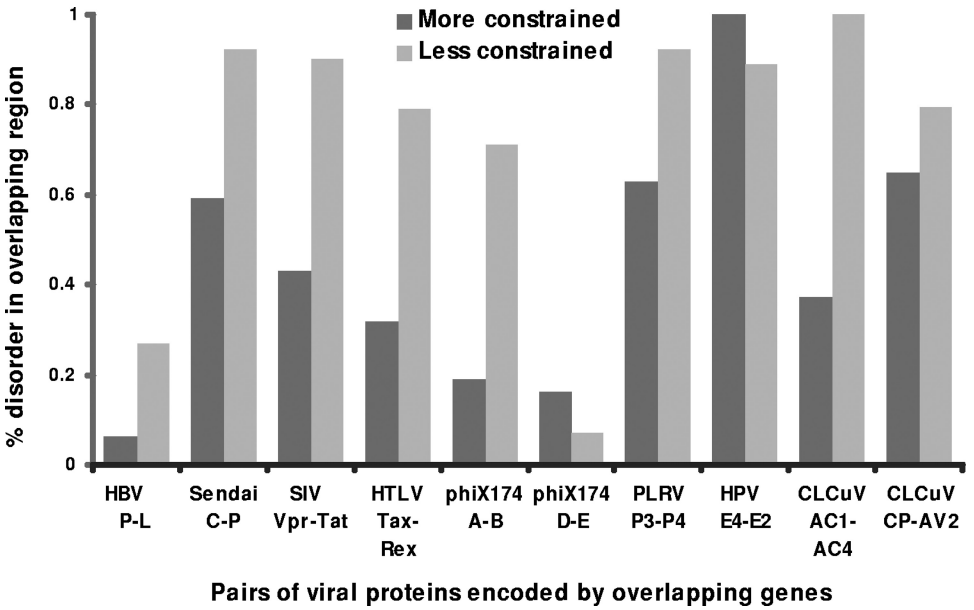

FIG. 7. Evolutionary constraints of overlapping protein regions and their disorder content. Predicted disorder content is plotted for overlapping protein pairs from several viruses, listed below the graph. In each pair, the first protein listed is the more constrained. Bars indicate the percentage of disorder in the overlapping parts of these proteins. Abbreviations: HBV, hepatitis B virus; CLCuV, cotton leaf curl virus; SIV, simian immunodeficiency virus; HTLV, human T-lymphotropic virus; ϕX174, coliphage ϕX174; PLRV, potato leafroll virus; HPV, human papillomavirus.

lower number of viral genera sampled in the previous work (68).

Recent work on (uncurated) protein products of overlapping genes of RNA viruses has made interesting connections between their relative frames, their ages, and the modes of creation of the overlap (8). Our data set of ancestral/novel protein regions is too small to reliably analyze their findings, but we plan to do so once a larger data set is created.

Why structural disorder in protein products of overlapping genes? In the introduction, we proposed two nonexclusive hypotheses to explain the increased occurrence of disorder in proteins encoded by gene overlaps: either (i) the newborn protein in each pair tends to be disordered or (ii) the presence of disorder in either protein encoded by overlapping genes lessens evolutionary constraints. In fact, our results are compatible with both hypotheses.

Indeed, almost two-thirds of novel, overlapping protein regions are disordered (Fig. 4), compared with fewer than one-fourth of nonoverlapping protein regions (Fig. 3), which is compatible with the first hypothesis. However, these results should be validated by further studies, since we could determine novel/ancestral status for only 21 overlaps out of 52.

The analysis summarized in Fig. 7 is also compatible with the second hypothesis. A number of studies have shown that overlapping genes most often encode one heavily constrained protein and another one that is much less constrained (28, 32, 37, 59, 63, 64, 67, 77, 98). Our analysis of a limited data set formed with the proteins studied in these works suggests that the less constrained proteins are generally the more disordered, which is consistent with the second hypothesis.

Thus, it is possible that both factors invoked in the two hypotheses actually contribute to the increased disorder content of overlapping gene products. A simple and attractive explanation would be that the novel proteins of each pair generally are the less constrained ones. Further studies will be needed to address this question.

Insights for viral bioinformatics. This work establishes several methodological points.

It is possible, with a reasonable effort, to make a thorough bioinformatics structural analysis with a large number (~100) of proteins involved in a given biological question. At present, this kind of analysis is quite rare (see, e.g., reference 31), although it obviously adds great value when compared to global statistics (e.g., compare Fig. 3 and 4). Furthermore, such analyses are feasible for bench virologists, thanks to the availability of user-friendly web-based tools such as the MPI toolkit (11).

Our work also suggests that viral ORFs overlapping a known coding sequence and encoding hypothetical proteins with highly biased sequence composition, which are often considered noncoding (99) and are discarded, might in fact encode a protein. Indeed, recent exciting discoveries of overlapping genes using a systematic approach (26) suggest that overlapping genes in viruses might be even more common than previously thought.

Most studies aimed at determining the ancestral protein encoded by a gene overlap did not take into account domain organization, with a few exceptions (28, 64, 67). However, the present work makes it clear that overlapping gene products are often composed of several domains that might have different evolutionary histories. For instance, the overlapping parts of the *Capillovirus* replicase and movement protein are each composed of several domains, as is the overlapping part of the *Tymovirus* replicase (Fig. 4). Thus, analyses of overlapping gene evolution should be carried out by studying domains separately.

The study of de novo proteins should enhance our knowledge of protein space. At present, it is thought that proteins adopt fewer than 10,000 structural folds in nature, much less than expected from our understanding of biophysics (115). This discrepancy has brought about two main hypotheses: (i) some structural folds are favored by nature for unknown biophysical or functional reasons, and (ii) most proteins are descended from a limited set of ancestors by duplication (for a review, see reference 116).

All solved structures of overprinting proteins presented here and elsewhere correspond to previously unobserved folds (53, 60). This constitutes a challenge to the first hypothesis above and even suggests that we might underestimate the number of folds created in nature, because of our limited knowledge of the 3D structures of proteins created de novo. Solving them (as advocated by Keese and Gibbs, remarkably, more than 15 years ago [43]) might thus help to improve methods to predict the 3D structures of proteins from their sequences, a central problem of bioinformatics which crucially depends on knowing the diversity of protein folds (33).

De novo protein creation: a significant factor in evolution? We noted in Results that the great majority of novel proteins are "accessory" (i.e., neither structural nor enzymatic), most often overprinting a structural or enzymatic protein, confirming an earlier observation (8). "Accessory" does not mean that they are dispensable in vivo; on the contrary, most novel regions play an important role in viral pathogenicity or spread (Table 7), as noticed by Li and Ding (53). Thus, de novo protein creation appears to be a significant factor in viral evolution, in particular in the evolution of pathogenicity, which is poorly understood at present.

Is it limited to overprinting by viruses? At the time that this article was submitted, two systematic studies of de novo protein creation in eukaryotes (from noncoding sequences and thus not generating overlapping genes) were published. They indicate that de novo protein creation occurs at a significant and unexpected rate, having generated between 5% and 20% of orphan proteins of primates (95) and about 12% of orphan proteins of the genus *Drosophila* (118). Reciprocally, almost all de novo-created viral proteins that we identified are orphans at the genus level, i.e., are restricted to one genus at most (see Table 7). Thus, these works and ours provide numerous examples of orphan proteins created de novo, as opposed to having diverged beyond recognition from other relatives (see the introduction).

ACKNOWLEDGMENTS

We thank S. Longhi, B. Canard, and B. Henrissat for support; V. Uversky for useful advice; R. Belshaw, N. Chirico, and V. Brechot for useful comments on the manuscript; and F. Ferron, J. Grimes, R. Esnouf, and D. Glaser for support in the latest stages. D.K. thanks A. Gibbs and P. Keese for their inspirational work. We also thank all the authors of the excellent freely available programs and databases mentioned in this work.

C.R. gathered and classified all complete, unspliced RNA viral genomes and extracted the overlapping genes. M.K. performed the order-disorder prediction and initial analysis of the genomic data set. A.K.D. coordinated the disorder prediction study. P.R.R. supervised the disorder prediction study, performed statistical analysis on the genomic data set, gathered the data, analyzed the relationship between evolutionary constraints and intrinsic disorder, and cowrote the manuscript. D.K. conceived and coordinated the study, curated the overlapping gene data set, performed the remaining bioinformatics analyses, and cowrote the manuscript.

REFERENCES

1. **Abramowitz, J., D. Grenet, M. Birnbaumer, H. N. Torres, and L. Birnbaumer.** 2004. XLalphas, the extra-long form of the alpha-subunit of the Gs G protein, is significantly longer than suspected, and so is its companion Alex. Proc. Natl. Acad. Sci. USA **101:**8366–8371.
2. **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.
3. **Andreeva, A., D. Howorth, J. M. Chandonia, S. E. Brenner, T. J. Hubbard, C. Chothia, and A. G. Murzin.** 2008. Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res. **36:**D419–D425.
4. **Bairoch, A., R. Apweiler, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi, and L. S. Yeh.** 2005. The Universal Protein Resource (UniProt). Nucleic Acids Res. **33:**D154–D159.
5. **Ball, L. A.** 2007. Virus replication strategies, p. 119–139. *In* D. M. Knipe and P. M. Howley (ed.), Fields virology, 5th ed., vol. 1. Lippincott Williams & Wilkins, Philadelphia, PA.
6. **Bao, Y., S. Federhen, D. Leipe, V. Pham, S. Resenchuk, M. Rozanov, R. Tatusov, and T. Tatusova.** 2004. National center for biotechnology information viral genomes project. J. Virol. **78:**7291–7298.
7. **Beck, J., and M. Nassal.** 2007. Hepatitis B virus replication. World J. Gastroenterol. **13:**48–64.
8. **Belshaw, R., O. G. Pybus, and A. Rambaut.** 2007. The evolution of genome compression and genomic novelty in RNA viruses. Genome Res. **17:**1496–1504.
9. **Bennett-Lovsey, R. M., A. D. Herbert, M. J. Sternberg, and L. A. Kelley.** 2008. Exploring the extremes of sequence/structure space with ensemble fold recognition in the program Phyre. Proteins **70:**611–625.
10. **Berman, H. M., T. Battistuz, T. N. Bhat, W. F. Bluhm, P. E. Bourne, K. Burkhardt, Z. Feng, G. L. Gilliland, L. Iype, S. Jain, P. Fagan, J. Marvin, D. Padilla, V. Ravichandran, B. Schneider, N. Thanki, H. Weissig, J. D. Westbrook, and C. Zardecki.** 2002. The Protein Data Bank. Acta Crystallogr. D **58:**899–907.
11. **Biegert, A., C. Mayer, M. Remmert, J. Soding, and A. N. Lupas.** 2006. The MPI bioinformatics toolkit for protein sequence analysis. Nucleic Acids Res. **34:**W335–W339.
12. **Bozarth, C. S., J. J. Weiland, and T. W. Dreher.** 1992. Expression of ORF-69 of turnip yellow mosaic virus is necessary for viral spread in plants. Virology **187:**124–130.
13. **Brown, C. J., S. Takayama, A. M. Campen, P. Vise, T. W. Marshall, C. J. Oldfield, C. J. Williams, and A. K. Dunker.** 2002. Evolutionary rate heterogeneity in proteins with long disordered regions. J. Mol. Evol. **55:**104–110.
14. **Callebaut, I., G. Labesse, P. Durand, A. Poupon, L. Canard, J. Chomilier, B. Henrissat, and J. P. Mornon.** 1997. Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives. Cell Mol. Life Sci. **53:**621–645.
15. **DiGiammarino, E. L., I. Filippov, J. D. Weber, B. Bothner, and R. W. Kriwacki.** 2001. Solution structure of the p53 regulatory domain of the p19Arf tumor suppressor protein. Biochemistry **40:**2379–2386.
16. **Dillon, P. J., and K. C. Gupta.** 1989. Expression of five proteins from the Sendai virus P/C mRNA in infected cells. J. Virol. **63:**974–977.
17. **Ding, H., T. J. Green, and M. Luo.** 2004. Crystallization and preliminary X-ray analysis of a proteinase-K-resistant domain within the phosphoprotein of vesicular stomatitis virus (Indiana). Acta Crystallogr. D **60:**2087–2090.
18. **Dolja, V. V., V. P. Boyko, A. A. Agranovsky, and E. V. Koonin.** 1991. Phylogeny of capsid proteins of rod-shaped and filamentous RNA plant viruses: two families with distinct patterns of sequence and probably structure conservation. Virology **184:**79–86.
19. **Dosztanyi, Z., M. Sandor, P. Tompa, and I. Simon.** 2007. Prediction of protein disorder at the domain level. Curr. Protein Pept. Sci. **8:**161–171.
20. **Duffy, S., L. A. Shackelton, and E. C. Holmes.** 2008. Rates of evolutionary change in viruses: patterns and determinants. Nat. Rev. Genet. **9:**267–276.
21. **Dyson, H. J., and P. E. Wright.** 2005. Intrinsically unstructured proteins and their functions. Nat. Rev. Mol. Cell Biol. **6:**197–208.
22. **Feltens, R., M. Gossringer, D. K. Willkomm, H. Urlaub, and R. K. Hartmann.** 2003. An unusual mechanism of bacterial gene expression revealed for the RNase P protein of Thermus strains. Proc. Natl. Acad. Sci. USA **100:**5724–5729.
23. **Ferron, F., S. Longhi, B. Canard, and D. Karlin.** 2006. A practical overview of protein disorder prediction methods. Proteins **65:**1–14.
24. **Finn, R. D., J. Mistry, B. Schuster-Bockler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, S. R. Eddy, E. L. Sonnhammer, and A. Bateman.** 2006. Pfam: clans, web tools and services. Nucleic Acids Res. **34:**D247–D251.
25. **Finn, R. D., J. Tate, J. Mistry, P. C. Coggill, S. J. Sammut, H. R. Hotz, G. Ceric, K. Forslund, S. R. Eddy, E. L. Sonnhammer, and A. Bateman.** 2008. The Pfam protein families database. Nucleic Acids Res. **36:**D281–D288.

26. **Firth, A. E., and J. F. Atkins.** 2008. Bioinformatic analysis suggests that a conserved ORF in the waikaviruses encodes an overlapping gene. Arch. Virol. **153:**1379–1383.
27. **Fischer, D., and D. Eisenberg.** 1999. Finding families for genomic ORFans. Bioinformatics **15:**759–762.
28. **Fujii, Y., K. Kiyotani, T. Yoshida, and T. Sakaguchi.** 2001. Conserved and non-conserved regions in the Sendai virus genome: evolution of a gene possessing overlapping reading frames. Virus Genes **22:**47–52.
29. **Fukuda, Y., Y. Nakayama, and M. Tomita.** 2003. On dynamics of overlapping genes in bacterial genomes. Gene **323:**181–187.
30. **Gibrat, J. F., T. Madej, and S. H. Bryant.** 1996. Surprising similarities in structure comparison. Curr. Opin. Struct. Biol. **6:**377–385.
31. **Ginalski, K., L. Rychlewski, D. Baker, and N. V. Grishin.** 2004. Protein structure prediction for the male-specific region of the human Y chromosome. Proc. Natl. Acad. Sci. USA **101:**2305–2310.
32. **Guyader, S., and D. G. Ducray.** 2002. Sequence analysis of Potato leafroll virus isolates reveals genetic stability, major evolutionary events and differential selection pressure between overlapping reading frame products. J. Gen. Virol. **83:**1799–1807.
33. **Hardin, C., T. V. Pogorelov, and Z. Luthey-Schulten.** 2002. Ab initio protein structure prediction. Curr. Opin. Struct. Biol. **12:**176–181.
34. **Harrison, S. C.** 2007. Principles of virus structure, p. 59–98. *In* D. M. Knipe and P. M. Howley (ed.), Fields virology, 5th ed., vol. 1. Lippincott Williams & Wilkins, Philadelphia, PA.
35. **Helgstrand, C., S. Munshi, J. E. Johnson, and L. Liljas.** 2004. The refined structure of Nudaurelia capensis omega virus reveals control elements for a T = 4 capsid maturation. Virology **318:**192–203.
36. **Hong, J. R., H. Y. Gong, and J. L. Wu.** 2002. IPNV VP5, a novel anti-apoptosis gene of the Bcl-2 family, regulates Mcl-1 and viral protein expression. Virology **295:**217–229.
37. **Hughes, A. L., K. Westover, J. da Silva, D. H. O'Connor, and D. I. Watkins.** 2001. Simultaneous positive and purifying selection on overlapping reading frames of the *tat* and *vpr* genes of simian immunodeficiency virus. J. Virol. **75:**7966–7972.
38. **Hurst, L. D.** 2002. The Ka/Ks ratio: diagnosing the form of sequence evolution. Trends Genet. **18:**486.
39. **Jaroszewski, L., L. Rychlewski, Z. Li, W. Li, and A. Godzik.** 2005. FFAS03: a server for profile-profile sequence alignments. Nucleic Acids Res. **33:** W284–W288.
40. **Jordan, I. K., B. A. T. Sutter, and M. A. McClure.** 2000. Molecular evolution of the Paramyxoviridae and Rhabdoviridae multiple-protein-encoding P gene. Mol. Biol. Evol. **17:**75–86.
41. **Karlin, D., F. Ferron, B. Canard, and S. Longhi.** 2003. Structural disorder and modular organization in Paramyxovirinae N and P. J. Gen. Virol. **84:**3239–3252.
42. **Karlin, D., S. Longhi, V. Receveur, and B. Canard.** 2002. The N-terminal domain of the phosphoprotein of Morbilliviruses belongs to the natively unfolded class of proteins. Virology **296:**251–262.
43. **Keese, P. K., and A. Gibbs.** 1992. Origins of genes: "big bang" or continuous creation? Proc. Natl. Acad. Sci. USA **89:**9489–9493.
44. **Kendall, A., M. McDonald, W. Bian, T. Bowles, S. C. Baumgarten, J. Shi, P. L. Stewart, E. Bullitt, D. Gore, T. C. Irving, W. M. Havens, S. A. Ghabrial, J. S. Wall, and G. Stubbs.** 2008. Structure of flexible filamentous plant viruses. J. Virol. **82:**9546–9554.
45. **Koonin, E. V., A. E. Gorbalenya, M. A. Purdy, M. N. Rozanov, G. R. Reyes, and D. W. Bradley.** 1992. Computer-assisted assignment of functional domains in the nonstructural polyprotein of hepatitis E virus: delineation of an additional group of positive-strand RNA plant and animal viruses. Proc. Natl. Acad. Sci. USA **89:**8259–8263.
46. **Krakauer, D. C.** 2000. Stability and evolution of overlapping genes. Evolution **54:**731–739.
47. **Kretzschmar, E., R. Peluso, M. J. Schnell, M. A. Whitt, and J. K. Rose.** 1996. Normal replication of vesicular stomatitis virus without C proteins. Virology **216:**309–316.
48. **Krishnamurthy, K., M. Heppler, R. Mitra, E. Blancaflor, M. Payton, R. S. Nelson, and J. Verchot-Lubicz.** 2003. The Potato virus X TGBp3 protein associates with the ER network for virus cell-to-cell movement. Virology **309:**135–151.
49. **Krissinel, E., and K. Henrick.** 2004. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr. D **60:**2256–2268.
50. **Kulkarni-Kale, U., S. G. Bhosle, G. S. Manjari, M. Joshi, S. Bansode, and A. S. Kolaskar.** 2006. Curation of viral genomes: challenges, applications and the way forward. BMC Bioinformatics **7**(Suppl. 5)**:**S12.
51. **Lee, S. K., and D. L. Hacker.** 2001. In vitro analysis of an RNA binding site within the N-terminal 30 amino acids of the southern cowpea mosaic virus coat protein. Virology **286:**317–327.
52. **Letunic, I., R. R. Copley, B. Pils, S. Pinkert, J. Schultz, and P. Bork.** 2006. SMART 5: domains in the context of genomes and networks. Nucleic Acids Res. **34:**D257–D260.
53. **Li, F., and S. W. Ding.** 2006. Virus counterdefense: diverse strategies for evading the RNA-silencing immunity. Annu. Rev. Microbiol. **60:**503–531.
54. **Liang, X. Z., A. P. Lucy, S. W. Ding, and S. M. Wong.** 2002. The p23 protein of hibiscus chlorotic ringspot virus is indispensable for host-specific replication. J. Virol. **76:**12312–12319.
55. **Long, M., E. Betran, K. Thornton, and W. Wang.** 2003. The origin of new genes: glimpses from the young and old. Nat. Rev. Genet. **4:**865–875.
56. **Magden, J., N. Takeda, T. Li, P. Auvinen, T. Ahola, T. Miyamura, A. Merits, and L. Kaariainen.** 2001. Virus-specific mRNA capping enzyme encoded by hepatitis E virus. J. Virol. **75:**6249–6255.
57. **Malik, H. S., and T. H. Eickbush.** 2001. Phylogenetic analysis of ribonuclease H domains suggests a late, chimeric origin of LTR retrotransposable elements and retroviruses. Genome Res. **11:**1187–1197.
58. **Mayo, M. A., and A. L. Haenni.** 2006. Report from the 36th and the 37th meetings of the Executive Committee of the International Committee on Taxonomy of Viruses. Arch. Virol. **151:**1031–1037.
59. **McGirr, K. M., and G. C. Buehuring.** 2006. Tax & rex: overlapping genes of the Deltaretrovirus group. Virus Genes **32:**229–239.
60. **Meier, C., A. R. Aricescu, R. Assenberg, R. T. Aplin, R. J. Gilbert, J. M. Grimes, and D. I. Stuart.** 2006. The crystal structure of ORF-9b, a lipid binding protein from the SARS coronavirus. Structure **14:**1157–1165.
61. **Melcher, U.** 2000. The '30K' superfamily of viral movement proteins. J. Gen. Virol. **81:**257–266.
62. **Mills, R., M. Rozanov, A. Lomsadze, T. Tatusova, and M. Borodovsky.** 2003. Improving gene annotation of complete viral genomes. Nucleic Acids Res. **31:**7041–7055.
63. **Mizokami, M., E. Orito, K. Ohba, K. Ikeo, J. Y. Lau, and T. Gojobori.** 1997. Constrained evolution with respect to gene overlap of hepatitis B virus. J. Mol. Evol. **44**(Suppl. 1)**:**S83–S90.
64. **Narechania, A., M. Terai, and R. D. Burk.** 2005. Overlapping reading frames in closely related human papillomaviruses result in modular rates of selection within E2. J. Gen. Virol. **86:**1307–1313.
65. **Obradovic, Z., K. Peng, S. Vucetic, P. Radivojac, C. J. Brown, and A. K. Dunker.** 2003. Predicting intrinsic disorder from amino acid sequence. Proteins **53**(Suppl. 6)**:**566–572.
66. **Pavesi, A.** 2000. Detection of signature sequences in overlapping genes and prediction of a novel overlapping gene in hepatitis G virus. J. Mol. Evol. **50:**284–295.
67. **Pavesi, A.** 2006. Origin and evolution of overlapping genes in the family Microviridae. J. Gen. Virol. **87:**1013–1017.
68. **Pavesi, A., B. De Iaco, M. I. Granero, and A. Porati.** 1997. On the informational content of overlapping genes in prokaryotic and eukaryotic viruses. J. Mol. Evol. **44:**625–631.
69. **Peng, K., P. Radivojac, S. Vucetic, A. K. Dunker, and Z. Obradovic.** 2006. Length-dependent prediction of protein intrinsic disorder. BMC Bioinformatics **7:**208.
70. **Ribrioux, S., A. Brungger, B. Baumgarten, K. Seuwen, and M. R. John.** 2008. Bioinformatics prediction of overlapping frameshifted translation products in mammalian transcripts. BMC Genomics **9:**122.
71. **Rogozin, I. B., A. N. Spiridonov, A. V. Sorokin, Y. I. Wolf, I. K. Jordan, R. L. Tatusov, and E. V. Koonin.** 2002. Purifying and directional selection in overlapping prokaryotic genes. Trends Genet. **18:**228–232.
72. **Rozanov, M. N., E. V. Koonin, and A. E. Gorbalenya.** 1992. Conservation of the putative methyltransferase domain: a hallmark of the 'Sindbis-like' supergroup of positive-strand RNA viruses. J. Gen. Virol. **73:**2129–2134.
73. **Rui, E., P. R. Moura, A. Goncalves Kde, and J. Kobarg.** 2005. Expression and spectroscopic analysis of a mutant hepatitis B virus onco-protein HBx without cysteine residues. J. Virol. Methods **126:**65–74.
74. **Sadreyev, R. I., M. Tang, B. H. Kim, and N. V. Grishin.** 2007. COMPASS server for remote homology inference. Nucleic Acids Res. **35:**W653–W658.
75. **Sander, C., and G. E. Schulz.** 1979. Degeneracy of the information contained in amino acid sequences: evidence from overlaid genes. J. Mol. Evol. **13:**245–252.
76. **Sanger, F., G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, C. A. Fiddes, C. A. Hutchison, P. M. Slocombe, and M. Smith.** 1977. Nucleotide sequence of bacteriophage phi X174 DNA. Nature **265:**687–695.
77. **Sanz, A. I., A. Fraile, J. M. Gallego, J. M. Malpica, and F. Garcia-Arenal.** 1999. Genetic variability of natural populations of cotton leaf curl geminivirus, a single-stranded DNA virus. J. Mol. Evol. **49:**672–681.
78. **Schlessinger, A., M. Punta, and B. Rost.** 2007. Natively unstructured regions in proteins identified from contact predictions. Bioinformatics **23:** 2376–2384.
79. **Schneider, P. A., A. Schneemann, and W. I. Lipkin.** 1994. RNA splicing in Borna disease virus, a nonsegmented, negative-strand RNA virus. J. Virol. **68:**5007–5012.
80. **Scholthof, H. B.** 2006. The Tombusvirus-encoded P19: from irrelevance to elegance. Nat. Rev. Microbiol. **4:**405–411.
81. **Shi, J., T. L. Blundell, and K. Mizuguchi.** 2001. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J. Mol. Biol. **310:**243–257.
82. **Sickmeier, M., J. A. Hamilton, T. LeGall, V. Vacic, M. S. Cortese, A. Tantos, B. Szabo, P. Tompa, J. Chen, V. N. Uversky, Z. Obradovic, and A. K. Dunker.** 2007. DisProt: the Database of Disordered Proteins. Nucleic Acids Res. **35:**D786–E793.

83. **Skalka, A. M., and S. P. Goff.** 1993. Reverse transcriptase. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
84. **Smith, T. F., and M. S. Waterman.** 1981. Overlapping genes and information theory. J. Theor. Biol. **91:**379–380.
85. **Smith, T. F., and M. S. Waterman.** 1980. Protein constraints induced by multiframe encoding. Math. Biosci. **49:**17–26.
86. **Soding, J., A. Biegert, and A. N. Lupas.** 2005. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. **33:**W244–W248.
87. **Stricher, F., L. Martin, and C. Vita.** 2006. Design of miniproteins by the transfer of active sites onto small-size scaffolds. Methods Mol. Biol. **340:** 113–149.
88. **Su, C. T., C. Y. Chen, and C. M. Hsu.** 2007. iPDA: integrated protein disorder analyzer. Nucleic Acids Res. **35:**W465–W472.
89. **Suzuki, N., M. Sugawara, D. L. Nuss, and Y. Matsuura.** 1996. Polycistronic (tri- or bicistronic) phytoreoviral segments translatable in both plant and insect cells. J. Virol. **70:**8155–8159.
90. **Szilagyi, A., D. Gyorffy, and P. Zavodszky.** 2008. The twilight zone between protein order and disorder. Biophys. J. **95:**1612–1626.
91. **Taliansky, M., I. M. Roberts, N. Kalinina, E. V. Ryabov, S. K. Raj, D. J. Robinson, and K. J. Oparka.** 2003. An umbraviral protein, involved in long-distance RNA movement, binds viral RNA and forms unique, protective ribonucleoprotein complexes. J. Virol. **77:**3031–3040.
92. **Taliansky, M. E., and D. J. Robinson.** 2003. Molecular biology of umbraviruses: phantom warriors. J. Gen. Virol. **84:**1951–1960.
93. **Tan, D. Y., M. Hair Bejo, I. Aini, A. R. Omar, and Y. M. Goh.** 2004. Base usage and dinucleotide frequency of infectious bursal disease virus. Virus Genes **28:**41–53.
94. **Taylor, J. S., and J. Raes.** 2004. Duplication and divergence: the evolution of new genes and old ideas. Annu. Rev. Genet. **38:**615–643.
95. **Toll-Riera, M., N. Bosch, N. Bellora, R. Castelo, L. Armengol, X. Estivill, and M. M. Alba.** 2009. Origin of primate orphan genes: a comparative genomics approach. Mol. Biol. Evol. **26:**603–612.
96. **Tompa, P., and M. Fuxreiter.** 2008. Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. Trends Biochem. Sci. **33:**2–8.
97. **Torrance, L., and M. A. Mayo.** 1997. Proposed re-classification of furoviruses. Arch. Virol. **142:**435–439.
98. **Torresi, J.** 2002. The virological and clinical significance of mutations in the overlapping envelope and polymerase genes of hepatitis B virus. J. Clin. Virol. **25:**97–106.
99. **Upton, C.** 2000. Screening predicted coding regions in poxvirus genomes. Virus Genes **20:**159–164.
100. **Uversky, V. N.** 2002. Natively unfolded proteins: a point where biology waits for physics. Protein Sci. **11:**739–756.
101. **Uversky, V. N., P. Radivojac, L. M. Iakoucheva, Z. Obradovic, and A. K. Dunker.** 2007. Prediction of intrinsic disorder and its use in functional proteomics. Methods Mol. Biol. **408:**69–92.
102. **Vacic, V., V. N. Uversky, A. K. Dunker, and S. Lonardi.** 2007. Composition Profiler: a tool for discovery and visualization of amino acid composition differences. BMC Bioinformatics **8:**211.
103. **van der Heijden, M. W., and J. F. Bol.** 2002. Composition of alphavirus-like replication complexes: involvement of virus and host encoded proteins. Arch. Virol. **147:**875–898.
104. **van Eyll, O., and T. Michiels.** 2002. Non-AUG-initiated internal translation of the L* protein of Theiler's virus and importance of this protein for viral persistence. J. Virol. **76:**10665–10673.
105. **Vargason, J. M., G. Szittya, J. Burgyan, and T. M. Hall.** 2003. Size selective recognition of siRNA by an RNA silencing suppressor. Cell **115:**799–811.
106. **Wang, L. F., W. P. Michalski, M. Yu, L. I. Pritchard, G. Crameri, B. Shiell, and B. T. Eaton.** 1998. A novel P/V/C gene in a new member of the *Paramyxoviridae* family, which causes lethal infection in humans, horses, and other animals. J. Virol. **72:**1482–1490.
107. **Watters, A. L., and D. Baker.** 2004. Searching for folded proteins in vitro and in silico. Eur. J. Biochem. **271:**1615–1622.
108. **Weber, S., D. Fichtner, T. C. Mettenleiter, and E. Mundt.** 2001. Expression of VP5 of infectious pancreatic necrosis virus strain VR299 is initiated at the second in-frame start codon. J. Gen. Virol. **82:**805–812.
109. **Wehner, T., A. Ruppert, C. Herden, K. Frese, H. Becht, and J. A. Richt.** 1997. Detection of a novel Borna disease virus-encoded 10 kDa protein in infected cells and tissues. J. Gen. Virol. **78:**2459–2466.
110. **Wilson, G. A., N. Bertrand, Y. Patel, J. B. Hughes, E. J. Feil, and D. Field.** 2005. Orphans as taxonomically restricted and ecologically important genes. Microbiology **151:**2499–2501.
111. **Xie, H., S. Vucetic, L. M. Iakoucheva, C. J. Oldfield, A. K. Dunker, V. N. Uversky, and Z. Obradovic.** 2007. Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. J. Proteome Res. **6:**1882–1898.
112. **Yamauchi, A., T. Yomo, F. Tanaka, I. D. Prijambada, S. Ohhashi, K. Yamamoto, Y. Shima, K. Ogasahara, K. Yutani, M. Kataoka, and I. Urabe.** 1998. Characterization of soluble artificial proteins with random sequences. FEBS Lett. **421:**147–151.
113. **Yang, Z. R., R. Thomson, P. McNeil, and R. M. Esnouf.** 2005. RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. Bioinformatics **21:**3369–3376.
114. **Yokoo, H., and T. Oshima.** 1979. Is bacteriophage φX174 DNA a message from an extraterrestrial intelligence? Icarus **38:**148–153.
115. **Yooseph, S., G. Sutton, D. B. Rusch, A. L. Halpern, S. J. Williamson, K. Remington, J. A. Eisen, K. B. Heidelberg, G. Manning, W. Li, L. Jaroszewski, P. Cieplak, C. S. Miller, H. Li, S. T. Mashiyama, M. P. Joachimiak, C. van Belle, J. M. Chandonia, D. A. Soergel, Y. Zhai, K. Natarajan, S. Lee, B. J. Raphael, V. Bafna, R. Friedman, S. E. Brenner, A. Godzik, D. Eisenberg, J. E. Dixon, S. S. Taylor, R. L. Strausberg, M. Frazier, and J. C. Venter.** 2007. The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. PLoS Biol. **5:**e16.
116. **Zeldovich, K. B., and E. I. Shakhnovich.** 2008. Understanding protein evolution: from protein physics to Darwinian selection. Annu. Rev. Phys Chem. **59:**105–127.
117. **Zhang, Y., B. Stec, and A. Godzik.** 2007. Between order and disorder in protein structures: analysis of "dual personality" fragments in proteins. Structure **15:**1141–1147.
118. **Zhou, Q., G. Zhang, Y. Zhang, S. Xu, R. Zhao, Z. Zhan, X. Li, Y. Ding, S. Yang, and W. Wang.** 2008. On the origin of new genes in Drosophila. Genome Res. **18:**1446–1455.
119. **Zhou, T., Z. F. Fan, H. F. Li, and S. M. Wong.** 2006. Hibiscus chlorotic ringspot virus p27 and its isoforms affect symptom expression and potentiate virus movement in kenaf (Hibiscus cannabinus L.). Mol. Plant-Microbe Interact. **19:**948–957.