doi:10.1016/j.jmb.2006.03.043 J. Mol. Biol. (2006) 359, 496–507

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Ig-Like Domains on Bacteriophages: A Tale of Promiscuity and Deceit

James S. Fraser[1], Zhou Yu[1], Karen L. Maxwell[1] and Alan R. Davidson[1,2]*

[1]*Department of Molecular and Medical Genetics, University of Toronto, Toronto, Ont., Canada M5S 1A8*

[2]*Department of Biochemistry University of Toronto, Toronto Ont., Canada M5S 1A8*

**Corresponding author*

The immunoglobulin (Ig) fold is one of the most important structures in biology, playing essential roles in the vertebrate immune response, cell adhesion, and many other processes. Through bioinformatic analysis, we have discovered that Ig-like domains are often found in the constituent proteins of tailed double-stranded (ds) DNA bacteriophage particles, and are likely displayed on the surface of these viruses. These phage Ig-like domains fall into three distinct sequence families, which are similar to the classic immunoglobulin domain (I-Set), the fibronectin type 3 repeat (FN3), and the bacterial Ig-like domain (Big2). The phage Ig-like domains are very promiscuous. They are attached to more than ten different functional classes of proteins, and found in all three morphogenetic classes of tailed dsDNA phages. In addition, they reside in phages that infect a diverse set of Gram negative and Gram positive bacteria. These domains are deceptive because many are added to larger proteins through programmed ribosomal frameshifting, so that they are not always detected by standard protein sequence searching procedures. In addition, the presence of unrecognized Ig-like domains in a variety of phage proteins with different functions has led to gene misannotation. Our results demonstrate that horizontal gene transfer involving Ig-like domain encoding DNA has occurred commonly between diverse classes of both lytic and temperate phages, which otherwise display very limited sequence similarities to one another. We suggest that phage may have been an important vector in the spread of Ig-like domains through diverse species of bacteria. While the function of the phage Ig-like domains is unknown, several lines of evidence suggest that they may play an accessory role in phage infection by weakly interacting with carbohydrates on the bacterial cell surface.

Keywords: bacteriophage; Ig-like domain; morphogenetic proteins; misannotation; frameshifting

Introduction

Recent studies have demonstrated that bacteriophages are the most abundant and most rapidly reproducing biological entities on the planet. Consequently, they have a profound impact on the ecology and evolution of their bacterial hosts.[1,2] This realization, coupled with the rapid increase in the number of sequenced phage genomes, has led to an explosion of interest in the field of phage genomics.[3] Phage genomes have been found to be highly mosaic as a result of rampant horizontal exchange of sequences through both homologous and non-homologous recombination. Due to the extreme divergence of phage genomes, it is often difficult to detect sequence similarity between genes that are likely homologous. However, the strong conservation of gene order in these genomes can facilitate the identification of their identification. The complexity and great relevance to bacterial biology of phage genomes makes them a fascinating subject for bioinformatic investigation.

The sequence diversity of phage genomes has made the identification of protein domains or functions that are conserved in large numbers of

Abbreviations used: ds, double-stranded; ss, single-stranded; ORF, open reading frame; Ig, immunoglobulin; HOC, highly immunogenic outer capsid; MTP, major tail protein; MHP, major head protein.

E-mail address of the corresponding author: alan.davidson@utoronto.ca

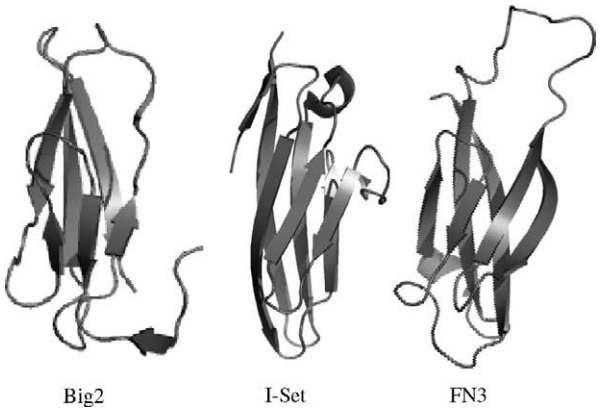

Figure 1. Structures of Ig-like domains representative of the Big 2, I-Set and FN3 families. The Big 2 domain structure (1F00) is from Intimin, the I-Set domain structure (1FHG_A) is from Telokin, and the FN3 domain structure (1GH7_A) is from the extracellular domain of the beta-common receptor of Il-3, Il-5, and Gm-Csf.

different phages quite difficult.[4] Furthermore, there have been few studies in which phage genomes have been systematically searched for specific conserved protein features. Here, we have performed a thorough examination of the occurrence and potential function of Ig-like domains in phages. The Ig-like fold, which is one of the most common and widely dispersed folds in nature,[5] is comprised of at least seven β-strands arranged into two distinct sheets packed in a parallel manner (Figure 1).[6] Although Ig-like domains are defined by a common basic topology, their amino acid sequences are highly diverged, and thus often appear unrelated at the sequence level. While the general functional role of Ig-like domains is predominantly in binding reactions, the specific reactions mediated by these domains vary widely. They can bind small molecules, hormones, or large protein complexes through homo- and heterophilic interactions.[7] The binding sites on these domains are located on the surfaces of the sheets or in the loops that connect the strands. Whether all Ig-like domains evolved from a single ancestor or if many unrelated domains converged on this common structure is not known.

Proteins possessing the Ig-like fold are commonly found in bacteria,[8] and are most often involved in cell-cell adhesion,[9] or extracellular glycohydrolysis.[10] Although Ig-like domains have been previously recognized to occur in some bacteriophage proteins,[11,12] the prevalence and possible roles of Ig-like domains in phage have not been investigated. Here, through bioinformatic studies, we demonstrate that Ig-like domains are commonly found in structural proteins of tailed double-stranded (ds) DNA phages. The properties and potential functions of these domains are described.

Results

Ig-like domains are found commonly in structural proteins of tailed dsDNA phages

Our laboratory is currently pursuing structural and biophysical studies on the major tail protein (MTP) of phage lambda, gpV. Previous electron microscopic and genetic experiments on this 246 amino acid residue protein indicated that its C-terminal third comprised a distinct domain.[13] A PSI-BLAST[14] search performed to aid in delineating this domain revealed that the C-terminal portion of gpV showed significant sequence similarity to a variety of immunoglobulin (Ig)-like domains of the bacterial Ig-like domain family (Big 2 family, as defined in the PFAM database[15]). Based on these BLAST results, the last 86 residues of gpV were expressed in *Escherichia coli*, purified, and characterized in our laboratory. This region of gpV was found to form an independently folded domain, and determination of its three-dimensional structure by NMR spectroscopy has demonstrated that it does indeed possess an Ig-like fold (unpublished results).

The PSI-Blast search results with gpV revealed significant sequence similarities between its Ig-like domain and many other phage proteins. Surprisingly, many of these proteins were completely unrelated to gpV except in the region containing the Ig-like domain, and they were from phages infecting diverse genera of Gram positive and negative bacteria, such as Yersinia, Bacillus, Staphylococcus, and Listeria. In order to ascertain how frequent the occurrence of Ig-like domains might be in all bacteriophage genomes, and what types of phage proteins might contain Ig-like domains, we performed exhaustive searches of the completely sequenced phage genomes for Ig-like domains. Hidden Markov Models (HMMs) derived from 36 families of Ig-like domains defined in the PFAM database (Table 1) were used to query 246 sequenced phage genomes. Since one Ig-like domain found in our initial PSI-BLAST search with gpV is added to a head protein of phage T3 through a translational frameshifting mechanism,[16] we searched DNA sequences translated in all reading frames using protein profiles to ensure the identification of unrecognized or unannotated Ig-like domains that might be appended by frameshifting.

Through our sequence searching protocol, we ultimately identified 68 Ig-like domains located in 54 proteins in 41 different tailed dsDNA phage (*Caudovirales*) genomes. Interestingly, no Ig-like domains were discovered in either ssDNA or RNA phages, or in other classes of dsDNA phages. Ig-like domains were found in all three families of *Caudovirales*: those with long non-contractile tails (*Siphoviridae*), those with contractile tails (*Myoviridae*), and those with short tails (*Podoviridae*). They occur in both lytic and temperate phages, and

Table 1. PFAM HMMs used to search for Ig-like domains in phage genomes

SCOP superfamily	PFAM families
Immunoglulin	V-set (PF07686), I-set (PF07679), C2-set (PF05790), C1-set (PF07654), Ig (PF00047), ICAM_N (PF03921)
E-Set	Alpha_amylase_N (PF02903), arrestin_N (PF00339), arrestin_C (PF02752), CelD_N (PF02927), peptidaseC25 (PF03785), TIG (PF01833), RHD (PF00554), DUF291 (PF03442), filamin (PF00630), He_Pig (PF05345)
Fibronectin Type III	FN3 (PF00041), tissue_fac (PF01108), lep_receptor_Ig (PF06328)
PKD	PKD (PF00801), PPC (PF04151), HYR (PF02494)
β-Galactosidase/glucuronidase	Glycol_hydro_2 (PF00703)
Cu,Zn superoxide dismutase-like	Sod_Cu (PF00080)
PapD-like	Pili_assembly_C (PF02753), pili_assembly_N (PF00345)
Invasin/intimin cell-adhesion fragments	Big_1 (PF02369), Big_2 (PF02368), Big_3 (PF07523), Big_4 (PF07532)
Clathrin adaptor appendage domain	Alpha_adaptin_C2 (PF02883)
Transglutaminase N-terminal domain	Transglut_N (PF00868)
Cadherin-like	Cadherin domain (PF00028)
Actinoxanthin-like	Neocarzinostatin family (PF00960)
CBD9-like	Domain of unknown function (PF06452)
Lamin A/C globular tail domain	Intermediate filament tail domain (PF00932)

approximately one quarter of the sequenced tailed dsDNA phage genomes encode at least one Ig-like domain. It is important to note that these phage genomes are spectacularly diverse,[2] and the detection of similar sequences within their genomes is very unusual. Although Ig-like domains are found widely in the tailed dsDNA phages, they are found predominantly in only five functional classes of proteins: tail fiber, baseplate wedge initiator, major tail, major head, and highly immunogenic outer capsid (HOC) proteins (Figure 2(a)). These proteins are all structural components of the phage particle. Details on each of the phage Ig-like domains that we discovered are given in Table 2.

Phage Ig-like domains fall within three distinct sequence families

The 68 phage Ig-like domains can be broadly classified into three sequence families based on similarities to the PFAM sequence families designated bacterial Ig-like domain 2 (Big 2), Immunoglobulin superfamily (I-Set), and fibronectin type III domain (FN3). As is seen in Figure 3, these sequence families are quite distinct from one another, and we observed no cross-hits between them in any BLAST searches. While there are clusters of closely related domains within each sequence family, the overall diversity of the phage Ig-like

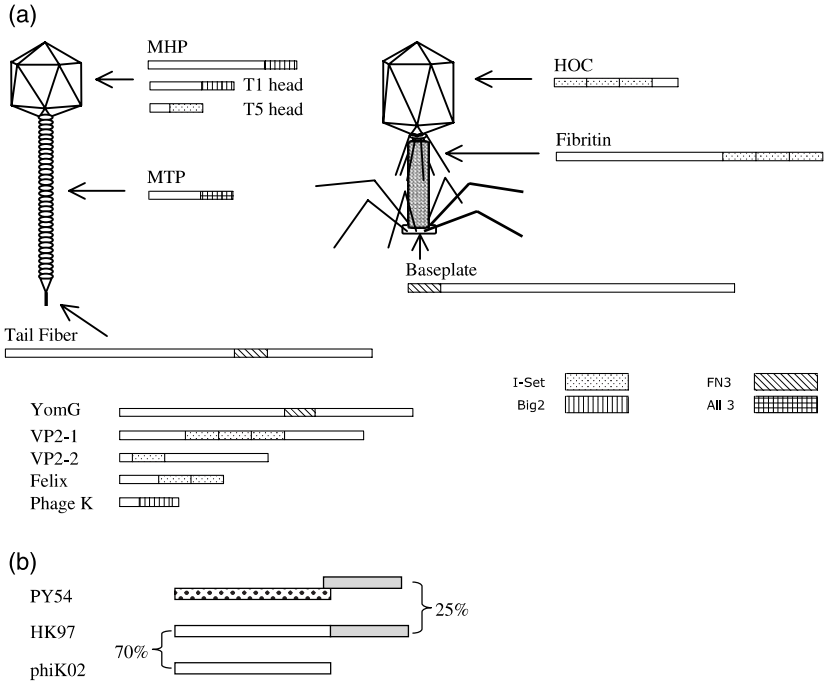

Figure 2. The variety of phage proteins possessing one or more Ig-like domains. (a) Schematic diagrams of a typical Siphovirida (left) and Myovirida (right) are shown, and the locations of various Ig-like domain-containing proteins are shown. Ig-like domains are shaded as indicated. Proteins with unknown locations are listed at the bottom. Explanations of protein names are found in Table 2. (b) The interesting relationships between the MTPs of phages PY54, HK97, and phiK02 are illustrated.

Table 2. Summary of data for phage proteins containing Ig-like domains

GI number	Ig-like domain family	Ig-like domain number	Abbreviation	Protein function	Phage	Phage class	Bacterial host	Host Gram staining	Protein length	Frameshift mechanism
30267419	I-Set	3	FIB-RB43	Fibritin	Enterobacteria phage RB43	Myo	*Escherichia coli*	–	762	
46401880	I-Set	1	HP-T5	Head protein[a]	Bacteriophage T5	Sipho	*Escherichia coli*	–	164	
37651653	I-Set	1	HOC-44	HOC	Bacteriophage 44RR2.8t	Myo	*Aeromonas salmonicida*	–	180	
33620536	I-Set	3	HOC-RB49	HOC	Enterobacteria phage RB49	Myo	*Escherichia coli*	–	404	
32453675	I-Set	4	HOC-RB69	HOC	Enterobacteria phage RB69	Myo	*Escherichia coli*	–	471	
9632750	I-Set	3	HOC-T4	HOC	Enterobacteria phage T4	Myo	*Escherichia coli*	–	376	
48696642	I-Set	3	LM1-VP2	Likely morphogenetic[b]	Vibriophage VP2	Sipho	*Vibrio cholerae*	–	741	
48696686	I-Set	3	LM1-VP5	Likely morphogenetic[b]	Vibriophage VP5	Sipho	*Vibrio cholerae*	–	743	
48696646	I-Set	1	LM2-VP1	Likely morphogenetic[b]	Vibriophage VP2	Sipho	*Vibrio cholerae*	–	460	
50282965	I-Set	1	LM2-VP5	Likely morphogenetic[b]	Vibriophage VP5	Sipho	*Vibrio cholerae*	–	444	
9634159	I-Set	1	MTP-97	Major tail protein	Bacteriophage HK97	Sipho	*Escherichia coli*	–	234	
49523647	I-Set	1	MTP-47	Major tail protein	Phage phi 4795	Sipho	*Escherichia coli*	–	238	
33770520	I-Set	1	MTP-PY	Major tail protein	Bacteriophage PY54	Sipho	*Yersinia enterocolitica*	–	236	−1
38707688	I-Set	2	UN-FE	Unknown[c]	Bacteriophage Felix 01	Myo	*Salmonella*	–	294	
45686313	Big2	1	LM-T1	Likely morphogenetic[b]	Enterobacteria phage T1	Sipho	*Escherichia coli*	–	255	
12248115	Big2	1	MHP-GA	Major head protein	Bacillus phage GA-1	Podo	*Bacillus subtilis*	+	472	
9626396	Big2	1	MHP-29	Major head protein	Bacillus phage phi29	Podo	*Bacillus subtilis*	+	448	
22855155	Big2	1	MHP-B103	Major head protein	Bacteriophage B103	Podo	*Bacillus subtilis*	+	449	
9634035	Big2	1	MHP-YE	Major head protein	Bacteriophage phiYeO3-12	Podo	*Yersinia enterocolitica*	–	426	−1
17570826	Big2	1	MHP-T3	Major head protein	Bacteriophage T3	Podo	*Escherichia coli*	–	433	−1
40806231	Big2	1	MHP-A2	Major head protein	*Lactobacillus casei* bacteriophage A2	Sipho	*Lactobacillus casei*	+	485	−1
21700277	Big2	1	MHP-UL	Major head protein	*Lactococcus lactis* bacteriophage ul36	Sipho	*Lactococcus lactis*	+	372	−1
41189535	Big2	1	MTP-77	Major tail protein	Bacteriophage 77	Sipho	*Staphylococcus aureus*	+	302	−1
30043945	Big2	1	MTP-N3	Major tail protein	Staphylococcus phage phiN315	Sipho	*Staphylococcus aureus*	+	302	−1
22217807	Big2	1	MTP-A2	Major tail protein	*Lactobacillus casei* bacteriophage A2	Sipho	*Lactobacillus casei*	+	286	−1
16798795	Big2	1	MTP-A1	Major tail protein	Bacteriophage A118	Sipho	*Listeria monocytogenes*	+	231	+1
9626256	Big2	1	MTP-LAM	Major tail protein	Bacteriophage lambda	Sipho	*Escherichia coli*	–	246	
9630478	Big2	1	MTP-N15	Major tail protein	Bacteriophage N15	Sipho	*Escherichia coli*	–	246	
45775052	Big2	1	MTP-T5	Major tail protein	Bacteriophage T5	Sipho	*Escherichia coli*	–	464	
9635177	Big2	1	MTP-PV	Major tail protein	*Staphylococcus aureus* bacteriophage PVL	Sipho	*Staphylococcus aureus*	+	317	
29028706	Big2	1	MTP-13	Major tail protein	*Staphylococcus aureus* phage phi 13	Sipho	*Staphylococcus aureus*	+	317	
29028654	Big2	1	MTP-12	Major tail protein	*Staphylococcus aureus* phage phi 12	Sipho	*Staphylococcus aureus*	+	395	+1
12719438	Big2	1	MTP-SL	Major tail protein	*Staphylococcus aureus* temperate phage phiSLT	Sipho	*Staphylococcus aureus*	+	395	+1
48696482	Big2	1	UN-K	Unknown[c]	Staphylococcus phage K	Myo	*Staphylococcus aureus*	+	170	
37651623	FN3	1	BP-44	Baseplate wedge initiator	Bacteriophage 44RR2.8t	Myo	*Aeromonas salmonicida*	–	1019	
38640133	FN3	1	BP-AE	Baseplate wedge initiator	Bacteriophage Aeh1	Myo	*Aeromonas salmonicida*	–	1163	

Table 2 (*continued*)

GI number	Ig-like domain family	Ig-like domain number	Abbreviation	Protein function	Phage	Phage class	Bacterial host	Host Gram staining	Protein length	Frameshift mechanism
33620635	FN3	1	BP-RB49	Baseplate wedge initiator	Enterobacteria phage RB49	Myo	*Escherichia coli*	–	1028	
32453649	FN3	1	BP-RB69	Baseplate wedge initiator	Enterobacteria phage RB69	Myo	*Escherichia coli*	–	1032	
137900	FN3	1	BP-T4	Baseplate wedge initiator	Enterobacteria phage T4	Myo	*Escherichia coli*	–	1032	
2764866	FN3	1	MTP-SP	Major tail protein	Bacteriophage SPP1	Sipho	*Bacillus subtilis*	+	264	+1
29567005	FN3	1	MTP-BAR	Major tail protein[d]	Mycobacteriophage Barnyard	Sipho	*Mycobacteria*	+	281	
29565897	FN3	1	MTP-CJ	Major tail protein[d]	Mycobacteriophage CJW1	Sipho	*Mycobacteria*	+	324	
29566768	FN3	1	MTP-OM	Major tail protein[d]	Mycobacteriophage Omega	Sipho	*Mycobacteria*	+	334	
9634161	FN3	1	TF-97	Tail fibre	Bacteriophage HK97	Sipho	*Escherichia coli*	–	1296	
215125	FN3	1	TF-LAM	Tail fibre	Bacteriophage lambda	Sipho	*Escherichia coli*	–	1132	
3192704	FN3	1	TF-N15	Tail fibre	Bacteriophage N15	Sipho	*Escherichia coli*	–	1061	
38505402	FN3	1	TF-1026	Tail fibre	Bacteriophage phi1026b	Sipho	*Burkholderia pseudomallei*	–	1101	
17975182	FN3	1	TF-125	Tail fibre	Bacteriophage phiE125	Sipho	*Burkholderia pseudomallei*	–	1101	
46402107	FN3	1	TF-KO2	Tail fibre	Bacteriophage phiKO2	Sipho	*Klebsiella oxytoca*	–	3433	
33770530	FN3	1	TF-PY	Tail fibre	Bacteriophage PY54	Sipho	*Yersinia enterocolitica*	–	847	
9634142	FN3	1	TF-22	Tail fibre	Enterobacteria phage HK022	Sipho	*Escherichia coli*	–	1183	
45686327	FN3	1	TF-T1	Tail fibre	Enterobacteria phage T1	Sipho	*Escherichia coli*	–	1172	
49523655	FN3	1	TF-4795	Tail fibre	Phage phi 4795	Sipho	*Escherichia coli*	–	1158	
9630155	FN3	1	UN-SP	Unknown (YomG)	Bacteriophage SPBc2	Sipho	*Bacillus subtilis*	+	875	

[a] This previously misannotated protein was assigned by us as a head protein due to its genomic position. The gene encoding this protein is surrounded by other genes encoding head proteins.
[b] These proteins, some of which were misannotated, clearly lie within morphogenetic regions of the genome though their exact functions are unclear.
[c] The gene encoding this previously misannotated protein does not lie in a region of the genome that is clearly encoding morphogenetic proteins.
[d] These proteins, the functions of which were not previously annotated, have been annotated by us as MTPs due to the genomic positions of their genes and their possession of Ig-like domains.

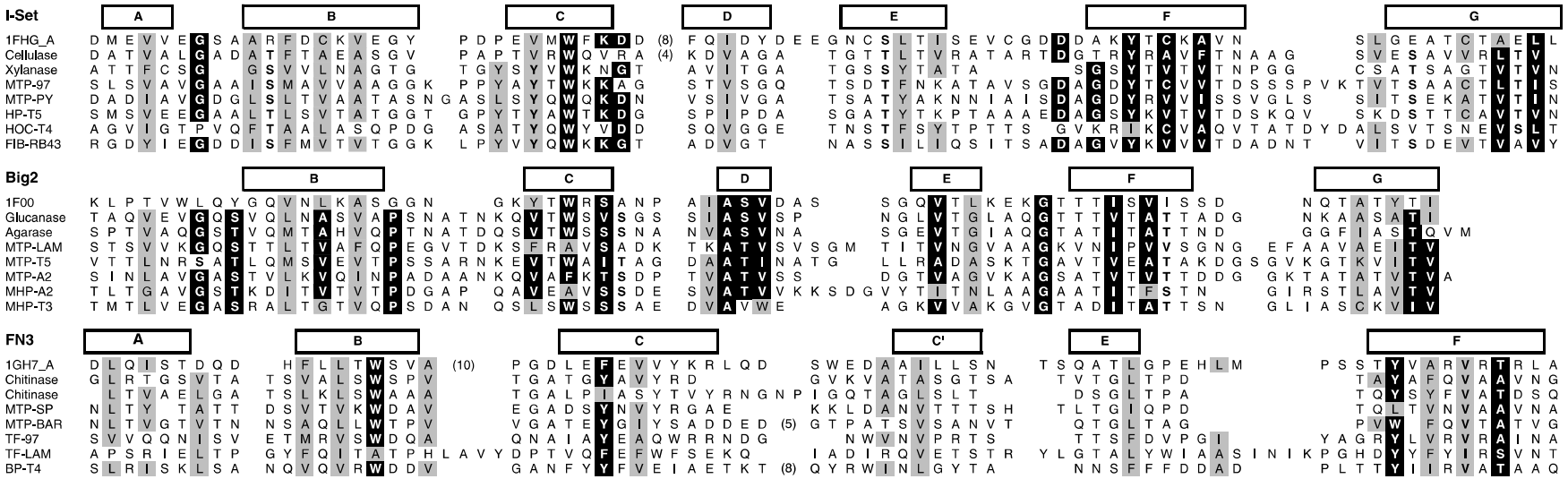

Figure 3. Representative sequence alignment of the three classes of Ig-like domains found in phage proteins. The sequences of representative domains possessing solved three-dimensional structures are also shown (see Figure 1). In addition, the sequences of Ig-like domains from various bacterial glycohydrolases are included. Positions that are highly conserved in the phage Ig-like domains and within the given Ig-like domain family are shaded in black. Conserved hydrophobic positions are shaded in gray. The positions of β-strands within the solved structures are indicated. These strands are named according to the convention for Ig-like domains.[6] In some cases, large sequence insertions have been left out. The length of these insertions is indicated in parentheses. These sequences were aligned by CLUSTALW[43] and manually adjusted to produce the optimal alignment.

domains is very high. For example, in the I-Set group of domains pairwise identities between individual domains range from over 60% to less than 10%, but the average pairwise identity in the whole group is only 17%. Within each family, most hydrophobic core positions are conserved as hydrophobic. In addition, in each alignment, several of the most conserved positions defining the given family of Ig-like domains are seen to be conserved among the phage sequences. The β-strands in these alignments are named to correspond with the conserved β-strands of the Ig-like fold as described by Bork *et al.*[6] The first and last strands are more difficult to align, as sequence and structural variation in these regions of Ig-like domains tends to be high. For example, in the PFAM alignments of Ig-like domains, the first and last strands are often not included.

Although the I-Set and FN3 family of domains are seen predominantly in eukaryotic proteins involved in the immune response, cell adhesion, or muscle contraction (e.g. titin), members of these families have also been observed in bacterial surface glycohydrolases, such as chitinases and cellulases. Representatives of these bacterial domains are shown in Figure 3. The Big 2 domain has been linked most often with bacterial adhesion proteins, such as intimins.[17] However, this domain is also found in various glycohydrolases. Interestingly, the phage Big 2 domain sequences resemble those from the glycohydrolases more closely than they resemble those from intimins (Figure 3). There is a correlation between the Ig-like domain family and the function of the phage protein in which the domain is found (Figure 2(a)). Tail fibre and baseplate proteins possess only FN3 family domains, HOC proteins and fibritin possess only I-Set family domains, and major head proteins (MHP) possess only Big 2 family domains. In contrast, major tail proteins (MTP) can possess any of the three families of Ig-like domains. While the Big 2 and FN3 domains are observed in phages infecting both Gram positive and negative bacteria, the I-Set domains are only seen in phage infecting Gram negative strains. Most of the phage proteins have only one Ig-like domain with only I-Set family domains present in multiple copies in some proteins (Figure 2(a)). It is striking that within all three classes of phage Ig-like domains, domains with similar sequences are observed to be attached to proteins that are completely unrelated in either sequence or function. For example, the Ig-like domain of the phage ul36 MHP is 35% identical with the Ig-like domain of the MTP of phage N15.

Phage Ig-like domains are often added by ribosomal frameshifting

Twelve of the Ig-like domain encoding regions identified in phage genomes were found to be overlapping with, but out of frame with respect to a nearby morphogenetic gene open reading frame (ORF). These domains are likely appended to the protein encoded by the previous ORF *via* a translational frameshifting mechanism. There have been a number of cases of programmed translational frameshifts characterized in bacteriophage, retroviral and bacterial genes reviewed.[18,19] They are caused by the ribosome shifting its reading frame at a specific site in a coding region from the one it initiated translating into a new reading frame. The mechanism of frameshifting involves specific mRNA signals that cause the ribosome to pause, allowing the ribosome-bound tRNAs to slip and subsequently pair their non-wobble bases with the out-of-frame codons. Leftward (−1) frameshifts are normally associated with a characteristic "slippery" heptanucleotide sequence located near the 3′ end of the in-frame protein coding sequence.[20] Rightward (+1) frameshifts occur less commonly and are more difficult to predict than −1 frameshifts, but can be explained with reference to "hungry" codon deficits.[21]

Five putative frameshifts that we identified involving phage Ig-like domains have been experimentally verified, including a +1 frameshift involving the MTP of bacteriophage SPP1 (P. Tavares, personal communication), and −1 frameshifts found in the MHPs of phages A2,[22] T3 [16] and phiYeO3-12[23] and the MTP of phage A2.[24] It was not previously recognized that any of these frameshifts leads to fusion with an Ig-like domain, except for the A2 major tail protein frameshift (we informed the authors of the existence of the Ig-like domain before publication of their paper). We believe that unreported frameshifts may append Ig-like domains to other phage proteins (Table 3). All three classes of Ig-like domains that we have identified in phages are seen to be added through ribosomal frameshifting, though only one case is observed for the FN3 and I-Set families. It is intriguing that frameshifting by the same mechanism can place Ig-like domains with similar sequences onto the C termini of totally unrelated proteins. For example, in the case of the ul36 head protein, its putative frameshifted Ig-like domain shares sequence identities of 34 and 40% with the Ig-like regions frameshifted onto the MHPs of phages A2 and T3, respectively. Despite this sequence similarity in their frameshifted Ig-like domains and the functional similarity of the total protein, the non-frameshifted portions do not display detectable sequence similarity.

The MTPs of phages HK97, PY54, and phiK02 provide a typical example of the complicated nature of Ig-like domain occurrence in phages (Figure 2(b)). The HK97 and PY54 MTPs possess related (25% identical) Ig-like domains, but the PY54 domain is likely added through ribosomal frameshifting.[25] The N-terminal domains of these proteins display no detectable sequence similarity. On the other hand the MTP of phiK02, a phage with morphogenetic genes similar to PY54, is 70% identical with the PY54 MTP N-terminal domain,

Table 3. Ig-like domains appended by a frameshifting mechanism

Annotation of adjacent gene	Phage	Ig-like domain class	Ig-like domain ORF identification	Frameshift direction	Likely framshift mechanism
MTP	A118	Big2	Separate ORF	+1	CCC->CCT hungry proline
MTP[a]	phi 12	Big2	Separate ORF	+1	CCC proline
MTP[a]	phiSLT	Big2	Separate ORF	+1	CCC proline
MHP	A2	Big2	Not identified[b]	−1	Slippery sequence CCCAAA
MHP[a]	phiYeO3-12	Big2	Separate ORF	−1	Slippery sequence CCCAAA
MHP[a]	T3	Big2	Separate ORF	−1	Slippery sequence CCCAAA
MHP	ul36	Big2	Not identified	−1	Unknown
MTP	A2	Big2	Not identified	−1	Slippery sequence CCCAAA
MTP[a]	77	Big2	Not identified	−1	Unknown
MTP[a]	phiN315	Big2	Not identified	−1	Unknown
MTP	SPP1	FN3	Separate ORF	+1	CCC->CCT hungry proline
MTP	PY54	I-Set	Separate ORF	−1	Slippery sequence CCCAAA

[a] These adjacent pairs of proteins are closely related (>90% identity).
[b] This frameshift is published though the ORF is not recognized in the database.

but it possesses no C-terminal Ig-like domain (Figure 2(b)). Rather, its next ORF begins directly after the end of the MTP ORF.

Phage Ig-like domains are a significant source of gene misannotation

Since the phage Ig-like domains are similar in sequence to many different bacterial cell adhesion proteins and extracellular glycohydrolases, they are a potential source of confusion for genome annotation. Given that the fast rate of evolutionary divergence in phages can often render homologues to phage proteins unrecognizable at the sequence level, the Ig-like domains may be the only part of a protein that displays a significant blast hit to other proteins. As described in Table 4, this phenomenon has led to several cases of misannotation in phage genomes. For example, genes from phages Felix 01, K, T5, and VP5 have all been annotated as "tail protein" or "putative major tail protein" apparently based solely on BLAST hits to Ig-like domains in tail genes of other phages. Closer analysis of these protein sequences and their positions in their respective genomes (phage morphogenetic genes are generally clustered and ordered according to their function), indicates that none of these annotations are likely to be correct. In another case, Mycobacteriophage Barnyard gp30 was described as a chitinase due to the sequence similarity of its Ig-like domain (FN3 family) to those found in these enzymes.[26] However, due to the position of its gene in the phage genome, we believe that it is actually the MTP. This gene lies next to a gene encoding a probable functional homologue of the lambda G-T protein, which is followed by the putative tail tape measure protein. The contiguous arrangement of MTP, lambda G-T homologue, and tape measure protein is highly conserved in Siphoviridae genomes.[27] Similar to other MTPs, the putative Barnyard MTP possesses a 180 residue N-terminal domain, and a 100 residue C-terminal Ig-like

Table 4. Database misannotations of proteins due to sequence similarities of their Ig-like domains to other phage proteins

GI Number	Ig-like domain	Phage	Database annotation	Our re-annotation and explanation
46401880	I-Set	T5	Putative tail protein	Head protein: its gene is located in the middle of the head gene cluster, and it has been demonstrated experimentally to be a head protein (44)[a]
48696686	I-Set	VP5	MTP	Unknown: no similarity to MTPs outside of Ig-like domains. Protein size and possession of 3 Ig-like domains is not consistent with other MTPs
48696646	I-Set	VP2	Outer capsid protein (HOC)	Unknown: annotation is based on similarity to Ig-like domains of HOC proteins, but Siphoviridae are not known to have HOC proteins. Gene location next to putative tail fibre gene suggests that this is a tail protein
38707688	I-Set	Felix 01	HK97 MTP	Unknown: this virus is in the Myoviridae class, so functionally meaningful sequence similarity to a Siphoviridae tail protein is extremely unlikely
48696482	Big2	Phage K	Putative MTP	Unknown: sequence similarities are to Siphoviridae tail protein Ig-like domains, but this is a Myoviridae.
29567005	FN3	Barnyard	gp30	MTP: likely an MTP due to genome position of gene. In publication describing this genome, protein was called a chitinase

These proteins show no significant sequence similarity to the proteins named in their database annotation except within their Ig-like domains.
[a] One of the three genome sequences (AY692264) of T5 in the database has correctly annotated this ORF as a head protein.

domain, which displays BLAST hits to many chitinases. However, this protein cannot be a chitinase because it possesses no region similar to the chitinase catalytic domain.[28]

In contrast to the cases described above, the presence of Ig-like domains can actually aid in the annotation of some genes. For example, gp18 of Mycobacteriophage CJW1 and gp31 of Mycobacteriophage Omega are unannotated, but their size and genomic positions are consistent with their being MTPs. Since they also possess Ig-like domains at their C termini, as do 16 other phage MTPs, the case for annotation of these genes as MTPs becomes quite strong.

Discussion

Through bioinformatic studies, we have demonstrated that Ig-like domains are found commonly in tailed dsDNA bacteriophages, with 68 Ig-like domains appearing in 41 different genomes. A remarkable aspect of this discovery is that related Ig-like domains are seen to be constituents of otherwise completely unrelated proteins in unrelated bacteriophages. In addition, tailed dsDNA phages are shown to possess three distinct families of phage Ig-like domains. These observations imply that the DNA sequences encoding Ig-like domains have not arisen from a single ancestral gene, but that Ig-like domain encoding DNA segments have been shuttled from one phage genome to other unrelated phage genomes through non-homologous recombination events. This type of "domain shuffling" has been described in bacteria,[29] and has been invoked for the spread of FN3 domains in bacterial glycohydrolases.[10]

The presence of diverse Ig-like domains in a wide variety of phages has ramifications for the evolution of phages and bacteria. The occurrence of domains with similar sequence in all three morphogenetic classes of tailed dsDNA phages, and in both lytic and temperate phages, demonstrates that all of these groups must to some extent share a common gene pool. Since these phages also infect a wide variety of both Gram negative and positive bacteria, our data suggest that phage may be an important vector for horizontal genetic exchange between diverse bacterial species. Thus, the common occurrence of Ig-like domains in phage genomes may partially account for their widespread occurrence in bacterial genomes.

It is intriguing that many of the phage Ig-like domains are added to the C termini of proteins through a ribosomal frameshifting mechanism. The common occurrence of this phenomenon is unlikely to be due solely to close relatedness among the frameshifted Ig-like domains for the following reasons: (a) all three classes of Ig-like domains display frameshifting, (b) within the Big 2 family of domains, where most frameshifting occurs, distinct -1 and $+1$ frameshifting mechanisms are observed, and (c) frameshifting occurs in unrelated proteins. Frameshifting provides a means to attach a unique C terminus to a fixed proportion of a given protein. For example, in the case of the phage T3 MHP, approximately 10% of molecules possess the Ig-like domain that is added through frameshifting.[16] Frameshifting may be selected for in some cases where placing a particular domain on every molecule of a given protein would be detrimental to its function, whereas placing the domain on a small proportion of molecules would be advantageous. In addition, attaching an inessential but marginally advantageous domain to a protein through frameshifting might provide a means to prevent mutations in the domain from destroying the function of the whole protein. In the case of the phage lambda MTP, a point mutation in its C-terminal Ig-like domain can cause a loss of phage viability even though this domain can be deleted with little effect.[13] The C-terminal domain mutation likely causes it to misfold, and leads to precipitation of the whole MTP. If this domain were added through a frameshifting mechanism, a point mutation within it would be much less likely to cause total inactivation of the protein. The common occurrence of frameshifting in phage genomes emphasizes the necessity of searching these genomes for domains of interest at the level of translated DNA, not annotated ORFs. Five of the frameshifted Ig-like domains that we identified were not picked up by standard PSI-BLAST or blastp methods because they are not annotated as proteins at all. Another five of the frameshifted Ig-like domains were annotated as separate ORFs even though they are unlikely to be such.

While we observed that unrelated phage proteins can possess similar Ig-like domains, there are also many examples of closely related pairs of phage proteins where one possesses an Ig-like domain and the other does not as shown in Figure 2(b) for the PY54 and phiK02 phages. A similar phenomenon is seen in the phage T7 and T3 MHPs, which are 80% identical in their N-terminal domains, yet the T3 MHP possesses an Ig-like domain at its C terminus, which is added by frameshifting, and T7 does not. Strangely, the T7 head protein maintains a frameshift at its C terminus, but this frameshift adds a different smaller domain. This is not simply a case of mutational decay of the Ig domain (i.e. there is no evidence of sequence similarity at the DNA level). The spacing between the end of the T3 Ig-like ORF and the T7 non-Ig-like ORF to the next protein is conserved between the two phages. The sporadic occurrence of Ig-like domains in proteins with very similar sequences and shared functions implies that the Ig-like domains are not generally essential for the function of these proteins. Consistent with this supposition, is the observation that the Ig-like domains of the phage lambda MTP and the phage T3 MHP are not essential for phage growth under standard laboratory conditions.[13,16] On the other hand, the Ig-like domain of the phage A2 MHP has been shown to be essential.[22] Similar to the pattern of Ig-like domain occurrence in phage proteins, FN3

domains have been observed to occur sporadically in closely related bacterial glycohydrolases.[10] In spite of the non-essential nature of the Ig-like domains in some phages as tested in the laboratory, the promiscuous spread of these domains among diverse phages suggests that they must confer some selective advantage to phages that possess them.

Although there has been no specific function yet demonstrated for a phage Ig-like domain, several lines of evidence suggest that these domains may aid in the attachment of phage to the cell surface. A significant feature of these domains is that all of the reliably annotated proteins in which they appear are part of the mature phage particle. Since the vast majority of Ig-like domains in all types of cells from humans to bacteria are located in proteins on the cell surface,[7] we expect that the phage Ig-like domains will be located in exposed positions on phage surfaces. Supporting this supposition are the observations that the Ig-like domain of the phage lambda MTP has been shown by electron microscopy to be located on the phage outside of the tail,[13] and that the phage A2 MHP form containing the Ig-like domain is highly immunogenic.[22] Cryoelectron microscopy studies have also shown that the T4 HOC protein is highly exposed on the phage head surface,[30] and that the Ig-like domain of the Phi29 MHP is in an accessible position on the phage surface.[12] Most of the bacterial Ig-like domains that are closest in sequence to the phage Ig-like domains are found in bacterial surface glycohydrolases, such as chitinases and cellulases. While the functions of the Ig-like domains in these enzymes are not well characterized, in one case the FN3 domain of a cellobiohydrolase has been shown to bind cellulose and promote its hydrolysis.[31] There are also other examples of both eukaryotic and prokaryotic Ig-like domains that bind carbohydrates.[32,33] We hypothesize that the phage Ig-like domains may aid in the adhesion of viral particles to bacterial cell surfaces by weakly binding to carbohydrate cell wall components such as peptidoglycan or lipopolysaccharide. A weak, and possibly non-specific interaction between the Ig-like domains and the bacterial cell wall could maintain the phage in close proximity to the cell in a gliding or bouncing manner until a specific receptor molecule is encountered. This mechanism might be similar to the manner in which relatively weak protein–carbohydrate interactions involving Selectin molecules induce leukocytes to "roll" on blood vessel walls as they approach a site of inflammation.[34,35] A relatively weak, non-specific binding function for the Ig-like domains would be consistent with their being located in a variety of phage surface proteins; it would not matter which protein the domains were in as long as they were on the phage particle surface. It is interesting to note that Ig-like domains on the surface of eukaryotic viruses have been found to participate in the process of cell attachment.[36,37]

Finally, this work adds an important note of caution for the annotation of both bacterial and phage genomes. We have demonstrated that the existence of promiscuous domains partaking of rampant horizontal transfer, such as Ig-like domains, can lead to the bioinformatic connection and subsequent misannotation of many genes that do not actually share common functions. Since misannotations tend to rapidly propagate,[38] the detection and cataloguing of all the promiscuous domains in bacteria and phage may be an important future bioinformatic goal. Possessing a clear picture of the functioning of promiscuous domains could both prevent misannotation and aid in the correct assignment of protein functions.

Materials and Methods

We compiled a nucleotide database of completed bacteriophage genomes from the NCBI genome database. This database included dsDNA, ssDNA, RNA, and unclassified bacteriophages. To seed our search for phage Ig-like domains, we searched our phage genome database with Hidden Markov Models[39] from a wide variety of PFAM sequence alignment database entries[15] corresponding to domains classified as immunoglobulin-like beta-sandwich folds in the SCOP database[40] (SCOP: 48725). The database was searched at the DNA level using the Wise2 suite.[41] Scores above 25 bits were considered significant, and scores from 18–25 bits were examined manually to determine if they were Ig-like. Domains identified in these searches were clustered using TRIBE-MCL[42] and aligned at the protein level using CLUSTALW.[43] These alignments were used to create further HMMs, which were used in subsequent searches. In this way, more distant relatives of the original hits were identified. This procedure was continued until no new hits were returned. To verify the efficacy of this approach, each Ig-like domain found in our searches was used to initiate a PSI-BLAST[14] search of a database of all returned bacteriophage Ig-like domains for five iterations or until convergence. The position-specific scoring matrix generated was used to initiate a PSI-TBlastN search of the bacteriophage DNA database. The results from both approaches were consistent. It should be noted that more examples of phage-related proteins containing Ig-like domains can be found in bacterial genomes within prophage sequences. We chose not to include these proteins in the current study due to the difficulty of classifying and interpreting prophage data.[44]

The final collection of phage Ig-like domain sequences were clustered and aligned. In addition, protein sequences encoded by genes surrounding the Ig-like domain encoding regions in each genome were clustered and aligned. For each sequence, we noted any annotations and references to literature. If the hypothesized function of these ORFs was unclear, we used a TBLASTN search of 2000 bases each upstream and downstream to attempt to assign a broad functional role as inferred from annotations of the surrounding ORFs (e.g. head morphogenesis if surrounding ORFs showed homology to head morphogenesis proteins).

Acknowledgements

This work was supported by an operating grant to A.R.D. from the Canadian Institutes of Health

research. We thank Paolo Tavares and Juan Suarez for communicating unpublished data.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2006.03.043

References

1. Hendrix, R. W. (1999). Evolution: the long evolutionary reach of viruses. *Curr. Biol.* **9**, R914–R917.
2. Hendrix, R. W. (2003). Bacteriophage genomics. *Curr. Opin. Microbiol.* **6**, 506–511.
3. Renaissance phage. *Nature Rev. Microbiol.* **2**, 922.
4. Rohwer, F. & Edwards, R. (2002). The phage proteomic tree: a genome-based taxonomy for phage. *J. Bacteriol.* **184**, 4529–4535.
5. Halaby, D. M., Poupon, A. & Mornon, J. (1999). The immunoglobulin fold family: sequence analysis and 3D structure comparisons. *Protein Eng.* **12**, 563–571.
6. Bork, P., Holm, L. & Sander, C. (1994). The immunoglobulin fold. Structural classification, sequence patterns and common core. *J. Mol. Biol.* **242**, 309–320.
7. Halaby, D. M. & Mornon, J. P. (1998). The immunoglobulin superfamily: an insight on its tissular, species, and functional diversity. *J. Mol. Evol.* **46**, 389–400.
8. Bateman, A., Eddy, S. R. & Chothia, C. (1996). Members of the immunoglobulin superfamily in bacteria. *Protein Sci.* **5**, 1939–1941.
9. Luo, Y., Frey, E. A., Pfuetzner, R. A., Creagh, A. L., Knoechel, D. G., Haynes, C. A. *et al*. (2000). Crystal structure of enteropathogenic *Escherichia coli* intimin–receptor complex. *Nature*, **405**, 1073–1077.
10. Little, E., Bork, P. & Doolittle, R. F. (1994). Tracing the spread of fibronectin type III domains in bacterial glycohydrolases. *J. Mol. Evol.* **39**, 631–643.
11. Bateman, A., Eddy, S. R. & Mesyanzhinov, V. V. (1997). A member of the immunoglobulin superfamily in bacteriophage T4. *Virus Genes*, **14**, 163–165.
12. Morais, M. C., Choi, K. H., Koti, J. S., Chipman, P. R., Anderson, D. L. & Rossmann, M. G. (2005). Conservation of the capsid structure in tailed dsDNA bacteriophages: the pseudoatomic structure of phi29. *Mol. Cell*, **18**, 149–159.
13. Katsura, I. (1981). Structure and function of the major tail protein of bacteriophage lambda. Mutants having small major tail protein molecules in their virion. *J. Mol. Biol.* **146**, 493–512.
14. Altschul, S. F. & Koonin, E. V. (1998). Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases. *Trends Biochem. Sci.* **23**, 444–447.
15. Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Finn, R. D. & Sonnhammer, E. L. (1999). Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucl. Acids Res.* **27**, 260–262.
16. Condreay, J. P., Wright, S. E. & Molineux, I. J. (1989). Nucleotide sequence and complementation studies of the gene 10 region of bacteriophage T3. *J. Mol. Biol.* **207**, 555–561.
17. Kelly, G., Prasannan, S., Daniell, S., Fleming, K., Frankel, G., Dougan, G. *et al*. (1999). Structure of the cell-adhesion fragment of intimin from enteropathogenic *Escherichia coli*. *Nature Struct. Biol.* **6**, 313–318.
18. Farabaugh, P. J. (1996). Programmed translational frameshifting. *Annu. Rev. Genet.* **30**, 507–528.
19. Gesteland, R. F. & Atkins, J. F. (1996). Recoding: dynamic reprogramming of translation. *Annu. Rev. Biochem.* **65**, 741–768.
20. Giedroc, D. P., Theimer, C. A. & Nixon, P. L. (2000). Structure, stability and function of RNA pseudoknots involved in stimulating ribosomal frameshifting. *J. Mol. Biol.* **298**, 167–185.
21. Harger, J. W., Meskauskas, A. & Dinman, J. D. (2002). An "integrated model" of programmed ribosomal frameshifting. *Trends Biochem. Sci.* **27**, 448–454.
22. Garcia, P., Rodriguez, I. & Suarez, J. E. (2004). A −1 ribosomal frameshift in the transcript that encodes the major head protein of bacteriophage A2 mediates biosynthesis of a second essential component of the capsid. *J. Bacteriol.* **186**, 1714–1719.
23. Pajunen, M., Kiljunen, S. & Skurnik, M. (2000). Bacteriophage phiYeO3-12, specific for *Yersinia enterocolitica* serotype O:3, is related to coliphages T3 and T7. *J. Bacteriol.* **182**, 5114–5120.
24. Rodriguez, I., Garcia, P. & Suarez, J. E. (2005). A second case of −1 ribosomal frameshifting affecting a major virion protein of the Lactobacillus bacteriophage A2. *J. Bacteriol.* **187**, 8201–8204.
25. Casjens, S. R., Gilcrease, E. B., Huang, W. M., Bunny, K. L., Pedulla, M. L., Ford, M. E. *et al*. (2004). The pKO2 linear plasmid prophage of *Klebsiella oxytoca*. *J. Bacteriol.* **186**, 1818–1832.
26. Pedulla, M. L., Ford, M. E., Houtz, J. M., Karthikeyan, T., Wadsworth, C., Lewis, J. A. *et al*. (2003). Origins of highly mosaic mycobacteriophage genomes. *Cell*, **113**, 171–182.
27. Xu, J., Hendrix, R. W. & Duda, R. L. (2004). Conserved translational frameshift in dsDNA bacteriophage tail assembly genes. *Mol. Cell*, **16**, 11–21.
28. Monzingo, A. F., Marcotte, E. M., Hart, P. J. & Robertus, J. D. (1996). Chitinases, chitosanases, and lysozymes can be divided into procaryotic and eucaryotic families sharing a conserved core. *Nature Struct. Biol.* **3**, 133–140.
29. Ochman, H., Lawrence, J. G. & Groisman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature*, **405**, 299–304.
30. Fokine, A., Chipman, P. R., Leiman, P. G., Mesyanzhinov, V. V., Rao, V. B. & Rossmann, M. G. (2004). Molecular architecture of the prolate head of bacteriophage T4. *Proc. Natl Acad. Sci. USA*, **101**, 6003–6008.
31. Kataeva, I. A., Seidel, R. D., 3rd, Shah, A., West, L. T., Li, X. L. & Ljungdahl, L. G. (2002). The fibronectin type 3-like repeat from the Clostridium thermocellum cellobiohydrolase CbhA promotes hydrolysis of cellulose by modifying its surface. *Appl. Environ. Microbiol.* **68**, 4292–4300.
32. Buts, L., Bouckaert, J., De Genst, E., Loris, R., Oscarson, S., Lahmann, M. *et al*. (2003). The fimbrial adhesin F17-G of enterotoxigenic *Escherichia coli* has an immunoglobulin-like lectin domain that binds *N*-acetylglucosamine. *Mol. Microbiol.* **49**, 705–715.
33. Zaccai, N. R., Maenaka, K., Maenaka, T., Crocker, P. R., Brossmer, R., Kelm, S. & Jones, E. Y. (2003). Structure-guided design of sialic acid-based Siglec inhibitors

and crystallographic analysis in complex with sialoadhesin. *Structure*, **11**, 557–567.
34. Kansas, G. S. (1996). Selectins and their ligands: current concepts and controversies. *Blood*, **88**, 3259–3287.
35. Wild, M. K., Huang, M. C., Schulze-Horsel, U., van der Merwe, P. A. & Vestweber, D. (2001). Affinity, kinetics, and thermodynamics of E-selectin binding to E-selectin ligand-1. *J. Biol. Chem.* **276**, 31602–31612.
36. Jin, D. Y., Li, Z. L., Jin, Q., Hao, Y. W. & Hou, Y. D. (1989). Vaccinia virus hemagglutinin. A novel member of the immunoglobulin superfamily. *J. Expt. Med.* **170**, 571–576.
37. Rey, F. A., Heinz, F. X., Mandl, C., Kunz, C. & Harrison, S. C. (1995). The envelope glycoprotein from tick-borne encephalitis virus at 2 Å resolution. *Nature*, **375**, 291–298.
38. Gilks, W. R., Audit, B., De Angelis, D., Tsoka, S. & Ouzounis, C. A. (2002). Modeling the percolation of annotation errors in a database of protein sequences. *Bioinformatics*, **18**, 1641–1649.
39. Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
40. Hubbard, T., Murzin, A. G., Brenner, S. E. & Chothia, C. (1997). SCOP: a structural classification of proteins database. *Nucl. Acids Res.* **25**, 236–239.
41. Birney, E., Clamp, M. & Durbin, R. (2004). GeneWise and Genomewise. *Genome Res.* **14**, 988–995.
42. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucl. Acids Res.* **30**, 1575–1584.
43. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22**, 4673–4680.
44. Casjens, S. (2003). Prophages and bacterial genomics: what have we learned so far? *Mol. Microbiol.* **49**, 277–300.

Edited by I. Wilson

(Received 27 December 2005; received in revised form 15 March 2006; accepted 17 March 2006)
Available online 6 April 2006