Research

Extensive domain shuffling in transcription regulators of DNA viruses and implications for the origin of fungal APSES transcription factors

Lakshminarayan M Iyer, Eugene V Koonin and L Aravind

Address: National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA.

Correspondence: L Aravind. E-mail: aravind@ncbi.nlm.nih.gov

Abstract

Background: Viral DNA-binding proteins have served as good models to study the biochemistry of transcription regulation and chromatin dynamics. Computational analysis of viral DNA-binding regulatory proteins and identification of their previously undetected homologs encoded by cellular genomes might lead to a better understanding of their function and evolution in both viral and cellular systems.

Results: The phyletic range and the conserved DNA-binding domains of the viral regulatory proteins of the poxvirus D6R/N1R and baculoviral Bro protein families have not been previously defined. Using computational analysis, we show that the amino-terminal module of the D6R/N1R proteins defines a novel, conserved DNA-binding domain (the KilA-N domain) that is found in a wide range of proteins of large bacterial and eukaryotic DNA viruses. The KilA-N domain is suggested to be homologous to the fungal DNA-binding APSES domain. We provide evidence for the KilA-N and APSES domains sharing a common fold with the nucleic acid-binding modules of the LAGLIDADG nucleases and the amino-terminal domains of the tRNA endonuclease. The amino-terminal module of the Bro proteins is another, distinct DNA-binding domain (the Bro-N domain) that is present in proteins whose domain architectures parallel those of the KilA-N domain-containing proteins. A detailed analysis of the KilA-N and Bro-N domains and the associated domains points to extensive domain shuffling and lineage-specific gene family expansion within DNA virus genomes.

Conclusions: We define a large class of novel viral DNA-binding proteins and their cellular homologs and identify their domain architectures. On the basis of phyletic pattern analysis we present evidence for a probable viral origin of the fungus-specific cell-cycle regulatory transcription factors containing the APSES DNA-binding domain. We also demonstrate the extensive role of lineage-specific gene expansion and domain shuffling, within a limited set of approximately 24 domains, in the generation of the diversity of virus-specific regulatory proteins.

Background

Large DNA viruses of bacteria and eukaryotes have complex life cycles with several distinct phases that involve diverse virus-host interactions. An array of regulatory systems mediate activation or repression of expression of specific batteries of viral genes that are required at different phases

of the life cycle. Other sets of regulatory genes directly interact with components of the host cell and modulate its response to the virus [1-3]. Studies on these viral regulatory systems have revealed a pivotal role of transcription and chromatin organization in the control of gene expression and have contributed to the basic understanding of these processes in various model systems [1,3]. Several viral DNA-binding regulators have conserved domains that are shared with cellular transcription factors. The classic helix-turn-helix (HTH)-domain proteins, which govern the switch between the lysogenic and lytic pathways of temperate bacteriophages, and repressors containing the MetJ/Arc domain are well-known examples of such regulatory DNA-binding proteins in prokaryotic virus-host systems [4-8]. Large eukaryotic DNA viruses, such as poxviruses, phycodnaviruses, phaeoviruses, asfarviruses, iridoviruses (all of which form the recently identified monophyletic clade of nucleo-cytoplasmic large DNA viruses (NCLDVs) [9]), baculoviruses and herpesviruses, also encode a number of transcription factors. Some of these share domains with regulatory proteins of their eukaryotic hosts, for example the FCS zinc finger and the TFIIS-like zinc ribbon [9].

Much less is known of the domain architecture and evolutionary history of those viral DNA-binding regulatory proteins whose cellular homologs have not (yet) been identified. For example, the baculovirus repeat ORFs (Bro) proteins are a family of DNA-binding proteins that may regulate both viral and host transcription or chromatin structure, but no homologs, cellular or otherwise, have been identified for these proteins [10,11]. The regulatory proteins of the poxviral variola D6R/Shope fibroma virus N1R family have been shown to bind DNA and probably regulate apoptosis of the host cell [12]. All these proteins contain a conserved amino-terminal domain, for which no homologs outside this family have as yet been reported. These proteins additionally contain a carboxy-terminal RING finger domain [13], and some of them also contain a single-stranded nucleic-acid-binding CCCH domain between the amino-terminal domain and the RING finger.

During the comparative analysis of DNA viruses of the NCLDV class [9], we observed that the amino-terminal domain of the D6R/N1R family and baculoviral Bro-family proteins have a wide range of homologs in both DNA viruses and cellular genomes. Typically, these domains were identified in multidomain proteins that additionally contained several previously undetected, evolutionarily mobile domains occurring in different contexts. These observations suggested the existence of a common, previously uncharacterized set of regulatory proteins (domains) encoded by numerous eukaryotic and bacterial DNA viruses, as well as some of their host genomes.

To gain a better understanding of the functions and evolution of these regulatory proteins, we initiated a detailed analysis of their sequences using position-specific-score matrix searches, sequence-structure threading, secondary-structure prediction and structure comparisons. Here we describe the functional predictions and evolutionary history of these viral regulatory proteins and their cellular homologs that were detected as a result of these analyses. The comparison of the domain architectures of these proteins points to a major general role for combinatorial shuffling of a small set of domains in the evolution in transcriptional regulators of eukaryotic and bacterial DNA viruses.

Results and discussion
Identification of the KilA-N domain in diverse viral proteins and its potential relationship to APSES and LAGLIDADG domains

To determine the evolutionary affinities of the D6R/N1R amino-terminal regions, we initiated a PSI-BLAST search of the non-redundant (NR) protein database (National Center for Biotechnology Information), which was seeded with the sequence of the corresponding region from the variola virus D6R protein. This search not only recovered homologous proteins from almost all poxviral proteomes, but also previously undetected homologs from the Chilo iridescent virus, a variety of γ-proteobacterial temperate phages (such as KilA of the phage BPP1), and chromosome-encoded proteins from *Neisseria meningitidis*, *Xylella fastidiosa*, *Salmonella paratyphi* and *Clostridium difficile* (Table 1, Figure 1a). The presence of this conserved region in the amino terminus of the BPP1 KilA protein, which is involved in killing the host cells [14], with another distinct conserved carboxy-terminal region (Figure 2), suggests that this region is a mobile domain that is present in different proteins in independent contexts. Accordingly, this domain was named the KilA-N (terminal) domain (Table 2). In all proteins shown to contain the KilA-N domain, it occurs at the extreme amino terminus accompanied by a wide range of distinct carboxy-terminal domains other than the RING finger or CCCH domains that are seen in the poxviruses (Figure 2 and see below).

PSI-BLAST searches seeded with the KilA-N domains from a diverse set of viral and bacterial proteins also consistently recovered fungal transcription factors that are involved in cell-cycle-specific gene expression and filamentation, with E-values of borderline statistical significance (0.05-0.15). These hits precisely mapped to the APSES DNA-binding domain, which is shared by a range of fungus-specific transcription factors, such as MBP1p, SWI4p, PHD1p and SOK2p from *Saccharomyces cerevisiae* and Stunted A from *Aspergillus nidulans* [15]. For example, a search initiated with the KilA-N domain of *Pseudomonas* phage D3 Orf11 (gi:9635595) recovered the APSES domain of *Saccharomyces* MBP1p with an E-value of 0.05 in iteration 3. Reciprocal searches with the profiles of APSES domains, similarly recovered the KilA-N domain with borderline E-values. As an example, a profile seeded with the MBP1p

Table 1

Distribution of KilA-N, Bro-N and associated conserved domains in viral and cellular proteomes

Organism	Host, if phage or virus	Domain architectures
NCLDVs, baculoviruses	Eukaryotes	KilA-N solo (AMV, FPV)
		KilA-N + Bro-C (AMV, CIV, FPV)
		KilA-N + T5ORF172 (CIV)
		KilA-N + RING finger (chordopoxviruses); ectromelia p28 [12]
		KilA-N + CCCH + RING finger (FPV)
		Bro-N solo (MSV, AMV, HzNV, LdNV, XnGV)
		Bro-N + Bro-N (CnBV, CpGV, PxGV)
		Bro-N + T5ORF172 (CIV, MSV, ESV, baculoviruses)
		Bro-N + Bro-C (AMV, baculoviruses); BroA, BroD of BmNV [10]
		Bro-N + Bro-N + Bro-C (HaNV)
		Bro-N + VSR nuclease (MSV, CIV)
		Bro-N + Bro-N (CnBV, CpGV)
		Bro-N + Bro-C (AMV, baculoviruses)
		MSV199 solo (MSV)
		MSV199 + T5ORF172 (AMV, MSV, CIV)
		MSV199 + UVRC (CIV)
		MSV199 + Bro-C (CIV)
		Bro-C solo (CIV, DpAV4, AcNV, BmNV, EpNV, LdNV, LsNV, MbNV, OpNV)
		T5ORF172 solo (CIV, LdNV)
Proteobacteria	-	KilA-N solo (*N. meningitidis, X. fastidiosa*)
		Bro-N solo (*N. meningitidis, X. fastidiosa, Pseudomonas aeruginosa*)
		Bro-N (2/3 copies) + XF1559 (*X. fastidiosa*)
		XF0704 + Bro-N (*X. fastidiosa*)
		Bro-N + P22AR-C (*Escherichia coli*)
		KilA-C solo (*E. coli*)
		φ31-ORF238N + P22AR-C (*Haemophilus influenzae*)
		φ31-ORF238N + wHTH (*N. meningitidis*)
		XF0704 solo (*X. fastidiosa*)
		XF1559 solo (*X. fastidiosa*)
		RHA solo (*Pasteurella multocida*)
		RHA + D3ORF11-C (*H. influenzae*)
		RHA + ASH (*Shigella flexneri*)
		D3ORF11-C + P22AR-N (*E. coli*)
		bIL285ORF6-N solo (*E. coli*)
		ASH solo (*E. coli, S. flexneri*)
		T5ORF172 solo (*N. meningitidis*)
Low-GC Gram-positive bacteria/ Firmicutes	-	Bro-N solo (*Streptococcus pyogenes*)
		Bro-N + HTH (*S. pyogenes*)
		Bro-N + KilA-C (*S. pyogenes, Staphylococcus aureus*)
		RHA solo (*Streptococcus pneumoniae*)
		φ31-ORF238N + KilA-C (*S. pyogenes, S. aureus*)
		φ31-ORF238N + φSLT-orf81a (*Clostridium acetobutylicum*)
		T5ORF172 solo (*Bacillus subtilis*)
		KilA-C solo (*Lactococcus lactis*)
High-GC Gram-positive bacteria	-	Bro-N solo (*Streptomyces coelicolor*)
Phages BP7888, BP933W, BPHK97, BPN15, BPHK620, BPHK022, BPφ-R73 BPP22, BPH-19B, BPP1, BPP4, BPP27, BPφ80, BPT5, BPVT2-Sa	*Escherichia coli*	Bro-N solo (BPN15)
		KilA-N + KilA-C; phage P1- KilA [14]
		KilA-N + D3ORF11-C (BPHK620)
		KilA-N + p63C (BP933W, BPHK97)
		RHA solo (BPφ80); BPP22 ORF 201 [47]
		RHA + KilA-C (BPHK022, BP933W, BPP1, BPH-19B, BPHK97, BPHK620, BPVT2-Sa); Reduced on IHF⁻ (ROI) [48]
		KilA-C solo; bacteriophage P1 antirepressor Ant1/Ant2 [49]
		φ31-ORF238N + P22AR-C (BP933W, BPVT2-Sa)
		D3ORF11-C + P22AR-N (BP933W, BP7888)
		ASH solo (BPP4, BPφ-R73, BPN15)
		bIL285ORF6-N + P22AR-C (BPVT2-Sa)
		T5ORF172 solo (BPT5, BPP27)
Phage P22	*Salmonella typhimurium*	P22AR-N + Bro-N + P22AR-C; BPP22 antirepressor [33,34]
		RHA + D3ORF11-C [46]

Table 1 *(continued)*

Organism	Host, if phage or virus	Domain architectures
Phage APSE-1	*Buchnera aphidicola*	RHA (solo) Bro-N (solo)
Phage D3	*Pseudomonas aeruginosa*	KilA-N + D3ORF11-C
BPφPV83, BPφETA, BPφSLT	*Staphylococcus aureus*	Bro-N + KilA-C (BPφPV83) φ31-ORF238N + KilA-C (BPφETA) φSLT-ORF81a solo (BPφSLT) RHA solo (BPφSLT)
BPbIL285, BPbIL286, BPbIL311, BPbIL309, BPpi3, BPφ31.1, BPTP901-1, BPRIT, BPBK5-T, BPLL-H, BPTuc2009	*Lactococcus lactis*	Bro-N + KilA-C (BPRIT, BPBK5-T, BPLL-H, BPbIL309) KilA-C solo (BPRIT) RHA solo (BpbIL310, BPbIL311) φ31-ORF238N + bIL285ORF6-C (BPpi3, BPTuc2009, BPbIL286, BPφ31.1) bIL285ORF6-N + bIL285ORF6-C (BPbIL285, BPTP901-1)
LcBPA2	*Lactobacillus casei*	Bro-N solo
BPTP-J34, BPSfi21	*Streptococcus thermophilus*	φ31-ORF238N + KilA-C
BPA118	*Listeria monocytogenes*	φ31-ORF238N + KilA-C
BPSpβc2	*Bacillus subtilis*	RHA + KilA-C
BPMx8 (Myxococcal phage)	*Myxococcus xanthus*	Bro-N + p63C

APSES domain and including all APSES domains in the NR database detects the *N. meningitidis* protein NMA1544 (gi:11290039) with an E-value of 0.1 in iteration 3. Given the availability of the three-dimensional structure of the APSES domain of the MBP1p protein [16,17], we investigated this potential relationship using the KilA-N domain sequences for sequence-structure threading of the PDB database with the 3DPSSM, PSIPRED and combined-fold prediction algorithms. Both 3DPSSM and PSIPRED gave hits (E-value approximately 0.05 for 3DPSSM and probability of matching approximately 0.8 for PSIPRED) that implied a 90% certainty of the KilA-N domain adopting the same fold as the APSES domain (PDB 1bm8/1bm1). Additionally, the 3DPSSM threading also suggested that the LAGLIDADG endonuclease domain [18] (E-value 0.2; approximately 80% certainty) shared a common fold with the KilA-N proteins. Threading with the combined-fold prediction algorithm also gave the MBP1p APSES domain as the best hit with a very high Z-score (Z = 60), suggesting that the KilA-N domain was highly likely to adopt the same fold as the APSES domain.

Secondary-structure prediction using the Jpred method [19], with a multiple alignment of the KilA-N domain used as the input, pointed to an $\alpha+\beta$ fold with four conserved strands and at least two conserved helices. This predicted secondary structure of the KilA-N domain, with a head-to-tail dyad of a 2-β-strand-α-helix unit (Figure 1a), is identical to the secondary structure seen in the conserved core of the APSES domains. Structural comparisons using the DALI program [20], indicated that the core fold shared by the KilA-N and APSES domains is more distantly related to the LAGLIDADG site-specific DNA endonucleases and the amino-terminal domain of the tRNA splicing endonucleases (TEN

domain) (Figure 1b) [21]. Three-dimensional superpositions of the three of these domains for which structures are available aligned the C-α atoms with a root-mean-square deviation of 3.2 Å or less for approximately 60 residues. A search with the APSES domain (PDB 1bm8) using the VAST program [22] detected significant structural alignments with both the LAGLIDADG and the TEN domains (*p* approximately 10^{-4}). The structural similarity between the TEN domain and LAGLIDADG nuclease domain has been previously noted [23], but the connection of both of these with the APSES domains has not been reported before to our knowledge. We noticed, however, that, although the TEN domain is related to the LAGLIDADG endonucleases, the active-site residues are not preserved in the former, and the TEN domain probably functions as a RNA-binding domain rather than a nuclease. These observations suggest that the APSES, LAGLIDADG and TEN domains, together with the KilA-N domain, define a novel nucleic-acid-binding fold with a conserved $(\beta_2\alpha)_2$ core (Figure 1b). Previously, it had been proposed that the APSES domains were related either to the winged HTH or to the basic helix-loop-helix (bHLH) domains [16,17,24]. These connections are not, however, recovered in any sensitive sequence-profile or structural similarity searches. A direct comparison of the structures also showed that the only domains that share a similar topology and conformation with the APSES domain are the LAGLIDADG and TEN domains. This implies that the previously proposed relationships for the APSES domain are unlikely to represent the true evolutionary connections.

A comparison of the multiple alignment of the KilA-N and APSES domains showed that several characteristic hydrophobic/aromatic residues were conserved between the
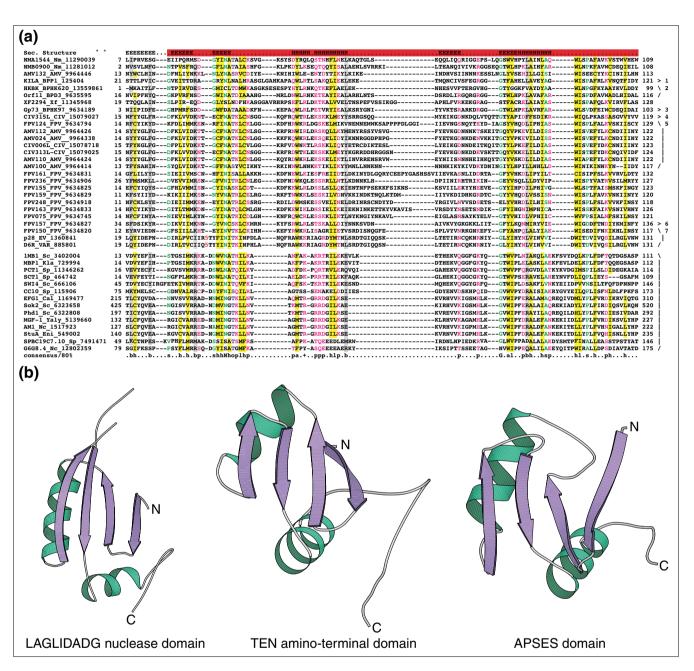
(a)

```
Sec. Structure  ° °    EEEEEEEE....EEEEEE.....EEEEE...............HHHHH.HHHHHHHHH.......................EEEEEE.........EEEEEHHHHHHHHH................
NMA1544_Nm_11290039    7 LIPRVESG---EIIPQRMSD---GYINATALCKSVG----KSYSDYRQLQSTNHFLNELKAQTGLS---------------EQQLIQQRIGGEPS-LQGSWVHPYLAINLAQ------WLSPAFAVKVSTWVHEW 109
NMB0900_Nm_11281012    2 NVSVLNFG---NTPVSFRQD---GFLNATAIASHFG----KLPKDYLKSEQTQQYISALAENLSVRRKI-----------LTEANQIVIVKRGGSEQGTWLHPKLAIHFAR-------WLNPKFAVWCDEQIEIL 108
AMV132_AMV_9964446    13 NYWCLHIN---DFNLIYNKKL--NLYNASRVCDIYE----KNIHIWLE-ENYDYTIKYLKIKE-------------------INDHVSIINNNKESSLNGLYVSEHILLGISI------WISEECYYKCINIILHN 113
KILA_BPP1_125404      21 STTLPVIC---GVEITTDRA---GRYNLNALHRASGLGAHKAPAQWLRTLSAKQLIEELEKE--------------------TMQNCIVSFEGRGG---GTFAHELLAVEYAG-------WISPAFRLKVNQTFIDY 121 > 1
HKBK_BPHK620_13559861  1 -MKAITLF---NTPIRVDES---GMICLTDMWKASGKSESESPYHYLRNKQTKEFLAELEK---------------------NHESVVFTERGVHG---GTYGGKFVAYDYAA------WLNPGFKYAAYKVLDDY  99 \ 2
Orf11_BPD3_9635595    16 NVIPFHYQ---GKPVRFNSD---GWINATDIAAAHG----MRLDNVLRNKETEAYIEALARHLNTS---------------DSRDLIRGQRGRGG---GTWLHPKLAVAFAR-------WISPDPAVWADLHIDAL 116 /
XF2294_Xf_11345968    19 TTQQLAIN---SLPIR-EQD---GLYSLNDFHKASGGAVRHRPSEFLRLDKTKALVVELTNSPEFVSSIKGG---------APHLFVRKEKGRAG---STFACRELAIAYAS-------WISPAFQLKVIRVFLAS 128
Gp73_BPHK97_9634189    3 NIIPIDFE---GHPMRFSDD---GWFDATAAKDKFN----KEPAQNLRLPETVRYIEALKSRYGNI------------------TYVKTSRARKDRGG---GTWLHPKLAVRFAR-------WLSVDFEINVDEQIDAI 103 > 3
CIV315L_CIV_15079027  15 NFYYGLFR---DFKLVVDKNT--ECFNATKICNSGG----KQFRQWTRLEKSKKLMEYSRRGSQQ---------------MYEIKGDNKDQLVTQTTGTYAPIDFFEDIKR------WIQLPKASSASGVVYVV 119 > 4
FPV124_FPV_9634794    14 RFCYIKYD---KFDLIMMKEN--RFINATKICKLGG----KDFHRWKRLDGSKELMIKVNEMNEMWKSAPPPPDLGGI----IIEVNGSNQYTEYD-IAGSYVHQDLIPHIAS-------WISPLFALKVSKIISCY 129 \ 5
AMV112_AMV_9964426    14 SYYYGLFG---DFKLVIDKTT--GCFNATKICNLSG----KRFRNWVIRLDRSKQLLKYMENYRSSYVSVG---------FYEVKGQNNNKTSKEITGQYVPKEVILDISS-------WISVEFYLKCNDIIINY 122
AMV024_AMV_9964338    14 SYYYGLFG---DFKLVIDKTN--GCFNATKICNLGG----KKFKQWKRLEKSQELIDYIKNNRGGDPHPG---------FYETKGDNKDENVKKITGCYVPKEVILDISS-------WISVEFYLKCNDIIINY 123
CIV006L_CIV_15078718  15 TFYKGLFG---DFPLIVDKKT--GCFNATKLCVLGG----KRFVDWNRTLRSKKLIQYYETRCDIKTESL----------LYEIKGDNNDEITKQITGTYLPKEFILDIAS-------WISVEFYDKCNNIIINY 123
CIV313L_CIV_15079025  15 NFYYGLFG---DFKLVVDKNT--ECFNATKICNSGG----KDFRKWTKLEKSKKLMEYSRRGSQQ---------------FYEVKGQNKDDEVSKTTGQYVKKELILDIAS-------WISTEFYDKCNNIVIDF 124
AMV110_AMV_9964424    14 SYYYGLFG---DFKLVIDKTN--GCFNATKICNLGG----KQYRDWKRLEKSKELIKTLINVVRRENSRVM---------EYNIISNNNHEIHKQYTGYYVSKDLILDIAS-------WIAPEFYLKCNDIIINY 122
AMV100_AMV_9964414    13 TFYSAHIN---SYQLVIDKKT--GCFNASYVCIKNY----RKINNNLNNKKTIKLIKYVIKDKNINGIYLHPILLNHLLD-----NNNKIKYKIVDKYDNINGIYLHPILLNHLLD------WINIKINNKYN--IIDY 117 /
FPV161_FPV_9634831    14 GFLILYYD---SIEIIVMSCN--HFNISALLAKKN----KDFNEWLKIESFREIIDTLDKINYDLGQRYCEEPYGASHSSVIIEVKASNLIDDRTA---GFYVHKDLIPYILT------CISIPFSLKVVRVLDTY 132
FPV236_FPV_9634906    26 YFMSMKLL---DVEVVIMRSN--GFVNITRLCNLEG----KDFHNWKQLESSRRILTEIENTNFPSEKKFSIKNS------KSVIILEKYYHEEVE--GYYIHPDILPHIVG-------WLSPTFAISMSKFINGY 123
FPV155_FPV_9634825    14 EFCYIQYS---GFHLVMMISN--CYINASKLCCT-----KDFKKWLRLDSSLSLLQEIENTNFPSEKKFSIKNS-------KSVIILEKYYHEEVE--GYYIHPDILPHIVG-------WLSPTFAISMSKFINGY 123
FPV159_FPV_9634829    11 KFSYIIYD---KIKIIIMKSN--NYVNATRLCELRG----RKFTNWKKLSESKIIVDYIKNNRGGDPHPG---------IIYVKDIDHKGRDTC--GYYIHPDLIPHVG-------WLSPLFAVKVNKIINYY 120
FPV248_FPV_9634918    11 NFCKLSYE---DIEIIMMKEN--EYINATKLCSRG----RDILDWMSKESSVELINELDRINRSCNDYYD---------YRGIVLNVVSDSETS--ELYVHRDLILHISH-------WISPLFSLKVVKFINSY 118
FPV163_FPV_9634833    14 HFCYIKYD---GITLIMMKCN--GYINATQLCMLGN----KDFKEWVRLDESSKLIKIETEKNINKETTKYVKAVIS----VRSDYYNSETSNDIK--GFYIHGNIMPHICA-------WISSKPAIKVSVIIVHNY 126
FPV075_FPV_9634745    14 NFCFINYA---NIEVIMLKYN--GYINATKICDLGN----KNFRQWCRLESSKKLIKTLNYKNGIYNKAVL---------EIGLASNSAYKYELV--GTYVHIDLVPHIIC-------WVFPSIALNFSKILNSY 121
FPV157_FPV_9634827    34 SFDSIKYR---DIKVIIMKNN--GYVNCSKLCKMRN----KYFSRWLRLSTSKALLDIYNNKSVDN---------------AIVKVYGKGKKLIIT--GTYLKQNMIRYVIE------WIGDDFTNDIYKMINFY 136 > 6
FPV150_FPV_9634820    12 EYRVIEDN---GFSIILLKHT--EYINVTKLCKIHN----KEFYRWKRLISAGRIIETVSRDISNQGFE-----------SPLVYVNRKGNKEFY--GFYAHPQLALYIAK------WISEDIFNKIKHLINSY 117 /
p28_EV_1360841        19 LQYIDEPN---DIRLPVCIIR5TYFINITKINPDLA----NQFRAWKRKIAGRDYMTNLSRDTGIQQSK---------LTETIRNCQKNRNIY--GLYIHYNLVINVVI-----DWITDVIVQSILRGLVNW 131 |
D6R_VAR_885801        19 LQYIDEPN---DIRLTVCIIQ5TYYINITKINPHLA----NQFRAWKRKIAGRDYMTNLSRDTGIQQSN---------LTETIRNCQKNRNIY--GLYIHYNLVINVVI-----DWITDVIVQSILRGLVNW 131 /

1MB1_Sc_3402004       13 VDVYEFIH---STGSIMKRKK-DDWVNATHILKA---------ANFAK-AKRTRILEKEVLK----------ETHEKVQGGFGKYQ---GTWVPLNTAKQLAEKFSVYDQLKPLFDFTQTDGSASP 111 \
MBP1_Kla_729994       14 VDVYEFIH---PTGSIMKRKA-DNWVNATHILKA---------AKFPK-AKRTRILEKEVIT----------DTHEKVQGGFGKYQ---GTWIPLELASKLAEKFEVLDELKPLFDFTQQEGSASP 112 |
PCT1_Sp_11346262      16 VEVYECFI---KGVSVMRRRR-DSWLNATQILKV---------ADFDK-PQRTRVLERQVQI----------GAHEKVQGGYGKYQ---GTWVPFQRGVDLATKYKVDGIMSPILSLDIDEGKAIA 114 |
SCT1_Sp_464742        14 VEVFEYTI---NGFPLMKRCH-DNWLNATQILKI---------AELDK-PRTRILEKFAQK----------GLHEKIQGGCGKYQ---GTWVPSERAVELAHEYNVFDLIQPLIEYS---GSAFM 109 |
SWI4_Sc_666106        45 TDVYECYIRGFETKIVMRRTK-DDWINITQVFKI---------AQFSK-TKRTKILEKESND----------MQHEKVQGGYGRFQ---GTWIPLDSAKFLVNKYEIIDPVVNSILTFQFDPNNPP 146 |
CC10_Sp_115906        75 MKYMELSC---GDNVALRRCP-DSYFNISQLLKL---------AGTSS-SENAKELDDIIES----------GDYENVQSKHPQID---GVWVPYDRAISIAKRYGVYEILQPLISFNLDLFPKFS 173 |
EFG1_Cal_1169477     215 TLCYQVDA---NNVSVVRRAD-NNMINGTKLLNV---------AQMTR-GRRDGILKSE----------KVRHVVKIGSMHLK---GVWIPFERALAMAQREQIVDMLYPLFVRDIKRVIQTG 310 |
Sok2_Sc_6323658      425 TLCYQVEA---NGISVVRRAD-NNMINGTKLLNV---------TKMTR-GRRDGILKAE----------KIRHVVKIGSMHLK---GVWIPFERALAIAQREKIADYLYPLFIRDIQSVLKQN 520 |
Phd1_Sc_6322808      197 TICYQVEA---NGISVVRRAD-NNMINGTKLLNV---------TKMTR-GRRDGILKSE----------KVREVVKIGSMHLK---GVWIPFERAYILAQREQIDHLYPLFVKDIESIVDAR 292 |
MGF-1_Yaly_5139660   132 TLCFQVEA---RGICVARRED-NNMINGTKLLNV---------AGMTR-GRRDGILKSE----------KLRHVVKAGAMHLK---GVWIPYDRALEFANKEKIIDLLFYPLFVRDISNVLYHP 227 |
AM1_Nc_1517923       127 SLCFQVEA---RGICVARRED-NAMINGTKLLNV---------AGMTR-GRRDGILKSE----------KVRHVVKIGPMHLK---GVWIPFERALDFANKEKITELLYPLFVHNIGALLYHP 222 |
StuA_Eni_549002      140 SLCYQVEA---KGVCVARRED-NGMINGTKLLNV---------AGMTR-GRRDGILKSE----------KVRNVVKIGPMHLK---GVWIPFDRALEFANKEKITDLLYPLFVRDISNLLYHP 235 /
SPBC19C7.10_Sp_7491471 49 LKCTNPES---KVPHFLMRMAK-DSSISATSMFRS----------AFPK-ATQEEEDLEMRW----------IRDNLNPIEDKRVA---GLWWPPADALALAKDYSMTPFINALLEASSTPSTYAT 146 |
G6G8.4_Nc_12802359    79 SGIFKSSP---PSYFLMRRSQ-DGYISATGMFKA----------TFPY-ASQEEEAERKY----------IKSIPTTSSEETAG---NVWIPPEQALILAEEYQITPWIRALLDPSDIAVTATD 175 /
consensus/80%            .bh...b....s..h.h.bp...shhNhoplhp...........pa.+..ppp.hlp.b...........................p....p....G.al..pbh..hsp......hl.s.h..ph...h...
```

(b)

LAGLIDADG nuclease domain   TEN amino-terminal domain   APSES domain

Figure 1
Sequence and structural analysis of the KilA-N domain. **(a)** Multiple alignment of KilA-N and APSES domains. Sequences are designated by their gene name, followed by species abbreviation and the Genbank index (gi) number. Species abbreviations are listed in Materials and methods. The coloring reflects the conservation profile at 80% consensus of amino acids. h, hydrophobic residues (L,I,Y,F,M,W,A,C,V in the single-letter amino-acid code); a, aromatic residues (F,H,Y,W); and l, aliphatic residues (L,I,A,V), all shaded yellow. c, charged residues (K,E,R,D,H); +, basic residues (K,R,H); -, acidic residues (D,E); and p, polar residues (S,T,E,D,R,K,H,N,Q), all colored magenta. s, small residues (S,A,C,G,D,N,P,V,T) colored green. b, big residues (L,I,F,M,W,Y,E,R,K,Q) shaded gray. Further grouping of sequences is based on the association of KilA-N with other domains as follows: 1, fused to KilA-C; 2, fused to D3ORF11-C; 3, fused to Mx8p63C; 4, fused to T5ORF172; 5, fused to Bro-C; 6, fused to a CCCH domain and a RING finger; 7, fused to a ring finger. **(b)** Structural comparison of the APSES, LAGLIDADG and tRNA splicing endonuclease (TEN) domains. The ribbon diagrams were drawn using Molscript.

two families (Figure 1a), suggesting that they have similar functional properties. The APSES domains principally bind specific DNA sequences associated with regulatory regions of numerous genes expressed in the G1 to S transition of the cell cycle in yeast [25,26]. The sequence relationship with the APSES domain, taken together with the evidence of DNA binding by the D6R/N1R protein [12], suggests that the KilA-N domain is a previously undetected DNA-binding
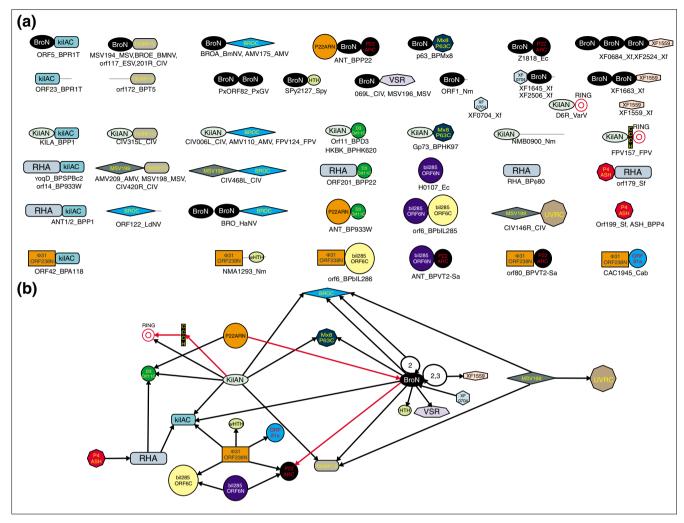
Figure 2
Domain architectures and architecture graphs of the KilA-N, Bro-N and other associated domains. **(a)** Domain architectures of the KilA-N, Bro-N and other associated domains. Gene names and species abbreviations are given below the architectures. Species abbreviations are listed in Materials and methods. **(b)** Domain architecture graph for the KilA-N, Bro-N and other associated domains. Each vertex represents a domain, and edges indicate domain combinations. Arrows point from the amino terminus to the carboxyl terminus of a multidomain protein. Architectures involving more than two colinear domains (see the three-domain proteins in (a)) are connected by red lines. Circular arrows indicate multiple copies of the same domain.

domain that is prevalent in the viral world. This prediction was also supported by the evidence from mutations to the conserved KilA-N domain of N1R that affected the localization of this protein to viral DNA-containing cytoplasmic virus maturation complexes (virus factories) [12].

The phyletic patterns of the KilA-N and APSES domains have interesting implications for the origin of the fungal transcription factors. Unlike several DNA-binding domains conserved throughout eukaryotes, such as the homeodomain, the bHLH, bZip, C2H2 zinc finger and the RFX domain, APSES domains are restricted to fungi [27,28]. Their closest relatives are the KilA-N domains, which are widespread in diverse DNA viruses and prophage derivatives from bacterial genomes. This suggests that APSES domains

probably emerged early in fungal evolution from a viral KilA-N-like precursor that was acquired by the host cell. An alternative scenario, that the viral KilA-N domains were acquired from the fungal APSES domain, is also imaginable. This appears less likely, however, because the APSES domains show limited sequence diversity in the fungi, compared with the much greater sequence diversity of the KilA-N domain in a wide range of DNA viruses that had probably already diverged by the time fungi emerged (Figure 1a). Furthermore, the viral provenance of the APSES-like domains is analogous to the recruitment of the transposon-derived BED-finger domain in cell-cycle-specific transcription factors such as BEAF-1 and DREF in the arthropods [29]. Both viruses and transposons are likely to derive selective advantages by evolving transcription factors that regulate

Table 2

Domains identified in this study

Domain name	Domain definition
Bro-N	Amino-terminal domain of baculovirus BRO proteins
KilA-N	Amino-terminal domain of phage P1 KilA
Bro-C	Carboxy-terminal domain of BmNV BroA and BroD
KilA-C	Carboxy-terminal domain of phage P1 KilA
T5ORF172	Domain present in ORF172 product of bacteriophage T5
P22AR-N	Amino-terminal domain of the phage P22 antirepressor
P22AR-C	Carboxy-terminal domain of the phage P22 antirepressor
Mx8P63C	Carboxy-terminal domain of the *Myxococcus* phage Mx8 p63
XF1559	Carboxy-terminal domain of the XF1559 protein of *Xylella fastidiosa*
XF0704	Carboxy-terminal domain of the XF0704 protein of *Xylella fastidiosa*
D3ORF11-C	Carboxy-terminal domain of ORF11 product of *Pseudomonas aeruginosa* bacteriophage D3
RHA	Domain present in the RHA protein of bacteriophage φ81
MSV199	Domain present in MSV199 of *Melanoplus sanguinipes* entomopoxvirus
bIL285ORF6-N	Amino-terminal domain of bacteriophage bIL2850 ORF6
bIL285ORF6-C	Carboxy-terminal domain of bacteriophage bIL2850 ORF6
P4ASH	Domain present in the ASH protein of bacteriophage P4
φ31ORF238-N	Amino-terminal domain of bacteriophage φ31.1 ORF238
ORF81a	Domain present in ORF81a of the temperate phage φSLT of *Staphylococcus aureus*

their genes in response to the host cell cycle. Hence, from the above observations, it is not unlikely that the host cells co-opt the transcription factors of their genomic parasites for their own cell-cycle-specific gene expression.

The potential higher-order structural relationship between the APSES and KilA-N domains with the LAGLIDADG site-specific DNA endonucleases and tRNA endonuclease amino-terminal domains has interesting implications for their evolutionary affinities and origin. A structural comparison of these proteins indicates that the nuclease active site of the LAGLIDADG domains is contained in a specific amino-terminal α-helical extension packed against the core $(\beta_2\alpha)_2$ domain common to all these proteins (Figure 1b) [30]. The KilA-N, APSES and TEN-terminal domains lack the equivalent residues of the specific active-site extension of the LAGLI-DADG nucleases, suggesting that the former domains do not possess nuclease activity. Thus, it appears plausible that these domains evolved from an ancestral nucleic-acid-binding module that, on one hand, gave rise to the nucleases through the acquisition of an amino-terminal helical extension that provided the active site, and on the other hand, diversified into distinct nucleic-acid-binding domains. Like the KilA-N domains that are mainly associated with DNA viruses, the LAGLIDADG endonucleases are predominantly encoded by mobile genetic elements [31]. Just as the APSES domain appears to be a derivative of a KilA-N domain captured by the cellular genome, the TEN domains appear to have been derived through the ancient cellular capture of an inactive LAGLIDADG-like domain [32]. The common ancestor of this

fold might have emerged in a mobile genomic symbiont or parasite and subsequently spread widely across viral and transposon genomes.

Conserved domains in the Bro proteins

The baculovirus Bro proteins are encoded by a multigene family and represent another class of virus-specific DNA-binding regulators whose evolutionary affinities are not yet understood. The typical Bro proteins that have been experimentally investigated are BroA, BroC and BroD from *Bombyx mori* nuclear polyhedrosis virus (BmNV) [11]. In addition to baculoviruses, we observed that the NCLDV class members, such as poxviruses and iridoviruses, also encoded homologs of the Bro proteins. Proteins such as FPV124 from fowlpox virus and three distinct proteins encoded by the entomopoxvirus AMV showed similarity only to the carboxy-terminal part of the baculovirus Bro proteins, whereas, on their amino terminus, they contain a KilA-N domain (Figure 2). In contrast, another group of entomopoxvirus proteins, such as MSV226 from MSV and AMV262 from AMV, showed similarity only to the amino-terminal part of the baculovirus Bro proteins. Yet another set of viral proteins, such as MSV194 from MSV, ORF117 from the phaeovirus ESV, and baculovirus proteins, such as BroE of BmNV, combine the region homologous to the amino-terminal segment of the typical Bro proteins with another distinct domain that occurs in a stand-alone form in the phage T5 ORF172 protein. This suggests that the typical Bro proteins contain distinct amino- and carboxy-terminal domains (Bro-N and Bro-C, respectively) that are present independently of

each other and in distinct contexts in a variety of other viral proteins. To uncover their entire range of diversity, we initiated PSI-BLAST sequence-profile searches, seeded separately with the sequences of the Bro-N and Bro-C domains. The Bro-C domain was essentially restricted to the eukaryotic viruses of the baculovirus and NCLDV classes (Table 1). In contrast, the Bro-N domain was more widely distributed, occurring in a stand-alone form or combined with other domains in proteins from temperate phages that infect Gram-positive bacteria and *Myxococcus xanthus*, and proteins encoded in the genomes of proteobacteria and Gram-positive bacteria. The P22 anti-repressor protein, which regulates phage transcription [33,34], is one of the previously characterized bacteriophage proteins in which the Bro-N domain was observed.

Studies on the BmNV BroA protein have shown that it binds DNA with high affinity and associates with the chromatin in the BmNV-infected cells [11]. Additionally, it has been shown that the DNA-binding determinants of the BroA protein map to the amino-terminal 80 amino acids that correspond to the Bro-N domain defined above [10]. Thus, the Bro-N domain appears to define a distinct superfamily of widespread viral DNA-binding domains. Multiple alignment-based secondary-structure prediction of the Bro-N domain reveals a core with two head-to-tail units of a β-hairpin followed by an α-helix (Figure 3). Thus, the Bro-N domain adopts an α+β fold; furthermore, the pattern of predicted secondary-structure elements was similar to that seen in the KilA-N domain and its relatives (Figure 3). The multiple alignment shows

Figure 3

Multiple alignment of Bro-N domains. The color scheme is as in Figure 1. The coloring reflects the conservation profile of amino acid residues at 85% consensus. Further groupings described reflect domain architectures as follows: 1, fused to KilA-C; 2, fused to Mx8P63C; 3, fused to P22AR-C; 4, *Xylella fastidiosa* specific Bro-N duplications and fusions; 5, fused to T5ORF172; 6, fused to Bro-C; 7, duplicated Bro-N fused to Bro-C; 8, fused to a VSR nuclease; 9, duplicated Bro-N; 10, fused to a HTH. Species abbreviations are listed in Materials and methods.

that the Bro-N domains contain two highly conserved aromatic or hydrophobic residues at the end of the second and fourth conserved strands, a pattern that is reminiscent of similarly conserved residues in the KilA-N and APSES domains (Figure 3). However, sequence profile searches do not detect any significant similarity between the Bro-N and KilA-N domains. The sequence-structure threading with the 3D-PSSM method recovered the APSES domain structure (PDB 1bm8/1bm1) as the best hit for the Bro-N domain, albeit with statistically insignificant E-values. Thus, given that the Bro-N domain is a DNA-binding domain, it appears plausible that it adopts a fold similar to that of the KilA-N domain, although sequence analysis and threading methods failed to provide strong support for this.

Extensive lineage-specific expansion and domain shuffling in DNA-binding regulators containing the KilA-N and Bro domains

There are several analogies between the KilA-N and Bro-N domains in terms of phyletic patterns, intragenomic distribution and domain architectures. Both these domains are widely prevalent in bacteriophages, bacteria and large eukaryotic viruses of the NCLDV and baculovirus classes. Both of them show expansions in particular viral genomes, for example KilA-N in FPV and Bro-N in certain baculoviruses, and the entomopoxvirus MSV. Phylogenetic tree construction for the KilA-N and Bro-N domains using the least-squares and maximum-likelihood methods showed that multiple versions of these domains from a given genome typically grouped together, to the exclusion of members from other genomes (data not shown). Thus, KilA-N and Bro-N domains probably have undergone lineage-specific expansions through amplification of a single ancestral gene. The genes encoding multiple copies of these domains do not necessarily occur in close proximity in the corresponding genomes, indicating that duplications were accompanied by extensive genome rearrangement resulting in dispersion of the paralogous genes.

The KilA-N and the Bro-N domains show additional parallels in the domain architectures of the corresponding proteins. Both domains almost always are located at the amino termini of these proteins and most often are fused to another distinct domain at the carboxyl terminus (Figure 2). On many occasions, the Bro-N and KilA-N domains are fused to the same carboxy-terminal domains (Table 1, Figure 2), such as the Bro-C, T5ORF172, KilA-C, and Mx8P63-C domains. Additionally, Bro-N domains show specific combinations with certain domains, such as HTH and the Vsr-superfamily endonuclease. These architectures suggest that both the KilA-N and Bro-N modules are, to a large extent, functionally equivalent and probably act as the principal DNA-binding moiety that recruits a specific activity purveyed by their carboxy-terminal domains to the target DNA sequences. These carboxy-terminal modules may be enzymes, such as the nuclease domains, or might mediate additional, specific

interactions with nucleic acids or proteins. Thus, the principal function of these proteins appears to be transcriptional regulation of viral or host genes. Consistent with this hypothesis, these proteins include antirepressors from phages such as BPP22, BPVT2 and BP933W.

The majority of the domains with which the KilA-N and Bro-N domains combine in multidomain proteins are restricted to proteins encoded by temperate bacteriophages and large eukaryotic DNA viruses (the exceptions are a few domains that are common in cellular proteomes, such as HTH, CCCH and RING finger). In order to identify the entire repertoire of domains involved in these domain-shuffling events, we systematically explored all domains that combined with the KilA-N and Bro-N domains and compiled a list of domains with which the latter combined in other multidomain proteins (Table 2, Figure 2a). This information is represented as a directed graph in Figure 2b, in which each vertex corresponds to a particular domain and the edges connect domains that combine to form multidomain proteins. The entire network consists of 24 domains, of which Bro-N and KilA-N domains show by far the greatest number of connections (12 and 7, respectively) compared to other domains. The RHA, φ31ORF238-N and MSV199 domains, while less versatile in their connections, tend to combine with the same domains, and in the same orientation, as Bro-N and KilA-N (Figure 2b), which suggests an analogous function, such as DNA binding.

Conclusions

We show here that KilA-N and Bro-N domains are two DNA-binding domains that are widespread in large DNA viruses infecting bacteria and eukaryotes. At least the former, and perhaps even the latter, appear to belong to a large class of nucleic-acid-binding domains that includes the APSES, LAGLIDADG endonuclease and TEN domains. The fungus-specific transcription factors containing the APSES domain appear to have evolved through capture of a viral KilA-N-like precursor early in fungal evolution. KilA-N and Bro-N domains combine with overlapping sets of carboxy-terminal domains, which, in turn, combine with several additional domains, resulting in an extensive network of 18 domains that are predominantly specific to large DNA viruses and six domains acquired by these viruses from the host genomes (18 + 6 = 24). These observations establish a major role for shuffling within a limited set of domains during evolution of viral DNA-binding regulatory proteins.

Materials and methods

The non-redundant (NR) database of protein sequences was searched using the BLASTP program [35]. Profile searches were conducted using the PSI-BLAST program with either a single sequence or an alignment used as the query, with a profile-inclusion expectation (E) value threshold of 0.01, and

were iterated until convergence [35,36]. If whole-length proteins were used, the searches were carried out using the composition-based statistics [37] in order to prevent the detection of spurious matches that could arise from low-complexity segments in the query or target proteins. In searches carried out with just the globular domain of a particular protein, the composition-based statistics was not used because this helps in improving the sensitivity of searches without a major risk of corruption of the profile with false positives [37]. Additional searches with hidden Markov models were performed using the HMMER package [38,39]. Previously known conserved protein domains were detected using the corresponding position-specific scoring matrices (PSSMs) constructed using PSI-BLAST [40].

Multiple alignments of protein sequences were constructed using the T-Coffee program [41], followed by manual correction based on the PSI-BLAST results. Protein secondary structure was predicted using a multiple alignment as the input for the JPRED program [19]. Sequence-structure threading was performed using the hybrid fold-prediction method, which combines multiple-alignment information with secondary-structure prediction [42], and the 3D-PSSM method [43]. Phylogenetic trees were constructed using neighbor-joining, least-squares and maximum-likelihood methods [44,45]. Species abbreviations used in this paper are as follows: AcNV, *Autographa californica* nucleopolyhedrosis virus; AgNV, *Anticarsia gemmatalis* nucleopolyhedrosis virus; AMV, *Amsacta moorei* entomopoxvirus; BPBK5T, *Lactococcus* phage BK5-T; BmNV, *Bombyx mori* nuclear polyhedrosis virus; BP933W, bacteriophage 933W; BPA118, bacteriophage A118; BPAPSE1, bacteriophage APSE1; BPbIL285, bacteriophage bIL285; BPbIL286, bacteriophage bIL286; BPbIL309, bacteriophage bIL309; BPD3, bacteriophage D3; BPφ80, bacteriophage φ80; BPHK620, bacteriophage HK620; BPHK97, bacteriophage HK97; BPLLH, bacteriophage LLH; BPMx8, bacteriophage Mx8; BPN15, bacteriophage N15; BPP1, bacteriophage P1; BPP22, bacteriophage P22; BPP4, bacteriophage P4; BPP27, bacteriophage P27; BPφPV, bacteriophage φPV; BPR1T, bacteriophage R1T; BPSpβC2,bacteriophage SpβC2; BPT5, bacteriophage T5; BPVT2-Sa, bacteriophage VT2; Cab, *Clostridium acetobutylicum*; Cal, *Candida albicans*; CIV, Chilo iridescent virus; CnBV, *Culex nigripalpus* baculovirus; CpGV, *Cydia pomonella* granulovirus; DpAV4, *Diadromus pulchellus* ascovirus; Ec, *Escherichia coli*; EpNV, *Epiphyas postvittana* nucleopolyhedrovirus; EV, ectromelia virus; Eni, *Emericella nidulans*; ESV, *Ectocarpus siliculosus* virus; FPV, fowlpox virus; HaEPV, *Heliothis armigera* entomopoxvirus; HaNV, *Heliocoverpa armigera* nucleopolyhedrovirus G4; HK97, bacteriophage HK97; HzNV, *Helicoverpa zea* single nucleocapsid nucleopolyhedrovirus; Kla, *Kluyveromyces lactis*; LcBPA2, *Lactobacillus casei* bacteriophage A2; LdNV, *Lymantria dispar* nucleopolyhedrovirus; LsNV, *Leucania separata* nuclear polyhedrosis virus; MbNV, *Mamestra brassicae* nucleopolyhedrovirus;

MSV, *Melanoplus sanguinipes* entomopoxvirus; Nc, *Neurospora crassa*; Nm, *Neisseria meningitidis*; OpNV, *Orgyia pseudotsugata* single capsid nuclear polyhedrosis virus; Pa, *Pseudomonas aeruginosa*; PxGV, *Plutella xylostella* granulovirus; Sa, *Staphylococcus aureus*; Sc, *Saccharomyces cerevisiae*; Scoe, *Streptomyces coelicolor*; SeNV, *Spodoptera exigua* nucleopolyhedrovirus; Sf, *Shigella flexneri*; Sp, *Schizosaccharomyces pombe*; SpLNV, *Spodoptera litura* nucleopolyhedrovirus; Spy, *Streptococcus pyogenes*; VAR, variola virus; Xf, *Xylella fastidiosa*; XnGV, *Xestia c-nigrum* granulovirus; Yaly, *Yarrowia lipolytica*.

Additional data files
Alignments of all viral specific domains identified in this work are available with this article online and at [46].

References
1. Lodish H, Baltimore D, Berk A, Zipursky SL, Matsudaira P, Darnell J: *Molecular Cell Biology.* New York: WH Freeman; 1995.
2. Fields BN: *Fields Virology.* Boston, MA: Lippincott, Williams & Williams;1996.
3. Cann A: *Principles of Molecular Virology.* San Diego: Academic Press; 2001.
4. Ptashne M: *A Genetic Switch: Phage λ and Higher Organisms.* Cambridge, MA: Blackwell Scientific Publications; 1992.
5. Aravind L, Koonin EV: **DNA-binding proteins and evolution of transcription regulation in the archaea.** *Nucleic Acids Res* 1999, **27:**4658-4670.
6. Sauer RT, Yocum RR, Doolittle RF, Lewis M, Pabo CO: **Homology among DNA-binding proteins suggests use of a conserved super-secondary structure.** *Nature* 1982, **298:**447-451.
7. Raumann BE, Rould MA, Pabo CO, Sauer RT: **DNA recognition by beta-sheets in the Arc repressor-operator crystal structure.** *Nature* 1994, **367:**754-757.
8. Sauer RT, Krovatin W, DeAnda J, Youderian P, Susskind MM: **Primary structure of the immI immunity region of bacteriophage P22.** *J Mol Biol* 1983, **168:**699-713.
9. Iyer LM, Aravind L, Koonin EV: **Common origin of four diverse families of large eukaryotic DNA viruses.** *J Virol* 2001, **75:**11720-11734.
10. Zemskov EA, Kang W, Maeda S: **Evidence for nucleic acid binding ability and nucleosome association of *Bombyx mori* nucleopolyhedrovirus BRO proteins.** *J Virol* 2000, **74:**6784-6789.
11. Kang W, Suzuki M, Zemskov E, Okano K, Maeda S: **Characterization of baculovirus repeated open reading frames (bro) in *Bombyx mori* nucleopolyhedrovirus.** *J Virol* 1999, **73:**10339-10345.
12. Brick DJ, Burke RD, Schiff L, Upton C: **Shope fibroma virus RING finger protein N1R binds DNA and inhibits apoptosis.** *Virology* 1998, **249:**42-51.
13. Senkevich TG, Koonin EV, Buller RM: **A poxvirus protein with a RING zinc finger motif is of crucial importance for virulence.** *Virology* 1994, **198:**118-128.
14. Hansen EB: **Structure and regulation of the lytic replicon of phage P1.** *J Mol Biol* 1989, **207:**135-149.
15. Aramayo R, Peleg Y, Addison R, Metzenberg R: **Asm-1+, a *Neurospora crassa* gene related to transcriptional regulators of fungal development.** *Genetics* 1996, **144:**991-1003.
16. Xu RM, Koch C, Liu Y, Horton JR, Knapp D, Nasmyth K, Cheng X: **Crystal structure of the DNA-binding domain of Mbp1, a transcription factor important in cell-cycle control of DNA synthesis.** *Structure* 1997, **5:**349-358.
17. Taylor IA, Treiber MK, Olivi L, Smerdon SJ: **The X-ray structure of the DNA-binding domain from the *Saccharomyces cerevisiae* cell-cycle transcription factor Mbp1 at 2.1 Å resolution.** *J Mol Biol* 1997, **272:**1-8.
18. Heath PJ, Stephens KM, Monnat RJ Jr, Stoddard BL: **The structure of I-CreI, a group I intron-encoded homing endonuclease.** *Nat Struct Biol* 1997, **4:**468-476.

19. Cuff JA, Barton GJ: **Evaluation and improvement of multiple sequence methods for protein secondary structure prediction.** *Proteins* 1999, **34**:508-519.
20. Holm L, Sander C: **Dictionary of recurrent domains in protein structures.** *Proteins* 1998, **33**:88-96.
21. Li H, Trotta CR, Abelson J: **Crystal structure and evolution of a transfer RNA splicing enzyme.** *Science* 1998, **280**:279-284.
22. Madej T, Gibrat JF, Bryant SH: **Threading a database of protein cores.** *Proteins* 1995, **23**:356-369.
23. Bujnicki JM, Rychlewski L: **Unusual evolutionary history of the tRNA splicing endonuclease EndA: relationship to the LAGLIDADG and PD-(D/E)XK deoxyribonucleases.** *Protein Sci* 2001, **10**:656-660.
24. Dutton JR, Johns S, Miller BL: **StuAp is a sequence-specific transcription factor that regulates developmental complexity in** *Aspergillus nidulans.* *EMBO J* 1997, **16**:5710-5721.
25. Koch C, Nasmyth K: **Cell cycle regulated transcription in yeast.** *Curr Opin Cell Biol* 1994, **6**:451-459.
26. Koch C, Moll T, Neuberg M, Ahorn H, Nasmyth K: **A role for the transcription factors Mbp1 and Swi4 in progression from G1 to S phase.** *Science* 1993, **261**:1551-1557.
27. Chervitz SA, Aravind L, Sherlock G, Ball CA, Koonin EV, Dwight SS, Harris MA, Dolinski K, Mohr S, Smith T, *et al.*: **Comparison of the complete protein sets of worm and yeast: orthology and divergence.** *Science* 1998, **282**:2022-2028.
28. Aravind L, Subramanian G: **Origin of multicellular eukaryotes - insights from proteome comparisons.** *Curr Opin Genet Dev* 1999, **9**:688-694.
29. Aravind L: **The BED finger, a novel DNA-binding domain in chromatin-boundary-element-binding proteins and transposases.** *Trends Biochem Sci* 2000, **25**:421-423.
30. Jurica MS, Monnat RJ Jr, Stoddard BL: **DNA recognition and cleavage by the LAGLIDADG homing endonuclease I-CreI.** *Mol Cell* 1998, **2**:469-476.
31. Belfort M, Reaban ME, Coetzee T, Dalgaard JZ: **Prokaryotic introns and inteins: a panoply of form and function.** *J Bacteriol* 1995, **177**:3897-3903.
32. Anantharaman V, Koonin EV, Aravind L: **Comparative genomics and evolution of proteins involved in RNA metabolism.** *Nucleic Acids Res* 2002, in press.
33. Botstein K, Lew KK, Jarvik V, Swanson CA: **Role of antirepressor in the bipartite control of repression and immunity by bacteriophage P22.** *J Mol Biol* 1975, **91**:439-462.
34. Susskind MM, Botstein D: **Mechanism of action of** *Salmonella* **phage P22 antirepressor.** *J Mol Biol* 1975, **98**:413-424.
35. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**:3389-3402.
36. Aravind L, Koonin EV: **Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches.** *J Mol Biol* 1999, **287**:1023-1040.
37. Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, Altschul SF: **Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements.** *Nucleic Acids Res* 2001, **29**:2994-3005.
38. Eddy SR: **Hidden Markov models.** *Curr Opin Struct Biol* 1996, **6**:361-365.
39. Eddy SR: **Profile hidden Markov models.** *Bioinformatics* 1998, **14**:755-763.
40. Schaffer AA, Wolf YI, Ponting CP, Koonin EV, Aravind L, Altschul SF: **IMPALA: matching a protein sequence against a collection of PSI-BLAST- constructed position-specific score matrices.** *Bioinformatics* 1999, **15**:1000-1011.
41. Notredame C, Higgins DG, Heringa J: **T-Coffee: A novel method for fast and accurate multiple sequence alignment.** *J Mol Biol* 2000, **302**:205-217.
42. Fischer D: **Hybrid fold recognition: combining sequence derived properties with evolutionary information.** *Pac Symp Biocomput* 2000, 119-130.
43. Kelley LA, MacCallum RM, Sternberg MJ: **Enhanced genome annotation using structural profiles in the program 3D-PSSM.** *J Mol Biol* 2000, **299**:499-520.
44. Hasegawa M, Kishino H, Saitou N: **On the maximum likelihood method in molecular phylogenetics.** *J Mol Evol* 1991, **32**:443-445.
45. Felsenstein J: **Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods.** *Methods Enzymol* 1996, **266**:418-427.
46. **Multiple alignments of domains, other than KilA-N and Bro-N, described in this study** [ftp://ftp.ncbi.nlm.nih.gov/pub/aravind/viral_domains.txt]
47. Henthorn KS, Friedman DI: **Identification of related genes in phages phi 80 and P22 whose products are inhibitory for phage growth in** *Escherichia coli* **IHF mutants.** *J Bacteriol* 1995, **177**:3185-3190.
48. Clerget M, Boccard F: **Phage HK022 Roi protein inhibits phage lytic growth in** *Escherichia coli* **integration host factor mutants.** *J Bacteriol* 1996, **178**:4077-4083.
49. Riedel HD, Heinrich J, Heisig A, Choli T, Schuster H: **The antire-pressor of phage P1. Isolation and interaction with the C1 repressor of P1 and P7.** *FEBS Lett* 1993, **334**:165-169.