BIOINFORMATICS DISCOVERY NOTE Vol. 24 no. 4 2008, pages 458–461
doi:10.1093/bioinformatics/btn007

Sequence analysis

BEN: a novel domain in chromatin factors and DNA viral proteins

Saraswathi Abhiman, Lakshminarayan M. Iyer and L. Aravind*

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

Received and revised on November 20, 2007; accepted on January 3, 2008

Advance Access publication January 18, 2008

Associate Editor: Alex Bateman

ABSTRACT

We report a previously uncharacterized α-helical module, the BEN domain, in diverse animal proteins such as BANP/SMAR1, NAC1 and the *Drosophila* mod(mdg4) isoform C, in the chordopoxvirus virosomal protein E5R and in several proteins of polydnaviruses. Contextual analysis suggests that the BEN domain mediates protein–DNA and protein–protein interactions during chromatin organization and transcription. The presence of BEN domains in a poxviral early virosomal protein and in polydnaviral proteins also suggests a possible role for them in organization of viral DNA during replication or transcription.

Contact: aravind@ncbi.nlm.nih.gov

Supplementary information: Supplementary data for this study can also be accessed at http://www.ncbi.nlm.nih.gov/CBBresearch/Lakshmin/BEN/

1 INTRODUCTION

Eukaryotes are distinguished by their complex chromatin, which directly and indirectly affects all key nuclear events such as DNA replication and repair, transcription and post-transcriptional regulation. The dynamics of eukaryotic chromatin is mediated by several distinctive, large protein complexes. These include enzymes that covalently modify histones (the histone code), such as acetylases and methylases, or remove modifications, such as deacetylases and demethylases, or remodel chromatin-using energy from ATP-hydrolysis, such as the SWI2/SNF2 and MORC ATPases (Allis *et al.*, 2006; Kouzarides, 2007). The specificity of these processes, as well as the interpretation of the histone code is facilitated by a wide array of domains that interact with other proteins or nucleic acids, or specifically bind covalently modified peptides. Evolutionary analyses suggest that the majority of these adaptors' domains and enzymes acquired their chromatin-related role only in the eukaryotes (Iyer *et al.*, 2008). Moreover, several of these domains and specific transcription factors (TFs) are only found in a limited set of eukaryotic taxa suggesting that lineage-specific innovations have been critical to the evolution of this system. Understanding the role and origins of these domains is necessary for generating a complete picture of the processes of transcriptional regulation and chromatin organization. In this study, we report a novel domain in animal TFs, chromatin proteins and polypeptide encoded by two unrelated groups of large animal DNA viruses. Based on conserved sequence features and contextual analysis, we predict this to function as an adaptor for the higher-order structuring of chromatin, and recruitment of chromatin modifying factors in transcriptional regulation.

*To whom correspondence should be addressed.

2 METHODS

Profile-based searches against the NR database were performed using the PSI-BLAST (Altschul *et al.*, 1997) and HMMER (Eddy, 1998) programs. The BLASTCLUST program was used for clustering protein sequences (ftp://ftp.ncbi.nih.gov/blast/documents/blastclust.html). Multiple alignments were constructed using the KALIGN program (Lassmann and Sonnhammer, 2005) followed by polishing using the PSI-BLAST HSPs. The multiple alignment was used to predict a protein secondary structure using the JPRED program (Cuff and Barton, 2000). The MEGA software was used to construct phylogenetic trees (Kumar *et al.*, 2004). For a detailed description of the methods and their application, refer to the Supplementary Material.

3 RESULTS

3.1 Identification, phyletic distribution and evolutionary history of the BEN domain

In course of a comprehensive analysis of domains involved in transcription regulation and chromatin function, we identified an uncharacterized, predicted globular region in the vertebrate POZ-domain protein NAC1 (overlapping partially with the DUF1172 in Pfam database). A search with this region retrieved homologous segments in diverse animal and viral proteins in profile-based PSI-BLAST and HMM searches. For example, PSI-BLAST searches with the C-terminal region of the human NAC1 as query (gi: 16418383, amino acids 300–500) retrieved, with significant *E*-values ($E<10^{-3}$) prior to convergence, human BANP/SMAR1, *Drosophila* Insensitive (CG3227), several proteins from polydnaviruses (e.g. CcBV_3.4 of *Cotesia congregata* bracovirus), a mod(mdg4) isoform in dipterans (e.g. in *Anopheles gambiae*; gi: 119112359), a potential insect TF (*Aedes aegypti* gi: 157104034) and several poorly characterized human proteins and their animal orthologs such as human C6orf65, CCDC4 and C1orf165. Homologous segments were also found in multiple tandem copies in several proteins in the cnidaria (e.g. *Nematostella* gi: 156383934: 5 copies, gi: 156383936: 3 copies) and vertebrates (e.g. human Cxorf20: 2 copies and human KIAA1553: 4 copies). Further transitive searches, such as with one of the copies from a *Nematostella*

Published by Oxford University Press (2008)

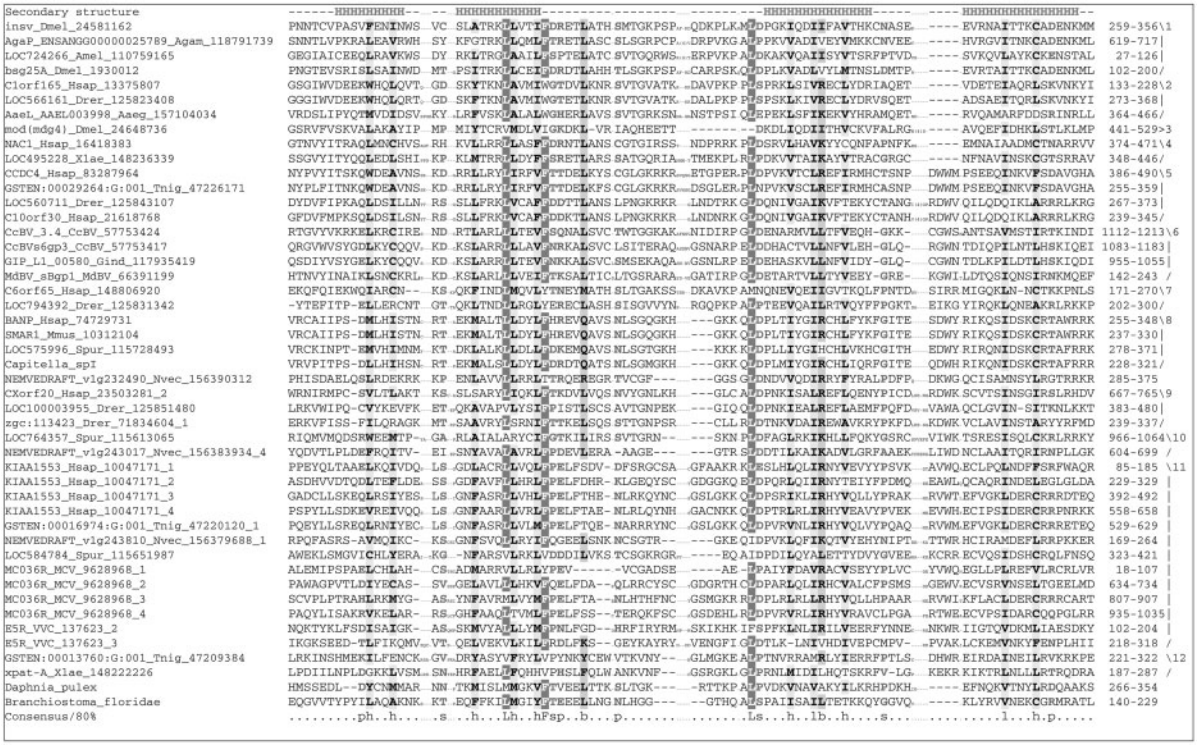

Fig. 1. Multiple sequence alignment of the BEN domain. Proteins are represented by their gene names, species abbreviations and gis. The sequence range and families, represented by numbers, are given to the right of the alignment. F 1, Drosophila insensitive; 2, C1orf165; 3, mod(mdg4); 4, NAC1; 5, CCDC4; 6, polydnavirus family; 7, C6orf65; 8, BANP/SMAR1; 9, Cxorf20; 10, NEMVEDRAFT_v1g243017; 11, E5R/KIAA1553; 12, Xpat. The coloring reflects the consensus at 80% conservation calculated from a more extensive alignment (Supplementary Material). Consensus abbreviations are h, hydrophobic; l, aliphatic; s, small; p, polar; b, big. Species abbreviations are as follows: Aaeg, *Aedes aegypti*; Agam, *Anopheles gambiae*; Amel, *Apis mellifera*; Btau, *Bos taurus*; CcBV, *Cotesia congregata* bracovirus; CpPV, *Cotesia plutellae* polydnavirus; Dmel, *Drosophila melanogaster*; Drer, *Danio rerio*; Gind, *Glyptapanteles indiensis*; Hsap, *Homo sapiens*; MCV, Molluscum contagiosum virus; MdBV, *Microplitis demolitor* bracovirus; Mmus, *Mus musculus*; Nvec, *Nematostella vectensis*; Spur, *Strongylocentrotus purpuratus*; Tcas, *Tribolium castaneum*, Tnig, *Tetraodon nigroviridis*; VVC, Vaccinia virus; Xlae, *Xenopus laevis*.

protein (gi: 156383936, region 380–540), retrieved multiple repeats with significant *E*-values in the vaccinia virus E5R and its orthologs from various chordopoxviruses and the *Xenopus* protein Xpat. Analysis of incompletely sequenced eukaryotic genomes revealed several copies of the domain in the cephalochordate *Branchiostoma*, the crustacean *Daphnia* and the mollusk *Lottia* and a single copy in the annelids *Capitella* (a polychaete) and *Helobdella* (a leech) (see Supplementary Material for a comprehensive list of sequences retrieved).

The shared region of conservation present in one or more copies in these proteins thus appears to define a novel domain that we refer to as the BEN domain after experimentally characterized proteins BANP, E5R and NAC1 in which it is present. While in NAC1 and some of its close relatives it overlapped partially with the alignment annotated as DUF1172 in PFAM, the relationships defined here were unnoticed in the majority of the proteins in which they were identified (Fig. 1). Furthermore, the boundaries of the BEN domain defined here accurately represent the actual region of homology shared by all the above proteins. Prediction of the secondary structure using the multiple alignment indicated an all α-fold, with four conserved helices, for the BEN domain (Fig. 1). Its conservation pattern revealed several conserved residues, most of which have hydrophobic side-chains and are likely to stabilize the fold through helix–helix packing (Fig. 1). The most characteristic signatures of the domain are a LhxxlFs motif (l, aliphatic; s, small; x, any residue) in helix 2 and an aliphatic residue (mostly leucine) at the beginning of helix 3 (Fig. 1).

In order to establish the phyletic patterns and evolutionary history of the BEN domain, we clustered the retrieved proteins using BLASTCLUST and further grouped them using shared sequence features and a phylogenetic tree. As a result we obtained 12 distinct families. Of these, the families typified by **E5R/KIAA1553** and **NEMVEDRAFT_v1g243017** appear to have been present from early in animal evolution, being present in the cnidarian *Nematostella*. Most others, including the family prototyped by **BANP/SMAR1** are present both in a wide range of invertebrates and vertebrates, whereas those typified by **NAC1**, **CCDC4** and **Cxorf20** appear to be restricted to chordates (Supplementary Material). Many families show a sporadic distribution: for example, that defined by C10orf165 is only present in vertebrates and *A.aegypti*, while orthologs of NEMVEDRAFT_v1g243017 are only detected in cnidarians and sea urchins. BEN domain proteins also appear to have been entirely lost in certain animal lineages, such as nematodes and urochordates.

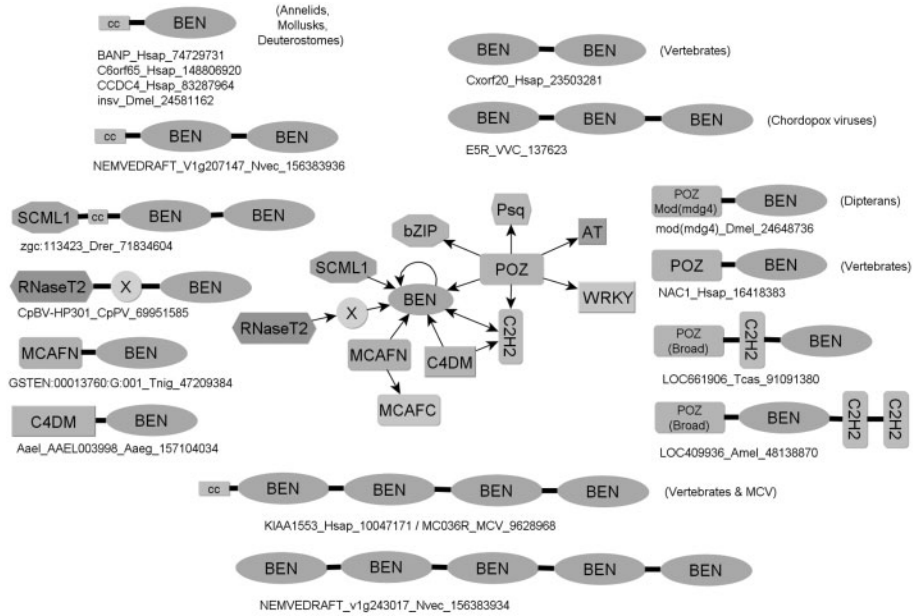

Fig. 2. Domain architectures and context graph. Domain architectures are labeled with the gene name, species abbreviation and gi numbers separated by underscores. The contextual graph in the center represents domain architectures of BEN and POZ domain containing proteins. The arrows represent the directionality of the domain organization with the arrowhead pointing to the C-terminus. Domains are denoted by their standard abbreviations. Additional domain abbreviations include: CC, coiled coil regions; SCML1, Sex comb on mid-leg1 N-terminal like domain; X, uncharacterized polydnavirus specific domain; Species abbreviations are as in Figure 1.

The E5R proteins with three tandem BEN domains are phylogenetically closest to the KIAA1553-like proteins of animals with quadruple-BEN domains, and are detected in ortho-, capri-, lepori- and yata-poxviruses. This suggests an early transfer from a vertebrate host before the radiation of the different chordopoxvirus lineages. Interestingly, in phylogenetic trees the multiple BEN domains of the Molluscum contagiosum virus (MCV) MC036R protein closely group with the multiple copies of the mammalian KIAA1553, rather than the other chordopoxviruses. Moreover, MCV has four tandem BEN domains like the KIAA1553 proteins (Fig. 2). This suggests that in MCV, the original E5R protein was displaced by another more recent transfer of a mammalian KIAA1553. The polydnavirus BEN domain family is related to the NAC1 and CCDC4 families suggesting an independent acquisition by these viruses. There is considerable diversity in the number of copies of this domain coded by different polydnaviruses: the *C.congregata* bracovirus has 11 BEN domain containing proteins coded by 7 of the 30 genomic DNA circles, while *Microplitis demolitor* bracovirus codes a single BEN domain. BEN domain-coding gene is also one of the polydnaviral genes transferred to host wasp genome (Desjardins *et al*., 2007). These phyletic distributions suggest that the BEN domain was an early lineage-specific innovation in the animals, either *de novo*, or from an unrecognized preexisting α-helical domain, followed by at least three independent transfers to two unrelated classes of animal viruses.

3.2 Functional predictions for the BEN domain from contextual analysis

Of the experimentally studied proteins with BEN domains, NAC1 interacts with CoRest and histone deacetylases and the region encompassing the BEN domain is one of the regions shown to be required for interaction with histone deacetylases HDAC3 and HDAC4 (Korutla *et al*., 2005, 2007). The transcriptional repressor and candidate tumor suppressor BANP/SMAR1 is a matrix attachment region (MAR)-interacting protein that also interacts with the MAR-binding protein Cux/CDP, and the SIN3-histone deacetylase complex. The region encompassing the BEN domain has been implicated in these interactions. The C-terminal region of the BEN domain also overlaps with the MAR-binding region in BANP/SMAR1 (Kaul-Ghanekar *et al*., 2004; Rampalli *et al*., 2005). The Xenopus BEN domain protein Xpat has been shown to be a nuclear- and germplasm-localized protein (Machado *et al*., 2005). In order to gain further functional insights, we analyzed the domain architectural contexts of the BEN domain.

BEN domains are also linked in polypeptides to other globular domains with functions related to transcriptional regulation and chromatin structure, such as POZ, C4DM, C2H2 fingers, MCAF N-terminal domain (MCAFN) and a domain that is also found N-terminal to the SAM domain in sex combs in midleg-like-1, a protein of the vertebrate polycomb complex (van de Vosse *et al*., 1998) (Fig. 2). The most striking of these is the association of BEN with the POZ domain, which appears to have occurred on multiple independent occasions: to a Nac1-like POZ domain in vertebrates a mod(mdg4) POZ in dipterans [e.g. *Drosophila* mod(mdg4) isoform C], and to a Broad complex-type POZ domain in a honeybee and a beetle protein (Fig. 2). Some of these proteins might also contain one or more C2H2 fingers (Fig. 2). The multiple independent fusions of distinct BEN domains to distinct POZ domains suggest an intimate functional association between the two domains.

POZ domains are protein–protein interaction domains found in a wide range of functional contexts including chromatin organization (Aravind and Koonin, 1999). POZ domains are often present N-terminal to DNA-binding domains such as C2H2 and WRKY(FLYWCH family) fingers, bZIP, AT-hooks and pipsqueak (Aravind and Koonin, 1999; Dorn and Krauss, 2003) (Fig. 2). Moreover, in both the mod(mdg4) and Broad complex loci, a single POZ domain participates in multiple isoforms via splicing of exons coding distinct DNA-binding domains such as WRKY, C2H2 fingers and AT-hook (Aravind and Koonin, 1999; Dorn and Krauss, 2003). Similarly, the BEN domain is also fused to a N-terminal C4DM domain (Fig. 2) that is usually present N-terminal to DNA-binding C2H2 fingers in several insect TFs (Lander *et al*., 2001). These contextual patterns derived from independent polypeptides would hint that the BEN domain is a DNA-binding domain occurring C-terminal to a POZ or C4DM domain. This is consistent with the region overlapping with the BEN domain interacting with MARs in BANP/SMAR1. Fishes possess a protein (e.g. *Tetraodon* gi: 47209384), where the BEN domain is fused to the MCAFN domain. The MCAFN domain has been shown to bind the histone methylase ESET (Ichimura *et al*., 2005), suggesting that the BEN domain might collaborate with it in recruiting chromatin-modifying activities. These observations, together with the role of the BEN domain in interactions with the histone deacetylase complex, suggest that it could alternatively function as adaptor domain in chromatin modification.

Several BEN domain proteins have 2–4 tandem copies of the domain (Fig. 2). In phylogenetic trees, tandem copies of a particular family are always closer to each other in sequence in comparison to those of other families, suggesting that the duplication events that led to the formation of tandem copies occurrences independently in different families. This propensity to form tandem copies might suggest an inherent property of the BEN domain to form multimeric assemblies through helix–helix interactions. Additionally, majority of the families of BEN domain proteins contain coiled-coil regions that might further assist in the multimerization of these proteins, with an extended interaction surface formed by the BEN domains (Fig. 2).

The independent acquisition of the BEN domain by two groups of large DNA viruses hints at a possible role in viral chromatin organization. The poxviral E5R protein, which is composed of 3–4 BEN domains, is an abundant early protein in the virosomes (Murcia-Nicolas *et al*., 1999). The virosome is a site of active DNA replication (Netherton *et al*., 2007), where the E5R might help in organization of viral DNA. Interestingly, multiple polydnaviral proteins show a fusion of the BEN domain to an RNaseT2 domain, suggesting that these proteins might also participate in an as yet unknown aspect of RNA processing in these viruses. However, it is also possible that the viral versions are used to modify host cell function by mimicking interactions of the endogenous versions.

4 CONCLUSIONS

Our investigations reveal a hitherto uncharacterized animal-specific domain found in several TFs, chromatin proteins and proteins from poxviruses and polydnaviruses. Sequence analysis and contextual information provide evidence that it might function as a DNA-binding protein or an adaptor recruiting chromatin-modifying complexes. We hope that these findings would provide the stimulus for further experimental studies to address precise roles of this domain.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the Intramural research program of the National Library of Medicine, National Institutes of Health, USA, for funding their research.

Conflict of Interest: none declared.

REFERENCES

Allis,C.D. *et al*. (2006) *Epigenetics*. Cold Spring Harbor Laboratory Press, New York.

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.

Aravind,L. and Koonin,E.V. (1999) Fold prediction and evolutionary analysis of the POZ domain: structural and evolutionary relationship with the potassium channel tetramerization domain. *J. Mol. Biol*., **285**, 1353–1361.

Cuff,J.A. and Barton,G.J. (2000) Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins*, **40**, 502–511.

Desjardins,C.A. *et al*. (2007) Structure and evolution of a proviral locus of Glyptapanteles indiensis bracovirus. *BMC Microbiol*., **7**, 61.

Dorn,R. and Krauss,V. (2003) The modifier of mdg4 locus in Drosophila: functional complexity is resolved by trans splicing. *Genetica*, **117**, 165–177.

Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Ichimura,T. *et al*. (2005) Transcriptional repression and heterochromatin formation by MBD1 and MCAF/AM family proteins. *J. Biol. Chem*., **280**, 13928–13935.

Iyer,L.M. *et al*. (2008) Comparative genomics of transcription factors and chromatin proteins in parasitic protists and other eukaryotes. *Int. J. Parasitol*, **38**, 1–31.

Kaul-Ghanekar,R. *et al*. (2004) SMAR1 and Cux/CDP modulate chromatin and act as negative regulators of the TCRbeta enhancer (Ebeta). *Nucleic Acids Res*., **32**, 4862–4875.

Korutla,L. *et al*. (2007) NAC1, a cocaine-regulated POZ/BTB protein interacts with CoREST. *J. Neurochem*., **101**, 611–618.

Korutla,L. *et al*. (2005) The POZ/BTB protein NAC1 interacts with two different histone deacetylases in neuronal-like cultures. *J. Neurochem*., **94**, 786–793.

Kouzarides,T. (2007) Chromatin modifications and their function. *Cell*, **128**, 693–705.

Kumar,S. *et al*. (2004) MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief Bioinform*., **5**, 150–163.

Lander,E.S. *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Lassmann,T. and Sonnhammer,E.L. (2005) Kalign–an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics*, **6**, 298.

Machado,R.J. *et al*. (2005) Xenopus Xpat protein is a major component of germ plasm and may function in its organisation and positioning. *Dev. Biol*., **287**, 289–300.

Murcia-Nicolas,A. *et al*. (1999) Identification by mass spectroscopy of three major early proteins associated with virosomes in vaccinia virus-infected cells. *Virus Res*., **59**, 1–12.

Netherton,C. *et al*. (2007) A guide to viral inclusions, membrane rearrangements, factories, and viroplasm produced during virus replication. *Adv. Virus Res*., **70**, 101–182.

Rampalli,S. *et al*. (2005) Tumor suppressor SMAR1 mediates cyclin D1 repression by recruitment of the SIN3/histone deacetylase 1 complex. *Mol. Cell. Biol*., **25**, 8415–8429.

van de Vosse,E. *et al*. (1998) Characterization of SCML1, a new gene in Xp22, with homology to developmental polycomb genes. *Genomics*, **49**, 96–102.