The BAH (bromo-adjacent homology) domain: a link between DNA methylation, replication and transcriptional regulation

Isabelle Callebaut[a,*], Jean-Claude Courvalin[b], Jean-Paul Mornon[a]

[a] *Systèmes moléculaires and Biologie structurale, LMCP, CNRS UMR 7590, Universités Paris 6 et Paris 7, case 115, 4 place Jussieu, F75252 Paris Cedex 05, France*

[b] *Institut Jacques Monod, CNRS, Université Paris 7, Tour 43, 2 place Jussieu, F75251 Paris Cedex 05, France*

Received 18 January 1999

Abstract Using sensitive methods of sequence analysis including hydrophobic cluster analysis, we report here a hitherto undescribed family of modules, the BAH (bromo-adjacent homology) family, which includes proteins such as eukaryotic DNA (cytosine-5) methyltransferases, the origin recognition complex 1 (Orc1) proteins, as well as several proteins involved in transcriptional regulation. The BAH domain appears to act as a protein-protein interaction module specialized in gene silencing, as suggested for example by its interaction within yeast Orc1p with the silent information regulator Sir1p. The BAH module might therefore play an important role by linking DNA methylation, replication and transcriptional regulation.

Key words: DNA methyltransferase; Orc1; Sir3; ASH1; Silencing

1. Introduction

DNA methylation of cytosine residues at the CpG dinucleotide sequence is an important mechanism of epigenetic regulation of genomic function. In particular, CpG methylation is involved in gene silencing by participating in the recruitment of histone deacetylases via the methyl-binding protein MeCP2, generating a chromatin structure that limits promoter accessibility [1]. Faithful clonal transmission of methylation patterns requires close coordination of replication and methylation, and is achieved by the recognition during the S-phase of hemimethylated CpG sites by a maintenance DNA methyltransferase. Several regions within the large regulatory NH_2-terminal domain of Dnmt1, the major maintenance DNA (cytosine-5) methyltransferase (DNA MTase) in mammals, appear to link the enzyme to the replication machinery (Fig. 2, MTDM identifier). First, a particular sequence encompassing amino acids (aa) 207–455, targets the mouse enzyme to replication foci [2]. Another short segment of the regulatory domain (aa 41–53 on the MTDM_HUMAN sequence), sharing similarities with DNA ligase I and the large subunit of replication factor C, binds the proliferating cell nuclear antigen (PCNA), an auxiliary factor for DNA replication and repair [3,4]. A third region of Dnmt1 (aa 574–846), designated PBHD (polybromo homology domain) due to sequence similarities with the chicken polybromo-1 protein (PB1, 23% identity in a 270 aa overlap), is also involved in the targeting of DNA MTase to sites of DNA replication [5]. Part of the PBHD region (aa 670–759) corresponds to a duplicated sequence in PB1 (aa 989–1072; aa 1188–1273), termed BAH for bromo-adjacent homology [6]. Here, we show that the BAH module is larger than initially described, present in a duplicated form in PB1 as well as in animal DNA MTases, and also found in several proteins participating to transcriptional regulation.z

2. Materials and methods

Searches within the non-redundant database (NR) were performed using BLAST2 and PSI-BLAST programs [7] at the National Center for Biological Information (NCBI, USA). Hidden Markov Model (HMM) searches were carried out using the HMMER package [8]. Guidelines to the use of bidimensional hydrophobic cluster analysis (HCzA) are described elsewhere [9,10]. HCA combines sequence comparison with secondary structure predictions and is particularly efficient at low levels of sequence identity (typically below 20–25% sequence identity). Secondary structure predictions were performed using the profile neural network prediction PHD program [11].

3. Results and discussion

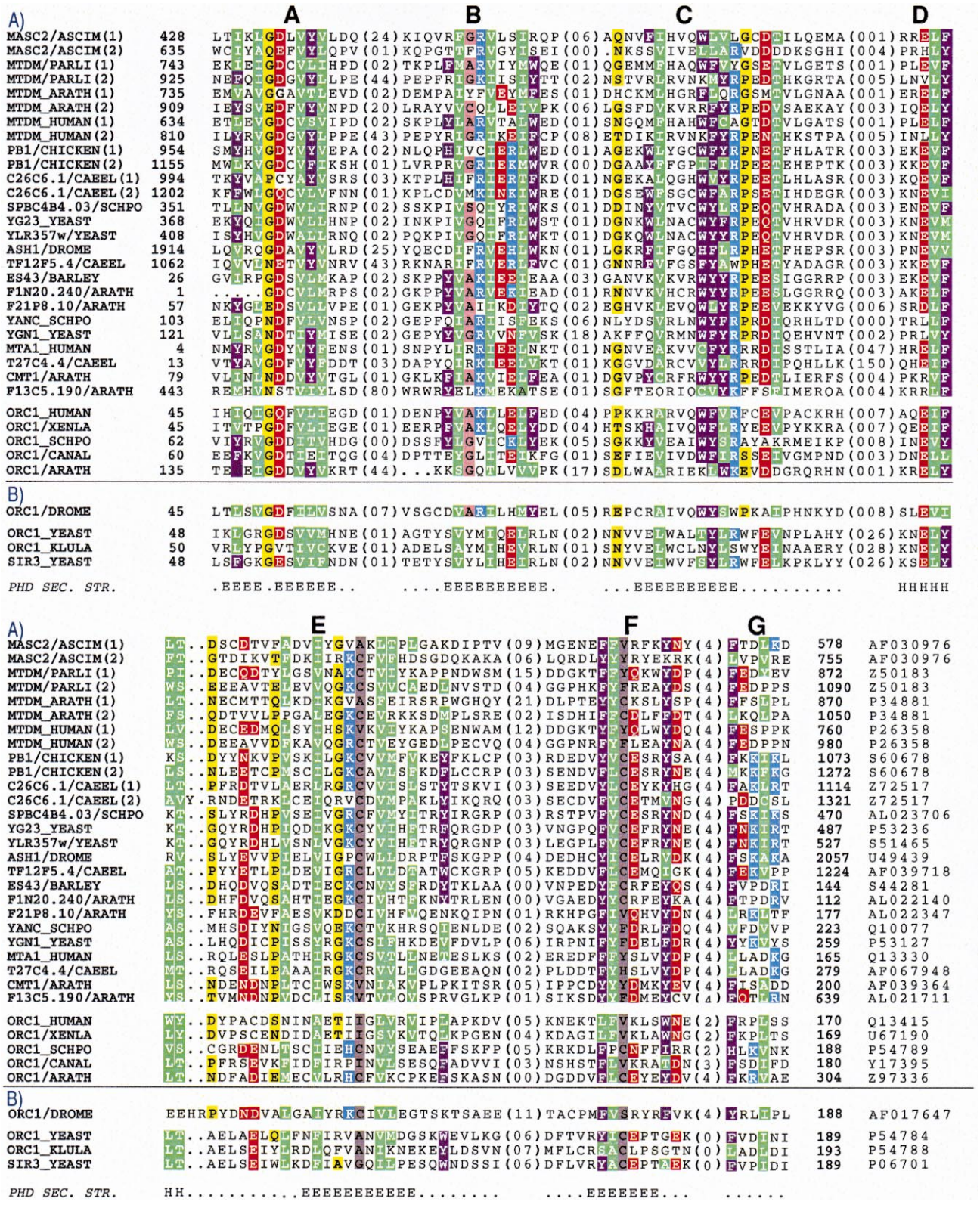
Using a combination of sensitive methods of sequence analysis such as BLAST2, PSI-BLAST [7] and bidimensional hydrophobic cluster analysis (HCA) [9,10], we found that the duplicated domain of PB1 was in fact larger than reported, due to the presence of two highly conserved motifs (designated A and B) upstream of motif C (the previously reported NH_2-terminal limit of the BAH module) (Fig. 1). The redefined BAH domain has a length of $\sim$120–140 amino acids (minimal length), with clear limits, especially when considering the whole BAH family (see below). By using the BLAST 2 sequences program (version 2.0.6, scoring matrix blosum 62), the similarity between the two BAH domains of PB1 was found to be significant (E value $= 4 \times 10^{-17}$, 37% identity over 115 aa).

Having defined new limits to BAH domains in PB1, we observed by HCA that these domains are also present in a duplicated form in MTDM, with the presence of amino acids stretches of variable length (24–44 aa) between motifs A and B of the second BAH domain (Fig. 1). Using the first module as query, the MTDM duplication was found to be significant relative to the whole BAH family in an iterative PSI-BLAST search (see below; E value 1×10^{-13} following convergence after six iterations).

We next used the two BAH domains of PB1 (aa 936–1278) as a query sequence in iterative PSI-BLAST searches on the NCBI non-redundant (NR) database (343 871 sequences; BLAST2 version 2.0.6, cutoff E value 0.001, scoring matrix blosum 62). Following convergence after four iterations, significant hits were obtained with proteins of panel A of Fig. 1

*Corresponding author: Fax: (33) (1) 4427 3785
E-mail: callebau@lmcp.jussieu.fr

PII: S0014-5793(99)00132-5



{*E* values ranging from 4×10^{-87} (MTDM_HUMAN; 16% identities over 373 residues-duplicated BAH domains) to 3×10^{-10} (MTA1_HUMAN; 21% over 157 residues-single BAH domain)}. The limits of the BAH domain can be clearly defined as they are either (i) preceded or followed by non-globular regions, or (ii) found at the very NH_2-terminus of some proteins (e.g. the metastasis-associated protein MTA1), or (iii) immediately followed by well defined modules such as PHD fingers (e.g. receptor-like protein ES43) (Fig. 2). Insertions of variable length may be observed between conserved blocks, corresponding to either (i) mainly non-globular regions (e.g. 150 aa between blocks C and D of *Caenorhabditis*

Fig. 1. Multiple alignment of the conserved motifs of the BAH domain. The full-length of the BAH domain and its limits were defined using hydrophobic cluster analysis (HCA). The alignment was constructed on the basis of PSI-BLAST results and refined using HCA. Only non-orthologous or highly divergent orthologous sequences are shown. The positions of the NH2- and COOH-termini and distances between aligned blocks are indicated in numbers of residues. Color shading indicates residues which are particularly conserved in the BAH family: green for bulky hydrophobic residues (V,I,L,M,F,Y,W), red for acidic residues (E,D,Q,N), blue for basic residues (K,R,H), purple for aromatic residues (F,Y,W), pink for small residues (G,A,C,T), orange for turn promoting residues (P,G,D,N,S). Other colors are used for specific properties (e.g. gray for the columns where conserved cysteine residues are frequently substituted by hydrophobic residues). A,C,T,S which can substitute for bulky hydrophobic amino acids, H which can substitute for aromatic amino acids, E and Q which can substitute for turn promoting residues are shaded accordingly. Secondary structure predictions using the PHD server [11] are found below the alignment (E: β strand, H: α helix). This figure has been prepared using the ESPript software (P. Gouet et al., in preparation). A comprehensive multiple alignment of BAH domain sequences is available at GenBank, EMBL, SwissProt or PIR accession numbers are given in the far right column. ARATH: *Arabidopsis thaliana*; ASCIM: *Ascolobus immersus*; BARLEY: *Hordeum vulgare*; CAEEL: *Caenorhabditis elegans*; CANAL: *Candida albicans*; CHICKEN: *Gallus gallus*; DROME: *Drosphila melanogaster*; HUMAN: *Homo sapiens*; KLULA: *Kluyveromyces lactis*; PARLI: *Paracentrotus lividus*; SCHPO: *Schizosaccharomyces pombe*; XENLA: *Xenopus laevis*; YEAST: *Saccharomyces cerevisiae*.
←

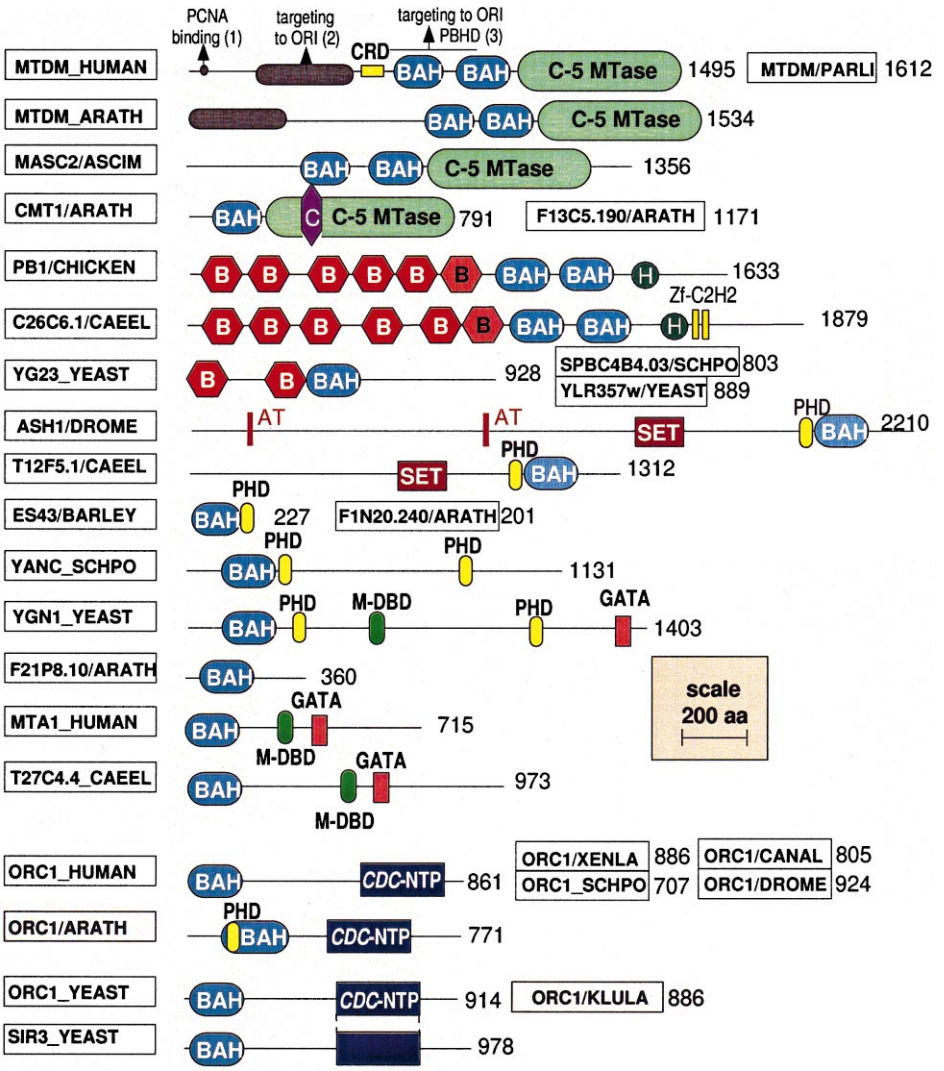


Fig. 2. Modular architecture of BAH-containing proteins. The protein sequences are drawn to scale. Abbreviations correspond to those of Fig. 1. Domains are designated as follows: CRD: cysteine-rich domain shared by PCM1 (protein containing methyl-CPG binding domain); mammalian HRX proteins and animal DNA-C5-methyltransferases [36]; C-5-MTase: DNA-C5-methyltransferase; C: chromo domain [21,22]; B: bromo domain [25]; H: HMG (high mobility group) box [37]; zf-C2H2: zinc finger, C2H2 type [38]; AT: AT-hook motif, a motif from HMG proteins involved in the binding to the AT-rich regions of double-stranded DNA minor groove [39,40]; SET: SET domain [28]; PHD: PHD finger [26]; M-DBD: Myb-like DNA binding domain [41]; GATA: GATA zinc finger [42]; *CDC*-NTP: *CDC* nucleotide triphosphate (NTP)-binding motif (sharing similarities with Cdc6p and Cdc18p). Although the carboxyl-terminal domain of Sir3p is clearly related to the *CDC*-NTP domain of Orc1p, the sequences which would be critical for nucleotide binding are not conserved (*CDC*-NTP-like). The three regions of mouse Dnmt1 (reported on the MTDM_HUMAN sequence) shown to be involved in the targeting of the protein to origins of replication are designated with arrows (1: [3]; 2: [2]; 3: [5]). A bromo-like domain of PB1, downstream of the five bromo domains of the chicken PB1 and *C. elegans* C26.6.1 sequences but not mentioned in [6] and [25], is shaded differently. Boxes located at the right of some proteins correspond to identifiers of proteins in Fig. 1 with a similar modular architecture.

elegans T27C4.4 protein), or to (ii) regular secondary structures (e.g. between blocks A and B in the second BAH domain of the MTDM sequences and in *Arabidopsis thaliana* F13C5.190 protein; between blocks C and D of human MTA1), or (iii) to well defined domains (e.g. PHD finger between blocks A and B of *A. thaliana* replication control protein homolog).

Interestingly, a BAH module was identified in the NH_2-terminus of Orc1 proteins (Orc1p, origin recognition complex, subunit 1) from *Schizosaccharomyces pombe* (*E* value 2×10^{-26}; 21% identity over 170 residues) and human (*E* value 1×10^{18}; 14% identity over 127 residues), both domains hitherto considered unrelated to each other and to Orc1 proteins of other species [12]. Significant *E* values were also found with the Orc1 protein sequences of *A. thaliana*, *Xenopus laevis* and *Candida albicans* (*E* values 6×10^{-26}, 9×10^{-19}, 7×10^{-18}, respectively). Furthermore, PSI-BLAST hits (just below the threshold of significance) were also observed within the NH_2-termini of Orc1 proteins from *Drosophila melanogaster* and *Saccharomyces cerevisiae*. This last sequence is otherwise related to the NH_2-termini (NTD) of *Kluyveromyces lactis* Orc1 protein and of *S. cerevisiae* Sir3 protein [12]. The presence of BAH domains in all Orc1 protein sequences, strongly supported by the examination of their HCA plots (data not shown), was further assessed by performing a HMM database search using the multiple alignment of the BAH domains of Fig. 1A and including intervening sequences between conserved blocks. *D. melanogaster* and *S. cerevisiae* Orc1p as well as *S. cerevisiae* Sir3p sequences were now significantly detected (scores of 26.8, 29.3 and 25.5, respectively). It is worth noting that the Orc1p sequence of *Candida albicans* appears to be a key linker between human and *S. cerevisiae* Orc1p BAH domains. Orc1p is one of the six polypeptides forming an oligomer, the origin replication complex (ORC), initially identified in *S. cerevisiae*. This complex binds origins of replication and directs DNA replication throughout the genome, as well as transcriptional silencing at the yeast mating-type loci *HML* and *HMR* [13]. The BAH domain is located at the NH_2-terminal domain (NTD) of Orc1p. This NTD is particularly well conserved between the *S. cerevisiae* Orc1p and the related silencing information regulator Sir3p (50% identity within the first 214 amino acids) [14], which is required to maintain transcriptional silencing at mating-type loci and telomeres [15]. The yeast Orc1p NTD appears to play a role in mating-type silencing, but not in DNA replication [14]. Interestingly, the yeast Orc1p NTD (aa 5–228) has been shown to interact with the carboxyl-terminal half of Sir1p in two-hybrid screen and in vitro binding studies, indicating that this silencing factor might be targeted to silencers by binding Orc1p, a phenomenon called 'targeted silencing' [16]. However, the comparable NTD of yeast Sir3p does not interact with Sir1p in analogous two-hybrid studies [16], and no direct interaction between this region of Sir3p and components of the silencing machinery has yet been identified [17]. Nonetheless, the Sir3p NTD has been observed to increase the frequency of telomere-proximal silencing and the mutations for the SIR3 suppressors of histone H4 and rap1 mutants all fall within this NTD [18,19]. On the other hand, the NH_2-terminal sequence of *Drosophila* Orc1 protein (aa 1–238), in which the BAH module is included, is able to bind HP1 (heterochromatin-associated protein 1) [20] which contains two copies of a small module, the chromo domain [21,22], respectively called chromo and chromo-shadow domains. Chromo domain-containing proteins such as HP1 or polycomb (PC) protein are present in inactive chromatin regions, suggesting a role for these domains in the packaging of the chromatin fiber, making it therefore transcriptionally inactive or unable to change its epigenetic state [23]. Both chromo and chromo-shadow domains of HP1 are necessary for the association with Orc1p [20]. The NH2-terminus (aa 1–413) of *X. laevis* Orc1p also interacts with HP1 [20]. These data suggest that the BAH module of Orc1p might participate in the recruitment of HP1 and associated heterochromatin factors to their target sites. However, there is no evidence for a direct interaction of BAH module with these proteins, as an adjacent sequence in Orc1p (aa 161–319), overlapping the carboxyl-terminal end of the *D. melanogaster* BAH module (blocks F and G), also interacts with HP1 [20]. It is worth noting that *Arabidopsis* chromomethylase CMT1 contains, in addition to a BAH domain, a chromo domain inserted within the catalytic domain [24] (Fig. 2). Thus, the data concerning Orc1p and the related Sir3p strongly suggest that the BAH module might play a crucial role in gene silencing and possibly transcriptional repression by binding to components of the silencing machinery and of heterochromatin.

The BAH module is frequently associated in proteins with other modules implicated in epigenetic mechanisms of gene regulation such as bromo and SET domains, as well as PHD fingers (Fig. 2). The BAH module is associated with multiple copies (2–6) of the bromo domain, a small domain present in several proteins involved in transcriptional regulation [25]. The PHD finger is a small zinc-binding motif (C_4-H-C_3) that mediates protein-protein interactions in chromatin-dependent transcriptional regulation [26]. A PHD finger is located immediately downstream of a BAH domain in the barley ES43 protein [27]. Interestingly, in *Drosophila* ASH1 protein (absent, small or homeotic disc1) which belongs to the trithorax group of activators (trx-G) [28], the BAH module is associated with a PHD finger and a SET domain. The SET (suvar 3–9 enhancer-of-zeste trithorax) domain is found in several proteins that also contribute to epigenetic mechanisms of gene regulation [29], and mutations within SET domains are related to a silencing defect [30,31]. SET domains also mediate interactions with myotubularin-type phosphatases, suggesting that these domains might connect the epigenetic regulatory machinery to signalling pathways [32]. Myb-DNA binding domains and GATA zinc-fingers are present with a BAH domain in Mta1 protein that may play a role in breast cancer invasion and metastasis [33].

Thus, using an extended sequence of the BAH domain as query, we have detected the presence of the BAH module either duplicated as in eukaryotic DNA MTases, or unique as in *A. thaliana* chromomethylase CMT1 (Fig. 2). The BAH module appears to be tightly associated with replication events, as it is involved in the targeting of mouse DNA MTase Dnmt1 to replication origins and is absent from the amino-terminal extension of Dnmt3 proteins, which are thought to correspond to de novo DNA methyltransferases [34]. The BAH protein family also contains Orc1p of all known species in addition to several proteins involved, or thought to be involved, in transcriptional regulation (Fig. 2). The BAH domain might therefore play an important role in coupling DNA methylation, replication and chromatin-mediated gene inactivation. The most direct evidence for a functional role of the

BAH domain comes from studies showing a direct interaction of *S. cerevisiae* Orc1p NTD with Sir1p [16,35]. We propose therefore that the BAH module might act as a protein-protein interacting module responsible for the targeting of this family of proteins to their sites of action.

Acknowledgements: This research was supported by the CNRS program 'Physique et Chimie du Vivant' (PCV). We thank Nicole Dalla Venezia for suggestions about DNA MTases.

References

[1] Razin, A. (1998) EMBO J. 17, 4905–4908.
[2] Leonhardt, H., Page, A.W., Weier, H.-U. and Bestor, T.H. (1992) Cell 71, 865–873.
[3] Chuang, L.S.-H., Ian, H.-I., Koh, T.-W., Ng, H.-H., Xu, G. and Li, B.F.L. (1997) Science 277, 1996–2000.
[4] Montecucco, A. et al. (1998) EMBO J. 17, 3786–3795.
[5] Liu, Y., Oakeley, E.J., Sun, L. and Jost, J.-P. (1998) Nucleic Acids Res. 26, 1038–1045.
[6] Nicolas, R.H. and Goodwin, G.H. (1996) Gene 175, 233–240.
[7] Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Nucleic Acids Res. 25, 3389–3402.
[8] Eddy, S. (1996) Curr. Opin. Struct. Biol. 6, 361–365.
[9] Gaboriaud, C., Bissery, V., Benchetrit, T. and Mornon, J.P. (1987) FEBS Lett. 224, 149–155.
[10] Callebaut, I., Labesse, G., Durand, P., Poupon, A., Canard, L., Chomilier, J., Henrissat, B. and Mornon, J.-P. (1997) Cell. Mol. Life Sci. 53, 621–645.
[11] Rost, B. and Sander, C. (1993) J. Mol. Biol. 232, 584–599.
[12] Gavin, K.A., Hidaka, M. and Stillman, B. (1995) Science 270, 1667–1671.
[13] Bell, S.P., Kobayashi, R. and Stillman, B. (1993) Science 262, 1844–1849.
[14] Bell, S.P., Mitchell, J., Leber, J., Kobayashi, R. and Stillman, B. (1995) Cell 83, 563–568.
[15] Stone, E.M. and Pillus, L. (1998) BioEssays 20, 30–40.
[16] Triolo, T. and Sternglanz, R. (1996) Nature 381, 251–253.
[17] Gotta, M., Palladino, F. and Gaser, S.M. (1998) Mol. Cell. Biol. 18, 6110–6120.
[18] Johnson, L.M., Kayne, P.S., Kahn, E.S. and Grunstein, M. (1990) Proc. Natl. Acad. Sci. USA 87, 6286–6290.
[19] Liu, C. and Lustig, A.J. (1996) Genetics 143, 81–93.
[20] Pak, D.T.S., Pflumm, M., Chesnokov, I., Huang, D.W., Kellum, R., Marr, J., Romanowski, P. and Botchan, M.R. (1997) Cell 91, 311–323.
[21] Aasland, R. and Stewart, A.F. (1995) Nucleic Acids Res. 23, 3168–3174.
[22] Koonin, E.V., Zhou, S. and Lucchesi, J.C. (1995) Nucleic Acids Res. 23, 4229–4234.
[23] Cavalli, G. and Paro, R. (1998) Curr. Opin. Cell Biol. 10, 354–360.
[24] Henikoff, S. and Comai, L. (1998) Genetics 149, 307–318.
[25] Jeanmougin, F., Wurtz, J.M., Le Douarin, B., Chambon, P. and Losson, R. (1997) Trends Biochem. Sci. 22, 151–153.
[26] Aasland, R., Gibson, T.J. and Stewart, A.F. (1995) Trends Biochem. Sci. 20, 56–58.
[27] Speulman, E. and Salamini, F. (1995) Plant Sci. 106, 91–98.
[28] Tripoulas, N., LaJeunesse, D., Gildea, J. and Shearn, A. (1996) Genetics 143, 913–928.
[29] Jenuwein, T., Laible, G., Dorn, R. and Reuter, G. (1998) Cell. Mol. Life Sci. 54, 80–93.
[30] Nislow, C., Ray, E. and Pillus, L. (1997) Cell 8, 2421–2436.
[31] Ivanova, A.V., Bonaduce, M.J., Ivanov, S.V. and Klar, A.J.S. (1998) Nature Genet. 19, 192–195.
[32] Cui, X., De Vivo, I., Slany, R., Miyamoto, A., Firestein, R. and Cleary, M.L. (1998) Nature Genet. 18, 331–337.
[33] Toh, T., Pencil, S.D. and Nicolson, G.L. (1994) J. Biol. Chem. 269, 22958–22963.
[34] Okano, M., Xie, S. and Li, E. (1998) Nature Genet. 19, 219–220.
[35] Gardner, K.A., Rine, J. and Fox, C.A. (1999) Genetics 151, 31–44.
[36] Cross, S.H., Meehan, R.R., Nan, X. and Bird, A. (1997) Nature Genet. 16, 256–259.
[37] Gehring, W.J. (1992) Trends Biochem. Sci. 17, 277–280.
[38] Bohm, S., Frishman, D. and Mewes, H.W. (1997) Nucleic Acids Res. 25, 2464–2469.
[39] Reeves, R. and Nissen, M.S. (1990) J. Biol. Chem. 265, 8573–8782.
[40] Aravind, L. and Landsman, D. (1998) Nucleic Acids Res. 26, 4413–4421.
[41] Aasland, R., Stewart, A.F. and Gibson, T.J. (1996) Trends Biochem. Sci. 21, 87–88.
[42] Omichinsky, J.G., Clore, M.G., Schaad, O., Felsenfeld, G., Trainor, C., Appela, E., Stahl, S.J. and Gronenborn, A.M. (1993) Science 261, 438–446.