The Crystal Structure of AF1521 a Protein from *Archaeoglobus fulgidus* with Homology to the Non-histone Domain of MacroH2A

Mark D. Allen, Ashley M. Buckle, Suzanne C. Cordell, Jan Löwe and Mark Bycroft*

MRC Centre for Protein Engineering and MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

MacroH2A is an unusual histone H2A variant that has an extensive C-terminal tail that comprises approximately two thirds of the protein. The C-terminal non-histone domain of macroH2A is also found in a number of other proteins and has been termed the macro domain. Here we report the crystal structure to 1.7 Å of AF1521, a protein consisting of a stand-alone macro domain from *Archaeoglobus fulgidus*. The structure has a mixed α/β fold that closely resembles the N-terminal DNA binding domain of the *Escherichia coli* leucine aminopeptidase PepA. The structure also shows some similarity to members of the P-loop family of nucleotide hydrolases.

© 2003 Elsevier Science Ltd. All rights reserved

Keywords: histone macroH2A; crystal structure; P-loop nucleotide hydrolases; RNA binding; X-chromosome inactivation

Introduction

The basic component of chromatin is the nucleosome, a complex of approximately 150 base-pairs of DNA wrapped around an octomer of the core histones H2A, H2B H3 and H4.1 The core histones contain a central histone fold domain, which is flanked by N- and C-terminal tails. Histone tails are the targets of several forms of covalent modification including acetylation, methylation, ubiquitination, ADP-ribosylation and phosphorylation. These modifications play a central role in regulating gene expression.2 They can influence nucleosome stability and thereby affect the accessibility of DNA to the transcriptional machinery and can also provide docking sites for regulatory proteins. In addition to the basic core histones a number of histone variants have been identified. These proteins have a central domain similar to the core histones but vary at their N and C termini. Their incorporation into nucleosomes offers an alternative means of modulating chromatin structure and function and variant histones have been implicated in the regulation of processes such as transcription and DNA repair.3 MacroH2A is the largest variant histone.4 It has an amino-terminal domain very similar in sequence to histone H2A followed by a large C-terminal non-histone domain which accounts for about two thirds of the protein. The human genome contains two genes that code for macroH2A histones. The MACROH2A1 gene encodes two subtypes MACROH2A1.1 and MACROH2A1.2 produced by alternative splicing. A second gene codes for MACROH2A2.5,6 These proteins appear to have a role in transcriptional silencing and are enriched in the chromatin of inactive X-chromosomes.7,8 In female mammals one of the X-chromosomes is transcriptionally inactivated to equalise the expression of X-linked genes between XY males and XX females.9 The inactivated X-chromosome has several features associated with gene silencing. Its DNA is hypermethylated within CpG islands and its chromatin is condensed and characterised by hypoacetylation of histones H3 and H4 and hypermethylation of histone H3. The inactivation process can be separated into several distinct stages. The initiation and establishment of inactivation requires the expression of a large non-coding RNA *Xist*, which coats, in cis, the chromosome that is to be inactivated.10,11 Once inactivation has been established *Xist* is not required for the maintenance of silencing. *Xist* is believed to function by recruiting proteins to the chromosome that modify chromatin.12 The role of macroH2A histones in
X-inactivation is unclear. Their incorporation into the chromatin of the inactivated X-chromosome is a late event that depends on Xist accumulation and it seems that these proteins play some role in the establishment of the inactivated state.

The C-terminal non-histone domain of macroH2A is also found alone or in multiple copies in a number of unrelated unrelated proteins and has been termed the macro domain (Figure 1). One of these proteins, BAL, has been shown to be a highly expressed risk factor in some aggressive lymphomas. Macro domains are also found in the non-structural proteins of several types of ssRNA viruses including rubella virus and hepatitis E virus. The presence of the macro domain in macroH2A histones and in proteins containing DNA and RNA binding domains would suggest a role in nucleic acid recognition. In other proteins, however, it is linked to domains that are known to have a role in protein–protein (WWE) or protein–lipid (Sec14) interaction.

The macro domain is also found on its own in a family of largely uncharacterised proteins from bacteria, archaea and eukaryotes. The wide distribution of this protein family suggests that it is involved in an important and ubiquitous cellular process. The only functional information available on this group of proteins has come from a genome wide screen of yeast proteins for biochemical activities related to tRNA splicing. In this study YBR022Wp, a protein consisting of an isolated macro domain, was shown to have ADP-ribose 1'-phosphate (Appr-1'p) processing activity. Appr-1'p is produced as a by product of the NAD\(^+\) dependent reaction that removes the 2' phosphate group from the 3', 5'-phosphodiester-2'-phosphomonoester linkage that is produced during tRNA splicing in yeast. Whether this activity is biologically significant and what relevance it has to the function of macro domains in other proteins has not been established. This result does, however, suggest that this family of proteins may act as ADP-ribose phosphoesterases.

To further our understanding of this family of proteins we have determined the structure of AF1521, a macro domain containing protein from the thermophile Archaeoglobus fulgidus. The structure reveals homology to the P-loop family of nucleotide hydrolases. The AF1521 structure is also very similar to the N-terminal domain of Escherichia coli PepA, a domain known to bind to DNA demonstrating that proteins with this fold are also capable of interacting with nucleic acids.

**Results and Discussion**

**Overall structure**

The structure of A. fulgidus AF1521 was solved using phases derived from a three-wavelength MAD data set collected from a single crystal of seleno-methionine-substituted protein (Table 1). The structure of the seleno-methionine derivative was then used to determine the structure of the non-substituted protein. The final model of the native protein containing all 192 residues of AF1521 was refined against 1.7 Å resolution data.

---

**Figure 1.** A schematic representation of proteins containing macro domains. H2A, histone H2A domain; DUF144, a domain described in the Pfam database involved in phage tail assembly; SNF2, SNF2 helicase like domain; AAA, AAA ATPase domain; SEC14, lipid-binding domain found in SEC14p and other proteins; HLH, helix-loop-helix DNA binding domain; ZF, zinc finger; RRM, RNA recognition motif; WWE, a protein–protein interaction domain containing conserved tryptophan and glutamate residues; PARP, poly-ADP-ribose polymerase domain. This domain was detected using the program Superfamily. The Figure was adapted from the SMART database.
Table 1. Summary of statistics for data collection

<table>
<thead>
<tr>
<th></th>
<th>Peak</th>
<th>Inflection</th>
<th>High-energy remote</th>
<th>Native</th>
</tr>
</thead>
<tbody>
<tr>
<td>Space group</td>
<td>P2,2,2</td>
<td>P2,2,2</td>
<td>P2,2,2</td>
<td>C2</td>
</tr>
<tr>
<td>Wavelength (Å)</td>
<td>0.9797</td>
<td>0.9799</td>
<td>0.9393</td>
<td>1.00</td>
</tr>
<tr>
<td>Resolution range (Å)</td>
<td>33.5–2.5</td>
<td>33.5–2.5</td>
<td>33.5–2.5</td>
<td>17.1–1.7</td>
</tr>
<tr>
<td>Unique reflections</td>
<td>7251</td>
<td>7251</td>
<td>7251</td>
<td>40692</td>
</tr>
<tr>
<td>R_m()</td>
<td>21.0(8.5)</td>
<td>21.6(8.5)</td>
<td>23.6(10.9)</td>
<td>17.6(3.0)</td>
</tr>
<tr>
<td>R_F()</td>
<td>0.060</td>
<td>0.058</td>
<td>0.055</td>
<td>0.047</td>
</tr>
<tr>
<td>Multiplicity^c</td>
<td>7.0</td>
<td>6.9</td>
<td>7.0</td>
<td>3.4</td>
</tr>
</tbody>
</table>

The first three columns refer to the Se derivative and the fourth column refers to the native protein.

^a Significance to noise ratio of intensities, highest resolution bin in brackets.

^b R_m = Σ_i [I(h, i) − I(0)] / Σ_i [I(h, i)] where I(h, i) are symmetry related intensities and I(0) is the mean intensity of the reflection with unique index h.

^c Multiplicity for unique reflections, for MAD datasets I(+) and I(−) are kept separate. Correlation coefficients of anomalous differences at different wavelengths for the MAD experiment: Peak versus Inflection, 0.55; Peak versus High-energy remote, 0.70; Inflection versus High-energy remote, 0.46.

(Table 2) and has an R_work of 20.3% (R_free = 22.7%). AF1521 is a compact, single-domain protein with a mixed α/β structure. It contains a single seven-stranded mixed β-sheet with strand order 1276354 and five α-helices (Figure 2). The two strands at the edge of the β-sheet are antiparallel to the others. The β-sheet is sandwiched between four helices, three packing onto one face and one onto the other. The remaining α-helix packs onto the edge of the sheet.

A structure-based alignment of representative macro domains is shown in Figure 3. Several macro domain only proteins are significantly shorter than AF1521. Based on this alignment it would appear that some of these proteins lack either the first strand of the β-sheet or the C-terminal α-helix. Both of these elements of secondary structure are at the periphery of the fold and their absence would not be expected to produce a substantial alteration in the structure. Based on the sequence alignment it would also appear that in some proteins helices 2 and 3 are shorter than in AF1521 and that most macro domains lack the Σ_1,0 helix in the turn between strands 4 and 5. With these exceptions deletions

and insertions are restricted to the loops between elements of secondary structure and it is likely that the majority of macro domains have a structure very similar to that of AF1521. It has been noted that the residues at the N terminus of the macro domain of macroH2A1 could form a leucine zipper. In AF1521 these residues form strands 1 and 2 of the β-sheet and a Σ_3,0 helix. Given the degree of sequence similarity between macroH2A1 and AF1521 it is likely that the equivalent regions of macroH2A1 adopt a similar structure.

A putative active site

The AF1521 structure has a deep cleft in the protein surface, a feature typical of an enzyme active site (Figure 4). AF1521 is a member of a family of proteins conserved in bacteria, archaea and eukaryotes that have a high degree of sequence similarity (Figure 3). Most of the conserved residues are located within this cleft. Many of them are solvent exposed, suggesting that they are functionally important and this region of the protein is a good candidate for the active site. Part of this cleft is occupied by a molecule of MES buffer. The residues surrounding the buffer molecule are particularly highly conserved suggesting that it is mimicking a natural ligand. These include a glycine rich sequence that forms part of the loop between strand 3 and helix 1, two asparagine residues in strand 3 and residues in the loop between strand 6 and helix 6 (Figure 4). Interestingly the alternative splicing of the MACROH2A1 mRNA alters the protein sequence between the end of strand 2 and the start of helix 1 (Figure 3). As this region of the protein forms part of the putative active site the two isoforms are likely to have significantly different properties.

AF1521 is a divergent member of the P-loop-containing nucleotide triphosphate hydrolase superfamily

The central portion of AF1521, including the putative active site, shows some structural
similarity to members of the P-loop-containing superfamily of nucleotide triphosphate hydrolases including the hexamerisation domain of N-ethylmaleimide-Sensitive Fusion Protein, RecA, helicase core domain and F1 ATPase. Given that a macro domain containing protein has been shown to hydrolyse Appr-1p the structural similarity to this protein family is of particular interest. The strongest similarity (Dali Z-score 5.2) is with CobA, an ATP:co(I)rrinoid adenosyltransferase which transfers an adenosyl moiety from MgATP to a variety of co(I)rrinoid substrates (Figure 5).

Both CobA and AF1521 contain five strands of parallel β-sheet hydrogen bonded in the order 32451. Although the loops and helices connecting them differ these strands can be superimposed with a root mean squared deviation of 1.3 Å for 30 pairs of aligned Cα atoms. The P-loop-containing nucleotide triphosphate hydrolases are an important and well-studied family of proteins and the residues involved in ligand binding and catalysis have been characterized in detail. The active site in these proteins is in the same part of the structure as the putative active site identified for AF1521. All of the blocks of amino acids conserved within the AF1521 family correspond to residues that are known to be functionally important in the P-loop-containing nucleotide triphosphate hydrolases. The combination of structural similarity and correspondence of active sites taken in aggregate strongly suggest that AF1521 is evolutionarily related to the P-loop-containing nucleotide triphosphate hydrolases. The available biochemical data suggest that macro domains hydrolyze ADP-ribose derivatives rather than nucleotide triphosphates. The P-loop containing nucleotide triphosphate hydrolases all interact with the triphosphate moiety of their substrates using a P-loop situated between a strand of parallel β-sheet and an α-helix that typically adopts the sequence pattern GxxxxG[KST] sometimes referred to as the Walker A motif. In AF1521 the equivalent loop is between the end of strand 2 and a 3_10 helix and contains a highly conserved GDIT motif. In AF1521 this region probably interacts with the diphosphate group of the substrate and the differences in structure and sequence in this part of the protein presumably reflects this change in specificity. CobA and other P-loop containing nucleotide triphosphate hydrolases bind to substrates using a pocket formed by loops and helices at the C-terminal end of the β-sheet (Figure 5).

The equivalent part of the AF1521 structure contains the highly conserved MES binding pocket suggesting that the buffer molecule is occupying a substrate-binding site. The group of conserved residues at the end of strand 5 in AF1521 is in the same position as the Walker B motif of the P-loop containing nucleotide triphosphate hydrolases. This sequence motif is involved in catalysis and it is likely that the corresponding residues in the AF1521 family have a similar role. Cleavage of the phosphodiester bond of the substrate therefore probably occurs at the junction between the MES binding pocket and the rest of the active site cleft.

AF1521 has the same fold as the DNA binding domain of PepA

When the AF1521 structure was searched against known structures using the program Dali the strongest structural similarity detected was to the N-terminal domain of leucine aminopeptidase (Z = 10.2). Leucine aminopeptidases catalyse the hydrolysis of amino acids from the amino terminus of polypeptide chains and are widely distributed in plants, animals and bacteria. The AF1521 structure is most similar to the N-terminal domain of

Figure 2. Overall structure of AF1521. The Figure was prepared using the program PyMOL.
E. coli PepA aminopeptidase. Although there is no significant sequence similarity between the two proteins, they can be superimposed with a root mean squared deviation of 1.6 Å for 106 pairs of aligned Cα atoms (Figure 6). This degree of structural similarity indicates that these proteins are descended from a common ancestor and would be classified as being members of the same homologous family in the SCOP database. The aminopeptidase active site is located entirely within the C-terminal domain. Mutagenesis studies have shown that the N-terminal domain is mainly responsible for DNA binding. It seems likely that the N-terminal domain is the site of DNA binding.

The structure of a Macro Domain

Figure 3. Structure based sequence alignment of selected macro domains. Residues conserved in the AF1521 family are highlighted. The protein name is followed by the species abbreviation. The limits of the domain are indicated by the gene bank GI identification number. The residues altered by the alternative splicing of the MACROH2A1 gene are underlined. YMDB_ECOLI and LRP16 are AF1521 orthologues from E. coli and mouse. Q96F23 is a human protein that is a member of a second family of macro domain only proteins found in eukaryotes. GDAP2 is a human protein with a N-terminal macro domain and a C-terminal domain similar in sequence to the lipid-binding domain of the S. cerevisiae protein Sec14p. Similar proteins are found in Drosophila melanogaster and Arabidopsis thaliana. Q9CF7 is a mouse protein with a N-terminal SNF2 like helicase domain and a C-terminal macro domain. Related proteins are found in Homo sapiens and in a variety of plants. The other proteins are referred to in the text.
The domain of PepA is derived from a fusion between a protein with peptidase activity and a macro domain containing protein that subsequently evolved a DNA binding activity.

**Sequence variation in macro domains**

The high degree of sequence similarity within the AF1521 family, particularly for residues that are likely to be involved in catalysis or substrate binding, suggests that these proteins are all orthologous that carry out the same ubiquitous cellular function.

The sequences of the macro domains of eukaryotic and viral proteins, however, diverge significantly from that of AF1521 particularly in
the putative active site suggesting that they do not retain the function of the parent protein family (Figure 3). This sequence variation may reflect a change in substrate specificity, as the residues that are probably involved in substrate binding are especially variable. In mammalian proteins the macro domain is often found in association with poly-ADP-ribose polymerase domains (Figure 1) and it is possible that the macro domains in these proteins are regulating protein–ADP ribosylation.15 This post-translational modification is catalyzed by ADP-ribosyl transferases that attach ADP-ribose to target proteins using NAD⁺ as a precursor. Protein–ADP ribosylation has been shown to play a role in DNA repair and replication, modulation of chromatin structure, and apoptosis.10 The sequences of some macro domains, however, vary so much from that of the AF1521 family that it is not clear if they still retain catalytic activity. It is possible that in these proteins, as is the case with the PepA N-terminal domain, the macro domain functions as a nucleic acid binding module.

Implications for the role of macroH2A histones in chromosome silencing

The recruitment of macroH2A to the inactivated X-chromosome is dependent on the accumulation of the non-coding RNA Xist and it has been suggested that there may be a direct interaction between the protein and the RNA mediated by the histone's C-terminal macro domain.14 The finding that a related protein can bind to nucleic acids supports this proposal. The role of macroH2A could simply be to consolidate the binding of Xist to the inactive X-chromosome. Alternatively the macro domain could also be acting as a DNA binding domain and may alter chromatin structure by interacting with linker DNA performing a role analogous to that of the N-terminal domain of PepA. It is also possible that the macro domain of macroH2A is involved in the chemical modification of chromatin. There is growing evidence that poly-ADP-ribosylation has a role in transcriptional regulation via the modulation of chromatin structure.11 Poly-ADP-ribosylation of linker histones has, for example, been shown to induce chromatin decondensation and protect CpG dinucleotides from full methylation in genomic DNA.32 The macro domain of macroH2A could therefore be acting as an ADP-ribose phosphoesterase. Poly-ADP-ribosylation is normally reversible. The enzyme poly-ADP-ribose glycohydrolase can remove poly-ADP-ribose chains leaving the protein free for future modification.33 It is possible that the macro domain of macroH2A cleaves poly-ADP-ribose chains in a way that blocks subsequent modification. This could then act as an epigenetic marker that helps to maintain the inactive state even in the absence of the variant histone. Although no link between X-chromosome inactivation and poly-ADP-ribosylation has yet been established this is a possibility worth exploring.

Conclusions

Both structural and biochemical evidence now strongly suggest that AF1521 and related proteins are divergent members of the P-loop containing superfamily of phosphoesterases that act on ADP ribose derivatives. The structural similarity to a known DNA binding domain demonstrates that this fold is also capable of interacting with nucleic acids. The ability of this scaffold to fulfil both a catalytic and binding role may explain the wide distribution of the macro domain. The structure described here can now be used to design experiments to help to further elucidate the role of this important module.

Materials and Methods

The gene coding for AF1521 protein was amplified from A. fulgidus DNA using PCR amplification and cloned into a pRSETA (Invitrogen) derivative that enables proteins to be expressed with a thrombin cleavable N-terminal hexa-histidine tag. The plasmid was transformed into C41(DE3) E. coli cells.34 Cells were grown at 37 °C in L-broth to mid log phase and induced with 1 mM IPTG. The temperature was then reduced to 25 °C and the cells were grown for a further 16 hours. Cells were lysed by sonication and the fusion protein was purified using a Ni-NTA Superflow affinity column. Following cleavage with thrombin (four hours at 30 °C) AF1521 was purified by gel filtration using a Superdex 75 HR column (Amersham Pharmacia) equilibrated in 50 mM potassium phosphate buffer (pH 8.0), 50 mM NaCl. The protein eluted from this column as a monomer. The seleno-methionine-substituted protein was prepared in an identical manner except that the cells were grown in M9 minimal media supplemented with seleno-methionine.

Crystallization and data collection

Native crystals were grown using the sitting drop vapour diffusion technique using 18% (w/v) PEG 5000 MME, 0.1 M MES pH 6.5, 0.2 M ammonium sulphate, and 25 mM CdCl₂ as the crystallization solution. Drops composed of 5 μl of protein at 10 mg/ml and 5 μl of crystallization solution were equilibrated for a minimum of three days at 17 °C. SeMet-substituted crystals were grown in the same manner as for the native protein but with 18% PEG 5000 MME, 0.1 M MES pH 6.5, 0.2 M ammonium sulphate, and 25 mM MnCl₂ for the crystallization solution, and additionally 5 mM DTT in the drop.

Crystals were flash-cooled in liquid nitrogen in the mother liquor with 20% PEG 200. The native protein crystals belong to space group C2 and have cell dimensions a = 119.93 Å, b = 55.72 Å, c = 62.74 Å, β = 115.15°. The SeMet-substituted crystals belong to space group P2₁2₁2 and have cell dimensions a = 62.85 Å, b = 56.1 Å, c = 56.1 Å. The native dataset was collected at beamline 14-112, Daresbury, UK and the
MAD datasets were collected at ID14-4 ESRF, Grenoble. X-ray diffraction data were indexed and integrated using the MOSFLM package\(^{39}\) and were further processed using the CCP4 package.

**Structural determination and refinement**

An initial 2.5 Å MAD density map was generated by locating five selenium sites in the data sets Peak, Inflection, High-energy remote using the program SOLVE\(^{40}\), which was also used to calculate phases. RESOLVE\(^{41}\) was used for solvent flattening, assuming a 48% solvent content. The structure was built using MAIN\(^{38}\) and refined using CNS\(^{37}\) with Engh and Huber stereochemical parameters.\(^{42}\) CNS was used for molecular replacement to determine the structure of the native protein in the C2 crystals. Details of the final model are summarized in Table 2.

**References**


Edited by R. Huber

(Received 2 December 2002; received in revised form 5 April 2003; accepted 7 April 2003)