Histone-DNA assemblies in archaea : shaping the genome on the edge of life

Henneman, B.

CHAPTER 2

Structure and function of archaeal histones

Part of this chapter is based on:

Bram Henneman, Clara van Emmerik, Hugo van Ingen & Remus T. Dame, (2018) *Structure and function of archaeal histones*. PLoS Genetics. doi: 10.1371/journal.pgen.1007582

Abstract

The genomes of all organisms throughout the tree of life are compacted and organized in chromatin by association of chromatin proteins. Eukaryotic genomes encode histones, which are assembled on the genome into octamers, yielding nucleosomes. Post-translational modifications of the histones, which occur mostly on their N-terminal tails, define the functional state of chromatin. Like eukaryotes, most archaeal genomes encode histones, which are believed to be involved in the compaction and organization of their genomes. Instead of discrete multimers, *in vivo* data suggest assembly of "nucleosomes" of variable size, consisting of multiples of dimers, which are able to induce repression of transcription. Based on these data and a model derived from X-ray crystallography, it was recently proposed that archaeal histones assemble on DNA into "endless" hypernucleosomes. In this chapter, we analyse archaeal histones from all archaeal domains and phyla. We identify archaeal histones differing from the consensus, which are expected to be unable to assemble into hypernucleosomes. Finally, we identify atypical archaeal histones with short N- or C-terminal extensions and C-terminal tails similar to the tails of eukaryotic histones, which are subject to post-translational modification. Based on the expected characteristics of these archaeal histones, we discuss possibilities of involvement of histones in archaeal transcription regulation.

Introduction

Architectural chromatin proteins are found in every domain of life. Bacteria express DNA-bending and DNA-bridging proteins, such as histone-like protein from *Escherichia coli* strain U93 (HU) and histone-like nucleoid-structuring protein (H-NS), to structure and functionally organize the genome and to regulate genome activity (11, 12). In eukaryotes and most archaeal lineages, histones are responsible for packaging and compaction of the DNA (**TABLE 2.1**). The histones found in archaea are widespread throughout the domain but are absent in some lineages, notably *Candidatus* (*Ca.*) Parvarchaeota and *Ca.* Marsarchaeota, and most Crenarchaeota and Thermoplasmata. They have the same histone fold as eukaryotic histones, but N-terminal histone tails have not been identified. Linker histones, homologous to eukaryotic H1, have not been found. Archaeal histones exist as dimers in solution, which have been shown to bend DNA (160, 193). These histone dimers can be homodimeric or heterodimeric (194), as many archaeal species express, or at least encode, more than one histone variant. In *Methanothermus fervidus* (class Methanobacteria), two histone variants are expressed at different levels and ratios at different growth phases, suggesting a distinct function for both proteins (193). In addition to binding as dimers, archaeal histones have been reported *in vivo* and *in vitro* to bind DNA as tetramers (160, 162, 195), wrapping the DNA once. However, micrococcal nuclease (MNase) digestion patterns of *Thermococcus kodakarensis* (class Thermococci) chromatin suggest that histone–DNA complexes consist of discrete multiples of a dimeric histone subunit (i.e., not limited to dimers and tetramers) *in vivo* without obvious dependence on the DNA sequence (196). Based on the latter observations, it was proposed that histone dimers multimerize and wrap DNA into a filament of variable length (140, 196). The crystallography study of Luger and coworkers on histone HMfB from *M. fervidus* indicates that these histones assemble into an endless left-handed rod *in vitro*, which we propose to call a "hypernucleosome" (**FIGURE 2.1**). Note that these complexes were assembled on SELEX-optimized DNA previously shown to favor tetrameric nucleosome assembly (197). The number of wraps in the hypernucleosome, which is the number of times that the DNA is bent by 360° around the histone multimer, scales linearly with the number of histone subunits, resulting in a tight packaging of DNA. The authors also provide evidence that mutation-directed perturbation of hypernucleosome function *in vivo* alters the response to nutrient change in *T. kodakarensis*, suggesting a role in transcription. Both eukaryotes and archaea encode histone proteins, which seem to be involved in response to environmental cues by their involvement in transcription regulation.

Table 2.1: phylogenetic subdivision of the archaeal domain

Superphylum	Phylum	Class	Histones
-	**Euryarchaeota**	Archaeoglobi	Yes
		Hadesarchaea	Yes
		Halobacteria	Yes
		Hydrothermarchaeota	Yes
		Methanobacteria	Yes
		Methanococci	Yes
		Methanomicrobia	Yes
		Methanonatronarchaeia	Yes
		Methanopyri*	Yes
		Theionarchaea	Yes
		Thermococci	Yes
		Thermoplasmata*	Yes
DPANN	*Ca.* Aenigmarchaeota		Yes
	Ca. Altiarchaeota		Yes
	Ca. Diapherotrites*		Yes
	Ca. Huberarchaeota		Yes
	Ca. Micrarchaeota		Yes
	Nanoarchaeota		Yes
	Ca. Nanohaloarchaeota		Yes
	Ca. Pacearchaeota*		Yes
	Ca. Parvarchaeota		No
	Ca. Woesearchaeota		Yes
TACK	*Ca.* Bathyarchaeota		Yes
	Crenarchaeota*		Yes
	Ca. Geothermarchaeota		Yes
	Ca. Korarchaeota		Yes
	Ca. Marsarchaeota		No
	Thaumarchaeota		Yes
	Ca. Verstraetearchaeota		Yes
Asgard Archaea	*Ca.* Heimdallarchaeota		Yes
	Ca. Lokiarchaeota		Yes
	Ca. Odinarchaeota		Yes
	Ca. Thorarchaeota		Yes

Division of archaea in superphyla and phyla, including the euryarchaeal classes. Presence or absence of histone-coding genes on the genome of at least one of the members of the phyla and classes have been indicated. *) of these groups, histones are found in a minority of sequenced genomes.

Histones are found in most newly discovered archaea

With the widespread use of metagenomic sequencing, entire new branches within the archaeal domain have been discovered. Genomes of the recently discovered archaeal superphylum Asgard archaea and phyla *Ca.* Bathyarchaeota, *Ca.* Woesearchaeota, *Ca.* Pacearchaeota, *Ca.* Aenigmarchaeota, *Ca.* Diapherotrites, *Ca.* Huberarchaeota, *Ca.* Verstraetearchaeota, *Ca.* Geothermarchaeota and *Ca.* Micrarchaeota encode histones (20-24, 26, 28, 29, 31, 39) (TABLE 2.1). With the publication of the genome sequences of these organisms, we were able to scrutinize the sequence divergence of histones by comparing sequences of histones from archaea throughout the domain (FIGURE 2.2). The selection of histones shown here is based on the presence of histone-coding genes in different phyla. Since many of those phyla were only discovered in the last five years, our selection includes a relatively large number of histones that have not yet been studied *in vivo* or *in vitro*.

We found that in *Methanosphaera* sp. SHI613 (class Methanobacteria), 11 different histones are encoded, which is the highest number of histones found in

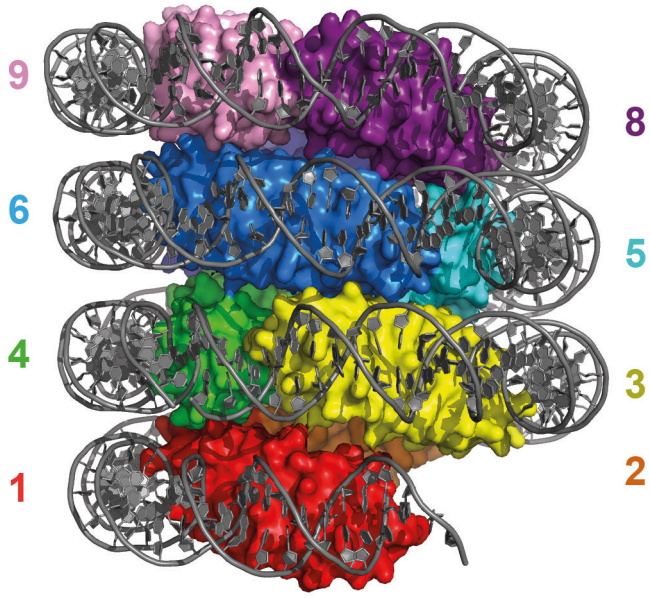


Figure 2.1: Overview of the hypernucleosome structure. HMfB dimers stack to form a continuous, central protein core that wraps the DNA in a left-handed superhelix. Nine HMfB dimers are shown, each dimer in surface mode and in rainbow colors. Numbering indicates position of the nine histone dimers; note that dimer 5 and 6 occlude the view of dimer 7. DNA is in gray and shown as cartoon. Image generated using Protein Data Bank (PDB) entry 5T5K (197).

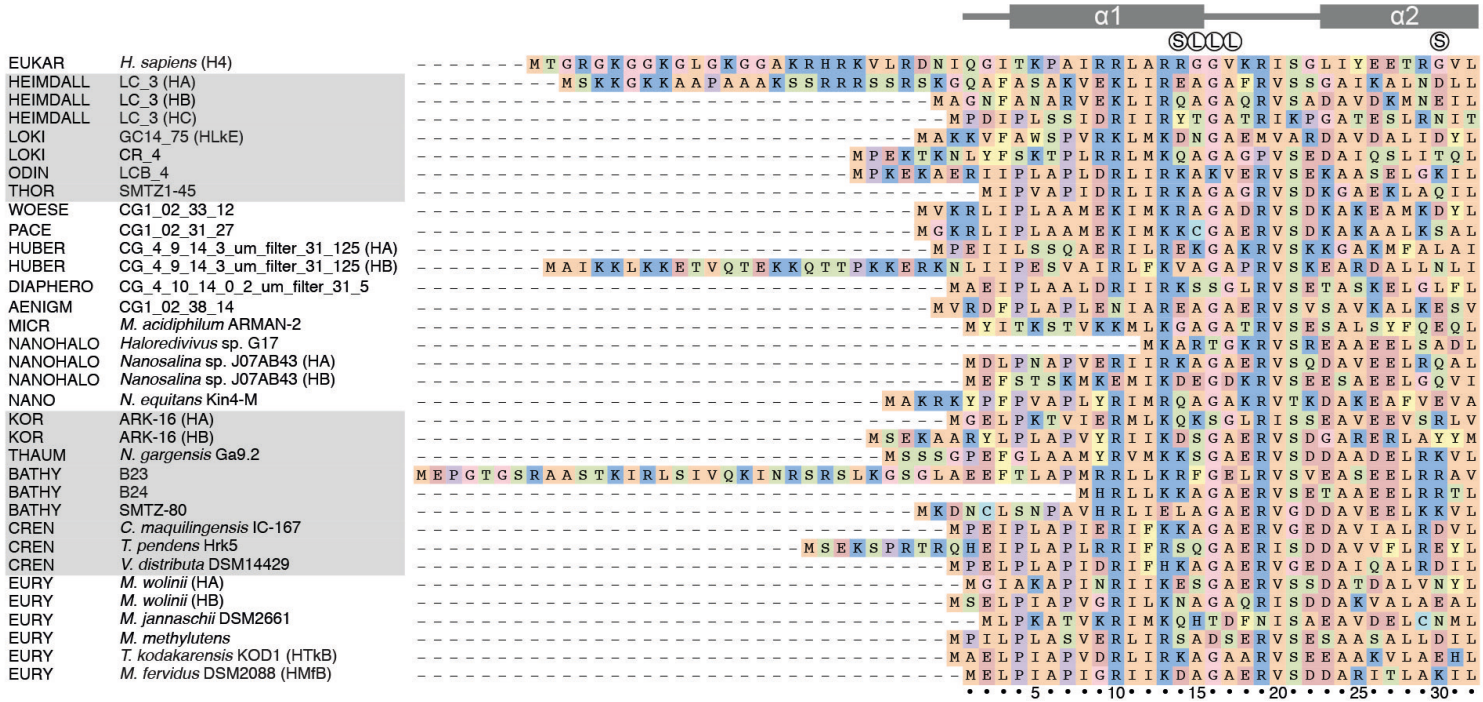
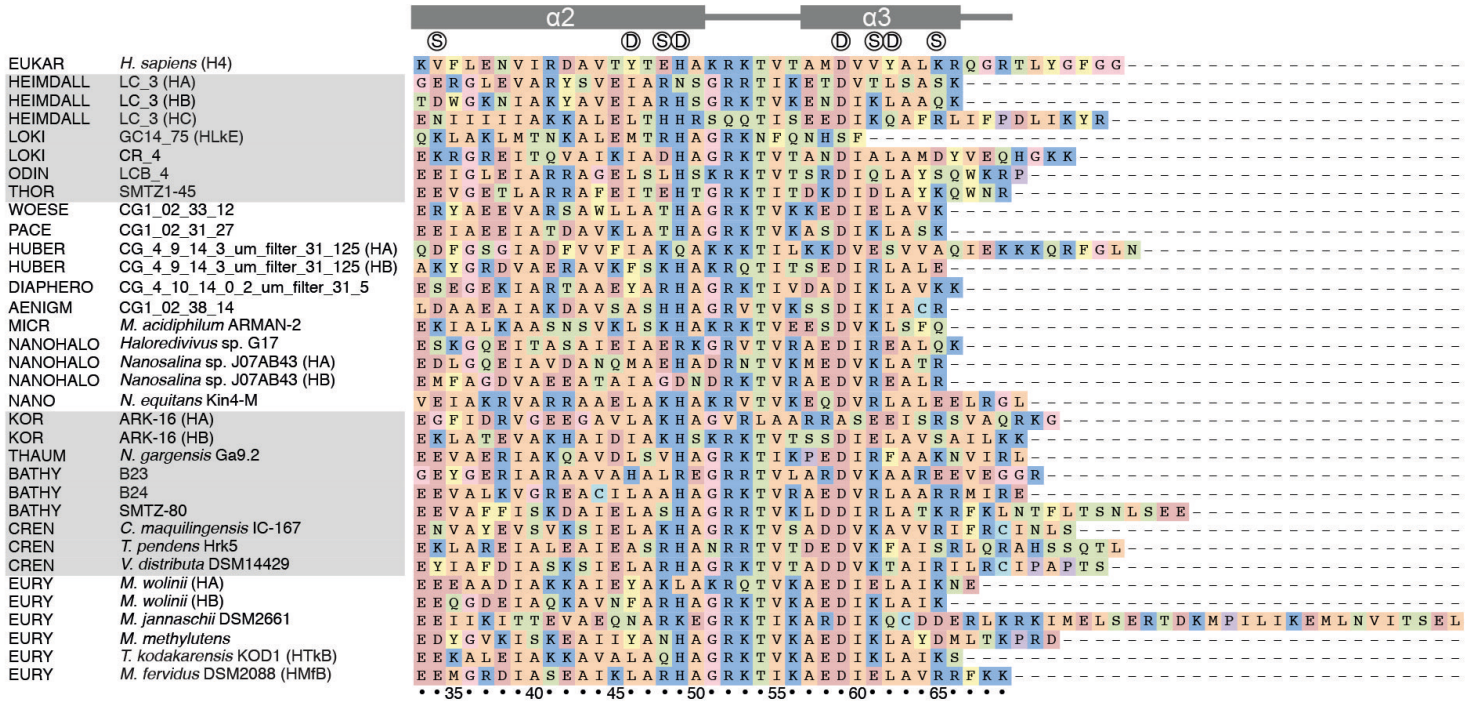


Figure 2.2. Alignment of histones from different archaeal species and human histone H4. Colors indicate the side chain group: R, H, K: blue; D, E: red; A, V, I, L, M: orange; F, Y, W: yellow; S, T, N, Q: green; C: turquoise; G: pink; P: purple. Symbols above the alignment indicate the dimer–dimer interface (D), the loop of the stacking interface (L), and the putative stacking interactions (S) (based on HMfB). Secondary structure and numbering of HMfB is used for reference. EUKAR H4, eukaryotic (human) histone H4 NP_724344.1; HEIMDALL LC_3 HA, HB, and HC, *Ca.* Heimdallarchaeota OLS22332.1, OLS24873.1, and OLS21974.1, respectively; LOKI GC14_75 HLkE and CR_4, *Ca.* Lokiarchaeota KKK41979.1 and OLS16336.1, respectively; ODIN, *Ca.* Odinarchaeota OLS18261.1; THOR, *Ca.* Thorarchaeota KXH71038.1; WOESE, *Ca.* Woesearchaeota OIO61677.1; PACE, *Ca.* Pacearchaeota OIO41945.1; HUBER, *Ca.* Huberarchaeota CG_4_9_14_3_um_filter_31_125 HA and HB, PJB03565.1, and PJB04497.1, respectively; DIAPHERO, *Ca.* Diapherotrites PJA17623.1;



(Figure 2.2, continued) AENIGM, *Ca.* Aenigmarchaeota OIN88081.1; MICR, *Ca.* Micrarchaeota *Micrarchaeum acidiphilum* ARMAN-2 EET90461.1; NANOHALO, *Ca.* Nanohaloarchaeota *Haloredivivus* sp. G17 and *Nanosalina* sp. J07AB43 HA and HB, EHK01841.1, EGQ42849.1, and EGQ43804.1, respectively; NANO, Nanoarchaeota *Nanoarchaeum equitans* Kin4-M AAR39197.1; THAUM, Thaumarchaeota *Nitrososphaera gargensis* Ga9.2 AFU59009.1; BATHY B23, B24, and SMTZ-80, *Ca.* Bathyarchaeota KYH36356.1, KYH37304.1, and KON27866.1, respectively; CREN, Crenarchaeota *Caldivirga maquilingensis* IC-167, *Thermofilum pendens* Hrk5, and *Vulcanisaeta distribute* DSM14429, ABW02527.1, ABL77757.1, and ADN51226.1, respectively; EURY, Euryarchaeota *Methanobrevibacter wolinii* HA and HB, *Methanocaldococcus jannaschii* DSM2661, *Methanococcoides methylutens*, *Thermococcus kodakarensis* KOD1, and *Methanothermus fervidus* DSM2066, WP_42707783.1, WP_42706862.1, AAB99668.1, KGK98166.1, BAD86478.1, and ADP77985.1, respectively.

one archaeal genome (198, 199). Also notable are genomes LC_3 of the phylum *Ca.* Heimdallarchaeota, and *Methanosphaera* sp. BMS, which both encode 10 histone paralogs (21). We have not found any histones in the genomes of phyla *Ca.* Parvarchaeota and *Ca.* Marsarchaeota, although it should be noted that abundance of available genomes and completeness of the genomes differs. The majority of available genomes from the phyla that do not seem to encode histones have an estimated completeness of between 40% and 99% (26, 30). This means that we cannot rule out the possibility that any of those genomes does contain one or more genes coding for histones. The absence of histones suggests that other NAPs may be involved in genome compaction. In that light, it is notable that in genomes from *Ca.* Parvarchaeota and *Ca.* Marsarchaeota, as well as in genomes from the phyla from Asgard Archaea, *Ca.* Woesearchaeota, *Ca.* Bathyarchaeota, *Ca.* Pacearchaeota, *Ca.* Verstraetearchaeota, *Ca.* Geothermarchaeota, *Ca.* Aenigmarchaeota and *Ca.* Micrarchaeota, genes coding for the DNA-bridging protein Alba1 (and in some cases, Alba2) are present. Like histones, Alba (or Sso10b) proteins are likely involved in transcription repression. They are highly abundant in nonhistone-coding Crenarchaeota (200), possibly taking the functional role of histones as found in other archaea. Some *Ca.* Parvarchaeota genomes encode Alba but not histones, and their genomes may therefore be shaped or regulated in a similar way as in Crenarchaeota. Furthermore, we found that only genomes of *Ca.* Thorarchaeota and Thermoplasmata contain an HU gene (a DNA-bending protein generally found in bacteria and archaea without histones (183)). Also, some genomes from Euryarchaeota and TACK encode Dps, a DNA protection protein that is found in bacteria. Dps has several mechanisms to protect the DNA in bacteria, such as strong DNA compaction and protection from hydrogen peroxide toxicity, but it is not known if archaeal Dps is able to function in a similar way. Additionally, the genomes of *Ca.* Huberarchaeota, *Ca.* Altiarchaeota and some Euryarchaeota encode an MC1 homologue, which is a monomeric DNA-bending protein often found in organisms from the euryarchaeal class Halobacteria. Genes coding for other known archaeal NAPs (201-203) were not found.

Some archaeal histones have eukaryote-like N-terminal tails

A striking finding based on the amino acid sequence comparison reveals that two histones from *Ca.* Heimdallarchaeota archaeon LC_3 (only Histone A [HA] shown in **FIGURE 2.2**), one from *Ca.* Huberarchaeota archaeon CG_4_9_14_3_um_filter_31_125 and one from *Ca.* Bathyarchaeota archaeon B23, contain an N-terminal tail, which was previously thought to exist only in eukaryotic histones and only recently reported for *Ca.* Heimdallarchaeota (197). In eukaryotes, these tails stabilize a higher order of compaction by interacting with either the DNA or

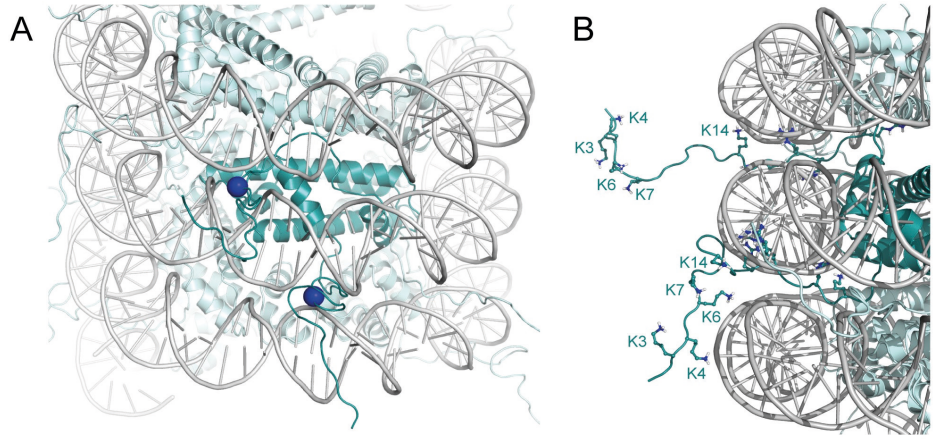


Figure 2.3: Model of a *Ca.* Heimdallarchaeota archaeon LC_3 hypernucleosome with N-terminal tails. A) View showing histone tails protruding through the DNA minor grooves. The R17 Cα-atom is shown as a blue sphere to mark the exit point of the tail. B) Close up of the histone tails with lysine and arginine residues shown as sticks, and N-terminal lysines are labeled. Homodimers of Heimdall LC_3 histone HA are shown in teal; one dimer is highlighted in darker colors. Models are based on the structure of HMfB (PDB entry 5T5K); the tail in the top (bottom) of panel B is modeled in the H3 (H4) tail conformation (PDB entry 1KX5).

another nucleosome. The tails of the two histones from *Ca.* Heimdallarchaeota and *Ca.* Huberarchaeota are of roughly the same length and sequence composition as eukaryotic H4 tails (*see* **FIGURE 2.2**). Prompted by the importance of the eukaryotic histone tails in modulating chromatin structure and function (150, 155), collaborator Clara van Emmerik constructed a molecular model of a hypernucleosome formed by Histone A (HA) from *Ca.* Heimdallarchaeota LC_3 to investigate its potential function (*see* **METHODS SECTION**).

The model illustrates how three subsequent arginines (R17–R19) could facilitate passing of the tails through the DNA gyres (**FIGURE 2.3**). The tails exit the hypernucleosome through DNA minor grooves, similar to eukaryotic histone tails, and might position their lysine side chains to bind to the hypernucleosomal DNA or to other DNA close by, facilitating (long-range) genomic interactions in trans. Like the H4 tail that is subject to acetylation of lysines K5, K8, K12, and K16 (204), lysines in the heimdallarchaeal histone tail may well be subject to acetylation. Archaeal genomes are known to have several candidate lysine acetyltransferase and deacetylase enzymes, including proteins belonging to the ELP3 superfamily, to which transcription elongation factor and histone acetyltransferase ELP3 belongs (205-207). Searches using the ProSite database* (208) and Protein Information Resource** (209) further reveal that the *Ca.* Heimdallarchaeota LC_3 genome

* http://prosite.expasy.org
** http://pir.georgetown.edu

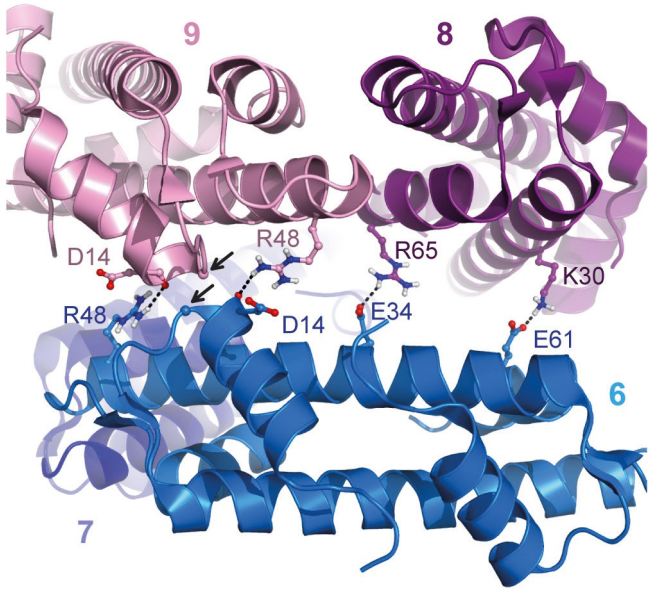


Figure 2.4: Stacking interface between HMfB dimers in the hypernucleosome. Each dimer *i* forms stacking interactions with dimer *i*+2 and *i*+3, shown here for dimer 6. Residues deemed important are shown in ball-and-sticks and labeled; close stacking of G16 in dimer 6 and 9 is indicated by arrows (Cα shown as spheres); hydrogen bonds are indicated with dashes. Image generated using PDB entry 5T5K (197)

contains multiple gene products containing the Gcn5-related N-acetyltransferase domain, which is present in many histone acetyltransferases (210). Interestingly, a potential 'reader' protein that binds modified lysines can also be identified. This protein, HeimC3_47440, contains a YEATS-domain, which has recently been shown to bind histone tails that carry acetylated or crotonylated lysines (211-214). Comparison with the closest homolog of known 3D structure, YEATS2 (35% identity, PDB-id 5IQL, (215)), shows that the binding site for the modified lysine side chain is strictly conserved in the archaeal protein. Notably, only *Ca.* Bathyarchaeota, which also features tailed histones, contains a detectable homolog of HeimC3_47440. The presence of lysine-containing N-terminal tails in combination with histone modification writers and readers suggests that archaea use post-translational modifications in a similar way to Eukaryotes as modulators of genome compaction and gene activity. The tail of the *Ca.* Huberarchaeota histone also contains lysine residues that are found at the same position as some of the lysines of the H4 tail. However, we have not identified any proteins resembling post-translational modification-related proteins from other organisms in terms of sequence in this phylum.

Other histones, for example from *Ca.* Lokiarchaeota CR_4, *Ca.* Odinarchaeota

LBC_4, *Nanoarchaeum equitans* (Nanoarchaeota), and *Thermofilum pendens* (Crenarchaeota), contain a short N-terminal tail of 5–10 residues. Also, histones with a C-terminal tail have been found. The histone from the euryarchaeal species *Methanocaldococcus jannaschii* (class Methanococci) has a 28-residue C-terminal tail, which seems to be unique among archaeal histones. Other C-terminal tails are up to 11 residues long (as compared to *Methanothermus fervidus* HMfB) and appear in *Caldiarchaeum subterraneum* (Thaumarchaeota), *Ca.* Bathyarchaeota SMTZ-80, *Ca.* Heimdallarchaeota LC_3, *Ca.* Lokiarchaeota CR_4, and all histones found in Crenarchaeota. These short C-terminal tails are similar in length to the H4 C-terminal tail, that is reported to play a role in the promotion of histone octamer formation in eukaryotes (216). The genomes of some archaeal species contain genes for histone truncates. The histone from *Haloredivivus* sp. G17, member of *Ca.* Nanohaloarchaeota, and the histone from *Ca.* Bathyarchaeota archaeon B24 both lack part of the N-terminal α-helix (α1), and one histone from *Ca.* Lokiarchaeota GC14-75 is reduced in length at the C-terminus. The remainder of the C-terminal amino acids likely does not form a C-terminal helix (α3) in this histone from *Ca.* Lokiarchaeota. Although histones of reduced length or containing tails lack part of the histone fold, they likely still possess DNA-binding properties. Therefore, they possibly have functional roles in the regulation of genes.

Multimerization of histones

Both eukaryotic histones and HMfB form dimers, a process that is driven by a hydrophobic core (involving residues A24, L28, L32, I39, and A43 in HMfB) as well as a crucial salt bridge for a stable histone fold (R52-D59 in HMfB) (217). These hydrophobic residues and the salt bridge are conserved among archaea. This indicates that archaeal histones have very similar tertiary structures (217, 218). Also, residues that play an important role in DNA binding are present in all examined histones, including the arginines that anchor archaeal histone dimers to the DNA minor grooves (R10 and R19 in HMfB) (217). Both eukaryotic H3-H4-dimers and HMfB dimers can form tetramers by hydrogen bonding of H49 and D59 (HMfB) and additional hydrophobic interactions in the interface (L46 and L62 in HMfB) (171), pairs of residues that, too, are generally conserved among archaeal histones (**FIGURE 2.2**).

Recently, a crystallography study on HMfB shed new light on the compaction of DNA by archaeal histones. The HMfB–DNA cocrystal structure reveals left-handed wrapping of DNA around a histone-multimer core (197) (**FIGURE 2.1**). This structure supports the model in which HMfB dimers multimerize along

DNA into an 'infinite' hypernucleosome, thereby linearly compacting the DNA approximately ten-fold. It is likely that hypernucleosomes grow or shrink by association or dissociation of dimers at both ends. The resolution of the crystal structure allowed us to identify several interacting residues between layers of dimers that may be important for stabilizing the complex (**FIGURE 2.4**). Based on this structural information, the propensity of different archaeal histones to multimerize can be predicted.

In **TABLE 2.2**, we set out three criteria for hypernucleosome formation by archaeal histones. Firstly, conservation of residues in the dimer–dimer interface (E42, L46, H49, D59, L62 and R66 in HMfB) is required, as forming a tetramer is the first step in multimerization. Secondly, residue G16, which is positioned at the stacking interface of the hypernucleosome (**FIGURE 2.4**), is crucial in permitting formation of the hypernucleosome (197). Bulkier residues at this position interfere with multimerization (197). Lastly, favorable interactions between histone dimers *i* and *i*+2 and *i*+3, here termed stacking interactions, will contribute to stability of the compacted hypernucleosome. The HMfB hypernucleosome crystal structure shows three stacking interactions, hydrogen bonds from K30 to E61, E34 to R65, and R48 to D14 (**FIGURES 2.2** and **2.4**).

Scrutiny of a selection of histone sequences reveals that most archaeal histones meet these criteria and are thus likely to form hypernucleosomes, (**TABLE 2.2**, marked +). We identified two to seven potential stacking interactions for this group of histones, which may affect hypernucleosome stability and compactness. Fewer interactions may allow for more 'breathing' of the hypernucleosome structure, yielding hypernucleosomes that are more flexible or 'floppy'. We predict such structures to be formed also by a number of archaeal histones that do not fully meet our criteria (**TABLE 2.2**, marked ±). For example, *Ca.* Heimdallarchaeota LC_3 HA and *Ca.* Lokiarchaeota GC14_75 HLkE have H49N and D59S substitutions, respectively, which likely weakens the crucial hydrogen bonding interaction at the dimer–dimer interface (171). Similarly, substitution of the hydrophobic residues 46 and 62 for more hydrophilic or bulkier ones would lead to a less stable dimer–

Caption table 2.2 (page 43): Dimer-dimer interactions in the tetrameric interface were assumed to be essential for hypernucleosomes formation. Absence of bulky residues in the first loop and a high number of potential hydrogen bonds in the stacking interface will enhance the compactness and stability of the hypernucleosome. Likely, uncertain and unlikely stacking ability is indicated with +, ± and -, respectively.
[a] Dimer-dimer interface includes residues at position 42, 46, 49, 59, 62 and 66.
[b] Stacking interface includes residues at positions 15-17.
[c] For all potential stacking interactions, residue numbering of HMfB was used according to the alignment in **FIGURE 2.2**.

Table 2.2: Assessment of possible hypernucleosome formation by archaeal histones.

		Dimer-dimer interface[a]	Stacking interface[b]	Potential stacking interactions[c]	Hypernucleosome formation	Histone features
Heimdall	LC_3 HA	±	+	3 (E14-R48, K26-E57, R41-E45)	±	N-terminal tail
Heimdall	LC_3 HB	+	+	5 (E30-K61, Q14-R48, R13-Q18, K27-E57, K37-E45)	+	
Heimdall	LC_3 HC	±	+	3 (N34-R65, T15-K41, Y14-Q53)	±	
Loki	GC14_75 (HLkE)	-	+	2 (D14-R48, K34-E45)	-	Truncated C-term.
Loki	CR_4	+	+	2 (Q14-D48, Q41-Q41)	+	
Odin	LCB_4	+	-	3 (K30-Q61, K14-E18, E38-R41)	±	
Thor	SMTZ1-45	+	+	5 (Q30-D61, E34-K65, K14-E48, E37-R41, E26-K58)	+	
Woese	CG1_02_33_12	+	+	4 (R14-T48, R34-E61, E26-K57, E37-R41)	+	
Pace	CG1_02_31_27	+	+	4 (S30-K61, E34-K65, K14-T48, E37-K45)	+	
Aenigm	CG1_02_38_14	+	+	5 (E30-K61, D34-R65, E14-H48, A15-K41, E37-K41)	+	
Micr	*M. acidiphilum* ARMAN-2	+	+	3 (E30-K61, K34-Q65, Y2-K48)	+	
Nanohalo	*Haloredivivus* sp. G17	±	-	2 (E27-R61, Q37-E45)	-	Truncated N-term.
Nanohalo	*Nanosalina* sp. J07AB43 (HA)	+	+	5 (Q30-K61, D34-R65, K14-E48, K14-E18, Q37-Q45)	+	
Nanohalo	*Nanosalina* sp. J07AB43 (HB)	-	±	2 (Q30-R61, D14-K18)	-	
Nano	*N. equitans* Kin4-M	+	+	4 (E30-R61, Q14-K48, Q14(bb)-R41, K37-E45)	+	
Aig	*C. subterraneum*	±	+	4 (K30-E61, K14-Q48, E27-K57, K41-E45)	±	
Thaum	*N. gargensis* Ga9.2	+	+	4 (E34-K65, K14-E18, E27-R61, E37-K41)	+	
Bathy	B23	±	±	4 (R14-V44, E34-K61, E37-R41, E26-R58)	±	N-terminal tail
Bathy	B24	+	+	3 (E34-R65, K14-E18, E27-R61)	+	Truncated N-term.
Bathy	SMTZ-80	+	+	3 (E34-K65, K41-E45, E27-R61)	+	
Cren	*C. maquilingensis* IC-167	+	+	4 (D30-K61, N34-R65, K14-E18, Y37-K48)	+	
Cren	*T. pendens* Hrk5	+	+	4 (E30-K61, S14-R48, R37-E45, R13-E18)	+	
Cren	*V. distributa* DSM14429	+	+	4 (D30-K61, Y34-R65, K14(bb)-R48, K14-E18)	+	
Eury	*M. wolinii* (HA)	±	+	4 (N30-E61, E34-K65, E14-K48, K41-E45)	±	
Eury	*M. wolinii* (HB)	+	+	5 (E30-K61, E34-K65, N14-R48, N14-Q18, Q41-Q41)	+	
Eury	*M. jannaschii* DSM2661	+	-	4 (N30-K61, Q14-R48, K37-Q45, D26-R58)	±	C-terminal tail
Eury	*M. methylutens*	±	-	2 (D30-K61, S14-E18)	-	
Eury	*T. kodakarensis* KOD1 (HTkB)	+	+	4 (E30-K61, E34-K65, K14-Q48, K26-E58)	+	
Eury	*M. fervidus* DSM2088 (HMfB)	+	+	3 (K30-E61, E34-R65, D14-R48)	+	

dimer interface, as for *Ca.* Heimdallarchaeota LC_3 HC and *Ca.* Bathyarchaeota B23. In the presence of the canonical dimer–dimer interface, bulky substitutions at position 16 likely also result in a more open hypernucleosome structure, as for *Ca.* Odinarchaeota LCB_4.

Three archaeal histone species fail multiple criteria in our analysis, indicating that these cannot form hypernucleosomes. These histone species are *Haloredivivus* sp G17, *Nanosalina* J07AB43 HB, and Euryarchaeal *Methanococcoides methylutens* (class Methanomicrobia) that all combine defects in the dimer interface with a bulky substitution at position 16 and few potential stacking interactions (**TABLE 2.2**, marked –). In particular, *Nanosalina* J07AB43 Histone B (HB) shows a H49D substitution and a glutamic acid at position 62, making the dimer surface highly negatively charged and thus very unlikely to interact with another dimer.

It is remarkable that most of the histones having N- or C-terminal tails or N- or C-terminal truncations additionally have substitutions in the dimer–dimer and/or stacking interface that will affect hypernucleosome formation. Histones with reduced ability to form compact hypernucleosomes are expected to exhibit different roles in shaping the genome, like simple DNA bending or site-specific interference with histone multimerization. Interestingly, the genomes of several organisms encode histones that we predict are able to multimerize as well as histones that probably do not multimerize. Histones have been hypothesized to bind to promoters, thereby making the promoter inaccessible to transcription factors that onset gene expression. The ability to multimerize into hupernucleosomes suggests that histones may also have an effect on gene expression via multimerization, thereby potentially silencing the gene by binding into the coding region.

Sequence analysis of archaeal histones

Although a sequence alignment of a selection of histones from throughout the archaeal domain as shown in **FIGURE 2.2** and a structural analysis as discussed in **TABLE 2.2** give a good overview of the multeity of structural and sequential elements, a more systematic sequence analysis of all known archaeal histones allows for a better characterization of phylum-specific features of archaeal histones. Here, we will discuss the histones of all archaeal phyla and all euryarchaeal classes. Sequences of these histones can be found in the supplementary Excel file (accessible via the Leiden Repository). Main points of discussion are length and isoelectric point (pI), the residues involved in DNA binding and dimerization, the well-conserved 'HAGRKT-motif' at the C-terminal linker, the stacking interface or AGA-loop situated at the N-terminal linker, and the stacking interactions (as

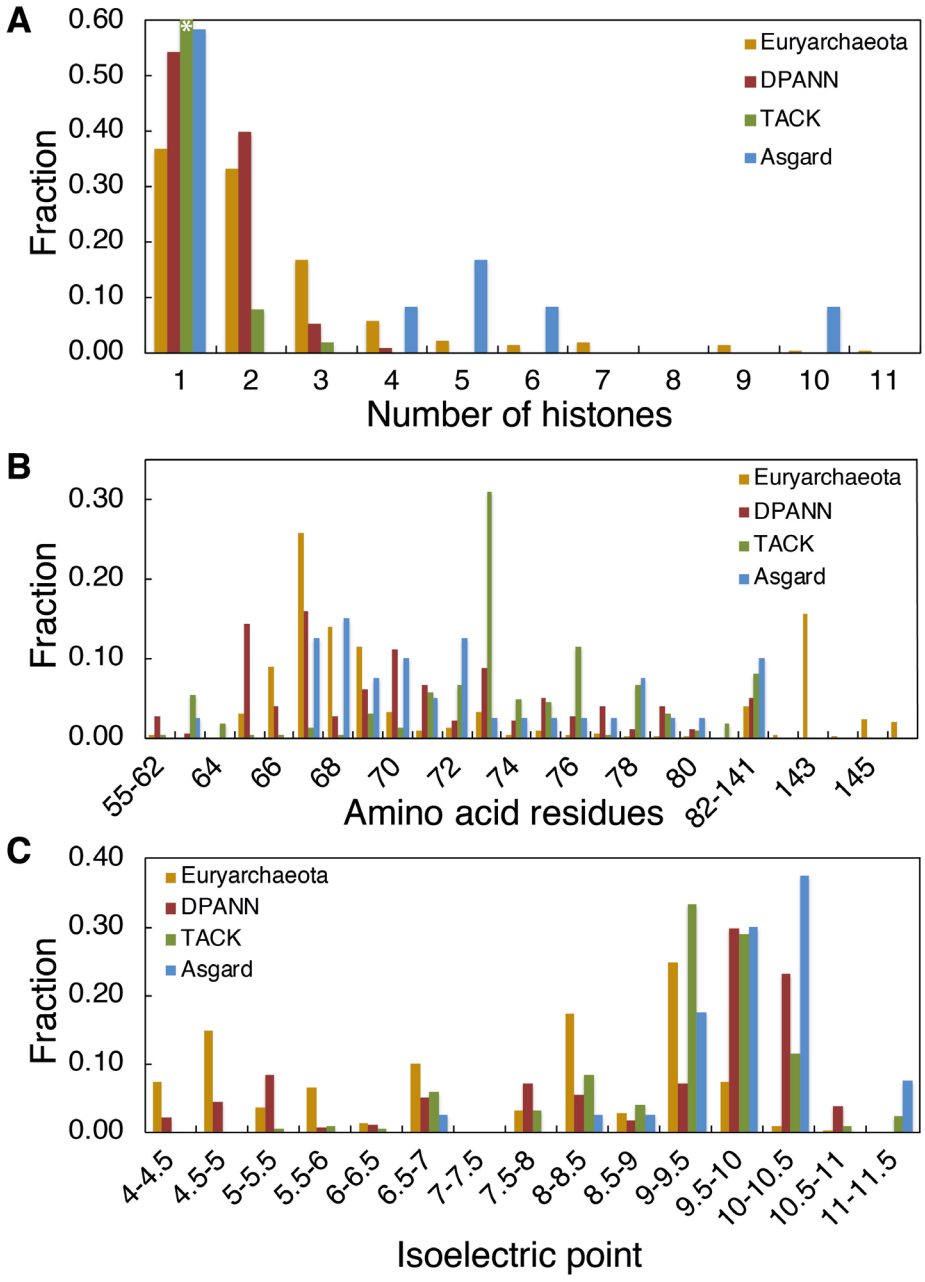


Figure 2.5: Characteristics of archaeal histones categorized by superphylum. A) Histones per genome, as fraction of the total number of genomes per category. The double-histone-fold histones of Halobacteria are not included in this analysis. *) The value of TACK at one histone is 0.90. Total number of genomes included: Euryarchaeota, 228; DPANN, 118; TACK, 203; Asgard archaea, 12. B) Division of histone lengths, as fraction of the total number of histones per category. Total number of histones: Euryarchaeota, 662; DPANN, 181; TACK, 225; Asgard archaea, 40. C) Division of theoretical isoelectric point, as fraction of the total number of histones per category. Total number of histones: *see* B.

found in HMfB). Histones from genomes that were not assigned to a phylum or class, are disregarded. Some histones have been identified as being part of the same histone type within a taxon. These histones are more similar to histones within the same type than histones from the same organism, and are identified using sequence alignment with Clustal Omega (*see also* **Methods**). Histone types are assigned independently of other types, and so histones of a certain type in one taxon do not have any relation with histones from another type with the same name in another taxon.

Euryarchaeota

In all superphyla and Euryarchaeota, the mode of number of histones encoded per genome is one. However, in Euryarchaeota, 63% of the genomes evaluated in this study encode two or more histones (**figure 2.5A**). Most euryarchaeal histones are between 65 and 70 amino acid residues long, with the double-histone-fold histones from Halobacteria being the main exception, at 142-150 amino acid residues (**figure 2.5B**). The theoretical isoelectric point (determined using ProtParam (219)) of euryarchaeal histones varies between 4 and 11 (**figure 2.5C**). The pI of histones from halophilic organisms is usually between 4 and 6, whereas the pI of histones from other organisms is mostly between 7.5 and 10.

Archaeoglobi

The majority of Archaeoglobi genomes encodes two histones, although some encode three histones. There is little variance between the histones, and unlike histones from for example Thermococci or *Ca.* Nanohaloarchaeota, no types can be identified. Therefore, it remains unclear how the two or three different histones in Archaeoglobi are functionally different. Likely, Archaeoglobi histones are able to bind DNA. The number of tetramerization interactions differs. Some histones can probably form five tetramerization interactions, while others can form only four. The AGA-loop is conserved among all histones. Stacking interactions can likely be formed by most histones in this class, which may result in formation of hypernucleosomes.

Hadesarchaea

The four genomes of Hadesarchaea included in this study encode one, two (twice) or three histones. The AGA-loop is found in only two histones, but four others have SGA instead, which may lead to a more flexible hypernucleosome. Two histone-coding genes do not encode an AGA-loop. The hadesarchaeal histones

without AGA or SGA are predicted to be able to bind DNA and tetramerize, like the others, but may do this with fewer interactions. Affinity for DNA might therefore be lower than for the other histones. The histones with AGA or SGA loop are able to form at least two stacking interactions, and may therefore (at least, if SGA does not sterically hinder it) assemble into hypernucleosomes. For the other two, this is unlikely, not only because the amino acid residues in their N-terminal loop probably cause sterical hindrance and thereby prevent dimerization, but also because these histones do not have enough stacking interactions.

Halobacteria

Halobacteria express a unqiue kind of histones, with two histone folds. These histones can be regarded as naturally linked dimers. Because of the double histone fold, they are much longer than other histones. Their length is 140-150 amino acid residues, with a mode of 143 residues. Within Halobacteria, histone sequence is highly conserved (FIGURE 2.6). The pI of halobacterial histones ranges from 4.24 to 5.37, which is very low compared to other archaeal histones, although these values are common for salinophilic archaea. Because of its two histone folds, one histone fold of the halobacterial histone is able to 'specialize' in certain interactions (182). Therefore, not all residues involved in DNA-

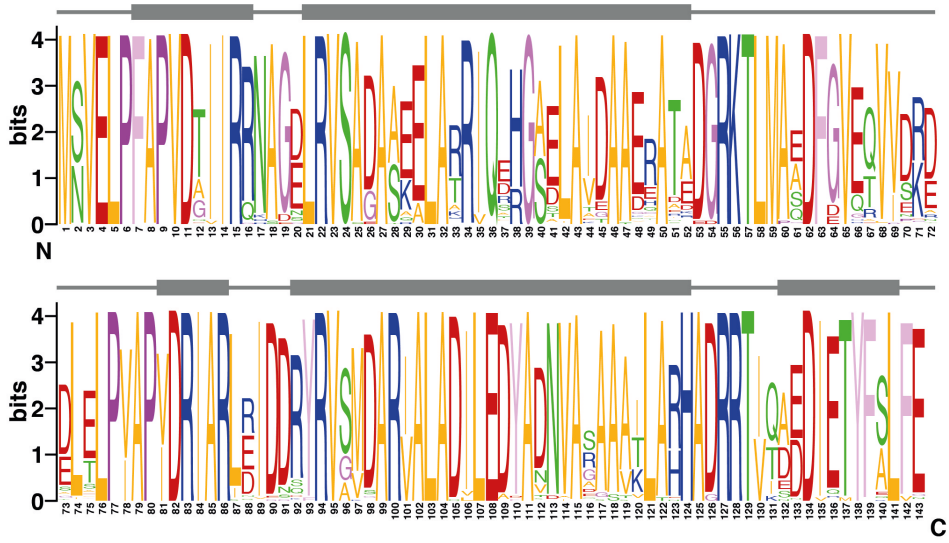


Figure 2.6: Sequence logo of histones from Halobacteria. The top panel represents the N-terminal histone fold, the bottom panel represents the C-terminal histone fold. 104 histones with a length of 143 residues were included in this figure. Grey bars represent predicted α-helices. Figure made using WebLogo (224).

binding- and tetramerization interactions are present in one of the histone folds, which makes it difficult to discuss those interactions here. From *in vivo* experiments, it is known that halobacterial histones bind DNA as tetramers, and probably do not form hypernucleosomes (220, 221). Histones from Halobacteria do not have an AGA-loop, but in the N-terminal histone fold, XGA is often present at positions 15-17. Only the second half of the HAGRKT-motif is found in the N-terminal histone fold, whereas the C-terminal histone fold usually contains a full (variant of the) HAGRKT-motif. The least conserved part of halobacterial histones is the central α-helix and C-terminus of the N-terminal histone fold. Positively charged and hydrophobic residues are generally well-conserved (**FIGURE 2.6**). *In vitro* studies suggest that histone HpyA from *Halobacterium salinarum* plays a minor role in DNA compaction, but is important for growth phase-dependent transcription regulation (185). MC1 may be the main DNA compacting protein in Halobacteria (222, 223).

Hydrothermarchaeota

Two histone-encoding genomes of Hydrothermarchaeota were identified, both encoding one histone. These histones are identical. Remarkably, this histone does not contain an AGA loop. Also, no potential stacking interactions could be identified, which would make hypernucleosome formation highly unlikely. Hydrothermarchaeal histones are likely able to bind DNA and to tetramerize, as residues generally involved in these processes are present. Most hydrothermarchaeal genomes encode Alba proteins (225). Based on the hypothesized inability of histones to form hypernucleosomes, Alba may be the main chromatin protein responsible for DNA compaction in Hydrothermarchaeota.

Methanobacteria

The histone distribution of histone-coding genes in Methanobacteria is highly diverse. Of the 42 genomes that we have examined, half encoded three histones. However, two genomes encoded a single histone, whereas one genome encoded ten histones and one genome encoded eleven histones. Methanobacterial histones as a group cannot be categorized into types, since histones are better conserved within a certain genus than within subgroups. The class Methanobacteria contains only one order, Methanobacteriales, which is subsequently subdivided into two families: Methanobacteriaceae and Methanothermaceae. The former is further subdivided into the histone-encoding genera *Methanobacterium*, *Methanobrevibacter*, *Methanothermobacter* and *Methanosphaera*, whereas the latter is made up of a single genus: *Methanothermus*.

Genomes from the genus *Methanobacterium* mostly encode three histones, with the main exception being *Methanobacterium bryantii* M.o.H., encoding seven histones. The histones of *Methanobacterium* seem to be equal in terms of DNA binding and stacking interactions, but based on tetramerization interactions, three types can be discerned, which form four, five or six stacking interactions. Presence of the AGA-loop and putative tetramerization and stacking interactions suggest that these histones can assemble into a hypernucleosome.

Most of the genomes from *Methanobrevibacter* encode three or four histones. The histones can be categorized into three types, A, B and C. Histones from type A likely form more stacking interactions than the others, which in theory could result in more stable hypernucleosome formation. However, the histones of type A lack H49, which is very important in tetramerization. Thus, whether type A histones are able to form hypernucleosome remains unclear. This is supported by the structural analysis of this chapter, in which we concluded that histone HA of *Methanobrevibacter wolinii* may or may not form hypernucleosomes. Histones of type B and C are likely able to multimerize into hypernucleosomes, and do not differ in residues that are important for DNA binding, tetramerization or stacking. Within *Methanobrevibacter*, large differences between the pIs of histones are observed (5.10-9.52). This may have an effect on transcription regulation in response to environmental cues.

All genomes of the genus *Methanothermobacter* included in this study encode three histones. These histones form three types, D, E, and F. Type D is probably slightly impaired in tetramerization, as it forms five tetramerization interactions while the histones from the other types form six tetramerization interactions. Unlike *Methanobrevibacter* histones, all types contain the important and well-conserved H49, and therefore tetramerization is likely not affected in *Methanothermobacter* histones. Interstingly, the histones from type E exhibit more potential stacking interactions than the others. Histones from type D or F do not differ at residues involved in any known interactions. The AGA-loop is also present, which together with the likely presence of tetramerization and stacking interaction leads to the conclusion that *Methanothermobacter* histones of any type probably form hypernucleosomes. The histones from *Methanothermobacter thermautotrophicus*, HMtA1 (type D), HMtA2 (type E) and HMtB (type F) were studied *in vitro*. Substitution R19I in HMtB, found in a laboratory strain, resulted in the loss of DNA binding of the HMtB homodimer. However, $HMtB_{R19I}$-HMtA2 heterodimers were able to bind DNA (226).

Methanosphaera is the genus of which the genomes encode on average the largest

number of histones: between four and eleven. These histones form seven loosely-defined types, and histones from these types are not equally distributed among the *Methanosphaera* genomes. Histones from most types exhibit six potential DNA-binding interactions, although some types are likely able to form just five DNA-binding interactions. Histones from most types form five tetramerization interactions, but this number varies from three to six. The histones of one type, expected to form just 3 tetramerization interactions, also lack H49, which means the histones belonging to this type likely cannot tetramerize or form hypernucleosomes. The number of potential stacking interactions varies even within types, and therefore it is impossible to make statements about the hypernucleosome formation abilities of the types. Similar to *Methanobrevibacter*, within *Methanosphaera*, large differences between the pIs of histones are observed (4.64-9.77), which may be of importance for transcription regulation in response to environmental cues.

Within the genus *Methanothermus*, only three species have been identified: *Methanothermus fervidus*, *Methanothermus sociabilis* and *Methanothermus jannaschii*. Histones are exclusively encoded by *M. fervidus*. These histones, HMfA and HMfB, are regarded as archaeal 'model histones', and are very well studied (*see also* **CHAPTER 4** and **6** of this thesis). Hypernucleosome formation was first shown for *M. fervidus* histones (197), and sequence and structural analysis in this chapter is based on experiments on HMfA and HMfB.

Methanococci

Of the 17 genomes included in this study, eight encode three histones, while the others encode two to five histones. Some histone-coding genes encode C-terminal tails with a length of ~32 residues, which are relatively well-conserved. Most genomes encode at least one histone with C-terminal tail, possibly forms an additional α-helix (identified with Jpred4 (227)). The function of this tail remains unclear. Remarkably, the histones with C-terminal tail likely have very different properties compared to Methanococci histones without a tail. The histones with tail have a poorly conserved HAGRKT-motif and lack an intact AGA-loop. They almost completely lack residues involved in tetramerization, and are probably unable to form stacking interactions. This means that it is very unlikely that histones with a C-terminal tail form hypernucleosomes. However, these histones can likely bind to DNA. In contrast, Methanococci histones without tail have an AGA-loop and can probably form DNA-binding-, tetramerization- and stacking interactions, which leads to the conclusion that they likely form hypernucleosomes. Most Methanococci genomes encode two histones without tail that are very similar. One of the two often contains a C-terminal methionine of which the function is unknown, whereas the other does not. The clear difference between

Methanococci histones with and without tail indicates that the histones without tail form hypernucleosomes, while histones with tail regulate hypernucleosome length or position hypernucleosomes onto specific loci.

Methanomicrobia

Of the 120 Methanomicrobia histones included in this study, more than half is the only histone encoded by its genome. ~75% of histones have an intact AGA-loop. Most histones are likely able to form DNA-binding- and tetramerization interactions, although a large minority of histones has one or more tetramerization defects. The histidine at position 49 is conserved in 99% of histones. Stacking interactions are found in Methanomicrobia histones, although the number varies from zero to three. Therefore, some histones likely form hypernucleosomes, some likely do not, and for some the analysis is inconclusive. The histone of *Methanococcoides methylutens* DSM 2657, the only histone encoded by this genome, are likely unable to form hypernucleosomes, as shown by the structural analysis (**table 2.2**). Thus, hypernucleosome formation is likely nonessential in Methanomicrobia. Some histones in this class have a highly charged C-terminus, which may be used for transcripton factor recruitment. *Methanosarcina*, a genus within Methanomicrobia of which the genome encodes one histone, also expresses MC1 (formerly known as HMb). In this genus, MC1 was identified as the major component of extracted nucleoprotein, which may indicate that this NAP is the most important protein involved in genome compaction (228-230).

Methanonatronarchaeia

One genome from Methanonatronarchaeia encoding one histone was included in this analysis. This histone is likely able to form DNA-binding-, tetramerization- and stacking interactions. Its AGA-loop is not fully conserved, but instead it has GGA at positions 15-17. The A15G substitution probably does not interfere with hypernucleosome formation, since glycine does not possess a side chain and therefore it likely does not sterically hinder multimerization of histones.

Methanopyri

Of the six Methanopyri genomes that have been identified, only *Methanopyrus kandleri* AV19 encodes histones. The three gene products annotated as histones are vastly different from each other. One gene product is 77 amino acid residues long, has a pI of 9.50 and lacks features that are associated with histones. The gene product is predicted to contain a histone fold (identified with Jpred4 (227)). It is

therefore questionable if this gene product really is a histone, or if it was wrongly annotated. Another histone is 91 residues long, has a pI of 9.15 and has a HAGRKT-motif. It exhibits some similarities the histones with tail from Methanococci, and also unlikely to tetramerize or form hypernucleosomes. The third histone has 154 residues, a pI of 4.91 and has been shown to contain two histone folds (231). This histone, referred to in literature as MkaH, thus somewhat resembles halobacterial histones, although they lack the AGA-loop and HAGRKT-motif. Although there is little sequence similarity to other archaeal histones, *in vitro* experiments have shown that MkaH is able to dimerize into structures that resemble eukaryotic (H3-H4)2 tetramers, and to compact DNA (232).

Theionarchaea

Although only two genomes of Theionarchaea are included in this study, combined they encode twelve histones, five from archaeon DG-70 and seven from archaeon DG-70-1. The AGA-loop is present, although it is substituted with GGA in some cases. This likely does not interfere with hypernucleosome formation. Most of the histones are expected to multimerize into hypernucleosomes. However, one histone of each genome probably cannot tetramerize and therefore cannot form hypernucleosomes. Four out of twelve histones are probably able to form sufficient stacking interactions in order to form stable hypernucleosomes. The other histones either cannot form hypernucleosomes because of the before-mentioned defects in tetramerization interactions, or cannot be predicted to form hypernucleosomes based on this analysis.

Thermococci

Over 75% of Thermococci genomes encode two histones. Of these two histones, one usually has a lower theoretical pI than the other, the difference being approximately 1. We categorized thermococcal histones into type A, histones with lower pI, and type B, histones with higher pI. Although length and sequence of the histones of both types are relatively similar when compared to histones from other classes or phyla, there are four clear differences. E19, well conserved in type A, is poorly conserved in type B and can be alanine, glutamine or proline (**FIGURE 2.7**). The aromatic tyrosines in type A at position 32 and 36, are replaced by the positively charged residues histidine and lysine in type B, respectively. Finally, the aromatic phenylalanine or tyrosine at position 47 is not present in type B, which instead has leucine at position 47. The first three residues do not have a known effect on DNA binding, tetramerization, or hypernucleosome formation, however, L47 is involved in tetramerization, but this substitution was shown to

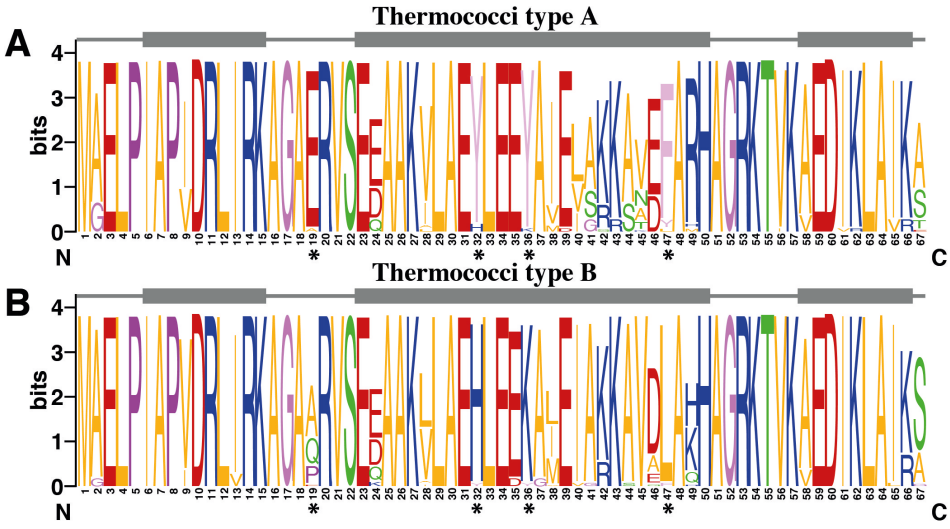


Figure 2.7: Sequence logo of the two types of thermococcal histones. A) Type A and B) type B. Type C was not included because of the heterogeneity in histone length. Grey bars represent predicted α-helices. Main differences between type A and type B are represented by asterisks. Figure made using WebLogo (224).

have no effect in electrophoretic mobility shift assays (EMSAs) (233). Despite these differences, histones from both types can likely multimerize into hypernucleosomes and bind DNA. This is supported by the finding that type A histone HTkA from *Thermococcus kodakarensis* forms hypernucleosomes *in vivo* (197). Most, but not all Thermococci histones can be placed in either types A or B. Type C, which represents the remaining histones, is probably less conserved than types A or B. The length of some type C histones does not exactly match the lengths of histones from type A or B, and the number of potential tetramerization interactions varies within type C. Additionally, some type C histones likely form weaker stacking interactions, which may have its effect on hypernucleosome formation. The genomes to which thermococcal type C histones belong, were all identified recently using metagenomic sequencing. The habitat from which they were retrieved does not differ from that of other Thermococci genomes (225).

Thermoplasmata

Most Thermoplasmata genomes do not encode histones. Currently, only five, all identified via metagenomic sequencing, encode a single histone. One of these, although annotated as histone, is missing all signature motifs and conserved elements, and is therefore probably not a real histone. It is therefore disregarded in this analysis. The remaining histones all have an AGA loop, but the number of potential stacking interactions differs, with some histones likely able to form

a hypernucleosome, while others are likely unable to do so. Also the expected number of DNA interactions and tetramerization interactions varies. It is therefore difficult to predict the function of Thermoplasmata histones. However, *Thermoplasma acidophilum* expresses HTa, a NAP that is a member of the bacterial HU protein family (234, 235). It is one of the most highly expressed proteins in *T. acidophilum*, and is a functional analog of histones (235, 236). Also, Alba is encoded by most Thermoplasmata genomes. HTa and Alba may together compact and regulate the genomes of Thermoplasmata.

DPANN

In DPANN, genomes that encode more than one histone, make up 56% of the data set. In contrast to Euryarchaeota, genomes encoding more than four histones were never found (**figure 2.5A**). Histones from DPANN come in a wide range of lengths, although most histones are between 65 and 70 amino acid residues long (**figure 2.5B**). Also, a broad range of theoretical pIs of histones can be found in DPANN, although the average pI is significantly higher than in Euryarchaeota (8.72 compared to 7.31) (**figure 2.5C**). Similar to Euryarchaeota, genomes found in high-salinity habitats, such as Nanoarchaeota and *Ca.* Nanohaloarchaeota, encode histones with low pIs (pI 4-6).

Ca. Aenigmarchaeota

Most of the twenty genomes of *Ca.* Aenigmarchaeota that encode histones, have one or two histone-coding genes. The AGA loop is usually intact, although some variations occur, such as YGA, LGA, NGA, VGA, CGA and SGA. It is not known what the effect of these AGA-loop variants is with respect to hypernucleosome formation. However, sequences of histones with an AGA-loop variant are likely less conserved in general. Potential stacking interactions are exhibited by a majority of aenigmarchaeal histones, which, together with the presence of the AGA-loop, makes it likely that those histones are able to form a hypernucleosome. The six residues involved in DNA binding are found in aenigmarchaeal histones, but none of the histones contains all six residues. This may result in distinct DNA binding affinities within this group of histones, and even within the same organism. Residues involved in tetramerization are usually found in aenigmarchaeal histones, although also here, some defects are found.

Ca. Altiarchaeota

Of the eleven genomes of *Ca.* Altiarchaeota included in this study, eight genomes

encode one histone. Of the AGA-loop, only the glycine is conserved in all histones, and the majority of histones has a motif that is different from AGA. A subset of histones may be able to form two stacking interactions and is therefore possibly able to form hypernucleosomes, whereas the remaining histones likely do not stack and thus cannot form hypernucleosomes. Also, the histones without stacking interface are partially defective in tetramerization. These histones are somewhat longer than the others, with an extension at the C-terminus. Since DNA interaction is likely possible for all histones, the non-stacking histones may function as a regulator for hypernucleosome extension. All genomes with more than one histone-coding gene encode one non-stacking histone.

Ca. Diapherotrites

The majority of *Ca.* Diapherotrites genomes does not encode histones. Only four histones from *Ca.* Diapherotrites are included, encoded by four different genomes. Remarkably, the AGA-loop is absent in all histones, although the variant GGA was found in one histone. Two histones likely form stacking interactions. All histones are likely to be able to tetramerize and bind DNA. Since no histone contains amino acid residues at positions 15-17 that cause no steric hindrance when multimerizing, and also is predicted to form stacking interactions, it is likely that none of the histones from *Ca.* Diapherotrites assemble into hypernucleosomes.

Ca. Huberarchaeota

Although we included twelve histones from seven huberarchaeal genomes, effectively there are only two histones included in this analysis, since all huberarchaeal histones are identical to either of those two histones. The histone of 91 amino acid residues contain an AGA-loop, whereas the 79-residue histone has a variant of which it is unknown if it allows for hypernucleosome formation. Stacking interactions are identified in both histones, albeit between different residues. The long histone is likely able to form a hypernucleosome, but the analysis is inconclusive for the short histone. Residues of the DNA-binding interface are best conserved in the short histone, whereas tetramerization interactions are best conserved in the long histone. The long histone has an N-terminal tail that is highly positively charged because of lysines, as discussed elsewhere in this chapter, whereas the short histone has a short C-terminal extension with positive charges due to the presence of lysines and arginines. We hypothesize that the long histone is involved in hypernucleosome formation, whereas the short histone is able to regulate the hypernucleosome by binding to its extremities and possibly positioning the hypernucleosome on the DNA.

Ca. Micrarchaeota

Most *Ca.* Micrarchaeota genomes encode two histones, while one-third of the micrarchaeal genomes encodes only one histone. We found a broad spectrum of micrarchaeal histones, with lengths ranging from 59 to 113 residues and theoretical pIs from 5.02 to 10.53. The histones can roughly be divided in two types: type A that is likely able to form a hypernucleosome via stacking interactions, with likely strong DNA-binding- and tetramerization interactions, and type B that lacks stacking interactions and likely has weaker DNA-binding- and tetramerization interactions. Types A and B are not correlated to histone length and pI. Also, the histones of the two types are not equally divided between the genomes, as some genomes encode two histones from the same type (FIGURE 2.8). This suggests contradiction with the hypothesis that the histones of a single organism have different functions, with one being able to form a hypernucleosome and one involved in a different process, for example regulation of the hypernucleosome. It should be noted though that differences on a sequence level between two histones of the same type may result in a yet unidentified from of differentiation. Also, encoding of multiple histones of the same type may be a mechanism for increasing histone expression levels. Still, the majority of genomes encoding two histones encode a histone from either type. Four type B histones have a shorter N-terminal α-helix, which may also contribute to a differential function in genome compaction or transcription regulation.

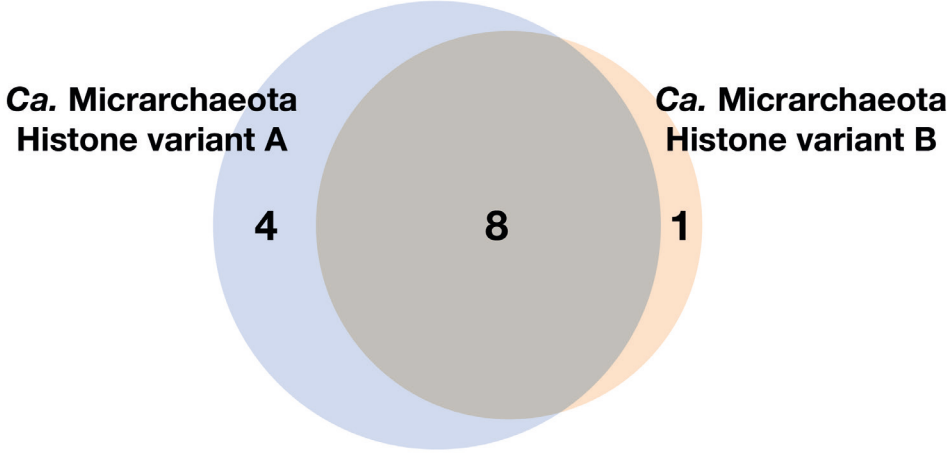


Figure 2.8: Venn diagram of micrarchaeal genomes encoding two histones, and the presence of histone type A (likely forming a hypernucleosome) and type B (likely unable to form a hypernucleosome).

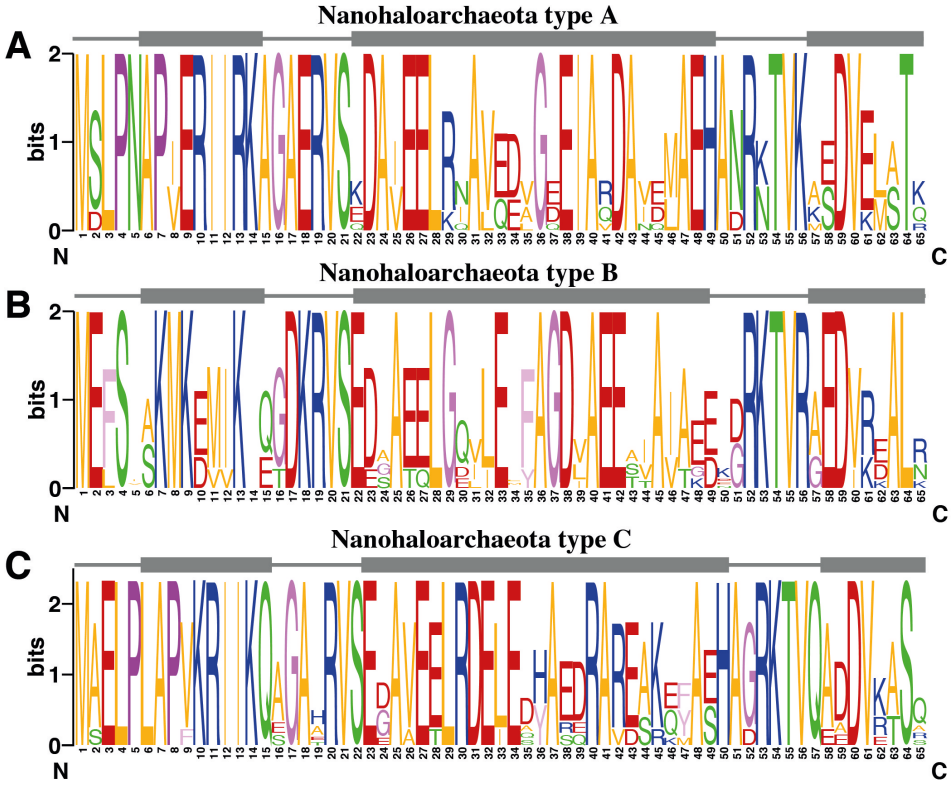


Figure 2.9: Sequence logo of the three types of nanohaloarchaeal histones. A) Type A, which is likely able to form hypernucleosomes. B) Type B, which is unlikely to form hypernucleosomes. C) Type C, which may form hypernucleosomes. Grey bars represent predicted α-helices. Figure made using WebLogo (224).

Nanoarchaeota

Like *Ca.* Micrarchaeota, within the five genomes of Nanoarchaeota that were included in this study, two types can be distinguished. Here, every genome encodes one histone of each type (with Nanoarchaeota archaeon B36_G17, encoding only one histone, being the exception). Histones of type A have a fixed length of 75 residues, whereas histones of type B are always longer. The types are not related to pI, DNA interactions or tetramerization interactions, but they do seem to be related to the presence of potential stacking interactions, with histones of type A generally being expected to form stacking interactions while histones of type B are not. Type B histones also have AGV instead of the AGA-loop, which may prevent hypernucleosome formation for the members of this type. Here, again, histones of type A are potentially involved in hypernucleosome formation, and

histones of type B likely have different structural properties. The type B histone of *Ca. Nanobsidianus stetteri* has a length of 103 residues and is predicted to contain an N-terminal tail. However, an internal methionine at exactly the location where other histones of the same type have the N-terminal methionine, suggests that this gene was misannotated. The internal methionine is in reality likely the N-terminal methionine, and the N-terminal tail is probably never translated.

Ca. Nanohaloarchaeota

Histones of *Ca.* Nanohaloarchaeota are, like histones from other (presumably) salinophilic species, characterized by their low (<6) theoretical pI. Most histones have a length of 65 amino acid residues. Within the nanohaloarchaeal histones, three types can be distinguished. Type A has an intact AGA loop, six DNA-binding interactions and six tetramerization interactions (**FIGURE 2.9A**). Also, it has several tetramerization interactions, although the exact number varies per histone. Therefore, histones of type A can likely form a hypernucleosome (*see also Nanosalina* sp. J07AB43 HA in **FIGURE 2.2** and **TABLE 2.2**). Histones of type B completely lack the AGA loop, and also most histones of type B cannot facilitate stacking (**FIGURE 2.9B**). Additionally, H49, important for tetramerization, is lacking. Therefore, it is highly unlikely that histones of this type are involved in hypernucleosome formation. Amino acids involved in DNA binding are present, although with five interactions, they are slightly smaller in number compared to type A. This may cause slightly weaker DNA binding affinities. Lastly, histones belonging to type C have an AGA loop in which at least the glycine and C-terminal alanine are well conserved (**FIGURE 2.9C**). Also, they are able to form stacking interactions, but less in number than histones of type A. Instead of six DNA-binding and tetramerization interactions, histones of type C probably form five of both. It is unclear if these histones are able to form a hypernucleosome, but if they do, type C hypernucleosomes are likely less stable than those of type A. In this phylum, the stability of the hypernucleosome may therefore be regulated by expressing histones of type A or C. Genomes that encode three histones, encode all three variants. Genomes encoding two histones encode either A and B, A and C or two histones of type B. Five nanohaloarchaeal histones were difficult to categorize. These histones are either shorter or longer than 65 residues, and in some cases their pIs are much different from the other histones.

Ca. Pacearchaeota

A minority (15%) of pacearchaeal genomes encodes histones, while Alba is encoded by most genomes belonging to *Ca.* Pacearchaeota. All pacearchaeal genomes investigated here encode a single histone. The AGA-loop is only present

in one histone, other histones have variants, such as LGA and CGA. Four out of six histones that were investigated here, are likely able to form stacking interactions. However, the histone with the AGA-loop is unlikely to form stacking interactions. This suggests that none of the pacearchaeal histones are able to assemble into a hypernucleosome. In contrast, we concluded from the structural analysis that at least the histone from genome Pacearchaeota CG1_02_31_27 is able to form a hypernucleosome, despite having CGA instead of AGA. We do not expect the other histones to form hypernucleosomes as well: only two out of six histones are likely able to form tetramers; the other four histones likely cannot form structures larger than dimers. Pacearchaeal histones lack potential DNA-binding interactions and are therefore probably unable to bind DNA.

Ca. Parvarchaeota

In the genomes of *Ca.* Parvarchaeota, no genes coding for histones have been identified to date.

Ca. Woesearchaeota

The genomes of *Ca.* Woesearchaeota encode one, or in some cases two histones. These histones can likely bind DNA and form tetramers. Some woesearchaeal histones are probably able to form stacking interactions, although the AGA-loop is conserved in less than half of the histones. Variants of the AGA motif similar to those in *Ca.* Pacearchaeota are often found, which means we cannot rule out the possibility of hypernucleosome formation for the majority of the histones of *Ca.* Woesearchaeota. The histones from this phylum cannot be grouped into types, as seen for some other DPANN histones. There are relatively few conserved residues as compared to other phyla, although length and pI are very constant, considering the variation in sequence.

TACK

Over 90% of TACK genomes encode a single histone; the maximum is three histones encoded by one genome (**figure 2.5A**). Histone length is variable, although almost a third of the TACK histones in this study consist of 73 amino acid residues (**figure 2.5B**). The pI of TACK histones is relatively high, mostly between 9 and 10.5, with an average of 9.19 (**figure 2.5C**).

Ca. Bathyarchaeota

50 out of 60 genomes of *Ca.* Bathyarchaeota in this study encode one histone. The remaining ten genomes encode two or three histone homologs. Bathyarchaeal histones range from 62 to 105 amino acid residues in length, although almost half of the histones in this study consist of exactly 73 residues. Also the most extreme theoretical pIs are far apart (from 5.06 to 11.30), but most pIs lie between 9 and 10.5. More than 85% of the histones have an AGA-loop. Those histones seem to share common features of archaeal histones: DNA binding, tetramerization, and multimerization into hypernucleosomes. The analysis suggests that the remaining group of histones form less DNA-binding interactions and less tetramerization interactions, and are not able to form stacking interactions. Remarkably, most of the histones without conserved AGA-loop are identified as the only histone encoded by the genome it belongs to. This suggests that these histones do not regulate transcription via hypernucleosome modulation, as hypothesized for *Ca.* Huberarchaeota and *Ca.* Nanohaloarchaeota, i.a.. A small subset of bathyarchaeal histones contains an N-terminal tail. The N-terminal histone tails within *Ca.* Bathyarchaeota do not resemble each other, nor do they resemble the tails of eukaryotic histones. Also, two histones have a short C-terminal tail, which also does not resemble any other known histone tail.

Crenarchaeota

Crenarchaeota were long known as the archaeal phylum in which histones do not occur. The chromosomes were thought to be compacted and regulated by Crenarchaeota-specific chromatin proteins such as Cren7 and Sul7 (*see also* **Chapter 1**). Metagenomic sequencing resulted in the identification of 48 genomes that do encode histones. Except for one, all genomes encode a single histone; only Thermoprotei archaeon B132_G9 encodes two. These histones likely bind to DNA and are able to tetramerize. Also, the AGA-loop is present in nearly all crenarchaeal histones, and most are able to form stacking interactions. This means that most histones may be able to assemble into a hypernucleosome. Of all histones with a single histone fold, the ones from Crenarchaeota are on average the longest, which is probably related to their short C-terminal tail. This tail is poorly conserved within the phylum. Additionally, some histone-coding genes encode short N-terminal extensions of unknown function.

Ca. Geothermarchaeota

In *Ca.* Geothermarchaeota, two genomes were identified to encode histones. One genome encodes a long, 83-residue histone, whereas the other encodes two

shorter histones. All geothermarchaeal histones are likely able to tetramerize. The AGA-loop is conserved only in the two short histones. One of these histones is probably also able to assemble into a hypernucleosome, whereas the other cannot do this because it lacks stacking interactions. The histone that likely cannot form a hypernucleosome has six DNA-binding interactions, which possibly results in a high DNA binding affinity. The hypernucleosome-forming histone encoded by the same genome likely has a lower binding affinity, as a result of having only five residues involved in DNA binding. The non-hypernucleosome-forming histone may therefore function as an inhibitor against extension of the hypernucleosome. The long histone is probably unable to multimerize into a hypernucleosome. However, its positively charged N-terminal tail and C-terminal tail may play a role in regulation of DNA compaction and transcription. The N-terminal tail does not resemble its eukaryotic counterpart.

Ca. Korarchaeota

Although most TACK histones do not have homologs in the same genome, *Ca.* Korarchaeota is the only phylum within TACK that has more genomes that encode two histones than those that encode one histone (six and five, respectively). No histone types could be identified. The AGA-loop is conserved in most korarchaeal histones. Stacking interactions are also likely to be formed by these histones, although the combination of residues forming a stacking interaction widely varies

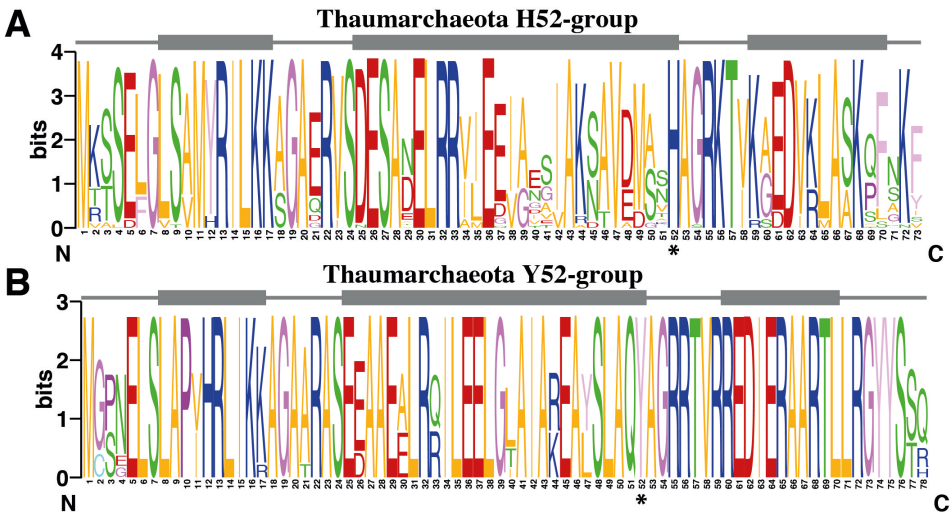


Figure 2.10: Sequence logo of the two groups of thaumarchaeal histones. A) The H52-group, which is possibly able to form hypernucleosomes. B) The Y52-group, which is unlikely to form hypernucleosomes. Asterisks indicate the residue at position 52. Grey bars represent predicted α-helices. Figure made using WebLogo (224).

in this phylum. Based on our analysis, DNA binding is likely conserved in most korarchaeal histones. There are generally less tetramerization interactions than in histones from other phyla, which may make tetramerization weaker in some histones. Therefore, it is not possible to predict hypernucleosome formation based on this analysis. Also, structural analysis proved to be inconclusive in terms of hypernucleosome formation (**TABLE 2.2**).

Ca. Marsarchaeota

In the genomes of *Ca.* Marsarchaeota, no genes coding for histones have been identified to date.

Thaumarchaeota

Of the 78 thaumarchaeal genomes encoding histones that were analysed in this study, 77 genomes encode one histone. Two histones are encoded by the remaining genome, and these histones are almost identical. The AGA-loop was found in most histones. The tetramerization-related histidine at position 52, which is well conserved in most histones throughout the domain, is in 13% of thaumarchaeal histones substituted by tyrosine (**FIGURE 2.10**). Here, remarkably, the histones with Y52 form a group that possibly are able to form stacking interactions, but their weakened tetramerization interactions may affect their ability to form hypernucleosomes. Also, histones of this group possess a short C-terminal extension. Examination of the metadata of the genomes to which the 'Y52-group' histones belong, showed that these genomes were found at deep sea hydrothermal vent sediments, whereas the genomes of the 'H52-group' histones are found in sea water and soil samples. Since this is the only clear difference between the two histone groups, and the Y52-group consists of a much smaller number of histones, we will not assign them to a type, as for other phyla. The vastly different habitat in which the organisms of the Y52-histones live, may demand for other histone properties than the H52-histones. Since the H52Y substitution in the Y52-group is the only element that possibly prevents hypernucleosome formation while all other elements required for forming hypernucleosomes are in place, it is not possible to be conclusive about the hypernucleosome formation abilities based on this sequence analysis. Structural analysis has shown that at least one member of the H52-group likely can assemble into hypernucleosomes (**TABLE 2.2**).

Ca. Verstraetearchaetoa

Of *Ca.* Verstraetearchaeota, only three histones from three different genomes are known. Two of those histones are identical, which effectively means that this

analysis includes only two verstraetearchaeal histones. In these histones, the AGA loop is conserved, and so are probably also its DNA-binding and tetramerization properties. Prediction of the ability of verstraetearchaeal histones to form hypernucleosomes is not possible, since identification of the stacking interactions via sequence analysis is challenging for these histones.

Asgard archaea

A small number of available Asgard archaea genomes leads to a small number of histones available for investigation. However, within this small population of genomes we found a relatively broad range of histones per genome and histone lengths (**FIGURE 2.5A** and **B**). The variety of lengths is partially caused by histone truncates and histones with N-terminal tails. With a value of 9.84, Asgard archaeal histones account for the highest average pI of all superphyla (**FIGURE 2.5C**). Some histones have a pI that comes close to the pI of eukaryotic histones (>11).

Ca. Heimdallarchaeota

Of three genomes of *Ca.* Heimdallarchaeota in this study, one genome encodes one histone, one genome encodes five histones and one genome encodes ten histones. Although few genomes are included, the number of histones per genome varies a lot. Additionally, three genomes were identified, which all encode one gene product annotated as histone. However, these gene products lack all main conserved amino acid residues, such as the HAGRKT-motif and the AGA-loop. Other parts do show resemblance to heimdallarchaeal histones, especially to the central α-helix, which is usually the least-conserved part of the histone. It is highly uncertain if these gene products are really histones, and therefore they were not included in the analysis. The pI of most heimdallarchaeal histones is very high, with an average of 10.05. The N-terminal tail of two of the histones encoded by *Ca.* Heimdallarchaeota archaeon LC_3 has been extensively discussed in this chapter. The histones with tail have a length of 91 residues, whereas the other heimdallarchaeal histones are between 67 and 76 residues long. Heimdallarchaeal histones exhibit the potential interactions that are involved in DNA binding and and tetramerization, except for the histones with N-terminal tail. They lack the histidine of the HAGRKT-motif, which is crucial in tetramerization. The HAGRKT-motif of other histones is conserved but in most cases the alanine is substituted by a serine. The consequenes are not clear, but DNA binding and tetramerization are likely not affected. In general, most heimdallarchaeal histones likely form hypernucleosomes, since they exhibit a stacking interface and potential stacking interactions. However, H49N substitution in histones of archaeon

LC_3 that likely express tails, may interfere with hypernucleosome formation. Also, two histones from archaeon LC_3 (which encodes ten histones) and one histone from archaeon J3-BM-08 (which encodes five histones) do not have an intact AGA-loop and cannot form stacking interactions. These histones likely cannot multimerize into hypernucleosomes.

Ca. Lokiarchaeota

Ca. Lokiarchaeota histones are possibly the group that least resembles histones from other phyla. These histones have mostly variants of the AGA-loop and HAGRKT-motif, and also they possess some but not all possible DNA-binding- and tetramerization interactions that are found in most other histones. Three lokiarchaeal genomes are included in this analysis, archaeons CR_4, GC14_75 and B53_G9, which encode four, five and six histones, respectively. Archaeon B53_G9 encodes three histones that are likely able to form hypernucleosomes and archaeon CR_4 encodes one histone that probably does the same, which is supported by the structural analysis discussed in this chapter. Histones from archaeon GC14_75 likely do not multimerize into hypernucleosomes (140), although these histones are very dissimilar to other histones, which makes it difficult to draw conclusions. Archaeon GC14_75 also encodes a truncated histone, which lacks part of the C-terminal α-helix. The N-termini of histones from archaeon B53_G9 are highly positively charged, and one histone from the same genome has a short N-terminal extension that contains three more positive charges. Also, one histone from archaeon GC14_75 and one from CR_4 have positively charged N-termini. The function of these charged regions remains unclear, but they may be involved in recruitment of other proteins or may play a role in DNA binding.

Ca. Odinarchaeota

Only one histone from *Ca.* Odinarchaeota has been identified to date. The AGA-loop is absent, its HAGRKT-motif is different from other histones (although the residues involved in tetramerization and DNA binding are not affected), and its stacking interface likely disables hypernucleosome formation. The only feature that likely is present, is DNA binding.

Ca. Thorarchaeota

Five genomes of *Ca.* Thorarchaeota encoding one histone have been found. These histones are quite different from each other. Two histones have an N-terminal tail that does not resemble the heimdallarchaeal or eukaryotic N-terminal tails. These histones have AKA instead of AGA at the N-terminal loop, which likely interferes

with hypernucleosome formation. Stacking interactions were identified, as well as DNA-binding interactions and tetramerization interactions. Two other histones do possess an intact AGA-loop, but are missing a histidine at position 49, which makes tetramerization and hypernucleosome formation unlikely. These histones are able to form stacking interactions and probably can bind to DNA. The last histone of five somewhat resembles the thorarchaeal histones with AGA-loop, but does have a conserved H49. This histone is the most likely candidate to form hypernucleosomes, while this is unlikely or uncertain for the other thorarchaeal histones.

Histones in genome regulation

MNase-seq experiments have shown that histones position upstream and downstream of a promoter region (221). This, in combination with knock-out studies showing both up- and down-regulation of transcription levels, leads to the hypothesis that histones are important for transcription regulation in the relatively well-studied phylum Euryarchaeota (117, 185, 237, 238) and may play a similar role in other histone-coding phyla. The exact mechanisms by which histones act in regulation are at this moment largely unknown. What is the mechanistic role of histones in the regulation of gene expression? Is the hypernucleosome, with a mechanism analogous to that in bacterial gene repression, able to block promoter regions and other regulatory elements, thereby making them inaccessible to the transcription machinery (239-242)? In Bacteria, such a mechanism exists for H-NS and partition protein B (ParB) proteins, in which filaments laterally spread from a nucleation site, often a high affinity DNA sequence (123, 243-245). Specific high affinity sites have been identified both *in vivo* and *in vitro* in archaea (164, 220, 221, 246). The role of such high affinity sites may be to position the hypernucleosome on the genome and could be a key feature in archaeal genome regulation. In archaea, cooperative lateral spreading of filaments has been reported for Alba proteins (116, 118, 247, 248). Also, promoter occlusion mechanisms and competitive binding of archaeal NAPs and transcription factors have been reported (117, 249, 250).

In addition, how dynamic are hypernucleosomes, and how does the cell control the size of the hypernucleosome in order for it to be functional? Is up- and down-regulation of histone expression important in fine tuning this process? Another option for control of hypernucleosome size is heteromerization of histone variants with different stacking propensity. Heteromerization of such histone variants, for instance a histone from type A and a histone from type B from *Ca.* Nanohaloarchaeota (**TABLE 2.2** and **FIGURE 2.9**), could restrict hypernucleosome size to fewer subunits. Distinct expression patterns of histone variants or histones

from different types at different growth phases or as a result of environmental cues such as osmolarity (237, 251), may alter the composition and size of the hypernucleosome. However, so far, histone variants and types of histones within a phylum or class have been poorly studied in archaea. The results of our predictions on hypernucleosome formation clearly point out the need for *in vitro* and *in vivo* studies explicitly addressing all of these questions.

Conclusion

Histones from archaea and eukaryotes are similar in tertiary but not in quaternary structure when bound to DNA. While eukaryotic histones form octamers on the DNA, archaeal histones form filaments of variable size: hypernucleosomes. Important residues responsible for DNA binding, dimer–dimer interactions, and stacking interactions are mostly conserved among archaea, including Asgard archaea, Bathyarchaeota, and other newly discovered archaea. In these recently discovered Archaeal phyla, histone tails and truncated histone variants were also found. In terms of evolution, it appears that, based on fragmentary data derived from extant lineages, the hypernucleosome has progressively become more flexible as histones with N-terminal and C-terminal tails and additional terminal helices (like in H2A and H2B in the nucleosome) developed. Furthermore, the appearance of additional DNA-binding residues and positively charged N-terminal tails may have increased the affinity of histones for DNA (252). These changes in dimer structure and DNA affinity may have stabilized octameric nucleosomes and disfavored multimerization. Specifically, the emergence of the eukaryotic H2A-H2B heterodimer blocked hypernucleosome formation since H2A lacks the dimer–dimer interface, and H2B contains an additional helix at its C-terminus that blocks the stacking interface.

The histone tails from *Ca.* Heimdallarchaeota are likely to function in similar ways as those of eukaryotic histones. They are lysine rich and potentially subject to post-translational modification, thereby possibly affecting the histone's interactions with other actors. Alternatively, they may provide stabilization of the hypernucleosome via interactions with DNA in cis or in trans. Since it is believed that eukaryotes share their latest common ancestor with *Ca.* Heimdallarchaeota, eukaryotic histones may have evolved from the predecessors of the tail-containing Heimdallarchaeal histones. As some histone proteins that have an N-terminal tail (*Ca.* Heimdallarchaeota LC_3 HA and *Ca.* Bathyarchaeota archaeon B23) seem to form less stable hypernucleosomes, these histones may represent an evolutionary transition towards a different mechanism of gene regulation, switching from regulation by multimerization and compaction toward regulation by histone tail modifications.

Although the hypernucleosome structure is suggestive of stacking interactions between dimers in adjacent turns, experimental evidence for such interactions has been lacking. However, we provide evidence for such interactions in chapter 4. Also, the functional role of tails, as well as truncates, has yet to be proven experimentally. *In vitro* hypernucleosome reconstitution experiments and *in vivo*

foot-printing assays of species expressing nonstandard histones combined with mutation of the residues proposed to be involved in stacking interactions could answer these questions. Lastly, the existence of post-translational modifications of residues in archaeal histone tails, as well as their effect on transcription regulation, remains to be discovered and would give an important insight into the evolution of transcription regulation and genome folding from archaea to eukaryotes.

Methods

Alignment of archaeal histone sequences

We have included histones from every histone-encoding (candidate) phylum within the archaeal domain in our analysis. We show different histones from the same organism if the predicted stacking properties are very dissimilar. Sequences were aligned with Clustal Omega (253) using default parameters, removing gaps.

Sequence analysis of archaeal histones

All sequences of archaeal histones that were available on 01-01-2019 via the NCBI protein database*. Only sequences of histones were included that had been annotated as histones in the database, and that originate from genomes that were assigned to a superphylum or phylum. A total of 1107 sequences were included, which are listed in the supplemental Excel file. Histone types were identified with the help of Clustal Omega (253) using default parameters.

Analysis of potential hypernucleosome formation

Structural analysis of the selected archaeal histones and assessment of potential hypernucleosome formation was done by inspecting the conservation of residues that are important for multimerization in the published HMfB hypernucleosome structure (197). Comparative multichain modeling was performed in MODELLER (254) using default parameters to construct dimer models of the archaeal histones. These models were superimposed onto HMfB dimers in the hypernucleosome crystal structure to assess whether alternative or additional interactions were possible in the different archaeal histone complexes.

Model of Heimdall HA tails in hypernucleosome

The molecular model of the histone HA dimer from the Heimdallarchaeota LC_3 genome was constructed by multitemplate modeling in MODELLER (254) using otherwise default parameters. The HMfB dimer in the hypernucleosome (197) was used as a structural template for the histone fold and eukaryotic histone H3 and H4 as structural templates for the N-terminal tails. An initial model for the Heimdall HA hypernucleosome was obtained by superimposing the HA dimer model onto HMfB in the hypernucleosome crystal structure, with either an H3-like or an H4-like tail conformation. To optimize the path of the tails through

* https://www.ncbi.nlm.nih.gov/protein

the DNA gyres and remove major steric clashes, the HA dimer model and surrounding DNA was excised from the initial model and water refined separately using High-Ambiguity Driven Docking (HADDOCK) (255), imposing ambiguous interaction restraints between HA residues 14–19 and the surrounding 3-bp section of DNA, using otherwise default parameters.