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

Henneman, B.

CHAPTER 1

Introduction

This chapter is based on:

Bram Henneman & Remus T. Dame, (2015) *Archaeal histones: dynamic and versatile genome architects*. AIMS Microbiology. doi: 10.3934/microbiol.2015.1.72

and

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

Organisms from the domain archaea - one of the three domains of life - share similarities with both bacteria and eukaryotes (1). Genomic comparisons demonstrate that bacteria and archaea share a common ancestor; eukaryotes are to date classified as being part of the archaeal branch (2-4). The archaeal domain comprises single-cellular organisms found in diverse habitats. Although archaea and bacteria have common features, such as a circular genome and the absence of a nucleus, at the genetic level archaea seem to be more related to eukaryotes. Amongst others, archaeal RNA polymerase, a key component of cellular life in all domains, is more similar to RNA polymerase from eukaryotes than bacterial RNA polymerase (5, 6). Archaeal ribosomes share their size and structural core with bacterial ribosomes, but are more similar to eukaryotic ribosomes when it comes to protein and rRNA sequence and some specific domains (7-9). Also, some cellular processes thought to be unique to eukaryotes, such as endosomal sorting and the ubiquitin system, have been identified in some archaea (10).

Chromatin proteins are found in every domain of life. Bacteria express DNA bending and DNA bridging proteins, such as HU and 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. Histones act together with other chromatin proteins, such as linker histones and histone variants in eukaryotes, Alba, MC1 and Sso10a in archaea, and SMC in both domains (13). These proteins jointly shape the genome, which has an effect on transcription regulation and has a role in response to environmental cues.

This thesis focuses on the histones of archaea and the roles that they have in chromatin architecture and transcription regulation. In this chapter, I will extensively introduce archaea as a domain of life, discussing their habitats and their metabolism. Next, I will discuss their genomic features and evolutionary relations to each other, as well as their mechanisms of transcription regulation. In CHAPTERS 2-6 I will further describe the evolution of histones and molecular mechanisms of genome compaction by histones.

Archaea as a domain of life

Evolution and the tree of life

Archaea were identified as a domain that is separate from bacteria by Carl Woese and George Fox (14). They proposed a three-domain model, in which bacteria, archaea and eukaryotes are monophyletic groups sharing a common ancestor (15). As the availability of information on phylogenetic and genomic features of archaea grew, the model was changed to a two-domain model, in which eukaryotes are a sister group of a phylum that is now known as Crenarchaeota (16). Together with Euryarchaeota, these phyla were the only known archaeal phyla for a long time. This changed with the discovery of Thaumarchaeota, *Candidatus* (*Ca.*) Aigarchaeota and *Ca.* Korarchaeota (17-19), which together form the TACK superphylum (an acronym of Thaumarchaeota, Aigarchaeota, Crenarchaeota and Korarchaeota). At that point, eukaryotes were hypothesized to branch from the TACK superphylum. Later, metagenomic sequencing boosted the discovery of new archaeal species and phyla. This resulted in the discovery of the DPANN superphylum (an acronym of Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota and Nanohaloarchaeota), and most recently the Asgard superphylum, which revealed that eukaryotes are most closely related to extant members of Asgard archaea. Recent additions to the archaeal branch of the tree of life include the plyla *Ca.* Heimdallarchaeota, *Ca.* Odinarchaeota, *Ca.* Thorarchaeota and *Ca.* Lokiarchaeota (belonging to Asgard archaea) (20, 21), *Ca.* Huberarchaeota, *Ca.* Micrarchaeota, *Ca.* Pacearchaeota, *Ca.* Woesearchaeota and *Ca.* Mamarchaeota (belonging to DPANN) (22-27), *Ca.* Bathyarchaeota, *Ca.* Geothermarchaeota (also

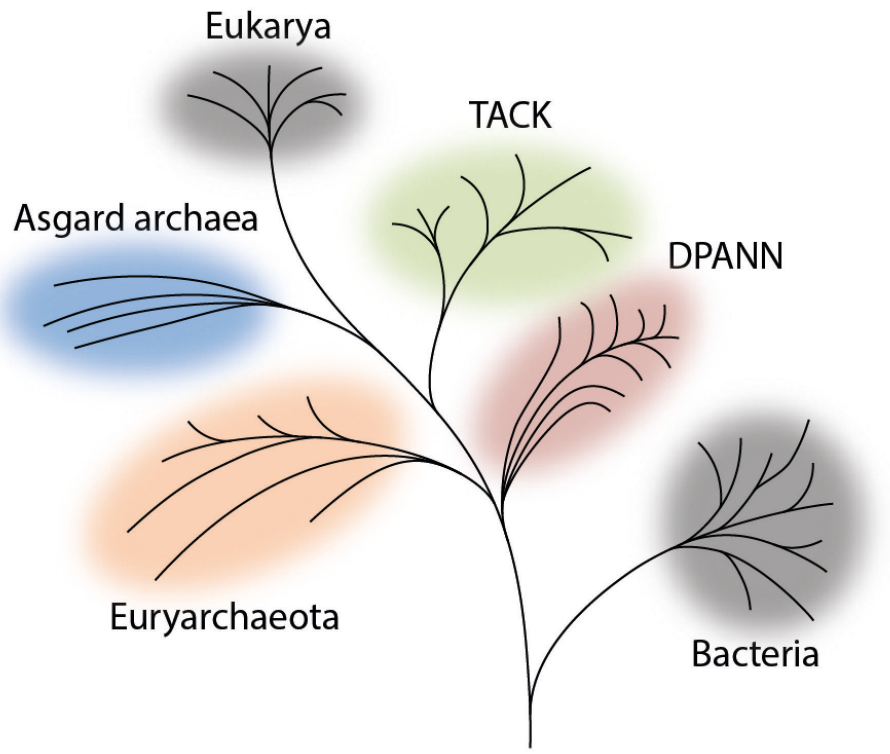

Figure 1.1: Cartoon of the tree of life, with emphasis on archaea. The tree is rooted at the last universal common ancestor (LUCA), and shows the two domains: bacteria and archaea/eukarya. The archaeal superphyla Asgard archaea, TACK and DPANN, and the phylum Euryarchaeota, together represent all archaeal genomes that are known to date. Modified from: Eme *et al.* Nat. Rev. Microbiol. 2017 (47).

referred to as *Ca.* Geoarchaeota), *Ca.* Verstraetearchaeota and *Ca.* Marsarchaeota (belonging to TACK) (28-31), and the classes Methanonatronarchaeia, Theoinarchaea and Hadesarchaea (belonging to Euryarchaeota) (32-34) (**FIGURE 1.1**). Most of the more recently discovered archaeal phyla have the *Candidatus* status, since their existence is thus far solely based on metagenomic sequencing, and they are categorized based on their 16S ribosomal RNA sequence. For now, only phyla with at least one member strain maintained in microbiological culture get a definitive phylum name (35-37). Also, the categorization of archaeal phyla is not set in stone, as new sequencing data and insights into archaeal evolution may change the layout of the archaeal tree of life. Especially the categorization of DPANN is subject to discussion (38). While most studies regard DPANN as a separate and deep-branching superphylum within archaea (39-44), others suggest that some or all DPANN lineages are part of Euryarchaeota, and that their categorization as part of the DPANN superphylum is an artifact (45). Also the position of *Ca.* Lokiarchaeota has been challenged (46).

The study of archaeal evolution is the key to understanding eukaryogenesis, the evolutionary events that have led to the emergence of eukaryotes from prokaryotes. Eukaryogenesis is best understood by identifying the first eukaryotic common ancestor (FECA) and the last eukaryotic common ancestor (LECA) (47). FECA can be defined as the archaeon that is the oldest ancestor of all eukaryotes that are living today. LECA is the organism of which all now-living eukaryotes are descendants. Although all present-day eukaryotes are descendants of both FECA and LECA, not all descendants of FECA and LECA are living today, and FECA may have had archaeal descendants that have gone extinct. FECA and LECA are phylogenetic entities, identified as nodes in the tree of life (47). The discovery of more archaea that are closely related to eukaryotes and deep-branching eukaryotes can add branches to the tree of life, and as a result, FECA and LECA can move closer together. The narrowing of the gap between FECA and LECA provides us with information about when and how the first eukaryote emerged. Also, it sheds light on the order in which eukaryotic characteristics, such as nucleus and mitochondria, first appeared in the evolutionary predecessors of eukaryotes. It is therefore important to identify common eukaryotic features, in order to determine the cellular processes of LECA. It has been established that LECA lived 1 to 1.9 billion years ago and had a nucleus containing linear chromosomes, encoding genes with introns (48-50). Histones played a role in the transcription regulation system, together with an RNAi system and small non-coding RNAs (48, 51). Cell motility machinery involving an actin- and tubulin-based cytoskeleton and motor proteins was present in LECA, just like mitochondria that allow for aerobic respiration (48). The mitochondria were descendants from alphaproteobacteria, which have once been engulfed by an ancestor of LECA, as described by the endosymbiotic theory (52, 53). The membrane of LECA was composed of phospholipids made of fatty acids and *sn*-glycerol-3-phosphate (G3P), just like present-day eukaryotes and bacteria (54). In contrast, archaeal membranes consist of phospholipids based on isoprenoids and *sn*-glycerol-1-phosphate (G1P). This 'lipid divide' was initially used as an argument for the three-domain theory, but it does not rule out the existence of archaea with a G3P-based membrane. It has been suggested that some archaea, including the Asgard archaea member phylum *Ca.* Lokiarchaeota, may have the ability to produce eukaryotic-like membranes, based on the presence of genes coding for enzymes involved in the metabolism of G3P and fatty acids (55-57). However, this has not been confirmed experimentally.

Like the composition of the membrane, other features of LECA may have been inherited from FECA and its ancestors. Of the estimated 4000 genes of LECA, 550-1100 genes are hypothesized to be of prokaryotic descendance; 18-37% of those genes have their origin in archaea. This does not seem like a large propor-

tion, although it should be noted that also some archaeal genes are of bacterial origin. Furthermore, the recent discovery of new phyla belonging to TACK and Asgard archaea has revealed that some components of processes thought to occur exclusively in eukaryotes, are also present in some archaea (47). Most Asgard archaea contain some, if not all, components of the eukaryotic trafficking and membrane remodeling machinery (including ESCRT-machinery), cytoskeleton, oligosaccharyl-transferase system and ubiquitin system, as well as some informational proteins and eukaryotic-like ribosomal proteins (21). Fluorescence *in situ* hybridization and catalyzed reporter deposition (CARD-FISH) experiments have shown that *Ca.* Lokiarchaeota cells occur in a wide spectrum of shapes and sizes, which probably reflects the expression of cytoskeletal and membrane remodeling proteins (58). To a lesser extent, the ribosomal, ubiquitin-related, trafficking-related and informational proteins were also identified in *Ca.* Verstraetearchaeota, Crenarchaeota, *Ca.* Korarchaeota, *Ca.* Bathyarchaeota and Thaumarchaeota (47, 59-62). This suggests that FECA likely contained the before-mentioned 'eukaryotic' components, and may have passed these on to its eukaryotic descendants.

Habitats and metabolism

Archaea were initially known as organisms that live at - from a human perspective - extreme conditions, such as hot springs, hydrothermal vents, hypersaline lakes and highly acidic water. Later, archaea were also found in oceans, soil and intestines of humans and cattle. Now, archaea are considered to inhabit nearly every possible habitat on earth, from permafrost to sewage and from ocean floors to goldmines (63). Furthermore, archaea occupy a large array of niches on or in the human body (64). The role of archaea in the human microbiome has largely been undiscovered, although knowledge of archaeal presence on or in humans is rapidly expanding. Recently, *Ca.* Woesearchaeota species were found in human lungs, Thaumarchaeota members were found on skin, and Crenarchaetoa and Euryarchaeota were found in intestines (65, 66). Surprisingly, to date, no real archaeal pathogens have been identified, but it was suggested that *Methanobrevibacter oralis* may be involved in periodontitis and brain abscesses (67, 68). Archaea are also suspected to be involved in inflammatory bowel diseases (69). Additionally, in their capacity of improving the microbiome, the role of archaea as probiotics is investigated, and archaea may help against trimethylaminuria and cardiovascular diseases (70). Other animals also live in symbiosis with archaea, as has been shown in studies on farmhouse animals, nonhuman primates, macropodidae, birds, reptiles, arthropods and sponges (66). Methanogens account for the vast majority of archaea in animal gastrointernal tracts, with *Methanobrevibacter*, *Methanobacterium* and *Methanosphaera* (all Euryarchaeota)

as main contributing families.

Archaea not only live in symbiosis with eukaryotes, some are co-dependent on other archaea. A well-studied example is *Nanoarchaeum equitans* (Nanoarchaeota), which can only be grown in co-culture with *Ignicoccus hospitalis* (Crenarchaeota) (71, 72). These organisms may exchange membrane lipids, amino acids and enzymes involved in several metabolic pathways (71, 72).

Archaea play an important role in biogeochemical cycles together with other life forms. Via their unique and highly variable metabolic pathways, they recycle carbon, nitrogen and sulfur, as well as other elements (73). Euryarchaeota, Crenarchaeota and Thaumarchaeota are relatively well characterized in terms of metabolism, as they can be cultured. In other phyla, some information about their metabolic pathways can be deduced from the enzymes that their genomes encode. Archaea are an important contributor to production and consumption of greenhouse gases such as carbon dioxide and methane (73). Significant amounts of methane are known to be produced by oceanic archaea (including archaea living inside oceanic plankton) and archaea living inside cattle. Most Euryarchaeota are anaerobic and autotrophic, producing organic compounds from carbon dioxide and bicarbonate (74). Also, some Crenarchaeota and Thaumarchaeota are autotrophs. Mixotrophic organisms can be found in the phyla Crenarchaeota and Euryarchaeota (73). Chemoorganotrophic archaea reduce organic compounds to carbon dioxide or partially oxidized compounds such as organic acids and alcohols. These types of archaea are identified as Euryarchaeota and Crenarchaeota. Nanohaloarchaeota, *Ca.* Korarchaeota, *Ca.* Geoarchaeota, *Ca.* Aigarchaeota, *Ca.* Micrarchaeota, *Ca.* Parvarchaeota and *Ca.* Marsarchaeota can likely also be regarded as chemoautotrophs, and can be anaerobic or (facultatively) aerobic (30, 73, 75). Phototrophic organotrophs, which get their energy from sunlight, have also been identified (76).

A large group of euryarchaeal species are methanogenic, producing methane with hydrogen gas as electron donor and carbon dioxide as electron acceptor (77). The reverse process, methane oxidation, is done by the anaerobic methanotrophs (ANME) group of Euryarchaeota (78, 79). ANME usually use sulfate as electron acceptor, but some use nitrite or iron- or manganese-containing compounds instead (80). Some archaea are thereby also coupled to biogeochemical cycles of metals.

Archaea are mainly involved in the nitrogen cycle via the oxidation of nitrate, ammonia or nitrogen-containing organic compounds to nitrite (81). However,

in the absence of those compounds, some anaerobic methane-oxidizing Euryarchaeota can use nitrogen gas as an electron donor (82). Denitrification, the reverse process (nitrite or nitrate to nitrogen gas), can also be done by some archaea (83). Few examples of archaea capable of producing nitrous oxide are currently known (84).

Some Crenarchaeota and Euryarchaeota contribute to the sulfur cycle via sulfidogenesis: the anaerobic production of hydrogen sulfide (85). A number of sulfur-containing compounds are used as electron acceptors, most notably elementary sulfur, sulfite, thiosulfate and sulfate. Sulfur oxidation, in which the sulfur compound is the electron donor, is observed in some crenarchaeal and euryarchaeal species, which use the iron-containing compounds pyrite, marcasite and chalcopyrite (86).

Transcription regulation in archaea

Like in any organism, archaea regulate transcription in order to adapt to their environment. The archaeal transcription machinery contains some elements that resemble eukaryotic transcription machinery, while other elements are more similar to their bacterial counterparts. Organisms from all domains express RNA polymerase (RNAP), a well-conserved and multi-subunit enzyme responsible for transcribing DNA into mRNA (87). The structure of archaeal RNAP resembles that of eukaryotic RNAPII, with most of the 13 subunits matching. Also, the initiation- and elongation-related stalk domain, absent in bacterial RNAP, is present in archaeal RNAP and resembles that of eukaryotes (88, 89). Transcription occurs in three stages: initiation, elongation and termination, together forming the transcription cycle. All three stages are regulated, although to date, the most information regarding transcription regulation is available on transcription initiation (88, 90). It should be noted that for historic reasons, almost all research regarding archaeal transcription regulation has been done on organisms belonging to Euryarchaeota and Crenarchaeota, which are also the easiest classes to cultivate. Although homologs of transcription machinery elements are present in genomes belonging to DPANN and Asgard archaea, the exact transcription mechanisms in these superphyla have yet to be unveiled.

Basal transcription regulation

On a basal level, archaeal regulation of transcription initiation is similar to its eukaryotic equivalent (90-92). Two DNA elements are required for transcription

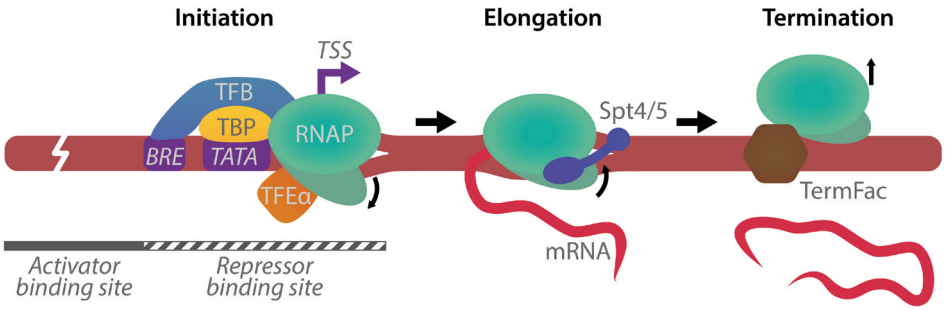

Figure 1.2: The archaeal transcription cycle. At the initiation stage, TATA-binding protein (TBP) binds the TATA-box (TATA), and is stabilized by transcription factor B (TFB) that binds to the B recognition element (BRE) and to TBP. This complex, situated just upstream of the transcription start site (TSS), recruits RNA polymerase (RNAP), which melts the DNA, forming the open complex. Open complex formation is sometimes assisted by transcription factor E (TFEα), which leaves during early elongation. Transcription factors with a repressive or activatory role can bind the promotor region at or (far) upstream of the initiation-related DNA elements, respectively. During elongation, elongation factors Spt4 and Spt5 stabilize the clamp domain of RNAP, thereby preventing dissociation of the elongation complex. Transcription is likely terminated by termination factors (TermFac) or by intrinsic termination due to weak base pairing at dedicated DNA sequences (not shown). RNAP dissociates from the DNA and can start a new transcription cycle. DNA elements are written in italics.

initiation: the TATA-box and the B recognition element (BRE). Both are located within the first 50 bp upstream of the transcription start site (TSS), in the promoter region (FIGURE 1.2). These elements require binding of general transcription factors (GTFs) in order to start transcription (90, 91). TATA-binding protein (TBP) binds to the TATA-box, and Transcription factor B (TFB) binds to the BRE and to TBP, thereby stabilizing the TBP-TATA complex (93). The assembly of TBP and TFB on DNA in the preinitiation complex (PIC) is a prerequisite for RNAP recruitment. Unlike eukaryotic transcription initiation, no energy is required by the PIC to separate DNA strands and form a transcription bubble (94). However, the non-essential GTF transcription factor E (TFEα) assists promoter opening by inducing the opening of the RNAP clamp (95, 96). Additionally, the promoter proximal DNA element (PPE), situated just downstream of the TSS, stabilizes TFB binding at weak promoters, and is able to provide additional promoter selectivity. Also, the initiator element (Inr) can increase the strength of the promoter, and in some cases it is bound by transcription activators. Both PPE and Inr are not essential for transcription initiation.

Not much is known about elongation regulation in archaea. The universally conserved protein Spt5 (also called Spt5 in eukaryotes; NusG in bacteria) with its conserved binding partner Spt4, and transcription factor S (TFS) are the only elongation factors that have been identified (97). TFS was hypothesized to be

the main elongation factor, with Spt4/5 as a secondary factor (98). TFS enhances RNAP's endonucleolytic cleavage activity, which allows removal of stalled elongation complexes, thereby making way for other elongation complexes. Spt4/5 replaces the GTFs at promoter escape or during early elongation and enhances processivity (99, 100). Also, Spt4/5 closes the clamp domain of RNAP, presumably resulting in a more stable elongation complex, which reduces dissociation from the DNA (95). In addition, both TFS and Spt4/5 allow for faster processivity of RNAP through histone-induced barriers (98). The interplay between elongation factors and histones provides a possible mechanism of transcription regulation.

Besides early transcription termination facilitated by elongation factors, regular transcription termination is regulated via DNA elements and likely also via termination factors. Intrinsic termination occurs when base pairing between the DNA template and the mRNA becomes too weak. Similar to eukaryotes, the main cue for archaeal intrinsic termination is a series of adenines, which results in a poly-U tract in the mRNA (101, 102). Subsequent destabilization of the RNA-DNA hybrid results in transcription termination. In some cases however, a stem-loop structure helps termination, like in bacteria (103). Not much is known about factor-mediated termination, because homologs of bacterial and eukaryotic termination factors have never been identified (90). Eta (euryarchaeal termination activity) from *Thermococcus kodakarensis* is the only known termination factor to date in archaea (104). Eta is conserved in most archaeal lineages, and therefore it may be functional as a transcription terminator throughout the domain.

Advanced transcription regulation

On a more advanced level, transcription is regulated by transcription factors (TFs), which, in addition to GTFs, are necessary for transcription initiation at specific genes and operons. These TFs resemble their bacterial counterparts (105). TFs function as activators or repressors, and together with the GTFs, they promote or block transcription initiation (106). The promotive or repressive function of a TF can be deduced from the position of their binding site: if it overlaps the BRE, TATA-box or TSS, it likely prevents PIC formation (91, 93). In case a TF's binding position is located more upstream of the BRE, it may recruit RNAP, TBP or TFB, resulting in transcription activation (**FIGURE 1.2**). However, more complex mechanisms of transcription regulation have also been found, which involve for example autoactivation and autorepression by the same TF, complex formation by repressors in order to block RNAP recruitment, and activation of alternative TSSs within the coding region (107-109). Many TFs are able to respond to environmental cues by binding ligands, such as sugars or amino acids (105). Ligand

binding subsequently enhances or decreases DNA binding affinity of TFs, allowing them to bind to or dissociate from repressor- or activator binding sequences. Some TFs are able to bind a large set of ligands, thereby gaining or losing affinity to defined promoters, while others bind only one ligand and operate more binary (110, 111). When groups of genes related to one physiological trait need to be activated or silenced, this is in general done by one TF, controlling the so-called regulon. Regulon-related TFs are almost always bifunctional, which allows them to promote one set of genes while repressing another set with an opposite physiological effect (91).

Like in bacteria, there are a few master regulators that control a large group of genes. All TFs and genes that have a transcriptional relation, operate in a gene regulatory network (GRN). In these networks, paralogs of GTFs play an important role. More than 70% of all archaeal genomes contain at least two genes encoding the same GTF: TBP or TFB (112). Specific combinations of GTF paralogs promote transcription of defined genes (113). Combining TBP and TFB paralogs results in a number of possible combinations that is equal to the product of the number of paralogs. This allows archaea to use some combinations for standard growth conditions, whereas other combinations may specifically be used during stress. Also, some promoters have unique GFT paralog combinations. In addition, also TFE has different paralogs in some species. Especially the genomes of the phylum Thaumarchaeota and the euryarchaeal class Halobacteria encode a large number of GTF paralogs (91).

Transcription regulation via chromatin architecture

Archaea have more options when it comes to gene regulation. Small RNAs (sRNA) are highly abundant and play a role in GRNs (114). sRNAs have been studied in Euryarchaeota, Crenarchaeota and Nanoarchaeota. Also, the covalent modification of proteins by small archaeal modifier proteins (SAMPs), a process known as SAMPylation, can be regarded as regulatory mechanism (115). Like ubiquitination in eukaryotes, SAMPylation targets proteins for degradation. Furthermore, post-translational modification (PTM) of proteins by adding or removing functional group can alter a protein's function. Another important regulatory factor is the modulation of chromatin architecture. Archaea express nucleoid-associated proteins (NAPs) and histones (together called chromatin proteins) that are able to change chromatin formation by bending, wrapping or bridging DNA (13, 116). Modulating genome architecture can result in activation or repression of genes and operons (117). For example, changes in chromatin conformation may result in blockage of the promoter region or inhibition of the transcription elongation

complex. In contrast, certain chromatin protein-induced DNA conformations may allow DNA-bound activators to facilitate RNAP recruitment, leading to transcription initiation. Some chromatin proteins are highly conserved among archaea, while others are unique to one or a few phyla or classes. Chromatin proteins are highly abundant and usually do not bind at specific DNA sequences (116). Therefore, they may affect transcription regulation globally, which would resemble transcription regulation via chromatin conformation in bacteria. However, structural data on archaeal genome architecture *in vivo* and its dynamics is mostly lacking.

A widespread NAP, encoded in a subset of genomes from all archaeal superphyla, is Alba (acetylation lowers binding affinity). Alba is able to bridge two DNA duplexes. Depending on the presence of a phenylalanine at position 60 (present in the variant Alba1, but absent in Alba2), Alba is able to form filaments along DNA via dimer-dimer interactions along the DNA (118). DNA binding by Alba can be modulated by trimethylation of the lysine at position 16, which lowers DNA-binding affinity (119, 120). In bridging mode, Alba may facilitate loop formation. In bacteria, loop formation by H-NS functionally organizes the genome, and in archaea, Alba may have a similar role (121). Its lateral filament mode may repress transcription via binding across promoter regions. Also, bridging by H-NS is able to cause repression of transcription, which is arguably the favored model for the involvement of H-NS in transcription regulation, as bridging can be modulated (122, 123).

In the TACK-phylum Crenarchaeota, many NAPs are expressed, such as Sul7, Cren7, CC1, Sso10a and Sso7c (124). Cren7 is best conserved within Crenarchaeota, whereas Sso10a can also be found in some classes of Euryarchaeota. The abundance of different NAPs in Crenarchaeota may be related to the absence of genes encoding histones in most genomes of the phylum. Cren7, Sul7, and likely also Sso7c, organize chromatin by DNA bending, whereas Sso10a has bending, bridging and lateral filament modes (125-127). The role of CC1 in chromatin architecture has yet to be elucidated (128).

Methanogen chromosomal protein 1 (MC1) is another NAP that is presumably involved in functional organization of the archaeal genome. MC1 induces sharp kinks into the DNA (129). It is encoded by genomes in the euryarchaeal classes Halobacteria, Methanomicrobia and Hadesarchaea, and by the genome of the DPANN-phylum Altiarchaeota.

Homologues of bacterial NAPs were identified in some archaea. Dps (DNA-

binding protein from starved cells), which protects DNA during oxidative stress by condensing chromatin in rosette-like structures, can be found in some species belonging to Euryarchaeota and Thaumarchaeota (130-132). Also, some genomes spread across the archaeal tree of life encode homologs of the bacterial DNA bender HU (133). It seems likely that these proteins have been acquired via horizontal gene transfer.

The most conserved protein related to chromatin architecture, however, is the histone. It has been found in almost every phylum of archaea (*see also* **CHAPTER 2** of this thesis), and is believed to be a major contributor to DNA conformation on a local and global level (134). The archaeal histone resembles the well-studied eukaryotic histone in both sequence and in structure of the histone fold, the core of the histone monomer. However, there are also some clear differences between histones from eukaryotes and archaea. For a better understanding of these differences, I introduce the eukaryotic histone first, before comprehensively discussing the archaeal histone.

The eukaryotic histone

In eukaryotes, octameric histone cores compact DNA by wrapping an approximately 150 bp unit twice around its surface, forming a nucleosome (135, 136). Nucleosomes interact with each other, yielding an additional level of DNA organization in the form of a fiber. Besides a role in compaction, histones also play roles in genome organization, DNA replication, DNA repair and gene expres-

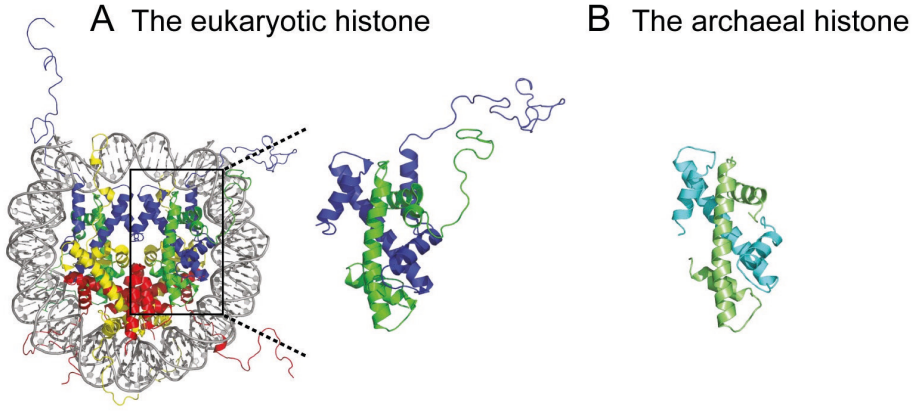

Figure 1.3: Eukaryotic and archaeal histones. A) Eukaryotic nucleosome consisting of DNA wrapped around a core of a (H3-H4)2 tetramer and two H2A-H2B dimers. Zoom-in of the black box shows a H3-H4 histone dimer. Yellow: H2A; red: H2B; blue: H3; green: H4. B) Archaeal histone homodimer of HMfB.

sion, which highlights the nucleosome as a very important complex affecting a vast array of cellular processes. Characteristic of core histone proteins is a common 'histone fold': two short and one long α-helix, separated by loops (137-141). In eukaryotes, the histone core consists of two H2A-H2B dimers and a H3-H4 tetramer, around which ~146 bp of DNA is wrapped 1.65 times (**FIGURE 1.3A**). It has been suggested that smaller histone assemblies, such as tetrasomes (H3-H4 tetramers), hexasomes (H3-H4 tetramers plus one H2A/H2B dimer) and hemisomes (a H3-H4 dimer plus one H2A/H2B dimer), have functional roles as intermediate structures during for example transcription elongation (142-145). The linker histone H1 (which lacks the characteristic histone fold) binds at the entry and exit points of the DNA wrapped around the octameric histone core (146, 147). The association of histone H1 constrains an additional 20 bp of DNA and helps formation of the 30-nm fiber, which results in tighter compaction (148, 149). Also, flexible N-terminal tails that protrude from eukaryotic histones contribute to a tighter DNA packaging. These tails may interact with either the DNA or the histone surface of another nucleosome, which facilitates close association of nucleosomes (150-152). Furthermore, post-translational modifications of amino acid residues in the N-terminal tails, such as acetylation, methylation, phosphorylation, ubiquitination and biotinylation, are a key instrument for the cell to regulate gene expression, the DNA damage response and many other processes (153-155). For instance, while heterochromatin (tightly packed DNA) is typically devoid of acetylated lysines, euchromatic (lightly packed) regions typically contain histones with acetylated lysines. Histone acetylation is believed to cause a locally less condensed chromatin structure *in vivo*, which is permissive to transcription. Thus, euchromatin contains actively transcribed genes. In particular the lysine-rich histone H4 tail seems to be crucial in the modulation of chromatin structure (150). *In vitro*, H4 tails are required for higher order chromatin folding (156-158), which can be disrupted by acetylation of K16 (150). Nucleosome function and level of genome compaction can be altered in a multitude of ways, providing flexible and versatile mechanisms for tuning the cell's chromatin structure and transcription regulation.

Archaeal histones: structure and alignment

Histones are the only archaeal chromatin proteins that resemble the eukaryotic histones H3 and H4 in structure and sequence. These proteins occur in almost every archaeal phylum, although histone-coding genes are not conserved throughout all orders of every phylum (159). The best-studied histones are HMfA and HMfB, two histones from the euryarchaeon *Methanothermus fervidus*, which are closely related in terms of sequence (160). These histones occur in solution

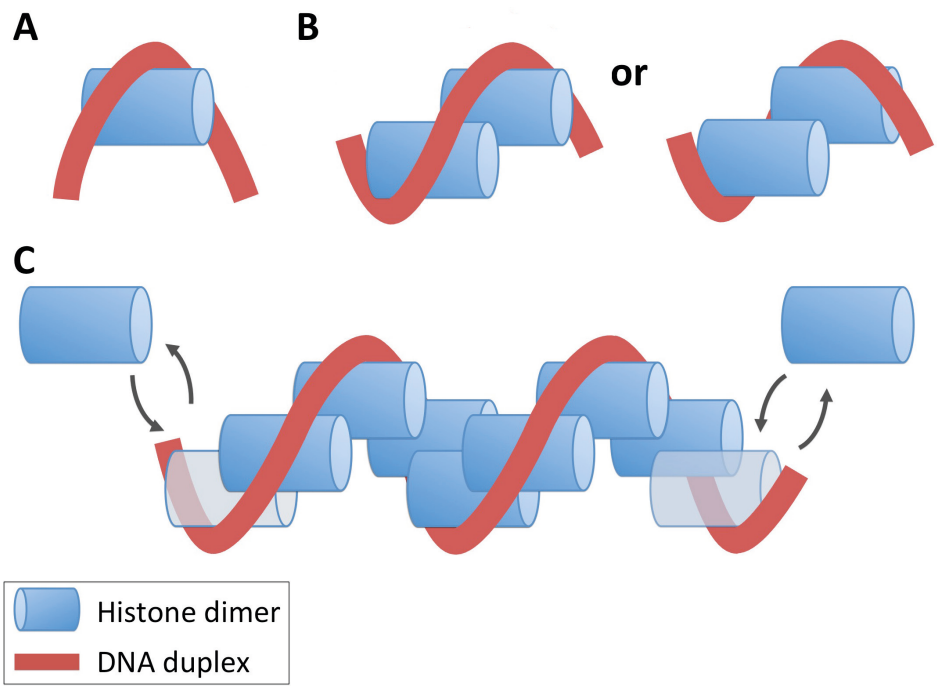

Figure 1.4: Model for multimerization of histones in *Thermococcus kodakarensis* and *Methanothermobacter thermautotrophicus*. Blue: histone dimers; red: DNA duplex. A) DNA is bent by a histone dimer, which covers 30 bp of DNA. B) A histone tetramer wraps the DNA in a left- or right-handed manner and is able to switch between these configurations, for example in response to changes in salt concentrations (172). C) Histone dimers can further associate to form a larger multimer. Every additional dimer wraps another 30 bp of DNA. At both extremities of the multimer, dimers can be added or removed. The left-handed wrapping depicted here is arbitrary, since also this large histone-multimer-DNA complex might accommodate both left-handed and right-handed wrapping.

as homo- or heterodimers (FIGURE 1.3B), but have been shown to bind to specific DNA sequences as tetramers, possibly resembling the (H3-H4)2 tetrasomes formed as intermediates during histone octamer assembly in eukaryotes (161). Archaeal histone tetramers were shown already in early studies to protect ~60 bp of DNA from micrococcal nuclease (MNase) digestion (162) and to assemble at specific high affinity sequences (163, 164). Conversely, MNase digestion of chromatin from the euryarchaeal species *Thermococcus kodakarensis* yields digests with fragments 60 bp in size, representing a tetrameric histone, but also fragments of 30 bp and multiples of 30 bp up to ~500 bp. This observation suggests that 'nucleosomes', which likely protected DNA from MNase digestion, formed by histone HTkB are not of defined size and that dimers, tetramers and larger multimers are functionally relevant (either in chromatin organization or in relation to transcription regulation) (165). Our lab proposed a model for multimerization

of histone proteins, in which histone dimers can be added and removed from both ends to assemble into a DNA-coiling multimer, similar in structure to the eukaryotic nucleosome, that consists of dimers that each cover 30 bp of DNA (140) (**FIGURE 1.4**). Studies on *Methanothermobacter thermautotrophicus* chromatin are in accordance with this observation, which suggests that the basic functional unit of archaeal histones might not be a tetramer, as generally assumed, but a dimer (165, 166). The observed differences between species might be due to differences in genomic sequence and/or protein sequence, altering dimer-dimer interactions. Although sequence homology between histones is sometimes limited, the histone fold is well conserved between species (167). Histones consist of three α-helices separated by short β-loops. Residues R10, K13, R19, K53, T54 and K56, located in the N-terminal α-helix, N-terminal β-loop and the C-terminal β-loop, were, based on crystal structure and mutagenesis studies, predicted to be mainly responsible for interacting with the DNA backbone. Dimer-dimer interaction, tetramerization, has been proposed to be mediated by residues E42, L46, H49, D59, L62 and R66 in the C-terminal and middle α-helix and the C-terminal β-loop (137, 160) Euryarchaeal histones have a high degree of sequence identity and are very similar in length, although some have an extended N-terminal loop and/or helix, or an extended C-terminus of ~30 amino acids. It has been proposed that these unconventional histone proteins are modulators of DNA binding by interaction with other histones (167, 168). Also, some histones from halophilic archaea consist of two histone domains fused together, whereas other histones assemble into dimers following synthesis (169). In these naturally occurring histone-histone fusion proteins, the histone fold at the C-terminus is more similar to the histone fold of other euryarchaeal histones than the one at the N-terminus (167, 170). HTkB and HMtA2, which likely bind along the genome as a dimer, tetramer or larger multimer in *T. kodakarensis* and *M. thermautotrophicus*, respectively, contain all residues identified as responsible for tetramerization in HMfB from *M. fervidus*. Therefore the difference in minimal size of the functional histone unit cannot be attributed to residues known to be involved in tetramerization.

Binding motif

Archaeal histones have a preference for binding GC-rich sequences with alternating (G/C)2/3 and (A/T)2/3 motifs, which are separated by half a helical turn. This compresses the minor and major groove on one side of the helix with A:T facing towards the histones, and G:C facing outwards, causing the DNA to bend. This mode of binding is very similar to that of eukaryotic histones. Studies on HMfB from *M. fervidus* reveal that binding motifs are more complex: AT and GC are not equal to TA and CG, respectively, in terms of facilitating archaeal

nucleosome assembly (164). Also, when analyzing substitution patterns in the DNA of nucleosomes in sister lineages of *Haloferax volcanii*, it was found that there is a preference for G:C near the dyad, whereas changes towards A:T are more common further away from the dyad, near the ends of the nucleosome DNA (172, 173). These sequence patterns are again very similar to human nucleosome positioning sequences. (174, 175).

In eukaryotes, a histone wraps 147 bp of DNA 1.65 times around its core, by means of which the DNA is compacted and supercoils are constrained (176). In archaea, tetrameric histones have been reported to wrap DNA without making a full turn, which results in a horseshoe-like conformation (177). However, interactions with the DNA may determine the extent of wrapping, which means DNA sequence motifs may affect the exact number of turns that is made by the DNA. HMf proteins from *M. fervidus* can wrap DNA around its tetramer core at high affinity sequences, but also are able to bend DNA as a dimer *in vitro* (178). Wrapping can be considered an advanced form of bending, rather than a separate mechanism, and may be an evolutionary consequence of recruitment of dimers at adjacent sites due to high affinity sequences or protein-protein interactions between dimers. *In vivo* bending and wrapping by histones is expected to occur in parallel at different locations along the genome. Not only wrapping by histones, but also loop formation by trans-acting elements such as Alba contributes to DNA organization and compaction. Loop formation may stabilize topologically isolated domains in which supercoiling is preserved, and in which a subset of genes can be co-regulated. (121, 179, 180).

Evolution of the histone protein

It has been suggested that eukaryotic histones evolved from archaeal histones (181). This hypothesis is supported by the high similarity at the amino acid sequence level and in secondary structure (182, 183). Suggestive of an archaeal origin of eukaryotic histones is also the dimeric nature of archaeal histones; archaeal histone complexes are built from dimers, but members of the archaeal class halobacteria express a 'tandem histone'. In these tandem histones, the histone folds are linked end-to-end (184-186). This implies that the histone folds always occupy the same position and role in the naturally linked dimer. This leads to the relaxation of evolutionary constraints in parts of the histone, an example of subfunctionalization (187, 188). According to this hypothesis, the histone folds further evolved in a divergent way, leading to an asymmetric dimer. This may have been an ancestor of H3-H4, which later separated to become two individual proteins and corresponding genes (182). It has been suggested that H2A and H2B have arisen

from H3 and H4 later on in histone evolution (182). Indeed, H3 and H4 are more similar to archaeal histones than H2A and H2B, supporting this hypothesis. From this point in time, eukaryotic histones have further evolved into histone variants: highly homologous substitutes of canonical eukaryotic histones, which often play a specialist role in a wide variety of cellular processes (189). Unlike canonical histones, which are mainly expressed during DNA replication, histone variants are expressed in a replication-independent manner (190, 191). Histone variants of H2A and H3 are widely known and studied, whereas only few examples have been found of diversified H2B and H4 (192). The evolutionary pressure for the evolution of dimer-based histones to octameric histones and their subsequent variants is believed to be DNA compaction (182). The fact that eukaryotic cells undergo mitosis, in which chromosomes are highly compacted, together with the abundance of gene-poor regions may have favored a histone conformation that wraps DNA twice (eukaryotic octamer) instead of once (archaeal tetramer), and that via its N-terminal tails has the ability to compact DNA at a higher order. By understanding the structural properties and functionality of archaeal histones, we can learn how evolution led to the eukaryotic histone.

In this thesis, I will further discuss archaeal histones in terms of structure, function, and their relation to DNA. **Chapter 2** highlights the interactions that histones have in a complex with other histones and DNA. Also, in **chapter 2**, all currently known (proposed) superphyla and phyla are discussed in terms of encoding and possibly expressing histones and histone variants. Furthermore, based on the protein sequences of these histones, I predict what kind of complexes they may form together with DNA, and what consequences this may have for the organism's genome architecture. In **chapter 3**, I describe a protocol for Tethered Particle Motion (TPM), a single molecule technique which is able to provide a read out of DNA conformation. Addition of histones and other DNA-binding proteins may change the DNA conformation. TPM is the main technique used in this thesis. Results of TPM experiments are described in **chapters 4**, **5** and **6**. In **chapter 4**, I focus on the largest histone-DNA assembly known so far: the hypernucleosome. Using TPM, as well as magnetic tweezers, I show at what concentrations hypernucleosomes of HMfA and of HMfB, the best-studied archaeal histones, form, and how they obtain their stability. **Chapter 5** describes histones from the candidate phylum Nanohaloarchaeota, which have never been reported to have been purified. I use TPM and microscale thermophoresis to characterize these histones. Finally, in **chapter 6**, I show how specific high affinity sites on the DNA alter the binding of histones to DNA using TPM. This chapter also provides an extensive overview of the possible functions of histones and histone variants in archaea. **Chapter 7** is a general discussion.