Core Histones of the Amitochondriate Protist, *Giardia lamblia*

Gang Wu,* Andrew G. McArthur,† András Fiser,‡ Andrej Šali,‡ Mitchell L. Sogin,† and Miklós Müller*  

*Laboratory of Biochemical Parasitology, The Rockefeller University, New York; †The Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, Massachusetts; and ‡Laboratory of Molecular Biophysics, The Rockefeller University, New York

Genes coding for the core histones H2a, H2b, H3, and H4 of *Giardia lamblia* were sequenced. A conserved organism- and gene-specific element, GRGCCGAGATTVGG, was found upstream of the coding region in all core histone genes. The derived amino acid sequences of all four histones were similar to their homologs in other eukaryotes, although they were among the most divergent members of this protein family. Comparative protein-structure modeling combined with energy evaluation of the resulting models indicated that the *G. lamblia* core histones individually and together can assume the same three-dimensional structures that were established by X-ray crystallography for *Xenopus laevis* histones and the nucleosome core particle. Since *G. lamblia* represents one of the earliest-diverging eukaryotes in many different molecular trees, the structure of its histones is potentially of relevance to understanding histone evolution. The *G. lamblia* proteins do not represent an intermediate stage between archaean and eukaryotic histones.

Introduction

Histones are basic structural proteins that play an important role in DNA organization and gene regulation in eukaryotes. Histone-like proteins play a similar role in archaea of the Euryarchaeota lineage (Isenberg 1979; Pereira and Reeve 1998). Histones are classified into five main types, the core histones H2a, H2b, H3, H4 and the linker histone H1. Two molecules each of the four core histone types are arranged in an octameric structure. Around this octamer, a 146-bp segment of DNA is coiled (Hayes, Clark, and Wolffe 1991). The amino-terminal tails of the core histones interact loosely with DNA and histone H1. The octameric histones and DNA make up the nucleosome core particle, the unit of chromatin organization (Luger et al. 1997). The three-dimensional (3D) structure of the vertebrate nucleosome core particle has been elucidated by X-ray crystallography at 2.8 Å resolution (Luger et al. 1997), confirming earlier results obtained at lower resolution (Arents et al. 1991).

The universal role of histones in eukaryotic chromatin is reflected by their remarkable conservation. The core histones are regarded as one of the most conserved protein families. All four core histones contain a region that forms the easily recognized histone fold, consisting of three α-helices connected by short loops (Luger et al. 1997). The histone folds represent a major part of these proteins. Behind this structural conservation lies an extreme sequence conservation (Isenberg 1979; Wells 1986; Thatcher and Gorovsky 1994; Makalowska et al. 1999). Vertebrate H3 and H4 sequences are almost identical, and the identities across most known H4 genes exceed 95%.

No histones are found in bacteria or in many archaea, which have a chromosome organization different from that of eukaryotes (Sandman, Pereira, and Reeve 1998). Histone-like proteins have been observed, however, in archaea belonging to the lineage Euryarchaeota, comprising methanogens and related organisms (Sandman, Pereira, and Reeve 1998). These proteins are smaller than eukaryotic histones and correspond essentially to the eukaryotic histone fold region. Even in this homologous domain, the two groups of proteins are markedly divergent. Archaean histone-like proteins do not show a differentiation into the four types observed in eukaryotes. Analysis of NMR data complemented with structural modeling revealed that while divergent in sequence, histone-like proteins of the archaeon *Methanothermus fervidus* form typical histone folds (Starich et al. 1996), which are essentially the same as the crystallographically determined 3D structures of all four vertebrate histone types (Luger et al. 1997). The limited but clear sequence similarity and essentially identical 3D structures of eukaryotic histones and archaeal histone-like proteins show that these proteins derive from common ancestral molecules (Sandman, Pereira, and Reeve 1998).

Although numerous eukaryotic histone molecules have been studied, those of unicellular eukaryotes (protists) have received limited attention (Makalowska et al. 1999). Since the dominant part of eukaryotic diversity is represented by protists (Patterson and Sogin 1992), a more detailed exploration of protists promises to shed light on the limits placed on divergence among eukaryotic histones. In this paper, we present the sequences, the predicted 3D structures and their energy evaluations, and the promoter of the core histones in the highly divergent amitochondriate eukaryote *Giardia lamblia* (Sogin et al. 1989; Adam 1991; Upcroft and Upcroft 1998). Molecular phylogenetic analyses have identified *Giardia*, and related diplomonads, as an early diverging eukaryotic lineage that may have retained some primitive characteristics of the first nucleated cells (Leipe et al. 1993; Stiller and Hall 1997; Roger et al.
1999; but see Embly and Hirt 1998; Stiller et al. 1998; Hirt et al. 1999). Here, we examine the predicted structure of *Giardia* DNA-binding proteins and find them to be similar to those of other eukaryotes and not intermediary between eukaryotes and Archaea.

**Materials and Methods**

**Organism and Genomic Clones**

*Giardia lamblia* strain WB clone 6 (ATCC 30957) was used throughout this study. The *G. lamblia* genome project (www.mbl.edu/Giardia; Henze et al. 1998) provided the gDNA clones used in this study.

**Clone Sequencing and Sequence Evaluation**

A number of random gDNA clones obtained by the *G. lamblia* genome project contained complete core histone genes as indicated by BLASTX (Altschul et al. 1997) annotation of single-pass sequencing reads on the project’s web page (www.mbl.edu/Giardia). One clone for each of the four core histones was resequenced on both strands and analyzed further. These clones were CI0342 for H2a, NF0363 for H2b, CI0136 for H3, and MD0903 for H4. Motif searches of all *G. lamblia* genome project sequences and NCBI nucleotide databases were performed using the GREF utility of the SEALS package (Walker and Koonin 1997).

**Alignment of the Individual Histone Genes**

The derived amino acid sequences of the *G. lamblia* open reading frames (ORFs) were first manually aligned with homologous sequences from public databases with the ED program of the MUST package (Philippe 1993). The alignments were subsequently refined with a method that considers the putative 3D structures of the molecules (Sali and Blundell 1993). In view of the high conservation of core histones within most eukaryotic lineages, only one sequence each was selected to represent the animals, plants, and fungi. Protists were sampled more exhaustively. Histone-like proteins from *Methanothermus fervidus* and *Pyrococcus* sp. were used as representatives of archaean homologs and were aligned only with H4, which is most similar to archaean histone-like proteins.

**Comparative Modeling and Model Evaluation**

Comparative 3D models of the histone subunits and the putative nucleosome core particle of *G. lamblia* were constructed by the program MODELLER-5 (Sali and Blundell 1993; Sánchez and Sali 1997) based on the crystallographic structure of the reconstituted *Xenopus laevis* nucleosome core particle (PDB code 1aoi) (Luger et al. 1997). The input to the program is the alignment of the target sequence to be modeled with the known template structures. The output obtained without any user intervention is a 3D model of the target with all nonhydrogen atoms. The multiple-sequence alignments for modeling were calculated by the MALIGN command of MODELLER-5. The accuracy of the models was subsequently evaluated with the program ProsaII (Sippl 1993). The Z score given by this program approximates the free energy of a model related to that of a random structure, expressed in units of standard deviation. The more negative the Z score, the more accurate is the model likely to be. Larger structures tend to have more negative Z scores. The terminal parts of the models that did not overlap with the template structures were omitted from the evaluations. To judge the significance of the Z scores of the model, Z scores of the crystallographic structures of the template *Xenopus* histones were also calculated.

**Sequence Availability**

The nucleotide sequence data reported here have been submitted to the GenBank database under accession numbers AF139873–AF139876.

**Results**

**Derived Amino Acid Sequences**

The ORFs, uninterrupted by introns, contained in the *G. lamblia* histone H2a, H2b, H3, and H4 genes corresponded to putative translation products of 124, 130, 146, and 99 amino acid residues, respectively (fig. 1). The derived sequences were similar in length to their eukaryotic homologs. As is the case for all histones, the corresponding proteins were rich in positively charged amino acids. Calculated pI values were 10.48 for H2a, 9.38 for H2b, 10.58 for H3, and 10.79 for H4.

The alignment of each gene with its homologs from other organisms was unambiguous. The only exceptions were the amino-termini of H2a and H2b and the carboxyl-terminus of H2a. Remarkably high amino acid conservation was noted in the part of the sequences corresponding to the three α-helices forming the histone fold (table 1). For each core histone, the fungus-plant-animal group showed the highest intragroup amino acid identity. The *G. lamblia* homologs, compared with this group, were consistently less similar. Other protists exhibited various, often intermediate, levels of amino acid identity. Histones H2a, H2b, and H3 of the kinetoplastid protist *Trypanosoma cruzi* showed the lowest identity with their *G. lamblia* homologs, revealing the greatest divergence in the sample studied. Interestingly, the identities of the *G. lamblia* and *T. cruzi* histones with the proteins from other organisms were in the same range. Alignment of the *G. lamblia* core histones with archaean histone-like proteins was very difficult but revealed the conservation of a number of residues (Pereira and Reeve 1998). Eukaryotic histones clearly represent a separate group from the archaean histone-like proteins.

Compared with other eukaryotic H2a sequences, *G. lamblia* histone H2a showed unique deletions of three amino acid residues at positions 113–115. The divergent regions were located mostly at two termini of the protein.

*Giardia lamblia* H2b had a six-unique-amino-acid-residue insertion in the loop between the α-1 and the α-2 helices. The rather divergent amino-terminal regions were of various lengths in different species but were
**Fig. 1.**—Alignment of the derived amino acid sequences of *Giardia lamblia* core histones with a selected set of homologs from other eukaryotes. Residues identical to those in the first sequence are indicated with dashes; deletions are indicated with empty spaces. The three α-helices forming the histone fold are underlined or marked at the top (H4). “V” designates an arginine side chain that in the *Xenopus* nucleosome is inserted into the DNA minor groove (Luger et al. 1997). These are shifted by one residue in *G. lamblia* H2b. In H3, one of these is replaced by a lysine. “O” designates residues that are involved in acetylation in the *Xenopus* nucleosome (Luger et al. 1997). Database accession numbers are given after the species names.
consistently rich in lysine and arginine. The carboxy-terminal region of the *G. lamblia* H2b, similar to the *Entamoeba histolytica* homolog (Sánchez, Enea, and Eichinger 1994), is somewhat longer than other H2b sequences.

The H3 sequences are similar over their entire length. In *G. lamblia* H3, there is an insertion of two residues, but in a somewhat uncertain position, in the loop between helices α-1 and α-2 and a one-residue deletion close to the carboxyl-terminus. For the best alignment, short gaps had to be introduced into the amino-terminal extensions of protist H3 histones. *Giardia lamblia* H3 is distinguished by an eight-amino-acid extension at its carboxyl-terminus.

Histone H4 is the most conserved core histone, and *G. lamblia* H4 was no exception. Single-residue gaps were required for the best alignment of the amino-terminal region. The 10-residue insertion in *E. histolytica* H4 (Binder et al. 1995) was not noted in other H4 sequences.

Of the seven arginine residues in *Xenopus* histones known to be inserted into the DNA minor groove (Luger et al. 1997), six were conserved in *G. lamblia* and one was replaced by lysine (fig. 1). Similarly, almost all lysines known to serve as acetylation sites (Luger et al. 1997) are present in *G. lamblia*.

### Protein Structure Modeling

Three-dimensional models of the four *G. lamblia* core histones were built by comparative modeling based on the crystallographic structure of the *X. laevis* nucleosome core particle (Luger et al. 1997) (fig. 2). The four available template folds from the vertebrate nucleosome core particle, corresponding to the four core histones, were used independently as templates to obtain four different models for each of the four histone sequences in *G. lamblia*. Accuracies of the models were quantified by the ProsaII Z score (Sippl 1993) (table 2). The actual structures of the *G. lamblia* subunits are expected to be most similar to those vertebrate structures that result in the best comparative models as evaluated by the Z score. Furthermore, because the best comparative models have Z scores close to those for the template structures, the

---

**Table 1**

Amino Acid Identities (as percentages) Between Histones H2a, H2b, H3, and H4 of Various Organisms

<table>
<thead>
<tr>
<th></th>
<th><em>Giardia lamblia</em></th>
<th><em>Trypanosoma cruzi</em></th>
<th><em>Tetrahymena thermophila</em></th>
<th><em>Saccharomyces cerevisiae</em></th>
<th><em>Triticum aestivum</em></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>H2a</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>T. cruzi</em></td>
<td>35.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>T. thermophila</em></td>
<td>54.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>S. cerevisiae</em></td>
<td>50.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>T. aestivum</em></td>
<td>53.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>H. sapiens</em></td>
<td>54.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>H2b</strong></td>
<td><em>G. lamblia</em></td>
<td><em>Entamoeba histolytica</em></td>
<td><em>T. cruzi</em></td>
<td><em>S. cerevisiae</em></td>
<td><em>T. aestivum</em></td>
</tr>
<tr>
<td><em>E. histolytica</em></td>
<td>50.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>T. cruzi</em></td>
<td>37.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>T. thermophila</em></td>
<td>48.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>S. cerevisiae</em></td>
<td>53.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>T. aestivum</em></td>
<td>45.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>H. sapiens</em></td>
<td>54.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>H3</strong></td>
<td><em>G. lamblia</em></td>
<td><em>Trichomonas vaginalis</em></td>
<td><em>E. histolytica</em></td>
<td><em>T. cruzi</em></td>
<td><em>S. cerevisiae</em></td>
</tr>
<tr>
<td><em>T. vaginalis</em></td>
<td>57.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>E. histolytica</em></td>
<td>63.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>T. cruzi</em></td>
<td>41.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>T. thermophila</em></td>
<td>64.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>S. cerevisiae</em></td>
<td>58.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>T. aestivum</em></td>
<td>63.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>H. sapiens</em></td>
<td>58.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>H4</strong></td>
<td><em>G. lamblia</em></td>
<td><em>T. vaginalis</em></td>
<td><em>E. histolytica</em></td>
<td><em>T. thermophila</em></td>
<td><em>S. cerevisiae</em></td>
</tr>
<tr>
<td><em>T. vaginalis</em></td>
<td>75.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>E. histolytica</em></td>
<td>74.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>T. thermophila</em></td>
<td>77.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>S. cerevisiae</em></td>
<td>79.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>T. aestivum</em></td>
<td>77.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>H. sapiens</em></td>
<td>77.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Methanococcus fervidae</em></td>
<td>21.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Pyrococcus sp.</em></td>
<td>GB-3A Hpy1</td>
<td>27.4</td>
<td>30.1</td>
<td>27.4</td>
<td>32.3</td>
</tr>
</tbody>
</table>

NOTE.—Matrices are calculated for the histone fold region from α-1 to α-3. The numbers of positions analyzed were 62 in H2a, 63 in H2b, 67 in H3, and 66 in H4. Histone-like proteins of archaea were compared with H4.
Fig. 2.—Comparative protein structure models of the *Giardia lamblia* core histones (red) based on the known structures of the *Xenopus* histones (blue) (Luger et al. 1997). The models and their evaluations indicate that the sequences of the *G. lamblia* are consistent with the structure of the corresponding *Xenopus* histones, with the exception of their terminal extensions. The noted longer insertion of a six-residue loop in H2b protrudes into the solvent, in the monomer as well as in the octamer complex, without any significant interactions with the rest of the protein.
suggested that those proteins have similar functions, as in other eukaryotes.

Although the *Giardia* histones show some divergent features, these only marginally exceed those observed in the histones of other protists (Bender et al. 1992; Sadler and Brunk 1992; Födinger et al. 1993; Sánchez, Enea, and Eichinger 1994; Binder et al. 1995; Marinets et al. 1996; Galanti et al. 1998). In essence, our results expanded the known sequence space explored by histones in eukaryotic diversification but indicated no unique position for the *Giardia* histones. Histone H1, which does not form part of the nucleosome core particle and plays a role in the linking of separate nucleosomes (Garrard 1991), has not been detected by the *Giardia* genome project so far.

The conservation of core histones is a consequence of their role in maintaining chromatin organization and gene regulation. Most of the sequence divergence was found in the amino- and some carboxyl-terminal regions, while the α-helix regions and the loops between α-helices retained highly conserved structures. The six-amino-acid insertion in *Giardia* H2b is located just after an α-1 helix without disrupting the main structure. The unique eight-residue extension at the carboxyl-terminus of H3 does not form an α-helix but may affect the interaction of H3 and H4.

Structure modeling shows that *Giardia* histones, individually and together, can assume 3D structures indistinguishable from those established for vertebrate histones and nucleosomes. Thus, the divergence noted in their amino acid sequences does not exceed the structural constraints imposed by their chromatin organization and gene regulation.

The upstream motif, GRGCCKETTTVGG, detected in all *Giardia* core histone genes differs from conserved upstream motifs that have been found in various other organisms such as the yeast *Schizosaccharomyces pombe* (Matsumoto and Yanagida 1985), the nematode *Caenorhabditis elegans* (Roberts, Emmons, and Childs 1989), and the green alga *Chlamydomonas reinhardtii* (Fabry et al. 1995). This motif is possibly organism- and gene-specific and may play a role in assuring a correlated transcription of the histone genes.

The protist *Giardia* is a typical eukaryote. Several of its features, however, reveal it to be one of the most divergent representatives of its group (Adam 1991; Upcroft and Upcroft 1998). It contains no morphologically or biochemically recognizable mitochondria and has a fermentative core metabolism, several enzymes of which are not found in typical mitochondriate eukaryotes (Brown et al. 1998; Müller 1998; Sánchez et al. 1999). In spite of this great biological divergence, the core histones of *Giardia* were found to be typical for a eukaryote. This finding supports the notion that the typical eukaryotic chromatin organization was present in the common ancestor of all eukaryotes and underwent only minor adjustments during their diversification (Starich et al. 1996). The highly unusual chromatin structure of dinoflagellates is an important exception, which probably arose secondarily (Vernet et al. 1990).

Histone-like proteins are found also in archaea of the Euryarchaeota lineage (Pereira and Reeve 1998). These proteins, which form nucleosomes (Starich et al. 1996), are smaller than eukaryotic histones and correspond to their central domain, i.e., the histone fold. The 3D structures of the archaeal histone-like proteins and of the nucleosome are virtually identical to those seen in eukaryotes (Starich et al. 1996). These structural similarities and the clear homology of archaeal and eukaryotic histones demonstrate their shared ancestry (Slesarev et al. 1998). At the same time, the lack of diversification...
into four separate types and the absence of amino- and carboxy-terminal extensions in archaeal histone-like proteins clearly separate this group of proteins from eukaryotic histones. So far, no extant organism has been found that would display characters intermediate between archaean and eukaryotic histones. The highly divergent protist *G. lamblia* is no exception.

## Acknowledgments

We thank Hilary G. Morrison (Woods Hole, Mass.) for providing the clones for sequencing, and Hervé Philippe and Philippe Lopez (Orsay, France) for the MUST package. We also thank Lidya B. Sánchez, Katrin Henze, and Jennifer A. Lee for help and advice. Oligonucleotide synthesis and DNA sequencing at the Rockefeller University were performed by the Nucleic Acid Sequencing Facility. This research was supported by U.S. Public Health Service National Institutes of Health grants AI11942 to M.M., AI43273 and GM32964 to M.L.S., and GM54762 to A.S. A.S. is a Burroughs Welcome Fellow. A.F. is a Sinsheimer Scholar and an Alfred P. Sloan Research Fellowship. M.L.S., and GM54762 to A.S. A.S. is a Sinsheimer Scholar and an Alfred P. Sloan Research Fellowship. This research was supported by National Science Foundation grant BIR-9601845 to A.S. A.F. is a Burroughs Welcome Fellow. Sinsheimer Scholar and an Alfred P. Sloan Research Fellowship. This research was supported by U.S. Public Health Service National Institutes of Health grants AI11942 to M.M., AI43273 and GM32964 to M.L.S., and GM54762 to A.S., as well as National Science Foundation grant BIR-9601845 to A.S. A.S. is a Sinsheimer Scholar and an Alfred P. Sloan Research Fellow. A.F. is a Burroughs Welcome Fellow.

## LITERATURE CITED


ROGER, A. J., O. SANDBLOM, W. F. DOOLITTLE, and H. PHILIPPE. 1999. An evaluation of elongation factor 1α as a


Geoffrey McFadden, reviewing editor

Accepted April 11, 2000