The Histone Database

Steven Sullivan, Daniel W. Sink[1], Kenneth L. Trout[1], Izabela Makalowska[1], Patrick M. Taylor, Andreas D. Baxevanis[1] and David Landsman*

Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Building 45, Room 6AN12J, 45 Center Drive, MSC 6510, Bethesda, MD 20892-6510, USA and [1]Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Building 49, Room 4A-22, Bethesda, MD 20892-4470, USA

Received September 20, 2001; Accepted September 28, 2001

ABSTRACT

Histone proteins are often noted for their high degree of sequence conservation. It is less often recognized that the histones are a heterogeneous protein family. Furthermore, several classes of non-histone proteins containing the histone fold motif exist. Novel histone and histone fold protein sequences continue to be added to public databases every year. The Histone Database (http://genome.nhgri.nih.gov/histones/) is a searchable, periodically updated collection of histone fold-containing sequences derived from sequence-similarity searches of public databases. Sequence sets are presented in redundant and non-redundant FASTA form, hotlinked to GenBank sequence files. Partial sequences are also now included in the database, which has considerably augmented its taxonomic coverage. Annotated alignments of full-length non-redundant sets of sequences are now available in both web-viewable (HTML) and downloadable (PDF) formats. The database also provides summaries of current information on solved histone fold structures, post-translational modifications of histones, and the human histone gene complement.

INTRODUCTION

Eukaryotic DNA associates periodically with histone proteins to form nucleosomes, the 'beads on a string' seen in classic electron micrographs of chromatin fibers (1). Each nucleosome consists of ~147 bp of DNA wrapped around a compact octameric core containing two molecules each of histones H2A, H2B, H3 and H4. Histones H2A/H2B and H3/H4 dimerize through the non-covalent interaction of a highly conserved structural motif found in each core histone called the histone fold, which also serves as the primary histone–DNA binding domain (2,3). Along with the handshake-like pairing of H2A/H2B and H3/H4 histone folds, other polar and hydrophobic interactions, notably between H2B and H4 and between two H3 molecules within the octamer complex, maintain the integrity of the octamer (4). Histones H1 and H5, which are structurally dissimilar to the core histones, bind as monomers to nucleosomes and internucleosomal linker DNA to facilitate higher-order DNA compaction (5).

Evolutionarily constrained by the profound functional importance of the nucleosome, the octameric core histones are among the most slowly evolving, highly conserved proteins known (6). At the extreme is H4, which differs by only two residues in cow and pea (7). Yet, core histone diversity is manifest at several levels. Histone H2A, H2B, H3 and H4 sequences form four distinct classes. Sequence-variant subclasses are discernable in all but H4 (reviewed in 8). These include the 'replacement' (i.e. replication-independent) histones that are functionally redundant with the major forms, and the functionally distinct H2A subclass, H2A.Z, which are specifically required in early development (9). On the other hand, inter- and intraspecies 'microvariation', small differences in largely identical sequences, is also evident and probably relates to the extensive histone gene duplication seen in many eukaryotes (10). Microvariant histones are assumed, though in most cases not proved, to be functionally equivalent.

Histone folds are also found in archaeal histones (11,12) and in several types of non-histone proteins, including the TATA-binding factor associated factors (TAFs) which are part of the transcription coactivator TFIID complex (reviewed in 13). Strikingly, it has been shown that four yeast TAFs form an octamer in an analogous fashion to the four core histones, though this TAF octamer does not appear to be DNA binding (14). The evolutionary relationships between histone fold-containing proteins (including the core histones) are obscured by the motif's roots in deep time; e.g. one cladistic analysis suggests that H2A and H4 histone folds had already diverged prior to the emergence of eukaryotes (15).

FEATURES OF THE HISTONE DATABASE

For several years we have maintained an online Histone Database (http://genome.nhgri.nih.gov/histones/) to serve those interested in the family of histone fold-containing proteins. The heart of the site is a searchable collection of histone and histone fold

*To whom correspondence should be addressed. Tel: +1 301 435 5981; Fax: +1 301 480 2918; Email: landsman@ncbi.nlm.nih.gov

Present address:
Izabela Makalowska, Department of Biology and The Life Sciences Consortium, The Pennsylvania State University, University Park, PA 16802–5301, USA

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

Table 1. Histone sequence content of Histone Database

	Redundant full-length	Redundant partial	Total redundant	Total non-redundant full-length	Species
H1/H5	400 (321)	37	437	178 (140)	71
H2A	458 (392)	56	514	167 (133)	72
H2B	451 (374)	40	491	178 (139)	71
H3	474 (426)	475	949	111 (99)	253
H4	324 (311)	106	430	72 (62)	99
	2107 (1824)	714	2821	706 (573)	330 (91)

Numbers in parentheses represent the content prior to the August 2001 update.
Species counts are not cumulative since one species can be represented in multiple histone classes.

protein sequences from GenBank. In this paper we report a complete revision of the sequence content by conducting new PSI-BLAST searches (16) of the protein databases with human and yeast query sequences to generate lists of redundant and non-redundant FASTA records for each histone type, as well as taxonomic information about the dataset (Table 1). GenBank and the protein sequence databases themselves are highly dynamic databases, where sequences are constantly being revised as well as added and deleted. Even the highly conserved core histones reflect this dynamism: the number of new non-redundant full-length (NFL) core sequences since our last update (17) ranges from 10 for H4 to 34 for H2A, while the total number of NFL histone sequences has increased by 133 (23%). Our presentation of non-redundant FASTA records has been revised to list all definition lines associated with a given sequence. As before, FASTA records are hotlinked to their GenBank files. With this revision, we have begun to include partial histone sequences, effectively increasing the taxonomic coverage of the database from 91 to 330 species. The sequence content of the Histone Database henceforth will be updated more frequently using software we have developed to scan GenBank updates specifically for histones.

We have also revised the presentation of the CLUSTALX sequence alignments (18). These have been replaced by new alignments of full-length non-redundant sets in HTML (for web viewing) and PDF (for downloading) formats. Physico-chemical properties of sequences are color-coded, and regions suspected of having arisen from frameshift errors are highlighted. In the HTML view, each aligned sequence is hotlinked to its full definition line in the non-redundant FASTA set.

We have retained multiple search modes (e.g. by type, species and sequence) for the site, and updated the lists of solved crystal structures of histone fold-containing proteins, post-translational modifications of histones, and the human histone gene complement. While we have also updated the list of eukaryotic non-histone histone fold proteins, archaeal histones have been replaced by a link to a web site devoted entirely to those proteins (http://www.biosci.ohio-state.edu/~microbio/Archaealhistones/index.html).

REFERENCES

1. Olins,A.L. and Olins,D.E. (1974) Spheroid chromatin units (v bodies). *Science*, **183**, 330–332.
2. Arents,G., Burlingame,R.W., Wang,B.C., Love,W.E. and Moudrianakis,E.N. (1991) The nucleosomal core histone octamer at 3.1 Å resolution: a tripartite protein assembly and a left-handed superhelix. *Proc. Natl Acad. Sci. USA*, **88**, 10148–10152.
3. Arents,G. and Moudrianakis,E.N. (1995) The histone fold: a ubiquitous architectural motif utilized in DNA compaction and protein dimerization. *Proc. Natl Acad. Sci. USA*, **92**, 11170–11174.
4. Luger,K., Mader,A.W., Richmond,R.K., Sargent,D.F. and Richmond,T.J. (1997) Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*, **389**, 251–260.
5. van Holde,K.E. (1989) *Chromatin*. Springer-Verlag , New York.
6. Thatcher,T.H. and Gorovsky,M.A. (1994) Phylogenetic analysis of the core histones H2A, H2B, H3, and H4. *Nucleic Acids Res.*, **22**, 174–179.
7. DeLange,R.J., Fambrough,D.M., Smith,E.L. and Bonner,J. (1969) Calf and pea histone IV. 3. Complete amino acid sequence of pea seedling histone IV: comparison with the homologous calf thymus histone. *J. Biol. Chem.*, **244**, 5669–5679.
8. Brown,D.T. (2001) Histone variants: are they functionally heterogeneous? *Genome Biol.*, **2**, e6.
9. Faast,R., Thonglairoam,V., Schulz,T.C., Beall,J., Wells,J., Taylor,H., Matthaei,K., Rathjen,P.D., Tremethick,D.J. and Lyons,I. (2001) Histone variant H2A.Z is required for early mammalian development. *Curr. Biol.*, **11**, 1183–1187.
10. Stein,G.S., Stein,J.L. and Marzluff,W.M. (1984) *Histone Genes: Structure, Organization, and Regulation*. John Wiley & Sons, New York.
11. Pereira,S.L. and Reeve,J.N. (1998) Histones and nucleosomes in Archaea and Eukarya: a comparative analysis. *Extremophiles*, **2**, 141–148.
12. Sandman,K., Soares,D. and Reeve,J.N. (2001) Molecular components of the archaeal nucleosome. *Biochimie*, **83**, 277–281.
13. Gangloff,Y.G., Pointud,J.C., Thuault,S., Carre,L., Romier,C., Muratoglu,S., Brand,M., Tora,L., Couderc,J.L. and Davidson,I. (2001) The TFIID components human TAF(II)140 and *Drosophila* BIP2 (TAF(II)155) are novel metazoan homologues of yeast TAF(II)47 containing a histone fold and a PHD finger. *Mol. Cell. Biol.*, **21**, 5109–5121.
14. Selleck,W., Howley,R., Fang,Q., Podolny,V., Fried,M.G., Buratowski,S. and Tan,S. (2001) A histone fold TAF octamer within the yeast TFIID transcriptional coactivator. *Nature Struct. Biol.*, **8**, 695–700.
15. Slesarev,A.I., Belova,G.I., Kozyavkin,S.A. and Lake,J.A. (1998) Evidence for an early prokaryotic origin of histones H2A and H4 prior to the emergence of eukaryotes. *Nucleic Acids Res.*, **26**, 427–430.
16. Schaffer,A.A., Aravind,L., Madden,T.L., Shavirin,S., Spouge,J.L., Wolf,Y.I., Koonin,E.V. and Altschul,S.F. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.*, **29**, 2994–3005.
17. Sullivan,S.A., Aravind,L., Makalowska,I., Baxevanis,A.D. and Landsman,D. (2000) The Histone Database: a comprehensive WWW resource for histones and histone fold-containing proteins. *Nucleic Acids Res.*, **28**, 320–322.
18. Thompson,J.D., Gibson,T.J., Plewniak,F., Jeanmougin,F. and Higgins,D.G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.*, **25**, 4876–4882.