272–273 *Nucleic Acids Research, 1997, Vol. 25, No. 1*

Histone and histone fold sequences and structures: a database

Andreas D. Baxevanis and David Landsman[1,*]

Genome Technology Branch, National Center for Human Genome Research, National Institutes of Health, Bethesda, MD 20892, USA and [1]Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Received October 1, 1996; Accepted October 8, 1996

ABSTRACT

A database of aligned histone protein sequences has been constructed based on the results of homology searches of the major public sequence databases. In addition, sequences of proteins identified as containing the histone fold motif and structures of all known histone and histone fold proteins have been included in the current release. Database resources include information on conflicts between similar sequence entries in different source databases, multiple sequence alignments, and links to the Entrez integrated information retrieval system at the National Center for Biotechnology Information (NCBI). The database currently contains over 1000 protein sequences. All sequences and alignments in this database are available through the World Wide Web at: http://www.ncbi.nlm.nih.gov/Baxevani/HISTONES/ .

INTRODUCTION

The central role of the histone proteins in the compaction and organization of eukaryotic chromosomes has made these proteins the objects of biological studies for many years. The four core histones (H2A, H2B, H3 and H4) form an octameric assembly around which 146 base pairs of DNA wraps to form nucleosomal core particles (1), representing the first order of compaction of DNA in eukaryotic nuclei. The linker histones (H1 and H5) bind to the DNA found between these nucleosomal core particles (2), playing a role in the stabilization and formation of higher-order chromatin structure. The histone proteins exhibit a very high degree of conservation of their primary sequences. While the core histones are highly conserved across their entire sequence, the linker histones only exhibit such conservation in their central, globular domain.

Crystallographic studies focusing on the structure of the core histone octamer identified a common folding motif within each of the component protein chains (3). This histone fold motif, which is an extended helix–strand–helix motif, drives the head-to-tail association of individual core histone proteins into heterodimers, which in turn form the core histone octamer. Using a motif searching method (4), a group of non-histone proteins containing the histone fold motif were identified (5). Most of these proteins are involved in protein–protein and protein–DNA interactions, as are the core histones themselves. Structural studies on the $TAF_{II}42/TAF_{II}62$ complex from *Drosophila* (6) and HMfB from *Methanococcus fervidus* (7), proteins identifed as containing the histone fold in the aforementioned searches, confirmed that a histone-like substructure exists in these proteins, with the individual proteins folding into the canonical histone fold motif.

The Histone Sequence Database represents a collection of all of the histone and histone fold protein sequences and structures available as of October 1996. The inclusion of non-histone proteins containing the histone fold motif, as well as detailed information on conflicts between sequence entries in the major public databases, make this the most comprehensive collection and annotation of such sequences available to the scientific community. (For database statistics, see Table 1.)

Table 1. Histone sequence database statistics

	Total sequence set	Non-redundant sequence set	Structures
Histone H1	198	74	1
Histone H2A	223	78	0
Histone H2B	205	70	0
Histone H3	226	82	0
Histone H4	178	59	0
Histone H5	13	5	1
Total Histone entries	1042	367	2
Histone fold proteins	44		1

DATABASE CONTENT

The histone protein sequences were compiled by searching the non-redundant (nr) protein sequence database at NCBI. This database is a compilation of entries from SWISS-PROT (8), the Protein Identification Resource (PIR) (9), the Protein Data Bank (PDB) and CDS translations from GenBank (10). In each case, the sequences from chicken or human sources were used as the basis for comparison; in the case of H5, only the chicken sequence

*To whom correspondence should be addressed. Tel: +1 301 496 2477; Fax: +1 301 480 9241; Email: landsman@ncbi.nlm.nih.gov

was used. Manually added to the histone H1 sequence set was the sequence of Lpi17p. This sequence was discovered by TBLASTN searches (11) against the recently-completed yeast genome sequence (http://www.genome-www.stanford.edu/Saccharomyces) in an open reading frame on chromosome 16 (12). Subsequent homology model building on Lpi17p supports the role of this protein as a yeast H1 (13).

For each of the five histone classes, there are two protein sequence files. The first protein sequence file contains all of the sequences found for that histone type and is, therefore, redundant. The sequence data is presented in FASTA format, with the definition line containing the accession number, description and histone code for each individual entry, each separated by a vertical bar. The second protein sequence file contains only one entry for each unique sequence from a particular organism or variant thereof, making these files non-redundant. These sequences are also in FASTA format, with only a histone code appearing in the definition line. The histone codes can be used to cross-reference entries in the complete, redundant set of protein sequences.

In the course of the database searches, cases were noted where there were conflicts between the individual histone sequence entries for a given histone. In citing sequence conflicts, a majority-rule approach was used. In the cases where there was no clear majority among the sequences, the differences are noted with respect to the entry found in SWISS-PROT, where available. A pairwise sequence alignment, generated using CLUSTAL W (14), is presented along with the sequence conflict information in each case. Cases where the protein sequences are incorrectly identified are also noted.

Multiple sequence alignments for each histone protein can be downloaded as PostScript files. In each alignment, the major human histone sequence (chicken in the case of H5) is shown at the top of the alignment, and the region comprising the histone fold motif is boxed and represents the result of manual alignment. Regions outside the histone fold region were aligned using CLUSTAL W (14), and alignments were formatted using ALSCRIPT (15).

As mentioned above, in addition to the histone sequence files, the database contains sequences of non-histone proteins identified as containing the histone fold motif (5). Two histone fold files are available, one containing the complete sequence of each protein, the second containing solely the histone fold motif portion of each of the sequences. Both of these files are in FASTA format. Multiple sequence alignments of these histone fold motifs are available in PostScript format. Finally, information is provided for all histone and histone fold proteins for which three-dimensional coordinate data has been published and made freely available.

DATABASE AVAILABILITY

The Histone Sequence Database is available through the World Wide Web at:

http://www.ncbi.nlm.nih.gov/Baxevani/HISTONES/ .

In order to increase the utility of the database, hyperlinks have been integrated into all of the FASTA-formatted sequence files. Clicking on the accession number for a particular entry produces the NCBI Entrez document report for that entry, including relevant links to MEDLINE and the sequence and structure databases. Hyperlinks from the table of structures connect to the MMDB structure summary, which provides the capability of viewing the structure using a utility called Cn3D [‘see in 3D,’ (16)]. In this fashion, users can take advantage of the integrated nature of Entrez to gather large amounts of useful information on a particular sequence or set of sequences (17).

The database files can also be downloaded directly from the FTP site at NCBI (ncbi.nlm.nih.gov, directory /pub/baxevanis/histones). For each histone type, there are two FASTA protein files, corresponding to the complete or redundant sequence set (*.raw) and to the non-redundant sequence set (*.nr). The histone codes used in the definition lines of these entries are in a text file (codes.txt) in the same directory. Two FASTA-format files of the histone fold protein sequences are in this directory as well: hf_seqs.txt contains the complete sequence of each protein, while hf_motif.txt contains only the histone fold motif portion of the each sequence.

Studies utilizing the data within this database, obtained either through the World Wide Web site or the anonymous FTP site, should cite this paper as the primary reference.

REFERENCES

1 Kornberg,R. and Thomas,J.O. (1974) *Science*, **184**, 865–868.
2 Noll,M. and Kornberg,R.D. (1977) *J. Mol. Biol.*, **109**, 393–404.
3 Arents,G. and Moudrianakis,E.N. (1993) *Proc. Natl. Acad. Sci. USA*, **90**, 10489–10493.
4 Tatusov,R.L., Altschul,S.F. and Koonin,E.V. (1994) *Proc. Natl. Acad. Sci. USA*, **91**, 12091–12095.
5 Baxevanis,A.D., Arents,G., Moudrianakis,E.N. and Landsman,D. (1995) *Nucleic Acids Res.*, **23**, 2685–2691.
6 Xie,X., Kokubo,T., Cohen,S.L., Mirza,U.A., Hoffman,A., Chait,B.T., Roeder,R.G., Nakatani,Y. and Burley,S.K. (1996) *Nature*, **380**, 316–322.
7 Starich,M.R., Sandman,K., Reeve,J.N. and Summers,M.F. (1996) *J. Mol. Biol.*, **255**, 187–203.
8 Bairoch,A. and Apweiler,R. (1996) *Nucleic Acids Res.*, **24**, 221–222.
9 George,D.G., Barker,W.C., Mewes,H.W., Pfeiffer,F. and Tsugita,A. (1996) *Nucleic Acids Res.*, **24**, 17–20.
10 Benson,D.A., Boguski,M., Lipman,D.J. and Ostell,J. (1996) *Nucleic Acids Res.*, **24**, 1–7.
11 Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) *J. Mol. Biol.*, **215**, 403–410.
12 Landsman,D. (1996) *Trends Biochem. Sci.*, **21**, 287–288.
13 Baxevanis,A.D. and Landsman,D. (1997) manuscript in preparation.
14 Higgins,D.G., Bleasby,A.J. and Fuchs,R. (1992) *CABIOS*, **8**, 189–191.
15 Barton,G.J. (1993) *Protein Engng.*, **6**, 37–40.
16 Hogue,C.W.V., Ohkawa,H. and Bryant,S.H. (1996) *Trends Biochem. Sci.*, **21**, 226–229.
17 Baxevanis,A.D. and Landsman,D. (1995) *FASEB J.*, **9**, 994.