

Homology Model Building of Hho1p Supports its Role as a Yeast Histone H1 Protein

Andreas D. Baxevanis¹ and David Landsman^{2,*}

¹*Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA, e-mail: andy@nhgri.nih.gov*

²*Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA, e-mail: landsman@ncbi.nlm.nih.gov*

ABSTRACT: Biochemical studies to date have not been able to identify the linker histone H1 protein in the budding yeast *Saccharomyces cerevisiae*. Database homology searching against the complete yeast genome has identified a gene, HHO1, (or YPL127C, formerly LPI17) which encodes a protein that has two regions that show similarity to the pea histone H1 globular domain. To determine whether Hho1p can assume the shape of an H1 protein, homology model building experiments were performed using the structure of chicken histone H5 globular domain as the basis for comparison. A statistically significant match between each of the two globular domains of Hho1p and the chicken histone H5 structure was obtained, and probability values indicate that there is a less than 1 in 100 chance that such a match would be the result of a random event. These findings support the proposal that Hho1p acts as an "H1 dimer" and could be responsible for the decreased linker DNA length observed between nucleosomal core particles.

KEY WORDS: Histone H1; HHO1; homology model building; protein structure prediction.

INTRODUCTION

Due to their critical role in the compaction of DNA and the sensitive nature of their interactions with other chromosomal proteins, the histones are amongst the most highly conserved proteins in nature. These highly basic proteins fall into two broad classes: the *core histones* (H2A, H2B, H3, and H4), which associate into an octameric complex that is the basis for the formation of nucleosomal core particles [1]; and the *linker histones* (H1 and H5), which bind to the linker DNA between nucleosome cores [2] and play a role in stabilizing higher-order chromatin structure.

With the exception of dinoflagellates, histones have been isolated from the somatic cells of all eukaryotic organisms studied. However, certain organisms do not possess the entire complement of histone proteins. The foremost example of this is the yeast *Saccharomyces cerevisiae*. While some immunological evidence exists arguing for the presence of histone H1 in yeast [3], it has not yet been isolated directly (*cf.* [4] for a review). Since H1 is the least conserved of the histone proteins, it is conceivable that the protein has indeed been isolated but has not been recognized as such. A further

* corresponding author

complication is that the nucleosomal repeat length in yeast is substantially shorter than that found in other eukaryotes [5,6,7], again arguing for an H1 that may be dissimilar to its counterparts in other organisms.

The search to identify the H1 protein in yeast was substantially advanced with the completion of the sequence for the *Saccharomyces* genome (<http://genome-www.stanford.edu/Saccharomyces>). The availability of extensive sequence data such as this allows for the application of computational techniques in biological discovery. The most widely-used and conceptually easiest to understand of these techniques is database homology searching [8], where sequence similarity can be used to assign putative functions to newly-found, unknown genes. In the case of the yeast histones, the protein sequence of H1 from pea was used to search against the complete yeast genome using the TBLASTN algorithm [9], the translation being performed in all six reading frames [10]. This study revealed a putative gene on chromosome XVI encoding a protein named Lpi17p (this locus has also been named YPL127C); this protein was subsequently renamed Hho1p. Similar search results were obtained by querying the open reading frame sequences in the *Saccharomyces* Genome Database (SGD) using the pea and human H1 sequences as the basis for comparison, identifying a gene at the HHO1 locus [11]. Based on the physical characteristics of the yeast protein, it was proposed that Hho1p is essentially an H1 "dimer", having two functional globular domains connected by a lysine-rich segment, and that this extended H1 structure could explain the observed lack of linker DNA in yeast.

The similarities found between the sequences of the known H1 proteins and Hho1p argue that there should be significant similarities in their three-dimensional structures, strengthening the hypothesis that these proteins perform a similar function within the cell. Since the structure of chicken H5 has been determined at a resolution of 2.0 Å [12], homology model building, or *threading* [13], can be used to test whether the sequence of Hho1p can adopt the H1 structure. In the present study, this threading analysis indicates that the two globular domains of Hho1p shows a statistically significant match of its sequence to the structure of H1, supporting the previous proposal of its role as the yeast H1 protein. A threading-based "structural alignment" of Hho1p to H1 is provided, indicating both structured core and variable loop regions within the structure.

MATERIALS AND METHODS

Based on the results of database searches described elsewhere [10], a data set containing the sequences of chicken H5, human H1, pea H1, *Xenopus* H1, and *Saccharomyces* Hho1p was created from the appropriate SWISS-PROT and GenBank entries, accession numbers for which are given in Table 1. Since the Hho1p sequence is proposed to contain two H1-like sequences [10], each domain within Hho1p is treated as a separate sequence in the following analysis.

Threading experiments were performed by the method of Bryant and Lawrence [13], with detailed derivations and methodology provided therein. All 3,240,300 possible alignments of the H1 and Hho1p query sequences with the X-ray structure of chicken H5 (1HST [12]), which is the avian analog of histone H1, were examined as described in the text. Six core segments (CS) were defined based on the crystal structure of chicken H5: CS1 spanned from residues 30 to 39, CS2 from residues 46 to 47, CS3 from residues 48 to 57, CS4 from residues 65 to 79, CS5 from residues 82 to 87, and CS6 from residues 92 to 95, the numbering corresponding to that in Fig. 1. These core residues correspond to each of the secondary structural elements as defined in the PDB file (α 1- β 1- α 2- α 3- β 2- β 3-). Intervening loop length constraints were systematically varied to allow for flexibility within the structure, as indicated in Fig. 1.

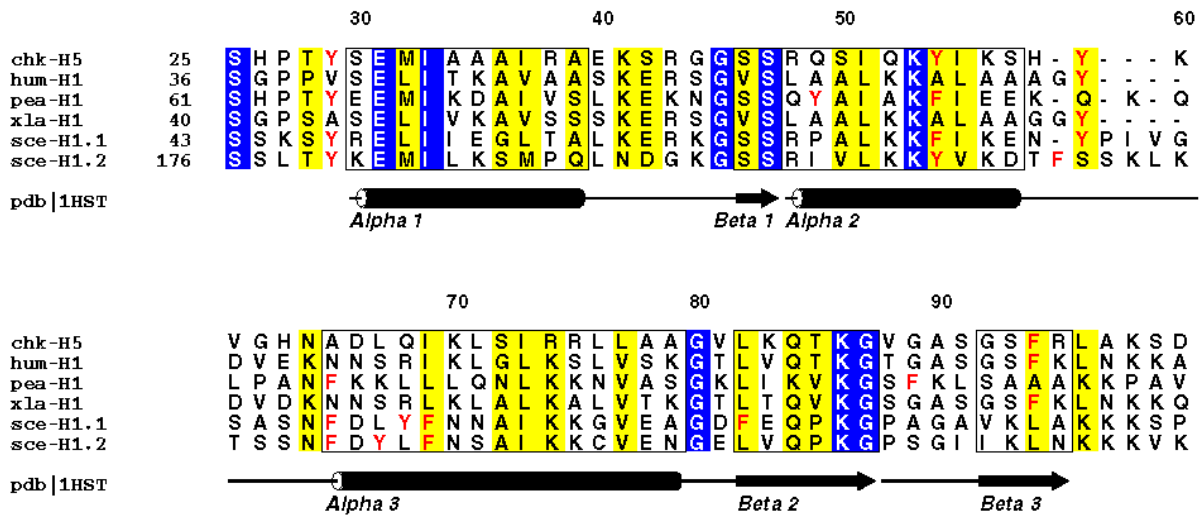


Fig. 1. Structural alignment of the globular domains of histone H1, H5, and yeast Hho1p. The abbreviation for each protein corresponds to those listed in Table 1. Decimals appended to the *Saccharomyces* sequences denote multiple H1 domains within the same protein, the numbers being assigned in the same order that these domains appear within the protein (N- to C-terminal). The numbering scheme at the top of the figure refers to the sequence of chicken H5 (the self-thread), while the numbers appearing to the left of each sequence identify the position of the first residue shown. Positions exhibiting absolute identity are shown in blue, while conserved positions are shown in yellow. Phenylalanine (F) and tyrosine (Y) residues are shown in red. The positions of the secondary structural elements, corresponding to the core elements in the threading experiments, are shown at the bottom of the figure. ALSCRIPT version 2.0 [21] was used to format the alignment.

For each possible alignment, individual pairwise residue interactions were determined based on chemical type and distance intervals with no use of arbitrary gap penalties [13]. Using these values, a conformational energy (ΔG_{RIM}), defined as the expected work for substitution of a specific sequence R for a random sequence with the same composition in the context of folding motif M , was then calculated for each alignment. Critical pairwise interactions are defined as those having a pairwise interaction energy < -1 kcal mol $^{-1}$. Z-scores (Z_{RIM}) and chance occurrence probabilities (E_{RIM}) were calculated to compare conformational energies for different alignments. Chance occurrence probabilities give the odds that a random sequence of the same length and amino acid composition would yield a threading energy as low as that for the query sequence R . A significant match of sequence to structure results when E_{RIM} for that thread is < 0.05 (5%). Calculation of energies and statistical values were performed using C and S-PLUS [14] subroutines. All energy scaffold figures were generated using the GRASP software package [15].

RESULTS AND DISCUSSION

Model structures were generated for each of the four H1 sequences and for the two globular regions within the yeast Hho1p protein by homology model building [13]. Each of the query sequences was threaded through the X-ray structure of chicken H5 at 2.0 Å resolution [12]. The three-dimensional structure itself was used to define core segments (regions of secondary structure) and intervening loop length limits. All possible placements of the core segments along the query sequence given the constraints of sequence length, core segment length, and loop length were considered. Conformational energies

(ΔG_{RIM}), defined as the expected work for substitution of a specific sequence R for a random sequence with the same composition in the context of folding motif M , were then calculated for each possible alignment. Threads with the most favorable conformational energies (*i.e.*, those with the lowest ΔG_{RIM}) were selected for further study.

The optimal threading energies for the H1 sequences are very close to one another, the extremes varying by only 9 kcal mol⁻¹ (Table 1). Given that the four H1 sequences are significantly different, with the fraction of residues being identical to that of chicken H5 being well below 50%, more variation would have been expected in the threading energies. The tight range of ΔG_{RIM} suggests that, despite a moderate level of conservation, these proteins do not exhibit a large net difference in their 3D structure. Quite interestingly, the two globular domains within Hho1p (sce-H1.1 and sce-H1.2) yielded more favorable ΔG_{RIM} values than the self-thread of the chicken H5 sequence through its own structure. Based on the chance occurrence probabilities, which correct for the effect of sequence length and composition, the threads for all of the H1 sequences and for the two Hho1p domains represent a statistically significant match of sequence to structure ($E_{\text{RIM}} < 0.05$). The probability values for sce-H1.1 and sce-H1.2 are 0.003 and 0.010, respectively, indicating that the odds are less than 1 in 100 that a random match of sequence to structure would yield a score this favorable.

Examination of the threading alignments indicates that there are no rigid sequence requirements for the formation of this structure, as there are no extensive regions of sequence identity across all of the sequences. Only eight positions out of the 70 within the threaded region show absolute identity, and only six of these eight are within regions of secondary structure (Fig. 1). 22 positions exhibit conservative substitutions as assessed by ALSCRIPT [16], most of which are in core regions. Taken together, conserved and identical positions comprise 38% of the threaded region.

In order to illustrate the network of pairwise interactions responsible for maintaining the structure of H1, a series of energy scaffolds were generated (Fig. 2). In the case of chicken H5, there are nine residues involved in critical pairwise interactions, defined as having an energy of < -1 kcal mol⁻¹ (Ser 50, Ile 51, Gln 52, Ile 55, Ile 69, Ile 73, Lys 83, Thr 85, and Phe 94). Interestingly, these residues are found in alpha-helices 2 and 3 and beta-sheets 2 and 3, all in the C-terminal region of the protein; no such strong interactions are seen in the N-terminal alpha-helix 1 and beta-sheet 2.

In the scaffolds for Hho1p, the N-terminal H1-like region also has nine residues involved in critical pairwise interactions (Leu 51, Lys 52, Ile 55, Phe 69, Ala 72, Ile 73, Gln 84, Lys 86, and Val 92). Ten such positions are seen in the C-terminal H1 like region of Hho1p (Val 50, Leu 51, Lys52, Val 55, Phe 69, Ala 72, Ile 73, Val 83, Pro 85, and Leu 94). Examining these critical positions as a group and correlating them back to the structural alignment (Fig. 1), the following pattern emerges for the five common positions identified as critical in all of the scaffolds: First, a neutral hydrophobic residue must be present at positions 51, 55, 69, and 73. Second, position 52 is interesting in that, while identified as critical across all scaffolds, no absolute assignment of residue type or size can be made, although the residue at position 52 is always seen to interact with a neutral hydrophobic (Phe or Ile) at position 69.

Both of these Hho1p sequences share one strong unfavorable interaction, between the arginine at position 48 and the asparagine at position 70 (+1.3 kcal mol⁻¹ in sce H1.1, +1.6 kcal mol⁻¹ in sce-H1.2). A second unfavorable interaction is also seen in sce-H1.2 between Arg 48 and Ile 92 (+1.8 kcal mol⁻¹). Based on the scaffolds and the alignment, it appears that the residue at position 48 is not responsible in itself for these unfavorable interactions, since the sequence for chicken H5 itself has an arginine at position 48. The substitution of a charged residue (Lys) by a neutral polar (Asn) at position 70 and the substitution of a neutral polar (Gly) by a neutral hydrophobic (Ile) at position 92 appear responsible for the large positive ΔG values. Despite this, these unfavorable interactions were not sufficient to disrupt the overall fit of the sequence to the structure of H5, the positive interactions being much more predominant and being distributed throughout the four C-terminal core regions. All the other sequences threaded here

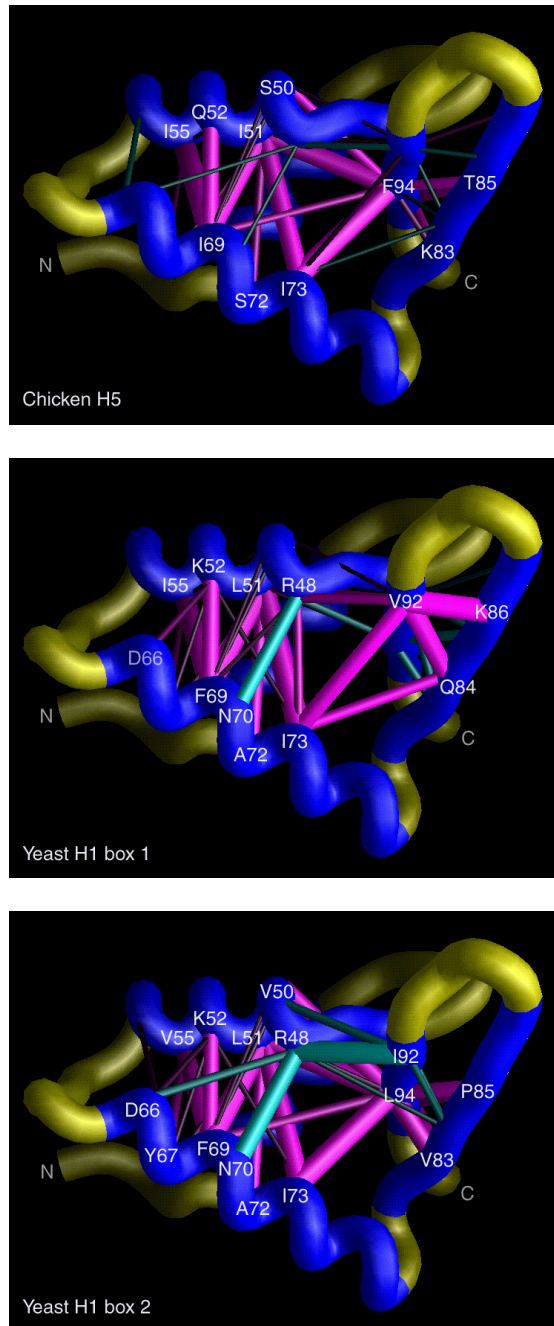


Fig. 2. Energy scaffolds for query sequences through the structure of chicken H5. The alpha-carbon backbone of the protein is depicted as a curving "worm". Within the backbone, core segments of the H5 structure are shown in blue, while the intervening loop regions are shown in yellow. Pairwise residue interaction energies between core residues are indicated by the thickness and coloring of the rods connecting alpha carbon positions on the protein backbone. Thick, magenta-colored cylinders indicate the most favorable interactions; thick, cyan-colored cylinders indicate the least favorable interactions. Intermediate colors and cylinder thicknesses represent interactions falling between these extremes. Residue numbering corresponds to the numbering in the multiple sequence alignment. Scaffolds were generated using the graphics program GRASP [15]. *Top*, chicken H5; *middle*, yeast H1 box 1; *bottom*, yeast H1 box 2.

contain either an Arg, Leu, or Gln at position 48 whereas at position 70 it is either a Lys, Leu, or Asn. However, none of these elicited an equally unfavorable interaction.

Based on these results and those previously published based on sequence analysis [10], it appears that the HHO1 gene product has the potential to carry out the functional role of histone H1 in yeast. The overall structure of the yeast H1 is consistent with the canonical structure of other H1s, where a central, globular domain is flanked by an N-terminal lysine-rich peptide and a C-terminal lysine, alanine, and proline-rich tail [17, 18]. Where yeast is dissimilar from other H1 proteins previously characterized is in that there are *two* globular domains connected by a lysine-rich segment. This makes the yeast H1 protein equivalent to an "H1 dimer", with one yeast H1 molecule possibly functioning the same as two H1 molecules from other organisms.

While histone H1 has been putatively identified in other Ascomycetes, namely *Neurospora* [19] and *Aspergillus* [20], assignments in these studies were made on the basis of relative mobility on polyacrylamide gels and by overall amino acid composition. To date, no primary sequence information is available in the public databases on H1 from either of these organisms. It is interesting to speculate whether the "dimeric H1" is a phenomenon seen solely in yeast or whether other fungi also possess such an H1 protein.

Table 1
Statistics for optimal threads of sequences through the structure of the globular domain of chicken histone H5

Accession	Sequence	Organism	ΔG (kcal/mol)	Z-score	E_{RIM}	Log odds	Fraction Identical
P02259	chk-H5	<i>Gallus gallus</i>	-26.39	2.44	0.007	0.00	1.00
P16403	hum-H1	<i>Homo sapiens</i>	-17.83	1.94	0.026	-1.27	0.42
P08283	pea-H1	<i>Pisum sativum</i>	-28.97	2.54	0.005	0.29	0.27
P06893	xla-H1	<i>Xenopus laevis</i>	-21.55	1.93	0.027	-1.30	0.37
P53551	sce-H1.1	<i>Saccharomyces cerevisiae</i>	-26.89	2.66	0.003	0.62	0.33
P53551	sce-H1.2	<i>Saccharomyces cerevisiae</i>	-28.88	2.30	0.010	-0.38	0.28

Sums of contact potentials are expressed as a conformational energy ΔG , the energy associated with non-local, non-bonded interactions. Z- scores, in standard deviation units, represent the variance from the mean of ΔG . The chance of occurrence probability E represents the probability of observing a maximum Z-score by chance within a certain number of alignments. Log odds presents the ratio of the chance occurrence probability of the native sequence to that of the non-native query sequence, that is, $E_{chk-H5} / E_{non-native}$. Fraction identical is with respect to chicken H5 within the threaded region only.

Acknowledgements

We would like to thank Joel Barnabas for his assistance in preparing the HTML version of this manuscript.

References

- [1] Kornberg, R. and Thomas, J.O. *Science* **184**, (1974) 865–868.
- [2] Noll, M. and Kornberg, R.D. *J. Mol. Biol.* **109**, (1977) 393–404.
- [3] Srebrena, L., Zlatanova, J., Miloshev, G. and Tsanev, R. *Eur. J. Biochem.* **165**, (1987) 449–454.
- [4] van Holde, K.E. *Chromatin*, Springer-Verlag, (1989) New York.
- [5] Lohr, D., Kovacic, R.T. and van Holde, K.E. *Biochemistry* **16**, (1977) 463–471.

- [6] Lohr, D., Corden, J., Tatchell, K., Kovacic, R.T. and van Holde, K.E. *Proc. Natl. Acad. Sci. USA* **74**, (1977) 79–83.
- [7] Hörz, W. and Zachau, H.G. *J. Mol. Biol.* **144**, (1980) 305–327.
- [8] Altschul, S.F., Boguski, M.S., Gish, W. and Wootton, J.C. *Nat. Genet.* **6**, (1994) 119–129.
- [9] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. *J. Mol. Biol.* **215**, (1990) 403–410.
- [10] Landsman, D. *Trends Biochem. Sci.* **21**, (1996) 287–288.
- [11] Ushinsky, S., Bussey, H., Ahmed, A., Wang, Y., Friesen, J., Williams, B. and Storms, R. *Yeast* **13**, (1996) 151–161.
- [12] Ramakrishnan, V., Finch, J.T., Graziano, V., Lee, P.L. and Sweet, R.M. *Nature* **362**, (1993) 219–223.
- [13] Bryant, S.H. and Lawrence, C.E. *Proteins* **16**, (1993) 92–112.
- [14] Becker, R.A., Chambers, J.M. and Wilks, A.R. *The New S Language - A Programming Environment for Data Analysis and Graphics*, (1988) Wadsworth, Pacific Grove, CA.
- [15] Nicholls, A., Sharp, K.A. and Honig, B. *Proteins* **11**, (1991) 281–296.
- [16] Livingstone, C.D. and Barton, G.J. *Comput. Appl. Biosci.* **9**, (1993) 745–756.
- [17] Aviles, F.J., Chapman, G.E., Kneale, G.C., Crane-Robinson, C. and Bradbury, E.M. *Eur. J. Biochem.* **88**, (1978) 363–371.
- [18] Chapman, G.E., Hartman, P.G., Cary, P.D., Bradbury, E.M. and Lee, D.R. *Eur. J. Biochem.* **86**, (1978) 35–44.
- [19] Goff, C.G. *J. Biol. Chem.* **251**, (1976) 4131–4138.
- [20] Felden, R.A., Sanders, M.M. and Morris, N.R. *J. Cell Biol.* **68**, (1976) 430–439.
- [21] Barton, G.J. *Protein Eng.* **6**, (1993) 37–40.