Characterization of sequence variability in nucleosome core histone folds

ARTICLE *in* PROTEINS STRUCTURE FUNCTION AND BIOINFORMATICS · AUGUST 2003

Impact Factor: 2.63 · DOI: 10.1002/prot.10441 · Source: PubMed

2 AUTHORS:

Steven A Sullivan
New York University
37 PUBLICATIONS **4,464** CITATIONS
SEE PROFILE

David Landsman
National Institutes of Health
129 PUBLICATIONS **9,307** CITATIONS
SEE PROFILE

PROTEINS: Structure, Function, and Genetics 52:454–465 (2003)

Characterization of Sequence Variability in Nucleosome Core Histone Folds

Steven A. Sullivan and David Landsman*
Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894

ABSTRACT **The three-helix, ~65-residue histone fold domain is the most structurally conserved part of the core histones H2A, H2B, H3, and H4. However, it evinces a notable degree of sequence variation within and between histone classes. We used two approaches to characterize sequence variation in these histone folds, toward elucidating their structure/function relationships and evolution. On the one hand we asked how much of the sequence variation seen in structure-based alignments of the folds maintains physicochemical properties at a position, and on the other, whether conservation correlates to structural importance, as measured by the number of residue-to-residue contacts a position makes. Strong physicochemical conservation or correlation of conservation to contacts would support the idea that functional constraints, rather than genetic drift, determines the observed range of variants at a given position. We used an 11-state table of physicochemical properties to classify each position in the core histone fold (CHF) alignments, and a public website (http://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/valdar/scorecons_server.pl) to score conservation. We found that, depending on histone class, from 38 to 77% of CHF positions are maximally conserved physicochemically, and that for H2B, H3, and H4 the degree to which a position is conserved correlates positively to the number of contacts made by the residue at that position in the crystal structure of the nucleosome core particle. We also examined the correlation between conservation and the type of contact (e.g., inter- or intrachain, histone-histone, or histone-DNA, etc.). For H2B, H3, and H4 we found a positive correlation between conservation and number of interchain protein contacts. No such correlation or statistical significance was found for DNA or intrachain contacts. This suggests that variations in the CHF sequences could be functionally constrained by requirements to make sufficient interchain histone contacts. We also suggest that inventory of histone residue variants can augment functional studies of histones. An example is presented for histone H3. Proteins 2003;52:454–465.** © 2003 Wiley-Liss, Inc.

Key words: sequence conservation

INTRODUCTION

The eukaryotic nucleosome is a highly conserved structure representing an ancient solution to the problem of DNA compaction within the cell. Its main protein component is the histone octamer, comprising two copies of each of the four basic core histones: H2A, H2B, H3 and H4.[1] The fundamental structural units of the octamer are heterodimers of H3|H4 and H2A|H2B, which are stabilized by the interaction of a ~65 amino acid motif called the histone fold. The three connected helices of the paired histone folds of the dimer interdigitate in a head-to-tail "handshake"-like fashion,[2] forming a compact structure. The association of these dimers to form first a tetramer and then an octamer creates a globular ramp around which DNA is supercoiled, with relatively unstructured N- or C-terminal regions of each histone extending from the core. Binding of the octamer to DNA is effected mainly through noncovalent bonding of DNA backbone phosphate groups to the positive ends of CHF helical dipoles and to polar CHF moieties such as main-chain amide nitrogens, basic side-chain groups, and hydroxyl side-chain groups.[3]

Histones are among the most highly conserved proteins known. The histone fold, in turn, is the most conserved region of each histone, with 100% sequence identity observed between plant and human sequences in some cases.[4] Histone fold sequences similar to eukaryotic core histones are also present in euryarchaeal DNA-binding proteins.[5] Such a degree of conservation across vast evolutionary distances indicates that sequence, structure, and function are highly interdependent in these proteins. Yet, sequence variation is still observed in histones.[6–8] The most variable regions are relatively unstructured and lie outside the histone folds.[3] They contain sites of post-transcriptional modification important for regulating higher-order DNA packing and, by extension, transcriptional accessibility (reviewed in Jenuwein et al.[9]). As such, these regions have been the focus of most studies of histone sequence variability. We have addressed this imbalance by characterizing core histone fold (CHF) sequence variation. The deposition of numerous homologous histone sequences in GenBank and other public databases[4] makes a compre-

*Correspondence to: David Landsman, NCBI/NIH, Building 38A, Room 6N601, Bethesda, MD 20894. E-mail:landsman@ncbi.nlm.nih.gov

Received 22 December 2002; Accepted 3 February 2003

hensive survey of histone variation possible. Moreover, now that the structure of the nucleosome has been solved at high resolution,[3,10,11] one can begin to analyze the positional variation in light of molecular contact information. Using these two resources, we have quantified histone variation both categorically and using empirical measures (i.e., profile-derived substitution matrices) and analyzed relationships between conservation and frequency of contacts as an indicator of functional importance.

METHODS

Sequence Sets

Core histone protein sequences were compiled from the online Histone Database (HDB) http://genome.nhgri.nih.gov/histones/).[4] The sequence datasets were updated to remove sequences that have been withdrawn or deleted from GenBank and to add sequences new to GenBank since the December 2001 update of the HDB. The dataset for each histone class, consisting of sequences in FASTA format, was made nonredundant using the "fauniq" command of the SEALS software package.[12] Sequence fragments that were subsequences of longer sequences were identified by all-against-all pairwise comparison and assigned to the appropriate cluster in the nonredundant dataset. Each dataset was aligned by CLUSTAL X[13] guided by a structural mask derived from the histone fold helical boundaries in the PDB record (PDB ID 1aoi) for the crystal structure of the *Xenopus* nucleosome at 2.8 Å resolution.[3] The aligned histone fold domain was extracted from the full-length alignment and adjusted by hand if necessary to transform multiple short gaps in the interhelical loops into one or two long gaps. A few sequences had gaps or deletions >1 residue in length in the center of helical regions, or had ambiguous residues, or had frameshifts in the histone fold, or were known to be products of pseudogenes; these were excluded from the datasets. For the four datasets, the numbers of redundant (r), CHF-containing sequences retrieved from the HDB relative to the numbers of nonredundant (nr) sequences from which CHFs were derived, were as follows: H2A: 467 r, 163 nr; H2B: 456 r, 170 nr; H3: 492 r, 113 nr; and H4: 322 r, 59 nr. The redundant and nonredundant datasets for each histone class, with sequences and subsequences clustered by identity, are available from http://www.ncbi.nlm.nih.gov/CBBresearch/Landsman/HistoneVariants.

Taxonomic Analysis

A taxonomic breakdown of the sequence sets was generated using the "tax_break" command of SEALS. Information about natural biodiversity, used for comparison to dataset biodiversity, was gleaned from the following online sources and references contained therein and from Hawksworth and Kalin-Arroyo[14]: the Phylogenomics site maintained by the BaNG group at the University of Edinburgh (http://nema.cap.ed.ac.uk/phylogenomics/seqspec.html), the Web Lift to Taxa site maintained by the Museum of Paleontology at UC Berkeley (http://www.ucmp.berkeley.edu/help/taxaform.html), the Wildlife Research Institute site (http://www.wri.org/wri/biodiv/f01-key.html), and the Tree of Life site edited by David Maddison (http://tolweb.org/tree/phylogeny.html).

Sequence Conservation Analysis

Sequence identity

Pairwise sequence comparisons of multiple-aligned sequences were performed on the Protein Sequence Analysis (PSAweb) server (http://www.imtech.res.in/raghava/psa/)[15] using an identity matrix to generate statistics for overall identity between CHF sequences.

Physicochemical conservation

An 11-state physicochemical property table (similar to that used by Taylor[16] and Livingston and Barton[17] was used to derive a physicochemical consensus for each column in an alignment. The 11 properties and their associated residues were as follows: negative, −, DE; Ser/Thr, *, ST; aliphatic, l, ILV; positive, +, HKR; tiny, t, AGS; aromatic, a, FHWY; charged, c, DEHKR; small, s, ACDGNPSTV; polar, p, CDEHKNQRST; big, b, EFHIKLMQRWY; and hydrophobic, h, ACFGHILMTVWY. These constitute the default classifications used in the CHROMA alignment coloring tool,[18] which we also used to generate a consensus property for each aligned column.

From the chart above it is apparent that residues can belong to multiple physicochemical classes. The maximum conservation threshold (MCT) is our term for the highest percentage of residues in a column that can be assigned to the same physicochemical class; e.g., a column containing only aliphatic and aromatic residues would have an MCT100 (because they can all be classed as hydrophobic), whereas one consisting of 80% negative and 20% tiny residues would have an MCT80, because these two classes do not overlap.

Solvent accessibility

The solvent accessibilities of protein residues in the PDB structure file 1aoi[3] were derived from a DSSP database server http://cubic.bioc.columbia.edu/services/DSSPcont/.[19] The relative solvent accessibility (RSA) = observed solvent accessibility (Å^2)/maximal accessibility (per Ref. 20). Buried residues were defined as having <9% RSA. Interfacial residues were defined as those whose RSA was <9% in the octamer (sans DNA) but ≥9% in the monomer.

Positional conservation scores

Conservation scores were assigned to each column in an alignment by the SCORECONS server (http://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/valdar/scorecons_server.pl). Four parameters of this server were set as follows:

1. Scoring method: valdar01,[21,22] which is a sum-of-pairs, substitution matrix-based, sequence-weighted method;
2. Substitution matrix: Pairwise Exchange Table (PET91)[23] modified so that values on the diagonal (e.g., A vs. A) are constant;

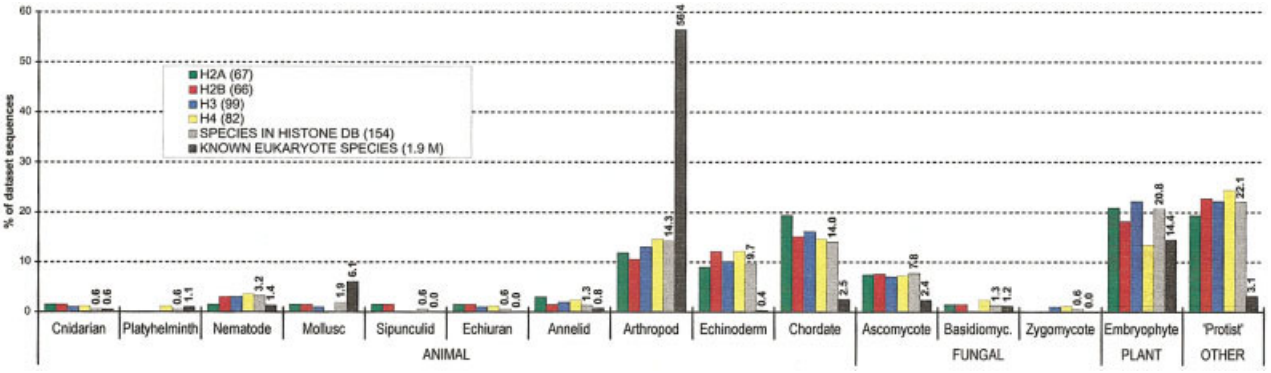

Fig. 1. Biodiversity of histone sequences in the datasets used in this study. Numbers in parentheses in the color code are the numbers of different species in each histone dataset. Species in histone db = tally of unique species in all four histone datasets. Known eukaryote species = approximate total number of known eukaryote species (see Methods for sources). For known histones and known eukaryote species data, the percentages represented by each phylum are noted explicitly above the bars (0.0 = <0.1%).

3. Matrix transformation to convert substitution matrix scores into a usable range: karlinlike[24];
4. Gapphilia: O (substitution of gaps for amino acids is strongly penalized).

Residue Contact Inventory

For each CHF side chain in the PDB nucleosome structure file 1aoi[3] we determined its connectivity, defined as the number of side chains and main chains it contacts, based on 2-D contact maps generated by WPBD.[25] A contact was defined as having occurred if at least one CHF side-chain atom (α-carbon inclusive) was within 4.5 Å of a side-chain or main-chain atom of another residue. The contacts could be made within the same histone molecule (intrachain), or with another histone or DNA (interchain). We observed that a CHF residue at sequence position n almost always makes contact with all residues in the range $n \pm 4$, as might be expected in a mostly α-helical peptide. Therefore, contacts with residues within $n \pm 4$ were excluded from the counts for position n.

Statistics

Nonparametric Spearman Rank Order correlation analysis and Mann-Whitney Rank Sum analysis of variance were performed using SigmaStat 2.03 (SPSS Inc.). Statistical significance was assumed when $P < 0.05$.

RESULTS AND DISCUSSION

Taxonomic Coverage of Dataset

Inferences about the natural sequence variation of a protein that are gleaned from a sample of all living taxa are more likely to be valid if the sample is taxonomically diverse. Our dataset of CHF sequences represents 154 species, from all four top-level eukaryote divisions (animal, fungal, plant, and unicellular organisms formerly grouped as "protist"). One hundred twenty species derive from 14 multicellular phyla, and 34 from unicellular taxa. Most are animal species (47%), followed by plant (22%), protist (21%), and fungal (10%)—a ratio of approximately 5:2:2:1. By contrast, the ~1.9 million identified living eukaryote species (see Methods) are distributed by kingdom at a ratio of 18:4:3:1. Thus, although the taxonomic coverage of our dataset is broad, and the rank order of percentages by kingdom in our dataset is natural (i.e., animal > plant or protist > fungal), the species percentages themselves are skewed in comparison to those found so far in nature. Animal species are underrepresented among histone classes on average by a factor of 1.4, whereas plant species are overrepresented by the same amount, and fungal and protist species are overrepresented by a factor of ~2. When the taxonomic distribution of histone sequences is broken down by phylum, reasons for this becomes more apparent (Fig. 1). Experimental bias toward "model" organisms has left many animal phyla underrepresented in the sequence databases (though the 10 of 35 animal phyla represented in the histone dataset is relatively large compared with other protein family datasets; S.A.S., personal observation). Among represented animal phyla, arthropod and mollusc sequences are highly underrepresented, whereas an experimental bias toward chordates, echinoderms, and nematodes is apparent. Fungal overrepresentation in the histone dataset is due mainly to the disproportionate number of ascomycete (e.g., yeast) sequences. Sequencing of genes from disease-causing organisms accounts for the protistan overrepresentation.

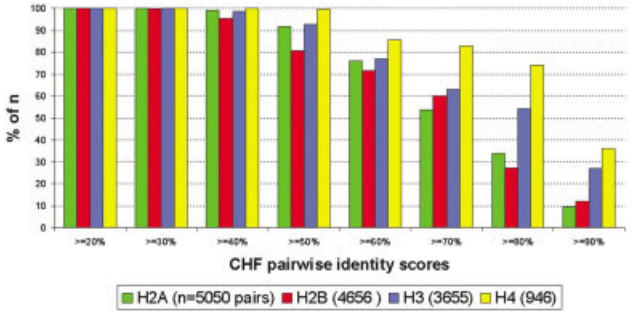

Fig. 2. Pairwise identity scores of CHFs. For the four histone fold multiple sequence alignments, the graph shows the percent of n pairwise sequence comparisons that produced percent identity scores greater than or equal to the bin values indicated.

TABLE I. Frequency of Strong Conservation in Core Histone and Myoglobin Folds[†]

					No. of positions at MCT (% length)			
Alignment	No. seq	Length	MCTmin	MCTavg	MCT100	MCT99	MCT98	MCT98–100
H2A	163	62	86	94	28 (45)	6 (10)	14 (23)	48 (78)
H2B	170	65	72	98	40 (62)	7 (11)	7 (11)	54 (84)
H3	113	68	78	97	26 (38)	11 (16)	7 (10)	44 (65)
H4	59	63	83	99	48 (76)	0 (0)	11 (17)	59 (94)
Myoglobin	35	113	57	86	23 (20)	0 (0)	0 (0)	23 (20)

[†]An 11-state physicochemical classification of amino acids was used (see Methods). The maximum conservation threshold (*MCT*) for a protein alignment column is the highest percentage of amino acid variants in the column that can be assigned to the same physicochemical class, e.g., the maximum value, MCT100, means that 100% of a sequence position's variants belong to the same class. MCTmin and MCTavg are the minimum and average MCT values observed across each alignment. For MCT98, MCT99, and MCT100 positions, the frequency of such positions, and the percent of the alignment length they comprise, are shown. The number of sequences in each alignment is also shown.

There are more species represented by the H3 and H4 sets (99 and 82, respectively) than the H2A and H2B sets (67 and 66), probably reflecting wider use of H3 and H4 in phylogenetic studies. Kingdom percentages among the histone classes are fairly consistent: 47–51% animal, 16–22% plant, 16–22% protist, and 8–11% fungal. We believe these data, taken together, indicate that our histone sequence datasets are sufficiently biodiverse for purposes of drawing general conclusions about histone sequence variation.

Sequence Conservation

As a first approach to quantifying CHF sequence variation we determined the percent identities of all pairwise comparisons of CHF sequences within each class (n = 5050, 4656, 3655, and 946 pairs for H2A, H2B, H3, and H4). The results confirm the extraordinary conservation characteristic of histone sequences and the histone fold domain in particular (Fig. 2). In each class more than half of the sequence pairs were at least 70% identical. The percentage of pairs that were >80% identical was notably higher for H4 and H3 than for H2A and H2B (73 and 54% of pairs versus 33 and 28%, respectively). Although derived from sets of longer, nonredundant histone sequences (see Methods), the sets of CHF subsequences used in this study contained redundancies. If these are removed, the histone class with the largest number of unique CHF domains is H2A (98), followed by H3 (91), H2B (58), and H4 (44) (data not shown, see http://www.ncbi.nlm.nih.gov/CBBresearch/Landsman/HistoneVariants).

Physicochemical Constraints on Histone Fold Sequence Variation

A priori physicochemical grouping of amino acids, i.e., grouping of amino acids by physicochemical criteria without reference to their frequency of co-occurrence in columns in sequence alignments, has been used by others to analyze variation along aligned sequence positions (e.g., Refs. 16, 17, 28, 29, and 30 and references therein).

We applied this approach to the histone fold alignments to assess how much histone sequence variation consists of physicochemically "acceptable" substitutions, rather than substitutions that might require significant functional change. Using an 11-state table of physicochemical properties (see Methods), for each column of an alignment we determined the maximum conservation threshold (MCT), that is, the highest percentage of residue variants for that position that can be assigned to the same physicochemical state. We found that MCT100 positions, in which all variants can be assigned to the same physicochemical state, tend to constitute a large proportion of CHFs (Tables I–V)—as much as three-quarters of positions in the case of H4). Moreover, if MCT99 and MCT98 positions are also considered, the variation in 65–94% of positions along CHFs is accounted for (Table I). By comparison, only 24% of sites across an alignment of myoglobin folds from fungal, plant, and metazoan globins[31] have MCT values ≥98 (Table I). The minimum MCT value observed among CHF positions (MCT72) is notably higher than the minimum observed in the myglobin fold (MCT57), as are the average MCT values for each of the four CHF sets versus the myoglobin fold (Table I). Such differences probably are not due to alignment of highly divergent globins, because differences between MCT statistics were at least as great when vertebrate CHF sequences were compared with vertebrate myoglobin sequences (data not shown). These data suggest that physicochemical constraint played a particularly strong role in delimiting sequence variation in CHFs as compared with other proteins.

Are any physicochemical groups predominant among MCT100 positions? When divided into 11 states, no physicochemical trend predominates. However, when the 11 states are reduced to 4 (big, small, polar, and hydrophobic), hydrophobic positions comprise the largest percentage (38–42%) of MCT100 positions in 3 of 4 histone classes (Table VI). Only in the H4 histone fold does the polar class predominate over the hydrophobic class of residues among MCT100 positions. Predominance of hydrophobic character in highly conserved positions is common among globular proteins, where close packing and exclusion of water from the interior maintain structural stability.[28] Interestingly, although histones are considered "basic," being among the most positively charged proteins known,[32] and generally a third or more of all arginine and lysine residues reside the CHFs, only H4 and H2B evince any "absolute" requirement for basic CHF residues, in 1 positively charged MCT100 position in H2B and 10 such positions in H4 (Tables II–IV). Indeed, the latter account

TABLE II. Core Histone Fold Sequence Variation and Associated Data—Histone H2A (Chains C, G)

												Residue contacts[g]									
													Inter							Total	
Pos[a]	Class[b]	MCT[c]	Access[d]	Score[e]	Variants[f]	Variants frequencies						Intra	Dimer	C|E	C|G	C|H	Prot	DNA	All	Prot	All
27	l	100	h	0.990	V I	162	1					6	0	0	0	0	0	0	0	6	6
28	t	100	e	0.943	G S A	159	3	1				1	0	0	0	0	0	2	2	1	3
					- A	162	1														
29	P	100	e	0.986	R Q	162	1					0	3	0	0	0	3	3	6	3	6
30	h	100	h	0.716	V I M L T	75	67	12	8	1		4	1	0	0	0	1	0	1	5	5
31	h	93	b	0.454	M G A L M S E K R T	89	25	20	12	6	5 3 1 1 1	4	0	0	0	0	0	0	0	4	4
32	p	96	e	0.812	R S G Y K	150	7	3	2	1		0	1	0	0	0	1	2	3	1	3
33	b	98	e	0.504	L Y F H M Q S I K	67	39	36	6	5	4 3 2 1	1	3	0	0	0	3	0	3	4	4
34	b	100	b (F)	0.886	L I M	148	12	3				5	2	0	0	0	2	0	2	7	7
35	b	100	e	0.793	R K Y	83	79	1				1	0	0	0	0	0	1	1	1	2
36	p	92	e	0.627	K R A N T S E Q G B	114	12	11	7	5	4 3 3 2 2	0	0	0	0	0	0	0	0	0	0
37	3	90	e	0.809	G R A B K T	145	14	1	1	1	1	0	0	0	0	1	1	0	1	1	1
38	p	95	e	0.466	N R H K T Q V A I S	70	25	17	17	15	11 5 1 1 1	0	0	0	4	1	5	0	5	5	5
39	h	87	e	0.644	Y P T F S O R	128	12	9	6	6	1 1	2	6	0	0	0	6	0	6	8	8
40	s	88	e	0.646	A S K G C N	115	21	18	6	2	1	0	3	0	1	0	4	0	4	4	4
41	b	90	e	0.452	E Q Y A H N K S A D I L P T	64	33	18	12	10	9 8 3 1 1 1 1 1	0	1	0	1	0	2	0	2	2	2
		0			- G M N Q S T	146	12	1	1	1	1 1										
42	R	98	e	0.960	R A P S	160	1	1	1			0	2	0	0	2	4	2	6	4	6
43	h	100	e	0.765	V I L Y A	110	46	4	2	1		3	1	0	0	0	1	1	2	4	5
44	t	98	e	0.891	G S R	152	9	2				1	1	0	0	0	1	1	2	2	3
45	s	98	e	0.648	A V S T G K P R	108	19	18	9	6	1 1 1	0	1	0	0	0	1	1	2	1	2
46	s	98	e	0.689	G T S K A D	133	17	9	2	1	1	0	1	0	0	0	1	0	1	1	1
47	s	100	b (F)	0.850	A G S D T	150	9	2	1	1		1	4	0	0	0	4	0	4	5	5
48	s	98	b	0.726	P A Q S V	132	26	2	2	1		5	0	0	0	0	0	0	0	5	5
49	h	100	b (F)	0.977	V I A L	159	2	1	1			2	3	0	0	0	3	0	3	5	5
50	a	100	b (F)	0.992	Y F	162	1					0	9	0	0	0	9	0	9	9	9
51	h	95	b (F)	0.675	L M S T A	112	37	7	5	2		2	3	0	0	0	3	0	3	5	5
52	s	99	b	0.905	A T S E	151	7	4	1			4	0	0	0	0	0	0	0	4	4
53	t	100	b (F)	0.965	A S	160	3					1	3	0	0	0	3	0	3	4	4
54	h	100	b (F)	0.893	V I T	147	15	1				0	4	0	0	0	4	0	4	4	4
55	h	100	b (F)	0.858	L M I T F V	140	12	7	2	1	1	2	2	0	0	0	2	0	2	4	4
56	E	100	e	1.000	E	163						3	1	0	0	0	1	0	1	4	4
57	Y	98	e	0.949	Y N	160	3					0	4	0	0	0	4	0	4	4	4
58	1	100	b (F)	0.987	L V	162	1					1	4	0	0	0	4	0	4	5	5
59	s	100	b (F)	0.687	T A C S V G	85	68	3	3	3	1	1	4	0	0	0	4	0	4	5	5
60	s	100	e	0.964	A T G	160	2	1				0	2	0	0	0	2	0	2	2	2
61	c	100	e	0.974	E K	161	2					0	1	0	0	0	1	0	1	1	1
62	l	100	b (F)	0.729	I V L	80	72	11				3	2	0	0	0	2	0	2	5	5
63	b	99	b (F)	0.958	L F M P	160	1	1	1			1	5	0	0	0	5	0	5	6	6
64	p	100	e	0.961	A D Q R	160	1	1	1			0	2	0	0	0	2	0	2	2	2
65	b	100	e	0.945	L M E	159	3	1				3	0	0	0	0	0	0	0	3	3
66	t	100	b	0.875	A S	153	10					4	0	0	0	0	0	0	0	4	4
67	s	98	b (F)	0.762	G V A R C E	146	9	4	2	1	1	1	1	0	0	0	1	0	1	2	2
68	p	98	e	0.853	N K I D	151	9	2	1			0	1	0	0	0	1	0	1	1	1
69	s	98	b	0.898	A T V E P R -	155	2	2	1	1	1 1	2	0	0	0	0	0	0	0	2	2
70	s	99	b	0.806	A S T C H	143	15	3	1	1		3	0	0	0	0	0	0	0	3	3
71	p	96	e	0.694	R K S A E N P Q	122	27	5	4	2	1 1 1	0	0	0	0	0	0	0	0	0	0
72	p	100	e	0.829	D Q E K N	149	9	3	1	1		0	0	0	0	0	0	0	0	0	0
73	p	89	e	0.615	N L S Q F H K T -	132	15	6	5	1	1 1 1 1 1	2	0	0	0	0	0	0	0	2	2
74	K	93	e	0.826	K A G S	152	5	5	1			0	0	0	0	0	0	1	1	0	1
75	p	88	e	0.718	K V E N P R S -	141	16	1	1	1	1 1 1	2	0	0	0	0	0	2	2	2	4
					- K E Q R	152	8	1	1	1											
					- P R	154	6	3													
76	p	86	e	0.461	T K S N G A R C I L Q	81	22	16	14	12	6 4 3 2 2 1	0	2	0	0	0	2	2	4	2	4
77	R	98	e	0.942	R V N	160	2	1				0	3	0	0	0	3	3	6	3	6
78	l	100	b (F)	0.831	I V L	140	14	9				5	5	0	0	0	5	0	5	10	10
79	h	92	e	0.545	I T V A N S K	81	37	23	9	6	5 2	0	2	0	0	0	2	0	2	2	2
80	p	100	b (F)	1.000	P	163						3	4	0	0	0	4	0	4	7	7
81	p	99	e	0.960	R L N	161	1	1				6	0	4	0	0	4	0	4	10	10
82	p	98	b	0.814	B T F L R -	150	9	1	1	1	1	6	0	0	0	0	0	0	0	6	6
83	l	100	b (F)	0.708	L I V	105	45	13				4	3	0	0	0	3	0	3	7	7
84	b	93	e	0.537	Q L T M C H D	107	39	7	5	2	2 1	4	0	0	0	0	0	0	0	4	4
85	h	100	e	0.958	L M T	160	2	1				3	0	0	0	0	0	0	0	3	3
86	s	100	b	0.984	A V	162	1					3	0	0	0	0	0	0	0	3	3

TABLE II. (Continued)

											Residue contacts[g]									
												Inter								Total	
Pos[a]	Class[b]	MCT[c]	Access[d]	Score[e]	Variants[f]	Variants frequencies					Intra	Dimer	C|E	C|G	C|H	Prot	DNA	All	Prot	All
87	h	99	b	0.747	**I** V A L P	84	74	3	1	1	4	0	0	0	0	0	0	0	4	4
88	+	88	e	0.794	**R** A H Y	142	18	2	1		5	0	0	0	0	0	0	0	5	5

[a]Position in sequence of solved *Xenopus* structure.[3]
[b]Physiocochemical class (*1* = aliphatic; *t* = tiny; *p* = polar; *h* = hydrophobic; *b* = big; *s* = small; *c* = charged; − negative; + = positive; * = Ser/Thr; *L,V, R, etc.* = single-residue conservation; see Methods for details)
[c]Maximum conservation threshold (see Table VI)
[d]Solvent accessibility in the octamer (without DNA) (*b* buried, <9% relative solvent accessibility (RSA) (see Methods); *e* exposed, ≥9% RSA; *b(F)* interfacial residue, <9% RSA in octamer but not monomer.)
[e]SCORECONS conservation score for aligned, non-gap column (see Methods)
[f]**Bold** residues occur in the solved *Xenopus* structure. Helical regions are boxed.
[g]Combined sidechain–sidechain and sidechain–mainchain contact counts. Numbers in each contact category are the count of 'unique' contacts from both cognates in the octamer (*e.g.,* dimer contact counts for H2A derive from counts of H2A|H2B and H2A′|H2B′ contacts; duplicate contacts are counted once). Chain designators: A = H3, B = H4, C = H2A, D = H2B, E = H3′, F = H4′, G = H2A′ H = H2B′. *inter* interchain; *intra* intrachain; *dimer* heterodimer formed by interdigitation of H3|H4 or H2A|H2B histone folds; *A|E, A|G, etc.,* nondimer interchain contacts (where *e.g., A|G* includes both *A|G* and cognate *E|C* contacts); *DNA* histone|*DNA* contacts; *prot* histone-histone contact (interchain or intrachain + interchain).

for more than half of the "polar" MCT100 positions noted above in the four-state analysis of the H4 fold (Table VI).

Sequence Conservation and Contacts

Physicochemical categorization of an aligned column of residues relies on assumptions about amino acid relatedness and function. It is desirable to test a putative relationship between conservation and functional importance on a more empirical basis. One approach is to score aligned positions for conservation based on substitution patterns observed in confidently aligned blocks of related sequences.[23,26,27] We used the SCORECONS web server (see Methods) to generate conservation scores for each histone fold position; these scores correlated significantly ($P < 0.001$) to MCTs (data not shown). As a measure of functional importance in the CHFs we counted the various types of contacts, for example, intrachain, interchain, protein-DNA, made by each side chain, reasoning that for a noncatalytic structural oligomer such as the histone octamer, the primary "function" is the maintenance of a compact tertiary and quarternary conformation conducive to reversible DNA packaging, and that this function is mediated significantly (though not exclusively) by the network of side chain-to-side chain and side chain-to-mainchain contacts.

For each category of contact, we tested for correlation between the number of residues contacted by a side chain at a given position (which we hereafter refer to as the connectivity of the position) and the conservation scores for that position (Table VII). We found that for histones H2B, H3, and H4, but not H2A, the connectivity of a position correlates significantly ($P < 0.05$) to the conservation score for that position. The correlation is positive; the connectivity of a side chain tends to increase with increasing conservation of that position. When different types of connectivity were analyzed, we found that conservation score also correlated positively to the number of protein-protein contacts—and specifically, interchain protein-protein contacts—made by a residue of H2B, H3, or H4. Numbers of intrachain contacts did not correlate significantly to conservation scores, nor did protein-DNA connectivity.

Of the various interchain interactions observed in the nucleosome, two are fundamental, based on biochemical properties[33] and the structure.[3] These are the interactions that form heterodimer pairs of H3|H4 and H2A|H2B (and their mirror-image cognates H3′|H4′ and H2A′|H2B′), and the association of those dimers to form tetramers. In particular, the formation of the H4|H3-H3′|H4′ tetramer is crucial, because it is the first subassembly of the nucleosome to associate with DNA after replication.[34] For histone H3, the number of tetramer-forming contacts with H3′ was found to correlate to conservation score; no such correlation was observed for the homologous, though more labile, interaction of H4 and H2B to form the H3|H4-H2B|H2A tetramer. The only dimer interaction to evince a correlation between contact number and conservation score was that involving H4 residues, but it was not reciprocal (i.e., H3 dimerization residues did not display the correlation).

In addition to correlations, we also tested for statistically significant differences between conservation scores when positions are characterized as belonging to one of two categories, for example, buried or exposed, interchain contact or not, dimer contact or not, etc. (Table VIII). Conservation scores were significantly higher for buried residues than for exposed residues in H2B and H3, though interfacial contact positions (surface residues that are buried in protein interfaces) were not significantly more conserved than other positions. In the H3 fold, but not in others, residues making any protein contacts, or only interchain protein contacts, or only H3-H3′ (i.e., tetramer-forming) contacts tended to be more conserved than positions that did not fall into those categories. In the H4 fold, but not in others, dimer contact residues had significantly higher conservation scores than residues not involved in dimer contacts. The conservation scores of DNA-protein and intrachain contact positions were not significantly more conserved than average in any histone fold.

TABLE III. Core Histone Fold Sequence Variation and Associated Data—Histone H2B (Chains D, H)[†]

																	Residue contacts[g]									
																		Inter							Total	
Pos[a]	Class[b]	MCT[c]	Access[d]	Score[e]	Variants[f]											Variant frequencies	Intra	Dimer	D|B	D|F	D|G	Prot	DNA	All	Prot	All
34	h	100	b (F)	0.824	Y	F	W	T								154 8 7 1	4	3	0	0	0	3	0	3	7	7
35	s	72	e	0.473	S	K	G	A	N	R	P	T				62 45 26 25 8 2 1 1	1	0	0	0	0	0	0	0	1	1
36	h	94	e	0.663	I	V	L	S	T	A	P	R				113 31 12 8 3 1 1 1	0	0	0	0	0	0	2	2	0	2
37	y	100	e	1.000	Y											170 0 4 0 0 0 4 1 5 4 5										
38	h	100	b (F)	0.779	I	V	F	L								114 51 4 1	2	2	0	0	0	2	0	2	4	4
39	h	94	e	0.622	Y	F	G	S	P	R	N					110 46 4 4 3 2 1	2	0	0	0	0	0	1	1	2	3
40	+	100	e	0.854	K	R										146 24	0	1	0	0	0	1	0	1	1	1
41	s	100	b (F)	0.867	V	S										163 7	0	6	0	0	0	6	0	6	6	6
42	b	100	b (F)	0.962	L	M										162 8	3	2	0	0	0	2	0	2	5	5
43	p	100	e	0.951	K	R	T									166 3 1	2	0	0	0	0	0	0	0	2	2
		0			-	S										169 1										
		0			-	E										169 1										
		0			-	N										169 1										
		0			-	I										169 1										
		0			-	R										169 1										
		0			-	S										169 1										
44	p	97	e	0.801	Q	A	S	E	H	N						158 5 4 1 1 1	0	1	0	0	0	1	0	1	1	1
45	h	100	e	0.875	V	I	T	A								157 9 3 1	0	3	0	0	0	3	0	3	3	3
46	p	98	e	0.822	H	N	D	F	R							158 7 2 2 1	0	4	0	0	0	4	0	4	4	4
47	s	97	e	0.796	P	A	Q	S	N	T	V					157 4 4 2 1 1 1	0	0	0	0	0	0	0	0	0	0
48	p	97	e	0.790	D	Q	G	E	H	A	K	N				155 5 3 2 2 1 1 1	1	0	0	0	0	0	0	0	1	1
49	h	100	e	0.609	T	I	M	L	V							111 40 8 7 4	1	2	0	0	0	2	0	2	3	3
50	t	100	e	0.845	G	S										159 11	1	2	0	0	0	2	1	3	3	4
51	h	100	e	0.822	I	V	M	L								145 14 7 4	2	2	0	0	0	2	0	2	4	4
52	*	100	e	0.943	S	T										166 4	0	4	0	0	0	4	1	5	4	5
53	p	95	e	0.657	S	G	Q	N	H	K	R	L	T			139 6 6 5 4 4 3 2	0	0	0	0	0	0	1	1	0	1
54	p	98	e	0.767	K	R	E	P	Q	H						132 30 3 2 2 1	0	1	0	0	0	1	0	1	1	1
55	s	100	b (F)	0.752	A	S	T	G								144 14 7 5	0	4	0	0	0	4	0	4	4	4
56	h	100	e	0.927	M	V	I									164 4 2	4	0	0	0	0	0	0	0	4	4
57	s	94	e	0.556	S	G	N	K	T	V	A	E	I	L		85 60 7 6 4 2 1 1	0	0	0	3	0	3	0	3	3	3
58	l	100	b (F)	0.915	I	V										157 13	0	4	0	2	0	6	0	6	6	6
59	h	100	b (F)	0.837	M	L	V	I	T							152 8 8 1 1	2	4	0	0	0	4	0	4	6	6
60	s	100	e	0.925	N	D										164 6	3	0	0	1	0	1	0	1	4	4
61	s	99	b (F)	0.969	S	M	N									168 1 1	0	0	0	5	0	5	0	5	5	5
62	b	100	b (F)	0.886	F	Y	L	M	I							160 5 2 2 1	0	4	0	1	0	5	0	5	5	5
63	h	100	b (F)	0.811	V	I	L	M								117 50 2 1	1	3	0	0	0	3	0	3	4	4
64	P	97	e	0.838	N	R	T	V	H	G	K					157 3 3 3 2 1 1	1	0	0	2	0	2	0	2	3	3
65	c	100	b (F)	0.982	D	H										169 1	0	1	0	2	0	3	0	3	3	3
66	h	97	b (F)	0.740	I	V	L	S	T	M	N					126 30 5 3 3 2 1	0	1	0	0	0	1	0	1	1	1
67	b	100	b (F)	0.874	F	M	L									162 5 3	1	6	0	0	0	6	0	6	7	7
68	p	99	e	0.950	E	D	G	Q								165 3 1 1	0	1	0	2	0	3	0	3	3	3
69	p	100	e	0.825	R	K	N									121 48 1	1	3	0	1	0	4	0	4	5	5
70	l	100	b (F)	0.785	I	L	V									122 39 9	4	2	0	0	0	2	0	2	6	6
71	s	100	b (F)	0.907	A	C	V									165 4 1	0	2	0	0	0	2	0	2	2	2
72	s	83	e	0.446	G	A	T	Q	S	L	M	R	C	K		65 34 23 17 17 5 3 2	0	1	0	0	0	1	0	1	1	1
73	c	100	b (F)	0.982	E	D	K									168 1 1	3	0	2	0	0	2	0	2	5	5
74	t	98	b	0.779	A	S	-									132 36 2	4	0	0	0	0	0	0	0	4	4
75	t	98	e	0.731	S	A	G	R	Y							136 27 5 1 1	1	1	0	0	0	1	0	1	2	2
76	p	98	e	0.628	R	K	S	N	T	F	H	I	L	Q		110 45 6 2 2 1 1 1 1 1	0	0	0	0	0	0	0	0	0	0
77	l	100	b (F)	0.902	L	I	V									163 6 1	3	0	2	0	0	2	0	2	5	5
78	s	100	b	0.711	A	T	V	S	C	P						134 11 11 10 3 1	3	0	0	0	0	0	0	0	3	3
79	p	94	e	0.598	H	R	Q	A	K	S	T	D	G	L		78 60 11 7 4 4 2 1 1 1 1	0	1	0	0	3	4	0	4	4	4
80	h	97	e	0.683	Y	A	F	I	N	S	V	Q				147 10 3 3 2 2 2 1	0	0	3	0	0	3	0	3	3	3
81	s	98	e	0.816	N	S	T	Y	A	H						156 5 5 2 1 1	0	0	2	0	0	2	0	2	2	2
82	p	100	e	0.892	K	R	N	Q								152 16 1 1	0	0	0	0	0	0	0	0	0	0
83	p	100	e	0.765	R	K	C	N	S							84 83 1 1 1	1	0	0	0	0	0	2	2	1	3
84	s	91	e	0.545	S	P	R	D	K	N	V	A	H	L		102 45 10 2 2 2 2 1 1 1 1 1	0	1	0	0	0	1	2	3	1	3
85	T	100	e	1.000	T											170	0	1	0	0	0	1	2	3	1	3
86	l	100	b (F)	0.799	I	L	V									142 16 12	3	5	0	0	0	5	0	5	8	8
87	s	100	e	0.666	T	S	G	N								119 42 8 1	0	1	0	0	0	1	0	1	1	1
88	s	99	e	0.789	S	A	V	K	T							154 12 2 1 1	1	3	0	0	0	3	0	3	4	4
89	b	100	e	0.942	R	K	W									165 4 1	0	0	3	0	0	3	0	3	3	3
90	-	100	b	0.969	E	D										167 3	2	0	1	0	0	1	0	1	3	3
91	h	100	b (F)	0.834	I	V	L	T								142 24 3 1	2	3	0	0	0	3	0	3	5	5
92	p	100	e	0.909	Q	E	R	K								163 4 2 1	2	1	0	0	0	1	0	1	3	3
93	h	99	h	0.910	T	M	A	S								164 3 2 1	1	0	1	0	0	1	0	1	2	2
94	s	100	b	0.859	A	S	V									147 22 1	2	0	0	0	0	0	0	0	2	2
95	h	100	b (F)	0.917	V	I	T	A	G							162 4 2 1 1	3	2	0	0	0	2	0	2	5	5
96	p	100	e	0.925	R	K	C									162 7 1	1	0	0	0	0	0	0	0	1	1
97	l	99	b	0.934	L	I	S									165 4 1	1	0	4	0	0	4	0	4	5	5
98	l	100	e	0.691	L	V	T									92 58 20	3	0	1	0	0	1	0	1	4	4

[†]See Table II for key.

TABLE IV. Core Histone Fold Sequence Variation and Associated Data—Histone H3 (Chains A, E)†

														Residue contacts[g]								
															Inter						Total	
Pos[a]	Class[b]	MCT[c]	Access[d]	Score[e]	Variants[f]	Variant frequencies								Intra	Dimer	A|E	A|G	Prot	DNA	All	Prot	All
64	p	99	e	0.794	K R C A H	98	9	4	1	1				2	0	0	0	0	1	1	2	3
65	h	96	e	0.684	L A S I K T	94	13	3	1	1	1			0	0	0	0	0	2	2	0	2
66	s	99	b (F)	0.933	P S R	110	2	1						5	1	0	0	1	2	3	6	8
67	F	100	b	1.000	F	113								1	2	0	0	2	0	2	3	3
68	p	95	e	0.625	Q S A C E I M R	96	9	3	1	1	1	1		0	0	0	0	0	0	0	0	0
69	b	99	e	0.947	R A K M	110	1	1	1					0	2	0	0	2	1	3	2	3
70	l	99	b (F)	0.913	L V E I	109	2	1	1					0	5	0	0	5	0	5	5	5
71	h	100	b (F)	0.955	V A M	110	2	1						2	2	0	0	2	0	2	4	4
72	p	100	e	0.935	R K Q	107	5	1						2	1	0	0	1	1	2	3	4
73	p	99	e	0.913	K O G S	106	5	1	1					0	4	0	0	4	0	4	4	4
74	h	100	b	0.750	I V F L	95	10	5	3					0	4	0	0	4	0	4	4	4
75	s	92	b (F)	0.606	A S I C T M E V	91	7	5	3	3	2	1	1	2	1	0	0	1	0	1	3	3
76	p	88	e	0.506	Q S M V T G H R A D E L N	84	6	5	4	3	2	2	2	1	1	1	1	1	2	0	0	2	0	2	3	3
77	p	92	e	0.526	D E K A Q S G T M V	74	12	9	5	4	3	2	2	1	1			0	0	0	0	0	0	0	0	0
78	b	95	e	0.627	F Y Q S L M G I K V	85	12	5	3	2	2	1	1	1	1			0	4	0	0	4	0	4	4	4
79	p	95	e	0.574	K S T L N V A C E H R	92	6	4	2	2	2	1	1	1	1	1		0	4	0	0	4	0	4	4	4
80	s	91	e	0.492	T S D E A G P R H N Q V	71	13	7	5	4	3	3	3	1	1	1	1	0	0	0	0	0	1	1	0	1
					- G K D F N P	99	7	3	1	1	1	1														
					- S D F R	108	2	1	1	1																
					- S	112	1																			
					- T V	111	1	1																		
					- D I	111	1	1																		
					- V D Q S	106	4	1	1	1																
81	p	93	e	0.720	D E G Q S T -	90	13	6	1	1	1	1		1	1	0	0	1	1	2	2	3
82	h	95	e	0.713	L I P F C M	98	5	5	3	1	1			2	4	0	0	4	0	4	6	6
					- L N	107	5	1																		
83	p	98	e	0.870	R N H I W	107	3	1	1	1				0	2	0	0	2	5	7	2	7
84	h	100	e	0.667	F W V I M	98	8	5	1	1				5	2	0	0	2	1	3	7	8
85	p	98	e	0.734	Q T R A S	102	5	3	2	1				0	4	0	0	4	2	6	4	6
86	p	92	e	0.642	S A E G T H I K N Q R	93	5	5	2	2	1	1	1	1	1	1		0	0	0	0	0	1	1	0	1
87	s	78	e	0.481	S A Q H G D E T M N V	59	18	13	8	6	2	2	2	1	1	1		0	1	0	0	1	1	2	1	2
88	s	100	b (F)	0.941	A S G T V	108	2	1	1	1				0	4	0	0	4	0	4	4	4
89	h	100	b	0.660	V I L F	63	36	12	2					3	0	0	0	0	0	0	3	3
90	h	89	e	0.441	L M G A S E Q V D I N R	35	35	22	6	4	3	2	2	1	1	1	1	1	1	0	0	1	0	1	2	2
91	h	100	b (F)	0.934	A C L M	110	1	1	1					0	3	0	0	3	0	3	3	3
92	h	100	b (F)	0.816	L M I A	102	6	3	2					3	4	0	0	4	0	4	7	7
93	p	100	b	0.977	Q K	112	1							4	0	0	0	0	0	0	4	4
94	-	98	b	0.930	K A D F	110	1	1	1					1	2	0	2	4	0	4	5	5
95	s	100	b (F)	0.803	A S V	101	11	1						0	2	0	0	2	0	2	2	2
96	s	99	b (F)	0.514	S A C T V R	44	40	10	9	9	1			1	3	0	0	3	0	3	4	4
97	E	100	b	1.000	E	113	3	2	0	0	2	0	2	5	5							
98	s	92	e	0.611	A S M D R C N T V Y	89	8	5	4	2	1	1	1	1	1			0	0	0	1	1	0	1	1	1
99	b	98	b (F)	0.852	Y F E L N T	102	7	1	1	1	1			3	3	0	0	3	0	3	6	6
100	h	99	b (F)	0.850	L I M V R	101	5	3	3	1				0	4	0	0	4	0	4	4	4
101	h	100	b (F)	0.809	V T I	104	6	3						0	3	0	1	4	0	4	4	4
102	s	89	e	0.535	G S Q A H N C E R	75	19	6	4	4	2	1	1	1				1	0	0	0	0	0	0	1	1
103	b	100	b	0.752	L R M H	104	5	3	1					4	1	0	0	1	0	1	5	5
104	b	100	b (F)	0.775	T L M	97	13	3						1	7	0	0	7	0	7	8	8
105	p	88	e	0.638	B A C D G Q R S V	95	11	1	1	1	1	1	1	1				0	1	0	1	2	0	2	2	2
106	s	98	e	0.903	D G H L N	109	1	1	1	1				2	0	1	0	1	0	1	3	3
107	s	100	b	0.705	T S A C G V	97	8	5	1	1	1			3	0	0	0	0	0	0	3	3
108	p	92	b (F)	0.695	N Y Q H M S	100	8	2	1	1	1			1	3	0	2	5	0	5	6	6
109	b	100	e	0.816	L M R I Q	104	5	2	1	1				3	0	2	1	3	0	3	6	6
110	h	100	b (F)	0.590	C A L I	93	10	9	1					4	0	3	0	3	0	3	7	7
111	s	100	b	0.686	A T C S G V	94	7	6	4	1	1			0	0	0	0	0	0	0	0	0
112	b	98	e	0.699	I K L M A E V	99	5	4	2	1	1	1		0	0	0	2	2	0	2	2	2
113	p	100	e	0.970	H K T	111	1	1						0	0	7	0	7	0	7	7	7
114	s	95	e	0.758	A R S G C	102	5	3	2	1				0	0	2	0	2	0	2	2	2
115	p	95	e	0.723	K N G R	96	7	5	5					0	0	0	0	0	1	1	0	1
116	p	100	e	0.985	R C	112	1							2	0	1	0	1	2	3	3	5
117	h	99	e	0.920	V I A S	106	5	1	1					0	2	0	1	3	1	4	3	4
118	*	99	e	0.967	T I S	111	1	1						0	3	0	0	3	3	6	3	6
119	l	100	b (F)	0.788	I L V	99	11	3						4	4	0	0	4	0	4	8	8
120	b	98	e	0.525	M Q E F I T Y	84	14	6	3	3	2	1		0	2	0	0	2	4	6	2	6
121	s	93	e	0.626	P V S T K A E L Q R	90	6	5	4	3	1	1	1	1	1			0	4	0	0	4	0	4	4	4
122	p	99	e	0.827	K R D P T	103	7	1	1	1				1	0	1	0	1	1	2	2	3
123	p	100	b	0.977	D E N	111	1	1						5	0	1	0	1	0	1	6	6
124	h	100	b (F)	0.681	I M V L F T	81	18	7	4	2	1			3	4	0	0	4	0	4	7	7
125	b	95	e	0.690	Q H A K E Y	93	8	5	4	2	1			1	1	0	0	1	0	1	2	2
126	b	100	b (F)	0.981	L F	112	1							1	0	3	0	3	0	3	4	4
127	h	99	b	0.757	A M T V G P Y	101	5	2	2	1	1	1		4	0	1	0	1	0	1	5	5
128	b	92	e	0.629	R L A K C H M Q T	92	6	5	3	2	2	1	1	1				5	2	0	0	2	0	2	7	7
129	b	96	e	0.721	R Y C S I L	102	5	2	2	1	1			2	0	1	0	1	0	1	3	3
130	l	100	b (F)	0.792	I L	92	21							0	0	5	0	5	0	5	5	5
131	p	97	e	0.740	R C G K L -	103	6	1	1	1	1			4	1	1	0	2	0	2	6	6

†See Table II for key.

TABLE V. Core Histone Fold Sequence Variation and Associated Data—Histone H4 (Chains B, F)[†]

								Residue contacts[g]									
									Inter							Total	
Pos[a]	Class[b]	MCT[c]	Access[d]	Score[e]	Variants[f]	Variant frequencies		Intra	Dimer	B|D	B|G	B|H	DNA	Prot	All	Prot	All
31	+	100	e	0.877	K R	55 4		2	0	0	0	0	0	0	0	2	2
32	s	100	e	0.735	P G C	54 4 1		0	0	0	0	0	2	0	2	0	2
33	s	100	b (F)	0.737	A C T V	53 4 1 1		0	2	0	0	0	0	2	2	2	2
34	l	100	b	0.877	I V	55 4		4	1	0	0	0	0	2	2	5	5
35	R	100	e	1.000	R	59		2	0	0	0	0	1	0	1	2	3
36	R	100	e	1.000	R	59		0	2	0	0	0	2	2	4	2	4
37	b	100	b (F)	0.856	L M	55 4		0	6	0	0	0	0	6	6	6	6
38	A	100	b (F)	1.000	A	59		2	1	0	0	0	0	1	1	3	3
39	+	100	b (F)	0.990	R K	58 1		3	5	0	0	0	2	5	7	8	10
40	+	100	b (F)	0.968	R H	58 1		0	6	0	1	0	0	7	7	7	7
41	G	100	b (F)	1.000	G	59		0	4	0	0	0	0	4	4	4	4
42	G	100	b (F)	1.000	G	59		0	2	0	0	0	0	2	2	2	2
43	V	100	b (F)	1.000	V	59		1	4	0	0	0	0	4	4	5	5
44	K	100	e	1.000	K	59		1	5	0	1	0	1	6	7	7	8
45	R	100	e	1.000	R	59		1	3	0	0	0	5	3	8	4	9
					- T	57 2											
46	l	100	e	0.952	I L	57 2		3	2	0	0	0	2	2	4	5	7
47	s	98	e	0.927	S L N	57 1 1		0	4	0	0	0	1	4	5	4	5
48	s	98	e	0.617	G S A T K N	41 10 4 2 1 1		0	0	0	0	0	1	0	1	0	1
49	b	91	e	0.537	L E F M T A D	44 4 3 3 3 1 1		0	1	0	0	0	0	1	1	1	1
50	l	100	b (F)	0.850	I V	51 8		0	5	0	0	0	0	5	5	5	5
51	Y	100	b	1.000	Y	59		4	0	0	0	0	1	0	1	4	5
52	P	100	e	0.882	E D Q	52 6 1		1	0	0	0	0	0	0	0	1	1
53	p	100	e	0.926	E D Q	55 3 1		0	4	0	0	0	0	4	4	4	4
54	s	98	b (F)	0.699	T V S I	49 6 3 1		1	3	0	0	0	0	3	3	4	4
55	+	100	b	0.986	R K	58 1		6	0	0	0	0	0	0	0	6	6
56	s	83	e	0.508	G N R A Q S T E I	36 5 5 3 3 3 2 1 1		0	0	0	0	0	0	0	0	0	0
57	v	100	b (F)	1.000	V	59		0	5	0	0	0	0	5	5	5	5
58	b	100	b (F)	0.945	L F R	57 1 1		1	4	0	0	0	0	4	4	5	5
59	+	100	e	0.962	K R	56 3		1	1	0	0	0	0	1	1	2	2
60	h	84	e	0.500	V S I A T L G H N Q	30 7 6 5 5 2 1 1 1 1		0	1	0	0	0	0	1	1	1	1
61	a	100	b (F)	0.897	F Y	55 4		3	3	0	0	0	0	3	3	6	6
62	h	100	b (F)	0.815	L V W	54 4 1		0	6	0	0	0	0	6	6	6	6
63	p	100	e	0.985	E Q	58 1		0	2	0	0	0	0	2	2	2	2
64	s	98	e	0.671	N D S G Q V	47 4 4 2 1 1		1	0	0	0	0	0	0	0	1	1
65	h	100	b	0.861	V I A	54 4 1		4	1	0	0	0	0	1	1	5	5
66	l	100	b (F)	0.803	I V L	48 10 1		0	5	0	0	0	0	5	5	5	5
67	b	100	e	0.932	R K W	56 2 1		0	1	0	0	0	0	1	1	1	1
68	s	98	b	0.698	D C Y	54 4 1		3	0	1	0	0	0	1	1	4	4
69	t	100	b	0.704	A S	42 17		4	0	0	0	0	0	0	0	4	4
70	h	100	b (F)	0.757	V T I L	53 4 1 1		1	3	0	0	0	0	3	3	4	4
71	h	100	b	0.813	T A M	54 4 1		0	1	1	0	0	0	2	2	2	2
72	Y	98	b (F)	0.956	Y N	58 1		5	0	4	0	0	0	4	4	9	9
73	s	100	b	0.907	T C	54 5		5	0	0	0	0	0	0	0	5	5
74	E	98	e	0.980	E G	58 1		0	2	0	0	0	0	2	2	2	2
75	a	100	b (F)	0.877	H Y	55 4		0	0	5	0	0	0	5	5	5	5
76	s	100	e	0.950	A G T	57 1 1		0	0	1	0	0	0	1	1	1	1
77	p	100	e	0.778	K R Q	37 21 1		0	0	1	0	0	1	1	2	1	2
78	b	100	e	0.852	R K I	54 4 1		3	1	0	0	0	1	1	2	4	5
79	c	98	e	0.917	K E P R	55 2 1 1		1	4	0	0	0	2	4	6	5	7
80	T	100	e	1.000	T	59		1	4	0	0	0	2	4	6	5	7
81	V	100	b (F)	1.000	V	59		3	5	0	0	0	0	5	5	8	8
82	T	100	e	1.000	T	59		0	2	0	0	0	0	2	2	2	2
83	t	100	b (F)	0.852	A S	51 8		1	3	0	0	0	0	3	3	4	4
84	h	94	e	0.591	M L C S Q	41 13 2 2 1		3	0	0	0	0	0	0	0	3	3
85	s	100	b	0.956	D A	58 1		5	0	0	0	0	0	0	0	5	5
86	l	100	b (F)	0.985	V I	58 1		3	3	0	0	0	0	3	3	6	6
87	h	98	b	0.936	V G I R	56 1 1 1		1	0	0	0	0	0	0	0	1	1
88	h	96	e	0.715	Y T D N	54 3 1 1		1	0	1	0	0	0	1	1	2	2
89	s	98	b	0.904	A S R V	55 2 1 1		4	0	0	0	0	0	0	0	4	4
90	L	100	b (F)	1.000	L	59		5	1	0	0	0	0	1	1	6	6
91	+	100	e	0.877	K R	55 4		2	0	1	0	1	0	2	2	4	4
92	b	100	e	0.841	R K H I	53 4 1 1		2	0	3	0	0	0	3	3	5	5
93	p	100	e	0.932	Q R S	57 1 1		4	0	0	0	0	0	0	0	4	4

[†]See Table II for key.

The foregoing results suggest that histone variation can be explained largely in terms of retention of physicochemical properties and (for H2B, H3, and H4) connectivity. In these folds, highly conserved positions tend to make many contacts—especially interchain protein contacts. The importance of the H4|H3-H3′|H4′ tetramer interface in par-

TABLE VI. Four-state Physicochemical Distribution of MCT100 Positions in Core Histone Fold Alignments

	# MCT100 positions (% MCT100 positions/% length)			
Class	H2A	H2B	H3	H4
Hydrophobic	11 (39/18)	15 (38/23)	11 (42/16)	13 (27/21)
Polar	5 (18/8)	12 (30/18)	6 (23/9)	17 (35/27)
Small	8 (29/13)	8 (20/12)	4 (15/6)	10 (21/16)
Big	4 (14/6)	5 (13/8)	5 (19/7)	8 (17/13)

TABLE VII. Results of Test for Correlation between Connectivity and Conservation Score, for Different Types of Contacts†

	Correlation w/conservation?			
Contact type[a]	H2A	H2B	H3	H4
All	no	yes	yes	yes
Intra	no	no	no	no
Dimer	no	no	no	yes
DNA	no	no	no	no
Prot	no	yes	yes	yes
Inter	no	no	yes	yes
Inter prot	no	yes	yes	yes
A|E	na[b]	na	yes	na
A|G	na	na	no	na
B|D	na	na	na	no
B|G	na	na	na	no
B|H	na	na	na	no
C|E	no	na	na	na
C|G	yes (−)[c]	na	na	na
D|F	na	no	na	na
D|G	na	no	na	na

†Connectivity refers to how many other residues are contacted by a sidechain. Conservation score is derived from the SCORECONS server analysis of alignments. Truth table was derived from Spearman Rank Order correlation analysis ($p < 0.05$).
[a]See Table II for key.
[b]na = not applicable.
[c]The correlation was negative.

TABLE VIII. Results of Test for Significant Difference between Conservation Scores of Classification Group vs. Background†

	Significant difference from background?			
Class[a]	H2A	H2B	H3	H4
Contact type				
Intra	no	no	no	no
Inter prot	no	no	yes	no
Dimer	no	no	no	yes
DNA	no	no	no	no
Inter	no	no	no	no
Prot	no	no	yes	yes
A|E	na	na	yes	na
A|G	na	na	no	na
B|D	na	no	na	no
B|G	na	na	na	no
B|H	na	na	na	no
C|E	no	na	na	na
C|G	yes (−)[b]	na	na	na
C|H	no	na	na	na
D|F	na	no	na	na
D|G	na	no	na	na
Accessibility				
Buried	no	yes	yes	no
b (F)	no	no	no	no

†Truth table derived from Mann-Whitney Rank Sum tests ($p < 0.05$). Background refers to scores for positions that do not make contacts of the specified type.
[a]See Table 2 for key.
[b]Scores were significantly lower than background.

ticular may have constrained variation at some H3 positions. On the other hand, in the H2A fold, connectivity did not relate to sequence variation in any way we could detect, perhaps because H2A sequences display the greatest number of CHF variants, and comprise the greatest number of functionally distinct subclasses (e.g., H2A, macroH2A, H2A.X and H2A.Z).[6,7] It is perhaps surprising that little or no importance could be attributed to DNA-protein contacts, dimer contacts, or intrachain contacts, made by any CHFs, as one might expect any of these contact types to be important enough to constrain variation. In the case of DNA-histone interaction, however, the more important contacts appear to be between the protein backbone and the double helix,[3] which were not counted in our analysis.

Connectivity is not the only possible impetus behind variational constraint, though it is convenient to categorize positions based on connectivity and conservation. Highly connected, highly conserved, and poorly connected, poorly conserved positions constitute two categories that are amenable to analysis by the methods we used. But what about the other two types of positions: highly connected but poorly conserved, and highly conserved but sparsely connected? In the former case, the connectivity itself may not be conserved—it requires comparison to nucleosome structures from other species, for example, yeast[11] to resolve the question. In the second case, the few connections the residue does make may be functionally crucial, or the residue has a "steric" function such as allowing the backbone to adopt a bend, as can be the case for glycine. A residue might also be conserved because it makes contacts with nonhistone proteins (e.g., side-chain modification enzymes and transcription factors).

Inventory and analysis of sequence variation should also be integrated with mutational studies of protein function. Such studies can benefit from foreknowledge of natural variations, which can help the investigator choose a site likely to have functional importance. Obvious examples in histones include the absolutely invariant positions of histone H4 (i.e., R35 and R36 in helix 1, G41 to R45 in interhelical loop 1, Y51 and V57 in helix 2, T80 to T82 in loop 2, and L90 in helix 3), H2A (E56 of helix 2 and P80 in helix 3), H2B (Y37 in helix 1 and T85 in loop 2), and H3 (F67 in helix 1 and E97 in helix 2), each of which allow occupancy by only one amino acid. Several of these positions have surface exposure and therefore could be involved in nucleosome-nucleosome or nucleosome-protein interactions (Fig. 3).

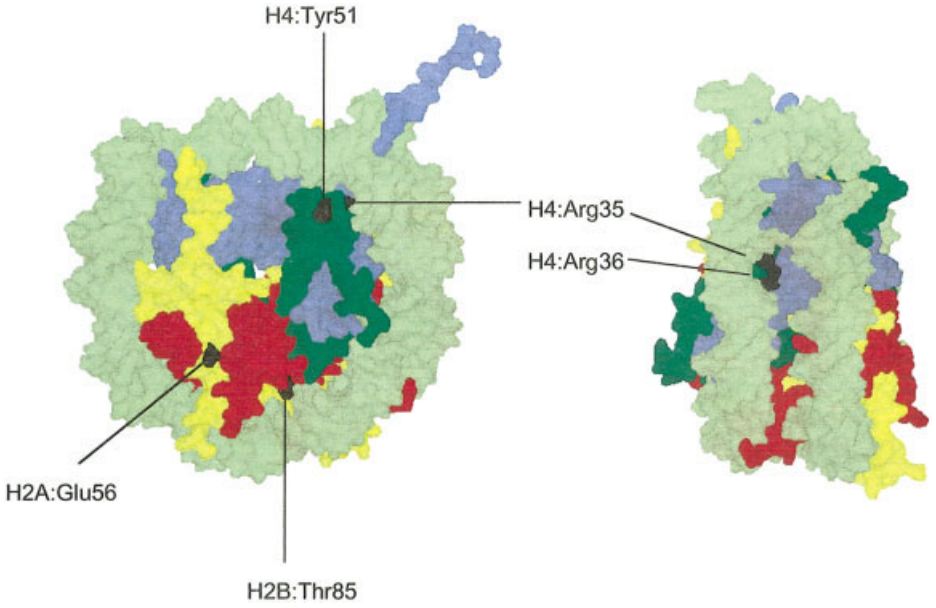

Fig. 3. Surface exposure of invariant residues. Selected solvent-accessible polar or charged invariant residues are illustrated in black on a surface map of the nucleosome core particle (left, view down the superhelical axis; right, side view, with H3 tail facing the reader). Light green, DNA; yellow, H2A; red, H2B; blue, H3; green, H4.

Examination of histone variation can also complement results of mutational studies. For example, Ng[35] recently reported that a substitution of A, P, or Q for methylated K79 in the first interhelical strand of a *Drosophila* H3 variant impairs telomeric gene silencing. Natural variants of H3 K79, a position with a low conservation score, are S, T, L, N, V, A, C, E, H, and N. It is implausible that all of these variants can be methylated, but examination of their source sequences reveals that all of the variants (but not K) occur in centromeric H3 proteins or in proteins that cluster with them in phylogenetic analysis (data not shown). Because these are unlikely to associate with telomeric nucleosomes, the natural variation accords with the hypothesis that K79 methylation is important for telomeric silencing. Another recent report[36] showed that mutation of any of three H3 CHF positions from the predominant variant to the H3.3 variant (S87→A, V89→I, and M90→G) releases H3 from strictly replication-dependent nucleosome deposition. In our inventory, these positions all display relatively low MCT values and conservation scores, as one might expect of positions that demarcate multiple functionally distinct paralogs. It would be interesting to see what happens when the other natural variants at these positions are used in such mutation studies.

REFERENCES

1. Thomas JO, Kornberg RD. Cleavable cross-links in the analysis of histone-histone associations. FEBS Lett 1975;58:353–358.
2. Arents G, Burlingame RW, Wang BC, Love WE, Moudrianakis EN. The nucleosomal core histone octamer at 3.1 A resolution: a tripartite protein assembly and a left-handed superhelix. Proc Natl Acad Sci USA 1991;88:10148–10152.
3. Luger K, Mader AW, Richmond RK, Sargent DF, Richmond TJ. Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature 1997;389:251–260.
4. Sullivan S, Sink DW, Trout KL, Makalowska I, Taylor PM, Baxevanis AD, Landsman, D. The histone database. Nucleic Acids Res 2002;30:341–342.
5. Pereira SL, Reeve JN. Histones and nucleosomes in Archaea and Eukarya: a comparative analysis. Extremophiles 1998;2:141–148.
6. Brown DT. Histone variants: are they functionally heterogeneous? Genome Biol 2001;2.
7. Ausio J, Abbott DW, Wang X, Moore SC. Histone variants and histone modifications: a structural perspective. Biochem Cell Biol 2001;79:693–708.
8. Von Holt C, Strickland WN, Brandt WF, Strickland MS. More histone structures. FEBS Lett 1979;100:201–218.
9. Jenuwein T, Allis CD. Translating the histone code. Science 2001;293:1074–1080.
10. Harp JM, Hanson BL, Timm DE, Bunick GJ. Asymmetries in the nucleosome core particle at 2.5 A resolution. Acta Crystallogr D Biol Crystallogr 2000;56 Pt 12:1513–1534.
11. White CL, Suto RK, Luger K. Structure of the yeast nucleosome core particle reveals fundamental changes in internucleosome interactions. Embo J 2001;20:5207–5218.
12. Walker DR, Koonin EV. SEALS: a system for easy analysis of lots of sequences. Proc Int Conf Intell Syst Mol Biol 1997;5:333–339.
13. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 1997;25:4876–4882.
14. Hawksworth DL, Kalin-Arroyo MT. Magnitude and distribution of biodiversity. In: Heywood VH, editor. Global biodiversity assessment. Cambridge: Cambridge University Press; 1995. p 107–191.
15. Raghava GPS. A graphical Web server for the analysis of protein sequences and alignment. Biotech Softw Int Rpt 2001;2:254–257.
16. Taylor WR. The classification of amino acid conservation. J Theor Biol 1986;119:205–218.
17. Livingstone CD, Barton GJ. Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. Comput Appl Biosci 1993;9:745–756.
18. Goodstadt L, Ponting CP. CHROMA: consensus-based colouring of multiple alignments for publication. Bioinformatics 2001;17:845–846.
19. Andersen CA, Palmer AG, Brunak S, Rost B. Continuum secondary structure captures protein flexibility. Structure (Camb) 2002; 10:175-184.
20. Rost B, Sander C. Conservation and prediction of solvent accessibility in protein families. Proteins 1994;20:216–226.

21. Valdar WS, Thornton JM. Protein-protein interfaces: analysis of amino acid conservation in homodimers. Proteins 2001;42:108–124.
22. Valdar, WS. Scoring residue conservation. Proteins 2002;48:227–241.
23. Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 1992;8:275–282.
24. Karlin S, Brocchieri L. Evolutionary conservation of RecA genes in relation to protein structure and function. J Bacteriol 1996;178: 1881–1894.
25. Shindyalov IN, Bourne PE. WPDB: a PC-based tool for analyzing protein structure. J. App. Cryst. 1995;28:847–852.
26. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 1992;89:10915–10919.
27. Dayhoff MO, Schwartz RM, Orcutt BC. A model of evolutionary change in proteins: matrices for detecting distant relationships. In: Dayhoff, MO, editor. Atlas of protein sequence and structure. Washington: National Biomedical Research Foundation; 1978. Vol. 5, p 345–358.
28. Patthy L. 1999. Protein evolution. Oxford: Blackwell Science; 1999. 228 p.
29. Zvelebil MJ, Barton GJ, Taylor WR, Sternberg MJ. Prediction of protein secondary structure and active sites using the alignment of homologous sequences. J Mol Biol 1987;195:957–961.
30. Livingstone CD, Barton GJ. Identification of functional residues and secondary structure from protein multiple sequence alignment. Methods Enzymol 1996;266:497–512.
31. Suzuki T, Imai K. Evolution of myoglobin. Cell Mol Life Sci 1998;54:979–1004.
32. Klyszejko-Stafanowicz L, Krajewska WM, Lipinska A. Histone occurrence, isolation, characterization, and biosynthesis. In: Hnilica LS, Stein GS, Stein JL, editors. Histones and other basic nuclear proteins. Boca Raton: CRC Press; 1989. p 17–72.
33. Eickbush TH, Moudrianakis EN. The histone core complex: an octamer assembled by two sets of protein-protein interactions. Biochemistry 1978;17:4955–4964.
34. Worcel A, Han S, Wong ML. Assembly of newly replicated chromatin. Cell 1978;15:969-977.
35. Ng HH, Feng Q, Wang H, Erdjument-Bromage H, Tempst P, Zhang Y, Struhl K. Lysine methylation within the globular domain of histone H3 by Dot1 is important for telomeric silencing and Sir protein association. Genes Dev 2002;16:1518–1527.
36. Ahmad K, Henikoff S. The histone variant H3.3 marks active chromatin by replication-independent nucleosome assembly. Mol Cell 2002;9:1191–1200.