The Molecular and Evolutionary Principles of Histone Folding in Eukarya and Archaea

Haiqing Zhao,[*,†,§] Hao Wu,[†] Alex Guseman,[‡] Dulith Abeykoon,[‡] Christina M. Camara,[‡] Yamini Dalal,[*,¶] David Fushman,[*,†,‡] and Garegin A. Papoian[*,†,‡]

†*Biophysics Program, Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742, United States*

‡*Department of Chemistry and Biochemistry, University of Maryland, College Park, Maryland 20742, United States*

¶*Laboratory of Receptor Biology and Gene Expression, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, United States*

§*Current address: Department of Systems Biology, Columbia University, New York, New York 10032, United States*

E-mail: hz2592@columbia.edu; dalaly@mail.nih.gov; fushman@umd.edu; gpapoian@umd.edu

(Last modified: October 15, 2022)

Abstract

Histones are the dominant chromosomal proteins that compact and store DNA in both Eukarya and Archaea. Despite their biological importance, the general mechanism that underlies and distinguishes the folding of histones and histone-like proteins has not been well established. Here, we used the Associative Memory, Water Mediated, Structure and Energy Model (AWSEM) and all-atom molecular dynamics (MD) simulations, AI-based structure prediction tool AlphaFold2, Nuclear Magnetic Resonance

(NMR) and circular dichroism (CD) experiments to show that histone monomers are mostly disordered in isolation but adopt ordered structure upon binding with their dimeric partner. This "folding upon binding" mechanism of histone is found to be conserved across eukaryotic and archaeal histones, and remarkably, even in histone-like transcription factors such as $dTAF_{II}$ and NF-Y. We unexpectedly found that histone cores without tails may form an alternative conformation, namely, a non-native, inverted dimer. We attributed the surprising energetic stability of this state to be a consequence of the ancient sequence symmetry underlying all histone proteins. Therefore, our results strongly support the previously proposed histone evolution hypotheses. Finally, AWSEM, AlphaFold2 and atomistic simulations indicate that the eukaryotic histones may also form stable homodimers, just as archaeal histones, whereas their disordered tails tip the balance towards formation of histone heterodimers.

INTRODUCTION

In Eukarya, histones are fundamental proteins for chromosome packaging. In the basic unit of chromosome, called nucleosome, histones are assembled by four pairs of heterodimers, among which two H3/H4 dimers form a tetramer and H2A/H2B participate as two dimers[1]. Besides the canonical histones composing the majority of eukaryotic nucleosomes, histone variants and other histone-like proteins carry out specific functions in the nucleus[2,3]. In Archaea, histones are encoded to package and compact DNA into continuous hypernucleosomes[4,5]. Among many different histone oligomerization variations, dimer is the smallest unit reported in Eukarya and Archaea, with regard to both structure and function. However, in Eukarya histones exist only as hetero-dimers containing long terminal tails, while in Archaea as both homo- and hetero-dimer forms, typically without tails.

Despite the functional and sequence diversities across species, histones and histone-like proteins possess the same structural motif, known as the histone-fold, which is comprised of three alpha helices connected by two loops[6]. Two histone-fold monomers assemble into

a "handshake" motif forming a dimer in an intertwined, head-to-tail manner[7] (Figure 1A). A scaling analysis of radius of gyration (R_g) as a function of the protein size suggests that histone dimer acts as the minimum folding unit, rather than its monomer (Figure 1B, supplemental session S1). Indeed, experimentally, it was known that eukaryotic histones fold and form complexes only in presence of their binding partner[8,9]. The stability of H2A/H2B dimers and $(H3/H4)_2$ tetramers were studied by a series of denaturation experiments[10–13]. Karantza *et al.* reported that during unfolding of either H2A/H2B or $(H3/H4)_2$, individual folded monomers were not detectable, indicating a direct transition from one folded histone dimer to two unfolded monomers. However, the molecular principles underlying this cooperative behavior have not been elucidated. In addition, how these principles apply to histone variants, archaeal histones, and other histone-like structures such as transcription factors, remain unknown.

Deeper insights into the folding dynamics of histone and histone-like proteins may shed light on the evolutionary basis and divergence of canonical and variant histones[2], and help understand their higher level structural organizations such as tetramer or octamer formation[14,15], nucleosome interactome[16,17], and chromosome organization[18–22]. In this work, we aim to gain mechanistic understanding of histone folding at molecular detail and more broadly, to compare the folding mechanisms of histones across Eukarya and Archaea.

To address these questions, we first employed molecular dynamics (MD) simulations based on a coarse-grained protein force field, AWSEM[24]. AWSEM has been successfully applied to predict structures of proteins and their complexes[24,25]. We used AWSEM to study folding of various histones related to H2A/H2B, finding that both canonical and variant histones are unstable as monomers. In agreement with the abovementioned prior experiments, our simulations show that two histone monomers cooperatively fold into a stable complex, revealing folding-upon-binding dynamics. We complemented our computational predictions by experimental investigations using Nuclear Magnetic Resonance (NMR) spectroscopy and circular dichroism (CD). These experimental data also indicate that histones adopt an ordered

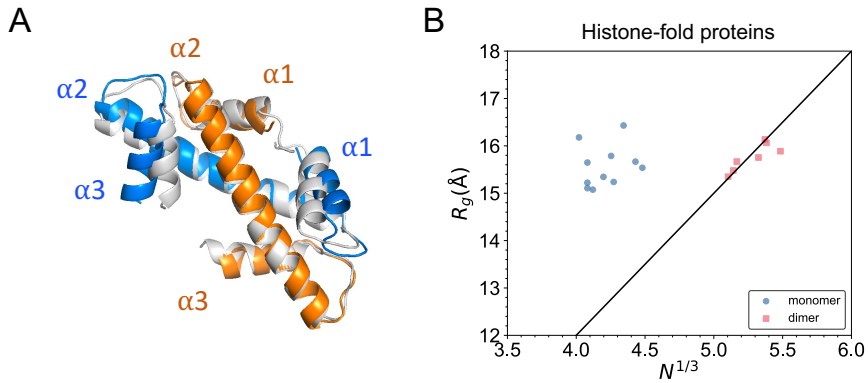


Figure 1: **Structure information of eukaryotic and archaeal histones.** (A) The main scaffold of an histone dimer consists of three α helices connected by two loops. Shown is the aligned native structures of eukaryotic histone H2A/H2B (blue/orange, PDB: 1AOI), and archaeal histone $(\text{HmfB})_2$ (grey, PDB: 1A7W). (B) Polymer scaling fit of histones suggests histone dimers act more as "monomeric" proteins. R_g is plotted as a function of the residue number N for the histone(-like) monomer and dimer proteins that are included in this study. Monomers and dimers are marked with different colors and marker types. The black line is the empirical relation of R_g and N from 403 globular monomeric proteins[23].

structure only upon binding their partners at equimolar stoichiometry.

Our calculations revealed an unexpected non-native conformation of the histone dimer which has low free energy. The free energy barrier between the native and non-native states was estimated to be around 9 kcal/mol. Subsequent analysis of related protein sequences indicated that our observation of non-native conformation is consistent with the well-known evolution hypotheses of the histone fold[26], which implies a hidden sequence symmetry in contemporary histones. Moreover, our simulations suggest that archaeal histones and other histone-like proteins including transcription factors also exhibit folding upon binding, with the abovementioned non-native dimer conformation potentially being a low energy state.

To further probe these ideas, we computationally constructed a hybrid structure of an archaeal histone heterodimer, and investigated its folding dynamics. We applied the AWSEM model as well as the deep-learning structure prediction algorithm AlphaFold2 (AlphaFold-Multimer)[49,50] to predict the structure of the hypothetical eukaryotic histone homodimer. Structures obtained from these two independent prediction methods were compared and their stabilities were examined through extensive all-atom simulations. Together, these sim-

ulations suggested the important role of histone tails in determining folding and structures of various histone dimers.

METHODS

MD Simulations and AlphaFold2

Coarse-grained MD simulations were carried out using LAMMPS package with AWSEM[24] model, under non-periodic shrink-wrapped boundary condition and Nose-Hoover thermostat control. AWSEM is a coarse-grained protein force field inspired by the free energy landscape theory[27]. Its Hamiltonian contains both physical interaction terms and a bioinformatics-inspired memory term, $V_{AWSEM} = V_{backbone} + V_{contact} + V_{burial} + V_{H-bond} + V_{FM}$. Details of every term are described in Davtyan *et al.*[24] and Papoian *et al.*[28]. AWSEM model has been applied to predict protein structures[24], protein-protein interactions[25], and in particular to histone dynamics such as chaperone-assisted histone dimer[29], tetramer and octamer[14], histone tails[30], nucleosome[31] and linker histone[32].

To find low-energy protein structures, simulated annealing[33] was executed by gradually decreasing the simulation temperature from 600 K to 200 K. The conformation that has the lowest energy is chosen as the prediction outcome of this optimization. The order parameter, Q , defined as $\frac{1}{N}\sum_{i<j-2} exp[-\frac{(r_{ij}-r_{ij}^{N})^2}{2\sigma_{ij}}]$, was used to quantify the structural similarity to the corresponding native structure. To compute a comprehensive binding free energy profile, we applied the umbrella sampling technique with Q parameter as the reaction coordinate in a harmonic potential form. The weighted histogram analysis method (WHAM)[34] was used to remove the potential bias to calculate the free energy profiles. Protein sequences of all studied systems are provided in supplemental information (SI session S2). More AWSEM simulation details are provided in the SI session S3.1.

The all-atom MD simulations were performed using the high-performance MD simulation toolkit OpenMM 7.6.0[35] on GPUs, with the atomic force field Amber ff14SB for proteins

and the TIP3P water model for solvent. Initial conformations were either from the crystal structure, or final predictions of AWSEM-MD and AlphaFold2. All the simulated systems were solvated in a 150 mM KCl solution. In total, eight molecular systems were separately simulated with two independent replicas for each system. After standard energy minimization and equilibration (details in SI), each replica was run for 800 ns for subsequent analyses. Results in the main context are from one replica. Complete all-atom simulation details are available in the SI (session S3.2).

The AlphaFold2 (AlphaFold-Multimer) predictions for protein-protein complex were carried out using ColabFold[36] on cloud service platform Google Colaboratory (colab pro), with the fast homology search method MMseqs2[37], the paired alignment and amber relaxation options.

NMR and CD Experiments

Unlabeled and ^{15}N labeled histones human H2A (type 2-A) and H2B (type 1-C) were expressed in *E. coli* and purified from inclusion bodies using cation exchange chromatography. Their correct mass was confirmed by mass spectrometry. All NMR experiments were performed at 23°C on a 600 MHz Bruker Avance-III NMR spectrometer equipped with TCI cryoprobe. Proteins were dissolved at 100-200 μM in 20 mM sodium phosphate buffer (pH 6.8) containing 7% D_2O and 0.02% NaN_3. NMR data were processed using TopSpin (Bruker Inc.)

Circular dichroism (CD) spectra of histone proteins at 0.40 mg/mL concentration in 20 mM sodium phosphate buffer at pH 6.8 were acquired on a Jasco J810 Spectro-Polarimeter using a Peltier-based temperature-controlled chamber at 25°C and a scanning speed of 50 nm/min. A quartz cell (1.0 mm path length) was used. All measurements were performed in triplicate. To determine the secondary structure content, the CD data were analyzed using the DichroWeb server[38]. Two methods were used in parallel: (i) CDSSTR, a singular value decomposition (SVD)-based approach employing two datasets (7 and SMP180) from the

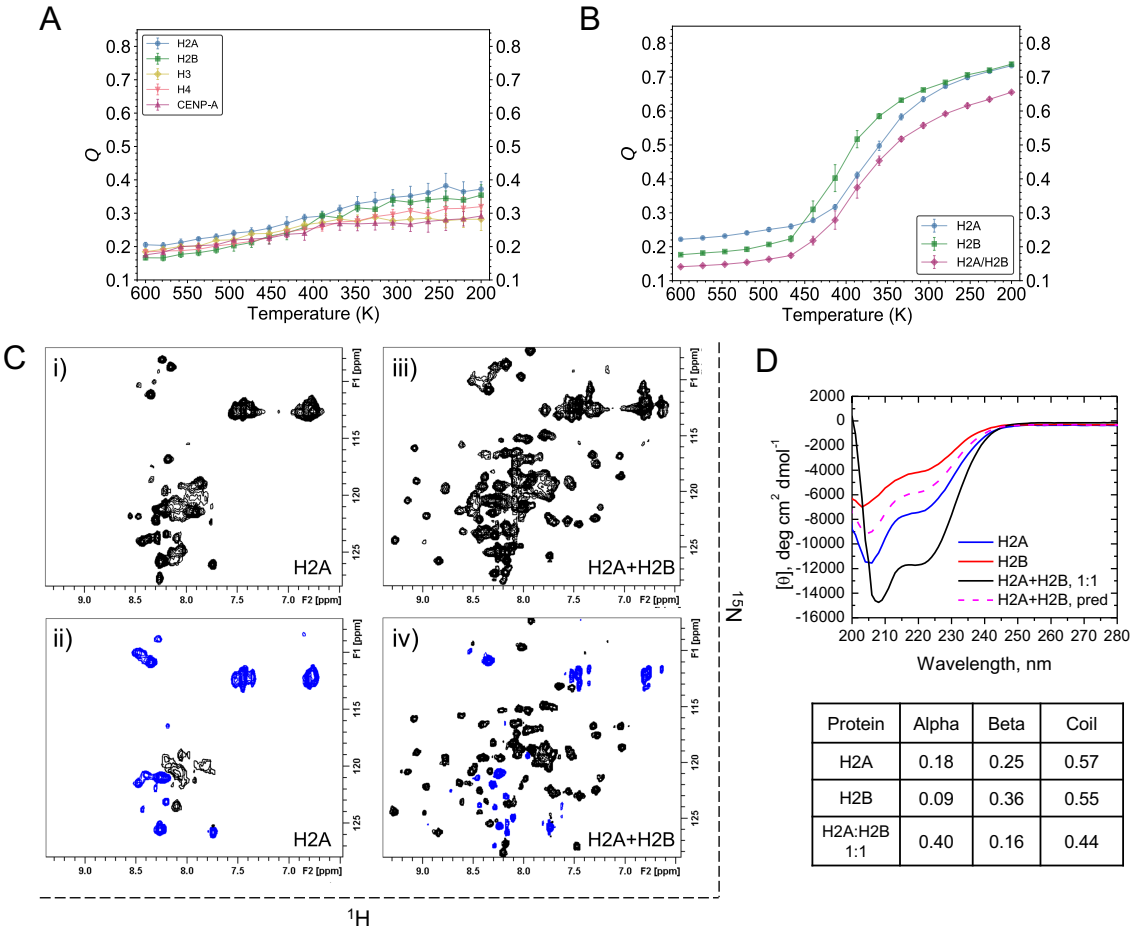
DichroWeb server and (ii) K_2D, a neural network-based algorithm trained using reference CD data[39].

RESULTS

Histones Only Fold Upon Binding with Their Partners

Individual folded monomers were not detectable in the past unfolding experiments of histone dimers or tetramers[10,11]. Our first aim was to investigate the molecular basis for this observation using AWSEM-MD. For that, we simulated canonical histone monomers H2A, H2B, H3, H4, and a centromeric variant histone type CENP-A. Ten independent temperature annealing runs were carried out. Our simulations indicate that histone monomers are highly disordered at the tertiary structure level, as evidenced by Q values less than 0.4 (Figure 2A). Structures of $Q < 0.4$ are indicative of far-from-native conformations. The corresponding root-mean-square deviations (RMSD) from the native structure are more than 8 Å. Moreover, we calculated pairwise RMSDs among the final ten conformations from different simulation runs and performed clustering analysis based on mutual structural similarity. Even the smallest RMSD for H2A/H2B simulations was 3.8 Å, indicating that no consistent tertiary structure emerges (Figure S2). On the other hand, histone monomers exhibit stable secondary structure elements, especially the long $\alpha 2$ helix, which are similar to the native structure. We noticed that without the stabilizing contacts from another histone monomer, the partially formed two short helices $\alpha 1$ and $\alpha 3$ of one histone are typically highly mobile, moving around the $\alpha 2$ helix, leading to significant tertiary disorder (Figure S3).

We next applied a similar annealing protocol to investigate folding of histones in the presence of their cognate binding partner. Taking H2A/H2B as an example, the contact map of the predicted structure shows that 98% of its native contacts, either intra- or inter-chain, were correctly predicted, in the absence of any bias towards these contacts by the AWSEM force field (Figure S5). We calculated the Q values of the entire dimer, and of its component



Protein	Alpha	Beta	Coil
H2A	0.18	0.25	0.57
H2B	0.09	0.36	0.55
H2A:H2B 1:1	0.40	0.16	0.44

Figure 2: **Histone heterodimers fold in MD simulations and NMR experiments, not monomers** (A) Q values as a function of the annealing temperature are plotted for histone monomers H2A, H2B, H3, H4, and CENP-A, with the mean and standard deviation displayed as circles and error bars. (B) Q analysis shows that H2A/H2B annealing simulations undergo cooperative folding and binding between H2A (blue) and H2B monomers (green). The Q values of H2A/H2B dimer are shown in magenta. (C-D) NMR and CD studies of H2A and H2B upon their complex formation. (C) ^{1}H-^{15}N NMR spectra of ^{15}N-labeled H2A alone (i) and in the presence of unlabeled H2B at a 1:1 molar ratio (iii). Heteronuclear steady-state ^{15}N{^{1}H} NOE spectra with amide proton presaturation recorded for ^{15}N-labeled H2A alone (ii) and in the presence of unlabeled H2B at an equimolar ratio (iv). In these spectra, contours with positive intensities are colored black while negative intensities are blue. (D) CD spectra of H2A (blue) and H2B (red) alone and in an equimolar mixture (black). The total concentrations of the proteins are the same in all three cases. Also shown (dashed magenta) is the expected CD spectrum of the equimolar mixture of H2A and H2B if there were no structural changes in either protein. The table shows the secondary structure composition of the proteins estimated from these experimental CD data using K_2D algorithm (see Methods for details).

monomers. As shown in Figure 2B, Q_{dimer} increases roughly concurrently with $Q_{monomer}$, indicating a clear structural transition wherein the two interacting monomeric chains cooperatively fold and bind. Similar simulations and analyses were undertaken for other histones, histone variants and histone-like proteins including H3/H4, CENP-A/H4, archaeal histones $(HMfA)_2$, $(HMfB)_2$ and HMfA/HMfB, transcription factors $dTAF_{II}42/dTAF_{II}62$ dimer and NF-YB/NF-YC dimer (Figure S4, S5, S6, S7). Our simulations suggest that all histones only fold upon binding with their partners. We note that the disordered histone tails were not included in these simulations. A previous experimental study of H2A/H2B indicated that the histone tails may accelerate the folding speed of histones[40]. The detailed effects of histone tails on H2A/H2B dimer binding and folding are discussed in a separate result session below.

To experimentally probe histone dimer folding, we carried out NMR and CD measurements on H2A and H2B. ^{1}H-^{15}N NMR spectrum of H2A alone (Figure 2C i) shows a narrow spread of NMR signals resulting in signal crowding in the region typical for amide signals of unstructured/unfolded proteins. The negative or close to zero signal intensities observed in the heteronuclear steady-state NOE spectrum of ^{15}N-labeled H2A recorded upon presaturation of amide protons (Figure 2C ii) are a clear indication that the protein is unstructured and highly flexible. Upon addition of unlabeled H2B, we observed a dramatic change in the ^{1}H-^{15}N NMR spectra of ^{15}N-labeled H2A, wherein new signals (corresponding to the bound state) emerge and increase in intensity until they saturate at ca. equimolar H2B:H2A ratio (Figure 2C iii). Concomitantly, the unbound signals reduce in intensity and practically disappear at the saturation point. This behavior of the NMR signals, which exhibit essentially no gradual shifts, indicates that the binding is in slow exchange regime on the NMR chemical shift time scale. In contrast to the unbound state, the signals of ^{15}N-labeled H2A in complex with H2B (Figure 2C iv) show a significant spread, indicating that the bound state of H2A is well structured. Also, many H2A signals in the heteronuclear steady-state NOE spectra recorded at these conditions have positive intensities, which are characteristic

of a well-folded state of the protein (see Fushman *et al.*[41]). A similar behavior was observed for ^{15}N-labeled H2B, which is unstructured in the unbound state and folds upon complex formation with H2A (Figure S9).

The NMR data presented above suggested that only with each other can H2A and H2B fold into a histone dimer with well-defined structures. To extend this analysis further, we performed CD measurements, which allow one to assess the helical content of a protein experimentally. The CD results demonstrate a significant increase in the helical content of these proteins upon formation of the H2A/H2B heterodimer (Figure 2D). Together, these experimental results indicate that in isolation H2A and H2B are disordered but adopt a well-defined tertiary structure upon binding to each other, which is consistent with a previous experimental study of H2A/H2B's thermodynamic stability.[10] The 40% helical content of the human H2A/H2B heterodimer observed here is in excellent agreement with that for *Xenopus laevis* H2A/H2B[42]. Overall, our experiments and the above elaborated simulations are in qualitative agreement.

Inverted Histone Dimer is Surprisingly Stable

Besides the native conformation, we observed an interesting non-native structure in our annealing simulations of core histone dimers. Compared to that of the native complex, the secondary structural elements of the non-native complex show very little differences but are arranged differently leading to an inverted tertiary structure of histone complex. The $\alpha1$ helix of one histone is in proximity of the $\alpha3$ helix of the other (Figure 3A, right), instead of the interacting two $\alpha1$ helices as in the native conformation (Figure 3A, left). If we define a direction pointing from N- to C-terminus for each histone, the angle between H2A $\alpha2$ and H2B $\alpha2$ in the inverted conformation is nearly complementary to that of the native structure (Figure 3A). The inverted conformation was found in the annealing dimer simulations of all the histone fold systems in this study, including eukaryotic and archaeal histones and transcription factors (Figure S7). Interestingly, the native and inverted H2A/H2B complexes

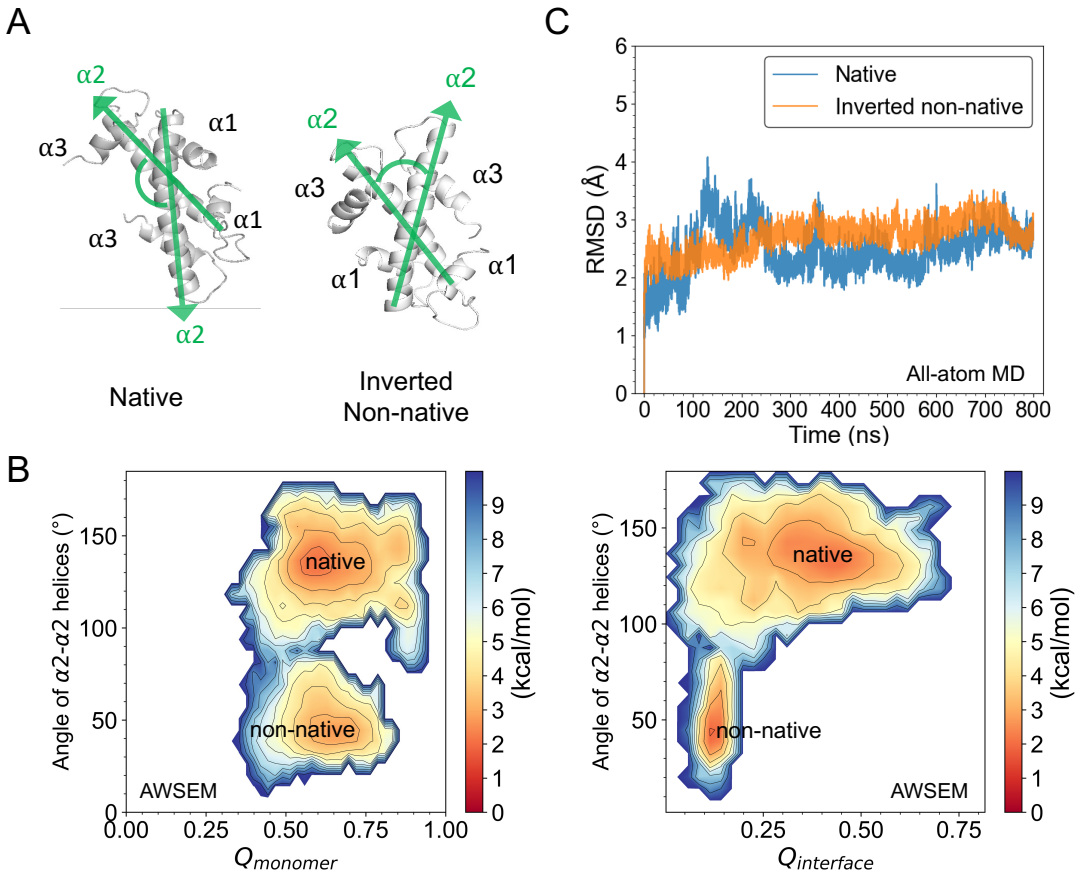


Figure 3: **Inverted non-native conformation is found to be a stable formation of histone dimer** (A) The native and non-native conformation of H2A/H2B found in AWSEM simulations are shown. Their major difference is measured by the angle between the α2-α2 helices (green) with each vector arrow pointing from N- to C-terminal. (B) Free energy profiles of H2A/H2B are projected on $Q_{monomer}$ and the angle of α2-α2 helices (left), and on $Q_{interface}$ and the α2-α2 helices angle (right). Two energy minima are found, corresponding to the native-like and inverted non-native conformation respectively. (C) All-atom simulations show comparable stability of inverted non-native conformation to that of native structure of histone dimer.

are energetically nearly comparable according to the AWSEM potential (Table S8).

To probe the free energy (FE) landscapes of native and inverted configurations, we performed umbrella sampling simulations of H2A/H2B using AWSEM. The calculated free energy profiles were further projected onto the following reaction coordinates: the angle between two $\alpha 2$ helices to describe the general inter-chain geometry (Figure 3A), $Q_{monomer}$ to show how well histone monomer is folded, and $Q_{interface}$ to show how well the dimeric interface is formed. In both of the FE landscapes (Figure 3B, left and right panels), there are two major FE basins. The first basin has the $\alpha 2$-$\alpha 2$ angle of 140°, high $Q_{monomer}$ and Q_{dimer}, all of which are consistent with the native conformation. The other basin has the $\alpha 2$-$\alpha 2$ angle of 40°, high $Q_{monomer}$ but low Q_{dimer} — consistent with the "inverted" structure. The representative structures at the two basins show that they are the native-like and the inverted non-native conformations as we found in previous annealing runs. The native basin seems to be significantly broader, suggesting entropic stabilization. Interestingly, the free energy barrier between the two states is relatively high at ∼8-9 kcal/mol.

To further test the structural stability of the newly-found non-native histone dimer, we took the final snapshot of its inverted conformation from AWSEM prediction and used that as the initial conformation to perform extensive all-atom MD simulations in explicit solvent (see details in SI). The subsequent RMSD analysis indicates that the non-native structure reached a steady state and maintained remarkable structural stability (Figure 3C), which is comparable to the simulation started from the native struture.

Sequence Symmetry Explains the Inverted Conformation

Next, we explored the molecular interactions that promote the non-native dimeric complex formation. In the native complexes, $\alpha 1$, $\alpha 2$, and $\alpha 3$ helices of one histone interact with the other histone through hydrophobic interactions (Figure 4A). We found that the above-mentioned inverted arrangement in the non-native complex still favors these interactions by swapping $\alpha 1$ and $\alpha 3$ helices. To further explain this phenomenon, we reversed the hi-

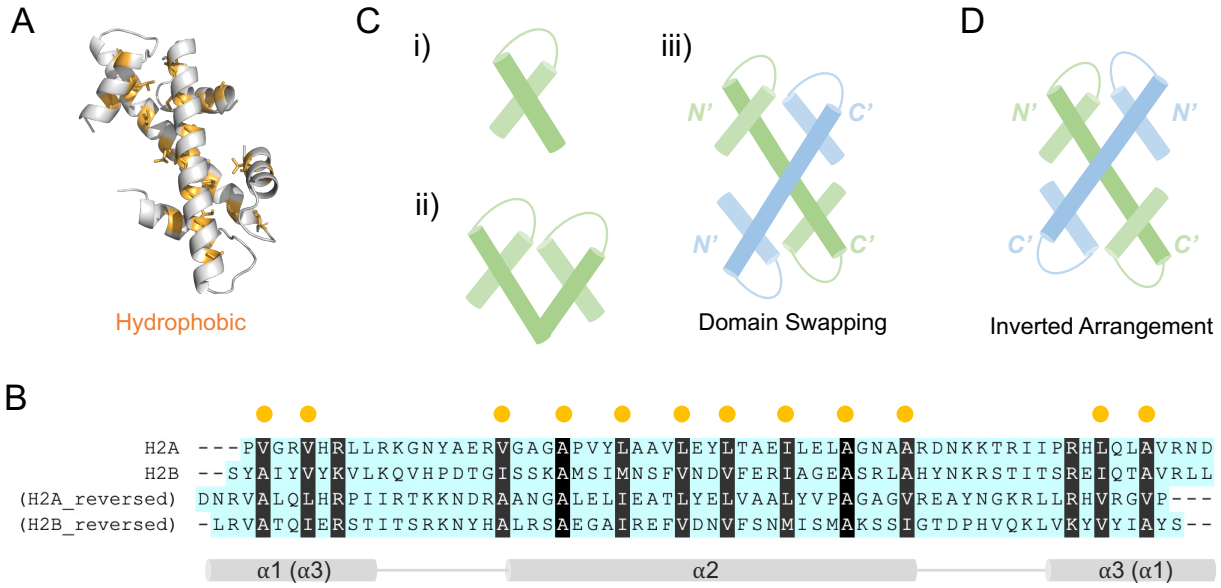


Figure 4: **Sequence symmetry of hydrophobic residues explains the predictability of inverted histone-fold structure.** (A) Conserved hydrophobic residues of H2A/H2B are shown with orange color in its native structure (silver). (B) Multiple sequence alignments of histones H2A, H2B and their sequences in a reversed order. Hydrophobic positions as in (A) are marked by orange dots on the top, and major structural elements are marked on the bottom. (C) A conceptual cartoon scheme illustrates previously proposed histone evolution hypotheses: i) histones may originate from one single helix-strand-helix structural motif; ii) duplication, differentiation and fusion of two helix-strand-helix peptides result in one protein; iii) domain swapping between two proteins (colored in green and blue) forms a histone-fold structure. (D) The inverted arrangement based on hydrophobic interactions could be an alternative formation of histone-fold structure.

stone sequences from C- to N-terminus, and carried out sequence alignments between the reversed and normal order histone sequences (Figure 4B). Strikingly, the multiple sequence alignments reveal a symmetric distribution of hydrophobic residues among histone and their reversed sequences. It is known that interactions between two histone monomers are mainly hydrophobic (Figure 4A). Thus, one may expect that in an N-C-terminal inverted arrangement, histones should still be able to maintain the required hydrophobic interactions for a stable formation due to the head-to-tail symmetry of hydrophobic positions.

This finding of sequence symmetry may be connected to the previously proposed histone evolution hypotheses. The latter posits that histones may have arisen through the duplication and differentiation of a primordial helix-strand-helix motif[7,43] (Figure 4C i-ii), and domain swapping may have triggered the subsequent dimerization of two histone monomers[26,44] (Figure 4C iii). Thus, it is possible that the conservation of hydrophobic positions found in histone sequences, no matter in a normal or reversed order, originated from the same ancestral peptide. Meanwhile, the conserved hydrophobicity in the four segments of two histones explains the rationale of two alternative ways of the crossing forming the handshake motif for the characteristic of the histone fold.

In summary, our finding of energetically favorable native and inverted dimers points to the hidden ancestral symmetry of histone protein sequences. Analogous sequence symmetry in an unrelated protein, the Rop dimer, was previously suggested to result in a double-funnel free energy landscape.[45,46] In this context, our results suggest that histone chaperone, or other molecular factors and regulators may be needed to facilitate correct folding of histone dimers, which, in turn, would ensure that the correct binding interface is exposed for their subsequent interactions with DNA and other histone proteins. In the next section we demonstrate that histone tails significantly influence the relative propensities of native and inverted conformations.

Histone Homodimer Formation and the Role of Histone Tails

AWSEM and AlphaFold2 Predict Eukaryotic Histone Homodimer

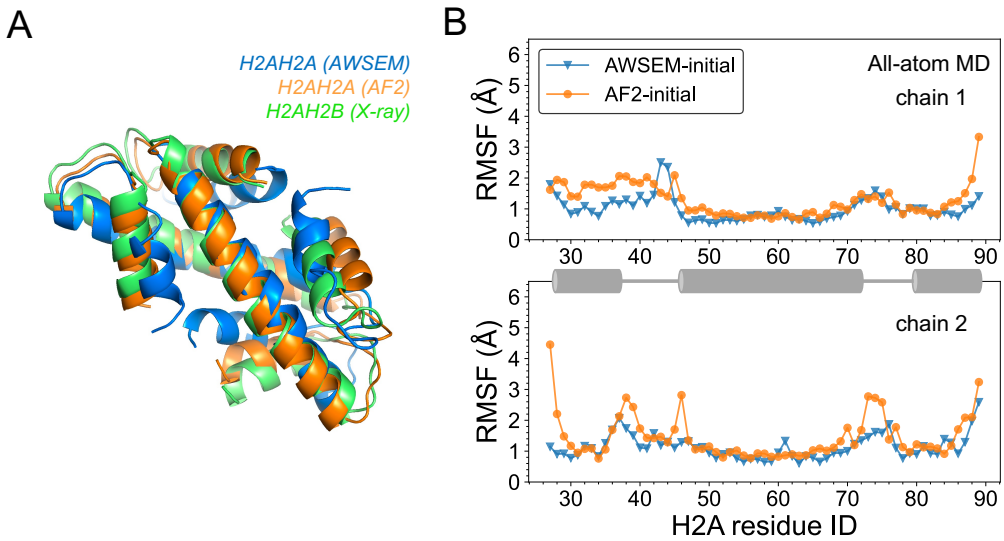


Figure 5: **Predicted structures of eukaryotic histone homodimer.** (A) AWSEM and AlphaFold2 predicted homodimer structures of H2A/H2A align well with the native heterodimer structure of H2A/H2B (colored in blue, orange, and green). (B) RMSF analysis of all-atom simulations demonstrates comparable stabilities of AWSEM- and AlphaFold2-predicted homo-complex structures (in blue and orange). RMSF of two chains are plotted separately and helix regions are animated by cartoon in grey.

We found the above observed inverted non-native histone-fold conformation in archaeal histones $(HmfA)_2$, $(HmfB)_2$, HmfA/HmfAB, and transcription factors $dTAF_{II}42/dTAF_{II}62$ dimer and NF-YB/NF-YC dimer (Figure S6). These results indicate a well-conserved folding mechnism of histone-fold structures regardless of species and function distinctions, which prompted to investigate the hypothetical eukaryotic homodimeric histones. As known, one significant difference of histone oligomers between Eukarya and Archaea is that eukaryotic histones exist as heterodimer *in vivo* while in Archaea both homodimer and heterodimer histones are prevalent. In early biochemical investigations of histone-histone interactions, sedimentation experiments showed that eukaryotic homotypic histones may exist at high histone and salt concentrations,[47,48] however, without additional structural elaboration.

Based on the above findings, we used AWSEM annealing simulations to predict the

structure of a putative H2A/H2A homodimer (histone core only, N- and C-termini were not included). Among the predicted structures, we found both native-like and the above mentioned inverted conformations. The RMSD between native-oriented H2A/H2A complex and the native H2A/H2B structure is 6.0 Å. We then applied AlphaFold-Multimer (AF2)[49,50] in ColabFold[36] to carry out analogous complex structure prediction for H2A/H2A. Interestingly, both native-oriented and inverted structures were predicted as well, consistent with our predictions from the AWSEM model. The top-scored prediction from AF2 is a native-like histone-fold structure, with a RMSD of 2.6 Å to the native heterodimer structure of H2A/H2B. The structural alignments to the H2A/H2B native structure (Figure 5A) indicate that predictions from two totally different methodologies, AWSEM and AF2, agree well. We note that we have used two models of AF2-based prediction tools. Details of the two versions' performance are discussed in the SI.

To examine the structural stability of the H2A/H2A histone homodimer, we further carried out explicit solvent all-atom simulations starting with the two initial conformations predicted by AWSEM, and AF2, respectively. We structurally aligned the simulation trajectory snapshots to the initial conformation, and calculated the root-mean-square-fluctuation (RMSF) of the C_α atoms to quantify local residual fluctuations (Figure 5B). The RMSF analysis displays comparable fluctuations between the two simulations on average within 3 Å. The RMSF analysis also shows that the loop regions of histone homodimer are in general more flexible than helical regions, as we expected.

The Role of Histone Tails in Regulating Homodimer Formation

The above results suggest that eukaryotic histone homodimers may be sufficiently stable in the absence of histone tails. This finding points to the possibility of histone tails tilting the balance between the homo- and heterodimer towards the latter. Hence, we next predicted the full-sequence histone homodimers using simulated annealing in AWSEM, and subsequently evaluated their structural similarities to the native heterodimer H2A/H2B (only for

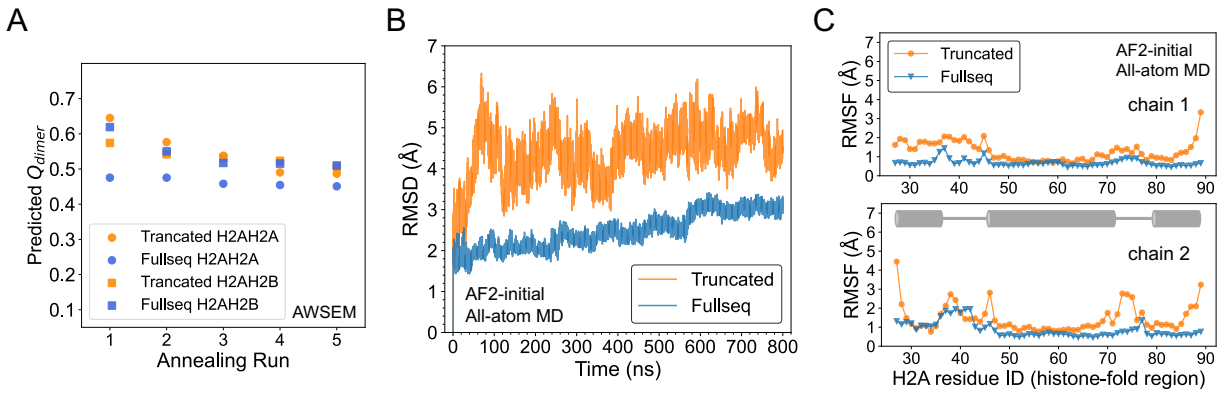


Figure 6: **Histone tails disfavor the formation of histone homodimer, yet stabilize the histone fold once formed.** (A) AWSEM-predicted trancated and full-sequence homodimer H2A/H2A are assessed by Q_{dimer} of the histone-fold region (circle in orange and blue) and compared with predictions of trancated and full-sequence heterodimer H2A/H2B as a control (square in orange and blue). (B-C) All-atom simulations of AlphaFold2-predicted truncated (orange) and full-sequence (blue) H2A/H2A homodimer are analyzed through root-mean-square-deviation (RMSD) (B) and root-mean-square-fluctuation (RMSF) (C) for the histone-fold region. In (C), the two chains of H2A/H2A are plotted separately and their helix regions are animated by a cartoon diagram.

the histone fold core part) via the structural similarity measure, Q_{dimer} (Figure 6A). From the best five predicted structures for each group, we can see that full-sequence H2A/H2A has significantly lower Q_{dimer} than H2A/H2A without histone tails (referred to "truncated H2A/H2A" hereafter), indicating a disrupting role of histone tails in the formation of the histone homodimer. This inhibitory effect of histone tails was not observed in the formation of H2A/H2B complex. Interestingly, instead, the full-length H2A/H2B shows slightly better Q_{dimer} values than what were found for truncated H2A/H2B. This result is consistent with a previous experimental study showing that H2A/H2B N-terminal tails may speed up the correct folding of H2A/H2B[40].

Moreover, we used AlphaFold2 to predict the full-sequence structure of H2A/H2A. The obtained structure is very similar to that of H2A/H2B, with a Q_{dimer} of 0.7. We then performed all-atom simulations using the predictions of AlphaFold2 for H2A/H2A homodimer with and without histone tail (truncated) as initial conformations. Interestingly, the full-sequence H2A/H2A has lower RMSD on average compared to its initial conformation

than truncated H2A/H2A (Figure 6B). RMSF analysis (Figure 6C) shows that the flexibility difference between the two systems mainly appears at the terminal residues of the histone-fold core and the loop regions. In summary, once the H2A/H2A has formed a histone-fold core, the disordered tails may potentially stabilize the histone-fold by forming interactions between tails and histone-fold core.

DISCUSSION

Biological Implication

From a biology perspective, our simulations and experiments reveal that both archaeal and eukaryotic histones cannot fold into a stable monomeric structure. Our study shows that the eukaryotic histone dimers not only inherited the main structural motif but also the folding mechanism from their ancestor proteins. However, an interesting consequence of eukaryotic histones having forked from the archaeal tree is that eukaryotic histones have tails and do not form homodimers. The length of histone tails and their post-translational modifications vary quite widely among variants found in eukaryotic organisms, playing important roles in determining the linker DNA and nucleosome interaction.[51] In this work we suggest that addition of histone tails impedes the formation of eukaryotic homotypic histones. In addition, histones' evolution not only expanded the structural variability but also dramatically enhanced the functional repertoire of histone dimers, for example in DNA repair[52], enhancer activity[53], and centromere function[54,55]. Possible new functions may arise from distinct post-translational modifications on each monomer[56–58].

In recent decades, a number of eukaryotic histone chaperones have been studied[59–63], such as Asf1 and MCM2 for H3/H4, Nap1 and MBD for H2A/H2B. Among them, Asf1 and HJURP have been reported to promote major conformational changes of histones H4 and CENP-A, respectively[29,60]. Our current finding of relatively stable inverted dimers suggests that chaperones, either protein, DNA or RNA, may be generally needed to assist proper

histone folding. Interestingly, no archaeal histone chaperone has yet been reported. Thus, from this work we anticipate the existence of archaeal histone chaperones or new chaperoning function from known proteins. However, we do not exclude the possibility that extreme ionic or temperature conditions which define the extremophile habitat, might somehow prevent the non-native association of archaeal histones. Furthermore, the ratio of native to non-native complexes varied among the different dimers that we simulated (Figure S7). This implies that the sequence and structural diversities of histones, the varying lengths of histone tails, and post-translational modifications may all affect their overall folding dynamics. For instance, previously we reported that histone monomers H3 and H4 play distinct structural and functional roles in their heterodimer H3/H4 complex[29]. It is possible that a chaperone would interact with only one histone monomer and primarily assist its folding. HJURP mainly interacting with CENP-A in the CENP-A/H4 dimer represents one such salient example. Consequently, one may anticipate close co-evolution between histone monomers and chaperones.

CONCLUSIONS

In this work, we used computational approaches supplemented with NMR and CD experiments to investigate the folding mechanism of histones. We found that a cooperative folding-upon-binding principle widely applies to canonical and variant eukaryotic histones, archaeal histones as well as histone-like transcription factors. Moreover, we report that eukaryotic histone cores without tails form a non-native dimeric complex, in which on the tertiary structure level, an inverted arrangement is energetically nearly competitive with the native conformation. Further sequence analysis indicates that this surprising non-naive stable formation could be a consequence of the ancient sequence symmetry underlying all histone proteins. Finally, to further explore non-native histone complexes, we used AWSEM, AlphaFold2 and atomistic MD simulations to predict possible formations of eukaryotic homodimers. Our re-

sults show that histone tails play an important role in regulating the formation of eukaryotic homodimer or heterodimer. Without tails, eukaryotic histones form stable homodimers in both AWSEM and AlphaFold2 predictions, while with tails AWSEM simulations indicate an impeded formation of histone homodimers whereas AlphaFold2 predicts a well-formed homodimer structure.

Acknowledgement

The authors thank Dr. Daniël Melters for helpful discussions, Dr. Carlos Castañeda for help with initial NMR measurements, and Dr. Tingting Yao for kindly providing plasmids for histone expression. This work was supported by the Amazon Web Services Artificial Intelligence Award (G.P.), the National Institutes of Health grant GM065334 (D.F.), and the Intramural Research Program of the National Institutes of Health (Y.D.). H.Z. also thanks for the NCI-UMD partnership for Integrative Cancer Research and the Ann G. Wylie Dissertation Fellowship at University of Maryland.

Supporting Information Available

The Supporting Information contains the following contents:.

- Method details of AWSEM and all-atom simulations
- Complete information and analyses for all simulated protein structures
- More results from NMR experiments
- Details and discussions about complex structure prediction

References

(1) Luger, K.; Mäder, A. W.; Richmond, R. K.; Sargent, D. F.; Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **1997**, *389*, 251.

(2) Henikoff, S.; Furuyama, T.; Ahmad, K. Histone variants, nucleosome assembly and epigenetic inheritance. *Trends in Genetics* **2004**, *20*, 320–326.

(3) Gangloff, Y.-G.; Romier, C.; Thuault, S.; Werten, S.; Davidson, I. The histone fold is a key structural motif of transcription factor TFIID. *Trends in Biochemical Sciences* **2001**, *26*, 250–257.

(4) Mattiroli, F.; Bhattacharyya, S.; Dyer, P. N.; White, A. E.; Sandman, K.; Burkhart, B. W.; Byrne, K. R.; Lee, T.; Ahn, N. G.; Santangelo, T. J.; Reeve, J. N.; Luger, K. Structure of histone-based chromatin in Archaea. *Science* **2017**, *357*, 609–612.

(5) Henneman, B.; Van Emmerik, C.; van Ingen, H.; Dame, R. T. Structure and function of archaeal histones. *PLoS genetics* **2018**, *14*, e1007582.

(6) Ramakrishnan, V. The histone fold: evolutionary questions. *Proceedings of the National Academy of Sciences* **1995**, *92*, 11328.

(7) Arents, G.; Moudrianakis, E. N. The histone fold: a ubiquitous architectural motif utilized in DNA compaction and protein dimerization. *Proceedings of the National Academy of Sciences* **1995**, *92*, 11170–11174.

(8) Isenberg, I. Histones. *Annual Review of Biochemistry* **1979**, *48*, 1304–1316.

(9) Holde, K. E. v. *Chromatin*; Springer-Verlag New York, 1989.

(10) Karantza, V.; Baxevanis, A. D.; Freire, E.; Moudrianakis, E. N. Thermodynamic studies of the core histones: ionic strength and pH dependence of H2A-H2B dimer stability. *Biochemistry* **1995**, *34*, 5988–5996.

(11) Karantza, V.; Freire, E.; Moudrianakis, E. N. Thermodynamic studies of the core histones: pH and ionic strength effects on the stability of the $(H3-H4)/(H3-H4)_2$ system. *Biochemistry* **1996**, *35*, 2037–2046.

(12) Karantza, V.; Freire, E.; Moudrianakis, E. N. Thermodynamic studies of the core histones: stability of the octamer subunits is not altered by removal of their terminal domains. *Biochemistry* **2001**, *40*, 13114–13123.

(13) Banks, D. D.; Gloss, L. M. Folding mechanism of the (H3–H4) 2 histone tetramer of the core nucleosome. *Protein Science* **2004**, *13*, 1304–1316.

(14) Zhao, H.; Winogradoff, D.; Dalal, Y.; Papoian, G. A. The oligomerization landscape of histones. *Biophysical Journal* **2019**, *116*, 1845–1855.

(15) Attar, N. et al. The histone H3-H4 tetramer is a copper reductase enzyme. *Science* **2020**, *369*, 59–64.

(16) Peng, Y.; Markov, Y.; Goncearenco, A.; Landsman, D.; Panchenko, A. R. Human Histone Interaction Networks: An Old Concept, New Trends. *Journal of Molecular Biology* **2020**,

(17) Skrajna, A.; Goldfarb, D.; Kedziora, K. M.; Cousins, E. M.; Grant, G. D.; Spangler, C. J.; Barbour, E. H.; Yan, X.; Hathaway, N. A.; Brown, N. G.; Cook, J. G.; Major, M. B.; McGinty, R. K. Comprehensive nucleosome interactome screen establishes fundamental principles of nucleosome binding. *Nucleic acids research* **2020**, *48*, 9415–9432.

(18) Grigoryev, S. A.; Arya, G.; Correll, S.; Woodcock, C. L.; Schlick, T. Evidence for heteromorphic chromatin fibers from analysis of nucleosome interactions. *Proceedings of the National Academy of Sciences* **2009**, *106*, 13317–13322.

(19) Di Pierro, M.; Cheng, R. R.; Aiden, E. L.; Wolynes, P. G.; Onuchic, J. N. De novo prediction of human chromosome structures: Epigenetic marking patterns encode genome architecture. *Proceedings of the National Academy of Sciences* **2017**, *114*, 12126–12131.

(20) Zhou, B.-R.; Jiang, J.; Ghirlando, R.; Norouzi, D.; Yadav, K. S.; Feng, H.; Wang, R.; Zhang, P.; Zhurkin, V.; Bai, Y. Revisit of reconstituted 30-nm nucleosome arrays reveals an ensemble of dynamic structures. *Journal of molecular biology* **2018**, *430*, 3093–3110.

(21) Stevens, K. M.; Swadling, J. B.; Hocher, A.; Bang, C.; Gribaldo, S.; Schmitz, R. A.; Warnecke, T. Histone variants in archaea and the evolution of combinatorial chromatin complexity. *Proceedings of the National Academy of Sciences* **2020**, *117*, 33384–33395.

(22) Bowerman, S.; Wereszczynski, J.; Luger, K. Archaeal chromatin 'slinkies' are inherently dynamic complexes with deflected DNA wrapping pathways. *bioRxiv* **2020**,

(23) Dima, R. I.; Thirumalai, D. Asymmetry in the shapes of folded and denatured states of proteins. *The Journal of Physical Chemistry B* **2004**, *108*, 6564–6570.

(24) Davtyan, A.; Schafer, N. P.; Zheng, W.; Clementi, C.; Wolynes, P. G.; Papoian, G. A. AWSEM-MD: protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing. *The Journal of Physical Chemistry B* **2012**, *116*, 8494–8503.

(25) Zheng, W.; Schafer, N. P.; Davtyan, A.; Papoian, G. A.; Wolynes, P. G. Predictive energy landscapes for protein–protein association. *Proceedings of the National Academy of Sciences* **2012**, *109*, 19244–19249.

(26) Alva, V.; Ammelburg, M.; Söding, J.; Lupas, A. N. On the origin of the histone fold. *BMC Structural Biology* **2007**, *7*, 17.

(27) Onuchic, J. N.; Luthey-Schulten, Z.; Wolynes, P. G. Theory of protein folding: the

energy landscape perspective. *Annual Review of Physical Chemistry* **1997**, *48*, 545–600.

(28) Papoian, G. A.; Wolynes, P. G. In *Coarse-Grained Modeling of Biomolecules*; Papoian, G. A., Ed.; CRC Press, Taylor & Francis Group: Boca Raton, FL, 2017; Chapter 4, pp 121–189.

(29) Zhao, H.; Winogradoff, D.; Bui, M.; Dalal, Y.; Papoian, G. A. Promiscuous histone mis-assembly is actively prevented by chaperones. *Journal of the American Chemical Society* **2016**, *138*, 13207–13218.

(30) Wu, H.; Wolynes, P. G.; Papoian, G. A. AWSEM-IDP: a coarse-grained force field for intrinsically disordered proteins. *The Journal of Physical Chemistry B* **2018**, *122*, 11115–11125.

(31) Zhang, B.; Zheng, W.; Papoian, G. A.; Wolynes, P. G. Exploring the free energy landscape of nucleosomes. *Journal of the American Chemical Society* **2016**, *138*, 8126–8133.

(32) Wu, H.; Dalal, Y.; Papoian, G. A. Binding dynamics of disordered linker histone H1 with a nucleosomal particle. *Journal of Molecular Biology* **2021**, *433*, 166881.

(33) Kirkpatrick, S.; Gelatt, C. D.; Vecchi, M. P. Optimization by simulated annealing. *Science* **1983**, *220*, 671–680.

(34) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *Journal of Computational Chemistry* **1992**, *13*, 1011–1021.

(35) Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L.-P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D., et al. OpenMM 7: Rapid

development of high performance algorithms for molecular dynamics. *PLoS computational biology* **2017**, *13*, e1005659.

(36) Mirdita, M.; Schütze, K.; Moriwaki, Y.; Heo, L.; Ovchinnikov, S.; Steinegger, M. ColabFold-Making protein folding accessible to all. **2021**,

(37) Steinegger, M.; Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology* **2017**, *35*, 1026–1028.

(38) Whitmore, L.; Wallace, B. A. Protein secondary structure analyses from circular dichroism spectroscopy: methods and reference databases. *Biopolymers: Original Research on Biomolecules* **2008**, *89*, 392–400.

(39) Andrade, M.; Chacon, P.; Merelo, J.; Moran, F. Evaluation of secondary structure of proteins from UV circular dichroism spectra using an unsupervised learning neural network. *Protein Engineering, Design and Selection* **1993**, *6*, 383–390.

(40) Placek, B. J.; Gloss, L. M. Three-state kinetic folding mechanism of the H2A/H2B histone heterodimer: the N-terminal tails affect the transition state between a dimeric intermediate and the native dimer. *Journal of Molecular Biology* **2005**, *345*, 827–836.

(41) Fushman, D.; Cahill, S.; Cowburn, D. The main-chain dynamics of the dynamin pleckstrin homology (PH) domain in solution: analysis of 15N relaxation with monomer/dimer equilibration 1. *Journal of Molecular Biology* **1997**, *266*, 173–194.

(42) Placek, B. J.; Gloss, L. M. The N-terminal tails of the H2A- H2B histones affect dimer structure and stability. *Biochemistry* **2002**, *41*, 14960–14968.

(43) Arents, G.; Moudrianakis, E. N. Topography of the histone octamer surface: repeating structural motifs utilized in the docking of nucleosomal DNA. *Proceedings of the National Academy of Sciences* **1993**, *90*, 10489–10493.

(44) Lupas, A. N.; Ponting, C. P.; Russell, R. B. On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? *Journal of structural biology* **2001**, *134*, 191–203.

(45) Levy, Y.; Cho, S. S.; Shen, T.; Onuchic, J. N.; Wolynes, P. G. Symmetry and frustration in protein energy landscapes: A near degeneracy resolves the Rop dimer-folding mystery. *Proceedings of the National Academy of Sciences* **2005**, *102*, 2373–2378.

(46) Wolynes, P. G. Symmetry and the energy landscapes of biomolecules. *Proceedings of the National Academy of Sciences* **1996**, *93*, 14249.

(47) Kornberg, R. D.; Thomas, J. O. Chromatin structure: oligomers of the histones. *Science* **1974**, *184*, 865–868.

(48) Sperling, R.; Bustin, M. Dynamic equilibrium in histone assembly. Self-assembly of single histones and histone pairs. *Biochemistry* **1975**, *14*, 3322–3331.

(49) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A., et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589.

(50) Evans, R.; O'Neill, M.; Pritzel, A.; Antropova, N.; Senior, A. W.; Green, T.; Žídek, A.; Bates, R.; Blackwell, S.; Yim, J., et al. Protein complex prediction with AlphaFold-Multimer. *BioRxiv* **2021**,

(51) Zalenskaya, I. A.; Zalensky, A. O.; Zalenskaya, E. O.; Vorob'ev, V. I. Heterogeneity of nucleosomes in genetically inactive cells. *FEBS letters* **1981**, *128*, 40–42.

(52) Giaimo, B. D.; Ferrante, F.; Herchenröther, A.; Hake, S. B.; Borggrefe, T. The histone variant H2A. Z in gene regulation. *Epigenetics & chromatin* **2019**, *12*, 1–22.

(53) Das, C.; Tyler, J. K. Histone exchange and histone modifications during transcription

and aging. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* **2012**, *1819*, 332–342.

(54) Henikoff, S.; Smith, M. M. Histone variants and epigenetics. *Cold Spring Harbor perspectives in biology* **2015**, *7*, a019364.

(55) Jin, J.; Cai, Y.; Li, B.; Conaway, R. C.; Workman, J. L.; Conaway, J. W.; Kusch, T. In and out: histone variant exchange in chromatin. *Trends in biochemical sciences* **2005**, *30*, 680–687.

(56) Bannister, A. J.; Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Research* **2011**, *21*, 381–395.

(57) Potoyan, D. A.; Papoian, G. A. Regulation of the H4 tail binding and folding landscapes via Lys-16 acetylation. *Proceedings of the National Academy of Sciences* **2012**, *109*, 17857–17862.

(58) Winogradoff, D.; Echeverria, I.; Potoyan, D. A.; Papoian, G. A. The acetylation landscape of the H4 histone tail: disentangling the interplay between the specific and cumulative effects. *Journal of the American Chemical Society* **2015**, *137*, 6245–6253.

(59) Hammond, C. M.; Strømme, C. B.; Huang, H.; Patel, D. J.; Groth, A. Histone chaperone networks shaping chromatin function. *Nature reviews Molecular cell biology* **2017**, *18*, 141–158.

(60) English, C. M.; Adkins, M. W.; Carson, J. J.; Churchill, M. E.; Tyler, J. K. Structural basis for the histone chaperone activity of Asf1. *Cell* **2006**, *127*, 495–508.

(61) Foltz, D. R.; Jansen, L. E.; Bailey, A. O.; Yates III, J. R.; Bassett, E. A.; Wood, S.; Black, B. E.; Cleveland, D. W. Centromere-specific assembly of CENP-a nucleosomes is mediated by HJURP. *Cell* **2009**, *137*, 472–484.

(62) Nye, J.; Sturgill, D.; Athwal, R.; Dalal, Y. HJURP antagonizes CENP-A mislocalization driven by the H3. 3 chaperones HIRA and DAXX. *PloS One* **2018**, *13*.

(63) Melters, D. P.; Pitman, M.; Rakshit, T.; Dimitriadis, E. K.; Bui, M.; Papoian, G. A.; Dalal, Y. Intrinsic elasticity of nucleosomes is encoded by histone variants and calibrated by their binding partners. *Proceedings of the National Academy of Sciences* **2019**, *116*, 24066–24074.